Technical
University
of Crete

DECONVOLUTION OF IR SIGNALS FROM OIL CHARACTERIZATION AND PREDICTION OF PROPERTIES

Michalopoulos Alessandro

A thesis submitted to the faculty at the Technical University of Crete in partial fulfillment of the requirements for the Master of Science in Petroleum Engineering at the Mineral Resources Engineering Department.

Chania 2020

Approved by:

Nikos Pasadakis

Vasileios Gaganis

Drosos Kourounis

ABSTRACT

Michalopoulos Alessandro:

(Under the direction of Nikos Pasadakis)

In order to interpret a Fourier Transform Infrared (FTIR) signal in its entirety, thousands of peaks would have to be identified requiring a significant amount of samples making the process not just expensive but actually impossible. The aim of this research is to instead focus on specific spectral bands in order to significantly reduce the number of required samples for interpretation.

A new algorithm is proposed to deconvolve the infrared spectrum of complex hydrocarbon mixtures in the 3000-2750 cm^{-1}, 1400-1330 cm^{-1} and 1000-700 cm^{-1} regions. The algorithm is developed based on the analysis of FTIR spectra of 33 oil fractions.

The experimentally derived spectra are deconvolved by fitting 5, 2 and 7 Lorentzian distributions, corresponding to aliphatic C-H stretching vibrations, aliphatic C-H scissoring/symmetric deformation vibrations and aromatic C-H out-of-plane bending vibrations for each of the above regions respectively.

The distribution of chemical structures in the oils were extracted, and correlations were studied between them and the measured saturates and aromatics percentage concentrations of the samples through the use of a linear regression model.

The validity of the results was interpreted by the Root Mean Square Error of Prediction (RMSEP) and through their comparison with the errors calculated from a previous student thesis. The reliability of the selected peaks is backed up by the small calculated NLP errors and the fitting of the first and second derivatives between the composite curve and the FTIR signal.

The algorithm facilitates the spectra modeling and the accurate estimation of the fitted methyl and methylene peak areas, which can be used for calculating specific compositional parameters of oil samples instead of the usually employed peak heights. Such modeling is extremely important for heavy petroleum fractions, where detailed compositional information is difficult to be obtained.

The results of the research showed that the connection between the peaks and the concentration values becomes clear enough through the usage of this procedure which, with the appropriate corrections, yields low RMSEP errors.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. What is FTIR spectroscopy?

In infrared spectroscopy, FTIR represents Fourier Transform Infrared, where IR radiation goes through a sample and the vibration characteristics of chemical compounds in it are detected. The range of the IR spectrum used depends on what the researcher is looking for in terms of compounds. Generally speaking it can be any wavenumber in the mid-infrared spectrum, between 4000-200 cm^{-1}. More information can be found in Table 1.

During the process, a portion of the infrared radiation gets consumed by the sample and some of it goes through. The generated graph represents the molecular absorption, creating an imprint of the example for which no two unique molecular structures can produce the same infrared spectrum. This characteristic makes infrared spectroscopy very useful for the detection of various compounds.

In general, through FTIR analysis it is possible to:

- identify unknown materials
- determine the quality or consistency of a sample
- determine the number of components in a mixture

Table 1, part of the FTIR spectrum and compounds detected in it

Wavelength (cm^{-1})	Curve Shape	Detection
3300	strong	≡C-H stretching
3100-3000	medium	=C-H stretching
3000-2800	strong	-C-H stretching
2830-2695	medium	-C-H stretching
1470-1400	medium	bending (CH_2), asymmetrical bending (CH_3)
1380	strong-medium	symmetrical bending (CH_3)
1300-600	medium-weak	stretching (CC), alicyclics, aliphatic chains
995-985	strong	C=C bending
980-960	strong	C=C bending
895-885	strong	C=C bending
840-790	medium	C=C bending
730-665	strong	C=C bending
880 ± 20	strong	C-H bending
880 ± 20	strong	C-H bending
810 ± 20	strong	C-H bending
780 ± 20	strong	C-H bending
755 ± 20	strong	C-H bending
750 ± 20	strong	C-H bending

Note. Reprinted from Infrared and Raman Spectroscopy (p. 190-191), by B. Schrader, 1995, Germany: VCH. Copyright 1995 by "VCH Verlagsgescllschaft mbH".

The mid-infrared spectrum can be divided in four regions and the nature of a group frequency may generally be determined by the region in which it is located. The regions are generalized as follows: the X-H stretching region (4000-2500 cm^{-1}), the triple bond region (2500-2000 cm^{-1}), the double bond region (2000-1500 cm^{-1}) and the fingerprint region (1500-600 cm^{-1}). This thesis generally focuses on the first and fourth regions.

In the first region, C-H stretching bands from aliphatic compounds occur in the ranges of 3000-2850 cm^{-1} or 3100-3000 cm^{-1} if the C-H bond is nearby a double bond or aromatic ring.

Even though, for the most part, specific types of deformations of molecules can be assigned to a band range which is the case most of the times, sometimes though the vibrations can vary by hundreds of wavenumbers. This applies to most single-bond stretching frequencies and bending (skeletal) vibrations, which absorb in the 1500-600 cm^{-1} region. Due to a lot of frequencies being present in this region, individual identification is not a viable option but collectively the absorption bands aid in identification.

1.1 Theory of FTIR

FTIR has been the most broadly accepted procedure for materials analysis in the laboratory for more than seventy years. The infrared spectrum of a sample, as referenced above, can be viewed as having its very own unique mark, with absorption peaks that correspond to the vibration frequencies between the sample's materials' atomic bonds. Every material is made up from a singular mixture of particles, which implies that no two mixtures can exist with an identical infrared spectrum. The size of the peaks likewise, is proportionate to the quantity/concentration of each compound. Thus FTIR can produce data with respect to a great deal of the material's qualitative and quantitative properties.

FTIR is favored over dispersive or filter techniques for infrared spectral analysis for a few reasons:

- It is a non-destructive technique
- Offers measurement precision which requires no external calibration
- Fast scanning times
- The sensitivity can be increased by adding together one second scans to remove random noise
- It is mechanically simple with only one moving part

FTIR spectrometry was created so as to surmount the limitations created by the instruments that existed then with the fundamental issue being the long scanning process. So an answer was found that utilized a simple optical device called an interferometer. Interferometers produce a one of a kind sort of sign which uses the infrared frequencies previously needed for each sample. The overall signal can be generated rapidly, for the most part within mere seconds, significantly improving upon the time constriction, in contrast to other devices that would take minutes made this a huge improvement.

When IR radiation interacts with matter, chemical bonds get stretched, compressed and bent. Subsequently, a chemical functional group will in general adsorb IR radiation in a particular wavenumber range without being affected by the structure of the remainder of the molecule. The wavenumber positions, where functional groups adsorb are constant, without being affected by temperature, pressure, sampling, or change in the molecular structure throughout the rest of the particle. Hence, there exists a connection between the wavenumber position and the compound structure that can be utilized to distinguish a functional group of a sample. The accumulation of specific functional groups can be observed by these types of infrared bands, which are called group wavenumbers.

Early-stage IR instruments were characterized as dispersive types that use a prism or a grating monochromator. The dispersive instrument is characteristic of a slow scanning. A Fourier Transform Infrared (FTIR) spectrometer obtains infrared spectra by first collecting an interferogram of a sample signal with an interferometer, which measures all the infrared frequencies simultaneously. It then acquires and digitizes the interferogram, performs the Fourier Transformation (FT), and outputs the spectrum.

An interferometer utilizes a beamsplitter to split the incoming infrared beam into two optical beams. One beam reflects off of a flat mirror which is fixed in place and another beam reflects off of a flat mirror which travels a very short distance (a few millimeters) away from the beamsplitter. The two beams reflect off of their respective mirrors and are recombined when they meet together at the beamsplitter. The resulting signal is called an interferogram, which has every infrared frequency encoded into it.

The beam finally arrives at the detector and is then measured. The detected interferogram cannot be directly interpreted, and the aforementioned FT technique is used. The computer can apply the FT method and present an infrared spectrum, which plots absorbance versus wavenumber. All the above mentioned mechanisms can be seen in Figure 1.

When an interferogram gets properly processed, a single beam spectrum is generated. A single beam spectrum is a plot of raw detector response versus wavenumber. A single beam spectrum obtained without a sample is called a background spectrum, which is induced by the instrument and the environments. A background spectrum must always be run when analyzing samples by FTIR. Since this could cause erroneous measurements to be displayed a simple process for the removal of this noise is utilized to eliminate these contributions. The sample single beam spectrum must be normalized against the background spectrum utilizing a few simple calculations with the transmittance, the intensity measured due to the sample and the intensity measured due to the background spectrum. The final spectrum should be devoid of all instrumental and environmental contributions, and only present the features of the sample.

1.2 Advantages of FTIR

Some of the major advantages of FTIR over the dispersive technique include:

- Speed (all the frequencies are measured simultaneously)
- Sensitivity (more sensitive detectors, lower noise levels)
- Mechanical Simplicity (small chance of mechanical failure due to few moving parts)
- Internally Calibrated

These major advantages allow measurements made by FTIR to be accurate and reproducible making it a very reliable technique for positive identification of any sample. The sensitivity benefits enable identification of even the smallest of contaminants which makes FTIR an invaluable tool for quality control or quality assurance applications. Furthermore, through the usage of a multitude of software algorithms and the increased sensitivity and accuracy of FTIR detectors, have helped promote the usage of infrared for quantitative analysis.

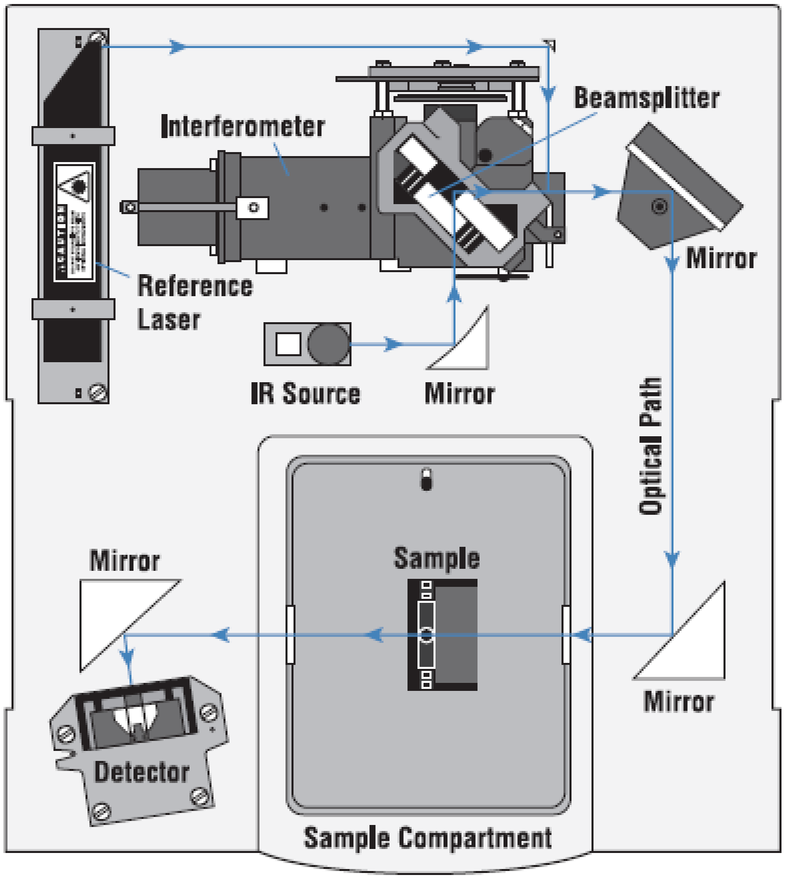


Figure 1, typical instrument overview of FTIR spectrometer

1.3 Oil characterization

When developing an oil fluid model in order to improve simulation, design, optimization, operation and prediction of petroleum processes, the only way to approach it is through oil characterization. It is the representation of desired oil properties through the study of other related measurements.

Because of the complex composition of oil, even regarding the lighter fractions, it is not possible to have an extensive analysis for all the compounds present in a sample due to increased time and cost. Hence, the correlation between other properties and the analyzed sample are utilized. Particularly, for heavy oil fractions similar to resins and asphaltenes, it is the sole manner of properties identification.

FTIR spectroscopy is one of the most widely employed methods in petroleum analysis given its ability to handle a wide variety of samples at any physical state, while still providing valuable information about its molecular structure in a quick and inexpensive manner. Additionally this technique has the ability to report on a sample's composition without affecting its "internal equilibrium", a characteristic frequently unavoidable when dealing with dynamic methods.

Any distillation cut is composed of a distribution of different chemical types and contains a varying amount of hydrocarbon structures and nitrogen, sulfur, and oxygen compounds. FTIR spectrometry can provide a quantitative determination of C−H bond types from very small samples and has often been used for the identification and characterization of organic compounds.

FTIR has been used for the chemometric correlation of properties in crude oils and petroleum products, such as diesel, fuel oils, and gasoline. Little research has been done though on the use of FTIR for property determination, while most of the work in this area has been focused on the H/C atomic ratio. The wide range of oil samples available has created the possibility to study into more detail the potential of FTIR for oil property classification.

The objective of this study is to study the possible existing correlations between the curve characteristics obtained from the FTIR analysis of a variety of fluids. The previously measured properties considered in this study are the concentrations of saturates and aromatics in the oil samples.

1.4 Important oil physical properties

Crude oil derives, by way of geological processing, from organic material initially buried in sediments at the bottom of ancient lakes and oceans. Crude oil formed at depth in a sedimentary basin migrates upward because of lower density. Many such migrations end with the oil collecting beneath a layer of impermeable rock, also referred to as a "trap," and forming a reservoir that can be tapped by drilling.

If the oil approaches the surface, it cools and comes in contact with groundwater. At the oil-water interface, anaerobic microorganisms degrade the oil in the absence of oxygen. The progressive loss of metabolizable molecules from the oil leads to an increase in viscosity and eventually results in a tarry residue that clogs the pores of the strata through which the oil had been migrating.

Over a long duration and with adequate sources of oil from below, enormous deposits of biodegraded oil residue can accumulate. This sequence is how the Alberta oil sands and other oil-sand deposits were formed. The substance that is then heated and removed from the rock or sand surface is called bitumen.

Bitumen and other crude oils contain a variety of substances with a range large enough that no two oil mixtures are the same. In general though, the different compounds present in the mixture can be divided into four basic groups. Saturated hydrocarbons, aromatic hydrocarbons, resins, and asphaltenes create the also known as SARA scheme.

Ideally, fluid properties such as bubblepoint pressure (P_b), solution gas/oil ratio (GOR), formation volume factor (FVF) and others are determined from laboratory studies designed to duplicate the conditions of interest.

However, experimental data are quite often unavailable because representative samples cannot be obtained or the producing horizon does not warrant the expense of an in-depth reservoir fluid study. In these cases, pressure-volume-temperature (PVT) properties must be determined by analogy or through the use of empirically derived correlations.

The calculation of reserves in an oil reservoir or the determination of its performance requires knowledge of the fluid's physical properties at elevated pressures and temperatures. Of primary importance are bubblepoint pressure, solution gas-oil ratio, and formation volume factor (FVF). In addition, viscosity and interfacial or surface tension must be determined for calculations involving the flow of oil through pipe or porous media.

Below the most important oil properties will be discussed along with relevant information. The key oil properties that are generally needed for understanding a reservoir and its producibility are:

- Bubblepoint pressure
- Solution gas oil ratio (GOR)
- Formation volume factor
- Viscosity
- Interfacial tension
- Isothermal compressibility

Oil bubblepoint pressure

In their original condition, reservoir oils include some natural gas in solution. The pressure at which this natural gas begins to come out of solution and form bubbles is known as the bubblepoint pressure.

Several studies provide statistical analyses for bubblepoint-pressure and solution GOR correlations and provide recommendations based on their findings; however, none of these references examines the full set of correlations.

Al-Shammasi compiled a databank of 1,243 data points from literature. This was supplemented by 133 samples available from a GeoMark Research database, bringing the total number of data points to 1,376. These data were then used to rank the bubblepoint pressure correlations.

The data were further grouped to examine the impact of crude oil gravity and GOR on the consistency of the correlations. Methods proposed by Lasater, Al-Shammasi, and Velarde et al. showed reliability over a wide range of conditions. The author has experienced good results from both the Standing and Glasø correlations, although they may not have ranked highly with this data set.

Solution gas oil ratio

The solution gas-oil ratio is a general term for the amount of gas dissolved in the oil. Heavy oils (lower API gravity) have lower capacity to contain dissolved gas than lighter oils. Solution GOR in black oil systems typically range approximately from 0 to 2000 scf / bbl. For most purposes, the solution GOR at the bubblepoint is the value of interest. At pressures above the bubble point pressure the oil is said to be undersaturated. Below the bubblepoint pressure, the gas begins to come out of solution and form a free gas phase, and the oil is said to be saturated.

Oil formation volume factor

The oil formation volume factor (FVF) relates the volume of oil at stock-tank conditions to the volume of oil at elevated pressure and temperature in the reservoir. Values typically range from approximately 1.0 bbl/STB for crude oil systems containing little or no solution gas to nearly 3.0 bbl/STB for highly volatile oils. For saturated systems, gas is liberated as pressure is reduced below the bubblepoint, this results in a corresponding shrinkage in oil volume.

Recent studies provide statistical analyses for bubblepoint oil FVF correlations and provide recommendations based on their findings; however, none of these references examines the full set of correlations. Al-Shammasi compiled a databank of 1,345 data points from the literature that was combined with 133 data points from the GeoMark

Research database to yield a total of 1,478 data points and these datasets were then used to rank the accuracy of the oil FVF correlation.

The data were further grouped to examine the impact of crude oil gravity and GOR on consistency of the correlations. Methods proposed by Al-Marhoun, Al-Shammasi, Farshad et al., and Kartoatmodjo and Schmidt showed reliability over a wide range of conditions. The author has experienced good results from both the Standing and Glasø correlations, although they may not have ranked highly with this data set.

Oil viscosity

Absolute viscosity provides a measure of a fluid's internal resistance to flow. For liquids, viscosity corresponds to the informal notion of "thickness". For example, honey has a higher viscosity than water.

Any calculation involving the movement of fluids requires a value of viscosity. This parameter is required for conditions ranging from surface gathering systems to the reservoir. Correlations for the calculation of viscosity can be expected to evaluate viscosity for temperatures ranging from 35 to 300°F.

The principal factors affecting viscosity are:

- Oil composition
- Temperature
- Dissolved gas
- Pressure

Typically, oil composition is described by API gravity only. The use of both the API gravity and the Watson characterization factor provides a more complete description of the oil. A characterization factor of 12.5 is reflective of highly paraffinic oils, while a value of 11.0 is indicative of naphthenic oil. Clearly, chemical composition, in addition to API gravity, plays a role in the viscosity behavior of crude oil. In general, viscosity characteristics are predictable. Viscosity increases with decreases in crude oil API gravity (assuming a constant Watson characterization factor) and decreases in temperature. The effect of solution gas is to reduce viscosity. Above saturation pressure, viscosity increases almost linearly with pressure.

Viscosity calculations for live reservoir oils require a multistep process involving separate correlations for each step of the process. Dead or gas-free oil viscosity is determined as a function of crude oil API gravity and temperature. The viscosity of the gas saturated oil is found as a function of dead oil viscosity and solution gas-oil ratio (GOR). Undersaturated oil viscosity is determined as a function of gas saturated oil viscosity and pressure above saturation pressure.

Interfacial tension

Interfacial or surface tension exists when two phases are present. These phases can be gas/oil, oil/water, or gas/water. Interfacial tension is the force that holds the surface of a particular phase together and is normally measured in dynes/cm. The surface tension between gas and crude oil ranges from near zero to approximately 34 dynes/cm. It is a function of pressure, temperature, and the composition of each phase.

Two forms of correlations for calculating gas/oil surface tension have been developed.

- The first form is a pseudocompositional black oil approach. Two components, gas and oil, are identified, and techniques used with compositional models are used to calculate surface tension.

- The second approach uses empirical correlations to determine surface tension.

Black oil correlations may provide less than accurate results because of the simplified characterization of the crude oil. Generally, the heavy end components of a crude oil may be made of asphaltic and surface active materials that have a measurable effect on surface tension.

Isothermal Compressibility of oil

Isothermal compressibility is the change in volume of a system as the pressure changes while temperature remains constant. The isothermal compressibility of undersaturated oil is defined as

$$c_o = -\frac{1}{V}\left(\frac{\partial V}{\partial p}\right)_T = -\frac{1}{B_o}\left(\frac{\partial B_o}{\partial p}\right)_T,$$

which reflects the change in volume with change in pressure under constant temperature conditions. Below the bubblepoint pressure, oil isothermal compressibility is defined from oil and gas properties to account for gas coming out of solution. The corresponding saturated oil compressibility is

$$c_o = -\frac{1}{B_o}\left[\left(\frac{\partial B_o}{\partial p}\right)_T - B_g\left(\frac{\partial R_s}{\partial p}\right)_T\right].$$

Above bubblepoint pressure, oil volume changes as a function of isothermal compressibility only. Oil formation volume factors (FVFs) for undersaturated crude oil are determined as a function of bubblepoint FVF, isothermal compressibility, and pressure above bubblepoint from

$$B_o = B_{ob} e^{\left[c_o (p_b - p)\right]}.$$

Other properties of interest

The concentrations in saturated and aromatic (BTEXs and PAHs) compounds can also have an effect on the properties of different oils. Mostly density and viscosity are correlated with the concentrations of the above compounds.

Saturated hydrocarbons are most abundant in light crude oils, which are the least dense and least viscous. Denser and more viscous crude oils have greater concentrations of other components, including resins and asphaltenes, which contain more polar compounds, often including “heteroatoms” of nitrogen, sulfur, and oxygen as well as carbon and hydrogen.

Under the anaerobic conditions prevailing during formation of the oil sands, the saturated hydrocarbons are mostly biodegradable, the aromatic hydrocarbons much less so, and the resins and asphaltenes not at all. A heavy crude, or the bitumen from an oil sand, is composed of the residue from a very protracted process whereby microbial action consumes most of the metabolizable saturates.

Therefore it can be mentioned that higher concentrations in saturated hydrocarbons mean lighter oils with usually smaller viscosities and densities, increased presence of aromatic hydrocarbons is generally interpreted as oils with higher values of viscosity and density, whereas increased amounts of resins and asphaltenes are mostly followed by even higher values.

Therefore it can be noticed that the ability to accurately predict any number of the above mentioned oil properties would drastically improve obtained results on the field or lab with a significant decrease in expenses by decreasing the required amount of samples for interpretation.

2. Regression Analysis

Regression analysis is the study of the correlation between a response (dependent) variable Y and one or more predictors (independent) variables X_n. When this relationship can be approximated by a straight line, it is said to be linear, and is therefore mentioned as linear regression. When the relationship follows a type of curve, it is then called a curvilinear regression.

To conduct a regression analysis a dependent variable needs to be defined first. This variable is supposed to be affected by one or more independent variables.

Subsequently a comprehensive dataset needs to be established in order to increase the reliability of the results that are produced in the end. If possible all the previously decided independent variables should be taken into account throughout the process.

The dependent variable should be plotted in the vertical y-axis whereas the independent ones should be plotted in the horizontal x-axis. At the end of the computing process a graph should be created similar to the ones in the figures of this chapter. A line should also be included within the plot that would represent the best explanation of the relationship between the dependent and independent variable.

In regards to this thesis, after having fitted the data points with Lorentzian curves, a connection is investigated between the characteristics of the curves (height and full width at half height) and the concentration of saturates and aromatic compounds separately using a linear regression model.

2.1 Linear Regression Models

The easiest example to comprehend for a regression model would be a straight line regression for just two variables Y and X through the following formula:

$$Y = B_0 + B_1 X \tag{1}$$

where, B_0 and B_1 are called parameters, which are known constants linking Y and X. B_0 is the y-intercept, B_1 is the slope. The relationship described in Equation 1 is exact. For a given X variable the Y variable can be precisely calculated. It's difficult to find such relationships in applied science though.

More often than not empirical approximations, created from observed data, are used. These kinds of relationships are represented as follows:

$$Y_i = B_0 + B_1 X_i \tag{2}$$

In this case Y_i and X_i are the i^{th} observed (known) values of the dependent and independent variable, respectively, whereas B_0 and B_1 are now the unknown parameter constants which must be estimated. In actuality, linear models include a wider range of models and not just the ones that are depicted from Equation 2. The main requirement is that the B coefficients of the model are linear themselves (e.g.$ln(Y_i) = B_0 + B_1 \, ln(X_i)$). An example of a linearly regressed sample can be seen in Figure 2.

As it can be easily seen a very small percentage of the data actually falls exactly on the line created by the regression model. As mentioned though, it can be helpful to create a first approximation like this since it’s easier to understand in general. There are more reasons that are also explained in the next section.

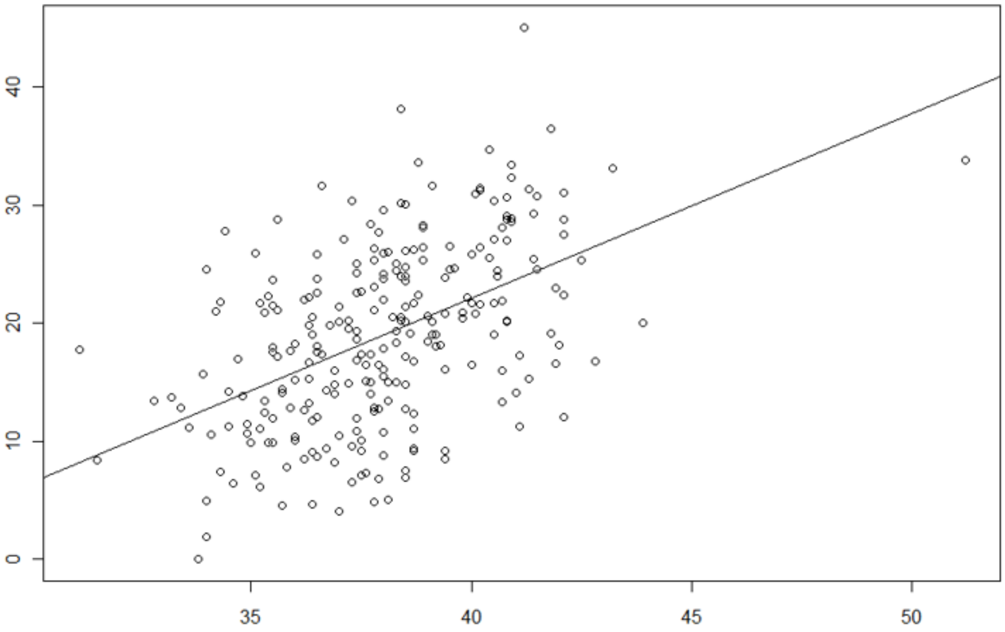


Figure 2, Linear Regression Model example

2.2 Nonlinear Regression Models

Nonlinear regression models are called those that do not have linear parameters in the first place, nor can they be linearized through transformation. An example of an additive model is show in Equation 3:

$$Y_i = B_0 e^{B_1(X_i)} \tag{3}$$

Linear models are generally preferred compared to nonlinear ones for a few main reasons. First and foremost, the linear model is mathematically easier to work with. The parameters can be calculated through explicit expressions. Nonlinear models are restricted to iterative schemes, which, more often than not, may converge to several solutions. Second, often the actual form of the relationship is not known and an approximation needs to be initially used making the linear model an obvious place to start.

An example for a nonlinear regression model can be seen below in Figure 3. It is clear that a straight line would not be able to correctly represent the data of such a sample; hence it would not be possible to create a very accurate model based on just a straight line.

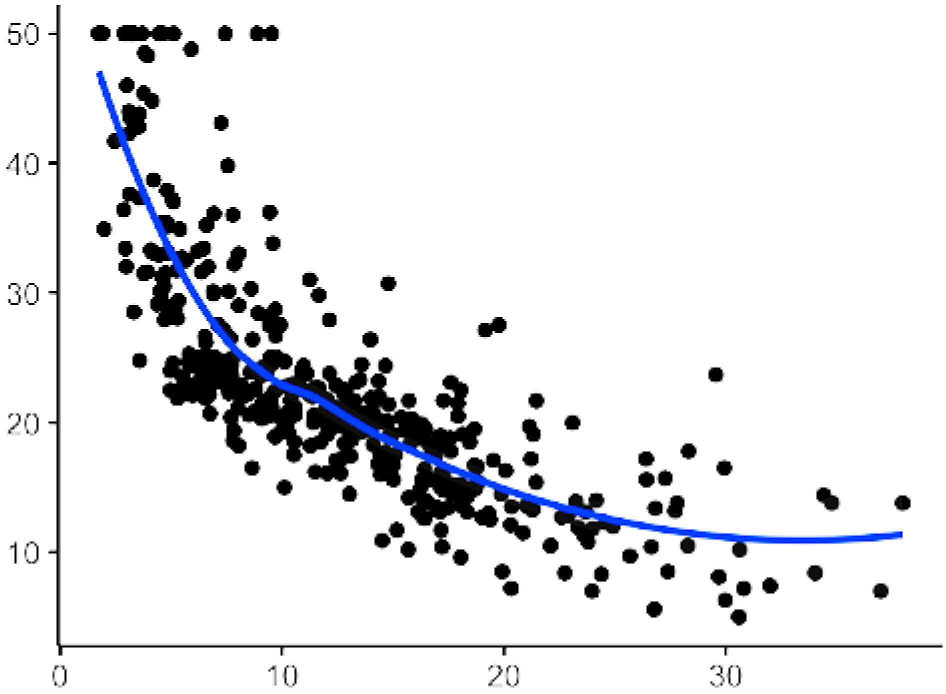


Figure 3, Nonlinear Regression example

2.3 Multiple Linear Regression (MLR)

MLR, also called multiple regression, is a statistical technique that, through the usage of multiple independent variables, predicts the outcome of a dependent variable. The goal of MLR is to model the linear relationship between these variables. Practically, this means that MLR is akin to the ordinary least squares (OLS) regression but with more independent variables involved.

The generic formula for MLR is shown in Equation 4:

$$Y_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + \cdots + B_n X_{in} \tag{4}$$

where, Y_i is the dependent variable, X_{in} is the independent variable, B_0 the y-axis intercept, B_n the slope coefficient for the respective independent variable.

In order to create an MLR model some basic assumptions need to be made:

- There is a linear relationship between the dependent and independent variables

- The independent variables are not too highly correlated with each other
- Y_i observations are selected independently and randomly from the population.
- Residuals should be normally distributed with a mean of 0 and variance σ.

2.4 Lorentzian Curve Fitting

Fitting a Lorentz or Cauchy function to given data points falls underneath the spectrum of nonlinear regression models. The general equation is described by Equation 5:

$$y = y_0 + 2\frac{a}{\pi}\frac{w}{4(x - x_c)^2 + w^2} \tag{5}$$

where, y and x are the two axis variables, y_0 is the y-values offset, x_c is the center of the distribution, a is the area and w is the full width at half height (FWHH) of the curve. An example of this type of model can be seen below in Figure 4.

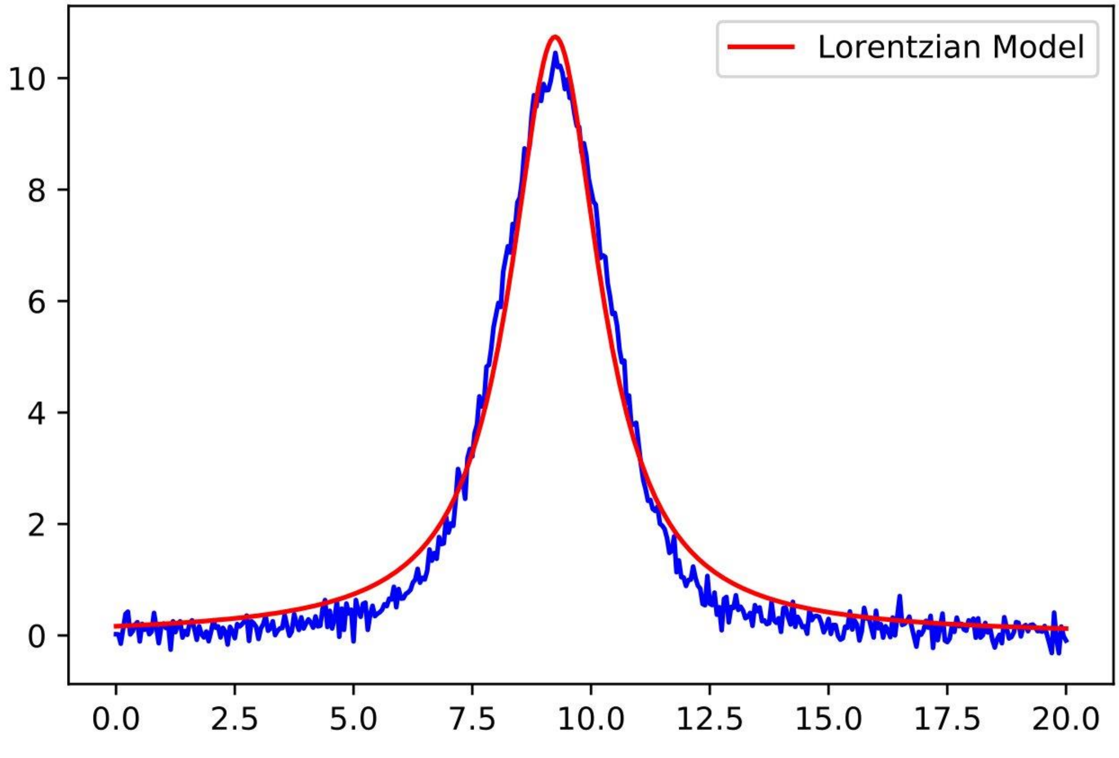


Figure 4, Lorentzian Regression example

It could then be assumed that the curves resulting from an FTIR analysis of a sample are able to be described and fitted by Lorentz curves by connecting the above equation variables with FTIR. The height of a peak (y_c) depends on the molecule concentration and its capability to absorb. In general, peak height is more often used to its ease of measurement. Peak fitting allows the area to be used, which can improve linearity of calibrations.

The location of a peak (x_c) depends on the natural vibrational frequency of the isolated molecule. In other words (as described in Chapter 1) this variable will have a different value depending on the molecular structure but should always have the same value for that structure. The peak location also depends upon interactions with the environment. If the molecule is hydrogen bound to neighbors it will spread the bond energy over more space, and the peak will red-shift. If the molecule is being repulsed, the peak blue shifts.

Line widths are reported as full width at half height. Despite width being an often overlooked parameter, it contains plenty of information within itself. Motion and energy loss affect the line width, with many theories being that different environments can have any effect on the line width. A primary example would be the relation of line width and the effective lifetime of excitation of a molecule τ. Swift decrease of the excitation (short τ) would result in broader peaks; an extended lifetime (long τ) would lead to narrower peaks. Hence, any action that would increase energy loss rates, e.g. molecule collisions, would result in the broadening of the peaks, whereas molecules in low pressure environments have reduced energy losses and therefore narrower peaks.

The reason for which Lorentzian curves are of interest in regards to this research is because it has been found that the Lorentzian profile works very well for liquids in many cases (Curve Fitting in Raman and IR Spectroscopy: Basic Theory of Line Shapes and Applications, Michael Bradley and Thermo Fisher n.d.)

As mentioned above, such a model was used for the first computational process of this thesis, created by Dr. Drosos Kourounis. The model created is based on the fitting of the given data points from the FTIR analysis of a number of samples for Lorentzian distributions.

The process created by Dr. Kourounis is much too complex to be explained in detail within this text and its contents belong to an entirely different scientific text altogether.

2.5 Error checking for a regression analysis

There is a multitude of tools to check the validity of a regression analysis. For the purpose of this research the following ones were used:

- Standard Error
- t-statistic
- P-value
- Root Mean Square Error of Prediction

The Standard Error or SE for short, of a regression calculates the absolute measure of the typical distance that the data points fall from the regression. It is an estimate of the standard deviation of the coefficient. It is very simple to calculate as it can be seen in equation 6:

$$SE = (Y_{obs} - Y_{pred}) \tag{6}$$

where Y_{obs} are the observed values and Y_{pred} are the values predicted by the model.

Because of the way it is calculated its given units are those of the variable of interest. The reason for which this error is used is because it gives a good first estimate of the distance between a data point and its predicted value through regression, but also because it can be used both in linear and nonlinear regressions.

Next is the t-statistic (t-stat). The t-stat is the calculated by dividing a coefficient by its SE. It describes how the mean of a sample with a certain number of observations is expected to behave although it can also be seen as a measure of the precision with which the regression coefficient is measured. The t statistic of the variable is then compared with values in the Student's distribution to determine the P-value.

As mentioned above, from the t-stat analysis the P-value is extracted. It examines the hypothesis that the coefficient is not affecting the dependent variable. A low p-value indicates that the hypothesis can be rejected, meaning that a predictor that has a low p-value is likely to play a meaningful role within a model. Conversely, a larger P-value suggests that changes in the predictor are not likely to be associated with changes in the response.

The Root Mean Square Error of Prediction (RMSEP) is the most commonly used unit in order to check the validity of a generated regression model. It is calculated as shown below in Equation 7:

$$RMSEP = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_{obs} - Y_{pred}\right)^{2}} \tag{7}$$

It is the standard deviation of the Standard Errors also called residuals. RMSEP is a measure of how spreads out those residuals are.

3. Multivariate analysis methods

Statistics are used in medicine for data description and inference. Inferential statistics are used to answer questions about the data, to test hypotheses (formulating the alternative or null hypotheses), to generate a measure of effect, typically a ratio of rates or risks, to describe associations (correlations) or to model relationships (regression) within the data and, in many other functions.

Usually point estimates are the measures of associations or of the magnitude of effects. Confounding, measurement errors, selection bias and random errors make unlikely the point estimates to equal the true ones. In the estimation process, the random error is not avoidable. One way to account for it is to compute p-values for a range of possible parameter values (including the null).

The range of values, for which the p-value exceeds a specified alpha level (typically 0.05) is called confidence interval. An interval estimation procedure will, in 95% of repetitions (identical studies in all respects except for random error), produce limits that contain the true parameters.

It is argued whether the pair of limits produced from a study contains the true parameter but could not be answered by the ordinary (frequentist) theory of confidence intervals. Frequentist approaches derive estimates by using probabilities of data (either p-values or likelihoods) as measures of compatibility between data and hypotheses, or as measures of the relative support that data provide hypotheses. Another approach, the Bayesian, uses data to improve existing (prior) estimates in light of new data. Proper use of any approach requires careful interpretation of statistics.

The goal in any data analysis is to extract from raw information the accurate estimation. One of the most important and common question concerning if there is statistical relationship between a response variable (Y) and explanatory variables (Xi). An option to answer this question is to employ regression analysis in order to model its relationship.

There are various types of regression analysis. The type of the regression model depends on the type of the distribution of Y; if it is continuous and approximately normal we use linear regression models; if dichotomous we use logistic regression; if Poisson or multinomial we use log-linear analysis; if time-to-event data in the presence of censored cases (survival-type) we use Cox regression as a method for modeling.

Modeling is the attempt to predict the outcome (Y) based on values of a set of predictor variables (Xi). These methods allow us to assess the impact of multiple variables (covariates and factors) in the same model.

3.1 MLR (Multivariate Linear Regression)

As the name implies, multivariate regression is a technique that estimates a single regression model with more than one outcome variable. When there is more than one predictor variable in a multivariate regression model, the model is a multivariate multiple regression.

A common problem in statistical analysis is the connection between a group of independent variables X within a group of dependent variables Y.

There practically are an infinite amount of ways to combine them (exponential, linear, polynomial, etc.) and are dependent on the type of data. For a greater amount of X variables, the process is called a Multiple Linear Regression. This should not be confused with the MLR term where multiple correlated variables are included in the process.

When the number of X variables is smaller than 20, MLR can give reliable results in general. It can also give unreliable results in case of strong correlations between the X variables.

MLR is a generalized form of the linear regression with the difference that it takes into consideration more than one independent variable.

3.2 PLS (Partial Least Squares)

In the PLS method the score values diagrams are studied for both X and Y variables and are depicted on a Score Values of X - Score Values of Y graph, created through the usage of weighted vectors.

The score values table is a dataset that is generated from variables that are linear combinations of the initial values.

It can be observed that as the orientation of the vectors changes then the structure of the previously mentioned diagram changes also. PLS rotates in such a way the weighted vectors so that the covariance of the X and Y score values gets maximized. The correlation of the variables is therefore maximized as well.

This way it is possible for a given X variable to calculate the score of X and through a previously generated scores graph to match it with the appropriate score of Y and so to predict the anticipated value of Y at that point. PLS is structurally similar to PCA but utilizes different criteria.

3.3 iPLS (Partial Least Squares with intervals)

iPLS has the ability to select a specific group of variables in order to localize the process in a specific data area so as not to have it contaminated by the rest of the sample.

This process calculates through the usage of PLS of multiple combinations of variables in order to select the group that would generate the least error for the cross-validation process it utilizes.

This technique, even though it picks variables at random, can improve the prediction error of the model. The risk in this case though is the possibility of leaving out of the calculation process important variables. By using a smaller amount of variables means that each of the selected variables increases in significance.

If any of those variables is removed there will be others that will take their place. This means that the detection of any errors becomes harder than before.

There is also a difficulty in the calculation and interpretation of the resulting errors from the iPLS process. Therefore when using such a method the final variables should always be double checked keeping in mind what the model needs to be appropriately interpreted.

3.4 Dimension Reduction

The process of reducing dimensions is utilized in order to study multidimensional data in an easier to comprehend method.

The aim in such cases is to edit the dataset so that the correlations within can be identified, as well as possible patterns that will lead to a statistical hypothesis, all while attempting to lose as little information as possible. The data is subsequently graphed on two dimensional plots while also getting clustered, smoothed, and classified and probability density diagrams are calculated.

In a system with a lot of variables that affect the studied properties, it would be incorrect to randomly remove data to decrease the dataset's dimensions since a lot of the information within the data would also be removed resulting in erroneous calculations. It would be possible though to create new variables as linear combinations of the initial variables in the case of a small dimension number.

There are many techniques for this like Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Non-negative Matrix Factorization, Factor Analysis, and Linear Discriminant Analysis.

3.5 PCR (Principal Component Regression)

PCR is one of the most used techniques in multifactorial statistical systems and is mainly based on the PCA and MLR methods. In such cases the variance and correlation of the variables for the given dataset are analyzed.

The data extracted from a Standard Linear Regression Model is analyzed and subsequently a PCA process is used to predict the unknown regression terms.

The PCA process can be executed through the decomposition of the eigenvalues of the data covariance, or from a correlation matrix, an SVD of the data table, after the standardization and smoothing or the use of z-scores of the data table for each property.

Whenever a dataset with too many dimensions is given, meaning they cannot be depicted on a simple two- dimensional (2D) or three-dimensional (3D) graph, the PCA technique should be applied.

In order for multidimensional data to be depicted on simple 2D or 3D plots, the dataset's dimensions have to be reduced so in a way where only the most important variables that affect the properties the most, while also minimizing information loss. In other words a new set is created that is generated from linear combinations of the initial variables.

If a vector "x" exists with enough "p" variables that are correlated, then it would be really time consuming to study the variables and all the correlations and co-variances. It would be simpler to just calculate the co-variance of every column from the dataset and selectively study the ones with the greater values. Those will be the ones majorly affecting the results while containing the most information for a PCA analysis.

While using PCA it is possible to discern data clusters with common characteristics in regards to a property. It is possible to analyze the data, understand its structure, to calculate the degree of co-linearity between the properties of the samples, to accentuate their differences and to find the various relations within the data. This kind of process is also called Classification and Discriminant Analysis.

PCA is used wherever it is possible to separate data groups with common properties. Just like PCA, Discriminant Function Analysis is based on the process of calculating the best possible linear combination of the initial variables.

Another kind of information that can be deduced from a PCA plot derives from the study of the principal components within the plot. In such cases it is important to note their position and whether they have positive or negative values.

PCR utilizes the principal components from PCA and applies Multiple Linear Regression (MLR). This is made possible by converting the initial table in a new set of principal component variables that have no correlation between them and are structured in a way where they will hold the most variance.

PCA can calculate the greatest variance within the data in regards to one dimension and arranges the eigenvectors of that direction in order to make the axes perpendicular to one another.

3.6 SVD (matrix analysis for particular values)

The PCA method can be correlated also with the SVD one, where the principal components of a dataset for a normal PCA process are projected.

A PCA analysis can be done from either an eigenvalue decomposition of a co-variance or data correlation matrix, by using the SVD method, through normalization, or the usage of z-scores. The results of PCA, meaning the principal components, are used for the creation of the PCA plot.

The SVD technique for the PCA process has a few drawbacks, since it could not take into account some variable properties like minimal differences and robustness.

4. Signal Deconvolution and Curve Fitting

4.1 What is Signal Deconvolution

The word "deconvolution" can have two meanings, which can lead to confusion. In the Oxford dictionary it is defined as a process of resolving something into its constituent elements or removing complication in order to clarify it, where it applies to Fourier deconvolution.

But the same word can also be sometimes used for the process of resolving or decomposing a set of overlapping peaks into their separate additive components by the technique of iterative least-squares curve fitting of a proposed peak model to the data set.

In mathematics, deconvolution is an algorithm-based process used to enhance signals from recorded data. Where the recorded data can be modeled as a pure signal that is distorted by a filter (a process known as convolution), deconvolution can be used to restore the original signal. The concept of deconvolution is widely used in the techniques of signal processing and image processing.

However, that process is actually conceptually distinct from Fourier deconvolution, since in Fourier deconvolution, the underlying peak shape is unknown but the broadening function is assumed to be known; whereas in iterative least-squares curve fitting it's just the reverse: the peak shape must be known but the width of the broadening process, which determines the width and shape of the peaks in the recorded data, is unknown.

Thus the term "spectral deconvolution" can be ambiguous: it might mean the Fourier deconvolution of a response function from a spectrum, or it might mean the decomposing of a spectrum into its separate additive peak components.

Fourier deconvolution is the converse of Fourier convolution in the sense that division is the converse of multiplication. If it is known that m times x equals n, where m and n are known but x is unknown, then x equals n divided by m. Conversely if you know that m convoluted with x equals n, where m and n are known but x is unknown, then x equals m deconvolved from n.

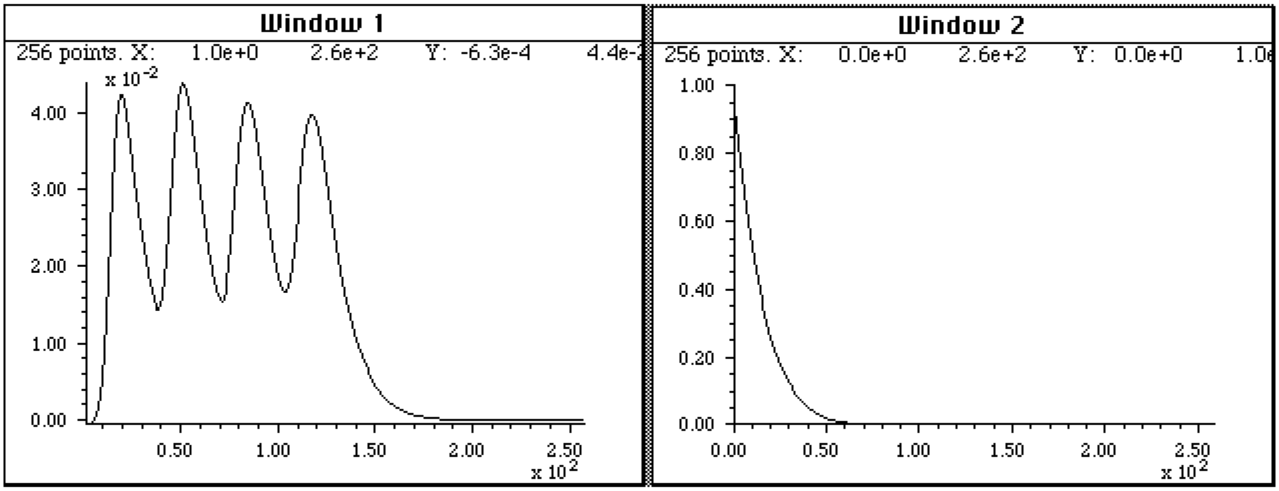
In practice, the deconvolution of one signal from another is usually performed by point-by-point division of the two signals in the Fourier domain, that is, dividing the Fourier transforms of the two signals point-by-point and then inverse-transforming the result. Fourier transforms are usually expressed in terms of complex numbers, with real and imaginary parts representing the sine and cosine parts. If the Fourier transform of the first signal is a + ib, and the Fourier transform of the second signal is c + id, then the ratio of the two Fourier transforms, by the rules for the division of complex numbers, is:

$$\frac{a+ib}{c+id} = \frac{ac+bd}{c^2+d^2} + i\frac{bc-ad}{c^2+d^2}$$

4.2 Practical significance of Signal Deconvolution

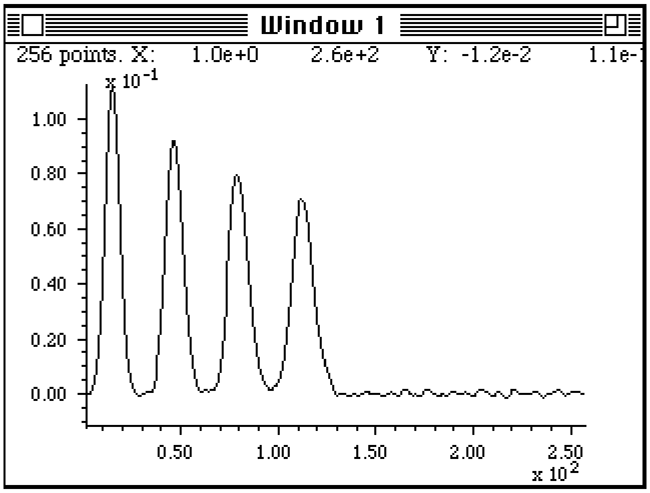
The practical significance of deconvolution in signal processing is that it can be used as a computational way to reverse the result of a convolution occurring in the physical domain, for example, in order to reverse the signal distortion effect of an electrical filter or of the finite resolution of a spectrometer.

In some cases the physical convolution can be measured experimentally by applying a single spike impulse ("delta") function to the input of the system, then that data used as a deconvolution vector. Deconvolution can also be used to determine the form of a convolution operation that has been previously applied to a signal, by deconvolving the original and the convolved signals. These two types of application of Fourier deconvolution are shown in the examples below in Figures 5 through 8.



```
256 points. X:    1.0e+0     2.6e+2     Y:  -6.3e-4      4.4e-2
256 points. X:    0.0e+0     2.6e+2     Y:   0.0e+0      1.0e
```

Figure 5, Instrument signal (left) and distorting influence of recorder signal (right)



```
256 points. X:    1.0e+0     2.6e+2     Y:  -1.2e-2      1.1e-
```

Figure 6, Deconvolved signal of example

Signal deconvolution here is used in order to remove the distorting influence of an exponential tailing response function from a recorded signal. The response function must be known and is usually either calculated on the basis of some theoretical model or is measured experimentally as the output signal produced by applying an impulse (delta) function to the input of the system.

The response function, with its maximum at x=0, is deconvolved from the original signal. The result shows a closer approximation to the real shape of the peaks; however, the signal-to-noise ratio is unavoidably degraded compared to the recorded signal, because the Fourier deconvolution operation is simply recovering the original signal before the low-pass filtering, including possible pre-existing noise.

It should be noted that this process has an effect that is visually similar to resolution enhancement, although the latter is done without specific knowledge of the broadening function that caused the peaks to overlap.

In this second example signal deconvolution a different application is presented that aims to reveal the nature of a data transformation function that has been applied to a dataset by the measurement instrument itself.

On the left of Figure 7 a UV-visible absorption spectrum recorded from a commercial photodiode array spectrometer is represented. The figure on the right of Figure 7 is the first derivative of this spectrum produced by an unknown algorithm in the software supplied with the spectrometer.

The signal on the left of Figure 8 is the result of deconvolving the derivative spectrum from the original spectrum. This therefore must be the convolution function used by the differentiation algorithm in the spectrometer's software. Rotating and expanding it on the x-axis makes the function easier to see (bottom right). Expressed in terms of the smallest whole numbers, the convolution series is seen to be +2, +1, 0, -1, -2. This simple example of "reverse engineering" would make it easier to compare results from other instruments or to duplicate these results on other equipment.

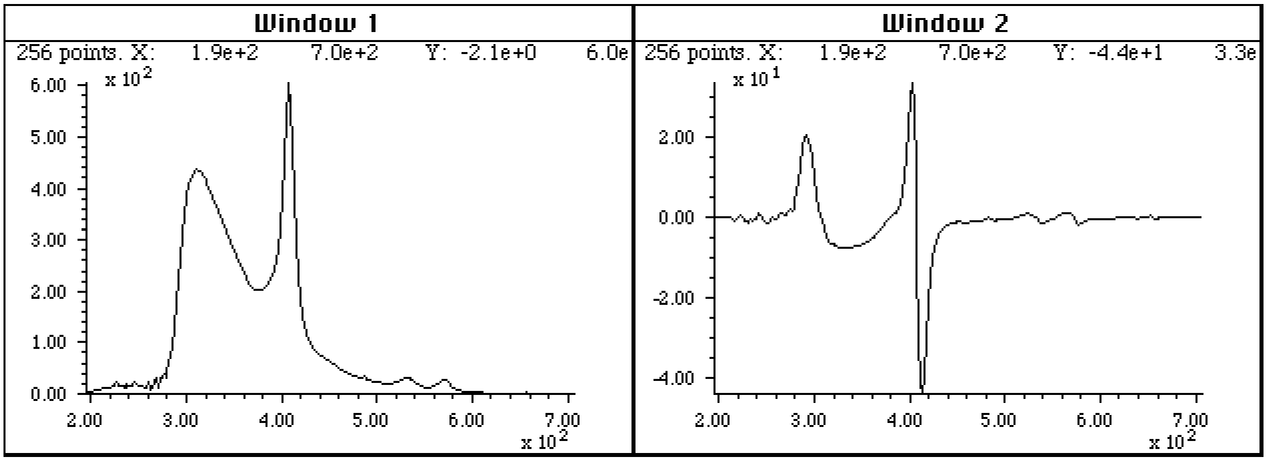


Figure 7, UV- visible wave absorption spectrum (left) and the first derivative of the spectrum (right)

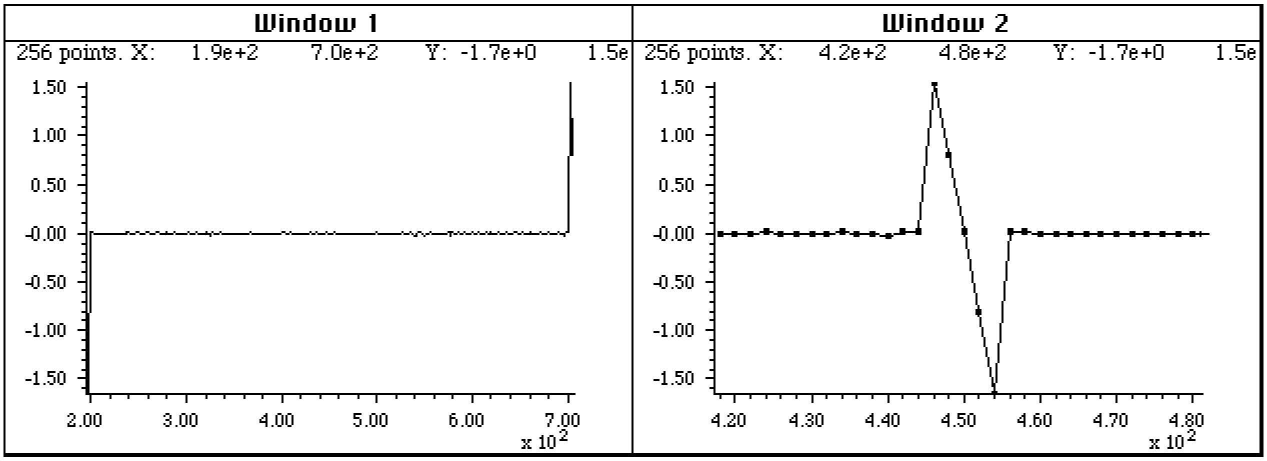


Figure 8, Deconvolution of the derivative spectrum with the original spectrum (left) and different representation of the same result (right)

When using signal deconvolution to experimental data, for example to remove the effect of a known broadening or low-pass filter operator caused by the experimental system, there are three serious problems that limit the utility of the method:

i. the convolution occurring in the physical domain might not be accurately modeled by a mathematical convolution

ii. the width of the convolution - for example the time constant of a low-pass filter operator or the shape and width of a spectrometer slit function - must be known, or at least adjusted by the user to get the best results

iii. a serious signal-to-noise degradation commonly occurs. Any noise added to the signal by the system after the convolution by the broadening or low-pass filter operator will be greatly amplified when the Fourier transform of the signal is divided by the Fourier transform of the broadening operator, because the high frequency components of the broadening operator (the denominator in the division of the Fourier transform) are typically very small, with some individual components often of the order of 10^{-12} or 10^{-15}, resulting in a huge amplification of those particular frequencies in the resulting deconvolved signal, which is called "ringing". The problem can be reduced either by low-pass filtering (smoothing) or even more simply by adding a small positive non-zero constant to the denominator, which increases the excessively small high-frequency members in the denominator without significantly increasing the much greater low-frequency members. Smoothing or filtering reduces the amplitude of the highest-frequency components, and denominator addition reduces the amplitude of the frequencies that are the most highly to be amplified by deconvolution. Both methods can have a similar effect, but they work in different ways and can sometimes be more effective when used together rather than separately.

In many cases, the width of the physical convolution is not known exactly, so the deconvolution must be adjusted empirically to yield the best results. Similarly, the width of the final smooth operation must also be adjusted for best results. The result will seldom be perfect, especially if the original signal is noisy, but it is often a better approximation to the real underlying signal than the recorded data without deconvolution.

5. Processing and Results

5.1 Signal Deconvolution

To accurately create a fluid model, so as to correctly simulate design and optimize its behavior, the oil should be properly characterized. In order to more easily characterize an oil, the properties that have to be studied need to be substituted with other observed values that can be directly correlated to the former. In this research the characterization is done by using data provided by an FTIR analyzer. Little research has been done in regards to physical property prediction through FTIR data.

Thanks to the extensive database available in this faculty's system, it is a chance to properly investigate the ability to utilize FTIR in oil characterization. Specifically, it is aimed at detecting possible relationships between the curve characteristics of the FTIR spectrum and the percent concentration of saturates and aromatics. In turn this would allow the, even if only roughly, prediction of the samples' physical properties allowing for a better simulation study in earlier production stages. As mentioned in a previous chapter three areas of FTIR spectra were measured from a variety of 33 oil samples. The spectral bands of interest are:

- 3000-2700 cm^{-1}
- 1400-1330 cm^{-1}
- 1000-700 cm^{-1}

Initially the algorithm developed by Dr. Kourounis is utilized to generate the Lorentzian distributions that best approach the FTIR datasets. Examples of these curves are shown in the Figures 9-11 below for each one of the spectral bands.

The 3 types of curves presented are sequentially:

- the algorithm generated Lorentzian distributions (colored), the their composite curve (solid black) and the FTIR generated curve (dashed blue)
- The algorithm's fit on the first derivative of the FTIR curve
- The algorithm's fit on the second derivative of the FTIR curve

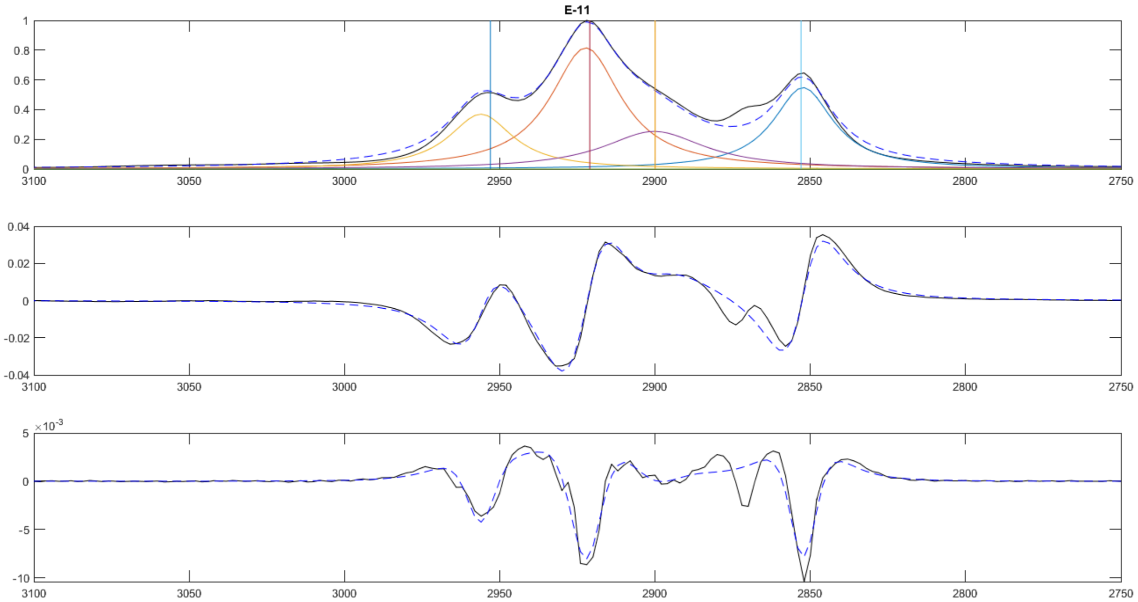


Figure 9, Fitting of Lorentzian distribution for the 3000-2700cm^{-1} band of sample E-11

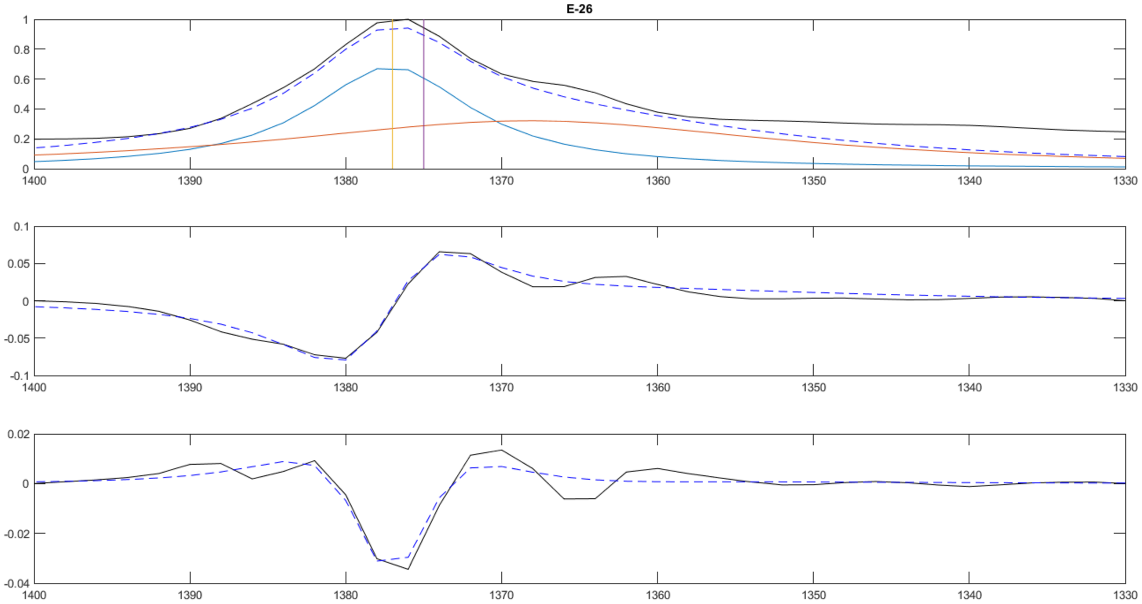


Figure 10, Fitting of Lorentzian distribution for the 1400-1330cm^{-1} band of sample E-26

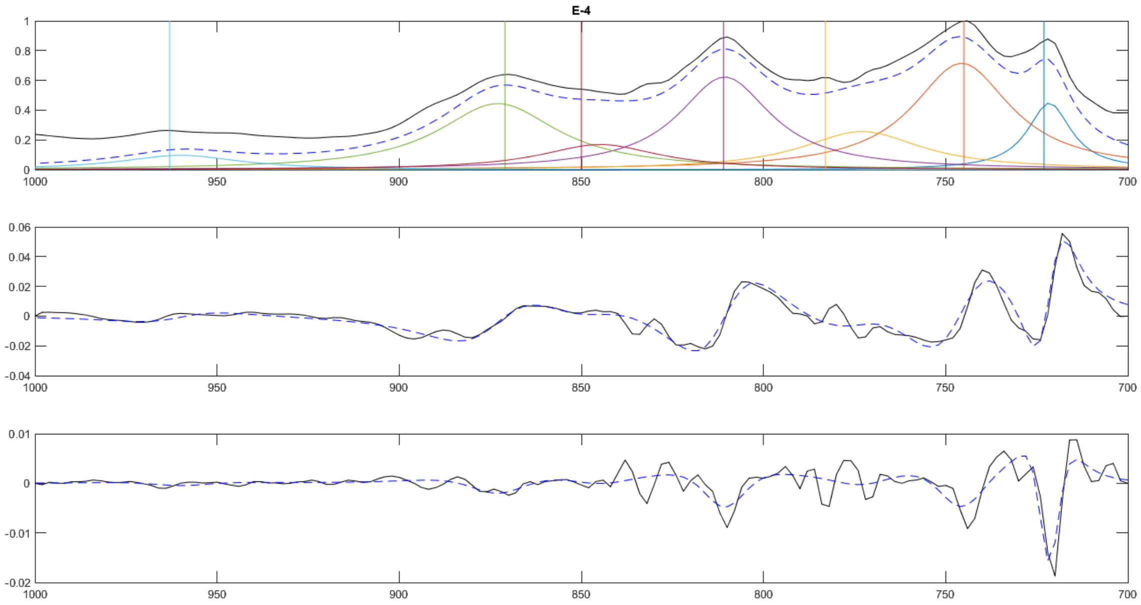


Figure 11, Fitting of Lorentzian distribution for the 1000-700cm^{-1} band of sample E-4

It should be noted that even though the margins between the composite curve and the FTIR generated curve appear extensive, they are in fact a matter of perspective.

From these curves their basic characteristics are taken as mentioned in chapter 2, namely their heights and full widths at half heights (see Tables 2 through 4 in Appendix A). Having extracted these values the next step is to regress them against physical values and check for computational errors through the use of the Standard Error (SE) histogram and the Root Mean Square Error of Prediction (RMSEP) value.

5.2 MLR Analysis

The regression is computed twice, once against the saturates and once against the aromatics percent concentration of the samples. This process is checked for all the possible spectral combinations of the three spectral bands being studied.

After checking the SE and RMSEP values it seems like the combination of all three spectral bands gives the best results. This regression model has an RMSEP of about 0.48, both for saturates and aromatics content, and the generated regression plots for the entire data range, as well as the marginal values that were picked in order to check the regression results, are shown along with the compared lab and generated values as well as the histograms of the SE below in Figures 12-19.

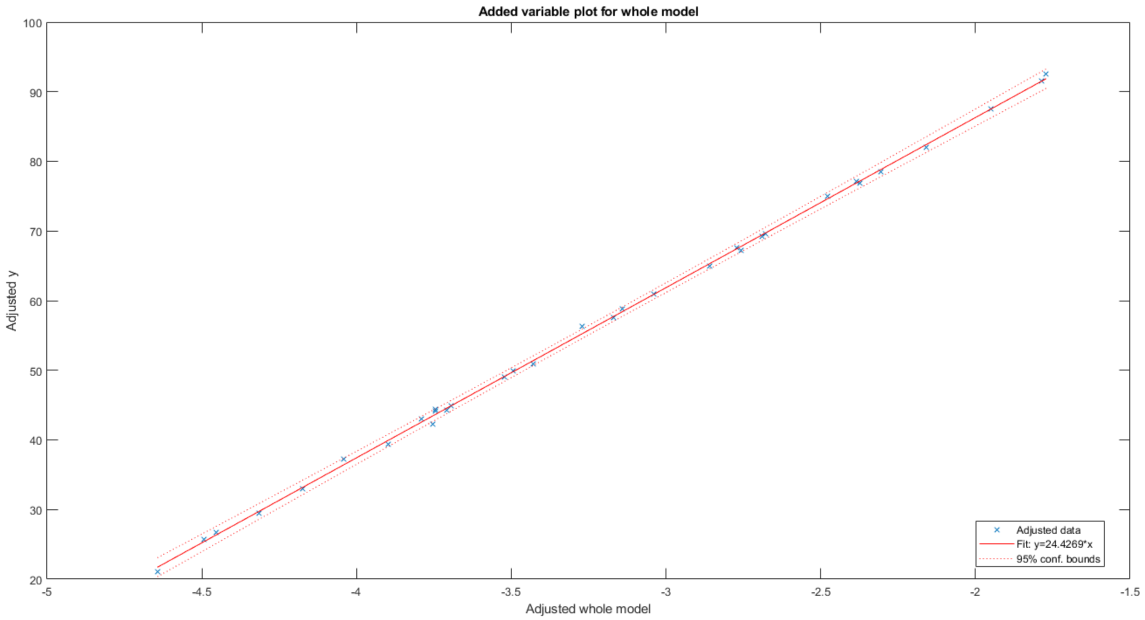


Figure 12, Regression model with respect to saturates concentration

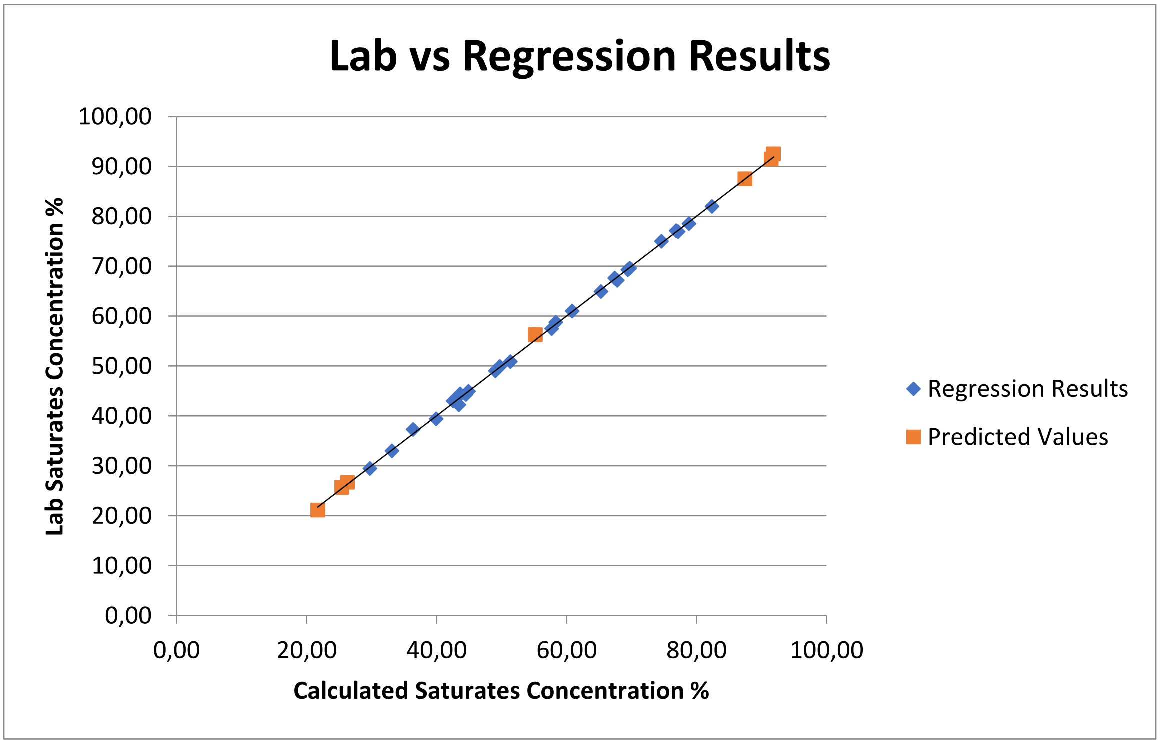


Figure 13, Regression model with respect to saturates concentration (highlighted marginal values)

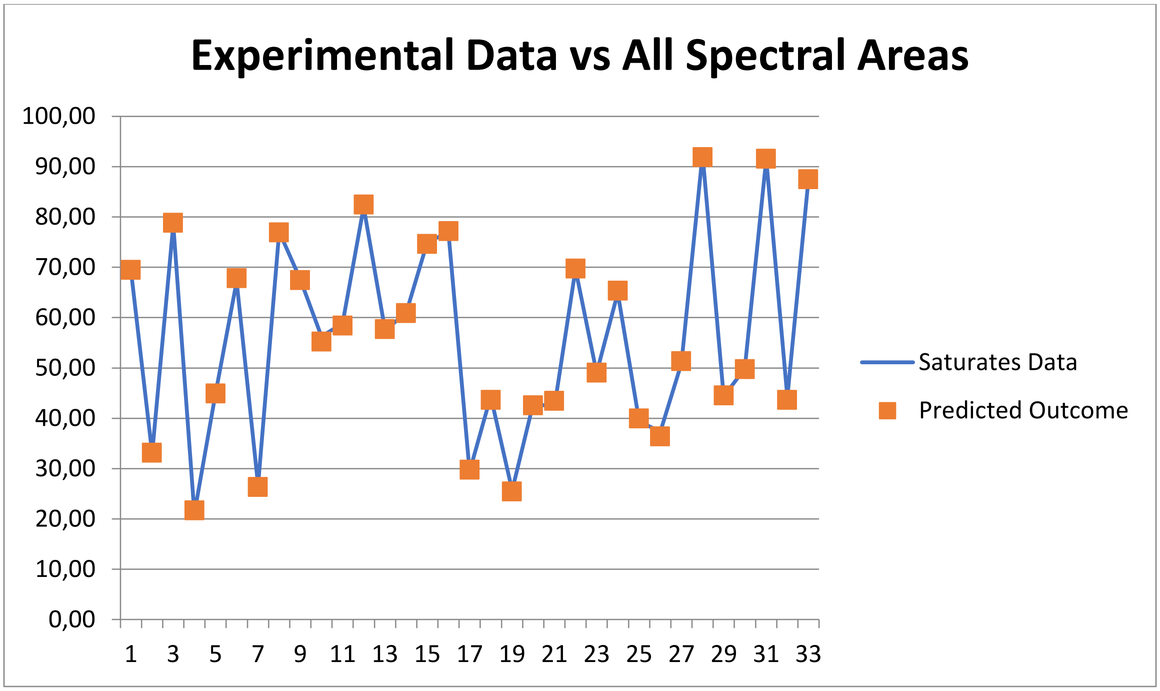


Figure 14, Comparison between Lab data and the regression generated data for saturates

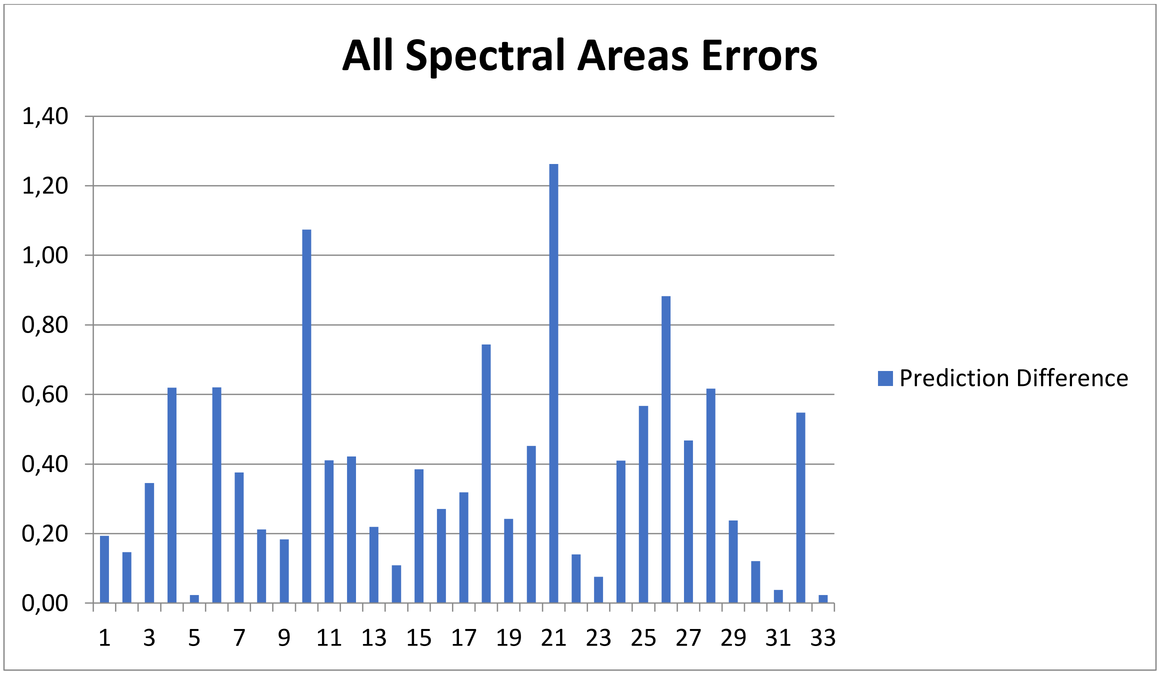


Figure 15, Calculated SE for saturates

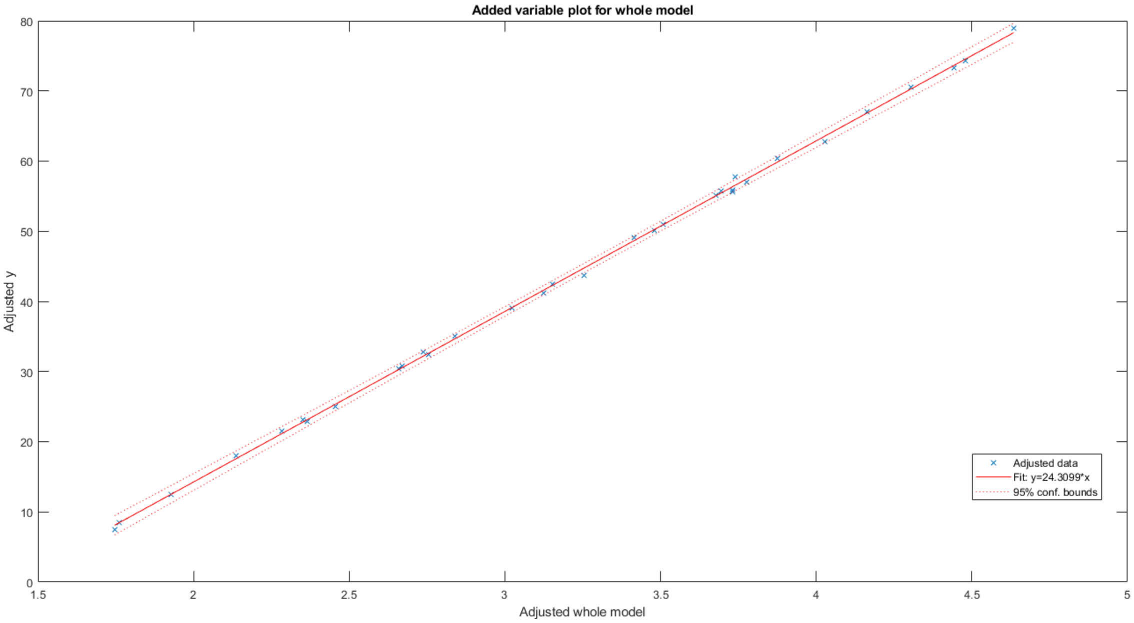


Figure 16, Regression model with respect to aromatics concentration

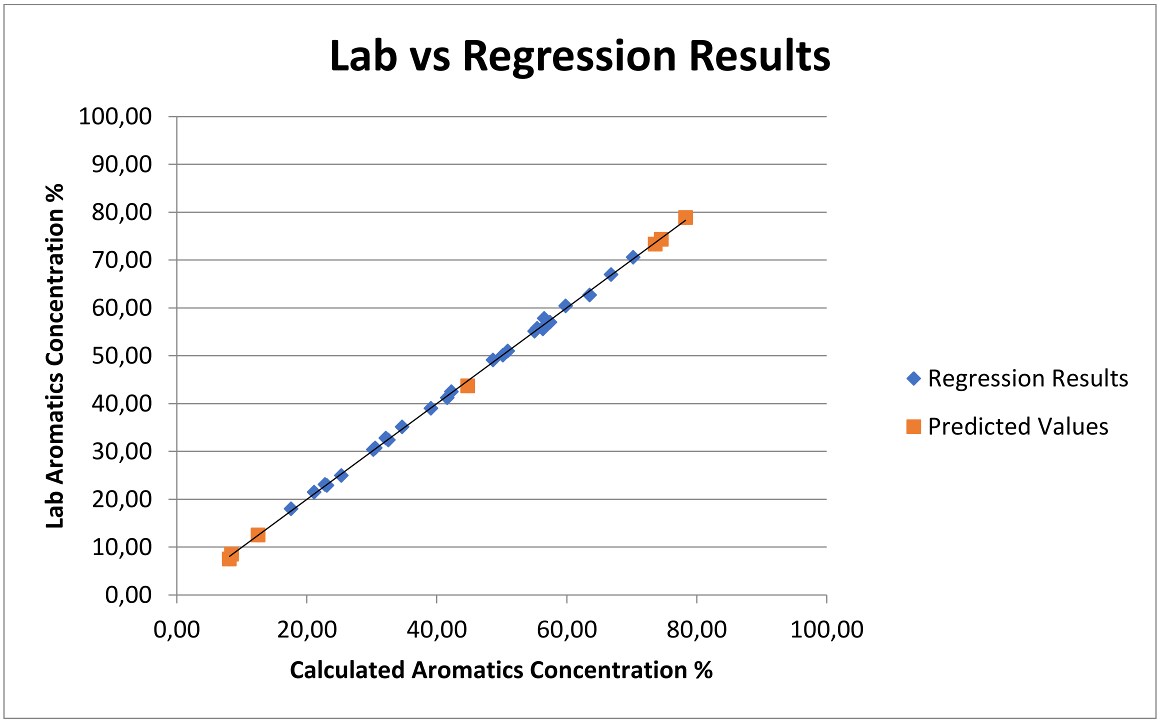


Figure 17, Regression model with respect to aromatics concentration (highlighted marginal values)

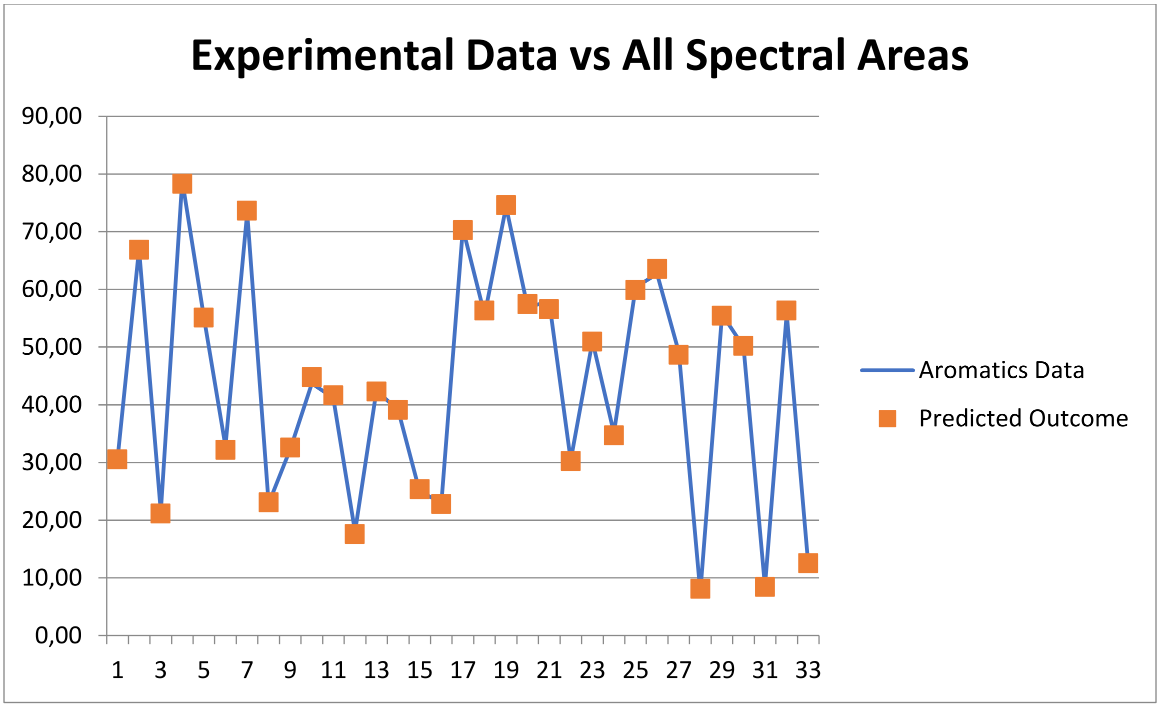


Figure 18, Comparison between Lab data and the regression generated data for aromatics

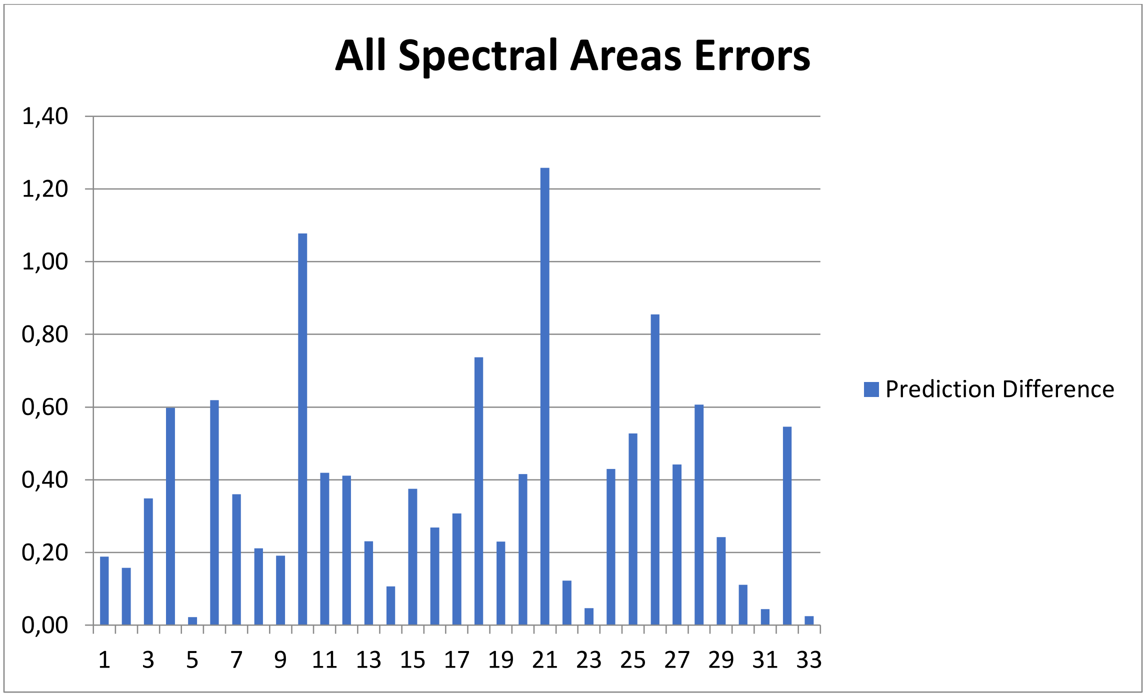


Figure 19, Calculated SE for aromatics

The axes depicted on the graphs of Figures 12 and 16 are corrected axes and they are created through the usage of the Frisch-Waugh-Lovell theorem. What that means is that the y and x axes presented are corrected with respect to their connection to all other variables, in order to be able to correctly represent the regression model on just one 2D plot.

That is the reason for which even though the estimates used are the ones calculated, the intercept is always zero for these plots.

Having extracted these values the next step is to regress them against the physical values. In this case the properties being studied are in regards to the concentration of saturates or aromatics of the samples.

By performing a linear regression between the curves values and the saturates percentage for the upper section of the spectrum then the graph from Figure 12 is generated. For the correlation of aromatics percentage within the same spectrum the graph of Figure 16 is generated instead.

The presented models show that there is a good relation between the terms that were taken into account for this paper. In addition the calculated errors for them back up that claim as well.

This means that there most likely is a connection between the curves taken from the three FTIR areas and the concentrations of saturates and aromatics of an oil sample.

For comparison purposes the computed and the lab obtained concentrations are presented below in Table 2.

Table 2, Comparison between regression and calculated substance concentration values

Saturates %	
Regression	Lab
21.72	21.10
26.32	26.70
55.23	56.30
25.42	25.66
91.88	92.50
91.54	91.50
87.48	87.50

Aromatics %	
Regression	Lab
78.30	78.90
73.66	73.30
44.78	43.70
74.57	74.34
8.11	7.50
8.46	8.50
12.52	12.50

5.3 Other Analysis Methods

The thesis that the next part of the code is based on for this chapter used methods like PCA, PLS and PCR in order to compare them with each other and to comment on their effective results, trying to minimize the error margin in each case.

In this study, all the methods that were analyzed in the older thesis were modified in order to fit the needs of the research, and the PLS analysis appeared to give the best results in this instance. Therefore, only those results will be presented and commented on in this paper.

By using the PLS approach the error margin both for aromatics and saturates was shown to have been minimized meaning that in this case it could be a very good approach as well in order to calculate the desired properties of samples with precision.

In order to generate an appropriate model 80% of the sample size (33 samples total) was used for the creation of the model and the remaining 20% for the prediction.

Specifically, the calculated errors for this scenario are shown below in Figures 20-21 for concentration of aromatics and Figures 22-23 for concentration of saturates. The model's efficiency is also reflected in the calculated RMSEP errors for both properties.

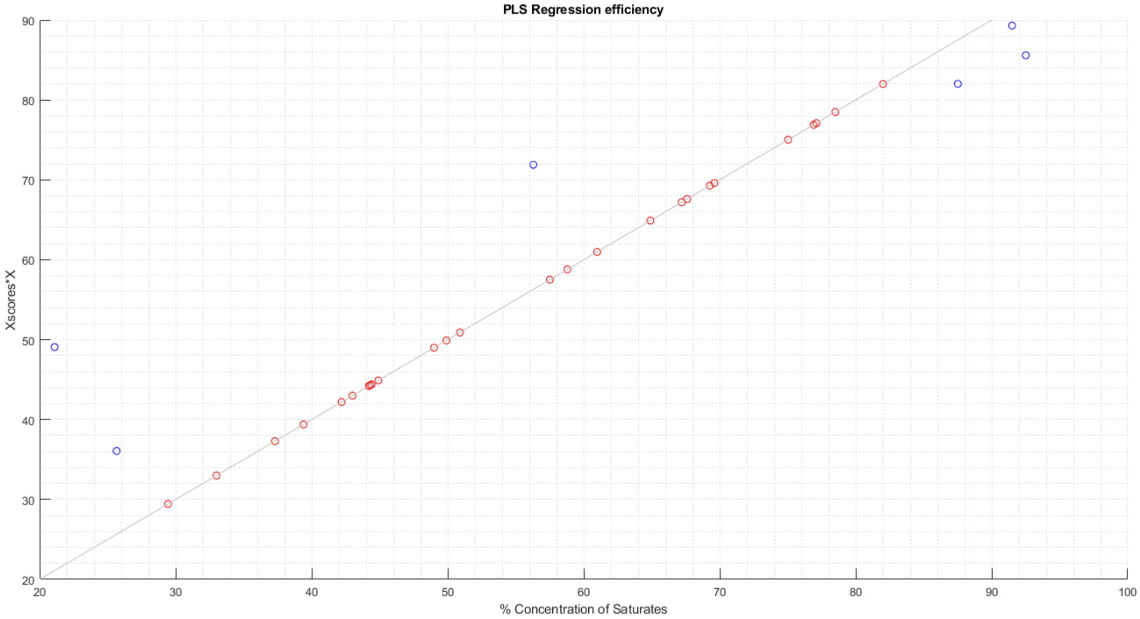


Figure 20, PLS regression efficiency for saturates %

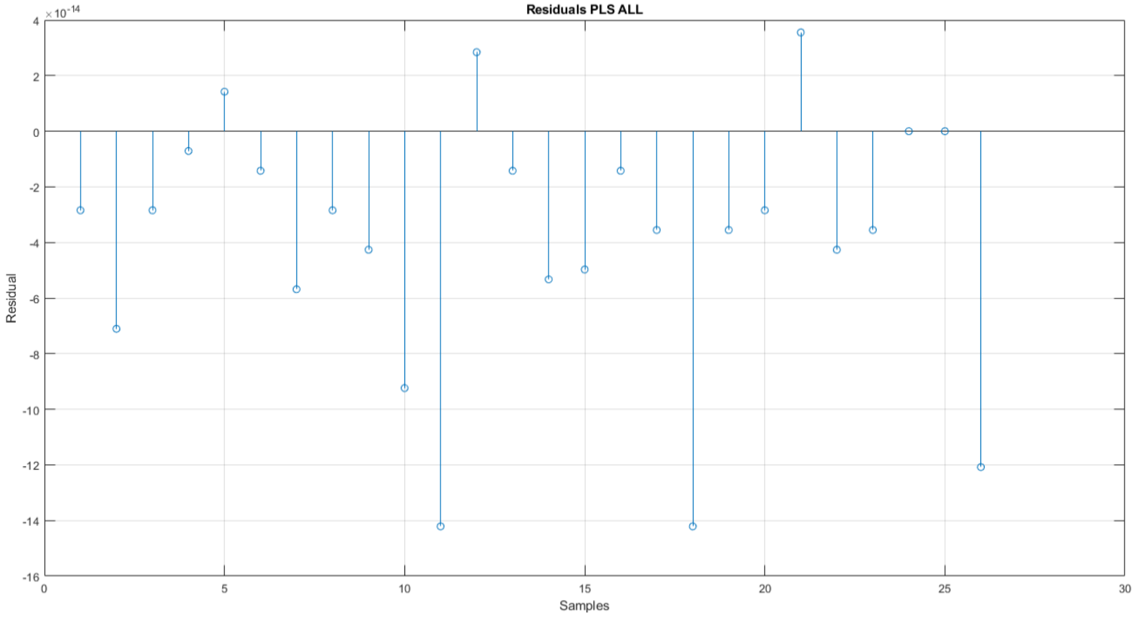


Figure 21, Calculated values deviation for saturates %

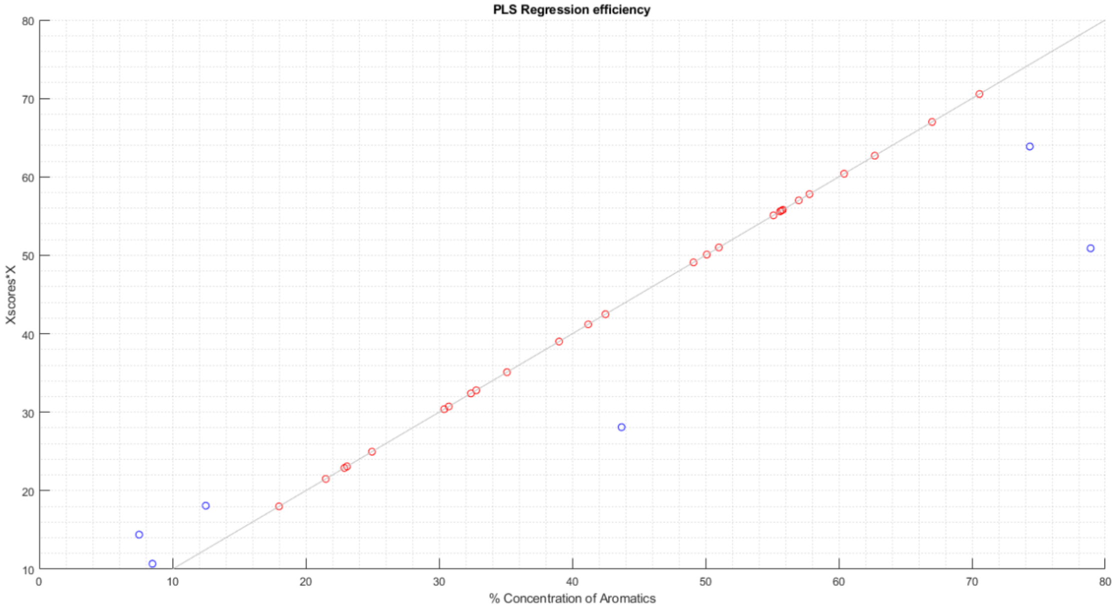


Figure 22, PLS regression efficiency for aromatics %

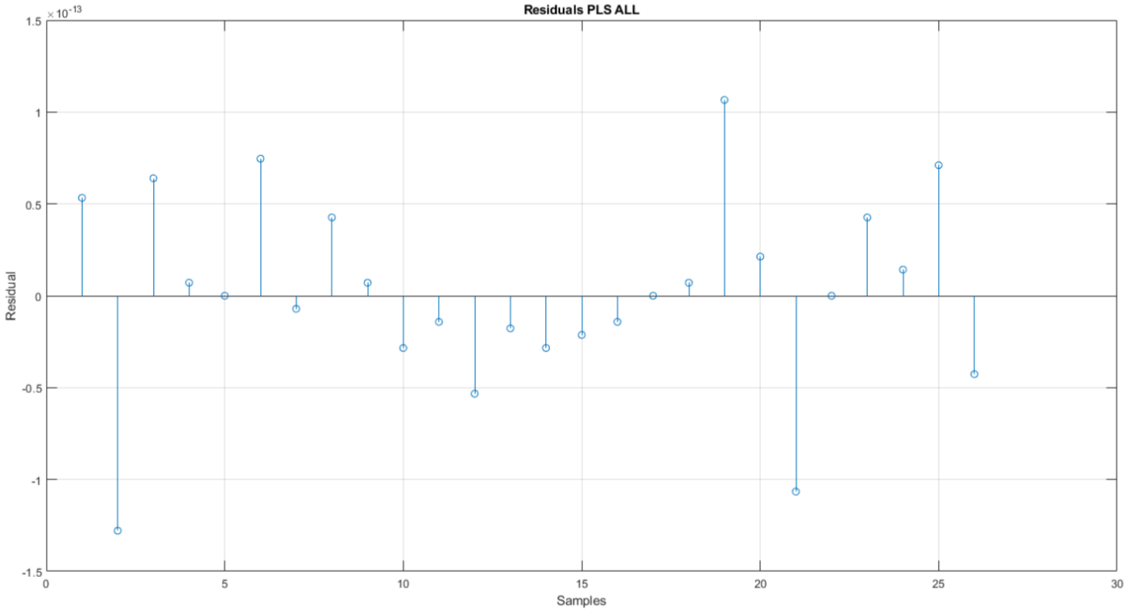


Figure 23, Calculated values deviation for aromatics %

The calculated RMSEP errors for the cases of aromatics and saturates respectively are:

- 5.91e-14
- 5.14e-14

In comparison the results generated by the PCR method are shown below in figures 24-25 and their calculated RMSEP errors for saturates and aromatics concentration are calculated as 0.35 and 0.36 respectively.

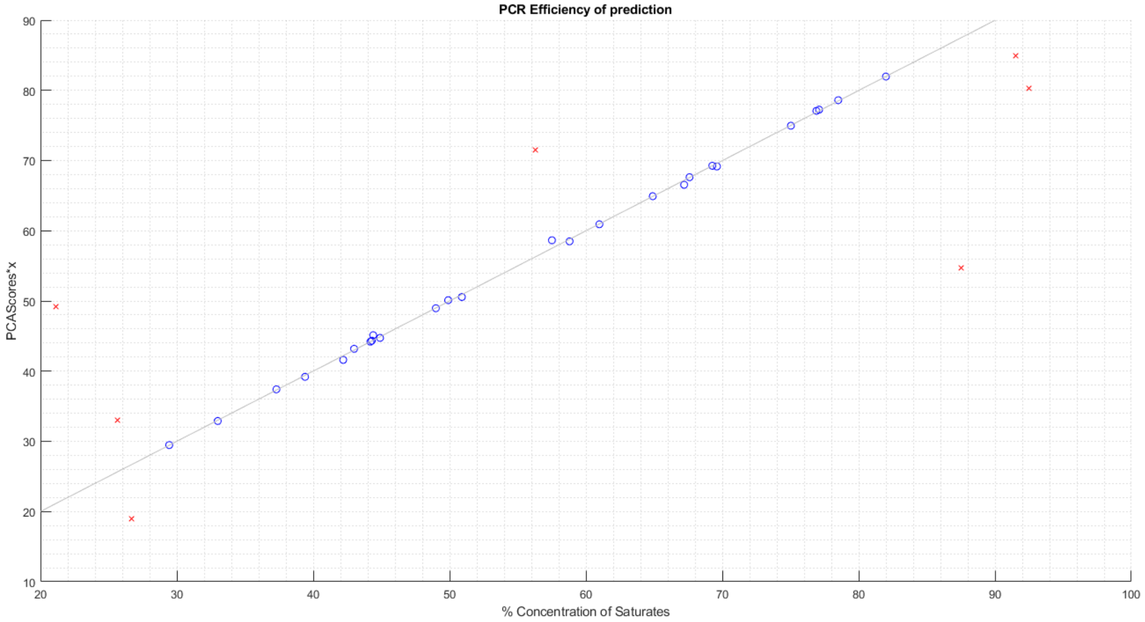


Figure 24, PCR analysis results for saturates %

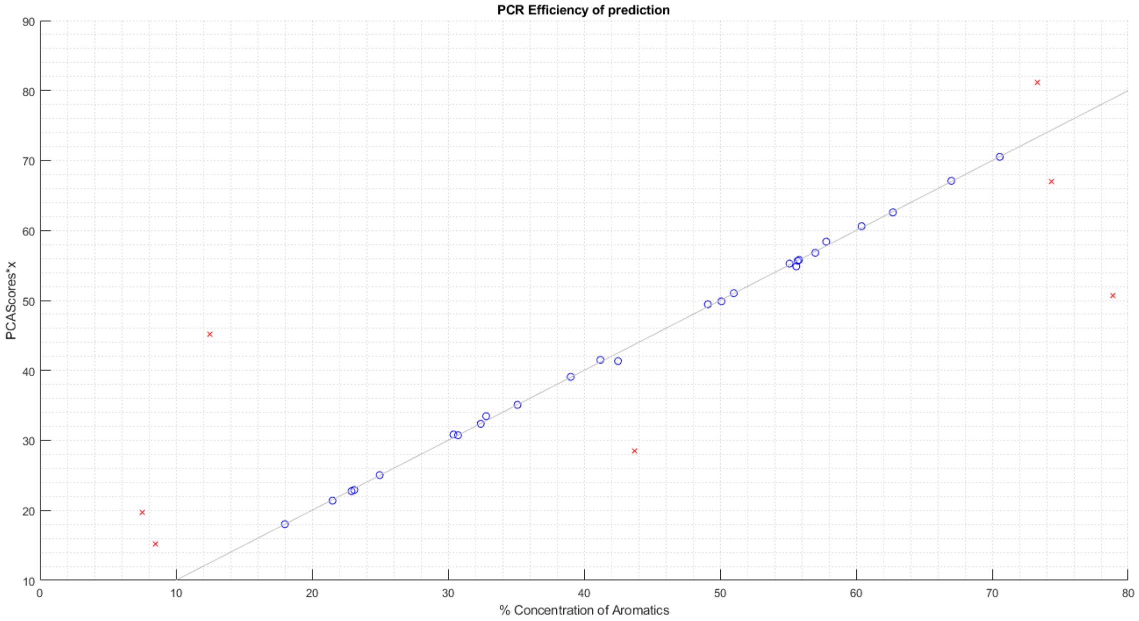


Figure 25, PCR analysis results for aromatics %

6. Conclusions

Summarizing, 33 oil samples' saturates and aromatics concentrations data from FTIR spectroscopy were used for this study. The spectrum got broken down into three parts namely the 3100-2750 cm^{-1}, 1400-1330 cm^{-1} and 1000-700 cm^{-1} bands.

Subsequently, a mathematical model for deconvolution of the signal into specific amounts of Lorentzian curves (five, two and seven respectively) was used and the properties of the generated curves were extracted, namely their heights and widths. These values were then regressed against the measured concentrations of saturates and aromatics of said samples, while also calculating the errors of the generated models.

At a later stage the above mentioned values where then processed with various methods, analyzed in a previous thesis, in order to calculate the errors and efficiency of other models for this scenario.

According to the results presented above it has been shown that the developed process' results are reliable enough for the prediction of the studied properties, while a method like PLS and PCR can also produce precise enough models for this dataset even though not seemly as accurate. The study's topic should be tested against a wider range of samples and properties like density, viscosity and others, seeing as such a capability would drastically decrease the amount of required samples to get a good estimate of the physical and chemical properties of a sample and the subsequent cost connected to it.

Further study is necessary in this field but it has been proven that it is a step towards the right direction. A greater sample range should be studied, for it would most likely help fortify the previous results and give a clearer image and better understanding as to the way that oil samples' compositions or other studied properties are distributed.

Seeing as this field of study is still relatively uncharted, further research is going to be necessary in order to make sure that the generated results are as accurate as possible and to therefore facilitate the future oil industry.

Furthermore other types of curves could be used that have similarly given very good results in this field of study like the Gaussian-Lorentzian (G-L) and the Voigt functions to see if a better result could be attained.

References

Brian C. Smith, Fundamentals of Fourier Transform Infrared spectroscopy, CRC press, Boca Raton, 1996.

"Introduction to Curve Fitting." *NCSS*, n.d., ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Introduction_to_Curve_Fitting.pdf.

Kenton, Will. "How Multiple Linear Regression Works." *Investopedia*, Investopedia, 29 Jan. 2020, www.investopedia.com/terms/m/mlr.asp.

"Built-in Fitting Models in the Models Module¶." *Built-in Fitting Models in the Models Module - Non-Linear Least-Squares Minimization and Curve-Fitting for Python*, n.d., lmfit.github.io/lmfit-py/builtin_models.html.

Gooch, Sophie, and Adrien. "Nonlinear Regression Essentials in R: Polynomial and Spline Regression Models." *STHDA*, 11 Mar. 2018, www.sthda.com/english/articles/40-regression-analysis/162-nonlinear-regression-essentials-in-r-polynomial-and-spline-regression-models/.

"IR Spectrum Table & Chart." *Sigma*, n.d., www.sigmaaldrich.com/technical-documents/articles/biology/ir-spectrum-table.html.

"Simple Linear Regression Model Fitting." *Simple Linear Regression Model Fitting*, n.d., sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression5.html.

Foley, Ben. "What Is Regression Analysis and Why Should I Use It?: SurveyGizmo Blog." *SurveyGizmo*, 19 Sept. 2019, www.surveygizmo.com/resources/blog/regression-analysis/.

Gallo, Amy, et al. "A Refresher on Regression Analysis." *Harvard Business Review*, 30 Nov. 2017, hbr.org/2015/11/a-refresher-on-regression-analysis.

Sánchez-Lemus, M. C., et al. "Characterization of Heavy Distillation Cuts Using Fourier Transform Infrared Spectrometry: Proof of Concept." *Energy & Fuels*, vol. 30, no. 12, 2016, pp. 10187–10199., doi:10.1021/acs.energyfuels.6b01912.

Boundless. "Boundless Statistics." *Lumen*, n.d., courses.lumenlearning.com/boundless-statistics/chapter/r-m-s-error-for-regression/.

Editor, Minitab Blog. "How to Interpret Regression Analysis Results: P-Values and Coefficients." *Minitab Blog*, n.d., blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients.

Frost, Jim, et al. “Standard Error of the Regression vs. R-Squared.” *Statistics By Jim*, 15 Mar. 2019, statisticsbyjim.com/regression/standard-error-regression-vs-r-squared/.

Pasadakis, Nikos, et al. “Deconvolving the Absorbance of Methyl and Methylene Groups in the FT-IR 3000-2800 cm^{-1} Band of Petroleum Fractions.” *Trends in Applied Spectroscopy*, vol. 10, 2013.

Bradley, Michael, and Thermo Fisher. “Curve Fitting in Raman and IR Spectroscopy: Basic Theory of Line Shapes and Applications.” *ThermoFisher Scientific*, n.d., assets.thermofisher.com/.

Pierce, John A., et al. “Combined Deconvolution and Curve Fitting for Quantitative Analysis of Unresolved Spectral Bands.” *Analytical Chemistry*, vol. 62, no. 5, 1990, pp. 477–484., doi:10.1021/ac00204a011.

“Chapter 3 Spectral Analysis.” *Infrared Spectroscopy: Fundamentals and Applications*, by Barbara H. Stuart, Wiley, 2008, pp. 45–48.

“4 Vibrational Spectroscopy of Different Classes and States of Compounds.” *Infrared and Raman Spectroscopy: Methods and Applications*, by B. Schrader and D. Bougeard, VCH, 1995, pp. 190–191.

Willard, Hobart Hurd. *Instrumental Methods of Analysis*. Wadsworth, 1988.

O'Haver, Tom. “A Pragmatic Introduction to Signal Processing.” *Intro. to Signal Processing:Deconvolution*, 17 May 2008, terpconnect.umd.edu/~toh/spectrum/Deconvolution.html.

Spills of Diluted Bitumen from Pipelines: a Comparative Study of Environmental Fate, Effects, and Response. The National Academies Press, 2016.

Alexopoulos E. C. (2010). Introduction to multivariate regression analysis. Hippokratia, 14(Suppl 1), 23–28.

“HOME.” IDRE Stats, stats.idre.ucla.edu/stata/dae/multivariate-regression-analysis/.

“Oil Fluid Properties.” PetroWiki, petrowiki.org/Oil_fluid_properties.

APPENDICES

APPENDIX A

Useful data tables regarding the computational process

In Tables 2-4 below are presented the curve characteristics extracted from the fitted curves for all the samples. This data was used for the creation of the regression model but where all yc values where scaled up by a factor of 10 for computational purposes. The data for the curves' center locations could not be generated by this algorithm.

Table 3, Curve properties for the upper section of the spectrum

Curves	**#1**		**#2**		**#3**		**#4**		**#5**	
Sample Name	**yc**	**w**	**yc**	**w**	**yc**	**w**	**yc**	**w**	**yc**	**w**
DAO15-1-2009	0.54	11.85	0.83	13.50	0.28	13.15	0.25	0.05	0.26	20.00
E-4	0.48	10.20	0.90	17.56	0.30	12.48	0.02	20.00	0.17	20.00
E-11	0.55	11.74	0.82	13.82	0.37	12.53	0.25	20.00	0.00	19.28
E-12	0.56	9.79	0.87	16.28	0.23	12.41	0.02	13.54	0.03	20.00
E-19	0.56	12.05	0.82	14.79	0.31	13.65	0.00	17.21	0.23	20.00
E-26	0.57	10.06	0.91	15.62	0.32	13.38	0.18	11.79	0.18	13.56
E-33	0.50	17.26	0.87	19.89	0.39	14.05	0.06	17.34	0.08	20.00
E-37	0.55	8.97	0.92	15.71	0.29	11.32	0.19	12.30	0.17	12.23
E-39	0.54	12.31	0.79	14.01	0.39	13.10	0.00	18.34	0.27	20.00
E-40	0.58	10.28	0.91	15.98	0.32	13.41	0.17	11.48	0.19	14.09
E-41	0.55	11.57	0.84	14.71	0.33	13.25	0.22	20.00	0.01	20.00
E-44	0.44	5.31	0.30	19.97	0.90	11.42	0.45	9.56	0.06	6.55
E-47	0.52	13.15	0.80	13.86	0.45	12.42	0.01	12.98	0.25	20.00
E-50	0.58	10.60	0.91	16.17	0.32	13.28	0.16	10.74	0.19	14.19
E-51	0.53	8.30	0.86	13.58	0.27	12.96	0.32	0.00	0.11	10.46
E-53	0.64	7.39	0.88	13.58	0.20	12.13	0.10	9.06	0.04	4.67
E-54	0.56	7.12	0.92	14.36	0.23	12.30	0.21	17.61	0.14	11.03
E-55	0.49	16.04	0.86	19.57	0.36	13.77	0.05	16.55	0.07	20.00
E-56	0.54	12.00	0.82	14.63	0.31	13.69	0.23	20.00	0.01	20.00
E-59	0.56	10.07	0.88	16.22	0.21	12.25	0.02	13.41	0.03	20.00
E-61	0.55	12.04	0.82	14.87	0.31	13.74	0.12	20.00	0.12	20.00
E-62	0.55	11.70	0.83	14.84	0.30	13.51	0.00	13.57	0.23	20.00
E-63	0.58	10.95	0.88	15.43	0.32	13.45	0.23	16.34	0.15	9.37
E-67	0.56	12.00	0.83	14.96	0.31	13.76	0.00	16.72	0.23	20.00
E-74	0.56	11.45	0.83	14.74	0.33	13.12	0.23	20.00	0.00	13.68
E-83	0.55	11.90	0.86	15.53	0.31	13.69	0.20	20.00	0.01	19.99
E-84	0.54	11.88	0.81	14.42	0.32	14.14	0.04	6.77	0.21	20.00
E-92	0.57	9.59	0.89	16.45	0.29	12.73	0.10	10.68	0.00	12.73

Table 2, Continued

Curves	**#1**		**#2**		**#3**		**#4**		**#5**	
Sample Name	**yc**	**w**	**yc**	**w**	**yc**	**w**	**yc**	**w**	**yc**	**w**
E-93	0.55	10.43	0.89	16.60	0.32	10.92	0.00	20.00	0.00	19.99
E-94	0.53	12.46	0.84	14.76	0.36	13.54	0.22	20.00	0.00	18.79
E-102	0.56	9.17	0.94	16.57	0.32	13.41	0.15	11.76	0.22	15.66
E-103	0.58	11.04	0.81	14.10	0.34	11.85	0.14	20.00	0.14	20.00
E-104	0.52	12.13	0.83	15.06	0.35	13.31	0.00	19.80	0.20	20.00
M-17	0.60	8.64	0.12	6.87	0.94	12.49	0.33	9.69	0.13	M-17
M-17-C	0.63	9.41	0.10	5.74	0.90	13.84	0.35	9.33	0.10	9.73
M-17-D	0.59	7.50	0.18	20.00	0.99	14.63	0.34	9.26	0.01	20.00
M-17-E	0.65	7.06	0.15	20.00	0.93	12.60	0.30	8.01	0.09	6.43
M-18-B	0.00	7.61	0.63	11.71	0.12	4.77	0.38	8.17	0.64	10.65
M-18-C	0.66	11.94	0.13	5.05	0.43	8.43	0.72	12.05	0.50	9.46
M-18	0.56	6.24	0.22	11.82	0.23	10.09	0.85	9.40	0.43	10.01
M-19-B	0.33	0.03	0.48	6.92	0.16	19.99	0.95	14.47	0.33	7.01

Table 4, Curve properties for the intermediate section of the spectrum

Curves	***#1***		***#2***	
Sample Name	**yc**	**w**	**yc**	**w**
DAO15-1-2009	0,37	4,75	0,71	6,30
E-4	0,56	6,37	0,30	20,00
E-11	0,11	20,00	0,81	6,87
E-12	0,56	6,42	0,30	20,00
E-19	0,76	7,02	0,00	4,93
E-26	0,68	6,24	0,32	20,00
E-33	0,12	20,00	0,74	8,11
E-37	0,81	6,89	0,24	0,02
E-39	0,91	8,10	0,25	20,00
E-40	0,67	6,27	0,30	20,00
E-41	0,90	8,37	0,27	20,00
E-44	0,05	2,85	0,04	2,38
E-47	0,12	20,00	0,86	7,06
E-50	0,89	8,57	0,28	20,00
E-51	0,68	6,19	0,33	20,00
E-53	0,61	5,90	0,30	20,00
E-54	0,69	5,84	0,31	17,14
E-55	0,63	7,10	0,21	20,00
E-56	0,63	6,35	0,29	20,00
E-59	0,73	8,33	0,13	7,81
E-61	0,61	6,31	0,35	20,00

Table 3, Continued

Curves	***#1***		***#2***	
Sample Name	**yc**	**w**	**yc**	**w**
E-62	0,87	9,41	0,33	20,00
E-63	0,78	7,09	0,00	7,16
E-67	0,62	6,24	0,37	20,00
E-74	0,71	5,74	0,27	17,21
E-83	0,34	20,00	0,87	11,40
E-84	0,34	20,00	0,87	11,48
E-92	0,67	6,30	0,35	20,00
E-93	0,83	5,29	0,32	20,00
E-94	0,31	20,00	0,88	10,78
E-102	0,67	6,33	0,32	20,00
E-103	0,89	6,03	0,18	5,16
E-104	0,31	20,00	0,88	10,80
M-17	0,10	4,08	0,21	20,00
M-17-C	0,09	4,24	0,18	20,00
M-17-D	0,09	4,30	0,18	20,00
M-17-E	0,09	4,42	0,20	20,00
M-18-B	0,07	3,91	0,19	20,00
M-18-C	0,00	13,43	0,68	20,00
M-18	0,08	4,01	0,20	20,00
M-19-B	0,00	12,38	0,07	4,96
M-19	0,07	4,15	0,00	13,66

Table 5, Curve properties for the lower section of the spectrum

Curves	***#1***		***#2***		***#3***		***#4***		***#5***		***#6***		***#7***	
Sample Name	**yc**	**w**	**yc**	**w**	**yc**	**w**	**yc**	**w**	**yc**	**w**	**yc**	**w**	**yc**	**w**
DAO15-1-2009	0,78	7,66	0,16	18,77	0,04	8,64	0,12	20,00	0,08	10,81	0,14	20,00	0,15	20,00
E-4	0,45	7,25	0,72	16,40	0,26	20,00	0,62	15,57	0,44	20,00	0,10	20,00	0,17	20,00
E-11	0,75	10,17	0,63	11,17	0,30	20,00	0,12	20,00	0,54	12,15	0,07	1,59	0,43	20,00
E-12	0,55	8,58	0,81	17,13	0,27	20,00	0,57	10,94	0,22	20,00	0,53	20,00	0,19	20,00
E-19	0,59	7,23	0,80	20,00	0,53	20,00	0,36	20,00	0,04	20,00	0,10	20,00	0,11	20,00
E-26	0,64	7,37	0,50	20,00	0,37	14,12	0,21	20,00	0,02	20,00	0,07	20,00	0,04	6,23
E-33	0,08	3,25	0,18	5,38	0,79	15,07	0,10	20,00	0,51	14,00	0,05	20,00	0,03	3,22
E-37	0,60	6,35	0,39	20,00	0,13	20,00	0,34	13,42	0,05	7,74	0,19	20,00	0,02	20,00
E-39	0,74	10,68	0,39	9,10	0,32	20,00	0,28	20,00	0,53	13,48	0,17	20,00	0,35	20,00
E-40	0,76	9,56	0,60	20,00	0,52	20,00	0,25	20,00	0,15	20,00	0,24	20,00	0,14	19,93
E-41	0,62	7,49	0,62	20,00	0,47	15,30	0,06	7,77	0,26	20,00	0,05	20,00	0,09	20,00
E-44	1,00	8,98	0,14	2,94	0,04	19,99	0,05	5,60	0,05	3,56	0,03	3,35	0,17	20,00
E-47	0,60	6,22	0,74	14,68	0,31	8,90	0,22	6,97	0,47	12,85	0,07	20,00	0,07	5,15

Table 4, Continued

Curves	***#1***		***#2***		***#3***		***#4***		***#5***		***#6***		***#7***	
Sample Name	**yc**	**w**	**yc**	**w**	**yc**	**w**	**yc**	**w**	**yc**	**w**	**yc**	**w**	**yc**	**w**
E-50	0,68	8,07	0,56	20,00	0,47	17,03	0,20	20,00	0,07	20,00	0,10	20,00	0,15	20,00
E-51	0,65	7,47	0,53	20,00	0,16	20,00	0,43	14,63	0,06	7,21	0,22	20,00	0,01	20,00
E-53	0,42	1,34	0,67	13,77	0,00	12,09	0,24	20,00	0,24	20,00	0,14	20,00	0,30	20,00
E-54	0,80	8,26	0,33	20,00	0,21	20,00	0,06	19,99	0,14	20,00	0,05	4,93	0,17	20,00
E-55	0,17	9,20	0,21	3,54	0,91	18,56	0,19	20,00	0,52	11,53	0,33	0,00	0,59	20,00
E-56	0,71	10,99	0,68	20,00	0,60	20,00	0,30	20,00	0,16	20,00	0,26	20,00	0,21	20,00
E-59	0,46	6,58	0,83	15,92	0,18	7,28	0,27	9,65	0,58	9,75	0,25	20,00	0,48	20,00
E-61	0,71	10,16	0,66	20,00	0,61	20,00	0,32	20,00	0,17	20,00	0,28	20,00	0,20	20,00
E-62	0,64	8,32	0,58	20,00	0,53	18,79	0,23	20,00	0,05	20,00	0,08	20,00	0,19	20,00
E-63	0,66	7,54	0,57	20,00	0,44	16,28	0,15	20,00	0,21	20,00	0,07	20,00	0,11	20,00
E-67	0,63	7,89	0,59	20,00	0,48	19,11	0,22	20,00	0,06	20,00	0,10	20,00	0,19	20,00
E-74	0,71	8,74	0,55	20,00	0,21	20,00	0,45	13,06	0,30	20,00	0,15	20,00	0,21	20,00
E-83	0,47	7,25	0,71	19,37	0,16	20,00	0,65	20,00	0,40	20,00	0,09	5,43	0,15	7,27
E-84	0,44	7,18	0,67	19,07	0,15	20,00	0,62	19,48	0,37	20,00	0,08	5,26	0,15	7,10
E-92	0,51	4,09	0,40	20,00	0,21	2,56	0,44	20,00	0,48	20,00	0,34	20,00	0,24	20,00
E-93	0,81	7,77	0,28	20,00	0,02	3,17	0,15	12,85	0,06	19,98	0,06	0,47	0,08	20,00
E-94	0,47	7,58	0,67	20,00	0,25	20,00	0,65	18,72	0,42	20,00	0,53	6,70	0,20	6,45
E-102	0,49	4,10	0,40	20,00	0,18	2,73	0,48	20,00	0,50	20,00	0,36	20,00	0,25	20,00
E-103	0,83	7,97	0,28	20,00	0,41	0,17	0,15	13,78	0,10	20,00	0,61	0,13	0,26	20,00
E-104	0,49	7,62	0,65	17,44	0,31	20,00	0,70	19,14	0,33	20,00	0,32	13,31	0,52	6,91
M-17	0,94	8,51	0,09	6,11	0,05	20,00	0,01	3,53	0,03	7,77	0,03	7,60	0,11	20,00
M-17-C	0,85	4,92	0,53	4,24	0,12	19,99	0,05	19,96	0,03	19,95	0,06	19,99	0,10	20,00
M-17-D	0,80	4,44	0,61	4,54	0,06	6,67	0,02	20,00	0,02	10,62	0,02	7,05	0,01	6,27
M-17-E	0,81	4,19	0,68	4,33	0,07	10,30	0,03	20,00	0,02	11,97	0,02	9,81	0,05	20,00
M-18-B	0,81	4,49	0,65	4,62	0,07	6,90	0,02	4,46	0,01	6,04	0,02	5,06	0,02	6,72
M-18-C	0,81	4,29	0,69	4,47	0,06	6,54	0,03	12,68	0,02	6,84	0,02	4,94	0,08	6,97
M-18	0,98	8,11	0,13	19,41	0,05	10,01	0,04	11,40	0,04	6,57	0,08	7,32	0,08	19,99
M-19-B	0,98	8,10	0,13	3,39	0,11	20,00	0,03	2,96	0,09	20,00	0,04	1,03	0,41	20,00
M-19	0,97	8,42	0,19	5,52	0,07	6,78	0,07	6,91	0,07	8,46	0,07	8,46	0,35	6,29

The percentage saturates data, percentage aromatics data and names for all the samples are shown below in Table 5:

Table 6, Samples' data used in the research

Sample Name	**Saturates %**	**Aromatics %**
DAO15-1-2009	69.27	30.73
E-4	33.0	67.0
E-11	78.5	21.5
E-12	21.1	78.9

Table 5, Continued

Sample Name	Saturates %	Aromatics %
E-19	44.9	55.1
E-26	67.2	32.8
E-33	26.7	73.3
E-37	77.1	22.9
E-39	67.6	32.4
E-40	56.3	43.7
E-41	58.8	41.2
E-44	99.8	0.2
E-47	82.0	18.0
E-50	57.5	42.5
E-51	61.0	39.0
E-53	75.0	25.0
E-54	76.9	23.1
E-55	29.4	70.6
E-56	44.4	55.6
E-59	25.7	74.3
E-61	43.0	57.0
E-62	42.2	57.8
E-63	69.6	30.4
E-67	49.0	51.0
E-74	64.9	35.1
E-83	39.4	60.4
E-84	37.3	62.7
E-92	50.9	49.1
E-93	92.5	7.5
E-94	44.3	55.7
E-102	49.9	50.1
E-103	91.5	8.5
E-104	44.2	55.8
M-17	99.6	0.4
M-17-C	99.6	0.4
M-17-D	99.6	0.4
M-17-E	99.6	0.4
M-18	99.5	0.5
M-18-B	99.5	0.5
M-18-C	99.5	0.5
M-19-B	87.5	12.5

APPENDIX B

Coding section

Below are given and explained the commands used in order to calculate the above results.

The “x” variables refer to the imported data regarding the deconvolved curves, “xi_y” for the heights and “xi_w” for the widths where “i” denominates the curve, and are vectors with a structure of n*1. This data is subsequently inserted in a matrix “X” for it to be processed by the fitlm command. The “sat” and “arom” commands are vectors that contain the same amount of data points as the x variables

fitlm generates a linear model between the “X” matrix and the “sat” or “arom” vector. This command generates all the needed values that could be used in order to check the validity of the results. Nonetheless, just as a safety measure the end result is double checked by constructing the y_predict vector that will calculate the expected outcomes for the y values.

Create regression model for saturates data

```
X_s = [x1_y x1_w x2_y x2_w x3_y x3_w x4_y x4_w x5_y x5_w x6_y x6_w x7_y x7_w x8_y x8_w
x9_y x9_w x10_y x10_w x11_y x11_w x12_y x12_w x13_y x13_w x14_y x14_w];
mdl_s = fitlm(X_s,sat);
b_s = mdl_s.Coefficients.Estimate;
y_predict_s = b_s(1,1) + b_s(2,1)*x1_y + b_s(3,1)*x1_w + b_s(4,1)*x2_y + b_s(5,1)*x2_w +
b_s(6,1)*x3_y + b_s(7,1)*x3_w + b_s(8,1)*x4_y + b_s(9,1)*x4_w + b_s(10,1)*x5_y +
b_s(11,1)*x5_w + b_s(12,1)*x6_y + b_s(13,1)*x6_w + b_s(14,1)*x7_y + b_s(15,1)*x7_w +
b_s(16,1)*x8_y + b_s(17,1)*x8_w + b_s(18,1)*x9_y + b_s(19,1)*x9_w + b_s(20,1)*x10_y +
b_s(21,1)*x10_w + b_s(22,1)*x11_y + b_s(23,1)*x11_w + b_s(24,1)*x12_y + b_s(25,1)*x12_w +
b_s(26,1)*x13_y + b_s(27,1)*x13_w + b_s(28,1)*x14_y + b_s(29,1)*x14_w ;
```

Create regression model for aromatics data

```
X_a = [x1_y x1_w x2_y x2_w x3_y x3_w x4_y x4_w x5_y x5_w x6_y x6_w x7_y x7_w x8_y x8_w
x9_y x9_w x10_y x10_w x11_y x11_w x12_y x12_w x13_y x13_w x14_y x14_w];
mdl_a = fitlm(X_a,arom);
b_a = mdl_a.Coefficients.Estimate;
y_predict_a = b_a(1,1) + b_a(2,1)*x1_y + b_a(3,1)*x1_w + b_a(4,1)*x2_y + b_a(5,1)*x2_w +
b_a(6,1)*x3_y + b_a(7,1)*x3_w + b_a(8,1)*x4_y + b_a(9,1)*x4_w + b_a(10,1)*x5_y +
b_a(11,1)*x5_w + b_a(12,1)*x6_y + b_a(13,1)*x6_w + b_a(14,1)*x7_y + b_a(15,1)*x7_w +
b_a(16,1)*x8_y + b_a(17,1)*x8_w + b_a(18,1)*x9_y + b_a(19,1)*x9_w + b_a(20,1)*x10_y +
b_a(21,1)*x10_w + b_a(22,1)*x11_y + b_a(23,1)*x11_w + b_a(24,1)*x12_y + b_a(25,1)*x12_w +
b_a(26,1)*x13_y + b_a(27,1)*x13_w + b_a(28,1)*x14_y + b_a(29,1)*x14_w ;
```

Again, even though the errors have been calculated beforehand by the command they are still double checked in order to verify and then subsequently plotted. The models are plotted as well.

Error calculation

```
n = length(x1_y);

error_s = (y_predict_s-sat);
RMSEP_s = sqrt(1/n*sum(error_s.^2));
figure(1)
histogram(error_s,n)

error_a = (y_predict_a-arom);
RMSEP_a = sqrt(1/n*sum(error_a.^2));
figure(2)
histogram(error_a,n)
```

Plot models

```
figure(3)
plot(mdl_s)

figure(4)
plot(mdl_a)
```

Further below is presented the code utilized for the model creation process through PLS and PCR analysis. This code, with the appropriate changes is viable for both saturates % and aromatics % data.

Values input

```
clear; clc;
D = load('ASTM_2549_modified.mat');
wavenumber = D.Wl;
Labels = D.Labels;
Absorbance = D.X';
Sat = xlsread('Data_input.xlsx',1,'C3:C43');
Arom = xlsread('Data_input.xlsx',1,'D3:D43');
format
```

Transmittance to Absorbance

```
[~, h1] = sort(Sat);
[~, h2] = sort(Arom);
```

Pretreatment of Absorbance Data

```
Mean_of_Absorbance = mean(Absorbance);
Standard_Deviation_of_Absorbance = std(Absorbance);

Standardised_Data_Matrix = (Absorbance - repmat(Mean_of_Absorbance,
[(size(Absorbance,1)) 1])) ./
repmat(Standard_Deviation_of_Absorbance,[(size(Absorbance,1)) 1]);

Training_Set = Standardised_Data_Matrix;

Training_Set = Training_Set/max(max(Training_Set));
```

Cross Validation: PCR vs PLSR

```
NumberofComponents = 40;

[Xloadings,Yloadings,Xscores,Yscores,betaPLSn,PLSctVar] =
plsregress(Training_Set,Sat,NumberofComponents);
[PCALoadings,PCAScores,PCAVar] = pca(Training_Set,'Economy',false);

yfitPLSn = [ones(size(Training_Set,1),1) Training_Set]*betaPLSn;

betaPCRn = regress(Sat-mean(Sat),PCAScores(:,1:NumberofComponents));
betaPCRn = PCALoadings(:,1:NumberofComponents)*betaPCRn;
betaPCRn = [mean(Sat)-mean(Training_Set)*betaPCRn; betaPCRn];
yfitPCRn = [ones(size(Training_Set,1),1) Training_Set]*betaPCRn;

plot (Sat,yfitPLSn,'bo',Sat,yfitPCRn,'rx');
title ('Fitting more Components')
xlabel ('observedResponse');
ylabel ('Fitted Response');
lsline
legend ({'PLSR' 'PCR'},'location','NW') ;

TSS = sum((Sat-mean(Sat)).^2);
RSS_PLS = sum((Sat-yfitPLSn).^2);
Correlation_Coeficient_PLS=1-RSS_PLS/TSS;
RSS_PCR = sum((Sat-yfitPCRn).^2);
Correlation_Coeficient_PCR=1-RSS_PCR/TSS;

plot (Sat,yfitPLSn,'bo',Sat,yfitPCRn,'rx');
title ('Fitting more Components')
xlabel ('observedResponse');
ylabel ('Fitted Response');
```

```
lsline
legend ({'PLSR' 'PCR'},'location','NW') ;

TSS = sum((Sat-mean(Sat)).^2);
RSS_PLS = sum((Sat-yfitPLSn).^2);
Correlation_Coeficient_PLS=1-RSS_PLS/TSS;
RSS_PCR = sum((Sat-yfitPCRn).^2);
Correlation_Coeficient_PCR=1-RSS_PCR/TSS;
```

Plot: Variance Explained in Training Set vs Number of Principal Components

```
principalcomponents = 40;

[PCALoadings,PCAScores,PCAVar] = pca(Training_Set,'Economy',false);
[Xloadings,Yloadings,Xscores,Yscores,betaPLS10,PLSPctVar] = plsregress
(Training_Set,Sat,principalcomponents);

plot (1:principalcomponents,100*cumsum(PLSPctVar(1,:)),'bo');
xlabel ('Number of Principal Components');
ylabel ('Percent Variance Explained in Training_Set');
legend ({'PLSR'},'location','SE');
grid on

hold on
[LOADINGS,SCORE,LATENT] = pca(Training_Set);
x = var(SCORE);
y = 100*x/(sum(var(SCORE)));

plot ((1:principalcomponents),cumsum(y(1,1:principalcomponents)),'rx')
grid minor
xlabel ('Number of PCs')
ylabel ('percent age of total variance')


Number_of_PCS = 40;

[X1,Y1,Xs,Ys,beta,pctVar,PLSmsep] =
plsregress(Training_Set,Sat,Number_of_PCS,'CV',Number_of_PCS);

plot (0:Number_of_PCS,PLSmsep(2,:),'bo');
title ('Choosing Number of PCs');
xlabel ('Number of components');
ylabel ('Estimated Mean Squared Prediction Error');
legend ({'PLSR'},'location','NE');
grid minor

minPLSmsep = min (PLSmsep(2,:));




hold on
for i=1:Number_of_PCS
```

```
[PCALoadings,PCAScores,PCAVar] =
pca(Training_Set,'Centered',false,'Economy',false,'NumComponents',i);
betaPCRn = regress(Sat-mean(Sat),PCAScores(:,1:i));
yfitPCRn = PCAScores*betaPCRn+mean(Sat);
TSS = sum((Sat-mean(Sat)).^2);
RSS = sum((Sat-yfitPCRn).^2);
Correlation_Coeficientn(i) = RSS/TSS;
end

plot (0:Number_of_PCS-1 ,100*Correlation_Coeficientn,'r^');
legend ({'PLS','PCR'},'location','NE');
```

Regression - PLS

```
[n,k] = size(Sat);
Number_of_PCs = 32;

S = [6,7,18,20,25,34,35,37];
A = Training_Set;
A ([S],:) = [];
B = Sat;
B([S],:) = [];
[n,k] = size(B);

[Xloadings,Yloadings,Xscores,Yscores,betaPLS10,PLSPctVar] =
plsregress(A,B,Number_of_PCs);
yfitPLSn = [ones(size(A,1),1) A]*betaPLS10;

scatter (B,yfitPLSn,'ro')
grid minor
lsline
xlabel ('% Concentration of Saturates');
ylabel ('Xscores*X');
title ('PLS Regression efficiency')
hold on

yfitPLSn2 = [ones(size(Training_Set(S,:),1),1)
Training_Set(S,:)]*betaPLS10;

scatter (Sat(S),yfitPLSn2,'bo')

errors = B-yfitPLSn;

No = (1:1:n)';
table (No,B,yfitPLSn,errors)

TSS = sum((B-mean(B)).^2);
RSS = sum((B-yfitPLSn).^2);
RMSEP = sqrt(RSS/n);
Correlation_Coefficient = 1-RSS/TSS;

figure()
```

```
residuals = B-yfitPLSn ;
stem (residuals)
xlabel ('Samples');
ylabel ('Residual');
grid on
title ('Residuals PLS ALL')
```

Regression - PCR

```
Number_of_PCs = 32;

[n,k] = size(Sat);

S = [6,7,18,20,25,34,35,37];
A = Training_Set;
A ([S],:) = [];
B = Sat;
B ([S],:) = [];
C = (Sat(S));

[PCALoadings,PCAScores,PCAVar] =
pca(A,'Centered',false,'Economy',false,'NumComponents',Number_of_PCs);

sat = (B-min(B))/(max(B)-min(B));

betaPCRn = regress (sat-mean(sat),PCAScores(:,1:Number_of_PCs));

sat_new = (PCAScores*betaPCRn+mean(sat))*(max(B)-min(B))+min(B);

% Plot 90% of Data
scatter (B,sat_new,'bo')
lsline
xlabel ('% Concentration of Saturates')
ylabel ('PCAScores*x')
grid minor
title ('PCR Efficiency of prediction')
hold on

AA = Training_Set(S,1:end);
zz = AA*PCALoadings;
sat_C_Predicted = (zz*betaPCRn+mean(sat))*(max(B)-min(B))+min(B);

TSS = sum((B-mean(B)).^2);
RSS = sum((B-sat_new).^2);
Correlation_Coeficient = 1-RSS/TSS;

RMSEP = sqrt(sum((sat_new-B).^2)/size(B,1));
m = sum(B-sat_new)/size(B,1);
SEP = sqrt((sum((B-sat_new-m).^2))/(size(B,1)-1));
Error = (sat_new-B);
```

```
% Plot 10% of Data
hold on
scatter (C,sat_C_Predicted,'rx');
error = (C-sat_C_Predicted);
No = (1:1:8)';
table (No,C,sat_C_Predicted,error)
```