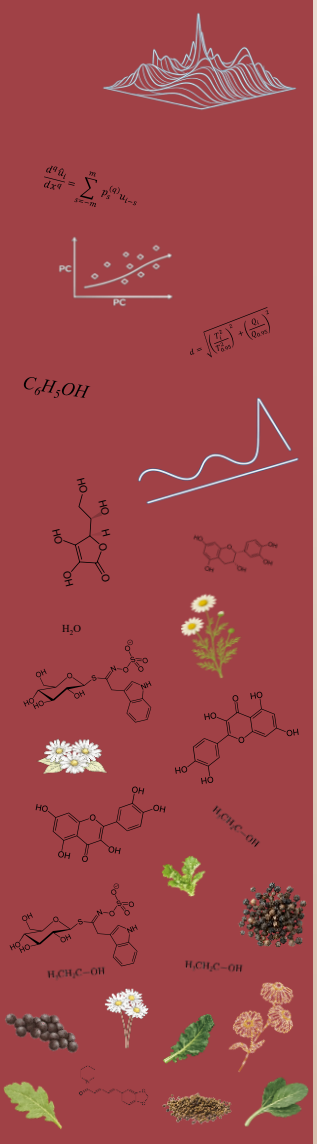
CLAUDIA SCAPPATICCI

Development of analytical and chemometric approaches for studying vegetable matrices to evaluate compositional differences and bioactive properties



UNIVERSITÀ DEGLI STUDI DELL'AQUILA
DIPARTIMENTO DI SCIENZE FISICHE E CHIMICHE

Dottorato di Ricerca in Scienze Fisiche e Chimiche

Curriculum Chimica Analitica

XXXVIII ciclo

Development of analytical and chemometric approaches for studying vegetable matrices to evaluate compositional differences and bioactive properties

SSD CHEM-01/A

Dottorando

Claudia Scappaticci

Coordinatore del corso

Prof. Massimiliano Aschi

Tutor

Prof. Angelo Antonio D'Archivio

Dr.ssa. Alessandra Biancolillo

a.a. 2024/2025

TABLE OF CONTENTS

CHAPTER 1

THE INCREASING INTEREST IN BIOACTIVE COMPOUNDS FROM PLANT MATRICES

CHAPTER 2

ANALYTICAL TECHNIQUES

CHAPTER 3

CHEMOMETRIC APPROACHES

Section 1

Brassicaceae

CHAPTER 4

MORPHOLOGICAL ANALYSIS AND OPTIMISATION OF EXTRACTION FOR THE ANALYSIS OF THE SPECTROCHROMATOGRAPHIC PROFILE OF SPECIES BELONGING TO THE BRASSICACEAE FAMILY AND EVALUATION OF ANTIOXIDANT ACTIVITY

Section 2

Chamomile (*Matricaria chamomilla L.*)

CHAPTER 5

OPTIMISATION OF COLD EXTRACTION OF BIOACTIVE SUBSTANCES FROM CHAMOMILE FLOWERS THROUGH CHEMOMETRIC APPROACHES

Section 3

Black pepper

CHAPTER 6

COMPARISON OF THE ANTIOXIDANT ACTIVITY OF DIFFERENT VARIETIES OF PEPPER

CHAPTER 7

Conclusion and final remarks

CHAPTER 8

Way forward

Abstract

This research aim to develop and apply multi-analytical and chemometric approaches to plant matrices belonging to the Brassicaceae, Asteraceae and Piperaceae families, in order to evaluate their compositional differences and bioactive properties. The study focused on the identification, extraction and characterisation of bioactive compounds, in particular polyphenols, flavonoids and glucosinolates, which are responsible for antioxidant activity and beneficial properties for human health.

Non-destructive techniques such as digital microscopy (E-Eye) and attenuated total reflection infrared spectroscopy (ATR-FT-IR) were used to characterise different varieties of Brassicaceae grown in the same experimental field, eliminating soil and climate variability. SIMCA and PLS-DA chemometric algorithms enabled the classification and discrimination of local varieties, with particular attention to Mugnoli di Pettorano sul Gizio, whose hybrid nature was confirmed by multivariate analysis.

A D-Optimal design was applied to optimise the extraction parameters (sample quantity, temperature, time and composition of the extracting mixture) of Brassicaceae leaves using HPLC-DAD and UV-Vis analysis. The correlation between chromatographic data and antioxidant activity was modelled using the multi-way N-PLS algorithm, obtaining a robust predictive model.

Subsequently, a full factorial design was used to optimise the cold extraction of bioactive compounds from *Matricaria chamomilla* L. Chemometric analysis (PCA and ASCA) highlighted the effect of temperature, time and quantity on the polyphenolic profile and spectral characteristics. The optimal conditions (2.5 g, 25 °C, 32 min) produced an extract rich in hydroxybenzoic and hydroxycinnamic acids with high antioxidant activity.

Finally, a multi-analytical approach was applied to *Piper nigrum* L. to characterise its chemical composition, antioxidant properties, and geographical traceability. The study analysed black pepper samples from Kampot (Cambodia, Protected Geographical Indication, one of which also organic), Madagascar, and commercial products, along with a green pepper sample for comparison.

The aromatic profile, determined by headspace solid-phase microextraction coupled with gas chromatography–mass spectrometry (HS-SPME-GC-MS), revealed compositional differences related to geographical origin, dominated by terpenoid compounds such as β-caryophyllene, 3-carene, limonene, and sabinene. Principal Component Analysis (PCA) enabled the identification of distinctive volatile markers associated with each provenance.

Antioxidant activity, assessed through the DPPH assay, showed the highest radical-scavenging capacity in Madagascar pepper (IC_{50} = 7.94 mg/mL), whereas Kampot pepper exhibited higher IC_{50} values, indicating lower antioxidant efficiency.

Infrared spectroscopy (ATR-FT-IR), combined with chemometric methods (PLS-DA), allowed non-destructive classification of samples according to their geographical origin, achieving an accuracy greater than 90%.

Overall, the integrated approach demonstrated that combining volatile profile analysis, antioxidant evaluation, and vibrational spectroscopy provides an effective tool for the characterisation and geographical authentication of black pepper, highlighting the influence of origin and processing factors on its bioactive properties.

CHAPTER 1

THE INCREASING INTEREST IN BIOACTIVE COMPOUNDS FROM PLANT MATRICES

SUMMARY: 1. Bioactive compounds contained in plant matrices – 2. Plants of interest – 3. Evolution of analytical investigation – 4. Objectives of the research work

1. Bioactive compounds contained in plant matrices

Numerous plant species are known for their beneficial properties for humans thanks to the presence of pharmacologically active substances within them. Common species such as those belonging to the Brassicaceae family are rich in medicinal properties such as antioxidants and anti-tumour agents, while other species such as chilli peppers (*Capsicum spp.*)[1] are rich in antimicrobial properties and have anti-mutagenic and anti-fungal activity. They also have antidiabetic, anti-obesity, cardiovascular and anti-inflammatory effects due to the presence of phenolic compounds, ascorbic acid and carotenoids[2].

Since ancient times, commonly used vegetables such as garlic[3] and onions have been included in the diet because they are rich in phytonutrients such as

[1] H. Duranova, V. Valkova, and L. Gabriny, "Chili peppers (Capsicum spp.): the spice not only for cuisine purposes: an update on current knowledge," vol. 21, no. 4. Springer Netherlands, 2022.

[2] T. Ahmad *et al.*, "Phytochemicals in daucus carota and their health benefits—review article," *Foods*, vol. 8, no. 9, pp. 1–22, 2019.

[3] L. Bayan, P. H. Koulivand, and A. Gorji, "Garlic: a review of potential therapeutic effects," *Avicenna J. phytomedicine*, vol. 4, no. 1, pp. 1–14, Jan. 2014.

thiosulfinates, which prevent coronary heart disease and gastrointestinal disorders and treat hypercholesterolemia and hypertension[4][5].

Knowledge of the presence of bioactive compounds in plant species has encouraged the study of plants in order to make them primary sources of compounds beneficial to human health. In addition to the pharmacological aspect, the cosmetic and agri-food sectors have also been involved in order to identify substances useful for improving health conditions and for the treatment and prevention of numerous diseases.

Phytochemical profiles are influenced by climatic conditions. It is well known that plants of the same species grown in different geographical areas have different chemical compositions because they are subjected to different soil and climatic conditions and agronomic practices that affect their growth and, consequently, the development of substances within them[6]. Therefore, it is necessary to study their chemical composition, as heterogeneity has an impact on the biological efficacy of plant extracts, their reproducibility in cultivation and standardisation[7], and to apply advanced chemometric approaches to interpret the data obtained from the food matrices analysed[8].

For this reason, it is necessary to accurately characterise batches according to their origin in order to avoid problems of phytochemical heterogeneity, which directly affects biological activity. Isothiocyanates derived from glucosinolates[9], which are known for their anti-tumour and anti-inflammatory properties, can vary in structure and concentration depending on the temperature, humidity and phenological phase at the time of harvest. Following the identification of pharmacologically active compounds, it is possible to evaluate biological activities such as antioxidant activity, also in relation to variations between species belonging to the same family, and

---

[4] L. Bayan, P. H. Koulivand, and A. Gorji, "Garlic: a review of potential therapeutic effects," *Avicenna J. phytomedicine*, vol. 4, no. 1, pp. 1–14, Jan. 2014.

[5] V. Lanzotti, "The analysis of onion and garlic," *J. Chromatogr. A*, vol. 1112, no. 1–2, pp. 3–22, 2006.

[6] E. Cazzaniga, N. Cavallini, A. Giraudo, G. Gavoci, F. Geobaldo, M. Pariani, D. Ghirardello, G. Zeppa, and F. Savorani, "Lipids in a nutshell: Quick determination of lipid content in hazelnuts with NIR spectroscopy," Foods, vol. 12, p. 34, 2023.

[7] E. L. Connolly *et al.*, "Glucosinolates From Cruciferous Vegetables and Their Potential Role in Chronic Disease: Investigating the Preclinical and Clinical Evidence," *Front. Pharmacol.*, vol. 12, no. October, pp. 1–12, 2021.

[8] A. Laftouhi *et al.*, "Effect of Climate Change on the Phytochemical Constituents, Essential Oil Yield and Chemical Composition of Inula viscosa Leaves," *Trop. J. Nat. Prod. Res.*, vol. 8, no. 5, pp. 7073–7081, 2024.

[9] T. Bohinc and S. Trdan, "Environmental factors affecting the glucosinolate content in Brassicaceae," *J. Food, Agric. Environ.*, vol. 10, no. 2, pp. 357–360, 2012.

identify which molecules are responsible for these properties and specific behaviours.

2. Plants of interest

Plants are rich in bioactive compounds that are of great interest to many sectors, such as food, pharmaceuticals and cosmetics. It has been shown that plant extracts have significant biological properties, such as antioxidant and anti-inflammatory properties, thanks to the presence of secondary metabolites such as polyphenols, flavonoids and alkaloids[10]. Therefore, this study focused on the analysis of selected plant species belonging to three botanical families, namely Brassicaceae, Asteraceae and Piperaceae, in order to enhance their application potential by investigating their chemical characteristics and biological activities.

Polyphenols are among the most abundant compounds in plants and are responsible for antioxidant[11] and anti-inflammatory[12] activities in various food matrices, such as Brassicaceae, contributing to the different pigmentations assumed by foods. These are indicators of food freshness because, being subject to enzymatic degradation by polyphenol oxidase, they change colour as they age, which then contributes to their flavour. Among the phenolic compounds, the most commonly found in plant foods, and consequently within the Brassicaceae family, are flavonoids such as catechin and epicatechin, quercetin, luteolin, apigenin, which are among the most important antioxidants present in this plant species. The antioxidant action of these compounds varies according to their chemical structure, and increases with the presence of hydroxyl groups capable of neutralising free radicals[13]. Hydroxybenzoic acids such as p-coumaric acid, ferulic acid and, although in smaller quantities, gallic acid also contribute to health benefits. Therefore, the

---

[10] M. Muthusamy and S. I. Lee, "Abiotic stress-induced secondary metabolite production in Brassica: opportunities and challenges," *Frontiers in Plant Science*, vol. 14, Art. no. 1323085, Jan. 2024.

[11] R. Tsao, "Chemistry and biochemistry of dietary polyphenols," *Nutrients*, vol. 2, no. 12, pp. 1231–1246, Dec. 2010.

[12] E. Sangiovanni and M. Dell'Agli, "Special Issue: Anti-inflammatory activity of plant polyphenols," Biomedicines, vol. 8, no. 3, p. 64, Mar. 2020.

[13] K. E. Heim, A. R. Tagliaferro, and D. J. Bobilya, "Flavonoid antioxidants: Chemistry, metabolism and structure-activity relationships," *J. Nutr. Biochem.*, vol. 13, no. 10, pp. 572–584, 2002.

direct attribution of antioxidant capacity is mostly due to the total phenolic content[14]. Phenolic compounds also exhibits biological activities such as anti-inflammatory and vasoprotective properties. The presence of glucosinolates such as glucoraphanin, generated by the enzyme myrosinase during chewing, produces sulforaphane (belonging to the isothiocyanate class) which acts as an estrogen antagonist, reducing the proliferation of cancer cells[15].

Ascorbic acid contributes to maintaining oxidative balance, i.e. protecting cells from attack by free radicals. It keeps tissues and the immune system healthy by detoxifying reactive oxygen species, thus protecting them from oxidative stress.

Flavonoids and phenolic acids, responsible for antioxidant and anti-inflammatory activity, are common to plants and flowers belonging to the Asteraceae family, which, in addition to being rich in polyphenols including caffeic and chlorogenic acid, are also rich in essential oils and sesquiterpene lactones. The latter protect the species from attack by herbivores such as grazing animals and rodents due to their bitter taste. They also have spasmolytic and soothing properties. Brassicaceae, on the other hand, defend themselves by producing glucosinolates, which release pungent and irritating isothiocyanates capable of deterring herbivorous insects such as caterpillars and beetles. They also produce phenols and flavonoids with antimicrobial functions, useful for counteracting pathogens like fungi and bacteria.[16]. Climate change has inevitably led plants and flowers to adapt to survive in new conditions, thus reducing environmental stress caused by salinity, drought and extreme temperatures, by varying the production of these compounds, which involve genetic regulation[17].

Chamomilla recutita is a plant with soothing and digestive properties, whose extracts are commonly used to treat gastrointestinal disorders,

---

[14] C. A. Rice-Evans, N. J. Miller, and G. Paganga, "Structure-antioxidant activity relationships of flavonoids and phenolic acids," *Free Radic. Biol. Med.*, vol. 20, no. 7, pp. 933–956, 1996.

[15] M. Nestle, "Broccoli Sprouts in Cancer Prevention," *Nutr. Rev.*, vol. 56, no. 4, pp. 127–130, 1998.

[16] P. Kumar, D. Kumar, S. Pal, and S. Singh, "Plant secondary metabolites in defense against phytopathogens: Mechanisms, biosynthesis, and applications," *Physiol. Mol. Plant Pathol.*, vol. 138, p. 102639, 2025.

[17] K. Goura *et al.*, "Beyond Survival: The Role of Secondary Metabolites in Plant Defense Mechanisms," *J. Crop Heal.*, vol. 77, no. 4, p. 121, 2025.

rheumatic pain, insomnia and ulcers. The presence of essential oils[18], makes it a key ingredient in cosmetic and medicinal preparations[19].

Its main constituents include apigenin, luteolin, quercetin, lactones, terpenoids (which can act as antidepressants and anxiolytics)[20] and mucilage which, being hydrophilic substances, form a viscous gel with emollient and gastroprotective properties used in medicine to promote a feeling of satiety.

Chamomile, rich in bioactive substances, is used in phytotherapy[21] and included in major pharmacopoeias[22].

These properties have allowed herbal teas and infusions to be included among nutraceutical products. These are also considered in cosmetics[23] for their natural lightening and brightening, decongestant and anti-ageing effects, thanks to the presence of antioxidants that reduce oxidative stress, itching and irritation. The aromas of essential oils are used as natural flavourings in the confectionery industry and in the food industry for the production of yoghurt[24], bars and ready-made herbal teas that aid digestion, and to extend shelf life by reducing lipid oxidation and microbial growth[25].

Just as the aroma and beneficial properties of chamomile allow for applications in various fields, *Piper nigrum L.*, belonging to the Piperaceae family, is also one of the most widely used spices. The most common commercial forms are green, white and black pepper, which are derived from the same fruit but processed in different ways. The first is peeled pepper; green pepper is obtained by freeze-drying or treating immature berries with SO_2 (to maintain their colour), while black pepper is obtained by sun drying ripe green berries for several days.

---

[18] E. Lemberkovics, A. Kéry, G. Marczal, B. Simándi, and E. Szöke, "Phytochemical evaluation of essential oils, medicinal plants and their preparations," *Acta Pharm. Hung.*, vol. 68, no. 3, p. 141—149, 1998.

[19] K. H. C. Baser *et al.*, "The Essential Oil Constituents and Antimicrobial Activity of Anthemis aciphylla B OISS. var . discoidea BOISS," vol. 54, no. 2, pp. 222–225, 2006.

[20] D. L. McKay and J. B. Blumberg, "A review of the bioactivity and potential health benefits of chamomile tea (Matricaria recutita L.).," *Phytother. Res.*, vol. 20, no. 7, pp. 519–530, Jul. 2006.

[21] O. Singh, Z. Khanam, N. Misra, and M. K. Srivastava, "Chamomile (Matricaria chamomilla L.): An overview," *Pharmacogn. Rev.*, vol. 5, no. 9, pp. 82–95, 2011.

[22] M. M. El Joumaa and J. M. Borjac, "Matricaria chamomilla: A valuable insight into recent advances in medicinal uses and pharmacological activities," *Phytochem. Rev.*, vol. 21, no. 6, pp. 1913–1940, Dec. 2022.

[23] K. P. Gaikwad *et al.*, "Unwind with Beauty: Exploring Chamomile's Cosmetics Values," *Int. J. Pharm. Sci. Rev. Res.*, vol. 85, no. 3, pp. 73–78, 2025.

[24] C. Caleja *et al.*, "Development of a functional dairy food: Exploring bioactive and preservation effects of chamomile (Matricaria recutita L.)", *J. Funct. Foods*, vol. 16, pp. 114–124, 2015.

[25] A. El Mihyaoui, J. C. G. Esteves da Silva, S. Charfi, M. E. Candela Castillo, A. Lamarti, and M. B. Arnao, "Chamomile (Matricaria chamomilla L.): A Review of Ethnomedicinal Use, Phytochemistry and Pharmacological Uses," *Life*, vol. 12, no. 4, p. 479, Mar. 2022.

In addition to its use in food field as a seasoning and natural preservative, it is also used in medicine to treat muscle pain due to the presence of piperine, which has anti-inflammatory, antimicrobial, antioxidant and anti-cancer (inhibits the growth of colon cancer cells[26]) effects[27], reduces cardiovascular problems[28] and cholesterol uptake[29]. Furthermore, it has shown prebiotic potential, promoting the growth of *Lactobacillus* and *Bifidobacterium*, which prevent the proliferation of pathogenic species and contribute to the regulation of the gut microbiota. Many studies have contributed to identify numerous bioactive compounds contained in the species *Piper nigrum L.*, such as terpenes, flavonoids, flavanones and steroids[30]. The antioxidant properties of black pepper are attributed to the presence of polyphenols and flavonoids, which are able to neutralise reactive oxygen species and prevent oxidative damage to cells[31] [32]. Pepper has an aromatic composition that can also be influenced by treatment and preservation processes such as fruit dehydration, which involves the enzymatic oxidation of phenols, leading to the typical black colour[33]. The complex phytochemical matrix and biological activity of this spice can be confirmed, justifying its use in food, pharmaceuticals and cosmetics.

---

[26] R. Wu, J. Zhao, P. Wei, M. Tang, Z. Ma, Y. Zhao, L. Du, and L. Wan, "*Piper nigrum* Extract Inhibits the Growth of Human Colorectal Cancer HT-29 Cells by Inducing p53-Mediated Apoptosis," *Pharmaceuticals*, vol. 16, no. 9, p. 1325, Sep. 2023.

[27] A. S. POP, A. R. MARC, and C. C. MUREȘAN, "Therapeutic Insights into Black Pepper (*Piper nigrum*): Phytochemical Composition, Bioactive Properties, and Health Benefits," *Hop Med. Plants*, vol. 32, no. 1–2, pp. 139–152, 2024.

[28] D. Wang *et al.*, "Cardiovascular protective effect of black pepper (*Piper nigrum* L.) and its major bioactive constituent piperine," *Trends Food Sci. Technol.*, vol. 117, pp. 34–45, 2021.

[29] A. Duangjai, K. Ingkaninan, S. Praputbut, and N. Limpeanchob, "Black pepper and piperine reduce cholesterol uptake and enhance translocation of cholesterol transporter proteins," *J. Nat. Med.*, vol. 67, no. 2, pp. 303–310, 2013.

[30] V. S. Parmar *et al.*, "Phytochemistry of the genus Piper," *Phytochemistry*, vol. 46, no. 4, pp. 597–673, 1997.

[31] J.-G. Lee, Y. Chae, Y. Shin, and Y.-J. Kim, "Chemical composition and antioxidant capacity of black pepper pericarp," *Applied Biological Chemistry*, vol. 63, no. 35, pp. 1–9, 2020.

[32] M. E. Embuscado, "Spices and herbs: Natural sources of antioxidants – a mini review," *Journal of Functional Foods*, vol. 18, pp. 811–819, Apr. 2015.

[33] C. K. Mangalakkumari, V. P. Sreedharan, and A. G. Mathew, "Studies on Blackening of Pepper (*Piper nigrum*, Linn) During Dehydration," *Journal of Food Science*, vol. 48, no. 2, pp. 604–606, 1983.

3. Evolution of analytical investigation

Over time, the study of bioactive molecules in plants has become increasingly important in both academic and industrial circles, given their therapeutic, nutraceutical and preventive potential as an alternative to synthetic drugs[34].

Black pepper, one of the most widely used spices in the world, has seen an increase in consumption and market value, to the extent that studies using spectroscopic methods[35] have focused on fraud prevention to avoid the use of cheaper similar spices[36]. Since black pepper is often marketed in ground powder form, it is particularly susceptible to adulteration, i.e. the intentional addition or substitution of cheaper or foreign materials, such as papaya seeds, in order to reduce production costs and increase profits[37]. In some cases, it is also fraudulently labeled as a Protected Geographical Indication (PGI) product to obtain illegal advantages. Such adulteration can be effectively detected by combining Fourier transform infrared diffuse reflectance spectroscopy (DRIFTS) with chemometrics[38]. Fourier transform infrared spectroscopy (FT-IR) and Near-Infrared Spectroscopy (NIR) analysis combined with chemometric approaches have proven effective in highlighting the substitution of black pepper with papaya seeds which are commonly used as adulterants due to their structural similarity to peppercorns, low cost, and easy availability on the market. The volatile fractions of black pepper (*Piper nigrum* L.) were obtained both by supercritical fluid extraction (SFE)[39] and by headspace solid-phase

---

[34] A. Kumar, F. Ahmad, and S. Zaidi, "Importance of Bioactive Compounds Present in Plant Products and Their Extraction – A Review," *Agric. Rev.*, vol. 40, no. 04, 2019.

[35] S. Mayr, K. B. Beć, J. Grabska, E. Schneckenreiter, and C. W. Huck, "Near-infrared spectroscopy in quality control of *Piper nigrum*: A comparison of performance of benchtop and handheld spectrometers," *Talanta*, vol. 223, p. 121809, 2021.

[36] A. S. Wilde, S. A. Haughey, P. Galvin-King, and C. T. Elliott, "The feasibility of applying NIR and FT-IR fingerprinting to detect adulteration in black pepper," *Food Control*, vol. 100, pp. 1–7, 2019.

[37] K. Dhanya, S. Syamkumar, and B. Sasikumar, "Development and application of SCAR marker for the detection of papaya seed adulteration in traded black pepper powder," Food Biotechnology, vol. 23, no. 2, pp. 97–106, 2009.

[38] L. Hu, C. Yin, S. Ma, and Z. Liu, "Assessing the authenticity of black pepper using diffuse reflectance mid-infrared Fourier transform spectroscopy coupled with chemometrics," *Computers and Electronics in Agriculture*, vol. 154, pp. 491–500, 2018.

[39] A. Bensebia and A. Barth, "Supercritical CO_2 extraction of black pepper oil," *Food Research International*, vol. 35, no. 5, pp. 495–500, 2004.

microextraction (HS-SPME) [40], and subsequently analysed by gas chromatography-mass spectrometry (GC-MS). In addition, spectroscopic techniques combined with chemometric approaches were used for authentication and discrimination purposes[41]. They revealed the presence of monoterpenes and sesquiterpenes responsible for the characteristic aroma, antimicrobial and antioxidant properties[42]. The non-volatile fractions were analysed by high-performance liquid chromatography (HPLC) [43] to quantify piperidine alkaloids (piperine) [44], phenols, and flavonoids. Piperine was also isolated and analysed using various analytical techniques such as Ultraviolet-Visible spectroscopy and FT-IR[45]. In addition to piperine, alkaloids and amide derivatives that contribute to biological functions have been identified through analysis of ethanol and methanol extracts conducted using high-performance liquid chromatography with Diode-Array Detector (HPLC-DAD)[46]. Recently, studies based on FT-IR[47], UV-Vis[48], and fluorescence[49] spectroscopies combined with chemometric algorithms have been carried out to develop analytical approaches capable of classifying black pepper samples from different geographical origins and assessing their

---

[40] P. Yu *et al.*, "Effect of peeling and packaging methods on volatile compounds in pepper (*Piper nigrum* L.) during storage," *Int. J. Food Sci. Technol.*, vol. 59, no. 9, pp. 6533–6542, 2024.

[41] M. Wilde, L. Schumann, C. Haughey, and C. Elliott, "Authentication of black pepper (*Piper nigrum* L.) using spectroscopic techniques coupled with chemometric analysis," *Food Research International*, vol. 178, p. 113540, 2024.

[42] D. N. Do, "Hydrodistillation of essential oil from the whole black pepper and light berries black pepper (*Piper nigrum* L.) harvesting in Dak Nong, Vietnam on a pilot scale: The study on extraction process and analyses of essential components," *AIP Conference Proceedings*, vol. 2610, no. 060003, Aug. 2022.

[43] P. S. Variyar and C. Bandyopadhyay, "Estimation of phenolic compounds in green pepper berries (*Piper nigrum* L.) by high-performance liquid chromatography," *Chromatographia*, vol. 39, no. 11, pp. 743–746, 1994.

[44] V. Parab Gaonkar, V. K. Mannur, and K. Hullatti, "Quality assessment and Analytical Quality by Design-based RP-HPLC method development for quantification of Piperine in *Piper nigrum* L.," *Futur. J. Pharm. Sci.*, vol. 8, no. 1, p. 16, 2022.

[45] S. R. Pawar *et al.*, "Analytical Development of Piperine as Herbal Reference Material from Black Pepper for Quality Control Purposes: Liquid Chromatography-Quadrupole Time-of-Flight, Nuclear Magnetic Resonance, and Thermal Analysis," *J. Chromatogr. Sci.*, vol. 63, no. 6, p. bmaf036, 2025.

[46] J. Lee, Y. Chae, Y. Shin, and Y.-J. Kim, "Chemical composition and antioxidant capacity of black pepper pericarp and seed extracts," *Applied Biological Chemistry*, vol. 63, no. 1, pp. 79–90, 2020.

[47] L. Hu, C. Yin, S. Ma, and Z. Liu, "Assessing the authenticity of black pepper using diffuse reflectance mid-infrared Fourier transform spectroscopy coupled with chemometrics," *Computers and Electronics in Agriculture*, vol. 154, pp. 491–500, 2018.

[48] J.-G. Lee and Y.-J. Kim, "Comparison study of validation parameters and measurement uncertainty of rapid analytical methods for piperine in black pepper by ultraviolet spectroscopy and high-performance liquid chromatography," *Food Science and Biotechnology*, vol. 31, no. 8, pp. 1133–1143, 2022.

[49] Z.-X. Liu, S.-R. Xiong, S.-H. Tang, Y. Wang, and J. Tan, "A practical application of front-face synchronous fluorescence spectroscopy to rapid, simultaneous and non-destructive determination of piperine and multiple adulterants in ground black and white pepper (*Piper nigrum* L.)," *Food Research International*, vol. 167, p. 112654, 2023.

quality and authenticity[50]. These methods make it possible to discriminate adulterated or geographically distinct samples and to correlate spectral properties with chemical composition and biological activity.

During the 1990s, interest grew in the glucosinolates contained in Brassicaceae. This led to a shift from a traditional analysis technique based on thin-layer chromatography (TLC)[51] to an extraction and separation technique using HPLC-UV analysis, which ensured international comparability on a regulatory basis (ISO 9167:2019). The use of HPLC-DAD for the quantification of the main polyphenols showed that the polarity of the solvent significantly influences the yield and phenolic composition. Methanol provided the best extraction yield and the highest antioxidant activity, followed by ethanol and solvent mixtures[52]. Evolution led to the use of NIR to reduce analysis times and costs and enable rapid and economical detection of product quality[53] and variety discrimination[54] based on the presence of predominant phenolic acids (caffeic acid, ferulic acid and sinapic acid) and flavonoids (kaempferol and quercetin)[55]. It proved to be effective for the optical characterisation of leaves and the assessment of chlorophyll, carotenoid and polyphenol content[56] [57]. This has led to the combination of analysis techniques with chemometrics for a more accurate definition, thus

---

[50] M. H. Baky, I. M. Kamal, L. A. Wessjohann, and M. A. Farag, "Assessment of metabolome diversity in black and white pepper in response to autoclaving using MS- and NMR-based metabolomics and in relation to its remote and direct antimicrobial effects against food-borne pathogens," RSC Advances, vol. 14, pp. 10799–10813, 2024.

[51] H. M. Dietz and R. V Harris, "Novel and rapid methods of glucosinolate analysis with particular reference to their application to 00-rapeseed," *Food Control*, vol. 1, no. 2, pp. 84–97, 1990.

[52] S. Muzammil, Y. Wang, M. H. Siddique, E. Zubair, S. Hayat, M. Zubair, A. Roy, R. Mumtaz, M. Azeem, T. B. Emran, and M. Q. Shahid, "Polyphenolic composition, antioxidant, antiproliferative and antidiabetic activities of *Coronopus didymus* leaf extracts," *Molecules*, vol. 27, no. 19, p. 6263, Sep. 2022.

[53] R. Font, M. Del Río-Celestino, E. Rosa, A. Aires, and A. De Haro-Bailón, "Glucosinolate assessment in Brassica oleracea leaves by near-infrared spectroscopy," *J. Agric. Sci.*, vol. 143, no. 1, pp. 65–73, 2005.

[54] S.-I. Sohn *et al.*, "Discrimination of Transgenic Canola (Brassica napus L.) and their Hybrids with B. rapa using Vis-NIR Spectroscopy and Machine Learning Methods," *Int. J. Mol. Sci.*, vol. 23, no. 1, p. 220, Dec. 2021.

[55] M. M. Rahman, A. T. M. Abdullah, M. Sharif, S. Jahan, M. A. Kabir, M. Motalab, and T. A. Khan, "Relative evaluation of in-vitro antioxidant potential and phenolic constituents by HPLC-DAD of *Brassica* vegetables extracted in different solvents," *Heliyon*, vol. 8, no. 10, p. e10838, 2022.

[56] C. Pérez-Roncal, M. Álvarez-González, J. L. García-López, and J. A. Fernández, "Visible and near-infrared spectroscopy for non-destructive estimation of chlorophyll and polyphenol content in *Brassica* leaves," *Plant Physiology and Biochemistry*, vol. 190, pp. 38–49, 2022.

[57] D. Ceccarelli, M. R. Tofanelli, G. Lorenzini, and L. Lazzeri, "Laser-induced fluorescence spectroscopy for rapid assessment of antioxidant status in *Brassica oleracea* leaves," *Sensors and Actuators B: Chemical*, vol. 384, p. 133533, 2023.

enabling monitoring in the food and agricultural industries[58], focusing on the highly accurate prediction of total phenolic content, glucosinolate levels and antioxidant activity in leaves[59] [60], providing an overview of the metabolites present and the oxidative stress linked to environmental conditions and post-harvest processes[61]. Extraction processes were optimised using HS-SPME and GC-MS[62]. To reduce sample pretreatment and improve analysis accuracy, direct analysis using LC-ESI/MS/MS was proposed[63]. This provided an opportunity to conduct a comparative study between edible and inedible parts of vegetables. The effects of cultivation techniques such as the use of fertilisers and pesticides on secondary metabolites in Brassicaceae samples were also evaluated[64].

The efficacy and safety of chamomile extracts depend strictly on the raw material, solvent, process parameters and standardisation of the extract[65]. The antioxidant activity was studied using the 2,2-diphenyl-1-picrylhydrazyl (DPPH) assay, the toxicity and total polyphenolic content in aqueous extracts of chamomile, and the toxicity test showed that the lower the extraction temperature, the lower the toxicity of the extract[66].

---

[58] A. Ali Redha *et al.*, "Determination of glucosinolates in broccoli (Brassica oleracea var. italica) by combining mid-infrared (MIR) spectroscopy with chemometrics," *Int. J. Food Sci. Technol.*, vol. 58, no. 11, pp. 5679–5688, 2023.

[59] R. Pandiselvam, S. Subhashini, C. Banuu Priya, J. Kothakota, K. K. Manikantan, and R. Ramesh, "Application of hyperspectral imaging and chemometrics for the non-destructive determination of phenolic content in leafy vegetables," *Food Chemistry*, vol. 356, p. 129628, 2021.

[60] A. H. Al-Saeedi, M. A. Al-Marhobi, A. S. Al-Bulushi, and A. S. Al-Harrasi, "Prediction of antioxidant activity and phenolic compounds in *Brassica* leaves using Vis–NIR spectroscopy and PLS regression," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 276, p. 121211, 2022.

[61] H. Wu, Y. Zhang, X. Li, J. Chen, and L. Wang, "Hyperspectral imaging and machine learning for mapping glucosinolate distribution in *Brassica rapa* leaves," *Plant Methods*, vol. 20, no. 1, p. 15, 2024.

[62] M. H. Baky, S. N. Shamma, J. Xiao, and M. A. Farag, "Comparative aroma and nutrients profiling in six edible versus nonedible cruciferous vegetables using MS based metabolomics," *Food Chem.*, vol. 383, p. 132374, Jul. 2022.

[63] Q. Tian, R. A. Rosselot, and S. J. Schwartz, "Quantitative determination of intact glucosinolates in broccoli, broccoli sprouts, Brussels sprouts, and cauliflower by high-performance liquid chromatography–electrospray ionization–tandem mass spectrometry," *Anal. Biochem.*, vol. 343, no. 1, pp. 93–99, 2005.

[64] M. Lucarini *et al.*, "NMR-Based Metabolomic Comparison of Brassica oleracea (Var. italica): Organic and Conventional Farming," *Foods*, vol. 9, no. 7, p. 945, Jul. 2020.

[65] A. El Mihyaoui, J. C. G. Esteves da Silva, S. Charfi, M. E. Candela Castillo, A. Lamarti, and M. B. Arnao, "Chamomile (*Matricaria chamomilla* L.): A review of ethnomedicinal use, phytochemistry and pharmacological uses," *Life*, vol. 12, no. 4, p. 479, 2022.

[66] N. S. Sotiropoulou, S. F. Megremi, and P. Tarantilis, "Evaluation of Antioxidant Activity, Toxicity, and Phenolic Profile of Aqueous Extracts of Chamomile (Matricaria chamomilla L.) and Sage (Salvia officinalis L.) Prepared at Different Temperatures," *Appl. Sci.*, vol. 10, no. 7, p. 2270, Mar. 2020.

Subcritical water extraction methods followed by Ultra-High-Performance Liquid Chromatography coupled with Heated Electrospray Ionization Tandem Mass Spectrometry (UHPLC-HESI-MS/MS) analysis were also used, which once again highlighted the benefits and limitations in terms of bioactivity, which decreases as degradation increases[67].

The growing interest in the phytochemical characterisation of chamomile extracts has led to the development of analytical methods aimed at quantifying, identifying and evaluating the stability of metabolites of pharmacological and cosmetic interest. Analytical techniques such as HPLC and UV-Vis spectrophotometry are useful tools for determining the content of polyphenols and flavonoids. UHPLC-HESI-MS/MS was used to define the phytochemical profile and understand the variations due to solvents, temperature and extraction method. This made it possible to identify compounds such as apigenin, luteolin, quercetin, caffeic acid and ferulic acid, monitoring them in terms of thermal or oxidative degradation. It has been shown that, in water extraction processes, an increase in temperature leads to a progressive reduction in antioxidant activity due to the degradation of the most unstable phenolic compounds. Some studies based on HPLC-DAD and Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS) have shown phytochemical heterogeneity in ethanol extracts of commercial chamomile products, also revealing a direct correlation between phenolic concentration and cytotoxic potential, thus demonstrating that highly concentrated extracts or obtained with organic solvents may be biologically more potent but less safe for human health[68]. LC-MS was used to determine the phenolic composition of aqueous chamomile infusions at 80 °C, which proved to be the optimal temperature to maximise bioactivity. The main phenolic compounds identified following this analysis were rutin trihydrate, ferulic acid, chlorogenic acid and apigenin-7-O-glucoside[69].

---

[67] A. Cvetanović, J. Švarc-Gajić, Z. Zeković, J. Jerković, G. Zengin, U. Gašić, Ž. Tešić, P. Mašković, C. Soares, M. F. Barroso, C. Delerue-Matos, and S. Đurović, "The influence of the extraction temperature on polyphenolic profiles and bioactivity of chamomile (*Matricaria chamomilla* L.) subcritical water extracts," *Food Chemistry*, vol. 271, pp. 328–337, 2019.

[68] M. V. Catani, F. Rinaldi, V. Tullio, V. Gasperi, and I. Savini, "Comparative Analysis of Phenolic Composition of Six Commercially Available Chamomile (*Matricaria chamomilla* L.) Extracts: Potential Biological Implications," *International Journal of Molecular Sciences*, vol. 22, no. 19, Art. no. 10601, Sep. 30, 2021.

[69] N. S. Sotiropoulou, S. F. Megremi, and P. Tarantilis, "Evaluation of antioxidant activity, toxicity, and phenolic profile of aqueous extracts of chamomile (*Matricaria chamomilla* L.) and sage (*Salvia officinalis* L.) prepared at different temperatures," *Applied Sciences (Switzerland)*, vol. 10, 2020.

The increased interest in improving extraction methods is driven by growing industrial demand for greater safety and bioactivity in chamomile-based formulations, using environmentally friendly methods that minimise sample degradation[70]. The most effective extraction of essential oils is that which uses supercritical fluids (SFE) with carbon dioxide[71]. Microwave-assisted extraction techniques and ultrasonic extraction with pure water or environmentally friendly solvents were also used[72], although to a lesser extent. While these techniques reduce the use of volatile organic compounds, they do not reduce sample degradation and the risk of toxicity. It follows that extraction techniques significantly influence the bioactivity. Ethanolic extracts of commercial chamomile products can produce cytotoxic effects as high concentration, highlighting how potency and efficacy depend greatly on the method of preparation. Therefore, herbal medicines are not always risk-free[73]. On the contrary, aqueous extracts show antioxidant activity and a lower risk due to their lower essential oil content[74]. The presence of studies in the literature carried out using GC-MS has demonstrated that chamomile essential oils are rich in terpenoids such as α-bisabolol, chamazulene and bisabolol oxide A and B with soothing, anti-inflammatory and antibacterial properties[75]. A comparison between results obtained using GC-MS and GC-FID showed that extraction with supercritical fluids allows for richer aromatic profiles to be obtained[76]. In recent years, innovative and green techniques such as Raman spectroscopy and fluorescence have been applied for continuous online monitoring of extraction quality and for non-destructive evaluation of total phenolic

---

[70] G. Haghi, A. Hatami, A. Safaei, and M. Mehran, "Analysis of phenolic compounds in Matricaria chamomilla and its extracts by UPLC-UV," *Res. Pharm. Sci.*, vol. 9, no. 1, pp. 31–7, 2014.

[71] G. Zengin *et al.*, "A Comparative Study of Chamomile Essential Oils and Lipophilic Extracts Obtained by Conventional and Greener Extraction Techniques: Chemometric Approach to Chemical Composition and Biological Activity," *Separations*, vol. 10, no. 1, p. 18, 2022.

[72] J. Š. Žlabur *et al.*, "Effect of Different Green Extraction Methods and Solvents on Bioactive Components of Chamomile (Matricaria chamomilla L.) Flowers," *Molecules*, vol. 25, no. 4, 2020.

[73] M. V. Catani, F. Rinaldi, V. Tullio, V. Gasperi, and I. Savini, "Comparative Analysis of Phenolic Composition of Six Commercially Available Chamomile (Matricaria chamomilla L.) Extracts: Potential Biological Implications," *Int. J. Mol. Sci.*, vol. 22, no. 19, p. 10601, Sep. 2021.

[74] M. da G. Campos and K. R. Markham, "Structure information from HPLC and on-line measured absorption spectra: flavones, flavonols and phenolic acids" 2007.

[75] J. Lu, Z. Jiang, J. Dang, D. Li, D. Yu, C. Qu, and Q. Wu, "GC–MS combined with fast GC E-nose for the analysis of volatile components of chamomile (*Matricaria chamomilla* L.)," *Foods*, vol. 13, no. 18, p. 1865, 2024.

[76] J. Lu, Z. Jiang, J. Dang, D. Li, D. Yu, C. Qu, and Q. Wu, "GC–MS combined with fast GC E-nose for the analysis of volatile components of chamomile (*Matricaria chamomilla* L.)," *Foods*, vol. 13, no. 18, p. 1865, 2024.

content[77]. The combination of these techniques and chemometric approaches such as Principal Component Analysis (PCA) and Partial Least Squares-Discriminant Analysis (PLS-DA) makes it possible to distinguish between samples obtained with different solvents, predict their antioxidant activity, and classify the extracts according to their purity or botanical origin. The combination of the results obtained from the various studies has made it possible to construct a phytocomplex model of chamomile by correlating the chemical composition, extraction method and bioactivity, demonstrating that advanced analytical characterisation allows products to be standardised but also their quality to be controlled and their safety to be assessed[78].

4. Objectives of the research work

Based on the above considerations, this research project aims to develop non-destructive and low-cost analytical methods, coupled with efficient chemometric strategies, to characterise and identify pharmacologically active substances in plant species and their beneficial properties (such as antioxidant activity). In particular, plants belonging to the Brassicaceae, Asteraceae and Piperaceae, were characterized through the use of chromatographic techniques (in particular, HPLC-DAD), spectroscopic techniques (IR and UV-Vis), integrated with specific assays such as the DPPH test, and with non-destructive optical analysis techniques (E-Eye), aimed at evaluating bioactivity, combined with chemometric data analysis.

The study on Brassicaceae focused on the application and optimisation of advanced and non-destructive analytical methods for the chemical, morphological and spectroscopic characterisation of local varieties. In particular, a fingerprinting approach was applied to distinguish local varieties such as Mugnolo di Pettorano sul Gizio from other commercial cultivars. These methods were integrated with chromatographic techniques and spectrophotometric assays necessary for the extraction of bioactive compounds from the dried leaves of Brassicaceae species and the relative

---

[77] J. Zheng and L. He, "Surface-enhanced Raman spectroscopy for the chemical analysis of food," *Comprehensive Reviews in Food Science and Food Safety*, vol. 13, no. 3, pp. 317–329, 2014.

[78] J. K. Srivastava, E. Shankar, and S. Gupta, "Chamomile: A herbal medicine of the past with bright future," *Molecular Medicine Reports*, vol. 3, no. 6, pp. 895–901, 2010.

evaluation of the antioxidant activity of hydroalcoholic extracts, correlating it with spectrochromatographic data.

Following the study on Brassicaceae, the research was extended to another plant matrix of high pharmacological relevance, chamomile (*Matricaria chamomilla* L.), with the aim of optimising a green extraction process. Given that previous studies have shown that the quality of the extract increases as the extraction temperature decreases, and that organic solvents are the most commonly used approach for extraction[79], this study optimised the cold aqueous extraction of dried chamomile flower heads in order to guarantee a product that is healthy and safe for human health and the environment.

The high consumption of *Piper nigrum* L. (black pepper) and its remarkable beneficial properties prompted this study to investigate and characterise both the chemical and functional properties of selected peppercorn samples. The samples included black peppercorns from different geographical origins, namely Kampot (Cambodia, PGI, including one organic sample), Madagascar, and several commercially available products, as well as a green peppercorn sample for comparative purposes.

The study aimed to characterise the aroma profile and antioxidant activity of these samples and to evaluate whether their geographical origin could be predicted through non-destructive infrared spectroscopy. The aromatic profiles were determined using headspace solid-phase microextraction coupled with gas chromatography-mass spectrometry, while antioxidant activity was assessed through the DPPH radical scavenging assay. Finally, the data obtained from IR spectroscopy were processed using chemometric methods to enable the classification of the samples according to their geographical origin.

[79] M. da Silva Pinto, "Tea: A new perspective on health benefits," *Food Research International*, vol. 53, no. 2, pp. 558–567, 2013.

CHAPTER 2

ANALYTICAL TECHNIQUES

SUMMARY: 1. Chromatography: High Performance Liquid Chromatography with Diode-Array Detector and Gas Chromatography-Mass Spectrometry – 2. Spectroscopic approaches: UV-Vis and IR – 3. E-Eye techniques: digital microscopy – 4. DPPH radical scavenging assay

1. Chromatography: High Performance Liquid Chromatography with Diode-Array Detector and Gas Chromatography-Mass Spectrometry

Chromatography is a physical method of separation in which the components to be separated are distributed between two phases, one of which is stationary (stationary phase) while the other (the mobile phase) moves in a definite direction[80]. According to this definition, the mobile phase transports the samples to be analysed along the stationary phase and allows the components of a matrix to distribute themselves between the two phases and consequently separate according to their different affinities between the phases themselves, which causes a different elution speed depending on the compound. This produces a chromatogram in which each peak corresponds to the exit of a component from the system. In this way, the molecules with greater affinity for the stationary phase will be those that elute with a longer retention time than molecules that, having greater affinity for the mobile phase, move faster within the column.

This technique is used to separate, identify and purify compounds present within complex matrices or those rich in components, exploiting their ability to interact with the phases. It is also the most widely used method due to its versatility, high selectivity and efficiency, and is therefore used in the analysis

[80] https://goldbook.iupac.org/terms/view/C01075

of chamomile extracts and species belonging to the Brassicaceae family in order to separate and identify the compounds present within them. In these works, liquid chromatography was employed, a technique that can be performed in a column or on a planar surface and the modern inversion, which generally uses very small particles and high inlet pressure, is called high-performance liquid chromatography, or HPLC[81].

HPLC is a chromatographic technique in which the mobile phase advances along the column containing the stationary phase thanks to high pressure exerted by pumps (1800 psi). The sample to be analysed is placed inside an injection system and transported by the mobile phase inside the column, the core of the instrument, where separation takes place based on the interactions of the components with the two phases (mobile and stationary). Finally, a specific detector monitors the compounds eluting from the column, generating a signal that is acquired and processed by a computerised control system; this allows the operator to study the signal and make the relevant observations.

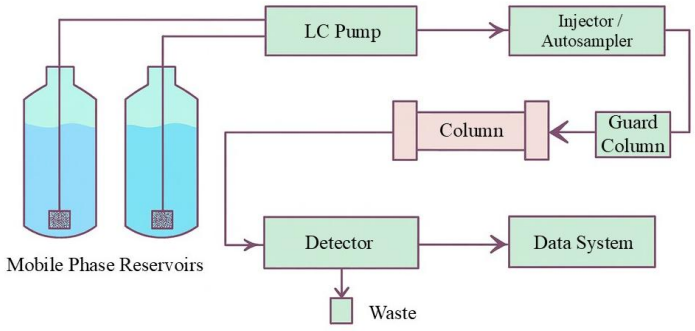


Figure 1. *Block diagram of HPLC instrumentation*

In this study, qualitative analyses were performed using an HPLC (Waters, Milford, MA, USA) equipped with a Kinetex C18 100 Å column (Phenomenex, Torrance, CA, USA), a Model 600 pump, a 600 pump controller module, and a 996 series photodiode array detector. The chromatography system was controlled by Empower Pro software supplied

[81] https://goldbook.iupac.org/terms/view/L03578

by Waters (Milford, MA, USA)[82]. The instrument is equipped with a binary pump system that pushes the two solutions constituting the mobile phase, which subsequently meet in a high-pressure mixing chamber. This process allows gradient operation, as the presence of two pumps ensures better efficiency in chromatographic runs where gradient elution of the mobile phase is required, especially in rapid gradients where the composition varies quickly. The mobile phase used was degassed using an Agilent 1200 degasser (Agilent Technologies, Waldbronn, Germany) to remove any dissolved gases, mainly carbon dioxide and oxygen, present in the solutions that make up the mobile phase before they enter the pumps and cause pressure variations, inconsistent flow and poor reproducibility, which could then generate ghost peaks in the chromatogram; whereas a gas-free mobile phase ensures more stable and repeatable separations.

Having to analyse bioactive compounds, including phenolic compounds and flavonoids, reversed-phase chromatography was used as the most suitable and widely used technique for analysing this type of compounds[83] [84] coupled with a DAD detector. The latter simultaneously records UV-Vis spectra over a wide wavelength range (210.3 - 496.9 nm), allowing both quantification and identification of compounds through their spectral profile. A C18 column was used, therefore a non-polar stationary phase and a polar mobile phase consisting of a gradient of acetonitrile (≥99.9%, gradient-grade for HPLC, Sigma-Aldrich) and water acidified with 0.1% phosphoric acid (pH≈2.25) was used to ensure that the polyphenols did not dissociate, guaranteeing their neutral form and preventing the formation of salts and secondary interactions with the residual silanol groups of the column, improving the shape of the peaks and reducing the tailing effect. One of the advantages of using Reversed-Phase High-Performance Liquid Chromatography (RP-HPLC) for the analysis of compounds of interest in this study is its wide applicability. In fact, it is suitable for the analysis of a large number of polar compounds such

---

[82] https://www.waters.com/waters/ en_US/Empower-Software-Solutions/nav.htm?cid=513188&locale=en_US

[83] Z. A. Temerdashev, N. A. Frolova, and I. A. Kolychev, "Determination of phenolic compounds in medicinal herbs by reversed-phase HPLC," *J. Anal. Chem.*, vol. 66, no. 4, pp. 407–414, Apr. 2011.

[84] A. Escarpa, M. D. Morales, and M. C. González, "Analytical performance of commercially available and unavailable phenolic compounds using real samples by high-performance liquid chromatography–diode-array detection," *Anal. Chim. Acta*, vol. 460, no. 1, pp. 61–72, Apr. 2002.

as secondary plant metabolites, drugs[85], environmental contaminants[86] like pesticides[87] or for monitoring water or air quality standards.

While HPLC is particularly suitable for the analysis of polar or thermolabile compounds, the study of volatile and semi-volatile organic molecules is more efficiently performed by gas chromatography (GC), especially when coupled with mass spectrometry (MS).

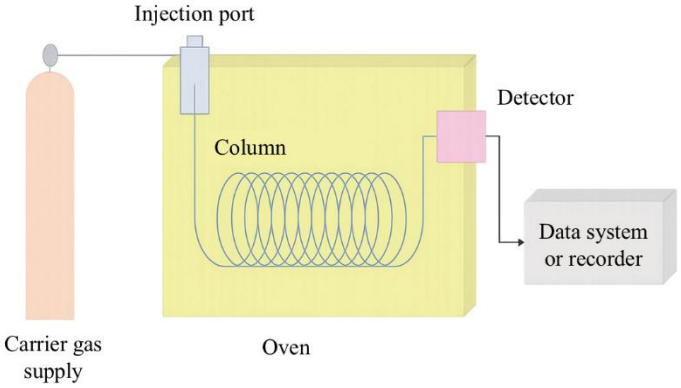


Figure 2.Schematic representation of a typical gas chromatograph

This technique combines the high separation capability of gas chromatography with the powerful identification ability of mass spectrometry.

In GC, the mobile phase consists of an inert gas, generally helium, which carries vaporized analytes through a capillary column internally coated with a thin film of the stationary phase. The separation of compounds depends on their volatility and interactions with the stationary phase: more volatile or less interactive compounds elute faster, whereas less volatile ones require longer retention times. The oven temperature, programmed in a gradual manner,

---

[85] D. Kumar, A. Kumar, V. Kumar, A. Raj, R. R. M. Rai, V. Baliyan, and N. Kumar, "A comprehensive review on analytical method development using RP-HPLC and recent advances in pharmaceutical applications," *J. Res. Appl. Sci. Biotechnol.*, vol. 2, no. 2, pp. 53–60, Apr. 2023.

[86] I. Baranowska and B. Kowalski, "Using HPLC method with DAD detection for the simultaneous determination of 15 drugs in surface water and wastewater," *Pol. J. Environ. Stud.*, vol. 20, no. 1, pp. 21–28, 2011.

[87] E. Watanabe, Y. Kobara, K. Baba, and H. Eun, "Aqueous acetonitrile extraction for pesticide residue analysis in agricultural products with HPLC–DAD," *Food Chem.*, vol. 154, pp. 7–12, 2014.

regulates the volatility of the compounds and enables optimal separation over a wide range of boiling points.

In the present study, analyses were performed using a GC-MS Trace 1300 ISQ LT system (Thermo Fisher Scientific, Waltham, MA, USA), equipped with a split/splitless injector and a J&W HP-5MS capillary column (5% phenyl-polymethylsiloxane, 30 m × 0.25 mm × 0.25 μm). Helium was used as the carrier gas. After chromatographic separation, the compounds eluting from the column were introduced into the mass spectrometer, where they were ionized (typically by electron impact), fragmented, and analysed according to their mass-to-charge ratio (m/z). Each analyte produced a characteristic mass spectrum, which represents a molecular "fingerprint". Comparison with spectral libraries, such as NIST 14 (NIST14. In Mass Spectral Database; NIST-National Institute of Standards and Technology: Gaithersburg, MD, USA, 2014), allowed unambiguous identification of the compounds present.

The coupling of GC with MS provides high sensitivity, selectivity, and identification capability, making this combination one of the most powerful techniques for qualitative and quantitative analysis of complex mixtures.

A crucial aspect for the quality of GC-MS analysis is sample preparation, which must pursue key objectives such as improving the chromatographic behavior and detectability of the analytes, separating them from interfering matrix components, and selectively enriching the compounds of interest.

Classical sample preparation methods include liquid-liquid extraction (LLE) for liquid samples and Soxhlet extraction for solid samples. Both methods are simple and widely used, but they generally require large amounts of sample and organic solvents. A more modern alternative to LLE is solid-phase extraction (SPE).

In recent decades, extraction, enrichment, and matrix separation methods have shown a clear evolution toward miniaturization, automated sample preparation, and higher throughput, thanks to reduced processing times, increased extraction yields, and improved reproducibility. At the same time, there has been a progressive reduction or complete elimination of organic solvents, in line with the principles of green chemistry.

In this context, the concept of solid-phase microextraction (SPME) was introduced in 1989. This approach combines solvent-free analytical

extraction using a coated silica fiber with the subsequent thermal desorption of the extracted analytes[88].

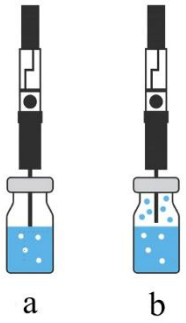


Figure 3. Two SPME extraction modes: a) Direct Immersion Solid-Phase Microextraction (DI-SPME); b) Headspace Solid-Phase Microextraction (HS-SPME).

SPME fibers can be used in two operational modes: direct immersion (DI-SPME), in which the fiber is immersed directly in the liquid sample (*Figure 3a*), and headspace extraction (HS-SPME), in which the fiber is placed above the sample in the gaseous phase (*Figure 3b*). After extraction, desorption occurs thermally in the GC injector. Since the fiber is fragile, it is housed in a syringe-like holder that protects it during penetration through the vial septum and the GC injector septum.

In the present study, sample preparation was performed using headspace solid-phase microextraction (HS-SPME), a technique that offers numerous advantages: it does not require organic solvents, reduces preparation time and the risk of contamination, improves reproducibility, and allows selective enrichment of volatile compounds. Moreover, it prevents the loss or transformation of thermolabile analytes and preserves the natural composition of the sample aroma, making it particularly suitable for studying the volatile profiles of foods[89] and spices[90].

---

[88] R. P. Belardi and J. Pawliszyn, "The application of chemically modified fused silica fibers in the extraction of organics from water matrix samples and their rapid transfer to capillary columns," *Water Pollution Research Journal of Canada*, vol. 24, pp. 179–191, 1989.

[89] S. J. Lehotay and J. Hajšlová, "Application of gas chromatography in food analysis," TrAC – Trends in Analytical Chemistry, vol. 21, nos. 9-10, pp. 686–697, 2002.

[90] Y. Sánchez-Cabrera and J. A. Pino, "Headspace solid-phase microextraction analysis of volatile compounds from spice essential oils in dry flavourings," International Journal of Food Science and Technology, vol. 46, no. 10, pp. 2118–2123, Oct. 2011.

Thanks to the combination of HS-SPME and GC-MS, it is possible to identify with high accuracy the main aromatic constituents of the analysed samples. The analytical power of this technique lies in its ability to couple the high-resolution separation capability of gas chromatography with the molecular identification capacity provided by mass spectrometry, allowing a comprehensive and detailed study of the chemical and aromatic characteristics of the analysed samples.

2. Spectroscopic approaches: UV-Vis and IR

Spectroscopic methods are based on the interaction between matter and electromagnetic radiation, which is absorbed or emitted in certain wavelength ranges, creating a characteristic spectrum that represents a fingerprint of the molecule. This is possible thanks to the presence of chromophores (functional groups, chemical bonding or unsaturations alternating with simple bonds) within the substances, which are able to interact with electromagnetic radiation. These techniques such as IR spectroscopy and UV-Vis spectrophotometry allow, through the study of the electromagnetic spectrum, to determine the composition, quantify and trace the chemical structure of the compound.

Molecules interact with electromagnetic radiation, and this interaction can be taken into account either in terms of absorption (moving from a lower energy state to a higher energy state) or emission (from higher energy states to lower energy states), obtaining structural information about matter.

Depending on the spectral region used, different instruments are required, leading to different analysis techniques. In fact, UV-Visible spectroscopy uses electromagnetic radiation which, when interacting with matter, stimulates transitions between the electronic valence states of atoms and molecules, this is the spectral range from 400 nm to 700 nm. Radiation employed in infrared spectroscopy, on the other hand, interacts with matter by stimulating its molecular vibrations, therefore it provides information on the types of chemical bonds present and is useful for molecular identification, and includes the zone from near infrared to far infrared (4000-400 cm^{-1}).

UV-Vis analysis was performed using a single-beam ONDA UV-30 SCAN UV spectrophotometer (Onda, Carpi, MO, Italy) equipped with two light

sources, a deuterium lamp for the UV region and a halogen-tungsten lamp for the visible region. The radiation emitted by the source is conveyed through a mirror to a monochromator (equipped with a diffraction grating), which separates the beam into the different wavelengths that compose it and selects a specific wavelength of interest that will pass through the sample contained in a quartz cuvette. Part of the radiation will be absorbed, while the transmitted radiation will reach the silicon photodiode detector, which will convert the light intensity into an electrical signal. This signal will be visualised using the software M.Wave Professional 2.0 (*Figure 4*).

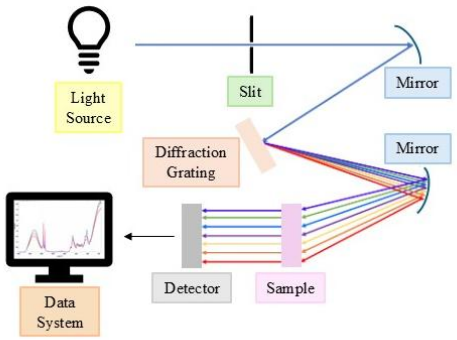


Figure 4. A simplified scheme of UV-Vis spectrophotometer.

By comparing the intensity of the incident radiation with that of the transmitted radiation, it is possible to determine parameters such as absorbance, transmittance or reflectance. Being single-beam, it does not allow the white spectrum to be recorded simultaneously. Therefore, before acquiring the measurement of a sample, it is necessary to measure the baseline which, for liquid samples (as in the case studies in this work), the background measurement is useful for taking into account the absorbance due to the cuvette and the solvent used. In this sense, if there is a variation in light intensity or in the optical performance of the system between the baseline and sample readings, the measurement may be less accurate. This inaccuracy could be due to the instability of the white, which may undergo changes over time, or to sample measurements that require a long time to prepare between samples and, consequently, between measurements. This means that during an analysis campaign, it is necessary to perform frequent and regular blank

measurements. The advantage lies in an instrument with fewer optical components, consisting of a single light beam, which results in smaller dimensions and lower costs.

IR spectroscopy allows working in transmittance and attenuated reflectance (ATR), enabling the analysis of solid, liquid and powder samples. The measurement is acquired in the form of an interferogram where absorption passes from the frequency domain to the time domain. This is processed mathematically using the Fourier transform, which allows the signal to be decomposed into its components and the infrared spectrum to be reconstructed as a function of the wave number.

The IR spectra were acquired using an attenuated total reflectance Fourier-transform infrared spectroscopy (ATR FT-IR) PerkinElmer Spectrum Two™ (PerkinElmer, Waltham, MA, USA). The spectrometer is equipped with a deuterated triglycine sulfate (DTGS) detector and a PerkinElmer Universal Attenuated Total Reflectance (uATR) accessory fitted with a single-bounce diamond crystal (*Figure 5*).

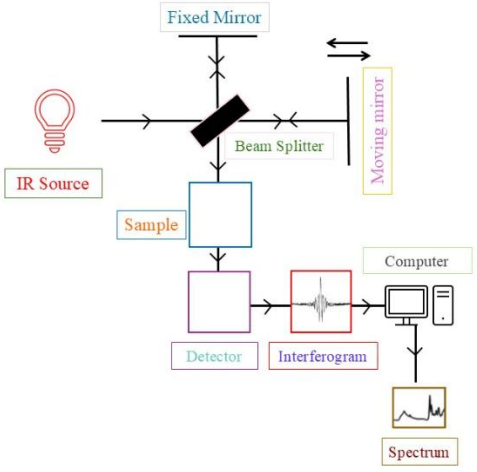


Figure 5. A scheme of the IR spectrometer

Fourier transform spectroscopy is an analytical technique that measures the absorption of infrared electromagnetic radiation by a sample, allowing it to be characterised by identifying vibration modes associated with the

presence of specific functional groups or specific molecular geometries in the compound.

After each analysis, the crystal and the plate in which it is inserted, as well as the sample compression tip, were cleaned using non-abrasive absorbent paper soaked in methanol. The background was recorded by exposing the diamond to air after cleaning it.

3. E-eye techniques: digital microscopy

Imaging techniques are a fundamental tool for studying morphological and structural characteristics in the food industry. They allow information to be acquired on the surfaces, colour and shape of samples, highlighting morphological differences. Compared to other analytical techniques, they are non-destructive and easy to use, but they do not provide chemical information because they are limited to describing the sample.

The E-Eye technique using the RSPro Wi-Fi USB Microscope (RS Components S.r.l., Milan, Italy) used in this research is advantageous because it is fast, non-invasive, low-cost, and useful for morphological analysis and visual documentation. It is limited to surface characteristics and therefore does not provide direct chemical-structural information, and its resolution is lower than that of more advanced microscopes. This device has a resolution of 1280×1024 pixels and a variable magnification of 10× to 160× with integrated LED lighting. Images can be collected while maintaining a constant distance between the optics and the sample, thus ensuring the reproducibility of observations.

4. DPPH radical scavenging assay

DPPH is a stable free radical and, thanks to this property, is commonly used as a reagent in the analysis of antioxidant activity.

The DPPH assay is a colorimetric assay based on the antioxidant's ability to reduce the free radical (DPPH•) by decreasing its concentration.

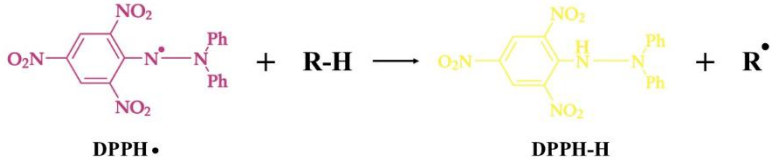


Figure 6. Reaction at the base of the DPPH radical assay

This decrease causes a colour change in the reaction solution, which turns from purple to pale yellow, depending on the concentration of antioxidant present[91].

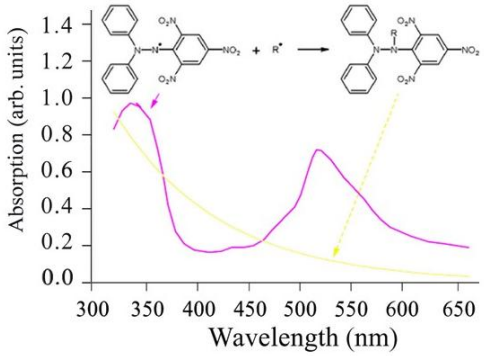


Figure 7. Variation in DPPH absorption with corresponding colour change of the solution

The change in absorbance, related to the reduction of DPPH, is evaluated by spectrophotometric analysis, with readings at maximum absorption namely at 516 nm[92]. The acquisition is performed over a range from 500 to 550 nm, with a wavelength interval of 2 nm.

To perform the assay, a stock solution of DPPH at 0.5 mg·mL^{-1} in methanol (≥99.0%, HPLC grade, Carlo Erba Reagenti, Milan, Italy) was prepared. The solution was left to stabilise at 4°C for 48 hours before use. From this solution,

---

[91] W. Brand-Williams, M. E. Cuvelier, and C. Berset, "Use of a free radical method to evaluate antioxidant activity," *LWT - Food Sci. Technol.*, vol. 28, no. 1, pp. 25–30, 1995.

[92] F. Poureini, M. Mohammadi, G. D. Najafpour, and M. Nikzad, "Comparative study on the extraction of apigenin from parsley leaves (*Petroselinum crispum* L.) by ultrasonic and microwave methods," *Chem. Pap.*, vol. 74, pp. 3857–3871, 2020.

a working solution at 0.025 $mg \cdot mL^{-1}$ was then prepared and used to perform the assay. The tests were performed on extracts of the matrices of interest; therefore, four dilutions were prepared for each extract. For each measurement, two separate solutions were prepared and analysed in parallel: the first was used to evaluate the absorbance in the absence of the samples under examination by taking 4.35 mL of DPPH working solution and adding methanol up to a volume of 5 mL (t_0); the second, a solution in which the reaction takes place, was obtained by adding 150 µL of the extract or its dilutions to 4.35 mL of DPPH working solution and brought to volume with methanol in a 5 mL flask (t_1). Both solutions (t_0 and t_1) were left to react for one hour in the dark before being analysed by spectrophotometer: solution t_0 to verify the stability of the DPPH reagent in the absence of the sample and to take into account any alterations in the reagent itself; the other (t_1) for the time necessary for the reaction.

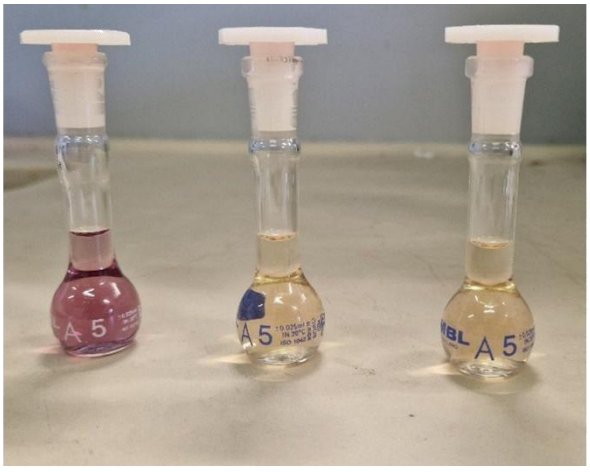


Figure 8. Photos of the solutions prepared for the antioxidant activity evaluation: on the left, the purple t_0 solution, used to verify the stability of reagent, and on the right, the two yellow t_1 solutions and their replicates, in which the reaction took place. The image was taken after one hour of reaction.

The percentage decrease in absorbance was calculated according to *Equation (0)*

$$\%A_{DPPH} = \frac{t_0 - t_1}{t_0} 100 \qquad (0)$$

where t_0 is the absorbance value of the working solution of DPPH at time zero (without the addition of the extract solution) while t_1 is the absorbance value recorded after one hour of reaction of the DPPH solution in the presence of the extract.

All measurements were carried out in duplicate.

For each extract, four dilutions plus the undiluted sample were considered and the IC_{50} value, the concentration of extract necessary to obtain a 50% reduction in absorbance, and the relative standard deviations were estimated from the dose-response curves obtained. The IC_{50} value indicates the concentration necessary to obtain a 50% reduction in DPPH free radicals.

Since a lower IC_{50} value indicates greater antioxidant activity (as a lower concentration is required to achieve 50% inhibition of oxidative capacity), a higher IC_{50} value indicates lower inhibitory efficacy[93].

The analysis and comparison of the IC_{50} values obtained highlighted differences between the various matrices or varieties in terms of antioxidant activity and, in some cases, correlated them with spectrochromatographic data.

[93] S. B. Kedare and R. P. Singh, "Genesis and development of DPPH method of antioxidant assay," *J. Food Sci. Technol.*, vol. 48, no. 4, pp. 412–422.

CHAPTER 3

CHEMOMETRIC APPROACHES

1.The value of chemometrics in analytical chemistry

The usefulness and necessity of chemometrics came to light in the early 1900s, when William Sealy Gosset, chief brewer at the Guinness brewery, needed to assess the quality of numerous samples as quickly as possible. The laboratory procedures of the time were lengthy and the instruments available were inefficient. This prompted William Sealy Gosset to take an interest in statistics, proposing the distribution known as Student's *t*-distribution. Later, in 1974, chemist Svante Wold coined the term chemometrics and gave it its first definition, which we now refer to as *the art of extracting chemically relevant information from data obtained in chemical experiments*.

The development of chemometrics has been aided by advances in increasingly sophisticated analytical instruments capable of obtaining more information simultaneously from the same sample (just think of a GC-MS and Inductively Coupled Plasma Mass Spectrometry (ICP-MS) analyses), in parallel with the evolution of specialized computational software.

Nowadays, in the field of analysis, the application of this discipline is essential, especially in campaigns involving the analysis of numerous

samples[94]. It is applicable in numerous areas, including air monitoring and production process monitoring in industry.

In this study, it was applied purely in the food sector to meet the need to create mathematical-statistical models capable of extracting relationships between the compositional profile and the variety of a species or the place of cultivation, which may highlight the differences due to different cultivation techniques.

Products are often analysed to assess their complexity and study their composition. This leads to the production of voluminous and complex data that necessarily involves the use of chemometrics to extract the information of interest from the raw data and also to understand which data is useful and which is less so in order to improve the interpretability of the results.

Understanding clearly and without confusion what chemometrics is currently capable of, with the growing development of machine learning, artificial intelligence and related methods, can be complicated. Chemometrics is a discipline that does not differ from the latter but uses mathematical and statistical tools to design optimal experimental procedures in order to extract the maximum relevant chemical information. Therefore, machine learning and data mining are not disciplines distinct from chemometrics, as they are differentiated by the nature of the data, their scope of application, their origin, and the fact that such data requires understanding in order to be interpreted correctly[95].

In particular, exploratory, regression and classification methods were applied to achieve the objectives of the various projects.

The data matrices that are obtained are complex and rich in information, requiring prior knowledge in order to best extract the relevant information. It is often necessary to preprocess the raw data in order to improve the quality and reliability of the models. Consequently, the nature of the data will also determine the necessary preprocessing.

---

[94] L. Eriksson, E. Johansson, T. Byrne, E. J. Johansson, R. Trygg, and C. Vikström, "Chemometrics in analytical chemistry—part II: modeling, validation, and applications," *Anal. Bioanal. Chem.*, vol. 410, no. 26, pp. 6691–6704.

[95] J. M. Amigo, "Data Mining, Machine Learning, Deep Learning, Chemometrics—Definitions, Common Points and Trends (Spoiler Alert: VALIDATE your models!)," *Braz. J. Anal. Chem.*, vol. 8, no. 32, pp. 45–61, 2021.

2. Data preprocessing

Preprocessing means performing preliminary data processing prior to the specific analysis (exploratory, regression, classification) of interest.

Preprocessing greatly influences the outcome of data analysis. Pretreatment procedures are applied to bring out information that may be "covered up" by instrumental effects or that may be caused by the presence of variables with very different scales. Data preprocessing may be required prior to any type of chemometric analysis.

Choosing the type of pretreatment is important because it must be suitable for the type of data: just consider a data matrix obtained following chromatographic analysis and spectral data matrices, which could both have the same dimensions but different nature of data, so the preprocessing to be applied will necessarily be different. In chromatography, it may be necessary to correct the baseline, while in spectroscopy, it may be necessary to reduce spectral noise, so it is necessary to apply data pretreatment procedures appropriate to each situation. Data preprocessing helps to remove the contribution of undesirable variability from the model, which, not originating from the chemical information to be modelled, could compromise the robustness and reduce the predictive capability of the model itself.

Following chemical analysis of the available samples, different pretreatments were applied to the data matrices analysed in this work, as necessary.

Using the *mean centring* as a data pretreatment allows you to remove the offset. It consists of bringing the centre of the data back to the origin of the axes, which very often does not coincide, and this would mean that the direction of maximum variability of the data is the first direction that coincides with the one that points from the origin of the axes towards the centre of the data itself. In this way, it will no longer be only the first direction that has the greatest variability, but the directions on which the data are projected will become relevant directions.

To do this, it is necessary to subtract the mean value from each column of the data matrix $\boldsymbol{X}$, as show in *Equation 1*

$$\boldsymbol{X}^{mc} = \boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^{T} \qquad (1)$$

in which $\boldsymbol{X}^{mc}$ represent the matrix of the mean-centered data, $\boldsymbol{X}$ the original data matrix and $\bar{\boldsymbol{x}}^T$ the row vector containing the means of each variable.

In this way, the data will be centred on the mean value and the new coordinates of the origin of the axes will be shifted according to the mean of all the experimental variables.

The result of this pretreatment, as shown in *Figure 9*, is that the origin of the axes coincides with the centroid of the data; in this way, this point becomes the new origin of the axes (0,0).

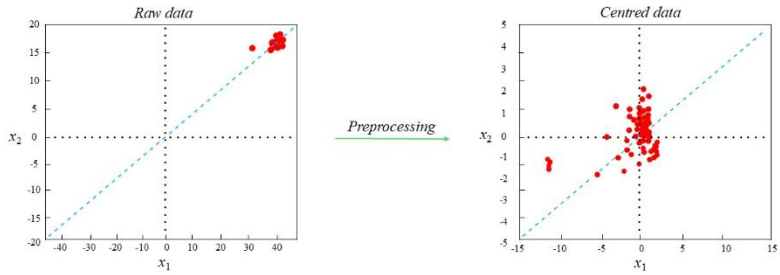


Figure 9. Graphical illustration of raw data and post-preprocessing data

Preprocessing that eliminates scale differences between variables has an important role in chemometric analysis. Without this operation, the contribution of each variable to the overall variance of the data could be altered: some variables, expressed in units of measurement with higher numerical values, would tend to dominate the modelling, while others would be underweighted regardless of their chemical relevance. To correct this effect and obtain a matrix without offset and with unit variance, *autoscaling* is applied, a procedure that first involves centring the data (subtracting the mean of each variable) and then normalising it by dividing by the corresponding standard deviation (*Equation 2*). In this way, all variables contribute equally to the construction of the multivariate model.

$$\boldsymbol{X}^{as} = \frac{\boldsymbol{X} - 1\bar{\boldsymbol{x}}^T}{\sigma_j} \qquad (2)$$

where $\boldsymbol{X}^{as}$ is the autoscaled data matrix, $\boldsymbol{X}$ is the original data matrix (raw data) and $\overline{\boldsymbol{x}}^{T}$ is the transposed vector of means for each variable.

In this way, in addition to bringing the centre of the data back to the origin of the axes, it ensures that the variance of each column is made equal to that of the others, specifically equal to 1 (*Figure 10*). Therefore, this normalisation means that each variable has a unit variance.

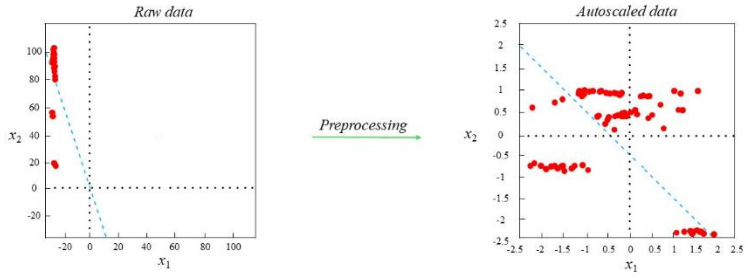


Figure 10. Graphical representation of autoscaling preprocessing

This eliminates spurious contributions to the total variance due to the different measurement scales used to express the results. In this way, all variables are given equal weight.

However, spectral data are also often affected by light scattering phenomena, which can introduce baseline shifts and multiplicative effects unrelated to chemical composition.

Part of the variability observed in the spectra does not depend on chemical composition, but on optical and physical phenomena related to the interaction of radiation with the sample matrix.

In particular, light scattering and differences in optical path length cause systematic variations due to parameters such as particle shape, size and distribution.

These purely physical effects can account for a significant portion of the total variability, manifesting themselves as baseline shifts (additive effects), changes in intensity or slope (multiplicative effects) and, in some cases, as non-linearity in the spectral response.

To minimise these contributions and improve the comparability between spectra, dispersion correction pre-processing techniques are used, including Standard Normal Variate (SNV) and derivative filters, which are used to

compensate for distortions introduced by scattering and the optical path and to highlight the chemical information of real interest[96].

Among the most effective correction methods, SNV [97] allows signals to be aligned and the effects of scattering to be reduced without requiring a common reference spectrum.

It is therefore a pretreatment carried out on each sample and then on each row of the matrix, as shown in *Equation 3*:

$$\boldsymbol{X}^{SNV} = \frac{\boldsymbol{X} - \bar{x}}{\sigma_i} \qquad (3)$$

In this equation $\boldsymbol{X}^{SNV}$ indicates the normalised spectrum using the SNV method, $\boldsymbol{X}$ the original data matrix, $\bar{x}$ the average of values in the sample row and σ_i the standard deviation of the spectrum of the i-th sample.

This type of preprocessing is applied to spectroscopic data, particularly when acquired in reflectance mode, as it reduces light scattering (the impact of light diffusion due to the physical nature of the sample, such as particle size, shape and distribution) that causes baseline shifts and also compensates for variations due to the optical path of radiation through the sample, which can also be affected by particle size and distribution. The correction is necessary because these variations are not due to the chemical composition of the sample. SNV therefore allows additive and multiplicative effects to be corrected simultaneously, improving the comparability between spectra and reducing non-chemical variability.

An additional tool widely used for spectroscopic data that allows the correction of systematic artefacts and highlight chemical information of interest are *derivatives*. In particular, the first derivative (D1) is used to correct baseline shifts, while the second derivative (D2) allows for the correction of both shifts and baseline drifts, as well as improving the resolution between partially overlapping spectral bands, highlighting small differences between very similar profiles. The disadvantage of using these pretreatments lies in the reduction of the S/N ratio, as the calculation of derivatives also tends to

---

[96] Å. Rinnan, F. van den Berg, and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," *Trends in Analytical Chemistry*, vol. 28, no. 10, pp. 1201–1222, 2009.

[97] R. J. Barnes, M. S. Dhanoa, and S. J. Lister, "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra," *Appl. Spectrosc.*, vol. 43, no. 5, pp. 772–777, 1989.

amplify the noise present in the data[98]. The solution to this drawback lies in the use of the Savitzky-Golay filter, which allows the derivative to be estimated and, at the same time, performs a signal smoothing operation, reducing noise without losing relevant information[99]. This filter approximates a portion of the spectrum using polynomial functions of a predetermined order calculated on a moving window of consecutive signal points, as shown in *Equation 4*

$$U(x) = \sum_{q=0}^{i} c_q x^q \tag{4}$$

where $U(x)$ is the approximate value of the signal in the spectrum, i the order of the polynomial used for approximation, q the degree of the polynomial term, c_q the polynomial coefficients determined using the least squares method x the independent variable (e.g. wavelength).
and for each point, the smooth or derived values are recalculated using a least squares convolution[100] (*Equation 5*).

$$\frac{d^q \hat{u}_i}{dx^q} = \sum_{s=-m}^{m} p_s^{(q)} u_{i-s} \tag{5}$$

where i indicates the point of the signal at which the derivative is being calculated and is defined as $i = m + 1 \ldots, N - m - 1$ (in which N represents the total number of points in the spectrum) and m represents half of the moving window.

This demonstrates the usefulness of applying least squares convolution to a set of equally spaced points centred around a point u_i which allows the determine a set of coefficients $p_s^{(q)}$ that provide the smoothed value of the q-th derivative.

---

[98] V.-M. Taavitsainen, "Denoising and signal-to-noise ratio enhancement: derivatives," in *Comprehensive Chemometrics*, S. D. Brown, R. Tauler, and B. Walczak, Eds., vol. 2. Oxford, U.K.: Elsevier, 2009, pp. 57–66.
[99] H. H. Madden, "Comments on the Savitzky–Golay convolution method for least-squares-fit smoothing and differentiation of digital data," *Anal. Chem.*, vol. 50, no. 9, pp. 1383–1386.
[100] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639.

The key to applying this algorithm is in the choice of polynomial order and number of window points, since a polynomial of too low an order would not adequately represent the shape of the signal, while one of too high an order may adapt to noise rather than actual data. Similarly, a window that is too narrow may not reduce the noise sufficiently, while one that is too wide may cause excessive smoothing, resulting in the loss of important information.

A correct compromise between these parameters allows stable derivatives and meaningful information to be obtained.

It is often useful to combine pretreatments sequentially by performing SNV+D1 or SNV+D2 derivative calculations after SNV in order to optimise accuracy and reduce unwanted variability in model performance.

Preprocessing aims to reduce irrelevant or non-chemical variability, such as baseline shifts, scattering effects, or instrumental noise, while enhancing the chemical information of interest. This step improves the linear relationship between spectral data and the analytical properties to be predicted.

However, improper or excessive application can lead to the loss of useful information, so it is essential to select the most appropriate combination for the model and limit pre-processing operations to the minimum necessary in order to preserve the simplicity and robustness of the model.

3. Calibration and validation

For the construction of a chemometric model, calibration and validation are essential. The first defines the relationship between the independent variables in $\boldsymbol{X}$ and the responses $\boldsymbol{Y}$ using reference samples; in the validation, however, the model's reliability is verified by assessing its capacity to provide correct predictions on data not used in the calibration phase[101]. When the number of samples collected during an analysis campaign is insufficient to divide the samples into separate data sets, i.e. training samples (used for calibration) and test (used for external validation), cross-validation is used, in which the data is iteratively divided into several subsets, with part of the

[101] F. Westad, M. Bevilacqua, and F. Marini, "Chapter 4 – Regression," in *Chemometrics in Food Chemistry*, F. Marini, Ed., vol. 28 of *Data Handling in Science and Technology*. Amsterdam, Netherlands: Elsevier, 2013, pp. 127–170.

samples being used as a training set and the remainder as a test set for each iteration. Repeating this iteration several times provides an estimate of the model's performance, reducing dependence on a single division of the data. If, on the other hand, a sufficient data set is available to allow the data to be divided into training and test sets, cross-validation can be used to calibrate the model, which will then be validated with the external data set (test set). This procedure is necessary both for calibration models, in which validation serves to assess both the model's ability to correctly assign samples to their class, and in regression models in which cross-validation can be used (using the *Leave-One-Out* (LOO) approach, in which one sample at a time is excluded to be considered a test sample, or using the *k-fold* approach in which several samples are excluded simultaneously in an iterative manner) to determine the optimal number of latent variables, as in PLS, and to estimate the predictive quality of the model using the Root Mean Square Error of Cross-Validation (RMSECV) parameters if no test set samples are available, and the Root Mean Square Error of Prediction (RMSEP), as well as R^2 and the bias. This allows evaluating the ability to adequately describe the calibration data and the predictive power of the regression model.

In chemometrics, it is essential to validate a model in order to assess its reliability and predictive ability. Without proper validation, the model would have no statistical significance.

To perform significant validation, the available data must be divided into two representative subsets namely a calibration set (used to build the model) and a test set (used to evaluate its performance).

A common approach to achieving this is the *Duplex* algorithm[102] which allows the division of the data set into a calibration set and a test set by iteratively selecting samples in order to ensure that in both subsets the samples are distributed evenly so as to cover the variability of the multivariate space, namely the distribution of samples in the multivariate space of independent variables. The test set must contain a sufficient number of samples to cover all regions of the multivariate space without impoverishing the training, to this end, the typical distribution includes 20-30% of the samples in the test and the rest in calibration, thus obtaining reliable validation.

---

[102] R. D. Snee, "Validation of regression models: Methods and examples," *Technometrics*, vol. 19, no. 4, pp. 415–428, 1977.

4. Exploratory analysis

In the food sector, the interest in satisfying consumer needs and the focus on production practices has increased. In this context, food characterisation involves the need to extract information and obtain models capable of deducing the underlying relationships that link the compositional profile and process conditions to general food properties such as antioxidant properties and the link with the territory. For this purpose, exploratory data analysis is particularly suitable, as it is defined as the operational approach to data analysis aimed at improving the understanding and accessibility of results. Exploratory analysis (EA) reveals hidden and unknown information from the data in a graphical representation that allows for immediate and direct understanding, reliably highlighting similarities and differences between the samples being analysed, as well as the presence of any clusters within them, outliers and trends[103] [104] [105].

The complexity of the data obtained following chemical analysis of food samples, which nowadays allows a large number of variables to be obtained in a short time, requires a multivariate approach to interpretation.

The multivariate approach allows for an assessment of the simultaneous effect of all variables characterising a system based on the relationships between its samples.

4.1 Principal Component Analysis

Multivariate analysis based on the projection of samples onto a new reduced space defined by latent variables offers a better view for highlighting the characteristics between samples, and this is precisely what happens in PCA, where the space of the acquired variables is reduced to a lower-dimensional space defined by new principal components (in the direction of maximum data variance), and the samples are projected into a

---

[103] S. Sharma, H. Rani, G. Kaur, S. Kumar, S. Sheikh, and M. K. Samota, "Comprehensive overview of glucosinolates in crucifers: occurrence, roles, metabolism, and transport mechanisms—a review," *Phytochem. Rev.*, vol. 24, no. 3, pp. 2417–2444.

[104] M. Li Vigni, C. Durante, and M. Cocchi, "Chapter 3 – Exploratory data analysis," in *Chemometrics in Food Chemistry*, F. Marini, Ed., vol. 28 of *Data Handling in Science and Technology*. Amsterdam, Netherlands: Elsevier, 2013, pp. 55–126.

[105] M. Komorowski, D. Marshall, J. Salciccioli, and Y. Crutain, "Exploratory data analysis," in *Secondary Analysis of Electronic Health Records*, Springer, 2016, pp. 185–203.

reduced-dimensional space defined by new directions called principal components, which provides final graphs that are easy to interpret and suitable for human visualisation.

PCA is therefore a technique that allows large amounts of data to be condensed into a few parameters called principal components, which capture the similar and different properties between the samples and variables that make up the data examined, representing multivariate data in a low-dimensional space. This is done by means of a linear transformation with the constraint of preserving variance and keeping the latent variables orthogonal to each other so that they are uncorrelated, while allowing simultaneous visualisation of the sample space (scores plot) and variables (loadings plot) or through a simultaneous representation of both (biplot)[106].

Following the representation in the new space of orthogonal variables, the data matrix $\boldsymbol{X}$ will undergo bilinear decomposition according to *Equation 6*

$$\boldsymbol{X}_{N,V} = \boldsymbol{T}_A \boldsymbol{P}_A^T + \boldsymbol{E}_{N,V} \qquad (6)$$

where $\boldsymbol{T}$ is the score matrix, which contains the coordinates of the samples in the new component space defined as $N \times A$, in which N corresponds to the number of samples and A to the number of principal components; $\boldsymbol{P}$ is the matrix of loadings containing the coordinates of the original variables in the new space defined by the principal components ($A \times V$) and finally $\boldsymbol{E}$, that represents the matrix of residuals and contains the part of information that the model cannot explain, therefore it has the same dimensions as $\boldsymbol{X}$ ($N \times V$), as illustrated in *Figure 11*

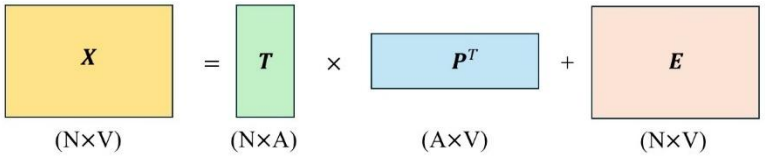


Figure 11. Schematic representation of PCA

[106] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., Springer Series in Statistics. New York, NY, USA: Springer, 2002.

The principal components are linear combinations of the original variables and are described by *Equation 7*:

$$\boldsymbol{t}_a = \boldsymbol{X}\boldsymbol{p}_a \qquad (7)$$

where t_a is the vector of scores and describes the projection of the samples onto the new *a*-th component, while p_a is the vector of loadings that collects the weight of the linear combination.

Since $\boldsymbol{X}$ contains the greatest variance useful for solving the problem, the objective is to maximise the variance captured in $\boldsymbol{t}_a$ by choosing optimal weights ($\boldsymbol{p}_a$) for the linear combination. This can be done by normalising the loading vectors (*Equation 8*)

$$\boldsymbol{p}_a^T\boldsymbol{p}_a = 1 \qquad (8)$$

and performing an orthogonalisation (*Equation 9*)

$$\boldsymbol{p}_a^T\boldsymbol{p}_{\alpha+1} = 0 \qquad (9)$$

To do this, it is possible to apply two different algorithms: *Non-linear Iterative Partial Least Squares* (NIPALS)[107] particularly suitable when only the first principal components are required, as it extracts them sequentially, iteratively optimising the weights until the convergence criteria are achieved; *Single Value Decomposition* (SVD)[108] in which all the main components are obtained simultaneously and not sequentially, describing the data as orthogonal vectors while keeping the results of the information linearly independent; *Eigenvalue Decomposition* (EVD)[109] decomposition to eigenvalues that unlike NIPALS operates on the matrix $\boldsymbol{X}^T\boldsymbol{X}$.

---

[107] H. Wold, "Estimation of principal components and related models by iterative least squares," in *Multivariate Analysis*, P. R. Krishnaiah, Ed. New York, NY, USA: Academic Press, 1966, pp. 391–420.

[108] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," in *Handbook for Automatic Computation, Volume II: Linear Algebra*, F. L. Bauer, Ed. Berlin, Germany: Springer, 1971, pp. 134–151.

[109] Y. Miyashita, T. Itozawa, H. Katsumi, and S.-I. Sasaki, "Comments on the NIPALS algorithm," *J. Chemom.*, vol. 4, pp. 97–100, 1990.

PCA is an unsupervised method that allows data to be analysed by capturing the maximum information contained in the samples through the most significant variability, separating it from noise and less relevant variability.

4.2 ANOVA-Simultaneous Component Analysis

L'ANOVA-Simultaneous Component Analysis is an algorithm that originates from the combination of ANOVA and Simultaneous Component Analysis (SCA), developed to address the need for performing variance analysis within a complex multivariate context[110]. ASCA therefore allows to study the effect of individual variables on the spectrum. To do this, it is necessary to break down the data matrix *X* into several independent effect matrices using ANOVA and then perform principal component analysis on the individual matrices of interest and then on the individual variables of interest.

The model used in the ASCA algorithm can be represented as follows (*Equation 10*):

$$\boldsymbol{X}_i = \boldsymbol{X}_\alpha + \boldsymbol{X}_\beta + \boldsymbol{X}_\gamma + \boldsymbol{X}_{\alpha\beta} + \boldsymbol{X}_{\alpha\gamma} + \boldsymbol{X}_{\beta\gamma} + \boldsymbol{E} \qquad (10)$$

$\boldsymbol{X}_\alpha$, $\boldsymbol{X}_\beta$, $\boldsymbol{X}_\gamma$ represent the matrices that take into account the main effects of the factors, while $\boldsymbol{X}_{\alpha\beta}$, $\boldsymbol{X}_{\alpha\gamma}$, $\boldsymbol{X}_{\beta\gamma}$ capture the interaction effects between the factors, and finally $\boldsymbol{E}$ represent the residual matrix[111].

The second step, which consists of analysing the principal components separately on each effect matrix, can be represented as (*Equation 11*):

$$\boldsymbol{X}_i = \boldsymbol{T}_i \boldsymbol{P}_i^T \qquad (11)$$
$$\text{with } i = \alpha, \beta, \gamma, \alpha\beta, \alpha\gamma, \beta\gamma$$

where $\boldsymbol{T}_i$ and $\boldsymbol{P}_i$ indicate the matrices of scores and loadings for each factor, respectively. The model is validated by performing permutation testing

[110] J. J. Jansen, H. C. J. Hoefsloot, J. Van Der Greef, M. E. Timmerman, J. A. Westerhuis, and A. K. Smilde, "ASCA: Analysis of multivariate data obtained from an experimental design," *J. Chemom.*, vol. 19, pp. 469–481, 2005.

[111] J. A. Westerhuis, E. J. van Velzen, H. C. Hoefsloot, and A. K. Smilde, "Statistical validation of megavariate effects in ASCA," *BMC Bioinformatics*, vol. 8, p. 322.

and bootstrap procedures112. The permutation test allows the evaluation of the significance of effects and consists of randomly reassigning labels to factors and recalculating the model a certain number of times (permutations) and then comparing the variance explained by the original model associated with a factor with that associated with the null distribution. If the observed value is greater than that of the null distribution, then the effect of the factor is defined as statistically significant and not significant by chance.

The second (bootstrap) indicates which variables are influenced and by which factors. It is performed by evaluating the stability of the loadings, which requires resampling with reinsertion of the original data into numerous subsets and recalculating the eigenvectors $\boldsymbol{P}_i$ each time. This results in an empirical distribution of the loadings (*Equation 12*):

$$\boldsymbol{X}_i^b = \boldsymbol{T}_i^b \boldsymbol{P}_i^b \qquad (12)$$
$$b = 1, 2, \dots, B$$

where $\boldsymbol{X}_i^b$ is the resampled matrix at bootstrap b, while $\boldsymbol{T}_i^b$ and $\boldsymbol{P}_i^b$ re the matrices of the scores and loadings estimated at bootstrap b, respectively. Finally, B is the number of bootstraps113.

Through this procedure, the stability of the loadings is examined, and if a variable is consistently significant in all resamples, then it can be stated that it is influenced by the factor under consideration.

5. Regression

Regression allows to describe one or more dependent variables $\boldsymbol{Y}$ starting from a set of independent variables $\boldsymbol{X}$.

The purpose of regression models is to find a functional relationship between two or more sets of predictive variables.

---

[112] C. Bertinetto, J. Engel, and J. Jansen, "ANOVA simultaneous component analysis: A tutorial review," *Anal. Chim. Acta X*, vol. 6, p. 100061, 2020.

[113] H. Babamoradi, F. van den Berg, and Å. Rinnan, "Bootstrap based confidence limits in principal component analysis—A case study," *Chemom. Intell. Lab. Syst.*, vol. 120, pp. 97–105, 2013.

The regression approximates the responses provided by predictors contained in $\boldsymbol{X}$, which best fit the data according to the least squares criterion (*Equation 13*):

$$\boldsymbol{Y} = f(\boldsymbol{X}) + \boldsymbol{E} = \widehat{\boldsymbol{Y}} + \boldsymbol{E} \qquad (13)$$

where f indicates a general mathematical relationship and $\boldsymbol{E}$ is the residual matrix representing the difference between the true values of $\boldsymbol{Y}$ and those predicted by the model $\widehat{\boldsymbol{Y}}$.

In many cases in chemometrics, a linear relationship described by *Equation 14* is assumed:

$$f(\boldsymbol{X}) = \boldsymbol{XB} \qquad (14)$$

$\boldsymbol{B}$ is the regression coefficient matrix.

The simplest case is when there are two data matrices, but in analytical chemistry problems are often encounter whit a higher number of them.

When the number of predictive variables $\boldsymbol{X}$ is very high and often highly correlated, it is necessary to introduce the *Partial Least Squares* (PLS) algorithm, which also uses $\boldsymbol{Y}$ information in the construction of latent variables. This improves both the interpretation and predictive power of the model. PLS searches for latent variables that balance the ability to explain the variability of $\boldsymbol{X}$ and, at the same time, to predict $\boldsymbol{Y}$[114].

The PLS model is a bilinear model described by *Equations 15* and *16*:

$$\boldsymbol{X} = \boldsymbol{TP}^T + \boldsymbol{E}_X \qquad (15)$$

$$\boldsymbol{Y} = \boldsymbol{TCQ}^{\boldsymbol{T}} + \boldsymbol{E}_Y \qquad (16)$$

In the *Equation 15* $\boldsymbol{X}$ represents the experimental data matrix, $\boldsymbol{T}$ the scores matrix, $\boldsymbol{P}^T$ the transposed loadings matrix and $\boldsymbol{E}_X$ the residuals matrix while in *Equation 16* $\boldsymbol{Y}$ is the matrix of observed responses; $\boldsymbol{T}$ is the matrix of scores; $\boldsymbol{C}$ is a diagonal matrix holding the coefficients of the PLS inner relation; $\boldsymbol{Q}$ are the Y-loadings and $\boldsymbol{E}_Y$ is the matrix of response residuals.

---

[114] F. Westad, M. Bevilacqua, and F. Marini, "Chapter 4 – Regression," in *Chemometrics in Food Chemistry*, F. Marini, Ed., vol. 28 of *Data Handling in Science and Technology*. Amsterdam, Netherlands: Elsevier, 2013, pp. 127–170.

In many cases of chemical data analysis, information is not only organised in two-dimensional matrices but also in higher-order structures. In this case, the traditional PLS algorithm is not sufficient, and the multiway N-PLS method is used, which stands for N-way Partial Least Squares, allowing multidimensional data to be managed while maintaining the predictive logic of PLS. Many data analyses have the need to identify relationships between two two-dimensional data matrices in order to allow both correlation analysis and prediction analysis of the values of one matrix based on the other. However, when analysing data with a multiway (multidimensional) structure, it is necessary to use algorithms such as *Multilinear PLS* (N-PLS) that are capable of handling higher-order arrays.

5.1 N-PLS

N-PLS is a PLS that uses multilinear models to analyse data, allowing it to take multiple data dimensions into account simultaneously[115] [116]. This algorithm is an extension of PLS to cases where $\boldsymbol{X}$ and/or $\boldsymbol{Y}$ are multidimensional arrays. N-PLS also maximises the covariance between the scores of $\boldsymbol{X}$ and those of $\boldsymbol{Y}$ by performing a decomposition that is done in a multilinear manner, respecting the different dimensions of the arrays[117]. Therefore, the N-PLS model is constructed by simultaneously adapting a multilinear model for each of the two data matrices ($\boldsymbol{X}$ and $\boldsymbol{Y}$) *Equation 17* and *18* and a regression model that relates the corresponding scores (*Equation 19*)

$$\boldsymbol{X}^{(I\times JK)} = \boldsymbol{T}(\boldsymbol{W}^{K} \odot \boldsymbol{W}^{J}) + \boldsymbol{E}_{X} \qquad (17)$$

$$\boldsymbol{Y}^{(I\times LM)} = \boldsymbol{P}(\boldsymbol{Q}^{M} \odot \boldsymbol{Q}^{L})^{T} + \boldsymbol{E}_{Y} \qquad (18)$$

$$\boldsymbol{P} = \boldsymbol{TB} \qquad (19)$$

---

[115] R. Bro, "Multi-way calibration. Multi-linear PLS," *Journal of Chemometrics*, vol. 10, no. 1, pp. 47–61, 1996.

[116] R. Bro, A. K. Smilde, and S. de Jong, "On the difference between low-rank and subspace approximation: Improved model for multi-linear PLS regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 1, pp. 3–13, 2001.

[117] J. M. Amigo and F. Marini, "Chapter 7 – Multiway methods," in *Chemometrics in Food Chemistry*, F. Marini, Ed., vol. 28 of *Data Handling in Science and Technology*. Amsterdam, Netherlands: Elsevier, 2013, pp. 265–313.

The main advantage of using N-PLS over unfolding followed by classical PLS lies in its ability to preserve the multiway structure of the data, maintaining the relationships among all modes. The latent components simultaneously describe the variation along each dimension of the tensor, thereby reducing information loss and improving model interpretability compared to the two-dimensional unfolded approach.

If the response is a single response variable ($\boldsymbol{y}$), this falls within the N-PLS1 case. Two sets of weights are sought for each component, $\boldsymbol{w}^K$ and $\boldsymbol{w}^J$, along the second and third modes defined in *Equation 20* such that the scores ($\boldsymbol{t}$) (*Equation 21*), obtained by projecting $\boldsymbol{X}$ on the weight vectors, have maximum covariance with $\boldsymbol{y}$ (*Equation 22*)

$$\boldsymbol{w} = \boldsymbol{w}^K \otimes \boldsymbol{w}^J \tag{20}$$

$$\boldsymbol{t} = \boldsymbol{X}^{(I\times J\times K)}\boldsymbol{w} \tag{21}$$

$$\max_{\boldsymbol{w}} cov\,(\boldsymbol{t}, \boldsymbol{y}),\ \|\boldsymbol{w}\| = 1 \tag{22}$$

To identify the optimal weights ($\boldsymbol{w}^K$ and $\boldsymbol{w}^J$) to satisfy *Equation 22,* it is necessary to calculate the $\boldsymbol{Z}$ matrix according to *Equation 23*

$$z_{jk} = \sum_{i=1}^{I} y_i x_{ijk} \tag{23}$$

where i represents the index of the samples; j and k represent the indices of the second and third modes of $\boldsymbol{X}$; y_i represents the value of the response variable for the i-th sample; and finally x_{ijk} indicates the value observed for sample i corresponding to modes j and k of $\boldsymbol{X}$.

The resulting matrix $\boldsymbol{Z}$ is a two-dimensional matrix representing the correlation between the predictive variables, distributed along modes j and k, and the response. In this way, it is not necessary to explore the entire multiway array, but it is sufficient to search for the first left and right eigenvectors of $\boldsymbol{Z}$ (using Singular Value Decomposition[118]) to obtain the desired weights ($\boldsymbol{w}^K$ and $\boldsymbol{w}^J$), as show in *Equation 24*.

[118] The *Singular Value Decomposition* (SVD) performed on $\boldsymbol{Z}$ orders the possible directions, which derive from the combination of weights, according to the explained covariance. Therefore, the eigenvectors associated with the

$$svd(\boldsymbol{Z}) = \boldsymbol{w}^J \zeta (\boldsymbol{w}^K)^T \qquad (24)$$

ζ is the first singular value representing the covariance between $\boldsymbol{X}$ and $\boldsymbol{y}$ captured by the pair of weights $\boldsymbol{w}^K$ and $\boldsymbol{w}^J$. Since the algorithm is sequential, before the extraction of each subsequent component, it is necessary to perform a deflation.
The advantage of using N-PLS for multiway data lies in the fact that, unlike approaches based on unfolding methods, where the data tensor is rearranged into a two-dimensional matrix (thus losing its natural multiway structure), N-PLS preserves the intrinsic structure of the data and operates directly on it. In this way, the joint correlations among modes are exploited, reducing sensitivity to noise and shifts, and improving both the interpretability and robustness of the model.

6. Classification methods

To make a classification means to find a criterion to assign an object (sample) to one category (class) based on a set of measurements performed on the object itself. This yields discrete responses that can be coded by a nominal variable or a natural number, since in classification the categories are defined a priori[119].

When talking about classification, the purpose is to find a criterion to make a prediction on the samples, that is, to assign these samples to a class on the basis of the analyses carried out on the sample itself. This allows finding a way to identify separation surfaces in the multivariate space so that the sample is located in a specific region of space (representing the class).

The ultimate goal is to be able to classify an unknown sample using a model built on the basis of samples belonging to a training set for which, in addition to the measured data, the classes to be characterised are also known.

---

maximum singular values represent the directions along which the covariance between the response and the scores of X is maximum and for this reason they are chosen for the construction of the N-PLS model.

[119] M. Bevilacqua, R. Bucci, A. D. Magrì, A. L. Magrì, and R. Nescatelli, "Classification and class-modeling," in *Chemometrics in Food Chemistry*, F. Marini, Ed. Oxford, U.K.: Elsevier, 2013, pp. 171–233.

The starting hypothesis is the presence of sample classes within the data. From here, an attempt is made to correlate the measured information with the response, that is with the information on the categorisation of the samples to be predicted.

At this point, it is possible to proceed using two main strategies, discriminative or modelling, which, being two similar approaches but with different characteristics and results, reflect the two different ways in which the classification problem can be addressed[120].

One of the main requirements of classification approaches, being predictive models, is to have a representative set of samples for each of the categories to be defined so that they can be used for model construction and validation. Classification can also be described as the identification of decision surfaces, which act as boundaries for the different categories where each sample characterised by a vector of measurements can be geometrically represented as a point in multivariate space.

It is possible to classify data if it contains the information necessary to distinguish between classes otherwise, no model will function correctly. The objective of a multivariate model is to predict the response value of new samples that were not included in the construction of the model itself.

6.1 Discriminant approach

Discriminant approaches focus on constructing decision surfaces, thus identifying and tracing, in the best possible way, surfaces that divide the multivariate space, defined by the variables being measured, into as many regions as there are classes represented in the training set.

[120] C. Albano, W. Dunn III, U. Edlund, E. Johansson, B. Nordén, M. Sjöström, and S. Wold, “Four levels of pattern recognition,” *Anal. Chim. Acta*, vol. 103, pp. 429–443, 1978.

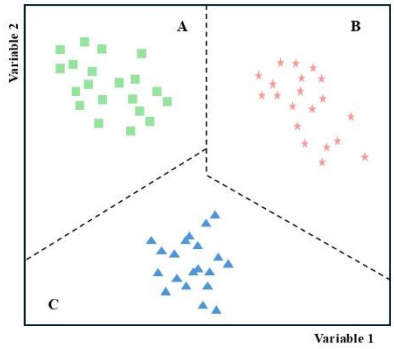


Figure 12. Example of subdivision of multivariate space into regions representing classes in the case of discriminant approaches

Discriminant techniques are characterised by the fact that they use samples from the training set to identify the best separation surfaces between classes, namely the best surfaces that divide the multivariate space into as many regions as there are classes, providing the most accurate predictions possible. These regions are very often contiguous, but there may also be disjoint regions, resulting in the existence of sub-regions associated with the same category. Discriminant methods do not offer any possibility of detecting outliers, that is, of highlighting samples that are not well represented by the modelled classes. In the case of discriminant classification, establishing a certain method and calculating the coefficients of the chosen model in order to minimise the total classification error corresponds to applying Bayes' rule[121]. In all cases, this rule is the one that explicitly or implicitly leads, given the chosen method, to the lowest classification error.

An example of a discriminant classification approach is the *Partial Least Squares-Discriminant Analysis* (PLS-DA) algorithm. This algorithm is useful

[121] This rule allows to predict the class of the unknown sample, corresponding to the category to which the unknown sample is most likely to belong. This presupposes being able to calculate the posterior probability that any sample belongs to each of the various classes so that making a prediction based on the values of these probabilities and then assigning the sample to the class for which this probability is highest.

for analysing highly correlated data, such as data obtained from spectroscopic analyses[122] [123] [124].

Belonging to the category of supervised methods, this approach allows the transformation of a classification problem into a regression problem, which is then solved using PLS (combining multivariate regression and classification, resulting in a robust method in the presence of collinearity). The class membership is represented by a binary matrix $\boldsymbol{Y}$ (called Dummy) of dimension $N \times F$ where N is the number of samples and F the number of classes. Therefore, each sample will have a value of 1 in the column corresponding to the class to which it belongs and 0 in the others.

The generated model is described as follows (*Equation 25*):

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E} \qquad (25)$$

$\boldsymbol{X}$ is the training data matrix, $\boldsymbol{B}$ the regression coefficients and $\boldsymbol{E}$ the residual matrix.

Solving this equation yields a predicted vector $\widehat{\boldsymbol{Y}}$ which will not be binary but will contain real values within a continuous interval representing the estimated probability of belonging to each class. To classify the samples, *Linear Discriminant Analysis* (LDA) is applied to the predicted values. Operating on $\widehat{\boldsymbol{Y}}$, LDA classifies the samples according to Bayes' rule, which maximises a posteriori membership of the class.

$$p(c|\boldsymbol{x}_i) \propto p(\boldsymbol{x}_i|c)p_0(c) \qquad (26)$$

The proportionality relation (*26*) defines how the posterior probability that the i-th ample belongs to class c is estimated. This probability is proportional to the product of the likelihood of the class ($p(\boldsymbol{x}_i|c)$) and the prior probability of the class itself ($p_0(c)$)[125]. In this case, the dummy response matrix Y used

[122] M. Barker and W. Rayens, "Partial least squares for discrimination," *Journal of Chemometrics*, vol. 17, no. 3, pp. 166–173, 2003.

[123] H. Nocairi, E.-M. Qannari, E. Vigneau, and D. Bertrand, "Discrimination on latent components with respect to patterns: Application to multicollinear data," *Computational Statistics & Data Analysis*, vol. 48, no. 1, pp. 139–154, 2005.

[124] U. G. Indahl, H. Martens, and T. Næs, "From dummy regression to prior probabilities in PLS-DA," *Journal of Chemometrics*, vol. 21, no. 12, pp. 529–536, 2007.

[125] N. F. Pérez, J. Ferré, and R. Boqué, "Calculation of the reliability of classification in discriminant partial least-squares binary classification," *Chemom. Intell. Lab. Syst.*, vol. 95, pp. 122–128, 2009.

in the PLS-DA model has a number of columns equal to the number of classes minus one. The construction of the model requires the selection of the optimal number of principal components, which is carried out using cross-validation. The models obtained are evaluated primarily using the confusion matrix, thus placing greater attention on the quality of the representation of each class.

6.2 Modeling approach

Class modelling techniques are a family of tools that address the classification problem by modelling one category at a time[126]. They focus on the similarities between samples belonging to a particular category and operate by defining a multivariate boundary that identifies the region of space where individuals of that particular class are likely to be found.

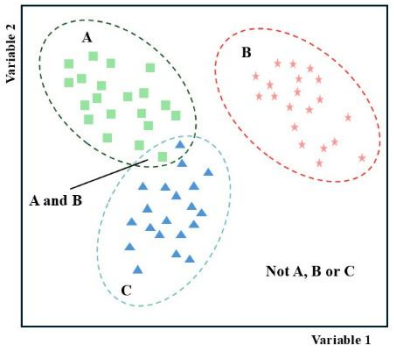


Figure 13. Example of class space identification in the case of modelling approaches

With the training set, a model is constructed to define the region of space in which samples belonging to that specific class can be found. Unlike discriminant methods, the advantage is that if you are only interested in a particular class, it is not necessary to model the other classes as well[127]. Furthermore, the regions of space associated with a particular class are closed

[126] P. Oliveri and G. Downey, "Multivariate class modeling for the verification of food-authenticity claims," *TrAC, Trends Anal. Chem.*, vol. 35, pp. 74–86, 2012.
[127] M. Forina, P. Oliveri, S. Lanteri, and M. Casale, "Class-modeling techniques, classic and new, for old and new problems," *Chemom. Intell. Lab. Syst.*, vol. 93, pp. 132–148, 2008.

regions. Since class models are defined independently of each other, it may happen that samples are confused, meaning that they belong to multiple classes whose models overlap. These types of models are evaluated using figures of merit such as sensitivity, specificity and efficiency. These represent, respectively, the percentage of samples correctly accepted by the model and the percentage of samples correctly rejected by the model; efficiency is the geometric mean of the two previous figures.

Among the various modelling classification algorithms, the *Soft Independent Modeling of Class Analogies* (SIMCA) approach was applied, according to which an unknown sample will be assigned to a class only if, when projected into the multivariate subspace of the model, it shows sufficient similarity to the samples belonging to that particular class. To do this, Principal Component Analysis is used to approximate the typical variability of the class of interest and evaluate its similarity[128].

The sample will then be accepted or rejected by the class model depending on its distance from the class space itself (*Equation 27*):

$$d^2 = \left(\frac{T_i^2}{T_{0.95}^2}\right)^2 + \left(\frac{Q_i}{Q_{0.95}}\right)^2 = \left(T_{i,red}^2\right)^2 + \left(Q_{i,red}\right)^2 \qquad (27)$$

here T^2 represents the Mahalanobis distance of the sample from the centre of the class model space, while Q represents its orthogonal distance from the model. These values are divided by the limit values ($T_{0.95}^2$ e $Q_{0.95}$) of the distributions of T^2 e Q. This gives the reduced distances $T_{i,red}^2$ e $Q_{i,red}$.

The normalisation thus gives limit values of T^2 and Q equal to 1, so it can be said that if $d \leq \sqrt{2}$, the sample will be accepted by the model, otherwise if $d \geq \sqrt{2}$, the sample will be rejected by the model.

7. Design of Experiment

Contrary to the traditional *One Variable At a Time* (OVAT) method, which involves evaluating one variable at a time, ignoring possible interactions

[128] R. Vitale et al., "Class modelling by Soft Independent Modelling of Class Analogy: Why, when, how? A tutorial," Analytica Chimica Acta, vol. 1270, art. no. 341304, 2023.

between factors, *Design of Experiment* (DoE) is able to simultaneously estimate the main effects, interactions and, if necessary, quadratic terms. It does this by varying several variables at the same time in order to take into account the various effects due to the interaction between them, with the advantage of requiring fewer experiments to be carried out[129].

A classic example is the full factorial design which, in the case of three levels, for k variables requires 3^k experiments. This factorial design explores the vertices of a cube in the factor space and allows to estimate a model of the type (*Equation 28*):

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_{12}X_1X_2 + b_{13}X_1X_3 + b_{23}X_2X_3 + b_{123}X_1X_2X_3 \quad (28)$$

This enables the estimation of both linear effects and higher order interactions. The results derived from an experimental design are interpreted through the analysis of significant factors and also through the construction of response surfaces and iso-response diagrams that allow the behaviour of the system in the experimental space to be represented graphically. These graphical tools allow regions of maximum or minimum response to be identified immediately, identifying the conditions for optimising the process by visualising the phenomenon in a global way. To ensure the reliability of the model, it is necessary to evaluate the statistical significance of the coefficients using a *t*-test, considering the degrees of freedom given by $N_{replicate} - 1$ and calculating the standard deviation of the coefficients as shown in *Equation 29*:

$$S_c = \sqrt{S^2_{pooled} C_{ii}} \quad (29)$$

S^2_{pooled} represents the combined experimental variance estimated from independent measurements and C_{ii} represents the variance associated with the coefficient, that is the diagonal element of the variance-covariance matrix.

To validate the model, one can use Analysis of Variance (ANOVA), which allows one to quantify the overall significance of the model and of the individual factors, the lack of fit, which assesses the adequacy of the model with respect to the experimental data (it compares the error of the model with

[129] R. Leardi, "Experimental design in chemistry: A tutorial," *Anal. Chim. Acta*, vol. 652, pp. 161–172, 2009.

the error due to experimental replication, but in the absence of replicates it is not possible to calculate it), and the predictive ability of the model measured in terms of its ability to correctly estimate new observations within the experimental domain. These tools ensure that the model obtained not only accurately describes the data collected but is also capable of making predictions in experimental and industrial contexts.

As the number of variables increases, full factorial designs quickly become burdensome. For this reason, alternative designs such as *D-Optimal design*, which is based on information theory, are often used. This approach selects the subset of experiments that maximises the determinant of the information matrix ($\boldsymbol{X}^T\boldsymbol{X}$), reducing the correlation between the estimated coefficients and ensuring the highest possible accuracy of the model with fewer tests than a full factorial design, while maintaining predictive power over the experimental domain.

The DoE therefore offers the possibility of conducting an empirical exploration of factors as well as performing true multivariate modelling, with advantages in terms of reducing the number of experiments to be carried out, cost, time and increasing the quality of information with the ability to solve complex method optimisation problems.

The experimental design allows for the optimisation of experimental procedures by evaluating the interaction effects between the variables considered and the quadratic effects of the factors.

The experimental design allows for the systematic exploration of the entire experimental domain and the creation of an empirical mathematical model capable of predicting the response across the entire space.

In this work, all of this was done with the help of CAT software[130].

[130] R. Leardi, C. Melzi, and G. Polotti, *CAT (Chemometric Agile Tool)*.

Section 1

Brassicaceae

CHAPTER 4

MORPHOLOGICAL ANALYSIS AND OPTIMISATION OF EXTRACTION FOR THE ANALYSIS OF THE SPECTROCHROMATOGRAPHIC PROFILE OF SPECIES BELONGING TO THE BRASSICACEAE FAMILY AND EVALUATION OF ANTIOXIDANT ACTIVITY

1. Aim of the study on Brassicaceae varieties

The Brassicaceae have attracted considerable interest due to their rich chemical composition, particularly for the presence of antioxidants and anti-inflammatory compounds, as well as their antimutagenic and potentially antioxidant activity[131] [132].

[131] D. Lučić, I. Pavlović, L. Brkljačić, S. Bogdanović, V. Farkaš, A. Cedilak, L. Nanić, I. Rubelj, and B. Salopek-Sondi, "Antioxidant and antiproliferative activities of kale (*Brassica oleracea* L. var. *acephala* DC.) and wild cabbage (*Brassica incana* Ten.) polyphenolic extracts," *Molecules*, vol. 28, p. 1840, 2023.

[132] A. Basit, S. Ahmad, K. U. R. Khan, H. Y. Aati, A. E. Sherif, C. Ovatlarnporn, S. Khan, H. Rao, M. A. Arshad, M. N. Shahzad, *et al.*, "Evaluation of the anti-inflammatory, antioxidant, and cytotoxic potential of *Cardamine amara* L. (*Brassicaceae*): A comprehensive biochemical, toxicological, and in silico computational study," *Front. Chem.*, vol. 10, p. 1077581, 2022.

The Brassicaceae family includes approximately 4,000 known species, rich in key antioxidants such as vitamin C and flavonoids, which are beneficial to human health[133], as well as glucosinolates[134] which have anti-inflammatory and cardioprotective properties[135] and act as chemopreventive agents both in vitro and in vivo[136]. Plants belonging to this family are common and cultivated in Italy as well as throughout the world, together with numerous local ecotypes.

In Italy, several varieties of Brassicaceae are commercially available, including traditional local cultivars with unique characteristics and genetic diversity; consequently, there is a growing need to protect and recognise these landraces in order to preserve biodiversity.

In Italy, there are many cultivated varieties and a wide range of landraces with particular morphological and agronomic characteristics, including local variants adapted to withstand low temperatures and harsh winter conditions in mountainous regions. Interest in the latter is growing, both for scientific reasons and for the purpose of preserving genetic biodiversity. The preservation is of interest because these traditional species have been discarded by seed companies in favour of more marketable plants, such as those that produce more uniform crops (simultaneous rather than staggered harvesting) and are less prone to hybridisation, so as to maintain homogeneous and uniform production for easier mechanical harvesting and reusable seeds. Today, institutions provide support to growers of local varieties to ensure their protection and conservation.

These traditional species often possess valuable agronomic traits, such as strong natural defence mechanisms. This natural defence ability is due to the presence of glucosinolates, which undergo hydrolysis when the plant is

---

[133] H. E. Khalil, M. F. Abdelwahab, P. M. Emeka, L. I. Badger-Emeka, S. M. N. Abdel Hafez, K. A. AlYahya, A.-S. F. Ahmed, A. F. Anter, N. M. Abdel-Wahab, K. Matsunami, *et al.*, "Chemical composition and valorization of broccoli leaf by-products (*Brassica oleracea* L. variety: *Italica*) to ameliorate reno-hepatic toxicity induced by gentamicin in rats," *Appl. Sci.*, vol. 12, p. 6903, 2022.

[134] J. S. Jo, S. R. Bhandari, G. H. Kang, Y. K. Shin, and J. G. Lee, "Selection of broccoli (*Brassica oleracea* var. *italica*) on composition and content of glucosinolates and hydrolysates," *Sci. Hortic. (Amsterdam)*, vol. 298, p. 110984, 2022.

[135] K. M. Favela-González, A. Y. Hernández-Almanza, and N. M. la Fuente-Salcido, "The value of bioactive compounds of cruciferous vegetables (*Brassica*) as antimicrobials and antioxidants: A review," *J. Food Biochem.*, vol. 44, e13414, 2020.

[136] X. Wang, Y. S. Chan, K. Wong, R. Yoshitake, D. Sadava, T. W. Synold, P. Frankel, P. W. Twardowski, C. Lau, and S. Chen, "Mechanism-Driven and Clinically Focused Development of Botanical Foods as Multitarget Anticancer Medicine: Collective Perspectives and Insights from Preclinical Studies, IND Applications and Early-Phase Clinical Trials," *Cancers*, vol. 15, p. 701, 2023.

damaged, transforming into bioactive isothiocyanates, volatile compounds known for their broad-spectrum antimicrobial and insecticidal activities. As a result, the cultivation of Brassicaceae has proven effective in reducing the need for pesticides or other chemical inputs and is therefore considered an ecologically sustainable crop[137].

In this study, various local and commercial varieties of leaves belonging to the Brassicaceae family were analysed using non-destructive techniques such as E-Eye and attenuated total reflectance Fourier transform infrared spectroscopy (ATR-FT-IR). At the end of the non-destructive investigation of the leaves, they were stored at -20°C and subsequently analysed using chromatographic techniques such as HPLC-DAD.

2. Non-destructive analysis on leaves

In this first part of the non-destructive study of varieties, the focus is on Mugnoli, a typical local ecotype from Abruzzo grown in Pettorano sul Gizio (Abruzzo, Italy).

Table 1 lists the species considered in this study.

Varieties	Class
Mugnoli, accession 1	A
Mugnoli, accession 2	B
Mugnoli, accession 3	C
Guardiagrele Turnip	D
Curly Kale	E
Rapa Senza Testa	Γ
Cima dell'Osento	F
Cima 90° San Marzano	G
Cima Grande	H

[137] Flor-Weiler, L.B.; Behle, R.W.; Berhow, M.A.; McCormick, S.P.; Vaughn, S.F.; Muturi, E.J.; Hay, W.T. Bioactivity of brassica seed meals and its compounds as ecofriendly larvicides against mosquitoes. Sci. Rep. 2023, 13, 3936.

Table 1. List and respective classes of the species studied

Curly kale is the only international species not typical of the area and, like Mugnoli belongs to the species *Brassica oleraceae* L. The other landraces such as Guardiagrele turnip, Rapa Senza Testa and Cima dell'Osento, together with Cima 90° San Marzano and Cima Grande are varieties belonging to the species *Brassica rapa*.

The nine different varieties of Brassicaceae were grown in the same experimental field and subsequently analysed using non-destructive tools such as multivariate image analysis and IR spectroscopy followed by chemometric approaches.

Mugnoli are an ancient local variety from Pettorano sul Gizio, which most likely evolved from the hybridisation of kale, belonging to the *Brassica oleracea L*. species, with turnip, belonging to the *Brassica rapa* L. family. Mugnoli were included in this study because they are a hybrid species, exhibiting traits of both parent species, with some individuals within the population predominantly displaying cabbage characteristics and others predominantly displaying turnip characteristics. Furthermore, as they are easily hybridisable species, individuals generated through hybridisation with other *Brassica* species cultivated in neighbouring areas are also common. Like other Brassicaceae, this plant is distinguished by its nutritional value[138] and, among its various benefits, it is particularly rich in glucosinolates[139].

The Guadiagrele turnip is a landrace cultivated in the locality of the same name, which displays characteristics of the turnip (*Brassica rapa*).

The Curly Kale of Lama dei Peligni also takes its name from the place where it is grown, belonging to the *Brassica oleracea* var. *sabellica* cultivated at the foothills of the Maiella (Abruzzo, Italy).

Rapa Senza Testa is a *Brassica rapa* that does not develop a main stem but is rich in leaves.

Cima dell'Osento is a variety of *Brassica napus* cultivated in the Osento river plain.

---

[138] X. Zhang, Q. Jia, X. Jia, J. Li, X. Sun, L. Min, Z. Liu, W. Ma, and J. Zhao, "Brassica vegetables—an undervalued nutritional goldmine," *Hortic. Res.*, vol. 12, no. 2, p. uhae302, 2024.

[139] M. P. Argentieri, R. Accogli, F. P. Fanizzi, and P. Avato, "Glucosinolates profile of 'mugnolo,' a variety of *Brassica oleracea* L. native to southern Italy (Salento)," *Planta Med.*, vol. 77, pp. 287–292, 2011.

Finally, Cima 90° San Marzano and Cima Grande are two commercial varieties investigated alongside local varieties for comparison.

The three Mugnoli categories differ in terms of the different accessions that come from different producers.

The conservation activities regarding the Mugnoli ecotype involve the cultivation and propagation of seeds, which are not commercially available, by local farmers who maintain the accessions by passing them on from year to year from one crop to another, thus contributing to the conservation of this local variety.

Among the cultivars considered, Lama dei Peligni kale is a local Abruzzo variety that is easily found on the market, but only classes G and H are commonly sold throughout Italy, while the others are all niche varieties.

The above varieties were studied using two non-destructive approaches: multivariate image analysis (MIA)[140]via E-Eye analysis and IR technique. The data acquired using these techniques were analysed chemometrically using exploratory analysis, class modelling and discriminant classification approaches.

These species cultivated in Abruzzo (Italy) are the subject of the study, which is based on the characterisation and classification of these local varieties in order to protect and promote them. In order to evaluate similarities and differences between samples, all cultivated in an experimental field under the same soil and climate conditions, so that any differences between species would be due solely to differences in variety and not to the influence of soil and climate conditions, but mainly related to inter- and intra-landrace differences. *Figure 14* shows the subdivision and organisation of the rows in the experimental field.

[140] A. Antonelli, M. Cocchi, P. Fava, G. Foca, G. C. Franchini, D. Manzini, and A. Ulrici, "Automated evaluation of food colour by means of multivariate image analysis coupled to a wavelet-based classification algorithm," *Anal. Chim. Acta*, vol. 515, pp. 3–13, 2004.

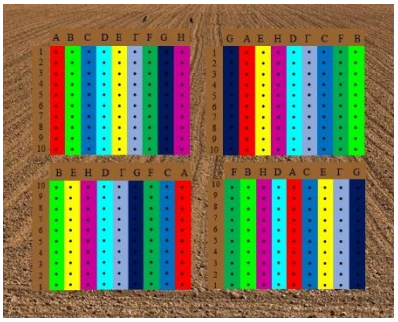


Figure 14. Organization of the experimental field

From *Figure 14*, it is possible to understand the division of the cultivation land into four sub-fields, each consisting of nine rows (one for each class), containing 10 plants. The order of the rows of plants was randomised for each sub-field. The period from sowing to harvest lasted five months (8 September 2022, and the plants were harvested on 6 February 2023). Once harvested, the leaves were stored at 4°C and analysed within three days of harvesting.

The cultivation in an experimental field thus ensured a classification based exclusively on their genetic background.

Figure 8 shows actual photographs of the leaf samples. All show a strong resemblance, which does not allow for immediate distinction and requires in-depth botanical expertise.

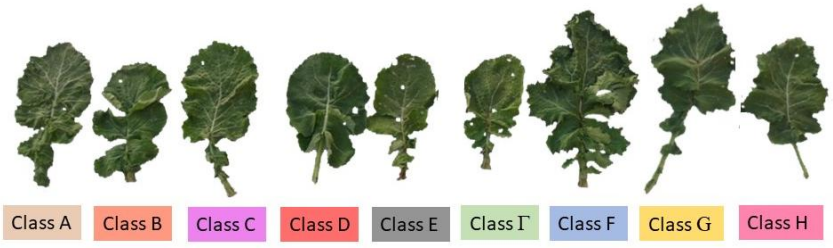


Figure 15. Real photo of the Brassicaceae leaves

The cultivation of Rapa Senza Testa (Class Γ) was unsuccessful and it was not possible to obtain a statistically significant number of samples; consequently, it was excluded from this study.

2.1 E-Eye Analysis

The images of the leaves were acquired using the
RS Pro Wi-Fi USB microscope. *Figure 16* shows an example of an image collected on one of the leaves analysed.

Figure 16. Example of E-Eye image

For each Brassicaceae class studied, 30 images were collected, for a total of 240 images. The colorgrams were created using MATLAB software[141], following the MIA approach.

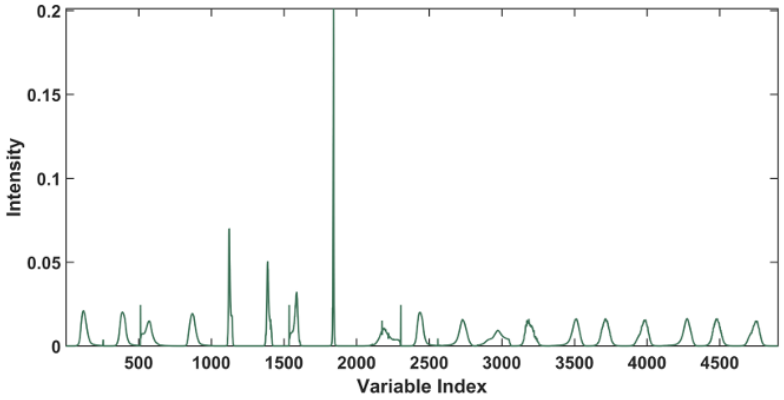


Figure 17. Mean colorgram calculated on the signal of all samples

[141] The MathWorks, Inc., *MATLAB, Version R2021b*. Natick, MA, USA: The MathWorks, Inc., 2021.

The different peaks of the colour diagram visible in *Figure 17* can be associated with specific characteristics of the image. Variables from 1 to 2560 are associated with RGB channels, hue and saturation. In particular, the first three peaks represent the distribution curves of red (R), green (G) and blue (B) values; the next signal is the distribution curve of brightness values (L=R+G+B), followed by the relative values of R, G and B and the distribution curve of hue values. From variable 2561 onwards, the colorgram includes the distribution curves of the scores of the three principal components obtained by PCA applied to the unfolded RGB matrix, calculated on raw, mean-centred and autoscaled data; finally, the normalised loading vectors and the eigenvalues of the three PCA models are reported (raw, mean-centred, autoscaled). Each distribution profile was generated using 256 bins, covering the full variability range, and no normalization was applied to the acquired RGB images prior to colorgram calculation.

The collected colorgrams were divided into a training set of 160 samples (20 for each class) and a test set of 80 objects (10 for each category), according to the Duplex algorithm to ensure representativeness, in order to perform external validation of the models built to execute a class-modelling approach according to SIMCA.

Before calculating the models, the data were autoscaled.

For the calculation of SIMCA models, it is necessary to define the optimal number of PCs to be extracted. To avoid overly optimistic estimates, the optimal complexity of the model was determined through a seven-fold cross-validation procedure. For each class, ten different calibration models were calculated by progressively increasing the number of latent variables from 1 to 10. The quality of each model was evaluated by considering the three cross-validated parameters: sensitivity ($Sens_{cv}$), specificity ($Spec_{cv}$) and efficiency (Eff_{cv}). The analysis of the efficiency in cross-validation revealed the appropriate number of PCs as shown in *Table 2*, numbers of PCs selected for each model and the relative values of sensitivity, specificity and efficiency in cross-validation.

Class	PCs	$Sens_{cv}$	$Spec_{cv}$	Eff_{cv}
A	6	55.0	45.0	39.7
B	4	90.0	86.9	88.4
C	4	75.0	75.4	75.2
D	6	55.0	39.2	46.4
E	4	80.0	69.1	74.3
F	6	75.0	61.3	63.1
G	4	75.0	60.8	67.5
H	4	80.0	68.2	73.8

Table 2. Selected PC numbers for each SIMCA class model and relative sensitivity, specificity and efficiency values in cross-validation.

From the values shown in the *Table 2*, it can be seen that all models, with the exception of those referring to Class A and Class D, show high sensitivity and good specificity, which consequently lead to satisfactory efficiency values. Subsequently, all calibration models were used to predict the test set, thus obtaining the sensitivity ($Sens_{pred}$) and specificity ($Spec_{pred}$) values in prediction as reported in *Table 3*.

Class	PCs	$Sens_{pred}$	$Spec_{pred}$
A	6	30.0	37.1
B	4	100.0	92.8
C	4	70.0	77.1
D	6	70.0	32.8
E	4	90.0	68.6
F	6	80.0	68.6
G	4	90.0	67.1
H	4	80.0	65.7

Table 3. Sensitivity and specificity values of SIMCA models in prediction

These values confirm the trend already observed in calibration validated in cross-validation; in fact, the predictive models show high sensitivity and good specificity for all classes except Class A and, to a lesser extent, Class D.

In *Table 4* the specificities of each model are reported with respect to each of the other categories.

Class	$Spec_{wrtA}$	$Spec_{wrtB}$	$Spec_{wrtC}$	$Spec_{wrtD}$	$Spec_{wrtE}$	$Spec_{wrtF}$	$Spec_{wrtG}$	$Spec_{wrtH}$
A	-	0.0	30.0	80.0	50.0	20.0	30.0	50.0
B	100.0	-	70.0	100.0	100.0	90.0	90.0	100.0
C	90.0	0.0	-	90.0	100.0	70.0	100.0	90.0
D	80.0	30.0	50.0	-	20.0	20.0	10.0	20.0
E	100.0	80.0	80.0	90.0	-	70.0	40.0	20.0
F	100.0	10.0	60.0	80.0	70.0	-	80.0	80.0
G	90.0	90.0	90.0	70.0	30.0	70.0	-	30.0
H	100.0	70.0	80.0	80.0	20.0	60.0	50.0	-

Table 4. Prediction of SIMCA model in the test set

This allows identifying which classes were confused with each other. The specificity with respect to another category represents the percentage of samples belonging to the other class that were correctly rejected. This means that if, after modelling a class, all samples belonging to another category are rejected, the specificity with respect to that modelled class will be 100.0%. Consequently, *Table 4* reveals which classes are most similar from the E-Eye perspective.

These results highlight the fact that the low accuracy provided by the Class A model is mainly due to the fact that it incorrectly accepts all samples belonging to Class B, 80.0% of Class F and 70.0% of individuals in Classes C and G. This result was expected for Class B as it is an accession of Mugnoli. The confusion with the other classes mentioned could be attributed to the fact that Mugnoli and classes C, G, and F have darker leaves and, at the same time, are the richest in incisions.

The possibility of modelling samples belonging to classes A, B and C as a single Mugnoli class was tested. This led to low sensitivity and specificity values in the prediction, 53.3% and 58.0% respectively, indicating that the variability between samples belonging to classes A, B and C is significant and that combining these three categories into a single class is not reasonable. To better examine the overlap of these three Mugnoli accessions, a PCA was performed on the colorgrams and, as shown in *Figure 18*, there is a strong overlap of these three classes.

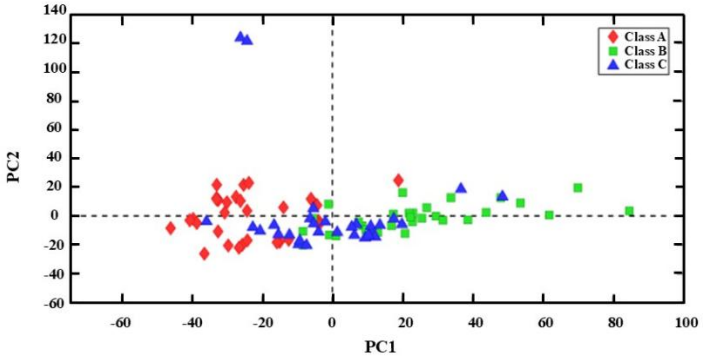


Figure 18. Plot of the PCA performed on the colorgrams of samples belonging to classes A, B and C

In particular, samples belonging to Class A have negative values for the first principal component (PC1); those belonging to Class B have purely positive values for PC1, while those belonging to Class C are distributed along this component. Only two samples belonging to Class C have highly positive values on the second principal component (PC2), but inspection of the T^2-Q plot did not allow them to be considered outliers and they were therefore maintained.

This analysis strategy demonstrated that combining MIA with the class-modelling approach allows for highly accurate classification of the different categories of plants considered. In fact, all class models correctly accepted a percentage greater than 70%, with the exception of one Mugnoli accession that was confused with the other two classes due to the strong hybridisation capacity typical of Brassicaceae.

2.2 IR analysis

The IR analysis, as already mentioned, was also performed on eight of the nine Brassicaceae cultivars grown and harvested in the experimental field due to the failure of the cultivation of Class Γ.

The infrared spectra were recorded using the PerkinElmer Spectrum Two™ FT-IR spectrophotometer. They were acquired in a spectral range from

400 cm^{-1} to 4000 cm^{-1} with an instrumental resolution of 4 cm^{-1}, and sixteen scans were averaged for each spectral replica.
The background was acquired with the crystal exposed to air. The ATR-FT-IR spectra were acquired after cleaning the leaves with absorbent paper to remove any soil residues. The diamond was cleaned using methanol and precision wipes after each sample analysis, and particular attention was paid to ensuring complete evaporation of the methanol before placing a new sample on the ATR device.

60 IR spectra were acquired for Class A, 58 IR spectra for Class B, and 45 IR spectra for each of Classes C, D, E, F, G, and H (*Figure 19*).

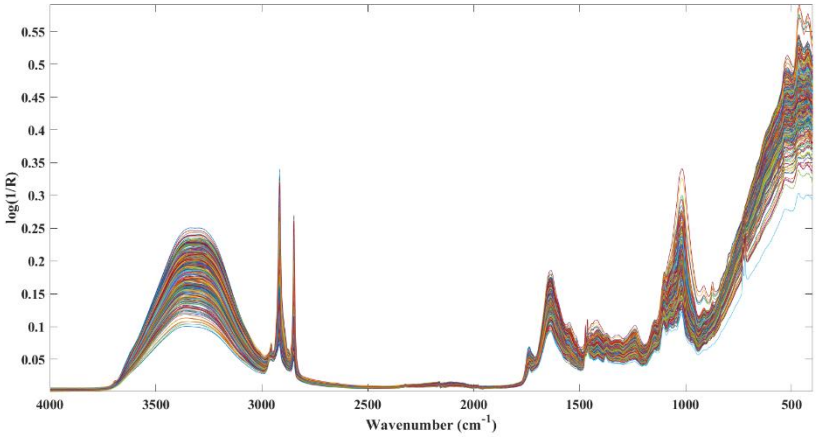


Figure 19. IR spectra acquired for all the analysed matrices

Figure 20 shows the average spectra for each of the classes considered.

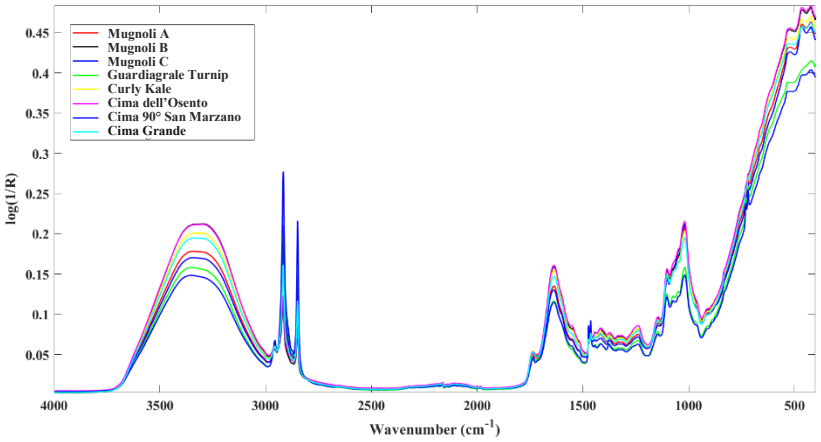


Figure 20. Average IR spectra acquired for each class

Since in the absence of spectrum processing it is not possible to make a direct comparison of intensities, it is possible to recognise some differences in the intensities of the peaks in the fingerprint zone from 1100 cm^{-1} to 500 cm^{-1}.

Given the complexity of the matrix and the strong influence of water on the spectroscopic signals, detailed interpretation of the profiles reported is complex. The profiles reported reflect the expected composition of the samples (water, proteins, sugars, carbohydrates, glucosinolates, crude fibre, phenolic compounds, chlorophylls and porphyrin carotenoids, vitamins (A, C and K), choline and betaine). The band centred at 3360 cm^{-1} indicates that one of the main constituents is water, attributed mainly to the stretching of the O-H bond, overlapping with that of proteins, the N-H bond of sugars as well as that of phenolic compounds and vitamin C, and the indole portion of glucosinolates[142].

The peaks between approximately 2920 cm^{-1} and 2860 cm^{-1} represent the symmetric and antisymmetric stretching vibrations of C-H, respectively. The signal at 1740 cm^{-1} is related to the C=O stretching of chlorophylls and pectins. The broad spectral band at 1640 cm^{-1} results from several combined effects, such as O-H bending, H-O-H bending of water, the Amide I band of

[142] Q. V. Vo, C. Trenerry, S. Rochfort, J. Wadeson, C. Leyton, and A. B. Hughes, "Synthesis and anti-inflammatory activity of indole glucosinolates," *Bioorg. Med. Chem.*, vol. 22, pp. 856–864, 2014.

proteins and the C=N sulphate portion, which produce strong IR bands between 1630 cm^{-1} and 1690 cm^{-1}.
The signal at 1515 cm^{-1} is attributed to amide II proteins. However, in the 1500-1200 cm^{-1} range, the absorptions are associated with the bending modes of cellulose CH_2 and various vibrational modes of pectins and chlorophylls. Signals in the 1200-950 cm^{-1} range can be found in signals linked to the S=O group of glucosinolates. Furthermore, signals at 1600 cm^{-1} and 1100 cm^{-1} indicate the presence of aromatic compounds, flavonoids and phenols.
The collected spectra were subjected to classification analysis using SIMCA and PLS-DA algorithms to assess whether it was possible to classify the Mugnoli by considering the samples belonging to classes A, B and C as a single class compared to the other ecotypes investigated (classes D, E, F, G and H), which were considered as four independent categories, one of which included classes G and H, considered to be commercial classes.
Subsequently, the classification was tested considering the Mugnoli as three independent classes to verify whether the variability between the three classes was sufficiently high to allow individual modelling.
Finally, using the PLS-DA algorithm, the possibility of discriminating between the three classes of Mugnoli considered as a single category was verified with respect to individuals belonging to the classes most commonly found on the market (Classes E, G and H).

Before proceeding with the classification analysis, the spectral data were preprocessed using different preprocessing methods: Mean Centring (MC), SNV followed by MC; D1 and D2, and combinations of multiple preprocessing methods such as SNV+D1+MC and SNV+D2+MC. For each classification problem addressed and for each algorithm applied, the different models were calculated according to the 7-fold cross-validation procedure.

The outcome of the internal validation was then used to define the optimal data preprocessing.

Table 5 contains the cross-validation efficiencies ($\%Eff_{CV}$) obtained from the analysis, together with the optimal number of PCs to be retained and the relative preprocessing referred to the Mugnoli classification (including classes A, B and C considered as a single class) compared to the other ecotypes investigated.

Preprocessing	MC	SNV	D1	D2	SNV+D1	SNV+D2
% Eff_{CV}	77.3	79.4	77.1	75.3	68.5	48.5
PCs	15	14	14	14	14	14

Table 5. SIMCA analysis comparing Mugnoli class (ABC) with the other ecotypes. Table show the cross-validation efficiencies (%Eff_{CV}) and the corresponding number of Principal Components (PCs) selected for each model

From the results shown in *Table 5* it can be seen that the models built on data simply subjected to simple MC and the models built on data after subjecting them to D1 provide the highest and most comparable efficiencies, and since the latter required fewer PCs, it was considered the optimal model (*Figure 21*).

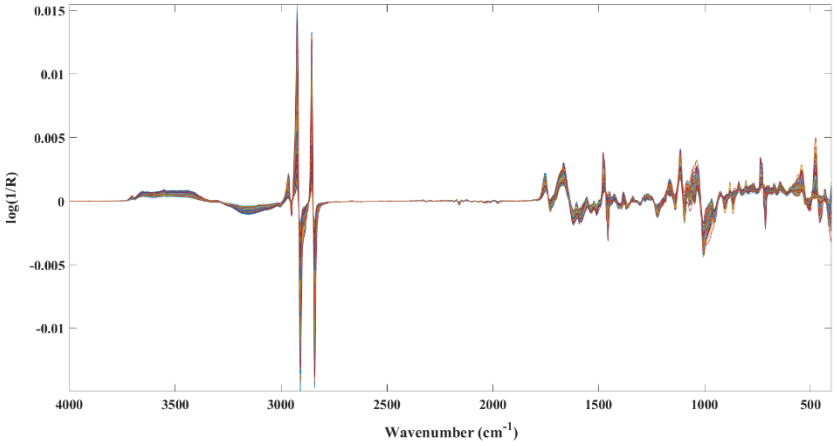


Figure 21. Spectra preprocessed with D1

This model was used to make predictions on the test set, providing a sensitivity of 82.0% and a specificity of 70.7%.

The results are not entirely satisfactory, which could be attributed to a possible high variability between accessions, making Mugnoli's class (ABC) less well defined and compromising the predictive capabilities of the model.

For this reason, the three classes of Mugnoli were modelled separately to assess whether this would lead to an improvement in predictive performance. The results of this classification are shown in *Table 6*.

Class	Preprocessing	MC	SNV	D1	D2	SNV+D1	SNV+D2
A	% Eff_{CV}	81.8	81.0	80.6	76.2	72.4	62.6
	PCs	8	8	6	8	6	9
B	% Eff_{CV}	79.3	78.1	83.3	82.0	83.8	70.4
	PCs	6	5	6	5	6	8
C	% Eff_{CV}	80.7	80.4	78.6	83.8	80.3	71.6
	PCs	5	5	6	6	6	6

Table 6. SIMCA analysis comparing Mugnoli classes (A, B and C) with the other ecotypes. Reported are the cross-validation efficiencies (%EffCV) and the corresponding number of Principal Components (PCs) selected for each model

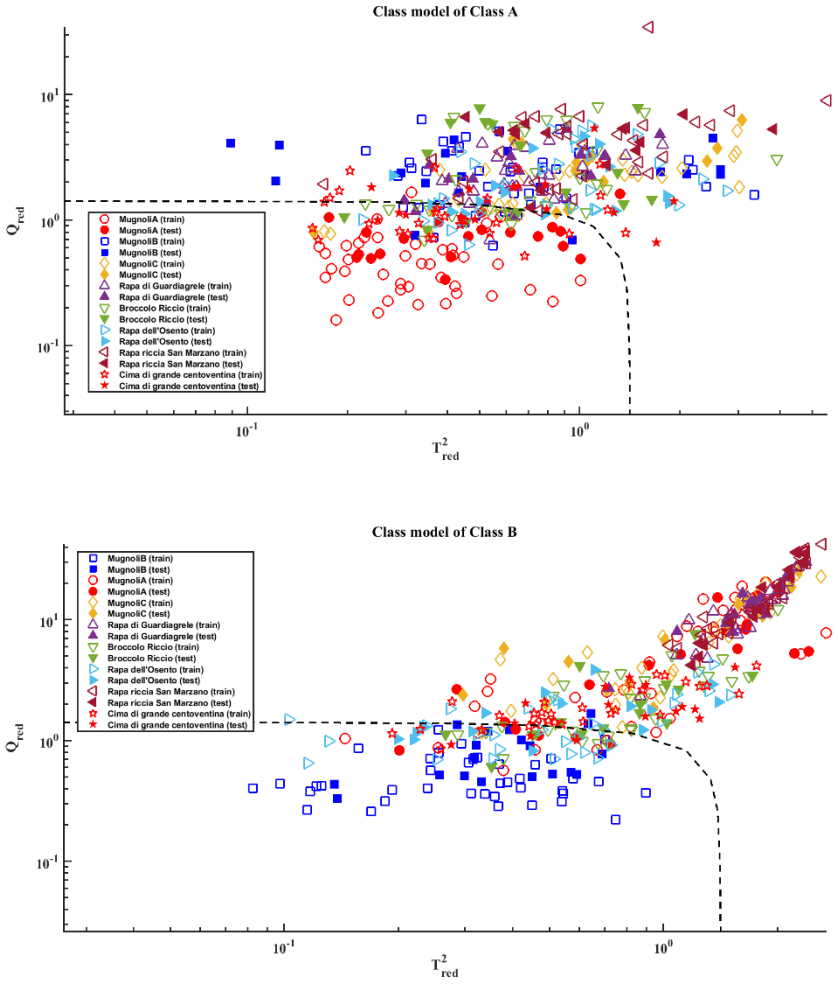
The data shown in *Table 6*, indicate the results of the models comparing pretreatment by pretreatment. An improvement can therefore be seen in this case by modelling the three Mugnoli accessions as three distinct classes. The best calibration model chosen for predicting the test set for Class A was the one built on data centred on the mean, for Class B the one modelled on signals preprocessed with SNV and D1, and for Class C the one on spectra preprocessed with second derivative. *Table 7* shows the sensitivity and specificity values of the above models.

Class	Preprocessing	Sensitivity	Specificity
A	MC	95.0	81.6
B	SNV+D1	89.5	82.7
C	D2	93.3	83.3

Table 7. Sensitivity and specificity values for the models selected with the preprocessing yielding the highest %Eff_{CV}

The results in *Table 7* show that the model performances are better than those obtained when the Mugnoli were modelled as a single class and are satisfactory overall. It is therefore clear that the different accessions should be considered as three independent entities, thus allowing for better predictive modelling. The cause of the significant differences shown between the three accessions, despite belonging to the same variety and being grown in the same experimental field, could be linked to the intense hybridisation processes characteristic of these plants.

Figure 22 shows the graphical representation of this result by displaying the T^2 *vs.* Q plots of the Class A, Class B, and Class C models.



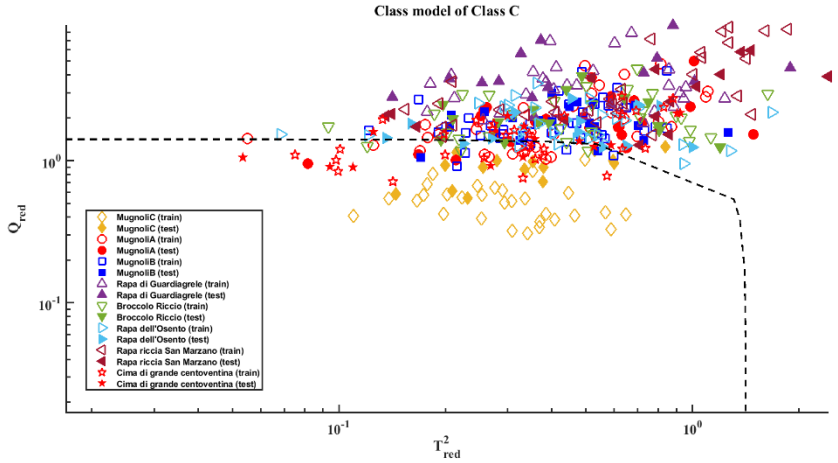


Figure 22. SIMCA results for Mugnoli's classes A, B and C reported as projections of the training set (full symbols) and test set (empty symbols) samples onto the T^2_{red} vs. Q_{red} space

In all figures, the full symbols represent training samples, while the empty symbols represent test samples. The black dotted line represents the acceptance threshold, meaning that samples below this line are accepted by the model and therefore predicted as belonging to that modelled class, while all others are rejected.

From the plot of class A model most of the samples belonging to this class fall within the acceptance zone, resulting in a high sensitivity value, while about 20% of test samples belonging to other categories, in particular B and H, were incorrectly accepted by the model.

As regards the Class B model, however, the incorrect attribution of test samples mainly concerned the Mugnoli belonging to Class A and the samples from Cima dell'Osento (Class F), which the model incorrectly accepted. Finally, the Class C model proved to be the most accurate in terms of prediction, despite the fact that most of the incorrectly accepted samples belong to Classes A and B.

Also in the case of classification using the PLS-DA algorithm, the spectral data were centred on the mean using mean centring and preprocessed using SNV, D1, D2, and combinations of preprocessing such as SNV+ D1 and SNV + D2 before model construction.

The calibration models were optimised through cross-validation and by selecting optimal parameters such as the best data pretreatment and the

optimal number of latent variables (LVs), based on the cross-validated correct classification rate (%CCRcv) for the Mugnoli class (ABC) or for classes A, B, and C of interest.

Initially, the Mugnoli class (ABC), comprising the three individual classes, was discriminated from all other ecotypes such as Classes D, E, F, G and H.

Table 8 shows the optimal number of LVs and the cross-validation correct classification rate (%CCRcv) of the Mugnoli class (ABC) for each pretreatment.

Preprocessing	**MC**	**SNV**	**D1**	**D2**	**SNV+D1**	**SNV+D2**
%CCRcv	77.0	82.3	81.4	89.4	71.7	86.7
LVs	16	17	17	15	13	17

Table 8. Results of the PLS-DA classification in which class ABC was classified with respect to the other classes. The table shows the best correct classification rate values in cross-validation and the number of LVs.

From the values shown in *Table 7*, it can be deduced that the best model is the one obtained by preprocessing the data with second derivatives.

Table 9 displays the percentage prediction values (%CCRpred) of the previously selected optimal model applied to the test samples.

Pretreatment	**LVs**	**%CCRpred ClassABC**	**%CCRpred ClassD**	**%CCRpred ClassE**	**%CCRpred ClassF**	**%CCRpred ClassG**	**%CCRpred ClassH**
D2	15	84.0	100.0	73.3	80.0	93.3	73.3

Table 9. Results obtained in prediction on the test set whit PLS-DA algorithm, where Mugnoli class (ABC) was classified against all other ecotypes

From the values reported, it can be seen that the model built on preprocessed data with second derivative provides a %CCRpred value of 84.0%.

Subsequently, in an attempt to improve predictive capabilities, Mugnoli's three classes were considered as three different classes, namely Class A, Class B and Class C, and classified with respect to all other ecotypes, which were also considered independent. The results of this test are shown in *Table 10*.

Preprocessing	LVs	%CCR$_{CV}$ ClassA	%CCR$_{CV}$ ClassB	%CCR$_{CV}$ ClassC	%CCR$_{CV}$ ClassD	%CCR$_{CV}$ ClassE	%CCR$_{CV}$ ClassF	%CCR$_{CV}$ ClassG	%CCR$_{CV}$ ClassH	Mean %CCR$_{CV}$
MC	17	72.5	92.3	83.3	96.7	80.0	90.0	90.0	89.7	86.8
SNV	15	75.0	94.9	76.7	93.3	83.3	80.0	83.3	79.3	83.2
D1	15	72.5	97.4	86.7	96.7	73.3	90.0	90.0	86.2	86.6
D2	18	90.0	89.7	90.0	96.7	83.3	93.3	96.7	86.2	90.7
SNV+D1	15	82.5	84.6	83.3	96.7	86.7	90.0	96.7	86.2	88.3
SNV+D2	16	77.5	48.6	86.7	96.7	80.0	80.0	93.3	79.3	84.8

Table 10. Results of the PLS-D classification in which class A, B, and C was classified in relation to the other classes. The table shows the best correct classification rate values in cross-validation and the number of LVs for each pretreatment

In *Table 10* the values of the correct classification rate in cross-validation (%CCRcv) and the optimal number of LVs for the models calculated using the various preprocessing methods are reported. Once again, the best model in cross-validation is the one provided by the data preprocessed with second derivative, as it has the highest percentage of correct classification. For this reason, it was used to predict the samples belonging to the test set, as shown in *Table 11*. The latter reports the prediction results.

Preprocessing	LVs	%CCR$_{pred}$ ClassA	%CCR$_{pred}$ ClassB	%CCR$_{pred}$ ClassC	%CCRpred ClassD	%CCR$_{pred}$ ClassE	%CCR$_{pred}$ ClassF	%CCR$_{pred}$ ClassG	%CCR$_{pred}$ ClassH
D2	18	100.0	100.0	93.3	100.0	93.3	93.3	100.0	80.0

Table 11. Results obtained in predictions where Mugnoli classes were considered separately (A, B and C) and classified in relation to other varieties

The model allows for the correct classification of Class A and Class B test samples with a classification rate of 100.0% and 93.3% for Class C, of which only one sample was incorrectly assigned to Class H. According to the model, a single Class F sample was confused with Class C, while the two commercial Class H samples were assigned to Class F and one sample to Class E.

These high %CCRpred values indicate that the spectral data are highly informative, to the extent that they can discriminate between varieties and between different Mugnoli accessions.

In accordance with the results obtained by applying the SIMCA algorithm, the %CCRcv and %CCRpred values are better when the three classes of Mugnoli are considered as three distinct classes (Class A, Class B and Class

C). This confirms that the local variety of Mugnoli from Pettorano sul Gizio (Abruzzo, central Italy) are hybrid species, therefore the different accessions could be influenced by other Brassicaceae grown nearby.

Finally, the PLS-DA algorithm was used to address a more specific classification problem in order to assess the possibility of distinguishing Mugnoli di Pettorano sul Gizio (Abruzzo, central Italy) from the most common commercial Brassicaceae such as Cavolo Riccio (Class E), Cima 90° San Marzano (G) and Cima Grande (H). Therefore, only Classes A, B, C, E, G and H were taken into consideration.

Table 12 shows the results relating to the construction of the calibration models validated according to cross-validation.

Preprocessing	LVs	%CCR$_{CV}$ Class A	%CCR$_{CV}$ Class B	%CCR$_{CV}$ Class C	%CCR$_{CV}$ Class E	%CCR$_{CV}$ ClassG	%CCR$_{CV}$ ClassH	Mean %CCR$_{CV}$
MC	18	82.5	94.9	90.0	100.0	100.0	93.1	91.8
SNV	15	85.0	94.9	83.3	93.3	93.3	86.2	86.6
D1	15	85.0	94.9	96.7	96.7	96.7	93.1	93.3
D2	17	90.0	94.9	90.0	90.0	96.7	100.0	93.6
SNV+D1	15	82.5	92.3	96.7	96.7	100.0	96.6	92.5
SNV+D2	17	95.0	92.3	93.3	93.3	100.0	89.76	93.4

Table 12. Results obtained in calibration by PLS-DA models to classified class A, B, C, E, G, H

The correct classification mean rate in cross-validation (Mean %CCRcv) value is used to compare different types of preprocessing in order to obtain an overall assessment and understand which pretreatment produces the most accurate classification on average.

In this case too, the second derivative proved to be the best preprocessing method, validated using the external test set as shown in *Table 13*.

Preprocessing	LVs	%CCR$_{pred}$ ClassA	%CCR$_{pred}$ ClassB	%CCR$_{pred}$ ClassC	%CCRpred ClassE	%CCR$_{pred}$ ClassG	%CCR$_{pred}$ ClassH
D2	17	100.0	100.0	93.3	100.0	93.3	100.0

Table 13. Prediction results on the test samples using D2 preprocessing

This classification problem was solved by classifying Classes A, B classes with a sensitivity of 100.0% and Class C with 93.3%, for which only one sample was assigned to Class A. The other Brassicaceae varieties most commonly found on the market were classified with a correct classification rate in prediction of 100.0% for curly kale (Class E), 93.3% for Cima 90° San Marzano (Class G), according to which a single sample was confused with Class A, and 100.0% for Class H referring to the Cima Grande variety.

Applying linear discriminant analysis (LDA) to the values obtained from this PLS-DA model yields the plotted results shown in *Figure 23* e *Figure 24*.

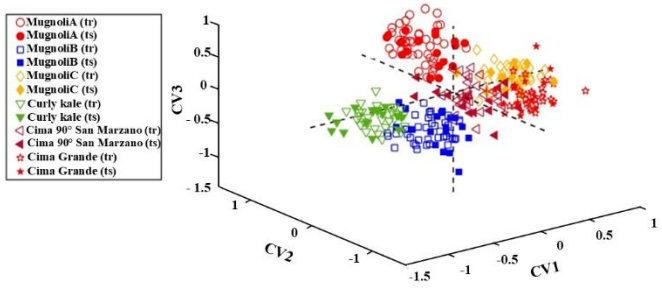


Figure 23. PLS-DA plot results obtained projection of the first three Canonical Variates (CV1, CV2 and CV3). Samples are distinguished according to the variety and the training (full symbol) and test (empty symbol) sets

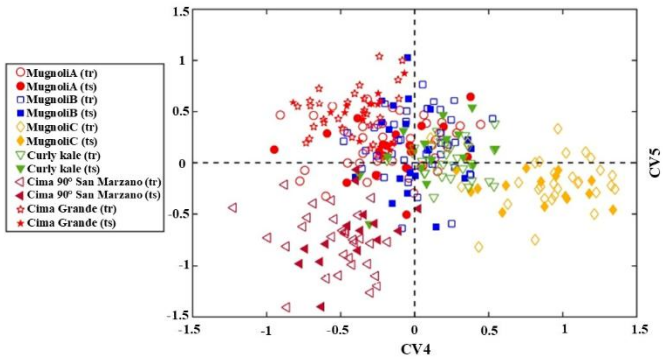


Figure 24. PLS-DA plot results obtained projection of the last two Canonical Variates (CV4 and CV5). Samples are distinguished according to the variety and the training (full symbol) and test (empty symbol) sets

LDA is a supervised classification technique that projects the data on new axes, called canonical variables (CV), chosen to maximise the separation between predefined classes. Unlike unsupervised exploratory methods, LDA directly takes class membership into account, allowing for a clearer visualisation of the differences between groups. In this case, its application is advantageous because it allows the positions of Mugnoli accessions to be highlighted in relation to commercial Brassicaceae. This provides an additional level of interpretation, facilitating the appreciation of overlaps (Class B and Cavolo Riccio) and distinctions (Class B compared to Classes A and C) that might not be evident from PLS-DA plots alone.

In fact, it can be seen that the Class B samples, which partly overlap with Cavolo Riccio in *Figure 23*, are different from the Mugnoli samples in Classes A and C, which instead have positive CV1 scores, as do Cima Grande and Cima 90° San Marzano.

In *Figure 24*, on the other hand, the dispersions of the Cima Grande, Cima 90° San Marzano and Mugnoli C classes are shown, which are distinct from each other. Meanwhile, Mugnoli A is very close to Cima Grande and Mugnoli C, which accounts for the few classification errors for the last model considered.

To identify the variables that contribute most to the separation of classes, Variable Influence Projection (VIP) analysis[143] was applied to this model, as shown in *Figure 25* below:

[143] S. Wold, E. Johansson, and M. Cocchi, "PLS-Partial least-squares projections to latent structures," in *3D QSAR in Drug Design*, Dordrecht, The Netherlands: Kluwer ESCOM Science Publisher, 1993, pp. 523–550.

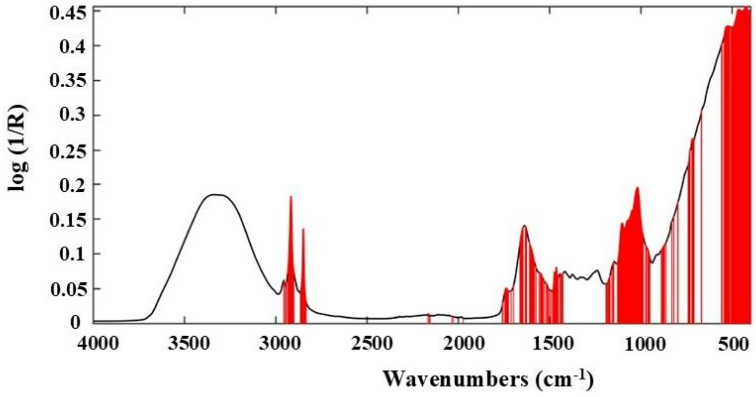


Figure 25. VIP analysis

This approach allows VIP scores to be calculated for each variable and indicates their importance in the model's discriminating capacity. In the figure, the variables showing a VIP index greater than 1 are highlighted in red, while the black solid line represents the average spectrum. VIP analysis is useful in interpreting the results in terms of the most significant variables, considering both their relevance in explaining the variance in space and the correlation between the responses. Since, by construction, the average VIP value is equal to 1, this can be set as a threshold. In this way, spectral variables with a VIP index greater than 1 will be those that most influence the model and are therefore most relevant in the classification process.

It can be seen that the peaks at approximately 2900 cm^{-1} and at approximately 2850 cm^{-1}, attributable to the symmetric and antisymmetric stretching vibrations of C-H, are significant, as is the contribution of the interval between approximately 1700 cm^{-1} and 1500 cm^{-1} associated with the presence of chlorophylls and pectins, the Amide I band of proteins and the C=N-sulphate group. Finally, the presence of significant variables from 1200 cm^{-1} to 900 cm^{-1} highlights the importance of aromatic compounds, flavonoids and phenols in the classification of these categories under study.

This study focused purely on the characterisation and classification of the three accessions of Mugnoli di Pettorano sul Gizio, typical of the Abruzzo region compared to other local and commercial cultivars. All these varieties were grown under the same soil and climate conditions to highlight their

intrinsic differences. It therefore emerged that considering the samples from the three different accessions as three different classes improves the predictive power of the models.

The data were processed using exploratory and classification analyses. The former highlighted similarities and differences between the various populations studied, while the latter showed high sensitivity in predicting the external set. Although the plants studied belong to different local populations, they show strong similarities. This is mainly due to the ability of Brassicaceae to hybridise. Despite this peculiarity, the combination of non-destructive techniques such as E-Eye and IR with SIMCA allowed the samples to be classified with excellent accuracy.

The study highlights and demonstrates the importance and applicability of this specific non-destructive method as a support for the management and conservation of genetic resources linked to local varieties.

The rapid, non-destructive and green IR technique has made it possible to obtain results quickly while keeping the analysed sample intact.

These chemometric algorithms have made it possible to classify the various classes of samples with good efficiency[144] [145], enriching the study conducted so far on landraces, which are interesting for their rich chemical composition, nutritional value and genetic heritage[146].

3. Analysis of extracts

The focus of the first part of this analysis was to identify the optimal extraction conditions, using chromatographic analysis, in order to evaluate the bioactivity of the plant species. To do this, commercial turnip tops from the same batch were taken into consideration. The leaves were cleaned with absorbent paper to remove any soil residues. They were stored at -20°C to

---

[144] A. Biancolillo, F. Marini, C. Ruckebusch, and R. Vitale, "Chemometric strategies for spectroscopy-based food authentication," *Appl. Sci.*, vol. 10, p. 6544, 2020.

[145] M. Cocchi, A. Biancolillo, and F. Marini, "Chemometric methods for classification and feature selection," in *Data Analysis for Omic Sciences: Methods and Applications*, J. Jaumot, C. Bedia, and R. Tauler, Eds., *Comprehensive Analytical Chemistry*, vol. 82. Amsterdam, The Netherlands: Elsevier, 2018, pp. 265–299.

[146] C. Cartea, M. Francisco, P. Soengas, and P. Velasco, "Functional ingredients from Brassicaceae species: Overview and perspectives," *Int. J. Mol. Sci.*, vol. 21, no. 6, p. 1998, 2020.

ensure that they received the same treatment as those grown in the same experimental field.

Two methods of drying were tested, one using a glass chamber vacuum dryer (BÜCHI Labortechnik AG, Switzerland), which allows the temperature to be adjusted and the process times to be monitored, ensuring gentle and controlled conditions for vacuum drying or solvent removal. This instrument consists of a heated borosilicate glass chamber. The cylindrical shape of the glass allows for even heat distribution and the possibility of visually monitoring the sample during the process and evaluating its effect. It allows the removal of even high boiling point solvents at low temperatures, safeguarding the sample by preventing the thermal or oxidative degradation of sensitive compounds. It consists of a temperature control that allows the process time to be adjusted and monitored. These features make it suitable for drying natural extracts or compounds that require delicate treatment conditions.

The second method is drying using a ventilated oven designed for drying and heat treatment of samples under controlled temperature conditions. Internal ventilation ensures uniform heat distribution within the chamber, reducing temperature variations to guarantee uniform thermal conditions and improve reproducibility. It is equipped with a thermally insulated internal steel chamber and electric heating elements that allow temperatures of up to 250°C to be reached and maintained thanks to a thermoregulator that controls the temperature.

The drying tests were carried out at 50°C, a temperature sufficient to avoid compromising the integrity of the sample and to allow complete evaporation of the water in a gradual manner.

The weights of the leaves were monitored hourly and, for the samples dried in the oven, weight stabilisation was achieved after 2 hours, while with the vacuum dryer, weight stabilisation was achieved after 6 hours. At the end of these tests, oven drying was identified as the most suitable method given the shorter times and the possibility of treating larger quantities of leaves in a single session.

Preliminary solvent evaporation tests were carried out to verify that the quantity of extracting mixture used did not suffer significant losses due to evaporation during the process. Tests were carried out to determine the sample mass that could be effectively wetted and kept in uniform and constant agitation in 10 mL of solvent.

This limit was identified as 300 mg. The volume of 10 mL was chosen in relation to the use of 20 mL calibrated flasks, in which the solution remained confined to the widest part, preventing it from rising into the neck. This ensured homogeneous agitation and, at the same time, allowed the geometry of the narrow neck of the flask to act as a natural barrier, reducing any evaporation of the solvent. The flask was also capped and covered with aluminium foil to protect it and prevent photodegradation of light-sensitive compounds.

Subsequent extractions were performed in a 10 mL flask using 5 mL of extracting solution for a maximum of 150 mg of sample. A D-Optimal design was planned in which the effects of four variables at three levels were investigated.

D-Optimal design allows to reduce the number of experiments by evaluating the Minimum Inflactor Factor and the log(Normalised determinant). These two indices are considered as criteria for selecting the minimum number of experiments that allows for very low uncertainty in parameter estimation. The first allows to evaluate how much less accurate the parameter estimates are, while the second allows to identify the minimum number of experiments for which the uncertainty in the parameter estimates is minimal. These two parameters coincide and have made it possible to reduce the number of experiments to 26, which in a full factorial design would have been 81.

The levels investigated for the Temperature (T) variable are: 20°C, 40°C and 60°C. The time interval considered ranges from 15 minutes to 45 minutes, while the extraction solution consists of 100% Milli-Q water, water produced by Millipore (Bedford, MA, USA), a mixture of water and ethanol (W:E) in a ratio of 80:20 and another in a ratio of 60:40, as shown in *Table 14*

Quantity (Q) mg	**Temperature (T) °C**	**Time (t) min**	**Extraction mixture (W:E) Water:Ethanol**	**Levels in the DoE**
30	20	15	100:00	-1
90	40	30	80:20	0
150	60	45	60:40	1

Table 14. Variables and levels investigated in the D-Optimal design

By inserting replicates as close as possible to the extremes of the design, a total of 34 replicates were achieved, 26 planned by the DoE and performed randomly, with the addition of 8 replicates as reported in *Table 15* and *Figure 26*.

Sample	Q(mg)	T (°C)	t (min)	W:E
23	150	20	45	60:40
20	150	20	30	60:40
9 R	150	60	45	100:0
26 R	150	60	45	60:40
4	150	60	15	100:0
24	30	40	45	60:40
8 R	30	60	45	100:0
6 R	30	20	45	100:0
1 R	30	20	15	100:0
21	30	60	30	60:40
17	150	40	15	60:40
7	150	20	45	100:0
3	30	60	15	100:0
22	30	20	45	60:40
25	90	60	45	60:40
12	150	60	30	80:20
16	90	20	15	80:20
5	90	40	30	100:0
10	150	20	15	80:20
14	30	60	45	80:20
2 R	150	20	15	100:0
19 R	150	60	15	60:40
18	30	60	15	60:40
11	30	40	15	80:20
13	90	20	45	80:20
15 R	30	20	15	60:40

Table 15. List of experiments carried out and related experimental conditions. R indicates the replication carried out on the sample.

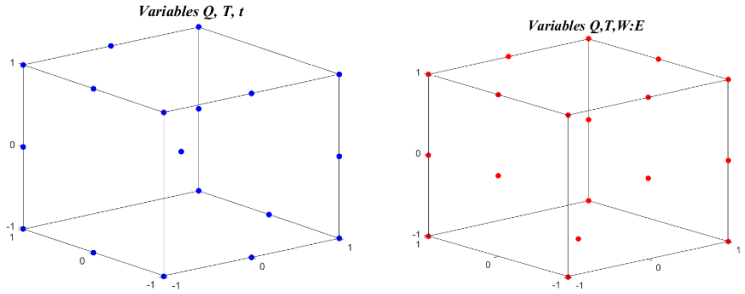


Figure 26. Position of experiments within the experimental domain as a function of variables Q (1), T (2), t (3) and W:E (4)

The leaves were dried for 2 hours at 50°C, ground for 2 minutes at 5000 rpm and stored in a flat-bottomed plastic vial with a screw cap, sealed with Parafilm, at 4°C.

The extracts were prepared according to the experimental design, sieving the matrix with a 1 mm sieve and placing it in a flask in a thermostatic water recirculation bath, protecting the extract from light and keeping it under agitation by keeping the flask capped.

Once the extraction was complete, the mixture was transferred into a plastic vial and centrifuged for 5 minutes at 4020 rpm. The supernatant was then filtered through nylon filters with an average membrane pore diameter of 0.22 µm. The filtered extract was stored in a parafilm-covered plastic vial at -4 °C for 24 hours. Subsequently, the extract was brought back to room temperature and analysed by HPLC-DAD and UV-Vis, and the DPPH assay was performed.

For HPLC-DAD analysis, the sample was diluted 1:2 before injection.

The chromatographic run was performed in gradient mode, as shown in *Table 16*

Time	Flow	%A	%B
	1.00	95.0	5.0
37.00	1.00	5.0	95.0
40.00	1.00	5.0	95.0
45.00	1.00	95.0	5.0
57.00	1.00	95.0	5.0

Table 16. Gradient elution program used for the chromatographic analysis

With a flow rate of 1 mL/min, the gradient started with an excess of water acidified with 0.01% phosphoric acid (Solution A), passing through 95% acetonitrile (Solution B) and then returning after 45 minutes to the starting condition and remaining there for 12 minutes to ensure that the initial conditions of the column were restored before proceeding with the analysis of the next sample.

Below, *Figure 27* shows an example of a chromatogram obtained following the chromatographic analysis of the samples, extracted at 340 nm:

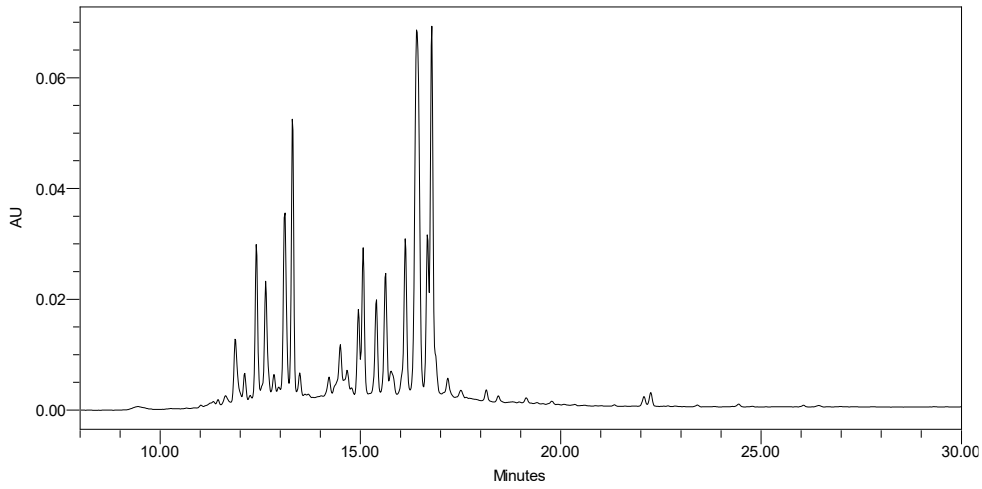


Figure 27. Example of a chromatogram obtained by HPLC analysis of the extracts

Once the analyses for the 12 replicate samples were completed, the values of the areas of all the peaks present in the chromatogram were collected, each extracted near the maximum absorption of each chromatographic peak.

A matrix was constructed in which the rows corresponded to the samples analysed and the columns to the variables considered (peak areas). On this matrix, an exploratory analysis was performed using PCA to evaluate the influence of the variables on the samples and to highlight the differences in the relative quantities of the compounds present in the samples on the autoscaled values.

Figure 28 shows the graph of the scores where the first principal component explains 72.1% of the variance while PC2 explains 12.0%.

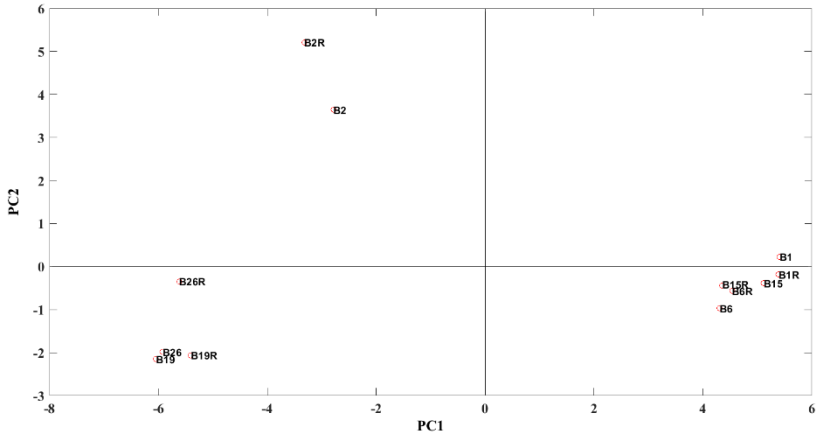


Figure 28. Scores plot

From this plot, three clusters of samples can be observed. Comparing it with the loadings graph (which shows the variables, in other words the chromatographic areas representing the compounds contained in the extracts) shown in *Figure 29,* in which the variables have been numbered according to their elution order, we can see sets of areas that differ from each other in terms of their different weights along the two components.

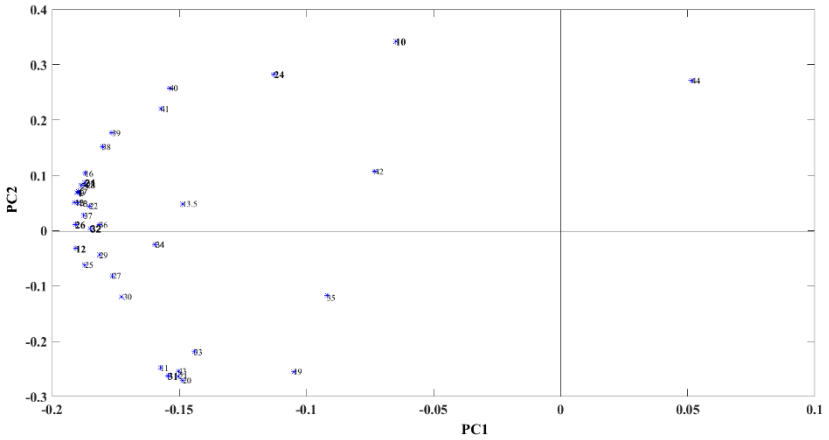


Figure 29. Loadings plot in which the variables of interest are highlighted in bold

It is possible to note a set of variables at the bottom left of the graph that have negative loadings on both PC1 and PC2, others such as 12 and 26 on the left that have no component on PC2 but high weight along PC1, and others such as 10 and 24 that have high loadings on PC2, which are responsible for the separation of sample B2 from all other samples.

From this preliminary exploratory analysis, the areas that best represented the sample clusters were evaluated, and those that represented the trend of all the interrelated areas were sampled, as highlighted in bold in *Figure 29*. Among the 37 total areas, 7 were identified: area 10 for the set of variables that has positive loadings for PC2; 12, 26 and 32 for the variables with loadings of 0 on PC2 but very negative on PC1; and finally areas 2, 24 and 31 for the group of variables with negative loadings on both components.

A new matrix containing the 32 total samples and the seven variables was constructed by re-running a new PCA, which showed a trend consistent with the previous one, as can be seen from the scores plot and loadings plot graphs shown in *Figure 30* and *Figure 31* respectively.

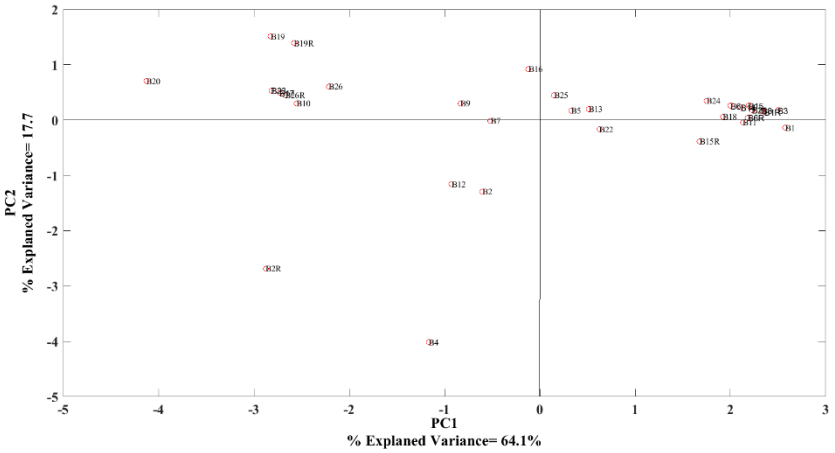


Figure 30. Scores plot

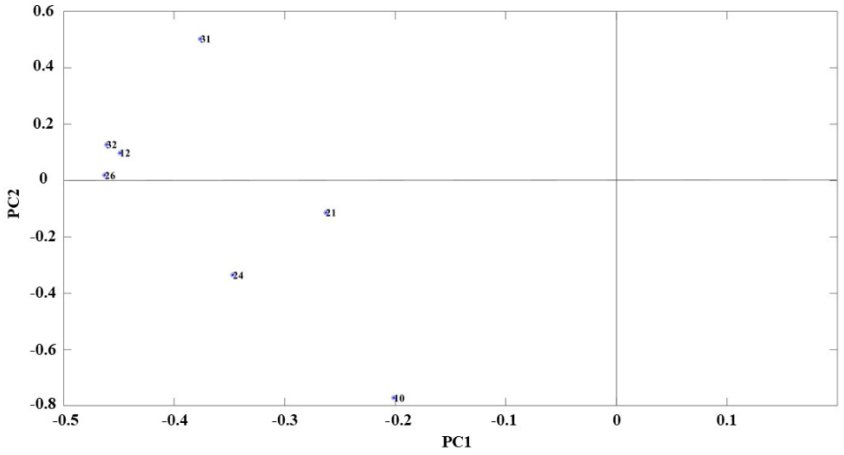


Figure 31. Loadings plot

The area of variable 10 has no weight on the first principal component, unlike the second, on which it has a higher weight. Samples 19, 20 and 26, which have negative scores on PC1, have a greater number of areas of all variables with negative loadings than samples 1, 15 and 16, which have positive scores. Looking at the scores plot (*Figure 30*) it can be said that among the extraction conditions of extracts B19 and B26, as well as between

B1 and B6, the only factor that changes is time, and since they do not show large differences in scores on PC1, it can be concluded that time is highly likely to be an insignificant and irrelevant factor.

The coefficient plots on PC1 shown in *Figure 32* confirm this observation. In fact, the main variability of the data (92.4% of the total explained variance) is mainly influenced by the sample quantity (*Q*), which has the highest and statistically significant coefficient, and by the *W:E* ratio of the extraction mixture, while *t* contributes only secondarily. After eliminating the insignificant coefficients (*Figure 33*), the simplified model maintains a high descriptive capacity, reflected in 93.8% of the explained variance, highlighting how the distribution of samples along PC1 is explained mainly by the sample quantity and the composition of the extraction mixture, with a marginal contribution from the other factors.

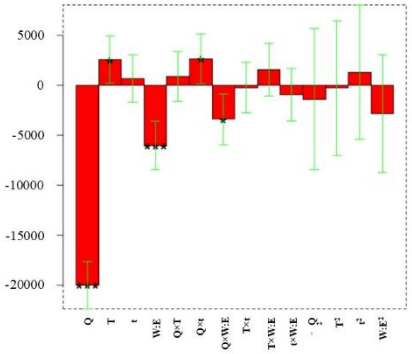


Figure 32. Coefficients scores on PC1

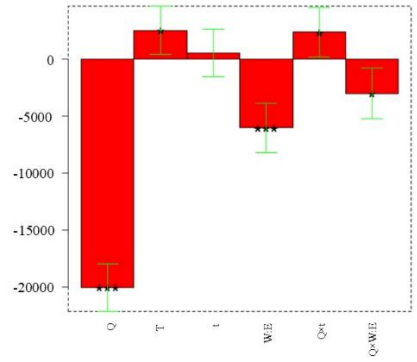


Figure 33. Coefficients scores on PC1 after deleting the non-significant coefficients

The response surface plot reported in *Figure 34* represents the variation in scores on the first principal component as factors *Q* and *W:E* vary, while the others remain constant (temperature fixed at level 1 (60 °C) and time at level 0 (30 min)). The surfaces show a clear negative effect of the sample quantity (Q) on the PC1 scores, in agreement with what is shown in the coefficient plot (*Figure 33*), while the extraction mixture (W:E) has a more moderate influence. This representation allows the combined effect of the significant factors to be visualised and confirms the trends identified in the coefficient analysis.

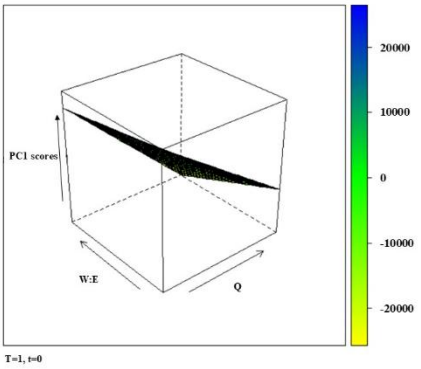


Figure 34. Response surface plot on PC1

This will be confirmed by the construction of the models.

Based on the points of the experimental design, a polynomial model was generated in which the area of the substance identified as representative of the cluster is related to the factors evaluated in the DoE.

The plot given in *Figure 35* show how area 10, attributed to a luteolin derivative (spectrum extracted at 250 nm with an elution time of 12.27 min), showed a very low percentage of explained variance equal to 48.2% and for this reason the model was not considered significant.

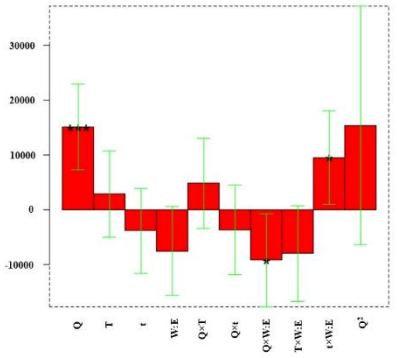


Figure 35. Coefficients plot of area 10

Similarly, area 21, attributed to chlorogenic acid, shows a poor model, as shown in *Figure 36* explaining a variance of 31.2%.

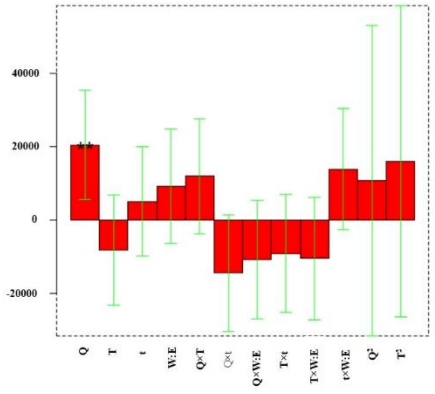


Figure 36. Coefficients plot of area 21

As well as for area 24, attributed to an apigenin derivative whose spectrum was extracted at 230 nm, the model showed an explained variance of 37.8%, a value to be considered unsatisfactory (*Figure 37*).

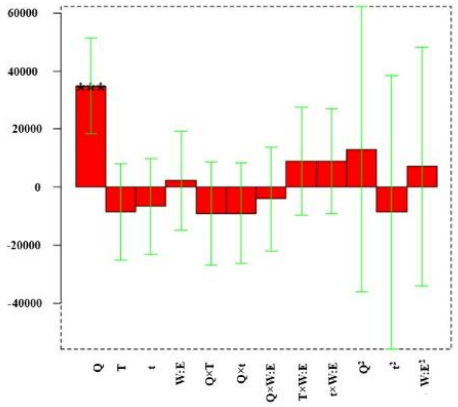


Figure 37. Coefficients plot of area 24

Area 12 is an area related to an apigenin derivative compound, which has high loadings on PC1 and, in fact, the model explains 94.1% of the variance. From the construction of the response surfaces (*Figure 38b*) it can be seen that the optimum is achieved with 150 mg of sample extracted at 40°C for 45 minutes with an extraction mixture consisting of water and ethanol in a ratio of 60:40.

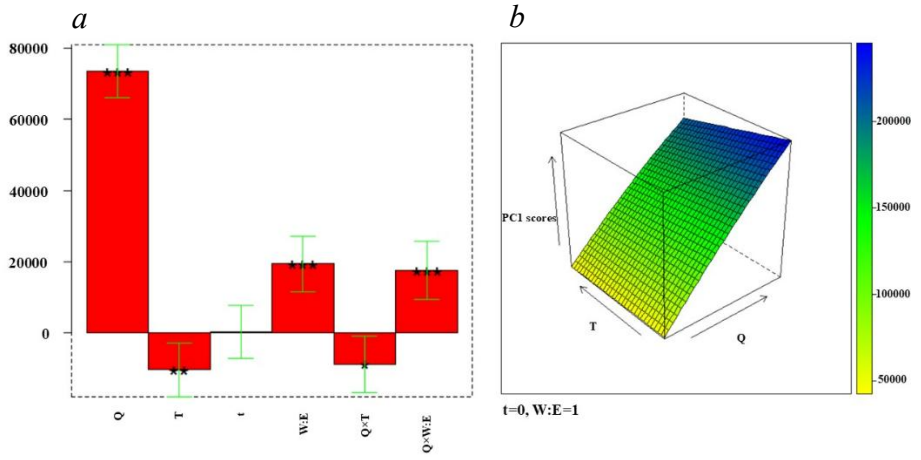


Figure 38. Coefficients plot and response surface of area 12, respectively *a* and *b* plot

The spectrum of this compound was identified at 12.5 min and extracted at a wavelength of 340 nm.

Area 26 (*Figure 39*), given by another apigenin derivative compound, whose spectrum was extracted at 290 nm, is the one that shows the highest loadings. In fact, the model has a percentage of explained variance of 96.6% due not only to the weight of this single variable, which is the one with the highest loadings, but also to the overall ability of the model to describe the distribution of the experimental data observed. It was eluted with a retention time of 15.6 min.

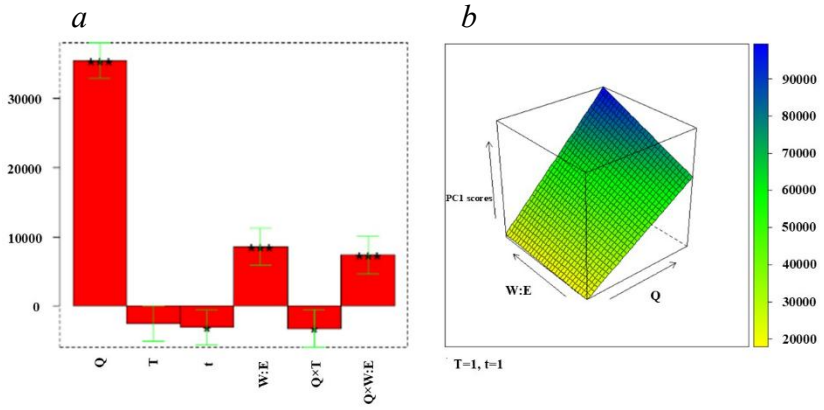


Figure 39. Coefficients plot and response surface of area 26, respectively *a* and *b* plot

The optimum extraction (*Figure 39b*) results in a smaller quantity (90 mg) but at a higher T (60°C) compared to the model identified in area 12. Therefore, 90 mg extracted at 60°C for 45 minutes with a W:E mixture in a 60:40 ratio, coinciding with the extraction conditions of sample B25.

The model for area 31, whose spectrum was extracted at 270 nm and attributed to a luteolin derivative, with a retention time of 16.86 min, showed a percentage of explained variance of 77.4% and, from the response surfaces, the identified extraction optimum corresponds to 90 mg at a temperature of 60 °C for 45 minutes with a 60:40 mixture of H_2O:EtOH. In this case, the optimum is the same as the model constructed on the basis of area 26 and again coincides with the extraction conditions of sample B25 (*Figure 40*).

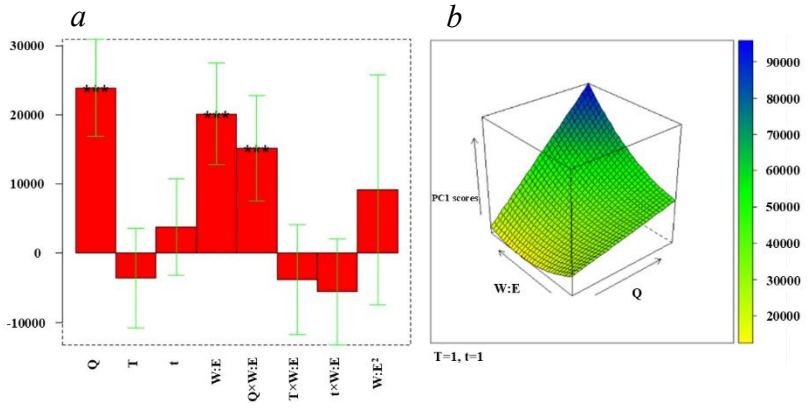


Figure 40. Coefficients plot and response surface of area 31, respectively *a* and *b* plot

For the luteolin compound corresponding to area 32 (*Figure 41*), extracted at 380 nm at 16.9 min, the model had an explained variance of 91.9%, and the same considerations made for the model of area 31 apply.

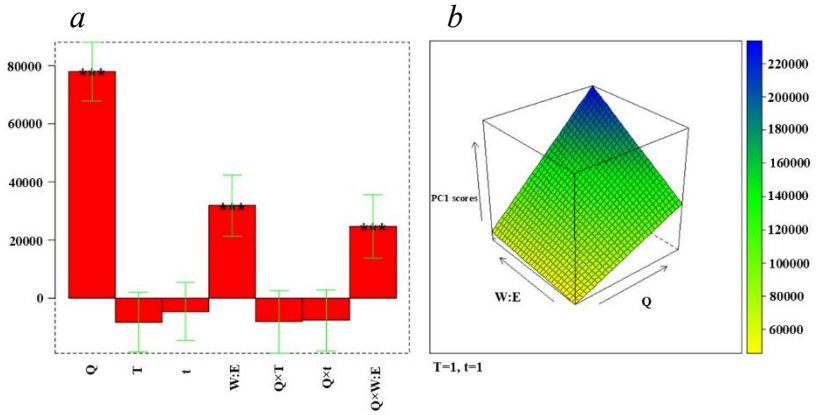


Figure 41. Coefficients plot and response surface of area 32, respectively *a* and *b* plot

In conclusion, the areas that appear to be most influenced by the factors studied in the DoE are mainly luteolin and its derivatives (areas 31 and 32) as well as apigenin derivatives (areas 12 and 26), as they have the greatest

influence on the separation of sample clusters along the first principal component. When the model obtained for a given variable is not significant, it indicates that the area relating to that chemical species does not depend significantly on the factors considered (Q, T, t and W:E). In other words, the extraction of this compound is not influenced by the experimental conditions studied, but could be determined by other factors not included in the model.

The construction of the models confirmed what had already been suggested by the PCA, namely that quantity and extraction mixture are the factors that most influence extraction. Furthermore, the optimum was identified for the combination of factors involving the extraction of 90 mg of Brassicaceae extracted at 60 °C for 45 minutes with an extraction mixture consisting of water and ethanol in a 60:40 ratio, coinciding with the extraction conditions of sample B25 (coordinates 011 in the DoE).

Once the optimal extraction conditions had been identified, the leaves of the native samples stored at -20°C, i.e. those of Class A, Class D and Class E, were dried for 3 hours at 50°C. The samples stored at -20°C required a longer drying time due to the amount of water they contained. Subsequently, the extracts were prepared under the optimal conditions identified in the DoE. HPLC-DAD analyses were performed on the extracts diluted 1:2, and based on these preliminary analyses, it was necessary to modify the elution gradient of the mobile phase to improve peak resolution.

Maintaining the flow rate at 1 mL/min, the elution program was refined as shown in *Table 17*

Time	Flow	%A	%B
	1.00	98.0	2.0
45.00	1.00	10.0	90.0
48.00	1.00	10.0	90.0
58.00	1.00	98.0	2.0
68.00	1.00	98.0	2.0

Table 17. Elution program of the mobile phase

Starting from an initial condition that included 98% solution A and 2% solution B, followed by a 45 minute gradient until reaching 10% solution A and 90% solution B, which was able to separate the polyphenols more

efficiently, followed by a 3-minute isocratic phase, then returning to the initial conditions in ten minutes. An additional 10-minutes of isocratic elution were added to restore the correct conditions of the column before proceeding with the analysis of the next sample. The total analysis time was 68 minutes per sample.

Simultaneously with the HPLC-DAD analysis, the extracts were subjected to UV-Vis analysis by diluting the extract 50-fold and considering as a blank a solution consisting of the extraction mixture (60:40 W:E) also diluted 50 -fold in water, for a total of 8 samples analysed as shown in *Figure 42.*

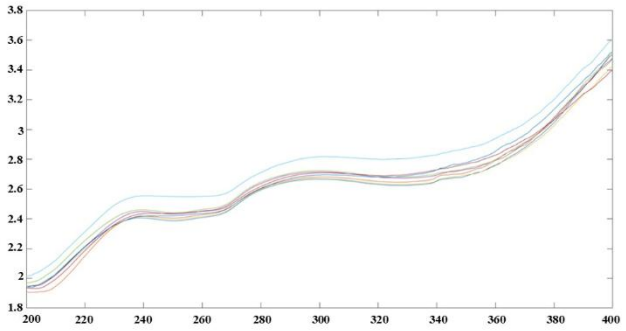


Figure 42. UV-Vis spectra obtained from the analysis of the 8 samples

A PCA was performed to assess whether this extraction procedure could effectively distinguish between the three classes of native species samples (A, D and E) analysed. The spectral region ranging from 200 nm to 400 nm was considered the most significant and with the least noise.

From the scores plot performed on all samples pretreated with D1 (*Figure 43*) it is possible to distinguish Class A from the rest of the other classes.

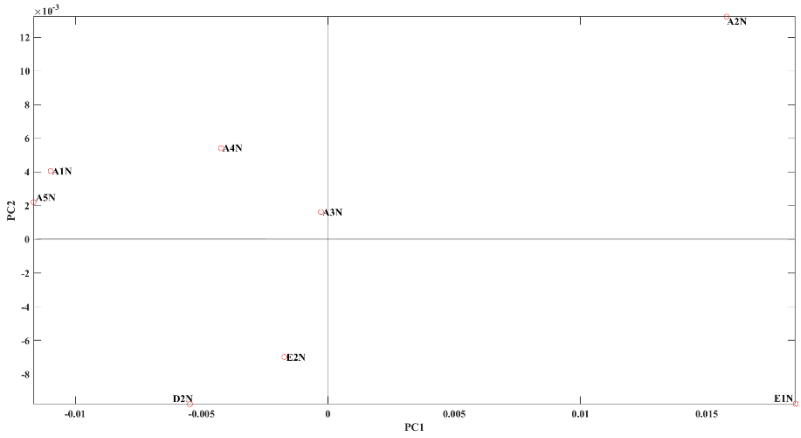


Figure 43. Scores plot performed on all samples pretreated with D1

From here, the idea was to correlate antioxidant activity with the polyphenolic profile using a spectrochromatogram. To do this, in addition to the varieties grown in the experimental field, three commercial varieties of turnip greens were sourced from Marche and Lazio, and a variety of Mugnoli from Abruzzo. The leaves belonging to these varieties were cleaned to remove any soil residues with absorbent paper and stored at -20 °C. They were then dried in an oven for 3 hours at 50 °C, ground and stored at 4 °C in flat-bottomed plastic vials with screw caps, sealed with Parafilm, as were the leaves of the varieties grown in the experimental field.

Extracts were prepared for all varieties under the same optimal conditions identified previously. The IC_{50} value was calculated in duplicate for each variety, then averaged, and the average value was attributed to the spectrocromatogram obtained from the analysis of the same extract.

Two samples extracted under the same conditions but using water as the extraction solvent (sample 41 and sample 13), rather than the water:ethanol mixture, were added to these samples. Furthermore, for the diluted samples (sample 4 and sample 6), the extract was prepared by doubling the doses (180 mg of extract in 10 mL of extracting mixture in a 20 mL flask) in order to have enough solution to make the dilutions. After being centrifuged and filtered (NY 0.22µm), all extracts were stored at -20°C and analysed only after 24 hours, with the exception of the extracts prepared in water, which were analysed at the end of the extraction process.

After conservation at -20°C, the samples were brought back to room temperature and divided into two aliquots, the first destined for HPLC-DAD analysis after re-filtering with a NY 0.20 μm filter and the other subjected to DPPH assay for the determination of IC_{50}.

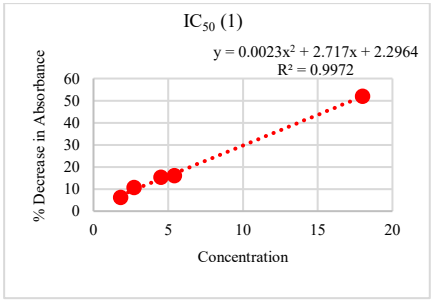
For the chromatographic analysis, 20 μL were manually injected into the HPLC-DAD system through the injector loop, while for the DPPH analysis, in addition to the extract as is, four dilutions were prepared in duplicate and analysed according to the procedure specified in the assay. Once the tests were completed, the IC_{50} values were calculated as reported below for each sample.

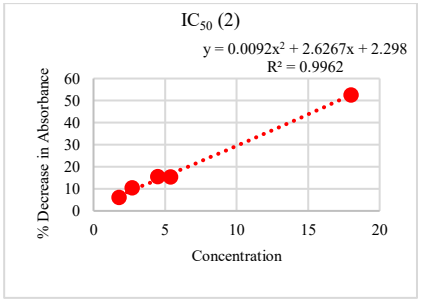
A total of 54 samples were analysed, but following an initial analysis, 13 of them were found to be outliers and were therefore excluded from the model. The values are shown below with the respective IC_{50} curves and the model constructed with only 41 samples.

To correlate the antioxidant activity (IC_{50}) with the spectrochromatographic profiles obtained from the analysis of the extracts, a regression model was developed using the multiway N-PLS (N-way Partial Least Squares) algorithm. This approach was chosen because it allows the natural three-dimensional structure of the data (samples × time × wavelength) to be preserved, simultaneously modelling the relationship between spectrochromatographic information and the antioxidant response.

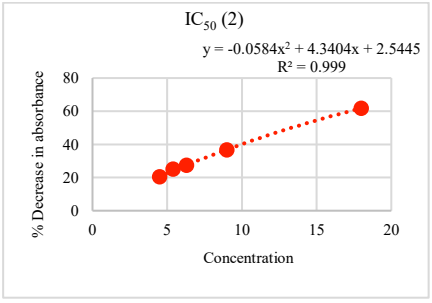
The IC_{50} value for each replicate was obtained by non-linear regression. IC_{50} values were determined by solving the quadratic regression equation adapted to the inhibition data.

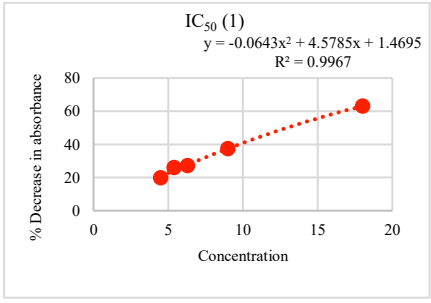
The mean IC_{50} values were determined by considering the weighted mean of the calculated values, taking into account the error in the estimate for each regression. Since the N-PLS model showed no differences, the final IC_{50} values attributed to the sample will be reported as the mean ± standard deviation on which the multiway model was constructed.



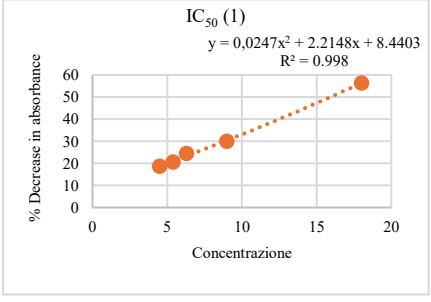


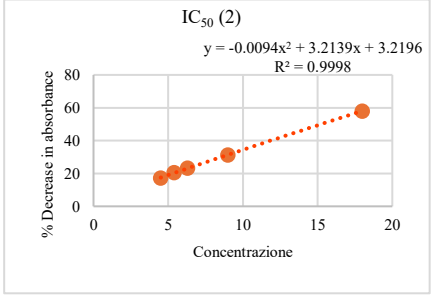
IC_{50} Sample 1	17.2±0.1



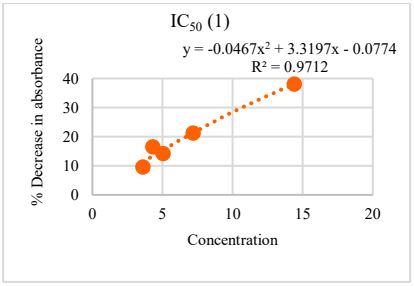


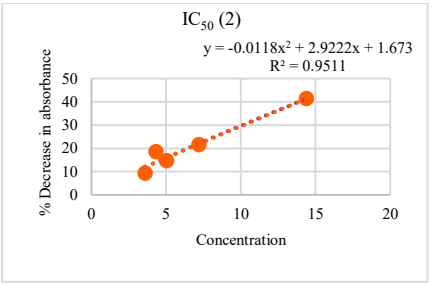
IC_{50} Sample 2	13.1±0.3



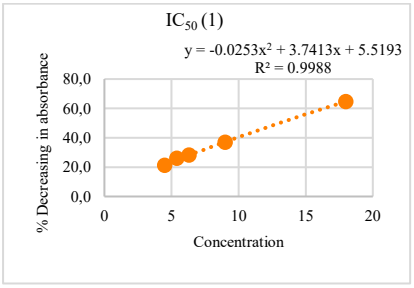


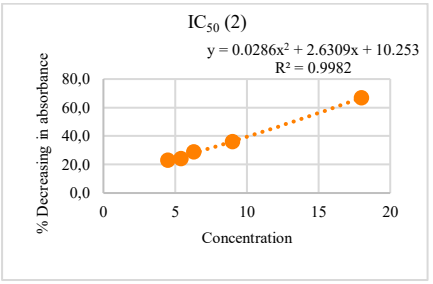
IC_{50} Sample 3	15.6±0.4



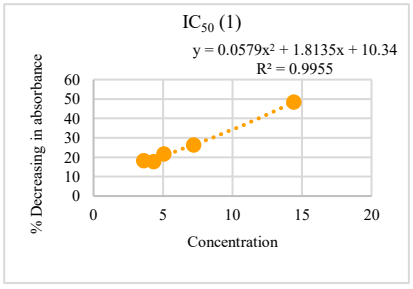


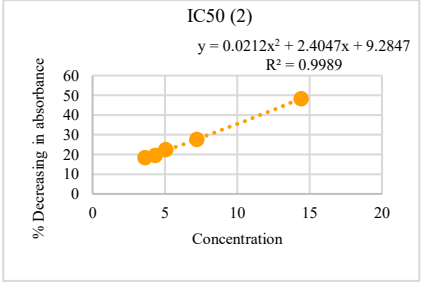
IC_{50} Sample 4	19.9±2.8



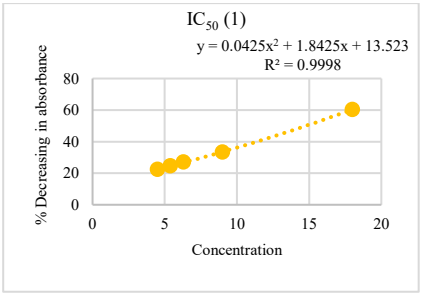


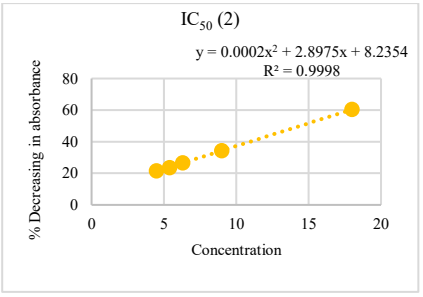
IC_{50} Sample 5	13.1±0.14



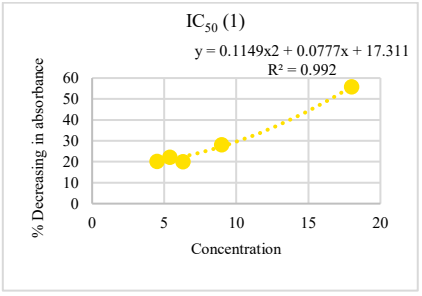


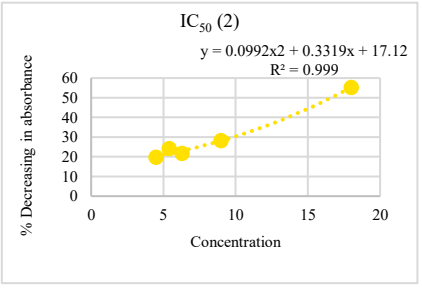
IC_{50} Sample 6	14.9±0.1



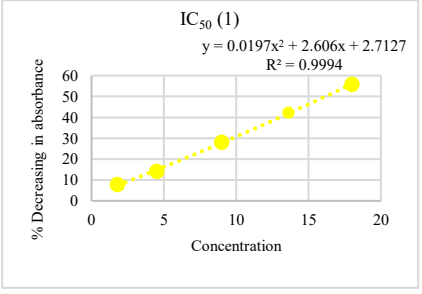


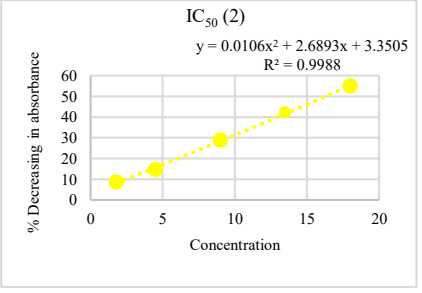
IC_{50} Sample 7	14.4±0.6



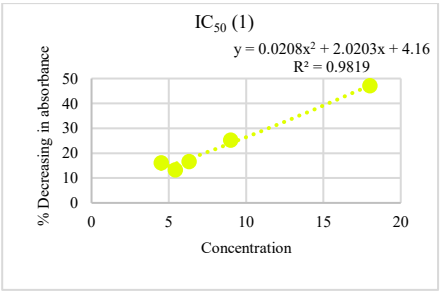


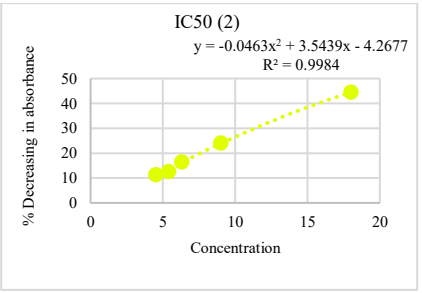
IC_{50} Sample 8	16.58±0.07



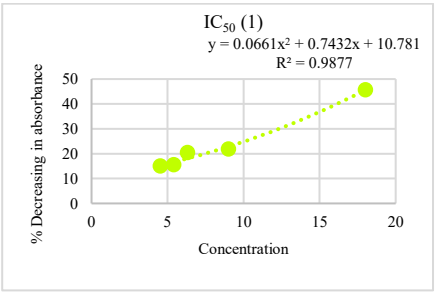


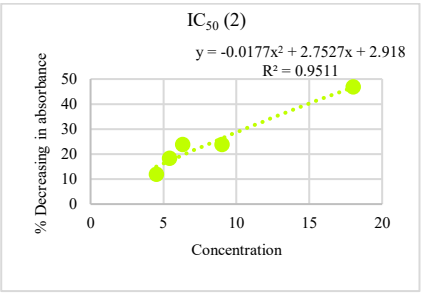
IC_{50} Sample 9	16.2±0.1



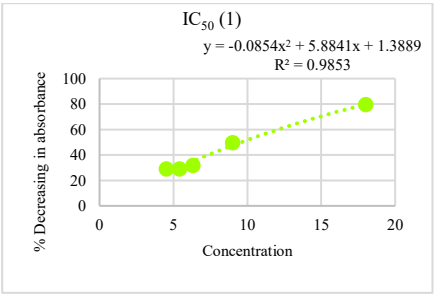


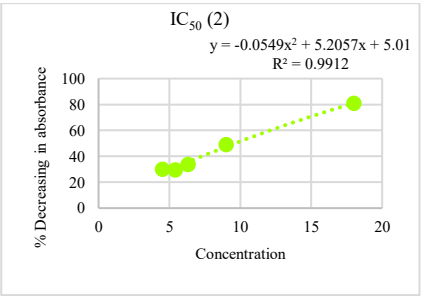
IC_{50} Sample 10	20.1±1.6



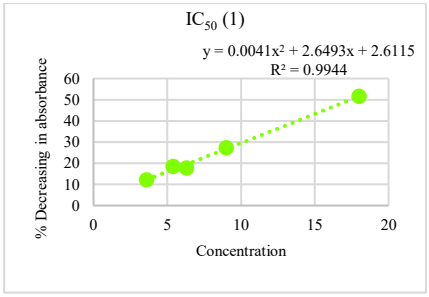


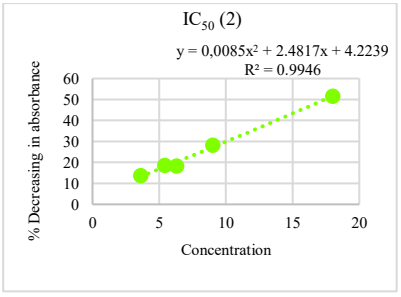
IC_{50} Sample 11	19.4±0.1



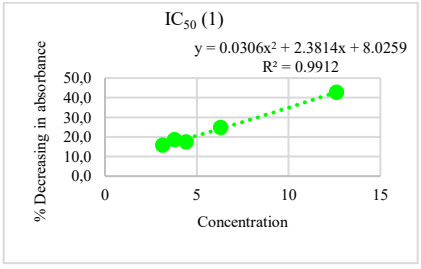


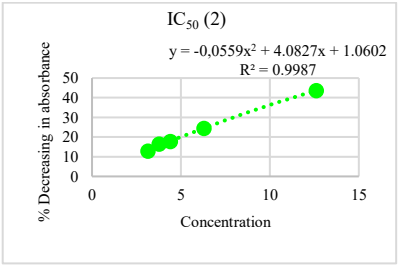
IC_{50} Sample 12	9.6±0.0



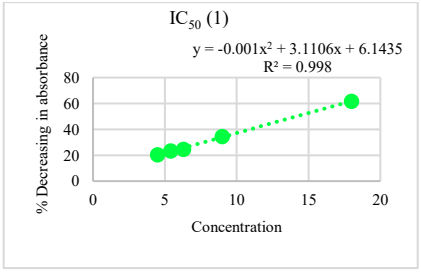


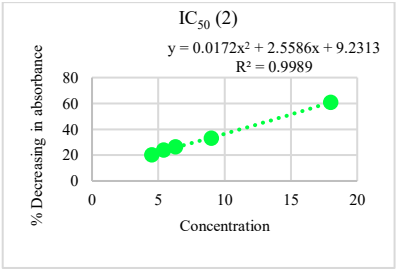
IC_{50} Sample 13	17.5±0.2



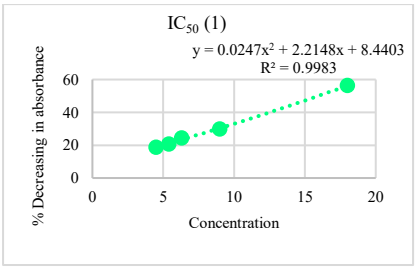


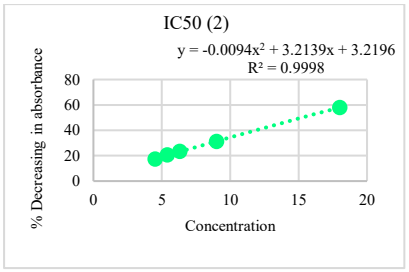
IC_{50} Sample 14	15.0±0.2



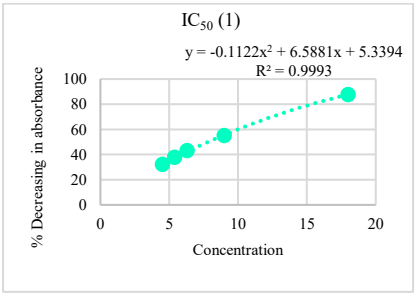


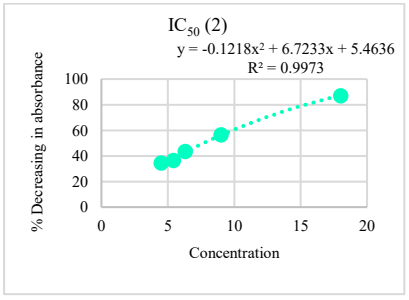
IC_{50} Sample 15	14.3±0.4



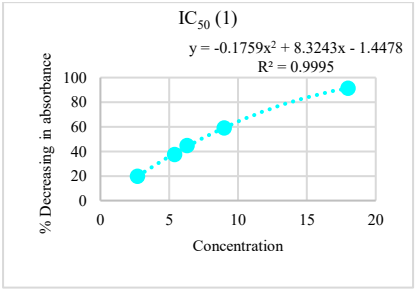


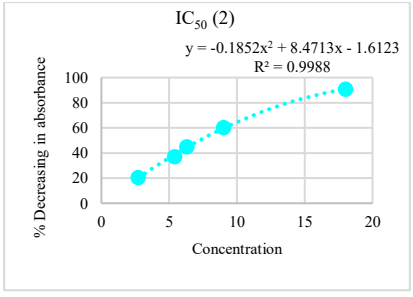
IC_{50} Sample 16	15.6±0.4



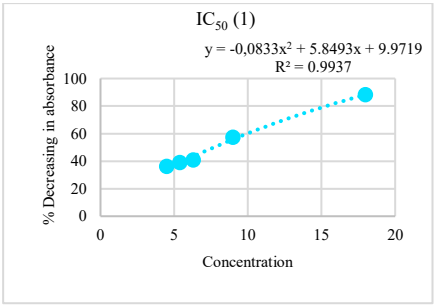


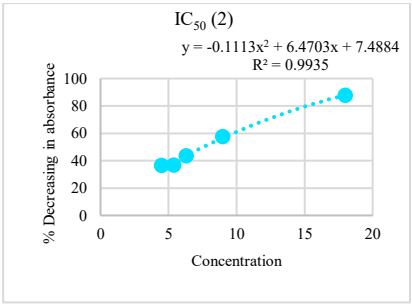
IC_{50} Sample 17	7.7±0.0



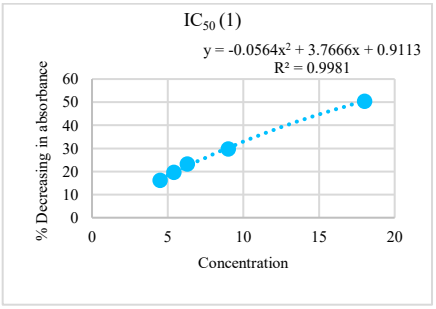


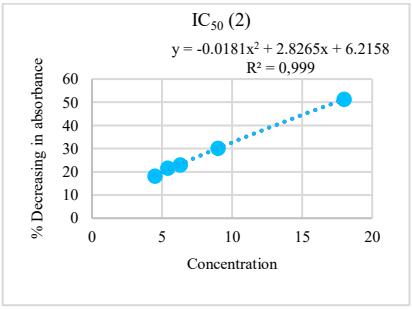
IC_{50} Sample 18	7.30±0.07



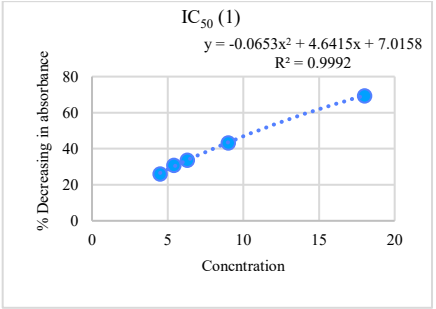


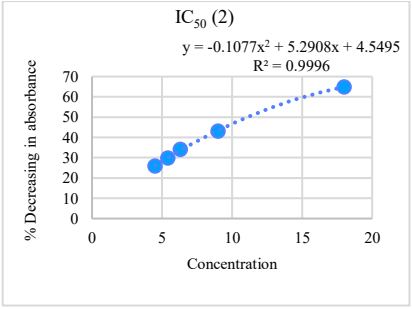
IC_{50} Sample 19	7.7±0.07



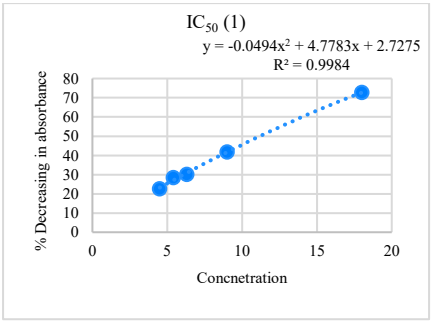


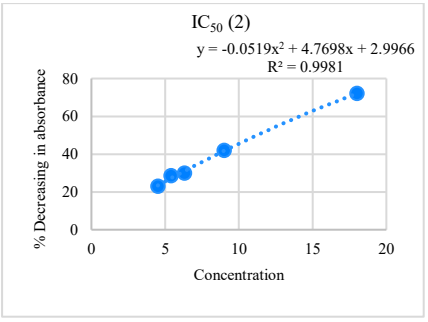
IC_{50} Sample 20	18.6±1.1



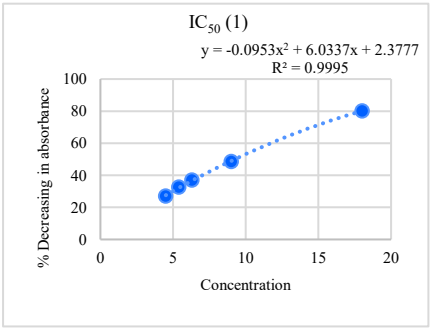


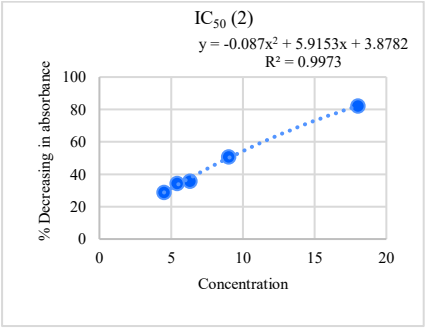
IC_{50} Sample 21	11.10±0.07



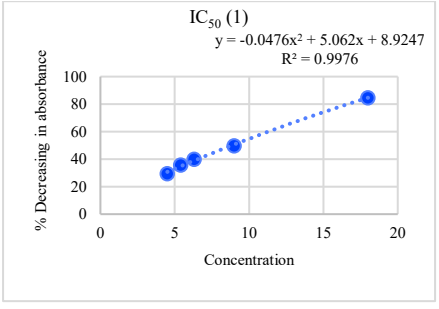


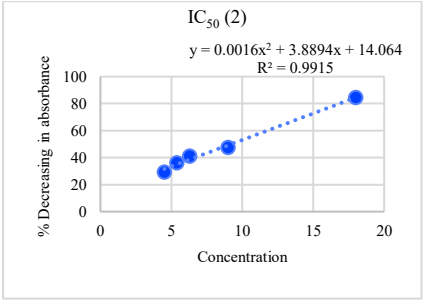
IC_{50} Sample 22	11.2±0.0



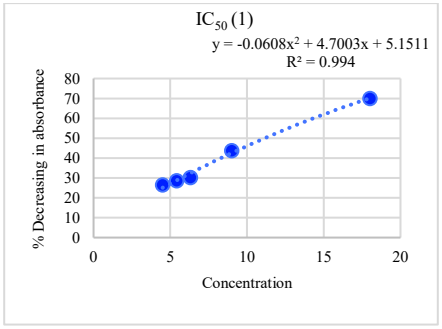


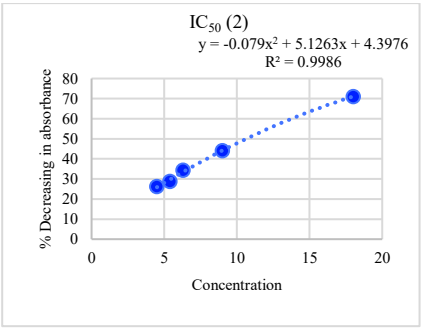
IC_{50} Sample 23	9.1±0.1



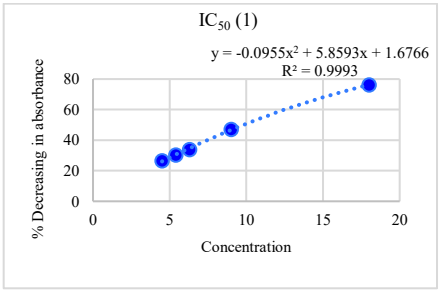


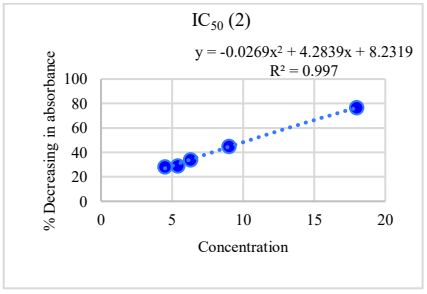
IC_{50} Sample 24	8.90±0.07



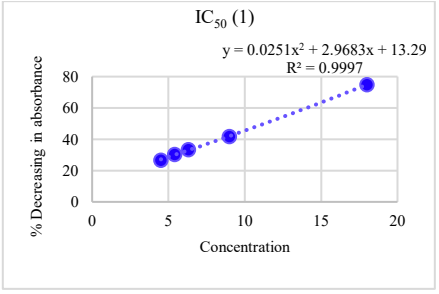


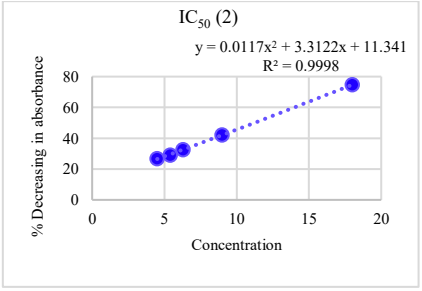
IC_{50} Sample 25	10.9±0.4



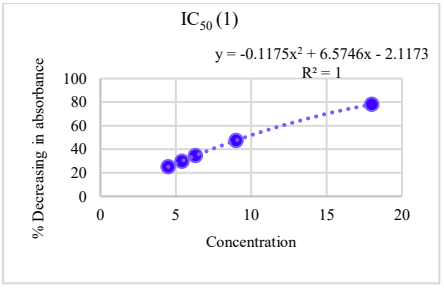


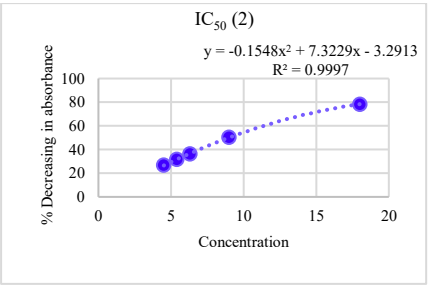
IC_{50} Sample 26	10.2±1.3



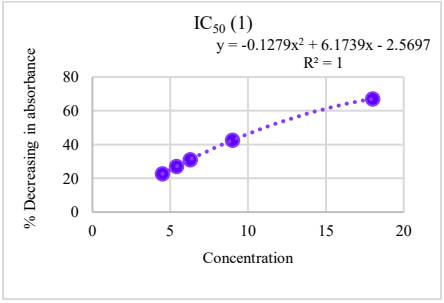


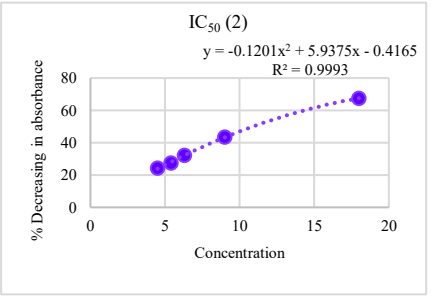
IC_{50} Sample 27	11.30±0.07



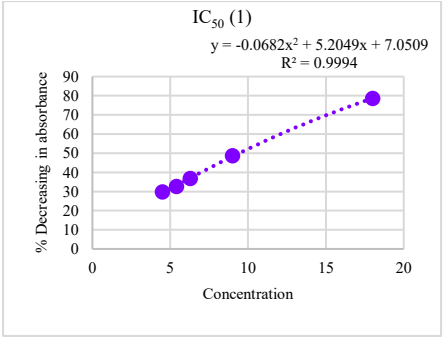


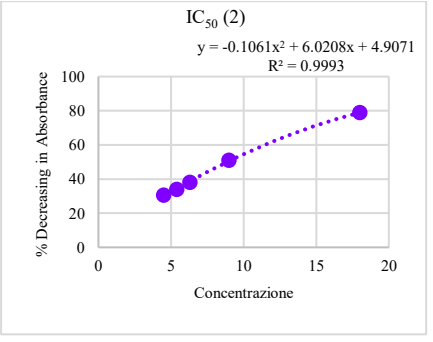
IC_{50} Sample 28	9.3±0.4



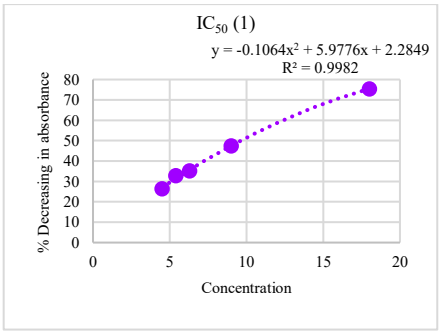


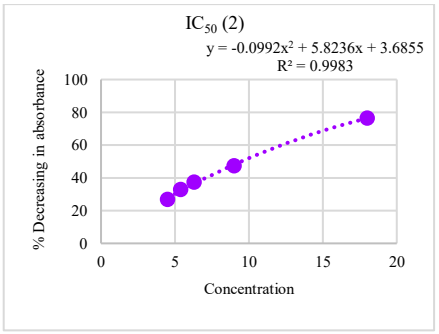
IC_{50} Sample 29	11.0±0.1



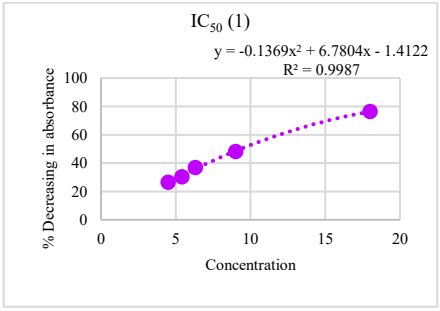


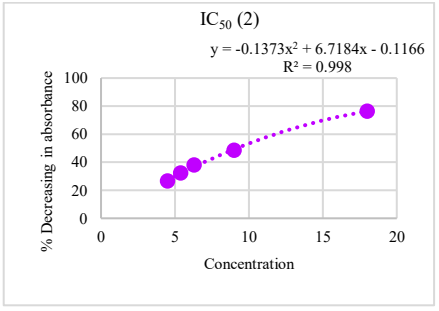
IC_{50} Sample 30	9.2±0.4



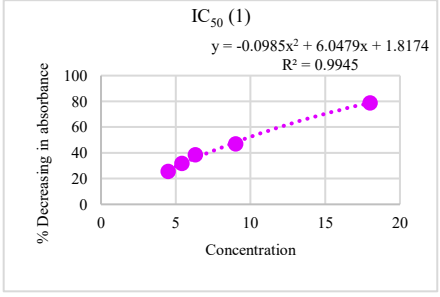


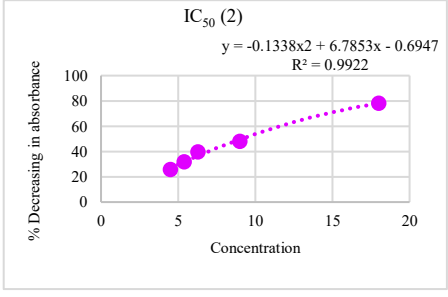
IC_{50} Sample 31	9.55±0.07



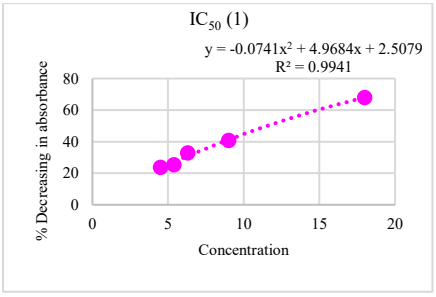


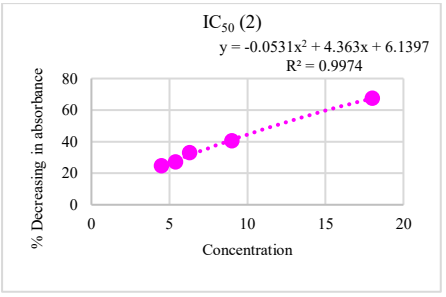
IC_{50} Sample 32	9.3±0.1



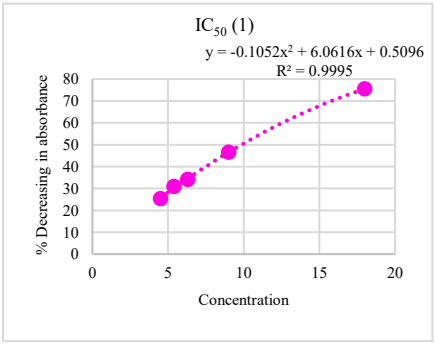


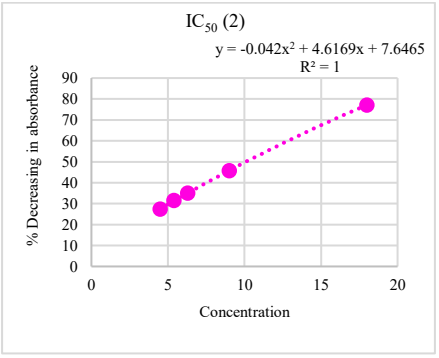
IC_{50} Sample 33	9.3±0.2



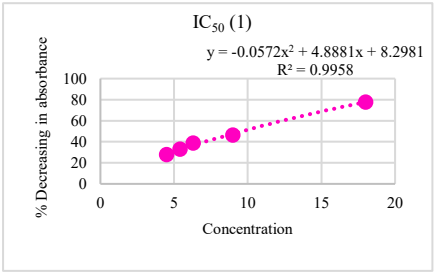


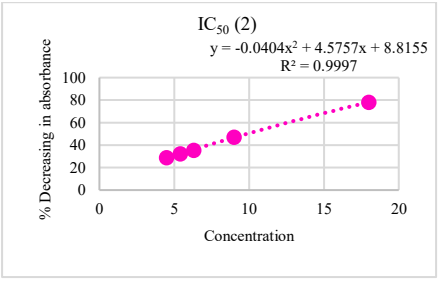
IC_{50} Sample 34	11.6±0.1



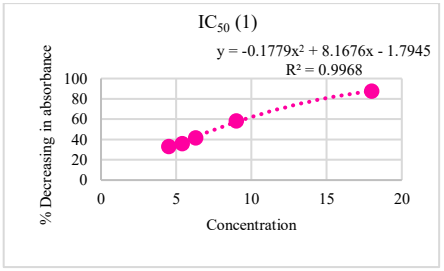


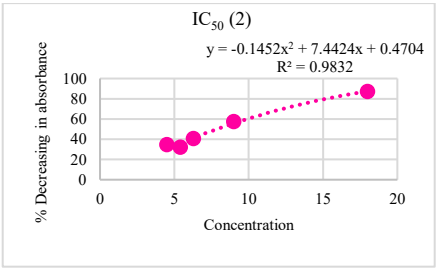
IC_{50} Sample 35	10.0±0.18



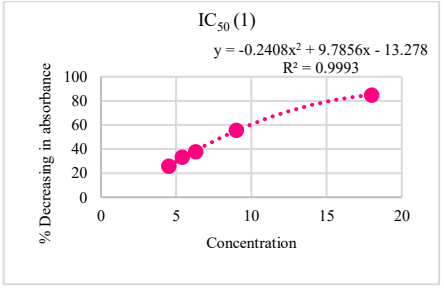


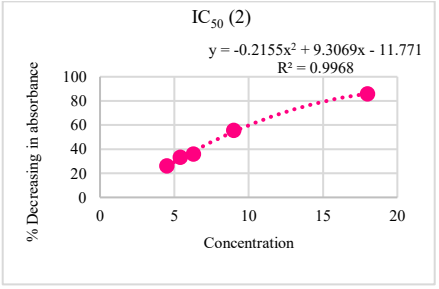
IC_{50} Sample 36	9.7±0.2



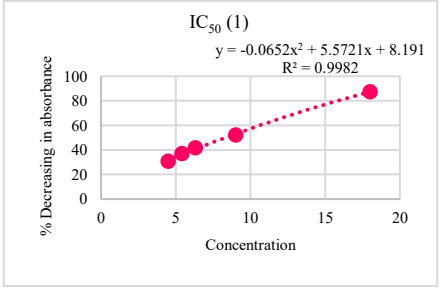


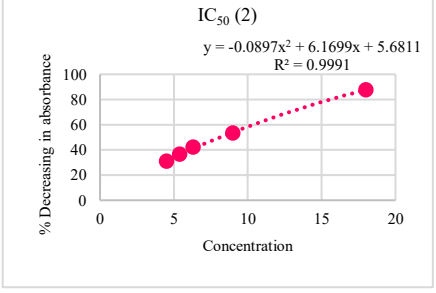
IC_{50} Sample 37	7.8±0.2



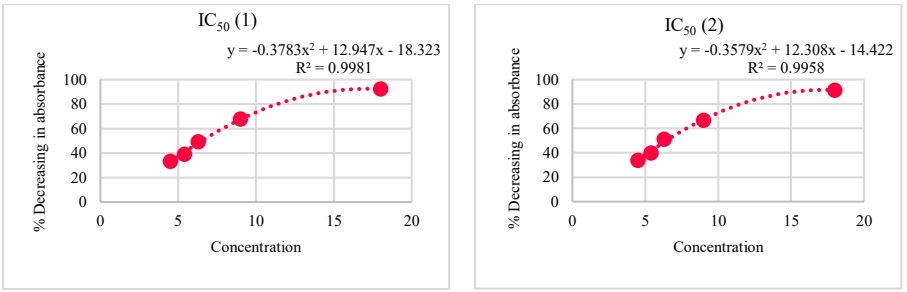


IC_{50} Sample 38	8.20±0.07

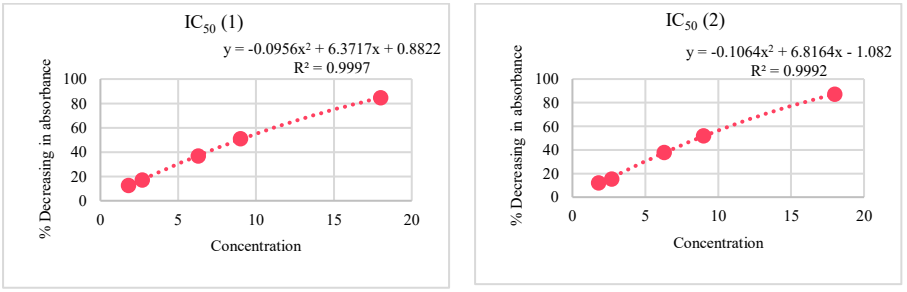




IC_{50} Sample 39	8.2±0.0



IC_{50} Sample 40	6.50±0.07



IC_{50} Sample 41	8.8±0.1

Figure 44. Dose-response curves and average IC_{50} values with relative standard deviation for each sample considered in the calibration and validation of the model

At the end of the IC_{50} value calculations, these values, with the spectrochromatograms (an example is displayed in *Figure 45*) were processed using multiway analysis according to the N-PLS algorithm.

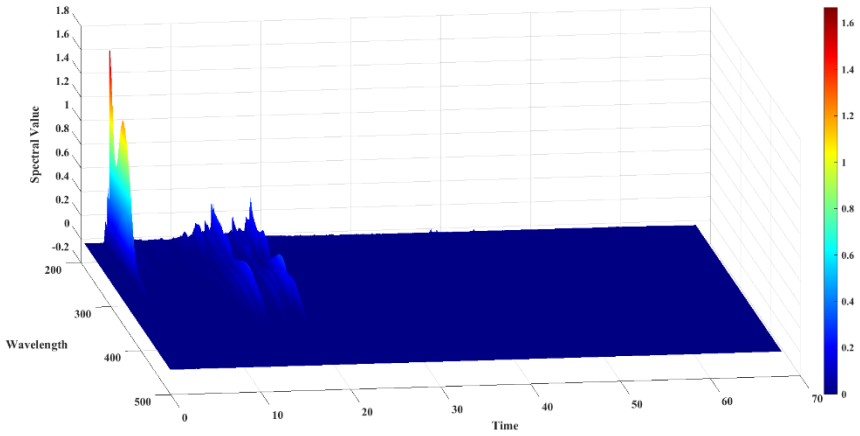


Figure 45. Example of spectrochromatogram

The data cube (sample × time × wavelength) was constructed by combining the spectrochromatograms of the analysed samples as shown in *Figure 46* and obtaining the final model using the N-PLS algorithm.

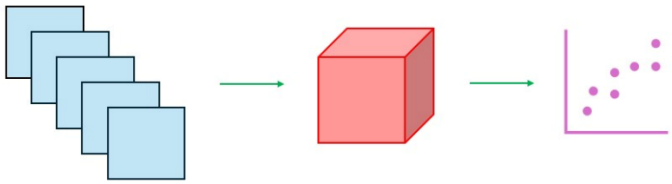


Figure 46. Simplified scheme of the merging of spectrochromatograms, cube construction and subsequent N-PLS analysis.

The time ranges were then cut down to consider only the most interesting part of the spectrum from 10 to 20 minutes and the spectral region between 210.3 nm and 446.1 nm. This resulted in a reduced spectrochromatogram as shown in *Figure 47*.

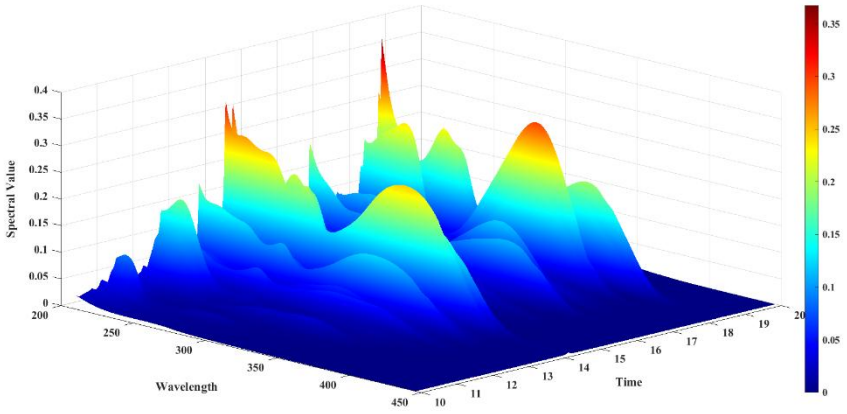


Figure 47. Example of a cut spectrochromatogram

The three-dimensional structure of the data was preprocessed by centring along the first mode to eliminate offset effects mainly due to baseline differences between spectra and autoscaling along the other two modes (time and wavelength) to normalise the variables and make their contribution homogeneous, allowing the N-PLS model to exploit the correlation between all dimensions without scale bias. The values were collected in a Y matrix, which was only preprocessed with centring, as being a unit response variable, they do not require autoscaling.

The cube is the result of combining the spectrochromatograms of the 41 total samples analysed.

The dataset was divided into a training set and a test set with 29 samples and 12 samples respectively.

The raw and preprocessed data relating to the samples belonging to the training set are shown in *Figure 48* and *49* below.

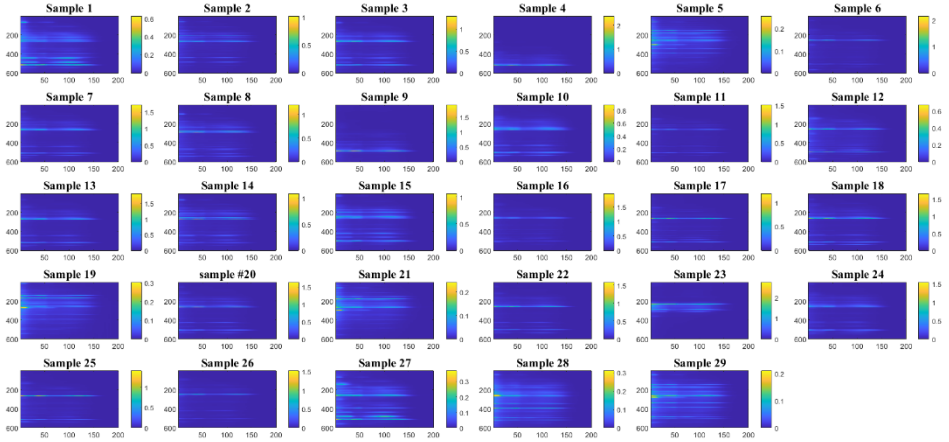


Figure 48. Raw data of training set (time × wavelength)

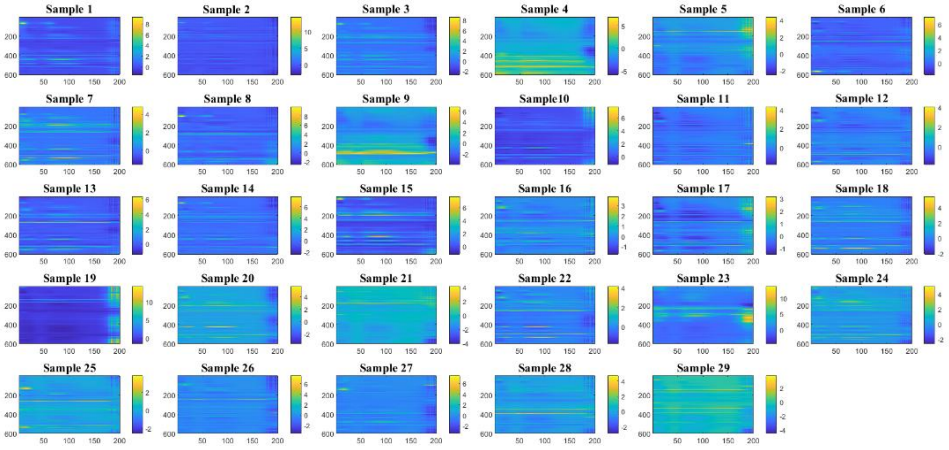


Figure 49. Preprocessed data of training set (time × wavelength)

Figure 50 and *Figure 51* instead show the raw and preprocessed data for the 12 samples belonging to the test set.

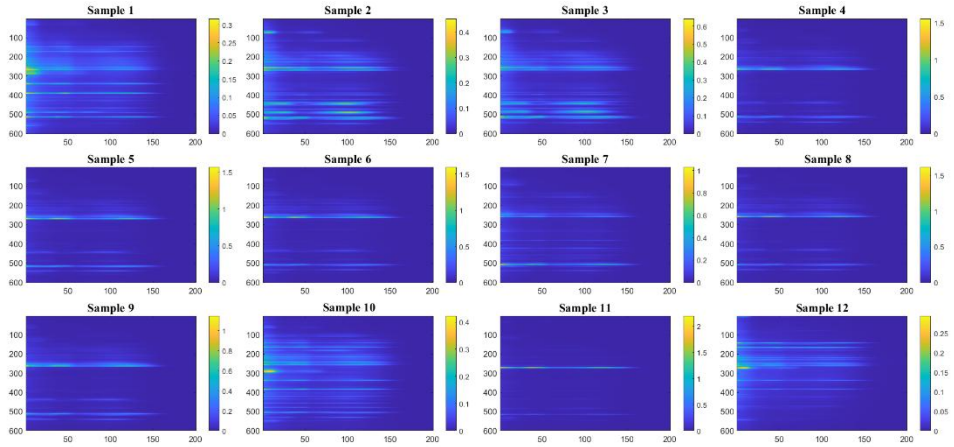


Figure 50. Raw data of test set (time × wavelength)

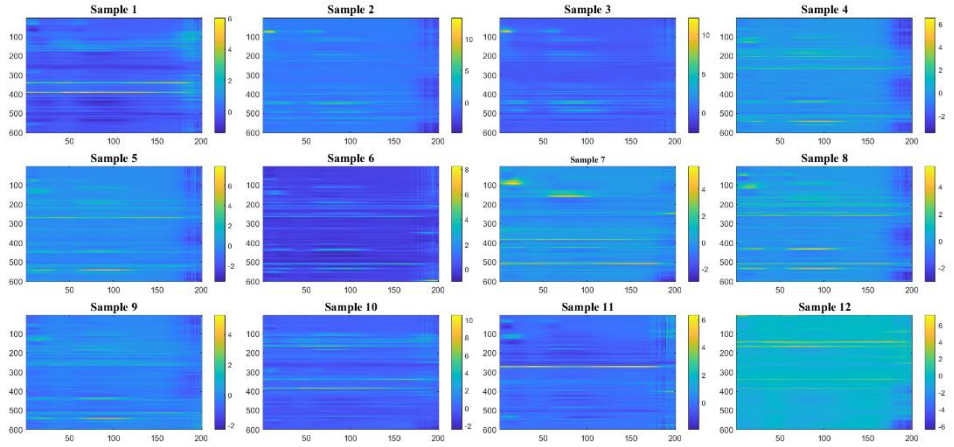


Figure 51. Preprocessed data of test set

The model was optimised through cross-validation, choosing 12 LVs based on the minimum Root Mean Square Error of Cross-Validation (RMSECV) (*Figure 52*). In this way, a calibration coefficient of determinant R^2_{cal} of 99.96% and a validation coefficient of determination (R^2_{val}) of 81.8% were obtained with an RMSECV of 1.8. The coefficients of determination, R^2_{cal} ed R^2_{val}, represent the percentage of variance explained by the model, in calibration and validation, respectively, of the $\boldsymbol{y}$ data. Although less relevant for predictive purposes, for the sake of completeness, the percentage of

variance explained in the ***X*** three-way array is also reported, corresponding to 75.8% in calibration and 47.2% in validation, indicating the fraction of spectral variance captured by the model.

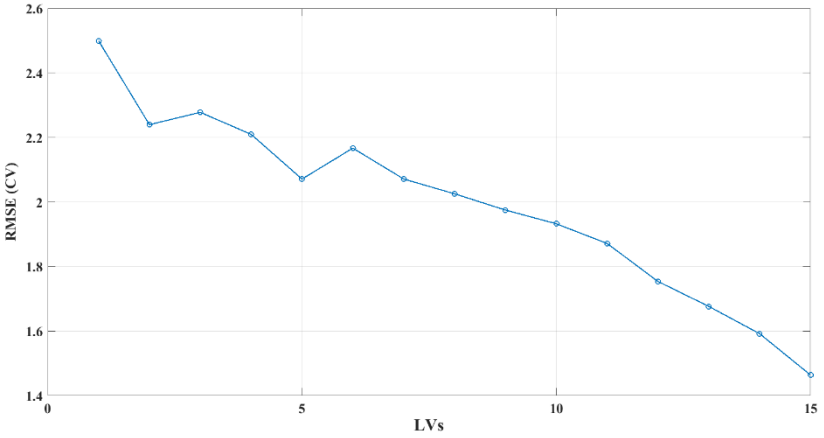


Figure 52. RMSE (CV) values as a function of the number of N-PLS components

From the residual plot in *Figure 53* it can be seen that there is no apparent systematic pattern among the errors, indicating that they are randomly distributed around zero. This suggests that the model adequately captures the systematic structure of the variance in the data and that there are no significant residual biases.

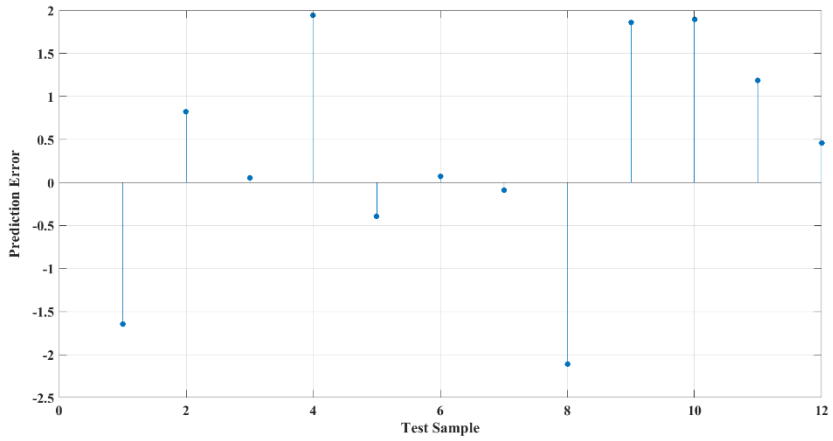


Figure 53. Plot of residuals

Finally, the model was validated on an external test set, providing a prediction coefficient of determination $R^2_{pred} = 0.81$ and a Root Mean Square Error of Prediction (RMSEP) of 1.3, values that indicate good predictive capacity of the model, as shown in *Figure 54*.

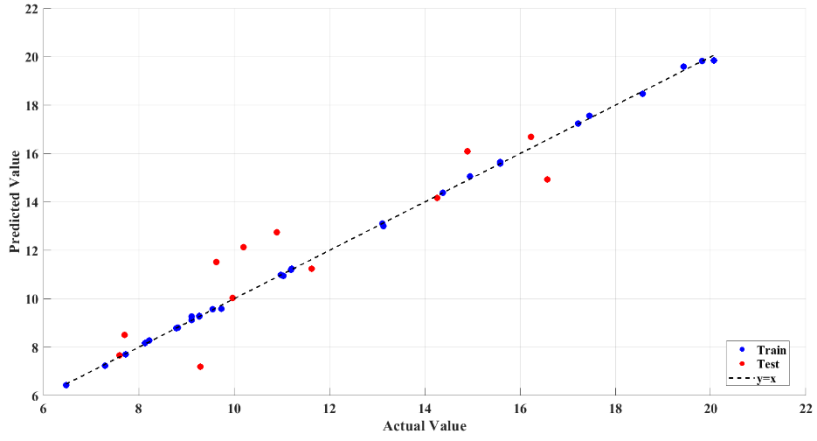


Figure 54. N-PLS prediction model

From the plot, it is evident that the training samples are well distributed along the regression line, while the test samples show greater dispersion,

especially for intermediate IC_{50} values. This behaviour reflects both the intrinsic limitations of the model and the experimental variability of the DPPH assay, which is known to be less reproducible than other techniques for measuring antioxidant power. Despite this, overall the model shows a good fit, as can be deduced from the R^2 and RMSE values, and is robust and has good predictive power, as shown by the graph in *Figure 54*, the absence of patterns in the residuals and the R^2 value in external validation.

The advantage of using N-PLS is that it allows the information distributed across the three dimensions (sample × time × wavelength) to be exploited jointly, preserving the original structure of the data and reducing the loss of information that would occur if the cube were unfolded into a two-dimensional matrix. Furthermore, it reduces sensitivity to noise and chromatographic retention shifts and allows the latent relationships between spectral characteristics and antioxidant activity (IC_{50}), to be highlighted, providing a more robust and interpretable model.

IC_{50} measurements were performed over time to assess the stability of the antioxidant activity of the assay, both by preparing individual aliquots and storing them for different numbers of days, and by preparing a single aliquot and analysing a portion of it after several days. The absence of significant variations in DPPH values allows to conclude that the extracts are stable.

The analysis of the same commercial variety from different regions of central Italy (Lazio, Abruzzo and Marche) showed that soil and climate conditions, as well as any differences in cultivation, can affect antioxidant activity. The samples from the Marche region were found to have the highest amount of antioxidant compounds with an IC_{50} of 7.7±0.0 which is not very different from those from Lazio (IC_{50}= 8.2± 0.01), while for the Mugnoli from Abruzzo, an IC_{50} of 11.9 ±0.1 was obtained.

The difference between the IC_{50} values of the replicates carried out on the same variety may lie in the non-homogeneity of the leaf samples, especially in the matrices stored at -20 °C for longer periods of time, as they showed clear signs of ageing due to their age.

Section 2

Chamomile (*Matricaria chamomilla L.*)

CHAPTER 5

OPTIMISATION OF COLD EXTRACTION OF BIOACTIVE SUBSTANCES FROM CHAMOMILE FLOWERS THROUGH CHEMOMETRIC APPROACHES

1. Objective of the work carried out on chamomile

In this study, the aim was to identify the optimal conditions for cold aqueous extraction of bioactive substances from the dried flower heads of *Matricaria chamomilla L.* chamomile flowers, through the design of experiments using a full factorial design, which allows for the preparation of a number of laboratory tests taking into account the effects of individual variables and their interactions, reducing the overall number of experiments compared to a traditional approach. The extracts were analysed using HPLC-DAD and UV-Vis analysis coupled with chemometric approaches such as PCA to explore patterns and reduce data complexity, and ASCA to evaluate the effects of experimental variables. The DPPH assay was also used to evaluate the antioxidant capacity of the extract obtained under the optimal conditions identified through the experimental design.

All this stems from the increase in consumer demand for natural and nutraceutical products, stimulating industrial growth, which in turn has

encouraged research into the safety and bioactivity of chamomile-based formulations[147] [148].

Chamomile is one of the most widely consumed herbs thanks to its therapeutic properties, including analgesic, sedative and anti-inflammatory effects[149], therefore, it is important in phytotherapy[150] and is included in numerous pharmacopoeias and used in various fields of application such as cosmetics, food and pharmaceuticals[151] [152].

In recent years, research attention has also extended to methodological aspects related to its extraction and characterization, with a growing interest in green approaches, i.e. unconventional methods that minimise sample degradation. The aim was to develop a safe and eco-friendly method that preserves the native composition of water-soluble compounds in chamomile using chemometric techniques.

An important aspect to consider is the strong influence of environmental factors such as soil and climate conditions, exposure to diseases and pests, seasonal variations, drying and storage[153]. These factors significantly affect the native composition of medicinal herbs. In this study, such variability was minimised by using a single batch of commercial dried chamomile, focusing on the effects of the experimental variables.

The results discussed in the present work refer to the commercial dried chamomile sample as received, without any further processing or thermal treatment, in order to preserve its native composition. A complete factorial design was developed to evaluate the effects of extraction variables, such as

---

[147] L. K. Mailänder, P. Lorenz, H. Bitterling, F. C. Stintzing, R. Daniels, and D. R. Kammerer, "Phytochemical characterization of chamomile (*Matricaria recutita* L.) roots and evaluation of their antioxidant and antibacterial potential," *Molecules*, vol. 27, no. 23, Art. no. 8508, 2022.

[148] N. Tsivelika, M. Irakli, A. Mavromatis, P. Chatzopoulou, and A. Karioti, "Phenolic profile by HPLC-PDA-MS of Greek chamomile populations and commercial varieties and their antioxidant activity," *Foods*, vol. 10, no. 10, Art. no. 2345, 2021.

[149] J. K. Srivastava, E. Shankar, and S. Gupta, "Chamomile: A herbal medicine of the past with a bright future (review)," *Mol. Med. Rep.*, vol. 3, no. 6, pp. 895–901, 2010, doi: 10.3892/mmr.2010.377.

[150] O. Singh, Z. Khanam, N. Misra, and M. K. Srivastava, "Chamomile (*Matricaria chamomilla* L.): An overview," *Pharmacogn. Rev.*, vol. 5, no. 9, pp. 82–95, 2011.

[151] A. El Mihyaoui, J. C. G. Esteves da Silva, S. Charfi, M. E. C. Castillo, A. Lamarti, and M. B. Arnao, "Chamomile (*Matricaria chamomilla* L.): A review of ethnomedicinal use, phytochemistry and pharmacological uses," *Life*, vol. 12, no. 4, Art. no. 479, 2022.

[152] M. M. El Joumaa and J. M. Borjac, "*Matricaria chamomilla*: A valuable insight into recent advances in medicinal uses and pharmacological activities," *Phytochem. Rev.*, vol. 21, no. 6, pp. 1913–1940, Dec. 2022.

[153] R. Baranauskienė, P. R. Venskutonis, and O. Ragažinskienė, "Valorisation of Roman chamomile (*Chamaemelum nobile* L.) herb by comprehensive evaluation of hydrodistilled aroma and residual non-volatile fractions," Food Research International, vol. 160, p. 111715, 2022.

extraction temperature (T), extraction time (t) and the quantity of chamomile used in the preparation of the extracts (Q), on the polyphenolic profile of the samples by chromatographic analysis.

The latter was conducted using an HPLC-DAD (with a 717 Plus autosampler thermostated at 10°C) to collect information on the polyphenolic profile of chamomile at extraction temperatures between 15°C and 25°C. At the same time, a UV-Vis analysis was performed to evaluate the impact of experimental factors on spectroscopic variables by applying the ASCA algorithm, which provided relevant information for characterising the bioactive compounds[154]. Through ASCA, it was possible to evaluate the UV-Vis profile, offering an untargeted method for understanding the effects of the variables. Using the extraction optimum, identified with the factorial design, the IC_{50} value was calculated.

In this context, it was therefore possible to highlight how combining analysis techniques with chemometrics can improve the quality of extracts and characterise the influence of process variables in conventional cold water extraction of chamomile.

While flavonols were detected in the chamomile extracts, they were not fully resolved and thus were not considered in the final analysis.

2. Performance of laboratory experiments

After identifying the variables that most influenced the quality of the extract in terms of polyphenolic profile, a full factorial design was devised to study the simultaneous effect of three factors: extraction time, quantity of chamomile used to obtain the extract, and extraction temperature, at three levels.

For the HPLC-DAD analysis, an experimental domain reported in *Figure 55* was explored.

---

[154] N. Sarma, R. Gogoi, T. Begum, M. Lal, K. Perveen, N. A. Alsahikh, and J. A. Alsulami, "A study on the chemical profile of cultivated chamomile (*Matricaria chamomilla* L.) flower essential oil from North East India with special emphasis on its pharmacological importance," *J. Essent. Oil Bear. Plants*, vol. 26, no. 3, pp. 745–760, 2023.

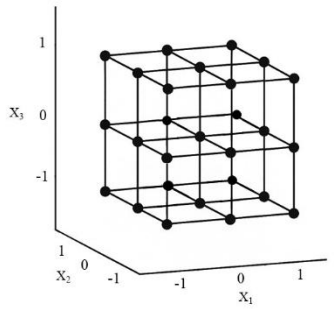


Figure 55. Explored experimental domain

As shown in *Table 18*, for temperature, the levels investigated were 15°C, 20°C and 25°C; for extraction time, 32 min, 52 min and 72 min; and for chamomile quantity, 1.5 g, 2 g and 2.5 g.

Levels	-1	0	1
Temperature (°C)	15	20	25
Time (min)	32	52	72
Quantity (g)	1.5	2	2.5

Table 18. Levels of the variables investigated

More specifically, *Figure 56* illustrates the samples prepared and analysed by simultaneously and randomly varying the three experimental factors at different levels within the experimental domain.

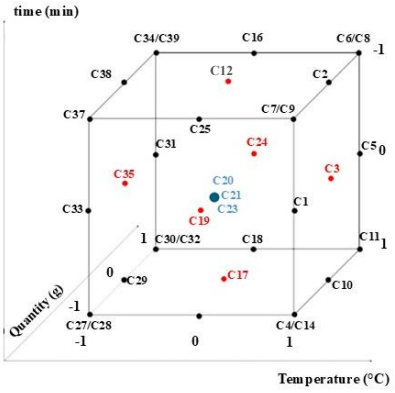


Figure 56. Samples prepared and analysed within the experimental domain.

Figure 56 shows the replicas of the samples at the centre of the experimental domain in blue (C20, C21, C23), the tests at the centre of the cube faces in red (C3, C17, C19, C26, C24, C35) and finally the replicates at the points at the extremes of the domain (C4/C14, C6/C8, C27/C28, C30/C32 and C34/C39) are shown. T*able 19* lists the samples analysed.

Sample	Temperature – *T* (°C)	Time – *t* (min)	Quantity – *Q* (g)
C1	25	52	1.5
C2	25	72	2
C3	25	52	2
C4	25	32	1.5
C5	25	52	2.5
C6	25	72	2.5
C7	25	72	1.5
C8	25	72	2.5
C9	25	72	1.5
C10	25	32	2
C11	25	32	2.5
C12	20	32	1.5
C14	25	32	1.5
C16	20	72	2.5
C17	20	32	2
C18	20	32	2.5
C19	20	52	1.5
C20	20	52	2
C21	20	52	2
C23	20	52	2
C24	20	52	2.5
C25	20	72	1.5
C26	20	72	2
C27	15	32	1.5
C28	15	32	1.5
C29	15	32	2
C30	15	32	2.5
C31	15	52	2.5
C32	15	32	2.5
C33	15	52	1.5
C34	15	72	2.5
C35	15	52	2
C37	15	72	1.5
C38	15	72	2
C39	15	72	2.5

Table 19. List of experiments conducted

The chromatographic runs corresponding to points [0 -1 -1] and [0 1 0] respectively C16 and C36, as well as samples C13 and C22 respectively [1 0 1] and [0 0 1] of the experimental domain were unsuccessful and therefore, given their marginal position in the experimental domain (as the optimum was far from these points), they were not taken into consideration since the orthogonality of the design was not significantly affected. This was calculated using the Variance Inflation Factor (a measure of multicollinearity between independent variables in a multiple regression model), which indicated that the variables were independent of each other without risk of collinearity, and it was possible to disregard the two unsuccessful experiments. For this reason, a total of 33 extracts were analysed.

The extracts were prepared from commercially available dried chamomile flower heads (*Matricaria chamomilla* L.) from the same batch, which were ground for 90 seconds at 5000 rpm using a Tube Mill 100 control. After grinding, the chamomile powder was sieved with a 1 mm mesh sieve and weighed in quantities to obtain different mass/volume ratios after the addition of 100 mL of water, as established by the experimental design. The extraction temperature was kept constant using a thermostatic water bath with circulating water, inside which special sterile disposable containers were placed where the extraction took place, keeping everything in the dark and stirring the solution with the aid of a magnetic stirrer for the entire duration of the extraction time defined by the DoE. Once the extraction was complete, the extract was centrifuged for 10 minutes at 3000 rpm and then filtered using cellulose acetate (CA) filters with a filtering surface area of 13 mm and pores with a diameter of 0.22 μm. The extracts were then analysed by HPLC-DAD and, only after a 140-fold dilution in suprapur water, they were also analysed by UV-Vis spectroscopy.

The qualitative analysis of the aqueous extracts was performed by HPLC-DAD.

Chromatographic analysis was conducted with a constant flow rate of 1 mL/min and a sample injection volume of 10 μL. The mobile phase consisted of a 0.1% H3PO4 solution (A), prepared using HPLC-grade acid supplied by Thermo Scientific Chemicals, (Waltham, MA, USA), and a second solution (B) of 100% acetonitrile (≥99.9%, gradient-grade for HPLC, Sigma-Aldrich) was eluted under gradient elution conditions as shown in *Table 20*.

Time	Flow	%A	%B
	1.0	95.0	5.0
10.0	1.0	95.0	5.0
35.0	1.0	5.0	95.0
45.0	1.0	5.0	95.0
50.0	1.0	95.0	5.0
60.0	1.0	95.0	5.0

Table 20. Mobile phase elution gradient during HPLC-DAD analysis of extracts

Before each analysis, the column was equilibrated by isocratic elution for 60 minutes using a mobile phase composition of 95% A and 5% B.

At the end of the analysis of each sample, the composition of the mobile phase was kept constant for 10 minutes under the final gradient conditions coinciding with the initial ones, i.e. 95% A and 5% B, to restore the conditions at the start of the analysis with column conditioning in isocratic elution followed by analysis of the next sample.

HPLC-grade standards obtained from Sigma-Aldrich (St. Louis, MO, USA) were analysed under the same chromatographic conditions as the samples. The standards analysed are shown in *Table 21*

Phenolic class	Compound	% of purity
Hydroxybenzoic acid derivatives	Protocatechuic acid	> 97.0
	Gentisic acid	> 98.0
	Syringic acid	> 95.0
	Gallic acid	97.5 – 102.5
	Vanillic acid	> 97.0
Hydroxycinnamic acids	Caffeic acid	> 98.0
	Trans-ferulic acid	99.0
	p-cumaric acid	> 98.0
	3-O-caffeoylquinic acid	95.0
Flavones	Apigenin	95.0
	Luteolin	> 97.0
Flavonols	Quercetin	> 95.0
	Rutin trihydrate	> 95.0
Phenols	Tyrosol	98.0

Table 21. Standard compounds analysed

The stock solutions were prepared at a concentration of 1 $mg \cdot mL^{-1}$ by dissolving the standards in a mixture of acetonitrile (≥99.9%, gradient-grade for HPLC, Sigma-Aldrich) and methanol (≥99.0%, HPLC grade, Carlo Erba Reagenti, Milan, Italy) in a ratio of 70:30. From these, standard solutions diluted to 5 $mg \cdot mL^{-1}$ were then prepared and analysed.

As already mentioned, dilution was required before performing the UV-Vis analysis of the chamomile extracts.

Quartz cuvettes with an optical path length of 1 cm were used. The reference blank was recorded using pure water. The spectra were acquired in a spectral range between 200 nm and 500 nm and with a data interval of 1 nm.

As a final step, to perform the DPPH colorimetric assay in order to evaluate the antioxidant activity and calculate the IC_{50} on the chamomile extract obtained from the optimal extraction conditions identified by DoE, UV-Vis was used, monitoring a single wavelength of 516 nm and using methanol (≥ 99.9%, HPLC grade, Sigma-Aldrich) as a reference blank.

Five dilutions of the extract were then prepared at concentrations of 1.5 $mg \cdot mL^{-1}$, 1.88 $mg \cdot mL^{-1}$, 3.0 $mg \cdot mL^{-1}$, 3.75 $mg \cdot mL^{-1}$ and 7.5 $mg \cdot mL^{-1}$ and subjected to the DPPH assay. The absorbance values obtained from the two

replicates were averaged and this value was used for the final calculation of the IC_{50}.

3. Results

3.1 HPLC-DAD results

Following HPLC-DAD analysis performed on chamomile extracts, chromatograms were obtained as shown in *Figure 57*.

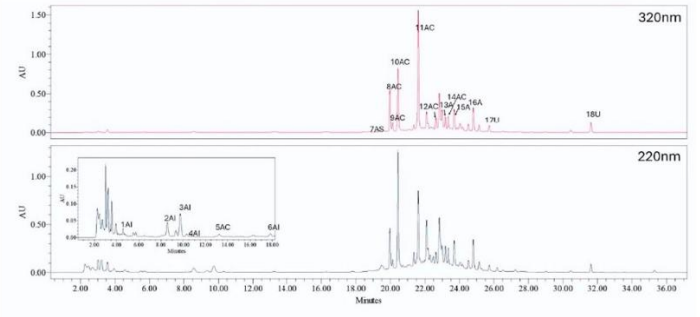


Figure 57. Example of chromatogram obtained from the HPLC analysis extracted at 320 nm and 220 nm. The peaks are labeles with abbreviation corresponding to the DoE response vectors.

In *Figure 57,* the attribution of the 18 monitored peaks is shown, indicating the order of elution and attribution to the polyphenolic class to which they belong, based on the UV-Vis spectra associated with each peak and the analytical standards available. Attribution was made only for the 18 peaks that were well resolved at their maximum absorption, as reported in *Table 22*.

Abbreviation	Compound
1AI	*p*-hydroxybenzoic acid[155]
2AI	GentisicAcid[156]
3AI	Protocatechuic Acid[146]
4AI	Tyrosol/Tryptophan Derivative[157]
5AC	3-O-Caffeoylquinic acid[146]
6AI	*p*-hydroxybenzoic acid derivative[147]
7AS	Salicylic acid derivative[145]
8AC	5-O-Caffeoylquinic acid[147]
9AC	4-O-Caffeoylquinic acid[147]
10AC	*cis*-2-Hydroxy-4-methoxycinnamic acid 2-O-glucopyranoside[145]
11AC	*trans*-2-Hydroxy-4-methoxycinnamic acid 2-O-glucopyranoside[145]
12AC	Dicaffeoylquinic acid (derivative 1)[146]
13A	Apigenin-7-O-glucoside[147]
14AC	Dicaffeoylquinic acid (derivative 2)[145]
15A	Apigenin-7-O-hexoside (derivative 1)[147]
16A	Apigenin-7-O-hexoside (derivative 2)[147]
17U	Unknown
18U	Unknown

Table 22. Attempts to attribute compounds to the chromatographic peaks obtained from chamomile extracts

In this table, AI indicates the class belonging to hydroxybenzoic acids; AC indicates the class of hydroxycinnamic acids; A indicates the class of phenols in general; U indicates unassigned peaks. In *Table 22*, derivative 1 and derivative 2 indicate the order of elution of the derivatives.

Since UV-Vis profiles can be easily recognised and attributed to specific polyphenolic families, available aglycone standards were used to accurately assign chromatographic signals to the corresponding polyphenol classes. After determining the class to which the compounds belonged, an attempt was made to identify the specific derivative compound by consulting the literature. This involved considering only the elution order and absorption

[155] Attempt to attribute the compound to the chromatographic peak based on the literature.

[156] Attribution carried out using reference standards.

[157] Attempt at attribution based on the standard aglycone and by comparison with the literature.

maxima of the compounds, as the samples were not subjected to any mass spectrometry analysis.

The 18 compounds were then considered as response variables to be optimised, and this was done by performing a PCA on the autoscaled data matrix, consisting of 33 rows, each representing a chamomile extract analysed as specified in the experimental design, and 18 columns, namely 18 chromatographic variables, as shown in *Figure 58* which reports the biplot obtained by projecting scores and loadings into the same subspace.

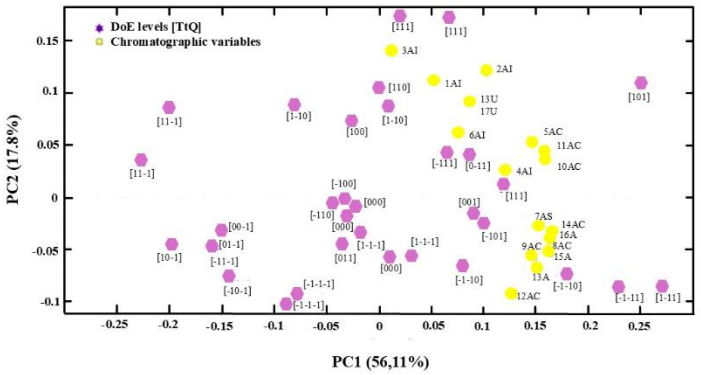


Figure 58. Biplot obtained from PCA of HPLC data

Detailed observation of the biplot allows to evaluate the correlation trends between samples and identify which variables most influence these correlations, enabling us to synthesise the 18 chromatographic variables into a single representative response in the DoE because there is a common trend, or main direction of variation, that accurately summarises the responses of the samples.

It can be observed that most of the replicated samples tend to be close in the subspace defined by PC1 and PC2 compared to samples obtained by varying the levels of the experimental variables.

Analysis of the biplot in *Figure 58* shows that all 18 chromatographic areas (indicated by yellow symbols) are characterised by positive loadings along the first principal component (PC1). The latter explains 56.1% of the total variance, suggesting that samples with high and positive scores along the first principal component (PC1) contain higher amounts of all the compounds

examined than samples with lower or negative scores located on the left side of the graph, on this axis (PC1), which have lower amounts.

The second principal component (PC2), which represents a direction complementary to PC1, explains 17.8% of the total variance and provides further useful information for better understanding the distribution of samples within the subspace. This second component better reflects the differences in the relative composition of chamomile extracts, and in fact, the trend of the scores along PC2 allows the samples to be distinguished according to the different proportions of the various classes of polyphenols of which they are composed, rather than by total quantity, as was the case for PC1. PC1 mainly describes an overall trend in the quantitative variations of the compounds analysed, while PC2 is useful for a qualitative assessment that highlights the variation in the internal composition of the extracts. This proved useful for differentiating samples which, despite having a similar overall polyphenolic content, were discriminated by PC2 due to the specific chemical families that characterised them.

From the biplot analysis, for the optimisation phases, the PC1 scores obtained after projecting the samples onto PC1 were selected as the only response variable. This choice was made considering the fact that almost all response variables, except 1AI and 3AI, show positive and high loadings on PC1. For this reason, the linear combination of the chromatographic areas represented by PC1 can be considered an indicator of the overall quality of chamomile extracts. Considering PC1 as a parameter that summarises the overall polyphenolic content, samples with high scores can be considered to be of higher quality in terms of greater total content of the substances of interest.

In this study, the best extract is identified in sample C11, which shows higher and more positive scores on PC1. C11 is the sample that occupies a position at the top of the experimental domain, thus at the extreme end of the domain explored, characterised by the use of a maximum quantity of chamomile, i.e. 2.5 g, extracted at a temperature of 25°C for the shortest extraction time (32 minutes).

Following the information obtained from the preliminary multivariate exploratory analysis, a regression model was constructed to further explore and understand the effects of the experimental variables within the studied range. *Figure 59a* presents a bar chart highlighting the direction, statistical significance of the estimated effects and their magnitude. The direction distinguishes between the positive and negative impacts of the factors on the

responses. The statistical significance of each term, shown in the diagram by asterisks (*Figure 59a*), provides an immediate indication of which factors have a significant effect in the model and which can be considered marginal. It is also possible to assess the uncertainty in the estimation of the coefficients (shown in green in *Figure 59a*). This dual combination allows for a clearer interpretation of the impact of individual terms on the overall behaviour of the system and an assessment of the accuracy of whether an effect can be considered statistically significant or not. Consequently, all terms that were not statistically significant were not considered in the construction of the model according to the hierarchical rules of simplification, as was the case for the terms t^2, Q^2 and the t^*Q interaction[158]. This helps to reduce the excessive complexity of the model and the risk of overfitting, resulting in a more robust and effective model for describing the response trend in the experimental domain under consideration (so as not to include terms with no relevant statistical significance in the model).

The graph shown in *Figure 59b* shows the three-dimensional response surface graph, which allows for a quick and clear visualisation of the model's predictive trend. This graphical representation shows how, by varying time and temperature while keeping the amount of chamomile at the maximum level constant, it is possible to observe the trend of the predicted response in the experimental space, identifying the combination of conditions that maximise the response, or the overall polyphenol content in the extract. This visualisation allows the identification of the region of the experimental domain where the highest response values are obtained.

[158] N. R. Draper and H. Smith, *Applied Regression Analysis*, 3rd ed. New York, NY, USA: John Wiley & Sons, 1998.

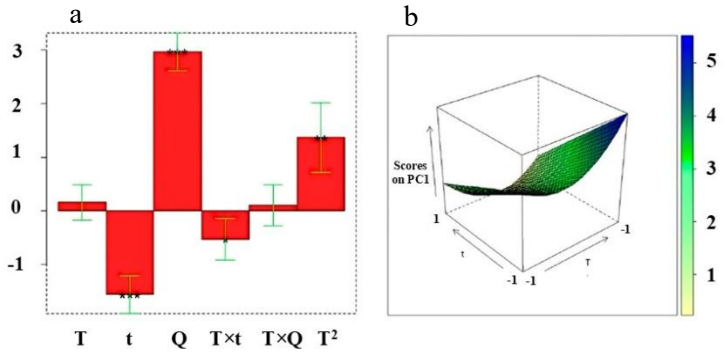


Figure 59. a) bar plot of estimated coefficients effects, showing their direction, statistical significance and uncertainty; b) 3D response surface illustrating the effect of time and temperature on the predicted polyphenol content

Looking at the bar plot, we can confirm that only quantity Q and time t showed a significant linear effect on the response, while temperature T influenced the response through the second-order term and its interaction with time. The response surface graphically identified the optimal extraction conditions T=25°C, t=32 min and Q=2.5g, which correspond to a score higher than 5.

The optimal extraction conditions yielded a response value greater than 5. It is important to note that this value does not coincide with that reported on the PC1 axis of the biplot in *Figure 58*, as the latter is the result of a scaling necessary to simultaneously represent scores and loadings in the same subspace without altering the relative distances between the samples. The model provides a corrected R^2 of 82.87%. To verify the reliability of the model and reduce the risk of overfitting, an ANOVA test was performed, the results of which are reported in *Table 23*. This includes both the significance of the regression and the assessment of the lack of fit, which evaluates the discrepancy not explained by the model with respect to experimental variability.

Origin	DF	Sum of Squares	Mean Squares	F-Ratio
Regression	6	253.357	42.226	26.735
Residuals (E)	26	41.066	1.579	Prob
Total	32	294.423		<0.0001
Lack of Fit	18	36.421	2.023	36.421
Pooled Error	8	5.006	6.626	Prob
Residuals	26	41.427		0.045

Table 23. ANOVA results

It has thus been demonstrated that the model not only identifies an optimum but is also statistically reliable. According to the ANOVA analysis values reported in *Table 23*, the very low p-value (<0.0001) indicates that the regression model is highly significant, explaining a substantial portion of the variability in the data.

This suggests that the mathematical model fits the experimental data well. However, from the p-value of 0.045 for lack of fit, it can be inferred that the model contains an unexplained error contribution and, consequently, a significant marginal lack of fit.

Using this model, the influence of experimental factors on the polyphenolic profile of chamomile extracts was evaluated, which would require mass spectrometry analysis for accurate identification.

3.2 UV-Vis analysis results

The chamomile extracts, prepared according to the list of experiments designed with the DoE, were further characterised using a rapid, inexpensive and untargeted UV-Vis method.

In parallel with the HPLC-DAD analysis, a separate aliquot of the extract was diluted 140-fold in water and subsequently analysed by UV-Vis spectroscopy over the wavelength range 220-400 nm. The spectra were pretreated according to SNV to reduce scattering and baseline shift effects, in order to ensure a more accurate and reliable analysis.

Figure 60 shows the UV-Vis spectra of the chamomile extracts acquired.

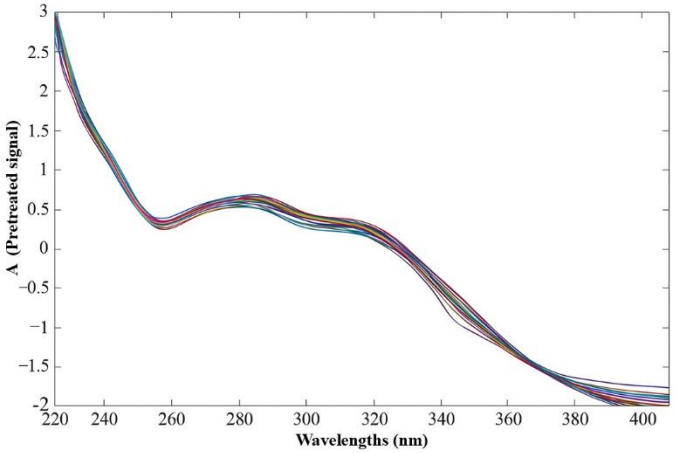


Figure 60. All the UV-Vis spectra of the produced extracts pretreated with SNV

The application of the ASCA algorithm provided interpretation of the multivariate signal, allowing us to understand the impact of experimental factors on the spectroscopic profile.

Figure 61 shows the ASCA results for the effect of temperature.

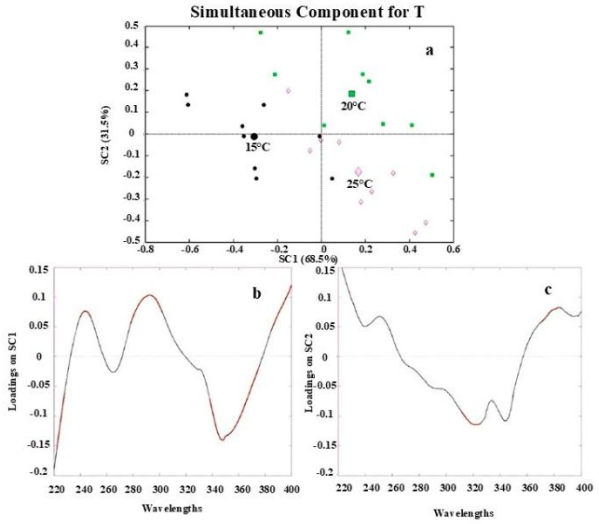


Figure 61. ASCA results for the effect of temperature. a) score plot for the effect of temperature; b) loadings profile for the loadings plot on SC1; c) loadings profile for the loadings plot on

In particular, *Figure 61a* shows the score plot for the experimental factor of temperature, which explains 16% of the total variance in the data. In the figure, each symbol larger than the others represents the average effect for each temperature level, while the smaller symbols indicate the residual of each sample with respect to the average effect.

The score plot highlights a distinct cluster for extracts obtained at 15°C, which have negative scores on SC1, compared to extracts at 20°C and 25°C, which have positive scores. This trend indicates a significant difference in the spectroscopic profiles as a function of the extraction temperature.

Observation of the loadings plot on SC1 (*Figure 61b*) shows which spectroscopic variables are significantly influenced by temperature (highlighted in red), evaluated through 100 bootstrapping cycles. The variables with negative loadings in the range between 330 nm and 370 nm show greater intensity in extractions at 15°C. This suggests that lower temperatures are more effective in solubilising flavones and flavonols, which are probably highly glycosylated and therefore have good solubility in water.

Contrary to this, most samples extracted at 20°C and 25°C have positive scores along SC1. This suggests greater absorbance signal intensity around 250 nm and 290 nm, regions of the spectrum in which there are positive loadings on SC1 at the characteristic wavelengths of hydroxybenzoic acid derivatives. Furthermore, according to *Figure 61c*, the extracts at 25°C, which have negative scores on SC2, are richer in hydroxycinnamic acid derivatives than those at 20°C, as evidenced by the significant negative loadings on SC2 centred at 320 nm. Based on these observations, it can be said that temperature has a substantial influence on the polyphenolic profile of chamomile extracts. In fact, extraction at 15°C favours the extraction of flavones and glycosylated flavonols, while higher temperatures such as 25°C increase the extraction of hydroxybenzoic and hydroxycinnamic acid derivatives, diversifying the polyphenolic classes present. This highlights how temperature variation significantly influences the quantity and type of polyphenolic compounds present in the extracts, allowing products rich in different chemical substances to be obtained according to requirements.

Figure 62, on the other hand, illustrates the influence of the extraction time effect.

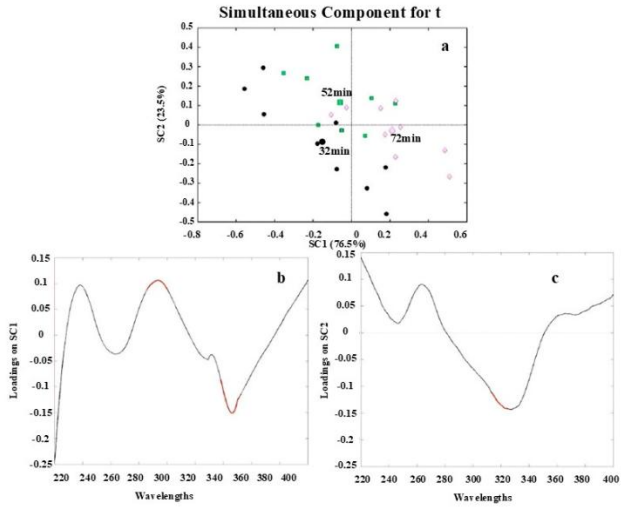


Figure 62. ASCA results for the effect of time. a) score plot for the effect of time; b) loadings profile for the loadings plot on SC1; c) loadings profile for the loadings plot on SC2.

As regards the extraction time effect, the samples extracted for 32 min and 52 min are rich in flavonoids, as indicated by the negative loadings along SC1, in the regions highlighted in red in *Figure 62b*, compared to the samples extracted for 72 minutes, which instead have a higher quantity of hydroxycinnamic acids (positive loadings). SC2 (*Figure 62a*) differentiates the samples extracted for 32 minutes from those extracted for 52 minutes, with the former being richer in hydroxycinnamic acids (in red in panel *c* of *Figure 62*) than the samples extracted for 52 minutes.

For ASCA and UV-Vis analysis, the optimal conditions reported above can be interpreted as follows: 25°C is better than 20°C because it is richer in hydroxycinnamic acid derivatives. 15°C and 25°C generate extracts with different relative amounts of polyphenolic compounds (15°C in flavones and flavonols and 25°C in hydroxybenzoic acid derivatives). Consequently, the maximum amount of chamomile extracted at room temperature (25°C) for 32 minutes seems to represent the best compromise for obtaining a high-quality extract, characterised by a good amount of the main polyphenolic classes present in chamomile, such as flavonols (more abundant in extracts at 32 min), hydroxycinnamic acids and hydroxybenzoic acids.

The aim of this work was to optimise the conditions for extraction in cold water, with the aim of preserving the water-soluble compounds in chamomile.

The results obtained following HPLC-DAD analysis of chamomile extracts confirmed the possibility of obtaining a rich polyphenolic profile even at room temperature with an extraction time and concentration four times lower than in other studies[159] [160], showing how T, t and their interaction significantly influence the polyphenolic profile of chamomile extracts[161].

Further qualitative information emerged following ASCA analysis applied to UV-Vis spectroscopic profiles. For the first time, untargeted multifactorial and multivariate data were used to characterise and evaluate the effects of experimental variables on the spectroscopic profile of aqueous chamomile extracts.

3.3 DPPH assay and IC_{50} results

From the results obtained following the DPPH assay conducted to evaluate the antioxidant activity of chamomile extracts prepared under optimised conditions, it can be observed that, for the five dilutions prepared, the relative decreases in absorbance are those reported in *Table 24*:

Concentration (mg·mL-1)	***% decrease in absorbance[162]***
1.50	25±3
1.88	28±1
3.00	36.6±0.3
3.75	50±3
7.50	90.5±0.3

Table 24. Percentage values of the decrease in absorbance obtained for the five solutions analysed

Based on these data, a least squares regression was performed, obtaining an R^2 value of 0.992, and the IC_{50} value of 3.85 ±0.3 mg·mL^{-1} was obtained

[159] N. S. Sotiropoulou, S. F. Megremi, and P. Tarantilis, "Evaluation of antioxidant activity, toxicity, and phenolic profile of aqueous extracts of chamomile (*Matricaria chamomilla* L.) and sage (*Salvia officinalis* L.) prepared at different temperatures," *Appl. Sci.*, vol. 10, no. 7, p. 2270, 2020.

[160] N. Karaaslan Ayhan, "Investigation of antioxidant properties of chamomile consumed as herbal tea," *J. Food Process. Preserv.*, vol. 45, p. e15327, 2021.

[161] J. Šic Žlabur, I. Žutić, S. Radman, M. Pleša, M. Brnčić, F. J. Barba, G. Rocchetti, L. Lucini, J. M. Lorenzo, R. Domínguez, et al., "Effect of different green extraction methods and solvents on bioactive components of *Matricaria chamomilla* L. flowers," *Molecules*, vol. 25, no. 4, Art. no. 810, 2020.

[162] Values given as percentage decrease in absorbance ± standard error on estimate.

by interpolation. From this value, it can be stated that a chamomile extract at a concentration of 3.85±0.3 $mg \cdot mL^{-1}$ is capable of inhibiting 50% of the antioxidant capacity of the DPPH radical.

From the results of the DPPH assay, it was confirmed that extracts prepared at 25°C for 32 minutes with 2.5g of dried, ground and sieved chamomile showed substantial antioxidant activity, reaching an IC_{50} value of 3.8 $mg \cdot mL^{-1}$.

The optimal extraction conditions, identified as 2.5 g of chamomile extracted for 32 minutes at 25°C, produced extracts rich in polyphenolic compounds such as hydroxybenzoic acids, hydroxycinnamic acids and flavonol derivatives. This represents an ideal compromise for the extraction of the main water-soluble polyphenolic classes. UV-Vis analysis provided an overview of the absorbances influenced by the experimental variables without the need to identify individual compounds. The spectra proved useful in the qualitative evaluation of the different classes of polyphenols present and in the interpretation of compositional trends in the extracts.

The combination of these techniques with chemometric analysis highlighted the significant impact of temperature, time and their interaction on the polyphenolic profile.

It emerged that lower temperatures favour the solubilisation of glycosylated flavones and flavonols, while higher temperatures lead to an increase in hydroxybenzoic and hydroxycinnamic acids in the extracts. Therefore, extraction at 25°C for a short time allows extracts with greater antioxidant activity to be obtained.

Section 3

Pepper

CHAPTER 6

COMPARISON OF THE ANTIOXIDANT ACTIVITY OF DIFFERENT VARIETIES OF PEPPER

SUMMARY: 1. Aim of the work based on the pepper varieties – 2. Method of conducting the experiment – 2.1 DPPH assay – 2.2 IR analysis – 2.3 GC-MS analysis – 3. Experimental results and interpretations – 3.1 Analysis of antioxidant activity and IC_{50} results – 3.2 Evaluation of IR spectra – 3.3 Volatile fraction analysis – 4. Summary of results

1. Aim of the work based on the pepper varieties

Black pepper (*Piper nigrum* L.) is one of the species belonging to the Piperaceae family and is among the most widely used and distributed spices in the world.

The three most common forms, green, white, and black pepper, originate from the same fruit, which is harvested at different stages of ripeness and subjected to distinct processing methods. Green pepper is obtained from unripe berries that are either lyophilized or treated with sulfur dioxide to preserve their bright green colour; black pepper is produced by sun-drying mature green berries for several days, during which polyphenol oxidase reacts with phenolic compounds, darkening the pericarp and giving the final product its characteristic dark colour and intense aroma[163]; finally, white pepper is obtained from fully ripe (red) berries that are soaked in water to facilitate the removal of the pericarp; the remaining seeds are then dried, yielding a lighter product with a more delicate flavor compared to black pepper.

[163] S. Martini, A. Cattivelli, A. Conte, and D. Tagliazucchi, "Black, green, and pink pepper affect differently lipid oxidation during cooking and in vitro digestion of meat," Food Chemistry, vol. 350, p. 129246, 2021.

The interest in studying this plant species arises from the presence of bioactive compounds that play an essential role as therapeutic agents[164]. In fact, *Piper nigrum* L. is an effective medicinal plant traditionally used for the treatment of stomach pain and rheumatoid arthritis.

Analyses of its extracts have revealed the presence of substances with antimicrobial, anti-inflammatory, antimalarial, and antioxidant activities[165].

Piper nigrum L. contains high levels of flavonoids which, together with bioactive compounds such as piperine and phenols, enhance its ability to reduce the presence of free radicals[166].

The antioxidant activity of various pepper samples, characterized to assess potential differences related to their geographical origin, was evaluated using the DPPH radical scavenging assay and the calculation of the IC_{50} value, both performed on hydroalcoholic extracts of ground black pepper.

Species from different geographical areas were used, including high-quality samples such as two from the Kampot region in Cambodia, one with protected geographical indication PGI (Kampot PGI) and the other with organic certification (Kampot), pepper from Madagascar and two additional commercial pepper samples (Commercial 1 and Commercial 2).

The analysis of the volatile profile of the same black pepper varieties, compared with a green pepper sample (Green), revealed a composition rich in terpenes, particularly caryophyllene, 3-carene, limonene, β-pinene, and copaene, which were found to be the most abundant compounds. Seasonal and anatomical variations strongly influence the composition[167], the presence

---

[164] Barh, N. Barve, K. Gupta, S. Chandra, N. Jain, S. Tiwari, N. Leon-Sicairos, A. Canizalez-Roman, A. Rodrigues dos Santos, S. S. Hassan, S. Almeida, R. Thiago Jucá Ramos, V. Augusto Carvalho de Abreu, A. Ribeiro Carneiro, S. de Castro Soares, T. Luiz de Paula Castro, A. Miyoshi, A. Silva, A. Kumar, *et al.*, "Exoproteome and secretome derived broad spectrum novel drug and vaccine candidates in *Vibrio cholerae* targeted by *Piper betel* derived compounds," *PLoS ONE*, vol. 8, no. 1, p. e52773, 2013.

[165] E. E. Mgbeahuruike, T. Yrjönen, H. Vuorela, and Y. Holm, "Bioactive compounds from medicinal plants: Focus on *Piper* species," *S. Afr. J. Bot.*, vol. 112, pp. 54–69, 2017.

[166] N. Pradubyat, T. Wunnakup, R. Praparatana, S. Wongwiwatthananukit, S. Jongrungruangchok, T. Songsak, F. Madaka, and T. Sudsai, "Evaluation of antioxidant and anti-inflammatory properties, bioactive compound profiling, and molecular mechanisms of a multicomponent Thai herbal formulation," *Phytomedicine Plus*, vol. 4, no. 4, p. 100662, 2024.

[167] B. Feitosa, O. O. Ferreira, C. Franco, J. P. De, H. Karakoti, R. Kumar, M. M. Cascaes, R. D. Jawarkar, S. N. Mali, J. N. Cruz, I. C. De Menezes, M. S. De Oliveira, and E. H. D. A. Andrade, "Chemical composition of *Piper nigrum* L. cultivar guajarina essential oils and their biological activity," Molecules, vol. 29, no. 5, p. 947, 2024.

of bioactive compounds, and consequently the beneficial properties for humans[168].

Due to the frequent economic adulterations, it is essential to ensure the authenticity and accurate characterization of high-value foods such as black pepper.

Therefore, samples of black pepper from different geographical regions were characterized using IR spectroscopy and analysis of the aromatic profile composition through HS-SPME followed by GC-MS. These two analyses were performed on five varieties of black pepper and one green pepper sample for comparison. Finally, chemometric modeling of the data was carried out using PCA and PLS-DA in order to identify specific patterns related to the geographical origin of the samples.

2. Method of conducting the experiment

2.1 DPPH assay

The pepper extracts were prepared by grinding the grains with a ceramic mortar. The powder obtained was then sieved with a 1 mm sieve, then 250 mg of powder was weighed and placed in a 20 mL calibrated flask. To this, 10 mL of the extracting mixture consisting of water and ethanol in a 60:40 ratio was added. Protected from light and capped, the sample was subjected to magnetic stirring in a thermostatic circulating water bath in order to maintain a constant temperature of 60 °C for the entire duration of the extraction (45 min). Once this process was complete, the samples were centrifuged for 5 min at 1100 rpm and filtered through a nylon membrane with a pore diameter of 0.22 µm.

Four dilutions were prepared for each extract and analysed together with the undiluted solution. The solutions were subjected to the DPPH assay and analysed using a UV-Vis spectrophotometer. All measurements were performed in duplicate. Subsequently, the IC_{50} values with their respective

[168] K. Charoensedtasin, W. Kheansaard, S. Roytrakul, and D. Tanyong, "Piperine, a black pepper compound, induces autophagy and cellular senescence mediated by NF-κB and IL-6 in acute leukemia," BMC Complementary Medicine and Therapies, vol. 24, no. 1, p. 343, 2024

standard deviations from the mean were estimated from the dose-response curve equation.

2.2 IR analysis

Infrared spectra were acquired using the ATR-FT-IR technique in the spectral range of 4000-500 cm^{-1}. Before conducting the analysis, a background spectrum was recorded by exposing the diamond crystal to air after cleaning it with methanol. For each variety, 100 spectra were collected on individual pepper grains, resulting in a total of 600 acquisitions. After each scan, the ATR crystal was cleaned with absorbent paper soaked in methanol before acquiring the next spectrum of the peppercorns.

2.3 GC-MS analysis

All analyses were performed using the GC-MS Trace 1300 ISQ LT system (Thermo Fisher Scientific, Waltham, MA, USA) operating in splitless mode. For headspace sampling, peppercorns were ground using a ceramic laboratory mortar to obtain a homogeneous sample. A portion of 200 mg of ground pepper was placed in a 4 mL glass vial sealed with a Mininert cap suitable for SPME analysis. The vials were then placed in an oil bath maintained at 40 °C under constant magnetic stirring to promote equilibrium between the solid and gaseous phases. The extraction of volatile compounds was performed by headspace solid-phase microextraction (HS-SPME) using a fiber coated with polydimethylsiloxane (PDMS, 100 µm thickness; Supelco, Bellefonte, PA, USA), which was exposed to the headspace of the sample for 5 minutes. After extraction, the fiber was retracted into the needle, removed from the vial, and inserted into the GC injection port for thermal desorption at 280 °C for 5 minutes. Five replicates were carried out for each type of pepper. After each analysis, the fiber was reconditioned in the GC injection port at 280 °C for 15 minutes, followed by a blank run to verify the cleanliness of the sorbent and to prevent carry-over phenomena. Helium (5.5 IP) was used as the carrier gas at a constant flow rate of 1.1 mL/min. The oven temperature program was as follows: initial temperature of 40 °C held for 5 minutes, increased to 100 °C at 25 °C/min, then to 150 °C at 12 °C/min, to 165 °C at 15 °C/min, to 200 °C

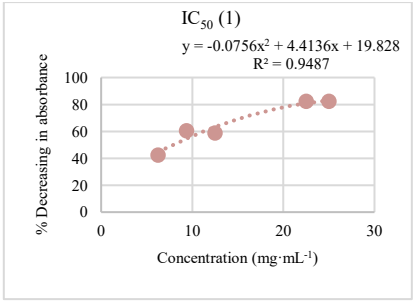
at 5 °C/min, and finally to 270 °C at 125 °C/min, held for 2 minutes. Volatile compounds were identified by comparing their mass spectra with those in the NIST 14 library and by calculating their linear retention indices using a mixture of n-alkanes from C7 to C40 injected under the same chromatographic conditions.

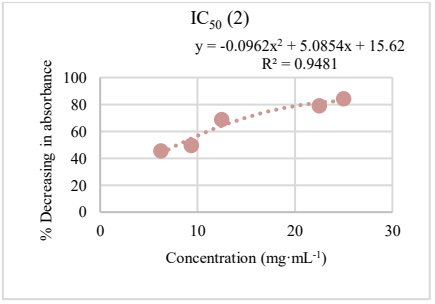
3. Experimental results and interpretations

3.1 Analysis of antioxidant activity and IC_{50} results

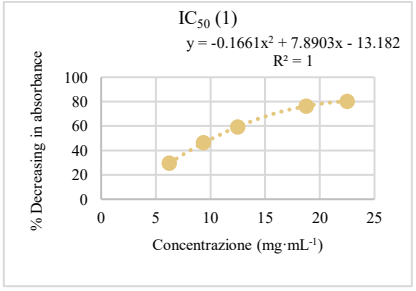
The dose-response curves for the pepper samples studied are shown below, showing the percentage of radical inhibition as a function of sample concentration. The IC_{50} values for the two replicates were calculated from the corresponding regression equations of the fitted mathematical models and finally averaged.

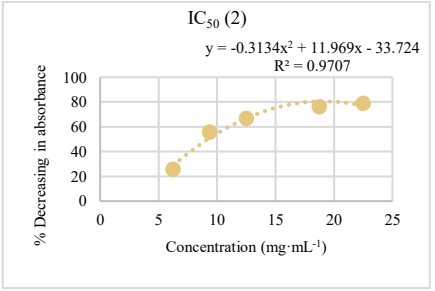
(a)



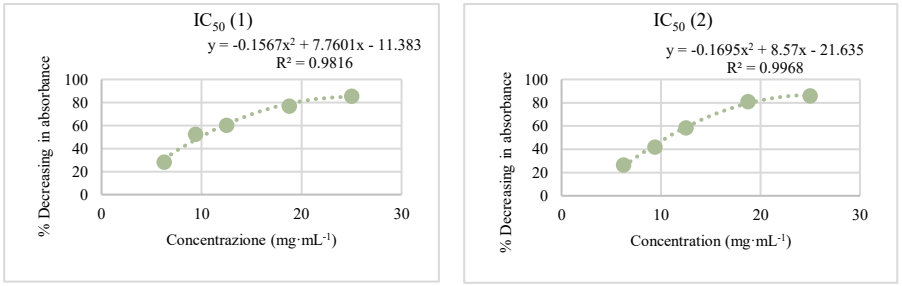


(b)

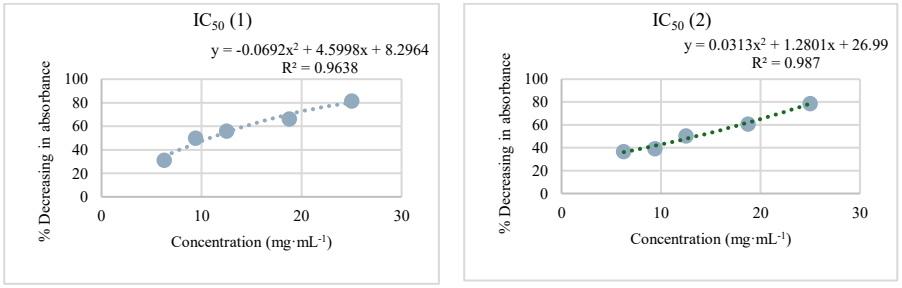




(c)



(d)



(e)

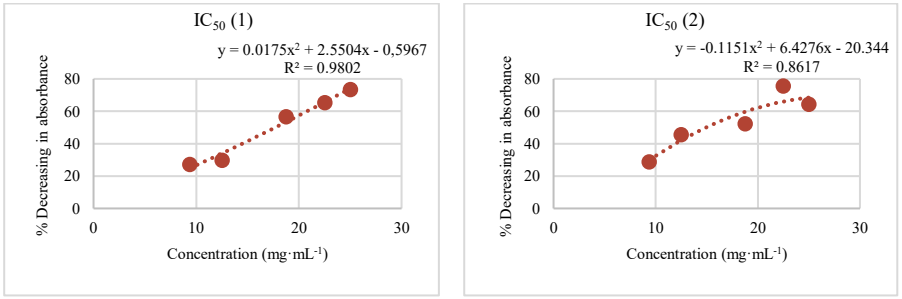


Figure 63. Dose-response curves used to determine the IC_{50} value (mean ± standard deviation) of Madagascar pepper (a), Commercial 1 (b), Kampot PGI (c), Commercial 2 (d) and Kampot (e)

Table 25 shows the results obtained in terms of dry pepper concentration, reflecting the antioxidant potential.

Samples of pepper	IC_{50} (mg·mL^{-1})
Madagascar	7.94 ± 0.04
Commercial 1	9.71 ± 0.69
Kampot PGI	10.23 ± 0.49
Commercial 2	12.11 ± 1.89
Kampot	16.32 ± 1.94

Table 25. *IC_{50}* results for pepper varieties

Based on the interpretation of the calculated IC_{50} values, it can be stated that Madagascar pepper showed the lowest IC_{50} value (7.94 ± 0.04 mg·mL^{-1}), making it the most effective variety in terms of inhibitory activity among those tested. Conversely, Kampot pepper showed the highest IC_{50} value (16.32 ± 1.94 mg·mL^{-1}), proving to be the variety with the lowest antioxidant efficacy compared to the other varieties analysed, which showed intermediate IC_{50} values.

These results suggest that factors such as the processing method, geographical origin (including soil, climate, cultivation methods), post-harvest treatment techniques and storage conditions significantly influence the biological activity of pepper, making some varieties more concentrated in bioactive compounds, responsible for antioxidant activity, than others. In particular, Madagascar pepper, with the highest antioxidant efficacy, can be considered richer in bioactive compounds than other varieties.

Antioxidant activity, evaluated using the DPPH assay, showed marked differences between varieties, with Madagascar pepper showing the greatest radical scavenging capacity, suggesting a higher content of bioactive compounds. Among the different varieties, the choice of Cambodian pepper was dictated by the fact that it is one of the most prized pepper varieties in the world, recognised as a PGI, while Madagascar pepper, although not certified, is highly appreciated by consumers and is therefore its direct competitor in terms of commercial relevance.

3.2 Evaluation of IR spectra

To distinguish the different pepper samples according to their geographical origin, a chemometric classification analysis was carried out using the PLS-DA algorithm. This approach enabled the classification of the black and

the green pepper samples based on the data obtained from the spectroscopic analyses.

Figure 64 presents the average raw ATR-FT-IR spectra obtained for each class of analysed samples.

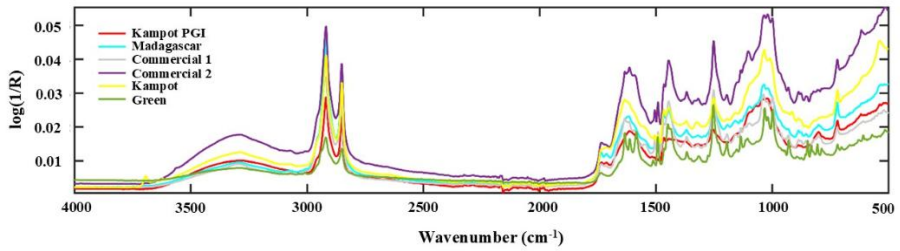


Figure 64. Average raw spectra of the analysed pepper classes

The interpretation of the ATR-FTIR spectra focuses on the main spectral regions characteristic of the functional groups present in the pepper samples.

In the spectral region between 3700 and 3000 cm^{-1}, a broad and weak absorption band centred around 3300 cm^{-1} can be observed, attributed to O–H stretching vibrations typical of hydroxyl groups found in alcohols and carboxylic acids.

Between 3000 and 2800 cm^{-1}, two distinct peaks appear within the 2950-2850 cm^{-1} range, corresponding to aliphatic C–H stretching vibrations, which suggest the presence of aliphatic hydrocarbon chains commonly found in terpenes and lipids.

In the 1700-1650 cm^{-1} region, a pronounced peak near 1700 cm^{-1} is characteristic of C=O stretching vibrations of carbonyl groups, associated with compounds such as piperine, the main component responsible for the pungency of black pepper.

The bands observed between 1600 and 1400 cm^{-1} can be related to aromatic ring vibrations, linked to the presence of piperine and other aromatic constituents of the essential oils.

Finally, the region between 1300 and 900 cm^{-1} shows intense bands attributed to C–O–C stretching vibrations characteristic of polysaccharides and starches, which represent structural components commonly found in pepper.

Variations are evident in both the band intensity and slight shifts in their positions, particularly in the 1700-1600 cm^{-1} (C=O stretching) and 1300-900 cm^{-1} (C–O–C vibrations) regions, which may reflect compositional differences related to the geographical origin of the samples.

Although these variations are relatively subtle, they indicate that the spectral features can provide useful information for distinguishing between different types of pepper, consistent with the classification results obtained through the PLS-DA model. The latter was developed by optimising its classification performance through the evaluation of different spectral data preprocessing strategies, as reported in *Table 26*.

Preprocessing	LVs	Average correct classification rate (%CV)
MC	12	81.5
SNV	12	89.3
D1	12	90.5
D2	12	90.5
SNV+D1	11	91.0
SNV+D2	12	90.0

Table 26. Cross-validated (CV) mean correct classification rates (%) and corresponding number of latent variables (LVs) for calibration models obtained using different spectral preprocessing methods.

The preprocessing methods assessed comprised SNV, D1, D2, and the combined techniques SNV+D1 and SNV+D2.

For each of these approaches, a PLS-DA model was developed using 7-fold cross-validation on the training dataset, and the classification error rates were subsequently evaluated.

The best classification performance was achieved using the Standard Normal Variate combined with the first derivative (SNV+D1), which resulted in an average correct classification rate of 91.0% in cross-validation. *Table 27* reports the confusion matrix and the correct classification rates obtained from the application of the model to the external test set, which resulted in a global classification error of 5.0%.

	Kampot PGI	Madagascar	Commercial 1	Commercial 2	Kampot	Green	% Correct classification rate
Predicted Kampot PGI	63	1	0	0	0	0	98.4
Predicted Madagascar	0	56	4	1	0	0	91.8
Predicted Commercial 1	0	0	69	0	0	1	98.6
Predicted Commercial 2	0	3	5	52	0	2	83.9
Predicted Kampot	0	0	2	0	66	0	97.0
Predicted Green	0	0	8	0	0	69	89.6

Table 27. Confusion matrix and Correct Classification rates in prediction on the external test set

The results highlight the importance of preprocessing steps in spectroscopic classification processes and confirm that the SNV+D1 combination represents the most effective approach for the geographical authentication of pepper based on IR analysis.

Figure 65 shows the projection of the samples in the space defined by the first three canonical variates (CV1, CV2, and CV3) extracted from the PLS-DA model developed using the preprocessed ATR-FT-IR spectra.

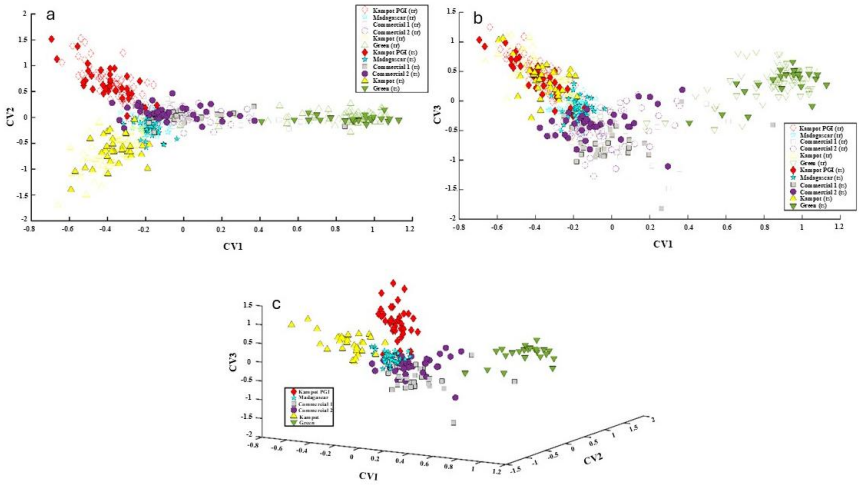


Figure 65. Representation of the samples projected in the spaces defined by (a) the first and second canonical variates (CV1 and CV2); (b) the first and third canonical variates (CV1 and CV3); and (c) all extracted canonical variates obtained from the PLS-DA model. The sample assignment is indicated by *tr* for the training set and *ts* for the test set.

Figure 65a shows the projection of the samples onto the plane defined by the canonical variates CV1 and CV2. It can be observed that both Kampot pepper classes, regardless of their PGI certification, are mainly clustered at negative CV1 values, whereas the other samples are distributed at positive values of the same component.

The remaining samples exhibit greater dispersion along the CV1 axis, with the green pepper samples located at the extreme positive side.

Figure 65b displays the projection of the samples onto the CV1-CV3 plane. In this case, a partial overlap is observed between the two Kampot classes (both showing positive CV1 values) and between the commercial and Madagascar samples, which tend to cluster around the origin of the axes. As expected, the green pepper samples form a distinct and well-separated group, positioned at negative CV1 values and clearly detached from the others.

In the three-dimensional plot (*Figure 65c*), the distribution of the samples reveals a clear separation among the different classes, confirming the ability of the PLS-DA model to capture compositional differences relevant to classification.

The 3D plot is consistent with the trends observed in the CV1-CV2 and CV1-CV3 projections: commercial samples tend to cluster near the origin or at positive CV1 values, close to the Madagascar samples, while those from Cambodia exhibit negative values for the same variate.

The green pepper samples remain clearly distinguishable from all other categories, confirming their distinct spectral composition.

The limited overlap observed between the classes, together with the compact grouping within each class, particularly evident for high-quality, single-origin samples, highlights the ability of the PLS-DA model to extract geographical information from the IR spectral data. These results confirm the value of combining vibrational spectroscopy with supervised multivariate modelling for the reliable classification and geographical authentication of pepper.

3.3 Volatile fraction analysis

The chromatographic profiles obtained through GC-MS analysis revealed a general similarity among the different samples, while showing subtle yet significant variations in the relative abundances of certain specific compounds. In total, 44 volatile compounds were identified, as reported in *Table 28*, all belonging to the terpenoid family, which includes terpene hydrocarbons as well as oxygenated and aromatic derivatives.

			Retention Index		Commercial 1	Commercial 2	Madgascar	Kampot PGI	Kampot	Green Pepper
n	**RT min**	**Compound**	**[1]Exp**	**[2]Lit**	**Mean A%±se**	**Mean A%±se**	**Mean A%±se**	**Mean A%±se**	**Mean A%±se**	**Mean A%±se**
1	7.68	α-Thujene	930	929±2	1.45±0.23	4.65±0.29	0.25±0.03	0.31±0.04	0.33±0.05	3.16±0.21
2	7.79	α-Pinene	938	937±3	9.08±0.14	8.42±0.23	12.51±0.30	6.57±0.22	7.25±0.52	6.41±0.13
3	7.98	Camphene	954	952±2	0.54±0.03	0.39±0.03	1.26±0.11	0.27±0.02	0.32±0.06	0.19±0.02
4	8.29	Sabinene	976	974±2	7.08±0.94	16.30±0.37	1.10±0.27	0.55±0.04	0.64±0.11	14.44±0.53
5	8.38	β-Pinene	979	979±2	12.40±0.07	11.41±0.54	14.27±0.20	9.28±0.25	10.11±0.68	6.77±0.14
6	8.46	β-Myrcene	992	991±2	6.01±0.40	4.56±0.08	6.68±0.35	5.35±0.34	6.04±0.89	2.48±0.13
7	8.69	α-Phellandrene	1007	1005±2	2.22±0.48	2.21±0.13	2.49±0.40	2.20±0.26	2.60±0.22	2.68±0.22
8	8.72	3-Carene	1014	1011±2	17.40±0.78	7.34±0.74	16.38±0.30	20.49±0.49	19.05±0.52	8.32±0.12
9	8.79	α-Terpinene	1021	1017±2	0.35±0.03	0.37±0.05	0.28±0.04	0.21±0.02	0.26±0.04	0.26±0.02
10	8.83	o-Cymene	1024	1022±2	0.35±0.01	0.022±0.001	0.08±0.01	0.29±0.01	0.23±0.01	0.002±0.001
11	8.88	p-Cymene	1029	1025±2	3.22±0.35	0.62±0.10	1.13±0.11	3.27±0.18	2.72±0.33	0.28±0.01
12	8.97	Limonene	1033	1030±2	16.91±0.28	16.09±0.72	16.82±0.29	13.64±0.51	12.35±0.77	10.23±0.25
13	9.02	β-Phellandrene	1041	1031±2	0.97±0.10	2.85±0.56	2.19±0.14	1.12±0.45	2.75±0.30	3.36±0.08
14	9.10	trans-β-Ocimene	1048	1049±2	0.22±0.04	0.70±0.09	0.49±0.05	0.18±0.01	0.20±0.03	0.41±0.03
15	9.26	γ-Terpinene	1063	1060±3	0.47±0.01	0.69±0.04	0.50±0.04	0.45±0.03	0.54±0.07	0.59±0.04
16	9.40	Sabinene hydrate	1075	1077±9	0.09±0.02	0.34±0.04	0.018±0.001	0.012±0.002	0.008±0.001	0.232±0.004
17	9.52	p-Mentha-2.4(8)-diene	1086	1086±3	0.92±0.10	0.24±0.04	0.95±0.05	1.12±0.07	1.23±0.13	0.24±0.01
18	9.57	p-Mentha-1.4(8)-diene	1090	1088±2	1.80±0.16	0.92±0.06	2.19±0.09	2.09±0.12	2.43±0.23	0.78±0.03
19	9.61	1-methyl-4-(1-methylethenyl)-benzene	1094	1090±2	0.04±0.01	0.014±0.002	0.030±0.001	0.07±0.01	0.041±0.004	0.004±0.001
20	9.68	Linalool	1100	1099±2	0.20±0.02	0.16±0.02	0.29±0.02	0.84±0.03	0.73±0.05	0.358±0.005
21	9.75	cis-β-Terpineol	1106	1144±1	0.06±0.01	0.17±0.02	0.009±0.001	0.011±0.001	0.015±0.003	0.16±0.01
22	10.30	Camphor	1156	1145±2	0.043±0.004	0.009±0.001	0.060±0.003	0.003±0.001	0.004±0.001	0.006±0.001
23	10.65	Terpinen-4-ol	1188	1182±0	0.07±0.02	0.20±0.02	0.018±0.001	0.022±0.002	0.020±0.001	0.42±0.01
24	10.79	α-Terpineol	1201	1190±3	0.042±0.005	0.036±0.003	0.09±0.01	0.028±0.002	0.024±0.002	0.060±0.005
25	12.31	δ-Elemene	1344	1338±2	1.43±0.02	1.12±0.12	2.38±0.12	1.20±0.08	1.16±0.19	3.90±0.10
26	12.43	α-Cubebene	1355	1351±2	0.22±0.02	0.21±0.02	0.08± 0.01	0.27±0.02	0.13±0.02	0.046±0.004

27	12.71	Isoledene	1383	1375±2	0.09±0.01	0.09±0.01	0.15±0.01	0.052±0.003	0.03±0.01	0.056±0.003
28	12.75	α-Copaene	1386	1376±2	2.66±0.20	2.34±0.22	0.09±0.02	0.35±0.02	0.18±0.03	0.11±0.01
29	12.87	β-Elemene	1398	1391±2	0.66±0.05	0.85±0.04	0.80±0.06	2.03±0.14	1.12±0.19	1.12±0.04
30	13.10	α-Gurjunene	1404	1409±2	0.09±0.01	0.21±0.03	0.09±0.01	0.45±0.02	0.22±0.05	0.16±0.01
31	13.30	β-Caryophyllene	1419	1419±3	10.34±0.58	12.41±0.62	11.91±0.78	18.04±0.88	20.70±2.57	25.25±0.98
32	13.37	α-Guaiene	1444	1439±2	0.12±0.01	0.24±0.02	0.58±0.05	1.28±0.08	1.12±0.19	0.047±0.004
33	13.45	(E)-β-Famesene	1451	1457±2	0.07±0.01	0.15±0.02	0.027±0.002	0.18±0.02	0.10±0.02	0.63±0.03
34	13.55	Cadina-3.5-diene	1461	1458± n/a	0.04±0.01	0.04±0.01	0.011±0.001	0.008±0.001	0.003±0.001	0.017±0.003
35	13.65	α-Caryophyllene	1470	1454±3	0.88±0.06	1.25±0.11	1.29±0.10	2.39±0.15	2.38±0.37	2.36±0.10
36	13.92	Germacrene D	1494	1481±3	0.16±0.04	0.14±0.01	1.12±0.11	0.09±0.01	0.08±0.01	0.72±0.02
37	13.98	(Z.E)-α-Farnesene	1500	1491±3	0.010±0.004	0.07±0.03	n.d.	0.024±0.004	0.02±0.01	0.16±0.01
38	14.03	δ-Guaiene	1504	1505±3	0.25±0.01	0.58±0.06	0.72±0.07	2.46±0.16	1.41±0.27	0.54±0.03
39	14.10	α-Selinene	1510	1494±3	0.22±0.02	0.40±0.05	0.45±0.04	1.63±0.12	0.88±0.18	0.16±0.01
40	14.13	β-Bisabolene	1512	1509±3	0.35±0.07	0.55±0.06	0.04±0.01	0.77±0.05	0.39±0.07	1.18±0.07
41	14.29	δ-Cadinene	1525	1524±2	0.35±0.04	0.33±0.03	0.028±0.004	0.032±0.005	0.015±0.002	0.027±0.003
42	14.38	α-Panasinsene	1533	1527±8	0.010±0.001	0.08±0.01	0.023±0.004	0.15±0.01	0.08±0.02	0.09±0.01
43	14.88	Germacrene B	1573	1557±3	0.02±0.01	0.10±0.02	0.039±0.003	0.027±0.002	0.02±0.01	1.41±0.08
44	15.18	Caryophylene oxide	1598	1581±2	0.069±0.003	0.12±0.03	0.06±0.01	0.19±0.03	0.10±0.02	0.18±0.01

Table 28. Volatile profiles of pepper samples. For each compound, the following information is provided: peak number (n), retention time (RT, min), identified structure, experimental and literature retention indices, and the mean relative peak area (%) ± standar
[1]Exp: experimental retention indices calculated using a C7-C40 linear alkane mixture.
[2]Lit : semistandard non polar retention indices reported in the NIST14 mass spectra database.
n.d: not detected
n/a: not available

Among these, 24 compounds (from n. 1 to 24 in *Table 28*) were classified as monoterpenoids, which were predominant in the aromatic profile of all samples. They accounted for approximately 80% in Commercial 1, Commercial 2, and Madagascar black pepper; about 70% in Kampot and Kampot PGI samples; and 62% in Green Pepper.

The main monoterpenes, i.e., those showing a relative abundance greater than 5% in at least one of the analysed samples, included α-pinene, sabinene, β-pinene, β-myrcene, 3-carene, and limonene.

With regard to sesquiterpenoids, β-caryophyllene was the only compound consistently detected at high concentrations across all samples. Its mean relative abundance was around 12% in Commercial 1, Commercial 2, and Madagascar black pepper samples, while in Kampot, Kampot PGI, and Green Pepper it reached an average of 21%, peaking at 25.25% in the latter.

Each sample exhibited a distinctive aromatic profile. Commercial 2 was characterized by a high sabinene content (16.30%), whereas Commercial 1 showed a marked predominance of 3-carene (17.40%). Madagascar black pepper stood out for its elevated limonene level (16.82%). The Kampot and Kampot PGI samples displayed very similar profiles, both rich in 3-carene and β-caryophyllene, with the latter reaching 20.70% and 18.04%, respectively. Green Pepper, on the other hand, was distinguished by its exceptionally high β-caryophyllene content (25.25%) and a substantial amount of sabinene (14.44%).

Comparatively, β-caryophyllene emerged as the most representative compound across all samples, particularly in Green Pepper. 3-Carene was predominant in Kampot and Commercial 1, but less abundant in Green Pepper. Sabinene occurred at high levels only in Commercial 2 and Green Pepper, while it was nearly absent in the other samples. Both β-pinene and α-pinene were found in moderate amounts across all varieties, with the highest concentrations observed in Madagascar pepper. Limonene showed particularly high values in Commercial 1 and Madagascar, whereas β-myrcene remained consistently low, reaching its minimum level in Green Pepper.

Although the relative abundances of the main volatile compounds provide an overall picture of the aromatic characteristics of the samples, these data alone are not sufficient to clearly distinguish between the different pepper varieties. In fact, the major compounds are common to all analysed samples, and their quantitative variations do not fully reflect the chemical complexity required for reliable discrimination.

To gain a deeper understanding of the possible differences among samples, a larger number of observations would have been desirable to effectively apply supervised modeling techniques. However, given the limited size of the dataset obtained from the GC-MS analysis, it was not possible to perform a

reliable supervised classification such as that applied to the IR data. Therefore, an exploratory approach based on PCA was adopted to identify potential latent patterns and grouping trends among the samples.
The Principal Component Analysis (PCA) was performed on the data matrix comprising the relative percentage areas of the volatile compounds detected in all pepper samples. Even at an exploratory level, the analysis revealed a clear differentiation among the various pepper varieties. Two preprocessing approaches, autoscaling and mean-centering, were evaluated; the PCA scores plot (*Figure 66*) obtained using mean-centering provided the most distinct grouping of samples, suggesting marked differences in the volatile profiles among the analysed peppers.

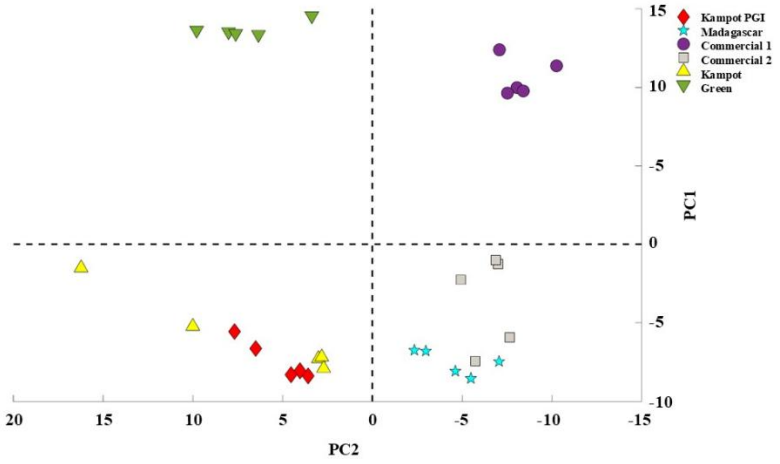


Figure 66. PCA score plot of volatile compounds. Legend: Kampot PGI (red diamonds), Madagascar (cyan stars), Commercial 1 (purple dots), Commercial 2 (gray squares), Kampot (yellow triangles), and Green Pepper (green downward-pointing triangles).

The Kampot and Kampot PGI pepper samples are located in the region characterized by negative PC1 and positive PC2 values, indicating a strong similarity between their aromatic profiles and a clear differentiation from the other groups. The Madagascar and Commercial 2 samples, on the other hand, display negative values for both principal components, suggesting similarities in their volatile profiles. Commercial 1 is positioned in the area with positive

PC1 and negative PC2 values, while Green Pepper stands out with positive scores on both components, confirming its unique aromatic profile.

Overall, clear grouping tendencies can be observed: Kampot samples consistently fall within the region of negative PC1 and positive PC2 values, whereas Madagascar and Commercial samples are mainly distributed in the area of negative PC2 values.

The loadings plot (*Figure 67*) identified the variables, corresponding to the volatile compounds, that contribute most significantly to the separation and clustering of the different samples.

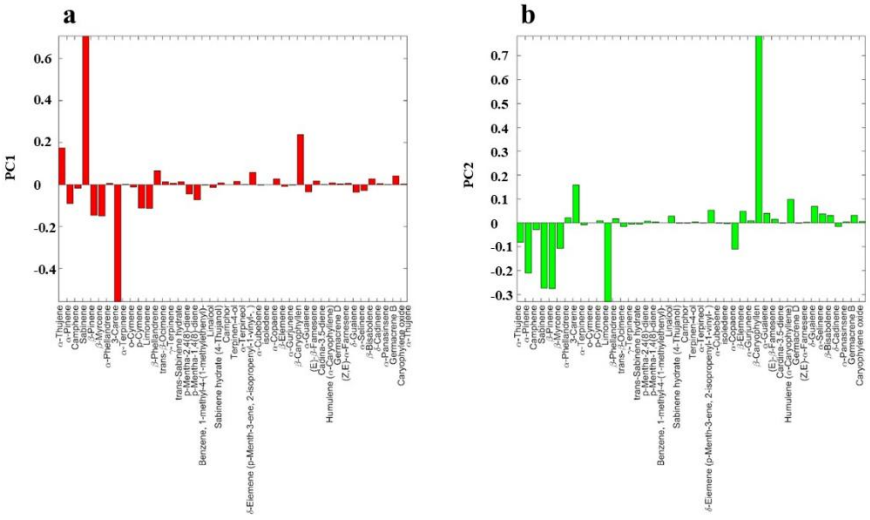


Figure 67. PCA loading plots for volatile compounds, showing PC1 (a) and PC2 (b).

Sabinene is among the compounds that contribute most to the positive scores of PC1 and is therefore associated with the Commercial 1 and Green Pepper samples. Conversely, 3-carene is mainly responsible for the negative PC1 scores, characterizing the Kampot PGI, Kampot, Madagascar, and Commercial 2 samples.

Regarding PC2, β-caryophyllene, which shows positive values, can be considered a potential distinctive marker for the Kampot, Kampot PGI, and Green Pepper samples, while limonene, exhibiting negative values along the same component, is primarily associated with the Commercial 1 samples.

The identification of these markers is particularly interesting in relation to their known biological properties. Sabinene contributes to the pepper's spicy note and has antifungal and anti-inflammatory activity[169]. Limonene is known for its anti-inflammatory, antioxidant, antitumor, gastroprotective, and immunomodulatory properties[170]. Finally, β-caryophyllene exhibits anti-inflammatory, antioxidant, neuroprotective, and analgesic effects[171].

4. Summary of results

In this study, a comprehensive multi-analytical approach was applied to characterize different samples of black pepper (*Piper nigrum* L.) originating from various geographical areas, including Kampot (PGI and organic certified), Madagascar, and two commercial samples, with a green pepper sample included for comparative purposes. The combination of functional, chemical, and spectroscopic analyses provided a detailed overview of the compositional and bioactive characteristics of the samples, as well as their potential for geographical authentication.

The evaluation of antioxidant activity using the DPPH radical scavenging assay revealed significant differences among the samples. The black pepper from Madagascar exhibited the highest antioxidant capacity, suggesting a higher concentration of bioactive constituents such as phenolic and terpenoid compounds. These results highlight the functional quality of this variety and confirm the influence of geographical and processing factors on the accumulation of antioxidant compounds in pepper.

ATR-FT-IR spectroscopy, combined with PLS-DA, proved to be highly effective in classifying the samples according to their geographical origin. Among the tested data preprocessing strategies, the combination of SNV + D1 produced the best results, allowing clear class separation and high predictive accuracy. The final PLS-DA model achieved an overall

---

[169] V. Tyagi, V. K. Singh, P. K. Sharma, and V. Singh, "Essential oil-based nanostructures for inflammation and rheumatoid arthritis," *Journal of Drug Delivery Science and Technology*, vol. 60, p. 101983, 2020.

[170] M. F. Nagoor Meeran, A. Goyal, S. Suchal, S. Sharma, and S. S. Ojha, "Therapeutic potential of limonene: anti-inflammatory and antioxidant properties," *Inflammopharmacology*, vol. 29, no. 3, pp. 871–894, 2021.

[171] A. Al-Shudifat et al., "Antidepressant potential of β-caryophyllene in maternal separation-induced depression-like in mice: A focus on oxidative stress and nitrite levels," *Phytomedicine Plus*, vol. 4, no. 4, p. 100624, 2024.

classification accuracy of 95%, demonstrating the robustness and reliability of the method. These results confirm that IR spectroscopy, combined with chemometrics, represents a rapid and non-destructive analytical tool for the authentication and traceability of black pepper.

The volatile profile, analysed through HS-SPME-GC-MS, complemented the spectroscopic and functional data by providing detailed information on the aromatic composition of the samples. A total of 44 volatile compounds were identified, mainly terpenoids, with monoterpenes as the predominant group. Compounds such as sabinene, 3-carene, β-caryophyllene, and limonene were found to be key discriminant markers among the varieties, contributing both to their characteristic aroma and to the potential bioactive properties of the samples. PCA revealed clear grouping tendencies, with similar profiles for Kampot and Kampot PGI, and marked differences for Madagascar and green pepper samples.

Overall, this work demonstrates the effectiveness of integrating complementary analytical techniques, antioxidant assays, volatile profiling by GC-MS, and IR spectroscopy supported by chemometrics, to achieve a comprehensive characterization of black pepper. The proposed approach allows simultaneous evaluation of functional quality, chemical composition, and origin-related attributes within a single analytical framework. This methodology not only enhances the understanding of the relationship between composition and provenance but also highlights the potential of rapid, cost-effective, and non-destructive techniques for quality control and authentication in the spice supply chain. The workflow developed in this study may contribute to the development of traceability and certification systems for high-value agricultural products such as black pepper.

CHAPTER 7

Conclusion and final remarks

The decision to focus this research project on commonly used plant matrices was driven by the wealth of bioactive compounds they contain. As already explained, Brassicaceae are a rich source of polyphenols and antioxidant compounds, which has led to increasing research interest in their study. Techniques such as IR and E-Eye allow non-destructive analysis of the leaves of varieties belonging to this family. This helps to prevent fraud in a commercial context and encourages the cosmetics and pharmaceutical industries to increasingly use natural products in the production of beauty formulations, medicines and supplements.

The bioactive substances present in plants can vary depending on growing conditions, the soil in which they are grown (pH, texture, nutrient content) and the climate. just think of the interaction between genotype and environment that influences the production of glucosylates, which is strongly regulated by the interaction between environmental signals and genetic loci[172]. Therefore, varieties grown within the same experimental field have been studied to eliminate differences due to different soil and climate conditions. The experimental field was necessary to eliminate differences in cultivation such as altitude, microclimate, exposure, and agronomic practices (irrigation, fertilisation, and treatments) that would have influenced the accumulation and presence of secondary metabolites such as glucosinolates[173], flavonoids, and phenolic acids. For this reason, when

[172] S. Kattel and G. F. Antonious, "Glucosinolates in Cruciferous Vegetables: Genetic and Environmental Regulation, Metabolic Pathways, and Cancer-Preventive Mechanisms," *Int. J. Plant Biol.*, vol. 16, no. 2, p. 58, 2025.
[173] F. Langston, A. A. Redha, G. R. Nash, J. R. Bows, L. Torquati, M. J. Gidley, and D. Cozzolino, "Qualitative analysis of broccoli (*Brassica oleracea* var. *italica*) glucosinolates: Investigating the use of mid-infrared spectroscopy combined with chemometrics," *J. Food Compos. Anal.*, vol. 123, p. 105532, 2023.

studying local varieties, it is essential to document the cultivation conditions and area of origin for the traceability of plant products and the reproducibility of analytical data. The geographical and environmental characterisation of plant matrices is not only of agronomic or taxonomic interest but is also a key element in the standardisation and enhancement of the differences between plant species and their pharmacological activity. For these reasons, the choice of the experimental field for the cultivation of the species characterised in the first part of the work on Brassicaceae was essential.

Having eliminated the differences listed above, which would have involved the analysis of samples from different geographical areas, classification using non-destructive techniques such as infrared spectroscopy (ATR-FT-IR) and E-Eye analysis was based mainly on chemical information due to interaction with infrared radiation, i.e. on the metabolic profile, and in the second case on external appearance, focusing on morphological and chromatic parameters such as colour, homogeneity and surface texture.

Both techniques were able to classify the samples of interest with high efficiency thanks to the application of chemometric approaches.

The data obtained following spectroscopic analysis were analysed by applying modelling classification approaches for classes with the SIMCA algorithm and discriminant analysis with the PLS-DA algorithm, achieving accurate classification with particular attention to the unique varieties of Abruzzo Mugnoli.

The data obtained following E-Eye analysis were processed using EA and SIMCA. The EA highlighted similarities and differences between the populations investigated, while the SIMCA approach allowed classification with a predictive sensitivity of over 70% on an external validation test for seven of the eight classes considered, despite the strong tendency of these plants to hybridise, which leads to strong similarities between classes. The combination of MIA for colorgrams analysis and a class-modelling approach such as SIMCA allowed for highly efficient classification of the varieties investigated, with the exception of the Mugnoli class accessions, which were confused with each other or with the commercial categories due to their strong hybridisation capacity.

The study using non-destructive techniques highlighted the applicability of these approaches for the assessment of genetic biodiversity and the conservation of accessions of different local Brassicaceae varieties.

From these aspects, obtained from non-destructive analyses, it can be stated that the development and use of non-destructive approaches allows the characterisation and classification of Brassicaceae varieties.

As regards the destructive analysis performed on leaf samples of the varieties of interest stored at -20°C through the design of a DoE, it was possible to identify the optimal conditions for the extraction of polyphenolic compounds contained in the species. The optimum extraction was found to coincide with sample B25, i.e. 90 mg of sieved dry ground leaf extracted in 5 mL of extracting mixture consisting of water and ethanol in a 60:40 ratio for 45 minutes at 60 °C. The aim was to correlate the antioxidant activity determined by the DPPH assay with the chromatographic profile using the entire spectrochromatogram obtained following HPLC-DAD chromatographic analysis. This was done by developing a multiway model such as N-PLS in order to construct a predictive model of antioxidant activity based on the spectrochromatogram obtained following analysis of the extract.

Despite the disadvantages of the assay used, including its low specificity due to the main mechanism of DPPH radical reduction by the antioxidant, based solely on the transfer of electrons or hydrogen atoms, and possible chromatic interference, being photosensitive, the assay needs to be performed quickly with the risk of introducing random errors, the model showed good predictive performance, higher than expected.

Also for the optimisation of cold water extraction of bioactive compounds from dried flower heads of *Matricaria chamomilla* L. using a totally green and safe approach for human use, designing experiments a priori with a full factorial design to study the simultaneous effect, across the entire experimental domain, of the variables of interest with a limited number of laboratory tests to be performed.

From the experiments conducted, using flower heads from the same batch to reduce differences due to soil and climate factors and treatment and storage processes that influence the composition of plant species of phytotherapeutic interest, the optimal extract was found to be that extracted using 2.5 grams of ground and sieved chamomile at 25°C for 32 minutes.

The extracts were characterised by HPLC-DAD analysis simultaneously with UV-Vis analysis.

The results were evaluated with the support of chemometrics, highlighting the negative influence of extraction time on the quality of the extract as well as the interaction between time and temperature factors. On the other hand, a

significant positive quadratic effect due to temperature and a positive coefficient indicating a favourable effect on extraction for the amount of chamomile used emerged. The use of ASCA was fundamental in the evaluation of the UV-Vis profile, providing an alternative untargeted method for understanding the effects of the variables. Being used as an untargeted technique, it highlighted the absorbances most influenced by the experimental variables considered without necessarily having to identify the compounds, resulting in an effective and useful method for the qualitative evaluation and interpretation of the compositional characteristics of the extracts. The extraction conducted under optimal conditions produced samples rich in hydroxybenzoic acids, hydroxycinnamic acids and flavonol derivatives. Chemometric analysis highlights how the polyphenolic profile varies significantly with changes in temperature, time and their interaction. Lower temperatures contributed to enriching the composition of the extract with flavones and glycosylated flavonols, while higher temperatures favoured the extraction of hydroxycinnamic and hydroxybenzoic acids. Consequently, conducting an extraction combining a temperature of 25°C with a short extraction time favours the preparation of extracts with greater antioxidant activity. The antioxidant activity of the optimal extract was evaluated using the DPPH assay, with subsequent calculation of the IC_{50} value. The application of chemometric approaches has shown how they can be useful in improving the quality of extracts and evaluating how variables influence the chamomile extraction process, with the only disadvantage being that, as it is not a hot infusion, it does not reduce the microbial risk that could cause gastrointestinal infections.

The combination of chemical characterization and advanced data analysis proved to be highly effective not only for the study of complex matrices in general but also for widely consumed plant species such as *Piper nigrum* L. This species is well known for its diverse phytochemical composition and associated functional properties. In the present study, the antioxidant activity of black pepper samples from different geographical origins was evaluated using the DPPH radical scavenging assay, and the results were expressed as IC_{50} values to enable quantitative comparison among varieties.

The comparison of antioxidant capacities revealed clear differences among samples, reflecting the influence of geographical origin on the chemical composition of the spice. In particular, Madagascar black pepper exhibited the highest radical scavenging capacity, suggesting a greater abundance of

bioactive constituents, especially phenolic compounds and terpenoids. These findings underline how environmental and agronomic factors, such as climate, soil composition, and post-harvest processing, can significantly affect the biosynthesis and accumulation of antioxidant molecules.

When considered alongside the ATR-FT-IR and GC-MS analyses, these results further support the link between the chemical and functional attributes of peppercorns and their geographical provenance. While FT-IR spectroscopy combined with PLS-DA successfully distinguished the samples according to their origin with a classification accuracy of 95%, the volatile compound profiling performed by HS-SPME-GC-MS provided a detailed overview of the aromatic fingerprint. Although the limited size of the GC-MS dataset did not allow for supervised classification, exploratory PCA analysis highlighted distinct clustering patterns and identified key volatile markers, such as sabinene, 3-carene, β-caryophyllene, and limonene, that contribute both to the characteristic aroma and to the bioactive potential of the samples.

Taken together, these results confirm the value of integrating functional, compositional, and spectroscopic analyses with chemometric tools to achieve a comprehensive understanding of black pepper's chemical diversity. This integrated approach not only enhances product characterization but also offers a promising framework for quality control, authentication, and traceability of spices and other high-value agricultural products.

CHAPTER 8

Way forward

To resolve challenges related to genetic diversity and the hybrid nature of Brassicaceae, the combination of analytical techniques and chemometric approaches has proven to be of fundamental importance, providing a basis for future studies on the conservation of agricultural biodiversity and functional foods. The confusion generated by the classification of the three Mugnoli classes suggests that future research should focus on DNA analysis of these plants to define the genetic relationships between the different accessions, confirm the results obtained from phenotypic analyses in combination with chemometrics, and develop optimal strategies for the conservation and enhancement of these particular local varieties.

The use of more advanced analytical techniques such as cyclic voltammetry could be effective in constructing a multiway predictive model in which the spectrochromatographic profile is correlated with the antioxidant power of the varieties considered, offering a more complete and robust assessment of their bioactive potential; at the same time, the use of standard mixtures would allow the identification of the compounds that most influence antioxidant activity. Furthermore, the leaves could be analysed using hyperspectral imaging as an additional non-destructive approach, capable of providing complementary information on compositional variability and functional properties.

Furthermore, the presence of anti-cancer substances in Brassicaceae extracts such as glucobrassicin[174], identified among the metabolites present in the extracts, sinalbine, with potential anti-proliferative activity[175], glucoraphanin and other glucosinolates known for their anti-tumour

[174] L. Mandrich and E. Caputo, "Brassicaceae-Derived Anticancer Agents: Towards a Green Approach to Beat Cancer," *Nutrients*, vol. 12, no. 3, p. 868, Mar. 2020.

[175] Boscaro, L. Boffa, A. Binello, G. Amisano, S. Fornasero, G. Cravotto, and M. Gallicchio, "Antiproliferative, Proapoptotic, Antioxidant and Antimicrobial Effects of *Sinapis nigra* L. and *Sinapis alba* L. Extracts," *Molecules*, vol. 23, no. 11, p. 3004, Nov. 2018.

properties[176] particularly in the gastrointestinal tract, colon-rectum[177], prostate and breast[178] sees numerous interesting future prospects, including the targeted isolation of these molecules and their encapsulation in innovative delivery systems, such as Metal-Organic Frameworks (MOFs), in order to deliver them selectively and enhance their chemopreventive and anti-cancer efficacy.

As regards the chamomile matrix, using a targeted assay to evaluate anti-inflammatory activity, such as the inhibition of cyclooxygenase (COX), could enrich the considerations made so far with specific complementary information on the anti-inflammatory activity of chamomile. Similarly, mass spectrometry analysis could enrich the work with the accurate identification of the polyphenolic compounds present in the extracts.

The use of high-resolution analytical techniques, such as liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) and HPLC-DAD, would allow for more accurate identification of the molecules responsible for antioxidant activity in black pepper extracts. Integrating the data obtained in this way with the results of antioxidant assays would allow, through the application of multiway algorithms and advanced chemometric models, the chemical composition to be correlated with the biological activity value, providing a more thorough understanding of the mechanisms that determine the bioactive potential of the matrix studied.

In the case of black pepper (*Piper nigrum* L.), future studies could also explore the influence of climatic and geographical factors on its chemical composition and bioactive profile. By correlating environmental parameters such as temperature, humidity, rainfall, altitude, and soil composition with the variation in volatile, phenolic, and alkaloid contents, it would be possible to better understand how the growing conditions affect the quality and functional characteristics of the spice. Combining such environmental data with chemometric analysis could lead to predictive models capable of describing the relationship between geographical origin and the accumulation

---

[176] K. Grosser and N. M. van Dam, "A straightforward method for glucosinolate extraction and analysis with high-pressure liquid chromatography (HPLC)," *J. Vis. Exp.*, no. 121, p. e55425, Mar. 2017.

[177] M. Peña, A. Guzmán, R. Martínez, C. Mesas, J. Prados, J. M. Porres, and C. Melguizo, "Preventive effects of Brassicaceae family for colon cancer prevention: A focus on in vitro studies," *Biomed. Pharmacother.*, vol. 151, p. 113145, 2022.

[178] R. Ilahy, I. Tlili, Z. Pék, A. Montefusco, M. W. Siddiqui, F. Homa, C. Hdider, T. R'Him, H. Lajos, and M. S. Lenucci, "Pre- and Post-harvest Factors Affecting Glucosinolate Content in Broccoli," *Front. Nutr.*, vol. 7, p. 147, 2020.

of bioactive compounds, thereby supporting sustainable cultivation practices and quality certification processes. This line of investigation would also provide valuable information on how climate change might influence the aromatic and nutritional attributes of black pepper in the future, contributing to both the protection of local cultivars and the preservation of agricultural biodiversity.

Scientific publications

C. Scappaticci, S. Spera, A. Biancolillo, and F. Marini, “Detection and quantification of alprazolam added to long drinks by near infrared spectroscopy and chemometrics,” Molecules, vol. 27, no. 19, p. 6420, 2022. doi: 10.3390/molecules27196420.

A. Biancolillo, C. Scappaticci, M. Foschi, C. Rossini, and F. Marini, “Coupling of NIR spectroscopy and chemometrics for the quantification of dexamethasone in pharmaceutical formulations,” Pharmaceuticals, vol. 16, no. 2, p. 309, 2023. doi: 10.3390/ph16020309.

A. Biancolillo, R. Ferretti, C. Scappaticci, M. Foschi, A. A. D’Archivio, M. Di Santo, and L. Di Martino, “Development of a non-destructive tool based on E-Eye and agro-morphological descriptors for the characterization and classification of different Brassicaceae landraces,” Applied Sciences, vol. 13, no. 11, p. 6591, 2023. doi: 10.3390/app13116591.

C. Scappaticci, M. Foschi, A. Plaku, A. Biancolillo, and A. A. D’Archivio, “Enhancing traceability of Italian almonds through IR spectroscopy and chemometric classifiers,” Applied Sciences, vol. 13, no. 23, p. 12765, 2023. doi: 10.3390/app132312765.

L. Di Martino, A. Biancolillo, C. Scappaticci, M. Foschi, and A. A. D’Archivio, “Non-destructive characterization of Italian local Brassicaceae cultivars using ATR-FT-IR and chemometrics,” Applied Sciences, vol. 14, no. 3, p. 1277, 2024. doi: 10.3390/app14031277.

M. Foschi, L. Marsili, I. Luciani, G. Gornati, C. Scappaticci, F. Ruggieri, A. A. D’Archivio, and A. Biancolillo, “Optimization of the cold water extraction method for high-value bioactive compounds from Chamomile (Matricaria chamomilla L.) flower heads through chemometrics,” Molecules, vol. 29, no. 20, p. 4925, 2024. doi: 10.3390/molecules29204925.

M. Foschi, V. Colantoni, S. Reale, C. Scappaticci, A. A. D’Archivio, and A. Biancolillo, “Analysis of volatile organic compounds in textiles: Insights from GC–MS with metal content assessment using ICP–MS,” Applied Sciences, vol. 15, no. 3, p. 1572, 2025. doi: 10.3390/app15031572.

A. Biancolillo, M. Rossi, S. Reale, C. Scappaticci, M. Foschi, and A. A. D'Archivio, "Multi-analytical approach for the characterization and geographical authentication of black pepper: volatile analysis, antioxidant activity, and spectroscopic profiling," Journal of Food Composition and Analysis, submitted for publication, under review, 2025.

Conference Participation

Oral Contributions

A. Biancolillo, D. Tanzilli, C. Scappaticci, F. Marini, Class-modeling revisited: the algorithms' new clothes, The 20th International conference on NIR – NIR2021, Beijing (Cina), October 18-21, 2021 (Co-author);

A. Biancolillo, D. Tanzilli, C. Scappaticci, F. Marini, Class-modeling for multi-block data: something borrowed, something new, SciX2021, Providence (USA), September 26 – October 1, 2021 (Co-author);

C. Scappaticci, A. Biancolillo, F. Marini, SIMCA: unified framework for single- and multi-block data. SYNC - First Symposium for Young Chemists: Innovation and Sustainability, Rome, June 20-23, 2022;

C. Scappaticci, A. Biancolillo, F. Marini, SIMCA: A Unified Framework for Single-Block and Multi-Block Data Analysis, Chemometrics Workshop, L'Aquila (Italy), May 30 – June 1, 2022;

M. Foschi, G. Gornati, C. Scappaticci, L.Marsili, A..A. D'Archivio, F. Ruggieri, A.Biancolillo, optimization of water-based extraction of chemomile (Matricaria Chamomilla L.) by analytical and chemometric approach. Interregional Congress of the Tuscany, Umbria, Marche and Abruzzo Sections of the Italian Chemical Society (TUMA 2023), Francavilla al Mare (CH, Italy), June 22-23, 2023 (Co-author);

C. Scappaticci, L. Di Martino, A. Biancolillo, M.Foschi, A. A. D'Archivio, Use of non-destructive analytical techniques and chemometric tools for the authentication of typical Italian vegetables species "Mugnoli di Pettorano sul Gizio". Congresso interregionale delle sezioni Toscana, Umbria, Marche e Abruzzo della società chimica italiana (TUMA 2023), Francavilla al Mare (CH, Italy), June 22-23, 2023;

C. Scappaticci, A. Biancolillo, M. Foschi, A. A. D'Archivio, Determination of geographical origin of Italian almond varieties using ATR-FT-IR spectroscopy and chemometric tools. TMS Day of Sciences, virtual attendance, September 14, 2023;

M. Foschi, A. Biancolillo, C. Scappaticci, A. A. D'Archivio, Chemometric analysis for traceability of high-value Italian rice cultivars: a multi-platform approach. XXX Congresso della Divisione di Chimica Analitica della Società Chimica Italiana, Città del Vasto (CH, Italy), September 17-21, 2023 (Co-author);

C. Scappaticci, A. Biancolillo, A. A. D'Archivio, S. Fiore, C. Filauro, M.Foschi, Chemometric analysis for the traceability of high-quality Italian rice cultivars. Merck Young Chemists' Symposium XXII edition (MYCS 2023), Rimini (Italy), November 13-15, 2023;

M. Foschi, C. Scappaticci, L. Marsili, A. A. D'Archivio, F. Ruggieri, A. Biancolillo, Optimization of water-based extraction of Chamomile (Matricaria Chamomilla L.) by analytical and Chemometric approach. Second Symposium for Young Chemists (SYNC 2024), Rome (Italy), June 24-28, 2024 (Co-author);

C. Scappaticci, A. Biancolillo, A. A. D'Archivio, S. Fiore, C. Filauro, M. Foschi, Traceability of high quality Italian rice cultivars through chemometric approaches. Second Symposium for Young Chemists (SYNC 2024), Rome (Italy), June 24-28, 2024;

C. Scappaticci, A. Biancolillo, A. A. D'Archivio, S. Fiore, C. Filauro, M. Foschi, Use of chemometric approaches for the traceability of high-quality Italian rice cultivars. TMS Day of Sciences, September 9, 2024;

C. Scappaticci, A. Biancolillo, M. Foschi, A. A. D'Archivio, Exctraction of pharmacologically active substances from typical Italian plant species such as "Mugnoli di Pettorano sul Gizio using analytical technique and chemometric approaches. Merck Young Chemists' Symposium XXIII edition (MYCS 2024), Rimini (Italy), November 13-15, 2024.

Poster Contributions

C. Scappaticci, A. Biancolillo, F. Marini ,SIMCA framework for multi-block class modeling, XVIII Chemometrics in Analytical Chemistry (CAC2022), Rome (Italy), August 29 – September 2, 2022;

M. Foschi, A. Biancolillo, C. Scappaticci, A.A. D'Archivio, ICP-MS and chemometrics for the characterization and discrimination of three different species of edible insects, XI Colloquium Chemometricum Mediterraneum, Padua (Italy), June 27-30, 2023 (Co-author);

C. Scappaticci, A. Biancolillo, M. Foschi, A. A. D'Archivio, Development of non-destructive tools for the authentication of the typical Italian vegetable species "Mugnoli di Pettorano sul Gizio", XI Colloquium Chemometricum Mediterraneum, Padua (Italy), June 27-30, 2023;

C. Scappaticci, A. Biancolillo, M. Foschi, A. A. D'Archivio, Application of IR spectroscopy and class modeling technique for almond traceability. XXX Congresso della Divisione di Chimica Analitica della Società Chimica Italiana, Città del Vasto (CH, Italy), September 17-21, 2023;

C. Scappaticci, A. Biancolillo, F. Marini, Application of NIR spectroscopy for detecting adulteration in insect flours. 22nd International Conference on Near Infrared Spectroscopy (NIR 2025- Light through centuries), Rome (Italy), June 8-12, 2025.

Awards

Best Oral Presentation Award at the Second Symposium for Young Chemists (SYNC 2024), Rome (Italy), June 24-28, 2024, C. Scappaticci, A. Biancolillo, A. A. D'Archivio, S. Fiore, C. Filauro, M. Foschi, Traceability of high quality Italian rice cultivars through chemometric approaches;

Third Prize for Best Presentation at the TMS Day of Sciences, September 19, 2024, C. Scappaticci, A. Biancolillo, A. A. D'Archivio, S. Fiore, C. Filauro, M. Foschi, Use of chemometric approaches for the traceability of high-quality Italian rice cultivars;

Winner of the Chemometrics Group Grant for participation at the 22nd International Conference on Near Infrared Spectroscopy (NIR 2025 – Light Through Centuries), Rome (Italy), June 8–12, 2025, C. Scappaticci, A. Biancolillo, F. Marini, Application of NIR spectroscopy for detecting adulteration in insect flours.

Trainings

Training Internship at Eurolab Srl, Laboratory of Chemical Analyses and Quality Control: Performed analytical tests on water samples (drinking, wastewater, etc.), agricultural products, and soil.

Information and Training Course on Chemical Risk at Department of Chemistry, Sapienza University of Rome:

Health and safety risks; Prevention and protection measures.

Doctoral School: School of Chemometrics, University of Genoa
Module: Experimental Design.

JMP Workshop: Getting Started with Design of Experiments (DoE) and Use of JMP.

XIX Chemometrics in Analytical Chemistry (CAC 2024), Virtual Pre-Conference Course:
Module 1: Multivariate Curve Resolution (MCR);
Module 2: Targeted and Untargeted Anomaly Detection Approaches in Hyperspectral Imaging;
Attended virtually the XIX Chemometrics in Analytical Chemistry Conference.

Symposium of Young Chemometricians, University Centre of Bertinoro (FC, Italy):
PCA and PLS: Introduction, Examples, Interpretation and Issues;
Model Validation: General Concepts and PCA Cross-Validation;
PLS Validation: CV, dCV in PLS and PLS-DA; Discussion on Test-Set Validation;
Variable Selection: General Concepts, Sparse PCA, Group-Wise PCA and Interpretation;
PLS Variable Selection: sPLS and dCV;
Data Fusion: General Concepts and Low-Level PCA Fusion;
PLS Data Fusion: Low-Level and Mid-Level Algorithms.

FT-NIR Analyses for Process Control in the Food Industry.

Rethinking Mid-IR Gas Sensors Towards More Sustainable Approaches.

Membrane Technology.

Chemometrics Open Day, Italian Chemical Society, Division of Analytical Chemistry, Chemometrics Group: Teaching and Learning Chemometrics.

Workshop: Problem Solving with Design of Experiments, Navigating Complex Challenges in R&D and Manufacturing.

Claudia Scappaticci was born in Sora (FR, Italy) on November 24, 1997. She obtained a diploma as a Chemical Expert from the Tulliano Institute in Arpino (FR, Italy), I.T.I.S. section in Chemistry and Materials. Subsequently, she pursued her university studies at Sapienza University of Rome, where she obtained her Bachelor's degree in Chemistry in 2019 and her Master's degree in Analytical Chemistry in 2021. In 2021, she obtained the qualification to practice as a professional chemist. From 2021 to 2022, she carried out a research fellowship at Sapienza University of Rome in the field of Chemometrics. She is currently a Ph.D. student in Chemistry at the University of L'Aquila, where she conducts scientific research focused on the application of chemometric methods for the characterization and valorization of food matrices. She will soon begin a research fellowship at the National Research Council (CNR) in Rome, continuing her training and specialization in the field of chemical research.

Finanziato dall'Unione europea-*NextGeneration EU*
Missione 4, Componente 1 – CUP E13C22001060006


UNIVERSITÀ
DEGLI STUDI
DELL'AQUILA

DSFC
Dipartimento
di Scienze Fisiche
e Chimiche

VITALITY