ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

APPLICATION OF MULTIVARIABLE CALIBRATION TECHNIQUES TO DETERMINE PHYSICAL-CHEMICAL PROPERTIES AND QUALITY OF GASOLINE PRODUCTS

M.Sc. THESIS

Ümit Ayna

Department of Chemistry

Chemistry Programme

DECEMBER 2015

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

APPLICATION OF THE MULTIVARIABLE CALIBRATION TECHNIQUES TO DETERMINE THE PHYSICAL-CHEMICAL PROPERTIES AND QUALITY OF GASOLINE PRODUCTS

M.Sc. THESIS

Ümit Ayna
(509091061)

Department of Chemistry

Chemistry Programme

Thesis Advisor: Prof. Dr. Mustafa Özcan
Co-Advisor: Prof. Dr. Uğur Akman

DECEMBER 2015

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

BENZİN ÜRÜNLERİNİN KİMYASAL – FİZİKSEL ÖZELLİKLERİNİN VE KALİTESİNİN BELİRLENMESİNDE ÇOK DEĞİŞKENLİ KALİBRASYON TEKNİKLERİNİN UYGULANMASI

YÜKSEK LİSANS TEZİ

Ümit Ayna
(509091061)

Kimya Anabilim Dalı

Kimya Programı

Tez Danışmanı: Prof. Dr. Mustafa Özcan
Eş Danışmanı: Prof. Dr. Uğur Akman

ARALIK 2015

Ümit AYNA, a **M.Sc.** student of ITU **Graduate School of Science Engineering and Technology** student ID 509091061, successfully defended the **thesis** entitled "**APPLICATION OF MULTIVARIABLE CALIBRATION TECHNIQUES TO DETERMINE PHYSICAL-CHEMICAL PROPERTIES AND QUALITY OF GASOLINE PRODUCTS**", which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor : **Prof. Dr. Mustafa ÖZCAN** ....
İstanbul Technical University

Co-advisor : **Prof.Dr. Uğur AKMAN** ....
Boğaziçi University

Jury Members : **Prof.Dr. Süleyman AKMAN** ....
İstanbul Technical University

Prof.Dr. İsmail YILMAZ ....
İstanbul Technical University

Prof.Dr. Abdürrezzak BOZDOĞAN ....
Yıldız Technical University

Date of Submission : 26 November 2015
Date of Defense : 24 December 2015

To my spouse and family,

FOREWORD

This master study is supported by Istanbul Technical University, Institute of Science and Technology and Tupras Izmit Refinery Quality Control Laboratory.

Firstly, I would like to express my gratitude to my thesis supervisor, Prof. Dr. Mustafa Özcan and co-advisor Prof. Dr. Uğur Akman for offering help, continuous encouragement and helpful critics through this research. I thank to them for their patience and time spent.

I would like to thank Tupras Izmit Refinery Quality Control Laboratory personnel for their great support.

I also thank to my wife for her support and to my family for their encouragement through all my education.

December 2015

Ümit AYNA
(Chemist)

TABLE OF CONTENTS

Page

FOREWORD .... **ix**
TABLE OF CONTENTS .... **xi**
ABBREVIATIONS .... **xiii**
LIST OF TABLES .... **xv**
LIST OF FIGURES .... **xvii**
SUMMARY .... **xxi**
ÖZET .... **xxiii**
1. INTRODUCTION **1**
  1.1 Purpose of Thesis 4
  1.2 Literature Review 5
2. PETROLEUM AND PETROLEUM REFINING **15**
  2.1 Petroleum and Crude Oil 15
  2.2 Gasoline 18
    2.2.1 Octane Number 20
    2.2.2 Distillation 22
    2.2.3 Hydrocarbon Composition 24
    2.2.4 Density 27
3. NIR SPECTROSCOPY **29**
  3.1 Electromagnetic Radiation and Energy Levels 29
  3.2 NIR Spectroscopy 34
    3.2.1 Overtones and combinations 37
    3.2.2 Advantages and disadvantages of NIR 40
4. CHEMOMETRICS **43**
  4.1 Calibration 44
    4.1.1 Univariate Calibration 44
      4.1.1.1 Classical Calibration 45
      4.1.1.2 Inverse Calibration 46
    4.1.2 Multivariate Calibration 48
      4.1.2.1 Classical Least Squares (CLS) 49
      4.1.2.2 Multiple Linear Regression 50
      4.1.2.3 Principal Component Analysis (PCA) 51
      4.1.2.4 Principal Component Regression (PCR) 54
      4.1.2.5 Partial Least Squares (PLS) 56
  4.2 Developing Chemometrics based Analytical Methods 59
5. EXPERIMENTATION and INSTRUMENTATION **61**
  5.1 Experimentation 61
  5.2 Instrumentation 62
    5.2.1 NIR 62
    5.2.2 Octane Number – RON and MON 63
    5.2.3 Distillation 65

5.2.4 Hydrocarbon Types 66
5.2.5 Density 67
5.3 Data Analysis 67
6. RESULTS & DISCUSSION 69
6.1 Data pre-processing 69
6.2 Principles and Essential Practices used for Regression Analysis 73
6.3 Evaluation of NIR Spectra 76
6.4 Principal Component Regression – PCR 78
6.4.1 RON – Research Octane Number 79
6.4.2 MON – Motor Octane Number 82
6.4.3 Aromatics 85
6.4.4 Olefins 88
6.4.5 Benzene 91
6.4.6 Density 94
6.4.7 E150 – Evaporated at 150 °C 97
6.5 Partial Least Squares - PLS 100
6.5.1 RON – Research Octane Number 101
6.5.2 MON – Motor Octane Number 104
6.5.3 Aromatics 107
6.5.4 Olefins 110
6.5.5 Benzene 113
6.5.6 Density 116
6.5.7 E150 –Evaporated at 150 °C 119
6.6 Results and Summary 122
7. CONCLUSION AND RECOMMENDATIONS FOR FUTURE RESEARCH 125
REFERENCES 127
APPENDICES 133

ABBREVIATIONS

ANN : Artificial Neural Network
ATR : Attenuated Total Reflectance
bi-PLS : Backward interval PLS
CLS : Classical Least Squares
dyn bi-PLS : Dynamic backward interval PLS
FTIR : Fourier Transform Infrared Spectroscopy
E150 : Evaporated at 150 ^{o}C
EPDK : Energy Market Regulatory Authority
ETBE : Ethyl tert-Butyl Ether
FAME : Fatty Acid Methyl Ester
GA : Genetic Algorithm
GC : Gas Chromatography
GR : Genetic Regression
GCLS : Genetic Classical Least Squares
GILS : Genetic Inverse Least Squares
HPLC : High Pressure Liquid Chromatography
LDA : Linear Discriminant Analysis
LSRN : Light Straight Run Naphtha
LV : Latent Variable
mboe/d : Million barrels of oil equivalent per day
MIR : Middle Infrared Region
MLP : Multilayer Perceptron
MON : Motor Octane Number
MTBE : Methyl-tertiary-butyl Ether
MW : Molecular weight
mw-PLS : Moving window PLS
NIR : Near Infrared Spectroscopy
NMR : Nuclear Magnetic Resonance Spectroscopy
OPEC : Organization of the Petroleum Exporting Countries
p.a. : Per annum
PCA : Principal Component Analysis
PLS : Partial Least Squares
PoLiSh : Smoothed PLS
RON : Research Octane Number
SIMCA : Soft Independent Modelling of Class Analogy
TAN : Total Acid Number
App : Appendix

LIST OF TABLES

Page

Table 1.1: World Supply of Energy [1] 2
Table 1.2: Gasoline requirements and test methods [3]. 3
Table 1.3: Strategies for IR spectroscopic analysis of petroleum (physical characterization and chemical structure) [45]. 7
Table 2.1: Crude oil element composition [5]... 15
Table 2.2: Distillation ranges for fractions obtained from crude oil refining [5]. 16
Table 2.3: Domestic fuel sales [2]... 19
Table 2.4: Global Petroleum Products Demand [1]. 19
Table 2.5: Octane test method conditions [10, 11]. 22
Table 3.1: Electromagnetic spectrum [14]. 30
Table 3.2: NIR bands of some vibrational groups [17].. 38
Table 4.1: Eigen values and cumulative in percentage [36]. 54
Table 4.2: SIMPLS algorithm formulated by Jong [42] 58
Table 4.3: Essential steps for building chemometric prediction model [18]. 60
Table 5.1: Basic statistical data for 45 gasoline sample analysis results 62
Table 5.2: Distillation analysis parameters [7].. 65
Table 5.3: Gas Chromatography parameters [64]. 66
Table 6.1: PCR-RON: Measured, Predicted results and Residuals. 80
Table 6.2: PCR-MON : Measured, Predicted results and Residuals......................... 83
Table 6.3: PCR-Aromatics: Measured, Predicted and Residuals, % v/v. 86
Table 6.4: PCR-Olefins: Measured, Predicted and Residuals, % v/v. 89
Table 6.5: PCR-Benzene: Measured, Predicted and Residuals, % v/v. 92
Table 6.6: PCR-Density: Measured, Predicted and Residuals, kg/l......................... 95
Table 6.7: PCR-E150: Measured, Predicted and Residuals, ^{o}C. 98
Table 6.8: PLS-RON: Measured, Predicted results and Residuals. 102
Table 6.9: PLS-MON: Measured, Predicted results and Residuals. 105
Table 6.10: PLS-Aromatics: Measured, Predicted results and Residuals, % v/v. .. 108
Table 6.11: PLS-Olefins: Measured, Predicted results and Residuals, % v/v. 111
Table 6.12: PLS-Benzene: Measured, Predicted results and Residuals, %v/v. 114
Table 6.13: PLS-Density: Measured, Predicted results and Residuals, kg/l.......... 117
Table 6.14: PLS-E150: Measured, Predicted results and Residuals, ^{o}C. 120
Table 6.15: Summary of values for PCR. 123
Table 6.16: Summary of values for PLS. 124

LIST OF FIGURES

Page

Figure 1.1: Percentage shares of oil demand by sector in 2010 and 2035 [1] 1
Figure 2.1: Fractions of crude oil distillation [Url-1]. 17
Figure 2.2: Refining flow scheme [Url-2]. 18
Figure 2.3: Research Octane Number for hydrocarbon groups [53]........................ 21
Figure 2.4: Correlation of gasoline performance with distillation profile [54]. 23
Figure 2.5: Distillation apparatus [12]. 24
Figure 2.6: Gas Chromatography Scheme [Url-9]. 25
Figure 2.7: A typical Gas Chromatography chromatogram [Url-10]. 26
Figure 2.8: Typical chromatogram of a gasoline sample containing MTBE [13]. ... 26
Figure 3.1: Oscillating electromagnetic waves [15]. 29
Figure 3.2: Molecular Vibration types; Bending, Symmetric and Asymmetric stretching [17]. 32
Figure 3.3: Harmonic and anharmonic models for potential energy of a diatomic molecule (d_e: equilibrium distance) [18]. 33
Figure 3.4: Summary of IR region spectroscopy techniques [16]. 34
Figure 3.5: Types of measurements in NIR Spectroscopy. a) Transmittance, b)Transflectance, c) Diffuse Transflectance, d) Interactance, e)Transmittance through scattering medium [18]................................ 35
Figure 3.6: A design of monochromator scanning instrument [16]. 36
Figure 3.7: NIR spectrum of chloroform [14]. 38
Figure 3.8: NIR spectra, between 5000 – 4000 cm^{-1} of normal paraffin and iso-paraffin hydrocarbons [21]. 39
Figure 3.9: NIR spectra of naphtha, n-hexane, toluene and cyclohexane (offset for clarity) [4]. 40
Figure 4.1: Relation between other disciplines with respect to chemometrics [19]. 43
Figure 4.2: Errors in Classical calibration (a) and Inverse calibration (b) [34]. 46
Figure 4.3: a) Some spectra of multiple samples without any interferences, b) univariate calibration of samples in a). c) Some spectra of multiple samples with interferences, d) univariate calibration of samples in c) [34]. 49
Figure 4.4: Data reduction in PCA [30]. 51
Figure 4.5: A schematic explanation of PCA and data reduction [30]. 53
Figure 4.6: A schematic of Principal Component Regression [37]. 55
Figure 4.7: A schematic presentation of Partial Least Squares (PLS) [34]. 57
Figure 4.8: Process analytical chemistry team approach [61]. 59
Figure 5.1: a) NIR Instrument and b) Flow Cell apparatus from laboratory. 63
Figure 5.2: CFR Engine from laboratory. 64
Figure 5.3: Atmospheric Distillation instrument from laboratory. 65
Figure 5.4: Picture of Gas Chromatography from laboratory. 66
Figure 5.5: Density Meter form laboratory. 67

Figure 6.1: NIR Spectra without baseline correction. 70
Figure 6.2: NIR Spectra with baseline correction. 71
Figure 6.3: Residual vs Cross-Validation Residual Plot. 73
Figure 6.4: Hotelling's T^2 vs Q Residuals – Scores on LV1 vs Scores on LV2 – Measured vs Predicted – Leverage vs Studentized Residual. 74
Figure 6.5: NIR Spectra between 1100 – 1550 nm. 77
Figure 6.6: NIR Spectra between 1800 - 2200 nm. 77
Figure 6.7: Graphical User Interface (GUI) for PCR in PLS Toolbox. 78
Figure 6.8: PCR-RON: **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance. 79
Figure 6.9: PCR-RON: Residuals vs All Sample Set. 80
Figure 6.10: PCR-RON: Measured vs. predicted results for a) all sample set, b) calibration set and c) validation set. 81
Figure 6.11: PCR-MON: **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance. 82
Figure 6.12: PCR-MON : Residuals vs. All Sample Set. 83
Figure 6.13: PCR-MON: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set. 84
Figure 6.14: PCR-Aromatics a) PC vs. RMSEC-RMSECV, b) PC vs. X Cumulative variance. 85
Figure 6.15: PCR-Aromatics: Residuals vs. All Sample Set. 86
Figure 6.16: PCR-Aromatics: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set. 87
Figure 6.17: PCR-Olefins: **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance. 88
Figure 6.18: PCR-Olefins: Residuals vs. All Sample Set. 89
Figure 6.19: PCR-Olefins: Measured vs. predicted results for a) all sample set, b) calibration set and c) validation set. 90
Figure 6.20: PCR-Benzene: **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance. 91
Figure 6.21: PCR-Benzene: Residuals vs. All Sample Set. 92
Figure 6.22: PCR-Benzene: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set. 93
Figure 6.23: PCR-Density: **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance. 94
Figure 6.24: PCR-Density: Residuals vs. All Sample Set. 95
Figure 6.25: PCR-Density: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set. 96
Figure 6.26: PCR-E150: **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance. 97
Figure 6.27: PCR-E150: Residuals vs. All Sample Set. 98
Figure 6.28: PCR-E150: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set. 99
Figure 6.29: PLS Toolbox GUI for PLS method. 100
Figure 6.30: PLS-RON: **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance. 101
Figure 6.31: PLS-RON: Residuals vs All Sample Set. 102
Figure 6.32: PLS-RON: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set. 103

Figure 6.33: PLS-MON **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. Cumulative variance. 104
Figure 6.34: PLS-MON: Residuals vs All Sample Set. 105
Figure 6.35: PLS-MON: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set. 106
Figure 6.36: PLS-Aromatics **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance. 107
Figure 6.37: PLS-Aromatics: Residuals vs All Sample Set. 108
Figure 6.38: PLS-Aromatics: Measured vs. predicted results for a) all sample set, b) calibration set and c) validation set. 109
Figure 6.39: PLS-Olefins: **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance. 110
Figure 6.40: PLS-Olefins: Residuals vs All Sample Set. 111
Figure 6.41: PLS-Olefins: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set. 112
Figure 6.42: PLS-Benzene: **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance. 113
Figure 6.43: PLS-Benzene: Residuals vs All Sample Set. 114
Figure 6.44: PLS-Benzene: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set. 115
Figure 6.45: PLS-Density **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance. 116
Figure 6.46: PLS-Density: Residuals vs All Sample Set. 117
Figure 6.47: PLS-Density: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set. 118
Figure 6.48: PLS-E150 **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance. 119
Figure 6.49: PLS-E150: Residuals vs All Sample Set. 120
Figure 6.50: PLS-E150: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set. 121

APPLICATION OF THE MULTIVARIABLE CALIBRATION TECHNIQUES TO DETERMINE THE PHYSICAL-CHEMICAL PROPERTIES AND QUALITY OF GASOLINE PRODUCTS

SUMMARY

Gasoline is one of the main fuels used in transportation and is produced in petroleum refineries via distillation of crude oil, followed by some other processes. For every batch of gasoline prepared at petroleum refineries, all the parameters given in TS EN 228 specifications are analyzed according to reference method published by ISO 17025 accredited laboratories to determine if the product meets the limits.

Multivariate calibration techniques, a subject in chemometrics, take an important role here. Multivariate calibration method is a technique in which spectroscopic measurement is combined with the reference test methods results to obtain a model for future prediction.

In this thesis work, 46 gasoline samples, obtained from the Quality Control Laboratory of Tupras Izmit Refinery, were studied to establish multivariate calibration models to see if they can be used instead of reference test methods and the prediction error is determined to compare. Octane number (RON and MON), distillation, hydrocarbon types (aromatics, olefins and benzene) and density parameters were the parameters studied, and all gasoline samples were analyzed by using reference test methods for these parameters. Same samples were analyzed by NIR spectroscopy in wavelength range of 800 – 2500 nm. After baseline correction, mean centering and removal of outlier, PCR and PLS multivariate calibration methods used to obtain calibration models.

One sample is removed from data set as outlier. 30 samples were used as a calibration data set and 15 samples were used as a validation data set, which are not used in calibration data set. Predicted results, residuals, root mean squared error of prediction (RMSEP) were calculated, and some useful plots were plotted for both calibration and validation data sets.

According to the results, all the residuals calculated for MON and Aromatics by using PCR and for RON, MON and Aromatics by using PLS are smaller than reproducibility of reference test method. The PLS model results indicated that more than 90% of residuals are even smaller than half of the reproducibility (R/2) for RON, MON, Aromatics and E150. It can be concluded that the PLS model can be safely used as a replacement for the reference test methods for RON, MON, Aromatics and E150 with the condition of remaining in studied range and with the condition of using samples of similar molecular structure. The residuals and RMSEP values were found higher for olefins and density parameters, a detailed study with different parameters needed in order to get better results.

BENZİN ÜRÜNLERİNİN KİMYASAL – FİZİKSEL ÖZELLİKLERİNİN VE KALİTESİNİN BELİRLENMESİNDE ÇOK DEĞİŞKENLİ KALİBRASYON TEKNİKLERİNİN UYGULANMASI

ÖZET

Dünya enerji ihtiyacının % 85 i fosil yakıtlar kullanılarak karşılanmaktadır. Fosil yakıtlar, elektrik üretimi, ulaşım ve ısınma gibi temel ihtiyaçlar için kullanılan en önemli kaynaklardır. British Petroleum (BP) un Haziran 2015 te yayınladığı Dünya Enerji Değerlendirme raporuna göre, global enerji tüketimi 2014 yılında 2013 yılına göre % 0,9 artmıştır. Bu artış son 10 yılın artış oranı olan % 2,0 a göre düşük olmasına rağmen, daha önceki yıllarda olduğu gibi 2014'te de yıllık enerji tüketiminde artış görülmüştür. Benzer durum Türkiye için de geçerlidir. Aynı raporda verilen verilere göre, Türkiye'de 2014 yılı enerji tüketimi bir önceki yıla göre % 2,7 artmıştır. Toplam enerji tüketiminin, yenilenemeyen ve yenilenebilir enerji kaynakları oranları şu şekildedir; petrol % 27,0, doğal gaz % 34,9, kömür % 28,7, hidroelektrik % 7,3 ve yenilenebilir enerji % 2,2. 2013 yılı ile 2014 yılı arasındaki enerji kaynakları tüketim oranlarındaki değişiklik raporda şu şekilde verilmektedir; petrol + %0,6, doğal gaz + %6,3, kömür + %13,6, hidroelektrik - %32,0 ve yenilenebilen enerji + %21,0. Tüketimdeki en büyük artış yenilenebilir enerji alanında olduğu görülse de yenilenebilir enerji tüketiminin toplam tüketimin sadece % 2,2 si olduğu göz önünde bulundurulmalıdır. Fosil yakıtların tüketimindeki artış göze çarpmaktadır.

Yukarıda verilen oranlardan da görüleceği üzere, fosil yakıtlar Türkiye'nin 2014 yılı enerji tüketiminin % 90 ını oluşturmaktadır ve petrol bu miktarın % 27 sini oluşturmaktadır. Miktar olarak bakıldığında, 2014 yılında 33,8 milyon ton petrol (petrol ürünleri) tüketilmiştir.

Benzin, petrolün damıtılması sonucu direk olarak üretilen saf bir ürün değildir. Nafta, isomerat, reformat, gibi bazı ara ürünlerin karışımı sonucu elde edilen bir üründür. Benzin içindeki temel hidrokarbon türleri parafinler, aromatikler ve olefinler olup,karbon sayısı C_4 ten başlayarak C_{12}'ye kadar devam etmektedir. Benzinin tipik kaynama aralığı 30 – 200 ^oC'dir. Benzin, ulaşım amaçlı kullanılan yakıtların başında gelmektedir.

Bahsedilen gereksinimleri karşılamak için benzin standardında bulunan analizler, ISO 17025 akreditasyon belgesine sahip laboratuvarlar tarafından analiz edilip, TS EN 228 standardında verilen limitlere uygunluğu kontrol edilir. Bu analizlerin laboratuvarda yapılması için genel olarak cihaz, gerekli aparatlar, kimyasal ve teknik yeterliliğe sahip personel gerektirmektedir. Ayrıca analizin sürekli olarak yapılması için sarf malzemeler ve cihazların düzenli bakımı için gerekli olan masraflar bu analizlerin maliyetini arttırmaktadır. Ek olarak, analiz süresi ve analiz öncesi yapılması gereken ön hazırlık aşamaları, bazı analizler için oldukça fazla olabilmektedir.

Bu aşamada, çok değişkenli kalibrasyon teknikleri, konvansiyonel analiz yöntemleri yerine geçebilecek modeller oluşturma ve bu modellerle konvansiyonel analiz

yapmadan hızlı spektroskopik analiz verisi kullanarak sonuç elde etme imkanı vermektedir. Çok değişkenli kalibrasyon, kemometride kullanılan yöntemlerden sadece bir tanesidir.

Kemometri, analitik kimyanın alt dalı olup istatistik ve matematiksel modelleme tekniklerinin kullanıldığı disiplinler arası bir bilim dalıdır. Kemometri, çok karmaşık ve fazla veri içeren kimyasal verileri analiz ederek anlamlı kimyasal bilgiyi elde etmek için oldukça önemli bir araçtır. Deneysel tasarım, modelleme, kalibrasyon, resim işleme, kemometrinin en önemli konu başlıkları arasındadır. Bilgisayar, yazılım ve uygulamalı matematik alanlarındaki gelişmelerin artması ile kemometriye olan ilgi ve bu alanda yapılan çalışmaların sayısı oldukça artmaktadır. Gıda, ilaç, kimya sektörü kemometrik uygulamaların oldukça yaygın kullanıldığı alanlardır.

Bu çalışmada, 46 adet benzin numunesi kullanılarak çok değişkenli kalibrasyon yöntemleri geliştirilmiştir. Bu yöntemler ile referans olan konvansiyonel yöntemler karşılaştırılarak sonuçlar değerlendirilmiştir. Referans analizlere alternatif olarak kullanılan yöntem oluşturmak için, çok değişkenli kalibrasyon metotları yardımıyla çeşitli spektroskopik analizlerden ve referans test metotlarının sonuçlarından elde edilen bilgi kullanılarak çeşitli parametreler için modeller oluşturulur.

Benzinin ulaşım amaçlı kullanılabilmesi için Türk Standartları Enstitüsü tarafından belirlenen ve oldukça sıkı limitler içeren, TS EN 228 – Otomotiv Yakıtları – Kurşunsuz Benzin – Özellikleri ve Deney Yöntemleri standardının gerekliliklerini karşılamalıdır. Bu standardın içeriği, yakıtın çeşitli performans özelliklerini karşılamak ve ülkede uygulanan çevresel düzenlemelere uyumunu sağlamak için birçok parametreden oluşmaktadır. Uçuculuk, vuruntu yapmama (antiknock) ve yakıt ekonomisi, yakıt performansını etkileyen önemli ölçütlerdir. Bu ölçütleri kontrol etmek ve düzenlemek için TS EN 228 standardında yoğunluk, distilasyon, oktan sayısı, buhar basıncı parametreleri ile ilgili limit değerler yer almaktadır. Ayrıca, benzinin yanması ile ortaya çıkan karbon monoksit (CO), uçucu organik karbonlar (VOC), azot oksitler (NO_x) ve partiküller gibi kirleticilerin miktarlarını olması gereken seviyelerin altında tutmak için, standartta aromatik, benzen, kurşun içeriği gibi analizlerle ilgili limit değerler de bulunmaktadır.

Benzin numuneleri RON, MON, Aromatik, Olefin, Benzen, Distilasyon ve Yoğunluk analizleri için TS EN 228 standardında verilen test metotlarına uygun olarak analiz edilmiştir. Aynı numuneler, 800 – 2500 nm dalga boyu aralığından NIR spektroskopi ile de analiz edilmiştir. Bu numunelerin 30 adeti kalibrasyon seti, 15 adeti de validasyon seti olarak ayrılmıştır. Burada kalibrasyon ve validasyon numunelerinin birbirinden bağımsız numuneler olması, elde edilen modellerin doğruluğunun kontrolü için oldukça önemlidir. Modelleme çalışmasından önce baseline düzeltmesi, merkezileştirme işlemleri gibi bazı önişlemler yapılmıştır. Kalibrasyon seti numunelerinin referans ve NIR sonuçları kullanılarak, tüm parametreler için ayrı ayrı PCR ve PLS yöntemleri ile kalibrasyon modelleri elde edilmiştir. Bu modeller kullanılarak validasyon seti numuneleri için ilgili parametreler tahmin edilip, referans analiz yöntemleri değerleri ile istatistiksel olarak karşılaştırılmıştır. Ayrıca, kalibrasyon ve validasyon veri seti için, referans yöntem ile analiz edilerek elde edilen ve çok değişkenli model ile tahmin edilen değerlerin grafikleri çizilerek değerlendirilmiştir.

Elde edilen sonuçlara göre, kalibrasyon ve validasyon veri setleri kullanılarak, MON ve Aromatik parametreleri için PCR yöntemi ile elde edilen artıkların hepsi referans test metodunun uyarlık değerinden küçüktür. Yine kalibrasyon ve validasyon veri setleri kullanılarak, RON, MON ve Aromatik parametreleri için PLS yöntemi ile elde edilen artıkların hepsi referans test metodunun uyarlık değerinden küçüktür. Ayrıca, PLS modeli ile tahmin edilen sonuçların % 90 ından fazlası RON, MON, Aromatik ve E150 (distilasyon) parametreleri için uyarlık değerinin yarısından küçüktür. PCR ile elde edilen model MON ve Aromatik parametreleri, PLS ile elde edilen model RON, MON, Aromatik ve E150 parametreleri için kullanılan referans analiz yöntemlerinin yerine kullanılabilir. Olefinler ve Yoğunluk parametreleri için hem PCR hem de PLS modelleri ile elde edilen tahmin sonuçları başarılı olmamıştır. Bu parametreler için ilave çalışmalar yapılması gereklidir.

1. INTRODUCTION

Crude oil is refined and processed in refineries in order to produce petroleum products. Firstly, crude oil is separated into fractions by fractional distillation according to boiling points of hydrocarbon types. Then, the fractions are further processed with many different types of processes to get final petroleum products such as LPG, naphtha, gasoline, kerosene, diesel, fuel oil, bitumen. Petroleum products are playing very critical, important and indispensable role in human life. They are almost everywhere in our daily life, like transportation, power generation, heating, raw material used in petrochemical plants, etc. In Figure 1.1, the distribution of oil use in daily life in 2010 and expected use in 2035 is given. The transportation and industry take the first and second place. Oil is the main component that acts as an energy source with a percentage of 32.2, as given in Table 1.1.

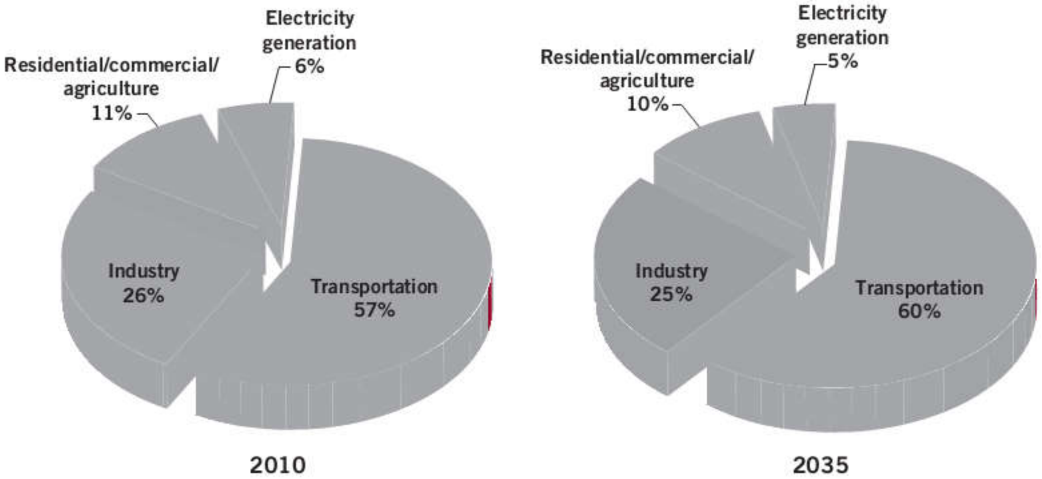


Figure 1.1: Percentage shares of oil demand by sector in 2010 and 2035 [1]

Gasoline is one of the main petroleum products produced in refineries and used as second preferred transportation fuel after diesel, globally and in Turkey.

Major investments in worldwide refineries are mainly focused on producing gasoline and diesel products having ultra-low sulfur concentration, that is below 10 parts per million (ppm) by weight, in order to comply very tightened regulations about exhaust emission limits for sulfur dioxide, SO_2. In addition to sulfur dioxide, carbon dioxide, CO_2 is also very important parameter that must be controlled to avoid air pollution. Oxygenated fuels, produced as a mixture of MTBE (methyl tert-butyl ester), ETBE (ethyl tert-butyl ester), FAME (fatty acid methyl ester) and ethanol with fuels, are primary solutions for reducing carbon dioxide emissions. Reduction of benzene concentration and aromatics concentration in gasoline will follow in following years [1].

Table 1.1: World Supply of Energy [1]

		Levels		Growth		Fuel shares	
		mboe/d		% p.a.		%	
	2010	2020	2035	2010–35	2010	2020	2035
Oil	81.2	89.7	100.2	0.8	32.2	30	26.3
Coal	69.8	84.9	104	1.6	27.7	28.4	27.2
Gas	54.8	69	99.8	2.4	21.7	23.1	26
Nuclear	14.3	16	21.6	1.7	5.7	5.4	5.7
Hydro	5.8	7.4	10.1	2.3	2.3	2.5	2.6
Biomass	24.4	28	35.2	1.5	9.7	9.4	9.2
Other renewables	1.8	3.6	10.7	7.5	0.7	1.2	2.8
Total	251.9	298.6	381.7	1.7	100	100	100

In 2013, gasoline production is 4.307.303 tons and diesel production is 7.636.794 tons in Turkey; in percentages as 19% and 33% for gasoline and diesel, respectively. [2]

In Turkey, the gasoline fuel specification “TS EN 228: Automotive Fuels - Unleaded Petrol - Requirements and Test Methods”, is published by TSE (Turkish Standards Institute) with the latest revision issued in 2013. Table 1.2 indicates the complete requirements for unleaded gasoline.

Table 1.2: Gasoline requirements and test methods [3].

Property	**Test Unit**	**Guarantee**	**Limit**	**Test Method**
Appearance		Clear and Bright		Visual
Corrosion, Copper Strip (3 h at 50 ºC)		No.1	Max	TS 2741 EN ISO 2160
Density at 15 ºC	kg/m^3	720-775		TS 1013 EN ISO 3675 TS EN ISO 12185
Distillation				TS EN ISO 3405
Evaporated at 70 ºC				
Summer grade (a)	vol%	20-48		
Winter grade (b)	vol%	22-50		
Evaporated at 100 ºC				
Summer grade (a)	vol%	46-71		
Winter grade (b)	vol%	46-71		
Evaporated at 150 ºC	vol%	75	Min	
End Point	ºC	210	Max	
Residue	% vol	2	Max	
Gum, Existent (Washed)	mg/100 ml	5	Max	TS EN ISO 6246
Oxidation Stability	minutes	360	Min	TS 2646 EN ISO 7536
Research Octane Number	RON	95	Min	TS EN ISO 5164
Motor Octane Number	MON	85	Min	TS EN ISO 5163
Lead	mg/l	5	Max	TS EN 237
Sulfur	mg/kg	10	Max	TS EN ISO 20846 TS EN ISO 20884
Vapor Pressure (DVPE)				TS EN 13016-1
Summer grade (a)	kPa	45-60		
Winter grade (b)	kPa	60-90		
VLI (Vapor Lock Index)**	Index	1150	Max	
Transition period for				
Summer and Winter grade				
Benzene	vol %	1.0	Max	TS 7088 EN 238 TS EN 12177 TS EN ISO 22854
Olefins	vol %	18.0	Max	TS EN ISO 22854 TS EN 15553
Aromatics	vol %	35.0	Max	TS EN ISO 22854 TS EN 15553

Table 1.2 cont.: Gasoline Requirements and test methods [3].

Oxygen	wt %	2.7	Max	TS 11413 EN 1601
				TS EN 13132
				TS EN ISO 22854
Oxygenates content				TS 11413 EN 1601
				TS EN 13132
				TS EN ISO 22854
-Ethers	vol %	15	Max	
-Methanol	vol %	3	Max	
-Ethanol	vol %	5	Max	
-Iso-propyl alcohol	vol %	10	Max	
-Iso-butyl alcohol	vol %	10	Max	
-Tert-butyl alcohol	vol %	7	Max	
-Other oxygenates	vol %	10	Max	

As seen from above table, there are almost 20 parameters that should be met in order to sale gasoline product in the local market. Similar parameters and associated limits are also valid for EU countries.

In order to do these analyses, there should be a very well designed, established and accredited laboratory with required instruments and well-educated technicians and chemists. The above test methods require long analysis time, expensive instruments to be invested in and high maintenance cost.

1.1 Purpose of Thesis

RON, MON, aromatics content, benzene content, and distillation points are some of the essential components for gasoline-type hydrocarbon products. CFR (Cooperative Fuel Research) test engines, gas chromatography and physical distillation instruments are used to perform RON, MON, aromatics content,, benzene content, and distillation analyses correspondingly in a petroleum-testing laboratory. As a very good and practical alternative to these conventional test methods, “Near-Infrared (NIR) spectroscopy is an excellent analytical method for the identification of petroleum products because it is fast, rugged, and provides highly reproducible results with minimal maintenance” (Choi et al, 1999, p.1021). NIR spectra of a sample consist of spectroscopic information like absorbance, transmittance etc. in a very large frequencies (or wave-lengths). Only the NIR spectroscopy data is not enough to get quantitative results for above-mentioned parameters. Some multivariate statistical calibration techniques are needed to extract the valuable information from very large and complex spectral data.

The main goal of this thesis work is to predict the essential physicochemical and quality-related components of the gasoline products, which are RON, MON, aromatics content, olefins content, benzene content, density and E150 via the NIR spectroscopy analysis combined with multivariate calibration techniques by utilizing the data obtained in the Quality Control Laboratory of TÜPRAŞ Izmit Refinery. The predicted results will also be compared with the results obtained from the conventional methods, which are CFR engine analysis, multidimensional gas chromatography, atmospheric distillation, automated densitometer. A good prediction ability of the model proposed in this work will be very useful to refiners to adjust and control the process for the production of final products like gasoline with ease and confidence. The valuable outputs of this work will be saving time and money, taking immediate actions in case of any deviation from production set points and safe process operations.

1.2 Literature Review

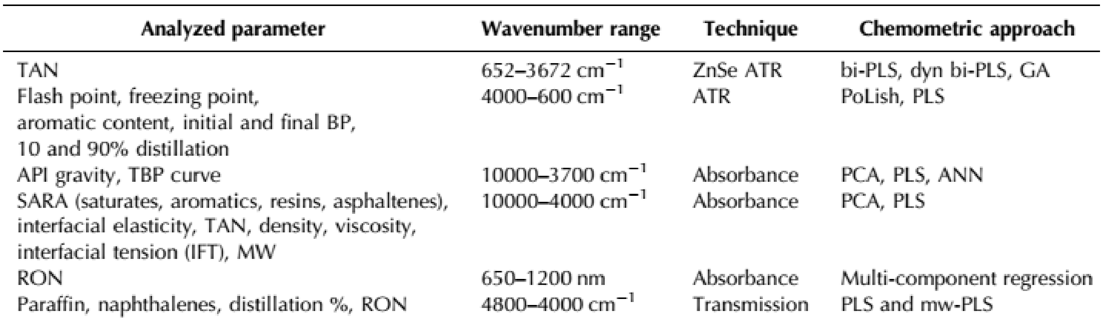
Chemometrics (the science of getting chemical information from a chemical system by using mathematical based approach), combined with NIR spectroscopy has been widely used, starting at mid-1970's by many academicians and industry practitioners. As technology extends, meaning better computer systems and spectroscopy instruments with very high resolution and fast analysis time, it became inevitable for people to use these techniques in their studies and work environment. Although combination and overtone bands in near-infrared region, that is 800 – 2500 nm, are very broad and difficult to interpret, they give very valuable structural information which is not available in MIR, middle infrared region. Main chemometric methods are experimental design, classification, pattern recognition, clustering and multivariate calibration. Multivariate calibration is a good way of identifying the relationship between the measured property and the concentration of component in the sample of interest by following the Beer's Law. Multivariate calibration is used when simple univariate calibration is not enough for complicated systems, like having several components in a sample that absorbs at a given wavelength. Partial least squares (PLS) is one of the best multivariate calibration techniques used widely. There are other techniques like multiple linear regression (MLR), principal component regression (PCR), genetic algorithm (GA). Having these calibration techniques, coupled with NIR spectroscopy as a fast and reliable spectroscopic analysis technique, revealed many application areas for chemometrics, like agriculture, food, pharmaceutical, and refinery/petrochemical. The following review is based on the literature studies related to chemometric techniques, mostly including multivariate calibration methods used in different applications, and is especially focused on the hydrocarbon-type samples produced and analyzed in the refinery/petrochemical areas.

Determination of gasoline quality parameters by applying various types of multivariate calibration techniques has been studied by many researchers in order to control the process in a production facility continuously and to find an alternative testing method to replace the traditional test method. Measurement of octane numbers (RON and MON) in a laboratory requires very large investment cost initially to purchase the instrument, constant and expensive maintenance, long time of analysis (around 45 minutes) and very experienced and trained personnel to perform the analyses. These difficulties led people to look for alternatives and thus, there have been many studies attempting to correlate the octane number to chemical structure of gasoline.

The role of fossil fuels in daily life and estimation of quality parameters about these fossil fuels are emphasized, along with a review of the chemometric methods for the determination of characteristics of petroleum-based products, by Khanmohammadi et al. (2012).

The main target is given as estimation of physico-chemical parameters of petroleum products especially gasoline and diesel by establishing a relation between the property interested and spectroscopic response (e.g. absorption, reflection and transmission) at spectral region [45]. API, octane number, TBP (true boiling point) curve, benzene, aromatics and olefins are the main parameters considered for gasoline and diesel. In addition, various chemometric approaches like PCA (principal component analysis), PLS, PCR, ANN are explained in [45]. A summary given in Table 1.3 shows the spectroscopic techniques, spectral regions and chemometric methods to predict fuel quality parameters determined with reference methods.

Table 1.3: Strategies for IR spectroscopic analysis of petroleum (physical characterization and chemical structure) [45].



Analyzed parameter	Wavenumber range	Technique	Chemometric approach
TAN	652–3672 cm^{-1}	ZnSe ATR	bi-PLS, dyn bi-PLS, GA
Flash point, freezing point, aromatic content, initial and final BP, 10 and 90% distillation	4000–600 cm^{-1}	ATR	PoLish, PLS
API gravity, TBP curve	10000–3700 cm^{-1}	Absorbance	PCA, PLS, ANN
SARA (saturates, aromatics, resins, asphaltenes), interfacial elasticity, TAN, density, viscosity, interfacial tension (IFT), MW	10000–4000 cm^{-1}	Absorbance	PCA, PLS
RON	650–1200 nm	Absorbance	Multi-component regression
Paraffin, naphthalenes, distillation %, RON	4800–4000 cm^{-1}	Transmission	PLS and mw-PLS

(TAN: total acid number, MW: molecular weight, ATR: attenuated total reflectance, bi-PLS: backward interval PLS, dyn bi-PLS: dynamic backward interval PLS, PoLiSh: smoothed PLS, mw-PLS: moving window PLS)

In addition to estimation of quality parameters, structural analysis and classification, especially to detect adulteration, is also discussed and spectroscopic techniques and chemometric methods are reviewed [45]. Detection of adulteration and identification of external materials are very difficult tasks. IR spectroscopy is one of adulteration-detecting instrumental analysis and it is very powerful to separate adulterated and non-adulterated fuel both qualitatively and quantitatively. MIR (middle infrared region) and NIR are very good at to determine adulteration of diesel and biodiesel with vegetable oil when combined with chemometric methods PCR, PLS, ANN and SIMCA (soft independent modelling of class analogy). Gasoline is another petrol-based fuel that is very open to adulteration by using cheaper chemical like kerosene, diesel oil, petrochemical thinner and turpentine.

H^+ NMR (Nuclear Magnetic Resonance) spectra at 100 MHz was studied by Meyer et al. (1975) to correlate the octane numbers with chemical structure by using linear regression analysis [23]. The main interested regions are due to methylene protons and methyl protons in NMR spectra. Study revealed that the relationship between octane numbers and chemical structure was not linear. Another similar study was established by using individual integrated areas of aliphatic, allylic and aromatic regions of NMR spectra to establish a linear relation to get a good correlation with octane numbers of gasoline samples. Dolbear (1972) found the following linear equations;

$$RON = AromaticH + 1.55 \times OlefinicH + 76.5 \qquad \textbf{(1.1)}$$

$$MON = 0.55 \times AromaticH + 0.27 \times OlefinicH + 71.6 \qquad \textbf{(1.2)}$$

Chemical information taken from NIR analysis combined with mathematical methods, i.e. multivariate calibration, provides very good and reliable results. NIR spectroscopic region has very good structural information for gasoline samples since the absorption bands in this region are the overtones and combinations of C-H stretching vibrations of the hydrocarbon molecules. The absorptivity of C-H stretching of methyl, methylene, olefinic and aromatic groups have different absorptivity in this region than other components in gasoline. Kelly et al. (1989) studied 65 gasoline samples with a range of RON from 91.7 to 98.4 and MON from 82.0 to 87.4, each sample was analyzed according to ASTM reference test methods. Then, SW-NIR (Short Wave NIR) spectra were recorded between 660 – 1215 nm wavelengths with a 2.00 cm path length. The reason to choose SW-NIR is that the range was just enough because of overtones of symmetric and antisymmetric C-H stretching vibrations. MLR and Partial Least Squares (PLS) techniques were used to correlate the SW-NIR spectra and ASTM reference test methods results. MLR technique lead to R^2 values of 0.979 for RON and 0.957 for MON with a linear regression equation calculated at 3 different wavelengths chosen as 932 nm, 1164 nm, 896 nm for RON and 930 nm, 1012 nm, 940 nm for MON, respectively. For the same validation sample set, 4 variable PLS regression gave R^2 values of 0.949 for RON and 0.993 for MON. It was concluded that although the weak absorptions by various C-H bonds overlapping in this region, MLR gave good correlation, but PLS was better. Both MLR and PLS results were satisfying and showed smaller variations then ASTM reference methods [25].

Another study with gasoline samples were done by Bohacs et al. (1998), which consists of 350 gasoline samples at 3 different research octane number (RON) grade as 91, 95 and 98, including summer and winter grades. The gasoline sample set was different than Kelly et al. (1989) in the number of samples and variety of seasonal and RON requirement grades. 12 different chemical and physical properties of gasoline including RON, MON, benzene, methyl-tertier-butyl-ether (MTBE), Sulphur content, distillation points, Reid vapor pressure (RVP) and density at 15 ^{o}C were correlated by using PLS regression as multivariate calibration method. NIR spectra of gasoline samples were recorded from 900 to 1700 nm spectral range with 10 mm quartz cuvette and baseline correction was adopted. After recording spectra, there is a data pre-processing step which is different from that of the previous reference, Kelly et al. (1989) the transmittance values were converted to absorbance values, first and second derivative of the absorbance values were calculated in order to remove shifts. After regression analysis, the authors found that the NIR methods developed to predict RON, MON, benzene and MTBE were very successful and they could be used as substitute for the reference methods. The R^2 and SEP (standard error of prediction) values are as following for these four quality parameters; 0.975 and 0.34 for RON, 0.972 and 0.30 for MON, 0.970 and 0.13 %v/v for benzene, 0.999 and 0.2 %v/v for MTBE. Prediction models developed for other gasoline properties showed poor correlation and gave higher standard error than the reference methods [26]. Although both studies, Bohacs et al. (1998) and Kelly et al. (1989), have similar regression coefficients and good predictions for RON and MON, the latter has a more robust prediction model since samples were collected at different RON grades and seasonal variety was also considered. Since the specification of gasoline has some difference in summer and winter grade, meaning that the volatility, which is vapor-pressure property, is limited to max. 60 kPa in summer season, starting from late March to October. In winter season, vapor pressure should be min. 60 kPa and max. 90 kPa (further details are in Table 1.3). This change in volatility also changes the molecular structure of gasoline and causes changes in the NIR spectra collected. In addition to this, variations in RON grades also affect molecular structure of gasoline. Having these variations in calibration sample set makes the prediction model more robust in real-life applications. It is very important to have a valid prediction method that gives correct results according to seasonal and process changes.

PLS is very powerful tool to be used in chemometric studies and there are many PLS algorithms studied so far. Felicio at al. (2005) presented a work to compare different PLS algorithms based on MIR (4000 cm^{-1} – 600 cm^{-1}) and NIR (9400 cm^{-1} – 4500 cm^{-1}) spectra. RMSEP (root mean square error of prediction) and %95 confidence intervals were used for comparison of PLS algorithms for diesel sample flash point and gasoline benzene and RON quality parameters. 249 gasoline samples and 128 gas oil samples were analyzed by MIR and NIR instruments and after eliminating some regions that could lead erroneous results in MIR and NIR spectra and pre-processing, PLS algorithms were applied where randomly chosen %80 of data used as calibration and %20 of data used as validation. PLS (single PLS), MB-PLS (multi block PLS) and S-PLS (serial PLS) are three PLS algorithms compared and RMSEP and confidence intervals were given for PLS-MIR, PLS-NIR, MB [46]. Following results were obtained; RMSEP for flash point is 2.95 ^{o}C obtained by S-PLS algorithm applied to MIR spectra, RMSEP for benzene is 0.0641 % vol. obtained by PLS algorithm applied to MIR spectra and RMSEP for RON is 0.52 obtained by PLS algorithm applied to NIR spectra. The confidence intervals are 1.83 – 5.16 oC, 0.0459 – 0.0847 % vol. and 0.42 – 0.63 correspondingly [46]. Because of this study, spectroscopic technique, chemometric modelling method and sample & parameters to be predicted are three important points that should be determined and combined according to prediction needed or targeted.

Besides having very common multivariate calibration techniques like MLR and PLS, there is another technique called Genetic Regression (GR), which is a calibration technique that optimizes linear regression models using a genetic algorithm. GR is an implementation of Genetic Algorithm (GA) selects an optimum linear combination of wavelengths and simple mathematical operators to build a linear combination model using the simple least squares method [27]. GA consists of several steps, such as the initialization of gene population, evaluation of the population, selection of parent genes for breeding and mating, crossover and mutation, and replacing parents with their offspring. Ozdemir (2005) studied a data set obtained from a web source that consist of 60 gasoline samples with known octane numbers collected using diffuse reflectance NIR analysis in the range of 900 to 1700 nm. 60 samples were divided into three sets, one for calibration, one for prediction and one for validation purposes. He applied three different genetic multivariate calibration methods, genetic regression (GR), genetic classical least squares (GCLS), and genetic inverse least squares (GILS). The regression coefficients for GR, GCLS and GILS are 0.9931, 0.9538, and 0.9962 respectively [27]. As a result of this study it is very obvious that GA improves the prediction power of CLS and ILS multivariate calibration techniques.

As stated at the beginning, spectroscopy measurements combined with multivariate calibration has various applications for hydrocarbon products. There are many studies for naphtha, kerosene, gas oil, diesel, crude oil and residue hydrocarbon products in literature. Ultraviolet absorption spectra of 114 gasoil and diesel fuels were collected in the range of 200 – 400 nm with a 1 cm flow-through cell (Wentzell et al. 1999). Supercritical fluid chromatography with flame ionization detection was used to quantify saturates %, monoaromatics %, diaromatics %, and poly aromatics %. All spectra data first were analyzed for outliers and then mean centered. CLS, FS-MLR

(Forward Selection), SS-MLR (Stepwise Selection), PCR, and PLS multivariate calibration techniques were applied. Except the CLS technique to predict polyaromatics property, all calibration techniques were very successful to evaluate four properties in interest. A detailed table of findings are given in Wentzell et al. As a result, the UV spectroscopic measurements used with multivariate calibration methods are good enough to predict saturates and aromatics content of diesel type of hydrocarbon products successfully [28].

Another interesting study to predict asphaltenes in crude oil by using ATR-IR spectroscopy together with ANN was carried out (Colaicco and Farrera, 2008). Asphaltenes are asphalt like substances found in crude oil and bitumen products, which have high asphaltene concentration, are mostly used in paving materials on road and waterproof coatings [Url-5]. Determination of asphaltene in crude oil is very difficult, time-consuming laboratory test method. The test method, that is IP-143, contain many wet chemistry steps like reflux, precipitation, filtering extraction and weighing [47]. Qualified laboratory personnel analyzed 19 Venezuelan crude oil samples with reference test method (IP-143) and ATR-IR spectroscopy (spectral range from 10000 to 650 cm^{-1} with an accumulation of 128 scan and resolution of 2 cm^{-1}), the these data used for modeling by three layer neural network configuration. The reason to use ANN as modeling tool is that Wild et al. (1998) and Aske et al. (2001) tried to predict asphaltene content by using NIR and IR data with PLS modeling but the error found was higher than standard laboratory test method precision [49,50]. The study revealed very good results and plot of ATR-IR-ANN predicted asphaltene concentration for 19 samples (16 of 19 are calibration and 3 of 19 are validation samples) versus reference test methods results has a correlation, R^2 of 0.996 and standard error of calibration (SEC) is 0.37 wt % [48]. The model established is very successful for asphaltene concentrations lower than 10 wt%. The highest residue value is -0.7 wt% for a reference value of 17.8 wt% whereas predicted value is 17.1 wt%. Standard test method, IP-143, has a reproducibility, R, value of 3.56 wt% for 17.8 wt% asphaltene result. The residue is 5 times smaller than test method precision value. Considering low number of total samples studied, only 4 samples have asphaltene concentration higher than 10 wt%, better prediction values most probably would be obtained when more samples included in study with high asphaltene concentration.

Besides predicting quality parameters of hydrocarbon type products, chemometric techniques are widely used for classification purposes as reviewed before in this section (Khanmohammadi et al, 2012). Balabin and Safieva (2007) studied for classification of 382 gasoline samples by source (refinery and process) and type (regarding to octane number). They divided all gasoline samples into three sets as A, B, C. Set A was used for classification by source (refinery), and it has three classes called as Refinery 1, Refinery 2 and Refinery 3. Set B was used for classification by source (process), and it has six classes called as straight-run, reformate, catalysate, isomerizate, hydrocracking gasoline and mixture. Set C was used for classification by type, and it has three classes called normal, regular and premium. Linear discriminant analysis (LDA), Multilayer perceptron (MLP) and Soft independent modelling of class analogy (SIMCA) are three tools used for classification by combining with NIR spectra of gasoline samples. Prediction efficiency was determined by calculating error that is the ratio of number of wrongly classified samples to total number of samples in the data set. Predicted classification results according to source by refinery has errors of 13%, 14% and 8% for LDA, SIMCA and MLP correspondingly. Similar errors calculated according to source by process are 35%, 30% and 18% for LDA, SIMCA and MLP correspondingly. These results show that it is difficult to classify gasoline according to source by process. Lastly, the prediction errors for classification according to types are 12%, 10% and 9% for LDA, SIMCA and MLP correspondingly. Because of this study, it is concluded that MLP is more very powerful tool than LDA and SIMCA for classification of gasoline samples in relation to source (refinery and process) and type (octane number).

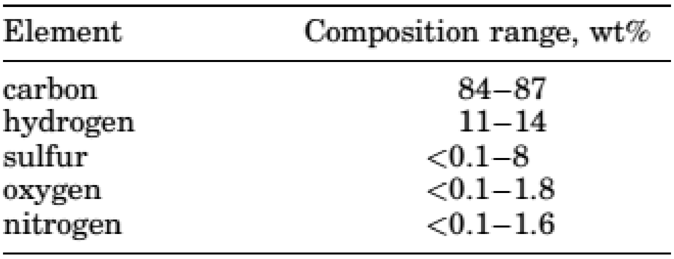
In light of the above-mentioned studies and many others in the literature, predicting hydrocarbon product properties and classification of these hydrocarbons using spectroscopic measurements, mainly the NIR, combined with multivariate calibration techniques are of interest to many researchers and provide successful prediction models. There are many variations used in all these studies such as type of spectroscopic measurement (NIR, NMR, UV), wavelength range chosen, type of samples, pre-processing techniques (first derivative, second derivative, mean centering), multivariate calibration methods (MLR, CLS, PLS, ILS, GA, ANN). Depending on the sample and property to be predicted, above possible options are chosen and optimized. Including sample production and property variations to calibration sample sets provides more robust prediction models.

2. PETROLEUM AND PETROLEUM REFINING

2.1 Petroleum and Crude Oil

Petroleum is naturally occurring mixture containing of carbon and hydrogen as main elements. In addition to carbon and hydrogen, nitrogen, sulfur, oxygen and some smaller amounts of vanadium, nickel elements are also available in petroleum. Petroleum can include three phases: gaseous (natural gas), liquid (crude oil), and solid or semisolid (bitumen, asphalt, tars, and pitches) [5]. Crude oils are a highly complex combination of hydrocarbons, heterocyclic compounds of nitrogen, oxygen and sulfur, organometallic compound, inorganic sediment, and water (Giles and Mills, 2010). Common crude oils have elemental composition range given in Table 2.1.

Table 2.1: Crude oil element composition [5].



Element	Composition range, wt%
carbon	84–87
hydrogen	11–14
sulfur	<0.1–8
oxygen	<0.1–1.8
nitrogen	<0.1–1.6

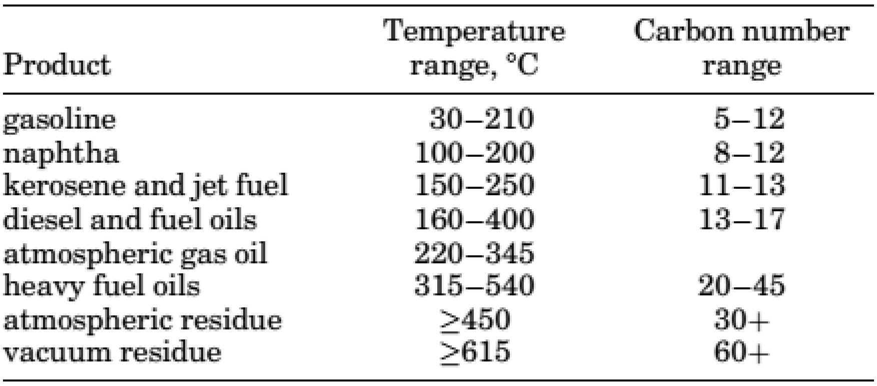
Many of these compounds are identified but there are some compounds, which are not identified yet [5].

Crude oil has different properties like odor and color, usually depending on its origin. The main properties to classify the crude oil are API gravity and sulfur content. API gravity is special function of relative density (specific gravity) 60/60°F, as given in equation 2.1 [6].

$$^{o}API = \left(\frac{141.5}{specific\ gravity\ 60/60^{o}F} \right) - 131.5 \quad \textbf{(2.1)}$$

Crude oils are called as light and heavy according to the density property. Light crude oils are low in density and have light hydrocarbons, thus they have paraffinic hydrocarbons. On the other hand, heavy crude oils have high density, high viscous asphalt like molecules. Distillation, pour point, viscosity, and element content are also important to process crude oil in oil refineries. Crude oil is used as raw material in petroleum refining industry. In petroleum refining, the fractions in the crude oil are separated via distillation according to their boiling point. As a fundamental principle, longer hydrocarbon chains boil at higher temperature. This provides the separation of lighter hydrocarbons from the heavier ones. By using distillation as a separation technique, many products are produced by distillation of crude oil in refining process. LPG, naphtha, gasoline, kerosene, diesel oil, lubrication oil, fuel oil, residue and bitumen are main fractions found after crude oil distillation. A general scheme of distillation column is given in Figure 2.1 and related distillation ranges for these fractions are given in Table 2.2.

Table 2.2: Distillation ranges for fractions obtained from crude oil refining [5].



Product	Temperature range, °C	Carbon number range
gasoline	30–210	5–12
naphtha	100–200	8–12
kerosene and jet fuel	150–250	11–13
diesel and fuel oils	160–400	13–17
atmospheric gas oil	220–345	
heavy fuel oils	315–540	20–45
atmospheric residue	≥450	30+
vacuum residue	≥615	60+

These fractions are intermediate products and it is not possible to use them as final product. There are many processes in which these fractions are further processed and become ready for use. LPG treating, hydrotreating, isomerization, reforming, sweetening, hydrocracking, and FCC are main methods to further process crude oil fractions. A modern refinery scheme consisting of many processes is available in Figure 2.2. As seen, gasoline production is done by mixing various intermediate products coming from different processes like isomerization, reforming, gasoline treater, hydrocracking, fluid catalytic cracking etc.

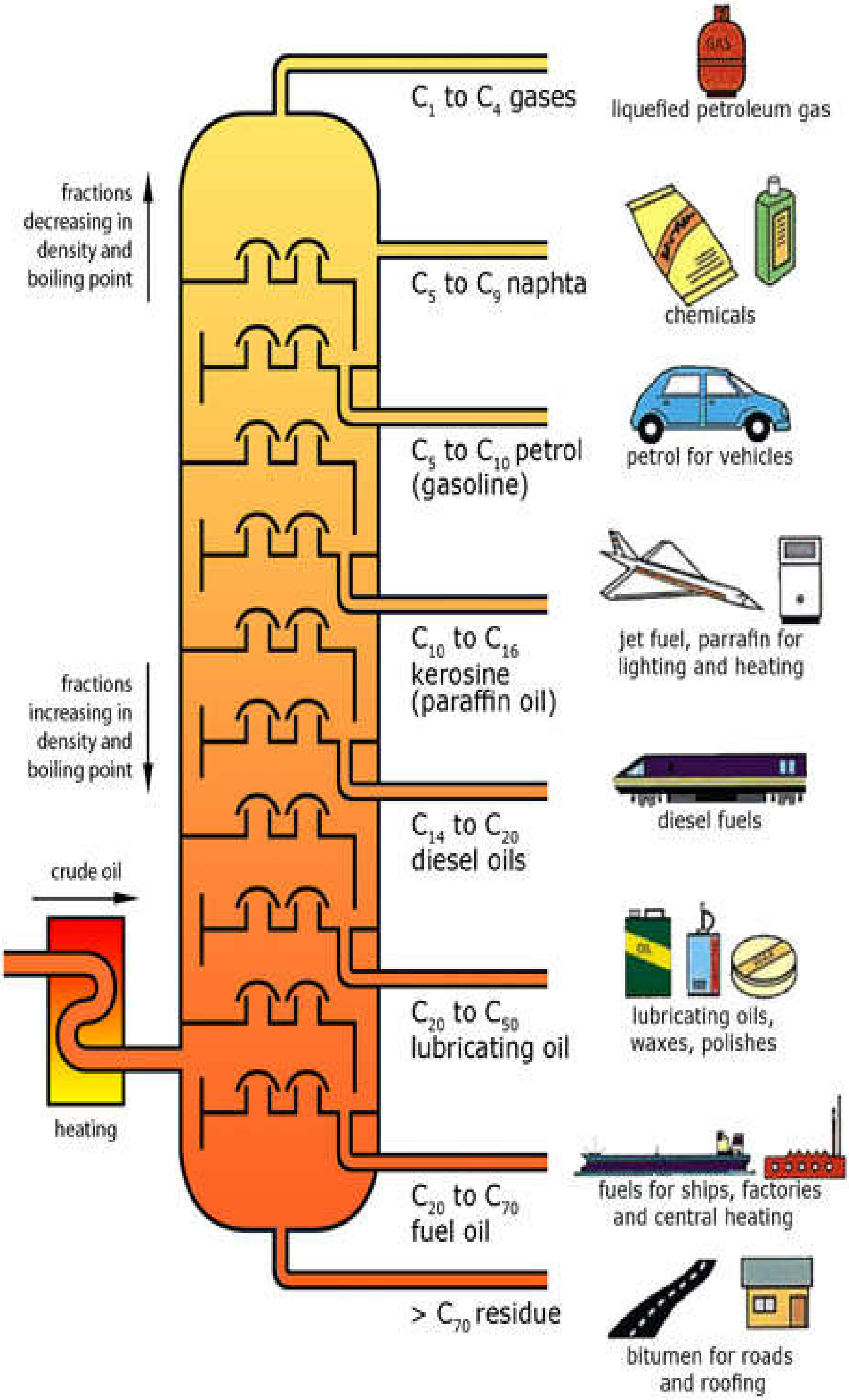

C_1 to C_4 gases
liquefied petroleum gas
fractions decreasing in density and boiling point
C_5 to C_9 naphta
chemicals
C_5 to C_{10} petrol (gasoline)
petrol for vehicles
jet fuel, parrafin for lighting and heating
C_{10} to C_{16} kerosine (paraffin oil)
fractions increasing in density and boiling point
diesel fuels
C_{14} to C_{20} diesel oils
crude oil
C_{20} to C_{50} lubricating oil
lubricating oils, waxes, polishes
heating
fuels for ships, factories and central heating
C_{20} to C_{70} fuel oil
> C_{70} residue
bitumen for roads and roofing

Figure 2.1: Fractions of crude oil distillation [Url-1].

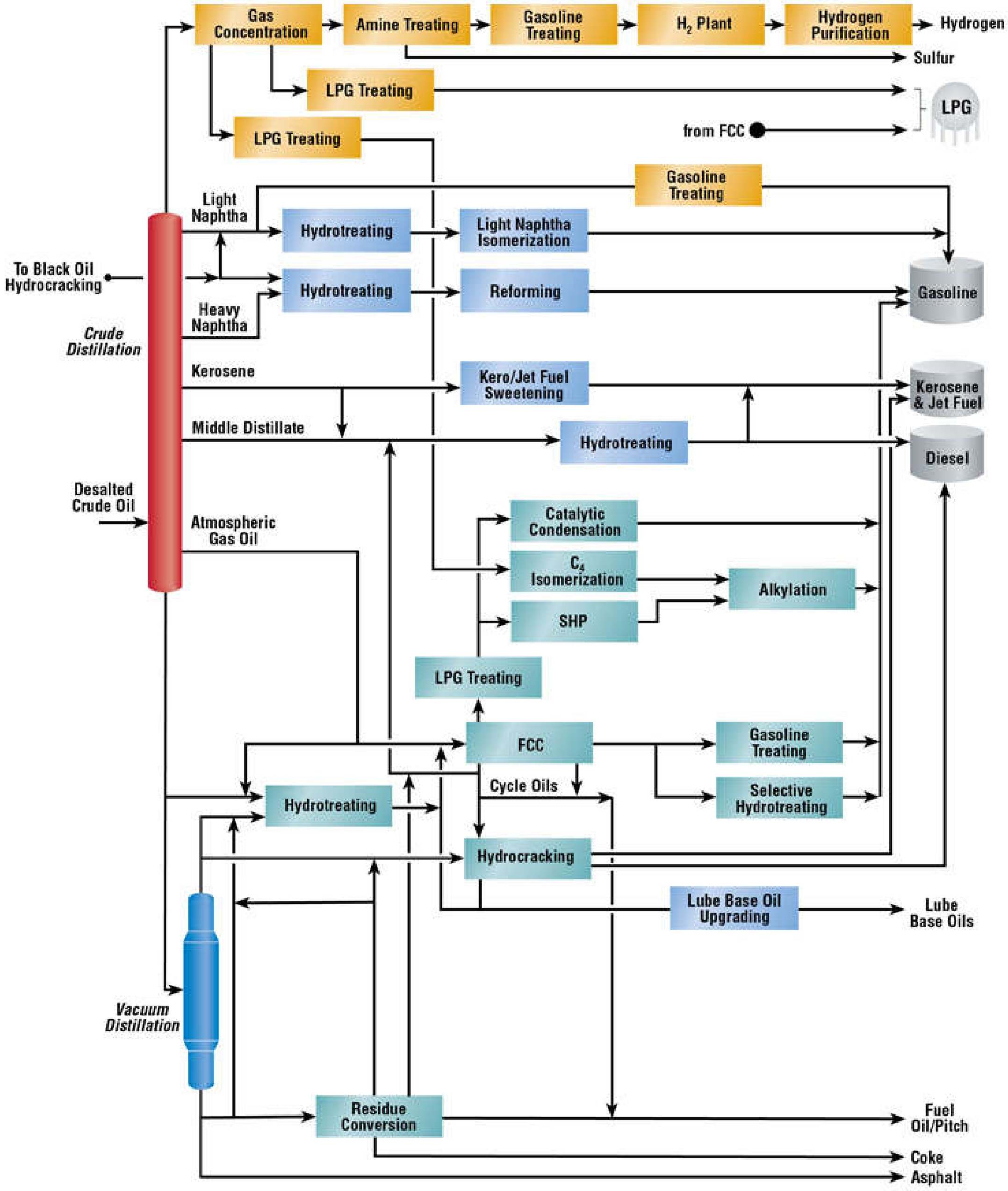
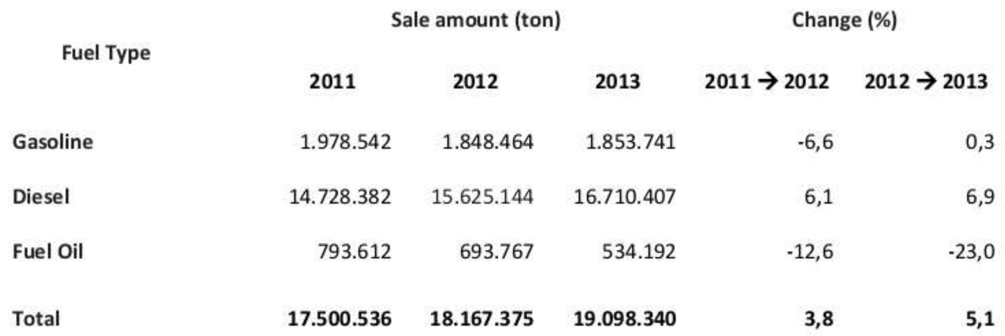


Figure 2.2: Refining flow scheme [Url-2].

2.2 Gasoline

Gasoline is one of the main petroleum products produced in oil refineries and mainly used for transportation purposes in internal combustion and spark ignition engines. Gasoline is not a pure substance, it is a homogenous mixture of various hydrocarbon intermediate products like naphtha, isomerate, reformate, MTBE, ethanol etc. Typical distillation range is in the rage of 30 – 200 °C, containing paraffins, aromatics and olefins as major hydrocarbon types with an overall carbon number range of C_4 – C_{12}. As seen in Table 2.3 below, total gasoline sales in 2013 by retailers is 1,853,741 tons, according to Turkey Energy Market Regulatory Authority 2013 report [2]. It seems that gasoline consumption is less than diesel consumption, mainly because of tax policy and high fuel prices.

Table 2.3: Domestic fuel sales [2].



Fuel Type	Sale amount (ton)			Change (%)	
	2011	2012	2013	2011 → 2012	2012 → 2013
Gasoline	1.978.542	1.848.464	1.853.741	-6,6	0,3
Diesel	14.728.382	15.625.144	16.710.407	6,1	6,9
Fuel Oil	793.612	693.767	534.192	-12,6	-23,0
Total	**17.500.536**	**18.167.375**	**19.098.340**	**3,8**	**5,1**

According to global petroleum products, demand given in World Oil Outlook 2013 report in Table 2.4, there will be 1 million barrels per day (mb/d), approximately 118,343 tons, gasoline demand increase for every 5 years starting from 2015 up to 2035 with a total demand of 27.5 mb/d.

Table 2.4: Global Petroleum Products Demand [1].

	2012	2015	2020	2025	2030	2035
Light products (mb/d)						
Ethane/LPG	9.7	10	10.5	10.9	11.2	11.5
Naphtha	5.9	6.2	6.8	7.3	7.9	8.5
Gasoline	22.7	23.3	24.4	25.5	26.5	27.5
Middle distillates (mb/d)						
Jet/Kerosene	6.5	6.7	7.1	7.4	7.7	8.1
Diesel/Gasoil	25.8	27.3	30	32.2	34.1	36
Heavy products (mb/d)						
Residual fuel*	8.2	7.8	7.1	6.6	6.3	6
Other**	10	10.2	10.5	10.7	10.8	10.9
Total (mb/d)	88.9	91.6	96.3	100.7	104.6	108.5
* Includes refinery fuel oil ** Includes bitumen, waxes, still gas, coke, sulphur etc.						

As seen from Table 2.3 and 2.4, gasoline is currently second main transportation fuel with an increasing demand for future.

In order to use gasoline in transportation vehicles, the limits for components given in TS EN 228 - specification for automotive gasoline (Table 1.2) must be fulfilled to meet fuel performance factors such as volatility, antiknock, and fuel economy. Another very important factor is environmental policies and regulations and they take place in specification for automotive gasoline. Main vehicle emissions produced by gasoline powered transportation cars are Carbon Monoxide (CO), Hydrocarbons (HC), Oxides of Nitrogen (NO_x) and Particulates (PM_{10} - particles have diameter less than 10 micron). These emissions are directly contribute to the air pollution. Hydrocarbons are main source for Volatile Organic Compounds (VOCs). Interaction of NO_x, organic gases and sunlight produce Ozone (O_3) and it is one of main air pollutants. Sulfur in the fuel contributes as Sulfur Dioxide (SO_2) after combustion in engine. CO is another main pollutant comes from vehicles after burning the fossil-based fuel. The limits for air pollutants and more details are given in Ambient Air Quality Evaluation and Management Directive [55]. In addition, The Clean Air Act Amendments of 1990 classified following pollutants related to gasoline vehicles; benzene, formaldehyde, acetaldehyde, Polycyclic organic matter (POM), and 1,3-butadiene [54]. Benzene is found in gasoline (max. 1% vol.) and other pollutants are formed during combustion. All above pollutants have effects on human diseases about respiratory tract, cancer, heart attacks etc.

The concentration of components given in specification for gasoline have direct effects on emissions and air pollutants, thus quality control of gasoline plays an important and critical role in human life.

It would be meaningful to give detailed information about essential properties of gasoline, which are subject to this thesis study.

2.2.1 Octane Number

Octane is the most well known component of gasoline. It is a measure of combustion characteristics of gasoline. Gasoline octane number is measured by depending on its knocking tendency. To have a better understanding about knocking, the following information is given: “The fuel-air mixture undergoes chemical reactions that may cause it to auto-ignite and detonate the entire remaining mixture. Instead of being pushed down smoothly on the power stroke, the piston is given a hard instantaneous rap to which it cannot respond because of the large mechanical inertia present in the crankshaft and other pistons. This rapid energy release causes pressure fluctuations in the cylinder which result in a loud metallic noise commonly called knock” [5].

There are two laboratory test methods (TS EN ISO 5164 and TS EN ISO 5163) to measure octane numbers called Research Octane Number (RON) and Motor Octane Number (MON). RON is measured under low engine speed and MON is measured under high engine speed. RON is always higher than MON and the difference between these two gives information about the sensitivity of gasoline to changes in operating conditions [54]. RON is more common than MON and it is usually RON when octane number of gasoline is mentioned.

One should note that knocking tendency is very dependent on chemical structure of the fuel and it decreases in the following order: alkanes > branched chain alkanes > cycloalkanes > alkenes > aromatics, thus aromatics have higher octane number compared to other hydrocarbon types [52]. In Figure 2.3, RON variation is given with respect to different hydrocarbon types. There are also some additives used as octane improver, like oxygenates and patented chemicals.

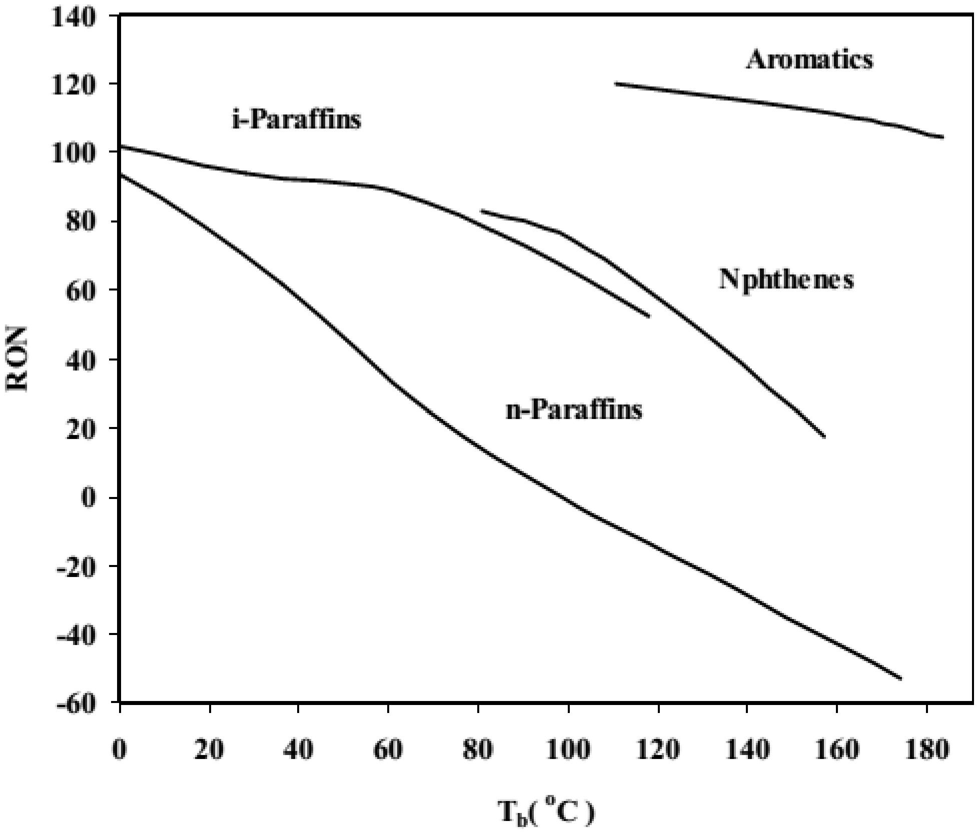


Figure 2.3: Research Octane Number for hydrocarbon groups [53].

Both octane numbers (RON and MON) are measured by using a knock engine in the laboratory with different test conditions. The standard knock engine is a single cylinder cooperative fuels research (CFR) engine. To have similar driving conditions on the road, there are to different test method conditions, given in Table 2.5, to determine RON and MON. It is possible to run the engine with sample fuel and the reference fuel. According to reference fuel knocking tendency, it is possible to find the octane number of sample fuel. There are two main reference fuels, iso-octane (2, 2, 4-trimethyl pentane) and n-heptane, having octane number of 100 and 0 respectively. Any volumetric mixture of these two reference fuels gives us the reference fuel needed for intermediate octane numbers between 0 and 100. Comparison of reference fuel and sample fuel knocking tendency is key to measure octane number.

Table 2.5: Octane test method conditions [10, 11].

Condition	Research octane no. (ASTM D 2699)	Motor octane no. (ASTM D 2700)
engine speed, RPM	600	900
inlet air temperature, °C	51.7	38
mixture temperature, °C		149
spark advance, °BTDC[a]	13	14–26

[a] Spark advance for Motor method is a function of compression ratio.
BTDC = before top dead center.

Detailed information regarding to octane number analysis can be easily found in TS EN ISO 5164 (or ASTM D 2699) and TS EN ISO 5163 (or ASTM D 2700) test methods for RON and MON [8-11].

2.2.2 Distillation

Since gasoline is a mixture of many intermediate petroleum products, it boils over a range of temperatures. Instead of having a single boiling point, gasoline has a distillation curve starting from initial boiling point (IBP), which is the first drop recovered after boiling starts, up to final boiling point (FBP), which is the temperature when there is not more recovered liquid. Distillation range is very important for gasoline performance such as cold start, hot start, fuel economy, power, acceleration and exhaust emissions. The relation between gasoline performance indicators and distillation curve can easily be seen in Figure 2.4.

100 ml of sample is distilled at a rate between 4 – 5 ml/min by applying heat. After heating and start of boiling, vapors of sample are condensed in condenser kept at a temperature around 10 °C and collected in receiver. Sample collected in receiver is measured volumetrically and temperature is recorded at the same time.

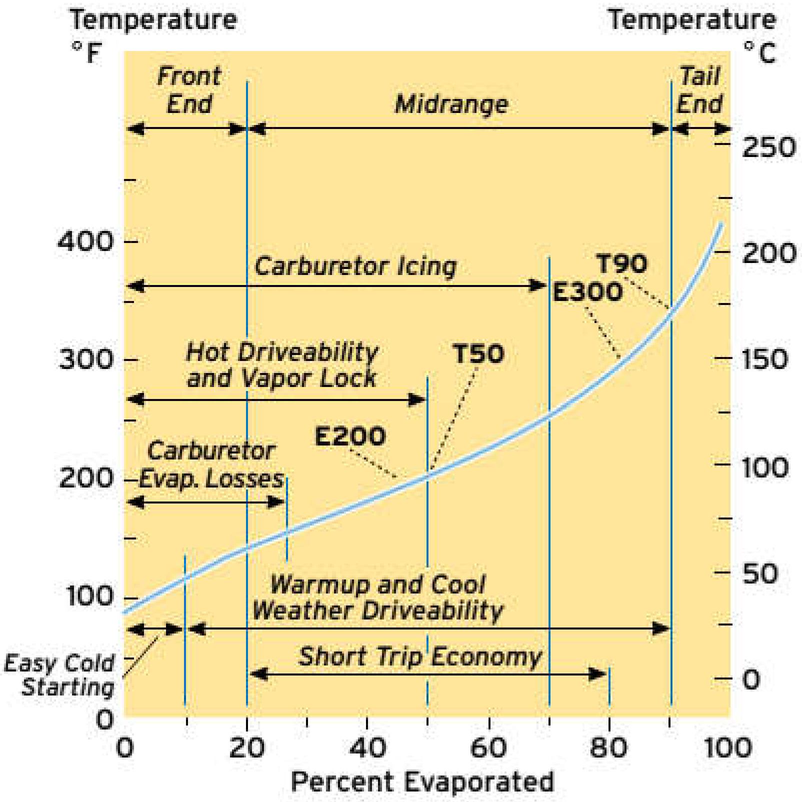
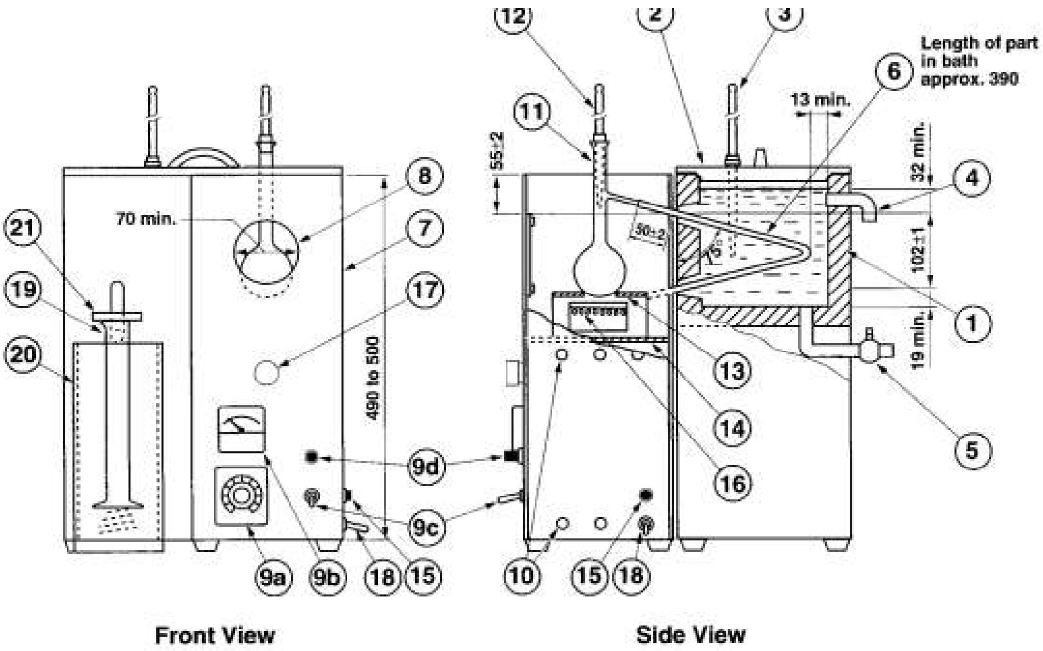


Figure 2.4: Correlation of gasoline performance with distillation profile [54].

Temperature data is recorded at each volume collected in receiver after condensation. Also IBP and FBP are recorded. By plotting temperature versus percent sample evaporated and collected in receiver, a distillation curve is established. In order to avoid any sample loss during distillation, gasoline sample temperature should be below 18 °C and receiver and condenser temperatures should be at around 15 °C. A typical example of distillation apparatus is given in Figure 2.5 [12].

Detailed information for distillation test method is in TS EN ISO 3405 [7]. In gasoline specification given in Table 1.3, distillation analysis points are given as evaporated at 70 °C (E70), evaporated at 100 °C (E100) and evaporated at 150 °C (E150). These points are for evaporated volume of sample at the specific temperatures.



1–Condenser bath
2–Bath cover
3–Bath temperature sensor
4–Bath overflow
5–Bath drain
6–Condenser tube
7–Shield
8–Viewing window
9a–Voltage regulator
9b–Voltmeter or ammeter
9c–Power switch
9d–Power light indicator
10–Vent
11–Distillation flask
12–Temperature sensor
13–Flask support board
14–Flask support platform
15–Ground connection
16–Electric heater
17–Knob for adjusting level of support platform
18–Power source cord
19–Receiver cylinder
20–Receiver cooling bath
21–Receiver cover

Figure 2.5: Distillation apparatus [12].

2.2.3 Hydrocarbon Composition

Gasoline contains of many types of hydrocarbons starting from C_5 to C_{12}. Parafins, naphtenes, aromatics, olefins are essential hydrocarbon groups in gasoline. When gasoline specification (Table 1.2) is examined, there are some components related to hydrocarbon types. In order to determine concentration of hydrocarbon types (paraffins, aromatics, olefins, benzene) and oxygenates in automotive-motor gasoline, multidimensional gas chromatography method, that is TS EN ISO 22854, is the primary method used [13].

Gas chromatography (GC) is a very well known and most widely used laboratory technique for the separation and analysis of volatile compounds with a history of more than 60 years. Separation method is used to identify the components in a sample, and then the concentrations of these components are calculated by using reference standards for the calibration of GC. A gas chromatography system consists of following parts, carrier gas source helps component to move in the column, sample inlet to vaporize sample, column to have separation, detector to measure output and computer to collect **data [55]. A very common scheme is given in Figure 2.6.**

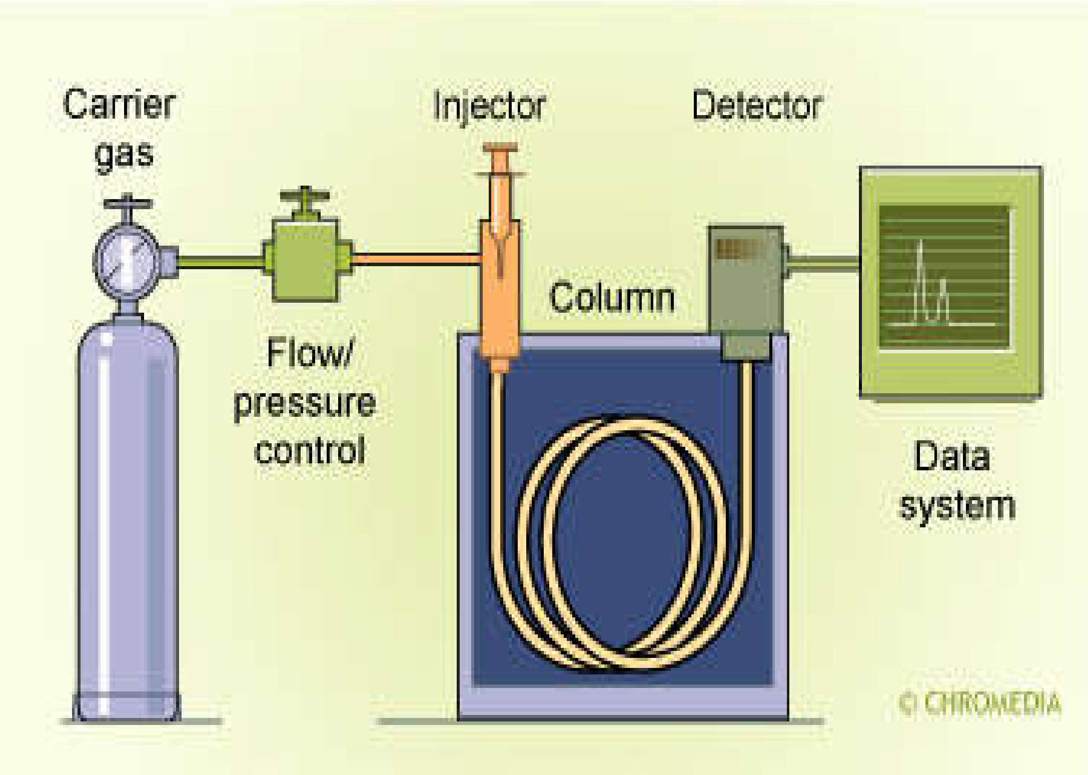


Figure 2.6: Gas Chromatography Scheme [Url-9].

A small amount of sample is injected to the column, where separation in time is achieved by retaining components inside the column with the help of special materials coated to internal surface of column, and then separated components are detected at the detector. The detector output is converted to a chromatogram by using computer and software. A typical chromatogram is given in Figure 2.7. Individual components have unique retention time, the interval between injection of sample and detection of component at detector, used to identification and the area under the peak is used for determination of concentration.

Gasoline sample is separated into hydrocarbon groups (also according to carbon numbers) by gas chromatographic analysis using column-coupling and column switching procedures. There are columns, traps and valves inside gas chromatography system with a flame ionization detector. A typical chromatogram of gasoline with MTBE analysis according to TS EN ISO 22854, is given in Figure 2.8 [13].

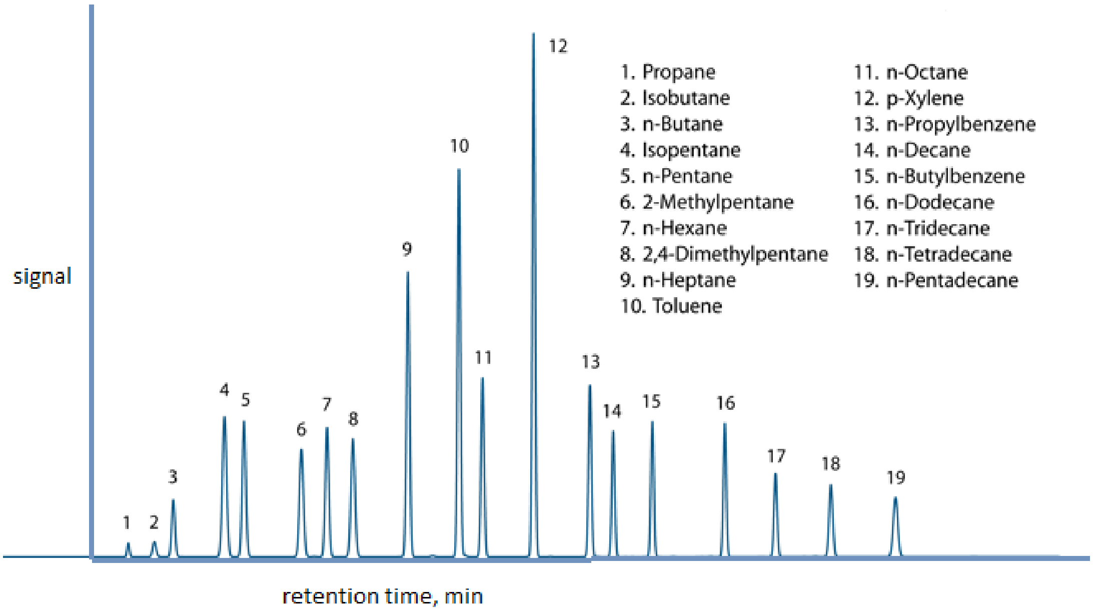


Figure 2.7: A typical Gas Chromatography chromatogram [Url-10].

It should be noted that in this chromatogram the hydrocarbons are grouped and separated by carbon number. Combination of traps specific to hydrocarbon group types, switching valves, boiling point column results in a non-typical chromatogram as given above Figure 2.8.

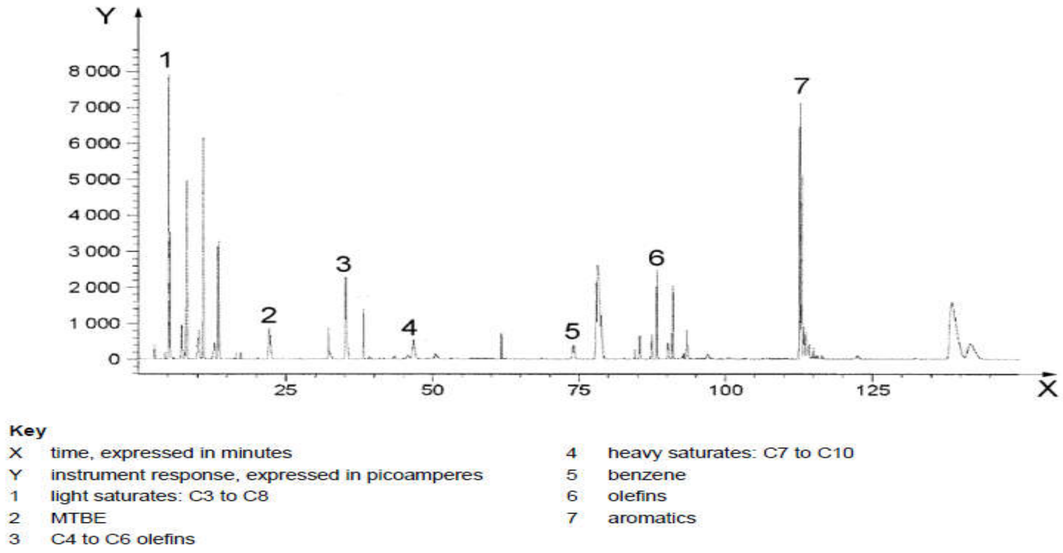


Figure 2.8: Typical chromatogram of a gasoline sample containing MTBE [13].

2.2.4 Density

Density is defined as mass per unit volume at a specified temperature; it is usually given in kg/l or kg/m^3 for petroleum products [57]. Sample temperature is set to density measurement temperature if possible; otherwise, 20 ^oC above the pour point of sample might be used. An important point in this analysis is that the temperature of sample and the measurement instrument should be stable. Sample should be clean, free of particles and air bubbles. A small volume of liquid is injected into an oscillating tube and the change in oscillation frequency caused by the change in mass of tube is correlated to calibration [58]. The result is provided as an average of three consecutive measurements and it is recorded with ±1kg/m^3 and temperature value. Density is determined according to given equation 2.2 [58].

$$d = d_w + K_1 \times (T_s^2 - T_w^2) \tag{2.2}$$

where d is density at test temperature, kg/l or kg/m^3, K_1 is instrument constant for density, T_w is observed period of oscillation for cell containing water (used for calibration), T_s is observed period of oscillation for cell containing sample and d_w is density of water at test temperature.

3. NIR SPECTROSCOPY

NIR is a form of molecular spectroscopy and the near-infrared region of electromagnetic spectrum is from 700 nm to 2500 nm. NIR energy is firstly found by William Herschel in 19th century [Url-11]. Different from mid-IR spectroscopy, NIR has molecular overtone and combination of vibrations, which are forbidden transitions according to quantum mechanics selection rules. Broad bands found in NIR spectra are complex and difficult to assign a molecular structure. Researchers overcame this difficulty by using chemometric models combining with spectroscopy data.

3.1 Electromagnetic Radiation and Energy Levels

Electromagnetic radiation in terms of classical theory is the flow of energy at the speed of light through free space or through a material medium in the form of electric and magnetic field. Electromagnetic radiation consists of electromagnetic waves which can be defined as oscillating waves of electric and magnetic fields (Figure 3.1).

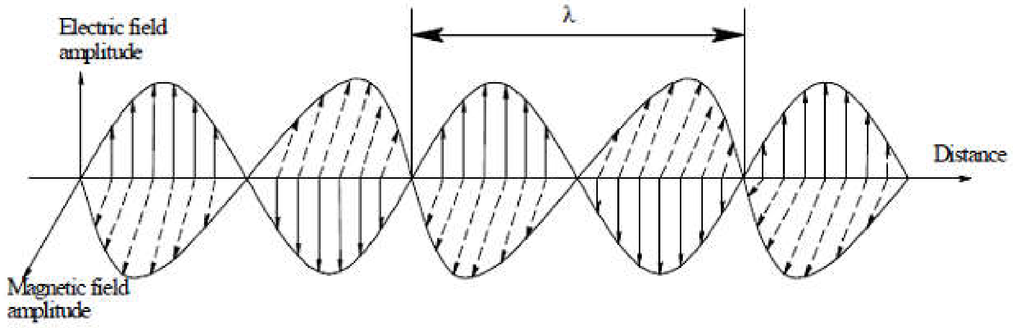
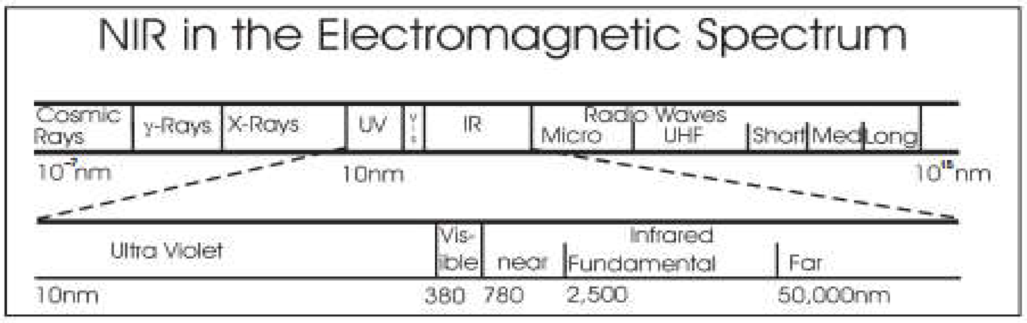


Figure 3.1: Oscillating electromagnetic waves [15].

Oscillations in different wavelength or frequency form the electromagnetic spectrum and, according to wavelength, there are some regions of energy levels; radio waves, microwaves, infrared radiation, visible light, ultraviolet radiation, X-rays and gamma rays as given in Table 3.1. The oscillations of the two fields are perpendicular to each other and perpendicular to the direction of energy and wave propagation. Relation between frequency and wavelength is given as Equation 3.1, where c is speed of light, ν is frequency and λ is wavelength [15].

$$\lambda = \frac{c}{\nu} \tag{3.1}$$

Table 3.1: Electromagnetic spectrum [14].



According to modern quantum theory, electromagnetic radiation is the flow of photons through space. Photons are called as wave-like particles and they can be treated as the "carriers" and "transferers" of energy [15]. Energy of photon is given by Equation (3.2), also called Bohr equation where h is Planck constant ($h=6.626 \times 10^{-34}$ Js) and ν is frequency.

$$E = h\nu \tag{3.2}$$

Matter is anything that has mass and takes up space. It is made up of atoms and atoms are made up of protons (positively charged), electrons (negatively charged) and neutrons (no charge). Protons and neutrons are located in the center of atom, that is nucleus, but electrons are around the nucleus, that is called as orbital. Orbitals are particular energy levels around the nucleus and they have particular energies with individual distance from the nuclei.

Since molecular vibrations are the basis for complete IR region spectroscopy, it would be useful to understand energy of levels (orbitals) by looking harmonic and anharmonic oscillator models in view of classical mechanical and quantum mechanical models.

A diatomic molecule can be treated as two spherical masses (m1 and m2) attached to one another with a spring given force constant (k).

According to Hook's Law, the energy (E) of this system is given by Equation (3.3);

$$E = \frac{h}{2\pi}\sqrt{\frac{k}{\mu}} \qquad \textbf{(3.3)}$$

where μ is reduced mass;

$$\mu = \frac{m1m2}{m1 + m2} \qquad \textbf{(3.3a)}$$

The potential energy is given by Equation (3.4);

$$V = \frac{1}{2}kq^2 \qquad \textbf{(3.4)}$$

where q is displacement between r (inter-nuclear distance during vibration) and r_e (inter-nuclear distance at equilibrium).

$$q = r - r_e \qquad \textbf{(3.4a)}$$

Quantum mechanically, it is well known that electrons can stay in specific energy levels, not in between two energy levels, that also means that electrons stay in specific orbitals. Equation (3.5) gives the energy of these levels, where E_n is molecule vibrational energy, n is (0, 1, 2, 3 ...), h is Plank's constant, k is force constant and μ is reduced mass.

$$E_n = \left(n + \frac{1}{2}\right)h\nu \qquad \textbf{(3.5)}$$

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \qquad \textbf{(3.5a)}$$

Regarding to absorption frequency, it has another form in terms of wavenumber in Equation (3.6) where c is the speed of light.

$$\bar{\nu}, cm^{-1} = \frac{1}{2\pi c}\sqrt{\frac{k}{m}} \tag{3.6}$$

In harmonic oscillation, energy levels given in above equations are in equal distance and transitions are only allowed between neighboring energy levels, Δn=±1. According to the Boltzmann distribution, most molecules at room temperature populate the ground level n=0, and consequently the allowed, so-called fundamental, transitions between n=0 and n=1 dominate the vibrational absorption spectrum . These fundamental vibrations is mainly Mid-IR (MIR) spectra region. MIR region, from 8500 to 12.500 nm, is very characteristic for molecules and called as fingerprint region. Thus, mid-IR is mostly used for qualitative analysis. A change in the dipole moment of the molecule needed to get absorption of infrared radiation for any type of vibration [16].

According to Stuart (2004), "A molecule can only absorb radiation when the incoming infrared radiation is of the same frequency as one of the fundamental modes of vibration of the molecule" (p. 8). Vibrations are in form of stretching (change in bond length) and bending (change in bond angle). There are symmetrical and asymmetric stretching, in-phase and out-of-phase respectively. These vibrations are given in Figure 3.2 [17].

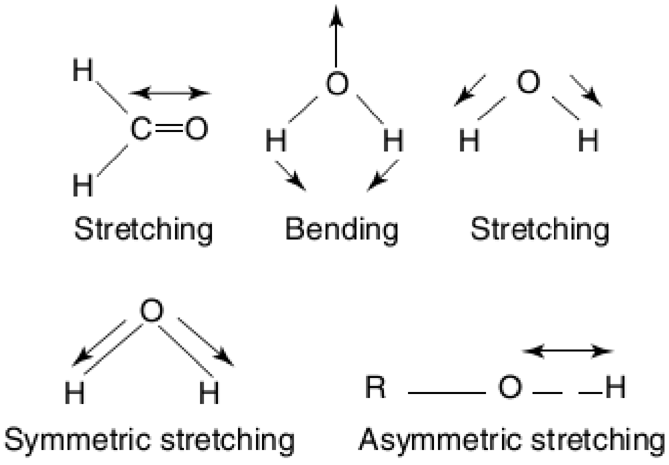


Figure 3.2: Molecular Vibration types; Bending, Symmetric and Asymmetric stretching [17].

Harmonic oscillation phenomena is very good to understand fundamental vibrations, but it is not enough to explain energy level transition like Δn=±2 or more where overtone bands exist. According to harmonic oscillation model, combination of vibrations is not possible because of restriction rules. However, it is well known that there are overtones and combinations in NIR region.

Anharmonic oscillator model comes into place where "The model considers some non-ideal behaviors of the oscillator which account for repulsion between electronic clouds when the atomic nuclei approach (the potential energy rises fasten than in the harmonic model) and a variable behavior of the bond force when the atoms move apart from one another" (Pasquini, 2003, p. 202). For anharmonic oscillator model, the energy levels do not have equal distance and the transitions to more than on level is possible, that is Δn=±2, ±3 etc. These are called as first, second etc. overtones. In addition to overtones, any combination of vibrational transitions form combinations. NIR region contains absorption bands for overtone and combination absorptions. Then the vibrational energy levels are given in Equation (3.7).

$$E_n = \left(n + \frac{1}{2}\right) h\nu_0 - \left(n + \frac{1}{2}\right)^2 h\nu_0 \chi \qquad \textbf{(3.7)}$$

where χ is anharmonicity constant of vibration. Energy levels for harmonic and anharmonic models are shown in Figure 3.3.

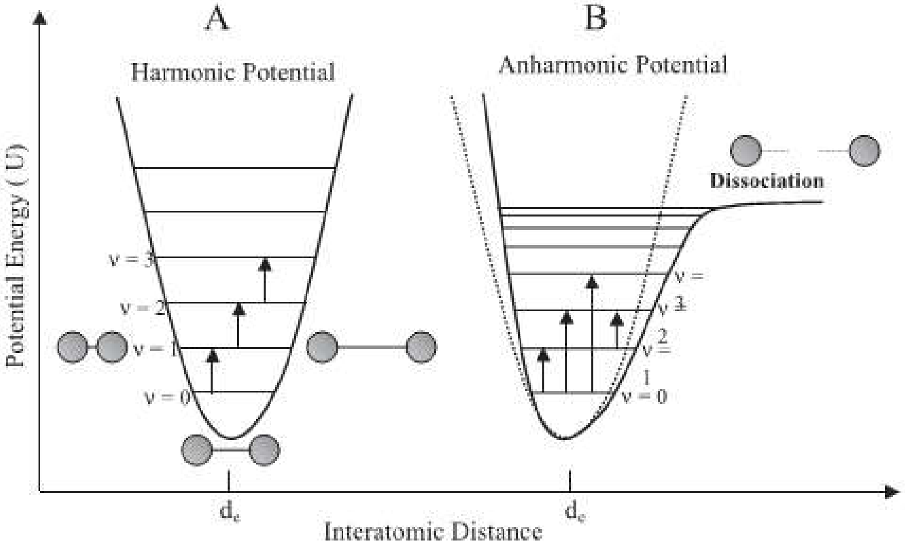


Figure 3.3: Harmonic and anharmonic models for potential energy of a diatomic molecule (d_e: equilibrium distance) [18].

Figure 3.4, given below, is also a good summary of above discussions.

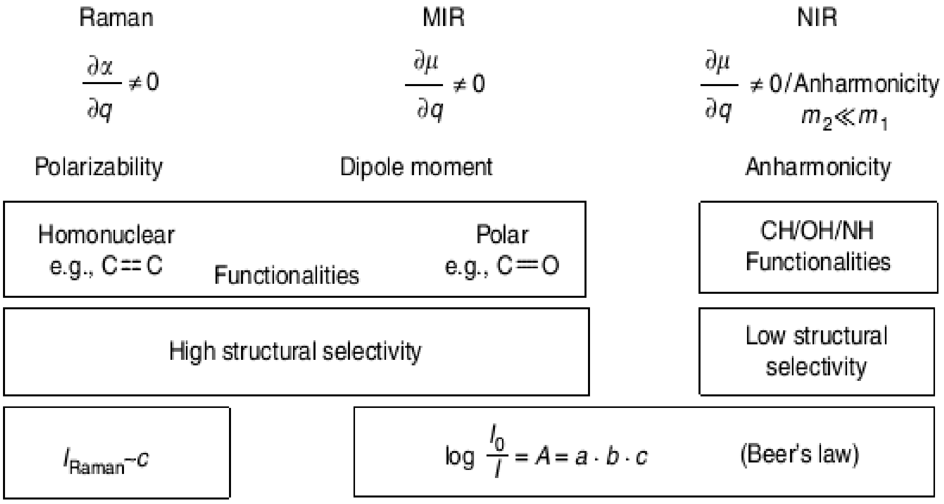

Raman
MIR
NIR
$\frac{\partial \alpha}{\partial q} \neq 0$
$\frac{\partial \mu}{\partial q} \neq 0$
$\frac{\partial \mu}{\partial q} \neq 0$/Anharmonicity $m_2 \ll m_1$
Polarizability
Dipole moment
Anharmonicity
Homonuclear e.g., C==C
Functionalities
Polar e.g., C=O
CH/OH/NH Functionalities
High structural selectivity
Low structural selectivity
I_{Raman}~c
$\log \frac{I_0}{I} = A = a \cdot b \cdot c$
(Beer's law)

Figure 3.4: Summary of IR region spectroscopy techniques [16].

3.2 NIR Spectroscopy

Near infrared radiation is first discovered by William Herschel who was a successful musician and astronomer. He realized that there is heating effect beyond the visible light range. He called this as radiant heat and the thermometrical spectrum. However, he mistakenly defined this radiant heat different than light [16]. In 1835, Ampere contributed to Herchel's studies that NIR had same properties with UV light except the wavelength.

NIR is a form of molecular spectroscopy and it utilizes the spectral range from 780 to 2500 nm (12,500 and 4,000 cm^{-1}) and provides complex structural information related to the vibration behavior of combinations of bonds in molecules. Similar to UV, visible and mid-IR spectroscopy, NIR also follows Beer's Law. Same philosophy, that is the frequency of light matches the frequency of a suitable molecular vibration, then the light can be absorbed, applies for NIR spectroscopy. The major bands in the NIR region are second or third harmonics of fundamental O−H, C−H, and N−H stretching vibrations found in the mid-IR region, detailed in section 3.2.1. Monochromatic light produced by an NIR instrument interact with material as reflection, refraction, absorption, diffraction, and transmission. A representative scheme is given in Figure 3.5 for NIR radiation for various different type of samples [18].

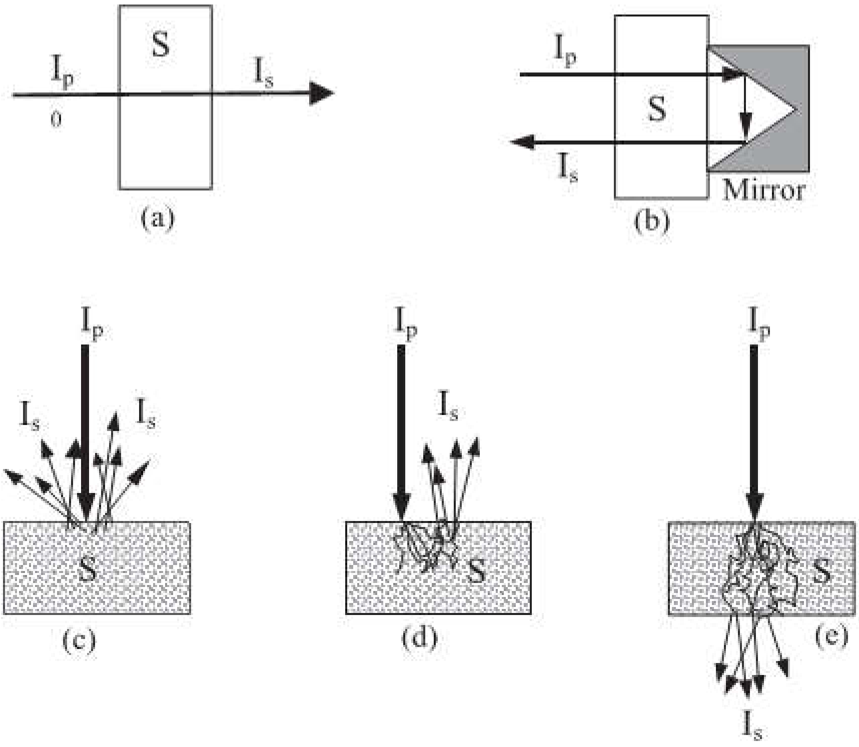


Figure 3.5: Types of measurements in NIR Spectroscopy. a) Transmittance, b)Transflectance, c) Diffuse Transflectance, d) Interactance, e)Transmittance through scattering medium [18].

These are different techniques to get the spectra from sample. They are different for the different positioning of the light source and of the measurement sensor around the sample. Transmittance (absorption) and reflectance, given in equation 3.8 and 3.9, are most important and used techniques among others [18].

Transmittance is based on the measuring the light that goes through whole sample, and at this time some of it absorbed by sample. This type of analysis can give information about internal structure of sample. Since the light from source should pass through the sample, there is need a high intensity light source and high sensitive measuring detector. Transmittance is a measure of light intensity in terms of wavelength remaining after the absorption of light by sample. The sample takes place in between the light source and the detector. Either transmittance (T) or absorbance (A) is determined because of measurement [17,18].

$$T = \frac{I}{I_0} \tag{3.8}$$

where I = intensity of transmitted radiation and I_0 = intensity of incident radiation.

$$A = -\log_{10} T \quad A = \log_{10}\left(\frac{1}{T}\right) \quad A = \log_{10}\left(\frac{I_0}{I}\right) \tag{3.9}$$

Diffuse Reflectance (R), given in equation 3.10, is measured by component of radiation reflected from the sample. An incident light penetrates into sample and some part of light absorbed and some part reflected back again. Measuring the reflected light gives information about the relation between reflected light intensity and analyte concentration in sample. Reflectance technique is better for solid samples [18].

$$R = \frac{I}{I_r} \quad A_R = \log_{10}\left(\frac{1}{R}\right) \quad A = \log_{10}\left(\frac{I_r}{I}\right) \tag{3.10}$$

where I = intensity of light diffusively reflected from the sample and I_r = intensity of light reflected from background or reference reflected surface.

This technique is mostly used with solid samples. Baseline is acquired by the radiation reflected from a background reference, I_r. By using same source, reflectance of sample is also measured. These two spectra used to obtain related absorbance value [18]. A design of monochromator scanning instrument is given in Figure 3.6.

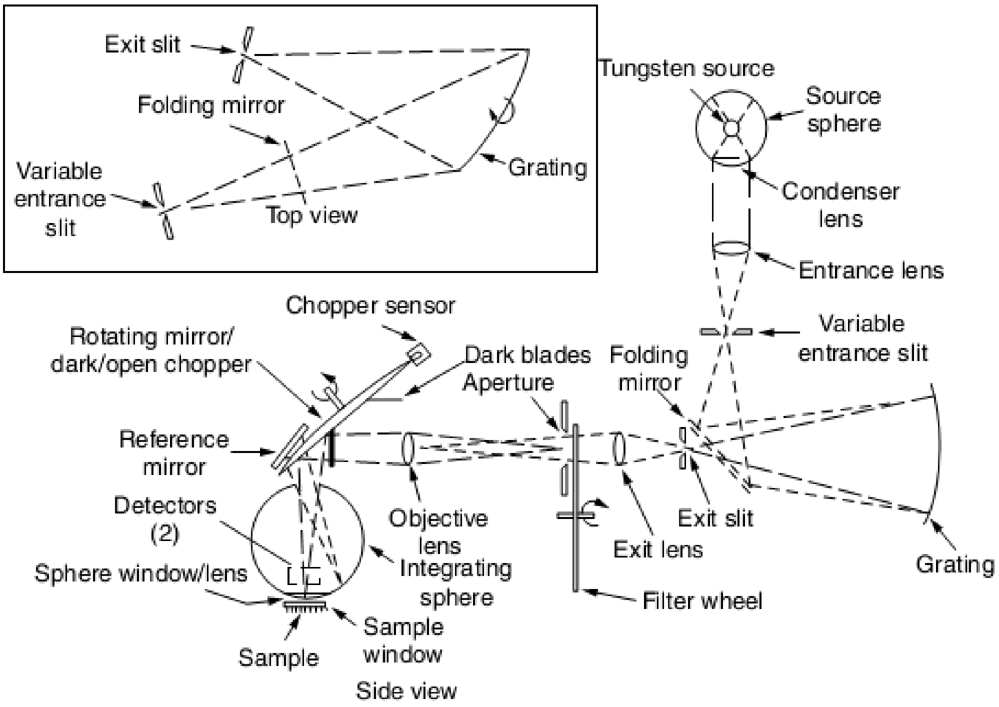


Figure 3.6: A design of monochromator scanning instrument [16].

In the interactance measurement type, light is interacted with sample more than the other techniques. Thus, we gather more information about the sample composition. Interactance as shown in Figure 3.5 e), mainly for solid samples and it is a good technique for quantitative determination in pharmaceutical studies. The path length that light travel through sample is 65 times greater than the thickness of the drug tablet [18].

3.2.1 Overtones and combinations

As discussed earlier, together with the vibrational bands, there are overtone bands, rising from transitions for more than one energy level and combination bands, rising from combination of two or more fundamental vibrations. In order to make these overtones or combinations, the energy absorbed by the molecule should exactly be same as energy levels between two transition levels. In addition to this, there should be a dipole moment change because of vibrational motion of the molecule. When considering combinations are allowed by anharmonicity, it is possible that one specific combination of vibrations is infrared active (causing a change in dipole moment) and this can only be displayed in NIR spectrum, not in MIR spectrum. The intensity of absorption band is directly related with the degree of change in dipole moment [18].

NIR spectra are dominated by hydrogen which are mainly overtone and combination bands of some fundamental groups containing C-H, O-H, and N-H bonds. Common NIR bands are given in Table 3.2 with respect to vibrational groups. Some detailed information about NIR bands with structure information is given in Table A.1 in Appendix A.

Table 3.2: NIR bands of some vibrational groups [17].

Wavelength (nm)	Assignment
2200–2450	Combination C–H stretching
2000–2200	Combination N–H stretching, combination O–H stretching
1650–1800	First overtone C–H stretching
1400–1500	First overtone N–H stretching, first overtone O–H stretching
1300–1420	Combination C–H stretching
1100–1225	Second overtone C–H stretching
950–1100	Second overtone N–H stretching, second overtone O–H stretching
850–950	Third overtone C–H stretching
775–850	Third overtone N–H stretching

By looking at the NIR spectra of a simple and a complex structure, as given in the following figures, very big differences are observed between spectra according to the complexity.

In Figure 3.7, NIR spectrum of Chloroform – $CHCl_3$ where only one Hydrogen atom causes the absorption. There are almost no broad typical NIR bands.

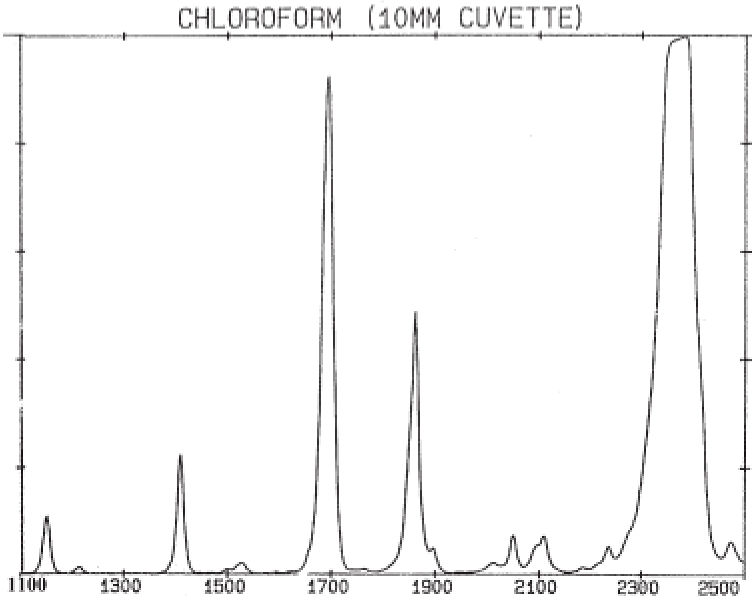


Figure 3.7: NIR spectrum of chloroform [14].

In order to understand structural differences in absorption bands, it is valuable to analyze NIR spectrum of pure hydrocarbons. One can easily observe different absorption peaks between normal paraffin and iso-paraffin hydrocarbons as given in Figure 3.8 [21].

Straight chain normal paraffins which are n-Hexane, n-Heptane and n-Octane have similar NIR spectra since their molecular structure are similar except the number of methylene groups, $-CH_2-$. But NIR spectra of branched hydrocarbons which are 2,2-Dimethylpentane and 2,2-Dimethylbutane have different absorption peaks at around 4400 cm^{-1} and they have more different spectral features compared with normal paraffins [21].

NIR spectra of naphtha, n-hexane – C_6H_{14}, toluene – C_7H_8, cyclohexane – C_6H_{12} are shown in Figure 3.9 Since the structure becomes more complex, the spectra include broad NIR bands.

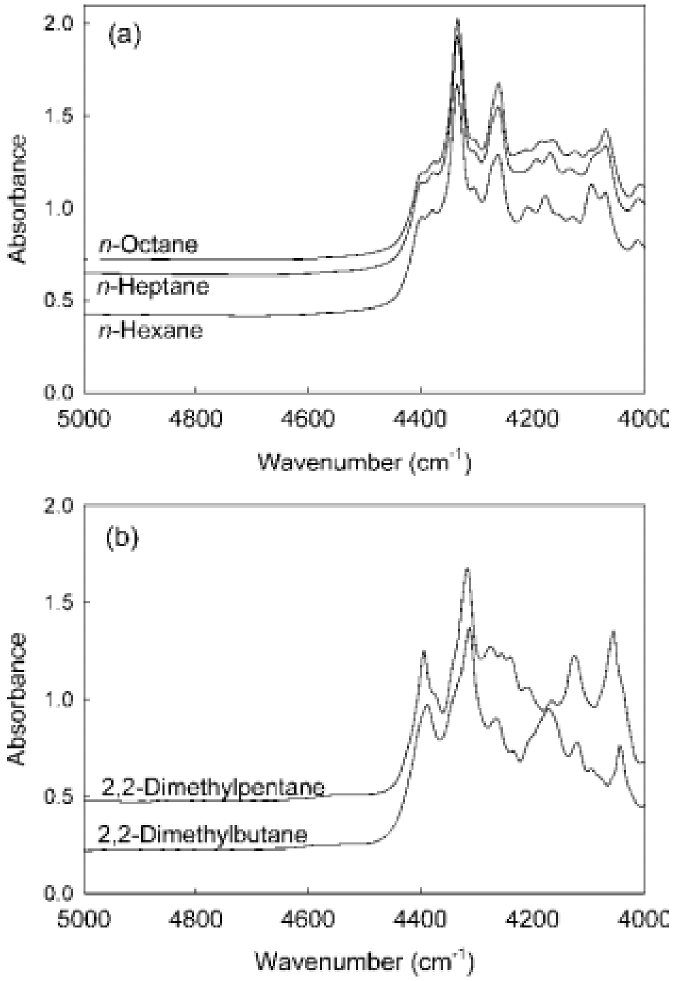


Figure 3.8: NIR spectra, between 5000 – 4000 cm^{-1} of normal paraffin and iso-paraffin hydrocarbons [21].

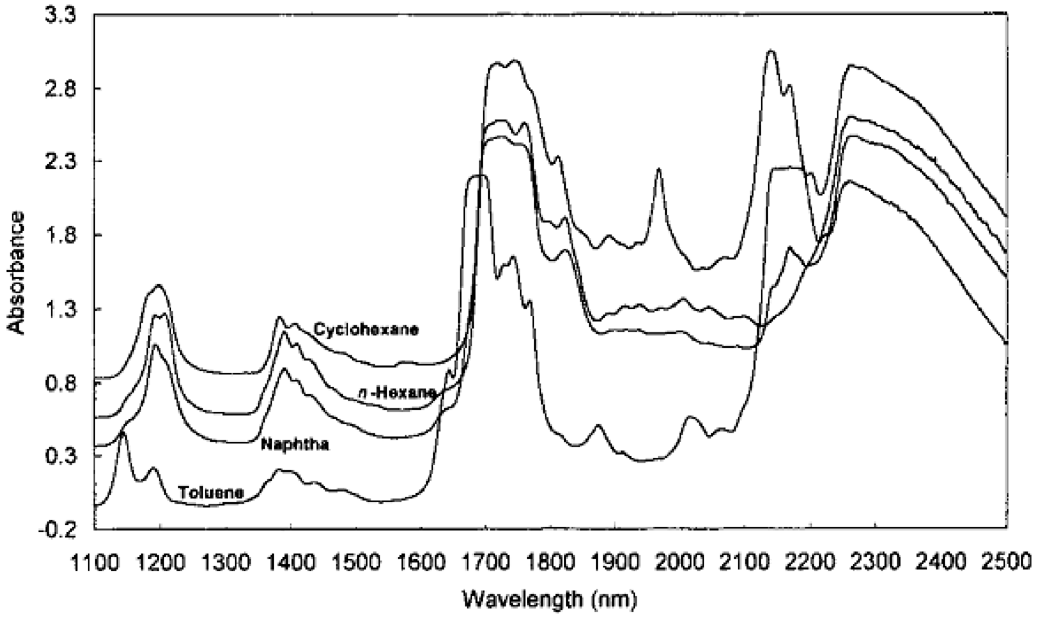


Figure 3.9: NIR spectra of naphtha, n-hexane, toluene and cyclohexane (offset for clarity) [4].

According to Ku et al. (1998), very valuable spectral information can be seen in 1100-1650 nm and 1800-2100 nm regions. The 1650-1800 nm and 2100-2500 nm ranges do not contain valuable information because of strong and saturated absorption bands form a long optical path length. According to structures of individual samples, differences in NIR absorption bands are observed. The bands around 1200 nm stand for the second overtone of fundamental CH stretching band at 3000-2700 cm^{-1} in MIR region. The second overtone of a fundamental absorption at 2870 cm^{-1} is 3 x 2870 cm^{-1}, that is 8610 cm^{-1} or 1161 nm. The bands around 1400 nm stand for the combination bands with the first overtone of CH stretching band at 3000-2700 cm^{-1} range and CH_3/CH_2 bending around 1450 cm^{-1}. This combination band wavelength is found as 2 x 2870 cm^{-1} + 1450 cm^{-1}, that is 7190 cm^{-1} or 1390 nm. The bands around 1800-2100 nm range stand for the shoulder of the strongly absorption combination band at 2300 nm, that is formed by combination of CH stretching band at 3000-2700 cm^{-1} range and CH_3/CH_2 bending around 1450 cm^{-1}. This combination results in an absorption at 2870 cm^{-1} + 1450 cm^{-1}, that is 4320 cm^{-1} or 2314 nm [4].

3.2.2 Advantages and disadvantages of NIR

As detailed in previous given information, NIR spectroscopy has many advantages in quantitative analysis and structural identification of various type of samples in the form of solid and liquid.

NIR application is mostly chosen because of being a very rapid and nondestructive analysis. The sample after NIR analysis can be used for other purposes since there is not any chemical or physical reactions take place. Since no chemicals are used, there is not any need for sample disposal. No sample preparation before NIR analysis is required and this makes it very time saving and thus requires less man-hour for the analyses. By using the advantage of fiber optics, NIR spectroscopy can be used in many industrial areas which are in exproof or non-exproof conditions. This provides different application in many processes like drug, oil, chemical etc. Lastly, very successful calibrations and studies based on chemometrics make NIR an indispensable tool. All these advantages are good reasons to choose NIR instead of traditional methods.

However, calibration task is not easy for NIR. There should be separate calibrations for each property to be quantitatively analyzed. Also someone should take care about the calibrations to monitor the accuracy. According to process and sample, the calibrations should be updated periodically. Although chemometrics with NIR is very powerful tool, chemometrics is difficult and sophisticated to study. It requires some level of training and time to spend.

4. CHEMOMETRICS

According to International Chemometrics Society, “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” is stated as the definition of chemometrics. Advanced instrumentation on measurement with enormous valuable data and increased power in computing have taken attention of many people who study on spectroscopy, process control, food, drug, environmental chemistry, statistics etc. Then chemometrics is recognized as a sub-discipline within analytical chemistry. Although it is known as a sub-discipline within analytical chemistry, it takes an integral part of many scintific disciplines. The relation between different disciplines is given in Figure 4.1 [19].

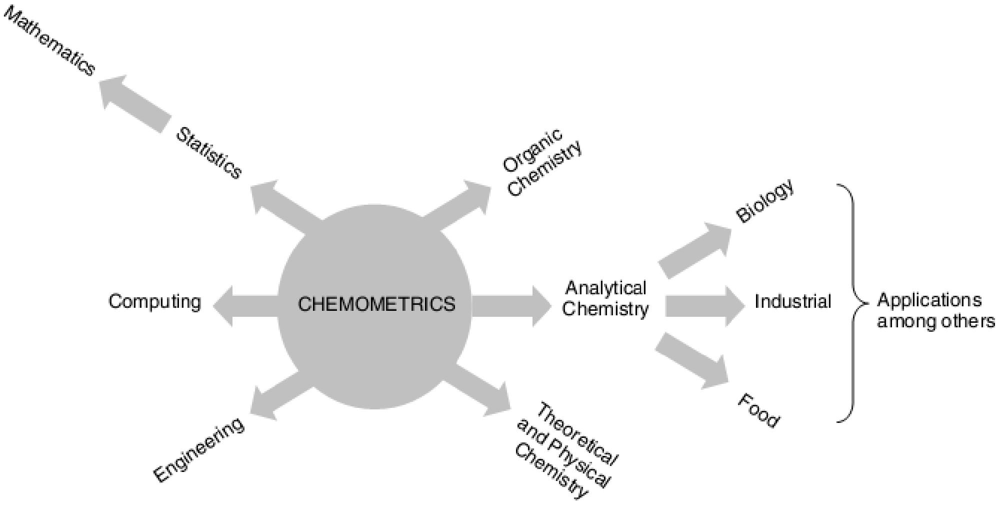


Figure 4.1: Relation between other disciplines with respect to chemometrics [19].

Chemometrics is an interdisciplinary science that involves statistics, mathematical modelling, computer science and analytical chemistry. Major application areas are calibration, validation, optimization of chemical measurements and experimental procedures, and getting most out of chemical information from analytical analysis.

The first word “Chemometrics” was mentioned in the 1970s and mainly involved in multivariate analytical data derived from analytical chemistry analysis data. In the 1980s, chemometrics as a discipline became organized in various journals, societies, books, meetings and it was separated from other disciplines, like computational and theoretical chemistry. In these years, chemometricians also started to use Matlab instead of FORTRAN that was used initially. Currently, with the emerging technology in computers and analytical instruments used for analysis, data sets became much bigger in size and applications in chemometry became more critical. Application areas have been expanded starting from 1970s; the first studies were in food and pharmaceutical chemistry areas. Petrochemical, environmental and bioinformatics are the other areas studied by people in both academic life and industry [19].

4.1 Calibration

Calibration is one of the most important tasks required to perform a quantitative analysis in chemistry. According to Lavine and Workman (2008), "Calibration involves relating, correlating, or modeling a measured response based on the amounts, concentrations, or other physical or chemical properties of a set of analytes." By establishing relationship between response and property, then it is possible to determine the amount of property because of obtained response from analytical instrument. A scientific law formed of mathematical formulas mostly describes this relationship; a very famous example is Beer's Law that relates the attenuation of light to the properties of the material through which the light is traveling [Url-3].

There are mainly two types of calibration, univariate and multivariate calibration. In univariate calibration, one signal response is correlated to a specific analyte property, but in the multivariate calibration, multiple responses correlated to one or more property of interest. Although there are some solutions to avoid the interferences if there is any, it is obvious some more study should be carried out in order to find the correlation in case of having a very broad NIR spectrum over a wide wavelength range. Multivariate calibration is a good solution for this type of solution to establish a prediction model.

4.1.1 Univariate Calibration

A very common and known example of univariate calibration is to determine the concentration of a single compound by using the response of single detector, correlation of concentration of compound of interest to a peak found at a fixed wavelength for a spectroscopic measurement. In order to apply univariate calibration, the signal or response must be selective for the interested analyte property. According to IUPAC "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" [33]. Thus univariate calibration is also named as single-component analysis. There are two methods of getting univariate calibration, classical calibration and inverse calibration.

4.1.1.1 Classical Calibration

Evaluating the correlation of a single component concentration to a response is a simple explanation of classical univariate calibration. Beer's Law is a very good example to apply this rule. Absorbance is considered as a function of absorbance from a spectroscopic measurement. It is simply given as;

$$A = \varepsilon bc \tag{4.1}$$

where A is absorbance, ε is molar absorptivity, c is concentration. In vector notation, we have following equation.

$$\mathbf{a} \approx \mathbf{c}\ .\ \mathrm{s} \tag{4.2}$$

where **a** is the vector of absorbances (response) at one wavelength for a number of samples, and **c** is the vector of corresponding concentrations. These two vectors have same size, which is the number of samples. The scalar coefficient s is related with these parameters and can be calculated (via pseudo-inverse) by solving 4.2 as follows:

$$\begin{gathered}\mathbf{c}'.\mathbf{a} \approx \mathbf{c}'.\mathbf{c}\ .\ \mathrm{s} \\ (\mathbf{c}'.\mathbf{c})^{-1}.\mathbf{c}'.\mathbf{a} \approx (\mathbf{c}'.\mathbf{c})^{-1}.(\mathbf{c}'.\mathbf{c})\ .\ \mathrm{s} \\ s \approx (\mathbf{c}'.\mathbf{c})^{-1}.\mathbf{c}'.\mathbf{a} = \frac{\sum_{i=1}^{I} c_i a_i}{\sum_{i=1}^{I} c_i^{\,2}}\end{gathered} \tag{4.3}$$

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

When s is calculated, the prediction model is as follows:

$$\hat{a} \approx \hat{c}.s \tag{4.4}$$

where $\hat{a}$ and $\hat{c}$ are refer to prediction. Since s is calculated from the model, for a given absorbance, the corresponding concentration can be calculated, or for a given concentration, the corresponding absorbance can be calculated.

In order to see the prediction error (e), also named as residual, the difference between observed and predicted values is calculated [34].

$$e \approx a - \hat{a} \tag{4.5}$$

4.1.1.2 Inverse Calibration

The response is predicted in univariate calibration, but in most cases, predicting concentration from the response is more of interest, especially in analytical chemistry. Another reason to use inverse calibration is the source of prediction error. The prediction error in classical calibration comes from the error in response, which is essentially instrumental, but with improving technology, the instruments are very reliable. The instrumental errors very small when compared to errors occurred by the measurement of concentration. Weighing and dilution steps, equipment used (flask, pipette etc.), and human factor are very common sources of contamination used in prediction model. Thus, inverse calibration is better to predict the concentration and that is why it is used extensively. Classical calibration fits a model so that the errors are in the response, while inverse calibration fits a model so that the errors are in the concentration [34]. A schematic view of error source for two calibration types is in Figure 4.2.

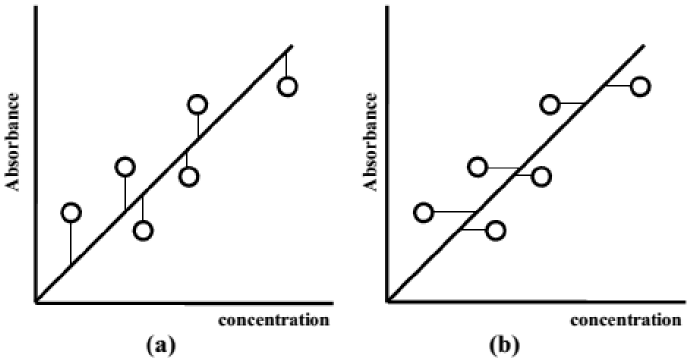


Figure 4.2: Errors in Classical calibration (a) and Inverse calibration (b) [34].

Since concentration is a function of absorbance than we have following equations in vector notation:

$$\mathbf{c} \approx \mathbf{a}.b \tag{4.6}$$

$$\mathrm{b}{=}(\mathbf{a}'.\mathbf{a})^{-1}.\mathbf{a}'.\mathbf{c} = \frac{\sum_{i=1}^{I} a_i.c_i}{\sum_{i=1}^{I} a_i^{\,2}} \tag{4.7}$$

$$\hat{c} \approx \hat{a}.b \tag{4.8}$$

where b is scalar coefficient. The error in prediction of concentration is given in equation 4.9.

$$e \approx c - \hat{c} \tag{4.9}$$

b is only approximately inverse of s, since each calibration has different approach to error calculation as given in Figure 4.2. For a good data, both models should give similar prediction results and errors. Intercept, non-linearity, noise can affect the data and the prediction differs between two calibration methods [19].

In both regressions, the regression line is forced through zero since it is assumed that the intercept is zero. This assumption gives poor fit, because of other components in the sample also absorbs.

4.1.1.3 Intercept

Mostly an intercept is added to inverse calibration model, as follows:

$$c \approx b_0 + b_1.a \tag{4.10}$$

and matrix/vector notation is given in equation 4.11.

$$\mathbf{c} \approx \mathbf{A}.\mathbf{b} \tag{4.11}$$

where **c** is vector of concentrations, **b** is vector consisting of b_0 as intercept and b_1 as slope. **A** is a matrix of two columns, first column of 1's and second column is responses. Same pseudo-inverse rule is applied to calculate coefficients, b_0 and b_1.

$$\mathbf{b} \approx (\mathbf{A}'.\mathbf{A})^{-1}.\mathbf{A}'.\mathbf{c} \tag{4.12}$$

The predicted concentrations are calculated by using below equation given in matrix notations [34].

$$\hat{\mathbf{c}} \approx \mathbf{A}.\mathbf{b} \tag{4.13}$$

4.1.2 Multivariate Calibration

Multivariate calibration is used in many applications and it depends on a vector or matrix of data for each sample, such as a full wavelength range of spectrum, whereas univariate calibration depends on a single scalar measurement for each sample. There are many different calibration techniques available; they only differ from each other in methods used to calculate regression coefficients.

Interferences are very common problems to establish a calibration according to Beer's Law. If there is a specific response for interested specific property or component in a sample, then univariate calibration is good enough. In case of having difficult samples and corresponding interferences in the spectra, then univariate calibration does not respond very well, and we need multivariate calibration to overcome this problem. Instead of correlating the concentration to a single wavelength, using full spectrum or important wavelengths in calibration as multivariate calibration, is very successful. Thus, the difficulties caused by interferences are removed and it is very time saving method when compared to removing interferences by applying difficult and long pre-analysis techniques. A schematic explanation of this advantage is given in Figure 4.3. As seen in the graph, it is not possible to have a reliable calibration when there are interferences between the components in the sample for analyzed wavelength range [39].

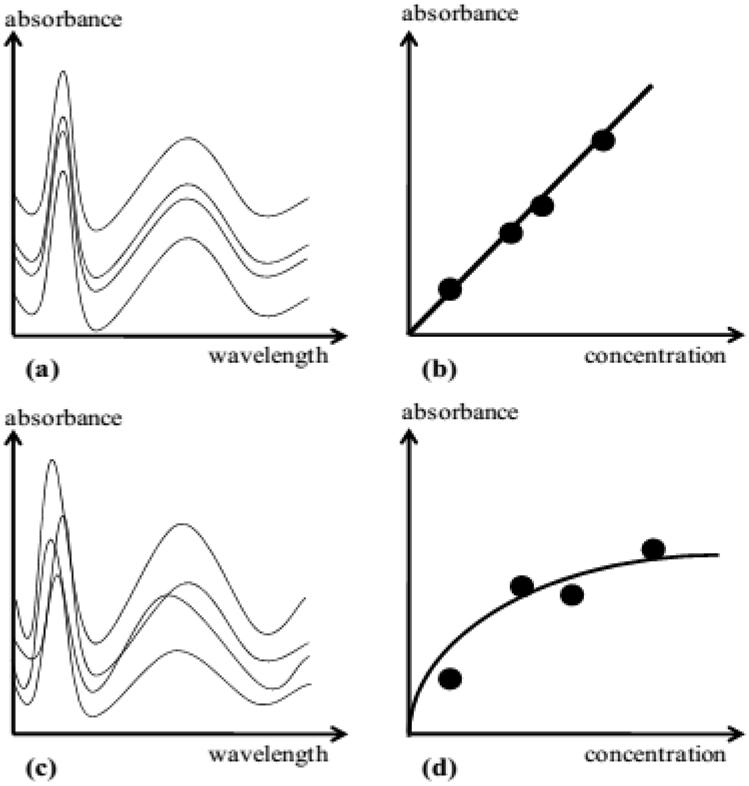


Figure 4.3: a) Some spectra of multiple samples without any interferences, b) univariate calibration of samples in a). c) Some spectra of multiple samples with interferences, d) univariate calibration of samples in c) [34].

A very general calibration procedure consists of following steps. A set of samples is collected and the compositional range is included. Then a spectroscopic measurement (or any other interested measurement to get a data set) is performed and corresponding spectrum is collected in a given wavelength range. In addition to spectroscopy measurement, reference analysis of interest is executed to have concentrations for each sample. Lastly, the sample set is divided into two sets, one set for calibration and one set for validation studies, and one of multivariate calibration methods is applied with calibration set [28].

4.1.2.1 Classical Least Squares (CLS)

In classical least squares, also known as K-matrix method, the response is considered as a function of concentrations, where response is treated as dependent variable whereas the concentration is independent variable. A very common example is Beer's

Law, absorbance is a function of concentration. Matrix notation form of CLS is given as;

$$\mathbf{A} \approx \mathbf{C}.\mathbf{B} \tag{4.14}$$

where $\mathbf{A}$ is a *mxn* matrix of spectra, m samples measured at n wavelength, $\mathbf{C}$ is a *mxp* matrix of concentrations of p components for each m samples. $\mathbf{B}$ is calculated by solving linear equations.

$$\hat{\mathbf{B}} = (\mathbf{C}'.\mathbf{C})^{-1}.\mathbf{C}'.\mathbf{A} \tag{4.15}$$

CLS is a good calibration technique if all components in sample contributing to the spectra are identified and their concentrations are known. The concentration of the interested property must be related with the defined spectrum. Otherwise some properties might have interferences with another property that affects the spectrum. Another condition for CLS is that the number of components must be less than or equal to number of experiments or wavelengths.

4.1.2.2 Multiple Linear Regression

Multiple linear regression is also called as inverse least squares where concentration is modeled as function of response, i.e. spectrum. It is not required anymore to identify all components in the sample. Therefore, it is possible to apply this methods only property of interest. The difficulty with the MLR is that the number of samples in calibration set must be greater than the number of wavelength measured. It is impossible to have this condition with high technology spectroscopy instruments; a smaller set of wavelength must be selected.

The matrix notation of MLR is given as below;

$$\mathbf{C} \approx \mathbf{A}.\mathbf{B} \tag{4.16}$$

where **C** is mxp matrix of concentrations of p properties for m samples, and **A** is mxn matrix of responses of n wavelengths of m samples. **B** is nxp matrix of regression coefficients for n wavelengths and p properties. Once this equation is solved, **B** matrix is found as;

$$\mathbf{B} = (\mathbf{A}'.\mathbf{A})^{-1}.\mathbf{A}'.\mathbf{C} \tag{4.17}$$

Since the number of samples must be greater than the number of wavelengths, there are some algorithms to reduce the number of wavelengths to the most meaningful ones in order to use inverse method. In order to do this, very good knowledge is required about the errors occurred. Then the regression equation takes the following form for the i_{th} property.

$$c_i = b_{0i} + b_{1i}A_1 + b_{2i}A_2 + b_{3i}A_3 + \ldots \tag{4.18}$$

4.1.2.3 Principal Component Analysis (PCA)

A very common problem with the multivariate calibration is the difficulty to see the relation within the variables since the data is very large. Latest technology spectrometers can produce data as a result of an analysis in a very wide wavelength range. The aim to use principal component analysis is to reduce the dimension of the data, by capitalizing on the colinearity of the data, in order to make the evaluation without losing any valuable information. In addition to this, it is a very useful tool about identifying the relation between variables. PCA, a schematic view is given in Figure 4.4, is used for many different multivariate calibration techniques as an initial step before establishing prediction model. Two most common PCA algorithms are NIPALS and SVD (Singular Value Decomposition).

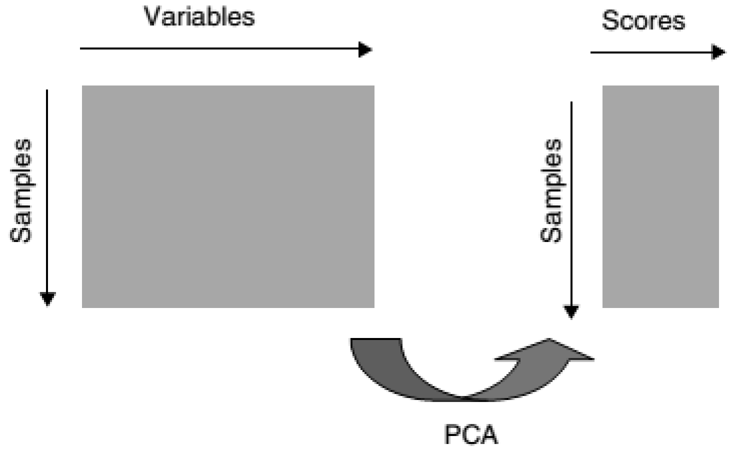


Figure 4.4: Data reduction in PCA [30].

The original data, in form of a matrix, is transformed as following in PCA analysis,

$$\mathbf{X}_{MxN} = \mathbf{T}_{MxA}.\mathbf{P}_{AxN} + \mathbf{E}_{MxN} \qquad \textbf{(4.19)}$$

where **X** is orginial data matrix, **T** are the scores, **P** are the loadings, and **E** error matrix associated from transformation. Scores and loadings matrices can only be calculated if their dimensions are smaller or equal to smallest dimension of original matrix, which is also the maximum number of principal components to be calculated. The number of rows in the original data matrix is also the number of rows in score matrix (usually number of samples). The number of columns for score matrix are limited to number of principal components. For loadings matrix, the number of columns equals to the number of columns in the original data matrix (response i.e. detectors, wavelengths). The number of rows is determined by the number of principal components and each row corresponds to one principal component. First principal component, PC1, is defined by a loading vector as follows, where m is the number of variable.

$$\mathbf{p}_1 = (p_1, p_2, p_3, \ldots, p_m) \qquad \textbf{(4.20)}$$

The corresponding score vectors are linear combinations of loadings and the variables. For sample i, score t_{i1} for PC1 is as following;

$$t_{i1} = x_{i1}.p_1 + x_{i1}.p_2 + \ldots + x_{im}.p_m \qquad \textbf{(4.21)}$$

For all n samples (objects) arranged as rows in X matrix, the score vector $\mathbf{t}_1$ of PC1 is obtained by equation 4.22.

$$\mathbf{t}_1 = \mathbf{X}.\mathbf{p}_1 \qquad \textbf{(4.22)}$$

where $\mathbf{p}_1$ is given as;

$$\mathbf{P} = \begin{pmatrix} p_1 & p_3 \\ p_2 & p_4 \end{pmatrix} \qquad \textbf{(4.23)}$$

$$PC1 = \mathbf{p}_1 = [p_1, p_2]; \; PC2 = \mathbf{p}_2 = [p_3, p_4]$$

In mathematical terms, the principal components are the eigenvectors of the covariance matrix of original data matrix. Eigen analysis is applied to find eigen value and corresponding eigenvector. Eigen value gives the amount of variation in the data set and the eigen value of PC1 has larger variation.

Once loading matrix (principal component eigenvectors) **P** and score matrix **T** are calculated, **X** matrix can be predicted back.

$$\hat{\mathbf{X}} = \mathbf{T}.\mathbf{P}' \; ; \; \mathbf{X} = \mathbf{T}.\mathbf{P}' + \mathbf{E} \; ; \; \mathbf{E} = \mathbf{X} - \hat{\mathbf{X}} \tag{4.24}$$

As stated above, the maximum number of PCs must be smaller or equal to minimum dimension of original data matrix. All PCs are not required to be used in PCA analysis and further multivariate calibration methods. There are a few ways of determining the number of PCs to be used in analysis, like comparing the eigen values for each PC (eigen vector) and cross validation.

Eigen value of a PC (eigen vector) is mentioned as sum of squares of scores, with g_a is a_{th} eigen value and I is number of rows (objects, samples etc) in original matrix.

$$g_a = \sum_{i=1}^{I} t_{i1}^{\;2} \tag{4.25}$$

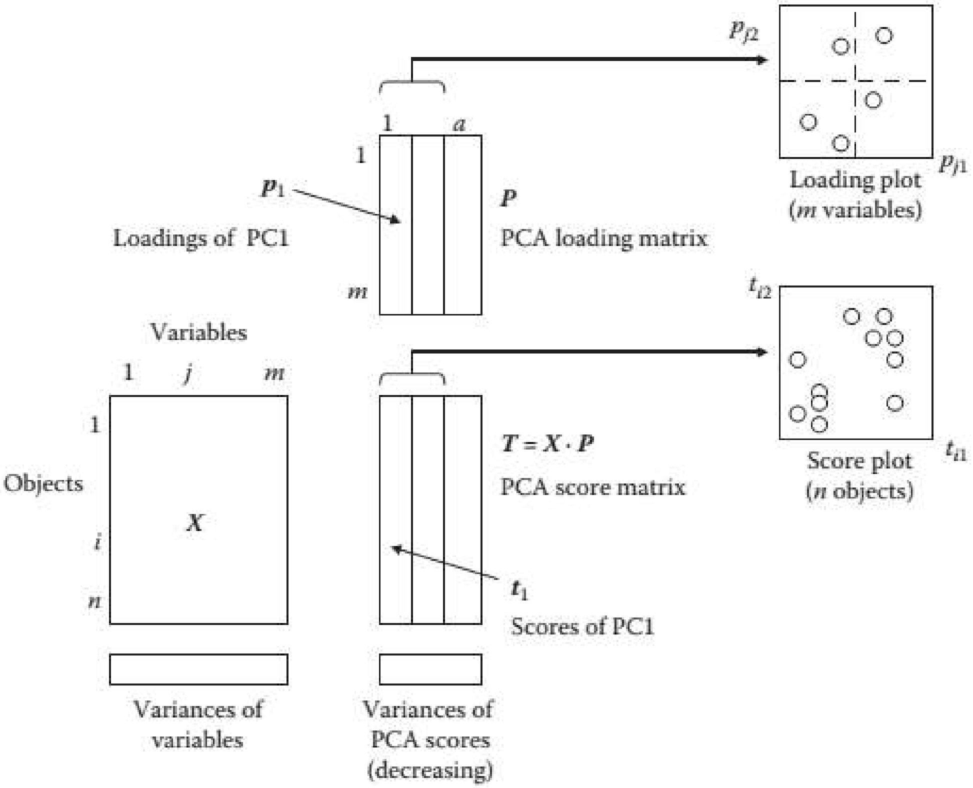


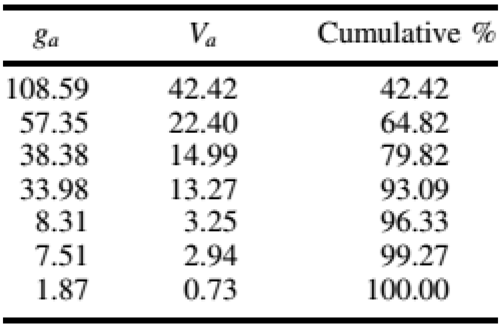
Figure 4.5: A schematic explanation of PCA and data reduction [30].

Eigen values are also analyzed in percentages and cumulative percentages in order to see the contribution of each PC to variation. A graph of number PCs vs cumulative percentage is very common.

$$V_a = 100 \frac{g_a}{\sum_{i=1}^{I} \sum_{j=1}^{J} x_{ij}^{2}} \tag{4.26}$$

Once the contribution to variation by adding another PC becomes less than %5 to cumulative variation, and then it is good to keep number of PCs instead of increasing more. A random example is given in Table 4.1. Thus, 4 PCs that has % cumulative of %93.09, would be enough since there is not significant contribution with 5th PC. A summary of PCA is given in Figure 4.5.

Table 4.1: Eigen values and cumulative in percentage [36].



g_a	V_a	Cumulative %
108.59	42.42	42.42
57.35	22.40	64.82
38.38	14.99	79.82
33.98	13.27	93.09
8.31	3.25	96.33
7.51	2.94	99.27
1.87	0.73	100.00

4.1.2.4 Principal Component Regression (PCR)

Contrary to multiple linear regression, that requires information about all components in the sample, principal component regression (PCR) method only requires the concentration information of interested components (or property, compounds) in the sample. This makes PCR very important in use of spectroscopic measurement data, in which the spectra collected in a very large/wide wavelength range but limited information about the compounds (property) in the sample, except the interested ones. Since the original data is very large, PCA analysis is necessary to calculate scores matrix, **T**, and loadings matrix, **P**. The next step is regression (transformation or rotation) by finding a relation between scores matrix, **T**, and the concentrations of interested components in the sample. The concentrations of interested components are determined by reference analysis methods in a trusted laboratory. The accuracy of the concentration results are very critical in the prediction error of regression model.

$$\mathbf{c}_a = \mathbf{T}.\mathbf{r}_a \tag{4.27}$$

The concentration of interested component n is $\mathbf{c_a}$ and $\mathbf{r_a}$ is a column vector (also called as transformation or rotation vector) having equal number of rows with the number of principal components. It is preferred to choose the number of PCs equal to the number of components in the sample [19,34].

The vector $\mathbf{r_a}$ is calculated by equation 4.28 since **c** and **T** are known.

$$\mathbf{r}_n = (\mathbf{T}'.\mathbf{T})^{-1}.\mathbf{T}'.\mathbf{c}_n \tag{4.28}$$

Finally, it is possible to estimate the concentrations of components interested by knowing scores matrix, **T** [34]. A schematic explanation of PCR is given in Figure 4.6. If there is concentrations to be calculated more than one, we have the following matrix form of above equations.

$$\mathbf{C} = \mathbf{T}.\mathbf{R} \tag{4.29}$$

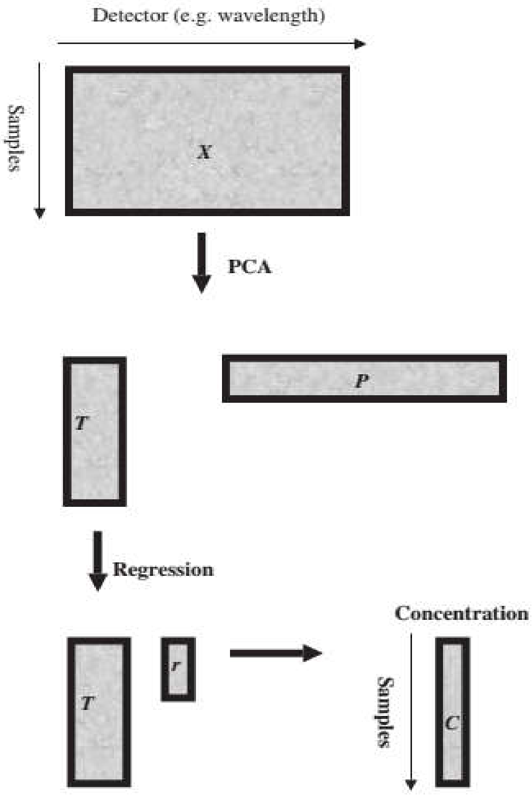


Figure 4.6: A schematic of Principal Component Regression [37].

4.1.2.5 Partial Least Squares (PLS)

Partial least squares (PLS) is mostly accepted as major multivariate calibration technique. Similar to PCR, PLS is also forms the linear combinations of predictor variables (concentration), but the way of choosing the linear combinations is different. In PLS, the errors coming from both concentration estimate and spectra are used, whereas PCR assumes that the concentration predictions do not have any errors. While determining the concentration of interested component in a laboratory, there are many source of errors like sample preparation, personal, instrument etc. These errors are much larger than instruments error in spectra. Then, it is meaningful to consider the errors in concentration. In PLS, the covariance between both variables' spectra (**X**) and concentration (**c**) is minimized. Modelling concentration data and spectra data together is main principle of PLS. Thus, modelling the concentration data is also important as modelling the spectral data [37]. A schematic view is given in Figure 4.7.

There are several algorithms for PLS, it is expressed mainly by the equations given below.:

$$\mathbf{X} = \mathbf{T}.\mathbf{P} + \mathbf{E} \tag{4.30}$$

$$\mathbf{c} = \mathbf{T}.\mathbf{q} + \mathbf{f} \tag{4.31}$$

where **q** is similar to a loadings vector.

T and **P** are for estimating the spectra data and **T** and **q** are estimating the concentration data. As seen above, **T** score matrix is same for both set of data. The PLS eigenvalues, sum of squares of each component, is different than PCA eigenvalues since both concentration and spectra are taken into account. For each PLS, there are spectral scores vector **t**, spectral loadings vector **p**, and concentration loadings scalar q [34]. The concentration of compound n is predicted by the below equations.

$$\hat{c}_{in} = \sum_{a=1}^{A} t_{ian} q_{an} + \overline{c_i} \tag{4.32}$$

$$\mathbf{c_n} = \mathbf{T_n}.\mathbf{q_n} + \overline{\mathbf{c}}_{\mathbf{n}} \tag{4.33}$$

where $\overline{\mathbf{c}}_{\mathbf{n}}$ is the vector of average concentration.

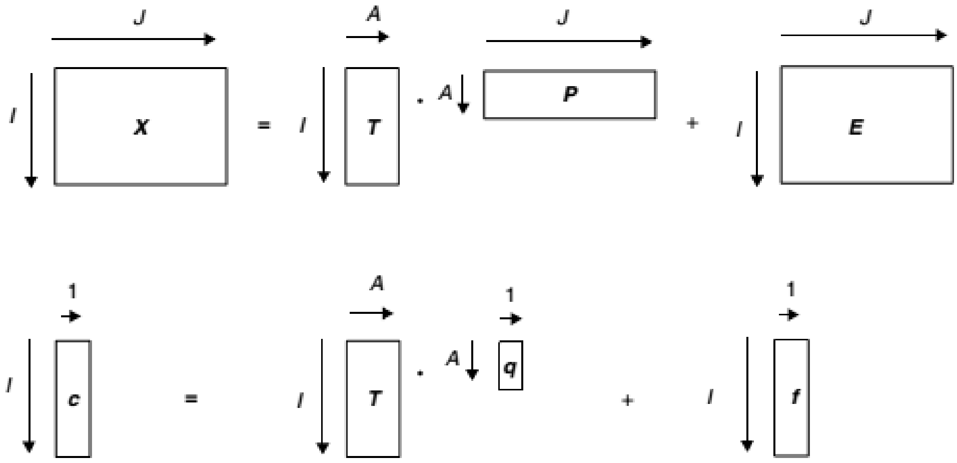


Figure 4.7: A schematic presentation of Partial Least Squares (PLS) [34].

PLS and PCA are similar to each other in dimension reduction. PLS reduces the dimension of a data set by projecting the data onto components of maximum variance with a second data set, which is y. Matrix notation of this definition is as following.

$$\mathbf{X} = \mathbf{T}.\mathbf{P}' + E_X \tag{4.34}$$

$$\mathbf{Y} = \mathbf{U}.\mathbf{Q}' + F_Y \tag{4.35}$$

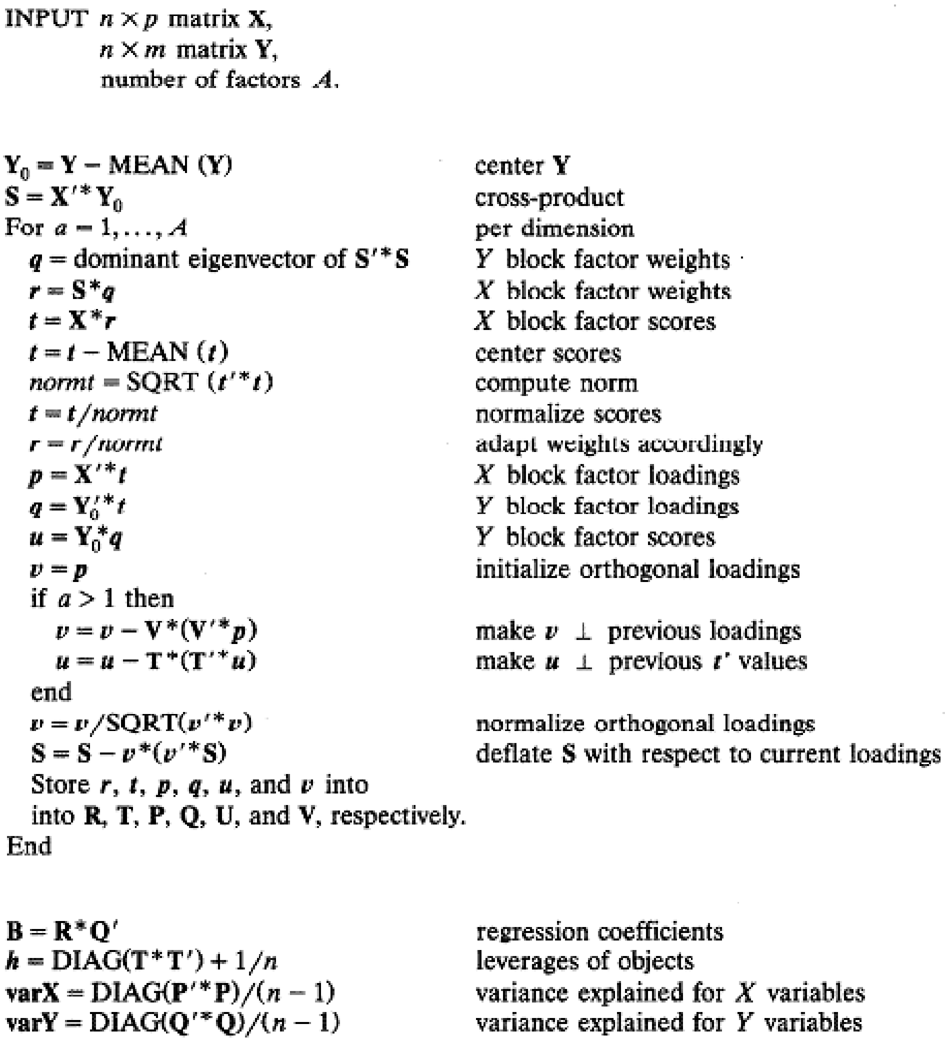
where **X** is a matrix of predictors (i.e. NIR spectra), **Y** is corresponding matrix of responses (i.e. reference analysis results), **T** and **U** are matrices which are projections of **X** and **Y**, score matrices. P and Q are orthogonal loading matrices. E and F terms are errors. In PLS, the covariance between score matrices is maximized.

As PLS regression followed, one gets couple of factors, which used to explain data further. Well knowns are loading graphs, residual graphs and score graphs, and they are used for group identification, outlier detection, and data analysis [40]. As mentioned above, PLS regression is also used for multivariate calibration tool to get prediction models. A regression vector, composed from regression coefficients, shows the interaction between raw data and predicted data. If there is only one y variable to predict, the algorithm used to make prediction model is called as PLS1, and if there are multi-variable to predict, it is as PLS2 algorithm. For quantitative analysis using spectroscopy data, usually PLS1 is used since it is possible to use different number of PLS factors (PCs) for each y variable.

There are many algorithms used for PLS1 regression and they give very close results. Some differences may arise from chosen number of factors or significant factors used in calculations. SIMPLS [42] and NIPALS (non-linear iterative partial least squares) [43] are very well known two alternative algorithms. In this thesis, SIMPLS algorithm is used for PLS regression. For univariate regression, SIMPLS algorithm is same as PLS1 algorithm. SIMPLS algorithm formulated by Jong [42] is given in Table 4.2.

For both PCA and PLS, **R** (combination of **r** vectors) defines the transformation of **X** to **T**. In PCA, **T** explains the variance of **X** but in PLS **T** explains covariance between **X** and **Y**.

Table 4.2: SIMPLS algorithm formulated by Jong [42]



```
INPUT $n \times p$ matrix $\mathbf{X}$,
      $n \times m$ matrix $\mathbf{Y}$,
      number of factors $A$.

$\mathbf{Y}_0 = \mathbf{Y} - \text{MEAN}(\mathbf{Y})$                  center Y
$\mathbf{S} = \mathbf{X}'^{*}\mathbf{Y}_0$                              cross-product
For $a = 1, \ldots, A$                                                  per dimension
  $q$ = dominant eigenvector of $\mathbf{S}'^{*}\mathbf{S}$             $Y$ block factor weights
  $r = \mathbf{S}^{*}q$                                                 $X$ block factor weights
  $t = \mathbf{X}^{*}r$                                                 $X$ block factor scores
  $t = t - \text{MEAN}(t)$                                              center scores
  $normt = \text{SQRT}(t'^{*}t)$                                        compute norm
  $t = t/normt$                                                         normalize scores
  $r = r/normt$                                                         adapt weights accordingly
  $p = \mathbf{X}'^{*}t$                                                $X$ block factor loadings
  $q = \mathbf{Y}_0'^{*}t$                                              $Y$ block factor loadings
  $u = \mathbf{Y}_0^{*}q$                                               $Y$ block factor scores
  $v = p$                                                               initialize orthogonal loadings
  if $a > 1$ then
    $v = v - \mathbf{V}^{*}(\mathbf{V}'^{*}p)$                          make $v \perp$ previous loadings
    $u = u - \mathbf{T}^{*}(\mathbf{T}'^{*}u)$                          make $u \perp$ previous $t'$ values
  end
  $v = v/\text{SQRT}(v'^{*}v)$                                          normalize orthogonal loadings
  $\mathbf{S} = \mathbf{S} - v^{*}(v'^{*}\mathbf{S})$                   deflate S with respect to current loadings
  Store $r$, $t$, $p$, $q$, $u$, and $v$ into
  into R, T, P, Q, U, and V, respectively.
End

$\mathbf{B} = \mathbf{R}^{*}\mathbf{Q}'$                                regression coefficients
$h = \text{DIAG}(\mathbf{T}^{*}\mathbf{T}') + 1/n$                      leverages of objects
$\mathbf{varX} = \text{DIAG}(\mathbf{P}'^{*}\mathbf{P})/(n-1)$          variance explained for $X$ variables
$\mathbf{varY} = \text{DIAG}(\mathbf{Q}'^{*}\mathbf{Q})/(n-1)$          variance explained for $Y$ variables
```

4.2 Developing Chemometrics based Analytical Methods

Multivariate calibration techniques are widely used in industry as qualitative and quantitative analytical methods performed in both laboratory and process [60, 62, 63]. When fast response of NIR spectroscopy is combined with chemometrics modelling, it is very useful to use it in order to reduce cost and analyze more samples for a quality control laboratory. While doing this, keeping the precision and accuracy of results obtained by prediction models same or even lower than the reference method is one of the main advantages. Besides performing chemometrics based analytical methods off-line in the laboratory, having them on-line and/or in-line process systems makes life easier for companies which have continuous manufacturing process. There are mainly two chemometric-related systems for manufacturing process named as Process Analytical Technology (PAT) is in mostly pharmaceutical business and Process Analytical Chemistry (PAC) is in mostly other business segments [59, 60]. PAT is defined by US Food and Drug Administration (FDA) as “a system for designing, analyzing, and controlling manufacturing through timely measurements (i.e., during processing) of critical quality and performance attributes of raw and in-process materials and processes with the goal of ensuring final product quality” [Url-12]. Although it is defined for PAT but it is same for every industry that uses similar chemometric methods to control the process.

In order to have reliable chemometric prediction models, there should be a dedicated team and very well written procedures about development, evaluation and maintenance of methods. An example of groups involved this type of study is given in Figure 4.8.

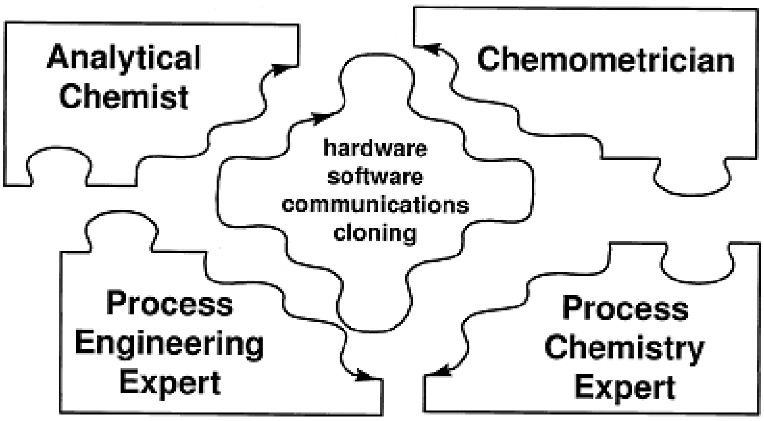


Figure 4.8: Process analytical chemistry team approach [61].

The essential steps to produce chemometric prediction models are summarized in below Table 4.3. Selection of samples before starting the study is very critical, since the samples must cover the range to be studied or planned to study. The rule of thumb is that the variation in concentration must be ±5 R (reproducibility) of the reference method. Then, performing reference analysis and spectroscopic measurements must be done by qualified personnel very precisely. Model building, validating and predicting steps play very important role since the model will be used as replacement for the reference method. Chemometrician must apply best techniques for the lowest prediction error. As a result of continuous manufacturing processes, the process is not stable and samples collected for the process are subject to change, thus the models always should be updated by collecting new samples not included at the initial steps.

Table 4.3: Essential steps for building chemometric prediction model [18].

1. Selection of calibration and validation test
2. Determination of the concentration for property interested by using reference test method
3. NIR spectra analysis
4. Development and optimization of the multivariate calibration model
5. Validation of calibration model
6. Predicting unknown samples by using calibration model
7. Maintenance of the calibration model

5. EXPERIMENTATION and INSTRUMENTATION

5.1 Experimentation

Commercial gasoline samples, produced in Tupras Izmit Refinery as non-oxygenate and with oxygenate containing MTBE, were collected with two grades in RON property, which are 95 RON and 97 RON. Samples were kept at 4 °C temperature and in dark place before analysis to avoid any evaporation and possible interferences as a result of direct light. All gasoline samples were tested in Tupras Izmit Refinery Quality Control Laboratory, accredited from ISO/IEC 17025:2005 (General requirements for the competence of testing and calibration laboratories), according to standard test methods given in TS EN 228 - Automotive Fuels - Unleaded Petrol Requirements and Test Methods - Specification.

Basic descriptive statistical information of analyses results and precision data for test methods are given in Table 5.1.

Repeatibility, r and reproducibility, R values are two very important precision data for any test method. Repeatibility is defined by ASTM as the difference between successive test results, obtained by the same operator using the same apparatus under constant operating conditions on identical test material, would in the long run, in the normal and correct operation of test method. Reproducibility is defined as the difference between two single and independent test results, obtained by different operators working in different laboratories on identical test material, would in the long run, in normal and correct operation of test method. R and r values were calculated by using average values for each property.

A brief table including all the analysis results are given in Appendix Table A.2. These analyses results were used as reference results to be modeled by multivariate calibration techniques. Only the distillation E150 property is used to get a prediction model for distillation characteristics.

Table 5.1: Basic statistical data for 45 gasoline sample analysis results

	min	max	average	std deviation	r	R
RON	94.9	97.4	95.5	0.8	0.2	0.7
MON	85.1	86.9	85.9	0.4	0.2	0.9
Aromatics, %(v/v)	29.7	39.9	33.8	2.3	0.5	1.7
Olefins, %(v/v)	1.2	9.8	5.5	2.0	0.2	1.2
Benzene, %(v/v)	0.57	0.95	0.78	0.09	0.02	0.04
Density, kg/l	0.7291	0.7491	0.7398	0.0055	0.0003	0.002
Distillation - E70, deg C	33.8	48.6	40.9	3.9	0.9	2.2
Distillation - E100, deg C	55.3	66.8	61.6	2.9	0.7	1.8
Distillation - E150, deg C	89.3	92.9	91.2	0.9	0.5	1.2

5.2 Instrumentation

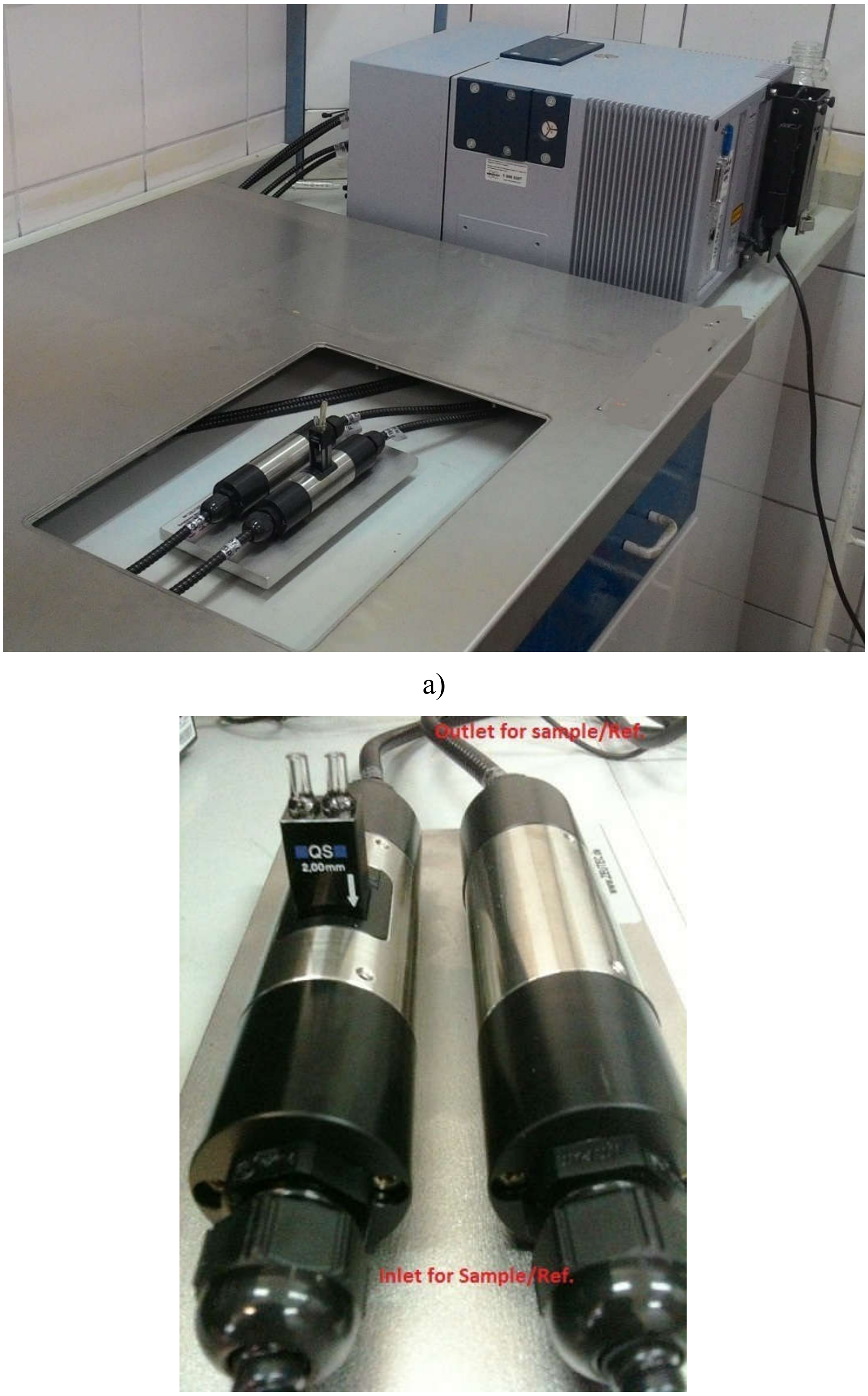
All reference analysis were done by the Quality Control Laboratory.of Tupras Izmit Refinery. A brief information about the analysis and the pictures of instruments and apparatus were given as following.

5.2.1 NIR

45 gasoline samples were analyzed by FT-NIR spectroscopy (MATRIX-F, Bruker-Germany) in wavelength rage of 800 – 2500 nm. For regression analysis, the spectral region of 1100 – 2200 nm was used. Before starting the analysis cycle, a background against air was taken. All analysis were done at room temperature 23 – 25 ^oC.

NIR instrument, set-up and flow cell apparatus are given in Figure 5.1.

Absorbance spectrum was collected by single channel NIR, using 10 mm quarts cuvette with InGaAs amplified photodetector with thermo-electric cooled (Te-InGaAs). The instrumental parameters used in NIR analysis are given as follows; resolution 8 cm^{-1}, background and sample scan time 16 scans, scanner velocity 10 KHz, open aperture. Fourier Transform parameters are as give follows; phase resolution 128, phase correction mode: power spectrum, apodization function: Blackman-Harris 3-term, zero-filling factor: 2.



a)

b)

Figure 5.1: a) NIR Instrument and b) Flow Cell apparatus from laboratory.

5.2.2 Octane Number – RON and MON

RON and MON analysis were done by using Cooperative Fuel Research (CFR) engines, Figure 5.2, (Waukesha, CFR Engines Inc., USA) which testing capability in the 40-120 octane number range. Main parts of the CFR engine are variable compression ratio cylinder (4:1 to 18:1) and sleeve assembly, four-bowl falling level carburetor, CFR crankcase, intake air humidity equipment, exhaust surge system, and knock meter [Url-13].

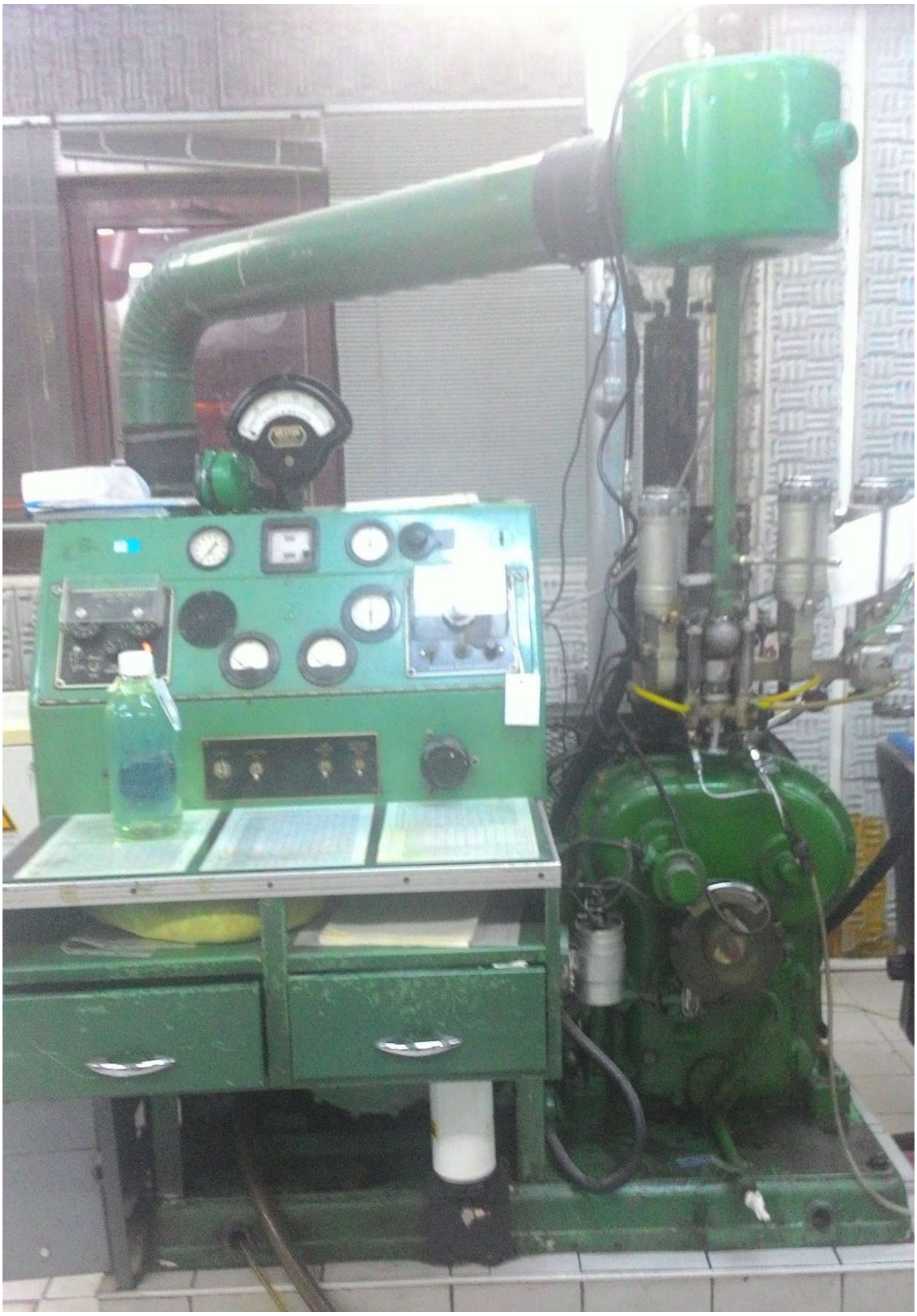


Figure 5.2: CFR Engine from laboratory.

5.2.3 Distillation

Distillation analysis were done by using atmospheric distillation instruments, Figure 5.3, (OptiDist, PAC, USA) according to TS EN ISO 3405 (ASTM D 86) standard test method. Following analysis parameters, given in Table 5.2, were set while performing distillation analysis.

Table 5.2: Distillation analysis parameters [7].

Sample temperature: < 10°C
Temperature of cooling bath: 0 – 1 °C
Temperature of bath around receiving cylinder: 13 – 18 °C
Distillation rate: 4 – 5 ml/min.

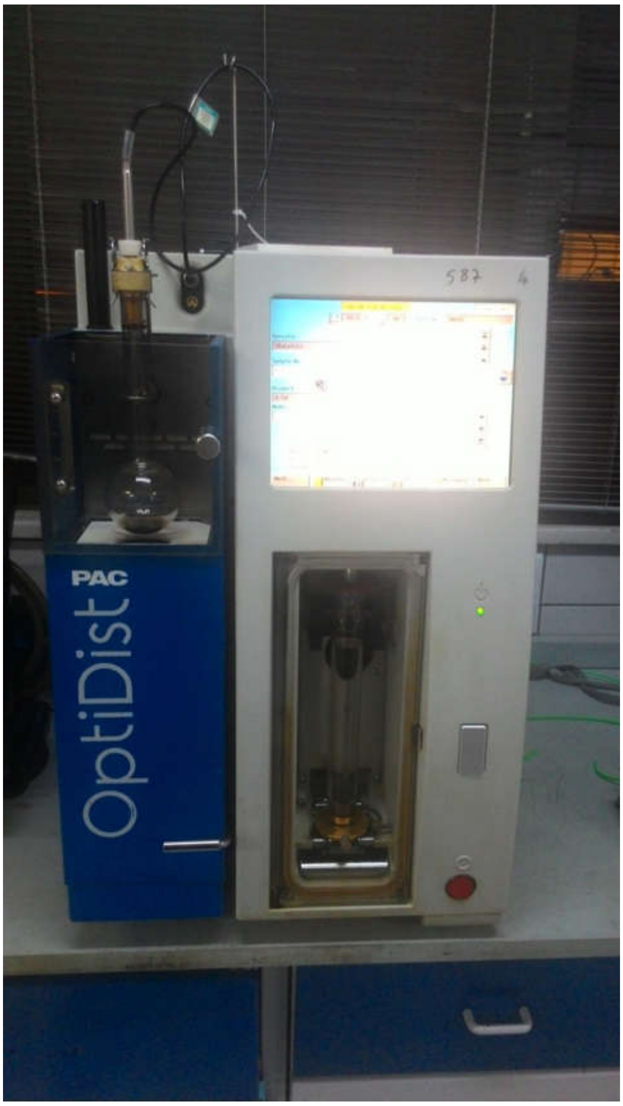
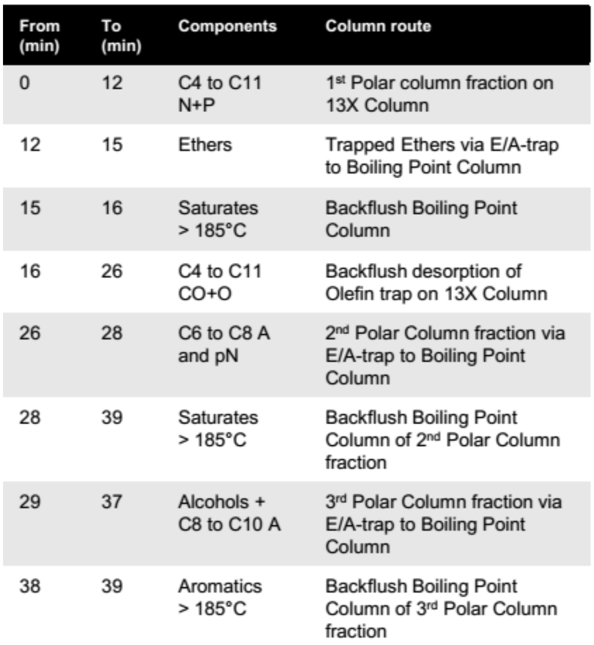


Figure 5.3: Atmospheric Distillation instrument from laboratory.

5.2.4 Hydrocarbon Types

Hydrocarbon types, aromatics, olefins and benzene, were analyzed by multidimensional gas chromatography, Figure 5.5, (AC Reformulyzer, PAC, USA) according to TS EN ISO 22854 standard test method. The instrument has benefits from using auto injector, FID detector, capillary/micropacked columns and traps to get a good separation and save analysis time. In Table 5.3, the parameters and usage of traps and columns are given [64]. The picture of gas chromatography and flow diagram of application are given in Figure 5.4 and Appendices Figure B.1 [64], correspondingly.

Table 5.3: Gas Chromatography parameters [64].



From (min)	To (min)	Components	Column route
0	12	C4 to C11 N+P	1st Polar column fraction on 13X Column
12	15	Ethers	Trapped Ethers via E/A-trap to Boiling Point Column
15	16	Saturates > 185°C	Backflush Boiling Point Column
16	26	C4 to C11 CO+O	Backflush desorption of Olefin trap on 13X Column
26	28	C6 to C8 A and pN	2nd Polar Column fraction via E/A-trap to Boiling Point Column
28	39	Saturates > 185°C	Backflush Boiling Point Column of 2nd Polar Column fraction
29	37	Alcohols + C8 to C10 A	3rd Polar Column fraction via E/A-trap to Boiling Point Column
38	39	Aromatics > 185°C	Backflush Boiling Point Column of 3rd Polar Column fraction

Figure 5.4: Picture of Gas Chromatography from laboratory.

5.2.5 Density

Density measurements were done by density meter, Figure 5.6, (DMA 5000M, Anton Paar, USA) according to TS EN ISO 12185 (ASTM D 4052) standard test method. Measurement temperature was set to 15.6 °C as reference temperature.

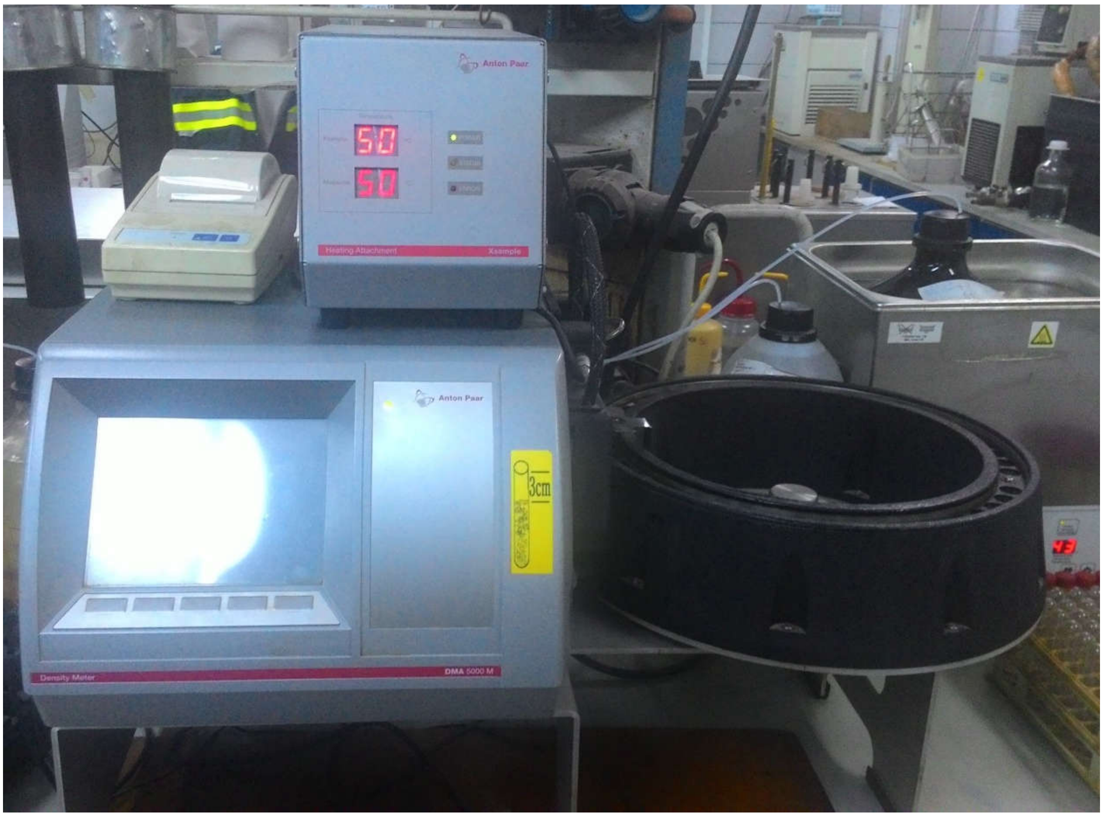


Figure 5.5: Density Meter form laboratory.

5.3 Data Analysis

The collected spectra were transferred in ASCII format and were combined in Microsoft® Excel® 2013. Then the data converted to text file for regression analysis. MATLAB R2014b (MathWorks Inc., MA) with PLS_Toolbox Graphical User Interface (GUI) (Eigenvector Research Inc., USA) was used for PCR and PLS regression analysis.

6. RESULTS & DISCUSSION

6.1 Data pre-processing

Before applying the regression techniques, baseline correction is applied by using *baseline* function of PLS_Toolbox, version 8.0.2 and release 18015, (Eigenvector Research Inc., USA). The methodology here is 'The Weighted Least Squares' which determines the spectral regions due to baseline only. Baseline correction was done in order to remove baseline offsets from raw data. This application is very helpful where there is signal variation due to baseline or background. Some specific baseline references are required, where there is no spectral information, and they are used as basis to eliminate baseline effects for the whole spectrum. The range used in this NIR spectrum is as following; 1280 – 1330, 1540 – 1580 and 1900 – 1950 nm. 46 raw NIR spectra without baseline correction and with baseline correction are given in Figure 6.1 and Figure 6.2 correspondingly

After correcting baseline, another most common pre-processing technique which is 'mean centering' was applied to NIR spectra data. After mean centering is applied to data, each row of mean-centered data includes how that individual row is different from the average sample in the original data matrix. Mean centering is applied as given in Equation 6.1.

$$b_{ij} = a_{ij} - \frac{1}{n}\sum_{j=1}^{n} a_{ij} \qquad \mathbf{(6.1)}$$

where a_{ij} is original row entry and b_{ij} is the mean centered entry.

For the $\mathbf{X}_{\mathbf{mxn}}$ matrix, i is from 1 to m (number of samples) and j is from 1 to n (number of wavelengths). By applying same formula to each entry in the matrix, we can get mean centered matrix. Thus the mean-centered matrix is 46 x 1179, that is the NIR spectra of 46 samples with absorbance for 1179 different wavelengths.

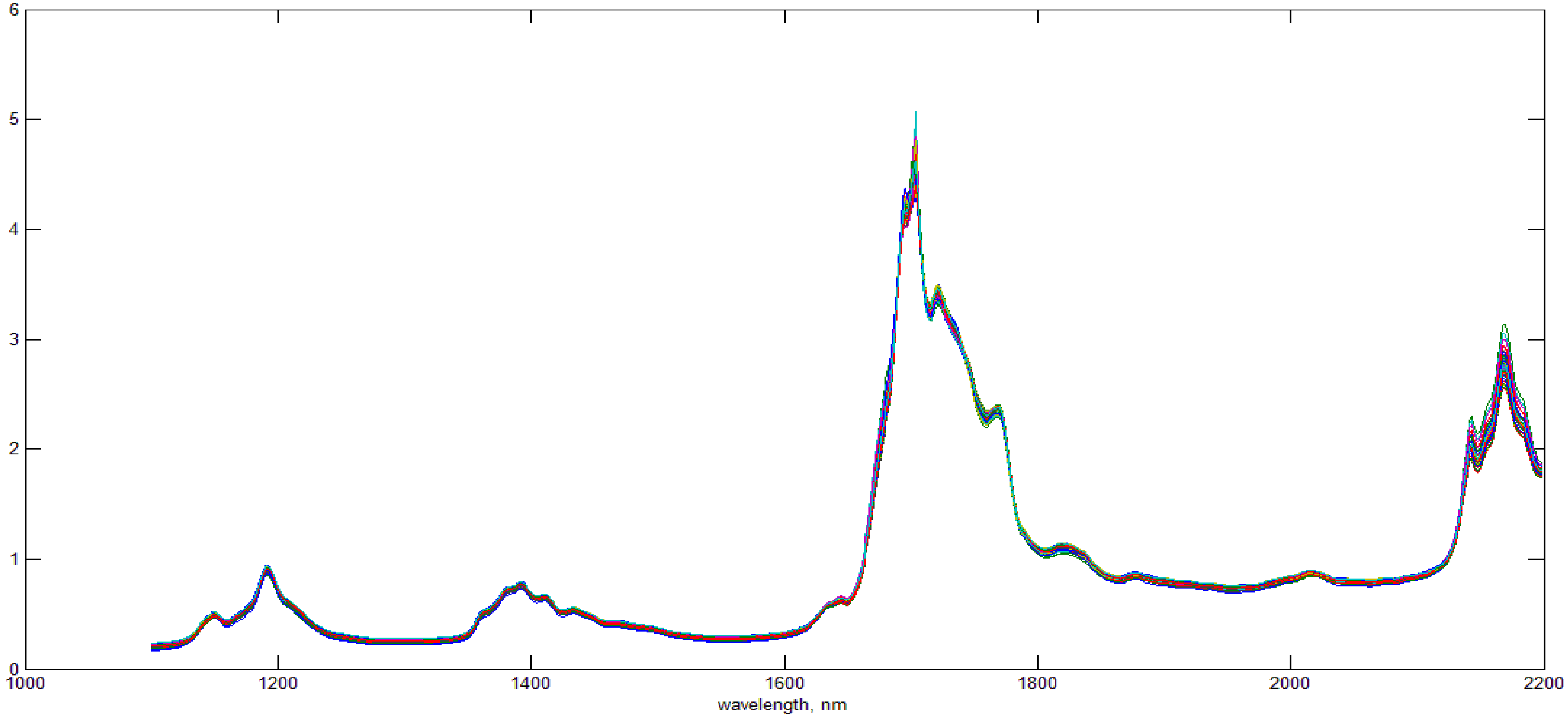


Figure 6.1: NIR Spectra without baseline correction.

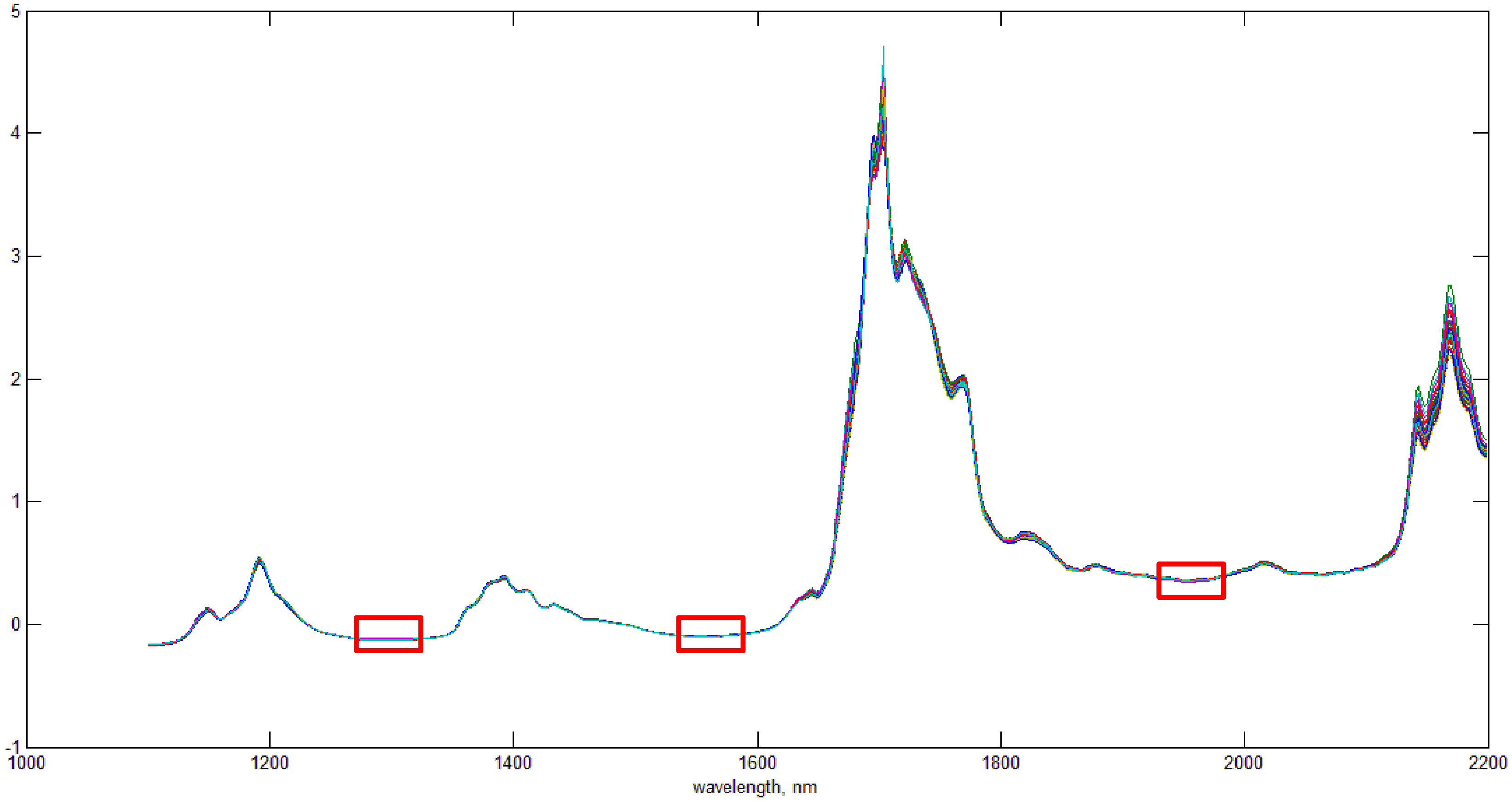


Figure 6.2: NIR Spectra with baseline correction.

As a last step before starting regression, all 46 gasoline sample NIR spectra were checked against any outlier in the data set. Some useful plots, which are scores on Latent Variable 1 (Principal Component 1) and scores on Latent Variable 2 (Principal Component 2), Q residuals and Hotelling's T^2, measured and predicted were used to identify any outlier after mean centering the data after determining the principal components for spectral data by using PLS_Toolbox Graphical User Interface (GUI).

Q residuals are calculated as the sum of squares of each row of error matrix, **E**, given in equation 4.17 and 4.26, resulting from prediction and are the measure of difference -between sample and its projection [Url-4].

$$Q_i = e_i e_i^T \tag{6.2}$$

where Q_i is the Q residual for i^{th} sample and e_i is the i^{th} row of error matrix, **E**.

Hotelling's T^2 is a form of Student's t distribution for multivariate analysis and Hotelling's T^2 values represent a measure of variation in each sample in model. Hotelling's T^2 is the sum normalized squares of scores given in scores matrix.

$$T_i^2 = t_i \lambda^{-1} t_i^T \tag{6.3}$$

where t_i is the i^{th} term of score matrix, **T**, from the model and λ is diagonal matrix of eigenvalues [Url-4].

Hotelling's T^2 and Q residuals are very helpful statistics used to explain how a model is describing a given sample data set.

As seen in Figures 6.3 and 6.4, data with sample number 2 is clearly an outlier and it was removed from data set before regression analysis. In Figure 6.4, Studentized Residual vs Leverage graph (at top-right corner) is given. Leverage shows the influence of a sample on the model, smaller the leverage, better model fit for the inspected sample. Sample 2 has a big leverage and away from the 0 line at studentized axis. Thus, after the outlier removal, the mean-centered data matrix is 45x1179 that is the NIR spectra of 45 sample with absorbance for 1179 different wavelenghths.

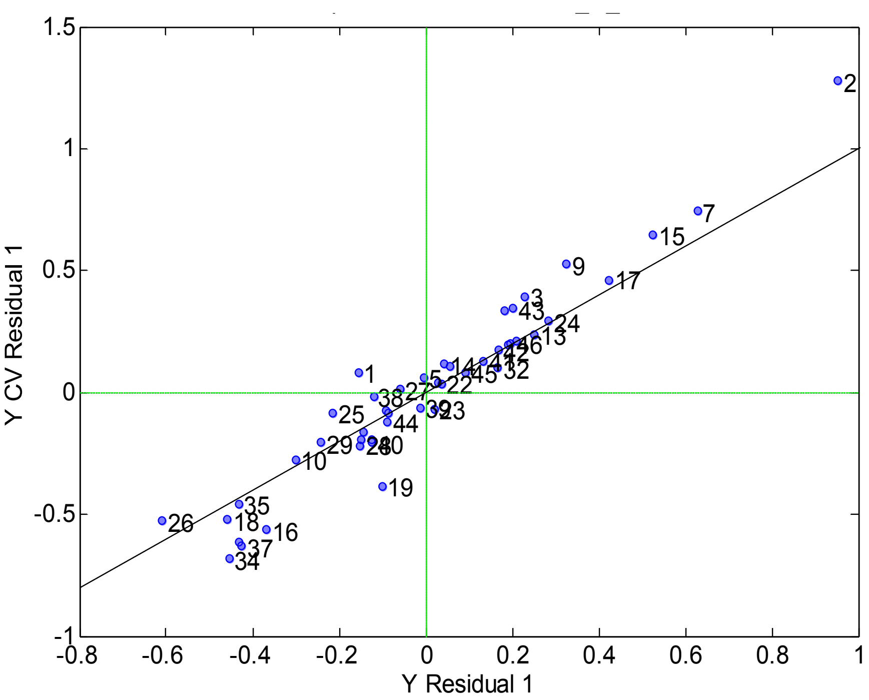


Figure 6.3: Residual vs Cross-Validation Residual Plot.

6.2 Principles and Essential Practices used for Regression Analysis

After data pre-processing the sample data set, which is composed of 45 gasoline NIR spectra after removing outlier, the main data set was divided into two data sets as calibration data set and validation data set, sample 1- 30 and sample 31 – 45 correspondingly. $X^{calibration}$ has size of 30x1179 and $X^{validation}$ has size of 15x1179 for each property to be modelled. The calibration data set was used for establishing the prediction model for the related regression technique and the validation set was used to predict the properties related to the NIR spectral data by using the prediction model to assess the predictive power of the developed model.

Any regression algorithm requires a certain number of variables, i.e. Principal Components or Latent Variables, to be included in the prediction model. Choosing the best number of variables is very important and a few common rules of thumb should be followed. At this point, the plots gathered after regression analysis are very helpful to determine the number of variables according to some critical parameters. Using external set of samples to be predicted by the model and cross-validation procedures are two main tools to estimate the number of variables.

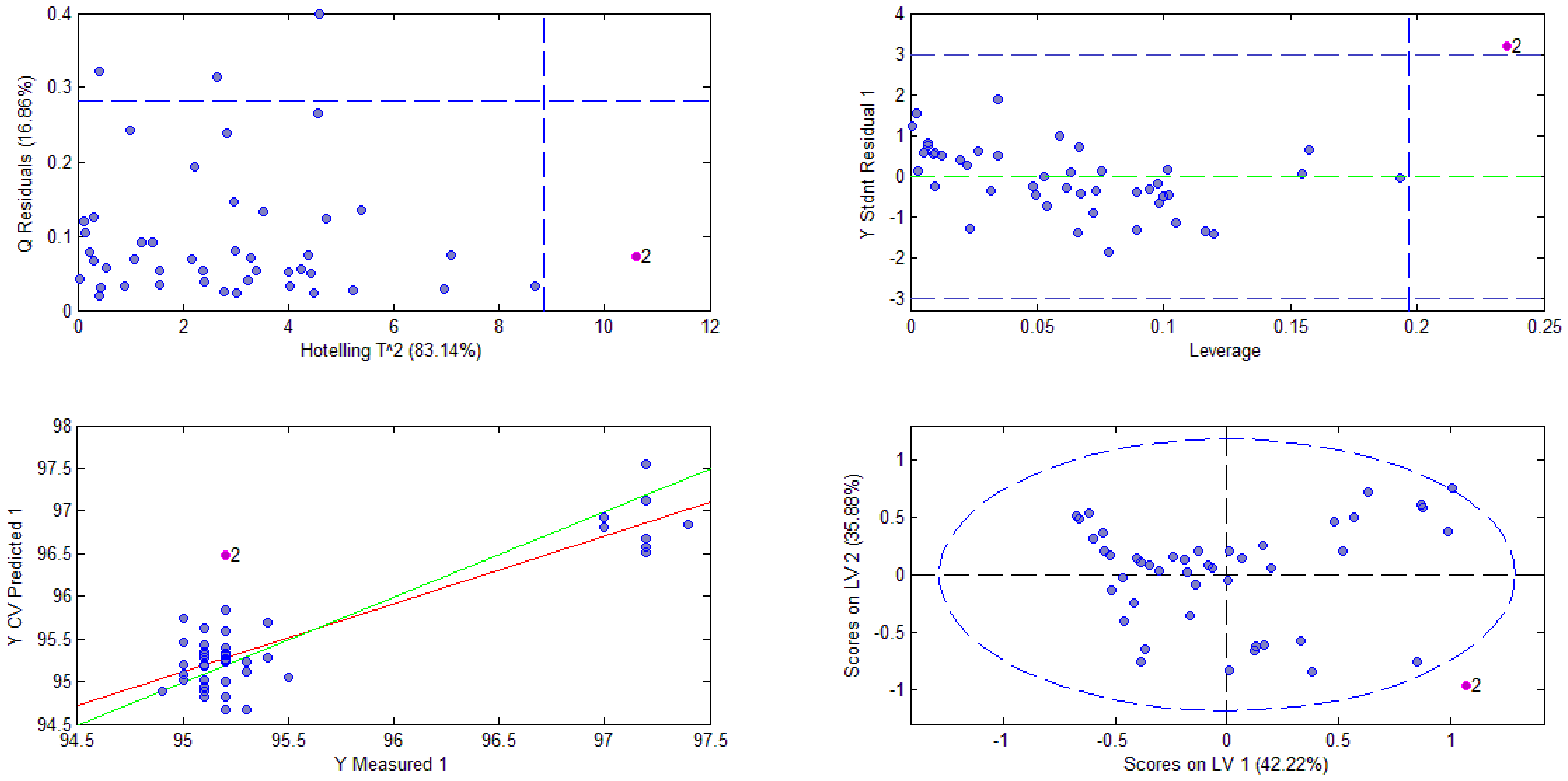


Figure 6.4: Hotelling's T^2 vs Q Residuals – Scores on LV1 vs Scores on LV2 – Measured vs Predicted – Leverage vs Studentized Residual.

The cross-validation will remove one sample or a given number of samples from the calibration data set and construct the model without this data and the sample(s) remaining is predicted by the model, resulting with a calibration error. Then another sample or set of samples are removed from the calibration data set and prediction and error calculation repeated in similar manner. After iterating this procedure for every sample or sample set, a parameter called 'RMSECV - Root Mean Square Error of Cross Validation' is determined.

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad \textbf{(6.4)}$$

where $\hat{y}_i$ is the predicted sample results for sample not used for calibration and y_i is the reference data.

In case of using external data set for prediction, the error resulting from the predicted result is calculated by using reference results. In this case, the errors is called as 'RMSEP – Root Mean Square Error of Prediction'. The equation is same with equation 6.4 but the predicted y values comes from determining external data set values by using prediction model.

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad \textbf{(6.5)}$$

where $\hat{y}_i$ is the predicted results for external data set and y_i is the reference result. After determining these parameters, plot of PCs vs RMSECV/RMSEP is very helpful for determining the number of variables to include the model.

A very critical point in determining the number of variable is to avoid increasing the number of variable without a valid reason. A good rule of thumb is not to include any additional variable unless it improves the RMSECV by at least %2. Also considering the cumulative RMSECV is also important; covering %95 of variability is usually valid. Another very common rule is keeping the number of variables as low as possible in order to avoid complexity. If more variables than required are put into the model, one may get very good results for calibration set but poor results with predicting the external data set. This issue is called as ‘over-fitting’ and obviously it should be avoided [18, 44].

RMSEP parameter is also used for evaluating the performance of prediction model in validation step since it is determined by using and independent external data set from calibration data set.

6.3 Evaluation of NIR Spectra

As explained in detail in Section 3, near-infrared region includes weak absorption bands by mainly overtones and combinations of hydrocarbon C-H bonding, also N-H, and O-H bonding.

According to Ku et al. (1998), very valuable spectral information can be seen in 1100-1650 nm and 1800-2100 nm regions. The 1650-1800 nm and 2100-2500 nm ranges do not contain valuable information because of strong and saturated absorption bands form a long optical path length. The bands around 1200 nm correspond to second overtone of methyl and methylene CH band at 3000 – 2700 cm^{-1} MIR range and bands around 1400 nm stands for the combination bands between the first overtone of CH stretching at 3000 – 2700 cm^{-1} and CH_3/CH_2 bending at 1450 cm^{-1}. There is also a 2nd overtone of aromatic C-H band at around 1145 nm.

In Figure 6.2, the regions without any spectral information were shown in red colored rectangles. This spectral information can easily be seen in NIR spectra plot of collected gasoline samples in Figure 6.5, which shows the wavelengths range that are informative and that will be used in subsequent regressions.

NIR Spectra between 1100 – 1550 nm. The absorption bands in the 1800 – 2100 nm range (Figure 6.6) corresponds to coupling between the C-H stretching at 3000 – 2700 cm-1 and CH3/CH2 bending at 1450 cm-1 [4].

As seen in NIR spectra in these figures, gasoline sample is a very complex hydrocarbon mixture (C5 – C10 in carbon number) and this results in many overlapping bands in the spectrum. Although there are some regions are available to interpret such as specific overtones and combinations, it is required to use multivariate calibration techniques to solve these overlapping issues in order to get quantitative information interested property.

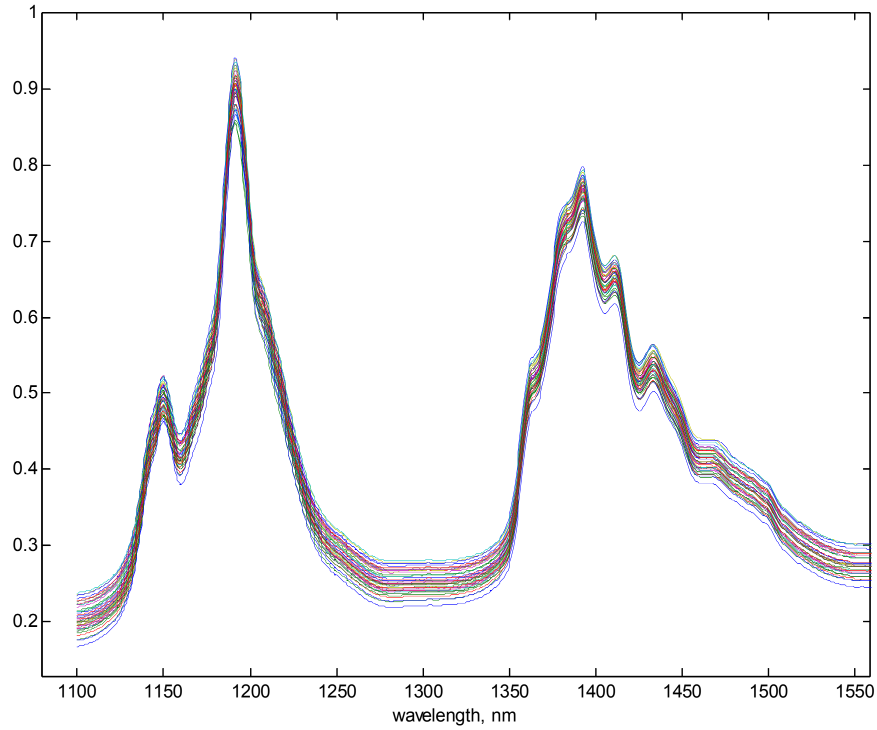


Figure 6.5: NIR Spectra between 1100 – 1550 nm.

Detailed information regarding to NIR spectral bands can be found in Section 3 with given Table 3.2 and Table A.1 in Appendix A.

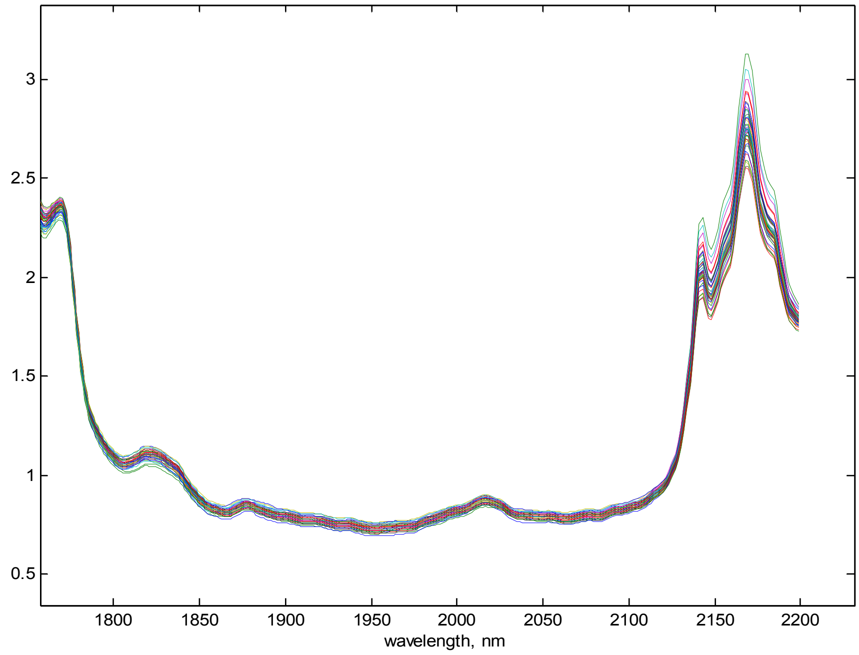


Figure 6.6: NIR Spectra between 1800 - 2200 nm.

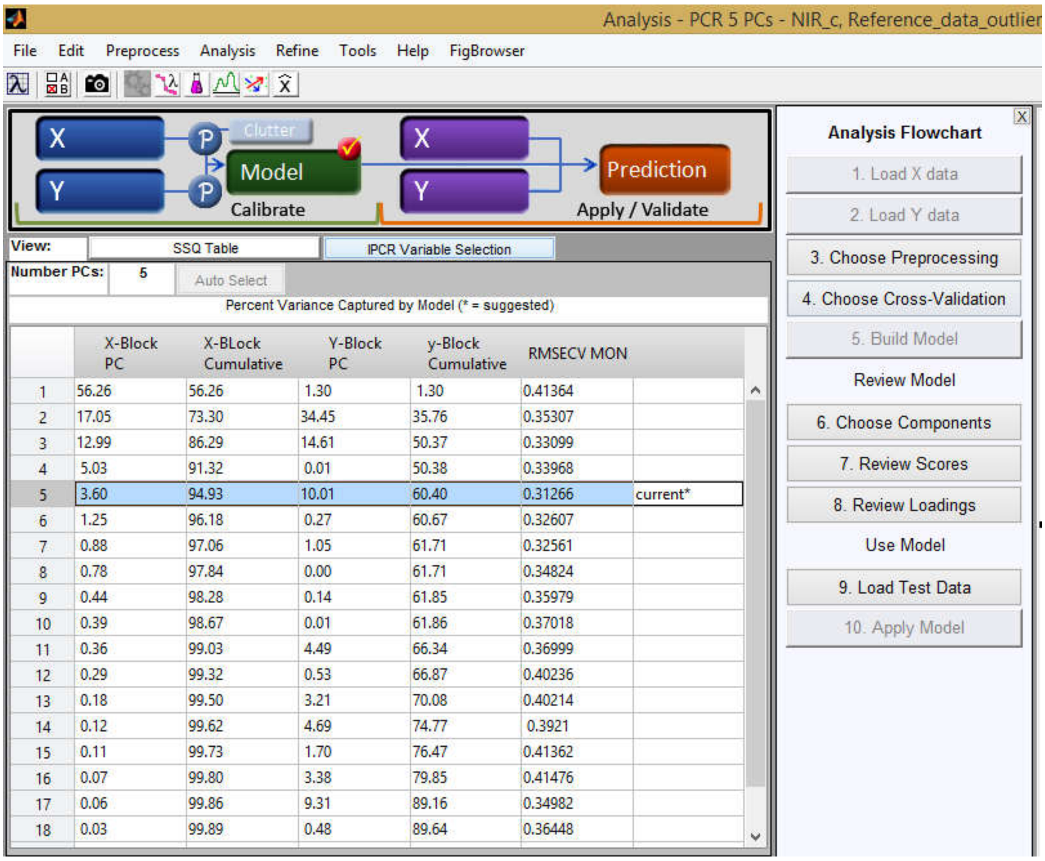
In addition, Figure 3.8 is also very informative in order to understand the NIR bands for pure normal linear hydrocarbons and C6 structural isomers [4, 14, 16, and 17].

6.4 Principal Component Regression – PCR

After pre-processing raw NIR spectra, PCR was applied to establish prediction models (using the PLS_Toolbox GUI given in Figure 6.7) for RON, MON, aromatics, olefins, benzene, density, and E150, individually.

The reason for applying the PCR to each of the measurement individually is the flexibility of choosing the number of PCs independently for each measurement. In order to determine the number of principal components, leave-one-out principle was followed. The PC vs RMSEC plot is helpful to determine the number of PCs. Samples 1 – 30 are used as calibration data set and samples 31 – 45 are used as validation data set, after removing the outlier from complete data set.

Some useful plots and a few critical parameters (RMSEC, RMSEP, R^2-calibration, and R^2-validation) will be shown in order to evaluate the performance of each prediction model. RMSEC and RMSEP values are compared with reproducibility, R, values of standard test methods. In addition, the predicted and reference values are given in corresponding tables for each property.



View: SSQ Table

Number PCs: 5

Percent Variance Captured by Model (* = suggested)

	X-Block PC	X-BLock Cumulative	Y-Block PC	y-Block Cumulative	RMSECV MON	
1	56.26	56.26	1.30	1.30	0.41364	
2	17.05	73.30	34.45	35.76	0.35307	
3	12.99	86.29	14.61	50.37	0.33099	
4	5.03	91.32	0.01	50.38	0.33968	
5	3.60	94.93	10.01	60.40	0.31266	current*
6	1.25	96.18	0.27	60.67	0.32607	
7	0.88	97.06	1.05	61.71	0.32561	
8	0.78	97.84	0.00	61.71	0.34824	
9	0.44	98.28	0.14	61.85	0.35979	
10	0.39	98.67	0.01	61.86	0.37018	
11	0.36	99.03	4.49	66.34	0.36999	
12	0.29	99.32	0.53	66.87	0.40236	
13	0.18	99.50	3.21	70.08	0.40214	
14	0.12	99.62	4.69	74.77	0.3921	
15	0.11	99.73	1.70	76.47	0.41362	
16	0.07	99.80	3.38	79.85	0.41476	
17	0.06	99.86	9.31	89.16	0.34982	
18	0.03	99.89	0.48	89.64	0.36448	

Figure 6.7: Graphical User Interface (GUI) for PCR in PLS Toolbox.

6.4.1 RON – Research Octane Number

According to Figure 6.8, the optimum number of principal components for 30 gasoline sample NIR spectra, that is X matrix in vector notation given in section 4.1.2.4, was determined as 6 PCs. As seen in Fig 6.8 a) and b), after the 6th PC, there is no more significant contribution to the variance from PC 7, 8 etc. 6 PCs are good enough to cover 96.18% of X data variance. Furthermore, the RMSECV (cross-validation error) values start to increase beyond 6 PCs.

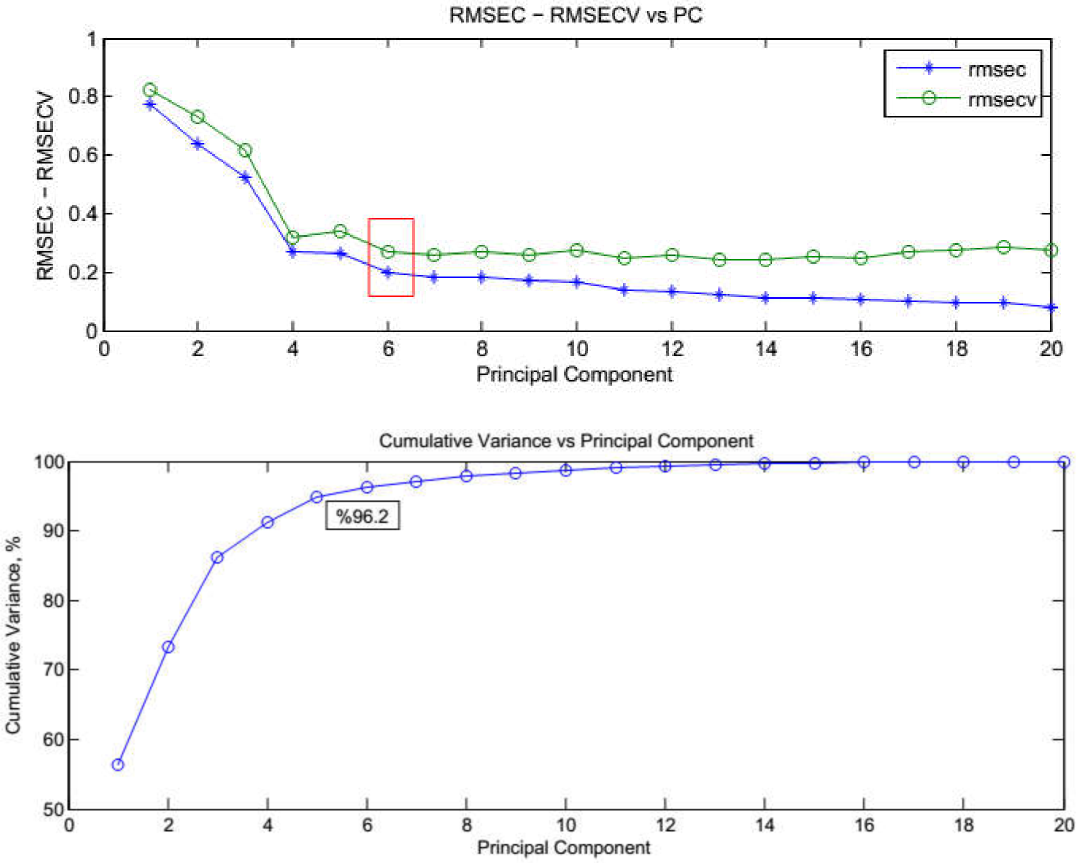


Figure 6.8: PCR-RON: **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance.

The reproducibility of TS EN ISO 5164 analysis is given as 0.7 in. Sample 34 has a residual value higher than this reproducibility value. The residuals for the all sample set are plotted in Figure 6.9.

In parallel to this, the residuals for the validation set are larger than the residuals of the calibration set, as expected. The measured, predicted and residual values are given in Table 6.1 for both the calibration and validation data sets.

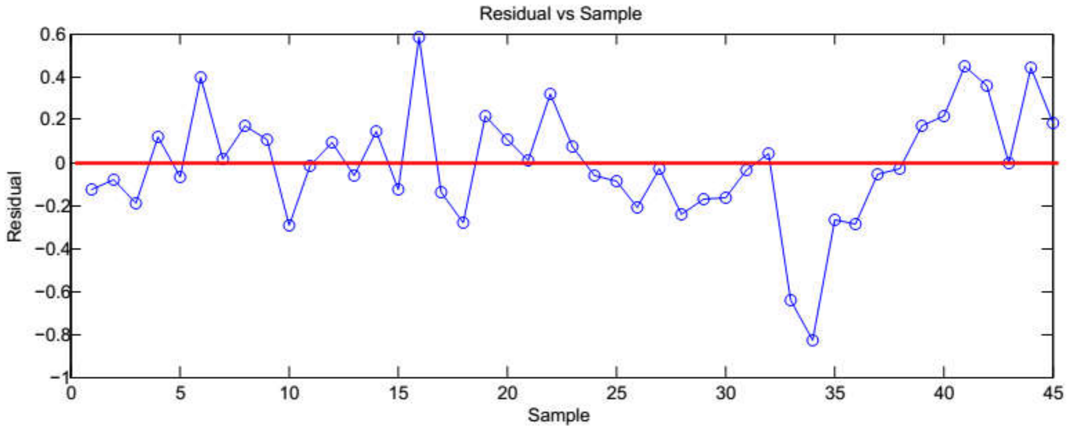


Figure 6.9: PCR-RON: Residuals vs All Sample Set.

The measured values are plotted against the predicted values for the all sample set, calibration set and validation set in Figure 6.10 a), b) and c). The R^2 value for the measured versus predicted regression line for complete sample set is 0.940.

Table 6.1: PCR-RON: Measured, Predicted results and Residuals.

Calibration Set (1-31)

Sample	Measured	Predicted	Residual
1	95.10	94.97	-0.13
2	95.20	95.12	-0.08
3	97.20	97.01	-0.19
4	95.20	95.32	0.12
5	95.20	95.13	-0.07
6	95.00	95.39	0.39
7	95.10	95.11	0.01
8	95.10	95.27	0.17
9	95.10	95.21	0.11
10	97.00	96.71	-0.29
11	95.30	95.29	-0.01
12	95.10	95.19	0.09
13	95.20	95.14	-0.06
14	95.20	95.34	0.14
15	97.40	97.28	-0.12
16	95.00	95.58	0.58
17	97.20	97.06	-0.14
18	95.20	94.92	-0.28
19	95.10	95.31	0.21
20	95.20	95.30	0.10
21	95.20	95.21	0.01
22	97.00	97.32	0.32
23	95.40	95.48	0.08
24	97.20	97.14	-0.06
25	95.20	95.11	-0.09
26	95.00	94.79	-0.21
27	95.10	95.07	-0.03
28	95.20	94.96	-0.24
29	95.10	94.93	-0.17
30	95.30	95.13	-0.17
31	95.10	95.06	-0.04

Validation Set (32-45)

32	95.10	95.15	0.05
33	97.20	96.56	-0.64
34	95.50	94.67	-0.83
35	95.10	94.83	-0.27
36	95.30	95.01	-0.29
37	94.90	94.85	-0.05
38	95.30	95.27	-0.03
39	95.20	95.37	0.17
40	95.20	95.42	0.22
41	95.10	95.55	0.45
42	97.20	97.56	0.36
43	95.40	95.39	-0.01
44	95.00	95.44	0.44
45	95.00	95.19	0.19

When the validation set was predicted by applying the model, RMSEP is 0.3552 and R^2 is 0.765, that is lower than R^2 for the calibration set. For the calibration set, RMSEC and R^2 are 0.1986 and 0.7654, correspondingly.

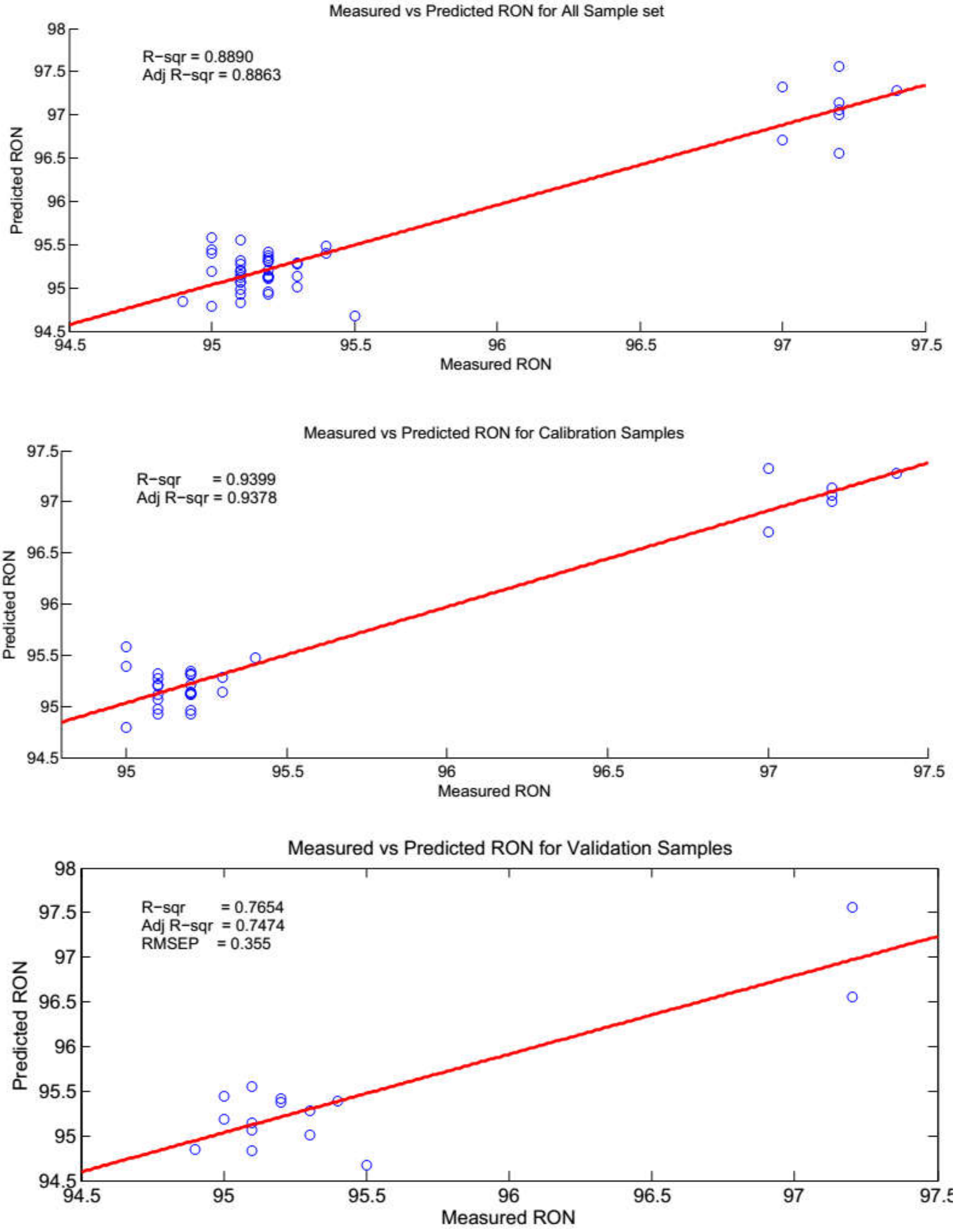


Figure 6.10: PCR-RON: Measured vs. predicted results for a) all sample set, b) calibration set and c) validation set.

6.4.2 MON – Motor Octane Number

As seen in Fig 6.11 a) and b), after the 5th PC, there is no more significant contribution to the variance from PC 6, 7 etc. 5 PCs are good enough to cover 94.93% of X data variance. Furthermore, the RMSECV (cross-validation error) values start to increase beyond 5 PCs.

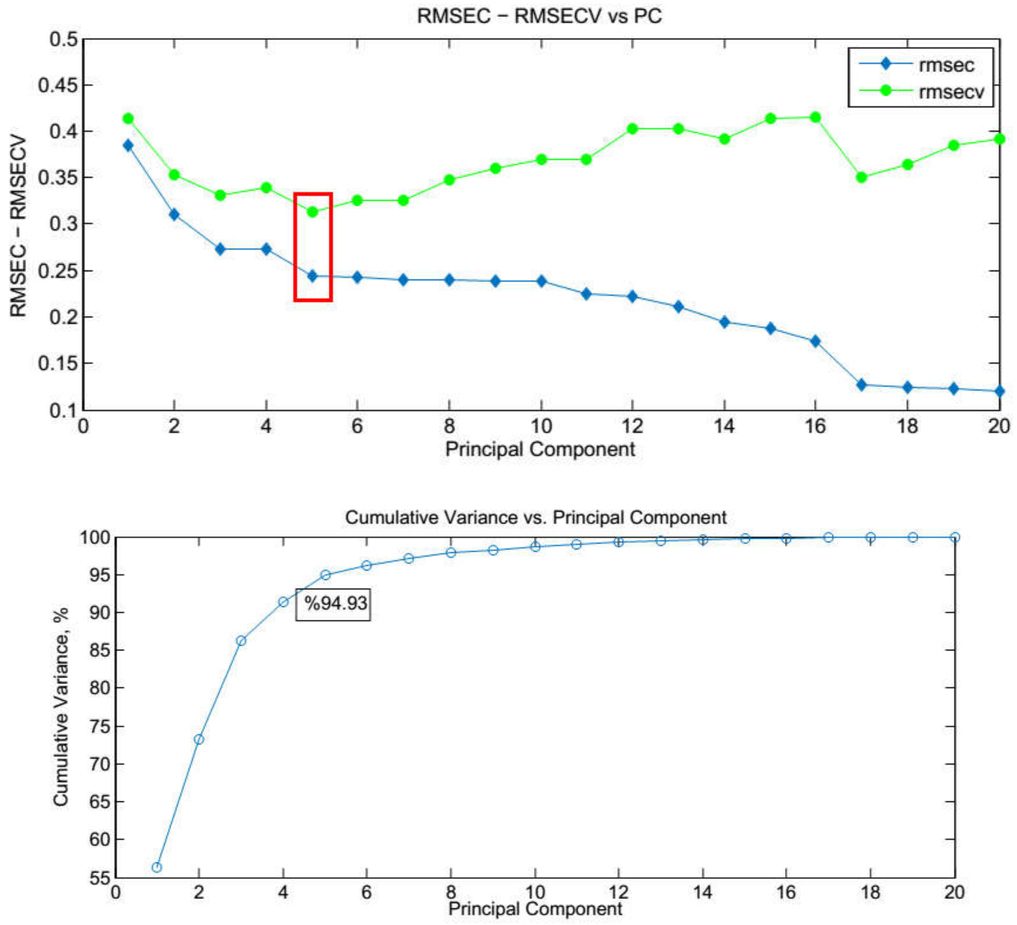


Figure 6.11: PCR-MON: **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance.

The reproducibility of TS EN ISO 5163 analysis is given as 0.9 in. There is not any sample has residual higher than this reproducibility value. The residuals for the all sample set are plotted in Figure 6.12.

In parallel to this, the residuals for the validation set are larger than residuals of calibration set, as expected. The measured, predicted and residual values are given in Table 6.2 for both the calibration and validation data sets.

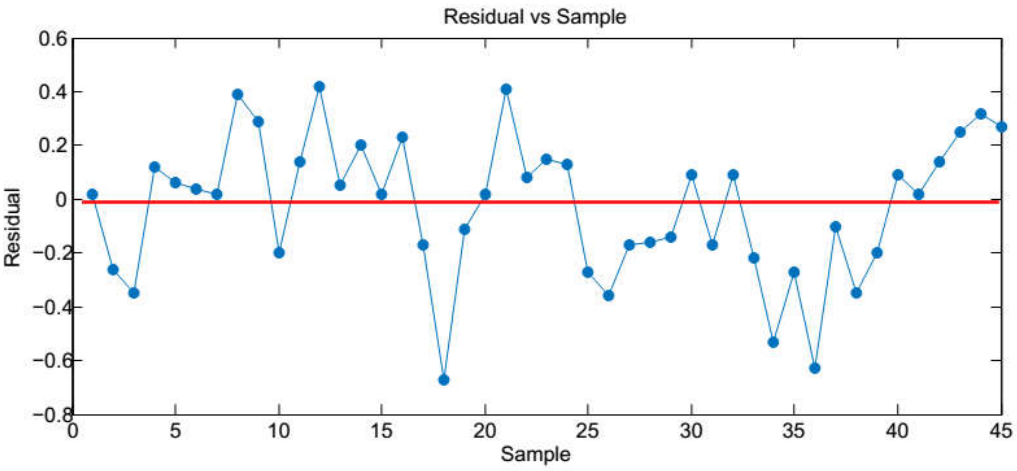


Figure 6.12: PCR-MON : Residuals vs. All Sample Set.

The measured values are plotted against the predicted values for the all sample set, calibration set and validation set in Figure 6.13 a), b) and c). The R^2 value for the measured versus predicted regression line for complete sample set is 0.5668.

Table 6.2: PCR-MON : Measured, Predicted results and Residuals.

Calibration Set (1-31)

Sample	Measured	Predicted	Residual
1	86.00	86.02	0.02
2	86.20	85.94	-0.26
3	86.20	85.85	-0.35
4	85.50	85.62	0.12
5	86.00	86.06	0.06
6	86.10	86.14	0.04
7	86.10	86.12	0.02
8	85.30	85.69	0.39
9	85.10	85.39	0.29
10	86.80	86.60	-0.20
11	85.30	85.44	0.14
12	85.30	85.72	0.42
13	85.90	85.95	0.05
14	85.80	86.00	0.20
15	86.30	86.32	0.02
16	85.60	85.83	0.23
17	86.30	86.13	-0.17
18	86.10	85.43	-0.67
19	86.30	86.19	-0.11
20	86.00	86.02	0.02
21	85.20	85.61	0.41
22	86.00	86.08	0.08
23	85.60	85.75	0.15
24	86.20	86.33	0.13
25	85.60	85.33	-0.27
26	86.20	85.84	-0.36
27	85.70	85.53	-0.17
28	85.80	85.64	-0.16
29	85.90	85.76	-0.14
30	85.90	85.99	0.09
31	85.90	85.73	-0.17

Validation Set (32-45)

Sample	Measured	Predicted	Residual
32	85.80	85.89	0.09
33	86.90	86.68	-0.22
34	86.40	85.87	-0.53
35	86.20	85.93	-0.27
36	85.80	85.17	-0.63
37	85.60	85.50	-0.10
38	86.10	85.75	-0.35
39	85.70	85.50	-0.20
40	85.60	85.69	0.09
41	85.50	85.52	0.02
42	86.20	86.34	0.14
43	85.60	85.85	0.25
44	85.30	85.62	0.32
45	85.60	85.87	0.27

When the validation set was predicted by applying the model, RMSEP is 0.29 and R^2 is 0.5267, that is lower than R^2 for the calibration set. For the calibration set, RMSEC and R^2 are 0.2436 and 0.6040, correspondingly.

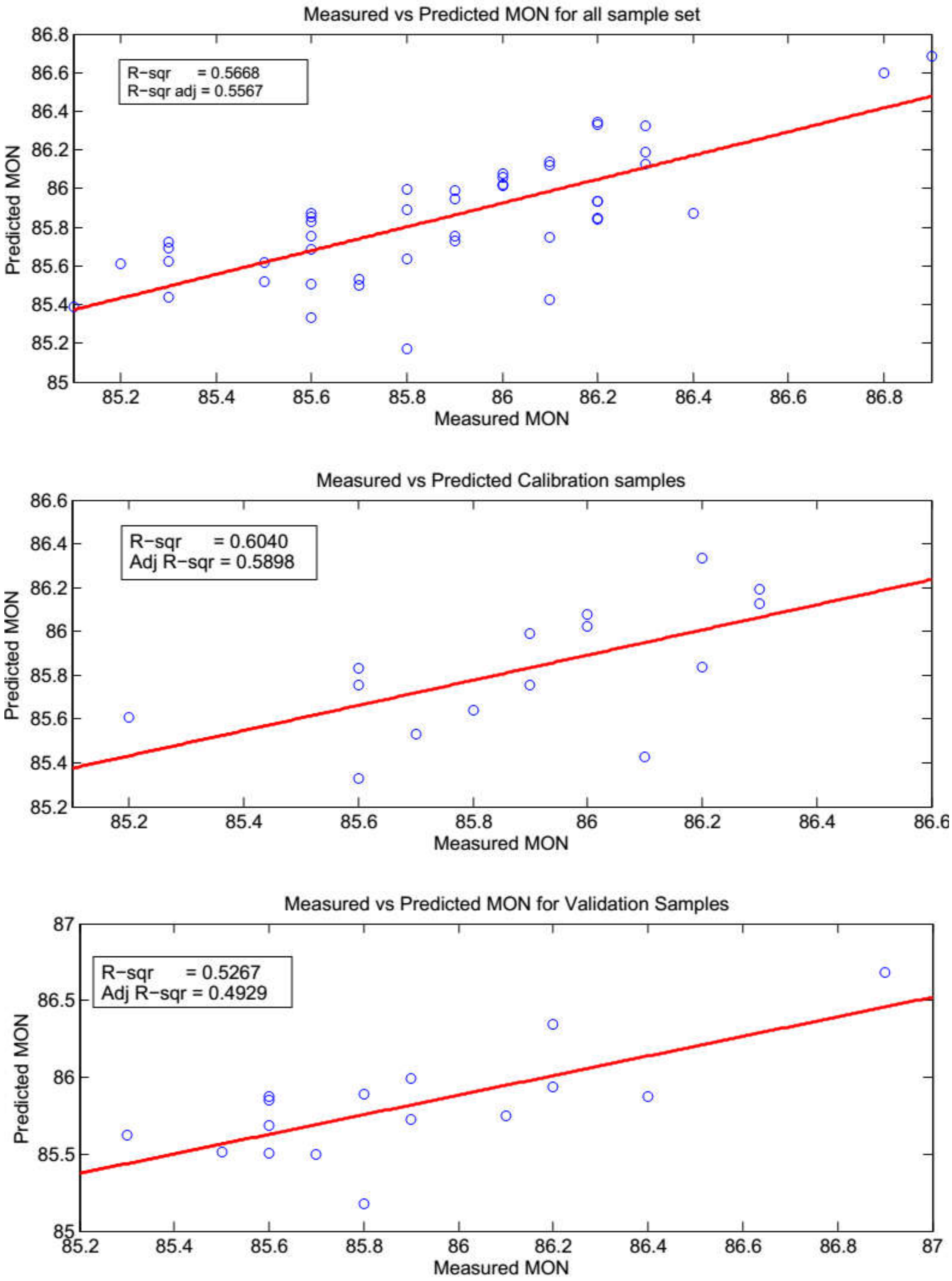


Figure 6.13: PCR-MON: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set.

6.4.3 Aromatics

As seen in Fig 6.14 a) and b), after the 6th PC, there is no more significant contribution to the variance from PC 7, 8 etc. 6 PCs are good enough to cover 96.18% of X data variance. Furthermore, the RMSECV (cross-validation error) values start to increase beyond 6 PCs.

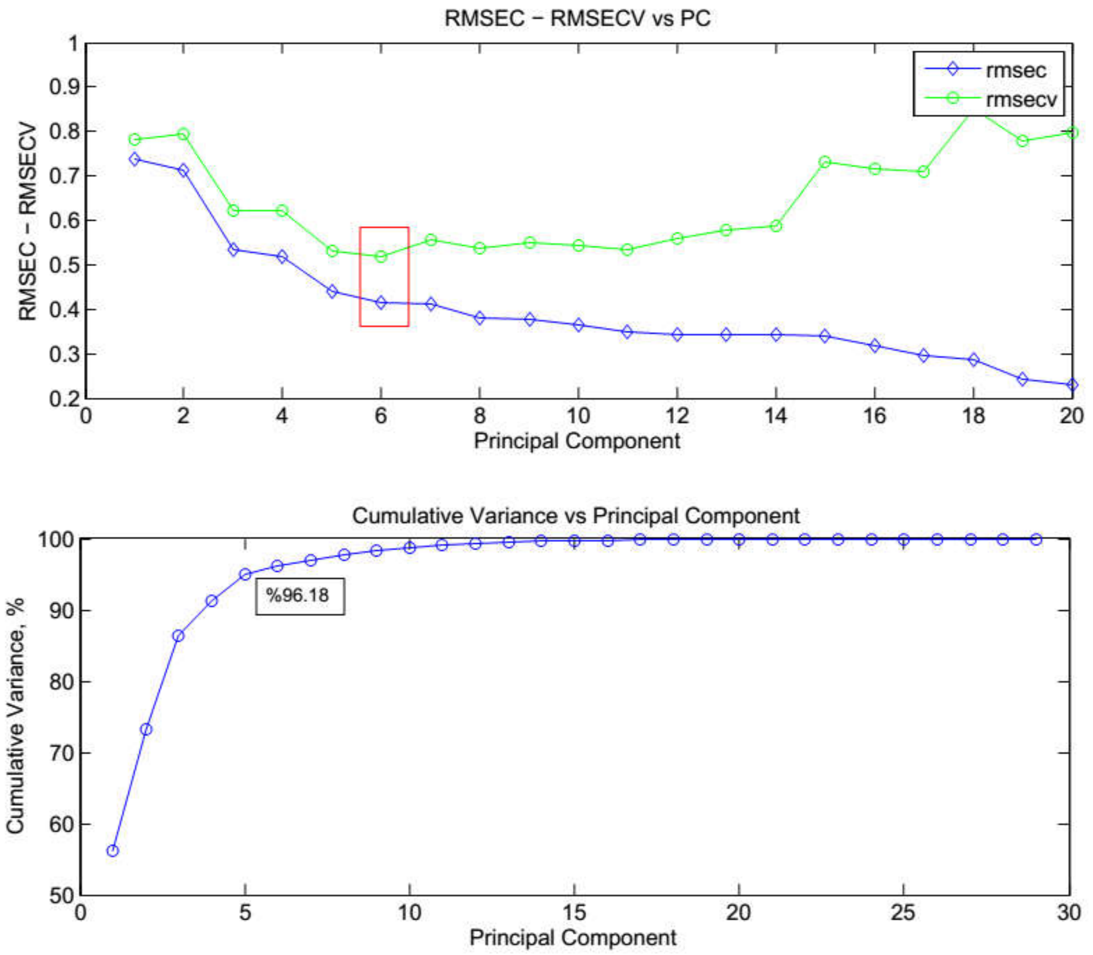


Figure 6.14: PCR-Aromatics a) PC vs. RMSEC-RMSECV, b) PC vs. X Cumulative variance.

The reproducibility, R, formula for Aromatics given in TS EN ISO 22854 is described by the below equation.

$$R = 0.045 \times X + 0.1384 \tag{6.6}$$

All samples have residuals smaller than calculated reproducibility value for test results. The residuals for the all sample set are plotted in Figure 6.15.

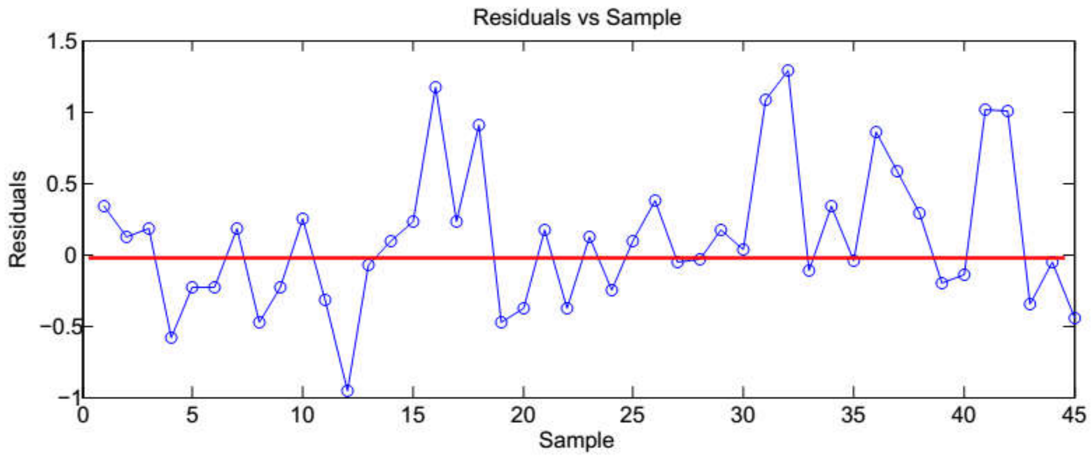


Figure 6.15: PCR-Aromatics: Residuals vs. All Sample Set.

In parallel to this, the residuals for the validation set are larger than the residuals of the calibration set. The measured, predicted and residual values are given in Table 6.3 for both the calibration and validation data sets.

Table 6.3: PCR-Aromatics: Measured, Predicted and Residuals, % v/v.

Calibration Set (1-31)

Sample	Measured	Predicted	Residual
1	33.20	33.54	0.34
2	29.70	29.83	0.13
3	34.80	34.98	0.18
4	38.30	37.72	-0.58
5	33.60	33.37	-0.23
6	33.70	33.47	-0.23
7	33.50	33.68	0.18
8	31.90	31.42	-0.48
9	37.90	37.66	-0.24
10	34.20	34.45	0.25
11	34.50	34.18	-0.32
12	34.40	33.44	-0.96
13	34.00	33.93	-0.07
14	33.60	33.69	0.09
15	34.30	34.53	0.23
16	33.00	34.17	1.17
17	33.80	34.03	0.23
18	33.60	34.51	0.91
19	32.70	32.22	-0.48
20	30.40	30.02	-0.38
21	32.90	33.07	0.17
22	33.80	33.42	-0.38
23	34.90	35.02	0.12
24	33.30	33.05	-0.25
25	35.90	36.00	0.10
26	30.00	30.38	0.38
27	37.00	36.95	-0.05
28	37.20	37.17	-0.03
29	30.90	31.07	0.17
30	30.90	30.94	0.04
31	31.70	32.79	1.09

Validation Set (32-45)

32	31.00	32.29	1.29
33	33.60	33.48	-0.12
34	32.10	32.44	0.34
35	31.00	30.96	-0.04
36	35.30	36.16	0.86
37	37.20	37.78	0.58
38	39.90	40.19	0.29
39	39.20	39.00	-0.20
40	32.70	32.56	-0.14
41	34.20	35.22	1.02
42	33.30	34.31	1.01
43	33.00	32.65	-0.35
44	33.10	33.05	-0.05
45	33.80	33.36	-0.44

The measured values are plotted against the predicted values for the all sample set, calibration set and validation set in Figure 6.16 a), b) and c). The R^2 value for the measured versus predicted regression line for complete sample set is 0.9538.

When the validation set was predicted by applying the model, RMSEP is 0.67 and R^2 is 0.9543, that is lower than R^2 for the calibration set. For the calibration set, RMSEC and R^2 are 0.4138 and 0.9615, correspondingly.

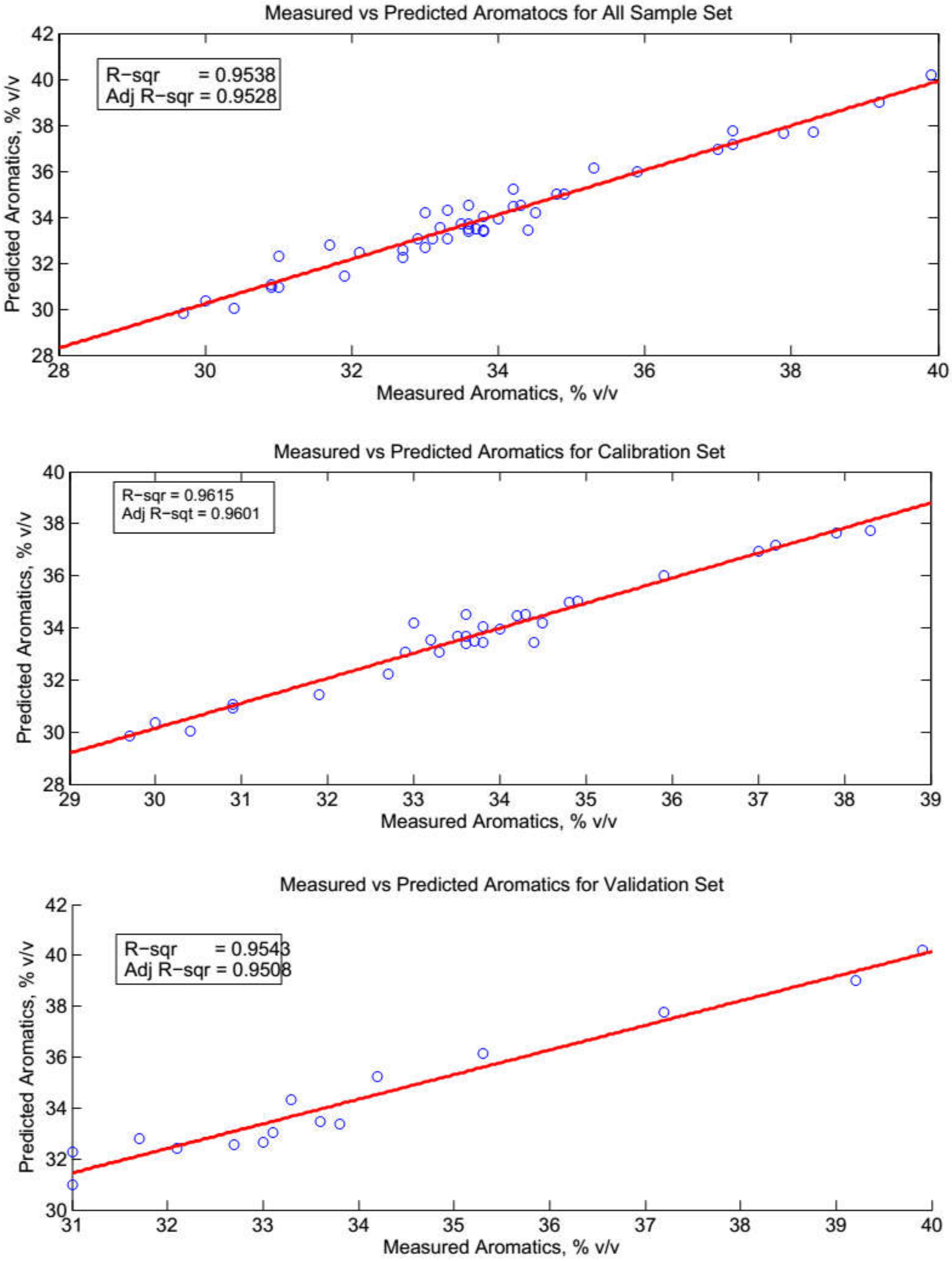


Figure 6.16: PCR-Aromatics: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set.

6.4.4 Olefins

As seen in Fig 6.17 a) and b), after the 6th PC, there is no more significant contribution to the variance from PC 7, 8 etc. 6 PCs are good enough to cover 96.8% of X data variance. Furthermore, there is not more contribution to the RMSECV (cross-validation error) values beyond 6 PCs.

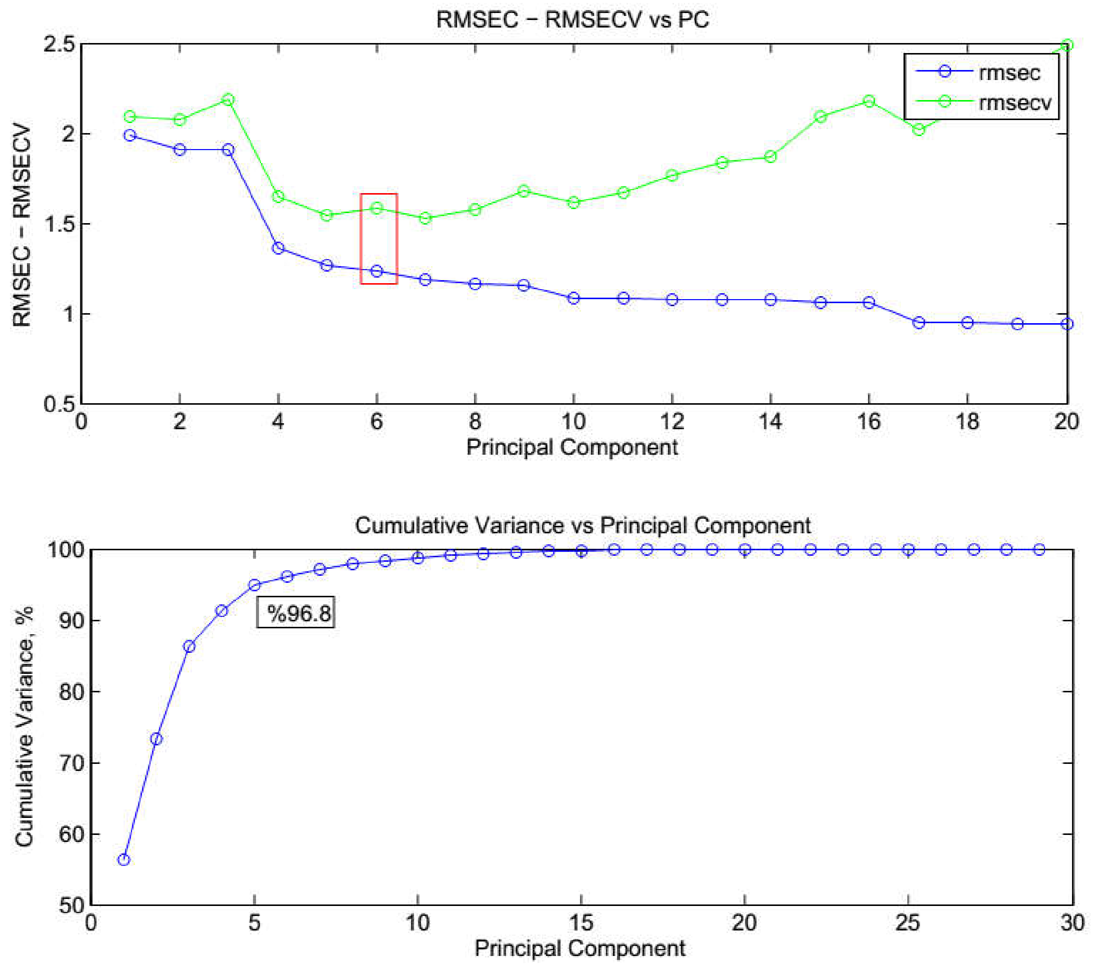


Figure 6.17: PCR-Olefins: **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance.

The reproducibility, R formula for Olefins given in TS EN ISO 22854 is described by the below equation.

$$R = 0.1176 \times X + 0.5118 \tag{6.7}$$

where X is measured test result. Samples having residuals higher than calculated reproducibility value are 1, 8, 12, 14, 18, 19, 21, 30, 35, 43 and 45. The residuals for the all sample set are plotted in Figure 6.18.

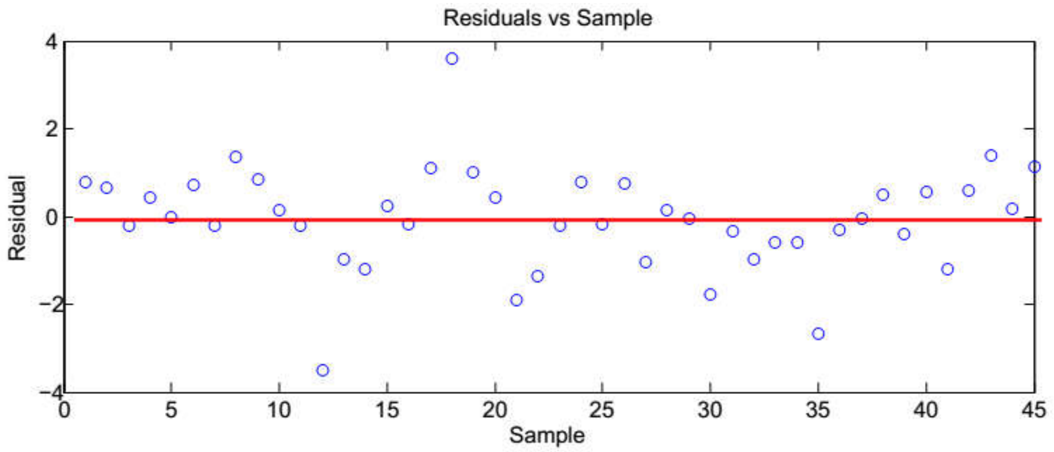


Figure 6.18: PCR-Olefins: Residuals vs. All Sample Set.

The measured, predicted and residual values are given in Table 6.4 for both the calibration and validation data sets.

Table 6.4: PCR-Olefins: Measured, Predicted and Residuals, % v/v.

Calibration Set (1-31)

Sample	Measured	Predicted	Residual
1	2.00	2.79	0.79
2	5.70	6.35	0.65
3	8.30	8.11	-0.19
4	3.80	4.22	0.42
5	4.20	4.18	-0.02
6	2.80	3.51	0.71
7	3.10	2.90	-0.20
8	6.00	7.37	1.37
9	4.80	5.65	0.85
10	2.80	2.94	0.14
11	8.10	7.88	-0.22
12	9.00	5.50	-3.50
13	5.80	4.81	-0.99
14	5.90	4.69	-1.21
15	5.40	5.65	0.25
16	5.60	5.43	-0.17
17	5.70	6.80	1.10
18	3.90	7.49	3.59
19	2.90	3.90	1.00
20	6.10	6.53	0.43
21	9.60	7.70	-1.90
22	9.80	8.45	-1.35
23	5.10	4.89	-0.21
24	5.40	6.19	0.79
25	7.10	6.94	-0.16
26	4.50	5.27	0.77
27	5.30	4.26	-1.04
28	3.60	3.74	0.14
29	5.80	5.74	-0.06
30	7.60	5.82	-1.78
31	6.20	5.87	-0.33

Validation Set (32-45)

32	6.30	5.32	-0.98
33	3.10	2.52	-0.58
34	5.10	4.52	-0.58
35	7.40	4.74	-2.66
36	7.90	7.59	-0.31
37	3.90	3.86	-0.04
38	1.20	1.69	0.49
39	4.90	4.50	-0.40
40	6.40	6.95	0.55
41	8.40	7.20	-1.20
42	5.40	6.00	0.60
43	4.20	5.59	1.39
44	7.20	7.36	0.16
45	3.70	4.84	1.14

The measured values are plotted against the predicted values are for the all sample set, calibration set and validation set in Figure 6.19 a), b) and c). The R^2 value for the measured versus predicted regression line for complete sample set is 0.6602.

When the validation set was predicted by applying the model, RMSEP is 0.99 and R^2 is 0.7409, that is higher than R^2 for the calibration set that is unusual. For the calibration set, RMSEC and R^2 are 1.23 and 0.6264, correspondingly.

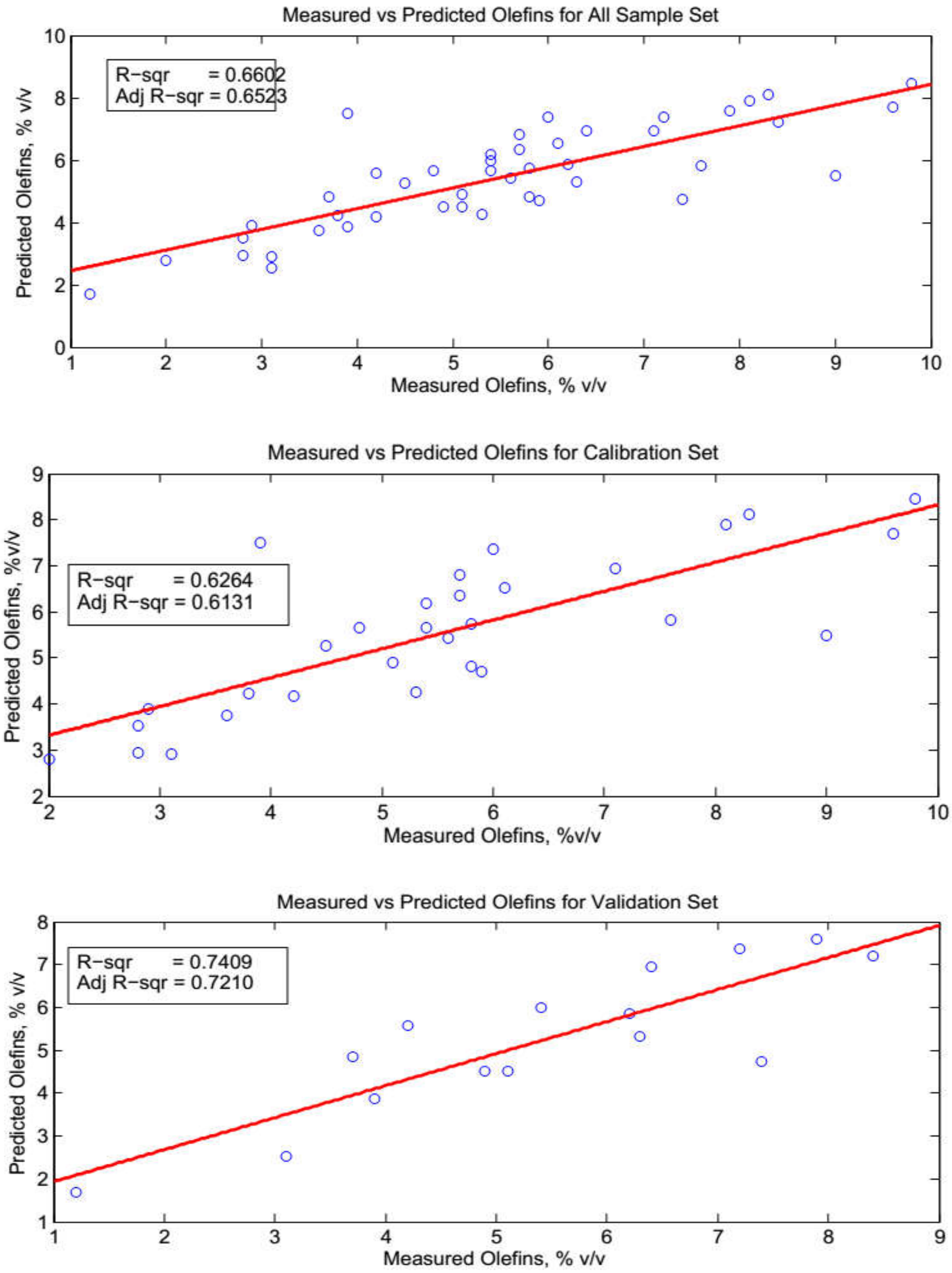


Figure 6.19: PCR-Olefins: Measured vs. predicted results for a) all sample set, b) calibration set and c) validation set.

6.4.5 Benzene

As seen in Fig 6.20 a) and b), after the 6th PC, there is no more significant contribution to the variance from PC 7, 8 etc. 6 PCs are good enough to cover 96.18% of X data variance. Furthermore, the RMSECV (cross-validation error) values start to increase beyond 6 PCs.

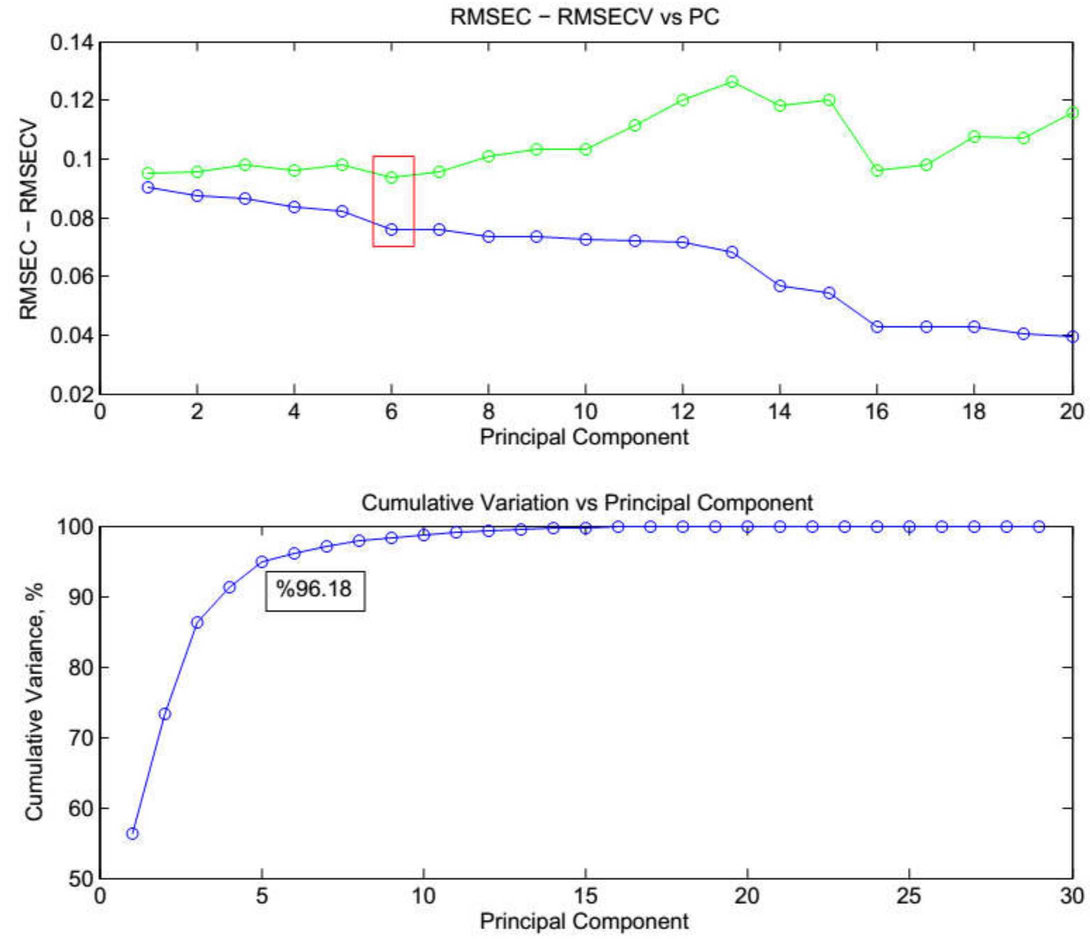


Figure 6.20: PCR-Benzene: **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance.

The Reproducibility, R formula for Benzene given in TS EN ISO 22854 is described by the below equation.

$$R = 0.0777 \times X - 0.025 \tag{6.8}$$

where X is measured test result. Samples having residuals higher than calculated reproducibility value are 1, 8, 9, 11, 12, 15, 16, 18, 19, 20, 21, 22, 23, 25, 26, 30, 31, 32, 34, 35, 38, 39, 41, 43, 44 and 45. The residuals for the all sample set are plotted in Figure 6.21.

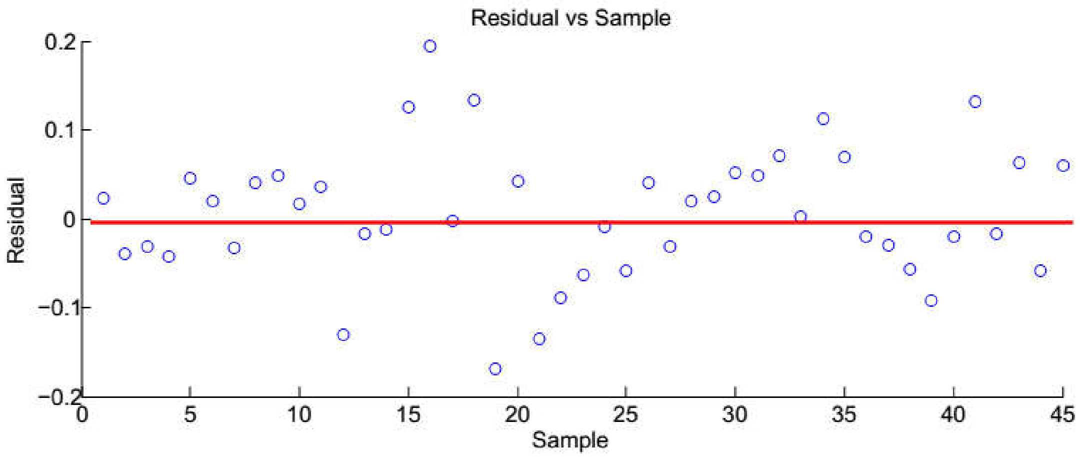


Figure 6.21: PCR-Benzene: Residuals vs. All Sample Set.

The measured, predicted and residual values are given in Table 6.5 for both the calibration and validation data sets.

Table 6.5: PCR-Benzene: Measured, Predicted and Residuals, % v/v.

Calibration Set (1-31)

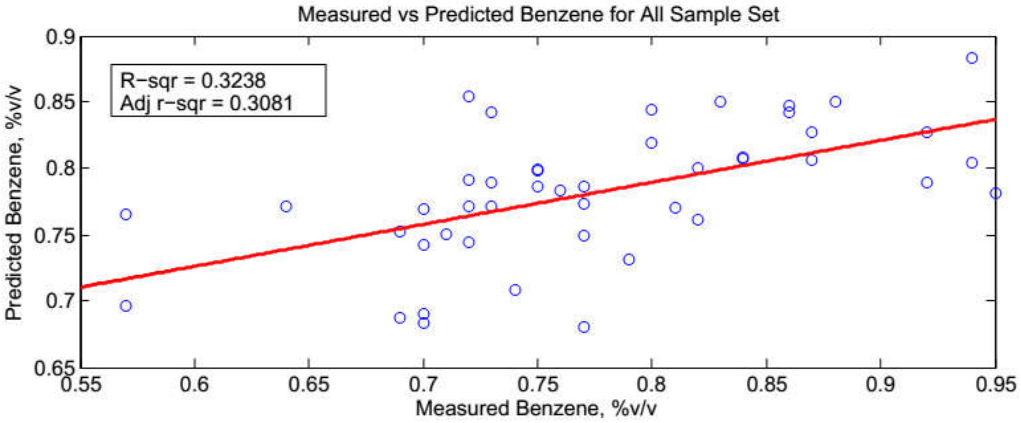
Sample	Measured	Predicted	Residual
1	0.76	0.78	0.02
2	0.81	0.77	-0.04
3	0.74	0.71	-0.03
4	0.87	0.83	-0.04
5	0.80	0.84	0.04
6	0.80	0.82	0.02
7	0.84	0.81	-0.03
8	0.73	0.77	0.04
9	0.75	0.80	0.05
10	0.77	0.79	0.02
11	0.75	0.79	0.04
12	0.92	0.79	-0.13
13	0.86	0.84	-0.02
14	0.86	0.85	-0.01
15	0.57	0.70	0.13
16	0.57	0.77	0.20
17	0.69	0.69	0.00
18	0.72	0.85	0.13
19	0.95	0.78	-0.17
20	0.70	0.74	0.04
21	0.94	0.80	-0.14
22	0.77	0.68	-0.09
23	0.87	0.81	-0.06
24	0.70	0.69	-0.01
25	0.82	0.76	-0.06
26	0.71	0.75	0.04
27	0.84	0.81	-0.03
28	0.83	0.85	0.02
29	0.72	0.74	0.02
30	0.72	0.77	0.05
31	0.75	0.80	0.05

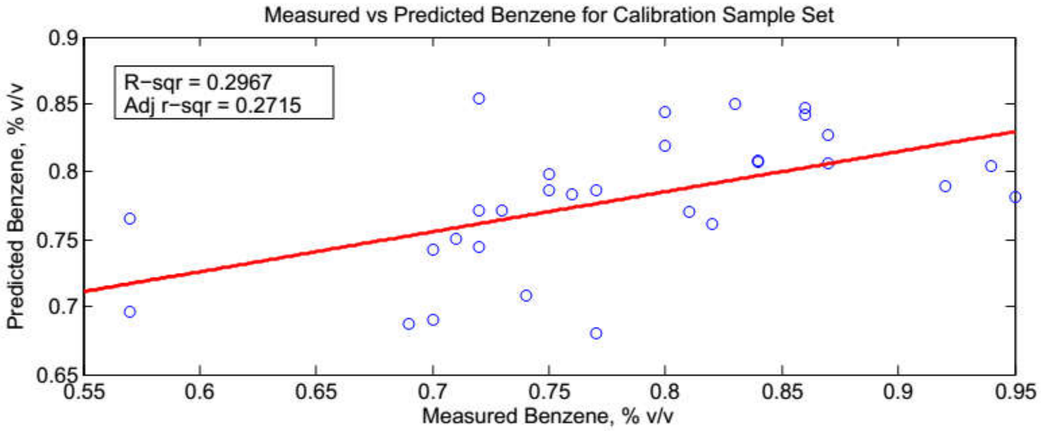
Validation Set (32-45)

Sample	Measured	Predicted	Residual
32	0.72	0.79	0.07
33	0.77	0.77	0.00
34	0.73	0.84	0.11
35	0.70	0.77	0.07
36	0.82	0.80	-0.02
37	0.88	0.85	-0.03
38	0.94	0.88	-0.06
39	0.92	0.83	-0.09
40	0.77	0.75	-0.02
41	0.64	0.77	0.13
42	0.70	0.68	-0.02
43	0.69	0.75	0.06
44	0.79	0.73	-0.06
45	0.73	0.79	0.06

The measured values are plotted against the predicted values are for the all sample set, calibration set and validation set in Figure 6.22 a), b) and c). The R^2 value for the measured versus predicted regression line for complete sample set is 0.3238.

When the validation set was predicted by applying the model, RMSEP is 0.067 and R^2 is 0.4156, that is higher than R^2 for the calibration set that is unusual. For the calibration set, RMSEC and R^2 are 0.076 and 0.2967, correspondingly.





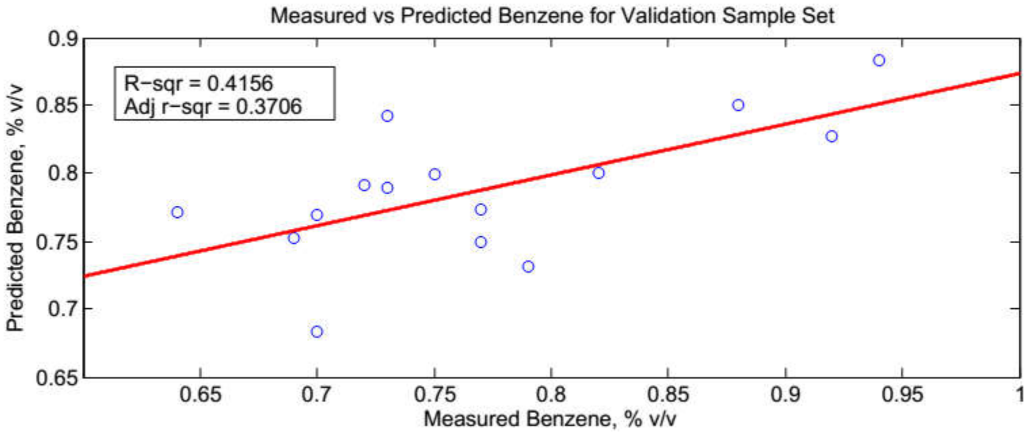


Figure 6.22: PCR-Benzene: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set.

6.4.6 Density

As seen in Fig 6.23 a) and b), after the 5th PC, there is no more significant contribution to the variance from PC 6, 7 etc. 5 PCs are good enough to cover 94.93% of X data variance. Furthermore, the RMSECV (cross-validation error) values start to increase beyond 5 PCs.

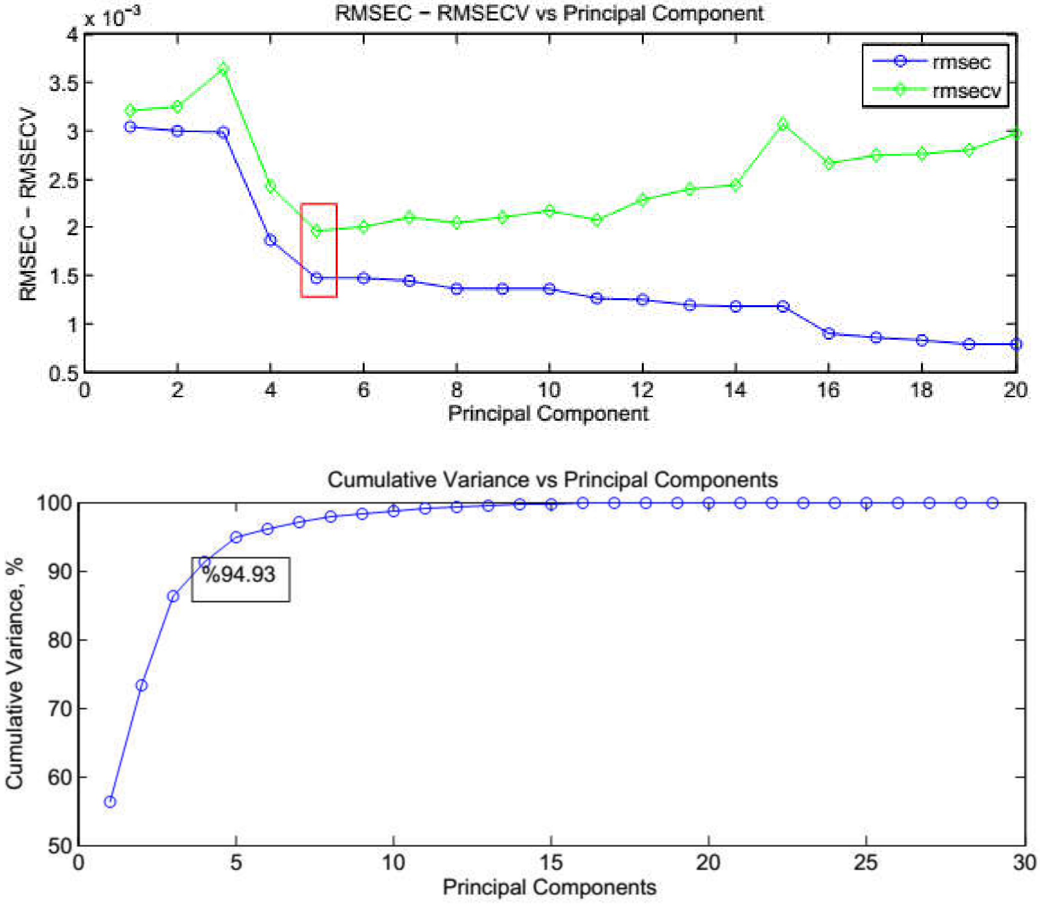


Figure 6.23: PCR-Density: **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance.

The Reproducibility, R, formula for Density given in ASTM D 4052 is described by the below equation.

$$R = 0.00195 - 0.0315 \times (X - 0.75) \qquad \textbf{(6.9)}$$

where X is measured test result. Samples having residuals higher than calculated reproducibility values are 1, 3, 8, 12, 18, 31, 32, 36, 40, 42 and 45. The residuals for the all sample set are plotted in Figure 6.24.

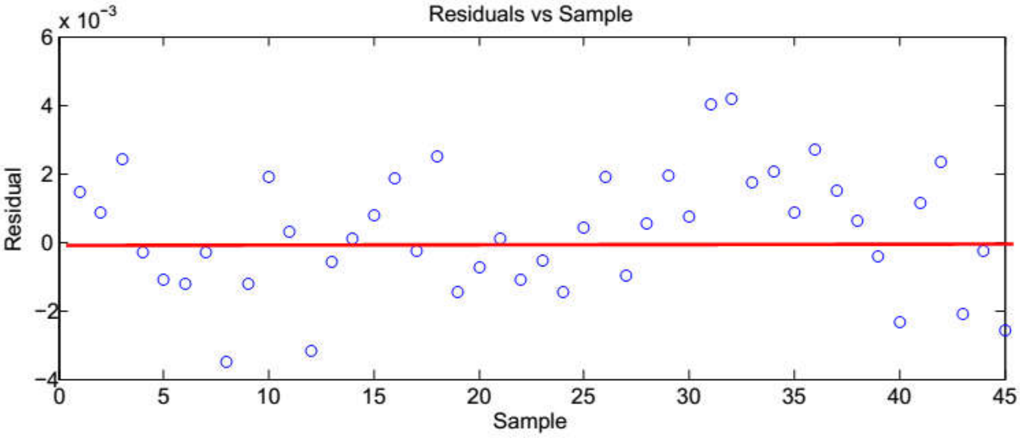


Figure 6.24: PCR-Density: Residuals vs. All Sample Set.

In parallel to this, the residuals for the validation set are larger than the residuals of the calibration set. The measured, predicted and residual values are given in Table 6.6 for both the calibration and validation data sets.

Table 6.6: PCR-Density: Measured, Predicted and Residuals, kg/l.

Calibration Set (1-31)

Sample	Measured	Predicted	Residual
1	0.7336	0.7351	0.0015
2	0.7311	0.7320	0.0009
3	0.7466	0.7490	0.0024
4	0.7462	0.7459	-0.0003
5	0.7387	0.7376	-0.0011
6	0.7391	0.7379	-0.0012
7	0.7372	0.7369	-0.0003
8	0.7395	0.7360	-0.0035
9	0.7473	0.7461	-0.0012
10	0.7405	0.7424	0.0019
11	0.7420	0.7423	0.0003
12	0.7419	0.7387	-0.0032
13	0.7398	0.7392	-0.0006
14	0.7394	0.7395	0.0001
15	0.7462	0.7470	0.0008
16	0.7386	0.7404	0.0018
17	0.7465	0.7462	-0.0003
18	0.7396	0.7421	0.0025
19	0.7360	0.7345	-0.0015
20	0.7337	0.7330	-0.0007
21	0.7398	0.7399	0.0001
22	0.7481	0.7470	-0.0011
23	0.7429	0.7424	-0.0005
24	0.7455	0.7440	-0.0015
25	0.7428	0.7432	0.0004
26	0.7291	0.7310	0.0019
27	0.7439	0.7429	-0.0010
28	0.7428	0.7434	0.0006
29	0.7307	0.7327	0.0020
30	0.7323	0.7331	0.0008
31	0.7332	0.7372	0.0040

Validation Set (32-45)

Sample	Measured	Predicted	Residual
32	0.7310	0.7352	0.0042
33	0.7382	0.7399	0.0017
34	0.7327	0.7348	0.0021
35	0.7308	0.7316	0.0008
36	0.7421	0.7448	0.0027
37	0.7432	0.7447	0.0015
38	0.7483	0.7489	0.0006
39	0.7491	0.7487	-0.0004
40	0.7407	0.7383	-0.0024
41	0.7430	0.7441	0.0011
42	0.7455	0.7478	0.0023
43	0.7392	0.7371	-0.0021
44	0.7399	0.7396	-0.0003
45	0.7400	0.7374	-0.0026

The measured values are plotted against the predicted values are for the all sample set, calibration set and validation set in Figure 6.25 a), b) and c). The R^2 value for the measured versus predicted regression line for complete sample set is 0.8940.

When the validation set was predicted by applying the model, RMSEP is 0.0023 and R^2 is 0.8674, that is lower than R^2 for the calibration set, as expected. For the calibration set, RMSEC and R^2 are 0.00147 and 0.9193 correspondingly.

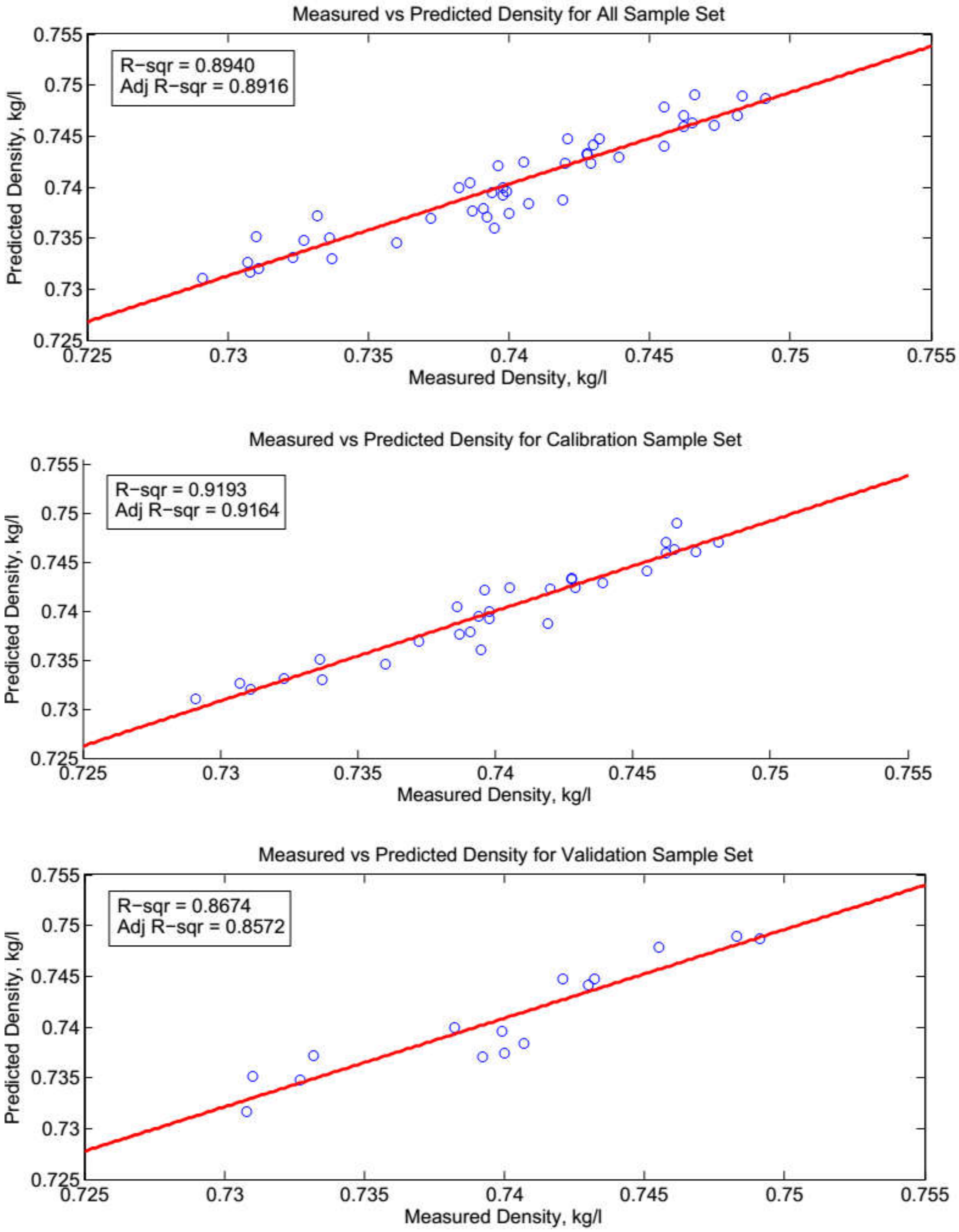


Figure 6.25: PCR-Density: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set.

6.4.7 E150 – Evaporated at 150 ºC

As seen in Fig 6.26 a) and b), after the 6th PC, there is no more significant contribution to the variance from PC 7, 8 etc. 6 PCs are good enough to cover 96.18% of X data variance. Furthermore, there is not significant contribution to the RMSECV (cross-validation error) values beyond 6 PCs.

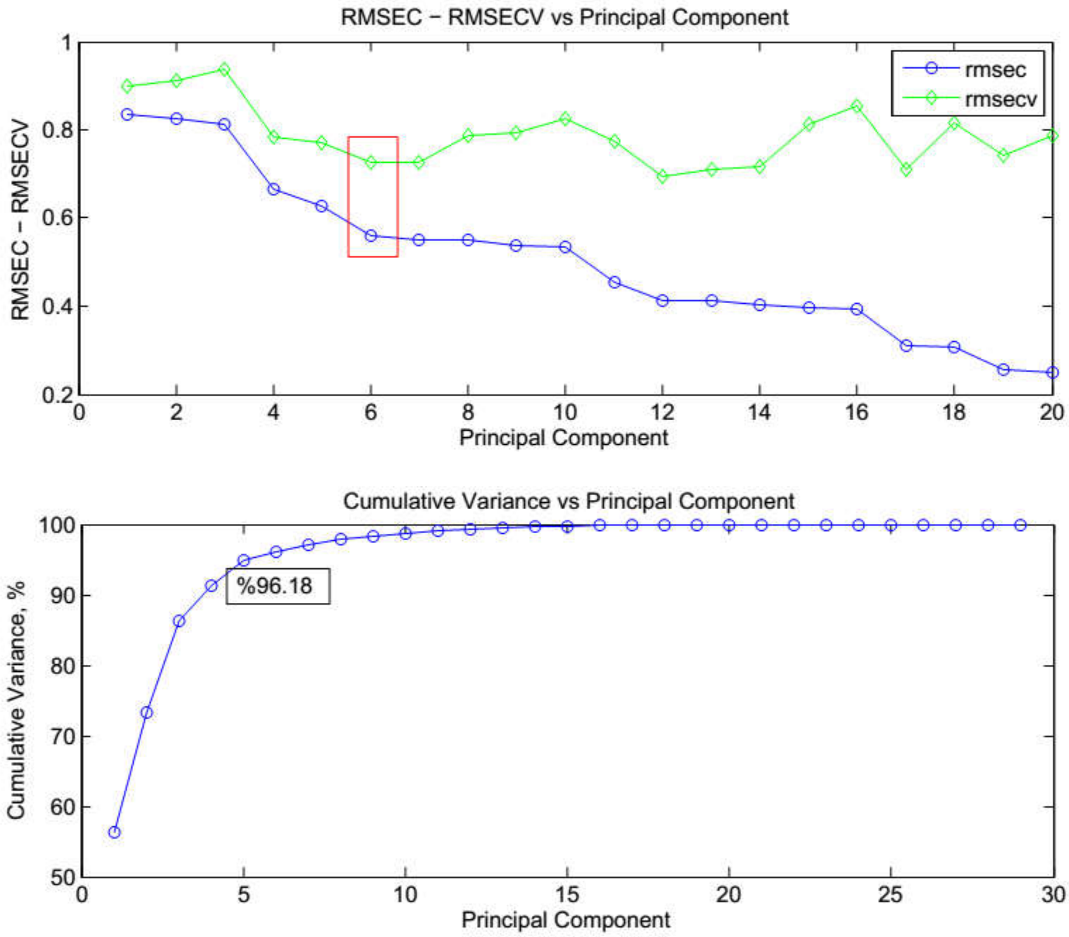


Figure 6.26: PCR-E150: **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance.

The reproducibility, R formula for E150 – Evaporated at 150 ºC given in ASTM D 86 is described by the below equation.

$$R = 0.02 \times (150 - X) \tag{6.10}$$

where X is measured test result. Samples having residuals higher than calculated reproducibility values are 12, 18, 34, 36 and 37. The residuals for the all sample set are plotted in Figure 6.27.

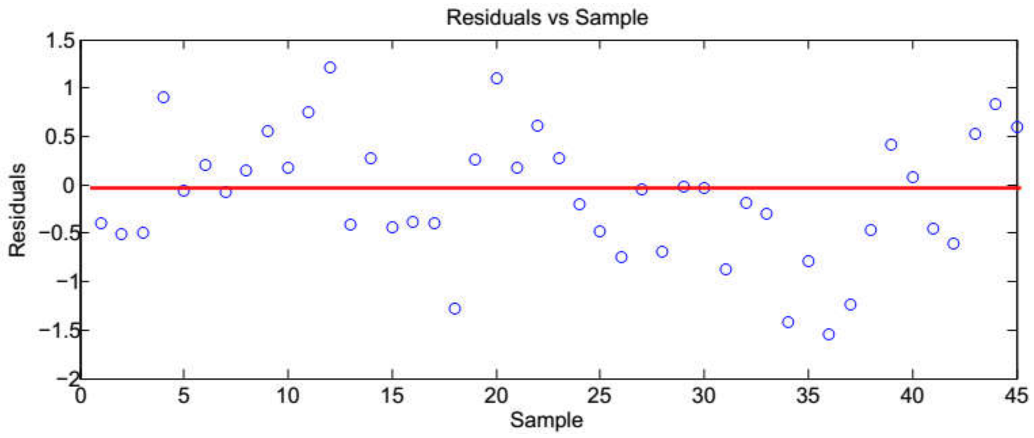


Figure 6.27: PCR-E150: Residuals vs. All Sample Set.

In parallel to this, the residuals for the validation set are larger than the residuals of the calibration set, as expected. The measured, predicted and residual values are given in Table 6.7 for both the calibration and validation data sets.

Table 6.7: PCR-E150: Measured, Predicted and Residuals, ^{o}C.

Calibration Set (1-31)

Sample	Measured	Predicted	Residual
1	92.70	92.31	-0.39
2	92.30	91.79	-0.51
3	90.50	90.00	-0.50
4	89.60	90.50	0.90
5	91.10	91.04	-0.06
6	91.20	91.41	0.21
7	91.40	91.33	-0.07
8	91.30	91.44	0.14
9	89.90	90.45	0.55
10	91.90	92.08	0.18
11	89.30	90.05	0.75
12	89.70	90.92	1.22
13	91.10	90.69	-0.41
14	90.50	90.78	0.28
15	91.10	90.67	-0.43
16	91.70	91.31	-0.39
17	90.90	90.50	-0.40
18	91.00	89.73	-1.27
19	92.00	92.26	0.26
20	90.40	91.50	1.10
21	90.10	90.28	0.18
22	89.80	90.41	0.61
23	90.30	90.57	0.27
24	91.20	91.00	-0.20
25	91.20	90.73	-0.47
26	92.70	91.95	-0.75
27	91.40	91.35	-0.05
28	91.50	90.81	-0.69
29	92.10	92.08	-0.02
30	92.10	92.07	-0.03

Validation Set (32-45)

Sample	Measured	Predicted	Residual
31	92.00	91.12	-0.88
32	92.30	92.12	-0.18
33	92.50	92.21	-0.29
34	92.60	91.17	-1.43
35	92.90	92.11	-0.79
36	91.40	89.86	-1.54
37	91.50	90.27	-1.23
38	90.80	90.33	-0.47
39	89.70	90.12	0.42
40	91.00	91.08	0.08
41	90.70	90.25	-0.45
42	91.20	90.59	-0.61
43	91.10	91.62	0.52
44	89.90	90.73	0.83
45	90.90	91.50	0.60

The measured values are plotted against the predicted values are for the all sample set, calibration set and validation set in Figure 6.28 a), b) and c). The R^2 value for the measured versus predicted regression line for complete sample set is 0.5146.

When the validation set was predicted by applying the model, RMSEP is 0.81 and R^2 is 0.4121, that is lower than R^2 for the calibration set, as expected. For the calibration set, RMSEC and R^2 are 0.56 and 0.6130 correspondingly.

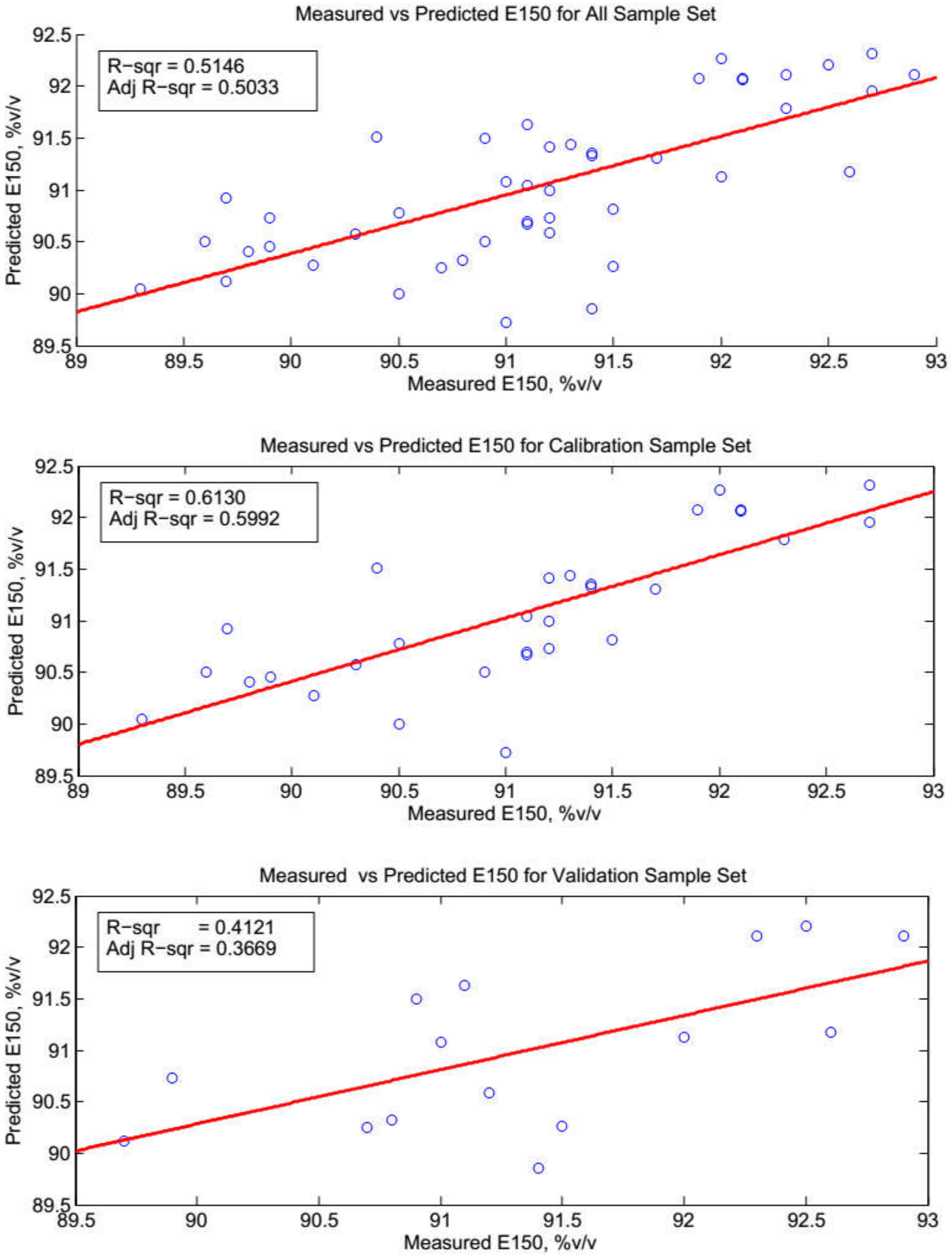
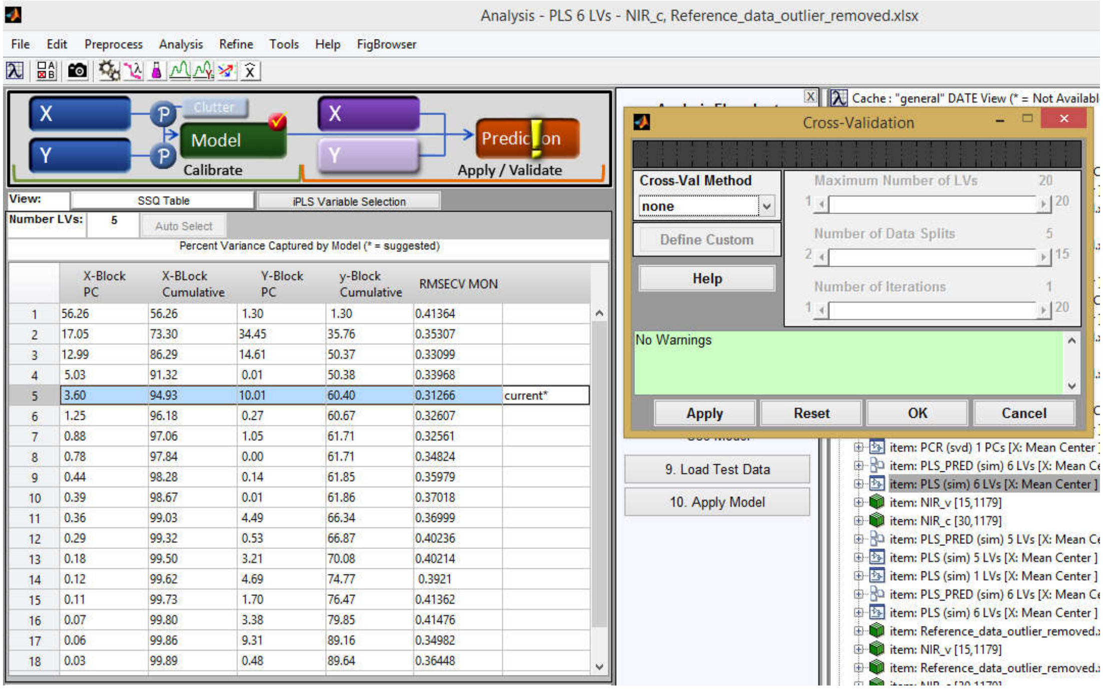


Figure 6.28: PCR-E150: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set.

6.5 Partial Least Squares - PLS

After pre-processing raw NIR spectra, Partial Least Squares (PLS) was applied to establish prediction models (using the PLS_Toolbox GUI given in Figure 6.29) for RON, MON, aromatics, olefins, benzene, density, and E150, individually.



Analysis - PLS 6 LVs - NIR_c, Reference_data_outlier_removed.xlsx

Number LVs: 5

Percent Variance Captured by Model (* = suggested)

	X-Block PC	X-BLock Cumulative	Y-Block PC	y-Block Cumulative	RMSECV MON	
1	56.26	56.26	1.30	1.30	0.41364	
2	17.05	73.30	34.45	35.76	0.35307	
3	12.99	86.29	14.61	50.37	0.33099	
4	5.03	91.32	0.01	50.38	0.33968	
5	3.60	94.93	10.01	60.40	0.31266	current*
6	1.25	96.18	0.27	60.67	0.32607	
7	0.88	97.06	1.05	61.71	0.32561	
8	0.78	97.84	0.00	61.71	0.34824	
9	0.44	98.28	0.14	61.85	0.35979	
10	0.39	98.67	0.01	61.86	0.37018	
11	0.36	99.03	4.49	66.34	0.36999	
12	0.29	99.32	0.53	66.87	0.40236	
13	0.18	99.50	3.21	70.08	0.40214	
14	0.12	99.62	4.69	74.77	0.3921	
15	0.11	99.73	1.70	76.47	0.41362	
16	0.07	99.80	3.38	79.85	0.41476	
17	0.06	99.86	9.31	89.16	0.34982	
18	0.03	99.89	0.48	89.64	0.36448	

Cross-Val Method: none

Maximum Number of LVs: 20

Number of Data Splits: 5

Number of Iterations: 1

No Warnings

Cache: "general" DATE View (* = Not Availabl

item: PCR (svd) 1 PCs [X: Mean Center

item: PLS_PRED (sim) 6 LVs [X: Mean Ce

item: PLS (sim) 6 LVs [X: Mean Center]

item: NIR_v [15,1179]

item: NIR_c [30,1179]

item: PLS_PRED (sim) 5 LVs [X: Mean Ce

item: PLS (sim) 5 LVs [X: Mean Center]

item: PLS (sim) 1 LVs [X: Mean Center]

item: PLS_PRED (sim) 6 LVs [X: Mean Ce

item: PLS (sim) 6 LVs [X: Mean Center]

item: Reference_data_outlier_removed.

item: NIR_v [15,1179]

item: Reference_data_outlier_removed.

Figure 6.29: PLS Toolbox GUI for PLS method.

In order to determine the optimum number of principal components to retain in the model, cross-validation with method of leave-one-out was followed, cross-validation is applied by using tools in the PLS Toolbox GUI.

The PC vs RMSEC plot is helpful to determine the number of PCs. Samples 1 – 30 are used as calibration data set and samples 31 – 45 are used as validation data set, after removing the outlier from complete data set.

Some useful plots and a few critical parameters (RMSEC, RMSEP, R^2-calibration, and R^2-validation) will be shown in order to evaluate the performance of each prediction model. RMSEC and RMSEP values are compared with reproducibility, R, values of standard test methods. In addition, the predicted and reference values are given in corresponding tables for each property.

6.5.1 RON – Research Octane Number

According to Figure 6.30, the optimum number of principal components for 30 gasoline sample NIR spectra, that is X matrix in vector notation given in section 4.1.2.4, was determined as 6 PCs. As seen in Fig 6.30 a) and b), after the 6th PC, there is no more significant contribution to the variance from PC 7, 8 etc. 6 PCs are good enough to cover 95.41% of X data variance. Furthermore, the RMSECV (cross-validation error) values start to increase beyond 6 PCs.

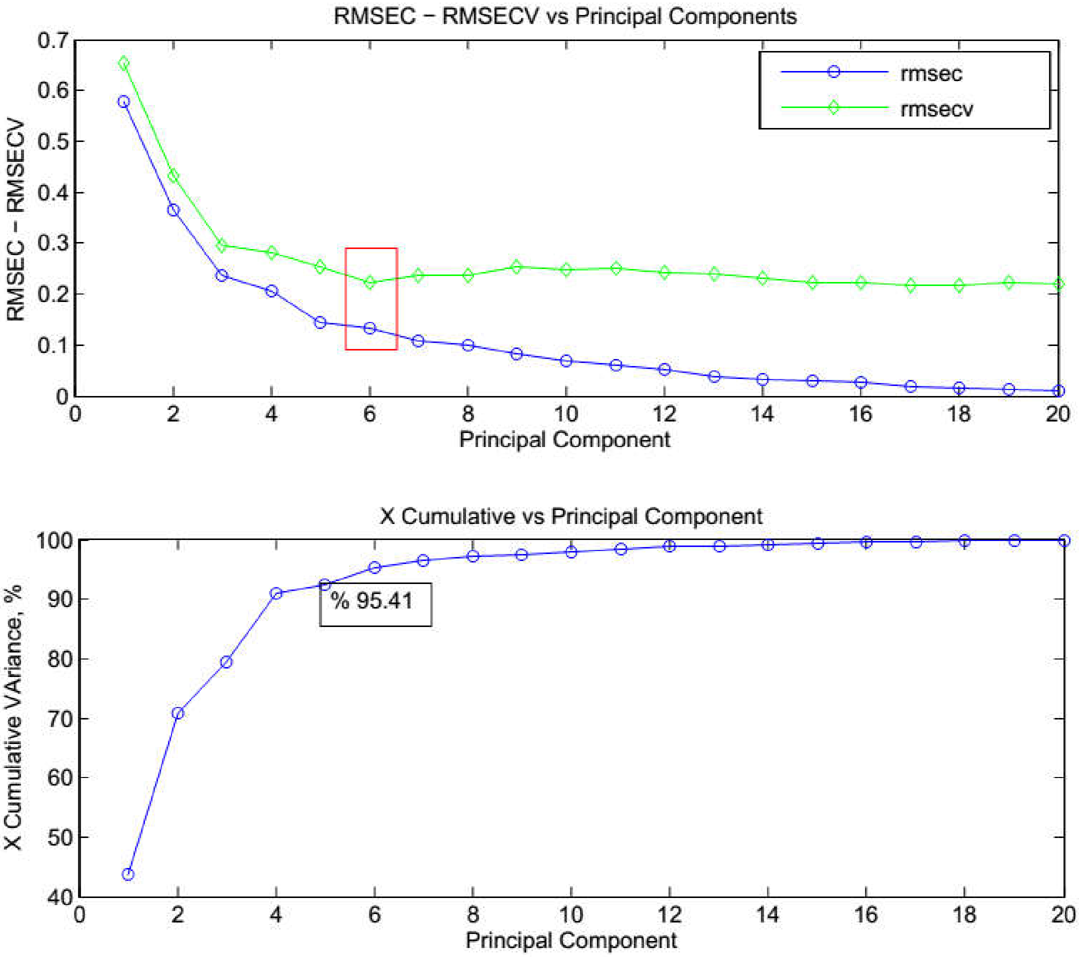


Figure 6.30: PLS-RON: **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance.

The reproducibility of TS EN ISO 5164 analysis is given as 0.7. None of the samples has residual value higher than this reproducibility value. The residuals for the all sample set are plotted in Figure 6.31.

In parallel to this, the residuals for the validation set are larger than the residuals of the calibration set, as expected. The measured, predicted and residual values are given in Table 6.8 for both the calibration and validation data sets.

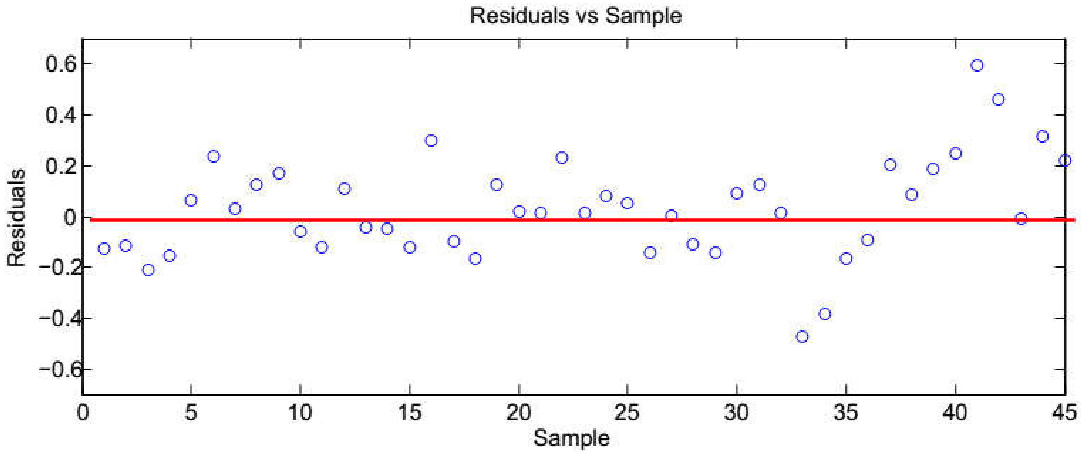


Figure 6.31: PLS-RON: Residuals vs All Sample Set.

The measured values are plotted against the predicted values are for the all sample set, calibration set and validation set in Figure 6.32 a), b) and c). The R^2 value for the measured versus predicted regression line for complete sample set is 0.9354.

Table 6.8: PLS-RON: Measured, Predicted results and Residuals.

Calibration Set (1-31)

Sample	Measured	Predicted	Residual
1	95.10	94.97	-0.13
2	95.20	95.08	-0.12
3	97.20	96.99	-0.21
4	95.20	95.05	-0.15
5	95.20	95.26	0.06
6	95.00	95.24	0.24
7	95.10	95.13	0.03
8	95.10	95.22	0.12
9	95.10	95.27	0.17
10	97.00	96.94	-0.06
11	95.30	95.18	-0.12
12	95.10	95.21	0.11
13	95.20	95.16	-0.04
14	95.20	95.15	-0.05
15	97.40	97.28	-0.12
16	95.00	95.30	0.30
17	97.20	97.10	-0.10
18	95.20	95.03	-0.17
19	95.10	95.23	0.13
20	95.20	95.22	0.02
21	95.20	95.21	0.01
22	97.00	97.23	0.23
23	95.40	95.41	0.01
24	97.20	97.28	0.08
25	95.20	95.26	0.06
26	95.00	94.85	-0.15
27	95.10	95.10	0.00
28	95.20	95.09	-0.11
29	95.10	94.96	-0.14
30	95.30	95.39	0.09
31	95.10	95.23	0.13

Validation Set (32-45)

32	95.10	95.11	0.01
33	97.20	96.72	-0.48
34	95.50	95.11	-0.39
35	95.10	94.93	-0.17
36	95.30	95.21	-0.09
37	94.90	95.11	0.21
38	95.30	95.39	0.09
39	95.20	95.39	0.19
40	95.20	95.45	0.25
41	95.10	95.70	0.60
42	97.20	97.66	0.46
43	95.40	95.39	-0.01
44	95.00	95.32	0.32
45	95.00	95.22	0.22

When the validation set was predicted by applying the model, RMSEP is 0.29 and R^2 is 0.8475, that is lower than R^2 for the calibration set, as expected. For the calibration set, RMSEC and R^2 are 0.13 and 0.9736 correspondingly.

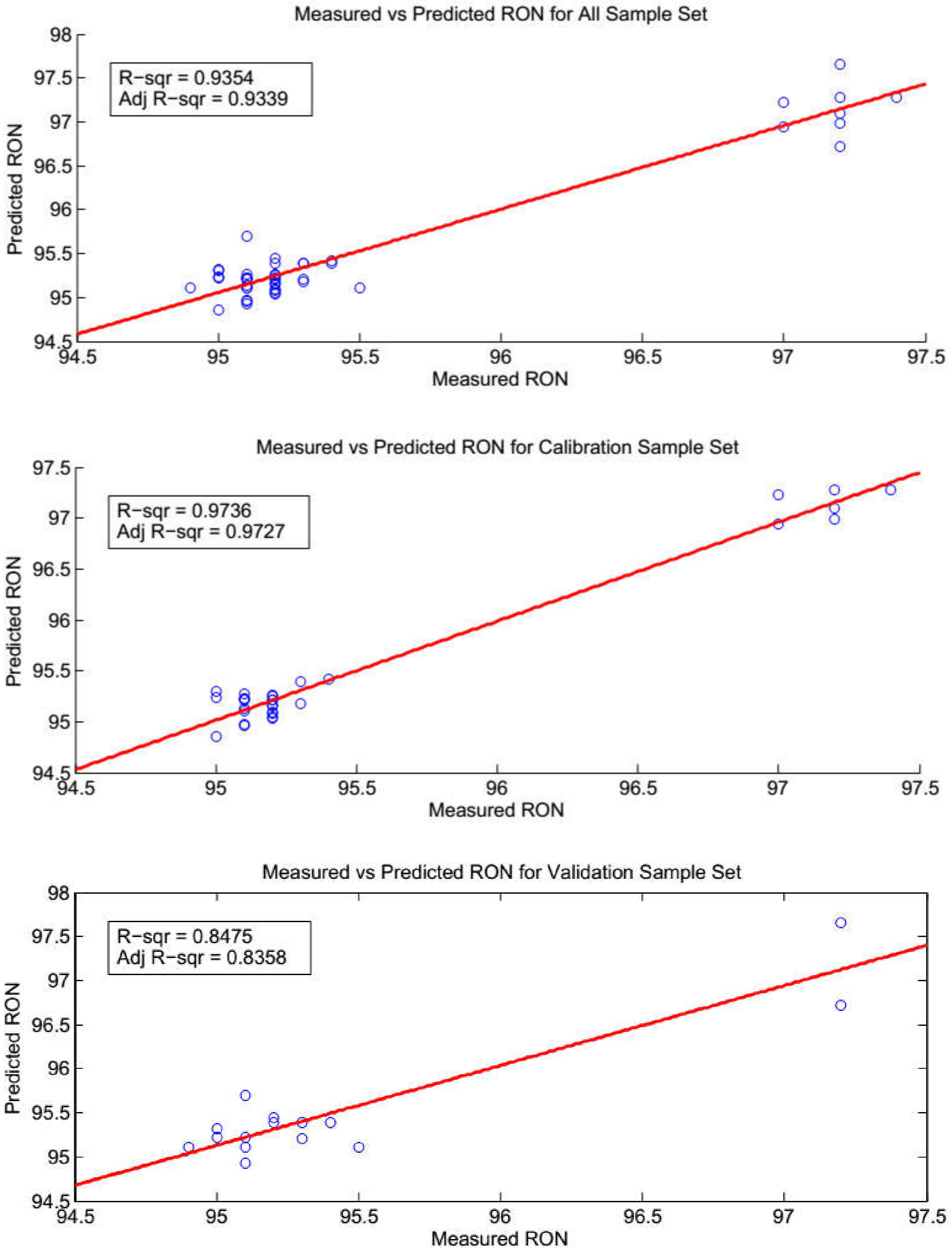


Figure 6.32: PLS-RON: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set.

6.5.2 MON – Motor Octane Number

As seen in Fig 6.33 a) and b), after the 5th PC, there is no more significant contribution to the variance from PC 6, 7 etc. 5 PCs are good enough to cover 90.49% of X data variance. Furthermore, the RMSECV (cross-validation error) values start to increase beyond 5 PCs.

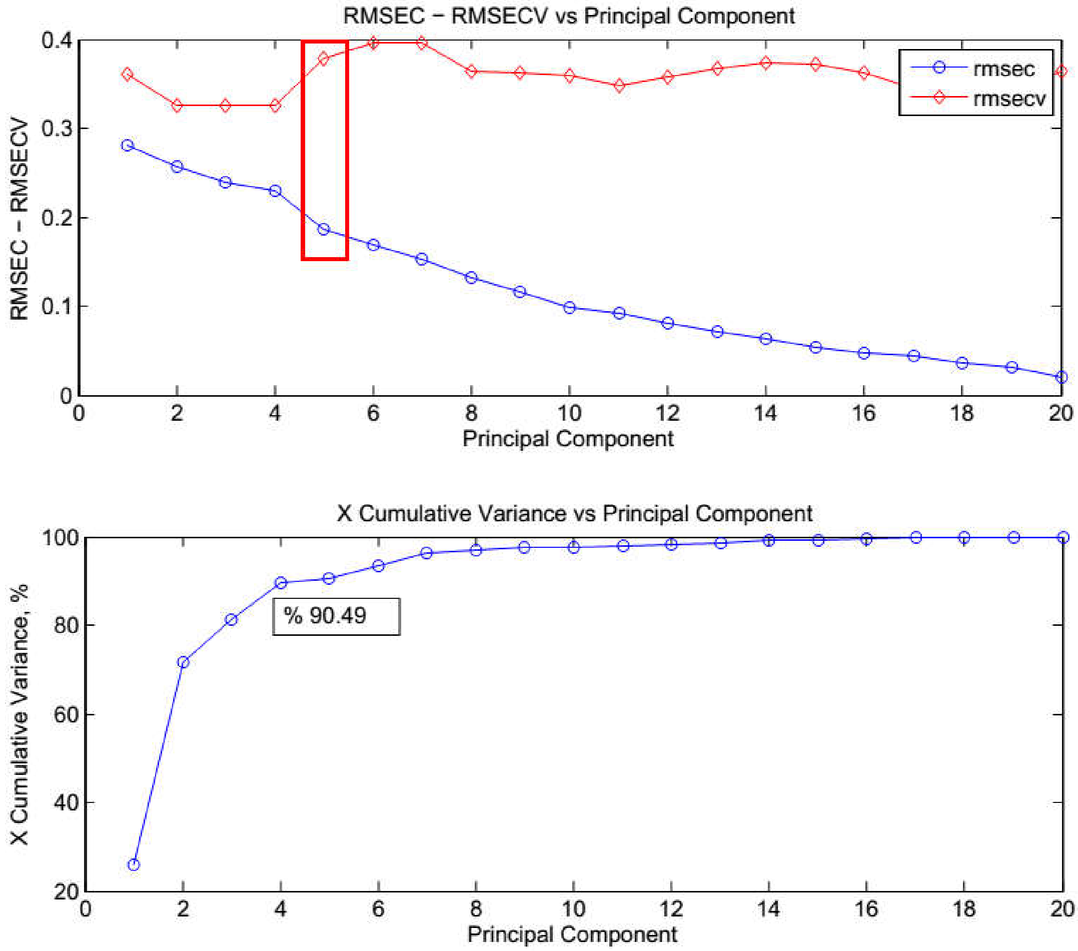


Figure 6.33: PLS-MON **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. Cumulative variance.

The reproducibility of TS EN ISO 5163 analysis is given as 0.9. There is not any sample has residual higher than this reproducibility value. The residuals for the all sample set are plotted in Figure 6.34.

In parallel to this, the residuals for the validation set are larger than the residuals of the calibration set, as expected. The measured, predicted and residual values are given in Table 6.9 for both the calibration and validation data sets.

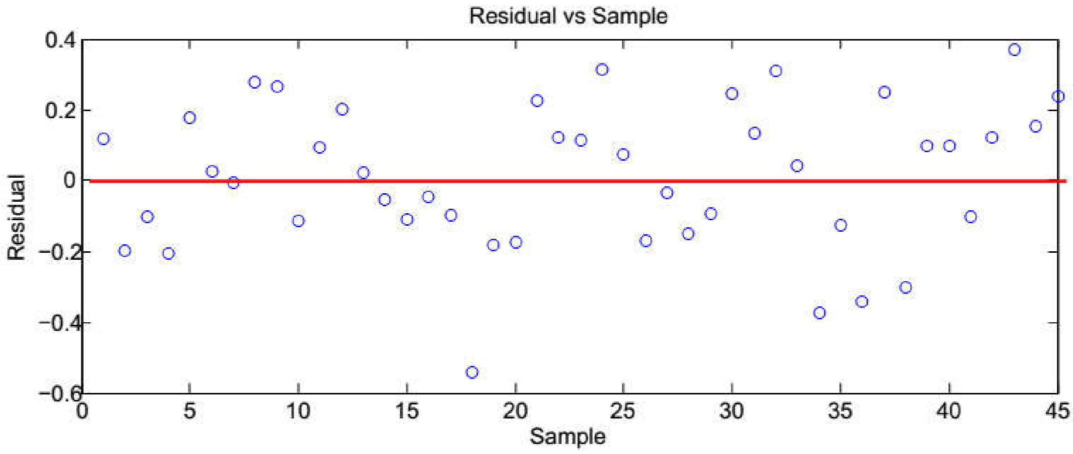


Figure 6.34: PLS-MON: Residuals vs All Sample Set.

The measured values are plotted against the predicted values are for the all sample set, calibration set and validation set in Figure 6.35 a), b) and c). The R^2 value for the measured versus predicted regression line for complete sample set is 0.7366.

Table 6.9: PLS-MON: Measured, Predicted results and Residuals.

Calibration Set (1-31)

Sample	Measured	Predicted	Residual
1	86.00	86.12	0.12
2	86.20	86.00	-0.20
3	86.20	86.10	-0.10
4	85.50	85.30	-0.20
5	86.00	86.18	0.18
6	86.10	86.13	0.03
7	86.10	86.09	-0.01
8	85.30	85.58	0.28
9	85.10	85.37	0.27
10	86.80	86.68	-0.12
11	85.30	85.40	0.10
12	85.30	85.50	0.20
13	85.90	85.92	0.02
14	85.80	85.75	-0.05
15	86.30	86.19	-0.11
16	85.60	85.55	-0.05
17	86.30	86.20	-0.10
18	86.10	85.56	-0.54
19	86.30	86.12	-0.18
20	86.00	85.83	-0.17
21	85.20	85.43	0.23
22	86.00	86.12	0.12
23	85.60	85.71	0.11
24	86.20	86.51	0.31
25	85.60	85.67	0.07
26	86.20	86.03	-0.17
27	85.70	85.66	-0.04
28	85.80	85.65	-0.15
29	85.90	85.81	-0.09
30	85.90	86.15	0.25
31	85.90	86.03	0.13

Validation Set (32-45)

32	85.80	86.11	0.31
33	86.90	86.94	0.04
34	86.40	86.03	-0.37
35	86.20	86.07	-0.13
36	85.80	85.46	-0.34
37	85.60	85.85	0.25
38	86.10	85.80	-0.30
39	85.70	85.80	0.10
40	85.60	85.70	0.10
41	85.50	85.40	-0.10
42	86.20	86.32	0.12
43	85.60	85.97	0.37
44	85.30	85.45	0.15
45	85.60	85.84	0.24

When the validation set was predicted by applying the model, RMSEP is 0.23 and R^2 is 0.6862, that is lower than R^2 for the calibration set as expected. For the calibration set, RMSEC and R^2 are 0.19 and 0.7958 correspondingly.

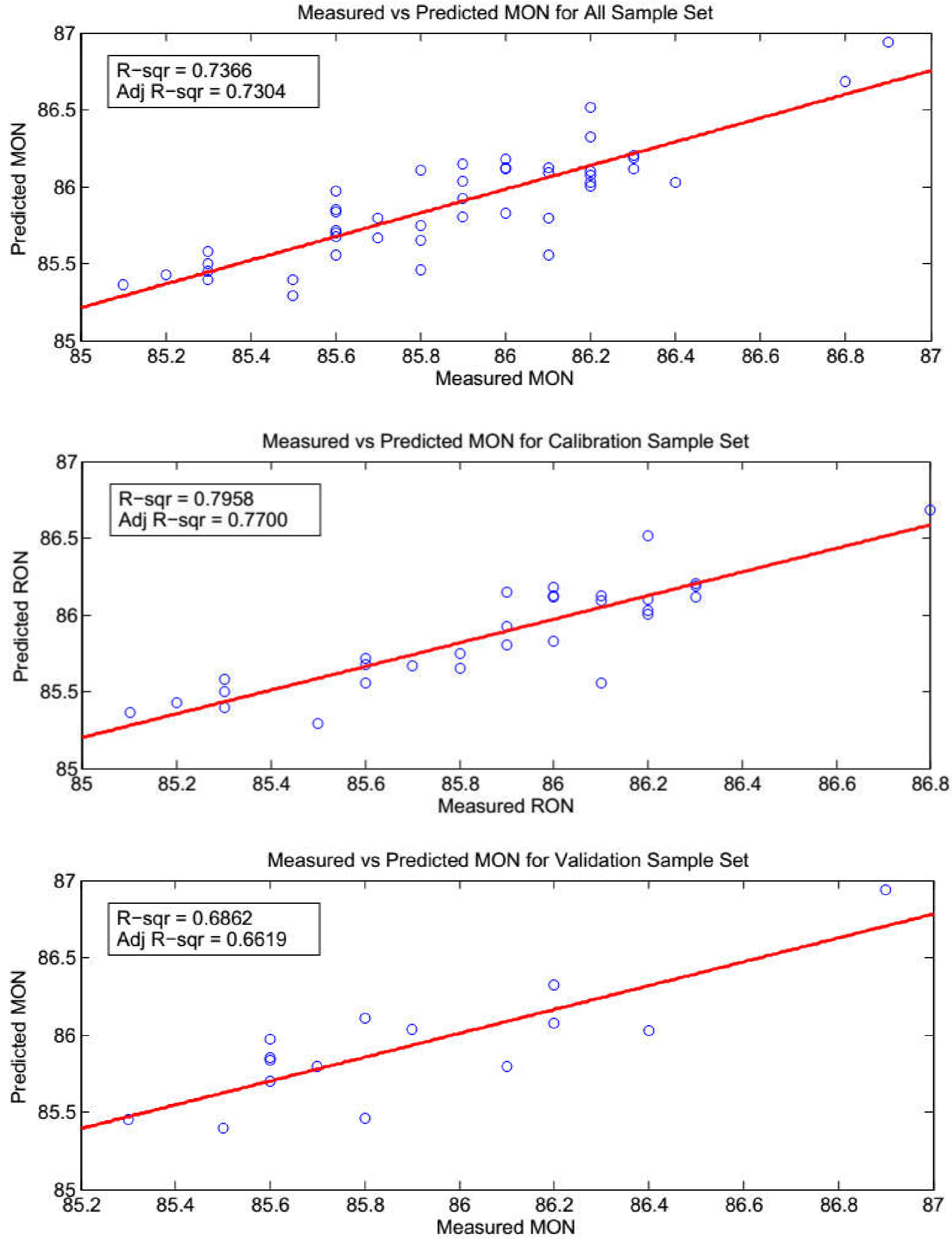


Figure 6.35: PLS-MON: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set.

6.5.3 Aromatics

As seen in Fig 6.36 a) and b), after the 6th PC, there is no more significant contribution to the variance from PC 7, 8 etc. 6 PCs are good enough to cover 95.74% of X data variance. Furthermore, the RMSECV (cross-validation error) values start to increase beyond 6 PCs.

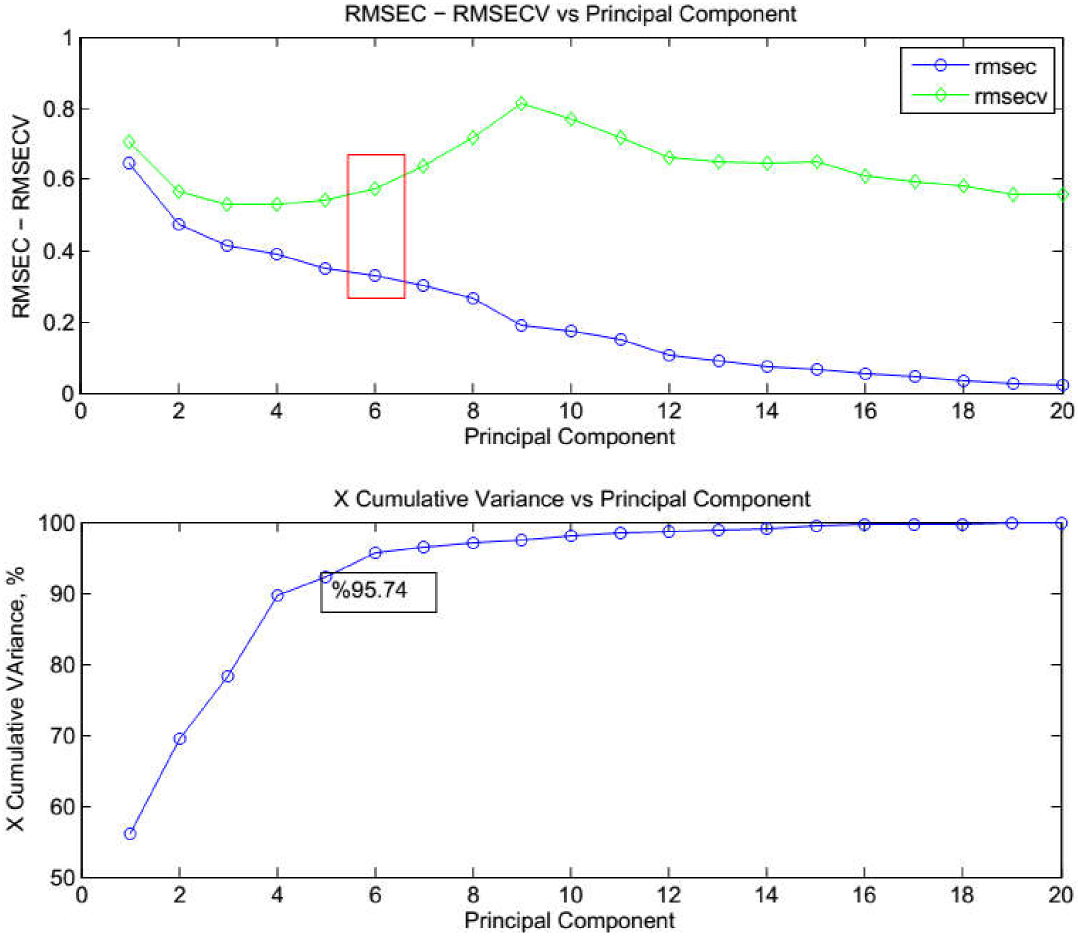


Figure 6.36: PLS-Aromatics **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance.

The reproducibility, R, formula for Aromatics is given in Equation 6.6. All samples have residuals smaller than calculated reproducibility value for test results. The residuals for the all sample set are plotted in Figure 6.37.

In parallel to this, the residuals for the validation set are larger than the residuals of the calibration set, as expected. The measured, predicted and residual values are given in Table 6.10 for both the calibration and validation data sets.

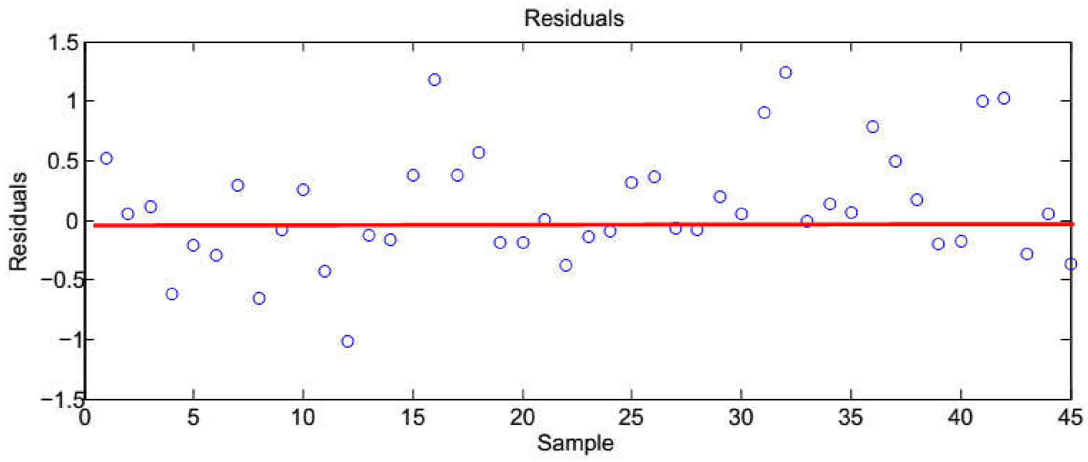


Figure 6.37: PLS-Aromatics: Residuals vs All Sample Set.

The measured values are plotted against the predicted values are for the all sample set, calibration set and validation set in Figure 6.38 a), b) and c). The R^2 value for the measured versus predicted regression line for complete sample set is 0.9571.

Table 6.10: PLS-Aromatics: Measured, Predicted results and Residuals, % v/v.

Calibration Set (1-31)

Sample	Measured	Predicted	Residual
1	33.20	33.72	0.52
2	29.70	29.76	0.06
3	34.80	34.91	0.11
4	38.30	37.69	-0.61
5	33.60	33.39	-0.21
6	33.70	33.41	-0.29
7	33.50	33.79	0.29
8	31.90	31.25	-0.65
9	37.90	37.82	-0.08
10	34.20	34.46	0.26
11	34.50	34.08	-0.42
12	34.40	33.39	-1.01
13	34.00	33.88	-0.12
14	33.60	33.43	-0.17
15	34.30	34.68	0.38
16	33.00	34.18	1.18
17	33.80	34.18	0.38
18	33.60	34.17	0.57
19	32.70	32.51	-0.19
20	30.40	30.21	-0.19
21	32.90	32.91	0.01
22	33.80	33.42	-0.38
23	34.90	34.76	-0.14
24	33.30	33.21	-0.09
25	35.90	36.22	0.32
26	30.00	30.37	0.37
27	37.00	36.93	-0.07
28	37.20	37.12	-0.08
29	30.90	31.09	0.19
30	30.90	30.96	0.06
31	31.70	32.61	0.91

Validation Set (32-45)

32	31.00	32.24	1.24
33	33.60	33.59	-0.01
34	32.10	32.24	0.14
35	31.00	31.07	0.07
36	35.30	36.09	0.79
37	37.20	37.70	0.50
38	39.90	40.07	0.17
39	39.20	39.00	-0.20
40	32.70	32.52	-0.18
41	34.20	35.21	1.01
42	33.30	34.33	1.03
43	33.00	32.71	-0.29
44	33.10	33.16	0.06
45	33.80	33.43	-0.37

When the validation set was predicted by applying the model, RMSEP is 0.63 and R^2 is 0.9611, that is lower than R^2 for the calibration set as expected. For the calibration set, RMSEC and R^2 are 0.33 and 0.9614 correspondingly.

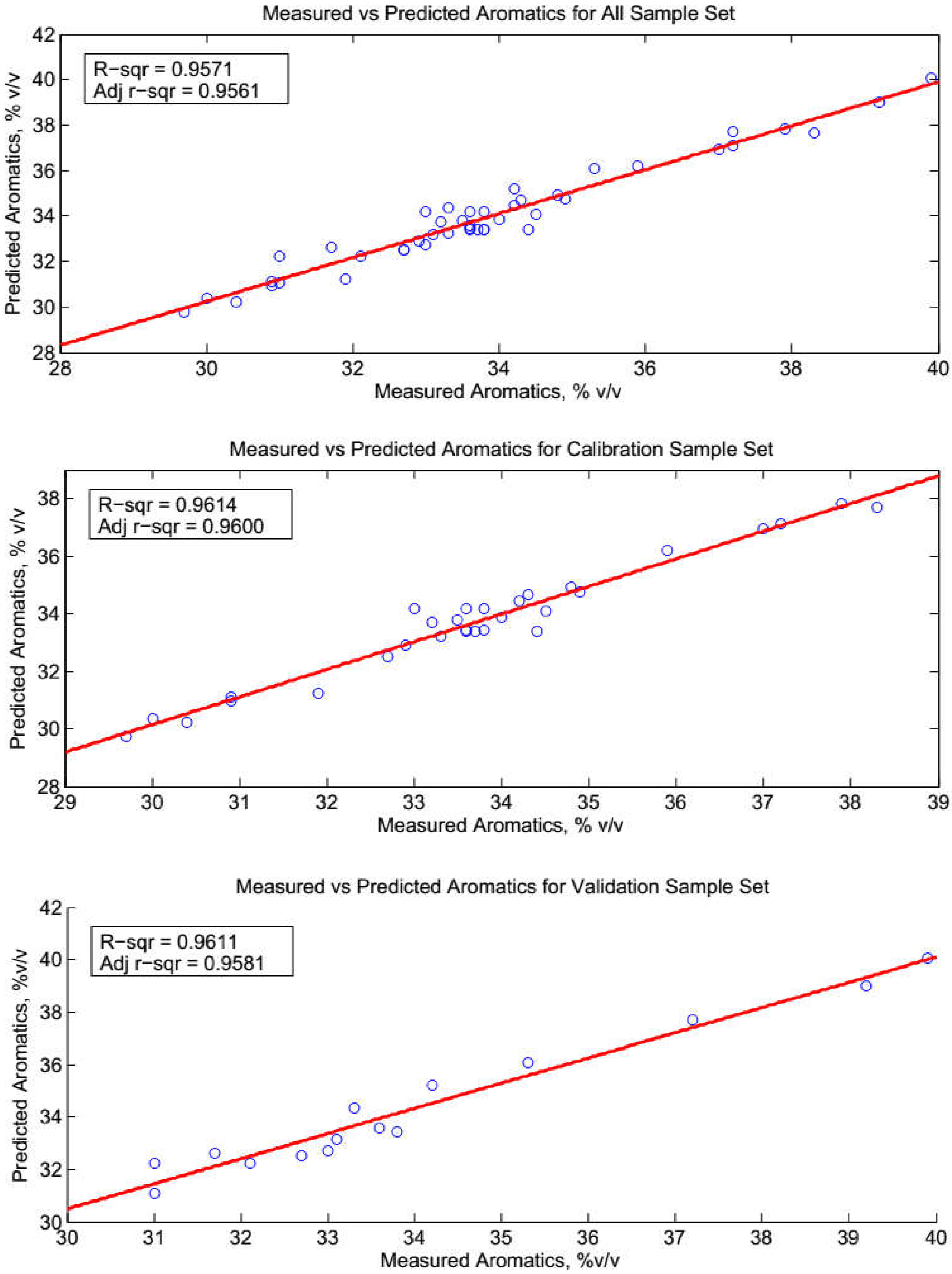


Figure 6.38: PLS-Aromatics: Measured vs. predicted results for a) all sample set, b) calibration set and c) validation set.

6.5.4 Olefins

As seen in Fig 6.39 a) and b), after the 3rd PC, there is some contribution to the variance from PC 4, 5 but for this component when PCs get higher RMSEP gets also higher. Thus 3 PCs are taken for further calculations. 3 PCs are enough to cover 78.15% of X data variance. Furthermore, the RMSECV (cross-validation error) values start to increase beyond 3 PCs.

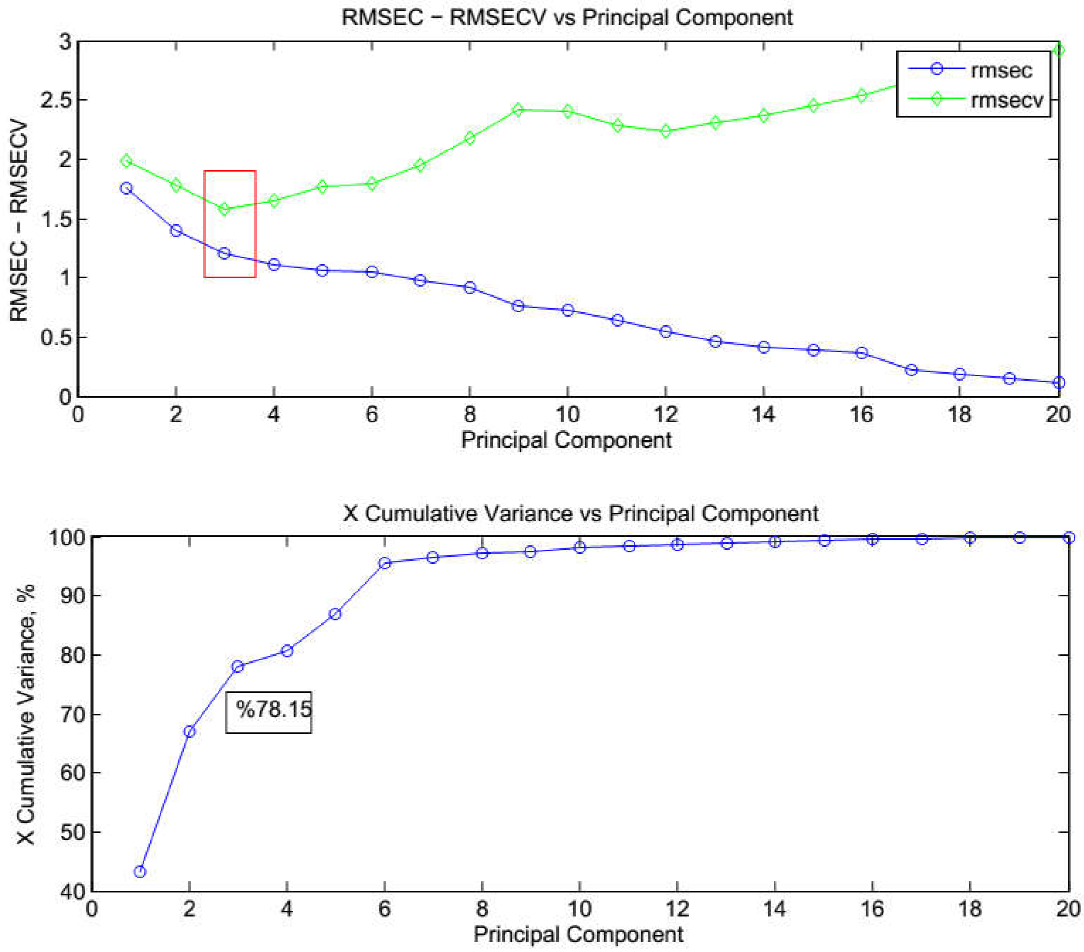


Figure 6.39: PLS-Olefins: **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance.

The Reproducibility given in Equation 6.7. Samples having the residuals higher than calculated reproducibility values are 12, 18, 27, 30 and 35. The residuals for the all sample set are plotted in Figure 6.40.

The measured, predicted and residual values are given in Table 6.11 for both the calibration and validation data sets.

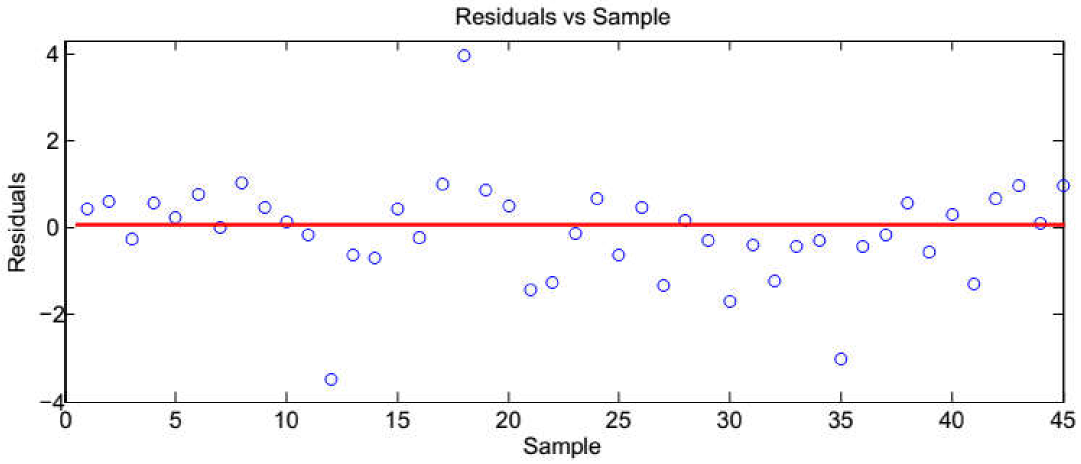


Figure 6.40: PLS-Olefins: Residuals vs All Sample Set.

The measured values are plotted against the predicted values are for the all sample set, calibration set and validation set in Figure 6.41 a), b) and c). The R^2 value for the measured versus predicted regression line for complete sample set is 0.6665.

Table 6.11: PLS-Olefins: Measured, Predicted results and Residuals, % v/v.

Calibration Set (1-31)

Sample	Measured	Predicted	Residual
1	2.00	2.43	0.43
2	5.70	6.29	0.59
3	8.30	8.04	-0.26
4	3.80	4.37	0.57
5	4.20	4.45	0.25
6	2.80	3.57	0.77
7	3.10	3.10	0.00
8	6.00	7.02	1.02
9	4.80	5.25	0.45
10	2.80	2.92	0.12
11	8.10	7.94	-0.16
12	9.00	5.52	-3.48
13	5.80	5.18	-0.62
14	5.90	5.21	-0.69
15	5.40	5.83	0.43
16	5.60	5.38	-0.22
17	5.70	6.69	0.99
18	3.90	7.87	3.97
19	2.90	3.76	0.86
20	6.10	6.60	0.50
21	9.60	8.16	-1.44
22	9.80	8.53	-1.27
23	5.10	4.98	-0.12
24	5.40	6.08	0.68
25	7.10	6.48	-0.62
26	4.50	4.96	0.46
27	5.30	3.96	-1.34
28	3.60	3.76	0.16
29	5.80	5.49	-0.31
30	7.60	5.89	-1.71
31	6.20	5.80	-0.40

Validation Set (32-45)

Sample	Measured	Predicted	Residual
32	6.30	5.09	-1.21
33	3.10	2.66	-0.44
34	5.10	4.79	-0.31
35	7.40	4.37	-3.03
36	7.90	7.46	-0.44
37	3.90	3.75	-0.15
38	1.20	1.77	0.57
39	4.90	4.35	-0.55
40	6.40	6.72	0.32
41	8.40	7.10	-1.30
42	5.40	6.06	0.66
43	4.20	5.17	0.97
44	7.20	7.31	0.11
45	3.70	4.65	0.95

When the validation set was predicted by applying the model, RMSEP is 1.03 and R^2 is 0.7196 that is higher than R^2 for the calibration set as unusual. For the calibration set, RMSEC and R^2 are 1.20 and 0.6444 correspondingly.

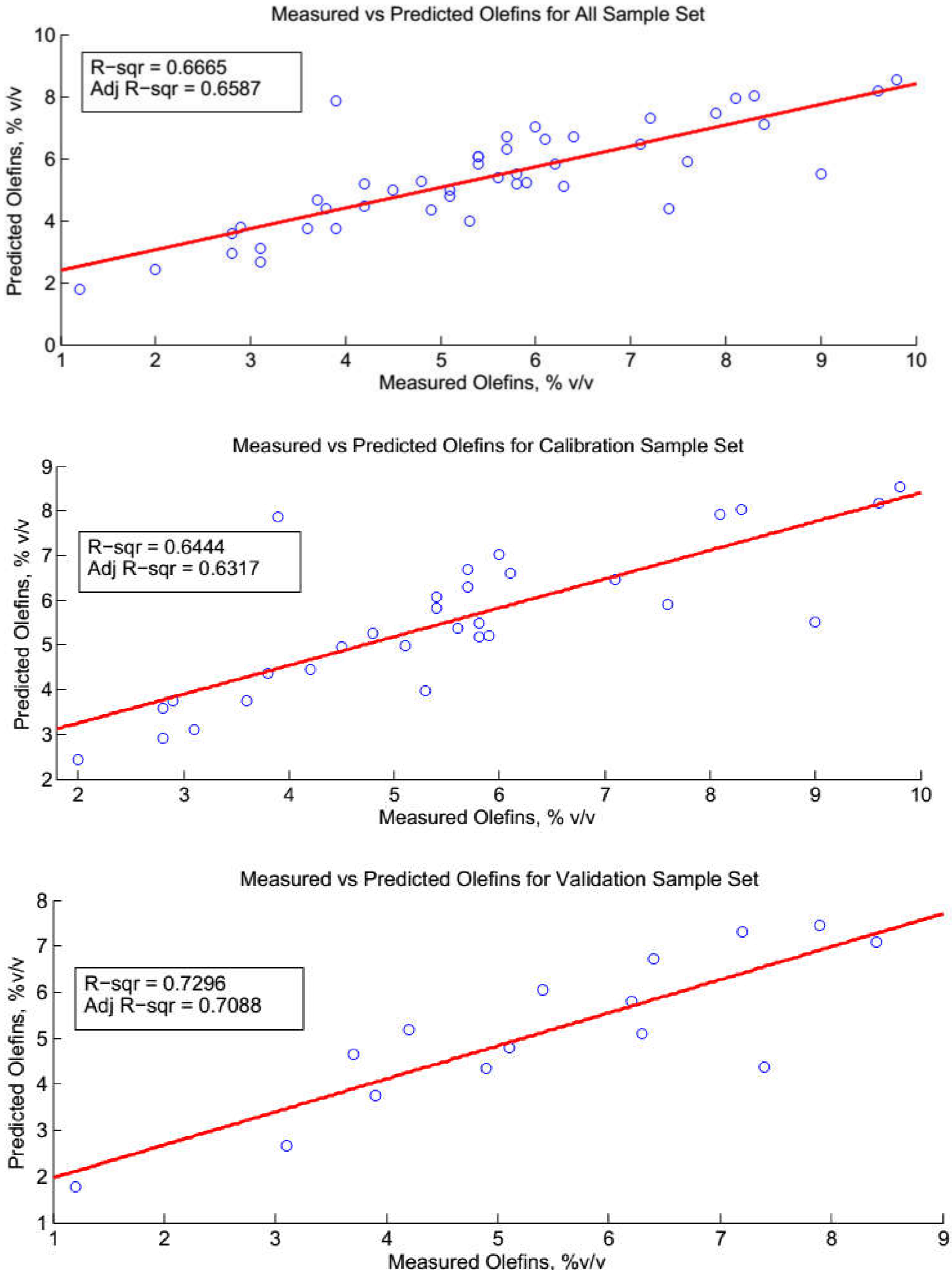


Figure 6.41: PLS-Olefins: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set.

6.5.5 Benzene

As seen in Fig 6.42 a) and b), after the 11th PC, there is no more significant contribution to the variance from PC 12, 13 etc. 11 PCs are good enough to cover 98.45% of X data variance. Furthermore, the RMSECV (cross-validation error) values start to increase beyond 11 PCs.

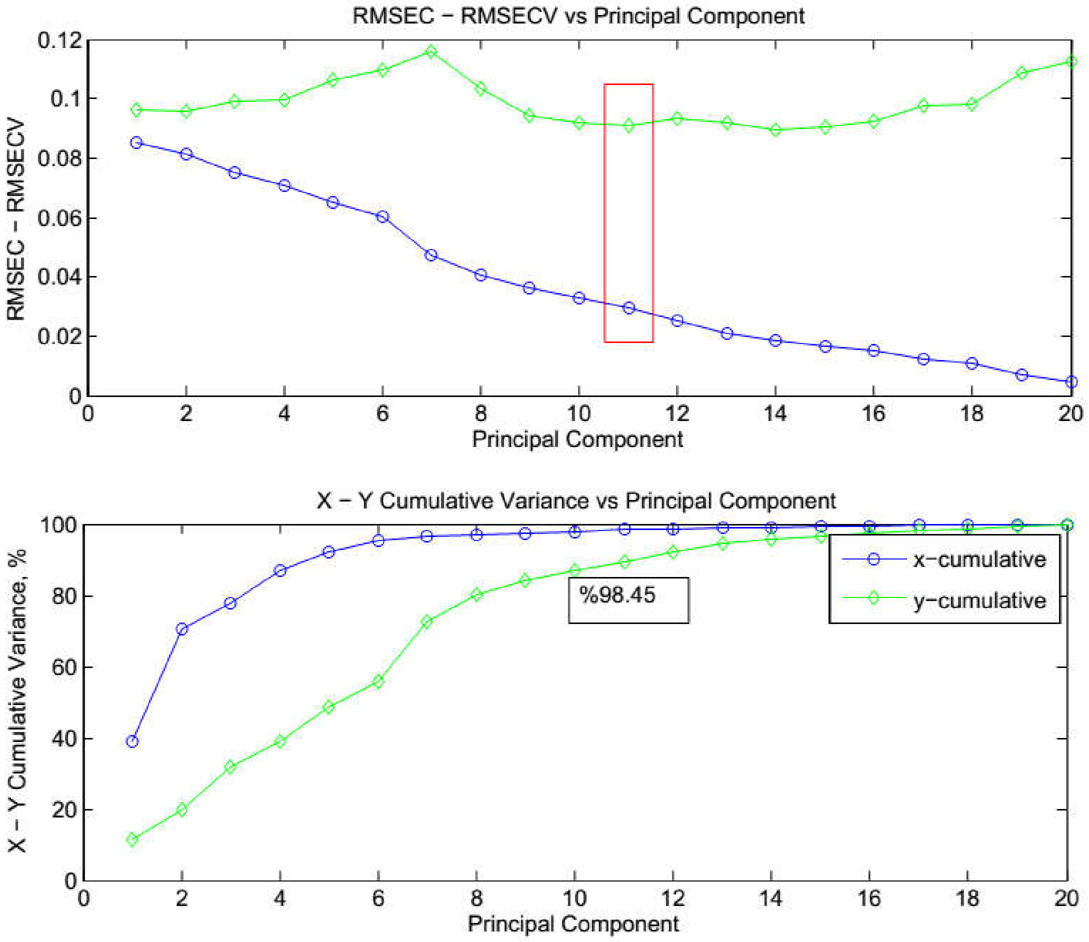


Figure 6.42: PLS-Benzene: **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance.

The reproducibility, R is given in Equation 6.8. Samples having the residuals higher than calculated reproducibility values are 3, 12, 15, 17, 18, 24, 33, 34, 35, 36, 38, 39, 40, 41, 42 and 45. The residuals for the all sample set are plotted in Figure 6.43.

In parallel to this, the residuals for the validation set are larger than the residuals of the calibration set, as expected. The measured, predicted and residual values are given in Table 6.12 for both the calibration and validation data sets.

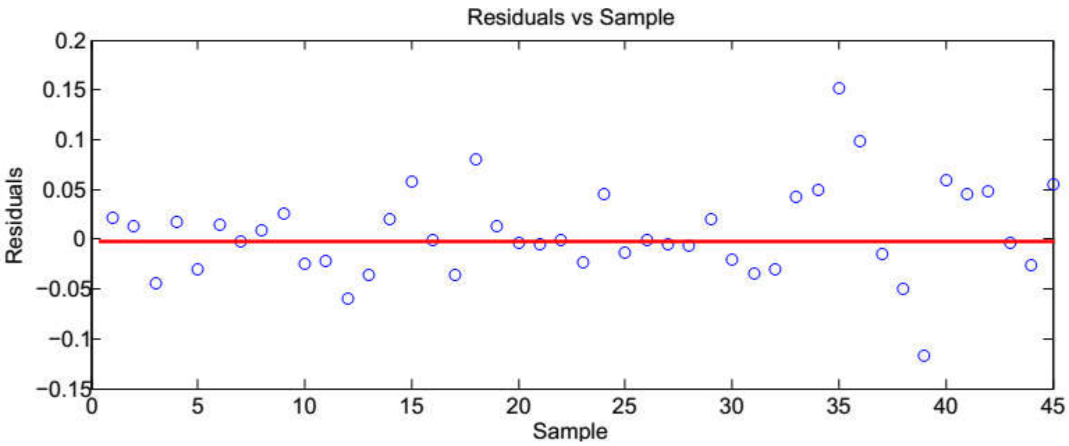


Figure 6.43: PLS-Benzene: Residuals vs All Sample Set.

The measured values are plotted against the predicted values are for the all sample set, calibration set and validation set in Figure 6.44 a), b) and c). The R^2 value for the measured versus the predicted regression line for complete sample set is 0.7461.

Table 6.12: PLS-Benzene: Measured, Predicted results and Residuals, %v/v.

Calibration Set (1-31)

Sample	Measured	Predicted	Residual
1	0.76	0.78	0.02
2	0.81	0.82	0.01
3	0.74	0.70	-0.04
4	0.87	0.89	0.02
5	0.80	0.77	-0.03
6	0.80	0.81	0.01
7	0.84	0.84	0.00
8	0.73	0.74	0.01
9	0.75	0.78	0.03
10	0.77	0.74	-0.03
11	0.75	0.73	-0.02
12	0.92	0.86	-0.06
13	0.86	0.82	-0.04
14	0.86	0.88	0.02
15	0.57	0.63	0.06
16	0.57	0.57	0.00
17	0.69	0.65	-0.04
18	0.72	0.80	0.08
19	0.95	0.96	0.01
20	0.70	0.70	0.00
21	0.94	0.93	-0.01
22	0.77	0.77	0.00
23	0.87	0.85	-0.02
24	0.70	0.74	0.04
25	0.82	0.81	-0.01
26	0.71	0.71	0.00
27	0.84	0.83	-0.01
28	0.83	0.82	-0.01
29	0.72	0.74	0.02
30	0.72	0.70	-0.02
31	0.75	0.72	-0.03

Validation Set (32-45)

Sample	Measured	Predicted	Residual
32	0.72	0.69	-0.03
33	0.77	0.81	0.04
34	0.73	0.78	0.05
35	0.70	0.85	0.15
36	0.82	0.92	0.10
37	0.88	0.87	-0.01
38	0.94	0.89	-0.05
39	0.92	0.80	-0.12
40	0.77	0.83	0.06
41	0.64	0.69	0.05
42	0.70	0.75	0.05
43	0.69	0.69	0.00
44	0.79	0.76	-0.03
45	0.73	0.79	0.06

When the validation set was predicted by applying the model, RMSEP is 0.07 and R^2 is 0.4511, that is lower than R^2 for the calibration set, as expected. For the calibration set, RMSEC and R^2 are 0.03 and 0.8949 correspondingly.

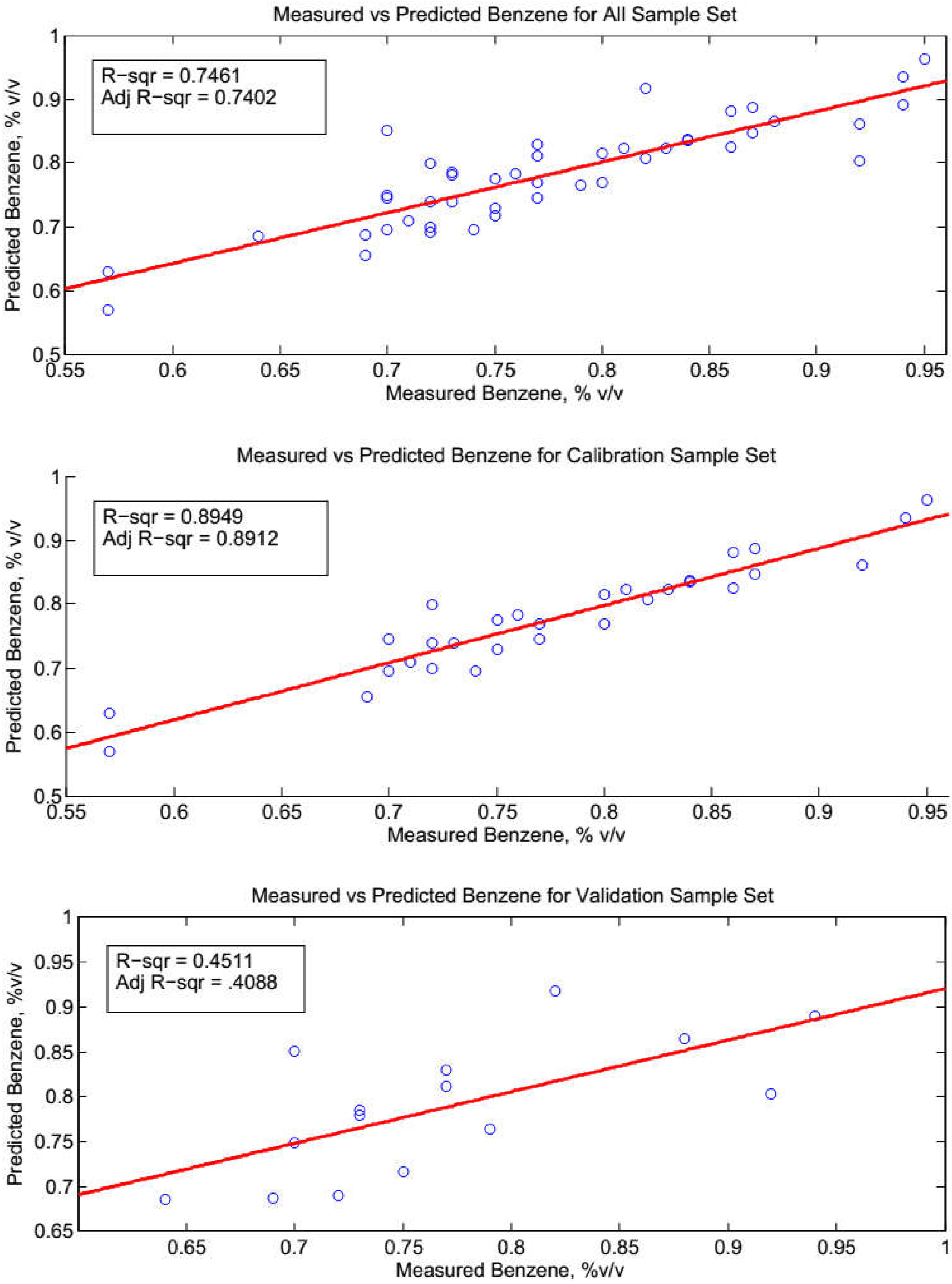


Figure 6.44: PLS-Benzene: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set.

6.5.6 Density

As seen in Fig 6.45 a) and b), after the 4th PC, there is no more significant contribution to the variance from PC 5, 6 etc. 4 PCs are good enough to cover 86.03% of X data variance. Furthermore, the RMSECV (cross-validation error) values start to increase beyond 4 PCs.

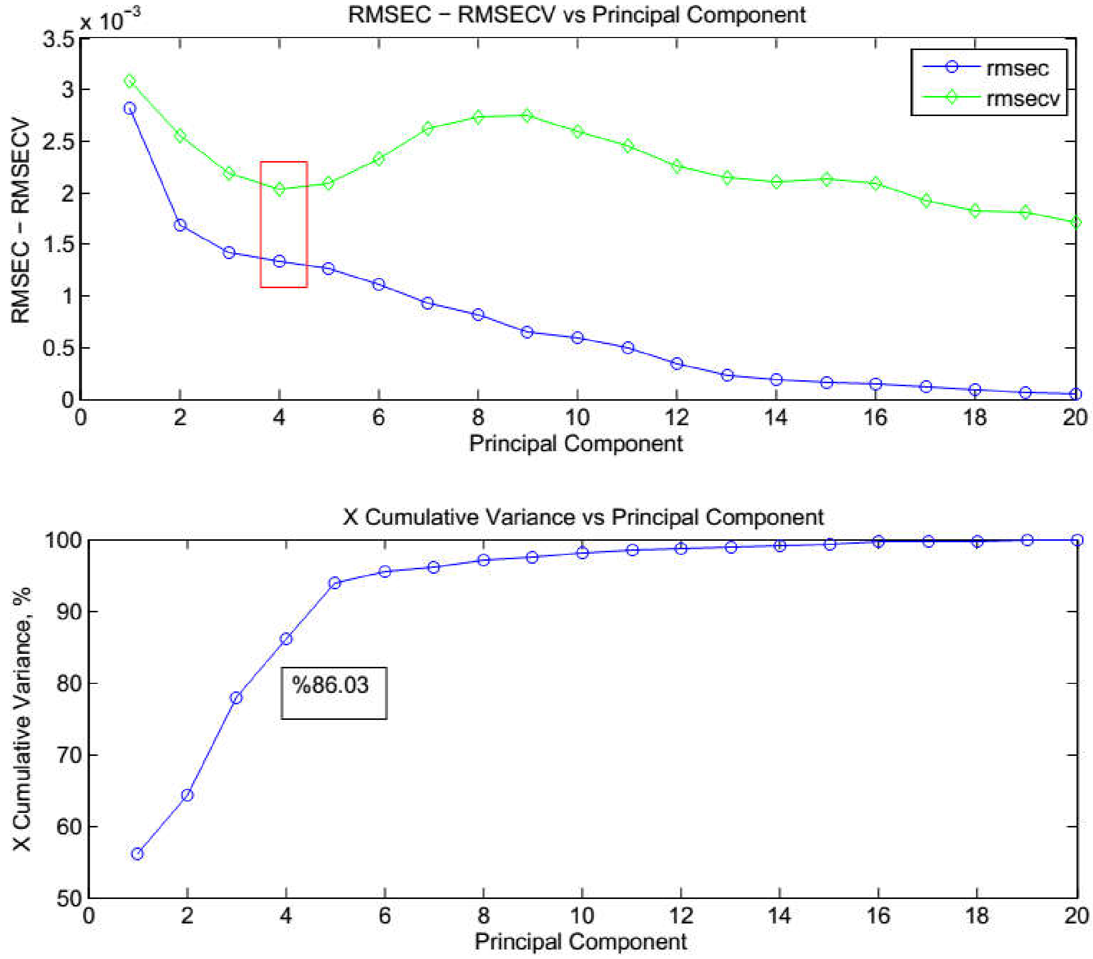


Figure 6.45: PLS-Density **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance.

The reproducibility is given in below Equation 6.7. Samples having residuals higher than this reproducibility are 3, 8, 12, 18, 31, 32, 36 and 42. The residuals for all sample set are plotted in Figure 6.46.

In parallel to this, the residuals for the validation set are larger than the residuals of the calibration set, as expected. The measured, predicted and residual values are given in Table 6.13 for both the calibration and validation data sets.

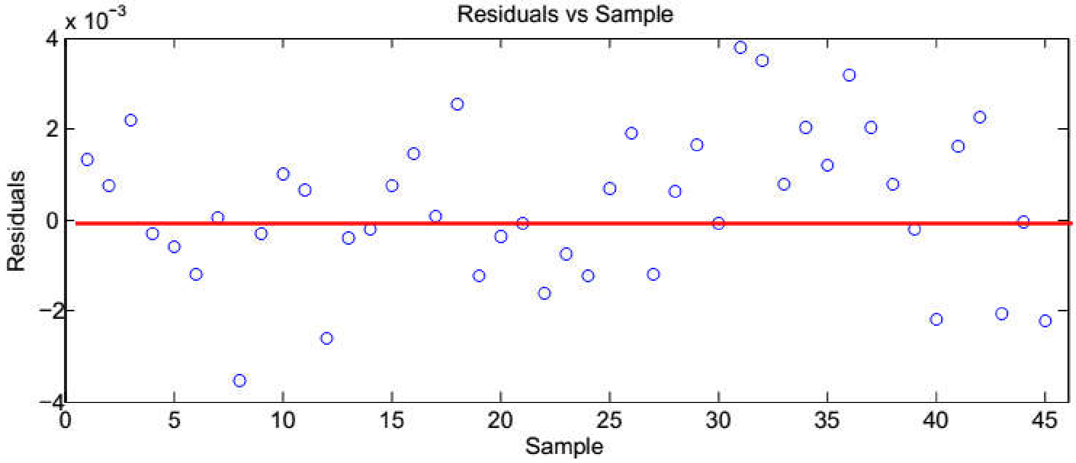


Figure 6.46: PLS-Density: Residuals vs All Sample Set.

The measured values are plotted against the predicted values are for the all sample set, calibration set and validation set in Figure 6.47 a), b) and c). The R^2 value for the measured versus predicted regression line for complete sample set is 0.9077.

Table 6.13: PLS-Density: Measured, Predicted results and Residuals, kg/l.

Calibration Set (1-31)

Sample	Measured	Predicted	Residual
1	0.7336	0.7349	0.0013
2	0.7311	0.7319	0.0008
3	0.7466	0.7488	0.0022
4	0.7462	0.7459	-0.0003
5	0.7387	0.7381	-0.0006
6	0.7391	0.7379	-0.0012
7	0.7372	0.7372	0.0000
8	0.7395	0.7360	-0.0035
9	0.7473	0.7470	-0.0003
10	0.7405	0.7415	0.0010
11	0.7420	0.7427	0.0007
12	0.7419	0.7393	-0.0026
13	0.7398	0.7394	-0.0004
14	0.7394	0.7392	-0.0002
15	0.7462	0.7469	0.0007
16	0.7386	0.7400	0.0014
17	0.7465	0.7466	0.0001
18	0.7396	0.7421	0.0025
19	0.7360	0.7348	-0.0012
20	0.7337	0.7333	-0.0004
21	0.7398	0.7397	-0.0001
22	0.7481	0.7465	-0.0016
23	0.7429	0.7421	-0.0008
24	0.7455	0.7443	-0.0012
25	0.7428	0.7435	0.0007
26	0.7291	0.7310	0.0019
27	0.7439	0.7427	-0.0012
28	0.7428	0.7434	0.0006
29	0.7307	0.7324	0.0017
30	0.7323	0.7322	-0.0001
31	0.7332	0.7370	0.0038

Validation Set (32-45)

Sample	Measured	Predicted	Residual
32	0.7310	0.7345	0.0035
33	0.7382	0.7390	0.0008
34	0.7327	0.7347	0.0020
35	0.7308	0.7320	0.0012
36	0.7421	0.7453	0.0032
37	0.7432	0.7452	0.0020
38	0.7483	0.7491	0.0008
39	0.7491	0.7489	-0.0002
40	0.7407	0.7385	-0.0022
41	0.7430	0.7446	0.0016
42	0.7455	0.7478	0.0023
43	0.7392	0.7371	-0.0021
44	0.7399	0.7398	-0.0001
45	0.7400	0.7378	-0.0022

When the validation set was predicted by applying the model, RMSEP is 0.0022 and R^2 is 0.8833 that is lower than R^2 for the calibration set as expected. For the calibration set, RMSEC and R^2 are 0.0013 and 0.9333 correspondingly.

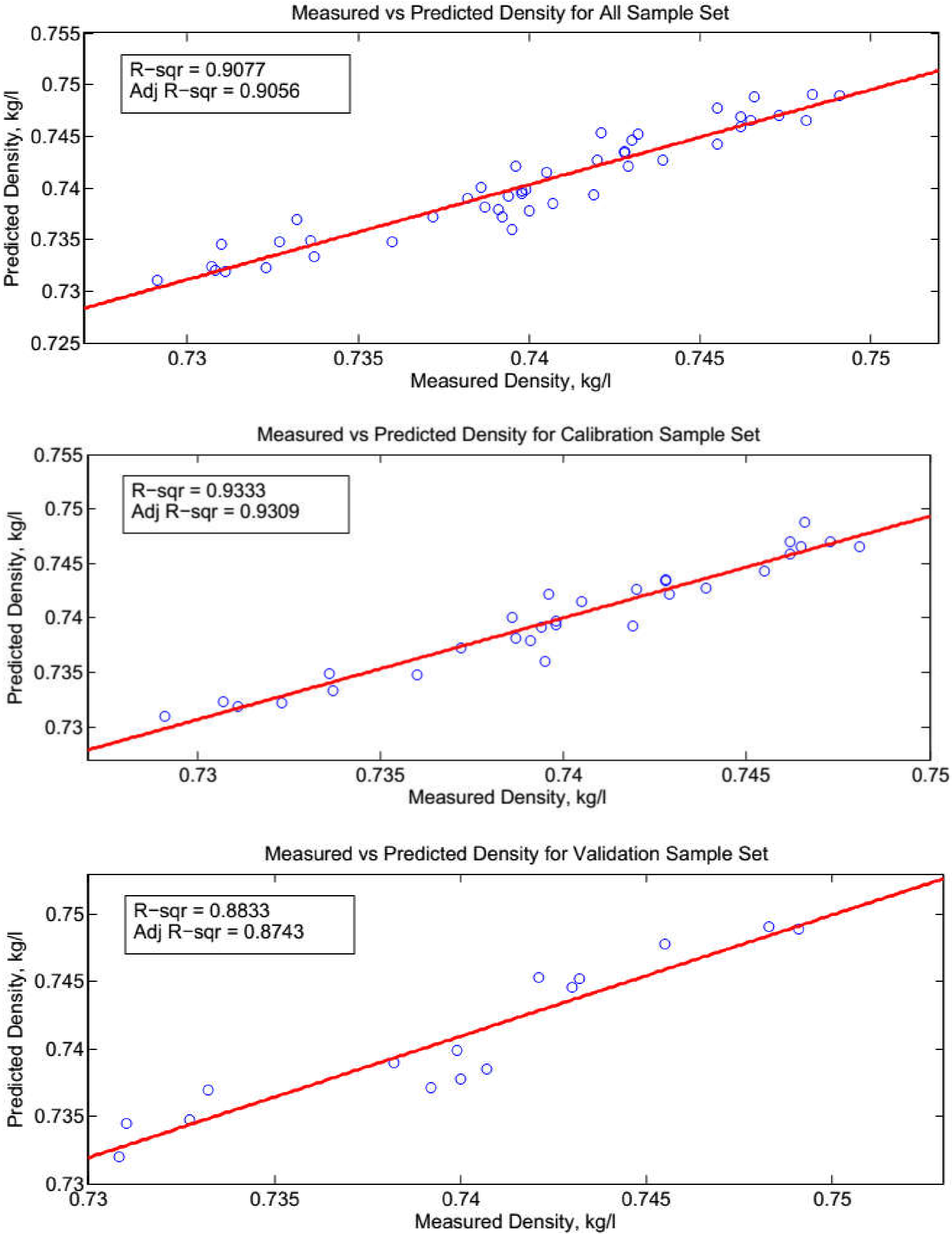


Figure 6.47: PLS-Density: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set.

6.5.7 E150 –Evaporated at 150 °C

As seen in Fig 6.48 a) and b), after the 6th PC, there is no more significant contribution to the variance from PC 7, 8 etc. 6 PCs are good enough to cover 95.62% of X data variance. Furthermore, the RMSECV (cross-validation error) values start to increase beyond 6 PCs.

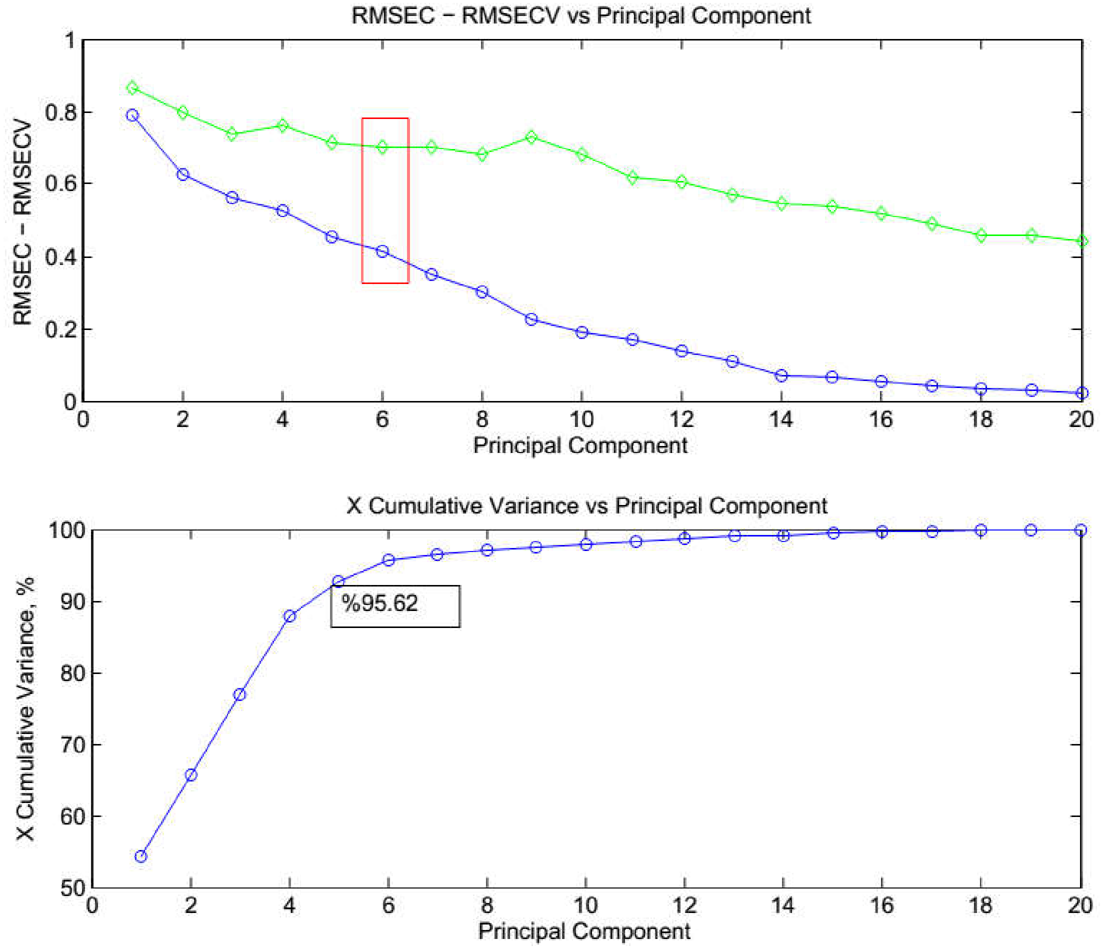


Figure 6.48: PLS-E150 **a)** PC vs. RMSEC-RMSECV, **b)** PC vs. X Cumulative variance.

The reproducibility is given in Equation 6.10. Sample having residual higher than this reproducibility is 36. The residuals for the all sample set are plotted in Figure 6.49.

In parallel to this, the residuals for the validation set are larger than the residuals of the calibration set, as expected. The measured, predicted and residual values are given in Table 6.14 for both the calibration and validation data sets.

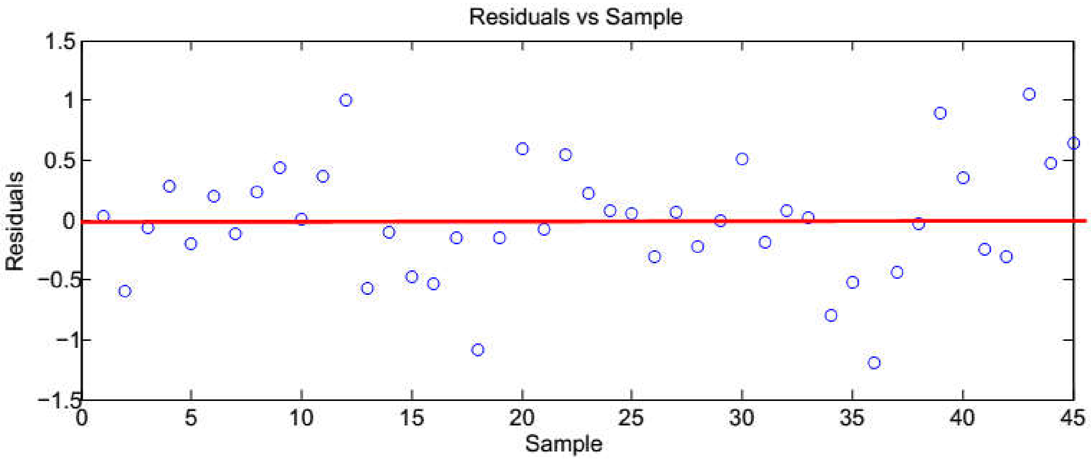


Figure 6.49: PLS-E150: Residuals vs All Sample Set.

The measured values are plotted against the predicted values are for the all sample set, calibration set and validation set in Figure 6.50 a), b) and c). The R^2 value for the measured versus predicted regression line for complete sample set is 0.7226.

Table 6.14: PLS-E150: Measured, Predicted results and Residuals, °C.

Calibration Set (1-31)

Sample	Measured	Predicted	Residual
1	92.70	92.73	0.03
2	92.30	91.71	-0.59
3	90.50	90.43	-0.07
4	89.60	89.88	0.28
5	91.10	90.90	-0.20
6	91.20	91.40	0.20
7	91.40	91.28	-0.12
8	91.30	91.54	0.24
9	89.90	90.34	0.44
10	91.90	91.91	0.01
11	89.30	89.67	0.37
12	89.70	90.71	1.01
13	91.10	90.53	-0.57
14	90.50	90.40	-0.10
15	91.10	90.62	-0.48
16	91.70	91.16	-0.54
17	90.90	90.75	-0.15
18	91.00	89.92	-1.08
19	92.00	91.85	-0.15
20	90.40	91.00	0.60
21	90.10	90.02	-0.08
22	89.80	90.35	0.55
23	90.30	90.53	0.23
24	91.20	91.28	0.08
25	91.20	91.25	0.05
26	92.70	92.40	-0.30
27	91.40	91.46	0.06
28	91.50	91.27	-0.23
29	92.10	92.10	0.00
30	92.10	92.61	0.51
31	92.00	91.82	-0.18

Validation Set (32-45)

Sample	Measured	Predicted	Residual
32	92.30	92.38	0.08
33	92.50	92.51	0.01
34	92.60	91.81	-0.79
35	92.90	92.38	-0.52
36	91.40	90.21	-1.19
37	91.50	91.06	-0.44
38	90.80	90.76	-0.04
39	89.70	90.59	0.89
40	91.00	91.36	0.36
41	90.70	90.46	-0.24
42	91.20	90.89	-0.31
43	91.10	92.15	1.05
44	89.90	90.38	0.48
45	90.90	91.54	0.64

When the validation set was predicted by applying the model, RMSEP is 0.60 and R^2 is 0.5801, that is lower than R^2 for the calibration set as expected. For the calibration set, RMSEC and R^2 are 0.41 and 0.7886 correspondingly.

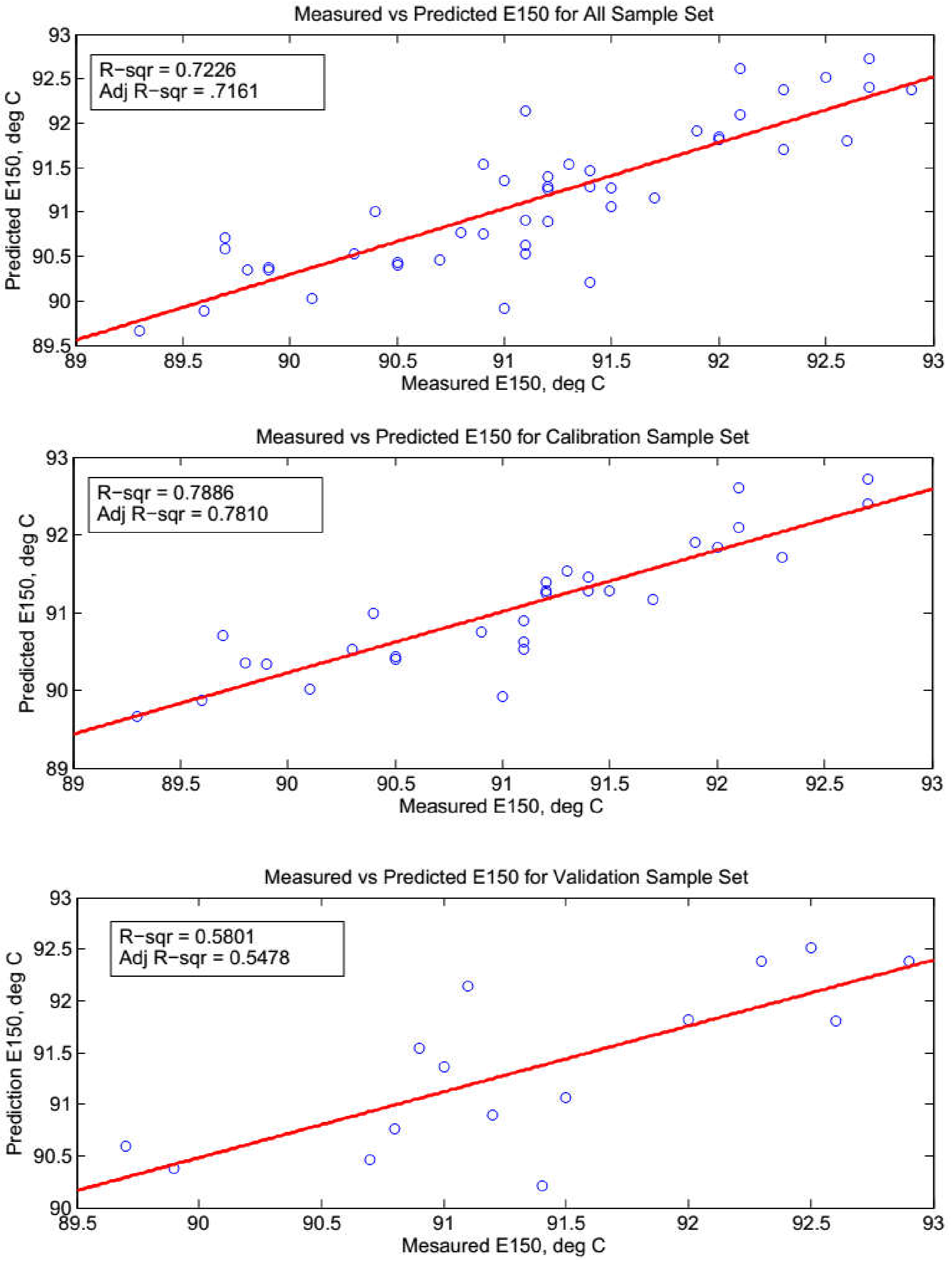


Figure 6.50: PLS-E150: Measured vs. predicted results for **a)** all sample set, **b)** calibration set and **c)** validation set.

6.6 Results and Summary

As a result of the predicted values, according to PCR and PLS algorithms, for all the components, given in the previous sections, it is very clear that the chemometric models are very powerful tools in getting the expected results. By combining the spectroscopic information and the chemometric modelling techniques, it is possible to avoid the reference analyses in the laboratory which requires more time and cost. A summary of the results and precision data calculated for PCR and PLS are given in Table 6.15 and 6.16 correspondingly.

The residuals calculated by using PLS for RON, MON and Aromatics components are smaller than the reproducibility precision value given in corresponding reference test method. For E150, there is only one sample that has higher residual than the reproducibility. More than 90% of residuals are even smaller than half of the reproducibility (R/2) for RON, MON, Aromatics and E150, which means the prediction model and reference test methods are almost similar.

The residuals calculated by using PCR for MON and Aromatics components are smaller than the reproducibility precision value given in corresponding reference test method. For RON, there is only one sample has higher residual than the reproducibility. The PCR model is also very good at predicting the RON, MON, Aromatics values.

It is also observed that PLS is more powerful than PCR for RON, MON, Aromatics, Benzene, Density and E150. The residuals are smaller for samples predicted by PLS than that of PCR. Only for olefins, PCR has better precision values than PLS.

Both PCR and PLS techniques are not very good in prediction for benzene concentration in gasoline samples. All samples have benzene concentration lower than 1% and this is obviously a reason for having high residuals.

For density and olefins properties, the RMSEP values are higher than reference values. A further study would be helpful to see the real performance of two (PCR and PLS) techniques by varying the number of calibration sample set, by changing the pre-processing methods, by utilizing a variable selection algorithm.

Table 6.15: Summary of values for PCR.

	RON	**MON**	**Aromatics**	**Olefins**	**Benzene**	**Density**	**E150**
number of PC	6	5	6	6	6	5	6
X variance, cumulative %	96.18	94.93	96.18	96.18	96.18	94.93	96.18
RMSEC	0.1986	0.2436	0.4138	1.2323	0.0762	0.0015	0.5591
RMSECV	0.2669	0.3127	0.519	1.5841	0.0939	0.002	0.3717
RMSEP	0.3552	0.2912	0.6608	0.9898	0.0668	0.0022	0.8071
R^2-all	0.889	0.5668	0.9528	0.6602	0.3238	0.894	0.5146
R^2-cal	0.9399	0.604	0.9615	0.6264	0.2967	0.9193	0.613
R^2-val	0.7654	0.5267	0.9543	0.7409	0.4156	0.8674	0.4121

Table 6.16: Summary of values for PLS.

	RON	**MON**	**Aromatics**	**Olefins**	**Benzene**	**Density**	**E150**
number of PC	6	5	6	3	11	4	6
X variance, cumulative %	95.41	90.49	95.74	78.15	98.45	86.03	95.62
y variance, cumulative %	97.36	77	97.52	64.44	89.49	93.33	78.86
RMSEC	0.1316	0.1856	0.3317	1.2023	0.02943	0.0013	0.4132
RMSECV	0.2232	0.3785	0.5738	1.5768	0.09111	0.002	0.7006
RMSEP	0.2934	0.2312	0.6304	1.0334	0.06668	0.0022	0.5979
R^2-all	0.9354	0.7366	0.9571	0.6665	0.7402	0.9056	0.7226
R^2-cal	0.9736	0.7958	0.9614	0.6444	0.8949	0.9333	0.7886
R^2-val	0.8475	0.6861	0.9611	0.7296	0.4511	0.8833	0.5801

7. CONCLUSION AND RECOMMENDATIONS FOR FUTURE RESEARCH

In this study, the multivariate calibration models were developed by combining the Near Infrared Spectroscopy data and the calibration techniques for 45 gasoline samples collected at Tupras Izmit Refinery. The gasoline samples were analyzed at the Tupras Izmit Refinery Quality Control Laboratory (accredited from ISO/IEC 17025:2005) by the reference test methods and modelled for RON, MON, Aromatics, Olefins, Benzene, Density and E150 distillation components. Same samples were analyzed by NIR Spectroscopy. PCR and PLS (SIMPLS algorithm) techniques were used to obtain calibration models by using a calibration sample set (30 samples), then a validation sample set (15 samples) was used for prediction. The RMSEP values were calculated and compared with the R value of reference test method.

According to the results, all the residuals calculated for MON and Aromatics by using PCR and for RON, MON and Aromatics by using PLS are smaller than reproducibility of reference test method. This means that PCR and PLS are successful to predict given properties. The RMSEP values for these components are smaller than half of the reproducibility (R/2), that is accepted as a performance indicator to compare measured and predicted results [65].

In addition, the PLS model results indicated that more than 90% of residuals and the RMSEP values are even smaller than half of the reproducibility (R/2) for RON, MON, Aromatics and E150. This means that PLS model can be safely used as a replacement for the reference test methods for RON, MON, Aromatics and E150 with the condition of remaining in studied range and with the condition of using samples of similar molecular structure.

Mainly the PLS models have smaller RMSEP values than the PCR models, with the exception of olefins property. This results is also expected as explained in the text.

It is also concluded that a detailed study with different parameters needed in order to get better results for density and olefins properties.

For future studies on this subject, there are mainly three issues to concentrate on. Firstly, increasing the number of samples in the calibration and validation sets and extending the range for individual properties will be helpful. Secondly, using different multivariate calibration techniques both linear and non-linear with alternative pre-processing and variable selection methods is also worth to study. As given in the Literature Review section, multivariate calibration techniques have unique advantages and disadvantages. The best technique should be found for each interested property.

Lastly, in addition to NIR spectroscopy, some other spectroscopic measurement methods can be used, i.e. Infrared spectroscopy and Nuclear Magnetic Resonance (NMR) spectroscopy. Each measurement system has unique properties and this will be helpful to get most form the sample as input to calibration model. Especially, NMR spectroscopy is becoming more popular since it gives very important information about the chemical structure different than NIR. Another advantage is that dark samples can be analyzed by NMR whereas it is not possible to do it by NIR.

REFERENCES

[1] OPEC, (2013). 2013 World Oil Look, *OPEC Secretariat,* Vienna.

[2] EPDK, (2013). 2013 Petrol Piyasasi Sektor Raporu, EPDK, Ankara.

[3] TS EN 228, 2013. Automotive fuels - Unleaded petrol - Requirements and test methods, *Turkish Standards Institution,* Ankara.

[4] Ku, M., Chung, H., Lee, J. (1998). Rapid Compositional Analysis of Naphtha by Near-Infrared Spectroscopy, *Bull. Korean Chem. Soc.,* Vol.**14**, no.19, pp. 1189-1193

[5] Wiley, (2007). Wiley Critical Content: Petroleum Technology, *Wiley-Interscience.*

[6] ASTM D 1298, 2012. Standard Test Method for Density, Relative Density, or API Gravity of Crude Petroleum and Liquid Petroleum Products by Hydrometer Method*, ASTM International,* West Conshohocken, PA.

[7] TS EN ISO 3405, 2011. Petroleum products - Determination of distillation characteristics at atmospheric pressure, *Turkish Standards Institution,* Ankara.

[8] TS EN ISO 5164, 2014. Petroleum products - Determination of knock characteristics of motor fuels - Research method*, Turkish Standards Institution,* Ankara.

[9] TS EN ISO 5163, 2014. Petroleum products - Determination of knock characteristics of motor and aviation fuels - Motor method (ISO 5163:2014), *Turkish Standards Institution,* Ankara.

[10] ASTM D 2699, 2014. Standard Test Method for Research Octane Number of Spark-Ignition Engine Fuel, *ASTM International*, West Conshohocken, PA.

[11] ASTM D 2700, 2014. Standard Test Method for Motor Octane Number of Spark-Ignition Engine Fuel, *ASTM International*, West Conshohocken, PA.

[12] ASTM D 86, 2012. Standard Test Method for Distillation of Petroleum Products at Atmospheric Pressure, *ASTM International,* West Conshohocken, PA.

[13] TS EN ISO 22854, 2014. Liquid petroleum products - Determination of hydrocarbon types and oxygenates in automotive-motor gasoline and in ethanol (E85) automotive fuel - Multidimensional gas chromatography method (ISO 22854:2014), *Turkish Standards Institution,* Ankara.

[14] Davies, A. M. C., 2005. An Introduction to Near infrared spectroscopy, *The Newsletter of the International Council for Near Infrared Spectroscopy,* Vol. 16, no. 7, pp. 9-11.

[15] Eldin, A. B., 2011. *Near Infra Red Spectroscopy,* Wide Spectra of Quality Control, InTech, retrieved from http://www.intechopen.com/books/wide-spectra-of-qualitycontrol/near-infra-red-spectroscopy.

[16] Burns, A.D., Ciurczak, E. W., 2008. Handbook of Near-Infrared Analysis, 3rd Edition, *CRC Press,* Boca Raton, FL.

[17] Stuart, B. H., 2004. Infrared Spectroscopy: Fundamentals and Applications, *Wiley.*

[18] Pasquini, C., 2003. Near Infrared Spectroscopy: Fundamentals, Practical Aspects and Analytical Applications, *J. Braz. Chem. Soc.,* Vol. 14, no. 2, pp. 198-219.

[19] Brereton, R. G., 2003. Chemometrics Data Analysis for the Laboratory and Chemical Plant, *Wiley,* England.

[20] Metrohm, 2013. NIR Spectroscopy Monograph, A guide to near-infrared spectroscopic analysis of industrial manufacturing processes, Metrohm, Switzerland.

[21] Chung, H., 2007. Applications of Near-Infrared Spectroscopy in Refineries and Important Issues to Address, *Applied Spectroscopy Reviews,* Vol. 42, pp. 251-285.

[22] European Pharmacopoeia, 2005. 2.2.40. Near-Infrared Spectrophotometry.

[23] Myers, M. E., Stollsteimer, J. J., Wims, A.M., 1975. Determination of Gasoline Octane Numbers from Chemical Composition, *Analytical Chemistry,* Vol. 47, no. 13, pp. 2301-2304.

[24] Kelly, J. J., Barlow, C. H., Jinguji, T. M., Callis, J. B., 1989. Prediction of Gasoline Octane Numbers from Near-Infrared Spectral Features in the Range 660 – 1215 nm, *Analytical Chemistry,*Vol. 61, no. 4, pp. 313-320.

[25] Dolbear, G. E., 1972. Method for Determining Octae Ratings for Gasoline, *U.S. Patent 3,693,071.*

[26] Bohacs, G., Ovadi, Z., 1998. Prediction of Gasoline Properties with Near Infrared Spectroscopy, *J. Near Infrared Spectrosc.,* Vol. 6, pp. 341-348.

[27] Ozdemir, D., 2005. Determination of Octane Number of Gasoline Using Near Infrared Spectroscopy and Genetic Multivariate Calibration Methods, *Petroleum Science and Technology,* Vol. 23, pp. 1139-1152.

[28] Wentzell, P. D., Andrews, D.T., Walsh, J. M., Cooley, J. M., Spencer, P., 1999. Estimation of hydrocarbon types in light gas oils and diesel fuels by ultraviolet absorption spectroscopy and multivariate calibration, *Canadian Journal of Chemistry,* Vol. 77, no. 3, pp. 391-400.

[29] Lavine, B., Workman, J., 2008. Chemometrics, *Anal. Chem.,* Vol. 80, pp. 4519-4531

[30] Varmuza, K., Filzmoser, P., 2009. Introduction to Multivariate Statistical Analysis in Chemometrics, *CRC Press,* UK.

[31] **Bro, R., 2003.** Multivariate Calibration What is in chemometrics for the analytical chemist?, *Analytica Chimica Acta,* Vol. 500, pp. 185-194

[32] **Olivieri, A., C., Faber, N., M., Ferre, J., Boque, R., Kalivas, J., Mark, H., 2006.** Uncertainty estimation and figures of merit for multivariate calibration, *Pure Appl. Chem.,* Vol. 78, no. 3, pp. 633–661.

[33] **Vessman, J., Stefan, R. I., Staden, J. F., Danzer, K., Lindner, W., Burns, D., T., Fajgelj, A., Muller, A., 2001.** *Pure Appl. Chem.,* Vol.73, pp. 1381.

[34] **Brereton, R., G., 2000.** Introduction to multivariate calibration in analytical chemistry, Analyst, Vol. **125**, pp. 2125-2154.

[35] **Karaman, I., 2008,** Prediction of Extractive and Lignin Contents of Anatolian Black Pine (Pinus nigra Arnold. var pallasiana) and Turkish Pine (Pinus brutia Ten.) Trees Using Infrared Spectroscopy and Multivariate Calibration, *M.S. Thesis,* Izmir.

[36] **Miller, N., J., Miller, J., C., 2005.** Statistics and Chemometrics for Analytical Chemistry, 5th edition, *Pearson Education Limited,* UK.

[37] **Brereton, R., G., 2007.** Applied Chemometrics for Scientists, *Wiley,* UK.

[38] **Daszykowski, M., Serneels, S., Kaczmarek, K., Van Espen, P., Croux, C., Walczak, B., 2007.** TOMCAT: a MATLAB toolbox for multivariate calibration techniques, *Chemometrics and Intelligent Laboratory Systems,* Vol. **85**, pp.269-277.

[39] **Chung, H., Choi, H., Ku, M., 1999.** Rapid Identification of Petroleum Products by Near-Infrared Spectroscopy, *Bull. Korean Chem. Soc.,* Vol. **20**, no.9, pp. 1021-1025.

[40] **Andersson, M., 2009.** A comparison of nine PLS1 algorithms, *Journal of Chemometrics,* Vol. **23**, p.p 518-529.

[41] **Hall, S., 2014.** Implementation and Verification of a Robust PLS Regression Algorithm, *Master's Thesis in Engineering Mathematics and Computational Science,* Department of Mathematical Sciences Mathematical Statistics, Chalmers University of Technology, Sweden.

[42] **Jong de, S., 1993.** SIMPLS: an alternative approach to partial least squares regression, *Chemometrics and Intelligent Laboratory Systems,* Vol. 18, p.p 251-263.

[43] **Wold, S., Ruhe, A., Wold, H., Dunn III, W.J., 1984.** The collinearity problem in linear regression, The partial least squares approach to generalized inverses, *SIAM J. Sci. Stat. Comput*. Vol. **5**, p.p 735–743.

[44] **Wise, B., Gallagher, N. B., Bro, R., Shaver, J. M., Windig, W., Koch, R. S., 2006**. Chemometrics Tutorial for PLS_Toolbox and Solo, *Eiegenvector Research Inc.,* WA, USA.

[45] **Khanmohammadi, M., Garmarudi, A. B., DelaGuardia, M., 2012.** Characterization of petroleum based products by infrared spectroscopy and chemometrics, *Trends in Analytical Chemistry,* Vol. 35, p.p 135-149.

[46] **Felicio, C.C, Bras, L.P., Lopes, J.A., Cabrita, L., Menezes, J.C., 2005.** Comparison of PLS algorithms in gasoline and gas oil parameter

monitoring with MIR and NIR, *Chemometrics and Intelligent Laboratory Systems,* Vol.78, p.p 74-80.

[47] IP 143, 2001. Determination of asphaltenes (heptane insolubles) in crude petroleum and petroleum products, *Energy Institute,* London.

[48] Colaiocco, M.S., Farrera, M., 2008. Determination of asphaltene content in crude oil by attenuated total reflectance infrared spectroscopy and neural network algorithms, *Journal of Process Analytical Chemistry,* Vol.8-1, p.p 23-26.

[49] Wilt, B.K., Welch, W.T., 1998. Determination of Asphaltenes in Petroleum Crude Oils by Fourier Transform Infrared Spectroscopy, *Energy Fuels,* Vol.12(5), p.p 1008-1012.

[50] Aske, N., Kallevik, H., Sjoblom, J., 2001. Determination of Saturate, Aromatic, Resin, and Asphaltenic (SARA) Components in Crude Oils by Means of Infrared and Near-Infrared Spectroscopy, *Energy Fuels,* Vol.15(5), p.p 1304-1312.

[51] Balabin, R.M., Safieva, R.Z., 2007. Gasoline Classification by Source and Type based on Near Infrared (NIR) Spectroscopy Data, *Fuel,* Vol.87, p.p 1096-1101.

[52] Sivasankar, B., 2008. Engineering Chemistry, *Tata McGraw-Hill Publishing,* New Delphi.

[53] Albahri, T.A., Riazi, M.R., Alqattan, A.A., 2002. Octane number and Aniline point of petroleum products, *Kuwait University, Chemical Engineering Department, Fuel Chemistry Division Reprints,* Vol. 47(2), p.p 710-711.

[54] Motor Gasolines Technical Review, 2009. *Chevron Corporation,* USA.

[55] Air Quality Evaluation and Management Directive (Hava Kalitesi Degerlendirme ve Yonetimi Yonetmeligi), 2008. *Resmi Gazete,* No. 26898.

[56] Fundamentals of Gas Chromatography, 2002. Agilent Technologies Inc., USA.

[57] TS EN ISO 12185, 2007. Crude petroleum and petroleum products – determination of density – oscillating U-tube method, *Turkish Standard,* Ankara.

[58] ASTM D 4052, 2011. Standard Test Method for Density, Relative Density, and API gravity of Liquids by Digital Density Meter, *ASTM,* USA.

[59] Pomerantsev, A.L., Rodionova, O.Y., 2012. Process Analytical Technology: a critical view of the chemometricians, *Journal of Chemometrics,* Vol. 26, p.p 299-310.

[60] Pell, R.J., Seasholtz, M.B., Beebe, K.R., Koch, M.V., 2013. Process Analytical Chemistry and chemometrics, Bruce Kowalski's legacy at The Dow Chemical Company, *Journal of Chemometrics,* Special issue article.

[61] Seasholtz, M.B., 1999. Making money with chemometics, *Chemometrics and Ingtelligent Laboratory Systems,* Vol. 45, p.p 55-63.

[62] Miletic, I., Quinn, S., Dudzic, M., Vaculik, V., Champagne, M., 2004. An industrial perspective on implementing on-line applications of multivariate statistics, *Journal of Process Control,* Vol. 14, p.p 821-836.

[63] Zachariassen, C.B., Larsen, J., Berg, F., Engelsen, S.B., 2005. Use of NIR spectroscopy and chemometrics for on-line process monitoring of ammonia in Low Methoxylated Amidated pectin production, *Chemometrics and Ingtelligent Laboratory Systems,* Vol. 76, p.p 149-161.

[64] Analysis of Gasoline Blends and Finished Gasoline using Reformulyzer® M4 according to ASTM D6839 & EN ISO 22854, *Application Note 00.00.239 2013/1*, PAC, retrieved from https://b2b.paclp.com/HTML/item_master/links/m4%20Gasoline%20mode.pdf.

[65] SMS 2965, 2008. Development and Validation of Multivariate Calibration Models for Quantitative Near-Infrared Analysis**,** *Shell Method Series,* Netherlands.

[66] BP Statistical Review of World Energy 2015, June 2015, *BP*, UK.

Url-1 <*http://geosci.uchicago.edu/~moyer/GEOS24705/2011/*>, date retrieved 20.12.2014.

Url-2 < *http://www.uop.com/refining-flowscheme-2/*>, date retrieved 20.12.2014.

Url-3 < *http://en.wikipedia.org/wiki/Beer%E2%80%93Lambert_law*>, date retrieved 15.3.2014.

Url-4 <*http://wiki.eigenvector.com/index.php?title=T-Squared_Q_residuals_and_Contributions*>, date retrieved 29.04.2015.

Url-5 < *https://en.wikipedia.org/wiki/Asphaltene*>, date retrieved 31.10.2015.

Url-6 < *http://wiki.eigenvector.com/index.php?title=Pcr*>, date retrieved 11.11.2015.

Url-7 < *http://wiki.eigenvector.com/index.php?title=Pls*>, date retrieved 11.11.2015.

Url-8 < *http://wiki.eigenvector.com/index.php?title=Crossval*>, date retrieved 11.11.2015.

Url-9 < *http://www.chromedia.org*>, date retrieved 14.11.2015.

Url-10 < *https://public.wasson-ece.com/applications/?p=351* >, date retrieved 14.11.2015

Url-11 < *https://en.wikipedia.org/wiki/Near-infrared_spectroscopy*>, date retrieved 14.11.2015.

Url_12<*http://www.fda.gov/AboutFDA/CentersOffices/OfficeofMedicalProductsandTobacco/CDER/ucm088828.htm*>, date retrieved 15.11.2015.

Url_13< *http://www.waukeshacfr.com/f1-f2/*>, date retrieved 15.11.2015.

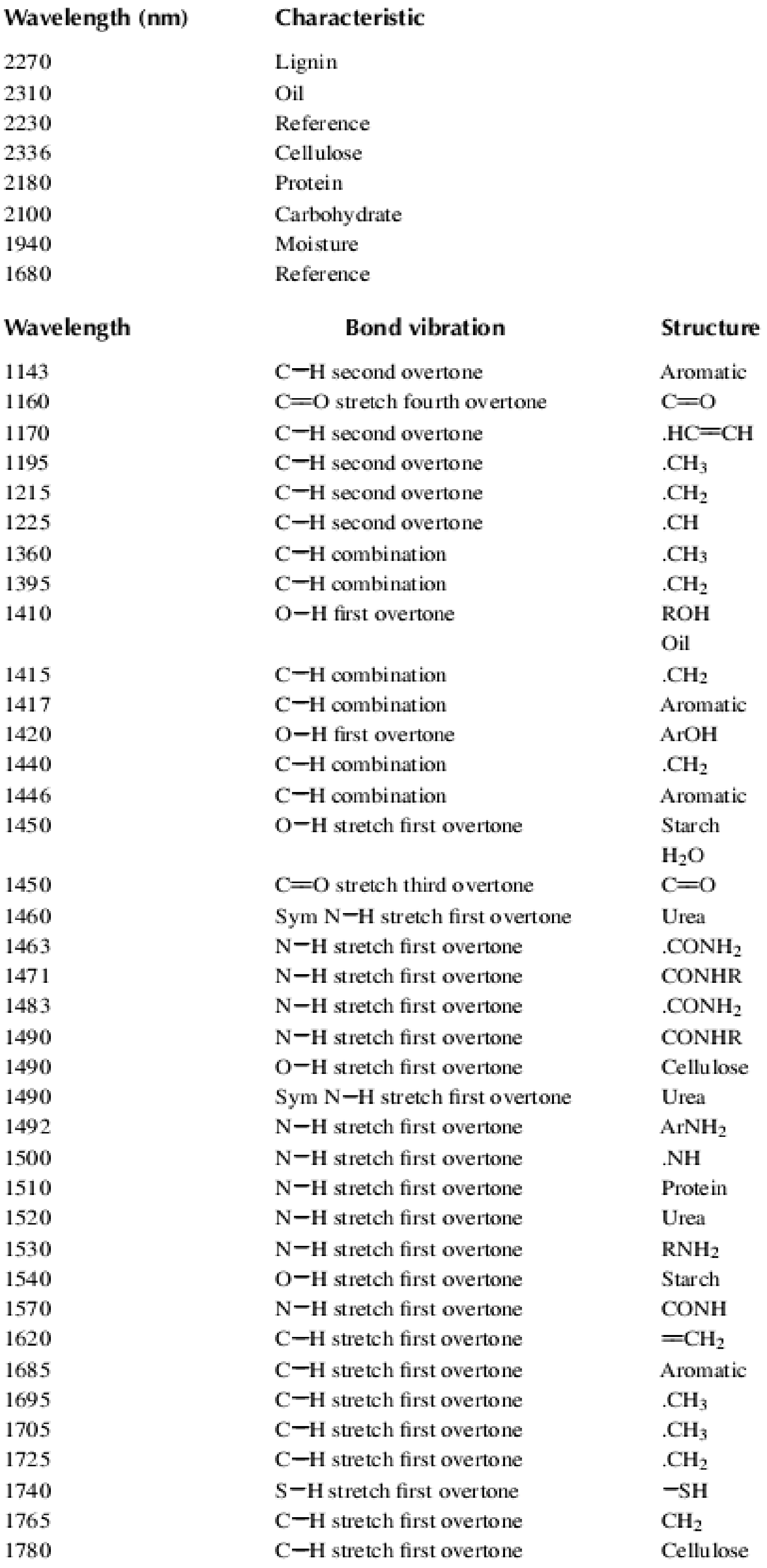
APPENDICES

APPENDIX A: Tables

APPENDIX B: Figures

APPENDIX A: Tables

Table A.1 : Characterizing wavelengths in the NIR region [16].



Wavelength (nm)	Characteristic
2270	Lignin
2310	Oil
2230	Reference
2336	Cellulose
2180	Protein
2100	Carbohydrate
1940	Moisture
1680	Reference

Wavelength	Bond vibration	Structure
1143	C—H second overtone	Aromatic
1160	C=O stretch fourth overtone	C=O
1170	C—H second overtone	.HC=CH
1195	C—H second overtone	.CH_3
1215	C—H second overtone	.CH_2
1225	C—H second overtone	.CH
1360	C—H combination	.CH_3
1395	C—H combination	.CH_2
1410	O—H first overtone	ROH Oil
1415	C—H combination	.CH_2
1417	C—H combination	Aromatic
1420	O—H first overtone	ArOH
1440	C—H combination	.CH_2
1446	C—H combination	Aromatic
1450	O—H stretch first overtone	Starch H_2O
1450	C=O stretch third overtone	C=O
1460	Sym N—H stretch first overtone	Urea
1463	N—H stretch first overtone	.$CONH_2$
1471	N—H stretch first overtone	CONHR
1483	N—H stretch first overtone	.$CONH_2$
1490	N—H stretch first overtone	CONHR
1490	O—H stretch first overtone	Cellulose
1490	Sym N—H stretch first overtone	Urea
1492	N—H stretch first overtone	$ArNH_2$
1500	N—H stretch first overtone	.NH
1510	N—H stretch first overtone	Protein
1520	N—H stretch first overtone	Urea
1530	N—H stretch first overtone	RNH_2
1540	O—H stretch first overtone	Starch
1570	N—H stretch first overtone	CONH
1620	C—H stretch first overtone	=CH_2
1685	C—H stretch first overtone	Aromatic
1695	C—H stretch first overtone	.CH_3
1705	C—H stretch first overtone	.CH_3
1725	C—H stretch first overtone	.CH_2
1740	S—H stretch first overtone	—SH
1765	C—H stretch first overtone	CH_2
1780	C—H stretch first overtone	Cellulose

Wavelength	Bond Vibration	Structure
1780	C—H stretch/HOH deformation combination	Cellulose
1790	O—H combination	H_2O
1820	O—H stretch/C—O stretch second overtone combination	Cellulose
1860	C—Cl stretch sixth overtone	C—Cl
1900	C═O stretch second overtone	$-CO_2H$
1908	O—H stretch first overtone	P—OH
1920	C═O stretch second overtone	CONH
1930	O—H stretch/HOH deformation combination	Starch Cellulose
1940	O—H bend second overtone	H_2O
1950	C═O stretch second overtone	$-CO_2R$
1960	O—H stretch/O—H bend combination	Starch
1980	Asym N—H stretch/amide II[b] combination	$CONH_2$
1990	N—H stretch/N—H bend combination	Urea
2030	C═O stretch second overtone	Urea
2050	N—H/Amide II[b] or	CONH
	N—H/Amide III[b] or combination	$CONH_2$
2055	Sym N—H stretch/amide I[b] combination	Protein
2060	N—H bend second overtone or N—H bend/N—H stretch combination	Protein
2070	N—H deformation overtone	Urea
2070	O—H combination	Oil
2090	O—H combination	.OH
2100	O—H bend/C—O stretch combination	Starch
2100	Asym C—O—O stretch third overtone	Starch or cellulose
2140	C—H stretch/C═O stretch combination or sym C—H deformation	? Oil NC═CH
2170	Asym C—H stretch/C—H deformation combination	HC═CH
2180	N—H bend second overtone	Protein
	C—H stretch/C═O stretch combination	Protein
	C═O stretch/amide III[b] combination	Protein
2200	C—H stretch/C═O stretch combination	$-CHO$
2270	O—H stretch/C—O stretch combination	Cellulose
2280	C—H stretch/CH_2 deformation	Starch
2300	C—H bend second overtone	Protein
2310	C—H bend second overtone	Oil
2322	C—H stretch/CH_2 deformation combination	Starch
2330	C—H stretch/CH_2 deformation combination	Starch
2335	C—H stretch/C—H deformation	Cellulose
2352	CH_2 bend second overtone	Cellulose Protein
2380	C—H stretch/C—C stretch combination	Oil
2470	C—H combination	$.CH_2$
2470	Sym C—N—C stretch first overtone	Protein
2488	C—H stretch/C—C stretch combination	Cellulose
2500	C—H stretch/C—C and C—O—C stretch	Starch
2530	Asym C—N—C stretch first overtone	Protein

[a] Original work.

[b] Amide I: C═O stretch. Amide II: N—H in-plane bend; C—N stretch. Amide III: N—H in-plane bend; C—N stretch.

Table A.2 : Gasoline samples reference analysis results.

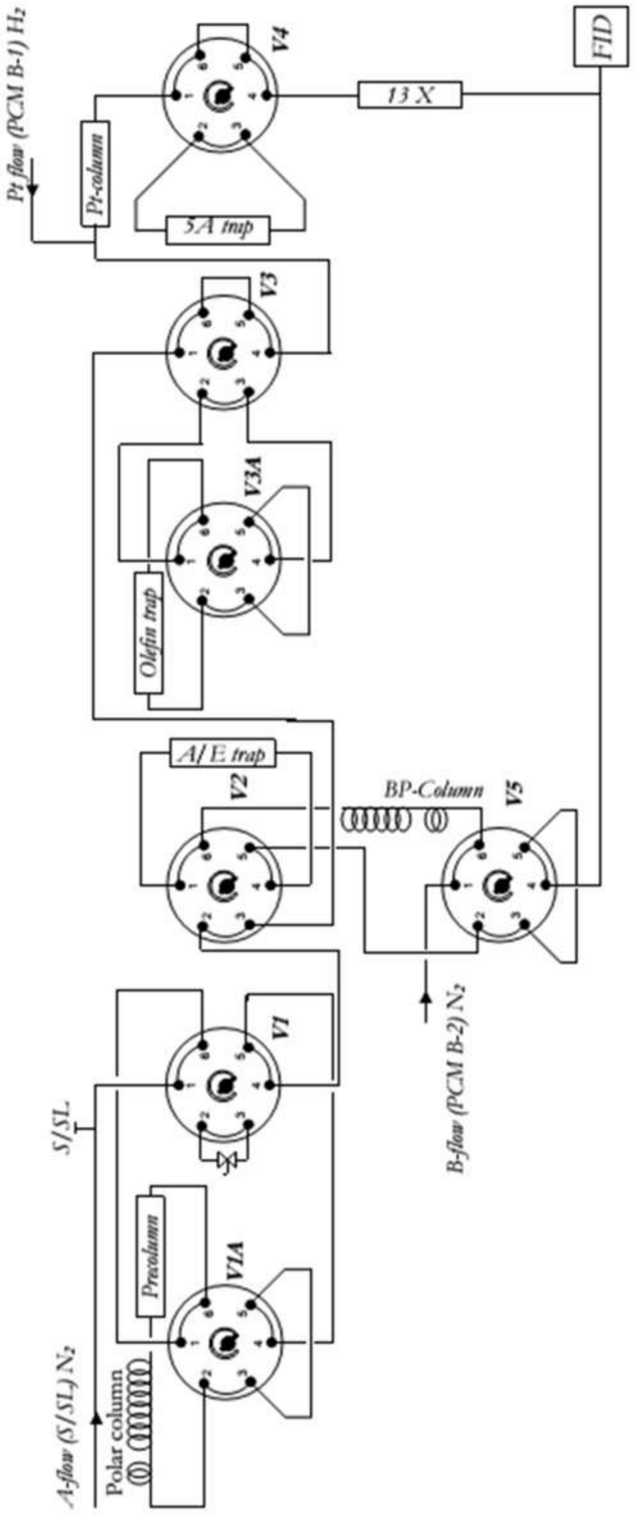
	RON	MON	Aromatics , %(v/v)	Olefins , %(v/v)	Benzene , %(v/v)	Density , kg/l	Distillation - E70, deg C	Distillation E100, deg C	Distillation E150, deg C
Sample 1	95.1	86	33.2	2	0.76	0.7336	45.2	64.3	92.7
Sample 2	95.2	85.9	31.5	5.7	0.75	0.7328	45.4	64.5	91.9
Sample 3	95.2	86.2	29.7	5.7	0.81	0.7311	47.3	66.7	92.3
Sample 4	97.2	86.2	34.8	8.3	0.74	0.7466	37.7	59.9	90.5
Sample 5	95.2	85.5	38.3	3.8	0.87	0.7462	34	56.8	89.6
Sample 6	95.2	86	33.6	4.2	0.8	0.7387	41.7	62.1	91.1
Sample 7	95	86.1	33.7	2.8	0.8	0.7391	41.3	63	91.2
Sample 8	95.1	86.1	33.5	3.1	0.84	0.7372	42.3	63.1	91.4
Sample 9	95.1	85.3	31.9	6	0.73	0.7395	39.6	62.3	91.3
Sample 10	95.1	85.1	37.9	4.8	0.75	0.7473	34	55.7	89.9
Sample 11	97	86.8	34.2	2.8	0.77	0.7405	42.9	63	91.9
Sample 12	95.3	85.3	34.5	8.1	0.75	0.742	38.3	59.2	89.3
Sample 13	95.1	85.3	34.4	9	0.92	0.7419	37.1	58.8	89.7
Sample 14	95.2	85.9	34	5.8	0.86	0.7398	39.1	61.3	91.1
Sample 15	95.2	85.8	33.6	5.9	0.86	0.7394	40.1	61	90.5
Sample 16	97.4	86.3	34.3	5.4	0.57	0.7462	40.2	61.5	91.1
Sample 17	95	85.6	33	5.6	0.57	0.7386	40.6	61.8	91.7
Sample 18	97.2	86.3	33.8	5.7	0.69	0.7465	38.7	61.4	90.9
Sample 19	95.2	86.1	33.6	3.9	0.72	0.7396	42	62.5	91
Sample 20	95.1	86.3	32.7	2.9	0.95	0.736	44.8	64.9	92
Sample 21	95.2	86	30.4	6.1	0.7	0.7337	45.7	64.7	90.4
Sample 22	95.2	85.2	32.9	9.6	0.94	0.7398	39.5	60.3	90.1
Sample 23	97	86	33.8	9.8	0.77	0.7481	34.5	58.7	89.8
Sample 24	95.4	85.6	34.9	5.1	0.87	0.7429	38.3	60.2	90.3
Sample 25	97.2	86.2	33.3	5.4	0.7	0.7455	39.5	61.6	91.2
Sample 26	95.2	85.6	35.9	7.1	0.82	0.7428	38.6	58.7	91.2
Sample 27	95	86.2	30	4.5	0.71	0.7291	48.6	66.8	92.7
Sample 28	95.1	85.7	37	5.3	0.84	0.7439	37.9	58	91.4
Sample 29	95.2	85.8	37.2	3.6	0.83	0.7428	39.4	58.6	91.5
Sample 30	95.1	85.9	30.9	5.8	0.72	0.7307	46.6	65	92.1
Sample 31	95.3	85.9	30.9	7.6	0.72	0.7323	45.3	64.3	92.1
Sample 32	95.1	85.9	31.7	6.2	0.75	0.7332	44	63.9	92
Sample 33	95.1	85.8	31	6.3	0.72	0.731	46.1	65	92.3
Sample 34	97.2	86.9	33.6	3.1	0.77	0.7382	44.7	63.9	92.5
Sample 35	95.5	86.4	32.1	5.1	0.73	0.7327	46.1	64.8	92.6
Sample 36	95.1	86.2	31	7.4	0.7	0.7308	47.6	66.2	92.9
Sample 37	95.3	85.8	35.3	7.9	0.82	0.7421	38.9	58.8	91.4
Sample 38	94.9	85.6	37.2	3.9	0.88	0.7432	38.7	58.8	91.5
Sample 39	95.3	86.1	39.9	1.2	0.94	0.7483	36.5	56.5	90.8
Sample 40	95.2	85.7	39.2	4.9	0.92	0.7491	33.8	55.3	89.7
Sample 41	95.2	85.6	32.7	6.4	0.77	0.7407	39.3	61.6	91

Table A.2 cont. Gasoline samples reference analysis results.

Sample 42	95.1	85.5	34.2	8.4	0.64	0.743	37.6	58.5	90.7
Sample 43	97.2	86.2	33.3	5.4	0.7	0.7455	39.5	61.6	91.2
Sample 44	95.4	85.6	33	4.2	0.69	0.7392	42.7	62.6	91.1
Sample 45	95	85.3	33.1	7.2	0.79	0.7399	38.6	60.7	89.9
Sample 46	95	85.6	33.8	3.7	0.73	0.74	42.6	62.6	90.9

APPENDIX B: Figures

Figure B.1 Flow Diagram [64].



CURRICULUM VITAE

Name Surname: Ümit AYNA

Place and Date of Birth: ISTANBUL - 09.11.1980

E-Mail: umitayna@gmail.com

EDUCATION:

B.Sc.: Bilkent University, Faculty of Science, Department of Chemistry

PROFESSIONAL EXPERIENCE AND REWARDS:

Process Analyzers Superintendent - Izmit Refinery - Turkish Petroleum Refineries Corporation

Laboratory Superintendent – Izmit Refinery - Turkish Petroleum Refineries Corporation

Process Analyzer Supervisor - Izmit Refinery – Turkish Petroleum Refineries Corporation

Laboratory Supervisor – Izmit Refinery – Turkish Petroleum Refineries Corporation

Chemist – Izmit Refinery – Turkish Petroleum Refineries Corporation

Copenhagen University, Department of Chemistry – Copenhagen – Denmark

Internship and Research study with Biophysical NMR Group

Intern - Analytical Chemistry Laboratory – TUBITAK (The Scientific and Technical Research Council of Turkey)

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

OTHER PUBLICATIONS, PRESENTATIONS AND PATENTS:

Jensen, M.R., Hansen, F.D., Ayna, U., Dagil, R., Hass, M.A.S., Christensen, H.E.M., Led, J.J., 2006. On the use of pseudocontact shifts in the structure determination of metalloproteins, *Magnetic Resonance in Chemistry,* Vol.44. p.p 294-301.