CLASSIFICATION AND QUANTIFICATION OF BEET & CANE SUGAR BY USING OPTICAL SPECTROSCOPY AND CHEMOMETRICS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

HİLMİ ERİKLİOĞLU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
FOOD ENGINEERING

SEPTEMBER 2022

Approval of the thesis:

CLASSIFICATION AND QUANTIFICATION OF BEET & CANE SUGAR BY USING OPTICAL SPECTROSCOPY AND CHEMOMETRICS

submitted by **HİLMİ ERİKLİOĞLU** in partial fulfillment of the requirements for the degree of **Master of Science in Food Engineering, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Serpil Şahin
Head of the Department, **Food Engineering**

Assoc. Prof. Dr. Halil Mecit Öztop
Supervisor, **Food Engineering, METU**

Assist. Prof. Dr. Ali Can Karaca
Co-Supervisor, **Computer Engineering, Yıldız Technical University**

Examining Committee Members:

Prof. Dr. Behiç Mert
Food Engineering, METU

Assoc. Prof. Dr. Halil Mecit Öztop
Food Engineering, METU

Assist. Prof. Dr. Ali Can Karaca
Computer Engineering, Yıldız Technical University

Prof. Dr. Deniz Çekmecelioğlu
Food Engineering, METU

Prof. Dr. Marena Manley
Food Science and Technology, Stellenbosch University

Date: 02.09.2022

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Last name : Hilmi Eriklioğlu

Signature :

ABSTRACT

CLASSIFICATION AND QUANTIFICATION OF BEET & CANE SUGAR BY USING OPTICAL SPECTROSCOPY AND CHEMOMETRICS

Eriklioğlu, Hilmi
Master of Science, Food Engineering
Supervisor: Assoc. Dr. Halil Mecit Öztop
Co-Supervisor: Assist. Prof. Dr. Ali Can Karaca

September 2022, 113 pages

Sucrose is one of the main ingredients used in food industry. It is obtained mainly from two different sources; *sugar beet and sugar cane*. Due to govermental regulations, cane sugar is not allowed to be produced in Turkey. On the other hand, cane sugar can be illegally sold as beet sugar. Since molecular structure of sucrose is same, it is difficult to differentiate sources by using chemical methods. Therefore, developing more practical and affordable methods would be valuable for the food industry. Optical spectroscopy (UV-VIS-NIR) can be a promising technique for detection of differences. In this thesis, sucrose samples (cane, beet) were collected from nine countries to prepare 25% (w/w) sucrose water solutions and their absorbances were recorded (200-1380nm). Results showed that, spectral signature differences were observable between 200-600 nm. It is known that improving the prediction accuracy requires chemometrics, such as linear discriminant analysis (LDA), soft independent modelling of class analogy (SIMCA), k-nearest neighbors (KNN) and classification and regression trees (CART). All methods showed high performance, but LDA gave 100% correct classification with a simple interpretation. In addition, binary mixtures of these sugar were also prepared for quantification analysis. Multiple linear regression (MLR) with *Savitsky Golay* (SG) and the first derivative, gave the most acceptable results of root mean square error of calibration (RMSEC), prediction (RMSEP) and residual predictive deviation (RPD) values of 2.956, 3.026 and 10.251 respectively. The obtained results seemed promising for the plant source of sucrose to be detected by using UV region and chemometrics.

Keywords: Sucrose, sugar cane, sugar beet, optical spectroscopy, chemometrics

ÖZ

OPTİK SPEKTROSKOPİ VE KEMOMETRİ KULLANARAK PANCAR VE KAMIŞ ŞEKERİNİN SINIFLANDIRILMASI VE KARIŞIMLARDAKİ MİKTARININ TESPİTİ

Eriklioğlu, Hilmi
Yüksek Lisans, Gıda Mühendisliği
Tez Yöneticisi: Doç. Dr. Halil Mecit Öztop
Ortak Tez Yöneticisi: Dr. Öğretim Üyesi Ali Can Karaca

Eylül 2022, 113 sayfa

Sükroz, şeker endüstrisinde kullanılan ana maddelerden biridir. Dünya çapında sükroz, şeker pancarı ve şeker kamışı olmak üzere iki farklı kaynaktan üretilmektedir. Ancak, hükümet politikaları nedeniyle şeker kamışının Türkiyede üretilmesine izin verilmemiştir. Öte yandan, şeker kamışı yasadışı olarak pancar şekeri olarak satılabilir.. Sükrozun moleküler yapısı aynı olduğu için kimyasal yöntemlerle kaynağını ayırt etmek zordur. Bu nedenle daha pratik ve ekonomik yöntemlerin geliştirilmesi gıda endüstrisi için değerli olacaktır. Optik spektroskopi (UV-VIS-NIR), farklılıkların tespiti için umut verici bir teknik olabilir. Bu araştırmada, farklı bitki kaynaklarından (kamış, pancar) %25 (w/w) sükroz su solüsyonları hazırlamak için dokuz ülkeden sükroz örnekleri toplanmış ve absorbansları 200-1380nm arasında kaydedilmiştir. Sonuçlar 200-600 nm arasında spektral imza farklılıkları olduğunu ortaya koymuştur. Tahmin doğruluğunun iyileştirilmesinin LDA, SIMCA, KNN ve CART gibi kemometrik yaklaşımları gerektirdiği bilinmektedir. Sonuçlar, birkaç yöntemin yüksek performans gösterdiğini, ancak LDA'nın en basit yorumla %100 doğru sınıflandırma verdiğini göstermektedir. Ayrıca, kantifikasyon analizi için değişen konsantrasyonlar yoluyla bu şekerlerin ikili karışımları hazırlanmıştır. En kabul edilebilir sonuçları veren çoklu doğrusal regresyon (MLR), Savitsky Golay filtresi ve birinci türev ile birlikte, sırasıyla 2.956, 3.026 and 10.251 olan, RMSEC, RMSEP ve RPD değerlerini verdi. Elde edilen sonuçlar, sakarozun bitki kaynağının UV bölgesi ve kemometrik yöntemler kullanılarak tespit edilebileceği konusunda umut verici görünmektedir.

Anahtar Kelimeler: Sükroz, şeker pancarı, şeker kamışı, optik spektroskopi, kemometri

Dedicated to my family and friends.

ACKNOWLEDGMENTS

The author wishes to express his deepest gratitude to his supervisor Assoc. Prof. Dr. Halil Mecit Öztop and co-supervisor Assist. Dr. Ali Can Karaca for their guidance, advice, criticism, encouragements and insight throughout the research.

I would also like to thank Prof. Dr. Rosario Castillo, Prof. Dr. Marena Manley, for their valuable suggestions and comments, Dr. Mikhail Khodasevich for his knowledge transfer and support about instruments, PhD. St. Esmanur İlhan for her technical support and academic guide.

This work has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101008228.

TABLE OF CONTENTS

CHAPTERS

APPENDICES

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

1.1 Sugar

Sugar, being one of the main energy sources of human metabolism, has been obtained from various sources throughout the history. It is a product that carries a strategic role all over the world. Other than being a fundamental food product, sugar is a protected commodity for all countries because of its contributions to the agricultural production, by-product evaluation and employment opportunities (Eştürk, 2018). In confectionary industry, invert sugar, corn syrup and sucrose are used as sugar sources.

Corn syrup is a starch-based sugar and one of the most preferred sugar sources in the industry since its production cost is low when compared to the other alternatives. High fructose corn syrup is a sweeter version of corn syrup, and it is obtained by converting glucose contained by corn syrup to the fructose.

Sucrose also known as table sugar is mostly produced from sugar beet and sugar cane. It is a disaccharide that contains glucose and fructose subunits (Figure 1.1). It has a relatively low glycemic index of 65 because of the fructose subunit (Wolever, 2006). Fructose has a very small effect on increasing blood glucose levels (Wolever, 2006). The molecular formula of sucrose is $C_{12}H_{22}O_{11}$.

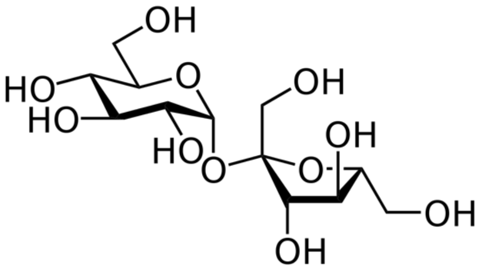


Figure 1.1 Molecular structure of sucrose.

1.2 Sources of Sucrose

Worldwide, sucrose is mainly produced by using either sugar cane or sugar beet. According to *'World, the EU and Turkey Sugar Stats'* report published by PANKOBİRLİK, which is the cooperation of Turkish beet producers, between 2016 and 2017 cane sourced sugar covered 77.60% of world's sucrose production meanwhile beet sugar was only 22.40% (PANKOBİRLİK, 2017). Another report prepared by Turkish Sugar Factories in 2017-2018 stated that 78% of annual production was obtained from sugar cane and 22% from sugar beet (Türkiye Şeker Fabrikaları, 2018). Due its high production volume, prices of the world sugar market are determined based on sugar cane, (Kaya, 2015).

Geographical locations of the countries play the most important role on source of sugar (Sefaoğlu et al., 2016). While sugar cane (*Saccharum officinarum*) grows in the tropical climate zone, sugar beet (*Beta vulgaris saccharifera L.*) grows in different climatic zones and regions located between 30° south-60° north latitudes. From one decare of sugar cane, 2-4 times higher sucrose yields can be obtained when compared to sugar beet. Due to this efficiency, the production and raw material costs of cane sugar are much lower than beet sugar. However, because of the climatic conditions, for some countries, it is not possible to grow sugar cane.

Due to the geographical conditions, some countries such as, Turkey, Russia, Ukraine, Belarus and EU are producing their sugar from sugar beet and countries like USA, Japan, China are producing from both sources. Brazil, India, Thailand, Mexico, Pakistan, Australia and a high number of other countries are producing from cane (Thow et al., 2021). In (Table 1) and (Table 2), sugar beet and sugar cane production amount of different countries are presented.

Table 1. Top 15 sugar cane producing countries (Thow et al., 2021)

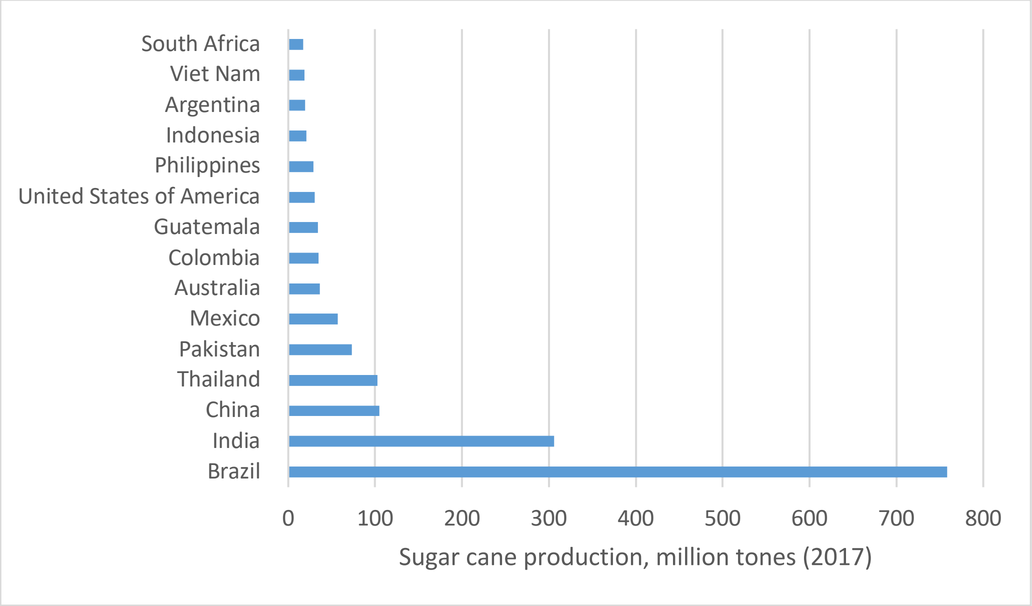
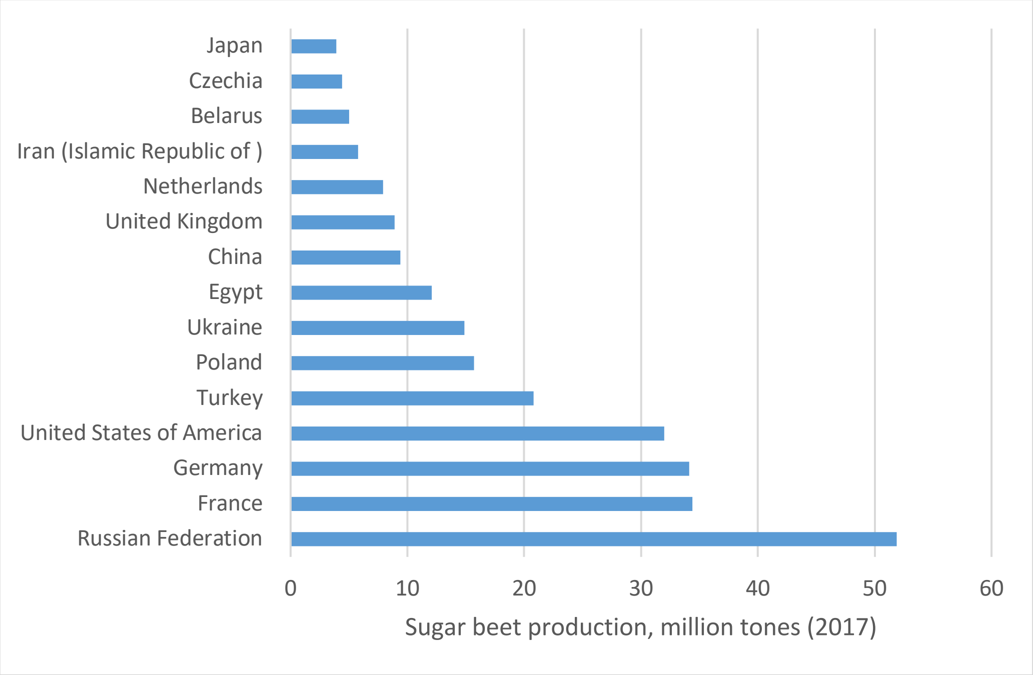


Table 2. Top 15 sugar beet producing countries (Thow et al., 2021)



When production efficiency is considered, producing sugar from sugar cane is by far more advantageous than using sugar beet as source. However, in some countries such as Turkey it is forbidden to produce sugar from sugar cane due to some governmental policies. Maintaining sugar production by using sugar beet might not be economical in one perspective, but there are other aspects such as foreign dependency, employment of farmers and regional factory workers, protecting natural balance and ongoing social and economic life.

Sugar beets are not the source of sucrose only, but also with its by-products such as beet pulp, pectin and cellulose, which are natural additives, contribute to the food production ecosystem. Molasses, which is obtained after processing beet is used as an animal feed and one of the main raw materials of yeast industry. With above-mentioned benefits, use of sugar beet to produce sugar is helpful for sustainable life by contributing to meat and milk production (Türkiye Şeker Fabrikaları, 2018). Also, in Germany and France, using sugar beet as a source of sucrose is being supported by government.

Because of the differences in production efficiency, hence high gap between the cost of white sugar obtained from sugar beet and cane, cane sugar could sometimes enter to the country with illegal ways. When the news in the media is examined, no significant case has been found regarding cane sugar smuggling into the country for the last five years, nevertheless possibility of such incident is always there, because of the reasons mentioned above.

In this regard, it is expected from the scientists to find easily applicable and quick analysis methods, which require less expertise and low-cost materials to differentiate source of the sucrose.

1.3 Differences Between Cane and Beet Sugar

There is no significant difference between sugars obtained from beet and cane sources as both of them have greater than 99.8% sucrose as the final product (Asadi, 2007). However, there are some differences in the production step. Sugar beets are refined in same factory with the addition of SO_2, whereas sugar canes go under different refining process. First, raw cane sugar is produced in factory and then transported to a separate refinery (Asadi, 2007).

There are studies that investigated the differences between sugars obtained from beet and cane. Chemical structures of sucrose for both beet and cane are same however, there are differences in the sensory properties, aroma profiles, thermal behaviors and small levels of chemical compound disparities (Asadi, 2007; Godshall et al., 1995; Lu et al., 2017; Urbanus et al., 2014).

Urbanus et al. (2014) suggested that the differences between plant sources can be detected by sensory analysis panels by making an experienced panelist to try cane and beet sugars, which were produced by different companies located in USA (Urbanus et al., 2014). In another study conducted by the same team, the sugar used in the first study were tested in different food formulations (*Messupé cake (pavlova); sugar syrup; cookies; pudding; whipped cream; iced tea*). While there were differences observed for meringue cake and sugar syrup, no statistical difference was detected for other formulations ($p>0.05$) (Urbanus et al., 2014).

One of the most important differences between cane and beet plants are their photosynthesis mechanisms. CO_2 fixation mechanisms during dark phase of photosynthesis are different for the two plants (Lu et al., 2017). Sugar beet is a C3 plant (Calvin-Benson mechanism) whereas sugar cane is a C4 plant (Hatch-Slack mechanism). Adaptation to drier and warmer environments of C4 plants are high whereas, C3 plants are more adaptive to cooler and high moisture environments. C3 plants produce three-carbon compounds as their first stable product. In C4 plants photosynthesis occur 2 times more than C3 plants.

The carbon isotopes ratio ($^{13}C/^{12}C$) of both plant is different, it is 25% for beet sugar and 11% for cane sugar (Bubnik et.al., 1995). The reason of the difference is that C4 plants metabolize almost all ^{13}C which are coming from CO_2. However, C3 plants are not as efficient as C4, they lose more ^{13}C from their leaves during photosynthesis (O'leary, 1988), High resolution NMR Spectroscopy and Isotope Ratio (IR)-Mass Spectrometry use this approach for differentiating cane and beet sugar as will be described in the latter section.

Another marker that can be used for beet and cane sugar differentiation is the presence of raffinose and theanderose. Theanderose is only present in cane sugar, and it is considered as a natural constituent (Moreldu Boil, 1996). Raffinose is present in both cane and beet sugar however in beet sugar, raffinose levels are higher compared to cane (Vaccari & Mantovani, 1995; Morel du Boil, 1997). It is also known that both theanderose and raffinose influence morphology and crystal growth (Liang et al., 1989; Morel du Boil, 1992).

Melting behavior and crystallinity of sucrose have also been studied for many years. Sucrose, a crystalline solid, was expected to have stable and non-changing melting point, nevertheless, melting behavior of sucrose was reported in wide variety of results. In a study conducted by Shah and Chakradeo (1936), impurity of the samples was stated as the only reason of those variances. Powers (1956, 1958) suggested that the reason is the water inclusions inside mother syrup. On the other hand, beet and cane sugar were shown to have different thermal behaviors when examined by differential scanning calorimeter (DSC). In the DSC thermograms, number of peaks was reported as two for cane samples and one for beet and also thermal stability degree of beet was found to be greater than cane. Moreover, heating rate dependency was found higher for cane compared to beet (Lu et al., 2017). However, it was stated in the studies that further research was necessary to understand the reason of differences, and to assess how impurities play a significant role in those variations.

In another study, differences in terms of ash conductivity, pH range and moisture content were examined. It was reported that the pH range of beet sugar was wider than cane sugar, being 6.5 to 8.0 and 6.2 to 6.7 for beet sugar and cane sugar respectively, while there was no significant difference detected for ash conductivity and moisture content (Godshell, 2013). A study was published about detecting the polysaccharides amount, that could not be removed after all clarifications, in the final sucrose product. Total polysaccharides amount residing in sugar cane was reported as 169 ppm (solids) and 77 ppm for beet samples. Amounts were initially 8238 ppm and 4067 ppm for cane and beet raw juices respectively (Godshell, 2002). In terms of the whole impurities, beet juice contained 2.5% (w/w) whereas cane juice contained 5% (w/w) of non- sugar compounds. Finally, fibers in whole fruit were studied and the ratio for beet was found as 5% (w/w) and 10% for cane (Asadi, 2007).

Such diversities are also the reason why different production approaches are necessary to produce sucrose from two sources (Asadi, 2007). Overall, studies showed that even after all clarification processes, some non-sucrose compounds can remain inside the end-product and the amount of impurities can be different for sugar cane and sugar beet.

For many years, considering the differences that were mentioned above, high number of studies were conducted to detect beet and cane sources of sucrose. In the following section the different approaches used for differentiation purposes will be explained.

1.4 Methods Used for Differentiation of the Source of Sugar (Sucrose)

Based on the differences mentioned above, several methods have been studied in the literature including sensory analysis, ion chromatography profiles, nuclear magnetic resonance spectroscopy (NMR), differential scanning calorimetry (DSC) and isotope ratio mass spectroscopy (IRMS).

1.4.1 Sensory Analysis

Based on the differences in aroma profiles and flavors, sensory analysis could be a differentiating method for the source of the sucrose. In a study, 100 panelists who are 23% male and 73% female, ages ranging from 18 to 55 years, attended a tetrad test. Samples were given in three sets, each corresponding to a different condition. Following the test, the panelists were able to distinguish between cane and beet sources of sugar. Cane sugars were characterized with sweet aftertaste and fruity aroma by-mouth, on the other hand, beet sugars were associated with earth, barnyard, oxidized and off-diary, off-flavors. Beet sugars were also characterized by burnt sugar aroma by-mouth with an aftertaste (Urbanus et al., 2014).

Some analytical techniques were also tested to understand the effects of various volatile fatty acid compounds, and it was stated that the rejected beet sugars possess high amount of volatile fatty acids which then, suggested as the reason of above-mentioned off-flavors (Godshell et.al. 1995; Moore et. al., 2004). And the presence of volatile components inside beet sugar were associated with decomposition of some parts of the plant, microbial contamination of the beet root and the soil itself (Marsili et. al., 1994; Godshell et.al., 1995; Lu et. al. 2003).

1.4.2 Ion Chromatography Profiles

Ion chromatography was proposed as a differentiation tool of plant origin of sucrose and quantification of the cane and beet mixtures. Chromatographic method focuses on the above-mentioned raffinose and theanderose contents of beet and cane originated sugars. It was suggested that, by using ion chromatography with integrated pulsed amperometric detection (IC-IPAD), also known as high performance anion-exchange chromatography (HPAEC), 20% white cane sugar adulterated beet sugar was successfully detected (Eggleston et. al., 2005).

According to Eggleston (2005), before applying more sophisticated techniques such as Nuclear Magnetic Resonance (NMR), Isotope Ratio Mass Spectroscopy (IRMS) and Differential Scanning Calorimetry, IC-IPAD can be used as a screening tool for industrial approaches. In the same study, it was also proposed that, using further chemometric methods can enhance the performance of the IC-IPAD.

1.4.3 Nuclear Magnetic Resonance Spectroscopy (NMR)

It has been known that by using high resolution (HR) NMR spectroscopy, it is possible to classify and quantify sugar beet and cane samples. This NMR method is based on using the difference of (^{13}C) carbon isotope content of C1, C4, C5, C6 positions of fructosyl and C1, C2 and C3 positions of glucosyl moieties of the sucrose molecule. A study suggested that NMR spectroscopy can be an alternative for semi-quantification and routine detection of cane sugar adulteration with IRMS methods (Monakhova & Diehl, 2016).

Another study that used HR-NMR for a discrimination tool was published in 1991. Chemometric approaches such as principal component analysis (PCA), principal discriminant analysis (PDA) and hierarchical clustering, was used to analyze isotropic variables in the form of a multidimensional space. By using site-specific natural isotope fractionation by deuterium NMR with mentioned chemometric approaches, a successful discrimination between beet and cane originated white sugars, was achieved (Martin et al., 1991).

1.4.4 Differential Scanning Calorimetry (DSC)

As mentioned above, several studies were conducted about different thermal behaviors of cane and beet sugars. A study conducted in 2017 suggested that beet and cane sugars possess substantial differences when thermal behaviors are considered (Lu et al., 2017). Beet and cane sugars differentiated from each other by parameters such as heating rate dependency, degree of thermal stability and number of DSC peaks (one peak for beet samples and two peaks for cane samples) (Lu et al., 2017).

1.4.5 Isotope Ratio Mass Spectroscopy (IRMS)

There are two types of isotopes, radioactive isotopes and stable isotopes. If an isotope of an atom is going under radioactive decay by emitting rays such as α-, β- and γ-rays, it is called as a radioactive isotope and if the isotope does not radioactively decay, then it is categorized as a stable isotope. Numerous elements such as carbon, hydrogen, oxygen, nitrogen and sulfur have two or more stable isotopes. Stable isotopes used for biological system analysis are mainly ^{1}H and ^{2}H, ^{14}N and ^{15}N, ^{12}C and ^{13}C, ^{18}O and ^{16}O. On the other hand, isotopes such as ^{3}H, ^{14}C, ^{32}P, ^{35}S, ^{125}I, ^{131}I are used for the same type of analysis as radioactive isotopes (Türkiye Atom Enerjisi Kurumu, 2015).

By measuring stable isotope ratios of carbon, nitrogen and oxygen elements, geographical origin of foodstuffs can be determined. Analyzing isotopic carbon, provides information about photosynthesis group of the plant. Since sugar beet and sugar cane follow different photosynthesis pathways, stable isotope ratios can be used as an indicator of sugar origin (Bubnik et.al., 1995).

Many studies have been carried out to determine the type of sugar by the stable isotope method. It was used to detect the adulteration of honey with beet sugar (Gonzalez Martin et al., 1998), to decide whether mulberry molasses is mixed with sugar syrup (Tosun, 2014) and in the determination of sugars for vinegar production (Perini et al., 2018). IRMS is now considered as a widely accepted method used for authentication of sugar plant origin. However, it requires an advance level of expertise with equipment, operation, and maintenance costs. In this study, referring to that problem, a new approach which includes equipment that is relatively cheap with very low operating costs, easy to use interface, will be introduced.

1.5 Optical Spectroscopy

Spectroscopic analysis techniques are based on interaction between matter and radiated energy to detect different properties of materials. Nature of this interaction can be named as absorption, reflection and transmission. Being an inexpensive and fast analysis method that requires minimal sample preparation, optical spectroscopic techniques have been widely used by researchers for many purposes including agricultural and food traceability (Fanelli et al., 2021).

Optical spectroscopy is used for both classification and quantification purposes. Main principle of differentiation between materials by optical spectroscopy is the interaction of the light with different materials, chemical bonds or physical surfaces. When a compound is excited with radiation, chemical bonds in organic materials give responses such as changing absorption or emission characteristics for different wavelengths. From those spectral fingerprints, which are determined by the type of the bond, mass of the atoms, shape of the molecule or other valuable information is extracted and used for classification or quantification (Magnus et al., 2021).

Lately UV-VIS spectra, which corresponds to (200-800nm) wavelengths of electromagnetic spectrum has gained a high number of interests among food scientists and utilized for food analysis purposes, because of its easy application, relatively low equipment costs and minimum sample preparation requirements. For significant number of studies, methods based on UV spectrum was used for authentication of food materials and to detect adulteration (Alamprese et al., 2013; Boggia et al., 2017; Dankowska & Kowalewski, 2019; Fanelli et al., 2021). Moreover, implementation of UV spectroscopy for routine analysis is also possible (Suhandy & Yulia, 2021).

An article published in 2017 suggests that, by using UV-VIS region spectrum and pattern recognition methods, such as class modelling and principal component analyses (PCA) which are mentioned later, it is possible to authenticate a new category of plant food supplements (Boggia et al., 2017). In this method, extracts of plant tissues coming from different species were studied in terms of absorbance values for different wavelengths. And those absorption profiles were related with genetically determined phytocomplexes of plant tissues (Boggia et al., 2017).

Another study was conducted in 2019, to classify tea types by using combination of UV-VIS, fluorescence and NIR spectroscopic techniques (Dankowska & Kowalewski, 2019). With the implementation of several chemometric approaches results were quite promising. All green, white, yellow, dark, oolong and black tea samples gave different spectral signatures obtained from mentioned regions. As a conclusion, different teas were classified with less than 3.3% error, it was suggested that these promising results can be utilized for routine analysis techniques with success (Dankowska & Kowalewski, 2019).

A quantitative study by using UV-VIS Spectroscopy was published in 2017. Where Concentration of the blends obtained from Coffea arabica and Coffea canephoa var. robusta, was examined by using UV-VIS region for quantification purposes (Dankowska et al., 2017). It was shown that there was a significant difference between UV-VIS spectral patterns between water extracts of mentioned coffee derivatives. Best predictive ability was obtained by applying multiple linear regression (MLR) for 60nm interval with the values of 3.6% and 7.9% for RMSEC and RMSEV respectively (Dankowska et al., 2017).

Classification of wine samples were studied with UV spectroscopy by the implementation of chemometric methods in 2013. Sauvignon Blanc Wines was collected in Argentina from 2 vintages and 3 different geographical origins. It was concluded that with the help of the chemometric methods, UV-VIS spectroscopy, showed correct classification and correlation between wine samples (Azcarate et al., 2013). Same researchers also referred to a complication that is also an important factor for this study. Some parameters such as number of samples, similarities which came from harvested soil, climate and several characteristics could limit the performance of the classification problem (Azcarate et al., 2013). Since, if the similarities or differences in the other parameters, affect the selected spectrum significantly, classification could be misled. Solution to this problem could be overcome by monitoring the results with other reference methods, finding the direct relation between target variance and spectrum, and increasing the sample set.

Up to now, numerous supportive publications and methods to detect differences were investigated in above parts. Now as the final section of the introduction, chemometric data analysis methods, which includes qualitative and quantitative analysis approaches, will be briefly introduced.

1.6 Chemometrics

Chemometrics was emerged when computers started to be used in chemical analysis methods. In 1970s, some groups already applied some of the todays known chemometric methods. However, the name chemometrics was first introduced by Swede Svante Wold and American Bruce R. Kowalski in 1972 during a conference. After several years, multiple conference series and journals mentioned the name chemometrics (Otto, 2017). The definition of chemometrics is as follows:

It is the chemical discipline which uses the statistical and chemical methods, (a) to design or select optimal measurement procedures and experiments, and (b) to provide maximum chemical information by analyzing related chemical data (Massart et al., 1988). In Figure 1.2, relation between chemometrics and different disciplines can be seen.

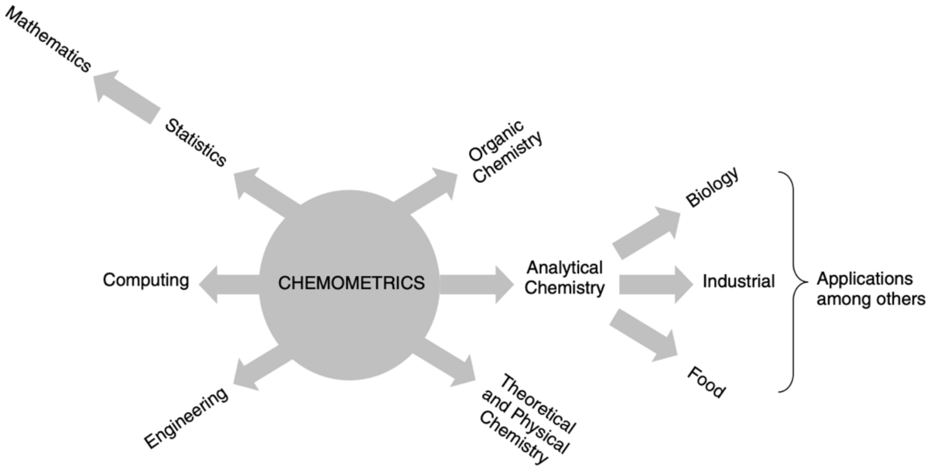


Figure 1.2. Relation of chemometrics with other disciplines (Brereton, 2003).

Application of chemometric methods for complex matrices such as food materials is a well interested topic for both food scientist and analytical chemists. For many years, numerous studies were conducted with the integration of spectroscopic methods having complex matrices and high number of variables which require multivariate data analysis tools such as chemometrics (Cámara et al., 2010; Dankowska & Kowalewski, 2019; Magnus et al., 2021; Suhandy & Yulia, 2021; Torrecilla et al., 2008). Almost all publications, which includes spectroscopy and multivariate data analysis methods, suggested that application and for some cases, combination of several chemometric approaches, improved model capability, by decreasing misclassification rates and lowering regression errors (Diniz et al., 2016;

Gad et al., 2013; Liu et al., 2010; Suhandy & Yulia, 2017; vanden Branden & Hubert, 2005).

Preprocessing methods are important tools for chemometric applications. Preprocessing is used to correct scattering effects, baseline shifts, temperature, particle size, texture deviations. Savitsky-Golay is one of the important preprocessing techniques that is used to remove unwanted spectral effects and smooth data. With first derivative it was used in food industry to increase data quality and prediction ability (Ren et al., 2021).

In following section, several chemometric methods that were used in this study are introduced.

1.6.1 Unsupervised Methods

The term unsupervised comes from '*not having a target value that will be included for the assessment of the model quality*', simply meaning the lack of Y matrix used for predictions while performing supervised methods. Unsupervised methods such as principal component analysis (PCA) and hierarchical clustering are useful for analyzing clusters that do not require a sample labeling as concentration or class. With those algorithms, one can observe groupings and patterns that are otherwise unlikely to be visualized. Human brain is more comfortable in understanding 3-dimensional space. But when the number of variables increases, it is hard to comprehend hidden properties even with high level of expertise. With concepts based on distances or dimensionality reduction, above mentioned algorithms are making multivariate data more understandable for researchers.

1.6.1.1 Principal Component Analysis (PCA)

PCA is a chemometric tool to analyze multivariate data by using dimensionality reduction approach. PCA was first introduced by Pearson (1901). It was described as *lines and planes of the closest fit to systems of points in space* (S. Wold et al., 2007). PCA is also mentioned briefly by Fisher and Mackenzie as being more convenient than variance analysis for response data modelling (Fisher & Mackenzie, 1923). They also outlined NIPALS algorithm that developed PCA to the present version, however it was later rediscovered by Herman Wold (Wold, 1973).

In short, it approximates a data matrix, X, as a product of T and P^T matrices which are smaller than original X matrix, containing information about essential patterns of X data. Moreover, dominating object patterns of original data can be seen by plotting columns of T and variable patterns are observable in the rows of P^T (Esbensen & Geladi, 2009). From Figure 1.3, structure of a PCA model can be observed, where E is a residual matrix.

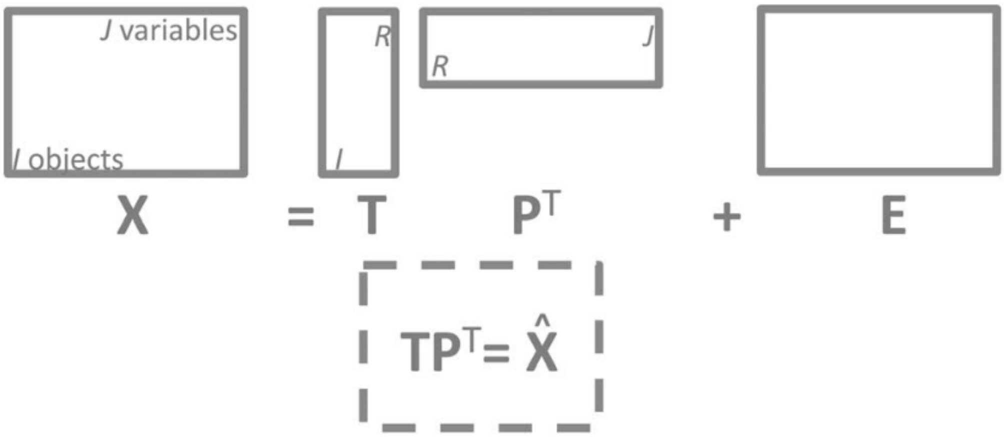


Figure 1.3. Visualization of a PCA model, where residuals (E) have the identical structure as the data (X) (Bro & Smilde, 2014).

PCA basically interprets complex datasets, which contain high number of variables with collinearities and remodels it, to be expressed as several factors called as components. Spectroscopic data require such chemometric tools to increase its descriptive performance. Without any supervised method application, PCA can be performed to see the clusters and different groupings reside in the data matrix. PCA with ultraviolet spectral measurements used for food authentication and regression models for many years and still is a topic that attracts scientist attention (Boggia et al., 2017; Dankowska et al., 2017; Magnus et al., 2021; Suhandy & Yulia, 2021).

1.6.1.2 Hierarchical Cluster Analysis (HCA)

HCA was originated from taxonomy where biological systems ordering has been made with their phenomenological similarities. In this method, hierarchical aggregation was used to differentiate objects. HCA analysis was applied according to the distances such as Euclidean, Mahalanobis and similarities such as single linkage, complete linkage between samples (Otto, 2017). Several researchers have applied HCA for unsupervised classification (Lin et al., 2015; Patras et al., 2011).

1.6.2 Supervised Methods

1.6.2.1 Linear Discriminant Analysis (LDA)

LDA is a multivariate probabilistic classification method, which works under the principle that for all considered classes there is a normal distribution with the identical covariance-variance matrix (Pouliarekou et al., 2011). Several studies in terms of food analysis were conducted including, classification of three different botanical origins of *Curcumae Rhizoma* by using UV spectroscopy (Ren et al., 2021), cocoa bean cultivars authentication with non-destructive methods (Teye et al., 2016), quality grading of in-situ cocoa beans by using optical spectroscopy (Kutsanedzie et al., 2017), quantification of coffees in blends using UV spectra (Dankowska et al., 2017).

1.6.2.2 Classification and Regression Trees (CART)

The aim of the classification analysis is to build an accurate classifier or disclosing the predictive structure of the problem. Purpose of algorithm can be expressed as finding the interactions between variables to obtain the simplest characterization conditions (Breiman Leo et al., 1984). In a study, which investigated the spectroscopic data of several food materials such as beer, oil and olive; it was concluded that using CART for classification and regression analysis was efficient, simple and could cope with data collinearity problems that are observed in spectral datasets (Kucheryavskiy, 2018). Here, in (Figure 1.4) classification nodes for an example study is shown.

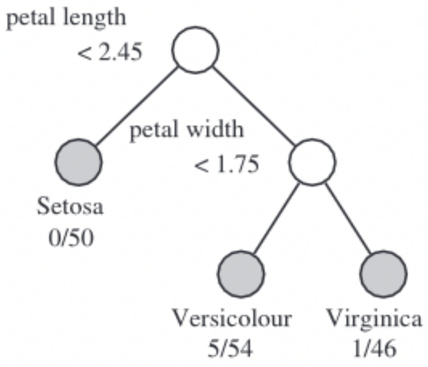


Figure 1.4. Classification tree model for iris data (Loh, 2014).

1.6.2.3 Soft independent modelling of class analogy (SIMCA)

SIMCA was first introduced by Svante Wold, as a method, which uses similarity and analogy terms to analyze chemical data (S. Wold & Sjöström, 1977). SIMCA operates, by applying PCA to the classes separately, this approach gives more information related to the classes about separation measures and relation between different variables (vanden Branden & Hubert, 2005). Application of the PCA in SIMCA was quite useful since with this way working with high number of variables is possible (Bicciato et al., 2003). Soft, implies that the model can classify multiple classes and if any of the samples were not overlapping with assigned classes, model can identify them as not-classified. It can be used with different number of PC's.

SIMCA as a classification method used for various studies, to classify tea infusions according to their variety and geographical origin with the UV spectrum integration (Diniz et al., 2016), to detect adulteration in pistachios (Menevseoglu et al., 2021) and to discriminate pea berry coffee using 190 and 400nm regions of optical spectroscopy (Suhandy & Yulia, 2017).

1.6.2.4 K-Nearest Neighbors (KNN)

KNN can be considered as a non-parametric and simple classification method, firstly introduced by Fix and Hodges in 1951. KNN simply assigns a test subject to a class or group by following the majority vote procedure. The assigned class is basically the most representative one in training the object of k nearest. When assessing the similarity of two objects, commonly used distance methods are *Mahalanobis distance and Euclidean distance* (Wu & Massart, 1997). In (Figure 1.5) a visual representation of the method is given. Target object is assigned to the class by selecting minimum distances (Otto, 2017).

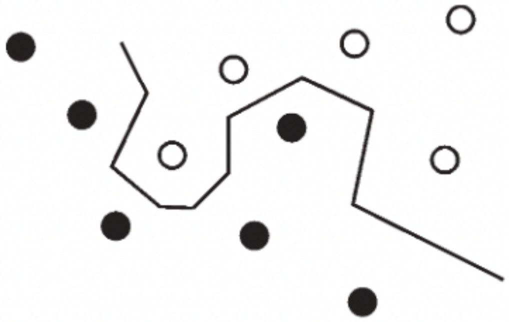


Figure 1.5. Separation boundary of two classes (Otto, 2017).

1.6.2.5 Partial Least Square Regression (PLSR)

PLS method was initially proposed around 1975 by scientist Herman Wold to model complex datasets in the form of matrices and chains and then further developed by algorithms such as Non-linear Iterative Partial Least Squares (NIPALS) (Wold, Sjostrom, et al., 2001). PLSR is a multivariate data analysis tool that is widely used for quantitative analysis on spectral data and relates X matrix, which is absorbance to the y vector, the concentration for this study. With the PLSR models, highly informative results can be achieved about the relation of X matrix and y vector, when compared with traditional regression methods like multiple linear regression, ANOVA and t-tests (Wold, Sjostrom, et al., 2001). In (Figure 1.6) geometric representation of PLSR method is depicted.

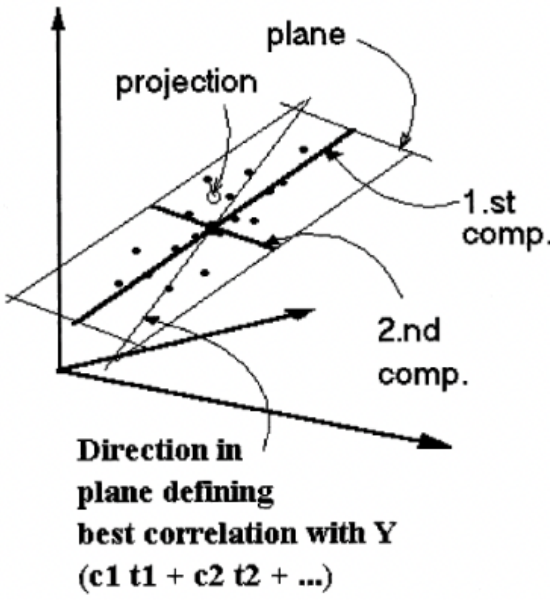


Figure 1.6. Geometric representation of PLSR (Wold, Sjostrom, et al., 2001).

PLSR has been utilized for several regression approaches such as, to predict chemical properties of fish muscle (Cheng & Sun, 2017), quantification of *Curcumae Rhizoma* by using UV-VIS spectroscopy and HPLC (Ren et al., 2021), fast detection of adulteration, made by adding carob flour to the cocoa powder (Quelal Vásconez et al., 2018).

1.6.2.6 Multiple Linear Regression (MLR)

Linear regression is a well-known and a popular regression tool applied for numerous quantification problems. By using LD algorithm, also known as multiple linear regression (MLR) if number of variables is more than one, regression function can be modelled as a linear combination of predictors. Therefore, parameters of the model can be easily interpreted. While estimating model parameters, there are several methods including maximum likelihood, Bayesian approach, robust estimation, least squares and ridge regression (Su et al., 2012). Hovewer, there is a challenge with using linear regression about ill-posed datasets. If number of variables are higher than number of samples MLR cannot be calculated. To solve that problem regularization techniques can be utilized such as ridge regression (Wu et al., 1996). MLR was used for several food analysis applications such as to quantify blends of Coffea arabica and Coffea canephora var. robusta by using UV-VIS spectroscopy and synchronous fluorescence (Dankowska et al., 2017), quantitative analysis of histamine by using raman spectroscopy (Xiao-ying et al., 2019).

In (Table 3), recent applications of above-mentioned chemometric methods which were utilized for food analysis compiled, with preprocessing methods and wavelength intervals.

Table 3. Recent chemometric applications covering UV-VIS-NIR.

Method	Task	Wavelength	Preprocessing	Reference
PCA	Authenticate plant food supplements (Bud extracts)	190-1100 nm	-	(Boggia et al., 2017)
	Improving food classification by combining optical spectroscopy and machine learning	200 -1700 nm	Sequential forward search (SFS)	(Magnus et al., 2021)
	Authentication of Indonesian Honeys by using UV spectroscopy	190-400	Smoothing, SG	(Suhandy & Yulia, 2021)
LDA	Classification of different botanical origins of Curcumae Rhizoma	200-400 nm	FD, SD, SG, SNV, MSC	(Ren et al., 2021)
	Cocoa bean cultivars authentication	10000-4000 cm^{-1}	SNV	(Teye et al., 2016)
	quality grading of in-situ coca beans	900-2500 nm	SNV, MSC	(Kutsanedzie et al., 2017)
Simca	Classify tea infusions according to their variety and geographical origin with the UV spectrum integration	190-800 nm	Successive projections algorithm (SPA)	(Diniz et al., 2016)
	Detecting adulteration in pistachios	200-800 nm	-	(Menevseoglu et al., 2021)
	Discriminating pea berry coffee	190-400 nm	-	(Suhandy & Yulia, 2017)
PLSR	Quantification of Curcumae Rhizoma	200-400 nm	FD, SD, SG, SNV, MSC	(Ren et al., 2021)
	Fast detection of adulteration, made by adding carob flour to the	400-2500 nm	SD, SG	(Quelal Vásconez et al., 2018)
	Predicting chemical properties of fish muscle	400-1700 nm	SNV, MSC	(Cheng & Sun, 2017)
MLR	Quantificatying blends of Coffea arabica and Coffea robusta	190-700 nm	-	(Dankowska et al., 2017)
	Quantitative analysis of histamine	1167-1474 cm^{-1}	-	(Xiao-ying et al., 2019)

1.7 Aim of the Study

Sucrose is produced from both sugar cane and sugar beet. However, some countries are not able to grow sugar cane because of their climatic conditions. Producing sugar cane is more efficient, thus less costly when compared to beet and for this reason, several incidents have occurred throughout the history about the illegal use of sugar cane. Differentiating sugar cane and beet with chemical methods is challenging and suggested methods are expensive, complicated or slow. Therefore, a quick and affordable analysis method that requires less expertise while operation is a need in the industry. The objective of this thesis is to apply *optical spectroscopy* that covers 200-1400 nm range with the implementation of several chemometric methods and to classify and quantify beet and cane sources of sugar.

Hypothesis of the study is formed as; *Using optical spectroscopy in UV-VIS-NIR region with chemometric approaches, to observe spectral signature differences, which occur as a result of impurities, one can distinguish and quantify different sucrose sources, as beet or cane.*

CHAPTER 2

MATERIALS AND METHODS

In this section, properties of materials and description of the methods which are used for this study, are explained. Different chemometric methods were utilized firstly, to differentiate sugar beet and sugar cane sources (*referred as classification*). Secondly, to predict concentrations of binary mixtures which were prepared by using beet and cane sugars (*referred as quantification*) regression methods were applied.

2.1 Materials

Different sucrose samples originated from sugar cane and sugar beet plants were collected from 9 different countries and 23 different brands. For classification purpose, only known source samples and white sugars were used. In order to prepare a solution from sucrose crystals and powders, distilled water was used. In (Table 4), one can see sample information.

Table 4. Grouping of sucrose by its origin (*Code; first two letters are country, third letter is color of sugar, fourth letter is source*)

Sr. No.	**Country**	**Code**	**Source**	
S_01	Pakistan	PKWC	Sugarcane	
S_02	Portugal	PTBC	Sugarcane	
S_03	Portugal	PTWC	Sugarcane	
S_04	Poland	PLW		Unknown
S_05	Poland	PLW		Unknown

Table 4 (cont’d)

S_06	Poland	PLBC	Sugarcane		
S_07	Poland	PLW		Unknown	
S_08	Poland	PLW		Unknown	
S_09	Poland	PLW		Unknown	
S_10	Romenia	ROW		Unknown	
S_11	Italy	ITWC	Sugarcane		
S_12	Italy	ITBC	Sugarcane		
S_13	Serbia	RSWB			Sugarbeet
S_14	Serbia	RSWB			Sugarbeet
S_15	Serbia	RSWB			Sugarbeet
S_16	Serbia	RSWB			Sugarbeet
S_17	Belarus	BLWC	Sugarcane		
S_18	Belarus	BLWB			Sugarbeet
S_19	Belarus	BLWB			Sugarbeet
S_20	Poland	PLW		Unknown	
S_21	Ukraine	UKW		Unknown	
S_22	Colombia	COBC	Sugarcane		
S_23	Belarus	BLWC	Sugarcane		
S_24	Belarus	BLWB			Sugarbeet
S_25	Belarus	BLWB			Sugarbeet

*(For example, PKWC means, sugar collected from Pakistan, it is a white cane sugar, W means white, and B means brown).

2.2 Methods

A summary of the chemometric methods used in the study is given in (Figure 2.1).

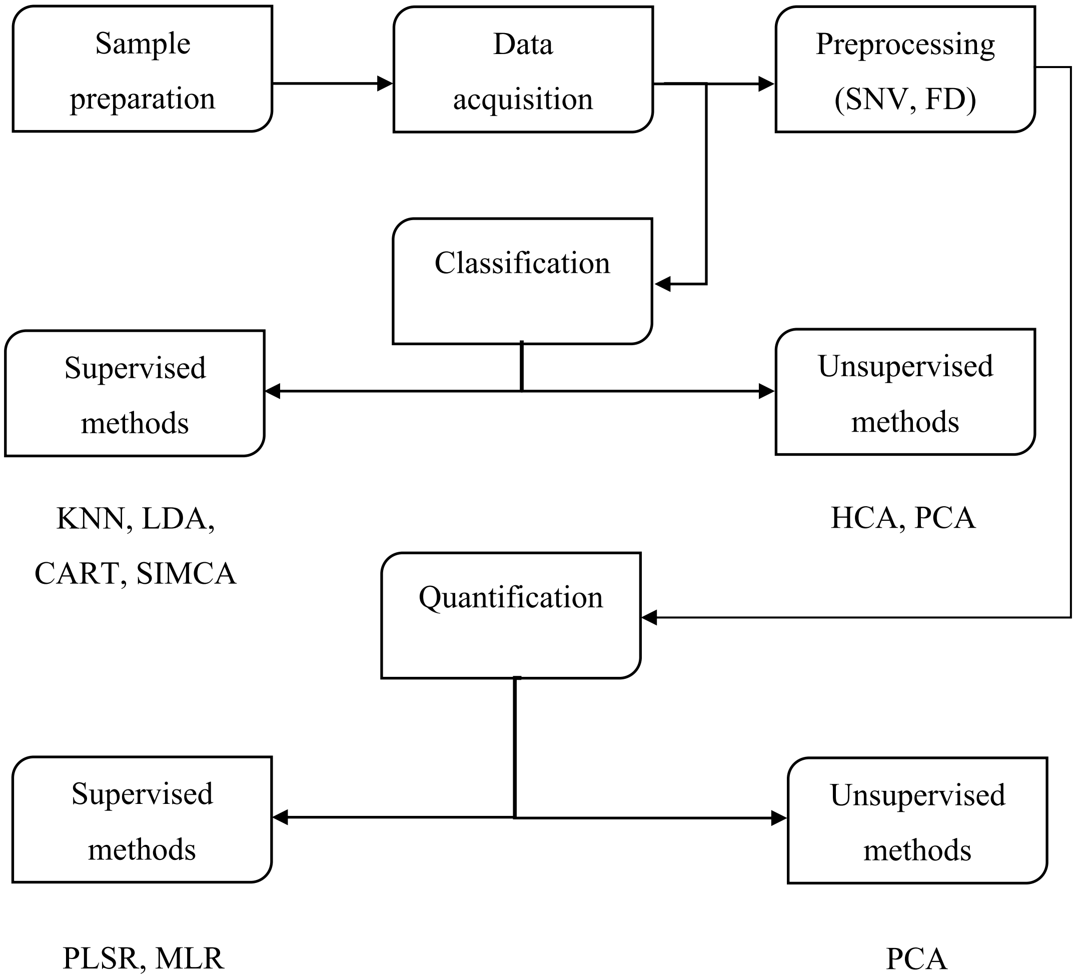


Figure 2.1. Chemometric analysis flowchart.

2.2.1 Preparation of Sucrose Solutions for Spectral Analysis

For classification part, from every sucrose bag, two different 25% (w/w) sucrose solutions were prepared and 5 replicates from each were taken for further analysis.

In quantification part, selected beet and cane sugars were mixed at different concentrations from 0% (w/w) beet sugar to 100% (w/w) beet sugar with 5% (w/w) increments to obtain a binary mixture that has a final concentration of 25% (w/w) sucrose. All samples were mixed properly before adding water since spectroscopy analysis requires good representative sampling. Samples were stirred with magnetic stirrers after addition of water inside beakers for about 10 minutes and following sucrose dissolution, they were placed into quartz cuvettes (10 mm path lenght).

2.2.2 Spectral Measurements

Absorbance data were recorded by using UV-VIS-NIR scanning spectrophotometer UV-3101PC (Shimadzu, Japan) that covered a spectral rage from 200nm to 3000nm. For classification of beet and cane samples, wavelength ranging from 200nm to 1300nm were used with 1 nm spectral resolution. All measurements were conducted at a slow scan speed, with 1nm of slit width and 1 nm of sampling interval.

In quantification part, the wavelength interval was 200nm-600nm with the same 1nm spectral resolution. In classification part of the study, every measurement was recorded with an air reference. However, in quantification part both cuvette holders were used and for the reference, a cuvette filled with distilled water was used to remove the effect of the solvent.

2.2.3 Principal Component Analysis

PCA was mainly used for classification purposes to easily interpret and visualize the multivariate data. For this study, the number of PCs were selected as two for easy visualization of scores in a 2-D graph. And the two PCs were already capable of explaining most of the variation in the data set as discussed later. PCA was applied by using MATLAB® Release 2022a, The MathWorks, Inc., Natick, Massachusetts, United States.

2.2.4 Hierarchical Cluster Analysis (HCA)

HCA was used to have an opinion about the methods which used '*distance-based approaches*', before supervised classification methods. Both Euclidian and Mahalonobis distances were used, and mean centering was applied to the spectroscopic data as a preprocessing. Linkage method was selected as incremental and algorithm was run as sample oriented.

2.2.5 Data Treatment and Preprocessing

Obtained UV-VIS raw spectra were preprocessed with different methods to remove noise and eliminate unnecessary variables. In addition, *first derivate* (FD) was used to remove baseline differences under Savitsky Golay algorithm with 5 windows and 2nd order polynomial. These preprocessing methods were only used for quantification part of the study.

At first, to easily visualize the methods on 2-D graphs, PCA was applied to the dataset and first two components were used for classification methods. Then all variable (wavelength) data were included, and results were obtained. For all chemometric analysis, first 5-fold cross validation (CV) was applied to the whole dataset. Furthermore, the samples were randomly divided into two different forms as train and test samples. In total 124 samples were examined in whole dataset for classification, and later 85 samples were selected for train and 39 as a separate test set. All results are presented in figures and tables in the later sections.

2.2.6 Linear Discriminant Analysis

For classification purposes, linear discriminant analysis (LDA) was used. Since LDA operates well under a specific ratio between sample number and variable number, high number of variables in spectral data, compared to sample number could make the model unstable. To solve that problem first, PCA was applied to decrease the number of factors that was used for LDA and then 5 different wavelengths determined by trial-and-error, based on maximum classification performance.

2.2.7 Decision Trees

Classification and regression tree (CART) provide structural mapping that consists of binary selection (Kotsiantis, 2013). By selecting the variables between numerous input data, algorithm grows tree like shapes with root nodes. Any root that is added to the algorithm is based on an appointed value for one variable, which is also called univariate split. These splits are basically threshold values selected from variables, which are used to differentiate between samples. Main aim of the algorithm is to improve the model performance by adding one split with the least split number possible. For this study, only two roots were applied on PC1 and PC2 scores.

2.2.8 K-Nearest Neighbors (KNN)

Another method of classification is the *k nearest neighbors* (kNN). Classes of this study was assigned by using the majority vote procedure. As a distance method, Mahalonobis distance was used, since it can tackle with collinearity problem, which spectral data have. Number of neighbors was selected as 10 and none of the row or column preprocessing was applied. While assessing the model 5-fold cross validation and separate train and test dataset were used.

2.2.9 Soft Independent Modelling of Class Analogy (SIMCA)

SIMCA is a method that works with PCA, for dimensionality reduction from the dataset PCA is applied. For this study, optimum number of PCs selected as 2. Algorithm tested different number of PCs and shows which one gave the maximum sensitivity, specificity (will be explained in section below) and minimum error.

2.2.10 Evaluation of Classification Models

Evaluation of the classification models were done as seen in (Figure 2.2).

	cane	beet
cane	True Positive (TP)	False Positive (FP)
beet	False Negative (FN)	True Negative (TN)

Figure 2.2. Confusion matrix.

Where;

TP = Sugar cane identified as sugar cane.

TN = Sugar beet identified as sugar beet.

FP = Sugar cane identified as sugar beet.

FN = Sugar beet identified as sugar cane.

Sensitivity = TP/(TP+FN)

Specificity = TN/(TN+FP)

Precision = TP/(TP+FP)

Accuracy = TP+TN/(TP+FP+FN+TN)

2.2.11 Partial Least Squares Regression (PLSR)

Similar to PCA, PLSR also works with component but in this case, they are called latent variables/factors. Also, in PLSR, while applying regression for the chosen dataset, X matrix decomposition is guided by the variance in y vector, thus the main purpose is to increase the co-variation between y and X. Basic linear model can be shown as:

$$y = Xb + e \qquad (1)$$

Where e is residual and b is the vector containing coefficients of regression obtained after model calibration. Operating with latent variables gives the ability to work with spectral data that contains high collinearity (S. Wold, Trygg, et al., 2001). For this study, different number of latent variables (LV) were used such as 2,5 and 8 to increase the models' ability with least number of LV's. Different number of variables were tested, and results are given in Chapter 3.

2.2.12 Variable Selection with PLS

PLS was applied with also using '*searching combination moving window interval* 'PLS (scmwiPLS), which is a variable selection method (Du et al., 2004). In this method, after selecting the size of the windows (number of variables in one window), algorithm decides the best combination of windows with the least RMSEP. Furthermore, different number of windows were compared and selected according to their performance result, which is RMSEP.

In this study, scmwiPLS was used for finding the combinations of informative bands to increase prediction capability of the PLS model. For unprocessed data, 3 windows were used with 6 wavelengths as window size and model was calculated with 5 factors. For the preprocessed data, 7 windows were used with 9 wavelengths as window size and model was calculated with 8 factors.

2.2.13 Multiple Linear Regression

Multiple Linear Regression (MLR) was used to build a quantification model. Ridge regularization was applied for tackling with multicollinearity problem, since MLR gives biased prediction results if independent variables are highly correlated. Also, different wavelengths were tested.

2.2.14 Evaluations of Linear Regression (LR) and Partial Least Squares (PLS) Regression Models

In regression part both linear regression and PLS regression was applied to the dataset. To calculate the performance of the models, dataset was split into two subgroups as test and train samples and following equations were used for RMSEP (also RMSEC), residual prediction deviation (RPD) and coefficient of determination (R^2).

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{2}$$

$$R^2 = 1 - \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}}} \tag{3}$$

$$RPD = \frac{STD}{RMSEP} \tag{4}$$

Where n is the number of samples in the validation set, y_i is reference result for sample i, $\bar{y}$ is mean of y values and $\hat{y}_i$ is the results that were estimated by the algorithms, which corresponds to test sample i Eq. (2). Correlation coefficients between known and predicted values were calculated by Eq. (3). STD means standard deviation and RPD values were calculated by Eq. (4).

Preprocessing methods and spectral regions were selected by considering minimum RMSECV for whole dataset, RMSEP and high R^2 values.

While performing PLS, number of latent variables which covers the adequate variance of data were kept as low as possible to achieve high performing predictions.

2.2.15 Software and Algorithms

PLS without variable selection and linear regression with and without Savitsky Golay and first derivative preprocessing methods was conducted with called Orange (Demšar et al., 2013). ScmwiPLS and all classification methods were performed with MATLAB (The Mathworks, Natick, MA, USA). While performing classification, ‘classification toolbox’ of MATLAB was used (Ballabio & Consonni, 2013).

CHAPTER 3

RESULTS AND DISCUSSION

Sucrose samples, originated from sugar beet and sugar cane were investigated by using a spectrophotometer that covers UV-VIS-NIR regions of the electromagnetic spectrum. Firstly, raw spectra of the samples were examined to extract valuable information from the samples and then different chemometric techniques were applied for a selected subset of samples for classification of sugar beet and cane originated sucrose. Finally, a regression analysis was conducted on sugar beet and cane samples for completion of the method. Results are given by different statistical analysis methods with their evaluation parameters, which are:

- absorbance for different wavelengths in raw spectra,
- sensitivity and specificity for classification of beet and cane samples, and
- RMSECV, RMSEC, RMSEP, R^2 and RPD values for regression analysis.

3.1 Classification of Beet and Cane Sugar

3.1.1 Raw Spectral Data & Sample Selection

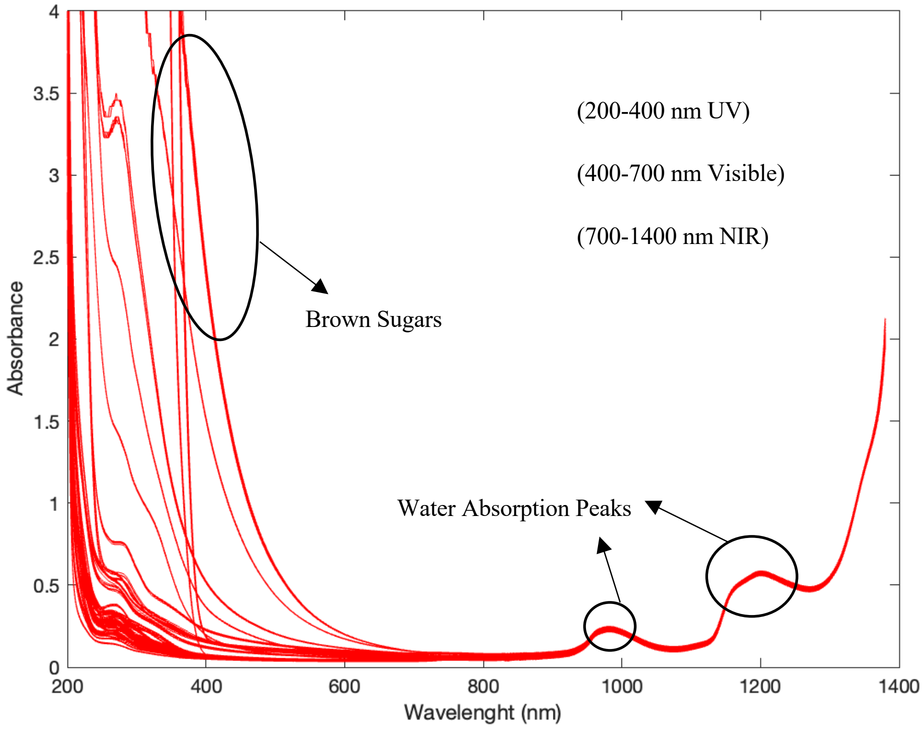


Figure 3.1. Spectral data of all 25 sugar beet and cane samples without preprocessing.

As can be seen from (Figure 3.1), most of the differences were observed in the UV region of the spectra. Possible reasons for this case exist and discussed later. But one can easily understand that by just looking at the UV region, for some samples UV absorbance values were too high for spectrometer to read (detector was saturated), and for other samples absorption data increased drastically. Those samples belonged to brown sugar crystals, since in UV region, colored compounds are highly absorbed. After analyzing all samples including brown sugars, it was decided that such high color deviations should be excluded from spectral data to effectively differentiate sugar beet and cane. The reason for this is some parameters such as color, which are significantly affecting spectral signatures in UV region hide the valuable information required to make regression or classification analysis.

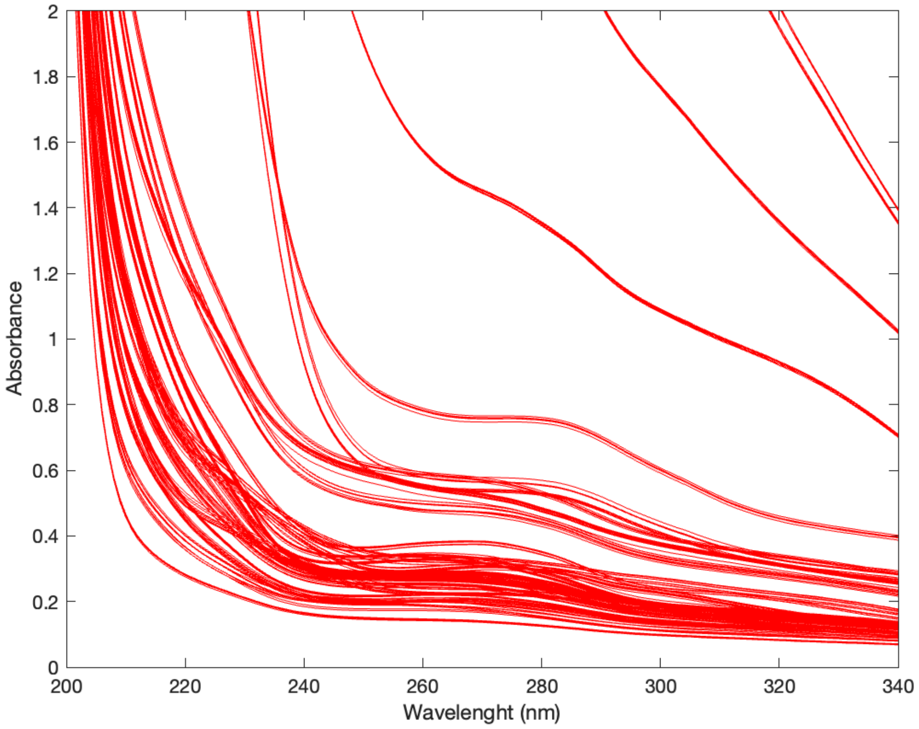


Figure 3.2. Spectral data of all 25 sugar beet and cane samples without preprocessing from 200nm to 340nm.

Following the elution of brown color sugar samples, the spectra were reinvestigated. As can be seen from (Figure 3.2), there were still some outliers, and these were the finely powdered sugars like ‘icing sugar’. During the production of powder sugars, different ingredients are added for anticaking purposes, which may cause blurry sugar solutions (Hollenbach et al., 1982). The presence of starch makes the solution turbid and even after waiting a day for starch to precipitate, results were not promising for most of the cases. Only two powdered sugars could be kept in the experiment dataset which gave reasonable absorbance values after precipitation.

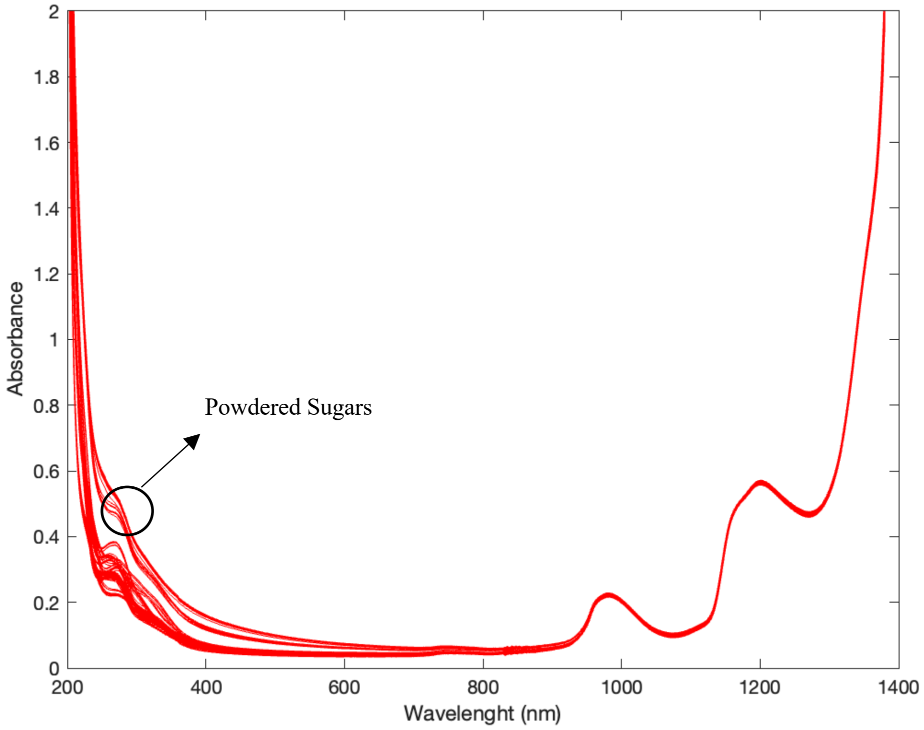


Figure 3.3. Wavelength vs absorbance after outliers were excluded.

Collected raw data from spectrometer after outliers were excluded is given in Figure 3.3. At first, it was expected to have differences in NIR region (800-1400) of the spectra, since there were two significant absorption bands that could be observed in those areas which belonged to water (Palmer & Williams, 1974). First one of those water absorption peaks was around 980 nm and the second one was around 1200 nm. Initially it was thought from a theoretical point that there could be different processing step and strategy while extracting and purifying sucrose molecules from their sources. Those processes might have caused impurities or some molecular level changes that could interact with water and then would change water bands in NIR region so that it would be possible to obtain those spectral signature diversities. However, results showed that there was an observanle difference in only UV region (220-340) not in NIR, which can be easily seen in (Figure 3.4).

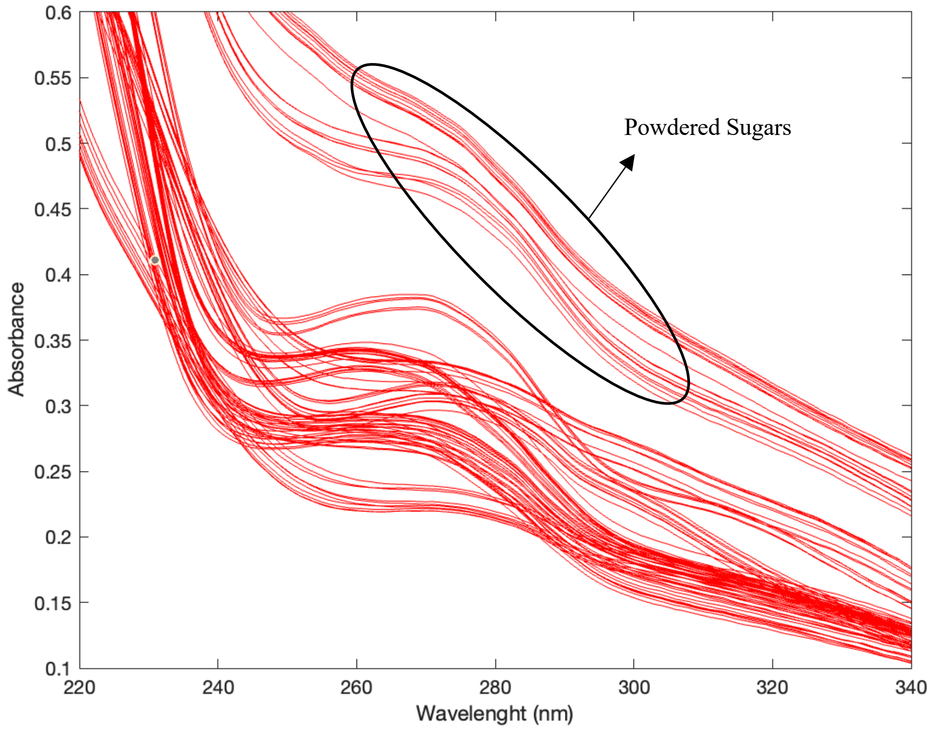


Figure 3.4. Wavelength vs absorbance data of collected sugar samples between 220 nm to 340 nm.

The reason behind that could be the presence of impurities, which can be present in the raw material or mixed during different processing steps Since chemically both cane and beet sugar are quite indifferentiable, both being sucrose with same molecular structure, minor impurities could be a good differentiator under diverse spectroscopic methods (Lu et al., 2017).

3.1.2 Unsupervised Methods

3.1.2.1 Principle Component Analysis

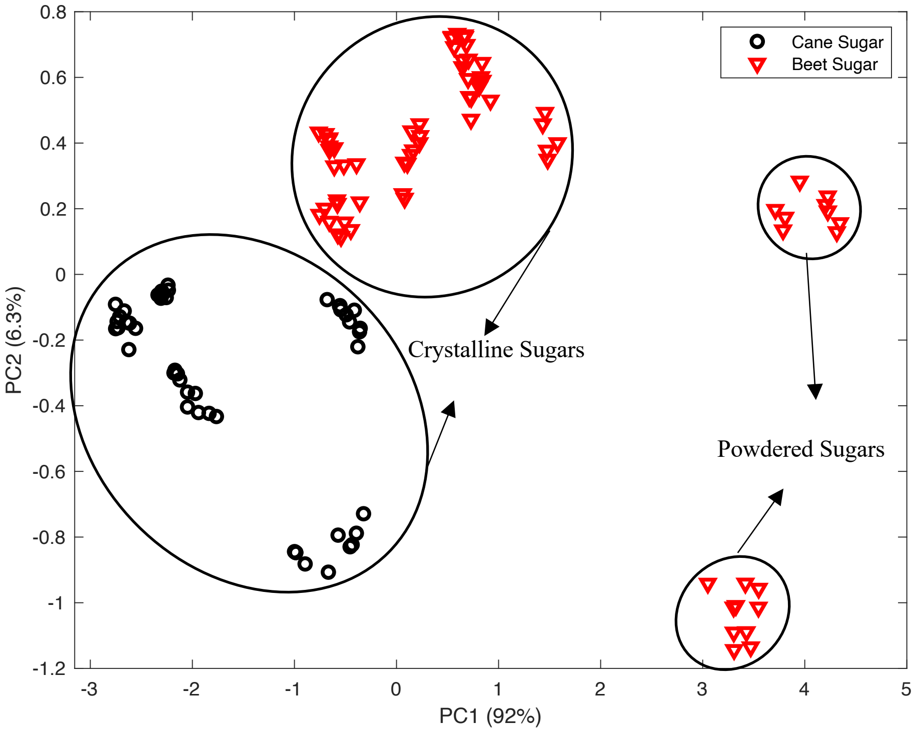


Figure 3.5. PCA score plot of cane and beet sugar.

For qualitative analysis, first PCA was applied to the dataset to detect if any outliers were present and to see if any data clusters existed. PC1 and PC2 explained 98.4% of variance in the dataset which was considerably high. As can be seen from (Figure 3.5) just plotting PC1 and PC2 gave an idea about the differences of cane and beet, regions, and brands. Samples collected from different countries and brands showed small data clusters without a specific pattern about their country. However, samples which belong to the same brand showed similar PC scores that can be observed in (Figure 3.5). It was expected since raw materials that were used for production of the same brand were most likely undergo similar processing steps with similar equipment and the plants sources have been harvested from geographically close regions. Each could be a parameter affecting the spectral signatures of samples with same brands. Short distance between the data points for same brand samples also showed that sampling process was done effectively, since otherwise some large deviations would be observed.

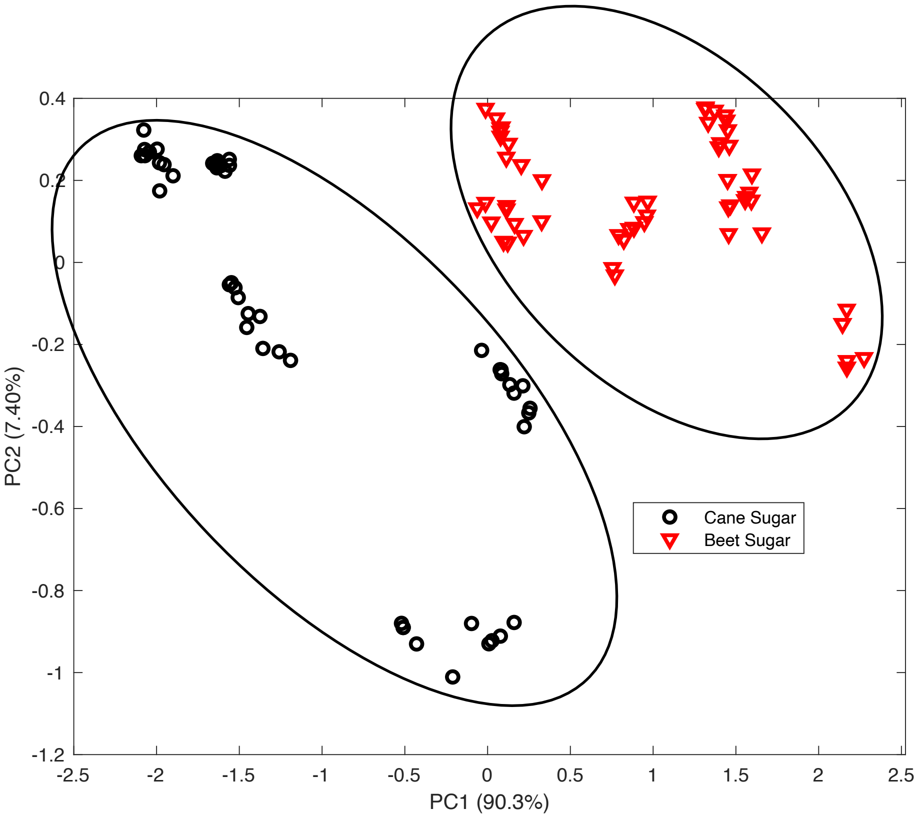


Figure 3.6. PCA score plot of cane and beet sugar (powders excluded).

Another outcome of plotting a PC score graph was the easy detection of powdered sugars. They were separated from crystalline samples significantly even for the same origins (beet & cane). As stated earlier powdered sugars caused spectral shifts because of the additives present for anticaking purposes. Results showed that for powdered sugars when included in the same dataset with crystal sugars, separation by using chemometric methods was still possible (Figure 3.5).

This outcome is of promising since with UV absorbance values one can detect the presence of impurities or foreign materials added during productions steps. On the other hand, if the goal is beet or can differentiation of it is also a challenge because if contamination level is high, it can also mask the differences caused by sugar source and then separation by plant type is compromised. It is a problem than can be solved by some preprocessing; filtering, decoloring etc. to remove impurities that effect the UV absorbance values significantly. Here in (Figure 3.6) after exclusion of powdered samples another PCA was applied and the distance between the clusters of beet and cane samples became more descriptive.

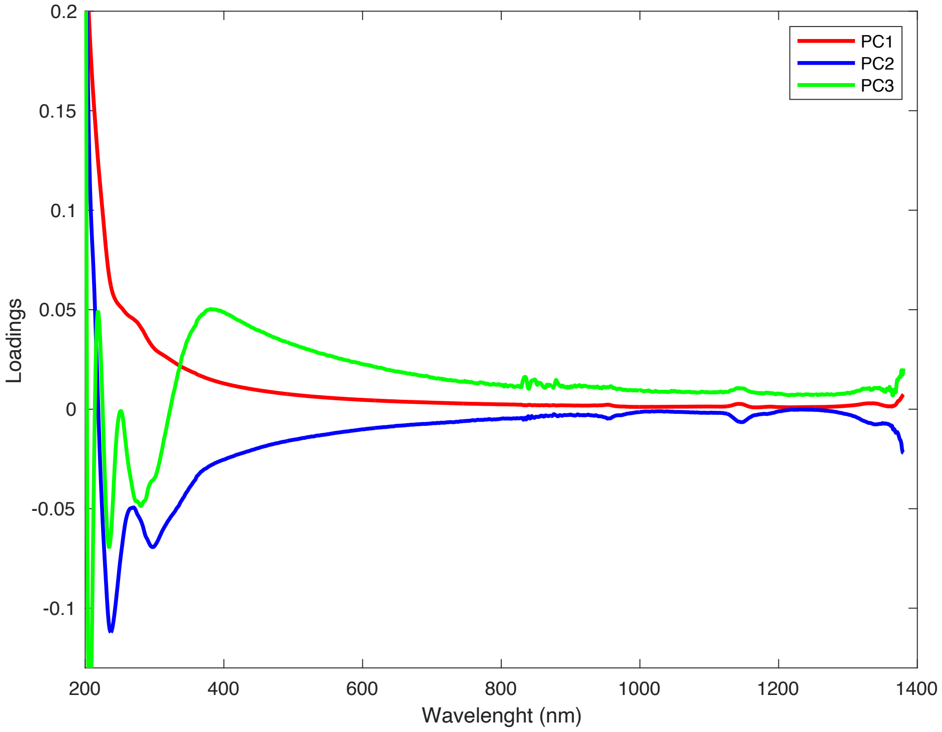


Figure 3.7. PCA loading plot of cane and beet sugar.

PCA loading plots explain the level of importance of the variables to PC scores. For this study, loadings are basically linear combination of wavelengths represented by different principal components (Bro & Smilde, 2014). Here in (Figure 3.7) it is clear that for PC1, the most important wavelengths start at 200nm and as wavelength increases loadings decrease. Findings in the raw spectra were also supported with the loading graphs. PC2 on the other hand gave high loadings around 230 nm, which was similar to PC1. Both principal components was helpful to differentiate sugar brands. PC3 was not included since the goal of this study was fulfilled with PC1 and PC2.

3.1.2.2 Hierarchical Clustering

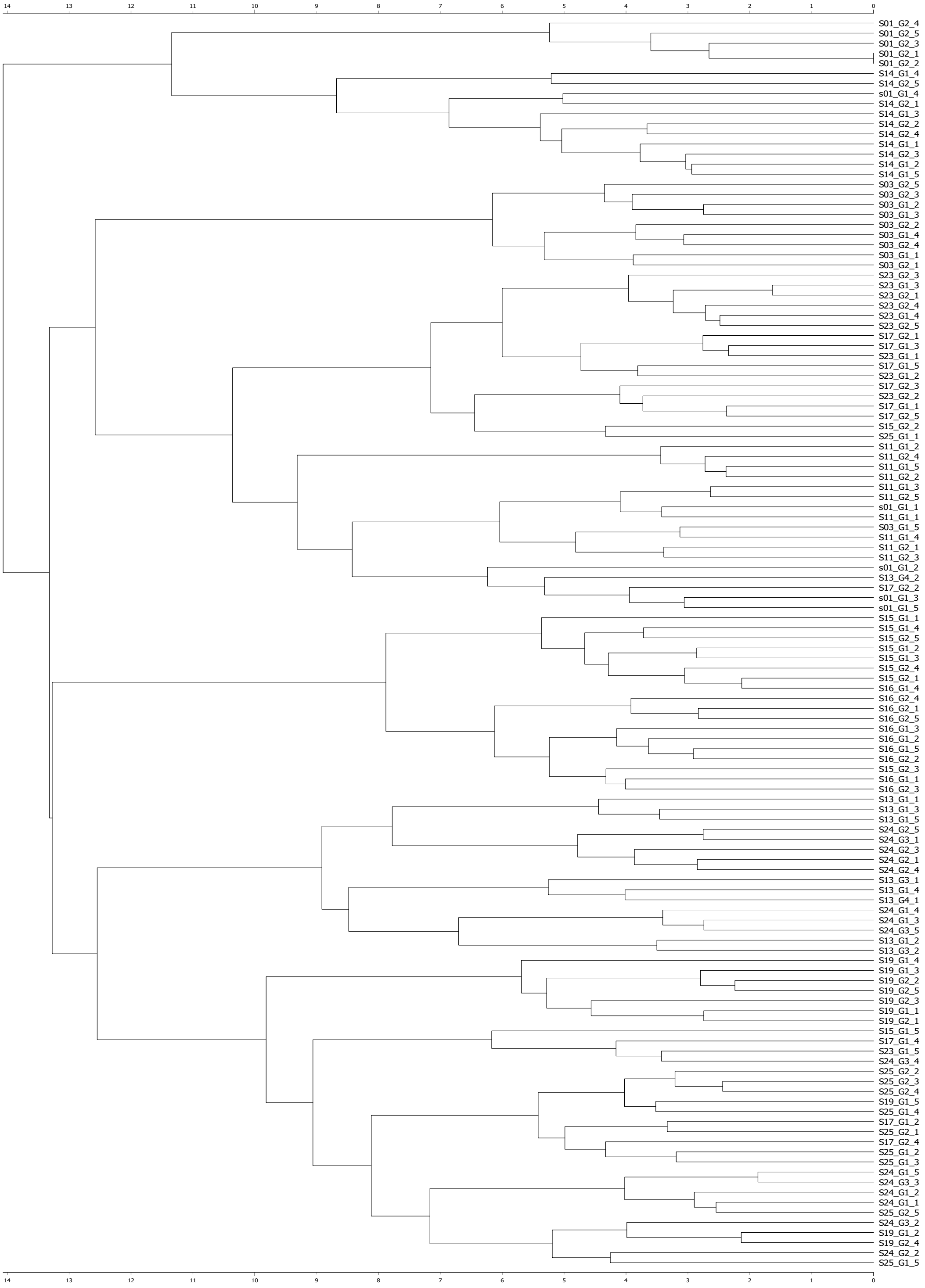


Figure 3.8. HCA clusters of beet and cane sugars.

Hierarchical clustering was also applied to the whole dataset as an unsupervised method. From (Figure 3.8), it can be seen that a good separation was obtained from HCA, as the number of classes started to decrease only after their distance were so small, indicating high similarity. However, with HCA some of the beet and cane sugars were found in the same clusters meaning it was not as successful as *dimensionality reduction* approaches like PCA. Moreover, it was perfect for separating the powdered sugar samples confirming that when it comes to detect high differences such as additives or contaminations, HCA can be suggested.

3.1.3 Supervised Methods

In this part of the study, different chemometric approaches were tested such as CART, LDA, SIMCA and KNN.

3.1.3.1 Linear Discriminant Analysis

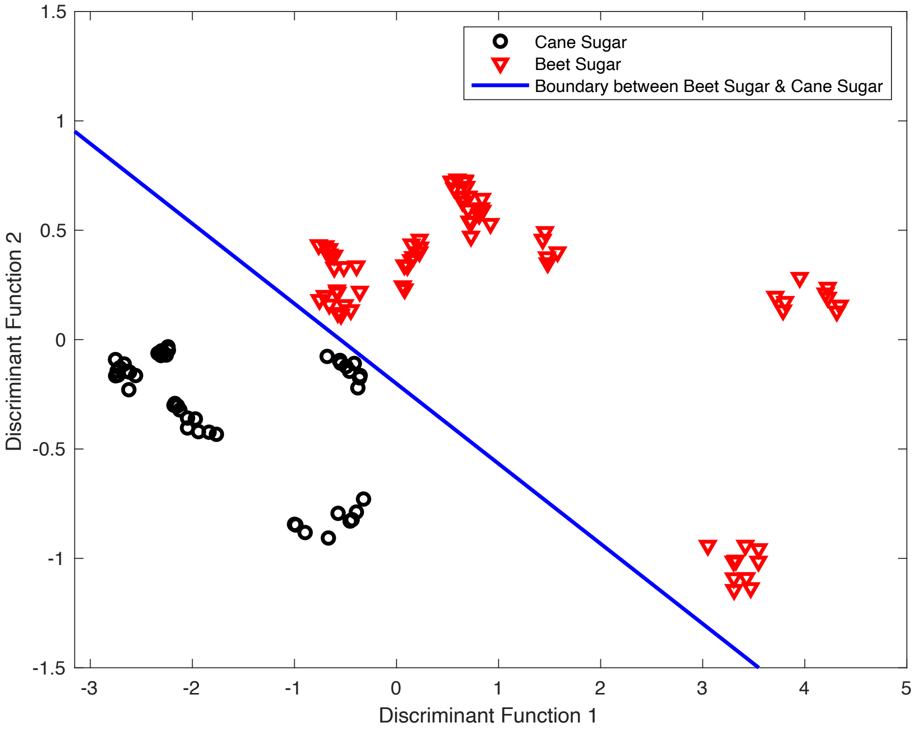


Figure 3.9. Linear discriminant analysis between beet and cane using score values of first two PCs.

As the first approach, LDA was selected, since it is one of the simplest chemometric methods that could be applied to a multivariate dataset. As the name implies, it works with a *linear approach*. If it is applied with two variables as in this case, it is quite intuitive to make comments on the classification. In (Figure 3.9) it was quite clear that by just drawing a linear line between cane and beet samples, it was possible to obtain a differentiation between groups of sugars. However, it has some challenges, since LDA cannot be applied for datasets, which have a greater number of variables than the sample number. Thus, in this study first it was applied by using score values of the first 2 PCs as variables as stated by other researchers (de Luca et al., 2012) . Then, it was applied for 5 selected wavelengths which were 230 nm, 250nm, 255nm, 270nm and 320nm to see if it could easily be applied in industry. Since if one can build a system with less variables, in this case few wavelengths, equipment costs decrease significantly, and model becomes more applicable for industrial purposes. The sensitivity and specificity values were both 1 for both variable selection approaches, meaning all samples were classified correctly. Also results for both whole dataset cross validation and separate train-test sample selection were same, and the algorithm correctly differentiated the samples.

LDA were selected over quadratic discriminant analysis (QDA) because if difference between class covariance matrices was small, LDA performed often better when compared with QDA since low number of estimates were calculated (Wu et al., 1996). In this study they were all sucrose samples hence variances between classes were expected to be small.

3.1.3.2 Classification and Regression Trees (CART)

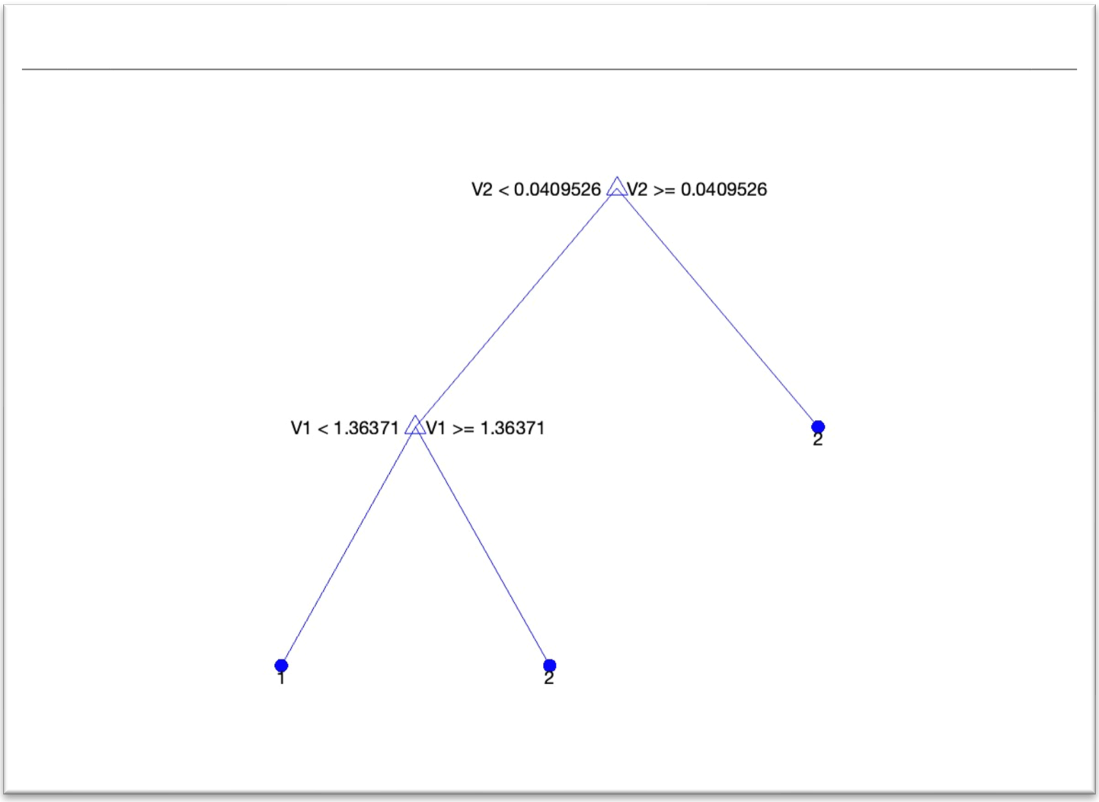


Figure 3.10. Classification nodes of CART analysis (V1 = PC1 & V2 = PC2).

Here in (Figure 3.10) the binary selection of a structural mapping is shown. The first two score values of PCA are used and the algorithm decides a threshold that can separate between two groups and continues in that way. For our dataset, by using only two separation nodes, all 124 samples were classified correct with 5-fold cross-validation. 2 means beet and 1 means cane sugar.

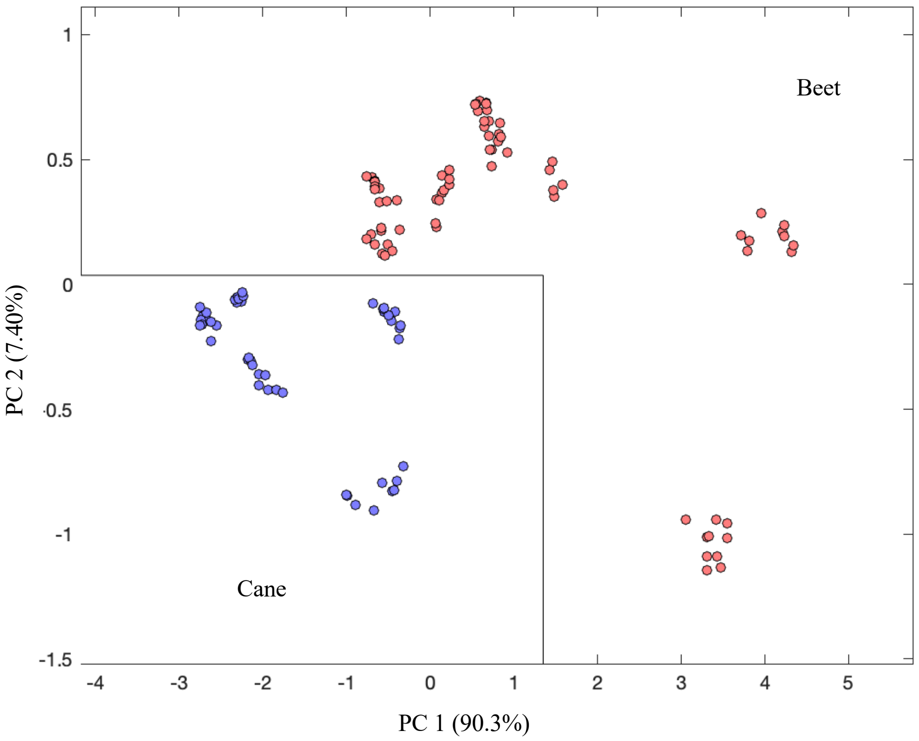


Figure 3.11. Class boundaries of classification tree.

Here in (Figure 3.11) with two differentiation nodes (PC 1 and PC 2) projected to the x-y plane and all samples were differentiated with respect to their sources. PC1 was successful on differentiating powdered samples with the crystal ones. Both PC 1 and PC 2 was quite beneficiary to separate beet sugar from cane.

CART was also tested including all wavelength variables without applying PCA by 5-fold cross validation and results are shown in (Table 5) and (Table 6).

Table 5. Classification measures for whole dataset.

	Sensitivity	Specificity	Precision	Error Rate	Accuracy
Training					
cane	1.00	0.99	0.98	0.01	0.99
beet	0.99	1.00	1.00	0.01	0.99
5-fold CV					
cane	0.98	0.97	0.96	0.02	0.98
beet	0.97	0.98	0.99	0.02	0.98

Table 6. Confusion matrix for whole dataset.

real/predicted	cane	beet	not assigned
cane	49	1	0
beet	2	72	0

As can be seen from (Table 5) and (Table 6), two of the beet samples were classified as cane and one of the cane samples were classified as beet. There was a small error, and it was the only error for all trials in this study.

After the application of CART to the whole data, samples were divided into two as test and train as mentioned above. The results for prediction on the external set was successful for all test samples.

3.1.3.3 K-Nearest Neighbors (KNN)

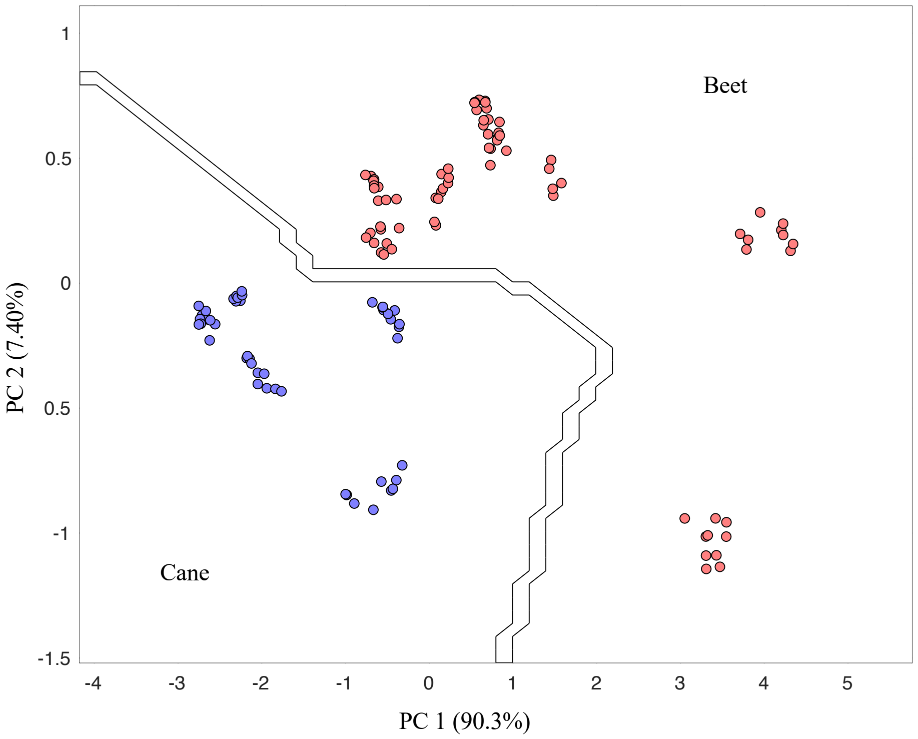


Figure 3.12. K-nearest neighbors class boundaries with PCA 2 components.

In (Figure 3.12), another classification method was presented which classified samples according to the distances (Wu & Massart, 1997). Euclidean distance was the most commonly used similarity approach used for KNN. However, in this research, it was not preferred since Euclidean distance could cause some distortions when there are correlated variables. Highly correlated variables could be found in the techniques such as optical spectroscopy with high spectral resolution (Massart, 1988). Because of the covariance problem, Mahalanobis distance was used, since this approach can cope with correlated variables in the objects found in class.

Nevertheless, it cannot be used with the whole wavelength interval, since KNN with Mahalanobis distance only works if number of variables is less than number of samples, otherwise distance cannot be calculated because of the singularity of the covariance matrix (Wu & Massart, 1997). To solve the high number of variables problem, 15 wavelengths were manually selected from 280-294 nm, which gave the highest classification performance for separate test sets. Again, two different evaluation techniques were tested, and all samples were correctly classified for CV, and separate train dataset. When it comes to separate test samples, results were successful, but 3 misclassifications were observed. Performance measures can be seen in (Table 7) and (Table 8).

Table 7. Classification measures for separate test group.

	Sensitivity	Specificity	Precision	Error Rate	Accuracy
Training					
cane	1.00	1.00	1.00	0	1
beet	1.00	1.00	1.00	0	1
Test					
cane	1	0.90	0.77	0.07	0.92
beet	0.90	1	1	0.07	0.92

Table 8. Confusion matrix for test samples.

real/predicted	cane	beet
cane	10	0
beet	3	26

Also, using PCA with the KNN can solve both collinearity and high number variables problems by reducing the number of dimensions. In this part, first PCA was applied to the whole dataset in (Figure 3.12) and then first two score values were used as variables for KNN classification. After applying KNN with 5-fold CV and train and test sample separation, all samples were identified correctly for all sample sets. These results showed that, some models can work with conditional requirements in the dataset, such as number of variables should be lower than sample number. Even for those methods, successful results were obtained after applying different combinations of chemometric methods.

3.1.3.4 Soft Independent Modelling of Class Analogy (SIMCA)

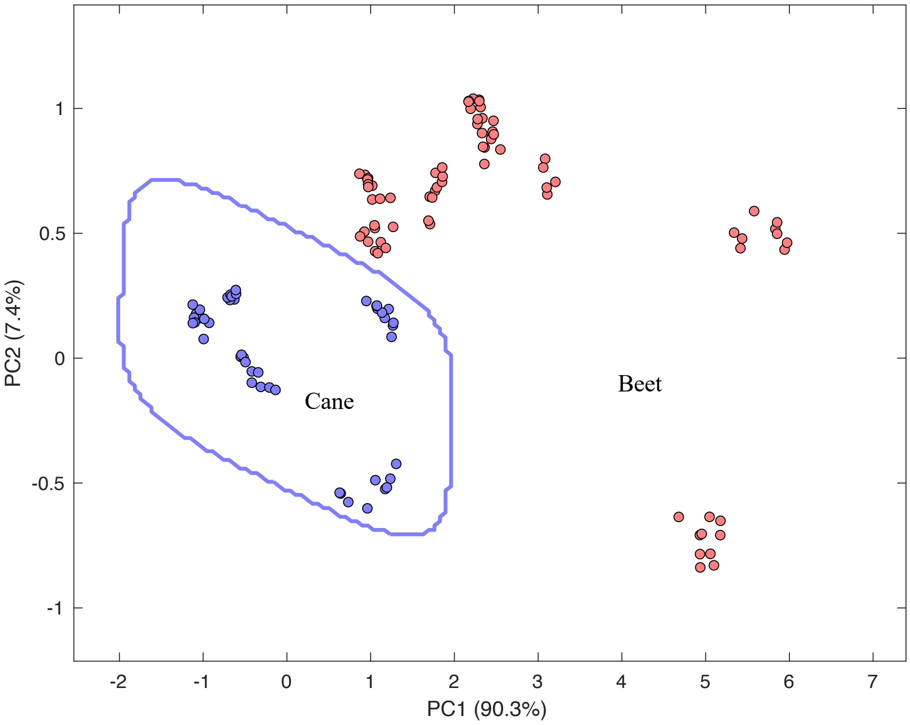


Figure 3.13. Simca class boundaries (one PCA on the whole dataset).

As mentioned in Chapter 2, there are two different ways of applying PCA. SIMCA operates well even with high number of variables. The reason for that is, SIMCA algorithm applies PCA to the groups separately, thus provides a dimensionality reduction. All samples were classified correctly with specificity and sensitivity of 1. However, for easy representation of data in (Figure 3.13) PCA model was applied for whole dataset and then those PCs operated with SIMCA gave also a successful classification. (Figure 3.13), belongs to all dataset without sample separation as test and train and performance of this model is calculated with 5-fold cross validation.

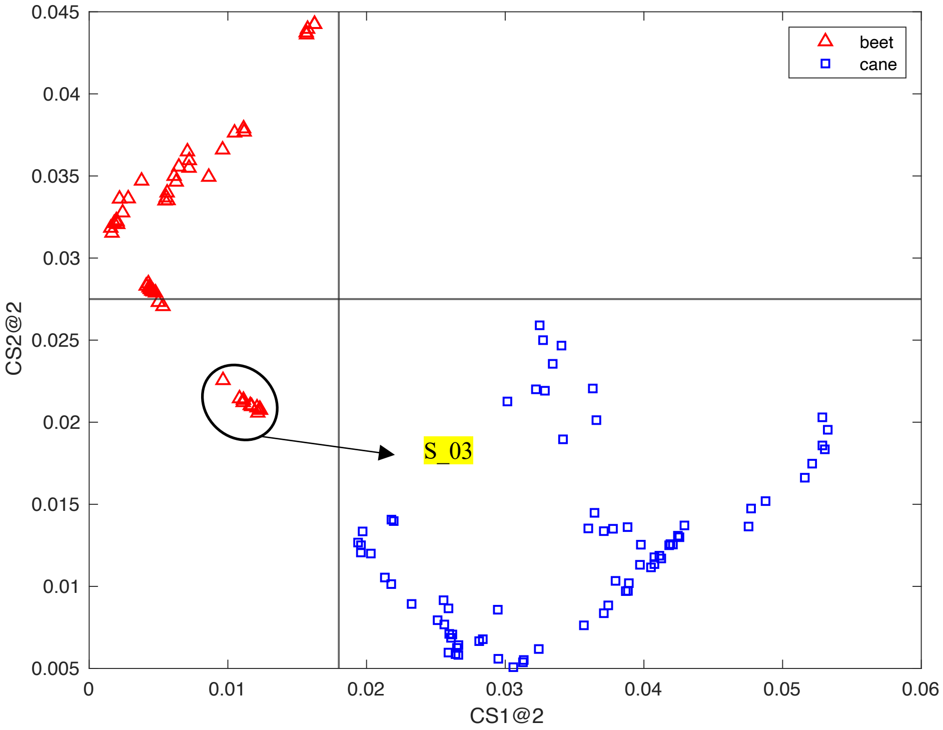


Figure 3.14. Class distances of the model.

(Figure 3.14) shows the class distances of the model and there is a successful separation for all classes. However, even it was classified correctly as beet, circled sample has a possibility to be assigned as a member of both classes. Those results are also complementary with the findings after PCA model was applied.

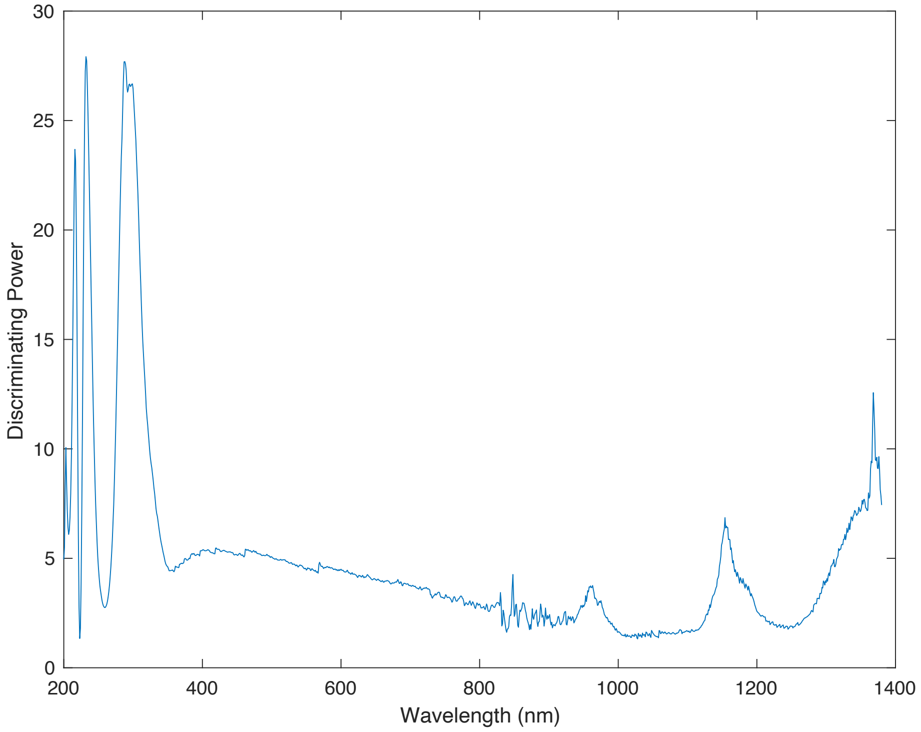


Figure 3.15. Discriminating power (classification power).

Classification power of SIMCA was also complementary with PCA loadings. As can be seen from (Figure 3.15), UV wavelengths have higher discrimination ability when compared with visible and NIR regions.

After applying SIMCA to the whole dataset, samples were also divided into two as train and test and results. All samples were classified correctly for both test and training sets. Thus, by applying SIMCA for two different approaches, a successful classification was achieved.

3.2 Quantification of Beet and Cane Sugar

3.2.1 Raw Spectra

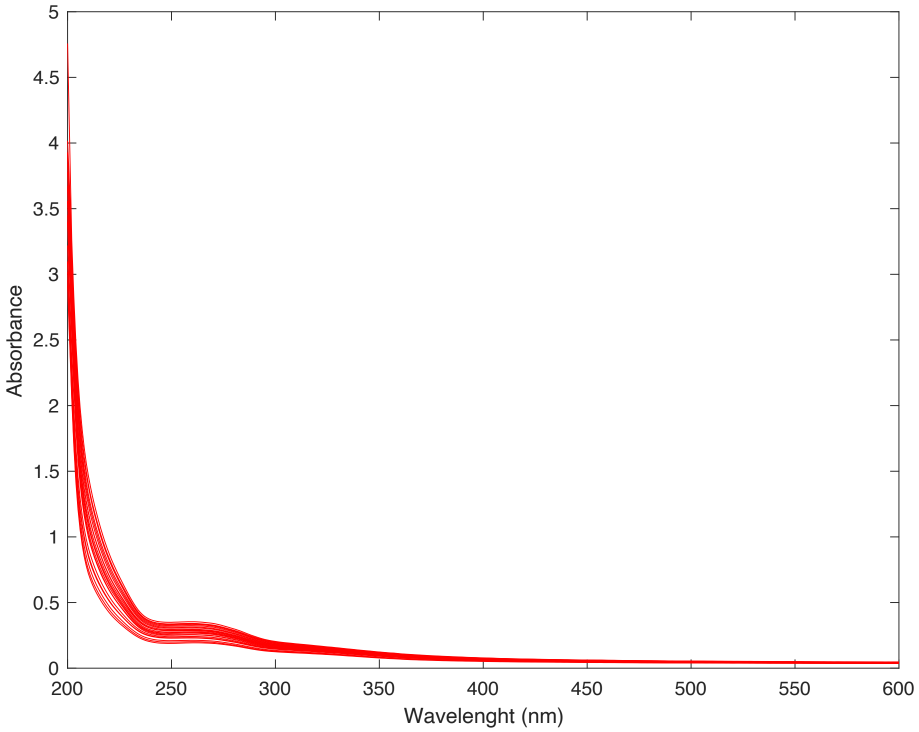


Figure 3.16. Raw spectra of binary beet and cane sugar mixture.

In the regression part of the study, selected wavelength interval was between 200-600 nm, since spectral differences in NIR region were not observable as mentioned. Also, in this part it was observed that after 390 nm which is the end of UV region there were not observable differences compared to shorter wavelength regions. (Figure 3.16) belongs to raw spectra of binary mixture.

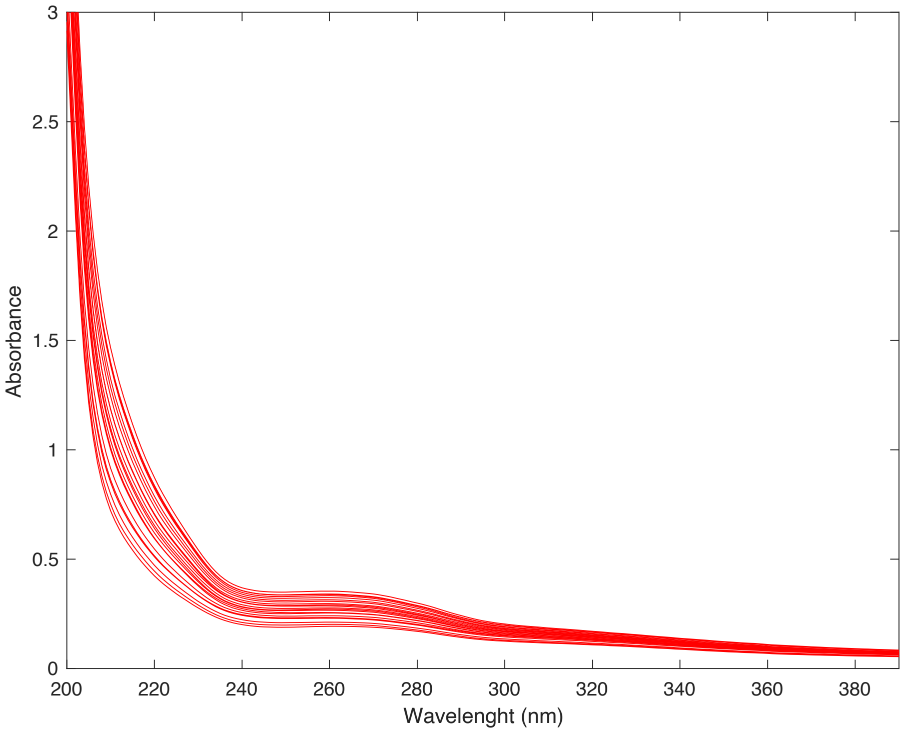


Figure 3.17. Wavelength limited raw spectra of binary beet and cane sugar mixture.

From the (Figure 3.17), it is quite clear that around 270 nm there is a band which was observed also in classification measurements. On the other side around 220 nm, spectral signatures differ in an observable manner. Quantification part of the study was conducted by considering the stated differences with spectral preprocessing and wavelength selection methods.

3.2.2 Principal Component Analysis (PCA)

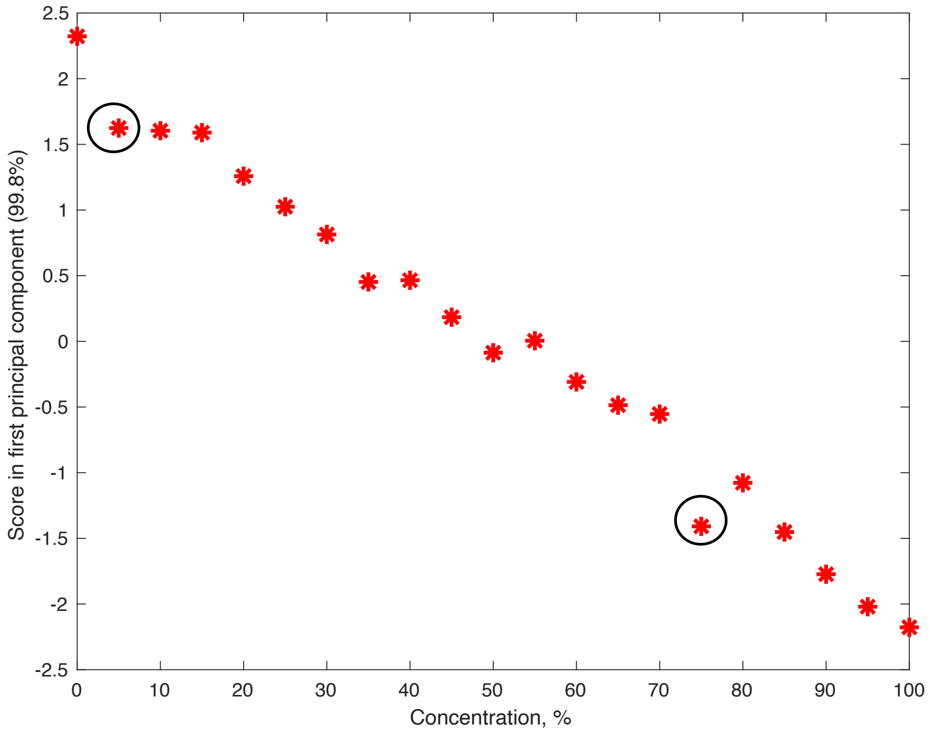


Figure 3.18. Concentration vs score in first component of PCA analysis figure.

A PCA analysis can be performed for a regression problem to see how parameters affect the PC scores and to detect outliers (S. Wold et al., 2007). In (Figure 3.18) the first component was able to explain the variance by 99.8%. This means almost all the variance can be explained by using only one component. And it is also quite clear there is an inverse correlation between concentration and score of the first component.

By applying PCA, also outlier detection is possible. In (Figure 3.18) '*black circled*' concentration values, which belong to 75% and 5% (w/w) beet sugar samples deviated from the sample set. They can be outliers and excluded from the dataset by an outlier analysis. However, in this case since sample number is limited and deviations are not relatively high all samples were decided to be kept for further chemometric analysis. It is important to highlight that, removing samples from dataset could cause overfitting problems and might decrease the robustness of model.

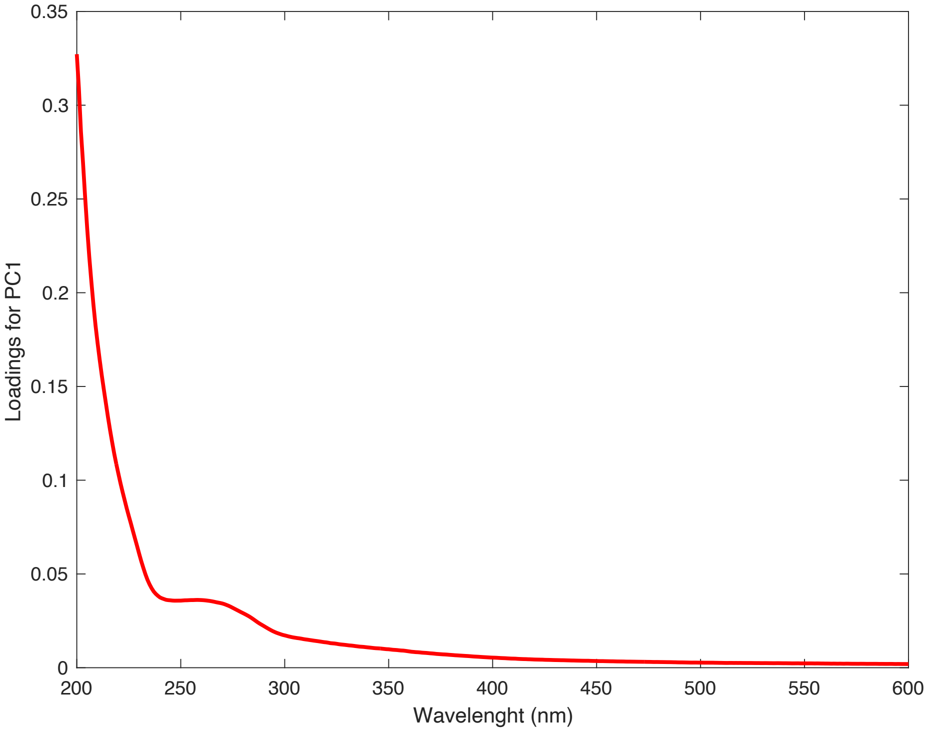


Figure 3.19. Loading graph of first principal component.

From Figure 3.19 as it was expected from the raw spectra, contributions of UV wavelengths to the first principal component, which explained the 99.8% of the variance in whole dataset are the highest. In following sections some manual wavelength selections were applied considering the findings in Figure 3.19.

3.2.3 PLS Regression

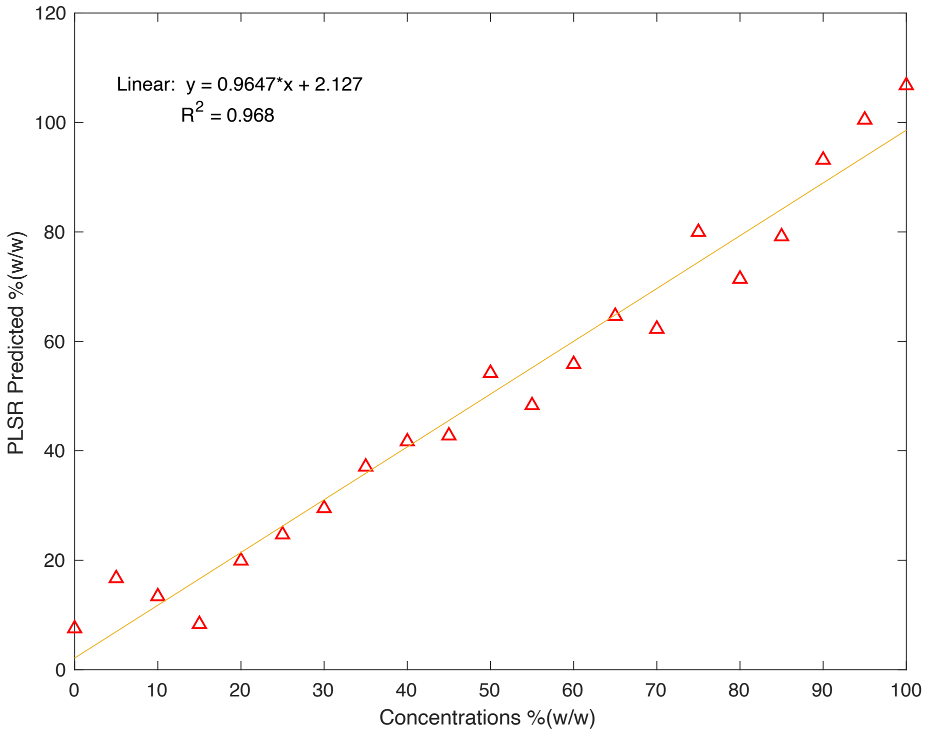


Figure 3.20. PLSR prediction vs concentration graph (5-fold CV unprocessed data).

At first, no preprocessing was performed and only 5-fold cross validation was applied to the whole dataset without any data treatment to see the initial results. From Figure 3.20, it was observed that fit was successful with regression measures; 0.968, 5.433, 5.712 which were *R^2, RMSECV, RPD* respectively. When comparing results with other studies *RPD* is more reliable since it changes with standard deviation. If *RPD* is higher than 3, it is generally accepted that the model is successful.

Following the cross validation for all samples, 6 samples were assigned as test and 15 were to train and *RMSEC, RMSEP* and *RPD* were calculated as 4.191, 6.455, 4.809 respectively. Even if the results were promising, difference between *RMSEC* and *RMSEP* should be low for a good regression model. For that reason, data preprocessing such as Savitsky Golay (SG) and first derivative was applied to the dataset with wavelength selection to enhance the model quality. Also, other methods of preprocessing such as standard normal variate (SNV), normalization, mean centering, second derivative (SD) and Gaussian smoothing were tested but best outcomes were obtained with the mentioned methods, thus their results were discussed.

Application of SG (5 window, second polynomial order) with the first derivative smooths the data and removes baseline effects. Scattering of light might be problem while making spectral measurements that can deviate the results however, with preprocessing methods effects of such deviations could be removed. Results were 2.155, 3.192 and 9.724 for *RMSEC, RMSEP* and *RPD* respectively. With preprocessing, errors were decreased and *RPD* was increased significantly, also difference between *RMSEC* and *RMSEP* became smaller.

After applying preprocessing, wavelength selection could be a good option since wavelengths which are not related with the target outcome can cause misevaluations. After selection of the wavelengths between 200-300 nm results were 2.834, 2.225 and 13.941 for *RMSEC, RMSEP* and *RPD* respectively. The difference between *RMSEC*, *RMSEP* were decreased and *RPD* was increased which means preprocessing and wavelength selection strategies were increased the model capabilities. The R^2 values of above-mentioned models are given in (Table 9).

3.2.4 Searching Combination Moving Window Interval PLSR

When it comes to wavelength selection methods, scmwiPLS is one of the novel ones. After selecting window size, which is number of variables in one window, algorithm tries different number of windows for different variables and selects the one which gives the less *RMSEP*. Number of windows selection can be seen on (Figure 3.21). It is a very successful automized method since one cannot try all combinations by hand and assess all the results in such short times. To apply method, first a PLS without any wavelength selection was applied, to see latent structure number, which gives minimum *RMSECV* and then window length was selected by adding one to the component number. The reason was after several trials from different datasets, this application gave best results.

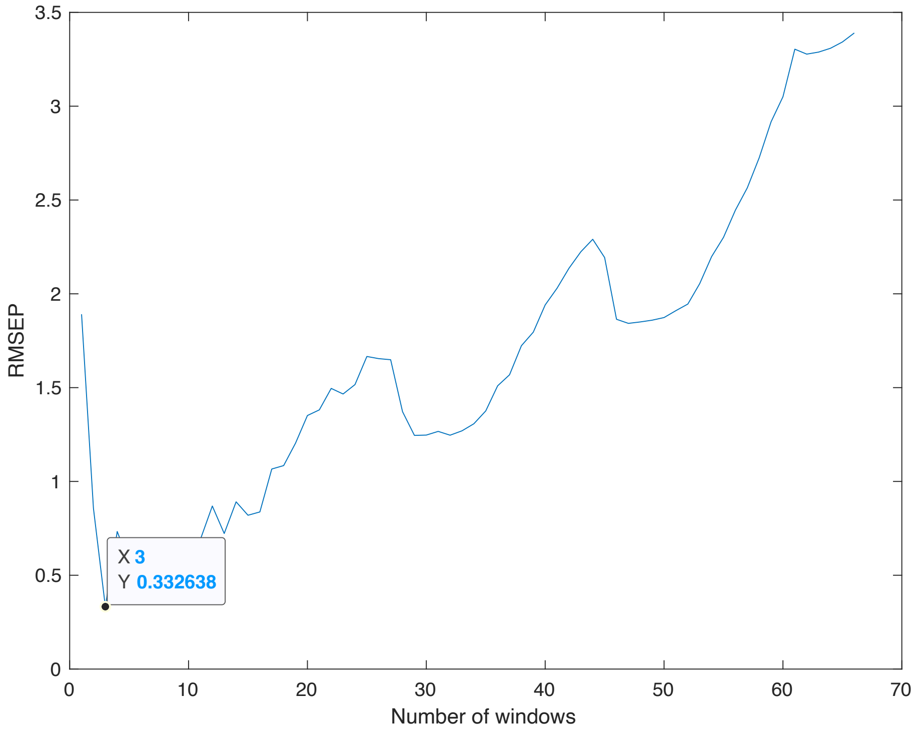


Figure 3.21. Number of windows vs RMSEP.

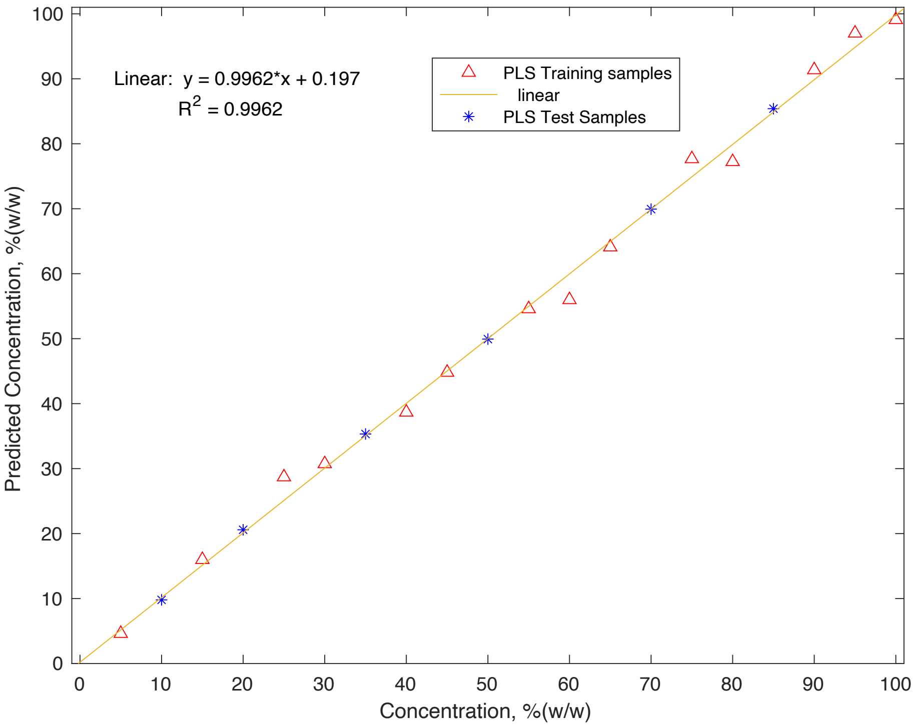


Figure 3.22. ScwmiPLS results without any preprocessing.

As can be seen from (Figure 3.22) PLS was successful in terms of predicting the test samples with *RMSEP* of 0.332 and with the highest *RPD* of 93.433 and with *RMSEC* of 1.936. This model gave the highest results so far. However, there was a challenge with these high-performance models. One of them was since wavelengths were selected for the lowest *RMSEP*, the difference between *RMSEC* vs *RMSEP* could be relatively high which was also the case in this model. This difference can decrease robustness of models, for this reason, scwmiPLS was not selected as the best option in current work.

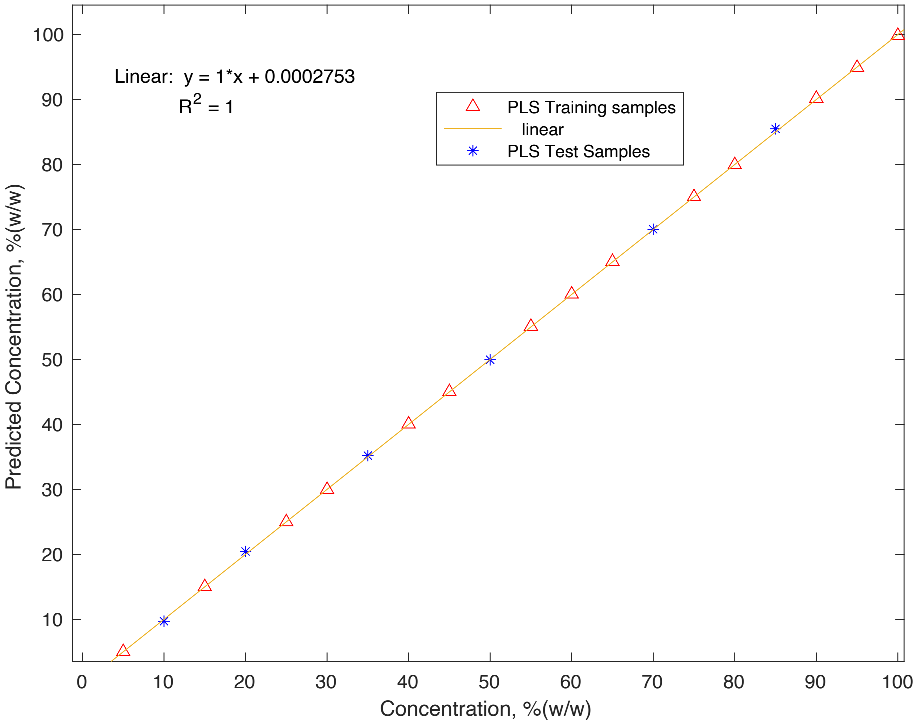


Figure 3.23. ScmwiPLS concentration vs prediction results with SG and First Derivative.

Results obtained from ScmwiPLS can be enhanced with the preprocessing methods. From (Figure 3.23) it was quite clear that the best fit was obtained with preprocessing methods. Evaluation results were 0.072, 0.308 and 100.714 for *RMSEC, RMSEP* and *RPD,* respectively. With preprocessing methods, difference between calibration set and prediction set decreased with even higher *RPD*. However, results obtained from ScmwiPLS could be evaluated as overfitting since possible deviations from other parameters in such perfectly fitted models are most likely ignored. In chemometrics such overfitted models should always be considered and robustness of models is more important for future applications when compared with high success for a specific dataset.

3.2.5 Linear Regression

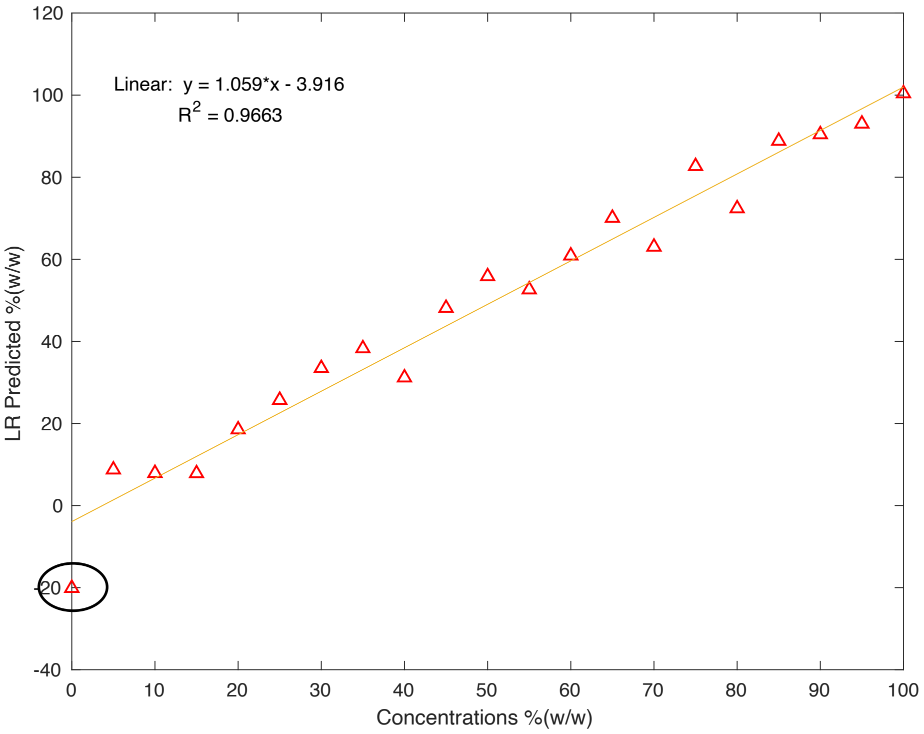


Figure 3.24. Linear regression analysis without any preprocessing, whole dataset.

Linear regression is a method which works well under well-posed datasets which have less variables than sample amount and works well if multicollinearity does not exist in the dataset. However, spectroscopic data with whole wavelength interval was not very convenient for the above conditions. In (Figure 3.24) ridge regularization was applied to solve ill-posed data problem and results were obtained. *RMSECV* with 5-fold CV for whole dataset was 6.318 and *RPD* was 4.909.

Results were very similar but less successful than the PLSR-CV model which had entire wavelength ranges (200-700 nm) and no pretreatments. However, there was a huge prediction error shown in (Figure 3.24) for sample which was circled in black, and the reason is most probably the spectral noise and scattering affects. This problem could have been solved by applying spectral preprocessing methods as described in the following sections.

Samples were then split into two as test and train datasets and other data analysis strategies were applied. When data were split, evaluation results were 1.567, 3.784 and 8.197 for *RMSEC, RMSEP* and *RPD* respectively. Difference between calibration and prediction datasets was high thus further preprocessing methods were also tested. First SG with first derivative was tried and the results were better in terms of consistency, having 2.956, 3.026 and 10.251 for *RMSEC, RMSEP* and *RPD* respectively. As expected by applying preprocessing, stability between test and train datasets can be achieved since noise can be excluded from the spectrum with preprocessing methods.

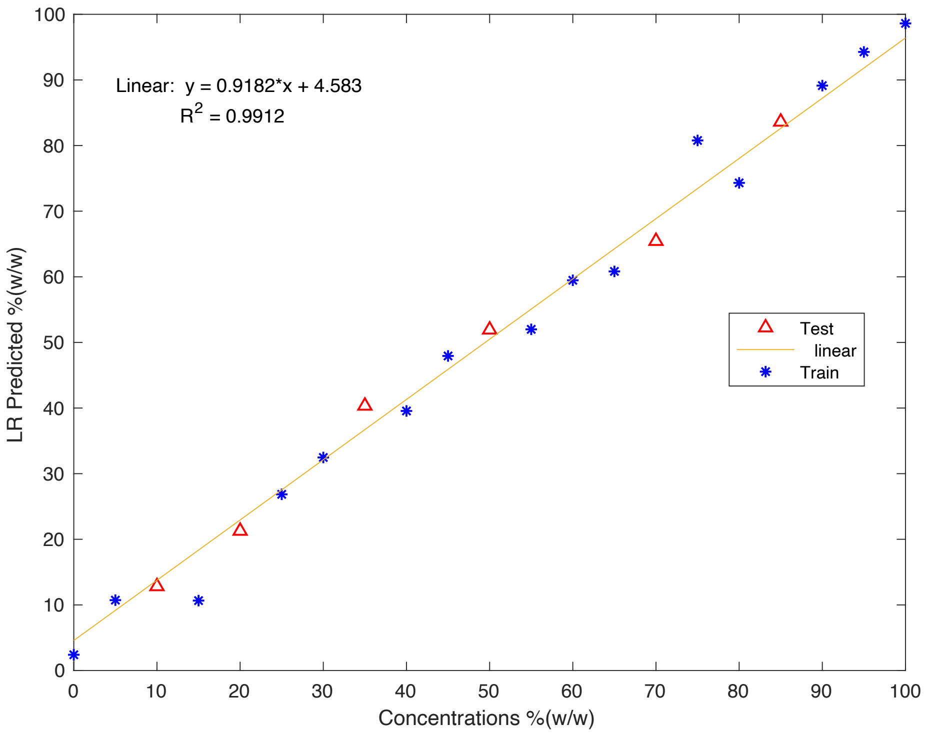


Figure 3.25. Linear regression analysis with SG and FD for 6 selected wavelengths.

Finally, after applying the preprocessing methods, 6 wavelengths were selected from UV spectrum for easy application of the method, which were 220, 221, 222, 223, 224, 225 (*determined by trial and error*). The numbers were 3.382, 3.279 and 9.460 for *RMSEC, RMSEP* and *RPD* respectively. Concentration vs predicted result can be seen in (Figure 3.25). Results for narrow wavelength interval seemed promising for the industrial applications.

Table 9. Regression results

Model	# of Variables	Factors	Calibration Set		Prediction Set		RPD
			Rc^2	RMSEC	Rp^2	RMSEP	
PLSR	Full	2	0.982	4.191	0.940	6.455	4.809
PLSR (SG + 1stDerivative)	Full	2	0.995	2.156	0.985	3.192	9.724
PLSR (SG + 1stDerivative)	100	2	0.992	2.834	0.993	2.225	13.941
scmwiPLSR	18	5	-	1.936	0.996	0.332	93.433
scmwiPLSR (SG + 1stDer)	72	8	-	0.0724	1	0.3085	100.714
Linear Regression	Full	-	0.998	1.567	0.980	3.784	8.197
Linear Regression (SG + 1stDer)	Full	-	0.991	2.956	0.987	3.026	10.251
Linear Regression (SG + 1stDer)	6	-	0.988	3.382	0.985	3.279	9.460

CHAPTER 4

CONCLUSION

Sucrose is mainly produced from either sugar beet or cane. Due to geographical and climatic conditions, some countries can only produce sucrose from one of these sources. Some government policies are restricting the use of cane sugar because of strategic reasons. Also, sensory properties of sucrose produced from sugar beet or sugar cane can be different. Due to these reasons, differentiation of the plant origin of sucrose becomes important.

Since many years, methods such as IRMS, HR-NMR and sensory analysis are being used to differentiate the source of the sucrose. These methods focused on differences between isotropic ratios, sensory properties, impurities etc... On the other hand, sucrose, which is obtained from either source are identical in chemical perspective. Due to this reason, with conventional methods such as HPLC, it is not likely to differentiate sources.

Current work focuses on developing a new method that requires low maintenance and operating costs, less expertise, minimal sample preparation with user friendly interface when compared with previous techniques.

Optical spectroscopy with chemometric methods, provided quite promising results for many studies in differentiating origins of food materials. However, for our knowledge, UV spectroscopy was not studied to authenticate sucrose sources in the literature, which made current work crucial. Results of this thesis showed that UV region of the electromagnetic spectrum was highly sensitive for impurities that could be used to diversify sources of sugar.

As a conclusion, all supervised classification methods, including SIMCA, LDA, KNN and CART, showed high performance to authenticate the source of the sucrose. In addition to that, LDA with only 5 selected wavelengths provided 100% classification with the simplest interpretation. On the other hand, for regression analysis, MLR with Savitsky Golay (SG) and first derivative preprocessings, gave the most stable results of RMSEC, RMSEP by being close to each other 2.956, 3.026 respectively. Moreover, MLR also provided high RPD value of 10.251. The obtained results seem promising that the plant source of sucrose can be differentiated by using UV spectra and chemometric methods.

REFERENCES

Alamprese, C., Casale, M., Sinelli, N., Lanteri, S., & Casiraghi, E. (2013). Detection of minced beef adulteration with turkey meat by UV-vis, NIR and MIR spectroscopy. *LWT - Food Science and Technology*, *53*(1), 225–232. https://doi.org/10.1016/j.lwt.2013.01.027

Asadi, M. (2007). Beet-Sugar Handbook. Wiley-Interscience.

Azcarate, S. M., Cantarelli, M. Á., Pellerano, R. G., Marchevsky, E. J., & Camiña, J. M. (2013). Classification of argentinean sauvignon blanc wines by UV spectroscopy and chemometric methods. *Journal of Food Science*, *78*(3). https://doi.org/10.1111/1750-3841.12060

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: Linear models. PLS-DA. In *Analytical Methods* (Vol. 5, Issue 16, pp. 3790–3798). https://doi.org/10.1039/c3ay40582f

Bicciato, S., Luchini, A., & di Bello, C. (2003). PCA disjoint models for multiclass cancer analysis using gene expression data. *Bioinformatics*, *19*(5), 571–578. https://doi.org/10.1093/bioinformatics/btg051

Boggia, R., Turrini, F., Anselmo, M., Zunin, P., Donno, D., & Beccaro, G. L. (2017). Feasibility of UV–VIS–Fluorescence spectroscopy combined with pattern recognition techniques to authenticate a new category of plant food supplements. *Journal of Food Science and Technology*, *54*(8), 2422–2432. https://doi.org/10.1007/s13197-017-2684-7

Breiman Leo, Friedman Jerome H., Olshen Richard A., & Stone Charles J. (1984). *Classification And Regression Trees*.

Brereton, R. G. (2003). *Chemometrics : data analysis for the laboratory and chemical plant*. John Wiley & Sons.

Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, *6*(9), 2812–2831. https://doi.org/10.1039/c3ay41907j

Bubník, Z., Kadlec, P., Urban, D., Bruhns, M. (1995). Chemical and physical data forsugar manufacturers and users. In: Sugar Technologists Manual, eighth ed. Bartens Pub. Co, Berlin, Germany, p. 417.

Cámara, M., Torrecilla, J. S., Caceres, J. O., Cortes Sánchez Mata, M., & Fernandez-Ruiz, V. (2010). Neural network analysis of spectroscopic data of lycopene and β-carotene content in food samples compared to HPLC-UV-Vis. *Journal of Agricultural and Food Chemistry*, *58*(1), 72–75. https://doi.org/10.1021/jf902466x

Cheng, J. H., & Sun, D. W. (2017). Partial Least Squares Regression (PLSR) Applied to NIR and HSI Spectral Data Modeling to Predict Chemical Properties of Fish Muscle. *Food Engineering Reviews*, *9*(1), 36–49. https://doi.org/10.1007/s12393-016-9147-1

Dankowska, A., Domagała, A., & Kowalewski, W. (2017). Quantification of Coffea arabica and Coffea canephora var. robustaconcentration in blends by means of synchronous fluorescence and UV-Visspectroscopies. *Talanta*, *172*, 215–220. https://doi.org/10.1016/j.talanta.2017.05.036

Dankowska, A., & Kowalewski, W. (2019). Tea types classification with data fusion of UV–Vis, synchronous fluorescence and NIR spectroscopies and chemometric analysis. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, *211*, 195–202. https://doi.org/10.1016/j.saa.2018.11.063

de Luca, M., Terouzi, W., Kzaiber, F., Ioele, G., Oussama, A., & Ragno, G. (2012). Classification of moroccan olive cultivars by linear discriminant analysis applied to ATR-FTIR spectra of endocarps. *International Journal of Food Science and Technology*, *47*(6), 1286–1292. https://doi.org/10.1111/j.1365-2621.2012.02972.x

Demšar, J., Erjavec, A., Hočevar, T., Milutinovič, M., Možina, M., Toplak, M., Umek, L., Zbontar, J., & Zupan, B. (2013). Orange: Data Mining Toolbox in Python Tomaž Curk Matija Polajnar Laň Zagar. In *Journal of Machine Learning Research* (Vol. 14).

Diniz, P. H. G. D., Barbosa, M. F., de Melo Milanez, K. D. T., Pistonesi, M. F., & de Araújo, M. C. U. (2016). Using UV-Vis spectroscopy for simultaneous geographical and varietal classification of tea infusions simulating a home-made tea cup. *Food Chemistry*, *192*, 374–379. https://doi.org/10.1016/j.foodchem.2015.07.022

Du, Y. P., Liang, Y. Z., Jiang, J. H., Berry, R. J., & Ozaki, Y. (2004). Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares. *Analytica Chimica Acta*, *501*(2), 183–191. https://doi.org/10.1016/j.aca.2003.09.041

Eggleston, G., Pollach, G., Triche, R. (2005). The use of ion chromatography profiles as a screening tool to differentiate cane white sugar from beet white sugar. Zuckerindustrie. 130(8):611-616.

Esbensen, K. H., & Geladi, P. (2009). Principal Component Analysis: Concept, Geometrical Interpretation, Mathematical Background, Algorithms, History, Practice. *Comprehensive Chemometrics*, *2*, 211–226. https://doi.org/10.1016/B978-044452701-1.00043-0

Eştürk, Ö. (2018). Türkiye' de Şeker Sektörünün Önemi ve Geleceği Üzerine Bir Değerlendirme. *Anadolu İktisat ve İşletme Dergisi*, *2*(1), 67–81.

Fanelli, V., Mascio, I., Miazzi, M. M., Savoia, M. A., de Giovanni, C., & Montemurro, C. (2021). Molecular approaches to agri-food traceability and authentication: An updated review. In *Foods* (Vol. 10, Issue 7). MDPI AG. https://doi.org/10.3390/foods10071644

Gad, H. A., El-Ahmady, S. H., Abou-Shoer, M. I., & Al-Azizi, M. M. (2013). A modern approach to the authentication and quality assessment of thyme using UV spectroscopy and chemometric analysis. *Phytochemical Analysis*, *24*(6), 520–526. https://doi.org/10.1002/pca.2426

Godshall MA, Grimm CC, Clarke MA. 1995. Sensory properties of white beet sugars. Int SugarJ 97(1159B):296–343.

González Martín I, Marqués Macías E, Sánchez Sánchez, J., &amp; González Rivera, B. (1998). Detection of honey adulteration with Beet Sugar using stable isotope methodology. Food Chemistry, 61(3), 281–286. https://doi.org/10.1016/s0308-8146(97)00101-5

Hollenbach, A. M., Peleg, M., & Rufner, R. (1982). Effect of four anticaking agents on the bulk characteristics of ground sugar. Journal of Food Science, 47(2), 538–544. https://doi.org/10.1111/j.1365-2621.1982.tb10119.x

Kaya, F. (2015). Küresel ve Bölgesel Şeker Politikalarının Türkiye Şeker Fabrikalarına Etkilerine Bir Örnek ; Ağrı Şeker Fabrikası. *İstanbul Üniversitesi Edebiyat Fakültesi Coğrafya Dergisi*, *31*.

Kotsiantis, S. B. (2013). Decision trees: A recent overview. In *Artificial Intelligence Review* (Vol. 39, Issue 4, pp. 261–283). https://doi.org/10.1007/s10462-011-9272-4

Kucheryavskiy, S. (2018). Analysis of NIR spectroscopic data using decision trees and their ensembles. *Journal of Analysis and Testing*, *2*(3), 274–289. https://doi.org/10.1007/s41664-018-0078-0

Kutsanedzie, F. Y. H., Chen, Q., Sun, H., & Cheng, W. (2017). In situ cocoa beans quality grading by near-infrared-chemodyes systems. *Analytical Methods*, *9*(37), 5455–5463. https://doi.org/10.1039/c7ay01751k

Lin, Y. K., Ho, Y. L., Zhao, Y., & Chang, Y. S. (2015). Quality assessment of Fritillariae Thunbergii Bulbus sold in Taiwan markets using a validated HPLC-UV method combined with hierarchical clustering analysis. *Journal of Food and Drug Analysis*, *23*(1), 130–135. https://doi.org/10.1016/j.jfda.2014.06.004

Liang, B., Hartel, R.W., Berglund, K.A. (1989). Effects of raffinose on sucrose crystal growth kinetics and rate dispersion. Aiche. J. 35 (12), 2053e2057.

Liu, Y., Sun, X., & Ouyang, A. (2010). Nondestructive measurement of soluble solid content of navel orange fruit by visible-NIR spectrometric technique with PLSR and PCA-BPNN. *LWT - Food Science and Technology*, *43*(4), 602–607. https://doi.org/10.1016/j.lwt.2009.10.008

Loh, W. Y. (2014). Fifty years of classification and regression trees. *International Statistical Review*, *82*(3), 329–348. https://doi.org/10.1111/insr.12016

Lu, Y., Thomas, L., & Schmidt, S. (2017). Differences in the thermal behavior of beet and cane sucrose sources. *Journal of Food Engineering*, *201*, 57–70. https://doi.org/10.1016/j.jfoodeng.2017.01.005

Magnus, I., Virte, M., Thienpont, H., & Smeesters, L. (2021). Combining optical spectroscopy and machine learning to improve food classification. *Food Control*, *130*. https://doi.org/10.1016/j.foodcont.2021.108342

Marsili RT, Miller N, Kilmer GJ, Simmons RE. (1994). Identification and quantitation of theprimary chemicals responsible for the characteristic malodor of beet sugar by purge-and-trapGC-MS-OD techniques. J Chromatogr Sci 32(5):165–71.

Martin, G. J., Danhol, D., & Vallet, C. (1991). Natural Isotope Fractionation in the Discrimination of Sugar Origins. In *J Sci Food Agric* (Vol. 56).

Massart, D. L. (1988). *Chemometrics: A textbook*. Elsevier.

Menevseoglu, A., Aykas, D. P., & Adal, E. (2021). Non-targeted approach to detect green pea and peanut adulteration in pistachio by using portable FT-IR, and UV–Vis spectroscopy. *Journal of Food Measurement and Characterization*, *15*(2), 1075–1082. https://doi.org/10.1007/s11694-020-00710-y

Monakhova, Y. B., & Diehl, B. W. K. (2016). Authentication of the origin of sucrose-based sugar products using quantitative natural abundance 13C NMR. *Journal of the Science of Food and Agriculture*, *96*(8), 2861–2866. https://doi.org/10.1002/jsfa.7456

Morel du Boil, P.G. (1992). Theanderose - a contributor to c-axis elongation in canesugar processing. Int. Sugar J. 94 (1120), 90−94.

Morel du Boil, P.G. (1996). Theanderose - a characteristic of cane sugar crystals. Proc.S. Afr. Sug. Technol. Ass. 70140−70144.

Morel du Boil, P.G. (1997). Theanderose - distinguishing cane and beet sugars. Int.Sugar J. 99 (1179), 102−106.

O'Leary, M. H. (1988). Carbon isotopes in photosynthesis. BioScience, 38(5), 328–336. https://doi.org/10.2307/1310735

Otto, M. (2017). *Chemometrics: statistics and computer application in analytical chemistry*. Wiley-VCH.

Palmer, K. F., & Williams, D. (1974). OPTICAL PROPERTIES OF WATER IN THE NEAR INFRARED. *J Opt Soc Am*, *64*(8), 1107–1110. https://doi.org/10.1364/JOSA.64.001107

PANKOBİRLİK. (2017). *Dünya, AB ve Türkiye Şeker İstatistikleri*.

Patras, A., Brunton, N. P., Downey, G., Rawson, A., Warriner, K., & Gernigon, G. (2011). Application of principal component and hierarchical cluster analysis to classify fruits and vegetables commonly consumed in Ireland based on in vitro antioxidant activity. *Journal of Food Composition and Analysis*, *24*(2), 250–256. https://doi.org/10.1016/j.jfca.2010.09.012

Perini, M., Nardin, T., Camin, F., Malacarne, M., & Larcher, R. (2018). Combination of sugar and stable isotopes analyses to detect the use of nongrape sugars in balsamic vinegar must. *Journal of Mass Spectrometry*, *53*(9), 772–780. https://doi.org/10.1002/jms.4211

Pouliarekou, E., Badeka, A., Tasioula-Margari, M., Kontakos, S., Longobardi, F., & Kontominas, M. G. (2011). Characterization and classification of Western Greek olive oils according to cultivar and geographical origin based on volatile compounds. *Journal of Chromatography A*, *1218*(42), 7534–7542. https://doi.org/10.1016/j.chroma.2011.07.081

Quelal Vásconez, M. A., Pérez-Esteve, É., Arnau-Bonachera, A., Barat, J. M., & Talens, P. (2018). Rapid fraud detection of cocoa powder with carob flour using near infrared spectroscopy. *Food Control*, 183–189.

R. A. Fisher, & W. A. Mackenzie. (1923). Studies in crop variation. II. The manurial response of different potato varieties. *Journal of Agricultural Science*, *13*, 311–320.

Ren, X., He, T., Wang, J., Wang, L., Wang, Y., Liu, X., Dong, Y., Ma, J., Jia, J., Song, R., Fan, Q., Wei, J., Yu, A., Wang, X., & She, G. (2021). UV spectroscopy and HPLC combined with chemometrics for rapid discrimination and quantification of Curcumae Rhizoma from three botanical origins. *Journal of Pharmaceutical and Biomedical Analysis*, *202*. https://doi.org/10.1016/j.jpba.2021.114145

Savitzky, A., & Golay, M. J. E. (1951). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. In *Z. Physiol. Chem* (Vol. 40, Issue 2). https://pubs.acs.org/sharingguidelines

SEFAOĞLU, F., KAYA, C., & KARAKUŞ, A. (2016). Farklı Tarihlerde Hasat Edilen Şeker Pancarı Genotiplerinin Verim ve Verim Unsurlarının Belirlenmesi. *Tarla Bitkileri Merkez Araştırma Enstitüsü Dergisi*, *25*(ÖZEL SAYI-2), 61–61. https://doi.org/10.21566/tarbitderg.281846

Shah, S.V., Chakradeo, Y.M. (1936). A note on the melting point of cane sugar. Curr.Sci. 4 (9), 652-653.

Su, X., Yan, X., & Tsai, C. L. (2012). Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(3), 275–294. https://doi.org/10.1002/wics.1198

Suhandy, D., & Yulia, M. (2017). Peaberry coffee discrimination using UV-visible spectroscopy combined with SIMCA and PLS-DA. *International Journal of Food Properties*, *20*, S331–S339. https://doi.org/10.1080/10942912.2017.1296861

Suhandy, D., & Yulia, M. (2021). The use of UV spectroscopy and SIMCA for the authentication of Indonesian honeys according to botanical, entomological and geographical origins. *Molecules*, *26*(4). https://doi.org/10.3390/molecules26040915

Teye, E., Uhomoibhi, J., & Wang, H. (2016). *Nondestructive Authentication of Cocoa Bean Cultivars by FT-NIR Spectroscopy and Multivariate Techniques*. *2*(3). https://doi.org/10.21859/focsci-020247

Thow, A. M., Lencucha, R. A., Rooney, K., Colagiuri, S., & Lenzen, M. (2021). Implications for farmers of measures to reduce sugars consumption. *Bulletin of the World Health Organization*, *99*(1), 41–49. https://doi.org/10.2471/BLT.19.249177

Torrecilla, J. S., Cámara, M., Fernández-Ruiz, V., Piera, G., & Caceres, J. O. (2008). Solving the spectroscopy interference effects of β-carotene and lycopene by neural networks. *Journal of Agricultural and Food Chemistry*, *56*(15), 6261–6266. https://doi.org/10.1021/jf8005239

Tosun, M. (2014). Detection of adulteration in mulberry pekmez samples added various sugar syrups with 13C/12C isotope ratio analysis method. *Food Chemistry*, *165*(2–3), 555–559. https://doi.org/10.1016/j.foodchem.2014.05.136

Türkiye Atom Enerjisi Kurumu. (2015). *Gıda İzlenebilirliğinde Kararlı İzotop Ölçümlerinin Uygulanması.*

Türkiye Şeker Fabrikaları. (2018). *Şeker Sektörü Raporu.*

Urbanus, B. L., Cox, G. O., Eklund, E. J., Ickes, C. M., Schmidt, S. J., & Lee, S. Y. (2014). Sensory Differences Between Beet and Cane Sugar Sources. *Journal of Food Science*, *79*(9), S1763–S1768. https://doi.org/10.1111/1750-3841.12558

Vaccari, G., Mantovani, G. (1995). Sucrose crystallization. In: Mathlouthi, M., Reiser, P.(Eds.), Sucrose Properties and Applications, first ed. Bishopbriggs: Blackie Ac-ademic and Professional, pp. 33−72

vanden Branden, K., & Hubert, M. (2005). Robust classification in high dimensions based on the SIMCA Method. *Chemometrics and Intelligent Laboratory Systems*, *79*(1–2), 10–21. https://doi.org/10.1016/j.chemolab.2005.03.002

Weldegergis, B. T., de Villiers, A., & Crouch, A. M. (2011). Chemometric investigation of the volatile content of young South African wines. *Food Chemistry*, *128*(4), 1100–1109. https://doi.org/10.1016/j.foodchem.2010.09.100

Wold, H. (1973). Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments. In *Multivariate Analysis–III* (pp. 383–407). Elsevier. https://doi.org/10.1016/b978-0-12-426653-7.50032-6

Wold, S., Esbensen, K., & Geladi, P. (2007). Principal Component Analysis (PCA) Principal Component Analysis (PCA). *Statistics*, *2*(June), 1–12.

Wold, S., & Sjöström, M. (1977). *SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy* (pp. 243–282). https://doi.org/10.1021/bk-1977-0052.ch012

Wold, S., Sjostrom, M., Eriksson, L., & Sweden°°, S. (2001). PLS-regression: a basic tool of chemometrics. In *Chemometrics and Intelligent Laboratory Systems* (Vol. 58). www.elsevier.comrlocaterchemometrics

Wold, S., Trygg, J., Berglund, A., & Antti, H. (2001). Some recent developments in PLS modeling. In *Chemometrics and Intelligent Laboratory Systems* (Vol. 58). www.elsevier.comrlocaterchemometrics

Wolever, T. M. S. (2006). Mechanisms by which Different Carbohydrates Elicit Different Glycaemic Responses. In The glycaemic index: A physiological classification of dietary carbohydrate. essay, CAB International.

Wu, W., Mallet, Y., Walczak, B., Penninckx, W., Massart, D. L., Heuerding, S., & Erni, F. (1996). Comparison of regularized discriminant analysis, linear discriminant analysis and quadratic discriminant analysis, applied to NIR data. *Analytica Chimica Acta*, *329*(3), 257–265. https://doi.org/10.1016/0003-2670(96)00142-0

Wu, W., & Massart, D. L. (1997). Regularised nearest neighbour classification method for pattern recognition of near infrared spectra. *Analytica Chimica Acta*, *349*(1–3), 253–261. https://doi.org/10.1016/S0003-2670(97)00285-7

Xiao-ying, G., Li, Q., Jin-jie, Z., Dan-ting, Y., Tang Chun-lan, Da-lun, X., Qiao-ming, L., Wen-ge, Y., & Qi-jie, H. (2019). Rapid and Quantitative Analysis of Histamine in Fish Using Surface Enhanced Raman Spectroscopy. *SPECTROSCOPY AND SPECTRAL ANALYSIS*, *39*(8), 2561–2567. https://doi.org/10.3964/j.issn.1000-0593(2019)08-2561-07

APPENDICES

A. Statistical Tables

Table 10. Classification measures of CART Analysis

	Sensitivity	Specificity	Precision	Error Rate	Accuracy
Training					
cane	1.00	0.99	0.98	0.01	0.99
beet	0.99	1.00	1.00	0.01	0.99
5-fold CV					
cane	0.98	0.97	0.96	0.02	0.98
beet	0.97	0.98	0.99	0.02	0.98

real/predicted	cane	beet	not assigned
cane	49	1	0
beet	2	72	0

Table 11. Classification measures of KNN

	Sensitivity	Specificity	Precision	Error Rate	Accuracy
Training					
cane	1.00	1.00	1.00	0	1
beet	1.00	1.00	1.00	0	1
Test					
cane	1	0.90	0.77	0.07	0.92
beet	0.90	1	1	0.07	0.92

real/predicted	cane	beet
cane	10	0
beet	3	26

Table 12. Regression results

Model	# of Variables	Factors	Calibration Set		Prediction Set		RPD
			Rc^2	RMSEC	Rp^2	RMSEP	
PLSR	Full	2	0.982	4.191	0.940	6.455	4.809
PLSR (SG + 1stDerivative)	Full	2	0.995	2.156	0.985	3.192	9.724
PLSR (SG + 1stDerivative)	100	2	0.992	2.834	0.993	2.225	13.941
scmwiPLSR	18	5	-	1.936	0.996	0.332	93.433
scmwiPLSR (SG + 1stDer)	72	8	-	0.0724	1	0.3085	100.714
Linear Regression	Full	-	0.998	1.567	0.980	3.784	8.197
Linear Regression (SG + 1stDer)	Full	-	0.991	2.956	0.987	3.026	10.251
Linear Regression (SG + 1stDer)	6	-	0.988	3.382	0.985	3.279	9.460

B. Matlab Codes

PCA Matlab Code (Background)

```
function [coeff, score, latent, tsquared, explained, mu] =
pca(x,varargin)

%PCA Principal Component Analysis (PCA) on raw data.

%   COEFF = PCA(X) returns the principal component coefficients for
the N

%   by P data matrix X. Rows of X correspond to observations and
columns to

%   variables. Each column of COEFF contains coefficients for one
principal

%   component. The columns are in descending order in terms of
component

%   variance (LATENT). PCA, by default, centers the data and uses
the

%   singular value decomposition algorithm. For the non-default
options,

%   use the name/value pair arguments.

%

%   [COEFF, SCORE] = PCA(X) returns the principal component score,
which is

%   the representation of X in the principal component space. Rows
of SCORE

%   correspond to observations, columns to components. The centered
data

%   can be reconstructed by SCORE*COEFF'.

%

%   [COEFF, SCORE, LATENT] = PCA(X) returns the principal component

%   variances, i.e., the eigenvalues of the covariance matrix of X,
in

%   LATENT.

%

%   [COEFF, SCORE, LATENT, TSQUARED] = PCA(X) returns Hotelling's
T-squared
```

```
%    statistic for each observation in X. PCA uses all principal
components
%   to compute the TSQUARED (computes in the full space) even when
fewer
%   components are requested (see the 'NumComponents' option below).
For
%   TSQUARED in the reduced space, use MAHAL(SCORE,SCORE).
%
%   [COEFF, SCORE, LATENT, TSQUARED, EXPLAINED] = PCA(X) returns a
vector
%    containing the percentage of the total variance explained by
each
%   principal component.
%
%   [COEFF, SCORE, LATENT, TSQUARED, EXPLAINED, MU] = PCA(X) returns
the
%    estimated mean, MU, when 'Centered' is set to true; and all
zeros when
%   set to false.
%
%    [...] = PCA(..., 'PARAM1',val1, 'PARAM2',val2, ...) specifies
optional
%     parameter name/value pairs to control the computation and
handling of
%   special data types. Parameters are:
%
%    'Algorithm' - Algorithm that PCA uses to perform the principal
%                  component analysis. Choices are:
%        'svd'   - Singular Value Decomposition of X (the default).
%       'eig'   - Eigenvalue Decomposition of the covariance matrix.
It
%                  is faster than SVD when N is greater than P, but
less
%                     accurate because the condition number of the
covariance
%                  is the square of the condition number of X.
%         'als'    - Alternating Least Squares (ALS) algorithm which
finds
%                    the best rank-K approximation by factoring a X
into a
%                    N-by-K left factor matrix and a P-by-K right
factor
%                    matrix, where K is the number of principal
components.
%                    The factorization uses an iterative method
starting with
%                    random initial values. ALS algorithm is designed
to
%                    better handle missing values. It deals with
missing
%                    values without listwise deletion (see {'Rows',
%                    'complete'}).
%
%     'Centered' - Indicator for centering the columns of X. Choices
are:
%           true    - The default. PCA centers X by subtracting off
column
%                    means before computing SVD or EIG. If X contains
NaN
%                    missing values, NANMEAN is used to find the mean
with
%                    any data available.
%          false   - PCA does not center the data. In this case, the
original
%                    data X can be reconstructed by X = SCORE*COEFF'.
%
%      'Economy'   - Indicator for economy size output, when D the
degrees of
%                    freedom is smaller than P. D, is equal to M-1,
if data
%                    is centered and M otherwise. M is the number of
rows
%                    without any NaNs if you use 'Rows', 'complete';
or the
```

```
%                    number of rows without any NaNs in the column
pair that
%                    has the maximum number of rows without NaNs if
you use
%                    'Rows', 'pairwise'. When D < P, SCORE(:,D+1:P)
and
%                      LATENT(D+1:P) are necessarily zero, and the
columns of
%                        COEFF(:,D+1:P) define directions that are
orthogonal to
%                    X. Choices are:
%          true   - This is the default. PCA returns only the first
D
%                    elements of LATENT and the corresponding columns
of
%                    COEFF and SCORE. This can be significantly faster
when P
%                    is much larger than D. NOTE: PCA always returns
economy
%                    size outputs if 'als' algorithm is specifed.
%          false  - PCA returns all elements of LATENT. Columns of
COEFF and
%                    SCORE corresponding to zero elements in LATENT
are
%                    zeros.
%
%      'NumComponents' - The number of components desired, specified
as a
%                    scalar integer K satisfying 0 < K <= P. When
specified,
%                    PCA returns the first K columns of COEFF and
SCORE.
%
%      'Rows'      - Action to take when the data matrix X contains
NaN
%                    values. If 'Algorithm' option is set to 'als,
this
%                    option is ignored as ALS algorithm deals with
missing
```

```
%                  values without removing them. Choices are:
%          'complete' - The default action. Observations with NaN
values
%                      are removed before calculation. Rows of NaNs
are
%                     inserted back into SCORE at the corresponding
%                      location.
%          'pairwise' - If specified, PCA switches 'Algorithm' to
'eig'.
%                       This option only applies when 'eig' method
is used.
%                       The (I,J) element of the covariance matrix
is
%                        computed using rows with no NaN values in
columns I
%                        or J of X. Please note that the resulting
covariance
%                      matrix may not be positive definite. In that
case,
%                      PCA terminates with an error message.
%         'all'      - X is expected to have no missing values. All
data
%                       are used, and execution will be terminated
if NaN is
%                      found.
%
%       'Weights'   - Observation weights, a vector of length N
containing all
%                  positive elements.
%
%     'VariableWeights' - Variable weights. Choices are:
%          - a vector of length P containing all positive elements.
%            - the string 'variance'. The variable weights are the
inverse of
%             sample variance. If 'Centered' is set true at the same
time,
%            the data matrix X is centered and standardized. In this
case,
%               PCA returns the principal components based on the
correlation
%            matrix.
%
%    The following parameter name/value pairs specify additional
options
%   when alternating least squares ('als') algorithm is used.
%
%       'Coeff0'   - Initial value for COEFF, a P-by-K matrix. The
default is
%                  a random matrix.
%
%       'Score0'   - Initial value for SCORE, a N-by-K matrix. The
default is
%                  a matrix of random values.
%
%       'Options' - An options structure as created by the STATSET
function.
%                  PCA uses the following fields:
%          'Display' - Level of display output.  Choices are 'off'
(the
%                      default), 'final', and 'iter'.
%          'MaxIter' - Maximum number of steps allowed. The default
is
%                           1000. Unlike in optimization settings,
reaching
%                      MaxIter is regarded as convergence.
%              'TolFun' - Positive number giving the termination
tolerance for
%                      the cost function.  The default is 1e-6.
%                'TolX' - Positive number giving the convergence
threshold
%                      for relative change in the elements of L and
R. The
%                      default is 1e-6.
%
```

```
%

%   Example:

%       load hald;

%               [coeff,  score,  latent,  tsquared,  explained]  =
pca(ingredients);

%

%    See also PPCA, PCACOV, PCARES, BIPLOT, BARTTEST, CANONCORR,
FACTORAN,

%   ROTATEFACTORS.


% References:

%       [1]  Jolliffe,  I.T.  Principal  Component  Analysis,  2nd
ed.,Springer,2002.

%   [2] Krzanowski, W.J., Principles of Multivariate Analysis, Oxford

%       University Press, 1988.

%    [3] Seber, G.A.F., Multivariate Observations, Wiley, 1984.

%   [4] Jackson, J.E., A User's Guide to Principal Components, Wiley,
1988.

%    [5] Sam Roweis, EM algorithms for PCA and SPCA, In Proceedings
of the

%       1997 conference on Advances in neural information processing

%        systems 10 (NIPS '97), MIT Press, Cambridge, MA, USA, 626-
632,1998.

%    [6] Alexander Ilin and Tapani Raiko. Practical Approaches to
Principal

%         Component Analysis in the Presence of Missing Values. J.
Mach.

%       Learn. Res. 11 (August 2010), 1957-2000, 2010.


%   Copyright 2012-2020 The MathWorks, Inc.




if nargin > 1

    [varargin{:}] = convertStringsToChars(varargin{:});
```

```
end


[n, p] = size(x);

internal.stats.checkSupportedNumeric('X',x,false,false,true,true);
% complex is accepted here


% Parse arguments and check if parameter/value pairs are valid

paramNames                                                          =
{'Algorithm','Centered','Economy','NumComponents','Rows',...

    'Weights','VariableWeights','Coeff0','Score0','Options'};

defaults   = {'svd',      true,     true,    p,         'complete',...

    ones(1,n,'like',x)  ,ones(1,p,'like',x),          [],        [],
statset('pca')};


[vAlgorithm,       vCentered,       vEconomy,       vNumComponents,
vRows,vWeights,...

    vVariableWeights, c0, s0, opts, setFlag]...

    = internal.stats.parseArgs(paramNames, defaults, varargin{:});

% Validate String value for  Algorithm value

AlgorithmNames = {'svd','eig','als'};

vAlgorithm                                                          =
internal.stats.getParamVal(vAlgorithm,AlgorithmNames,...

    '''Algorithm''');

% Validate boolean value for 'Centered' option

vCentered = internal.stats.parseOnOff(vCentered,'''Centered''');

% Validate boolean value for 'Economy' option

vEconomy = internal.stats.parseOnOff(vEconomy,'''Economy''');

% Validate the number of components option 'NumComponents'

if ~isempty(x) && ~internal.stats.isScalarInt(vNumComponents,1,p)

    error(message('stats:pca:WrongNumComponents',p));

end

% Validate value for 'Rows' option

RowsNames = {'complete','pairwise','all'};

vRows = internal.stats.getParamVal(vRows,RowsNames,'''Rows''');
```

```
switch vAlgorithm
    case 'svd'
        % Switch 'Algorithm' to 'eig' if 'Rows' set to 'pairwise'
        if strcmp(vRows,'pairwise')
            if setFlag.Algorithm
                warning(message('stats:pca:NoPairwiseSVD'));
            end
            vAlgorithm = 'eig';
        end
        %  Switch  algorithm  to  'als'  if  user  specify  'Coeff0'  and
'Score0'.
        if setFlag.Coeff0 || setFlag.Score0
            vAlgorithm = 'als';
        end
    case 'als'
        % If 'als' is chosen, force PCA to use ALS and to ignore the
        % Rows' option
        if setFlag.Rows
            warning(message('stats:pca:NoALSRows'));
        end
end

% Validate Weights Vectors
if isvector(vWeights) && isequal(numel(vWeights),n)
    vWeights = reshape(vWeights,1,n); % make sure it is a row vector
else
    error(message('stats:pca:WrongObsWeights', n));
end

if internal.stats.isString(vVariableWeights)
    WeightsNames = {'variance'};
    internal.stats.getParamVal(vVariableWeights,WeightsNames,...
        '''VariableWeights''');
    vVariableWeights                                                 =
1./classreg.learning.internal.wnanvar(x,vWeights,1);
elseif isnumeric(vVariableWeights) && isvector(vVariableWeights)...
        && (isequal(numel(vVariableWeights), p))
    vVariableWeights = reshape(vVariableWeights,1,p);
else
    error(message('stats:pca:WrongVarWeights', p));
end

if any(vWeights <= 0) || any(vVariableWeights <= 0)
    error(message('stats:pca:NonPositiveWeights'));
end
% end of checking input arguments


% Handling special empty case
if isempty(x)
    pOrZero = ~vEconomy * p;
    coeff = zeros(p, pOrZero, "like", x);
    coeff(1:p+1:end) = 1;
    score = zeros(n, pOrZero, "like", x);
    latent = zeros(pOrZero, 1, "like", x);
    tsquared = zeros(n, 1, "like", x);
    explained = zeros(0, "like", x);
    mu = zeros(0, "like", x);
    return;
end

nanIdx = isnan(x);
numNaN = sum(nanIdx, 2); % number of NaNs in each row
wasNaN = any(numNaN,2); % Rows that contain NaN
```

```
% Handling special cases where X is all NaNs:
if all(nanIdx(:))
    coeff = NaN("like",x);
    score = NaN("like",x);
    latent = NaN("like",x);
    tsquared = NaN("like",x);
    explained = NaN("like",x);
    mu = NaN("like",x);
    return;
end
% Handling special scalar case;
if isscalar(x)
    coeff = ones("like",x);
    notCentered = cast(~vCentered,"like",x);
    score = notCentered*x;
    latent = notCentered*x^2;
    tsquared = notCentered;
    explained = 100*coeff;
    mu = vCentered*x;
    return;
end

if strcmp(vRows,'all') && (~strcmp(vAlgorithm,'als'))
    if any(wasNaN)
        error(message('stats:pca:RowsAll'));
    else
        vRows = 'complete';
    end
end

if strcmp(vRows,'complete')
```

```
    % Degrees of freedom (DOF) is n-1 if centered and n if not
centered,
    % where n is the numer of rows without any NaN element.
    DOF = max(0,n-vCentered-sum(wasNaN));
elseif strcmp(vRows,'pairwise')
    % DOF is the maximum number of element pairs without NaNs
    notNaN = double(~nanIdx);
    nanC = notNaN'*notNaN;
    nanC = nanC.*(~eye(p));
    DOF = max(nanC(:));
    DOF = DOF-vCentered;
else
    DOF = max(0,n-vCentered);
end

if vCentered
    % Weighted sample mean:
    mu = classreg.learning.internal.wnanmean(x, vWeights);
else
    mu = zeros(1,p,'like',x);
end

% Compute by EIG if no weights are given
switch vAlgorithm
    case 'eig'
        % Center the data if 'Centered' is true.
        if vCentered
            x = x - mu;
        end

        % Use EIG to compute.
        [coeff,   eigValues]   =   localEIG(x,   vCentered,   vRows,
vWeights,...
```

```
            vVariableWeights);

        % When 'Economy' value is true, nothing corresponding to
zero
        % eigenvalues should be returned.
        if (DOF<p)
            if vEconomy
                coeff(:, DOF+1:p) = [];
                eigValues(DOF+1:p, :) = [];
            else % make sure they are zeros.
                eigValues(DOF+1:p, :) = 0;
            end
        end

        % Force small negative eigenvalues to zero because of
rounding error

eigValues((eigValues<0)&(abs(eigValues)<(eps(eigValues(1))*length(
eigValues)))) = 0;

        % Check if eigvalues are all positive
        if any(eigValues<0)

error(message('stats:pca:CovNotPositiveSemiDefinite'));
        end

        if nargout > 1
            score = x/coeff';
            latent = eigValues; % Output Eigenvalues
            if nargout > 3
                tsquared = localTSquared(score, latent, n, p);
            end
        end
```

```
    case 'svd' % Use SVD to compute
        % Center the data if 'Centered' is true.
        if vCentered
            x = x - mu;
        end

        [U,sigma, coeff, wasNaN] = localSVD(x, n,...
            vEconomy, vWeights, vVariableWeights);
        if nargout > 1
            score =  U.*(sigma');
            latent = sigma.^2./DOF;
            if nargout > 3
                tsquared = localTSquared(score,latent,DOF,p);
            end
            %Insert NaNs back
            if any(wasNaN)
                score = internal.stats.insertnan(wasNaN, score);
                if nargout >3
                    tsquared                                      =
internal.stats.insertnan(wasNaN,tsquared);
                end
            end
        end

        if DOF < p
            % When 'Economy' value is true, nothing corresponding
to zero
            % eigenvalues should be returned.
            if vEconomy
                coeff(:, DOF+1:end) = [];
                if nargout > 1
                    score(:, DOF+1:end)=[];
                    latent(DOF+1:end, :)=[];
                end
            elseif nargout > 1
                % otherwise, eigenvalues and corresponding outputs
need to pad
                % zeros because svd(x,0) does not return columns of
U corresponding
                % to components of (DOF+1):p.
                score(:, DOF+1:p) = 0;
                latent(DOF+1:p, 1) = 0;
            end
        end

    case 'als' % Alternating Least Square Algorithm

        vNumComponents = min([vNumComponents,n-vCentered,p]);  % ALS
always return economy sized outputs

        if isempty(s0)
            s0 = randn(n,vNumComponents,'like',x);
        elseif              ~isequal(size(s0),[n,vNumComponents])||
any(isnan(s0(:)))

error(message('stats:pca:BadInitialValues','Score0',n,vNumComponen
ts));
        end
        if isempty(c0)
            c0 = randn(p,vNumComponents,'like',x);
        elseif       ~isequal(size(c0),[p,vNumComponents])       ||
any(isnan(c0(:)))

error(message('stats:pca:BadInitialValues','Coeff0',p,vNumComponen
ts));
        end


[score,coeff,mu,latent]=alsmf(x,vNumComponents,'L0',s0,'R0',c0,...
'Weights',vWeights,'VariableWeights',vVariableWeights,...

'Orthonormal',true,'Centered',vCentered,'Options',opts);


        if nargout > 3
            % T-squared values are in reduced space.
            tsquared = localTSquared(score, latent,n-vCentered,
vNumComponents);
        end
end % end of switch vAlgorithm


% Calcuate the percentage of the total variance explained by each
principal
% component.
if nargout > 4
    explained = 100*latent/sum(latent);
end


% Output only the first k principal components
if (vNumComponents<DOF)
    coeff(:, vNumComponents+1:end) = [];
    if nargout > 1
        score(:, vNumComponents+1:end) = [];
    end
end



% Enforce a sign convention on the coefficients -- the largest
element in
% each column will have a positive sign.
[~,maxind] = max(abs(coeff), [], 1);
[d1, d2] = size(coeff);
colsign = sign(coeff(maxind + (0:d1:(d2-1)*d1)));
```

```
coeff = coeff.*colsign;
if nargout > 1
    score = score.*colsign; % scores = score
end

end % End of main function



%----------------Subfucntions-------------------------------------
-------


function          [coeff,          eigValues]=localEIG(x,vCentered,
vRows,vWeights,...
    vVariableWeights)
% Compute by EIG. vRows are the options of handing NaN when compute
% covariance matrix

% Apply observation and variable weights
OmegaSqrt = sqrt(vWeights);
PhiSqrt = sqrt(vVariableWeights);
x = x.*(OmegaSqrt');
x = x.*PhiSqrt;

xCov = ncnancov(x, vRows, vCentered);

[coeff, eigValueDiag] = eig(xCov);
[eigValues, idx] = sort(diag(eigValueDiag), 'descend');
coeff = coeff(:, idx);

coeff = coeff./(PhiSqrt');
end
```

```
function [U,sigma, coeff, wasNaN] = localSVD(x, n,...,
    vEconomy, vWeights, vVariableWeights)
%  Compute  by  SVD.  Weights  are  supplied  by  vWeights  and
vVariableWeights.


% Remove NaNs missing data and record location
[~,wasNaN,x] = internal.stats.removenan(x);
if n==1  % special case because internal.stats.removenan treats all
vectors as columns
    wasNaN = wasNaN';
    x = x';
end


% Apply observation and variable weights
vWeights(wasNaN) = [];
OmegaSqrt = sqrt(vWeights);
PhiSqrt = sqrt(vVariableWeights);
x = x.*(OmegaSqrt');
x = x.*PhiSqrt;


if vEconomy
    [U,sigma,coeff] = svd(x,'econ');
else
    [U,sigma, coeff] = svd(x, 0);
end


U = U./(OmegaSqrt');
coeff = coeff./(PhiSqrt');


if n == 1     % sigma might have only 1 row
    sigma = sigma(1);
else
    sigma = diag(sigma);
```

```
end
end

function tsquared = localTSquared(score, latent,DOF, p)
% Subfunction to calulate the Hotelling's T-squared statistic. It
is the
% sum of squares of the standardized scores, i.e., Mahalanobis
distances.
% When X appears to have column rank < r, ignore components that are
% orthogonal to the data.

if isempty(score)
    tsquared = score;
    return;
end

r = min(DOF,p); % Max possible rank of x;
if DOF > 1
    q = sum(latent > max(DOF,p)*eps(latent(1)));
    if q < r
        warning(message('stats:pca:ColRankDefX', q));
    end
else
    q = 0;
end
standScores = score(:,1:q)./(sqrt(latent(1:q,:))');
tsquared = sum(standScores.^2, 2);
end

function c = ncnancov(x,Rows,centered)
%    C = NCNANCOV(X) returns X'*X/N, where N is the number of
observations
%   after removing rows missing values.
```

```
%
%   C = NCNANCOV(...,'pairwise') computes C(I,J) using rows with no
NaN
%    values in columns I or J.  The result may not be a positive
definite
%    matrix. C = NCNANCOV(...,'complete') is the default, and it
omits rows
%   with any NaN values, even if they are not in column I or J.
%
%    C = NCNANCOV(...,true), C is normalized by N-1 if data X is
already
%   centered. The default is false.


if nargin <2
    Rows = 'complete';
end


d = 0;
if nargin>2
    d =  d + centered;
end


idxnan = isnan(x);


[n, p] = size(x);



if ~any(any(idxnan))
    c = x'*x/(n-d);
elseif strcmp(Rows,'complete')
    nanrows = any(idxnan,2);
    xNotNaN = x((~nanrows),:);
    denom = max(0, (size(xNotNaN,1)-d) );
    c = xNotNaN'*xNotNaN/denom;
```

```
elseif strcmp(Rows,'pairwise')
    c = zeros(p,class(x));
    for i = 1:p
        for j = 1:i
            NonNaNRows = ~any(idxnan(:,[i, j]),2);
            denom = max(0,(sum(NonNaNRows)-d));
            c(i,j) = x(NonNaNRows,i)'*x(NonNaNRows,j)/denom;
        end
    end
    c = c + tril(c,-1)';
end
end
```

PCA Application

```
clc;
clear all;
load BinMixData.mat
A=DataBinMix % all spectrum
P=ConcBinMix;%samples
B=A;%without first column - wavelength
% B=B'; %A has every example as column. In PCA A must have example
as row
[B_r,B_c]=size(B);%dimensions of A, number of rows and number of
columns
% Bc=B-ones(B_r,1)*mean(B);%normalization
% Aa=Ac./(ones(A_r,1)*sum((Ac.^2)/A_r).^(1/2));%auto-scaling
[coeff,score,latent,tsquared,explained,mu]=pca(B,'NumComponents',2
);
Dis = cumsum(explained);
figure(1);
```

```
plot(Dis(1:10,1),'ro-','LineWidth',2,'MarkerSize',5);
xlabel('Number of component');
ylabel('Explained, %');
%%

figure(1);
plot(WavelenghtBinMix,B,'r');
%%
figure(2);
plot(score(1:end,1),score(1:end,2),'rs','LineWidth',2,'MarkerSize'
,10)
xlabel('First principal component');
ylabel('Second principal component');

%%
figure(3);
plot(P,score(1:end,1),'r*','LineWidth',2,'MarkerSize',10);
xlabel(' Concentration, % ');
ylabel('Score in first principal component');

%%
figure(4);

plot(WavelenghtBinMix,coeff(:,1),'r','linewidth',2);

xlabel('Wavelenght (nm)');
ylabel('Loadings');
```

PLS Matlab Code

```
clear all;
```

```
clc;
%PCA
load BinMix6Test.mat;

A=DataBinMix;
M=ConcBinMix;
%B=A(377:382,2:end);
[B_r,B_c]=size(A);
Bc=A-ones(B_r,1)*mean(A);
[coeff,score,latent,tsquared,explained,mu]=pca(A,'NumComponents',5
);
figure();
plot(M,score(1:end,1),'r*','LineWidth',3,'MarkerSize',5)
ylabel('Score to PC1');
xlabel('concentration');
S=score(1:end,1);

figure();
Dis = cumsum(explained);
plot(Dis,'b*','LineWidth',3,'MarkerSize',5);
axis([1 10 98.5 100]);
ylabel('Total variance explained, %');
xlabel('Number of principal components');

%%

%PLS
X=TrainDataBinMix(:,1:401);%train
Y=TrainDataBinMix(:,402);%train
%X=X';
[n,p]=size(X);
%X=X-ones(n,1)*mean(X);
```

```
[u,v]=size(Y);
X1=TestDataBinMix(:,1:401);%test
Y1=TestDataBinMix(:,402);%test
%X1=X1';
[e,k]=size(X1);
%X1=X1-ones(e,1)*mean(X1);
[z,s]=size(Y1);

for i=1:10
[XL,YL,XS,YS,betaPLS,PLSPctVar,PLSmsep]=plsregress(X,Y,i);
yfitPLS = [ones(n,1) X]*betaPLS;
yfitPLS1 = [ones(e,1) X1]*betaPLS;
RMSEPLS(i)=sqrt(sum((Y-yfitPLS).^2)/size(Y,1));
RMSEPLS1(i)=sqrt(sum((Y1-yfitPLS1).^2)/size(Y1,1));
end;
figure();
plot(1:10,RMSEPLS1,'b+','LineWidth',5,'MarkerSize',2);
ylabel('RMSEPLS test');
xlabel('Number of latent structures');

%%
figure();
[XL,YL,XS,YS,betaPLS,PLSPctVar,PLSmsep]=plsregress(X,Y,8);
yfitPLS = [ones(n,1) X]*betaPLS;
yfitPLS1 = [ones(e,1) X1]*betaPLS;



q=(yfitPLS-Y)./Y;
w=(yfitPLS1-Y1)./Y1;
plot(Y,q,'r^',Y1,w,'b*');
legend({'PLS train' 'PLS test'},'location','SE');
xlabel('Measured mass, g');
```

```
ylabel('Difference between the predicted and measured mass');

%%

plot(Y,yfitPLS,'b^',Y1,yfitPLS1,'r*');
xlabel('Measured mass, g');
ylabel('Predicted mass, g');
legend({'PLS train' 'PLS test' },'location','NW');
```

LDA Code

```
%clear all;clc;

%load SugarClassLDA  %two matrices output one for variables, one for
species (vertical allignment)


PC1 = PCscoresofBeetCane(:,1);
PC2 = PCscoresofBeetCane(:,2);

h1 = gscatter(PC1,PC2,AllType,'krb','ov^',[],'off');
 h1(1).LineWidth = 2;
 h1(2).LineWidth = 2;
 %h1(3).LineWidth = 2;
 legend('Cane Sugar','Beet Sugar') %Change according to your data
 hold on
 X = [PC1,PC2];
 MdlLinear = fitcdiscr(X,AllType);
 MdlLinear.ClassNames([1 2])
 K = MdlLinear.Coeffs(1,2).Const;
 L = MdlLinear.Coeffs(1,2).Linear;
```

```
f = @(x1,x2) K + L(1)*x1 + L(2)*x2;

h2 = fimplicit(f,[-3.5 6 -1.5 1.5]); %Size of your graph, x and y axis, lenght and where they start

h2.Color = 'b';

h2.LineWidth = 2;

h2.DisplayName = 'Boundary between Beet Sugar & Cane Sugar';
```