Università degli Studi di Bologna

Dottorato di Ricerca in Scienze degli Alimenti – XIX Ciclo –

Settore Scientifico-Disciplinare CHIM. 06

Coordinatore: Chiar.mo Prof. Claudio Cavani

A CASE STUDY DEVELOPMENT OF CHEMICAL INDEXES, ORIGINATED FROM THE APPLICATION OF CHEMOMETRIC METHODS TO THE NUCLEAR MAGNETIC RESONANCE (NMR), IN THE ASSESMENT OF THE QUALITY AND OF THE GEOGRAPHICAL ORIGIN OF VEGETABLE PRODUCTS

PhD FINAL DISSERTATION

Presentato
dal Dottore:
FRANCESCO SAVORANI

Relatore:
Chiar.mo Prof.
GIUSEPPE PLACUCCI

Correlatore:
Chiar.mo Prof.
FRANCESCO CAPOZZI

A.A. 2006-2007

PREFACE

The present thesis work has been developed towards the direction of providing an objective method finalized to the definition of food quality, by identification of suitable chemical parameters among a wide range of choices. Since foodstuffs are complex mixtures of substances with a large dynamic range of concentrations, it is restrictive to give importance to a limited number of nutrients in the definition of quality. In fact, whatever is the preferred choice, arbitrariness is introduced in the weight that is attributed to every parameter, selected on the only basis of the current knowledge and on the available instrumentations.

Spectroscopies earn, from this point of view, notable success because they are able to give, in only one shot, a picture enough general of the chemical composition of foods. The limits of their widespread application are set by the fact that the most promising of them, e.g. Nuclear Magnetic Resonance, is also the least sensitive instrumental technique available to the analysts. However, as the limits of sensitivity are gradually going to diminish with the advent of cryoprobes, the attractiveness of NMR is raising more and more, also in virtue of the fact that such a defect is however compensated by the wealth of information. Many chemical parameters are deducible, at once, from a single NMR spectrum, otherwise obtainable with a long series of alternative analysis. In literature, indeed, the articles that describe one-shot polyvalent analysis of foods are largely proliferated in the last decade.

Since the information becomes more complex and overwhelming, it is more and more pressing the necessity to individuate statistic methods able to extract the useful information and to discard the useless data. Often, the statement "too much information means noise" is expressed to underline the difficulty to find discriminant features among plenty of data. The research work carried out during the three years of this thesis is an example of symbiotic integration between instrumental and multivariate analyses, the former giving spectra to be transformed in vectors of long sequence of numbers, the latter extracting from such strings the parameters necessary to objectively depict the quality of foods. The case study is focused on the definition of the differences that make the "pomodoro di Pachino" worth to be protected, with a European Quality Mark, from the attempts to use in a fraudulent way its name for tomatoes of different quality. For an Italian food scientist, it is mandatory to find the way to protect, from imitations, such a vegetable product which is, together with olive oil and mozzarella, the fundamental ingredients of the most famous Italian food: that is "pizza".

1 INTRODUCTION

1.1 FOOD QUALITY

The quality of foods is of primary importance for both consumers and industries, at all levels of the production process, from raw materials to finished products. Quality standards have been established through the requirement of quality labels, that specify the chemical composition of each product, and its authentication is essential to avoid unfair competition that can create a destabilized market and disrupt the regional and national economy. Thus, scientists researching in the field of food and beverage industry are faced with many different quality control tasks, e.g., making sure that flavors meet certain standards, identifying changes in process parameters that may lead to a change in quality, detecting adulteration in any ingredient, and identifying the geographical origin of raw materials.
Many of these quality control issues have traditionally been assessed by panels of experts, who are able to determine a product's quality by observing their organoleptic characteristics such as color, texture, taste, aroma, etc. However, it takes years of experience to acquire these skills. It would therefore be advantageous if there were a way for food scientists to measure the quality of a product by instrumental analyses.

1.1.1 Instrumental analyses for the evaluation of food quality

Many techniques are used and are usable to carry out analytical control with the final goal of food characterization or adulteration detection. High-performance liquid chromatography (HPLC), for example, represents one of the most widely used techniques as a quality control tool, because it can separate various chemical constituents of mixtures. Numerous chemical compounds have been extensively analyzed by HPLC either for food products characterization or to detect adulteration [1-8].
In addition, gas chromatography (GC) is used for separating volatile organic compounds and GC-MS coupling is the most widely used hyphenated technique even if the use of GC-FTIR (gas chromatography with Fourier transform infrared spectroscopy) is increasingly adopted because it gives some structural information on the chemical functional groups of the

molecules but at lower costs than GC-MS and with less maintenance. GC is generally used to discriminate among different varieties of the same product, adulteration detection, and organic compound authentication and identification [9-11].

The determination of nitrogenous content in food products such as cheese, milk, or honey is usually performed with numerous techniques based on the chemical properties of protein, peptides, and/or amino acids. Most of these are spectrophotometric or fluorometric techniques [12-15], also employed to characterize other classes of substances, such as fatty acids, sugars, vitamins, or mineral elements, entering in the composition of a wide variety of food products [16-19].

Others very useful techniques for the determination of the mineral content in food products are the Atomic Absorption/Atomic Emission (AAS/AES) and Inductively Coupled Plasma-Atomic Emission (ICP-AES) analytical techniques, that found their principal application in multi-element analysis of wines[20], sugar[21], fruit [22, 23], cheeses [24], and honeys [25, 26].

Furthermore, Isotope Ratio Mass Spectrometry (IRMS) has been widely applied for the determination of food authenticity and also for origin control in honey and many other food products [27]. Since the elemental isotopic abundance in biological products varies in relation to natural processes, IRMS provides important information for studies with different finalities. The isotopic signatures of bio-molecules depend upon geographical parameters, and seasonal effects [28]. For example, the measurements of stable isotope ratio of light elements ($^{2}H/^{1}H$, $^{15}N/^{14}N$, $^{13}C/^{12}C$, $^{18}O/^{16}O$, $^{34}S/^{32}S$) and of heavy element $^{87}Sr/^{86}Sr$ (a trace element), have been used to detect regional provenance and supply authenticity control information for products like fruit juices, wine, milk, butter and cheese [26, 28-32]. These analyses point out chemical differences between food samples originating from a specific region and, conversely, similarities between foodstuffs produced in different regions. In other words, stable isotope analysis enables the differentiation of chemically identical substances, but with different origins, through their specific isotopic fingerprints.

A complementary technique to the chemical analytical methods mentioned above is Differential Scanning Calorimetry (DSC). It is the most widely used among all thermal analysis techniques and has a great utility in quality assurance of food. This technique measures the difference in energy transfer to a sample and to a reference material, as a function of time or temperature, while the sample and the reference material are subjected to a controlled time-temperature program. So it provides qualitative and quantitative

information regarding transitions in materials that involve endothermic (e.g. melting) or exothermic processes (e.g. crystallization) or changes in heat capacity (e.g. glass transition).
Proteins are the main food components studied by thermal analysis including studies on conformational changes of these macromolecules, thermal denaturation of tissue proteins, food enzymes and enzyme preparations for the food industry, as well as effects of various additives on their thermal properties [33-35]. Frozen foods are also monitored by DSC to measure their thermal properties, and to estimate the state of their constituent water, naturally presents in foods or added as ingredient in food preparations. Moreover DSC is used to determine gelatinization behavior of starches and interaction of starch with other food components [36-41]; to characterize polysaccharides and to study their phase transitions during baking processes [42]; to observe change in lipid composition during fusion or crystallization of fats [43]; to predict oil stability during thermal oxidative decomposition of edible oils. At last research in food microbiology utilizes DSC in better understanding thermoadaptive mechanisms or heat killing of food-borne microorganisms.
Since multiple interactions can arise between food components and lead to some modifications of the thermal behavior of foods [44, 45], a good understanding of their thermal properties has great value for a good definition of product quality [46-55] and for detecting alteration [56] or adulteration [57-59].

1.2 NUCLEAR MAGNETIC RESONANCE (NMR) IN FOOD SCIENCE

In recent years, the constantly growing potentiality of Nuclear Magnetic Resonance (NMR) methods has found an increasing application in the field of food chemistry. During the past 50 years NMR spectroscopy first and Magnetic Resonance Imaging (MRI) later, has evolved from being expensive, academic and not appropriate for industrial applications into an extremely powerful analytical technique able to elucidate chemical structures, molecular conformations and dynamics of food constituents in the liquid or solid state.
The great advantage of NMR in complex systems such as foodstuffs is that the spectroscopic method properly filters the information that is observed by the spectrometer. At the most basic level, signals from only one kind of nucleus (e.g. ^{1}H, ^{13}C, ^{31}P, ^{15}N, ^{23}Na, etc.) are observed at a time.

Different nuclei give rise to characteristic spectral lines, in different position, depending on the different chemical environment. Therefore, it is possible to observe a specific nucleus in a particular environment even in complex structures, both in solution and in the solid state.
In literature, numerous applications in complex food systems are published: NMR is used to examine the chemical and physical characteristics of meat, fish, dairy products, vegetables, fruits, juices, pastries, cheese, wine and emulsified materials. Specific properties can be measured, including freeze-thaw, percent alcohol, fruit ripeness [60], sugar content, moisture content, state of water, oil/water ratio, saturate versus unsaturated fatty acid content, food adulteration [61]. NMR applications also include the determination of enzyme activity and water macromolecular interactions, the detection of bacterial spoilage and the study of rheology and mixing in multi-phase process streams. All of these measurable parameters ultimately lead to keep control of product quality and many of these characterizations can be made on-line and thus influence processes control. Nowadays, different NMR experiments and devices, such as low-resolution NMR, Site-Specific Natural Isotopic Fractionation–NMR (SNIF-NMR), Magnetic Resonance Imaging (MRI) and high-resolution NMR, are available and yield different information.

1.2.1 Low-Field NMR (LF-NMR) spectroscopy

Most low-field NMR applications involve the measurement of spin–spin (T_2) and/or spin–lattice (T_1) relaxation times, which are related to specific physical properties, such as viscosity, surface area and moisture content. Other applications involve analysis of the FID or spin–echo signals to yield quantitative information relating to the concentrations of individual components, which can be distinguished by virtue of different T_2 values [62].
Low-field, NMR spectroscopy has been extensively used in foodstuffs analyses, such as the determination of moisture [63-68], fat [69-71], hydrogen and fluorine contents [72, 73].
The spin–lattice relaxation time (T_1) has also been measured on hen egg to evaluate the change of quality during its first few days of storage [74]. Moreover, it has been explored as parameter to be measured, with on-line NMR sensor, for assessing the egg quality in a non-destructive and fast way. The study was focused mainly on the thick albumen trying to find correlations between its NMR properties and viscoelasticity, an important internal quality factor that, together with water content, decreases during storage [75]. The viscoelasticity of fresh thick albumen appears to be related to the existence of a cross-linked gel-like protein

network formed by the ovomucin–lysozyme complex. The change of viscoelasticity observed during the storage suggested that the nuclear relaxation times of water may be parallel to the modifications of the dynamic state of biopolymers' network. The gradual degradation of this network during the storage, accounting for the progressive decrease in viscoelasticity, was found to correlate with water proton T_1 measured at low-field, thus indicating that such an NMR parameter may be a good indicator of albumen transformation, a.k.a. the thinning phenomenon [74, 76]. Moreover an external unit was developed that permits using a commercial electromagnet spectrometer for measurements of T_1 at very low fields by lowering its frequency down to the 700 kHz-1 MHz range [77]. The application of such an instrument, at 700 kHz, was demonstrated useful with hen shell eggs, whose relaxation time measured on its water protons can give an indication of the freshness of the egg. The very low resonance frequency allowed by such an external unit permits the highest discrimination between fresh and timed albumen. The constructive layout of the simple probe gave the capability of containing large samples, thus giving the possibility of performing non-destructive analyses on whole shell eggs [77]. This feature, which also eliminates the risk of microbial contamination, makes the approach suitable for on-line applications.

1.2.2 NMR-MObile Universal Surface Explorer (NMR -MOUSE)

The concept of the NMR-MOUSE was first proposed in 1996 by Eidmann et al. [78] and further described by Blumich et al. [79] and Balibanu et al. [80]. The NMR-MOUSE is a small and portable LF-NMR system with a one-sided magnet layout that replaces the conventional magnet and probe on a bench-top LF-NMR system. This LF-NMR device allows unrestricted access to large intact samples, giving few restrictions to sample geometry.

The potential of the NMR-MOUSE applied to food systems has been firstly evaluated in a study that has measured the concentration of oil and water in emulsions [81]. In a successive study has been developed a method for fat content measurements in live or slaughtered fish using the Bruker Professional Mouse®, supporting how this technique has good potential for the quantitative analysis of intact food products [82]. Moreover, the NMR-MOUSE technique has been applied for measurements on beverages, contained in closed bottles, for determining the oxygen content in table superoxygenated water [83].

The version with a "semisingle-sided" sensor has been reputed an appropriate tool for nondestructive NMR relaxation measurements. The term "semisingle-sided" refers to the

open bay which receives the sample in a single-sided RF coil. This type of sensor has allowed much better sensitivity without sacrificing the necessary open access needed for measurements on entire food packages, such as bottles, making this sensor a tool for practical applications in food science.

1.2.3 SNIF NMR

SNIF-NMR®, stands for Site-Specific Natural Isotopic Fractionation studied by Nuclear Magnetic Resonance [84]. This type of spectroscopy, developed by G.J. Martin in 1980's [85, 86], is based upon the natural isotopic distribution of atoms inside molecules. The technique is able to determine the $^{2}H/^{1}H$ isotopic ratio of the different sites of a molecule. This information, supported by a robust database of geographical isotopic radio distribution, permits to individuate the botanic and geographic origin of natural substances such as sugars, ethanol, aromas, glycerol, fatty acids, etc. The main application of this technique has been so far the wine sector [EC Reg. n. 2676/90] but there are many other food and/or natural products, such orange juices [87], honey [88], olive oil and tobacco leaves [89] that are subjects of SNIF-NMR.

1.2.4 Magnetic Resonance for Imaging (MRI)

The NMR experiment, when performed in a magnetic field with a controlled gradient, provides spatial distribution of observed spins, i.e. the NMR image, rather than NMR spectrum. The NMR images give non-invasive pictures of cross-sections of biological objects. Combined with spectroscopy, NMR images give a detailed map of the physiological state of a studied model.

The most widespread use of MRI is as a diagnostic tool in the field of medicine. However, over recent years MRI has been applied extensively in the food industry to obtain both structural and dynamic information of a wide variety of foodstuffs [90, 91].

The images can be quantified to yield information about several processes and material properties, such as mass and heat transfer, fat and ice crystallization, gelation, water mobility, composition and volume changes, food stability and maturation, flow behavior, and temperature [92]. For example, this technique has been used to map the internal structure

of fruit and vegetables. Monitoring the tissue morphology has allowed to study the quality of food in terms of its ripeness[93, 94] and bruising [95] or in terms of alteration upon freezing [96].

MRI has also been used to asses the quality of cheese by mapping or imaging moisture and fat contents, providing real-time and real-location information on the distribution of moisture and fat in cheese blocks, while the cheese blocks are being cured or cooled or aged [97].

1.2.5 High Resolution Nuclear Magnetic Resonance (HR-NMR)

High-resolution nuclear magnetic resonance (NMR) spectroscopy is a powerful technique for the identification of pure organic compounds, but in the last fifteen years the use of NMR in chemically complex and highly heterogeneous systems, has consistently grown thanks to the development of 2D and multiD experiments [98, 99].

The success of this technique is due to the current availability of NMR equipments of high magnetic field strength (800-900 MHz) and to the development of cryogenic probe technologies. The latter have significantly increased the sensitivity of the systems, affecting experimental time and/or concentration of compounds needed to obtain adequate spectra [100].

These high-field devices, combined with the use of two-dimensional (2D) homonuclear and heteronuclear correlated techniques using pulsed field gradients, have facilitated the identification of specific compounds in complex mixtures without purification or severe and time-consuming extractions which generate qualitative and quantitative modifications [101].

Capitalizing its growing potentiality to solve spectra of complex mixtures and to quantify the corresponding components without chemical separation, high-field NMR methods have found an increasing application in the field of food chemistry [102-104].

The advantages of the NMR technique with respect to other analytical methods are the non invasive approach, preserving food structure, the high specificity and selectivity reachable, and the possibility to provide information on a wide range of metabolites in a single experiment. NMR also supports the food industry in its increasing need to understand and be innovative in products and process and provides a new method to enforce legislation and control quality [105-112]. A well-known example is the authentication of olive oil by using ^{13}C NMR spectroscopy [113-115]. As mentioned in the reference work, the fatty acid amount, as well as the saturated, monounsaturated and polyunsaturated fatty acid ratios has

been determined from a single ^{13}C NMR spectrum. Moreover the presence of unsaturated trans-isomers was detected, and the distribution of fatty acids on the glycerol chain was determined [114, 116, 117].
Finally, many relevant NMR studies have been published on high resolution NMR techniques, finding chemical details on different types of food and drinks which include wine [118-121], olive oil [113-115] [116, 117], coffee [122-124], fruit juices [125, 126], tomatoes [127], vegetables [128], milk and dairy products [129-132], meat [133] and flour [134].

1.3 NOTIONS ON NMR SPECTROSCOPY

NMR spectroscopy investigates on matter by studying its magnetic nuclei by aligning them along a very powerful external magnetic field and perturbing this alignment using an electromagnetic field. The resulting response to the external perturbing electromagnetic field is the phenomenon that is observed and exploited in nuclear magnetic resonance, both spectroscopy and imaging.

1.3.1 Discovery

The first step in the discovery of the NMR phenomenon begun in the 1930's with the pioneering works of the scientists Gorter, Stern, Gerlach and Rabi. Gorter attributed the coining of the phrase "nuclear magnetic resonance" to Rabi, in a publication which appeared in the Netherlands in 1942 [135]. Only in 1946, independently from each other, two scientists working in the United States described a physical-chemical phenomenon depending on the magnetic properties of certain nuclei. When introduced into a strong magnetic field, these nuclei would adsorb energy, by resonance, in the radio-frequency interval of electromagnetic radiations and re-emit this energy afterwards. The observation that different atoms within a single molecule resonate at different frequencies, at a given magnetic field strength, allows discovering of a wide range of information about the chemistry and the structure of the molecule.
The phenomenon was called nuclear magnetic resonance (NMR) because the magnetic field strength and the radio-frequency must match each other: nuclear because only the nuclei of the atoms react; magnetic because it happens in a magnetic field; and resonance because of

the direct dependence of field strength and frequency. For this discovery the two scientists, Felix Bloch and Edward M. Purcell, were awarded the Nobel Prize in Physics in 1952.
The concept of "chemical shift", which is at the basis of spectroscopy, was exposed for the first time in the article published by Proctor and Yu in 1950, titled: "The Dependence of a Nuclear Magnetic Resonance Frequency upon Chemical Compound"[136].

1.3.2 Continuous wave (CW) spectroscopy

At its beginning, and for the first few decades, nuclear magnetic resonance spectrometers utilized a technique called continuous-wave (CW) spectroscopy, even if this technique may be performed in two different manners: either the magnetic field is kept constant and the oscillating electromagnetic field is scanned in frequency to diagram the resonant frequencies of the nuclei present or, more often, the oscillating field is set at a fixed frequency (continuous wave) and the magnetic field is varied looking for all the frequencies involved in NMR phenomenon in the analyzed system. The limit of CW spectroscopy is that it investigates each frequency separately, in succession. This makes CW experiments rather slow and, because magnetic resonance phenomenon is, for the reasons below explained, intrinsically insensitive, the resulting spectra suffer from a poor signal-to-noise ratio (S/N). A way to overcome such a poor signal-to-noise ratio is that it can be improved by signal averaging. The latter is a method where the NMR signals from many successive scans are added together. The random character of the noise leads to average the noise itself toward the normal baseline oscillations, while the actual nuclei signal is constant and additive. For this reasons several scans of the sample submitted to NMR analysis are often required in order to obtain an adequate S/N ratio but obviously this determines an increase of the experiment time.

1.3.3 Pulsed NMR spectroscopy and Fourier's transform.

In the mid 1960's Richard R. Ernst successfully preformed the first experiments with the technique presently known as "Fourier transform nuclear magnetic resonance spectroscopy" (FT-NMR). For that he won the Nobel Prize in chemistry in 1991. A great consequence of this improved NMR technique was that the time required for a single scan was dramatically reduced by allowing a range of frequencies to be probed at once. Successively this technique

has been made more and more practical and affordable with the development of two technologies: the knowledge of how to create an array of frequencies at once and computers capable of performing the computationally-intensive mathematical transformation of the data from the time domain to the frequency domain to produce a spectrum. In FT-NMR the sample, put into a static external strong magnetic field, is irradiated with a very short (of the order of μs) square pulse of radiofrequency energy containing all the frequencies in the range of interest. This is possible because the Fourier decomposition of an approximate square wave, for the Heisenberg uncertainty principle, contains contributions from all the frequencies in the neighborhood of the main frequency. The polarized magnets of the nuclei, when exposed to this radiofrequency energy, start to precess together, determining an oscillation of the surrounding magnetic field that is observable because it will induce current in a surrounding coil of conductive material. When the radiofrequency pulse is ended and, consequently, the associate energy is not further provided, the precessions decay to the equilibrium state making the polarization vector aligning with the field.

The strength of the magnetic field determines the frequency at which magnetic resonance of the nucleus to be studied occurs. The stronger the magnetic field applied the higher the resonance frequency, the greater the spectral resolution, the higher the sensitivity of the instrument, and the better the spectrometer.

Modern NMR spectrometers have now superconducting magnets providing the strongest magnetic fields obtainable with the present technology. The most common nucleus to be observed by NMR is that of hydrogen, although most of the elements in the periodic table have at least one isotope sensitive to NMR.

The fact that not less than 99.9% of organic molecules and compounds contain at least one hydrogen proton make NMR spectroscopy the perfect tool for investigating natural organic products. This has led to the common use of the hydrogen resonant frequency for a given NMR spectrometer as a measure of the magnetic field strength of that spectrometer. NMR spectrometers that use permanent or electromagnets range from 10 MHz up to 100 MHz, while spectrometers with superconducting magnets range from 200 MHz to, until now, 900 MHz.

High resolution NMR (HR-NMR) spectrometers are still very expensive and require specialized technicians for their use and maintenance. However, the high spectral resolution achieved by these instruments is not always required in every research field and hence lately many low field instruments have been developed and commercialized. These instruments use permanent or electronic magnets compatible with small dimensions, similar to those of other

bench-top analytical instruments, and are resulted perfect for many research field among which also food chemistry.

The initial cost and maintenance of these instruments is affordable for most of research laboratories and their use is simpler than high filed spectrometer. Normally the range of their magnetic field intensity varies between 0.23 and 0.70 Tesla (T) corresponding to 10 to 30 MHz Larmor's frequency for ^{1}H protons. However, there are examples of applications suggested at much lower magnetic field like that one described for the home made external device used in the direct analysis of intact shell eggs (see Section 1.2.1).

Since they are cheaper, smaller, lighter and less sensitive to fluctuations in environment and magnetic field than HR Spectrometers they represent, very likely, the future of NMR for industrial on-line applications.

1.3.4 Theoretical principles

1.3.4.1 Nuclear Magnetic Resonance phenomenon

Every atomic nucleus possesses an electric charge and in some of them, because of the rotation (spinning) around the nuclear axis, this induces a magnetic dipole oriented along the rotational axis. Likewise electrons, these atomic nuclei present an intrinsic property called "spin angular momentum" such as that, if they are put into a magnetic field, they can assume only two different states or orientations denoted with the Greeks letters α and β (Figure 1.1).

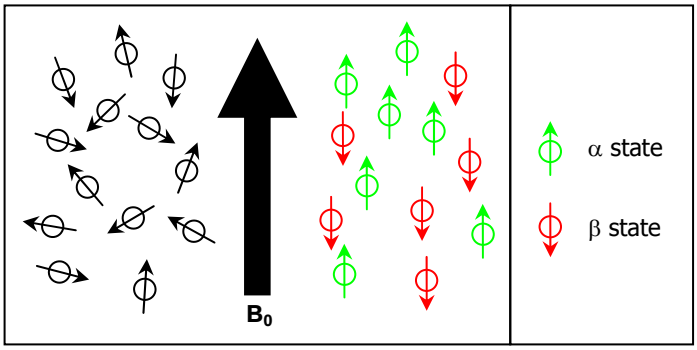


Figure 1.1: Effect of the external magnetic field on the orientation of the nuclear magnetic dipole

Spin angular momentum is a quantized property and it can assume different integer or half-integer values for a given nucleus. Protons and neutrons possess a spin angular momentum whose value can be +½ or –½ and inside a given atomic nucleus, protons can pair with other antiparallel protons in the same way that electrons pair in a chemical bond. Neutrons behave in the same way. The resulting net value of spin in the case of paired particles is equal to zero "0" but a nucleus with unpaired protons and neutrons will have a non-zero overall spin, with the number unpaired contributing ½ to the overall nuclear spin quantum number *I*.

Table 1.1: Magnetic properties of some nuclei [137]

I	Atomic mass	Atomic number	Example (I)
Half-integer	Odd	Odd or even	${}^{1}_{1}H(\frac{1}{2})$, ${}^{17}_{8}O(\frac{5}{2})$, ${}^{15}_{7}N(\frac{1}{2})$
Integer	Even	Odd	${}^{2}_{1}H(1)$, ${}^{14}_{7}N(1)$, ${}^{10}_{5}B(3)$
Zero	Even	Even	${}^{12}_{6}C(0)$, ${}^{16}_{8}O(0)$, ${}^{34}_{16}N(0)$

When this number is not equal to zero, a nucleus will present a spin angular momentum and an associated magnetic moment, μ, depending on the direction of the spin. It is this magnetic moment μ that is possible to exploit in every NMR experiment.

Examples of nuclei of this type, interesting for NMR applications, are:

^{1}H (Natural abundance = 99.98%)

^{13}C (Natural abundance = 1.1%)

^{15}N (Natural abundance = 0.37%)

^{31}P (Natural abundance = 100%)

^{17}O (Natural abundance = 0.038%)

^{23}Na (Natural abundance = 100%)

If a nucleus presenting this characteristics is put, as affirmed above, into a magnetic field of intensity B_0 (directed along the z-axis) then the two states α and β present different energy level and hence a ΔE is revealed. Since

$$\Delta E = h\,\nu \qquad \textbf{Equation 1.1}$$

and also

$$\Delta E = (h\gamma/2\pi)\,B_0 \qquad \textbf{Equation 1.2}$$

where h is the Plank's constant and B_0 represents the magnetic field intensity, it is possible to induce transitions between the two states using an electromagnetic radiation with opportune frequency ν.

Furthermore, since h, γ and π are constant, ΔE is proportional to B_0. For this type of energetic gap, the involved frequencies fall into the radio-frequencies interval between 100 and 1000 MHz, depending on the magnetic field intensity. The fundamental NMR equation that relates an involved radiofrequency ν_1 with the magnetic filed intensity is:

$$\nu_1 = (\gamma/2\pi)\, B_0 \qquad \textbf{Equation 1.3}$$

For instance if B_0 = 1.4 Tesla (14000 Gauss) then a value of ν_1 = 60 MHz (corresponding to a radio frequency with λ = 5 m) is required to fulfill the equation.

The constant γ is called gyromagnetic ratio and it is a fundamental nuclear constant proportionally related with the magnetic moment μ and the spin quantum number I:

$$\gamma = 2\pi\, \mu/h\, I \qquad \textbf{Equation 1.4}$$

Unlike UV and IR spectroscopies, in NMR the value of ΔE is extremely small: for this reason the α state population (N_α) (energetically more stable and then more crowded in agreement with Boltzmann's distribution) has only an excess of about 0.005 % with respect to the β state population (N_β).

$$N_\beta / N_\alpha = e^{-\Delta E/kT} \qquad \textbf{Equation 1.5}$$

This determines that the NMR spectroscopy technique has a relatively low sensitivity with respect to UV and IR techniques. Nevertheless, this lack is more than compensated because of the plenty of information that is possible to collect in a single NMR spectrum.

Increasing the magnetic filed B_0 increases the difference between the two Boltzmann levels and hence the sensitivity of the NMR technique as well as it becomes more sensitive at lower temperatures (cryoprobes equipped spectrometers) where the S/N ratio is increased by a much lower noise (Figure 1.2).

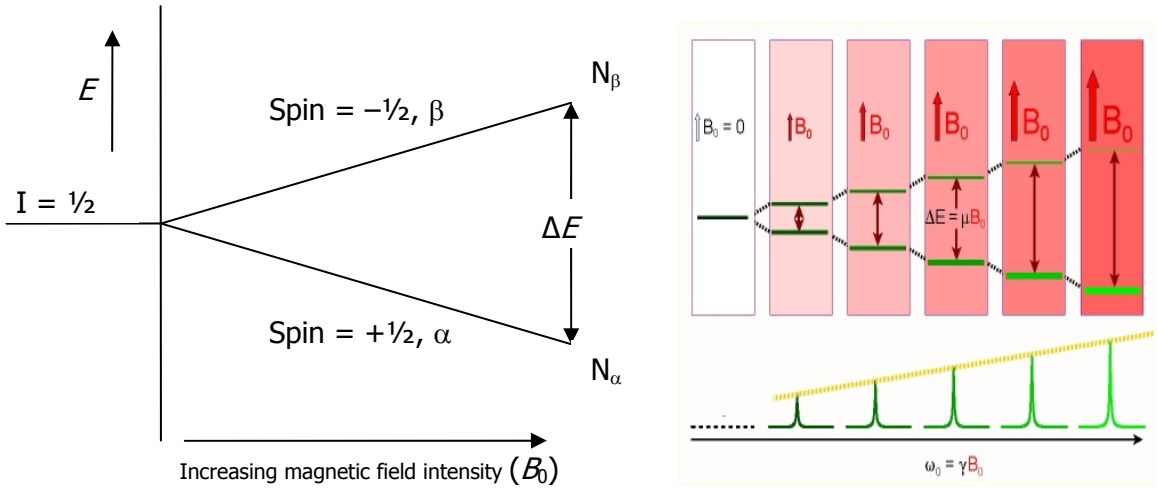


Figure 1.2: Difference between the two Boltzmann levels affecting the sensitivity of the NMR technique

If the nucleus is viewed as a rotating particle immersed in an external magnetic field, the magnetic axis of this nucleus precedes around the z-axis of the stationary magnetic field B_0 in the same way as a whipping-top precedes under the influence of the gravity (Figure 1.3).

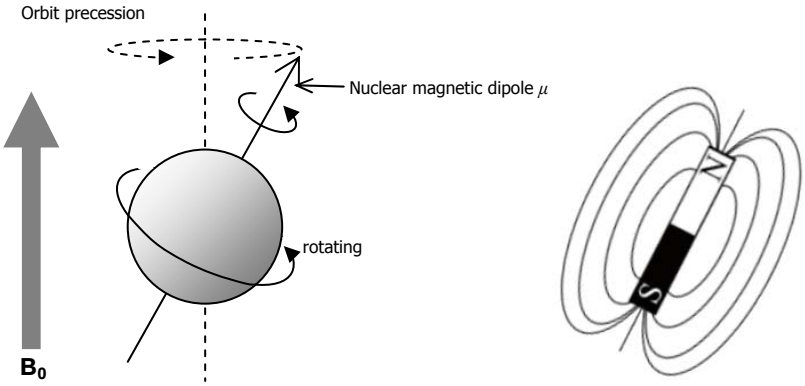


Figure 1.3: Precession of nuclear magnetic axis about the external magnetic field B_0 vector

An array of equivalent protons randomly preceding around z-axis produces a net macroscopic magnetization M_0 along this axis but not in the xy plane (Figure 1.4). When the provided radiofrequency ν_1 is equal to precession frequency of the equivalent protons (Proton Larmor Frequency ν_L, in MHz), the magnetic resonance is achieved and the NMR fundamental equation can be rewritten as:

$$\nu_L = \nu_1 = (\gamma/2\pi)\, B_0 \qquad \textbf{Equation 1.6}$$

This equation can be applied to an isolated array of protons. The aim of a radiofrequency pulse is to tilt the net magnetization vector onto the horizontal plane xy of the reference Cartesian system by which it is possible to measure the resulting magnetization component in such a plane.

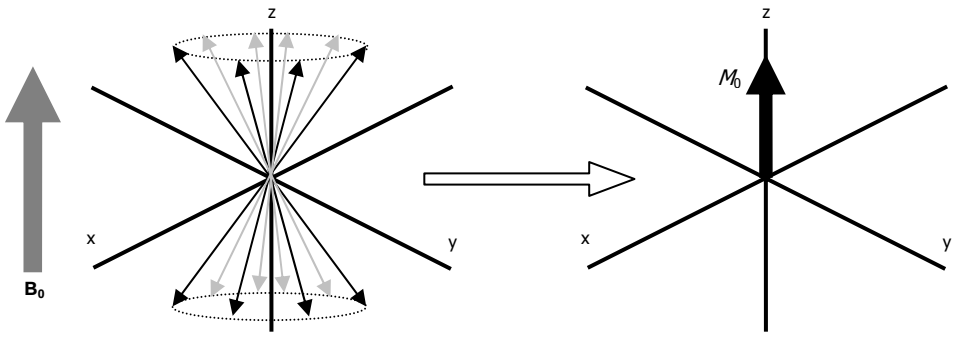


Figure 1.4: Macroscopic magnetization M_0 along the z-axis

If the sample is irradiated along the x-axis with a radiofrequency pulse (oscillating magnetic field B_1) containing the Larmor's frequencies of the examined nuclei (i.e. ^1H), then these nuclei will adsorb the associated energy with a consequent spin state transition. At a macroscopic level it's possible to observe that the net macroscopic magnetization vector M_0 rotate from the z-axis toward the xy horizontal plane, starting a precession rotation around the z-axis (M). Since the radiofrequency is not further provided to the system it will return the adsorbed energy and a receiving coil, disposed along the xy horizontal plane, will start to measure the oscillation of the y component of M. Because of relaxation phenomena, the adsorbed energy of an atom is transferred to the other surrounding atoms and the M vector, with a decreasing spiral of precession around the z-axis, slowly return to the initial position M_0 where the y component is equal to zero (Figure 1.5).

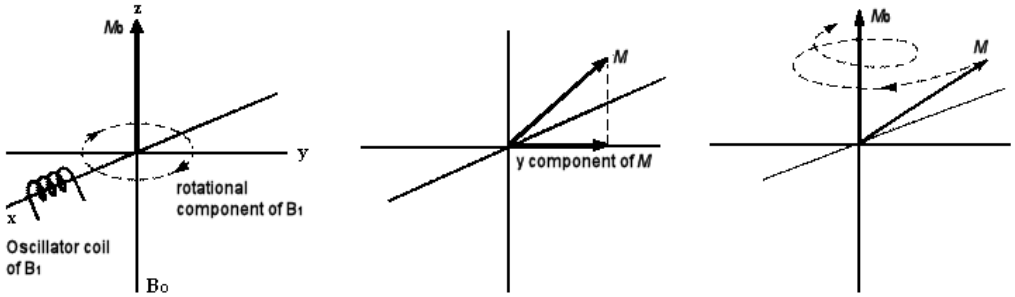


Figure 1.5: Perturbation of the macroscopic magnetization M_0 after a radiofrequency pulse and subsequent relaxation

The duration of the radiofrequency pulse, t(p), has to be accurately determined in order to obtain a strong NMR signal. If t(p) is strong enough to determine a 90° tilt of the vector M_0 then the maximum value of the y component will be achieved. Normally t(p) is of the order of μs. The signal acquired is an electromagnetic radiation with frequency ν, the Larmor's frequency of the nucleus examined, that tends to decay in the time (Figure 1.6). This characteristic oscillatory decay is known as the Free Induction Decay (FID).

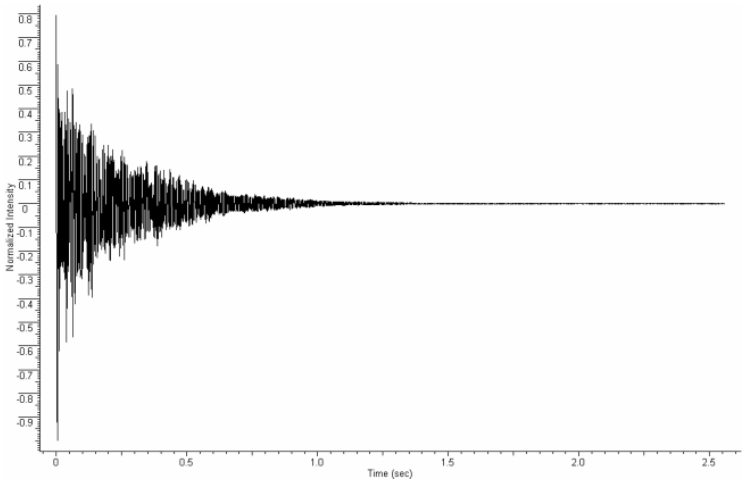


Figure 1.6: Oscillatory decay of the NMR signal known as Free Induction Decay (FID)

Successively, thanks to the calculation power of modern computers, the mathematical function known as Fourier transformation converts in real time this time-dependent pattern into a frequency-dependent pattern of nuclear resonances giving origin to characteristic nuclear magnetic resonance spectrum in which every different investigated nucleus presents its own signal (Figure 1.7).

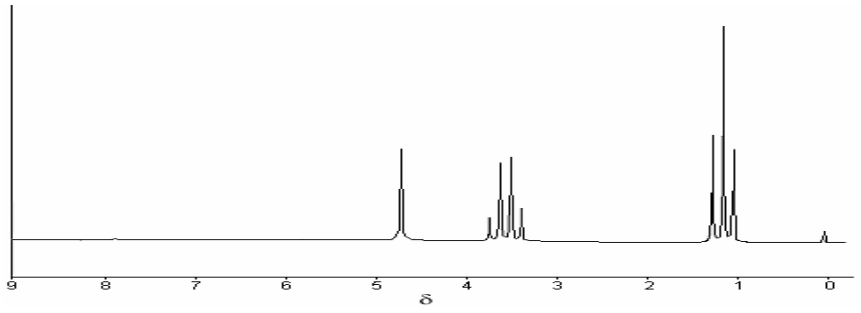


Figure 1.7: Fourier transformation of the FID signal: the spectrum of pure ethanol

Each magnetic nucleus that absorbs this radio-frequency energy will then radiate it back at a very specific frequency originating a specific signal in the NMR spectrum in agreement with the "chemical shift" theory.

1.3.4.2 Chemical shift

In molecules nuclei appear inside atoms and they are surrounded by electrons that react to B_0 giving rise to an additional induced magnetic field (B_{ind}) opposite to B_0. Each nucleus senses an actual magnetic field equal to $B_0 - B_{ind}$ or, as it is usually expressed, $(1-\sigma)B_0$, where σ is called the "shielding constant". Since the value of the shielding constant depends largely on the variation in electron density in the neighborhood of each nucleus, NMR-active atoms placed in different molecular locations may thus attain resonance at a slightly different frequency as compared to "naked" nuclei of the same kind. This phenomenon is at the basis of the great success of NMR because the resonance frequency of a nucleus is characteristic of a specific chemical environment in which it is located. For example, an electron withdrawing group decreases the electron density (hence the shielding constant) at a nearby nucleus. This "deshielded" nucleus resonates at an higher frequency than the one necessary for the same nucleus in a not electron withdrawing environment. Likewise, a proton bound to an sp^2 carbon is expected to be more deshielded than a proton bound to an sp^3 carbon, owing to the increased electronegativity of the former carbon resulting from bond orbitals hybridization with an higher s character.

However, a problem arises because the resonance frequency also depends linearly on the applied B_0. For this reason a "chemical shift" (σ) scale in units of part per million (ppm) is used in which the position of each line is given by difference (measured in Hz) from the resonance line of an internal standard divided by the operating frequency of the spectrometer expressed in MHz.

The chemical shift scale is independent of the external magnetic field and it is thus convenient for comparing data obtained with different spectrometers. The reference chemical shift standard (σ = 0 ppm) for ^1H and ^{13}C NMR spectroscopy is tetramethysilane (TMS), $(CH_3)_4Si$. TMS has been chosen because it is unreactive, readily soluble in most common solvents and easily removed owing to its low boiling point (28 °C). More importantly, it yields an NMR line that does not overlap with the NMR signals of most organic compounds. In fact, since silicon is less electronegative than carbon, it produces an electron

density increase at carbon and hydrogen and makes both nuclei more shielded than any other common organic compound. As a result, the ^{1}H and ^{13}C TMS lines appear at one edge of the corresponding NMR spectrum. A chart summarizing the chemical shifts of carbon and protons associated with the most common functional groups is provided in Figure 1.8.

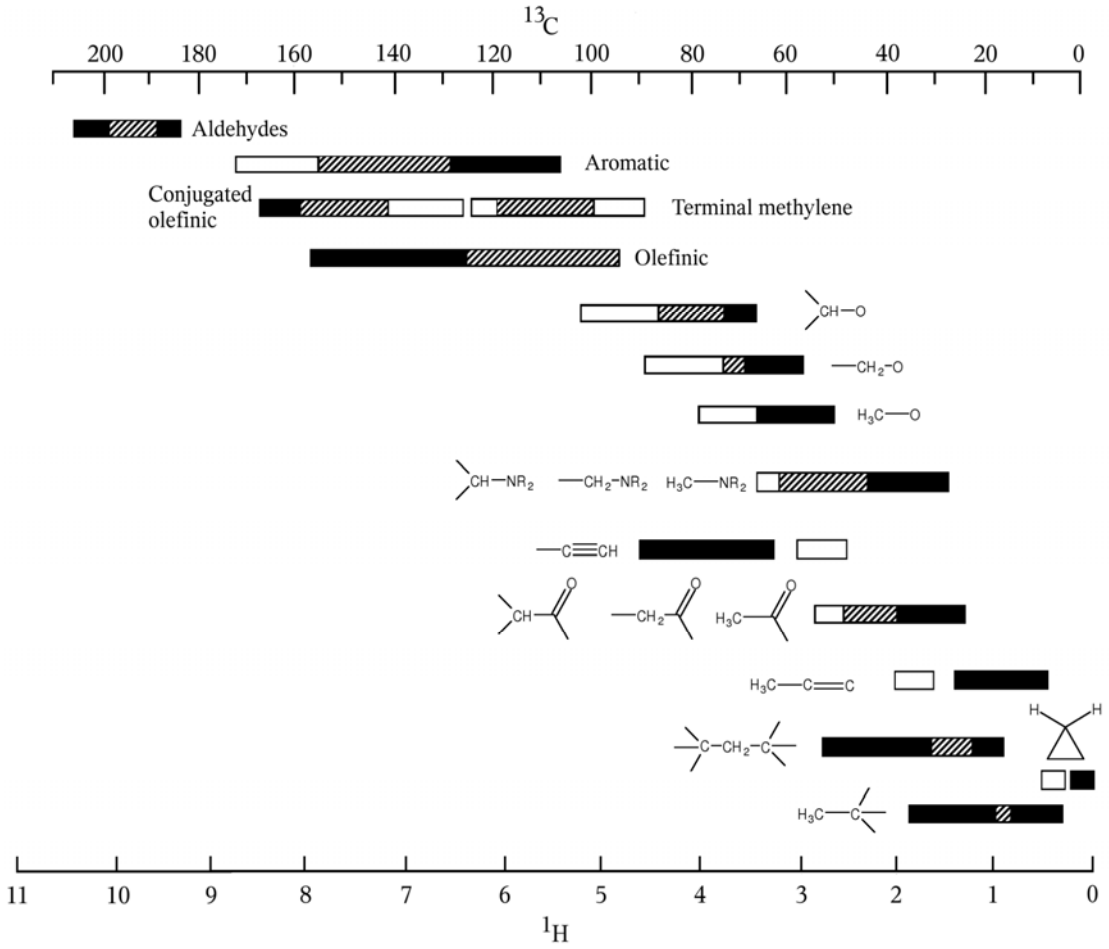


Figure 1.8: 1H and 13C chemical shift ranges for common proton (white fill) and carbon (black fill) environments in organic molecules. Hatched fill indicates overlap. [138]

1.3.4.3 Tuning, Locking and Shimming

Matching and tuning are operations that must be performed before the FID acquisition in order to optimize the instrumental conditions to the sample characteristics. Specifically, since the probe receives the sample within wires working like an antenna and radio-frequency waves cannot be efficiently sent over regular wires, the main effect is that the energy transfer requires pulses so long that become selective and no more able to cover all the spectral width. For this reason it is necessary to change the impedance of wires in order to prevent signal losses due to bad RF power transfer by matching and tuning.

Locking is a step, in the acquisition of spectra, which permits to keep the radiofrequency of the transmitter locked (referenced) to the ^{2}H frequency of a certain deuterium nucleus belonging to an abundant species, i.e. the solvent. When it is locked on that frequency, the spectrum peaks are not shifting any more during the data acquisition time, while magnet field is always drifting. Another reason of locking is that shimming is usually based on the lock signal.

Shimming is the operation that improves the magnetic field homogeneity in the sample by changing the current intensity in supplemental wires that function as magnetic lenses, thus aligning the force lines in the part of the sample comprised within the receiver coil. When the magnetic field homogeneity is optimized, it will be possible to get sharp and symmetric peaks. In order to better explain the effects of such operations on the quality of the NMR spectra three ^{1}H-NMR spectra have been recorded on the same sample of grape juice at 400 MHz, changing the instrumental conditions. The spectrum relative to the best instrumental conditions is shown in Figure 1.9. Only a restricted region of the spectra is shown to better pick the details of such an influence.

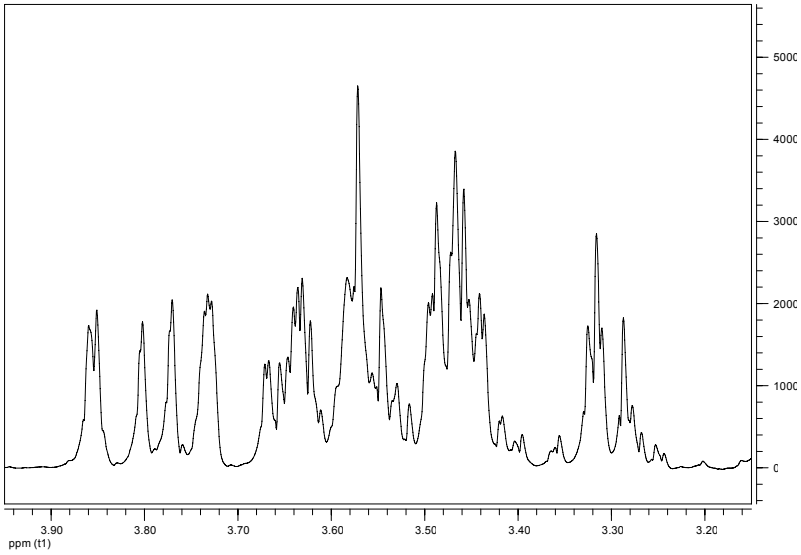


Figure 1.9: Sugars' region of the 1H-NMR spectrum of grape juice recorded in optimized shimming conditions.

A badly shimmed magnetic field produces the dramatic effect that is possible to observe in Figure 1.10.

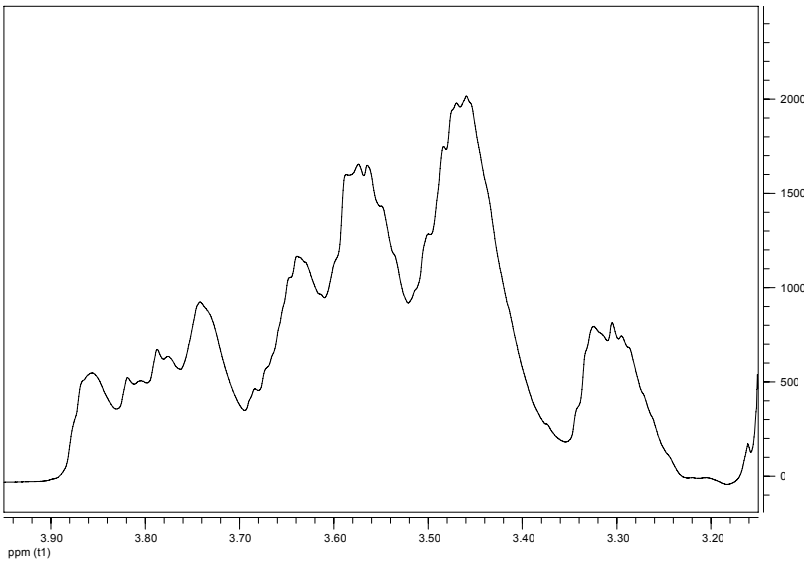


Figure 1.10: Sugars' region of the 1H-NMR spectrum of grape juice recorded in bad shimming conditions.

In this situation, since the frequency is tuned, signals are intense but very broad, determining a certain loss of the spectral features. A well shimmed magnetic field coupled with a non tuned frequency produces a spectra having a quite good resolution but showing a very low signal to noise rate. In such a way signals of the substances present in the smallest concentration inside the food sample are buried by the noise, thus lost. The absolute intensities of the signals are very low as it is possible to observe in Figure 1.11.

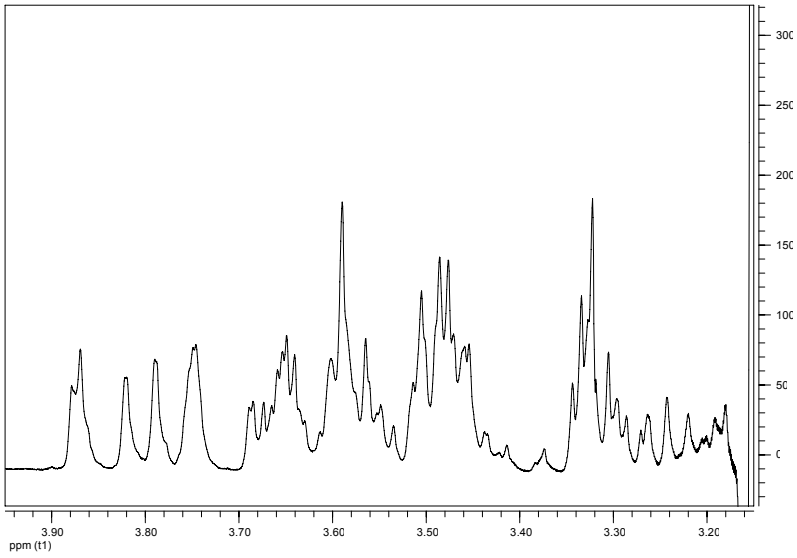


Figure 1.11: Sugars' region of the 1H-NMR spectrum of grape juice recorded in bad tuning conditions.

1.3.4.4 Fine structure in a NMR spectrum

In a hydrogen NMR spectrum, the presence of resonances explains first that the molecule of study contains hydrogen. Second, the number of signals in the spectrum shows how many different positions there are on the molecule to which hydrogen is attached. The frequency of a particular resonance in the NMR spectrum is referred to as the chemical shift. This is the most important measurable part of the NMR spectrum and contains information about the environment of each hydrogen atom and the structure of the compound under study. The third bit of information that an NMR spectrum provides is the ratio of the areas of the different bands, thus explaining the relative number of hydrogen atoms that exist at each position on a given molecule. This ratio is direct evidence of the structure of molecular structure and must correspond absolutely to any proposed structure before that structure may be considered correct.

Finally, the complex structure of the bands may contain information about the distance that separate the various hydrogen atoms through covalent bonds and the spatial arrangement of the hydrogen atoms attached to the molecule, including conformational diversity.

Normally this type of information is clear and understandable only for pure substances or for simple mixtures of them.

But, when the examined sample is a complex matrix, such as biofluids or food matrices, hundreds of molecules, and then thousands of protons, can participate to the final spectrum shape with the consequence of an unavoidable overlap of signals that makes the complex interpretation of the 1D-NMR spectrum not always (rarely) achievable. Therefore, the information associated would be unusable and the advantages of this spectroscopic technique would be precluded. Is there a solution for this type of cases? Fortunately in the last century mathematical and statistical tool have been developed in order to make easier the life for NMR scientists and help them in understanding the meaning of their too complex data. First of all a simplification of the data collected is required in order to better understand the information here contained, as well as an analysis of the data that take in account every variable in a multi-way manner. This is exactly what Chemometrics do!

In order to better understand how chemometrics works it is necessary an introduction to its statistical tools.

1.4 CHEMOMETRICS

In 1974 Svante Wold defined chemometrics as "The art of extracting chemically relevant information from data produced in chemical experiments…//...in analogy with biometrics, econometrics, etc." [139].

Chemometrics, like other "-metrics" techniques, make use of different kinds of mathematical models (high information models, ad hoc models, and analogy models) and requires knowledge of statistics, numerical analysis, operation analysis etc. In other words chemometrics is the field of extracting information from multivariate chemical data using tools of statistics and mathematics. However, the main problem for Wold is not mathematical but how to organize chemical data in a form that can be related to mathematical models. Following this criterion Wold, after 20 years, was not able to update the definition of chemometrics otherwise that: "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" [139].

It is relatively easy to generate a good deal of data in a short time by proper use of spectroscopies, but, as it has already explained, it is not always straightforward getting useful results from a set of spectral data. Determining the amounts of the components of a mixture can often be problematic without a prior separation step because of the overlap of spectral responses. Identifying the components of a mixture can also be challenging because of the similarity of many spectral responses. Initially, the solution to these problems has been that to increase spectral resolution or, as in the case of the quantitative analysis, to enhance the spectral resolution by means of a prior separation step. Many of these spectroscopic fixes, adopted in order to solve the problem of extracting understandable and significant results from data, work worst than one might expect, given only an apparent "information" in a spectral scan. Furthermore, huge amounts of data generate by spectroscopic measurements, often result redundant. Indeed, since the chemical and physical basis for the spectroscopic transition(s) observed are not perfectly unique to a single species and for a very isolated set of energies, the data generated have often a high correlation.

For this reasons, in the last decades, spectroscopists have increasingly turned to chemometrics for finding an effective help in dealing with spectral data. In fact, chemometric methods are efficient at extracting unique information from multichannel redundant data such as spectra. Moreover, these methods presume serial correlation that can be found in

spectra and are designed to use it to improve the precision of any estimation made from the data.

1.4.1 Preprocessing of raw data

Nuclear magnetic resonance analysis of complex samples such as foodstuff, produces data that frequently have to be arranged before their successive analysis. The original FIDs, representing NMR raw data, must be at first processed trough Fourier transformation, accurate spectral phasing and baseline correction, exact chemical shift referencing and scaling, in order to be considered suitable for statistical analysis. The differences on the spectral shape, due to tuning and shimming errors, must be checked and corrected, in the same way as the FID processing must not introduce any unwanted variance in the transformed spectra. Nevertheless, even when a correct FID processing is performed, still an eventual abscissa residual variance may produce unaligned spectra more difficult to be interpreted [140]. For example, temperature variations or differences of relative concentration in the background matrix of the sample may cause these variations [141]. The misalignment of spectral signals in different samples may be also associated to chemical shift variations of the signal assigned to a given nucleus belonging to the same substance but experiencing a different chemical environment, e.g. solvent polarity, ionic strength or different pH. In fact, molecules having protons sensitive to pH produce different ^{1}H-NMR chemical shifts of such protons depending on their ionization state. Other environmental effects can also influence chemical shifts, including metal ions concentrations, metabolite-protein binding, and chemical exchange phenomena. Obviously, any research study, based on such affected NMR data, may present missing and/or false information, among the true ones, and therefore the information inferred can lead to erroneous conclusions. For these reasons, chemometric tools able to preprocess data for making them suitable for successive multivariate analysis steps, have increasingly developed and utilized in the last years.

1.4.1.1 Data Alignment

"Binning" is, so far, one of the most common method, able to make a spectral dataset more homogeneous. It involves a data reduction performed through NMR signals integration, within standardized spectral regions whose width arbitrary ranges between 0.01 and 0.05 ppm [142, 143]. This operation results in a decrement of the errors due to either unwanted shifts and changes of the line shape of peaks. On the other hand, binning may hide significant variations in the concentration of low components when their signals overlap those of other molecules present at much higher concentrations. The effect is expressed as a loss of spectral resolution that often can results a drawback. In effect, the interpretation of derived chemometric models in terms of identified metabolic biomarkers, can not easily be obtained from the reduced data, but requires reexamination of the real NMR spectra to identify such metabolites responsible for the observed patterns. For this reason, when high definition is required in order to keep every useful spectroscopic information, it would be preferable to avoid binning and looking for some other chemometric tools. For example Needle Representation (NR) keeps all the information related to the true signals, but reducing the total amount of variables to submit to following (if any) improving algorithms [144]. An alternative, more complex approach is that of using automatic peak alignment algorithms, able to resolve the problem of signal position variation in ^{1}H-NMR spectra and allowing the use of the full spectral resolution for pattern recognition. For example, Stoyanova et al. [145] removed the signals' positional noise using Principal Component Analysis (see Section 1.4.2.1) to determine the misalignment across a series of biofluid NMR spectra. Furthermore, methods involving the application of a genetic algorithm to align segments of spectra have also been used, [146, 147]. Other examples of peak alignment methods are reported in literature such as that using reduced set mapping (PARS) [144, 148, 149], dynamic time warping [150, 151], correlation-optimized warping [152, 153], partial linear fit [154] and a method by Johnson et al. [155] again based on PCA results.

1.4.1.2 Data Normalization

Normalization (vertical scaling of data) is another crucial preprocessing step for all the "-omics" approaches [156]. It mainly corrects dilution errors occurring among samples but also some other due to instrumental variables, such as tuning of the instrument or radio-

frequency power settings. In addition, also a variable content of an unwanted species (e.g. water) on the samples can determine such a type of errors. Even if the use of an internal standard can often help to solves this problem, not all the time this can be used or, if used, can be not effective in solving the problem.
For this reason, a large variety of algorithms, able to perform a normalization of NMR data, have been published and, even if the "total integral to a Constant Sum (CS)" represent de facto the standard for most of recent studies [156], some other methods, such as "probabilistic quotient normalization" [157], demonstrate that normalization is context dependent and CS is not always the best solution.

1.4.1.3 Data Centering and Scaling

In many kinds of multivariate data analysis methods, it is common practice to perform preprocessing steps such as scaling and centering transformations [158, 159]. Without this preprocessing step, variables expressed by higher number or defined in a large range, would have more weight than variables expressed by smaller number or defined in a narrow range. To avoid this situation it is suitable to uniform all the variables in order to have the same weight *a priori*. The most common method are i) Centering that consists in subtracting the mean value from each variable (Equation 1.7)

$$Y_i = X_i - \overline{X} \qquad \textbf{Equation 1.7}$$

and ii) Scaling that is performed by dividing each centered variable by its standard deviation (Equation 1.8).

$$Y_i = \frac{X_i - \overline{X}}{\sqrt{\frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}{N-1}}} \qquad \textbf{Equation 1.8}$$

This can assist methods, such as PCA, that are sensitive to scale, by giving to each variable an equal chance of influencing the models parameters.

All the previous listed methods differ in their theoretical approach and computational complexity but they all have the same purpose to pre-process spectral data in order to make the successive statistical analysis more reliable and meaningful.

1.4.2 Principal Chemometric Tools

Chemometrics is typically used for one or more of three primary purposes:

- To explore patterns of association in data;
- To track properties of materials on a continuous basis;
- 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 and manageable form. Such a chemometric analysis can expose possible outliers and indicate whether there are patterns or trends in the data. Algorithms such as Principal Component Analysis (PCA) (see Section 1.4.2.1) and Hierarchical Cluster Analysis (HCA) are designed to reduce large and complex data sets into a series of optimized and interpretable views. These views emphasize the natural groupings of the data and show which variables most strongly influence those patterns [160].

In many applications, it is expensive, time consuming or difficult to directly measure a property of interest. Such cases require the analyst to predict something of interest based on related properties that are easier to measure. The goal of chemometric regression analysis is to develop a calibration model which correlates the information in the set of known measurements to the desired property. Chemometric algorithms for performing regression include Partial Least Squares (PLS) and Principal Component Regression (PCR) and are designed to avoid problems associated with noise and correlations in the data. Because the utilized regression algorithms are based in Factor Analysis (FA), the whole group of known measurements is considered simultaneously, and information about correlations among the variables is automatically built into the calibration model. Chemometric regression lends itself to the on-line monitoring and process control industry, where fast and inexpensive systems are needed to test, predict and take decisions about product quality.

Furthermore, many applications may require that samples be assigned to predefined categories, or "classes". This may involve determining whether a sample is good or bad, or predicting an unknown sample as belonging to one of several distinct groups. A classification model is used to predict a sample's class by comparing the sample to a previously analyzed experience set, in which categories are already known. K-Nearest Neighbor (KNN), Soft Independent Modeling of Alass Analogy (SIMCA) and Linear Discriminant Analysis (LDA) (see Section 1.4.2.2) are main chemometric workhorses. When these techniques are used to create a classification model, the answers provided are more reliable and include the ability

to reveal anomalous samples in the data. In all kind of classification methods is of critical importance the distance measure for data value. The most commonly used distance measures are the Euclidean Distance (ED) and the Mahalanobis Distance (MD) (see Section 1.4.2.3). The latter is a very useful way of determining the "similarity" of a set of values from an "unknown" sample to a set of values measured from a collection of "known" samples.

1.4.2.1 Principal Component Analysis (PCA)

Principal Components Analysis (PCA) is a multivariate procedure which rotates the data such that maximum variabilities are projected onto the axes [161]. Essentially, a set of correlated variables are transformed into a set of uncorrelated variables which are ordered by reducing variability. The uncorrelated variables are linear combinations of the original variables, and the last of these variables can be removed with minimum loss of real data.

The main use of PCA is to reduce the dimensionality of a data set while retaining as much information as is possible. It computes a compact and optimal description of the data set.

The first principal component is the combination of variables that explains the greatest amount of variation. The second principal component defines the next largest amount of variation and is independent to the first principal component. There can be as many possible principal components as there are variables.

What PCA performs, can be viewed as a rotation of the existing axes to new positions in the space defined by the original variables. In this new rotation, there will be no correlation between the new variables defined by the rotation. The first new variable contains the maximum amount of variation, the second new variable contains the maximum amount of variation unexplained by the first and orthogonal to the first, etc.

Principal component analysis is based on the statistical representation of a random variable. Supposing a random vector population x, where:

$$x = (x_1, ..., x_n)^T \qquad \textbf{Equation 1.9}$$

and the mean of that population is denoted by:

$$\mu_x = E\{x\} \qquad \textbf{Equation 1.10}$$

and the covariance matrix of the same data set is:

$$C_x = E\{(x-\mu_x)(x-\mu_x)^T\} \qquad \textbf{Equation 1.11}$$

The components of C_x, denoted by c_{ij}, represent the covariances between the random variable components x_i and x_j. The component c_{ii} is the variance of the component x_i. The variance of a component indicates the spread of the component values around its mean value. If two components x_i and x_i of the data are uncorrelated, their covariance is zero $(c_{ij} = c_{ji} = 0)$. The covariance matrix is, by definition, always symmetric.

From a sample of vectors $x_1,.....,x_M$ it is possible to calculate the sample mean and the sample covariance matrix as the estimates of the mean and the covariance matrix. From a symmetric matrix such as the covariance matrix, it is calculated an orthogonal basis by finding its eigenvalues and eigenvectors. The eigenvectors e_i and the corresponding eigenvalues λ_i are the solutions of the equation:

$$C_x e_i = \lambda_i e_i, i = 1,....n \qquad \textbf{Equation 1.12}$$

For simplicity it is assumed that the λ_i are distinct. These values can be found, for example, by finding the solutions of the characteristic equation:

$$\left|C_x - \lambda I\right| = 0 \qquad \textbf{Equation 1.13}$$

where the I is the identity matrix having the same order than C_x and the $\left|\ \right|$ denotes the determinant of the matrix. If the data vector has *n* components, the characteristic equation becomes of order *n*.

This is easy to solve only if *n* is small. Solving eigenvalues and corresponding eigenvectors is a non-trivial task, and many methods exist. One way to solve the eigenvalue problem is to use a neural solution to the problem [162]. The data is fed as the input, and the network converges to the wanted solution. By ordering the eigenvectors in the order of descending eigenvalues (largest first), one can create an ordered orthogonal basis with the first

eigenvector having the direction of largest variance of the data. In this way, it is possible to find directions in which the data set has the most significant amounts of energy.

Suppose one has a data set of which the sample mean and the covariance matrix have been calculated. Let A be a matrix consisting of eigenvectors of the covariance matrix as the row vectors.

By transforming a data vector x, it is obtained:

$$y = A(x - \mu_x) \qquad \textbf{Equation 1.14}$$

which is a point in the orthogonal coordinate system defined by the eigenvectors. Components of y can be seen as the coordinates in the orthogonal base. It is possible to reconstruct the original data vector x from y by:

$$x = A^T y + \mu_x \qquad \textbf{Equation 1.15}$$

using the property of an orthogonal matrix $A^{-1} = A^T$. The A^T is the transpose of a matrix A. The original vector x was projected on the coordinate axes defined by the orthogonal basis. The original vector was then reconstructed by a linear combination of the orthogonal basis vectors.

Instead of using all the eigenvectors of the covariance matrix, it is possible to represent the data in terms of only a few basis vectors of the orthogonal basis. If the matrix having the K first eigenvectors is denotet as rows by A_K, a similar transformation can be create as seen above:

$$y = A_K(x - \mu_x) \qquad \textbf{Equation 1.16}$$

and

$$x = A_K{}^T y + \mu_x \qquad \textbf{Equation 1.17}$$

This means that it projects the original data vector on the coordinate axes having the dimension *K* and transforms the vector back by a linear combination of the basis vectors. This minimizes the mean-square error between the data and this representation with given number of eigenvectors. If the data is concentrated in a linear subspace, this provides a way to compress data without losing much information and simplifying the representation. By

picking the eigenvectors having the largest eigenvalues it is lost as little information as possible in the mean-square sense.
One can e.g. choose a fixed number of eigenvectors and their respective eigenvalues and get a consistent representation, or abstraction of the data. This preserves a varying amount of energy of the original data. Alternatively, approximately the same amount of energy can be chooses together with varying amount of eigenvectors and their respective eigenvalues. This would in turn give approximately consistent amount of information in the expense of varying representations with regard to the dimension of the subspace.
At this point one is here faced with contradictory goals: on one hand, one should simplify the problem by reducing the dimension of the representation. On the other hand, one want to preserve as much as possible of the original information content. PCA offers a convenient way to control the trade-off between loosing information and simplifying the problem at hand. Then the question becomes: how many factors is wanted to be extracted? Note that as consecutive factors are extracted, they account for less and less variability. The decision of when to stop extracting factors depends on when there is only very little "random" variability left. The nature of this decision is arbitrary; however, various guidelines have been developed [163]. The two most useful methods are the Kaiser criterion and the scree test.

The Kaiser criterion retains only factors with eigenvalues greater than 1. In essence this is like saying that, unless a factor extracts at least as much as the equivalent of one original variable, it is dropped. This criterion was proposed by Kaiser [164], and is probably the one most widely used.

The scree test is a graphical method, first proposed by Cattell [165], suggesting to find the place where the smooth decrease of eigenvalues appears to level off to the right of the plot. To the right of this point, presumably, one finds only "factorial scree" ("scree" is the geological term referring to the debris, which collects on the lower part of a rocky slope).
Both criteria have been studied in detail [166-169]. Theoretically, one can evaluate those criteria by generating random data based on a particular number of factors. One can then see whether the number of factors is accurately detected by those criteria. Using this general technique, the first method (Kaiser criterion) sometimes retains too many factors, while the second technique (scree test) sometimes retains too few; however, both performed quite well under normal conditions, that is, when there are relatively few factors and many cases.

In practice, an additional important aspect is the extent to which a solution is interpretable. Therefore, one usually examines several solutions with more or fewer factors, and chooses the one that makes the best "sense".

1.4.2.2 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) easily handles the case where the within-class frequencies are unequal and their performances have been examined on randomly generated test data. This method maximizes the ratio of between-class variance to the within-class variance in any particular data set thereby guaranteeing maximal separability.
In general, an object is assigned to one of a number of predetermined groups based on observations made on the object. Note that the groups are known or predetermined and do not have order (i.e. nominal scale).
LDA is also closely related to PCA in that both look for linear combinations of variables which best explain the data. LDA explicitly attempts to model the difference between the classes of data. PCA on the other hand does not take into account any difference in class [170].
LDA works when the measurements made on each observation are continuous quantities. When dealing with categorical variables, the equivalent technique is called Discriminant Correspondence Analysis (LDC) [171].

LDA for two classes considers a set of observations x (also called features, attributes, variables or measurements) for each sample of an object or event with known class y. This set of samples is called the training set. The classification problem is then to find a good predictor for the class y of any sample of the same distribution (not necessarily from the training set) given only an observation x.
LDA approaches the problem by assuming that the probability density functions $p(\vec{x}|y=1)$ and $p(\vec{x}|y=0)$ are both normally distributed, with identical full-rank covariances

$$\Sigma_{y=0} = \Sigma_{y=1} = \Sigma \qquad \textbf{Equation 1.17}$$

It can be shown that the required probability $p(y|\vec{x}=0)$ depends only on the dot product

$$\vec{\omega} \cdot \vec{x} \text{ where } \vec{\omega} = \Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0) \qquad \textbf{Equations 1.18 and 1.19}$$

That is, the probability of an input x being in a class y is purely a function of this linear combination of the known observations.

A similar analysis that allows the covariances to be different is called quadratic discriminant analysis (QDA) [172]

The terms Fisher's linear discriminant and LDA are often used interchangeably, although Fisher's original article "The Use of Multiple Measures in Taxonomic Problems" actually describes a slightly different discriminant, which does not make some of the assumptions of LDA such as normally distributed classes or equal class covariances [173].

Supposing two classes of observations having means $\vec{\mu}_y = 0, \vec{\mu}_y = 1$ and covariances $\Sigma_{y=0}, \Sigma_{=1}$ the linear combination of features $\vec{\omega} \cdot \vec{x}$ will have means $\vec{\omega} \cdot \vec{\mu}_{y=i}$ and variances $\vec{\omega}^T \Sigma_{y=i} \vec{\omega}$ for i = 0, 1. Fisher defined the separation between these two distributions to be the ratio of the variance between the classes to the variance within the classes:

$$S = \frac{\sigma^2_{between}}{\sigma^2_{within}} = \frac{(\vec{\omega} \cdot \vec{\mu}_{y=1} - \vec{\omega} \cdot \vec{\mu}_{y=0})^2}{\vec{\omega}^T \Sigma_{y=1} \vec{\omega} + \vec{\omega}^T \Sigma_{y=0} \vec{\omega}} = \frac{(\vec{\omega} \cdot (\vec{\mu}_{y=1} - \vec{\mu}_{y=0}))^2}{\vec{\omega}^T (\Sigma_{y=0} + \Sigma_{y=1}) \vec{\omega}} \qquad \textbf{Equation 1.20}$$

This measure is, in some sense, a measure of the signal-to-noise ratio for the class labeling. It can be shown that the maximum separation occurs when

$$\vec{\omega} = (\Sigma_{=0} + \Sigma_{=1})^{-1} (\vec{\mu}_{y=1} - \vec{\mu}_{y=0}) \qquad \textbf{Equation 1.21}$$

When the assumptions of LDA are satisfied, the above equation is equivalent to LDA.

Multiclass LDA is an improved LD method that can deal with a multiple set of classes.

Supposing C classes having means $\vec{\mu}_i$ and covariance matrices Σ_i with i = 1,...,C the decision boundaries of these C classes if given by the quadratic terms

$$(\vec{x} - \vec{\mu}_i)^T \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i) + ln|\Sigma_i| = (\vec{x} - \vec{\mu}_j)^T \Sigma_j^{-1} (\vec{x} - \vec{\mu}_j) + ln|\Sigma_J| \qquad \textbf{Equation 1.22}$$

$\vec{x}$ is the vector of all samples $\vec{\mu}_i$ is a vector with the mean values of each feature

When the class distributions share the same covariance matrix, $\Sigma_i = \Sigma_j, \forall i \neq j$, the equations 1.22 results in linear boundaries. This led this extension of the work pioneered by

Fisher (which was originally only defined for the 2 class problem) to be known as Linear Discriminant Analysis.

Under the assumption of equal covariance matrices (also known as homoscedastic), LDA provides those C-1 features where the Bayes error is minimized. This operation can be shown to reduce to the following eigenvalue decomposition problem:

$$S_B V = \overline{\Sigma} \; V\Lambda \qquad \textbf{Equation 1.23}$$

where S_B is the between-class scatter matrix and $\overline{\Sigma}$ is the average over all Σ_i

LDA is not guarantee however to find the best optimal solution for a set of less than C-1 features (even if the data is homoscedastic), and is very sensitive to the number of samples used. A common way to improve the accuracy and robustness of this approach is to add a regularization term to the estimate of $\overline{\Sigma}$. In general this can be written as $\overline{\Sigma} + \varepsilon \mathbf{I}$, where $\mathbf{I}$ is the identity matrix and ε is small. Other generalizations of LDA have been defined to address the more general problem of heteroscedastic distributions (i.e., where the data distributions are not homoscedastic). These are usually called Heteroscedastic LDA. Alternatively, one can use QDA. In practice, the class means and covariances are not known. They can, however, be estimated from the training set. Either the maximum likelihood estimate or the maximum a posteriori estimate may be used in place of the exact value in the above equations. Although the estimates of the covariance may be considered optimal in some sense, this does not mean that the resulting discriminant obtained by substituting these values is optimal in any sense, even if the assumption of normally distributed classes is correct. Another complication in applying LDA and Fisher's discriminant to real data occurs when the number of observations of each sample exceeds the number of samples (and this is the case of the study considered in the present thesis). In this case, the covariance estimates do not have full rank, and so cannot be inverted. There are a number of ways to deal with this. One is to use a pseudo inverse instead of the usual matrix inverse in the above formulae. Another, called regularized discriminant analysis, is to artificially increase the number of available samples by adding white noise to the existing samples.

These new samples do not actually have to be calculated, since their effect on the class covariances can be expressed mathematically as

$$C_{new} = C + \sigma 2\mathbf{I} \qquad \textbf{Equation 1.24}$$

where I is the identity matrix, and σ is the amount of noise added, called in this context the regularization parameter. The value of σ is usually chosen to give the best results on a cross-validation set. The new value of the covariance matrix is always invertible, and can be used in place of the original sample covariance in the above formulae.

LDA can be generalized to multiple discriminant analysis, where c becomes a categorical variable with N possible states, instead of only two. Analogously, if the class-conditional densities $p(\vec{x}|c=i)$ are normal with shared covariances, the sufficient statistic for $P(c|\vec{x})$ are the values of N projections, which are the subspace spanned by the N means, affine projected by the inverse covariance matrix. These projections can be found by solving a generalized eigenvalue problem, where the numerator is the covariance matrix formed by treating the means as the samples, and the denominator is the shared covariance matrix.

1.4.2.3 Leave One Out (LOO) cross-validation

Though a powerful method, LDA has some drawbacks. Since it is a supervised technique, LDA has a tendency to overfitting [174] in small-sample-size problems, where the dimensionality is higher than the sample size. To avoid overfitting problem cross-validation methods, frequently the leave-one-out test (LOO) is applied.

Leaving-one-out is an elegant and straightforward technique for estimating classifier error rates. For a given method and sample size, n, a classifier is generated using (n - l) cases and tested on the single remaining case. This is repeated n times, each time designing a classifier by leaving-one-out. Thus, each case in the sample is used as a test case, and each time nearly all the cases are used to design a classifier. The error rate is the number of errors on the single test cases divided by n. Evidence for the superiority of the leaving-one-out approach is well documented [175]. The leave-one-out error rate estimator is an almost unbiased estimator of the true error rate of a classifier. This means that over many different sample sets of size n, the leaving-one-out estimate will average out to the true error rate. Because the leave-one-out estimator is unbiased, for even small sample sizes of over 100, the estimate should be accurate. The great advantage of cross-validation is that all the cases in the available sample are used for testing, and almost all the cases are also used for training the classifier.

1.4.2.4 Mahalanobis Distances (MD)

Mahalanobis Distance (MD) is a measure of distance based on correlations between variables by which different patterns can be identified and analyzed. It is a useful way of determining *similarity* of an unknown set of cases to a known one. It differs from Euclidean Distance (ED) in that it takes into account the correlations of the data set and is scale-invariant, i.e. not dependent on the scale of measurements.

Formally, the Mahalanobis distance from a group of values with mean $\mu = (\mu_1,\mu_2,\mu_3,...,\mu_p)^T$ and covariance matrix Σ for a multivariate vector $x = (x_1,x_2,x_3,...,x_p)^T$ is defined as:

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} \qquad \textbf{Equation 1.25}$$

Mahalanobis distance can also be defined as dissimilarity measure between two random vectors $\vec{x}$ and $\vec{y}$ of the same distribution with the covariance matrix Σ :

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T \Sigma^{-1} (\vec{x}-\vec{y})^T} \qquad \textbf{Equation 1.26}$$

If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. Moreover, if the covariance matrix is diagonal, then the resulting distance measure is called the *normalized Euclidean distance*:

$$d(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{p} \frac{(x_i - y_i)^2}{\sigma_i^2}} \qquad \textbf{Equation 1.27}$$

where σ_i is the standard deviation of the x_i over the sample set

To better understand how Mahalanobis distance works, consider that, when using Euclidean distance, the set of points equidistant from a given location is a sphere, that is to assume that the sample points are distributed about the center of mass in a spherical manner. Intuitively, the closer a test point in *N*-dimensional Euclidean space is to this center of mass, the more likely it is to belong to the set. But were the distribution is non-spherical, for instance ellipsoidal, the probability of the test point belonging to the set to depend not only on the distance from the center of mass, but also on the direction. In those directions where the ellipsoid has a short axis the test point must be closer, while in those where the axis is long the test point can be further away from the center. Putting this on a mathematical

basis, the ellipsoid that best represents the set's probability distribution can be estimated by building the covariance matrix of the samples. The Mahalanobis distance is simply the distance of the test point from the center of mass divided by the width of the ellipsoid in the direction of the test point.

1.4.3 Chemometrics in food science

The detailed study of food composition and quality attributes is of paramount importance to accomplish efficient quality control and improve properties of foodstuff. High-resolution NMR and hyphenated NMR methods give a valuable contribution to that study, as shown by recent work [176-178]. In addition, multivariate analysis of spectroscopic data can provide rapid information about quality related factors such as food geographical origin, processing conditions, and reproducibility within different production sites.

Spectroscopic NMR methods provide, in a single experiment, relevant information on a wide range of compounds present in the food matrix, offering advantages in terms of simplicity of sample preparation and rapidity of analysis. The speed with which NMR spectra can be obtained, often under automation, enables examination of many samples as required for most food composition, authenticity and quality control applications [179].

Because the richness of information often results in high spectral complexity, it calls for the use of multivariate analysis to study large numbers of spectra and extract meaningful information.

The application of chemometrics to high-resolution NMR data has been applied in some instances to address different issues of food authenticity and origin. For example, promising results have been obtained concerning the classification of apple juices according to variety [180], the detection of adulterations in orange juice [181, 182], the discrimination of coffee samples differing in their manufacturing process [183], the differentiation of olive oil according to cultivar, botanical and geographical origin [111, 184-187], the differentiation and characterization of grape cultivars [188, 189], the characterization of wine geographical origin [190-192][, the quality control of beer [193], the authentication of dairy products [194] and the metabolite profiling of transgenic tomato fruit [195].

In most of cases, the data patterns obtained from a significant number of samples are used to develop a predicting model. The latter is able to predict the features of unknown samples evaluated *a posteriori*, by means of their own quality parameters.

1.5 A CASE STUDY OF CHEMOMETRICS APPLIED TO NMR DATA

The present research work constitutes a case study of application on NMR data, sustained by chemometric analysis, applied to the assessment of the geographical origin of a food product which is protected by the European laws with the Protected Geographical Indication, a collective mark that need to be objectively associated to quality parameters in order to be actually protected.

1.5.1 Collective marks

The market, thanks to the effects of the globalization and the close economic relationships existing among the various Countries, is being enriching of a great variety of products, very different for characteristics and price, among which it is more and more difficult to orient. Following this situation, it has emerged for many producers the necessity to characterize and to valorize their own products with a mark legally recognized.

Agro-food typical products, that represent a segment of Quality Food Products (QFPs), are an example of the new concept consumer's choice in terms of genuinity and authenticity in front of food massification. These recent tendencies about food safety aspects, and the need of re-discovering the true values of agriculture strictly connected with the territory, have leaded to the creation of quality certifications marks. The latter have became a strategic instrument of differentiation that confers to the food products a commercial added value [196].

The consumer can immediately identify products that respect protocolled quality parameters, trough quality logos clearly reported in its package.

European logos, identifying Protected Designation of Origin (PDO) and Protected Geographical Indication (PGI), have been developed and used for individuating those European food products having the corresponding characteristics (Figures 1.12 and 1.13). The aim of such logos is to promote and protect genuine food products to the other having similar organoleptic characteristics but not responding to the established quality parameters. PDO and PGI were introduced by the EEC Reg. 2081/92 which has been recently upgraded by EC Reg. 510/2006. Their definition is reported below as well their respective logo.

Figure 1.12: PDO logo

PDO (Protected Designation of Origin):

"*Designation of Origin* means the name of a region, a specific place or, in exceptional cases, a country, used to describe an agricultural product or a foodstuff originating in that region, specific place or country, and the quality or characteristics of which are essentially or exclusively due to a particular geographical environment with its inherent natural and human factors, and the production, processing and preparation of which take place in the defined geographical area" [EEC Reg. 2081/92, article 2.2.a].

Thus PDO covers the term used to describe foodstuffs which are produced, processed and prepared in a given geographical area using recognised know-how.

Figure 1.13: PGI logo

PGI (Protected Geographical Indication):

"*Geographical Indication* means the name of a region, a specific place or, in exceptional cases, a country, used to describe an agricultural product or a foodstuff originating in that region, specific place or country, and which possesses a specific quality, reputation or other characteristics attributable to that geographical origin and the production and/or processing

and/or preparation of which take place in the defined geographical area" [EEC Reg. 2081/92, article 2.2.a].

In the case of PGI, the geographical link must occur in at least one of the stages of production, processing or preparation. Furthermore, the product can benefit from a good reputation [197].

PGI and POD are defined collective marks, and they identify products obtained by all the firms that respect certain environmental and productive conditions and that voluntarily accept to be submitted to a system of control effected by independent organisms. The fundamental document, at the basis of an application for a PDO or for a PGI certification, is the Disciplinary of Production, that specifies the criterions that the producers must follow in a peremptory way for the obtainment of the protected food product. Therefore it is a real certification of quality that accompanies a product responding to a collective mark. The activity of control is carried out by designate public authority or by private organisms, authorized with decree of the Office of the Agricultural and Forest Politics [198]. It allows to guarantee that the protected products respond to the requisite of the Disciplinary of Production. Both PDO and PGI certifications share the same normative system and the same procedures for the application. They give the same guarantees to consumers and the same rights to producers. The difference between these two quality certifications methods depend on how closely the specifications of the quality of the food product are linked to the geographical area whose it bears the name. The PDO certification is meant for those products which show an objective and very close link between their features and the area of which they bear the name (including human and natural factors, such as climate, soil quality and local know–how); the PGI certification also designates products linked to the area of which they bear the name but with a more flexible objective link. [199].The meaning of the idea standing behind these quality certifications marks is summarized in these paragraph of the EC Reg. 510/2006: "In view of the wide variety of products marketed and the abundance of product information provided, the consumer should, in order to be able to make the best choices, be given clear and succinct information regarding the product origin" [EC Reg. 510/2006].

The recognition of the greater quality grade of a product that these labels furnish it is accompanied to a great economic added value and therefore it makes the verification of the truthfulness of the declared characteristics necessary. Given the actual presence in the market of a great number of products very similar for chemical composition and organoleptic characteristics, the classical analytical techniques often result inadequate for a detailed

characterization and differentiation of the product. Therefore some advanced analytical methods have been lately applied, as already reported in Section 1.2, having characteristics of more accuracy and immediate analysis, as it is possible to find in NMR methods.

1.5.2 The "Pomodoro di Pachino" PGI product.

The Cherry Tomato of Pachino represents the characteristic agricultural production of a well delimited zone of the Italian region Sicily, collocated on the extreme southern cape of the isle, that covers the whole territories of Pachino and Portopalo di Capo Passero and a part of the district of Ispica (Ragusa) and Noto (Siracusa) (Figure 1.14).

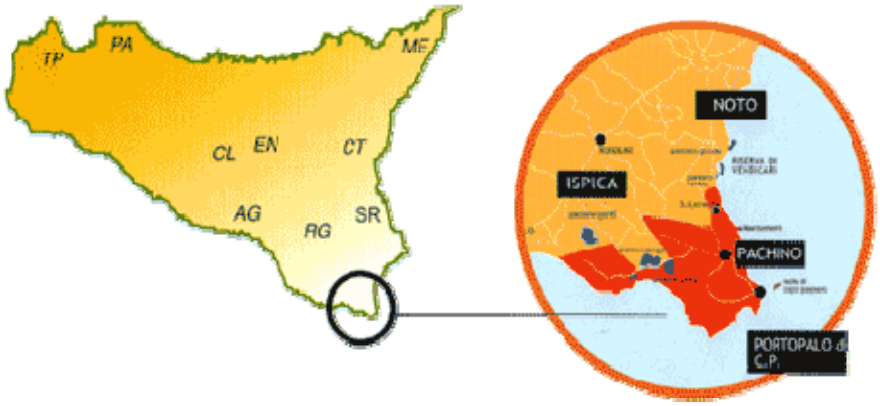


Figure 1.14: Detail of the district of Pachino

Thanks to its peculiar characteristics, due to a lucky combination of climate, salt water irrigation and cultivation techniques, the "Pomodoro di Pachino" (cherry tomato from Pachino) is the first Italian tomato which has obtained, on April 2003 [EC Reg. n.167/2003], the Italian PGI certification (Protected Geographic Indication), becoming well-known either in the national market and in the international one.

The pedological and micro-climatic conditions of the area where this cherry tomato is cultivated have peculiar characteristics, different from the other geographical zones indicated to produce this type of vegetable: the temperate-arid climate typical of the Mediterranean zone, very high temperatures, a huge quantity of solar radiations whose intensity is higher than 300000 lux, the proximity to the sea that determines a climate mitigation, the low

frequency of winter-spring frosts and the texture of the soil (usually sandy) are the factors that determine the sharpening of the qualitative characteristics of the product. But even more interesting is the effect due to the salinity of the irrigation waters: these are pulled out from salted groundwater and they are cause of reduction of the dimensions of tomatoes, which is the reason of a big increment of the organoleptic characteristics. The tomato, for the effect of water stress, increases his quantity of reducing sugars acquiring, in this way, a higher sweetness.

Picture 1.1: Cherry tomato of Pachino

For the cultivation conditions above described, which have as direct consequence an harvesting much lower than the possibilities of the employed cultivars, this agricultural production involves an economical expense obviously higher than the cost for traditional cultivations of cherry tomatoes, that determines an important enhancement of the price of the final product. For this reason is necessary to protect either producers and consumers from whom, not having the same expenses, introduces in the market a fake calling it with the protected name "pomodoro di Pachino".

Picture 1.2: "Pomodoro di Pachino" and European PGI logos

These are the reasons that have leaded the ATPTP (Associazione per la Tutela dei Prodotti Tipici di Pachino – Association for the Tutelage of the Typical Product of Pachino) to require the PGI certification; necessity due to the international notoriety of this product that, if on one hand represents a positive factor because it is a synonymous of prestige, on the other hand implicate the danger of an indiscriminate use of the denomination "Pomodoro di

Pachino IGP" also by the other cherry tomato producers of the other zones of Sicily or other areas of Italy. In fact, even though its great taste is hardly imitable, a lot of commercial fakes, with definitely lower organoleptic characteristics, are present on the Italian and international market. Therefore it was founded a organism of control, the Society of Certification "SoCert", which checks the whole production system of the tomato of Pachino through inspections and verification exams and draws up rules and procedures protocols. In spite of this, if on one hand this organism is able to certify the whole system of cultivation of the tomatoes, on the other hand it cannot take advantage of a scientific method able to discriminate with reliability an unknown sample from a sample coming from the district of Pachino. For this reason, lately, a large interest was focused on analytical techniques able to assess the origin of a tomato sample or, at least, to indicate whether or not it is originating from the Pachino area.

1.6 AIM OF THE RESEARCH

The research work of this PhD study is collocated in a project, financed by the MIUR (Ministero dell'Istruzione, dell'Università e della Ricerca) and by the MiPAF (Ministero delle Politiche Agricole e Forestali), whose aim is to be able to infer from the data, collected using the adopted analytical technologies, information relative to molecular markers, useful to asses the quality and to recognize whether or not the origin of the tomatoes is from Pachino. This project have found a collaboration in the ATPTP which provided the samples necessary to create a NMR spectra database, basic for the further statistic approach. In fact not all the molecular components are useful to determine the good quality of the cherry tomatoes of Pachino, but only some of these whose identification can lead to the recognition of the authenticity of the product under molecular scale. To such aim, ^{1}H Nuclear Magnetic Resonance spectroscopic analyses, together with an appropriated multivariate statistics analysis of the data, have been carried out on an elevated number of samples coming from the region of Pachino and on some other samples coming from other zones of Italy. As a further study, the effect of a different magnetic field strength has been tested by using two different spectrometers operating at 200 and 400, with the aim to test if a higher spectra resolution can make the difference on samples' classification.

2 MATERIALS

The following materials have been used along the present research work:

2.1 GLASSWARE AND DISPOSABLE MATERIALS

- 30 ml glass vials with sealing caps
- Plastic vessels for sample weighting
- Steel spatula for sample weighting (VWR™ International)
- Ceramic mortar (grinding vessel) and pestle
- Precision Microliter Pipettes PIPETMAN® Gilson (P1000, P200, P100, P20, P10) and their respective plastic tips
- 50 ml plastic centrifuge tubes with screw sealing caps, Sarstedt (62.547.254)
- 15 ml plastic centrifuge tubes (gamma sterilized) with screw caps, Oregon Scientific
- 1.5 ml safe-lock plastic tubes (centrifugation up 30000*g), Eppendorf (0030 120.086)
- NMR Sample Tubes (up to 700 MHz, 8" (203 mm), round bottom, with plastic caps), AmpolNMR (Cat.no AP5-600-8)

2.2 REAGENTS

- Milliq demineralized water produced by Helix5
- Glacial Acetic Acid (100% CH_3COOH, d=1.05 Kg/dm^3), Merk (UN 2789)
- HEPES (99.5% N-[2-Hydroxyethyl]piperazine-N'[2-ethanesulfonic acid], $C_8H_{18}N_2O_4S$) pKa=7.5 at 25°C, Sigma (H-3375)
- Sodium Hydrogen Carbonate ($NaHCO_3$, anhydrous) Carlo Erba (478537)
- UVASOL® Deuteriumoxid (99.9% D_2O, d= 1.11 Kg/dm^3), MERK (S4036166-519)
- Sodium Hydroxide (100% NaOH, INCOFAR) in chips

2.3 SOLUTIONS

Three different deuterated buffer solutions, at three different pH values, have been prepared being careful to minimize the introduction of protoned water in the sample. Therefore, salts for the buffer preparation have been used in a dry solid form (where possible) or in solutions at the highest possible concentration. All the buffer solutions have been prepared when necessary in small quantity each time (20 ml), so that they could be quickly used and replaced with others of new preparation in order to avoid an excessive ^{2}D exchange with the ^{1}H air humidity.

BUFFER SOLUTIONS:

- Buffer pH=4.0

For this value of pH, the Acetic-Acetate buffer has been chosen at a concentration of 100 mM. An amount of 114.3 μl of Glacial Acetic Acid is withdrawn with a Gilson micropipette p200 and diluted in 20 ml of deuterated water. The pH has been adjusted to 4.0 with the addition of small chips of pure NaOH (sodium hydroxide anhydrous), with the help of a digital pH-meter.

- Buffer pH=7.0

It has been prepared by dissolving 0.476 g of HEPES, weighted with the analytical balance, in 20 ml of deuterated water, up to a final concentration of 100 mM, thus the pH was adjusted to 7.0 by adding small chips of pure NaOH. The final correct value of pH has been reached with the help of a digital pH-meter.

- Buffer pH=10.0

For this value of pH, the carbonate-bicarbonate buffer has been chosen at a concentration of 100 mM. It has been prepared by dissolving 0.168 g of sodium bicarbonate ($NaHCO_3$), weighted with the analytical balance, in 20 ml of deuterated water and, as for the other two precedent buffers, the pH value has been corrected with the addition of small chips of NaOH under control of a digital pH-meter.

2.4 INSTRUMENTATIONS

2.4.1 Classical laboratory instrumentation

- ULTRA–LOW TEMPERATURE FREEZER (Lowest temp. -86°C, 328 lt), NUAIRE (NU6511)
- Technical balance (max 2200 g, d= 0.01 g), SCALTEC (SBA 52)
- Analytical balance (max 220g, d= 0.0001g), SCALTEC (SBC 31)
- Digital pH-meter, Jenway (model 3310)
- Combined mini-electrode (Reference electrode, saturated KCl / Combination pH/orp electrode 3.0 M KCl, saturated with AgCl), Jenway
- Thermostatic oven, INSTRUMENTS s.r.l. Bernareggio(MI)
- Orbital shaker (with thermostatic Cupola "Climatic Hood", mod.810), ASAL (mod.709)
- Heating magnetic stirrer, VELP® scientifica (model ARE)
- Micro-centrifuge (speed 0 to 14000 RPMs equipped with F241.5P rotor with 24 cavities for 1.5 ml vials), Beckman Coulter™ (Microfuge® 18)

2.4.2 200 MHz NMR spectrometer

For recording NMR FIDs at 200 MHz it has been used a Bruker Biospin (Karlsrhue, Germany) spectrometer operating at 200.132 MHz ^{1}H Larmor frequency and equipped with AC200 console, ASPECT3000 computer, 12 bit Analogical Digital Converter (ADC), broadband probe head from ^{15}N to ^{1}H and deuterium lock system. The spectrometer is interfaced with a PC computer with a Linux operating system for data transferring. The NMR system is shown in Picture 2.1 where also the Oxford Superconducting NMR Cryomagnet (200/54), generating a 4.7 Tesla constant magnetic field, is visible. In Picture 2.2 is shown a particular of the AC200 console and of the ASPECT3000 computer.

Picture 2.1: Bruker 200 MHz NMR spectrometer

Picture 2.2: A particular of the AC200 console.

The NMR system is also equipped with a Variable Temperature Unit (B-VT2000) and a Pneumatic Unit for probe head nucleus changing.

2.4.3 400 MHz NMR spectrometer

The NMR spectrometer utilized for recording FIDs at 400 MHz presents the following characteristics:

Varian Mercury Plus AS400/54 spectrometer, operating at 400.097 MHz ^{1}H Larmor frequency, equipped with:

- Mercury Plus Console
- 23 channels shimming unit
- Dual Fullband RF system
- Pulsed Field Gradient Driver (model L700)
- Variable Temperature Controller (-60°C to +100°C) (model L900)
- 5 mm PFG gradient 4-nuclei probe (1H/19F/13C/31P)
- 400MHz (9.4 Tesla) Oxford Active Shielded Superconducting Magnet System (400/54)
- Sun BLADE 150 Host Workstation with Solaris 10 Operating System, 80 GB System Disk, 512 Mb RAM, CD ROM SCSI Drive and VnmrJ Software 1.1D

Picture 2.3: Varian spectrometer operating at 400 MHz

Picture 2.4: Oxford superconducting magnet

The Picture 2.3 shows the full NMR system. A particular of the 9.4 Tesla Narrow Bore superconducting magnet is shown in Picture 2.4.

3 METHODS

3.1 PROTOCOLS

The following protocols have been developed for a standardized procedure.

3.1.1 Cherry tomato fruit extraction in D_2O buffer solutions

The milled freeze-dried samples of tomato to be submitted to analysis, maintained inside their glass vial in nitrogen atmosphere, have been withdrawn from the ultrafreezer, where they have been preserved at a temperature of -75°C, and placed at room temperature for about 10 minutes, until complete thawing. Then the sample has been weighted with a technical balance. In the case of not well freeze-dried or not well ground samples, fruit powders have been further manually milled by using ceramic mortar and pestle, to make it more homogeneous.

For each tomato sample, three 1.5 ml plastic tubes (Eppendorf), opportunely marked with the number of the sample and the relative pH value, have been prepared. Using the analytical balance, exactly 100 mg of freeze-dried sample have been accurately weighed inside every Eppendorf. These steps have been performed rather quickly in order to avoid the absorption of atmospheric damp and, successively, the freeze-dried powder has been rapidly reinserted inside its vial, under gaseous nitrogen insufflation, and immediately sealed and put back in ultra freezer to -75°C.

The freeze-dried sample weighted into the Eppendorf 1 has been extracted with 1 ml of pH 4.0 buffer solution, withdrawn with a Gilson p1000 micropipette. The same procedure has been also repeated for the Eppendorf 2 and 3 using respectively the buffer solutions at pH=7.0 and at pH=10.0 and then the three Eppendorf have been sealed with Parafilm and placed in a rack inside the thermostatic cupola of the orbital shaker kept at 30°C constant temperature.

After some initial tests, the extraction time of 2 hours at a 250 RPMs speed has been chosen in order to obtain an exhaustive extraction in the shortest time.

After the extraction, the suspensions contained in the Eppendorf microtubes have been spun in the microcentrifuge in order to achieve extract separation from the residual solid component. Centrifugation has been accomplished for 7 minutes at 15000 RPMs speed. After centrifugation the respective separated extracts have been transferred in three new Eppendorf, previously marked with the correspondent codes, with a Gilson p1000 micropipette, trying not to pick up solid particles. The quantity of clear extract usually recovered after this operation resulted to be about 900 µl.
Even if at this point the samples were ready for NMR analysis they have been frozen in ultrafeezer until their analysis moment, in order to prevent oxidation or hydrolysis phenomena.

3.1.2 NMR samples preparation

Picture 3.1: A sample into the NMR tube

The sample to be submitted to analysis has been withdrawn from the ultrafreezer few minutes before the analysis, until thawing and thermal equilibrium with ambient. Then it has been centrifuged at 15000 RPMs for 3 minutes with the purpose to separate the eventual solid particulate. Using a Gilson p1000 micropipette, equipped with a rigid plastic cannula inserted on the tip, 600 µl of the sample have been placed on the bottom of a 5 mm NMR tube, previously washed with distilled water and perfectly dried. At this point the sample has been inserted in the spinner and subsequently in the probe of the spectrometer, where it has spun to a speed of 20 RPMs. Before proceeding with tuning, locking and shimming operations around 15 minutes have been waited: in such a way the temperature of the sample reached the probe temperature selected for NMR analysis.
In Picture 3.1 is shown a sample extract inserted in the NMR tube with the spinner, ready to be placed inside the spectrometer probe head.
All the instrumental parameters referred to NMR analysis of the samples are reported in the following Section 3.2.1 distinguishing between the two different magnetic field strengths.

3.1.3 Protocol for 200 MHz NMR analysis

Before each 200 MHz NMR FID acquisition the procedures below listed have been followed.

- SIGNAL LOCK: with the sample inserted into the probe and spun at 20 RPMs, the Deuterium Lock signal has been found, centered and maximized, and then the radio-frequency has been locked.
- SHIMMING: axial magnetic field homogeneity has been regulated varying current circulating into its shim coils. Rarely also radial shim coil currents have been regulated.
- FID and SPECTRUM CHECK: before acquiring the complete FID an attempt has been performed acquiring only 8 scans after 4 dummy scans. The FID registered has been checked and FT transformed for signals validation: only if the ^{1}H α-D-Glucose downfield signal at 4.650 ppm presented a half-height width smaller than 0.6 Hz the spectrum has been considered good and the shimming step complete. Otherwise new shimming steps have been performed again and again until obtaining a good signal width.
- FID FILE SAVING: every FID has been saved with its univocal file name.

3.1.4 Protocol for 400 MHz NMR analysis

Before each 400 MHz NMR FID acquisition the procedures below listed have been followed.

- MATCHING AND TUNING OF THE MAIN FREQUENCY: with the sample inserted into the probe and spun at 20 RPMs, the best matching and tuning of the main frequency has been performed operating onto the apposite screws located under the probe head. A digital display on the foot of the instrument indicated a value to be minimized in order to achieve the best matching and tuning.
- SIGNAL LOCK: Using the apposite procedure included into the instrument software (VnmrJ) the Deuterium Lock signal has been found, centered, maximized and then the radio-frequency has been locked.
- SHIMMING: being the instrument equipped with a 23 channel gradient shimming unit with pre-charged shim maps, the shimming step has been performed in a totally automatic way. Since not every time this operation reached a satisfying result, in those cases it has been repeated several times until good result evaluated by measuring the half-height width of a signal successively specified. Sometime it has been necessary a manual fine shimming,

performed operating onto the axial shim coils through the specific shim coil control page of the software. Rarely also radial shim coil have been regulated.

● FID and SPECTRUM CHECK: before acquiring the complete FID a test step has been performed acquiring only 4 scans after 2 dummy scans. The FID registered has been checked and FT transformed for signals validation: only if the downfield peak of the ^{1}H signal of β-D-Glucose set at 4.650 ppm presented an half-height width smaller than 1.2 Hz the spectrum has been considered good and the shimming step complete. Otherwise new shimming steps have been performed again and again until a good signal width obtaining.

● FID FILE SAVING: every FID has been saved with its univocal file name.

3.1.5 Protocol for the temperature and pH dependence of signals

A single aqueous extract of cherry tomato sample coming from Pachino (ID number 225) has been investigated to study its spectral changes related to pH and temperature variability. The sample has been analyzed at three different pH values, each at three different temperatures. The pH has been changed by stepwise titration with a small quantity of NaOH. Three NMR spectra have been acquired at three different probe temperature on the sample prepared according to the standard procedure: 296, 298 and 300 °K. The parameters of the NMR analysis (see Section 3.2.2) have been kept like those of all other samples except for the number of acquired transients that has been reduced to 512 (half of the usual number) to shorten the total time of acquisition at 35 min.

Between a recording and the next one it has been necessary to attend about 10' for the stabilization of the set temperature, afterward a new round of magnetic field shimming and signal locking has been performed: this required a total time of about 20-25 min.

At the end of the first three acquisitions at different temperature, the sample has been drawn out from the NMR tube and accurately transferred into an Eppendorf microtube of 1.5 ml. After careful calibration of the pHmeter, equipped with a combined mini-electrode, the first pH value has been measured and annotated. The pH value has been then slightly changed introducing a very small quantity of NaOH picked up with the tip of a glass Pasteur pipette scraped over a NaOH chip and dipped inside the sample. The same Pasteur pipette has been also used for successive pH value correction in order to lose the smallest possible quantity of sample. For the same reason, the pH value has been measured only after that the 3 experiments at the selected temperatures have been completed.

3.2 INSTRUMENTAL PARAMETERS FOR NMR ACQUISITION

For this research work it has been chosen to not use any solvent suppression Pulse sequence in order to avoid losing signals falling near that of the solvent. Therefore it has been used the simplest pulse sequence either for FIDs acquired at 200 MHz and for those acquired at 400 MHz. Since two NMR spectrometers of different brand (Bruker and Varian) have been utilized, the way as the instrumental parameters are named slightly changes and then are briefly explained in the following dedicated sections.

3.2.1 ^{1}H-NMR spectra of Cherry tomato extracts recorded at 200 MHz

The parameters of a Bruker pulse sequence called "ZG" have been setup for recording FIDs at 200 MHz. The pulse sequence is illustrated in Figure 3.xx that also reports the times of each step.

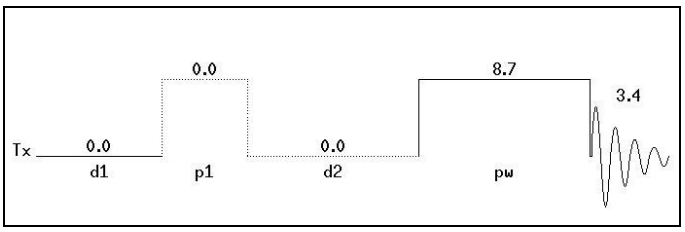


Figure 3.1: Pulse sequence scheme used for the 1D ^{1}H-NMR experiments at 200 MHz

The main Bruker acquisition parameters utilized are the following:

- Acquisition mode: SEC
- Dummy Scans (DS): 16
- Number of Scans (NS): 4096
- Time Domain (TD): 8 K (r) + 8 K (i)
- Data Size: 8 K
- Acquisition time (AQ): 3.4078 s
- Recycle Delay (d1): 0.0 s
- Preparation Pulse (p1): 0.0 μs
- Mixing Time (d2): 0.0 ms

- Acquisition pulse (pw): 8.70 µs
- Spectral Window (SW): 12 ppm
- Sample rotation (Spin): 20 Hz
- Temperature of acquisition (T): 298 °K
- Experiment Time (EXPT): 252 min

3.2.2 ^{1}H-NMR spectra of Cherry tomato extracts recorded at 400 MHz

The parameters of a Varian pulse sequence called "s2pul" have been setup for recording FIDs at 400 MHz. The pulse sequence is illustrated in Figure 3.xx that also reports the times of each step.

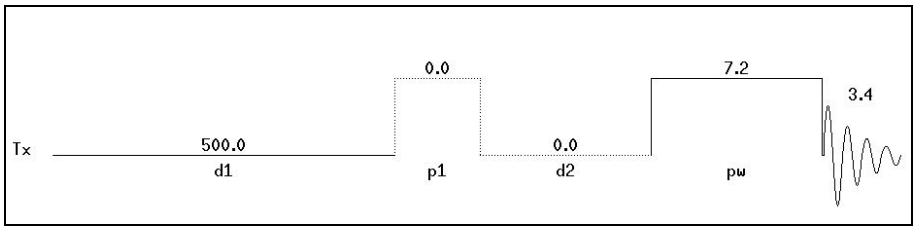


Figure 3.2: Pulse sequence scheme used for the 1D ^{1}H-NMR experiments at 400 MHz

The main Varian acquisition parameters utilized are the following:

- Steady Scans (ss): 16
- Number of Transient (nt): 1024
- Number of complex Points (np): 32768 (32K)
- Data Size: 16 K
- Acquisition time (at): 3.416 s
- Recycle Delay (d1): 500 ms
- Preparation Pulse (p1): 0.0 µs
- Mixing Time (d2): 0.0 ms
- Acquisition pulse (pw): 7.20 µs
- Spectral Window (sw): 12 ppm (4796.2 Hz)
- Sample rotation (Spin): 20 Hz
- Temperature of acquisition (T): 298 °K
- Experiment Time (exptime): 71 min

3.3 NMR DATA HANDLING

All FIDs acquired along the present thesis work have been transferred to a personal computer equipped with the software Mestre-C 4.9.8 (www.mestrec.com) for data processing [200]. Magnetic Resonance Companion (MestReC) is a software package that performs all necessary data processing, visualization, simulation and analysis of high resolution nuclear magnetic resonance (NMR) data. Since this software owns data filters for many types of NMR spectrometers, no preliminary data conversion has been necessary. An example of the user-friendly graphical interface of this program is shown in Figure 3.3

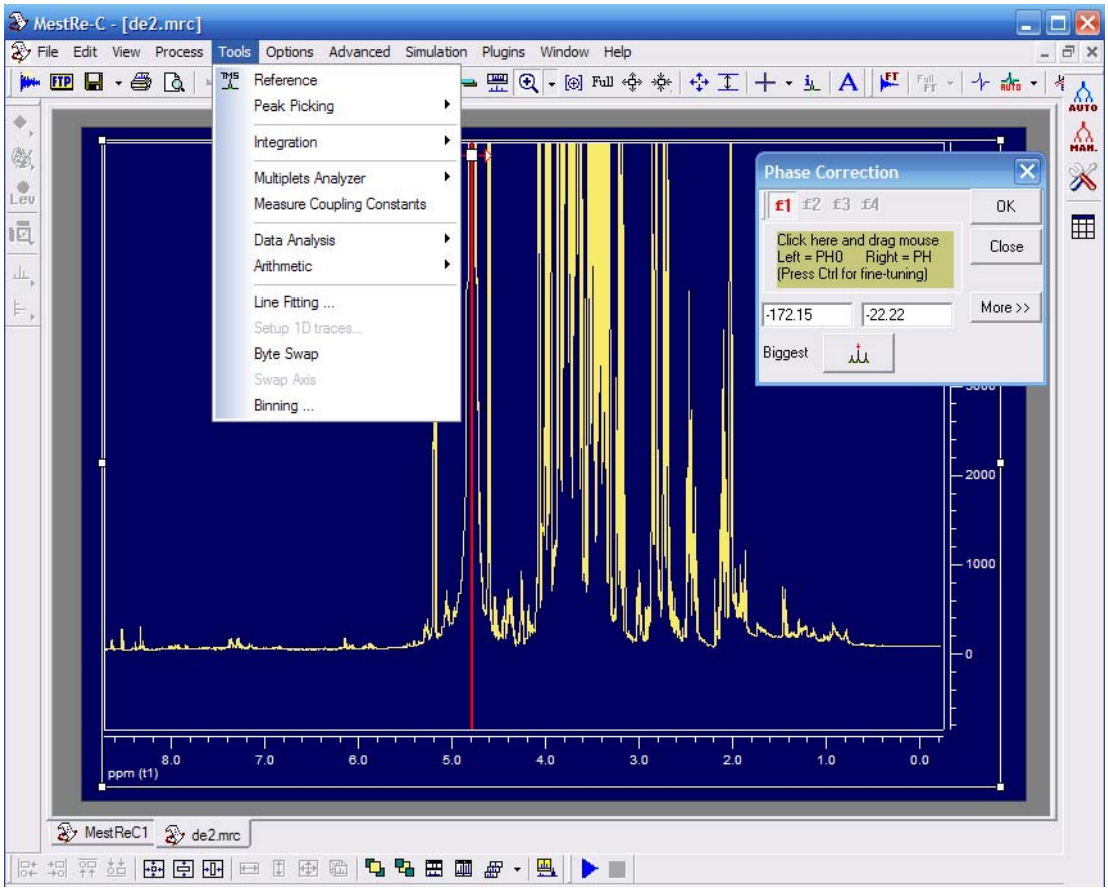


Figure 3.3: Mestre-C Software graphical interface.

3.3.1 FID processing for acquisition at 200 MHz

The 1D ^{1}H-NMR data recorded at 200 MHz have been processed with the following procedure:

- Fourier Transform (FT) along F1 with phase quadrature, automatic drift correction and 0.1 Hz exponential apodization (Line Broadening)
- Manual phasing (phase correction) of each spectrum along F1
- Raw automatic Baseline correction with Bernstein polynomials function of order 5
- Fine linear multipoint Baseline correction of the spectrum with selection of single points forced at 0.
- Chemical Shift referencing over methyl protons of acetate signal at 2.029 ppm for spectra registered at pH 4.0
- Chemical Shift referencing over characteristic protons signal of HEPES at 3.175 ppm for spectra registered at pH 7.0
- Chemical Shift referencing over ^{1}H α-D-Glucose downfield signal at 5.13 ppm for spectra registered at pH 10.0, arbitrary chosen because carbonate buffer solution doesn't present any NMR signal except the unusable water exchanging one
- Spectral region selection between 0.0 and 10.0 ppm
- ASCII conversion of the selected 10 ppm spectral region: 6820 intensity data points (from 8192 total data points) have been exported for each spectrum for successive chemometric processing (see Section 3.4.1)

3.3.2 FID processing for acquisition at 400 MHz

1D ^{1}H data in D_2O have been processed with the following procedure:

- Fourier Transform (FT) along F1 with phase quadrature, automatic drift correction and 0.5 Hz exponential apodization (Line Broadening)
- Manual phasing (phase correction) of each spectrum along F1
- Fine linear multipoint Baseline correction of the spectrum with selection of single points forced at 0.
- Chemical Shift referencing over methyl protons of acetate signal at 2.029 ppm for spectra registered at pH 4.0.
- ASCII conversion of the whole spectra: 16384 (16K) intensity data points have been exported for each spectrum for successive chemometric processing (see Section 3.4.2)

3.4 CHEMOMETRIC DATA PROCESSING

In this PhD thesis every statistical and chemometric analysis has been carried out with the free open source program "R" (version 2.4.0, www.r-project.org) [201].

A short description of the features of the program is reported below.

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. Among other things it has

• an effective data handling and storage facility,

• a suite of operators for calculations on arrays, in particular matrices,

• a large, coherent, integrated collection of intermediate tools for data analysis,

• graphical facilities for data analysis and display either directly at the computer or on hardcopy, and

• a well developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.

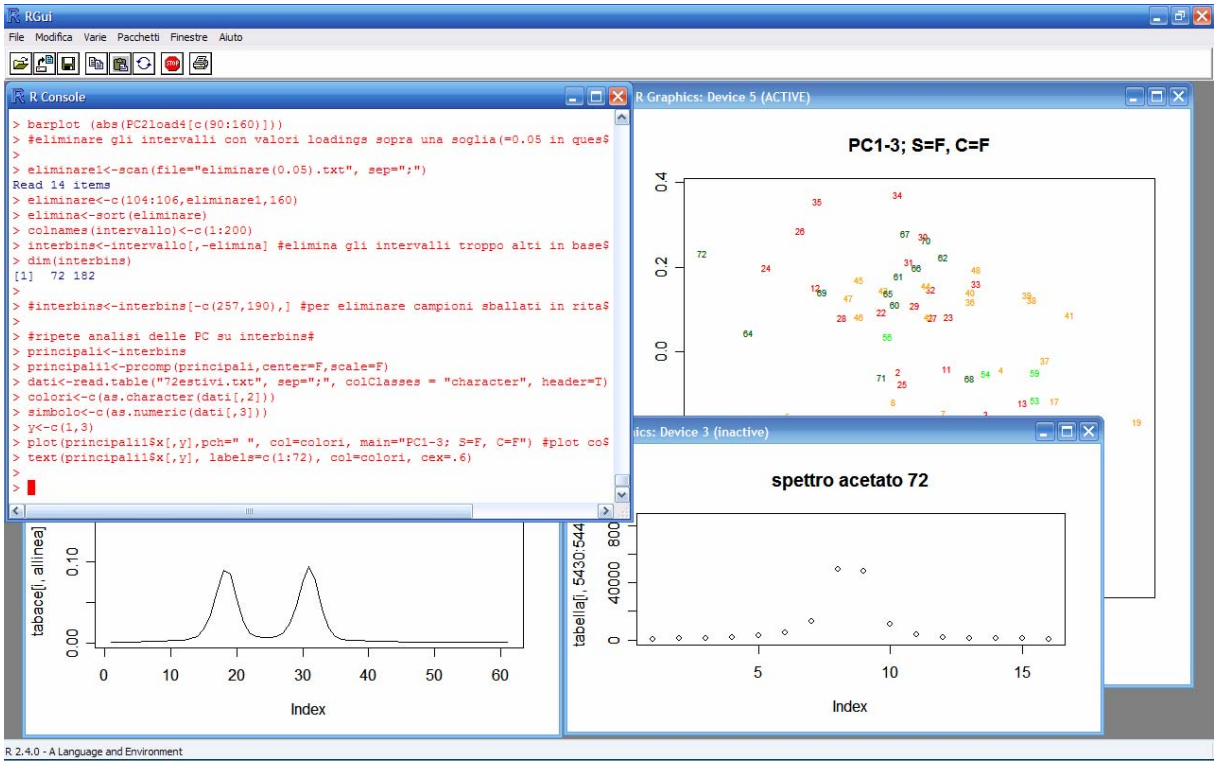


Figure 3.4: R environment with several devices opened during chemometric analysis of data

The term "environment" (Figure 3.X4) is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.
R is very much a vehicle for newly developing methods of interactive data analysis. Technically R is an expression language with a very simple syntax. Elementary commands consist of either expressions or assignments. An expression may be given as a command, and it is evaluated, printed, but the value is lost. An assignment also evaluates an expression and passes the value to a variable but the result is not automatically printed [202].
Several scripts in R language have been in-house written in order to handle spectroscopic data for chemometric analysis. The most important algorithms developed for data handling and chemometric analysis are briefly described below, taking into account the different data processing necessary for every spectroscopic dataset acquired.
The complete R scripts for data handling are reported in Appendix A.

3.4.1 Chemometric data processing of spectra recorded at 200 MHz

Spectral alignment, vertical normalization and binning of the dataset
Starting with single ASCII files, exported after 1D ^{1}H NMR FID processing, all the data have been arranged in a **S** x **V** matrix where **S** represents matrix rows consisting in NMR spectra and **V** represent matrix columns consisting in spectral variables (intensities for each data point). S can vary form case to case but V is always equal to 6820 data points corresponding to the spectral window exported by Mestre-C. Several text files listing alphanumeric strings relative to the samples to be processed have been preliminarily created, containing useful information about every sample, like spectrum file name, code number, harvesting season, cultivar, geographical region, etc. These files are read by the program at the beginning for initial dataset definition.
Data matrix creation is accomplished by a series of R language commands that take into account different situations depending on which sample group is going to be submitted to chemometric analysis (see Appendix A for script details)
Considering the necessary data preprocessing steps, preceding chemometric analysis, illustrated in Section. 1.4.1, some algorithms have been developed in R language in order to obtain reliable dataset to be submitted to multivariate analysis.

First of all every line of the NMR data matrix is plotted, in progressive order as in a sequence of frames during the projection of a movie, within a specific range that comprehends the ^{1}H α-D-Glucose doublet signals. In this way it is possible to take a look at the horizontal shifting of such a signal over all the NMR spectra. This operation helps to notice if there are anomalous spectra and if the dataset is enough homogeneous for preprocessing steps.

The same is made over acetate signal in a new plotting device thorough some script command lines that generate also a new data vector ("ace") containing index position of such a signal for every spectrum (matrix rows) of the data set.

The latter vector "ace" is used to perform a spectral normalization of every row in the data matrix accomplished dividing each of all data points of each row of the data matrix for the corresponding "ace" value. A new normalized data matrix ("tabace") is created.

Data normalization precedes horizontal signals alignment that can be performed in two different ways, according to data homogeneity, as described below:

- a first algorithm performs signals scanning considering a specific spectral region of a selected target spectrum, defined as reference spectrum, and sliding the same spectral region of the other samples/rows until the least square difference among the pair of compared spectra is reached. This corresponds to a best fitting algorithm between spectra. A vector containing points of shifting for every spectrum is written ("index") for the successive signal alignment;
- an alternative algorithm performs the signal scanning within a specific spectral region chosen in a selected target spectrum. The algorithm searches for the highest intensity in the narrow given range, where it is expected to find the signal on which all spectra will be aligned (reference peak). It must be the most intense signal in the chosen narrow region: in the present study, the downfield peak in the doublet signal of ^{1}H α-D-Glucose has been chosen as reference peak. Thus, the algorithm passes the position of the peak maximum to a variable; afterward, it calculates the difference between the position in each spectrum and that one in the target spectrum, storing such values in a vector called "maxvet".

"index" and "maxvet" vectors can be alternatively used to perform signal alignment for every spectrum (data matrix rows). Every row is aligned over the reference peak by sliding every spectrum by a number of datapoints equal to index or maxvet. An algorithm perform these operations returning a new matrix Sx6820 with aligned variables ("allineanomer")

Spectra dimension is reduced from 6820 to 6800 data points by cutting useless edge data containing only noise. A new reduced data matrix is obtained ("allineanomer1").

Since normalization over acetate signal has been seen to be not effective on obtaining the best vertical scaling of data an "Integral to a Constant Sum" (CS) [156] algorithm has been written and applied to the dataset obtaining a data matrix normalized in a reliable manner ("allineati.CS").
A binning step for variables reduction is now performed: the number of variables is lowered at 200, by performing summation of the intensities of 34 consecutive datapoints, constituting a bin of data, along the whole width of each spectrum. An Sx200 data matrix ("intervallo") represents the new binned dataset.
Before multivariate analysis 9 bins, containing water and acetate signals, obviously not significant for the successive steps, are removed from the dataset: a new data matrix ("perpca") is now ready to be processed with multivariate tools.

Principal Component Analysis

Principal Component Analysis of the data is carried out using a prebuilt R-language command ("*prcomp*"). Each variable (column) is centered and scaled during PCA analysis. The resulting scores and loadings matrices are produced in a single object ("principali4") where they are stored together ("principali4$x", "principali4$rotation"), respectively. Some commands have been written in order to generate meaningful plot of two dimensional PC scores spaces, in order to highlight differences among cherry tomatoes by using different labels and colors.

Linear Discriminant Analysis

LDA has been carried out using the corresponding R language command ("*lda*"). Canonical Variates matrix is produced in a R object from where it can be recalled and plotted ("zlda").
LD Scores vectors ("LDAx_scores" where x stands for every calculated LD space) have been then calculated by matrix product between LD Canonical Variates matrix and the selected PC score matrix. Further command lines have been written in order to generate meaningful plot of two dimensional and three dimensional LD scores spaces, in order to show samples grouping according to their geographical origin, by using different labels and colors.

3.4.2 Chemometric data processing of spectra recorded at 400 MHz

Starting with single ASCII files, exported after 1D ^{1}H NMR FID processing, all the data have been arranged, as for the 200 MHz dataset, in a **S** x **V** matrix where **S** represents matrix rows consisting in NMR single spectra and **V** represent matrix columns consisting in spectral variables (intensities). S can vary form case to case but V is always equal to 16K (16384) data points considering the whole spectral window. Several text files listing samples to be processed have been previously written, containing useful information about every sample like spectrum file name, code number, harvesting season, cultivar, geographical region, etc. These files are read by the software at the beginning for initial dataset definition.

Data matrix creation is performed by a series of R language commands that take into account different situations depending on which sample group is going to be submitted to chemometric analysis (see Appendix A for script details)

The data are hence organized in a raw-data matrix ("spettri") where every spectrum is represented in his initial conditions of horizontal signals' position and vertical scaling. This matrix represents the starting point for all successive chemometric processes.

For successive normalization and horizontal alignment of signals, chemometric processing needs a target spectrum. The scores matrix of a preliminary Principal Component Analysis of the samples allows selecting the one with average spectral features ("spref").

An algorithm analyzes each row of such a matrix and individuates the maximum value on the acetate region: this value results useful for successive initial data normalization: every point of every matrix row is then divided by such a maximum found for every spectrum, obtaining a new roughly normalized data matrix ("refacetato").

Considering the necessary data preprocessing steps, preceding chemometric analysis, illustrated in Section 1.4, some algorithms have been developed in R language in order to obtain reliable dataset to be submitted to multivariate analysis.

A first horizontal alignment of the variables is accomplished using as referencing point the variable index of the ^{1}H β-D-Glucose downfield signal at 4.650 ppm of a target spectrum, selected among all the others because the most similar to the average spectrum. This operation is carried out looking for the shifting of this signal for every row of the data matrix and generating a vector of length S ("maxvet") containing the number of points of shifting of this signal for each spectrum. Successively every row is aligned over this signal sliding every spectrum, according to his shifting value in "maxvet". An algorithm performs these operations returning a new matrix Sx13600 with aligned variables ("allineati")

A common data normalization "integral to a constant sum" [156] of the whole spectra is then carried out using a series of command that return a total sum of the intensities of each spectrum equal to 100. Since the information relative to real signal intensity want to be maintained, a successive vertical scaling of each spectral row is performed using, for every rows, an index ("up") calculated by the algorithm. A new matrix of normalized data is, in this way, stored in "allineati.CS".
The further preprocessing step consists in spectral variables reduction by means of a binning algorithm. It is clear that external regions of the spectrum do contain only noise and hence are not useful for successive chemometric analysis. Therefore, in order to obtain a number of useful variables V that can be divided in homogeneous groups (bins), these external points are cut from the original matrix obtaining a new matrix ("spettri") presenting V dimension equal to 13600 variables.
Each whole spectrum, consisting in 13600 data point,s has been divided in 200 integral regions (bins) of the same width equal to 68 points. This corresponds to 0.05 ppm on an NMR spectrum recorded at 400 MHz, where a single signal has average half-height width equal to 0.005 ppm (1÷10). Binning is therefore carried out on the whole matrix dataset obtaining a new Sx200 reduced matrix ("intervallo"). The latter contains also some spectral regions not useful for chemometric data processing. Therefore some script lines have been written in order to eliminate those regions relative to residual water (HDO) signal and acetate signal: in this way 6 intervals (bins) are cut from the dataset originating a new matrix dataset Sx194 ("perpca").

Bins containing pH dependent signal removing: 24 bins previously selected by pH dependence study (Section 4.2.1) as containing signals sensitive to pH small variations, and listed in a data file ("eliminare.txt") read by the script, are cut from the data matrix generating a new dataset ("interbins") Sx180 that is ready to be submitted to PCA analysis.

PCA analysis

Principal Component Analysis of the data is carried out using the relative R language commands ("*prcomp*"). In order to assign the same importance to each variable, each column is centered and scaled during PCA analysis. Resulting Scores and Loadings matrices are produced together in a single object from where they can be recalled and plotted ("principali4"). Some commands have been written in order to generate meaningful plot of

two dimensional PC scores spaces, in order to highlight differences among the samples by using different labels and colors.
A PCA scores calculation script, based on loadings matrix determined by PCA of a selected dataset of samples, has been also written in order to calculate PC scores of samples introduced posteriorly: this is made for projecting new unknown samples into the score plots in order to obtain information about their geographical origin.

LDA analysis

Also for 400 MHz data matrices Linear Discriminant Analysis is carried out on PCA scores. As it will be shown in Section 4.2.4, only PC spaces participating to a total variance explanation up to 90% have been kept by the algorithm to perform LDA analysis. New dataset matrices are assembled by R language script in order to obtain a dataset suitable for R LDA analysis, containing all the needful information ("LDAtable").
LDA is therefore carried out using the relative R language commands ("*lda*"). Canonical Variates matrix is produced in an R object from where they can be successively recalled and plotted ("zlda").
LD Scores vectors ("LDAx_scores" where x stands for every calculated LD space) are then calculated by matrix product between LD Canonical Variates matrix and the selected PC score matrix. Further commands have been written in order to generate meaningful plot of two dimensional and three dimensional LD scores spaces, in order to show samples grouping according to their geographical origin, by using different labels and colors.
In LD two dimensional score plots every determined group (class) is represented with a different color and each group is surrounded by ellipses representing its relative Mahalonobis distances calculated by the R command "*Mahalanobis*". In such a way, it visually becomes clearer whether a sample belongs to a group rather than another one.

3.5 FRUITS SAMPLING

3.5.1 Sampling and cataloguing cherry tomato fruits

Cherry tomato fruit samples analyzed in our laboratories for the present study have been harvested in Sicily by cooperating agricultural producers, within determined specific geographical zones accurately chosen at the beginning of the project. Every collected sample, consisting in about 200g of selected bunches of cherry tomato fruits, has been sent to the Experimental Institute for Plant Nutrition (INSP) in Rome where it has been freeze-dried and then homogenized by grinding. Although a standardized procedure has been searched, the final granulometry varies among samples depending on the freeze-drying performances. Such milled samples have been then put inside glass vials, immediately sealed in nitrogen atmosphere in order to prevent oxidative or degenerative processes. Every vial contains more or less 3 g of milled tomato sample, corresponding to around 30 g of fresh fruits. Five to seven glass vials has been obtained from every sample.
Firstly, the agricultural farms to be involved for these sampling procedures have been chosen during the project definition. Four different farms located in the Pachino area have been selected according to different salinity (measured by electrical conductibility parameter) of their irrigation water. In Table 3.1 are shown the average conductivity values of the irrigation waters of these first four farms.

	Farm 1	Farm 2	Farm 3	Farm 4
Conductivity (**µS/cm**)	5400	1500	2000	5500

Table 3.1: Electrical conductivity of irrigation waters for selected farms

For every selected farm have been individuated at least 3 different tomato plants (all belonging to the NAOMI cultivar), that have been marked with letters A, B and C, from which the fruits samples are withdrawn approximatively every three weeks during harvesting seasons. Some new different farms have been added during the whole time of the research project: some of these are inside the zone of Pachino while some other ones are located in different geographical regions successively specified. Not all of this fruit samples belong to Naomi cultivar: Shiren, Franchie and Rubino Top cultivars have been represented as well as market's cherry tomatoes and aspecific samples.

In order to test if the analytical approach adopted was suitable to predict the geographical origin of samples, a comparison between samples from Pachino and from other external areas was necessary. For samples harvested in summer it was possible to select another Italian region not in Sicily, where these vegetables are normally cultivated: these samples have been harvested in Sabaudia (Lazio).

For winter, at our knowledge, Italian cherry tomatoes originating from areas different from Pachino are not available except for those cultivated in other Sicilian areas, where the winter cultivation is still economically convenient. Thus, the choice of samples with origin different from the Pachino area fell on the neighboring Licata region. It is worthnoting here that both regions have similar pedoclimatic characteristics.

Both Licata and Sabaudia localization are represented in Figure 3.5.

Figure 3.5: Localization of the various geographical origins of samples.

For every plant the sampled fruits have been harvested from different bunch of the vegetables in order to collect berries at a homogenous, commercial ripening. This operation has been effected, for instance, withdrawing the first four cherry tomatoes of every cluster up to reach the weight of 200 g or directly withdrawing one or more clusters until to get around the same weight. Also for the samples coming from selected farm outside the zone of Pachino the same sampling procedure has been adopted.

Since the whole sampling project covered more than 3 years the samples have been subdivided in sampling groups. Every group of samples, after freeze-drying and grinding steps performed in ISNP in Rome, has been delivered to our laboratory where it has been submitted to integrity check, sorted, catalogued and stored in Ultrafreezer at -75°C until NMR sample preparation.

In Table 3.2 an example of the sampling catalogue is reported representing almost all the classes of samples involved in this research work.

For the complete list of samples collected see Appendix B.

ID	Code	Year	Season	Date	Arrival	Farm	Salinity	Plant	Origin	Cultivar
1	1Az.1A1a	2003	win. 03	10/03/03	1	1	5400	A	Pachino	Naomi
2	1Az.1A1b	2003	win. 03	10/03/03	1	1	5400	A	Pachino	Naomi
3	1Az.1A1c	2003	win. 03	10/03/03	1	1	5400	A	Pachino	Naomi
35	3Az.3A1a	2003	sum. 03	23/06/03	3	3	2000	A	Pachino	Naomi
36	3Az.3B1a	2003	sum. 03	23/06/03	3	3	2000	B	Pachino	Naomi
38	3Az.4A1a	2003	sum. 03	23/06/03	3	4	5500	A	Pachino	Naomi
39	3Az.4A1b	2003	sum. 03	23/06/03	3	4	5500	A	Pachino	Naomi
174	18Az.6SHA1a	2004	spr. 04	21/05/04	18	6	1500	SH	Pachino	Shiren
175	18Az.7SHA1a	2004	spr. 04	21/05/04	18	7	5600	SH	Pachino	Shiren
176	18Az.8SHA1a	2004	spr. 04	21/05/04	18	8	6000	SH	Pachino	Shiren
179	18 Aspecific3	2004	spr. 04	21/05/04	18	Aspecific	2200		Pachino	Shiren
180	18 Aspecific4	2004	spr. 04	21/05/04	18	Aspecific	2200		Pachino	Naomi
181	18 Aspecific 5	2004	spr. 04	21/05/04	18	Aspecific	1800		Pachino	Naomi
292	27Az18SHA3a	2005	sum. 05	30/06/05	26	18	1800	A	Pachino	Shiren
293	27Az.18SHB3a	2005	sum. 05	30/06/05	26	18	1800	B	Pachino	Shiren
501	1Licata NaomiA1a	2004	win. 04	16/01/04	1			A	(licata)	Naomi
506	1Licata ShirenC1a	2004	win. 04	16/01/04	1			C	(licata)	Shiren
706	1SabaudiaSHA2	2004	sum. 04	01/09/04	1			A	(Sabaudia)	Shiren
707	1SabaudiaSHB1	2004	sum. 04	01/09/04	1			B	(Sabaudia)	Shiren
912	2OrtoNaturaNaomi Gela	2005	win. 05	09/02/05		Ortonatura2000			np(Gela)	Naomi
916	Fondi(Lt)CID Corus	2005	win. 05	22/02/05		CIDCorus			np(Fondi)	Corus
917	Fondi(Lt)CID Piccadilly	2005	win. 05	22/02/05		CIDPiccadilly			np(Fondi)	Piccadilly
920	2OrtoPIU' Licata	2005	win. 05	18/03/05		OrtoPiù			np (OrtoPiù Licata)	cultivar sconosciuta

Table 3.2: an example of the sampling catalogue

The variety of the sampling catalogue is summarized in Table 3.3. and in Table 3.4, distinguishing between analysis carried out at 200 MHz and at 400 MHz.

Sample's origin	Winter season	Summer season	Total
Pachino (Naomi)	108	57	165
Pachino (Shiren)	17	15	32
Pachino (Aspecific)	9	6	15
Licata (AG)	12	0	12
Sabaudia (LT)	0	24	24
Markets	12	11	23
Total n. of samples	158	113	271
Samples suitable for chemometrics	155	112	267

Table 3.3: Samples analyzed at 200 MHz with classes specification

Sample's origin	Winter season	Summer season	Total
Pachino (Naomi)	112	50	162
Pachino (Shiren)	24	8	32
Pachino (Aspecific)	13	0	13
Licata (AG)	13	0	13
Sabaudia (LT)	0	24	24
Markets	14	13	27
Total n. of samples	176	95	271
Samples suitable for chemometrics	166	92	258

Table 3.4: Samples analyzed at 400 MHz with classes specification

3.5.2 Cherry tomatoes for inter-laboratory check test

For testing the general validity of the investigating method developed in this research study an inter-laboratory test on some representative samples of the cherry tomatoes cultivated in the region of Pachino has been performed following the procedures here described. For such intention 11 samples have been chosen and analyzed with NMR spectrometers operating at a Larmor's frequency of 400 MHz both in our laboratory and in an external laboratory. The latter has been individuated at the Experimental Institute for Plant Nutrition (ISNP) in Rome collaborating with us at the same project. Since the spectrometer in use in the laboratory of ISNP is a Bruker one, the experimental parameters have been aligned with those of Varian

spectrometer operating in our labs, in order to obtain comparable spectra suitable for a reliable successive chemometric analysis. Considering that the purpose of the test was that to verify if the use of different NMR spectrometers influences the statistical analysis of the data, and therefore their classification, the analyzed samples have to result identical in the preparation step. For this reason the 11 NMR samples of cherry tomato extract have been prepared and analyzed in the internal laboratory (following the protocols already described) and immediately brought to ISNP laboratory keeping them at 0°C in a chilled thermostatic chamber. In this way the samples analyzed in Rome are physically the same extracts that have been analyzed in our internal laboratory.

The cherry tomatoes selected for the inter-laboratory check test are listed in Table 3.5 (The increasing protocol number and the complete code indicating sample group, farm, plant and bunch are reported).

Winter season 2003	Summer season 2004
(35) 3Az.3A1a	(198) 21Az6SHA5a
(46) 4Az.3B2a	Spring season 2005
(82) 5Az4C6a	(223) 23Az12SHR(7/8)a
Winter season 2004	(257) 25Az9SHB7a
(107) 8Az.1C3a	Summer season 2005
(123) 10 Az.1A5a	(279) 26Az18SHC2a
(141) 12Az.1C2a	(285) 27Az1SHC4a

Table 3.5: Samples chosen for the inter-laboratory check test

How it is possible to infer from sample's code, every cherry tomato has been chosen either among winter and summer seasons in a period of time covering two years. Both Naomi cultivar and Shiren cultivar are represented. In Table 3.6 are shown other details relative to these samples.

ID n.	Sample official code	Year	Season	Date	Sample group	Farm	Salinity	Plant	Cultivar
35	3Az.3A1°a	2003	sum. 03	23/06/03	3	3	2000	A	Naomi
46	4Az.3B2a	2003	sum. 03	01/07/03	4	3	2000	B	Naomi
82	5Az.4C6a	2003	sum. 03	09/07/03	5	4	5500	C	Naomi
107	8Az.1C3a	2003	win. 04	23/12/03	8	1	5400	C	Naomi
123	10Az.1A5a	2004	win. 04	28/01/04	10	1	5400	A	Naomi
141	12Az.1C2a	2004	win. 04	10/02/04	12	1	5400	C	Naomi
198	21Az.6SHA5a	2004	sum. 04	21/06/04	21	9	3600	A	Shiren
223	23Az12ShR7/8a	2005	win. 05	15/03/05	23	12	4000	r	Shiren
257	25Az9ShB7a	2005	win. 05	29/04/05	25	9	3600	B	Shiren
279	26AZ.18SHC2a	2005	sum. 05	30/06/05	26	18	1800	C	Shiren
285	27AZ.1SHC4a	2005	sum. 05	30/06/05	26	1	5400	C	Shiren

Table 3.6: Samples chosen for the inter-laboratory check test

4 RESULTS AND DISCUSSION

NMR analysis of Cherry tomatoes

The proton spectrum of Naomi tomato juice, buffered at pH 4.0, recorded at 400 MHz is shown in Fig. 4.1. Several hundred signals are present in the spectrum, corresponding to a comparable number of tomato metabolites. The spectrum is dominated i) by the not suppressed residual peak of water; ii) by the intense singlet signal of the methyl protons of acetate which has been introduced with the buffer solution; and iii) by the resonances of the major components, such as α-D-glucose, β-D-glucose and other sugars. However, owing to the high dynamic range of the spectrometer, signals from minor components are even readily detectable (see inset in Fig. 4.1).

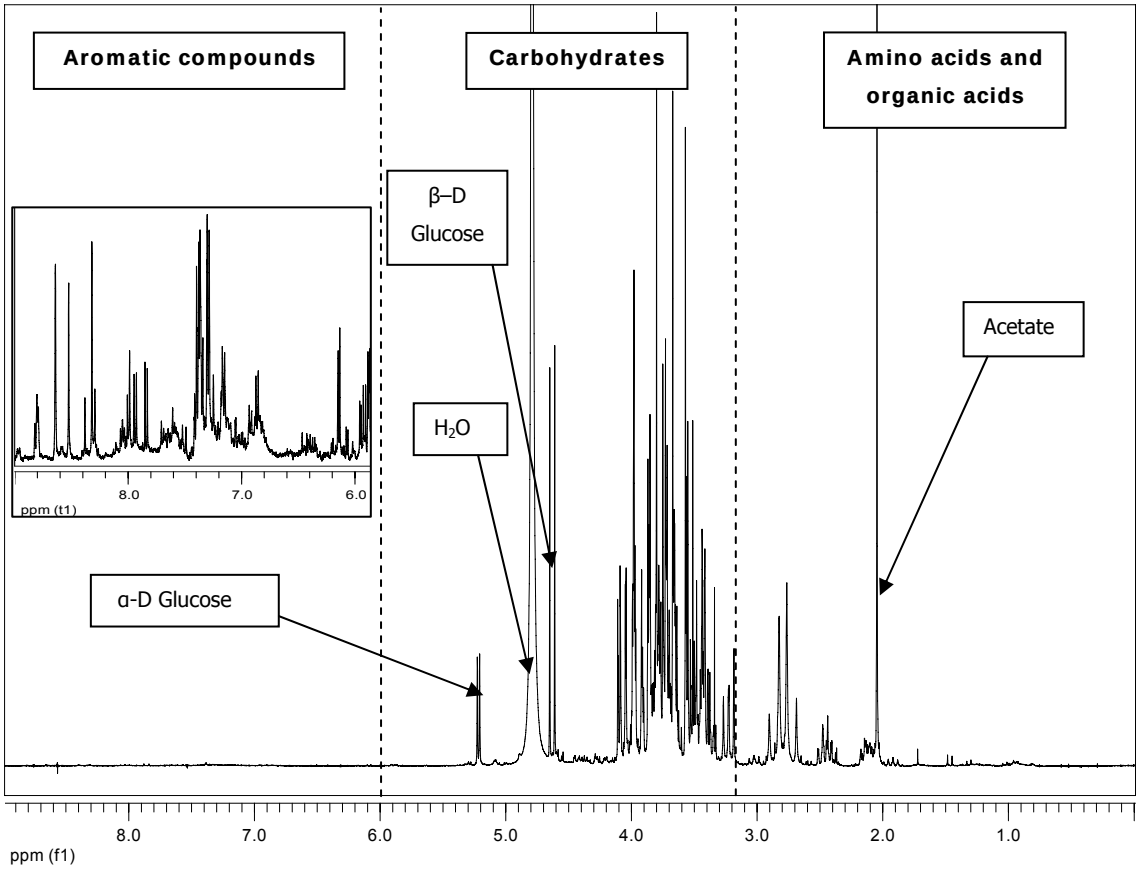


Figure 4.1: 400 MHz 1D ^{1}H spectrum (9.0 to 0.0 ppm) of a cherry tomato extract, Naomi cultivar, buffered at pH 4.0

Looking at the spectrum, the doublets corresponding to the anomeric protons of α and β D-glucose molecules are visible at both sides of the dominant residual signal of HOD, the latter at about 4.8 ppm. Also the dominant peak at 2.03 ppm, assigned to the methyl group of acetate, added to the solution as buffer system, is clearly evident. In addition, a coarse classification of all the signals is obtained by differentiating three characteristic regions in the whole spectral window: the region of the aromatic compounds (10.0 - 6.0 ppm), that one of carbohydrates (6.0 - 3.2 ppm) and that one collecting both aminoacids sidechains and organic acids (3.2 - 0.0 ppm).

The signal assignments of some molecules detected in cherry tomatoes extracts (pH 7.0), analyzed during a previous study by Segre et al., are reported in Table 4.1, with their respective multiplicity [203]. Although the chemical shifts are routinely adopted to identify molecules in NMR spectra, the actual pH of the analyzed solution is responsible of dramatic changes of their values, when ionization reactions take place. This is the case of organic acids, whose signals can span tenth of ppm depending on their ionization state, which in turn is imposed by the pH of their chemical environment. A study highlighting this behavior will be described below. Thus, only sugars may be identified in the NMR spectra of the present study by assigning signals according Table 4.1. On the contrary organic acids, including acetic acid, and aminoacids cannot be assigned following the chemical shift reported in the mentioned table, since they have different chemical shift at pH 4.0, which is the one adopted during most of the experiments of the present study.

The present chapter will discuss NMR spectra by differentiating results obtained at different magnetic field strength, because the information included in each spectrum is deeply depending on the sensitivity of the instrument, which makes available signals that are observed at higher fields but not at lower one. The effect is that the complexity of information changes, making the influence of the noise to be differently treated. The pre-processing of the spectral raw data, performed by using home made scripts written in a suitable programming language (R environment), will be discussed by referring at the algorithms described in Chapter 3.3 (Methods) and extensively reported in Appendix A. Finally, the statistical multivariate analysis performed on the pre-processed data will be discussed, still maintaining separated the data obtained at different magnetic field strength.

Table 4.1: Assignment of principal signals in aqueous extracts of tomatoes at pH 7.0

^{1}H shift	Multiplicity	Molecule
0.93	t	isoleucine
0.98	d	valine
1.00	d	isoleucine
1.03	d	valine
1.17	t	ethanol
1.26	m	isoleucine
1.32	d	threonine
1.46	m	isoleucine
1.47	d	alanine
1.89	m	γ-aminobutyrate
1.98	m	isoleucine
1.91	s	acetate
2.05	m	glutamate
2.12	m	glutamate
2.14	m	glutamine
2.26	m	valine
2.29	t	γ-aminobutyrate
2.34	m	glutamate
2.37	dd	malate
2.45	m	glutamine
2.53	d	citrate
2.66	dd	malate
2.66	d	citrate
2.68	dd	aspartate
2.80	dd	aspartate
2.87	dd	asparagine
2.95	dd	asparagine
3.00	t	γ-aminobutyrate
3.06	dd	tyrosine
3.13	dd	phenylalanine
3.18	dd	tyrosine
3.19	s	choline
3.24	dd	β-D-glucose
3.27	dd	phenylalanine
3.35	s	methanol
3.40	dd	β-D-glucose
3.41	dd	α-D-glucose
3.46	ddd	β-D-glucose
3.49	t	β-D-glucose
3.53	dd	α-D-glucose
3.54	d	β-D-fructofuranose
3.55	d	β-D-fructopyranose
3.59	m	threonine
3.59	d	β-D-fructofuranose
3.61	m	valine
3.65		α-D-fructofuranose
3.65		ethanol
3.65		α-D-fructofuranose
3.67		isoleucine
3.67	dd	β-D-fructofuranose
3.68		α-D-fructofuranose
3.70	dd	β-D-fructopyranose
3.71	td	α-D-glucose
3.71	d	β-D-fructopyranose
3.72	dd	β-D-glucose
3.75	dd	glutamate
3.76	dd	α-D-glucose
3.78	t	glutamine
3.78		alanine
3.79	dd	β-D-fructopyranose
3.79	dd	β-D-fructofuranose
3.80		α-D-fructofuranose
3.82	m	α-D-glucose
3.82	m	α-D-glucose
3.82	m	β-D-fructofuranose
3.89	dd	aspartate
3.89	dd	β-D-glucose
3.89	dd	β-D-fructopyranose
3.94		tyrosine
3.99	dd	phenylalanine
3.99	ddd	β-D-fructopyranose
3.99		α-D-fructofuranose
4.00	dd	asparagine
4.02	dd	β-D-fructopyranose
4.05		α-D-fructofuranose
4.11	m	β-D-fructofuranose
4.11	d	α-D-fructofuranose
4.25		threonine
4.29	dd	malate
4.64	d	β-D-glucose
5.23	d	α-D-glucose
6.90	d	tyrosine
7.18	d	tyrosine
7.19	t	tryptophan
7.27	t	tryptophan
7.32	dd	phenylalanine
7.37	t	phenylalanine
7.42	td	phenylalanine
7.54	d	tryptophan
7.72	d	tryptophan
8.45	s	formate

Abbreviations: s, singlet; d, doublet; t, triplet; m, complex multiplet; dd, doublet of doublets; ddd, doublet of doublets of doublets; td, triplet of doublets. Data taken from [203]

4.1 NMR ANALYSIS AT 200 MHz

In order to investigate the cherry tomato extracts without unwanted preconceptions, a preliminary study has been conducted aiming at verifying the influence of the pH of aqueous buffers on the extraction power and on the consequent quality of the information that is acquired with NMR spectroscopy. With this purpose in mind, three different system buffers were prepared for the extraction of metabolites from cherry tomato lyophilized powders, namely at pH 4.0, 7.0 and 10.0.

The overall spectrum recorded at 200 MHz resembles the one that has been already shown in Figure 4.1, recorded at 400 MHz at the same pH 4.0. The differences are noticeable only by inspecting in detail the spectrum by expanding it both in vertical and horizontal scales. In fact, the overall region subdivision is still valid, being only resolution and sensitivity affected by the magnetic field strength. The picture is also changed by working at different pH.

The spectrum of a Naomi cherry tomato aqueous extract, recorded at pH 7.0, is shown in Figure 4.2. Obviously, while most of the same main signals observed at pH 4.0 are still observable, there is not the intense signal of acetate at about 2.03 ppm.

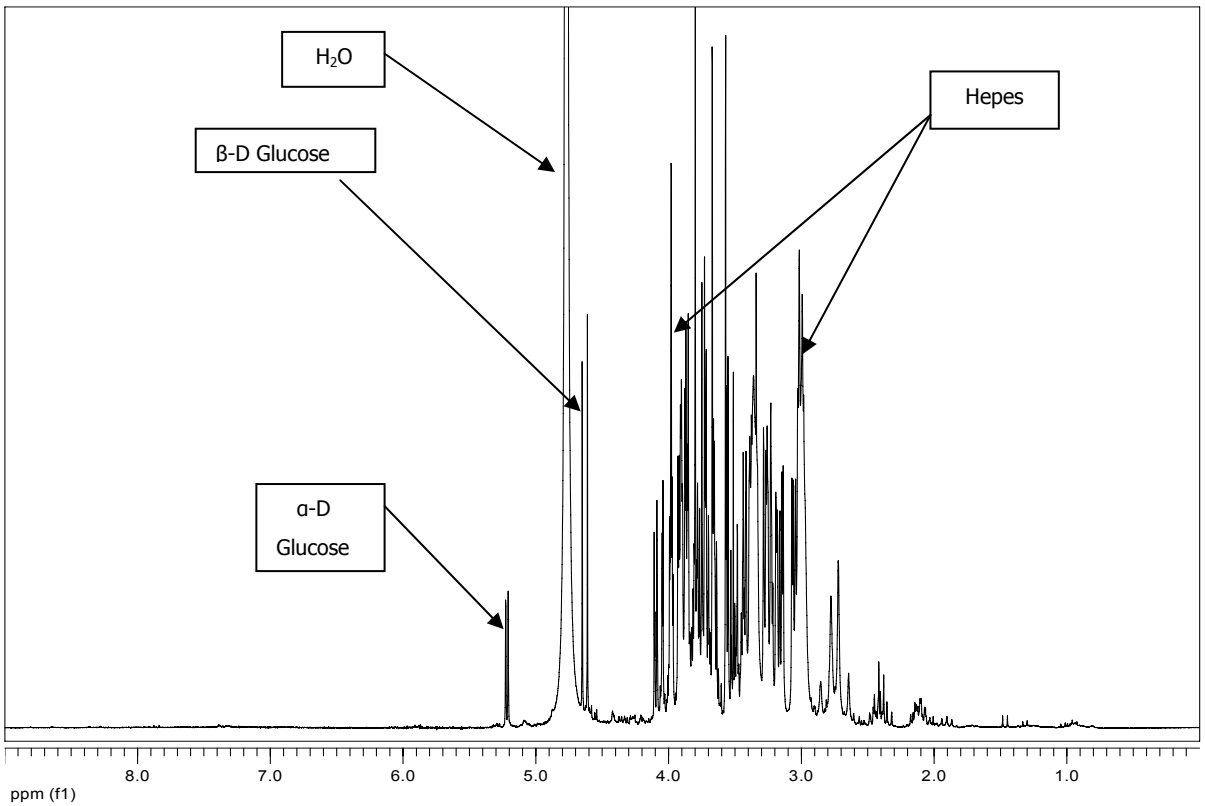


Figure 4.2: 200 MHz 1D ^{1}H spectrum (9.0 to 0.0 ppm) of a Naomi cherry tomato extract buffered at pH 7.0

Some new intense signals due to HEPES, used as the buffer system, are present in the spectrum at about 3 ppm. Such an intense multiplet signal can be easily assigned to the HEPES protons by recording a spectrum of a blank HEPES buffer solution at the same experimental conditions of the samples. The spectrum of HEPES at 200 MHz is shown in Figure 4.3, where it is possible to observe all signals assignable to its protons. Of course, such signals overlap a broad range of resonances belonging to the sample, affecting the successive chemometric analysis

In both spectra recorded at pH 4.0 and 7.0, the signals of buffer species are present, introducing other complexity to the system, although it has been found that acetate at pH 4.0 does not overlap any signal from abundant molecules. However, there are buffer systems consisting of species with only solvent exchangeable protons. For such species any other signal, besides the ones belonging to the tomato extracts, is not introduced in the proton NMR spectra.

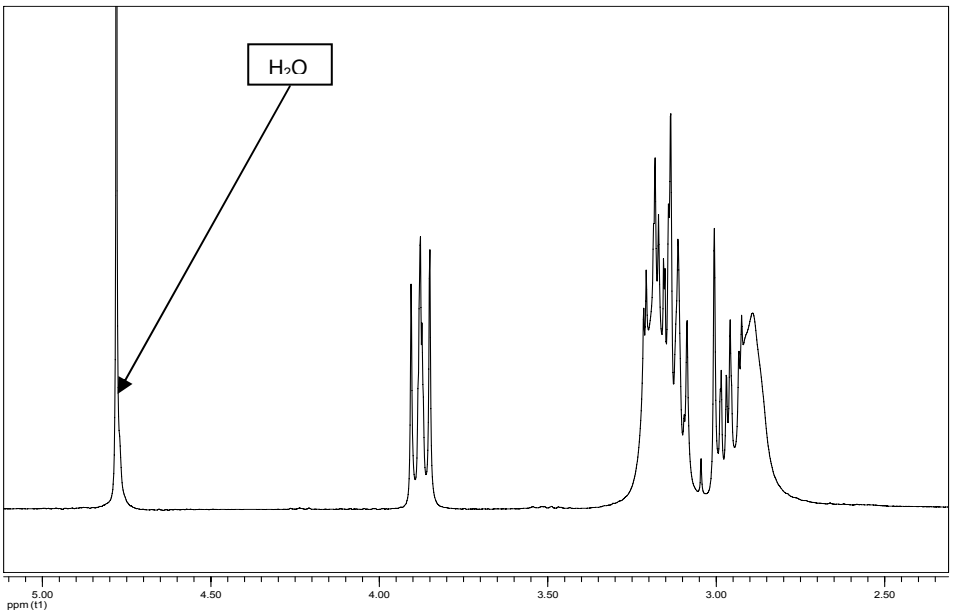


Figure 4.3: 200 MHz 1D ^{1}H spectrum (5.0 to 2.5 ppm) of HEPES buffer solution

The spectrum of a Naomi cherry tomato sample, recorded at pH 10.0, is reported in Figure 4.4. Here, the buffer signals are not present because bicarbonate is the only species of the buffer system possessing one proton, which is exchangeable with D_2O. Thus, it gives origin to a signal collapsed with that one of the residual HOD water solvent signal, anyhow present at about 4.8 ppm. For this reason, it is possible to affirm that the spectrum recorded at pH 10 represents that one referring to a tomato extract not contaminated with any other

external signal. For the same reason, since there is any substance with non exchangeable protons added at known concentration, it is impossible to select a signal as internal standard for area normalization in all spectra acquired at pH 10.0.

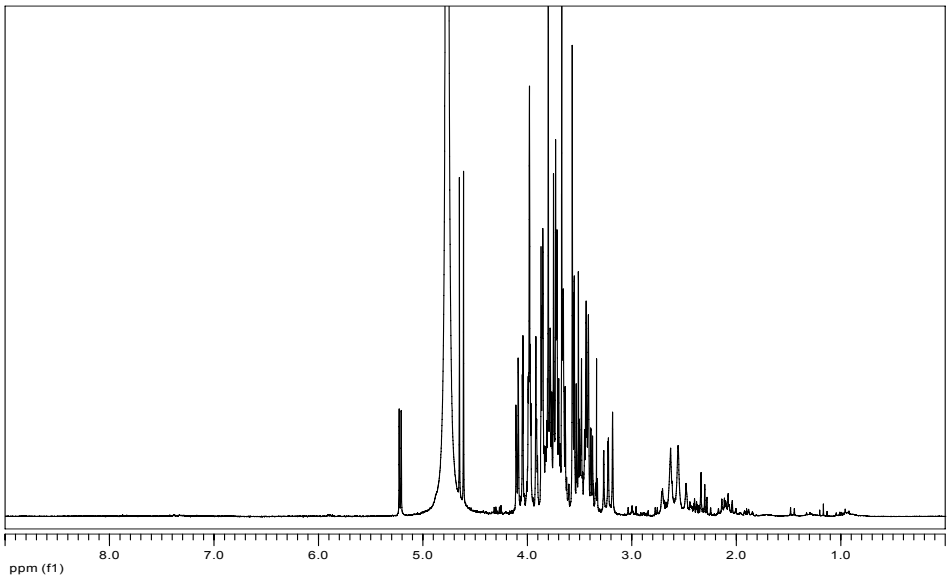


Figure 4.4: 200 MHz 1D ^{1}H spectrum (9.0 to 0.0 ppm) of a Naomi cherry tomato extract buffered at pH 10.0

By the direct comparison the three spectra recorded at the three different pH values, it is also possible to observe how signals falling on the acids region tend to shift toward higher fields when the pH value increases. This is due to the fact that many amino acids as well as the most common organic acids contain side-chains carrying functional groups interacting with the acid protons, and thus their chemical shift is affected by the pH value of the solution. In Figure 4.5 are shown the three spectra of the same cherry tomato sample analyzed at the three different values of pH that have been selected. This comparing picture permits to highlight the differences among the three spectra and also to observe the shifting of citrate signals, circled in blue in the spectrum acquired at pH 10.0, with respect to the other two spectra.

Since pH has a deep effect on the overall distribution of signals in the spectrum, a straight correlation among spectra recorded at different pH is not easily reachable. The absence of some signals in an NMR spectrum is, indeed, not necessarily the consequence of a reduced extraction power of the buffer solution, but instead the shift of its position in a different, not recognized, position. For this reason the study had to be definitely conducted at only one pH value, possibly one similar to the physiological pH, which ranges between 3.8 and 4.5 in cherry tomato fruits [204].

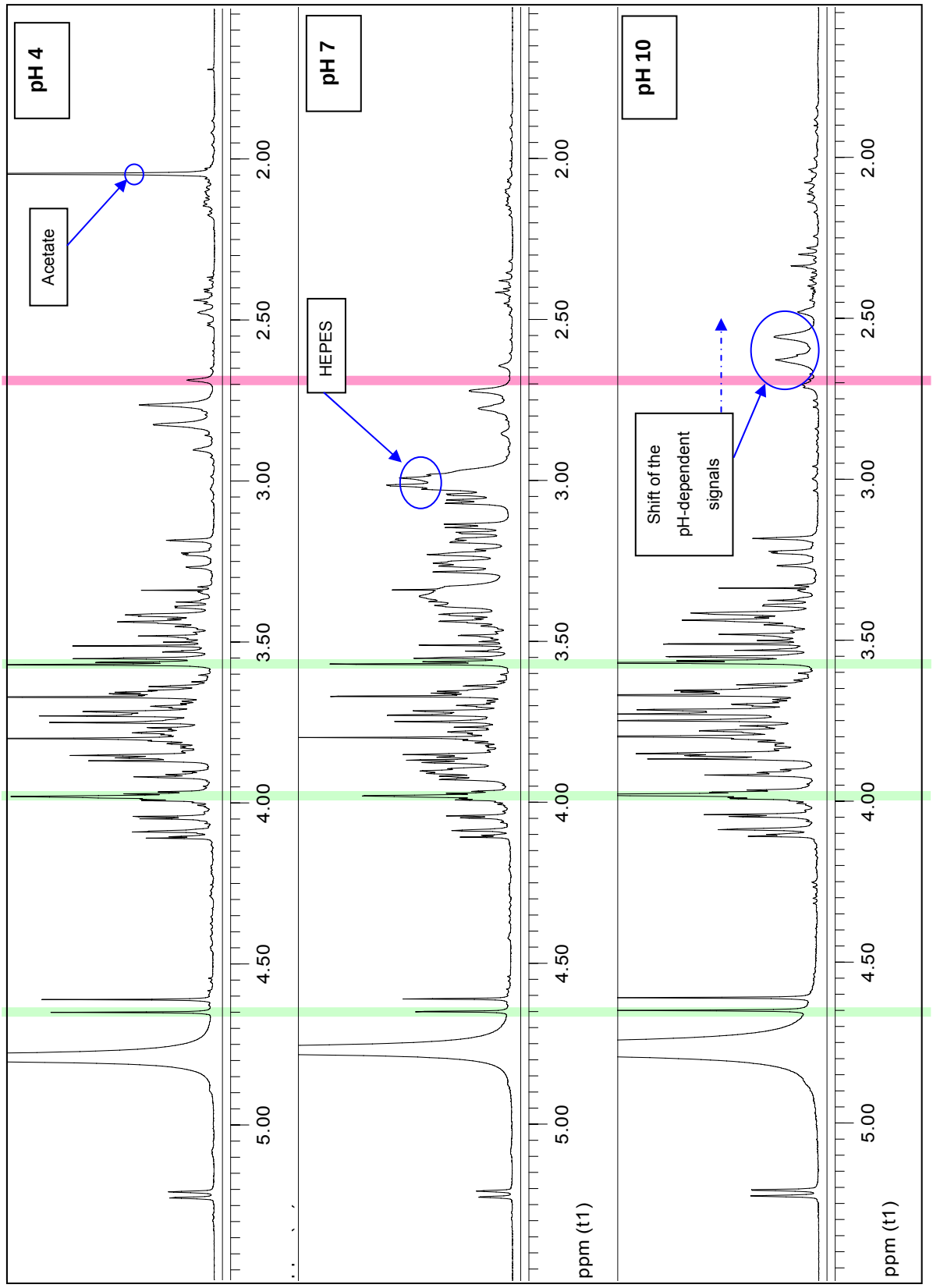


Figure 4.5: 200 MHz 1D ^{1}H spectra (9.0 to 0.0 ppm) of the same tomato sample extracted and analyzed at three different pH values with highlighting of signal alignment (for sugars signals in light green) and misalignment (for citrate signals in light fuchsia).

In this way it is expected that the chemical state of metabolites in the extracts is maintained as much as possible similar to that one naturally found in cherry tomato fruits. For this reason, all chemometric analysis discussed along the present thesis work will refer to the aqueous extracts with pH kept at about 4.0.

4.1.1 Chemometric data processing

271 cherry tomato samples have been extracted and analyzed by using a 200 MHz NMR Bruker spectrometer.

Once all the FIDs recorded at 200 MHz have been processed with Mestre-C, the spectra have been exported in ASCII files and arranged in a comprehensive spectral data matrix (271 x 6820) suitable for the subsequent chemometric analysis. Each row of such a raw data matrix represents one single spectrum and each column represents the same data point along all spectra. A simple data plot of 10 out of the 271 superimposed matrix rows restricted to the data points corresponding to the spectral region ranging from 1.80 to 4.30 ppm is shown in Figure 4.6.

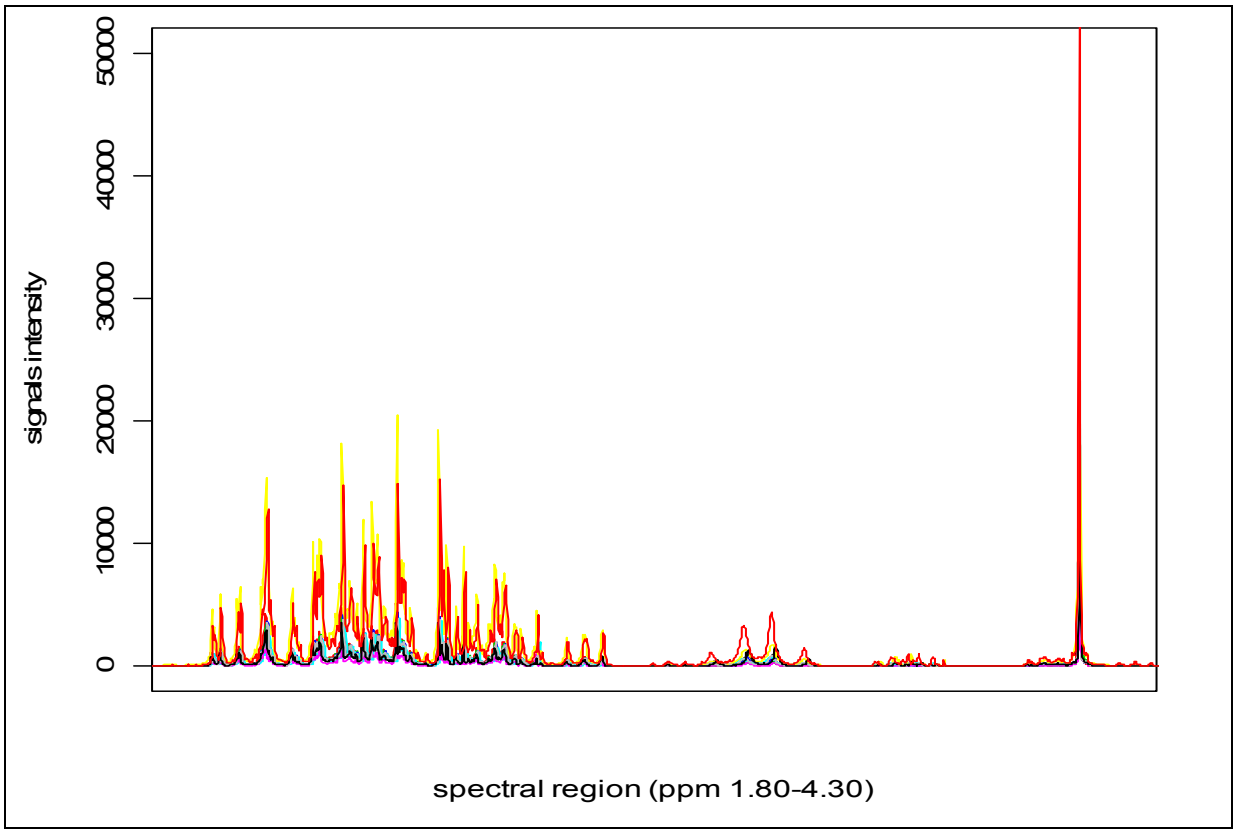


Figure 4.6: Superimposing of 10 rows of the whole 200 MHz raw data matrix.

At first glance, it appears that this spectral region shows a large horizontal shift of some signals, although recorded at the same pH value. Moreover, there are relevant vertical scale variations due to instrumental factors such as the probe temperature or the magnetic field homogeneity and/or casual errors in the samples preparation. These preliminary

observations become clear by zooming on some topic regions of the spectrum, some of which are reported in Figures 4.7-4.9.

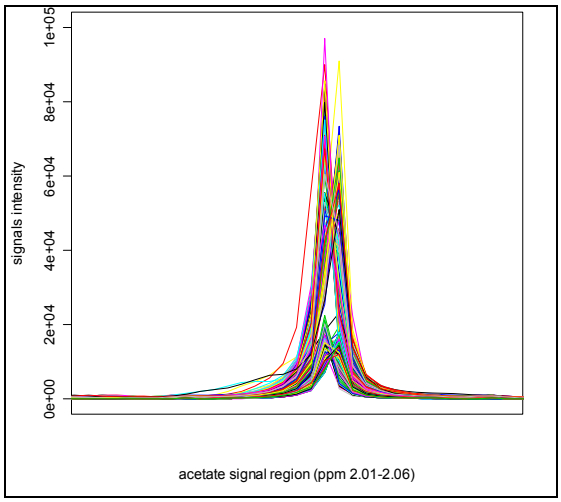


Figure 4.7: Acetate signal from 40 spectra.

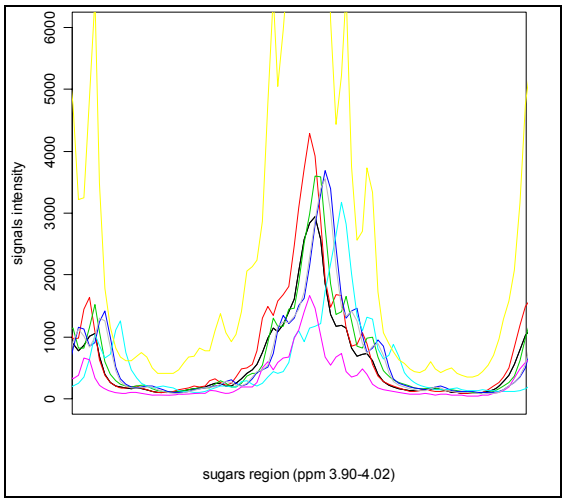


Figure 4.8: Sugars region signals.

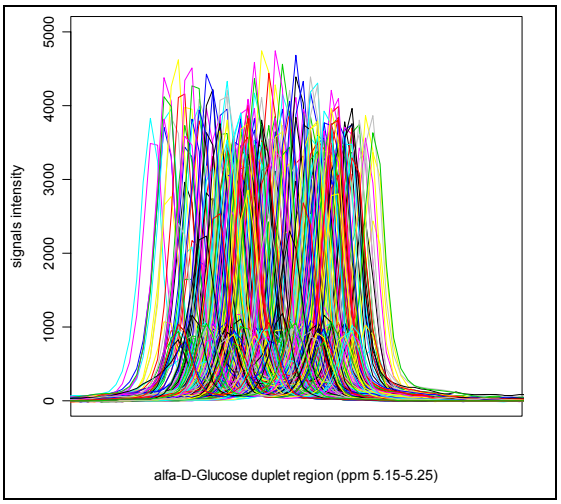


Figure 4.9: α-D-glucose doublet region

In these figures three different spectral regions, that need some explanation, are shown. Figure 4.7 depicts the superimposed signals of acetate in 40 samples. Since these signals are substantially aligned on the horizontal axis, it results clear that raw data have been well referenced over this signal in the previous FID processing step. Despite the good horizontal alignment, these signals are affected by a large variation along the vertical scale. Differences of the same order of magnitude are observed also for all the signals when an array of spectra restricted to the sugars' polyalcohol region is plotted, as evident in Figure 4.8. The α-D-glucose region of all 271 acquired spectra, relative to the anomeric doublet, reveals shifts along the horizontal axis as well as vertical scale variations (Figure 4.9). Such problems

presented by raw spectral data make them not reliable for successive chemometric analysis. Therefore further steps have been implemented in the developed series of algorithms, obtaining a homogenous dataset in order to minimize them.

In order to manage the vertical scale errors, a normalization algorithm has been developed by considering the signal of acetate being constant by intensity, since its concentration has been kept the same for all extracts at pH 4.0. In fact, acetate, which is present in every extract at the same concentration (100 mM) as buffer system, can be considered as an internal standard for the acquired spectra. For this reason, an algorithm able to normalize data over the intensity of such a signal in each spectrum has been written and applied, thus obtaining the results shown in Figures 4.10-4.11

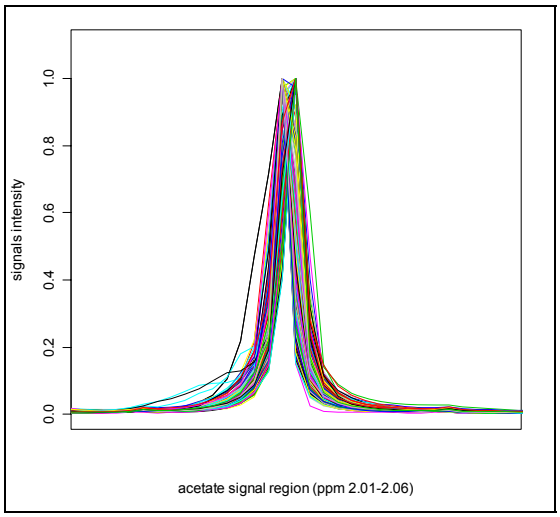


Figure 4.10: Acetate signal in 200 MHz after normalization step.

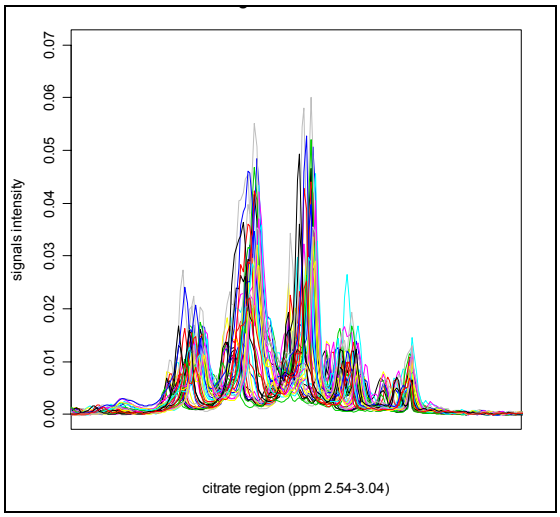


Figure 4.11: Citrate signal in 200 MHz after normalization step.

In Figure 4.10, the acetate spectral region of 40 spectra has been superimposed after normalization on the acetate methyl signal. Such a signal has been perfectly normalized by the algorithm and presents now the same intensity in each spectrum. As a consequence of this step, also other signals have been normalized but, since they can belong to substances present in different concentration in the different tomato samples, they can still show a relevant variation on its signals intensities. This is the case of the citrate signals of the same 40 samples shown in Figure 4.11. However, although acetate signal is now both normalized and horizontally aligned, the latter condition does not hold for other signals, such as those of citrate. This is due to the pH dependence of the acetate signal that is affected by small pH variations, causing a shift of the position of these signals in different samples, at a diverse extent with respect to that one observed for citrate. Therefore, it appeared clear that such a signal, even if intense and always present in all samples at similar concentrations, cannot be

used as chemical shift reference and, since no other internal standard has been used, another suitable signal have to be found to perform this step.

The requirements of a signal to be suitable for horizontal alignment (the chemical shift referencing) have been individuated in the following ones: i) the signal must be isolated and not affected by other adjacent signals; ii) it must be present at high and very similar concentration in all samples; iii) and, of course, it must be not affected by the pH fluctuations. For aqueous tomato extracts, analyzed by the 200 MHz spectrometer, such a signal has been individuated in the α-D-glucose doublet.

A best fit algorithm has been opportunely written to perform an alignment of such a signal over all spectra of the dataset, by sliding them in either left or right directions, until the best horizontal alignment with respect to the α-D-glucose doublet of a target spectrum is reached. The reference spectrum has been previously selected among those calibrated on acetate by identifying the one showing the anomeric signal in the position closer to its average position over all spectra.

The effects of the algorithm over not pH dependent signals are shown in Figures 4.12 and 4.13. Both the signals of the α-D-glucose doublet region and of another generic sugars' region appear perfectly aligned demonstrating the good results achievable by the algorithm.

In these figures, all the 271 acquired spectra are represented, showing that a significant difference in the intensities of the signals still exists. This, of course, is principally due to differences among samples on metabolites concentration but still another cause can be individuated: since the samples are not homogeneously freeze-dried, it happens that water concentration can vary a lot among samples determining fluctuations on the concentration of the metabolites extracted on the final sample submitted to NMR analysis. Because a constant weight of 100 mg of freeze-dried powder is withdrawn from each original milled sample, the amount of dry powder actually taken, depending on water content, can considerably affect the concentration of the extracted metabolites. Furthermore, in case of non homogeneous food matrices, and this is the case of freeze-dried cherry tomatoes powders, it is possible to collect, by chance, different anatomic parts of the fruit, such as seeds, peel or pulp with a lower content of extractable matter. Also this casual error can influence the actual concentration of metabolites on the resulting NMR spectrum, with the consequence of variations in the signals intensities of the molecules present in the mixture.

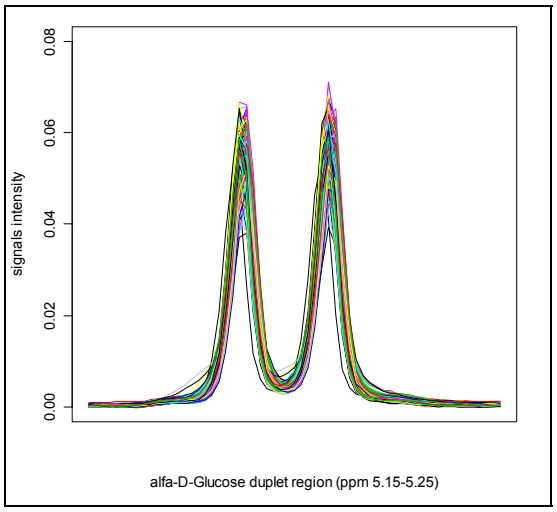


Figure 4.12: α-D-glucose doublet signal at 200 MHz after CS normalization step.

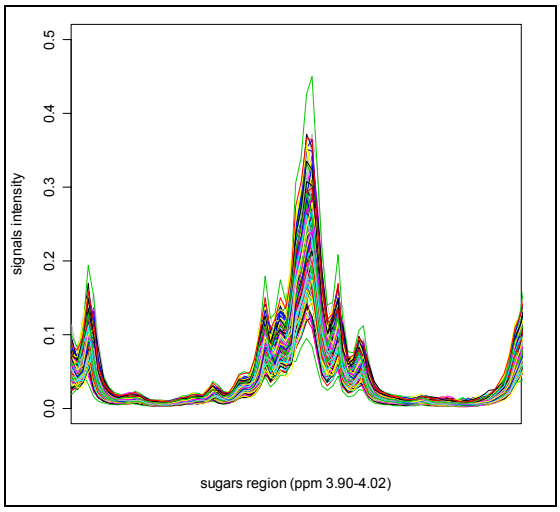


Figure 4.13: Sugars signal at 200 MHz after CS normalization step.

In order to evaluate the oscillations along the horizontal axis, the superimposition of all spectra aligned on the anomeric glucose signals are shown in Figures 4.14-4.15, revealing the effect of such an aligning algorithm on the pH dependent signals.

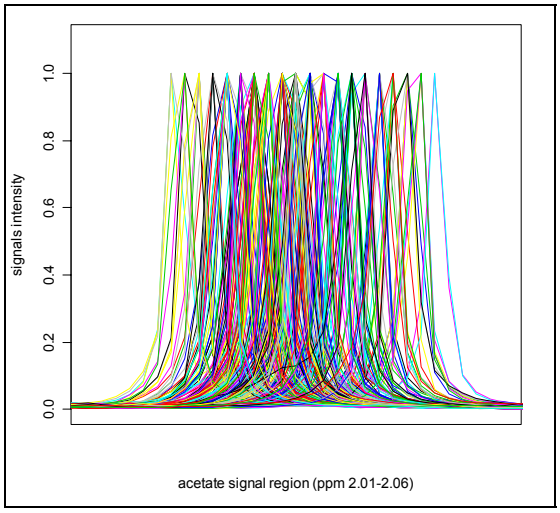


Figure 4.14: Acetate signal at 200 MHz after normalization and alignment on anomeric doublet of glucose.

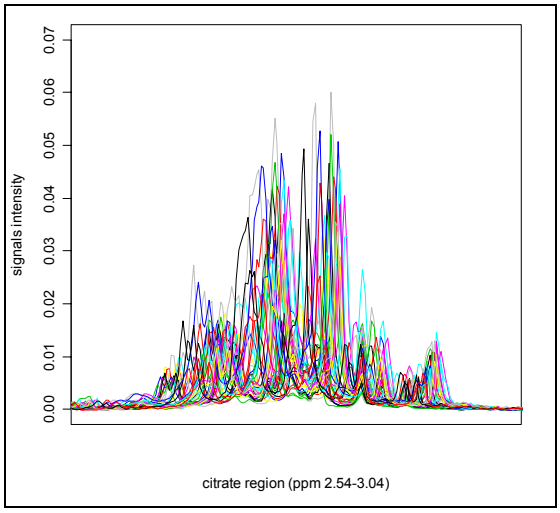


Figure 4.15: Citrate signal at 200 MHz after normalization and alignment on anomeric doublet of glucose.

The way how the acetate signal is shifted over all samples is shown in figure 4.14. Now, it results clear that such a signal is not suitable for chemical shift referencing. Furthermore, the citrate signals of 40 samples, that still present horizontal misalignment, are shown in Figure 4.15: this is due to the fact that also citrate signals are pH dependent but not in the same way as for acetate one.

As far as it concerns the correction of vertical scaling errors and of artifacts due to pH dependent shifts of signals, two different approaches have been applied, namely based on the "integral to a constant sum" and on the "binning" algorithms, respectively.
For solving the vertical scaling problems, often recurring on applications of chemometrics NMR data of food matrices, several methods have been developed and published in literature, like those briefly described in the Introduction. The most common and used way to solve such a problem is to normalize each spectrum to a constant integral of the whole area. An algorithm has been written in R language to perform this type of normalization, obtaining the results shown in Figures 4.16-4.19

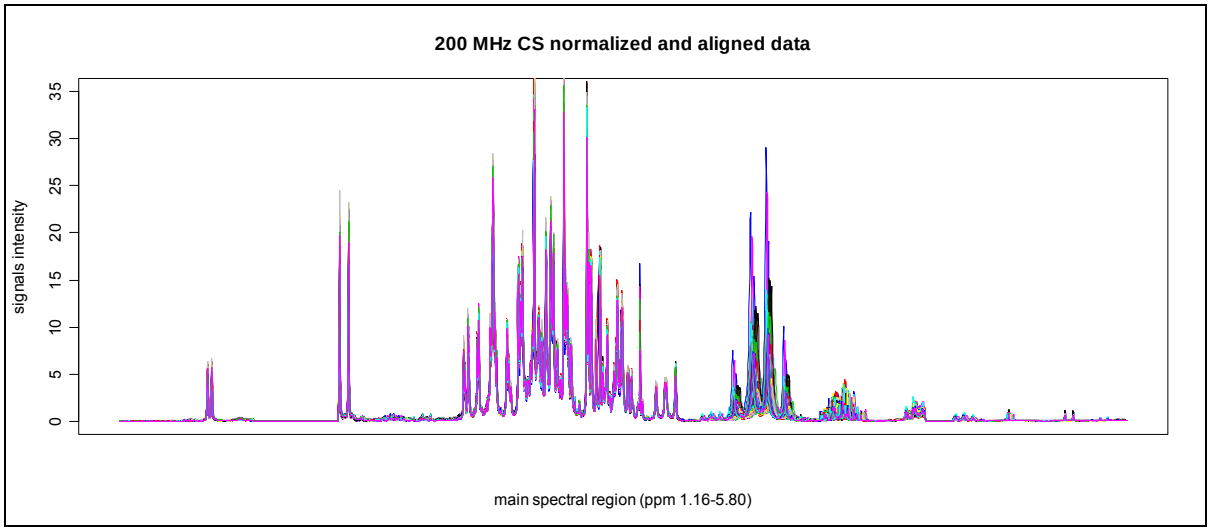


Figure 4.16: Superimposing of all the 271 rows of the whole 200 MHz aligned and CS normalized data matrix.

It is possible to capture the resulting homogeneity of data by plotting the whole spectral dataset on the same picture. For this reason, 271 spectra corresponding to all analyzed samples are plotted together in Figure 4.16. Water and acetate signals have been removed before normalization to a constant sum, in order to avoid their main influence on the calculation performed. In this way only the signal belonging to cherry tomato metabolites are responsible of the data normalization and it is possible to obtain a good consistency of data as it is shown in Figure 4.17. The α-D-glucose doublets of all the 271 spectra appear now definitely most similar in their intensities as expected for samples of tomato fruits collected at the same grade of ripeness.

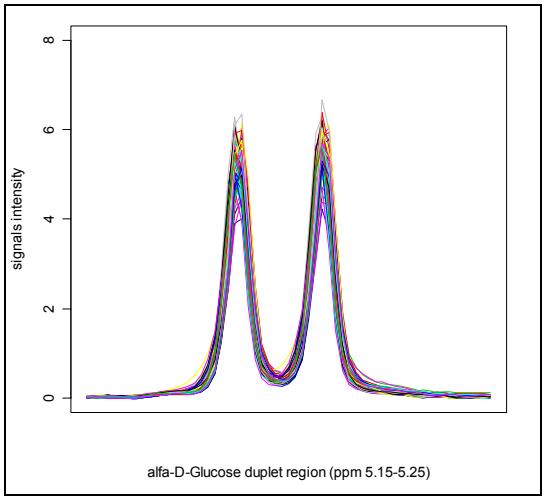


Figure 4.17: α-D-glucose doublet signal at 200 MHz after CS normalization step and horizontal alignment.

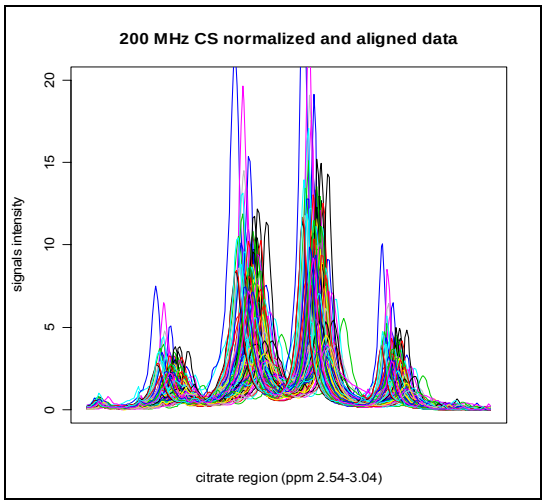


Figure 4.18: Citrate signal at 200 MHz after CS normalization and alignment on anomeric doublet of glucose.

The citrate AB double doublet presents positional misalignment, related to small pH differences, but also significant vertical oscillations, even though the normalizing algorithm was already applied, because the actual concentration of such a metabolite varies largely, also depending on the harvesting season as pointed out afterwards (Figure 4.18).

While the vertical changes due to differences among metabolite concentrations are searched by the chemometric tools for finding correlations and for expressing the actual variability within the sample population, the horizontal shift of the same signal in different spectra is an artifact introduced by very small pH errors in sample preparation. In order to minimize such errors, a binning algorithm has been applied to all spectra.

As previously described, this method permits to better handle the dataset by performing the reduction of variables, by dividing the whole spectral window into integral regions of defined wideness (bins). This operation performs also a further alignment of the variables, still misaligned after signals alignment, by forcing the inclusion of shifting signal in the same bin of all spectra. While this method permits a better data handling by condensing the data points in fewer integrals, it causes a loss of definition of spectral features. The latter represents the main disadvantage of this chemometric step but, if bins are narrow enough, the chance to have only one or few signals within each integral region is still high. By applying the binning algorithm, written in R language, each spectrum has been reduced from 6800 data points to 200 bins, each 0.05 ppm wide, representing the significant variables associated to each sample.

This is graphically represented in Figure 4.19, where the region of the most intense signals of a single NMR spectrum is placed over a number of yellow bars representing the integral bins, each one collecting few signals.

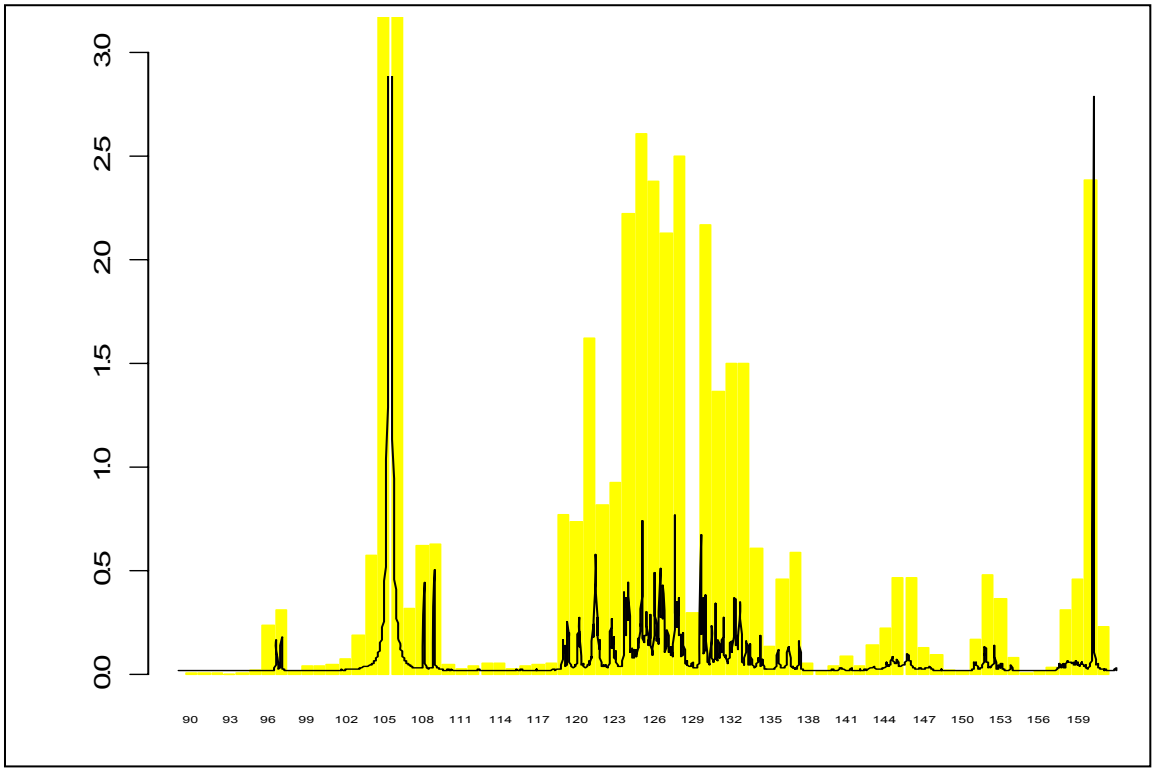


Figure 4.19: A 200 MHz NMR spectrum of a cherry tomato and its corresponding bins.

Not all the bins appear in Figure 4.19, because only a partial region of the whole spectrum is shown. It is worth noting here to observe that water signal falls within the bins numbered from 101 to 107, while the acetate signal falls in the bins numbered 160 and/or 161, depending on its pH dependent shift among samples.

Since these signals do not represent metabolites of the sample, their corresponding bins have been removed from data matrix before successive multivariate analysis obtaining a final data matrix of 271 samples (rows) for 191 variables (columns).

It is this pre-processed dataset that has been firstly submitted to Principal Component Analysis (PCA) after centering and scaling the whole data matrix.

The results obtained from this first PCA are shown in Figure 4.20, representing the plots of the third and of the second principal component scores (PC3 vs. PC2), where samples are labeled with their own cataloguing ID (See appendix B) in order to easily identify outliers.

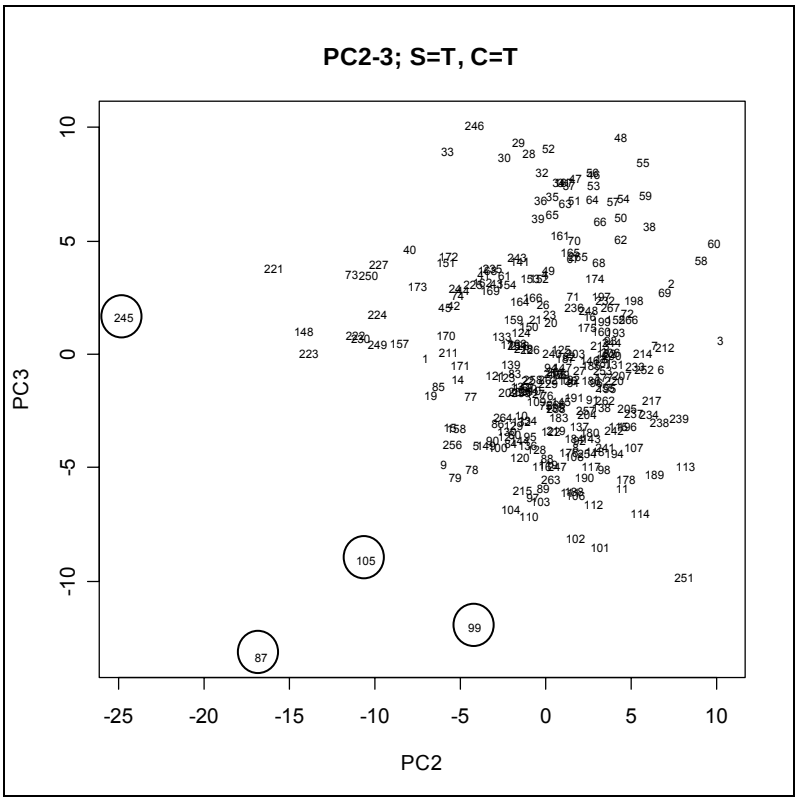


Figure 4.20 A: PC2-PC3 scores plot of all the samples labeled with ID number.

The samples evidenced by circles in the PC plot are outliers and this behavior is found also for other PC scores (data not shown here). A careful inspection of their spectra highlighted some anomalous features, such as the presence of broad hump or spurious signals, not present in any other sample, probably due to contamination. Such samples have been removed from the dataset. After four outliers have been removed, the new data matrix, consisting of 267 samples x 191 bins, has been resubmitted to PCA.

All next PC plots, along the thesis, will be presented without the use of numerical identifiers but, for the sake of clarity, only dots will afterward be adopted. Just the color labeling will be often employed as, for example, in Figure 4.21A, where samples are colored according to their geographical origin: red for samples of **Naomi** cultivar coming from the region of Pachino, orange for samples of **Shiren** cultivar coming from Pachino, black for aspecific randomly collected samples coming again from Pachino (hereafter **Aspecific**), magenta for samples coming from **Licata** (Sicily), green for samples coming from **Sabaudia** (Lazio) and brown for samples of uncertain origin that were purchased at the general **Markets**.

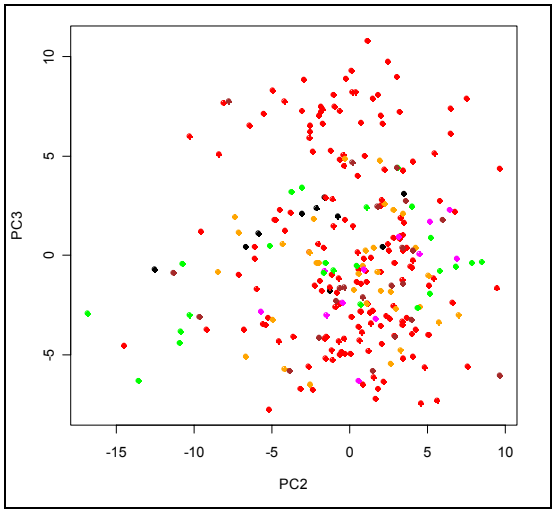


Figure 4.21 A: PC scores plot of all samples marked with colors corresponding to origin

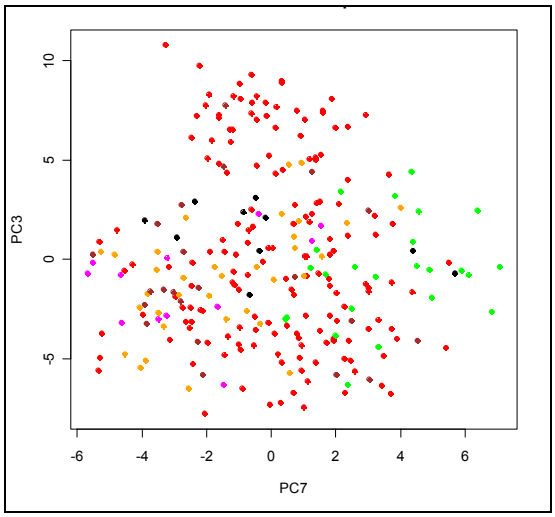


Figure 4.21 B: PC scores plot of all samples marked with colors corresponding to origin

It is immediately clear that PC scores do not permit a classification of the samples of this dataset in agreement with their geographical origin. In Figure 4.21A all classes are spread all over the PC spaces without any noticeable tendency. Searching for best combination of PC spaces able to distinguish among classes, the best result that can be achieved with the pair PC3/PC7 (shown in Figure 4.21B), where some tendency to cluster is visible for samples (green colored) relative to tomatoes harvested in the geographical region of Sabaudia.

However, a large amount of samples originating from Sabaudia and Licata are overlapped with samples coming from Pachino, the latter well spread all around.

Before admitting the inapplicability of PCA to capture the diversity of the geographical origin of cherry tomatoes as a function of their NMR data, it is necessary to individuate other sources of variance which are responsible of hiding the searched geographical effect. The seasonality of harvesting has been considered one of the most imputable effects, since the sampling duration was protracted along three years. The dots shown in the two PC plots represented in Figures 4.21 have been colored differently according to their harvesting season: blue for **winter** and red for **summer**. The results of such a new point of view of PCA analysis are shown in Figures 4.22 A and 4.22 B.

Even considering that it is impossible to clearly separate, from the meteorological point of view, winter from summer season, because of weather transitions during autumn and spring, and of the different behavior of clime along three years, an unequivocal seasonality nevertheless emerges from both PC scores plots. A line has been drawn in both PC plots of Figures 4.22, in order to better point out the separation between winter and summer samples.

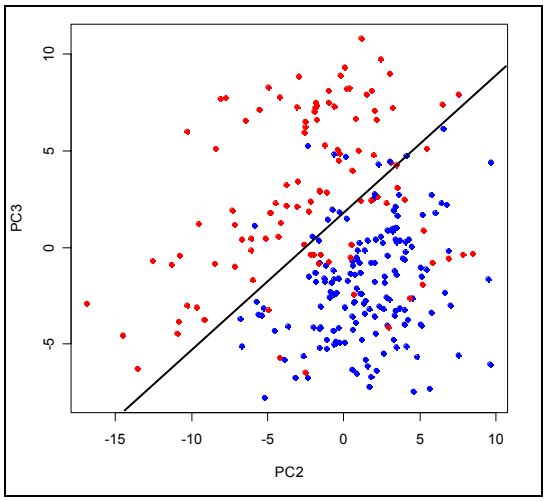


Figure 4.22 A: PC scores plot of all samples marked with colors corresponding to season

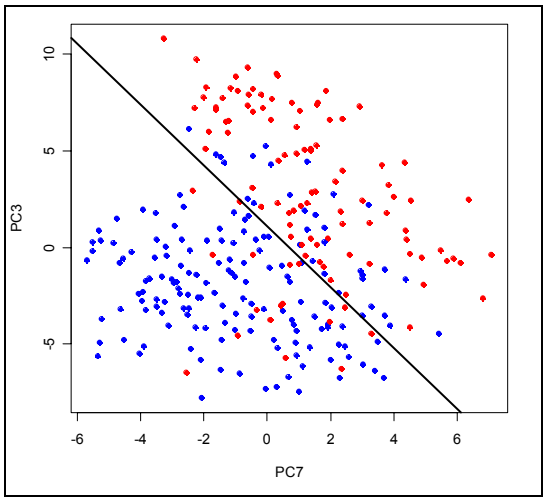


Figure 4.22 B: PC scores plot of all samples marked with colors corresponding to season

The distinction between the two groups of samples harvested during winter or summer has been obtained by first applying the chemometric approach to our data. However, also a spectral comparison may give a preliminary suggestion of such distinction, as emerging by the inspection of Figure 4.23, where the citrate region of all 271 spectra is plotted. After the normalization and the alignment steps described above were applied, and by coloring spectra according to their harvesting season, it is clear a tendency of the signals of such a substance to be considerably less intense in summer.

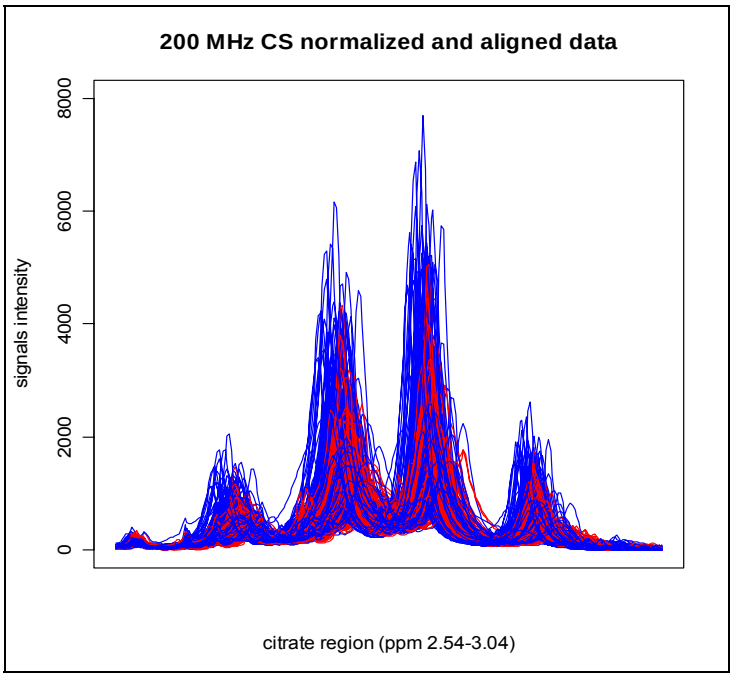


Figure 4.23: Superimposing of all the spectra in the citrate region

Thus, the right consequence of the precedent analysis is to split the whole dataset in two different data matrices, the one relative to summer and the other to winter, in order to avoid the seasonality effect in the chemometric analysis.

4.1.2 Processing of winter samples

The winter dataset of the spectra recorded at 200 MHz consists of 155 out of the 267 spectra previously accepted for statistical analysis after removal of outliers. The same algorithms have been applied to such a new dataset following all the previously described preprocessing steps.

The plot of the citrate region recorded in all summer samples is reported for completeness in Figure 4.24. The signals of such a metabolite, even if belonging to samples of the same harvesting season, still present a certain degree of vertical scaling oscillations, besides the expected horizontal misalignment associated to small pH variability. The amount of citrate, obviously, has also a variability which is independent from seasonality, even thought the latter has a larger effect.

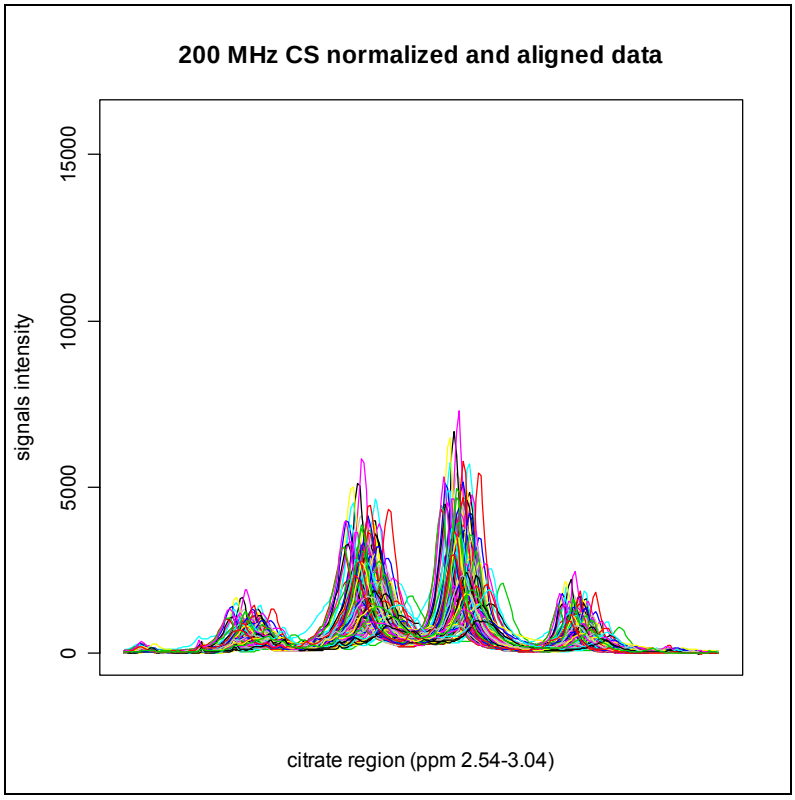


Figure 4.24: Superimposing of all the winter spectra in the citrate region

The results of the Principal Component Analysis conduced over the NMR data bins of the 155 winter samples are shown in Figures 4.25 and 4.26. The best combinations of PC scores that maximize the separation in the PC space among classes related to the geographical origin have been selected. The samples, in all PC plots describing the winter samples, are labeled with the following color coding: Pachino **Naomi** samples are blue, Pachino **Shiren** ones are cyan, **Licata** ones are magenta, and **Market** ones are brown. Finally, the **aspecific** samples, i.e. those harvested in Pachino without following any standard protocol for sampling, are represented as orange dots. Both PC3-4 and PC3-8 scores plots do not reveal any marked separation among classes, although a certain tendency to group may be vaguely seen for samples cultivated in Licata or belonging to the Shiren cultivar.

The samples extracted from Naomi tomatoes cultivated in Pachino appear spread all over the PC plot and overlapping with the other classes of samples. It is worth remembering here that samples from Licata are the only ones coming from a region different from Pachino. However, the growing pedoclimatic conditions of Licata are very similar to those findable in the adjacent Pachino area; this fact makes the differentiation among these two classes of samples PCA very difficult to be accomplished.

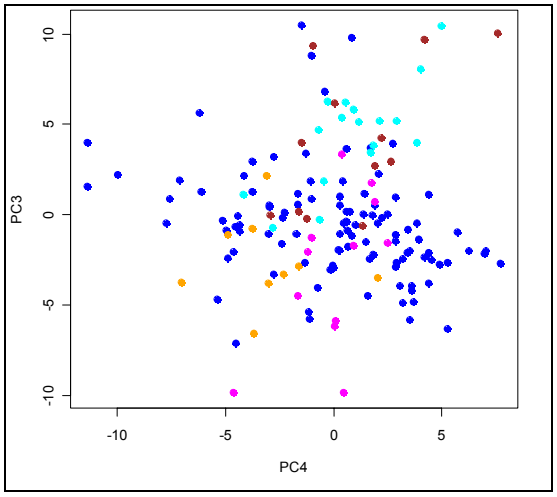


Figure 4.25: PC 3-4 scores plot of winter samples

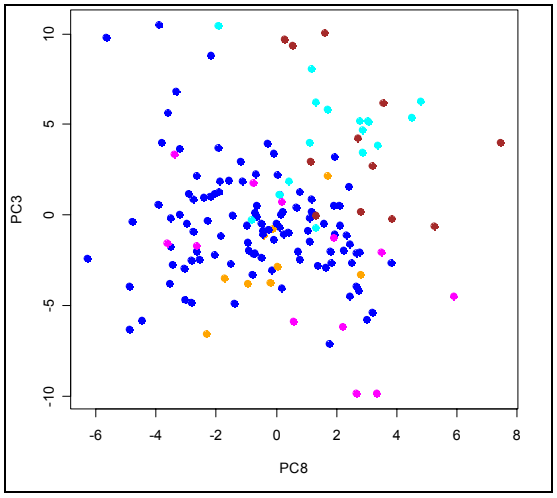


Figure 4.26: PC 3-8 scores plot of winter samples

Another class of samples, not associable to the Naomi cultivar from Pachino, comprises extracts of tomatoes purchased in the general markets whose origin was not declared, and are represented as brown dots. These samples are fairly separated from those of the Naomi Pachino but not from the samples of the Shiren cultivar (cyan colored dots). Orange dots representing aspecific samples, which have not satisfied any standard requirement for sampling, appear distributed among other samples of Pachino, where they also come from.

As already found for citrate, there are signals in the spectra that are subjected to chemical shift changes depending on the actual pH value of the extraction system. This variability constitutes an unwanted source of variation which is minimized but not completely removed from dataset. A further step in the direction of eliminating pH dependence in spectra consists in identifying those bins which vary their integral by only changing the pH of the solution, and by removing such variant bins from each row of the dataset matrix. A study directed towards such a goal, and described in Section 4.2.1, allowed us to remove 22 bins from the dataset, thus reducing the set to 169 variables (pH-free dataset). A new PCA analysis over such a reduced data matrix (155 x 169) has been performed obtaining the results shown in Figures 4.27 and 4.28.

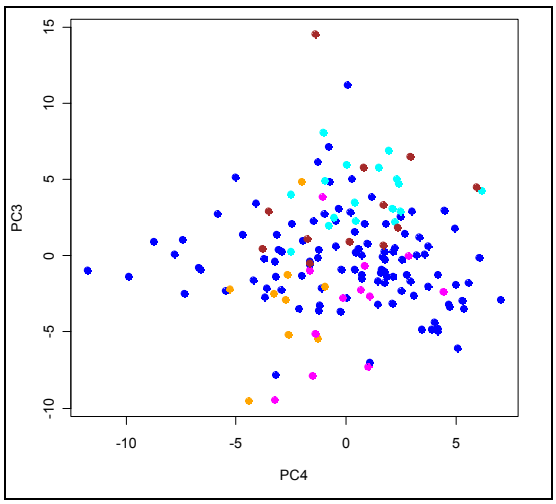


Figure 4.27: PC 3-5 scores plot of winter samples after removal of pH-dependent bins

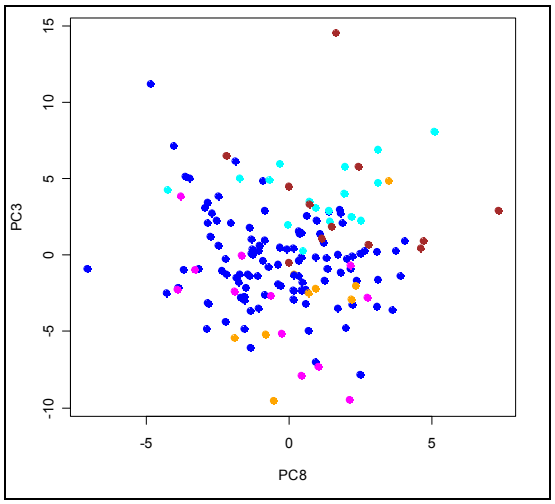


Figure 4.28: PC 3-8 scores plot of winter samples after removal of pH-dependent bins

As already found, the aspecific samples continue to be spread over all the PC space calculated on the pH-free dataset. Thus, they were removed from the winter dataset in order to avoid the effect of introducing samples not satisfying the standard of production although originating from the protected Pachino area of production. Thus, further chemometric analysis was performed on a reduced 146 samples x 191 bins matrix.

Although the gained improvement is not dramatic, it can be noticed that a lower group spreading and a tendency to color ticking can be seen. For this reason, this step has been considered useful, and the corresponding dataset has been utilized for further chemometric analysis. The tendency to cluster just seen, authorizes one to perform a Linear Discriminant Analysis over all the selected data.

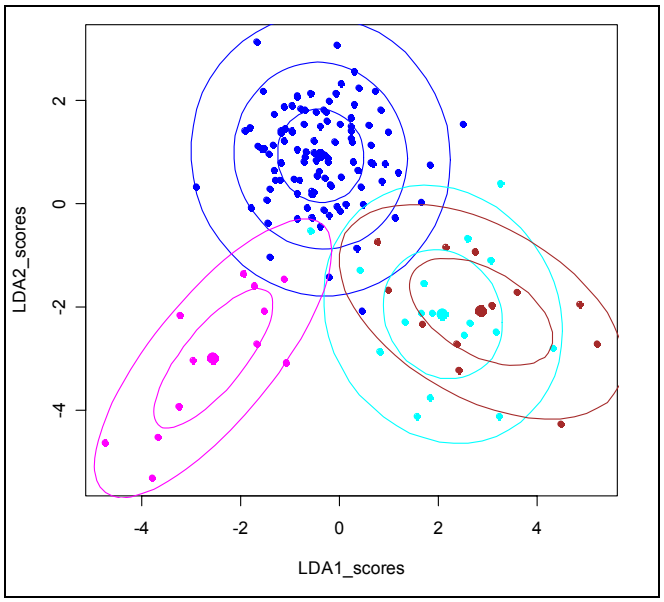


Figure 4.29: LDA 1-2 scores plot of winter samples with Mahalanobis distances

LDA analysis has been performed on the PC rotation matrix (loadings) obtained by PCA analysis of the winter data matrix, already purged of aspecific samples. LD 1-2 scores plot of the analyzed data matrix is shown in Figure 4.29. The PC space describing 90% of the total variance, corresponding to the first 19 PCs, has been involved in LDA analysis. The choice of 90% will be explained later in Chapter 4.2.4.

Classification performed by LDA appears to be quite good since the groups are clearly distinguishable. The overlap among classes still persists by observing data on a plane of two LDs. However, by looking at all the combination of the first three LDs, emerges a clear separation among all classes (Figures 4.30 and 4.31). A 3D scatterplot of all the three LD dimensions is shown in Figure 4.32.

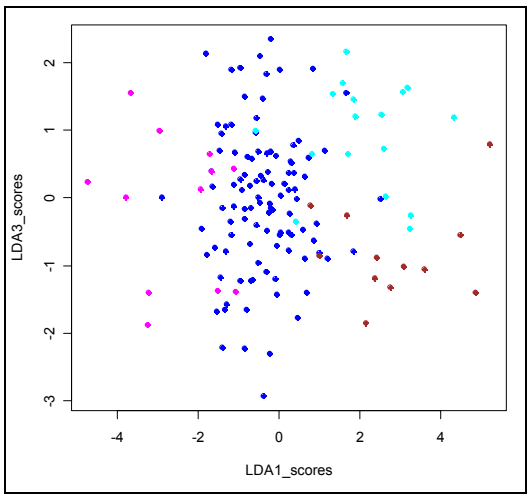


Figure 4.30: LDA 1-3 scores plot of winter samples

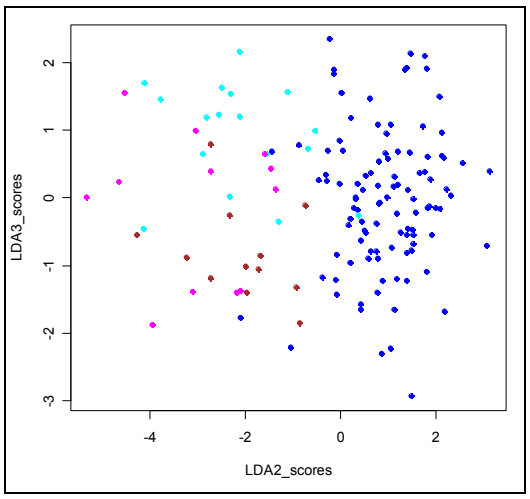


Figure 4.31: LDA 2-3 scores plot of winter samples

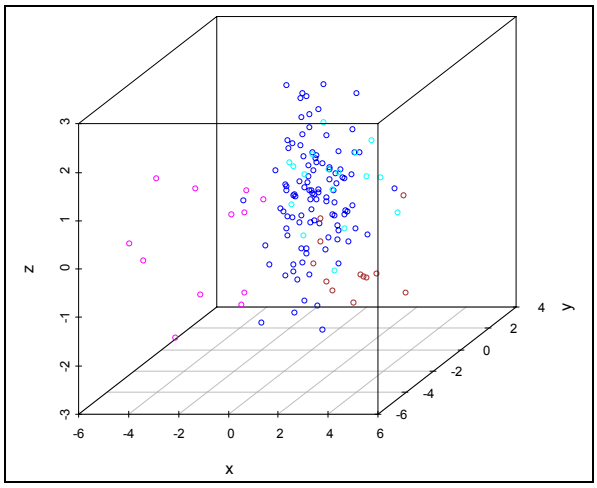


Figure 4.32: 3D plot of the three LD scores dimensions

In order to give an indicative estimation of the discriminative power of LDA, the ellipses corresponding to the Mahalanobis distances of each group is shown in Figure 4.29. The number of ellipses is limited to the highest possible that do not include the centroids of the other groups, except for Shiren and Market ones whose centroids are almost superimposed. It is noticeable that only samples originating from Pachino and belonging to Naomi cultivar have up to three Mahalanobis distances that do not overlap the other centroids.

According to such a criterion for class assignment, it is possible to observe that samples from Licata and those from Pachino are substantially separated. At this point, the "*a posteriori*" introduction of aspecific samples in the LDA space resulted in a large dispersion of the corresponding dots, thus justifying their preliminary removal.

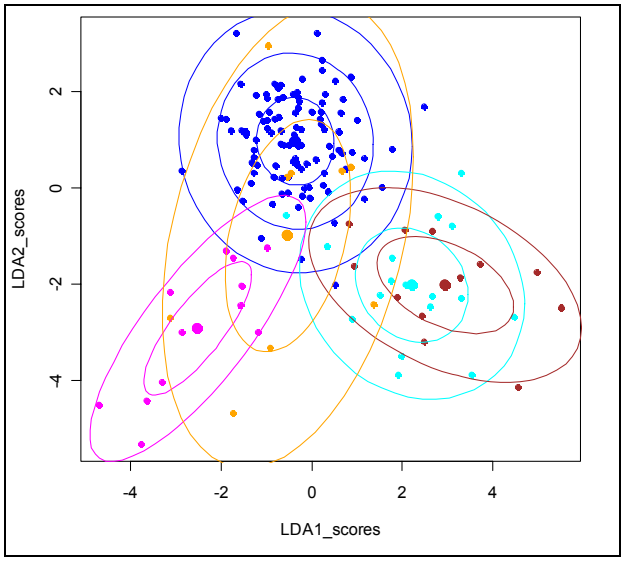


Figure 4.33: LD1-2 scores plot of winter samples with aspecific samples and Mahalanobis distances

In fact, Mahalanobis distances of aspecific samples intersect all the other classes determining a relevant ambiguity of such samples that does not permit their classification in a regional group (Figure 4.33).

One of the most used tests aiming at calculating the capability of LDA to predict the belonging of an unknown sample to one of the preconstituted categories is the lean-one-out test (LOO). By performing such a test on the 155 winter samples dataset, it has been found that 35 samples have been wrongly assigned to one of the five categories, giving a value of predictivity as low as 77.4%. For completeness of information, the category including the aspecific samples is forced to be considered a category although sampling conditions do not ensure a standardized quality. For this reason, the LLO test has been again conducted, without the 9 aspecific samples, giving a predictivity increased to 82.9%.

A study focused on salinity of irrigation water used for cultivation in the Pachino area has been performed in the LD spaces in order to inspect whether this parameter can influence the sample discrimination (Figure 4.34). All samples coming from areas external to Pachino have been grey colored. Samples from Pachino area have been colored according to their corresponding salinity parameters: purple has been used for samples irrigated with water having high salinity (≥ 5000 μS/cm of electrical conductivity) and violet has been used for samples irrigated with water having low salinity (≤ 4000 μS/cm).

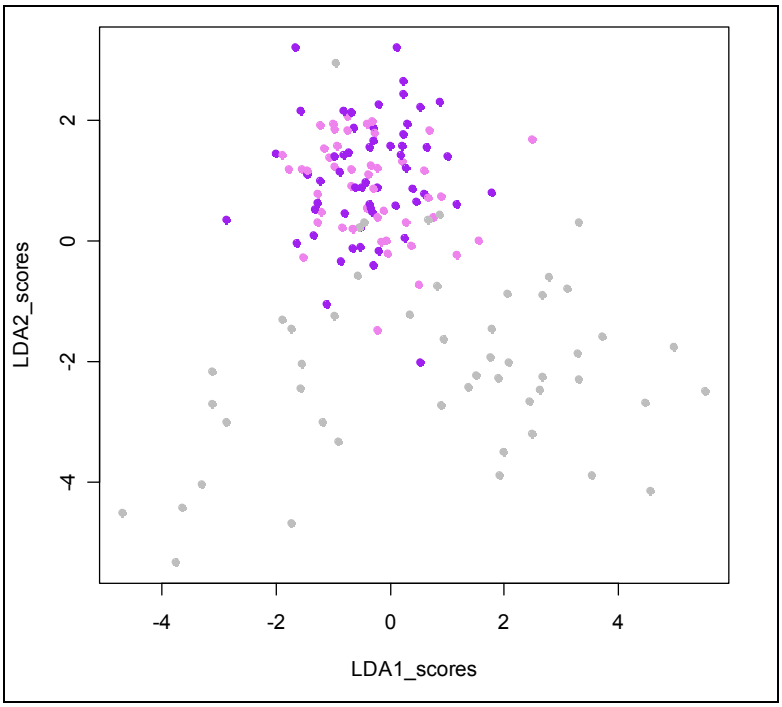


Figure 4.34: LD1-2 scores plot of winter samples with samples from Pachino area color labeled according their salinity.

As evidenced by inspection of Figure 4.34, samples belonging to different salinity grade do not show any tendency to group, thus demonstrating that this parameter does not interfere with the sample classification based on the production standards adopted by Pachino's farmers.

4.1.3 Processing of summer samples

The summer NMR dataset recorded at 200 MHz consists of 112 out of the 267 spectra previously accepted for statistical analysis, after removal of outliers. The same algorithms have been applied to such a new dataset following all the previously described preprocessing steps.

The results of the Principal Component Analysis conducted on the NMR data acquired on 112 summer samples are shown in Figures 4.35 and 4.36. Also for summer samples the best combination of PC scores spaces have been selected in order to maximize the separation among classes related to their geographical origin. The samples are color labeled according with their category: Pachino **Naomi** samples are red, Pachino **Shiren** ones are orange, **Sabaudia** ones are green, and **Market** ones are brown. Finally, the **Aspecific** samples, i.e. those harvested in Pachino without following any standard protocol for sampling, are represented as black dots.

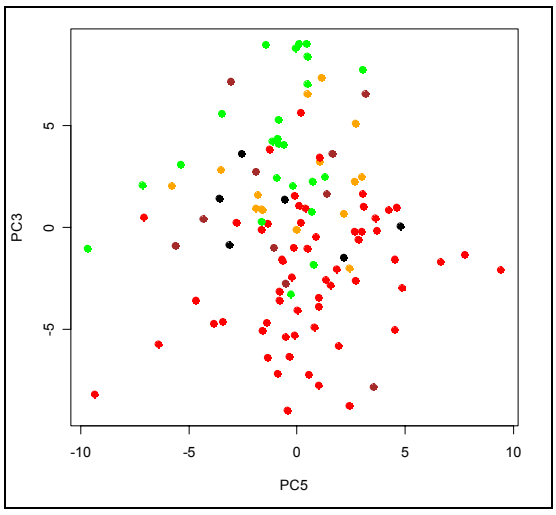


Figure 4.35: PC 3-4 scores plot of summer samples

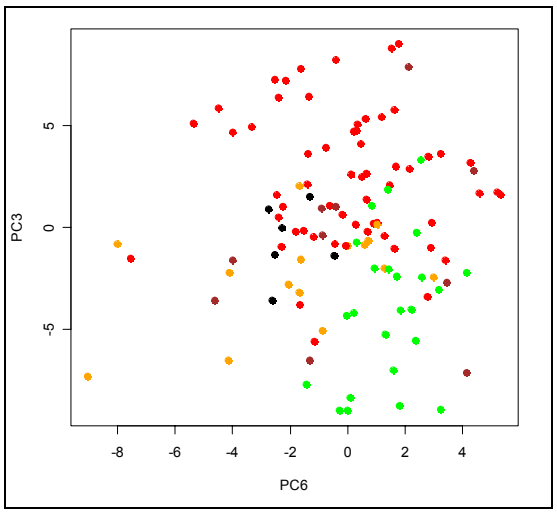


Figure 4.36: PC 3-8 scores plot of summer samples

Once again, aspecific samples are preferable to be discarded because of their random dispersion over the PC plots. Both PC3-5 and PC3-6 scores plots shows a better separation

among classes, if compared with those observed for winter samples. Above all, green samples, representing cherry tomatoes cultivated in Sabaudia (Lazio), appears substantially separated from the other samples along the PC3 axis.

Also for the summer data matrix, the bins containing the most pH dependent signals, identified through the same specific study conducted for the summer samples, have been removed from the dataset, thus reducing the involved variables from 191 to 169. PCA analysis conducted on such a reduced data matrix (112 samples x 169 bins) has been performed obtaining the results shown in Figures 4.37 and 4.38.

Once again, the obtained improvement is not dramatic, although it can be noticed a lower spreading of group colors. Therefore, the corresponding dataset has been utilized for further chemometric analysis.

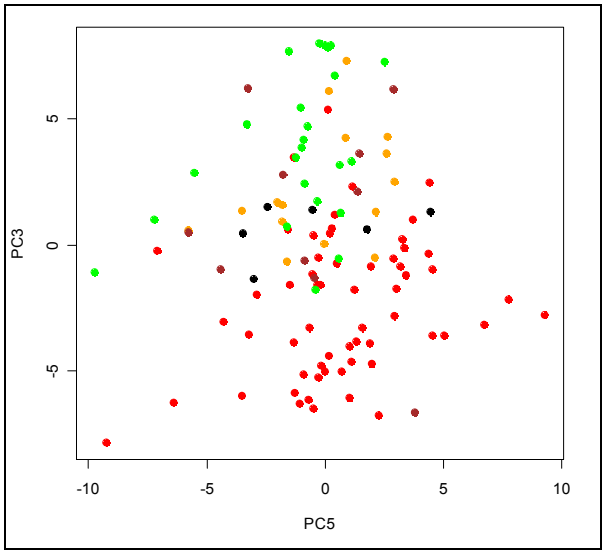


Figure 4.37: PC 3-5 scores plot of summer samples after removal of pH-dependent bins

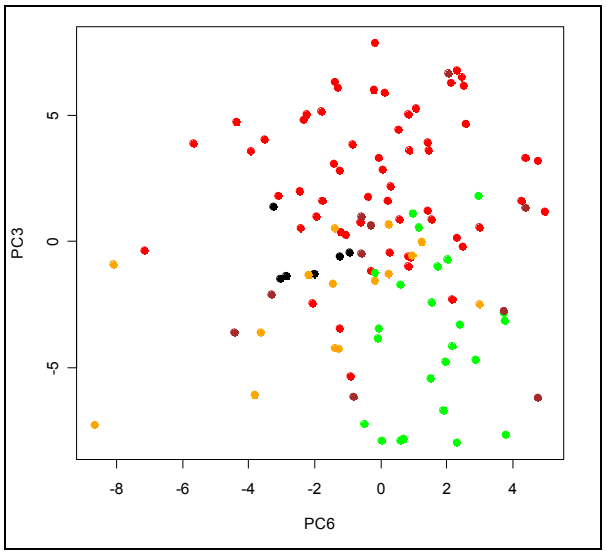


Figure 4.38: PC 3-8 scores plot of summer samples after removal of pH-dependent bins

LDA analysis has been performed on PC (loadings) rotation matrix obtained by PCA analysis of the data matrix constituted by the pH independent bins without aspecific samples.

By observing the Mahalanobis distances ellipsis in the space of the first two LDA scores, whose plot is shown in Figure 4.39, it is possible to outline some discrimination, especially for those samples that originate from areas with climatic conditions different from those of Pachino as observed for cherry tomatoes harvested in Sabaudia. As for winter samples at 200 MHz, these result has been obtained selecting only the PC spaces describing 90% of the total variance. In such a case only 17 of the 169 variables have been kept and used for LDA.

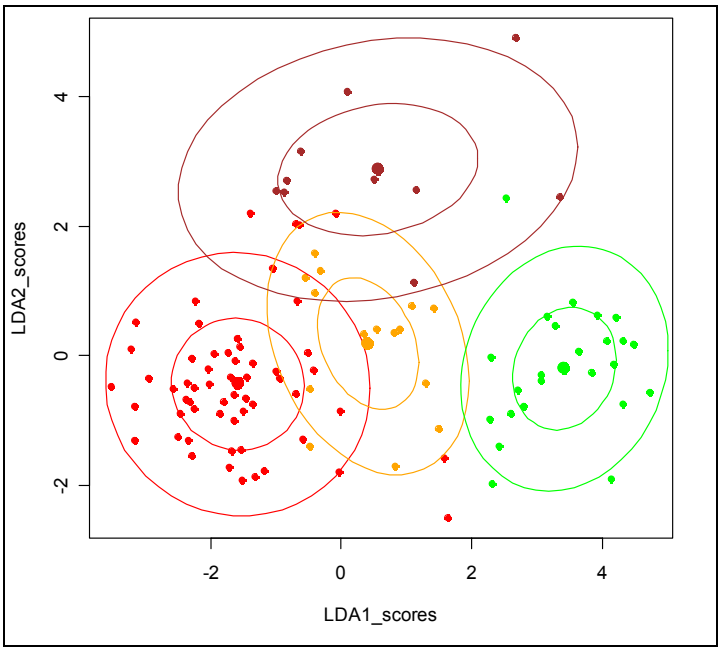


Figure 4.39: LD1-2 scores plot of summer samples with Mahalanobis distances

However, a clearer classification is even observable if also the space described by the third LD score it is taken into account (Figures 4.40 and 4.41). A 3D scatterplot of all the three LD dimensions is also reported in Figure 4.42. In this 3D space, it is easier to distinguish colored clouds representing classes of samples that appear sufficiently separated.

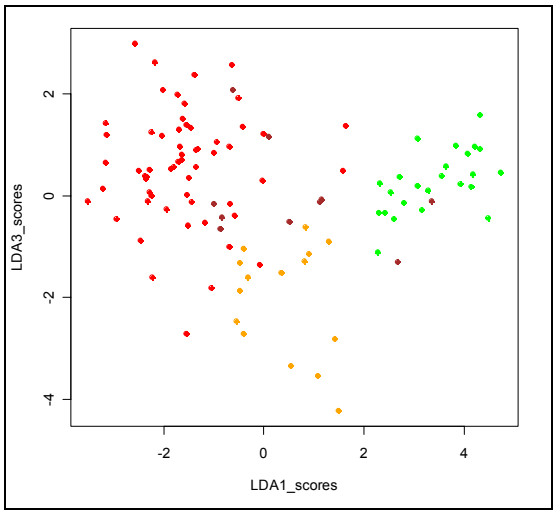


Figure 4.40: LDA 1-3 scores plot of summer samples

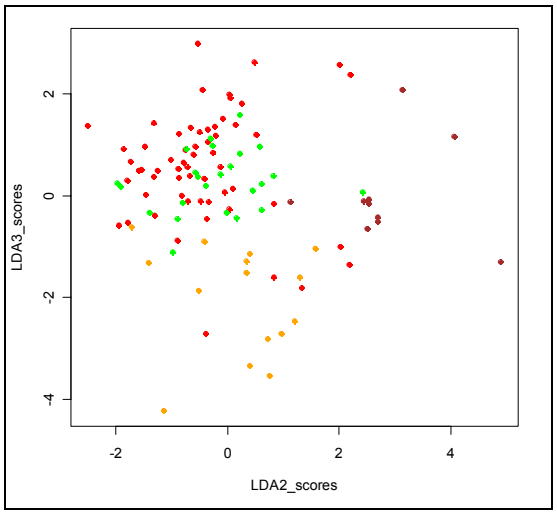


Figure 4.41: LDA 2-3 scores plot of summer samples

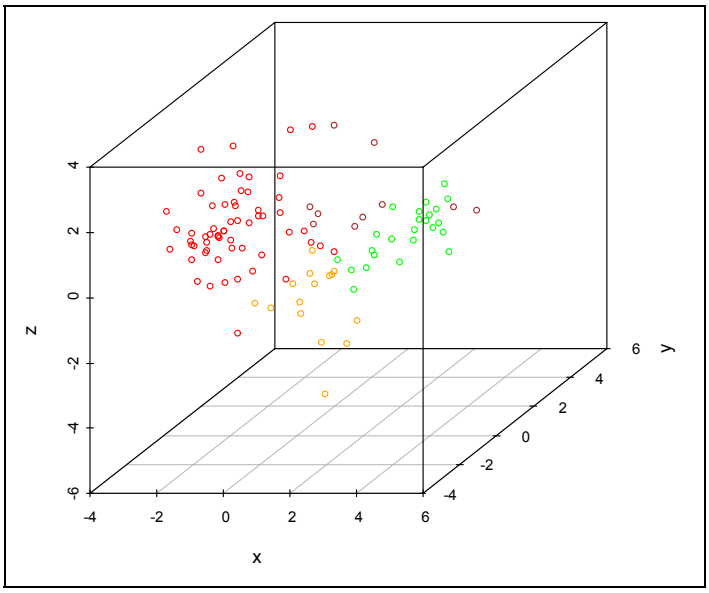


Figure 4.42: 3D plot of the three LD scores dimensions

Also for summer samples, the "*a posteriori*" introduction of aspecific samples in the LDA space resulted in a large dispersion of the corresponding dots, thus justifying their preliminary removal (Figure 4.43).

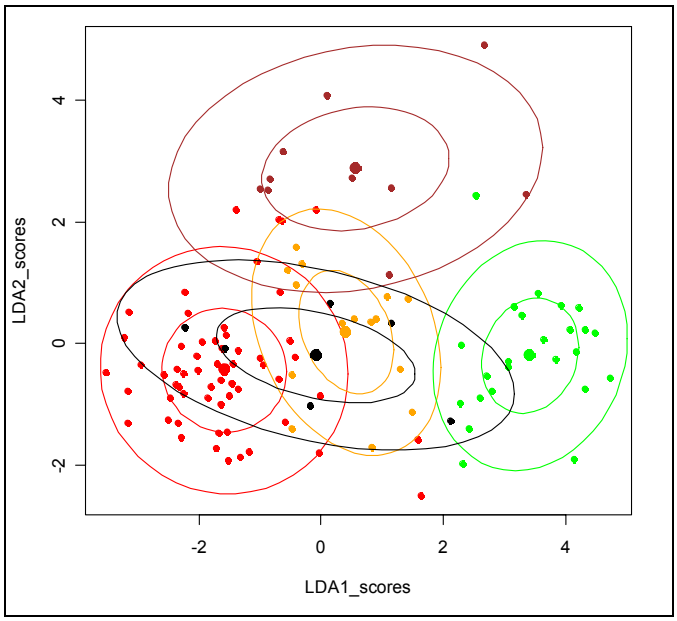


Figure 4.43: LD1-2 scores plot of summer samples with Mahalanobis distances represented and aspecific samples included

By performing the lean-one-out test (LOO) on the 112 summer samples dataset, it has been found that 19 samples have been wrongly assigned to one of the four categories, i.e. excluding the 6 aspecific samples, giving a value of predictivity as low as 82.1%. The latter

value is of the same magnitude of the one found for the winter samples (82.9%), although slightly worsen, confirming the pictorial idea that the categories of winter samples are better distinguishable than those defined for the summer ones.

Also for the summer sample, a study focused on the effect of the salinity of water used to irrigate samples cultivated in the Pachino area has been performed in LD. All samples not produced in Pachino area have been grey colored, while samples from Pachino have been colored according with their corresponding salinity parameters (Figure 4.44): purple has been used for samples irrigated with water having low salinity value (≤4000 μS/cm of electrical conductivity) and violet has been used for samples irrigated with water having high salinity value (≥ 5000 μS/cm).

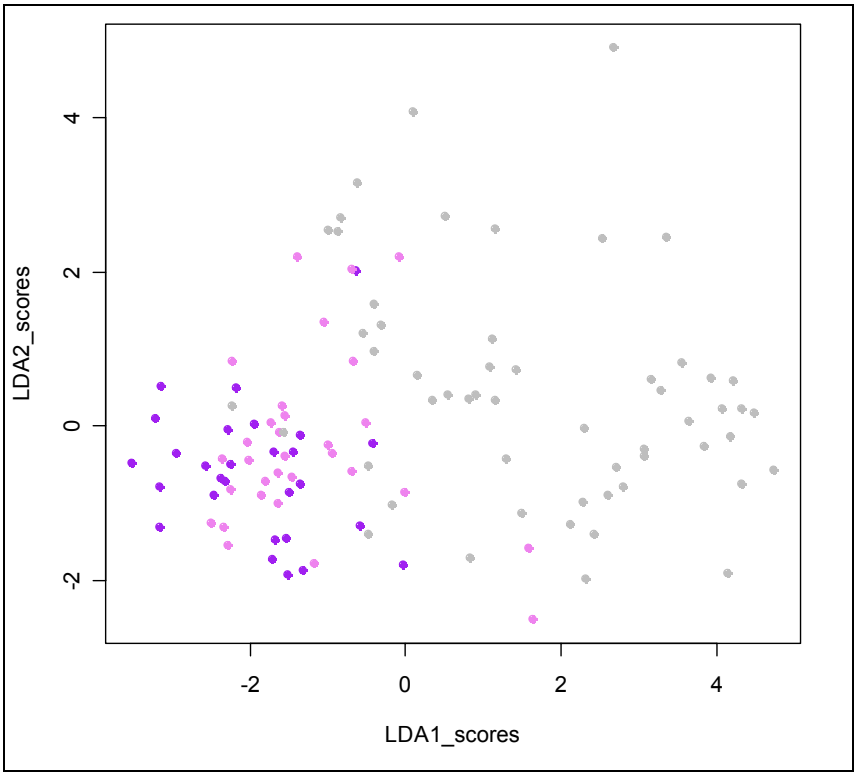


Figure 4.44: LD1-2 scores plot of summer samples with samples from Pachino area color labeled according their salinity.

As evidenced by inspection of Figure 4.44, also in the case of summer samples as for winter ones, a different salinity grade do not show any tendency to cluster, thus demonstrating that this parameter does not interfere with the sample classification based on the production standards adopted by Pachino's farmers.

4.2 NMR ANALYSIS AT 400 MHz

In order to give a preliminary glance to the deep improvement gained by using a most powerful instrument, the spectra acquired at 400 MHz and at 200 MHz on the same cherry tomato lyophilized powder, extracted with D_2O buffer solution at pH 4.0, are shown in Figures 4.45 and 4.46, respectively. Detailed expansion of the spectral regions, comprised between 5.7 and 6.2 ppm and between 3.1 and 3.3 ppm, are depicted in their insets.

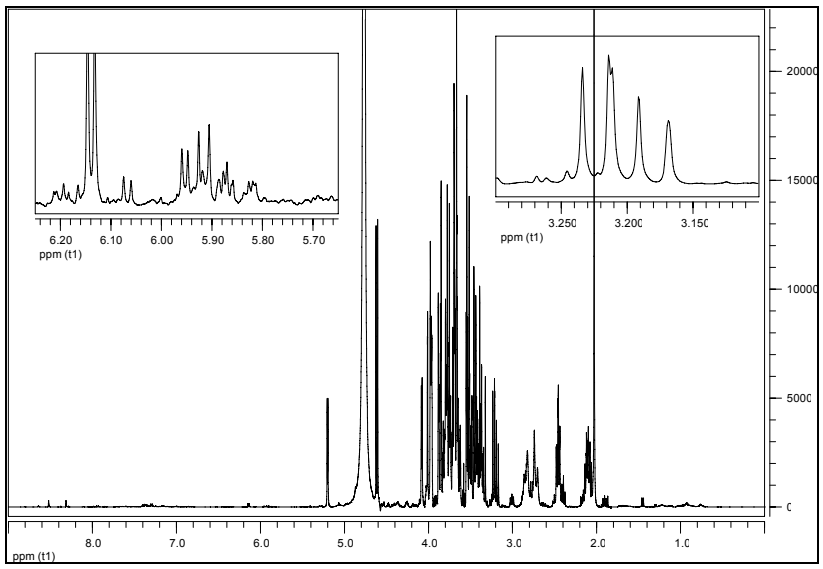


Figure 4.45: NMR spectra of cherry tomato extracts recorded at 400 MHz. The insets show some details pointing out sensitivity and resolution

It appears immediately clear that at higher magnetic field it is possible to observe more signals because of the higher sensitivity, and also the resolution of signals is deeply improved with respect to that one observed in spectra recorded at 200 MHz. This fact makes the spectra recorder on the same tomato extracts much richer in spectral information at 400 MHz than at 200 MHz.

All FIDs recorded at 400 MHz have been have been processed with Mestre-C and entirely exported in ASCII files, all containing the intensities relative to the 16384 data points. This file has been imported, by a set of home made scripts in the statistical software, and arranged in a data matrix for successive statistical pre-processing steps as described below.

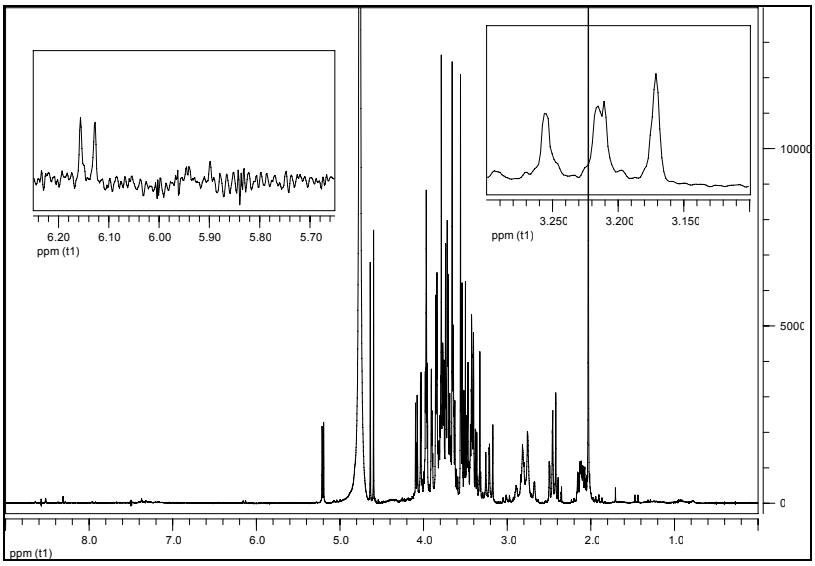


Figure 4.46: NMR spectra of cherry tomato extracts recorded at 200 MHz. The insets show some details pointing out sensitivity and resolution

Before any further chemometric analysis a detailed pH dependence study has been accomplished in order to find out what are, and in which measure, those signals affected by this parameter.

4.2.1 Study on the temperature and on pH dependence of signals.

A single aqueous extract of cherry tomato sample coming from Pachino (ID num. 225) has been investigated to study spectral changes related to pH variability. The sample has been analyzed at three different temperatures, and at three different pH values. Therefore, 9 FIDs have been acquired, all processed in the same way as for all other samples, and transformed in 9 ASCII files suitable for statistical analysis.

The 9 spectra acquired for this study have been labeled according to the following Table 4.2, with the colors used for all figures shown for the pH-dependence study.

Table 4.2: Experimental design for the study on the effect of temperature and pH on the signal position in NMR spectra

	Temp. = 23°C	Temp. = 25°C	Temp. = 27°C
pH = 3.86	(A1)	(A2)	(A3)
pH = 4.19	(B1)	(B2)	(B3)
pH = 4.44	(C1)	(C2)	(C3)

Signal referencing has been performed by setting the upfield peak of the β-D-glucose doublet signal at 4.65 ppm.

The spectra have been superimposed according to 4 different combinations, three along each column of the table, the fourth along the row at pH 3.86.

All superimpositions of spectra have been performed so that the β-D-glucose signals are aligned and scaled at the same intensity.

By looking at the comparison of spectra along the first row of the table, the influence of the temperature on the signal alignment is explored, permitting the individualization of the temperature-dependent signals. This is the case of HOD water signal, which is reported in Figure 4.47.

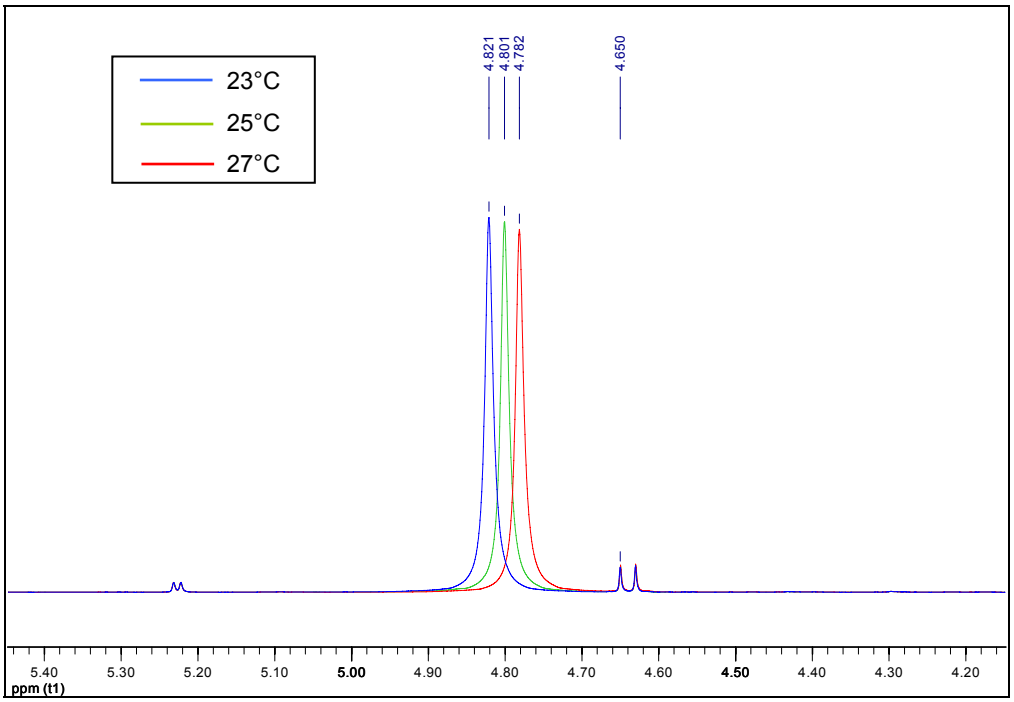


Figure 4.47: 1H-NMR spectra of cherry tomato extract recorded at different temperature. The highest signal is assigned to the residual HOD proton.

At first glance, the temperature dependence of the HOD residual resonance appears evident. The shift of this intense and broad signal may have a large influence also on the shape of small signals that may be found on the tail of the solvent peak, such as those of the anomeric protons of sugars.

The three superimposed spectra shows, by deep inspection, a very good alignment of all the signals falling in all regions of the spectrum. Some relevant spectral zones are shown in Figures 4.48-4.51.

Figure 4.51, illustrating the aromatic region comprised between 7.8 and 8.9 ppm, shows only a negligible horizontal shift of the signals belonging to the phenyl protons of aromatic compounds.

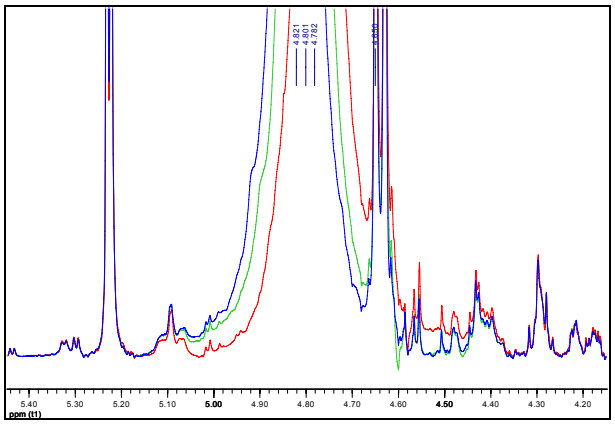


Figure 4.48: Superimposition of NMR spectra recorded at different temperatures

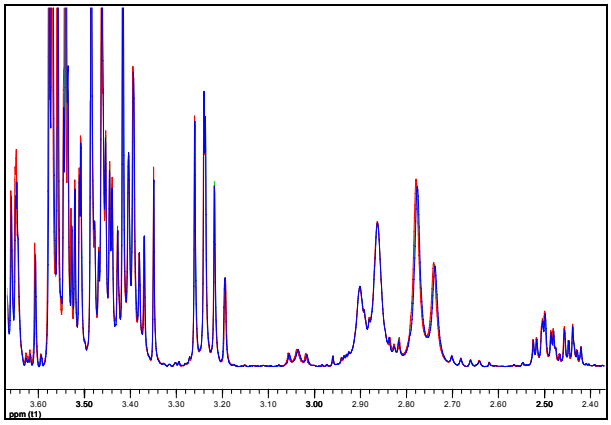


Figure 4.49: Superimposition of NMR spectra recorded at different temperatures

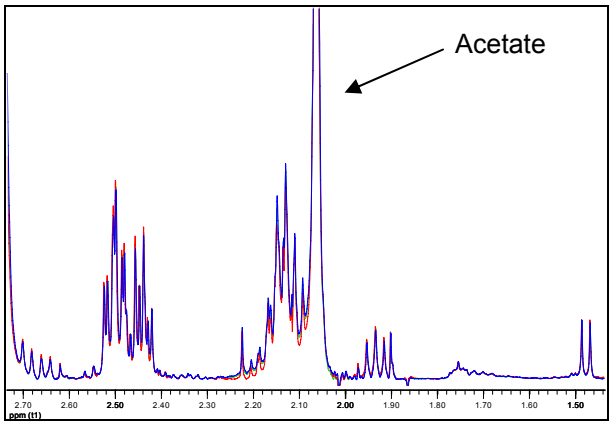


Figure 4.50: Superimposition of NMR spectra recorded at different temperatures

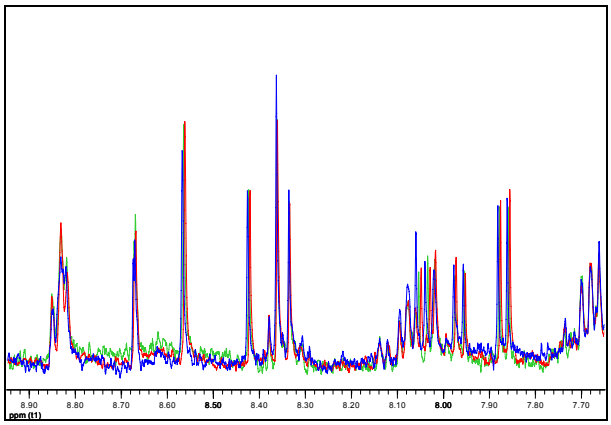
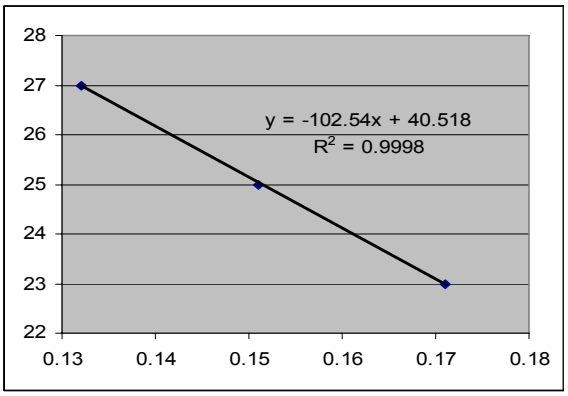


Figure 4.51: Superimposition of NMR spectra recorded at different temperatures

In the following Table 4.3, the chemical shifts of the residual HOD signal recorded at three different temperature are reported, together with their Δ with respect to the upfield peak of the doublet signal of β-D-glucose set at 4.65 ppm.

Table 4.3: Water proton chemical shift of HOD recorded at different temperatures and pH 3.86

Temp (°C)	chem. shift	Δ (ppm)
23	4.821	0.171
25	4.801	0.151
27	4.782	0.132



Graph 4.1: temperature dependence of HOD signal

It is found that the largest change of position among the main signals present in the spectrum is the one of the residual HOD signal. The chemical shift displacement recorded, for the same sample, at three different temperatures is reported in Graph 4.1. A linear dependence with temperature of such a signal results clear, and the effect is measurable as a shift of 0.01 ppm/°C. This value, compared with the width of bins (0.05 ppm), allowed us to neglect the effect of temperature in binned data, whenever a small temperature oscillation is admitted during the acquisition of the whole dataset.

Different results have been obtained by investigating the dependence of the chemical shift of signals with pH variations. An overview of the HOD resonance in three spectra recorded at 25 °C still shows a small dependence of its position with pH (Figure 4.52).

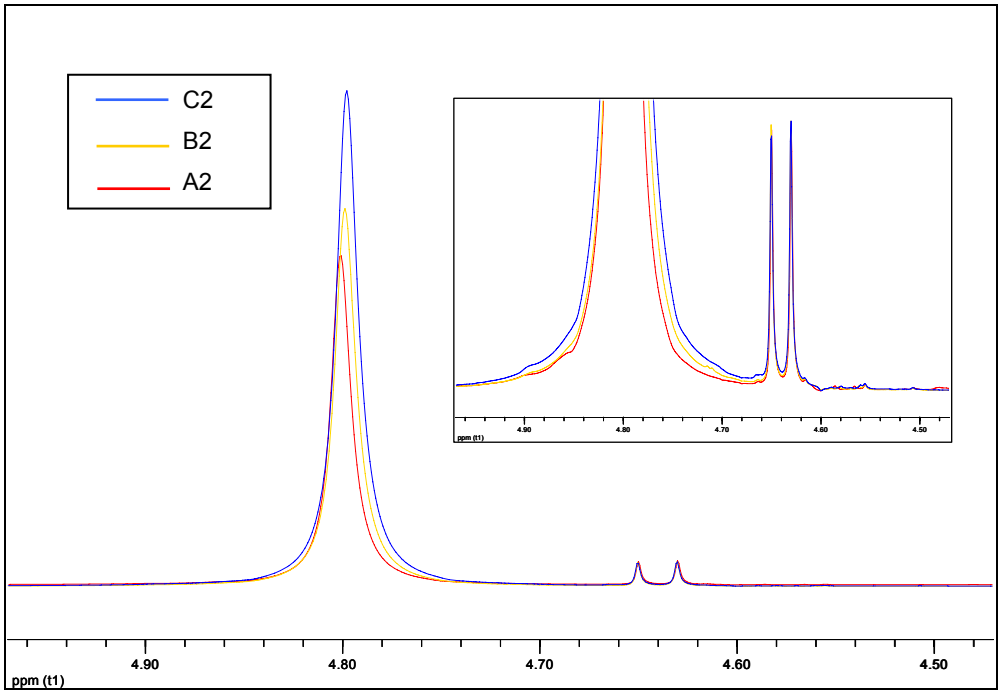


Figure 4.52: Temperature dependence of HOD signal

In this case water signal is not the one experiencing the largest shift with pH titration, but many others can be found along several spectral regions.

The whole spectra, from 10 to 0 ppm, have been shown in the following Figures 4.53-4.63, each covering about 1 ppm of spectra.

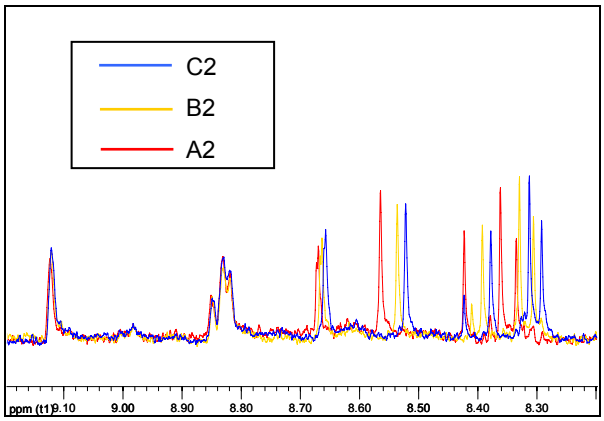


Figure 4.53: Superimposition of NMR spectra recorded at different pH

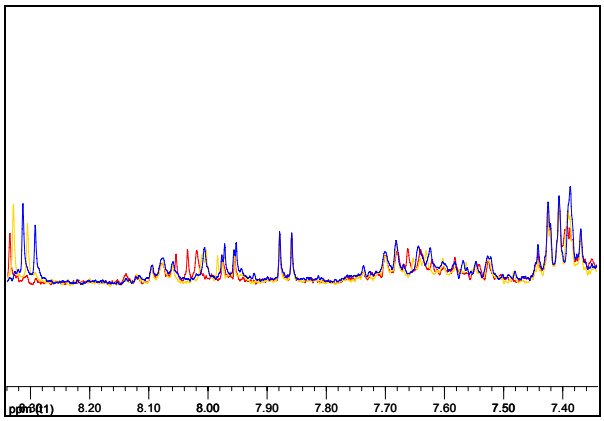


Figure 4.54: Superimposition of NMR spectra recorded at different pH

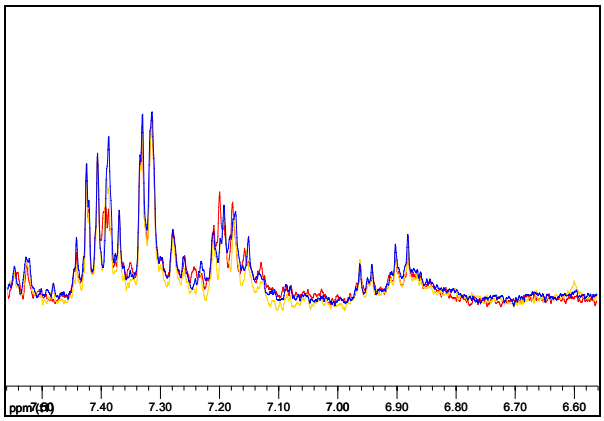


Figure 4.55: Superimposition of NMR spectra recorded at different pH

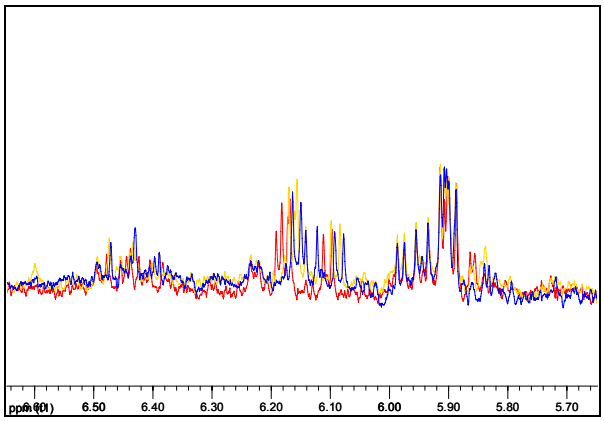


Figure 4.56: Superimposition of NMR spectra recorded at different pH

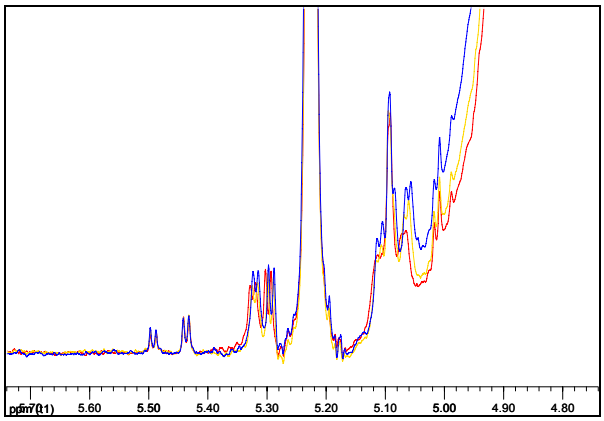


Figure 4.57: Superimposition of NMR spectra recorded at different pH

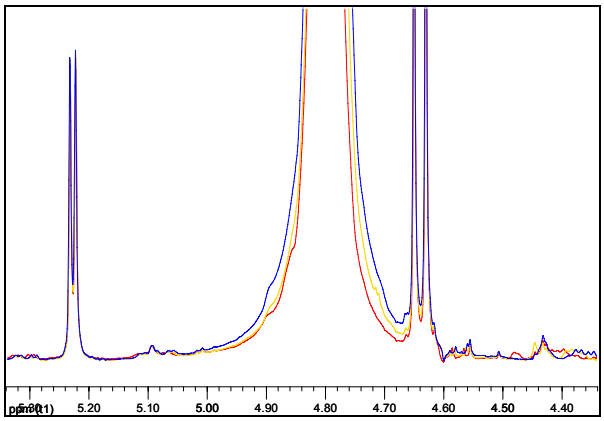


Figure 4.58: Superimposition of NMR spectra recorded at different pH

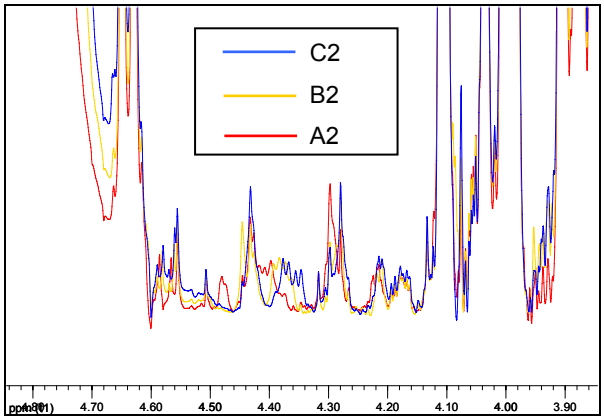


Figure 4.59: Superimposition of NMR spectra recorded at different pH

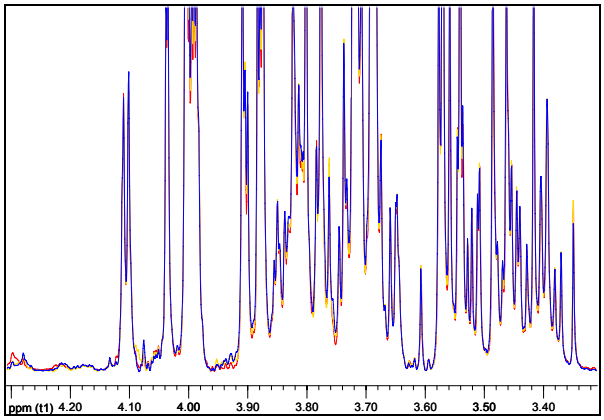


Figure 4.60: Superimposition of NMR spectra recorded at different pH

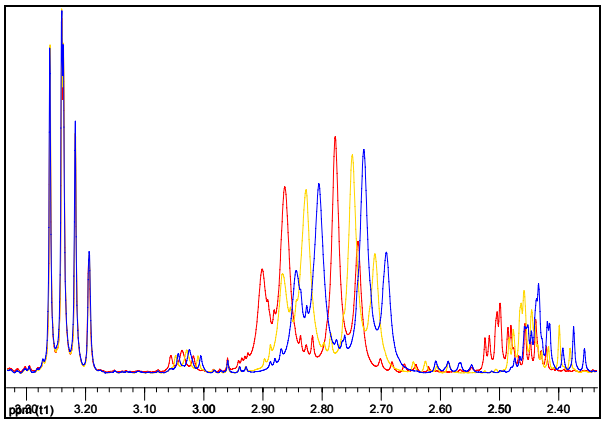


Figure 4.61: Superimposition of NMR spectra recorded at different pH

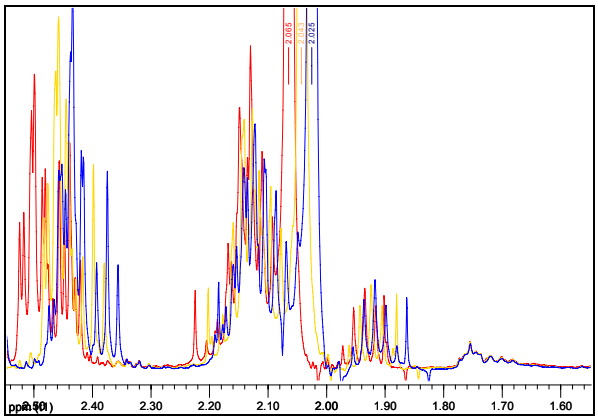


Figure 4.62: Superimposition of NMR spectra recorded at different pH

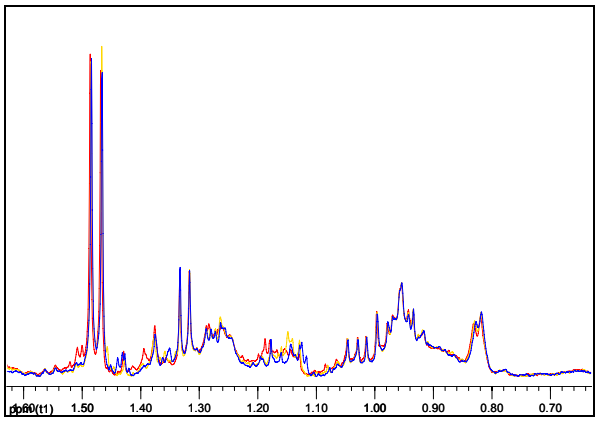


Figure 4.63: Superimposition of NMR spectra recorded at different pH

The vertical scale of the spectral regions represented in these figures has been individually adjusted by vertical expansion in order to better catch the positional variation of signals otherwise not observable.

Going on with the spectral inspection it is clearly noticeable that there are regions where the pH-dependent signals are side by side to other ones perfectly aligned in the three spectra.
For example, in Figure 4.61, the most upfield signals belonging to carbohydrates protons visible at about 3.25 ppm are clearly immobile, while the double doublet of citrate, centered at about 2.80 ppm, moves upfield with increasing pH.
A detailed study of the pH dependence of such a signal has been performed and reported below. A zoom on the citrate region of the three spectra has been shown in Figure 4.64, where also the chemical shift of each peak forming the citrate signal is reported for each spectrum. The chemical shifts of the 12 peaks assigned to the AB spin system of methylene groups in citrate are reported in Table 4.4 for the three superimposed spectra a different pH.

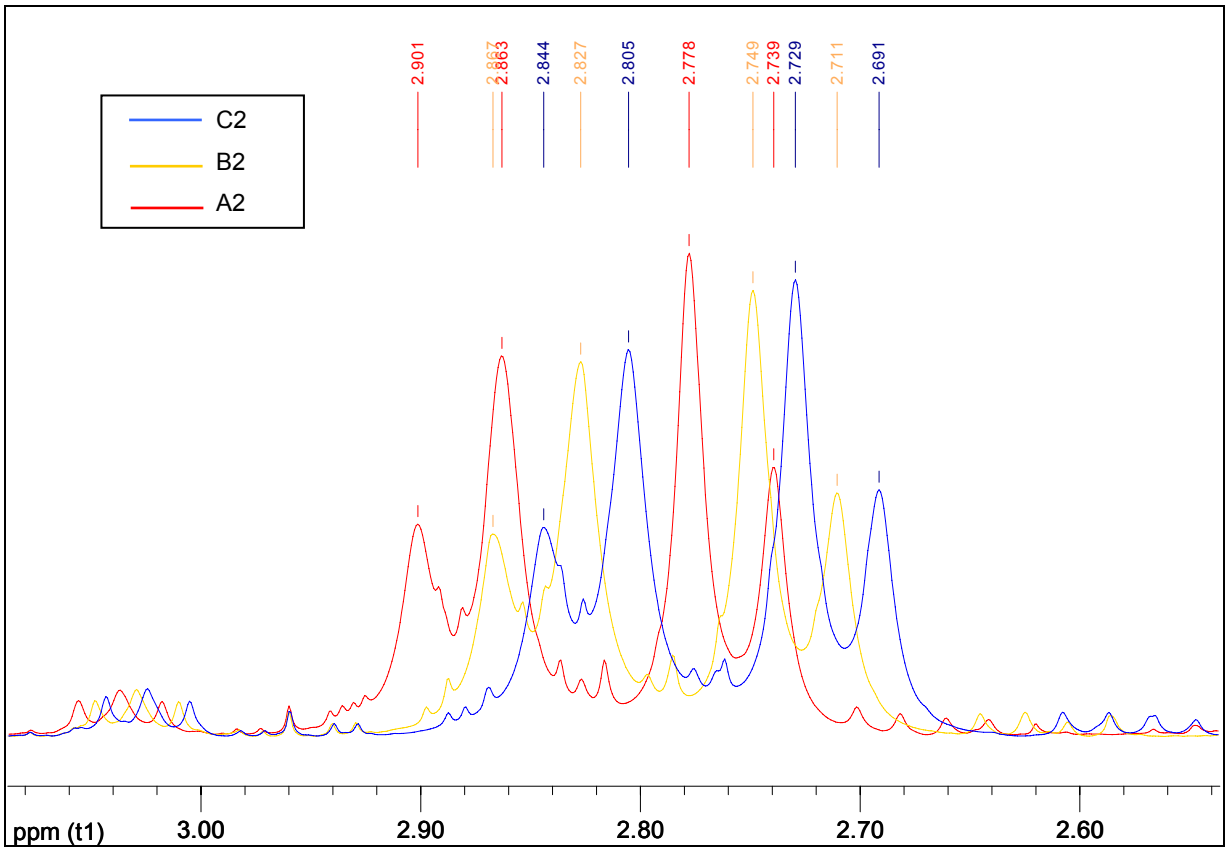
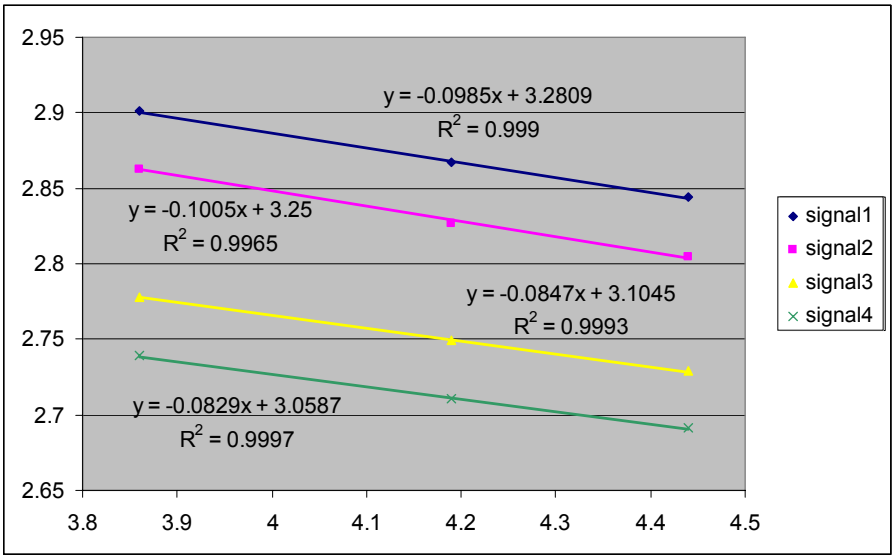


Figure 4.64: pH dependence of NMR signal of citrate

Table 4.4: proton chemical shift of citrate recorded at different pH and 25°C

Temp.=25°C	Signal1 shift (ppm)	Signal2 shift (ppm)	Signal3 shift (ppm)	Signal4 shift (ppm)
pH=3.86	2.901	2.863	2.778	2.739
pH=4.19	2.867	2.827	2.749	2.711
pH=4.44	2.844	2.805	2.729	2.691

There is a straight dependence with the pH of all chemical shifts of signal belonging to the AB system, as emerging from the linear regression, whose R^2 coefficients is reported in Graph 4.2. Such linearity is the result of citrate second pKa value (pKa1 = 3.1, pKa2 = 4.8, pKa3 = 5.4) being close to the pH of the extracts. The acidic titration is, indeed, in the linear portion of the sigmoid close to the flex point.



Graph 4.2: pH dependence of NMR signal of citrate

The same evaluation has been conducted on the acetate signal, whose results are reported in Figure 4.65 and summarized in Table 4.5 and Graph 4.3.

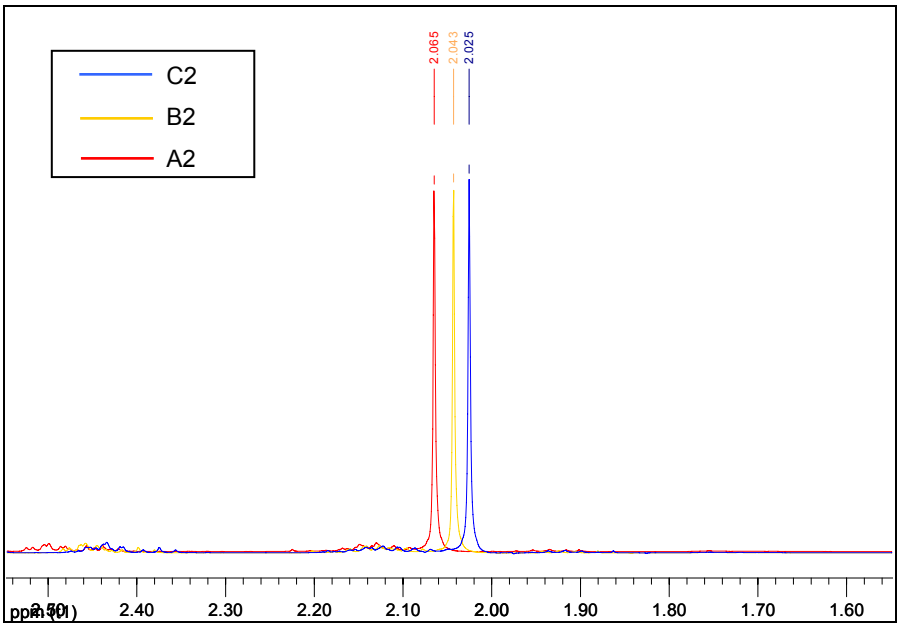
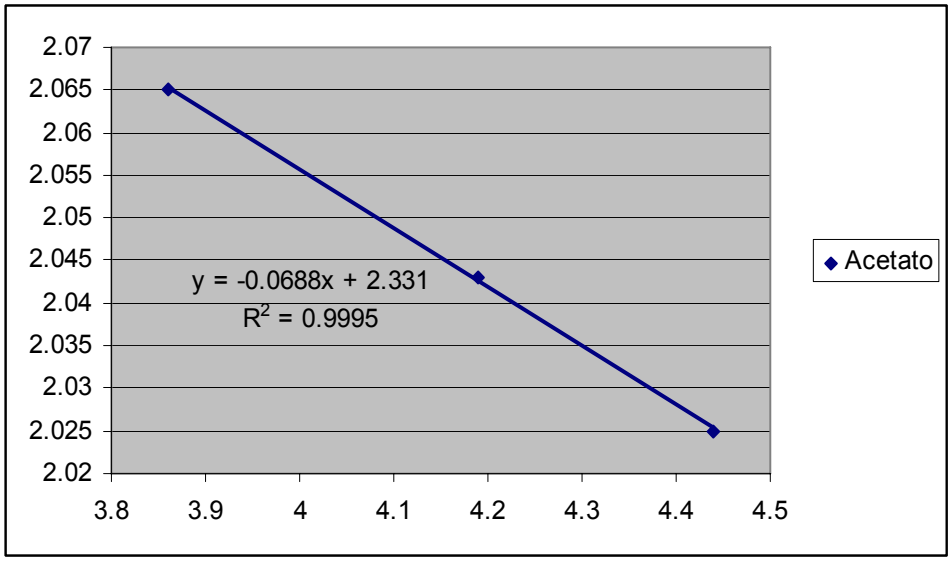


Figure 4.65: pH dependence of the NMR signal of acetate

Table 4.5: proton chemical shift of acetate recorded at 25°C and different pH

	pH=3.86	pH=4.19	pH=4.44
Acetate peak	2.065	2.043	2.025

Also in this case, there is a straight dependence with the pH of the chemical shifts of the signal belonging to the methyl proton of acetate, as emerging from the linear regression, whose R^2 coefficients is reported in Graph 4.3. Such linearity is the result of the acetate pKa value (pKa = 4.3) being close to the pH of the extracts. Also for acetate, the acidic titration is, indeed, in the linear portion of the sigmoid close to the flex point.



Graph 4.3: pH dependence of the NMR signal of acetate

As a result of this study, acetate and citrate signals have been demonstrated to shift upfield linearly with increasing pH. It is noticeable that, while the slope of the citrate signals is around -0.094 ppm/pH unit, the one relative to the acetate signal is about -0.7 ppm/pH unit. Thus, signals belonging to different titrable metabolites are usually pH dependent in a different manner, making such signals impossible to be aligned in the same spectrum without manipulating, in a non linear manner, the chemical shift scale. Such misalignment is even found at the level of a single multiplet, whose component peaks experience different pH dependence as found in citrate.

In the region between citrate and acetate signals, as well as upfield with respect to acetate, other signals appear to be titrable by pH, as emerging from inspection of Figure 4.62.

Summarizing the results obtained by the present study it is possible to infer that the region comprised between 3.1 and 4.2 ppm, i.e. the one relative to signals originating from sugars, results to be not influenced by pH fluctuations. Thus the superimposed spectra, within this region, appear essentially aligned.

In the light of these considerations, the acquired data have been submitted to statistical analysis following the same pre-processing steps already described for other cherry tomato samples. It results interesting to take a look at the horizontal shift of every spectrum represented by number of points of which each spectrum have to be slid for signal alignment with respect to a reference spectrum. These numbers are collected by the algorithm in a vector called "maxvet" that is plotted in Figure 4.66.

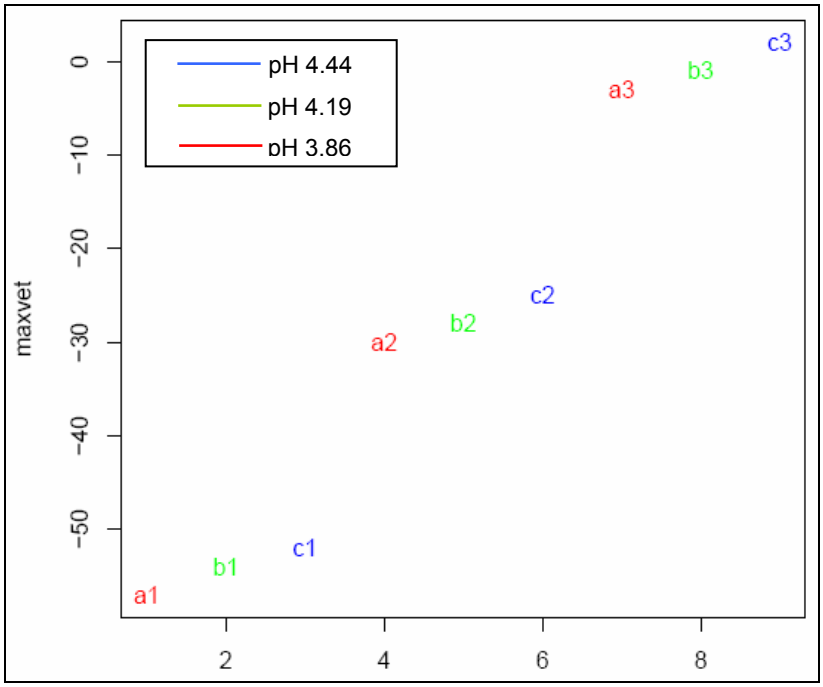


Figure 4.66: PCA performed on NMR data acquired on the same cherry tomato extracts at different pH and temperature

In the Y-axis of such a plot is reported the number of data points that separate β-D-glucose signal, set at 4.65 ppm in each spectrum, from the same signal positioned in a reference spectrum. Temperature has a bigger effect than pH on the position of this signal. This is due to the fact that the water signal is always kept at the middle of the spectral window by setting the transmitter to its frequency. Thus, the anomeric signal of glucose shifts its position with respect to the resonance of residual water, which is the signal that is actually dependent on the temperature variations, as seen before in Figure 4.52.

The original data points have been binned into 200 integral regions, each covering 0.05 ppm, and a Principal Component Analysis has been performed on the resulting data matrix, after removal of the bins containing the signals relative to residual HOD and acetate.

The results of the multivariate analysis are shown in Figure 4.67 and 4.68, where the same codes for samples identification have been respected, with colors indicating different values of pH from pH 3.86 in red to pH 4.44 in blue, through pH 4.19 in green.

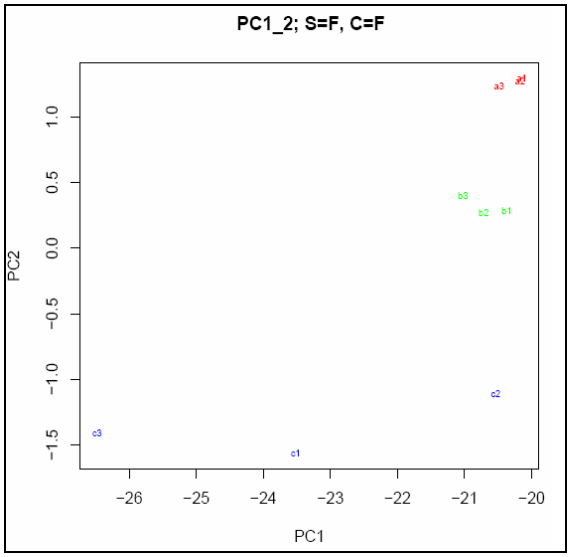


Figure 4.67: PCA performed on data acquired from samples at different pH and temperature

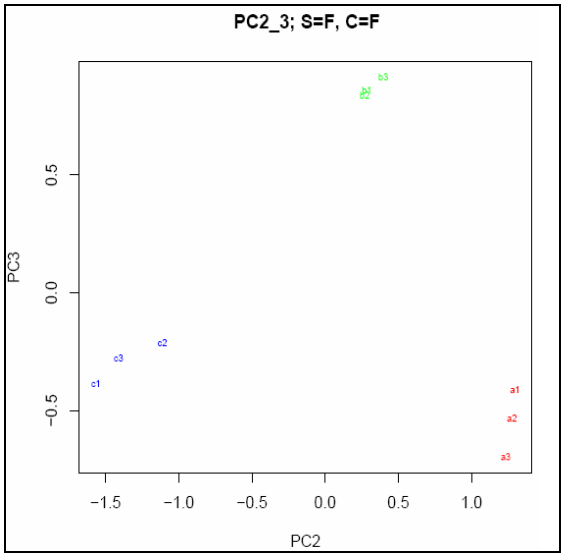


Figure 4.68:PCA performed on data acquired from samples at different pH and temperature

By comparison of both Figures 4.67-4.68, it is clear that the most important PC dimension able to separate samples recorded at different pH is PC2. It is correct to affirm that PC2 is affected by all bins experiencing variability upon pH titration and temperature modifications.

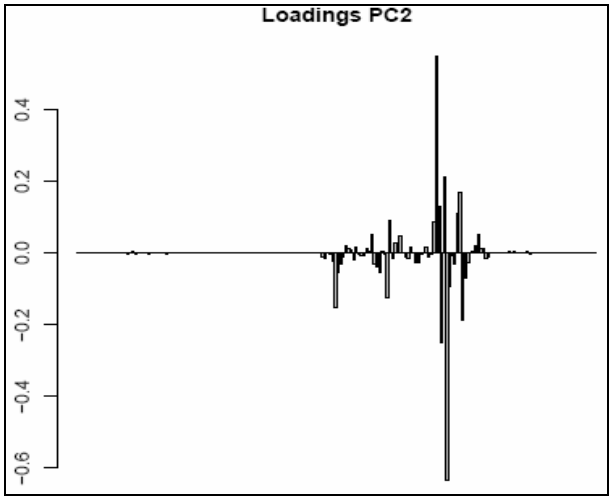


Figure 4.69: barplot representing the PC2 loading vector emerging from PCA on pH titrated samples

The PC2 loadings, bar-plotted in Figure 4.69, give a rank to the weight by which each bin influences the spectra upon titrations. By expanding the region containing the bins with the highest absolute values for loadings (Figure 4.70), it is possible to point out that bins from 138 to 144, corresponding to the signal belonging to citrate, are responsible of a great influence on the variability of the NMR spectrum, because citrate is the predominant organic acid in cherry tomato extracts.

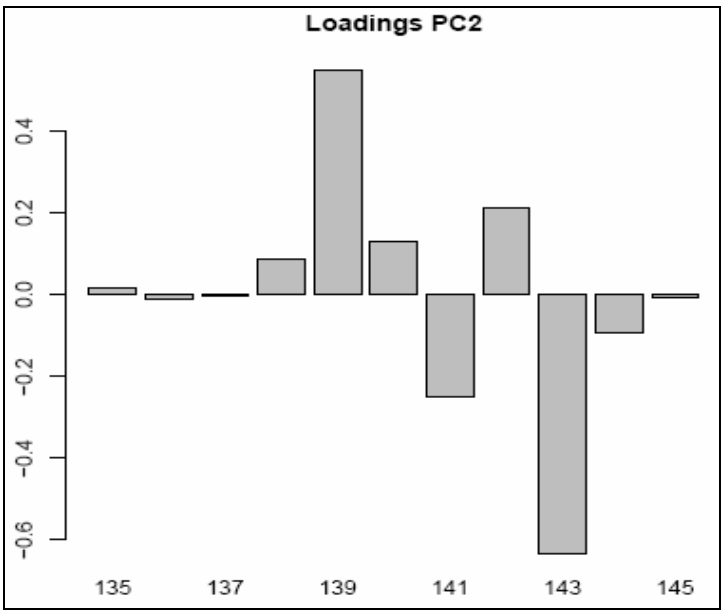


Figure 4.70: barplot representing the PC2 loading vector emerging from PCA on pH titrated samples (detail of Figure 4.69)

An algorithm, able to identify those bins having an absolute intensity higher than 0.025, has been used in order to select and remove them from the dataset before application of the successive chemometric processing of data. The removing of the same bins has been applied also to the NMR data recorded at 200 MHz, as mentioned in Paragraphs 4.1.2 and 4.1.3 for data shown in Figures 4.27, 4.28, 4.37 and 4.38.

4.2.2 Chemometric data processing of aqueous extracts

271 cherry tomato samples have been extracted and analyzed by the 400 MHz NMR Varian spectrometer. Once all the FIDs recorded at 400 MHz have been processed with Mestre-C, all spectra have been exported in ASCII files and arranged in a comprehensive spectral data

matrix (271 x 16384). The same steps followed to pre-process all spectra acquired at 200 MHz have been also applied to the comprehensive 400 MHz raw data matrix, in order to obtain a final data matrix suitable for the subsequent chemometric analysis.

First of all, a PCA analysis performed over raw data matrix has been exploited to point out outliers (like the one circled in Figure 4.71) and to find out the spectrum that has the most averaged features in the whole dataset (the one closer to the black cross in the same Figure 4.71). This "central" spectrum, with ID number 170 in the dataset, will be used as the target object during the alignment step (*spref* in the algorithm listed in Appendix A).

In Figures 4.71 samples are colored according with their geographical origin: red for **Naomi** samples coming from the region of Pachino, orange for **Shiren** samples coming from Pachino, black for **aspecific** samples still originating from Pachino, magenta for samples coming from **Licata** (Sicily), green for samples coming from **Sabaudia** (Lazio) and brown for samples of different origin purchased from general **markets**.

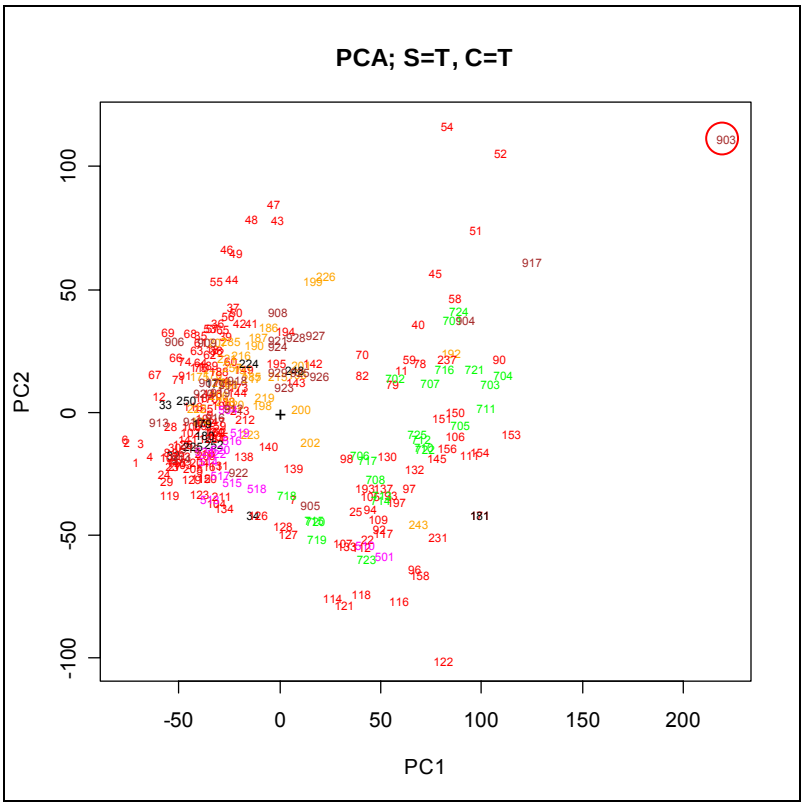


Figure 4.71: PCA performed on all data acquired at 400 MHz. See text for color coding.

After the preliminary multivariate analysis, 10 winter samples and 3 summer samples have been discarded because of their anomalous behavior, as pointed out by searching for outliers in those PCs explaining most of the variance. For examples, the sample with ID 903 was

considered a possible outlier when analyzing PC1 scores (circled label in Figure 4.71). An inspection of the spectra of all candidate outliers revealed that there were some problems in their NMR acquisition (mainly related to the magnetic field inhomogeneity due to bad shimming procedure) that made them not suitable for chemometric analysis. An example of such anomalous spectra is reported in Figure 4.72.

The methyl proton signal of acetate in 3 spectra is shown superimposed in the same figure to illustrate how poor spectra (red and green) differ from a good one (dark blue).

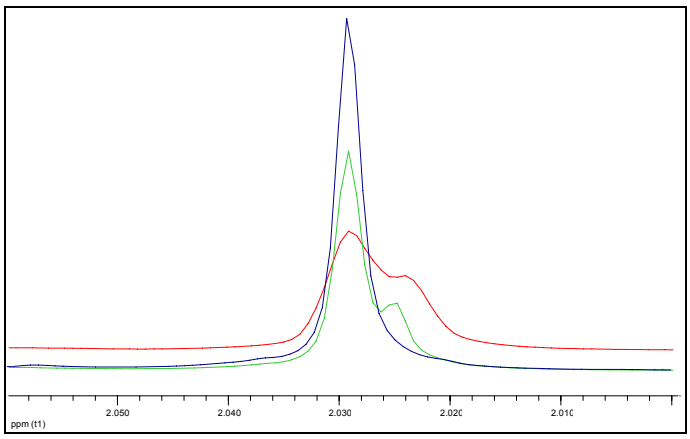


Figure 4.72: Signal of acetate in poor (green and red) and good (blue) spectra

The "central" spectrum (*spref*), to be used as the target object in the alignment step, becomes the one with ID number 164 in the dataset, after removal of the 13 outliers.

All spectra in the dataset have been horizontally aligned on the position of the downfield peak of β-D-glucose signal in the *spref* target spectrum. Vertical scale has been corrected for each data row using the integral to a Constant Sum algorithm previously described.

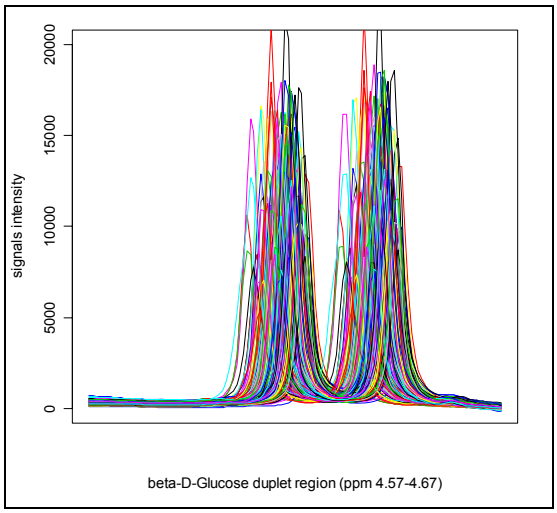


Figure 4.73: β-D-glucose doublet region.

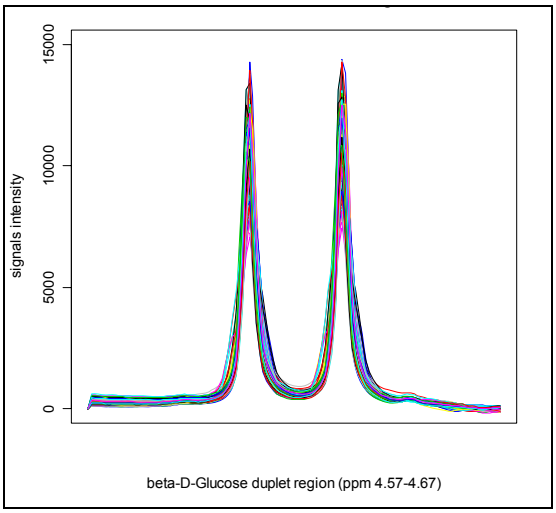


Figure 4.74: β-D-glucose doublet region.

Residual HOD and acetate signals have been removed (setting their spectral data to 0) before performing such a pre-processing step. The effects of the application of this algorithm are illustrated in Figure 4.73 (before application) and Figure 4.74 (after application).
Proceeding with data preprocessing, a new 258x16384 data matrix with signals horizontally aligned and vertically normalized has been obtained. Such a matrix still contains useless data relative to edge regions of the spectra containing only noise. For this reason a dimensional reduction is performed by cutting both edge regions and obtaining a 258x13600 data matrix suitable for successive binning step. As already seen for data acquired at 200 MHz, spectra have been split into 200 integral regions, 0.05 ppm wide, each containing the integral of 68 data points, and producing a new binned 258x200 data matrix. Such a matrix has been reduced to a 258x194 matrix by removing the bins corresponding to water and acetate signals, before submitting it to a new PCA analysis.
The results of the multivariate analysis involving all valid samples are reported in Figures 4.75-4.76.

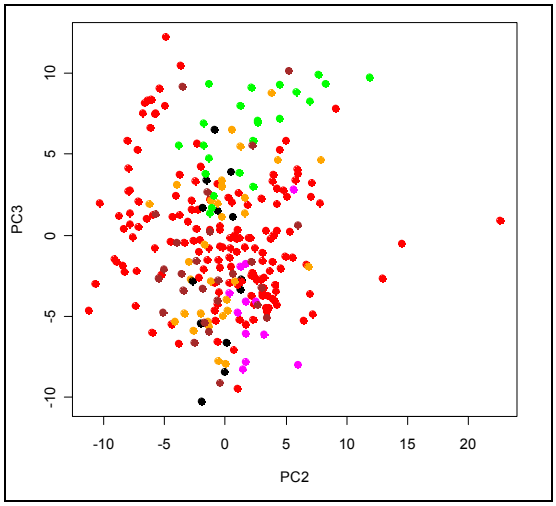


Figure 4.75: PCA performed on all samples

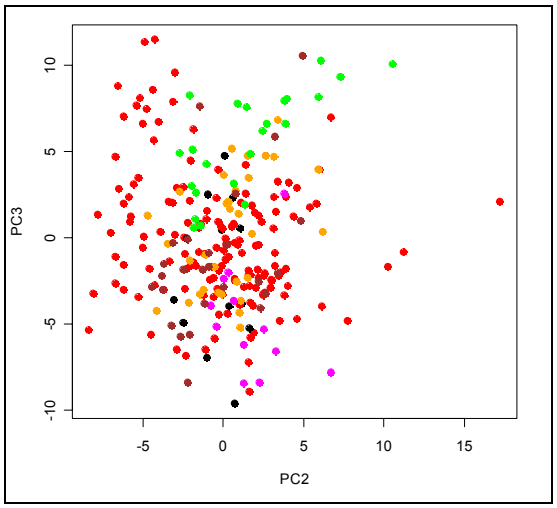


Figure 4.76: PCA performed on all samples

Only the samples from Licata and those from Sabaudia seem to be roughly clusterized, even though the two groups are overlapped to samples harvested in Pachino. Also the removal of the 14 bins selected by the pH dependence study does not improve significantly the distribution of samples as emerging, at first sight, from Figure 4.76, but their exclusion from analysis will be applied hereafter to avoid any possible unwanted experimental source of variance.

In Figure 4.77, the samples have been colored according with the four harvesting season: blue for **winter**, green for **spring**, red for **summer** and brown for **autumn**. It is noticeable that samples harvested in seasons with intermediate climatic conditions (spring and autumn) fall in between summer samples and winter ones.

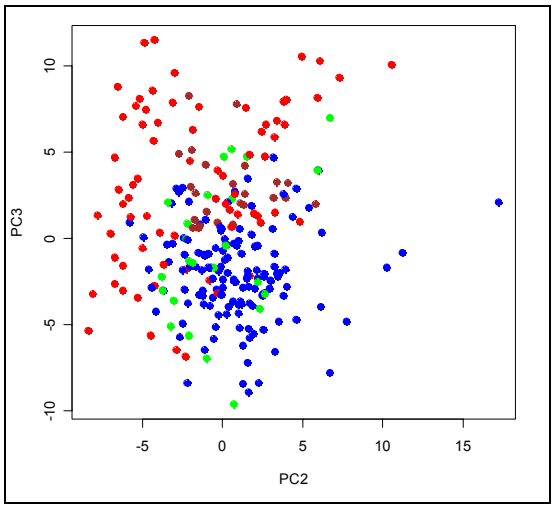


Figure 4.77: PCA performed on all samples

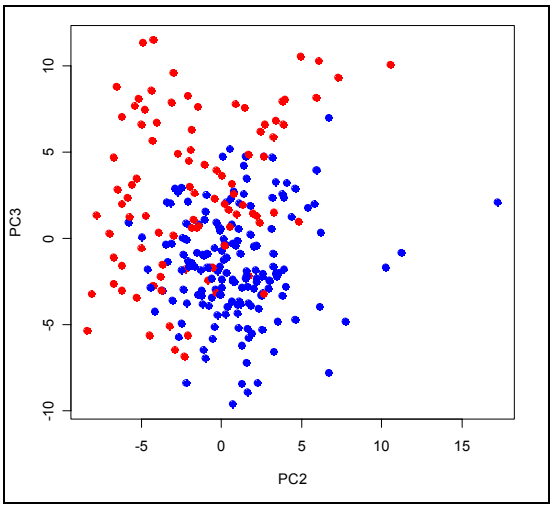


Figure 4.78: PCA performed on all samples

In order to make the PC plot more meaningful, the samples have been assigned to the two extreme winter and summer seasons according with their harvesting date. Considering that Sicily has warm temperature until October, summer samples range between the beginning of May and the end of October, while those assigned to winter span the remaining part of the year. The PC scores plot representing such a division is depicted in Figure 4.78.

As already seen for data acquired at 200 MHz, also for samples analyzed at 400 MHz a good separation between summer and winter classes exists. Obviously, an intersecting area of the two groups exists due to the intrinsic variability of the samples and to the temperature oscillations observable during intermediate seasons.

A Linear Discriminant Analysis has been performed on such a dataset in order to separate samples belonging to the four seasons. The results are shown in Figure 4.79 and 4.80. A quite good separation among the classes has been reached but a high grade of ambiguity still exists for many samples.

Since the variance due to the harvesting season exceeds all the other sources of variances that need to be investigated, the comprehensive dataset has been split in two different new data matrix relative to samples of cherry tomato harvested in winter or in summer.

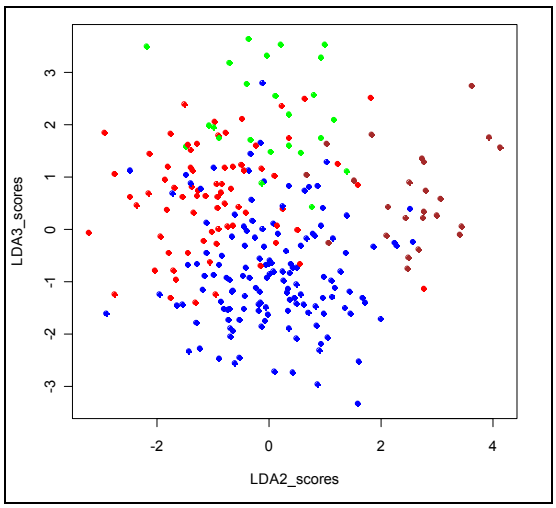


Figure 4.79: LDA performed on all samples

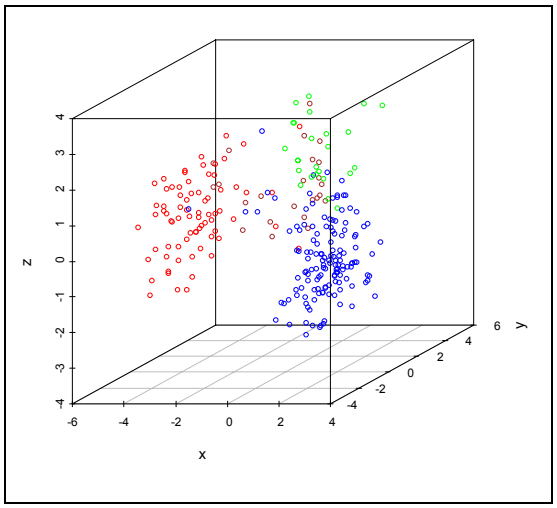


Figure 4.80: LDA performed on all samples

4.2.3 Processing of winter samples

The winter sample dataset of the spectra recorded at 400 MHz consists of 166 out of the 258 spectra previously accepted for statistical analysis after removal of outliers. The same algorithms have been applied to such a new dataset, following all the previously described preprocessing steps. The results of a preliminary PCA analysis of the raw data are shown in Figure 4.81, reporting the ID number of each sample. In this figure aspecific samples are not reported.

After the pre-processing steps and the removal of variables containing pH dependent signals, the resulting binned 166x180 data matrix has been statistically analyzed by the PCA chemometric tool.

Both PC2-3 and PC2-5 scores plots (Figure 4.82 and 4.83, respectively) do not reveal any clear separation among classes, even though a better tendency to group, with respect to the results obtained for data acquired at 200 MHz, may be observed for samples cultivated in Licata or belonging to the Shiren cultivar, as well as for the samples purchased at the market.

The samples extracted from Naomi tomatoes cultivated in Pachino appear spread all over the PC plot and overlapping with the other classes of samples. Again it is worth remembering here that the growing pedoclimatic conditions of Licata are very similar to those findable in the adjacent Pachino area. Also for data acquire at 400 MHz this fact makes the differentiation among these two classes of samples very difficult to be accomplished by PCA.

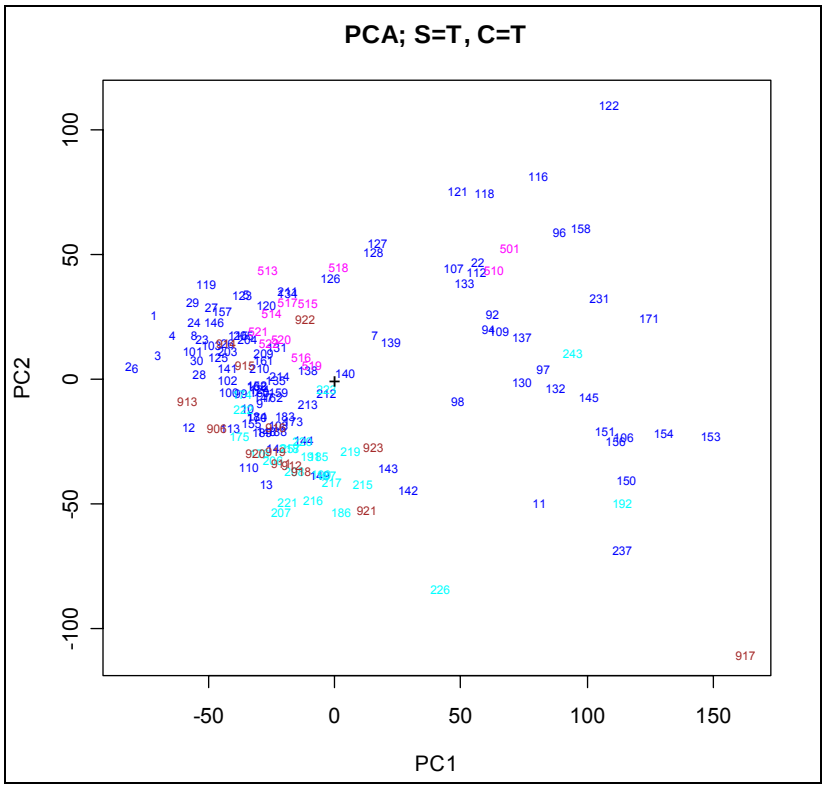


Figure 4.81: PCA performed on data acquired on winter samples at 400 MHz. See text for color coding.

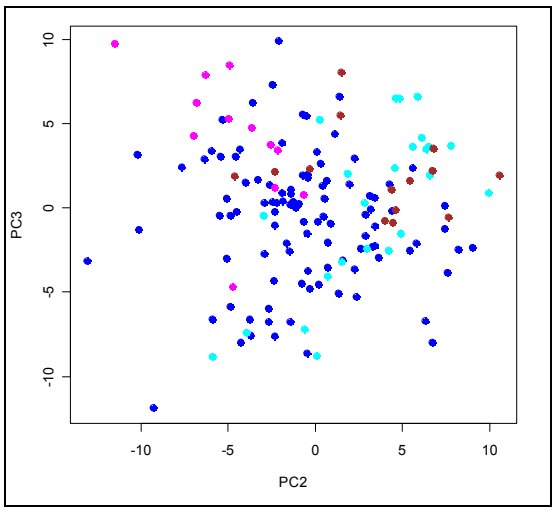


Figure 4.82: PCA performed on winter samples

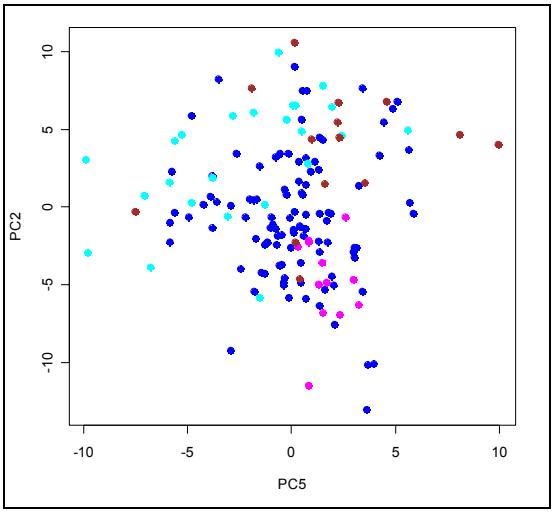


Figure 4.83: PCA performed on winter samples

Once again, adding *a posteriori* the aspecific samples on the PC spaces it results that such samples do not fall in any specific class and tend to have a random distribution (Figures 4.84 and 4.85).

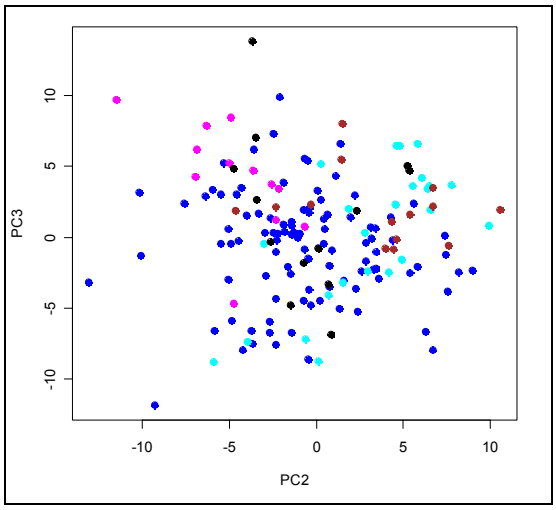


Figure 4.84: PCA performed on winter samples with aspecific samples

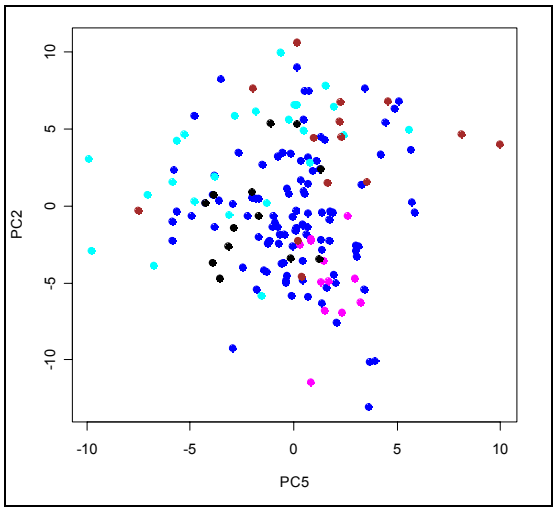


Figure 4.85: PCA performed on winter samples with aspecific samples

The tendency to cluster just seen, authorizes one to perform a Linear Discriminant Analysis over all the selected data, excluding only aspecific samples (153x180 data matrix).

LDA analysis has been performed on the PC rotation matrix (loadings) obtained by PCA analysis performed on the winter data matrix, already purged of the aspecific samples. LD1-2 scores plot of the analyzed data matrix is shown in Figure 4.86.

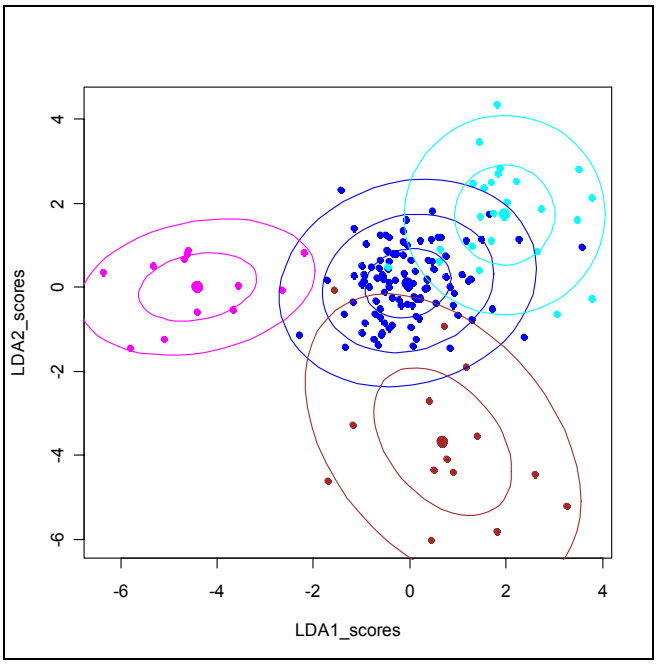


Figure 4.86: LDA performed on all winter samples without aspecific samples

The PC space describing 90% of the total variance, corresponding to the first 34 PCs, has been involved in this LDA analysis.

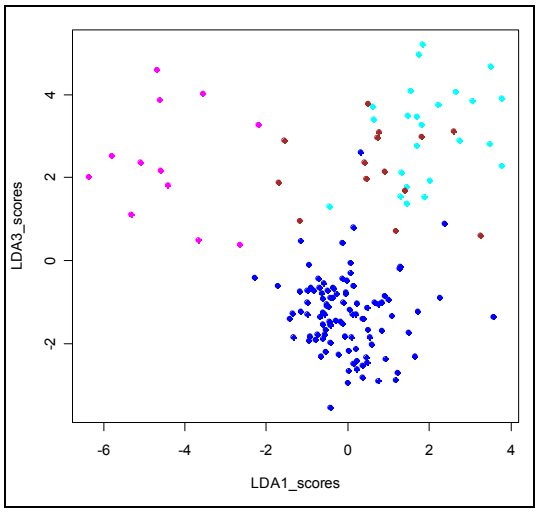


Figure 4.87: LDA 1-3 scores plot of winter samples

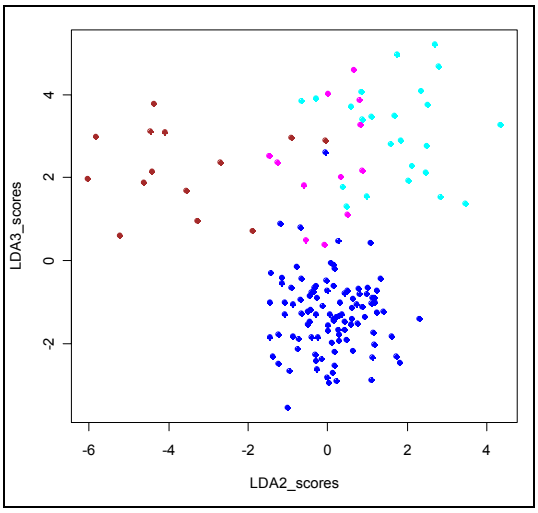


Figure 4.88: LDA 2-3 scores plot of winter samples

Classification performed by LDA appears to be quite good since the groups are clearly distinguishable. The overlap among classes still persists by observing data on a plane of two LDs. However, by looking at all the combination of the first three LDs, emerges a clear separation among all classes (Figures 4.87 and 4.88). A 3D scatterplot of all the three LD is shown in Figure 4.89, helping to have an overall sight on the localization of the four different classes in the space of the first three LDs.

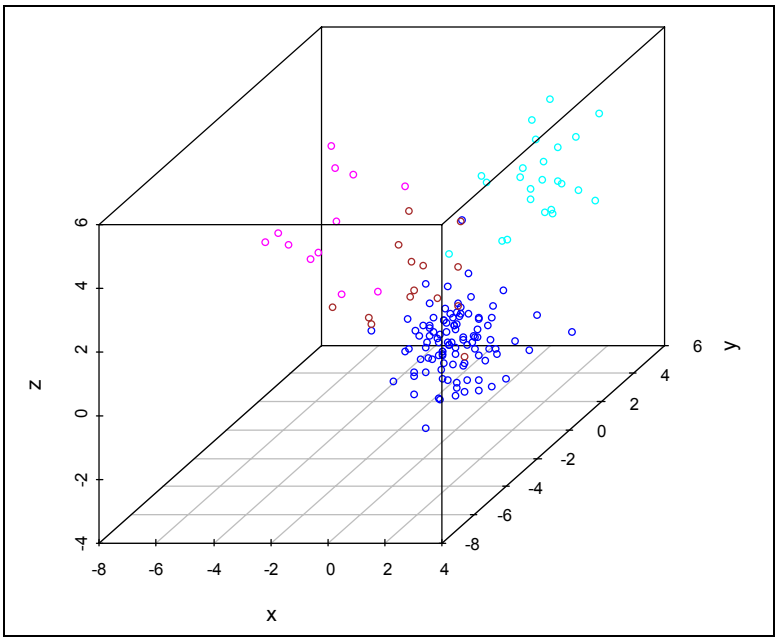
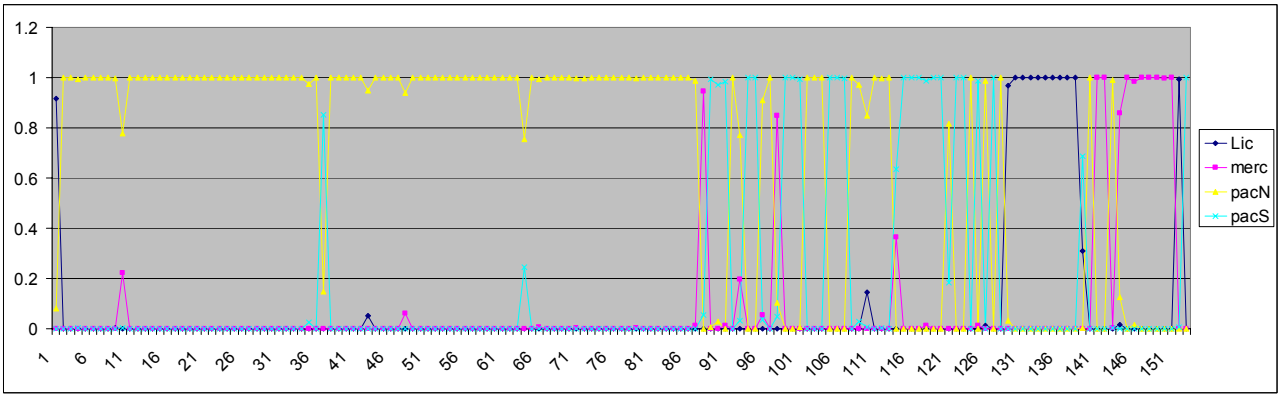


Figure 4.89 3D plot of the three LD scores dimensions

A leave-one-out test of such a statistical system has been performed in order to verify its predicting ability and whether it is affected by overfitting. The results of the test are shown in Graph 4.4. As much as 139 winter samples out of the 153 ones constituting the analyzed dataset have been rightly assigned to their own class. This means that only a marginal overfitting is present since as high as 90.85% of the samples have been correctly assigned.



Graph 4.4: Leave one out test performed on winter samples after LDA on PCs describing 90% of the total variance

Errors are mainly committed by confusing Shiren and Naomi cultivars both harvested in the Pachino area, suggesting that geographical origin, more than the cultivar, is the main discriminant factor for classification.

At this point, the aspecific samples previously excluded by the statistical multivariate analysis have been projected on the LD spaces. The following Figures 4.90 and 4.91 show in black the aspecific samples, together with all other samples.

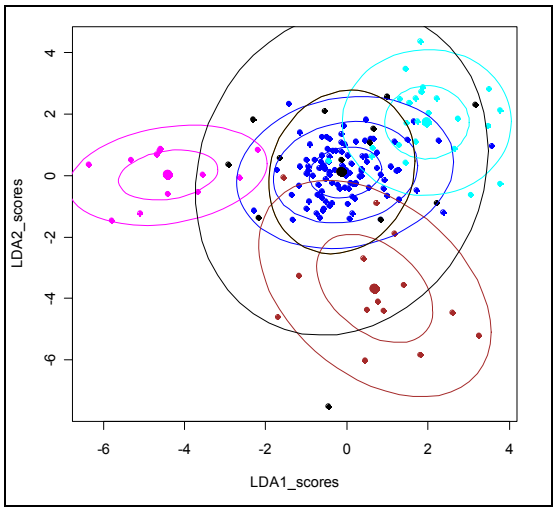


Figure 4.90: LD1-2 scores plot of winter samples with aspecific samples

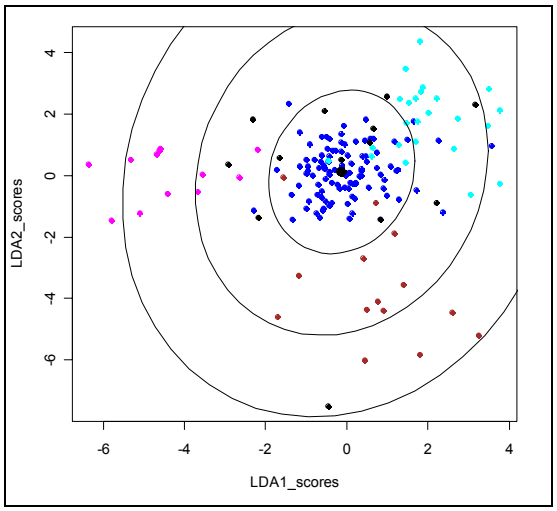


Figure 4.91: LD1-2 scores plot of winter samples with aspecific samples

Once again, their random dispersion is undoubted. Although the centroid of the group including the aspecific samples falls exactly over the centroid of the group of Naomi samples (in fact the origin of both classes is the Pachino area) their assignment to a specific class results unfeasible.

Naomi samples coming from Pachino, projected on the LD scores spaces, have been then marked in purple and violet colors according with the salinity grade of their irrigating water, as already seen for data recorded at 200 MHz. Other samples have been again gray colored in order to make the plot more significant. Figure 4.92 shows the results of the multivariate analysis on the effects of salinity of the irrigation water.

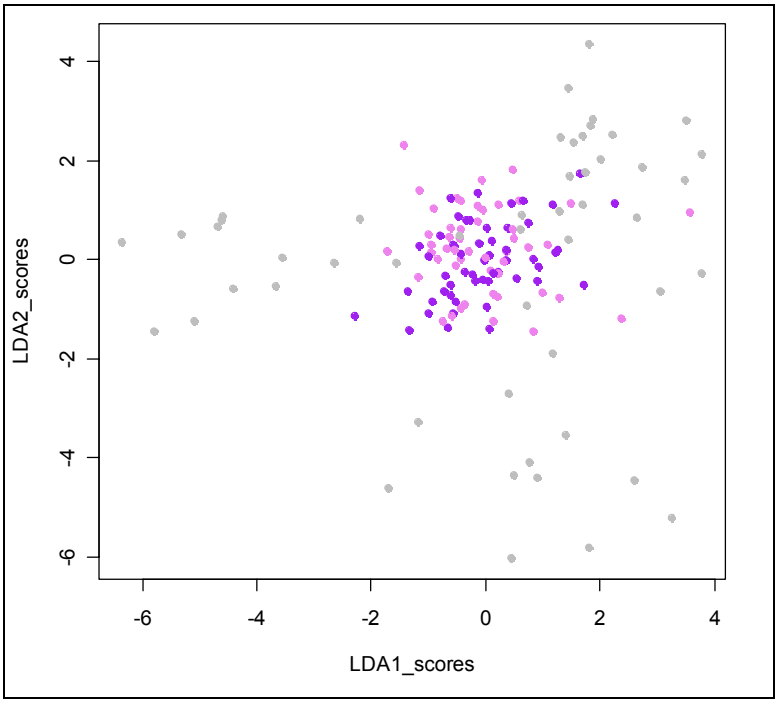


Figure 4.92: LD1-2 scores plot of winter samples with samples from Pachino area color labeled according to their salinity.

As emerging with data at 200 MHz, also those recorded at 400 MHz confirm that salinity does not influence in any significant way the distribution of samples, thus demonstrating that this parameter does not interfere with the sample classification.

4.2.4 Processing of summer samples

The summer dataset of the spectra recorded at 400 MHz consists of 92 out of the 258 spectra previously selected for statistical analysis after removal of outliers. The preprocessing steps already seen for winter samples have been performed on the 92 x 16384 matrix constituting the raw dataset of summer samples. A preliminary PCA analysis of the raw data is shown in Figure 4.93 reporting the ID number of each sample. During summer, any aspecific sample has been analyzed at 400 MHz; therefore, they are never present in this statistical step. For summer samples the reference spectrum, selected by the algorithm based on PCA, is that one having the ID number 50 ("spref"). In the following figures, summer samples have been yet colored according with their geographical origin: Pachino **Naomi** samples are red, Pachino **Shiren** ones are orange, **Sabaudia** ones are green, and **Market** ones are brown.

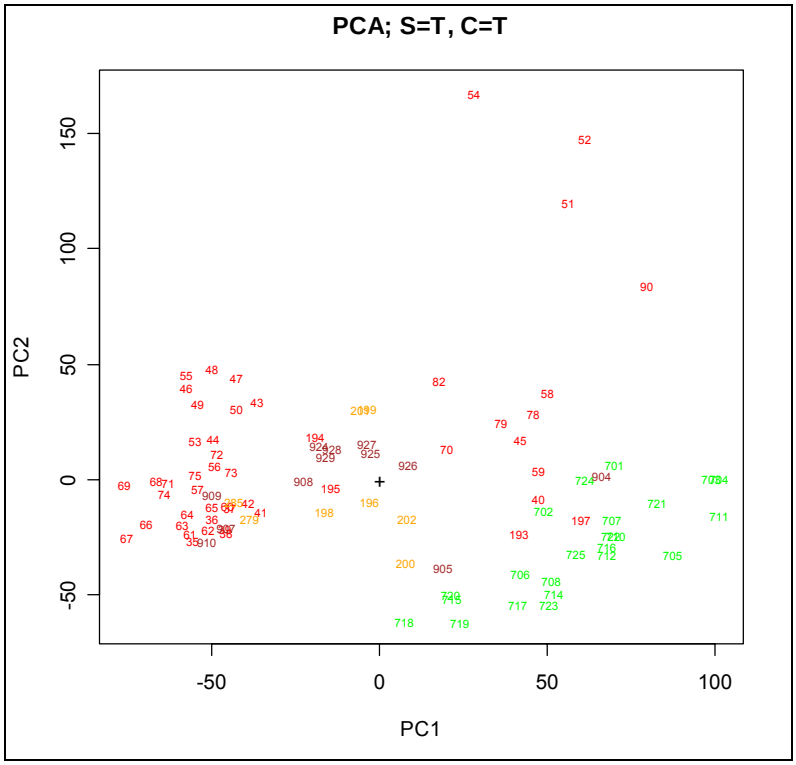


Figure 4.93: PCA performed on data acquired on summer samples at 400 MHz. See text for color coding.

Also for the summer data matrix, the bins containing the most pH dependent signals have been removed from the dataset, thus reducing the involved variables from 194 to 180. PCA

analysis conducted on such a reduced data matrix (92 x 180) has been performed obtaining the results shown in Figures 4.94 and 4.95.

Once again the best combinations of PC scores that maximize the separation in the PC space among classes related to the geographical origin have been selected.

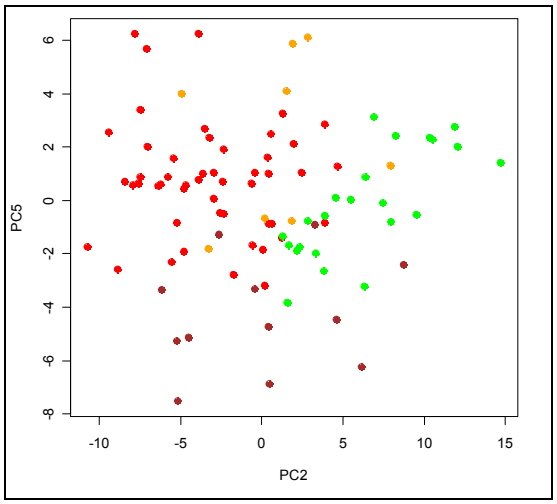


Figure 4.94: PCA performed on summer samples

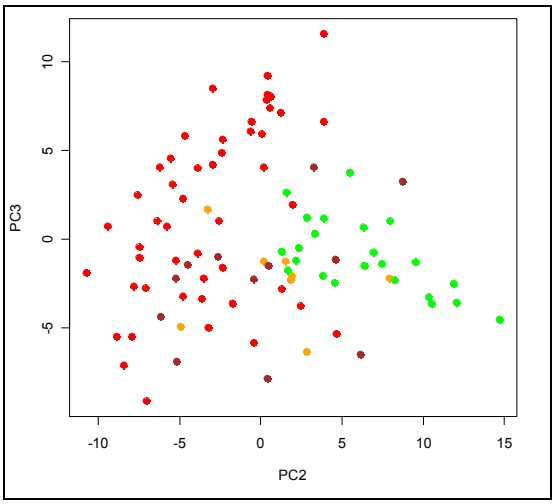


Figure 4.95: PCA performed on summer samples

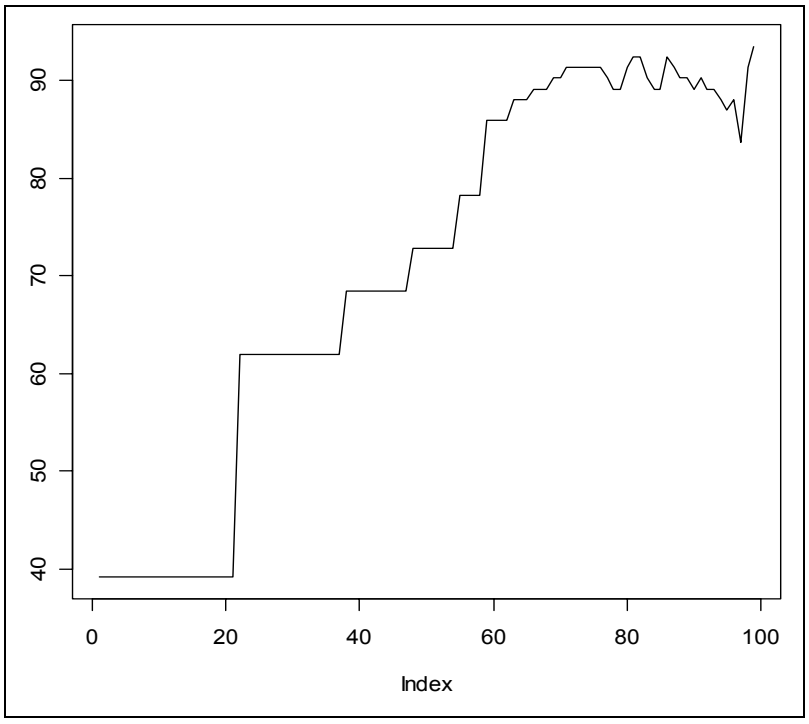
Since Shiren samples (orange) and Naomi ones (red) are both originating from Pachino it is not surprising that these two classes appear overlapped in the PC scores space. If these two groups are seen as belonging to a single class, thus the separation obtained in the PC space (PC2-PC5) results significantly good as emerging from Figure 4.94.

Samples of different origin appear now, for the firs time in this study, rather distinguished in the three classes, even before the LDA step. Only a small portion of samples is overlapped in the centre of the plot, where all the categories border on each others.

In order to perform a significant LDA on the data it has been necessary to accomplish a preliminary study for obtaining the best reachable conditions. Since LDA analysis suffers from data collinearity of variables, it is not convenient to perform this multivariate analysis directly on the data matrix. In fact, different bins may collect signals from the same molecule, resulting in a straight correlation among the involved bins.

The standard procedure transforms the space described by the bin variables in a rotated space of new variables called principal components which are chosen to be orthogonal each others. The first principal components have most of the information contained in the original data, whilst the last ones collect all the noise. For this reason it is convenient to select the number of PCs, being subjected to LDA, which retain all the useful information.

The number of PCs is also taken as lower as possible to avoid overfitting that usually happens in LDA analysis, when the number of variables used to explain a sample of investigated objects is much higher than that of objects [174]. Such a selection is based on Leave-One-Out studies performed on LDA results obtained by varying the number of PCs employed in such a way that the variance explained by PCs ranges between 1 and 99%. This has been accomplished by an algorithm that recursively repeats LDA analysis, and the consequent LOO tests, by varying the number of PC used. The results of this study are graphed in Graph 4.5.



Graph 4.5: percent of the explained variance as a function of the employed PCs

The percentage of correct assignments of samples to the proper class, as obtained by the corresponding LOO test, is reported in the ordinates, versus the percentage of total variance explained by the number of PCs used to perform LDA. It results clear that, in order to obtain a good predictivity, at least 60% of total variance have to be explained by the PCs selected. This result, alone, is not sufficient to select the right number of PCs for obtaining both a good separation among classes and a good predictivity of the system. For this purpose, the Mahalanobis distances have been calculated in order to reach a good compromise among high predictivity and low overfitting.

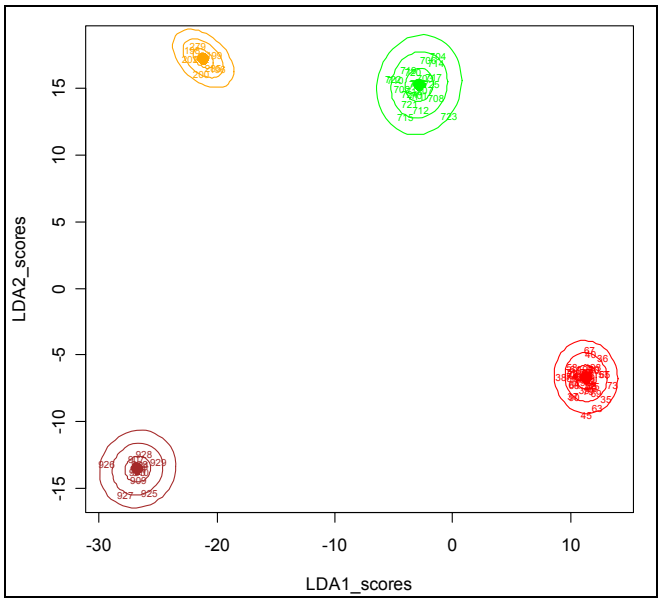


Figure 4.96: LDA performed by using 83 PCS explaining 99.9% of variance

When 83 PCs, explaining 99.9% of the total variance, are chosen out of 92 totally available, a clear overfitting affects the LDA, as evidenced by the as high as 60 mistakes obtained over 92 samples with the LOO test, corresponding to only a 65.2% predictivity.

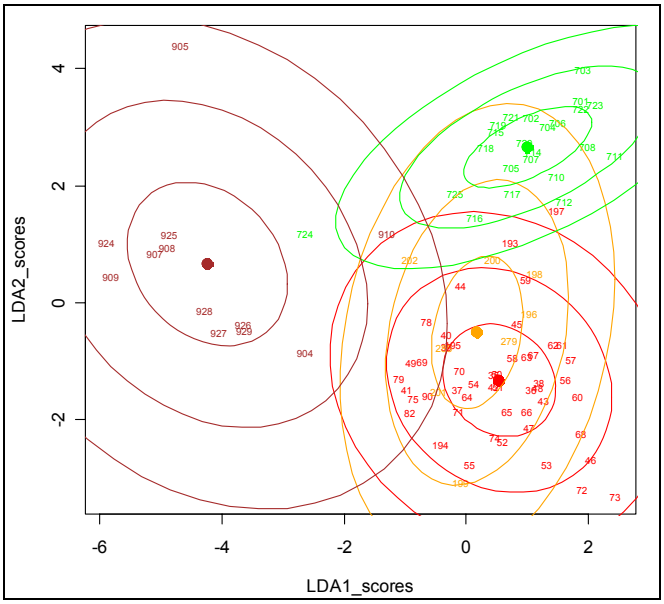


Figure 4.97: LDA performed by using 10 PCS explaining 71% of variance

On the other side, when only 10 PCs, explaining 71% of the total variance, are chosen for LDA a clear overlap of Sabaudia and Pachino classes is found, as evidenced by the Mahalanobis distances shown in Figure 4.97. It is worth noting here that predictivity takes

the main advantage by the low number of PCs, as it is as high as 91.3% with only 8 mistakes.

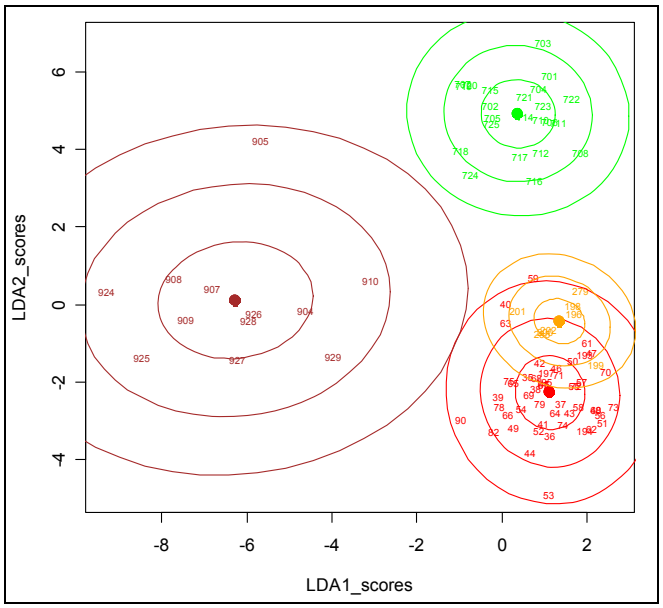
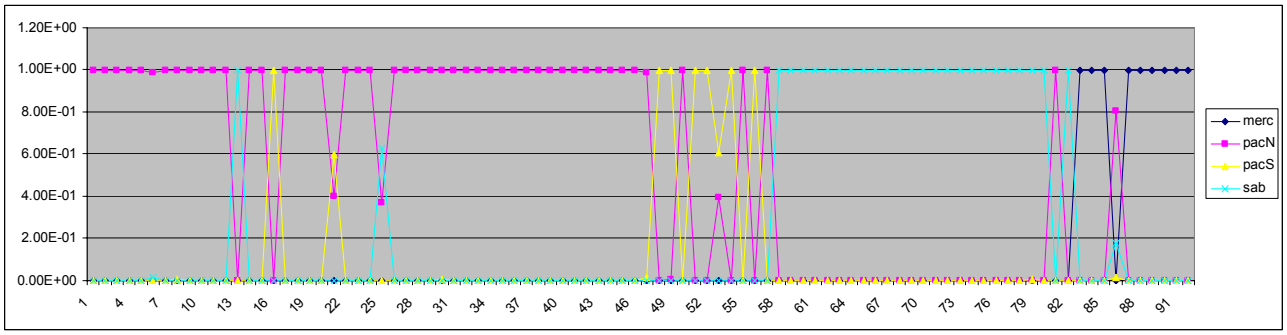


Figure 4.98: LDA performed by using 26 PCS explaining 90% of variance

The choice of 26 PCs explaining 90% of total variance resulted to be the best one among all those conducted on this study because, on one hand, it still presents a perfect separation among classes (Figure 4.98) and, on the other hand, maintains 89.1% predictivity, with only 10 mistakes (Graph 4.6), very close to the best one obtained with 10 PCs.



Graph 4.6: Leave one out test performed on summer samples after LDA on PCs describing 90% of the total variance

In most of the cases mistakes on the assignment of the class which the leaved-out sample belongs to are due to confusion between Naomi and Shiren samples of Pachino, thus not being a relevant errors since they really have the same geographical origin.

In the light of such results, all LDAs carried out along the present thesis work have been performed using always a number of PCs explaining 90% of total variance, as previously seen for winter data recorded at 400 MHz as well as for data collected at 200 MHz.

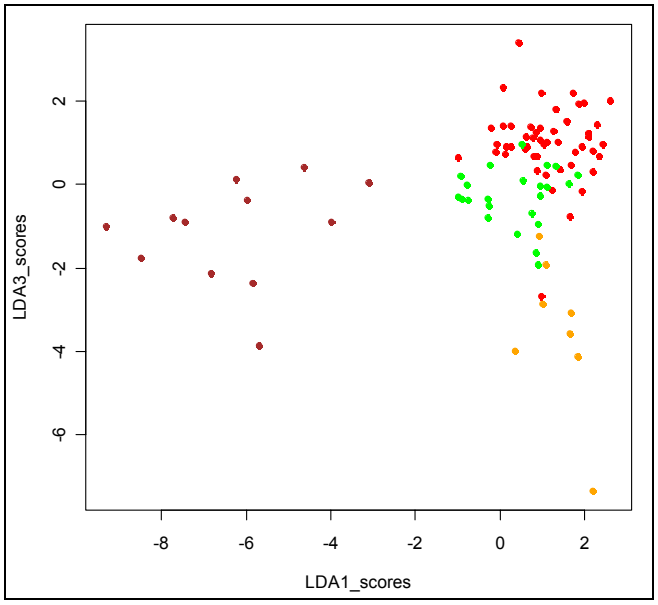


Figure 4.99: LD1-3 scores plot of summer samples

It is relevant to observe that even if Shiren and Naomi samples are forced to belong to different classes, LDA tends to overlap them in LD1-2 scores space (Figure 4.98), suggesting that these two groups are very similar, and indeed they are with respect to their geographical origin. However, LD3 scores differentiates the two groups, as emerges from inspection of Figures 4.99 and 4.100, prompting that another effect, e.g. the cultivar, determines certain diversity.

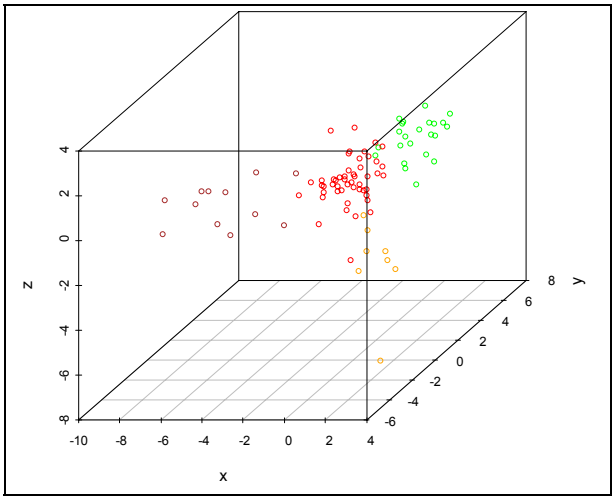
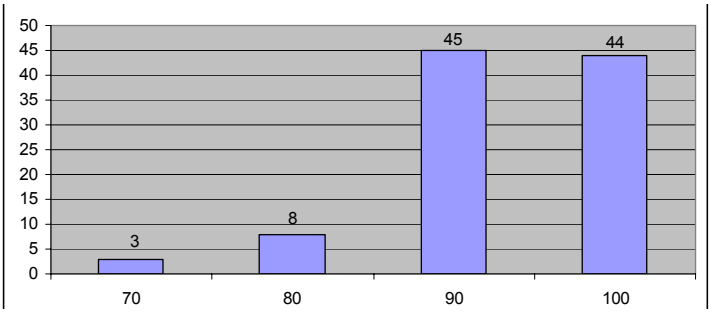


Figure 4.100: 3D plot of the three LD scores dimensions

In order to thoroughly test the predictivity of LDA, 100 "Leave-Ten-Out" tests have been performed, intending that 10 of the 92 summer samples have been randomly excluded from the LDA analysis and successively predicted using the Canonical Variates in this way calculated. The results of such a validation test are summarized in Graph 4.7.



Graph 4.7: Number of successes reached with the Leave-ten-out tests

It results that 89 out the 100 "Leave-Ten-Out" tests predict correctly more than 90% of the excluded samples (the weighted mean is 93%).

According to previous considerations, Naomi and Shiren samples may be considered as belonging to a single class; thus, the LDA has been performed on such a differently categorized dataset improving the predictivity of the statistical system from 89.13% to 93.48%, still using 26 PCs explaining 90% of the total variance. The discriminatory capability of LDA on the summer samples analyzed at 400 MHz is shown in Figure 4.101.

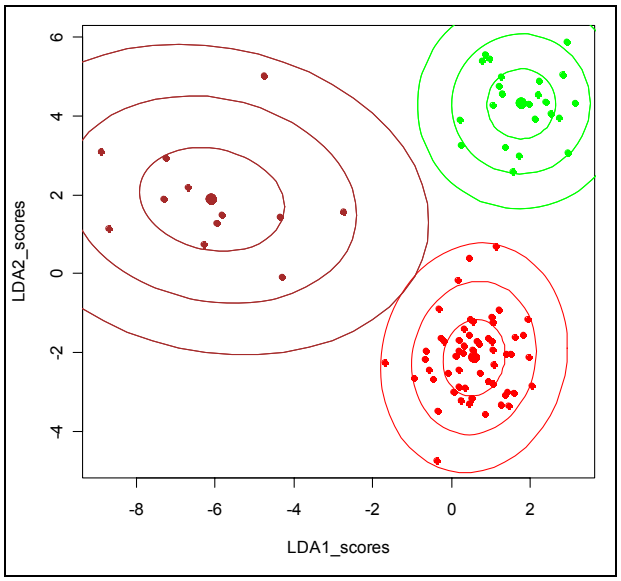


Figure 4.101: LDA performed by using 26 PCS explaining 90% of variance and keeping together Shiren and Naomi in the same category

Also for the summer samples, a study focused on the effect of the salinity of water used to irrigate cherry tomatoes cultivated in the Pachino area has been performed by LDA.

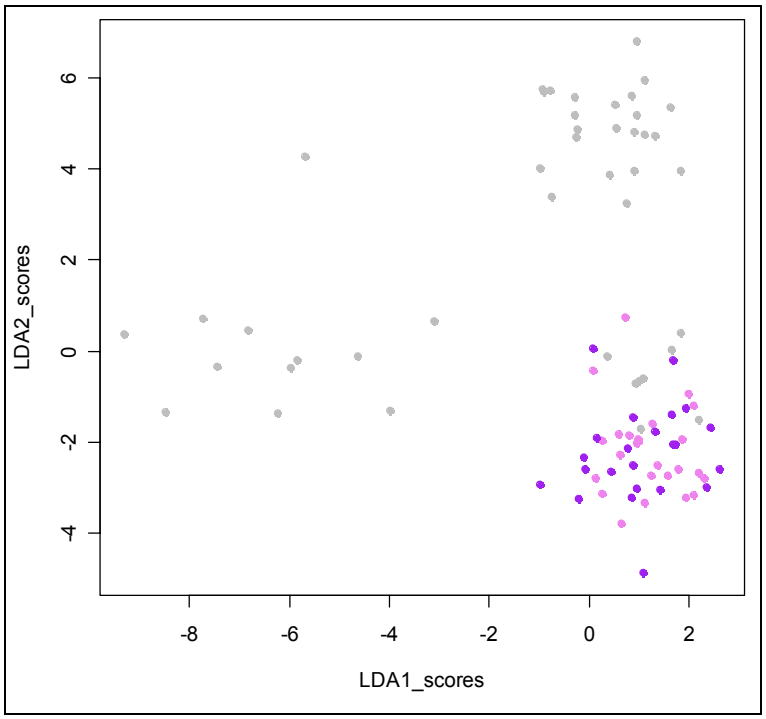


Figure 4.102: LD1-2 scores plot of summer samples with samples from Pachino area color labeled according to their salinity

Also in this case, the results shown in Figure 4.102 are identical to those already seen for other previously described studies performed on other dataset, confirming that salinity has no effect on sample distribution.

4.2.5 Inter-laboratory check test

A set of 11 cherry tomato extracts, already prepared and analyzed in our laboratory, all cultivated in the area of Pachino, have been analyzed, by another operator, with a 400 MHz NMR spectrometer located in an external laboratory. The results of such a check test are illustrated below, considering winter and summer samples separately.

4.2.5.1 Winter

5 samples, out of the 11 being considered for such a test, are extracted by tomatoes harvested in the winter season: 3 of these belong to the Naomi cultivar while the remaining 2 are chosen among the samples of the Shiren cultivar. The internal winter samples are labeled according to the codes already used for chemometric data processing described in section 4.2.3. The new added external winter samples are marked with crossed squares, colored in two different gradations of green: dark green for **Naomi** samples (to be compared with blue ones) and light green for **Shiren** samples (to be compared with cyan ones). FID processing has been performed on these external samples in the same way as for the internal samples. Once that data have been aligned and normalized, PCA analysis originated the results shown in Figure 4.103.

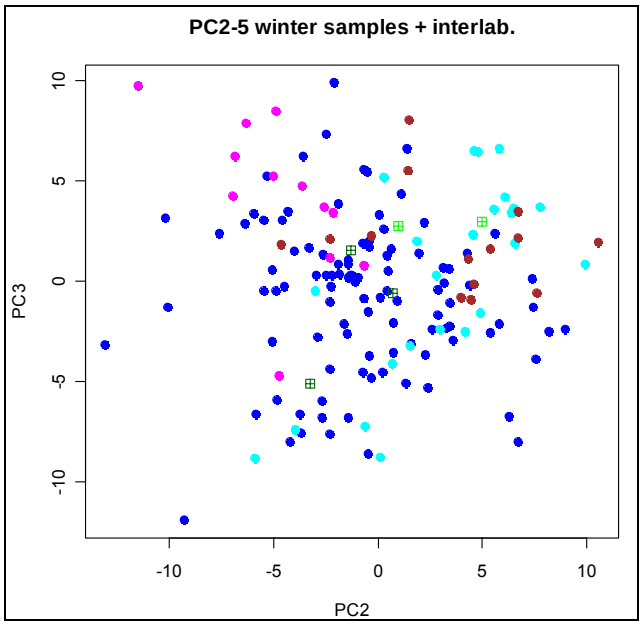


Figure 4.103: PCA of all winter samples together with external samples

It is noticeable that, even though a sharp separation is not present among classes, the external samples preferentially fall within the region preferentially occupied by the internal samples corresponding to the same category. This tendency, roughly seen with PCA, became more significant proceeding with the chemometric analysis and performing the LDA step.
The external samples have been projected on the LD scores spaces using the Canonical Variates matrix obtained by analyzing only the internal samples. In this way also the predictivity of the system has been tested since the test samples are used as unknown samples. The results of LDA analysis are shown in Figure 4.104.

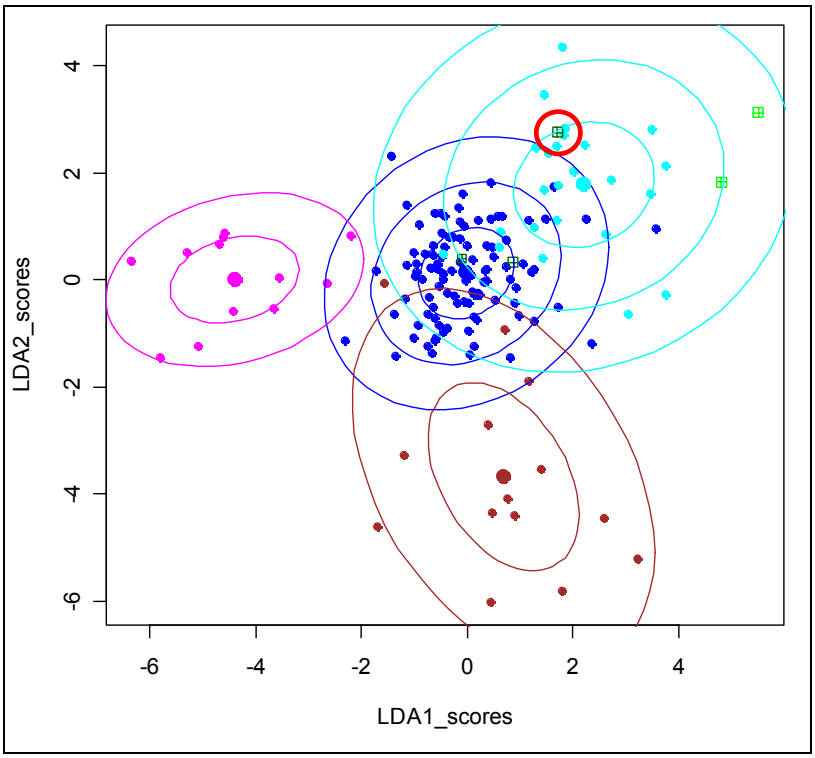
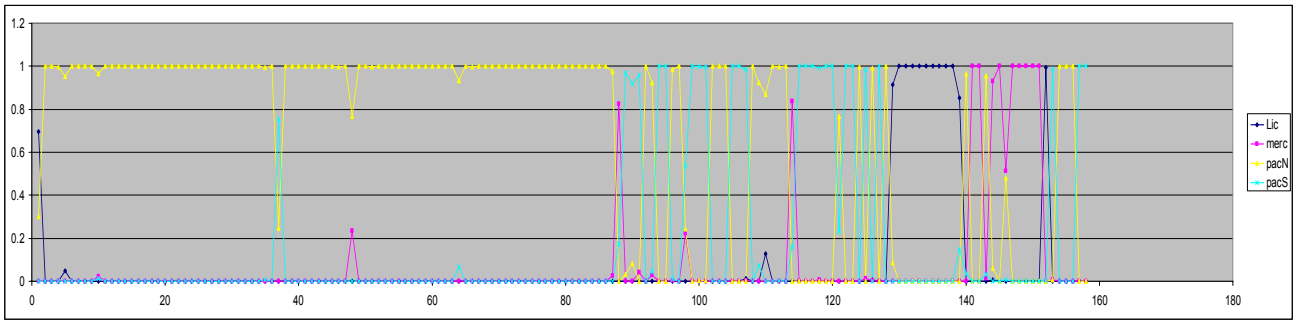


Figure 4.104: LDA of all winter samples together with external samples

Although one internal sample of the Naomi category (circled red in the figure) has been erroneously assigned to the Shiren group, all the external samples have been correctly predicted by considering all the LD scores space (only the plane LDA1-2 is shown in figure). The correct assignment has been predicted by performing the LOO test whose result is shown in Graph 4.8. In such a graph the external samples correspond to the last 5 points in the graph and all of them are rightly assigned. The classes of 144 samples out of the 158 used to perform LDA analysis have been correctly assigned, thus resulting in a predictivity equal to 91.1%.



Graph 4.8: Leave one out test performed on winter samples, including external samples, after LDA on PCs describing 90% of the total variance

4.2.5.2 Summer

The remaining 6 external samples, out of 11 analyzed in the external laboratory, originated from extracts of tomatoes harvested in summer: 3 of these belong to the Naomi cultivar, the other 3 are chosen among the samples of the Shiren cultivar.

Also in this case, the internal summer samples are labeled according the codes already described for the chemometric data processing of summer samples described in section 4.2.4. The new added external summer samples are marked with crossed squares colored in two different colors: blue for **Naomi** samples (to be compared with red ones) and cyan for **Shiren** samples (to be compared with orange ones).

FID processing has been performed on these external samples in the same way as for the internal samples. Once that data have been aligned and normalized, PCA analysis originated the results shown in Figure 4.105.

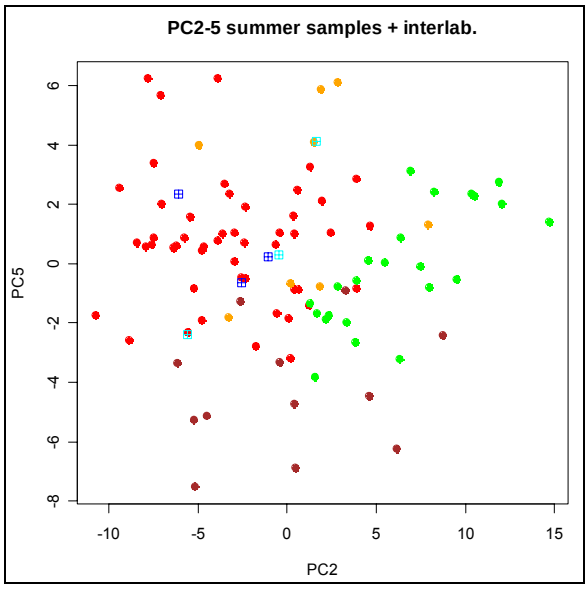


Figure 4.105: PCA of all summer samples together with external samples

Also for summer samples it is observable that the external samples preferentially fall within the region preferentially occupied by the internal samples corresponding to the same category. This inter-laboratory consistency, roughly seen with PCA, became more evident with LDA. The latter has been performed on summer samples in two different ways: i) the first one by including the internal samples in the data matrix being subjected to LDA, and ii) the second one by excluding them from data matrix before LDA, afterward they are just projected on the LD scores space. The results of such a multivariate analysis are shown in Figures 4.106 (i) and 4.107 (ii), respectively.

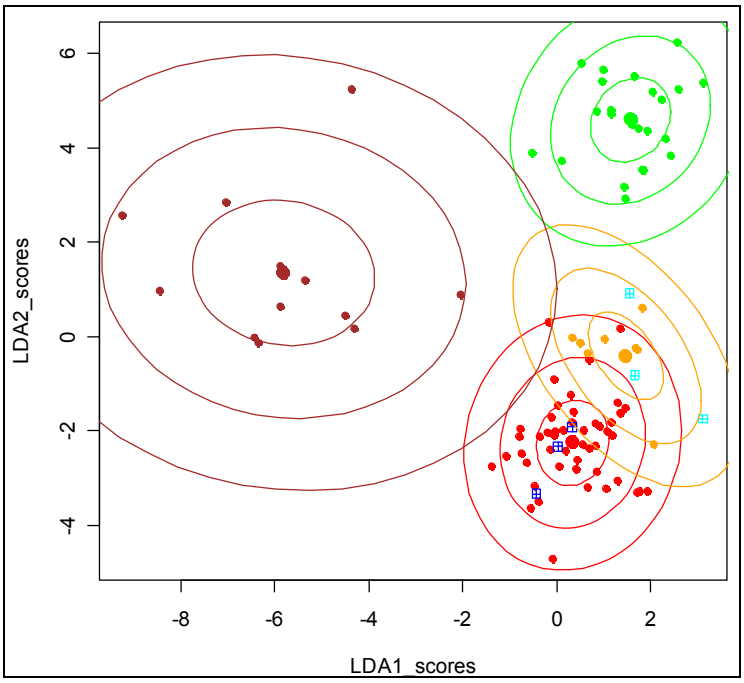


Figure 4.106: LDA performed on all summer samples together with external samples

As far it concerns the first approach, the correct inclusion of the external samples in the space of LDA confined to the second Mahalanobis distance of their own group is not surprising, being such samples included as trial set in LDA. The second approach, according which the external samples are excluded from the trial set and used only as test set, has been applied in order to test the equivalence of the internal and the external samples. In this way, this study should answer to the question whether samples are well recognized according to their original geographical origin, even if analyzed with a different NMR spectrometer.

Figure 4.107 illustrates the results of such second approach, which are almost coincident with those obtained by application of the first one, in terms of sample distribution over the LD space.

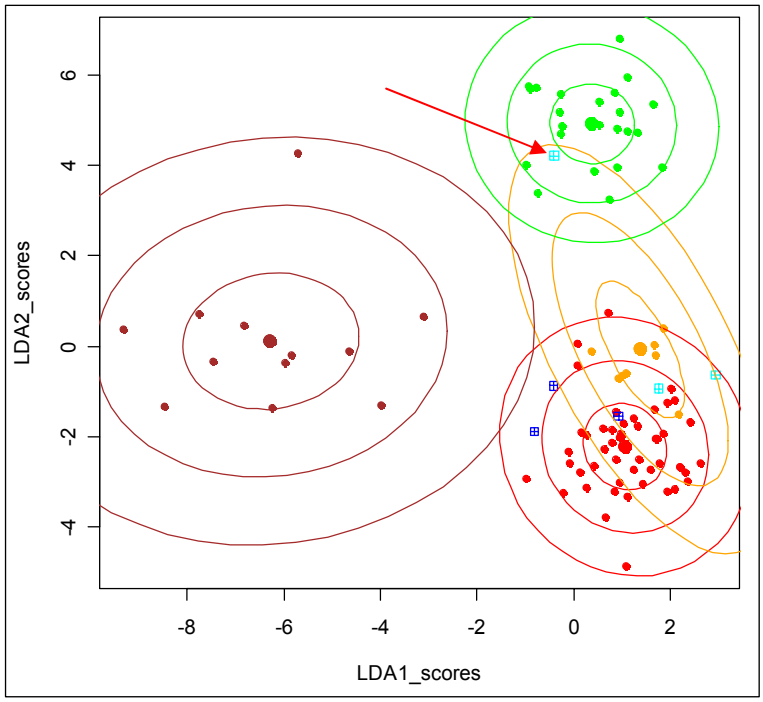
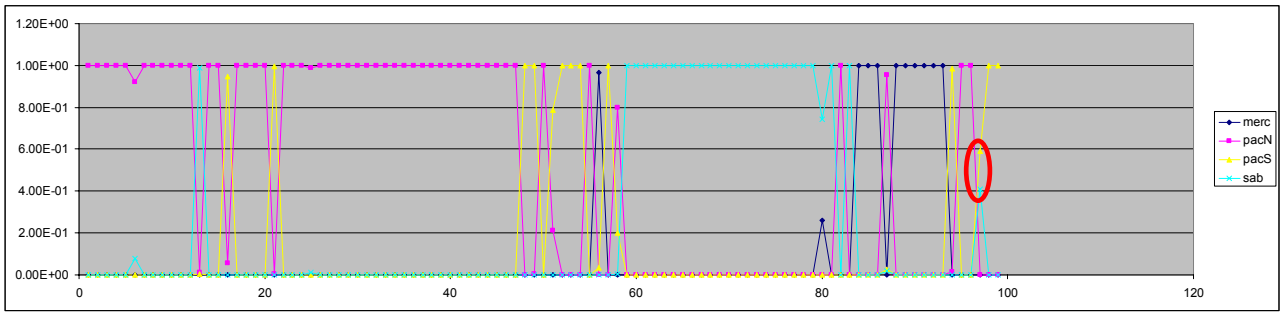


Figure 4.107: LDA performed on all summer samples together with external samples that have been used as test set

Differently from what has been found with external winter samples, not all the test samples result correctly assigned, since one of these falls in the class of the samples coming from Sabaudia (pointed by a red arrow in the figure). The same ambiguity on assigning the right category to the external samples is contained in Graph 4.9, showing the predictivity calculated with the LOO test.



Graph 4.9: Leave one out test performed on summer samples, including external samples, after LDA on PCs describing 90% of the total variance

In such a graph, the external samples are the last 6 ones. Here, even if sample n. 96 is rightly assigned to Shiren class, there is for it a relevant component of Sabaudia character as highlighted by the red ellipse. Nevertheless, the classes of 88 samples out of the 98 used to perform LDA analysis have been correctly assigned, resulting in predictivity equal to 89.80%. Although this result is appreciable, not all the external samples are predicted to belong to the same category of the corresponding internal one. This fact highlights the necessity to pay great attention to the instrumental set up as well as to the acquisition procedures which must be homogeneously set among different operators (shimming and tuning steps are annoying and more patient operators ensure better results on the spectra). The chemical modifications intervening during the gap of time elapsing between the internal and the external analyses may by excluded, since a kinetic study has been performed in order to verify the stability of extracts, showing that the spectra are perfectly superimposable at a distance of a week between two acquisitions (data not shown).

In the present study, the effect of different sample preparations among more operators has also taken into account: two operators have prepared two samples from the same tomato powder and analyzed on the same spectrometer. The comparison of the resulting spectra confirms that there is any appreciable difference, as predictable by considering the very simple method of sample preparation.

5 CONCLUSIONS

The research work described along the present thesis, is relative to the assessment of the quality in terms of chemical composition expressed as simple numerical parameters, and to the exploitation of such parameters to differentiate vegetable products with respect to their geographical origin.

The case study is referred to the "pomodoro di Pachino", a variety of cherry tomato cultivated in a restricted area of the southern coast of Sicily, protected by a European quality mark , which guarantee its geographical origin.

Often the mark is given to a product which is defined of good quality by means of criterions not always objectively measurable. In the present case, for example, the quality of tomatoes may be related to the way of its cultivation, which must satisfy some defined procedures. However, the mark is mainly related to its geographic origin, not necessarily confined to an areal with homogeneous pedoclimatic conditions, thus discriminating among neighboring regions which are only divided by political definitions. In this case, while it is credible that tomatoes cultivated in Sabaudia, an area in the central region of Italy, may be qualitatively different from those produced in the Pachino area, some doubts may arise for the legitimacy to discriminate between tomatoes cultivated in this latter area and those coming from Licata, another area remaining close to Pachino.

During the three years of the research work, whose results are presented in this thesis, a large amount of NMR spectroscopic data has been collected in a comprehensive dataset, becoming the basis for the development of a procedure which aims at the characterization of the quality of a unique vegetable product recognized and appreciated by consumers.

The plentiful spectroscopic dataset has been derived from the analysis of more than 270 samples of lyophilized cherry tomato fruits. Modern chemometric methods have been applied on this dataset to extract the necessary information for supporting, with objective measures, the decision previously taken without scientific bases, that tomatoes from Pachino are actually different from others of different origin. The results of the present research indicate that the tomatoes produced in this area, indeed, are different in terms of chemical quality from the other tomatoes studied for comparison and coming from different origins. It is also worth noting here that the method has allowed to establish that the quality of production is differentiable according to the climatic seasons of harvesting, as emerging along the three years of the study. This finding holds for both the product of Pachino and the other

tomatoes. Once the two seasons of production, summer and winter, are kept separate, non negligible differences emerge among tomatoes with different origin.

The results clearly point out that it is not a single molecular component to make the difference rather, like in an orchestra, it is the whole ensemble of instruments to play a good symphony. From the chemical point of view, it is the harmonious whole of substances to produce a differentiation in the quality of tomatoes having different origin.

Three additional fundamentals emerge from the present study:

1) it is necessary that tomatoes from Pachino are produced according to standardized procedures of cultivation and harvesting. In fact, the samples picked randomly without criteria of good quality, do not have a precise connotation in the space of parameters built through the use of chemical descriptors based on NMR. For this reason it is necessary, for a greater tutelage of the product and for its recognizability, that the standards of production are maintained constant;

2) the effect of the salinity of the irrigating water, so far characterizing the product even in the lowest limit of its range because it constitutes a selective factor for plant growth, has not a decisive influence on the quality within the same area of Pachino. The present research work has shown that there are not substantial differences among productions irrigated with waters with conducibility ranging between 1500 and 5000 μS/m. This allows to establish that 1500 μS/m is a value related to an extreme salinity already constituting a selective condition. These adverse growth conditions, which limit the development, are believed to bring the fruit to have a high concentration of components and a rich taste.

3) the discriminant power of the spectroscopic method is dependent from the strength of the spectrometer, as it emerges from the comparison of the results of the Linear Discriminant Analysis performed on the data recorded at 200 MHz and at the more powerful 400 MHz. With both magnetic field strength, a comparable capture of the season effect is obtained. However the weaker instrument gives the smaller predictivity (about 80%) in cataloguing the samples to the correct category of origin, whilst the most powerful spectrometer gives a correct prediction up to 90%. The method, still requiring to be optimized, however only marginally is affected by some inter-laboratory effects.

In conclusion, the present thesis describes a research work which adds, to the panorama of the foods of vegetable origin largely studied such as wine, oil and fruit's juices, another product that receives a good gastronomic reputation, and has an added value which is recognized by European Community with the quality mark. The same approach is applicable to other foods of vegetable origin.

In perspective, the research should continue in the identification of those molecular components responsible of the greatest part of the discriminating power evidenced by the multivariate analysis. Currently, non hyphenated methods are also studied which couple the NMR spectroscopy to other classical methods of separation such as HLPC and GC.

REFERENCES

[1] G.P.Blanch, M.Del Mar Caja, M.L.R.Del Castillo, M.Herraiz, Comparison of Different Methods for the Evaluation of the Authenticity of Olive Oil and Hazelnut Oil , *J. Agric. Food Chem.* **1998**, *46*, 3153.

[2] B.M.Silva, P.B.Andrade, G.C.Mendes, P.Valentao, R.M.Seabra, M.A.Ferreira, Analysis of phenolic compounds in the evaluation of commercial quince jam authenticity , *J. Agric. Food Chem.* **2000**, *48*, 2853.

[3] P.Mouly, E.M.Gaydou, A.Auffray, Simultaneous separation of flavanone glycosides and polymethoxylated flavones in citrus juices using liquid chromatography , *J. Chromatogr. A* **1998**, *800*, 171.

[4] M.L.Bengoechea, A.I.Sancho, B.Bartolome, I.Estrella, C.Gomez Cordoves, M.T.Hernandez, Phenolic Composition of Industrially Manufactured Pure?es and Concentrates from Peach and Apple Fruits , *J. Agric. Food Chem.* **1997**, *45*, 4071.

[5] A.M.Smith, S.Nakai, Classification of cheese varieties by multivariate analysis of HPLC profiles , *Can. Inst. Food Sci. Technol. J.* **1990**, *23*, 53.

[6] I.McDowell, S.Taylor, C.Gay, The phenolic pigment composition of black tea liquors - Part I: Predicting quality , *J. Sci. Food Agric.* **1995**, *69*, 467.

[7] I.McDowell, S.Taylor, C.Gay II, The phenolic pigment composition of black tea liquors - Part II: Discriminating origin , *J. Agric. Food Chem.* **1995**, *69*, 475.

[8] M.C.Parrilla, G.A.lez, F.J.Heredia, A.M.Troncoso, Differentiation of Wine Vinegars Based on Phenolic Composition , *J. Agric. Food Chem.* **1997**, *45*, 3487.

[9] H.K.Sivertsen, B.Holen, F.Nicolaysen, E.Risvik, Classification of French red wines according to their geographical origin by the use of multivariate analyses , *J. Sci. Food Agric.* **1999**, *79*, 107.

[10] J.M.F.Nogueira, A.M.D.Nascimento, Analytical characterization of Madeira wine , *J. Agric. Food Chem.* **1999**, *47*, 566.

[11] T.Konig, P.Schreier, Application of multivariate statistical methods to extend the authenticity control of flavour constituents of apple juice , *Zeitschrift fur Lebensmittel -Untersuchung und -Forschung* **1999**, *208*, 130.

[12] H.S.Soedjak, Colorimetric micromethod for protein determination with erythrosin B , *ANAL. BIOCHEM.* **1994**, *220*, 142.

[13] M.M.Bradford, A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein dye binding , *ANAL. BIOCHEM.* **1976**, *72*, 248.

[14] A.W.Johnson, D.J.McCaldin, The reaction of ninhydrin with cyclic R-imino acids , *J. Chem. Soc.* **1958**, 817.

[15] J.M.Wallace, P.F.Fox, Rapid spectrophotometric and fluorimetric methods for monitoring nitrogenous (proteinaceous) compounds in cheese and cheese fractions: A review , *Food Chem.* **1998**, *62*, 217.

[16] M.A.Fernandez-Muino, M.T.Sancho, S.Muniategui, J.F.Huidobro, M.P.nchez, J.Simal-Lozano, Direct enzymatic analysis of glycerol in honey: A simplified method , *J. Sci. Food Agric.* **1996**, *71*, 141.

[17] I.Mato, J.F.Huidobro, M.P.Sanchez, S.Muniategui, M.A.Fernandez-Muino, M.T.Sancho, Enzymatic Determination of Total D-Gluconic Acid in Honey , *J. Agric. Food Chem.* **1997**, *45*, 3550.

[18] A.Val, J.F.Huidobro, M.P.Sanchez, S.Muniategui, M.A.Fernandez-Muino, M.T.Sancho, Enzymatic Determination of Galactose and Lactose in Honey , *J. Agric. Food Chem.* **1998**, *46*, 1381.

[19] I.Mato, J.F.Huidobro, M.P.nchez, S.Muniategui, M.ndez, M.T.Sancho, Enzymatic determination of L-malic acid in honey , *Food Chem.* **1998**, *62*, 503.

[20] M.J.Baxter, H.M.Crews, M.J.Dennis, I.Goodall, D.Anderson, The determination of the authenticity of wine from its trace element composition , *Food Chem.* **1997**, *60*, 443.

[21] J.Leblebici, M.Volkan, Sample Preparation for Arsenic, Copper, Iron, and Lead Determination in Sugar , *J. Agric. Food Chem.* **1998**, *46*, 173.

[22] W.A.Simpkins, H.Louie, M.Wu, M.Harrison, D.Goldberg, Trace elements in Australian orange juice and other products , *Food Chem.* **2000**, *71*, 423.

[23] S.Prats-Moya, N.Teruel, V.Berenguer-Navarro, M.L.Martin-Carratala, Inductively Coupled Plasma Application for the Classification of 19 Almond Cultivars Using Inorganic Element Composition , *J. Agric. Food Chem.* **1997**, *45*, 2093.

[24] J.M.Fresno, B.Prieto, R.Urdiales, R.Sarmiento, J.Carballo, Mineral content of some Spanish cheese varieties. Differentiation by source of milk and by variety from their content of main and trace elements , *J. Sci. Food Agric.* **1995**, *69*, 339.

[25] J.L.Rodriguez-Otero, P.Paseiro, J.Simal, A.Cepeda, Mineral content of the honeys produced in Galicia (North-West Spain) , *Food Chem.* **1994**, *49*, 169.

[26] P.Vinas, I.Lopez Garcia, M.Lanzon, M.Hernandez Cordoba, Direct Determination of Lead, Cadmium, Zinc, and Copper in Honey by Electrothermal Atomic Absorption Spectrometry using Hydrogen Peroxide as a Matrix Modifier , *J. Agric. Food Chem.* **1997**, *45*, 3952.

[27] S.D.Kelly, Using stable isotope ratio mass spectrometry (IRMS) in food authentication and traceability , *Food authenticity and traceability* **2003**, 156.

[28] K.Rozanski, L.Araguas Araguas, R.Gonfiantini, Relation between long-term trends of oxygen-18 isotope composition of precipitation and climate , *Science* **1992**, *258*, 981.

[29] J.W.White, Internal standard stable carbon isotope ratio method for determination of C-4 plants sugars in honey: Collaborative study, and evaluation of improved protein preparation procedure , *J. Assoc. Off. Anal. Chem.* **1992**, *75*, 543.

[30] W.Meier-Augenstein, Applied gas chromatography coupled to isotope ratio mass spectrometry , *J. Chromatogr. A* **1999**, *842*, 351.

[31] J.W.White, K.Winters, Honey protein as internal standard for stable carbon isotope ratio detection of adulteration of honey , *J Assoc Off Anal Chem* **1989**, *72*, 907.

[32] J.W.White, K.Winters, P.Martin, A.Rossmann, Stable Carbon Isotope Ratio Analysis of Honey: Validation of Internal Standard Procedure for Worldwide Application , *J AOAC Int* **1998**, *81*, 610.

[33] V.Y.Grinberg, T.V.Burova, T.Haertle, V.B.Tolstoguzov, Interpretation of DSC data on protein denaturation complicated by kinetic and irreversible effects , *J. Biotechnol.* **2000**, *79*, 269.

[34] A.Kulmyrzaev, C.Bryant, D.J.McClements, Influence of sucrose on the thermal denaturation, gelation, and emulsion stabilization of whey proteins , *J. Agric. Food Chem.* **2000**, *48*, 1593.

[35] P.Relkin, Differential scanning calorimetry: A useful tool for studying protein denaturation , *Thermochim. Acta* **1994**, *246*, 371.

[36] P.L.M.Barreto, L.H.Beirao, M.S.Soldi, V.Soldi, Studies on Differential Scanning Calorimetry and Thermogravimetry of Tilapia (Oreochromis nilotica) Surimi, Surimi/Starch and Surimi/Starch/Carrageenan Systems , *J Food Sci Technol* **2000**, *37*, 265.

[37] E.Chiotelli, A.Rolee, M.Le Meste, Effect of sucrose on the thermomechanical behavior of concentrated wheat and waxy corn starch-water preparations , *J. Agric. Food Chem.* **2000**, *48*, 1327.

[38] K.Morikawa, K.Nishinari, Effects of concentration dependence of retrogradation behaviour of dispersions for native and chemically modified potato starch , *Food Hydrocolloids* **2000**, *14*, 395.

[39] K.Morikawa, K.Nishinari, Rheological and DSC studies of gelatinization of chemically modified starch heated at various temperatures , *Carbohydr Polym* **2000**, *43*, 241.

[40] K.Takahashi, H.Kondo, H.Kuroiwa, Y.Yokote, M.Hattori, Reversible Thermal Transition of Soluble Branched Chains from Slightly Acid-treated Potato Starch , *Biosci. Biotechnol. Biochem.* **2000**, *64*, 1365.

[41] Y.J.Wang, L.Wang, Structures and Properties of Commercial Maltodextrins from Corn, Potato, and Rice Starches , *starch* **2000**, *52*, 296.

[42] S.Chevallier, P.Colonna, Thermal analysis of protein-starch interactions at low moisture contents , *Sci. Aliments* **1999**, *19*, 167.

[43] V.K.Villwock, A.C.Eliasson, J.Silverio, J.N.BeMiller, Starch-lipid interactions in common, waxy, ae du, and ae su2 maize starches examined by differential scanning calorimetry , *Cereal Chem.* **1999**, *76*, 292.

[44] S.D.Clas, C.R.Dalton, B.C.Hancock, Differential scanning calorimetry: Applications in drug development , *Pharm. Sci. Technol. Today* **1999**, *2*, 311.

[45] A.John, P.N.Shastri, Studies on food macromolecules by differential scanning calorimetry : A critical appraisal , *J Food Sci Technol* **1998**, *35*, 1.

[46] J.M.Aguilera, H.Gloria, Determination of Oil in Fried Potato Products by Differential Scanning Calorimetry , *J. Agric. Food Chem.* **1997**, *45*, 781.

[47] P.Cornillon, Characterization of Osmotic Dehydrated Apple by NMR and DSC , *Food Sci. Technol.* **2000**, *33*, 261.

[48] R.Curini, G.D'Ascenzo, S.Materazzi, E.Chiacchierini, S.ngelis Curtis, M.C.Lucchetti, Characterization of coffees by thermoanalytical methods , *Ann. Chim.* **1991**, *81*.

[49] H.Gloria, J.M.Aguilera, Assessment of the Quality of Heated Oils by Differential Scanning Calorimetry , *J. Agric. Food Chem.* **1998**, *46*, 1363.

[50] H.D.Goff, The use of thermal analysis in the development of a better understanding of frozen food stability , *Pure Appl. Chem.* **1995**, *67*, 1801.

[51] W.L.Kerr, D.S.Reid, The use of stepwise differential scanning calorimetry for thermal analysis of foods , *Thermochim. Acta* **1994**, *246*, 299.

[52] E.A.Niediek, Amorphous sugar, its formation and effect on chocolate quality , *Manuf. Confect.* **1991**, *71*, 91.

[53] G.Roudaut, C.Dacremont, M.Le Meste, Influence of water on the crispness of cereal-based foods: Acoustic, mechanical, and sensory studies , *J. Texture Stud.* **1998**, *29*, 199.

[54] L.Sbarato, Fast analytical methods , *Ind. Aliment.* **1996**, *35*, 43.

[55] J.Farkas, C.csi-Farkas, Application of differential scanning calorimetry in food research and food quality assurance , *Journal of Thermal Analysis* **1996**, *47*, 1787.

[56] M.Fiala, K.O.Honikel, Application of differential calorymetry. Detection of additives in meat , *Fleischforschung* **1995**, *75*, 1013.

[57] E.Coni, M.Di Pasquale, P.Coppolelli, A.Bocca, Detection of animal fats in butter by differential scanning calorimetry: A pilot study , *JAOCS J Am Oil Chem Soc* **1994**, *71*, 807.

[58] C.Cordella, J.F.Antinelli, C.Aurieres, J.P.Faucon, D.Cabrol-Bass, N.Sbirrazzuoli, Use of differential scanning calorimetry (DSC) as a new technique for detection of adulteration in honeys. 1. Study of adulteration effect on honey thermal behavior , *J. Agric. Food Chem.* **2002**, *50*, 203.

[59] M.H.Tunick, DSC analysis of dairy products , *Proceedings of the 26th Conference of the North American Thermal Analysis Society, Cleveland* **1998**, *7*, 13.

[60] Q.X.Ni, T.M.Eads, Analysis by proton NMR of Changes in liquid-phase and solid-phase components during ripening of banana , *Journal of Agricultural and Food Chemistry«* **1993**, *41*, 1035.

[61] G.G.Martin, V.Hanote, M.Lees, Y.L.Martin, Interpretation of Combined 2H SNIF/NMR and 13C SIRA/MS Analyses of Fruit Juices to Detect Added Sugar , *J AOAC Int* **1996**, *79*, 62.

[62] G.E.Maciel, *NMR in industrial process control and quality control, Nuclear Magnetic Resonance in Modern Tecnology*, Kluver Academic,Netherlands, **1994**.

[63] T.M.Shaw, R.H.Elsken, Investigation of Proton Magnetic Resonance Line Width of Sorbed Water , *The Journal of Chemical Physics* **1953**, *21*, 565.

[64] W.Derbyshire, *Applications of low-resolution NMR, Analytical Application of Spettroscopy*, Royal Society of Chemistry, London, **1988**.

[65] M.C.Vackier, D.N.Rutledge, Influence of Temperature, pH, Water Content, Gel Strength and Their Interaction on NMR Relaxation of Gelatines. I- Analysis of the Calculated Relaxation Times , *J. Magn. Reson. Anal.* **1996**, *2*, 311.

[66] M.C.Vackier, D.N.Rutledge, Influence of Temperature, pH, Water Content, Gel Strength and Their Interaction on NMR Relaxation of Gelatines. I- Analysis of the Calculated Relaxation Times , *J. Magn. Reson. Anal.* **1996**, *2*, 321.

[67] M.C.Vackier, B.P.Hills, D.N.Rutledge, An NMR Relaxation Study of the State of Water in Gelatin Gels , *J. Magn. Reson.* **1999**, *138*, 36.

[68] S.I.Cho, C.H.Chung, Development of a nondestructive moisture sensor using proton NMR , *Transactions of the American Society of Agricultural Engineers* **1997**, *40*, 1129.

[69] R.B.Wettstrom, Wide-line NMR for product and process control in Sat industries , *American Oil Chemists' Society* **1971**, *48*, 15.

[70] T.F.Conway, *J. Am. Oil Chem. Soc.* **1971**, *48*, 54.

[71] P.Desbois, D.Le Botlan, Proton Low-Field NMR Measurements on Crackers , *Journal of Food Science* **1994**, *59*, 1088.

[72] "Resonance Instruments", **2007**.

[73] "Bruker NMR", **2007**.

[74] L.Laghi, M.A.Cremonini, G.Placucci, S.Sykora, K.Wright, B.Hills, A proton NMR relaxation study of hen egg quality , *Magn. Reson. Imaging* **2005**, *23*, 501.

[75] ND.Overfield, in ***Egg*** *quality - current problems and recent advances* (Eds: R.Wells, C.Belyavin), Butterworths, London, **1987**.

[76] S.Alessandri, F.Capozzi, M.A.Cremonini, C.Luchinat, G.Placucci, F.Savorani, M.Turano, in *Magnetic Resonance in Food Science* (Eds: S.B.Engelsen, P.S.Belton, H.J.Jakobsen), The Royal Society of Chemistry,Thomas Graham house, Science park, Milton Road, Cambridge, **2005**.

[77] F.Capozzi, M.A.Cremonini, C.Luchinat, G.Placucci, C.Vignali, A Low Frequency 1H-NMR External Unit for the Analysis of Large Foodstuff Samples , *J. Magn. Reson.* **1999**, *138*, 277.

[78] G.Eidmann, R.Savelsberg, P.mler, B.mich, The NMR MOUSE, amobile universal surface explorer , *Journal of Magnetic Resonance - Series A* **1996**, *122*, 104.

[79] B.Blumich, P.Blumler, G.Eidmann, A.Guthausen, R.Haken, U.Schmitz, K.Saito, G.Zimmer, The NMR-MOUSE: Construction, excitation, and applications , *Magn. Reson. Imaging* **1998**, *16*, 479.

[80] F.Balibanu, K.Hailu, R.Eymael, D.E.Demco, B.Blumich, Nuclear Magnetic Resonance in Inhomogeneous Magnetic Fields , *J. Magn. Reson.* **2000**, *145*, 246.

[81] H.T.Pedersen, S.Ablett, D.R.Martin, M.J.D.Mallett, S.B.Engelsen, Application of the NMR-MOUSE to food emulsions , *J. Magn. Reson.* **2003**, *165*, 49.

[82] E.Veliyulin, C.van der Zwaag, W.Burk, U.Erikson, *In vivo* determination of fat content in Atlantic salmon (*Salmo salar*) with a mobile NMR spectrometer , *J. Sci. Food Agric.* **2005**, *85*, 1299.

[83] H.Stork, A.Gadke, N.Nestle, Single-sided and semisingle-sided NMR sensors for highly diffusive samples: Application to bottled beverages , *J. Agric. Food Chem.* **2006**, *54*, 5247.

[84] G.J.Martin, M.L.Martin, **2007**.

[85] G.J.Martin, C.Guillou, N.Naulet, S.Brun, Y.Tep, J.C.Cabanis, M.T.Cabanis, P.Sudrad, Control of origin and enrichment of wine by specific isotope analysis – study of different methods for wine enrichment , *Sci. Aliment* **1986**, *6*, 386.

[86] G.J.Martin, S.Brun, Application de la résonance magnétique nucléaire du deutérium au contrôle des moúts, des moûts concentrés, du sucre de raisin et des vins , *Bull. O. I. V.* **1987**, 671.

[87] G.G.Martin, R.Wood, G.J.Martin, Detection of Added Beet Sugar in Concentrated and Single Strength Fruit Juices by Deuterium Nuclear Magnetic Resonance (SNIF-NMR1 Method): Collaborative Study , *J AOAC Int* **1996**, *79*, 917.

[88] P.Lindner, E.Bermann, B.Gamarnik, Characterization of Citrus Honey by Deuterium NMR , *J. Agric. Food Chem.* **1996**, *44*, 139.

[89] E.Jamin, N.Naulet, G.J.Martin, Multi-element and multi-site isotopic analysis of nicotine from tobacco leaves , *PLANT CELL ENVIRON.* **1997**, *20*, 589.

[90] L.F.Gladden, Review article number 46. Nuclear magnetic resonance in chemical engineering: Principles and applications , *Chemical Engineering Science* **1994**, *49*, 3339.

[91] M.J.McCarthy, K.L.McCarthy, Applications of magnetic resonance imaging to food research , *Magn. Reson. Imaging* **1996**, *14*, 799.

[92] S.J.Schmidt, X.Sun, J.B.Litchfield, Applications of Magnetic Resonance Imaging in Food Science , *Crit. Rev. Food Sci. Nutr.* **1996**, *36*, 357.

[93] C.J.Clark, L.N.Drummond, J.S.MacFall, Quantitative NMR imaging of kiwifruit (Actinidia deliciosa) during growth and ripening , *J. Sci. Food Agric.* **1998**, *78*, 349.

[94] B.Zion, M.J.McCarthy, P.Chen, Real-time detection of pits in processed cherries by magnetic resonance projections , *Lebensm. -Wiss. Technol.* **1994**, *27*, 457.

[95] B.Zion, P.Chen, M.J.McCarthy, Detection of bruises in magnetic resonance images of apples , *Computers and Electronics in Agriculture* **1995**, *13*, 289.

[96] S.L.Duce, T.A.Carpenter, L.D.Hall, Nuclear magnetic resonance imaging of fresh and frozen courgettes , *J Food Eng* **1992**, *16*, 165.

[97] R.Ruan, K.Chang, P.L.Chen, R.G.Fulcher, E.D.Bastian, A Magnetic Resonance Imaging Technique for Quantitative Mapping of Moisture and Fat in a Cheese Block , *J. Dairy Sci.* **1998**, *81*, 9.

[98] P.S.Belton, I.J.Colquhoun, B.P.Hills, Applications of NMR to Food Science , *Annu. Rep. NMR Spectrosc.* **1993**, *26*, 1.

[99] A.M.Gil, P.S.Belton, B.P.Hills, Applications of NMR to food science , *Ann. Reps NMR Spectrosc* **1996**, *32*, 1.

[100] H.Kovacs, D.Moskau, M.Spraul, Cryogenically cooled probes - A leap in NMR technology , *Prog. Nucl. Magn. Reson. Spectrosc.* **2005**, *46*, 131.

[101] W.R.Croasmun, R.M.K.Carlson, *Two-Dimensional NMR Spectroscopy: Applications for Chemists and Biochemists Methods in Stereochemical Analysis Edited by*, VCH, New York, **1994**.

[102] J.B.Lambert, E.P.Mazzola, *Nuclear Magnetic Resonance Spectroscopy: An Introduction to Principles, Applications, and Experimental Methods*, Pearson Prentice-Hall, New Jork, **2004**.

[103] T.W.M.Fan, Metabolite profiling by one- and two-dimensional NMR analysis of complex mixtures , *Prog. Nucl. Magn. Reson. Spectrosc.* **1996**, *28*, 161.

[104] O.Fiehn, Metabolomics - The link between genotypes and phenotypes , *Plant. Mol. Biol.* **2002**, *48*, 155.

[105] S.Ablett, Overview of NMR Application in Food Science , *Trends in Food Science and Technology* **1992**, *3*, 246.

[106] H.J.C.Berendsen, Rationale for using NMR to study water relations in foods and biological tissues , *Trends Food Sci. Technol.* **1992**, *3*, 202.

[107] A.J.Fischman, R.G.Tompkins, Perspective on the applications of NMR in nutrition , *Trends in Food Science and Technology* **1992**, *3*, 220.

[108] M.J.Gidley, High-resolution solid-state NMR of food materials , *Trends in Food Science and Technology* **1992**, *3*, 231.

[109] C.Guillou, G.Remaud, G.J.Martin, Applications of NMR to the characterization and authentication of foods and beverages , *Trends in Food Science and Technology* **1992**, *3*, 197.

[110] J.O'Brien, Authentication and quality assessment of food products , *Trends Food Sci. Techn.* **1992**, *3*, 19.

[111] R.Sacchi, L.Mannina, P.Fiordiponti, P.Barone, L.Paolillo, M.Patumi, A.Segre, Characterization of Italian Extra Virgin Olive Oils Using 1H-NMR Spectroscopy , *J. Agric. Food Chem.* **1998**, *46*, 3947.

[112] N.Ogrinc, I.J.ir, J.E.Spangenberg, J.Kidric?, The application of NMR and MS methods for detection of adulteration of wine, fruit juices, and olive oil. A review , *Anal. Bioanal. Chem.* **2003**, *376*, 424.

[113] R.Sacchi, F.Addeo, I.Giudicianni, L.Paolillo, Analisys of the positional distribution of fatty acids in olive oil triacylglyceerols by high-resolution 13C-NMR of the carbonyl region , *Italian Journal of Food Science* **1992**, *2*, 117.

[114] R.Sacchi, F.Addeo, S.Spagna Musso, L.Paolillo, I.Giudicianni, A high-resolution 13C-NMR study of vegetable margarines , *Italian Journal of Food Science* **1995**, *1*, 27.

[115] R.Sacchi, F.Addeo, L.Paolillo, 1H and 13C NMR of virgin olive oil. An overview , *Magn. Reson. Chem.* **1997**, *35*, S133-S145.

[116] K.F.Wellenberg, Quantitative high-resolution !£C nuclear magnetic resonance of the olefinic and carbonyl carbons of edile vegetable oils , *American Oil Chemists' Society* **1990**, *67*, 487.

[117] M.S.F.Lie Ken Jie, J.Mustafa, High-resolution nuclear magnetic resonance spectroscopy - Applications to fatty acids and triacylglycerols , *Lipids* **1997**, *32*, 1019.

[118] H.E.Jingren, C.Santos-Buelga, A.M.S.Silva, N.Mateus, V.De Freitas, Isolation and structural characterization of new anthocyanin-derived yellow pigments in aged red wines , *J. Agric. Food Chem.* **2006**, *54*, 9598.

[119] A.Avenoza, J.H.Busto, N.Canal, J.M.Peregrina, Time course of the evolution of malic and lactic acids in the alcoholic and malolactic fermentation of grape must by quantitative 1H NMR (qHNMR) spectroscopy , *J. Agric. Food Chem.* **2006**, *54*, 4715.

[120] B.Baderschneider, P.Winterhalter, Isolation and characterization of novel benzoates, cinnamates, flavonoids, and lignans from Riesling wine and screening for antioxidant activity , *J. Agric. Food Chem.* **2001**, *49*, 2788.

[121] F.Goncalves, A.Heyraud, M.N.De Pinho, M.Rinaudo, Characterization of white wine mannoproteins , *J. Agric. Food Chem.* **2002**, *50*, 6097.

[122] A.D'Agostina, G.Boschin, F.Bacchini, A.Arnoldi, Investigations on the high molecular weight foaming fractions of espresso coffee , *J. Agric. Food Chem.* **2004**, *52*, 7118.

[123] Y.Okada, M.Semma, S.Kitahata, A.Ichikawa, Isolation and characterization of two positional isomers of novel heterogeneous branched cyclomaltohexaoses (?-cyclodextrins) having a D-galactobiosyl residue on the side chain , *Carbohydr. Res.* **2004**, *339*, 2875.

[124] S.Mika, G.Ratsch, J.Weston, B.Scholkopf, K.R.Muller, "Fisher discriminant analysis with kernels", August 23, 1999.

[125] W.W.Widmer, P.F.Cancalon, S.Nagy, Methods for determining the adulteration of citrus juices , *Trends Food Sci. Technol.* **1992**, *3*, 278.

[126] S.Li, C.Y.Lo, C.T.Ho, Hydroxylated polymethoxyflavones and methylated flavonoids in sweet orange (Citrus sinensis) peel , *J. Agric. Food Chem.* **2006**, *54*, 4176.

[127] S.Tiziani, S.J.Schwartz, Y.Vodovotz, Profiling of carotenoids in tomato juice by one- and two-dimensional NMR , *J. Agric. Food Chem.* **2006**, 6094.

[128] A.P.Sobolev, E.Brosio, R.Gianferri, A.L.Segre, Metabolic profile of lettuce leaves by high-field NMR spectra , *Magn. Reson. Chem.* **2005**, *43*, 625.

[129] F.D.Gunstone, Information on the composition of fats from their high-resolution 13C nuclear magnetic resonance spectra , *JAOCS J Am Oil Chem Soc* **1993**, *70*, 361.

[130] P.Kalo, A.Kemppinen, I.inen, Determination of positional distribution of butyryl groups in milkfat triacylglycerols, triacylglycerol mixtures, and isolated positional isomers of triacylglycerols by gas chromatography and 1H nuclear magnetic resonance spectroscopy , *Lipids* **1996**, *31*, 331.

[131] M.R.Van Calsteren, C.Barr, P.Angers, J.Arul, ^{13}C NMR of Triglycerides , *Bull. Magn. Reson.* **1996**, *18*, 175.

[132] G.Andreotti, E.Trivellone, R.Lamanna, A.D.Luccia, A.Motta, Milk identification of different species: 13C-NMR spectroscopy of triacylglycerols from cows and buffaloes' milks , *J. Dairy Sci.* **2000**, *83*, 2432.

[133] C.Beavallet, J.P.Renou, Applications of NMR spectroscopy in meat research , *Trends in Food Science and Technology* **1992**, *3*, 241.

[134] M.E.Amato, G.Ansanelli, S.Fisichella, R.Lamanna, G.Scarlata, A.P.Sobolev, A.Segre, Wheat flour enzymatic amylolysis monitored by in situ (1) H NMR spectroscopy , *J Agric Food Chem* **2004**, *52*, 823.

[135] C.J.Gorter, L.J.F.Broer, Negative results of an attempt to observe direct nuclear magnetic in solids , *Physica* **1942**, *9*, 591.

[136] W.G.Proctor, F.C.Yu, The dependence of a nuclear magnetic resonance frequency upon chemical compounds , *Physical Review* **1950**, *77*, 717.

[137] R.M.Silverstein, F.X.Webster, D.J.Kiemle, in *Identificazione spettrometrica di composti organici* (Ed: Casa Editrice Ambrosiana), Milano, **2006**, Ch. 3.

[138] M.A.Cremonini, G.Bonaga, in *Encyclopedia of Life Support Systems (EOLSS)*, Eolss, Oxford ,UK, **2004**.

[139] S.Wold, Chemometrics; what do we mean with it, and what do we want from it? *Chemometr. Intell. Lab. Syst.* **1995**, *30*, 109.

[140] J.Forshed, R.J.O.Torgrip, K.M.Aberg, B.Karlberg, J.Lindberg, S.P.Jacobsson, A comparison of methods for alignment of NMR peaks in the context of cluster analysis , *Journal of Pharmaceutical and Biomedical Analysis* **2005**, *38*, 824.

[141] T.Brekke, O.M.Kvalheim, E.Sletten, Prediction of physical properties of hydrocarbon mixtures by partial-least squares calibration of carbon-13 nuclear magnetic resonance data , *Anal. Chim. Acta* **1989**, *223*, 123.

[142] E.Holmes, J.K.Nicholson, A.W.Nicholls, J.C.Lindon, S.C.Connor, S.Polley, J.Connelly, The identification of novel biomarkers of renal toxicity using automatic data reduction techniques and PCA of proton NMR spectra of urine , *Chemometr. Intell. Lab. Syst.* **1998**, *44*, 245.

[143] M.Spraul, P.Neidig, U.Klauck, P.Kessler, E.Holmes, J.K.Nicholson, B.C.Sweatman, S.R.Salman, R.D.Farrant, E.Rahr, C.R.Beddell, J.C.Lindon, Automatic reduction of NMR spectroscopic data for statistical and pattern recognition classification of samples , *Journal of Pharmaceutical and Biomedical Analysis* **1994**, *12*, 1215.

[144] R.J.O.Torgrip, M.Aberg, B.Karlberg, S.P.Jacobsson, Peak alignment using reduced set mapping , *J. Chemometr.* **2003**, *17*, 573.

[145] R.Stoyanova, A.W.Nicholls, J.K.Nicholson, J.C.Lindon, T.R.Brown, Automatic alignment of individual peaks in large high-resolution spectral data sets , *J. Magn. Reson.* **2004**, *170*, 329.

[146] J.Forshed, I.Schuppe-Koistinen, S.P.Jacobsson, Peak alignment of NMR signals by means of a genetic algorithm , *Analytica Chimica Acta* **2003**, *487*, 189.

[147] G.C.Lee, D.L.Woodruff, Beam search for peak alignment of NMR signals , *Analytica Chimica Acta* **2004**, *513*, 413.

[148] K.M.Aberg, R.J.O.Torgrip, S.P.Jacobsson, Extensions to peak alignment using reduced set mapping: Classification of LC/UV data from peptide mapping , *J. Chemometr.* **2004**, *18*, 465.

[149] R.J.O.Torgrip, J.Lindberg, M.Linder, B.Karlberg, S.P.Jacobsson, J.Kolmert, I.Gustafsson, I.Schuppe-Koistinen, New modes of data partitioning based on PARS peak alignment for improved multivariate biomarker/biopattern detection in 1H-NMR spectroscopic metabolic profiling of urine , *Metabolomics* **2006**, *2*, 1.

[150] G.Tomasi, F.Van Den Berg, C.Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data , *J. Chemometr.* **2004**, *18*, 231.

[151] A.Kassidas, J.F.MacGregor, P.A.Taylor, Synchronization of Batch Trajectories Using Dynamic Time Warping , *AIChE Journal* **1998**, *44*, 864.

[152] D.Bylund, R.Danielsson, G.Malmquist, K.E.Markides, Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data , *J. Chromatogr. A* **2002**, *961*, 237.

[153] N.P.V.Nielsen, J.M.Carstensen, J.Smedsgaard, Aligning of single and multiple wavelength chromatographic profiles for chemometric data using correlation optimized warping , *J. Chromatogr. A* **1998**, *805*, 17.

[154] J.T.W.E.Vogels, A.C.Tas, J.Venekamp, J.Van Der Greef, Partial linear fit: A new NMR spectroscopy preprocessing tool for pattern recognition applications , *J. Chemometr.* **1996**, *10*, 425.

[155] K.J.Johnson, B.W.Wright, K.H.Jarman, R.E.Synovec, High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis , *J. Chromatogr. A* **2003**, *996*, 141.

[156] A.Craig, O.Cloarec, E.Holmes, J.K.Nicholson, J.C.Lindon, Scaling and normalization effects in NMR spectroscopic metabonomic data sets , *Anal. Chem.* **2006**, *78*, 2262.

[157] F.Dieterle, A.Ross, G.Schlotterbeck, H.Senn, Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in1H NMR metabonomics , *Anal. Chem.* **2006**, *78*, 4281.

[158] R.Bro, A.K.Smilde, Centering and scaling in component analysis , *J. Chemometr.* **2003**, *17*, 16.

[159] R.A.van den Berg, H.C.Hoefsloot, J.A.Westerhuis, A.K.Smilde, M.J.van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data , *BMC Genomics* **2006**, *7*, 142.

[160] C.W.Hanson III, E.R.Thaler, Electronic nose prediction of a clinical pneumonia score: Biosensors and microbes , *Anesthesiology* **2005**, *102*, 63.

[161] D.L.Massart, B.G.M.Vandeginste, S.N.Deming, Y.Michotte, L.Kaufman, *Chemometrics: A Textbook*, Elsevier Science Publisher, Amsterdam, **1988**.

[162] E.Oja, Principal components, minor components, and linear neural networks , *Neural Networks* **1989**, *5*, 927.

[163] J.C.Hayton, D.G.Allen, V.Scarpello, Factor Retention Decisions in Exploratory Factor Analysis: A Tutorial on Parallel Analysis , *Org. Res. Methods* **2004**, *7*, 191.

[164] H.F.Kaiser, The application of elecrtonic computers to factor analysis , *Edu. Psycol. Meas.* **1960**, *20*, 141.

[165] R.B.Cattel, The scree test for the number of factors , *Multivar. Behav. Res.* **1966**, *1*, 245.

[166] M.W.Browne, A comparison of factor analytic techniques , *Psychometrika* **1968**, *33*, 267.

[167] R.B.Cattel, J.Jaspers, A general plasmode for factor analytic exercises and research , *Multivar. Behav. Res. Monogr.* **1967**, *3*, 1.

[168] R.A.Hakstian, W.T.Rogers, R.B.Cattel, The behaviour of numbers-of-factors rules with simulated data , *Multivar. Behav. Res.* **1982**, *17*, 193.

[169] L.B.Tucker, R.F.Koopman, R.L.Linn, Evaluation of factor analytic research procedures by means of simulated correlation matrices , *Psychometrika* **1969**, *34*, 421.

[170] A.M.Martinez, A.C.Kak, PCA versus LDA , *IEEE Trans Pattern Anal Mach Intell* **2001**, *23*, 228.

[171] H.Abdi, in *Encyclopedia of Measurement and Statistics* (Ed: N.J.Salkind), Thousand Oaks (CA), **2007**.

[172] G.J.McLachlan, *Discriminant Analysis And Statistical Pattern Recognition*, Wiley-Interscience, **2004**.

[173] R.A.Fisher, The use of multiple measurements in taxonomic problems , *Annals of Eugenics* **1936**, *7*, 179.

[174] R.Wehrens, R.de Gelder, G.J.Kemperman, B.Zwanenburg, L.M.C.Buydens, Molecular challenges in modern chemometrics , *Analytica Chimica Acta* **1999**, *400*.

[175] G.C.Cawley, N.L.C.Talbot, Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers , *Pattern Recognition* **2003**, *36*, 2585.

[176] I.Duarte, A.Barros, P.S.Belton, R.Righelato, M.Spraul, E.Humpfer, A.M.Gil, High-resolution NMR spectroscopy and multivariate analysis for the caracterization of beer , *J. Agric. Food Chem.* **2002**, *50*, 2475.

[177] I.F.Duarte, M.Godejohann, U.Braumann, M.Spraul, A.M.Gil, Application of NMR spectroscopy and LC-NMR/MS to the identification of carbohydrates in beer , *J. Agric. Food Chem.* **2003**, *51*, 4847.

[178] A.M.Gil, I.F.Duarte, M.Godejohann, U.Braumann, M.Maraschin, M.Spraul, Characterization of the aromatic composition of some liquid foods by nuclear magnetic resonance spectrometry and liquid chromatography with nuclear magnetic resonance and mass spectrometric detection , *Analytica Chimica Acta* **2003**, *488*, 35.

[179] N.E.Tzouros, I.S.Arvanitoyannis, Agricultural produces: Synopsis of employed quality control methods for the authentication of foods and application of chemometrics for the classification of foods according to their variety or geographical origin , *Crit. Rev. Food Sci. Nutr.* **2001**, *41*, 287.

[180] P.S.Belton, I.J.Colquhoun, E.K.Kemsley, I.Delgadillo, P.Roma, M.J.Dennis, M.Sharman, E.Holmes, J.K.Nicholson, M.Spraul, Application of chemometrics to the ^{1}H NMR spectra of apple juices: discrimination between apple varieties , *Food Chem.* **1998**, *61*, 207.

[181] J.T.W.E.Vogels, L.Terwel, A.C.Tas, F.Van Den Berg, F.Dukel, J.Van Der Greef, Detection of adulteration in orange juices by a new screening method using proton NMR spectroscopy in combination with pattern recognition techniques , *J. Agric. Food Chem.* **1996**, *44*, 175.

[182] G.Le Gall, M.Puaud, L.J.Colquhoun, Discrimination between orange juice and pulp wash by 1H nuclear magnetic resonance spectroscopy: Identification of marker compounds , *J. Agric. Food Chem.* **2001**, *49*, 580.

[183] A.J.Charlton, W.H.H.Farrington, P.Brereton, Application of 1H NMR and multivariate statistics for screening complex mixtures: Quality control and authenticity of instant coffee , *J. Agric. Food Chem.* **2002**, *50*, 3098.

[184] A.D.Shaw, A.Di Camillo, G.Vlahov, A.Jones, G.Bianchi, J.Rowland, D.B.Kell, Discrimination of the variety and region of origin of extra virgin olive oil using 13C NMR and multivariate calibration with variable reduction , *Analytica Chimica Acta* **1997**, *348*, 357.

[185] C.Fauhl, F.Reniero, C.Guillou, 1H NMR as a tool for the analysis of mixtures of virgin olive oil with oils of different botanical origin , *Magn. Reson. Chem.* **2000**, *38*, 436.

[186] L.Mannina, M.Patumi, N.Proietti, D.Bassi, A.L.Segre, Geographical characterization of Italian extra virgin olive oils using high-field 1H NMR spectroscopy , *J. Agric. Food Chem.* **2001**, *49*, 2687.

[187] L.Mannina, G.Dugo, F.Salvo, L.Cicero, G.Ansanelli, C.Calcagni, A.Segre, Study of the cultivar-composition relationship in sicilian olive oils by GC, NMR, and statistical methods , *J. Agric. Food Chem.* **2003**, *51*, 120.

[188] L.Forveille, J.Vercauteren, D.N.Rutledge, Multivariate statistical analysis of two-dimensional NMR data to differentiate grapevine cultivars and clones , *Food Chem.* **1996**, *57*, 441.

[189] G.E.Pereira, J.P.Gaudillere, C.Van Leeuwen, G.Hilbert, O.Lavialle, M.Maucourt, C.Deborde, A.Moing, D.Rolin, 1H NMR and chemometrics to characterize mature grape berries in four wine-growing areas in Bordeaux, France , *J. Agric. Food Chem.* **2005**, *53*, 6382.

[190] I.J.Kosir, J.Kidric, Use of modern nuclear magnetic resonance spectroscopy in wine analysis: Determination of minor compounds , *Analytica Chimica Acta* **2002**, *458*, 77.

[191] M.A.Brescia, V.Caldarola, A.De Giglio, D.Benedetti, F.P.Fanizzi, A.Sacco, Characterization of the geographical origin of Italian red wines based on traditional and nuclear magnetic resonance spectrometric determinations , *Analytica Chimica Acta* **2002**, *458*, 177.

[192] M.A.Brescia, I.J.ir, V.Caldarola, J.Kidric, A.Sacco, Chemometric classification of Apulian and Slovenian wines using 1H NMR and ICP-OES together with HPICE data , *J. Agric. Food Chem.* **2003**, *51*, 21.

[193] I.F.Duarte, A.Barros, C.Almeida, M.Spraul, A.M.Gil, Multivariate Analysis of NMR and FTIR Data as a Potential Tool for the Quality Control of Beer , *J. Agric. Food Chem.* **2004**, *52*, 1031.

[194] I.S.Arvanitoyannis, N.E.Tzouros, Implementation of quality control methods in conjunction with chemometrics toward authentication of dairy products , *Crit. Rev. Food Sci. Nutr.* **2005**, *45*, 231.

[195] A.K.Mattoo, A.P.Sobolev, A.Neelam, R.K.Goyal, A.K.Handa, A.L.Segre, Nuclear magnetic resonance spectroscopy-based metabolite profiling of transgenic tomato fruit engineered to accumulate spermidine and spermine reveals enhanced anabolic and nitrogen-carbon interactions , *Plant Physiol.* **2006**, *142*, 1759.

[196] G.Belletti, T.Burgassi, A.Marescotti, S.Scaramuzzi, "The effects of certification costs on the success of a PDO/PGI", presented at Quality Management and Quality Assurance in Food Chains, University of Göttingen (Germany), May 2, 2005.

[197] european commission, "Quality products catch the eye: PDO, PGI and TSG", **2007**.

[198] L.Costato, *Compendio di diritto alimentare*, CEDAM, Padova, **2002**.

[199] G.Belletti, T.Burgassi, E.Manco, A.Marescotti, A.Pacciani, S.Scaramuzzi, in *La valorizzazione economica delle tipicità locali tra localismo e globalizzazione* (Ed: C.Ciappei), Florence University Press, Firenze, **2006**.

[200] MESTRELAB RESEARCH, "Mestre-C", **2007**.

[201] R Development Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Wien, **2004**.

[202] W.N.Venables, D.M.Smith, *An Introduction to R*, Network Theory Limited, Bristol, **1992**.

[203] A.P.Sobolev, A.Segre, R.Lamanna, Proton high-field NMR study of tomato juice , *Magn. Reson. Chem.* **2003**, *41*, 237.

[204] R.H.Dougherty, P.E.Nelson, Effects of pH on quality of stored tomato juice , *Journal of Food Science* **1974**, *39*, 254.

Appendix A: Algorithms in R language

In this section the most important R language scripts, developed during my PhD, are reported. Obviously not all the command employed in any situation of data processing have been illustrated but only the most significant. For further material or utilization advices please refer to f.savorani@unibo.it

200 MHz data processing

```
rm(list = ls())
library(MASS)
#nomi<-read.table("72estivi.txt", sep=";", colClasses = "character", header=T)
#nomi<-read.table("267_ok_origine.txt", sep=";", colClasses = "character", header=T)
#nomi<-read.table("114_est_LDA.txt", sep=";", colClasses = "character", header=T)
nomi<-read.table("155inv_LDA.txt", sep=";", colClasses = "character", header=T)
#nomi<-read.table("108_estnornd_LDA.txt", sep=";", colClasses = "character", header=T)
#nomi<-read.table("271peranalisi.txt", sep=";", colClasses = "character", header=T)
#nomi<-read.table("prova281.txt", sep=";", colClasses = "character", header=T)
#nomi<-read.table("274_origine.txt", sep=";", colClasses = "character", header=T)
#nomi<-read.table("167inv_azienda.txt", sep=";", colClasses = "character", header=T)
#nomi<-read.table("142invernali.txt", sep=";", colClasses = "character", header=T)
num<-nomi
ncamp<-nrow(nomi)
campioni<-nomi

################################
###### spectral regions for alignment ######
################################
all1<-3250; all2<-3310 # c(3200:3350)(4700:5100)
reg1<-3250; reg2<-3310
allinea<-c(all1:all2)
regione1<-c(reg1:reg2)
inf1<-min(regione1); sup1<-max(regione1)

punti<-6820
rall1<-500
rall2<-5000
rall3<-50000
rall4<-350000

tabella<-matrix(nrow=ncamp, ncol=punti)
tabace<-matrix(data=0, nrow=ncamp, ncol=punti)

####################################
# spectra scanning in order to evaluate their shifting     #
# with respect to a target spectrum (sample 38, X=17) #
####################################
windows(5.7,3.55, xpos=5.8E2, ypos=0)
```

```
n<-c(1:punti)
for (z in 1:ncamp) {
spec<-read.table(nomi[z,1], sep=";", dec=".", header=F)
spectrum<-as.matrix(spec)
titolo<-paste("spettro",as.character(z))
tabella[z,]<-matrix(spectrum[n])
}
plot(tabella[z,],xlim=c(all1,all2), ylim=c(0,3000),type="l", main=titolo)
#lines(tabella[z,])
##########
# slow motion #
##########
a<-0
for (z in 1:rall4){
a<-a+1
}
#############
# slow motion end #
#############
}

windows(5.7,3, xpos=5.8E2, ypos=3.95E2)

#individuates acetate signal in any sample#

ace<-0
for (i in 1:ncamp) {
ace[i]<-max(tabella[i,5434:5440])
titolo<-paste("spettro acetato",as.character(i))
##########
# slow motion #
##########
a<-0
for (z in 1:rall3){
a<-a+1
}
#############
# slow motion end #
#############
plot(tabella[i,5430:5445], ylim=c(0,85000), type="p", main=titolo)
}

#########################################################
# scales any data point over the maximum of the acetate signal of the correspondent sample #
#########################################################

windows(5.7,3, xpos=0, ypos=3.95E2)
tabace<-tabella/ace
for (i in 1:ncamp) {
titolo<-paste("spettro",as.character(i))
plot(tabace[i,allinea], ylim=c(0,0.15), type="l", main=titolo)
##########
# slow motion #
##########
a<-0
for (z in 1:rall3){
```

```
a<-a+1
}
#############
# slow motion end #
#############
}

##########################################
# algorithm for signals alignment based on least square minimization #
##########################################

funtarget<-c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
minsqm<-matrix(nrow=1,ncol=ncamp)
windows(5.7,2.8, xpos=0, ypos=0)
index<-0
spettroref<-6
for (h in 1:ncamp) {
minore<-1E35
j<-0
for (z in -15:15){
j<-j+1
limite1<-all1+z
limite2<-all2+z
scalato1<-tabella[spettroref,allinea]/max(tabella[spettroref,allinea])
scalato2<-tabella[h,c(limite1:limite2)]/max(tabella[h,c(limite1:limite2)])
target<-scalato1-scalato2
funtarget[j]<-sum(abs(target))

#plot(scalato1, type="l", ylim=c(-1,1), main=as.character(h))
#lines(abs(target), type="l", col="red")
#lines(scalato2, type="l", col="red")

if (sum(abs(target))<minore){
minore<-sum(abs(target))
index[h]<-z
}
}
##########
# slow motion #
##########
a<-0
for (z in 1:rall3){
a<-a+1
}
#############
# slow motion end #
#############

plot(funtarget,type="l", main=as.character(h))
minsqm[h]<-min(funtarget)
}

barplot(index)
index1<- -index
```

```
##########################################
# algorithm for signals alignment over α-D-glucose downfield signal  #
##########################################

centro<-3278 #(signal position on the target spectrum)
maxvet<-matrix(data=0, ncol=1, nrow=ncamp)
reganomer<-c(3250:3310)
for (i in 1:ncamp){
for (j in reganomer){
if (tabace[i,j]==max(tabace[i,reganomer])) maxvet[i,]<- centro-j
}
}

index1<-maxvet

##############################
#Algorithm for generating aligned data matrix  #
##############################

allineanomer<- matrix(data=0, ncol=punti, nrow=ncamp)
for (i in 1:ncamp) {
shift<-index1[i]
if (shift>0) allineanomer[i,(1+shift):punti]<-tabace[i,1:(punti-shift)]
if (shift==0) allineanomer[i,]<-tabace[i,]
if (shift<0) allineanomer[i,1:(punti+shift)]<-tabace[i,(1-shift):punti]
}

##################################################
#Cutting useless external data points (keeping form point 11 to point 6810)#
##################################################

allineanomer1<-allineanomer[,c(11:6810)]

#################################
#Plotting of the aligned and reduced (6800 pts) spectra#
#################################

for (z in 1:ncamp) {
titolo<-paste("spettro",as.character(z))
plot(allineanomer1[z,], type="l", ylim=c(0,0.001), main=titolo)      #xlim=c(150,200), ylim=c(0,0.1)
#lines(allineanomer1[z,]

##########
# slow motion #
##########
a<-0
for (z in 1:rall4){
a<-a+1
}
#############
# slow motion end #
#############
}
####################
# average spectrum calculation  #
####################
```

```
spettromed<-apply(allineanomer1,2,mean)
barplot(spettromed, ylim=c(0,0.001))
allineati<-allineanomer1

################
# CS NORMALIZATION  #
################
riferiti<-allineati
riferiti[,c(3400:3650)]<-0
riferiti[,c(5400:5480)]<-0
coefCS<-rep(0,ncamp)
allineati.CS<-matrix(data=0, ncol=ncol(riferiti), nrow=ncamp)
for (t in 1:ncamp) {
rsum<-sum(riferiti[t,])
allineati.CS[t,]<-riferiti[t,]/rsum*1
coefCS[t]<-rsum/100
}
vedi<-apply(allineati.CS, 1, sum)
#max.CS<-which.max(allineati.CS)
 #up<-riferiti[max.CS]/allineati.CS[max.CS]

up<-mean(apply(riferiti, 1, sum))
allineati.CS<-allineati.CS*up

windows()
plot(allineati.CS[2,3000:6000], ylim=c(0,2e4), xaxt="n", type="l", main="200 MHz CS normalized and
aligned data", xlab="main spectral region (ppm 1.16-5.80)", ylab="signals intensity")
 for (i in 1:ncamp){
 lines(allineati.CS[i,3000:6000],col=i)
 }

windows()
plot(allineati.CS[2,(all1-10):(all2-10)], ylim=c(0,4000), xaxt="n", type="l", main="200 MHz CS
normalized and aligned data", xlab="alfa-D-Glucose doublet region (ppm 5.15-5.25)", ylab="signals
intensity")
 for (i in 1:ncamp){
 lines(allineati.CS[i,(all1-10):(all2-10)],col=i)
 }

windows()
plot(allineati.CS[2,4780:5050], ylim=c(0,8000), xaxt="n", type="l", main="200 MHz CS normalized
and aligned data", xlab="citrate region (ppm 2.54-3.04)", ylab="signals intensity")
 for (i in 1:ncamp){
 lines(allineati.CS[i,4780:5050],col=i)
 }

#########################################################
# data set binning into 200 integral regions of 34 data points each. Only values > 2e-4 are kept #
#########################################################

intervallo<-matrix(data=0, nrow=ncamp, ncol=200)
for (z in 1:ncamp) {
i<-1
for (i in 1:200){
for (j in 1:34){
k<-((i-1)*34)+j
```

```
a<-allineati.CS[z,k]
if (a>2e-4) intervallo[z,i]<-intervallo[z,i]+a
}
}
}

##############################
#REMOVING OF H2O AND ACETATE SIGNALS  #
##############################

scarto<-c(101:107,160,161)
perpca<-intervallo[,-scarto]

###########
# PCA analysis  #
###########

principali4<-prcomp(perpca,center=T,scale=T)
#nomi<-read.table("267_origine.txt", sep=";", colClasses = "character", header=T)
colori<-nomi[,3]
simbolo<-c(as.numeric(nomi[,4]))
y<-c(2,3)
windows()
plot(principali4$x[,y],pch=" ", col=colori, main="?")
text(principali4$x[,y], labels=c(1:ncamp), col=colori, cex=.6)


########################################
# pH dependent bins removing according to pH-dependence study #
# previously performed (data file "eliminareX(X.X).txt" needs)       #
########################################

eliminare1<-scan(file="eliminare(0.05).txt", sep=";")
eliminare<-c(104:106,eliminare1,160)
elimina<-sort(eliminare)
colnames(intervallo)<-c(1:200)
interbins<-intervallo[,-elimina]
dim(interbins)

####################
#PCA analysis over "interbins" #
####################

principali<-interbins
principali1<-prcomp(principali,center=F,scale=F)
dati<-read.table("72estivi.txt", sep=";", colClasses = "character", header=T)
colori<-c(as.character(dati[,2]))
simbolo<-c(as.numeric(dati[,3]))
y<-c(1,3)
plot(principali1$x[,y],pch=" ", col=colori, main="PC1-3; S=F, C=F") #plot con i numeri#
text(principali1$x[,y], labels=c(1:72), col=colori, cex=.6)
```

400 MHz data processing

```
# read sample's name from 'nomefile'.txt depending on sample's dataset

rm(list = ls())
library(MASS)
#campioni<-read.table("262bestcamp_stag.txt", sep=";", dec=".", header=T)
#campioni<-read.table("100campinv+lat2.txt", sep=";", dec=".", header=T)
#campioni<-read.table("67est_az3-4+sab2.txt", sep=";", dec=".", header=T)
campioni<-read.table("92bestcampest1.txt", sep=";", dec=".", header=T)
#campioni<-read.table("169campinv_random.txt", sep=";", dec=".", header=T)
#campioni<-read.table("166bestcampinv_rnd.txt", sep=";", dec=".", header=T)
#campioni<-read.table("153bestcampinv.txt", sep=";", dec=".", header=T)
#campioni<-read.table("158bestcampinv_inter.txt", sep=";", dec=".", header=T)
#campioni<-read.table("98campest_inter.txt", sep=";", dec=".", header=T)
#colClasses = "character"
names<-as.character(campioni[,1])
nomi<-as.matrix(names)
punti<-16384
pachini<-161

# reads all data of each sample and arranges them in a data matrix with ncamp rows and 16384
columns

ncamp<-nrow(nomi)
spettri<-matrix(data=0, nrow=ncamp, ncol=punti)
refacetato<-matrix(data=0, nrow=ncamp, ncol=punti)
coeface<-rep(0,ncamp)
for (z in 1:ncamp) {
tabella<-read.table(nomi[z,1], sep=";", dec=".", header=F)
# DEFINISCE GLI INTERVALLI E I VALORI PER LO SCALING SULL'ACETATO
tabella1<-as.matrix(tabella)
intace<-c(11711:12278)
acetato<-max(tabella1[intace])
coeface[z]<-acetato
spettri[z,]<-(tabella1)
# SCALE IN REFACETATO OVER ACETATE MAX
refacetato[z,]<-(tabella1/acetato)
}

#NOT USING REFACETATO BUT SPETTRI
refacetato<-spettri
#####################################
# VARIABLES ALIGNMENT OVER beta D GLUCOSE SIGNAL #
#####################################
centro<-8424 #( signal position on the target spectrum)
maxvet<-matrix(data=0, ncol=1, nrow=ncamp)
reganomer<-c(8360:8500)
for (i in 1:ncamp){
for (j in reganomer){
if (refacetato[i,j]==max(refacetato[i,reganomer])) maxvet[i,]<- centro-j
}
}
```

```
allineati<- matrix(data=0, ncol=punti, nrow=ncamp)
for (i in 1:ncamp) {
shift<-maxvet[i,1]
if (shift>0) allineati[i,(1+shift):punti]<-refacetato[i,1:(punti-shift)]
if (shift==0) allineati[i,]<-refacetato[i,]
if (shift<0) allineati[i,1:(punti+shift)]<-refacetato[i,(1-shift):punti]
}

##############################################
# Performs PCA over "riferiti" in order to individuate the average spectrum#
##############################################
noH2OAC<-c(1099:7900,8351:11899,12001:14698)
allineati2<-allineati[,noH2OAC]
pca<-prcomp(allineati2,center=T,scale=T)
y<-c(1,2)
centroide<-apply(pca$x[,y],2,mean)
stagione<-c(as.character(campioni[,3]))
camp<-campioni[,2]
colori<-c(as.character(campioni[,8]))
plot(pca$x[,y],pch=" ",cex=1.0, main="PCA; S=T, C=T")
text(pca$x[,y],labels=camp,col=colori, cex=.6)
text(centroide[1],centroide[2], labels="+")

matrice<-matrix(data=0, ncol=3, nrow=ncamp)
for (z in 1:ncol(matrice)) {
media<-mean(pca$x[,z])
for (n in 1:ncamp) {
scarto<-abs(pca$x[n,z]-media)
matrice[n,z]<-scarto
}
}

sumvet<-apply(matrice, 1, sum)
names(sumvet)<-c(1:ncamp)
best<-min(sumvet)
sumvet[sumvet>best]<-0
ermeio<-sumvet[sumvet!=0]
spref1<-names(ermeio)
spref<-as.numeric(spref1)

##########################################
# CS NORMALIZATION TO 100 and ORIGINAL VERTICAL SCALING #
##########################################
riferiti<-allineati
riferiti[,c(7901:8350)]<-0
riferiti[,c(11900:12000)]<-0
coefCS<-rep(0,ncamp)
allineati.CS<-matrix(data=0, ncol=ncol(riferiti), nrow=ncamp)
for (t in 1:ncamp) {
rsum<-sum(riferiti[t,])
allineati.CS[t,]<-riferiti[t,]/rsum*1
coefCS[t]<-rsum/100
}
vedi<-apply(allineati.CS, 1, sum)
#max.CS<-which.max(allineati.CS)       #if a scaling over the maximum is required
#up<-riferiti[max.CS]/allineati.CS[max.CS]
```

```
up<-mean(apply(riferiti, 1, sum))          #if a scaling over the mean is required
allineati.CS<-allineati.CS*up      #Coefficient for signal scaling up to real values

#######################################################
# Perform a difference minimization among spectra using spref spectrum as reference #
#######################################################
riferiti<-allineati
minimizza<-function(par){
sum(abs(ref-(test*par)))
}
colonna<-c(1:10)
riga<-c(1:ncamp)
Risutot<-matrix(0,ncamp,10,dimnames=list(riga, colonna))
limiti<-
c(1800,3500,3501,5500,5501,7700,8301,9110,9111,9810,9811,10200,10201,10660,10661,11220,112
21,11890,12000,14000)
zone<-matrix(limiti, nrow=2, ncol=10)
for (i in 1:ncamp) {
for (z in 1:10){
zona<-c(zone[1,z]:zone[2,z])
test<-riferiti[i,zona]
ref<-riferiti[spref,zona]
ris<-optimize(minimizza,c(-1e9,+1e9))
Risutot[i,z]<-ris$minimum
}
}

#plot(test[6155:6255], ylim=c(0,1000),type="l")
# write.table(Risutot,"Tab10coeff.txt",sep=";",dec=".")

medie<-apply(Risutot, 1,mean)
devst<-apply(Risutot, 1,sd)
risuscarti<-abs(Risutot-medie)
risubuoni<-(risuscarti-devst)
risubuoni[risubuoni>0]<-0
risubuoni[spref,]<-1
risubuoni[risubuoni<0]<-1
sumrisu<-apply(risubuoni, 2, sum)
sumgood<-sumrisu[sumrisu>=(ncamp*0.9)] #here a tolerance value can be set: now it is 10%
scelti<-c(as.numeric(names(sumgood)))

#####################################
# multiply each spectrum for the correspondent coefficient #
#####################################

preriferiti<-riferiti
coeff<-Risutot[,c(scelti)]
coefmedio<-apply(coeff,1,mean)
riferiti<-riferiti*coefmedio

####################
# PCA over every single point  #
####################
noH2OAC<-c(1099:7900,8351:11899,12001:14698)
perpca<-riferiti[,noH2OAC]
```

```
####################
# PRODUCES 200 BINS TABLE #
####################

n<-c(1099:14698)

#interesse<-riferiti[,n]
interesse<-allineati.CS[,n]

intervallo<-matrix(data=0, nrow=ncamp, ncol=200)
for (z in 1:ncamp) {
i<-1
for (i in 1:200){
for (j in 1:68){
k<-((i-1)*68)+j
a<-interesse[z,k]
intervallo[z,i]<-intervallo[z,i]+a
#if (a<0) intervallo[z,i]<-intervallo[z,i]+a #Use it if you want set a treshold for noise removing
}
}
}
colnames(intervallo)<-c(1:200)

                  ######
                 #PART I#
                  ######

###################################
#ALTERNATIVE 1: ALL THE SIGNAL ARE UTILIZED #
###################################

perpca<- intervallo

#####################################
#ALTERNATIVE 2: H2O and Acetate signals are excluded#
#####################################
noh2oacet<-c(1:101,107:159,161:200)
perpca<- intervallo[,noh2oacet]
perpca2<-perpca

###################################
#ALTERNATIVE 3: Only Bins selected by LDA are kept  #
###################################

perpca<- intervallo[,sort(PCbestLD)]

#####################################
#ALTERNATIVE 4: n Random bins are selected and taken #
#####################################
training<-sample(c(1:101,107:159,161:200), 33)
perpca<-intervallo[,training]

####################
# Principal Component Analysis  #
####################
```

```
principali4<-prcomp(perpca,center=T,scale=T)
colori<-c(as.character(campioni[,8]))
azienda<-c(campioni[,4])
camp<-campioni[,2]
stagione<-c(as.character(campioni[,3]))
cultivar<-campioni[,6]
y<-c(2,5)
windows()
plot(principali4$x[,y],pch=" ",col=colori,cex=1.2, main="?")
text(principali4$x[,y],labels=camp,col=colori, cex=.6)

#########################################################
# ALGORITHM FOR POSTERIOR SAMPLES PCA SCORES CALCULATION AND PLOTTING  #
#########################################################
> campioni2<-read.table("6campest_inter.txt", sep=";", dec=".", header=T)
> #campioni2<-read.table("5campinv_inter.txt", sep=";", dec=".", header=T)
> #campioni2<-read.table("13random_inv.txt", sep=";", dec=".", header=T)
> names<-as.character(campioni2[,1])
> nomi<-as.matrix(names)
> punti<-16384
> ncamp<-nrow(nomi)
> spettri<-matrix(data=0, nrow=ncamp, ncol=punti)
> for (z in 1:ncamp) {
> tabella<-read.table(nomi[z,1], sep=";", dec=".", header=F)
> tabella1<-as.matrix(tabella)
> spettri[z,]<-(tabella1)
> }

distanzePCA<-rep(0, nrow(campioni))
n<-c(1099:14698)
pcascores<-matrix(data=0, nrow=ncamp, ncol=ncol(principali4$x))
medie<-principali4$center
rms<-principali4$scale
noCS<-c(7901:8350,11900:12000)
noh2oacet<-c(1:101,107:159,161:200)


for (l in 1:ncamp) {
spettri[l,][spettri[l,]==0]<-0.00001
posto<-reganomer[1]+which.max(spettri[l,reganomer])-1
sposta<-centro-posto
if (sposta>0) spettru<-c(rep(0,sposta),spettri[l,1:(16384-sposta)])
if (sposta<0) spettru<-c(spettri[l,c((abs(sposta)+1):16384)],rep(0,abs(sposta)))
if (sposta==0) spettru<- spettri[l,]

#spettro<-spettru*coefmedio[l] #Using Regional Vertical Scaling normalization
spettru[noCS]<-0
spettro<-spettru*(100/sum(spettru))*up #Using Normalization to a constant sum
spettro<-spettro[n]

# integra i 200 intervalli formati ciascuno da 68 punti e crea nuovo vettore "interv"
interv<-rep(0,200)
i<-1
for (i in 1:200){
for (j in 1:68){
```

```
k<-((i-1)*68)+j
a<-spettro[k]
interv[i]<-interv[i]+a
}
}

#spettroZIP<-interv[noh2oacet]       #USE IF ALL BINS ARE SELECTED BUT THOSE IN "noh2oacet"
spettroZIP<-interv[sort(PCbestLD)]   #USE IF BINS ARE SELECTED THOROUGH LDA
#spettroZIP<-interv[sort(training)]  # USE IF BINS ARE SELECTED THOROUGH RANDOM PROCESS

pcascores[l,]<-((spettroZIP-medie)/rms) %*% principali4$rotation
distanzePCA[l]<-sqrt(sum((pcascores[l,y]-principali4$x[l,y])^2))
}
windows()

> colori<-c(as.character(campioni[,8]),as.character(campioni2[,8]))
> azienda<-c(campioni[,4],campioni2[,4])
> camp<-c(campioni[,2],campioni2[,2])
> pcascores<-rbind(principali4$x, pcascores)

plot(pcascores[,y],pch=" ",cex=1.2, main="194 BINS CS a mano") #, ylim=c(-6,7), xlim=c(-11,15))
text(pcascores[,y],labels=camp,col=colori, cex=.6)


                ###########
                #PART II #
                ###########

##########################################
# pH dependent bins removing according to pH-dependence study #
# previously performed (data file "eliminareX(X.X).txt" needs)       #
##########################################

eliminare1<-scan(file="eliminare2(0.025).txt", sep=";")
eliminare<-c(104:106,eliminare1,160)
#eliminare<-eliminare1
elimina<-sort(eliminare)

eliminapt<-rep(0, length(elimina)*68)
n<-0
for (g in 1:length(elimina)) {
j <- elimina[g]
n<-n+1
eliminapt[((68*n)-67):(n*68)]<-c(((j-1)*68+1):(j*68))
}

# FOR EVERY POINT
interbins<-interesse[,-eliminapt]

# FOR 200 BINS (INTERVALS)
#colnames(intervallo)<-c(1:200)
interbins<-intervallo[,-elimina] #elimina gli intervalli troppo alti in base ai loadings dei 3 pH#
dim(interbins)
```

```
#########################
# PRINCIPAL COMPONENT ANALYSIS  #
#########################

pc4<-prcomp(interbins,center=T,scale=T)
colori<-c(as.character(campioni[,8]))
azienda<-c(campioni[,4])
camp<-campioni[,2]
stagione<-c(as.character(campioni[,3]))
cultivar<-campioni[,6]
y<-c(2,5)
windows()
plot(pc4$x[,y],pch=" ",col=colori,cex=1.2, main="PC1_3 pH4; S=T, C=T")
#plot(pc4$x[,y],pch=azienda,col=colori,cex=1.2, main="PC1_2 pH4; S=T, C=T")
text(pc4$x[,y],labels=camp,col=colori, cex=.6) #COLORARE PER AZIENDA
#text(pc4$x[,y],labels=camp,col=stagione, cex=.6)  #COLORARE PER STAGIONE
#text(pc4$x[,y],labels=azienda,col=colori, cex=.6)

####################
# ANOVA and Tukey HSD tests  #
####################
xanova<-data.frame(pca$x[,2],campioni[,8])
#xanova[xanova=="orange"]<-"red"
colnames(xanova)<-c("values","ind")
oneway.test(values ~ ind, data=xanova, var.equal=T)
Anova<-aov(xanova[,1] ~ xanova[,2])
test<-TukeyHSD(Anova)
test
edit(test$xanova)


#################################
# Algorithms for interesting spectral regions study   #
#################################

pc<-2 #select here the PC space whose loadings want to be observed
spettropca<-filtro
ww<-1
for (i in 1:length(filtro)) {
if (filtro[i]==1) {
spettropca[i]<-pca$rotation[ww,pc]
ww<-ww+1
}
}

studiapc<-spettropca
studiapc[abs(studiapc)<0.05]<-0


sele<-c(9000:9500)  #Select here which data interval wants to be studied

esp<-10000 #Select here required vertical scaling
sens<-esp
 plot(riferiti[14,sele], type="l", ylim=c(-10,esp), col="red")
 lines(riferiti[58,sele], col="orange")
 lines(riferiti[60,sele], col="green")
```

```
 lines(riferiti[88,sele], col="brown")
 lines(studiapc2[sele]*sens, col="blue")

################################
# ANOVA and Tukey HSD test over "PCA" matrix #
################################

pca<-?????????? #indicate here which matrix has to be renamed "pca"

pc<-2
xanova<-data.frame(pca$x[,pc],campioni[,8])
#xanova[xanova=="orange"]<-"red"
colnames(xanova)<-c("values","ind")
whole<-oneway.test(values ~ ind, data=xanova, var.equal=T)
owt.R_G<-oneway.test(values ~ ind, data=xanova[c(1:48,50,59:81),], var.equal=T)
owt.R_O<-oneway.test(values ~ ind, data=xanova[c(1:58),], var.equal=T)
owt.R_B<-oneway.test(values ~ ind, data=xanova[c(1:48,50,82:93),], var.equal=T)
owt.G_B<-oneway.test(values ~ ind, data=xanova[c(59:93),], var.equal=T)
owt.G_O<-oneway.test(values ~ ind, data=xanova[c(49,51:81),], var.equal=T)
owt.O_B<-oneway.test(values ~ ind, data=xanova[c(49,51:58,82:93),], var.equal=T)
pcnames<-c("PacN", "PacS", "Sab", "Merc")
tabanova<-matrix(data=1, ncol=4, nrow=4)
dimnames(tabanova)<-list(pcnames,pcnames)
tabanova[c(1,2),c(1,2)]<-owt.R_O$p.value
tabanova[c(1,3),c(1,3)]<-owt.R_G$p.value
tabanova[c(1,4),c(1,4)]<-owt.R_B$p.value
tabanova[c(2,3),c(2,3)]<-owt.G_O$p.value
tabanova[c(2,4),c(2,4)]<-owt.O_B$p.value
tabanova[c(3,4),c(3,4)]<-owt.G_B$p.value
diag(tabanova)<-0
tabanova.pc2<-tabanova

pc<-5
xanova<-data.frame(pca$x[,pc],campioni[,8])
#xanova[xanova=="orange"]<-"red"
colnames(xanova)<-c("values","ind")
whole<-oneway.test(values ~ ind, data=xanova, var.equal=T)
owt.R_G<-oneway.test(values ~ ind, data=xanova[c(1:48,50,59:81),], var.equal=T)
owt.R_O<-oneway.test(values ~ ind, data=xanova[c(1:58),], var.equal=T)
owt.R_B<-oneway.test(values ~ ind, data=xanova[c(1:48,50,82:93),], var.equal=T)
owt.G_B<-oneway.test(values ~ ind, data=xanova[c(59:93),], var.equal=T)
owt.G_O<-oneway.test(values ~ ind, data=xanova[c(49,51:81),], var.equal=T)
owt.O_B<-oneway.test(values ~ ind, data=xanova[c(49,51:58,82:93),], var.equal=T)
pcnames<-c("PacN", "PacS", "Sab", "Merc")
tabanova<-matrix(data=1, ncol=4, nrow=4)
dimnames(tabanova)<-list(pcnames,pcnames)
tabanova[c(1,2),c(1,2)]<-owt.R_O$p.value
tabanova[c(1,3),c(1,3)]<-owt.R_G$p.value
tabanova[c(1,4),c(1,4)]<-owt.R_B$p.value
tabanova[c(2,3),c(2,3)]<-owt.G_O$p.value
tabanova[c(2,4),c(2,4)]<-owt.O_B$p.value
tabanova[c(3,4),c(3,4)]<-owt.G_B$p.value
diag(tabanova)<-0
tabanova.pc5<-tabanova

Anova<-aov(xanova[,1] ~ xanova[,2])
```

```
test<-TukeyHSD(Anova)
test
edit(test$xanova)


#########################
# EUCLIDEAN DISTANCES calculation  #
#########################
library(fields)
xscore<-pca$x[,y[1]]
yscore<-pca$x[,y[2]]
#xscore<- -xscore
#yscore<- -yscore
matrix<-cbind(xscore,yscore)
pacN<-matrix[c(1:48,50),]
pacS<-matrix[c(49,51:58),]
sab<-matrix[c(59:81),]
merc<-matrix[c(82:93),]
pacN.mean<-apply(pacN,2,mean)
pacS.mean<-apply(pacS,2,mean)
sab.mean<-apply(sab,2,mean)
merc.mean<-apply(merc,2,mean)

pacN.centre<-matrix(pacN.mean, nrow=1, ncol=2)
pacS.centre<-matrix(pacS.mean, nrow=1, ncol=2)
sab.centre<-matrix(sab.mean, nrow=1, ncol=2)
merc.centre<-matrix(merc.mean, nrow=1, ncol=2)
dist.matrix<-rbind(pacN.centre[1,], pacS.centre[1,], sab.centre[1,], merc.centre[1,])

varPacN<-sd(rdist(pacN, pacN.centre))
varPacS<-sd(rdist(pacS, pacS.centre))
varSab<-sd(rdist(sab, sab.centre))
varMerc<-sd(rdist(merc, merc.centre))

varPN_PS<-varPacN*varPacS
varPN_S<-varPacN*varSab
varPN_M<-varPacN*varMerc
varPS_S<-varPacS*varSab
varPS_M<-varPacS*varMerc
varS_M<-varSab*varMerc


distanze<-rdist(dist.matrix)
colnames(distanze)<-pcnames
rownames(distanze)<-pcnames
a<-edit(distanze)
distanze.var<-distanze
distanze.var[c(1,2),c(1,2)]<-distanze.var[1,2]/varPN_PS
distanze.var[c(1,3),c(1,3)]<-distanze.var[1,3]/varPN_S
distanze.var[c(1,4),c(1,4)]<-distanze.var[1,4]/varPN_M
distanze.var[c(2,3),c(2,3)]<-distanze.var[2,3]/varPS_S
distanze.var[c(2,4),c(2,4)]<-distanze.var[2,4]/varPS_M
distanze.var[c(3,4),c(3,4)]<-distanze.var[3,4]/varS_M
diag(distanze.var)<-0
```

```
        ############################
        # MAHALANOBIS DISTANCES CALCULATION  #
        ############################


pacN.var<-var(pacN)
pacN.mean<-apply(pacN,2,mean)
pacN.mah<-mahalanobis(pacN,pacN.mean,pacN.var)
pacS.var<-var(pacS)
pacS.mean<-apply(pacS,2,mean)
pacS.mah<-mahalanobis(pacS,pacS.mean,pacS.var)
sab.var<-var(sab)
sab.mean<-apply(sab,2,mean)
sab.mah<-mahalanobis(sab,sab.mean,sab.var)
merc.var<-var(merc)
merc.mean<-apply(merc,2,mean)
merc.mah<-mahalanobis(merc,merc.mean,merc.var)

        ############################
        # MAHALANOBIS DISTANCES PLOTTING  #
        ############################


library(car)

#CAMPIONI DI PACHINO NAOMI
ellipse(pacN.mean, pacN.var, radius=1, col="red", lwd=1, add=TRUE)
ellipse(pacN.mean, pacN.var, radius=2, col="red", lwd=1, add=TRUE)
#ellipse(pacN.mean, pacN.var, radius=3, col="red", lwd=1, add=TRUE)

#CAMPIONI DI PACHINO SHIREN
ellipse(pacS.mean, pacS.var, radius=1, col="orange", lwd=1, add=TRUE)
ellipse(pacS.mean, pacS.var, radius=2, col="orange", lwd=1, add=TRUE)
#ellipse(pacS.mean, pacS.var, radius=3, col="orange", lwd=1, add=TRUE)

#CAMPIONI DI SABAUDIA
ellipse(sab.mean, sab.var, radius=1, col="green", lwd=1, add=TRUE)
ellipse(sab.mean, sab.var, radius=2, col="green", lwd=1, add=TRUE)
#ellipse(sab.mean, sab.var, radius=3, col="green", lwd=1, add=TRUE)

#CAMPIONI DEL MERCATO
ellipse(merc.mean, merc.var, radius=1, col="brown", lwd=1, add=TRUE)
ellipse(merc.mean, merc.var, radius=2, col="brown", lwd=1, add=TRUE)
#ellipse(merc.mean, merc.var, radius=3, col="brown", lwd=1, add=TRUE)
```

```
                    ####################################
                    # ALGORITHMS FOR LINEAR DISCRIMINANT ANALYSIS  #
                    ####################################

###################################
#Creates dataset useful for successive LDA and plotting#
###################################
LDAest<-campioni[,c(7,8)]

> LDAest2<-campioni2[,c(7,8)]       #Use for unknown samples inserting
> LDAest<-rbind(LDAest, LDAest2)    # Use for unknown samples inserting

#LDAest<-read.table("datiLDA67est.csv", sep=";", dec=".", header=T)

Principali4 <- ?????????? #Renames interested data matrix in "principali4"

ncamp<-nrow(campioni)
importance<-principali4$sdev^2/sum(principali4$sdev^2)
elenco<-rep(0, ncamp)
soglia<-100               # set % of variance explained by PC variables are going to be chosen
summa<-0
for (s in 1:ncamp){
summa<-summa+importance[s]
elenco[s]<-s
if (summa >= soglia/100) break
}
PCgood<-length(elenco[elenco!=0])
PCgood


pcgood<-principali4$x[,1:PCgood]
#pcgood<-principali4$x

binslda<-cbind(pcgood,LDAest) #to use it if you want to keep the first PCs with cumulative sdev >
soglia
#binslda<-cbind(perpca,LDAest) #to use if there are not bins removed for pH-dependence
#binslda<-cbind(interbins,LDAest)        # to use if there are bins removed for pH-dependence

#binslda<-binslda1[,2:ncol(binslda1)]

#write.table(binslda,"bins67est.csv", sep=";", dec=".")



##################
# LDA ANALYSIS 400 MHz #
##################

#Matrices for LDA definition#
col<-ncol(binslda)
LDAtable1<- binslda[,1:(col-1)]

#############################################################
#Use to perform a system training removing 2/3 of the samples randomly from dataset and creating#
# a new dataset to be submitted to LDA analysis "LDAtable"                                    #
#############################################################
```

```
samples<-nrow(LDAtable1)
sampling<-as.integer(samples*2/3)
sampling<-82 #If the number of trained samples want to be chosen a priori
train1 <- sample(1:samples, sampling)
train <- sort(train1)
table(LDAtable1$Sp[train]) #Shows classes of samples selected for training
table(LDAtable1$Sp[-train]) # Shows classes of samples excluded from training
tranames<-LDAtable1$Sp[-train]
#sampled<-c(1:samples)[-train]
sampled<-c(campioni[,2])[-train]
camp<-c(campioni[,2])[train]
names(sampled)<-tranames
#sampled

LDAtest  <- LDAtable1[-train,] #matrix containing samples excluded from training
LDAtable <- LDAtable1[train,] #matrix containing samples selected for training
colorLDA <- binslda[train,]

#USE IT ONLY IF YOU WANT TO USE ALL THE SAMPLE FOR LDA
LDAtable<-LDAtable1
colorLDA <- binslda

###################################
# LDA ANALYSIS and results storing in "zlda" object #
###################################
lda" CONTENENTE I RISULTATI
gruppi<-length(names(table(LDAest$Sp)))
#zlda <- lda(Sp ~ ., LDAtable, prior = c(1,1,1,1,1)/5)
zlda <- lda(Sp ~ ., LDAtable, prior = rep(1,gruppi)/gruppi)
#zlda <- lda(Sp ~ ., LDAtable, prior = c(1,1,1,1)/4  ,subset = train)

#zlda <- lda(Sp ~ ., LDAtable, prior = c(1,1,1,1)/4  ,CV = TRUE)
#write.table(zlda$posterior,"leaveoneout_33bin.txt", sep=";", dec=".")


####################
# MANUAL LEAVE-ONE-OUT #
####################
samples<-nrow(LDAtable1)
prediction<- rep(0,samples)
for (l in 1:samples) {
z2lda <-lda(Sp ~ ., LDAtable[-l,], prior = c(1,1,1,1)/4)
prediction[l]<-predict(z2lda, LDAtable[l,])$class
}
# SUCCESS ESTIMATION OF LEAVE-ONE-OUT STEP #
comparing<-rbind(prediction,LDAtable1$Sp)
 different<-apply(comparing,2, diff)
 good<-different[different==0]
 success<-length(good)/length(different)*100
 success



############################################################
```

```
############################
# SUCCESS ESTIMATION OF THE TRAINING  #
############################

 predetti<-predict(zlda, LDAtest)$class
 compara<-rbind(predetti,tranames)
 diversi<-apply(compara,2, diff)
 buoni<-diversi[diversi==0]
 successo<-length(buoni)/length(diversi)*100
 successo


######################
# UNKNOWN SAMPLES INSERTION #
######################

samplex<-scan("samplex.spt", sep=";", dec=".")
samplex[samplex==0]<-0.00001
# devo fare in modo che il max del picco dell'anomero beta del glucosio sia in un determinato punto
#samplex<-samplex*coefmedio[32] #dovrei tenere conto della normalizzazione (questo era per il
campione 66.spt)
samplex_fil<-samplex*filtro
samplexzip<-samplex_fil[samplex_fil!=0]

medie<-principali4$center
rms<-principali4$scale
pcscoresx<-matrix(data=0, nrow=1, ncol=ncamp)
for (n in 1:ncamp) {
pcscoresx[,n]<-sum(((samplexzip-medie)/rms)*principali4$rotation[,n])
}

# CONTROL
pcscores<-matrix(data=0, nrow=1, ncol=ncamp)
medie<-principali4$center
rms<-principali4$scale
for (n in 1:ncamp) {
pcscores[,n]<-sum(((bins[32,200:291]-medie)/rms)*principali4$rotation[,n])
}


#################
# PLOTTING OF THE DATA #
#################

 pachino<-LDAtable      #Use for showing only samples selected by training
 campionitot<-campioni[train,]

#pachino<-LDAtable1   #Use for showing also samples excluded from training
 colorLDA <- binslda      #Use only with the above command line
 camp<-campioni[,2]    # Use only with the above command line
 campionitot<-campioni # Use only with the above command line

 pachino3<-pachino[,1:PCgood]
 pachino4<-colorLDA
 ncamp<-nrow(pachino)
```

```
> interlab<-
cbind((pcascores[(ncamp+1):nrow(pcascores),1:PCgood]),campioni2[,((ncol(campioni2))-
1):ncol(campioni)])
> colnames(interlab)<-colnames(binslda)
> pachino2<-pachino[,1:PCgood]
> interlab2<-interlab[,1:PCgood]
> pachino3<- rbind(pachino2, interlab2)
> pachino4<-rbind(binslda,interlab)
> campionitot<-rbind(campioni,campioni2)
> camp<-campionitot[,2]


campLDA<-nrow(pachino4)
LDA1_scores<-rep(0,campLDA)
LDA2_scores<-rep(0,campLDA)
LDA3_scores<-rep(0,campLDA)
LDA4_scores<-rep(0,campLDA)
cvlda<-zlda$scaling

# categ<-as.character(pachino[,ncol(pachino)])

#gruppi<-5
for (n in 1:campLDA) {
LDA1_scores[n]<-sum(pachino3[n,]*cvlda[,1])
LDA2_scores[n]<-sum(pachino3[n,]*cvlda[,2])
if (gruppi > 3)
LDA3_scores[n]<-sum(pachino3[n,]*cvlda[,3])
if (gruppi > 4)
LDA4_scores[n]<-sum(pachino3[n,]*cvlda[,4])          #1:(ncol(pachino)-1)
}

colori<- as.character(pachino4[,col])

windows()
#plot(LDA1_scores,LDA2_scores,pch=categ, col=colori) #ylim=c(-30,30), xlim=c(-30,30)
plot(LDA1_scores,LDA2_scores,pch=" ", col=colori) #ylim=c(-30,30), xlim=c(-30,30)
text(LDA1_scores,LDA2_scores, labels=camp, col=colori, cex=.6)

>windows()
>plot(LDA1_scores,LDA3_scores,pch=" ", col=colori) #ylim=c(-30,30), xlim=c(-30,30)
>text(LDA1_scores,LDA3_scores, labels=camp, col=colori, cex=.6)
>windows()
>plot(LDA2_scores,LDA3_scores,pch=" ", col=colori) #ylim=c(-30,30), xlim=c(-30,30)
>text(LDA2_scores,LDA3_scores, labels=camp, col=colori, cex=.6)

windows()
barplot(zlda$scaling[,1])
> windows()
> barplot(zlda$scaling[,2])
> windows()
> barplot(zlda$scaling[,3])
> sumCV<-abs(zlda$scaling[,1])+abs(zlda$scaling[,2])+abs(zlda$scaling[,3])
> windows()
> barplot(sumCV) #per valutare le PC più importanti consideranto tutte le LD
```

```
####################
# THREE DIMENSIONAL PLOT #
####################
 x<-LDA1_scores
 y<-LDA2_scores
 z<-LDA3_scores
library(scatterplot3d)
windows()
 for (angle in 1:360) {
 scatterplot3d(x, y, z, color=colori, angle=angle)
 }
scatterplot3d(x, y, z, color=colori)


##############################################
# SELECTION OF THE MOST IMPORTANT BINS OF EACH LD SPACE AND#
# PLOTTING OF THE 194 BINS ACCORDING TO THEIR IMPORTANCE     #
##############################################

ldn<-194
PC<-5
percent<-0.70
matLD1<-t(principali4$rotation[,1:26])*zlda$scaling[,1]
summatLD1<-apply(matLD1,2, sum)
names(summatLD1)<-colnames(perpca)
windows()
barplot(summatLD1)
sortLD1<-sort(abs(summatLD1), decreasing=TRUE)
#bestLD1<-sortLD1[1:PC]
 bestLD1<-sortLD1[sortLD1>=max(sortLD1*percent)]
PCbestLD1<-as.numeric(names(bestLD1))
PCbestLD1

matLD2<-t(principali4$rotation[,1:26])*zlda$scaling[,2]
summatLD2<-apply(matLD2,2, sum)
names(summatLD2)<-colnames(perpca)
windows()
barplot(summatLD2)
sortLD2<-sort(abs(summatLD2), decreasing=TRUE)
#bestLD2<-sortLD2[1:PC]
 bestLD2<-sortLD2[sortLD2>=max(sortLD2*percent)]
PCbestLD2<-as.numeric(names(bestLD2))
PCbestLD2

matLD3<-t(principali4$rotation[,1:26])*zlda$scaling[,3]
summatLD3<-apply(matLD3,2, sum)
names(summatLD3)<-colnames(perpca)
windows()
barplot(summatLD3)
sortLD3<-sort(abs(summatLD3), decreasing=TRUE)
#bestLD3<-sortLD3[1:PC]
 bestLD3<-sortLD3[sortLD3>=max(sortLD3*percent)]
PCbestLD3<-as.numeric(names(bestLD3))
PCbestLD3

PCbestLD<-sort(union(union(PCbestLD1, PCbestLD2),PCbestLD3), decreasing=TRUE)
```

```
length(PCbestLD)

for (s in PCbestLD) {
titolo<-paste("Bins n°",s)
windows(13,3)
meno<-min(intervallo[,s])
barplot(intervallo[,s]-meno, col=colori, main=titolo)
}

#######################################
# SCRIPT FOR PLOTTING SELECTED REGIONS OF SPECTRA #
#######################################
selez<-c(9,38,62,78)
selez.CS<-allineati.CS[selez,]
for (w in (sort(PCbestLD))) {
detto<-paste("Bins n°",w)
selezbin<-selez.CS[,c((1098+(68*w-67)):(1098+(w*68)))]
alto<-max(selezbin)+0.1*max(selezbin)
windows()
plot(selezbin[1,], type="l", col="red", main=detto, ylim=c(-0.1,alto))
lines(selezbin[2,], col="orange")
lines(selezbin[3,], col="darkgreen")
lines(selezbin[4,], col="green")
}

        ##############################
        # MAHALANOBIS DISTANCES CALCULATION  #
        ##############################

ldaest<-cbind(LDA1_scores,LDA2_scores)

>ldaest<-cbind(LDA1_scores,LDA3_scores)

>ldaest<-cbind(LDA2_scores,LDA3_scores)

##########
# ESTIVI #
##########

ncamp<-nrow(campionitot)
classi<-campionitot$Sp
names(classi)<-c(1:ncamp)
#classe<-table(pachino$Sp)
classe<-table(classi)
nomi<-names(classe)
numclas<-length(nomi)
for (g in 1:numclas) {
nam <- paste(nomi[g],"LDA", sep=".")
assign(nam, as.numeric(names(classi[classi==nomi[g]])))
}

ldaestpacN<-ldaest[pacN.LDA,]
ldaestpacS<-ldaest[pacS.LDA,]
ldaestsab<-ldaest[sab.LDA,]
ldaestmerc<-ldaest[merc.LDA,]
ldainvpacR<-ldainv[pacR.LDA,]
```

```
ldaestpacN.var<-var(ldaestpacN)
ldaestpacN.mean<-apply(ldaestpacN,2,mean)
ldaestpacN.mah<-mahalanobis(ldaestpacN,ldaestpacN.mean,ldaestpacN.var)
ldaestpacS.var<-var(ldaestpacS)
ldaestpacS.mean<-apply(ldaestpacS,2,mean)
ldaestpacS.mah<-mahalanobis(ldaestpacS,ldaestpacS.mean,ldaestpacS.var)
ldaestsab.var<-var(ldaestsab)
ldaestsab.mean<-apply(ldaestsab,2,mean)
ldaestsab.mah<-mahalanobis(ldaestsab,ldaestsab.mean,ldaestsab.var)
ldaestmerc.var<-var(ldaestmerc)
ldaestmerc.mean<-apply(ldaestmerc,2,mean)
ldaestmerc.mah<-mahalanobis(ldaestmerc,ldaestmerc.mean,ldaestmerc.var)
ldainvpacR.var<-var(ldainvpacR)
ldainvpacR.mean<-apply(ldainvpacR,2,mean)
ldainvpacR.mah<-mahalanobis(ldainvpacR,ldainvpacR.mean,ldainvpacR.var)

        ############################
        # MAHALANOBIS DISTANCES PLOTTING  #
        ############################
library(car)

#SUMMER SAMPLES OF PACHINO NAOMI
ellipse(ldaestpacN.mean, ldaestpacN.var , radius=1,col="red", lwd=1, add=TRUE)
ellipse(ldaestpacN.mean, ldaestpacN.var , radius=2,col="red",lwd=1, add=TRUE)
ellipse(ldaestpacN.mean, ldaestpacN.var , radius=3,col="red",lwd=1, add=TRUE)

# SUMMER SAMPLES OF PACHINO SHIREN
ellipse(ldaestpacS.mean, ldaestpacS.var , radius=1,col="orange", lwd=1, add=TRUE)
ellipse(ldaestpacS.mean, ldaestpacS.var , radius=2,col="orange",lwd=1, add=TRUE)
ellipse(ldaestpacS.mean, ldaestpacS.var , radius=3,col="orange",lwd=1, add=TRUE)

# SUMMER SAMPLES OF SABAUDIA
ellipse(ldaestsab.mean, ldaestsab.var , radius=1, col="green",lwd=1, add=TRUE)
ellipse(ldaestsab.mean, ldaestsab.var , radius=2, col="green",lwd=1, add=TRUE)
ellipse(ldaestsab.mean, ldaestsab.var , radius=3, col="green",lwd=1, add=TRUE)

# SUMMER SAMPLES OF MERCATO
ellipse(ldaestmerc.mean, ldaestmerc.var , radius=1, col="brown",lwd=1, add=TRUE)
ellipse(ldaestmerc.mean, ldaestmerc.var , radius=2, col="brown",lwd=1, add=TRUE)
ellipse(ldaestmerc.mean, ldaestmerc.var , radius=3, col="brown",lwd=1, add=TRUE)

##############
# WINTER SAMPLES #
##############
ldainv<-cbind(LDA1_scores,LDA2_scores)
ncamp<-nrow(campionitot)
classi<-campionitot$Sp
names(classi)<-c(1:ncamp)
#classe<-table(pachino$Sp)
classe<-table(classi)
nomi<-names(classe)
numclas<-length(nomi)
for (g in 1:numclas) {
nam <- paste(nomi[g],"LDA", sep=".")
assign(nam, as.numeric(names(classi[classi==nomi[g]])))
}
```

```
ldainvpacN<-ldainv[pacN.LDA,]
ldainvpacS<-ldainv[pacS.LDA,]
ldainvLic<-ldainv[Lic.LDA,]
ldainvmerc<-ldainv[merc.LDA,]
ldainvpacR<-ldainv[pacR.LDA,]


ldainvpacN.var<-var(ldainvpacN)
ldainvpacN.mean<-apply(ldainvpacN,2,mean)
ldainvpacN.mah<-mahalanobis(ldainvpacN,ldainvpacN.mean,ldainvpacN.var)
ldainvpacS.var<-var(ldainvpacS)
ldainvpacS.mean<-apply(ldainvpacS,2,mean)
ldainvpacS.mah<-mahalanobis(ldainvpacS,ldainvpacS.mean,ldainvpacS.var)
ldainvLic.var<-var(ldainvLic)
ldainvLic.mean<-apply(ldainvLic,2,mean)
ldainvLic.mah<-mahalanobis(ldainvLic,ldainvLic.mean,ldainvLic.var)
ldainvmerc.var<-var(ldainvmerc)
ldainvmerc.mean<-apply(ldainvmerc,2,mean)
ldainvmerc.mah<-mahalanobis(ldainvmerc,ldainvmerc.mean,ldainvmerc.var)
ldainvpacR.var<-var(ldainvpacR)
ldainvpacR.mean<-apply(ldainvpacR,2,mean)
ldainvpacR.mah<-mahalanobis(ldainvpacR,ldainvpacR.mean,ldainvpacR.var)

        ############################
        # MAHALANOBIS DISTANCES PLOTTING  #
        ############################

library(car)

# WINTER SAMPLES OF PACHINO NAOMI
ellipse(ldainvpacN.mean, ldainvpacN.var , radius=1,col="blue", lwd=1, add=TRUE)
ellipse(ldainvpacN.mean, ldainvpacN.var , radius=2,col="blue",lwd=1, add=TRUE)
ellipse(ldainvpacN.mean, ldainvpacN.var , radius=3,col="blue",lwd=1, add=TRUE)

# WINTER SAMPLES OF PACHINO SHIREN
ellipse(ldainvpacS.mean, ldainvpacS.var , radius=1,col="cyan", lwd=1, add=TRUE)
ellipse(ldainvpacS.mean, ldainvpacS.var , radius=2,col="cyan",lwd=1, add=TRUE)
ellipse(ldainvpacS.mean, ldainvpacS.var , radius=3,col="cyan",lwd=1, add=TRUE)

# WINTER SAMPLES OF LICATA
ellipse(ldainvLic.mean, ldainvLic.var , radius=1, col="magenta",lwd=1, add=TRUE)
ellipse(ldainvLic.mean, ldainvLic.var , radius=2, col="magenta",lwd=1, add=TRUE)
ellipse(ldainvLic.mean, ldainvLic.var , radius=3, col="magenta",lwd=1, add=TRUE)

# WINTER SAMPLES OF MERCATO
ellipse(ldainvmerc.mean, ldainvmerc.var , radius=1, col="brown",lwd=1, add=TRUE)
ellipse(ldainvmerc.mean, ldainvmerc.var , radius=2, col="brown",lwd=1, add=TRUE)
ellipse(ldainvmerc.mean, ldainvmerc.var , radius=3, col="brown",lwd=1, add=TRUE)

# WINTER SAMPLES OF PACHINO RANDOM
ellipse(ldainvpacR.mean, ldainvpacR.var , radius=1,col="orange", lwd=1, add=TRUE)
ellipse(ldainvpacR.mean, ldainvpacR.var , radius=2,col="orange",lwd=1, add=TRUE)
ellipse(ldainvpacR.mean, ldainvpacR.var , radius=3,col="orange",lwd=1, add=TRUE)
```

400 MHz temperature and pH dependence study

```
#INTRODUCTION:
#THIS SCRIP READS THE 3 SPECTRA ACQUIRED AT 25 °C AT 400 MHz ACCORDING TO
#pH-DEPENDENCE STUDY AND PERFORMS THE FOLLOWING STEPS:
#1) SCALING OVER RELATIVE ACETATE SIGNAL MAX
#2) DATA REDUCING IN 200 BINS (OF A REGION COVERING 10 PPM CONSTITUTED BY 13600 pt.)
#3) RESIDUAL WATER AND ACETATE SIGNALS REMOVING
#4) PCA ANALYSIS (SCALE=F, CENTER=F)
#5) DETERMINATION OF PC2 LOADINGS HAVING VALUE HIGHER THAN A TRESHOLD (0.025)
#6) DETERMINATION AND CREATION OF A FILE LISTING THE BINS TO BE REMOVED IN THE PCA
#ANALYSIS OVER ALL SAMPLES


# read spectra filenames from 'nomefile'.txt

rm(list = ls())
library(MASS)
campioni<-read.table("test25.txt", sep=";", dec=".", header=T)
#colClasses = "character"
names<-as.character(campioni[,1])
nomi<-as.matrix(names)
punti<-16384
#num<-read.table("numcamp.txt", sep="",col.names=F)

# read every spectrum and arrange them in a matrix of ncamp rows and 13600 (n°points in 10 ppm)
#columns

ncamp<-nrow(nomi)
spettri<-matrix(data=0, nrow=ncamp, ncol=punti)
refacetato<-matrix(data=0, nrow=ncamp, ncol=punti)
for (z in 1:ncamp) {
tabella<-read.table(nomi[z,1], sep=";", dec=".", header=F)
# defines intervals and values for data normalization over acetate signal
tabella1<-as.matrix(tabella)
intace<-c(11911:12000)
acetato<-max(tabella1[intace])
spettri[z,]<-(tabella1)
# scales in "refacetato" using acetate max
refacetato[z,]<-(tabella1/acetato)
}

####################
#PLOT FOR SIGNALS CONTROL #
####################

plot(refacetato[1,c(8360:8500)], type="l",)
plot(refacetato[2,c(8360:8500)], type="l",)
plot(refacetato[3,c(8360:8500)], type="l",)

refacetato16384<-refacetato
 plot(refacetato16384[3,],xlim=c(11920,12000),ylim=c(0,10), type="l")
 lines(refacetato16384[1,], col="blue")
 lines(refacetato16384[2,], col="red")
```

```
####################################
# VARIABLES ALIGNMENT OVER beta D GLUCOSE SIGNAL #
####################################
centro<-8424 #(rispetto al campione + rappresentativo che è il 153 (NC))
maxvet<-matrix(data=0, ncol=1, nrow=ncamp)
reganomer<-c(8360:8500)
for (i in 1:ncamp){
for (j in reganomer){
if (refacetato[i,j]==max(refacetato[i,reganomer])) maxvet[i,]<- centro-j
#if (allineati[i,j]==max(allineati[i,reganomer])) maxvet[i,]<- centro-j
}
}
allineati<- matrix(data=0, ncol=punti, nrow=ncamp)
for (i in 1:ncamp) {
shift<-maxvet[i,1]
if (shift>0) allineati[i,(1+shift):punti]<-refacetato[i,1:(punti-shift)]
if (shift==0) allineati[i,]<-refacetato[i,]
if (shift<0) allineati[i,1:(punti+shift)]<-refacetato[i,(1-shift):punti]
}

#########################
# PRODUCES 200 BINS DATA MATRIX #
#########################

n<-c(1099:14698)
interesse<-allineati[,n] #USE WHEN THE SPECTRUM HAVE BEEN ALIGNED
#interesse<-refacetato[,n] # USE WHEN THE SPECTRUM HAVE NOT BEEN ALIGNED

intervallo<-matrix(data=0, nrow=ncamp, ncol=200)
for (z in 1:ncamp) {
i<-1
for (i in 1:200){
for (j in 1:68){
k<-((i-1)*68)+j
a<-interesse[z,k]
intervallo[z,i]<-intervallo[z,i]+a
#if (a<0) intervallo[z,i]<-intervallo[z,i]+a #USE IF A TRESHOLD FOR NOISE REMOVING NEEDS
}
}
}
colnames(intervallo)<-c(1:200)


#ALTERNATIVE 1: ONLY BINS CONTAINING RESIDUAL WATER SIGNAL ARE REMOVED
noh2o<-c(1:103,107:200)
perpca<- intervallo[,noh2o]

#ALTERNATIVE 2: ARE REMOVED BINS CONTAINING BOTH WATER AND ACETATE SIGNALS
noh2oacet<-c(1:103,107:159,162:200)
perpca<- intervallo[,noh2oacet]

#ALTERNATIVE 3: ALL THE BINS ARE KEPT
perpca<-intervallo

####################
# Principal Component Analysis  #
```

```
####################

principali4<-prcomp(perpca,center=F,scale=F)
#colors<-read.table("colori.txt",row.names=1, sep=";", dec=".", header=T)
colori<-c(as.character(campioni[,3]))
azienda<-c(campioni[,4])
camp<-campioni[,2]
y<-c(1,2)
plot(principali4$x[,y],pch=" ",col=colori,cex=1.2, main="PC1_2 pH4; S=T, C=T")
#plot(principali4$x[,y],pch=azienda,col=colori,cex=1.2, main="PC1_2 pH4; S=T, C=T")
text(principali4$x[,y],labels=camp,col=colori, cex=.6)
#text(principali4$x[,y],labels=azienda,col=colori, cex=.6)

#windows()
 PC1load4<-principali4$rotation[,1]
 barplot(-PC1load4)

PC2load4<-principali4$rotation[,2]
numeri<-colnames(perpca)
names(PC2load4)<-numeri
windows()  #IN ANOTHER DEVICE Loadings2 ARE SHOWN
barplot(PC2load4)
windows()  # IN ANOTHER DEVICE ARE SHOWN ZOOMED bins 160 e 161
barplot(PC2load4[c(147:167)], cex.names=.5)

#REMOVES BINS HAVING LOADINGS 2 VALUES HIGHER THAN A TRESHOLD (=0.025 in this case)
positivi<-abs(PC2load4)
soglia<-0.025
ciccio<-positivi[positivi>=soglia]
noshift<-positivi[positivi<=soglia]
eliminare<-names(ciccio)
eliminare1<-as.numeric(eliminare)

#WRITES A FILE CONTANING ALL THE BINS TO BE REMOVED FROM DATASET
write(eliminare1, file = "eliminare2b(0.025).txt",append = FALSE, sep = ";")

#remove from Loadings2 barplot the rejected intervals
vettore<- rep(1,200)
vettore[eliminare1]<-0
load2<-PC2load4*vettore
windows()
names(load2)<-names(PC2load4)
barplot(load2)
```

Appendix B: Cherry tomato samples catalogue

The following table represents the whole sampling catalogue of cherry tomatoes. For each sample the main information are reported. "200 MHz" and "400 MHz" columns illustrate which samples have been analyzed with these techniques: a missing value means that such a sample has not been analyzed with this technique.

Samples with anomalous FIDs recorded at 400 MHz that have been discarded during chemometric analysis are highlighted in yellow. The two FIDs of the samples highlighted in red, recorded at 200 MHz, have been discarded because of problems occurred during NMR acquisition. External samples, used for inter-laboratory check test, are highlighted in blue.

Table's rows are colored in different colors for better identification of samples geographical origin: Black rows for samples of **Pachino**, Magenta rows for samples of **Licata**, Green rows for samples of **Sabaudia** and Brown rows for samples coming from **markets**, according with the colors used for chemometric data plotting.

ID	Sample code	200MHz	400MHz	Year	Season	Date	group	Farm of Pachino	Salinity	Origin	Cultivar
1	1Az.1A1°a	1	1	2003	win. 03	10/03/03	1	1	5400	Pachino	Naomi
2	1Az.1A1°b	2	2	2003	win. 03	10/03/03	1	1	5400	Pachino	Naomi
3	1Az.1A1°c	3	3	2003	win. 03	10/03/03	1	1	5400	Pachino	Naomi
4	1Az.1B1°a	4	4	2003	win. 03	10/03/03	1	1	5400	Pachino	Naomi
5	1Az.1B1°b	5	5	2003	win. 03	10/03/03	1	1	5400	Pachino	Naomi
6	1Az.1B1°c	6	6	2003	win. 03	10/03/03	1	1	5400	Pachino	Naomi
7	1Az.1C1°a	7	7	2003	win. 03	10/03/03	1	1	5400	Pachino	Naomi
8	1Az.1C1°b	8	8	2003	win. 03	10/03/03	1	1	5400	Pachino	Naomi
9	1Az.2A1°a	9	9	2003	win. 03	10/03/03	1	2	1500	Pachino	Naomi
10	1Az.2A1°b	10	10	2003	win. 03	10/03/03	1	2	1500	Pachino	Naomi
11	1Az.2B1°a	11	11	2003	win. 03	10/03/03	1	2	1500	Pachino	Naomi
12	1Az.2B1°b	12	12	2003	win. 03	10/03/03	1	2	1500	Pachino	Naomi
13	1Az.2C1°a	13	13	2003	win. 03	10/03/03	1	2	1500	Pachino	Naomi
14	1Az.2C1°b	14	14	2003	win. 03	10/03/03	1	2	1500	Pachino	Naomi
22	2Az.1C2°b	22	22	2003	win. 03	14/03/03	2	1	5400	Pachino	Naomi
23	2Az.1C2°c	23	23	2003	win. 03	14/03/03	2	1	5400	Pachino	Naomi
24	2Az.2A2°a	24	24	2003	win. 03	14/03/03	2	2	1500	Pachino	Naomi
25	2Az.2A2°b	25	25	2003	win. 03	14/03/03	2	2	1500	Pachino	Naomi
26	2Az.2B2°a	26	26	2003	win. 03	14/03/03	2	2	1500	Pachino	Naomi
27	2Az.2B2°b	27	27	2003	win. 03	14/03/03	2	2	1500	Pachino	Naomi
28	2Az.2C2°a	28	28	2003	win. 03	14/03/03	2	2	1500	Pachino	Naomi
29	2Az.2C2°b	29	29	2003	win. 03	14/03/03	2	2	1500	Pachino	Naomi
30	2Az.2C2°c	30	30	2003	win. 03	14/03/03	2	2	1500	Pachino	Naomi
31	Random a	31	31	2003	win. 03	14/03/03	2	Random		Pachino	Naomi
32	Random b	32	32	2003	win. 03	14/03/03	2	Random		Pachino	Naomi
33	Random c	33	33	2003	win. 03	14/03/03	2	Random		Pachino	Naomi

ID	Sample code	200MHz	400MHz	Year	Season	Date	group	Farm of Pachino	Salinity	Origin	Cultivar
34	Random d	34	34	2003	win. 03	14/03/03	2	Random		Pachino	Naomi
35	3Az.3A1°a	35	35	2003	sum. 03	23/06/03	3	3	2000	Pachino	Naomi
36	3Az.3B1°a	36	36	2003	sum. 03	23/06/03	3	3	2000	Pachino	Naomi
37	3Az.3C1°a	37	37	2003	sum. 03	23/06/03	3	3	2000	Pachino	Naomi
38	3Az.4A1°a	38	38	2003	sum. 03	23/06/03	3	4	5500	Pachino	Naomi
39	3Az.4A1°b	39	39	2003	sum. 03	23/06/03	3	4	5500	Pachino	Naomi
40	3Az.4B1°a	40	40	2003	sum. 03	23/06/03	3	4	5500	Pachino	Naomi
41	3Az.4C1°a	41	41	2003	sum. 03	23/06/03	3	4	5500	Pachino	Naomi
42	3Az.4C1°b	42	42	2003	sum. 03	23/06/03	3	4	5500	Pachino	Naomi
43	4Az.3A2a	43	43	2003	sum. 03	01/07/03	4	3	2000	Pachino	Naomi
44	4Az.3A3a	44	44	2003	sum. 03	01/07/03	4	3	2000	Pachino	Naomi
45	4Az.3A3b	45	45	2003	sum. 03	01/07/03	4	3	2000	Pachino	Naomi
46	4Az.3B2a	46	46	2003	sum. 03	01/07/03	4	3	2000	Pachino	Naomi
47	4Az.3B3a	47	47	2003	sum. 03	01/07/03	4	3	2000	Pachino	Naomi
48	4Az.3C2a	48	48	2003	sum. 03	01/07/03	4	3	2000	Pachino	Naomi
49	4Az.3C3a	49	49	2003	sum. 03	01/07/03	4	3	2000	Pachino	Naomi
50	4Az.4A2a	50	50	2003	sum. 03	01/07/03	4	4	5500	Pachino	Naomi
51	4Az.4A3a	51	51	2003	sum. 03	01/07/03	4	4	5500	Pachino	Naomi
52	4Az.4B2a	52	52	2003	sum. 03	01/07/03	4	4	5500	Pachino	Naomi
53	4Az.4B3a	53	53	2003	sum. 03	01/07/03	4	4	5500	Pachino	Naomi
54	4Az.4C2a	54	54	2003	sum. 03	01/07/03	4	4	5500	Pachino	Naomi
55	4Az.4C3a	55	55	2003	sum. 03	01/07/03	4	4	5500	Pachino	Naomi
56	5Az.3A4a	56	56	2003	sum. 03	09/07/03	5	3	2000	Pachino	Naomi
57	5Az.3A4b	57	57	2003	sum. 03	09/07/03	5	3	2000	Pachino	Naomi
58	5Az.3A5a	58	58	2003	sum. 03	09/07/03	5	3	2000	Pachino	Naomi
59	5Az.3A5b	59	59	2003	sum. 03	09/07/03	5	3	2000	Pachino	Naomi
60	5Az.3A6a	60	60	2003	sum. 03	09/07/03	5	3	2000	Pachino	Naomi
61	5Az.3B4a	61	61	2003	sum. 03	09/07/03	5	3	2000	Pachino	Naomi
62	5Az.3B4b	62	62	2003	sum. 03	09/07/03	5	3	2000	Pachino	Naomi
63	5Az.3B5a	63	63	2003	sum. 03	09/07/03	5	3	2000	Pachino	Naomi
64	5Az.3B6a	64	64	2003	sum. 03	09/07/03	5	3	2000	Pachino	Naomi
65	5Az.3C4a	65	65	2003	sum. 03	09/07/03	5	3	2000	Pachino	Naomi
66	5Az.3C4b	66	66	2003	sum. 03	09/07/03	5	3	2000	Pachino	Naomi
67	5Az.3C5a	67	67	2003	sum. 03	09/07/03	5	3	2000	Pachino	Naomi
68	5Az.3C6a	68	68	2003	sum. 03	09/07/03	5	3	2000	Pachino	Naomi
69	5Az.3C6b	69	69	2003	sum. 03	09/07/03	5	3	2000	Pachino	Naomi
70	5Az.4A4a	70	70	2003	sum. 03	09/07/03	5	4	5500	Pachino	Naomi
71	5Az.4A5a	71	71	2003	sum. 03	09/07/03	5	4	5500	Pachino	Naomi
72	5Az.4A6a	72	72	2003	sum. 03	09/07/03	5	4	5500	Pachino	Naomi
73	5Az.4A6b	73	73	2003	sum. 03	09/07/03	5	4	5500	Pachino	Naomi
74	5Az.4A6c	74	74	2003	sum. 03	09/07/03	5	4	5500	Pachino	Naomi
75	5Az.4A6d	75	75	2003	sum. 03	09/07/03	5	4	5500	Pachino	Naomi
76	5Az.4B4a	76		2003	sum. 03	09/07/03	5	4	5500	Pachino	Naomi
77	5Az.4B5a	77		2003	sum. 03	09/07/03	5	4	5500	Pachino	Naomi
78	5Az.4B5b	78	78	2003	sum. 03	09/07/03	5	4	5500	Pachino	Naomi
79	5Az.4B6a	79	79	2003	sum. 03	09/07/03	5	4	5500	Pachino	Naomi
80	5Az.4C4a	80		2003	sum. 03	09/07/03	5	4	5500	Pachino	Naomi
81	5Az.4C5a	81		2003	sum. 03	09/07/03	5	4	5500	Pachino	Naomi
82	5Az.4C6a	82	82	2003	sum. 03	09/07/03	5	4	5500	Pachino	Naomi
83	Random A	83		2003	sum. 03	09/07/03	5	Random		Pachino	Naomi

ID	Sample code	200MHz	400MHz	Year	Season	Date	group	Farm of Pachino	Salinity	Origin	Cultivar
84	Random B	84		2003	sum. 03	09/07/03	5	Random		Pachino	Naomi
90	6Az.4BFSa	90	90	2003	sum. 03	09/07/03	6	4	5500	Pachino	Naomi
91	6Az.4CFSa	91	91	2003	sum. 03	09/07/03	6	4	5500	Pachino	Naomi
92	7Az.1A2a	92	92	2003	aut. 03	09/12/03	7	1	5400	Pachino	Naomi
93	7Az.1A2b	93	93	2003	aut. 03	09/12/03	7	1	5400	Pachino	Naomi
94	7Az.1B2a	94	94	2003	aut. 03	09/12/03	7	1	5400	Pachino	Naomi
95	7Az.1B2b	95	95	2003	aut. 03	09/12/03	7	1	5400	Pachino	Naomi
96	7Az.1C2a	96	96	2003	aut. 03	09/12/03	7	1	5400	Pachino	Naomi
97	7Az.1C2b	97	97	2003	aut. 03	09/12/03	7	1	5400	Pachino	Naomi
98	7Az.2A2a	98	98	2003	aut. 03	09/12/03	7	2	1500	Pachino	Naomi
99	7Az.2A2b	99	99	2003	aut. 03	09/12/03	7	2	1500	Pachino	Naomi
100	7Az.2B2a	100	100	2003	aut. 03	09/12/03	7	2	1500	Pachino	Naomi
101	7Az.2B2b	101	101	2003	aut. 03	09/12/03	7	2	1500	Pachino	Naomi
102	7Az.2C2a	102	102	2003	aut. 03	09/12/03	7	2	1500	Pachino	Naomi
103	7Az.2C2b	103	103	2003	aut. 03	09/12/03	7	2	1500	Pachino	Naomi
104	8Az.1A3a	104	104	2003	win. 04	23/12/03	8	1	5400	Pachino	Naomi
105	8Az.1B3a	105	105	2003	win. 04	23/12/03	8	1	5400	Pachino	Naomi
106	8Az.1B3b	106	106	2003	win. 04	23/12/03	8	1	5400	Pachino	Naomi
107	8Az.1C3a	107	107	2003	win. 04	23/12/03	8	1	5400	Pachino	Naomi
108	8Az.1C3b	108	108	2003	win. 04	23/12/03	8	1	5400	Pachino	Naomi
109	8Az.2A3a	109	109	2003	win. 04	23/12/03	8	2	1500	Pachino	Naomi
110	8Az.2A3b	110	110	2003	win. 04	23/12/03	8	2	1500	Pachino	Naomi
111	8Az.2B3a	111	111	2003	win. 04	23/12/03	8	2	1500	Pachino	Naomi
112	8Az.2C3a	112	112	2003	win. 04	23/12/03	8	2	1500	Pachino	Naomi
113	8Az.2C3b	113	113	2003	win. 04	23/12/03	8	2	1500	Pachino	Naomi
114	9Az.1A4a	114	114	2004	win. 04	09/01/04	9	1	5400	Pachino	Naomi
115	9Az.1A4b	115	115	2004	win. 04	09/01/04	9	1	5400	Pachino	Naomi
116	9Az.1B4a	116	116	2004	win. 04	09/01/04	9	1	5400	Pachino	Naomi
117	9Az.1C4a	117	117	2004	win. 04	09/01/04	9	1	5400	Pachino	Naomi
118	9Az.1C4b	118	118	2004	win. 04	09/01/04	9	1	5400	Pachino	Naomi
119	9Az.2A4a	119	119	2004	win. 04	09/01/04	9	2	1500	Pachino	Naomi
120	9Az.2A4b	120	120	2004	win. 04	09/01/04	9	2	1500	Pachino	Naomi
121	9Az.2B4a	121	121	2004	win. 04	09/01/04	9	2	1500	Pachino	Naomi
122	9Az.2C4a	122	122	2004	win. 04	09/01/04	9	2	1500	Pachino	Naomi
123	10Az.1A5a	123	123	2004	win. 04	28/01/04	10	1	5400	Pachino	Naomi
124	10Az.1B5a	124	124	2004	win. 04	28/01/04	10	1	5400	Pachino	Naomi
125	10Az.1C5a	125	125	2004	win. 04	28/01/04	10	1	5400	Pachino	Naomi
126	10Az.2A5a	126	126	2004	win. 04	28/01/04	10	2	1500	Pachino	Naomi
127	10Az.2B5a	127	127	2004	win. 04	28/01/04	10	2	1500	Pachino	Naomi
128	10Az.2B5b	128	128	2004	win. 04	28/01/04	10	2	1500	Pachino	Naomi
129	10Az.2C5a	129	129	2004	win. 04	28/01/04	10	2	1500	Pachino	Naomi
130	11Az.1A6a	130	130	2004	win. 04	03/02/04	11	1	5400	Pachino	Naomi
131	11Az.1A6b	131	131	2004	win. 04	03/02/04	11	1	5400	Pachino	Naomi
132	11Az.1B6a	132	132	2004	win. 04	03/02/04	11	1	5400	Pachino	Naomi
133	11Az.1C6a	133	133	2004	win. 04	03/02/04	11	1	5400	Pachino	Naomi
134	11Az.2A6a	134	134	2004	win. 04	03/02/04	11	2	1500	Pachino	Naomi
135	11Az.2B6a	135	135	2004	win. 04	03/02/04	11	2	1500	Pachino	Naomi
137	12Az.1A2a	137	137	2004	win. 04	10/02/04	12	1	5400	Pachino	Naomi
138	12Az.1A2b	138	138	2004	win. 04	10/02/04	12	1	5400	Pachino	Naomi
139	12Az.1B2a	139	139	2004	win. 04	10/02/04	12	1	5400	Pachino	Naomi

ID	Sample code	200MHz	400MHz	Year	Season	Date	group	Farm of Pachino	Salinity	Origin	Cultivar
140	12Az.1B2b	140	140	2004	win. 04	10/02/04	12	1	5400	Pachino	Naomi
141	12Az.1C2a	141	141	2004	win. 04	10/02/04	12	1	5400	Pachino	Naomi
142	12Az.2A1a	142	142	2004	win. 04	10/02/04	12	2	1500	Pachino	Naomi
143	12Az.2B1a	143	143	2004	win. 04	10/02/04	12	2	1500	Pachino	Naomi
144	12Az.2C1a	144	144	2004	win. 04	10/02/04	12	2	1500	Pachino	Naomi
145	13Az.1A3a	145	145	2004	win. 04	25/02/04	13	1	5400	Pachino	Naomi
146	13Az.1B3a	146	146	2004	win. 04	25/02/04	13	1	5400	Pachino	Naomi
147	13Az.1C3a	147	147	2004	win. 04	25/02/04	13	1	5400	Pachino	Naomi
148	13Az.2A2a	148	148	2004	win. 04	25/02/04	13	2	1500	Pachino	Naomi
149	13Az.2B2a	149	149	2004	win. 04	25/02/04	13	2	1500	Pachino	Naomi
150	13Az.2C2a	150	150	2004	win. 04	25/02/04	13	2	1500	Pachino	Naomi
151	14Az.1A4a	151	151	2004	win. 04	10/03/04	14	1	5400	Pachino	Naomi
152	14Az.1B4a	152	152	2004	win. 04	10/03/04	14	1	5400	Pachino	Naomi
153	14Az.1C4a	153	153	2004	win. 04	10/03/04	14	1	5400	Pachino	Naomi
154	14Az.2A3a	154	154	2004	win. 04	10/03/04	14	2	1500	Pachino	Naomi
155	14Az.2B3a	155	155	2004	win. 04	10/03/04	14	2	1500	Pachino	Naomi
156	14Az.2C3a	156	156	2004	win. 04	10/03/04	14	2	1500	Pachino	Naomi
157	15Az.1A5a	157	157	2004	win. 04	19/03/04	15	1	5400	Pachino	Naomi
158	15Az.1B5a	158	158	2004	win. 04	19/03/04	15	1	5400	Pachino	Naomi
159	15Az.1C5a	159	159	2004	win. 04	19/03/04	15	1	5400	Pachino	Naomi
160	15Az.2A4a	160	160	2004	win. 04	19/03/04	15	2	1500	Pachino	Naomi
161	15Az.2B4a	161	161	2004	win. 04	19/03/04	15	2	1500	Pachino	Naomi
162	15Az.2C4a	162	162	2004	win. 04	19/03/04	15	2	1500	Pachino	Naomi
168	16Az.2C5a	168	168	2004	win. 04	30/03/04	16	2	1500	Pachino	Naomi
170	17Az.2B6a	170	170	2004	spr. 04	16/04/04	17	2	1500	Pachino	Naomi
171	17Az.2C6a	171	171	2004	spr. 04	16/04/04	17	2	1500	Pachino	Naomi
173	18Az.6A1a	173	173	2004	spr. 04	21/05/04	18	6	1500	Pachino	Naomi
174	18Az.6SHA1a	174	174	2004	spr. 04	21/05/04	18	6	1500	Pachino	Shiren
175	18Az.7SHA1a	175	175	2004	spr. 04	21/05/04	18	7	5600	Pachino	Shiren
176	18Az.8SHA1a	176	176	2004	spr. 04	21/05/04	18	8	6000	Pachino	Shiren
177	18 Random 1	177	177	2004	spr. 04	21/05/04	18	Random1	3600	Pachino	Shiren
179	18 Random3	179	179	2004	spr. 04	21/05/04	18	Random3	2200	Pachino	Shiren
180	18 Random4	180	180	2004	spr. 04	21/05/04	18	Random4	2200	Pachino	Naomi
181	18 Random 5	181	181	2004	spr. 04	21/05/04	18	Random5	1800	Pachino	Naomi
183	19Az.5A2a	183	183	2004	sum. 04	04/06/04	19	5	5500	Pachino	Naomi
184	19Az.6A2a	184	184	2004	sum. 04	04/06/04	19	6	1500	Pachino	Naomi
185	19Az.6ShA2a	185	185	2004	sum. 04	04/06/04	19	6	1500	Pachino	Shiren
186	19Az.8ShA2a	186	186	2004	sum. 04	04/06/04	19	8	6000	Pachino	Shiren
187	19Az.9SHA2a	187	187	2004	sum. 04	04/06/04	19	9	3600	Pachino	Shiren
188	20Az.5A3a	188	188	2004	sum. 04	18/06/04	20	5	5500	Pachino	Naomi
189	20Az.6A3a	189	189	2004	sum. 04	18/06/04	20	6	1500	Pachino	Naomi
190	20Az.6ShA3a	190	190	2004	sum. 04	18/06/04	20	6	1500	Pachino	Shiren
191	20Az.8SHA3a	191	191	2004	sum. 04	18/06/04	20	8	6000	Pachino	Shiren
192	20Az.9SHA3a	192	192	2004	sum. 04	18/06/04	20	9	3600	Pachino	Shiren
193	21Az.5A4a	193	193	2004	sum. 04	21/06/04	21	5	5500	Pachino	Naomi
194	21Az.5A5a	194	194	2004	sum. 04	21/06/04	21	6	1500	Pachino	Naomi
195	21Az.6A4a	195	195	2004	sum. 04	21/06/04	21	6	1500	Pachino	Shiren
196	21Az.6SHA4a	196	196	2004	sum. 04	21/06/04	21	6	1500	Pachino	Shiren
197	21Az.6A5a	197	197	2004	sum. 04	21/06/04	21	8	6000	Pachino	Shiren
198	21Az.6SHA5a	198	198	2004	sum. 04	21/06/04	21	9	3600	Pachino	Shiren

ID	Sample code	200MHz	400MHz	Year	Season	Date	group	Farm of Pachino	Salinity	Origin	Cultivar
199	21Az.8SHA4a	199	199	2004	sum. 04	21/06/04	21	5	5500	Pachino	Naomi
200	21Az.8SHA5a	200	200	2004	sum. 04	21/06/04	21	6	1500	Pachino	Naomi
201	21Az.9SHA4a	201	201	2004	sum. 04	21/06/04	21	8	6000	Pachino	Shiren
202	21Az.9SHA5a	202	202	2004	sum. 04	21/06/04	21	9	3600	Pachino	Shiren
203	22Az5A2a	203	203	2005	win. 05	26/02/05	22	5	5500	Pachino	Naomi
204	22Az5B2a	204	204	2005	win. 05	26/02/05	22	5	5500	Pachino	Naomi
205	22Az5C2a	205	205	2005	win. 05	26/02/05	22	5	5500	Pachino	Naomi
206	22Az9ShA1a	206	206	2005	win. 05	26/02/05	22	9	3600	Pachino	Shiren
207	22Az9ShB1a	207	207	2005	win. 05	26/02/05	22	9	3600	Pachino	Shiren
208	22Az9ShC1a	208	208	2005	win. 05	26/02/05	22	9	3600	Pachino	Shiren
209	23Az5A3a	209	209	2005	win. 05	15/03/05	23	5	5500	Pachino	Naomi
210	23Az5B3a	210	210	2005	win. 05	15/03/05	23	5	5500	Pachino	Naomi
211	23Az5C3a	211	211	2005	win. 05	15/03/05	23	5	5500	Pachino	Naomi
212	23Az5A4a	212	212	2005	win. 05	15/03/05	23	5	5500	Pachino	Naomi
213	23Az5B4a	213	213	2005	win. 05	15/03/05	23	5	5500	Pachino	Naomi
214	23Az5C4a	214	214	2005	win. 05	15/03/05	23	5	5500	Pachino	Naomi
215	23Az9ShA2a	215	215	2005	win. 05	15/03/05	23	9	3600	Pachino	Shiren
216	23Az9ShB2a	216	216	2005	win. 05	15/03/05	23	9	3600	Pachino	Shiren
217	23Az9ShC2a	217	217	2005	win. 05	15/03/05	23	9	3600	Pachino	Shiren
218	23Az9ShA3a	218	218	2005	win. 05	15/03/05	23	9	3600	Pachino	Shiren
219	23Az9ShB3a	219	219	2005	win. 05	15/03/05	23	9	3600	Pachino	Shiren
220	23Az9ShC3a	220	220	2005	win. 05	15/03/05	23	9	3600	Pachino	Shiren
221	23Az10ShR4a	221	221	2005	win. 05	15/03/05	23	10	5000	Pachino	Shiren
222	23Az11ShR8a	222	222	2005	win. 05	15/03/05	23	11	3600	Pachino	Shiren
223	23Az12ShR7/8a	223	223	2005	win. 05	15/03/05	23	12	4000	Pachino	Shiren
224	23Az13R1a	224	224	2005	win. 05	15/03/05	23	13	2000	Pachino	
225	23Az14R4a	225	225	2005	win. 05	15/03/05	23	14	2200	Pachino	
226	23Az15ShR1a	226	226	2005	win. 05	15/03/05	23	15	1800	Pachino	Shiren
231	24 Az5A5a	231	231	2005	spr. 05	05/04/05	24	5	5500	Pachino	Naomi
237	24Az9ShA4a	237	237	2005	spr. 05	05/04/05	24	9	3600	Pachino	Shiren
243	24Az17ShR3a	243	243	2005	spr. 05	05/04/05	24	17	4000	Pachino	Shiren
248	24Az19R8/9a	248	248	2005	spr. 05	05/04/05	24	19	5000	Pachino	Naomi
250	24Az21ShR6a	250	250	2005	spr. 05	05/04/05	24	21	1500	Pachino	Shiren
252	24Az23ShR6a	252	252	2005	spr. 05	05/04/05	24	23	2200	Pachino	Shiren
257	25Az9ShB7a		257	2005	spr. 05	29/04/05	25	9	3600	Pachino	Shiren
279	26AZ.18SHC2a		279	2005	sum. 05	30/06/05	26	18	1800	Pachino	
285	27AZ.1SHC4a		285	2005	sum. 05	30/06/05	26	1	5400	Pachino	
501	1Licata NaomiA1a	501	501	2004	win. 04	16/01/04	1			Licata	Naomi
502	1Licata NaomiB1a	502		2004	win. 04	16/01/04	1			Licata	Naomi
503	1Licata NaomiC1a	503		2004	win. 04	16/01/04	1			Licata	Naomi
504	1Licata ShirenA1a	504	504	2004	win. 04	16/01/04	1			Licata	Shiren
505	1Licata ShirenB1a	505		2004	win. 04	16/01/04	1			Licata	Shiren
506	1Licata ShirenC1a	506		2004	win. 04	16/01/04	1			Licata	Shiren
507	2Licata NaomiA3a	507		2004	win. 04	13/02/04	2			Licata	Naomi
508	2Licata NaomiB3a	508		2004	win. 04	13/02/04	2			Licata	Naomi
509	2Licata NaomiC3a	509		2004	win. 04	13/02/04	2			Licata	Naomi
510	2Licata ShirenA3a	510	510	2004	win. 04	13/02/04	2			Licata	Shiren
511	2Licata ShirenB3a	511		2004	win. 04	13/02/04	2			Licata	Shiren
512	2Licata ShirenC3a	512		2004	win. 04	13/02/04	2			Licata	Shiren
513	3Licata NaomiA4a		513	2004	win. 04	02/03/04	3			Licata	Naomi

ID	Sample code	200MHz	400MHz	Year	Season	Date	group	Farm of Pachino	Salinity	Origin	Cultivar
514	3Licata NaomiA4b		514	2004	win. 04	02/03/04	3			Licata	Naomi
515	3Licata NaomiB4a		515	2004	win. 04	02/03/04	3			Licata	Naomi
516	3Licata NaomiB4b		516	2004	win. 04	02/03/04	3			Licata	Naomi
517	3Licata NaomiC4a		517	2004	win. 04	02/03/04	3			Licata	Naomi
518	3Licata NaomiC4b		518	2004	win. 04	02/03/04	3			Licata	Naomi
519	3Licata ShirenA4a		519	2004	win. 04	02/03/04	3			Licata	Shiren
520	3Licata ShirenB4a		520	2004	win. 04	02/03/04	3			Licata	Shiren
521	3Licata ShirenB4b		521	2004	win. 04	02/03/04	3			Licata	Shiren
522	3Licata ShirenC4a		522	2004	win. 04	02/03/04	3			Licata	Shiren
701	1Sabaudia RubinoTopB1	701	701	2004	sum. 04	01/09/04	1			Sabaudia	Rubino top
702	1Sabaudia RubinoTopB2	702	702	2004	sum. 04	01/09/04	1			Sabaudia	Rubino top
703	1Sabaudia RubinoTopA1	703	703	2004	sum. 04	01/09/04	1			Sabaudia	Rubino top
704	1Sabaudia RubinoTopA2	704	704	2004	sum. 04	01/09/04	1			Sabaudia	Rubino top
705	1SabaudiaSHA1	705	705	2004	sum. 04	01/09/04	1			Sabaudia	Shiren
706	1SabaudiaSHA2	706	706	2004	sum. 04	01/09/04	1			Sabaudia	Shiren
707	1SabaudiaSHB1	707	707	2004	sum. 04	01/09/04	1			Sabaudia	Shiren
708	1SabaudiaSHB2	708	708	2004	sum. 04	01/09/04	1			Sabaudia	Shiren
710	1SabaudiaSH1R2	710	710	2004	sum. 04	01/09/04	1			Sabaudia	Shiren
711	1SabaudiaSH2R1	711	711	2004	sum. 04	01/09/04	1			Sabaudia	Shiren
712	1SabaudiaSH2R2	712	712	2004	sum. 04	01/09/04	1			Sabaudia	Shiren
713	2SabaudiaSHC3	713	713	2004	aut. 04	28/09/04	2			Sabaudia	Shiren
714	2SabaudiaSHC4	714	714	2004	aut. 04	28/09/04	2			Sabaudia	Shiren
715	2SabaudiaSHC5	715	715	2004	aut. 04	28/09/04	2			Sabaudia	Shiren
716	2SabaudiaSHD2	716	716	2004	aut. 04	28/09/04	2			Sabaudia	Shiren
717	2SabaudiaSHD3	717	717	2004	aut. 04	28/09/04	2			Sabaudia	Shiren
718	2SabaudiaSHD4	718	718	2004	aut. 04	28/09/04	2			Sabaudia	Shiren
719	2SabaudiaSHD5	719	719	2004	aut. 04	28/09/04	2			Sabaudia	Shiren
720	2SabaudiaSH3R6	720	720	2004	aut. 04	28/09/04	2			Sabaudia	Shiren
721	2Sabaudia RubinoTop B3	721	721	2004	aut. 04	28/09/04	2			Sabaudia	Rubino top
722	2Sabaudia RubinoTop B4	722	722	2004	aut. 04	28/09/04	2			Sabaudia	Rubino top
723	2Sabaudia RubinoTop B5	723	723	2004	aut. 04	28/09/04	2			Sabaudia	Rubino top
724	2Sabaudia Franchie1R4	724	724	2004	aut. 04	28/09/04	2			Sabaudia	Franchie
725	2SabaudiaFranchie2R5	725	725	2004	aut. 04	28/09/04	2			Sabaudia	Franchie
903	Lazio	903	903	2003	sum. 03	20/07/03				market	
904	Puglia	904	904	2003	sum. 03	20/07/03				market	
905	Gela	905	905	2003	sum. 03	Luglio		Ortonatura2000		market	Naomi
906	Pachino	906	906	2003	win. 04	06/12/03				market	
907	Gela	907	907	2004	sum. 04	19/06/04		One Ortofrutta		market	Naomi
908	Ciliegini domestici	908	908	2004	sum. 04	16/08/04				market	
909	Cieliegini Maccarese	909	909	2004	sum. 04	28/08/04				market	
910	OrtoPIU' Licata	910	910	2004	sum. 04	16/06/04		OrtoPiù		market	
911	1OrtoNaturaNaomi	911	911	2005	win. 05	01/02/05		Ortonatura2000		market	
912	2OrtoNaturaNaomi	912	912	2005	win. 05	09/02/05		Ortonatura2000		market	Naomi
913	3OrtoNaturaNaomi	913	913	2005	win. 05	14/02/05		Ortonatura2000		market	Naomi
914	4OrtoNaturaNaomi	914	914	2005	win. 05	16/02/05		Ortonatura2000		market	Naomi
915	5OrtoNaturaNaomi	915	915	2005	win. 05	19/02/05		Ortonatura2000		market	Naomi
916	Fondi(Lt)CIDCorus	916	916	2005	win. 05	22/02/05		CIDCorus		market	Corus
917	Fondi(Lt)CIDPiccadilly	917	917	2005	win. 05	22/02/05		CIDPiccadilly		market	Piccadilly
918	6OrtoNaturaNaomi	918	918	2005	win. 05	02/03/05		Ortonatura2000		market	Naomi
919	7OrtoNaturaNaomi	919	919	2005	win. 05	08/03/05		Ortonatura2000		market	Naomi

ID	Sample code	200MHz	400MHz	Year	Season	Date	group	Farm of Pachino	Salinity	Origin	Cultivar
920	2OrtoPIU' Licata	920	920	2005	win. 05	18/03/05		OrtoPiù		market	
921	Piccadylli Sicilia	921	921	2005	win. 05	30/03/05				market	Piccadilly
922	3OrtoPIU'Licata	922	922	2005	spr. 05	01/04/05		OrtoPiù		market	
923	1OrtofrutticolaACESE	923	923	2005	spr. 05	04/04/05		OrtofruttaACESE		market	
924	8OrtoNaturaNaomi	924	924	2005	spr. 05	21/04/05		Ortonatura2000		market	Naomi
925	9OrtoNaturaNaomi	925	925	2005	spr. 05	22/04/05		Ortonatura2000		market	Naomi
926	10OrtoNaturaNaomi	926	926	2005	spr. 05	23/04/05		Ortonatura2000		market	Naomi
927	11OrtoNaturaNaomi	927	927	2005	spr. 05	26/04/05		Ortonatura2000		market	Naomi
928	12OrtoNaturaNaomi	928	928	2005	spr. 05	27/04/05		Ortonatura2000		market	Naomi
929	13OrtoNaturaNaomi	929	929	2005	spr. 05	28/04/05		Ortonatura2000		market	Naomi

Publications and congress participations.

Alessandri S., Capozzi F., Cremonini M. A., Luchinat C., Placucci G., Savorani F., Turano M. (2005) PARAMAGNETIC CHALLENGES IN NMR MEASUREMENTS OF FOODS in Magnetic Resonance in Food Science. The Multivariate Challenge. Edited by S. Engelsen; P.S. Belton; H.J. Jakobsen (2005) ISBN 0-85404-648-8

Savorani F., Capozzi F., Placucci G. (2005). IDENTIFICATION OF MOLECULAR MARKERS FOR CHARACTERIZATION OF IGP CHERRY TOMATOES OF PACHINO USING HR-NMR SPECTROSCOPY. Congress Acta, "XXXV Italian National Congress on Magnetic Resonance", Roma, A68

Savorani F., Placucci G. (2005). Congress Acta "10th Workshop on the Developments in the Italian PhD Research in Food Science and Technology", Foggia 7-9 September 2005, 828

Savorani F., Capozzi F., Placucci G. (2006). NMR AND HPLC COUPLING IN CHEMOMETRICS: TOWARD THE IDENTIFICATION OF MOLECULAR MARKERS. Congress Acta, "XXXVI Italian National Congress on Magnetic Resonance", Vietri sul Mare (SA) 20-23 Settembre 2006, 94

Savorani F., Placucci G. (2006). A NUCLEAR MAGNETIC RESONANCE APPLICATION FOR THE ASSESSMENT OF THE MOLECULAR DIVERSITY AMONG TOMATOES WITH DIFFERENT GEOGRAPHIC ORIGIN. Congress Acta, "11th Workshop on the Developments in the Italian PhD Research in Food Science and Technology", Mosciano S.Angelo (TE) 27-29 September 2006, 249-252

Savorani F., Capozzi F., Placucci G. (2007). APPLICATION OF ^{1}H NMR TOCSY FILTERING FOR COUPLING BETWEEN NON HYPHENATED 1D ^{1}H NMR SPECTRA AND HPLC. In prep.