Chemometric data treatment of spectroscopic measurements for characterization of monovarietal extra-virgin olive oil from Marche

Rosangela Leone

Thesis to obtain the Master of Science Degree in

Chemistry

Supervisor: Dr. Paolo Conti

Examination Committee

Chairperson: Professora Doutora Matilde Marques

Supervisor: Professora Doutora Isabel Maria Marrucho Ferreira

Members of the Committee: Professor Doutor João Almeida Lopes

[October 2017]

INDEX

"A me stessa, per essere riuscita ad ottenere anche questo nuovo traguardo! "

INTRODUCTION

By virtue of its high nutritional and organoleptic properties, extra virgin olive oil is an expensive product that can cost even 4–5 times more than other edible vegetable oils[I]. Extra virgin olive oil is, therefore, a product with high added value and as such must provide guarantees of quality to justify, in the eyes of the consumer, its higher cost. The term "quality" is used in this case to indicate a set of meanings, and can best be described as "the totality of characteristics of a product that give that entity the ability to satisfy expressed or implicit needs" (ISO 8402): in this sense, it includes the genuineness of a product, the safety of use, the typicity, the absence of adulterations, and so on[II]. In particular, as far as extra virgin olive oils are concerned, it is widely reported in the literature, that quality characteristics and taste are largely related to their origin, both geographical and varietal, as well as to the agronomic techniques and the extraction and mixing procedures used[III].

In order to protect and preserve the specificity of many traditional foods, which owe their peculiar characteristics to the area of origin or to the local techniques of production, the EU has established the so-called "European brands" for food and agricultural products (Reg. EEC 2081/92 and 2082/92)[IV],i.e., special "regional" designations of origin, assigned according to the characteristics of the product which are linked to its geographical origin[V]. In particular, PDO status (protected designation of origin) indicates a food whose production, processing and preparation take place in a particular geographical area and characterized by a recognized and certified know-how; PGI (protected geographical indication), on the other hand, indicates a product of some repute, for which the link with the territory is present in at least one of the stages of production, processing or preparation of the finished product[VI]. To be able to boast of such quality status, food must undergo a strict production regulation, disciplining the borders of the geographical area, raw materials and processing techniques used, the most important physicochemical, microbiological and/or organoleptic characteristics of the product, etc.

The analytical control of labelling compliance plays, therefore, a role of fundamental importance for the protection of a high quality food from unfair competition from other products with the same trade name but a value, and consequently a sale price, significantly lower. In this sense, the verification of the origin is a key factor in establishing the authenticity of oil[VII]. It should be remembered that the verification of the authenticity of a product is a complex problem that takes into consideration the food in its entirety, and then requires an assessment far more articulate than simply measuring a parameter associated with some properties of the food in exam. For this reason, the problems of food authentication often need a multivariate approach for their resolution, and find a very effective investigative tool in the use of chemometric classification techniques[VIII].

ORIGINS and production OLIVE OIL

The olive tree (Olea Europea), traces its origins back to Eastern Mediterranean countries and the Middle East. It is believed that the first olive trees were planted by the Semite-Hamitic tribes which inhabited the southern slopes of the Caucasus, the Western Iranian plateaus, Syria and Palestine. From there, the cultivation of the olive tree expanded to Egypt and, later, to the Greek islands, in particular to Cyprus, Rhodes and Crete in the Mediterranean, the history of both mankind and the olive tree go hand in hand. Besides being the most important tree, in the Mediterranean, it has come to symbolize this area. The climate which is most suited to growing olive trees is that of the temperate zone, between the 30 and 45 northern parallel and the corresponding belt located in the southern hemisphere. A characteristic of the olive tree is that it must be subjected to a certain degree of cold during the winter months, otherwise it will neither flower nor bear fruit. Olive trees may reach a height ranging between 10-13 meters. The tree is very resistant and stands up well to humidity. If, for some reason, the trunk should die, the tree is able to regenerate itself into a new plant which grows at the roots. The olive tree lives for a great number of years; some have lived for more than 1,000 years.

HARVESTING THE OLIVES

TRADITIONAL HARVESTING METHODS

There are two traditional methods of harvesting the olives. The first, is known as the "brucatura" in which the olives are hand-picked using ladders to reach the higher branches. The other method is that which either leaves the olives to fall to the ground either by letting them ripen and fall off the branches or by shaking the trunk or by striking it with bamboo canes causing the ripe olives to fall to the ground. Large nets are laid out below the tree in order to facilitate gathering the olives. The so called "brucatura" method represents the ideal way of harvesting the olives because the olives are picked when they are mature and they haven't been bruised by falling to the ground. However, this is a very time-consuming and costly method. In fact, in an hour, a picker is only able to harvest 5-6 kg. of olives.

The “raccattatura” method (letting the olives fall naturally to the ground) is the poorest method in that the olives which fall to the ground are usually those which are over-ripe and once they hit the ground they bruise and begin to rapidly deteriorate, resulting in a poor quality olive oil.

MECHANICAL HARVESTING METHODS

These methods have been introduced mainly in an attempt to lower production costs. Both the “brucatura” method and the “raccattatura” method (picking up the olives from the ground after having shaken the tree), can be mechanized. One way is by using “combing” machines, which pass large-toothed blades through the branches, literally combing the tree, together with suction equipment which sucks up the olives from the ground and shoots them with a jet of air into the bin. The more sophisticated vacuum equipment separates the leaves and twigs from the olives. However, the most interesting piece of equipment is the “vibrating” harvesting machines. This machine has a long arm which clamps around the trunk of the tree and shakes the tree. The advantage of this machine is enormous, and with the new series of machines the fear of ruining the root system of the tree by repeatedly vibrating the trunks has disappeared. Obviously, the use of this machine becomes more cost-efficient when the olive grove lends itself to this mechanical means of harvesting. However, the quality of the olives is not the same as when the olives are hand-picked because the machine isn’t able to select the mature olives from those which are under or over ripe. On the other hand, the use of this type of machine yields a better product than that obtained by letting the mature olives fall to the ground.

STORING THE OLIVES

Usually, the olives aren’t crushed right away mainly because it takes longer to crush the olives than to pick them and this creates a back-log. This is a very delicate moment due to the fact that if the olives are not stored properly, the resulting quality of the olive oil produced would be poor due to the increased acidity of the bruised fruit.. Moreover, even in ideal conditions, the olives should not be stored for more than one week. Olives which are very ripe or that have been damaged or bruised must be crushed within a couple of days, at the most.

PRODUCING OLIVE OIL

The age old process of extracting the oil from the olives is comprised of separate stages: the crushing stage, the kneading or mixing stage, and the extraction stage. The extraction stage may be carried out in several ways. It is crucial, however, that the capacity of all the equipment is such that the process is carried out harmoniously with no delays and without leaving the semi-finished product for long periods. In fact, the olive paste is even more delicate than the olives and is subject to undergo changes which will compromise the quality of the final product.

EXTRACTION METHOD

Prepare and wash with water to remove leaves and impurities that alter the taste and the aroma of the oil, the drupe are subjected to the milling phase in order to break the cell wall and freeing the cellular juices and the oil contained inside. The operation can be accomplished through a discontinuous classic stone millstone system, or through a modern breeder with hammers, knives, or teeth, which makes the breaking of pulp through the progressive shock of rotating devices at high speed. It is a critical phase for the quality of oil, because during this process, with the breaking of the cell walls, activates the enzymes present in both the pulp and in seed, they play a dual action: one, of a positive nature that consists in extracting the volatile substances responsible for the aroma and the other, of a negative nature that it consists of the oxidation of the oleic matrix. Under the basic quality aspect (mainly in terms of acidity and number of Peroxides) the two processes tend to coexist but may have marked influences on the Organoleptic characteristics; The classic method has a low degree of emulsion, an average higher organoleptic rating and a less spicy and less bitter taste because the polyphenol content is slightly lower on average. Conversely, the violent action of the crushing causes a degree of emulsion pushed between water and oil by having an extra polyphenol extraction capacity that leads to obtain and an average oil more rich in chlorophyll in a lesser time. Ultimately, the classic method is best suited for cultivating cultivars that have potentials of high quality, obtaining high quality oils whose price can recover the major processing costs; Conversely, the crushing system is more suitable for the production of oils of slightly inferior quality.

THE PRODUCTION OF OLIVE OIL

FILTERING

The oil which is obtained from the paste must be filtered in order to get rid of any small particles of pulp and vegetal water which is a natural component of the olive. This is done by decanting the oil and transferring it to clean stainless steel tanks. Once it was necessary to transfer the oil three or four times in order to obtain a clear oil. Nowadays, with the aid of the centrifugal separator, the oil needs to be transferred only once. The filtered oil isn't necessarily transparent, in fact, a characteristic of the more prized oils is the opaque color.

STORAGE

Traditionally the olive oil was stored in terracotta jars which were enameled on the inside. These jars were called "orci". Today, almost all large bottlers use stainless steel tanks which are much easier to clean and tend to block out more light which can be detrimental to the olive oil causing it to go rancid. Olive oil has a shelf life of 18-24 months as long as it is protected from direct light and temperatures which are either extremely high or extremely low. The ideal temperature for olive oil is around 14-15 C.

BOTTLING

The best containers of olive oil are those made of either colored or transparent glass and aluminum or stainless steel tins. Out of these, the best choice is probably glass as it satisfies both hygienic needs as well as those pertaining to the marketing and selling of the product because it is important that the potential customer can see the olive oil through the glass bottle.

THE TYPES OF OLIVE OIL

By law, olive oil may be classified as follows, each representing a different processing method:

1. Extra Virgin Olive Oil is the result of the simple crushing of the olives which have been washed and separated from the leaves. This is by far the best product offering the widest range of perfect flavors and aromas with a maximum acidity of 1% (0.8 gram/100g of free oleic acid). Extra Virgin Olive Oil must meet the highest standards of flavor and aroma.

2. Virgin Olive Oil results from pressing olives which are over-ripe or have been bruised, and therefore have a high acidity. This type of oil has not been treated chemically.

3. Ordinary Olive Oil (sometimes known as “Pure”) is a blend of refined olive oil and virgin olive oil. Refined olive oil is an oil which has undergone chemical manipulation in order to permit the human consumption. This is blended with some virgin olive oil (usually 10%).

4. Olive Pomace Oil: The olive paste, which is left after the oil has been removed, still contains a small amount of oil which is extracted by using chemical solvents. The resulting oil, after having been treated to remove any chemical residue, is then blended with virgin olive oil.

DIFFERENCE BETWEEN FILTERED AND UNFILTERED OILS

Extra virgin olive oil may be consumed either in a filtered or unfiltered state. Filtration is the process by which the microscopic bits of the fruit of the olive are removed from the oil. Unfiltered oil will be cloudy until it settles to the bottom. Some consider unfiltered oil superior because of the added flavor from the fruit, while others say it shortens the oil’s shelf life.

Extra Virgin olive oils are not all the same. Like wines, extra virgin olive oils can vary dramatically in taste, depending upon the type and quality of the fruit that is pressed, the time of harvest, the weather during the growing season, and the region from which the olives were produced.

Connoisseurs generally use the following adjectives in appraising extra virgin olive oils: mild, semi-fruity and fruity depending on the flavor of the olive that can be detected. Further, some oils, such as the finer oils from Tuscany and Southern Italy, have a peppery finish that many appreciate.

CULTIVAR OF MARCHE

The Marche is one of the hilly regions of Italy: the hills comprise 69% of the territory (6,462.90 km²). The remaining 2,902.96 km² are mountainous. The plains are confined to a narrow coastal strip and to the parts of the valleys closest to the mouth of the rivers, but represent an negligible percentage. The region is characterized by three different climatic ranges: coastal strip, mid-low hills and mountain range (high hills). The coastal strip is characterized by a sub continental climate with temperature fluctuations from season to season: latitude and low altitude make the climate a whole mild, with warm summers but refreshed by the gentle sea breeze and cold winters with regular rains of the season. In mountainous areas there are more fresh summers and very cold winters with wide possibilities of snow; Winter is also stiff in indoor hilly areas where low temperatures can occur. As for the climate, the shape of the soil is quite varied: there are areas with a presence of limestone that makes the soil soft and aerated, but also sandy, clayey and chalky areas.

The combination of climatic conditions and orographic characteristics allowed for the development of olive cultivation, even in narrow areas, while the presence of microclimate of different olive breeds that settled in the Marche region give rise to the production of oils with very different organoleptic characteristics all of them remarkable value. What is passed on from olive to olive is a heritage not only genetic, but ancestral: every variety of oil speaks of its original area, its climate, its orographic features and its tradition: from the database of the project "oil Italian monovarietals " shows that only 31 varieties of oil are present in the Marche region[IX]. However, this number does not capture the significance of the quantity of oil that is produced: it is a largely mountainous and not fully cultivated territory and most of the Marche companies that invest a portion of the land on olive cultivation are medium to small and small ones. What you are looking for in the Marches is the quality both in the nutritional and in the gastronomic style, with distinctive character traits.

The value of olive production is 20 million euros. The Marche also have two DOPs in the olive oil sector (Cartoceto Dop Oil and Ascolana del Piceno Dop Olive Oil) and also IGP Marche.

PIANTONE DI MOGLIANO

The plants of the Piantone di Mogliano, hereafter simply named Mogliano, plant variety, which have their original nucleus in the municipality of Mogliano[X], tile the hills of the whole province of Macerata in a discreet manner, until to reach the most internal areas, at altitudes above 600m. Plants of this variety have a reduced vegetative development, characterized by a widespread and languid growth, the trunk rich in growth, the branched vegetation, the small, elongated leaves of dark gray green in the upper and light gray green in the lower one.

The remarkable diffusion of this variety throughout the province of Macerata is linked mainly to the interesting agronomic characteristics. Trees, with good cold resistance, come into production very early (at three years old they are already loaded with fruits) and keep good productivity for a long time.

Oil of a light fruity, tenderness sweet, of yellow gold, with a good content in oleic acid and a high unsaturated ratio.

Fatty acids (% ± E.S.):

- Palmitic acid 10.64 ± 0.73
- Palmitoleic acid 0.71 ± 0.12
- Stearic acid 1,90 ± 0,10
- Oleic acid 79.50 ± 1.01
- Oleic acid 79.50 ± 1.01
- Linoleic acid 5.96 ± 0.38
- Linoleic acid 5.96 ± 0.38
- Linolenic acid 0.57 ± 0.06

Unsaturated / saturated fatty acid ratio (n ± E.S.): 7.20 ± 0.51

Total polyphenols (mg / kg ± E.S.): 382.18 ± 27.29

Total chlorophyll (ppm ± E.S.): 5.31 ± 0.88

The organoleptic characteristics of the product obtained, summarized in **Figure 1**, describe a tendency of yellow, harmonious and balanced oil, with a light fruity and pleasant herbaceous notes combined with apple, tomato and almond hints. that highlight a sweet character with mild spicy and bitter feelings.

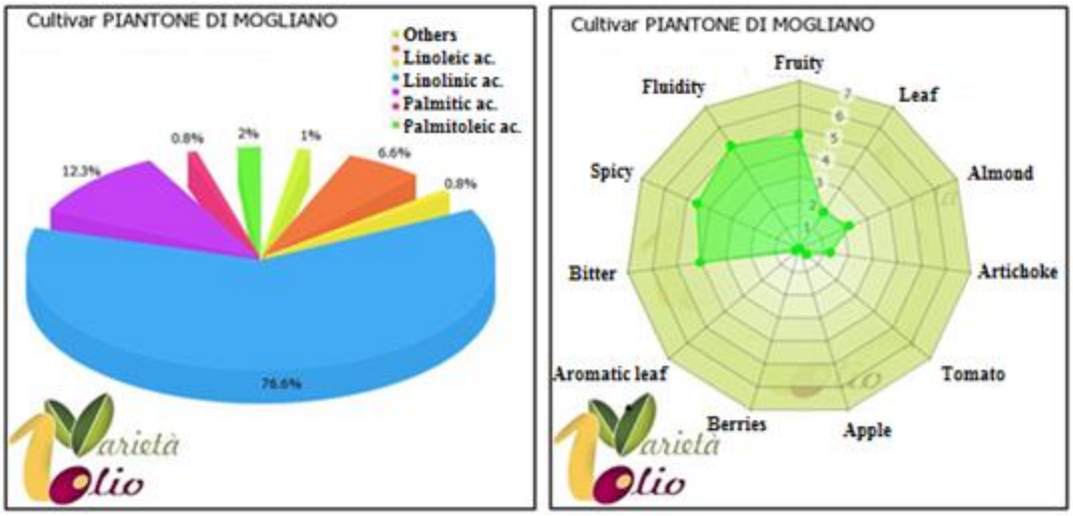


Figure 1 Chemical and Organoleptic composition of the Cultivar Piantone del Mogliano

MIGNOLA

The Mignola variety, originating in the area of Cingoli (MC), it is spread all over the hinterland between the provinces of Macerata and Ancona, till to the sea, finding it at the foot of Mount Conero and descending to the province of Ascoli Piceno, thanks also to its interesting productive features and environmental adaptation. Its tall, vigorous, quasi-monumental trees show massive, thick, blooming patches, large and expanded leaves with dark green on the top and light green in the lower one. The fruit is a very small drupe (an olive weighs just over a gram) that is difficult to collect, asymmetrical ovoid shape and color that changes from intense green to black ink with maturation. The pulp in the mature olives tends to become purplish and to soften to such an extent that it can be disintegrated in the hands at harvest time. Oil yield is very high, with accumulation also in this case of early oil.

The oil has medium fruity and a characteristic very bitter taste, yellow Gold color, high polyphenols content.

Fatty acids (% ± E.S.):

- Palmitic acid 14.44 ± 0.66
- Palmitoleic acid 1.75 ± 0.14
- Stearic acid 1,53 ± 0,14
- Oleic acid 72.88 ± 1.0
- Linoleic acid 8.60 ± 0.95

- Linolenic acid 0.43 ± 0.05

Unsaturated / saturated fatty acid ratio (n ± E.S.): 5.23 ± 0.18

Total polyphenols (mg / kg ± E.S.): 694.02 ± 69.67

Total chlorophyll (ppm ± E.S.): 9.11 ± 2.95

In the following **Figure 2** the chemical and organoleptic composition of the Cultivar Mignola.

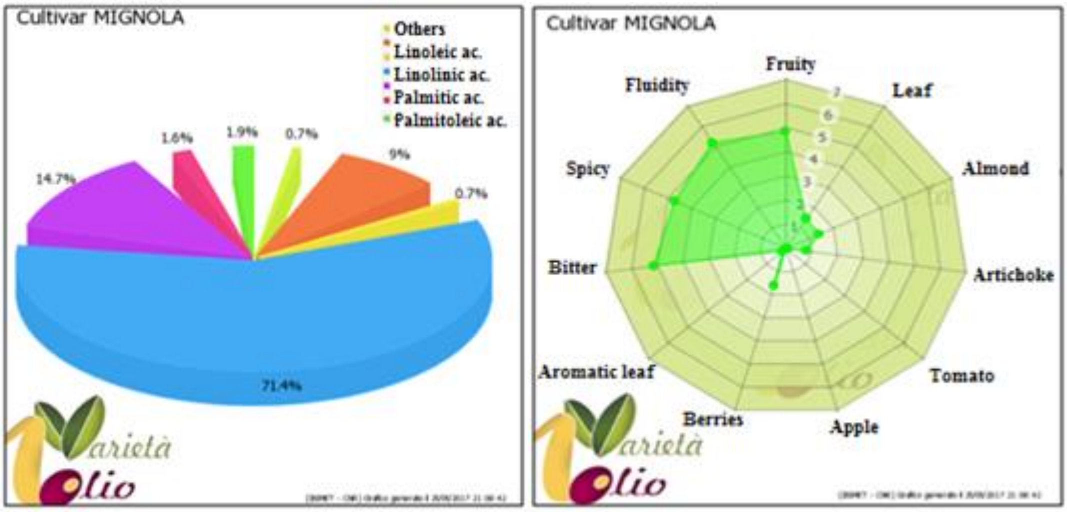


Figure 2 Chemical and Organoleptic composition of the Cultivar Mignola

RAGGIA

Variety of oil suitable for low-density plants, for manual or pneumatic harvesting. Also suitable for mechanical harvesting after appropriate pruning operations. Widespread cultivar of Marche in the province of Ancona, with more concentration in Ostra, Monte San Vito, Morro d'Alba. The productive characters are: early plant production, medium size (2 - 2.5 g), of elongated and asymmetrical ovoid shape and high yield in medium-high oil. Variety of oils similar to ‘Frantoio’ which occasionally repeats the anatomical aspects..

Oil of a good fruity almond, slightly bitter and spicy, greenish to yellow, with good unsaturated ratio.

Fatty acids (% ± E.S.):

- Palmitic acid 10,42 ± 0,14
- Palmitoleic acid 0,81 ± 0,52
- Stearic acid 1,70 ± 0,29
- Oleic acid 76,44 ± 3,41

- Linoleic acid 8,97 ± 1,65
- Linolenic acid 0,61 ± 0,06

Unsaturated / saturated fatty acid ratio (n ± ES): 7, 41 ± 0,14

Total polyphenols (mg / kg ± ES): 425,75 ± 134,75

Total chlorophyll (ppm ± ES): 7,34 ± 0,12

In the following **Figure 3** is reported the chemical and organoleptic composition of the cultivar Raggia.

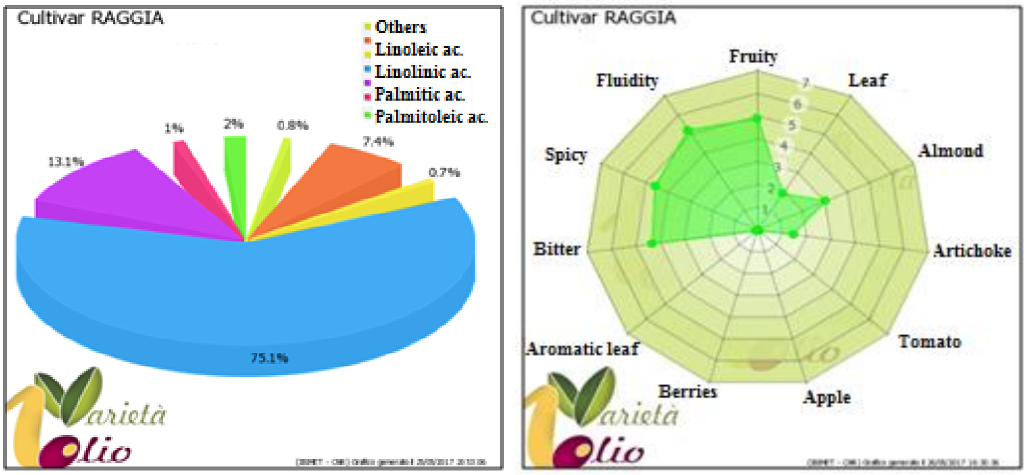


Figure 3 Chemical and Organoleptic composition of the Cultivar Raggia

CORONCINA

The Coroncina, also called Coronella, as a result of a specific project of valorization, has been defined as the variety of the five municipalities being predominantly localized in the areas of Caldarola (MC), Serrapetrona (MC), Belforte del Chienti (MC) , Camporotondo of Fiastrone (MC) and Cessapalombo (MC). Its presence was also found in neighboring municipalities, up to altitudes greater than 600 m. The tree is moderately vigorous, with a longitudinal and slightly branched growth. The leaves are of intense green color and the fruit of medium size and ovoid shape whose color changes from bright green to light green to violet red. The pulp remains consistent up to the high stage of ripening of the 'drupe'; This aspect, coupled with high oleuropein content, discourages the fly's oviduction in fruits. The high drought resistance allows cultivation in poor and gravel soils, even if with yields in oil significantly inferior to the abovementioned varieties. In addition, good resistance to cold has so far ensured survival in the area, though not without difficulty.

Like the oil of Mignola but for different reasons, the oil produced by the Coroncina variety differs from the classic type of Marche's oil.

High quality, very fruity, bitter and biting oil with a hint of artichoke, color Intense green, high in polyphenols and chlorophyll and good relationship unsaturated / saturated.

Fatty acids (% ± E.S.):

- Palmitic acid 10.54 ± 0.27
- Palmitoleic acid 0.62 ± 0.17
- Stearic acid 1,60 ± 0,17
- Oleic acid 78.37 ± 1.60
- Linoleic acid 7.34 ± 0.45
- Linolenic acid 0.63 ± 0.09

Unsaturated / saturated fatty acids (n ± E.S.) ratio: 7.08 ± 0.33

Total polyphenols (mg / kg ± E.S.): 670.42 ± 43.73

Total chlorophyll (ppm ± E.S.): 16.55 ± 1.90.

In the following ***Figure 4*** is reported the chemical and organoleptic composition of the cultiva rCoroncina.

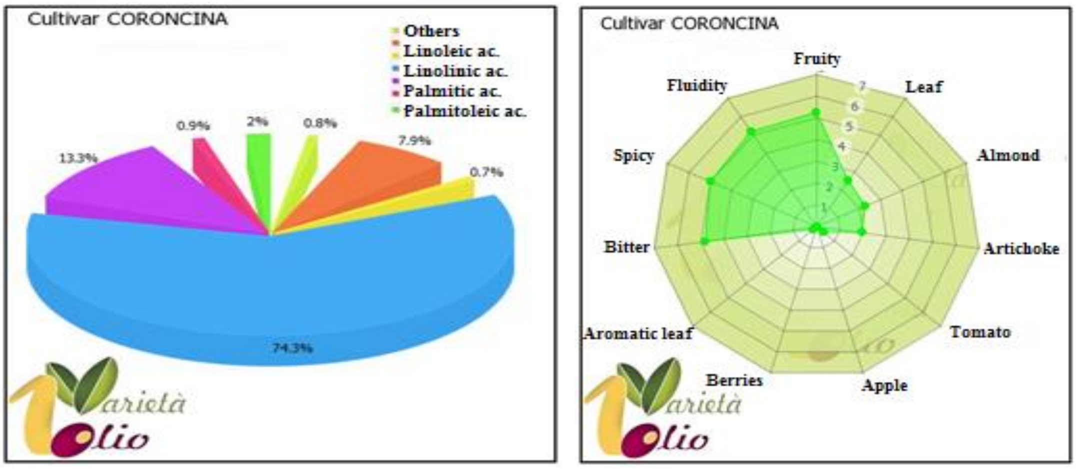


Figure 4 Chemical and Organoleptic composition of the Cultivar Coroncina

ASCOLANA TENERA

The origins of the Ascolana Tenera, hereafter named Ascolana, are very ancient, and it was especially appreciated by the Romans who called it "picena", and they imported it through Salaria. Sometimes elongated elliptical shape has a weight varying from 4 to 8 grams, with juicy and tender flesh (hence its name), with original (unique and particular) flavor and representing 88-92% by weight of ‘drupe’. The diffusion is in Province of Ascoli Piceno. Good resistance to the cold and the pathologies of the olive, except the oleaginous fly. The yield in oil is poor and produces a herb of gentle herbaceous fruity with delicate notes of bitter and spicy that make it as a whole a harmonious product. Oils characterized by a medium-intense level of fruity, from an average level of bitter and sparkling, with prevailing herbs / leaf and tomato scents and reading sensations of artichoke and fresh almond.

The organoleptic characteristics and the main chemical characteristics are described in ***Figure 5***.

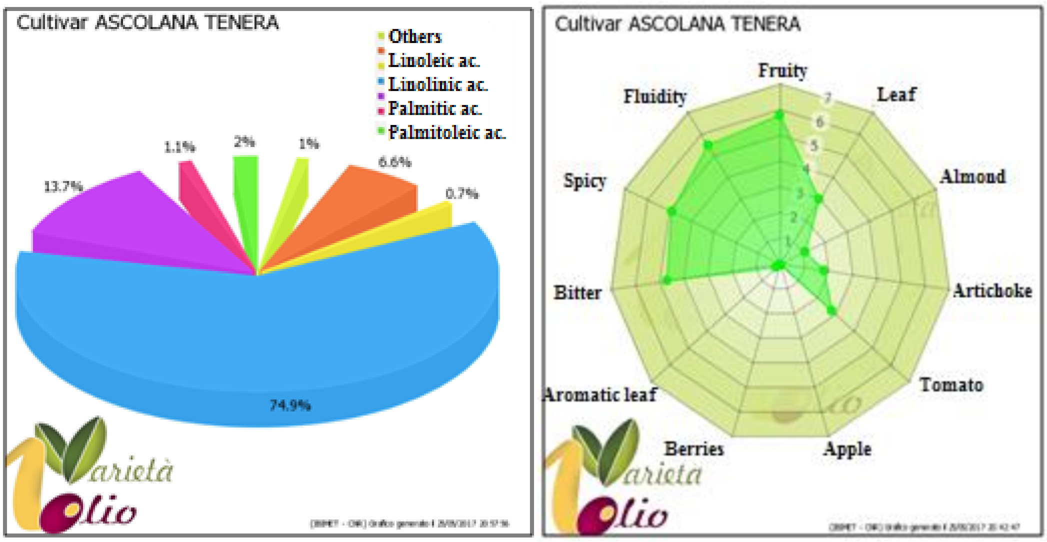


Figure 5 Chemical and Organoleptic composition of the Cultivar Ascolana

CHARACTERISTICS OF OLIVE OIL

CHEMICAL AND PHYSICAL CHARACTERISTICS

The characterization of olive oil is based on:

- Organoleptic characteristics
- Physical and chemical characteristics

A panel Test determine the organoleptic characteristics, they concern the aroma, the smell and the taste of the oil. Physical and chemical characteristics allow to classify oils on the bases of various parameters measured by chemical-physical methods. The analyses made on the Oil samples have three main goals:

- Tests for ensuring the quality of the oil, they are: acidity, peroxides, spectrophotometric analysis in ultraviolet, fatty acid composition, sterol composition, the content of halogenated solvents, and Gas chromatographic analysis;
- Tests to ensure the preservation of the oil: number of peroxides, spectrophotometric analysis in ultraviolet, acidity, panel test, time induction;
- Tests to ensure the genuineness of oil: acid composition, sterol, spectrophotometric analysis in ultraviolet, solvent analysis halogenated, difference ECN 42 HPLC and ECN 42 theoretical calculation

The analytical parameters determined on the oil and the techniques used are described in the EEC Regulation no. 2568/91 and subsequent amendments[XI].

The olive oil is a fat that at room temperature (20° C) is a liquid with density of about 916 g/liter. It consists of 98-99% of a mixture of triglycerides, known as the "saponifiable" fraction, and the remaining part from a set of compounds which represent the "unsaponifiable" fraction, consisting of chemicals that comprise several classes such as, for example, aliphatic and triterpenic alcohols, sterols, hydrocarbons, volatile compounds and antioxidants (Figure 6).

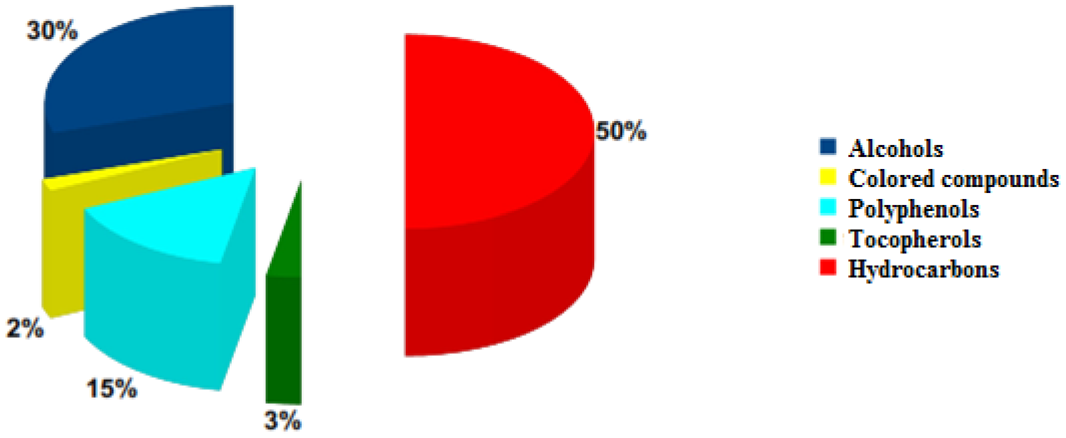


Figure 6 Main constituents of the unsaponifiable fraction of olive oil

SAPONIFIABLE FRACTION

Triglycerides are part of the saponificable fraction, these are molecules derived from the natural esterification of three fatty acid molecules with a glycerol molecule. These, treated at warm with a base (NaOH or KOH), react by forming the soaps, alkali salts of fatty acids, according to the reaction shown in ***Figure 7***.

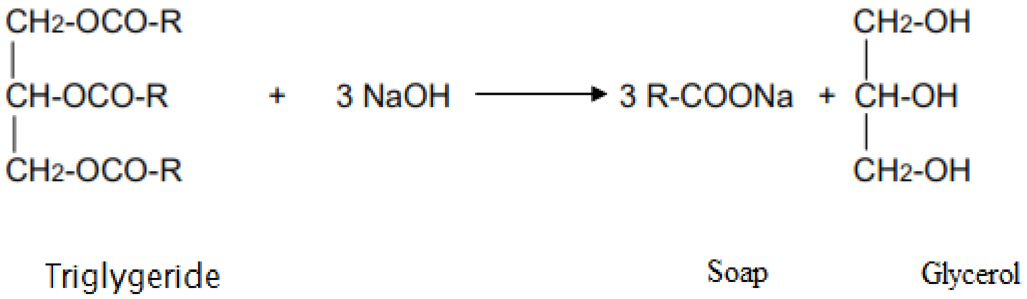


Figure 7 Saponification reaction of a triglyceride

Glycerides differ each other because of the nature and number of fatty acids linked to the glycerine frame. Fatty acids, in turn, are classified in saturated and unsaturated fatty acids, according to whether or not there are double bonds in their structure; they can be mono or polyunsaturated depending on the number of double bonds. Saturated fatty acids are solid, at room temperature, while those unsaturated are liquid.

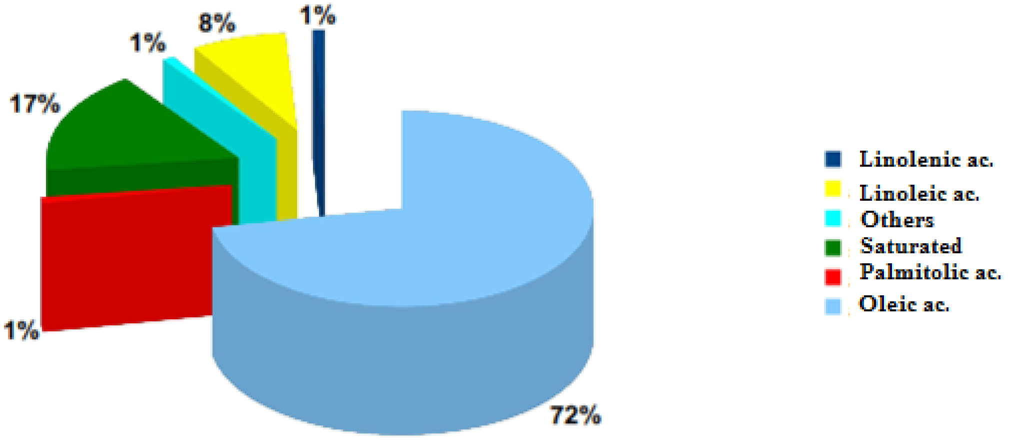


Figure 8 shows the mean distribution, in percentage, of the fatty acids present in the virgin olive oils.

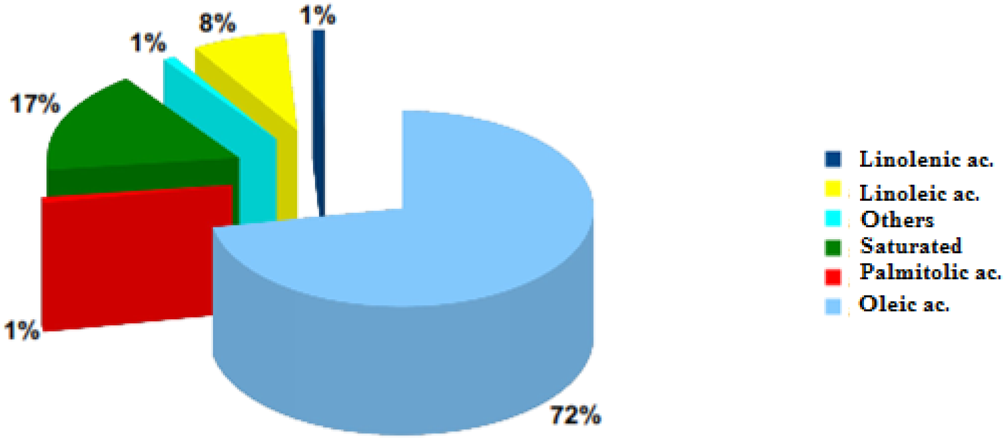


Figure 8 Distribution of saturated and unsaturated fatty acids in olive oils

looking to the fatty acids composition of the virgin olive oils we note that only 16-17% of them are saturated fatty acids such as palmitic and stearic acid. The monounsaturated fraction is the most abundant, about 70-80% of the total, mainly due to oleic and, to a lesser extent, to palmitolic acid (***Figure 9***). Olive oil contains also significant amounts of polyunsaturated fatty acids such as linoleic and linolenic acid that are "essential fatty acids", that is, fatty acids indispensable for the growth and functioning of the human tissues.

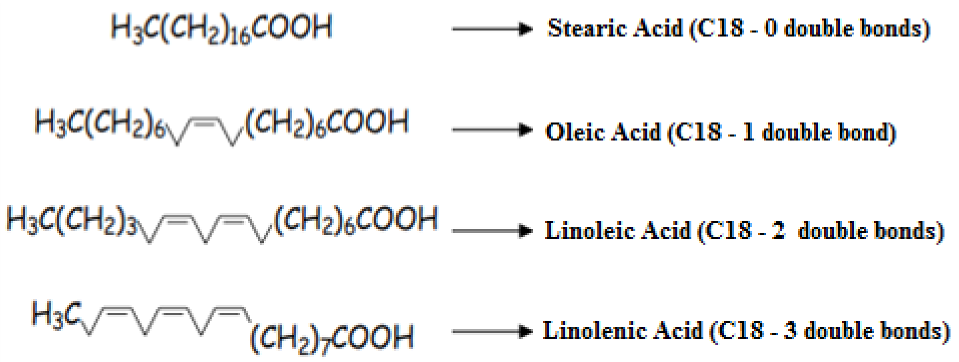

Stearic Acid (C18 - 0 double bonds)
Oleic Acid (C18 - 1 double bond)
Linoleic Acid (C18 - 2 double bonds)
Linolenic Acid (C18 - 3 double bonds)

Figure 9 Essential fatty acids

The presence of unsaturated bonds in fatty acids gives the oils particular biological qualities, but it also makes them subject to oxidative process, initiated or accelerated by several causes: light, heat, radiation, prolonged exposure to oxygen of the air. ***Figure 10*** shows the scheme of a radical reaction activated by light.

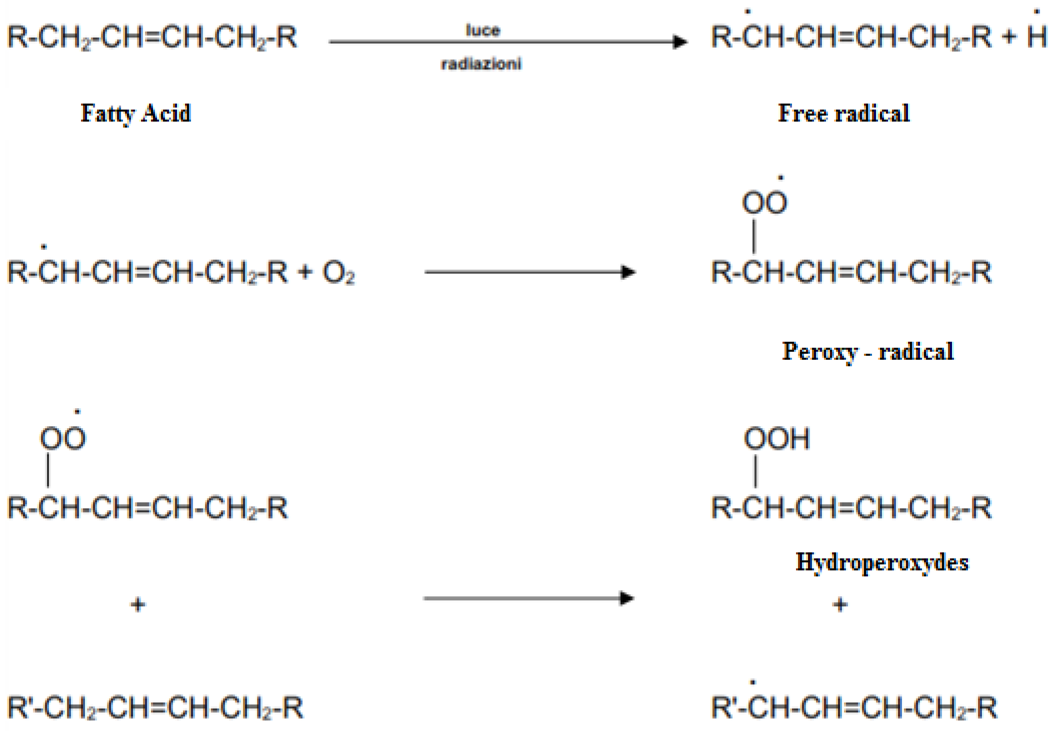


Figure 10 Unsaturated fatty acid irradiation process

This phenomenon proceeds with a speed proportional to the number of existing double bonds, but is opposed to the presence of antioxidant substances; these substances allow the oil to have some stability against rancidity[XII] .

UNSAPONIFIABLE FRACTION

The unsaponifiable fraction is about 1-2% of the olive oil. In spite of the minor amount, it is of great importance because the molecules that make up it are responsible for several interesting features:

Organoleptic properties (scents, odours and flavours) and biological properties (Useful antioxidant capacity for oil preservation).

Among the substances most present in the unsaponifiable fraction there are:

- Hydrocarbons including squalene that is the main compound of the unsaponifiable fraction
- Sterols, present in considerable quantities
- Alcohols, of which in very small amounts aliphatic alcohols and in larger quantities triterpenic alcohols
- Colour pigments such as carotenoids and chlorophyll
- Liposolubile vitamine, provitamine A, vitamine C, D, E and F
- Polyphenols, in the form of glycosides and esters
- Other more or less volatile compounds: terpenes, aldehydes, esters, aliphatic ketones that affect the aromatic oil note and are involved in his sensory evaluation.

Hydrocarbons: are about 50% of the unsaponifiable fraction. The main compound of this class is the squalene an unsaturated hydrocarbon (***Figure 11***) that affects the nutritional properties of oils.

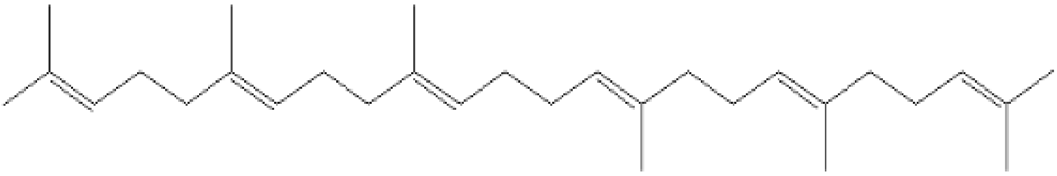


Figure 11 Squalene structure

Phenolic compounds: mainly represented by glycosides and esters, Phenolic compounds of olive oil are antioxidant substances useful to prevent the rancidity of oils over time because they undergo the oxidation process in place of the fatty acids. The amount of phenolic compounds depends on the cultivar (Carotene, for example, is the most rich one) and from the ripening grade at harvest time. Their presence is perceived by bitter and spicy taste and by fruity smell. Polyphenols undergo, especially during oil extraction processes, enzymatic attacks that cause a splitting of polyphenols (Figure 12) in simpler molecules: the phenols.

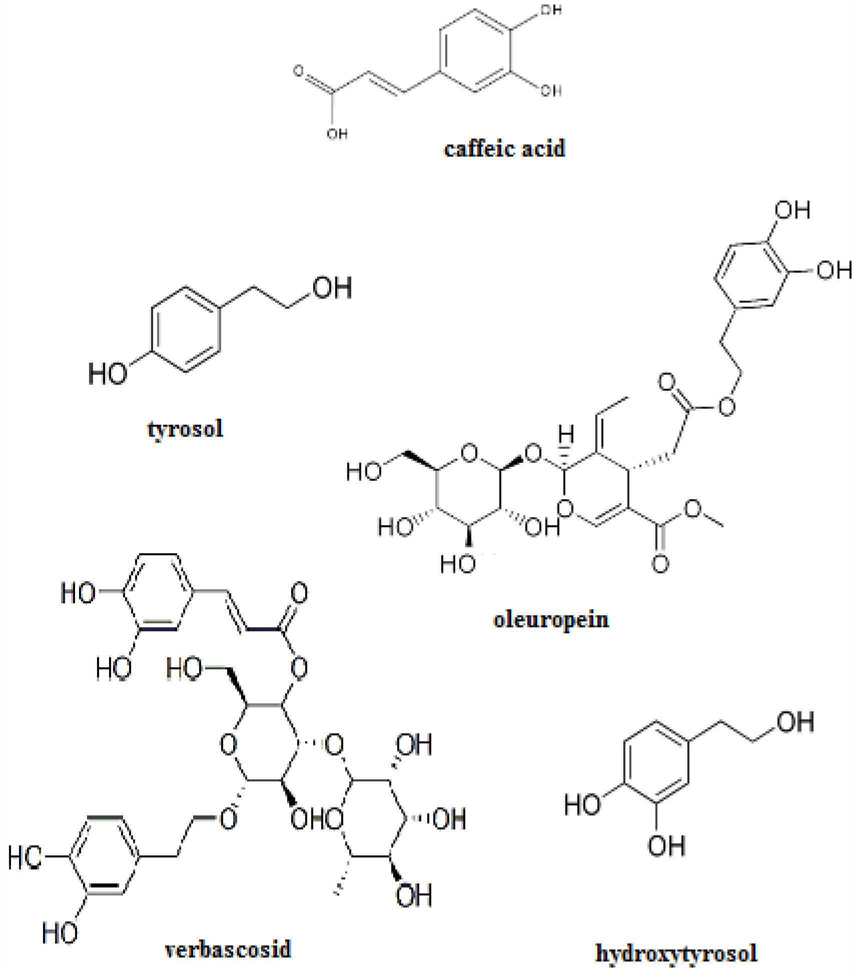
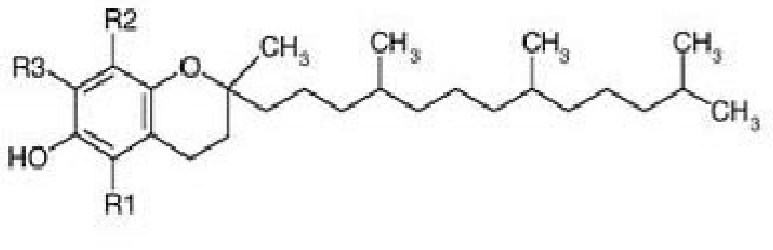


Figure 12 Examples of polyphenols present in olive oils

Tocopherols (***Figure 13***) (molecules of the group of Vitamin E): They are present in α, β, γ and δ form, although it actually exists a clear prevalence, 98%, of the α form that is the biologically most active. These substances, as phenolic compounds, have antioxidant activity, useful both as a food preservative to prevent oils from going rancid and to protect the cell from oxidative stress.

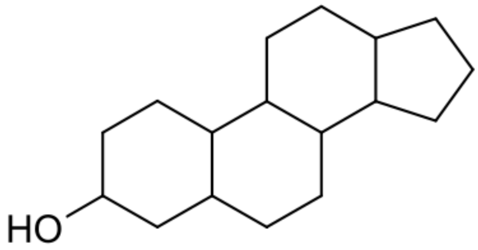


$R_1 = R_2 = R_3 = CH_3$

Figure 13 Vitamin E (Tocopherol Form α)

Sterols: are alcohols or esters of steroids, an important and vast group of compounds present in animal and plant tissues. These are substances biologically very important: just think about the physiological activity of the cholesterol, the best known among sterols, which among other things is also the forerunner of many steroid hormones. The percentage composition of sterolic fraction, in oleaginous plants, depends on the species. Olive oil contains several sterols, the major ones are: camp sterol, stigma sterol and β-sitesterol that is the most abundant (***Figure 14***).

General structure of sterols



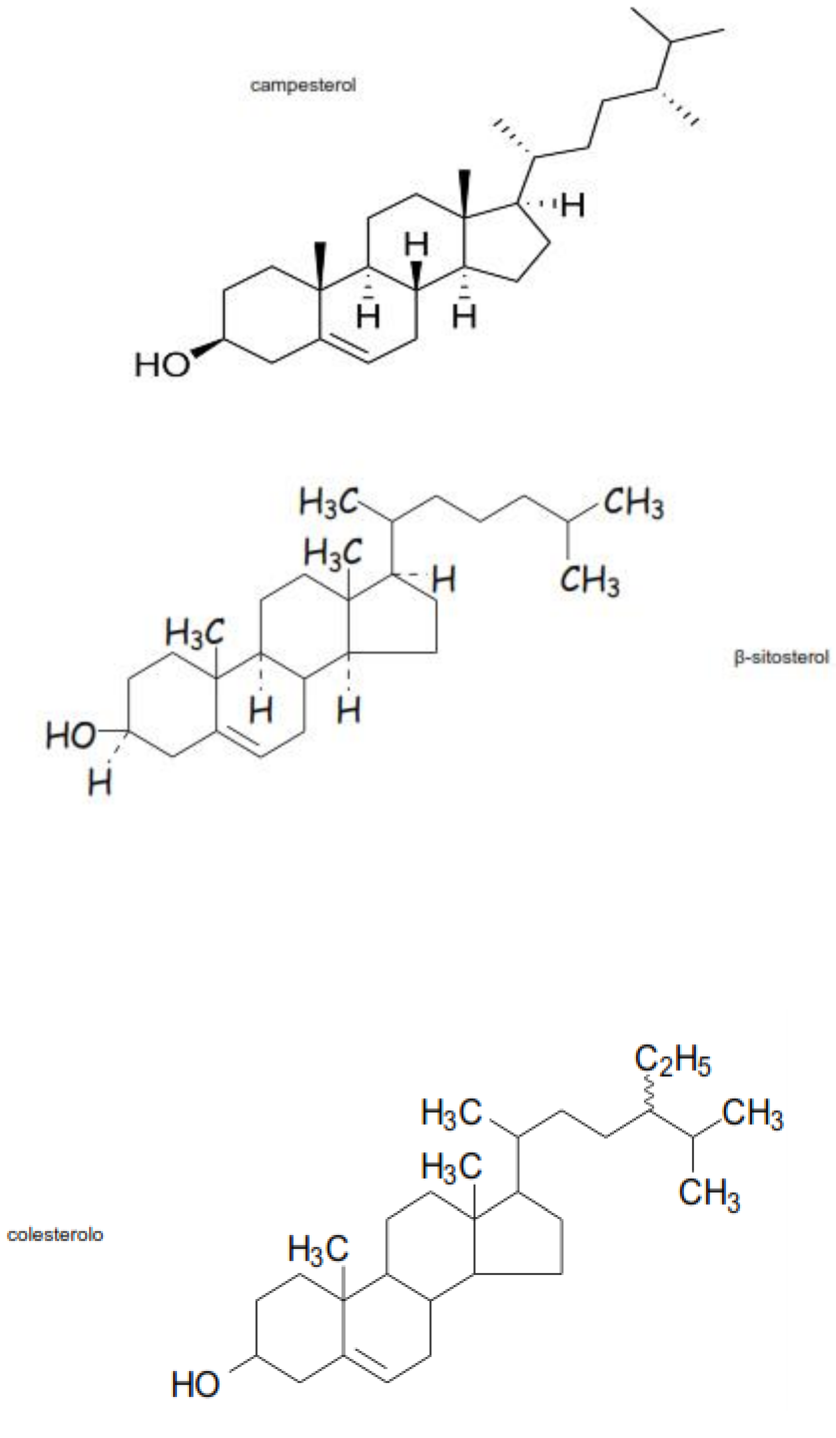


Figure 14 Examples of sterols present in olive oils

Coloured pigments: are the carotenoids and chlorophylls, responsible for the colouring characteristics of olive oil. Chlorophylls a and b (Figure 15) give the oils just extracted an intense green colour and their content may vary in relation to the cultivar and the aging stage of the olives. The coloured pigments, in the presence of light, act on oil as pro-oxidants, while in the dark, in synergy with the phenols, they protect it from oxidation.

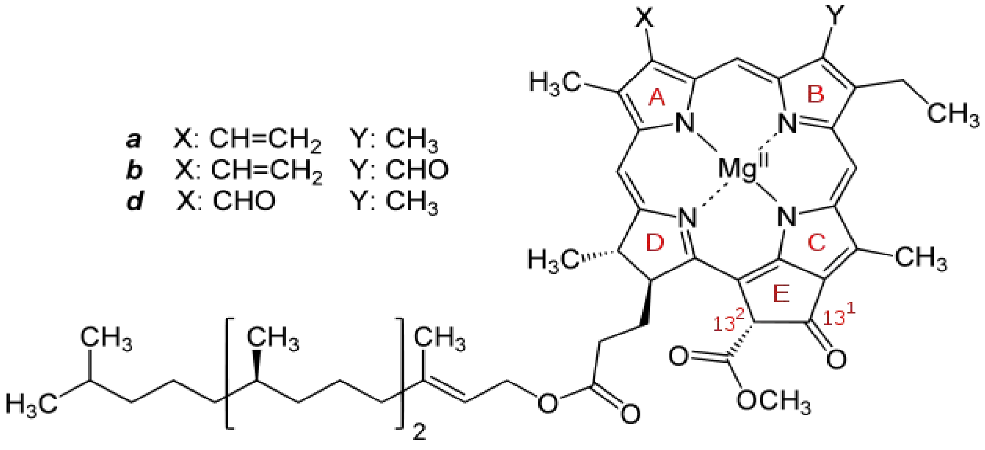


Figure 15 Chlorophyll structure a, b, d

Alcohols: Aliphatic and triterpenic alcohols are identified. Those aliphatic are quite volatile molecules that, at low temperatures, contribute to characterizing the smell of an oil. Other compounds such as, aldehydes, esters, aliphatic and aromatic ketones affect the aromatic oil note and are involved in its sensorial evaluation. The notes found, in olive oil, during tasting are due to volatile active substances present at more or less high concentrations. These generally are highly volatile compounds of small dimensions, generally with 5 (C5) or 6 (C6) carbon atoms skeleton, having high vapour pressure, so that they tend to go in vapour phase: this is why they easily contact the olfactory cells soliciting the odorous feeling. These compounds contributing to the smell of the oil are responsible of the pleasant aroma but also of stink indicating the presence of defects^{XIII}.

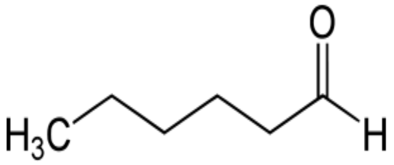


Figure 16 Hexanal

Figure 17 shows some of the numerous molecules responsible of the oil aroma.

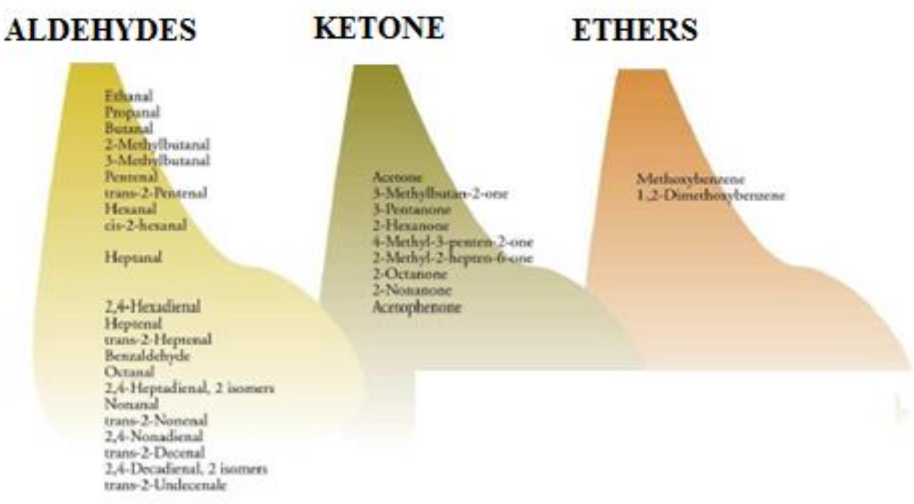


Figure 17 Major components responsible for the aroma of oil

Most aromatic compounds form during grinding of olives by enzymes released due to tissue breaks. The C5 and C6 compounds, especially the linear unsaturated C6 aldehydes, are the most important volatile components. There are numerous routes involved in the formation of such molecules, in particular the C6 and C5 compounds are obtained starting from polyunsaturated fatty acids via the lipooxygenase pathway (LOX), (Figure 18). The final concentration depends on the quantity and activity of each enzyme involved in the path of LOX. In a first step occurs of 9- and 13-hydroperoxide of linoleic and linoleic acids. Such products undergo then splitting by an extremely specific hydro-peroxide that which leads to C6 aldehydes; At this level species can isomerize from form (Z)-3 to the most stable (E)-2. Aldehydes can be reduced by an alcohol-dehydrogenase to the corresponding alcohol which, thanks to the action of acetal-transferase alcohol. The volatile compounds formed by chemical oxidation are the principal responsible for the negative notes in the aroma of oils, like the rancid smell[XIV].

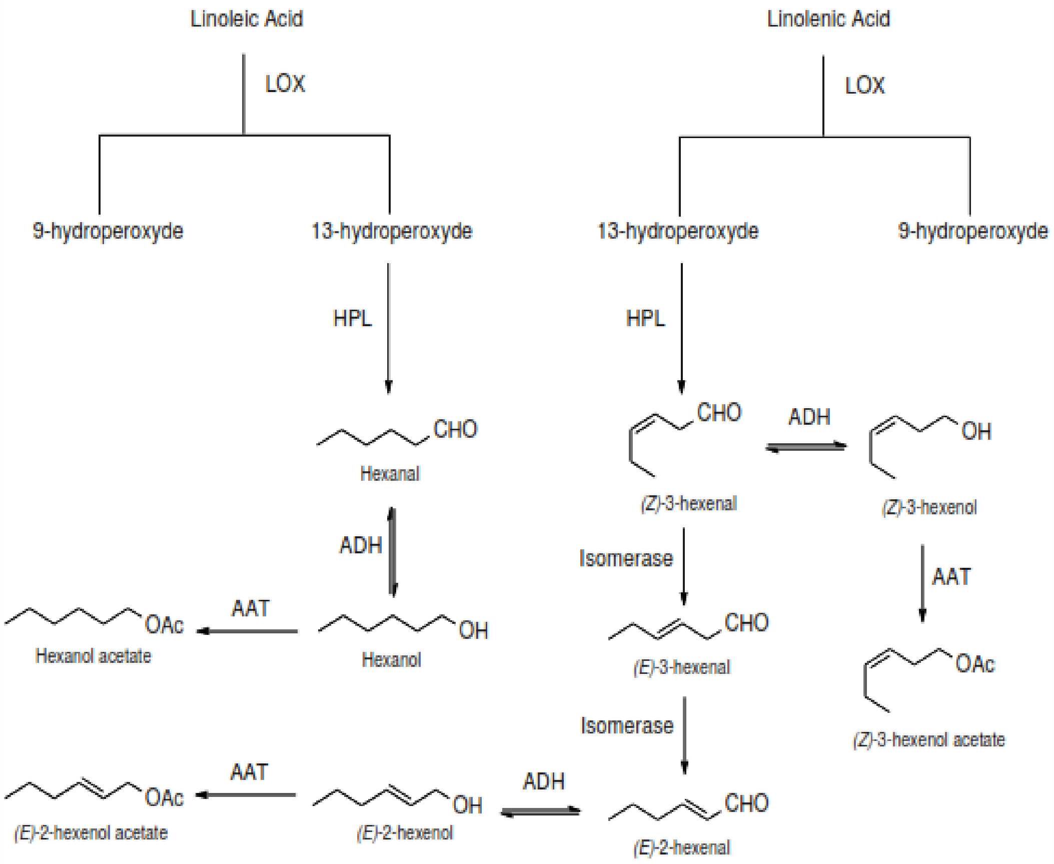


Figure 18 Biogenesis via the pathway of lipoxygenase (LOX: lipoxygenase; HPL hydroperoxide lyase; ADH: Alcohol dehydrogenases; AAT: Alcohol Acetyltransferase).

PURPOSE AND DEVELOPMENT OF WORK

The survival of the national agricultural system is increasingly linked to the typical nature of the products. The actors in our agro-industrial system are losing the International competition with producers that have broader agricultural extensions and lower labour costs. However, the Italian production system is still rich in typical high quality products, which are also highly demanded in the most advanced markets. These products, with high added value, suffer commercial frauds including counterfeiting.

It is therefore important, for proper protection, of our specialties, the development of methods that, in an economic and easy way, allow to characterize the products. The origin of products is, as a rule, a guarantee of the specificity and quality of the food product.

In this context, the aims of the present study were to find a chemometric data treatment of spectroscopic measurements that permits the characterization of monovarietal extra-virgin olive oil from Marche. We devote our efforts to the recognition of five cultivars characteristic of Marche production: Ascolana, Coroncina , Mignola, Mogliano and Raggia.

WORKING HYPOTHESIS AND ANALYTICAL PROCEDURE

The data processing techniques have been designed to discriminate the varietal origin of extra-virgin olive oil. The narrowness of the area subjected to sampling for each variety should guarantee the homogeneity of the chemical/organoleptic characteristics of samples.

There are two analytical methods[XV] for trying to characterize an oil; one uses specific analytical methods to quantify each single substance[XVI], in this case the choice of analytical parameters depends on the purpose of the analysis[XVII]. Several works apply this procedure as for example measuring acidy, fatty acids, content of the different kind of compounds[XVIII] (polyphenols, sterols[XIX], chlorophylls, carotenoids, ...).

A totally different method is one that uses fast analytical-instrumental methods[XX], which provide an aspecific response related to many of the substances that define the characteristics of oil[XXI].The readings, in this type of analysis, give no direct information but treating them with appropriate chemometric methods it is possible to disclose the information of interest. Chemometric methods are statistical techniques that need a sufficient number of samples to represent the overall product variability.

The experimental procedures have to be able to "neutralize" effects due to factors that are not being investigated. For example, in our case, we want to identify the origin of the product, so we should consider a sampling procedure that randomizes the effect of the climatic variability so that it confounds with noise. Samples can come from a single agricultural year if seasonal effects have to be neglected, or we have to sample several years to study the effect of seasons.

In this study the choice falls on non-specific instrumental techniques[XXII], non-invasive, not destructive and easy to apply even on a large scale, capable of providing complete and global information about the sample being examined. An increasing number of works mention spectroscopic analysis, such as UV-VIS[XXIII], IR[XXIV], to satisfy these requirements. Our study aims to develop a method for characterization of monovarietal extra-virgin olive oil, produced by the cultivars Ascolana, Coroncina, Mignola, Piantone di Mogliano and Raggia. To achieve this goal, non-specific measurements have been carried out, and particularly in my case: UV-VIS and IR spectra.

The UV-VIS and IR spectra were chemometrically investigated to respond to our needs. As each spectrum has a large number of points the first step was applying a method for reducing them. Several pre-treatment were investigated in order to find the one that combined with PCA and other methods enable us to reduce the dimensionality of the data and discriminate among categories (varieties of monovarietal extra-virgin olive oils). We applied several techniques of modelling and classification: PLS-DA, SIMCA, UNEQ, in order to find an analytical-chemometric model capable to classify each sample. It was subsequently proceeded to a selection of variables with SELECT, and then it has been verified the accuracy of the reduced model.

We have always adopted validation methods to estimate both the classification and prediction ability of the methods mentioned above.

Partial least square (PLS) calibration was used, with samples of year 2015, to study the correlation between the spectroscopic measurements (FTIR) and volatiles, phenols and sensorial analytical parameters.

The total energy, E_{mol}, of a molecule can be described as a sum of the external and internal energies of the molecule, including translational energy, the energy for the electronic transition, vibrational motions and rotational motions, i.e.,

$$E_{mol} = E_{trans} + E_{ele} + E_{vib} + E_{rot} \qquad \textbf{Equation 1}$$

where:

E_{Trans}: the translational energy of a molecule, ignored in this discussion because it is essentially not quantized;

E_{Ele}: energy for electronic transition of atoms in a molecule;

E_{Vib}: energy for vibration between two atoms in the molecule;

E_{Rot}: energy for molecular rotation.

Figure 19 shows a scheme of these energies.

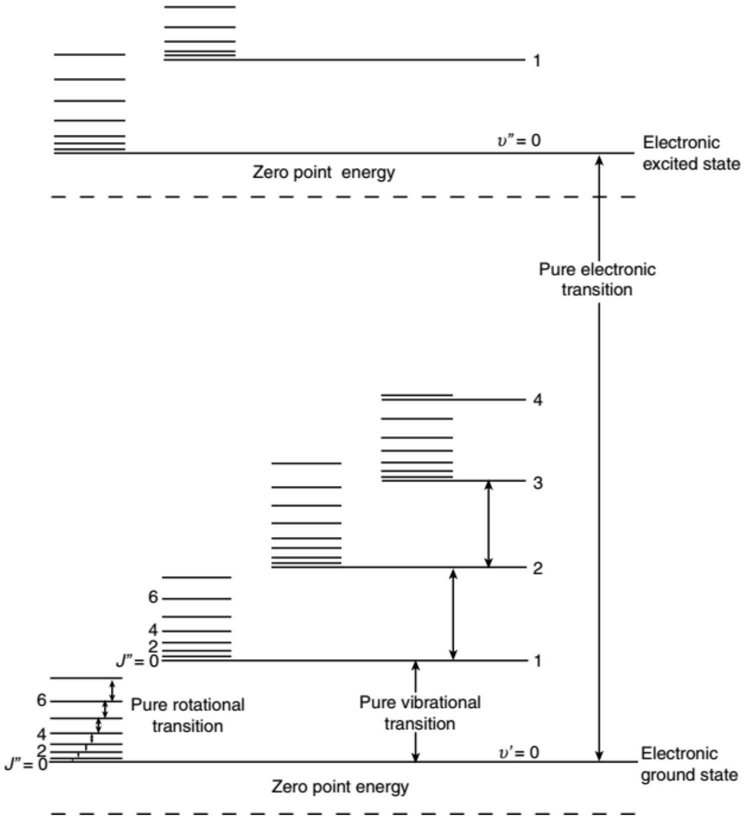


Figure 19: Schematic view of the energies of a molecule

As the electric field of an incident light can interact with the electric field of a chemical bond of a molecule during vibration, the molecule can thus absorb the photon and be brought to a higher vibrational state, i.e., the vibrational excited state.

INFRARED SPECTROSCOPY

Vibrational spectroscopy based methodologies have been widely recognized by the several advantages they offer: they are non-destructive and environmental friendly techniques, requiring minimal or no sample preparation, enabling the estimation of several properties from a single measurement in a short period of time (few seconds).

Infrared Spectroscopy is based on the interaction of the electromagnetic radiation with matter. The bonds in a molecule can be considered as oscillators with quantized energy levels. Generally, stronger bonds and light atoms will vibrate at higher stretching frequency (wavenumber). Two nuclei bonded covalently vibrate similarly to two balls attached by a spring. Double and triple bonds might be thought of as stronger springs.

Every bond can be approximated to a harmonic oscillator whose frequency is:

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$$

Where k is strength constant and *m* the reduced mass of the atoms giving the bond.

$$m = \frac{m_1 m_2}{m_1 + m_2}$$

However the energy levels permitted are quantized so that the molecule emits or absorbs only when the energy associated to the frequency correspond to the permitted ones on the base of the following

$E_{vib} = \left(n + \frac{1}{2}\right)h\nu$ that became $E_{vib} = \left(n + \frac{1}{2}\right)\frac{h}{2\pi}\sqrt{\frac{k}{m}}$

Where *n* is an integer number varying as 0, 1, 2, … and *h* is the Planck constant.

This means that are permitted only those transitions whose frequency are compatible with the chemical nature of the oscillator (*k*, *m*,..) so that $\Delta n = \pm 1$: first selection rule. The theory based on an ideal harmonic oscillator produce equally spaced quantized energy states however, the picture of the harmonic oscillator cannot be retained at larger amplitudes of vibration owing to:

- Repulsive forces between the vibrating atoms.
- The possibility of dissociation when the vibrating bond is strongly extended.

Accordingly, the allowed energy levels for an anharmonic oscillator have to be modified as shown in ***Figure 20***.

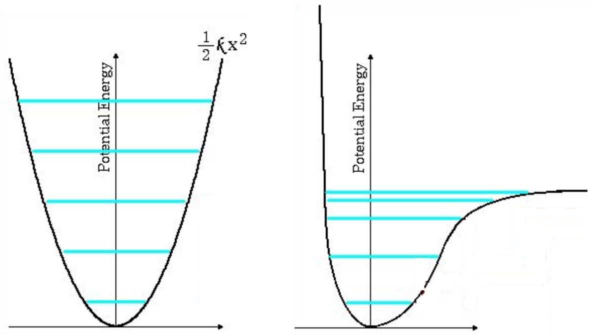


Figure 20: potential energy curves of an harmonic (left) oscillator and an anharmonic (right) one

Unlike the harmonic oscillator, energy levels are no longer equidistant and the strict selection rule $\Delta n = \pm 1$ is expanded to transitions over more than one energy level so that $\Delta n = \pm 1, \pm 2, \pm 3, \ldots$ are now also allowed and are called first, second, and so on, overtones.

Interaction of infrared radiation with a vibrating molecule, however, is only possible if the electric vector of the radiation oscillates with the same frequency as the molecular dipole moment. Thus, a vibration is infrared active only if the molecular dipole moment is modulated by the vibration. The requirement of a dipole moment change during the vibration makes MIR spectroscopy specifically sensitive to polar functionalities.

No net change in dipole moment occurs during the vibration or rotation of "homonuclear species", such as O_2, N_2 or $Cl_{2,}$ because of this reason, these molecules will not absorb IR radiation. They are IR-inactive whereas that of a heteropolar diatomic molecule is IR-active.

Molecular vibrations can be described into two categories:

1. Stretching, change bond lengths
2. Bending, change the bonding angles between the atoms

This permit to a same bond to absorb energy at more than one frequency depending on whether the bond is experiencing a stretching or a bending mode of vibration.

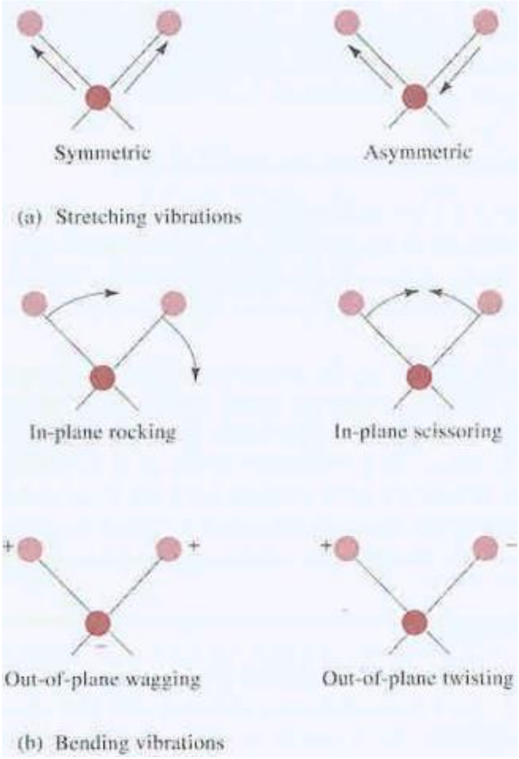


Figure 21: types of molecular vibration

Table 1 shows the spectral ranges of interest for the IR methods. The MIR spectrum results from the fundamental stretching, bending, and rotating vibrations employed for the elucidation of molecular structure, while the NIR spectroscopy measures the broad overtone and combination bands of some of the fundamental vibrations (only the higher frequency modes) and is an excellent technique for rapid, accurate quantitative determination.

Table 1: Spectral regions of interest in the IR spectroscopies.

techniques	Wavenumber (cm^{-1})	Wavelength (µm)	Wavelength (nm)
near-infrared (NIR)	14000 – 4000	0.71 – 2.5	714.29 – 2500
mid-infrared (MIR)	4000 – 400	2.5 – 25	2500 – 25000
far-infrared or Terahertz	400 – 10	25 – 1000	25000 – 1000000

The most commonly used is the middle infrared region, since molecules can absorb radiations in this region to induce the vibrational excitation of functional groups. In the infrared region the rotational energy levels are superimposed on the vibrational energy levels giving rise to many closely spaced transitions that are generally not resolved. Applications of near infrared spectroscopy have also been developed.

The infrared spectrum originate from passing infrared light through a sample and measuring the absorption or transmittance of incident light at each frequency. Peaks appear in the spectrum in correspondence to the frequency where active molecular bond absorb radiation. Since all groups have their characteristic vibrational frequencies, information regarding molecular structure can be gained from the spectrum.

Due to the high information content of its spectrum, infrared spectroscopy has been a very common and useful tool for structure elucidation and substance identification.

FACTORS INFLUENCING VIBRATIONAL FREQUENCIES

The vibrational frequency of a bond, being part of a molecule, is significantly affected by the electronic and steric factors of the surroundings, in addition to the bond strength and atomic masses discussed above. When two bond oscillators share a common atom, they seldom behave as individual oscillators where the individual oscillation frequencies are widely different. The mechanical coupling interactions between two oscillators are responsible for these changes. For example, the carbon dioxide molecule, which consists of two C=O bonds with a common carbon atom, has two fundamental stretching vibrations – an asymmetrical and a symmetrical stretching mode. The symmetrical stretching mode produces no change in dipole moment and is IR inactive. Asymmetric stretching mode is IR active and appears at a higher frequency (2350 cm^{-1}) than observed for a carbonyl group in aliphatic ketones (1715 cm). The carbonyl stretching frequency in $RCOCH_3$ (~1720 cm^{-1}) is lower than acid chloride RCOCl (1750-1820 cm^{-1}). This change in frequency of the C=O stretching can be due to (i) difference in mass between CH_3 and Cl (ii) the inductive or mesomeric influence of Cl on the C=O bond (iii) coupling interactions between C=O and C-Cl bonds (iv) change in bond angles arising due to steric factors etc. It is usually impossible to isolate one effect from the other. However, the appropriate emphasis can be placed on those features that seem to be most responsible in explaining the characteristic appearance and position of group frequencies.

INSTRUMENTATION

Most commonly used instruments in infrared spectroscopy are dispersive infrared spectrometer and Fourier transform infrared spectrometer.

Dispersive infrared spectrometer

Dispersive infrared spectrometer is mainly composed of radiation source, monochromator and detector. For mid-infrared region, Globar (silicon carbide), Nernst glower (oxides of zirconium, yttrium and erbium) and metallic helices (chromiumnickel alloy or tungsten) are frequently used as radiation sources. Tungsten-halogen lamps and metallic conductors coated with ceramic are utilized as sources for near-infrared region. A mercury high-pressure lamp is suitable for far-infrared region. Monochromator in conjunction with slits, mirrors and filters separates the wavelengths of light emitted. The dispersive elements within monochromator are prisms or gratings. Gratings have gradually replaced prisms due to their comparatively low cost and good quality. As shown in ***Figure 22***, radiation passes through both a sample and a reference path. Then the beams are directed to a diffraction grating (splitter), which disperses the light into component frequencies and directs each wavelength through a slit to the detector. The detector produces an electrical signal and results in a recorder response.

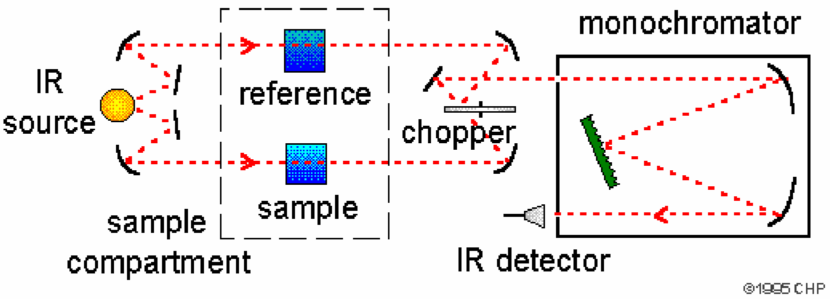


Figure 22 Schematic illustration of dispersive infrared spectrometer.

Two types of detectors are employed in dispersive infrared spectrometer, namely, thermal detectors and photon detectors. Thermal detectors include thermocouples, thermistors, and pneumatic devices, which measure the heating effect generated by infrared radiation. Photon detectors are semiconductor-based. Radiation is able to promote electrons in photon detectors from valence band to conduction band, generating a small current. Photon detectors have faster response and higher sensitivity than do thermal detectors but are more susceptible to thermal noise.

Fourier transform infrared spectrometer

Dispersive infrared spectrometer has many limitations because it examines component frequencies individually, resulting in slow speed and low sensitivity. Fourier transform infrared (FTIR) spectrometer is preferred over dispersive spectrometer, since it is capable of handling all frequencies simultaneously with high throughput, reducing the time required for analysis. The radiation sources used in dispersive infrared spectrometer can also be used in FTIR spectrometer.

In contrast with the monochromator in dispersive spectrometer, FTIR spectrometer as shown in ***Figure 20*** employs an interferometer. The beam splitter within the interferometer splits the incoming infrared beam into two beams, one of which is reflected by a fixed mirror, while the other one reflected by a moving mirror perpendicular to the fixed one. The length of path one beam travels is fixed and that of the other one is changing as the mirror moves, generating an optical path difference between the two beams. After meeting back at the beamsplitter, the two beams recombine, interfere with each other, and yield an interferogram. The interferogram produces inference signal as a function of optical path difference. It is converted to a spectrum of absorbance or transmittance versus wavenumber or frequency by Fourier transform[XXV].

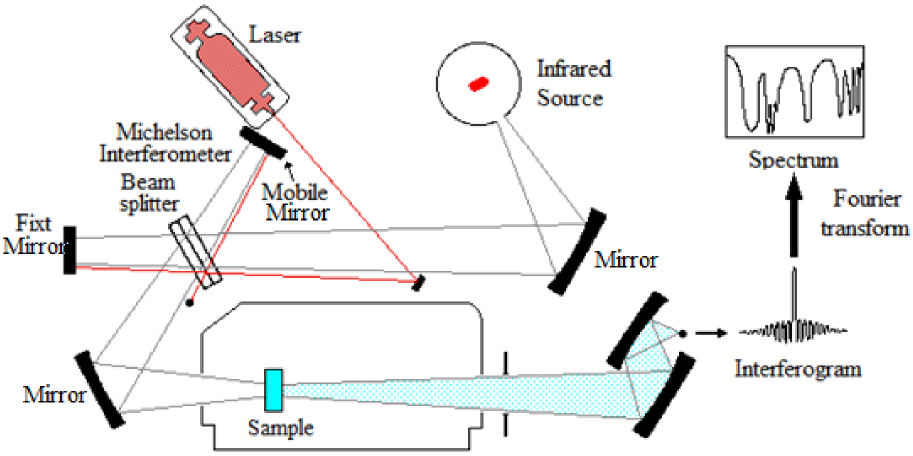


Figure 23 Schematic representation of Fourier transform infrared spectrometer.

Detectors used in FTIR spectrometers are mainly pyroelectric and photoconductive detectors. The former are constructed of crystalline materials (such as deuterated triglycine sulfate) whose electric polarization rely on temperature. The change in temperature leads to change in charge distribution of the detector and electric signal is produced. The latter (such as mercury cadmium telluride) provide better sensitivity and faster speed than do pyroelectric detectors over a broad spectral range. However, liquid nitrogen is needed for cooling of photoconductive detectors.

UV-VIS SPECTROSCOPIC

The portion of the electromagnetic spectrum ranging from the near ultraviolet to the very near infrared, between 180 and 1100 nm, is named ‘UV/Visible’ because it includes radiation perceptible to the human eye. Vacuum or far ultraviolet, operating in a vacuum, can attain 150 nm with samples in the gaseous state. UV/Visible spectroscopy, generally, yields little structural information but is very useful for quantitative measurements. This region of the spectrum is conventionally divided (***Figure 25***) into three sub-domains termed near UV (185-400 nm), visible (400-700 nm) and very near infrared (700-1100 nm)[XXVI].

When a molecule absorbs a photon from the UV/Vis region, the corresponding energy is captured by one (or several) of its outermost electrons. As a consequence there occurs a modification of its electronic energy E_{ele}, a component of the total energy of the molecule (see Equation 1).

A modification of E_{ele} will result in alterations for both E_{rot} and E_{vib} resulting in a vast collection of possible transitions obtained in all three cases (***Figure 19***).

The electronic transitions observed in organic compounds involve electrons engaged in σ or π or non-bonding *n* electron orbitals of light atoms such as H, C, N, O). The energy necessary to excite σ electrons to σ* orbitals is greater than that available in the UV region, because of this alkanes and other saturated compounds do not absorb UV radiation. UV radiation, however, is sufficient to excite electrons in π orbitals to π* antibonding orbitals or *n* electrons to π* or σ* antibonding orbitals.

Both organic and inorganic molecules may exhibit absorption and emission of UV/ VIS radiation. Molecular groups that absorb visible or UV light are called **chromophores**.

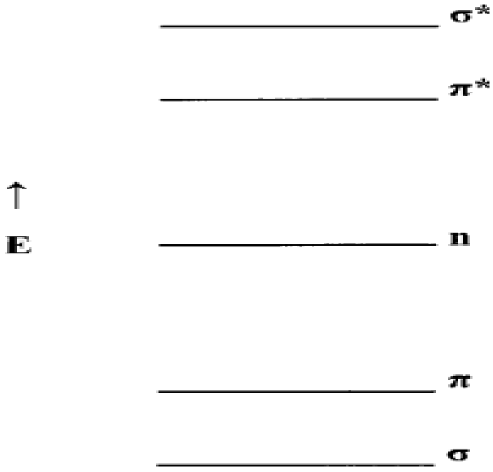


Figure 24 Energy dispersion of available orbital electrons in a generic molecule. At the fundamental state, they will occupy the lowest available energy states and hence all available orbital linkages σ, π, and non-binding.

For example, for a π→π* transition to occur, a molecule must possess a chromophore with an unsaturated bond, such as C═C, C═O, C═N, and so on. Compounds with these types of chromophores include alkenes, amides, ketones, carboxylic acids, and oximes, among others. The other transition that commonly occurs in the UV/VIS region is the *n*→π* transition, so organic molecules that contain atoms with nonbonded electrons should be able to absorb UV/VIS radiation. Such atoms include nitrogen, oxygen, sulfur, and the halogen atoms.

The UV/Vis spectrum of a molecule consists of broad absorption bands over wide wavelength ranges. The absorption bands are broad because each electronic energy level has multiple vibrational and rotational energy levels associated with it. Excitation from the ground electronic state can occur to more than one vibrational level and to more than one rotational level. Even though each separate transition is quantized, the close energy spacing of the vibrational levels and the even more closely spaced rotational sublevels cause the electronic transition to appear as a broad band.

The absorption regions of different electronic transitions, involving bonding and "lone-pair" electrons in an organic molecule, are summarized in ***Figure 25***.

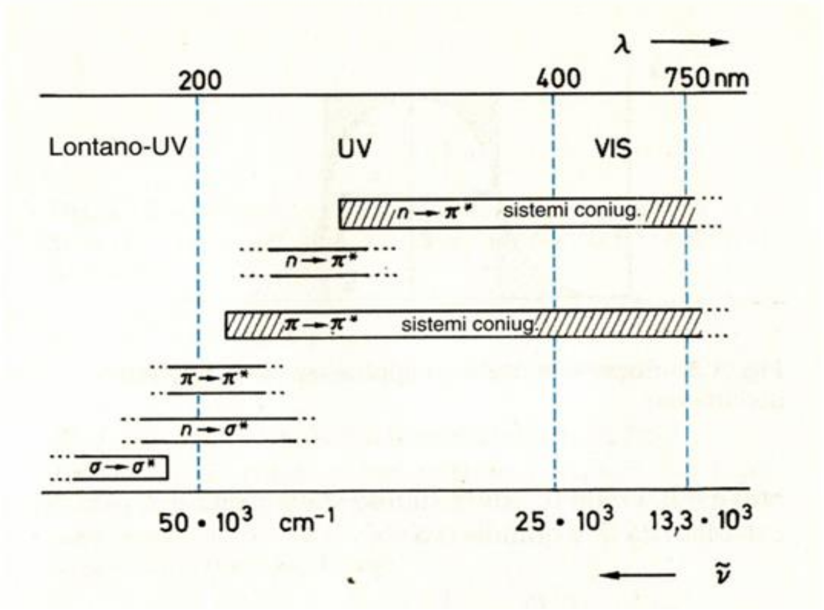


Figure 25 Absorption wavelengths of the different electronic transitions.

The UV/Vis domain has been widely exploited in quantitative analysis. The measurements are based upon the Lambert‑Beer law which, under certain conditions, links the absorption of the light to the concentration of a compound in solution. This is stated as:

$A = \varepsilon_{\lambda} l C$ ***Equation 2***

where *A* stands for absorbance, a dimensionless optical parameter measurable with a spectrophotometer, *l* is the thickness (in cm) of the solution passed through by the incident light, I_0. *C* is the molar concentration and ε_{λ} the molar absorption coefficient (or molar absorptivity) $\mathrm{Lmol^{-1}cm^{-1}}$ at *wavelength λ*, at which the measurement is made. The molar absorptivity, is characteristic of the compound being analysed and depends among other things upon the temperature and the nature of the solvent.

According to Lambert's hypothesis, the intensity I of *monochromatic radiation* decreases by dI passing through a thickness dx of a material whose *absorption coefficient* is k for the wavelength λ chosen, being:

$-\frac{dI}{dx} = kI$ chosen λ

Beer proposed, for a low concentration solution dissolved in a transparent medium (non-absorbing), that k is proportional to the molar concentration C of this compound: $k = k^{*}C$

If an incident radiation I_0 pass through a medium of thickness *l* whose absorption coefficient is k, the emerging light will be

$-\frac{dI}{I} = k^{*}Cdx$ integrating we obtain the transmitted intensity I, in fact $\int_{I_0}^{I}\left(-\frac{dI}{I}\right) = \int_{0}^{l}\left(k^{*}Cdx\right)$ that is $\ln\left(\frac{I}{I_0}\right) = -k^{*}Cl$ from this we see that the transmitted light is $I = I_0 e^{-k^{*}Cl}$

Absorbance is the common logarithm of the ratio of incident to transmitted radiant power through a material that is

$$A = \log\left(\frac{I_0}{I}\right) = -\varepsilon_{\lambda}Cl$$

Alternatively, the transmittance, indicated by T, can be considered, that is, the fraction of incident light that is passed through the sample in logarithmic term

$$A = \log\left(\frac{I_0}{I}\right) = \log(\frac{1}{T})$$

Modern instruments allow to use an incident polychromatic light so permitting the recording of the entire spectrum of the molecule. The light sources for UV/Vis spectroscopy are:

- for the visible region of the spectrum, an incandescent lamp fitted with a tungsten filament housed in a silica glass;
- for the UV region a deuterium arc lamp working under a slight pressure to maintain an emission continuum (<350nm);
- alternatively, for the entire region 200 to 1100 nm, a xenon arc lamp can be used for routine apparatuses.

The light emerging from the sample is decomposed, generally with a grating, in its components and sent to a set of photodiodes (diode arrays) that reads simultaneously the absorbance at each wavelength. These instruments use only a single light beam, so a reference spectrum need to be previously recorded and stored in memory in order to be used to calculate transmittance or absorbance spectra after recording the sample spectrum.

The information thus obtained is a kind of fingerprint of the analyte that depends on the concentration of the analyte and the chemical-physical characteristics of the molecule.

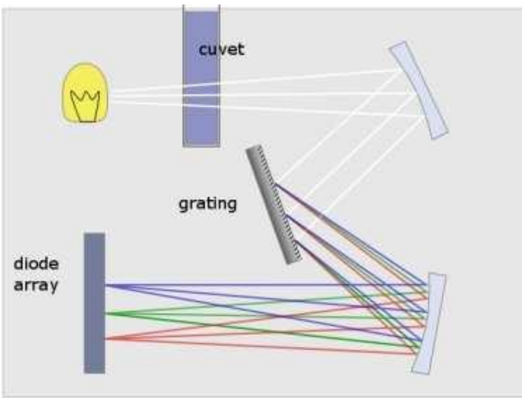


Figure 26 Layout of a Generic UV-VIS Spectrophotometer with Diode Series. The source radiation transmitted from the sample is decayed by a lattice and sent to the detector.

It is useful to remind that the position of a generic absorption peak does not depends solely on the nature of the chromophore, but also from its chemical environment, that is, substituents present on the molecule and solvent.

The environment of the chromophore can cause a shift of its absorption wavelength:

bathochromic shift or red shift, when the shift is to longer wavelengths;

hypsochromic shift or blue shift if the shift is to shorter wavelengths.

The transitions that give rise to UV/VIS absorption by organic molecules are the $n \rightarrow \sigma^*$, $\pi \rightarrow \pi^*$, and $n \rightarrow \pi^*$ transitions.

An **auxochrome** is a substituent that contains unshared (nonbonding) electron pairs, such as OH, NH, and halogens. An auxochrome attached to a chromophore with π electrons shifts the absorption maximum to longer wavelengths.

Even solvents can induce the shift; the reason for the shift in wavelength is related to the energy level of the excited state. Attractive dipole–dipole interactions and hydrogen-bonding affect excited π^* state more than the unexcited π state. Therefore, if a molecule is dissolved in a polar solvent, the energy level of the π^* antibonding orbital will decrease more than the energy level of the π bonding orbital. The energy difference required for the electronic transition in the polar solvent is less than that necessary when the molecule is in a nonpolar solvent. As a consequence, the absorption maximum is shifted to a longer wavelength in a polar solvent.

The energy levels of *n* electrons decrease significantly in a solvent that has the ability to form hydrogen bonds. The result is an increase in the energy difference between the *n* orbital and the *π** orbital. This causes a shift in the absorption maximum of an *n*→π* transition to shorter wavelengths, a blue shift, by as much as 25‑50 nm. The *n*→π* transition also shows a hypsochromic shift as solvent polarity increases even in non-hydrogenbonding solvents. This is due to the increased solvation of the *n* electrons; the energy level of the solvated electrons is lowered.

UV/Vis spectroscopy is simple, reliable and cheap, however, it has limitations on selectivity, due to the overlapping of characteristic peaks that are often observed.

An UV/Vis spectrum can’t be unambiguously assigned to a molecule if it is in a complex mixture. This is because spectra are made of broad bands hiding the fine structure of the absorption peaks. Therefore the absorption bands of different substances can largely overlap making it difficult to assign the contributions of each single compound[XXVII].

The spectrum of a mixture, however, carries an unspecific information on sample composition that can be used with advanced processing techniques.

CHEMIOMETRIC TECHNIQUES

Nowadays analytical instruments permit to acquire a large amount of data in a short time and at a reduced cost. This is an advantage as it allows to acquire a large amount of information in a single step; on the other hand, this abundance of data is difficult to interpret with classical methods. Classical methods of analysis are generally univariate, the knowledge of the behaviour of a system or the quantification of an analyte is based on a chemical/physical characteristic (variable).

An analyte can be quantified, for example, correlating its concentration to a measurable physical property by a physical law that regulates the phenomenon. The measurement of absorbance at a selected wavelength λ after appropriate sample treatment can be used to quantify an analyte using the Lambert-Beer law. The information about the system, in this way, is partially exploited in the sense that it is forced to study its behaviour by considering one characteristic at a time as if it were the only property of the system.

However, analytical systems are generally complex and their characteristics often correlate with each other. The amount of data collected (multivariate) with modern tools (for example, not the absorbance at a single wavelength but the whole spectrum) solicits the use of a survey approach able to take advantage of all the information in them contained.

These data, by their very nature multivariate, need to be elaborated for make the information contained in them readable and interpretable. The processing have to be able to eliminate the majority of duplicate information, correlations between variables (in the example of the spectrum, it is easily conceivable that the absorbance of very close wavelengths are due to the same analyte and the same feature of the system), and retain in synthetic form, the information useful to the interpretation of the phenomenon.

The experimental researcher is accustomed to making the most of our brains' abilities which is particularly efficient in recognizing figures and shapes; This feature comes exploited every time we use a chart to illustrate experimental results. Unfortunately, however, our brain has a high ability when referred to a three-dimensional space, while multivariate data is generally described in n-dimensional spaces (e.g. n = number of wavelengths of the spectrum). The purpose of the work thus becomes the reduction of the dimension of the space where represent the system under examination still containing all useful information.

NOMENCLATURE FOR CHEMOMETRY

Before dealing with the different chemometric methods we define some common terms and procedure.

First, we define what is a Validation procedure. When we calculate a mathematical model from some experimental points, we do it by optimizing the coefficients of the model so to minimize the differences between the estimated values and the corresponding experimental (error of model, residues). This means we fit the mathematical function to the experimental data we have, these data make up the ***training set*** of which we have complete knowledge both of the predictors and of the target quality/property values, all these knowledge are used in calculus of the model. Generally the main goal of estimating a mathematical model is the possibility to use it for predicting unknown value for outer data. The error of prediction estimated by the residues obtained for the training data are, because of the fitting optimization, lower than what we can expect for outer experimental data. We need therefore a method for estimating the prediction error for unknown, to this goal ***validation procedure*** exists.

Prediction by the model need data of which we know the predictor values but not the target quality/property values, these last will be estimated through the model.

To perform a validation procedure we have to have a complete knowledge of the data but the target quality/property values are used only at the end to compare their experimental value to the estimated ones, a comparison that permits to estimate the error of prediction. The data used for this goal are named “***Evaluation set***“ if they are alternatively used in the training set or “***External set***“ when they never are in the training set.

The daily use of the model need data of which we know the predictors values but nothing about the target values; these are “***Test set***” data.

VALIDATION METHODS

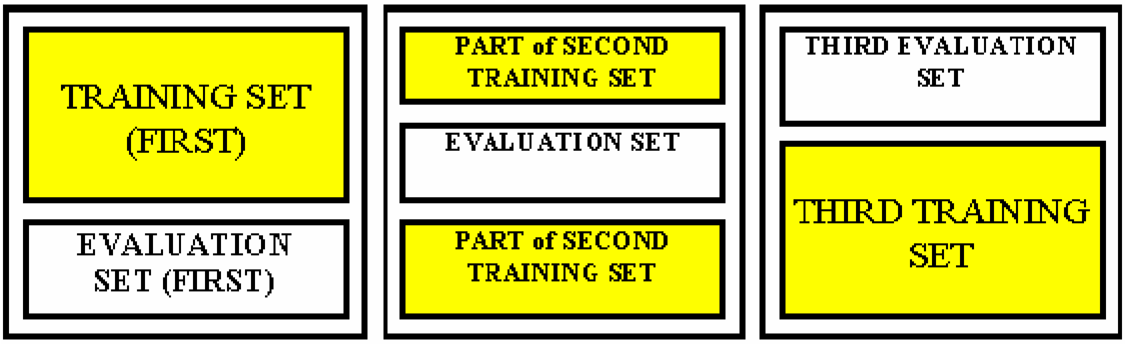
The validation methods differ on the base of the method used to select the data used for the estimation of the prediction error:

Single evaluation set

Make use of the external set to estimate the prediction error.

CROSS-VALIDATION

At the beginning, divide the data in C *cancellation groups*. C is a parameter fixed by the user that need optimization. The algorithm performs C runs every time using one of the cancellation groups as evaluation set and all the other as training set. Every sample is only one time in the evaluation set and C-1 times in the training set. The sum of the residues over all data gives the prediction error estimator. The assignment is made systematically:



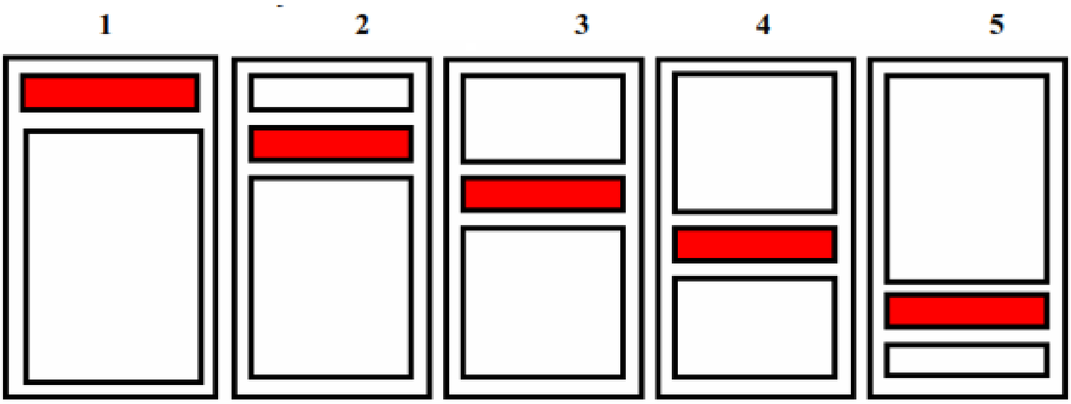
ADVANTAGE: the prediction error estimation is obtained for ALL the OBJECTS in the calibration set.

DRAWBACK: the estimated prediction error (predictive ability) depends on the number of cancellation groups.

LEAVE-ONE-OUT (jackknife)

LEAVE-ONE-OUT is a limiting case of cross-validation procedure. The number C of cross-validation cancellation groups, in LEAVE-ONE-OUT, equals the number of objects so that algorithm cycles as many times as the number of objects. Every time only one sample is in the evaluation set.

In the evaluation set object:



Relatively long time

Too low perturbation of the training set (optimistic)

REPEATED evaluation set or MONTECARLO validation.

The classification rule is computed a lot of times, L. Evaluation sets are created by random extraction, with a selected probability for the assignment. The number of objects in the evaluation sets is not exactly the same. An object can be assigned to the evaluation set many times, generally with different companions.

With very large L, an excellent exploration of prediction. Very long computing time. The validation can be repeated with different percentage of objects in the evaluation sets.

TRANSFORMATIONS

Transformation (often referred as pre-processing) apply to data to reduce or remove effects that do not carry relevant information for the knowledge of the system. The target of spatial and spectral pre-processing is to avoid the influence of undesirable phenomena affecting the measurement, like areas of the scanned surface without interest (e.g. background), light scattering and shadows, particle size effects or detector artefacts. Transformations include the application of derivatives to spectral data to reduce baseline offset, while accentuating small spectral differences. Scattering corrections are often used as transformations to diffuse reflectance spectra to reduce differences such as light scatter and path length. These transforms can only be performed on numerical data. Some of them cannot be performed when there is missing data.

CENTERING OF ROW

$$y_{i,v} = x_{i,v} - \bar{x}_i$$

Eliminates systematic location differences between subjects, for example it may mitigate systemic phenomena such as signal drift; In this case, the variability between objects.

CENTERING OF COLUMN

$$y_{i,v} = x_{i,v} - \bar{x}_v$$

Eliminates systematic location differences between variables by highlighting only variations of the variables in the analysed data group. Transformed variables will have zero average.

AUTOSCALING OF COLUMN

$$y_{i,v} = \frac{x_{i,v} - \bar{x}_v}{s_v}$$

Combining centring and column scaling, it combines the characteristics of the transforms; Eliminates the location effects (offsets) and brings the variables back to the same scale. The transformed variables will have all mean 0 and unit variance.

AUTOSCALING OF ROW (SNV)

Standard Normal Variate is the analog of the column autoscaling in the object space.

$$y_{i,v} \quad = \quad \frac{x_{i,v} - \bar{X}_i}{s_i}$$

It is a transformation usually applied to spectroscopic data, to remove scatter effects by centring and scaling each individual. This transformation cannot be applied to non-numeric data.

SAVITSKY-GOLAY DERIVATIVE ALGORITHM

It is a technique that allows to highlight overlapping peaks and locate the position of peak even in cases less-resolved. The first derivative of a peak shaped generic curve has value zero in correspondence to the peak maximum while the maximum and minimum of the derivative correspond to the flexes of the original curve.

Taking the second Derivatives of the peak shaped curve well-defined peaks are obtained; the derivative is minimum where original curve has a maximum, the derivative is maximum in correspondence to the minimum of the peaked curve.

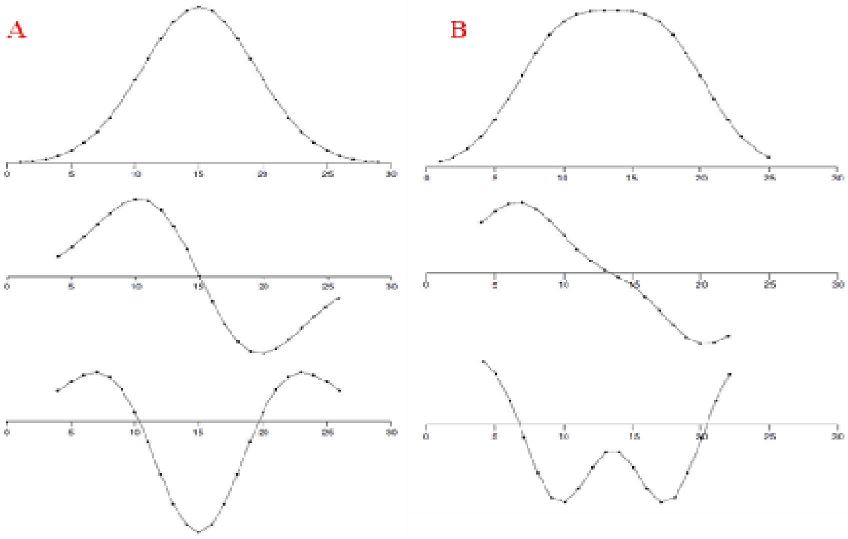


Figure 27 A) First and Second Derivatives of a Generic Gaussian Curve: The maximum point drop to zero on the derivative I allowing an easier detection of the peak position while on the derivative II becomes a minimum but reduces its dimensionality. B) The derivative II allows the detection and resolution of two superimposed peaks that in the first analysis could be confused for one signal.

Derivatives help to identify overlapping peaks and to eliminate the background curve. The higher the degree of the derivative the higher the degree of the background curve that can be eliminated. For example using the first degree derivative we can eliminate a linear background effect (when there is a straight line as a base line) using a derivative of the 2nd degree can delete a background line approximate with a second degree polynomial and so on.

It should be taken into account, however, that the numerical derivative of a signal affected by experimental noise could increase the noise so that in some cases the use of derivative becomes almost useless as well as deleterious. There are, however, some algorithms capable of reducing the influence of noise on the numeric derivative; The most used is the Savitsky-Golay[XXVIII] that combine the derivation with a smoothing algorithm. The technique is applicable only on equally spaced points: hypothesis generally satisfied with modern instrumental techniques. The method involve the selection of a 2m+1 points window over which to fit a n degree polynomial. The central point of the window is replaced by the value computed from the polynomial. The calculus repeats after shifting the window of a point; calculus ends when all points have been spanned. Both the size of the window (m) that the polynomial degree (n) must be optimized for the particular application. This algorithm allows to calculate up to (n-1)th derivatives for a n degree polynomial.[XXIX]

PRINCIPAL COMPONENT ANALYSIS (PCA)

Large data tables usually contain large amounts of information. This information hides when the data are too complex to be easily interpreted. Examples of such large tables include spectroscopic data collected on modern instrumentation, chromatographic data and the large data sets generated by biology research groups, in particular, metabolomics.

PCA can be used to reveal the hidden structure within large data sets. It provides a visual representation of the relationships between samples and variables and provides insights into how measured variables cause some samples to be similar to, or how they differ from each other.

PCA is a bilinear modeling method that provides an interpretable overview of the main information contained in a multidimensional table. It is also known as a *projection* method, because it takes information carried by the original variables and projects them onto a smaller number of latent variables called *Principal Components (PC)*. Each PC explains a certain amount of the total variance (information) contained in the dataset. The variance explained by every the PCs decreases passing from the first to the last.

To understand how PCA works, one must remind that *information* exists when there is *variation*, i.e. when a measured variable exhibits large systematic variation. Random error, however, affects the measurements too, so that *variation* (variance) has contribution from both *information* and *random error*. The variation due to random error is generally a minor contribution with respect to that due to information. PCA aims to extract the information from a data table and disregard the noise.

PCA perform an orthogonal rotation that permit to represent the data in new orthogonal (that means independent) variables, in matrix form we can write it as:

$\mathbf{X} = \mathbf{TP}^t$ where **T** contains the components in the new directions (eigenvectors) and **P** store the weights of the original variables in the new ones that could be named eigen vectors.

Here the number of eigenvectors will be the same as the number of variables, however, taking into account that they are sorted in decreasing order of retained variance and that the lower variance can be attributed to noise or in any case to useless information, we can discard some of the eigenvectors preserving only the A of them explaining the useful information.

Eigenvectors rename to *Principal Components (PC)* and the transform can be written:

$$\mathbf{X}_{N,V} = \mathbf{T}_{N,A}\mathbf{P}^t_{V,A} + \mathbf{E}_{N,V} \quad \textbf{\textit{Equation 3}}$$

where **E** is residue matrix. We use subscripts to highlight the new dimensions of the matrix.

Plot of the scores of the most important PCs can show grouping of samples. Comparing the plot of scores with the corresponding of loadings important sample-variable relationships can reveals.

PARTIAL LEAST SQUARES REGRESSION DISCRIMINANT ANALYSIS (PLS-DA)

Partial squares regression^{XXX} is one of the techniques most widely used in multivariate data processing. Basically by combining the principles of PCA and multiple regression, PLS is particularly useful when there is a need to predict a set of dependent variables (**Y** matrix) from a broad set of independent variables (predictors, **X** matrix).

Different algorithms permit to calculate the regression between the independent **X** variables and a single dependent variable with PLS1, or multiple dependent variables with PLS2[XXXI].

In PLS, the aim is therefore to predict the matrix of **Y** responses from matrix **X** (Predictors). PLS techniques, in contrast to the multivariate least squares regression, allow to calculate regression even when the number of predictors is much greater than the number of observations. This is because PLS, similarly to PCA, calculates a small number of latent variables from independent variables.

While in the case of the PCA, the selection of PCs is made to select those PCAs components that explain most of the variability of the data contained in the matrix **X**, in the PLS selection involves selecting "relevant" components (also called latent variables) both for the matrix **X** and for the matrix **Y**. More specifically, such components have to maximize the variance retained by the matrix **X** and at the same time must also maximize covariance with matrix **Y**. The predictive ability of the models is determined by calculating some parameters: root mean square error of calibration, RMSEC and cross-validation, RMSECV, multiple correlation coefficients for calibration, R^2cal and cross-validation, R^2cv. multiple correlation coefficient provides an idea about the prediction efficiency and both calibration and validation R^2 must be close to one for a good model[XXXII] . RMSEC and RMSECV values are related with the error between measured value and predicted value at each calibration step and cross-validation step, respectively. It is expected that the differences between RMSEC and RMSECV values should be small and close to zero since each of these values is attributed to the error; therefore, the main idea of good prediction is the minimization of the error. Comparison of RMSEC and RMSECV values reveals whether the calibration model is over-fitted or not[XXXIII] . When evaluating the results of a prediction model all of these parameters must be taken into consideration.

In PLS1 the matrix **X** decomposes into PCs, so that $\mathbf{X} = \mathbf{T}\,\mathbf{P}^T + \mathbf{E}$, in which the matrices **E** indicates errors made in using the main components instead of the original variables.

In PLS1 latent variables are also optimized to explain their correlation to vector of dependent variable **y** according to $\mathbf{y}_{N,1} = \mathbf{T}_{N,A}\,\mathbf{D}_{A,1.}$

This is because it is not said that the main components that retain most of the prediction variance (as in the case of PCA) are also the most important for regression purposes.

When dealing with more than one dependent variable at a time, we can use PLS2 where the two **X** and **Y** matrices are decomposed into main components according to

$$\mathbf{X} = \mathbf{T}\,\mathbf{P}^{T} + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{U}\,\mathbf{C}^{T} + \mathbf{F}$$

In these, the matrices **E** and **F** indicate the errors made in using the principal components, respectively the matrix of the predictors and that of the dependent variables, instead of the respective original variables.

These main components, better known as latent variables, are optimized with the internal relation $\mathbf{U}_{N,A} = \mathbf{T}_{N,A}\ \mathbf{D}_{A,A}$ which allows to correlate the matrix PCs of the dependent variables with those of the independent matrix to maximize their covariance.

The PLS algorithm provides latent orthogonal variables but optimized to represent the correlation between the **X** and **Y** matrix data.

PLS algorithms calculate the matrix of the coefficients **B** that allow to state predictors **Y** in closed form according to expression:

$$\mathbf{Y} = \mathbf{X}\,\mathbf{B}$$

A chance to get a model capable of representing and "interpolating" the best data through which it was built is to use a large number of latent variables. This choice, however, often leads to a phenomenon called overfitting that, as the number of latent variables used increases, the model tends to interpolate training set data more and more, but loses predictive capacity.

The validation techniques help to select a PLS model that optimize both fitting and prediction ability.

PLS methods therefore allow not only to correlate a group of independent variables **X** with a group of dependent variables **Y** but also to find the optimal dimension (significant latent variables) to represent the data and calculate the model.

PLS2 algorithm can be used for classification purposes[XXXIV] by using coded dummy variables as dependent variables. To this goal a new dummy variables is used for each category, it has code 0 if the object is not of the class otherwise its value is 1. When predicting the coded values it happen that they have decimal values so to proceed to the assignment it is necessary a threshold value, in our case 0.5, above which the object is considered assigned to the class while when below this threshold the object is out of class.

CLASSIFICATION OF CLASSIFICATION AND MODELING TECHNIQUES

Classification techniques aim to build, on the basis of a number of independent variables, a model that can identify the class that belongs to an object. It is important to establish a priori classes (based on theoretical considerations or the definition of a categorical variable or clustering methods). These are techniques that provide a quantitative response. Therefore the purpose of the classification techniques is to estimate the posterior probability to belong to the class c of the object x, which according to the Bayes theorem is obtained by

$$\mathrm{p(c/x)} = \frac{\boldsymbol{p(c)\,p(x/c\,)}}{\sum \mathrm{p(c)\,p(x/c)}} \qquad \textbf{\textit{Equation 4}}$$

where p(x/c) is a priori probability of the object x in the class c and p(c) is a priori probability of the class c. Considering for example, two classes ci and ck , the object x is assigned to the category ci if:

P(ci/x)>a•P(ck/x) vice versa to ck when P(ck/x) > a•P(ci/x) ; a, if is different than 1, defines an indecision interval. When P(ci/x) ≤ a•P(ck/x) and P(ck/x) ≥ a•P(ci/x) the object cannot be attributed to of the two categories. When P(ci/x) = a•P(ck/x) you have the same probability for x object to belong to each of the two classes.

There are several techniques able to performing, classifications and are generally distinguished in classification techniques and modeling techniques. The first calculate open models for each class so that each object must be assigned to a class even when the sample is completely foreign to the problem. Techniques of this type are: the linear discriminant analysis (LDA), K-Neares-Neighbours (KNN) and PLS-DA. Modeling techniques calculate closed models for each class. In this way, some objects may not be assigned to either class, among these methods there are, for example the techniques of quadratic discriminant analysis (UNEQ-QDA) and SIMCA modeling.

LDA

Linear Discriminant Analysis is a parametric probabilistic classification technique based on the hypothesis that the data follows the normal probability distribution and that normal distributions of each class have the same variance. The classes have to have non-coinciding barycentre. Using these hypotheses, it can detect a distinction between the i and k classes on the basis of their normal distribution; indeed being p(i/x) ∝ p(i)f(x/i) we can write:

$$\frac{p(i)}{(2\pi)^{v/2}|\mathbf{P}|^{1/2}}\exp\left[-(\mathbf{x}-\overline{\mathbf{x}}_i)^t\frac{\mathbf{P}^{-1}}{2}(\mathbf{x}-\overline{\mathbf{x}}_i)\right]=\frac{p(k)}{(2\pi)^{v/2}|\mathbf{P}|^{1/2}}\exp\left[-(\mathbf{x}-\overline{\mathbf{x}}_k)^t\frac{\mathbf{P}^{-1}}{2}(\mathbf{x}-\overline{\mathbf{x}}_k)\right]$$

The previous equation defines the equiprobability condition for object x with respect to the two classes considered. In this the matrix **P** is the pooled variance covariance (pooled), and $\overline{\boldsymbol{x}}_{\boldsymbol{i}}$ and $\overline{\boldsymbol{x}}_{\boldsymbol{k}}$ are the barycentre of classes i and k respectively. The terms p(i) and p(k) are the a priori probabilities of classes i and k respectively. From the previous one, simplifying, we get the formula for computing the discriminant score $\ln[p(i)]+2\overline{\mathbf{x}}_i^t\mathbf{P}^{-1}\mathbf{x}-\overline{\mathbf{x}}_i^t\mathbf{P}^{-1}\overline{\mathbf{x}}_i$ of each object **x** with respect to the class i. The object will be therefore assigned to the class for which its discriminating score is greater.

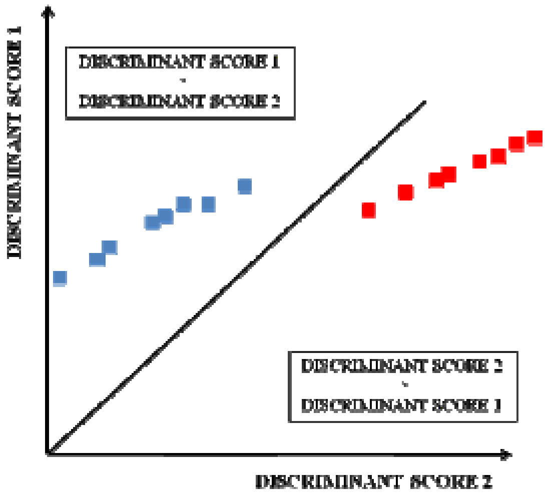


Figure 28 LDA discriminant score plot of two classes. Black line highlight the discriminant border where the discriminant scores with respect to the two classes match.

Considering all the points that the discriminating scores with respect to the categories i and j are equals a separation line, called straight discriminant (see ***Figure 28***), with respect to which is the maximum variance between the classes and at the same time the minimum intra-class variance.

UNEQ

The modeling version of QDA is known as UNEQ^{XXXV} (from “Unequal class models”). Quadratic discriminant analysis (QDA) is a probabilistic parametric classification technique, based on the following statistical hypotheses:

a) each class has a multivariate normal distribution;

b) the dispersion (measured by the variance-covariance matrix Vc) is different (rarely equal) in the classes;

c) the location (measured by the centroid) is (but not necessarily) different.

UNEQ is based on the Hoteling T^2 statistics, the multivariate generalization of Student statistics. It is based on the fact that the exponential operator in the equation of *f(**x**/c), normal probability density function in class c:*

$$f(\mathbf{x}/c) = \frac{1}{(2\pi)^{V/2}|\mathbf{V}_c|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}_c)^T \mathbf{V}_c^{-1}(\mathbf{x}-\mathbf{x}_c)}$$

Equation 4

i.e. the squared Mahalanobis distance:

$$M^2 = (X - X_c)^T V_C^{-1}(X - X_c) \quad \textbf{Equation 5}$$

is a χ^2 distributed variable, with ν = V (number of variables) degree of freedom.

Vc is the variance-covariance matrix of the class population and **x** is the centroid of the class population.

Equation5 is the equation of an isothetic (the same density of probability) ellipse, ellipsoid, hyperellipsoid. So the critical value of the χ^2 distribution, at a selected confidence level, gives the equation of the boundary of a confidence space of the class.

The mathematical CLASS MODEL is the CLASS SPACE around the CENTROID of the class.

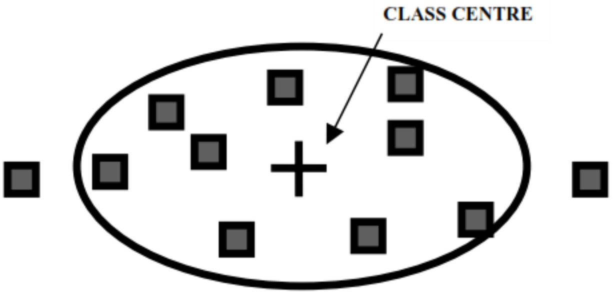


Figure 29: example of QDA class model computed fixing the confidence level

The equation 5 can be estimated using the training set from which it is possible to assess the class variance-covariance matrix, $\hat{V}_c$, and of the class centroid, $\hat{X}_c$.

UNEQ has some tolerance for not-informative variables, but this method cannot be applied when the number of objects in each class is less than the number of variables.

KNN

This is a non-parametric method based on distances between objects. This method classify an object according to the category of the majority of its K closest objects.

$$d_{i,k} = (\mathbf{x}_i - \mathbf{x}_k)'(\mathbf{x}_i - \mathbf{x}_k) = \sqrt{\sum_{v=1}^{V}(\mathbf{x}_{i,v} - \mathbf{x}_{k,v})^2}$$

The method start computing the distances of the object under exam with respect to each of all the other in the training, then it sorts these distances in ascending order.

Now examine the K closest distances, the object x_i under exam will be assigned to the class most represented in the K distances considered (criterion of majority voting). The procedure will be repeated until all the points have been processed. The number of points considered K is a critical parameter that needs to be optimized for the problem.

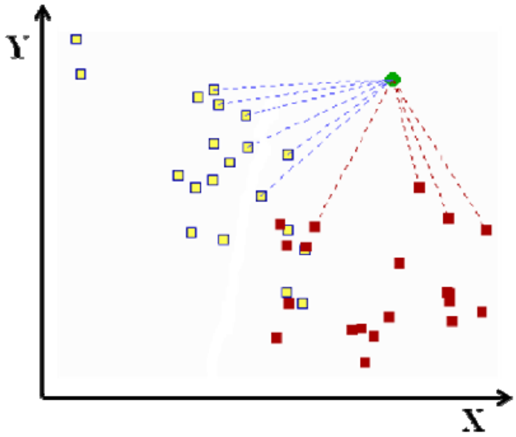
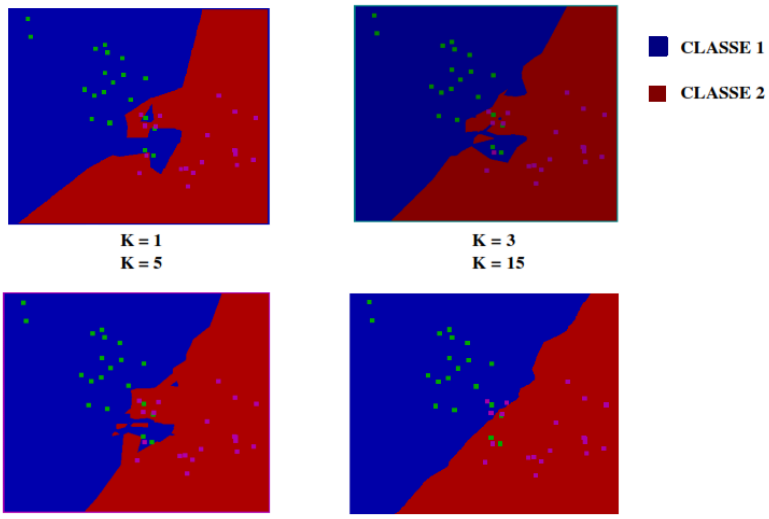


Figure 30 The green object is the incognito. Considering K=10 nearest, it is assigned to the yellow class because 6 of the nearest objects are yellow.

The value K is chosen by the operator; Generally low K values lead to a more and more marked classification as well as to the definition of composite categories even of a single object (space of a category 1 object within category 2 space, for example).

A technique for optimizing and verifying the predictive capacity of the technique can be the variation of the value of K and the study of the resulting classification.



SIMCA

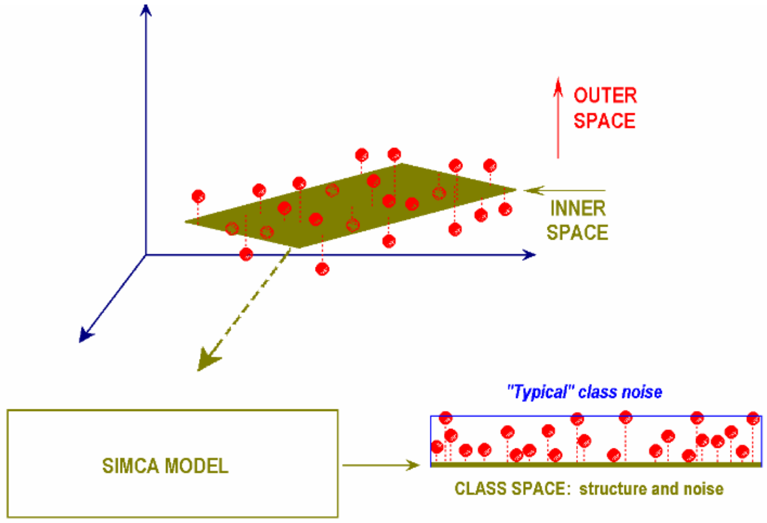
SIMCA (Soft Independent Modeling of Class Analogy)^{XXXVI} was the first class modeling technique introduced (by Svante Wold) in chemistry.

Soft: no hypothesis on the distribution of variables;

Independent: also in the case of a problem with many categories, each category model is developed independently;

Modeling: the mathematical model of the category is based on the principal components of the category; generally these are obtained as eigenvectors of the correlation coefficient matrix of the category, i.e. by using the data after separate category autoscaling. The models of more categories can be compared, for differences and analogies.

For each class, the number of significant components A is obtained by double-cross validation. It can be different for each category. A defines the dimensionality of the *inner space*, the space of the structured information; the other V-A components are the *outer space*, of the noise.



Using the training set, SIMCA calculates the main components independently for every class so that a different number of PCs can be retained for each of them.

The PCs range occupied by the category objects and the relative variance of the residues is established. For each object we can calculate, on the base of the relationship between its intra-category residue variance and its total residue variance, an F-Fisher distributed coefficient. This coefficient, according to a F Fisher test, allows to assign the probability of belonging of the object to the class.

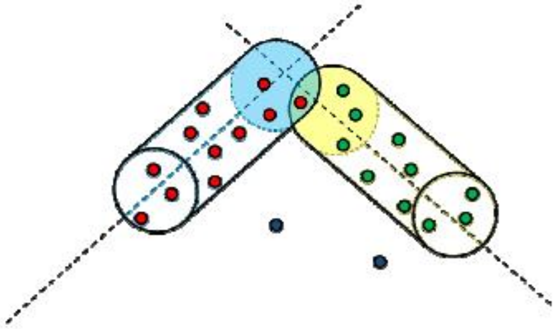


Figure 31 Graphic representation of modeling through SIMCA

This is the same of fixing a limiting value for PCs beyond which objects can no longer be considered belonging to the category. Globally, this means locating closed spaces of the category as graphically represented in the ***Figure 31***.

An object is accepted by a SIMCA model when its distance from the model is less than a critical distance, defined by means of Fisher statistics. Such critical distance is influenced by uninformative variables so that, when their number is large, many objects of non-pertinent classes may be accepted by a given model, whose specificity consequently decreases. For this reason, SIMCA is very sensitive to the presence of variables with no or small discriminant power.

SELECT A METHOD FOR VARIABLE SELECTION

SELECT[XXXVII] is a stepwise algorithm, based on the decorrelation of variables, that permits to select the most useful uncorrelated variables. It examines a variable at a time starting with that having the highest correlation. Linear correlation coefficient is the correlation index when the algorithm is applied with regression method while the Fisher classification weight is the correlation index when we want to optimize the classification. The variables are both decorrelated and selected on the base of the correlation index, upgraded at each step. The algorithm iterates until a fixed number of variables is chosen or no further selection happened in the last cycle.

EXPERIMENTAL PART

OILS SAMPLING

We gathered samples of monovarietal extra-virgin olive oils produced in Marche with the aim of identifying a specific fingerprint of the typical product. Several cause affect the characteristics of extra-virgin olive oils, between them we can highlight the season, the pedology, weather condition, processing method and storage condition. In order to take into account of these causes we sampled over 4 years (2009, 2010, 2015, 2016); the samples come from wide area in Marche including the typical areal of each variety. Sampling was carried out at some producers, selected with the help of the staff "Osservatorio regionale Suoli - Servizio Agricoltura Regione Marche" and of" Studio Agronomico Demetrio Ruffini di Colmurano". The producers guarantee the quality of the samples and the variety of the olive used. We used a total of 162 samples of oil, related to 7 categories: Ascolana, Coroncina, Mignola, Mogliano, Raggia, Leccino, Frantoio, ***Table 2*** shows the main characteristics of the samples.

Table 2 Samples used in the work.

Category	Year(2009)	Year(2010)	Year(2015)	Year(2016)
Ascolana	0	0	6	8
Coroncina	6	17	6	7
Mignola	6	15	6	7
Mogliano	7	26	9	8
Raggia	0	0	7	12
Leccino	0	3	0	6
Frantoio	1	0	0	2

Samples, gathered in autumn/winter of each year, were stored for as long as necessary to carry out the analyses in a dry place, in the dark at room temperature.

We decided not to use the samples of the Frantoio and Leccino categories, since they are not many and are not present all over the years.

INSTRUMENTS AND DATA TREATMENT

The samples were filtered on a 5 µm PVC filter before every spectroscopic measurements for removing any suspended material which may give rise to absorption and light diffusion phenomena during the spectrophotometric experiment.

IR MEASUREMENTS

IR spectra of each sample were obtained, on previously filtered oil, by a spectrophotometer Perkin Elmer spectrum 100 FT-IR equipped with ATR system, it was operated in transmission mode, resolution 0.5 cm^{-1} averaging on 11 scans to obtain the spectrum.

Each spectrum was acquired in the spectral range between 4000 and 650 cm^{-1} with subtraction of the background that was measured in air before every sample measurement. Two spectra for each sample were acquired by repeating the measurement procedure. We cleaned the sample holder by alcohol and lab paper before delivering 40 μL of oil on the measuring crystal.

UV/VIS SPECTRA

We prepared two solutions for each sample of filtered oil, at 10% and 1% respectively, carefully weighing suitable aliquots then taking to volume (volumetric dilution, 25 mL flask) with isooctane((2,2,4-Trimethylpentane) spectrophotometrically pure.

Two instruments were used for the spectroscopic measurements:

in years 2009-2010 the UV-Vis spectra were acquired from 200 to 700 nm every 0.5 nm by a Varian Cary 50 Scan spectrophotometer, spectra were subtracted of the pure isooctane spectrum; every measurement was repeated twice;

in years 2015-2016 the spectra were acquired from 190 to 1100 nm every 1 nm by a Hewlett-Packard 8453 UV-visible Spectroscopy System, spectra were subtracted of the pure isooctane spectrum. Four measurements on each solution were performed. Code A in the file name stand for 1% solution while B indicates 10% solution. Focusing on a solution the measurement procedure is as follow: Clean the cuvette with isooctane, fill it with solution get two replicas (ex. A1, A2) then clean again ad repeat other two measurements (ex. A3, A4) after them restart the procedure with another solution. All that was done for increasing the number of objects in the dataset to consider the variability due to experimental noise. In the UV-Vis spectroscopy the absorbance is linear with the concentration for values between 0.1 and 1 and the measurements are unusable since they have no analytical significance for values greater than 2.

SENSORIAL ANALYSIS

The sensory evaluation was conducted at the Assam Laboratory for Sensory Analysis by a panel of official tasters in accordance with the procedure laid down in Annex XII of the EEC commission Regulation n. 2568/91. This kind of evaluation is available only for samples of year 2015.

PHENOLS ANALYSIS

Phenols were determined on the samples of year 2015 at the HPLC–MS Laboratory of the University. A 5g of oil containing a fixed aliquot of internal standard (IS syringic acid) were dissolved in 5 mL of hexane and extracted in a separating funnel with 4 × 5 mL of methanol:water (60:40, v/v). Afterwards, 10 mL of hexane were added to the methanolic extracted solution, vortexed and centrifuged for 5 min at 5000 rpm. The methanolic solution was collected and evaporated to dryness under vacuum (60 mbar) at 30 C. The extract was reconstituted with 2.5 mL of HPLC-grade methanol and filtered through a 0.45 μm PTFE filter before HPLC-DAD-ESI/MS analysis.

HPLC-DAD-ESI/MS (ion trap) measurements were performed using an Agilent 1100 (Santa Clara, CA, USA) series instrument, made from an auto sampler, a binary solvent pump, with a diode–array detector (DAD) and a mass spectrometer detector (MSD) Trap SL equipped with an electrospray ionization (ESI) source operating in negative ionization (NI) mode. HPLC-DAD analysis, that was used for the quantification, was performed monitoring different wavelengths: 260 nm for vanillic acid, 280 nm for hydroxytyrosol, tyrosol, and secoiridoids derivatives, pinoresinol, acetoxypinoresinol and syringic acid; 310 nm for p-coumaric acid, 325 nm for ferulic acid, 338 nm for apigenin and 350 nm for luteolin. The results have been expressed in terms of concentration (mg/Kg).

VOLATILE COMPOUNDS DETERMINATION

The volatile composition was determined at the GC–MS Laboratory of the University, only for samples of year 2015, following the extraction procedure SPME and GC-MS analysis. The results have been expressed in terms peaks’s area.

Extraction SPME:

1. Put a 50 ml standard solution in a 10 ml vial, then bring a dry under nitrogen to remove the hexane.

2. Add 1.5 g of oil to the vial;

3. Insert a magnetic anchor into the vial, then immerse it in a water tank at 40 ° C and leave in agitation.

4. Expose the SPME fiber into the head space of the vial for the porous septum and leave in exposure for 30 minutes;

5. Inject GC-MS.

The method used has the following parameters:

Injector: The injector was set to splitless mode. The splitless time is 4.00 minutes

The injector temperature is 260° C. The carrier gas used is helium.

Column: The column used is an HP-88 (88% - Cyanopropylaryl-polysiloxane) of Agilent. The column has been kept at a speed of constant flow (1.2 ml / min).

Detector: The ionization source is electronically impacted (70 eV), the analyzer is a single Quadrupole, the detected mass range is 29.0-400.0. The temperature of the source is 230 ° C and the mass analyzer is 150° C.

DATA TREATMENTS SOFTWARES

The experimental data were then processed with the chemometric softwares: Unscrambler[XXXVIII] and V-Parvus[XXXIX].

PROCESSING IR DATA FOR CLASSIFICATION

The matrix **X** of IR data used with pattern recognition methods are made of IR spectra. A row of this matrix store an IR spectrum every column of it contains transmittance values at a given wavenumber for all the samples (2015 year). Replicated measurements appear as a new row in the matrix that is it is the same as a new sample. The matrix has 72 rows but, because each sample of oil is measured twice, the true oil samples are 36. The presence of the replicas permits us to have a larger number of objects and to evaluate the effect of noise on the applied data treatment.

Looking at the spectra (***Figure 32***) it is easy to see that all them are very similar, characterized from the alternation of peaks and basically no absorption area in the spectral range between 4000 and 1500 cm^{-1}. Going down in the lowest wavenumbers, we see a fingerprint area which represents a strong area of information about the nature of the compounds.

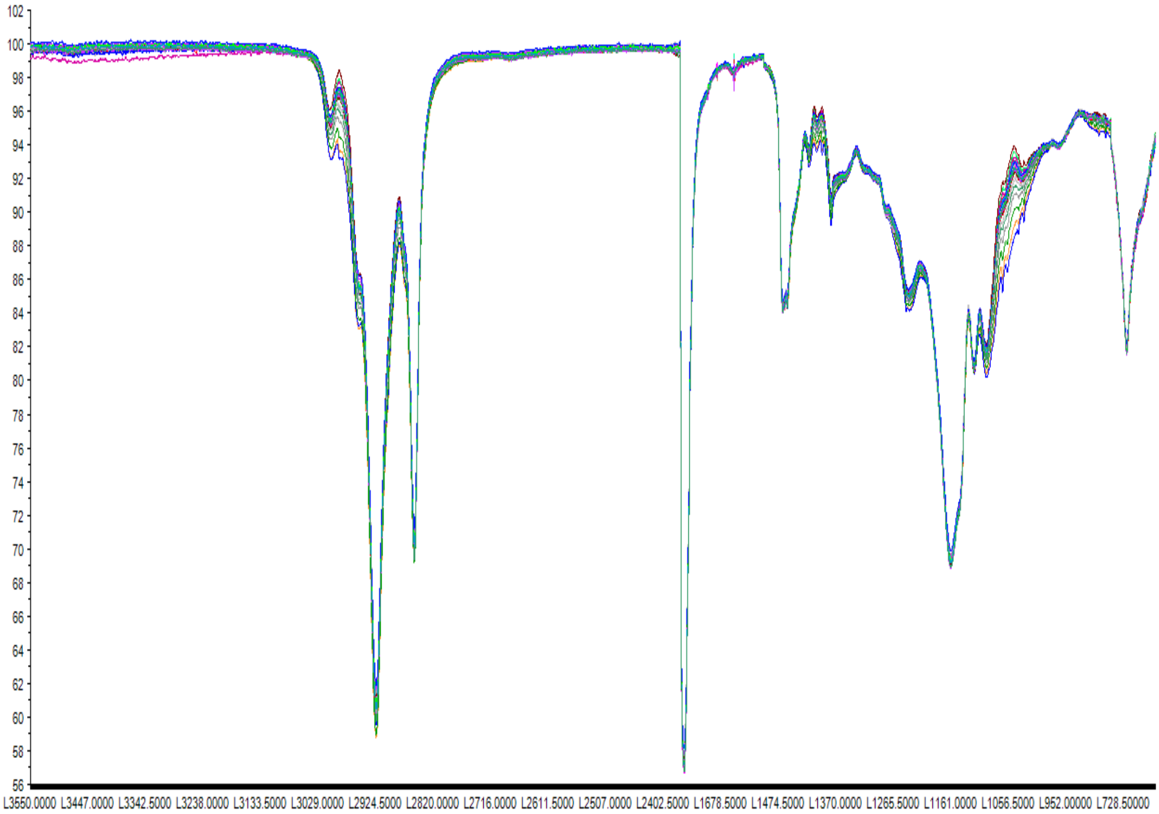


Figure 32 Overlapped FTIR Spectra of Examined Samples.

Based on these considerations, a set of wavelengths is defined, a groups of spectral areas most affected by absorption processes. That range is composed of the sub-ranges 3550-2650.5, 1750-1600,1500-870, 750.5-670 cm^{-1}, we call this IRset1 that include 3524 variables.

Several pre-treatments and transforms of the data were investigated in pairs with the different pattern recognition methods. The following pre-treatments were applied on this dataset (IRset1): SNV followed by first derivatives computed by means of third-order polynomials through eleven points Savitzky-Golay method; the result was then autoscaled by column. The use of SNV, and first derivative was effective in eliminating unwanted variations, like global intensity effects and baseline shifts. Whereas autoscaling for eliminating the scatter effects and scaling on the data.

The following chemometric methods were applied on the data matrices:

a) Principal component analysis (PCA) as a display method, in order to visualize the data structure.

b) PLS-DA in order to investigate the possibility to predict the classification of the different categories of extra-virgin olive oil.

c) UNEQ and SIMCA as class-modelling techniques, in order to build models for Monovarietal extra-virgin olive oil from Marche.

PCA

At first we used principal component analysis to “see” if there is some kind of grouping hidden in the data. Principal components analysis using only the objects of the year 2015 is performed on the IRset1 variables set including 3524 wavenumbers selected as above explained. Pre-treatment used for getting the next plot are: SNV, First derivatives and autoscaling, how we said before. Then we applied a Systematic(112233) Cross Validation method with 20 segments.

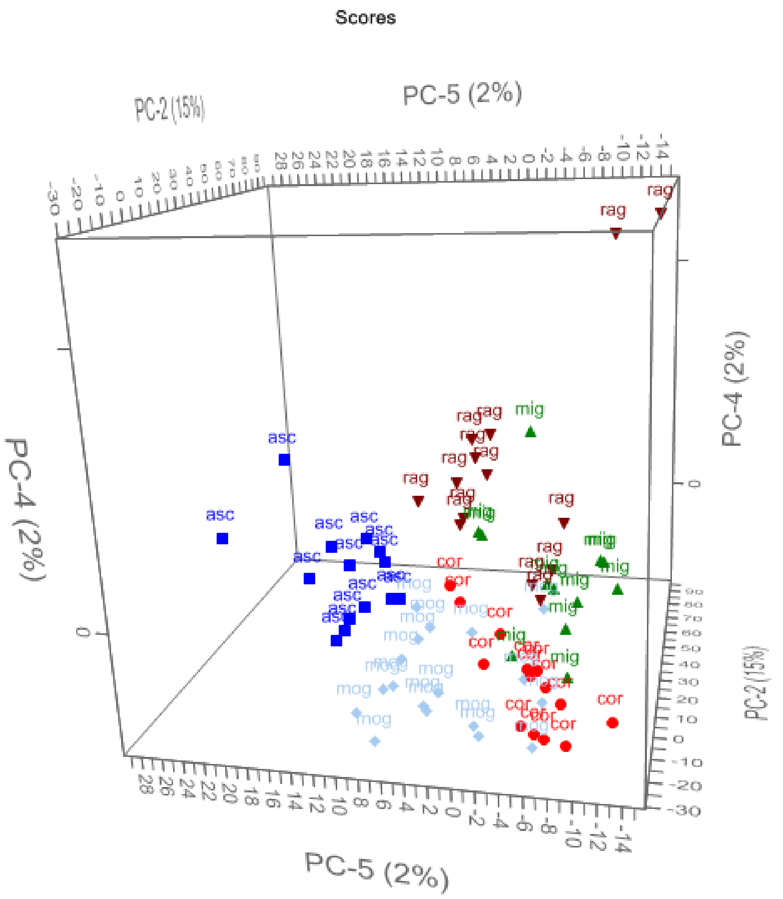


Figure 33 PCA on IRset1, score plot of PC2 (15%) PC4 (2%) and PC5 (2%)

The three-dimensional score plot of objects (**Figure 33**) show that the objects group on the base of their category if projected on the PC2, PC4 and PC5. Ascolana is the best separated category but even Mogliano and Raggia are well apart.

PLS-DA

The encouraging results obtained with PCA suggest that PLS-DA, a method based on a similar algorithm, could be profitably used to classify the oil varieties. Here variables in IRset1 were analysed using all the objects of matrix **X.** We evaluated the prediction ability of the method by 5 groups’ cross-validation. Figure 34 shows the score plot of

Factor-2 (22%,4%), Factor-3 (4%, 12%) and Factor-4 (4%,12%) where is evident the separation between categories.

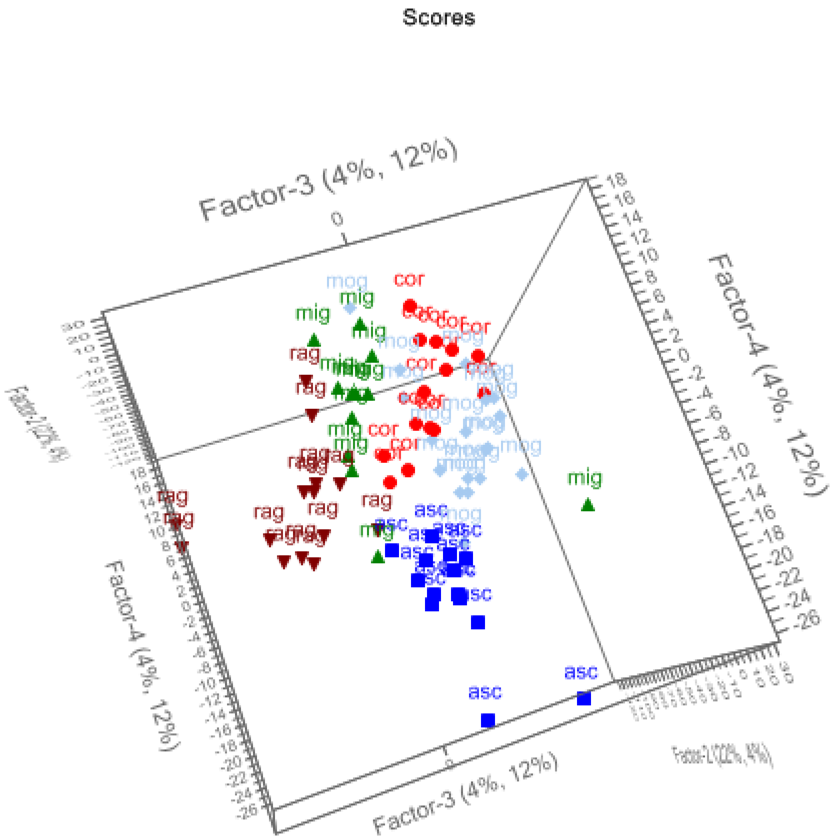


Figure 34 PLS-DA on IRset1, Score plot of Factor-2 (22%,4%), Factor-3 (4%, 12%) and Factor-4 (4%,12%).

The method, whose optimal number of latent variables is 6 shows a very good classification performance only for Mignola (see ***Table 3***), while its prediction is unsatisfactory for all categories.

Table 3 Classification and Prediction ability with PLS-DA with 6 latent variables. Calculus with all the objects of year 2015 and variables of IRset1

PLS-DA IR set1 *(6 factors)*	% CLASSIFICATION	% PREDICTION
ASCOLANA	35	14
CORONCINA	0	0
MIGNOLA	92	25
MOGLIANO	5	0
RAGGIA	50	14

The large number of variables in IRset1 could include correlated ones so that the useful information could be masked by noise and/or useless variance. To reduce this risk we tried to reduce the number of variables to use; for the goal we preferred SELECT^{XL} because it is based on the decorrelation of variables. The algorithm used on the IRset1, that is 3524 variables at the beginning, selected 10 variables here indicated with their wavenumber:

3354, 3095, 1732, 1461, 1451, 1273.5, 1136, 1132.5, 1123.5, 1107.

Hereafter this group of variables will be named ***IRselect***.

These wavenumbers can be connected with the following structural feature of the molecules:

- ✓ **1273.5**, **1461** and **1451** cm^{-1} are related to the alkylic chains, in particular the first wavenumber due to the terminal (CH_3) groups symmetric bending, while the other two wavenumbers can be assigned to the scissoring vibration of the CH_2 groups.
- ✓ **1732** cm^{-1} corresponding to the carbonyl (C=O) stretching vibration, in this case associated with the triglyceride ester bond, as well as the carboxylic group of free fatty acids.
- ✓ **3354** cm^{-1} can be associated to the phenol (O-H) stretching vibration.
- ✓ **1136**, **1132.5, 1123.5** and **1107** cm^{-1} is related to the (-C-C-O-H) antisymmetric stretch C-O compared to the Stretch C-C.
- ✓ **3094** cm^{-1} corresponding to the alkene (C-H) stretching vibration.

Repeating the PLS-DA (cross-validation with 5 groups) analysis using this variable set (IRselect) we are still able to group the objects on the base of their olive variety. ***Figure 35*** highlight a better differentiation (model with 6 latent variables) between categories with respect to what previously obtained.

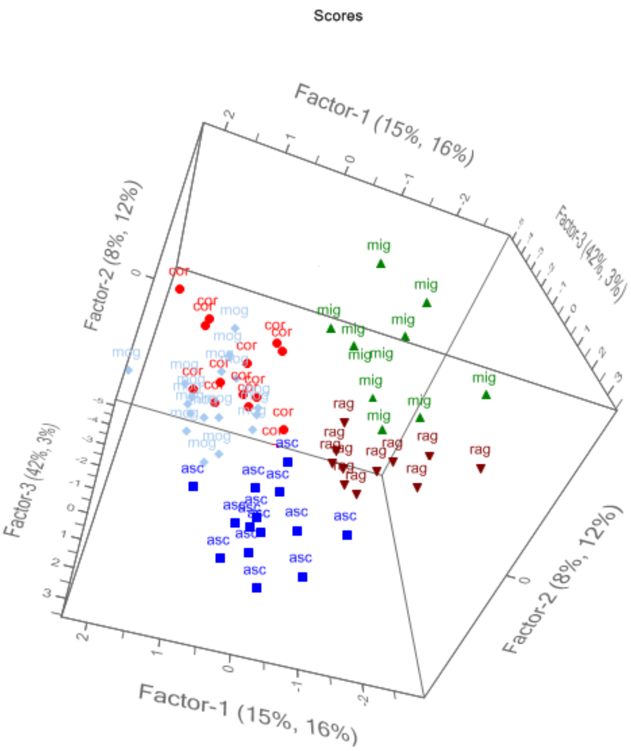


Figure 35 PLS-DA on IRselect, Score plot of Factor-1 (15%,16%), Factor-2 (8%, 12%) and Factor-3(42%,3%).

Table 4 shows the results obtained with a 6 latent variables model, classification ability diminish for Mignola while it increase quite a lot for Ascolana and then remain almost constant for the other categories. Prediction ability however increase for almost all the category.

Table 4 Classification and Prediction ability with PLS-DA(IR select) with 6 latent variables calculated with all 2015 the objects and variables of (IRselect)

PLS-DA IR select (6 factors)	**% CLASSIFICATION**	**% PREDICTION**
ASCOLANA	80	50
CORONCINA	0	0
MIGNOLA	75	33
MOGLIANO	11	6
RAGGIA	50	45

This result even if not completely satisfactory shows a potential utility of this method if we consider only Mignola, Ascolana and perhaps Raggia even if prediction ability is weak but it is useless with respect to Coroncina and Mogliano. It is to be pointed out, that the reduction of variables has improved the models.

MODELING TECHNIQUES QDA-UNEQ AND SIMCA

Since our problem is to recognize the categories of Monovarietal extra-virgin olive oil from Marche, we try to classify them with modelling methods, because they define a closed space for the model of category permitting in this way to discard even outlier samples. The classification methods, on the other hand, define an open model, because they look for a function able to divide the space so that a sample is in or out with respect to it, in so doing even outliers could be classified.

The model algorithms here investigated are UNEQ and SIMCA computed by The Unscrambler or V-Parvus.

The prediction and classification ability of a method, as well as by the confusion matrix, can best be judged if other indices such as sensitivity, specificity, and efficiency are also considered.

- **Sensitivity** is the percentage of objects in the investigated class that is getting recognized as belonging to the class model. It is the experimental measure of the confidence level. Sensitivity of 100% means that the system recognizes all objects of class C_i as such.
- **Specificity** is the percentage of the objects of the categories different from the modelled one rejected by the class model. Specificity of 100% means that the system recognizes all objects NOT of class C_i as such.
- **Efficiency** takes into account both Specificity and Sensitivity, it is the geometric mean of these two parameters.

  $$Efficiency = \sqrt[2]{(Specificity * Sensitivity)}.$$

  The comprehensive parameter *Efficiency* may vary between 0, when either sensitivity or specificity are null, and 100 (the ideal case), when both parameters have the maximum value of 100.

In the method optimization we have to find a compromise between specificity and efficiency, because real cases where we can maximize both are rare. Therefore, it depends on the specific situation which of them to improve most.

Training set and External set for modelling methods

To optimize both modelling methods (SIMCA and UNEQ) the data matrix X has been divided in two sets. The Training (see **Table 5**) include 54 objects, actually 27 oil samples, all produced in 2015, the remaining objects of the matrix X (18 objects) were used as external set together to objects related to all others years (2009,2010,2016). (see ***Table 6***).

Table 5 Training set (54 objects)

Category	Year (2015)
Ascolana	10
Coroncina	10
Mignola	10
Mogliano	14
Raggia	10

Table 6 External set (258 objects)

Category	Year(2009)	Year(2010)	Year(2015)	Year(2016)
Ascolana	0	0	4	16
Coroncina	12	34	4	14
Mignola	14	28	2	14
Mogliano	14	52	4	18
Raggia	0	0	4	24

SIMCA AND UNEQ MODELS WITH 5 CATEGORIES

SIMCA

The SIMCA model here discussed were computed retaining three principal components in each class at 95% confidence interval.

Table 7 shows that the sensitivity with the training set is poor only for Mogliano but good or excellent for other classes, even specificity is excellent (see **Table 8**) except for Mogliano that seems to have some degree of superposition with Coroncina, however the prediction shows, **Table 9**, very poor sensitivity for all the categories. Specificity is still high, **Table 10**, so that the efficiency (see **Table 11**) assume values indicating the inappropriateness of the method.

Table 7 Sensitivity Training set of Simca model.

SENSITIVITIES MODEL (Level = 95%)		
Ascolana	70%	(Accepted 7/ 10)
Coroncina	100%	(Accepted 10/10)
Mignola	80%	(Accepted 8/ 10)
Mogliano	43%	(Accepted 6/ 14)
Raggia	100%	(Accepted 10/10)
Mean sensitivity 76%		

Table 8 Specificity of Training set Simca model.

SPECIFICITIES MODEL (Level = 95 %)	
Ascolana	**Coroncina**
Coroncina 100.0% (Rejected 10/10)	Ascolana 100.0% (Rejected 10/10)
Mignola 100.0% (Rejected 10/10)	Mignola 100.0% (Rejected 10/10)
Mogliano 100.0% (Rejected 14/14)	Mogliano 85.7% (Rejected 12/14)
Raggia 100.0% (Rejected 10/10)	Raggia 100.0% (Rejected 10/10)
Mean specificity for Ascolana 100%	**Mean specificity for Coroncina 95.4%**
Mignola	**Mogliano**
Ascolana 100.0% (Rejected 10/10)	Ascolana 100.0% (Rejected 10/10)
Coroncina 100.0% (Rejected 10/10)	Coroncina 30.0% (Rejected 3/10)
Mogliano 100.0% (Rejected 14/14)	Mignola 100.0% (Rejected 10/10)
Raggia 60.0% (Rejected 6/10)	Raggia 100.0% (Rejected 10/10)
Mean specificity for Mignola 90.9%	**Mean specificity for Mogliano 82.5%**
Raggia	
Ascolana 100.0% (Rejected 10/10)	
Coroncina 100.0% (Rejected 10/10)	
Mignola 100.0% (Rejected 10/10)	
Mogliano 100.0% (Rejected 14/14)	
Mean specificity for Raggia 100.0%	

Table 9 Sensitivity of Ext-set Simca model.

SENSITIVITIES EXT SET (Level = 95%)		
Ascolana	25%	(Accepted 5 / 20)
Coroncina	2%	(Accepted 1/64)
Mignola	19%	(Accepted 11/58)
Mogliano	7%	(Accepted 6/88)
Raggia	14%	(Accepted 4/28)
Mean sensitivity 10.5%		

Table 10 Specificity of Ext- set Simca model.

SPECIFICITIES EXT SET (Level = 95 %)	
Ascolana	**Coroncina**
Coroncina 98.0% (Rejected 63/64)	Ascolana 100.0% (Rejected 20/20)
Mignola 100.0% (Rejected 58/58)	Mignola 100.0% (Rejected 58/58)
Mogliano 98.0% (Rejected 86/88)	Mogliano 100.0% (Rejected 88/88)
Raggia 96.0% (Rejected 27/28)	Raggia 100.0% (Rejected 28/28)
Mean specificity for Ascolana 98%	**Mean specificity for Coroncina 100%**
Mignola	**Mogliano**
Ascolana 100.0% (Rejected 20/20)	Ascolana 100.0% (Rejected 20/20)
Coroncina 92.0% (Rejected 59/64)	Coroncina 93.7% (Rejected 60/64)
Mogliano 98.8% (Rejected 87/88)	Mignola 100.0% (Rejected 58/58)
Raggia 93.0% (Rejected 26/28)	Raggia 92.9% (Rejected 26/28)
Mean specificity for Mignola 96%	**Mean specificity for Mogliano 96.5%**
Raggia	
Ascolana 100.0% (Rejected 20/20)	
Coroncina 100.0% (Rejected 64/64)	
Mignola 98.3% (Rejected 57/58)	
Mogliano 97.7% (Rejected 86/88)	
Mean specificity for Raggia 98.7%	

Table 11 Efficiency of Ext-set Simca model.

EFFICIENCY EXT SET (Level = 95%)	
Ascolana	50%
Coroncina	14%
Mignola	43%
Mogliano	26%
Raggia	37%

UNEQ

Because UNEQ is less sensitive than SIMCA to the presence of uninformative variables we check if it is better responding with respect to our problem. The data matrix **X** was identically partitioned as in the previous SIMCA data treatment (see **Table 5** and **Table 6** for details).

The UNEQ model here discussed were computed retaining three principal components (autoscaled) in each class at 95% confidence interval.

Comparing **Table 13** and **Table 8** we note a better performance of UNEQ with respect to SIMCA, in particular, it is able to well classify all the classes, however **Table 13** shows that the specificity is poor in all classes.

Table 12 Sensitivity of Training set Uneq model.

SENSITIVITIES MODEL (Level = 95%)		
Ascolana	90%	(Accepted 9/ 10)
Coroncina	90%	(Accepted 9/ 10)
Mignola	80%	(Accepted 8/ 10)
Mogliano	86%	(Accepted 12/14)
Raggia	90%	(Accepted 9 / 10)
Mean sensitivity 87%		

Table 13 Specificity of Training set Uneq model.

SPECIFICITIES MODEL (Level = 95 %)					
Ascolana			**Coroncina**		
Coroncina	30%	(Rejected 3/10)	Ascolana	20%	(Rejected 2/10)
Mignola	30%	(Rejected 3/10)	Mignola	40%	(Rejected 4/10)
Mogliano	50%	(Rejected 7/14)	Mogliano	14%	(Rejected 2/14)
Raggia	60%	(Rejected 6/10)	Raggia	60%	(Rejected 6/10)
Mean specificity for Ascolana 43.2%			**Mean specificity for Coroncina 31.8%**		
Mignola			**Mogliano**		
Ascolana	0%	(Rejected 0/10)	Ascolana	20%	(Rejected 2/10)
Coroncina	10%	(Rejected 1/10)	Coroncina	10%	(Rejected 1/10)
Mogliano	0%	(Rejected 0/14)	Mignola	70%	(Rejected 7/10)
Raggia	0%	(Rejected 0/10)	Raggia	90%	(Rejected 9/10)
Mean specificity for Mignola 2.27%			**Mean specificity for Mogliano 47.5%**		
Raggia					
Ascolana	20%	(Rejected 2/10)			
Coroncina	50%	(Rejected 5/10)			
Mignola	20%	(Rejected 2/10)			
Mogliano	79%	(Rejected 11/14)			
Mean specificity for Raggia 45.5%					

Prediction of the external test confirm the good sensitivity (**Table 14**) of all classes however specificity (**Table 15**) remains low with some class exception as Mogliano and Raggia for which the efficiency (**Table 16**) is significant.

Table 14 Sensitivity of Ext- set Uneq model.

SENSITIVITIES EXT SET (Level = 95%)		
Ascolana	75%	(Accepted 15/ 20)
Coroncina	88%	(Accepted 56/ 64)
Mignola	98%	(Accepted 57/ 58)
Mogliano	86%	(Accepted 76/ 88)
Raggia	82%	(Accepted 23/28)
Mean sensitivity 88%		

Table 15 Specificity of Ext-set Uneq model.

SPECIFICITIES EXT SET (Level = 95 %)	
Ascolana	**Coroncina**
Coroncina 47% (Rejected 30/64)	Ascolana 5% (Rejected 1/20)
Mignola 43% (Rejected 25/58)	Mignola 84% (Rejected 49/58)
Mogliano 42% (Rejected 37/88)	Mogliano 14% (Rejected 12/88)
Raggia 36% (Rejected 10/28)	Raggia 39% (Rejected 11/28)
Mean specificity for Ascolana 43%	**Mean specificity for Coroncina 38%**
Mignola	**Mogliano**
Ascolana 0% (Rejected 0/10)	Ascolana 10% (Rejected 2/20)
Coroncina 2% (Rejected 1/64)	Coroncina 13% (Rejected 8/64)
Mogliano 6% (Rejected 5/88)	Mignola 85% (Rejected 49/58)
Raggia 0% (Rejected 0/28)	Raggia 54% (Rejected 15/28)
Mean specificity for Mignola 3%	**Mean specificity for Mogliano 44%**
Raggia	
Ascolana 35% (Rejected 7/20)	
Coroncina 52% (Rejected 33/64)	
Mignola 69% (Rejected 40/58)	
Mogliano 60% (Rejected 53/88)	
Mean specificity for Raggia 58%	

Table 16 Efficiency of Ext-set Uneq model

EFFICIENCY EXT SET (Level = 95%)	
Ascolana	57%
Coroncina	58%
Mignola	17%
Mogliano	62%
Raggia	69%

SIMCA AND UNEQ MODELS WITH 3 CATEGORIES

As we can see UNEQ but even worst SIMCA does not work very well, however we used 5 class data in the training set when data from 2009 to 2010, included in the external test, have not oils of category Ascolana and Raggia. This fact introduces an high unbalance between classes with respect to their validation number of objects.

Since the external set is balanced enough on three categories, Coroncina, Mignola and Mogliano, we decided to build a new model, with the data of year 2015 (see **Table 17**), considering only these three categories for seeing if we can improve the predictability.

Table 17 Training set (34 objects)

Category	Year (2015)
Coroncina	10
Mignola	10
Mogliano	14

Table 18 External set (210 objects)

Category	Year(2009)	Year(2010)	Year(2015)	Year(2016)
Coroncina	12	34	4	14
Mignola	14	28	2	14
Mogliano	14	52	4	18

The following pre-treatments were applied to all data (Training set and External set) for modelling methods (SIMCA and UNEQ): Standard Normal Variate (SNV) followed by First derivatives(Savitzky-Golay method) and column autoscaling.

Before performing classification algorithms we reapply SELECT to select the most important variables. The procedure use IRset1 (3524 variables) of which 8 variables are selected:

2917.5, 2860.5, 2824.5, 2798, 1235.5, 1226, 1110.5, 1101

Name it **IRselect1**

Even in this new case, it becomes useful to trace which compounds affected the absorption at these specific wavelengths.

- ✓ **2917.5, 2860.5, 2824.5 and 2798** cm^{-1} are related to the CH_2 and CH_3 stretching modes vibration of the aliphatic moiety of the fatty acids.
- ✓ **1235.5, 1226 and 1110.5** cm^{-1} correspond to C-O stretching vibrations of the aliphatic esters.
- ✓ **1101** cm^{-1} can be associated to the O-H bending of the secondary alcohol.

Both SIMCA and UNEQ models were computed with **IRselect1** variables and a Training set of 34 objects (**Table 17**), actually 17 samples. Their prediction ability was estimated using an external set (**Table 18**) of 210 objects belonging to the years 2009, 2010, 2015, 2016.

SIMCA

Comparison of data (in ***Table 19*** and ***Table 20)*** highlights the good performance obtained both as sensitivity and specificity. As witnessed even by the efficiency (**Table 21**) this method could be good enough for a practical implementation.

Table 19 Classification and prediction Sensitivity of SIMCA 3 category models

SENSITIVITIES MODEL (Level = 95%)			**SENSITIVITIES EXT SET (Level = 95%)**	
Coroncina	90%	(Accepted 51/ 64)	80%	(Accepted 9/ 10)
Mignola	90%	(Accepted 44/ 58)	76%	(Accepted 9/ 10)
Mogliano	100%	(Accepted 64/ 88)	74%	(Accepted 14/14)
Mean sensitivity 94%			**Mean sensitivity 76%**	

Table 20 Classification and prediction Specificity of SIMCA 3 category models

SPECIFICITIES MODEL (Level = 95 %)	**SPECIFICITIES EXT SET (Level = 95 %)**
Coroncina	**Coroncina**
Mignola 80% (Rejected 8/ 10)	Mignola 91.4% (Rejected 53/ 58)
Mogliano 71.5% (Rejected 10/ 14)	Mogliano 79.5% (Rejected 70/ 88)
Mean specificity for Coroncina 75%	**Mean specificity for Coroncina 82%**
Mignola	**Mignola**
Coroncina 50% (Rejected 5/ 10)	Coroncina 40% (Rejected 25/ 64)
Mogliano 64.3% (Rejected 9/ 14)	Mogliano 72% (Rejected 63/ 88)
Mean specificity for Mignola 58.3%	**Mean specificity for Mignola 58%**
Mogliano	**Mogliano**
Coroncina 20% (Rejected 2/ 10)	Coroncina 52% (Rejected 33/ 64)
Mignola 50% (Rejected 5/ 10)	Mignola 72% (Rejected 42/ 58)
Mean specificity for Mogliano 35%	**Mean specificity for Mogliano 62%**

Table 21 Prediction Efficiency of SIMCA 3 category models

EFFICIENCY EXT SET (Level = 95%)	
Coroncina	81%
Mignola	66%
Mogliano	68%

As the specificity of some oil, as Mogliano or Mignola, is not satisfying we check if better results are possible by UNEQ.

UNEQ

The modelling by UNEQ produces very high sensitivities (**Table 22**), even in prediction, for all the categories but very poor specificity (**Table 23**) with respect to those of SIMCA model.

Table 22 Classification and prediction Sensitivity of UNEQ 3 category models .

SENSITIVITIES MODEL (Level = 95%)			SENSITIVITIES EXT SET (Level = 95%)	
Coroncina	90%	(Accepted 9/ 10)	83%	(Accepted 53/ 64)
Mignola	80%	(Accepted 8/ 10)	98%	(Accepted 57/ 58)
Mogliano	93%	(Accepted 13/14)	92%	(Accepted 81/88)
Mean sensitivity 88%			**Mean sensitivity 91.4%**	

Table 23 Classification and prediction Specificity of UNEQ 3 category models.

SPECIFICITIES MODEL (Level = 95 %)	SPECIFICITIES EXT SET (Level = 95 %)
Coroncina	**Coroncina**
Mignola 50% (Rejected 5/ 10)	Mignola 33% (Rejected 19/ 58)
Mogliano 14% (Rejected 2/ 14)	Mogliano 10% (Rejected 9/ 88)
Mean specificity for Coroncina 30%	**Mean specificity for Coroncina 19%**
Mignola	**Mignola**
Coroncina 0% (Rejected 0/ 10)	Coroncina 0% (Rejected 0/ 64)
Mogliano 0% (Rejected 0/ 14)	Mogliano 0% (Rejected 0/ 88)
Mean specificity for Mignola 0%	**Mean specificity for Mignola 0%**
Mogliano	**Mogliano**
Coroncina 10% (Rejected 1/ 10)	Coroncina 8% (Rejected 5/ 64)
Mignola 20% (Rejected 2/ 10)	Mignola 7% (Rejected 4/ 58)
Mean specificity for Mogliano 15%	**Mean specificity for Mogliano 7%**

Table 24 Efficiency of Ext-set Uneq model.

EFFICIENCY EXT SET (Level = 95%)	
Coroncina	40%
Mignola	0%
Mogliano	25%

The efficiency (**Table 24**)shows the weakness of the method.

As we see from these results, by building a model on three categories, we have got good results with SIMCA that did not work well when we used 5 categories. In fact in this case we have both, a good Sensitivity (prediction mean value 76%), so we are able to recognize the most part of the samples, and also a good Specificity (prediction mean value 65%), this means that each of the three class models exclude samples that are not belonging to it fairly well. This is particularly noticeable in the case of Coroncina where the prediction sensitivity if 90% and the prediction specificity attains 82%.

With UNEQ we got that the 3 class models are able to recognize quite well all samples (prediction mean sensitivity= 91%) but they don't succeed to exclude samples that do not belong to them, in fact the specificity parameter is worse than when modelling with 5 categories.

PREDICTION OF CHEMICAL PARAMETERS OF OLIVE OIL WITH FTIR

This study aims to determine some correlation between the spectroscopic measurements (FTIR) and volatiles and phenols substances or sensorial features of extra-virgin olive oils. The samples are those of year 2015 described in ***Table 2***.

Partial least square (PLS) calibration models were computed in order to reveal possible correlation between quality parameters and spectral data. The evaluation of the performance of these models was based on the multiple coefficient of determination (R^2), the root mean square error of calibration (RMSEC) and root mean square error of cross validation (RMSEC). The **Table 25** summarise the variable hereafter investigated for correlation with IR data, they results as the best correlated in an exploratory study including even other compounds.

Table 25 Selected volatile, phenolic and sensory variables used in the correlation study with IR variables.

Sensory notes	Volatile compounds	Phenolic compounds
FRUITY	2-HEXENAL	P-COUMARIC ACID
BITTER	HEXANAL	PINORESINOL

PLS regression was used to relate 10 of the IR spectral data (IRselect) with the analytical measurements of important chemical parameters of olive oils, those shown in **Table 25**. In order to increase the predictive ability of the PLS model area normalization was applied to the 10 of the spectral FTIR data.

Models were constructed for each response separately, and to validate the developed PLS model, Leave-One-Out (LOO) cross-validation method was used[XLI].

The Unscambler software was used for PLS calculus.

The best correlation, among sensory parameters, was found for Fruity (see **Figure 36**) with R^2cal = 0.99 and R^2cv = 0.99 when using a model with six latent variables. The high value of R^2 and the lowest of RMSEC (0.47) and RMSECV (0.52) indicate the good performance and precision of PLS model. The Weighted Regression Coefficients plot (***Figure 37***) shows that the Fruit note is mainly influenced by variables: 1273.5 cm^{-1}, related to the alkylic chains due to the terminal (CH_3) groups symmetric bending, and 1123.5 cm^{-1}, related to the (-C-C-O-H) antisymmetric stretch C-O compared to the Stretch C-C.

The Bitter (see ***Figure 38***) taste, modelled with six latent variables, also correlated well to IR; the values R^2cal = 0.97 and R^2cv = 0.97 indicates good stability of the model.

This is also supported by the values of tolerable differences between RMSEC (0.67) and RMSECV (0.62). ***Figure 39***, plot of the weights of variables, shows that Bitter taste is highly connected with IR variables 1273.5 and 1132.5 cm^{-1}.The variable 1132.5 cm^{-1} is related to the (-C-C-O-H) antisymmetric stretch C-O compared to the Stretch C-C.

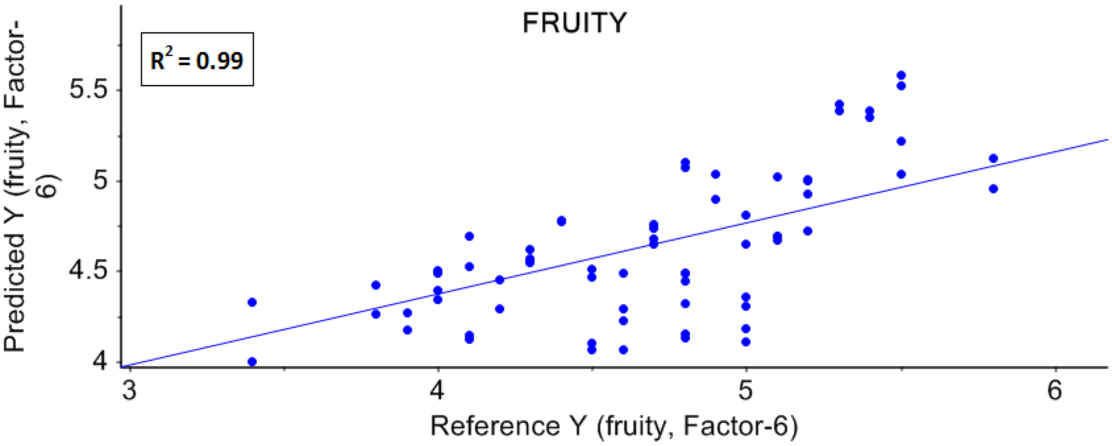


Figure 36 Predicted vs Reference (Fruity taste, 6 factors).

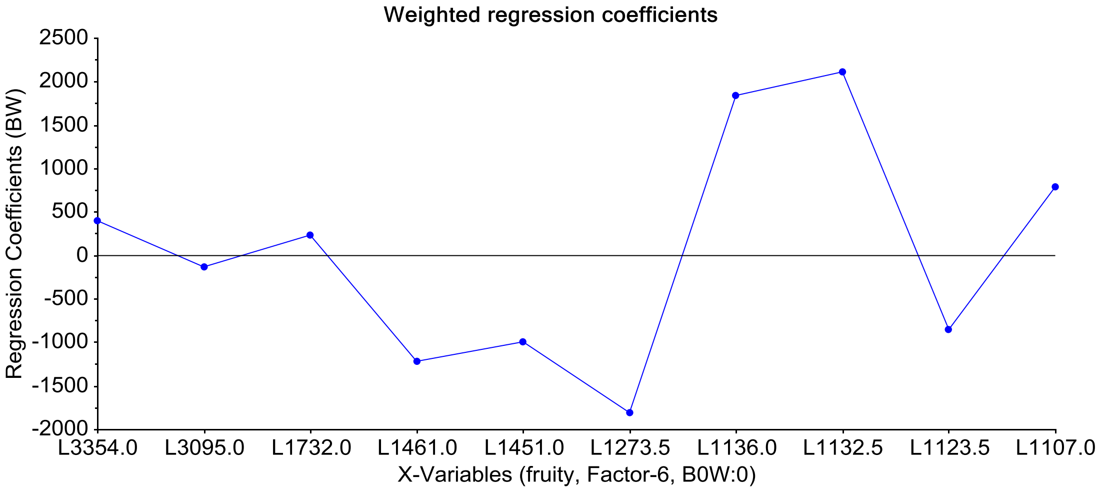


Figure 37 Weighted regression coefficients (Fruity taste, 6 factors).

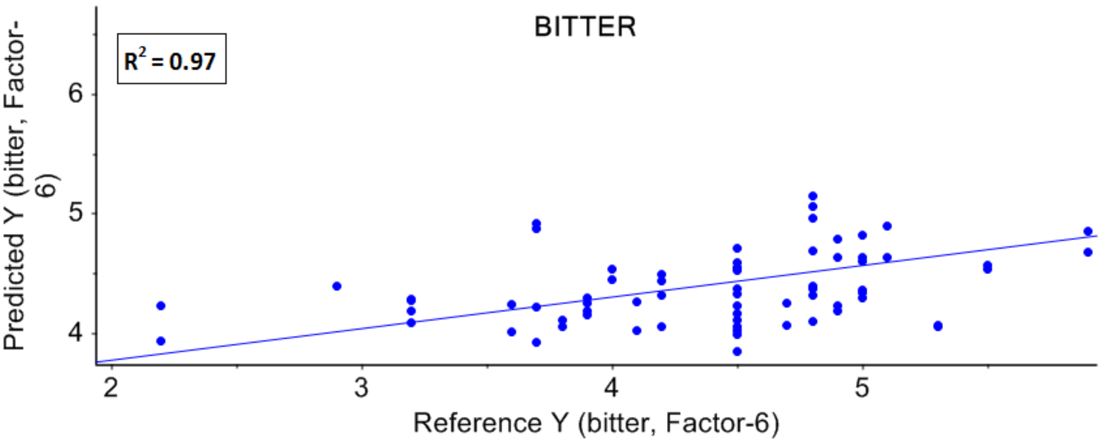


Figure 38 Predicted vs Reference (Bitter taste, 6 factors)

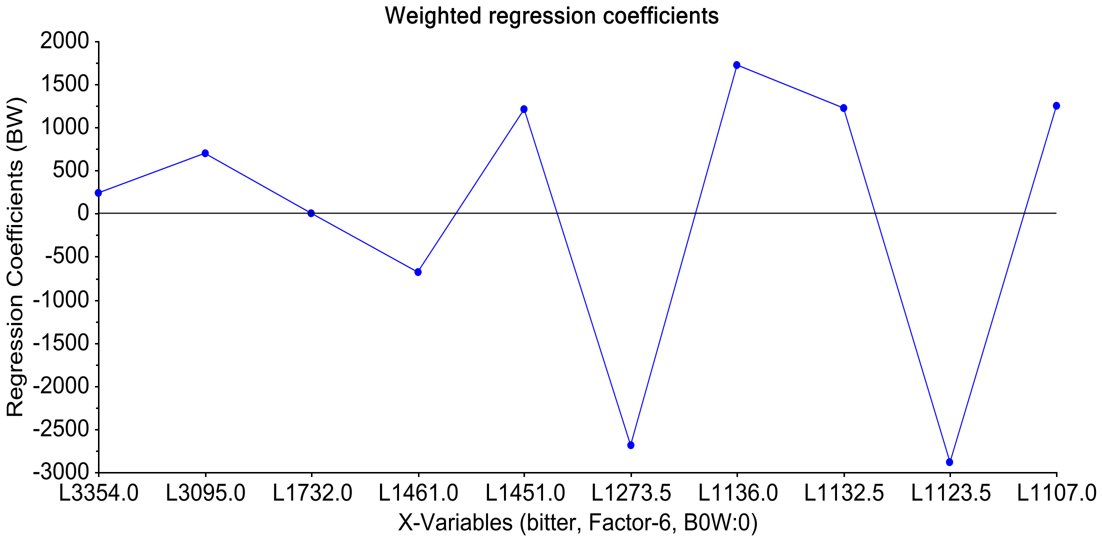


Figure 39: Weighted regression coefficients (Bitter taste, 6 factors).

The phenolic compounds P-Coumaric Acid correlate well enough to IR features, R^2cal = 0.91 and R^2cv = 0.89 indicates a good model (***Figure 40***). This is also supported by the low RMSE (that are RMSE=RMSECV=0.03).

Besides P-Coumaric acid, another phenlic compound, Pinoresinol shows high correlation with IR variables, with R^2cal= 0.88 and R^2cv = 0.85. This is also supported by the low difference between RMSEC(2.8) and RMSECV(3.1). Comparing the Weighted Regression Cofficients in ***Figure 41*** with those in ***Figure 43*** we see that the wavenumber 1451 cm-1 is highly significant for both phenolic compound models. This wavelength is related to the scissoring vibration of the CH2 groups[XLII] of the alkylic chains. The other important IR variable (1732 cm-1) for P-Coumaric acid corresponding to the carbonyl (C=O) stretching vibration, in this case associated with the triglyceride ester bond, as well as the carboxylic group of free fatty acids. The other IR variable (1123.5 cm^{-1}) in the case Pinoresinol model (***Figure 43***) is also one of the important wavelengths in the Bitter sensory model (**Figure 39**), in fact how we can see from tables (**Table 26, Table 27**), Bitter also is influenced by phenols. The peak 1451 cm^{-1} is due to the scissoring vibration of the CH2 groups[XLIII] of alkylic chains.

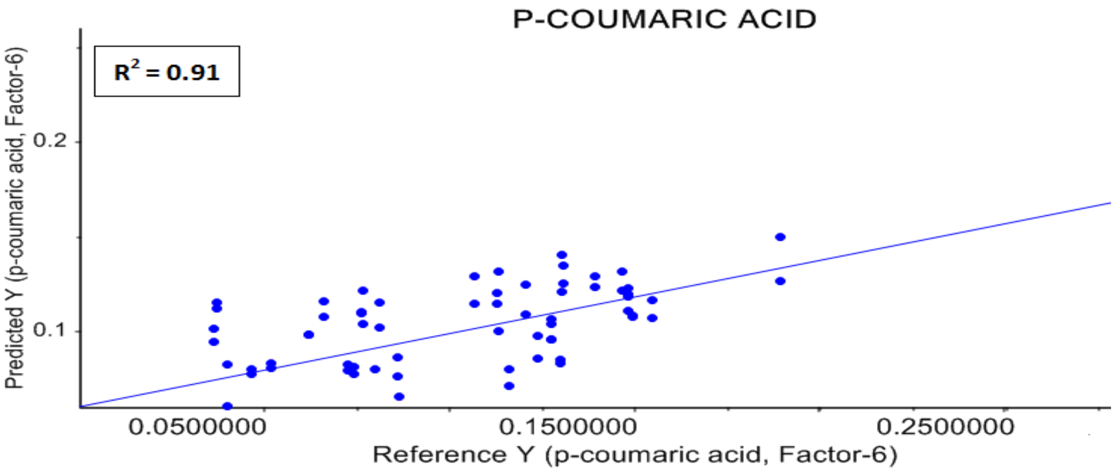


Figure 40 Predicted vs Reference (P-Coumaric acid, 6 factors).

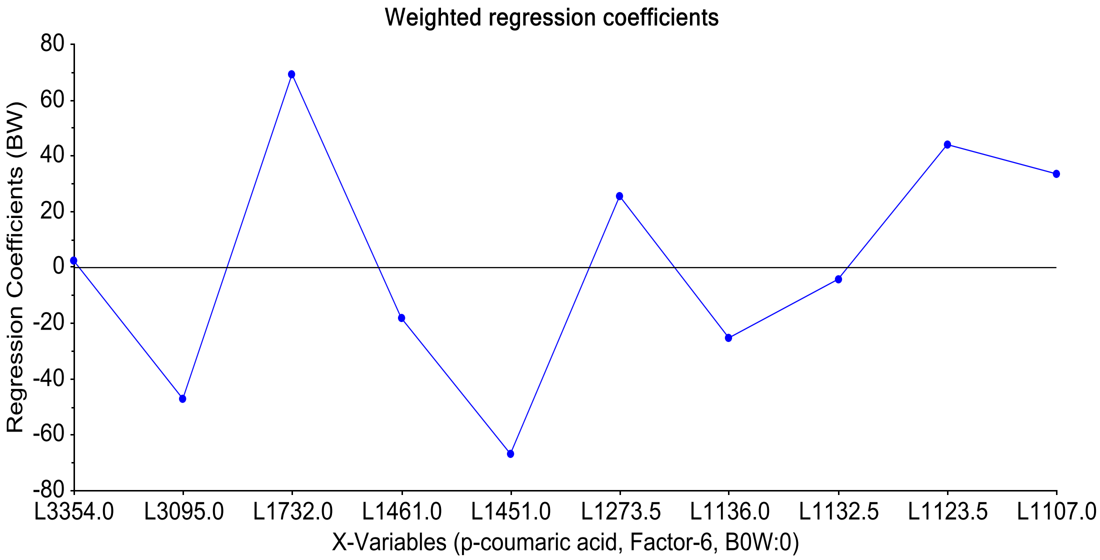


Figure 41 Weighted regression coefficients (P-Coumaric, 6 factors).

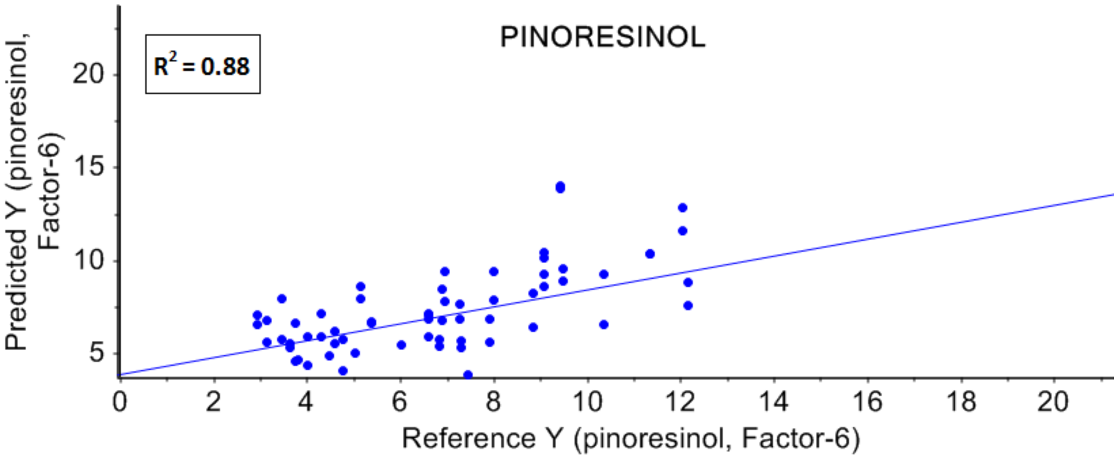


Figure 42 Predicted vs Reference (Pinoresinol, 6 factors).

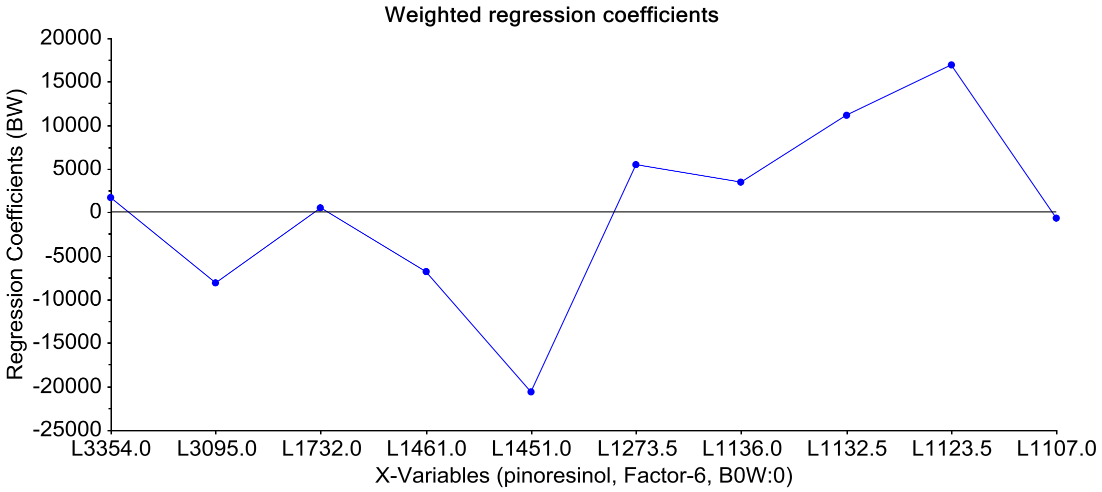


Figure 43 Weighted regression coefficients (Pinoresinol, 6 factors).

Volatile compouds such as: ethanol, penten dimers, 3-pentanone, 1-penten-3-one, 1-penten-3-ol, 1-hexanol, (Z)-3-hexen-1-yl acetate, (Z)-3-hexen-1-ol and (E)-3-hexen-1-ol are not correlated to IR wavelengths so well as the predict as the phenolic compounds. We observed, however, that 2-Hexenal and Hexanal correlate well enough to IR with 6 latent PLS model. We got a good values of regression coefficient (***Figure 44*** and ***Figure 46***) both calibration and validation R^2cal (0.83, 0.77) and R^2cv (0.79, 0.73) values provides an idea about the correlation with IR wavelength investigated. RMSEC and RMSECV values are related with the error between measured value and predicted value at each calibration step and cross-validation step, respectively are quite high, but the differences between them is small, 0.8 (in the case of Hexenal) and 0.15 (in the case of 2-Hexenal). A comparison of the " Weighted Regression Coefficients" plots for 2-Hexenal (see **Figure 45**) with those for Bitter (see **Figure 39**) permit so to see that wavenumbers 1273.5, 1123.5 cm^{-1} are important in both models. The result agrees with the two tables below, where **Table 26** shows relationships among aromatic chemical compounds and sensory reviewed in literature[XLIV]. **Table 27** shows relationships among aromatic chemical compounds and sensory notes but also include odor and taste thresholds found by several authors[XLV].

A comparison of the most important weighted regression coefficient reported in **Figure 37** with those in **Figure 47** permit highlight that Hexanal is correlated with Fruit as is also reported in **Table26** and **Table 27**.

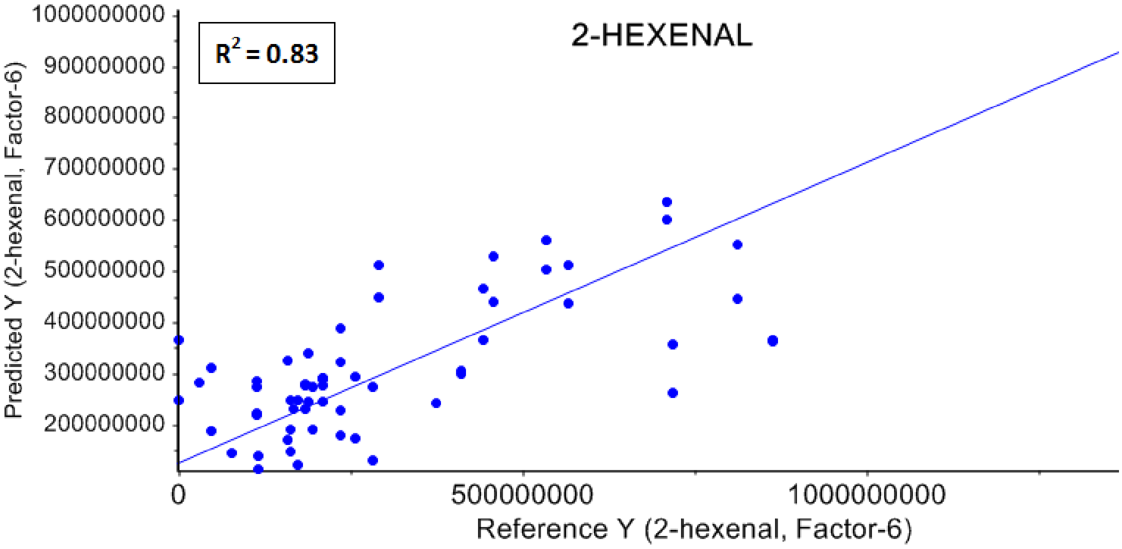


Figure 44 Predicted vs Reference (2-Hexenal, 6 factors)

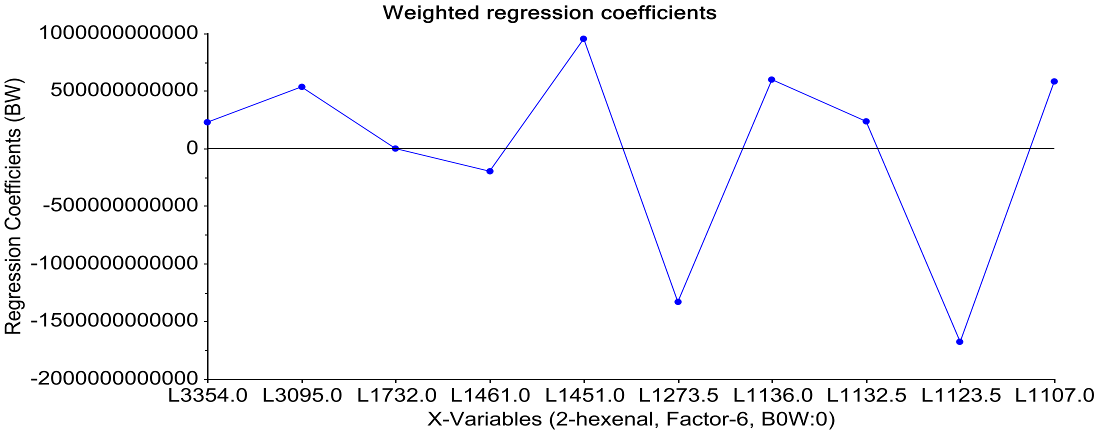


Figure 45 Weighted regression coefficients (2-hexenal, 6 factors)

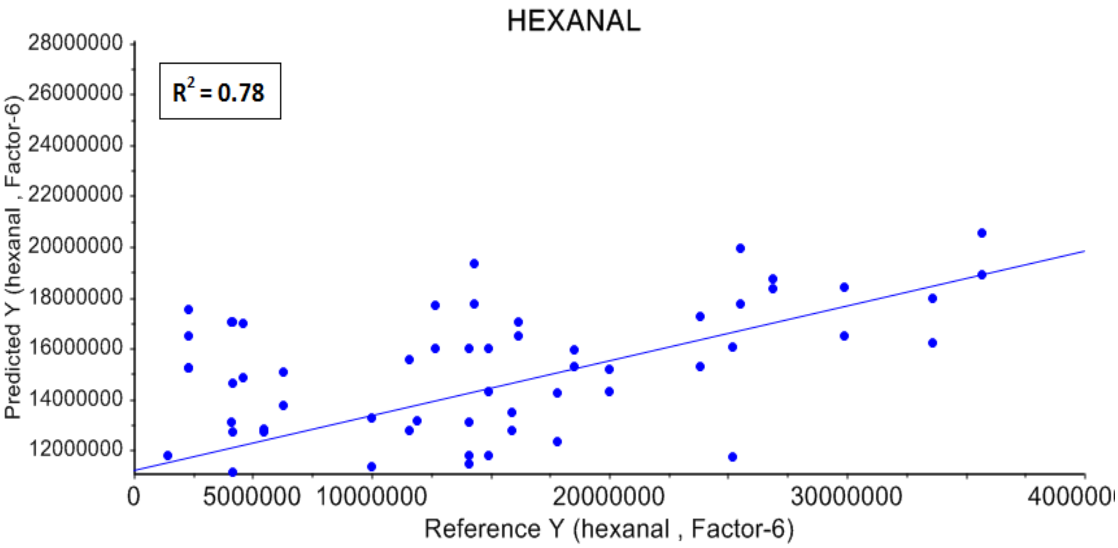


Figure 46 Predicted vs Reference (Hexanal, 6 factors).

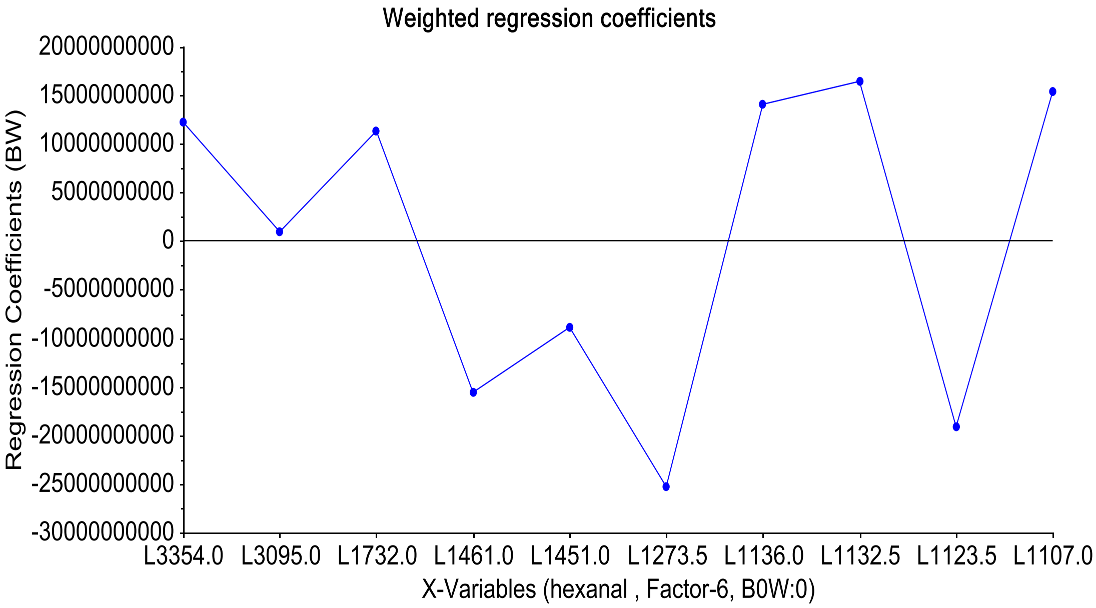
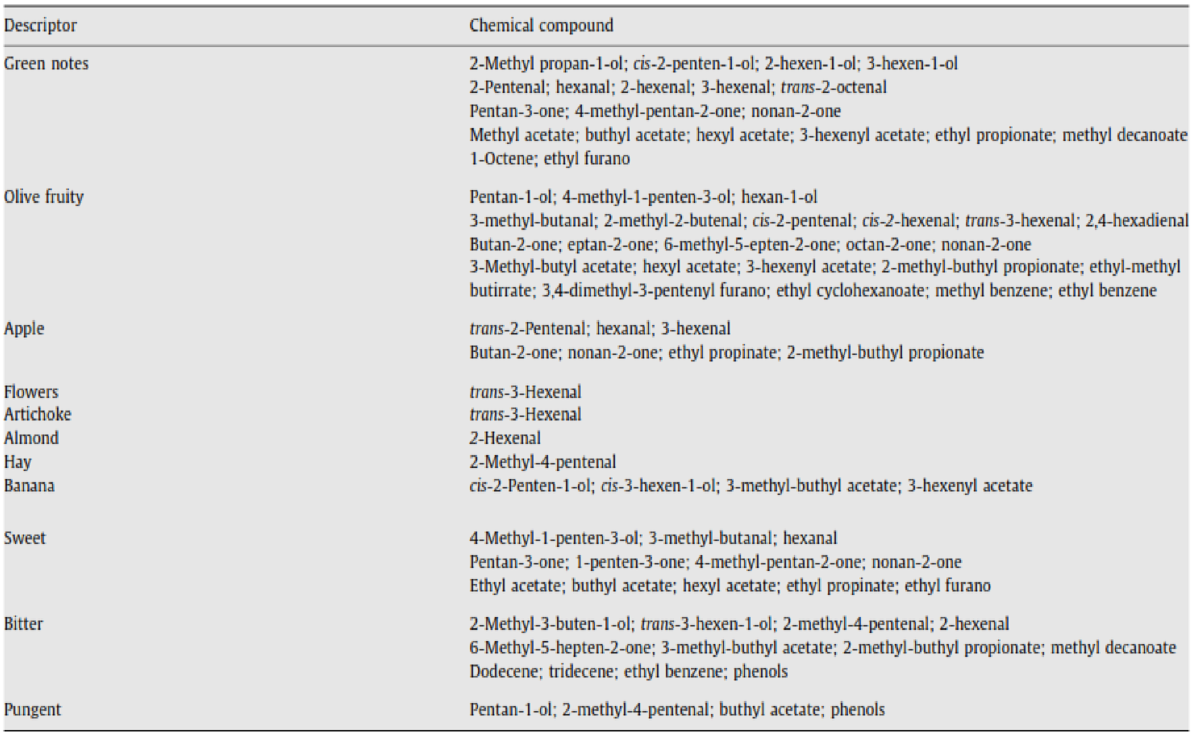


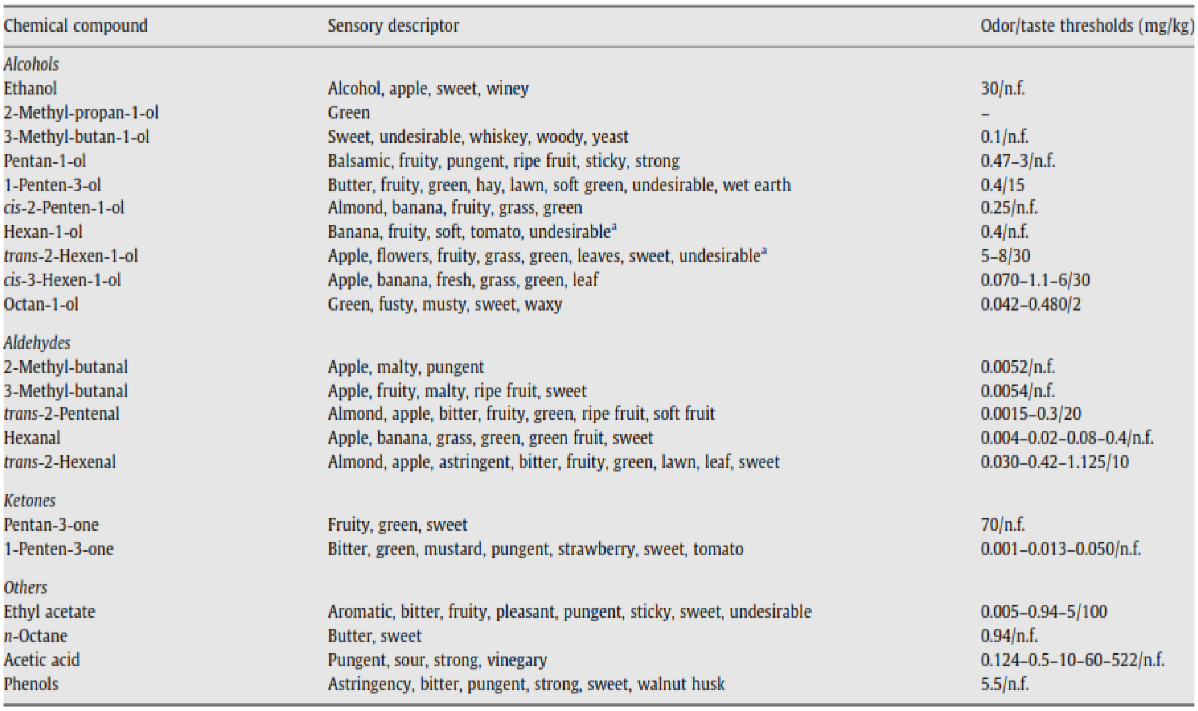
Figure 47 Weighted regression coefficients (Hexanal, 6 factors).

Table 26 Relationships among sensory notes and aromatic chemical compounds in olive oils found by several authors (Burdock, 2002; Kalua et al., 2007; Angerosa et al., 2004)



Descriptor	Chemical compound
Green notes	2-Methyl propan-1-ol; *cis*-2-penten-1-ol; 2-hexen-1-ol; 3-hexen-1-ol 2-Pentenal; hexanal; 2-hexenal; 3-hexenal; *trans*-2-octenal Pentan-3-one; 4-methyl-pentan-2-one; nonan-2-one Methyl acetate; buthyl acetate; hexyl acetate; 3-hexenyl acetate; ethyl propionate; methyl decanoate 1-Octene; ethyl furano
Olive fruity	Pentan-1-ol; 4-methyl-1-penten-3-ol; hexan-1-ol 3-methyl-butanal; 2-methyl-2-butenal; *cis*-2-pentenal; *cis*-2-hexenal; *trans*-3-hexenal; 2,4-hexadienal Butan-2-one; eptan-2-one; 6-methyl-5-epten-2-one; octan-2-one; nonan-2-one 3-Methyl-butyl acetate; hexyl acetate; 3-hexenyl acetate; 2-methyl-buthyl propionate; ethyl-methyl butirrate; 3,4-dimethyl-3-pentenyl furano; ethyl cyclohexanoate; methyl benzene; ethyl benzene
Apple	*trans*-2-Pentenal; hexanal; 3-hexenal Butan-2-one; nonan-2-one; ethyl propinate; 2-methyl-buthyl propionate
Flowers	*trans*-3-Hexenal
Artichoke	*trans*-3-Hexenal
Almond	2-Hexenal
Hay	2-Methyl-4-pentenal
Banana	*cis*-2-Penten-1-ol; *cis*-3-hexen-1-ol; 3-methyl-buthyl acetate; 3-hexenyl acetate
Sweet	4-Methyl-1-penten-3-ol; 3-methyl-butanal; hexanal Pentan-3-one; 1-penten-3-one; 4-methyl-pentan-2-one; nonan-2-one Ethyl acetate; buthyl acetate; hexyl acetate; ethyl propinate; ethyl furano
Bitter	2-Methyl-3-buten-1-ol; *trans*-3-hexen-1-ol; 2-methyl-4-pentenal; 2-hexenal 6-Methyl-5-hepten-2-one; 3-methyl-buthyl acetate; 2-methyl-buthyl propionate; methyl decanoate Dodecene; tridecene; ethyl benzene; phenols
Pungent	Pentan-1-ol; 2-methyl-4-pentenal; buthyl acetate; phenols

Table 27 Relationships among aromatic chemical compounds and sensory notes with odor and taste thresholds found by several authors (Burdock, 2002; Kalua et al., 2007; Morales et al., 2005)



Chemical compound	Sensory descriptor	Odor/taste thresholds (mg/kg)
Alcohols		
Ethanol	Alcohol, apple, sweet, winey	30/n.f.
2-Methyl-propan-1-ol	Green	–
3-Methyl-butan-1-ol	Sweet, undesirable, whiskey, woody, yeast	0.1/n.f.
Pentan-1-ol	Balsamic, fruity, pungent, ripe fruit, sticky, strong	0.47–3/n.f.
1-Penten-3-ol	Butter, fruity, green, hay, lawn, soft green, undesirable, wet earth	0.4/15
cis-2-Penten-1-ol	Almond, banana, fruity, grass, green	0.25/n.f.
Hexan-1-ol	Banana, fruity, soft, tomato, undesirable[a]	0.4/n.f.
trans-2-Hexen-1-ol	Apple, flowers, fruity, grass, green, leaves, sweet, undesirable[a]	5–8/30
cis-3-Hexen-1-ol	Apple, banana, fresh, grass, green, leaf	0.070–1.1–6/30
Octan-1-ol	Green, fusty, musty, sweet, waxy	0.042–0.480/2
Aldehydes		
2-Methyl-butanal	Apple, malty, pungent	0.0052/n.f.
3-Methyl-butanal	Apple, fruity, malty, ripe fruit, sweet	0.0054/n.f.
trans-2-Pentenal	Almond, apple, bitter, fruity, green, ripe fruit, soft fruit	0.0015–0.3/20
Hexanal	Apple, banana, grass, green, green fruit, sweet	0.004–0.02–0.08–0.4/n.f.
trans-2-Hexenal	Almond, apple, astringent, bitter, fruity, green, lawn, leaf, sweet	0.030–0.42–1.125/10
Ketones		
Pentan-3-one	Fruity, green, sweet	70/n.f.
1-Penten-3-one	Bitter, green, mustard, pungent, strawberry, sweet, tomato	0.001–0.013–0.050/n.f.
Others		
Ethyl acetate	Aromatic, bitter, fruity, pleasant, pungent, sticky, sweet, undesirable	0.005–0.94–5/100
n-Octane	Butter, sweet	0.94/n.f.
Acetic acid	Pungent, sour, strong, vinegary	0.124–0.5–10–60–522/n.f.
Phenols	Astringency, bitter, pungent, strong, sweet, walnut husk	5.5/n.f.

The success of this and other similar studies[XLVI] in the literature indicates that FTIR spectroscopy in combination with chemometric techniques have potential of predicting quality parameters of olive oil in a shorter analysis time compared to chromatographic techniques[XLVII.].This could be interesting for the industry, in fact rapid analyses of these chemical components would provide better control of quality during processing and storage and allow in determining the authenticity of the product.

PROCESSING UV-VIS SPECTRA

Since spectra obtained with 10% solutions possess absorbance greater than 2 in the UV zone, while spectra on 1% samples are found to have low absorbance in the visible, the two spectra were combined in order to respect the above-mentioned requirements. The spectra of all the samples were assembled into a matrix, which we call **X**, which shows the spectrum of a sample on a line. The spectra has the absorbance recorded for solutions at 10% at wavelengths between 690 and 285 nm and the values of absorbance of samples at 1% from 284 to 245 nm. We have decided to not consider the spectral readings between 244 and 190 nm because they are out of scale, even in diluted solutions. We build a model only with the data of year 2015.

The first investigation was the calculation of PCA (Principal Component Analysis) on the data after they had been column autoscaled. Column autoscaling needs because we have to mitigate scale and background differences due to the use of absorbances acquired in different experimental condition, that is concentration and spectra. The number of components was evaluated based on a Scree-plot. The score projection of PC1 vs PC3 (**Figure *48***) allows to highlight some degree of separation between the five varieties of oil, particularly it results well characterized, Mignola.

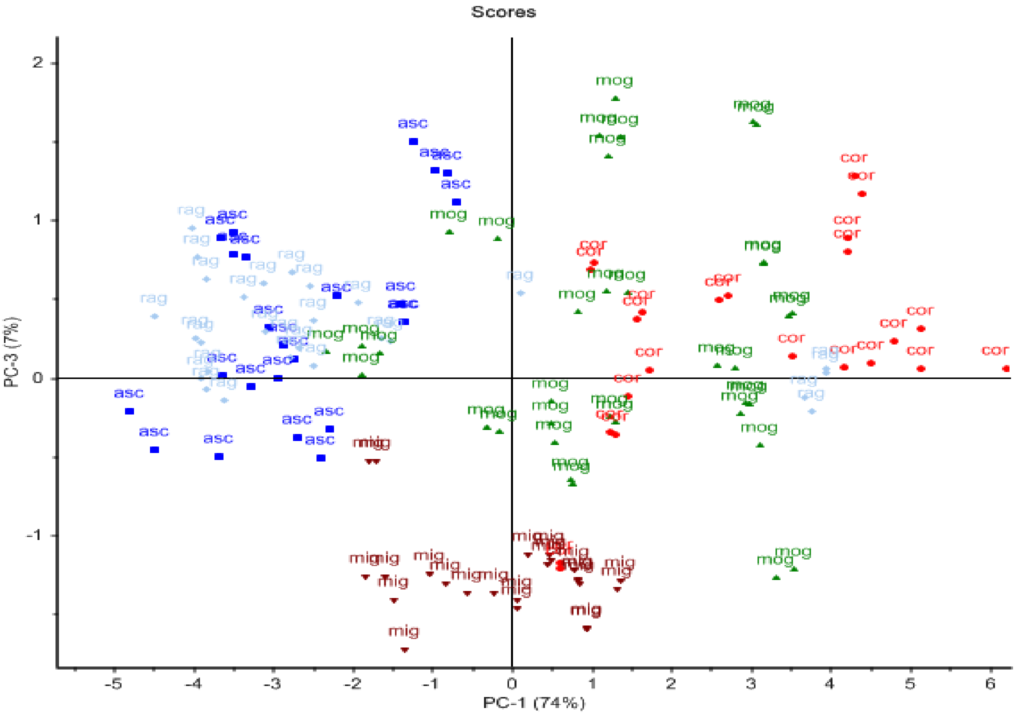


Figure 48 PCA Score plot after column autoscaling with PC1 (74%) vs PC3 (7%).

Spectrometric measures often have light diffusion phenomena that introduce noise that can worsen multivariate analysis. Several types of data pretreatment can be used to reduce this problem; one of the most used is the line autoscaling (SNV). We applied it separately to the data of the two combined spectra. The **Figure 49** shows the scores plot obtained after application of SNV followed by column autoscaling. We note the improvement in grouping ability of the homogeneous data, particularly evident for Mogliano, Raggia and Ascolana.

.

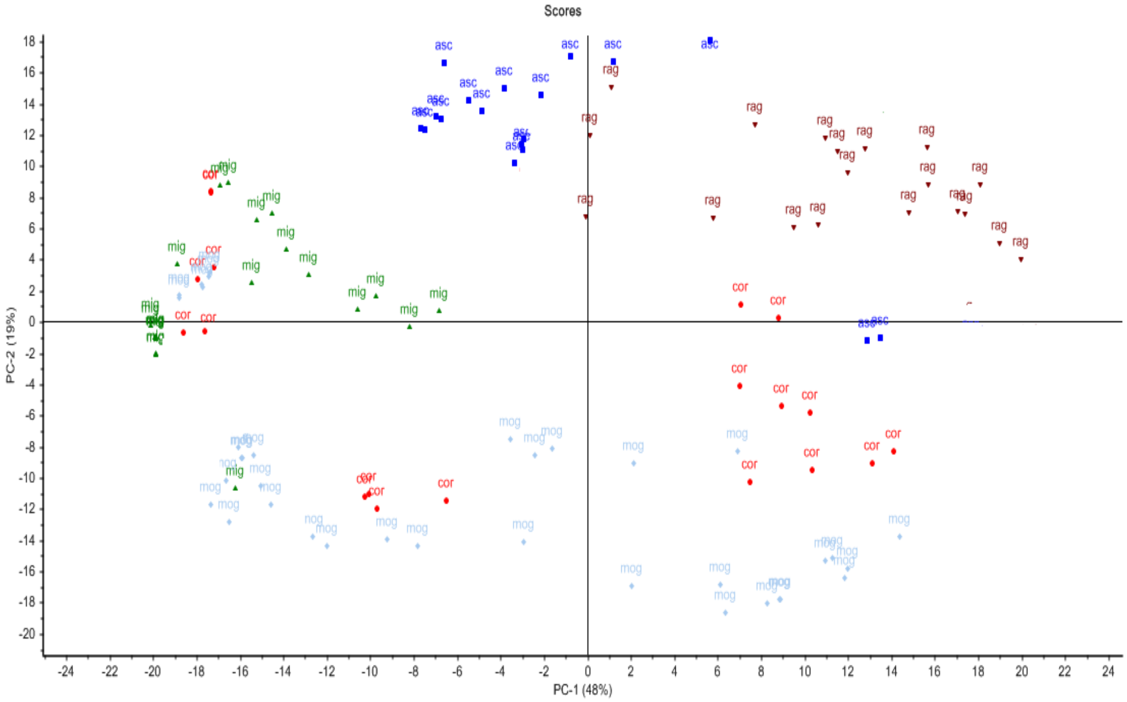


Figure 49 Score plot after SNV and column autoscaling with PC1 (48%) vs PC2 (19%).

To optimize the model, it was decided to do the selection of variables by using SELECT.

Accordingly from 450 variables, it has selected to 17 variables

432, 424 423, 421, 400, 331, 307,303,301, 291,284,283, 280, 278,277,265, 253

Name them UVselect

The wavelengths 432, 423, 424 421, 400, can be related to dyes and pigments. Several works[XLVIII] report the coloured pigments in the olive oil, have maximum absorption within spectral ranges [690-400]nm, in particular (410 phaeophytin a, 430 chlorophyll a, 435 phaeophytin b, 440 xanthophylls, 454 β-carotene, 466 and 650 chlorophyll b, 660-670 chlorophyll a and phaeophytin a).

The wavelengths **331**, **307**,**303**,**301**, **291**,284,**283**, **280**, **278**,**277**,**265**, **253** whereas can give us information about polyphenol compounds.

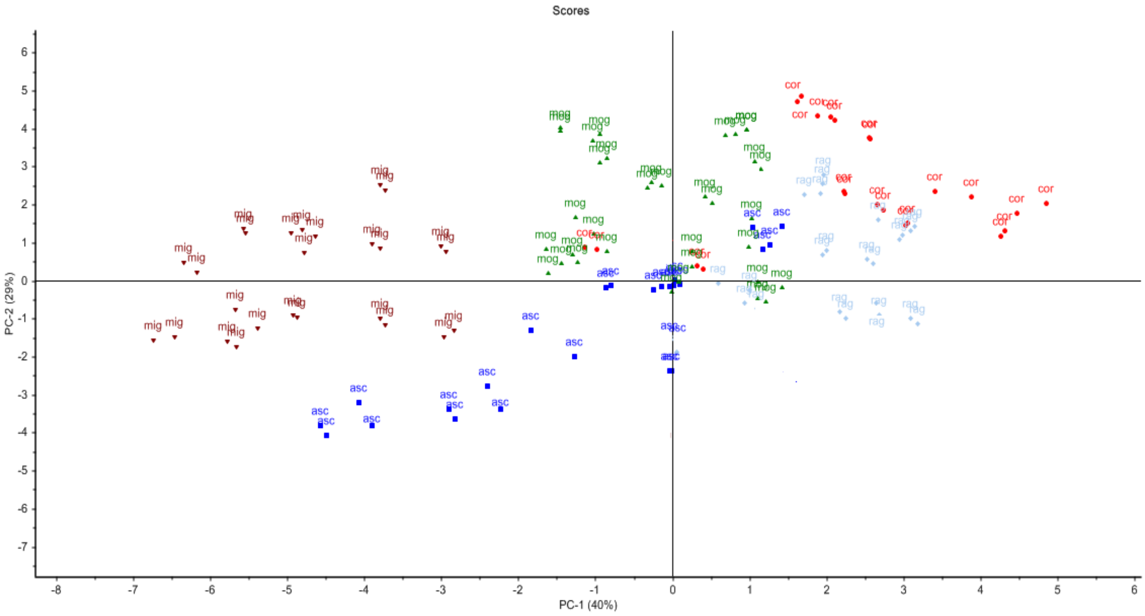


Figure 50 Score plot of 14 variables with PC1 (40%) vs PC2 (29%).

Figure 50 shows the projection on the first two PCs (retained variance 69%) obtained from UVselect pretreated as above. The variable reduction has not decreased the grouping ability of the calculated models.

Since my problem is to recognize the categories of Monovarietal extra-virgin olive oil from Marche, we try to classify them through modeling methods.

TRAINING SET AND EXTERNAL SET FOR MODELLING METHODS

The samples comes from dataset of several harvesting season. They were collected and measured in the autumn/winter of each of the years: 2009, 2010, 2015, and 2016. It is noteworthy to remind that, as explained above, measurements on samples of years 2009-2010 were performed in a different laboratory with a different instrument and procedure with respect to those of years 2015-2016. The class models have been built with a Training set of 92 objects (detail in **Table 28**), actually with 23 samples all of year 2015, because each sample present four replicas, in order to increase the number of datasets and consider the variability due to the analytical noise. It has been used two external set, “Ext-set1” (see **Table 29**) made of 2009 and 2010 objects plus some of 2015 objects for the categories Ascolana and Raggia, in total there are 170 objects. The samples of 2009 and 2010 are measured twice while those of 2015 have

4 replicas. “Ext-set2” (see ***Table* 30**) are samples of 2016 (172 objects), here each sample is present with 4 replicas. These two external set correspond to different analytical instrument, that could affect the results. This has been applied for both modeling methods(SIMCA and UNEQ).

Table 28 Samples used for the training set (92 objects)

Category	***Year (2015)***
Ascolana	16
Coroncina	16
Mignola	16
Mogliano	24
Raggia	20

Table 29 Samples used for the Ext-set1 (170 objects)

Category	**Year (2009)**	**Year (2010)**	**Year (2015)**
Ascolana	0	0	8
Coroncina	12	34	0
Mignola	14	28	0
Mogliano	14	52	0
Raggia	0	0	8

Table 30 Samples used for the Ext-set2 (172 objects)

Category	**Year (2016)**
Ascolana	32
Coroncina	28
Mignola	28
Mogliano	36
Raggia	48

MODELING TECHNIQUES QDA-UNEQ AND SIMCA

Classification and Prediction ability have been evaluated also with V-Parvus for modeling techniques QDA-UNEQ and SIMCA.

The SIMCA model here discussed were computed retaining three principal components in each class at 95% confidence interval.

The UNEQ model here discussed were computed retaining three principal components (autoscaled) in each class at 95% confidence interval.

This class model has been built with a Training set of 92 objects (**Table 28**); "Ext-set1" (**Table 29)** of 2009-2010 plus some of 2015 objects for the categories Ascolana and Raggia (170 objectss) and "Ext-set2" (**Table 30**) of 2016 (172 objects).

The result are reported below:

SIMCA

Table 31 and **Table 32** show the good sensitivity and even better specificity of the method for all the categories

Table 31 Sensitivity of Simca model

SENSITIVITIES MODEL (Level = 95%)		
Ascolana	81%	(Accepted 13/ 16)
Coroncina	81%	(Accepted 13/16)
Mignola	94%	(Accepted 15/ 16)
Mogliano	88%	(Accepted 21/ 24)
Raggia	80%	(Accepted 16/20)
Mean sensitivity 85%		

Table 32 Specificity of Simca model.

SPECIFICITIES MODEL (Level = 95 %)	
Ascolana	**Coroncina**
Coroncina 100% (Rejected 16/16)	Ascolana 100% (Rejected 16/16)
Mignola 100% (Rejected 16/16)	Mignola 100% (Rejected 16/16)
Mogliano 100% (Rejected 24/24)	Mogliano 100% (Rejected 24/24)
Raggia 100% (Rejected 20/20)	Raggia 100% (Rejected 20/20)
Mean specificity for Ascolana 100%	**Mean specificity for Coroncina 100%**
Mignola	**Mogliano**
Ascolana 75% (Rejected 16/16)	Ascolana 100% (Rejected 16/16)
Coroncina 100% (Rejected 16/16)	Coroncina 34% (Rejected 16 /16)
Mogliano 100% (Rejected 24/24)	Mignola 100% (Rejected 24/24)
Raggia 100% (Rejected 20/20)	Raggia 90% (Rejected 20/20)
Mean specificity for Mignola 100%	**Mean specificity for Mogliano 82%**
Raggia	
Ascolana 100% (Rejected 16/16)	
Coroncina 100% (Rejected 16/16)	

Mignola	100% (Rejected 16/16)
Mogliano	100% (Rejected 18/20)
Mean specificity for Raggia 100%	

The prediction, however, show trouble of the method, sensitivity (**Table 33**) of Coroncina, Mignola and Mogliano, the categories most present in Ext-set1, is poor with this external set but Specificity (**Table 34**) is still high. These results justify the mediocre efficiency (**Table 35)** for Coroncina ,Mignola and Mogliano. The apparently nice results of the Ascolana are obtained on few samples of external set a thing that weaken the result.

Table 33 Sensitivity of Ext-set1 Simca model.

SENSITIVITIES EXT-SET1 (Level = 95%)		
Ascolana	63%	(Accepted 5 / 8)
Coroncina	11%	(Accepted 5/46)
Mignola	15%	(Accepted 6/42)
Mogliano	0%	(Accepted 0/66)
Raggia	25%	(Accepted 2 / 8)
Mean sensitivity 11%		

Table 34 Specificity of Ext-set1 Simca model.

SPECIFICITIES EXT-SET1 (Level = 95 %)	
Ascolana	**Coroncina**
Coroncina 100% (Rejected 46/46)	Ascolana 100% (Rejected 8 / 8)
Mignola 100% (Rejected 42/42)	Mignola 100% (Rejected 42/42)
Mogliano 100% (Rejected 66/66)	Mogliano 100% (Rejected 66/66)
Raggia 100% (Rejected 8 / 8)	Raggia 100% (Rejected 8 / 8)
Mean specificity for Ascolana 98%	**Mean specificity for Coroncina 100%**
Mignola	**Mogliano**
Ascolana 100% (Rejected 8 / 8)	Ascolana 38% (Rejected 3 / 8)
Coroncina 100% (Rejected 42/42)	Coroncina 87% (Rejected 40/46)
Mogliano 100% (Rejected 66/66)	Mignola 81% (Rejected 34/42)
Raggia 100% (Rejected 8/ 8)	Raggia 75% (Rejected 6 / 8)
Mean specificity for Mignola 100%	**Mean specificity for Mogliano 80%**
Raggia	
Ascolana 100% (Rejected 8 / 8)	
Coroncina 96% (Rejected 44/46)	
Mignola 100% (Rejected 42/42)	
Mogliano 100% (Rejected 66/66)	
Mean specificity for Raggia 99%	

Table 35 Efficiency of Ext-set1 Simca model

EFFICIENCY EXT-SET1 (Level = 95%)	
Ascolana	79%
Coroncina	33%
Mignola	39%
Mogliano	0%
Raggia	50%

As we can see, SIMCA does work well , **Table 35** (Efficiency mean= 40%), but this can be justifiable by the fact that we built a model on 5 classes, and then we go to predict it with an external set that does not actually contain the same number of samples for all 5 classes, and so there is some kind of data disequilibrium. It is noteworthy that the most part of the samples present in Ext-set1 are those measured with a different instrument than that used for measuring the samples of the training set.

Table 36 Sensitivity of Ext-set2 Simca model.

SENSITIVITIES EXT-SET2 (Level = 95%)		
Ascolana	3%	(Accepted 1 / 32)
Coroncina	28%	(Accepted 3 /28)
Mignola	18%	(Accepted 6/32)
Mogliano	70%	(Accepted 22/32)
Raggia	8%	(Accepted 4/48)
Mean sensitivity 25.4%		

Table 37 Specificity of Ext-set2 Simca model.

SPECIFICITIES EXT-SET2 (Level = 95 %)	
Ascolana	**Coroncina**
Coroncina 100% (Rejected 28/28)	Ascolana 100% (Rejected 32 / 32)
Mignola 100% (Rejected 32/32)	Mignola 100% (Rejected 32 / 32)
Mogliano 100% (Rejected 32/32)	Mogliano 100% (Rejected 32 / 32)
Raggia 100% (Rejected 48/48)	Raggia 100% (Rejected 48 / 48)
Mean specificity for Ascolana 100%	**Mean specificity for Coroncina 100%**
Mignola	**Mogliano**
Ascolana 100% (Rejected 32/32)	Ascolana 63% (Rejected 20/32)
Coroncina 100% (Rejected 28/28)	Coroncina 57% (Rejected 16/28)
Mogliano 100% (Rejected 32/32)	Mignola 63% (Rejected 20/32)
Raggia 100% (Rejected 48/48)	Raggia 92% (Rejected 44/48)
Mean specificity for Mignola 100%	**Mean specificity for Mogliano 72%**
Raggia	
Ascolana 100% (Rejected 32 / 32)	
Coroncina 100% (Rejected 32 / 32)	
Mignola 100% (Rejected 32 / 32)	
Mogliano 100% (Rejected 32 / 32)	
Mean specificity for Coroncina 100%	

Table 38 Efficiency of Ext-set2 Simca model.

EFFICIENCY EXT-SET2 (Level = 95%)	
Ascolana	17%
Coroncina	53%
Mignola	42%
Mogliano	71%
Raggia	28%

When Ext-set2 is used we have an increase of the sensitivity (**Table 36)** with very high specificity (**Table 37**). The efficiency table (**Table 38**) highlight a good result for the category Mogliano while other categories still show unsatisfactory results. We built a model on 5 classes with a too small training set. This can be a weakness point for the robustness and stability of the SIMCA model.

UNEQ

Table 39 Sensitivity of Training set Uneq model.

SENSITIVITIES MODEL (Level = 95%)		
Ascolana	100%	(Accepted 16/ 16)
Coroncina	100%	(Accepted 16/16)
Mignola	100%	(Accepted 16/ 16)
Mogliano	100%	(Accepted 24/ 24)
Raggia	100%	(Accepted 20/20)
Mean sensitivity 100%		

Table 40 Specificity of Training set Uneq model.

SPECIFICITIES MODEL (Level = 95 %)	
Ascolana	**Coroncina**
Coroncina 25% (Rejected 4 /20)	Ascolana 88% (Rejected 14/16)
Mignola 63% (Rejected 10/16)	Mignola 75% (Rejected 12/16)
Mogliano 58% (Rejected 14/24)	Mogliano 100% (Rejected 24/24)
Raggia 90% (Rejected 18/20)	Raggia 65% (Rejected 13/20)
Mean specificity for Ascolana 61%	**Mean specificity for Coroncina 83%**
Mignola	**Mogliano**
Ascolana 75% (Rejected 12/16)	Ascolana 25% (Rejected 4 /16)
Coroncina 75% (Rejected 12/16)	Coroncina 0% (Rejected 0 / 16)
Mogliano 100% (Rejected 24/24)	Mignola 63% (Rejected 10/16)
Raggia 90% (Rejected 18/20)	Raggia 50% (Rejected 10/20)
Mean specificity for Mignola 87%	**Mean specificity for Mogliano 35%**
Raggia	
Ascolana 100% (Rejected 16 /16)	
Coroncina 0% (Rejected 0 /20)	
Mignola 75% (Rejected 12 /20)	
Mogliano 100% (Rejected 24/24)	
Mean specificity for Raggia 72%	

Table 39 and **Table 40** show the encouraging result of modelling the classes by UNEQ.

Table 41 Sensitivity of Ext-set1 Uneq model.

SENSITIVITIES EXT-SET1 (Level = 95%)		
Ascolana	75%	(Accepted 6 / 8)
Coroncina	28%	(Accepted 13 /46)
Mignola	26%	(Accepted 11/42)
Mogliano	97%	(Accepted 64/66)
Raggia	50%	(Accepted 4 / 8)
Mean sensitivity 58%		

Table 42 Specificity of Ext-Set1 Uneq model.

SPECIFICITIES EXT-SET1 (Level = 95 %)	
Ascolana	**Coroncina**
Coroncina 24% (Rejected 11/46)	Ascolana 63% (Rejected 5 / 8)
Mignola 21% (Rejected 9/46)	Mignola 45% (Rejected 19/42)
Mogliano 75% (Rejected 50/66)	Mogliano 44% (Rejected 29/66)
Raggia 0% (Rejected 0 / 8)	Raggia 25% (Rejected 2 / 8)
Mean specificity for Ascolana 45%	**Mean specificity for Coroncina 44%**
Mignola	**Mogliano**
Ascolana 63% (Rejected 5 / 8)	Ascolana 0% (Rejected 0 / 8)
Coroncina 45% (Rejected 19/42)	Coroncina 4% (Rejected 2 /46)
Mogliano 44% (Rejected 29/66)	Mignola 3% (Rejected 1 /42)
Raggia 100% (Rejected 8/ 8)	Raggia 0% (Rejected 8 / 8)
Mean specificity for Mignola 72%	**Mean specificity for Mogliano 3%**
Raggia	
Ascolana 38% (Rejected 3 / 8)	
Coroncina 74% (Rejected 34/46)	
Mignola 48% (Rejected 20/42)	
Mogliano 33% (Rejected 22/66)	
Mean specificity for Raggia 49%	

Table 43 Efficiency of Ext-Set1 Uneq model.

EFFICIENCY EXT-SET1 (Level = 95%)	
Ascolana	57%
Coroncina	35%
Mignola	43%
Mogliano	17%
Raggia	50%

The evaluation of prediction ability of the model performed on the Ext-set1 has some interesting results, in the case of the category Mogliano, there is a very high sensitivity=97% (**Table 41)**, but with a too inclusive model as shown by the low specificity (**Table 42**) and efficiency (**Table 43)**. Categories Ascolana and Raggia have a bit lower sensitivity, respectively 75% for Ascolana and 50% for Raggia, and even a poorer specificity. These results justify the mediocre efficiency in **Table 43**.

Table 44 Sensitivity of Ext-Set2 Uneq model.

SENSITIVITIES EXT-SET2 (Level = 95%)		
Ascolana	94%	(Accepted 30 / 32)
Coroncina	68%	(Accepted 19/28)
Mignola	13%	(Accepted 4/32)
Mogliano	88%	(Accepted 28/32)
Raggia	33%	(Accepted 16/48)
Mean sensitivity 56.4%		

Table 45 Specifity of Ext-Set2 Uneq model.

SPECIFICITIES EXT SET (Level = 95 %)	
Ascolana	**Coroncina**
Coroncina 7% (Rejected 2 /28)	Ascolana 50% (Rejected 16 / 32)
Mignola 6% (Rejected 2 /32)	Mignola 78% (Rejected 26/36)
Mogliano 13% (Rejected 4 /32)	Mogliano 38% (Rejected 12/32)
Raggia 17% (Rejected 8 /48)	Raggia 75% (Rejected 36 / 48)
Mean specificity for Ascolana 11%	**Mean specificity for Coroncina 62%**
Mignola	**Mogliano**
Ascolana 69% (Rejected 22/32)	Ascolana 0% (Rejected 0/32)
Coroncina 72% (Rejected 20/28)	Coroncina 0% (Rejected 0/28)
Mogliano 88% (Rejected 28/32)	Mignola 13% (Rejected 4 /32)
Raggia 92% (Rejected 44/48)	Raggia 0% (Rejected 0 /48)
Mean specificity for Mignola 81%	**Mean specificity for Mogliano 3%**
Raggia	
Ascolana 56% (Rejected 18/32)	
Coroncina 54% (Rejected 15 /28)	
Mignola 69% (Rejected 22/32)	
Mogliano 59% (Rejected 19/32)	
Mean specificity for Raggia 60%	

Table 46 Efficiency of Ext-set2 Uneq model.

EFFICIENCY EXT-SET2 (Level = 95%)	
Ascolana	32%
Coroncina	65%
Mignola	32%
Mogliano	16%
Raggia	45%

Prediction of EXT-set2 shows that the models works quite well to recognize samples of classes Ascolana (94%) , Coroncina (68%) and Mogliano (88%) (**Table 44**). It is interesting the Coroncina for which even specificity is moderately high (**Table 45**). The efficiency table (**Table 46)** shows that the best model is obtained for Coroncina.

RESULTS AND DISCUSSION

The non-selective information obtained from UV–VIS and FTIR spectroscopy allowed to build reliable models for the characterization of monovarietal extra-virgin olive oil from Marche.

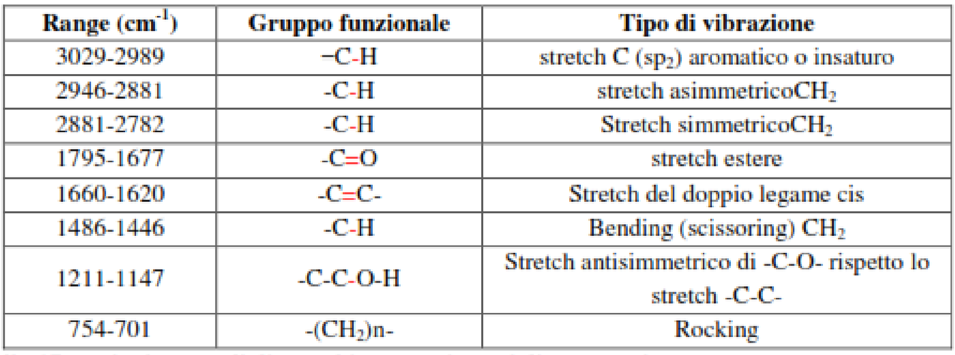
The class-modelling results obtained using the IR spectroscopic data were better than those obtained with the UV-VIS.

We were able, using IR spectra, to classify the varieties Coroncina, Mignola, Mogliano, Ascolana and Raggia. Select algorithm enabled us to choose 10 wavelengths of the spectra with which to classify the oil varieties, really models of classes. These 10 variables can be used both with SIMCA and UNEQ modelling methods. The prediction ability of the methods has been investigated with a large external set including samples of several years. In spite of this large test UNEQ modelling method permits to recognize quite well all samples belong to the 5 categories (average Sensitivity=88%). The method show also moderate specificity, in particular the category models Coroncina, Mogliano and Raggia succeed to reject the majority of samples belong to the category Mignola and the samples of the other of these categories. Coroncina and Mogliano show low specificity among them.

SIMCA modelling method performs worse in the same condition than UNEQ with 5 categories. We used as training set the most part of the samples of year 2015 were there are 5 classes, however two of these classes Raggia and Ascolana were not present in the data of the years preceding 2015. The external set have unbalanced number of objects that can negatively affect the results. To check this we made a training set including only the samples of year 2015 belonging to the categories: Coroncina, Mignola and Mogliano. The prediction results showed an improvement of both Sensitivity and Specificity. While the sensitivity is high both with SIMCA and UNEQ, the first method performs much better with respect to specificity. Even if specificity is high for all three categories it is noteworthy the case of Coroncina.

It is interesting, from a chemical point of view, to identify which compounds are responsible for varietal differentiation.

In spectrophotometric IR analysis by superimposing spectra is clear that all of them are strongly characterized by a spectral profile common to all samples, dominated by highly absorbed areas attributable to different functional groups[XLIX], reported below:



Range (cm^{-1})	Gruppo funzionale	Tipo di vibrazione
3029-2989	−C-H	stretch C (sp_2) aromatico o insaturo
2946-2881	-C-H	stretch asimmetricoCH_2
2881-2782	-C-H	Stretch simmetricoCH_2
1795-1677	-C=O	stretch estere
1660-1620	-C=C-	Stretch del doppio legame cis
1486-1446	-C-H	Bending (scissoring) CH_2
1211-1147	-C-C-O-H	Stretch antisimmetrico di -C-O- rispetto lo stretch -C-C-
754-701	-(CH_2)n-	Rocking

Table 47 Spectral absorption regions and assignment hypotheses.

The hypotheses of recognition of the different functional groups are justified and in line with the chemical nature of the basic composition of each extra virgin olive oil, which, regardless of their origin and variety, is predominantly composed of TAG made up of fatty acids common to the whole family of extra virgin olive oil: oleic palmitic, palmitoleic, linoleic, which show the features attributed to the table.

However, as the results obtained through the processing show, there is a difference in the composition able, through the developed model, to characterize the samples depending on variety. It is difficult, however, to be able to clearly define if and what are the specific features, and therefore the families of compounds responsible for differentiation of variety and in relation also to the fact that some of the variables selected through SELECT identify characteristic wavelengths of the fingerprint area.

Subsequently it was aimed at determining some correlation between these 10 variables IR spectroscopic and volatiles, phenols and sensorial analysis of virgin olive oils by using Partial least square calibration models.

Interestingly, Bitter note is correlated to 2-Hexenal, in fact it is characterized by the same IR variables (1273.5, 1123.5 cm^{-1}) of the 2-hexenal.

Bitter note is also influenced by phenols. In fact it is correlated to the Pinoresinal compound by the variable (1123.5 cm^{-1}).

For the same reason of common Ir variables (1273.5, 1132.5 cm-1) we can also see a correlation between Hexanal compound and Fruit note.

The phenolic compounds (P-Coumaric and Pinoresinol) have one IR variable in common (1451 cm^{-1}) are related to the alkylic chains, in particular can be assigned to the scissoring vibration of the CH_2 groups.

The success of this and other similar studies[L] in the literature indicates that FTIR spectroscopy in combination with chemometric techniques have potential of predicting other quality parameters of olive oil in a shorter analysis time compared to chromatographic techniques[LI].This could be interesting for the industry, in fact rapid analyses of these chemical components would provide better control of quality during processing and storage and also help in determining the authenticity of the product.

The modelling of oil types by SIMCA and UNEQ was attempted even with UV-Vis spectra. The training set consists of samples of year 2015. The elaborations produced model able to classify with good results the five varieties of extra-virgin olive oil. The prediction efficiency, estimated by 5 deletion groups cross-validation, for SIMCA and UNEQ models is, respectively, 90% and 85%. Prediction estimated by external set has good sensitivity for all categories with both methods (SIMCA and UNEQ) whereas the specificities are poor. In fact in SIMCA method only the category Mogliano succeeds to recognize well all samples and exclude most of those not belonging to it; whereas in UNEQ only the category Coroncina do.

The discussed data treatment results were obtained on a subset of 17 wavelengths of the UV-vis spectra selected by the Select algorithm. It is interesting to analyse which compounds affect the absorption at the wavelength selected for the modelling of oil types.

The wavelengths **432**, **423**, **424, 421**, **400**, can be related to dyes and pigments. Several works [LII] report the coloured pigments in the olive oil, have maximum absorption within spectral ranges [690-400]nm, in particular (410 phaeophytin a, 430 chlorophyll a, 435 phaeophytin b, 440 xanthophylls, 454 β-carotene, 466 and 650 chlorophyll b, 660-670 chlorophyll a and phaeophtin a).

As can be seen from the data reported above, however, it is not possible to discriminate or advance hypotheses about the most interesting and influential pigment family, since both families (Carotenoid and chlorophylls), show concatenated absorption and even often common in the ranges examined.

The wavelengths **331**, **307**, **303**, **301**, **291**, 2**84**, **283**, **280**, **278**, **277**, **265**, **253** can give information about polyphenol compounds.

The composition of olive oil is influenced by many factors, such as olive varieties or cultivar, soil composition and climate, the procedure for collection of olives, pesticide treatments, month of collection, production technology, transport and storage conditions before and after milling. Some of these factors are related to the specific production area while others cause a variation between oils of the same location as, for instance, month of collection, year of production, ripening or harvesting. In order to capture this variation and to build stable and strong models, this study could be extended with a higher number of samples, because it would be very helpful in safeguarding and promoting monovarietals olive oils in our territory.

RINGRAZIAMENTI

Dopo due anni intensi, finalmente il grande giorno è arrivato!!! È stato un periodo di profondo apprendimento, non solo a livello scientifico, ma anche personale. Vorrei spendere due parole di ringraziamento nei confronti di tutte le persone che mi hanno sostenuto e aiutato durante questo periodo.

Prima di tutto, vorrei ringraziare i miei genitori, con il loro dolce e instancabile sostegno, sia morale che economico, mi hanno permesso di arrivare fin qui davanti a voi oggi, contribuendo alla mia formazione personale. GRAZIE INFINITE!

Vorrei ringraziare il prof. P. Conti, relatore di questa tesi di laurea, oltre che per l'aiuto fornitomi e la grande conoscenza che mi ha donato nel campo della chemiometria, per la disponibilità e umanità dimostratemi durante tutto il periodo di stesura. La stimo molto!

Alle mie amate sorelle, sempre pronte ad ascoltarmi e a darmi consigli. A cercare in ogni occasione di far salire la mia autostima, insegnandomi a camminare ogni giorno a testa alta senza aver paura dei giudizi degli altri. Ogni volta che ho bisogno di loro, nonostante i chilometri che ci separano, sono sempre presenti. Vi Amo!

Un grazie a Martina, la migliore compagna di avventure che abbia mai avuto in questi anni perché è sempre stata presente senza essere né invadente né inopportuna e che mi ha sopportata e apprezzata per come sono. Un'Amica sincera e leale a cui posso confidare senza "riserve" i miei pensieri e le mie emozioni. Sei Speciale!

Un ringraziamento di cuore va anche alla mia Mariangela, amica di sempre, entrata nella mia vita circa due anni prima di concludere il mio primo traguardo di studi (…Ingegneria…), sono stati gli anni più duri ma anche i più belli e intensi. Anche se ormai le occasioni di vedersi sono sempre minori, ogni volta che ci sentiamo o vediamo è come se il tempo si fosse fermato per noi .Grazie per far parte della mia vita e per essere presente in tutte le occasioni. Ti Adoro!

Per ultimi ma non meno importanti, ringrazio tutti i miei amici (vicini e lontani)che hanno avuto un peso determinante nel conseguimento di questo risultato, punto di arrivo e contemporaneamente di partenza della mia vita. Grazie per aver condiviso con me questa esperienza importante della mia vita. Vi Voglio Bene!

BIBLIOGRAFIA

[I] European Commission, Working paper of the directorate-general for agriculture: the olive oil and table olives sector, http://ec.europa.eu/agriculture/ markets/olive/reports/rep en.pdf.

[II] F. Marini, R. Bucci, A.L. Magrì, A.D. Magrì, in: V.R. Preedy, R.R. Watson (Eds.), Olives and Olive Oil in Health and Disease Prevention, Elsevier, Amsterdam, The Netherlands, 2010, pp. 569–579.

[III] European Commission, Agriculture and rural developments fact-sheet 2003,1–6

[IV] European Commission, Regulation (EC) no. 2081/1992, Off. J. Eur. Union L208 (1992)

[V] European Commission, Regulation (EC) no. 2082/1992, Off. J. Eur. Union L208 (1992) 9.

[VI] European Commission, Regulation (EC) no. 510/2006, Off. J. Eur. Union L93 (2006) 12–25

[VII] European Commission, Regulation (EC) no. 882/2004, Off. J. Eur. Union L165 (2004)1-141.

[VIII] M. Forina, S. Lanteri, C. Armanino, Top. Curr. Chem. 141 (1987) 91–143.

[IX] Siti del progetto "oli monovarietali italiani" http://www.olimonovarietali.it/database

[X] *Pannelli g., alfei b., santinelli a., 2001. "varietà di olivo nelle marche", assam. nuova edizione.*

[XI] Anon Journal of the American Oil Chemist Society 75, 673 681

[XII] Tous, J. and L. Ferguson. 1996. Mediterranean fruits. p. 416-430. In: J. Janick (ed.), Progress in new crops. ASHS Press, Arlington, VA.

[XIII] Nutritional composition of Mediterranean crops (per 100 g of edible portion). Source: Goulart (1980); Sawaya et al. (1983); Fernandez Diez (1983); IBPGR (1986); Morton (1987); Cantwell (1994).

[XIV] A.K. Kiritsakis Flavor Components of Olive Oil -- A Review, School of Food Technology and Nutrition, Technological Educational Institution (TEI) of Thessaloniki, Sindos Thessaloniki, Greece

[XV] Aparicio R., "Characterisation: mathematical procedures for chemical analysis" In Handbook of Oli e Oil: Analysis and Properties; Harwood J., Aparicio, R., Eds.; Aspen Publications: Gaithersburg, MD, 2000

[XVI] Vichi S., Pizzale L., Conte,L. S., Buxaderas S., Lopez-Tamames E., "Solid-phase microextraction in the analysis of virgin olive oil volatile fraction: characterization of virgin olive oils from two distinct geographical areas of northern Italy" J. Agric. Food Chem. 51 (2003) 6572 6577

[XVII] Angerosa F., Basti C., Vito R., "Virgin olive oil volatile compounds from lipoxygenase pathway and characterization of some Italian cultivars" J. Agric. Food Chem. 47 (1999) 836-839.

[XVIII] L. Matos, S. Cunha, J. Amaral, J. Pereira, P. Andrade, R. Seabra, B. Oliveira, 102 (2007Food Chemistry 406–414

[XIX] Lerma-García, María Jesús. Concha-Herrera, Victoria. Herrero-Martínez, José Manuel. Simó-Alfonso, Ernesto Francisco, "Classification of Extra Virgin Olive Oils Produced at La Comunitat Valenciana According to Their Genetic Variety Using Sterol Profiles Established by High-Performance Liquid Chromatography with Mass Spectrometry Detection", J. Agric. Food Chem. 57 (2009) 10512-10517

[XX] C. Oliveros, R. Boggia, R. Casale, M. Casale, C. Armanino, M. Forina "Optimisation of anew headspace mass spectrometry instrument. Discrimination of different geographicalorigin olive oils" Journal of Chromatography A 1076(1–2) (2005) 7–15

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

[XXII] María E. Escuderos, Sebastián Sánchez, Antonio Jiménez; "Quartz Crystal Microbalance (QCM) sensor arrays selection for olive oil sensory evaluation" Food Chemistry 124 (2011) 857–862

[XXIII] Forina, M., Boggia, R., & Casale, M. (2007). "The information content of visible spectra oextra virgin olive oil in the characterization of its origin" Annali di Chimica, 97(8), 615–693

[XXIV] Tapp, Henri S. Defernez, Marianne. Kemsley, E Katherine. “FTIR Spectroscopy and Multivariate Analysis Can Distinguish the Geographic Origin of Extra Virgin Olive Oils”J. Agric. Food Chem. 51(2003) 6110-6115

[XXV] F. A. Settle "Handbook of instrumental techniques for analytical chemistry" Prentice Hall PTR 1997

[XXVI] Rouessac, F. & Rouessac, A. Chemical Analysis: Modern Instrumentation Methods and Techniques. Second Edition Chemical Analysis Modern Instrumentation Methods and Techniques 2nd ed (Wiley 2007)-Francis Rouessac, Annick Rouessac.pdf John Wiley, 2007, 1 , ISBN 978-0-470-85902-5 (cloth : alk. paper) ; ISBN 978-0-470-85903-2 (pbk. : alk. paper)

[XXVII] D. Luthria, S. Mukhopadhyay, R. Robbins, J. Finley, G. Banuelos, J. Harnly J. Agric. Food Chem. 56 (2008)5457–5462

[XXVIII] Savitsky, A. Golay M. J. E., “Smoothing and Differentiation of Data by Simplified Least. Squares Procedures.” Anal. Chem., 36 (1964) 1627-1639, DOI: 10.1021 /ac60214a047

[XXIX] Steinier, J. Termonia, Y. Deltour, Jules “Smoothing and differentiation of data by simplified least square procedure” Anal. Chem. 44(1972) 1906-1909, DOI: 10. 1021 /ac60319a045

[XXX] Harald Martens, Tormod Naes, “Multivariate Calibration” John Wiley&Sons 1989

[XXXI] Richard G. Brereton, “Chemomerics data analysis for the laboratory and chemical plant” John Wiley&Sons 2003

[XXXII] Bauer, R., Nieuwoudt, H., Bauer, F. F., Kossmann, J., Koch, K. R., & Esbensen, K. H. (2008). FTIR spectroscopy for grape and wine analysis. Analytical Chemistry, 80, 1371e1379.

[XXXIII] Muik, B., Lendl, B., Molina-Díaz, A., Perez-Villarejo, L., & Ayora-Canada, M. J. (2004). Determination of oil and water content in olive pomace using near infrared and Raman spectrometry. A comparative study. Analytical and Bioanalytical Chemistry, 379,35e41.

[XXXIV] Ståhle, L. & Wold, S. Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study J. Chemom., John Wiley & Sons, Ltd., 1987, 1, 185-196 , 10.1002/cem.1180010306

[XXXV] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.L. Lewi, J. Smeyers-Verbeke,
Handbook of Chemometrics and Qualimetrics: Part B, Elsevier, Amsterdam, 1998, p. 213.

[XXXVI] S. Wold, M. Sjostrom, in: B.R. Kowalski (Ed.), Chemometrics, Theory and Application, ACS Symposium Series 52, American Chemical Society, Washington, DC, 1977, pp. 243–255.

[XXXVII] M. Forina, S. Lanteri, M. Casale, M.C. Cerrato Oliveros, Chemom. Intell. Lab. Syst. 87 (2007) 252–261.

[XXXVIII] The Unscrambler 9.8, Camo Software AS, 1986-2008

XXXIX Forina, M., Lanteri, S., Armanino, C., Casolino, C., & Casale, M. 2007. V-PARVUS: "An extendable package of programs for explorative data analysis, classification and regression analysis" Dip. Chimica e Tecnologie Farmaceutiche ed Alimentari, University of Genova. Free available at <http://www.parvus.unige.it>

[XL] B.R. Kowalski,C.F.Bender,PatternRecognit.8(1976)1–10.

[XLI] Barros, A., Freire, I., Gonçalves, B., Bacelar, F., Gomes, S., Lopes, J., et al. (2012). Evaluation of chemical and phenotypic changes in Blanqueta, Cobrançosa and Galega during ripening. CYTA – Journal of Food. http://dx.doi.org/10.1080/19476337.2012.705329.

[XLII] INE (2012). Inquérito à produção de azeite. International Olive Oil, key-figures on the world market for olive oils, Data adopted at the 100th session of the IOOC (Madrid, Spain), 19–23 November 2012).

[XLIII] INE (2012). Inquérito à produção de azeite. International Olive Oil, key-figures on the world market for olive oils, Data adopted at the 100th session of the IOOC (Madrid, Spain), 19–23 November 2012).

[XLIV] Bassi, D., Failla, O., Pedo, S., Tura, D., Mariani, L., Minelli, R., Geuna, F., Gigliotti, C.,2003. The olive cultivars in Lombardy: agronomic proprieties and oil characteristics. Quaderni della ricerca no. 25, Regione Lombardia, Italy. http://www.agricoltura.regione.lombardia.it/sito/tmpl_action.asp?DocumentoId=1292&SezioneId=2812000000&action=Documento

[XLV] Bassi, D., Tura, D., Geuna, F., Failla, O., Pedo, S., 2002. Characterisation of local olive (Olea europea L.) accessions by oil composition, morphological and molecular markers methods. Acta Hortic. 586, 57–60.

[XLVI] Oguz Uncu, Banu Ozen, Prediction of various chemical parameters of olive oils with Fourier transform infrared spectroscopy, Izmir Institute of Technology, Department of Food Engineering, Urla-Izmir, Turkey.

[XLVII] Irene Gouvinhas, José M.M.M. de Almeida, Teresa Carvalho , Nelson Machado, Ana I.R.N.A. Barros Discrimination and characterisation of extra virgin olive oils from three cultivars in different maturation stages using Fourier transform infrared spectroscopy in tandem with chemometrics. CITAB – CITAB, University of Trás-os-Montes and Alto Douro, 5001-801 Vila Real, Portugal. Department of Physics, University of Trás-os-Montes and Alto Douro, 5000-801 Vila Real, Portugal.

[XLVIII] L. Cerretani, M. J. Motilva, M. P. Romero, A. Bendini, G. Lercker Eur Food Research oTechnology 226 (2008):1251–1258

[XLIX] M. J. L. Garcia, G. R. Ramos, J. M. H. Martinez, E. F. S. Alfonso, Food Chemistry 118 (2010) 78-83.

[L] Oguz Uncu, Banu Ozen, Prediction of various chemical parameters of olive oils with Fourier transform infrared spectroscopy, Izmir Institute of Technology, Department of Food Engineering, Urla-Izmir, Turkey.

[LI] Irene Gouvinhas, José M.M.M. de Almeida, Teresa Carvalho , Nelson Machado, Ana I.R.N.A. Barros Discrimination and characterisation of extra virgin olive oils from three cultivars in different maturation stages using Fourier transform infrared spectroscopy in tandem with chemometrics. CITAB – CITAB, University of Trás-os-Montes and Alto Douro, 5001-801 Vila Real, Portugal. Department of Physics, University of Trás-os-Montes and Alto Douro, 5000-801 Vila Real, Portugal.

[LII] L. Cerretani, M. J. Motilva, M. P. Romero, A. Bendini, G. Lercker Eur Food Research oTechnology 226 (2008):1251–1258.