UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Agronomia Animali Alimenti Risorse Naturali
E Ambiente
Dipartimento di Medicina Animale, Produzioni E Salute

Corso di Laurea Magistrale in SCIENZE E TECNOLOGIE ANIMALI

Assessment of Italian honey origin using Near InfraRed Spectroscopy and multivariate modeling

Relatore
Prof. Severino Segato

Correlatore
Dott.ssa Sara Khazzar

Laureanda
Silvia Zanotto
Matricola n. 2091606

Anno Accademico 2023/2024

INDEX

RIASSUNTO

Il miele è un dolcificante sintetizzato naturalmente dalle api grazie alla fioritura di fiori e piante, le cui proprietà nutrizionali dipendono unicamente dalle stesse origini floreali. Può fornire vari benefici per la salute umana, ponendo il suo valore di mercato in posizioni più favorevoli rispetto ad altri dolcificanti. Di conseguenza, il miele è stato bersaglio di adulterazione, rendendo la sua autenticità una preoccupazione per i ricercatori e i produttori di tutto il mondo.

Il più antico e popolare metodo per la determinazione dell'origine floreale e geografica del miele è la melissopalinologia, eseguita analizzando il polline contenuto nel miele. Tuttavia, questa tecnica comporta diverse limitazioni in termini di durata prolungata delle analisi e di costi elevati, oltre che alla necessità di specialisti per la sua esecuzione. Infine, essa non è in grado di rilevare la contaminazione fraudolenta del polline.

Negli ultimi decenni la *Near InfraRed (NIR) Spectroscopy (NIRs)* è stata ampiamente applicata nell'agricoltura e nell'industria alimentare rivelandosi una valida tecnica analitica esplorativa e predittiva per determinare la qualità dei prodotti. Il NIR offre numerosi vantaggi in termini di prestazioni essendo un metodo rapido, non distruttivo, affidabile, economico e che non richiede la preparazione preliminare dei campioni.

Lo scopo di questa tesi di laurea è stato quello di valutare la fattibilità dell'utilizzo della spettroscopia NIR come strumento analitico rapido per determinare l'origine geografica di un pool di campioni di miele millefiori prodotto nel contesto italiano. I campioni (n = 227) di miele sono stati classificati in base alle loro regioni e macroaree di produzione: *SL = South Lowland*, al di sotto dei 600 m sul livello del mare (slm); *NM = North Mountain*, sopra i 600 m slm; *NL = North Lowland*, al di sotto dei 600 m slm.

I campioni sono stati analizzati utilizzando tre strumenti portatili: uno spettrofotometro operante nella regione del visibile (VIS), uno spettrofotometro VIS-NIR e uno spettrofotometro NIR. I dati spettrali raccolti dai tre strumenti analitici VIS, VIS-NIR e NIR sono stati pre-elaborati e utilizzati per definire quattro modelli di classificazione, quali *Random Forest, K-Nearest Neighbor (KNN), Support Vector Machine (SVM) linear e SVM radial*, applicando algoritmi di machine learning (ML).

I dati sono stati divisi casualmente in un *training test* (70% dei dati, n = 162), necessario per l'addestramento dei quattro modelli di classificazione, e un *test set* (30% dei dati, n = 65), dove si è valutata la loro performance predittiva mediante l'utilizzo di matrici di confusione.

I risultati mostrano che l'approccio multi-strumentale, che combina spettroscopia NIR e analisi colorimetrica, presenta una scarsa accuratezza (0.49) nel distinguere l'origine del miele millefiori da regioni specifiche, ma si rivela più efficace nel classificare il miele in aree geografiche più ampie, con un'accuratezza del 0.58. Le basse performance dei modelli predittivi sono probabilmente dovute all'elevata variabilità intrinseca del prodotto miele millefiori, che è influenzata da molteplici fattori come ad esempio la composizione chimico-fisica, il contenuto minerale e le caratteristiche colorimetriche. Queste condizioni possono variare anche all'interno della stessa regione a causa della diversità botanica e di fioritura delle specie vegetali interessate dalle api, dal variare delle condizioni climatiche e delle pratiche di apicoltura. Inoltre, essendo le aree geografiche suddivise per altitudine, possono comprendere territori più o meno vasti, contribuendo così ad una maggiore variabilità intrinseca. L'uso di strumenti portatili, con un range di lunghezza d'onda limitato, ha posto maggiori sfide, rendendo più difficile una significativa distinzione dell'origine del miele.

I risultati suggeriscono che, nonostante la spettroscopia NIR e approcci di *multivariate modeling* siano largamente utilizzati nell'industria alimentare, ulteriori miglioramenti tecnologici sono necessari al fine di sviluppare un sistema analitico integrato capace di autenticare e discriminare in modo accurato l'origine geografica del miele millefiori italiano.

ABSTRACT

Honey is a natural sweetener synthesized by bees thanks to the blooming of flowers and plants, whose nutritional properties depend solely on the same floral origins. It can provide various benefits for human health, which places honey's market value in a better position than other sweeteners. As a result, honey has always been adulterated, making its authenticity a concern for researchers and producers worldwide.

The most ancient and popular method for determining the floral and geographical origin of honey is melissopalynology, performed by analyzing the pollen contained in honey. However, this method has several limitations in terms of both the lengthy duration of the analyses and high costs, as well as the need for specialists for its execution. Finally, it is unable to detect fraudulent pollen contamination.

In recent decades, Near InfraRed (NIR) Spectroscopy (NIRs) has been widely applied in agriculture and the food industry, revealing to be a valid predictive and explorative analytical technique to determine the quality of products. NIR offers numerous performance advantages being a fast, non-destructive, reliable, cost-effective method and does not require preliminary preparation of samples.

The purpose of this thesis is to evaluate the feasibility of using Near-Infrared (NIR) spectroscopy as a rapid analytical tool to determine the geographical origin of a pool of polyflower honey samples produced in the Italian context. The samples (n = 227) of honey were classified according to their regions and macro-areas of production: SL = South Lowland, below 600 m above sea level (asl); NM = North Mountain, above 600 m asl; NL = North Lowland, below 600 m asl.

The samples were analyzed using three portable instruments: a visible spectrophotometer (VIS), a VIS-NIR spectrophotometer and a NIR spectrophotometer. The spectral data collected by the three analytical instruments VIS, VIS-NIR and NIR were pre-processed and used to define four classification models, such as Random Forest, K-Nearest Neighbor (KNN), Support Vector Machine (SVM) linear and SVM radial, applying machine learning (ML) algorithms.

The data were randomly divided into a training test (70% of the data, n = 162), necessary for the training of the four classification models, and a test set (30% of the data, n = 65), where their predictive performance was evaluated using confusion matrices. The results show that the multi-instrumental approach, which combines NIR spectroscopy and colorimetric analysis, has a poor accuracy (0.49) in distinguishing the origin of polyflower honey from specific

regions, but is more effective in classifying honey in wider geographical areas, with an accuracy of 0.58. The low performance of the predictive models is probably due to the high intrinsic variability of the polyflower honey product, which is influenced by multiple factors such as the chemical-physical composition, mineral content, and colorimetric characteristics. These conditions can also vary within the same region due to the botanical and flowering diversity of the plant species affected by bees, varying climatic conditions and beekeeping practices. Moreover, since geographical areas are divided by altitude, they can include more or less vast territories, thus contributing to greater intrinsic variability. The use of portable instruments, with a limited wavelength range, posed greater challenges, making it more difficult to meaningfully distinguish the origin of honey. The results suggest that, although NIR spectroscopy and multivariate modeling approaches are widely used in the food industry, further technological improvements are needed in order to develop an integrated analytical system capable of accurately authenticating and discriminating the geographical origin of Italian polyflower honey.

CHAPTER 1

INTRODUCTION

There are several ways to define honey. From a biological perspective, honey is considered a reserve food because bees, which are feed only with nectar and pollen, need to accumulate food stocks with a long shelf life. As a dietary source, it consists of simple sugars, which is why it is regarded as a highly energetic and sweetening food.

Honey is unique as it is the only food that requires no processing to transition from nature to our table. The legal definition clearly identifies it: *"Honey refers to the natural sweet substance that bees (Apis mellifera) produce from the nectar of plants, or from secretions of the living parts of plants, or from exudates of sucking insects found on living parts of plants. They collect, transform by combining them with specific substances of their own, deposit, dehydrate, store, and allow it to mature in the honeycombs of the hive"* (- 2004. G.U. n. 168 del 20 luglio 2004. Decreto legislativo 21 maggio 2004, n. 179. Attuazione della direttiva 2001/110/CE concernente la produzione e la commercializzazione del miele).

The legal definition also stipulates that nothing is added to or removed from honey. Consequently, it should not be treated to eliminate any components.

Honey is the primary product of bees, and its physicochemical composition, particularly the pollen content, reflects the floristic complexity of the production area, which is influenced by topographical, climatic, pedological, and anthropogenic factors. For this reason, honey is closely linked to its production area and mirrors the plant diversity of its environment of origin through its organoleptic and palynological characteristics. These properties can be detected through physicochemical, sensorial, and melissopalynological laboratory analyses.

By combining knowledge of the local flora with the evaluation and comparison of analyses performed on a large number of honey samples collected from a specific study area, it is possible to identify the geographical origin of the honey. Specifically, the identification of the "*pollen spectrum*", the set of all pollen grains present in a sample of honey, and the determination of characteristic pollen associations, as well as the presence of indicator pollen typical of a region, contribute to pinpointing the geographical origin.

The ability to distinguish honeys based on their botanical origin and production area is crucial from a commercial standpoint to ensure the quality of the product for consumers. Thoroughly investigating the characteristics and geographical origin of honey also protects beekeepers from the market incursion of foreign honeys, particularly from Eastern Europe, which are often adulterated and lack traceability.

In terms of production, Italy ranks fifth in Europe and boasts another record: with approximately 7000 producers and almost one and a half million apiaries, it is the country with the most diverse range of honeys.

Regarding product types, the honey's market features both table honey and industrial honey. The former is primarily used for domestic consumption as a spread, to sweeten dishes and drinks, or in desserts prepared at home.

The latter is utilized in the food industry for preparing baked goods, candies, cereals, and beverages, and as a sweetener or flavoring agent in the pharmaceutical and cosmetic industries, as well as in the tobacco sector. About 60% of the honey sold on the national market is table honey, with the remaining 40% being industrial honey (Contessi, 2004).

1.1 THE POLYFLOWER HONEY

During production, it is often possible to produce honeys derived exclusively from the nectar (or honeydew) of a single plant species: these are known as monofloral honeys, and their production is particularly prevalent in Italy. Monofloral honeys can be harvested when a flowering event is sufficiently widespread, profuse, and not coinciding with other blooms. The market favors these honeys due to their unique organoleptic and quality characteristics, which are influenced by the specific plant source from which they originate.

On the other hand, when bees collect nectar from several blooms occurring simultaneously, it is not feasible to isolate the different sources. This results in the creation of multifloral honeys, often referred to as "*polyflower honey*."

This type of honey is notable for its considerable variation in taste, colour, aroma, and consistency, due to the diverse types of flowers involved and the honey's specific regional origin. In such cases, it is possible to anticipate particular characteristics for honeys from certain regions. For instance, throughout the Alps, there is a common type of polyflower honey produced in the lower valleys, distinguished by a robust aroma and a moderately bitter taste. The polyflower honeys from the Marche region often exhibit a yellow shade and rapid crystallization, resulting from the high presence of sunflower and a coconut scent due to the coriander nectar. In Salento, polyflower honeys frequently possess a spicy odor given by fenugreek. Meanwhile, the summer polyflower honeys from the interior of Sicily are characterized by intense aromas derived from plants in the fennel and carrot families.

Owing to these characteristics, a variety of polyflower honeys are available on the market, ranging from light to dark, crystallized to liquid, and with flavor profiles that can be bitter, very sweet, or subtly nuanced (Piana and Naldi, 2020).

1.1.1 THE CRUCIAL ROLE OF BEES THROUGH THE BIODIVERSITY

The bees' presence in the environment contributes to the production of honey and other products, but it also exerts a much more incisive and effective action. In fact, bees carry out the transport of the pollen of most cultivated and spontaneous plant species, which is very important because it allows the fruiting and preservation of the plant cover. It is for this reason that the primordial relationship between man and the bee continues to retain an unalterable appeal.

In the current historical context, in front of the increasingly pressing alarm of the rarefied biodiversity, bees are the living species that most preserve this essential environmental value. Both domestic and wild bees are accountable for about 70% of the pollination of all plant species and they contribute to approximately 35% of global food production.

Pollinators play a vital role in nature: they ensure the conservation of biodiversity. In Europe, almost half of pollinating insect species are in serious decline, due to habitat change and environmental pollution. In particular, the intensification of agriculture over the last six decades and the widespread and unstoppable use of synthetic pesticides is one of the main factors in the decrease in populations and biodiversity loss of pollinating insects in recent times.

The restoration of natural habitats, a drastic reduction of agro-chemical inputs and the implementation of conservation agriculture practices are probably the most effective way to avoid further decreases of pollinating insects but also, more generally, to improve the profitability of the farm (Gallai et al., 2009).

1.1.2 PRODUCTION

The production of honey begins with nectar, an aqueous solution of sugars, amino acids, proteins, lipids, minerals, and other components produced by the nectary cells of plants. Its exact composition varies depending on the environment and the plant species, influencing the flavor and quality of the resulting honey, particularly in the case of monofloral honeys.

The most present sugars in nectar are sucrose, glucose, and fructose in highly variable proportions; some nectars contain almost exclusively sucrose, while others may have equal amounts of the three polysaccharides. A worker bee can collect and store up to 25 mg of nectar in a special "*sac*" at the end of its esophagus.

During collection, the nectar mixes with secretions from the bee's salivary and hypopharyngeal glands. Enzymes in these secretions initiate the nectar's chemical transformation. Inside the hive, the nectar is transferred from the foraging worker bee to younger worker bees, who continue the transfer cycle multiple times, introducing additional enzymes and completing the transformation process.

Eventually, the transformed nectar is deposited into a honeycomb cell. As more processed nectar is added, the cell fills up and is eventually sealed with wax, a process called "*capping the honeycombs*". The conversion from nectar to honey takes between one and three days and results from two primary processes:

1. The conversion of sucrose into glucose and fructose, primarily by enzymatic action.
2. The evaporation of excess water, facilitated by the airflow generated by the bees' wings.

Once the honey is mature, beekeepers proceed with the extraction process, resulting in the honey that is typically sold on the market.

This process involves several sequential steps, which, depending on production volumes, can be conducted using automated machinery.

1. Removal of supers.
2. Uncapping: involves removing the wax seal from the cells.
3. Honey extraction: accomplished using special honey extractors that utilize centrifugal force to extract honey from the honeycombs. These machines enable cold processing, preventing any alterations in the product due to temperature increases.

4. Filtering: carried out with specialized fine metal meshes to eliminate wax, honeycomb residues, or other materials.
5. Decanting and skimming: the honey is decanted for a period ranging from days to weeks to remove air bubbles generated during extraction and allow any remaining residues to settle at the bottom or surface; any residues forming a foam on the surface are then removed.
6. Packaging in airtight jars: this sealing is crucial as honey is a hygroscopic product, prone to absorbing moisture from the air, potentially leading to bacterial growth (Piana, 1994).

1.1.3 CHEMICAL COMPOSITION

The exact chemical composition of honey depends on its botanical origin, production region and processing methods, but it can essentially be considered as a saturated solution of sugars, which can make up to 80% of the total. The water content, on the other hand, is normally below 20%, making honey scarcely susceptible to the development of bacteria if properly stored. In addition to an articulated mixture of carbohydrates, honey contains acids (both amino acids and organic acids), enzymes, minerals, vitamins, aromatic substances, pigments, and other minor compounds.

Carbohydrates

Honey is an extremely complex food, with a very specific profile. In general, the most common sugars are glucose and fructose, which together account for around 75% of the total. They are extremely important for the definition of certain properties of honey, including sweetening power and consistency.

In fact, fructose tends to make the final product sweeter and more liquid due to its high solubility in water, while glucose has a lower sweetening power and a greater tendency to crystallization, which occurs because of the passage of time and storage temperatures. In addition to the monosaccharides mentioned, there are also disaccharides such as maltose and sucrose, to which minor tri- and oligosaccharides are then added, which do not affect the organoleptic properties of honey, but are important to help determine its botanical origin.

Some of them, such as turanose, isomaltose, palatinose, kojibiose, nigerose and maltulose are present in constant concentrations in the different botanical origins, while others are characteristic of types of honey.

These include, for example, eroliose, which is more abundant in acacia, rhododendron, and honeydew honeys, melezitose, a trisaccharide typical exclusively of honeydew honeys, and gentiobiose, exclusive to honeydew.

Organic acids

The main acid present in honey is gluconic acid, which is produced by the effect of the enzyme glucose oxidase secreted by the glands of bees. This acid has a role in the aroma of honey, but it is also the main responsible for its pH, which is on average around 3.9.

The acidity of honey depends a lot on the time that elapses between the collection of nectar by the bees and the final storage in the cells of the hive and it is a very important factor to keep under control because it contributes to the stability of the final product towards microorganisms. Finally, the other acids present are butyric, acetic, succinic, lactic, maleic, oxalic, malic, and formic.

Amino acids and Proteins

Honey can generally be considered as a food low in protein and amino acids. The latter seem to derive mainly from bees and their secretions and make up 1% by weight of the components of honey. The proportion of the different amino acids depends on the botanical origin, but in general it is proline that is present in greater quantities. As far as protein is concerned, the honey produced by *Apis Mellifera* generally contains between 0.2 and 1.6%, but their amount also varies depending on the type of bee.

Water content

The water content contributes to define the quality. It depends firstly on the botanical origin, atmospheric and environmental conditions, prior to and following extraction, the intensity of nectariferous flow, the season of production, the beekeeper’s intervention techniques and preservation conditions. Around the 17% is considered an optimal value, indeed, a higher content would trigger fermentative processes, whereas levels below a certain limit would alter workability.

Minerals and Vitamins

Honey is a food without a significant percentage of minerals and vitamins, as these are present in very small proportions compared to other components. The primary minerals found in honey include potassium (K), sodium (Na), calcium (Ca), magnesium (Mg), chloride (Cl), sulfur (S), phosphorus (P), silicon (Si), iron (Fe), manganese (Mn), and copper (Cu). Their quantities vary greatly, particularly in relation to the geographical origin of the honey, since they originate from the minerals present in the soil, which are absorbed by plants through their sap. Regarding vitamins, their concentrations in honey are typically measured in parts per million (ppm) and are mainly derived from nectar. These vitamins predominantly belong to the group of water-soluble vitamins (Kunat-Budzynska et al., 2023).

Aroma and Colour

A particular characteristic of honey is certainly its colour, which can vary greatly with the season, the source of nectar and processing. In general, the colour of honey can be assimilated as that of a diluted caramel solution, i.e. an amber hue that ranges from almost transparent to a very dark tone. What is usually noticed is that the colour changes drastically from light to dark going from spring flowering honeys to summer ones, up to blooms closer to autumn.

Of course, it is rather difficult to identify with absolute certainty a type of honey by its colour, as it is also influenced by the presence and size of any glucose crystals: they, being white, can reflect light and thus change the colour compared to a more liquid honey. What exactly the colour of honey comes from is not perfectly clear, the most accredited hypotheses concern the presence of phenolic compounds and products of non-enzymatic browning reactions between amino acids and sugars.

The aroma is another characteristic element of honey. It, along with flavor, is attributed to volatile compounds that include esters of aliphatic and aromatic acids, aldehydes, ketones, and alcohols. Among the compounds that most seem to smell and taste like honey is phenylacetaldehyde, but then there are several compounds that are typical of the botanical origins. There are seven families of smells and aromas used to describe honey (vegetable, animal, floral, fruity, warm, aromatic, and chemical (Naldi and Pizzirani, 2015).

1.1.4 FOOD FRAUD

The term "*food fraud*" refers to the production, possession, trade, sale, and administration of food that does not comply with the laws in force. It may be of a health nature if it poses a threat to the health of the consumer or if it damages him economically without necessarily causing damage to his health. In addition, a distinction can be made between frauds that alter the intrinsic quality of products (alterations, adulterations, adulteration) and frauds that affect the marketing (counterfeiting, falsification).

For the honey, there are several ways in which such illegal practices are implemented, especially to reduce the costs of the product and increase the productivity of the hives.

The most common offences are:

- Adulteration: the sale of foodstuffs with characteristics different from those declared.
- Counterfeiting: defined as the presentation of a food in a way different from how it is in its natural constitution, or the creation of a product from scratch that is apparently like the original one.

A first source of adulteration is represented by the addition of syrups, with a sugar composition like that of genuine honey, to dilute it and therefore make it cheaper; some possible examples of syrups used are invert sugar or high fructose (HFCS), but many others could be cited.

Another common practice is to increase the productivity of hives through forced feeding of bees, which can also come from syrups, invert sugar and sugar pastes. It is allowed in the winter months to ensure the survival of insects; however, the practice must be suspended before the start of honey production, otherwise it becomes adulteration.

If we talk about counterfeits, the most common is the production of artificial honey, which is a cream consisting mainly of invert sugar and water, to which dextrins, sucrose, HMF and other compounds are added that well simulate the aroma and colour of natural honey.

Finally, more and more often we must deal with the marketing of honeys of botanical origin other than that declared and the sale of non-EU honeys for EU honeys.

As a result of treatments such as those described, honey undergoes changes, which are increasingly difficult to demonstrate, especially if the controls carried out stop at marginal details or if the techniques used are not modernized to keep up with those used to modify the product (Mateo et al., 2021).

1.2 NEAR INFRARED SPECTROSCOPY

The Near InfraRed (NIR) spectroscopy (NIRs) technique is an analytic method that exploits the interaction between the matter and the near infrared radiation. NIR is used in food science and technology because it is an accurate, rapid, non-destructive, reliable, and inexpensive analytical technique.

This technique takes advantage of the specific ability of each chemical compound to absorb, transmit or reflect light radiation. The combination of the absorbent properties, combined with the light energy scattering properties, results in the diffuse reflectance of the light, which contains information about the chemical composition of the sample. The time taken for a single analysis varies from a few seconds to a few minutes.

This method allows a rapid investigation of numerous samples with a significant reduction in time and cost compared to a traditional technique. This is mainly due to the simplicity of the preparation of the operations and the variety of analyses that can be carried out.

The possibility of reusing the sample and the absence of reagents completes the picture of advantages. The concentrations of the macro-elements of a food sample such as water, protein, fat, and carbohydrate can be determined using the classical absorption spectroscopy. However, for most food samples, this chemical information is hidden by changes in the spectra caused by physical properties, such as the particle size of the powders (Osborne and Fearn, 1993).

This means that NIR spectroscopy becomes a secondary method that requiring calibration against a reference method for the constituent of interest: the choice of the right algorithm to be used for the interpretation of the data and the accurate calibration of the sophisticated equipment, are of considerable complexity. The method is not applicable to a complete analysis of all constituents such as, for example, mineral elements (Chen et al., 2014).

Regarding this case study, currently, most of the technologies used to identify honey quality are inefficient and costly. But NIR has the potential to be a valid technique for the assessment of honey authentication domain including constituents, adulteration, botanical origin, and geographical origin.

Future research has to focus on increasing model robustness and a NIR-based integrated technology system on honey quality assessment. (Chen et al., 2014).

1.2.1 BACKGROUND

The first NIR spectrum was recorded in the 1800s when Herschel proceeded his measurements of solar emission heat beyond the "*red*" portion of the visible spectrum. However, starting in 1881, Abney and Festing began to evaluate the spectra of some organic liquids in the near infrared between 700 and 1200 nm using photographic instruments. They recognized the importance of hydrogen bonds in the formation of absorptions in this spectral region.

The foundations for modern analyses in this spectral area began around the 50s, when the Department of Agriculture of the United States of America was involved in a research program with the aim of developing chemical-physical methods to obtain rapid qualitative evaluations of food and, from here, numerous studies took place. The enhancement of NIR Spectroscopy subsequently followed advances in various fields of technology such as: optics, electronics, hardware, and software and, above all, chemometrics. But modern instrumentation owes its existence mainly to a microprocessor in which all calibrations are stored and which can convert spectral data into analytical results.

Initially, NIR instruments were used for food analysis because the low absorbances of the absorption bands were compatible with moderately concentrated samples and because the distances of the longer trajectories were similar to those of the mid infrared. This allowed measurement by transmission through intact materials and allowed for rapid, non-destructive analyses. The NIR spectra of intact, opaque, and biological samples can also be obtained by diffuse reflectance, so that they do not require the use of special cells. In addition, the optical materials that make up a NIR instrument and the low absorbance of water make this spectral region particularly suitable for the analysis of samples that contain a high percentage of water. Therefore, the growing interest of the NIR technique in the agri-food sector is probably a direct result of its major advantages over other analytical techniques (Osborne and Fearn, 1993).

For examples, an easier sample preparation without any pre-treatment, the ability to repeat multiple measurements on the same sample, and the prediction of chemical and physical parameters from a single spectrum.

1.2.2 ABSORPTION BANDS

The NIR region covers from the upper limit of visible wavelengths, around 750 nm, up to 2500 nm. Absorption bands within this region are harmonic or combination bands of fundamental vibrational stretch bands falling in the range of 3000 to 1700 cm^{-1}, typically involving C-H, N-H, and O-H bonds. Due to their harmonic or combination nature, the molar extinction coefficients of these bands are low, limiting their detection to approximately 0.1%. Molecular vibrations can be categorized as stretching or bending, with stretching involving a continuous change in interatomic distance along the bond axis.

At temperatures of 20-25 °C, molecules primarily exist in their fundamental vibrational energy states. Atoms or groups of atoms participating in chemical bonds oscillate relative to each other at a frequency determined by the bond strength and mass of the bonded atoms or groups. These vibrations have amplitudes of a few nanometers, which increase with additional energy input.

Transparent samples are typically analyzed in glass or quartz cuvettes with optical paths ranging from 1 to 50 mm, dependent on the spectral region of interest. The optical path length decreases as the wavelength transitions from higher-order overtones to the combination region near 2200 nm. In transflectance, unlike a simple transmittance measurement, doubling the optical path in measurements involves the radiation beam passing through the sample twice (Pasquini, 2003).

1.2.3 NIR INSTRUMENT

A NIR spectrophotometer is generally composed of a light source, a monochromator, a sampler, or an interface for the presentation of the sample and a detector for the measurement of reflectance and transmittance, all managed by a computer.

The light source is typically a tungsten halogen lamp. Detectors can be Silicon, Lead Sulfite (PbS) and Indium Gallium Arsenide (InGaAs). Silicon detectors are faster and highly sensitive from the visible region to 1100 nm. PbS detectors are slower but very common since they are sensitive from 1100 to 2500 nm and have a good signal-to-noise ratio. The most expensive is InGaAs, which combines the speed and features of silicon detectors with the wavelength range of the PbS detector (Reich, 2005).

There are various optical configurations that can be used to separate the NIR polychromatic spectral region into "*monochromatic*" frequencies (Osborne and Fearn, 1993).

1.2.4 SPECTROPHOTOMETER

This chapter focuses on NIR instruments for spectroscopic analysis. While a visible spectrophotometer is not traditionally considered an NIR instrument, it is included here for a more comprehensive understanding of the techniques involved in the trial.

A visible spectrophotometer is a high-precision instrument used to measure the reflectance or transmittance of a material across the visible spectrum, to assess the colorimetric characteristics of our samples, which ranges typically from 400 to 780 nanometers. This type of instrument is commonly used to analyze the optical properties of materials, such as transparency, absorption, and reflection of light. Specifically, a visible spectrophotometer is used to determine the concentration of a substance in a solution by measuring the absorption of light at a specific wavelength.

While spectrophotometers provide accurate and reliable measurements, they differ from colorimeters, in their operating principles and applications. Colorimeters are designed to measure the tristimulus values of a sample values X, Y, and Z, which are used to calculate the CIELAB or HUNTERLAB colour coordinates (L* Lightness, a* redness, b* yellowness).

Whereas spectrophotometers measure the entire reflectance or transmittance spectrum of a material, allowing for a more detailed analysis of its colorimetric properties.

Furthermore, the spectrophotometer instrument is based on the SCI (Specular Component Included) method, which considers the specular reflection of the material, whereas the SCE (Specular Component Excluded) method, commonly used in colorimetry, excludes this component. The choice of the SCI method in this study allows a more accurate representation of the colorimetric properties of our samples, particularly in the case of materials with high gloss or specular reflection. [1]

[1] https://www.plastix.it/author/maurizio-messa-konica-minolta/

1.2.5 CHEMOMETRICS

Before performing a quantitative analysis, it is necessary to develop a specific calibration for the matrix to be analyzed using multivariate modeling.

Chemometrics is a scientific discipline that deals with developing and using mathematical and statistical models to interpret chemical data obtained from analytical techniques such as spectroscopy, chromatography, and mass spectrometry, to extract quantitative and qualitative information about the analyzed samples.

It is widely applied in analytical chemistry but also in fields such as biochemistry, biology, medicine, food industry and environmental engineering.

The chemometric approach is used to relate the physico-chemical properties of the samples to be analyzed with the absorption of radiation, in the wavelength range of NIR. That allows predictions to be made about their chemical origin (Jimenez-Carvelo et al., 2019).

This approach is based on some fundamental principles:

- Optimization: A key aspect of chemometrics is the optimization of experimental conditions and analysis algorithms to extract the maximum information from the available data. So, it is necessary to reduce the dimensionality and the complexity of data while retaining the most relevant information (Principal Component Analysis, PCA).
- Validation and Calibration: Chemometric models must be rigorously validated to ensure they are reliable. This involves in testing the model performance with a reference set composed by samples that must contain all possible variations that might be found in unknown samples. The calibration of the instruments is fundamental to ensure accurate and reproducible results. Chemometrics provides methods for multivariate calibration, such as Partial Least Squares Regression (PLS) or Principal Component Regression (PCR).
- Spectral data analysis: Chemometrics is particularly useful in the analysis of spectral data (i.e., spectroscopic data), where complex patterns and relationships can be extracted through advanced analytical techniques.
- Multivariate analysis: Chemometrics focuses on the analysis of complex relationships between multiples variables which they are measured simultaneously using

mathematical models. These methods can be either quantitative (prediction of chemical concentrations) or qualitative predictions (discrimination of product categories). The first are called exploratory or “*data mining*” methods, which in an unsupervised manner can identify trends or clustering in data. Secondly there are supervised methods, in which the models are based on the input and output data (chemical composition of the sample to be predicted).

Chemometrics can be seen as a subset of the ML domain, which is itself included in the AI field (Barthès et al., 2019; Roussel et al., 2011).

1.2.6 MACHINE LEARNING

Machine learning (ML) deals with the development of algorithms and models that allow computers, by learning from data sets, to identify tendencies, to automate complex processes and to make decisions or provide more accurate predictions autonomously.

The two main applications of ML are Regression, to predict a continuous variable, and Classification, to assign each sample into one or more categories.

It can be applied to different types of data, such as graphs, trees, curves, discrete data, and instrumental data, that resulting from spectroscopy (Visible, Near InfraRed, InfraRed Spectroscopy). Generally, these databases are built of many observations and of a large number of observed variables.

Among supervised methods, there are: decision tree methods (Random Forest), K-Nearest Neighbors classification (KNN), and Support Vector Machines methods (SVM).

Random Forest

The Random Forest algorithm is a ML technique based on ensemble learning, which involves aggregating multiple models to achieve a better overall predictive performance. Specifically, Random Forest relies on the aggregation of decision trees.

To create a Random Forest, a certain number of decision trees (in our case 500 combination) are generated. Each tree is trained on a random subset of the training data and using a random subset of the features.

Each tree is trained independently of the other trees, using a recursive splitting process of the training data to create decision nodes that separate the data based on features.

To make a prediction on new data, the Random Forest combines the predictions of all the decision trees. In classification tasks, the mode (i.e., the most common class predicted by the trees) can be used, while in regression tasks, the average of the trees' predictions can be used.

The Random Forest algorithm can handle both classification and regression problems and is particularly effective for complex datasets with many features and noisy data. Additionally, the Random Forest can automatically handle the selection of the most relevant features for prediction, reducing the risk of overfitting.

K-Nearest Neighbors

The K-Nearest Neighbors (KNN) algorithm is a supervised learning algorithm used for classification and regression problems. KNN uses the similarity and proximity of data to make predictions. It assumes that similar points tend to be close together in feature space.

The algorithm works by identifying the "*Nearest Neighbors*" of a test data, looking at the most similar data points in the training set. The algorithm then classifies the test data point based on the majority of its Nearest Neighbors. The value of K represents the number of neighbors to consider for classifying the test data point.

First, KNN calculates the distance between the test point for which a prediction is desired and all points in the training dataset. Distances can be calculated using different metrics, such as the Euclidean distance or the Manhattan distance.

To make predictions, KNN calculates the distance between the test point and all points in the training dataset. It then selects the K-Nearest points to the test point based on the calculated distance (Nearest Neighbors).

Finally, for classification problems, it determines the majority class among the K-Nearest Neighbors and assigns this class to the test point. For regression problems, it predicts a target value based on the average of the target values of the KNN.

This approach is simple and intuitive, but it can be influenced by the choice of the K value and the distance metric used.

Support Vector Machines

Support Vector Machines (SVM) is primarily used for non-linear or complex problems, aiming to find the optimal hyperplane that separates data into different classes. It focuses on searching for boundaries that separate classes while maximizing the margin between support points of different classes. This means that only a subset of training samples, known as support vectors, is utilized to define the boundaries.

SVM is effective in handling complex and non-linear data thanks to the "*kernel trick*" technique, which transforms data into a higher-dimensional space where they become linearly separable.

Although originally created for classification, SVM methods have been extended to regression, in particular two methods: SVM linear and SVM radial (Kerser, 2023).[2-3]

[2] https://www.akkio.com/post/5-types-of-machine-learning-classification-algorithms/

[3] https://ondalys.fr/en/scientific-resources/machine-learning-methods/

CHAPTER 2

AIM OF THE DISSERTATION

The aim of this final MSc dissertation was to evaluate the geographical origin of Italian polyflower honey using a set of spectroscopic techniques and a machine learning classification approach. This purpose was achieved using three spectroscopy based portable instruments for the rapid analysis of honey. The objective was pursued by comparing the predictive performances of the three portable spectroscopy instruments based on VIS and NIR spectra data for a rapid assessment of the ecological and geographical origin of honey polyflower harvested in Italy during the summer season of 2022. The authentication of the honey terroir productive was performed according the first-level administrative divisions of the Italian Republic (so called regions) and furthermore based on an ecological clustering corresponding to plains and hills of the North and of the South Italy and high mountain of the North Italy.

The objective was earned by applying four supervised multivariate classification models named Random Forest, KNN, SVM linear and SVM radial, and their accuracy in classifying the samples was assessed using a set of descriptive statistics applied to the related confusion matrices.

CHAPTER 3

MATERIALS AND METHODS

3.1 SAMPLE COLLECTION

The study considered a total of two hundred and twenty-seven (n = 227) samples of Italian polyflower honey from various beekeepers honeyed during the spring-summer season in 2022. All the honeys were collected during the Italian national competition named *Tre Gocce d'Oro-Grandi Mieli d'Italia*, a competition organized by the National Honey Observatory (*Osservatorio Nazionale del Miele*) in September 2022.

The botanical composition (*polyflora*) and the geographical origin (Italian region and specific location of the apiaries) were declared by trained experts of the Italian National Honey Observatory. Which experts made use of the producers' declarations on the label and in case of absence had contacted the producer for the admission to the competition. According to Council Directive 2001/110/EC of 20 December 2001 on honey, which supplements the general EU rules on food labelling, laid down in Regulation (EU) n. 1169/2011, it is mandatory to indicate the name of sale of the product and the country of origin of the honey harvest on the label. Therefore, you had certainty in composing the dataset following the Italian region of origin and the ecological areas. For the purpose of the ecological area classification, the honey samples were classified based on the altitude of the apiaries from which they originated:

- samples from South Lowland and North Lowland apiaries situated at low altitudes (<600 m asl), which include the possibility of apiaries from both plains and hills.
- samples from North Mountain (NM) apiaries located at high elevations (>600 m asl).

All samples were delivered to the laboratory of Experimental Chemistry of the *Istituto Zooprofilattico Sperimentale delle Venezie (IZSVe),* where 200 g aliquots were sampled in conic tubes (Falcon 352098, 50 ml). Then, the aliquots were delivered to the LabCNX of the Department of Animal Medicine, Production and Health (MAPS) of the University of Padova (Campus of Agripolis).

The samples were recorded and identified with the progressive analysis codes of the laboratory and then stored at 4 ± 2 °C until they were processed for analytical purposes. A classification of the geographical origin of the dataset is given in Tables 3.1., 3.2. and Figure 3.3. The indications provided by the National Honey Observatory were followed for both subdivisions.

Table 3.1. *Honey origin and sampling size according to the Italian regions.*

Region	Sample number	Region	Sample number	Region	Sample number
Abruzzo	19	Lazio	5	Sardegna	11
Basilicata – Molise	25	Lombardia	15	Sicilia	22
Calabria	6	Marche	8	Toscana – Umbria	10
Campania	4	Piemonte – Liguria	21	Trentino Alto Adige	10
Emilia Romagna	19	Puglia	16	Valle d'Aosta	7
Friuli Venezia Giulia	6	San Marino	3	Veneto	16

Table 3.2. *Honey origin and sampling size according to three Italian ecological areas. South Lowland (SL) and North Lowland (NL), honey samples from apiaries located below 600 m above sea level (asl); North Mountain (NM) honey samples from apiaries above 600 m asl.*

Ecological Areas	Involved Regions	Sample Number
SOUTH LOWLAND (SL)	Abruzzo, Basilicata, Calabria, Campania, Lazio, Marche, Molise, Puglia, San Marino, Sardegna, Sicilia, Toscana, Umbria	133
NORTH MOUNTAIN (NM)	Friuli-Venezia Giulia, Trentino-Alto Adige, Lombardia, Piemonte, Valle d'Aosta, Veneto	35
NORTH LOWLAND (NL)	Emilia-Romagna, Friuli-Venezia Giulia, Trentino-Alto Adige, Liguria, Lombardia, Piemonte, Valle d'Aosta, Veneto	59

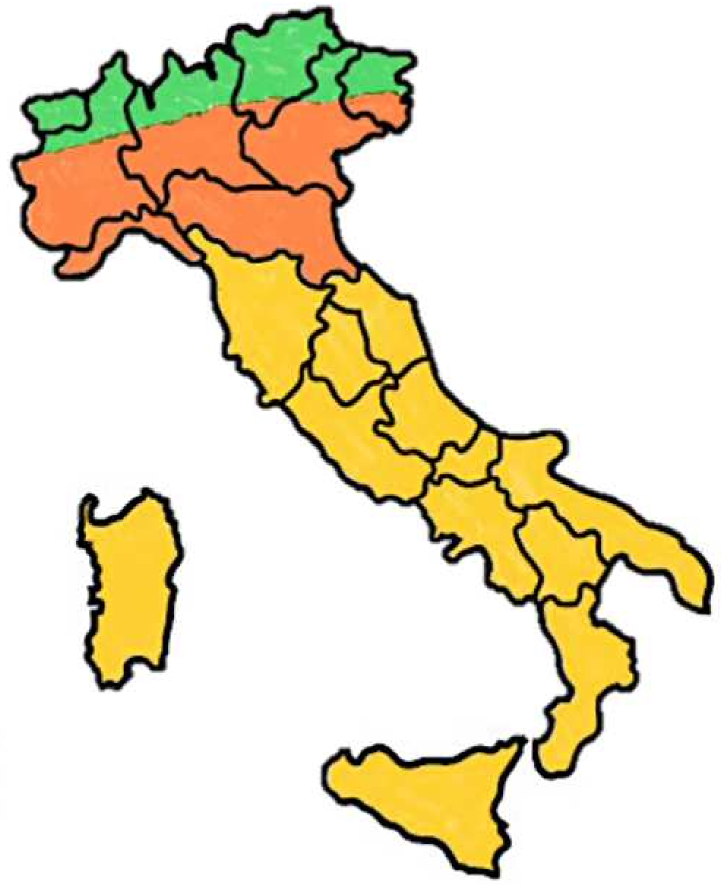


Figure 3.3. *Italian regions according to three ecological areas: South Lowland (SL) in yellow, North Mountain (NM) in green and North Lowland (NL) in orange. Lowland, honey samples from apiaries located below 600 m above sea level (asl); high mountain samples from apiaries above 600 m asl.*

3.2 SPECTROSCOPIC ANALYSIS

For this experimental trial, 227 honey samples were analyzed by three instruments: a portable VIS spectrophotometer (CM-600d Spectrophotometer, Konica Minolta Ramsey Sensing, Inc., Japan), a portable VIS-NIR (PolispecLITE, ITPhotonics, Fara Vicentino, Italy), and a portable NIR (PolispecNIR, ITPhotonics, Fara Vicentino, Italy). The VIS spectrophotometer operates in a range of 450-700 nm, with an interval of 10 nm. The VIS-NIR instrument operates in the spectral range of 770-1080 nm, with a gap of 2 nm between data points. The NIR instrument consists of a diode array operating in a region of 902-1680 nm, with a gap of 2 nm between data points.

The samples were first retrieved from their storage box, which maintained darkness and ensured a stable temperature (4 ± 2 °C). Six samples at a time were then heated in an oven at a controlled temperature of 40°C ± 2°C for 30 minutes ± 10 minutes. If any sample appeared unevenly mixed before analysis, it was gently mixed with a spatula to achieve homogeneity. Care was taken to avoid introducing air bubbles during this process. Spectrophotometer, VIS-NIR, and NIR analyses were performed within one hour of the sample's heat treatment in the oven.

For the VIS spectrophotometer, an aliquot (5 ± 1 g) of the sample was taken with a spatula, washed in hot water to avoid contamination, and poured into a glass cell of about 9.1 cm^2 of area. The spectrophotometer was fixed in a vertical device, and the cuvette was placed on top of the illuminant source. The measurement was conducted in a dark room, using a D65 light source, a 10° observation angle and 8 mm area. The cuvette was rotated by approximately 120° after each scan and it was thoroughly washed with hot water and dried following the three measurements. Ultimately, the data were exported in HUNTER L*a*b* system, where L* denotes the lightness, a* represents the redness and b* represents the yellowness (Segato et al., 2019).

For VIS-NIR and NIR instruments, an aliquot of each sample (1 g ± 0.5 g) was taken with a spatula and placed on a ring cell with a quartz window that allows the irradiation of about 12.6 cm^2 of area. A gold reflector with a 0.5 mm optical path was used to scan the sample in transflectance mode. To ensure a uniform distribution (homogeneity) of the sample across the entire quartz window, gentle pressure was applied. Care was taken to avoid introducing air bubbles during this process. Once the preparation was completed, the cell was placed first on the NIR and then on the VIS-NIR with the quartz window above the illuminant source.

Spectra were collected on average of three replicas of the same sample, and for each replica, three points were acquired. After the first reading, the ring cell was washed with hot water and dried, then a new aliquot of the same sample was put in to read the second repetition, and the same was done for the third replication. In conclusion, nine spectra for every sample were obtained. The spectra recorded with both apparatuses were exported in CSV. format, creating a dataset for further use (Woodcock et al., 2007; Bisutti et al., 2019).

3.3 STATISTICAL ANALYSIS

The data sets were reviewed to facilitate multivariate analysis, with each sample assigned a unique code that represents its region of origin and altitude (SL = South Lowland, below 600 m asl; NM = North Mountain, above 600 m asl; NL = North Lowland, below 600 m asl). Before any calculation, for all instruments, averages of the three replicates of the three aliquots per sample were calculated to obtain a single spectrum per sample.

Three datasets were obtained, each related to a spectrophometer, VIS-NIR, and NIR instrument, further fused in a single dataset including all three instruments arranged in the following order: spectrophotometer, VIS-NIR and NIR, where the overlapping wavelengths have been eliminated. The features within the fused dataset have all undergone standardization (mean = 0, standard deviation = 1). This standardization step was necessary to address the varying scales of the three instruments used for data collection. The standardization ensures that each feature has an equal weight in calculations, preventing any single instrument's scale from dominating the analysis.

To assess the use of portable instruments for determining honey origin using ML classification, three algorithms were tested on the fused dataset: a Random Forest (RF), K-Nearest Neighbors (KNN), and Support Vector Machines (SVM) with either linear or radial kernels.

The R software (version 4.0.2) was used to apply these models to spectral data. All algorithms were tuned using a grid of possible values for their parameters, according to the lower classification errors. All data sets were randomly divided into 70% (n = 162) of training (Tr) and 30% (n = 65) of testing (Ts) sets. The four ML models were trained on the Tr data and then evaluated on the Ts data to assess their performance.

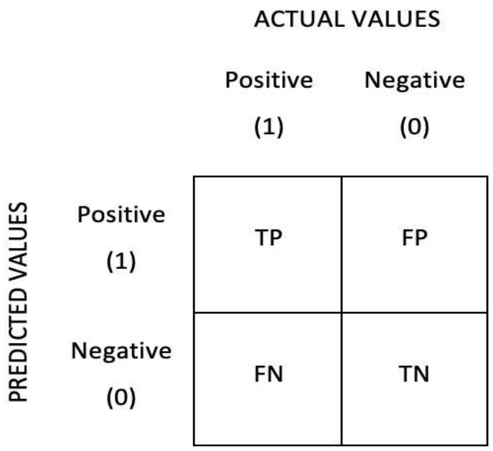
To obtain robust results, the process of randomly splitting the data into training and testing sets, followed by training, and testing the ML algorithms, was repeated 100 times for regional classifications and 500 times for ecological area classifications.

The average values of the performance metrics were then used to evaluate the effectiveness of the algorithms. Finally, the goodness of the tested models has been determined by creating a confusion matrix for each model.

The confusion matrix assesses the classification performance of these algorithms by comparing the actual values with the predicted values and then conveying its descriptive statistics. It permits the calculation of:

- True Positives (TP): number of examples that are classified as belonging to the positive class and actually belong to that class.
- True Negatives (TN): number of examples that are classified as belonging to the negative class and actually belong to that class.
- False Positives (FP): number of examples that are classified as belonging to the positive class but actually belong to the negative class.
- False Negatives (FN): number of examples that are classified as belonging to the negative class but actually belong to the positive class.

Table 3.4. *Example of a confusion matrix.*



		ACTUAL VALUES	
		Positive (1)	Negative (0)
PREDICTED VALUES	Positive (1)	TP	FP
	Negative (0)	FN	TN

Specifically, according to Bisutti et al., 2019, the descriptive statistics evaluated the sensitivity (the percentage of positive cases the model is able to detect), the specificity (the percentage of negative cases the model is able to detect), the accuracy (proportion of correctly identified cases as both true positive and true negative), the precision (proportion of correctly identified true positive cases belonging to the actual true positive and false positive group), and balanced accuracy (the mean of sensitivity and specificity, used to deal with imbalanced data), as following:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = 1 - (\frac{FP + FN}{TP + TN + FP + FN})$$

$$Precision = \frac{TP}{TP + FP}$$

$$Balanced\ Accuracy = \frac{Sensitivity - Specificity}{2}$$

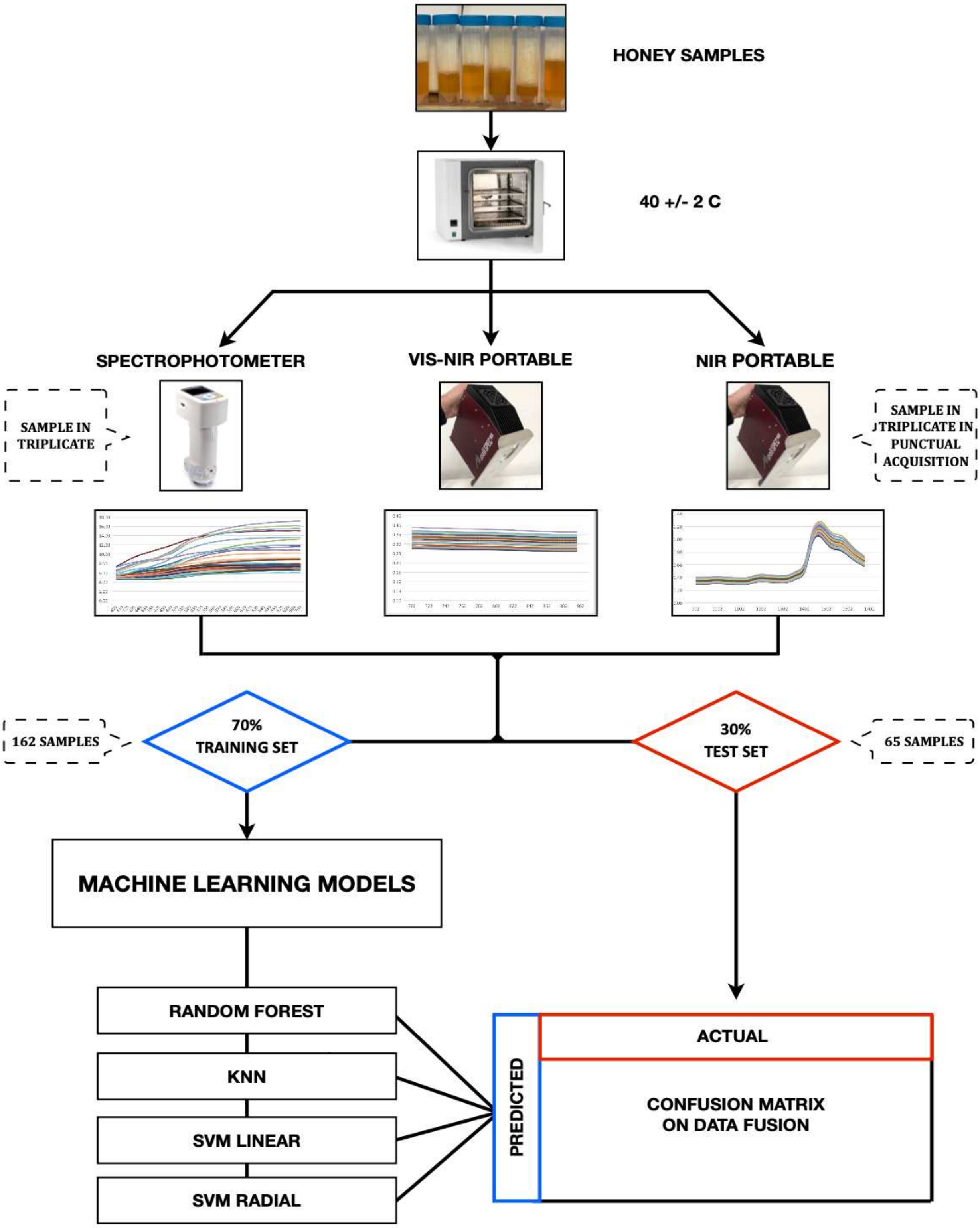


Figure 3.5. *Flowchart of the experimental design for the assessment of the origin of polyflower honey (n = 227). Spectrophotometer VIS, portable NIR, and VIS-NIR pre-treated data of the training set (n = 162) were used to build the supervised classification models. The predictive performance of classifying algorithms was evaluated by an external validation on a test set (n = 65) by means of a set of confusion matrices.*

CHAPTER 4

RESULTS AND DISCUSSION

4.1 NIR SPECTRA ANALYSIS

In Figure 4.1., 4.2. and 4.3. are reported the spectra collected by the three instruments. As stated in the chapter of Materials and Methods, honey samples were mild heated, stirred and the scanned with the portable NIR instruments. The preliminary heating treatment was done to avoid that the informative spectra NIR data might be affected by the presence of crystals of sugars that can interfere with light scattering. This methodological approach was suggested by a previous trial performed by Segato et al., 2019, which evidenced that low temperature treatment (40°C ± 1) was chosen as sample pre-analysis treatment, mainly because it helps homogenizing the honeys, but at the same time, it should not alter their physical-chemical properties.

The Standard Normal Variate (SNV) pre-processing was used for the VIS-NIR, NIR and fused spectra, for scatter correction, together with the first derivative and smoothing.

In each Figure 4.1., 4.2. and 4.3. the vertical axis represents the absorbance (log (1/R)) value from the honey samples, which is a measure of the amount of light absorbed by the sample at a given wavelength. The horizontal axis shows the usable spectral range, spanning from visible light to the NIR region. These wavelengths values are expressed in nanometers (nm) and covers a range from 400 to 1700 nm. Figure 4.1. evidences the light absorbance in the visible range for all the samples, which result in a quite similar trend for most of the samples. In a minority of samples, the absorbance is higher (from 8 to 17) throughout the absorption spectra compared to the majority that is from 4 to 8. This is probably due to the difference in colour that the polyflower honey samples present. The coloration ranged from straw yellow to bright orange (see Figures 6.3.A., 6.3.B. and 6.3.C. on the appendix).

Figures 4.2. and 4.3. show the VIS-NIR and NIR spectra of all the samples, which follow a similar profile with continued overlapping among the samples. The VIS-NIR (Figure 4.2.) show very low absorption peaks, proceeding in a linear manner. The NIR spectra shapes resemble the ones observed in Woodcock et al., 2007 and Bisutti et al., 2019, confirming the presence of dominant absorbance bands, which corresponds to O-H, C-H, and C-H_2 deformations. In fact, the Figure 4.3. exhibit a prominent peak in absorbance values, ranging especially around 1450-1550 nm, which is attributed to the O-H stretching vibrations of water molecules present in the honey.

Also, two weaker absorption bands are observed around 1200-1300 nm and 1600-1650 nm, which may be related to the C-H and C-H_2 stretching vibrations of fructose and glucose compounds. The observed differences in absorption can be attributed to the composition and presence of specific molecules in different samples (Woodcock et al., 2007; Bisutti et al., 2019).

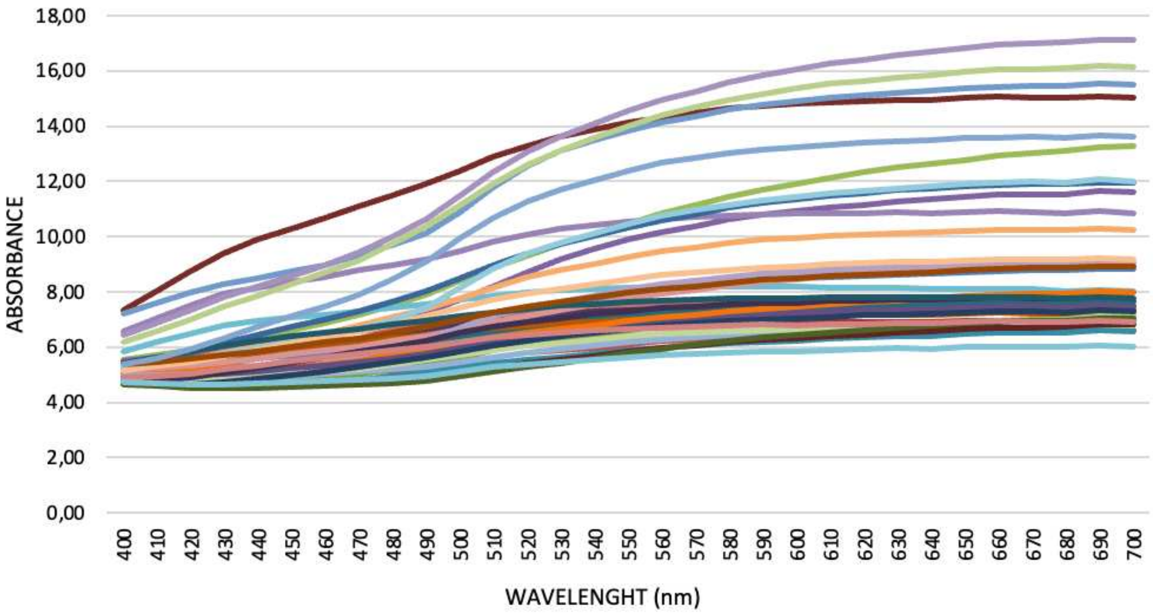


Figure 4.1. *Average NIR spectra of polyflower honey related to spectrophotometer NIR instrument.*

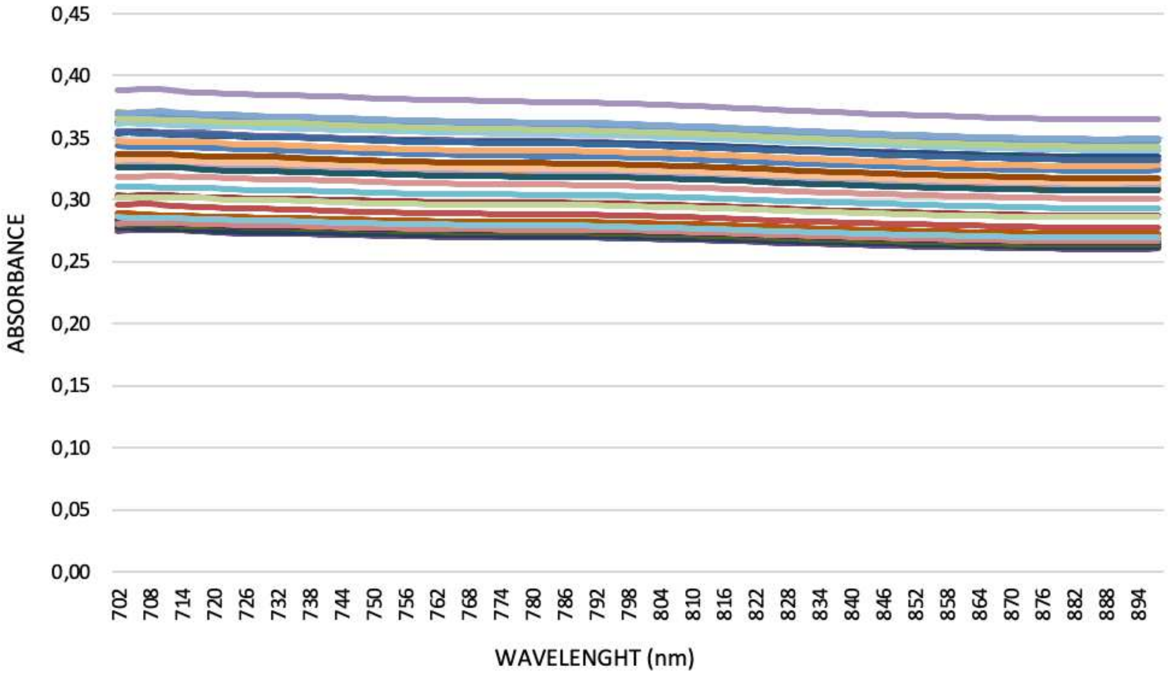


Figure 4.2. *Average NIR spectra of polyflower honey related to VIS-NIR instrument.*

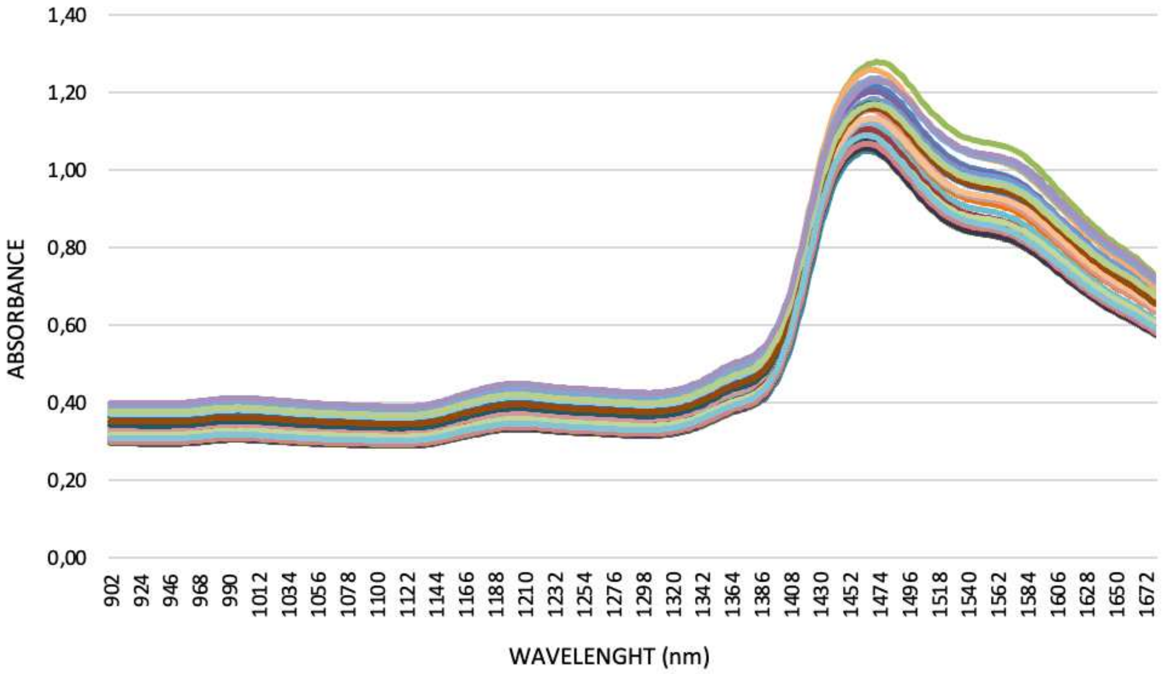


Figure 4.3. *Average NIR spectra of polyflower honey related to NIR instrument.*

4.2 RANDOM FOREST, KNN AND SVM MODELS

To assess the origin of the honey through VIS-NIR absorption, four supervised machine learning algorithms were used: Random Forest, KNN, SVM linear and SVM radial.

The tables Table 4.4. and Table 4.5. would allow comparing the performance of each ML model on the three different datasets and permit to identify the best-performing model and dataset combination for classifying honey samples by ecological areas and regions.

The performances are expressed in terms of accuracy and Cohen's Kappa (Kappa).

- Accuracy: indicates the percentage of correct predictions out of the total predictions made by the model; a higher accuracy value indicates better performance.
- Cohen's Kappa: is a measure that considers the agreement between the predictions and the actual values, correcting for chance agreement; a higher Kappa value indicates greater reliability of the classification model.

In Table 4.4. are reported the metrics for the classification of the ecological areas. All classifier models have similar performances in terms of accuracy, ranging between 0.51 and 0.60, which indicates that no model significantly stands out from the others. The Kappa value varies between 0.04 and 0.14, with SVM linear and SVM radial methods showing slightly better values compared to others. However, even these Kappa values are relatively low, indicating that the agreement between the model predictions and the actual data is not far from what would be obtained by chance.

For ecological areas, the supervised models have marginally acceptable performance in terms of accuracy, but the Kappa measure indicates that agreement beyond chance is limited. This suggests that although there is still a considerable amount of uncertainty in the predictions, the models can correctly predict more than half of the time.

In Table 4.5. are reported the metrics for the classification of the regions, the accuracy values are generally lower than the ecological areas (Table 4.4.), with a range between 0.10 and 0.49. In particular, VIS-NIR has a very low accuracy value (0.10) with KNN algorithm, which indicates that for this dataset and the classification method used, the prediction is slightly better than a random choice. The highest accuracy (0.49) was achieved by the Random Forest and SVM linear models on the data fusion.

Note that the maximum Kappa value in Table 4.5. is 0.10, which is still considered a low value. This implies that the concordance between the model predictions and the real data is not strong.

For Italian regions, the models generally have poor performance, with accuracy values that are significantly lower compared to ecological areas. Kappa values are also low, suggesting that the models are not as reliable. According to Bisutti et al., 2019, the imbalance in terms of group size, could affect the capability of the algorithm in the discrimination. In fact, the composition in terms of samples per regions is very variable (Table 3.1.).

Looking at both tables, these observations might indicate that the task of classifying regions is more complex or perhaps the available data are not sufficiently informative. Moreover, the differences in results between ecological areas and regions are may attributable to a high intra-regional variability. That could be due to differences in the features of the data or other factors such as the presence of unbalanced classes.

Besides, the fact that polyflower honey is a set of different blooms may have concurred in reducing the classification performances of the algorithms tested. For the development of this hypothesis, it will be needed to consider a larger number of samples for a classification of honey by region of origin (Bisutti et al., 2019).

Table 4.4. *Comparison of the predictive performance (Accuracy and Kappa) of the four investigated algorithms in the honey samples training set for ecological areas.*

	DATA FUSION	**VIS-NIR**	**NIR**

RANDOM FOREST	Accuracy = 0.58
	Kappa = 0.07
		Accuracy = 0.60
		Kappa = 0.11
			Accuracy = 0.58
			Kappa = 0.10

KNN	Accuracy = 0.51
	Kappa = 0.06
		Accuracy = 0.56
		Kappa = 0.14
			Accuracy = 0.55
			Kappa = 0.12

SVM LINEAR	Accuracy = 0.58
	Kappa = 0.07
		Accuracy = 0.58
		Kappa = 0.12
			Accuracy = 0.57
			Kappa = 0.06

SVM RADIAL	Accuracy = 0.57
	Kappa = 0.04
		Accuracy = 0.59
		Kappa = 0.09
			Accuracy = 0.58
			Kappa = 0.04

(Random Forest; KNN, K-nearest neighbor; SVM linear, Support Vector Machine linear; SVM radial, Support Vector Machine radial).

Table 4.5. *Comparison of the predictive performance (Accuracy and Kappa) of the four investigated algorithms in the honey samples training set for regions.*

	DATA FUSION	**VIS-NIR**	**NIR**

RANDOM FOREST	Accuracy = 0.49
	Kappa = 0.04
		Accuracy = 0.15
		Kappa = 0.06
			Accuracy = 0.18
			Kappa = 0.10

KNN	Accuracy = 0.11
	Kappa = 0.04
		Accuracy = 0.10
		Kappa = 0.02
			Accuracy = 0.48
			Kappa = 0.04

SVM LINEAR	Accuracy = 0.16
	Kappa = 0.08
		Accuracy = 0.14
		Kappa = 0.06
			Accuracy = 0.49
			Kappa = 0.04

SVM RADIAL	Accuracy = 0.49
	Kappa = 0.04
		Accuracy = 0.15
		Kappa = 0.06
			Accuracy = 0.14
			Kappa = 0.06

(Random Forest; KNN, K-nearest neighbor; SVM linear, Support Vector Machine linear; SVM radial, Support Vector Machine radial).

Based on the previous tables, the ML models were able to achieve moderate success in classifying honey samples by ecological areas, with the data fusion of the three instrument's absorption data (spectrophotometer, VIS-NIR, and NIR) providing the best performance. Also, classifying honey samples by specific region proved to be more challenging due to the models performing less well. The results suggest that the ML models better classify honey samples by ecological areas than by regions. This may be because ecological areas are a broader territory class (geographic category) encompassing different Italian regions, and the models can capture more general patterns in the data.

Table 4.6. summarizes the cross-validation results of the four supervised models achieved with the combination of the three instruments for ecological areas. The confusion matrix is a visual representation helpful in evaluating the performance of a classification model that shows how class instances have been classified correctly or incorrectly. The number of honey samples that have been extracted from the data set and used in validation are 30% (n = 65). They are represented in rows, the predicted classes that the model has attributed to the honey samples, and columns, which represent the actual classes of the honey samples.

For each algorithm, the following parameters were considered:

- Accuracy: represents the proportion of correctly classified honey samples.
- Sensitivity: represents the proportion of true positive honey samples (TP) relative to the number of true positive and false negative honey samples (FN).
- Specificity: represents the proportion of true negative honey samples (TN) relative to the number of true negative and false positive honey samples (FP).
- Balanced accuracy: represents the average of sensitivity and specificity for each class. This metric is useful when the classes have a non-uniform distribution.

In Table 4.6. Random Forest, SVM linear e il SVM radial shows higher overall accuracy than the table 4.7. (0.58, 0,58 and 0.57, respectively). While these accuracies may seem moderate, it's essential to consider the complexity of the task, because honey samples can exhibit similar spectral signatures across different ecological areas. The KNN model achieved an accuracy of 0.51, which is slightly lower than the others. This suggests that the KNN model may be more prone to overfitting or less effective at classifying the samples (Herrero Latorre et al., 2013).

The models were more able to better classify the SL and NL samples compared to the NM samples. This probably because the NM class is less numerous than the other ones. In fact, looking at the confusion matrix (Table 4.6.), the majority of the errors were attributed to the most numerous class (SL). This is due to the numerical imbalance of the three ecological areas (SL, n = 39; NM, n = 10; NL, n = 16). Models in classifying NM honey, found more likelihood in SL regions, so it can be inferred that ecological areas have similarities. The SL ecological area is very large and includes productions strictly from the plains and hills (<600 m asl) and this has contributed to widening the intrinsic variability of honey.

In Table 4.7., only the classification by Italian regions was considered, selecting a restricted pool of regions according to sample size (Table 3.1.) to avoid the presence of unbalanced classes that can lead to errors in the discrimination. Despite this, the results suggest that the models used have limitations and are unable to easily distinguish between the different Italian regions. The Random Forest and SVM radial models show the best overall performance, with a classification accuracy of 0.49 and a Kappa value of 0.04. The SVM linear and KNN models show much poorer performance, with classification accuracies of 0.16 and 0.11, respectively, and Kappa values of 0.08 and 0.04, respectively. The poor performance was likely due to an excessive intrinsic variability of honey characteristics within each region or a lack of discriminant features between different regions. Therefore, there is no prevalence of one model over another, and cataloguing honey by regions has little significance.

Honey quality characteristics (i.e., proximate composition, elements, chemical traits, colour coordinates) are affected by many factors and its intrinsic variability may be related to physicochemical properties and organoleptic characteristics. Polyflower honey can contain different proportions of polysaccharides, organic acids, minerals, phenolic and aromatic compounds depending on the species of plants that contributed to the nectar. According to Tedesco et al., 2022, honey from the Mediterranean region tends to have higher fructose content and carotenoids and aromatic compounds derived from orange-blossom and lemon-blossom, while honey from the northern regions has a higher glucose content and flavonoids like quercetin and kaempferol. These natural characteristics can vary even within the same region, depending on the type of flowers present in the nectar, the soil, beekeeping practices, processing, and storage conditions.

In a study carried out by Truzzi et al., 2014, polyflower and monofloral honeys typically from plain, hills and mountain territory (acacia, chestnut, honeydew, and sunflower) were compared according to the following parameters: pH, water content, acidity, electrical conductivity, proline, and hydroxy methyl furfural (HMF) content. The present study found that polyflower honey had not proven significantly different values from monofloral species, but rather occupies an intermediate position. Only the HMF content as higher values than the other species. This suggests that polyflower honey may have a shorter shelf-life than monofloral honeys.

A further explanation of the large variability both across regions and ecological areas could be related to the different concentrations of metals in honey samples. A study of De Alda-Garcilope et al. 2012, found a significant effects of content of metals in polyflora hones according to the geographical origin in terms of soil composition and floral type of specific ecological sites in Spain. Moreover, the mineral composition affects the colour of honey, with pale samples having a lower mineral concentration than the darker ones (Chudzinska and Baralkiewicz, 2010). The geographical discrimination of polyflower and monofloral (acacia, sulla, chestnut, orange-blossom, and lemon-blossom) honeys from Sicily and Calabria (Italy) was performed according to the concentration of a pool of elements, especially K, Ca, Mg but also some elements in trace like Cu, Pb, Zn, Cr, Ni (Di Bella et al., 2015). Also, a recent study assessing the origin authentication of honey from some geographical Mediterranean regions (including Slovenia, Croatia, and Bulgaria) reported that the mineral profile of samples showed a wide differentiation, and the concentrations of elements varied greatly from country to country and from the botanical source of honey (Pavlin et al., 2023). As regarding the polyflora honeys, the variation in the mineral content is more pronounced for the major elements, such as potassium, sodium, and calcium, compared with the minor and trace elements. These findings can be due to a complex of factors, such as soil chemistry, mineral soil bioavailability, depth of the root system of plants, which affected the concentration of these minerals in pollens of the different botanical species of flowers, and as a consequence in honey.

Environmental factors such, as temperature, humidity, and sunlight, play a fundamental role in nectar ripening and maturation, as well as the activity of the bees (Di Rosa et al., 2018). According to the 2022 report by the National Honey Observatory (Naldi and Pappalardo, 2022), the honey production under study, which occurred during the spring and summer of 2022, was influenced by above-average temperatures across the entire national territory.

This caused inadequate vernalization of plants and, consequently, prolonged drought, resulting in non-uniform plant blooming.

This intrinsic complexity made very ambitious the goal of discriminating honey using rapid and portable instruments. Furthermore, the study suggests that the combined use of portable and bench-top instruments could improve the classification’s accuracy .

Overall, while the portable instruments offer several advantages, their limitations in honey’s classification suggest the necessity of further technological and methodological improvements.

Table 4.6. *SPECTROPHOTOMETER, VIS-NIR and NIR – DATA FUSION*
Classification performance of the confusion matrices related to the four supervised models in discriminating honey from the three ecological areas according by validation on the spectra data fusion from the three portable instruments.

RANDOM FOREST	Accuracy = 0.58 Kappa = 0.07			Actual		
Predicted	Sensitivity	Specificity	BA	SL	NM	NL
SL	0.90	0.16	0.53	**35 (90%)**	9 (90%)	13 (81%)
NM	0.00	1.00	0.50	0 (0%)	**0 (0%)**	0 (0%)
NL	0.19	0.90	0.55	4 (10%)	1(10%)	**3 (19%)**
Total (n = 65)				39	10	16
Accuracy				0.60	0.85	0.82

KNN	Accuracy = 0.51 Kappa = 0.06			Actual		
Predicted	Sensitivity	Specificity	BA	SL	NM	NL
SL	0.72	0.35	0.54	**28 (72%)**	8 (80%)	10 (60%)
NM	0.04	0.91	0.47	3 (8%)	**0 (0%)**	1 (10%)
NL	0.30	0.80	0.55	8 (20%)	2 (20%)	**5 (30%)**
Total (n = 65)				39	10	16
Accuracy				0.57	0.78	0.66

SVM LINEAR	Accuracy = 0.58 Kappa = 0.07			Actual		
Predicted	Sensitivity	Specificity	BA	SL	NM	NL
SL	0.90	0.16	0.53	**36 (92%)**	9 (90%)	15 (94%)
NM	0.00	1.00	0.50	0 (0%)	**0 (0%)**	0 (0%)
NL	0.19	0.90	0.55	3 (8%)	1 (10%)	**1 (6%)**
Total (n = 65)				39	10	16
Accuracy				0.58	0.83	0.71

SVM RADIAL	Accuracy = 0.57 Kappa = 0.04			Actual		
Predicted	Sensitivity	Specificity	BA	SL	NM	NL
SL	0.91	0.12	0.52	**36 (92%)**	9 (90%)	14 (87%)
NM	0.00	1.00	0.50	0 (0%)	**0 (0%)**	0 (0%)
NL	0.14	0.91	0.52	3 (8%)	1 (10%)	**2 (13%)**
Total (n = 65)				39	10	16
Accuracy				0.60	0.85	0.72

Bold values represent the samples classified correctly. The percentages of assignment by class are expressed into parentheses. Random Forest; KNN, K-nearest neighbour; SVM linear, Support Vector Machine linear; SVM radial, Support Vector Machine radial; BA, Balanced Accuracy; SL, South Lowland; NM, North Mountain; NL, North Lowland).

Table 4.7. *SPECTROPHOTOMETER, VIS-NIR and NIR – DATA FUSION Performance of the supervised models in classifying honey from Italian regions according by validation on the data fusion of the three instruments (for brevity in table is reported the comparison only for four regions).*

RANDOM FOREST	Accuracy = 0.49
	Kappa = 0.04		

Predicted	Sensitivity	Specificity	BA
Piemonte	0.39	0.17	0.52
Abruzzo	0.77	0.88	0.56
Basilicata	0.04	0.40	0.53
Sicilia	0.58	0.56	0.65
Best Region (Sicilia)	0.58	0.56	0.65

KNN	Accuracy = 0.11
	Kappa = 0.04		

Predicted	Sensitivity	Specificity	BA
Piemonte	0.09	0.91	0.50
Abruzzo	0.16	0.90	0.53
Basilicata	0.21	0.89	0.55
Sicilia	0.20	0.89	0.55
Best Region (Sardegna)	0.20	0.96	0.58

SVM LINEAR	Accuracy = 0.16
	Kappa = 0.08		

Predicted	Sensitivity	Specificity	BA
Piemonte	0.21	0.91	0.56
Abruzzo	0.22	0.89	0.55
Basilicata	0.36	0.86	0.61
Sicilia	0.39	0.85	0.62
Best Region (Sicilia)	0.39	0.85	0.62

SVM RADIAL	Accuracy = 0.49
	Kappa = 0.04		

Predicted	Sensitivity	Specificity	BA
Piemonte	0.37	0.18	0.51
Abruzzo	0.77	0.89	0.56
Basilicata	0.07	0.37	0.54
Sicilia	0.56	0.56	0.59
Best Region (Sicilia)	0.56	0.56	0.59

The classification of the Regions is made according to the number of its samples. Best Region chooses based on the highest value close to 1. (Random Forest; KNN, K-nearest neighbor; SVM linear, Support Vector Machine linear; SVM radial, Support Vector Machine radial; BA, Balanced Accuracy).

CHAPTER 5

CONCLUSIONS

NIR spectroscopy is a multi-analytical technique that offers numerous advantages, particularly in the food sector, where it is widely used to ensure product traceability and authenticity, which are of fundamental importance. However, the use of portable instruments with a more limited wavelength range has place greater challenges in determining the geographical origin of the samples.

The results of the experimental trial related to the present MSc dissertation show that classifying the samples based on individual Italian regions is more complex. Specifically, none of the machine learning models used (Random Forest, KNN, linear SVM, and SVM radial) demonstrated an adequate classification ability based on the observed accuracies. Indeed, the results show similar accuracy levels, denoting obstacles in the exam of honey’s variability using predictive models. This suggests that the spectral differences between samples from different Italian regions are less marked, making it more difficult to achieve significant discrimination of honey based on regional level. Ecological areas, defined as regions with the same altitude, seem to better reflect the characteristics of honey. In this context, therefore, it was more logical and reasonable to classify honey by ecological areas rather than by regions, because the models have achieved a higher accuracy.

This study also highlighted the importance of adopting a more holistic approach to the classification of polyflower honey. As a natural product, honey's intrinsic complexity makes it challenging to control, even within a single region. Moreover, honey's composition and properties are difficult to analyze, and they can differ considerably due to the unique characteristics of each bloom. This research not only provides hints into the complexity of honey’s classification but also highlights the importance of further innovation of analytical techniques to backing the authenticity and traceability of Italian honey.

CHAPTER 6

APPENDIX – Graphical abstract and experimental pictures

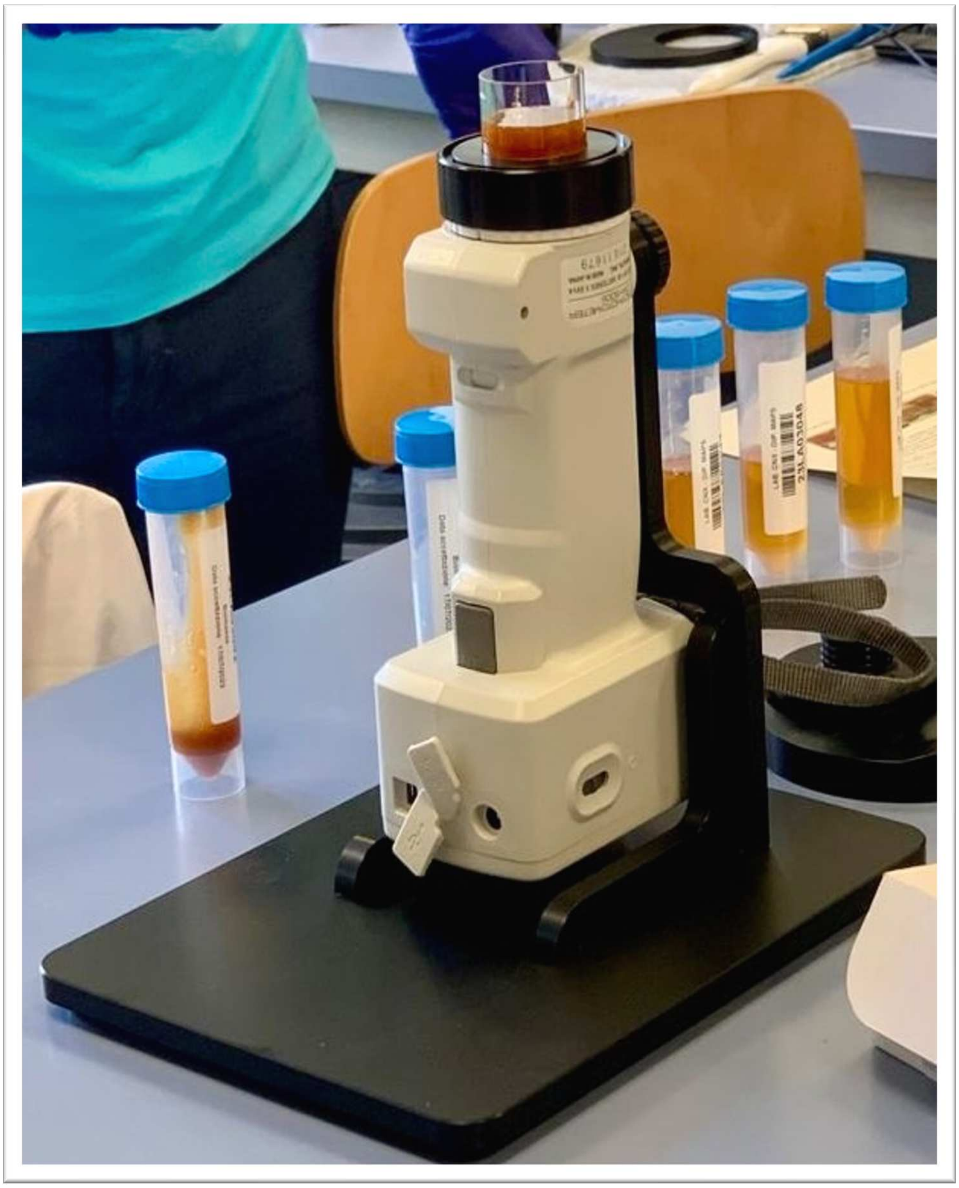


Figure 6.1. *Portable VIS spectrophotometer CM-600d, (Konica Minolta Ramsey Sensing, Inc). Picture shows the spectrophotometer fixed in a vertical device, and the cuvette placed on top of the illuminant source. The measurement was conducted in a dark room and the cuvette was rotated by approximately 120° after each scan.*

(credit: Silvia Zanotto)

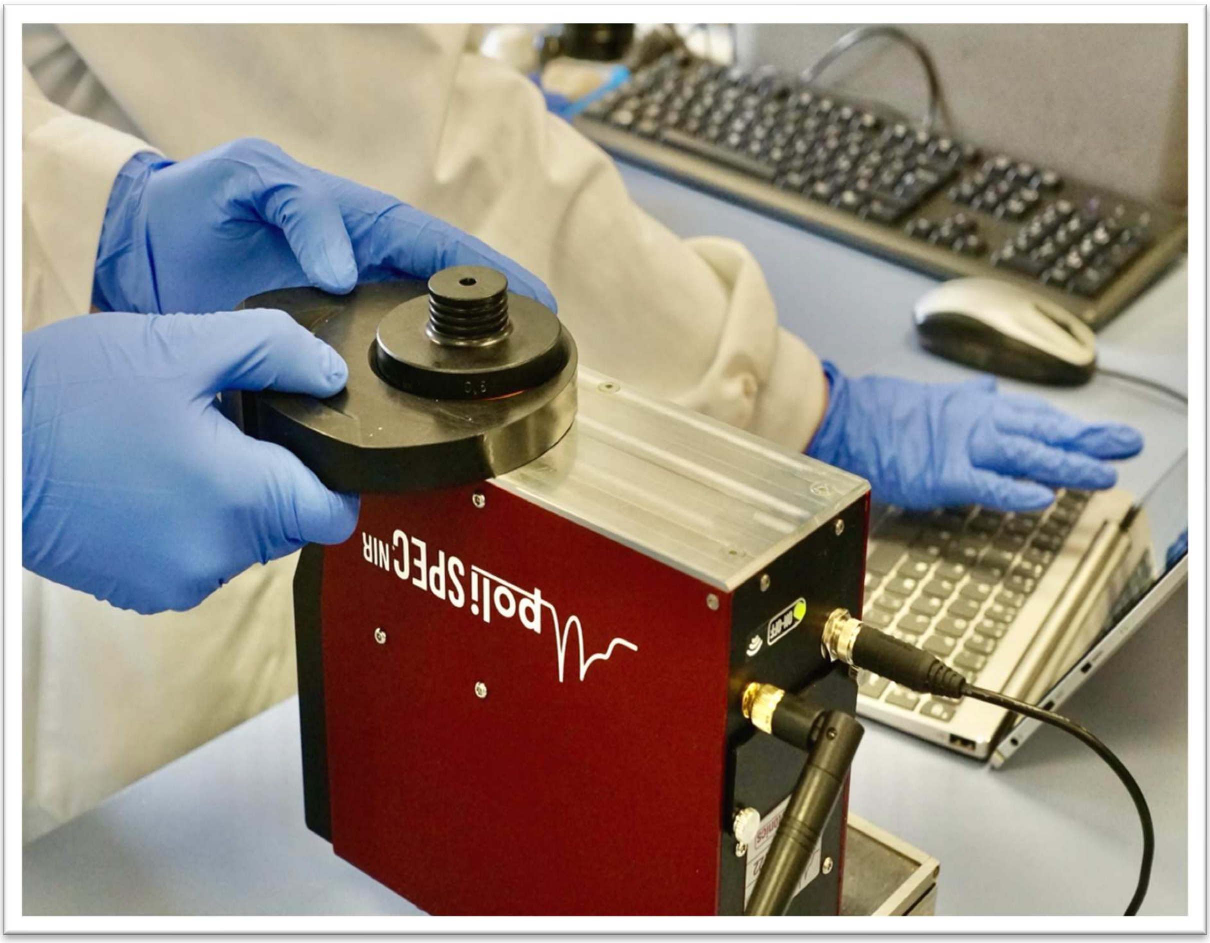


Figure 6.2. *Portable NIR (PolispecNIR, ITPhotonics).*

Picture shows a sample placed on a ring cell with the quartz window above the illuminant source. The analysis was conducing first on the NIR and then on the VIS-NIR.

(credit: Silvia Zanotto)

The following ***Figures 6.3.A., 6.3.B.*** *and* ***6.3.C.*** *show some honey samples in chromatic order ready to be analyze, after the heat pre-treatment. Data were exported in HUNTER L*a*b* system; where L* is lightness, a* is redness, and b* is yellowness. (credit: Silvia Zanotto)*

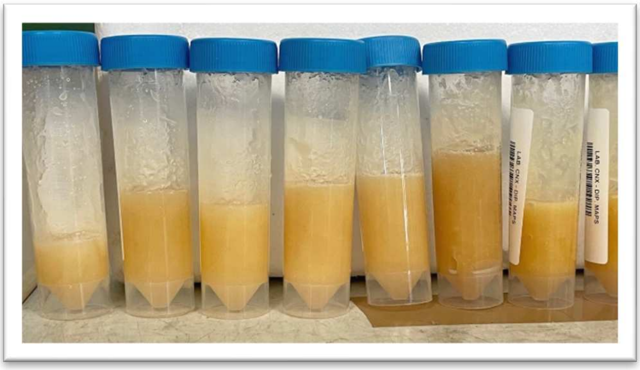


Figure 6.3.A. *The samples had on average these values: L* 45.1, a* 0.4, b* 9.2 .*

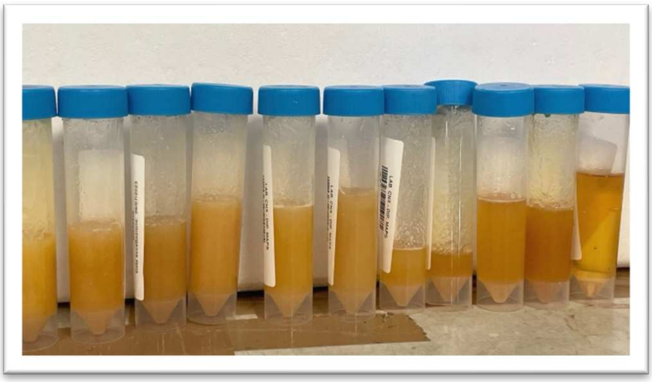


Figure 6.3.B. *The samples above had on average these values: L* 28.2, a* 1.1, b* 6.5 .*

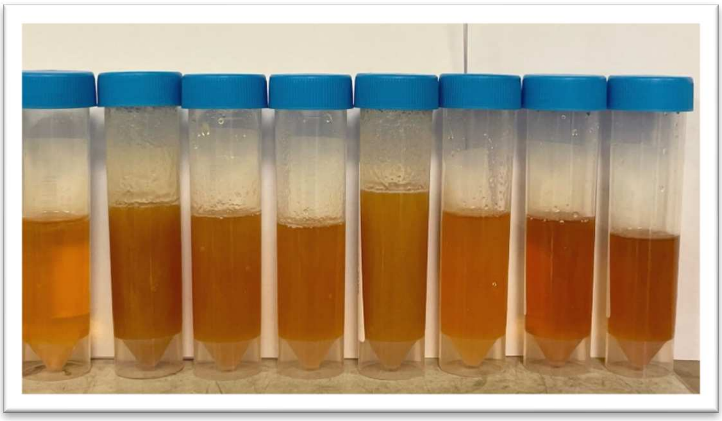


Figure 6.3.C. *The samples had on average these values: L* 28.9, a* 0.6, b* 4.3 .*

ACKNOWLEDGEMENTS

Me, supervisor, and co-author are grateful to **Lucia Piana** and **Giancarlo Naldi** from Osservatorio Nazionale Miele for their valuable collaboration in honey sample collection.

We are also thankful to Dr. **Roberto Piro**, **Alessandra Tata**, **Marco Bragolusi** and **Carmela Zacometti** of Istituto Zooprofilattico Sperimentale delle Venezie, Experimental Chemistry Laboratory for their valuable technical support in honey collection and sampling and classification.

We are truly grateful to Dr. **Lorenzo Serva** of Department of Animal Medicine, Production and Health (MAPS) of Padova University for his precius technical support in data processing and modeling.

We are appreciative to Dr. **Elisabetta Garbin**, **Sandro Tenti** and **Massimo Mirisola** of LabCNX of Dept. Animal Medicine, Production and Health (MAPS) of Padova University for their valuable technical support in VIS, VIS-NIR and NIR data collection.

We are also grateful to Dr. **Sarah Currò** and prof. **Luca Fasolato** of Dept. Comparative Biomedicine and Food Science (BCA) of Padova University for their valuable technical support in VIS-NIR and NIR data collection.

REFERENCES

- 2004. G.U. n. 168 del 20 luglio 2004. Decreto legislativo 21 maggio 2004, n. 179. Attuazione della direttiva 2001/110/CE concernente la produzione e la commercializzazione del miele.

- 2011. Regolamento (UE) n. 1169/2011 del Parlamento Europeo e del Consiglio del 25 ottobre 2011 relativo alla fornitura di informazioni sugli alimenti ai consumatori.

Barthès B., Kouakoua E., Clairotte M., Lallemand J., Chapuis-Lardy L., Rabenarivo M., Roussel S. (2019). Performance comparison between a miniaturized and a conventional near infrared reflectance (NIR) spectrometer for characterizing soil carbon and nitrogen. – Geoderma, Volume 338, Pages 422-429. https://DOI.org/10.1016/j.geoderma.2018.12.031

Bisutti V., Merlanti R., Serva, L., Lucatello L., Mirisola M., Balzan S., Tenti S., Fontana F., Trevisan G., Montanucci L., Contiero B., Segato S., Capolongo F. (2019). Multivariate and machine learning approaches for honey botanical origin authentication using near infrared spectroscopy. Journal of Near Infrared Spectroscopy, 27(1), 65-74. https://DOI.org/10.1177/0967033518824765

Chen G., Huang Y., Chen K. (2014). Recent advances and applications of near infrared spectroscopy for honey quality assessment. Advance Journal of Food Science and Technology, 461-467, 6(4). https:// DOI.org/10.19026/ajfst.6.55

Chudzinska, M., Baralkiewicz, D. (2010). Estimation of honey authenticity by multielements characteristics using inductively coupled plasma-mass spectrometry (ICP-MS) combined with chemometrics. Food and Chemical Toxicology, 48, 284–290. DOI: 10.1016/j.fct.2009.10.011

Contessi A. (2004). Le api: biologia, allevamento, prodotti . Editrice Edagricole, Bologna.

Currò S., Fasolato L., Serva L., Boffo L., Ferlito J. C., Novelli E., Balzan S. (2022). Use of portable near infrared tool for rapid on-site inspection of freezing and hydrogen peroxide treatment of cuttlefish (Sepia officinalis). Food Control, 132. 108524. https://DOI.org/10.1016/j.foodcont.2021.108524

De Alda-Garcilope C., Gallego-Picó A., Bravo-Yagüe J. C., Garcinuño-Martínez R.M., Fernández-Hernando P. (2012). Characterization of Spanish honeys with protected designation of origin Miel de Granada according to their mineral content. Food Chemistry, Volume 135, Issue 3,Pages 1785-1788. https://doi.org/10.1016/j.foodchem.2012.06.057.

Di Bella G., Lo Turco V., Potortì A. G., Bua G. D., Fede M. R., Dugo G. (2014). Geographical discrimination of Italian honey by multi-element analysis with a chemometric approach. Journal of Food Composition and Analysis, Volume 44, Pages 25-35. https://doi.org/10.1016/j.jfca.2015.05.003

Di Rosa A. R., Leone F., Cheli F., & Chiofalo V. (2018). Novel approach for the characterization of Sicilian honeys based on the correlation of physico-chemical parameters and artificial senses. Italian Journal of Animal Science, 18(1), 389–397. https://DOI.org/10.1080/1828051X.2018.1530962

Frausto-Reyes C., Casillas-Peñuelas R., Quintanar-Stephano J. L., Macías-López E., Bujdud-Pérez M. J., Medina-Ramírez I. (2017). Spectroscopic study of honey from Apis mellifera from different regions in Mexico. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy. 178, 212–217. http://dx.DOI.org/10.1016/j.saa.2017.02.009

Gallai N., Salles J. M., Settele J., Vaissière B. E. (2009). Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. Economia ecologica, vol. 68, no. 3, 810-821. https://DOI.org/10.1016/j.ecolecon.2008.06.014

Herrero Latorre C., Peña Crecente R.M., García Martín S., Barciela García J. (2013). A fast chemometric procedure based on NIR data for authentication of honey with protected geographical indication. Food Chemistry 141, 3559–3565. http://dx.DOI.org/10.1016/j.foodchem.2013.06.022

Jimenez-Carvelo A. M., Gonzalez-Casado A., Bagur-Gonzalez M. G., Cuadros-Rodriguez L. (2019). Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review. Food Research International 122, 25–39. https://DOI.org/10.1016/j.foodres.2019.03.063

Keserer E. (2023). 8 Types of Machine Learning Classification Algorithms. www.akkio.com (access 3/5/24)

Kunat-Budzynska, M., Rysiak, A., Wiater, A., Graz, M., Andrejko, M., Budzynska, M., Brys, M.S., Sudzinski, M., Tomczyk, M., Gancarz, M., Rusinek, R., Ptaszynska, A.A. (2023). Chemical Composizione and Antimicrobial Activity of New Honey Varietals. Int. J. Environ. Res. Public Health, 20(3), 2458.
https://DOI.org/10.3390/ijerph20032458

Lanza I., Conficoni D., Balzan S., Cullere M., Fasolato, L., Serva L., Contiero B., Trocino A., Marchesini G., Xiccato G., Novelli E., Segato S. (2021). Assessment of chicken breast shelf life based on bench-top and portable near-infrared spectroscopy tools coupled with chemometrics. Food Quality and Safety, 5, 1–11.
DOI: 10.1093/fqsafe/fyaa032

Lanza I., Currò S., Segato S., Serva L., Cullere M., Castellani P., Fasolato L., Pasotto D., Dalle Zotte A. (2023). Spectroscopic methods and machine learning modelling to differentiate table eggs from quails fed with different inclusion levels of silkworm meal. Food Control, 147. 109589.
https://DOI.org/10.1016/j.foodcont.2022.109589

Maione C., Barbosa Jr F., Melgaço Barbosa R. (2019). Predicting the botanical and geographical origin of honey with multivariate T data analysis and machine learning techniques: A review. Computers and Electronics in Agriculture. 157, 436–446.
https://DOI.org/10.1016/j.compag.2019.01.020

McGratha T. F., Haugheya S. A., Pattersona J., Fauhl-Hassekb C., Donarskic J., Alewijnd M., van Ruthd S., Elliotta C. T. (2018). What are the scientific challenges in moving from targeted to non-targeted methods for food fraud testing and how can they be addressed? – Spectroscopy case study. Food Science & Technology 76, 38–55.
https://DOI.org/10.1016/j.tifs.2018.04.001

Massaro A., Zacometti C., Bragolusi M., Bucek J., Piro R., Tata A. (2024). Authentication of the botanical origin of monofloral honey by dielectric barrier discharge ionization high resolution mass spectrometry (DBDI-HRMS). Breaching the 6 s barrier of analysis time. Food Control, 160, 110330.
https://DOI.org/10.1016/j.foodcont.2024.110330

Mateo F., Tarazona A., Mateo E. M. (2021). Comparative Study of Several Machine Learning Algorithms for Classification of Unifloral Honeys. Foods, 10, 1543.
https://doi.org/10.3390/foods10071543

Messa M. (2018). La differenza tra spettrofotometri e colorimetri. Konica Minolta. https://www.plastix.it/author/maurizio-messa-konica-minolta/ (access 4/5/24)

Naldi G. (2020) I MIELI ITALIANI: un patrimonio unico di qualità e tipicità. Da oltre mille analisi i punti di forza per la valorizzazione. Pubblicazione multimediale, Sulla base di dati analitici dei mieli partecipanti al CONCORSO TRE GOCCE D'ORO – GRANDI MIELI D'ITALIA. Osservatorio Nazionale del Miele.

Naldi G., Pizzirani C. (2015). Dall'ape ai mieli, Piccola guida per conoscere e gustare. Osservatorio Nazionale del Miele.

Ondalys (2019) https://ondalys.fr/en/scientific-resources/machine-learning-methods/ (access 11/05/24)

Osborne B.G., Fearn T. (1993). Near infrared Spectroscopy in Food Analysis. BRI Australia Ltd, North Ryde, Australia. https://DOI.org/10.1002/9780470027318.a1018

Pappalardo S., Naldi G. (2023). IL VALORE DELLA TERRA: agricoltura e nuova ruralità, economia e sostenibilità, qualità e consumo consapevole. Rivista Multimediale no. 1/2023. Osservatorio Nazionale del Miele.

Pasquini C. (2003). Near Infrared Spectroscopy: fundamentals, practical aspects and analytical applications. J. Braz. Chem. Soc, 14 (2). https://DOI.org/10.1590/S0103-50532003000200006

Pavlin, A., Kocˇar, D., Imperl, J., Kolar, M., Marolt, G., Petrova, P. (2023) Honey Origin Authentication via Mineral Profiling Combined with Chemometric Approaches. Foods 2023, 12, 2826. https://doi.org/10.3390/ foods12152826

Piana L. (1994). Miele di qualità: tecniche di produzione e lavorazione. Temi di apicoltura moderna. Regione Toscana, 210-223.

Piana L., Naldi G. (2020). Cos'è il miele? Guida ai Mieli d'Italia, un patrimonio unico al mondo . Rivista multimediale. Osservatorio Nazionale del Miele.

Singh P., Pandey S., Manik S. (2024). A comprehensive review of the Dairy Pasteurization Process using machine learning models. Food Control, 110574. https://DOI.org/10.1016/j.foodcont.2024.110574

Segato, S., Merlanti, R., Bisutti, V., Montanucci L., Serva L., Lucatello L., Mirisola M., Contiero B., Conficoni D., Balzan S., Marchesini G., Capolongo F. (2019). Multivariate and machine learning models to assess the heat effects on honey physicochemical, colour and NIR data. European Food Research and Technology, 245: 2269–2278. DOI:10.1007/s00217-019-03332-x

Reich G. (2005).Near-infrared spectroscopy and imaging: Basic principles and pharmaceutical applications. Advanced Drug Delivery Reviews, Volume 57, Issue 8, 2005, Pages 1109-1143, ISSN 0169-409X, https://doi.org/10.1016/j.addr.2005.01.020.

Roussel S. A., Igne B., Funk D. B., Hurburgh C. R., (2011). Noise Robustness Comparison for Near Infrared Prediction Models. J.NIRS, Volume: 19, Issue: 1, Pages: 23. https://DOI.org/10.1255/jnirs.916

Tao Y., Tian H., Zhao K., Yu Y., Liu G., Bai X. (2024). High-precision Discrimination of Maize Silage Based on Olfactory Visualization Technology Integrated with Chemometrics Analysis. BioResources 19(2), 3597-3613. DOI: 10.15376/biores.19.2.3597-3613

Tedesco R., Scalabrin E., Malagnini V., Strojnik L., Ogrinc N., Capodaglio G. (2022). Characterization of Botanical Origin of Italian Honey by Carbohydrate Composition and Volatile Organic Compounds (VOCs). Foods. 2022; 11(16):2441. https://DOI.org/10.3390/foods11162441

Truzzi C., Illuminati S., Annibaldi A., Finale C., Rossetti M., Scarponi G. (2014).Physicochemical Properties of Honey from Marche, Central Italy: Classification of Unifloral and Multifloral Honeys by Multivariate Analysis. Natural Product Communications. ;9(11). DOI:10.1177/1934578X1400901117

Woodcock T., Downey G., Kelly J. D., O’Donnell C. (2007). Geographical Classification of Honey Samples by Near-Infrared Spectroscopy: A Feasibility Study. Journal of Agricultural and Food Chemistry 2007 55 (22), 9128-9134. DOI: 10.1021/jf072010q