# Integration of machine learning techniques in chemometrics practices

Politecnico di Bari

# LIBERATORIA PER L'ARCHIVIAZIONE DELLA TESI DI DOTTORATO

Al Magnifico Rettore
del Politecnico di Bari

Il sottoscritto Maurizio Triggiani nato a Bari il 30/07/1985 residente a Bari in via Abate Giacinto Gimma, 140 e-mail maurizio.triggiani@poliba.it iscritto al 3° anno di Corso di Dottorato di Ricerca in Ingegneria Elettrica e dell'Informazione ciclo XXXIV ed essendo stato ammesso a sostenere l'esame finale con la prevista discussione della tesi dal titolo: "Integration of machine learning techniques in chemometrics practices"

**DICHIARA**

1) di essere consapevole che, ai sensi del D.P.R. n. 445 del 28.12.2000, le dichiarazioni mendaci, la falsità negli atti e l'uso di atti falsi sono puniti ai sensi del codice penale e delle Leggi speciali in materia, e che nel caso ricorressero dette ipotesi, decade fin dall'inizio e senza necessità di nessuna formalità dai benefici conseguenti al provvedimento emanato sulla base di tali dichiarazioni;
2) di essere iscritto al Corso di Dottorato di ricerca Ingegneria Elettrica e dell'Informazione ciclo XXXIV, corso attivato ai sensi del "*Regolamento dei Corsi di Dottorato di ricerca del Politecnico di Bari"*, emanato con D.R. n.286 del 01.07.2013;
3) di essere pienamente a conoscenza delle disposizioni contenute nel predetto Regolamento in merito alla procedura di deposito, pubblicazione e autoarchiviazione della tesi di dottorato nell'Archivio Istituzionale ad accesso aperto alla letteratura scientifica;
4) di essere consapevole che attraverso l'autoarchiviazione delle tesi nell'Archivio Istituzionale ad accesso aperto alla letteratura scientifica del Politecnico di Bari (IRIS-POLIBA), l'Ateneo archivierà e renderà consultabile in rete (nel rispetto della Policy di Ateneo di cui al D.R. 642 del 13.11.2015) il testo completo della tesi di dottorato, fatta salva la possibilità di sottoscrizione di apposite licenze per le relative condizioni di utilizzo (di cui al sito http://www.creativecommons.it/Licenze), e fatte salve, altresì, le eventuali esigenze di "embargo", legate a strette considerazioni sulla tutelabilità e sfruttamento industriale/commerciale dei contenuti della tesi, da rappresentarsi mediante compilazione e sottoscrizione del modulo in calce (Richiesta di embargo);
5) che la tesi da depositare in IRIS-POLIBA, in formato digitale (PDF/A) sarà del tutto identica a quelle **consegnate**/inviate/da inviarsi ai componenti della commissione per l'esame finale e a qualsiasi altra copia depositata presso gli Uffici del Politecnico di Bari in forma cartacea o digitale, ovvero a quella da discutere in sede di esame finale, a quella da depositare, a cura dell'Ateneo, presso le Biblioteche Nazionali Centrali di Roma e Firenze e presso tutti gli Uffici competenti per legge al momento del deposito stesso, e che di conseguenza va esclusa qualsiasi responsabilità del Politecnico di Bari per quanto riguarda eventuali errori, imprecisioni o omissioni nei contenuti della tesi;
6) che il contenuto e l'organizzazione della tesi è opera originale realizzata dal sottoscritto e non compromette in alcun modo i diritti di terzi, ivi compresi quelli relativi alla sicurezza dei dati personali; che pertanto il Politecnico di Bari ed i suoi funzionari sono in ogni caso esenti da responsabilità di qualsivoglia natura: civile, amministrativa e penale e saranno dal sottoscritto tenuti indenni da qualsiasi richiesta o rivendicazione da parte di terzi;
7) che il contenuto della tesi non infrange in alcun modo il diritto d'Autore né gli obblighi connessi alla salvaguardia di diritti morali od economici di altri autori o di altri aventi diritto, sia per testi, immagini, foto, tabelle, o altre parti di cui la tesi è composta.

Bari, 11/04/2022 Firma _____

Il sottoscritto, con l'autoarchiviazione della propria tesi di dottorato nell'Archivio Istituzionale ad accesso aperto del Politecnico di Bari (POLIBA-IRIS), pur mantenendo su di essa tutti i diritti d'autore, morali ed economici, ai sensi della normativa vigente (Legge 633/1941 e ss.mm.ii.),

**CONCEDE**

- al Politecnico di Bari il permesso di trasferire l'opera su qualsiasi supporto e di convertirla in qualsiasi formato al fine di una corretta conservazione nel tempo. Il Politecnico di Bari garantisce che non verrà effettuata alcuna modifica al contenuto e alla struttura dell'opera.
- al Politecnico di Bari la possibilità di riprodurre l'opera in più di una copia per fini di sicurezza, back-up e conservazione.

Bari, 11/04/2022 Firma _____

Politecnico di Bari

Department of Electrical and Information Engineering

ELECTRICAL AND INFORMATION ENGINEERING Ph.D. Program

SSD: ING-INF/05 – Information Processing Systems

**Final Dissertation**

# Integration of machine learning techniques in chemometrics practices

by

Maurizio Triggiani

**Supervisors**:

Prof. Tommaso Di Noia

Prof. Vito Gallo

*Coordinator of Ph.D. Program:*

*Prof. Mario Carpentieri*

*Course n°34, 01/11/2018-31/01/2022*

Politecnico di Bari

Department of Electrical and Information Engineering
ELECTRICAL AND INFORMATION ENGINEERING Ph.D. Program
SSD: ING-INF/05 – Information Processing Systems

**Final Dissertation**

# Integration of machine learning techniques in chemometrics practices

by
Maurizio Triggiani

_____
*Firma leggibile e per esteso*

**Referees**:

Prof. Danilo Caivano

Prof. Cristina Airoldi

**Supervisors:**

Prof. Tommaso Di Noia

_____
*Firma leggibile e per esteso*

Prof. Vito Gallo

_____
*Firma leggibile e per esteso*

*Coordinator of Ph.D. Program:*
*Prof. Mario Carpentieri*

_____
*Firma leggibile e per esteso*

*Course n°34, 01/11/2018-31/01/2022*

# Sommario

## Abstract

Food safety is a key objective in all the development plans of the European Union. To ensure the quality and the sustainability of the agricultural production (both intensive and extensive) a well-designed analysis strategy is needed. Climate change, precision agriculture, green revolution and industry 4.0 are areas of study that need innovative practices and approaches that aren't possible without precise and constant process monitoring. The need for product quality assessment during the whole supply chain is paramount and cost reduction is also another constant need. Non targeted Nuclear Magnetic Resonance (NMR) analysis is still a second-choice approach for food analysis and monitoring, one of the problems of this approach is the big amount of information returned. This kind of data needs a new and improved method of handling and analysis.

Classical chemometrics practices are not well suited for this new field of study. In this thesis, we approached the problem of food fingerprinting and discrimination by the means of non-targeted NMR spectroscopy combined with modern machine learning algorithms and databases meant for the correct and easy access of data. The introduction of machine learning techniques alongside the clear benefits introduces a new layer of complexity regarding the need for trusted data sources for algorithm training and integrity, if this kind of approach proves is worth in the global market, we'll need not only to create a good dataset, but we'll need to be prepared to defend against also more clever attacks like adversarial machine learning attacks.

Comparing the machine learning results with the classic chemometric approach we'll highlight the strengths and the weakness of both approaches, and we'll use them to prepare the framework needed to tackle the challenges of future agricultural productions.

# Introduction

Producing food in a developed country today is no easy task, from the seed plantation to the final industrial transformation each product must be monitored, tested, verified and approved. Each step of the chain adds costs and needs optimization, speed is key ingredient to maintain the production profitable and competitive against other producers especially against foreign player with a different and less strict set of regulations. Each step of the chain usually requires a laboratory test to verify the components of a sample, so affordable and reliable testing are needed to guarantee an active and healthy production chain, chemometrics is the field of chemistry that explores the mathematical instruments needed to evaluate the information gathered during those experiments, historically in chemometrics machine learning algorithms are not used in favor of more simple regression-based algorithms like PCA or PLS.[1], [2]

The food production is a process that requires a rich mosaic of products, techniques, and raw materials to create the perfect balance in food quality and economic return. One example of this complex environment is the importance of soil organic matter. Farmers use exogenous sources of organic matter and apply management practices to minimize SOM decay[3]. But despite this awareness, intensive agricultural practices have resulted in a decline in soil fertility and SOM across European regions. Decreases in SOM appear mainly because of intensive arable cropping systems and an underestimation of the relevance of soil organic matter.[4], [5]

Therefore, it is necessary that future agricultural technologies be compatible with solutions to such problems from those of conventional agricultural technologies. [6] In this regard P.A.S.C.Qua project (Produzioni Agricole Sostenibili Compost Qualità) focus on development of the chemometric supports, using nuclear magnetic resonance (NMR) spectroscopy on the agri-food samples, in order to correlate metabolic characteristics of the fruit and vegetable product to the agronomic practice adopted for its production. In the study, 1H-NMR data with chemometric[7] [8]methods were coupled, comparing multivariate analysis (Principal Component Analysis (PCA), Partial Least Squares Discriminant Analysis (PLS-DA)) and with expert systems [8]such as Decision tree, Random Forest and Artificial neural networks (ANN).

Another field of application of analytics is food fingerprinting, by studying the metabolic profile of food product it's possible to track the most important metabolomic characteristic and distinguish between

the same kind of product. The ability of recognize a specific family of product in a mix of other of the same kind is a key feature in the fraud prevention and quality certifications enforcement, labels like DOC or DOP for Italy for example are a great business and subject of fraud. A study with excellent results has been conducted on grapevines (Vitis sp.) that are one of the most important fruit species worldwide due to the use of their fruit in the production of wine. Grapes are also popular as fresh table grapes or dried as currants and raisins. The Vitaceae family consists of almost 1000 species, and among them Vitis vinifera ssp. vinifera is currently the most cultivated around the world [9].

Consequently, grapevines are characterized by high biodiversity and in Italy, over the past decade, research and experimentation in agriculture has focused attention on biodiversity safeguard and genetic improvement, considering them strategic topics for the greater typicality and for the improvement of local wine production.

NMR spectroscopy, offers direct identification and quantification of a broad range of metabolites. The advantages of NMR include potential for minimal or no sample pretreatment (other than clarification and pH adjustment), non-destructive analysis, the suitability for quantitative analysis and the ability to advance the structural identification of unknown primary and secondary metabolites with a single measure. [2], [4]

After spectra NMR acquisition and relative pre-processing [10], statistical comparisons between biological replicates, treatments and identification of relevant metabolites usually involves exploration of the dataset through clustering (principal component analysis, PCA), classification techniques (partial least squares discriminant analysis, PLS-DA, orthogonal-PLS, O-PLS), using non-targeted [2], [4] and/or targeted metabonomic approaches. Unbiased approaches of metabolomics are now providing thorough information about many different groups of compounds, also in grape juices and wines, and are therefore more advantageous than traditional targeted analysis [11] .

In particular, the fingerprints of the grape varieties from autochthonous grapevines, to establish a classification/prediction model, an non targeted metabonomic approach was performed by employing multivariate statistical analysis (PCA and PLS-DA) and specific classifiers [8] such as Decision tree, Random Forest and Artificial neural networks (ANN).

This work will be focused on the technique need to establish a concrete bridge between classical chemometrics and new approach based on the evolution of the machine learning techniques. In the first

part we'll explore the needed prerequisites that needs to be met to create a consistent and representative dataset starting with a classic NMR laboratory experiment, then we'll study some practical application of the machine learning techniques comparing their result with the classic conclusion of chemometrics procedures.

# Materials

Building a good classifier requires not only good data but also a deep understanding of the generation of the data itself. In this work we used different kind of products that needed to be analyzed via Nuclear Magnetic Resonance [12].

The preparation and analysis of the sample is in the domain of competence of the chemists that need to prepare the sample and analyze it. The procedure to create a good sample to be analyzed is a really complex subject on its own, but to understand this work is essential to understand at least the basics of the process of the sample preparation and also how the NMR analysis works to produce the data we use to feed the algorithms.

This first section will be used to explain and show how chemist works to obtain a good and repeatable measure. We'll start with a basic explanation to create an understanding of the process then go into detail in the next chapters.

A measurement starts with the raw material, the sample, it can be of any kind, food, minerals, pieces of cloth or biological samples. Regardless of the nature of the sample a precise cataloguing of the sample know properties is needed. In this work we'll focus on food so basic information on the variety of the product or the geographical origin are some of the classic metadata collected. [13]–[15]

After the cataloguing, the sample it needs to be 'prepared', the preparation process is a strict procedure, precisely tailored for each product and element. This procedure is usually well described in the bibliography associated with the product, and for new products is needed to create a specified process. [16]This process also can variate deeply according to the information we need to extract from the sample.

When the sample is ready is put in the test tube and then it can be analyzed by the NMR spectrometer, here the sample, simply speaking, is bombarded with know magnetic signals and with the analysis of the response is possible to determine the composition of the sample.

The obtained spectrum also needs a manual correction from an operator, then it's "quantized" or "bucketed" to be finally used as a single row in our training set.

These steps have to be repeated for each sample, usually multiple times to ensure a good measure, it's easy to understand how difficult and expensive is to build a good and reliable dataset. In this work we

attempted to create a good model to export the data and use the to feed classifiers usually not used in the chemometric space.

## *Targeted and not-targeted analysis*

In order to better understand how the sample investigation work, we need to differentiate between targeted and non-targeted analysis. [17] The non-targeted analysis consists of data generation aimed at providing a comprehensive description of the complex matrix under investigation, allowing for the detection of known and unknown compounds in the mixture. This kind of analysis provides a fingerprint of the sample that may be primarily intended not for the identification of the analytes, but for the definition of a pattern to trace the sample and to identify it by suitable classification tools. If appropriate, the number of data coming from the spectral analysis can be still examined to identify the nature and the features of each analyte contained in the mixture. Despite the clear advantages offered by this analytical approach, there are crucial issues to be faced [18], [19]. One of the main drawbacks of the non-targeted approach is the lack of official guidelines governing the standardization and harmonization of the method development and validation [20] [21]. Only recently, the United States Pharmacopeia (USP) introduced some first guidelines to assist scientists in the development and validation of nontargeted methods, from sample preparation to spectral data management (US Pharmacopoea). On the other hand, targeted analysis focuses the preparation and analysis only to assess the presence and the quantity of a given element is present in the sample. This kind of approach usually is really precise but can't extract 'new information' form the sample, therefore is not suitable for the building of a dataset aimed to train a classifier.

## The metabolomics, how a food sample is composed

The rapidly emerging field of metabolomics combines strategies to identify and quantify cellular metabolites using sophisticated analytical technologies with the application of statistical and multi-

variant methods for information extraction and data interpretation. In the last two decades, huge progress was made in the sequencing of a number of different organisms.

Simultaneously, large investments were made to develop analytical approaches to analyze the different cell products, such as those from gene expression (transcripts), proteins, and metabolites. All of these so-called 'omics' approaches, including genomics, transcriptomics, proteomics, and metabolomics, are considered important tools to be applied and utilized to understand the biology of an organism and its response to environmental stimuli or genetic perturbation.

Metabolites are considered to "act as spoken language, broadcasting signals from the genetic architecture and the environment"[22], and therefore, metabolomics is considered to provide a direct "functional readout of the physiological state" of an organism [23]. A range of analytical technologies has been employed to analyze metabolites in different organisms, tissues, or fluids [24]. fasts spectrometry coupled to different chromatographic separation techniques, such as liquid or gas chromatography or NftR, are the major tools to analyze a large number of metabolites simultaneously. Although the technology is highly sophisticated and sensitive, there are still a few bottlenecks in metabolomics. Due to the huge diversity of chemical structures and the large differences in abundance, there is no single technology available to analyze the entire metabolome. Therefore, several complementary approaches must be established for extraction, detection, quantification, and identification of as many metabolites as possible .

Another challenge in metabolomics is to extract the information and interpret it in a biological context from the vast amount

of data produced by high-throughput analyzers. The application of sophisticated statistical and multi-variant data analysis tools, including cluster analysis, pathway mapping, comparative overlays, and heatmaps, has not only been an exciting and steep learning process for biochemists, but has also demonstrated that current thinking needs to change to deal with large data sets and distinguish between noise and real sample-related information.

## Sample preparation

The first step for an analysis is the sample preparation, due to the fine nature of the NMR analysis is needed a really disperse and pure sample, in the metabolomics experiment, sampling provides a picture or snapshot of the metabolome at one point in time, although NMR measurements that have been reported recently do not require sampling before analysis [25]. The requirement of sampling and sample preparation that is not biased towards groups of metabolites provides challenges, which currently have not been fully resolved. The time and method of sampling can greatly influence the reproducibility of the analytical sample. Diurnal and dietary influences can have major effects on the composition of the metabolome [26], [27], as can the section of a plant sampled [28]. Finally, the storage of samples is important, as the continued freeze/thawing of samples can be detrimental to stability and composition [29]. All these influence the precision, accuracy and reproducibility of results.

Strategies of sampling and sample preparation vary. Both invasive and non-invasive sampling can be performed. Extra-cellular metabolites, such as metabolic footprint or urine, depict a picture over a period of metabolic activity and are normally stored at low temperatures to inhibit metabolic reactions. The extraction of intra-cellular metabolites provides a snapshot of the metabolome, can be time consuming, and is subject to certain difficulties when compared to other sampling strategies. Metabolic processes are rapid (reaction times less than 1 s), so rapid inhibition of enzymatic processes is required, generally by freeze clamping or freezing in liquid nitrogen after harvesting, and subsequent storage at −80°C. Freezing provides specific issues, such as loss of metabolites [30]. The application of acidic treatments using perchloric or nitric acid has been used but can result in a severe reduction in the number of metabolites detected and degradation compounds not stable at extreme pH. Polar/non-polar extractions are the most frequently applied method and are performed by physical/chemical disruption of the cells, removal of the cell pellet by centrifugation and distribution of metabolites to polar (methanol/water) and non-polar (chloroform) solvents. Hot alcoholic extractions are also performed. Metabolic fingerprinting has recently been used in microbial metabolomics to exclude the extraction procedure and provide rapid, high-throughput sampling . Here, metabolites naturally excreted from intra-cellular volumes to the extra-cellular supernatant are analysed. Sampling and collection of volatile compounds from plants has been discussed elsewhere [31], and a procedure for extraction and separation of metabolome, proteome and transcriptome has also been reported [32].

Further sample preparation depends on the sample and metabolomics strategy employed. In many applications, no further isolation of metabolites from the sample matrix is performed, and samples are diluted and analysed directly or analysed after chemical derivatisation. However, some isolation of metabolites can be undertaken, especially with biofluids, where an initial preparation stage is protein precipitation with organic solvents. Further isolation from the sample matrix can be used including solid phase extraction (SPE) or liquid–liquid extraction (LLE). These can especially be observed in the analysis of pharmaceuticals and related metabolites or for metabolite target analysis.

## Needed steps to assure repeatability

To be sure to correctly classify our future samples we need to be sure that the data obtained by the NMR measurement will be compatible with the previous investigation even if the analysis was made years before. In this context, recently, by means of interlaboratory comparisons (ILCs), we demonstrated that NMR spectroscopy can provide statistically equivalent signals when the same sample is analyzed by spectrometers that are different in terms of magnetic field strength, manufacturer, hardware configurations and age [1], [17] . Indeed, the exclusive correlation between the resonance frequency of a signal and the type of nuclei associated to that signal, makes NMR spectroscopy a powerful technique for structural determination and quantification. Since the area of a NMR signal is linearly proportional to the number of NMR active nuclei generating the signal, the response factor (ratio between the signal produced by the analyte and the quantity of analyte which produces the signal) is independent of the molecule and the analyte quantification can be achieved directly by calculating integral of the NMR signal [33] . Moreover, the design of new pulse sequences for FIDs acquisition [34], [35] and novel algorithms for data processing enhanced the capability of NMR for discriminating among very similar compounds contained in complex mixtures, as pharmaceutical, natural products, agrochemicals, foodstuff, and biofluids. Nevertheless, to date few official protocols have been reported which employ NMR technique for purity assessment and quantification purposes [36]. While the experimental conditions (pulse sequence, acquisition parameters, postprocessing strategy) assuring the intra-laboratory repeatability are well established [37], [38], still few studies are available that discuss the reproducibility assessment of quantitative NMR (qNMR) data obtained when the same sample is

analysed by different operators and/or by spectrometers with variable features (manufacturer, $B_0$ field strength)[39]–[41]. Taking into consideration the extensive application of qNMR in different fields of chemical science, it appears as a matter of urgency to overcome this significant shortcoming and make qNMR an internationally accepted standard analytical technique.

## *The NMR machine for sample analysis*

Almost all the data elaborated in this thesis has been processed within the Polytechnic of Bari NMR spectrometer, a Bruker Avance 400 spectrometer Figure 1. This machine is composed with several elements, the most recognizable is the huge Magnet Dewar where the actual sample is placed and investigated. The strength of the magnet is graded according to the frequency of the NMR signals emitted by hydrogen atoms. The stronger the magnet field, the higher this hydrogen frequency [42]. For example, with a 500 MHz magnet (11.7 T), when a chemical sample is placed in the magnet for analysis, the 1H atoms in the sample will emit signals with a frequency very close to 500 MHz. Magnets are available in the range of 200-1000 MHz.

Superconducting magnets are electromagnets, and as such make use of the fact that an electric current produces a magnetic field. The magnet core consists of a large coil of current carrying wire in the shape of a solenoid. At the center of the coil a very intense static magnetic field exists. The sample to be analyzed is placed inside this magnetic field.

At very low temperatures certain materials show the remarkable property of superconductivity. A superconducting wire carries electricity without the need for any driving energy (i.e. battery or mains supply). Once a current is started in a superconducting loop it will continue forever. Magnets consist of such a superconducting loop. The only maintenance required on the magnet is to ensure that the coil is kept immersed in liquid helium.

The magnet consists of several sections. The outer casing of the magnet is evacuated and inner surfaces are silvered (this is the same principle as a Thermos). Next comes a bath of nitrogen which reduces the temperature to 77.35K (-195.8°C) and finally a tank of helium in which the superconducting coil is immersed in. This tank is thermally isolated against the nitrogen bath by a second evacuated section.

This magnet is further controlled by a shim system mounted into the lower end of the magnet, is a set of current carrying coils (known as shims) used to maximize field homogeneity [43] by offsetting any existing in homogeneities. The currents in these room temperature shims (so called as they are not cooled by being immersed in a bath of liquid helium) are controlled by the Smart Magnet System and may be adjusted from the Smart Magnet System display to optimize the NMR signal. This has a major effect on signal resolution and sensitivity. This action of adjusting the currents in the room temperature shims is referred to as shimming the magnet.

The sample is held in the magnet by the probe designed to transmit radio frequency signals which excite the sample and receive the emitted response. The transmission and reception is achieved by using specially designed RF coils.

The probe is inserted into the bottom of the magnet and sits inside the room temperature shims. Coaxial cables carry the excitation signals from the console amplifiers to the probe and the NMR signal back from the sample to the receiver.

Probes come in different sizes and types. The size of the probe is given in terms of the sample tube sizes it can hold, with 5mm and 10mm sample tube diameters the most popular. Different types of probes are used depending upon the type of experiment. Selective probes are specially designed to observe specific nuclei, e.g. 13C, while multinuclear (X-BB or broadband) probes may be used to analyze a wide range of nuclei. The number and design of the internal coils are what physically distinguishes one type of probe from another. In addition, the outer diameter and length of the probe is built to the specifications of the various magnets.


**Figure 1. The Bruker Avance 400 spectrometer we used in most of our experiments**

Signals enter and leave the coils of the probe via connectors which are clearly labeled and located at the base of the probe. [44], [45]The same cable is used to carry the signal to and from the probe. Each probe has an inner coil (the observe coil). This coil is located closest to the sample volume to maximize sensitivity. The color coding of the inner coil BNC follows a simple rule. It always has the same color as the rectangular strip located directly above the BNC connectors.

Another important component of the machine is the control console in the NMR spectrometer we used is composed by a TwoBay unit and it houses most of the electronic hardware associated with a modern digital spectrometer. The principal units are the IPSO (Intelligent Pulse Sequence Organizer), the BSMS (Bruker Smart Magnet System), the VTU (Variable Temperature Unit) as well as various amplifiers.

- **IPSO**: The various units within the IPSO generate the radio frequency pulses used to excite the sample and receive, amplify and digitize the NMR signals emitted by the sample. Once the data is received and digitized, the information is transferred to the host computer for further

processing and storage. The main link with the host computer is via Ethernet. It is important to emphasize that the IPSO has total control over spectrometer operation within the duration of an experiment. This is to ensure uninterrupted operation and so guarantee the integrity of the acquisition. The rack contains a set of digital and analogue slot-in type boards that prepare the signal to be transmitted and receive, amplify and digitize the NMR signal. A detailed description of these boards is beyond the scope of this manual.

- **BSMS**: This system is controlled via software and is used to operate the lock and shim system as well as controlling the sample lift and spin.
- **VTU**: Its function is to vary the sample temperature in a controlled manner or maintain it at a constant value.

Amplifiers, also known as Transmitters. Signals of relatively large amplitude are often required to excite the NMR sample and hence the need for amplifiers. Amplifiers can be internal (incorporated into the IPSO rack) or external (separate stand-alone units). Cables running directly from the amplifier outputs to the HPPR (High Performance Preamplifier) [46] carry the RF signal to the sample. Although there is a wide range of available amplifiers (including solid amplifiers) the two main categories are:
Selective amplifiers (also known as 1H or proton amplifiers) are specifically designed to amplify the higher frequencies associated with 1H and 19F.
Broadband amplifiers (also known as X amplifiers) are designed to amplify a wide range of frequencies (excluding 1H and 19F).

## How NMR works

NMR is a technique used to analyze the structure of many chemical molecules, primarily organic compounds. A typical compound might consist of carbon, hydrogen and oxygen atoms.

In its simplest form, an NMR experiment consists of three steps:

1. Place the sample in a static magnetic field.
2. Excite nuclei in the sample with a radio frequency pulse.
3. Measure the frequency of the signals emitted by the sample.

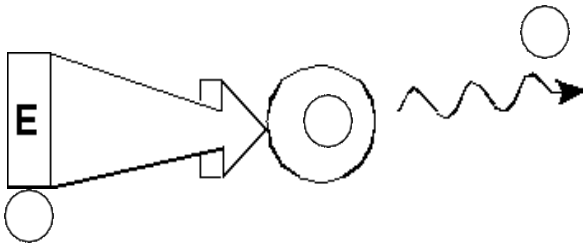



Figure 2.

From the emitted frequencies analysts can deduce information about the bonding and arrangement of the atoms in the sample. The NMR active nuclei in the sample resonate at different frequencies which are called "resonance frequencies ". Figure 2 These are the frequencies emitted by the nuclei when they are excited by the incoming radio frequency pulse. The value of a resonance frequency depends on two factors:

- **Type of Nucleus:** Every isotope has a particular combination of protons and neutrons in its nucleus. The nuclear structure largely determines the value of the resonance frequency. Thus, every isotope displays a "basic resonance frequency". 13C nuclei will have a different basic resonance frequency compared to that of 1H etc.

- **Local Atomic Environment:** Superimposed on the basic resonance frequency is an effect due to the local atomic environment in which an isotope is situated. The precise value of the resonance frequency of a 1H nucleus in a particular compound will depend upon the atoms it is bonded to and surrounded by. [47], [48]The nucleus is surrounded by electrons which may be viewed as moving electrical charges with associated magnetic fields. These electrons act as a source of magnetic shielding for the nucleus. The extent of the shielding will depend on the precise local atomic environment. The size of typical local field variations (which will result in a frequency variation) will depend on the isotope and the strength of the magnetic field in which the sample is placed. The table below shows the typical frequency variation for two of the most widely used NMR nuclei, 1H and 13C. The local atomic environment has a relatively small effect on the basic resonance frequency.

NMR signals are usually plotted as spectra and analyzed with respect to two features, frequency and intensity. It is conventional in NMR to plot frequency on the horizontal axis and increasing towards the left.

As mentioned above, the frequency yields qualitative information regarding the local atomic environment. The integrated intensity of a signal is a measure of signal strength and is determined by integrating the area under the signal peak. The integral will be directly proportional to the number of nuclei contributing to a signal at a particular frequency (if all nuclei are equally excited) and hence will provide quantitative information regarding chemical structure.

To excite a given nucleus in an NMR experiment, the frequency of the excitation pulse should closely match the resonance frequency of the nucleus. This frequency is referred to as the carrier frequency. Thus, if experiments are carried out using a 11.7 T magnet, the 1H nuclei would require a carrier frequency of approximately 500 MHz, whereas 13C nuclei would require a carrier frequency close to 126 MHz. The carrier frequency is specified by the parameter SFO1. The nucleus that is excited by this carrier frequency is referred to as the observe nucleus.

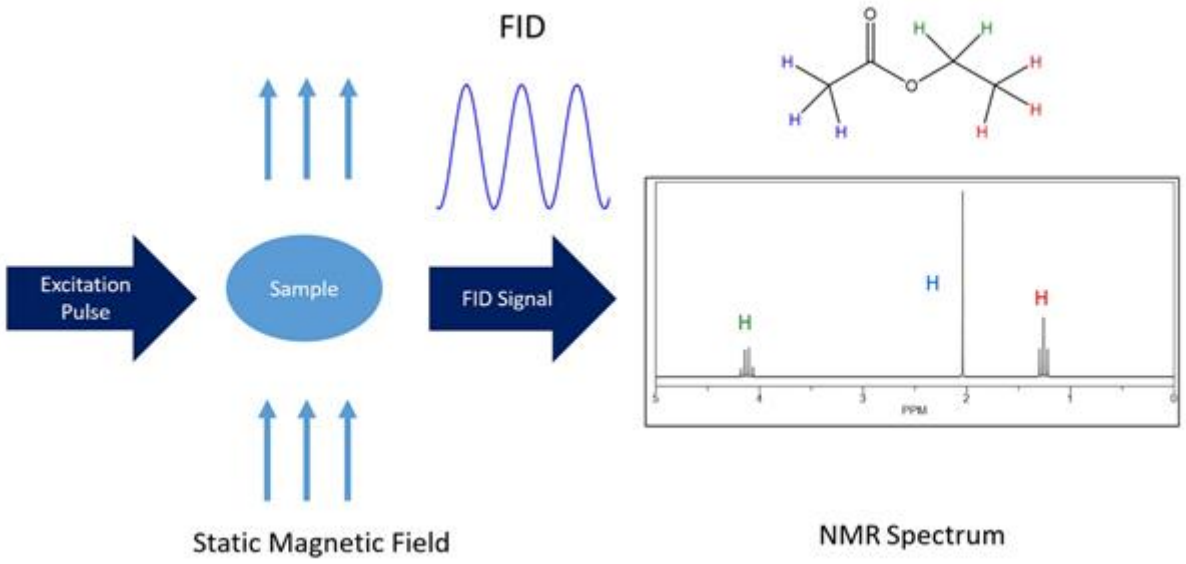



**Figure 3 Simplified layout of an NMR experiment.**

Not all isotopes will respond to radio frequency pulses, i.e. not all are NMR active. Three isotopes of the element hydrogen are found in nature: 1H (hydrogen), 2H (deuterium), and 3 H (tritium, radioactive!).

The natural abundance of these isotopes are 99.98%, 0.015%, and 0.005% respectively. All three are NMR active, although as can be seen in table 3.1, they all display a large variation in resonance frequency. To analyze a sample for hydrogen, the 1H isotope is excited, as this isotope is by far the most abundant. Of the carbon isotopes found in nature, only one is NMR active. By far the most common isotope, 12C (98.89% natural abundance) is inactive. Hence, NMR analysis of organic compounds for carbon rely on the signals emitted by the 13C isotope, which has a natural abundance of only 1.11%. Obviously, NMR analysis for carbon is more difficult than that of, for example, 1H (there are other factors which affect sensitivity, these will be discussed in the next sections of this chapter).

## The NMR Sectra post processing

The signals emitted by the excited atoms in the sample are received by the spectrometer and Fourier transformed by the software of the data station computer. The process of receiving the NMR signals is called an acquisition. Data is said to be acquired. A distinction should be made between the two terms; "FID" (time domain) and its associated "spectrum" (frequency domain). [49]
When an acquisition is carried out, "raw" data is acquired and the received signal is called an FID (Free Induction Decay). A typical FID is illustrated in the Figure 4.

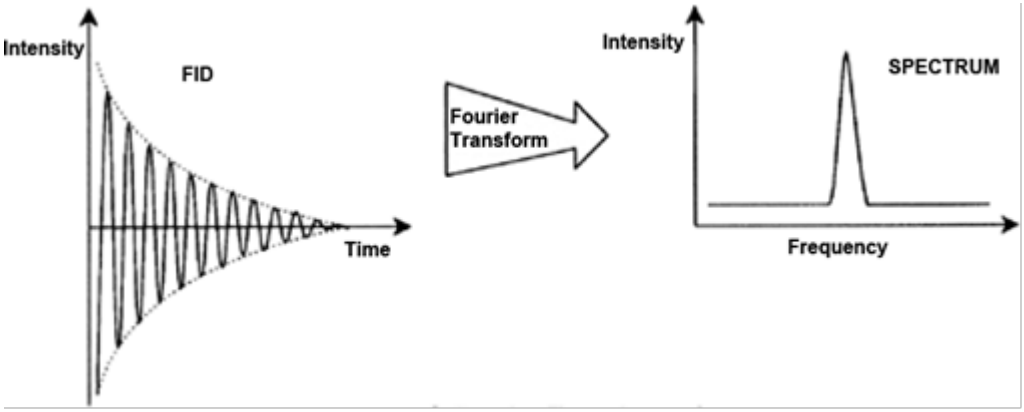



**Figure 4. Typical FID**

Before an FID can be usefully analyzed it must first be transformed to the frequency domain. This is achieved by applying a Fourier transformation. A Fourier transformation is a mathematical operation which converts the FID into a frequency spectrum. An FID is a signal whose intensity varies with time whereas a spectrum displays how intensity varies with frequency. Fourier transformation is the most important one of several processing operations that is normally carried out on raw data. While most other types of spectroscopic data can be subjected to chemometrics analysis directly from the spectrometer, NMR data often need to be preprocessed in several ways in order to conform to the prerequisites for chemometrics data analysis:

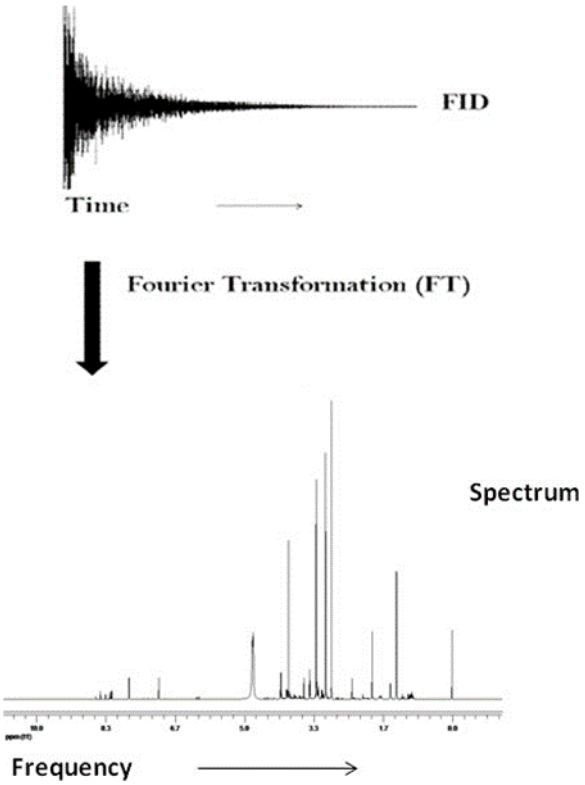



Figure 5. Conversion to time domain to frequency domain of a NMR spectrum

The most simplified layout of an NMR experiment. The NMR spectrum of ethyl acetate Table 1. H chemical shift assignments is shown as an example.

**Fourier transformation** (Keeler, 2011) (Figure 5).

In NMR spectroscopy, a Fourier transformation (FT) is required to convert the time domain data (free induction decay or FID, an electrical signal oscillating at the NMR frequency), obtained from the spectrometer, to the frequency domain (NMR spectrum). Naturally, quantitative methods require that parameter settings for the Fourier transform (choice of zero-filling and apodization function) are equal for all samples to be evaluated, since they may influence the finer details in the spectra.

**Phase errors** (Figure 6).

A difficult problem encountered with NMR data is the existence of phase errors of two orders: one and zero. In the real experiment, after FT, the spectrum line shapes are a mixture of absorptive and dispersive signals. They are related to the delayed FID acquisition that is commonly called first order phase. The delayed acquisition is a consequence of the minimum time required to change.

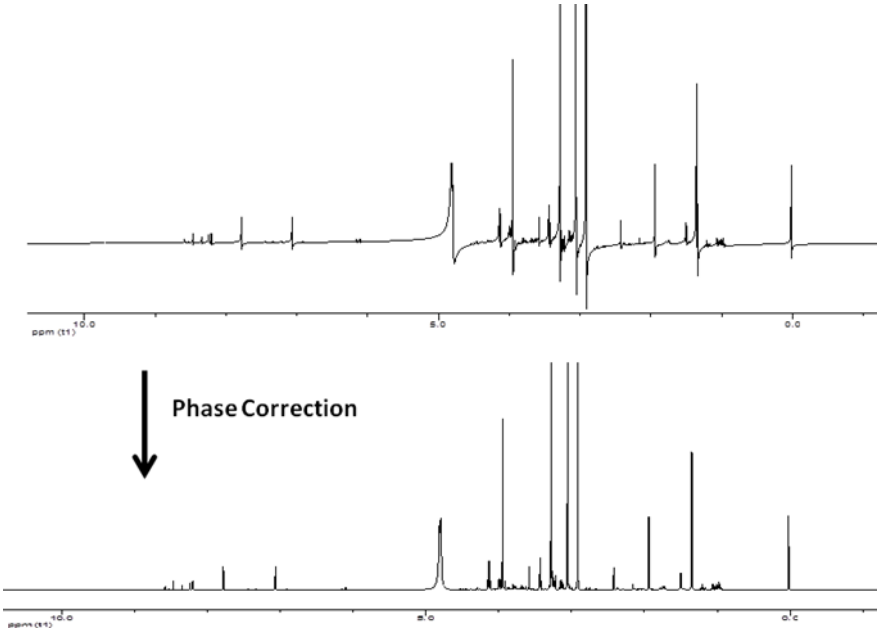



Figure 6. Phase errors

the spectrometer from transmit to receive mode. During this delay, the magnetization vectors process according to their chemical shift frequencies.

The zero order phase error arises because of the phase differences between the magnetization vectors and the receiver. Manual phase correction is usually implemented in the instrument software, but this process is very time consuming, especially for the large data sets that are often analyzed using chemometrics. More importantly, manually phase-correcting a series of spectra will lead to suboptimal results due to the subjective evaluation of the correction necessary for individual spectra.

**Data normalization** (Figure 7).

Data normalization is an important step for any statistical analysis.

The objective of data normalization is to allow meaningful comparisons of samples within the dataset. It is a row operation that is applied to the data from each sample and comprises methods to make the data from all samples directly comparable with each other [50], [51]. In this way it is possible to minimize most of the differences introduced with the effect of variable dilution and spectral data acquisition and processing.

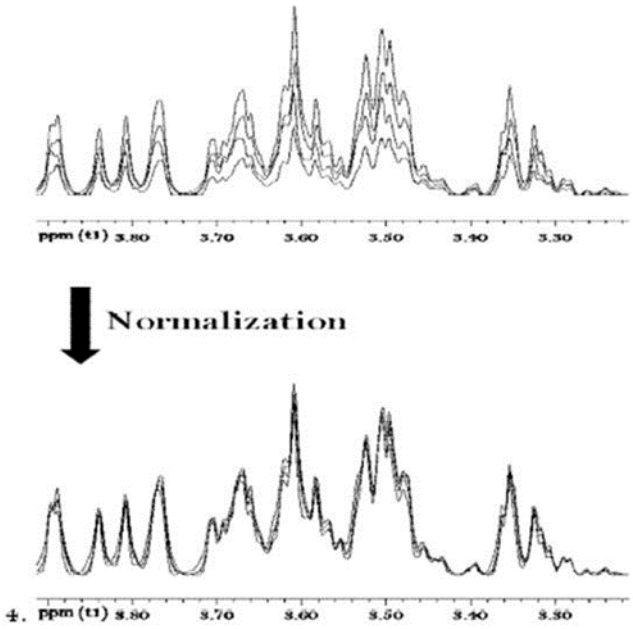



Figure 7. Data normalization

Normalization can be done using an internal "housekeeping" metabolite for example, an inner standard like TMS or, in this case, normalize each spectrum to (divide each variable by) the sum of the absolute value of all variables for the given sample. It returns a vector with an unit area (area = 1) "under the curve".

One of its common applications is to remove or to minimize the effects of variable dilution of the samples [52]

**Chemical shift variations**: the last preprocessing problem to be mentioned here and which occurs only in HR-NMR spectroscopy is the chemical shift variations that may occur from sample to sample or even from peak to peak. The overall sample-to-sample variations are due to small variations in spectrometer frequency, while the peak to peak chemical shift variations are due to variations in, for example, pH. In this last case, a data reduction in the form called binning or bucketing [50] is a pragmatic solution to the problem.

## Bucketing

In order to process the data and to create a usable dataset for a classification algorithm a bucketing procedure is needed. Bucketing performs a data reduction by grouping spectral responses, not being strictly a method to align data. [53]n the conventional method, the spectra are divided into evenly spaced windows, named bins or buckets, whose width commonly ranges between 0.01 and 0.04 ppm. The intensities inside each bin are summed, so that the area under each spectral region is used instead of individual intensities. Therefore, a new smaller set of variables (each one is the result of the sum of intensities) is created and, as the width of the buckets is set to cover the chemical shift variability around the peaks, the misalignment tends to be overcome. [54], [55]

that some areas from the same resonance signal can appear in two or more bins, splitting the chemical information in question. This occurs because conventional bucketing uses rigid boundaries. Despite this, several papers in the literature [56], [57] effectively use this methodology.

Figure 8 (a) shows a set of simulated misaligned NMR spectra. Figure 8 (b) illustrates the conventional bucketing procedure. As can be seen in Figure 8 (b), which presents the average simulated spectrum with the bucket boundaries denoted by vertical lines, the bucketing with 0.01 ppm width (Figure 8) is unable to properly isolate the signals. As result, in Figure 8 (c), where the buckets' values for each sample are shown through the colored bars (the bar heights are related to the values of the integrals that were normalized to the total sum equal to one), five important variables are observed, containing the principal information on the data set, which actually has three signals. Hence, this could seriously hamper interpretations, for example, when principal component analysis (PCA) is used.

The drawback cited above can be overcome by having the bin boundaries adjustable to minima, to provide optimized buckets of different sizes. In fact, a similar type of solution has already been proposed in the literature, as for example, the methodology for binning implemented in the commercial software ACD/Labs™ (Toronto, Canada) named intelligent bucketing [55]. In this method, the software chooses integral divisions based on local minima, thus searching for a better way of slicing up the spectra and avoiding the problem of the conventional bucketing. However, the software is not open source and the method for finding minima has not been reported. In other work, Davis et al. [58] proposed a

methodology named adaptive binning, where undecimated wavelet transform is used for denoising and to find all minima in a reference spectrum (maximum over each sample) performing the integration between these minima for each individual spectrum. However, in the decomposition a predefinition of both wavelet level and basis functions is necessary. Thus, there is a dependence on the number of levels in the decomposition, besides the threshold for denoising. Other alternatives for the traditional bucketing have been proposed in the recent literature, named Gaussian binning [59] and dynamic adaptive binning [60], but as the method proposed by Daves et al., these methodologies require a higher level of user expertise, being more complex than the algorithm presented here.

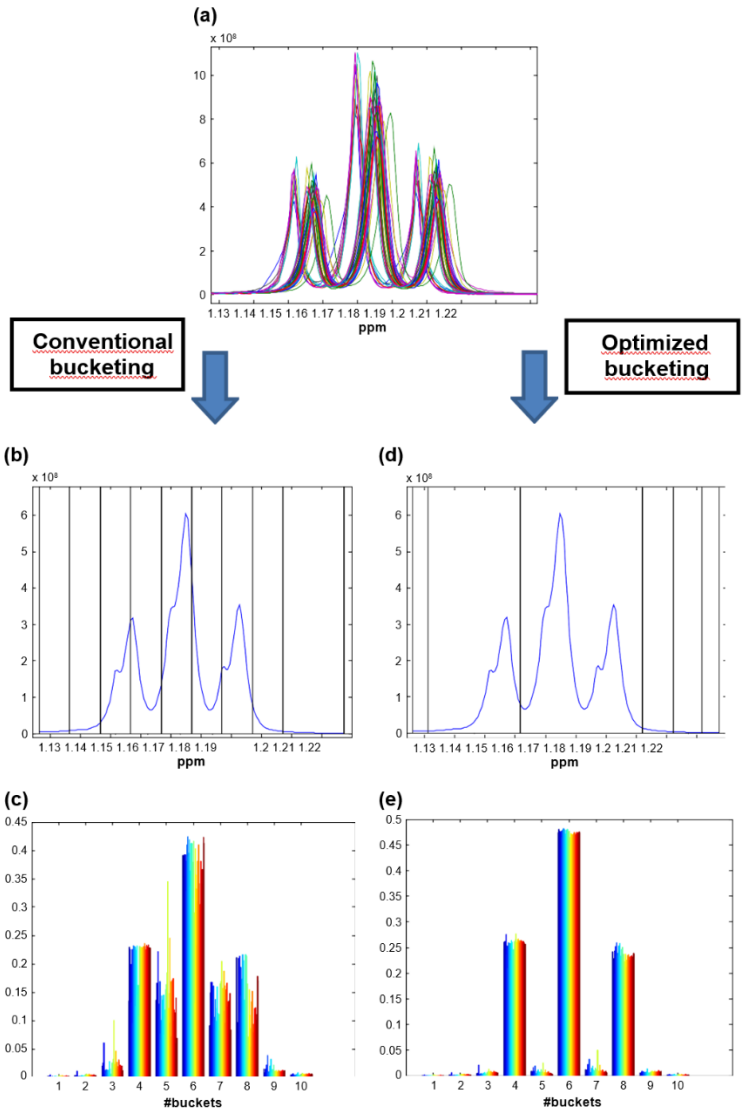



**Figure 8. . Scheme of the conventional and optimized bucketing procedures. (a) Simulated NMR spectra with misalignments. (b) Average simulated spectrum and the bucket boundaries (vertical lines) delimited by conventional bucketing, with buckets of size 0.01 ppm.**

## The case of wine grape in ReGeViP project

Before performing the Multivariate Statistical Analysis of all the spectra, these were subjected to alignment with respect to the TSP and a bucketing procedure. In fact, the input data of the analysis consist of matrices (bucket tables) formed by n rows and p columns. Each row indicates a sample and each column a variable. The variables are obtained from the bucketing procedure, ie the subdivision of the entire spectral region into intervals.

The bucketing is variable (see Figure 9) if we consider only some portions of the spectrum of amplitude proportional to some signals taken into consideration, instead it is regular (see Figure 10), like the one applied for the present work, if the entire spectrum is divided into equal regular intervals (0.05 to 0.01 ppm).

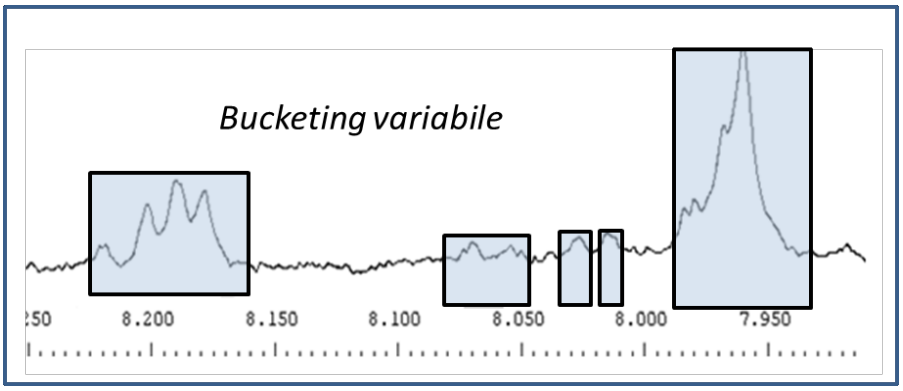



**Figure 9. Variable bucketing representation**

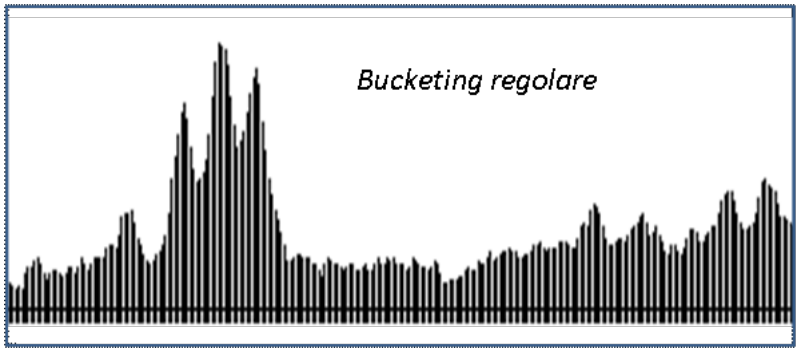



**Figure 10. Regular bucketing representation**

## *The challenges of data storage and conservation*

The data processing is completed using the proprietary software bundled with NMR Spectrometer or by the software MestreNova, a third-party solution which provides more advanced function in the elaboration of the data.

For a robust and efficient classification, the quantity and quality of data used to train the dataset is a paramount priority, as assessed in the previous chapters the quality of the dataset is assured by rigid protocols implemented during the phase of preparation or measurement. [16] We have yet to tackle the problem of storage and handling of the data. In the various experiments working with the file generated by the NMR Spectrometer a critical problem arises concerning the structure of the experiment results data structure.

Storing only the bucketed version of the spectrum obtained by the NMR experiment may seem an efficient and quite simple endeavor but storing the original experiment data is a key element to prepare for future test, experiment and research, hidden data can be still in the experiment result waiting for new approaches to reveal it, so efficiently storing of this data is an important task.

Actually, the BRUKER data storage is not very efficient, the entire system id file based and if on one side it makes easy to transfer the data between different machine and backup very easy is not ideal for batch operations.

Usually, each experiment is stored in a single folder, where the fid file is stored alongside other file regarding the experiment parameters. The most important files after the food file are the acqu and acqus file. Acqu stores the requested experiment parameters as set in the NMR control board and the acqus file stores the actual parameters as measured during the experiment. For example, the experiment may be set to happen at 0.0°C but due to the refrigerator system the experiment may happen at 0.2°C, this second parameter is stored for further analysis by the research team.

The problem with this approach is that a lot of batch investigation must be done in rudimental way.

The experiment parameter are almost 350 and is hard to store efficiently all of these data, but is also important to take care of these information.

Another problem with the raw experiment data is the storage and backup of the files, the file based approach makes easy to create a lot of duplicates, different data processing creates new file datasets starting from the same feed and this is not tracked. We observed a proliferation of copy of dataset only

to differentiate different analysis approach. Each experiment usually occupied 1.5mb space uncompressed (the compress ratio achievable with these data is almost 40%-50%), and each experiment consist of 20 files.

A simple experiment run can produce hundreds of spectra, each with 20 files inside and with a significative amount of data. With this approach is easy to miss some data or have some problem. We are actively working to create a centralized data collection system and integrity checking solution.

# Methods

## *Chemometrics*

An NMR measurement is capable to detect all metabolites in a sample. It is full of information (variables), part of which most of the time results to be redundant. For this reason, it is important in a metabolomics research to compress these variables in order to have only those that contain the useful information.

This kind of approach is commonly called chemometrics (approach) and it can be defined as "How to get chemically relevant information out of measured chemical data, how to represent and display this information, and how to get such information into data" [61].

Chemometrics is the field of extracting information from multivariate chemical data using the tools of statistics and mathematics. It is typically used for one or more of three primary purposes:

1) to explore patterns of association in data;
2) to track properties of samples;
3) to prepare and use multivariate classification models.

Exploratory data analysis can reveal hidden patterns in complex data by reducing the information to a more comprehensible form. Such chemometrics analysis can expose possible outliers and indicate whether there are patterns or trends in the data. Exploratory algorithms such as principal component analysis (PCA) and hierarchical cluster analysis (HCA) are designed to reduce large complex data sets into a series of optimized and interpretable views. These views emphasize the natural groupings in the data and show which variables greatly influence those patterns.

Formally, PCA is a way of identifying patterns in data, expressing them in such a way as to highlight their similarities and/or differences[62], [63]. The advantage of this technique is the capability to reduce multidimensional data set (a data matrix) into a new set of uncorrelated (i.e., orthogonal) variables by performing a covariance analysis (ANCOVA) between factors.[64]

The PCA works by decomposing the X-matrix (containing the original data set) as the product of two smaller matrices, which are called loading and score matrices.

The loading matrix (V) contains information about the variables: it is composed of a few vectors (Principal Components, PCs) which are (obtained as) linear combinations of the original X-variables.

The score matrix (U) contains information about the objects. Each object is described in terms of its projections onto the PCs, (instead of the original variables). The information not contained in these matrices remains as "unexplained X-variance" in a residual matrix (E) which has exactly the same dimensionality as the original X-matrix.

The PCs, among many others, have two interesting properties:

1. they are extracted in decreasing order of importance. The first PC always contains more information than the second, the second more than the third and so on.
2. they are orthogonal to each other. There is absolutely no correlation between the information contained in different PCs.

In PCA, is possible to decide how many PCs should be extracted (the number of significant components, i.e. the dimensionality of the model).

Each new PC extracted further increases the amount of information (variance) explained by the model. However, usually the first four or five PCs explain more than 90% of the X-variance. Anyway, there is not a simple nor unique criterion which decides how many PC to extract so two kinds of considerations should be considered. From a theoretical point of view, it is possible to use cross validation techniques to decide the number of PCs to include. Since data patterns can be hard to find in data of high dimension, like spectroscopic ones, where most of the time the information is redundant, PCA is a powerful tool for analyzing them.

In $^{1}$H NMR, redundancy means that some of the variables are correlated with one another because they are measuring the same construct (different picks for the same molecule). Therefore, this redundancy can reduce the observed variables into a smaller number of artificial variables (principal components or

latent factors) that are a linear combination of the original ones and will account for most of the variance in the observed variables without losing information.

In this way, by using a few components, each sample (spectrum) can be represented by relatively few numbers instead through values for thousands of variables (spectral data points). Then, samples can be plotted making it possible to visually assess similarities and differences between samples and determine whether samples can be grouped. [65]

There are a few common plots which are always used in connection with PCA: 1) the scores/scores plot (left part of the Figure below) and

2) the corresponding loading/loading plot (Figure 11)

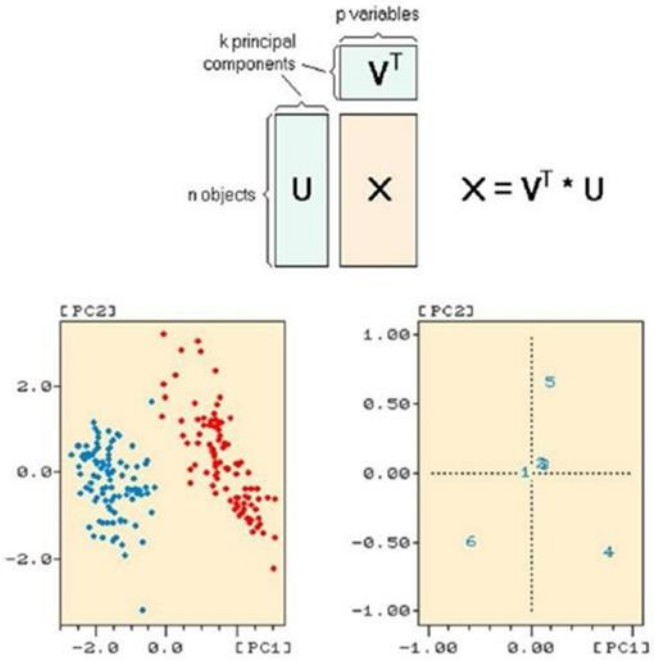



**Figure 11. the two matrices V and U are orthogonal. The matrix V is usually called the loadings matrix and the matrix U is called the scores matrix.**

As a clustering technique, PCA is most commonly used to identify how one sample is different from another, which variables contribute most to this difference, and whether those variables contribute in the same way (i.e. are correlated) or independently (i.e. uncorrelated) from each other.

In contrast to PCA, **PLS** and **PLS-DA** [66], [67] are supervised classification techniques that can be used to enhance the separation between groups of observations by rotating PCA components so that a maximum separation among classes is obtained. [68]

The purpose of Discriminant Analysis is to classify objects (people, customers, foods, genes, things, etc.) into one of two or more groups based on a set of features that describe the objects (e.g. gender, age, income, weight, preference score, genotypes, metabolites' content etc.). In general, is assigned an object to one of several predetermined groups based on observations made about the object. For example, if one wants to know whether a soap product is good or bad, this judgment is based on several measurements of the product such as weight, volume, people's preferential score, smell, color contrast etc. The object here is soap. The class category or the group "good" and "bad") is what is looked (it is

also called dependent variable). Each measurement on the product is called features that describe the object (it is also called independent variable). Thus, indiscriminant analysis, the dependent variable (Y) is the group, and the independent variables (X) are the object features that might describe the group. The dependent variable is always category (nominal scale) variable while the independent variables can be any measurement scale (i.e. nominal, ordinal, interval or ratio).

Partial Leas t Squares (PLS) is useful when a (very) large set of independent variables have to be predicted. It originates in the social sciences but becomes popular also in all branches basing on chemometrics methods, including food science [69]. It is a multivariate regression method allowing establishing a relationship between one or more dependent variables (U) and a group of descriptors (T).

Tand U-variables are modeled simultaneously to find the latent variables (LVs) in T that will predict the latent variables in U and at the same time account for the largest possible information present in T; Figure 12 gives a schematic outline of the method.

The overall goal (shown in the lower box of Figure 2.6) is to use the factors to predict the responses in the population. This is achieved indirectly by extracting latent variables T and U from sampled factors and responses, respectively.

The extracted factors T (also referred to as X-scores) are used to predict the Yscores U, and then the predicted Y-scores are used to construct predictions for the responses (Manetti et al., 2004).

Hence the PLS method is popular in industries that collect correlated data on many x-variables, known as predictors. For example, multivariate calibration in analytical chemistry; spectroscopy in chemometrics. The PLS method extracts orthogonal linear combinations of predictors, known as factors (**T** or **X-Scores**), from the predictor data that explain variance in both the predictor variables and the response (**U** or **YScores**) variable(s) (Figure is adapted from [70], [71]

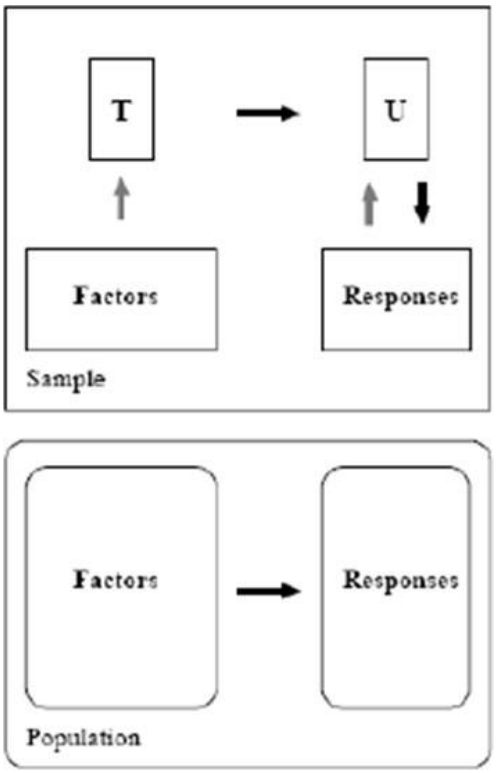



**Figure 12. schematic outline of PLS method**

So, in this case the latent variables are selected on the basis of explaining contemporarily both descriptors and predictors. These latent variables are similar to the principal components calculated from PCA. The first one accounts for the largest amount of information followed by the other components that account for the maximum residual variance. As for PCs, the last LVs are mostly responsible for random variations and experimental error. The optimal number of LVs, i.e. modelling information in X useful to predict the response Y but avoiding overfitting, is determined on the basis of the residual variance in prediction.

**Cross validation techniques** are adopted for evaluating the predictive ability and for selecting the optimal number of latent variables.

PLS was contrived to model continuous responses, but it can be applied even for classification purposes by establishing an appropriate Y related to each sample belonging to a class. In this case it is called Partial Least Squares Discriminant Analysis (PLS-DA). [72]–[74]

In the case of proteomic data, one response variable for each group of samples is usually adopted. Each response variable is assigned a 1 value for the samples belonging to the corresponding class and a 0 value for the samples belonging to the other classes.

In general, a PLS analysis [75] consists of the stages:

1. calculate a PLS model using a high number of factors (more than is likely to be required);
2. determine the number of factors to include in a fitted model by either:
    - analysing information calculated during the process of extracting factors;
    - calculating a prediction accuracy estimate based on, e.g. , cross validation;
3. fit the model with the determined number of factors by calculating parameter estimates of the linear regression;
4. given a set of predictors and responses used to fit a PLS model, and a suitable number of factors used to calculate parameter estimates, estimate response values to new predictor data.

As noted previously, chemometrics approaches like PCA and PLS-DA, on their own, do not permit the direct identification or quantification of compounds. In the other approach to metabonomics (quantitative metabolomics or targeted profiling), the focus is on attempting to identify and/or to quantify as many compounds in the sample as possible.

This is usually done by comparing the spectroscopic data (obtained from the sample's NM or MS ones) spectroscopic data reference library obtained from pure compounds [76]–[78]. Once the constituent compounds are identified and quantified, the data are then statistically processed (using PCA or PLS-DA) to identify the most important biomarkers or informative metabolic pathways [77].

Depending on the objectives and instrumental capacity, quantitative metabolomics may be either targeted (selective to certain classes of compounds) or comprehensive (covering all or almost all detectable metabolites).

## PCA

Among the multivariate analysis techniques, PCA [79] is the most frequently used because it is a starting point in the process of data mining [80], [81]. It aims at minimizing the dimensionality of the data. Indeed, it is common to deal with a lot of data in which a set of n objects is described by a number p of variables. The data is gathered in a matrix X, with n rows and p columns, with an element $x_{ij}$ referring to an element of X at the $i^{th}$ line and the $j^{th}$ column. Usually, a line of X corresponds with an "observation", which can be a set of physicochemical measurements or a spectrum or, more generally, an analytical curve obtained from an analysis of a real sample performed with an instrument producing analytical curves as output data. A column of X is usually called a "variable". With regard to the type of analysis that concerns us, we are typically faced with multidimensional data n x p, where n and p are of the order of several hundreds or even thousands. In such situations, it is difficult to identify in this set any relevant information without the help of a mathematical technique such as PCA. This technique is commonly used in all areas where data analysis is necessary; particularly in the food research laboratories and industries, where it is often used in conjunction with other multivariate techniques such as discriminant analysis.

## PLS

The PLS approach was originated around 1975 by Herman Wold for the modelling of complicated data sets in terms of chains of matrices (blocks), so-called path models. This included a simple but efficient way to estimate the parameters in these models called NIPALS (Non-linear Iterative Partial Least Squares). This led, in turn, to the acronym PLS for these models (Partial Least Squares). This relates to the central part of the estimation, namely that each model parameter is iteratively estimated as the slope of a simple bivariate regression (least squares) between a matrix column or row as the y-variable, and another parameter vector as the x-variable. So, for instance, the PLS weights, **w**, are iteratively re-estimated as **X**'**u**/(**u**'**u**). The "partial" in PLS indicates that this is a partial regression, since the **x**-vector (**u** above) is considered as fixed in the estimation. This also shows that we can see any matrix–vector multiplication as equivalent to a set of simple bivariate regressions. [82] This provides an intriguing connection between two central operations in matrix algebra and statistics, as well as giving a simple way to deal with missing data.

Gerlach et al. [83] applied multi-block PLS to the analysis of analytical data from a river system in Colorado with interesting results, but this was clearly ahead of its time.

Around 1980, the simplest PLS model with two blocks (**X** and **Y**) was slightly modified by Svante Wold and Harald Martens [84] to better suit to data from science and technology and shown to be useful to deal with complicated data sets where ordinary regression was difficult or impossible to apply. To give PLS a more descriptive meaning, H. Wold et al. have also recently started to interpret PLS as Projection to Latent Structures.

## SIMCA

Soft independent modelling of class analogy (SIMCA) [85], [86] is an established method for multivariate classification. Disjoint Principal Component Analysis (PCA) [87] models are fitted for each class, and model residuals are utilised to classify unknown observations to no class, one class or several classes [88]. The method effectively handles a multitude of classes demonstrating high within-class variability and has been utilised in numerous fields, such as metabonomics [89]–[91] and transcriptomics [92].

However, each disjoint PCA model is generated based on the direction in the data demonstrating the highest variation, which might be distinctly different from the direction separating the classes. Consequently, maximum class-separation is not explicitly the objective function of the method. Furthermore, due to the usage of several local PCA models, information regarding between-class differences is not easily accessible, which hampers the quality of interpretation (transparency) of the classification model.

In SIMCA classification, the residuals of several disjoint PCA models are utilised to assign an observation to one or several of the available classes. During the training of each class-specific PCA model, a distribution of the residuals for each class is generated. Given this class-specific residual distribution, any given observation can subsequently be assigned a probability of equal variance compared to the model residuals according to a F-test. The probability assignment is then ultimately used to accept or reject the observation to or from each class, which is essentially a tool for detecting model outliers.

## OPLS

OPLS [93], [94] is an extension to the supervised PLS regression method featuring an integrated OSC-filter. In simple terms, OPLS uses information in the Y matrix to decompose the X matrix into blocks of structured variation correlated to and orthogonal to Y, respectively. The block containing the correlated variation, also referred to as the predictive variation, can also be derived from the normalised PLS regression vectors followed by a procedure called 'target rotation' developed by Kvalheim and Karstang [95]. OPLS can, analogously to PLS-DA, be used for discrimination (OPLS-DA) which has been demonstrated in a recent metabonomic study by Cloarec et al. [96]. The integrated

OSC-filter introduces a noteworthy advance compared to PLS regression, mainly related to the transparency of the generated models.

The main benefit in interpretation using OPLS-DA compared to PLS-DA thus lies in the ability of OPLS-DA to separate predictive from non-predictive (orthogonal) variation. This advantage can be demonstrated using a simple two-class scenario based on spectral data. For PLS-DA, two components $t_1$ and $t_2$ are required to find a perfectly discriminating plane between the two classes as shown in Figure 1A. The

corresponding loading vectors p1 and p2 will contain a mixture of both the discriminatory properties as well as the non-discriminatory properties that are mainly confounded with the direction of $t_2$. In the same example, OPLS-DA effectively separates the discriminatory direction in tp,1 from the Y-orthogonal direction to,1 making the corresponding predictive loading vector pp,1 straightforward to interpret. The parts of the spectra responsible for the remaining variation can be identified from the Y-orthogonal po,1 loading vector, which is mainly related to high within-class variance of one of the classes.
Predictions from the OPLS-DA model are in the form of the categorical variables used for estimation of the OPLS-DA model. To make objective classification decisions using the predicted Y-matrix, henceforth denoted as $Y^{hat}$, there is a need to develop some decision rules for this purpose. For PLS-DA models, fixed or optimised boundaries have commonly been used to assign class membership based on the predicted values [97]–[99]. An alternative decision rule is simply to assign class membership according to the $Y^{hat}$ exhibiting the largest numerical value.

## *Machine learning algorithms*

There are different algorithms that can be used in chemometrics to describe and to predict the data, such as the pre-processing methods of variable selection, variable extraction, data dimensionality reduction, and data modeling methods (clustering, classification, and regression). Some of the algorithms have emerged from computer science and artificial intelligence, making them largely used in data mining and artificial intelligence research.

Most of the food analysis papers begin with exploratory data analysis (unsupervised techniques), followed by classification or regression algorithms (supervised techniques). Principal component analysis (PCA) and cluster analysis (HCA) are the most used exploratory techniques in food science [6], [100]. These techniques are usually employed to explore similarities and hidden patterns among samples and try to establish the existence of groups in data. Nevertheless, sometimes, PCA has been erroneously applied as a classification technique rather than as an exploratory method in the evaluation of different foods, such as rice [101], fats and oils [102], and other food products.

Regression models are used in many applications of analytical chemistry as predictive modeling or calibration, where the regression model establishes a relationship between a chemical and/or physical characteristic of a sample with cheaper and easier instrumental signals. The most commonly used regression techniques are partial least squares regression (PLS), principal component regression (PCR), and multiple linear regression (MLR). [103]

On the other hand, the goal of classification problems is to generate a function which maps a sample into one of a set of pre-defined classes. The classification algorithm can be divided into two categories: discriminant analysis and class modeling. The discriminant algorithms are two or multiclass classifiers, whereas the class modeling algorithms are one-class classification models. The most common class modeling algorithms are the soft independent modeling of class analogy (SIMCA), and unequal dispersed classes (UNEQ), which both algorithms verify whether a sample is compatible or not with the characteristics of a single class of interest. [104]

The discriminant algorithms establish boundaries among the analyzed classes defined by the training samples. The most common algorithms are linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA), quadratic discriminant analysis (QDA), k nearest neighbors (kNN), [105]support vector machines (SVM). Additionally, there are other multivariate classification methods based on decision rules, such as decision trees (DT), random forests (RF), and artificial neural networks (ANN).

Jiménez-Carvelo et al. (2019) classified the most used multivariate methods in food analysis as conventional ones (PCA, PLS-DA, LDA, KNN, PARAFAC, SIMCA) and the algorithms that emerged from computer science as alternative data mining/machine learning methods (SVM, ANN, CART, J48, C5.0, Random Forest). Methods that are alternative to conventional ones are required as the complexity of data increases due to the development of advanced analytical methods, which can provide physical-chemical parameter information with a high level of detail.

Different decision trees algorithms (CART, J48, C5.0, Random Forest), SVM, and ANNs emerged from computer science and some of them are computationally intense. Furthermore, in some cases, these algorithms do not have a reproducible solution. Moreover, these algorithms can be used on both classification and regression problems. However, the application of these algorithms in food science is still scarce, although they are largely used in other areas of artificial intelligence applications. [106] A detailed consideration of the respective mathematical models is not in the perspective of this review and has already been covered in the literature. [107], [108]

## Random Forest

Random forests are a combination machine learning algorithm. Which are combined with a series of tree classifiers, each tree cast a unit vote for the most popular class, then combining these results get the final sort result. RF posses high classification accuracy, tolerate outliers and noise well and never got overfitting. RF has been one of the most popular research methods in data mining area and information to the biological field. In China there are little study on RF, so it is necessary to systemic summarize the down to date theory and application about RF. [109]

RF can be used to deal with micro-information data, and the accuracy of RF is higher than those traditional predictions. So in recently 10 years, Random Forest has been got a rapid development, and widely used in many areas, such as bioinformatics, medicine, management science, economics. In bioinformatics, Smith et al. studied the tracking data on bacteria by RF and compared with Discriminant Analysis method. Alonso et al. use biomarkers parasite to discriminate fish stocks; In medicine, Using RF technology such as Lee to help lung CT images of lung nodules automatic detection, and also in the RF (CAC). In China, Jia FuCang,Li Hua, Using RF to the Dhoop magnetic resonance image segmentation, and that the RF has fast speed and high accuracy, is a promising muti-channel image segmentation method. The main application in economic management field, is predicating the loss degrees of customers. Bart used RF in customer relationship management, found that the effect of RF is better than ordinary linear regression and Logistic model. Coussement et al. [110] compared the predictive ability of SVM, logistic model and the RF in loss of customers, found that RF is always better than the SVM. Burez et al. applied

weighted RF in loss of customers, comparing with the RF, and found the weighted RF has better prediction. [111]

Today, the range of application of RF is very broad, in addition to the above mentioned application, the RF also used in ecology, remote sensing geography terms, customer's loyalty forecasting; Lessmann etc. also use Random Forest predict horse racing winning, and that the predictions of the Random Forest is superior to traditional forecasting methods can bring in huge commercial profits. [112]

## Multilayer Perceptron

Environmental modelling involves using a variety of approaches, possibly in combination. Choosing the most suitable approach depends on the complexity of the problem being addressed and the degree to which the problem is understood. Assuming adequate data and computing resources and if a strong theoretical understanding of the problem is available then a full numerical model is perhaps the most desirable solution. However, in general, as the complexity of a problem increases the theoretical understanding decreases (due to ill-defined interactions between systems) and statistical approaches are required. Recently, the use of neural networks, and in particular the multilayer perceptron, have been shown to be effective alternatives to more traditional statistical techniques. Primarily it has been shown [113] that the multilayer perceptron can be trained to approximate virtually any smooth, measurable function. Unlike other statistical techniques the multilayer perceptron makes no prior assumptions concerning the data distribution. It can model highly non-linear functions and can be trained to accurately generalise when presented with new, unseen data. These features of the multilayer perceptron make it an attractive alternative to developing numerical models, and also when choosing between statistical approaches. As will be seen the multilayer perceptron has many applications in the atmospheric sciences. [114]

The multilayer perceptron consists of a system of simple interconnected neurons, or nodes, as illustrated in Figure 13, which is a model representing a nonlinear mapping between an input vector and an output vector. The nodes are connected by weights and output signals which are a function of the sum of the inputs to the node modified by a simple nonlinear transfer, or activation, function. It is the superposition of many simple nonlinear transfer functions that enables the multilayer perceptron to approximate extremely non-linear functions. If the transfer function was linear then the multilayer perceptron would only be able to model linear functions. Due to its easily computed derivative a commonly used transfer function is the logistic function, as shown in Figure 14. The output of a node is scaled by the connecting weight and fed forward to be an input to the nodes in the next layer of the network. This implies a direction of information processing, hence the multilayer perceptron is known as a feed-forward neural network. [115] The architecture of a multilayer perceptron is variable but in general will consist of several layers of neurons. The input layer plays no computational role but merely serves to pass the input vector to the network. The terms input and output vectors refer to the inputs and outputs of the multilayer perceptron and can be represented as single vectors, as shown in Figure 13. A multilayer perceptron may have one or more hidden layers and finally an output layer. Multilayer perceptrons are described as being fully connected, with each node connected to every node in the next and previous layer. [116]

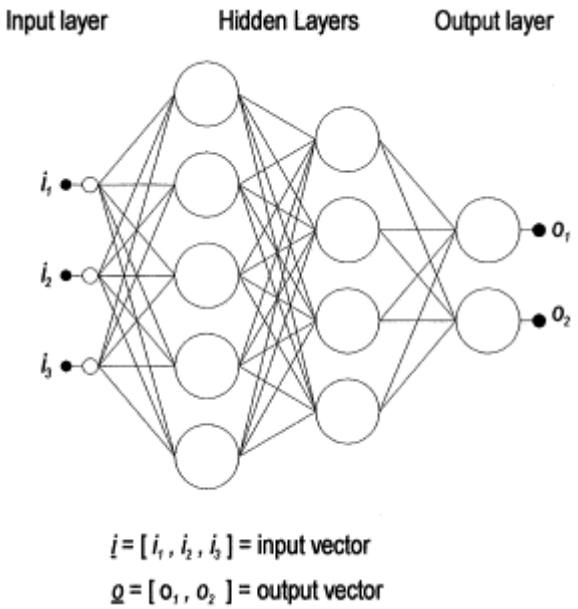



**Figure 13. interconnected neurons**

By selecting a suitable set of connecting weights and transfer functions, it has been shown that a multilayer perceptron can approximate any smooth, measurable function between the input and output vectors [113]. Multilayer perceptrons have the ability to learn through training. Training requires a set of training data, which consists of a series of input and associated output vectors. During training the multilayer perceptron is repeatedly presented with the training data and the weights in the network are adjusted until the desired input–output mapping occurs. Multilayer perceptrons learn in a supervised manner. During training the output from the multilayer perceptron, for a given input vector, may not equal the desired output. An error signal is defined as the difference between the desired and actual output. Training uses the magnitude of this error signal to determine to what degree the weights in the network should be adjusted so that the overall error of the multilayer perceptron is reduced. There are many algorithms that can be used to train a multilayer perceptron. Once trained with suitably representative training data the multilayer perceptron can generalise to new, unseen input data.

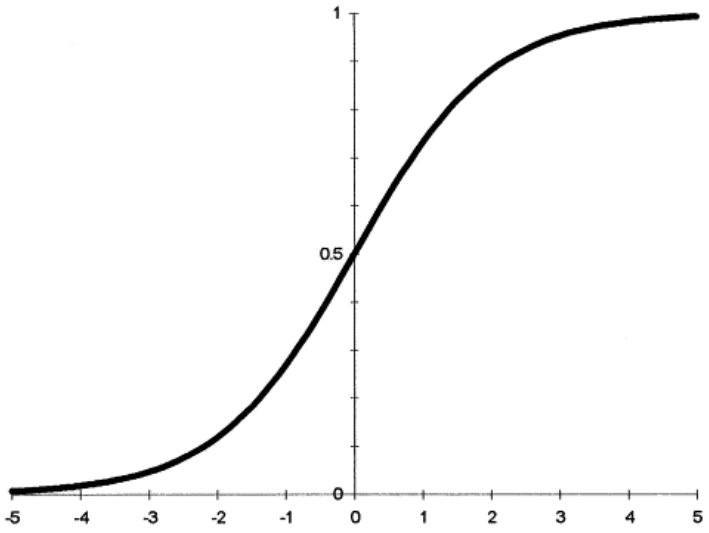


**Figure 14. The logistic function y=1(1−exp(−x)).**

3. Training a multilayer perceptron—the back-propagation algorithm

Training a multilayer perceptron is the procedure by which the values for the individual weights are determined such that the relationship the network is modelling is accurately resolved. At this point we will consider a simple multilayer perceptron that contains only two weights. [117] For any combination of weights the network error for a given pattern can be defined. By varying the weights through all possible values, and by plotting the errors in three-dimensional space, we end up with a plot like the one shown in Figure 15. Such a surface is known as an error surface. The objective of training is to find the combination of weights which result in the smallest error. In practice, it is not possible to plot such a surface due to the multitude of weights. What is required is a method to find the minimum point of the error surface.

Fig. 4. An error surface for a simple multilayer perceptron containing only two weights.

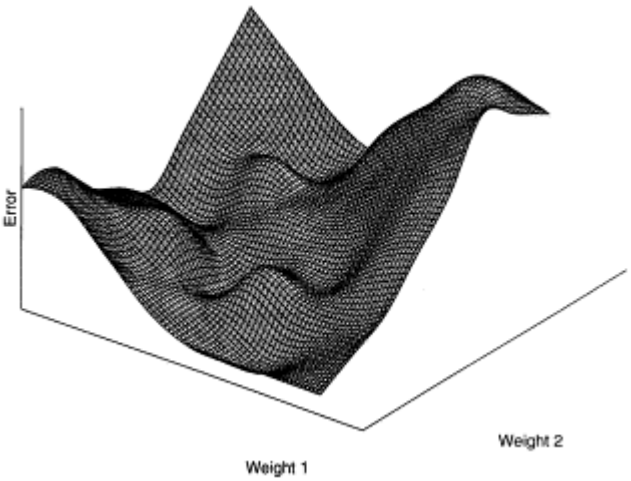



**Figure 15. plot of the errors in three-dimensional space**

One possible technique is to use a procedure known as gradient descent. The backpropagation training algorithm uses this procedure to attempt to locate the absolute (or global) minimum of the error surface. The backpropagation algorithm [118] is the most computationally straightforward algorithm for training the multilayer perceptron. Backpropagation has been shown to perform adequately in many applications; the majority of the applications discussed in this paper used backpropagation to train the multilayer perceptrons. A full mathematical derivation of this algorithm can be found in almost all NEURAL NETWORK textbooks so only the essential components of the algorithm will be discussed here. Backpropagation only refers to the training algorithm and is not another term for the multilayer perceptron or feed-forward neural networks, as is commonly reported.

The weights in the network are initially set to small random values. This is synonymous with selecting a random point on the error surface. The backpropagation algorithm then calculates the local gradient of the error surface and changes the weights in the direction of steepest local gradient. Given a reasonably smooth error surface, it is hoped that the weights will converge to the global minimum of the error surface.

The backpropagation algorithm is summarised below. Implementation details can be found in most neural network books [119]

1. initialize network weights,
2. present first input vector, from training data to the network,
3. propagate the input vector through the network to obtain an output,
4. calculate an error signal by comparing actual output to the desired (target) output,
5. propagate error signal back through the network,

6. adjust weights to minimise overall error,
7. repeat steps 2–7 with next input vector, until overall error is satisfactorily small.

The above implementation of the backpropagation algorithm is known as on-line training whereby the network weights are adapted after each pattern has been presented. The alternative is known as batch training, where the summed error for all patterns is used to update the weights. The benefits of each approach are discussed in [120]. In practice, many thousands of training iterations will be required before the network error reaches a satisfactory level—determined by the problem being addressed. As will be discussed later, training should be stopped when the performance of the multilayer perceptron on independent test data reaches a maximum, which is not necessarily when the network error is minimised.

The error surface in Figure 15 contains more than one minimum. It is desirable that the training algorithm does not become trapped in a local minimum. The backpropagation algorithm contains two adjustable parameters, a learning rate and a momentum term, which can assist the training process in avoiding this. The learning rate determines the step size taken during the iterative gradient descent learning process. If this is too large then the network error will change erratically due to large weight changes, with the possibility of jumping over the global minima. Conversely, if the learning rate is too small then training will take a long time. The momentum term is used to assist the gradient descent process if it becomes stuck in a local minimum. By adding a proportion of the previous weight change to the current weight change (which will be very small in a local minimum) it is possible that the weights can escape the local minimum.

**J48**

The Decision tree is one of the classification techniques which is done by the splitting criteria. The decision tree is a flow chart like a tree structure that classifies instances by sorting them based on the feature (attribute) value. Each node in a decision tree represents a feature in an instance to be classified. All branches denote an outcome of the test, each leaf node hold the class label. The instances are classified from starting based on their feature value. Decision tree generates the rule for the

classification of the data set. Three basic algorithms are widely used that are ID3, C4.5, and CART. ID3 is an iterative Dichotomer 3. [121] It is an older decision tree algorithm introduced by Quinlan Ross in 1986 [122]. The basic concept is to make a decision tree by using the top-down greedy approach. C4.5 is the decision tree algorithm generated by Quinlan [24]. It is an extension of ID3 algorithm.. C4.5 algorithm is widely used because of its quick classification and high precision. CART stands for Classification Regression Tree introduced by Bremen [123]. The property of CART is that it is able to generate the regression tree. In Regression tree leaf node contains a real number instance of a Class. A decision tree classifier is built in two phases:

- A growth phase
- A prune phase

After the preliminary tree has been built that is ‘growth phase', a sub-tree is created with the least estimated error rate, that is the 'prune phase'. The process of pruning the preliminary tree consists of removing small, deep nodes of the tree resulting from 'noise' contained in the training sample thus decreasing the risk of 'over fitting' and ensuring in a more precise classification of unknown data.

As the decision tree is being built, the goal at each node is to decide the split attribute (feature) and the split point that best divides the training instances belonging to that leaf. The value of a split point depends on how well it separates the classes. Numerous splitting indices have been proposed in the precedent to evaluate the quality of the split. The below Figure 16 shows the decision tree of weather prediction data base.

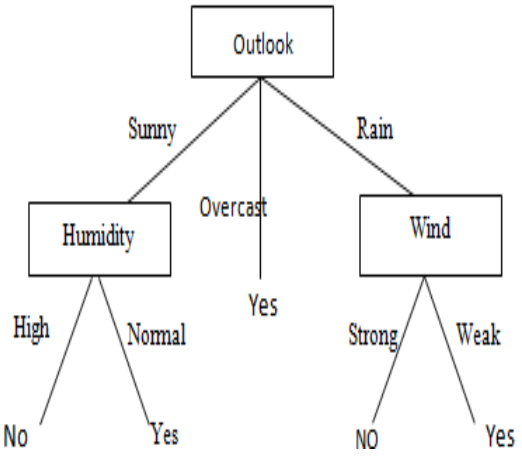



**Figure 16. decision tree**

Jiang Su and Harry Zhang [124] have discussed the decision tree method and also proposed a fast decision tree. They build a tree based on the conditional dependence assumption. Author shows that the performance and accuracy of the new approach is better than the C4.5 and less complexity as compare to C4.5 decision tree

In 2008 Tomasz Maszczyll and WlodzislawDuch [125] modify the C4.5 algorithm based on Tasallis and renyi entropy. After comparative analysis author is found that the modified C4.5 algorithm is better than the Shannon entropy based C4.5 algorithm. At basic of the decision tree algorithm ID3 are very famous and easy to classify but if classifying attribute which have many values then this algorithm are not beneficial.

Christiane Ferreria Lemos Lima et al [126] describes the comparative study of the use of Shannon, Renyi and Tsallis entropies for designing decision tree. The goal of that paper is to find more efficient alternative entropy for the intrusion Tolerant system. The author show, the resultant tsallis and renyi entropy can be used to construct more compact and efficient decision tree.

In 2011 Mosonyi et al [127] proposed the Quantum Renyi Relative formula and relative capacity formula. The Shannon entropy is sensitive to noise sample and doesn't work well in real work applications. So introduce the other measures of feature quality called the Rank mutual information.

Qinghua Hu et al [128][129] proposed the rank entropy-based decision tree for monotonic classification. They apply rank mutual information which combined with the Shannon entropy. Author shows that if the training sample is consisting monotonically then performance is still good with presence of noisy data.

C4.5 is based on the information gain ratio that is evaluated by entropy. The information gain ratio measure is used to select the test features at each node in the tree. Such a measure is referred to as a feature (attribute) selection measure. The attribute with the highest information gain ratio is chosen as the test feature for the current node.

## *Dataset consistency*

Unfortunately, as mentioned above, the application of NMR spectroscopy as a non-targeted testing tool is still hampered due to the absence of harmonized and internationally agreed procedures. The European Federation of National Associations of Measurement, Testing and Analytical Laboratories (EUROLABs) issued two technical reports paving the way to the development of NMR methods suitable

for non-targeted analysis[130]. However, in the absence of officially agreed procedures, single laboratories develop and validate their own NMR protocols. [131], [132]

A universal criterion to assess laboratory performance during spectra generation might help to overcome the lack of official non-targeted procedures. Indeed, even though sensitivity rate and specificity rate introduced as unbiased performance assessment indexes by USP work well when a number of samples are analyzed by a single instrument, they cannot be applied when comparing spectra recorded by different spectrometers. As a consequence, to date, for a single analytical technique, no reference conditions for non-targeted methods are available. Among the spectroscopic analytical techniques suitable for the development of non-targeted methods, based on theory, nuclear magnetic resonance offers the unique opportunity to generate statistically equivalent signals. Thus, when a single sample is analyzed by different spectrometers, the resulting NMR spectra can be compared, provided that the following conditions are fulfilled: (i) the statistical distribution of a selected spectral feature (e.g., signal intensity) should be normal; (ii) the standard deviation (σ) characterizing the normal distribution of the selected feature should be as low as possible with respect to its mean value (μ) (International Organization for Standardization (ISO) 2012). In this respect, the coefficient of variation (CV = σ/μ) of the selected feature represents an appropriate parameter to evaluate the statistical distribution of the selected spectral feature. Once these conditions are fulfilled, the performance of the laboratories can be assessed by several procedures. We reported on the generation of statistically equivalent NMR signals suggesting the use of a quality control index, namely, the Qp-score, to assess the performances of the laboratories. [133] For such purpose, a first interlaboratory comparison (ILC1) involving a number of NMR data sources (spectrometers which were different in hardware configuration, manufacturer, magnetic field, etc.) was organized, and the statistical equivalence of the calibration lines, built to quantify analytes contained in a five-component model mixture, was demonstrated. Importantly, this new index can find wide and easy applicability when calibration lines are available. Moving to the fingerprint of complex mixtures, for which no calibration lines are developed, performance can be assessed by referring to z-score calculated for selected signals.

Based on these considerations, food fraud fighting can be empowered by designing a data-driven system, where NMR spectra of authentic samples are collected in a single and shared database (a). Such a database stores spectra generated by NMR laboratories whose performance are previously assessed

using Qp-score and/or z-score. The available spectra are then exploited to optimize suitable classifiers (b). Finally, a number of end-users (NMR laboratories eligible because of satisfactory Qp-score and/or z-score) can enquire the developed classifiers by submitting NMR spectra of unknown samples (c).

To date, in the context of the comparison of NMR spectra recorded by different spectrometers, neither procedures for the selection of a spectral feature of a complex mixture nor indications of reference values for the corresponding CV have been introduced. Aiming at investigating the statistical equivalence of NMR signals for complex mixtures and introducing reference values for CV of the signals intensities, a second interlaboratory comparison (ILC2) was organized. In this case, the hardest challenge was to overcome the drawbacks deriving from the different field-strengths of the spectrometers which are expected to affect considerably the line width of the signals. Indeed, for given intervals, variation of signal intensities due to different field-strengths may prevent reliable

performance of the multivariate approaches. As a proof of concept, a set of aqueous extracts of wheat and flour belonging to different cultivars (cv Pietrafitta and cv Simeto, respectively) was considered. A study, including univariate statistics applied to 7 selected signals (targeted approach) and multivariate statistics applied to the whole spectrum (non-targeted approach), was carried out. [8]

## The need for large datasets requires uniform data

The development and maintenance of a food authenticity database can be very resource intensive. Therefore, the most important consideration that needs to be addressed before any sample collection or analysis can begin, is the definition of the specific purpose of a particular database. A database designed without a specific use in mind is likely to overlook considerations that are crucial for specific problems. Conversely, in databases created using non-targeted analysis, given the expense associated with creating them and the increasing use of 'bigdata', it is prudent to ensure metadata that is not directly relevant to the purpose of the database is also recorded. A balance needs to be reached, such that the recording of additional metadata is not too onerous. For example, where secondary analytical checks are performed on samples before inclusion into any database (e.g. fatty acid profiling of vegetable oils),[134] for example to verify samples of being typical or authentic by established approaches, it is recommended these metadata are included as part of the sample descriptors, where this is possible without additional cost.

The specific primary purpose of the food authenticity database will determine its applicability and inform what samples need to be analysed and the method in which they will be analysed. The reference samples, which are analysed to create the database, will define its scope. Consideration must be given to assure that samples used to populate the database are comparable to those that will be ultimately challenged by the database. For example, to create a database that is only designed to differentiate between Scottish and English beef [135], representative samples of only English and Scottish beef would be needed to populate the database. Although this is still a considerable task, it can be achieved relatively easily. If the samples are collected only during the period of 2017 though, the scope of the database would only be for beef slaughtered in the year 2017. Therefore, such a database would only be valid to challenge beef samples that are either Scottish or English that were slaughtered in 2017. If this database was challenged with English beef slaughtered in 2018 one of three outputs would occur: the challenge sample would be correctly classified as English, the challenge sample would be erroneously classified as Scottish, or the database would report that the sample could not be classified. In this case, the scope of a database could be increased through assumptions which are then validated. For example, if the factors that lead to the differentiation of English and Scottish beef remain consistent in 2018, then it is highly likely that the database will remain applicable. This can be validated by challenging the database with a limited number of English and Scottish beef samples, slaughtered in 2018. If these samples are correctly predicted, with an accuracy that is not significantly different to samples from 2017, then the assumptions and the use of this database in 2018 can be considered as valid.

If this same database was challenged with a sample from a non-English or non-Scottish origin slaughtered in 2017, one of three outputs would occur, that sample would be erroneously classified as English, the sample would be erroneously classified as Scottish, or the database would report that the sample could not be classified. A food authenticity database has no scope for identifying samples that were not considered during its creation, and in this specific example to increase the scope would be a very significant task. As a minimum, it would need to be demonstrated that the factors that led to differentiation between English and Scottish beef samples enabled discrimination between England, Scotland and all other countries. In general, truly global geographical discrimination of food samples may not be achievable. In cases where discrimination of one sample type from all others is possible,

these statements can only be considered valid for the samples that have been used to create or challenge the database. It is therefore common for food authenticity databases to be created and used to confirm whether a food is consistent with expected properties. In these cases, a specific property of a sample is measured, and suspect samples are challenged against this database, samples that do not fit the expected profile are rejected.

## *The interlaboratory comparison*

Interlaboratory comparisons (ILCs) are organised either to check the ability of laboratories to deliver accurate testing results to their customers or to find out whether a certain analytical method performs well and is fit for its intended purposes. The former is usually termed 'proficiency testing' and the latter 'collaborative method validation study'. Only those methods passing the stringent requirements of such a study can be subject to standardisation by European or international standardisation bodies (European Committee for Standardization, CEN, and International Organization for Standardization, ISO). [136]

Laboratories involved in official control activities are required to provide evidence for their competence in carrying out testing. This process is called accreditation. Accredited laboratories shall preferentially use standardised methods of analysis and are required to participate to proficiency tests for demonstrating their technical competence to their customers and to ensure comparability and acceptability of the testing results produced by them.

EU food and feed control legislation requires official control laboratories in the EU member states to use standardised methods, e.g. those issued by CEN, whenever available. The availability of standardised methods of analysis and sampling is therefore of great interest to all food chain stakeholders as it supports the uniform implementation of legislation in the EU Member States, in particular in cases where regulatory limits have been specified to ensure food safety.

In addition to protecting the well-being of consumers, standardised methods enable the free movement of goods within the EU and avoid duplication of analytical work commissioned by trading partners. When

preparing documentary standards for food safety and quality, the JRC focuses on submitting collaboratively validated methods for the determination of regulated substances in food and feed.

Examples of standardised methods of analysis developed and validated by JRC are methods for the analysis of several mycotoxins in feed and food, food additives (sweeteners), heavy metals, coccidiostats, animal-by products, and foreign fats in chocolate. The validated methods submitted to the various standard developing organisations (ISO, CEN, AOAC International) are in part the result of pre-normative research activities linked to the operation of the EU reference laboratories (EURLs) hosted by JRC.

Next to collaborative method validation studies JRC organises proficiency tests through the operation of:

- Regular European Interlaboratory Measurement Evaluation Program (REIMEP) for nuclear measurements;
- Interlaboratory Measurement Evaluation Program for Nuclear Signatures in the environment (NUSIMEP);
- International Measurement Evaluation Program (IMEP).

The results of proficiency studies inform European Commission departments about the state-of-the-art in certain areas relevant for policy making, e.g. whether the currently available measurement techniques are suitable for enforcing an envisaged regulatory limit. The European co-operation for Accreditation (EA) profits as well as accreditors get an overview of the comparability of test results provided by accredited laboratories in the different countries; lastly, laboratories have a means to identify problem areas which allows them to take the necessary actions to remedy the shortcomings. [137]

## Methods to uniform the datasets

In order to establish whether the inter-laboratory CV% represents a valid threshold for assessing the statistical equivalence of the scaled NMR signals, we explored a non-targeted approach by applying both unsupervised and supervised multivariate statistical analyses [138]. In place of the signals S1–S7, 190

regular intervals (buckets) of 0.05 ppm width, deriving from NMR spectrum segmentation, were analyzed. The underlying area of each bucket was calculated and normalized to TSP signal integral [0.5, –0.5] ppm. At this stage, it is important to outline that the comparison of the areas calculated for the same bucket may be strongly affected by the different signal resolutions deriving from the different field strength of the spectrometers (400 to 700 MHz).

Multivariate analysis was performed according to the following approach. For each NMR tube, the spectra provided by 36 spectrometers were divided randomly into two groups. The first group of spectra, including all the repetitions recorded by 28 spectrometers, was used to build the classification models, while the second group, including all the repetitions analyzed by the 8 remaining spectrometers, was used to validate the model. Among the repetitions of the first group, 3 repetitions of each of the 28 spectrometers were used as training set and repetitions as test-set. The 5 repetitions provided by the remaining 8 spectrometers were employed as an external validation set. Summarizing, for each sample, 84 spectra were used for training the model, 56 for testing and 40 for validation.

Initially, PCA was performed on the training sets to compare wheat samples B (cv. Pietrafitta) and C (cv. Simeto), and flour samples D (cv. Pietrafitta) and E (cv. Simeto). In both cases, significant clustering of the two classes of spectra was observed along the first two principal components t(1) and t(2) (Fig. 4). Next, the PCA-class analysis was performed (see supporting information for further details), and the resulting local PCA models were used to classify the test and validation sets by applying the soft independent modelling of class analogy (SIMCA). As represented in the Coomans' plots reported in Fig. 5, the majority of the observations fell within the class membership limit (DCrit), thus fitting the class models. Specifically, when SIMCA was applied to the test set, a satisfactory fit with the model was observed. When the validation set was subjected to SIMCA, [139] a more noticeable scatter was detected. Such a result was not surprising, as the spectra included in the validation set did not derive from the spectrometers involved in the classification model building. Importantly, in all cases, no class confusion was operative since no class overlapping occurred.

Suitable multivariate statistics for creating models are described, in detail in the literature[140], [141]. In brief, multivariate treatment of the data will produce a model able to classify samples as authentic or non-authentic, or into other classes (e.g. different origins). Two stages are normally completed: a data compression step which reduces the size and complexity of the original data, and a modelling step that

is carried out on these selected features. The principal statistical methods for creating models are discriminant analysis and class modelling. Discriminant analysis is appropriate for determination of a value (e.g. the oxidation level in olive oil [142]), whereas class modelling is used either to define normality for a single class (e.g. is a sample authentic) or for multiple classes (e.g. is an olive oil Italian, Spanish or Greek).

Validation of a database includes both the data within it, and its ability to satisfactorily complete the role for which it was designed. All data used to create the database, must be validated, i.e. reliable. This implies that the laboratory producing the data must demonstrate competence and accuracy. For example, in the case of the EU-Wine databank, laboratories providing data must be accredited according to ISO17025 and must participate in a proficiency test. The best way therefore to validate the data is for laboratories participate in interlaboratory comparison exercises. Where appropriate, available proficiency tests can be used and when these are not available AD hoc round robin tests can be organised. When laboratories upload their data to databases, which are created by several organisations, they should provide indications about their performance in the inter-laboratory comparison exercise in terms of a z-score or deviation from the target value.

For non-targeted databases, the organisation and use of interlaboratory comparison exercise is very challenging [143]. Hence, metadata describing the measurement device parameters and protocols would be more useful to be included in the database, to demonstrate the reliability of the laboratory and provided data.

## A community build calibration line

Relative or absolute quantification of an analyte in a matrix consists of a sequence of operations carried out under defined and agreed methods, which are developed according to technical specificity of the analytical tool, the analyte and the matrix. Among the operations required to quantify a substance, calibration processes deserve special consideration. [2] According to the International Bureau of Weights and Measures, calibration is defined as “Operation that, under specified conditions, in a first step, establishes a relation between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties (of

the calibrated instrument or secondary standard) and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication" [144]. Usually, a calibration curve is developed to find the optimal equation correlating the response of the selected analytical technique and the concentration of a set of standard samples of the analyte with known concentration. Such equation is used to derive the concentration of the analyte contained in an unknown sample. The calibration conditions vary with time and instrumental use thus, depending on the specific analytical issue, a periodical update is required. Therefore, the calibration process has an important impact on both cost and execution time of the analysis.

Establishing metrological traceability is a prerequisite to obtain a reliable metrological comparability of the measurement results produced at different laboratories and at different times. As stated by De Bièvre et al. achieving metrological comparability of measurement results requires the definitions of concepts of calibration hierarchies providing metrological traceability chains, which enable the establishment of metrological traceability of measured quantity values to a common and stable metrological reference. [145] Ideally, in order to produce reliable and traced measurement results with the corresponding measurement uncertainties for a given method, the whole analytical pipeline should be based on a community-built system able to simultaneously manage calibration and traceability steps for many

operators. Such a system should i) collect calibration data produced by many operators, ii) process data to develop a community-built reference calibration curve and iii) provide results (exploiting the community built reference calibration curve) to many operators after submission of data regarding unknown samples. Many advantages may derive from using such system. First, a number of community-built reference calibration curves and, thus, reference concentration values for different analytes in different matrices can be developed. Moreover, such community-built database could be continuously made more robust by introduction of new calibration data, produced also by laboratories not directly involved in the initial calibration curve building. Therefore, the reference calibration data (curve parameters and predicted analyte concentration) will become more precise and accurate and could be ultimately used as reference values to test the performance of the laboratories.

The only requirement for creating this community-built system is the use of an analytical technique able to generate, for a given sample, statistically equivalent signals. In other words, any sample should produce the same instrumental response when analysed by different instruments.

In this context, recently, by means of interlaboratory comparisons (ILCs), we demonstrated that NMR spectroscopy can provide statistically equivalent signals when the same sample is analysed by spectrometers that are different in terms of magnetic field strength, manufacturer, hardware configurations and age [1], [4]. Indeed, the exclusive correlation between the resonance frequency of a signal and the type of nuclei associated to that signal, makes NMR spectroscopy a powerful technique for structural determination and quantification[146], [147]. Since the area of a NMR signal is linearly proportional to the number of NMR active nuclei generating the signal, the response factor (ratio between the signal produced by the analyte and the quantity of analyte which produces the signal) is independent of the molecule and the analyte quantification can be achieved directly by calculating integral of the NMR signal. Moreover, the design of new pulse sequences for FIDs acquisition [148]–[152] and novel algorithms for data processing [36], [153]–[155] enhanced the capability of NMR for discriminating among very similar compounds contained in complex mixtures, as pharmaceutical, natural products, agrochemicals, foodstuff, and biofluids. Nevertheless, to date few official protocols have been reported which employ NMR technique for purity assessment and quantification purposes. While the experimental conditions (pulse sequence, acquisition parameters, postprocessing strategy) assuring the intra-laboratory repeatability are well established [156], still few studies are available that discuss the reproducibility assessment of quantitative NMR (qNMR) data obtained when the same sample is analysed by different operators and/or by spectrometers with variable features (manufacturer, B0 field strength) [39]–[41], [157]. Taking into consideration the extensive application of qNMR in different fields of chemical science, it appears as a matter of urgency to overcome this significant shortcoming and make qNMR an internationally accepted standard analytical technique.

In this paper, we provide a pipeline to assess the reproducibility of NMR data produced for a given matrix by spectrometers from different manufacturers, with different magnetic field strengths, age and hardware configurations. Moreover, we introduce a community-built quantification system able to perform quantitative analysis and to assess performance of the laboratories. The aspects affecting the interlaboratory reproducibility are also discussed. Specifically, by exploiting the big amount of

spectroscopic data produced during an interlaboratory comparison involving 65 spectrometers from 12 countries, the concentrations of four selected metabolites (alanine, arginine, glucose, and fructose) contained in the grape juice (cv. Primitivo) are predicted VIA calibration lines developed by standard addition method. A sequence of appropriate chemometric tests (Figure 17, reference calibration system development) are applied to the predictive models developed individually by the ILC participants. Upon evaluation of the likely sources of error, it was established a strategy for assessing the performance of the laboratories during the different stages of the quantitative analysis. The well-performing models were tested to predict the unknown concentrations of the metabolites contained in a test sample and the obtained data were evaluated in terms of reproducibility, allowing for the identification and validation of statistical equivalent signals in the NMR spectrum (Figure 17, performance assessment).

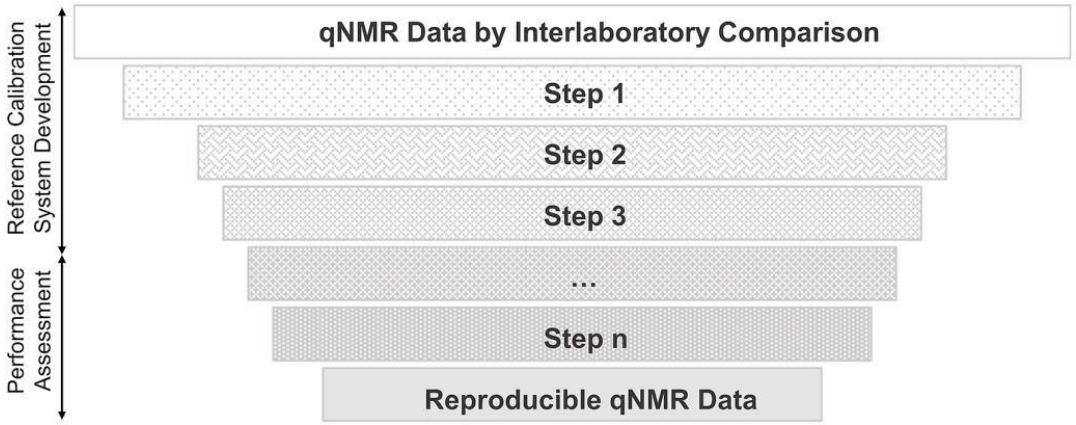



**Figure 17. Schematic representation of the selection process for the evaluation of the prediction models designed by 65 ILC participants**

## Different machine same languages

Food control has been historically achieved through a direct approach, namely by identification and quantification of a primary marker indicated as responsible for a food authenticity issue according to specific legal limits (targeted approach). Nevertheless, the possibility to obtain a larger amount of data more rapidly made the use of nontargeted approaches progressively more common for food control thanks also to the many advances in the analytical techniques and in the chemometric applications [138], [158], [159] . Non-targeted methods offer the possibility to extract rapidly and in non-destructive

way information which can be advantageously used to unveil the compounds that may affect the authenticity of the food sample under investigation. Such analytical methods can be performed according to two alternative approaches, namely the profiling and the fingerprinting. In the first case (profiling) the identity of the compounds of interest is well known and established before the statistical data elaboration. Conversely, in the second case (fingerprinting) the analysis is performed with no a priori identification of the compounds contained in the sample mixture [160]. Both the aforementioned approaches can produce a large amount of data that can be exploited to assess the authenticity of a big variety of food products. Nuclear Magnetic Resonance (NMR) spectroscopy is gaining growing attention in this field, as demonstrated by an increasing number of applications reported in the recent literature [161], [162] The interest in non-targeted NMR methods is mainly due to its ability to generate highly reliable instrumental responses [163]. Indeed, when a single sample is analyzed by different NMR spectrometers, statistically equivalent NMR spectra are generated. This aspect opens the way to the creation of a community-built system containing NMR spectra which can be safely compared and can be exploited to solve many analytical issues. For instance, for a given food fraud problem, as schematically represented in Figure 17, NMR spectra of several samples, suitably selected to represent a class of a food product, may be provided either by a single spectrometer or by different instruments according to an agreed and validated procedure (including sampling, sample preparation, spectra acquisition, and processing details). The repeatability and the reproducibility of the produced spectra should be verified upon the application of opportunely defined criteria (Figure 17, step 1). Then, only the laboratories producing comparable NMR spectra should be eligible for feeding the database containing NMR spectra of food samples (Figure 17, step 2). The stored NMR spectra would be exploited to develop a classifier properly designed to unveil the fraud (Figure 17, step 3). Finally, the same laboratories which resulted eligible to feed the database (admitted to step 2) could test the classifier by submission of the NMR spectra of an unknown sample. As a result, the commodity class, and, ultimately, the authenticity of the unknow sample should be established (Figure 17, step 4).

Despite the great interest in the described non-targeted NMR method to date no standardized procedures (protocols and materials) have been introduced to apply routinely this analytical strategy for the detection of food counterfeits and determining the authenticity of food products. In the context of an ongoing project, we gave a contribution to the harmonization of the experimental procedures of the

NMR methods in food control. Based on the large amount of data produced by interlaboratory comparisons (ILCs) [133], we demonstrated that targeted and non-targeted NMR methods can provide comparable results when the same sample is analyzed by spectrometers that are different in terms of magnetic field strength, manufacturer, hardware configurations and age. In particular, two selection criteria were adopted to assess the statistical equivalence of the spectra produced by different spectrometers during an interlaboratory comparison: a quality parameter, Qp-score, and the interlaboratory coefficient of variation, CV% [2]. Besides, exploiting the unique capability of NMR spectroscopy compared to other analytical techniques to generate equivalent signal intensity regardless of the spectrometer configuration [164] ,we developed an NMR-based community-built calibration system which was able to assess the performance of the laboratories and to perform quantitative analysis (qNMR) [2]. Nevertheless, one inherent issue observed when the same sample is analyzed by different spectrometers is that, while the intensity of the NMR signal is usually independent on the spectrometer configuration, the shape and the resolution of the signal is subjected to small variations which, not surprisingly, can affect the reliability of the non-target analysis. Indeed, the magnetic field strengths and the procedures adopted for the pre-treatment of data (normalization, peak alignment, scaling) play a crucial role to obtain high levels of repeatability and reproducibility of statistical results. [165]–[167] In the present paper, we explored the effect of data-pre-treatment (buckets size and data scaling) on the performance of a class-discrimination system upon the statistical elaboration of the large number of data produced during an interlaboratory comparison. As proof of concept, samples of grape juice extracted from two different cultivars (cv.), Primitivo and Negroamaro, were analyzed by 65 different spectrometers applying the same protocol. Both the profiling and the fingerprinting approaches were explored and the chemometric analysis was based on i) a training set constituted of 100 NMR spectra recorded by a single spectrometer for 50 grape juice cv. Primitivo and 50 grape juice cv. Negroamaro and ii) a test set constituted of 650 NMR spectra produced by 65 different NMR spectrometers for one grape juice sample cv. Primitivo and one grape juice sample cv. Negroamaro (5 repetitions per sample per spectrometer). This study should demonstrate that the judicious pre-treatment of data is crucial to make the spectra produced by different spectrometers statistically equivalent. Only in this case, they may be used for the development of classifiers able to predict the commodity class of a food sample and, thus, allow to assess its authenticity. Considering the high

throughput of non-targeted NMR methods, this potentiality is of great interest to the scientific community involved in food control.

# Case studies

## Project REGEVIP

### Project presentation

Puglia is a region rich in agricultural bioversity, in particular viticultural, for historical and geographical reasons.
By virtue of its position in the Mediterranean, Puglia has been a crossroads of trade and a meeting point for people and peoples for thousands of years. For these reasons and for the vocation of the territories, it is the home of many native or local vines, often in danger of extinction (minor vines), little known, forgotten or known with wrong names, sometimes related or synonyms of more important vines elsewhere. [168]

Recently the market is rewarding those few entrepreneurs who have successfully invested in some ancient genetic resources, bringing them back to cultivation and labeling.

The integrated project for the recovery of the Apulian Viticultural Germplasm has the objective of ensuring the conservation of the intraspecific and intravarietal viticultural biodiversity, improving the knowledge on the productive and technological characteristics of the Apulian vines, restoring and registering the propagation materials in the National Register of Vine Varieties to allow its use in accordance with legislation.

The project activity was based on the acquisition of spectra obtained with the analytical technique of Nuclear Magnetic Resonance (NMR) from wine grapes. The construction of a database of NMR profiles and the development of an expert system for the automatic classification on a varietal basis of the grapes of unknown native vines concerned the following subordinate activities:
the search for quality control criteria of the analytical data generated by the NMR spectrometer;

the search for comparison criteria of NMR data generated by different software applications.
Sampling, storage and preparation for NMR analysis of the wine grape samples were carried out using the laboratory's previous experience in analyzing table grape samples.
Given the standardized procedures, in the laboratory activities of Innovative Solutions it was decided to work only on an analytical replicate but, with a high number of biological replicates of each sample of wine grapes that represented the different geographical origin for each variety.
The setup of the user spectrometer for the execution of one-dimensional 1H NMR experiments (Bruker Avance 400 NMR spectrometer with automatic sampler) has been optimized and first verified of each analysis in accordance with the requirements of the Quality Management System of Innovative Solutions. The quality of instrumental performance was assessed in accordance with the defined Operating Procedures.[169]
The realization objective, to which the activities carried out refer, therefore provides for the application of the NMR technique for the definition of the fingerprint of table grapes of the cultivars to be used for the production of typical Apulian wines. This objective arises from the need to fill the gap in analytical methods capable of indicating the varietal and geographical origin of the products.

## Materials

The European vine (vitis vinifera) is one of the species of the vitis genus, which contributes to the use of table grapes for fresh consumption.
The popular varieties have berries of medium or large size, crunchy consistency with thin skin, non-astringent, bright in color with a balanced sugar-acid ratio, with a good level of sugar (14-15% in the juice) and a fruity aroma.
In typically Mediterranean climatic conditions, table grapes, if adequately supported by irrigation and correct agronomic practices, reach the most balanced ripening profiles with the highest taste levels.
From a functional point of view, the elementary vegetative unit of the plant is the fertile shoot. It originates from a hibernating bud, so called because it sprouts the spring following the season in which it was formed, after having spent a winter in quiescence. When the hibernating bud undergoes the process of flower differentiation, the bud it gives rise to is fertile and generates one or more clusters.

The bunch of grapes has a structure consisting of a central axis whose basal part (peduncle) splits into two branches, one of which, the rachis, represents the central axis of the bunch, while the other can develop either in tendril precociously caducous, or also in the rachis.

The cultivars selected to produce table grapes, compared to those for wine grapes, are characterized by the ability to develop large clusters often characterized by the fertility of the branching of the peduncle, by the large development of the basal branches, as well as by the high length of the axes of the rachis and its branches.

From a botanical point of view, the fruit of the vine is a berry and consists of the skin or dermal system, the pulp and a thin internal epidermis that separates the pulp from the seminal chambers. The growth of the berry both in weight and in volume is described by a double sigmoid curve that can be divided into at least three successive phases

• **Phase 1** also called herbaceous in which the berry maintains vegetative characteristics and accumulates malic and tartaric acids

• **Phase 2** or stasis, in which a reduction in biosynthetic activities occurs and the production of malic acid is the main activity

• **Phase 3** or maturation, during which the berry profoundly modifies the mechanical and composition characteristics, and the accumulation of sugars occurs

After the third phase, if the fruit remains on the plant for a long time, the over-ripening phase can also be identified, in which there is a gradual loss of water and consequently the concentration of all metabolites.

The quality of table grapes, understood as the set of nutritional characteristics, depends on the entire ripening process of the fruit but the phase that offers the possibility of easily evaluating the metabolic profile and therefore those qualities of the fruit that are perceived by the consumer. is certainly the ripening phase.

Throughout the maturation there is a progressive accumulation of simple sugars, glucose and fructose, which, when ripe, can be present in quantities close to 20% of the fresh weight of the berry juice. The titratable acidity of the juice, expressed in grams of tartaric acid contained in 100 ml of juice and which can be determined using automatic titrators, is reduced and there is a progressive increase in pH. Tartaric acid, the production of which occurs mainly in the herbaceous phase, is diluted during ripening

following the growth of the berry which practically consists in the accumulation of water. The number of metabolites present in table grape juice varies in well-defined intervals, typical for each cultivar and its determination can be done in a very detailed manner through the use of nuclear magnetic resonance spectroscopy.

The characteristic profiles of the metabolites that emerged from the studies supported the idea of being able to classify table grapes by analyzing the characteristic concentrations of the metabolites themselves, present in the juice of the ripe berry.

## Chemometrics Results

All data acquired by NMR were subjected to Multivariate Statistical Analysis, a data processing technique used to search for relationships between different blocks of variables. Among the multivariate analysis techniques, the Principal Component Analysis (PCA) is a method frequently used in the first phase of data processing because it serves to:

- give an overview of the problem;
- understand the relationships between the samples and / or the classes considered;
- provide a preliminary indication on the role of the variables, possibly highlighting the possibility of eliminating some which, being closely related to each other, carry similar information and can therefore be considered redundant.

PCA is a factorial method, as it allows the reduction of the number of variables through the construction of new synthetic variables, called principal components, obtained from linear combinations of the initial variables by means of "factors".

As output, the PCA gives us two graphs:

- Loadings plot in which the measured variables are represented in the form of vectors.
- Scores plot in which the samples are represented.

In our case, the bucketing procedure, necessary for PCA, was carried out using the following method: Single rectangular buckets; Left border: 10.5; Right border: 0.5; Bucket width: 0.05; Integration mode: sum of intensities; Scaling: scale to total intensity; Exclusions: 5.20-4.10; 2.25-2.20.

Below are two examples of varietal discrimination between 3 white wine grape cultivars and 3 black wine grape cultivars, using signals related to the amino acid spectral region.

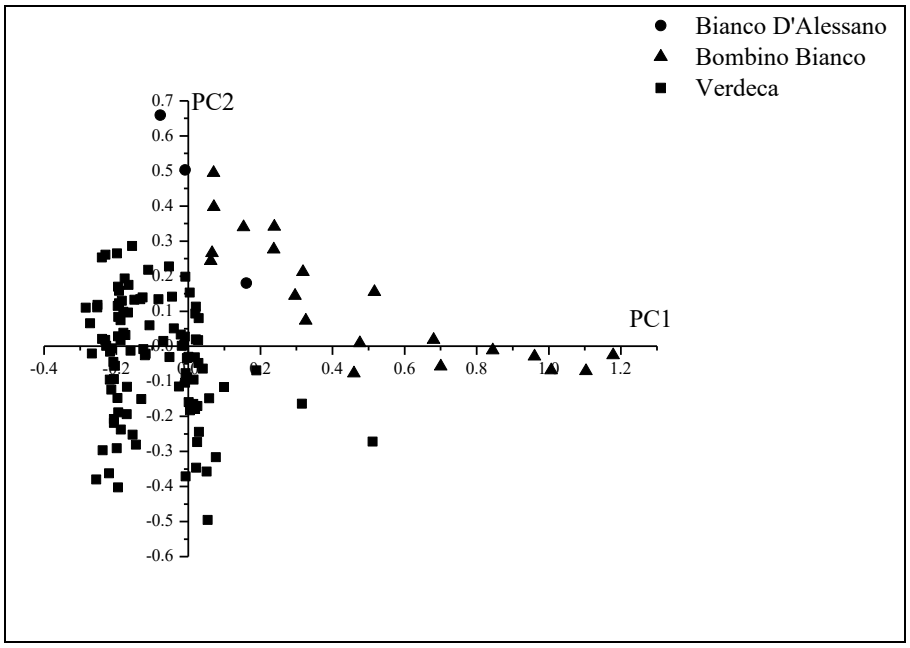



**Figure 18. PCA of 3 cultivars of white wine grapes**

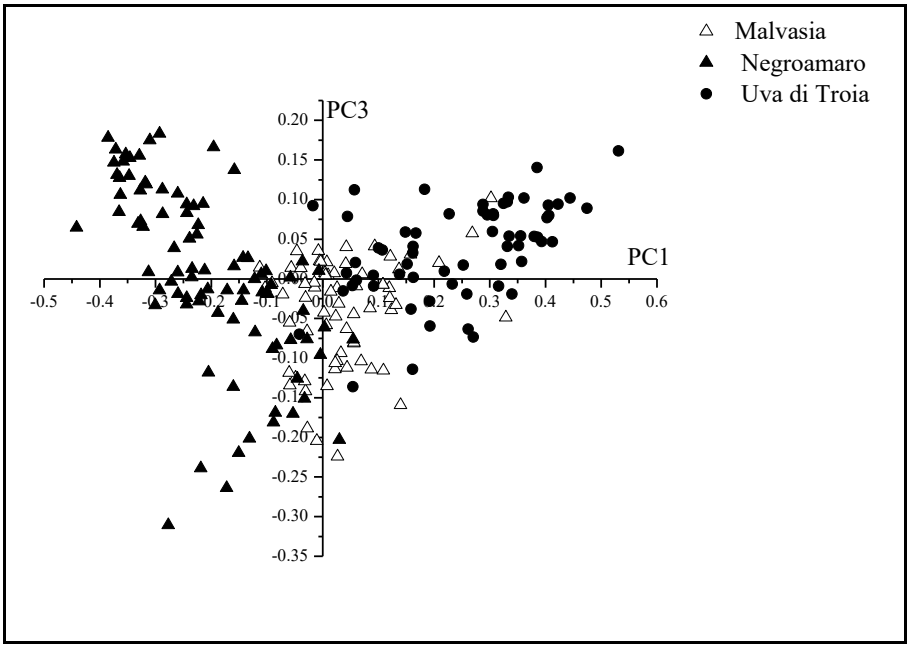



**Figure 19. PCA of 3 black wine grape cultivars.**

The NMR analysis made it possible to identify the metabolic profile of the selected wine grapes, highlighting the intra and inter variety variations for the harvest years. In particular, the primary and most abundant metabolites such as: leucine, isoleucine, valine, alanine, arginine, fructose, glucose, formic acid, citric acid, have been characterized by means of literature studies, consultation of NMR databases and use of reference standards. malic acid, tartaric acid, lactic acid and ethanol.

In the metabolomic study each acquired spectrum has an internal reference (TSP) whose signal falls to 0 ppm, furthermore each spectrum has been divided into three areas of interest:

1. 10 6 ppm the area of aromatic compounds;
2. 6 3 ppm the carbohydrate zone;
3. 0.5 ppm the amino acid zone.

Indicator signals have been identified for each area of interest:

- TSP (between 0.1 and -0.1 ppm);
- Valine_leucine (between 1.1 and 0.6 ppm);
- Glucose (between 5.3 and 5.1 ppm);

These signals were integrated, quantified and repeatability and reproducibility tests were performed on them.

## Machine learning results

After all the analyzes carried out, it is evident the need to summarize the performances obtained in order to indicate an effective methodology for the classification of wine grapes.
We have seen from the results that even with as many as 10 different cultivars the classification algorithms are able to successfully discern all the qualities, and from the confusion matrices it can be seen how the confusions concentrate on wines that are however proven to be similar.
Finally, the almost perfect ability to distinguish two different wine grapes (Negoamaro and Primitivo) is a further confirmation of the goodness of the method.

**Table 1. Machine learning alghoritms performance sintetic results**

| | Classificatore | | |
|---|---|---|---|
| | *j48* | *RF* | *ANN* |
| | *Acc %* | *Acc %* | *Acc %* |
| Cultivar Uve Raccolto 2013 | 53 | 63 | 89 |
| Cultivar Uve Raccolto 2014 | 72 | 80 | 88 |
| Cultivar Uve Raccolto 2016 | 56 | 71 | **88** |
| Cultivar 2016 - SubSpettro 3-1 | 52 | 69 | 81 |
| Cultivar 2016 - SubSpettro 3-1 - PCA | 48 | 71 | 72 |
| Cultivar 2016 - CFSSUBSETEVAL | 59 | 76 | 85 |
| Cultivar 2016 - PCA | **62** | **81** | 87 |
| Cultivar 2016 Primitivo-Negroamaro | 84 | 95 | **95** |
| Cultivar 2016 Primitivo-Negroamaro - PCA | 83 | 91 | 88 |
| Cultivar 2016 Primitivo-Negroamaro - SubSpettro 3-1 | 86 | 87 | 93 |
| Cultivar 2016 Primitivo-Negroamaro - SubSpettro 3-1 - PCA | **87** | 87 | 92 |
| Cultivar 2016 Primitivo-Negroamaro - CFSSUBSETEVA | 84 | **96** | 93 |
| Cultivar Uve Raccolti 2013-2014-2016 | 57 | 77 | 90 |
| Colore Uve Raccolti 2013-2014-2016 | 76 | 85 | 95 |
| Anni di Raccolta 2013-2014-2016 | 82 | 95 | 99 |

Project IntelliTrace

## Project presentation

**WP18. Improving comprehensive artificial intelligence, validation and harmonization methods, as "functional bridge" between non-targeted analytical approaches and tracking/authentication of food within the Food Integrity field** (*INTELLItrace*)

The analytical authentication/traceability of foods, considered in its many forms and directions (species/variety authentication; purity/integrity authentication; geographical traceability), is a complex problem that affects food safety, having a substantial impact on the consumer. Other than targeted methods, non-targeted analytical methods (e.g. complex fingerprint profiles generated by NMR, MS analysis, genomic sequencing, chromatography, FTIR and other techniques) can be considered new advanced tools that help to protect the EU consumer. According to the Gap n.1 of the Food Integrity Project, a specific process of validation of non-targeted methods is still lacking, and strictly required. Moreover, since non-targeted methods deal with large datasets, it is fundamental to have at disposal the most powerful post-analytical data processing tools, by exploring the application of advanced algorithms (e.g. Artificial Intelligence), with the aim of selecting the best performing non-targeted approach. [170]

Main goals of the INTELLItrace Project will be:

- applying advanced data mining techniques on old and new datasets generated by non-targeted fingerprinting. Old databases will concern wheat and honey, while the new datasets will derive from the analysis of rice, fish, honey and saffron;

- statistically validating the best performing approach, selected from the pool of all the analytical procedures employed during this Project to address the traceability/authentication issue;

- drafting a White Paper that will provide a guideline to all the stakeholders for approaching the statistical validation process, potentially opening new scenarios for the certification of international analytical protocols.

A skilled, well inter-connected and interdisciplinary Consortium (formed by public and private Research Centers, as well as private Companies) will work in collaboration with FI Official Partners in order to reach these goals and to respond to the request of the FI Procurement Call.

## Materials

## Chemometrics Results

The selected NMR signal integrals were scaled to the TSP integral and the corresponding ($I_{signal}/I_{TSP}$) 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. ($I_{signal}/I_{TSP}$) values were uploaded reporting at least four decimal places. The five ($I_{signal}/I_{TSP}$) 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, $I_{signal}/I_{TSP}$ 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 $I_{signal}/I_{TSP}$ 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 $I_{signal}/I_{TSP}$ ratio, according to the flowchart suggested by Horwitz,[i] 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 $I_{signal}/I_{TSP}$ 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 $I_{signal}/I_{TSP}$ value (Average),[ii] the inter-laboratory standard deviation (□), the coefficient of variation (CV%), the repeatability variance, the reproducibility limits and other statistical parameters.

**Results**

The results are summarized in the *scores* plot (Figure 20 and Figure 21) obtained by PLS (partial least squares) regression and in the confusion matrixes (Table 2 and Table 3). Figure 20 and Table 2 refer to the wheat extracts, while Figure 21 and Table 2 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.

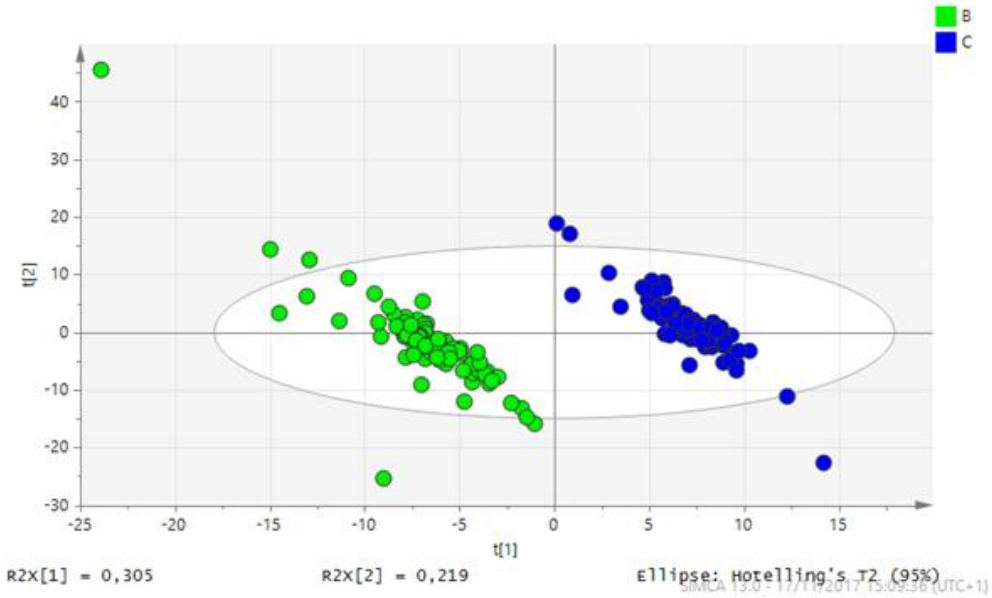



**Figure 20. Scores plot deriving from PLS regression applied to NMR spectra of tubes B and C (B and C refers to two different aliquots of the same wheat sample). Variables were scaled to unit variance.**

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

| | NMR Spectra | | $B_{true}$ | $C_{true}$ | Accuracy |
|---|---|---|---|---|---|
| | 73 | $B_{predicted}$ | 73 | 0 | |
| | 73 | $C_{predicted}$ | 0 | 73 | |
| Total | 146 | | 73 | 73 | **100%** |

Accuracy is calculated according to this equation.

$$Accuracy = \frac{(B_{correctly\ predicted} + C_{correctly\ predicted})}{Total\ population} \cdot 100$$

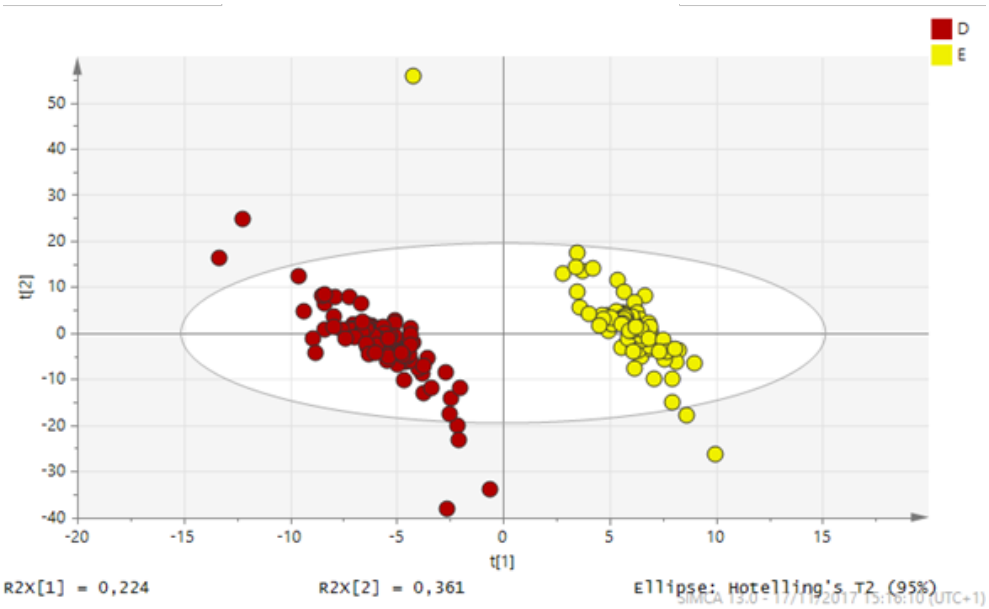



**Figure 21. Scores plot deriving from PLS regression applied to NMR spectra of tubes D and E (D and E refers to two different aliquots of the same flour sample). Variables were scaled to unit variance**

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

| NMR Spectra | | $D_{true}$ | $E_{true}$ | Accuracy |
|---|---|---|---|---|
| 73 | $D_{predicted}$ | 73 | 0 | |
| 73 | $E_{predicted}$ | 0 | 73 | |

| | | | | |
|---|---|---|---|---|
| Total | 146 | 73 | 73 | **100%** |

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

## Machine learning results

The statistical analyses are performed on the dataset prepared for the validation.

The dataset has been used to train and to validate a set of classifiers as:

- Decision trees as Random Forest (RF)
- Multi Layer Perceptrons (ANN)

**Table 4. Machine learning algorithms performance synthetic results**

| | *RF* | *ANN* |
|---|---|---|
| Wheat or Flour | 100,0% | 100,0% |
| Wheat Lot | 98,9% | 100,0% |
| Wheat Cultivar | 97,4% | 99,7% |
| Flour Lot | 94,9% | 99,1% |
| Flour Cultivar | 94,7% | 97,8% |
| Wheat Lot from Flour | 94,9% | 99,1% |
| Wheat Cultivr from Flour | 94,7% | 97,9% |

10-Fold cross-validation was selected as cross-validation strategy for all the tests, and subsequently the complete datasets have been analysed with the classifiers mentioned above.

All the classifiers are trained and validated on this dataset, searching for the classifier setting, for each classifier, 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, a comparison between the performances of the used classifiers is needed, in terms of the validation parameters outcomes provided as shown in Table 4.

The results obtained in the test are quite outstanding. Considering the first discrimination, a 100% accuracy can be expected, we can easily consider that the difference between raw wheat an processed flour is quite deep and pose no challenge to the algorithms to discriminate between the two different state of the sample.

Proceeding to the other test, we can see that the results are quite high across the board, an interesting result on the table is the link we trace between the processed wheat and the origining lot.

This kind of connection implies that a real paperless certification is possible, as the characteristic of the originating lot are ‘engraved’ inside the composition of returned flour, this can be used to track large batches of products across international shipments where some degree of sophistication is possible and the classic paper certification may be not enough.

The trained algorithms have been implemented in a web interface (Figure 22) where is possible to examine a demo of the system. This kind of approach can be used and expanded to meet industries and client’s needs.

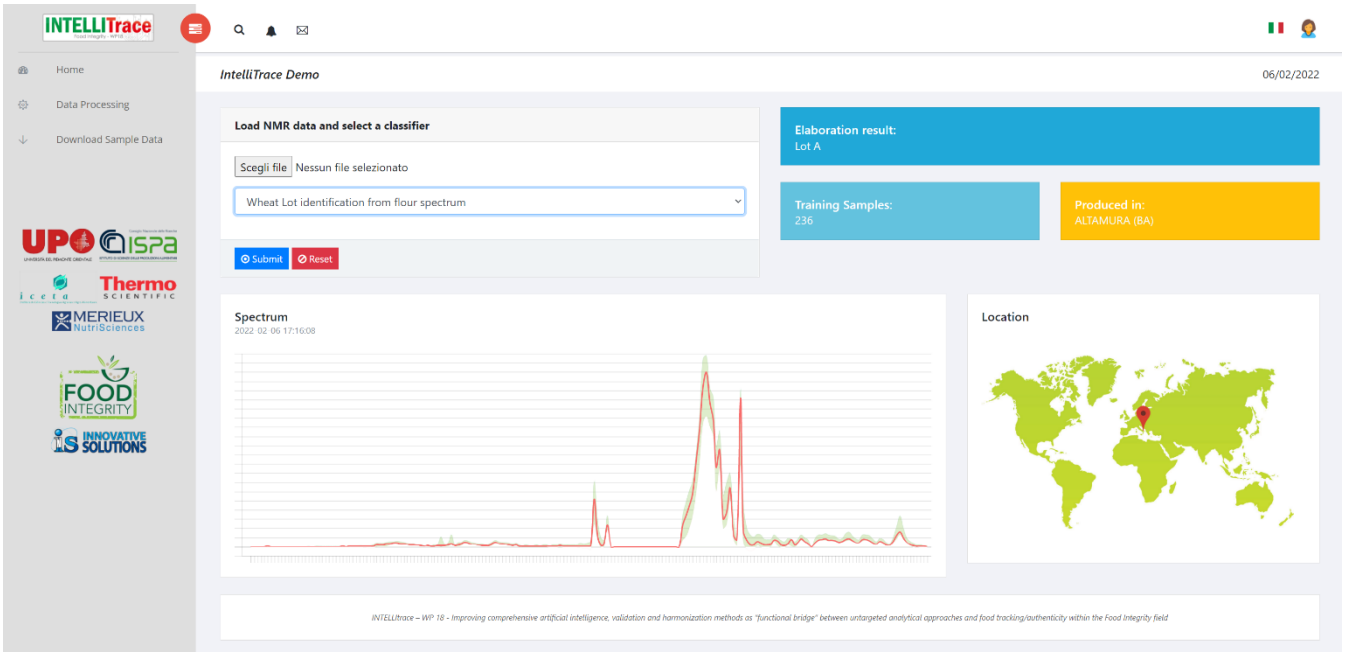



**Figure 22. Intellitrace Processing Interface Demo http://www.innovative-solutions.it/Intellitrace/index.php**

Project PASQUA

## Project presentation

The project stems from the increasingly pressing need for greater eco-sustainability of agricultural production and greater traceability of the agri-food product (required by both the consumer and the producer). The transformation of organic waste into mixed composted soil conditioner (or quality compost) to be used as organic fertilizer could be a positive process for environmental protection. In fact, the advantages are many: soil fertility is preserved (structure, ability to absorb and release water, ability to retain nutrients in an easily assimilable form, useful biological activities), reducing the use of chemical fertilizers and, at the same time, recovering a certain volume of waste.

Already widespread in gardening and in the floricultural sector, the use of compost is rather limited in the fruit and vegetable sector, mainly due to the scarce amount of scientific data about the qualitative and quantitative effects on production. [171]

The main objective of this project was to provide the agricultural production sector with an advanced analytical system, which would allow to evaluate the effect of the use of compost on the metabolic characteristics of agri-food products compared to traditional agronomic practices and to discriminate between products obtained with practices different. To this end, an innovative approach was used, the metabolomic one, based on nuclear magnetic resonance (NMR) spectroscopy, associated with the creation of a classification software. The NMR analytical technique is reliable, allows robust and repeatable analyzes with reduced sample manipulation, and produces a spectrum that represents the metabolic footprint of the tested sample (metabolite fingerprinting) as it reveals all the observable metabolites contained therein.

## Materials

## Chemometrics Results

### 1.1 Data collected and processed by standard soil analyzes.

All the analyzes of the soils (pre and post-compost) were collected in an Excel file and divided by crop and type of agronomic practice.

Values of 2-3% of organic matter are considered normal for a good soil. The soils examined (except those intended for tomato cultivation) were found to have values below the limit of 2% before the administration of the compost. Therefore, these soils needed continuous fertilization using fertilizers with a high S.O. like compost.

The collected data highlighted that the percentage content of organic matter (% S.O.) increased significantly after the administration of the compost, especially in the land destined for the cultivation of grapes and wheat (Figure 23).

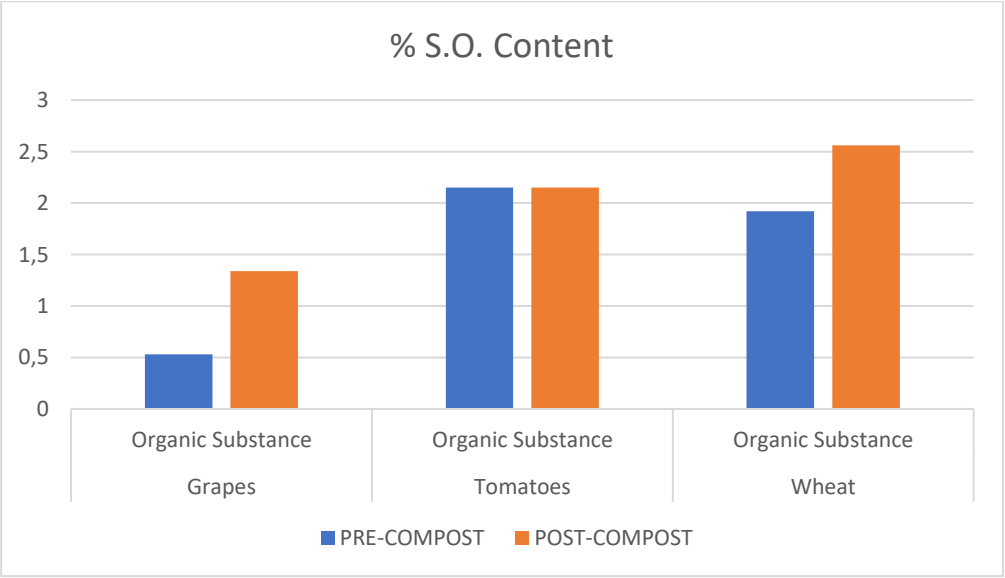



**Figure 23. Content in percentage of organic substance (% S.O.) before (blue) and after (red) the administration of the compost.**

### NMR Data

All the NMR spectra obtained were processed, aligned, normalized, and digitized in the form of a numerical matrix (bucket-table) for subsequent analysis and entry into the database.

All files (acquisition and processing) have been saved on the hard disk; in addition, all the information on each sample was entered into the IS-Tracer platform previously developed by IS (Figure 24). In summary, the insertion procedure is shown. From the "Data Input" command a window appears in

which you can add a new record. A unique ID number is assigned to the record and a screen opens with blank fields to fill in, including sample supplier, matrix type, sample description, sampling and storage temperature, acceptance temperature, test method, date and time of sample preparation, date and time of the experiment NMR, in addition to the "bucketized" NMR spectrum.

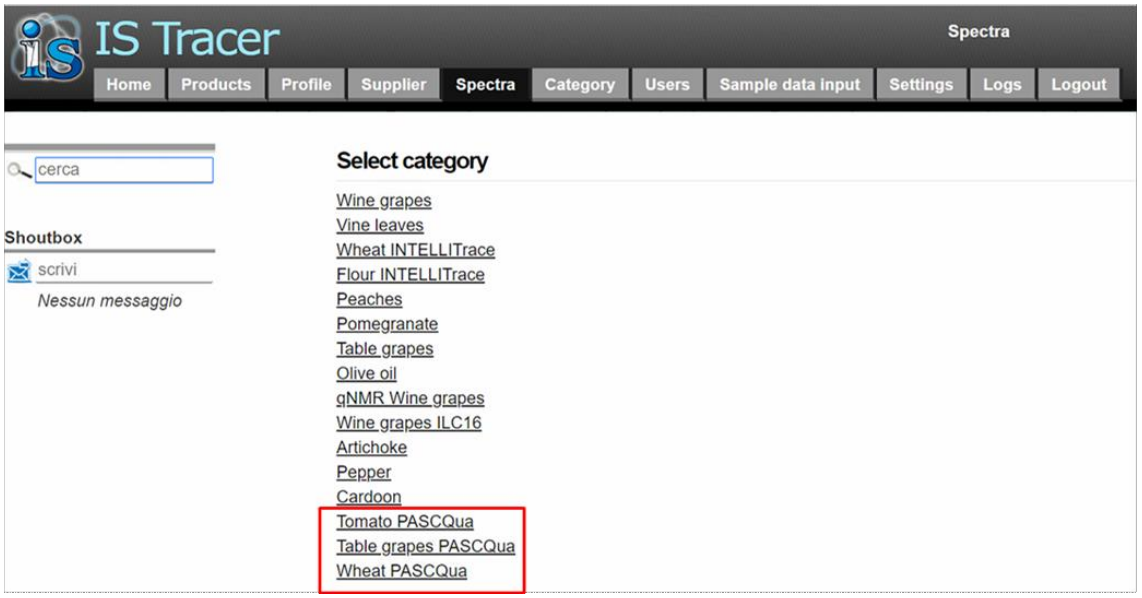



**Figure 24. IS-Tracer platform with profiles of fruit and vegetables activated**

The metabolic characterization, i.e. the recognition of NMR spectrum signals, was carried out through the use of scientific literature and by comparison with spectra present in databases available online for free.

The analytical data were subjected to multivariate statistical analysis to create classification models and find associations between the metabolic characteristics of the fruit and vegetable product and the agronomic practice adopted for its production.

**Processing of Table Grape NMR Data**

The Principal Component Analysis (PCA) of the NMR data of the grape juice samples was carried out.

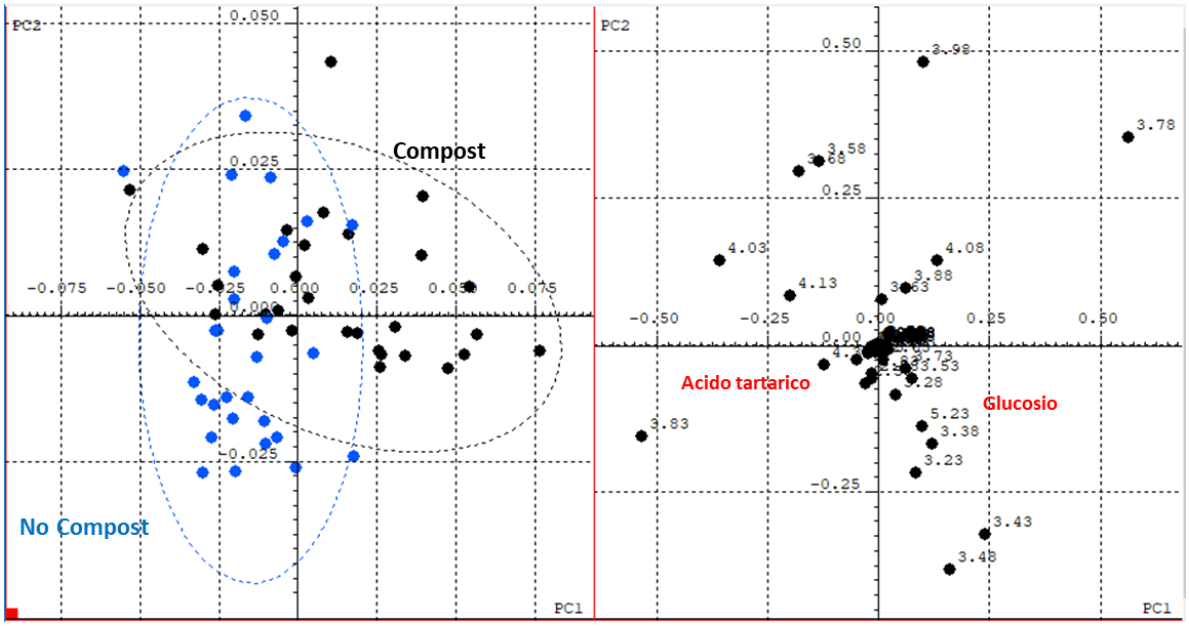



**Figure 25. Score plot and loading plot resulting from the PCA on the NMR data of the grape juice samples, relating to the first two PCs.**

The score plot generated by the PCA and shown in Figure 25(left) shows a certain grouping of the C samples deriving from the treatment with compost. From the loading plot, which reports the weights (loadings) that the variables have on the main components and therefore on the model, it can be seen that tartaric acid is more abundant in the NC grape samples, while the C samples have a higher glucose content.

### 1.1.1 Processing of Tomato NMR Data

PCA of the NMR data of the tomato samples was performed. The PCA analysis did not show separation between the C and NC samples, as can be seen from the score plot (Figure 26).

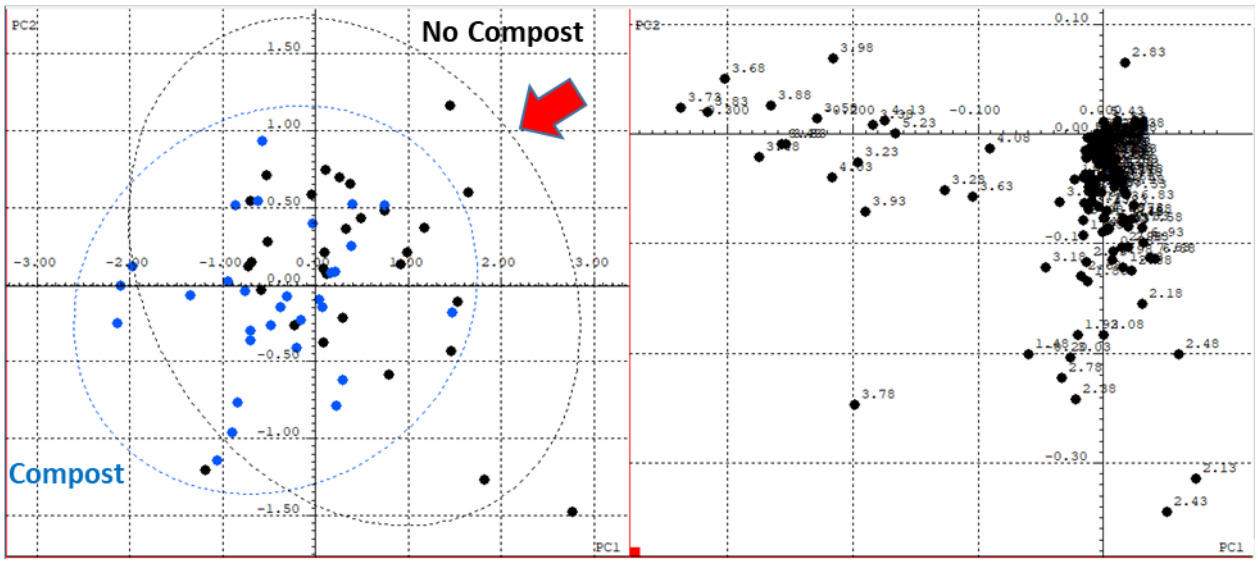



**Figure 26. Score plot and loading plot resulting from the PCA on the NMR data of the tomato samples relative to the first two PCs.**

In particular, the tomato samples obtained by treatment with compost showed a greater uniformity in the metabolic profile, as can be seen in Figure 27.

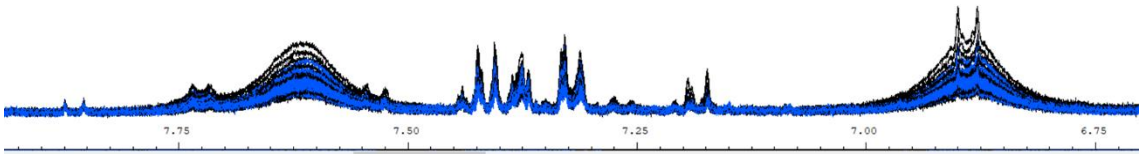


**Figure 27.Aromatic zone of the NMR spectra of C (blue) and NC (black) tomato samples**

Considering the poor results obtained in the PCA and PLS analyzes, it was decided to experiment with a new protocol for preparing tomato samples. Therefore, a protocol from the literature was optimized (Hohmann M. et al, J. Agric. Food Chem. 2015, 63 (43), pp 9666–9675) based on freeze-drying. After validation in terms of reproducibility and repeatability, the new protocol was used to reprepare the samples. The tomato samples were freeze-dried for 48 hours. For each sample, 150 mg of lyophilisate was dissolved in 1.5 mL of buffer containing sodium azide. The solution was vortexed for 1 min and then centrifuged at 6000 rpm (4700 g) for 10 min. The supernatant was transferred into a 2 mL amber glass vial, and placed in the autosampler. The tubes were prepared by taking 630 µL of solution and 70 µL of D2O solution containing TSP.

The procedures for the acquisition and processing of NMR spectra have been left unchanged.

The spectrum obtained using this protocol turned out to be very good, particularly rich in well-resolved signals and without criticality. Statistical analyzes were conducted again. Therefore, the NMR spectra obtained, aligned and normalized with respect to the TSP signal, were analogously digitized in the form of bucket-tables.

Already the PCA, an unsupervised MVA analysis approach, has highlighted a certain distinction between samples obtained from soil that has been amended (C) and not amended with compost (NC).

The application of the PLS, a super-examined method, has improved the separation between the two classes: in the score plot relating to the first two main components t1 vs t2 reported in Figure 28 a, it is observed that the C and NC samples tend to separate along the diagonal. From the examination of the loadings it was possible to identify the discriminating variables (metabolites): the NC tomato samples were found to contain greater amounts of glutamine and citric acid, while the C samples were found to contain more fructose and glucose (Figure 28 b-d). These are primary metabolites, which certainly influence the organoleptic characteristics of the product.

The PLS-DA model obtained is quite robust, having a percentage of correct answers in classification of 90% (Table 5) and a fair value of "goodness of prediction in cross-validation" Q2 = ~ 0.5.

In addition, using the new sample preparation protocol, lower variability in the content of different metabolites was again found among the C samples compared to the NC samples (Figure 29).

In conclusion, by varying the NMR analysis protocol, it was also possible for the tomato:

- highlight a metabolic pattern that can be traced back to agronomic practice;
- build a robust statistical model for the classification of samples based on agronomic practice.

Then, the data obtained with this second protocol were used in the creation of the classification software, described in Action 5.

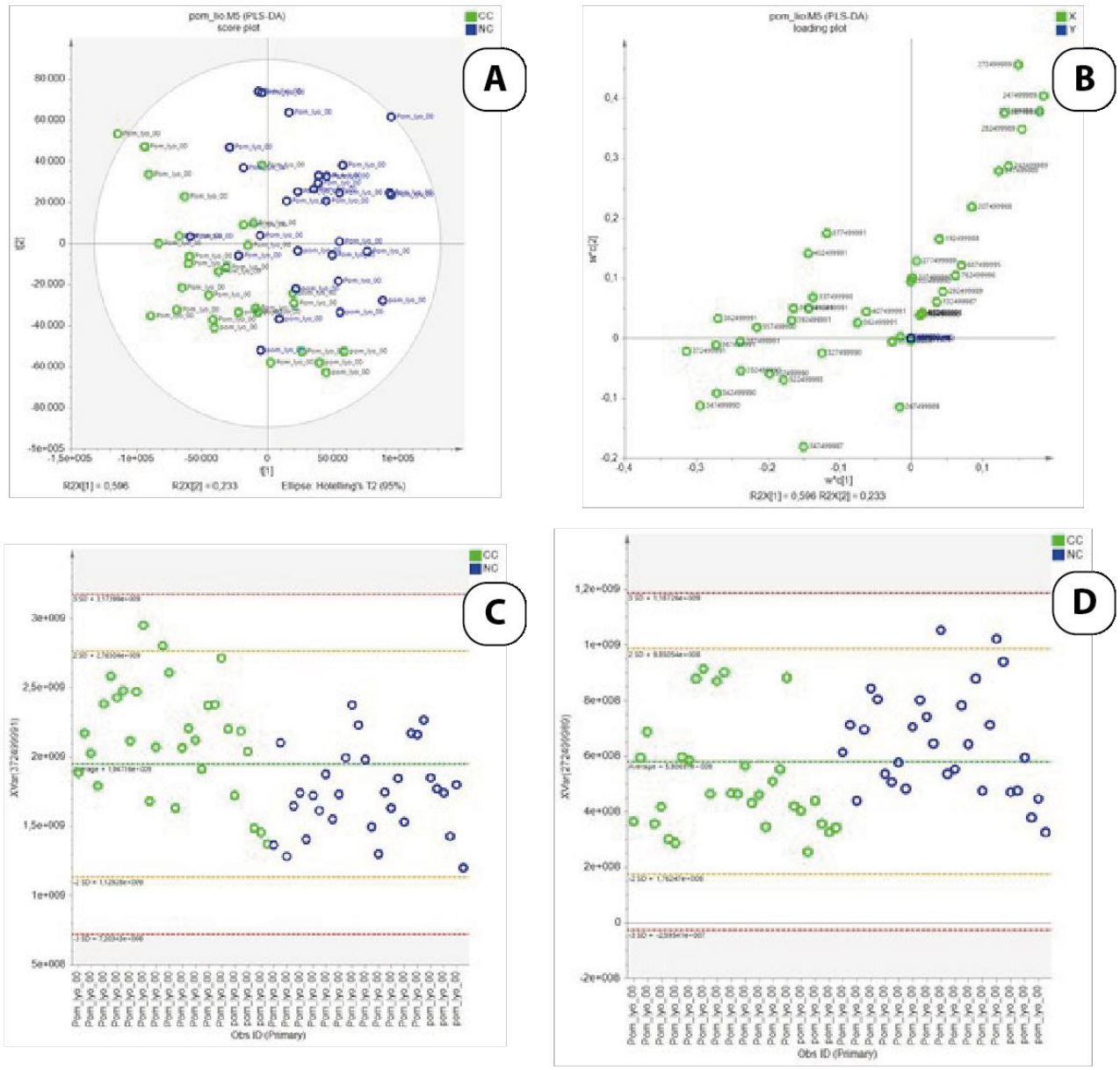



**Figure 28. PLS-DA analysis: above, score plot (a) and loading plot (b) relating to the first two main components t1 vs t2; below, graphs of the values of the two buckets with the greatest influence on the model, XVar (3.72) in c. and XVar (2.72) in d., correspond**

**Table 5. Confusion matrix with the classification results of the 60 tomato samples, obtained with compost C and without compost NC: in green the number of samples classified correctly, in yellow the number of samples not correctly classified.**

| | Samples | % correct classification | Compost Prediction | Non Compost Prediction |
|---|---|---|---|---|
| **Compost Observed** | 30 | 90% | 27 | 3 |
| **Non compost Observed** | 30 | 90% | 3 | 27 |
| total | 60 | 90% | 30 | 30 |
| Fishers prob. | 1.40E-10 | | | |

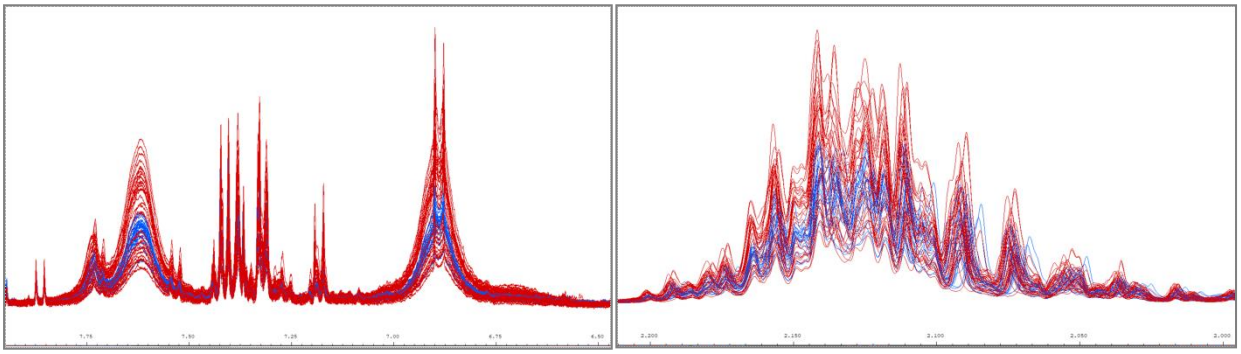


**Figure 29. Overlapping NMR spectra for NC samples in red and C in blue. On the left, the aromatic region; right, the glutamine region.**

### Processing of Wheat NMR Data

The PCA analysis on the NMR data of the wheat samples showed a clear separation of the C and NC samples, mainly due to the different choline content, higher in the C wheat samples (Figure 30).

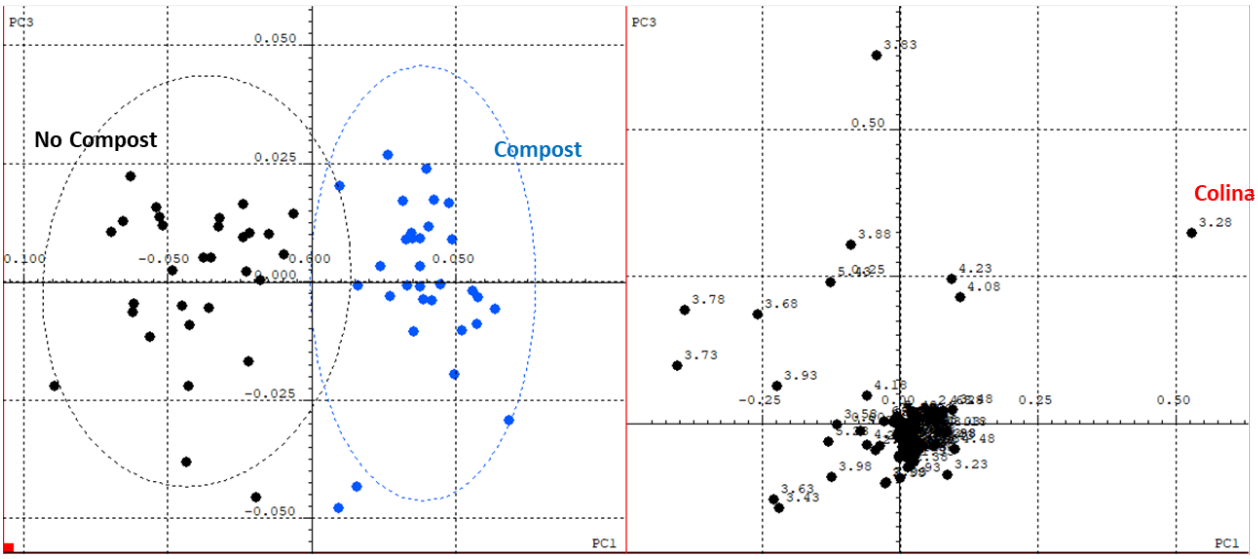
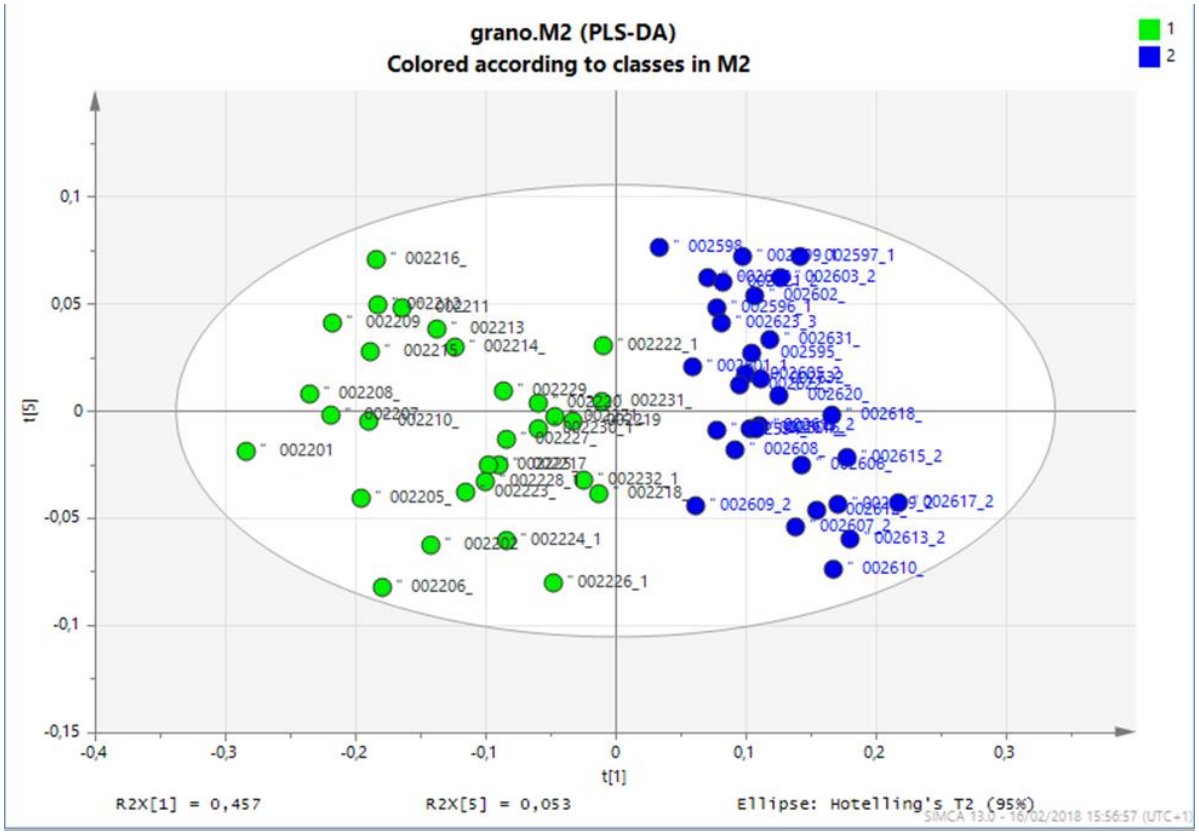



**Figure 30. Score plot and loading plot resulting from PCA on NMR data of wheat samples relative to PC1 vs PC2.**

The application of the PLS allowed to build a fairly robust classification model, with a percentage of correct answers in classification of 100% and an excellent value of "goodness of prediction in cross-validation" Q2 = ~ 0.97, as shown in Figure 31.




| | Members | Correct | 1 | 2 | No class (YPred < 0) |
|---|---|---|---|---|---|
| 1 | 30 | 100% | 30 | 0 | 0 |
| 2 | 31 | 100% | 0 | 31 | 0 |
| No class | 0 | | 0 | 0 | 0 |
| Total | 61 | 100% | 30 | 31 | 0 |
| Fishers prob. | 4,3e-018 | | | | |

| $R^2Y$(cum) | $Q^2$(cum) |
|---|---|
| 0,989 | 0,970 |

Cross Validation 10%

**Figure 31. Above, the score plot relating to the first and fifth main component t1 vs t5, obtained by PLS analysis of the NMR data of the wheat samples; below, the confusion matrix with the results of the prediction in cross-validation.**

The good results of the statistical processing carried out on NMR data have laid stable foundations for the development of a software for the classification and quality control of products obtained through different agronomic techniques (with compost C and without compost NC).

## Machine learning result

An application has been created that can operate both in “stand alone” and on the network. The software was created in such a way as to allow a non-specialized operator to:

- to classify the products in question;
- manage the updating of analytical data automatically;
- evaluate the quality parameters of the fruit and vegetable product;
- produce valid documentation for the voluntary certification of quality and agronomic practices and for the traceability of the product.

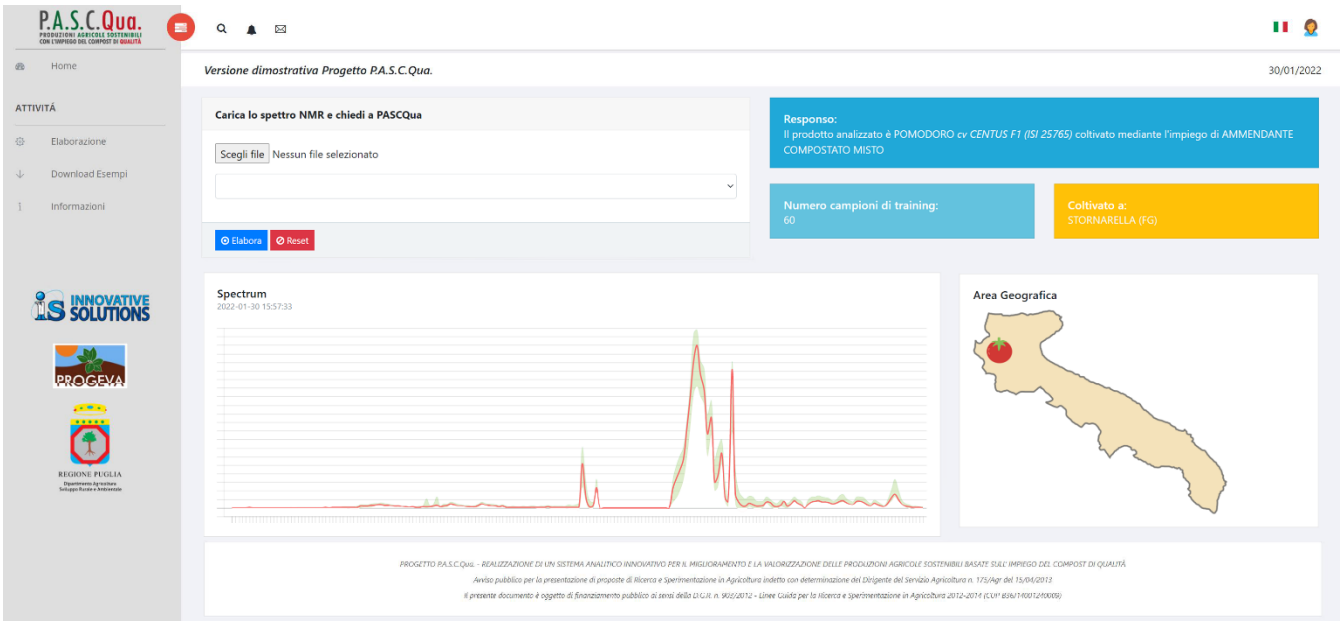



**Figure 32. P.A.S.C.Qua project web demo interface**

Two different algorithms were used for the creation of the classifier: neural network (Artificial Neural Network, ANN) based on Error Back Propagation (EBP) and Random Forest (RF) algorithm. The ANN algorithms were trained with 1000 training epochs, while the Random Forest algorithms were evaluated with the use of 500 trees. All results were verified with the n-fold validation technique, with n equal to 10.

The objective of the analysis was to find the kind of soil used to grow the products, the first batch of samples grew on a standard soil and the second batch was grown on a compost enriched soil.

**Table 6. Machine learning algorithms performance synthetic results**

| | ANN | RF |
|---|---|---|
| **Grapes** | 96,6% | 81,3% |
| **Tomato** | 86,6% | 83,3% |
| **Wheat** | 98,3% | 86,4% |

The model was able to identify the products based on the environment the samples grown within.

In general, the worst classification performance was obtained on tomato NMR data (in agreement with the statistical analyzes previously described), and the best classification results were produced by applying the ANN algorithm with a percentage of correct answers of about 98% for wheat and 96% for grapes. On the table n.1 we can see the results of the tests.

## A data model for artificial intelligence

The Bruker FID is a raw data format. It consists of complex point pairs, real and imaginary. Each data point is a 32bit integer which allows for complex averaging of the 16bit or higher acquired data point, 8bytes for each complex point. The number of complex points are listed in the acqp file, an ASCII text file, describing the scan acquisition parameters. The number of data points is listed in the ACQ_size array. Using this information, is possible to check the correct number of points and size by calculating the ACQ size points * 4 bytes * 2 = size of the fid file.

The information produced by the NMR spectrometer are not suited to a direct training of an algorithm, the data file as said before is composed by over 32.000 elements due to the nature of the experiment. The data must be processed in the frequency domain and then it can be used and stored for our purposes. Storing a 256 (at 0.04 ppm) or 1024 (at 0.01 ppm) float integer array is not a hard task by any means, but the true problem is to create a versatile environment to store the metadata accompanying this information.

For each sample is needed to record all available information, not only related to the nature of the sample but also regarding the nature of the experiment. A consistent method of acquisition is important to assure the consistency of the prediction.
We have also to deal with the fact that each kind of sample has his own peculiar information, for each sample we have to deal with the information supplied at the time of the sample collection, sometime this information is not available and therefore we have to deal with incomplete information on specified sample.
To deal with all these problem in the storage and the management of the NMR data, a dedicated information system has been developed. Figure 33

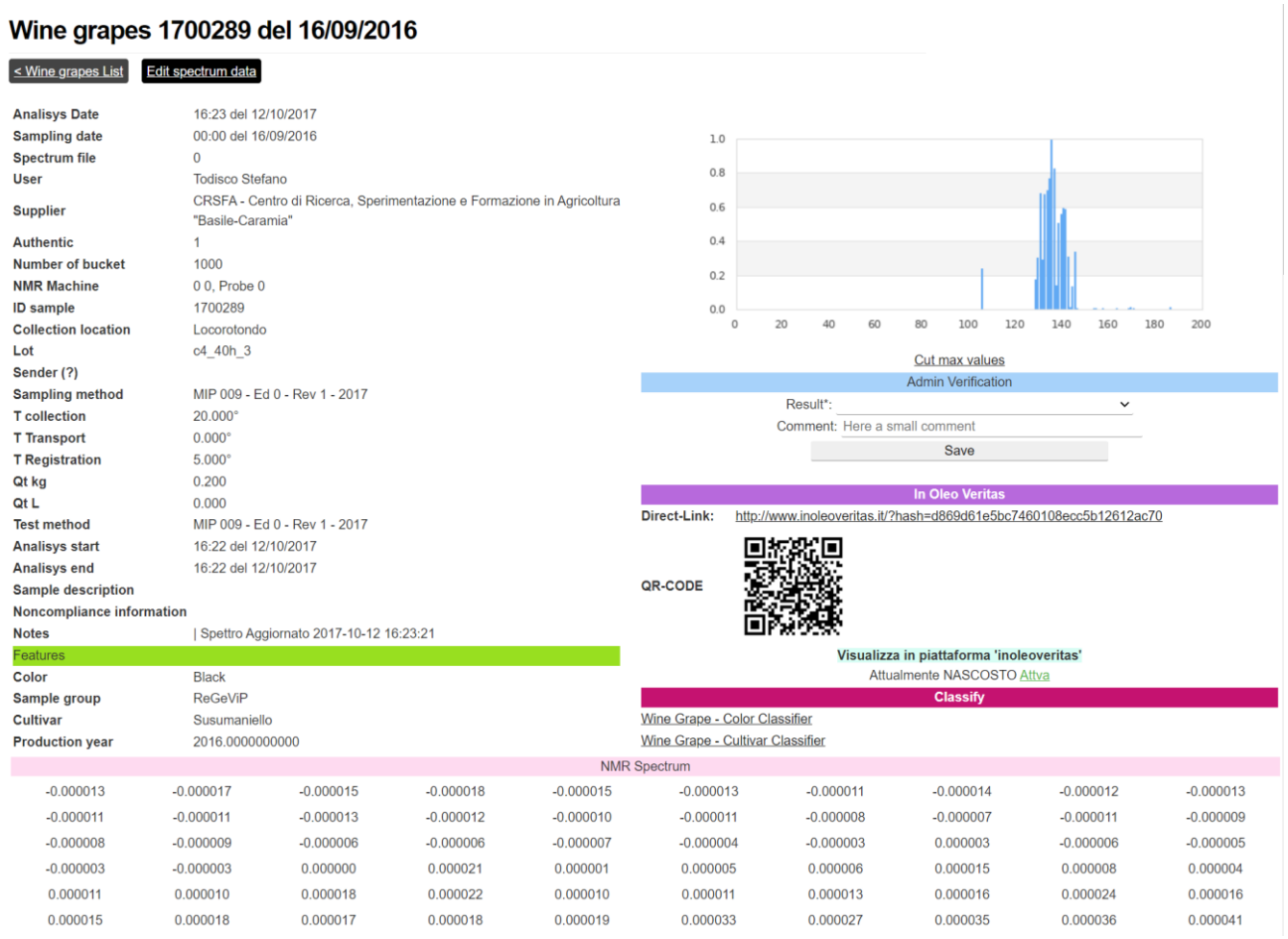



**Figure 33. IS-Tracer sample detail UI**

This custom application has been developed to create an hub for the collection and management of the information gathered in all the experiments we used to work on the course of this work.

The platform is structured to record standard information about the sample, the machine used and the sample preparation and analysis procedure.

The fixed information that can be recorded are the following

- Analisys Date - date when the analysis occurred
- Sampling date - date of collection of the sample
- Spectrum file - Fid file obtained from the experiment
- User – operator with performed the experiment
- Supplier – sample supplier

- Authentic – True/False, this value indicates if the metadata associated with the sample are supplied by an accredited body
- Number of bucket – indicated the obtained buckets after processing
- NMR Machine – information of the machine and the probe used in the experiment
- ID sample – unique sample id
- Collection location – geographical location of the sample
- Lot – sample production lot
- Sender – here can be specified any middleman that handled the sample collection
- Sampling method – here is recorded the protocol used for the sampling procedure
- T collection – temperature of the sample at the collection
- T Transport – temperature of the sample during transport
- T Registration – temperature of the sample when it reached the laboratory
- Qt kg – weight of the sample
- Qt L – volume of the sample for liquids
- Test method – here is recorded the protocol used for the NMR Experiment
- Analysis start – start time of the analysis
- Analysis end – end time of the analysis
- Sample description
- Noncompliance – if the sample is deemed not authentic or not suitable for the training of the algorithm here can be added a brief explanation
- Notes - further notes on the sample

Other than these standard information in the database is possible to create an additional set of parameters tailored on the sample nature and information.

For example, for the IntelliTrace related investigation the extra metadata used in the database was the cultivar obtained.

This platform also permits the extraction of data in ARFF format useful to use with weka suite.

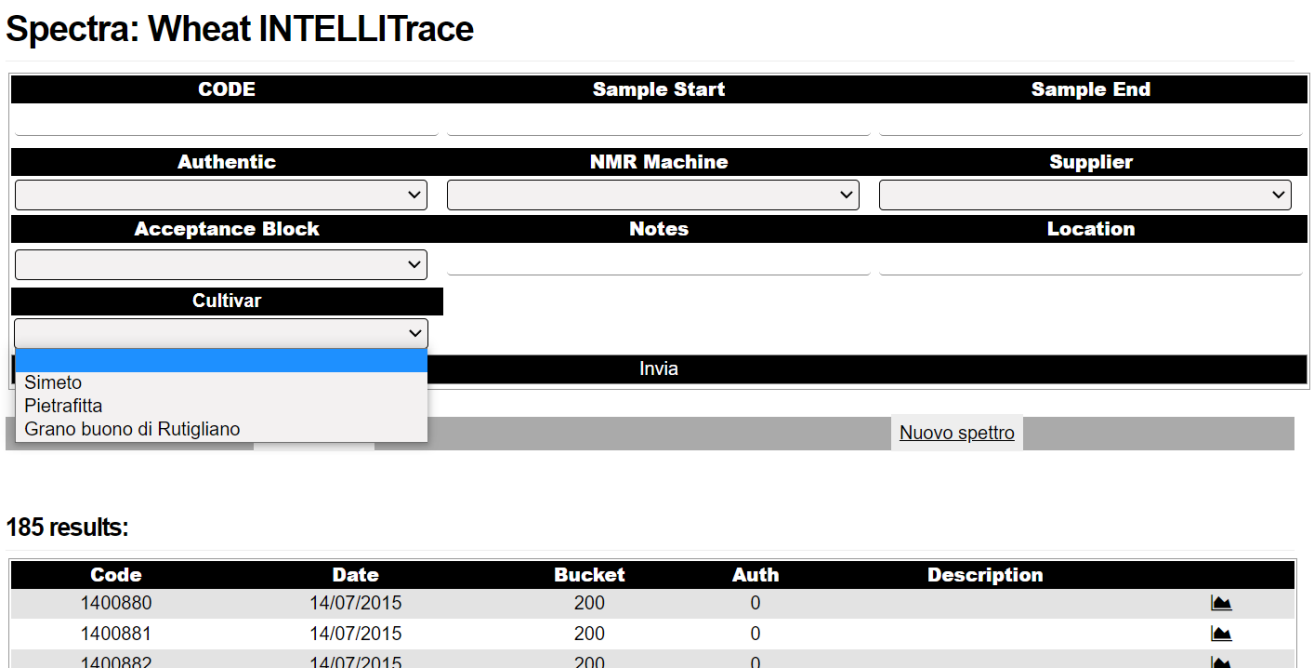


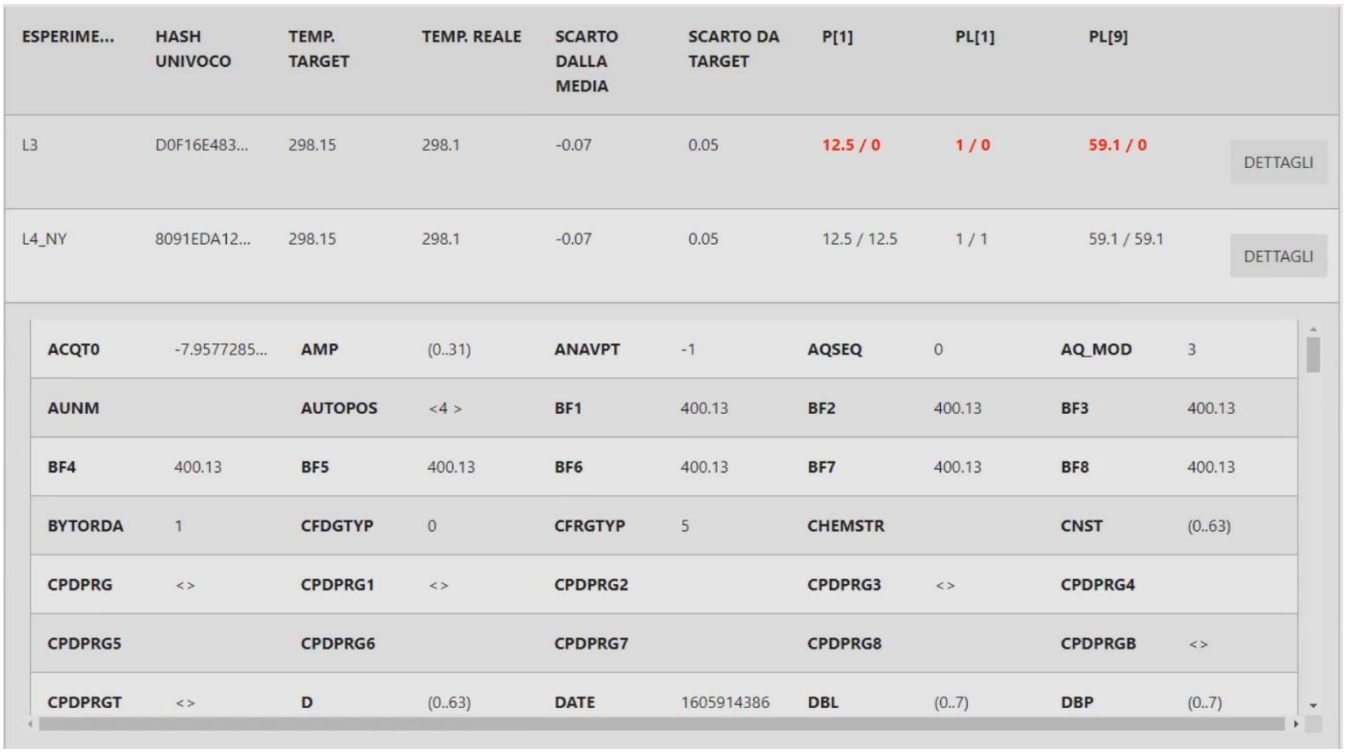

**Figure 34. IS-Tracer data export interface**

The obtained training file can be exported filtering for each parameter available.

Actually the database contain more than 4400 NMR spectrum related to a lot of different food matrix Wine grapes, Vine leaves, Flour, Peaches, Pomegranate, Table grapes, Olive oil, Artichoke, Pepper, Cardoon, Wheat and others.

## *How to store and organize NMR data*

A simple system has been built to navigate these parameters as needed. It uses the file as source for a simple sql-lite database for basic research.



| ESPERIME... | HASH UNIVOCO | TEMP. TARGET | TEMP. REALE | SCARTO DALLA MEDIA | SCARTO DA TARGET | P[1] | PL[1] | PL[9] | |
|---|---|---|---|---|---|---|---|---|---|
| L3 | D0F16E483... | 298.15 | 298.1 | -0.07 | 0.05 | 12.5 / 0 | 1 / 0 | 59.1 / 0 | DETTAGLI |
| L4_NY | 8091EDA12... | 298.15 | 298.1 | -0.07 | 0.05 | 12.5 / 12.5 | 1 / 1 | 59.1 / 59.1 | DETTAGLI |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ACQT0 | -7.9577285... | AMP | (0..31) | ANAVPT | -1 | AQSEQ | 0 | AQ_MOD | 3 |
| AUNM | | AUTOPOS | <4 > | BF1 | 400.13 | BF2 | 400.13 | BF3 | 400.13 |
| BF4 | 400.13 | BF5 | 400.13 | BF6 | 400.13 | BF7 | 400.13 | BF8 | 400.13 |
| BYTORDA | 1 | CFDGTYP | 0 | CFRGTYP | 5 | CHEMSTR | | CNST | (0..63) |
| CPDPRG | <> | CPDPRG1 | <> | CPDPRG2 | | CPDPRG3 | <> | CPDPRG4 | |
| CPDPRG5 | | CPDPRG6 | | CPDPRG7 | | CPDPRG8 | | CPDPRGB | <> |
| CPDPRGT | <> | D | (0..63) | DATE | 1605914386 | DBL | (0..7) | DBP | (0..7) |

**Figure 35. Gui interface for acqs file parsing**

This system has proven to be useful and faster than the traditional methods but is in his early stages of development and needs more polishing.

# Developments

## *A tentative guide to dataset generation for agri-food machine learning*

Once the scope of the database has been designed, consideration needs to be given to the analytical method, which will enable discrimination of authentic from non-authentic samples. The most appropriate approach is to consider the scope of the database, and to consult experts in the field of the foodstuff in question to list the physical and chemical distinctions that differentiate the products. Factors to consider include, geographical origin, temperature, age of material, ingredients, production methods and consistency. There is a significant risk that models, created without careful consideration of the analytical differences, will fail. Databases, that seek to differentiate samples from distinct geographical

origins, often rely on methods that use some form of stable isotope analysis, as these have shown to be influenced by factors such as altitude and temperature. Databases that seek to differentiate samples by their chemical differences often rely on either targeted or non-targeted methodologies. Non-targeted methods have the advantage of being able at least in principle to also identify new chemical markers, which were originally neither considered nor known. Considerations need to be given to the analytical method and its long-term ability to produce reproducible data. Stable isotope analysis is a significant field and a range of calibration standards and proficiency test materials are available. Spectroscopic technologies (e.g. NMR, FTIR, Raman) are considered to be reproducible technologies and are not influenced by drift or changes in sensitivity over time and also have calibration standards available. Hyphenated techniques, such as GC-MS and LC-MS require more consideration. Chromatography columns deteriorate through usage, which can lead to changes in retention times of analytes. MS detectors are also susceptible to fouling, which can impact on signal response. These factors need to be accounted for in analyses and require the use of randomised sample analysis and in-house reference material, which is used for intra and inter batch corrections [172]. More detailed information about methods of analysis is provided in the accompanying scientific opinions produced from the FoodIntegrity project [173].

Once a suitable analytical method has been identified, a small-scale study to confirm the validity of any assumptions made is recommended. Samples are rarely analysed directly; some sort of extraction is usually required. Therefore, considerations must be given to the reproducibility of the extraction method and the reproducibility of the instrumental method, and these should be related to the observed discrimination, which is the scope of the database. If instrument and sample extraction variability can be shown to be minimal, compared to the observed discrimination between groups, replicate extraction of samples and replicate analysis of extracts can be avoided. More detailed information about sampling can be found in the literature [174].

The act of performing a small study can also highlight any factors that have not been considered (e.g. difficulties in obtaining reference materials).

Identification of the point in the supply chain, where samples should be taken will ensure that the collected samples are fit for the analytical technique being used and the database is representative for the target product. The position in the supply chain, where samples are collected can influence both (i)

the quality of the analytical data and (ii) the integrity of the database. For stable isotope ratio analysis, processing or cooking of a raw material and the addition of other ingredients can affect the isotopic composition to an extent that it is no longer comparable to a database of 'raw' or 'unprocessed' material. It is necessary for non-targeted applications to understand the production process of the material of interest. This ensures that the database is representative of expected analytes, which are introduced during production. Dependent on the application, it may be necessary to perform validation to determine the effect of processing on the analyte of interest.

## Conclusions

Working on this matter has shown the need of an interdisciplinary collaboration between different kinds of professional and scholars.

This study in the years of making has spawned different collaborations, the most recent is a collaboration between our department and the Italian department ICQRF (Ispettorato centrale repressione frodi - Mipaaf) that supported the creation of authentic lentil and wheat database supplying authentic samples to be examined and stored.

The different machine learning approaches used in this thesis have proven the efficiency and reliability of their approach on data analysis, chemometric practices can heavily rely on this new approach to deliver more accurate and robust predictions.

The great amount of information generated by the chemical analysis must be well stored and managed. Non-targeted experiments are a crate of information that need to be handled with their story in mind, each experiment is influenced by environmental, electronic and human interference and a great amount of expertise is needed to create a working and secure model.

The amount of stored data can also be restudied in the future using new understanding or technology that today is unknown or not yet suited for this kind of elaboration, we need to be as ready as possible to handle the data also for the future researchers and not only for our actual needs.

The explored machine learning techniques are a starting set of tools needed to process the data provided by the NMR spectroscopy. A large amount of data needs to be constantly elaborated to discover new relations between the results and the field condition. For example, testing the rule-based

algorithms like the J48 allows extracting the regions of interest from the metabolomic spectra that can be studied by analytical chemists and can provide new insight into the relation of the metabolites with some product characteristic traits. ANN algorithms have proven to be highly effective but have a high computational cost, those algorithm can be used in the production pipeline where the update cycle of the model can be slower but accurate predictions are needed. In the end, also the random forest algorithms can be used to experiment with a faster algorithm some new correlation, the accuracy of the prediction and the training speed is a great value in the preliminary work of building a working model, then the results can be used to train a more robust Neural Network algorithm.

All these algorithms have proven their value in various projects, but the significant investment required to generate a food authenticity database and the sensitivity of the data that they contain, can reduce the willingness of organizations to share their data. Where proprietary food authenticity databases are offered as a commercial service to the food industry, it is recommended that the non-sensitive aspects discussed in this opinion piece are made available to provide confidence that the database is appropriate for its specified use. This study and the results obtained can prove that machine learning can improve prediction and elaboration on the NMR spectra enable researchers to further investigate the nature of food products and develop new and, hopefully, groundbreaking applications by the means of this approach.

# Bibliography


[1] V. Gallo *et al.*, "Performance Assessment in Fingerprinting and Multi Component Quantitative NMR Analyses," *Analytical Chemistry*, vol. 87, no. 13, 2015, doi: 10.1021/acs.analchem.5b00919.

[2] B. Musio *et al.*, "A community-built calibration system: The case study of quantification of metabolites in grape juice by qNMR spectroscopy," *Talanta*, vol. 214, 2020, doi: 10.1016/j.talanta.2020.120855.

[3] R. G. Cong, K. Hedlund, H. Andersson, and M. Brady, "Managing soil natural capital: An effective strategy for mitigating future agricultural risks?," *Agricultural Systems*, vol. 129, pp. 30–39, Jul. 2014, doi: 10.1016/J.AGSY.2014.05.003.

[4] V. Gallo *et al.*, "A Contribution to the Harmonization of Non-targeted NMR Methods for Data-Driven Food Authenticity Assessment," *Food Analytical Methods*, vol. 13, no. 2, 2020, doi: 10.1007/s12161-019-01664-8.

[5] A. A. Aksenov *et al.*, "Volatile organic compounds (VOCs) for noninvasive plant diagnostics," *ACS Symposium Series*, vol. 1141, pp. 73–95, 2013, doi: 10.1021/BK-2013-1141.CH006.

[6] R. Ragone *et al.*, "Development of a food class-discrimination system by non-targeted NMR analyses using different magnetic field strengths," *Food Chemistry*, vol. 332, p. 127339, Dec. 2020, doi: 10.1016/J.FOODCHEM.2020.127339.

[7] R. G. Brereton, "Pattern recognition in chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 149, pp. 90–96, Dec. 2015, doi: 10.1016/j.chemolab.2015.06.012.

[8] V. Bevilacqua, M. Triggiani, V. Gallo, I. Cafagna, P. Mastrorilli, and G. Ferrara, *An expert system for an innovative discrimination tool of commercial table grapes*, vol. 7390 LNAI. 2012. doi: 10.1007/978-3-642-31576-3_13.

[9] P. This, T. Lacombe, and M. R. Thomas, "Historical origins and genetic diversity of wine grapes," *Trends in Genetics*, vol. 22, no. 9, pp. 511–519, Sep. 2006, doi: 10.1016/J.TIG.2006.07.008.

[10] F. Savorani, G. Tomasi, and S. B. Engelsen, "icoshift: A versatile tool for the rapid alignment of 1D NMR spectra," *Journal of Magnetic Resonance*, vol. 202, no. 2, pp. 190–202, Feb. 2010, doi: 10.1016/J.JMR.2009.11.012.

[11] F. R. Pinu, "Grape and Wine Metabolomics to Develop New Insights Using Untargeted and Targeted Approaches," *Fermentation 2018, Vol. 4, Page 92*, vol. 4, no. 4, p. 92, Nov. 2018, doi: 10.3390/FERMENTATION4040092.

[12] D. Rodrigues, C. H. Santos, T. A. P. Rocha-Santos, A. M. Gomes, B. J. Goodfellow, and A. C. Freitas, "Metabolic profiling of potential probiotic or synbiotic cheeses by nuclear magnetic resonance (NMR) Spectroscopy," *Journal of Agricultural and Food Chemistry*, vol. 59, no. 9, pp. 4955–4961, May 2011, doi: 10.1021/JF104605R.

[13] D. A. Kidwell and D. L. Blank, "Comments on the paper by W.A. Baumgartner and V.A. Hill: sample preparation techniques," *Forensic Science International*, vol. 63, no. 1–3, pp. 137–143, 1993, doi: 10.1016/0379-0738(93)90267-E.

[14] W. A. Baumgartner and V. A. Hill, "Sample preparation techniques," *Forensic Science International*, vol. 63, no. 1–3, pp. 121–135, Dec. 1993, doi: 10.1016/0379-0738(93)90266-D.

[15] "Sample preparation techniques - ScienceDirect." https://www.sciencedirect.com/science/article/abs/pii/037907389390266D (accessed Jan. 30, 2022).

[16] J. D. McCord, E. Trousdale, and D. D. Y. Ryu, "An Improved Sample Preparation Procedure for the Analysis of Major Organic Components in Grape Must and Wine by High Performance Liquid Chromatography," *American Journal of Enology and Viticulture*, vol. 35, no. 1, 1984.

[17] V. Gallo *et al.*, "A Contribution to the Harmonization of Non-targeted NMR Methods for Data-Driven Food Authenticity Assessment," *Food Analytical Methods*, vol. 13, no. 2, pp. 530–541, Feb. 2020, doi: 10.1007/S12161-019-01664-8.

[18] S. Esslinger, J. Riedl, and C. Fauhl-Hassek, "Potential and limitations of non-targeted fingerprinting for authentication of food in official control," *Food Research International*, vol. 60, pp. 189–204, 2014, doi: 10.1016/j.foodres.2013.10.015.

[19] T. F. 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," *Trends in Food Science & Technology*, vol. 76, pp. 38–55, Jun. 2018, doi: 10.1016/J.TIFS.2018.04.001.

[20] J. Riedl, S. Esslinger, and C. Fauhl-Hassek, "Review of validation and reporting of non-targeted fingerprinting approaches for food authentication," *Analytica Chimica Acta*, vol. 885, pp. 17–32, 2015, doi: 10.1016/J.ACA.2015.06.003.

[21] F. Westad and F. Marini, "Validation of chemometric models – A tutorial," *Analytica Chimica Acta*, vol. 893, pp. 14–24, Sep. 2015, doi: 10.1016/J.ACA.2015.06.056.

[22] M. C. Jewett, G. Hofmann, and J. Nielsen, "Fungal metabolite analysis in genomics and phenomics," *Current Opinion in Biotechnology*, vol. 17, no. 2, pp. 191–197, Apr. 2006, doi: 10.1016/J.COPBIO.2006.02.001.

[23] C. Gieger *et al.*, "Genetics Meets Metabolomics: A Genome-Wide Association Study of Metabolite Profiles in Human Serum," *PLOS Genetics*, vol. 4, no. 11, p. e1000282, Nov. 2008, doi: 10.1371/JOURNAL.PGEN.1000282.

[24] C. Birkemeyer, A. Luedemann, C. Wagner, A. Erban, and J. Kopka, "Metabolome analysis: The potential of in vivo labeling with stable isotopes for metabolite profiling," *Trends in Biotechnology*, vol. 23, no. 1, pp. 28–33, 2005, doi: 10.1016/j.tibtech.2004.12.001.

[25] W. B. Dunn and D. I. Ellis, "Metabolomics: Current analytical platforms and methodologies," *TrAC Trends in Analytical Chemistry*, vol. 24, no. 4, pp. 285–294, Apr. 2005, doi: 10.1016/J.TRAC.2004.11.021.

[26] E. M. Lenz, J. Bright, I. D. Wilson, S. R. Morgan, and A. F. P. Nash, "A 1H NMR-based metabonomic study of urine and plasma samples obtained from healthy human subjects," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 33, no. 5, pp. 1103–1115, Dec. 2003, doi: 10.1016/S0731-7085(03)00410-2.

[27] K. S. Solanky *et al.*, "Application of biofluid 1H nuclear magnetic resonance-based metabonomic techniques for the analysis of the biochemical effects of dietary isoflavones on human plasma profile," *Analytical Biochemistry*, vol. 323, no. 2, pp. 197–204, Dec. 2003, doi: 10.1016/j.ab.2003.08.028.

[28] F. Ma and L. Cheng, "The sun-exposed peel of apple fruit has higher xanthophyll cycle-dependent thermal dissipation and antioxidants of the ascorbate-glutathione pathway than the

shaded peel," *Plant Science*, vol. 165, no. 4, pp. 819–827, Oct. 2003, doi: 10.1016/S0168-9452(03)00277-2.

[29] O. Al-Jowder, E. K. Kemsley, and R. H. Wilson, "Mid-infrared spectroscopy and authenticity problems in selected meats: A feasibility study," *Food Chemistry*, vol. 59, no. 2, pp. 195–201, Jun. 1997, doi: 10.1016/S0308-8146(96)00289-0.

[30] C. Wittmann, J. O. Krömer, P. Kiefer, T. Binz, and E. Heinzle, "Impact of the cold shock phenomenon on quantification of intracellular metabolites in bacteria," *Analytical Biochemistry*, vol. 327, no. 1, pp. 135–139, Apr. 2004, doi: 10.1016/j.ab.2004.01.002.

[31] W. Weckwerth, K. Wenzel, and O. Fiehn, "Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks," *Proteomics*, vol. 4, no. 1, pp. 78–83, Jan. 2004, doi: 10.1002/PMIC.200200500.

[32] K. Weinberger and A. Graber, "Using Comprehensive Metabolomics to Identify Novel Biomarkers.," *Screening Trends in Drug Discovery*, vol. 6, 2005.

[33] S. K. Bharti and R. Roy, "Quantitative 1H NMR spectroscopy," *TrAC - Trends in Analytical Chemistry*, vol. 35, pp. 5–26, May 2012, doi: 10.1016/j.trac.2012.02.007.

[34] E. Kupče and T. D. W. Claridge, "Molecular structure from a single NMR supersequence," *Chemical Communications*, vol. 54, no. 52, pp. 7139–7142, Jun. 2018, doi: 10.1039/C8CC03296C.

[35] S. Zhang, C. Zheng, I. R. Lanza, K. S. Nair, D. Raftery, and O. Vitek, "Interdependence of signal processing and analysis of urine1H NMR spectra for metabolic profiling," *Analytical Chemistry*, vol. 81, no. 15, pp. 6080–6088, Aug. 2009, doi: 10.1021/AC900424C.

[36] L. Castañar, G. D. Poggetto, A. A. Colbourne, G. A. Morris, and M. Nilsson, "The GNAT: A new tool for processing NMR data," *Magnetic Resonance in Chemistry*, vol. 56, no. 6, pp. 546–558, Jun. 2018, doi: 10.1002/MRC.4717.

[37] Y. B. Monakhova, H. Schäfer, E. Humpfer, M. Spraul, T. Kuballa, and D. W. Lachenmeier, "Application of automated eightfold suppression of water and ethanol signals in 1H NMR to provide sensitivity for analyzing alcoholic beverages," *Magnetic Resonance in Chemistry*, vol. 49, no. 11, pp. 734–739, Nov. 2011, doi: 10.1002/MRC.2823.

[38] Y. B. Monakhova, H. Schäfer, E. Humpfer, M. Spraul, T. Kuballa, and D. W. Lachenmeier, "Application of automated eightfold suppression of water and ethanol signals in 1H NMR to provide sensitivity for analyzing alcoholic beverages," *Magnetic Resonance in Chemistry*, vol. 49, no. 11, pp. 734–739, Nov. 2011, doi: 10.1002/MRC.2823.

[39] S. Monsonis Centelles *et al.*, "Toward Reliable Lipoprotein Particle Predictions from NMR Spectra of Human Blood: An Interlaboratory Ring Test," *Analytical Chemistry*, vol. 89, no. 15, pp. 8004–8012, Aug. 2017, doi: 10.1021/ACS.ANALCHEM.7B01329.

[40] J. L. Ward *et al.*, "An inter-laboratory comparison demonstrates that [1H]-NMR metabolite fingerprinting is a robust technique for collaborative plant metabolomic data collection," *Metabolomics*, vol. 6, no. 2, pp. 263–273, Jun. 2010, doi: 10.1007/S11306-010-0200-4.

[41] P. Giraudeau, I. Tea, G. S. Remaud, and S. Akoka, "Reference and normalization methods: Essential tools for the intercomparison of NMR spectra," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 93, pp. 3–16, 2014, doi: 10.1016/j.jpba.2013.07.020.

[42] J. L. Jungnickel and J. W. Forbes, "Quantitative Measurement of Hydrogen Types by Integrated Nuclear Magnetic Resonance Intensities," *Analytical Chemistry*, vol. 35, no. 8, pp. 938–942, Jul. 1963, doi: 10.1021/AC60201A005.

[43] S. Tewari, T. O'Reilly, and A. Webb, "Improving the field homogeneity of fixed- and variable-diameter discrete Halbach magnet arrays for MRI via optimization of the angular magnetization distribution," *Journal of Magnetic Resonance*, vol. 324, p. 106923, Mar. 2021, doi: 10.1016/J.JMR.2021.106923.

[44] M. Decorps, P. Blondet, H. Reutenauer, J. P. Albrand, and C. Remy, "An inductively coupled, series-tuned NMR probe," *Journal of Magnetic Resonance (1969)*, vol. 65, no. 1, pp. 100–109, Oct. 1985, doi: 10.1016/0022-2364(85)90378-6.

[45] J. A. Norcross *et al.*, "Multiplexed NMR: An automated CapNMR dual-sample probe," *Analytical Chemistry*, vol. 82, no. 17, pp. 7227–7236, Sep. 2010, doi: 10.1021/AC101003F.

[46] L. T. Wurtz and W. P. Wheless, "Design of a High-Performance, Low-Noise Charge Preamplifier," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 40, no. 8, pp. 541–545, 1993, doi: 10.1109/81.242329.

[47] V. Heine, "Electronic Structure from the Point of View of the Local Atomic Environment," *Solid State Physics - Advances in Research and Applications*, vol. 35, no. C, pp. 1–127, Jan. 1980, doi: 10.1016/S0081-1947(08)60503-2.

[48] "Electronic Structure from the Point of View of the Local Atomic Environment - ScienceDirect." https://www.sciencedirect.com/science/article/abs/pii/S0081194708605032 (accessed Jan. 30, 2022).

[49] "Sviluppo di metodologie per la valutazione della freschezza del pesce mediante applicazioni metabonomiche." http://amsdottorato.unibo.it/5498/ (accessed Jan. 30, 2022).

[50] A. Craig, O. Cloarec, E. Holmes, J. K. Nicholson, and J. C. Lindon, "Scaling and Normalization Effects in NMR Spectroscopic Metabonomic Data Sets," *Analytical Chemistry*, vol. 78, no. 7, pp. 2262–2267, Apr. 2006, doi: 10.1021/AC0519312.

[51] A. D. Maher, J. M. Fonville, M. Coen, J. C. Lindon, C. D. Rae, and J. K. Nicholson, "Statistical total correlation spectroscopy scaling for enhancement of metabolic information recovery in biological NMR spectra," *Analytical Chemistry*, vol. 84, no. 2, pp. 1083–1091, Jan. 2012, doi: 10.1021/AC202720F.

[52] F. Capozzi *et al.*, "Normalization is a Necessary Step in NMR Data Processing: Finding the Right Scale Factors "CHANCE KBBE FP7-(GA 266331)-Low cost technologies and traditional ingredients for the production of affordable, nutritionally correct foods Improving Health in Population groups at risk of poverty View project Food Kinetics by Spectroscopy View project NORMALIZATION IS A NECESSARY STEP IN NMR DATA PROCESSING: FINDING THE RIGHT SCALING FACTORS," 2014, doi: 10.1039/9781849732994-00147.

[53] S. A. A. Sousa, A. Magalhães, and M. M. C. Ferreira, "Optimized bucketing for NMR spectra: Three case studies," *Chemometrics and Intelligent Laboratory Systems*, vol. 122, pp. 93–102, Mar. 2013, doi: 10.1016/J.CHEMOLAB.2013.01.006.

[54] M. M. Koek, R. H. Jellema, J. van der Greef, A. C. Tas, and T. Hankemeier, "Quantitative metabolomics based on gas chromatography mass spectrometry: Status and perspectives," *Metabolomics*, vol. 7, no. 3, pp. 307–328, Sep. 2011, doi: 10.1007/S11306-010-0254-3.

[55] H. K. Choi, Y. H. Choi, M. Verberne, A. W. M. Lefeber, C. Erkelens, and R. Verpoorte, "Metabolic fingerprinting of wild type and transgenic tobacco plants by 1H NMR and multivariate analysis technique," *Phytochemistry*, vol. 65, no. 7, pp. 857–864, Apr. 2004, doi: 10.1016/j.phytochem.2004.01.019.

[56] E. M. Lenz, J. Bright, I. D. Wilson, S. R. Morgan, and A. F. P. Nash, "A 1H NMR-based metabonomic study of urine and plasma samples obtained from healthy human subjects," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 33, no. 5, pp. 1103–1115, Dec. 2003, doi: 10.1016/S0731-7085(03)00410-2.

[57] W. B. Dunn, I. D. Wilson, A. W. Nicholls, and D. Broadhurst, "The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans," *Bioanalysis*, vol. 4, no. 18, pp. 2249–2264, Sep. 2012, doi: 10.4155/BIO.12.204.

[58] R. A. Davis, A. J. Charlton, J. Godward, S. A. Jones, M. Harrison, and J. C. Wilson, "Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform," *Chemometrics and Intelligent Laboratory Systems*, vol. 85, no. 1, pp. 144–154, Jan. 2007, doi: 10.1016/J.CHEMOLAB.2006.08.014.

[59] P. E. Anderson, N. v. Reo, N. J. DelRaso, T. E. Doom, and M. L. Raymer, "Gaussian binning: A new kernel-based method for processing NMR spectroscopic data for metabolomics," *Metabolomics*, vol. 4, no. 3, pp. 261–272, 2008, doi: 10.1007/S11306-008-0117-3.

[60] P. E. Anderson, D. A. Mahle, T. E. Doom, N. v. Reo, N. J. DelRaso, and M. L. Raymer, "Dynamic adaptive binning: An improved quantification technique for NMR spectroscopic data," *Metabolomics*, vol. 7, no. 2, pp. 179–190, Jul. 2011, doi: 10.1007/S11306-010-0242-7/FIGURES/3.

[61] S. Wold, "Chemometrics; what do we mean with it, and what do we want from it?," *Chemometrics and Intelligent Laboratory Systems*, vol. 30, no. 1, pp. 109–115, Nov. 1995, doi: 10.1016/0169-7439(95)00042-9.

[62] M. Á. Rodríguez-Delgado, G. González-Hernández, J. E. Conde-González, and J. P. Pérez-Trujillo, "Principal component analysis of the polyphenol content in young red wines," *Food Chemistry*, vol. 78, no. 4, pp. 523–532, Sep. 2002, doi: 10.1016/S0308-8146(02)00206-6.

[63] B. M. Silva, P. B. Andrade, R. C. Martins, R. M. Seabra, and M. A. Ferreira, "Principal component analysis as tool of characterization of quince (Cydonia oblonga Miller) jam," *Food Chemistry*, vol. 94, no. 4, pp. 504–512, Mar. 2006, doi: 10.1016/J.FOODCHEM.2004.11.045.

[64] H. J. Keselman *et al.*, "Statistical Practices of Educational Researchers: An Analysis of their ANOVA, MANOVA, and ANCOVA Analyses:," *http://dx.doi.org/10.3102/00346543068003350*, vol. 68, no. 3, pp. 350–386, Jun. 2016, doi: 10.3102/00346543068003350.

[65] M. Ringnér, "What is principal component analysis?," *Nature Biotechnology 2008 26:3*, vol. 26, no. 3, pp. 303–304, Mar. 2008, doi: 10.1038/nbt0308-303.

[66] O. E. de Noord, "Multivariate calibration standardization," *Chemometrics and Intelligent Laboratory Systems*, vol. 25, no. 2, pp. 85–97, Nov. 1994, doi: 10.1016/0169-7439(94)85037-2.

[67] H. Martens and T. Naes, "Multivariate calibration." Wiley, 1989.

[68] "Application of PLS-DA in multivariate image analysis - Chevallier - 2006 - Journal of Chemometrics - Wiley Online Library." https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.994 (accessed Jan. 30, 2022).

[69] E. Szłyk, A. Szydłowska-Czerniak, and A. Kowalczyk-Marzec, "NIR spectroscopy and partial least-squares regression for determination of natural α-tocopherol in vegetable oils," *Journal of Agricultural and Food Chemistry*, vol. 53, no. 18, pp. 6980–6987, Sep. 2005, doi: 10.1021/JF050672E.

[70] R. D. Tobias, "An Introduction to Partial Least Squares Regression".

[71] A. J. Hopfinger, "Chemometric Methods in Molecular Design. Methods and Principles in Medicinal Chemistry, Volume 2 Edited by Han van de Waterbend (Hoffman-LaRoche Ltd., Basil, Switzerland). VCH: New York. 1995. xix + 359 pp. $110. ISBN 3-527-30044-9.," *Journal of the American Chemical Society*, vol. 118, no. 11, pp. 2774–2774, Jan. 1996, doi: 10.1021/JA9551998.

[72] R. G. Brereton and G. R. Lloyd, "Partial least squares discriminant analysis: taking the magic away," *Journal of Chemometrics*, vol. 28, no. 4, pp. 213–225, Apr. 2014, doi: 10.1002/CEM.2609.

[73] K. H. Liland and U. G. Indahl, "Powered partial least squares discriminant analysis," *Journal of Chemometrics*, vol. 23, no. 1, pp. 7–18, Jan. 2009, doi: 10.1002/CEM.1186.

[74] "Partial least squares discriminant analysis: taking the magic away - Brereton - 2014 - Journal of Chemometrics - Wiley Online Library." https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/full/10.1002/cem.2609 (accessed Jan. 30, 2022).

[75] M. Tenenhaus and V. E. Vinzi, "PLS regression, PLS path modeling and generalized Procrustean analysis: a combined approach for multiblock analysis," *Journal of Chemometrics*, vol. 19, no. 3, pp. 145–153, Mar. 2005, doi: 10.1002/CEM.917.

[76] J. L. Spratlin, N. J. Serkova, and S. G. Eckhardt, "Clinical Applications of Metabolomics in Oncology: A Review," *Clinical Cancer Research*, vol. 15, no. 2, 2009.

[77] A. M. Weljie, J. Newton, P. Mercier, E. Carlson, and C. M. Slupsky, "Targeted pofiling: Quantitative analysis of1H NMR metabolomics data," *Analytical Chemistry*, vol. 78, no. 13, pp. 4430–4442, Jul. 2006, doi: 10.1021/AC060209G.

[78] D. S. Wishart, "Metabolomics: applications to food science and nutrition research," *Trends in Food Science & Technology*, vol. 19, no. 9, pp. 482–493, Sep. 2008, doi: 10.1016/J.TIFS.2008.03.003.

[79] C. B. Y. Cordella, "PCA: The Basic Building Block of Chemometrics," *Analytical Chemistry*, Nov. 2012, doi: 10.5772/51429.

[80] R. Leardi, "Chemometrics: From classical to genetic algorithms," *Grasas y Aceites*, vol. 53, no. 1, pp. 115–127, Mar. 2002, doi: 10.3989/GYA.2002.V53.I1.294.

[81] K. J. Siebert, "Chemometrics in Brewing—A Review," *https://doi.org/10.1094/ASBCJ-59-0147*, vol. 59, no. 4, pp. 147–156, 2018, doi: 10.1094/ASBCJ-59-0147.

[82] I. Currie and A. Korabinski, "Some Comments on Bivariate Regression," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 33, no. 3, pp. 283–293, Sep. 1984, doi: 10.2307/2988232.

[83] R. W. Gerlach, B. R. Kowalski, and H. O. A. Wold, "Partial least-squares path modelling with latent variables," *Analytica Chimica Acta*, vol. 112, no. 4, pp. 417–421, Dec. 1979, doi: 10.1016/S0003-2670(01)85039-X.

[84] H. Martens, "Reliable and relevant modelling of real world data: A personal account of the development of PLS Regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 85–95, Oct. 2001, doi: 10.1016/S0169-7439(01)00153-8.

[85] M. Bylesjö, M. Rantalainen, O. Cloarec, J. K. Nicholson, E. Holmes, and J. Trygg, "OPLS discriminant analysis: Combining the strengths of PLS-DA and SIMCA classification," *Journal of Chemometrics*, vol. 20, no. 8–10, pp. 341–351, Aug. 2006, doi: 10.1002/CEM.1006.

[86] S. Wold, "Pattern recognition by means of disjoint principal components models," *Pattern Recognition*, vol. 8, no. 3, pp. 127–139, Jul. 1976, doi: 10.1016/0031-3203(76)90014-5.

[87] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1–3, pp. 37–52, 1987, doi: 10.1016/0169-7439(87)80084-9.

[88] C. Albano *et al.*, "Four levels of pattern recognition," *Analytica Chimica Acta*, vol. 103, no. 4, pp. 429–443, Dec. 1978, doi: 10.1016/S0003-2670(01)83107-X.

[89] E. H. J. L. J. G. A. V. M. S. J. N. BM Beckwith-Hall, "NMR-based metabonomic studies on the biochemical effects of commonly used drug carrier vehicles in the rat," *Chemical Research in Toxicology*, vol. 15, p. 1136, 2002, doi: 10.1021/tx020020+.

[90] J. Shockcor and E. Holmes, "Metabonomic applications in toxicity screening and disease diagnosis," *Current Topics in Medicinal Chemistry*, vol. 2, no. 1, p. 35, Mar. 2002, doi: 10.2174/1568026023394498.

[91] E. Holmes *et al.*, "Chemometric Models for Toxicity Classification Based on NMR Spectra of Biofluids," *Chemical Research in Toxicology*, vol. 13, no. 6, pp. 471–478, Jun. 2000, doi: 10.1021/TX990210T.

[92] S. Bicciato, A. Luchini, and C. di Bello, "Marker Identification and Classification of Cancer Types Using Gene Expression Data and SIMCA," *Methods of Information in Medicine*, vol. 43, no. 1, pp. 4–8, Feb. 2004, doi: 10.1055/S-0038-1633413/ID/BR1633413-4.

[93] T. Verron, R. Sabatier, and R. Joffre, "Some theoretical properties of the O-PLS method," *Journal of Chemometrics*, vol. 18, no. 2, pp. 62–68, Feb. 2004, doi: 10.1002/CEM.847.

[94] J. Trygg and S. Wold, "Orthogonal projections to latent structures (O-PLS)," *Journal of Chemometrics*, vol. 16, no. 3, pp. 119–128, Mar. 2002, doi: 10.1002/CEM.695.

[95] O. M. Kvalheim and T. v. Karstang, "Interpretation of latent-variable regression models," *Chemometrics and Intelligent Laboratory Systems*, vol. 7, no. 1–2, pp. 39–51, Dec. 1989, doi: 10.1016/0169-7439(89)80110-8.

[96] O. Cloarec *et al.*, "Statistical Total Correlation Spectroscopy: An Exploratory Approach for Latent Biomarker Identification from Metabolic 1H NMR Data Sets," *Analytical Chemistry*, vol. 77, no. 5, pp. 1282–1289, Mar. 2005, doi: 10.1021/AC048630X.

[97] H. C. Keun *et al.*, "Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling," *Analytica Chimica Acta*, vol. 490, no. 1–2, pp. 265–276, Aug. 2003, doi: 10.1016/S0003-2670(03)00094-1.

[98] D. v. Nguyen and D. M. Rocke, "Tumor classification by partial least squares using microarray gene expression data," *Bioinformatics*, vol. 18, no. 1, pp. 39–50, Jan. 2002, doi: 10.1093/BIOINFORMATICS/18.1.39.

[99] J. Trygg, "O2-PLS for qualitative and quantitative analysis in multivariate calibration," *Journal of Chemometrics*, vol. 16, no. 6, pp. 283–293, Jun. 2002, doi: 10.1002/CEM.724.

[100] "Trends in Chemometrics: Food Authentication, Microbiology, and Effects of Processing - Granato - 2018 - Comprehensive Reviews in Food Science and Food Safety - Wiley Online Library." https://ift.onlinelibrary.wiley.com/doi/full/10.1111/1541-4337.12341 (accessed Jan. 30, 2022).

[101] C. Maione and R. M. Barbosa, "Recent applications of multivariate data analysis methods in the authentication of rice and the most analyzed parameters: A review," *https://doi.org/10.1080/10408398.2018.1431763*, vol. 59, no. 12, pp. 1868–1879, Jul. 2018, doi: 10.1080/10408398.2018.1431763.

[102] P. de La Mata-Espinosa, J. M. Bosque-Sendra, R. Bro, and L. Cuadros-Rodríguez, "Olive oil quantification of edible vegetable oil blends using triacylglycerols chromatographic fingerprints and chemometric tools," *Talanta*, vol. 85, no. 1, pp. 177–182, Jul. 2011, doi: 10.1016/j.talanta.2011.03.049.

[103] G. K. Uyanık and N. Güler, "A Study on Multiple Linear Regression Analysis," *Procedia - Social and Behavioral Sciences*, vol. 106, pp. 234–240, Dec. 2013, doi: 10.1016/J.SBSPRO.2013.12.027.

[104] P. Oliveri and G. Downey, "Multivariate class modeling for the verification of food-authenticity claims," *TrAC - Trends in Analytical Chemistry*, vol. 35, pp. 74–86, May 2012, doi: 10.1016/J.TRAC.2012.02.005.

[105] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN Model-Based Approach in Classification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2888, pp. 986–996, 2003, doi: 10.1007/978-3-540-39964-3_62.

[106] A. M. Jiménez-Carvelo, A. González-Casado, M. G. Bagur-González, and L. Cuadros-Rodríguez, "Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review," *Food Research International*, vol. 122, pp. 25–39, Aug. 2019, doi: 10.1016/J.FOODRES.2019.03.063.

[107] L. Mutihac and R. Mutihac, "Mining in chemometrics," *Analytica Chimica Acta*, vol. 612, no. 1, pp. 1–18, Mar. 2008, doi: 10.1016/J.ACA.2008.02.025.

[108] L. A. Berrueta, R. M. Alonso-Salces, and K. Héberger, "Supervised pattern recognition in food analysis," *Journal of Chromatography A*, vol. 1158, no. 1–2, pp. 196–214, Jul. 2007, doi: 10.1016/J.CHROMA.2007.05.024.

[109] Y. Liu, Y. Wang, and J. Zhang, "New Machine Learning Algorithm: Random Forest," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7473 LNCS, pp. 246–252, 2012, doi: 10.1007/978-3-642-34062-8_32.

[110] K. Coussement and D. van den Poel, "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques," *Expert Systems with Applications*, vol. 34, no. 1, pp. 313–327, Jan. 2008, doi: 10.1016/J.ESWA.2006.09.038.

[111] E. Bauer, P. Chan, S. Stolfo, and D. Wolpert, "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants," *Machine Learning 1999 36:1*, vol. 36, no. 1, pp. 105–139, 1999, doi: 10.1023/A:1007515423169.

[112] S. Lessmann, M. C. Sung, and J. E. V. Johnson, "Alternative methods of predicting competitive events: An application in horserace betting markets," *International Journal of Forecasting*, vol. 26, no. 3, pp. 518–536, Jul. 2010, doi: 10.1016/J.IJFORECAST.2009.12.013.

[113] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, Jan. 1989, doi: 10.1016/0893-6080(89)90020-8.

[114] M. W. Gardner and S. R. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmospheric Environment*, vol. 32, no. 14–15, pp. 2627–2636, Aug. 1998, doi: 10.1016/S1352-2310(97)00447-0.

[115] C. Yu, M. T. Manry, J. Li, and P. Lakshmi Narasimha, "An efficient hidden layer training method for the multilayer perceptron," *Neurocomputing*, vol. 70, no. 1–3, pp. 525–535, Dec. 2006, doi: 10.1016/J.NEUCOM.2005.11.008.

[116] H. Ramchoun, "Multilayer Perceptron: Architecture Optimization and Training multi-criteria learning and nonlinear optimization View project," *Article in International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, pp. 1–26, 2016, doi: 10.9781/ijimai.2016.415.
[117] M. Cilimkovic and M. Jl, "Back Propagation Algorithm Neural Networks and Back Propagation Algorithm".
[118] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature 1986 323:6088*, vol. 323, no. 6088, pp. 533–536, 1986, doi: 10.1038/323533a0.
[119] R. Rojas, "The Backpropagation Algorithm," *Neural Networks*, pp. 149–182, 1996, doi: 10.1007/978-3-642-61068-4_7.
[120] R. Battiti, "First- and Second-Order Methods for Learning: Between Steepest Descent and Newton's Method," *Neural Computation*, vol. 4, no. 2, pp. 141–166, Mar. 1992, doi: 10.1162/NECO.1992.4.2.141.
[121] F. García López, M. García Torres, B. Melián Batista, J. A. Moreno Pérez, and J. M. Moreno-Vega, "Solving feature subset selection problem by a Parallel Scatter Search," *European Journal of Operational Research*, vol. 169, no. 2, pp. 477–489, Mar. 2006, doi: 10.1016/J.EJOR.2004.08.010.
[122] J. R. Quinlan, "Induction of decision trees," *Machine Learning 1986 1:1*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.
[123] W. R. Burrows, M. Benjamin, S. Beauchamp, E. R. Lord, D. McCollor, and B. Thomson, "CART Decision-Tree Statistical Analysis and Prediction of Summer Season Maximum Surface Ozone for the Vancouver, Montreal, and Atlantic Regions of Canada," *Journal of Applied Meteorology and Climatology*, vol. 34, no. 8, pp. 1848–1862, Aug. 1995, doi: 10.1175/1520-0450(1995)034.
[124] J. Su and H. Zhang, "A Fast Decision Tree Learning Algorithm Introduction and Related Work", Accessed: Jan. 31, 2022. [Online]. Available: www.aaai.org
[125] T. Maszczyk and W. Duch, "Comparison of shannon, renyi and tsallis entropy used in decision trees," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5097 LNAI, pp. 643–651, 2008, doi: 10.1007/978-3-540-69731-2_62.
[126] C. F. L. Lima, F. M. de Assis, and C. P. de Souza, "Decision tree based on shannon, rényi and tsallis entropies for intrusion tolerant systems," *5th International Conference on Internet Monitoring and Protection, ICIMP 2010*, pp. 117–122, 2010, doi: 10.1109/ICIMP.2010.23.
[127] M. Mosonyi and F. Hiai, "On the quantum Rényi relative entropies and related capacity formulas," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2474–2487, Apr. 2011, doi: 10.1109/TIT.2011.2110050.
[128] J. Havrda, F. C.- Kybernetika, and undefined 1967, "Quantification method of classification processes. Concept of structural -entropy," *dml.cz*, vol. 3, no. 1, pp. 30–35, 1967, Accessed: Jan. 31, 2022. [Online]. Available: https://dml.cz/bitstream/handle/10338.dmlcz/125526/Kybernetika_03-1967-1_3.pdf
[129] Q. Hu, X. Che, L. Zhang, D. Zhang, M. Guo, and D. Yu, "Rank entropy-based decision trees for monotonic classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 11, pp. 2052–2064, 2012, doi: 10.1109/TKDE.2011.149.
[130] D. Lachenmeier, "Guide to NMR Method Development and Validation-Part II: Multivariate data analysis".

[131] T. Gödecke *et al.*, "Validation of a Generic Quantitative 1H NMR Method for Natural Products Analysis," *Wiley Online Library*, vol. 24, no. 6, pp. 581–597, Nov. 2013, doi: 10.1002/pca.2436.
[132] G. F. Pauli, T. Gödecke, B. U. Jaki, and D. C. Lankin, "Quantitative 1H NMR. Development and potential of an analytical method: An update," *Journal of Natural Products*, vol. 75, no. 4, pp. 834–851, Apr. 2012, doi: 10.1021/NP200993K.
[133] V. Gallo *et al.*, "Performance assessment in fingerprinting and multi component quantitative NMR analyses," *ACS Publications*, vol. 87, no. 13, p. 41, Jul. 2015, doi: 10.1021/acs.analchem.5b00919.
[134] V. Dubois, S. Breton, M. Linder, J. Fanni, and M. Parmentier, "Fatty acid profiles of 80 vegetable oils with regard to their nutritional potential," *European Journal of Lipid Science and Technology*, vol. 109, no. 7, pp. 710–732, Jul. 2007, doi: 10.1002/EJLT.200700040.
[135] J. Donarski, F. Camin, C. Fauhl-Hassek, R. Posey, and M. Sudnik, "Sampling guidelines for building and curating food authenticity databases," *Trends in Food Science & Technology*, vol. 90, pp. 187–193, Aug. 2019, doi: 10.1016/J.TIFS.2019.02.019.
[136] "ISO - CEN - European Committee for Standardization." https://www.iso.org/organization/250321.html (accessed Feb. 01, 2022).
[137] "Interlaboratory comparisons | EU Science Hub." https://ec.europa.eu/jrc/en/interlaboratory-comparisons (accessed Jan. 31, 2022).
[138] A. C. Olivieri, "Analytical advantages of multivariate data processing. One, two, three, infinity?," *Analytical Chemistry*, vol. 80, no. 15, pp. 5713–5720, Aug. 2008, doi: 10.1021/AC800692C.
[139] S. Wold, M. Sjostrom, and B. Kowalski, "SIMCA: a method for analyzing chemical data in terms of similarity and analogy." ACS, 1977.
[140] K. Shikha Ojha, D. Granato, G. Rajuria, F. J. Barba, J. P. Kerry, and B. K. Tiwari, "Application of chemometrics to assess the influence of ultrasound frequency, Lactobacillus sakei culture and drying on beef jerky manufacture: Impact on amino acid profile, organic acids, texture and colour," *Food Chemistry*, vol. 239, pp. 544–550, Jan. 2018, doi: 10.1016/J.FOODCHEM.2017.06.124.
[141] D. Granato *et al.*, "Trends in Chemometrics: Food Authentication, Microbiology, and Effects of Processing," *Comprehensive Reviews in Food Science and Food Safety*, vol. 17, no. 3, pp. 663–677, May 2018, doi: 10.1111/1541-4337.12341.
[142] V. Lagouri, A. G.-F. R. B. and Medicine, and undefined 2016, "Optical Non Destructive UV-VIS-NIR Spectroscopic Tools and Chemometrics in the Monitoring of Olive Oil Phenolic Compounds and Oxidation," *Elsevier*, Accessed: Jan. 31, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0891584916307316
[143] J. Riedl, S. Esslinger, and C. Fauhl-Hassek, "Review of validation and reporting of non-targeted fingerprinting approaches for food authentication," *Analytica Chimica Acta*, vol. 885, pp. 17–32, 2015, doi: 10.1016/j.aca.2015.06.003.
[144] "International Vocabulary of Metrology–Basic and General Concepts and Associated Terms," *Chemistry International -- Newsmagazine for IUPAC*, vol. 30, no. 6, pp. 21–22, Nov. 2008, doi: 10.1515/CI.2008.30.6.21.
[145] P. de Bièvre, R. Dybkaer, A. Fajgelj, and D. B. Hibbert, "Metrological traceability of measurement results in chemistry: Concepts and implementation (IUPAC Technical report)," *Pure and Applied Chemistry*, vol. 83, no. 10, pp. 1873–1935, 2011, doi: 10.1351/PAC-REP-07-09-39.

[146] "Understanding NMR Spectroscopy - James Keeler - Google Libri." https://books.google.it/books?hl=it&lr=&id=WUmCpq30pygC&oi=fnd&pg=PT13&dq=%5B6%5D+J.+Keeler,+Understanding+NMR+Spectroscopy,+John+Wiley+and+Sons,+2010.&ots=SsBBVdcQSV&sig=5Jd1vkKvojgf8SEGCZteTU8Xey8#v=onepage&q&f=false (accessed Jan. 31, 2022).

[147] "NMR Spectroscopy Explained: Simplified Theory, Applications and Examples for ... - Neil E. Jacobsen - Google Books." https://books.google.it/books?id=KCkiiQ0uefoC&printsec=frontcover&dq=nmr%20+theory&hl=en&sa=X&ved=%200ahUKEwjzgJCUv9fiAhUEjqQKHckbBi4Q6AEIKjAA#v=onepage&q=nmr%20theory&f=false (accessed Jan. 31, 2022).

[148] K. Chen, "Mehdi Mobli and Jeffrey C. Hoch (Eds.): Fast NMR data acquisition: beyond the Fourier transform," *Analytical and Bioanalytical Chemistry*, vol. 410, no. 6, pp. 1615–1616, Feb. 2018, doi: 10.1007/S00216-017-0846-0.

[149] T. Parella, "Towards perfect NMR: Spin-echo versus perfect-echo building blocks," *Magnetic Resonance in Chemistry*, vol. 57, no. 1, pp. 13–29, Jan. 2019, doi: 10.1002/MRC.4776.

[150] E. Kupče and T. D. W. Claridge, "Molecular structure from a single NMR supersequence," *Chemical Communications*, vol. 54, no. 52, pp. 7139–7142, 2018, doi: 10.1039/C8CC03296C.

[151] M. Foroozandeh, R. W. Adams, N. J. Meharry, D. Jeannerat, M. Nilsson, and G. A. Morris, "Ultrahigh-resolution NMR spectroscopy," *Angewandte Chemie - International Edition*, vol. 53, no. 27, pp. 6990–6992, Jul. 2014, doi: 10.1002/ANIE.201404111.

[152] D. Raftery, "High-throughput NMR spectroscopy," *Analytical and Bioanalytical Chemistry*, vol. 378, no. 6, pp. 1403–1404, Mar. 2004, doi: 10.1007/S00216-003-2437-5.

[153] Y. Liu, J. Cheng, H. Liu, Y. Deng, J. Wang, and F. Xu, "NMRSpec: An integrated software package for processing and analyzing one dimensional nuclear magnetic resonance spectra," *Chemometrics and Intelligent Laboratory Systems*, vol. 162, pp. 142–148, Mar. 2017, doi: 10.1016/j.chemolab.2017.01.005.

[154] A. Mohamed, C. H. Nguyen, and H. Mamitsuka, "NMRPro: An integrated web component for interactive processing and visualization of NMR spectra," *Bioinformatics*, vol. 32, no. 13, pp. 2067–2068, Jul. 2016, doi: 10.1093/BIOINFORMATICS/BTW102.

[155] C. Cobas, I. Iglesias, and F. Seoane, "NMR data visualization, processing, and analysis on mobile devices," *Magnetic Resonance in Chemistry*, vol. 53, no. 8, pp. 558–564, Aug. 2015, doi: 10.1002/MRC.4234.

[156] U. S. Ellison Secretary *et al.*, "EURACHEM/CITAC Guide Quantifying Uncertainty in Analytical Measurement Composition of the Working Group* EURACHEM members A Williams Chairman A Brzyski R Kaus E Amico di Meane M Rösslein A Fajgelj IAEA Vienna," 2009.

[157] M. R. Viant *et al.*, "International NMR-based environmental metabolomics intercomparison exercise," *Environmental Science and Technology*, vol. 43, no. 1, pp. 219–225, Jan. 2009, doi: 10.1021/ES802198Z.

[158] S. Medina, R. Perestrelo, P. Silva, J. A. M. Pereira, and J. S. Câmara, "Current trends and recent advances on food authenticity technologies and chemometric approaches," *Trends in Food Science & Technology*, vol. 85, pp. 163–176, Mar. 2019, doi: 10.1016/J.TIFS.2019.01.017.

[159] D. Granato, A. Koot, and S. M. van Ruth, "Geographical provenancing of purple grape juices from different farming systems by proton transfer reaction mass spectrometry using supervised statistical techniques," *Journal of the Science of Food and Agriculture*, vol. 95, no. 13, pp. 2668–2677, Oct. 2015, doi: 10.1002/JSFA.7001.

[160] N. Z. Ballin and K. H. Laursen, "To target or not to target? Definitions and nomenclature for targeted versus non-targeted analytical food authentication," *Trends in Food Science and Technology*, vol. 86, pp. 537–543, Apr. 2019, doi: 10.1016/J.TIFS.2018.09.025.
[161] R. Consonni and L. R. Cagliani, "The potentiality of NMR-based metabolomics in food science and food authentication assessment," *Magnetic Resonance in Chemistry*, vol. 57, no. 9, pp. 558–578, Sep. 2019, doi: 10.1002/MRC.4807.
[162] U. K. Sundekilde, N. Eggers, and H. C. Bertram, "NMR-Based Metabolomics of Food," *Methods in Molecular Biology*, vol. 2037, pp. 335–344, 2019, doi: 10.1007/978-1-4939-9690-2_18.
[163] A. H. Emwas *et al.*, "Nmr spectroscopy for metabolomics research," *Metabolites*, vol. 9, no. 7, Jul. 2019, doi: 10.3390/METABO9070123.
[164] S. K. Bharti and R. Roy, "Quantitative 1H NMR spectroscopy," *TrAC Trends in Analytical Chemistry*, vol. 35, pp. 5–26, May 2012, doi: 10.1016/J.TRAC.2012.02.007.
[165] A. Craig, O. Cloarec, E. Holmes, J. K. Nicholson, and J. C. Lindon, "Scaling and normalization effects in NMR spectroscopic metabonomic data sets," *Analytical Chemistry*, vol. 78, no. 7, pp. 2262–2267, Apr. 2006, doi: 10.1021/AC0519312.
[166] L. R. Euceda, G. F. Giskeodegård, and T. F. Bathen, "Preprocessing of NMR metabolomics data," *Scandinavian Journal of Clinical and Laboratory Investigation*, vol. 75, no. 3, pp. 193–203, May 2015, doi: 10.3109/00365513.2014.1003593.
[167] R. A. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, "Centering, scaling, and transformations: Improving the biological information content of metabolomics data," *BMC Genomics*, vol. 7, Jun. 2006, doi: 10.1186/1471-2164-7-142.
[168] "Recupero del Germoplasma Viticolo Pugliese – Frutti Antichi Puglia." http://www.fruttiantichipuglia.it/il-progetto/regevip/ (accessed Jan. 31, 2022).
[169] Á. D. Nyitrainé Sárdy, M. Ladányi, Z. Varga, Á. P. Szövényi, and R. Matolcsi, "The Effect of Grapevine Variety and Wine Region on the Primer Parameters of Wine Based on 1H NMR-Spectroscopy and Machine Learning Methods," *Diversity*, vol. 14, no. 2, p. 74, Jan. 2022, doi: 10.3390/D14020074.
[170] "FoodIntegrity:" https://secure.fera.defra.gov.uk/foodintegrity/index.cfm?sectionid=21 (accessed Jan. 31, 2022).
[171] "Progetto P.A.S.C.Qua. Archives - Progeva Srl." http://www.progeva.it/category/news/progetto-p-a-s-c-qua/ (accessed Jan. 31, 2022).
[172] M. Rusilowicz, M. Dickinson, A. Charlton, • Simon O'keefe, and J. Wilson, "A batch correction method for liquid chromatography-mass spectrometry data that does not depend on quality control samples", doi: 10.1007/s11306-016-0972-2.
[173] D. Cavanna, L. Righetti, C. Elliott, and M. Suman, "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," *Trends in Food Science & Technology*, vol. 80, pp. 223–241, Oct. 2018, doi: 10.1016/J.TIFS.2018.08.007.
[174] "Sampling and Sample Preparation in Field and Laboratory: Fundamentals and ... - Google Libri." https://books.google.it/books?hl=it&lr=&id=LQ2wl4eRtbsC&oi=fnd&pg=PR31&dq=Pawliszyn,+2002&ots=QXLgb4K_sJ&sig=dD5Lr7XN0CipRgcpcoONTjxjI0U#v=onepage&q=Pawliszyn%2C%202002&f=false (accessed Jan. 31, 2022).