DEVELOPMENT OF CHEMOMETRIC MULTIVARIATE CALIBRATION MODELS FOR SPECTROSCOPIC QUALITY ANALYSIS OF BIODIESEL BLENDS

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
Murat BAĞCIOĞLU

July 2011
İZMİR

ACKNOWLEDGEMENTS

I would like to express my gratefully appreciation to my supervisor, Assoc. Prof. Durmuş ÖZDEMİR for his supervision and advice from the early stage of this research as well as giving me extraordinary experiences throughout the work. Without his guidance and persistent help, this dissertation would not have been possible.

Special thanks to Dr. Gökmen TAYFUR for participating as committee member and to Dr. Erol ŞEKER for reviewing my work as committee members with their kindly helps.

In addition, I would like gratefully to thank İbrahim KARAMAN for his constructive comments on this thesis and for being an exceptional collaborator.

I am pleased to pay my thanks to Technological Research Council of Turkey (TÜBİTAK) for funding the project (108T783) and providing scholarship; İzmir Institute of Technology (İYTE) for giving chance to work on this project.

I was delighted to interact with all my lab mates at the Chemometrics Research Laboratory who made it a warm place to work and thanks to all other folks in İYTE, had inspired me in research and life through our interactions during the long hours in the lab.

It is a pleasure to express my gratitude wholeheartedly to my family for their inseparable support and prayers. My mother, Kadriye, is the one who sincerely raised me with her caring and gently love also; my deepest appreciation goes to my father, Naci, who is always on my side with his spiritual presence. My older sister, İncigül, whose dedication, love and persistent confidence in me. I owe them for being unselfishly let their intelligence, passions, and ambitious collide with mine.

Eventually, I would like to thank everybody being supportive and caring who was important to successful realization of this thesis.

ABSTRACT

DEVELOPMENT OF CHEMOMETRIC MULTIVARIATE CALIBRATION MODELS FOR SPECTROSCOPIC QUALITY ANALYSIS OF BIODIESEL BLENDS

The fact that the biodiesel is produced from renewable resources and environmentally friendly when compared to the fossil-based petroleum diesel, biodiesel has gained an increasing interest. It is mainly produced from a variety of different animal fat and vegetable oil combined with an alcohol in the presence of a homogeneous catalyst and the determination of the quality of the produced biodiesel is as important as its production. Industrial scale biodiesel production plants have been adopted the chromatographic analysis protocols some of which are standard reference methods proposed by official bodies of the governments and international organizations. However, analysis of multi component mixtures by chromatographic procedures can become time consuming and may require a lot of chemical consumption. For this reason, as an alternative, spectroscopic methods combined with chemometrics offer several advantages over classical chromatographic procedures in terms of time and chemical consumption. With the immense development of computer technology and reliable fast spectrometers, new chemometric methods have been developed and opened up a new era for processing of complex spectral data.

In this study, laboratory scale produced biodiesel was mixed with methanol, commercial diesel and several different vegetable oils that are used to prepare biodiesels and then several different ternary mixture systems such as diesel-vegetable oil-biodiesel and methanol-vegetable oil-biodiesel were prepared and gas chromatographic analysis of these samples were performed. Then, near infrared (NIR) and mid infrared (FTIR) spectra of the same samples were collected and multivariate calibration models were constructed for each component for all the infrared spectroscopic techniques. Chemometric multivariate calibration models were proposed as genetic inverse least square (GILS) and artificial neural networks (ANN). The results indicate that determination of biodiesel blends quality with respect to chemometric modeling gives reasonable consequences when combined with infrared spectroscopic techniques.

ÖZET

BİYODİZEL KARIŞIMLARININ SPEKTROSKOPİK KALİTE ANALİZİ İÇİN KEMOMETRİK ÇOK DEĞİŞKENLİ KALİBRASYON MODELLERİ GELİŞTİRİLMESİ

Günümüzde, biyodizel gerek yenilenebilir kaynaklardan üretilmesi gerekse çevreci bir yakıt olması bakımından, diğer fosil yakıtlardan daha popüler hale gelmiştir. Çeşitli bitkisel ve hayvansal yağlarla, alkol ve homojen katalizör eşliğinde yürüyen bir tepkimenin ürünü olan biyodizelin kalitesinin tayini de gün geçtikçe önem kazanmaktadır. Endüstride bu amaç için yaygın olarak kromatografik yöntemler kullanılmaktadır. Ancak çok bileşenli karışımların kromatografik analizleri zaman ve kimyasal israfına yol açtığından, spektroskopik yöntemler zaman ve malzeme tasarrufu açısından daha avantajlı olmaktadır. Son yıllarda bilgisayar sektöründeki gelişmeler kemometrinin hızla gelişimini sağlamış ve spektral verilerin değerlendirilmesinde üreticiye hem zaman, hem malzeme tasarrufu, hem de kalitenin doğru tayin edilmesi bakımından avantajlar sunmuştur.

Bu çalışmada, laboratuvar ortamında üretilen biyodizeller, metanol, ticari olarak satılan dizel ve çeşitli bitkisel yağlar kullanılarak dizel-bitkisel yağ-biyodizel ve metanol-bitkisel yağ-biyodizel olmak üzere çeşitli üçlü karışımlar hazırlanıp gaz kromatografisi (GC) ile örneklerin analizleri yapılmıştır. Daha sonra, aynı örneklerin yakın infrared (NIR) ve orta infrared (MIR) ölçümleri alınarak çok değişkenli kemometrik kalibrasyon modelleri oluşturulmuştur. Kemometrik kalibrasyon modelleri oluşturulurken, genetik algoritmalara dayalı çok değişkenli kalibrasyon yöntemlerinin yanısıra, yapay sinir ağlarına dayalı çok değişkenli kalibrasyon yöntemleri de kullanılmıştır. Deney sonuçlarında elde edilen verilere göre çok değişkenli kalibrasyon yöntemleri infrared spektroskopik teknikler yardımı ile kullanıldığında biyodizelin kalite tayininde kullanılabilir olduğunu göstermektedir.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Throughout the history, foremost energy demand has been tremendously subjected by fossil fuels – with coal, oil and gas currently accounting for 81 percent of the total demand in world (Figure 1.1). The total energy demand has increased up to 11,400 million tones of oil equivalent per year. However, biomass that is including agricultural and forest products, organic wastes made use of 10 percent (IEA 2007). In recent times, the ratio of the volume and transport demand has increased for liquid biofuels. The main reason for this situation was to support the use of biofuels because of its protection of the environment and the fuel deliver. Given the rapid rise in crude oil and increasingly geopolitical uncertainties, the security of energy supply does not govern politics around the world.

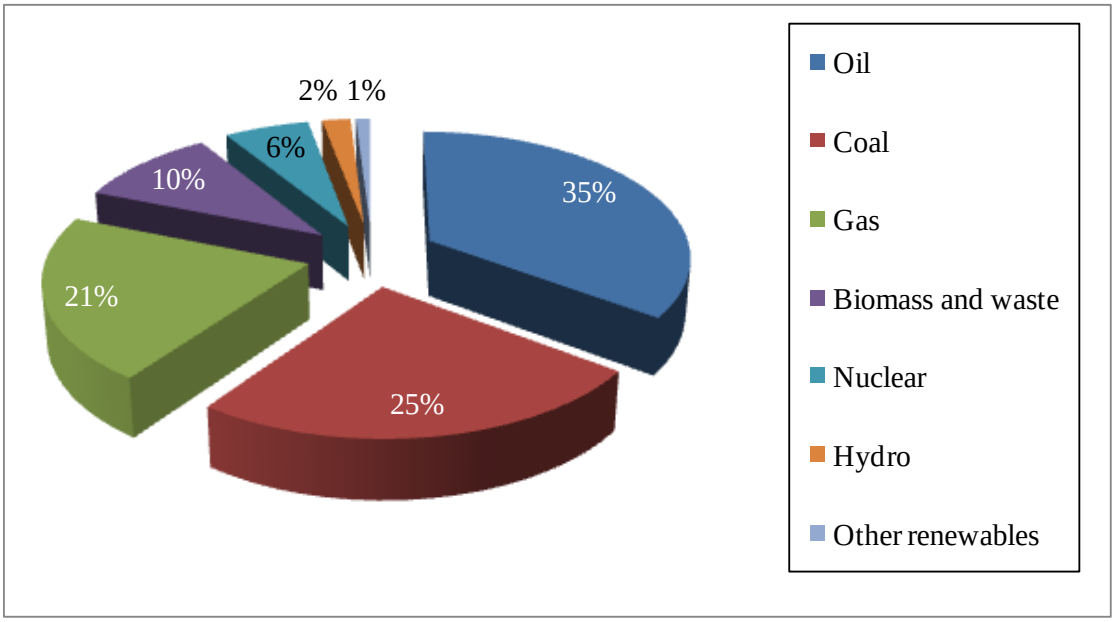


Figure 1.1. World primary energy demand
(Source: IEA 2007)

Different circumstances have proposed for biofuels, which is obtained from biomass sources in the future energy scheme. In the most probable scenario, by 2050 efficient biomass energy will supply about one-half of the total energy offer in developing countries (IPCC 2007). Biofuels are considered to decrease imported petroleum addiction with its related political and financial susceptibility, such as diminishing greenhouse gas production and other pollutants, and regenerate the economy by rising demand for agricultural crop. Presently, biodiesel use is principally dominant in Germany whereas it is produced from soybeans in the United States. The European Union has preferred biodiesel for renewable liquid fuel (Demirbas 2008).

The fact that the biodiesel is produced from renewable resources and environmentally friendly when compared to the fossil-based petroleum diesel, biodiesel has gained an increasing interest in recent years. Biodiesel is mainly produced from a variety of different animal fat and vegetable oil combined with an alcohol in the presence of a homogeneous catalyst and the determination of the quality of the synthesized biodiesel is as important as its production process.

Biodiesel is defined as the mono alkyl esters of long chain fatty acids derived from renewable lipid sources. Biodiesel is widely documented among the alternative fuels industry by the Department of Energy (DOE), the Environmental Protection Agency (EPA) and the American Society of Testing and Materials (ASTM). On the other hand, as other materials, which are tree oil derivatives, other woody products, or even biological slurries, have occasionally been submitted to as "biodiesel." Even though these other resources are natural, and are an alternate for diesel fuel, they are not considered as biodiesel acknowledged by the NBB, DOE, ASTM.

Biodiesel is primarily synthesized by the reaction of a vegetable oil or animal fat with methanol in the presence of a catalyst in order to yield glycerin and methyl esters. The reaction is demonstrated in Figure 1.2 that is given below.

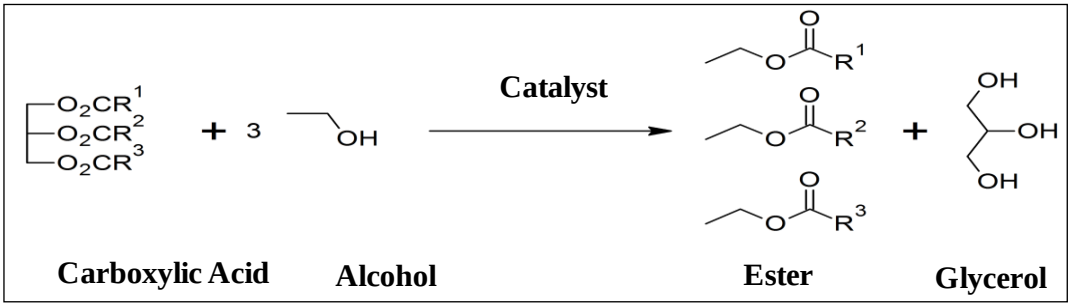


Figure 1.2. General mechanism of transesterification process

Given that this reaction is reversible, excess amount of alcohol is necessary for shifting the equilibrium to the right. In the case of excess alcohol, the forward reaction is a pseudo-first order reaction and the reverse reaction is a second-order reaction.

Biodiesel is an alternative fuel that can be utilized in neat form, or blended with petroleum diesel. Its physical and chemical properties as it relates to operation of diesel engines are similar to petroleum based diesel fuel. These properties are described in Table 1.1.

Table 1.1. Biodiesel ASTM Standards and Properties
(Source: Tyson et al. 2004)

Property	**ASTM**	**Value**	**Unit**
Specific Gravity	D1298	0.86-0.90	g/g H_2O
Gross Heating Value	D2382	11.4 min	kW-hr/kg
Cloud Point	D2500	+3 max	^oC
Pour Point	D97	-3 max	^oC
Flash Point (open cup)	D92	149 min	^oC
Viscosity @ 40^oC	D445	4.00-5.50	Cst
Sulfur	D129	0.02 max	%mass
Carbon Residue	D524	0.1 max	%mass
Cetane number	D613	48 min	
Ash	D482	0.02 max	%mass
Neutralization	D4739	1 max	mg OH/g
Methanol	*G.C	0.2 max	%mass
Free Glycerine	*G.C	0.03 max	%mass
Total Glycerine	*G.C	0.2 max	%mass
Oil Ester	*G.C	97.5 min	%mass

*G.C: Gas Chromatography

These physical and chemical properties make biodiesel quality diesel fuel substitute. However, petroleum diesel is a mixture of hydrocarbon molecules which is derived from crude oil that is supplied from natural resources (Table 1.2). Unadulterated biodiesel restrained up to 10-12 % weight of oxygen, whereas petroleum diesel has approximately 0 % oxygen. The presence of oxygen allocates more complete combustion, which diminishes hydrocarbons, carbon monoxide, and particulate matter emission. However, higher oxygen content increases nitrogen oxides (NO_x) emissions.

Table 1.2. Hydrocarbon Contents in Crude Oil
(Source: ATSDR 1995; OTM 1999)

HYDRO CARBONS	GENERAL FORMULA	CHAIN TYPE	STATE (Room temp)	SAMPLE EXAMPLES
Paraffins (Aliphatic)	C_nH_{2n+2} (n:1 to 20)	Linear or Branched	Gas or Liquid	Methane Propane Hexane
Aromatic	C_6H_5-Y	One or More Benzene Rings with Long Chains	Liquid	Benzene Napthalene
Napthenes (Cycloalkanes)	C_nH_{2n}	One or More Cycloalkane Rings	Liquid	Cyclohexane Methyl Cyclohexane
Alkenes (Olefin)	C_nH_{2n}	Liner or Branched One or More Double Bond	Gas or Liquid	Ethylene Butene Isobutene
Dienes and Alkynes	C_nH_{2n+2}	Triple Bond	Gas or Liquid	Butadiene Acetylene

The main reason for biodiesel is an alternative fuel instead of petroleum fuels can be seen the cetane number which indicates the ignition quality of a diesel fuel. It measures a fuel's ignition delay, which is a period between the start of injection and start of combustion of the fuel. Fuels, which have higher cetane number have shorter ignition delays, providing more time for the fuel combustion, process to be completed.

The cetane number term is considered as a straight chain alkane with 16 carbons ($C_{16}H_{34}$), hexadecane or cetane that is demonstrated in Figure 1.3.

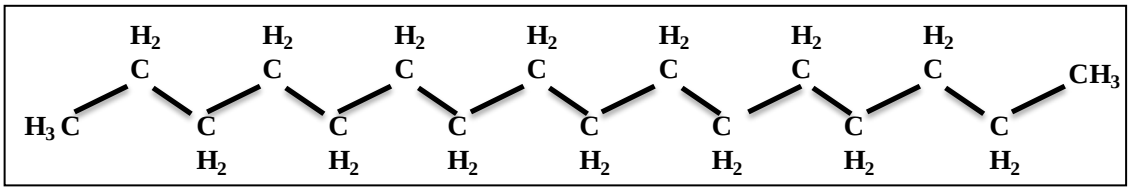


Figure 1.3. Molecular structure of hexadecane structure

This long unbranched hexadecane is regarded as the high quality about the cetane scale and they have called as 100 cetane number. However, highly branched alkanes are low quality compounds on the cetane scale and have low cetane numbers. Methyl esters which are considered as biodiesel have long chain fatty acids with number of carbons varying from 14 to 22 (Figure 1.4). This structure leads to biodiesel as an unconventional diesel fuel (Gerpen et al. 2004).

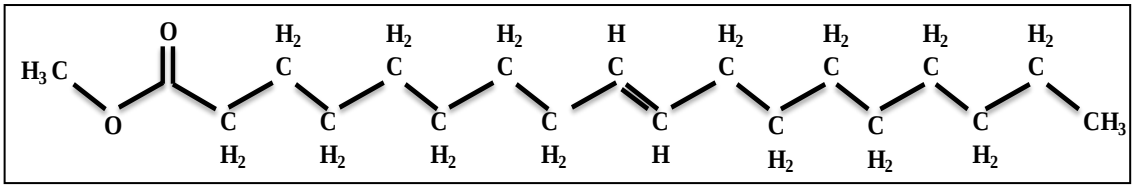


Figure 1.4. Fatty acid methyl ester structure

Biodiesel can be blended at any level with petroleum diesel to produce biodiesel blends, which are proper for compression-ignition engines without any alteration. Biodiesel blends are symbolized as "BXX" with the prefix 'B' indicating Biodiesel and the "XX" demonstrating the percentage of Biodiesel contained in the blend, for example, B20 biodiesel blend contains 20% Biodiesel, 80% petroleum diesel. General blends consist of B2 (2% Biodiesel, 98% petroleum diesel), B5, and B20. B2 and B5 preferred due to their safety in diesel engines also, B20 has established important ecological profits. In developing countries, biodiesel blends have been common for the applications in agricultural equipment such as generators, ships, heating and lighting, etc (ASTM 2010).

However, Rudolph Diesel in fact used fuel derived from peanut oil in the first compression ignition engines, which is later called as diesels. Thus, the studies have been performed on different vegetable oils including soybean oil, sunflower oil, cotton seed oil, corn oil, canola oil, in addition to waste vegetable oil for their feasibility as diesel fuel (Stavarache et al. 2007). The results of these studies have shown that vegetable oils in crude form can be used as a diesel engine fuel with small amounts of power loss when compared to diesel fuel (Knothe 2001). However, using vegetable oils directly in diesel engines may cause some engine problems due to their high viscosity, which is 10–20 times higher than petroleum diesel (Stavarache et al. 2007). There have been developed some methods to reduce viscosity by modifiying the engine or blending vegetable oils with diesel fuels illegally by which adding unconverted vegetable oils to biodiesel feedstock (Wang et al. 2006).

Minimum flash points of both biodiesel and petrodiesel are essential for fire safety requirements. For instance, flash point for petroleum diesel (70 °C) is smaller than for pure biodiesel (160 °C). That smaller flash point is acquired by excess methanol, which can be detached during the production process, due to the fact that methanol leads to diminish the flash point. In addition, existence of methanol in biodiesel can also affect the engine parts such as fuel pumps, seals and elastomers, for this reason it can outcome in poor combustion properties (ASTM 2010)

Furthermore, if excess water presents in the fuel, not only cause to corrosion but also it can promote the enlargement of microbes and germs. In addition, cetane number is used for determination of combustion quality under compression. For this reason, sufficient cetane number is essential for fine engine performance. In addition, cloud point is significant to make certain better engine performance in cold temperatures and carbon residue evaluates the affinity of a fuel to indicate carbon deposits in engine parts. Besides this, acid number is known as an indicator of free fatty acids in biodiesel and it enlarges in case of oxidative degradation.

In the lights of these facts, in order to synthesize biodiesel in a good way, it takes more time and financial problems. Thus, in order to compete with diesel fuel and survive in the market, lower-cost feedstocks are preferred, including waste cooking oil (WCO), grease, soapstocks, since feedstocks costs are more than 85% of the total cost of biodiesel production (Wang et al. 2007). In addition, increasing attention has begun relating to algae-based biodiesel (Campbell 2008). Nevertheless, biodiesel has many rewards against with petroleum diesel; the production cost has turned out to be the principal barrier to its commercialization. Biodiesel unit price is 1.5-3.0 times higher than that of petroleum derived diesel fuel depending on feedstock at present (Demirbas 2008).

In addition to this, the ratio of illegal marketing has been increasing in biodiesel industry such as adulteration of unconverted vegetable oils onto biodiesel/diesel blends (Divya and Mishra 2007). However, the other important circumstance for qualified biodiesel is concerning the monitoring of transesterification reaction for biodiesel synthesis. This typically controls long analysis terms and correlations. Working with process monitoring, enormous data are obtained from the measurement results. In practice, an alternative approach is often needed to use and interpret all the information stored in our database. One of them is the use of models based on statistical principles. The data processing and modeling can be quickly done at the same time using these principles by the help of modern powerful computers.

Recently, the studies related to biodiesel blends concern the topics of chemometric multivariate calibration techniques (Lira et al. 2010; Ferrao et al. 2011; Rio et al. 2010; Gaydou et al. 2011). Various multivariate calibration methods were used to analyze spectra and to construct calibration models. In some studies, mid-infrared (MIR) spectroscopy is used for rapid determination of chemical compositions biodiesel blend species (Oliveira et al. 2004). Near-infrared (NIR) spectroscopy is being used for measuring chemical properties such as cetane number, high combustion value, viscosity with artificial neural networks (Ramadhas et al. 2005). Also, there are some studies for determination of vegetable oils content for biodiesel/diesel blends with partial least squares (PLS), principal component regression (PCR) in literature (Oliveira et al. 2004).

The ultimate goal of this study is to construct multivariate calibration models such as based on genetic algorithm inverse least squares (GILS) and artificial neural networks (ANN) approach for biodiesel blends by using near-infrared and mid-infrared

spectroscopy. Firstly, one of the aims is related to monitor the transesterification reaction of biodiesel synthesis to determine the methanol in reaction medium and investigate the conversion of vegetable oils to methyl esters (i.e. biodiesels). Secondly, the other purpose concerned to find out the illegal marketing of biodiesel/diesel blends in such a way that determining the added unconverted vegetable oils to diesel blends and exactly agree on the amount of BXX blends in market. Thus, one can save time, effort and money by using this type of calibration models for different biodiesel blends.

CHAPTER 2

GAS CHROMATOGRAPHY

2.1. The principle

Chromatography has been popular in separation science in chemistry, which is arisen from adsorption, and partitioning after sample introduction between two phases which are stationary phase and mobile phase. If in mobile phase a gas is used, it is called as gas chromatography (GC). GC is sorted out into gas-solid chromatography (GSC), in which an adsorbent is applied as the stationary phase, and gas-liquid chromatography (GLC), where a partitioning mediator is formed by coating a proper sustain with a liquid.

2.2. The Instrumentation

Primarily, a gas chromatograph consists of the elements that are flowing mobile phase, an injection port (provides to bring in the sample into the flowing mobile phase), and a separation column including the stationary phase, a detector, and a data recording system (Figure 2.1). The sample which is hold by the mobile phase gases such as hydrogen (H_2), helium (He), nitrogen (N_2) or argon (Ar), is injected and immediately vaporized at the column inlet (Schomburg et al. 1990). Then, the vaporized sample is carried through the column by the carrier gas. When passing from beginning to end of the column, constituents in the sample are adsorbed to the stationary phase due to their distinctive concentration fraction. Thus, concentration equilibration achieved repetitively between the stationary, solid and mobile phase. Consequently, the level of adsorption or partition for each component leads differentiations in the rate of association for each component inside the column. The components then elute individually from the column outlet.

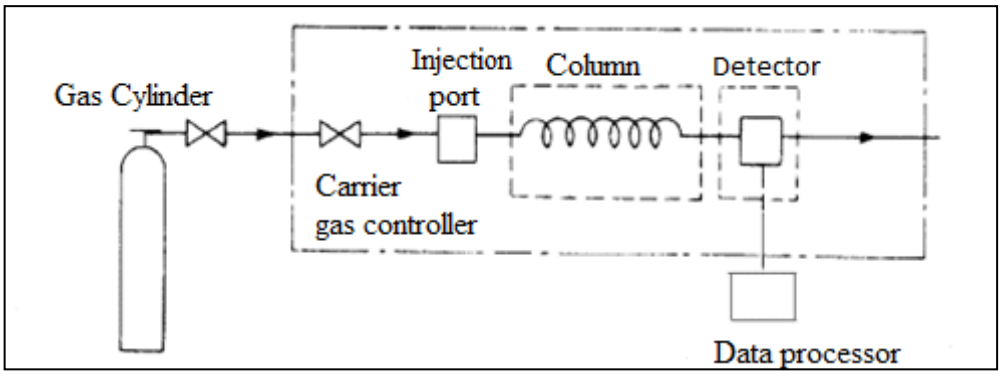


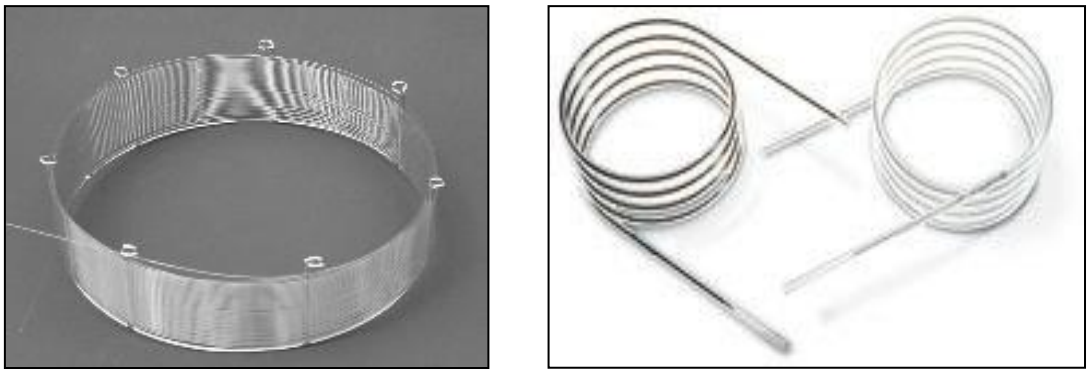
Figure 2.1 Fundamental schematic representation of a gas chromatograph (Source: SHU 2011)

2.3. Columns

The column in a GC is controlled in an oven, the temperature of which is specifically controlled by machine. The most commonly preferred separation columns are the sort of packed columns and capillary columns. Packed columns are 1.5 - 10 m in length containing an internal diameter of 2 - 4 mm. The tubing is usually made of stainless steel or glass and including a set of finely separated inert, solid support material, which is treated with a liquid or solid stationary phase. (Figure 2.2)

Nowadays, the application of capillary columns, which are made-up from stainless steel or quartz and up of fused silica tubing with including an inner diameter of about 30-500 µm, and a length of 10, 30, 60 meters, have been improved because of their high efficiency which has definition of the number of theoretical plates per unit length and temperature constancy.

The stationary phase which solely coats the inner surface is typically a thin film of thermally stable immobilized methylpolysiloxane (OV-1, DB 1, CP-Sil 5, SE-54 etc) (Schomburg et.al. 1990).



(a) (b)

Figure 2.2. Column types (a) Capillary columns, (b) Packed columns
(Source Sigma-Aldrich 2011)

2.4. Detectors

There are numerous diverse types of detectors prevalent to gas chromatography instruments. The general class of compounds being analyzed determines the choice of detector and the sensitivity required. Flame ionization detectors (FIDs) are the most commonly used detectors for organic samples. FIDs use an air/hydrogen flame to pyrolyze the effluent sample. The pyrolysis of the compounds in the flame creates ions and for that reasons the following on current, depends on the flame conditions and the characteristics of the molecule in issue. In other words, the detector demonstrates a diverse reply to each compound that is why, distinct calibrations should be utilized for each compound being analyzed.

FID detector is especially susceptible to organic molecules (10^{-12} g/s, linear range: 10^{6} –10^{7}), however insensible to a few small molecules e.g. N_2, NO_x, H_2S, CO, CO_2, H_2O. when suitable quantity of hydrogen/air are mixed, the combustion does not pay for any ions whereas when the other components are introduced which is including carbon atoms cations are created in the effluent stream (Figure 2.3). In other words, it is mentioned that the more carbon atoms are in the molecule, the more fragments are formed and so the more sensitive the detector is for this compound. Besides this, some gases are typically needed to activate a FID: hydrogen, oxygen (compressed air), and carrier gas.

A further commonly used detector is thermal conductivity detector (TCD) (Figure 2.3). This detector is less susceptible than the FID which has 10^{-5}-10^{-6}g/s, linear range: 10^{3}-10^{4}, however is suitable for preparative purposes, since the sample is not damaged. This type of detector system has the fundamental principle on the comparison of two gas streams, at which one containing only the carrier gas and the carrier gas with the compound. Logically, a carrier gas with a high thermal conductivity e.g. helium or hydrogen is preferred to increase the temperature difference (and thus the difference in resistance) between two thin tungsten wires.

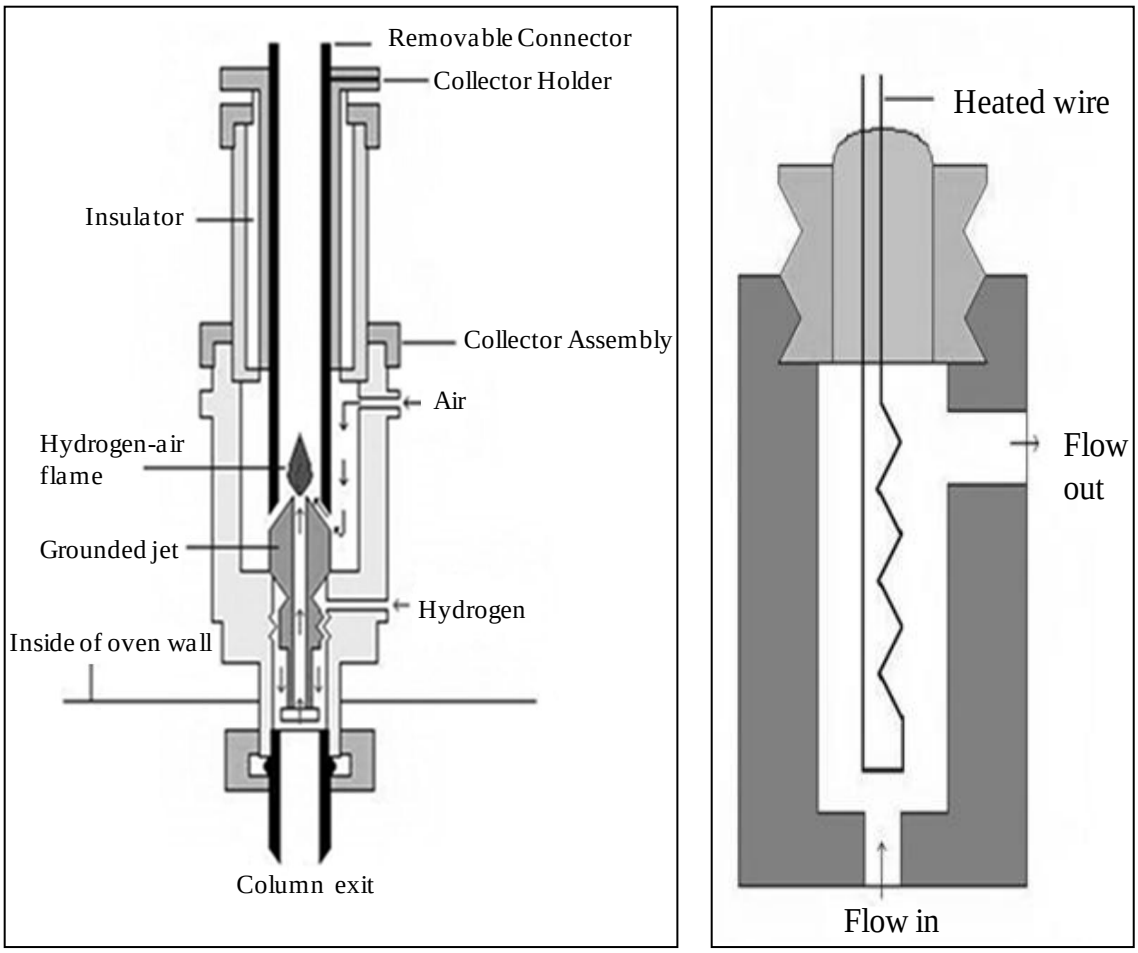


Figure 2.3. Sort of detectors broadly used in Gas Chromatograph (a) Flame Ionization Detector (FID), (b) Thermal Conductivity Detector (TCD) (Source: SHU 2010)

2.5. Sample Introduction: Injection Port

A sample port is essential for initiating the sample at the head of the column. Current injection methods often taking up the use of heated sample ports through which the sample can be injected and vaporized. A calibrated microsyringe is used to deliver a sample volume in the range of a few microliters through a rubber septum and into the vaporization chamber (Figure 2.4). Mainly separations involve only a small fraction of the initial sample volume and a sample splitter is preferred to direct excess sample to waste. Profitable gas chromatographs frequently allocate for both split and splitless injections when flashing between packed columns and capillary columns. The vaporization chamber is usually heated 50 °C above the lowest boiling point of the sample and consequently mixed with the carrier gas to transport the sample into the column.

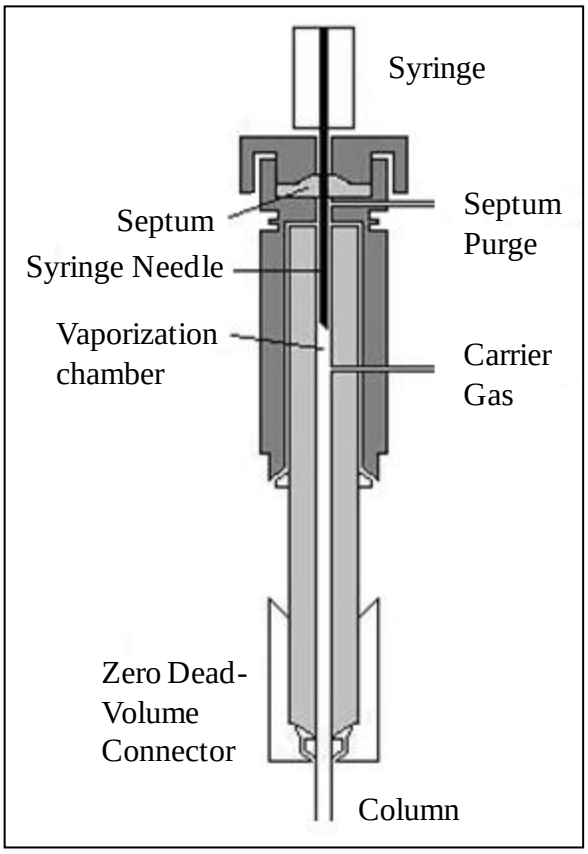


Figure 2.4. The diagram of a split/splitless injector
(Source: SHU 2010)

2.6. Temperature Programming

Temperature programming is defined as a technique that is applied principally in gas chromatography to speed up the elution rate of late peaks that, or else, would take a exceptionally extended time to elute. It is attained by incessantly lift up the column temperature, typically as a linear function of time, throughout the elution process. The retention time of a solute is relative to the distribution coefficient that, consecutively, enlarges as the negative promoter of the standard energy of distribution divided by the product of the gas constant and the absolute temperature. The standard energy is equal to the sum of the standard enthalpy and the product of the standard entropy and the absolute temperature. It is seen that retention is a slightly multifaceted function of temperature. The net effect of temperature programming on solute elution is comparable to the effect of gradient elution in liquid chromatography. In practice, program limits can be as low as 5 ^{o}C and as high as 250 ^{o}C and under convinced situation even higher. Temperature programming is a critical feature for nearly all gas chromatography analyses and so programming services are standard on virtually all gas chromatographs.

CHAPTER 3

INFRARED SPECTROSCOPY

3.1. Infrared Region

Infrared spectroscopy is defined as the study of interaction of infrared light with matter. Infrared radiation extents over electromagnetic spectrum having wavenumbers from 13,000 to 10 cm^{-1}, or wavelengths from 0.78 to 1000 μm. It is surrounded by the red end of the visible region at high frequencies and the microwave region at low frequencies where the region is alienated into three sub-regions due to several varying purposes and instrumentations as it is seen in the Table 3.1 given below (Skoog 1998).

Table 3.1. Infrared spectral regions
(Source: Skoog 1998)

Region	Wavelength Range, μm	Wavenumber, cm^{-1}
Near (NIR)	0.78 – 2.5	12,800 – 4,000
Middle (MIR)	2.5 – 50	4,000 – 200
Far (FIR)	50 – 1000	200 – 10

The far IR involves the use of particular optical equipment and resources. It is utilized for analysis of organic, inorganic, and organometallic compounds concerning heavy atoms such as mass number over 19. It affords constructive information in order to structural revisions for instance lattice dynamics of samples (Sherman et.al 1997). Near IR spectroscopy is requested nominal sample preparation. It recommends high-speed quantitative analysis without utilization or demolition of the sample. Hence, near IR spectroscopy has expanded its attention, particularly in process control purposes. Mid-infrared (MIR) spectroscopy is generally performed for both qualitative and quantitative analysis such as frequently functional to recognize organic, inorganic, biochemical species, biotechnology (Arnold et al. 2000) and pharmaceutical industry (Tran et al. 2004) at which the region around between 900 cm^{-1} and 1300 cm^{-1} described as fingerprint region that is specially for individual composites (Griffiths 1978; Koenig

1975). For instance, although MIR spectra of 1-propanol and 2-propanol are similar to each other, yet, it demonstrates variations in fingerprint region (DeThomas et al. 1994).

The main sort of molecular vibrations are stretching and bending. The various types of vibrations are shown in Figure 3.1. Infrared radiation is absorbed and the associated energy is changed into these type of motions. The absorption occupies distinct, quantized energy levels. Conversely, other rotational motions frequently is an adjunct to the entity vibrational motion. These combinations cause the absorption bands, but not the discrete lines, frequently seen in the mid IR region.

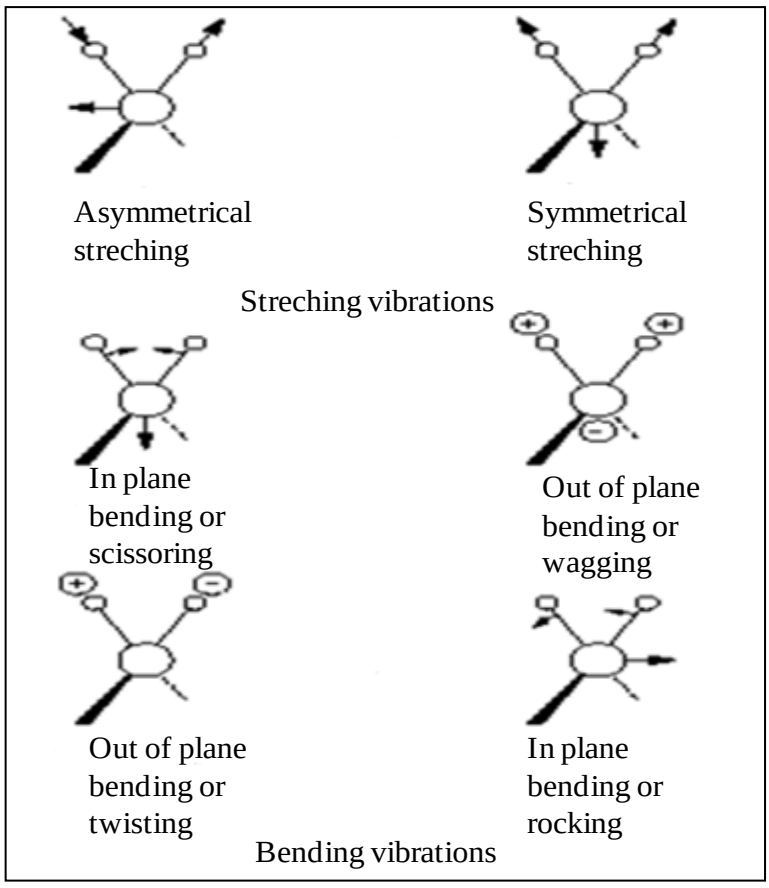


Figure 3.1. Major vibrational modes for a nonlinear group, CH_2. (+ indicates motion from the plane of page toward reader; – indicates motion from the plane of page away from reader.) (Source: Silverstein et al. 1981)

Infrared radiation proposes both rotational and vibrational motions to a molecule where rotational motion has low energy and rotational spectroscopy displays in FIR region. In addition, In MIR region, study fundamental vibrations and rotational-vibrational structures observe while, NIR region where radiation with higher energy fabrication is generally performed to work overtone and combination vibrations. Consequently, in MIR region vibrational quantum number varies in ±1 while in NIR region, vibrational quantum number alters in ±2, ±3, ±4, etc. Table 3.2 reviews the molecular interactions attached with infrared regions.

Table 3.2. Molecular interactions related to infrared regions

Name of region	Featured transitions
Near IR	Both of overtone and fundamental molecular vibrations
Mid IR	Fundamental molecular vibrations and rotations
Far IR	Molecular rotations

The entire number of absorption bands is commonly diverse from the whole number of fundamental vibrations. Since some modes are not IR active and a single frequency leads more than one mode of motion, it is diminished. On the other hand, supplementary bands are created by the form of overtones (integral multiples of the essential absorption frequencies), combinations of fundamental frequencies, differences of fundamental frequencies, coupling interactions of two fundamental absorption frequencies, and coupling interactions between vibrations and overtones or combination bands (Fermi resonance). The intensities of overtone, combination, and difference bands are less than those of the fundamental bands. The combination and blending of all the factors therefore produce a distinctive IR spectrum for each compound (Skoog 1998).

3.2. Infrared Instruments

In simple terms, IR spectra are attained by adjusting in transmittance or absorption intensity as a function of frequency. Mainly commercial instruments detach and evaluate IR radiation by means of dispersive spectrometers or Fourier transform spectrometers.

3.3. Dispersive Spectrometers

Dispersive spectrometers pioneered in the mid-1940s and made available for the robust instrumentation and widespread applications. Non-dispersive instruments are known as filter or non-dispersive photometers that are planned for quantitative analysis yet non-complex, easy to use and not expensive (Skoog 1998).

An IR spectrometer mainly composes of three basic components, which are radiation source, monochromator, and detector. A schematic diagram of a typical dispersive spectrometer is displayed in Figure 3.2. In a typical dispersive IR spectrometer, radiation from a broadband source passes through the sample and is dispersed by a monochromator into component frequencies (Figure 3.2). In that case, the beams fall on the detector, which generates an electrical signal and results in a recorder response.

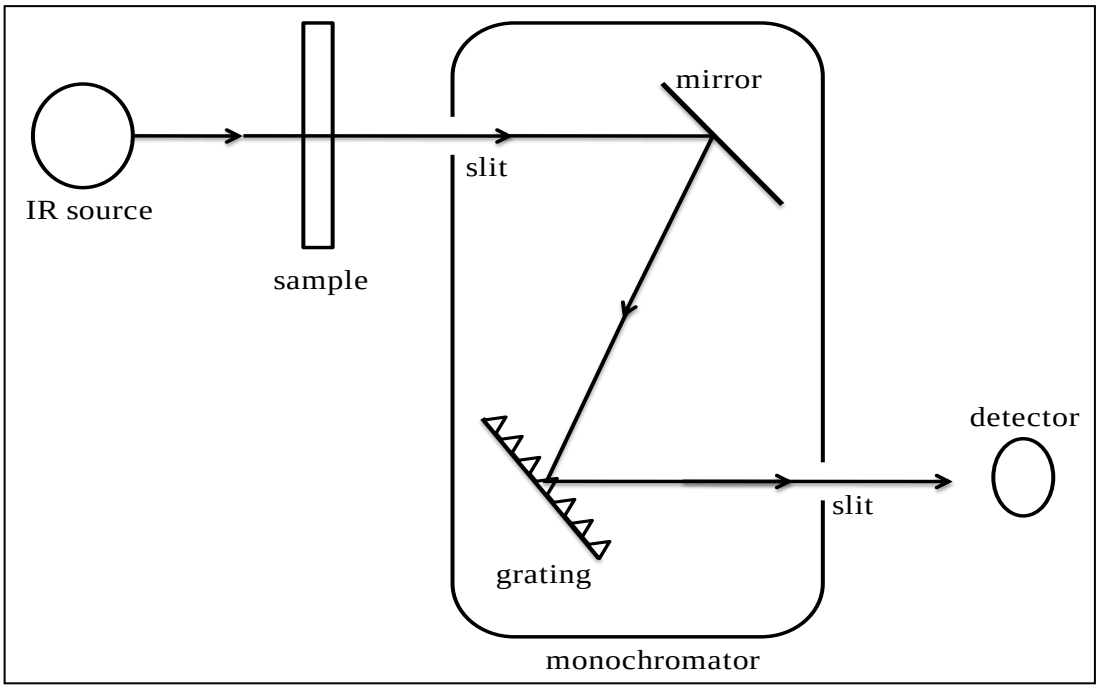


Figure 3.2. Schematic illustration of a commercial dispersive IR instrument

The frequent radiation resource for the IR spectrometer is an inert solid, which is heated up to 1000-1800 °C. Three popular kinds of basis are Nernst glower (created of rare-earth oxides), Globar (constructed of silicon carbide), and Nichrome coil. The entire produce continuous radiations, however with different radiation energy profiles (Skoog 1998).

The monochromator is a device applied to disperse a broad spectrum of radiation and makes available a continuous calibrated series of electromagnetic energy bands of determinable wavelength or frequency range. Prisms or gratings are the dispersive components utilized in cooperation with variable-slit mechanisms, mirrors, and filters. Narrower slits facilitate the instrument in order to better distinguish more closely spaced frequencies of radiation, resulting in better resolution. However, wider slits allocate further light to attain the detector and present enhanced system sensitivity (Skoog 1998).

Dispersive IR spectrometers detectors sorted out into two modules: thermal detectors and photon detectors. Thermal detectors contain thermocouples, thermistors, and pneumatic devices (Golay detectors). They determine the heating effect produced by infrared radiation. A variety of physical property changes is quantitatively determined: expansion of a nonabsorbing gas (Golay detector), electrical resistance (thermistors), and voltage at junction of dissimilar metals (thermocouple).

Nonconducting electrons are excited to a conducting state (Skoog 1998). Therefore, a small current or voltage can be generated. Thermal detectors supplies a linear response over a wide range of frequencies, on the other hand, they reveal slower response times and lower sensitivities rather than photon detectors.

3.3.1. Fourier Transform Spectrometers

Fourier transform spectrometers have preferred instead of dispersive instruments owing to their rapid analysis time and enhanced sensitivity. They have been applied to many areas, which are tricky or unfeasible to analyze, by dispersive instruments because of the fact that Fourier transform infrared (FTIR) spectroscopy scan the entire frequencies simultaneously.

FT system has mainly three basic spectrometer components which are radiation source, interferometer, and detector. Optical presentation of a classic FTIR spectrometer is demonstrated in Figure 3.3.

The similar types of radiation sources are applied for both dispersive and Fourier transform spectrometers. Yet, the source is generally water-cooled system in FTIR instruments in order to supply enhanced stability and power.

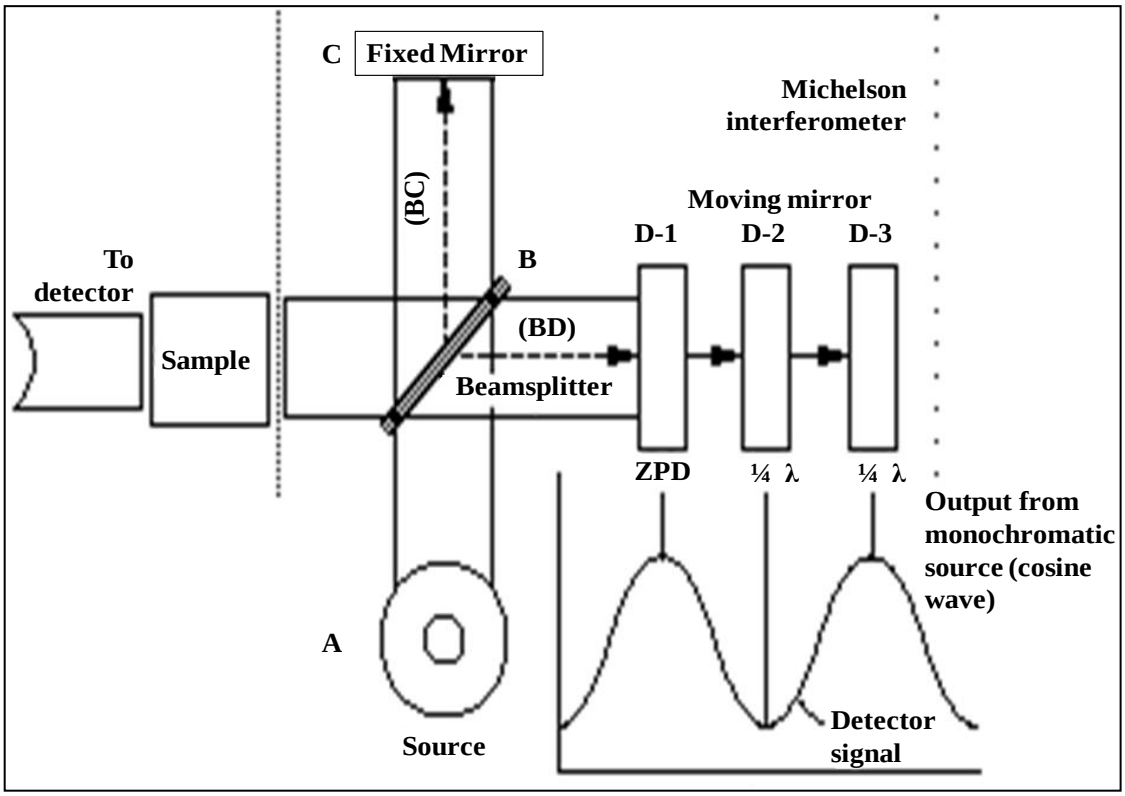
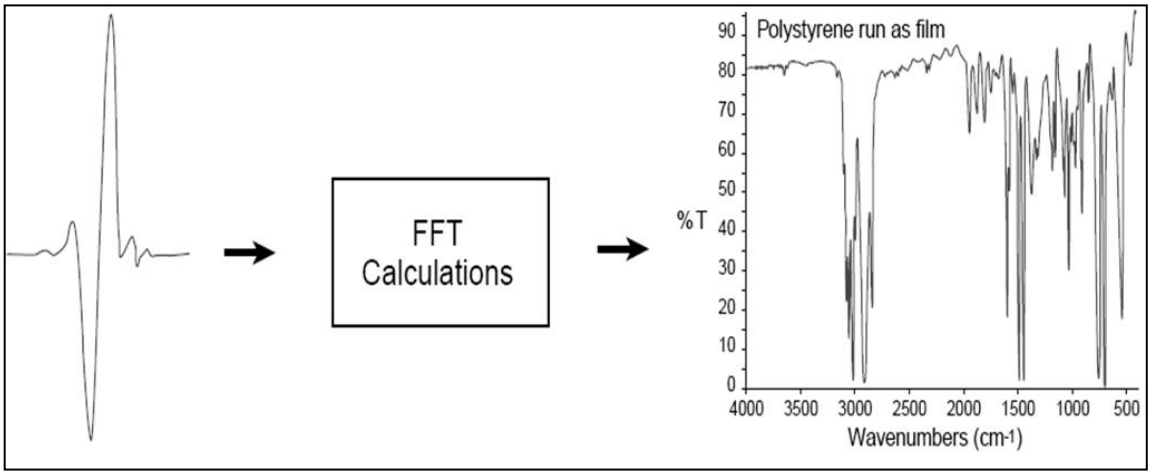


Figure 3.3. Basic optical illustration of a typical FTIR spectrometer
(Source: Newport Corporation 2011)

In other words, the detector response for a single-frequency component from the IR source is primarily reflected on. This imitates a desired situation at which the source is monochromatic, like a laser source. As mentioned before, differences in the optical paths among the two split beams are produced by changeable the relation position of moving mirror to the fixed mirror. The two beams are entirely in phase with each other; accordingly, they obstruct profitably and cause a maximum in the detector response. That position of the moving mirror is defined as the point of zero path difference (ZPD). While the moving mirror travels in either direction by the distance λ/4, the optical path (beamsplitter–mirror–beamsplitter) is changed by 2 (λ/4), or λ/2. The two beams are 180° out of phase with each other, and so interfere destructively. As the moving mirror travels another λ/4, the optical path difference is now 2 (λ/2), or λ. The two beams are performed in phase and result in another practical interference. While the mirror is moved at a constant velocity, the intensity of radiation of the detector changes in a sinusoidal behavior and constructed the interferogram output shown in Figure 3.4

However, diverse advance is considered in an FTIR spectrometer to differentiate and compute the absorption at component frequencies. The monochromator is altered by an interferometer, which divides radiant beams, generates an optical path difference between the beams, and then recombines them in order to produce repetitive interference signals measured as a function of optical path difference by a detector. As its name implies, the interferometer produces interference signals, which contain infrared spectral information generated after passing through a sample.

The most abundant used interferometer is a known as Michelson interferometer which includes information for the total IR region where the detector is receptive. Furthermore, Fourier transformation that is known as a mathematical operation switches the interferogram which is defined as a time domain spectrum displaying intensity versus time within the mirror scan, to the final IR spectrum, that is known as the familiar frequency domain spectrum showing intensity versus frequency. That makes obvious how the term *Fourier transform infrared* spectrometry is produced.



Interferogram Spectrum

Figure 3.4. Schematic representation of an interferogram and a spectrum.
(Source: ThermoNicolet 2009)

The detector signal is utilized at particular intervals throughout the mirror scan. An internal reference is used to modulate sampling rate and a controlled monochromatic beam obtained from helium neon (HeNe) laser focused on a distinct detector. The two mainly trendy detectors for a FTIR spectrometer are deuterated triglycine sulfate (DTGS) and mercury cadmium telluride (MCT). The response times of various detectors such as thermocouple and thermistors detectors selected in dispersive IR instruments are excessively slow for the quick scan times like 1 second or less. Since DTGS detector uses the altering in temperature instead of the value of temperature, it leads to retort in quickly. In addition, the MCT detector is known as photon detector, which uses the quantum nature of radiation and also reveals very fast responses. While DTGS detectors work at room temperature, MCT detectors should be controlled at liquid nitrogen temperature (77 °K) to be effective. In most cases, the MCT detector is considered as rapid and more perceptive rather than the DTGS detector.

Primarily FTIR spectrometers are obtained as single-beam instruments at which does not attain transmittance or absorbance IR spectra in real time.

A usual working process is mentioned as following steps given below:

i. Firstly, background spectrum (Figure 3.5) is gathered with the help of an interferogram where the raw data is pursued by dealing out the data by Fourier transform conversion. That refers to response curve of the spectrometer and takes account of the combined performance of source, interferometer, and detector. The background spectrum contain the contribution from any ambient water (two irregular groups of lines at about 3600 cm^{-1} and about 1600 cm^{-1}) and carbon dioxide (doublet at 2360 cm^{-1} and sharp spike at 667 cm^{-1}) present in the optical working range.

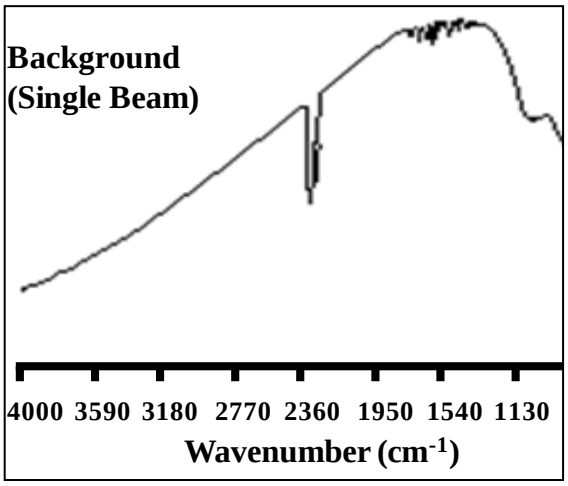


Figure 3.5. A single-beam IR spectrum background, which is showing contribution from trace amount of ambient water and carbon dioxide (Source: Shermann 1997).

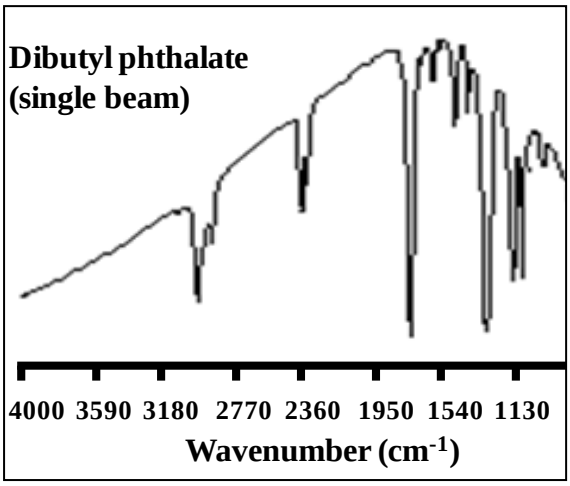


Figure 3.6. A single-beam IR spectrum of dibutyl phthalate (a liquid sample) (Source: Shermann 1997).

ii. Second, a single-beam sample spectrum is required to be gathered (Figure 3.6). and absorption bands from the sample and the background which is air or solvent.

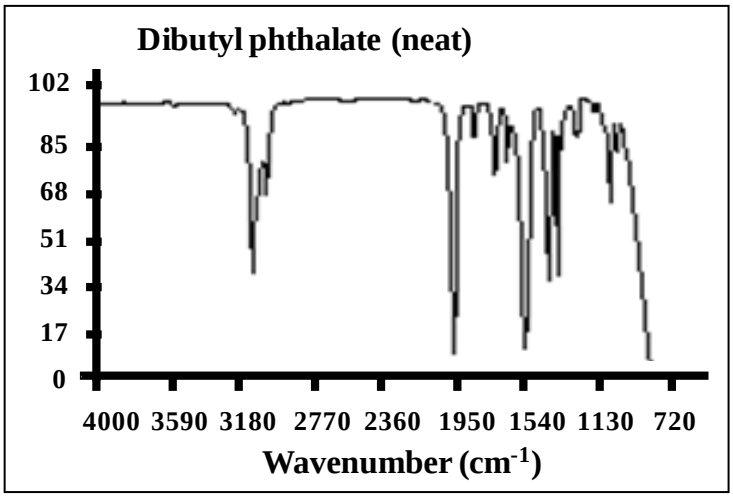


Figure 3.7. The "double-beam" IR spectrum of dibutyl phthalate, produced by ratio of the corresponding single-beam sample spectrum against the single-beam background spectrum. (Source: Shermann 1997)

To diminish effect of background absorption from water and carbon dioxide in the environment (Figure 3.7), inert gas or with dry, carbon dioxide–scrubbed air purges optical bench. Spectrometer alignment, which includes optimization of the beamsplitter angle, is recommended as part of a periodic maintenance or when a sample accessory is changed.

3.3.2. FTIR Advantages

Rapid analysis time and enhanced sensitivity (*Felgett advantage*). Full spectrum can be gathered through a single scan of the moving mirror, whereas the detector monitors all frequencies at the same time. An FTIR instrument can achieve the same signal-to-noise (S/N) ratio of a dispersive spectrometer in a fraction of the time (1 sec or less versus 10 to 15 min). Since multiple spectra can be easily obtained in 1 min or less.

Improved optical throughput (*Jaquinot advantage*). Most abundantly, circular optical slit is preferred in FTIR instruments. The beam area of an FT instrument is generally 75 to 100 times larger than the slit width of a dispersive spectrometer. Therefore, further radiation energy is made accessible which composes advantage for energy-limited samples.

Internal laser reference (*Connes advantage*). The use of a helium neon laser as in FTIR systems presents routine calibration in an accuracy of better than 0.01 cm^{-1} which removes the requirements for external calibrations.

Prevailing database station. Current FTIR spectrometers are usually equipped with a sophisticated data system, which executes broad range of data processing works like Fourier transformation, baseline correction, smoothing, integration, and library searching (Shermann 1997).

3.4. Sample Techniques

It is potential to get an IR spectrum from samples in several forms, such as liquid, solid, and gas. Nevertheless, most materials are resistance to IR radiation and should lead to be dissolved or diluted in a transparent matrix so as to gain spectra. Otherwise, it is achievable to get reflectance or emission spectra from opaque samples. Recent popular sampling techniques and accessories are discussed here.

Liquid cells are used for dilute solutions of solid and liquid samples which are dissolved in relatively IR-transparent solvents. The commonly used solvents are carbon tetrachloride for the region between 4000 and 1330 cm^{-1} and carbon disulfide for the region between 1330 and 625 cm^{-1}. Polar solvents such as water and alcohols are rarely used since they absorb strongly in the mid IR range and react with alkali-metal halides, such as NaCl, commonly used for cell windows. IR spectra of aqueous samples needed to use of special types of liquid cells such as thin cells of BaF2, AgCl, or KRS-5 (a mixed thallium bromide–thallium iodide).

Pellets are used for solid samples, which are not easy to melt or dissolve in any appropriate IR-transmitting solvents. The sample which has amount about 0.5 to 1.0 mg is thinly ground and thoroughly mixed with about 100 mg of dry potassium bromide (or any other alkali halides) powder. Grinding can be performed with an agate mortar and pestle, or lyophilization. The mixture is then pressed into a transparent disc in an evacuable die at sufficiently high pressure. To decrease band distortion because of scattering of radiation, the sample should be ground to particles of 2 μm or less in size. The IR spectra produced by the pellet technique often exhibit bands at 3450 cm^{-1} and 1640 cm^{-1} due to absorbed moisture (Skoog 1998).

Gas cells can be preferred to determined gases or low-boiling liquids. These cells including of a glass or metal body, two IR-transparent end windows, and valves for filling gas. They supplies vacuum-tight light paths from a few centimeters to 120 m. and longer path lengths are used to reflect the IR beam repetitively through the sample by using internal mirrors located at the ends of the cell. Sample gas pressure needed to obtain reasonable spectra depending on the sample absorbance and the cell's path length. Classically, a good spectrum can be obtained at a partial pressure of 50 torr in a 10-cm cell. Analysis of multicomponent gas samples at parts-per-billion levels can be effectively achieved.

Attenuated total reflectance (ATR) accessories are especially useful for obtaining IR spectra of difficult samples that cannot be readily examined by the normal transmission method. They are suitable for studying thick or highly absorbing solid and liquid materials, including films, coatings, powders, threads, adhesives, polymers, and aqueous samples. ATR requires little or no sample preparation for most samples and is one of the most versatile sampling techniques. ATR occurs when a beam of radiation enters from a more-dense (with a higher refractive index) into a less-dense medium (with a lower refractive index). The fraction of the incident beam reflected increases when the angle of incidence increases and entire incident radiation is totally reflected at the interface at which the angle of incidence is larger rather than the critical angle, which is a function of refractive index. The beam infiltrates in small distance further than the interface and into the less-dense medium prior to the complete reflection occurs. This penetration is defined as the evanescent wave and characteristically is at a depth of a few micrometers (μm) whose its intensity is attenuated by the sample in regions of the IR spectrum where the sample absorbs. Figure 3.8 demonstrates the basic ATR principle (Skoog 1998).

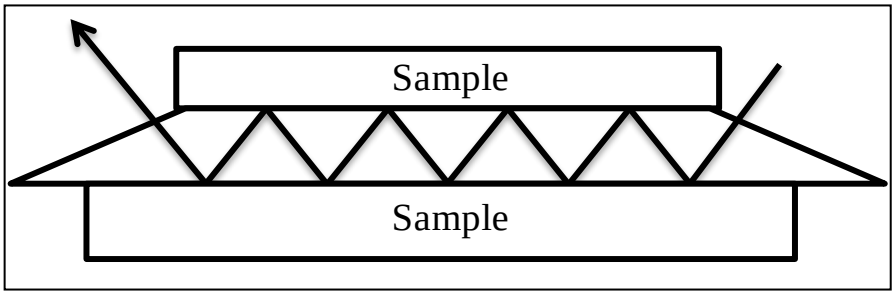


Figure 3.8. Basic illustration of multiple internal reflection effect in ATR (Source: Shermann 1997)

The sample is usually located with high-refractive-index crystal such as zinc selenide, thallium bromide–thallium iodide (KRS-5), or germanium. Several types of ATR accessories are commercially available, such as 25 to 75° vertical variable-angle ATR, horizontal ATR, and Spectra-Tech Cylindrical Internal Reflectance Cell for Liquid Evaluation (CIRCLE) cell (Shermann 1997).

The resulting ATR-IR spectrum is similar to the conventional IR spectrum. Even though, the absorption band positions are alike in the two spectra, but the relative intensities of analogous bands are different.

CHAPTER 4

MULTIVARIATE DATA ANALYSIS METHODS

The International Union of Pure and Applied Chemistry (IUPAC) states that "In general, calibration is an operation that relates an output quantity to an input quantity for a measuring system under given conditions" (Danzer et al. 1998) while according to the International Vocabulary of Basic and General Terms in Metrology (VIM), calibration is an "operation establishing the relation between quantity values provided by measurement standards and the corresponding indications of a measuring system, carried out under specified conditions and including evaluation of measurement uncertainty"

The improvement of the discipline *chemometrics* is concerned to the usage of computers in chemistry. Scientists in the 1970s were previously dealing with statistical and mathematical methods, which are defined, recently to chemometric methods.

The International Chemometrics Society (ICS) is defined the term "Chemometrics" as the following words; it is the science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods. The topics of chemometrics are also related to problems of the computer-based laboratory, to methods for handling chemical or spectroscopic databases and to methods of artificial intelligence. Chemometric techniques are arranged for collecting good data e.g. optimization of experimental parameters, design of experiments, calibration, signal processing and for getting information from these data e.g. statistics, pattern recognition, principal component analysis. In addition, chemometricians contribute to the development of all these methods. As a rule, these developments are dedicated to particular practical requirements, such as the automatic optimization of chromatographic separations or in prediction of the biological activity of a chemical compound (Source: IUPAC 2010). This chapter is focused on the calibration modeling techniques that are used in this study.

4.1. Overview

Several chemical applications of chemometrics related to the calibration, which is defined as a process model created to obtain a relationship within the output of an instrument and properties of samples. In addition to this, prediction is known as a process where the constructed model is used to predict the properties of a sample, whose instrument output is set. The model is constructed by measuring instrument responses and concentration levels of certain chemical contents of the samples. Then, this model is used to predict the concentration of an unknown content sample in the future (Chemometrics: A Practical Guide 2008). In this study, responses which are absorbance values obtained via instruments refer to MIR and NIR spectra, and concentration levels refer to biodiesel blends and their individual concentration of samples.

To find patterns in data and to assign samples, materials or in general, objects, to those patterns, calibration methods of data analysis are applied. Generally, for not sophisticated instruments, merely one response is taken from instrument and this response is correlated to the concentration of the chemical component of a sample. This technique is entitled as *univariate calibration* due to number of instrumental response for each sample is only one. However, the process requires a calibration or training data set, which includes reference values for the properties of interest for prediction, and the measured attributes believed to correspond to these properties. For instances, one can assemble data from a number of samples, including concentrations for an analyte of interest for each reference sample corresponding infrared spectrum. The process that relates multiple instrument responses to one or more properties of a sample is known as *multivariate calibration*. The sample can be multi-component and the goal is to predict the concentrations of the components from, for example, UV-Vis absorption measurements. (Chemometrics: A Practical Guide 2008)

4.2. Univariate Calibration

When deploying a univariate method to predict for unknown samples, one essentially assumes that the signal is highly selective for the analyte of interest. According to the latest IUPAC recommendation, "selectivity refers to the extent to which the method can be used to determine particular analytes in mixtures or matrices without interferences from other components of similar behavior" (Vessman et.al 2001).

Generally, in chemical analysis, this kind of calibration modeling has been preferred. In establishing the univariate calibration function, defined as the functional relation between the expected instrumental responses and analytes concentrations, the proper calibration design has to be taken into account. For absorption or chromatography studies, absorption at a wavelength or a peak area is communicated with the concentration of a sample. If the model is considered as linear, the model could be either classical calibration or inverse calibrations, which are based on Beer's law. According to the Beer's law, the absorptivity coefficient is directly proportional to absorbance at a wavelength, light path length and concentration.

4.3. Classical Calibration

One of the simplest problems is to determine the concentration of a single compound using the response at a single detector, for example a single spectroscopic wavelength or a chromatographic peak area. In this type of calibration models, absorbance at a spectroscopic wavelength of a chromatographic peak area is related as a function of concentration. Mathematically, the general formula of classical calibration is

$$\mathbf{a} \approx \mathbf{c} \cdot s \tag{4.1}$$

where, in the simplest case, $\mathbf{a}$ is the vector of absorbance at one wavelength for a number of samples and $\mathbf{c}$ is the vector of corresponding concentrations. The scalar coefficient s is related with these parameters and can be determined by the following equation:

$$s \approx \left(\mathbf{c}' \cdot \mathbf{c}\right)^{-1} \cdot \mathbf{c}' \cdot \mathbf{a} \tag{4.2}$$

where the $\mathbf{c}'$ is the transpose of the concentration vector.

After determining s, the prediction model for an unknown is constructed as,

$$\hat{c} \approx \hat{a} / s \tag{4.3}$$

where the hat symbol for scalars a and c refer to prediction.

Eventually, residuals are calculated in order to control whether the prediction model is qualified or not. Residuals or errors are considered as the difference between the actual and predicted concentration values ($c - \hat{c}$). It is always useful, however, to check the original graph just to be sure, and this percentage appears reasonable. Therefore; residuals value should be the least value as much as possible in order to construct a better model (Brereton 2003).

$$e = c - \hat{c} \tag{4.4}$$

The less the residuals mean the better the model

4.3.1. Inverse Calibration

Notwithstanding classical calibration is widely used in literature, it is not always the most appropriate approach in chemistry, for two main reasons. First, the ultimate aim is usually to predict the concentration (or independent variable) from the spectrum or chromatogram (response) rather than vice versa. The second relates to error distributions. Furthermore, the response errors are arisen from instrumental performance, on the other hand the ratio of reliability of instruments have been increased due to gravimetrically determination of concentration values, which lead to larger than instrumental error. In Figure 4.1 given below demonstrates the difference between errors stem from instrument and concentration.

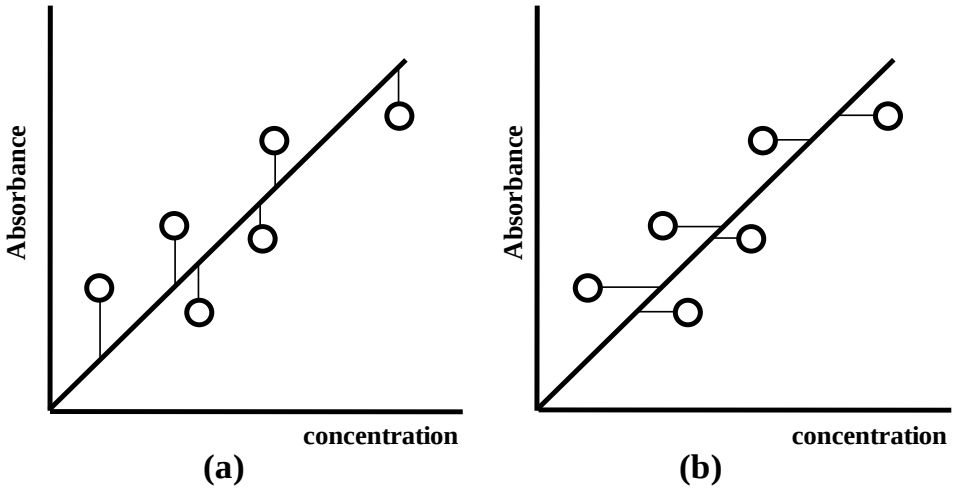


Figure 4.1. Difference between errors in (a) classical and (b) inverse calibration

Inverse calibration can be modeled as the following equations,

$$\mathbf{c} \approx \mathbf{a} \cdot b \tag{4.5}$$

at which b is a scalar coefficient and inverse of s due to variation on errors for every model. Then, b is calculated the formula given below,

$$b \approx \left(\mathbf{a}' \cdot \mathbf{a}\right)^{-1} \cdot \mathbf{a}' \cdot \mathbf{c} \tag{4.6}$$

and determination of unknown sample can be carried out simply

$$\hat{c} \approx \hat{a} \cdot b \tag{4.7}$$

It is also useful to realize that similar methods can be applied to classical calibration, the details being omitted for brevity, as it is recommended that inverse calibration is performed in normal circumstances (Brereton 2003).

4.4. Multivariate Calibration Methods

Multivariate calibration can be seen as a promising mathematical approach to the ubiquitous selectivity problem. Chemometricians have fostered multivariate calibration methods ever since the foundation of chemometrics as an independent branch of chemistry in the early 1970s. The goal is to construct determination of major and also minor components of mixtures and for various instrument types. Therefore multivariate calibration can give rise to the development of new analytical instruments. In addition, it can enhance the analytical capacity and reliability of traditional instruments. (Martens et al. 2004)

Multivariate calibration has some advantages over univariate calibration.

1) When the aim is to see whether a spectrum of a mixture can be employed to determine individual concentrations and may be to replace a slow and expensive chromatographic method by a rapid spectroscopic approach, multivariate could give better approximations.
2) Another different aim might be impurity monitoring: how well the concentration of a small impurity can be determined, for example, buried within a large chromatographic peak. Simultaneous analysis of multiple components in a sample is possible. By univariate method, there has to be one measurement for each component. Thus, spent time will be more. (Brereton 2003)
3) Precision in the prediction can be enhanced by repeating a measurement and calculating the mean. These are consequence of reduction in the standard deviation of the mean, which is called signal averaging. (Brereton 2003)
4) Furthermore, multivariate calibration has fault-detection capabilities. That means unknown interferences in the sample can be overcome by multivariate calibration. In univariate calibration, the presence of interferences may cause wrong prediction of concentration of analyte. To avoid this problem, physical separation of analyte from interfering material or using selective measurements is needed and this means necessity of more effort. Figure 4.2 demonstrates how the calibration curve is affected by the interferences. By multivariate calibration,

nonlinearities caused by the interferences can be reduced by selecting more variables and chance of obtaining better calibration curve can be increased. Therefore, time and effort spent to remove interferences physically is respectably decreased. So relying on multiple wavelengths will result better. (Öztürk 2003)

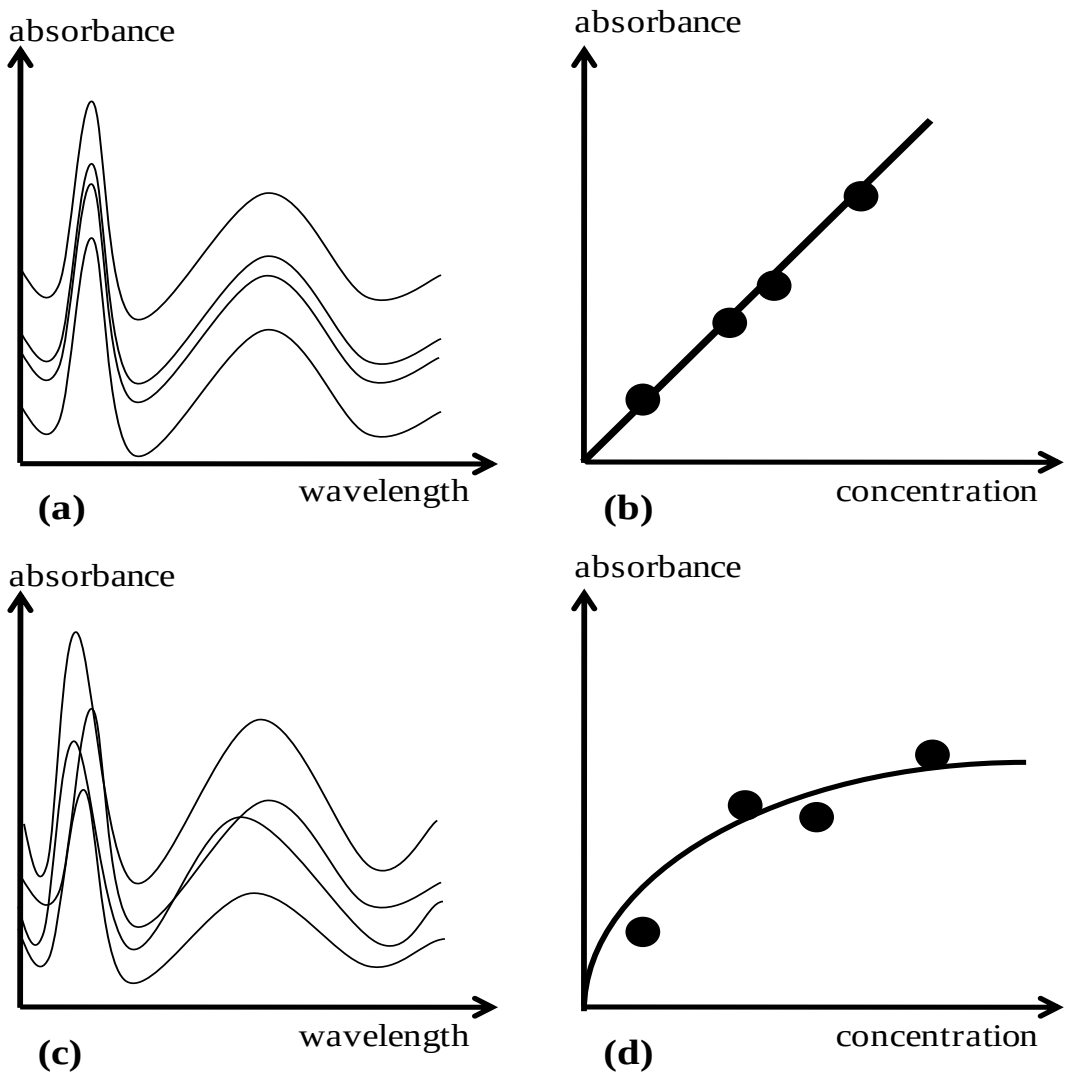


Figure 4.2. (a) Spectra of a sample in different concentrations which has no interference and its calibration curve (b) by univariate calibration; (c) spectra of a sample in different concentrations which has interfering materials and its calibration curve (d) by univariate calibration

In multivariate calibration modeling technique, the equations can be improved in two such cases, likewise in the classical calibration case, absorbance is a function of concentration and the other is likewise in the inverse calibration case, concentration is a function of absorbance. In addition to this, multivariate calibration applies the absorbance full spectral data. Thus, more than one component can be used at which concentration vector becomes a matrix.

In this study, calibration methods based on genetic inverse least squares and artificial neural networks are used. Prior to discussing these methods, it is necessary to give details about classical least squares and inverse least squares methods as a preface of the multivariate calibration methods.

4.4.1. Classical Least Squares (CLS)

This spectroscopic quantitation method, which is also known as *K-Matrix*, is founded in using the Beer Lambert Law to extend the calculation of the absorptivity coefficients across a much larger portion of the spectrum than the much simpler Least Squares Regression method. For instance, Beer's law that is a classical least squares method can be modeled as,

$$\mathbf{A} = \mathbf{C} \times \mathbf{K} + \mathbf{E} \tag{4.8}$$

where $\mathbf{C}$ is the matrix which consists of concentrations of multi-component samples but in case of one component, it is denoted as a vector $\mathbf{c}$. $\mathbf{A}$ is the matrix which consists of absorbance values of the samples at different wavelengths and $\mathbf{E}$ is the error matrix. Each row $\mathbf{C}$ of $\mathbf{A}$ and correspond to one sample, each column represents different component and different absorption values, respectively. $\mathbf{K}$ is the matrix of absorptivity coefficients multiplied by path length. Each member of this matrix corresponds to absorptivity coefficient of an absorption value at a certain wavelength. $\mathbf{K}$ matrix can be determined by the following formula

$$\mathbf{K} = \left(\mathbf{C}' \cdot \mathbf{C}\right)^{-1} \cdot \mathbf{C}' \cdot \mathbf{A} \tag{4.9}$$

So as to carry out prediction, an unknown sample spectrum is measured ($\mathbf{r}$). Given $\mathbf{r}$ and $\mathbf{K}$, concentration can be predicted by using simple matrix algebra:

$$\hat{\mathbf{c}} = \hat{\mathbf{a}} \cdot \mathbf{K}' \cdot \left(\mathbf{K} \cdot \mathbf{K}'\right)^{-1} \tag{4.10}$$

Noted that, prediction elements are vector not scalar as in the univariate calibration since in one unknown sample has more than one component and absorbance value as well.

The difference between the reference and predicted concentration values is known as residual and represents as,

$$\mathbf{e} = \mathbf{c} - \hat{\mathbf{c}} \tag{4.11}$$

In briefly, the CLS method can be applied to such simple systems that the whole pure-component spectra can be measured. Also, to construct the CLS model, the pure-component spectra are measured for each analyte in the sample and it leads to figure out spectral matrix which helps to predict the concentrations in unknown samples.

Furthermore, CLS method has advantageous due to modeling on Beer's Law, relatively fast, proper for moderately complex mixtures and wavelength selection is not a requirement. However, It has to be very susceptible to baseline effects since equations assume the response at a wavelength is due entirely to the calibrated constituents and require knowing the complete composition (concentration of every constituent) of the calibration mixtures.

4.4.2. Inverse Least Squares (ILS)

One of the most widely used spectroscopic quantitation methods is Inverse Least Squares, also known as *Multiple Linear Regression* and *P-Matrix*. In some cases, Most methods based on Beer's Law assume that there is little or no interference in the spectrum between the individual sample constituents or that the concentrations of all the constituents in the samples are known ahead of time. In real world samples, it is very unusual, if not entirely impossible to know the entire composition of a mixture sample. There have been approaches on this purpose and they give some guidance (Haalan et al. 1988).

The relationship between the measurements and concentrations is modeled as in CLS but in this case the concentrations are treated as a function of absorbance values, as shown in the following equation

$$\mathbf{C} = \mathbf{A} \times \mathbf{P} + \mathbf{E} \tag{4.12}$$

whereby $\mathbf{C}$ is the concentration matrix and $\mathbf{A}$ is the absorbance matrix as in CLS. The matrix $\mathbf{P}$ contains the model coefficients and can be determined by

$$\mathbf{P} = \left(\mathbf{A}' \cdot \mathbf{A}\right)^{-1} \cdot \mathbf{A}' \cdot \mathbf{C} \tag{4.13}$$

A predicted concentration of a multi-component sample can be obtained by

$$\hat{\mathbf{c}} = \hat{\mathbf{a}} \cdot \mathbf{P} \tag{4.14}$$

The residual is, same as in the CLS model, known as the difference between the reference and predicted concentration values

$$\mathbf{e} = \mathbf{c} - \hat{\mathbf{c}} \tag{4.15}$$

In ILS, the averaging effect gained by selecting many wavelengths in the CLS method is effectively lost. Therefore, wavelength selection is critically important to building an accurate ILS model. Ideally, there is a crossover point between selecting enough wavelengths to compute an accurate least squares line and selecting few enough so that the calibration is not overly affected by the colinearity of the spectral data.

4.4.3. Genetic Inverse Least Squares (GILS)

GILS can be interpreted as a customized method of ILS in which genetic algorithms (GA) are used as a tool for wavelength selection. GA's are global search and optimization methods based on the principles of natural evolution and selection as developed (Wang et al. 1991) at which defined as evolution, individuals who fit better to the environment are more likely survive and breed, thus are able to pass their genetic information to their offspring. However, individuals who do not fit and unable to adapt will eventually be eliminated from the population. This process progresses slowly over a long period (or may never end) through generations and the species will evolve into

better and fit forms. In the last couple of decades, scientists have been trying to take advantages of the natural evolutions as an improvement concept in the process of solving large-scale optimization problems. In the 1960's biologists have begun to perform the simulation of genetic systems experiments with computer. The initial work in genetic algorithms was done by Holland who developed a GA in his research on adaptive systems in the early 1960's and is considered the father of the field. (Gilbert et al. 1997) Over the years, GA have attracted attention and have been applied to various global optimization problems in many areas including chemometrics. (Fontain et al. 1993; Kateman et al. 1991) In terms of calibration, there have been several applications of GA to wavelength selection. (Lucasius et al. 1994; Williams et al. 1996; Paradkar et al. 1997; Ozdemir et al: 1998)

Computationally the implementation of a typical GA is quite simple and consists of five basic steps including initialization of gene population, evolution of the population, and selection of the parent genes for breeding and mating, crossover and mutation, and replacing the parents with their offspring. These steps have taken their names from the biological foundation of the algorithm. The implementation of a typical GA is shown in Figure 4.3.

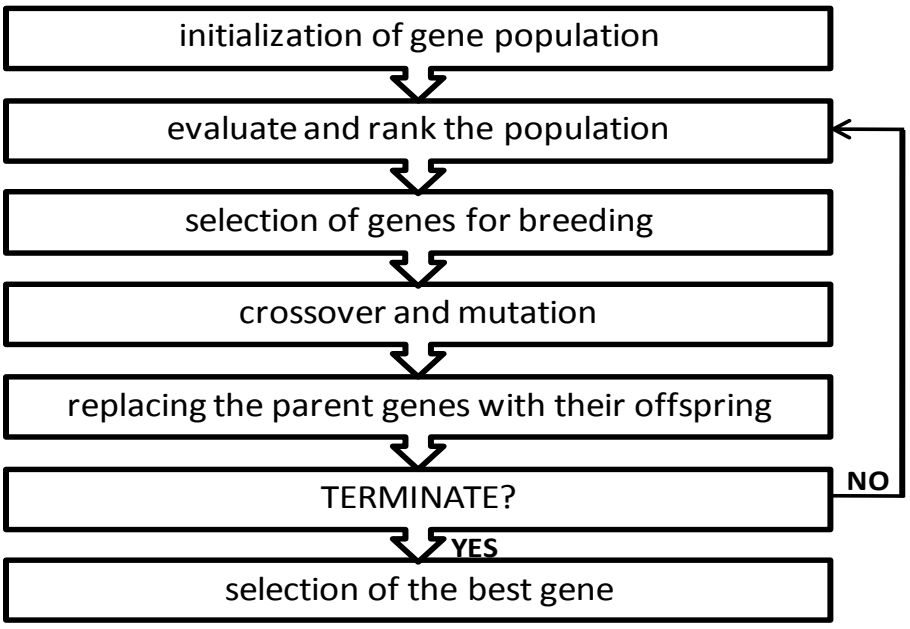


Figure 4.3. Flow chart of genetic algorithm used in GILS.

4.4.3.1. Initialization

In the initialization step, an even number of genes are formed from full spectral data matrix and each gene are used to form model at which a gene is defined as a potential solution to a given problem. The exact form of a gene may vary from application to application and depends upon the problem being investigated. The term population is used to describe the collection of individual genes in the current generation.

In the initial gene pool, a gene consists of absorbance values at randomly chosen wavelengths between 2 and 30. An example of a gene is as the following:

$$S = [A_{8420}, A_{6730}, A_{5987}]$$

where S is so-called a gene, A is the absorbance measured at the indicated wavelength. The chosen absorbance value at one wavelength is a vector of samples. Figure 4.4 shows the schematic representation of the gene for a biodiesel sample NIR spectra. Then, the population is formed according to the number of genes initially entered as an input of the software.

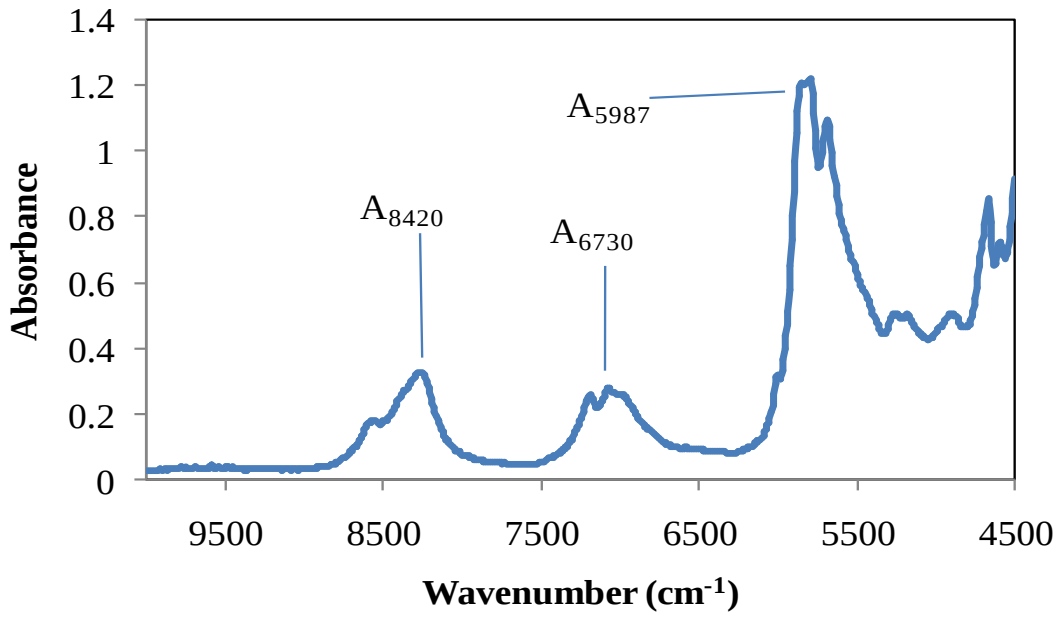


Figure 4.4. Schematic representation of the gene for a biodiesel sample NIR spectra.

4.4.3.2. Evaluate and Rank the Population

This step involves the evaluation of the genes using fitness function, which is the inverse of standard error of calibration (SEC), which is considered from the ILS model in which absorbance values from the selected wavelengths are used to construct the model. SEC is calculated from:

$$SEC = \sqrt{\frac{\sum_{i=1}^{m}\left(c_i - \hat{c}_i\right)^2}{m-2}} \tag{4.16}$$

whereby c_i is the reference and $\hat{c}_i$ are the predicted values of concentration of i^{th} sample and m is the number of samples. Noted that, degrees of freedom is considered as $m-2$ due to a linear model in which there are only two parameters to be extracted which are the slope of the actual vs. reference concentration plot and the intercept. In each step, increase in the fitness value is targeted.

4.4.3.3. Selection of Genes for Breeding

This step involves the selection of the parent genes from the current population for breeding according to their fitness value. The goal is to give higher chance to those genes with high fitness so that only the best performing members of the population will survive in the long run and will be able to pass their information to the next generations. Here, it is expected that the genes better suited for the problem will generate even better offspring. The genes with the low fitness values will be given lower chance to breed and hence most of them will be unable to survive. There are number of selection methods that can be used for parent selection (Wang et al. 1991). Top down selection is one of the simplest methods for parent selection. After genes are ranked in the current gene pool, they are allowed to mate in a way that the first gene mates with the second gene, third one with the forth one and so on. All the members of the current gene are given a chance to breed. Roulette wheel selection method, which is used in GILS, is the one where the chance of selecting gene is directly proportional to its fitness. In this method, each slot in the roulette wheel represents a gene. The gene with the highest fitness has

the biggest slot and the gene with the lowest fitness has the smallest slot. Therefore, when the wheel is rotated, there is a higher chance of being selected for a gene with high fitness than for a gene with a low fitness. There will also be the genes, which are selected multiple times, and some of the genes will not be selected at all and will be thrown out from the gene pool. After all the parent genes are selected, they are allowed to mate top-down, whereby the first gene (S_1) mates with the second gene (S_2), S_3 with S_4 and so on until all the genes mate. Since no ranking is done for the roulette wheel selected genes, the genes with low fitness have a chance to mate with better performing genes after being selected, thus resulting in an increased possibility of recombination.

4.4.3.4. Crossover and Mutation

The genetic algorithm does most of its work in the breeding/mating step. The step involves breaking the genes at random points and cross-coupling them as illustrated in the following example:

Consider S1 and S2 are parent genes which are breeding, S3 and S4 are their corresponding off-springs.

$$S_1 = \left[A_{4255}A_{5732} \oplus A_{9237}A_{4890}\right]$$
$$S_2 = \left[A_{5123}A_{8457}A_{9743}A_{7832} \oplus A_{8922}\right]$$
$$S_3 = \left[A_{4255}A_{5732}A_{8922}\right]$$
$$S_4 = \left[A_{5123}A_{8457}A_{9743}A_{7832}A_{9237}A_{4890}\right]$$

Here the first part of S_1 is combined with the second part of the S_2 to give the S_3, likewise the second part of the S_2 to give S_4. This process is called single point crossover and it is the one used in GILS. The symbol $\oplus$ is used to indicate the place where crossover takes place. There are also other types of crossover methods such as two point crossover and uniform crossover, each having their advantages and disadvantages. In the uniform case, each gene is broken at every possible point and many possible combinations are possible in the mating step, thus resulting in more exploitation. However, it is more likely to destroy good genes. Single point crossover will not provide different offspring if both parent genes are identical, which may happen

in the roulette wheel selection, and broken at the same point. To avoid this problem, two points crossover, where each gene is broken in two points and recombined, can be used. Single point crossover generally does not disturb a good gene but it provides as many recombinations as other types of crossover schemes. Also mating can increase or decrease the number of base pairs in the offspring.

Mutation, which introduces random deviations into the population, can be also introduced into the algorithm during the mating step at a rate of 1% as is typical in GA's. Replacing one of the wavelengths in an existing gene with a randomly generated new wavelength usually does this. However, it is not used in GILS in this study.

4.4.3.5. Replacing the parent genes by their off-springs

After crossover, the parent genes are replaced by their offsprings. The ranking process based on their fitness values follows the evolution step. Then the selection for breeding/mating starts again. This is repeated until a predefined number of iterations are reached.

At the end, the gene with the lowest SEC (highest fitness) is selected for model building. This model is used to predict the concentrations of component being analyzed in the validation set. The success of the model in the prediction of the validation set is evaluated using standard error of prediction (SEP) which is calculated as:

$$SEP = \sqrt{\frac{\sum_{i=1}^{m} \left(c_i - \hat{c}_i\right)^2}{m}} \tag{4.17}$$

whereby *m* is now, in this case, the number of validation samples.

4.4.3.6. Termination

The termination of the algorithm is done by setting predefined iteration number for the number of breeding/mating cycles. However, no extensive statistical test has been done to optimize it, though it can also be optimized. Since the random processes are heavily involved in the GILS, the program has been set to run predefined number of times for each component in a given multi-component mixture. The best run, i.e. the

one generating the lowest SEC for the calibration set and at the same time obtained SEP for the validation set that is in the same range with SEC was subsequently selected for evaluation and further analysis.

GILS has some major advantages over the classical univariate and multivariate calibration methods. First of all, it is quite simple in terms of the mathematics involved in the model building and prediction steps, but at the same time it has the advantages of the multivariate calibration methods with a reduced data set since it uses the full spectrum to extract genes. By selecting a subset of instrument responses, it is able to eliminate nonlinearities that might be present in the full spectral region.

4.4.4. Artificial Neural Networks (ANN)

A neural network is defined as a massively parallel-distributed processor which has a natural susceptibility for storing experiential knowledge and making it available for use. It is similar to the brain in terms of two purposes (Haykin et al. 1998):

i. Knowledge is acquired by the network through a learning process.

ii. Interconnection strengths known as synaptic weights are used to store the knowledge.

Learning is a process by which the free parameters (i.e., synaptic weights and bias levels) of a neural network are adapted through a continuing process of stimulation by the environment in which the network is embedded. The type of learning is determined by the manner in which the parameter changes take place. In a general sense, the learning process may be classified as follows:

i. Learning with a teacher, also referred to as supervised learning

ii. Learning without a teacher, also referred to as unsupervised learning

There are a large number of different types of networks, but they all are characterized by the following components: a set of nodes, and connections between nodes, computational units, which receive inputs, and process them to obtain an output.

One type of network sees the nodes as ‘artificial neurons’. These are called artificial neural networks (ANNs). An artificial neuron is a computational model inspired in the natural neurons. Natural neurons receive signals through synapses located on the dendrites or membrane of the neuron. When the signals received are strong enough (surpass a certain threshold), the neuron is activated and emits a signal

though the axon. This signal might be sent to another synapse, and might activate other neurons.

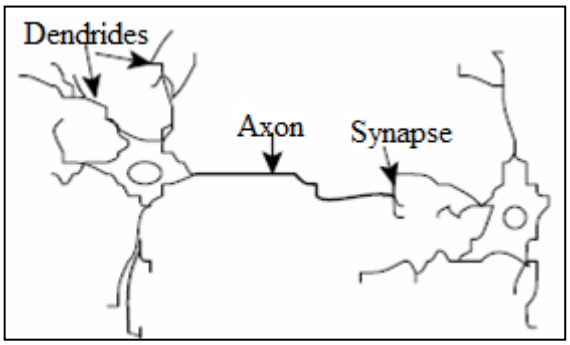


Figure 4.5. An illustration of natural neuron

The complexity of real neurons is highly abstracted when modeling artificial neurons. These consist of inputs (like synapses), which are multiplied by weights (strength of the respective signals), and then computed by a mathematical function which determines the activation of the neuron. Another function (which may be the identity) computes the output of the artificial neuron (sometimes in dependence of a certain threshold). ANNs combine artificial neurons in order to process information.

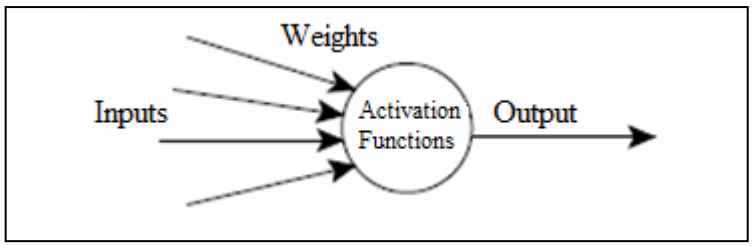


Figure 4.6. The schematic illustration of artificial neuron

The higher a weight of an artificial neuron is, the stronger the input, which is multiplied by it, will be. Weights can also be negative, so that the signal is inhibited by the negative weight. Depending on the weights, the computation of the neuron will be different. The output is obtained by adjusting the weights of an artificial neuron for specific inputs. However, when an ANN of hundreds or thousands of neurons, it would be quite complicated to find by hand all the necessary weights. Thus, it should be found algorithms which can adjust the weights of the ANN in order to obtain the desired

output from the network. This process of adjusting the weights is called learning or training.

4.4.4.1. Supervised Learning

This form of learning assumes the availability of a labeled (i.e., ground-truthed) set of training data made up of *N* input – output examples as shown in equations (4.18):

$$T = \{(x_i, d_i)\}\left\langle \begin{matrix} N \\ i = 1 \end{matrix} \right\rangle \tag{4.18}$$

whereby $\mathbf{x}_i$ = input vector of *i*th example, d_i = desired (target) response of *i*th example, assumed to be scalar for convenience of presentation and *N* = sample size.

Given the training sample *T*, the requirement is to compute the free parameters of the neural network so that the actual output y_i of the neural network due to $\mathbf{x}_i$ is close enough to d_i for all i in a statistical sense. For example, we may use the mean-square error as the index of performance to be minimized by following equation.

$$E(n) = \frac{1}{N}\sum_{i=1}^{N}(d_i - y_i)^2 \tag{4.19}$$

4.4.4.2. Feed-forward networks

The field of neural networks envelops a broad range of diverse network methods that is improved for and performed to very different situations. The "feed-forward" network structure is especially appropriate for treatment non-linear relationships between "input" and "output" variables, when the focus is prediction.

A feed-forward network is a known as function where the information from the input data utilizes through from intermediate variables to the output data. The input data (X) is often referred the input layer and the output data (Y) is known to be the output layer. Between these two layers are the hidden variables that are gathered in one or more hidden layers. The nodes in the hidden layers can be thought of as sets of

intermediate variables similar to the latent variables in bilinear regression methods such as PLS and PCR. A representation of a feed-forward network with one output variable and one hidden layer is shown in Figure 4.7.

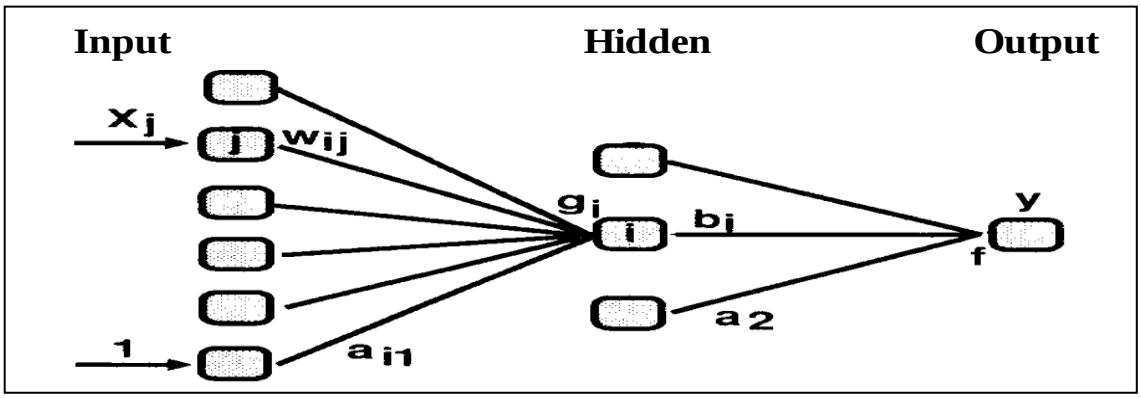


Figure 4.7. Basic illustration of a simple feed-forward network with one hidden layer and one output node.

The information from all input variables pass through to every nodes in the hidden layer and entire hidden nodes are linked to the single variable in the output layer in each case. The contributions from all nodes or elements are multiplied by constants and added prior to probable transformation occurs within the node. The transformation is usually a sigmoid function, on the other hand it can theoretically be any function. The sigmoid signal processing in a node is mentioned in Figure 4.8.

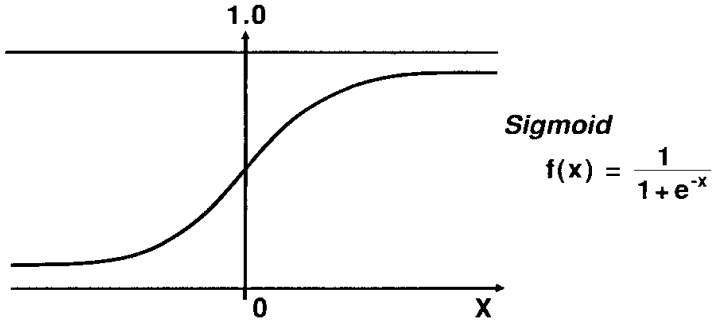


Figure 4.8. Representation of the signal processing in a sigmoid function

The feed-forward neural network in Figure 4.7 obtains a regression equation of the form

$$y = f\left[\sum_{f=1}^{I} b_i\, g_i\left(\sum_{f=1}^{J} w_{if}\, x_f + a_{f1}\right) + a_2\right] + e \tag{4.20}$$

whereby y is the output variable, the x's are the input variables, e is a random error term, g_i and f are functions and b_i, w_{ij} , a_{f1} and a_2 are constants to be determined.

The constants w_{ij} are the weights that each input element must be multiplied by before their contributions are added in node i in the hidden layer. In this node, the sum over j all elements $w_{ij}x_j$ is used as input to the function g_i. Then, each function g_i is multiplied by a constant b_i before summation over i and it is used as input for the function f. More than one hidden layer can also be used, resulting in a similar, but more complicated function. Note that for both the hidden and the output layer, there are constants, a_{i1} and a_2, respectively, that are added to the contribution from the rest of the variables before the transformations take place. These constants play the same role as the intercept (constant) term in a linear regression model (Barron et al. 1988).

As can be seen from equation given above, an artificial feed-forward neural network is simply a non-linear parametric model for the relationship between y and all the x-variables. There are functions g_i and f that have to be selected and parameters w_{ij} and b_i that must be estimated from the data. The process by which these parameters are determined is, in the terminology of artificial neural computing, called "learning". The best choice for g_i and f can in practice be found by trial and error, in that several options are tested and the functions that result in the best prediction ability are selected. (Martens et al. 1989)

As with linear calibration methods, network models must be constructed with consideration of two important effects: underfitting and overfitting. If a model that is too simple or too rigid is selected underfitting is the result, and if a model that is too complex is used, overfitting can be the consequence. The optimal model complexity is usually somewhere between these two extremes. (Barron et al. 1991)

4.4.4.3. The Backpropagation algorithm

The idea of the backpropagation algorithm is to reduce this error, until the ANN learns the training data. The training begins with random weights, and the goal is to adjust them so that the error will be minimal. The activation function of the artificial neurons in ANNs implementing the backpropagation algorithm is a_i weighted sum (the sum of the inputs x_i multiplied by their respective weights w_{ji}):

$$A_j(\overline{x}, \overline{w}) = \sum_{i=0}^{n} x_i w_{ji} \tag{4.21}$$

Noted that the activation depends only on the inputs and the weights. If the output function would be the identity (output=activation), then the neuron would be called linear. But these have severe limitations. The most common output function is the sigmoidal function.

$$Q_j(\overline{x}, \overline{w}) = \frac{1}{1 + e^{A_j(\overline{x}, \overline{w})}} \tag{4.22}$$

The sigmoidal function is very close to one for large positive numbers, 0.5 at zero, and very close to zero for large negative numbers. This allows a smooth transition between the low and high output of the neuron (close to zero or close to one) also; the output depends only in the activation, which in turn depends on the values of the inputs and their respective weights.

Now, the goal of the training process is to obtain a desired output when certain inputs are given. Since the error is the difference between the actual and the desired output, the error depends on the weights, and needs to adjust the weights in order to minimize the error function for the output of each neuron:

$$\boldsymbol{E_j(\overline{x}, \overline{w}, d) = (O_j(\overline{x}, \overline{w}) - d_j)^2} \tag{4.23}$$

Then, the square of the difference between the output and the desired target because it will be always positive, and because it will be greater if the difference is big and lesser if the difference is small. The error of the network will simply be the sum of the errors of all the neurons in the output layer:

$$E_j(\overline{x}, \overline{w}, \overline{d}) = \sum_j (O_j(\overline{x}, \overline{w}) - d_j)^2$$

(4.24)

The backpropagation algorithm now calculates how the error depends on the output, inputs, and weights. After we find this, we can adjust the weights using the method of gradient descendent:

$$\Delta w_{ji} = -\eta \frac{\partial E}{\partial w_{ji}}$$

(4.25)

This formula can be interpreted in the following way:

the adjustment of each weight (Δw_{ji}) will be the negative of a constant eta (η) multiplied by the dependence of the previous weight on the error of the network, which is the derivative of E in respect to w_i. This is, if the weight contributes a lot to the error, the adjustment will be greater than if it contributes in a smaller amount (4.25) is used until it is found appropriate weights (the error is minimal).

After that, first, calculate how much the error depends on the output, which is the derivative of E in respect j to O (4.24) and

$$\frac{\partial E}{\partial O_j} = 2(O_j - d_j)$$

(4.26)

Second, how much the output depends on the activation, which in turn depends on the weights (4.22) and (4.23)):

$$\frac{\partial O_j}{\partial w_{ji}} = \frac{\partial O_j}{\partial A_j} \frac{\partial A_j}{\partial w_{ji}} = O_j(1 - O_j)x_i$$

(4.27)

As it is seen in (from equation (4.26) and (4.27)):

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial O_j} \frac{\partial O_j}{\partial w_{ji}} = 2(O_j - d_j)O_j(1 - O_j)x_i$$

(4.28)

Therefore, the adjustment to each weight will be (from (4.25) and (4.28)):

$$\Delta w_{ji} = -2\eta(O_j - d_j)O_j(1 - O_j)x_i \tag{4.29}$$

Equation (4.29) as it is for training an ANN with two layers. Now, for training the network with one more layer to make some considerations. First adjust the weights (called them v_{ik}) of a previous layer, then calculate how the error depending not on the weight, but in the input from the previous layer. This is just done by just need to change x_i with w_{ji} in (4.27), (4.28), and (4.29). However, it is needed to see how the error of the network depends on the adjustment of v_{ik}. Thus,

$$\Delta v_{ik} = -\eta \frac{\partial E}{\partial v_{ik}} = -\eta \frac{\partial E}{\partial x_i}\frac{\partial x_i}{\partial v_{ik}} \tag{4.30}$$

whereby,

$$\frac{\partial E}{\partial w_{ji}} = 2(O_j - d_j)O_j(1 - O_j)w_{ji} \tag{4.31}$$

Then, assuming that there are inputs u_k into the neuron with v_{ik} (from (4.27)):

$$\frac{\partial x_i}{\partial v_{ik}} = x_i(1 - x_i)v_{ik} \tag{4.32}$$

If it is asked for adding yet another layer, same steps mentioned above would be performed, calculating how the error depends on the inputs and weights of the first layer. For practical reasons, ANNs implementing the backpropagation algorithm do not have too many layers, since the time for training the networks grows exponentially. In addition, there are refinements to the backpropagation algorithm, which allow a faster learning. (Rojas et al. 1996)

In principle, ANN can approximate any linear or non-linear dependence between the input and output data with an appropriate choice of its architecture (structure) and free parameters (weights). Therefore, ANN is one of the most effective techniques for non-linear data analysis in almost all fields of chemistry from quantum theory to petroleum chemistry (Balabin et al. 2008). The main disadvantage of the ANN approach is its computational complexity and stochastic nature (results of ANN training depend on initial parameters). It also requires a much larger data set for training.

4.4.5. Principal Component Analysis (PCA)

Principal component analysis is a full spectral and soft modeling method which is based on the decomposition of data matrix into two separate and smaller matrices. These two kinds explain the relationships between the variables and the relationships between the objects. In addition, this division makes the dimensionality reduction for the large data matrix. (Kowalski et al. 1983) For instance, spectral data contain hundreds of wavelengths with their corresponding absorbance values and it is hard to visualize this data matrix in hundreds of dimensionality. As this, it is not possible for dimensions larger than three, generally pictures or graphs that are used to explain the distributions of samples or variables should have three or less dimension in a space.

4.4.5.1. Singular Value Decomposition (SVD)

Singular value decomposition (SVD) and nonlinear iterative partial least squares (NIPALS) are most commonly used algorithms in PCA analysis. In this study, SVD based principal component analysis was used. In this algorithm, the training set A with *m* samples and *n* variables is decomposed into the principal component scores (U), matrix of singular values (S), and V matrix whose rows are eigenvectors of A. Equation 1 shows the mathematical expression of SVD. As it seen from the equation, the singular values matrix of S is a square diagonal matrix that has elements are different from zero on diagonal. Eigenvalues of corresponding training set are calculated using the singular value matrix. The larger the eigenvalue is the more significant information. Generally, the principal components (PC) are calculated according to this significance.

$$\mathbf{A}_{\mathbf{mxn}} = \mathbf{U}_{\mathbf{mxm}}\mathbf{S}_{\mathbf{mxn}}\mathbf{V}^{\mathbf{T}}_{\mathbf{nxn}} \tag{4.33}$$

Often the Equation 4.33 is given in only two matrices that is shown in below:

$$\mathbf{A}_{\mathbf{mxn}} = \mathbf{T}_{\mathbf{mxn}}\mathbf{V}^{\mathbf{T}}_{\mathbf{nxn}} \tag{4.34}$$

where $\mathbf{T}$ $(= \mathbf{U}_{\mathbf{mxh}}\mathbf{S}_{\mathbf{hxn}})$ is the *score matrix* and proportional to the size of the training set contains the information about the objects, $\mathbf{V}^{\mathbf{T}}$ is the loading matrix that has the knowledge of variables. Each row of original data matrix is linear combinations of

loading vectors. The first PC generally is the best straight line in multidimensional space (Brereton 2003).

As it is mentioned before, multidimensional data contains information of the variables of samples or objects. PCA generally uses not only all the wavelengths in the spectra but also the variables that extracted from the spectral measurements. When data reduction term is used in PCA analysis, it means variable selection is done but, all the wavelengths in the spectra are used in the explanation of the relationships of variables. For the best selection, one can also need a reduction in the wavelength selection. As a result, the data interpretation of objects is done with the most useful wavelengths and their corresponding variable and the relationship between the samples can be seen clearly. GA are used for wavelength selection in this case as in the calibration part.

CHAPTER 5

MATERIALS AND METHODS

5.1. Experimentation

In this study, in order to construct calibration models three diverse set consisting ternary mixtures of vegetable oils, biodiesel produced from these vegetable oils, methanol set and vegetable oils – biodiesels – diesel set besides this, as a quaternary mixtures of vegetable oils – biodiesel – methanol – diesel set were prepared.

Commercially available sunflower, canola (from Tansaş Inc.), cottonseed oils (from Diasa Inc.) were used as the fatty acid feedstock. Homogeneous transesterification reaction to produce biodiesel in laboratory was carried out using sodium hydroxide (NaOH) and methanol (MeOH). Acetic acid was used to remove the unreacted NaOH from the biodiesel (Umdu 2008).

5.1.1. Biodiesel Synthesis

Commercially available sunflower, canola, cottonseed oils were used and transesterification was carried out under reflux condenser to avoid methanol loses and also the temperature was kept constant by a thermostatic bath during the reactions. The biodiesel synthesis includes the steps following:

i. Sodium hydroxide was added to methanol at room temperature and stirred at 1100 rpm for 10 min to form methoxide

ii. In hot water bath, temperature was increased up to 50^{o}C and maintained.

iii. Vegetable oil were added at 50^{o}C and left the reaction medium for 4h stirring at same speed and maintaining temperature constant in order to get totally conversion of vegetable oils to methyl esters (biodiesel).

iv. After 4h, two separated phases (bottom phase is yellow and dense, glycerol) were observed. These phases were rinsed with 5% (v/v) acetic acid whose amount was the 1/3 volume of the medium.

v. After acetic acid addition, stirring continues at 50^{o}C with 500 rpm. At this time aqueous phase is separated from methyl ester phase. Then, its pH is measured and rinsing process is continued until the pH of aqueous phase is equal to pH of acetic acid.

vi. Finally, methyl esters solution exposed to centrifuge process at 5000 rpm within during 10 min in order to get rid of supernatants. Then, it is left in the rotary evaporator instrument at 40^{o}C under the 100 mbar vacuum pressure to remove all water content in solution.

All biodiesels were synthesized according to the procedure mentioned above and store at +4 ^{o}C until the analysis time.

5.2. Instrumentation

After the biodiesels were synthesized, gas chromatographic analysis as a reference analysis method were performed in order to confirm the conversion of all vegetable oils to methyl esters by determining the FAME percentages of biodiesels by mass.

5.2.1. Gas Chromatography

The synthesized methyl esters were analyzed using a GC – 2010 (Shimadzu) instrument installed with FID detector TRB-WAX capillary column with a 60 m column length, 0.25 µm ID x 0.25 µm polyethylene glycol. Instrument parameters are optimized after some trials performed.

Table 5.1. Specific parameters of gas chromatography instrument.

Inlet temperature	250 ^{0}C
Split ratio	1:50
Volume of injection	1 µL
Column flow (N_2)	1.63 mL/min
FID temperature	250 ^{0}C
H_2 flow	40 mL/min
Dry air flow	400 mL/min
Make up flow (N_2)	30 mL/min
Oven temperature program	Column temperature: from 150 ^{0}C to 210 ^{0}C with 10 ^{0}C/min increment. Maintain at 210 ^{0}C for 5 min. Then, with 5 ^{0}C/min increment goes up to 230 ^{0}C and maintain at 230 ^{0}C for 25 min.
Solvent	Methanol

Optimization process was performed by using a FAME standard (F.A.M.E. Mix RM-3, O7256-1AMP, Supelco) solution which includes the fatty acids of C14:0 (0.996%), C16:0 (0.37%), C18:0 (2.990%), C18:1 (44.853%), C18:2 (14.936%), C18:3 (3.187%), C20:0 (3.008%), C22:0 (2.986%), C22:1 (19.992%) and C24:0 (3.013%) by mass percentages. For each of FAME, from six different points different calibration models were constructed. Standard solutions were prepared by diluting the 100mg FAME sample to 25 ml methanol and by using this 4000 mg/L feedstock solutions, 100, 500, 1000, 1500, 2000 and 2500 mg/L solutions for calibration models.

Approximately 100mg for each biodiesel samples (103.0 mg sunflower oil, 98.0 mg cottonseed, 104.8 mg canola oil) were diluted to 25 ml methanol. Since each of FAME solutions have different fatty acids content by mass percentages. For instance, oleic acid methyl ester (C18:1) content in biodiesel produced from canola oil is approximately 60% by mass whereas 30% by mass for biodiesel produced by cottonseed oil. For this reason, biodiesel solutions were diluted to ratio of 1:2, 1:4 and 1:8 again to calculate FAME percentages by mass from calibration plots. (Results are shown in Chapter 6.)

5.2.2. Infrared spectroscopy

In this project, infrared spectra were collected with mid infrared (MIR) and near infrared (NIR) spectrometers. Near-infrared spectroscopic analyses were performed with FTS-3000 NIR spectrometer (Bio-Rad, Excalibur, Cambridge, MA) and mid-infrared spectroscopic analyses were performed with Spectrum 100 FTIR spectrometer (Perkin Elmer, Waltham, MA). Configurations of the spectrometers are shown in below.

Table 5.2. Instrumental parameters used in the spectrometric analyses.

	NIR spectrometer	**MIR spectrometer**
source	tungsten-halogen lamp	nichrome wire
beam splitter	calcium fluoride	extended range KBr
detector	lead selenide	FR-DTS
resolution	16 cm^{-1}	4 cm^{-1}
# of scans	128	4
# of data points	780	3601
range	10,000 – 4,500 cm^{-1}	4,000 – 600 cm^{-1}

5.3. Data Analysis

The collected spectra were transferred in ASCII file format and were combined with Microsoft Excel program. Then, data files for multivariate analyses were prepared as text files. Genetic algorithm based calibration methods were written in MATLAB programming language Version 7.0 (MathWorks Inc., Natick, MA) and artificial neural networks approach calibration modeling data analysis were performed by Neural Network Toolbox GPU in Matlab.

However, prior to ANN analysis, since spectral data have large number of values, principal component analysis (PCA) applied with SVD algorithms as it is mentioned in Chapter 4. Also in Matlab 10 score matrices vector (PC) at which the variances lay upper 90% percentages. Thus, data matrix consists of 10 variables. Due to nature of sigmoid function, data have to be lies between 0 1 and 0.9. To do this, equation given below is used:

$$\boldsymbol{X_i = 0.1 + 0.8 * (X_i - X_{min_i})/(X_{max_i} - X_{min_i})} \quad (5.1)$$

whereby, X_{mini} and X_{maxi} represents the minimum and maximum values within the column, $X_{i,}$ respectively.

Data matrices for calibration and validation sets were arranged with feed forward back propagation algorithm which is mentioned detail given in Chapter 4. To do this, a single hidden layer consisting Log sigmoid (activation function), Traindgm (training rule), 10 neurons in a hidden layer and a goal error of 10^{-4} were utilized for optimization of the network. The constructed artificial neural network model can be seen in the Figure 5.1.

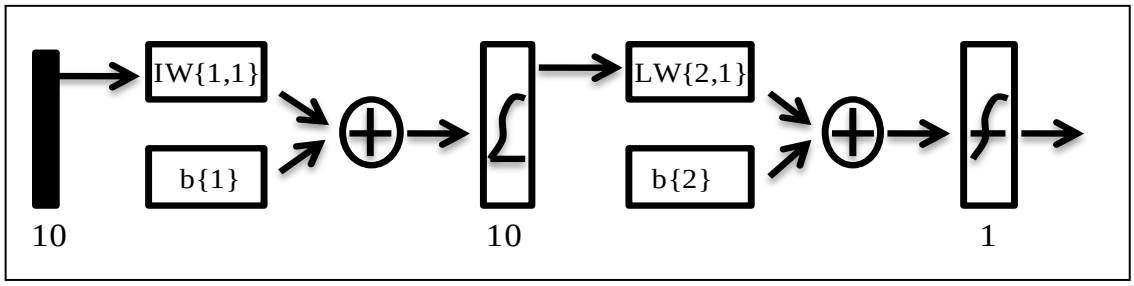


Figure 5.1. The schematic illustration of the prepared ANN block diagram.

CHAPTER 6

RESULTS AND DISCUSSION

In this study, two different supervised calibration methods, which are based on genetic algorithm least squares (GILS) and artificial neural networks (ANN) approach, were used. Both methods were examined in various spectral data which were obtained by near infrared (NIR) spectral and Fourier transform infrared (FTIR) spectral data matrices with respect to the measurement of different types of vegetable oil samples, biodiesel produced from these vegetable oils, methanol and diesel samples. In this chapter, all the calibration results will be discussed in detail for both supervised methods and all the sample types. Prior to discuss the infrared spectroscopic results along with chemometric methods, GC analysis results are conferred in order to investigate the biodiesel synthesis and FAME analysis.

6.1. Gas Chromatograms

GC analysis were performed in order to prepare calibration plots for FAME solutions at which 2500 mg/L standard solutions used for this purpose and chromatogram of this type of solution is shown in Figure 6.1. As can be seen from chromatogram, 10 peaks, which can be characterized and measurable, observed. The retention times that belong these peaks displayed at Table 6.1.

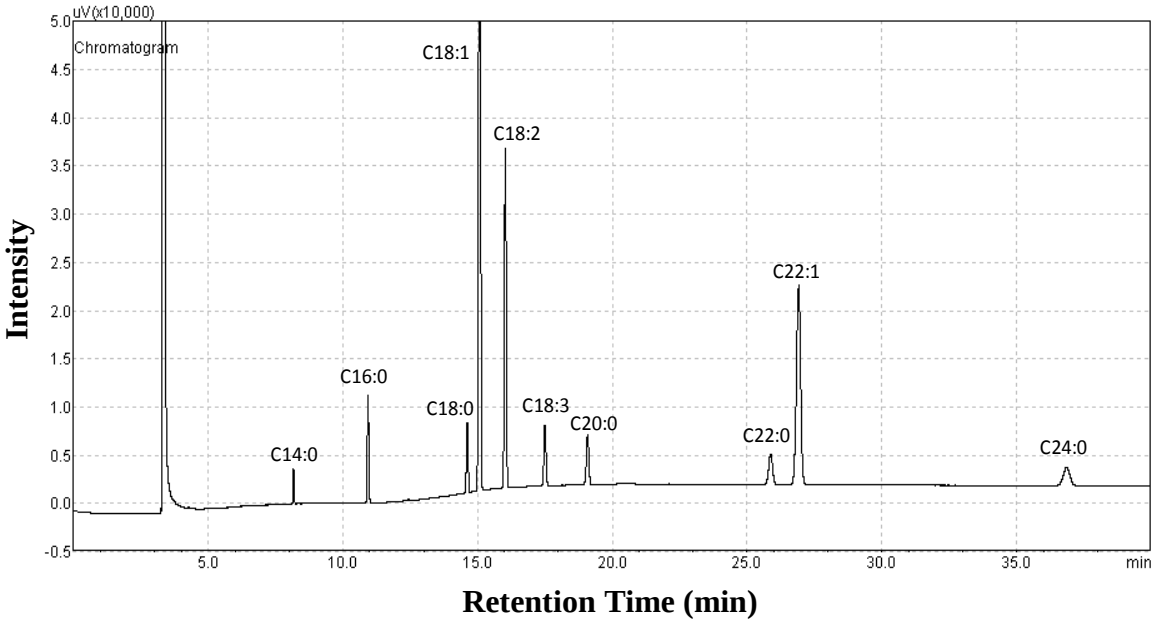


Figure 6.1. FAME standard solution chromatogram.

As can be seen from this chromatogram, primarily methanol leaves column at 4.6 min and then, fatty acid methyl esters according to their carbon number reaches detector. In addition, double bond numbers affect the retention time and the more double bond in structure the longer the retention times. In addition, there is no overlaping peaks on the chromatogram.

Table 6.1. Detailed information about FAME standard solution and retention times.

Fatty acid methyl esters (FAME)	**Retention Times (min)**
Myristic acid methyl ester, C14: 0	08.17
Palmitic acid methyl ester, C16: 0	10.94
Stearic acid methyl ester, C18: 0	14.62
Oleic acid methyl ester, C18: 1	15.90
Linoleic acid methyl ester, C18: 2	16.40
Linolenic acid methyl ester, C18: 3	17.51
Arachidic acid methyl ester, C20: 0	19.90
Behenic acid methyl ester, C22: 0	25.87
Erucic acid methyl ester, C22: 1	26.90
Lignoseric acid methyl ester, C24: 0	36.89

The values for percentages of FAME standard solution by mass which retention times were determined. FAME concentrations in solution were calculated and this concentration values were connected to the peak areas to construct calibration plots. Table 6.2 illustrates the values used in calibration model.

Table 6.2. Concentration values related to standard calibration solutions.

	STD1 (mg/L)	STD2 (mg/L)	STD3 (mg/L)	STD4 (mg/L)	STD5 (mg/L)	STD6 (mg/L)
C14:0	0.99	4.98	9.96	14.94	19.92	24.90
C16:0	4.04	20.18	40.37	60.55	80.74	100.92
C18:0	2.99	14.95	29.90	44.85	59.80	74.75
C18:1	44.85	224.26	448.53	672.79	897.06	1121.32
C18:2	14.94	74.68	149.36	224.04	298.72	373.40
C18:3	3.19	15.93	31.87	47.80	63.74	79.67
C20:0	3.01	15.04	30.08	45.12	60.16	75.20
C22:0	2.99	14.93	29.86	44.79	59.72	74.65
C22:1	19.99	99.96	199.92	299.88	399.84	499.80
C24:0	3.01	15.06	30.13	45.19	60.26	75.32

From Table 6.2, concentration values, which belong to the FAME components, were between 0.99 mg/L and 1121 mg /L. Figure 6.2 displays the calibration plots prepared by GC analysis of these standard solutions.

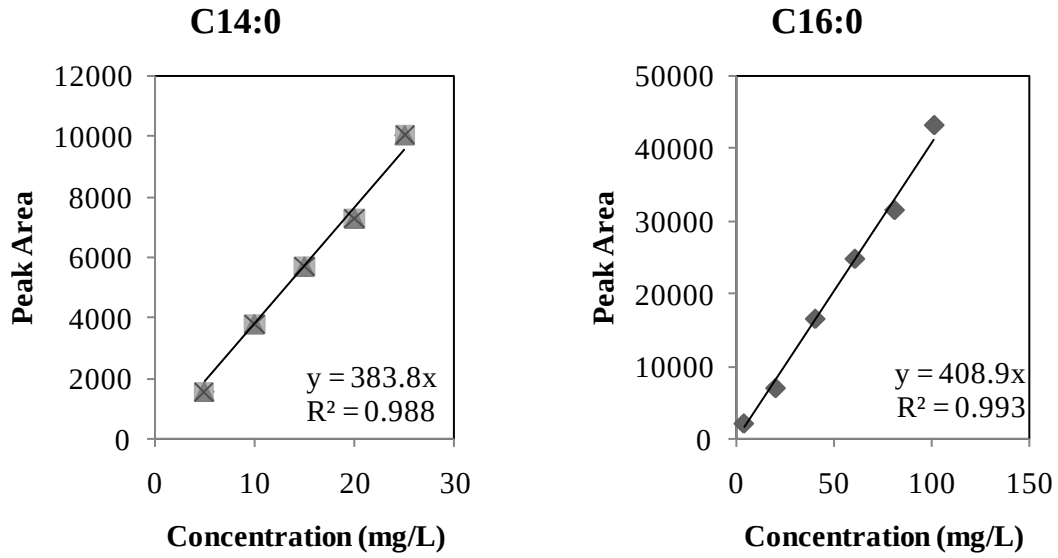


Figure 6.2. Standard calibration plots prepared by GC analysis for FAME components.

(cont. on next page)

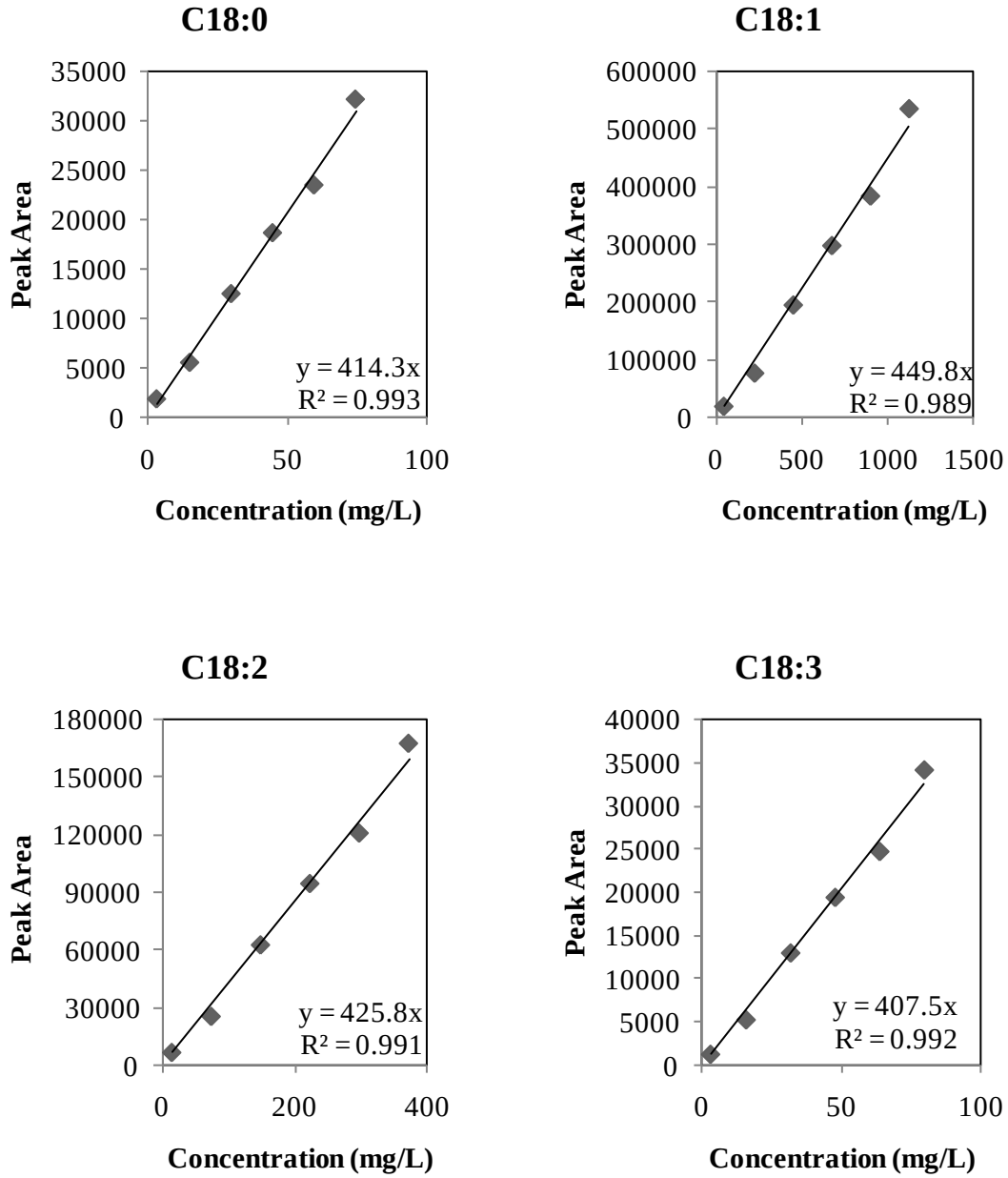


Figure 6.2. (cont.)

(cont. on next page)

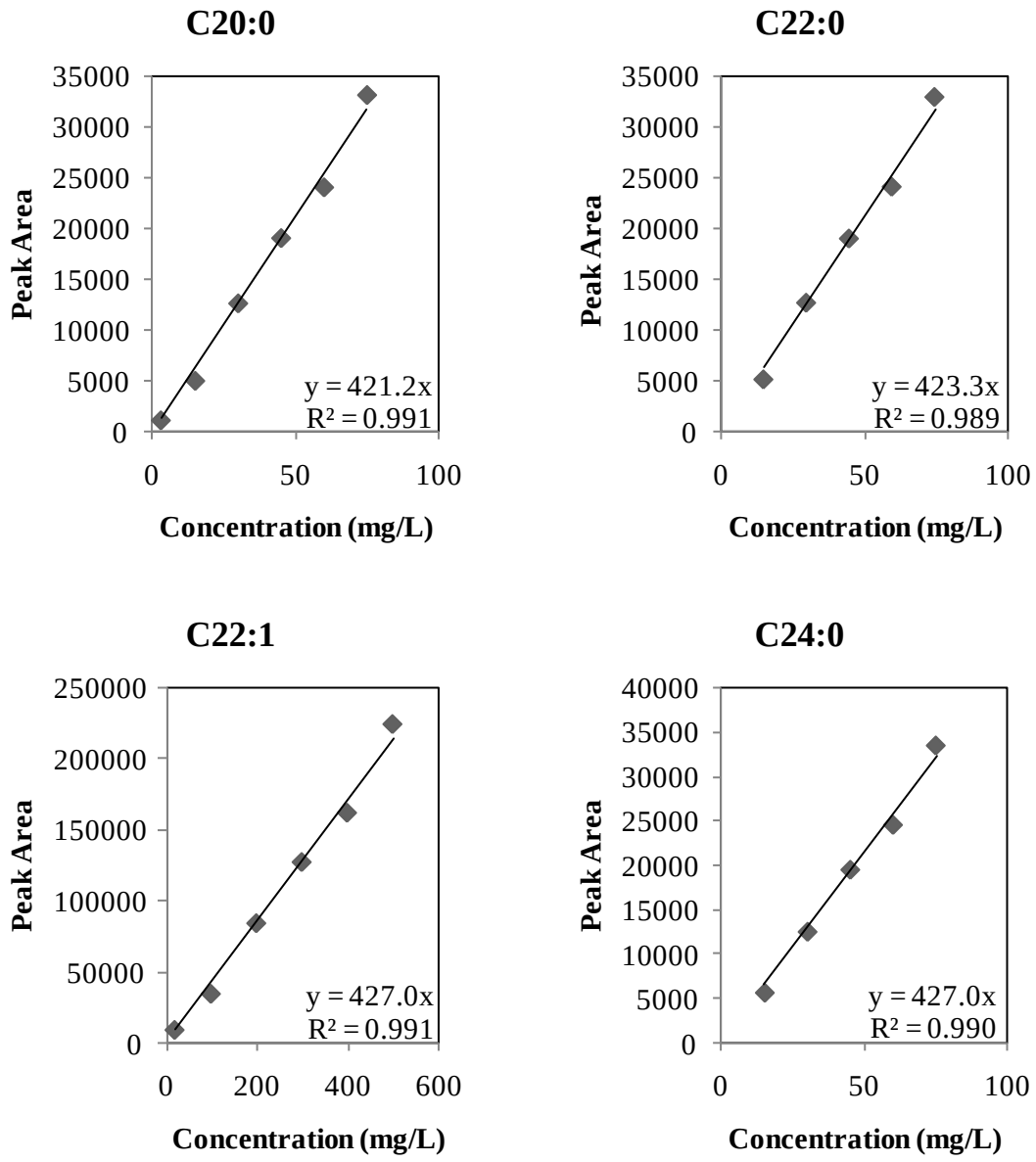


Figure 6.2. (cont.)

According to the Figure 6.2, regression values are between 0.98 and 1.00. However, there is no significant peak at the lowest concentration value for C14:0, C22:0 and C24:0 components, therefore, calibration plots have 5 points for these fatty acids. These calibration plots used for the FAME percentages of the synthesized biodiesels samples. Figure 6.3 illustrates the GC chromatogram of solutions prepared in 4 g/L methanol with synthesized biodiesels from three different types of vegetable oils.

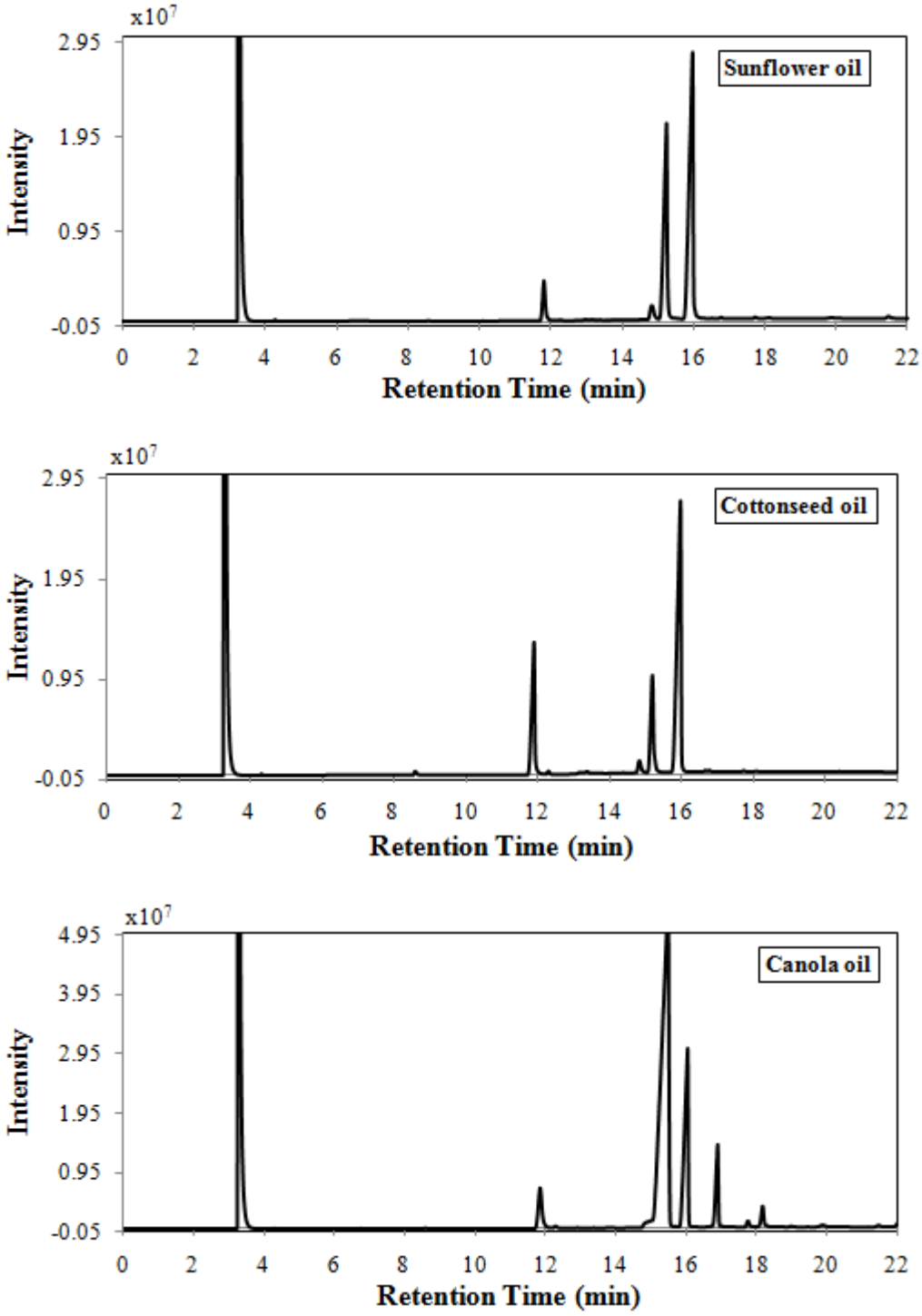


Figure 6.3. GC chromatogram of synthesized biodiesels from sunflower oil, cottonseed oil and canola oil, respectively.

As can be seen from chromatograms, sunflower oil consists of more C16:0, C18:0, C18:1 and C18:2 rather than the cottonseed and canola oils whereas cottonseed oil includes C18:3 and C20:0 fatty acid methyl esters as well. However, from

chromatogram of canola oil, it is seen that canola oil has C22:0, C22:1 and lower amount of C24:0 fatty acid methyl esters apart from the mentioned earlier.

GC analysis were performed for biodiesels from three different vegetable oils with respect to 20 samples, which were diluted to the ratio of 1:1, 1:2, 1:4, 1:8, and calibration plots were obtained from the software installed in GC automatically. Outlier values were not considered and FAME percentages by mass, which were proper for calibration plot for dilution factors, displayed in the Table 6.3.

Table 6.3. FAME percentages by mass from synthesized biodiesel using sunflower oil, cottonseed oil, canola oil.

FAME components	Biodiesel synthesized from sunflower oil (w/w %)	Biodiesel synthesized from cottonseed oil (w/w %)	Biodiesel synthesized from canola oil (w/w %)
C14:0	0.00	0.00	0.00
C16:0	6.74	11.47	4.97
C18:0	3.21	1.89	2.16
C18:1	35.45	27.91	64.79
C18:2	53.22	56.83	19.55
C18:3	0.17	1.06	6.27
C20:0	0.29	0.43	0.64
C22:0	0.68	0.16	0.27
C22:1	0.00	0.00	1.16
C24:0	0.23	0.25	0.18

After obtaining these values, they are compared with literature (Table 6.4, Knothe et al. 1997) at which there is wide range within the values and it is pointed out that they were similar to each other.

Table 6.4. FAME percentages by mass found in sunflower, cottonseed and canola oils (Source: Knothe et al. 1997)

	Sunflower Oil (w/w%)	Cottonseed Oil (w/w%)	Canola Oil (w/w%)
C14:0			
C16:0	3.50-6.50	7.00-13.0	4.00-5.00
C18:0	1.30-5.60	2.50-3.00	1.00-2.00
C18:1	14.0-43.0	30.50-43.0	55.0-63.0
C18:2	44.0-68.7	39.0-52.0	20.0-31.0
C18:3		1.00	9.0-10.0
C22:1			1.00-2.00

Consequently, construction of GC models for the synthesized biodiesel and production of biodiesels were completed successfully. In addition, this indicates the availability of the models for the determination of FAME percentages by mass in biodiesels.

6.2. Ternary mixtures of Biodiesel – Vegetable oil – Methanol Set

As stated before, the transesterification monitoring is an important issue to biodiesel quality control since some contaminants arise from this reaction also; such monitoring allows recognizing and correcting problems at an early stage. Generally, it is unfeasible to distinguish the infrared spectra of the mixture of vegetable oils and biodiesel produced from these oils especially because these species have almost the same chemical properties; it is incredibly difficult to select the wavelengths related to them in a spectroscopic study. However, there is no complexity with respect to chemometric studies.

All samples whose concentrations by mass percentages displayed in Table 6.5 and Table 6.6 were analyzed using NIR and MIR spectrometers and the data collected for the prediction. Each set corresponding to the transesterification reaction divided into two sets: one was for calibration and the other was for validation. Calibration set contained 27 samples spectra and validation set contained 12 samples spectra at which the samples used in calibration or validation set arranged in a random order. The range of concentrations each constituent in the sets were in the range of 0 – 100 % for both biodiesel and vegetable oil, 0 – 20 % for alcohol.

Table 6.5. Concentration profiles for calibration set of ternary mixture of biodiesel, sunflower oil, methanol set. All concentrations are given by mass percentages.

Sample No	Biodiesel (w/w%)	Sunflower Oil (w/w%)	Methanol (w/w%)	Sample No	Biodesel (w/w%)	Sunflower Oil (w/w%)	Methanol (w/w%)
1	79.87	20.13	0.00	15	83.24	6.15	10.61
2	75.66	23.60	0.75	16	78.28	10.75	10.97
3	67.78	30.28	1.94	17	70.10	17.93	11.97
4	63.77	33.73	2.50	18	65.71	21.75	12.54
5	55.76	40.93	3.31	19	56.42	30.37	13.21
6	48.94	47.38	3.68	20	64.67	19.91	15.42
7	44.31	50.66	5.03	21	73.45	10.47	16.09
8	39.81	54.89	5.29	22	75.64	7.97	16.40
9	32.29	61.72	5.99	23	20.09	79.91	0.00
10	65.40	27.62	6.98	24	85.11	0.00	14.89
11	72.10	20.05	7.94	25	0.00	0.00	20.00
12	74.59	15.66	9.75	26	0.00	100.00	0.00
13	81.90	8.20	9.90	27	100.00	0.00	0.00
14	86.66	4.15	9.19				

As it is seen in Table 6.5 calibration data set includes the pure form of sunflower oil, biodiesel, and methanol also.

Table 6.6. Concentration profiles for validation set of ternary mixtures of biodiesel, sunflower oil, methanol set. All concentrations are given by mass percentages.

Sample No	Biodiesel (w/w%)	Sunflower Oil (w/w%)	Methanol (w/w%)	Sample No	Biodiesel (w/w%)	Sunflower Oil (w/w%)	Methanol (w/w%)
1	60.27	37.10	2.63	7	87.50	2.25	10.25
2	47.73	47.91	4.36	8	74.01	13.80	12.19
3	36.07	58.16	5.78	9	60.28	26.63	13.10
4	69.19	23.75	7.06	10	77.53	7.03	15.45
5	79.05	12.12	8.83	11	71.62	11.05	17.33
6	87.79	2.07	10.15	12	73.34	6.39	20.27

6.2.1. Near Infrared Analysis

In the NIR spectral region, the absorbance bands are often broad and overlapping. Figure 6.4 demonstrates the near infrared spectra of biodiesel, sunflower oil, methanol and their ternary mixture between 4000-600 cm^{-1}.

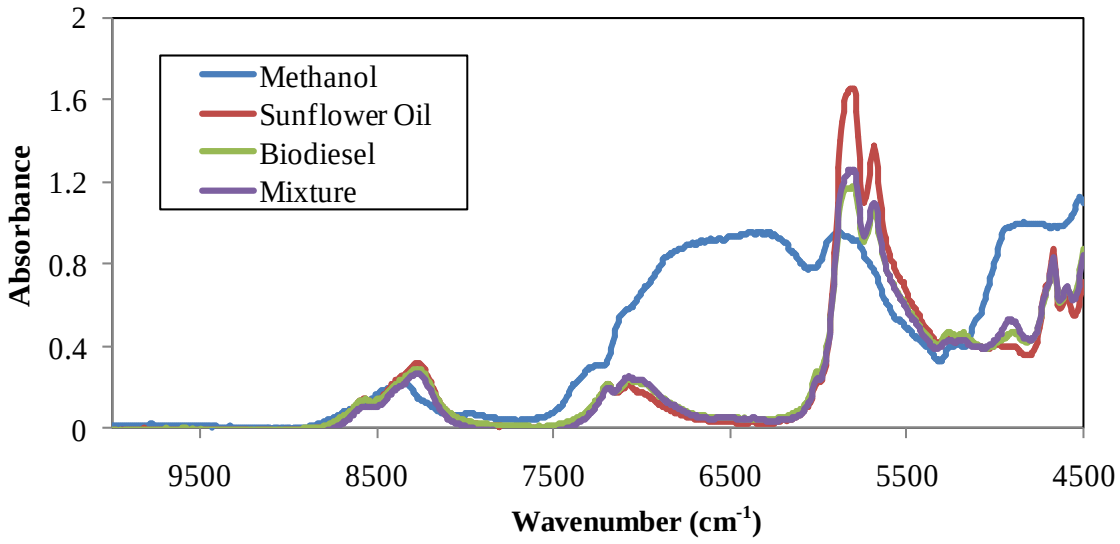


Figure 6.4. FT-NIR spectra of biodiesel, sunflower oil, and methanol along with their ternary mixture

As it is seen from these spectra, it is obvious that each constituent apart from methanol exhibits very similar spectral characteristics, which makes it necessary to use a multivariate calibration method to resolve the mixtures of these compounds. As can be seen in Figure 6.4, spectra show NIR absorption bands corresponding to the -CH=CH asymmetric stretching and C=C stretching at around 4600 cm^{-1}. Also, =C-H and C=C stretching belong to –CH=CH- group. The spectra show maximum absorbance at 5800 cm^{-1}. This maximum absorption band is related to first overtone of C-H bond which belongs to $–CH_2$ functional group while absorption band (only a shoulder) at 5680cm^{-1} is the first overtone of -C-H bond. Besides this, C-H stretching vibrations belongs to weak and broad $-CH_3$ functional group is seen at 7200 cm^{-1} whereas, the other weak and broad shoulder peak at around 7110 cm^{-1} indicates the C-H stretching which leads from CH_2 functional group. Another important peak in spectra is observed at 8285 cm^{-1} that belongs to C-H stretching and represents second overtone.

If the matrices of vegetable oils and biodiesels synthesized from these oils are considered, the NIR spectral changes that result from the varying concentration of the compounds in the transesterification reaction mixture are difficult to interpret visually. Even though gas chromatographic analysis has been used for determination of these mixtures, these are most abundantly time and cost consuming techniques as well. Thus, multivariate calibration techniques are preferred.

6.2.1.1. GILS Results

For each sample of both calibration and validation set, NIR absorbance spectral data matrices obtained and calibration models were constructed in terms of each component in ternary mixture, and then tested with validation set. GILS program run against 100 times along with 50 iterations and 30 genes. Figure 6.5 shows the actual sunflower oil, biodiesel, and methanol concentration values versus their GILS predicted concentration values based on FT-NIR spectral data. The standard error of calibration (SEC) values of each component were found between 0.89% (w/w) and 1.86% (w/w) and the standard error of prediction (SEP) values of each component were found between 0.91% (w/w) and 2.88% (w/w) by using GILS method for all components in set. SEC and SEP values obtained as 1.21% (w/w) and 2.75% (w/w) for sunflower oil and 0.89% (w/w) and 0.91% (w/w) for methanol, respectively. Also, R^2 values of regression lines for sunflower oil and methanol is around 0.997 but is 0.994 for biodiesel samples. Therefore if R^2, SEC and SEP values are examined, it is seen that values are compatible with each other, which demonstrates a good prediction for rapid monitoring the transesterification reaction of biodiesel synthesis to investigate production facilities at which the methanol was up to 20% by mass in medium.

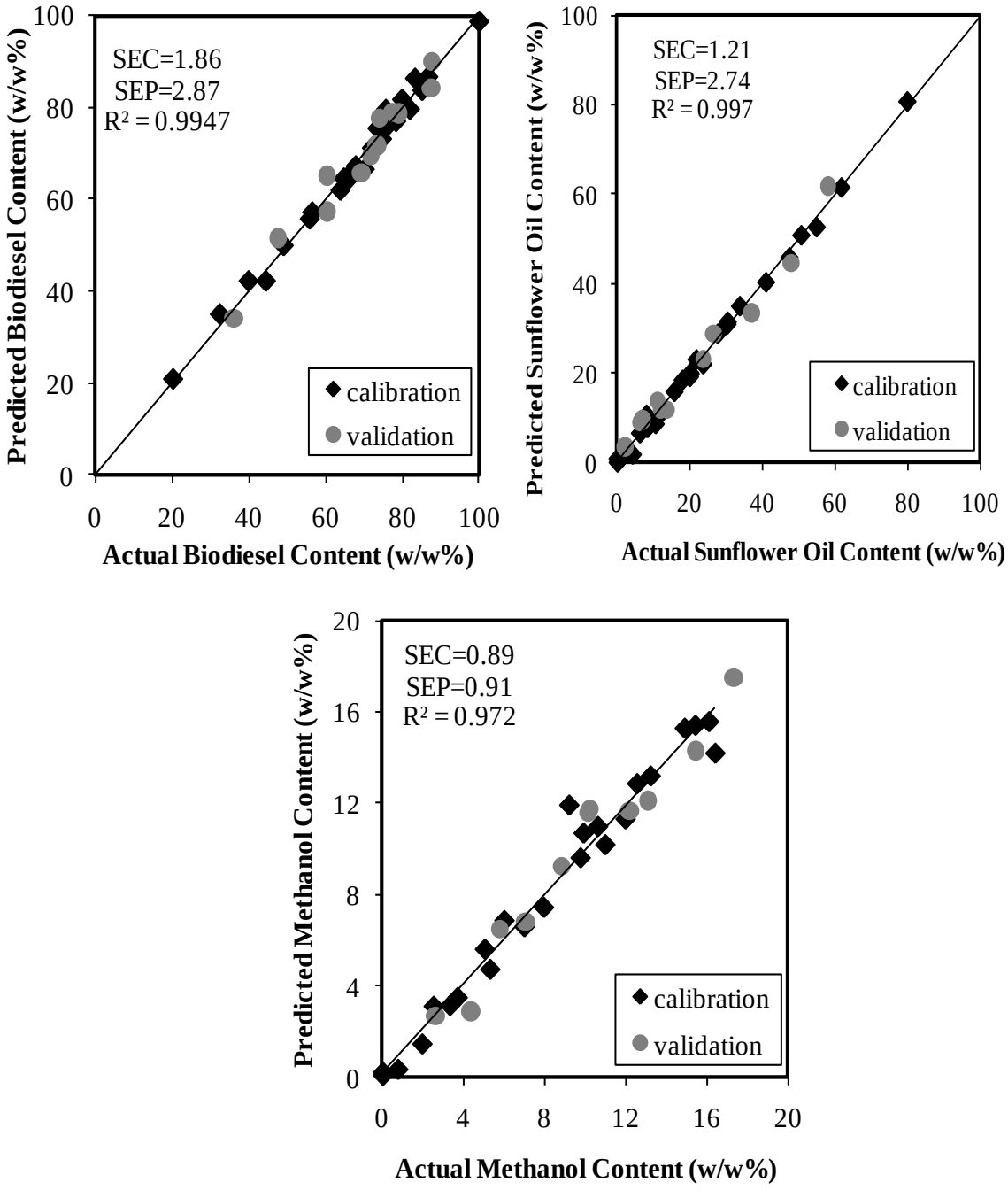


Figure 6.5. Actual versus predicted concentration plot for NIR data analysis obtained with GILS calibration method.

Since GILS is a method that based on wavelength selection, it is important to examine the distribution of selected wavelengths in multiple runs over the entire full spectral region. Figure 6.6 displays the frequency distribution of selected wavelengths in 100 runs with 30 genes and 50 iterations for biodiesel-sunflower oil-methanol set.

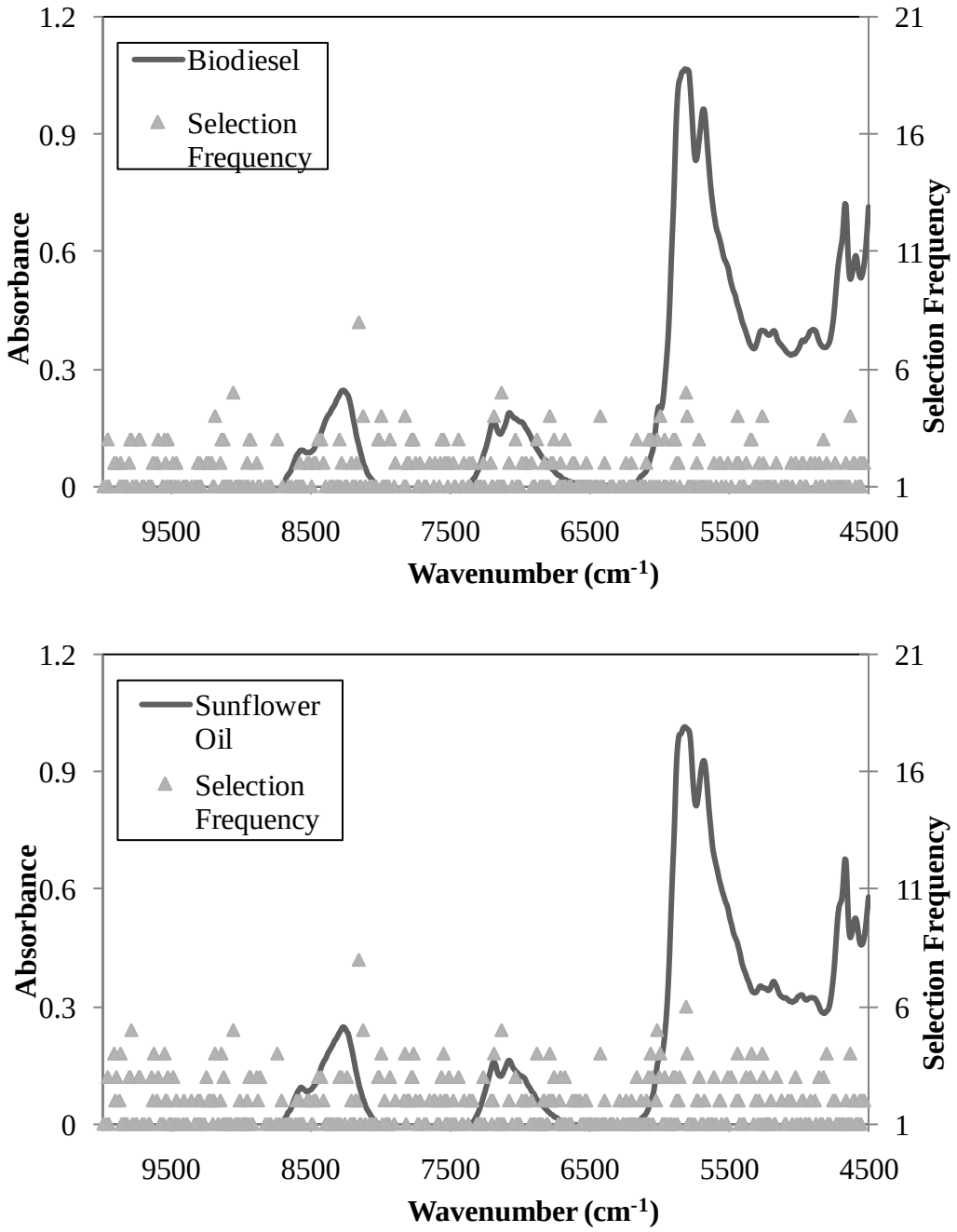


Figure 6.6. Wavelength selection frequency distribution of GILS method for ternary mixture of biodiesel-sunflower oil-methanol using NIR spectroscopy.

(cont. on next page)

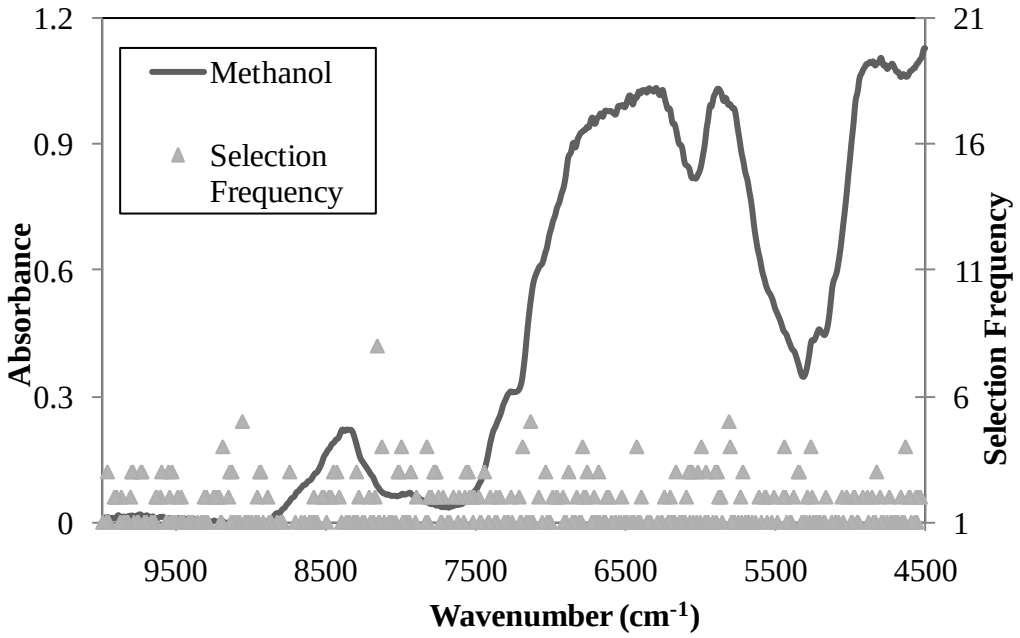


Figure 6.6. (cont.)

As can be seen from the figure there are a number of regions where selection frequencies are highly compared to the rest of the spectrum. The wavelength region around 5000 cm^{-1} for biodiesel and sunflower oil indicates a strong tendency for GILS method to select while for methanol content, around 7000 and 8500 cm^{-1} is the most frequently selected region. A significant difference between the frequency distribution of biodiesel and oil content is that the selected wavelengths are more distributed in the former and much more wavelengths are selected. This is a strong indication that the genetic algorithm incorporated GILS method is focusing on the regions where the most concentration related information is contained.

6.2.1.2. ANN Results

The ternary mixture of biodiesel-sunflower oil-methanol set was used for the construction of artificial neural network calibration models and optimization processes were performed with respect to their NIR spectral data matrices. After some trials to construct ANN calibration model, as an optimum conditions, learning rate and momentum value selected as 0.8 and 0.7, ”respectively”. Even though these values seem to be large for this type of model at first glance, mentioned in Chapter 5 in subsection of ANN, due to the fact that infrared spectra data have large number of data points in spectra matrices. Therefore, prior to ANN analysis, data reduction was performed with PCA-SVD algorithms to the full spectral data matrices. In addition, to calculate the total error, mse, which represents the term mean square error, applied with a goal of 0.0001 up to maximum value of 10000 epochs, which is a weight vector or iteration number.

After launching Neural Network Toolbox in Matlab 7.0.1 software programming in terms of the configurations as mentioned above; the user interface results for each component of ternary mixture modeled by ANN calibration are given in the Figure 6.7.

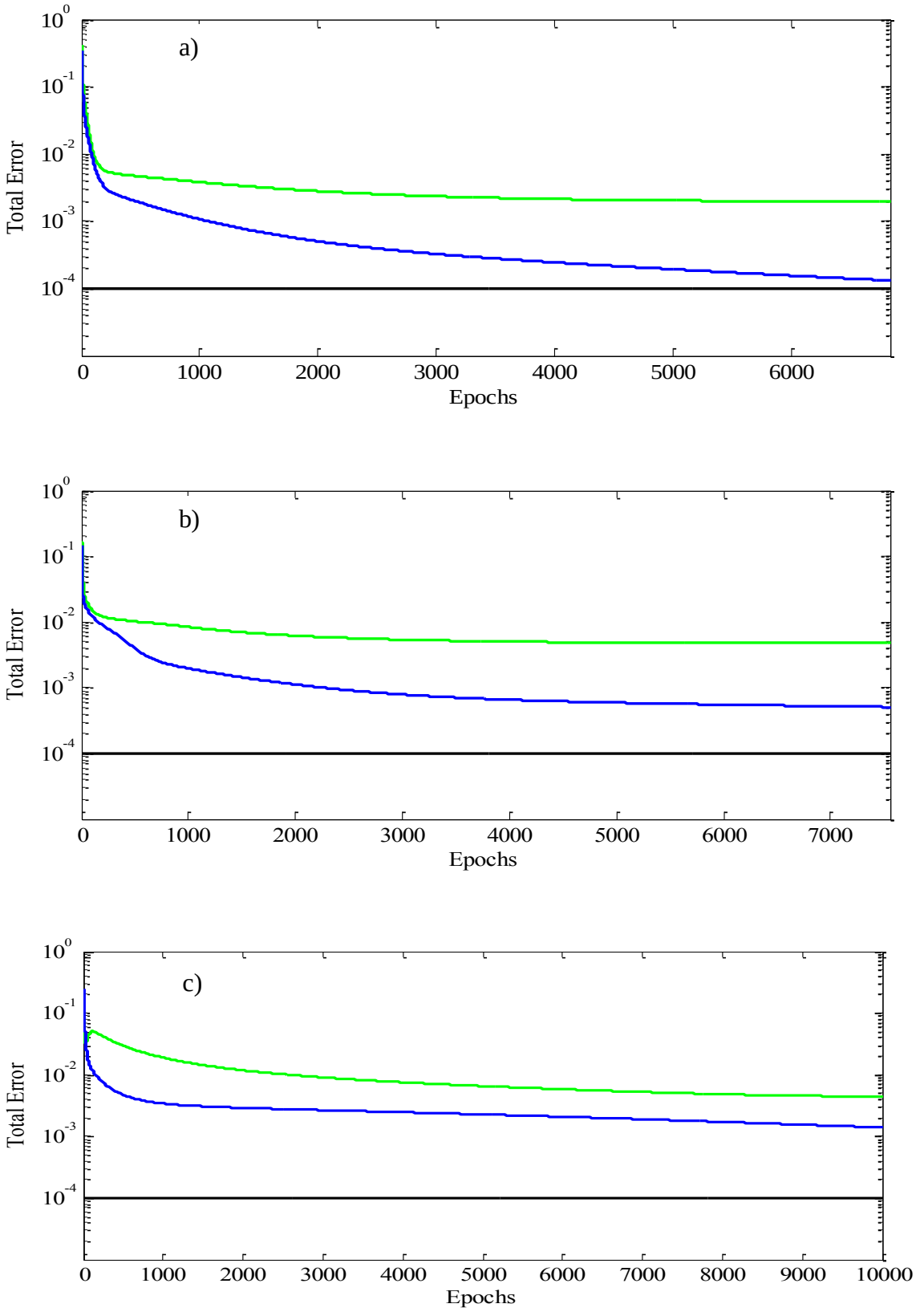


Figure 6.7. Error versus epochs plot of ternary mixture of a) sunflower oil b) biodiesel c) methanol (green line: validation, blue line: calibration, black line: target).

As can be seen from the Figure 6.7, for sunflower oil after 6834 iterations performed, the mse value reached about $13x10^{-5}$ for calibration, on the other hand, due to increasing of mse value of validation program was stopped. For biodiesel, 7556 iterations were necessary to achieve about $5x10^{-4}$ mse value and begin validation mse error again. For methanol, epochs reached the limit number, thus program stopped at which the performance was $13x10^{-4}$ mse value.

For a stable case, after few iterations total error value should begin to decrease sharply. As can be seen from Figure 6.7, total error value start to diminish immediately after program run and it is only mentioned to have an idea about the model apart from whether the attainment of the target value is get or not. If the calibration and validation lines are close to target line in a close way, it indicates that the calibration model constructed in well and total error is minimized. However, output data is required to revise from the 0 – 1 interval to the actual value since log-sigmoid function obtains results in from 0.1 to 0.9. After these conversion applied, it is possible to calculate the SEC and SEP values along with the regression values between actual and predicted by ANN model.

Furthermore, Figure 6.8 illustrates the R^2 values found between 0.96 and 0.99. In addition to this, SEC values obtained as 0.99% (w/w) for methanol, is 2.93% (w/w) for biodiesel and is 1.49% (w/w) for sunflower oil whereas SEP values is 1.69% (w/w) for methanol, is 8.63% (w/w) for biodiesel and is 5.58% (w/w) for sunflower oil. Thus, these values mention that the constructed model is applicable for calibration of biodiesel blends to determine the methanol and untreated vegetable oil content via ANN multivariate calibration modeling.

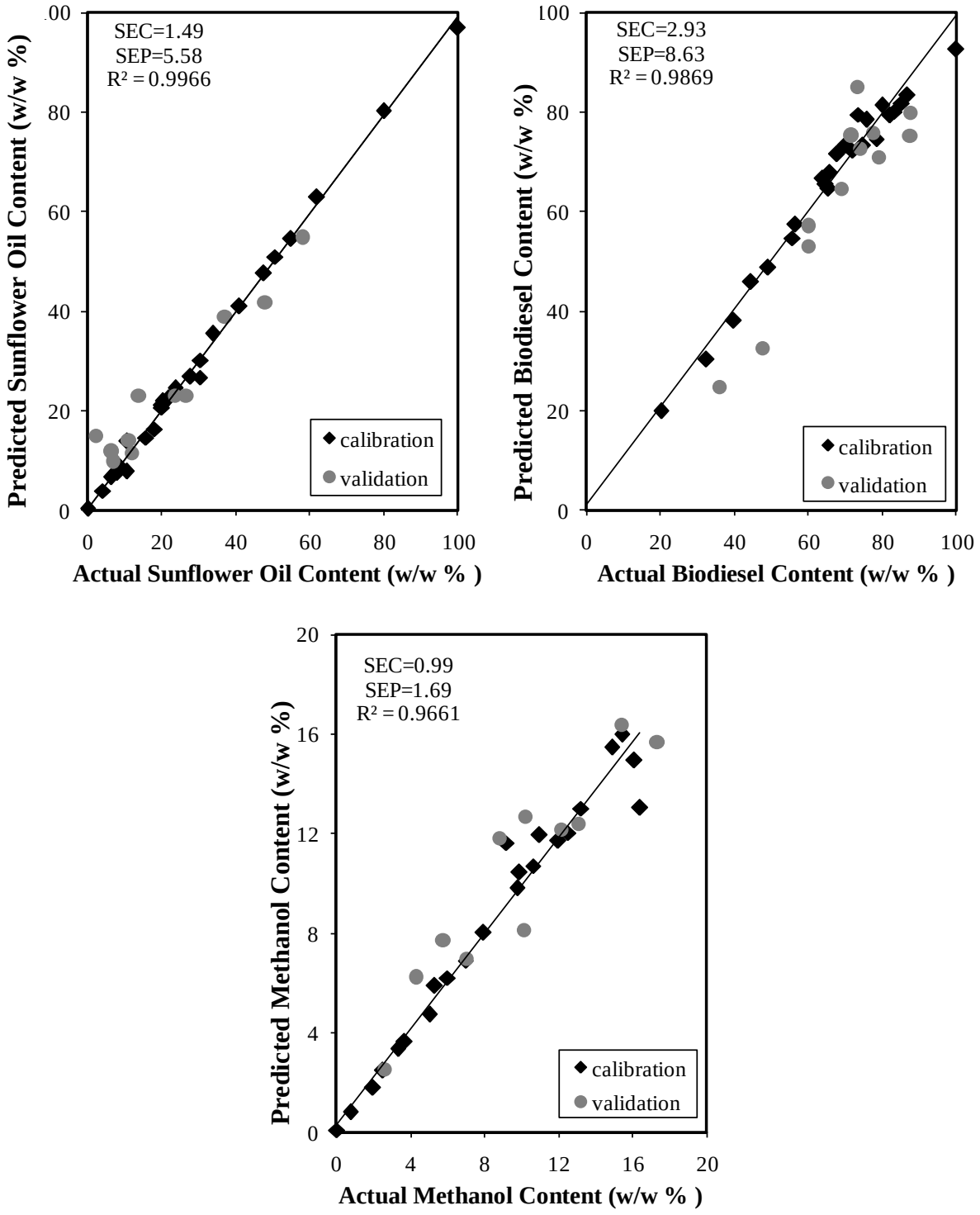


Figure 6.8. Actual versus predicted concentrations plot for NIR data analysis with ANN calibration method

6.2.2. Mid Infrared Analysis

Mid-infrared ATR spectra of ternary mixtures samples of biodiesel-sunflower oil-methanol and pure forms of components are shown in Figure 6.9. It is evident that the samples yield high absorbance values around 3400, 2900, and between the range 1750 and 1000 cm^{-1} wavenumbers. In addition, there is a peak around 2150 cm^{-1}.

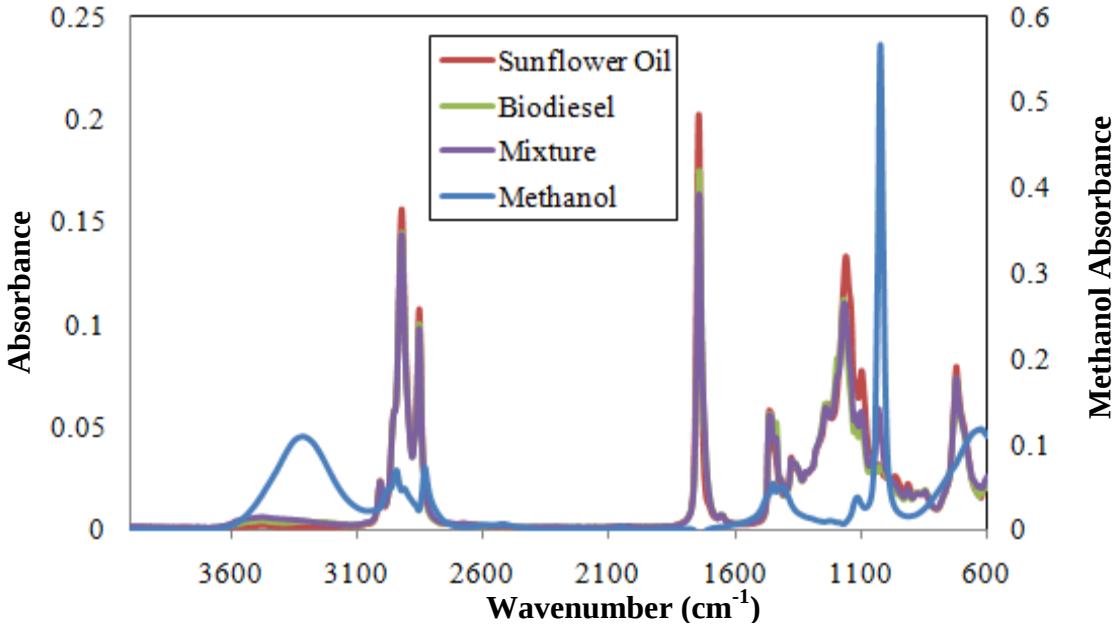


Figure 6.9. FTIR-ATR spectra of biodiesel, sunflower oil, and methanol along with their ternary mixture

As can be seen from Figure 6.9, biodiesel blends and its pure component have similar spectra except from methanol. The strong peaks, which observed at interval 1000-600 cm-1, related to C-H bonds of olefin or aromatics found in vegetable oils and biodiesels. The peaks at 1300-1000cm^{-1} displays the stretching of C-O bond and peaks at 1400-1250 cm^{-1} shows stretching for O-H bonds. The strongest peaks which overlap between the range of 2000-1500 cm^{-1} represents to C=C and C=O vibrations. Besides this, $-CH_2$ and $-CH_3$ bond stretching can be seen at 2950-2850 cm^{-1} interval. The peaks which were obtained in the range of 2900-2700 cm^{-1} are due to –CHO bond whereas the less strong peak at 3050cm^{-1} belongs to –C=C-H bond as well.

Consequently, as can be seen from Figure 6.9, there are very small differences between the spectra of oil, biodiesel and methanol with their ternary mixture.

Throughout the multivariate calibration process, it is expected these differences will reveal the information necessary to build successful calibration models otherwise almost impossible with univariate calibration methods. In order to construct MIR spectroscopic multivariate calibration models for biodiesel blends which contains biodiesel-sunflower oil-methanol the procedure followed in the NIR calibration is again used, i.e., ternary mixture calibration set were used again but NIR spectra were replaced with the MIR spectra with respect to GILS and ANN chemometric multivariate calibration modeling.

6.2.2.1. GILS Results

In order to prepare calibration models, 27 of 39 samples of the first ternary set were used to build calibration set and the remaining 12 sample were reserved for prediction set to test the performance of the models. GILS program run against 100 times along with 50 iterations and 30 genes to predict the concentration of ternary mixture of biodiesel, sunflower oil and methanol. Figure 6.10 shows the actual sunflower oil, biodiesel, and methanol concentration values versus their GILS predicted concentration values based on FTIR-ATR spectral data. Calibration models for methanol content determination gave standard error of calibration (SEC) and standard error of prediction (SEP) values as 1.08% (w/w) and 2.03% (w/w) for calibration and validation sets, respectively. In the case of sunflower oil and biodiesel content determination, the SEC and SEP values were 0.58% (w/w), 2.01% (w/w) and 1.33% (w/w), 2.61% (w/w) for calibration and prediction sets, respectively. In addition to this, when examining the correlation plots the R^2 value of regression lines for methanol was 0.995 and that for biodiesel and sunflower oil content was 0.993 and 0.999, respectively.

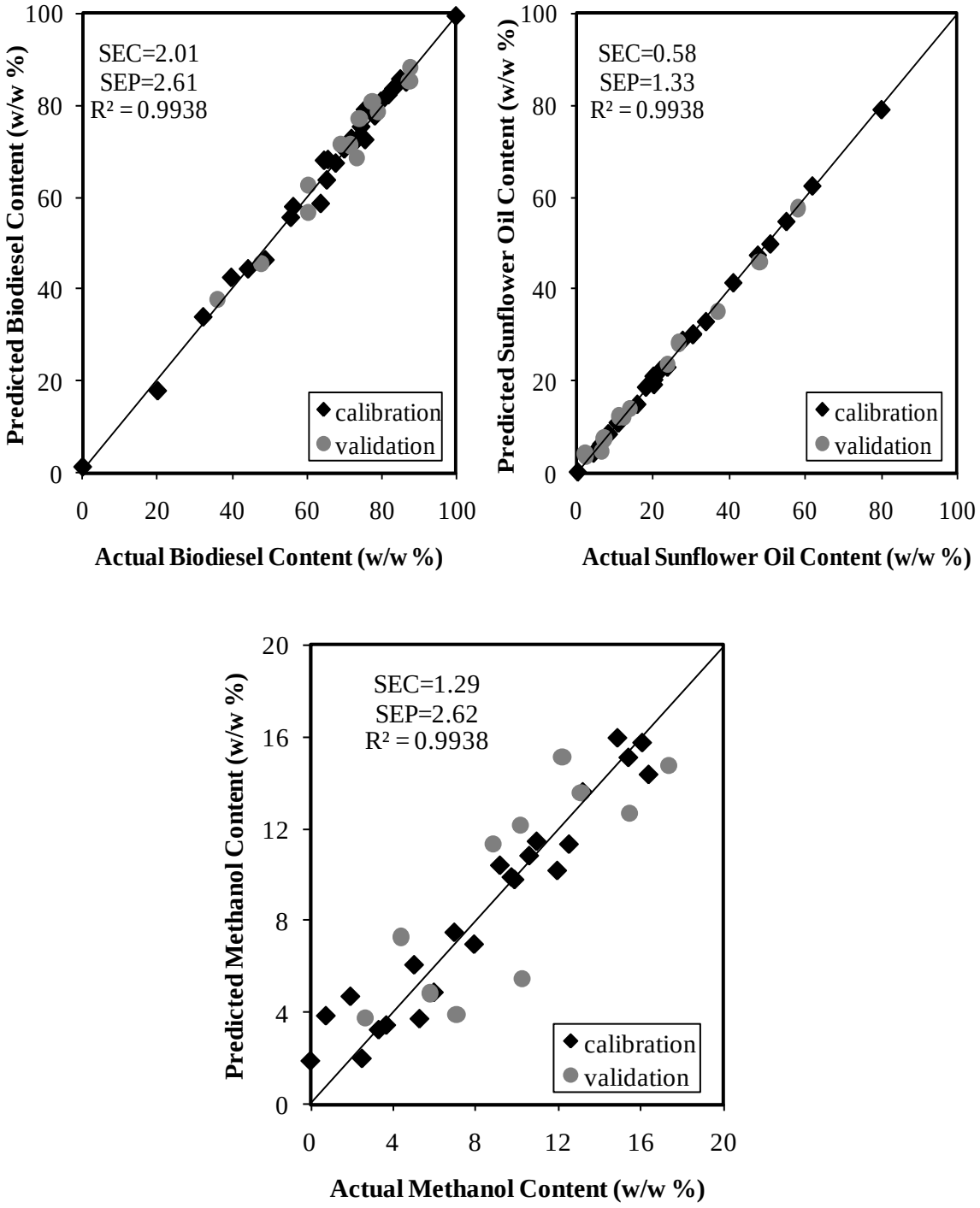


Figure 6.10. Actual versus predicted concentration plot for FTIR-ATR spectral data analysis with GILS calibration method.

As observed from SEC and SEP values that for biodiesel, sunflower oil and methanol contents they are still comparable. Nevertheless, the methanol values become higher than corresponding NIR results. Again, it must be realized that the GILS method is an iterative procedure due to the genetic algorithm used to select a subset of

wavelengths from the complete spectral range. However, when the overall calibration performance of the models examined, it is possible to state that the MIR spectra do contain quantitative information that correlated with sunflower oil and methanol contents of the biodiesel blends samples studied here. Figure 6.11 displays the frequency distribution of selected wavelengths in 100 runs with 30 genes and 50 iterations for biodiesel-sunflower oil-methanol set with respect to FTIR-ATR spectra data matrices.

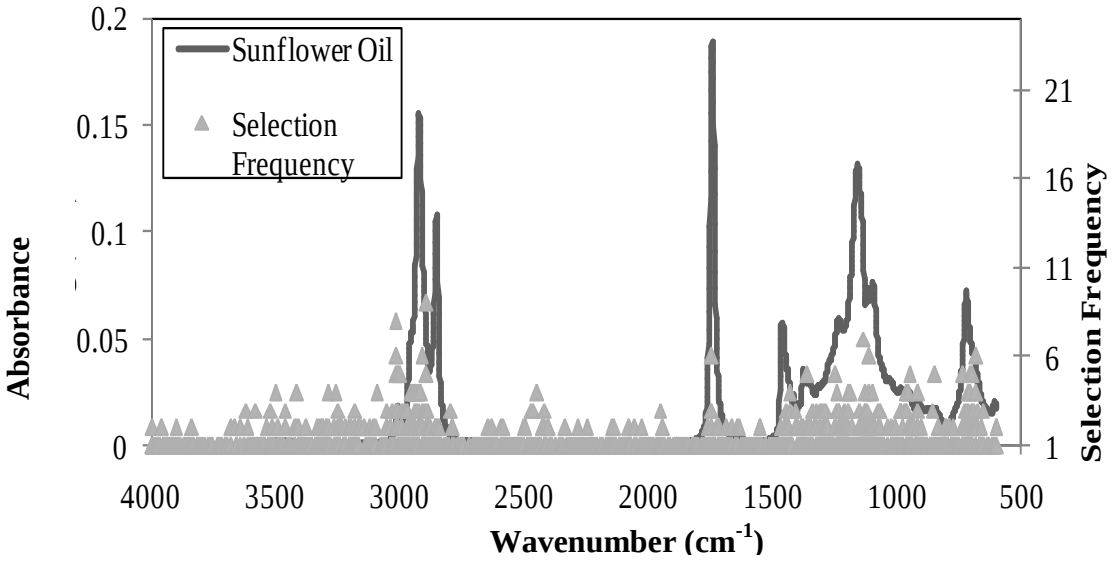
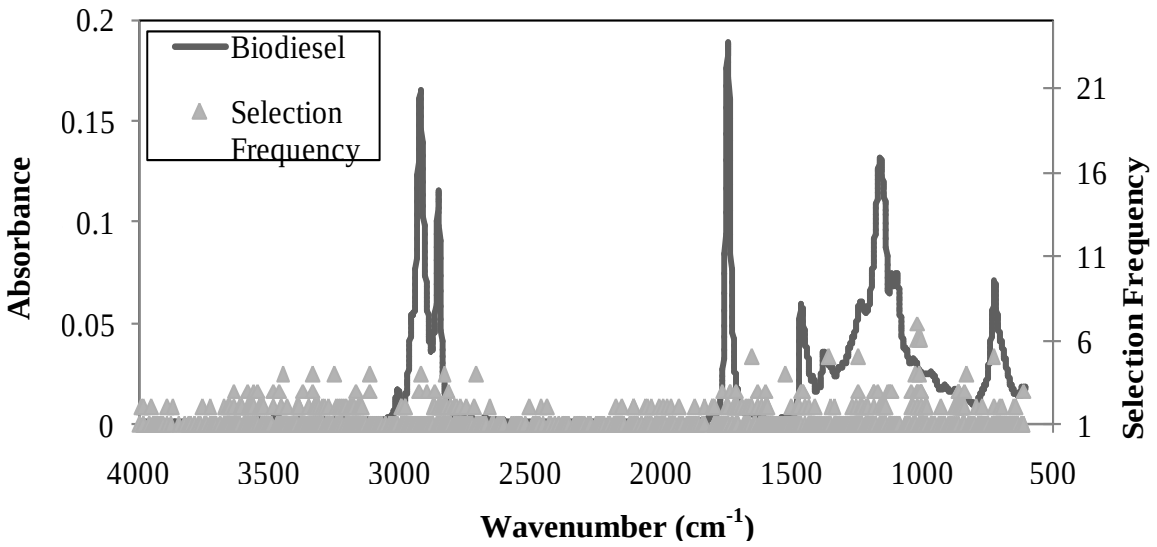


Figure 6.11. Wavelength selection frequency distribution of GILS method for ternary mixture of biodiesel-sunflower oil-methanol using FTIR-ATR spectroscopy.

(cont. on next page)



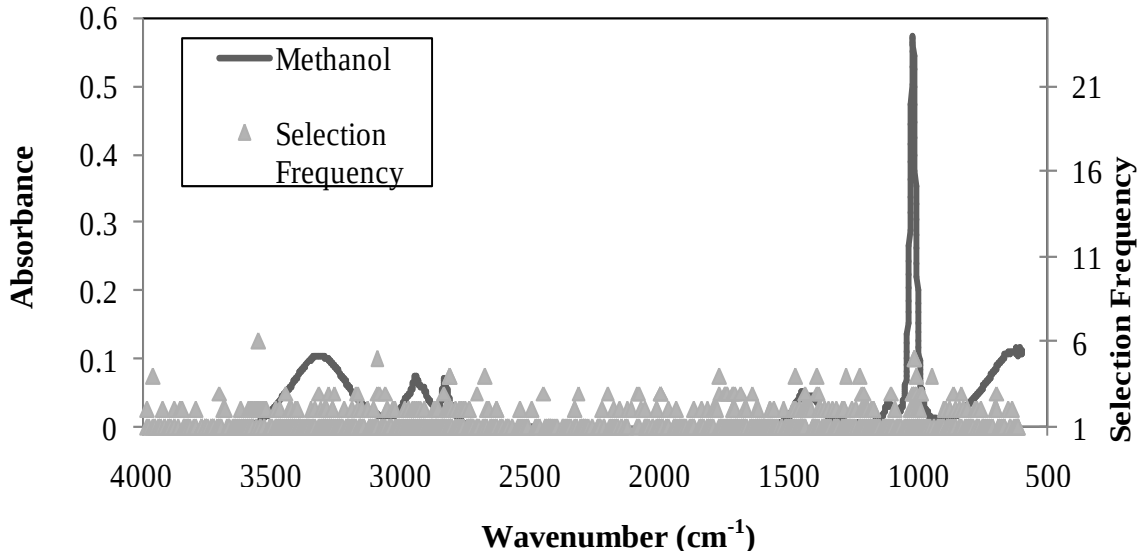


Figure 6.11. (cont.)

As can be seen from the Figure 6.11 there are a number of regions where selection frequencies are highly compared to the rest of the spectrum. The wavelength region around 1000 and 3000 cm^{-1} for all components indicates a strong tendency for GILS method to select while for biodiesel and sunflower oil content, around 500 cm^{-1} is the most frequently selected region. A significant difference between the frequency distribution of methanol and biodiesel, oil content is that the selected wavelength are more distributed in the former and much more wavelengths are selected.

6.2.2.2. ANN Results

In order to prepare calibration models for ANN, 27 of 39 samples of biodiesel, sunflower oil, methanol ternary set were used to calibration set and the remaining 12 samples were reserved for prediction set to test the performance of the models. Spectra were collected from each sample yielding a total of 39 spectra. The calibration model was tested with 27 spectra and then this model was tested with 17 independent prediction spectra which were not used in calibration step. Launching Neural Network Toolbox in Matlab 7.0.1 software programming in terms of the configurations as it is mentioned before; the user interface results for each component of ternary mixture modeled by ANN calibration are given in the Figure 6.12.

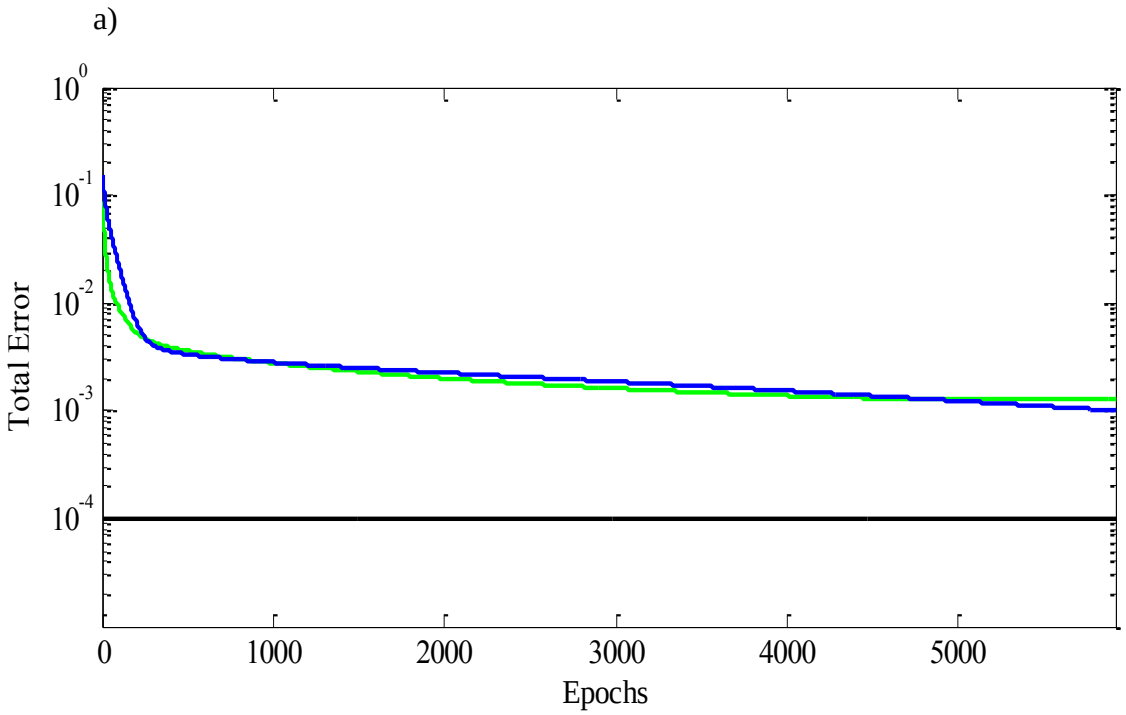


Figure 6.12. Error versus epochs plot of ternary mixture of a) sunflower oil b) biodiesel c) methanol (green line: validation, blue line: calibration, black line: target)

(cont. on next page)

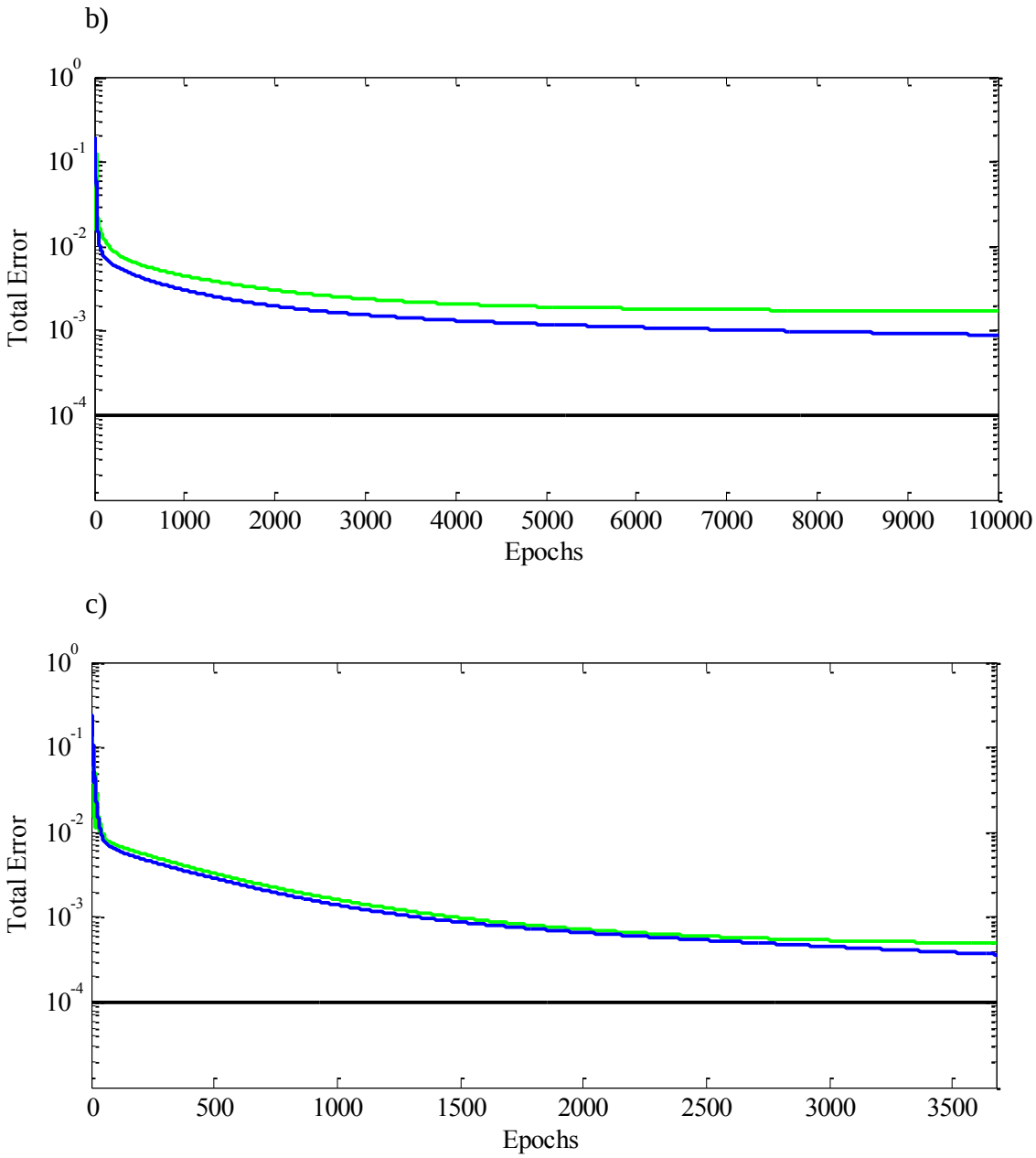


Figure 6.12. (cont.)

As can be seen in the Figure 6.12, ANN training process stopped at which the epochs number was 5917 for sunflower oil and mean square error value reached about 104×10^{-6} for calibration, on the other hand, due to increasing of mse value of validation program did not continued. Even in case of methanol, 3676 iterations were necessary to achieve about 366×10^{-6} mse value and validation mse value again increased. For biodiesel, epochs reached 10000, which is the limit number, thus program stopped at which the performance was 89×10^{-5} mse value.

Moreover, in training process, if the calibration and validation lines tend to be proceeding in an adjacent trajectory, it demonstrates that the model constructed successfully. Also for a stable case, after few iterations, total error value should begin to diminish and this trend was clearly seen in Figure 6.12. However, output data is required to revise from the 0-1 interval to the actual value since log-sigmoid function obtains results in from 0.1 to 0.9. After these conversion applied, it is possible to calculate the SEC and SEP values along with the regression values between actual and predicted by ANN model. Hence, Figure 6.13 shows the correlation graphs for FTIR analysis with ANN calibration of ternary mixture of biodiesel-sunflower oil-methanol.

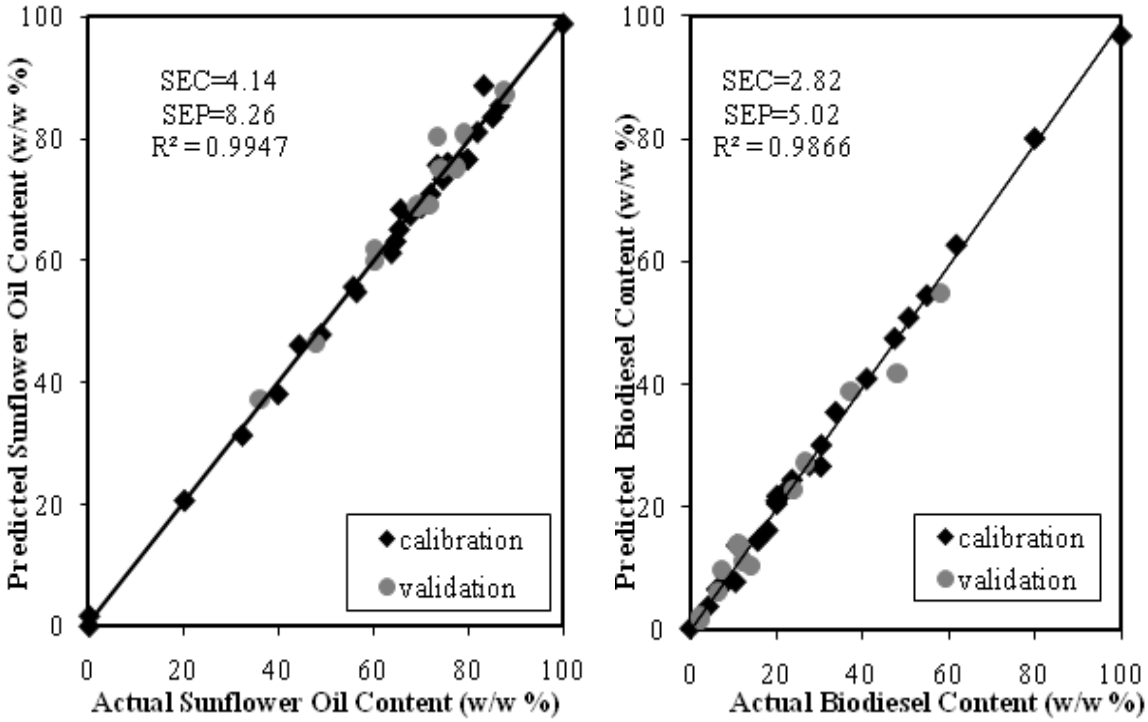


Figure 6.13. Actual versus predicted concentration plot for FTIR-ATR data analysis with ANN calibration method

(cont. on next page)

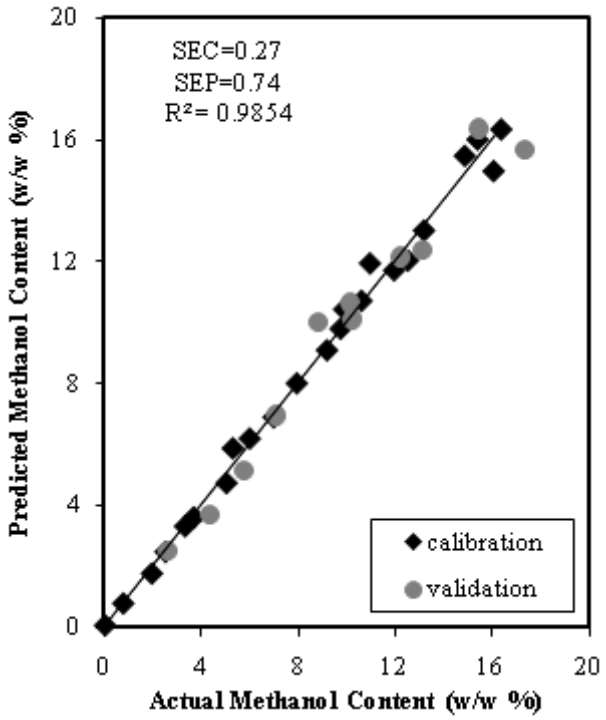


Figure 6.13. (cont.)

As can be seen from Figure 6.13, R^2 values found approximately 0.99. In addition to this, SEC values obtained 0.27% (w/w) for methanol, 2.82% (w/w) for biodiesel and 4.14% (w/w) for sunflower oil whereas SEP values 0.74% (w/w) for methanol, 5.02% (w/w) for biodiesel and 8.26% (w/w) for sunflower oil at which the methanol content was up to 20 percentages by mass whereas biodiesel and sunflower oil was up to 100 percentages by mass concentration. Thus, the constructed model is applicable for calibration of biodiesel blends to determine the methanol and untreated vegetable oil content via ANN multivariate calibration modeling when the reaction medium is considered. Especially to monitor the transesterification reaction of biodiesel synthesis by determining the methanol content can be possible by these chemometric multivariate calibration methods along with the infrared spectroscopy.

6.3. Ternary mixtures of Biodiesel – Vegetable oil – Diesel Set

As it is stated before, determining the amount of biodiesel content in petroleum diesel fuel in the presence of vegetable oils, which are not converted into biodiesel, have been important. Since biodiesel production process is expensive and time consuming diesel fuel may be illegally adulterated with raw or used frying vegetable oils before converting into biodiesel. Therefore, their determination with classic methods including ^{1}H-NMR spectroscopy (Knothe et al. 2001) and chromatography (Foglia et al. 2005) are including more expensive and time-consuming process, due to constraint of sample preparation.

Furthermore, infrared spectroscopy combined to multivariate calibration has been shown to be an alternative analytical technique to classic methods since it allows low cost, fast, and nondestructive determination without sample preparation (Pimentel et al. 2006).

All samples whose concentrations by mass percentages displayed in Table 6.7 and Table 6.8 were analyzed using NIR and MIR spectrometers and the data collected for the prediction. Each set divided into two sets: one was for calibration and the other was for validation. Calibration set contains 33 samples spectra and validation set contained 17 samples spectra at which the samples used in calibration or validation set arranged in a random order. The range of concentrations each constituent in the sets were in the range of 0 – 100 % for all biodiesel, vegetable oil and diesel.

Table 6.7. Concentration profiles for calibration set of ternary mixture of biodiesel, sunflower oil, diesel set. All concentrations are given by mass percentages

Sample No	Biodiesel (w/w%)	Sunflower Oil (w/w%)	Diesel (w/w%)	Sample No	Biodesel (w/w%)	Sunflower Oil (w/w%)	Diesel (w/w%)
1	100.00	0.00	0.00	18	9.88	14.88	75.24
2	0.00	100.00	0.00	19	10.41	67.20	22.40
3	0.00	0.00	100.00	20	57.83	20.17	22.00
4	56.69	9.59	33.72	21	37.39	10.94	51.68
5	37.26	22.92	39.82	22	39.30	20.22	40.47
6	31.38	25.85	42.76	23	35.72	22.97	41.30
7	17.43	24.39	58.18	24	67.50	24.78	7.72
8	5.44	12.92	81.64	25	9.75	38.69	51.56
9	59.70	9.82	30.48	26	16.76	77.99	5.26
10	44.31	23.25	32.43	27	45.33	8.30	46.38
11	72.14	9.01	18.85	28	27.14	24.44	48.42
12	15.69	83.15	1.16	29	33.62	59.07	7.31
13	4.80	40.66	54.53	30	50.81	15.12	34.07
14	6.43	63.15	30.42	31	35.52	39.59	24.89
15	16.57	20.38	63.05	32	2.25	46.25	51.50
16	42.30	57.10	0.61	33	83.01	1.23	15.75
17	41.35	26.54	32.11				

As it is seen in Table 6.7 calibration data set includes the pure form of sunflower oil and biodiesel with diesel components which pay the attentions on this set.

Table 6.8. Concentration profiles for validation set of ternary mixture of biodiesel, sunflower oil, diesel set. All concentrations are given by mass percentages

Sample No	Biodiesel (w/w%)	Sunflower Oil (w/w%)	Diesel (w/w%)	Sample No	Biodiesel (w/w%)	Sunflower Oil (w/w%)	Diesel (w/w%)
1	23.87	6.60	69.5	10	9.15	46.88	43.98
2	32.67	57.50	9.84	11	34.22	9.23	56.55
3	33.67	60.75	5.57	12	36.47	43.96	19.57
4	38.00	2.61	59.39	13	14.14	42.41	43.45
5	11.10	45.02	43.88	14	25.84	45.32	28.84
6	36.60	43.81	19.59	15	52.28	32.44	15.28
7	14.74	16.20	69.06	16	29.56	37.07	33.37
8	33.99	27.60	38.41	17	24.09	41.61	34.29
9	26.05	26.25	47.69				

6.3.1. Near Infrared Analysis

The NIR spectral changes that result from the varying concentration of the compounds in the biodiesel/diesel blends in vegetable oils are difficult to interpret visually. Generally, in the NIR spectral region the absorbance bands observe as broad and overlapping of which spectral bands makes the use of multivariate calibration necessary to resolve the components from the full spectral data which is impossible with univariate calibration. Figure 6.14 illustrates the NIR spectra of a ternary solution with pure forms of components of which are biodiesel, sunflower oil and diesel.

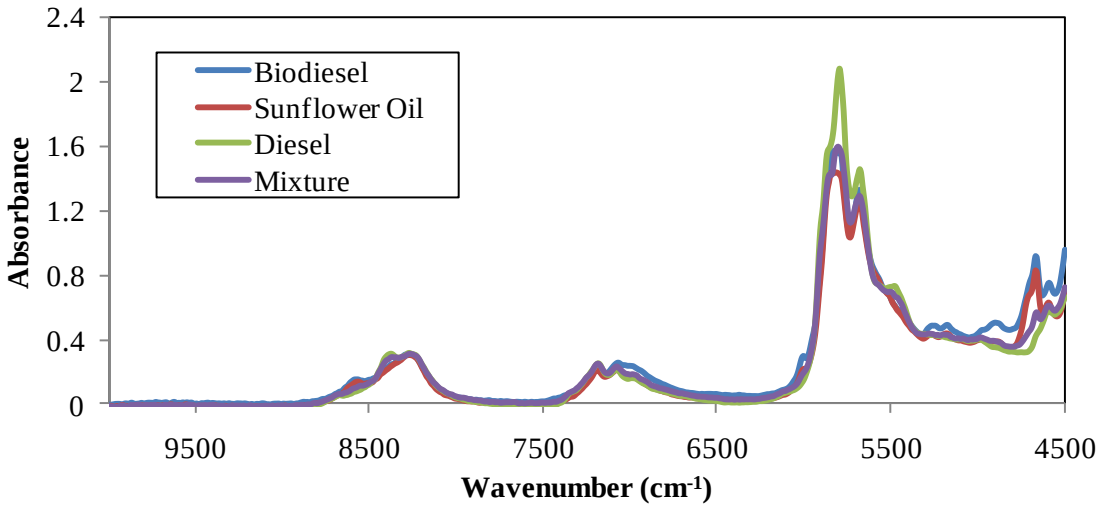


Figure 6.14. FT-NIR spectra of biodiesel, sunflower oil, diesel and their ternary mixture

As can be seen from Figure 6.14 spectra which has the interval of 10000-4500 cm-1 wavenumber, show NIR absorption bands corresponding to the -CH=CH, -C-H asymmetric stretching and C=C stretching at around 4650 cm-1. Also, =C-H and C=C stretching belong to –CH=CH- group. The spectra show maximum absorbance at 5810 cm-1. This maximum absorption band is related to first overtone of C-H bond, which belongs to –CH2 functional group while absorption band (only a shoulder) at 5680 cm^{-1} is the first overtone of -C-H bond. More to the point this, C-H stretching vibrations belongs to weak and broad -CH_3 functional group is seen at 7210 cm^{-1} whereas, the other weak and broad shoulder peak at around 7110 cm^{-1} indicates the C-H stretching which leads from CH_2 functional group. Another important peak in spectra is observed

at 8285 cm^{-1} that belongs to C-H stretching and represents second overtone. Specially, the differences in NIR spectra of oils, esters, and petrodiesel were observed for the region about 4500 cm^{-1} that can be assigned to CO and C–O stretching combinations.

From these spectra, it is noticeable that each constituent exhibits very similar spectral characteristics, which makes it necessary to use a multivariate calibration method to resolve the mixtures of these compounds.

6.3.1.1. GILS Results

FT-NIR absorbance spectral data matrices obtained and calibration models were constructed in terms of each component in ternary mixture, and then tested against 100 times along with 50 iterations and 30 genes. Figure 6.15 shows the actual sunflower oil, biodiesel, and diesel concentration values versus their GILS predicted concentration values based on FT-NIR spectral data. The standard error of calibration (SEC) values of each component were found between 0.90% (w/w) and 1.04% (w/w) and the standard error of prediction (SEP) values of each component were found between 1.07% (w/w) and 1.46% (w/w) by using GILS method for all components in set. SEC and SEP values for the calibration models of sunflower oil and diesel content were 1.04% (w/w), 0.94% (w/w) and 1.46% (w/w), 1.07% (w/w) respectively and for biodiesel determination were 0.90% (w/w) and 1.32% (w/w) for the data set. In addition to this, R^2 values of regression lines for all components were 0.99. Therefore, when R^2, SEC and SEP values are examined, it is seen that these values are compatible with each other, which demonstrates a good prediction for rapid determination of the adulteration of biodiesel/diesel blends in vegetable oils in illegal marketing.

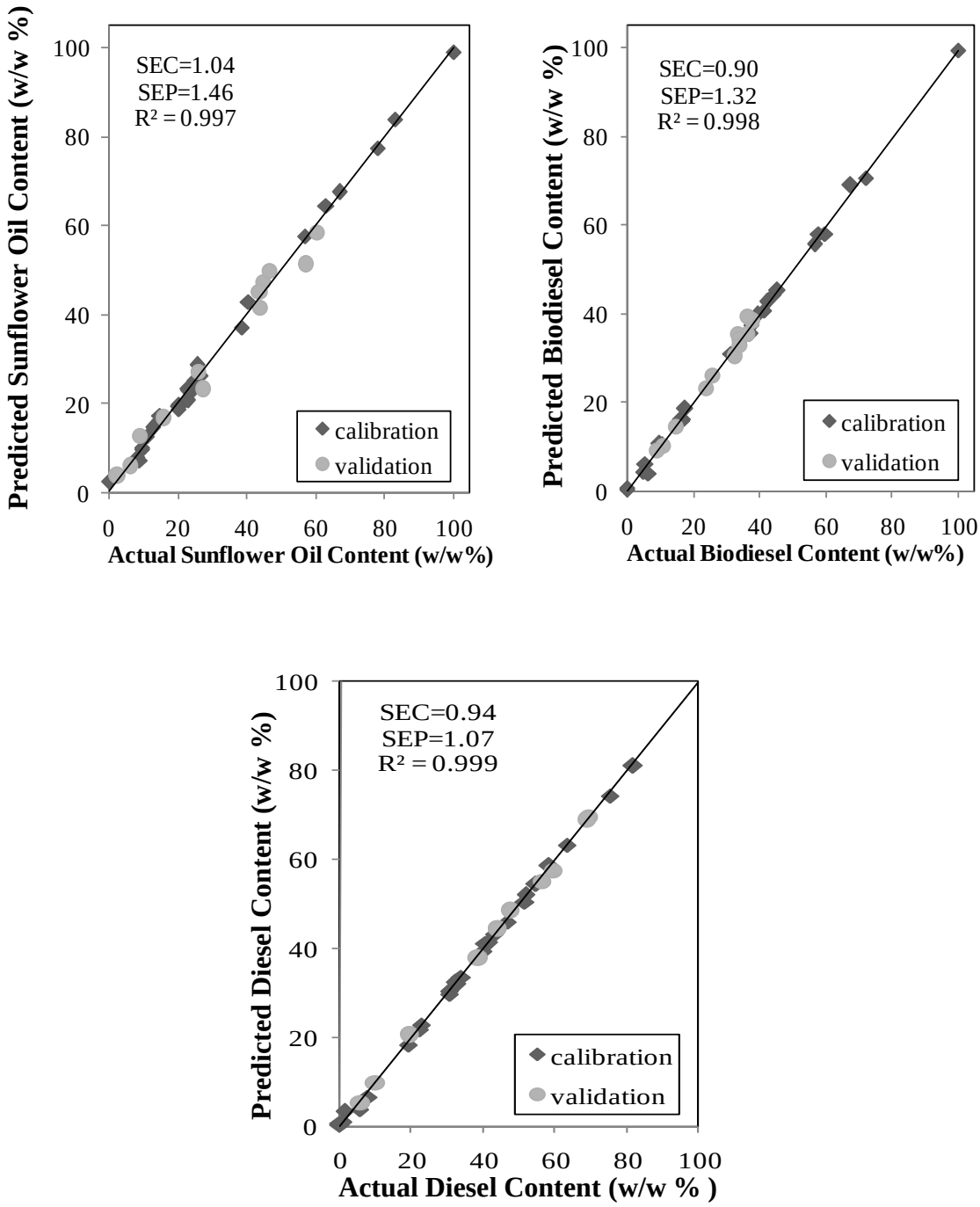


Figure 6.15. Actual versus predicted concentration plot for NIR data analysis with GILS calibration method.

Since GILS is a method that based on wavelength selection, it is important to examine the distribution of selected wavelengths in multiple runs over the entire full spectral region. Figure 6.16 displays the frequency distribution of selected wavelengths in 100 runs with 30 genes and 50 iterations for biodiesel-sunflower oil-diesel set.

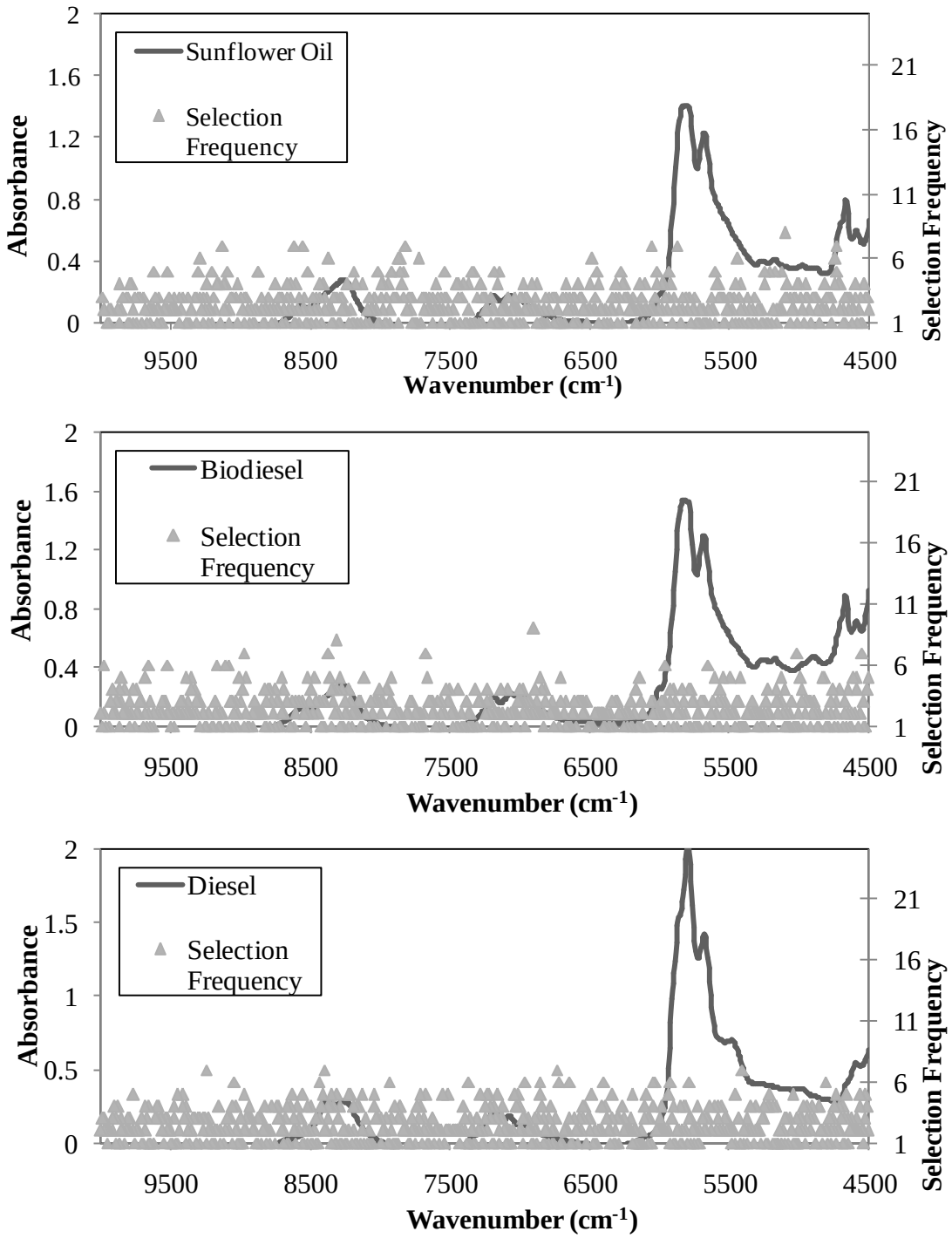


Figure 6.16. Wavelength selection frequency distribution of GILS method for ternary mixture of sunflower oil - biodiesel - diesel using FT-NIR spectroscopy.

As can be seen from the Figure 6.16, there are a number of regions where selection frequencies are highly compared to the rest of the spectrum. The wavelength region around 5000 cm^{-1} for biodiesel and sunflower oil indicates a strong tendency for GILS method to select while for diesel content, around 7000 and 8500 cm^{-1} is the most

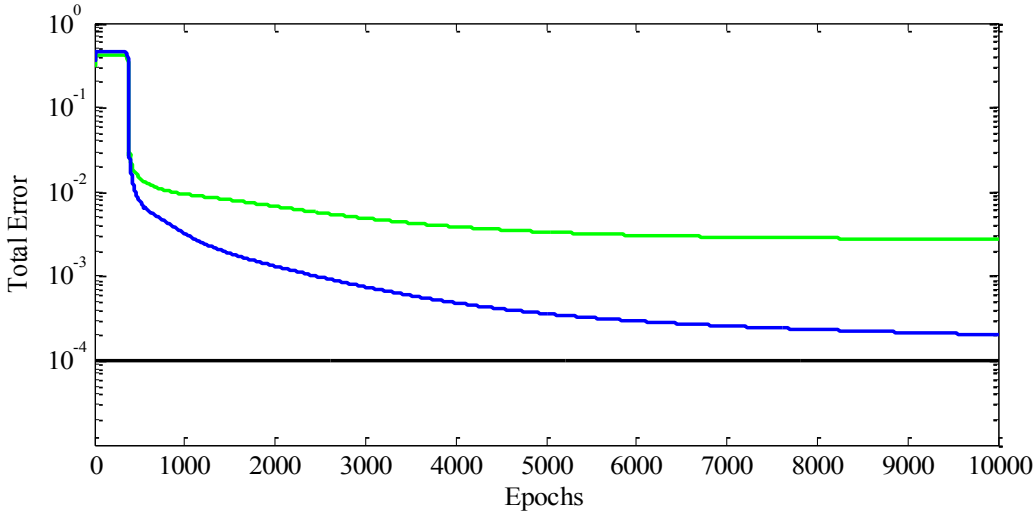
frequently selected region. As can be seen from the figures, the frequency of the selected cell numbers correspond to selected wavelengths is significantly higher around the peak maximum of each component. This shows that the GILS method selects the regions, where the most concentration related information is contained.

6.3.1.2. ANN Results

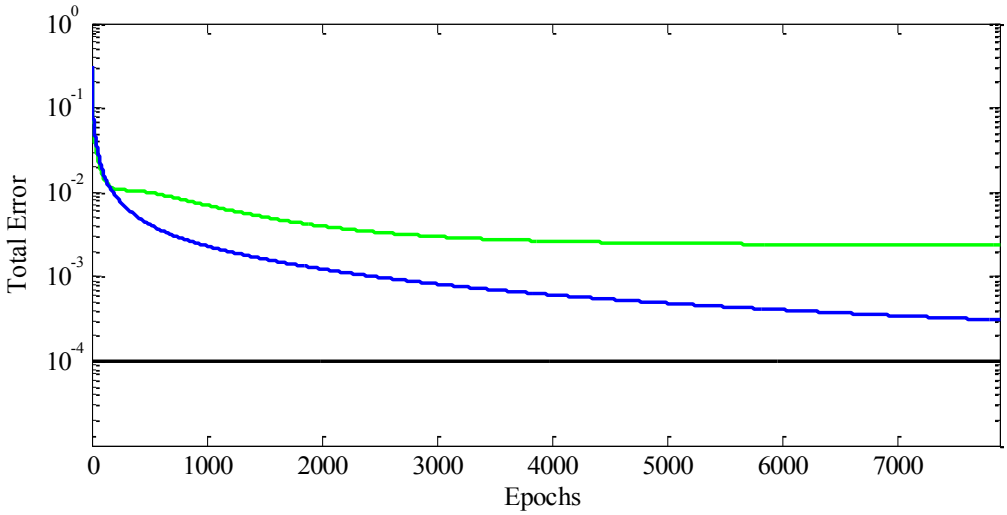
The ternary mixture of biodiesel-sunflower oil-diesel set was used for the construction of artificial neural network and optimization processes were performed with respect to their NIR spectral data matrices. After some trials to construct ANN calibration model, as an optimum conditions, learning rate and momentum value selected as 0.85 and 0.75, respectively. Since infrared spectra of ternary mixtures have large number of data points, prior to perform the ANN analysis, reduction of data by PCA-SVD algorithms is required. In addition, to calculate the total error, mse, which represents the term mean square error, applied with a goal of 0.0001 up to maximum value of 10000 epochs that means weight vectors or iteration number. Sample design data set and infrared spectral discussion were same as the previous section that is subtitled as GILS results

Launching Neural Network Toolbox in Matlab 7.0.1 software programming in terms of the configurations as it is pointed out earlier; the user interface results for each component of ternary mixture modeled by ANN calibration are given in the Figure 6.17. As can be seen from the Figure 6.17, for biodiesel component after 7889 iterations performed, the mse value reached about $306x10^{-6}$ for calibration; on the other hand, due to increasing of mean square error value of validation, program did not continued so far. For sunflower and diesel components, epochs reached the limit number which is determined as 10000, thus program stopped at which the performance was $295x10^{-5}$ and $20x10^{-5}$ mean square error values, respectively.

a)



b)



c)

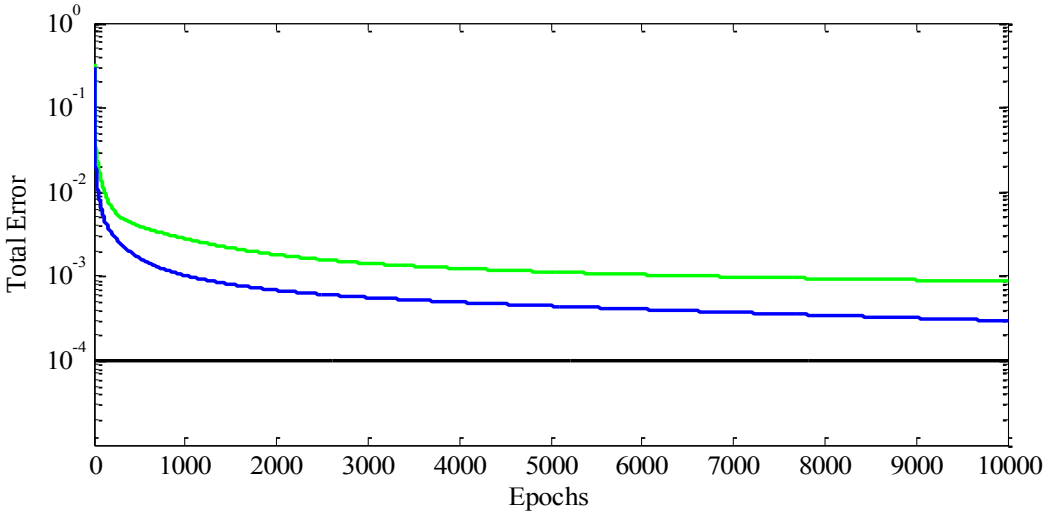


Figure 6.17. Error versus epochs plot of ternary mixture of a) sunflower oil b) biodiesel c) diesel (green line: validation, blue line: calibration, black line: target)

For ANN models, after some iteration, total error value should begin to diminish significantly and as can be seen from the Figure 6.17, error value immediately reduced after program run. Also, the more adjacent calibration and validation lines through the target line, the better constructed the model and this trend is seen as well. However, output data is required to revise from the 0-1 interval to the actual value since log-sigmoid function obtains results in from 0.1 to 0.9. After these conversion applied, it is possible to calculate the SEC and SEP values along with the regression values between actual and predicted by ANN model.

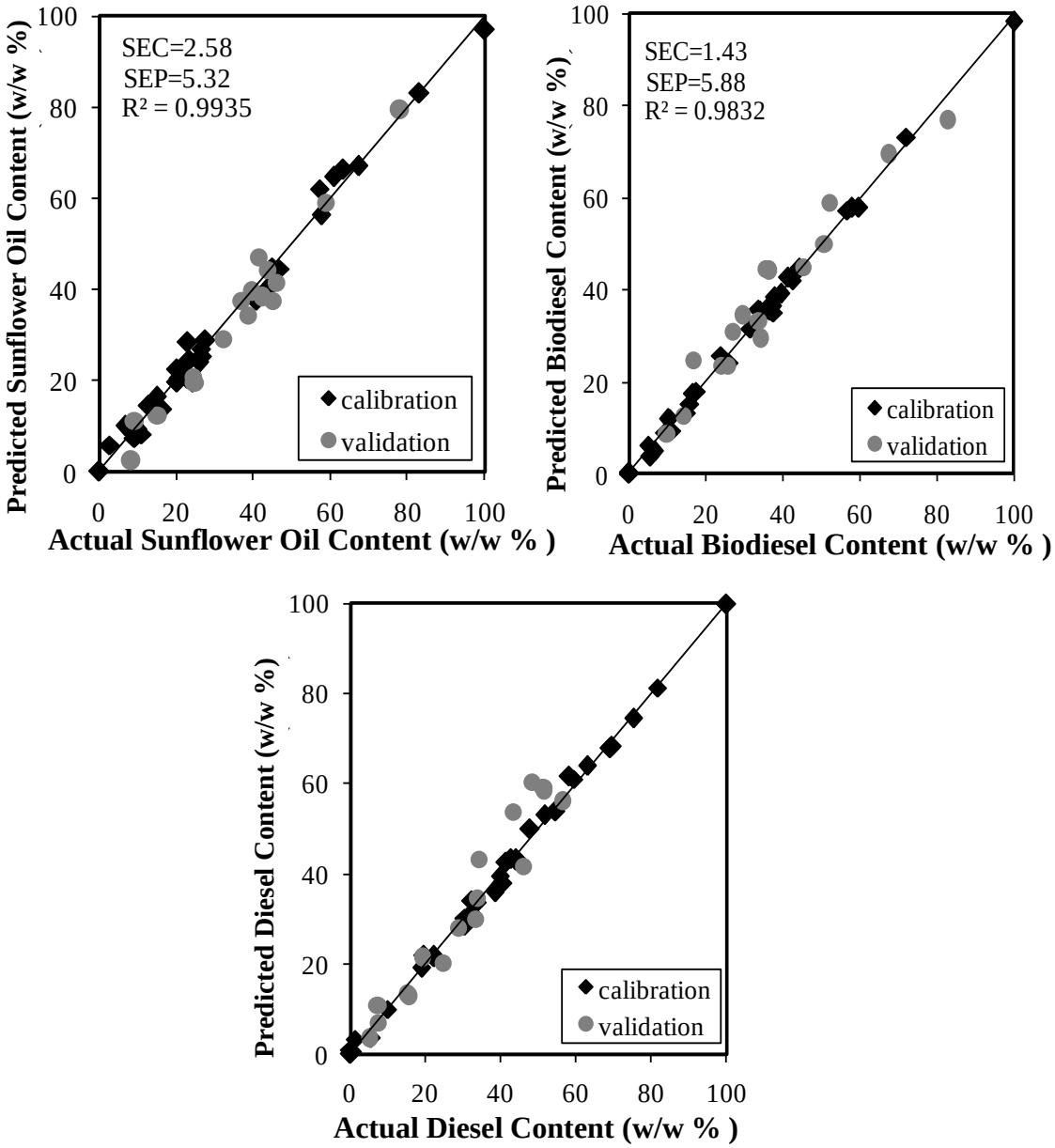


Figure 6.18. Actual versus predicted concentration plot for NIR data analysis with ANN calibration method.

As can be seen from Figure 6.18, R^2 values observed around 0.99. In addition, SEC values obtained 1.63% (w/w) for diesel, 1.43% (w/w) for biodiesel and 2.58% (w/w) for sunflower oil whereas SEP values 6.52% (w/w) for diesel, 5.88% (w/w) for biodiesel and 5.32% (w/w) for sunflower oil. According to these values, the constructed model is applicable for determination of adulteration of diesel, which is largely responsible for why many industrial machines and car engines using this product develop fault at regular intervals and finally break down after several repairs.

6.3.2. Mid Infrared Analysis

Mid-infrared ATR spectra of ternary mixtures samples of biodiesel-sunflower oil-diesel and pure forms of components are shown in Figure 6.19. It is obvious that the samples yield high absorbance values around 3400, 3000, and between the range 1750 and 1000 cm^{-1} wavelengths. In addition, there is a peak around 2150 cm^{-1}.

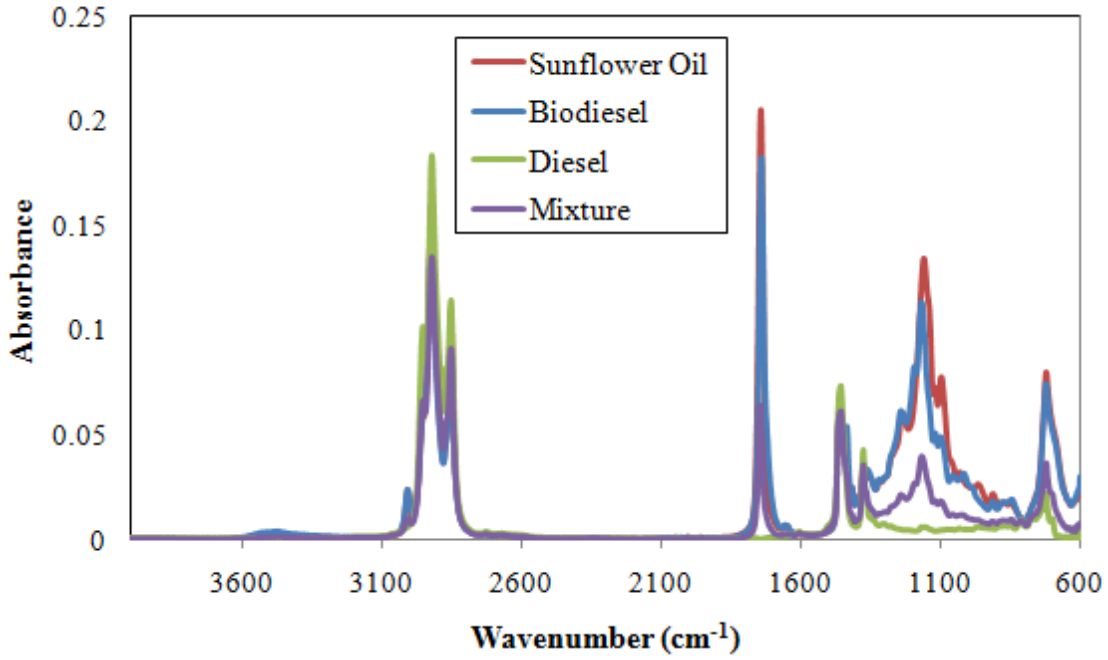


Figure 6.19. FTIR-ATR spectra of ternary mixtures of biodiesel, sunflower oil, diesel and their pure forms.

MIR spectra of the sunflower oil, their corresponding methyl ester, biodiesel were compared with petrodiesel, as shown in Figure 6.19. There are no absorption peaks from petrodiesel in the regions (3700 to 3000), (1900 to 1500) and (1800 to 800) cm^{-1}, while vegetable oils and their corresponding esters absorb well in those regions. Stretching vibration carbonyl bands, around 1750 cm^{-1}, for all the raw oils are overlapped with their corresponding esters. Peaks in the region (1300 to 800) cm^{-1} also indicate overlapped bands between oils and their corresponding esters. Peaks in the region (1000 to 900) cm^{-1} assigned to symmetric angular deformation out of plane of the C–H bonds of olefins. Peaks around 1200 cm^{-1} may be assigned to the axial deformation of CC(O)–O bonds of the ester, while peaks around 1183 cm^{-1} may be assigned to asymmetric axial deformation of O–C–C bonds. This region (1300 to 900) cm^{-1} is known as the "fingerprint" region of complex spectra that include many coupled vibration bands. These overlapped peaks indicate that univariate calibration models may cause significant prediction error to quantify ester concentration when raw oil is present. Those models are also inadequate for identifying the presence of raw oil in a spoiled blend either due to the illegal addition of raw oil. Therefore, multivariate calibration modeling via the peaks in these mid infrared regions, corresponding to the vibration of carbonyl groups to distinguish sunflower oil from its biodiesel with diesel blends along with the GILS and ANN chemometric modeling.

6.3.2.1. GILS Results

The sample design set generated from 33 of them as calibration set and the remaining 17 samples as validation samples performed as it is same in the NIR spectral data analysis.

GILS program run against 100 times along with 50 iterations and 30 genes to predict the concentration of ternary mixture of biodiesel, sunflower oil and diesel. Figure 6.20 shows the actual sunflower oil, biodiesel, and diesel concentration values versus their GILS predicted concentration values based on FTIR-ATR spectral data.

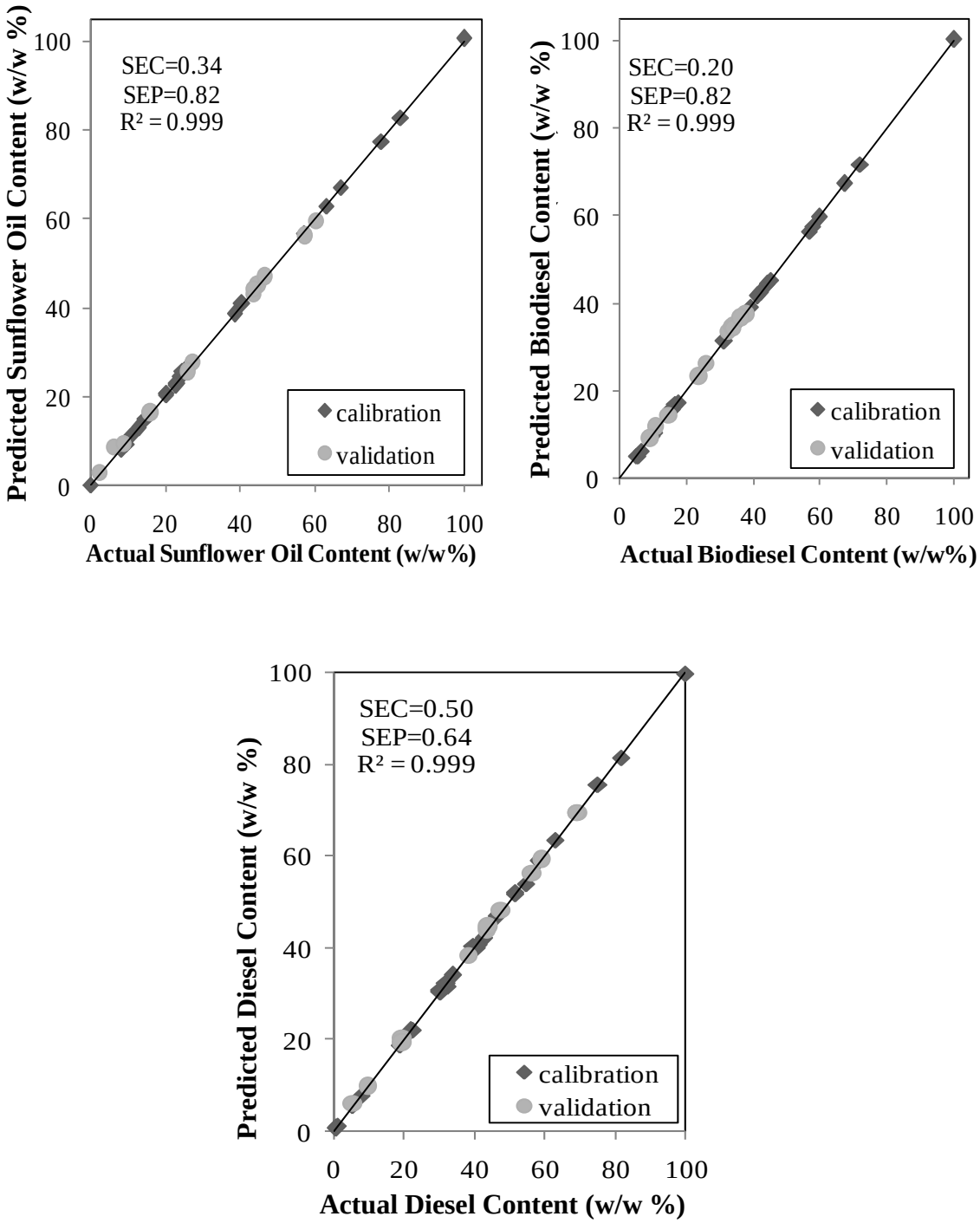


Figure 6.20. Actual versus predicted concentration plot for FTIR-ATR spectral data analysis with GILS calibration method.

As can be seen from Figure 6.20, calibration models for diesel content determination gave standard error of calibration (SEC) and standard error of prediction (SEP) values as 0.20% (w/w) and 0.34% (w/w) for calibration and validation sets. In the case of sunflower oil and biodiesel content determination, the SEC and SEP values were 0.34% (w/w), 0.82% (w/w) and 0.20% (w/w), 0.82% (w/w) for calibration and

prediction sets, respectively. In addition to this, when examining the correlation plots the R^2 value of regression lines for all components were 0.999 that indicates to be powerful multivariate calibration method when accompanied with proper wavelength selection methods.

However, it must be realized that the GILS method is an iterative procedure due to the genetic algorithm used to select a subset of wavelengths from the complete spectral range. Figure 6.21 displays the frequency distribution of selected wavelengths in 100 runs with 30 genes and 50 iterations for biodiesel-sunflower oil-diesel set with respect to FTIR-ATR spectra data matrices.

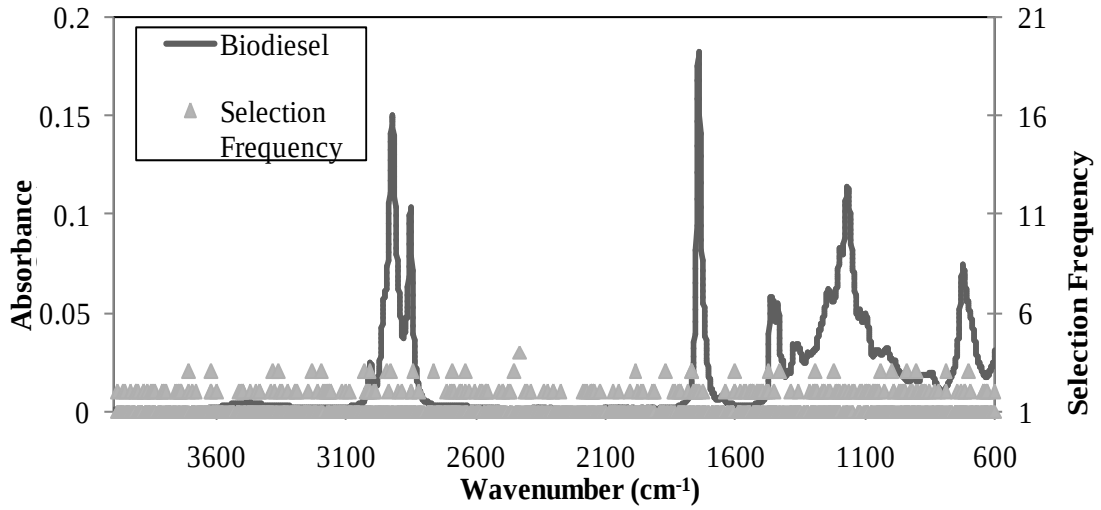


Figure 6.21. Wavelength selection frequency distribution of GILS method for ternary mixture of biodiesel-sunflower oil-methanol using FTIR-ATR spectroscopy.

(cont. on next page)

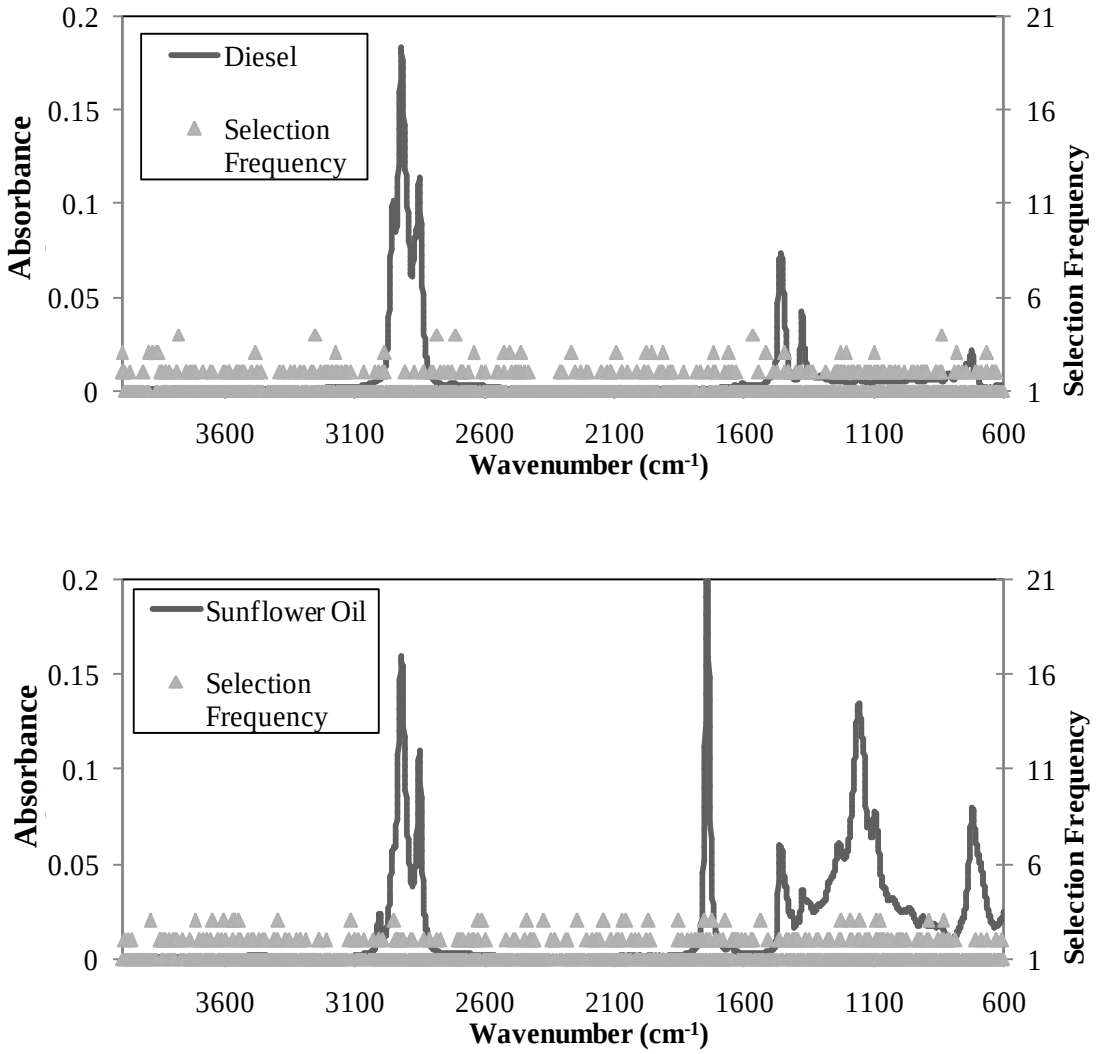


Figure 6.21. (cont.)

As can be seen from the Figure 6.21, there are a number of regions where selection frequencies are highly compared to the rest of the spectrum. The wavelength region around 1000 and 3000 cm^{-1} for all components indicates a strong tendency for GILS method to select while for biodiesel and sunflower oil content, around 500 cm^{-1} is the most frequently selected region.

6.3.2.2. ANN Results

Construction of artificial neural network calibration model is performed with 32 of 50 samples of biodiesel, diesel, sunflower oil ternary set and the remaining 18 were used to test the performance of the model. Spectra were collected from each sample yielding a total of 50 spectra and 32 spectra used to construct model while 18 spectra used to test. Launching Neural Network Toolbox in Matlab 7.0.1 software programming in terms of the configurations as it is discussed earlier, the user interface results for each component of ternary mixture modeled by ANN calibration are given in the Figure 6.22.

a)

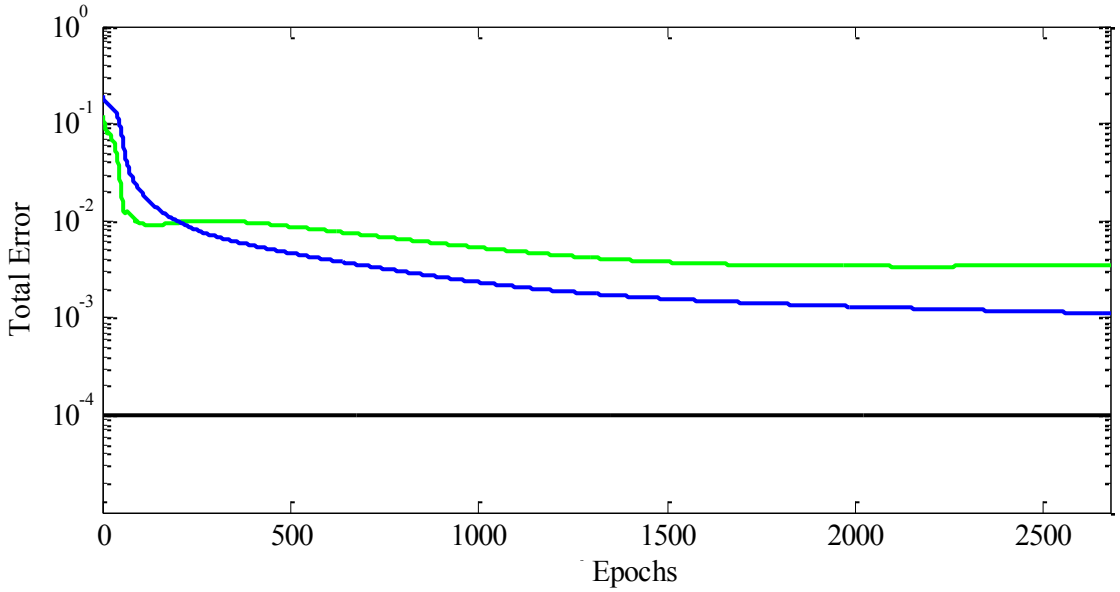
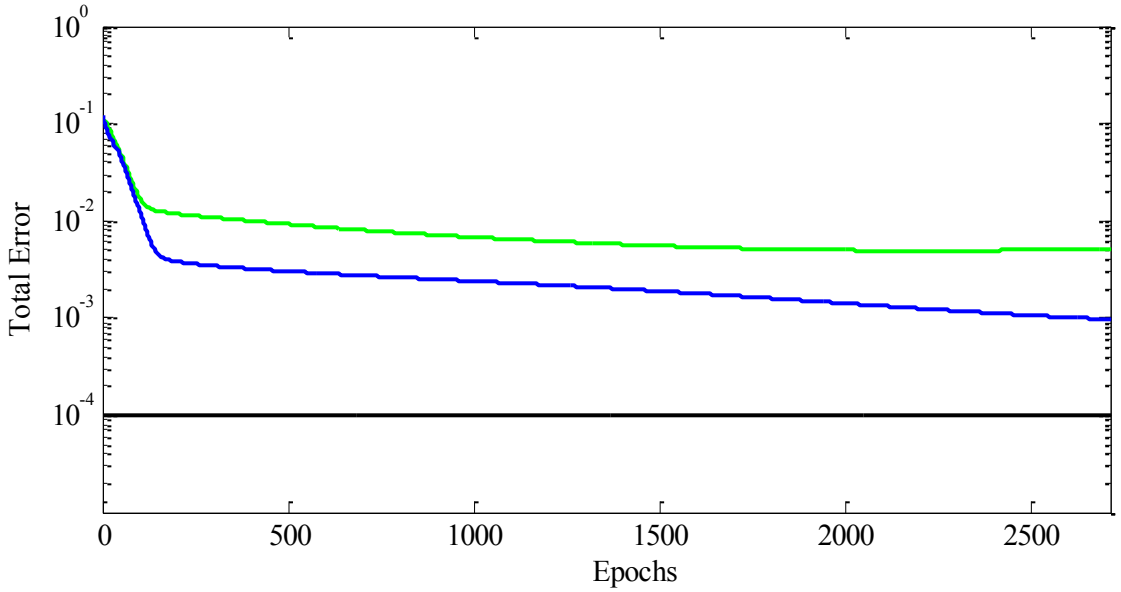


Figure 6.22. Error versus epochs plot of ternary mixture of a) sunflower oil b) biodiesel c) diesel (green line: validation, blue line: calibration, black line: target)

(cont. on next page)

b)



c)

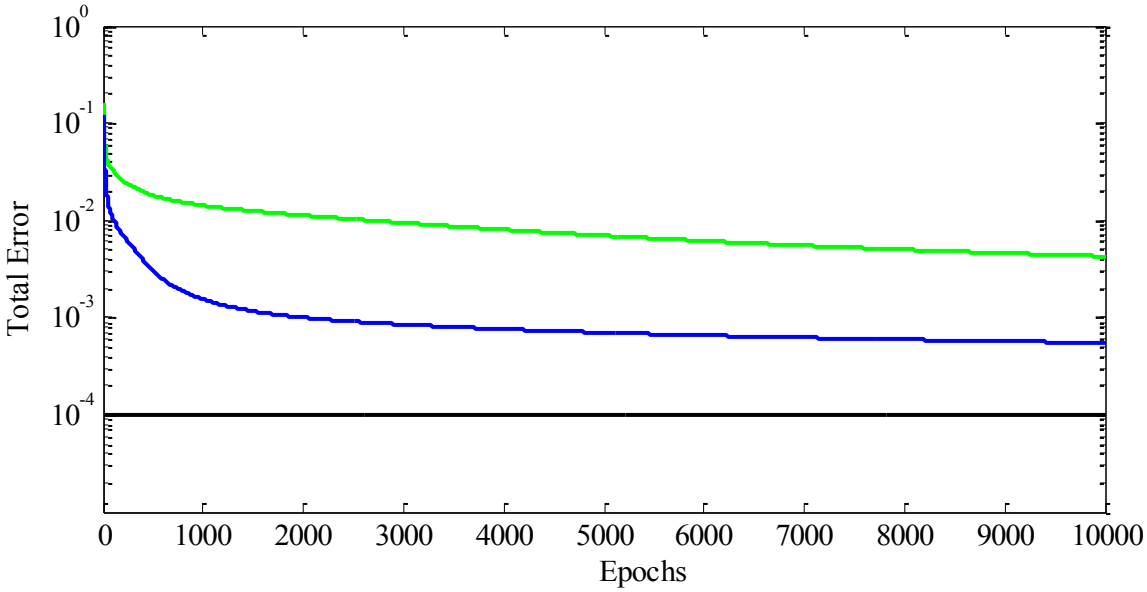


Figure 6.22. (cont.)

As can be seen in the Figure 6.22, ANN training process stopped at which the epochs number 2676 for sunflower oil and mean square error value reached about $111x10^{-6}$ for calibration, on the other hand, due to increasing of mse value of validation program did not continued. Even in case of biodiesel, 2713 iterations were necessary to achieve about $967x10^{-7}$ mse value and validation mse value again increased. For diesel,

epochs reached 10000, which is the limit number, thus program stopped at which the performance was $549x10^{-6}$ mse value.

In training process, if the calibration and validation lines tend to be proceeding in an adjacent trajectory, it demonstrates that the model was constructed successfully. This is an indication that further training would likely to result in the network over-fitting the training set. At this stage, the training process would be terminated. However, output data is required to revise from the 0-1 interval to the actual value since log-sigmoid functions obtain results in the range of 0.1 to 0.9. After these conversion applied, it is possible to calculate the SEC and SEP values for ANN model. Figure 6.23 shows the correlation graphs for FTIR analysis with ANN calibration of this ternary mixture set.

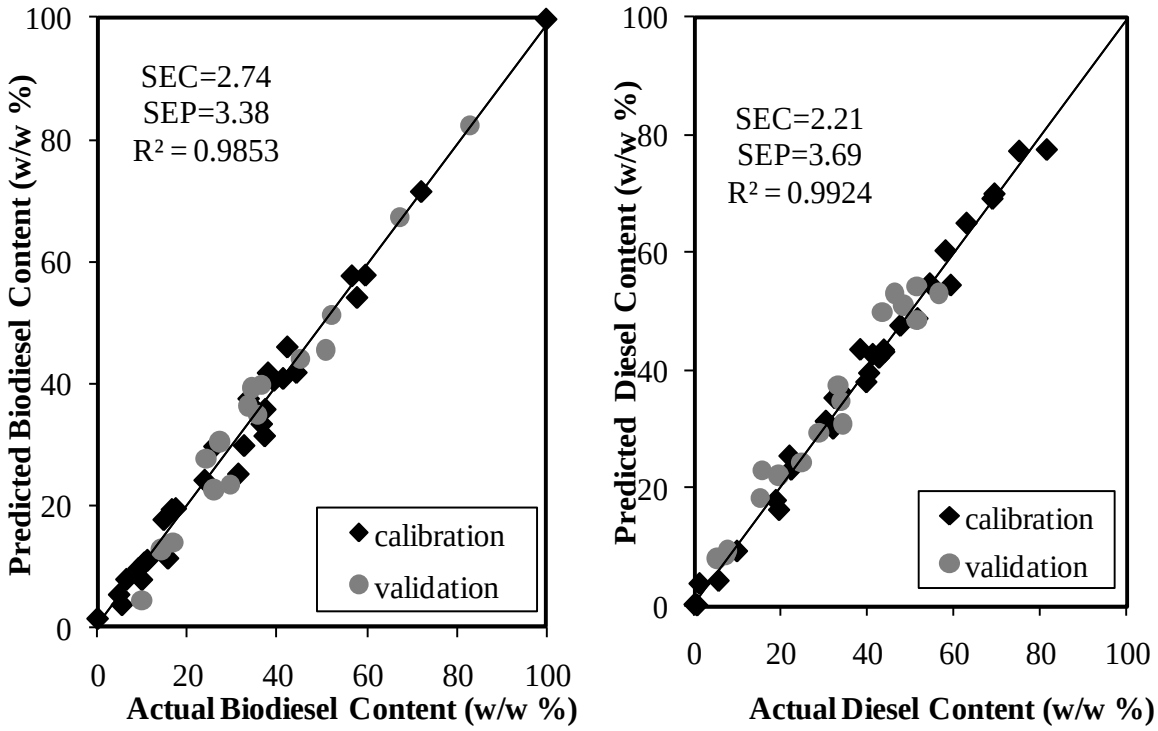


Figure 6.23. Actual versus predicted concentration plot for FTIR-ATR data analysis with ANN calibration method.

(cont. on next page)

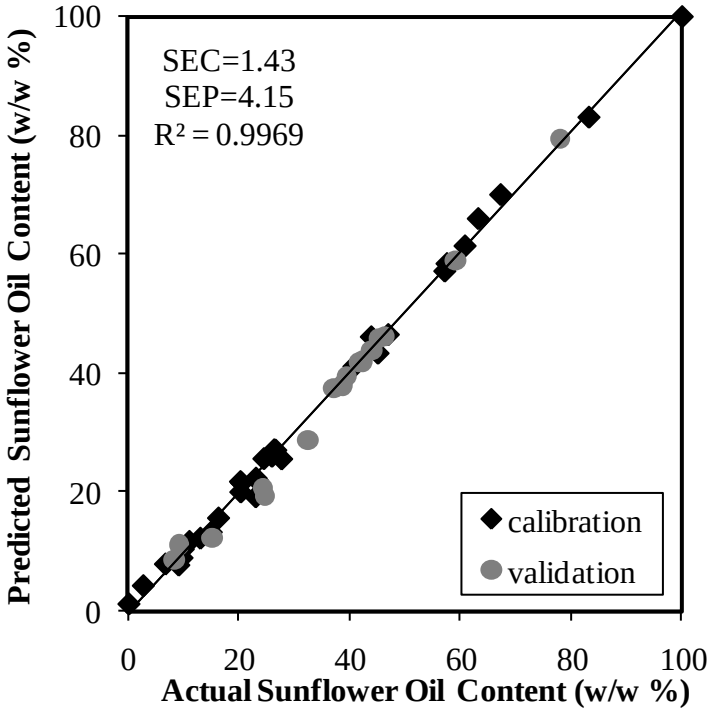


Figure 6.23. (cont.)

As can be seen from Figure 6.23, R^2 values found approximately 0.99. Besides this, SEC values obtained 2.47 for diesel, 3.27 for biodiesel and 2.25 for sunflower oil whereas SEP values 4.40 for diesel, 7.10 for biodiesel and 6.33 for sunflower oil at which the methanol content was up to 20 percentages by mass whereas biodiesel and sunflower oil was up to 100 percentages by mass concentration. Thus, the constructed model is applicable for calibration of biodiesel blends to determine the diesel and untreated vegetable oil content via ANN multivariate calibration modeling.

6.4. Quaternary Mixtures of Biodiesel, Vegetable Oil, Diesel, Methanol Set

As discussed previous sections, adulteration of highly valuable vegetable oils with lower-priced oils has begun to common, widespread, and illegal practice. In addition to this, monitoring residual methanol in biodiesel blends is a matter of safety since even small amounts of this material can reduce the flash point of the biodiesel. Moreover, residua methanol can affect fuel pumps, seals and elastomers an can result in poor combustion properties.

Even though as it is discussed in previous sections, ternary mixture solution of biodiesel in vegetable oils with diesel and methanol separately. However, combining two purposes in one model and determine the methanol and unconverted vegetable oils in biodiesel/diesel blends, with respect to chemometric multivariate calibration techniques along with infrared spectroscopy, have not been conferred in literature. To do this, a total of 60 samples of quaternary mixture of vegetable oils, biodiesels, methanol and diesel components were mixed. 42 samples used for calibration and 18 samples employed for validation. Furthermore, biodiesel, vegetable oils, diesel components consist of three different brand of the individual constituents. For instance, vegetable oils feedstock includes commercially available sunflower, cottonseed and canola oils while biodiesel feedstock has the biodiesels that were synthesized from those vegetable oils. In addition, diesel feedstock contains the mixture of three different brand of commercially available diesel sample.

The range of concentrations each constituent in the sets were in the range of 0 – 100 % for both biodiesel, vegetable oil, diesel by mass percentages but 0 – 20 % for alcohol by mass percentages. Table 6.9 and Table 6.10 show the concentration of the each constituent for biodiesel-vegetable oil-methanol-diesel quaternary set.

Table 6.9. Concentration profiles for calibration set of quaternary mixture of biodiesel, sunflower oil, methanol, diesel set. All concentrations are given by mass percentages. (No: sample number, VO: vegetable oil, B: biodiesel, D: diesel, M: methanol)

No	VO	B	D	M	No	VO	B	D	M
1	100.00	0.00	0.00	0.00	22	12.01	19.93	65.31	2.74
2	0.00	100.00	0.00	0.00	23	22.49	37.55	36.89	3.07
3	0.00	0.00	100.00	0.00	24	21.23	15.59	61.36	1.82
4	0.00	0.00	0.00	5.00	25	24.05	15.77	58.35	1.82
5	49.97	50.03	0.00	0.00	26	10.18	32.24	56.91	0.66
6	50.10	0.00	49.90	0.00	27	21.18	42.99	31.69	4.14
7	0.00	50.05	49.95	0.00	28	11.53	42.29	45.68	0.51
8	33.34	33.24	33.43	0.00	29	5.79	44.17	47.73	2.31
9	12.88	31.92	50.26	4.95	30	15.11	5.83	78.90	0.17
10	25.88	34.93	35.13	4.06	31	13.65	36.71	48.28	1.36
11	14.35	26.99	58.12	0.54	32	21.20	6.00	72.44	0.35
12	19.93	37.83	37.69	4.55	33	9.60	16.50	69.97	3.92
13	19.71	8.09	71.27	0.93	34	24.47	25.75	44.76	5.02
14	17.42	25.30	55.82	1.46	35	14.35	9.25	73.11	3.29
15	17.42	36.78	41.82	3.99	36	12.96	20.30	64.97	1.77
16	22.55	25.20	47.67	4.58	37	7.69	32.75	58.66	0.90
17	24.43	6.13	67.25	2.18	38	2.99	38.13	55.44	3.44
18	16.33	43.54	38.40	1.73	39	23.43	37.37	39.01	0.18
19	11.35	48.19	38.83	1.63	40	6.94	16.34	73.49	3.23
20	18.91	30.26	47.86	2.96	41	17.95	3.43	74.34	4.28
21	19.27	27.08	53.12	0.53	42	16.80	9.34	69.28	4.59

As it is seen in Table 6.10 calibration data set includes the pure form of sunflower oil and biodiesel, diesel with methanol components, in addition.

Table 6.10. Concentration profiles for validation set of quaternary mixture of biodiesel, sunflower oil, methanol, diesel set. All concentrations are given by mass percentages (No: sample no, VO: vegetable oil, B: biodiesel, D: diesel, M: methanol)

No	VO	B	D	M	No	VO	B	D	M
1	24.12	49.92	21.79	4.17	10	2.12	43.86	52.66	1.36
2	1.63	13.77	82.27	2.33	11	4.04	6.17	85.14	4.65
3	7.79	27.24	61.75	3.22	12	1.76	35.66	60.93	1.65
4	1.63	31.30	66.56	0.51	13	2.56	3.54	92.11	1.79
5	9.48	9.91	79.04	1.57	14	5.73	26.92	65.82	4.29
6	16.01	32.78	47.52	3.69	15	8.69	3.67	85.27	2.37
7	18.02	12.43	69.03	0.52	16	18.17	19.28	60.35	2.21
8	19.63	41.35	35.21	3.81	17	14.37	18.52	63.18	3.93
9	21.74	47.24	29.78	1.24	18	1.76	20.55	72.77	4.92

6.4.1. Near Infrared Analysis

NIR measurements of absorbance were performed using 128 scans in the range from 10000 cm^{-1} to 4500 cm^{-1} with a resolution of 16 cm^{-1}. Figure 6.24 illustrates the NIR spectra of quaternary solution with pure forms of components.

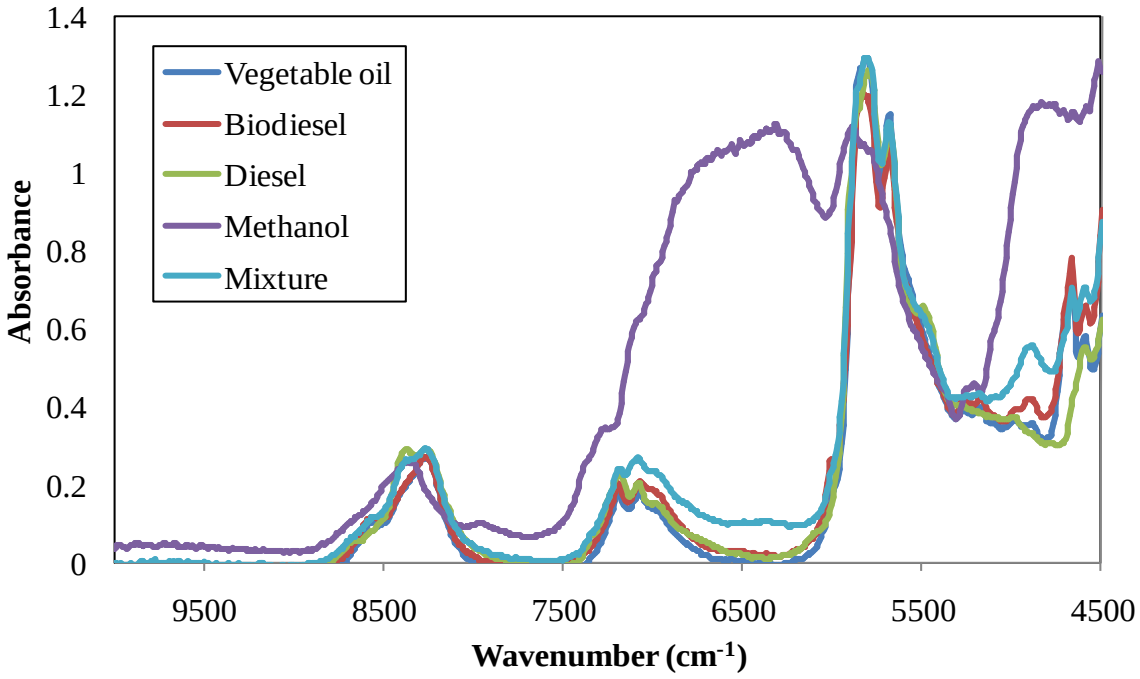


Figure 6.24. NIR spectra of biodiesel, vegetable oil, methanol, diesel and their quaternary mixture.

In the NIR spectral region, the absorbance bands are often broad and overlapping. First, methanol causes a relatively broad peak at 4480-4885 cm^{-1}, as well as a depression of the baseline in the range of 7000-6200 cm^{-1} due to –OH bonds in its structure and –CH bonds seen at 8500-8000 cm^{-1}. The peak at also, 4480-4885 cm^{-1} can be used for quantitation in the fashion discussed above by converting the spectra to absorbance with subsequent chemometric evaluation as described for earlier. In addition, NIR absorption bands corresponding to the -CH=CH, -C-H asymmetric stretching and C=C stretching at around 4600 cm^{-1}. Also, =C-H and C=C stretching belong to –CH=CH- group. The spectra show maximum absorbance at 5800 cm^{-1}. This maximum absorption band is related to first overtone of C-H bond that belongs to $–CH_2$ functional group while absorption band (only a shoulder) at 5680 cm^{-1} is the first overtone of -C-H bond. Besides this, C-H stretching vibrations belongs to weak and broad $-CH_3$ functional group is seen at 7200 cm^{-1} whereas, the other weak and broad shoulder peak at around 7110 cm^{-1} indicates the C-H stretching which leads from CH_2 functional group. Another important peak in spectra is observed at 8285 cm^{-1} that belongs to C-H stretching and represents second overtone. Furthermore, for diesel component the spectral interval from 3700 to 6500 cm^{-1} shows well behaved spectral features presenting absorptions bands that can be attributed largely to the combinations of vibrational modes for the C–H bonds (4500–4000 cm^{-1}) and to the first overtones of C–H bonds (5500–6250 cm^{-1}). The major differences in the FTNIR spectra of methyl esters, biodiesels are observed at 4670–4700 cm^{-1} (aliphatic CH stretching + CO stretching combination modes), 5550–6100 cm^{-1} (first overtone of CH stretching), 6900–7400 cm^{-1} (CH bending + CH stretching combination modes), and 8000–9000 cm^{-1} (second overtone of CH stretching).

6.4.1.1. GILS Results

GILS program run against 100 times along with 50 iterations and 30 genes. Figure 6.21 shows the actual sunflower oil, biodiesel, diesel and methanol concentration values versus their GILS predicted concentration values based on FT-NIR spectral data. The standard error of calibration (SEC) values of each component were found between 0.07 (w/w %) and 1.85 (w/w %) and the standard error of prediction (SEP) values of each component were found between 0.38 (w/w %) and 5.11 (w/w %) by using GILS

method for all components in set for NIR spectral data. SEC and SEP values for vegetable oil content were 1.85 (w/w %) and 5.46 (w/w %) and for methanol content were 0.07 (w/w %) and 0.38 (w/w %) while for diesel component were 1.06 (w/w%) and 4.83 (w/w%) and for biodiesel component were 1.57 (w/w%) and 5.11 (w/w%) respectively. In addition, R^2 values of regression lines for all components were almost 0.99. As a consequent, when these R^2, SEC and SEP values are considered, it is clear that these values are compatible with each other, which demonstrates a good prediction for rapid monitoring the transesterification reaction of biodiesel synthesis to investigate production facilities at which the methanol was up to 5.0 percentages by mass in medium.

In addition to this, to quantify the presence of vegetable oils in biodiesel blends and compare their respective accuracies with respect to adulteration in vegetable oils succeed along with this chemometric approach.

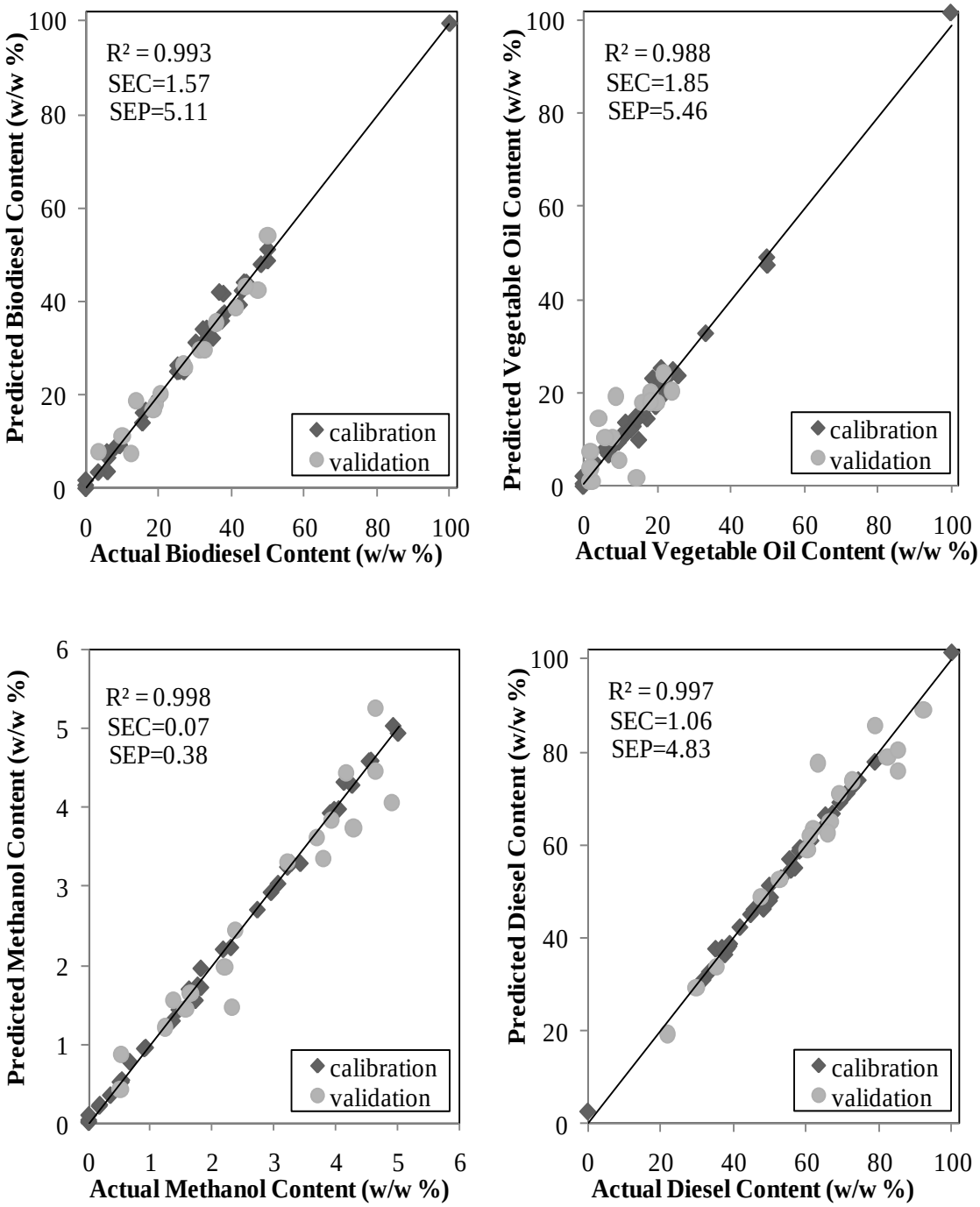


Figure 6.25. Actual versus predicted concentration plot for NIR data analysis with GILS calibration method.

Since GILS is a method that based on wavelength selection, it is significant to examine the distribution of selected wavelengths in multiple runs over the entire full spectral region. Figure 6.26 displays the frequency distribution of selected wavelengths in 100 runs with 30 genes and 50 iterations for biodiesel-vegetable oil-methanol-diesel set.

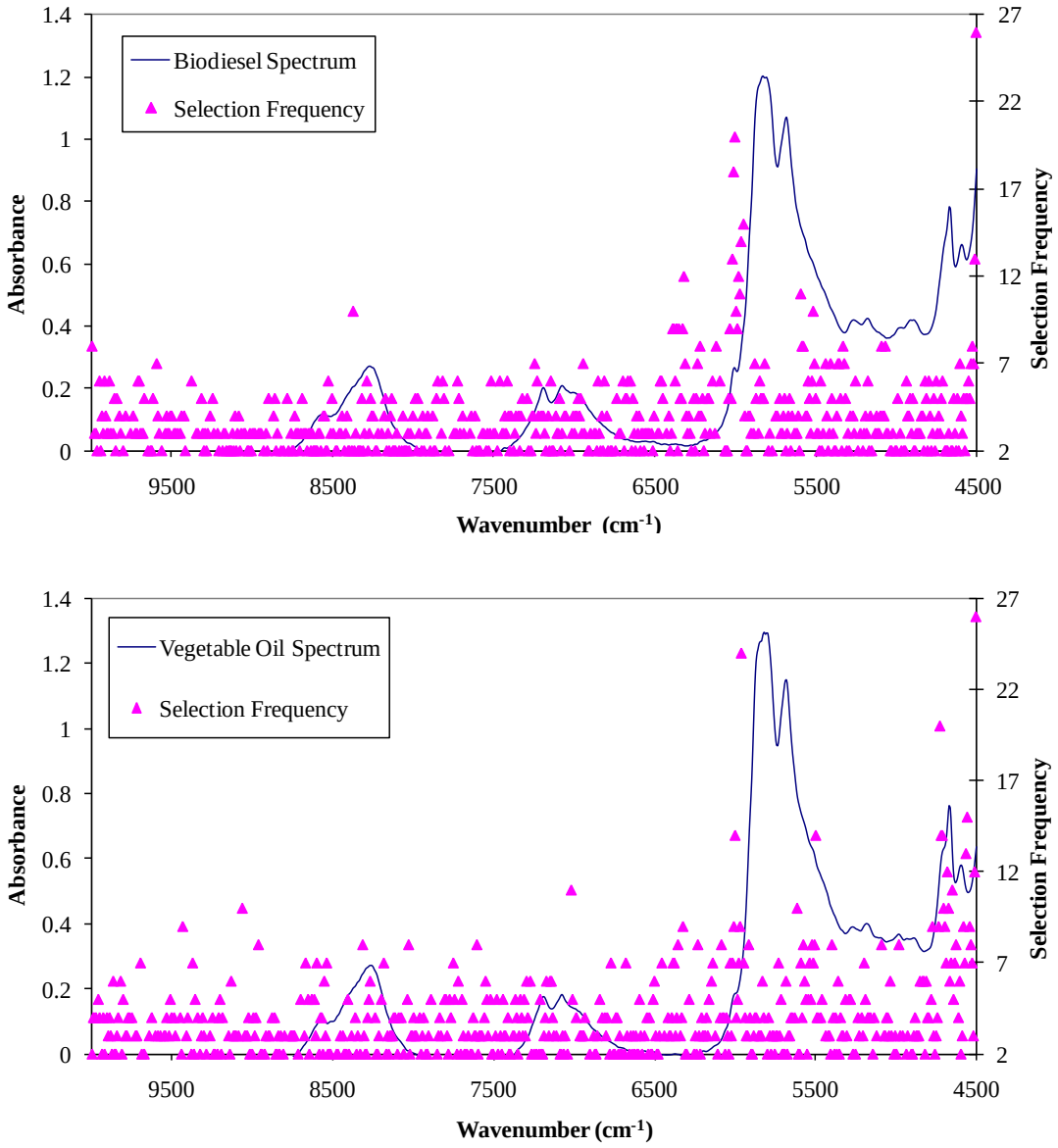


Figure 6.26. Wavelength selection frequency distribution of GILS method for quaternary mixture of biodiesel-vegetable oil-methanol-diesel using NIR spectroscopy.

(cont. on next page)

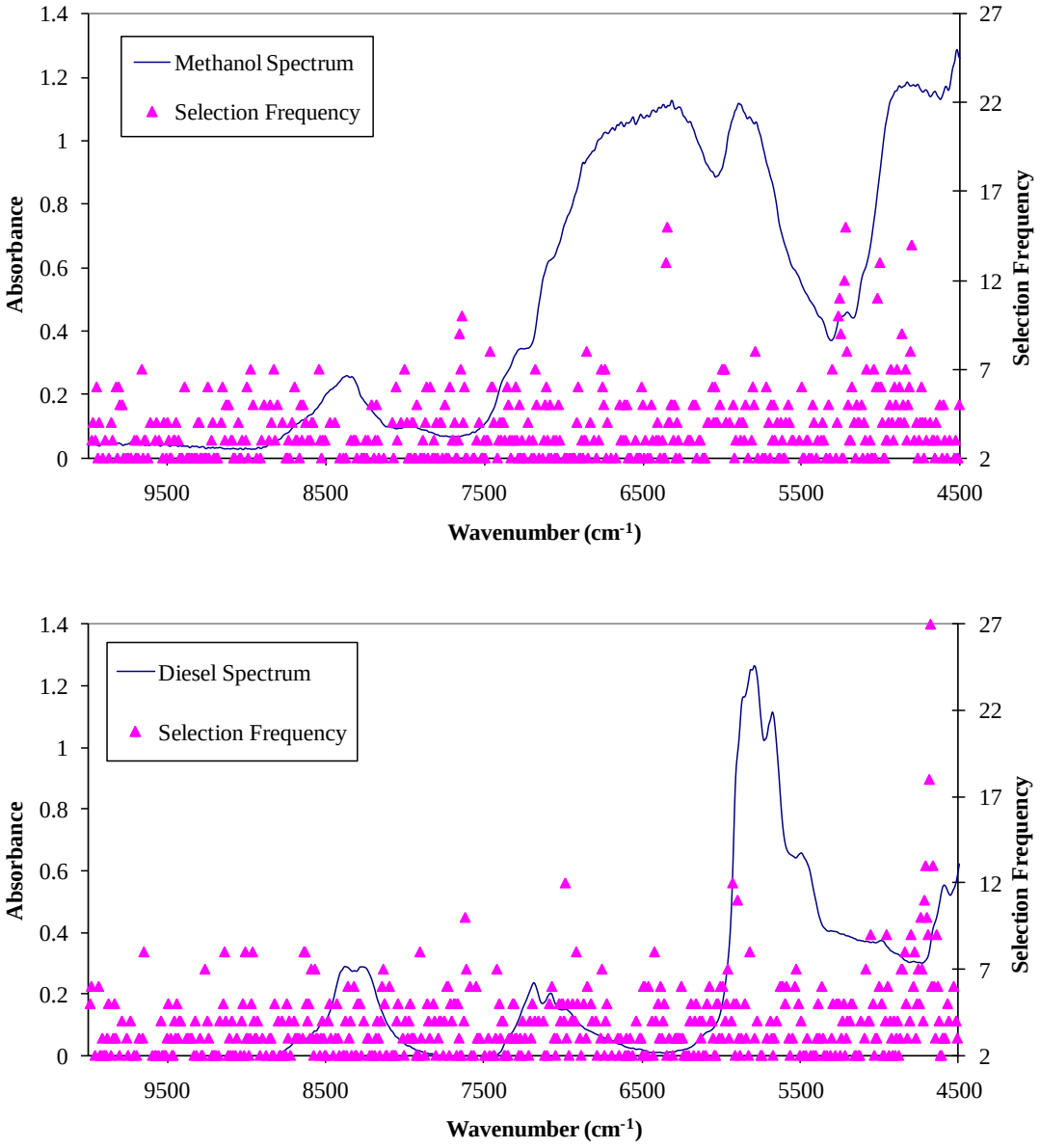


Figure 6.26. (cont.)

As can be seen from the Figure 6.26, there are a number of regions where selection frequencies are highly compared to the rest of the spectrum. The most frequently selected wavenumbers corresponding to the regions at about 5000 and 6000 cm^{-1} where strong peak is observed. This indicates that the genetic algorithm incorporated GILS method is focusing on the regions where most concentration related information is contained.

6.4.1.2. ANN Results

Sample design data set and infrared spectral discussion were same as the previous section that is subtitled as GILS results and also, the parameters that used for ANN calibration method was with a goal of 0.0001 up to maximum value of 10000 epochs numbers with learning rate and momentum value selected as 0.85 and 0.75. The user interface results for each component of quaternary mixture modeled by ANN calibration are given in the Figure 6.27.

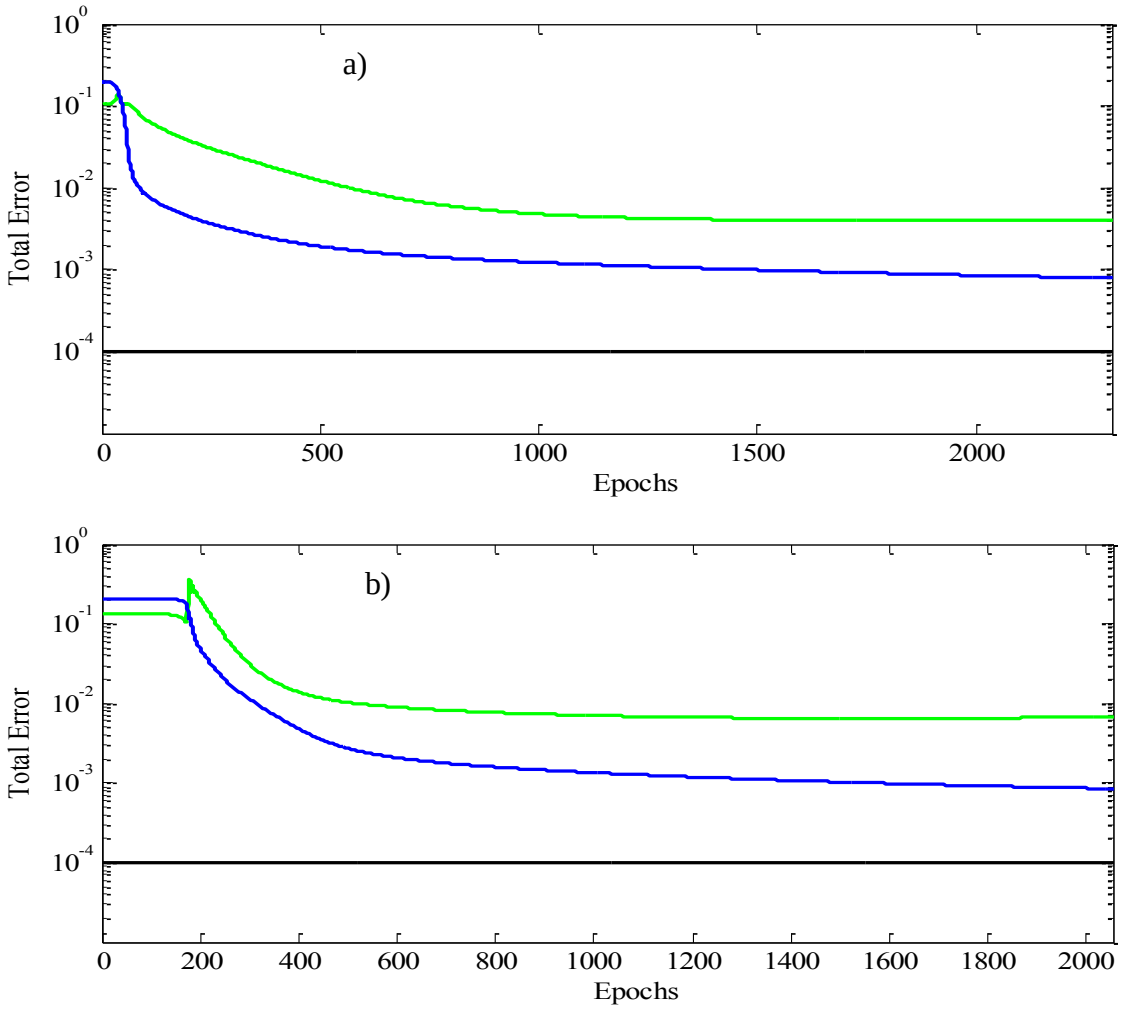


Figure 6.27 Error versus epochs plot of quaternary mixture of a) vegetable oil b) biodiesel c) methanol d) diesel (green line: validation, blue line: calibration, black line: target).

(cont. on next page)

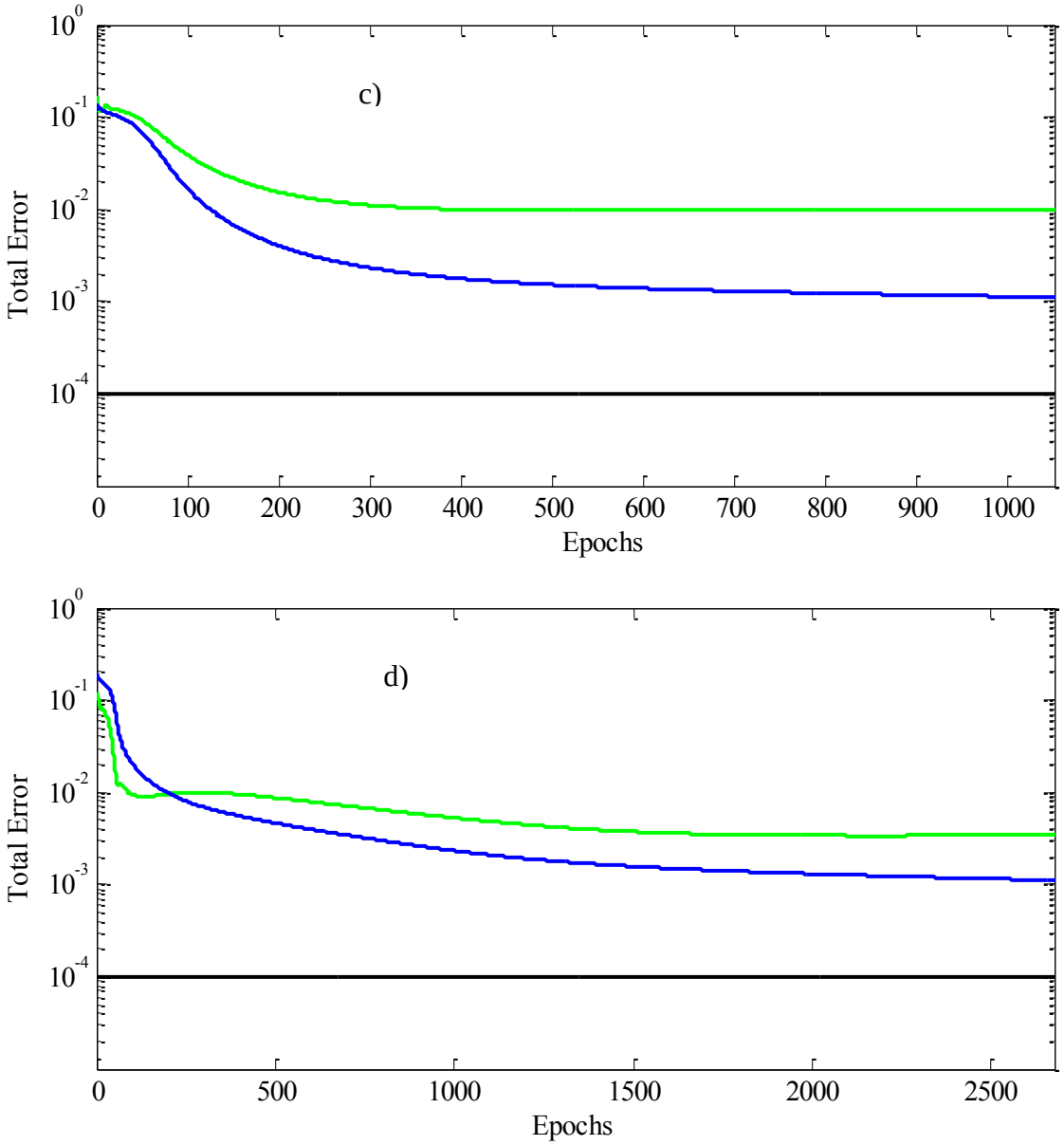


Figure 6.27. (cont.)

As can be seen from the Figure 6.27, for vegetable oil after 2310 iterations performed, the mse value reached about $788x10^{-6}$ for calibration, on the other hand, due to increasing of mse value of validation program was stopped. For biodiesel, 2053 iterations were necessary to achieve about $846x10^{-6}$ mse value and begin validation mse error again. For methanol, epochs reached the 1050, but due to increasing of validation mse value, program stopped at which the performance was $112x10^{-6}$ mse value whereas for diesel mse reached $251x10^{-6}$ at 2758 epochs.

In training process, for a stable case after some iteration, the calibration and validation lines tend to be proceeding in an adjacent trajectory that demonstrates the

model was constructed successfully. This is an indication that further training would likely to result in the network over-fitting the training set. At this stage, the training process would be terminated. However, output data is required to revise from the 0-1 interval to the actual value since log-sigmoid functions obtain results in the range of 0.1 to 0.9. After these conversion applied, it is possible to calculate the SEC and SEP values for ANN model. Figure 6.28 shows the correlation graphs for NIR analysis with ANN calibration of this quaternary mixture set

Furthermore, Figure 6.28 illustrates the R^2 values found between 0.96 and 0.99. In addition to this, SEC values obtained was 0.99 for methanol, was 2.93 for biodiesel and was 1.49 for sunflower oil whereas SEP values was 1.69 for methanol, was 8.63 for biodiesel and was 5.58 for sunflower oil. Consequently, simple for any business to take advantage of rapid reliable NIR analysis – a goal achieved with the ready to- use ANN calibrations.

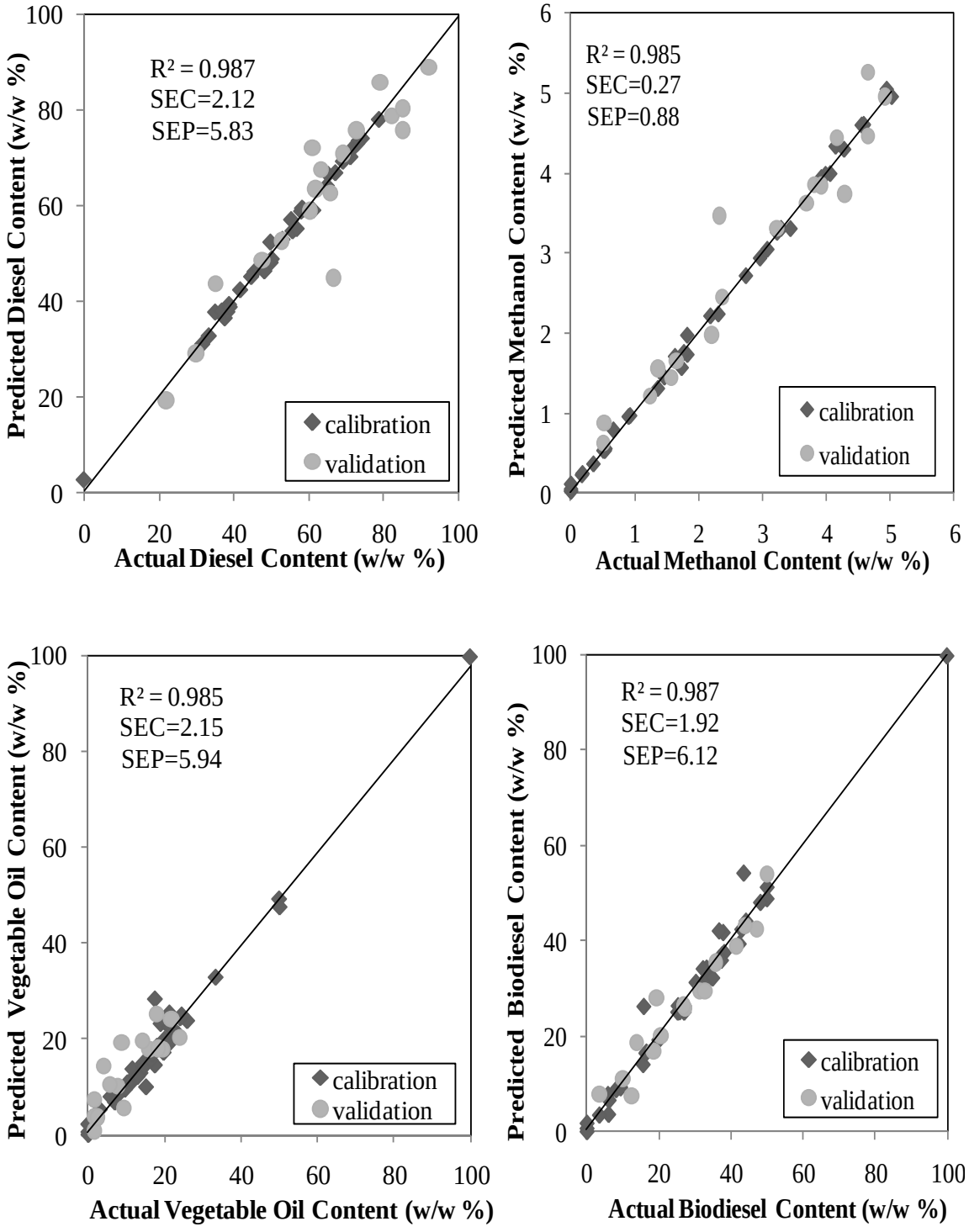


Figure 6.28. Actual versus predicted concentration plot for NIR data analysis with ANN calibration method.

6.4.2. Mid Infrared Analysis

The FTIR spectra of biodiesel blends in oil and methanol, which were shown in Figure. 6.29, are as expected very similar since the oil, biodiesel and diesel compounds have almost the same chemical groups. However, some differences are detectable. The position of the carbonyl band in FTIR is sensitive to substituent effects and to the structure of the molecule. The methoxycarbonyl group in biodiesel shows a different band position of the νC=O vibration when compared to the carbonyl band in the oil. The peak of this band changed from 1743 cm^{-1} in oil to 1740 cm^{-1} in biodiesel. The band due to the νC(=O)-O vibration shows a peak at 1235 cm^{-1} in the oil and at 1244 cm^{-1} in biodiesel. A new band at 1195 cm^{-1} was observed for biodiesel and was attributed to the ρMe vibration. The band observed at 1159 cm^{-1} in the oil is observed in biodiesel at 1169 cm^{-1}.

This band was attributed to methyl groups near carbonyl groups. There are no such absorptions due to C=O and C-O functional groups in petroleum diesel. In addition, the major changes are observed mainly at 3011 and 1654 cm^{-1}. These absorptions are assigned to ν(CH) and ν(CC) stretching modes characteristic of olefins. Note that the FTIR spectrum of the vegetable oil methyl ester presents the higher intensities at these spectral regions due to fact that the vegetable oil presents the largest content of unsaturated carbon atoms. Remarkable differences are also observed in the finger print region (1100–1500 cm^{-1}). The IR absorptions of different types of stretching, bending and out-of-plane vibrations of C-H, C=O, C-C-O, C-OH and C-O bonds are also useful for the multivariate analysis of biodiesel blends.

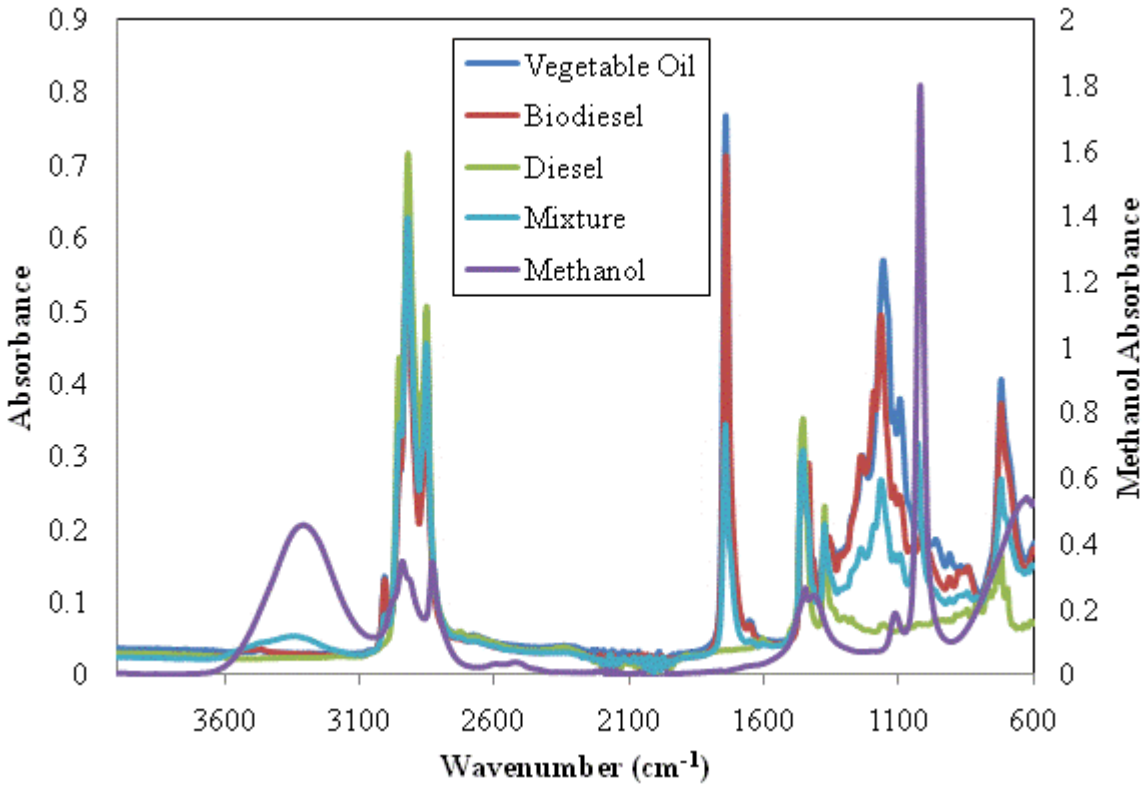


Figure 6.29. FTIR-ATR spectra of biodiesel, vegetable oil, methanol, diesel and their quaternary mixtures.

This spectral separation among the functional groups of vegetable oils, biodiesel and petroleum diesel forms the basis of characterization and quantitation of FAMEs in biodiesel and in blended biodiesel-diesel fuel through IR spectroscopy combined with the chemometric calibration methods such as based on GILS and ANN approach.

6.4.2.1. GILS Results

Since utilizing the GILS algorithm, method will decrease the effect of baseline shifts because it can select certain combination of wavelengths, which have maximum correlation with biodiesel blends. The sample design set generated from 42 of them as calibration set and the remaining 18 samples as validation samples performed as it is same in the FTIR-ATR spectral data analysis.

GILS program run against 100 times along with 50 iterations and 30 genes to predict the concentration of quaternary mixture of biodiesel, vegetable oil, diesel and methanol. Figure 6.30 shows the actual pure form of quaternary mixture concentration

values versus their GILS predicted concentration values based on FTIR-ATR spectral data. Calibration models for methanol content determination gave standard error of calibration (SEC) and standard error of prediction (SEP) values as 0.17 % (w/w) and 0.43% (w/w) respectively. In the case of vegetable oil and biodiesel content determination, the SEC and SEP values were 0.33% (w/w), 0.31% (w/w) and 0.34% (w/w), 0.39% (w/w) whereas for diesel component SEC and SEP values were 0.55% (w/w) and 0.51% (w/w) for calibration and prediction sets, respectively. In addition to this, when examining the correlation plots the R^2 value of regression lines was 0.99 which indicates the model succeed in a reasonable manner.

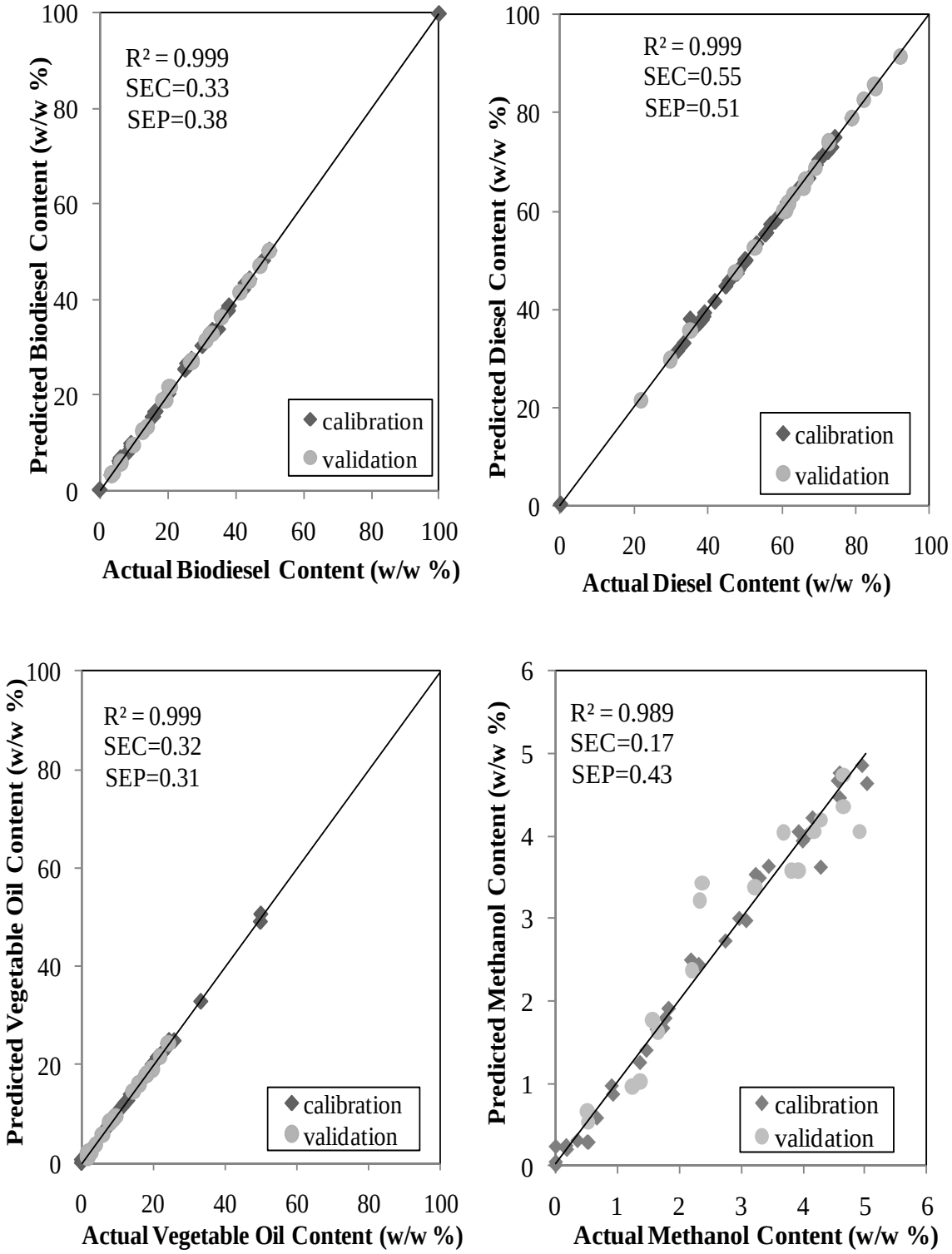


Figure 6.30. Actual versus predicted concentration plot for FTIR-ATR spectral data analysis with GILS calibration method.

However, GILS method is an iterative procedure due to the genetic algorithm used to select a subset of wavelengths from the complete spectral range and Figure 6.31 displays the frequency distribution of selected wavelengths in 100 runs with 30 genes and 50 iterations for biodiesel-vegetable oil-methanol-diesel set with respect to FTIR-ATR spectra data matrices.

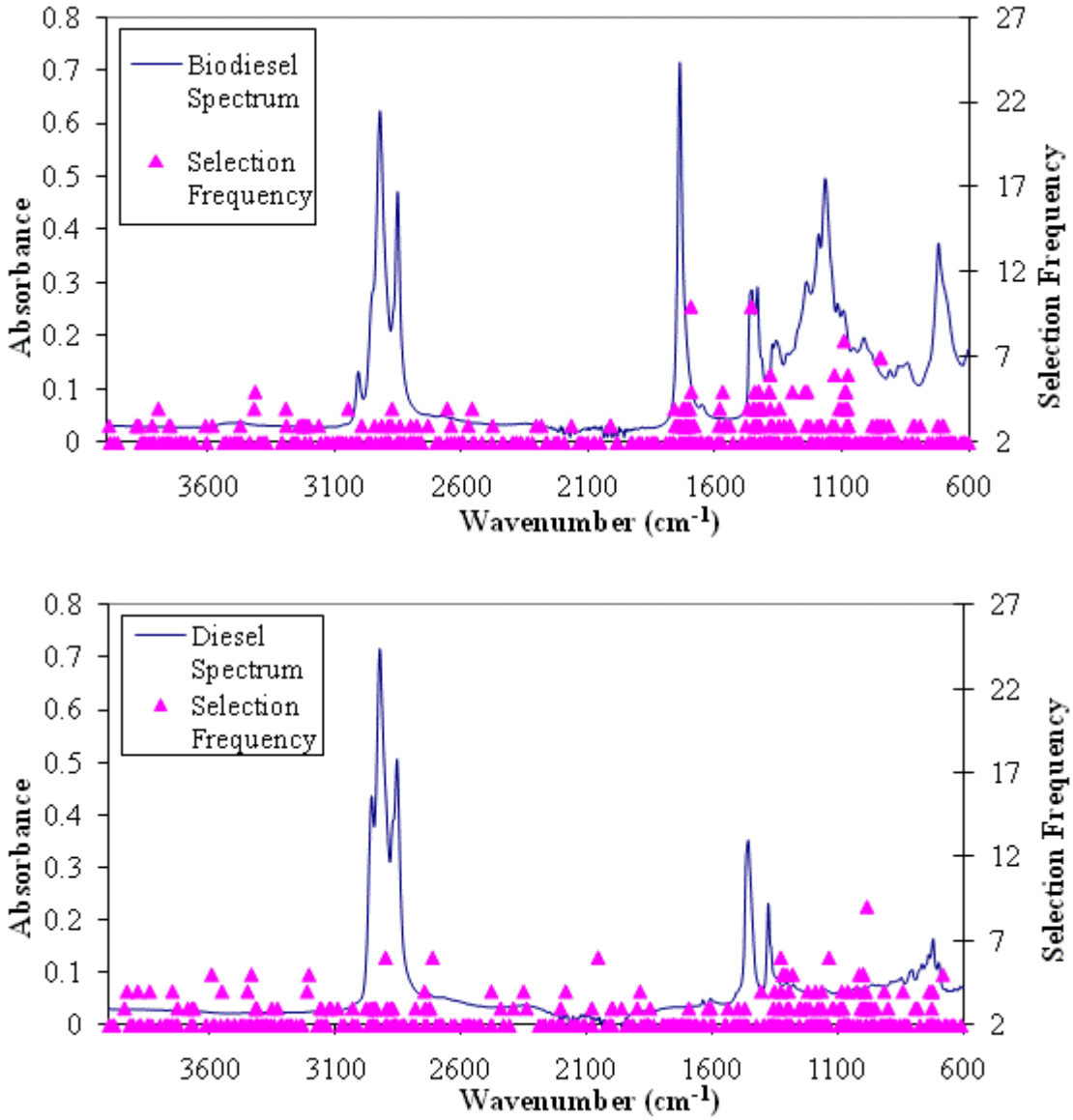


Figure 6.31. Wavelength selection frequency distribution of GILS method for quaternary mixture of biodiesel-vegetable oil-methanol-diesel using FTIR-ATR spectroscopy.

(cont. on next page)

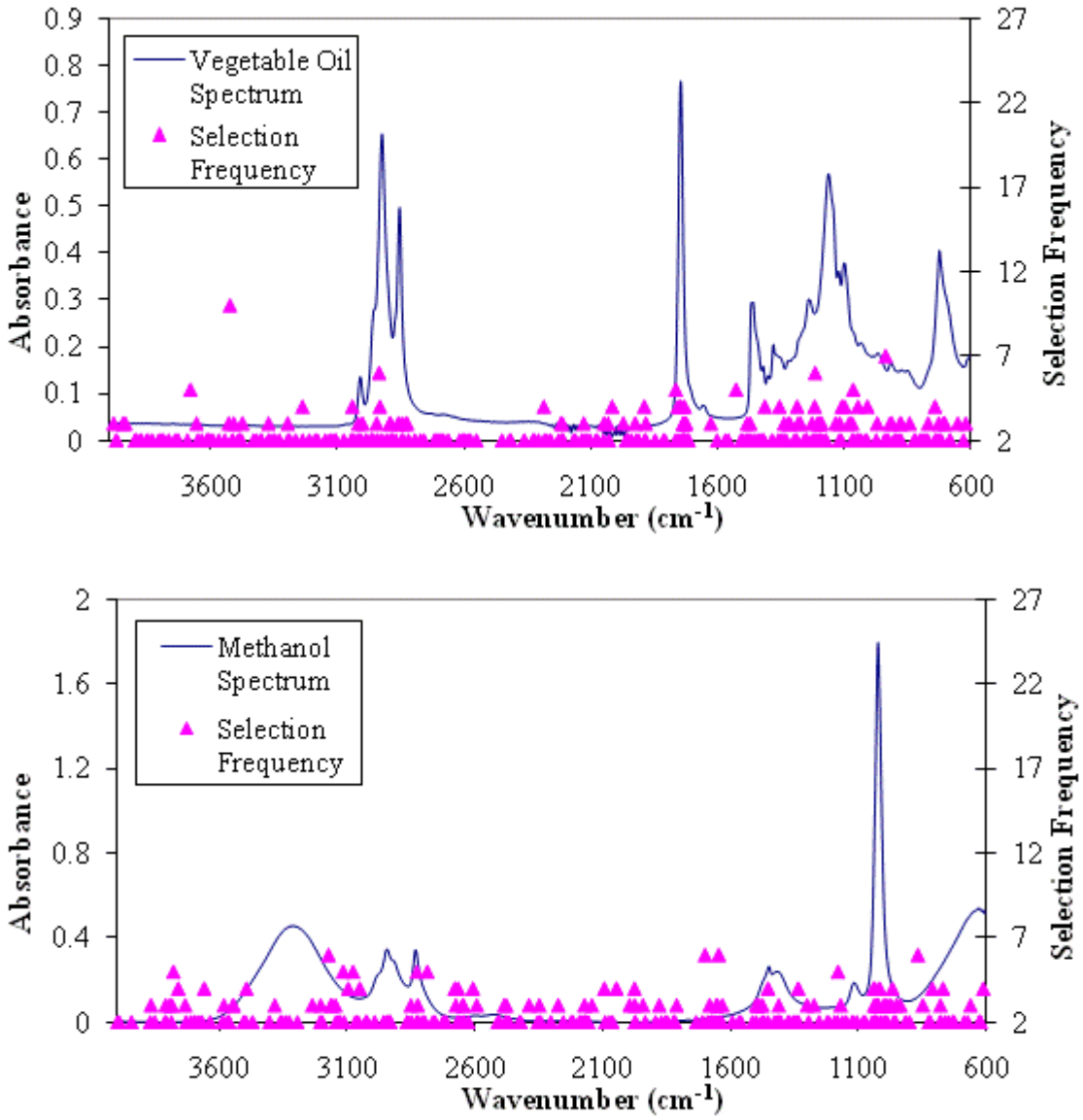


Figure 6.31. (cont.)

As can be seen from the Figure 6.31 there are a number of regions where selection frequencies are highly compared to the rest of the spectrum. The wavenumber region around 1000 and 3000 cm^{-1} for all components indicates a strong tendency for GILS method to select while for biodiesel and sunflower oil content, around 500 cm^{-1} is the most frequently selected region. This indicates that the genetic algorithm incorporated GILS method is focusing on the regions where most concentration related information is contained.

6.4.2.2. ANN Results

A total of 60 samples used for the ANN data analysis at which 42 biodiesel blends samples were used for calibration model optimization and training. The root mean squared error was used to characterize the prediction capacity of the created model and to optimize its parameters. 18 samples employed as validation data, which were used for early stopping for network and to check the multivariate model accuracy on an independent data set. Sigmoid function logsig generates outputs between 0 and 1 as the neuron's net input goes from negative to positive infinity and data matrices of quaternary mixtures as can be arranged in the NIR spectral data analysis as well as the same parameters utilized. Launching Neural Network Toolbox in Matlab 7.0.1 software programming in terms of the configurations as it is mentioned earlier, the user interface results for each component of quaternary mixture modeled by ANN calibration are given in the Figure 6.32.

In training process, for a stable case, if the calibration and validation lines tend to be proceeding in an adjacent trajectory, it demonstrates that the model was constructed successfully. This is an indication that further training would likely to result in the network over-fitting the training set. At this stage, the training process would be terminated. However, output data is required to revise from the 0-1 interval to the actual value since log-sigmoid functions obtain results in the range of 0.1 to 0.9. After these conversion applied, it is possible to calculate the SEC and SEP values for ANN model. Figure 6.33 shows the correlation graphs for FTIR analysis with ANN calibration of this ternary mixture set.

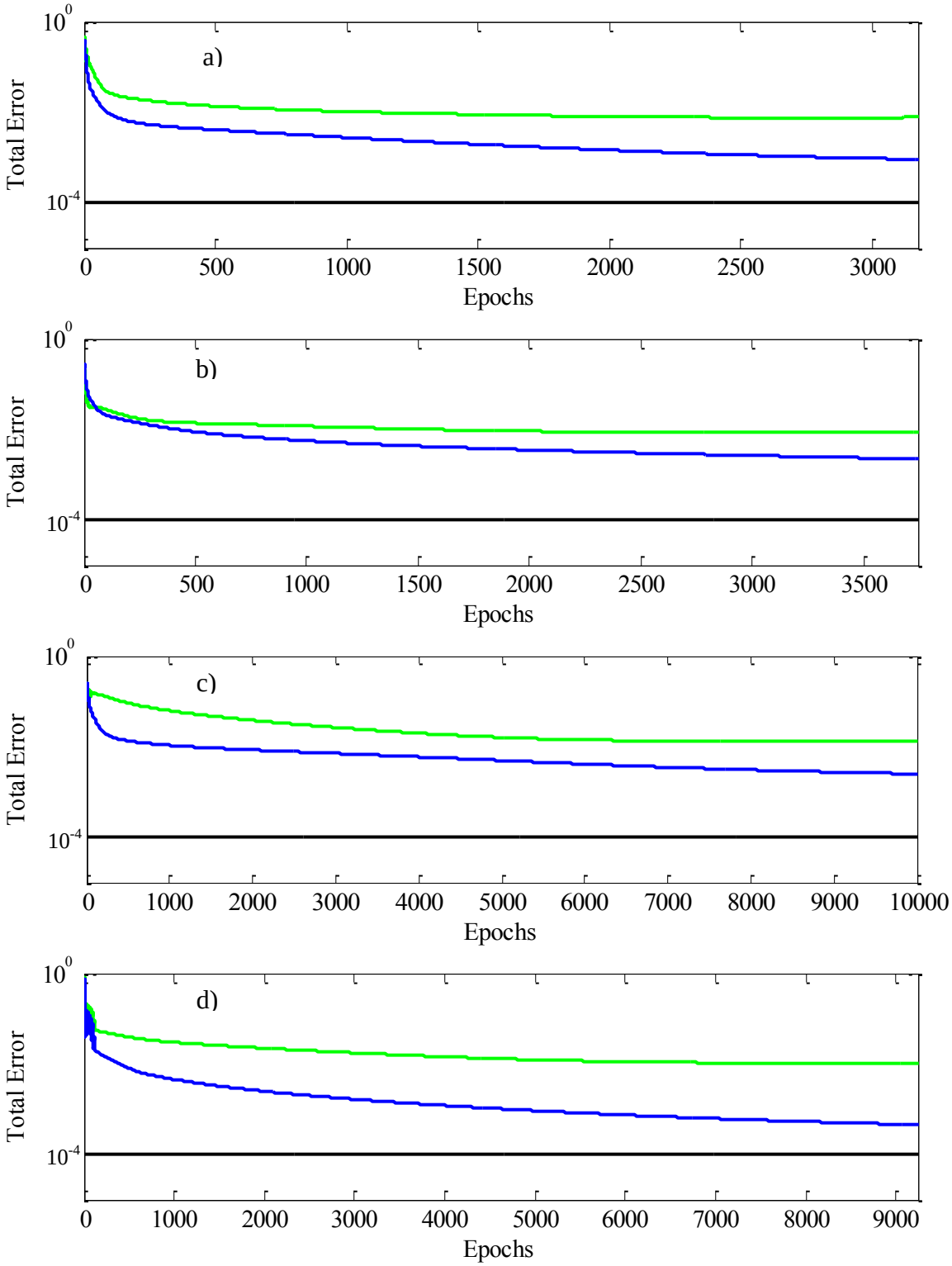


Figure 6.32. Error versus epochs plot of quaternary mixture of a) vegetable oil b) biodiesel c) methanol d) diesel (green line: validation, blue line: calibration, black line: target).

As can be seen in the Figure 6.32, ANN training process stopped at which the epochs number 3176 for vegetable oil and mean square error value reached about $906x10^{-6}$ for calibration, on the other hand, due to increasing of mse value of validation program did not continued. Even in case of biodiesel, 3741 iterations were necessary to achieve about $2241x10^{-6}$ mse value and validation mse value again increased. For methanol, epochs reached 10000, which is the limit number, thus program stopped at which the performance was $249x10^{-6}$ mse value whereas diesel reached $454x10^{-6}$ mse value after 9239 iterations.

Output data is required to revise from the 0-1 interval to the actual value since log-sigmoid function obtains results in from 0.1 to 0.9. After these conversion applied, it is possible to calculate the SEC and SEP values along with the regression values between actual and predicted by ANN model. Thus, Figure 6.33 shows the correlation graphs for FTIR analysis with ANN calibration of quaternary mixture of biodiesel blends.

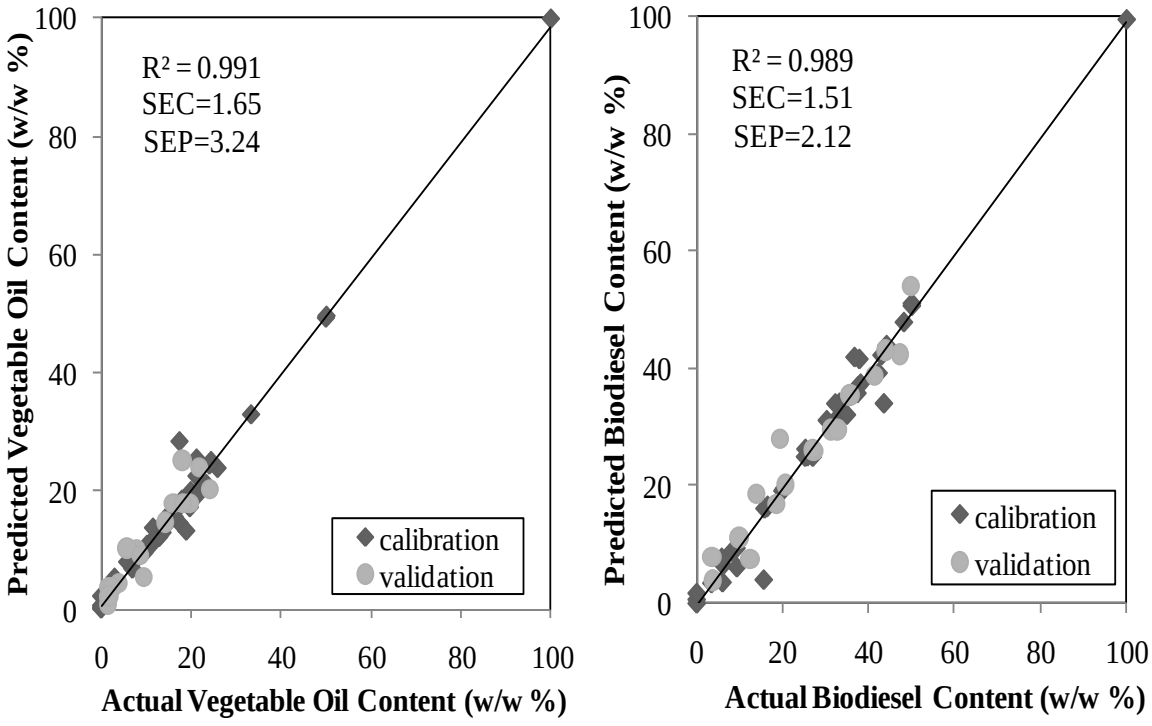


Figure 6.33. Actual versus predicted concentration plot for FTIR-ATR data analysis with ANN calibration method.

(cont. on next page)

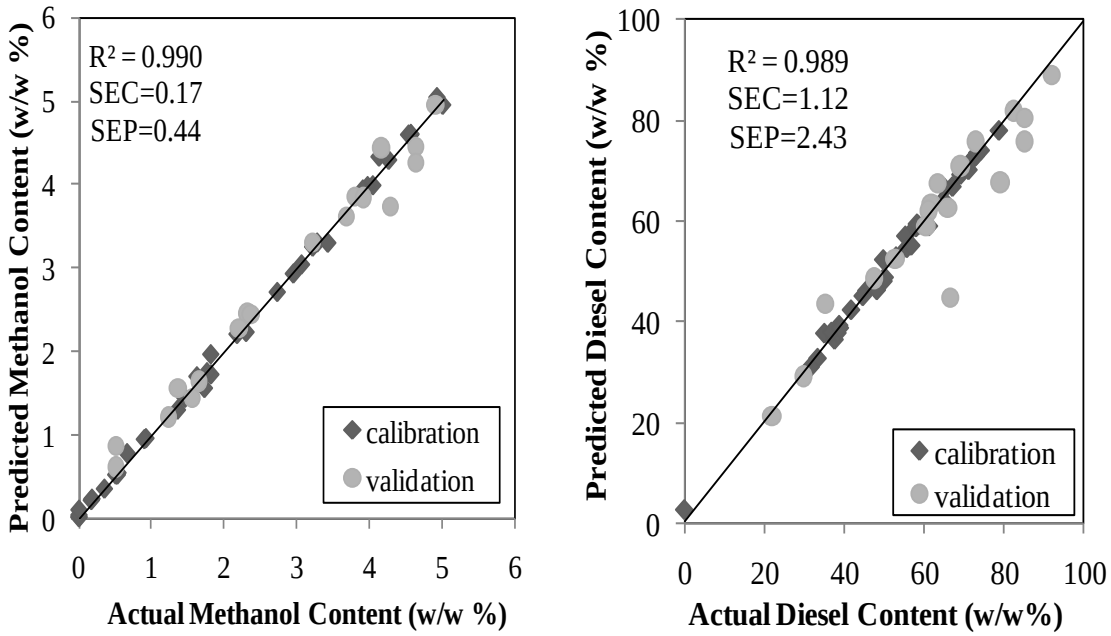


Figure 6.33. (cont.)

As can be seen from Figure 6.33, R^2 values found approximately 0.985. Besides this, SEC values obtained 0.17 for methanol, 1.51 for biodiesel, 1.65 for vegetable oil and 1.12 for diesel components whereas SEP values 0.44 for methanol, 2.12 for biodiesel, 2.43 for diesel and 3.24 for sunflower oil. Methanol content was only up to 5% by mass whereas biodiesel, vegetable oil and diesels were 100% by mass concentration in sample set. Thus, the constructed model is applicable for calibration of biodiesel blends to determine the methanol and untreated vegetable oil content via ANN that can be an effective multivariate calibration modeling for quality analysis of biodiesel blends.

6.5. Calibration Summary

To make a comparison between NIR spectroscopy and MIR spectroscopy along with the chemometric multivariate calibration techniques which based on GILS and ANN approach; SEC, SEP and R^2 values are given for all three data sets in Table 6.11, Table 6.12 and Table 6.13, respectively.

Table 6.11. Calibration summary for ternary mixture of biodiesel, oil and methanol data sets. All SEC and SEP values are given as mass percentages (w/w %).

Ternary mixtures	**Biodiesel**			**Oil**			**Methanol**		
Methanol set	SEC	SEP	R^2	SEC	SEP	R^2	SEC	SEP	R^2
NIR-GILS	1.86	2.88	0.997	1.21	2.75	0.994	0.89	0.91	0.997
NIR-ANN	2.93	8.63	0.986	1.49	5.58	0.995	0.99	1.69	0.966
FTIR-GILS	2.01	2.61	0.993	0.58	1.33	0.999	1.08	1.03	0.995
FTIR-ANN	1.87	2.45	0.994	1.48	2.61	0.996	0.43	0.95	0.993

Table 6.12. Calibration summary for ternary mixture of biodiesel, oil and diesel data sets. All SEC and SEP values are given as mass percentages (w/w %).

Ternary mixtures	**Biodiesel**			**Oil**			**Diesel**		
Diesel set	SEC	SEP	R^2	SEC	SEP	R^2	SEC	SEP	R^2
NIR-GILS	0.90	1.32	0.998	1.04	1.46	1.00	0.94	1.07	0.999
NIR-ANN	1.43	5.88	0.996	2.58	5.32	0.989	1.63	6.52	0.996
FTIR-GILS	0.20	0.82	0.999	0.34	0.82	0.999	0.20	0.34	0.999
FTIR-ANN	2.74	3.38	0.985	1.43	4.15	0.996	2.21	3.69	0.992

Table 6.13. Calibration summary for quaternary mixture of biodiesel, oil, methanol, diesel data sets. All SEC and SEP values are given as mass percentages.

Quaternary mixtures	**Biodiesel**			**Oil**			**Diesel**			**Methanol**		
	SEC	SEP	R^2	SEC	SEP	R^2	SEC	SEP	R^2	SEC	SEP	R^2
NIR-GILS	1.57	5.11	0.993	1.85	5.46	0.998	1.06	4.38	0.997	0.07	0.38	0.998
NIR-ANN	1.92	6.12	0.987	2.15	6.19	0.985	2.12	5.83	0.987	0.37	0.88	0.985
FTIR-GILS	0.33	0.38	0.999	0.32	0.31	0.999	0.55	0.51	0.999	0.17	0.3	0.989
FTIR-ANN	1.51	2.12	0.989	1.65	3.24	0.991	1.12	2.43	0.989	0.24	0.44	0.990

After taking into an account for all, it is bring to a close that biodiesel blends samples with MIR spectroscopic techniques much better than NIR measurements in the calibration. The sigmoid ANN method gave clear improvements in prediction ability. However, GILS does not only help to calibrate the samples but also select a few wavelengths that contain the necessary information. In the future, both the algorithms can be improved by adding classification steps after construction of models. Again, it must be realized that the GILS method is an iterative procedure due to the genetic algorithm used to select a subset of wavelengths from the complete spectral range. The effect of baseline fluctuation will be more since MIR region is very sensitive for quantitative analysis because absorbance changes become more than it becomes in NIR case. The reason can be that fundamental vibrations have more probability to be observed than overtones. Yet, when the overall calibration performance of the models examined, it is possible to state that the MIR spectra do contain quantitative information that correlated with diesel, vegetable oil and methanol contents of the biodiesel blends samples studied here.

CHAPTER 7

CONCLUSION

This work presents multivariate chemometric calibration methods based on genetic algorithm inverse least square (GILS) and artificial neural networks (ANN) for the determination of oils, diesels, biodiesel and methanol in blends form.

The analytical issues with biodiesels have two sources. The production facilities and terminal services need to ensure quality (completion of transesterification, glycerol removal, etc.) while testing labs and regulatory agents must ensure the labeled blend levels are present. The former are generally concerned with high FAME content materials (B100), while the latter may be exposed to a wide range of FAME content, from B2 and B20 up to B100. Since potential contaminants of biodiesel can arise during the transesterification reaction, it is important for biodiesel producers to be able to monitor the status of biodiesel production in order to recognize and correct any problems at an early stage. Infrared provides a rapid, precise and accurate tool for this analysis when these needs are taken into account.

Analytical methods have been developed using mid infrared (FTIR) and near infrared (NIR) spectroscopy conveniently to determine biodiesel content in the reaction mixture to monitor the transesterification reaction. It is also shown that it can be used to determine biodiesel content in biodiesel-petrodiesel blends. The method with small modifications can also be used to determine the oil content in the adulteration of biodiesel-petrodiesel blends. It is shown that the method can be used to measure the amount of biodiesel accurately to the extent of 98 % accuracy for biodiesel-oil mixtures and biodiesel content in the biodiesel-petrodiesel mixture (blend) with an accuracy of 99%. Transesterification reaction, which yields the methyl esters, can be monitored for completion by near infrared (NIR) spectroscopy using a fiber-optic probe or mid-infrared spectroscopy with ATR crystal attachment in future.

REFERENCES

Abu-Mostafa, Y. 1990. Learning from Hints in Neural Networks, *Journal of Complexity*, Vol. 6, pp. 192–198.

Agency for Toxic Substances and Disease Registry (ATSDR), 1995. Toxicological profile for fuel oils. Atlanta, GA: U.S. Department of Health and Human Services, Public Health Service.

Albrecht, R. F., Reeves C.R., Steele N.C. 1993. *Artificial Neural Nets and Genetic Algorithms*, Springer-Verlag, Vienna.

Aleksander, I., and Morton H. 1990. *An Introduction to Neural Computing*, Chapman and Hall, London.

Barron, A.R. 1998 *in Nonparametric Functional Estimation* (Ed. by G. Roussas) Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 1034–1053.

Beebe, K.R., Pell, R.J., and Seasholtz, M.B. 1998. *Chemometrics, a practical guide.* Wiley-Interscience: John Wiley & Sons, Inc.

Bos, A., Bos, M. and Linden, van der W.E. 1993. Artificial neural networks as a multivariate calibration tool: modeling the Fe-Cr-Ni system in x-ray fluorescence spectroscopy, *Analytica Chimica Acta*, 277, 289-295.

Brereton, R.G. 2000. Introduction to multivariate calibration in analytical chemistry. *The Analyst* 125:2125-2154.

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

Bro, R. 2003. Multivariate calibration What is in chemometrics for the analytical chemist?, *Analytica Chimica Acta* 500, 185–194.

Campbell M. 2008. Biodiesel: Algae as a Renewable Source for Liquid Fuel. *Guelph Engineering Journal* (1), 2 -7 ISSN: 1916-1107.

Cong, P. and Li, T. 1994. Numeric genetic algorithm part I. theory, algorithm and simulated experiments. *Analytica Chimica Acta* 293:191-203.

Danzer, K., Currie, L.A. 1998. *Pure Appl. Chem*. 70, 993.

Danzer, K, Otto, M, Currie, O. 2004. Guidelines for calibration in analytical chemistry, Part 2. Multispecies calibration, *Pure Appl. Chem.*, Vol. 76, No. 6, pp. 1215–1225.

Demirbas. A; 2008. Relationships derived from physical properties of vegetable oil and biodiesel fuels. *Fuel* 87, 1743–1748.

DeThomas, F.A., Hall, J.W., Monfre, S.L. 1994. Real-time monitoring of polyurethane production using near infrared spectroscopy. *Talanta* 41:425-431.

Divya, O. and Mishra, A.K. 2007. Combining synchoronus fluorescence spectroscopywith multivariate methods for the analysis of petrol-kerosene mixtures. Talanta 72:43-48.

Ferrão M.F., Viera M.S., Pazos R.E.P, Fachini D., Gerbase A.E., Marder L. 2011. Simultaneous determination of quality parameters of biodiesel/diesel blends using HATR-FTIR spectra and PLS, iPLS or siPLS regressions. *Fuel*, Volume 90, Issue 2, February 2011, Pages 701-706.

Fontain, E. 1992. The problem of atom-to-atom mapping. An application of genetic algorithms. *Analytica Chimica Acta* 265:227-232.

Gaydou V., Kister J., Dupuy N. 2011. Evaluation of multiblock NIR/MIR PLS predictive models to detect adulteration of diesel/biodiesel blends by vegetal oil. *Chemometrics and Intelligent Laboratory Systems*, Volume 106, Issue 2, 15 April 2011, Pages 190-197.

Gerpen, J, Shanks, B, Pruszko, R. 2004. Biodiesel Production Technology, National Renewable Energy Laboratory. Subcontractor Report, NREL/SR-510-36244.

Gilbert, R. J., Goodacre, R., Woward, A. N., and Kell, D.B. 1997. Genetic programming: a novel method for the quantitative analysis of pyrolysis mass spectral data. *Analytical Chemistry 69*:4381-4389.

Griffiths, P.R. 1978. *Chemical infrared Fourier transform spectroscopy*, New York: Wiley.

Haaland, D.M. and Thomas, E.V. 1988. Partial least-squares methods for spectral analyses relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry* 60:1193-1202.

Hibbert, D.B. 1993. Genetic algorithms in chemistry. *Chemometrics and Intelligent Laboratory Systems* 19:277-293.

IUPAC. Chemometrics, IUPACWiki. 2010. http://www.iupacterms.eigenvector.com/ index.php?title= chemometrics (accessed May 9, 2011).

International Energy Agency, IEA. Statistics. 2011. http://www.iea.org/stats/index.asp (accessed June 2, 2011).

Intergovernmental Panel on Climate Change, IPCC. 2007. http://www.ipcc.ch/ publications_and_data/publications_and_data.shtml (accessed June 12, 2011).

Koenig, J.L. 1975. Application of Fourier transform infrared spectroscopy to chemical systems. *Applied Spectroscopy* 29:293-308.

Kowalski, B.R. 1983. *Chemometrics Mathematics and Statistics in Chemistry*, NATO ASI Series C: Mathematical and Physical Sciences Vol. 138.

Li, T., Lucasius, C. B., Kateman, G. 1992. *Anal. Chim. Acta*; 268, 123.

Lira de F.L.B., Albuquerque de M.S., Pacheco J.G.A., Fonseca T.M., Cavalcanti de E.H.S., Stragevitch L., Pimentel M.F. 2010. Infrared spectroscopy and multivariate calibration to monitor stability quality parameters of biodiesel. *Microchemical Journal*, Volume 96, Issue 1, September 2010, Pages 126-131.

Lucasius, C.B. and Kateman, G. 1991. Genetic algorithms for large-scale optimization in chemometrics: an application. *Trends in Analytical Chemistry* 10:254-261.

Lucasius, C.B., Beckers, M.L.M., and Kateman, G. 1994. Genetic algorithms in wavelength selection: a comparative study. *Analytica Chimica Acta* 286:135-153.

Martens, H., and Naes, T. 1989. *Multivariate calibration*. Wiley: John Wiley & Sons Inc.

Massart, D.L., Vandeginste, B.G.M., Buydens, L.M.C., De Jong, S., Lewi, P.J., Smeyers-Verbeke, J. 1998. *Part B: Handbook of chemometrics and qualimetrics*. Elsevier.

Newport Corporation. 2010. Introduction to FT-IR Spectroscopy. http://www.new port.com/Introduction–to–FT–IRSpectroscopy/405840/1033/content.aspx (accessed May 10, 2011).

Nuopponen, M.H., Birch, G.M., Sykes, R.J., Lee, S.J., Stewart, D. 2006. Estimation of wood density and chemical composition by means of diffuse reflectance mid-infrared Fourier transform (DRIFT-MIR) spectroscopy. *Journal of Agricultural Food Chemistry*. 54:34-40.

OSHA Technical Manual, OTM. 1999. U.S. Department of Labor. Occupational Safety & Health Adminstration. TED 01-00-015.

Öztürk, B. 2003. *Monitoring the esterification reactions of carboxylic acids with alcohols using near infrared spectroscopy and multivariate calibration methods*. M.Sc. Thesis, Izmir Institute of Technology.

Özdemir, D., Mosley, R.M., and Williams, R.R. 1998. Hybrid calibration models an alternative to calibration transfer. *Applied Spectroscopy* 52:599-603(5).

Özdemir, D., Mosley, R.M., and Williams, R.R. 1998. Effect of wavelength drift on single- and multi-instrument calibration using genetic regression. *Applied Spectroscopy* 52: 1203-1209(7).

Özdemir, D. and Williams, R.R. 1999. Multi-instrument calibration with genetic regression in UV-visible spectroscopy. *Applied Spectroscopy* 53:210-217(8).

Paradkar, R.P. and Williams, R.R. 1996. Genetic regression as a calibration technique for solid-phase extraction of dithizone-metal chelates. *Applied Spectroscopy* 50:753-758(6).

Ramadhas A.S., Muraleedharan C., Jayaraj S. 2005. Performance and emission evaluation of a diesel engine fueled with methyl esters of rubber seed oil. *Renew Energy*. 2005: 30: 1789 -800.

Rio del V., Soledad M., Larrechi M., Callao P. 2010. Determination of sulphate in water and biodiesel samples by a sequential injection analysis—Multivariate curve resolution method. *Analytica Chimica Acta*, Volume 676, Issues 1-2, 31 August 2010, Pages 28-33.

Rojas, R., 1996. *Neural Networks: A Systematic Introduction*. Springer, Berlin.

Sheffield Hallam University, SHU. Gas Chromatography. http://teaching.shu.ac.uk/hwb/chemistry/tutorials/chrom/gaschrm.htm (accessed April 5, 2011).

Sherman, C.P. 1997. *Handbook of instrumental techniques for analytical chemistry (F. Settle, ed.)*, Chap. 15. New Jersey: Prentice Hall.

Skoog, D.A., Holler, F.J., Nieman, T.A. 1998. *Principles of instrumental analysis – fifth edition*. Philadelphia: Saunders College Publishing, Harcourt Brace College Publishers.

Sigma-Aldrich. Packed GC Columns. 2010. http://www.sigmaaldrich.com/analytical-chromatography/gas-chromatography/packed-gc-columns.html (accessed April, 2011).

Smith, B.C. 1996. *Fundamentals of Fourier transform infrared spectroscopy*. New York: CRC Press.

Stavarache, C., Vinatoru, M. and Maeda, Y. 2007. Aspects of ultrasonically assisted transesterification of various vegetable oils with methanol. Ultrasonics Sonochemistry 14:380-386.

Oliveira, D., Grishko, V.I., Tran, C.D. 2004. Determination of enantiomeric compositions of pharmaceutical products by near-infrared spectrometry. *Analytical Biochemistry* 325:206-214.

ThermoNicolet. 2009. Nicolet 6700 FT-IR Spectrometer. http://www.thermo scientific.com/ecomm/servlet/productsdetail?productId=11961710&groupType=PRODUCT&searchType=0&storeId=11152 (accessed May 2, 2011).

Tyson, K.S., Bozell, J., Wallace, R., Petersen, E., and Moens, L., 2004. Biomass Oil Analysis: Research Needs and Recommendations, National Renewable Energy Laboratory Technical Report, NREL/TP-510-34796.

Umdu, E.S, 2008. *Methyl ester production from vegetable oils on heteregenous basic catalysts.* M.Sc. Thesis, Izmir Institute of Technology.

Vessman, J., Stefan R. I., Staden van J. F., Danzer K., Lindner W., Burns D. T., Fajgelj A., Müller, H. 2001. *Pure Appl. Chem*. 73, 1381.

Wang, Y., Veltkamp, D. J., and Kowalski, B. R. 1991. Multivariate instrument standardization. *Analytical Chemistry* 63:2750-2756.