DETERMINATION OF HYDROCARBON COMPOSITION OF NAPHTHA BY USING FOURIER TRANSFORM INFRARED SPECTROSCOPY AND MULTIVARIATE CALIBRATION

A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in Chemistry

by
Selahattin ŞENTÜRK

December 2020
İZMİR

ACKNOWLEDGMENTS

I would like to express my sincere gratefulness to a group of people who made this thesis possible. This thesis could not be written without;

My supervisor, Prof. Dr. Durmuş ÖZDEMİR, mentored me with his great expertise. His great efforts and patience made this thesis possible to be created. His explanations to all of my questions clearly and simply made me encouraged to keep up with my hard work on a long way towards my career. Additionally, I would like to thank Prof. Dr. Durmuş Özdemir's research group for their efforts.

I would like to thank Prof. Dr. Şerife YALÇIN and Assoc.Dr. Hasan ERTAŞ for their support and insightful comments.

Finally, to my great family and friends for always being with me.

ABSTRACT

DETERMINATION OF HYDROCARBON COMPOSITION OF NAPHTHA BY USING FOURIER TRANSFORM INFRARED SPECTROSCOPY AND MULTIVARIATE CALIBRATION

Accurate monitoring of the charging and output of the refinery unit is required. These direct refineries need to provide a quick response to posts on crude oil compositions or directions to their latest request. Determining the physical properties of the intermediate products of the crude oil unit in the refinery based on conventional analytical methods requires time consuming and expensive processes. At this stage, multivariate calibration techniques, creating models that can replace conventional analysis methods and obtaining results using fast spectroscopic analysis. For this study, multivariate calibration techniques were used to determine the hydrocarbons in the naphtha product from crude oil distillation column. The results were evaluated by comparing with using the reference conventional method results. Parameters are Aromatics, Olefins, Benzene, Naphthenes, Paraffins, C7Plus (the sum of compounds with more than 7 carbons) and C6Minus (the sum of compounds with less than 6 carbons). Samples were analyzed by Fourier transform near infrared region spectroscopy between 10000 cm^{-1}- 4000 cm^{-1} wavenumbers. Calibration models were obtained by partial least squares and genetic inverse least squares methods. Using these models, the relevant parameters for the validation set samples were estimated and compared statistically with the values of the reference analysis methods. The results has been indicated that parameters has been successfully modelled with R^2 range from 0.917 to 0.998 for LSRN samples and R^2 range from 0.963 to 0.996 for HSRN samples.

ÖZET

NAFTANIN HİDROKARBON BİLEŞİMİNİN FOURIER DÖNÜŞÜMLÜ KIZILÖTESİ SPEKTROSKOPİSİ VE ÇOK DEĞİŞKENLİ KALİBRASYON İLE BELİRLENMESİ

Birçok farklı fiziksel ve kimyasal süreçler oluşan rafinerilerde yer almaktadır. Bu süreçler birbirleriyle ilişki içinde olan bir birim için son ürün olurken diğer birimin hammaddesi olmaktadır. Bu nedenle belirli spesifikasyonlara göre üretim yapılmaktadır. Rafineri planlamasında ve çeşitli süreçlerin karmaşıklığında gerekli esnekliği, ancak her rafineri biriminin değişimini ve son ürününün akışını sıkı bir şekilde gözlemleyerek sağlanabilir. Ham petrol damıtma ünitesi süreç koşullarının optimizasyonu her rafineride önemli bir parametredir. Analizdeki gecikmeler, süreç koşullarının ayarlanmasında gecikmelere neden olur. Rafinerideki ham petrol ünitesinin ara ürünleri için konvansiyonel analitik yöntemlere dayalı fiziksel özelliklerinin belirlenmesi zaman alıcı ve pahalı işlemler gerektirir. Rutin olarak izlenen özellikler arasından biri olan nafta ürünün içindeki hidrokarbon bileşimleri yer alır. Ham petrol ünitesinden üst katmanlarından alınan nafta olarak adlandırılan temel hidrokarbon karışımı olan ve içinde parafin, aromatik ve olefin gibi hidrokarbon moleküller bulunmaktadır. Nafta genel olarak 4 zincirli karbon ile başlayarak 10 zincirli karbon aralığındaki hidrokarbonlardan oluşmaktadır. Bu çalışmada 95 hafif naftanın ve 67 ağır nafta kullanılmıştır. Yakın kızıl ötesi spektroskopisi ile 10000 cm^{-1}- 4000 cm^{-1} dalga sayısı aralığında örneklerin spektrumları toplanmış ardından referans olan konvansiyonel yöntemlerden elde edilen parametrelerin konsantrasyonlarıyla çok değişkenli kalibrasyon modelleri oluşturulmuştur. Elde edilen sonuçları karıştırılarak bu modeller ile konvansiyonel yöntemin yerine kullanılabilecek daha hızlı, ucuz ve güvenilir alternatif yöntem geliştirme amaçlanmıştır. Kısmı en küçük kareler(PLS) ile genetik ters en küçük kareler(GILS) 2 farklı çok değişkenli kalibrasyon metodu kullanılmıştır. İlk aşamada örneklerin spektrumlara ön işleme tekniği olan Genişletilmiş Çarpımsal Saçılma Düzeltmesi(EMSC) uygulanmıştır. Sonuç olarak hem hafif nafta hem de ağır nafta için iki farklı modelleme tekniğiyle başarılı sonuçlar elde edilmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

Figure **Page**

LIST OF TABLES

Table **Page**

CHAPTER 1

INTRODUCTION

1.1. Petroleum Refineries

Crude oil or petroleum is a fossil fuel. It is the transformation of organic substances into hydrocarbons because of various chemical reactions after millions of years. Crude oil, usually buried in the form of a deposit, sometimes appears close on earth, so humankind has known this since ancient times. In the world, oil has been used in many fields such as construction and medical purposes. However, with the industrial revolution that took place in Europe and America in the middle of the 19th century, the use of petroleum began to increase very rapidly as it started to replace coal to meet its energy needs. The United States rushes into oil and the country becomes the world's largest oil producer. Originally, it was the light source using only distilled oil in lamps. Then it started to form a significant majority of the transportation, as it offers a better calorific value than coal. Today, crude oil appears to be an important and fundamental function in human life. Crude oil is present in every part of our lives such as transportation, energy production, heating and chemical factories. As seen in Figure 1.1, the distribution of petroleum products used in 2019 and the estimated use in 2045. Therefore, according to this Figure 1.1, crude oil will continue to be one of the most important energy sources of our lives in the coming years.

Figure 1.1. Daily usage of crude oil[1]

Petroleum oil is a mixture of different solid, liquid, and gaseous hydrocarbons. In addition, hydrocarbons, nitrogen, oxygen, sulfur, and a trace amount of some metals in Petroleum oil. This content can change according to the region and its formation conditions. Therefore, petroleum oil can find various combinations of these hydrocarbons and other materials. The lengths of these hydrocarbons reach from 1 to 60 carbon atom chain lengths and have different boiling points.

Refineries are more complex than other chemical industries. They have many different physical and chemical processes. Some processes like cracking isomerization, hydrogenation, desulfurization, aromatization, and blending that are linked to each other and can affect each other's charge. Refinery profitability is affected quickly and serious losses by unit shut down or out of control processes. Therefore, certain specifications can be very important for any process. Refineries often process different region's crude oil. Crude oil prices can be affected by almost every situation like capability limits of crude oil tanks, political instabilities of crude oil-exporting countries, regulations in product specifications. Crude oil blending is one of the most important actions of refineries to increase profit margin. Variations in the crude oil composition affect the planned production capacities in order to meet the final product quantities. Most refineries are designed to process crude oil in a certain specification. Generally, general properties and properties can be determined by the American Petroleum Institute gravity or API gravity shown in equation 1.1.

$$°API = \left(\frac{141}{specific\ gravity_{60/60°F}}\right) - 131.5 \qquad (1.1)$$

API gravity help to understand the weight of crude oil with comparison to water. If crude oil has API gravity more than 10 it refers to light crude oil which means have light hydrocarbons like a paraffinic hydrocarbon. Crude oil has API gravity less than 10, which means that longer hydrocarbon chains like asphalt molecules.

After the crude oil coming to refineries, it enters the fractional distillation column under atmospheric pressure. Products with different carbon numbers in the oil in this column begin to separate according to their boiling points as shown Figure 1.2.

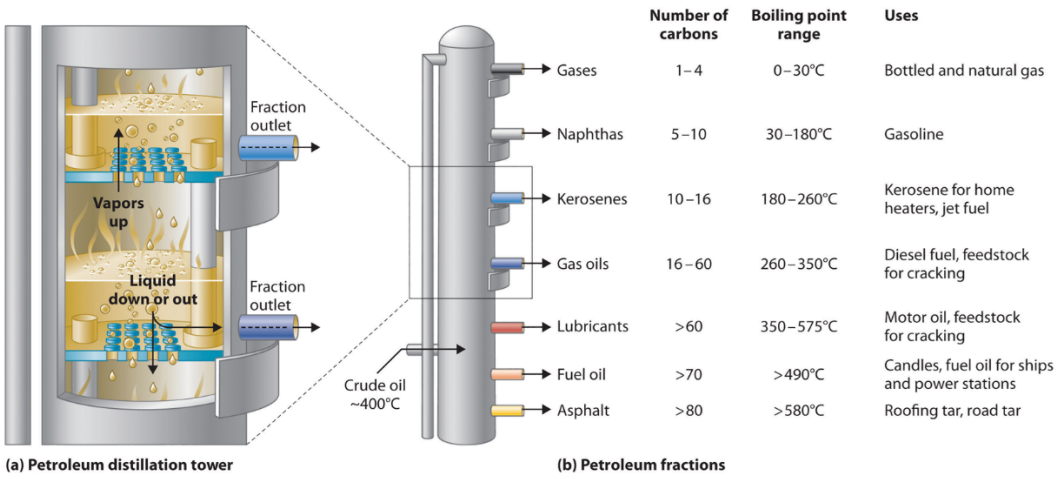


Figure 1.2. Atmospheric distillation column (Url-1)

As can be seen in Figure 1.2, Light products such as LPG and Naphtha are drawn from the top of the column. Products such as heavy kerosene, light diesel, and heavy diesel are gradually drawn through bottom of the column. The products drawn from here are generally referred to as intermediate products separated according to certain boiling points. These products are transformed into products such as gasoline, diesel fuel, jet fuel, and asphalt by passing through various processes. Light straight run naphtha (LSRN) consists of C5 to C6 carbon molecules while heavy straight run naphtha (HSRN) consists of C6 to C10 long carbon molecule. The naphtha product coming out of the column is converted to gasoline after adjusting octane number, benzene reduction etc. Since the composition of the oil coming to refineries can change continuously and this oil first enters the atmospheric column, different hydrocarbon ratios in the oil are constantly changing.

In this direction, it is very important to collect and analyze products such as naphtha daily, to determine the actions to be taken by the production management team according to the hydrocarbon components of the product to be produced, and to produce products with a certain standard. ASTM (American Society for Testing and Materials) helps manufacturers, sellers, buyers, developers, and users to describe all the features of the product they want with one or a few words by drawing a framework for the production method, chemical, and physical properties of these standards.

1.2. Quality Control Methods

Gas Chromatography, which is a standard analysis method, performs qualitative and quantitative analyzes according to the peak area and time by delivering the products passing through a long thin column to the detector at different times by taking advantage of the difference in boiling points. It is very well known and used as analysis of volatile compounds since the 1950s.

Gas chromatography uses carrier gas through a narrow tube known as a column with a volatile sample. Samples are moving with the carrier gas thought column and separate based on their boiling point difference and their interaction with specific column properties. Schematic representation of a typical gas chromatography instrument is shown Figure 1.3. Components in the sample have a unique retention time, so the time starting with the injection of the sample ends when it reaches the detector. This is used to identify the sample, and the area under the peak is used to determine the concentration.

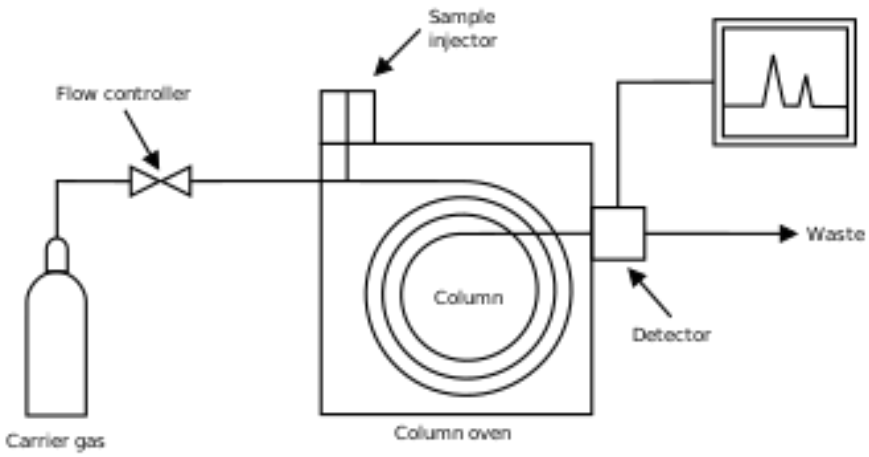


Figure 1.3 Gas Chromatography Scheme (Url-2)

The analysis of LSRN and HSRN products such as Paraffins, Olefins, Naphthenes, Aromatics, and benzene determine their concentration by the calculation of the area under their responsible peak. The American Petroleum Institute determine the standards of these products. A typical chromatogram of gasoline is given in Figure 1.4.

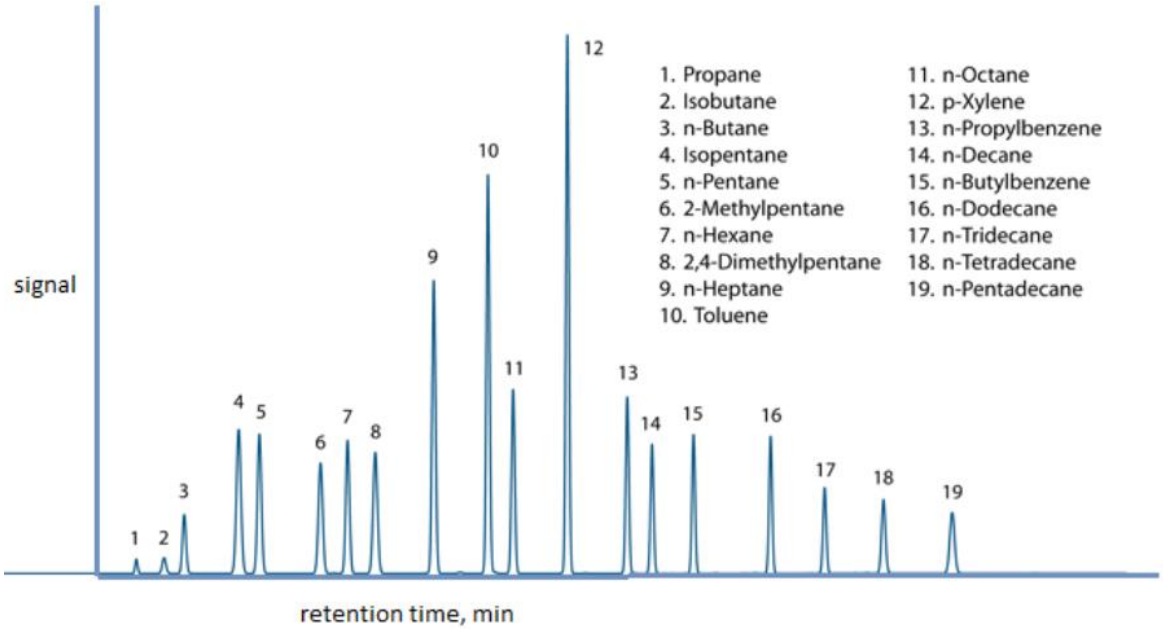


Figure 1.4 A typical Gas chromatogram of gasoline (Url-3)

1.3. Hydrocarbons Analysis

Naphtha includes various types of hydrocarbons from 5 carbons to 10 carbons. Paraffins, Olefins, Naphtenes and Aromatics (PONA) are important hydrocarbon groups. PONA analysis is the separation and characterization of these hydrocarbon mixtures, consisting of the initials of Paraffins, Olefins, Naphthenes and Aromatics fractions according to the carbon number or hydrocarbon type.

Paraffins

Their general formula is C_nH_{2n+2}. Some examples of paraffins can be given as Methane (CH_4), Ethane (C_2H_6), Propane (C_3H_8) and Pentane (C_5H_{12}) etc. that are shown in Figure 1.5.

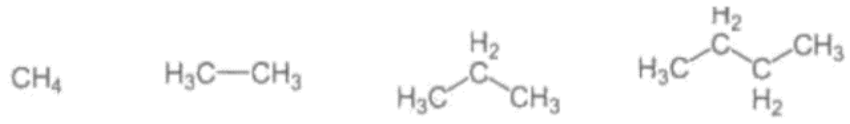


Figure 1.5. Structure formula of Methane, Ethane, Propane, and Butane

Naphthenes

Their general formula is C_nH_{2n}. (Provided that it is C≥3). These hydrocarbons are also known as cycloparaffins. Some examples of naphthenes can be given as Cyclopentane (C_5H_{10}), Cyclohexane (C_6H_{12}) etc. that are shown in Figure 1.6.

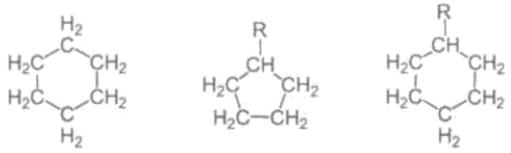


Figure 1.6. Structure formula of Cyclohexane, Cyclopentane with a functional group, Cyclohexane with the functional group

Aromatics

The general formula for aromatics is CnH2n-6. Benzene (C6H6) is the basic aromatics hydrocarbon. Derivatives of benzene molecular structure or those containing more than one benzene structure are called aromatic hydrocarbons. Some example of aromatic compounds are shown in 1.7.

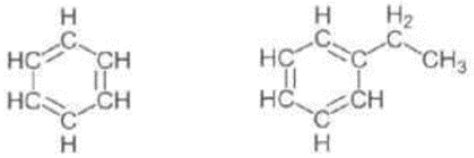


Figure 1.7. Structure formula of Benzene and ethylbenzene

Olefins

Olefins are known as unsaturated hydrocarbons that have double bonds between them. Their general formula is C_nH_{2n}. Some examples of olefins can be given as Ethylene, Propene, and 1-Butene that are shown in Figure 1.8.

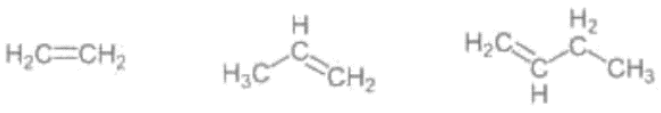


Figure 1.8. Ethylene, Propene, and 1-Butene

1.4. Literature Review

Researchers in many fields including academia and industry have used spectroscopy combining with chemometrics approaches since the 1970s. The acceleration of the computer with the developing technology and the advanced high-resolution spectrometers by providing opportunities to get faster results. The importance of chemometrics is increasing to that extent. To give examples of the main chemometrics methods are pattern recognition, classification, experimental design, clustering, and multivariate calibration. Using the multivariate calibration techniques are a very effective way of finding the concentration of a compound with the property measured in light of Beer's law. In general, multivariate calibration is used in complex chemical systems where univariate calibration is insufficient. NIR spectroscopy combined with multivariate calibration techniques give faster and more reliable results in various fields. The following review is based on relevant literature studies. When looking at the literature, better results were obtained in naphtha analysis, in which various methods based on chromatography were developed for refinery or petrochemical industry laboratories.[2-4] In addition, the ASTM D-5134[5] method, based on gas chromatography with flame ionization detector, is used for the determination of total Naphtha parameters by Tupras quality control laboratories. However, it was seen that the analysis was done in more than 2 hours. Accordingly, a long time makes it very difficult to optimize the analysis methods of continuous and large amounts of naphtha production in the process. It has been seen that this disadvantage can be overcome by using NIR spectroscopy and chemometrics tools with faster, more effective, and precise results[6-7].

In the study performed by Ku, Min-Sik[8], Naphtha samples like total Paraffins, Naphthenes, and Aromatics were determined by using NIR spectroscopy with PLS regression method. Parameters are predicted with different NIR spectral bands. All bands show excellent correlation with compare to conventional gas chromatography. According to the conclusion of the article, it is thought that NIR spectroscopy combined with multivariate calibrations may replace conventional gas chromatography. In the second publication of the author mentioned in the previous article, the quantitative analysis of naphtha products was compared using near infrared spectroscopy and Raman spectroscopy using PLS regression.[9] Chemical composition of total Paraffins with 6 naphtha parameters, total Naphthenes (cycloalkane), total Aromatics, C6 Paraffins,

benzene, and cyclopentane were used. In addition, specific gravity was used as physical parameter. PLS calibration models are built with two different methods, NIR and Raman spectroscopy. The predictive values showed good correlation with reference analysis in both methods. However, NIR was found to have better calibration performance. It has been concluded that the successful implementation of NIR can change the concept of process control and optimization in many refinery and petrochemical industries. Another study was performed with FT-NIR spectroscopy with PLS algorithm.[10] Naphtha samples were collected from control steam cracker processes. Many preprocess techniques are used for best-predicted performance. Nevertheless, control steam cracker process products are less impurity than the crude oil distillation products. In a different study for determination of the gasoline classification FT-NIR with different chemometrics approaches like support vector machine (SVM), K- nearest neighbor (KNN). 14000 cm^{-1}- 8000 cm^{-1} NIR spectral region was chosen. According to paper, with using NIR spectroscopy can help in the rapid and accurate classification.[11] In another study, a multivariate calibration was created based on the genetic algorithm established with diffuse reflection NIR spectra together with octane numbers of 60 gasoline samples taken over web source.[12] Divided into three sets. Genetic regression (GR), genetic classical least squares (GCLS) and genetic inverse least squares (GILS), which are three different genetic multivariate calibration methods, were used. Conclusion of this study that genetic algorithm with classical least square (CLS) and inverse least square (ILS) multivariate calibration techniques increase the prediction power of the model. In another article, it was stated that fast analysis has become an important trend in petroleum refining and more detailed molecular composition estimation of naphtha samples based on near infrared (NIR), which is a simple and effective analytical approach for rapid analysis, has been performed.[13] NIR spectra were collected with reference analysis of 101 naphtha samples, then the model was established with the chemometrics method Tchebichef curve moments (TCMs) and raw NIR spectra. Chemical composition of 26 hydrocarbons were used as parameter. It was concluded that the obtained 23 TCMs models reached excellent predictive quality. In this respect, it is concluded that NIR and TCMs method can be implemented quickly with simple, accurate and reliable analytical results.

1.5. Aim of the thesis

Reference analyses of naphtha samples analyzes in Tupras refineries are carried out by gas chromatography, which is the classical ASTM standard method. However, since this method takes a long time and is costly, performing analyses with a spectroscopic technique that will be an alternative to this method will provide faster results with less cost. LSRN and HSRN analyses using Near-Infrared Spectroscopy (NIR) coupled with chemometrics multivariate calibration methods can used to solve these problems.

In Figure 1.9, the modeling process, gathering NIR and GC analysis results then predict a successful model for every parameter.

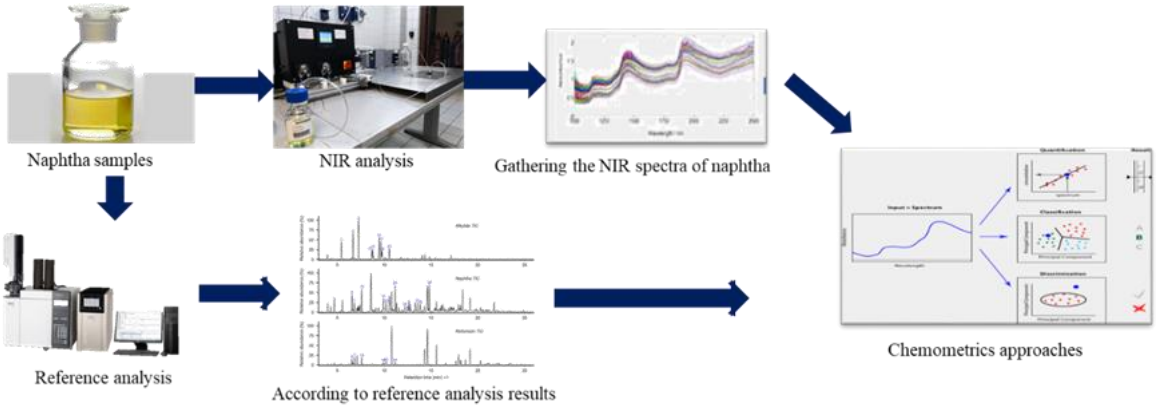


Figure 1.9. Chemometrics process

Figure 1.9 shows the schematic of the processes planned to be done in the thesis. The first step starts with the collection of naphtha samples. Subsequently, reference analysis and FT-NIR analysis of these samples are performed. Then, with these results, models are set up with FT-NIR spectra and reference analysis results using multivariate techniques. After the FT-NIR spectra of the new samples are obtained, estimation is performed for each parameters.

CHAPTER 2

NEAR INFRARED (NIR) SPECTROSCOPY

2.1. NIR Spectra and NIR Region

Fourier transform near-infrared spectroscopy (FT-NIR) is a spectroscopic method that focuses on the near-infrared region of the electromagnetic spectrum (10000 cm^{-1} to 4000 cm^{-1}). Near infrared (NIR) spectroscopy provides quantitative analysis of low molecular weight hydrocarbons without damaging the sample. Therefore, with the development of computer systems and faster result processing, NIR spectroscopy has become an important and popular method for chemical analysis. The discovery of infrared radiation is attributed to William Herschel, an astronomer who lived in the 19th century. (Url-4) Online analysis facility is an important factor as NIR provides. NIR spectra have broad bands that are complex and difficult to assign to a molecular structure. This challenge can be solved by using complex chemometrics models that combine spectroscopic data.

NIR is a form of molecular spectroscopy that provides complex structural information regarding the vibrational behavior between molecules. Like ultraviolet visible and mid-IR spectroscopy, the Beer Law applies to NIR. The main bands in the NIR region are the second or third harmonics of the fundamental, C - H, O - H and N - H stretching vibrations in the middle IR region.

2.2. Working Principle of FT-NIR

NIR spectroscopy based on measurement of transmittance, absorbance and reflectance are the most important modes.[14] Transmittance is calculated as the ratio of light passing through the sample to its loss. Information about the structure of the sample can be obtained from this type of analysis. The amount of light from the source passing through the sample is measured by how much light has passed along with the detector measurement. Transmittance is light in wavelengths remaining after the light has been

absorbed by the sample. The sample is between the light source and the detector. Transmittance (T) or absorbance (A) are given equation 2.1 and 2.2.[14-15]

$$T = \frac{I}{I_0} \tag{2.1}$$

Here I is the intensity of transmitted radiation and I_0 is the intensity of incoming radiation.

$$A = -\log_{10} T \qquad A = \log_{10}\left(\frac{1}{T}\right) \quad A = \log_{10}\left(\frac{I_0}{I}\right) \tag{2.2}$$

Diffuse Reflection (R) is measured in equation 2.3 by the reflection of the sample. The incoming light goes to the sample, some of the light is absorbed, and some is reflected again. The reflectance technique may give a better result for solid samples[15].

$$R = \frac{I}{I_r} \qquad A_R = \log_{10}\left(\frac{1}{R}\right) \quad A = \log_{10}\left(\frac{I_r}{I}\right) \tag{2.3}$$

Here I is the intensity of light reflected from the sample and I_r is the intensity of light reflected surface.

2.3. Advantages and disadvantages of NIR

NIR spectroscopy has an important advantage in providing quantitative analysis and structural determination of various sample types in solid and liquid form. NIR application is generally preferred because it is fast and easy to give results. After NIR analysis, the sample can be used for other analyzes as NIR does not cause any damage to the sample. In addition, there is no process such as sample preparation, which is an important step for rapid analysis. Taking advantage of fiber optics such as online analysis, NIR spectroscopy can be used in industrial areas. With these advantages, NIR is a very useful method compared to classical methods.

On the other hand, one of the challenges for NIR is the calibration task. For quantitative analysis, calibration models should be created separately for each feature. Calibrations by process and sample should be controlled occasionally. Combining

chemometrics with NIR is a powerful quantitative analysis method but also complex. It takes some training and time to spend.

2.4. Evaluation of NIR Spectra

According to L.G. Weyer[16], important spectral information can be seen in 14000 cm^{-1} – 4000cm^{-1} regions. However, spectra–structure correlations can be difficult than the mid-infrared (MIR). Aliphatic hydrocarbons are hydrocarbons based on chains of C atoms which can be found in the first overtones of C–H stretching occur between 5555–5882 cm^{-1}, the second overtones between 8264–8696 cm^{-1}, and the third overtones between 11 364 cm^{-1}- 10 929 cm^{-1} which can be related mostly the naphthenic compounds in crude oil. The Aromatics CH stretch can produce several bands at shorter wavelengths than the aliphatic CH absorptions. For example, benzene, the NIR absorption bands are at 8772 cm^{-1}, 5988 cm^{-1}), 4651–4587 $cm^{-1,}$ and 6065 cm^{-1}. The major peak at 5988 cm^{-1} has been assigned to the first overtone of the CH stretch. For the olefinic compounds, olefinic CH stretch first overtones occur in the NIR region. Appears from 6190–6110 cm^{-1}, while the symmetric $=CH_2$ appears at about 6110 cm^{-1}.

CHAPTER 3

MULTIVARIATE CALIBRATION

3.1. Overview

Chemometrics is a method for extract chemical information, both qualitative and quantitative, from the data produced by sophisticated chemical analysis or an experiment. With the collected data can be solved with the help of applied statistical and mathematical methods to turn it into a mathematical expression.[16]

Generally, data from the spectroscopic method can be used by the calibration method, where the known concentration of the sample is related to spectral information of the sample, the example of that absorbance. The model created in calibration step is then used to predict an concentration of unknown sample with spectral information.[17] There are two types of calibration approach namely Univariate Calibration and Multivariate Calibration.

3.2. Univariate Calibration

In univariate calibration, the concentration of a sample is determined by using the response of a single signal (i.e., chromatographic peak area) or a single spectroscopic wavelength. For the quantitative analysis, Beer-Lambert law is used for model creating, where the absorbance at a wavelength is directly linked to the absorptivity coefficient, light path length, and concentration, as described in Equation 3.1. Quantitative analysis of absorption spectroscopy begins with Beer-Lambert Law and the absorbance (A) at a single specified frequency is expressed as:

$$A = \varepsilon b c \qquad (3.1)$$

Where ε is the molar absorptivity at the frequency, b is the path length of the sample, and c is the concentration of the compound in solution. The law shows that the absorption intensity of a compound is linearly proportional to its concentration in the

homogenous mixture. In the case of a linear model, there are two options, namely, classical univariate calibration, and inverse univariate calibration.

3.2.1. Classical Univariate Calibration

The simplest calibration method is that the concentration of a compound is related to the absorbance value at one wavelength with the given model:

$$\mathbf{x} = \mathbf{c}.\,\mathrm{s} + \mathbf{e} \tag{3.2}$$

Where x is absorbance values at one wavelength for calibration samples, c is concentrations, and s; which is the multiplication of molar absorptivity and path length in Beer's law, is a scalar relating those vectors and determined by equation 3.2. Both x and c are vectors of the same size n.

The following mathematical operation is solving the scalar*s*:

$$\mathbf{c}'.\mathbf{x} = \mathbf{c}'.\mathbf{c}.\,\mathrm{s} \tag{3.3}$$

$$(\mathbf{c}'.\mathbf{c})^{-1}.\mathbf{c}'.\mathbf{x} = (\mathbf{c}'.\mathbf{c})^{-1}.(\mathbf{c}'.\mathbf{c}).\,\mathrm{s} \tag{3.4}$$

$$\mathrm{s} = (\mathbf{c}'.\mathbf{c})^{-1}.\mathbf{c}'.\mathbf{x} \tag{3.5}$$

Once *s* is solved, concentration of the unknown sample concentration can be calculated by equation (3.6), where the hat symbol shows the predicted values.

$$\hat{\mathbf{c}} = \frac{\hat{\mathbf{x}}}{\mathrm{s}} \tag{3.6}$$

3.2.2. Inverse Univariate Calibration

The classical calibration expects the error on predicted concentration is due to the instrumental response. However, the instruments become more sensitive and reproducible recent years and the instrumental errors are being reduced. The process of concentration

measurement involving human error which is (i.e., diluting, weighing) a great error source. Unlike the classical univariate calibration, inverse univariate calibration assumes that a property of the sample is a function of a response to minimize the errors due to concentration. The classical and inverse calibration is well schematized in Figure 3.1.

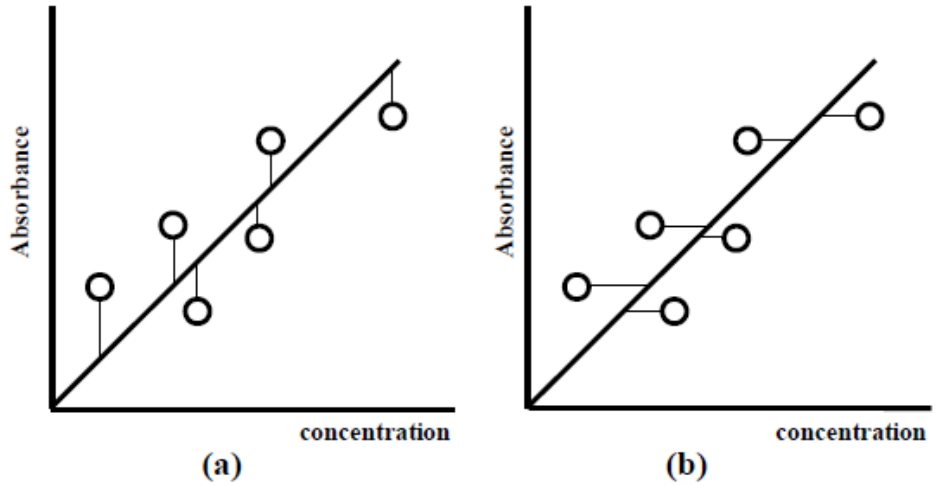


Figure 3.1. Errors in (a) Classical and (b) Inverse calibration[18]

The general inverse calibration is given below

$$\mathbf{c} = \mathbf{x}.\mathrm{b} \tag{3.7}$$

The different assumptions on the error distribution make the scalar b only approximately inverse of s. It is also evident in equation 3.8 that is very similar to the previous solution.

$$\mathrm{b} = (\mathbf{x}'.\mathbf{x})^{-\mathbf{1}}.\mathbf{x}'.\mathbf{c} \tag{3.8}$$

For a good set of data, the predictions of both classical and inverse calibration models should be in fair agreement. If not, there might be other factors such as non-linearity, outliers, and noise distributions.[23]

3.3. Multivariate Calibration

With the necessity of finding molar absorptivity for each pure component separately, the univariate calibration becomes a challenging task. Multivariate calibration allows the multiple components analysis in a sample simultaneously. It can be a more selective and reliable tool for fault detection ability and can analyze non-homogenous or contaminated analyte[21].

Using multiple wavelengths gives reliable results with the averaging of useful information in absorbance as well as noises. Unlike determining multiple components at once, multivariate calibration can be used.

In the following sections, different types of multivariate methods are provided from simple to more complex.

3.3.1. Classical Least Squares (CLS)

Classical least squares is very similar to Beer-Lambert Law, in which the absorbance values are expressed as a function of concentration. Like classical univariate, errors are supposed to be based on instrument responses. $\mathbf{X}$ CLS model for m calibration samples, l chemical compounds, and n wavelengths, is expressed in matrix as:

$$\mathbf{X} = \mathbf{CK} + \mathbf{E_X} \tag{3.9}$$

$\mathbf{X}$ represents an $m \times n$ matrix of calibration spectra, $\mathbf{C}$ $m \times l$ matrix of component concentrations, $\mathbf{K}$ $l \times n$ matrix of absorptivity path length constants, and $\mathbf{E_X}$ $m \times n$ matrix of spectral errors or residuals that are not fit by the model.

The estimation of $\mathbf{K}$ matrix is done by CLS with Equation 3.10:

$$\widehat{\mathbf{K}} = (\mathbf{C}'.\mathbf{C})^{-1}.\mathbf{C}'.\mathbf{X} \tag{3.10}$$

The concentrations of unknown samples can be predicted with their spectrum using the equation 3.11. Here, x represents the spectrum of an unknown sample.

$$\hat{\mathbf{c}} = (\widehat{\mathbf{K}}.\widehat{\mathbf{K}}')^{-1}.\widehat{\mathbf{K}}.\mathbf{x} \tag{3.11}$$

The drawback of CLS is that the concentrations of interfering species must be known apriority and included in the model.[20]

3.3.2. Inverse Least Squares (ILS)

Inverse least squares, like inverse univariate calibration, describe the properties of a sample as a function of the response as in equation 3.12. The model errors of ILS trust on the errors in the measurements of component concentrations.

$$\mathbf{C} = \mathbf{XP} + \mathbf{E_c} \tag{3.12}$$

where, $\mathbf{C}$ is the $m \times l$ concentration matrix, $\mathbf{X}$ $m \times n$ absorbance matrix, $\mathbf{E_c}$ $m \times l$ error matrix of concentrations that do not fit by the model and $\mathbf{P}$ is the calibration coefficients matrix with the size $n \times l$, relating component concentrations to the spectral intensities. If the elements in the $\mathbf{E_c}$ are assumed to be independent, identical analysis for each analyte can be done from equation 3.13, where a single component is modeled at a time.

$$\mathbf{c} = \mathbf{Xp} + \mathbf{e_c} \tag{3.13}$$

Here $\mathbf{p}$ is a $nx1$, $\mathbf{e_c}$ is a $mx1$, and thus $\mathbf{c}$ is a $mx1$ matrix. When making calibration, the least square solution of $\mathbf{p}$ in equation 3.13 yields:

$$\hat{\mathbf{p}} = (\mathbf{X^t}.\mathbf{X})^{-\mathbf{1}}.\mathbf{X}^{\mathbf{1}} \times \mathbf{c} \tag{3.14}$$

Finally, the concentration of the component in the unknown sample can be predicted as follows:

$$\hat{\boldsymbol{c}} = \mathbf{x}'.\hat{\mathbf{p}} \tag{3.15}$$

Where $\hat{c}$ the scalar estimated concentration of the analyzed component and x is the spectrum of unknown sample[20]. The strength of this method is that one does not need to know all the components in the sample. In addition, one can select as many variables, i.e.,

wavelengths, instead of using the full spectra. The main drawbacks of ILS are not being able to detect all the outliers, and not being very effective in selecting the optimal wavelength for predicted models. Moreover, adding more wavelengths to the model can lead to overfitting.[21]

3.3.3. Partial Component Analysis (PCA)

In multivariate calibrations, it is often difficult to see the relationship within the variables because the data are very large. However, the purpose of using principal component analysis is to reduce the size of data by taking its highest variance, while preserving important information from the data.[21]

Principal components define the variance in independent variables by mathematical transformation and can be formulated as Equation 3.16.

$$X = T * P + E \tag{3.16}$$

PCA divides the original X data matrix into different set of smaller size matrix. It represented by the score (T) and loading matrices (P) and E. The score matrix consists of column vectors, while the loading vectors are made up of row vectors.[22] As a result, the data matrix has been reduced to these terms for easy understanding; this is more applicable to reduced noise and gets more information from the data set. The data visualization can be explained with PC lines, these graphs help spectral understanding for discrimination and classification purposes.[22]

3.3.4. Partial Least Squares (PLS)

Partial Least Squares regression known one of the most used methods of multivariate calibration. Herman Wold first used this method in 1966 to model economic and social events. Kowalski et. al. have used the PLS method in the field of chemistry. It was started to be used after an initial study.[23]

PLS1 algorithm uses different number of PLS factors (PCs) for each concentration variables. It is supposed that errors can be caused by spectra or concentrations. The most

significant distinction of PLS from ILS is being less vulnerable to overfitting and become models that are more robust.

In PLS1, 2 model sets are constructed as follows:

$$\mathbf{X} = \mathbf{TP} + \mathbf{E} \tag{3.17}$$

$$\boldsymbol{c} = \mathbf{Tq} + \mathbf{f} \tag{3.18}$$

Where X dependent variable (e.g., absorbance data), c independent variable (e.g., concentration), scores matrix for T of PLS. Here, P and q are analogous to loadings vector with the multiplications T with P, and q are used to estimate spectral data and concentration. As Figure 3.2.

The absorbance and concentration matrices are represented as **X** and c, respectively. A crucial feature of PLS is that scores matrix, **T**, is common both for concentration and measurement. Here, P and q are analogous to loadings vector, and the multiplications **TP** and **Tq** are used to approximate spectral data and true concentration, respectively. Figure 3.2 represents the matrix operation. The sum of the squares of the scores of each component is called an eigenvalue. A more significant component has a greater eigenvalue.

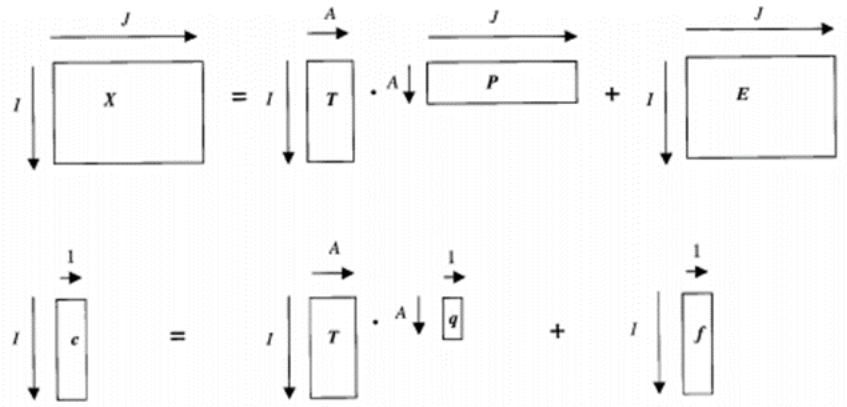


Figure 3.2. A schematized matrix view of the PLS1 algorithm

3.3.5. Genetic Inverse Least Squares (GILS)

ILS with Genetic Algorithm (GA) for selecting wavelengths to build multivariate calibration models with reduced data set. The genetic algorithm is a search and optimization technique that is from Darwin's theory of evolution and natural selection.[28]According to Darwin's theory, whoever has higher survival takes place by choosing the species and progressing. As this process continues over a long period, through generations, subsequent offspring will have a greater chance of survival than previous generations. Scientific researches started in the 1960s by biologists doing genetic systems experiments on a computer. The pioneer of the field is John Holland, who developed a GA in his research on adaptive systems in the early 1960s[25]. From that day to today optimization problems are solved by using GA tools, and applications found in calibration, more specifically on wavelength selection.[26-27]

Genetic Inverse Least Squares (GILS) is the calibration technique that combines GA and ILS.[32] The genetic algorithm consists of 5 main steps as follows;

Initialization

A gene is formed with a random selection of instrumental responses. It can be represented by the following expression, where S symbolizes the gene and A the absorbance measured wavelength in the subscript:

$$S = [A_{2678}A_{256}A_{1478}A_{560}]$$

A population is the collection of individual genes. In the initialization step, the first generation of genes is randomly generated with fixed population size. The random selection of responses enables the minimization of bias and maximization of the number of recombination. The population size is an important matter since it determines the time to complete an individual run of the algorithm, i.e., a larger size needs more time. The number of genes in the population needs to be an even number to allow breeding of the genes.

Moreover, there is a constraint in choosing the number of wavelength points in a gene, that is, it must be obtained randomly between a specified high and low limit. The

higher limit is chosen to prevent overfitting problems and to reduce the computation time, and the lower limit is set to 2 to allow single-point crossover.

Evaluation of the Population

In the second step, evaluation and ranking of the genes are done using a fitness function, which is defined as the inverse of the standard error of calibration (SECV) with cross-validation and shown in equation 3.1.

$$Fitness = \frac{1}{SECV} \tag{3.19}$$

SEC indicates the success of each gene and is calculated from equation 3.20.

$$SECV = \sqrt{\frac{\sum_{i=1}^{m}(c_i - \hat{c_i})^2}{m-2}} \tag{3.20}$$

Here, c_i is the reference concentration, $\hat{c_i}$ is the predicted concentration, $m-2$ is the degrees of freedom, while m is the number of samples, and 2 indicates the extracted parameters, which are the intercept and the slope between the reference and the predicted concentrations.

Selection of Genes for Breeding

This step relies on the natural evolution principle and the members with the highest fitness values are selected, who are to be replaced with parent genes. The highest fitness means a better-suited gene that can survive and transfer information to the next generation. There are several methods for parent selection. One of them is the top-down method, in which the genes are ranked in the pool, and they mate consequently, i.e., first gene (S1) mate with the second gene (S2) with the third (S3) and so on. This process gives all the genes a chance to breed. Another method is called roulette wheel selection, and it is illustrated in Figure 3.3.

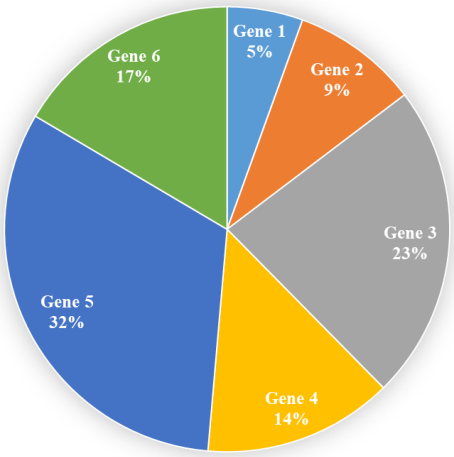


Figure 3.3. A visual representation of the roulette wheel

Each part of the wheel represents a gene, and each area in the wheel is proportional to the fitness of a gene. A gene occupying a higher area in the wheel has more chance of being selected. In this method, some genes can be selected multiple times, while others might not be even selected and thrown out of the pool. After the selection, parents that are selected are mated top-down. Since there is no ranking, the number of possible recombination increases.

Crossover and Mutation

Random points are chosen to create new offspring genes and gene comprises the fracture process by crosslinking. Most of the work of GA is on this step. In the following, the process is illustrated where the parents are S1 and S2 genes, and their offspring are S3 and S4.

Parents:

$$S1 = [A_{223}A_{3750}A_{8212} : A_{8123}A_{344}]$$
$$S2 = [A_{140}A_{7786} : A_{569}A_{4064}]$$

Offspring:

$$S3 = [A_{223}A_{3750}A_{8212}A_{569}A_{4064}]$$
$$S4 = [A_{140}A_{7786}A_{8123}A_{344}]$$

The parent genes are randomly cut from the parts indicated as to ⋮ where separation and crossover of the genes take place.

In addition, sometimes mutation can take place in this step by introducing random deviations into the population. One can realize this into the algorithm during mating at a rate of 1%, which is a typical rate. Generally, one of the wavelengths is replaced in an existing gene with another wavelength. However, the GILS studies today usually do not apply mutation.

Replacing the Parent Genes by Their Off-Springs

After the cross-over, the parent genes S1 and S2 are replaced by their offspring S3 and S4, which are evaluated and ranked. Now, the selection for breeding starts all over again with the new genes and repeated until the predefined iterations or the minimum tolerance value is reached.

Finally, the gene with the lowest SEC is selected for model building, which will then be used for the prediction of concentration of the sample being analyzed in the validation set. The success of the model in the prediction of validation set is determined by the standard error of prediction (SEP), as given by equation 3.21. Here, m represents the number of validation samples.

$$SEP = \sqrt{\frac{\sum_{i=1}^{m}(c_i - \hat{c}_i)^2}{m}} \qquad (3.21)$$

Once the predefined iteration number is reached, termination takes place. It can also be optimized by extensive statistical tests. Often, the decision of the best run is given when the lowest SEC for the calibration is quite close to the SEP value.

The GILS approach has various advantages over univariate and other multivariate calibration methods. It does not require any complex mathematical operation in the construction of the model and during the prediction process. In this study, GILS is used considering the advantages and accuracy of the method.

CHAPTER 4

EXPERIMENTATION AND INSTRUMENTATION

4.1. Experimentation

A total of 95 light straight run naphtha and a total of 67 heavy straight run naphtha samples were obtained for 2 years at Tupras İzmit Refinery from the crude oil atmospheric distillation unit. Samples were stored at 4 ^{o}C temperature in a dark place before analysis to prevent the evaporation of hydrocarbons and possible interferences as a result of direct light.

4.2. Instrumentation

Chemical composition of all samples was determined in Tupras Izmit Refinery Quality Control Laboratory, according to the ASTM 5134 (General requirements competence of testing and calibration laboratories). The concentration of each sample parameters measured by a multidimensional gas chromatography (AC Reformulyzer, PAC, USA), which has flame ionization detector. For analysis 0.2 µL to 1.0 µL of naphtha samples was injected and temperature programming from 35 °C to 200 °C in 1 °C/min. Near-Infrared absorption spectra of samples were recorded on MATRIX-F FT-NIR Spectrometer (Bruker-Germany) spectral range from 10000 cm^{-1} to 4000 cm^{-1} with using 2.0 mm pathlength quartz cell. Before each analysis, a background spectrum of air was recorded in with a clean dry cell. All FT-NIR spectra were collected with an average of 4 scans and resolution of 4 cm^{-1}. The room temperature was kept at 22±2 ^{o}C for spectrum acquisition.

4.3. Data analysis

The collected spectra in ASCII format were transferred to Microsoft® Excel® 2016. Then the data analysis are performed by chemometric calibration toolbox [29](OBA Quantifier, OBA kemometri Inc. Turkey) which is developed in the MATLAB R2018b (Math Works Inc., MA) environment. Genetic Inverse Least Squares (GILS) and Partial Least Squares Regression (PLSR) were performed for this study.

CHAPTER 5

RESULTS AND DISCUSSION

5.1. Near Infrared Spectra

Infrared ray has a longer wavelength and a lower frequency than visible light. The basic theory is that chemical bonds molecules absorb at different frequencies. Therefore, the sample with different structure will also have a different spectrum. In addition, the same sample with different ratios like different concentration, allow us to use multivariate calibration methods to make quantitative predictions and to find expected variation patterns. This conquers the restrictions of the conventional analytical chemical methodology, where non-analyte source, for examples, interfering compound constituents or physical phenomena, should have been eliminated physically in the samples before prediction. The two methodologies can be join into one, comprising of information-driven ' chemical ' or 'physical' pre-processing followed by a data-driven 'statistical' calibration modeling.

Near-Infrared spectra of 95 light straight run naphtha samples were collected for 2 years. Likewise, Near-Infrared spectra of 67 heavy straight run naphtha samples were collected for 2 years. Figure 5.1 and Figure 5.2 shows the raw NIR absorbance spectra, which were recorded in the 10,000 cm^{-1}– 4000 cm^{-1} wavenumber region for raw light straight run naphtha and Heavy straight run naphtha respectively.

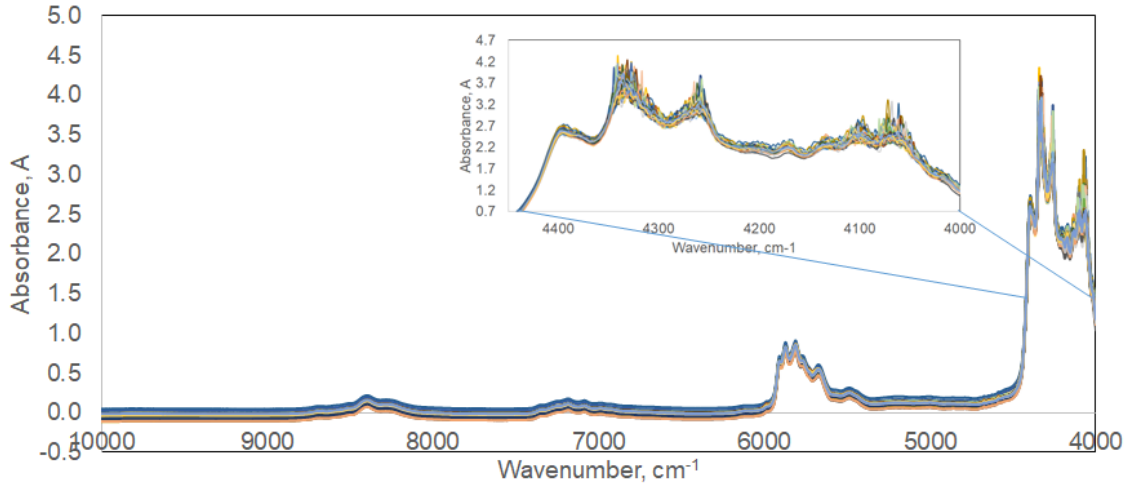


Figure 5.1. Raw FT-NIR spectra of a total of 95 Light straight run naphtha samples

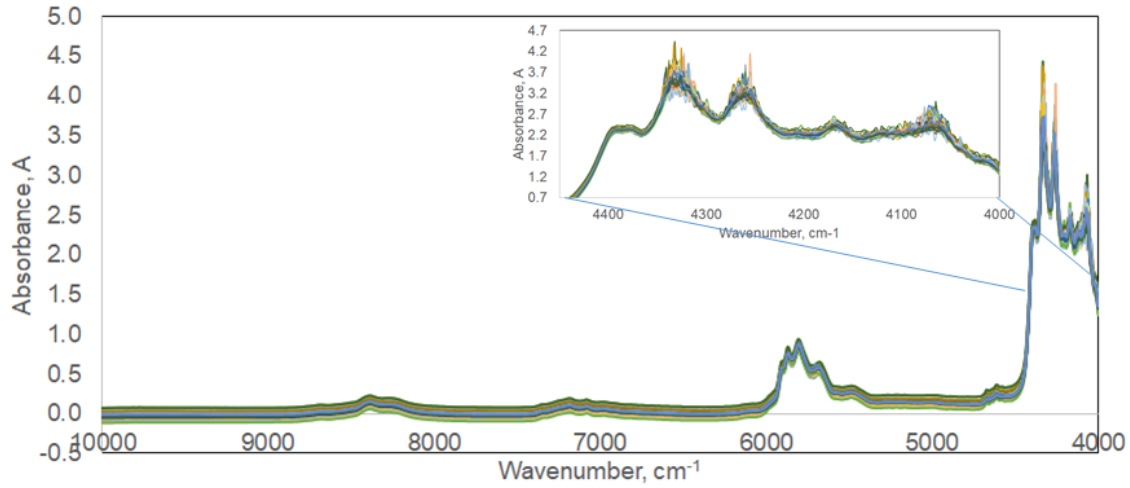


Figure 5.2. Raw FT-NIR spectra of a total of 67 Heavy straight run naphtha samples

As shown in Figure 5.1 and 5.2 absorbance value between 4450 cm^{-1} and 4000 cm^{-1} spectral region was removed because the absorbance value is found to be greater than 2 which can cause nonlinearity problems according to Beer's law rule. In addition, the small portion of spectra covering from 4450 cm^{-1} to 4000 cm^{-1} shows some noisy futures.

In Figure 5.3 and 5.4, the NIR spectra total of 95 of light straight run naphtha and total of 67 of Heavy straight run naphtha samples with narrowed intervals are shown, respectively.

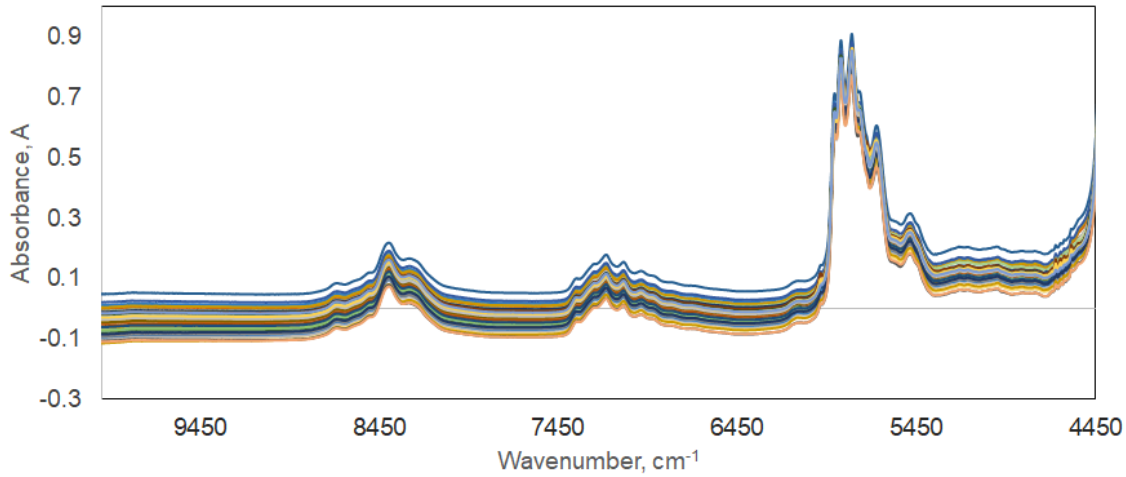


Figure 5.3. FT-NIR spectra of a total of 95 Light straight run naphtha with narrowed range

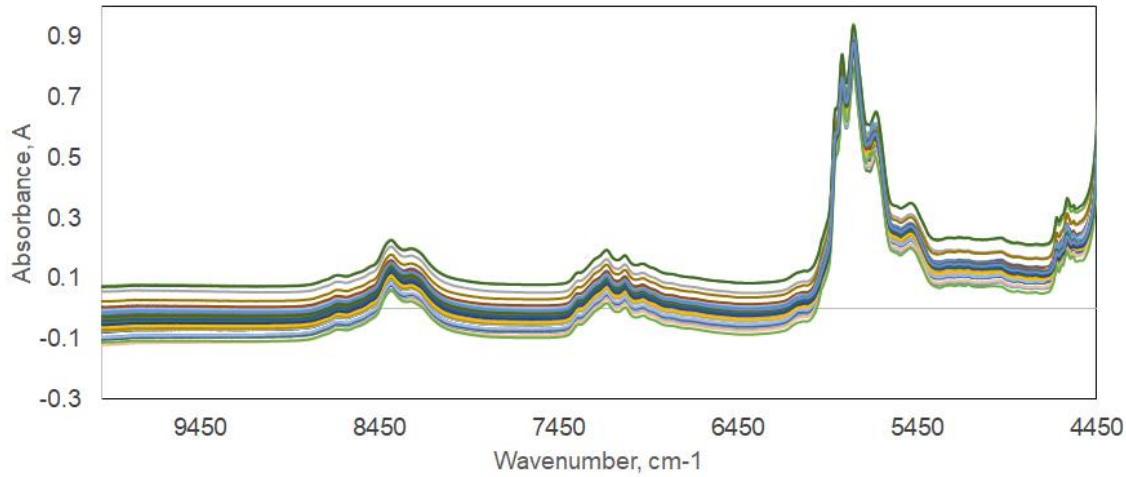


Figure 5.4. FT-NIR spectra of a total of 67 Heavy straight run naphtha with narrowed range

In analyzing samples from the NIR spectrometer, there is a great deal of chemical information to be used to identify the sample. However, due to the wide structure of the overtones in the NIR spectra, it is very difficult to associate these bands with the chemical bonds in the sample. Also, in spectral analyzes, deviations often occur, known as noise, that do not contain information about the sample or caused by scattering from the light source, sample cell, or particles heterogeneous mixtures in the sample during measurement. Therefore, qualitative and quantitative analysis of the sample will not be possible without applying any mathematical operation to the spectrum. In this direction,

by applying Extended Multiplicative Scattering Correction (EMSC), which is a preprocessing technique in the literature.

Extended Multiplicative Scattering Correction help to remove uncontrolled variation in light scattering due to uncontrolled physical variation by subtracting interferences at the preprocessing stage improves the interpretability of regression modeling. EMSC preprocessing effectively removed most of the path length and baseline effects, allowing the subsequent additive PLSR to work well. EMSC provides clear improvement over the traditional other preprocess methods[34]. The Extended Multiplicative Scattering Correction is specified in equation 5.1 and 5.2 given below;

$$X = a_i + \bar{X}_i * b_i + d_i * \lambda + e_i * \lambda^2 + \cdots + d_n * \lambda^n + E \quad (5.1)$$

$$X_{Corr} = \frac{\bar{X}_i - a_i - d_i\lambda - e_i * \lambda^2 - \cdots - d_n * \lambda^n}{b_i} \quad (5.2)$$

Where X is dependent variable (e.g., absorbance data), a_i is a constant baseline, λ is the wavelength vector; the coefficients a_i, b_i, d_i and e_i can be estimated by least squares regression of X. $\bar{X}_i$ is mean of dependent variable. After the calculation X_{Corr}, which is the corrected dependent variable, used in equation 5.2.[30]

By using these equations mentioned in equation 5.1 and equation 5.2. EMSC corrected NIR spectra of LSRN and HSRN samples are shown in Figure 5.5 and Figure 5.6, respectively.

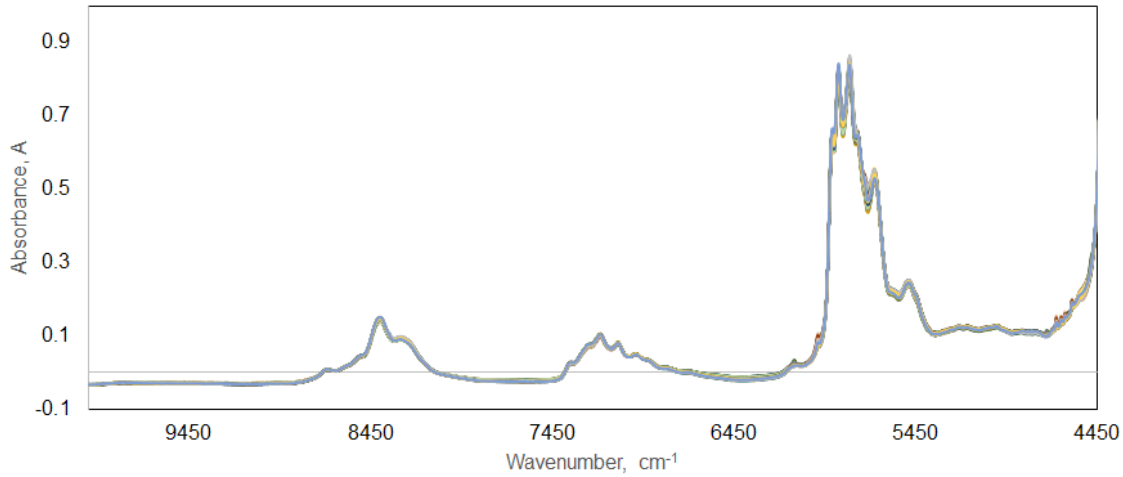


Figure 5.5. FT-NIR spectra of LSRN with EMSC preprocess and narrowed range

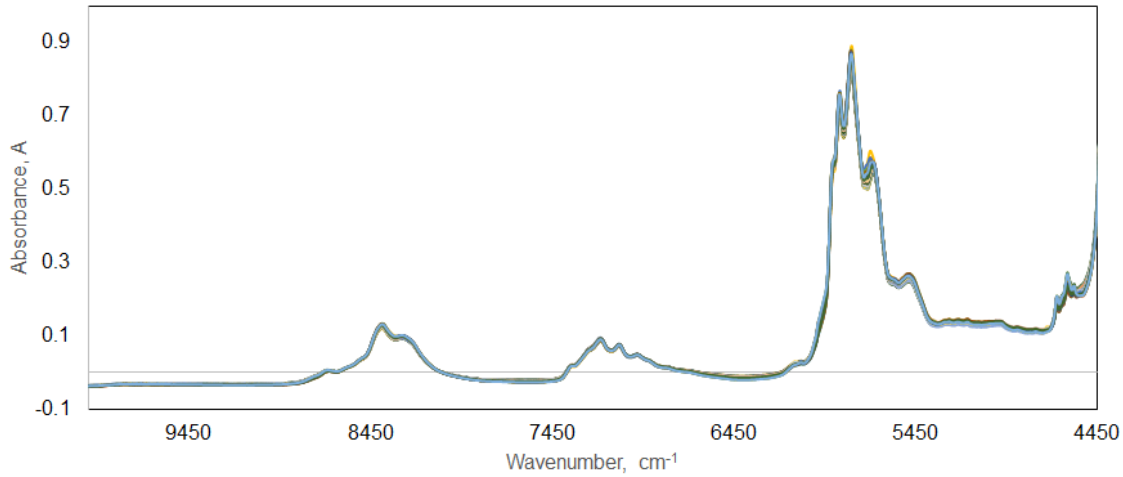


Figure 5.6. FT-NIR spectra of HSRN EMSC preprocess and narrowed range

As shown in Figure 5.6 and 5.7, after EMSC normalization in order to remove baseline shifts from raw data. EMSC helps to perform better interpretability of the LSRN and HSRN spectra and making calibration models statistically more robust.

Before multivariate calibration studies, Principal component analysis was performed in order to detect outliers.

5.2. Principal Component Analysis (PCA)

Principal Component Analysis is described in detailed explanation section 3.3.3. At this stage, PCA was used for outlier detection. After preprocessing, as the last step before

starting regression, the NIR spectra of all samples were checked against any outliers in the data set with PCA analysis. Charts with scores on Latent Variable 1 (Principal Component 1) and scores on Latent Variable 2 (Principal Component 2) measured and predicted were used to identify any outlier. Figure 5.7 and 5.8 shows the score plot of principal component 1 vs principal component 2 for LSRN and HSRN.

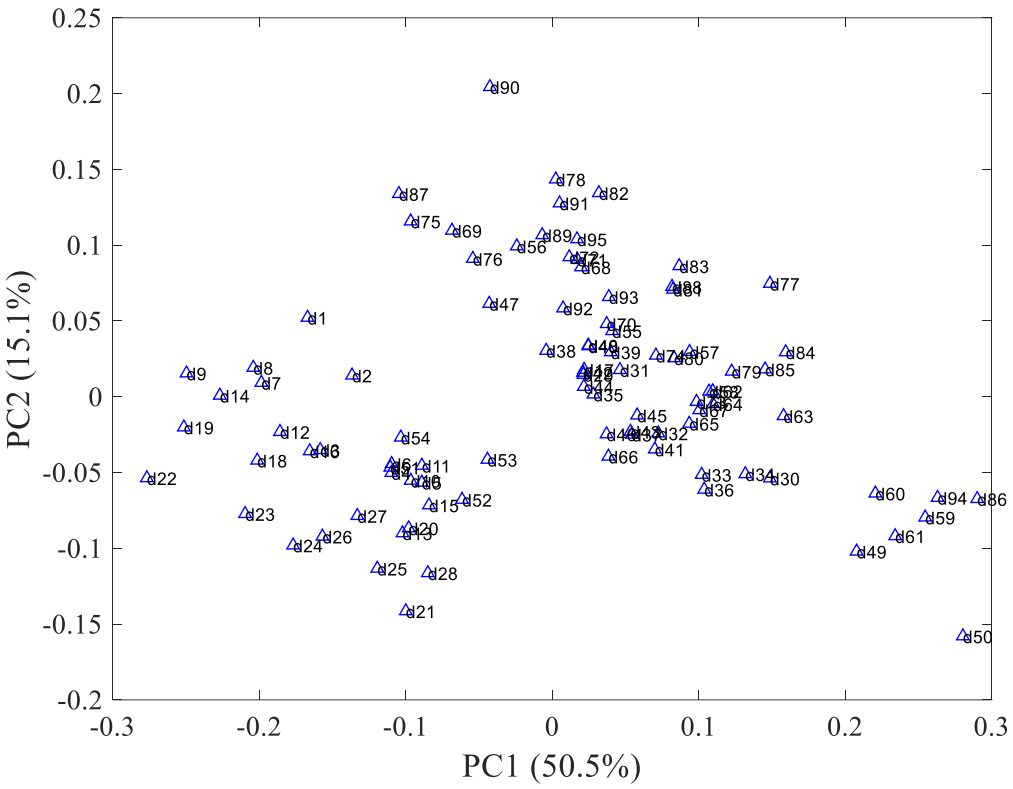


Figure 5.7 The scores plot of the first component (PC1) versus the second component (PC2) for LSRN using FT- NIR spectra

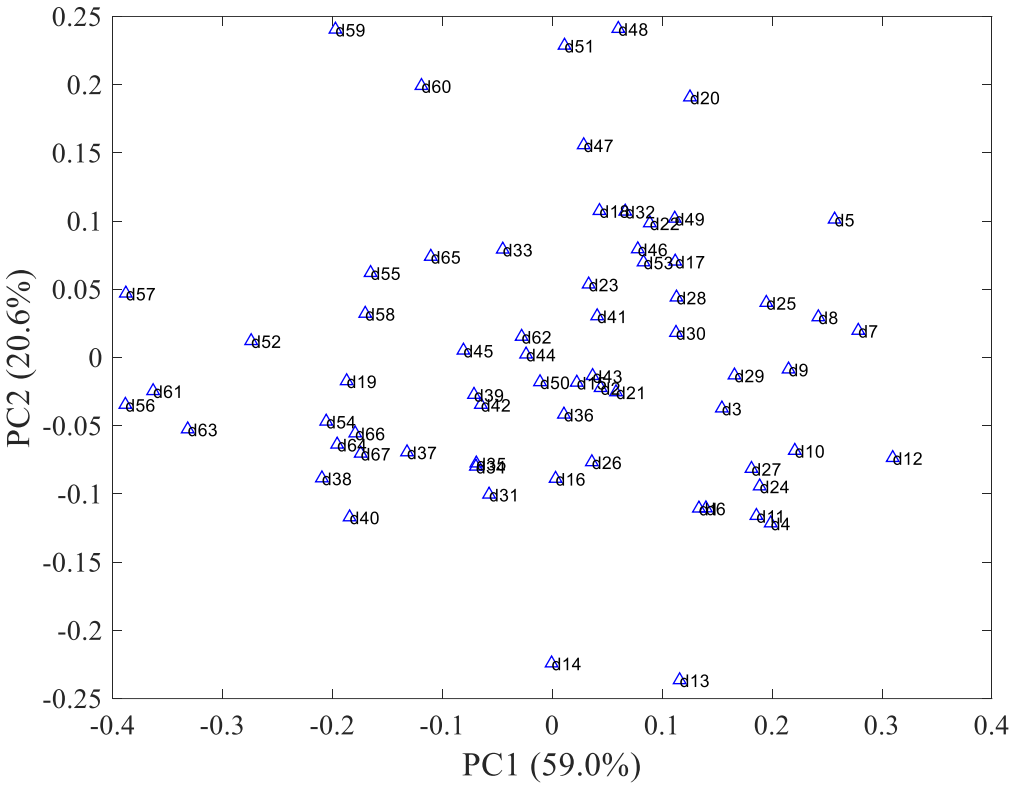


Figure 5.8. The scores plot of the first component (PC1) versus the second component (PC2) for HSRN using FT- NIR spectra

As shown in Figure 5.7, PCA scores of LSRN were plotted using the first 2 PCs, where the first and second PCs contain 50.5%, 15.1% (cumulative 65.6%) of the total variance respectively. For LSRN samples, there are two different cluster groups within themselves. The reason for this may be the grouping resulting from oil processing in two different crude oil was given distillation unit at that time. In Figure 5.8 PCA scores of HSRN were plotted using the first 2 PCs, where the first and second PCs contain 59.0%, 20.6% (cumulative 79.6%) of the total variance respectively, As shown in Figure 5.7 and Figure 5.8, LSRN and HSRN samples do not show any samples further away from their clusters center.

5.3. Multivariate Analysis

EMSC preprocessed spectra shown in Figure 5.5 and 5.6, two different calibrations approached were performed which are Genetic Inverse Least squares (GILS)

and Partial Least Squares (PLS) Regression. Multivariate calibration results for both methods are given in the following sections.

For both multivariate calibrations, a total of 63 light straight run naphtha samples were assigned as a calibration set to construct the model while the rest of 32 samples were chosen as an independent validation set to observe the predictive ability of the model. Also for heavy straight run naphtha of 45 samples assigned as a calibration set and the rest of 22 samples were chosen as an independent validation set to observe the predictive ability of the model. At least 3 samples each at the upper and lower limits were left for the calibration data to make the resulting model cover boundary conditions and account for most of the variance and all remaining samples were randomly distributed.

5.4. Partial Least Squares Regression

The partial least squares (PLS), which is one of the most widely used multivariate calibration methods in spectroscopic methods as regression method, is defined as the analysis method that can associate independent variables (spectrum) with dependent variables (in this case, the concentration of PONA products). In this context, PLS was applied to the data set, where pre-treatments were applied.

Prediction models both for LSRN and HSRN, models were created for Naphthenes, Paraffins, Olefins, Benzene, Aromatics, C7Plus (the sum of compounds with more than 7 carbons) and C6Minus (the sum of compounds with less than 6 carbons). A modeling study has not been done for the Olefins parameter in HSRN due to a lack of data. The prediction performances of the created models were evaluated by looking at the coefficient of determination (R^2) of the calibration and validation data set, the root mean square error of calibration (SEC), and the root mean square of validation errors (SEP) data. Models with low SEC, SECV, and SEP values and high determination coefficients were preferred at the stage of selecting the basic component numbers of the established models.

5.4.1. Light Straight Run Naphtha Results

For finding the best fitting number of LVs, Predicted residual error sum of squares (PRESS) values were calculated for the first 30 LVs and the results are given in Figure 5.10. For following part, the results which are Paraffins, Olefins, Naphthenes and Aromatics are given after that remaining three models which are given Benzene, C7 plus, and C6 minus PLS model results will be given for a clear explanation.

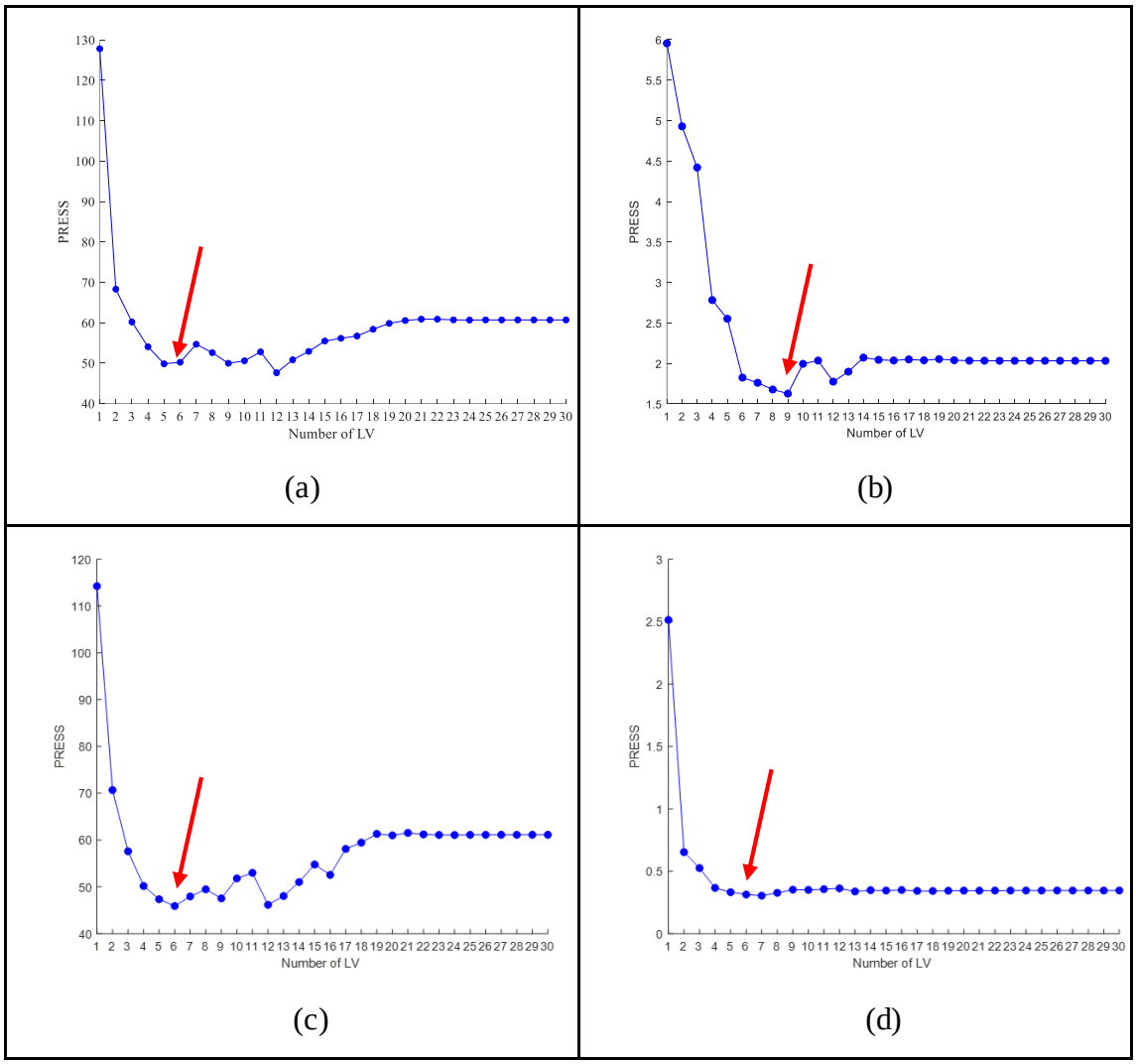


Figure 5.9. Number of PCs vs. PRESS plot for selecting the optimal number of LVs a) Paraffins b) Olefins c) Naphthenes d) Aromatics

By using Figure 5.9, Paraffins, Olefins, Naphthenes, Aromatics and Paraffins PLSR models with 6 LVs, 9 LVs 6 LVs and 6 LVs selected respectively. It can be seen that they are modelled with the relatively low number of component. Reference values

obtained from Reference analysis vs PLS model predicted values of FT-NIR spectra for each parameter are given in Figure 5.11. The standard error of calibration (SECV) and standard error of validation (SEP) are calculated for each parameter.

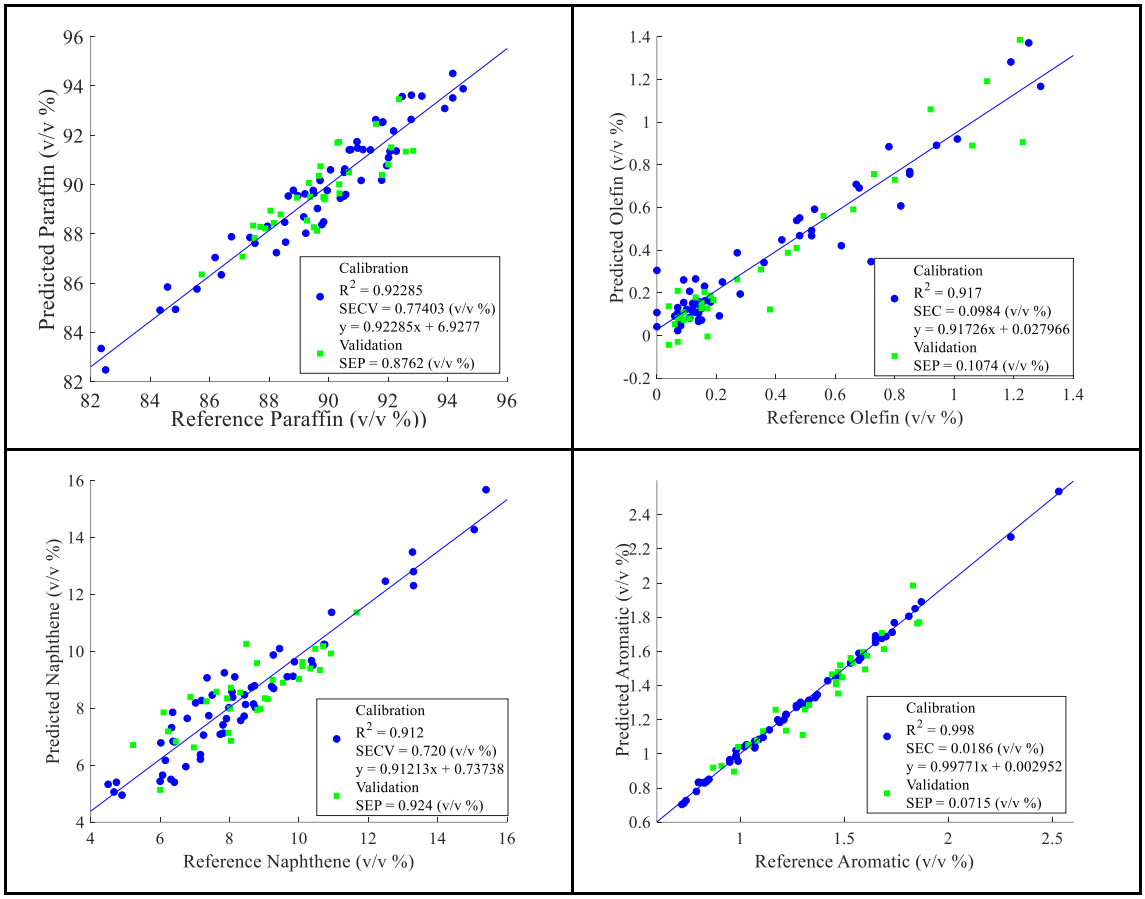


Figure 5.10. Actual concentrations vs. PLS predicted concentrations; Paraffins, Olefins, Naphthenes, Aromatics

As seen in Figure 5.10, the model performance is quite close for calibration and validation set predictions with calibration performance being only slightly better indicating no significant overfitting. For Paraffins, SECV and SEP values are found to be 0.774 (v/v %) and 0.876 (v/v %), respectively. The R^2 value for calibration set predictions are calculated as 0.923, and the R^2 value for the validation set is 0.751. The residuals for all samples are plotted in Figure 5.11 a. While most of the residuals are in the range of ±1.5 (v/v %). For Olefins, SECV and SEP values are found to be 0.0984 (v/v %) and 0.1074 (v/v %), respectively. The R^2 value for calibration set predictions are calculated as 0.917, and the R^2 value for the validation set is 0.926. The residuals for all samples are plotted in Figure 5.11 b. The residuals are in the range of ± 0.4 (v/v %). As seen in Figure 5.11 b the narrower, the range makes the ability to predict more difficult, the more

samples will improve the ability to predict. The range of reference concentrations is very narrow which makes the ability to predict more difficult, the more samples may improve the ability to predict. For Naphthenes, SECV and SEP values are found to be 0.720 (v/v %) and 0.924 (v/v %), respectively. The R^2 value for calibration set predictions are calculated as 0.912, and the R^2 value for the validation set is 0.682. The residuals for all samples are plotted in Figure 5.11 c. The most residuals are in the range of ± 1.5 (v/v %). For Aromatics, SECV and SEP values are found to be 0.0186 (v/v %) and 0.0715 (v/v %), respectively. The R^2 value for calibration set predictions are calculated as 0.998, and the R^2 value for the validation set is 0.942. The residuals for all samples are plotted in Figure 5.11 d. Most residuals are in the range of ± 0.15 (v/v %).

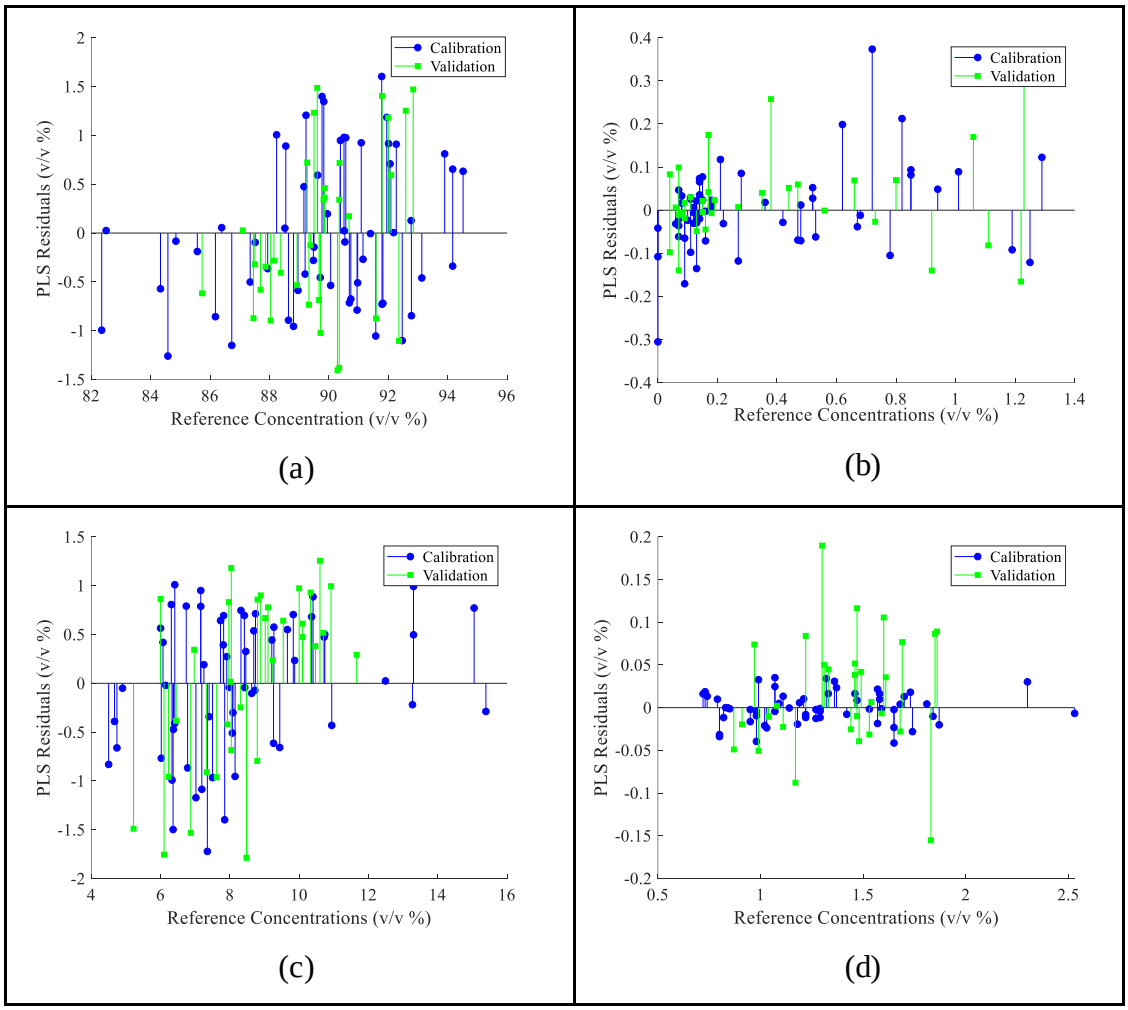


Figure 5.11. Reference concentrations vs. corresponding PLSR prediction residuals a) Paraffins b) Olefins c) Naphthenes d) Aromatics

For the following part, 3 parameters Benzene, C7 plus, and C6 minus PLS model results are shown in Figure 5.12.

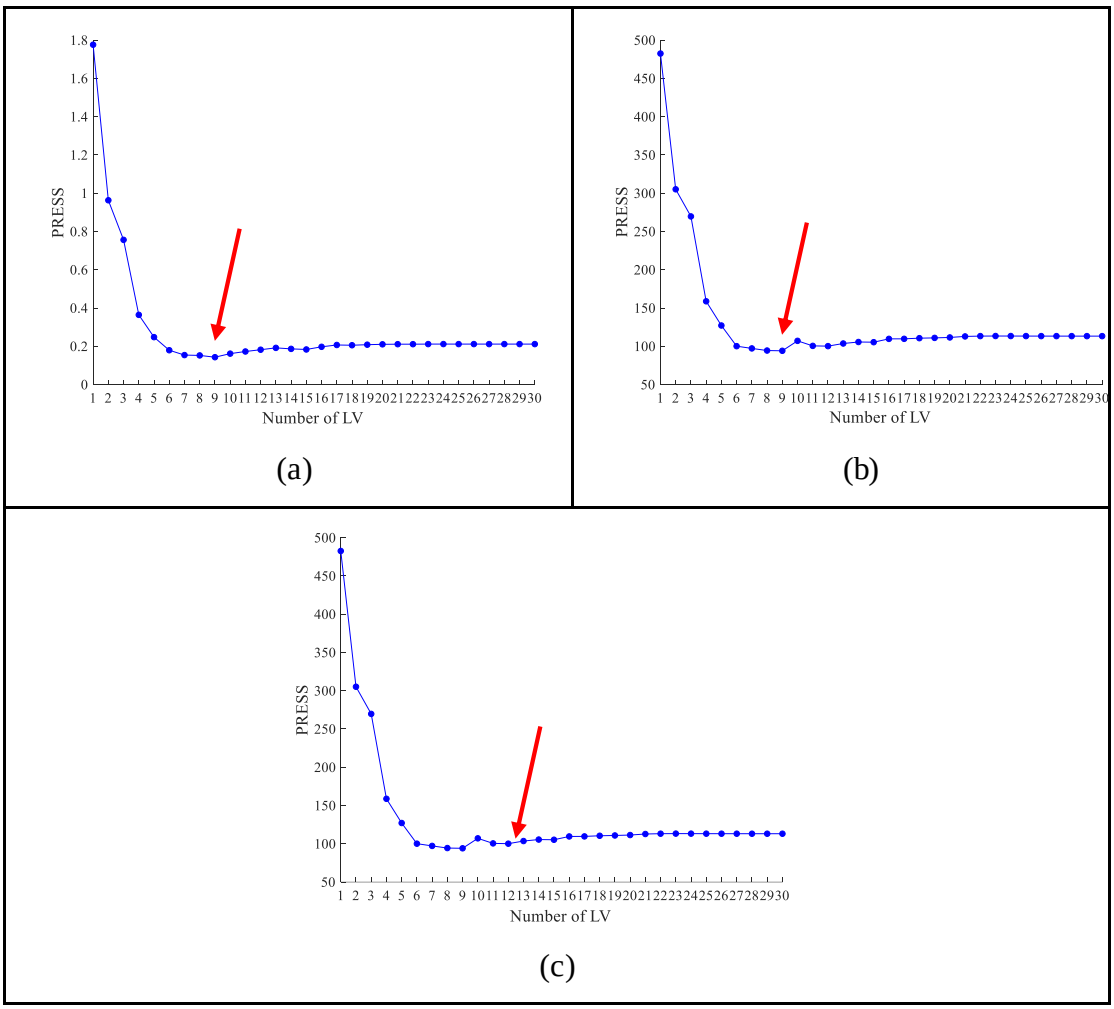


Figure 5.12. Number of LVs vs. PRESS plot for selecting the optimal number of LVs a) Benzene b) C7 plus c) C6 minus.

By using Figure 5.12, benzene, C7plus, and C6minus PLSR models with 9 LVs selected. Figure 5.13 present the reference values obtained from GC analysis versus predicted values obtained from PLS model. Standard error of calibration calculated and standard error of calculated for each parameter.

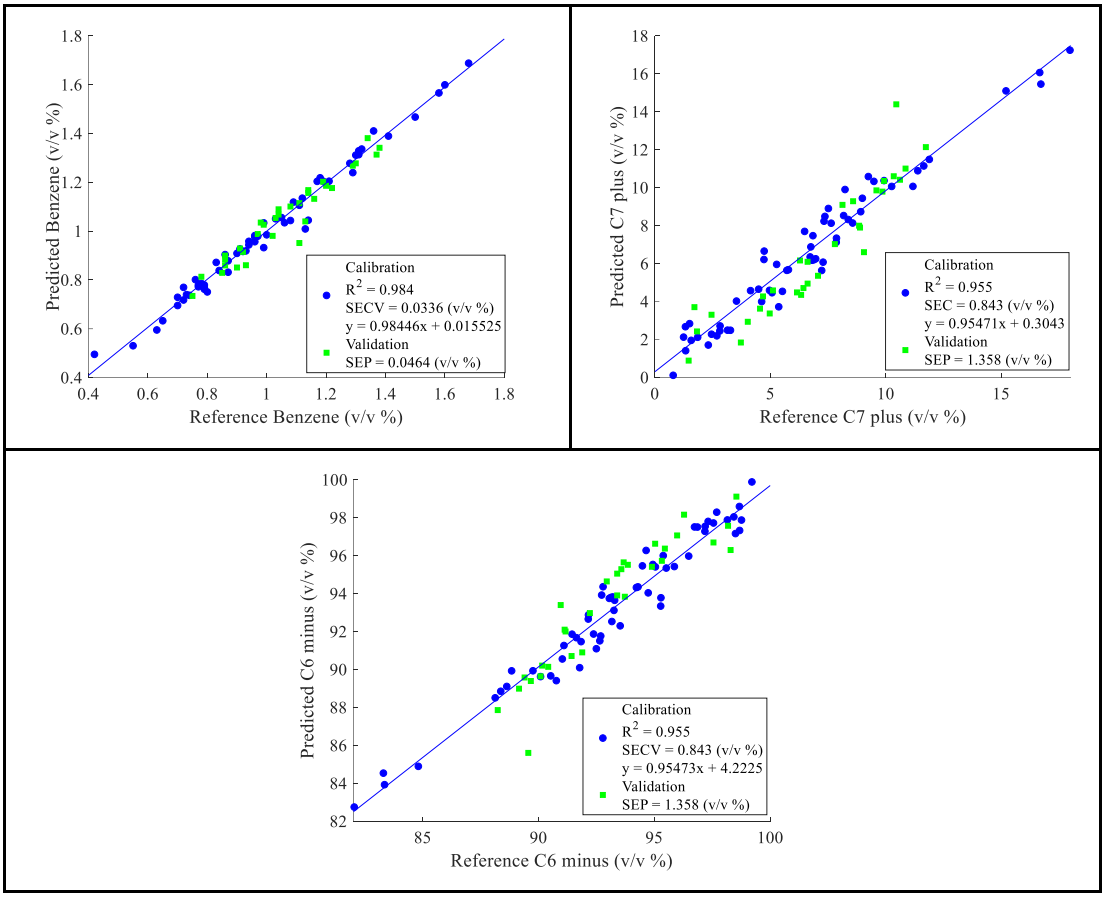


Figure 5.13. Actual concentrations vs. PLS predicted concentrations; Benzene, C7 plus, C6 minus

As seen in Figure 5.13, the model performance is quite close for calibration and validation set predictions with calibration performance being only slightly better indicating no significant overfitting. For Benzene, SECV and SEP values are found to be 0.0336 (v/v %) and 0.0464 (v/v %), respectively. The R^2 value for calibration set predictions are calculated as 0.984, and the R^2 value for the validation set is 0.930. The residuals for all samples are plotted in Figure 4.14 a. While most of the residuals are in the range of ±0.1 (v/v %). For C7 plus, SECV and SEP values are found to be 0.843 (v/v %) and 1.358 (v/v %), respectively. The R^2 value for calibration set predictions are calculated as 0.955, and the R^2 value for the validation set is 0.854. The residuals for all sample are plotted in Figure 4.14 b. The residuals are in the range of ± 2.0 (v/v %). For C6 minus, SECV and SEP values are found to be 0.843 (v/v %) and 1.358 (v/v %), respectively. The R^2 value for calibration set predictions are calculated as 0.955, and the R^2 value for the validation set is 0.855. The residuals for all sample are plotted in Figure 4.14 c. The most residuals are in the range of ± 3 (v/v %). The residuals corresponding to each concentration can be drawn in residuals plot for additional comments on the model.

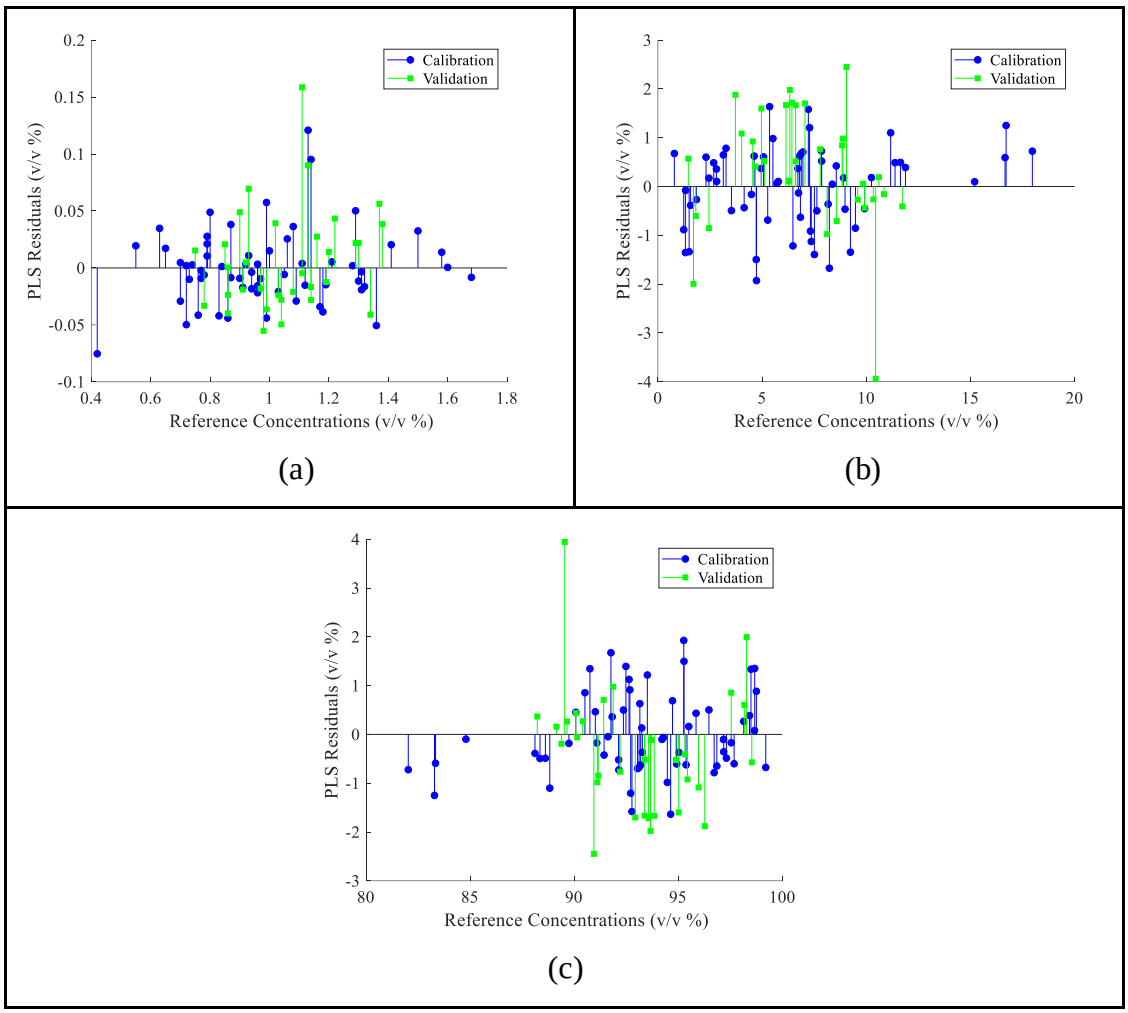


Figure 5.14. Reference concentrations vs. corresponding PLSR prediction residuals a) Benzene b) C7 plus c) C6 minus

As seen in Figure 5.14. all validation data are very close to calibration data which shows model prediction efficiency. However, it is seen that residual differences are higher in a few validation samples than others. It has been observed that this does not follow any pattern. It is thought that the NIR spectra of these samples are due to the composition change in the sample due to evaporation.

5.4.2. Heavy Straight Run Naphtha Results

For finding the best fitting number of LVs, PRESS values were calculated for the first 30 LVs and the results are given in Figure 5.15.

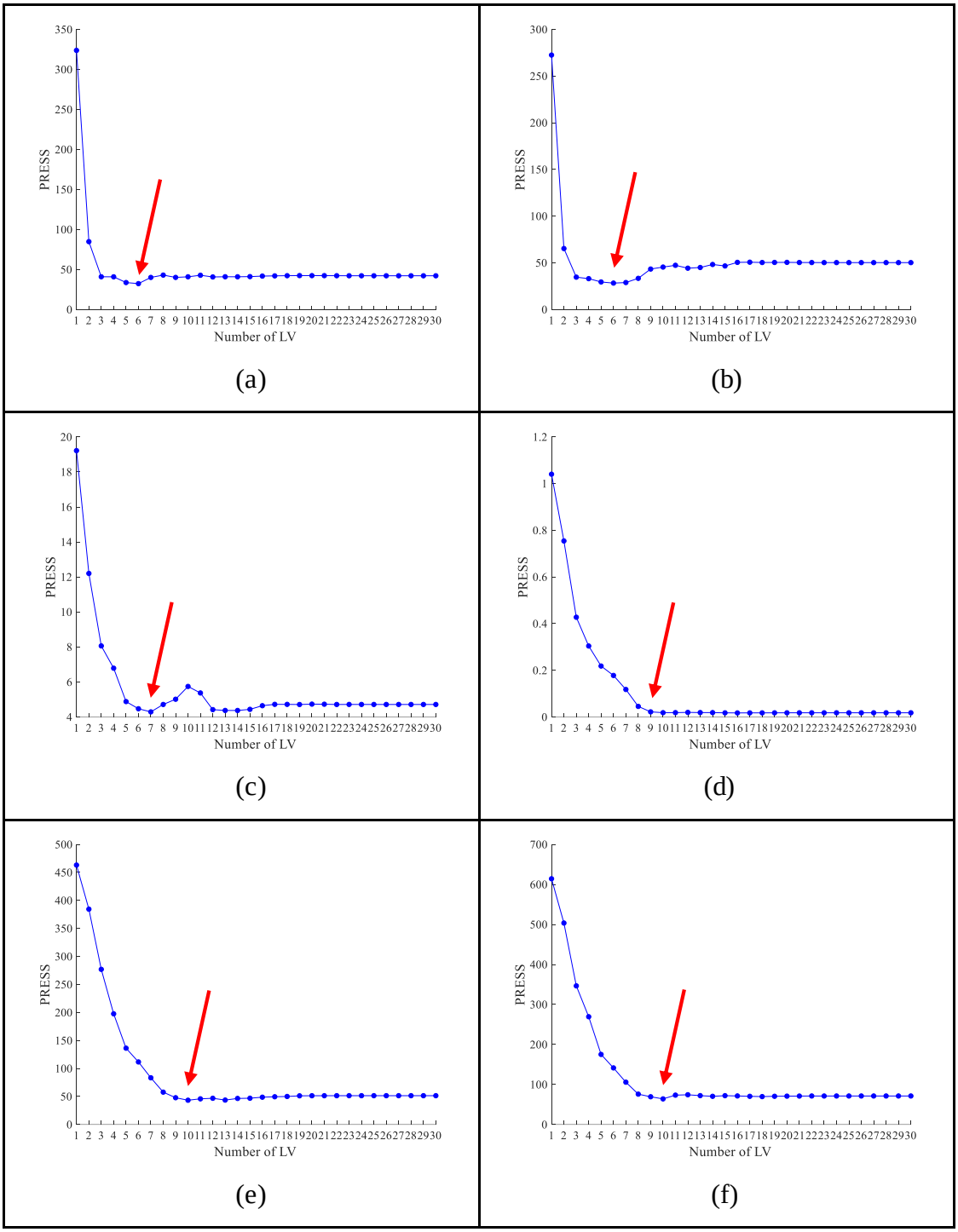


Figure 5.15. Number of LVs vs. PRESS plot for selecting the optimal number of LVs a) Paraffins b) Naphthenes c) Aromatics d) Benzene e) C7 plus f) C6 minus

By using Figure 5.15, Paraffins, Naphthenes, Aromatics, benzene, C7 plus and C6 minus PLSR model with 6 LVs, 6 LVs,7 LVs, 9 LVs, 10 LVs and 10 LVs selected respectively.

Figure 5.16 present the reference values obtained from GC analysis versus predicted values obtained from PLS model. Standard error of calibration calculated and standard error of calculated for each parameter.

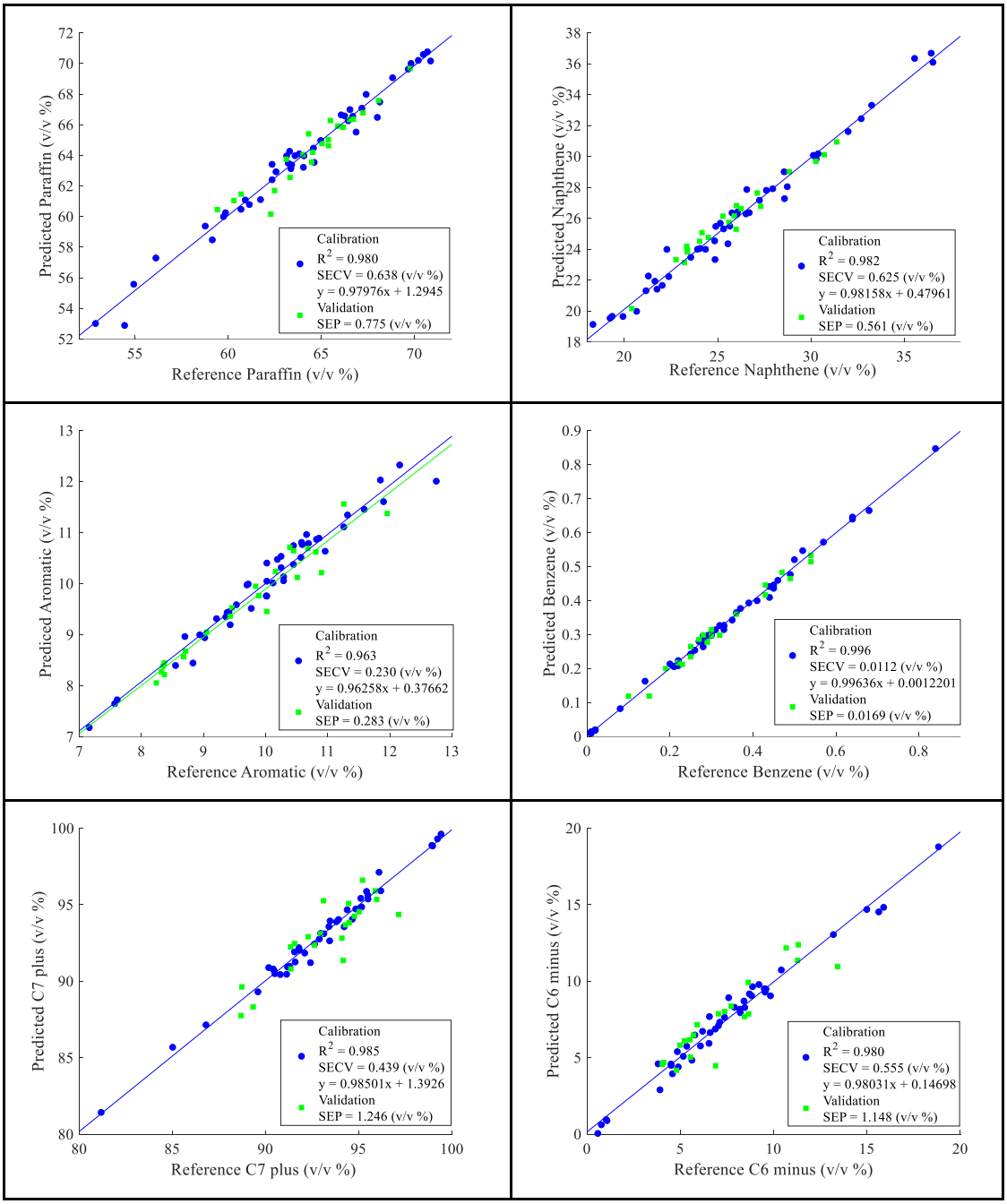
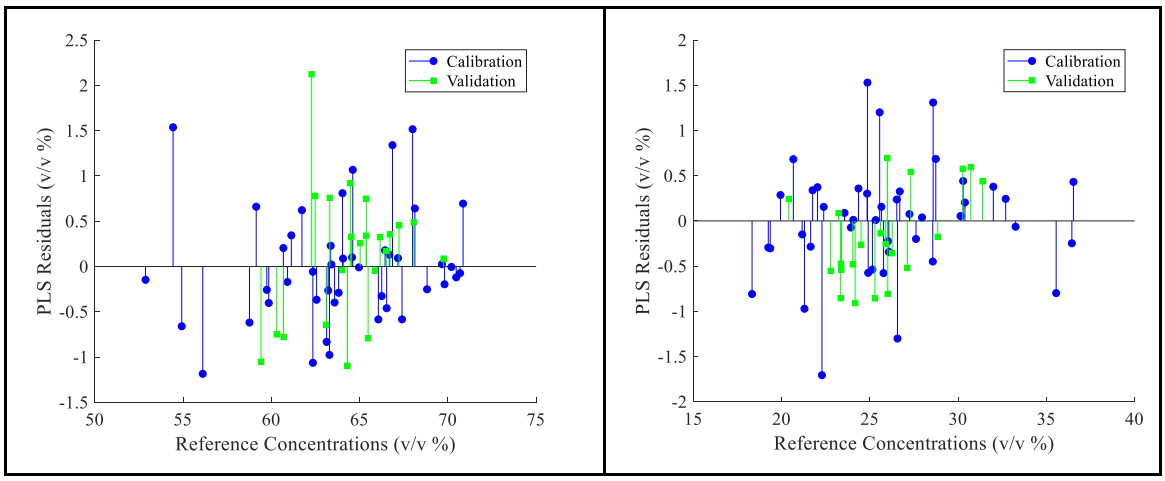


Figure 5.16. Actual concentrations vs. PLS predicted concentrations Paraffins, Naphthenes, Aromatics, Benzene, C7 plus, C6 minus

As seen in Figure 5.16, the model performance is quite close for calibration and validation set predictions with calibration performance being only slightly better indicating no significant overfitting. For Paraffins, SECV and SEP values are found to be

0.638 (v/v %) and 0.775 (v/v %), respectively. The R^2 value for calibration set predictions are calculated as 0.980, and the R^2 value for the validation set is 0.909. The residuals for all samples are plotted in Figure 5.17 a. While most of the residuals are in the range of ±1.5 (v/v %). For Naphthenes, SECV and SEP values are found to be 0.625 (v/v %) and 0.561 (v/v %), respectively. The R^2 value for calibration set predictions are calculated as 0.982, and the R^2 value for the validation set is 0.969. The residuals for all samples are plotted in Figure 5.17 b. The residuals are in the range of ± 2.0 (v/v %). For Aromatics, SECV and SEP values are found to be 0.230 (v/v %) and 0.283 (v/v %), respectively. The R^2 value for calibration set predictions are calculated as 0.963, and the R^2 value for the validation set is 0.940. The residuals for all samples are plotted in Figure 5.17 c. Most residuals are in the range of ± 0.6 (v/v %). For Benzene, SECV and SEP values are found to be 0.0112 (v/v %) and 0.0169 (v/v %), respectively. The R^2 value for calibration set predictions are calculated as 0.996, and the R^2 value for the validation set is 0.982. The residuals for all samples are plotted in Figure 5.17 d. While most of the residuals are in the range of ±0.03 (v/v %). For C7 plus, SECV and SEP values are found to be 0.555 (v/v %) and 1.148 (v/v %), respectively. The R^2 value for calibration set predictions are calculated as 0.980, and the R^2 value for the validation set is 0.825. The residuals for all samples are plotted in Figure 5.18 e. The residuals are in the range of ± 2.0 (v/v %). For C6 minus, SECV and SEP values are found to be 0.439 (v/v %) and 1.246 (v/v %), respectively. The R^2 value for calibration set predictions are calculated as 0.985, and the R^2 value for the validation set is 0.752. The residuals for all sample are plotted in Figure 5.17 e. The most residuals are in the range of ± 3 (v/v %). The residuals for the PLS model for HSRN samples are given in Figure 5.17.



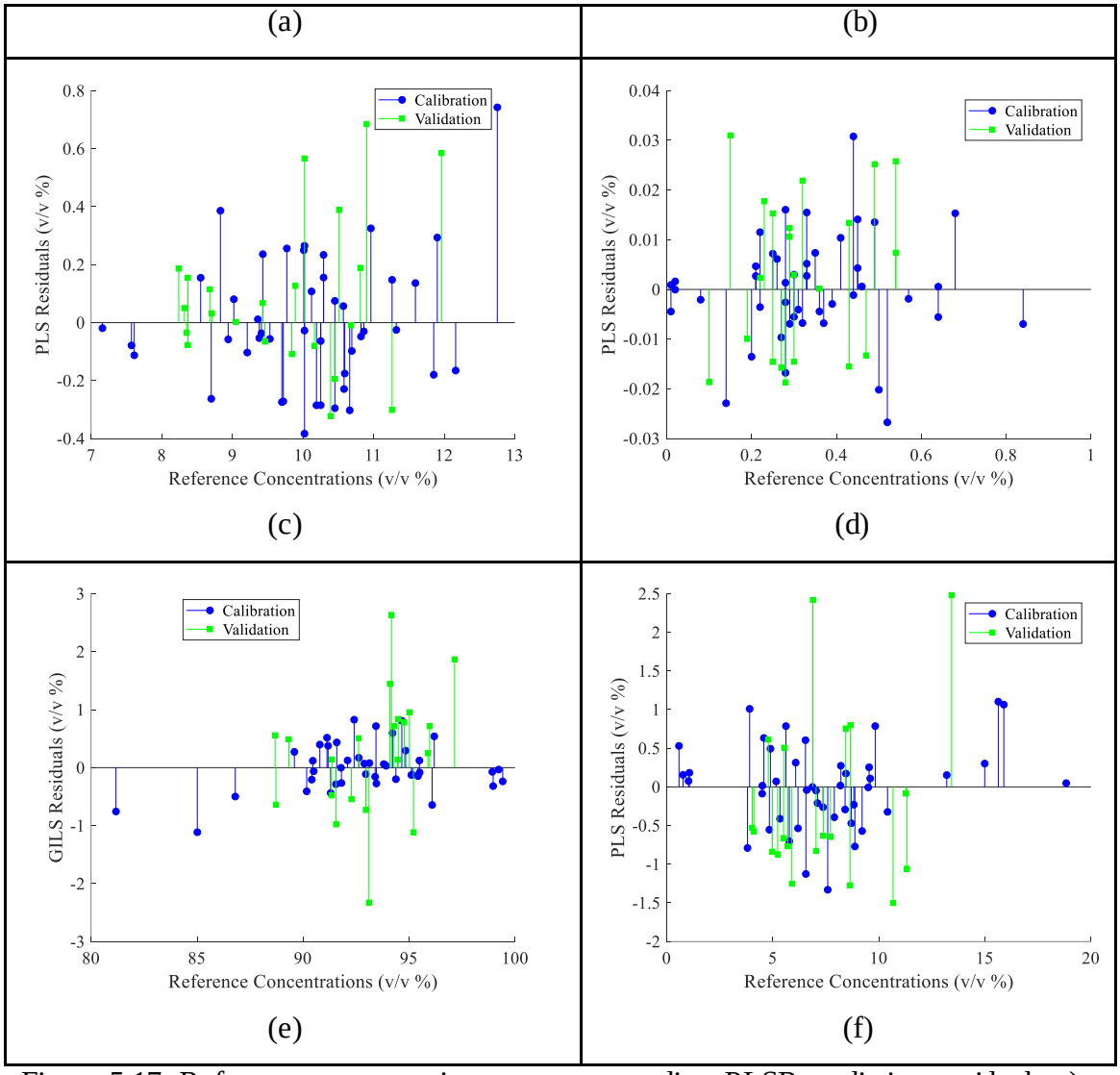


Figure 5.17. Reference concentrations vs. corresponding PLSR prediction residuals a) Paraffins b) Naphthenes c) Aromatics d) Benzene e) C7 plus f) C6 minus

As Shown Figure 5.17 for both validation and calibration sets. All validation data are very close to calibration data which shows model prediction efficiency.

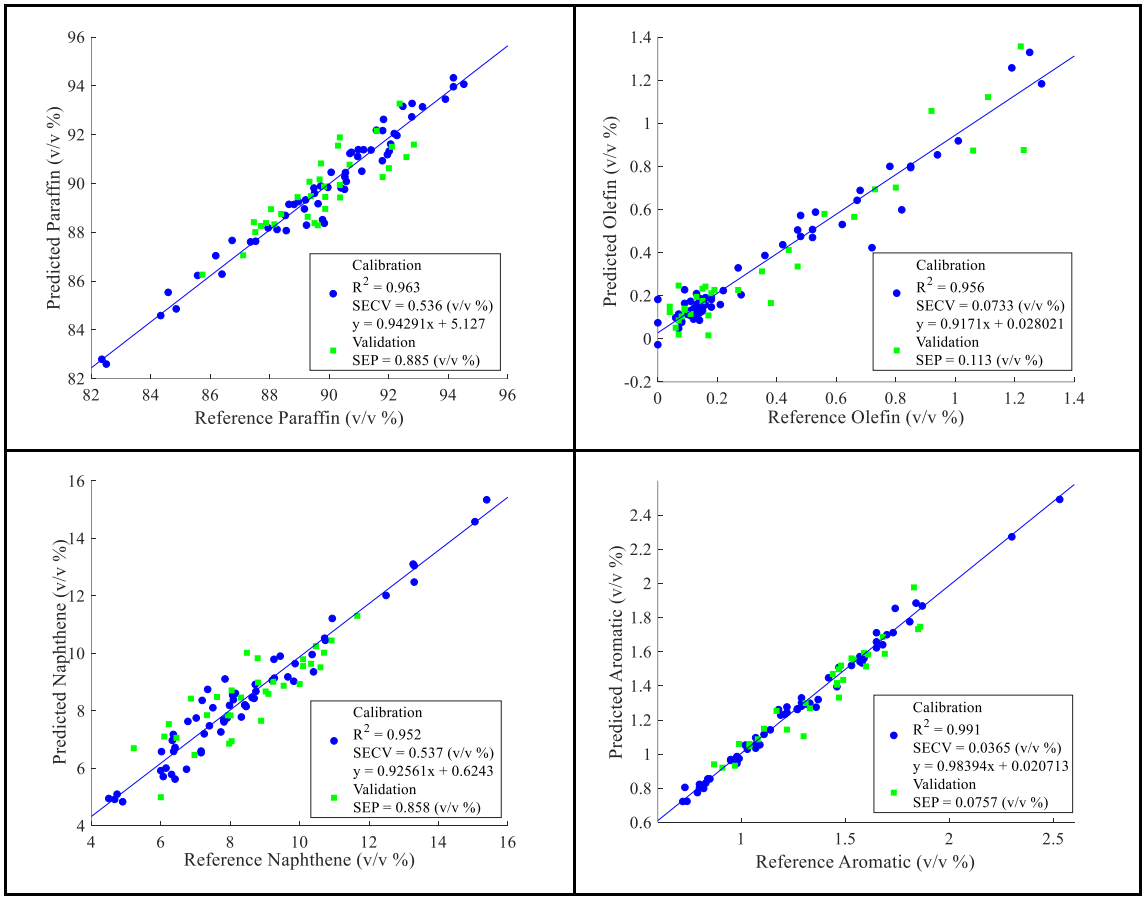
5.5. Genetic Inverse Least Square (GILS)

After pre-processing raw NIR spectra, Genetic Inverse Least Square (GILS) with 30 genes, 50 iterations, and 100 runs where R^2 threshold for selection of initial genes were 0.5 and 1-fold CV was used for determination of fitness. was applied to establish prediction models both for LSRN and HSRN, models were created as Naphthenes, Paraffins, Olefins, Benzene, Aromatics, C7Plus (the sum of compounds with more than

7 carbons) and C6Minus (the sum of compounds with less than 6 carbons). A modeling study has not been done for the Olefins parameter in HSRN due to a lack of data. The prediction performances of the created models were evaluated by looking at the coefficient of determination (R^2) of the calibration data set, the root mean square error of calibration (SEC), and the root mean square of validation errors (SEP) data. Models with low SEC, SECV, and SEP values and high determination coefficients were preferred at the stage of selecting the basic component numbers of the established models.

5.5.1. Light Straight Run Naphtha Results

Reference values obtained from GC analysis vs GILS model predicted values of FT-NIR spectra of LSRN samples treated by EMSC of each parameter are given in Figure 5.18.



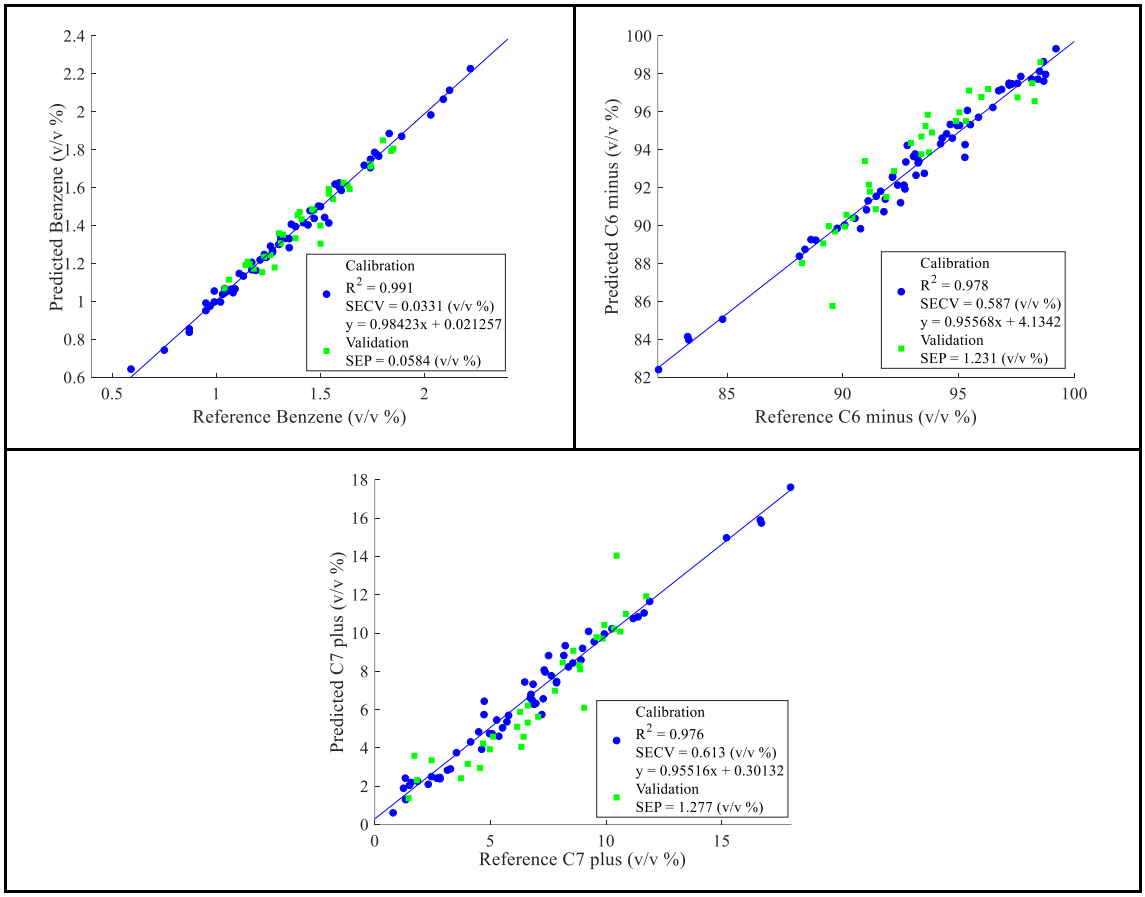
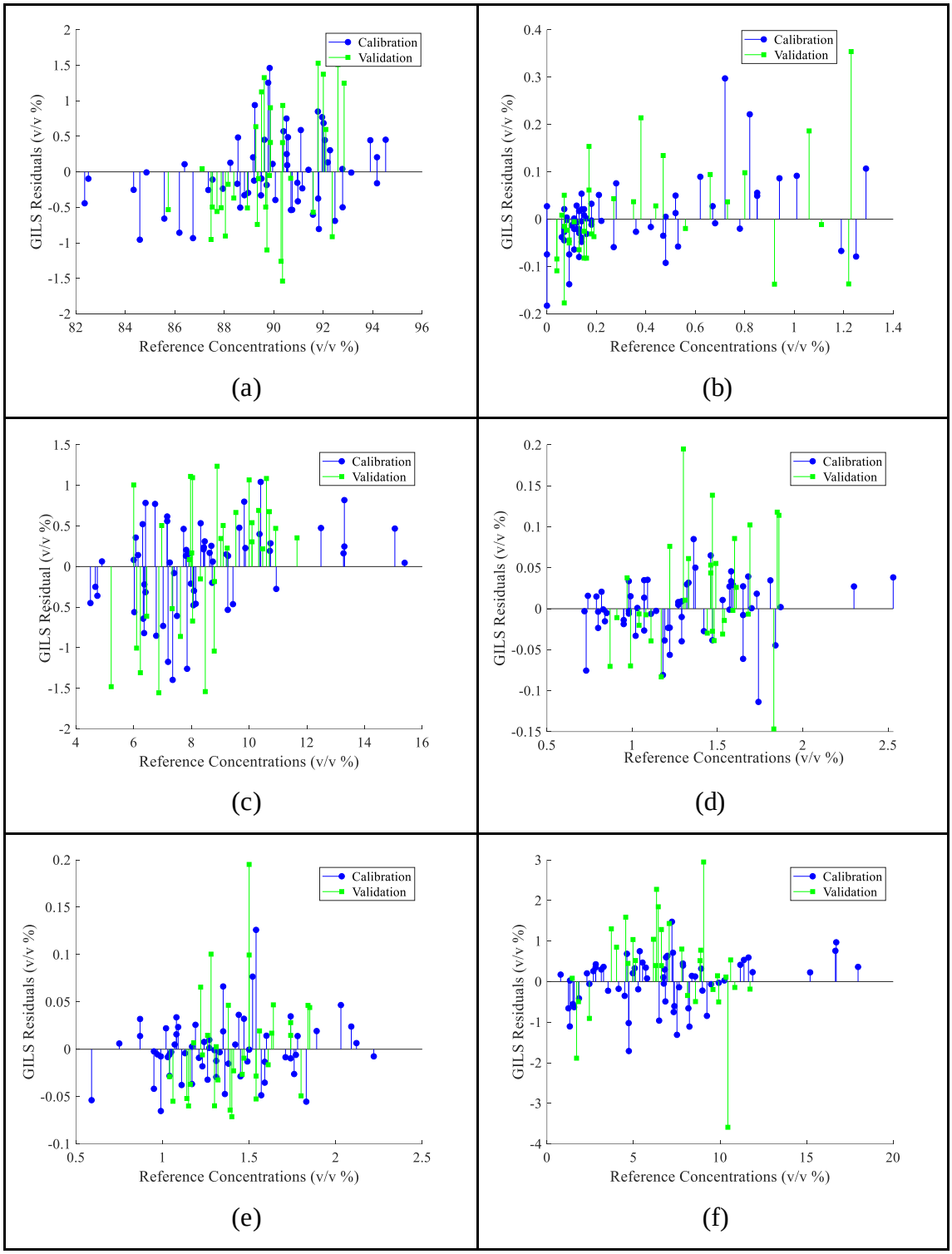


Figure 5.18. Actual concentrations vs. GILS predicted concentrations; Paraffins, Olefins, Naphthenes, Aromatics, Benzene, C7 plus, C6 minus

As seen in Figure 5.18, the model performance is quite close for calibration and validation set predictions with calibration performance being only slightly better indicating no significant overfitting. Therefore, ILS method can solves the overfitting problem with Genetic Algorithm. SECV and SEP values are given in Figure 5.18.

Generally, PLS result better than GILS result but for C6 minus and C7 plus Parameters are lower SEP value compare to PLS result.

In order to determine the error range and possible residual trends, the residue plot is given in Figure 5.19.


Calibration
Validation
GILS Residuals (v/v %)
Reference Concentrations (v/v %)
(a)
Calibration
Validation
GILS Residuals (v/v %)
Reference Concentrations (v/v %)
(b)
Calibration
Validation
GILS Residual (v/v %)
Reference Concentrations (v/v %)
(c)
Calibration
Validation
GILS Residuals (v/v %)
Reference Concentrations (v/v %)
(d)
Calibration
Validation
GILS Residuals (v/v %)
Reference Concentrations (v/v %)
(e)
Calibration
Validation
GILS Residuals (v/v %)
Reference Concentrations (v/v %)
(f)

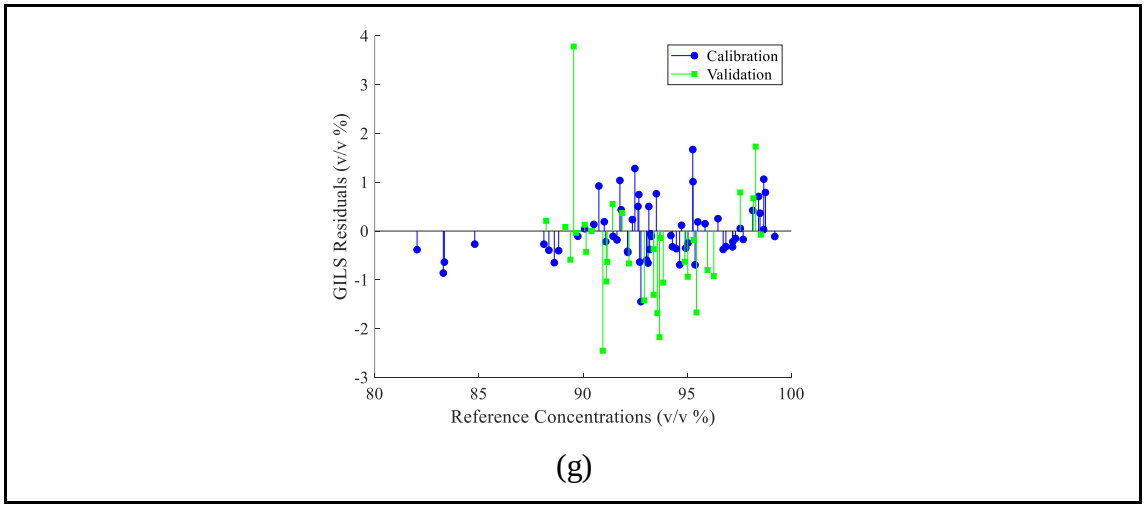
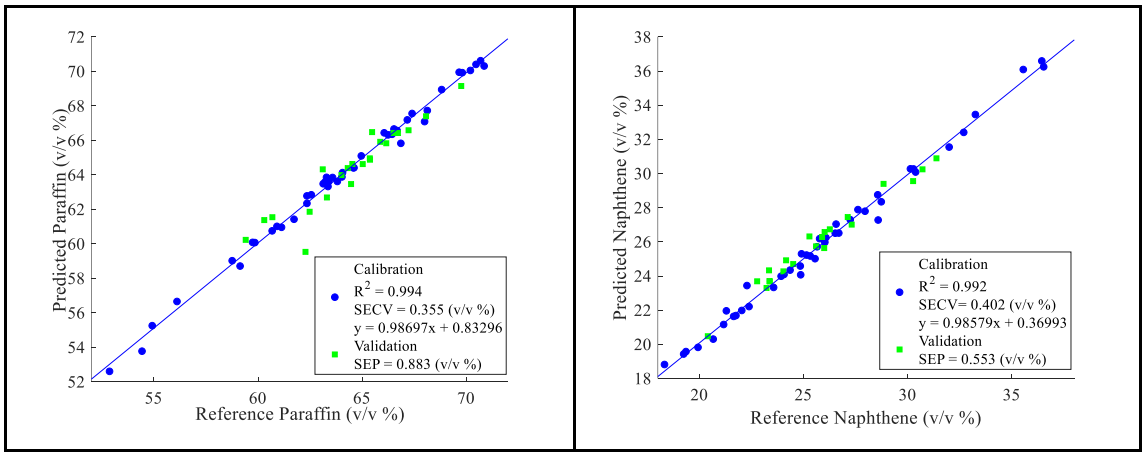


Figure 5.19. Reference concentrations vs. corresponding GILS prediction residuals a) Paraffins b) Olefins c) Naphthenes d) Aromatics e) Benzene f) C7 plus g) C6 minus

As Shown Figure 5.19 for both validation and calibration sets. All validation data are very close to calibration data which shows model prediction efficiency.

5.5.2. Heavy Straight Run Naphtha Results

Figure 5.20 present the reference values obtained from GC analysis versus predicted values obtained from PLS model. Standard error of calibration calculated and standard error of calculated for each parameter.



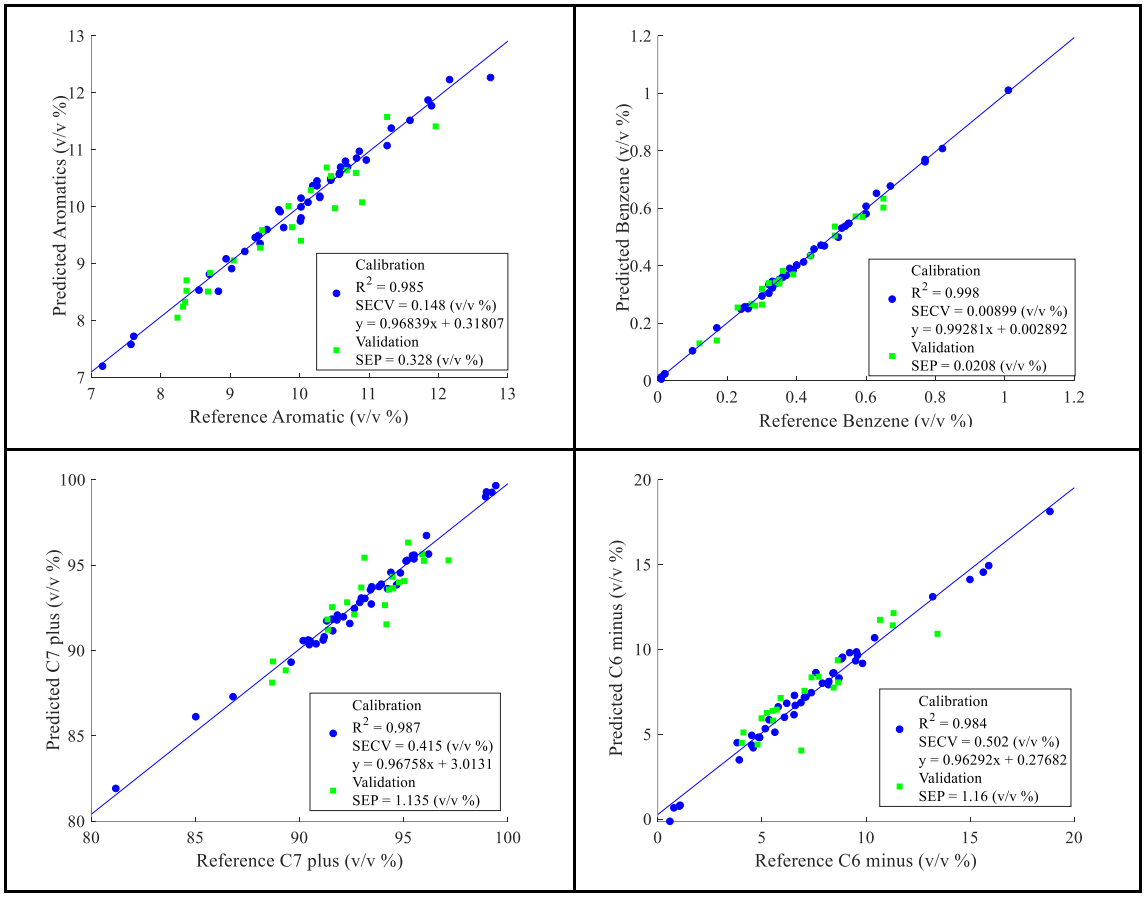


Figure 5.20. Actual concentrations vs. GILS predicted concentrations; Paraffins, Naphthenes, Aromatics, Benzene, C7 plus, C6 minus

The overall predictive performance of HSRN samples is quite well compare to LSRN samples. Since LSRN samples are lighter and have a lower boiling point than HSRN samples, LSRN samples may have affected the prediction ability due to evaporation during analysis. It may be caused by evaporation of samples when performing FT-NIR analysis after reference analysis of samples.

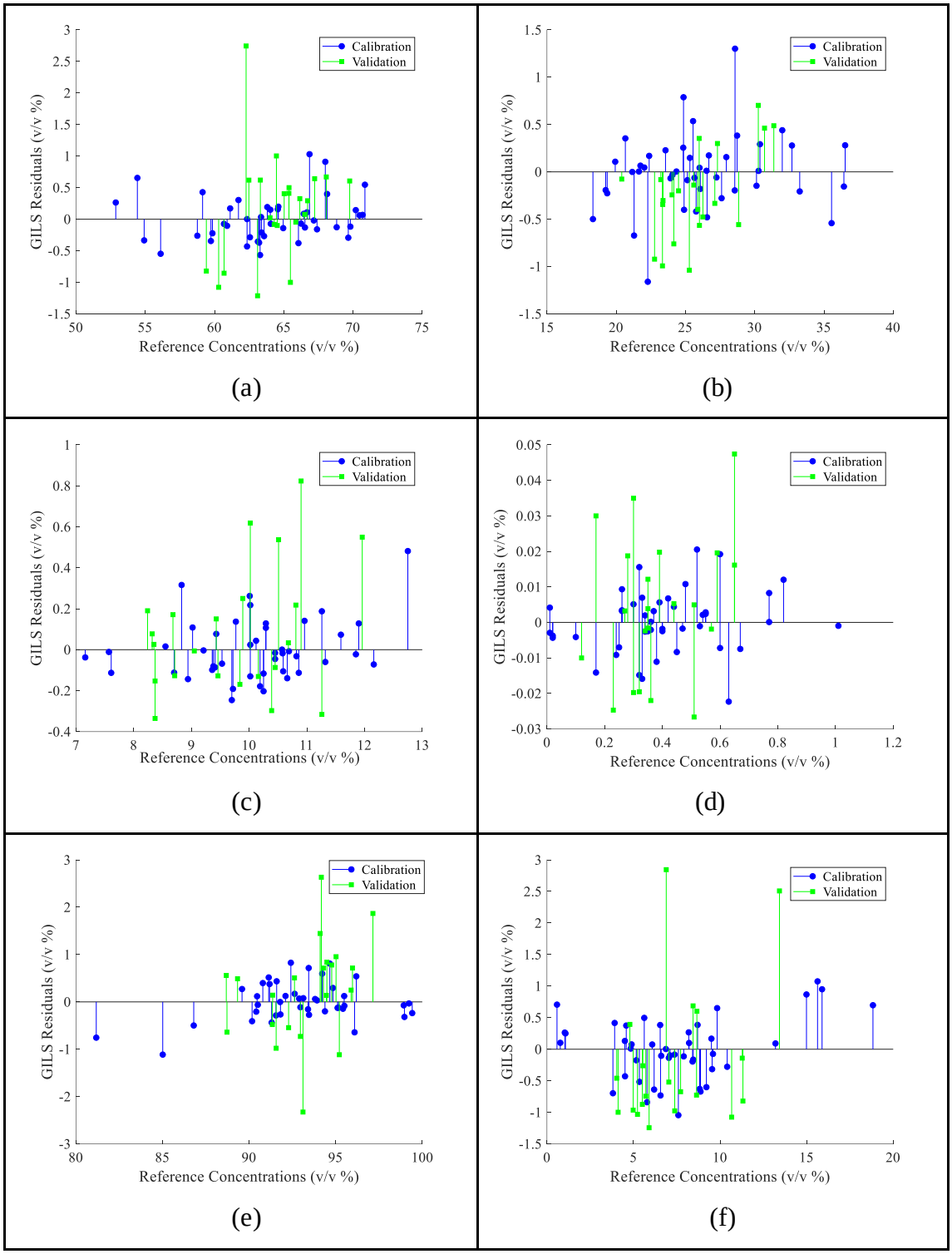


Figure 5.21. Reference concentrations vs. corresponding GILS prediction residuals a) Paraffins b) Naphthenes c) Aromatics d) Benzene e) C7 plus f) C6 minus

No visible residual pattern was observed in Figure 5.21. In fact, validation residuals are relatively higher than calibration residuals. This is actually an expected situation and the validation estimate does not get better than the calibration model.

5.6. Summary and Comparison of the Calibration Models

As a result of the predicted values, according to PLS and GILS calibration models with SECV (v/v %), SEP (v/v %), and R^2 values given in Tables 1 and 2 for LSRN and HSRN samples. According to results show SECV and SEP values indicating no severe overfitting. Both the PLS and GIRLS results are promising for all parameters. In all cases, using one of the 2 models can give fast and effective results. However, choosing the best model for each parameter gives even better results. The prediction performances of the created models were evaluated by looking at the coefficient of determination (R^2) of the calibration and validation data set, the root mean square error of calibration (SEC), and the root mean square of validation errors (SEP) data. Models with low SECV and SEP values and high determination coefficients were preferred at the stage of selecting the basic component numbers of the established models. For LSRN, The PLS model is better at predicting Paraffins, Olefins, Aromatics, and Benzene values than the GILS model. Also, in narrow range parameters, the PLS model gives a more successful result. GILS model is better at predicting Naphthenes, C7 plus, and C6 minus values than the PLS model.

Table 5.1. Summary of Values for LSRN

CDU	PLS Calibration Model Results				GILS Calibration Model Results			Data Range		
LSRN	SECV (v/v %)	SEP (v/v %)	R^2	Number of LVs	SECV (v/v %)	SEP (v/v %)	R^2	Max	Min	Interval
Paraffin	**0.774**	**0.876**	**0.923**	6	0.536	0.885	0.963	94.52	82.35	12.17
Olefin	**0.098**	**0.107**	**0.917**	9	0.073	0.112	0.956	1.29	0	1.29
Naphthene	0.719	0.924	0.912	6	**0.537**	**0.859**	**0.952**	15.39	4.5	10.89
Aromatic	**0.019**	**0.072**	**0.998**	6	0.036	0.076	0.991	2.53	0.72	1.81
Benzene	**0.034**	**0.046**	**0.984**	9	0.033	0.058	0.991	1.68	0.42	1.26
C7 plus	0.843	1.358	0.955	9	**0.613**	**1.277**	**0.976**	17.97	0.79	17.18
C6 minus	0.843	1.358	0.955	9	**0.587**	**1.231**	**0.979**	99.21	82.03	17.18

For HSRN, The PLS model is better predicting ability for Paraffins, Aromatics, Benzene, and C6 minus values. The GILS model can predict very good for Naphthenes and C7 plus. At the same time, the best model method chosen for the mentioned parameter is shown in bold.

Table 5.2 Summary of Values for HSRN

CDU	PLS Calibration Model Results				GILS Calibration Model Results			Data Range		
HSRN	**SECV (v/v %)**	**SEP (v/v %)**	**R^2**	**Number of LVs**	**SECV (v/v %)**	**SEP (v/v %)**	**R^2**	**Max**	**Min**	**Interval**
Paraffin	**0.638**	**0.775**	**0.980**	6	0.355	0.883	0.994	70.86	52.87	17.99
Naphthene	0.625	0.561	0.982	6	**0.402**	**0.553**	**0.992**	36.53	18.32	18.21
Aromatic	**0.230**	**0.283**	**0.963**	7	0.148	0.329	0.985	12.75	7.16	5.59
Benzene	**0.011**	**0.017**	**0.996**	9	0.009	0.021	0.998	1.01	0.01	1
C7 plus	0.439	1.246	0.985	10	**0.416**	**1.135**	**0.987**	99.42	81.17	18.25
C6 minus	**0.555**	**1.148**	**0.980**	10	0.502	1.160	0.984	18.83	0.58	18.25

A further study will help to see the usage information of the two (PLS and GILS) models, by increase the number of the calibration sample set, by changing the preprocessing methods, by utilizing a variable selection algorithm.

CHAPTER 6

CONCLUSION

As a result of the studies, the results of the new analysis method developed by using near-infrared (NIR) spectroscopy and chemometrics models for the estimation of the physical properties of naphtha samples show that the laboratory can be used instead of classical methods. All naphtha samples were collected at Tupras Izmıt Refinery. The naphtha samples were analyzed at the Tupras Izmit Refinery Quality Control Laboratory by the reference test methods. NIR when empowered with chemometrics tools which are PLS or GILS, gives a fast and effortless way of quantitatively determining naphtha parameters.

All of the modeling studies conducted have obtained very successful results. However, the most suitable model was chosen for each parameter. It was decided that the PLS model gave better estimation results for Paraffins, Olefins, Aromatics, and benzene parameters in LSRN samples. At the same time, despite the narrow data range of these parameters, PLS yielded very successful results. On the other hand, in GILS models, it is seen that Naphthenes, C7 plus, and c6 minus parameters have better prediction ability than PLS. It was concluded that the predictions of PLS models for the Paraffins, Aromatics, benzene, and c6 minus parameters in the heavier layer HSRN samples were more successful. On the other hand, GILS models offer better estimates in naphtha and c7 plus parameters than PLS. Naphthenic molecules, which are among the LSRN and HSRN parameters, have predicted the straight-chain 5 to 10 carbon structure better than the PLS algorithm quite successfully.

The industrial applications of chemometrics modeling approaches, which are widely used in the chemical industries in the world, are not seen much in our country. Within the scope of this project, with the fact that chemometrics model development based on molecular spectroscopic data will be applied for Tupras, the capabilities in statistical experimental design, spectroscopy, and chemometrics multivariate analysis methods have been developed. By collecting more samples in the future, the models that will be obtained may become stronger and more reliable. In addition, it seems that these studies can be worked with different parameters or with a different crude oil layer.

REFERENCES

1. Organization of the petroleum exporting countries World oil outlook 2045. https://woo.opec.org/ (accessed Dec 25, 2020).

2. Andrade, D. F.; Azevedo, D. A.; Troise, M. H. F.; Tristão, M. L.; Miranda, J. L.; D'Elia, E., Comparison of UOP-326, voltammetric and gas chromatographic/mass spectrometric methods for the determination of conjugated dienes in Brazilian Naphtha. *Fuel* 2006, *85* (7-8), 1024-1031.

3. Di Corcia, A.; Samperi, R.; Capponi, G., Gas chromatographic analysis of gasoline and pure naphtha using packed columns. *Journal of Chromatography A* 1978, *160* (1), 147-154.

4. Vendeuvre, C.; Bertoncini, F.; Espinat, D.; Thiébaut, D.; Hennion, M.-C., Multidimensional gas chromatography for the detailed PIONA analysis of heavy naphtha: Hyphenation of an olefin trap to comprehensive two-dimensional gas chromatography. *Journal of Chromatography A* 2005, *1090* (1-2), 116-125.

5. ASTM, D., 5134 98: Standard Test Method for Detailed Analysis of Petroleum Naphthas through n Nonane by Capillary Gas Chromatography. *1993 Annual Book of ASTM Standards* 1998, *5*, 03.

6. Macho, S.; Larrechi, M., Near-infrared spectroscopy and multivariate calibration for the quantitative determination of certain properties in the petrochemical industry. *TrAC Trends in Analytical Chemistry* 2002, *21* (12), 799-806.

7. Reboucas, M. V.; dos Santos, J. B.; Domingos, D.; Massa, A. R. C., Near-infrared spectroscopic prediction of chemical composition of a series of petrochemical process streams for aromatics production. *Vibrational Spectroscopy* 2010, *52* (1), 97-102.

8. 구민식; 정호일; 이준식, Rapid compositional analysis of naphtha by near-infrared spectroscopy. *Bulletin of the Korean Chemical Society* 1998, *19* (11), 1189-1193.

9. Ku, M.-S.; Chung, H., Comparison of near-infrared and Raman spectroscopy for the determination of chemical and physical properties of naphtha. *Applied spectroscopy* 1999, *53* (5), 557-564.

10. da Silva, V. H.; Reboucas, M. V.; Salles, A. R.; Pimentel, M. F.; Pontes, M. J. C.; Pasquini, C., Determination of naphtha composition by near infrared spectroscopy and multivariate regression to control steam cracker processes. *Fuel Processing Technology* 2015, *131*, 230-237.

11. Balabin, R. M.; Safieva, R. Z.; Lomakina, E. I., Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques. *Analytica Chimica Acta* 2010, *671* (1-2), 27-35.

12. Özdemir, D., Determination of octane number of gasoline using near infrared spectroscopy and genetic multivariate calibration methods. *Petroleum science and technology* 2005, *23* (9-10), 1139-1152.

13. Zhu, L.; Lu, S. H.; Zhang, Y. H.; Zhai, H. L.; Yin, B.; Mi, J. Y., An effective and rapid approach to predict molecular composition of naphtha based on raw NIR spectra. *Vibrational Spectroscopy* 2020, 103071.

14. Pasquini, C., Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. *Journal of the Brazilian chemical society* 2003, *14* (2), 198-219.

15. Stuart, B., Experimental methods. *Infrared spectroscopy: fundamentals and applications* 2004, 18-19.

16. Weyer, L.; Lo, S. C., Spectra–structure correlations in the near-infrared. *Handbook of vibrational spectroscopy* 2006.

17. Wold, S., Chemometrics; what do we mean with it, and what do we want from it? *Chemometrics and Intelligent Laboratory Systems* 1995, *30* (1), 109-115.

18. Skoog, D.; Holler, F.; Nieman, D., Principle of Instrumental Analysis, Reprint. *Thomson Brooks/Colepublication* 2004, 300-351.

19. Tobias, R. D., Chemometrics: a practical guide. Taylor & Francis: 1999.

20. Brereton, R. G., Introduction to multivariate calibration in analytical chemistry. *Analyst* 2000, *125* (11), 2125-2154.

21. Geladi, P., Chemometrics in spectroscopy. Part 1. Classical chemometrics. *Spectrochimica Acta Part B: Atomic Spectroscopy* 2003, *58* (5), 767-782.

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

23. Haaland, D. M.; Thomas, E. V., Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical chemistry* 1988, *60* (11), 1193-1202.

24. Özdemir, D.; Öztürk, B., Genetic multivariate calibration methods for near infrared (NIR) spectroscopic determination of complex mixtures. *Turkish Journal of Chemistry* 2004, *28* (4), 497-514.

25. Gilbert, R. J.; Goodacre, R.; Woodward, A. M.; Kell, D. B., Genetic programming: A novel method for the quantitative analysis of pyrolysis mass spectral data. *Analytical Chemistry* 1997, *69* (21), 4381-4389.

26. Li, T.-H.; Lucasius, C. B.; Kateman, G., Optimization of calibration data with the dynamic genetic algorithm. *Analytica Chimica Acta* 1992, *268* (1), 123-134.

27. Lucasius, C. B.; Kateman, G., Genetic algorithms for large-scale optimization in chemometrics: an application. *TrAC Trends in Analytical Chemistry* 1991, *10* (8), 254-261.

28. Özdemir, D.; Dinc, E., Determination of thiamine HCl and pyridoxine HCl in pharmaceutical preparations using uv–visible spectrophotometry and genetic algorithm based multivariate calibration methods. *Chemical and pharmaceutical bulletin* 2004, *52* (7), 810-817.

29. Akkoç, G. D. Development of chemometrics calibration toolbox and its application for determination of slep adulteration. Izmir Institute of Technology, 2018.

30. Afseth, N. K.; Kohler, A., Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemometrics and Intelligent Laboratory Systems* 2012, *117*, 92-99.

Url-1

<https://chem.libretexts.org/Bookshelves/Organic_Chemistry/Map%3A_Organic_Chemistry_(Vollhardt_and_Schore)/03._Reactions_of_Alkanes%3A_Bond-Dissociation_Energies_Radical_Halogenation_and_Relative_Reactivity/3-03_Conversion_of_Petroleum%3A_Pyrolysis> (accessed Dec 25, 2020).

Url-2 <https://en.wikipedia.org/wiki/Gas_chromatography> (accessed Dec 25, 2020).

Url-3 <https://applications.wasson-ece.com/?p=351> (accessed Dec 25, 2020).

Url-4 <https://en.wikipedia.org/wiki/Near-infrared_spectroscopy> (accessed Dec 25, 2020).