Modelling substances present in an amine scrubbing plant during CO_2 capture with aqueous monoethanolamines using multivariate data analysis and ATR-FTIR spectroscopy

Master Thesis in Process Technology

Henning M. Kvaløy

Department of Chemistry, University of Bergen

June 1st, 2018

Acknowledgment

I would like to thank TCM for providing samples, making this thesis possible and Bjørn Grung for supervising this master's thesis. You have provided excellent guidance through it all and have been very helpful with every question I have had.

I would like to thank my fellow students whom I share an office with, you have made this master's thesis a lot of fun, a special thanks to Helene I. Sjo for all the coffee breaks and discussions we have had on and off topic.

I would like to thank my friends and family for all their support and encouragement during this master's thesis. A special thanks to Louis Steigerwald for proof reading my thesis. Most of all, I would like to thank my fiancé Sofie M. Christiansen for proof reading my thesis and her unlimited support and encouragement.

Thank you,

Henning M. Kvaløy

Abstract

Since the 19th century, an increase in the earth's temperature has been recorded. The rise in temperature is caused mainly by anthropogenic greenhouse gases, such as CO_2. Industrial processes are the cause of 40% of the anthropogenic CO_2 emissions. To lower the concentration of greenhouse gases several carbon capture and storage systems have been introduced to power plants. Post-combustion carbon capture systems using chemical absorption by amines are the most promising solution as they can be retrofitted to already existing power plants.

This master's thesis is written in collaboration with Technology Center at Mongstad and aim to provide an inline model to predict CO_2 loading, monoethanolamine (MEA) concentration and the concentration of degradation products. Multivariate methods such as principal component analysis, partial least square and variable selection are applied to the data to build a model.

A model for inline analysis could potentially improve efficiency and decrease cost, as well as provide continuous monitoring of the process. This would optimize component regulations involved in the process, such as amine loading, the temperature of the stripper and flow rate of CO_2 into the system.

Total inorganic carbon, total alkalinity, and density have been investigated for both CO_2 -lean and CO_2-rich solutions. All models obtained have yielded in low root mean square error predictions (RMSEP), compared to the value of the response. The RMSEP values as a percentage of the average response value resulted in a maximum of 2,3 % for the total inorganic carbon model in the lean samples and a minimum of 0,003 % for the density model built on the rich samples.

List of abbreviations

ATR-FTIR	Attenuated Total Reflectance Fourier Transformed Infrared
CCS	Carbon Capture and Storage
EMSC	Extended Multiplicative Signal Correction
FTIR	Fourier Transformed Infrared
IR	Infrared
LV	Latent Variable
MEA	Monoethanolamine
PC	Principal Component
PCA	Principal Component Analysis
PLS	Principal Least Square
PRESS	Predicted Residual Error Sum of Squares
RMSECV	Root Mean Square Error Cross-Validation
RMSEP	Root Mean Square Error Prediction
RSD	Residual Standard Deviation
SR	Selectivity Ratio
TCM	Technology Center at Mongstad
TIC	Total Inorganic Carbon
TOT ALK	Total Alkalinity
VIP	Variable Importance Projection

Table of contents

1 Introduction

Carbon dioxide (CO_2) is one of the leading greenhouse gasses causing climate changes and global warming [1]. In the period from 1856 to 2005 the average warming rate per decade has been 0,045 °C, while in the period 1981 to 2005 the average warming rate per decade has been 0,177 °C [2, P.169]. The concentration of CO_2 has increased considerably over the last years, and industrial processes are responsible for about 40 % of the anthropogenic CO_2 emissions worldwide.

An increased global temperature results in melting of sea ice. Sea ice has a higher capability to reflect sunlight than sea water. Thus, by the melting of the sea ice more seawater can absorb energy and less sea ice can reflect it, resulting in more melting of the sea ice and rise in sea level [2, P.204].

To reduce the amount of CO_2 introduced into the atmosphere by industrial processes, several different Carbon Capture and Storage systems (CCS) have been developed and fitted to industrial power plants. In existing power plants, the most convenient form of CCS is to retrofit a post-combustion system in which CO_2 is removed from the flue gas. Chemical absorption techniques commonly employ the use of aqueous amine solutions. The most extensively investigated and employed amine for this purpose is monoethanolamine (MEA) [1].

The method currently employed in an amine scrubbing plant involves the collection of physical samples from the scrubbing apparatus, which are subsequently analyzed overnight. Unfortunately, a large amount of time and resources must be devoted to this procedure. Additionally, this procedure does not provide real-time process control, thus rendering optimization problematic [3].

An inline real-time measurement technique such as Attenuated Total Reflectance Fourier Transformed Infrared Spectroscopy, ATR-FTIR, coupled with multivariate methods could potentially improve the current situation drastically.

ATR-FTIR can measure samples in aqueous solutions in contrast to FTIR and is, therefore, the preferred measurement technique in aqueous solutions. ATR-FTIR is a fast measurement technique that simultaneously measures all wavelength. The spectrum of the in-line measurement can then be used in a multivariate model created by Partial Least Squares, PLS, to predict real-time measurements. This is a massive advantage as it provides real-time measurements for process control and optimization of the process.

2 Theory

2.1 Notation

Bold font and upper-case letter: $\mathbf{X}$ = matrix

Bold font and lower-case letter: $\mathbf{x}$ = vector

Bold font, upper-case letter and raised to the power of T: $\mathbf{X}^{\mathbf{T}}$ = matrix transposition

Bold font, lower-case letter and raised to the power of T: $\mathbf{x}^{\mathbf{T}}$ = vector transposition

2.2 CO_2 capture

Many different techniques are possible for CO_2 capture, and they can be applied at different stages in the process. In pre-combustion CO_2 capture, combustible gasses are created, and CO_2 is captured before the gasses are burned for power. The fossil fuel is gasified and reacted in a water gas shift reactor to create H_2 and CO_2 [4].

Post-combustion CO2 separation involves the capture of flue gasses produced by the combustion of fossil fuels. Both chemical and physical filtration methods have been developed. CO_2 is captured from a conventional energy generation of fossil fuels, therefore, these systems are particularly of interest as they can be retrofitted to already existing industrial plants.

2.2.1 CO_2 capture by amines

Amines are organic compounds where one or several of the hydrogen atoms of an ammonia molecule are replaced by an organic group. Amines can be divided into three groups; primary, secondary and tertiary. The classification of the amines is based on the quantity of organic groups connected to the nitrogen atom. Most amines are soluble in water if their organic group is not too large [5, P. 804-806].

Amines are the most commonly used chemical for CO_2 separation from flue gas, where the primary amine, monoethanolamine (MEA) has been the most extensively researched [1].

During the process of post-combustion separation, the flue gas is first cooled to 40-50 °C and then compressed before entering the absorber column to avoid pressure loss in the column. The flue gas enters the absorber column at the bottom where it will be met by aqueous amine solution from the top of the absorber column. At the bottom of the absorber column, the CO_2 rich amine solution is transported to the stripper, to regenerate the amines for recycling. The amines are regenerated in the stripper by thermal treatment with steam that has a temperature of 100-130 °C. This thermal treatment releases the CO_2, which then exits the top of the stripper column ready to be dried, compressed and stored. The CO_2 lean aqueous amine solution leaving the stripper is then cooled down in the heat exchanger before being recycled back to the absorber column. Heating of amines can form unwanted degradation products because they are unstable at high temperatures. Degradation of amines can also occur in the absorber column by oxidative degradation.

An illustration of an amine scrubbing plant can be seen in Figure 2-1.

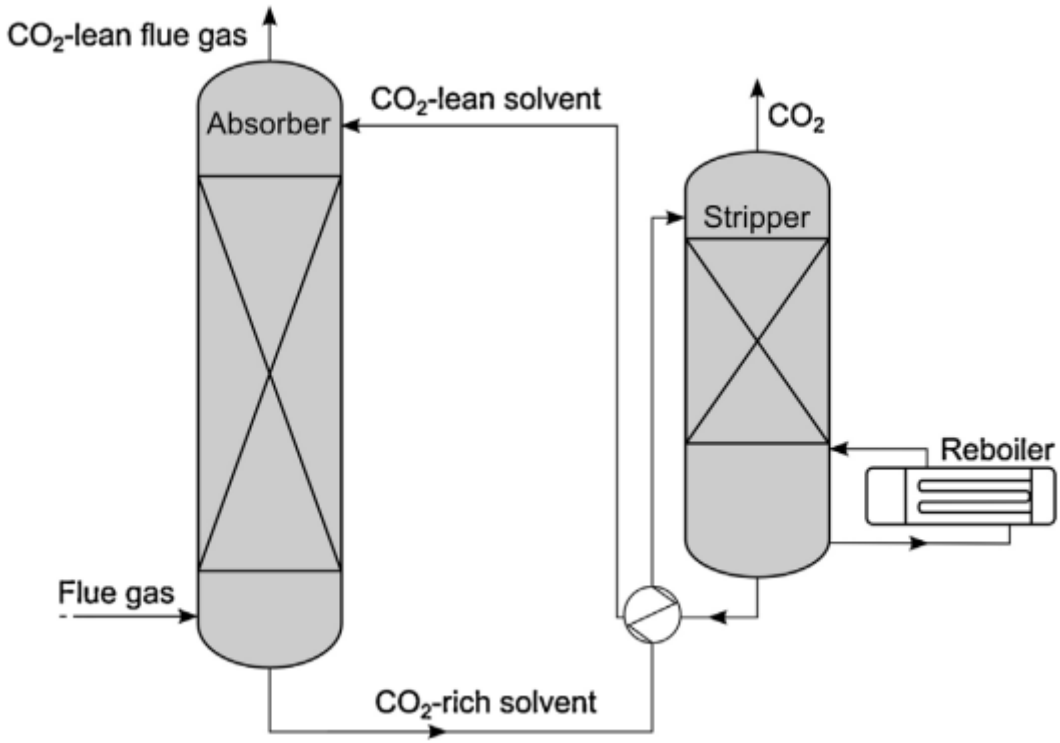


Figure 2-1: Illustration of an amine scrubbing plant [6]

Degradation of amines lowers the plant's capacity to absorb CO_2 due to the decreasing amount of amines capable of absorbing CO_2. MEA is the most used amine in aqueous amine carbon capture systems. Aqueous MEA solutions are highly CO_2 reactive and are therefore suitable for use with low CO_2 concentration (low partial pressure). The energy required by an aqueous amine carbon capture system is high due to the energy required to regenerate the amines. Aqueous amine carbon capture systems require steam for the regeneration of MEA and separation of CO_2, which lowers the plant's efficiency [7, 8].

Carbon capture utilizing an aqueous MEA solution involves either single-step (direct) or two-step zwitterion reactions. Two MEA molecules react with one CO_2 molecule to create, carbamate and protonated amine. The loading capacity of MEA is therefore 0,5 mol CO_2 per mol MEA [7].

The two-step process occurs when the CO_2 molecule's carbon attaches itself to the nitrogen of an amine molecule, forming a zwitterion intermediate formation. The zwitterion intermediate then reacts with a second amine molecule to form a carbamate and a protonated amine. In a single step reaction, the proton transfer and amine-CO_2 reaction happen at the same time.

A recent study from 2015 aimed to find the reaction mechanism of CO_2 capture with MEA. The study found that a zwitterion intermediate was created during the reaction between MEA and CO_2, thus the study found that the two-step zwitterion reaction mechanism occurs in the MEA and CO_2 reaction [9]. The underlaying reaction mechanism is presented in Appendix C. The overall reaction is presented in Equation 2-1.

$$2\,MEA + CO_2 \rightarrow MEACOO^- + MEAH^+ \qquad \text{Equation 2-1}$$

Illustration of these reaction is presented in Figure 2-2 and Figure 2-3.

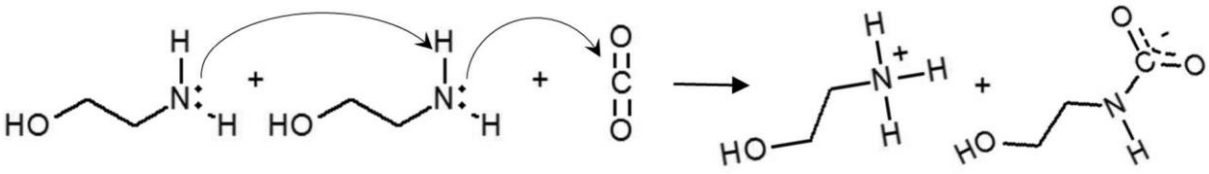
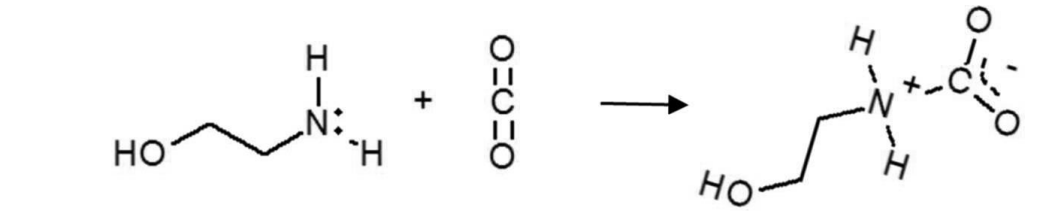
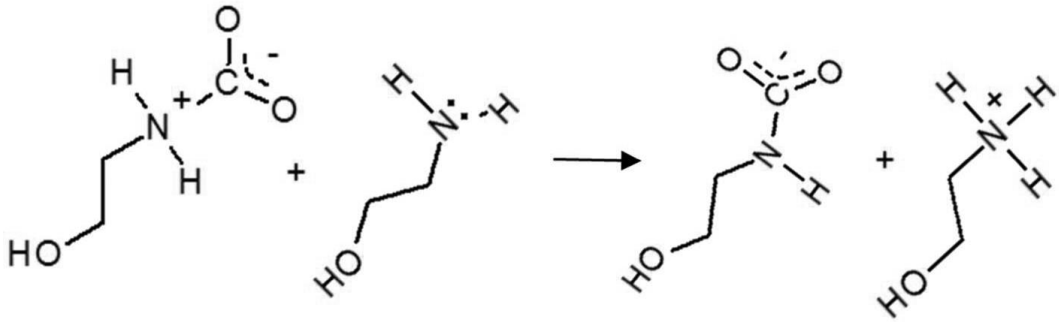


Figure 2-2: Illustration of a single step mechanism [7]



(a) Zwitterionic intermediate formation



(b) Deprotonation to form carbamate

Figure 2-3: Illustration of the two-step Zwitterion reaction [7]

2.3 The Electromagnetic Spectrum

The electromagnetic spectrum can be seen in Figure 2-4 and consists of all frequencies and wavelengths of electromagnetic waves that originate from a light source. A tiny portion of the electromagnetic spectrum is visible to the human eye, called the visible spectrum [10].

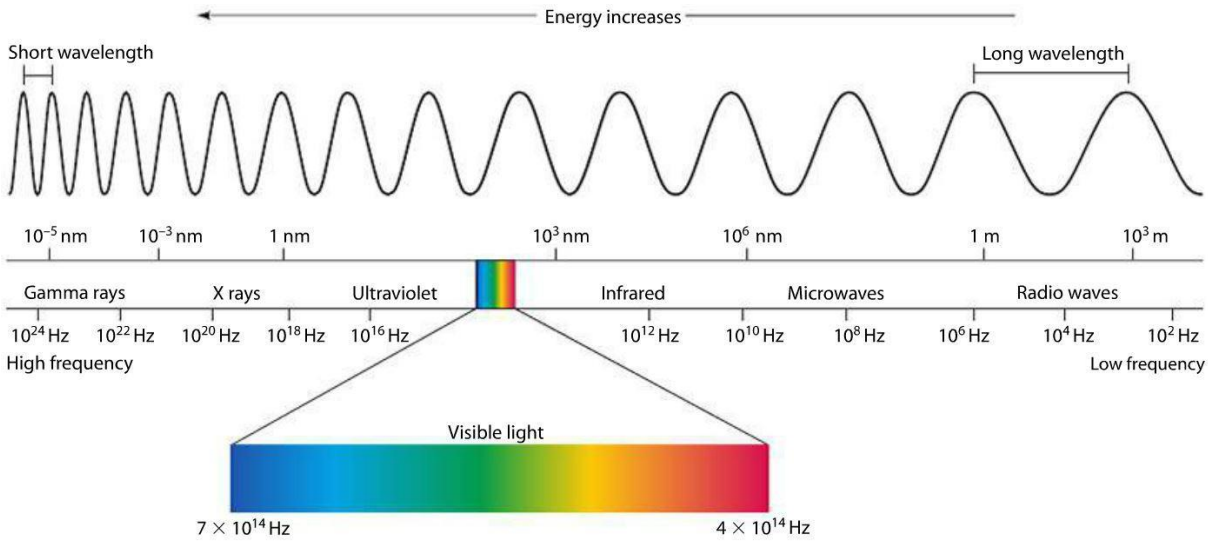


Figure 2-4 The Electromagnetic Spectrum [11]

2.4 Infrared Spectroscopy

The study of light and matter interaction is called spectroscopy [10]. Interaction of electromagnetic radiation with a sample can cause absorption, emission, transmission and reflection. If there is sample light interaction, the electron is excited to a higher energy state and the photon has been absorbed. If a photon is emitted while the excited electron relaxes to a lower energy state, emission is observed. The electromagnetic radiation can be reflected of the sample, the reflection of the sample depends on the physical properties of the sample surface where a medium difference plays a major role in reflection. Transmission is observed if the sample is transparent for a specific wavelength of light. The interactions between light and matter is illustrated in figure 2-5.

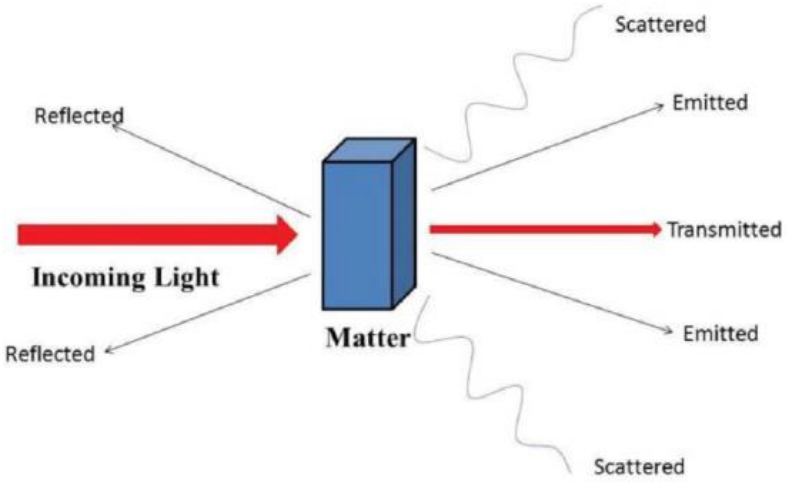


Figure 2-5: Illustration of the different interactions between light and matter [10]

Infrared spectroscopy, IR, is a measurement technique designed to identify and quantify the different molecules in a sample. Infrared spectroscopy or vibrational spectroscopy can investigate gases, liquids and solids that have a molecular dipole moment [10]. Two atoms bounded together within a molecule is never at rest but vibrates. The energy carried by a photon must be equal to a specific frequency mode of vibration in a molecule to create a change in dipole moment and be absorbed in IR [12, P. 14], bending and stretching are vibrational modes that can cause a change in the dipole moment. IR is a popular technique for analysis of samples due to its efficiency, it is non-destructive, sensitive and the sample preparation is easy [10].

Infrared radiation can be detected in the range of 14.000 to 10 cm^{-1} in the electromagnetic spectrum, [12, P. 13-14] where the most relevant region is the mid-infrared (4000-400 cm^{-1}), in which vibrational, rotational, bending and stretching modes is observed. Within the mid-infrared region, the fingerprint region is located. In the fingerprint region, each peak corresponds to one molecular vibration specific to a molecule [10,13].

From equation 2-2 we can see that the different wavelengths of light correspond to the different energy of the photons.

$$E = \frac{hc}{\lambda} \qquad \textit{Equation 2-2}$$

Where h is Planck's constant (h = 6,625 * 10^{-34} J s), c is the speed of light (300*10^{6} m/s) and λ is the wavelength. Wavenumber, $\bar{\nu}$, is extensively used in the field of IR and has the unit cm^{-1}, $\bar{\nu}$ is given by $1/\lambda$ [14].

Absorption of a photon by a molecule can only occur if the energy of the photon is equal to the energy level required to excite an electron to a higher energy level of the molecule. Thus, the chemical structure of a molecule can be determined by the absorbed frequencies of the IR radiation, and the different molecules in a sample can be determined [10, 15].

The major disadvantage of IR is that it is not suitable to measure aqueous samples due to the strong IR absorption of water [10].

2.5 Fourier Transformed Infrared Spectroscopy

Fourier transformed infrared spectroscopy, FTIR, performs better than dispersive IR spectroscopy [15, P. 148-151]. Dispersive IR techniques uses a prism to disperse the light and thus obtaining component wavelengths. Diffraction gratings can also be used in dispersive IR and consists of many very fine parallel lines in a transparent plate (can be several thousand per millimeter) that disperse the light [14]. In FTIR all wavelengths are measured simultaneously, thus reducing the measuring time required. This technique reduces noise and enhances sensitivity, thus producing a more favorable signal to noise ratio. This makes it possible to detect components at lower concentrations [16, P. 148-151].

FTIR is based on the interferometer. It uses the interference pattern of a measured sample in an interferometer and reconstructs the signal with Fourier transformation to a spectrum. The interferometer consists of an IR source, a beam splitter, a fixed mirror, a moving mirror and a detector (see Figure 2-6 for an illustration of the interferometer).

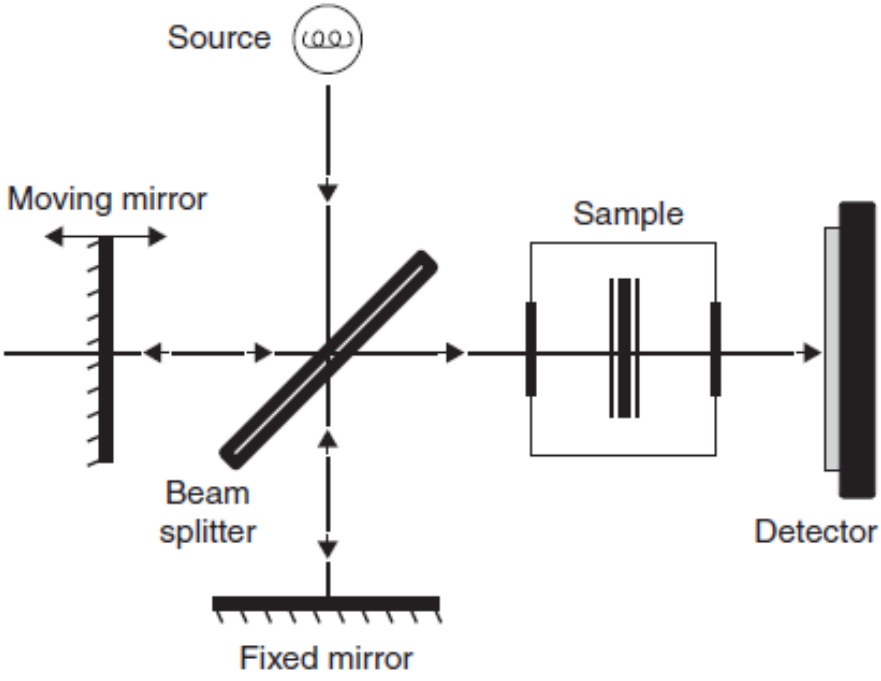


Figure 2-6: Illustration of an interferometer [16, P. 149]

The beam splitter lets half the light pass through to the fixed mirror and half passes through to the moving mirror. The moving mirror travels a length of Z to -Z. Thus the optical path length of the moving mirror will differentiate from the optical path length of the fixed mirror, except for zero path difference, where the optical path difference is zero. This change in path difference causes constructive and destructive interference once the two IR beams recombine at the beam splitter. The IR passes through the sample and is detected by a detector. An interferogram plot is then created by plotting the intensity of light, in volt, over the optical path difference [16, P. 149-150].

2.6 Attenuated Total Reflectance Fourier Transformed Infrared Spectroscopy

ATR-FTIR spectroscopy is based on absorption in the mid-infrared, (4000-400 cm^{-1}) region. The absorption of a photon causes the molecule to be excited to a higher vibrational state [17].

The energy required by a photon to excite a molecule is precisely equal to the energy difference between the higher vibrational state and the lower vibrational state. Thus, the absorbance of a photon at a given wavenumber provides information as to which molecules are in the sample.

In ATR-FTIR the measuring beam is not passed through the sample. Instead, the IR beam is reflected inside a highly refractive crystal on which the sample lies [10, 17]. Illustration of internal reflection in the crystal and the production of evanescent waves is presented in Figure 2-7.

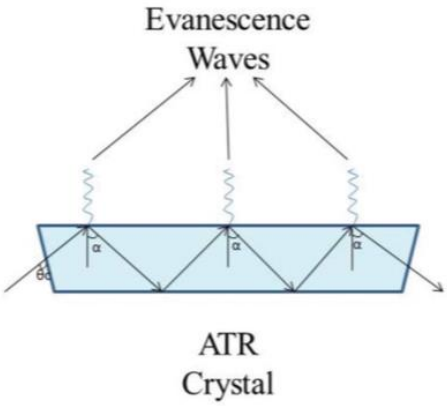


Figure 2-7: Illustration of ATR crystal and evanescence waves [10]

The internal reflections produce evanescent waves that penetrate the sample. This causes IR radiation and sample interaction, making it possible to obtain the spectrum of the sample.

This is a significant advantage over IR spectroscopy as samples with aqueous solutions can be measured. The preparation of a sample with ATR-FTIR is easier to do, as it only requires one drop of sample in the middle of the crystal. ATR-FTIR is also suitable to analyze the surface of a sample. Altering the reflection angle causes the penetration depth of evanescent waves to change [10].

ATR is less sensitive than transmission techniques. In a transmission technique light passes directly through the sample and a detector collects what comes out after there has been light-sample interaction. In ATR, evanescent waves interact with the sample, meaning that no light passes directly through it.

ATR can only measure the surface of a sample, due to the range of evanescent waves that are limited. If a sample requires depth to get accurate information the transmission technique is superior.

2.7 Multivariate Data Analysis

2.7.1 Pretreatment of data

The goal of pretreating data is to eliminate effects that do not reflect the chemical variation of the sample and to increase the signal to noise ratio. Such effects can be instrumental effects, light scattering effects, variation of the sample thickness or baseline variations [18,19, P. 129].

Pretreatment of the dataset can have an enormous effect on the result. However, the wrong pretreatment can destroy the information in the data set instead of improving it. Therefore, the choice of pretreatment should be considered based on knowledge of the data being investigated.

2.7.1.1 Extended multiplicative signal correction

The spectra of vibrational spectroscopy are subjected to many phenomena other than the chemical responses of the sample. These phenomena can be a challenge in the subsequent qualitative or quantitative analyses. The phenomena affecting the response can be random measurement noise, systematic errors, like interfering effects from unwanted physical and chemical variation or non-linear instrument response [20].

Extended multiplicative signal correction, EMSC, is a good method to correct the signal for multiplicative scaling effects, additive baseline effects and interference effects.
Baseline correction with EMSC is done by polynomial fitting to a reference specter, this is often the average specter, x_{ref}.

$$x_0 = b_0 + b_1 x_{ref} + b_2 \bar{v} + b_3 \bar{v}^2 + \cdots + b_n \bar{v}^{n-1} \qquad \textit{Equation 2-3}$$

Where the b's are the regression coefficients and $\bar{v}$ is the wavenumber.
The corrected spectra, x_c, is given by equation 2-4.

$$x_c = \frac{x_0 - b_0 + b_2 \bar{v} + b_3 \bar{v}^2 + \cdots + b_n \bar{v}^{n-1}}{b_1} \qquad \textit{Equation 2-4}$$

If done correctly, baseline correction can lead to simpler and better models and ease the interpretation of the data [19, p.130].

Equation 2 and 3 is the basic extension of the multiplicative signal correction and can be further extended. The regression coefficients are calculated using least squares [20].

2.7.1.2 Differentiation and smoothing

Savitzky-Golay is a numeric method for differentiation of a spectrum to eliminate additive and sloping baselines. A first-degree differentiation eliminates an additive baseline while a second-degree differentiation eliminates a sloping baseline [19, P. 131-132].

Differentiation and smoothing of a spectrum by Savitzky-Golay are performed by a moving window. The window size, w, can be altered to best accommodate various data types. In the moving window, a low-degree polynomial function based on least squares is fitted to the data.

The value of the center point, c_i, in the moving window is used in the polynomial function which is then differentiated to calculate the new value of the center point, c_i. The window is then moved one point, and a new value is calculated for the new center point, c_{i+1}. The operation only calculates new values for the center points, meaning that $^{w}/_{2}$ points are lost at the start and end of the spectra.

The polynomial of degree j can be described by equation 2-5.

$$g(\bar{v}) = b_0 + b_1\bar{v} + b_2\bar{v}^2 + \cdots + b_j\bar{v}^j \qquad \textit{Equation 2-5}$$

Where the regression coefficients, b, are calculated by the method of least squares. New values for the b's need to be calculated for each time the window is moved [18].

The signal-to-noise ratio is approximately improved by the square root of the window size. The Savitzky-Golay method increases the signal-to-noise ratio but makes it difficult to interpret the spectrum after transformation due to differentiation.

2.7.2 Variable and Object space

The data is collected in a matrix, **X,** with I rows and J columns. The matrix can be regarded as either being a composition by the $\mathbf{x}_j$ vectors (columns) or the $\mathbf{x}_i^T$ vectors (rows), this is illustrated in Figure 2-8. The $\mathbf{x}_j$ vectors are the variables and the $\mathbf{x}_i$ vectors are the objects, where each $\mathbf{x}_j$ is a vector in the I-dimensional space and each $\mathbf{x}_i$ vector is a vector in the J-dimensional space.

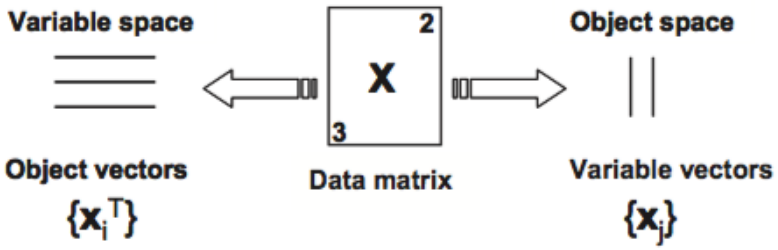


Figure 2-8: Variable and Object space [21]

The objects can then be plotted in the variable space to investigate the structure of the data and to find patterns, where the number of axes is equal to the number of variables. The same applies for the variables. They can be plotted in the object space where the number of axes is equal to the number of objects [21].

2.7.3 Latent Variables

Large datasets consisting of many variables make it difficult to interpret the data and build models. To ease this task, the number of variables can be reduced by latent variables, LV. LV can reduce the number of variables by creating new, and fewer variables that can describe most or all the variation in the original dataset [21].

Since it is difficult to handle several variables at once, a new vector is created, **t**, called scores, to replace the $\mathbf{x}_j$ vectors**.** Variation in matrix **X** is essential to solve a problem, it is therefore crucial that vector **t** explain as much as possible of the variation in the $\mathbf{x}_j$ vectors. If the amount of variation in **t** is sufficiently large, we can replace the $\mathbf{x}_j$ vectors with one vector **t**, where most of the variation is preserved.

The **t** vectors can then be described by equation 2-6.

$$\boldsymbol{t} = \boldsymbol{x_1} * \boldsymbol{w_1} + \ldots\ldots + \boldsymbol{x_j} * \boldsymbol{w_j} \qquad \textit{Equation 2-6}$$

Where **w** is a linear combination of all the measured variables. The score vector, **t**, is then found by projecting the objects on **w**. Vector **t** is a linear combination of the **x** variables in the same space as **x** and can be written as

$$\boldsymbol{t} = \boldsymbol{X} * \boldsymbol{w} \qquad \textit{Equation 2-7}$$

The next step is to maximize the variation of **t** by optimizing the weights $\mathbf{w_1,\ldots,w_j}$. An issue with this step is that the variance of **t** will increase if a large number is multiplied by **w**. It is therefore necessary to normalize the weights, **w**, to a constant sum, 1,0. This is the same as requiring that the sum of squared values equal 1,0. When **w** is optimized the first latent variable, LV, is found. How **w** is optimized is influenced by the procedure performed (PCA, PLS, etc.).

Now we introduce a new vector, **p**, called the loadings. Once the scores are optimized the loadings are found by projecting **X** on the scores, **t**. **X** can be written as

$$\boldsymbol{X} = \boldsymbol{t} * \boldsymbol{p^T} + \boldsymbol{E} = \boldsymbol{\hat{X}} + \boldsymbol{E} \qquad \textit{Equation 2-8}$$

$\mathbf{\hat{X}}$ is the product of $\mathbf{tp^T}$ and is a model of **X, E** is the residual matrix. **E** contain variation that is not explained by the latent variable and is found by Equation 2-9. Figure 2-9 illustrate the decomposition of the **X** matrix.

$$\boldsymbol{E} = \mathbf{X} - \boldsymbol{\hat{X}} \qquad \text{Equation 2-9}$$

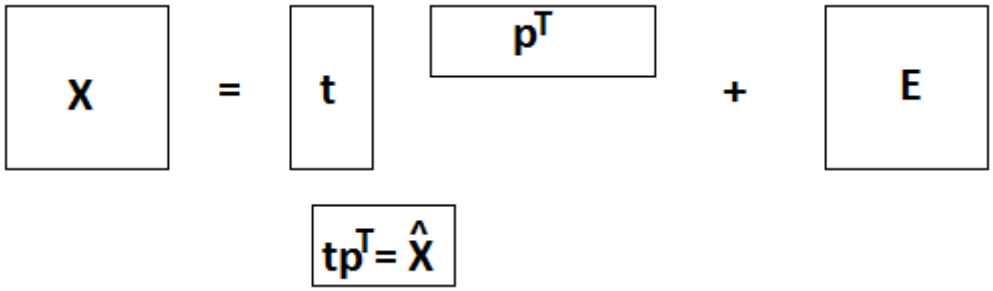


Figure 2-9: Illustration of decomposing matrix X. Based on Figure from [21]

The amount of variation explained by the LV can be calculated by Equation 2-10.

$$\frac{\|\mathbf{X}\| - \|\mathbf{E}\|}{\|\mathbf{X}\|} * \mathbf{100}\% \qquad \textit{Equation 2-10}$$

To find the second LV, the variation retained in the first LV must be subtracted from **X**. Equation 2-9 can thus be modified to generate Equation 2-11.

$$\boldsymbol{X_{new}} = \boldsymbol{X} - \boldsymbol{\hat{X}} \qquad \textit{Equation 2-11}$$

$\mathbf{X_{new}}$ is the new **X** matrix from which the second LV variable can be found. This operation can be repeated until all J latent variables in $\mathbf{X_{(I*J)}}$ is found [22]. Since none of the LV retain any of the same information their scalar product is zero and they are orthogonal to each other.

Vector **t** is the latent variable in the variable space and vector **p** is the latent variable in the object space. In Figure 2-10 the object and variable space with the latent variables are illustrated.

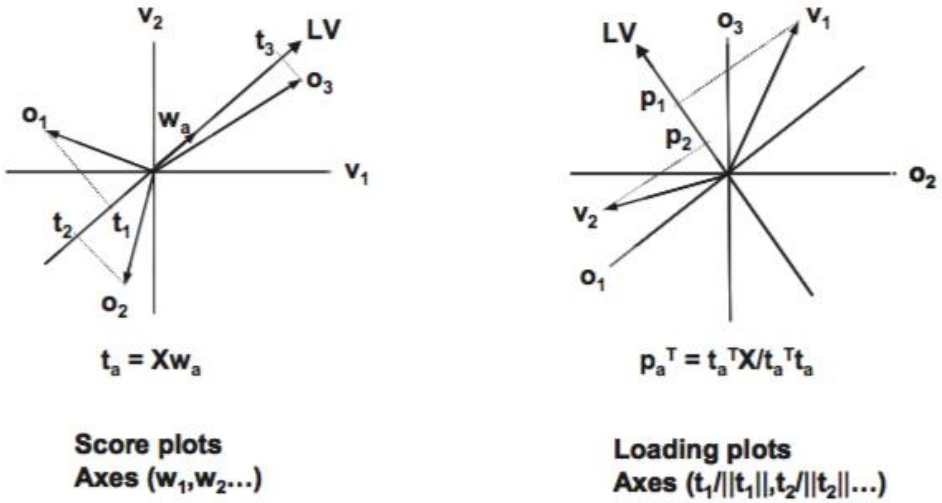


Figure 2-10: Illustration of the variable and object space with latent variables. [21]

2.7.4 Principal Component Analysis

Principal component analysis, PCA, is a method to reduce the number of variables to a minimum without losing important information. PCA uses latent variables to reduce the number of components, which are called principal components, PC. PCA is used to find the PCs that explain the most variation in the dataset [21, P.126, 21].

The NIPALS algorithm can be used to perform a PCA. NIPALS algorithm works by selecting a weight vector, $\mathbf{w_a}$, and projecting the objects on $\mathbf{w_a}$ to obtain the scores. See equation 2-12 [22].

$$\boldsymbol{t_a = X_a w_a} \qquad \textit{Equation 2-12}$$

The loadings are then found by Equation 2-13.

$$\boldsymbol{p_a^T = \frac{t_a^T X_a}{t_a^T t_a}} \qquad \textit{Equation 2-13}$$

The principal components in PCA explain the maximum possible variation and use the requirement that $\mathbf{w_a}$ equals $\mathbf{p_a}$. To find the maximum variation explained, the weight vector for PCA is calculated using Equation 2-14 [22].

$$\mathbf{w}_a = \frac{\mathbf{p_a}}{\|\mathbf{p_a}\|}$$ *Equation 2-14*

This new weight vector $\mathbf{w_a}$ is then used to calculate new scores and new loadings. The operation is performed until $\mathbf{w_a}$ and $\mathbf{p_a}$ converge. Once they have converged, the first PC has been found. The information retained in the first PC must be subtracted from $\mathbf{X_a}$ before the second PC can be found, see Equation 2-15.

$$\boldsymbol{X_{new}} = \boldsymbol{X_a} - \boldsymbol{t_a p_a^T}$$ *Equation 2-15*

Each object in the data set gets a score value on each PC, and each variable gets a loading value on each PC. The score and loading value of these PCs can be used to span a plane where the score values of the objects can be plotted in a score plot, and the loading values of the variables can be plotted in a loading plot. In the loading plot, two PCs span a plane and the loading values of the variables are projected on to this plane, creating a loading plot. The same applies for the score plot, the score value of two PCs span a plane where the score value of the objects are projected onto the plane, creating a score plot. In figure 2-11 a score-plot is illustrated, showing the new coordinate system of the score values of the objects plotted on PC 1 and PC 2. Score plots can uncover patterns, like groupings, outliers and trends. Similarities between two objects can be investigated in the score plot by the distance difference from the origin and by the angle between them. If the angle between two objects is zero degrees they are perfectly correlated, if the angle is 90 degrees they are not correlated at all and if the angle between them is 180 degrees they are perfectly negatively correlated. The same applies for variables in a loading plot [21, P.126, 22, 24].

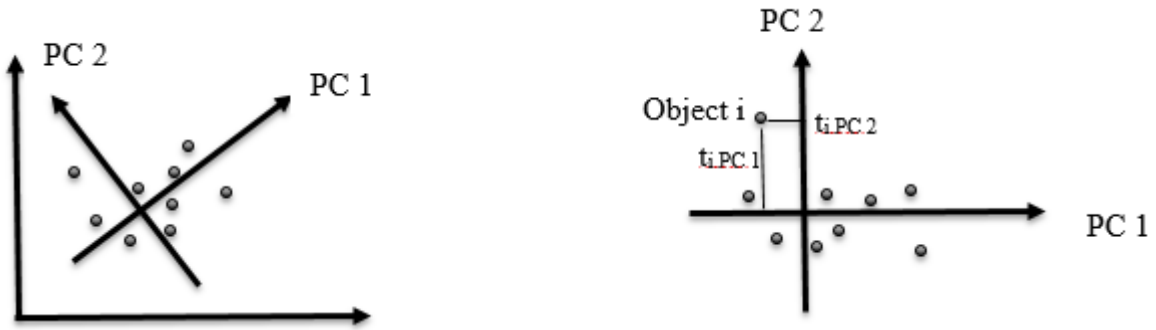


Figure 2-11: Illustration of the new coordinate system for the objects with PC 1 and PC 2. Based on figure from [21, P. 126]

2.7.5 Outliers

Outliers are samples that have a value that differs significantly from the rest of the sample values. Outliers can affect the result of a model. By including outliers, the model will be ''pulled'' towards the outliers and thus explain less variation in the samples that are not outliers. This causes the model's predictive power to decrease. The removal of outliers is therefore essential prior to model building [24].

PCA is a visualization technique to detect groupings, trends and outliers. PCA can be used to detect outliers using score-plot, a normal plot of the scores of the PCs and by a residual standard deviation versus leverage plot [22].

In the score-plot presented in Figure 2-12, an ellipse is provided by Sirius to detect outliers. The ellipse in Sirius is created by a Hoteling's T^2-test. Hoteling's T^2-test is a generalized version og the students t-test used for multivariate data, which utillize the object's score value to calculate how far each object is from the model center. The statistical limit calculated by the Hotelling's T^2-test is presented as an ellipse in the score-plot [25, 26].

Objects that are outside this ellipse are considered to have a too high deviation from the rest of the objects and are outliers [25]. The outliers will be removed prior to any model building. Example of outliers detected by the score-plot are object: 30 and 66.

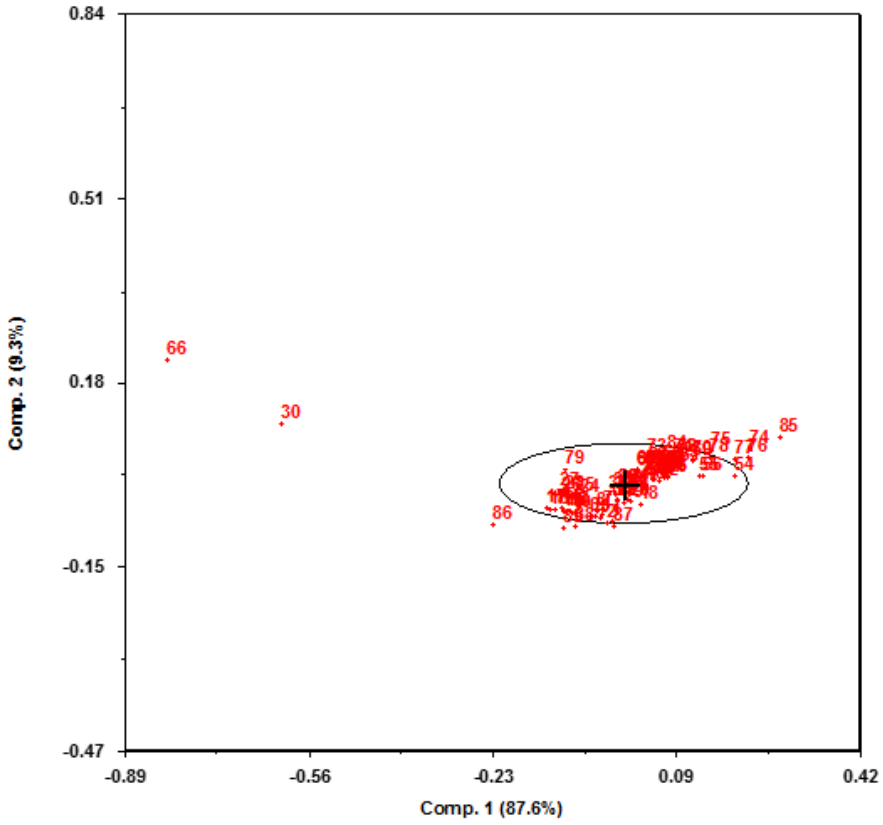


Figure 2-12: Illustrating outlier detection in a score-plot

In Figure 2-13 a normal plot of the scores of the first PC is provided.

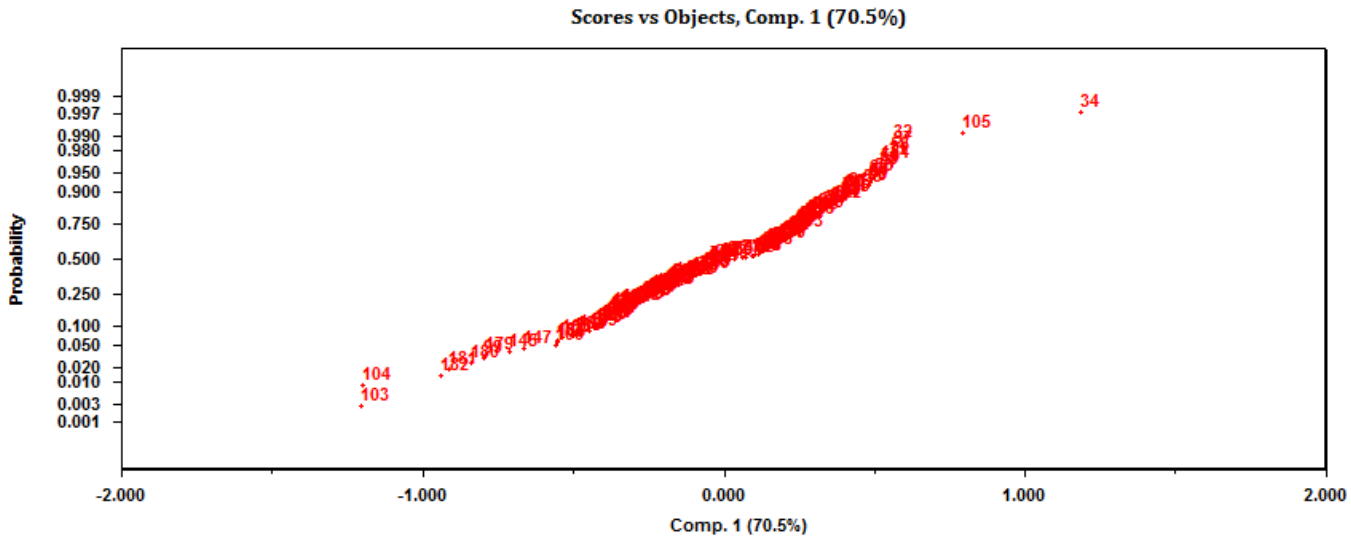


Figure 2-13: Illustration of outlier detection by a normal plot of the scores for the first PC

Outlier detection by a normal plot of the scores from the PC's is performed by looking for samples that do not fit the linear line. The most apparent outlier that can be seen in this

plot is sample 34. If the plot is not linear or close to linear, a pretreatment should be done to make them linear.

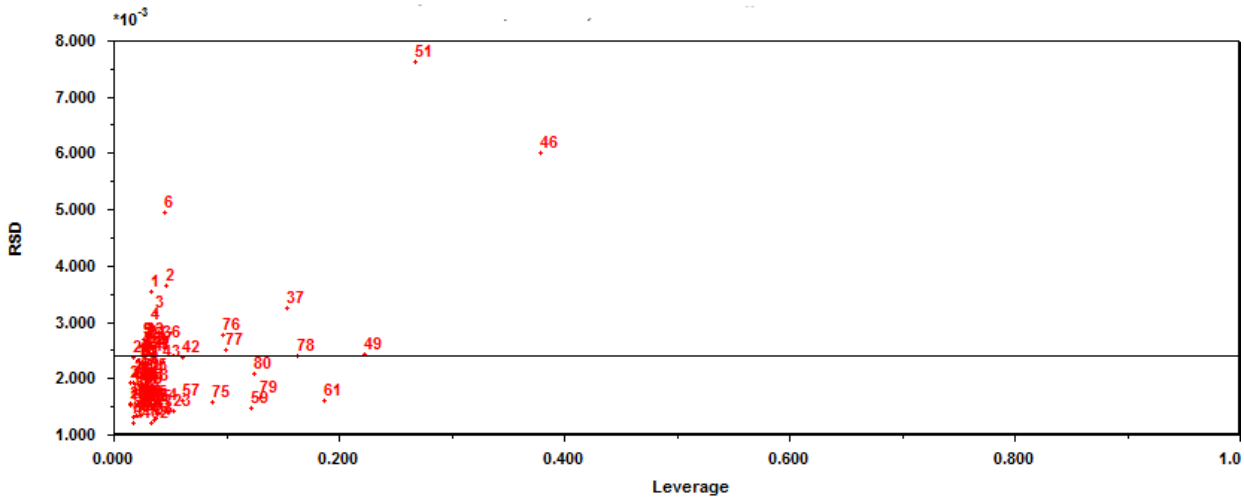


Figure 2-14: Illustration of outlier detection by a plot of RSD versus Leverage

The residual standard deviation (RSD) versus leverage plot in Figure 2-14 can be used to detect outliers by looking at the residual standard deviation, RSD, value and the leverage value. RSD is a measure of how good or bad a sample is fitted to the model; a high RSD value indicates poor agreement. Leverage is a measure of how much a sample influences the model, a high leverage value means that the object has a more significant influence on the model than the rest of the objects. Thus, a high leverage value and a low RSD value means that the outlier heavily affects the model and ''pulls'' the model towards the outlier [25]. The most apparent outliers in the plot are samples 51 and 46.

2.7.6 Partial Least Squares

Partial least squares (PLS) is a latent variable regression, which uses LVs to reduce the dimensions of the dataset and find the LVs that best explain the relation between **X** and **y**. PLS, was suggested as a better new alternative to principal component regression, PCR [22]. The issue with PCR is that the main latent variable could model variation in the **x**-variables which had little or no relevance to the response, **y**. PLS, like PCA, calculates latent variables to reduce the dimension. However, PLS does not use the same approach

as PCA for finding the LVs. In PCA, the LVs are calculated to explain the maximum amount of variation, while in PLS the LVs are calculated to find the best relation between **X** and **y**.

The weight vector for PLS is calculated by Equation 2-16.

$$\boldsymbol{w}_{\boldsymbol{PLS,a}}^{\boldsymbol{T}} = \frac{\boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{X}}{\|\boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{X}\|} \qquad \textit{Equation 2-16}$$

Scores for the PLS model is then calculated by projecting **X** on $\mathbf{w}^{T}_{PLS,a}$.

$$\boldsymbol{t}_{\boldsymbol{a}} = \boldsymbol{X}_{\boldsymbol{a}}\boldsymbol{w}_{\boldsymbol{PLS,a}}^{\boldsymbol{T}} \qquad \textit{Equation 2-17}$$

The loadings are found by the projection of **X** on the scores.

$$\boldsymbol{p}_{\boldsymbol{a}}^{\boldsymbol{T}} = \frac{\boldsymbol{t}_{\boldsymbol{a}}^{\boldsymbol{T}}\boldsymbol{X}_{\boldsymbol{a}}}{\boldsymbol{t}_{\boldsymbol{a}}^{\boldsymbol{T}}\boldsymbol{t}_{\boldsymbol{a}}} \qquad \textit{Equation 2-18}$$

The score and loading value of **y** also needs to be calculated, the scores are found by projecting **y** on $\mathbf{w}^{T}_{PLS,a}$.

$$\boldsymbol{u}_{\boldsymbol{a}} = \boldsymbol{y}_{\boldsymbol{a}}\boldsymbol{w}_{\boldsymbol{PLS,a}}^{\boldsymbol{T}} \qquad \textit{Equation 2-19}$$

The loadings, $\mathbf{q}_a$, are found by the projection of **y** on the scores, $\mathbf{u}_a$.

$$\boldsymbol{q}_{\boldsymbol{a}}^{\boldsymbol{T}} = \frac{\boldsymbol{u}_{\boldsymbol{a}}^{\boldsymbol{T}}\boldsymbol{y}_{\boldsymbol{a}}}{\boldsymbol{u}_{\boldsymbol{a}}^{\boldsymbol{T}}\boldsymbol{u}_{\boldsymbol{a}}} \qquad \textit{Equation 2-20}$$

The scores and loadings obtained from the first PLS component explain a part of **X** and **y**, this information needs to be removed for both **X** and **y** as shown in Equation 2-15 before the second PLS component can be calculated

2.7.7 Model validation

Overfitting is a strong possibility when the number of variables is significantly larger than the number of objects. When a model is overfitted, it models noise in the data and a false high correlation is observed resulting in poor predictive abilities. It is therefore necessary to test the predictive performance of the model. A validation set is used to test the model created by the training set [22].

The predictive ability of the model tested on the validation set can be investigated by calculating the root mean square error of prediction, RMSEP. A low value of RMSEP compared to the measured value means that the predictive ability of the model is good [27].

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \qquad \textit{Equation 2-21}$$

Where $\hat{y}$ is the predicted values from the model, y is the measured values and n is the number of samples.

Another essential aspect to consider in model validation, to ensure that the model is not overfitted, is the coefficient of multiple determination, R^2 and the adjusted coefficient of multiple determination, R^2_a. R^2 is the amount of variation in X that can be explained by the model [28, P. 686].

$$R^2 = 1 - \frac{SSE}{SST} \qquad \textit{Equation 2-22}$$

Where SSE is the error sum of squares and SST is the total sum of squares. SSE is calculated by Equation 2-23 and SST is calculated by Equation 2-24.

$$SSE = \sum(y_i - \hat{y}_i)^2 \qquad \textit{Equation 2-23}$$

Where $\hat{y}_i$ is the predicted value of y_i.

$$SST = \sum(y_i - \bar{y})^2$$ *Equation 2-24*

Where $\bar{y}$ is the mean value of the measured y-values.

R^2_a is calculated in such a way that if the number of components increase, the value of R^2_a decreases, however, if the increase of components leads to a lower sum of squared error the value of R^2_a increases.

$$R^2_a = 1 - \frac{n-1}{n-(k+1)}\frac{SSE}{SST}$$ *Equation 2-25*

Where n is the number of samples and k is the number of components in the model [28, P.631-633].

The value of R^2_a can never be larger than one and cannot excide the value of R^2. Thus, if the value of R^2 and R^2_a do not differ significantly, the model is not overfitted [28, P. 686].

2.7.7.1 Cross-Validation

Cross-validation is a method used to determine the number of components needed in a calibration model. Cross-validation works by leaving out one or more samples and PLS models with 1, 2,…,K number of components are calculated. Then, new samples are left out and new PLS models are calculated [27].

To determine how many components that should be included in the model the lowest predicted residual sum of squares, PRESS, is calculated for each model. The model with the lowest PRESS determines how many components that are needed.

$$PRESS_k = \sum_{i=1}^{I}\frac{(y_i - (\hat{y}_{(i),k})_i)^2}{I}$$ *Equation 2-26*

Where y_i is the ith element of $\mathbf{y}$ and $\hat{y}_{(i),k}$ is the estimate of $\mathbf{y}$ from PLS with k components when the ith observation has been eliminated [23].

From Equation 2-27 the Root Mean Square Error of Cross-Validation, RMSECV, is found.

$$RMSECV_k = \sqrt{\frac{PRESS_k}{n}}$$ *Equation 2-27*

The RMSECV values are plotted in an RMSECV plot as a bar graph. This is a visual presentation of how the number of components decreases the RMSECV value and improve the model. Once there is no significant decrease in RMSECV value of two components in the RMSECV plot, the number of components that should be used in the model is established [23].

A multitude of techniques exists for determining the optimal number of components. Among these, the most often utilized method in chemometrics is cross-validation [27].

2.8 Variable selection

Variable selection can improve model prediction, give an improved interpretation or lower the cost of measurements. Removal of variables that are irrelevant, noisy or unreliable can improve the predictive performance of the model or reduce the complexity of the model. To make the variable selection optimal, all combinations of variables should be tested. However, this is not practically possible due to the overwhelming amount of calculations needed and the risk of overfitting when the number of samples is not much higher than the number of variables [29].

To simplifiy variable selection, several methods have been developed to determine a suitable variable set. The choice of measurement instrument is the most critical variable selection to be made. The result of any modeling will be hugely affected by whether the right or the wrong instrument was used, due to the limitations of the selected instrument. An example of this is to use IR spectroscopy on aqueous solutions, the signal from the water can bury the signal from any other components due to the high absorption of water in IR spectroscopy and is therefore not suitable. ATR-FTIR can however measure samples in aqueous solutions.

When working with spectral data, the wavelengths are correlated to the neighboring wavelength. Thus, one wavelength cannot be chosen to explain one component because one component influence more than one wavelength. Several neighboring wavelengths or windows of wavelengths are better to use.

Outliers can be a pitfall for variable selection if not handled properly. Many methods of variable selection are based on small differences in model quality or statistics like the significance calculated from model parameters. This makes variable selection very sensitive to outliers and the wrong variables could be selected if outlier detection is not done correctly.

2.8.1 Variable Importance Projection

Variable importance projection, VIP, is a method for variable selection. It is a measure of how much a variable contributes to describe the dependent Y and the independent variables X. The VIP value of a variable is calculated by equation 2-28 [29].

$$VIP_j = \sqrt{\frac{\sum_{f=1}^{I} w_{ji}^2 * SSY_i * J}{SSY_{total} * I}} \qquad \textit{Equation 2-28}$$

w_{ji} is the weight value for variable j and component i. SSY_i is the sum of squares of explained variance for the i^{th} component. J is the number of variables. I is the number of components and SSY_{total} is the total sum of squares explained of the dependent variable.

Covariance between independent and dependent variables are reflected by the weights in a PLS model. VIP can reflect how well the dependent variable is described and how important that information is for the model of the independent variables due to the inclusion of the weights [25]. The VIP limit for non-important variables is usually set to 1,0. Instead of removing every variable with a VIP value below 1,0, the VIP limit is set low so that only a few variables are removed. The VIP procedure is repeated until there is no further model improvement. Exclusion of all variables with a VIP value below 1,0 could potentially remove variables that should be included in the model.

2.8.2 Selectivity Ratio

Selectivity ratio, SR, is a method of variable selection. SR describes the explained variance compared to the residual variance. The SR value of a variable is calculated using Equation 2-29 [30]

$$SR_j = \frac{v_{exp,j}}{v_{res,j}} \qquad \textit{Equation 2-29}$$

Where $v_{exp,\,j}$ is the explained variance of variable j, and $v_{res,\,j}$ is the residual variance of variable j.

A high SR value indicates variables with a high explained variance suitable to include in a model, low values indicate variables with low explained variance.

2.8.3 Manual Selection of Wavenumber Regions

MEA can be detected in the wavenumber region 3500-2500 cm^{-1}, where among C-H bonds can be detected at wavenumbers 2927 cm^{-1} and 2864 cm^{-1}. MEA can also be detected in the wavenumber region from around 2000-900 cm^{-1} where among C-N (1081 cm^{-1}) and C-O (1033 cm^{-1}) bonds can be detected. The absorption of MEA in the wavenumber region 3500-2500 cm^{-1} is however buried by the strong absorption of O-H bonds in water. The wavenumber region 2000-900 cm^{-1} is thus a better choice to be investigated and used for model building purposes of an MEA/CO_2 system [31]. In Figure 2-15 a single CO_2-lean spectrum is provided, showing some of the chemical bonds detected by ATR-FTIR is the CO_2-lean solution.

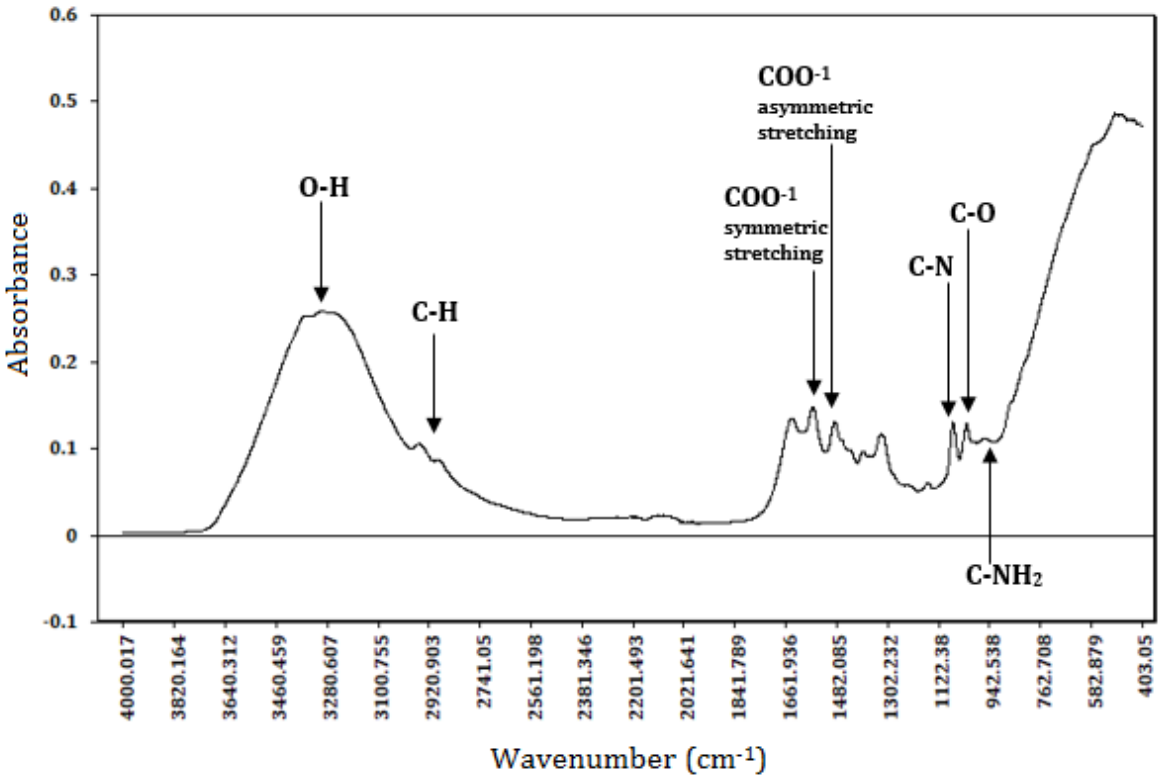


Figure 2-15: Illustration of chemical bonds detected by ATR-FTIR based on [31], created by Helene I. Sjo

3 Method

The data in this master thesis is provided by the Technology Center at Mongstad (TCM). Two main datasets are provided, mea2 and mea3. The mea2 dataset provides measurements from the 13th of July to the 19th 2of October 2015. The mea3 dataset provides measurements from the 13th of June to the 5ft of September 2017. There is no difference in procedures regarding the process or MEA concentration for these two datasets, therefore they have been merged to create a larger dataset providing more samples for each model. The new dataset is called mea2&3.

The samples provided can be classified into CO2-rich samples and CO2-lean samples. Measurements have been done on the aqueous MEA solution both before CO_2 interaction (lean samples) and after CO_2 interaction (rich samples) with the aqueous MEA solution. The number of samples available is different for each of the response variables investigated. The density model built from the lean samples had 154 samples, the total inorganic carbon model built from the lean samples had 202 samples and the total alkalinity model for the lean samples had 187 samples. The density model built from the rich samples had 90 samples, the total inorganic carbon model built from the rich samples had 92 samples and the total alkalinity model for the rich samples had 80 samples. The wavenumber region available for all the datasets is 4000 – 400 cm^{-1}. The number of variables in the wavenumber region is 2539 variables, which corresponds to wavenumbers.

The data from the measured quantity of the substances and the different techniques used were all in one document and had to be sorted. This was done by writing a MATLAB code to sort all data in one matrix (the code attached in Appendix A). The matrix (dimensions of I x J) consist of I samples and J responses, where each response refers to a measurement technique and the measured value of this. Thus, the point (1,1) in the matrix refers to the measured quantity of the first analysis of the first sample.

Spectral data was sorted in a similar manner. Each spectrum was in an individual file and had to be merged into one matrix. This operation was done by writing another MATLAB code. This code opens each file in turn and constructs a matrix (of dimensions M x N)

where there are M samples and N wavenumbers (the code is attached in Appendix B). The point (1,1) in the matrix represent the intensity measured at the first wavenumber for the first sample.

The data were then imported to Microsoft Excel where an excel file was created for each response variable. Each excel file contains sample numbers, analysis type, values of the response variable being investigated, wavenumbers and the measured absorption value corresponding to the wavenumber and sample.

The excel files were then imported to Sirius (version 11.0) to be studied. Firstly, the data had to be pretreated to eliminate effects that reflect non-chemical variation in the data, such as instrumental effects, variation of the sample thickness or baseline variations.

Multiple pretreatment techniques have been applied to each data set to investigate which one works best. EMSC and Savitzky-Golay are the two methods which have given the best results.

After pretreatment, a principal component analysis (PCA) was done to find patterns in the data and to find outliers. The outliers were found by score-plots, RSD vs. Leverage plot and by a normal plot of the scores for each PC.

When the outliers had been removed a training and validation set was created. The training set is used to build a model and consists of about half the samples available excluding outliers. The model is made by Partial Least Square (PLS) in Sirius, where the dependent variable is the response. The settings for PLS were made to be 100 iterations and a maximum number of components equal to 10. To select the required number of components, several factors were considered, such as explained variance for each component, the cross-validation value of each component and an RMSECV-plot. The RMSECV-plot provides a plot of how many components are needed in the model. The explained variance value of each component is a measure of how much variation in the dataset the components explain. The higher the value, the better. Cross-validation standard deviation is a measure of how good a component is to predict the measured value of the data used to build the model. A low value indicates that the prediction power

is good while a value over one indicates that the predictive power of the component is weak, and thus, should not be included in the model.

The number of components is then selected, and the model is built. Successful models are expected to generate residual values with a normal distribution. A normal probability plot of residuals is constructed using Sirius for verification. The plot should be close to linear and go through the point y = 0,5.

The model is then tested against the validation set, the data that were not used to create the model. To check if the model is a good fit the predicted value is plotted against the measured value. A successful model should result in a linear plot. If the model is not a good fit variable selection can be implemented to improve the model. Several methods for variable selection exist; this master's thesis incorporates VIP, SR and the manual selection of wavenumber based on which bonds that can be found using ATR-FTIR in the available region. The manual selection of wavenumbers gave the best results. Further variable selection by VIP and SR was performed to increase the model's predictive power.

After variable selection, a new model is built using PLS to assess if the model improved its predictive power. Variable selection can be done many times until there are no more improvements on the models.

3.1 Software

The spectra provided by TCM of the samples is measured using an ALPHA-P ATR-FTIR spectroscope. The measured range is from 4000 to 400 cm^{-1} with a resolution of 4 cm^{-1} and an interferogram size of 10 452 points.

The codes generated in this thesis is done with MATLAB R2017b (The MathWorks, Natick Massachusetts, USA)

The multivariate data analysis is performed by Sirius version 11.0 (Pattern Recognition System AS, Bergen, Norway)

4 Results and discussion

Of the variables provided by TCM, only total inorganic carbon (TIC), total alkalinity (TOT ALK) and density provided sufficient data to construct viable models. The models built on these variables are separated into lean and rich samples. As they are measured at different stages in the process the content of the sample will be different. Thus, models for lean and rich had to be created for each variable.

4.1 Lean samples

The lean samples are obtained with an ATR-FTIR spectroscope before flue gas has interacted with the amine solution. The CO_2 has been separated from the MEA solution in the stripper, an inline measurement is collected as the MEA solution is in route to the absorber cylinder. The best model obtained from the lean samples is presented in the results and is compared to models of similar quality. The plot of the lean samples is obtained in Sirius (11.0) by plotting the intensity against the wavenumbers, see Figure 4-1.

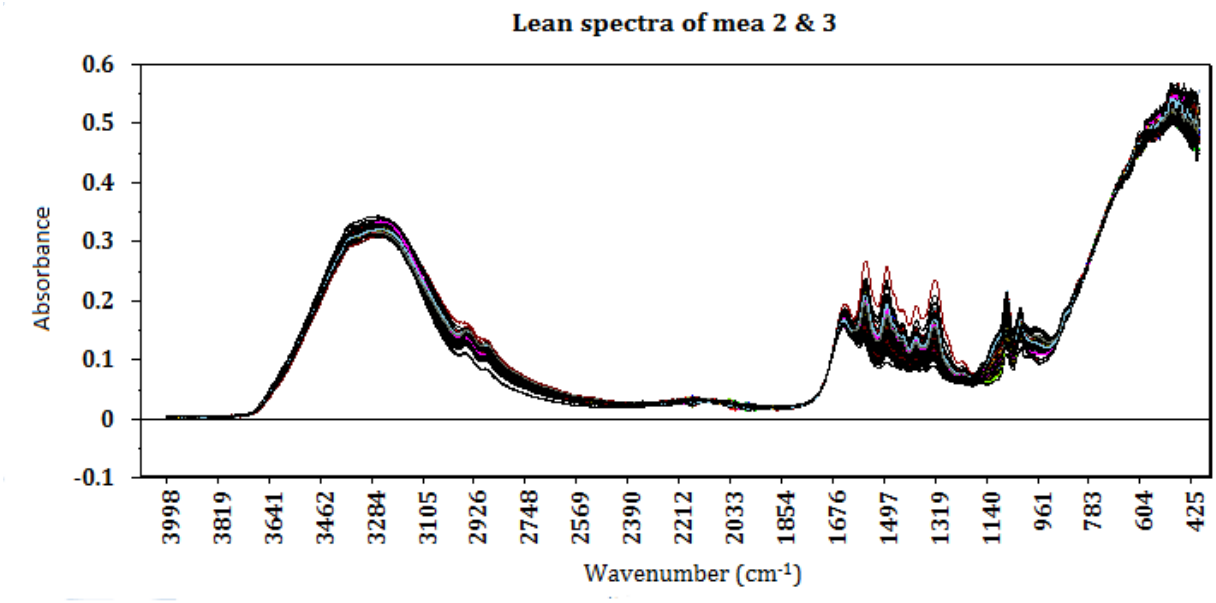


Figure 4-1: Lean Spectra of the mea2&3 dataset

As can be seen from figure 4-1, the organic C-H bonds at 3000 – 2800 cm^{-1} are not completely buried by the O-H bond absorption from water in the lean samples and it is possible to get information from this region.

4.1.1 Total Inorganic Carbon

The spectra of the lean samples from dataset mea 2 & 3 is obtained in Sirius (11.0) by plotting the intensity against the wavenumber. Several methods of pretreatment have been applied to the spectrum to make the results of a model better. The method that gave the best results is Savitzky-Golay pretreatment with a moving window of 21-points, 3rd degree polynomial fitting with a 2nd degree differentiation.

The size of the window has been tested for all values up to 25-points. Where a 21-point window seemed to be the best fit, and the 25-point was a good second. The 21-point and the 25-point window both gave an RMSEP value of 0.027 with three components. The choice to use a 21-point window is based on the number of outliers in the model, the 21-point window had six fewer outliers than the 25-point window and was therefore chosen.

2nd degree Extended Multiplicative Signal Correction, EMSC, also gave good results. The difference in root mean square error prediction, RMSEP, was only 0.001. Pretreatment with Savitzky-Golay produced fewer outliers in the model and in the predicted values. Thus, this model has better predictive power in a wider range of samples and for samples which might be weak outliers. The pretreated spectra are presented in Figure 4-2 and the selected wavenumber region is presented in Figure 4-3.

The response variable Total Inorganic Carbon (TIC) is used to determine the CO_2 concentration in the aqueous MEA solution, where the unit of TIC is given as moles/kg.

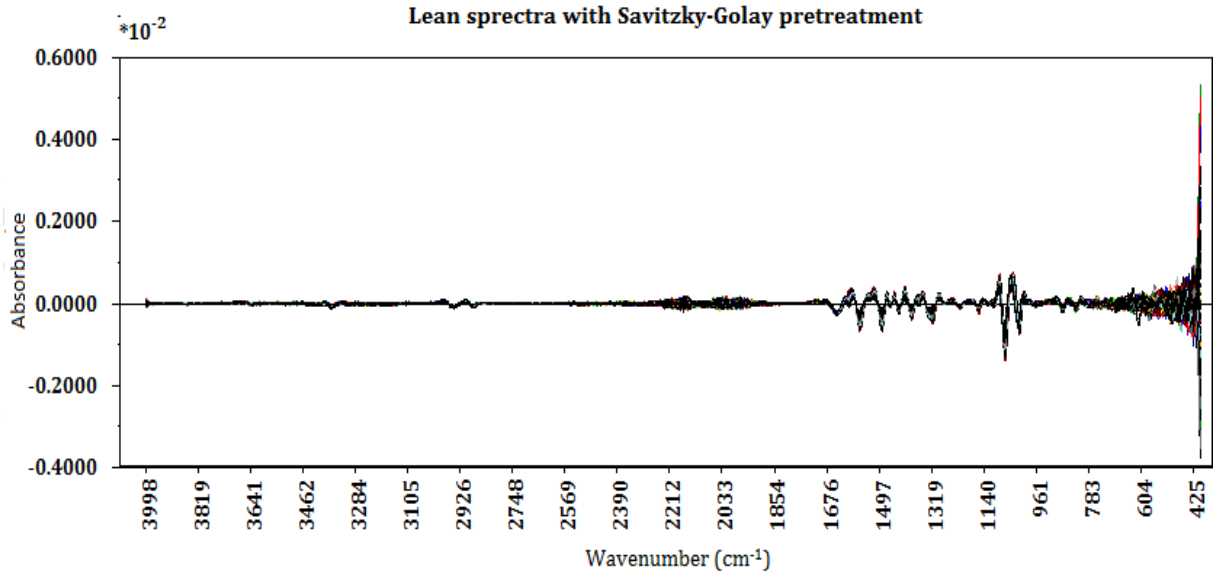


Figure 4-2: Lean spectra with a 21-point moving window, 3rd degree polynomial fitting and 2nd degree Savitzky-Golay pretreatment

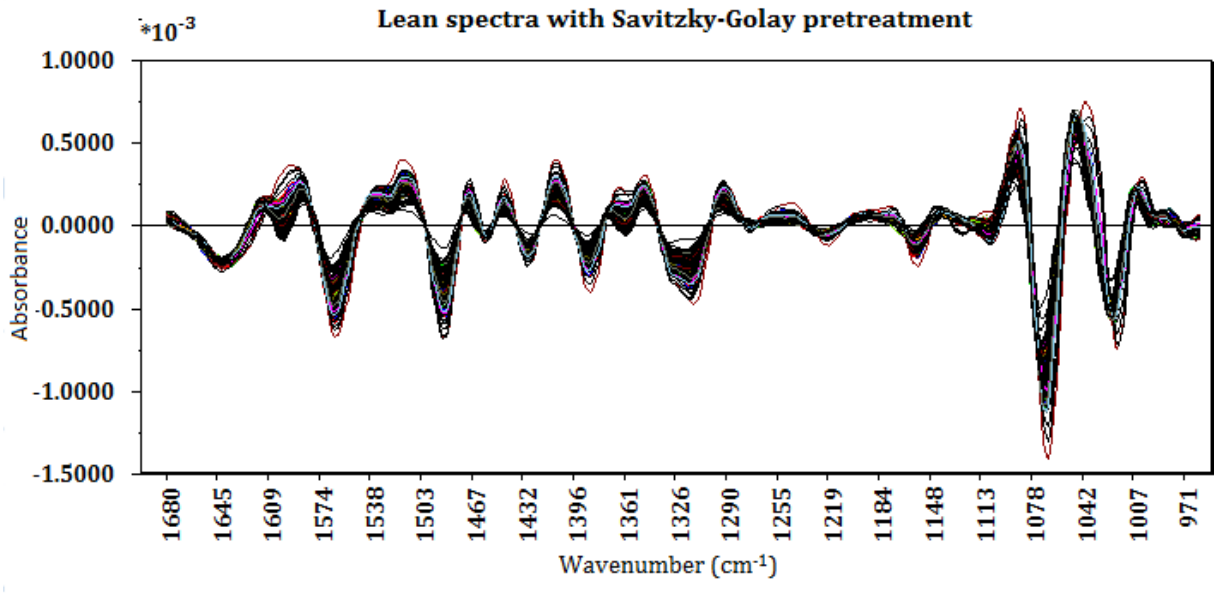


Figure 4-3: Wavenumber region 1680-960 of the pretreated spectra

Several methods for variable selection have been tested, such as VIP and SR. These methods test each individual variable, or wavelength, independently. This might not be the best variable selection process as the wavelength in the spectra is correlated to the neighboring wavelengths. The software does not always recommend an appropriate

scope of wavenumbers, therefore, additional manual selection was at times necessary. The variable selection that has yielded the best result in a Savitzky-Golay pretreatment when analyzing TIC was the manual selection of wavenumbers from 1680 to 960. The organic bonds found at 3000-2800 cm^{-1} is not included as a model for inorganic carbon is investigated. The score-plot of PC 1 and PC 2 is presented in Figure 4-4.

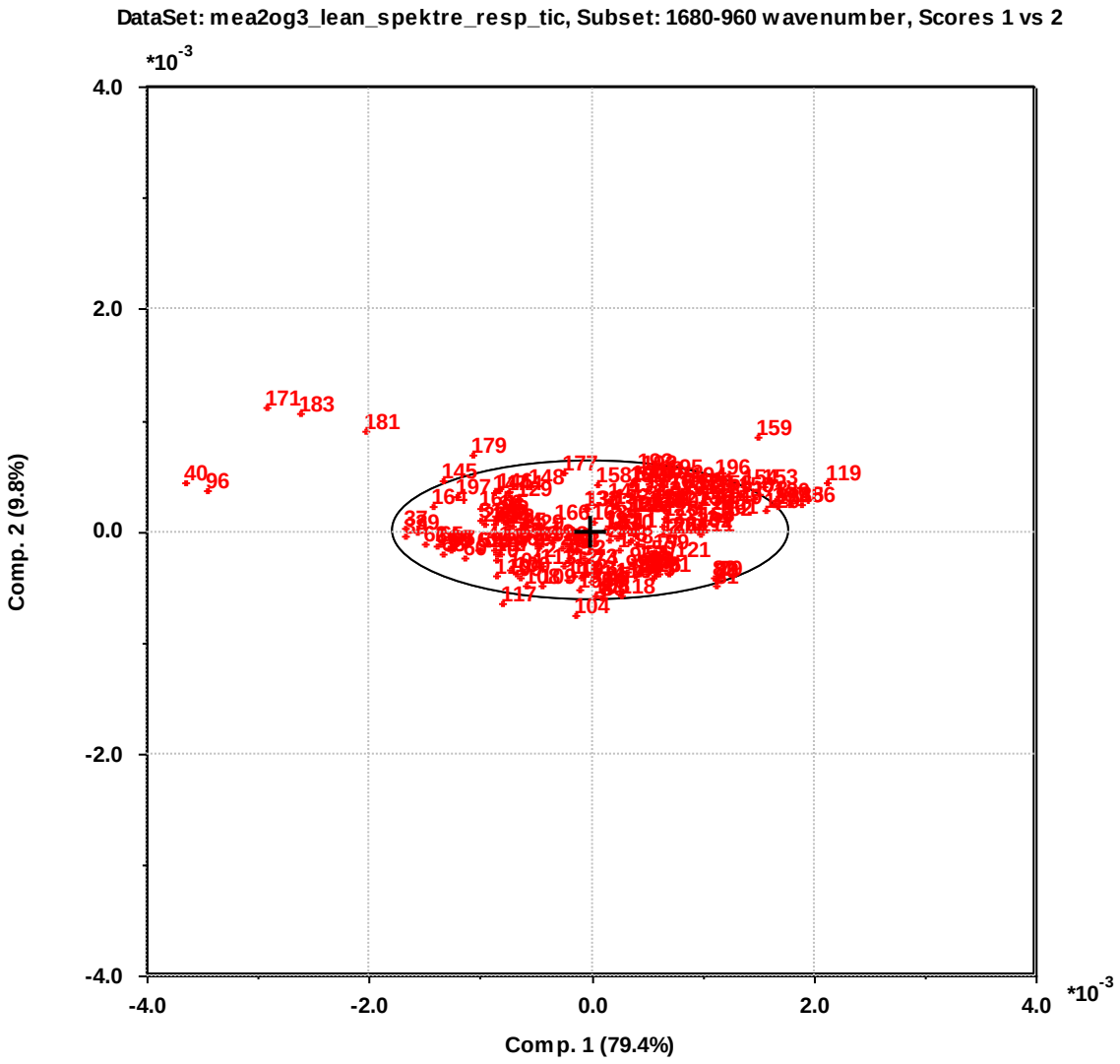


Figure 4-4: Score-plot of PC1 and PC2

As presented in the theory, the ellipse represents a measure for deviation of samples. Samples outside this ellipse have a higher deviation than accepted and are considered outliers. Together with the RSD versus Leverage plot and the normal plot of the scores for the PCs the outliers are detected and removed from the data set.

See Table 1 for a complete list of outliers. The largest outliers are removed first, then a new PCA is performed, and the largest outliers are removed again.

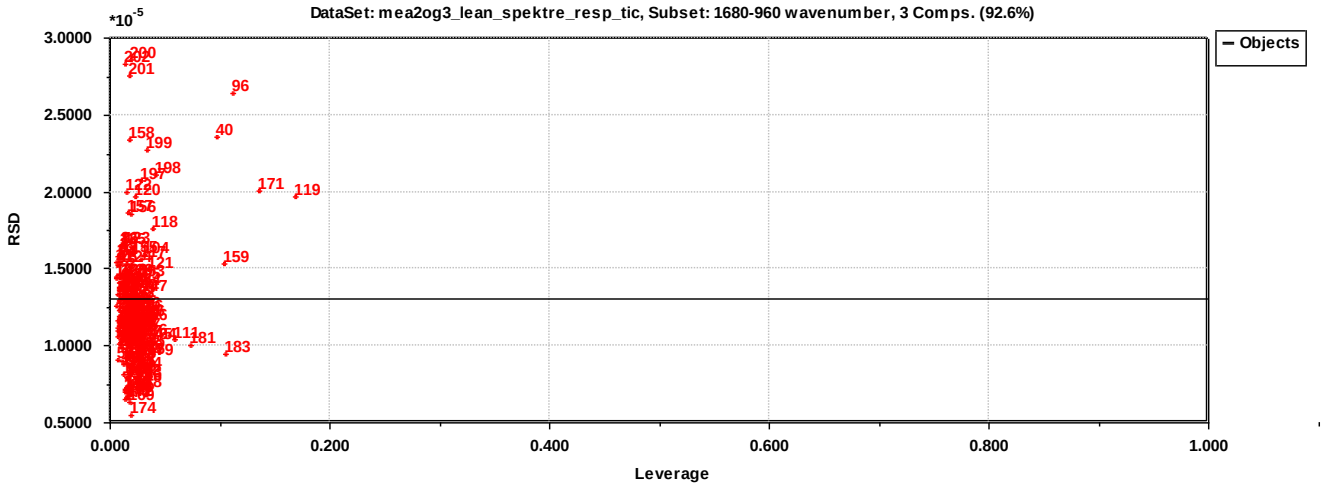


Figure 4-5: RSD versus Leverage for a three-component model

The residual standard deviation versus the leverage plot, Figure 4-5, does not indicate any outliers as there are no values with a low RSD and high Leverage.

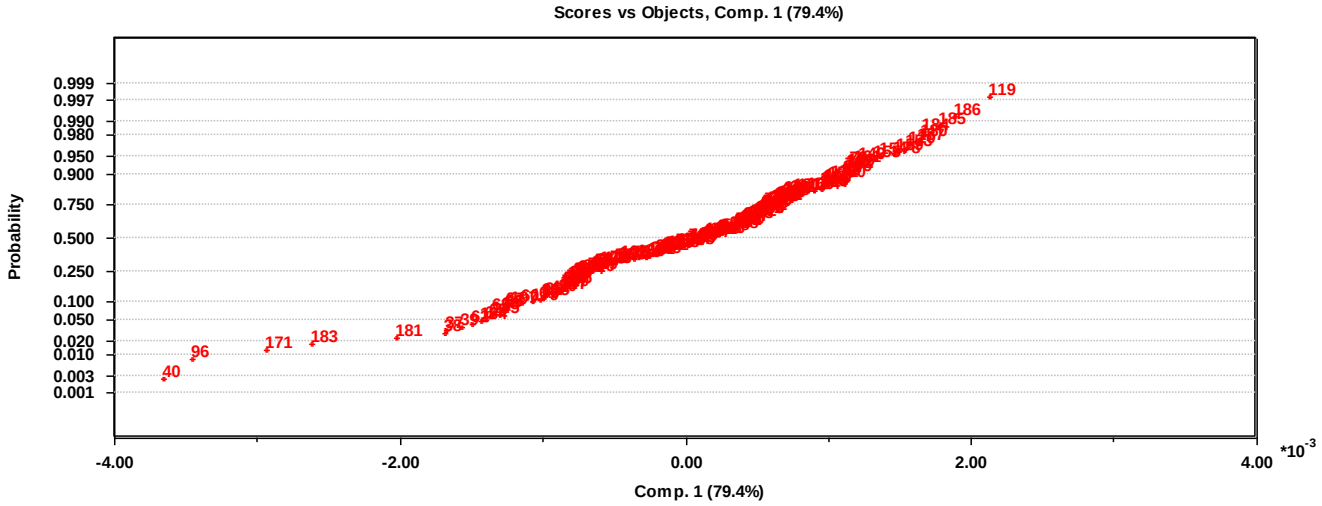


Figure 4-6: Normal plot of scores of the first component

In Figure 4-6, a normal plot of the scores for the first component is presented. There can be seen a ''tail'' at the beginning of the plot. This is an indication of outliers as the scores should be close to linear. Sample 40 and 96 are among the outliers that can be detected from this plot.

Table 1: Outliers in lean TIC model

Outliers
40, 70, 96, 119, 145, 159, 171, 173, 179, 181, 183.

Eleven samples are removed from the original 202 samples. Of the remaining 191 samples, 95 samples are used in the training set to build the model, and 96 samples are used to create a validation model. The training set is created in Sirius by selecting every odd-number sample in the data set, excluding outliers. The rest of the samples excluding outliers is used in the validation set to validate the model.

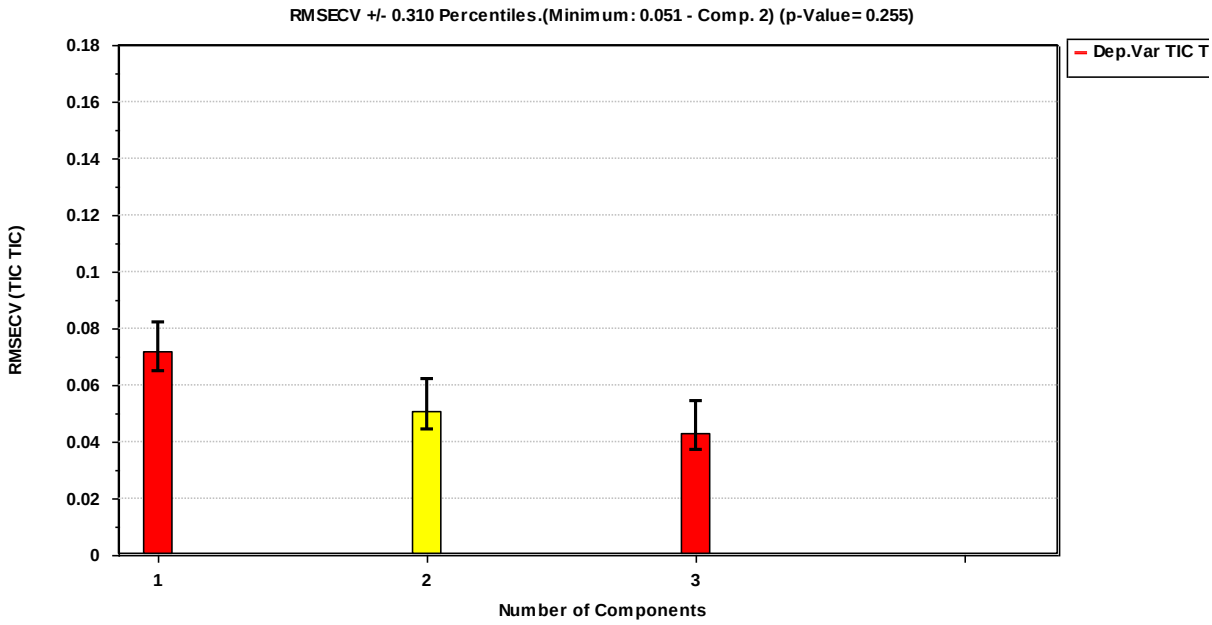


Figure 4-7: RMSECV plot

In Figure 4-7 the yellow bar in the RMSECV plot indicate how many components Sirius suggest using in the model, the RMSECV plot suggests using two components. A response residual plot was created for both two- and three-component model. The three-component model exhibited the greatest response residual improvement and was therefore chosen. Component information can be found in Table 2.

Table 2: Overview of components used in the model, their explained variation and cross-validation value:

Component number	Explained variation in % (Independent)	Explained variation in % (Dependent)	Cross-validation value
1	78,16	74,04	0,28
2	10,48	5,98	0,71
3	3,63	0,85	0,85

Component 3 has a very low explained variation, and it is possible that the variation explained by this component can be due to noise. The cross-validation value is below 1,0, indicating that the component does contribute to explaining the response. A further investigation of R^2 and R^2_a revealed that there is no reason to believe this third component is overfitting the model.

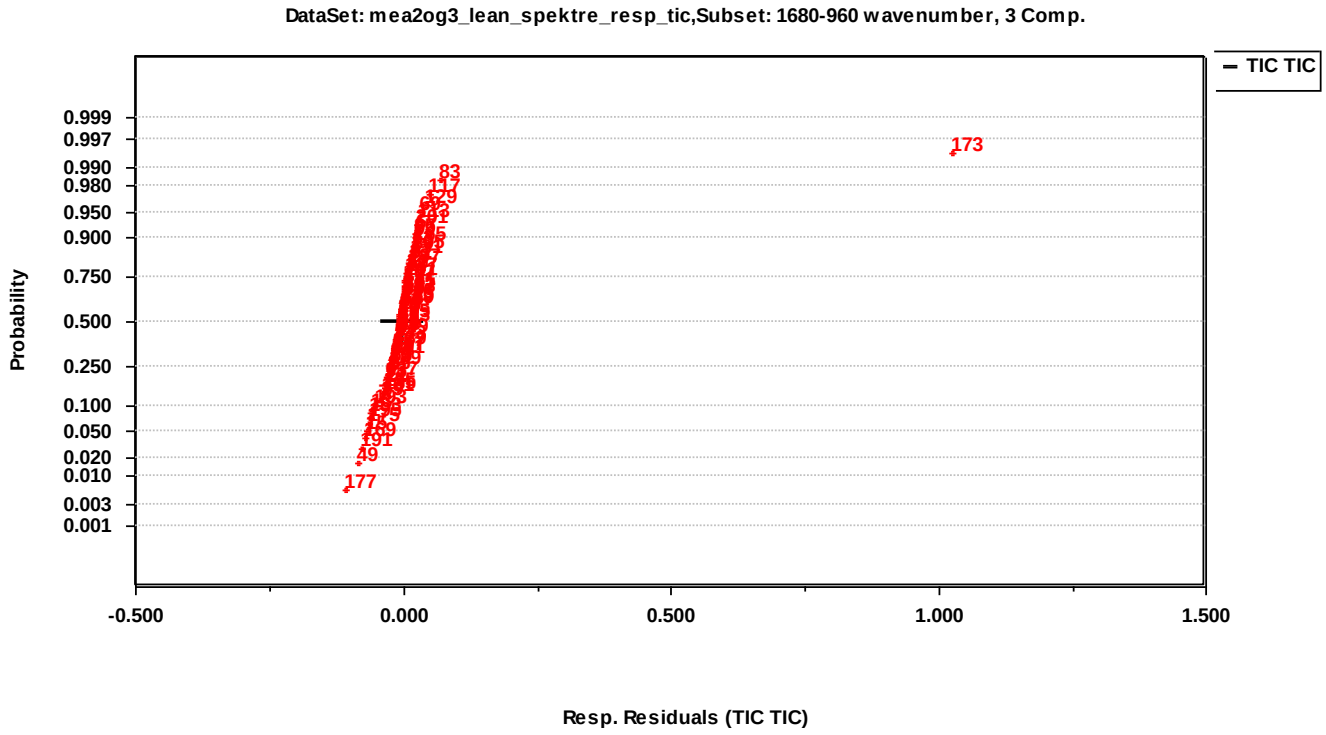


Figure 4-8: Response residuals for lean TIC model

From the response residuals in Figure 4-8 sample 173 is a clear outlier and is removed from the training set. The removal of sample 173 reveals several other weak outliers. These weak outliers were not removed and was not found to influence the models predictive abilities.

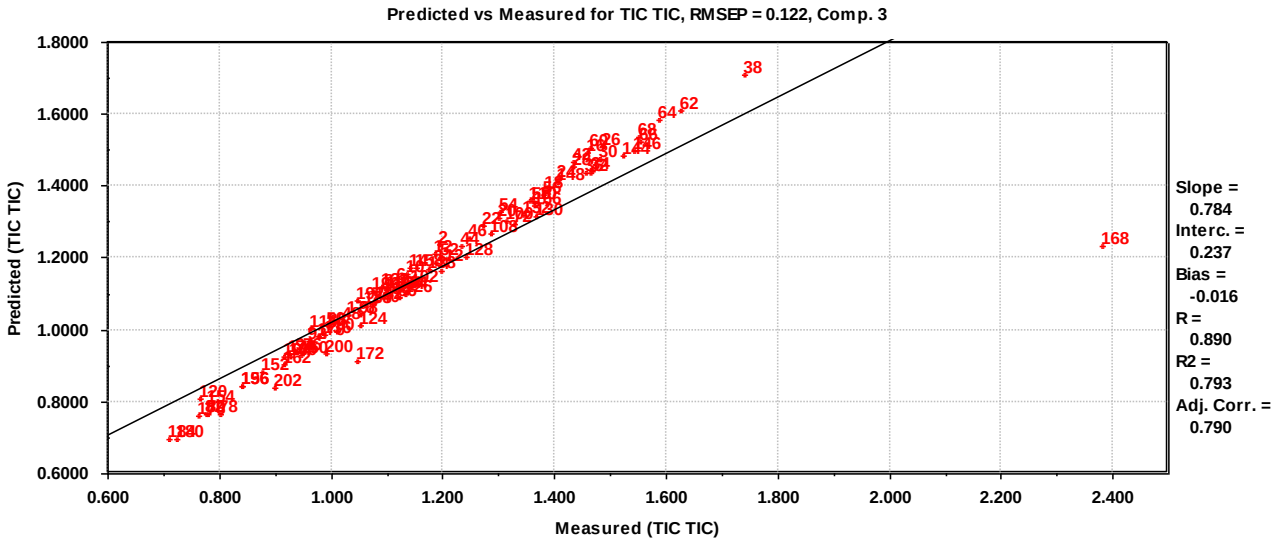


Figure 4-9: Predicted versus measured for lean TIC model

As can be seen in the predicted versus measured plot in Figure 4-9, sample 168 has a high deviation from the model and influences it significantly. Sample 168 is an extreme value and has a much higher value than any of the other samples. A new PCA is performed to evaluate sample 168. The sample is included in the training set and is not found to be an outlier. This may represent an unreliable data point due to erroneous error. The sample is removed and a new PLS is performed. The predicted versus measured plot without sample 168 is presented in Figure 4-10.

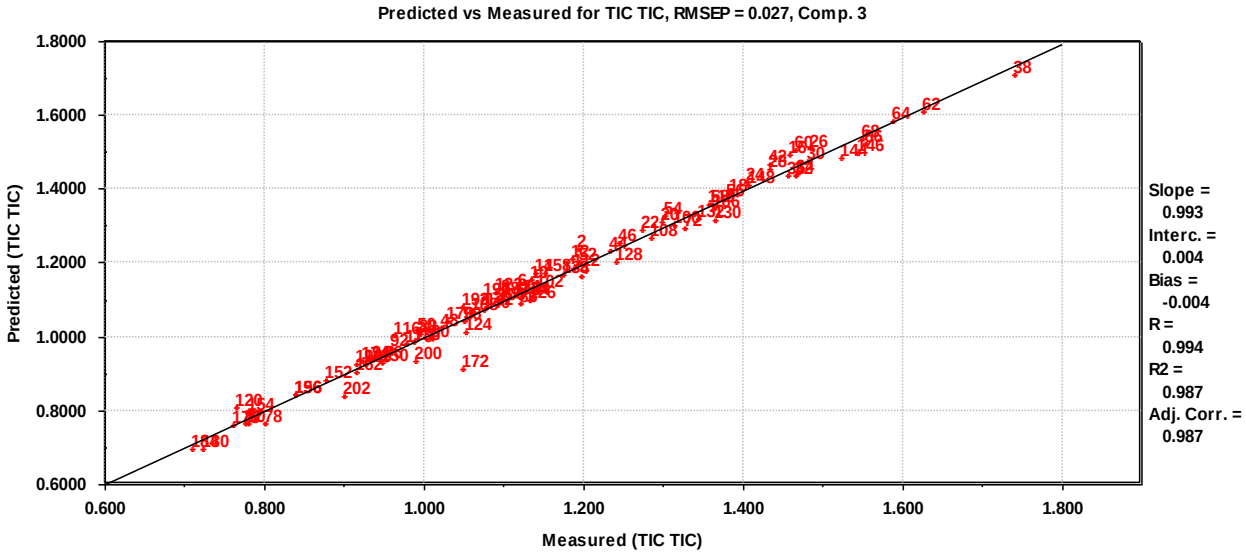


Figure 4-10: Predicted versus measured for lean TIC model afer removal of sample 168

The removal of sample 168 has significantly improved the model. The RMSEP value has dropped from RMSEP = 0,122 to RMSEP = 0,027. The average value of TIC is 1,153, thus the RMSEP value corresponds to 2,3 % of the average TIC value. The low RMSEP value compared to the typical value of TIC indicates that this model is reliable.

Table 3: Lean TIC model performance values

R	0,994
R^2	0,987
R^2_a	0,987
RMSEP	0,027

As can be seen from R^2 and R^2_a, there is no difference in value. Since the values are equal, there is no reason to believe that three components are overfitting the model and modeling noise in the data.

4.1.2 Total Alkalinity

Total alkalinity (TOT ALK) is a measure of alkaline substances dissolved in a solution, such as carbonates [32]. Alkaline substances can neutralize acids, thus, by titration of the solution by a strong acid such as hydrochloric acid the total alkalinity can be determined. Weak bases such as carbamate is produced is the process of CO_2 capture by amines. Thus, by calculating the TOT ALK of the solution and knowing the concentration of amines introduced to the system, the concentration of amines can be determined. TOT ALK is therefore a measure of the concentration of amines present is the solution, where the unit of TOT ALK is moles/kg.

When modeling TOT ALK, several methods of pretreatment have been tested such as Savitzky-Golay and EMSC. Both methods gave good values for the RMSEP. 2^{nd}-degree EMSC had a lower RMSEP value and a higher explained variation for both independent and dependent. The EMSC model is presented and the two models are compared at the end. The EMSC pretreated spectrum is presented in Figure 4-11 and the selected wavenumber region is presented in Figure 4-12.

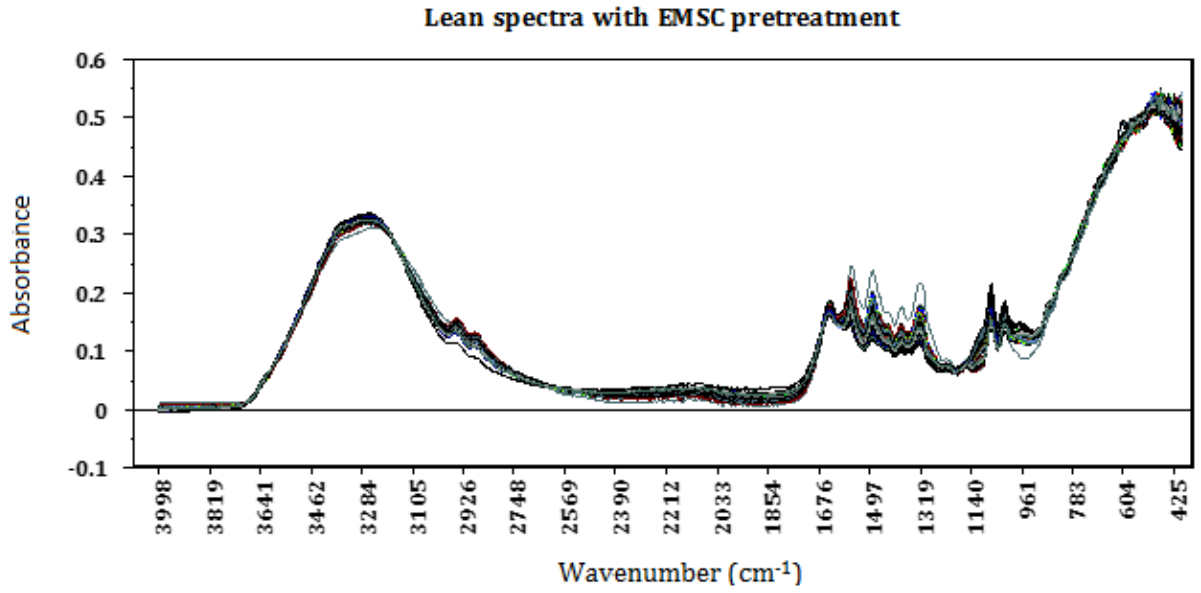


Figure 4-11: Spectra of mea 2 and 3 with EMSC 2nd degree pretreatment

Variable selection has been done to eliminate noise in the data so that the model would not be influenced by the noise. A region of wavenumbers from 3640-2750 and 1680-900 was chosen manually. Further variable selection has been applied to these regions after a complete model had been built to improve the models predictive power. The methods applied are SR and VIP. VIP and SR both decreased the models predictive power. Thus, only the selected region of wavenumbers was used to build the model.

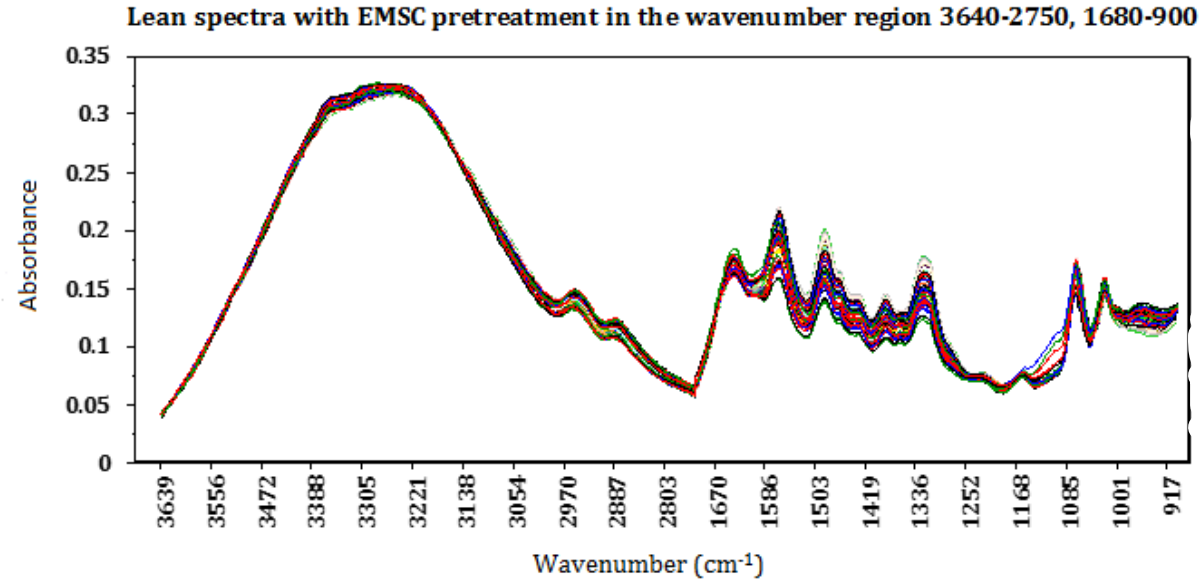


Figure 4-12: Spectra of mea 2 and 3 with 2nd-degree EMSC pretreatment in the wavenumber region 3640-2750 and 1680-900

Organic bonds can be found in the wavenumber region 3000 – 2800, therefore only noise in the spectral data has been removed.

An additional model was built for the wavenumber region 1680-960 but did not perform as well as the model with the wavenumber region 3640-2750 and 1680-900.

The outlier detection procedure is the same as explained in the theory and in the chapter on TIC modeling (4.1.1). A complete list of outliers detected by score-plot, normal-plot of scores of the PCs and RSD versus leverage can be found in Appendix D-1.

Table 4: Outliers in lean TOT ALK

Outliers:
27, 34, 52, 53, 84, 97, 99, 100, 103, 104, 105, 109, 129, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 154, 171, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182

36 outliers are removed from the original 186 samples. Of the remaining 150 samples, 76 samples are used in the training set to build the model and 74 samples are used in the validation set to validate the model. The training set is created in Sirius by selecting every odd-number sample in the data set, excluding outliers. The rest of the samples excluding outliers is used in the validation set to validate the model. The number of outliers in the model corresponds to 19 % of the samples. The score-plot of component one versus component two after the largest outliers have been removed is provided in Figure 4-13. From figure 4-13 a linear trend can be observed. Samples in this trend are mainly from the mea2 dataset from 2015. This linear trend could imply that there has been a change of some sort in the carbon capture system from 2015 to 2017, causing many of the samples to be outliers. Information provided by TCM does not indicate any apparent changes in the amine scrubbing plant from 2015 to 2017, therefore the cause of this linear trend is unknown.

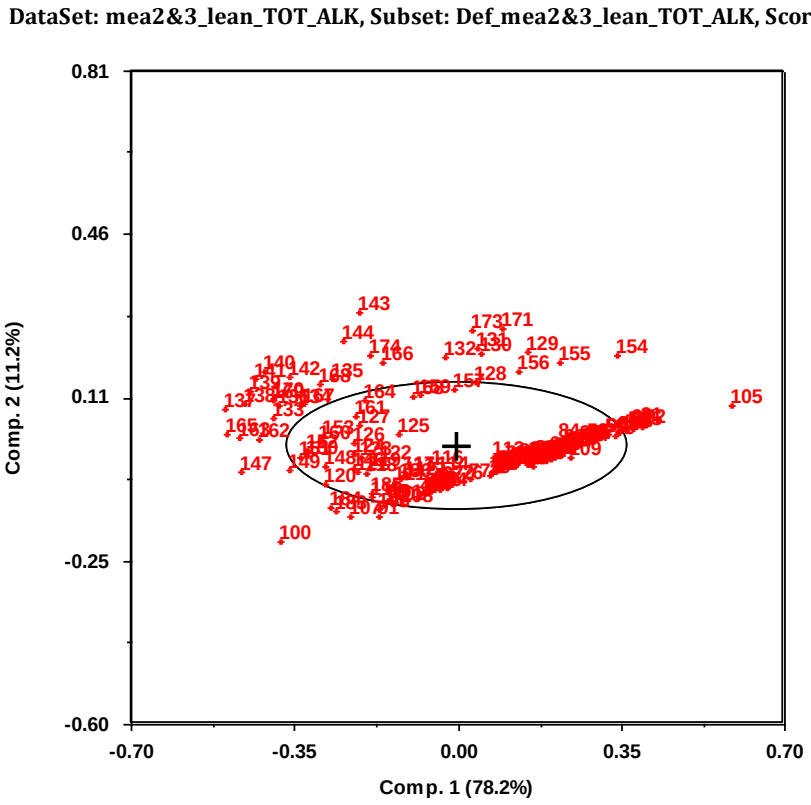


Figure 4-13: Score-plot of PC1 and PC2

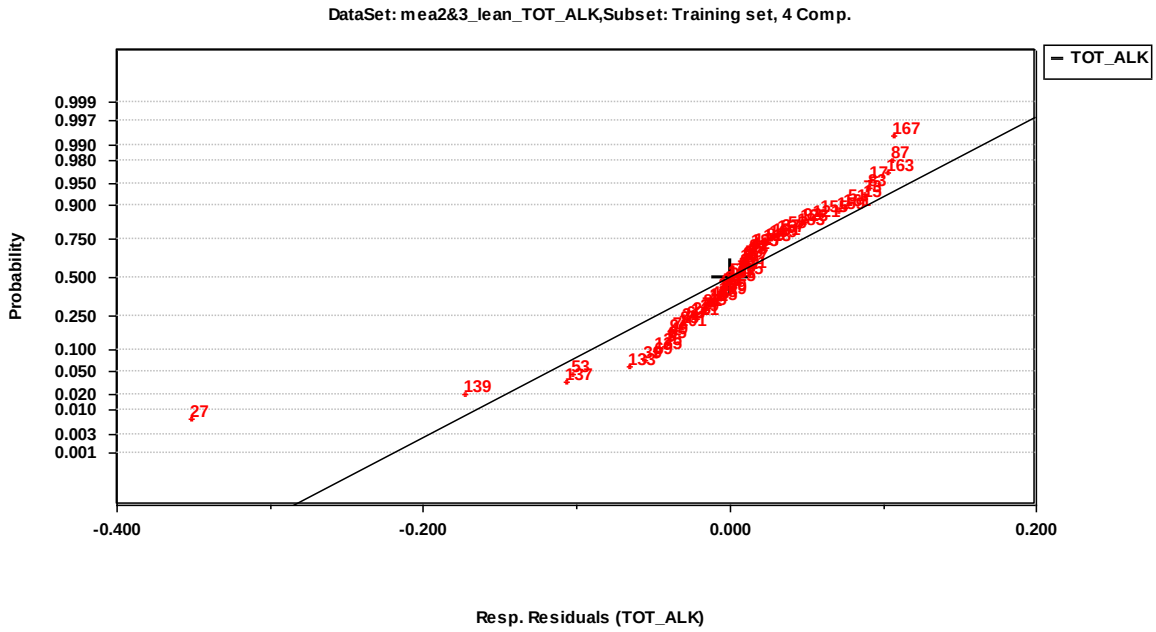


Figure 4-14: Response residuals in the CO_2-lean model for TOT ALK

From Figure 4-14 further outliers are detected such as sample 27 and 139. These are removed from the model and a new PLS model is created without these samples.

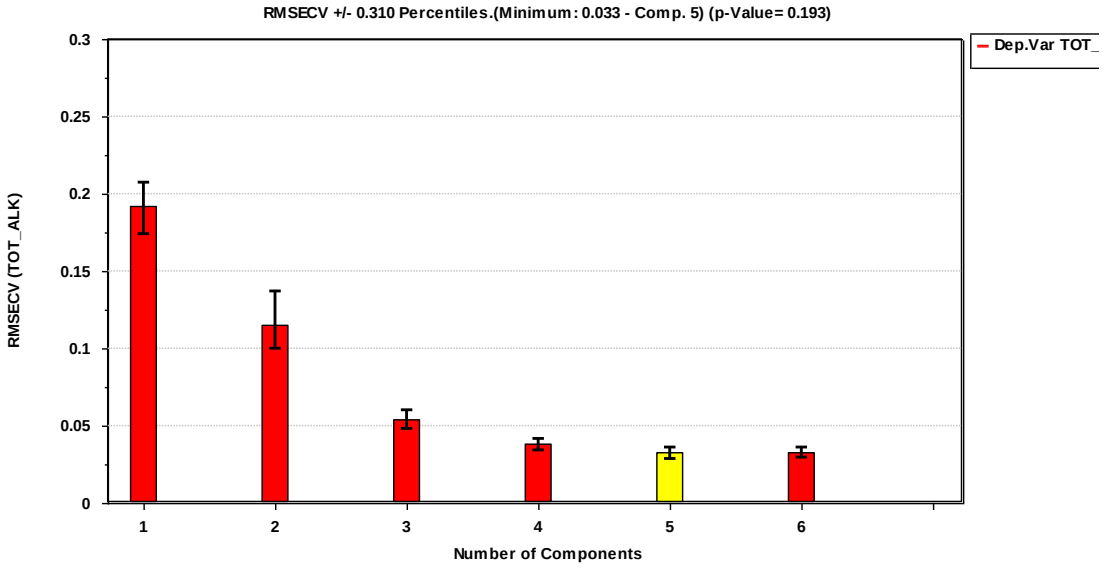


Figure 4-15: RMSECV plot showing how many components are suggested for the model by Sirius

Figure 4-15 shows that the RMSECV-plot in Sirius suggests using five components. Component information can be found in Table 5.

Table 5: Component information in the model

Component number	Explained variation (Independent), %	Explained variation (Dependent), %	Cross-validation value
1	82,36	24,69	0,88
2	6,94	48,85	0,60
3	4,67	20,77	0,47
4	5,51	3,13	0,70
5	0,07	1,07	0,87
6	0,04	0,62	0,98

Component 5 and 6 suggested by Sirius explain little variation. A four-components model was therefore investigated first. The four-component model resulted in poor response residuals. Both five- and six- component models were then tested. Even though component six has very little explained variance, the six-component model had the best

predictive power and the most normally distributed response residuals. The response residuals from the six-component model can be found in Figure 4-16.

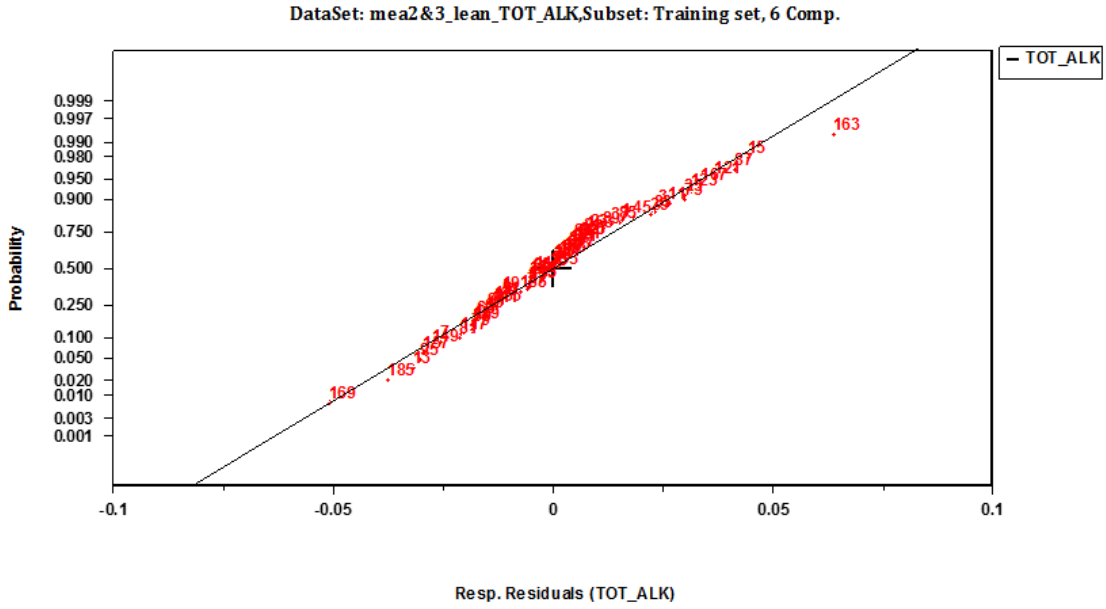


Figure 4-16: Response residuals

The response residuals are close to linear and are normally distributed as the response residuals pass through the point y = 0,5 and the center of the distribution is at x equal 0. Sample 163 deviate some from the rest of the response residuals. Sample 163 is not removed as the deviation is not significantly large. However, a PCA was performed on the training set to investigate the sample to check if the sample has the characteristics of an outlier in the score-plot and the RSD versus leverage plot. Sample 163 was not found to be an outlier in the subsequent PCA analysis and are therefore included in the model. The predicted versus measured plot is presented in Figure 4-17.

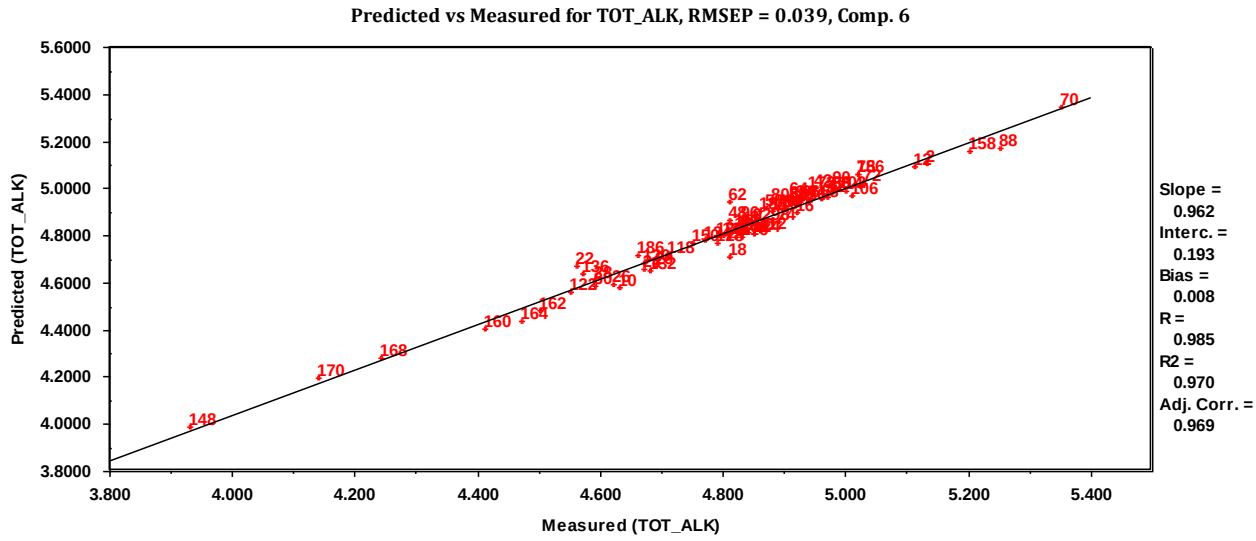


Figure 4-17: Predicted versus measured

A 21-point Savitzky-Golay 2nd-degree derivative resulted in an RMSEP value of 0,061, which is higher than the EMSC but Savitzky-Golay had fewer components and fewer outliers. The two models are compared in Table 6.

Table 6: Comparison of the two best models obtained for lean TOT ALK

	EMSC	Savitzky-Golay
Number of outliers	36	25
Number of components in the model	6	4
Explained variance Independent: Dependent:	 99,54 % 98,51 %	 95,55 % 95,06 %
RMSEP value	0,039	0,061
R	0,985	0,973
R^2	0,970	0,947
R^2_a	0,969	0,946

Both models performed well, exhibiting insignificant variation. The EMSC has a higher explained variance and lower RMSEP value, while the Savitzky-Golay has fewer components in the model and fewer outliers. As can be seen from R^2 and R^2_a, there is little difference in value between the models. There is therefore no reason to believe that either of these models is overfitted.

The average value of TOT ALK in the model built on lean samples is 4,76 and the corresponding RMSEP value of the EMSC model is 0,039. The RMSEP value corresponds to 0,82 % of the average value for TOT ALK. The RMSEP value is much smaller than the average value indicating that this model is reliable.

4.1.3 Density

The absorption of CO_2 in the aqueous MEA solution will increase the density of the solution, therefore the response variable density is used to determine the CO_2-loading (moles CO_2/mole amine) in the aqueous MEA solution, where the unit of density is given as kg/m^3. In the density model, a 2nd-degree EMSC resulted in the best model. The pretreated spectrum is presented in Figure 4-18 and the selected wavenumber region is presented in Figure 4-19.

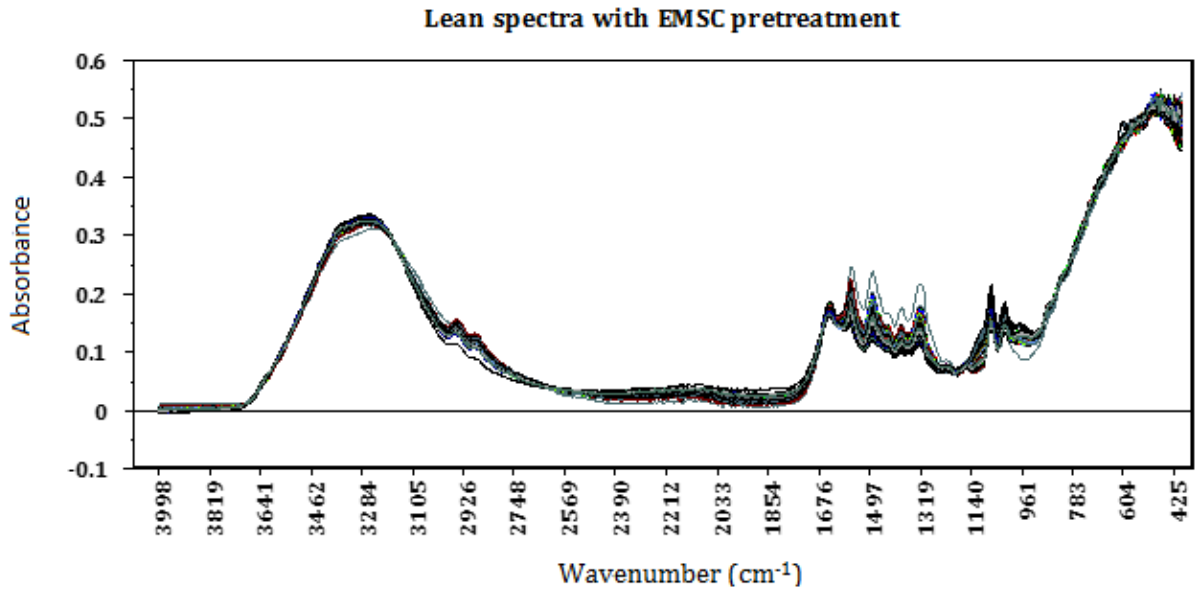


Figure 4-18: Spectra of mea 2 and 3 with EMSC 2nd degree pretreatment

Manual selection of wavenumbers has been made in the wavenumber region 3640-2750 and 1680-900 cm^{-1}.

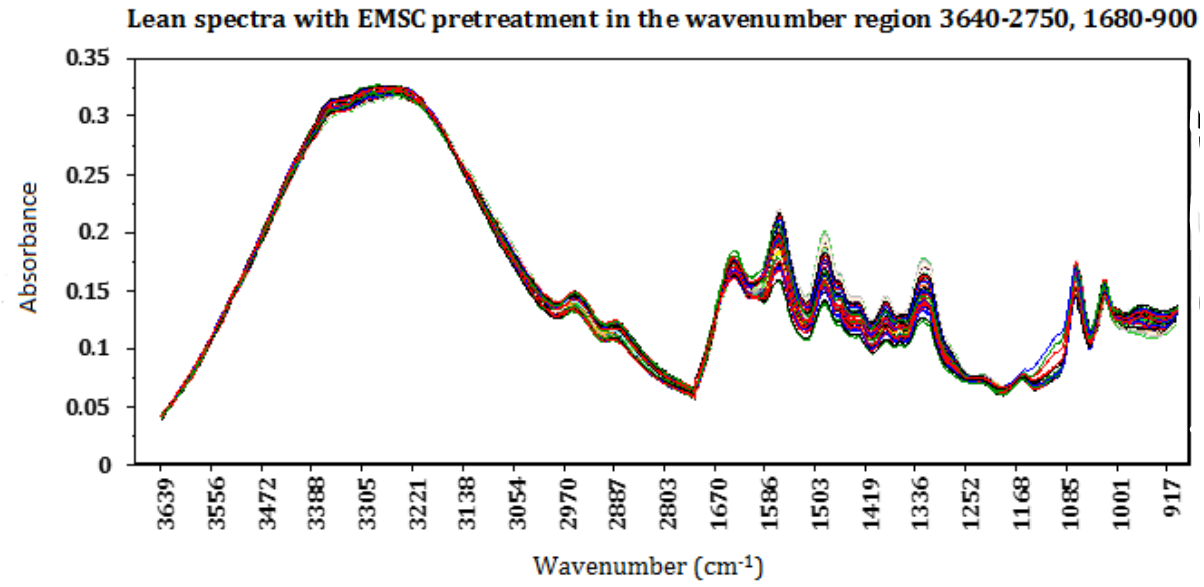


Figure 4-19: Spectra of mea 2 and 3 with 2nd-degree EMSC pretreatment in the wavenumber region 3640-2750 and 1680-900

The outlier detection procedure is the same as explained in the theory and in the chapter on TIC modeling (4.1.1). A complete list of outliers detected by score-plot, normal-plot of scores of the PCs and RSD versus leverage can be found in Appendix D-2.

Outliers found in the dataset are listed in Table 7.

Table 7: Outliers for lean density model.

Outliers
40, 92, 100, 113, 115, 117, 125, 126, 127, 128, 129, 130, 131, 134, 135, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149,

27 samples of the original 154 samples are removed. Of the remaining 127 samples, 64 samples are used in the training set to build the model and 63 samples are used to construct the validation model. The training set is created in Sirius by selecting every odd-number sample in the data set, excluding outliers. The rest of the samples excluding

outliers is used in the validation set to validate the model. A model of the training set was built and sample 113 and 115 as listed in table 7 were found to be outliers from the response residual plot, see figure 4-20

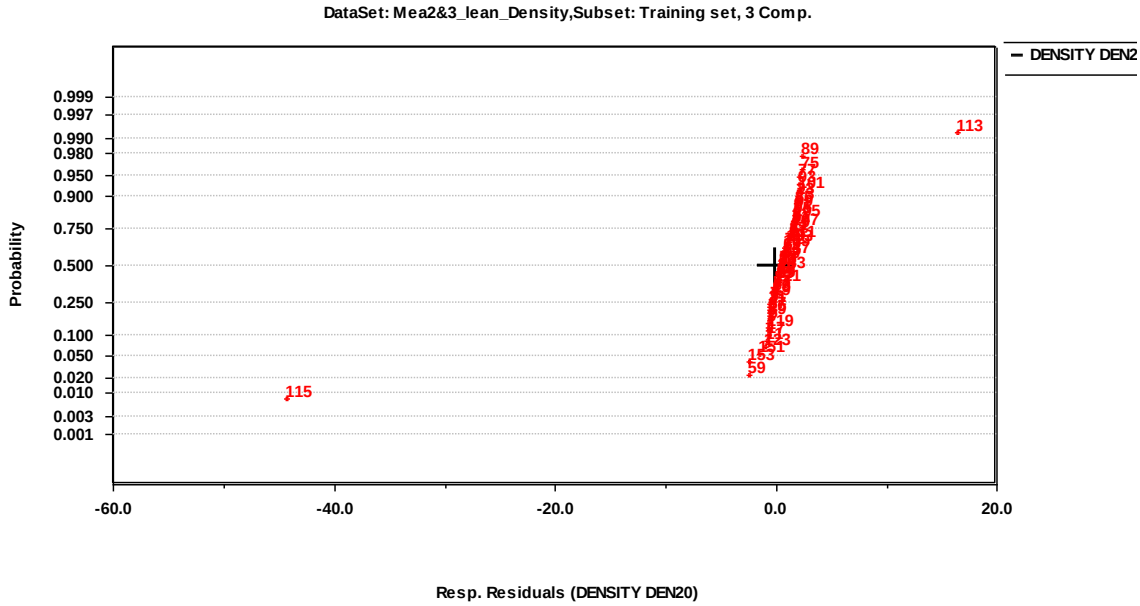


Figure 4-20: Response residuals of lean density model. Showing sample 113 and 115 as outliers.

These samples were removed and a new PLS model was built. Figure 4-21 shows how many components Sirius suggest and Table 8 shows the component information.

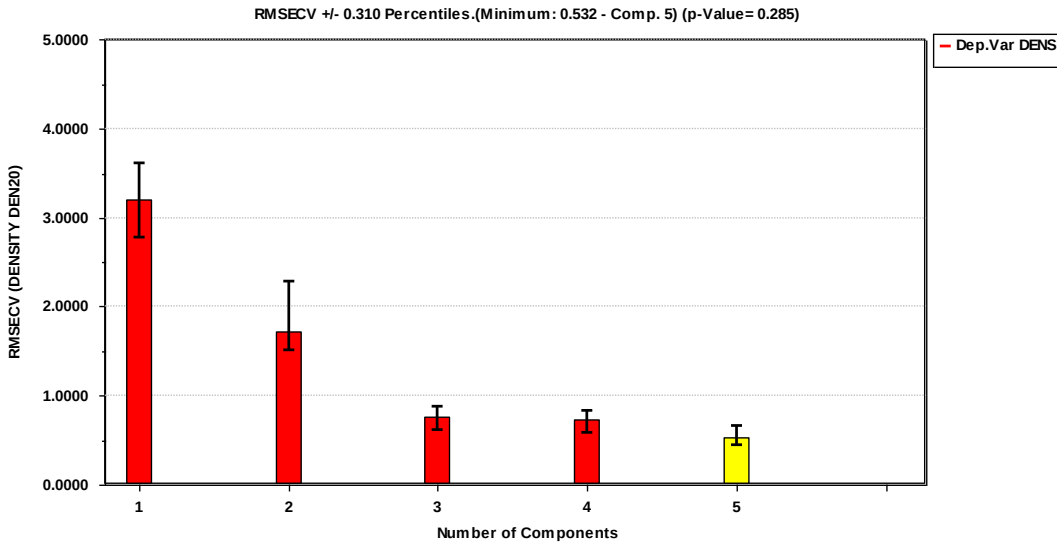


Figure 4-21: RMSECV plot. Showing how many components are suggested by Sirius.

The RMSECV plot suggests using five components even though the difference in RMSECV value for component three and four is very small. The insignificant difference in RMSECV value for component three and four suggest that a three-component model is adequate.

Table 8: Component information for lean density

Component number	Explained variation in % (Independent)	Explained variation in % (Dependent)	Cross-validation value
1	83,11	90,74	0,29
2	9,39	5,41	0,54
3	4,49	3,47	0,44
4	2,12	0,03	0,95
5	0,20	0,17	0,70

Component 4 and 5 have a very low amount of explained variance and was not included in the model at the beginning. The three-component model generated poor response residuals and a higher RMSEP value for the predicted values in the validation set than the four-component model. With five components the response residuals improved significantly, but the RMSEP value was the highest of the three models. The four-component model had a better RMSEP value, but not equally good normally distributed response residuals as the five-component model. The RMSEP values for each of the model can be found in Table 9.

Table 9: Comparison of the best models for lean density

Number of components in the model	RMSEP value
3	0,470
4	0,463
5	0,490

From Table 9 the RMSEP value differentiates very little compared to the high values of the density measurement. A five-component model is concluded to be reliable due to more normally distributed response residuals and higher explained variation by the model.

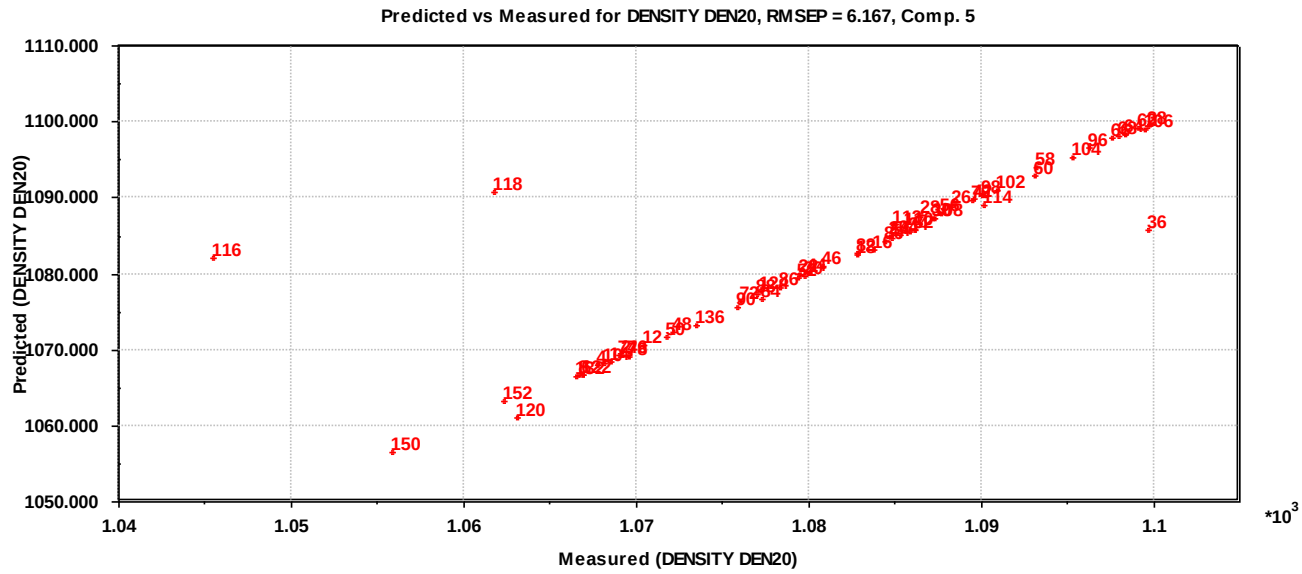


Figure 4-22: Predicted versus measure for lean density

The RMSEP value was slightly inflated for this model due to the three measurements that do not fit the model. These three samples were investigated further in a PCA by including them in the model building set. Sample 118 was found to be an outlier while sample 116 and 36 did not show any sign of being outliers.

PCA only investigate the how the objects and variables relate to one another respectively. Thus, any deviation in the response will go unnoticed by the PCA. Sample 36 and 116 are extreme values, meaning they either have a very high value or a very low value.

Sample 116 has a lower value than any other samples in this plot and sample 36 has one of the highest. This may represent unreliable data points due to erroneous error. Sample 36, 116 and 118 were removed and a new plot of the predicted versus measured plot was created. The plot is presented in Figure 4-23.

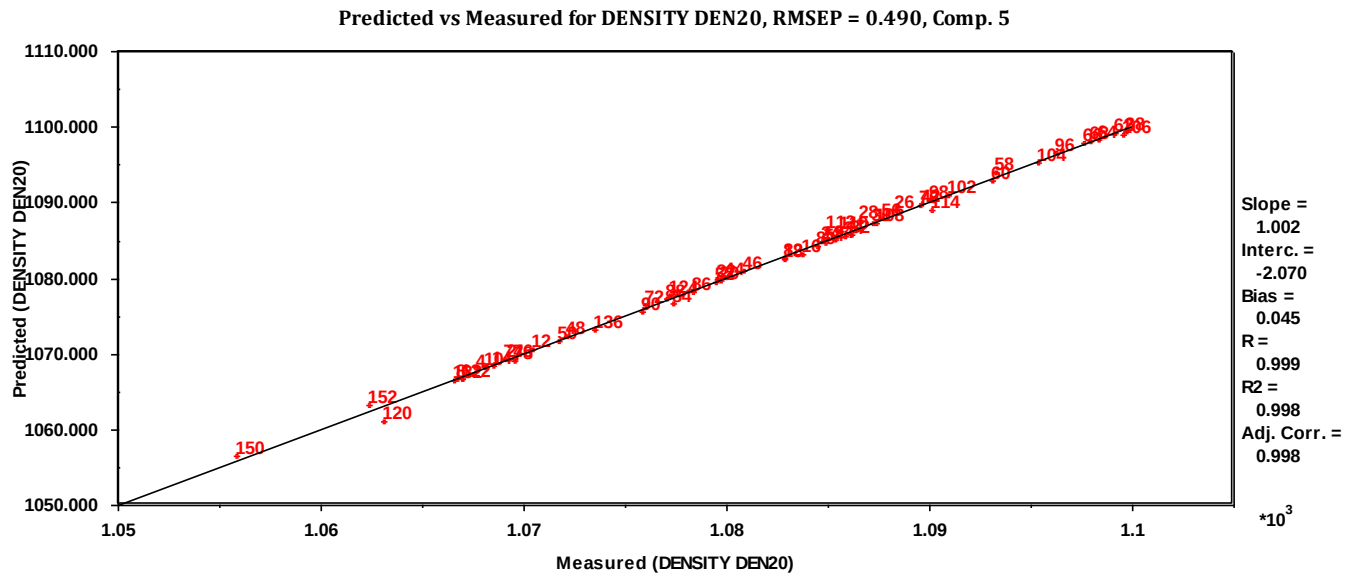


Figure 4-23: Predicted versus measured without sample 36, 116 and 118

The RMSEP value has significantly improved with the removal of sample 36, 116 and 118, from RMSEP = 6,167 to RMSEP = 0,490. The RMSEP values for this model are larger than for the other models. This is to be expected as the measured values are much higher. The average value of the density samples is 1082,51, thus the RMSEP value corresponds to 0,045 % of the average value. The low RMSEP value compared to the average value of density indicates that this model is reliable.

Table 10: Lean TOT ALK model performance values

R	0,999
R^2	0,998
R^2_a	0,998
RMSEP	0,490

As can be seen R^2 and R^2_a, there is no difference in value and the model is not overfitted.

Using a 2nd degree EMSC in the wavenumber region 3640-2750 and 1680-900 with multivariate methods has shown to be a acceptable method for predicting density.

4.2 Rich samples

The rich samples are obtained with an ATR-FTIR spectroscope after flue gas has interacted with the amine solution. An inline measurement is made when the CO_2-rich amine solution is on the way to the stripper. The plot of the rich samples is obtained in Sirius (11.0) by plotting the intensity against the wavenumbers.

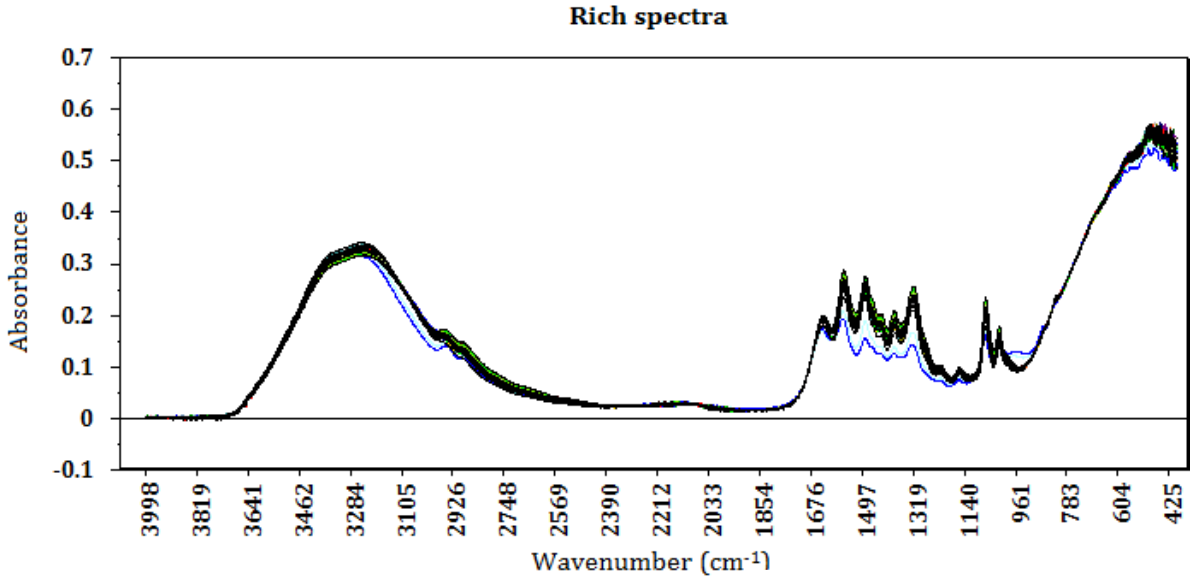


Figure 4-24: Rich spectra

As can be seen from the figure 4-24, the organic C-H bonds at 3000 – 2800 cm^{-1} is not entirely buried by the O-H bond absorption from water in the rich samples and it is possible to get information from this region.

The number of samples in the rich models are about half of the number of lean samples. This is due to very few rich measurements in the mea3 dataset from 2017. However, this has resulted in very few outliers in the rich models compared to the lean models, suggesting that there is some difference in the dataset from 2015 to 2017.

4.2.1 Total Inorganic Carbon

Building a model for TIC in the rich amine solution proved to be a challenge. Pretreatment with a 2nd degree EMSC and pretreatment with a Savitzky-Golay 3rd degree polynomial fit, 2nd degree differentiation with a window size range of 11-25 points have been tested. On each of these models, variable selection has been performed using VIP, SR and manual selection of different wavenumber regions.

The pretreated spectrum is presented in Figure 4-25. The model that resulted in the best RMSEP value was a 2nd-degree EMSC in the wavenumber region 1680-960, the spectra is found in figure 4-26. Even though the region 3000-2800 cm^{-1} does contain information, the molecular bonds found there are organic and are not included in the model for total inorganic carbon in the rich samples. The region has been included in an effort to improve the model but generated less reliable predictions.

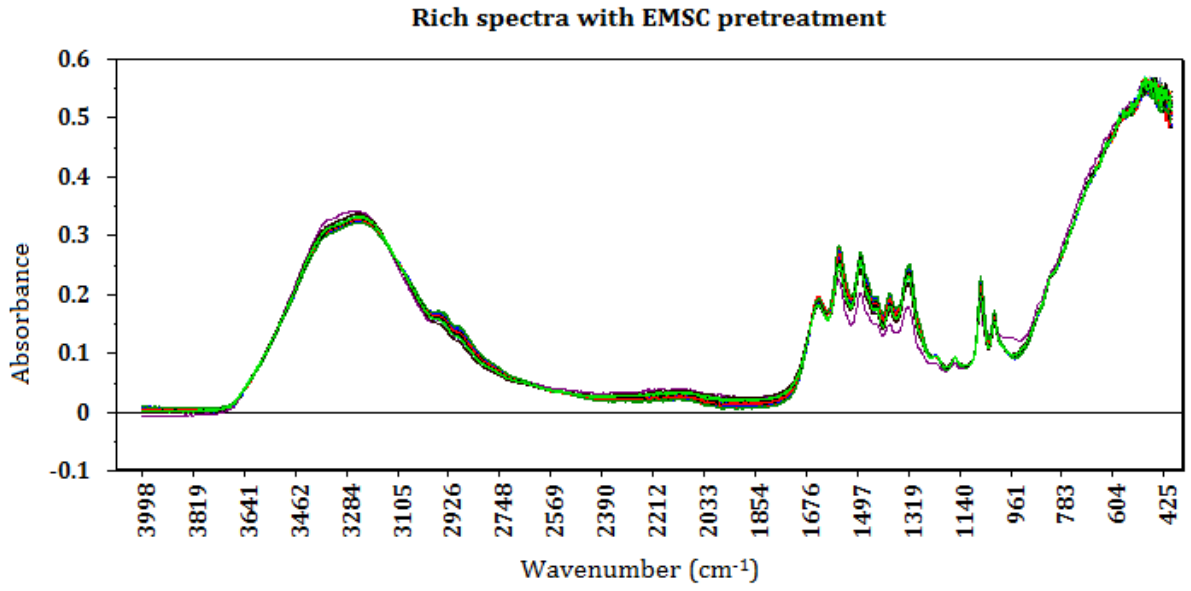


Figure 4-25: Rich spectra with a 2nd degree EMSC pretreatment

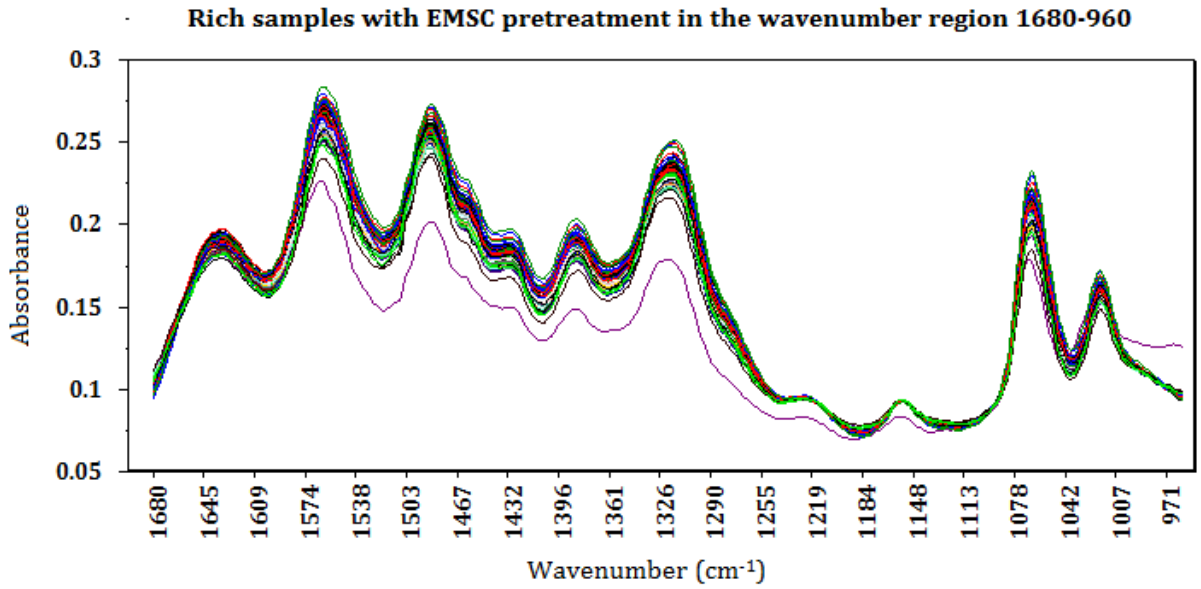


Figure 4-26: Manual variable selection of the wavenumber region 1680-960.

The outlier detection procedure is the same as explained in the theory and in the chapter on TIC modeling (4.1.1). A complete list of outliers detected by score-plot, normal-plot of scores for the PC's and RSD versus leverage can be found in Appendix D-3. A complete list of outliers is found in Table 11.

Table 11: Outliers in the rich TIC model

Outliers
23, 30, 39, 68, 76, 77, 78, 79, 87, 88, 90, 91

12 outliers of the original 90 samples have been removed. Of the remaining 78 samples, 39 samples are used in the training set to build the model and 39 samples are used in the validation set to validate the model. The training set is created in Sirius by selecting every odd-number sample in the data set, excluding outliers. The rest of the samples excluding outliers is used in the validation set to validate the model.

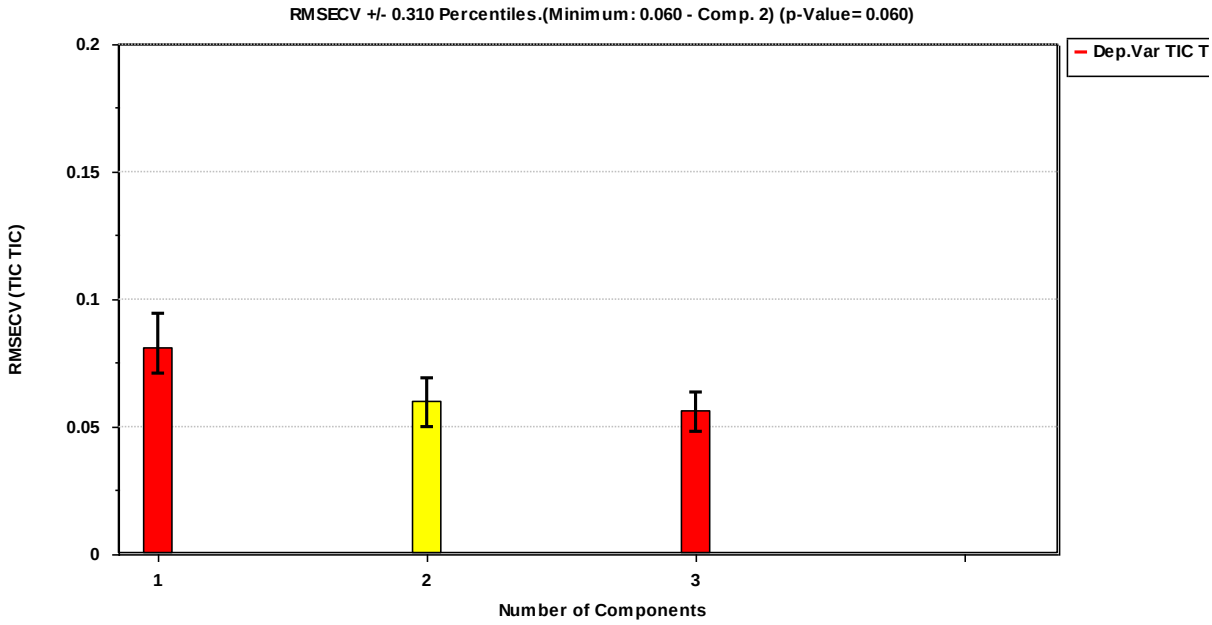


Figure 4-27: RMSECV plot for the rich TIC model

As seen in figure 4-27 the RMSECV-plot suggest using two components in the model. The component information is provided in Table 12.

Table 12: Component information for the rich TIC model

Component number	Explained variance (Independent), %	Explained variance (Dependent), %	Cross-validation value
1	68,79	31,87	0,81
2	27,23	37,63	0,74
3	2,32	5,00	0,94

Two components are suggested by the RMSECV-plot created by Sirius. However, a three-component model had a more normally distributed response residual and gave a lower RMSEP value for the predicted samples in the validation set.

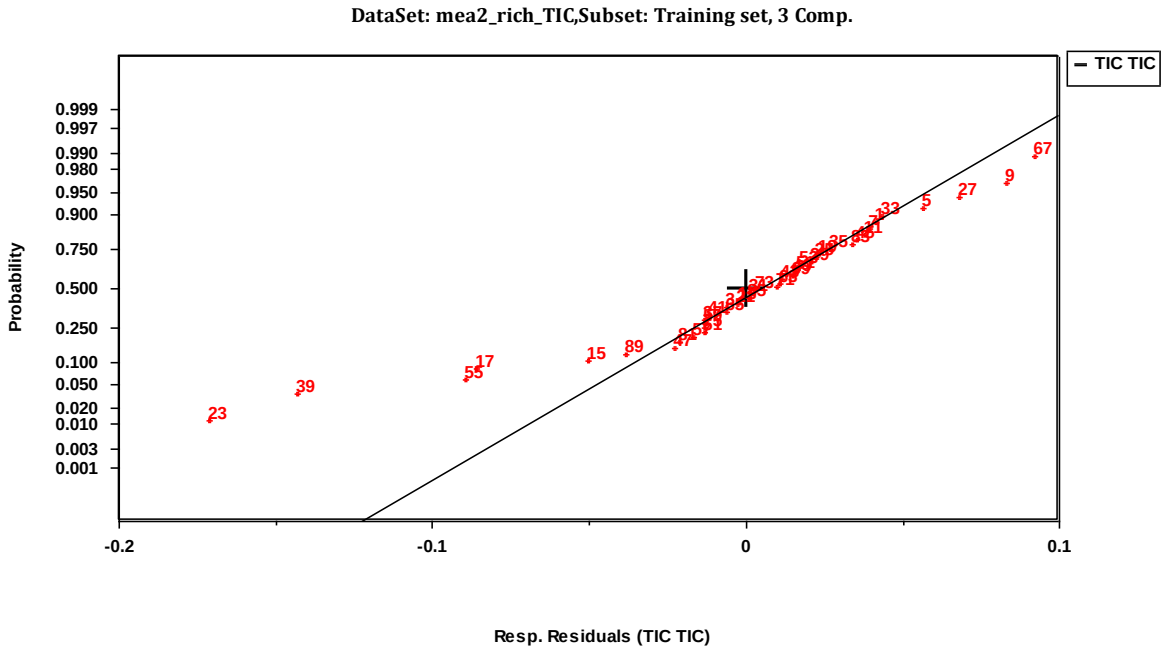


Figure 4-28: Response residuals for rich TIC model

In figure 4-28 a line has been fitted to the response residual plot where the most significant effects of the samples, the outliers, have been removed. Sample 23 and 39 are clear outliers and were removed from the training set prior to the prediction of the validation set. After removal of sample 23 and 39, sample 55 and 17 fit the line much better. There is still some deviation from these two samples, but they are not removed as the deviation is minimal. The new plot of the response residuals without sample 23 and 39 is presented in Figure 4-29.

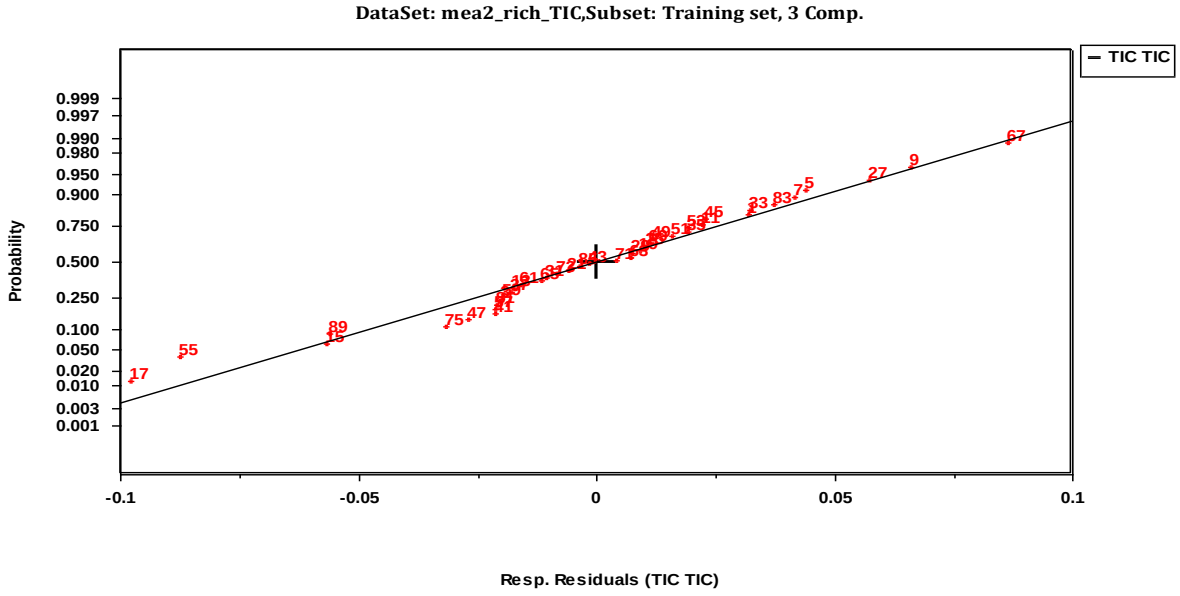


Figure 4-29: Response residuals without sample 23 and 39

The predicted versus measured values for the validation set is presented in Figure 4-30

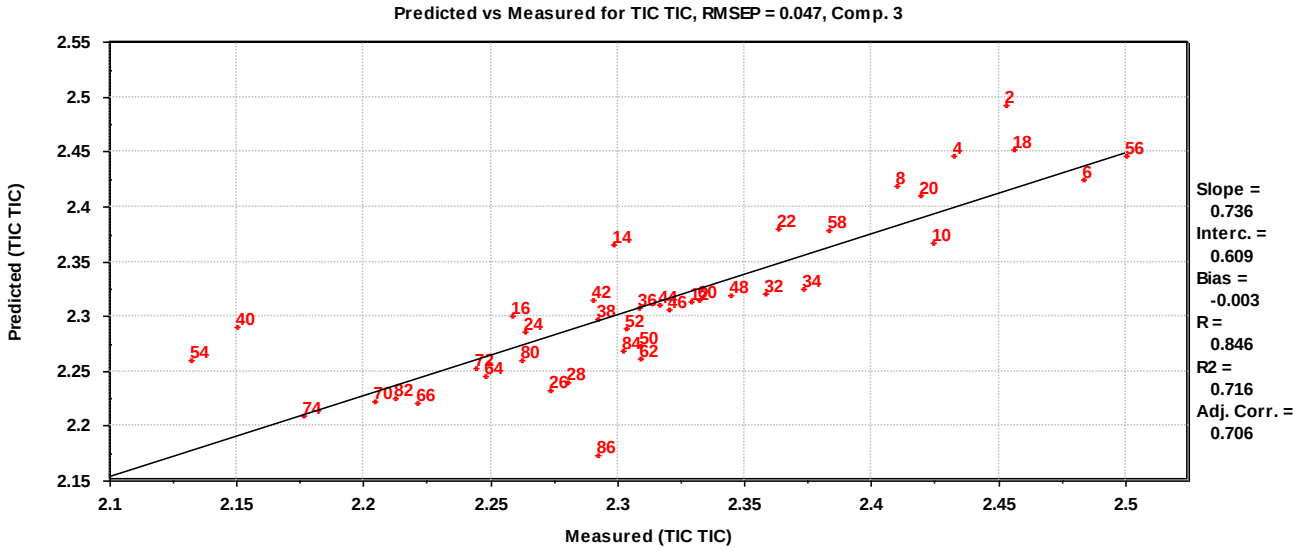


Figure 4-30: Predicted versus measured values for the validation set

The predicted versus measured plot for the rich TIC model is significantly more scattered than the other models. However, the scale of the plot is very small. The axis range in both x and y-direction is only 0,4. The average value of TIC for the rich model is 2,31.

Thus the RMSEP value is 2,03 % of the average TIC value. The RMSEP value is appreciable low, indicating that this model is reliable.

Table 13: Rich TIC model information

R	R^2	R^2_a	RMSEP
0,846	0,716	0,706	0,047

4.2.2 Total Alkalinity

The method for building a model for total alkalinity that gave the best results were a 2nd-degree EMSC pretreatment in the wavenumber region 3640-2750 and 1680-960. In

Figure 4-31 the pretreated spectrum is presented and in Figure 4-32 the pretreated selected wavenumber region is presented.

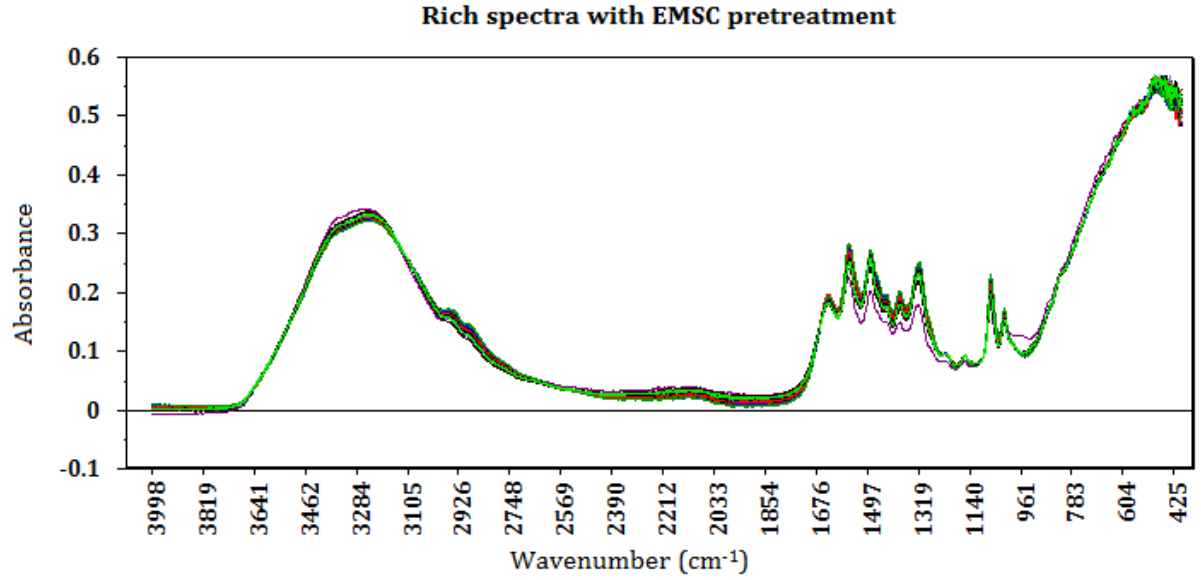


Figure 4-31: Rich spectra with 2nd degree EMSC pretreatment

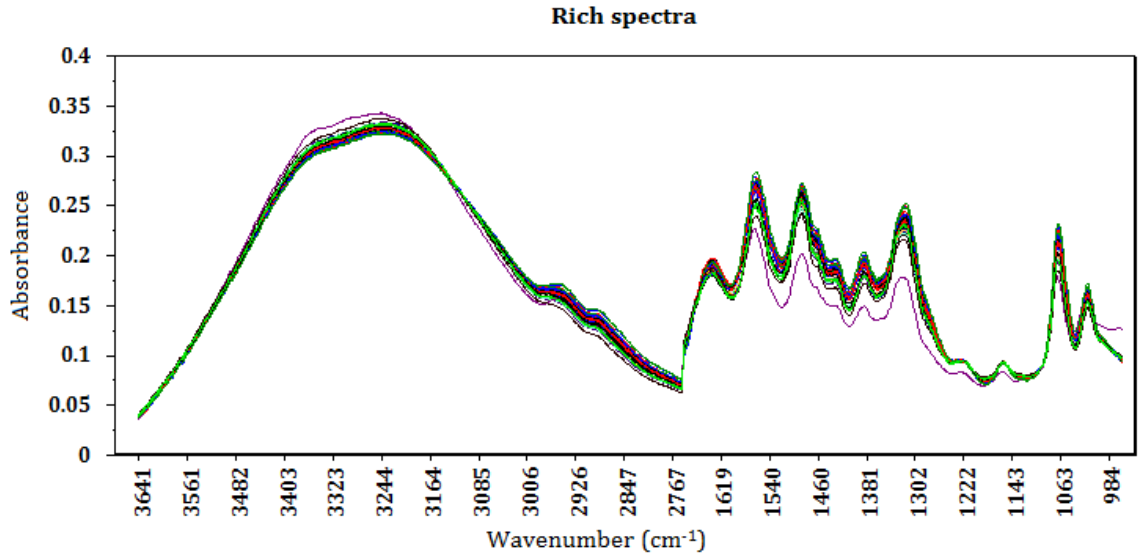


Figure 4-32: Wavenumber region 3640-2750 and 1680 - 960

From the spectra, it is obvious that one lean sample has not been removed from the rich TOT ALK dataset. Outliers were detected by a score-plot, RSD versus leverage plot and by a normal-plot of the scores of the PCs. The plots are presented in appendix D-4 and a complete list of outliers is presented in Table 14.

Table 14: Outliers in TOT ALK

Outliers
24, 75, 77.

Three outliers of the original 92 samples have been removed. Of the remaining 89 samples, 45 samples are used in the training set to build the model and 44 samples are used in the validation set to validate the model. The training set is created in Sirius by selecting every odd-number sample in the data set, excluding outliers. The rest of the samples excluding outliers is used in the validation set to validate the model.

Sample 24 is a clear outlier and was most likely a lean sample that had not been removed. After the removal of sample 24, a new PCA was performed to find outliers.
Only a few very weak outliers were detected but not removed. After building a model sample 75 and 77 turned out to be outliers in the model and was removed.

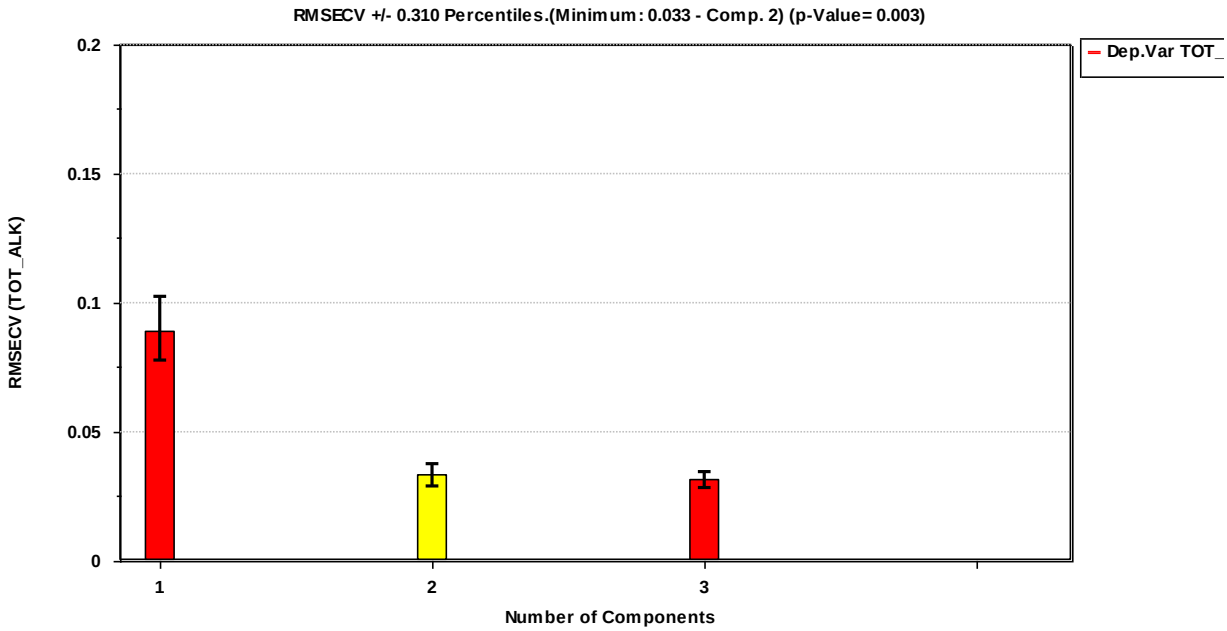


Figure 4-33: RMSECV plot

Two components are suggested by Sirius in the RMSECV plot in Figure 4-33, marked by the yellow bar. Both a two- and three-component model were tested. The three-components model performed better at predicting samples in the validation set and had

a more normally distributed response residual (see Figure 4-34). The three-component model was chosen, and the component information can be found in Table 15.

Table 15: Component information for rich TOT ALK

Component number	Explained variance (Independent), %	Explained variance (Dependent), %	Cross-validation value
1	87,51	52,04	0,72
2	8,92	42,39	0,38
3	1,89	1,40	0,94

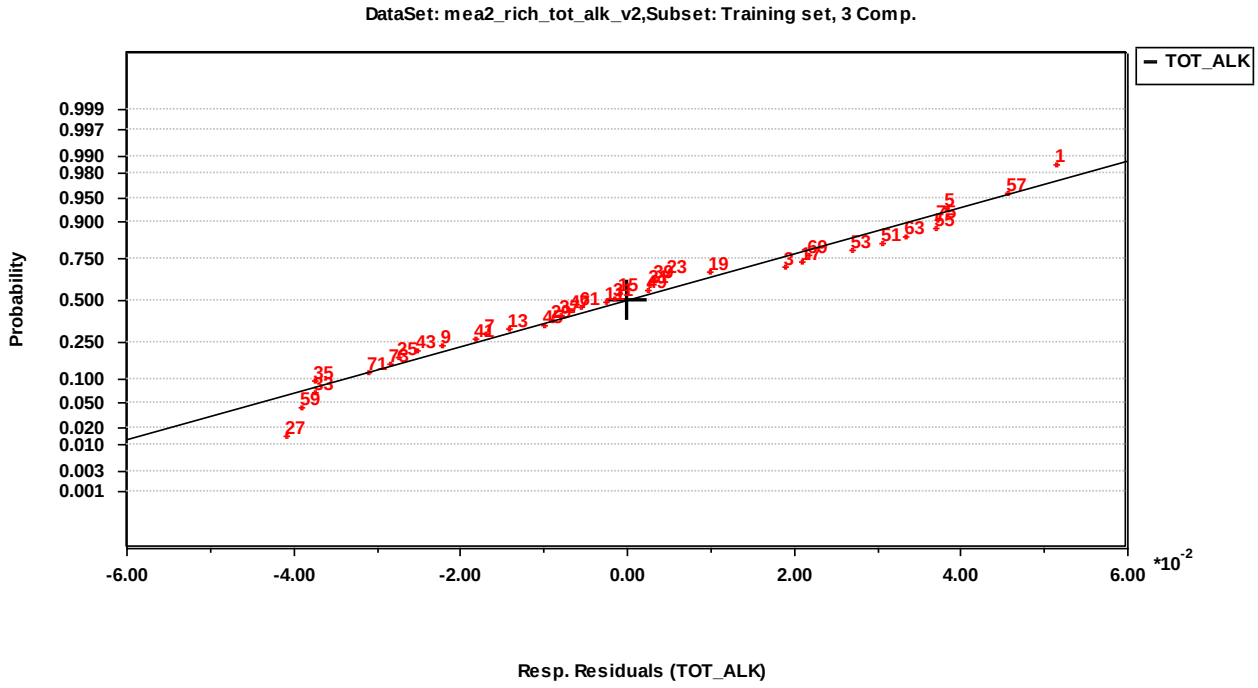


Figure 4-34: Response residuals for rich TOT ALK model

From the response residuals in Figure 4-34 sample 27 is a weak outlier in the model. Sample 27 is not removed from the model and were not found to influence the models

predictive ability. The predicted versus measured values for the TOT ALK samples in the rich model are presented in Figure 4-35.

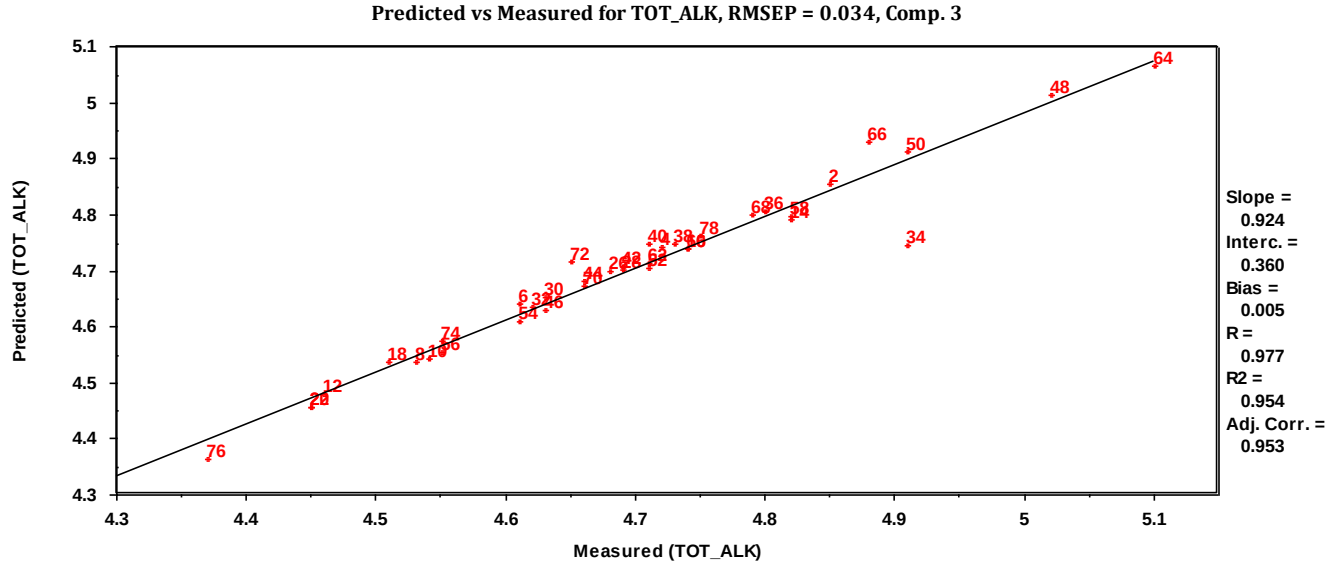


Figure 4-35: Predicted versus measured for rich TOT ALK model

Table 16: Comparison of two and three components to describe the model

Number of components	**R**	**R^2**	**R^2_a**	**RMSEP**
3	0,977	0,954	0,953	0,034
2	0,954	0,909	0,907	0,038

From Table 16 there can be seen very little difference in the two models. The three-component model performs marginally better than the two-component model. From R^2 and R^2_a, there is minimal divergence between the models. The three-component model is therefore concluded to be the best as the RMSEP value is lower and the difference in coefficient of multiple determination and the adjusted coefficient of multiple determination is low.

The average value of TOT ALK is 4,70, thus the RMSEP value corresponds to 0,72 % of the average value. The RMSEP value is much lower than the typical value of TOT ALK and the model is concluded to be satisfactory.

Sample 34 has a high response residual compared to the other samples and is a possible outlier. Further investigation of the validation set by PCA is performed and sample 34 is concluded not to be an outlier, but it might be an erroneous error.

4.2.3 Density

In the rich density model, a 2nd-degree EMSC pretreatment and a variable selection of wavenumbers in the region 1680-900 gave the best result. The pretreated spectrum is presented in Figure 4-36.

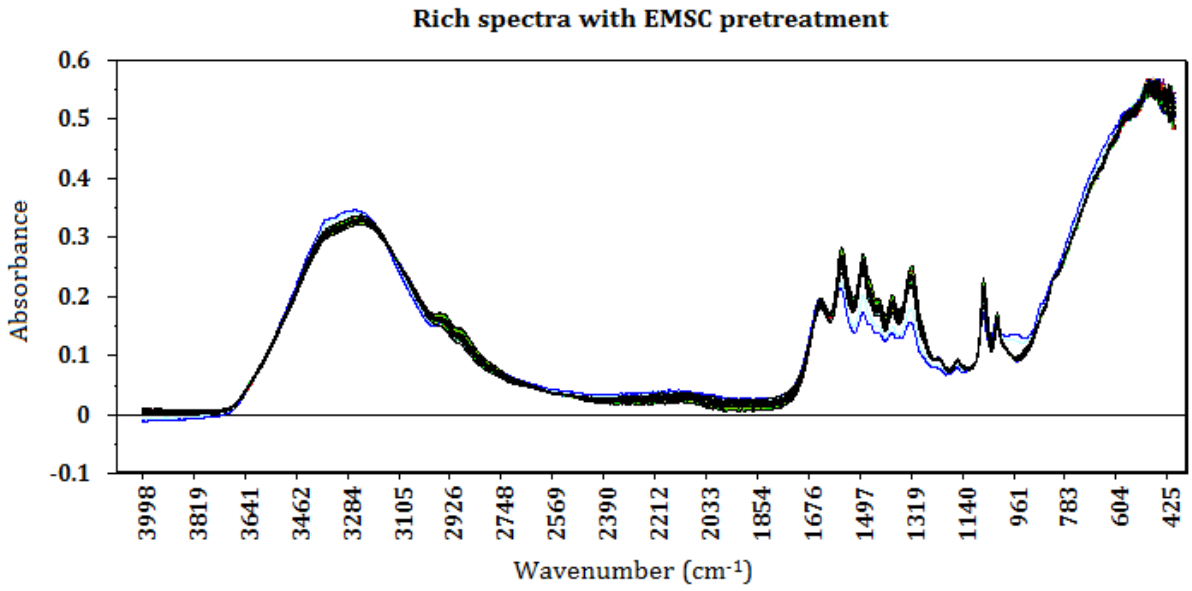


Figure 4-36: Rich spectra with 2nd-degree EMSC pretreatment

Unlike the density model built from the lean samples, the model for density built from the rich samples performed better when not including the wavenumber region 3640-2750 cm^{-1}. The selected pretreated wavenumber region is presented in Figure 4-37.

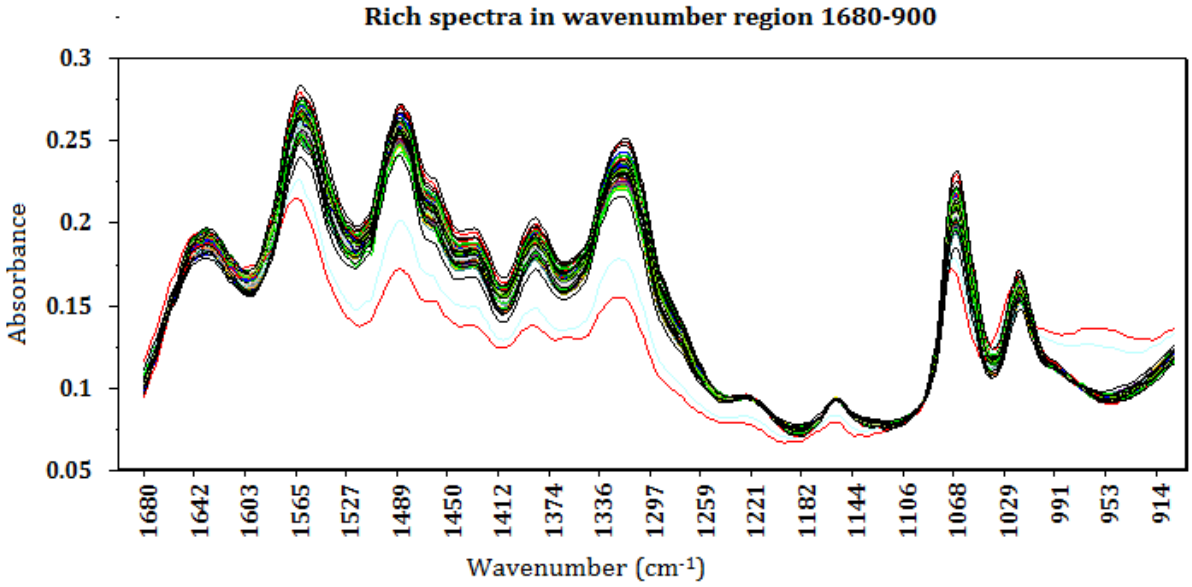


Figure 4-37: Wavenumber region 1680 - 960

From the spectra collected from all samples, two spectra differentiate significantly from the rest. A score-plot is provided to illustrate the severity of deviation.

Outlier detection is done by score-plots, RSD versus leverage and by normal-plots of the scores versus the objects. The plots can be found in Appendix D-5.

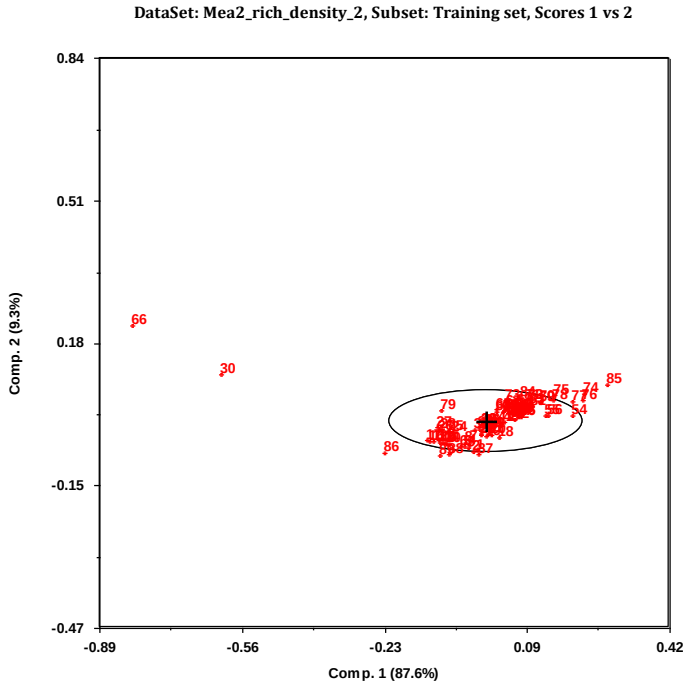


Figure 4-38: Score-plot of PC1 and PC2

As can be seen from the score-plot in Figure 4-38, the two samples 30 and 66 are obvious outliers. The reason for this large deviation can be seen in Figure 4-36. The spectra of sample 30 and 66 deviate from the rest, this is because these samples are lean samples that had not been removed along with the rest. These samples are removed first and then a new evaluation of the score-plot was made to identify the other outliers. A complete list of outliers can be found in table 17.

Table 17: Outliers in rich density

Outliers:
30, 66, 68, 73, 74, 79, 82, 85, 86, 87

Ten outliers of the original 90 samples have been removed. Of the remaining 80 samples, 41 samples are used in the training set to build the model and 39 samples are used in the validation set to validate the model. The training set is created in Sirius by selecting every odd-number sample in the data set, excluding outliers. The rest of the samples excluding outliers is used in the validation set to validate the model.

A PLS model was created and only two of the components had a cross-validation value below one. The component information is given in Table 18.

Table 18: Component information for rich density

Component number	Explained variance (Independent), %	Explained variance (Dependent), %	Cross-validation value
1	90,17	98,96	0,12
2	7,00	0,11	0,99

The response residuals are provided in figure 4-39 and as can be seen, the response residuals are linear and pass through the point y = 0,5. The response residuals are therefore normally distributed, indicating that the model is reliable.

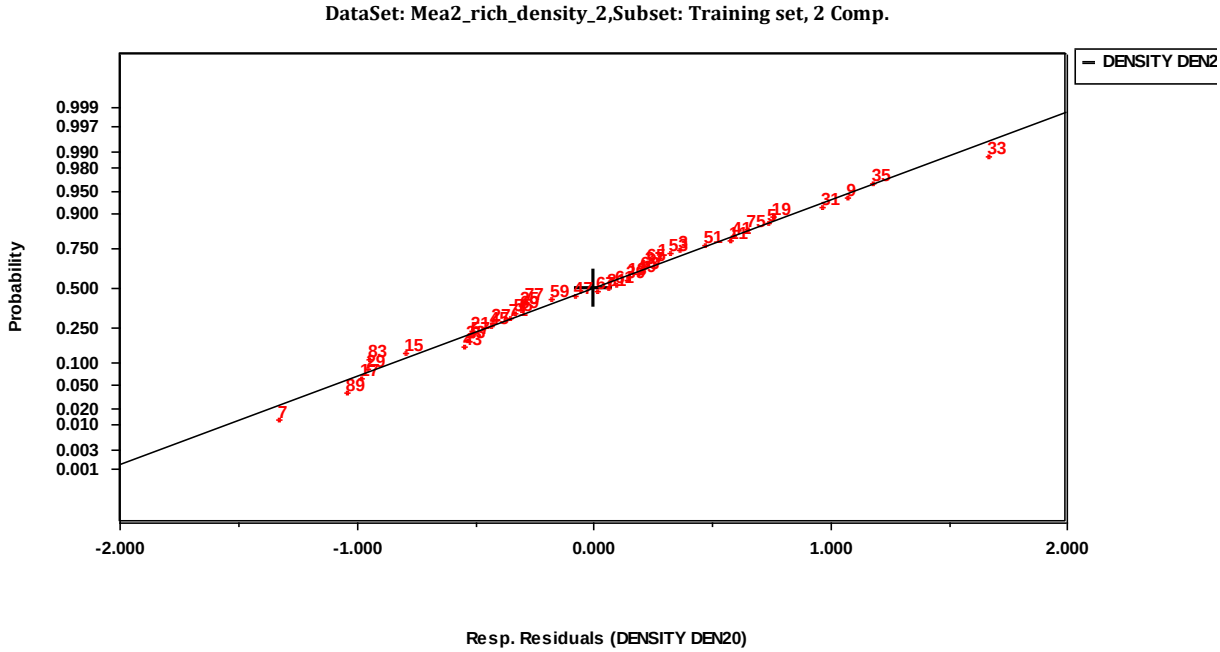


Figure 4-39: Response residuals for the model

The predicted versus measured plot for the validation set is provided in Figure 4-40.

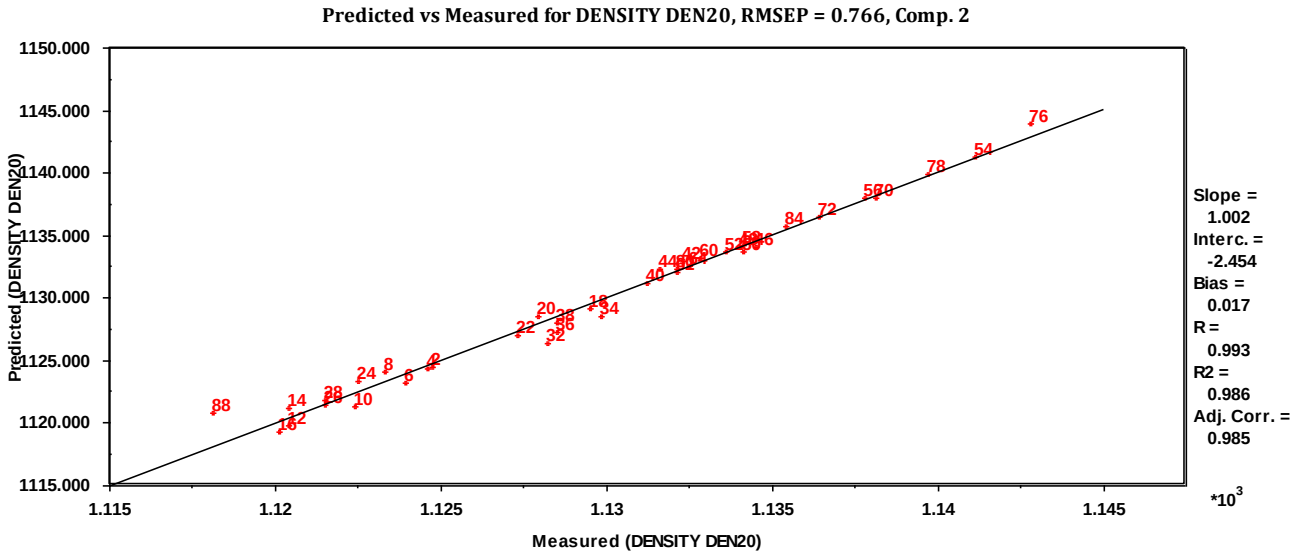


Figure 4-40: Predicted versus measured values

As can be seen from the predicted versus measured density plot the model is a good fit and can predict the values in the validation set with a low RMSEP value compared to the measured values. The average value of the density samples in the rich model is 1129,83, thus the RMSEP value (of 0,038) corresponds to 0,003 % of the average density value.

The low RMSEP value compared to the average value indicates that this model is reliable. Component information for the model is provided in table 19.

Table 19: Model information

Number of components	**R**	**R^2**	**R^2_a**	**RMSEP**
2	0,993	0,986	0,985	0,038

As can be seen from R^2 and R^2_a, there is little difference in value and the model is not overfitted.

5 Conclusion

This master thesis aimed to provide models for an in-line measurement with ATR-FTIR spectroscopy for an amine scrubbing plant. The data provided by TCM was not sufficient to model all the variables, as such, only models for TIC, TOT ALK and density are created. Principal Component Analysis has proven to be an excellent tool to find outliers and patterns in the data set. The elimination of outliers in the data improved the models significantly. A further investigation on the rich TIC model should be performed to increase the models predictive abilities. In general, the models generated demonstrated good predictive abilities.

Both a 2nd degree Extended Multiplicative Signal Correction and a 21-point Savitzky-Golay 3rd-degree polynomial fitting, 2nd-degree differentiation has performed good as pretreatment of the data, separately. All the models were tested for both pretreatments and the best model was chosen based on the Root Mean Square Error Prediction, number of outliers and number of components in the model.

In the model for total inorganic carbon, TIC, for the lean CO_2 samples, a Savitzky-Golay pretreatment with a 21-point window, 3rd-degree polynomial fitting and 2nd-degree differentiation resulted in the best model. The wavenumber region used to build the model is 1680-960 cm^{-1}. The model is built on two components and resulted in an RMSEP value of 0,027, where the average value of TIC is 1,153. The RMSEP value corresponds to 2,3 % of the average TIC value, indicating that the model is reliable.

The best model obtained for total alkalinity, TOT ALK, in the lean CO_2 samples was subjected to a 2nd degree EMSC pretreatment and used the wavenumber region 3640-2750 and 1680-960 cm^{-1}. The model is built on six components and resulted in an RMSEP value of 0,039 where the average TOT ALK value is 4,76. The RMSEP value is much lower than the average TOT ALK value and corresponds to 0,82 % of the average value. It is concluded that the model is reliable given how minor the RMSEP value is compared to the average TOT ALK value.

In the lean CO_2 samples, the best model for density was obtained by a 2nd-degree EMSC using the wavenumber region 3640-2750 and 1680-900 cm^{-1}. The model contained five components and resulted in an RMSEP value of 0,490 where the average density value is 1082,51. The RMSEP value for the density model is much lower than the average value. The RMSEP value corresponds to 0,045 % of the average value indicating that the model is reliable.

The TIC model built from the rich CO_2 samples has been pretreated with a 2nd degree EMSC and uses the wavenumber region 1680-960 cm^{-1}. Three components are included in the model and resulted in an RMSEP value of 0,047. The average value of TIC in the rich samples is 2,31 thus, the RMSEP value corresponds to 2,03 % of the average value. The RMSEP value is low compared to the typical values of TIC and the model is concluded to be reliable. The model should be further refined as there are some scattered results in the predicted versus measured plot of the validation set.

The TOT ALK model built from the rich samples has been pretreated with a 2nd degree EMSC and uses the wavenumber region 3640-2750 and 1680-960 cm^{-1}. Three components are included in the model and gave an RMSEP value of 0,034. The average value of TOT ALK in the rich samples is 4,70. The RMSEP value corresponds to 0,72 % of the average TOT ALK value, indicating that the model is reliable.

The best model obtained for density when using the rich samples were pretreated with a 2nd degree EMSC and used the wavenumber region 1680-900. Two components are included in the model and resulted in RMSEP value of 0,766, where the average value of density for the rich samples is 1129,83. The RMSEP value thus corresponds to 0,003 % of the average density value, indicating that the model is reliable.

An inline measurement with ART-FTIR coupled with multivariate methods has proven able to build models and predict values with a low RMSEP value compared to the value of the response variable.

6 Further work

More samples are needed to build a complete model of the degradation products produced in this process and to generate a reliable validation model. From the models created the lowest sample size was 80 samples. This was sufficient to build the rich CO_2 total alkalinity model and validate the model. Thus, a sample size of 80 should be enough to model the degradation products and validate the models.

An in-depth assessment of the net energy requirements of the power station to operate the carbon capture and storage system is still needed. This would presumably be associated with an economic feasibility study to determine the alterations required to return energy output capacity to original levels. To minimize degradation caused by high temperatures in the stripper, we suggest investigating the lower temperature threshold at which CO_2 separation from MEA is still viable.

If the temperature of the stripper is reduced, the rich CO_2-rich solution would require extended time in the stripper to separate the CO_2 from the aqueous MEA solution.
The negative impact to energy output capacity resulting from this reduction in efficiency would need to be thoroughly examined.

7 References

1. Luis, P. 2016. Use of monoethanolamine (MEA) for CO2 capture in a global scenario: Consequences and alternatives. *Desalination*, 380, pp.93–99.
2. Baird, C., Cann, M. 2012. *Environmental chemistry, fifth edition.* New York, W. H. Freeman and Company, pp. 169, 204.
3. Einbu, A. Ciftja, A. F. Grimstvedt, A. Zakeri, A. Svendsen, H. F. 2012. Online analysis of amine concentration and CO_2 loading in MEA solutions by ATR-FTIR spectroscopy. *Energy Procedia*, 23, pp.55–63.
4. Kenarsari, S. D., Yang, D., Jiang, G., Zhang, S., Wang, J., Russell, A. G., Wei, Q., Fan, M. 2013. Review of recent advances in carbon dioxide separation and capture. *RSC Advances*, 3(45), pp.22739–22773.
5. Ege, S. N., 2004. *Organic Chemistry: Structure and Reactivity, fifth edition.* Boston; Houghton Mifflin Company, pp. 804-806.
6. Kachko, A., Hamb, L. V., Bardowc, A., Vlugta, T. J. H., Goetheer, E. L. V. 2016. Comparison of Raman, NIR, and ATR FTIR spectroscopy as analytical tools for in-line monitoring of CO2 Concentration in an amine gas treating process. *International Journal of Greenhouse Gas Control*, 47, pp.17-24.
7. Stowe, H.M. and Hwang, G.S., 2017. Fundamental Understanding of CO 2 Capture and Regeneration in Aqueous Amines from First-Principles Studies: Recent Progress and Remaining Challenges. *Industrial & Engineering Chemistry Research*, 56(24), pp.6887–6899.
8. Cifre, P. G., Brechtel, K. Hoch, S. García, H., Asprion, N., Hasse, H., Scheffknecht, G. 2009. Integration of a chemical process model in a power plant modelling tool for the simulation of an amine based CO 2 scrubber. *Fuel*, 88(12), pp.2481–2488.
9. Hwang, G. S., Paek, E., Stowe, H. M., Manogaran, D. 2014. Reaction mechanisms of aqueous monoethanolamine with carbon dioxide: a combined quantum chemical and molecular dynamics study. *Physical Chemistry Chemical Physics*, 17(2), pp.831–839.
10. Koç, M. & Karabudak, E. 2017. History of spectroscopy and modern micromachined disposable Si ATR-IR spectroscopy. *Applied Spectroscopy Reviews*, pp.1–19.
11. https://www.miniphysics.com/electromagnetic-spectrum_25.html (accessed 14/4-2018)

12. Larkin, P. 2011. *Infrared and Raman spectroscopy : principles and spectral interpretation*. Elsevier, Amsterdam ; Boston, 1st edition.
13. Che Man, Y. B. , Syahariza, Z. A. Rohman, A. Rees, O. J. 2011. *Foriere Transform Infrared (FTIR) Spectroscopy: Development, Techniques, And Application In The Analyses Of Fats And Oils*. In: *Fourier Transform Infrared Spectroscopy: Developments, Techniques and Applications*, New York: Nova Science Publishers, Incorporated, pp.1-26.
14. Stuart, B. H. 2004. *Infrared spectroscopy: Fundamental and applications*. Chichester, West Sussex: Wiley.
15. Malainey, M.E. 2011. *A Consumer's Guide to Archaeological Science, Manuals in 23 Archaeological Method, Theory and Technique*. New York; Springer-Verlag New York Inc, pp.23-26.
16. Sun, D.-W. 2009. *Infrared spectroscopy for food quality analysis and control 1st edition*. Amsterdam: Academic Press/Elsevier.
17. Lindenberg, C. Cornel, J. Scho¨ll, J. Mazzotti, M. 2012. *Industrial Crystallization Process and Control, 1st edition, Chapter 9.* Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA.
18. Savitzky, A. & Golay, M.J.E., 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8), pp.1627–1639.
19. Karstang, K. V. 1996. Forbehandling av data. In: R., Nortvedt. F., Brakstad. O. M., Kvalheim and T., Lundstedt. *Andvendelse av Kjemometri innen forskning og industri.* Oslo: Tidsskriftforlaget Kjemi, pp.129-144.
20. Afseth. K., Kohler, A. 2012. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemometrics amd Intelligent Laboratory Systems*, 117, pp. 92-99
21. Grung, B. 1996. Det matematiske grunnlaget for latent-variabel metoder. In: R., Nortvedt. F., Brakstad. O. M., Kvalheim and T., Lundstedt. *Andvendelse av Kjemometri innen forskning og industri.* Oslo: Tidsskriftforlaget Kjemi, pp.121-128.
22. Rajalahti, T. Kvalheim, O. M. 2011. Multivariate data analysis in pharmaceutics: A tutorial review. *International Journal of Pharmaceutics*, 417(1), pp.280–290.
23. Bro, R. & Smilde, A.K., 2014. Principal component analysis. *Analytical Methods*, 6(9), pp.2812–2831.

24. Gabrielsson, J., Lindberg, N.O., Lundstedt, T., 2002. Multivariate methods in pharmaceutical applications. *Journal of Chemometrics*, 16(3), pp.141–160.
25. Pattern Recognition Systems. Sirius User Guide, 2009, Bergen
26. Mujica, L.E. et al., 2011. Q-statistic and T-2-statistic PCA-based measures for damage assessment in structures. *Structural Health Monitoring-An International Journal*, 10(5), pp.539–553.
27. Stordrange, L., Libnau, F. O., Malthe-Sørenssen, D., Kvalheim, O. M. 2002. Feasibility study of NIR for surveillance of a pharmaceutical process, including a study of different preprocessing techniques. *Journal of Chemometrics*, 16(8-10), pp.529–541.
28. Devore, J. & Berk, L., 2012. *Modern Mathematical Statistics with Applications*, New York, NY: Springer New York.
29. Andersen, C.M., Bro, R., 2010. Variable selection in regression—a tutorial. *Journal of Chemometrics*, 24(11-12), pp.728–737.
30. Kvalheim, O. M., Chan, H. Y., Benzie, I. F.F., Szeto, Y.T., Tzang, A. H. C., Mok, D. K. W., Chau, F. T. 2011. Chromatographic profiling and multivariate analysis for screening and quantifying the contributions from individual components to the bioactive signature in natural products. *Chemometrics and Intelligent Laboratory Systems*, 107(1), pp.98–105.
31. Jackson, P., Robinson, K., Puxty, G., Attalla, M. 2009. In situ Fourier Transform-Infrared (FT-IR) analysis of carbon dioxide absorption and desorption in amine solutions. *Energy Procedia*, 1(1), pp.985–994.
32. Dickson, A.G. 1981. An exact definition of total alkalinity and a procedure for the estimation of alkalinity and total inorganic carbon from titration data. *Deep Sea Research Part A, Oceanographic Research Papers*, 28(6), pp.609–623.

Appendix A

```
function [res, ana, idnr,lor, unit] = mea218(filename)

% This file can read and sort the different analysis methods,
% and put the right value of each measurement to the correct analysis
% and sample number.

% Removes information that is not needed.
m = readtable(filename);
m(:, [2, 4, 9, 10]) = [];

m1 = table2cell(m);

[n] = size(m1);

id = cell2mat(m1(:,1));

% Turns characters into numbers
res1 = char(m1(:,5));
res2 = string(res1);
res3 = str2double(res2);

v = char(m1(1,3));
v = string(v);

w = char(m1(1,4));
w = string(w);

v = [v, w];
v = join(v);

ana(1) = v;
minus = 0;
idnr(1,1) = id(1);
trekk = 0;
res(1,1) = res3(1);
lor(1,1) = m1(1,2);
unit(1,1) = m1(1,6);
for e = 2:n

   f = id(e-1,1);
   g = id(e,1);

   if f ~= g
      idnr(e-trekk,1) = g;
      lor(e-trekk,1) = m1(e,2);
      unit(e-trekk,1) = m1(e,6);
   else
```

```
      trekk = trekk + 1;
   end
end

for i = 2:n
   a = char(m1(i-1,3));
   a = string(a);

   b = char(m1(i,3));
   b = string(b);

   c = char(m1(i-1,4));
   c = string(c);

   d = char(m1(i,4));
   d = string(d);

   a = [a, c];
   a = join(a);

   b = [b, d];
   b = join(b);

   p = 0;
   p1 = 0;


   % This loop places the measurement to the right analysis
   l = length(ana);
   for k = 1:l

      if b ~= ana(k)
         p  = p + 0;

      else
         p = p + 1;
         co1 = k;
      end

   end
   if p > 0
      minus = minus + 1;
   else
      ana(1,i-minus) = string(b);
      co1 = i-minus;
   end


   % This loop places the sample to the right ID-tag (sample number)
```

```
    lid = length(idnr);
    for h = 1:lid
      if id(i) == idnr(h)
        co2 = h;
      end
    end

    % Coordinate for the measurement in the matrix
    res(co2,co1) = res3(i);

end

end
```

Appendix B

```
function [idnr, bolgetall, verdi, lor] = mea2_spekter(filename)

% This function can open multiple files and create a matrix consisting of
% wavenumbers, sample number and the intensity of each sample at each
% wavenumber. This file also produces a lean or rich vector to be able to
% separate the rich and lean samples.

a = importdata(filename);
a = string(a);

% This loop creates a matrix consisting of the intensities measured at each
% wavenumber for all the samples.

for i = 1:length(a)
   b = dlmread(a(i));
   bolgetall(:,i) = b(:,1);
   verdi(:,i)= b(:,2);
end

idnr = char(a);
idnr = idnr(:,1:5);
idnr = string(idnr);
idnr = double(idnr);

lor = char(a);
lor = lor(:,12:15);
lor = string(lor);

end
```

Appendix C

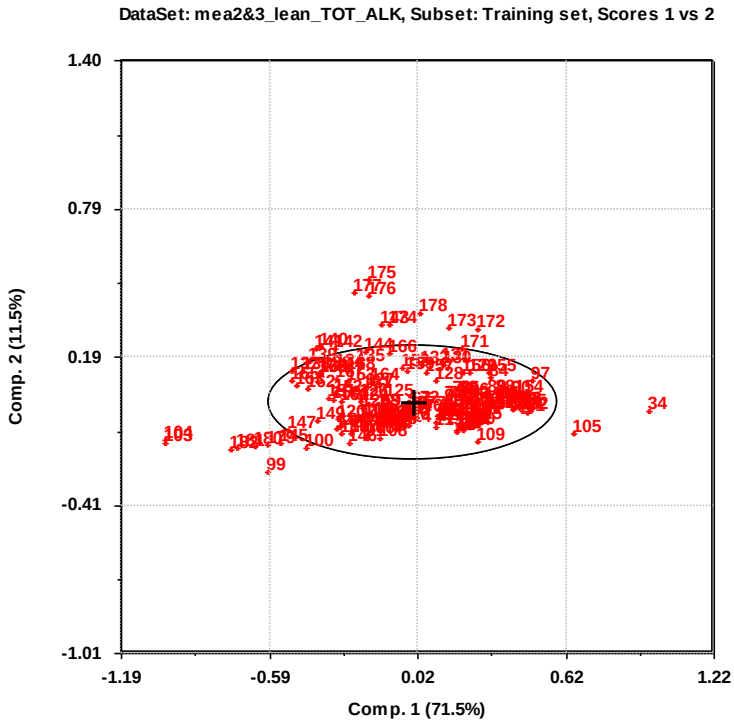
$$MEA + CO_2 \rightarrow MEA^+COO^-$$

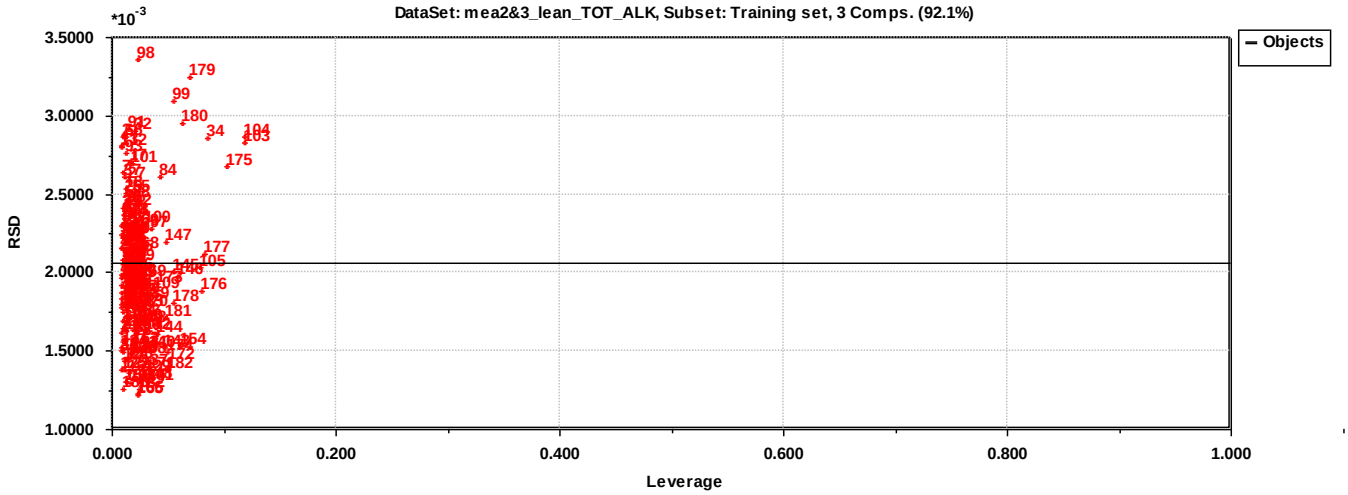
$$MEA^+COO^- + H_2O \rightarrow MEACOO^- + H_3O^+$$

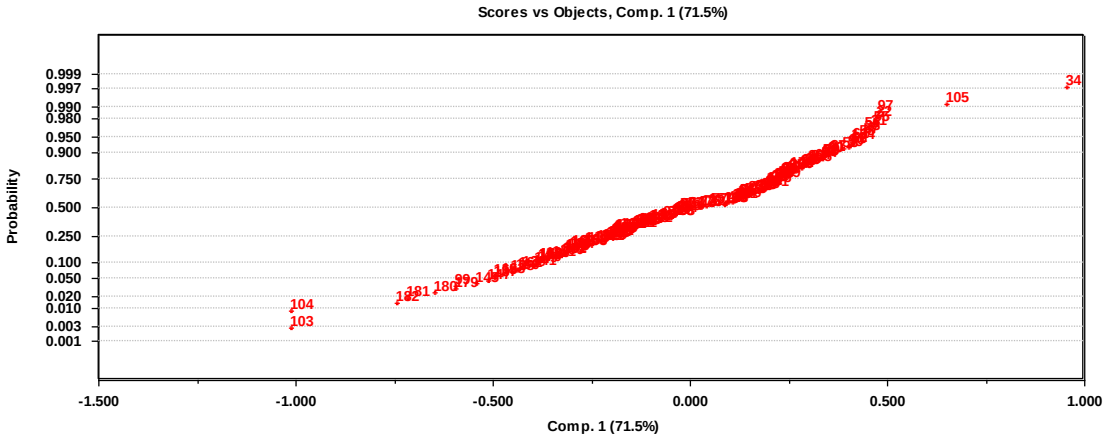
$$MEA + H_3O^+ + MEACOO^- \rightarrow MEAH^+ + H_2O + MEACOO^-$$

Appendix D

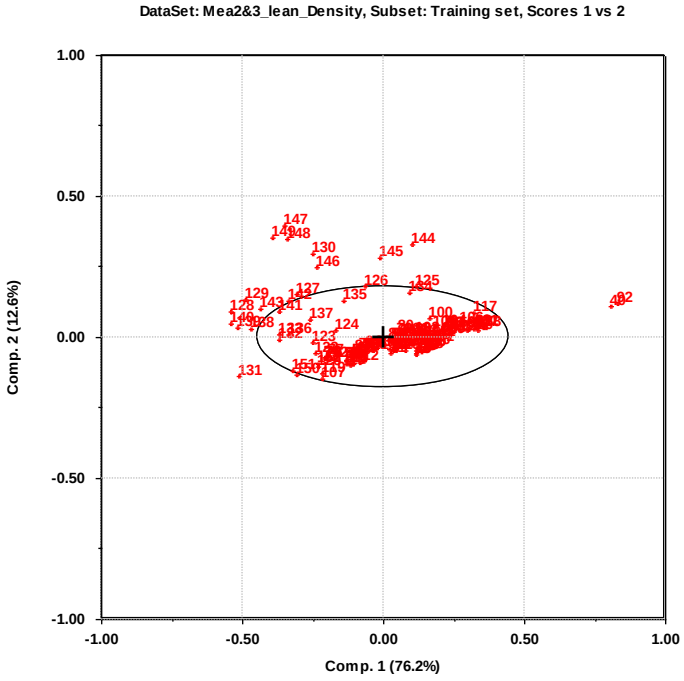
D-1. Lean TOT ALK

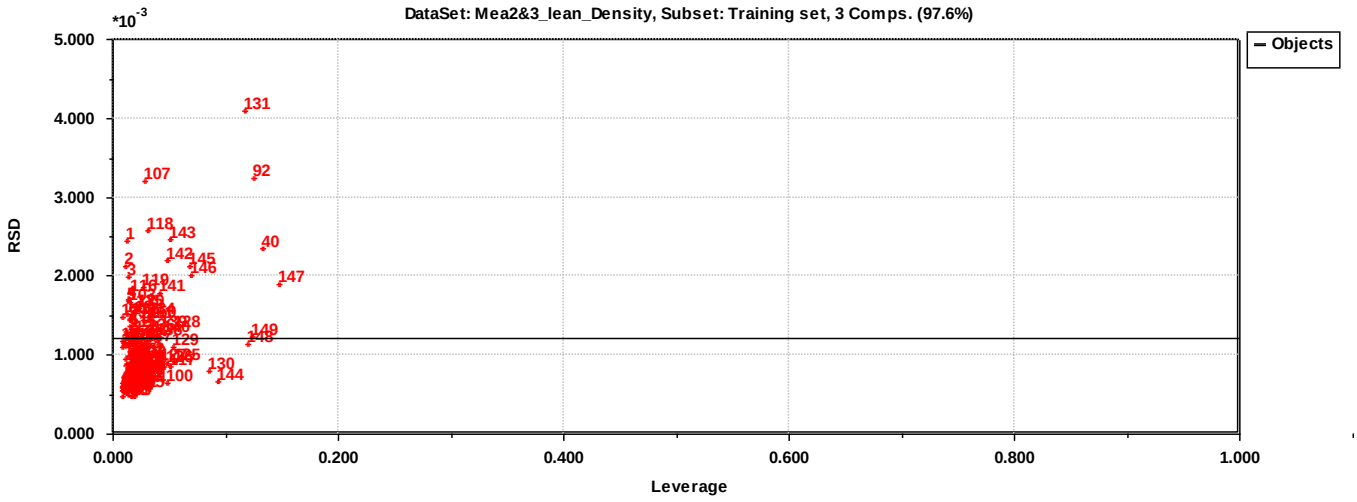


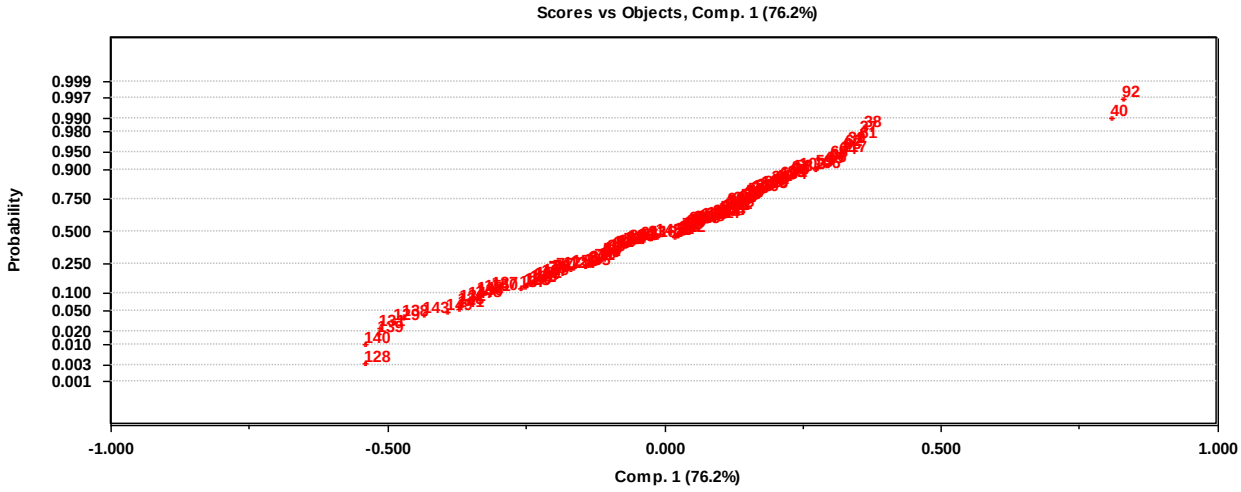




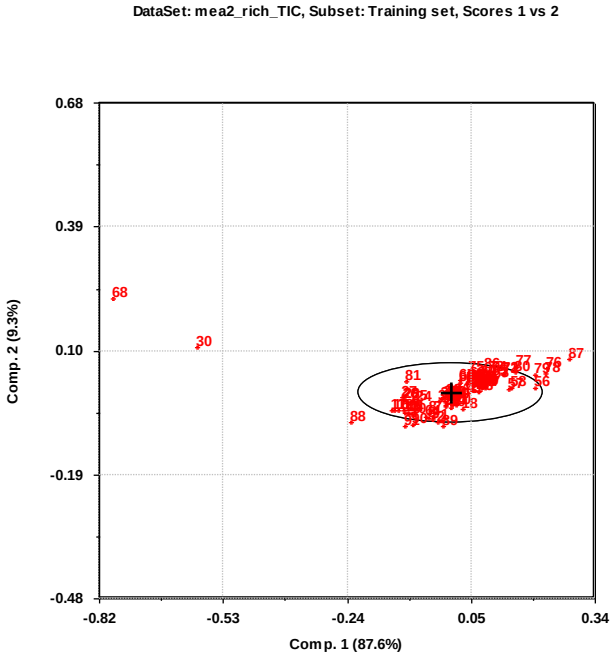
D-2. Lean Density

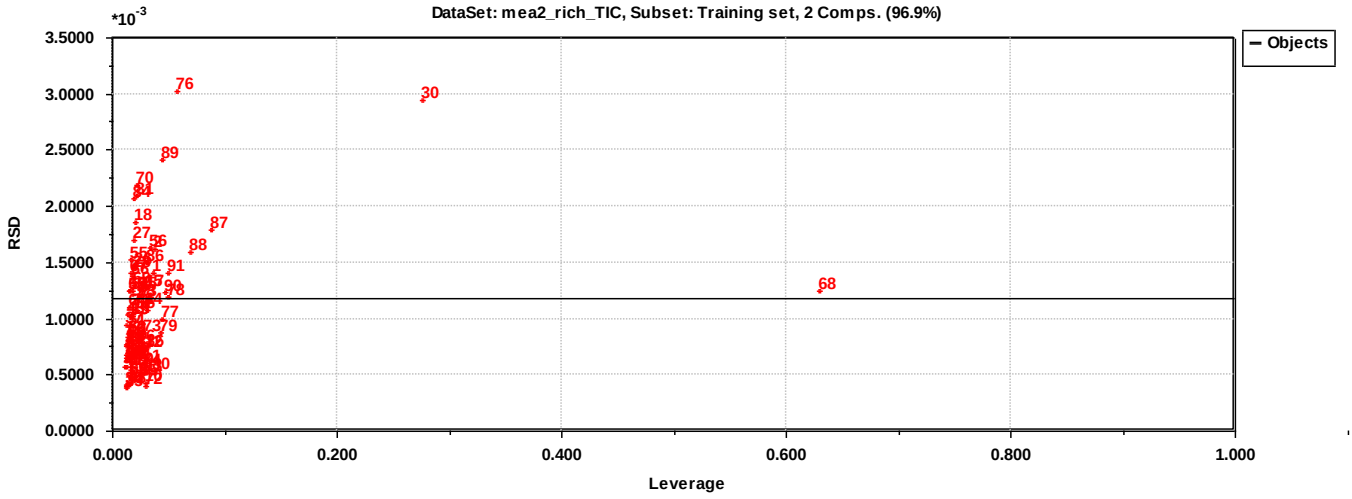


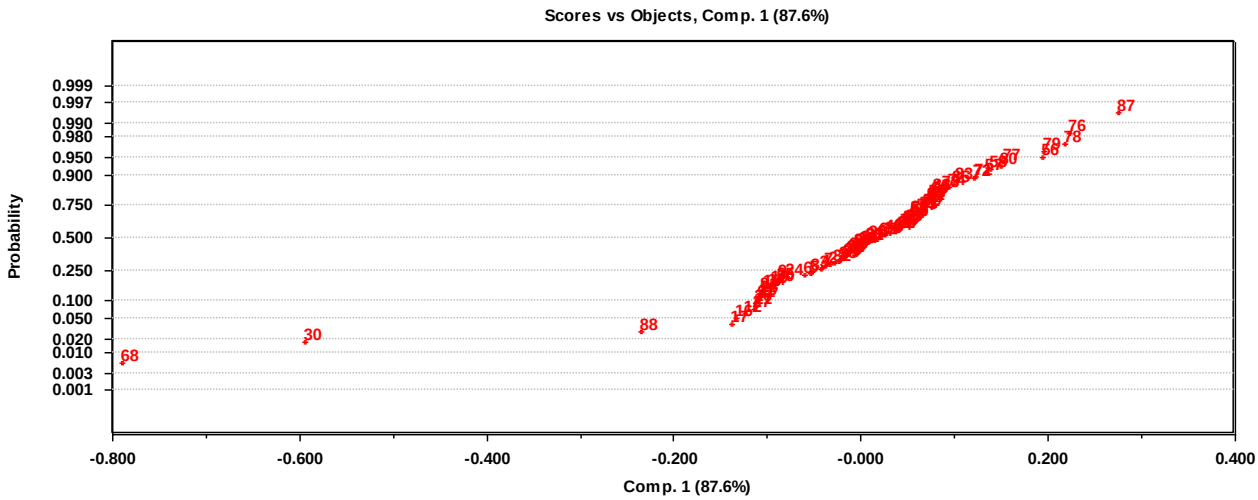



Scores vs Objects, Comp. 1 (76.2%)
Probability
0.999
0.997
0.990
0.980
0.950
0.900
0.750
0.500
0.250
0.100
0.050
0.020
0.010
0.003
0.001
92
40
140
128
-1.000
-0.500
0.000
0.500
1.000
Comp. 1 (76.2%)

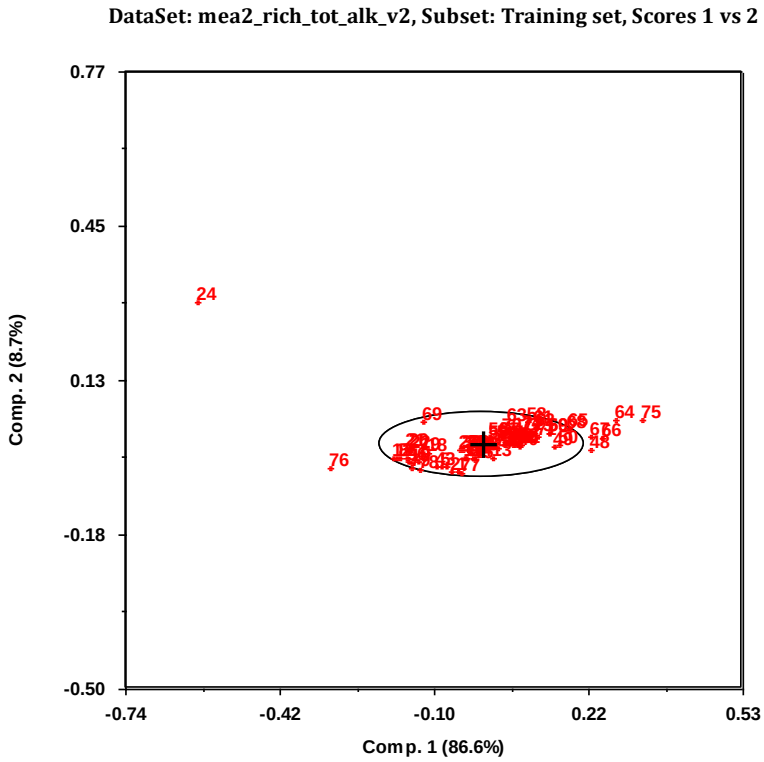
D-3. Rich TIC


DataSet: mea2_rich_TIC, Subset: Training set, Scores 1 vs 2
0.68
0.39
0.10
-0.19
-0.48
Comp. 2 (9.3%)
68
30
87
81
88
-0.82
-0.53
-0.24
0.05
0.34
Comp. 1 (87.6%)

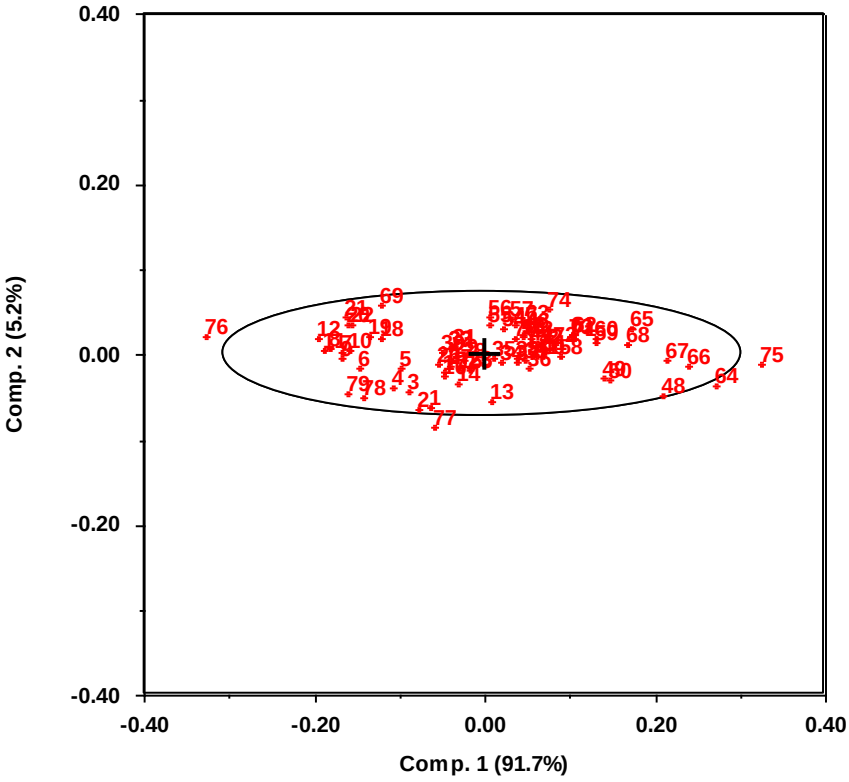

DataSet: mea2_rich_TIC, Subset: Training set, 2 Comps. (96.9%)
*10-3
Objects
RSD
3.5000
3.0000
2.5000
2.0000
1.5000
1.0000
0.5000
0.0000
76
30
89
70
18
87
88
91
68
0.000
0.200
0.400
0.600
0.800
1.000
Leverage

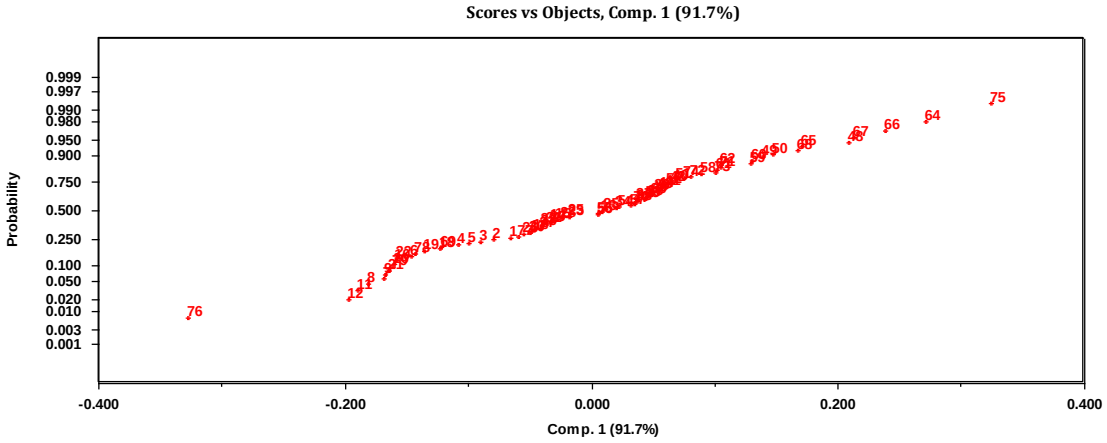


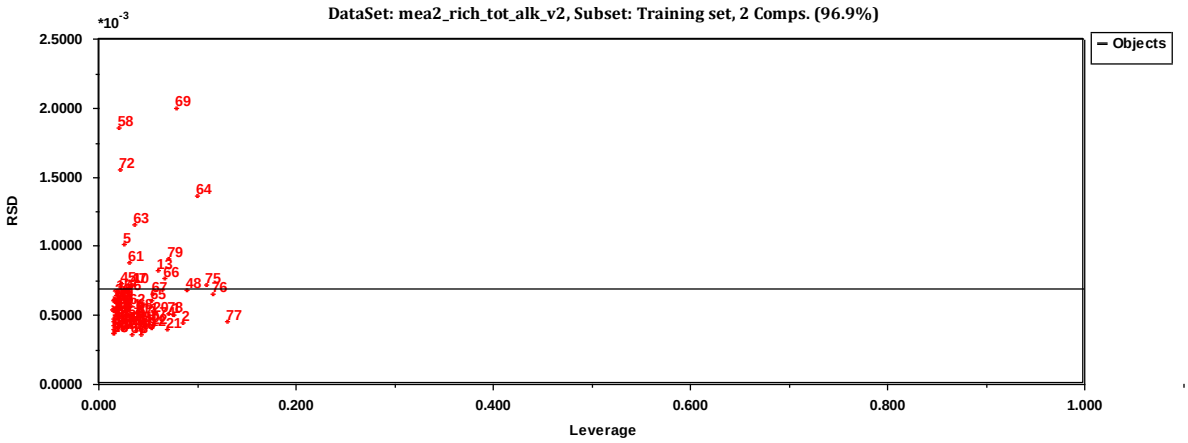
D-4. Rich TOT ALK



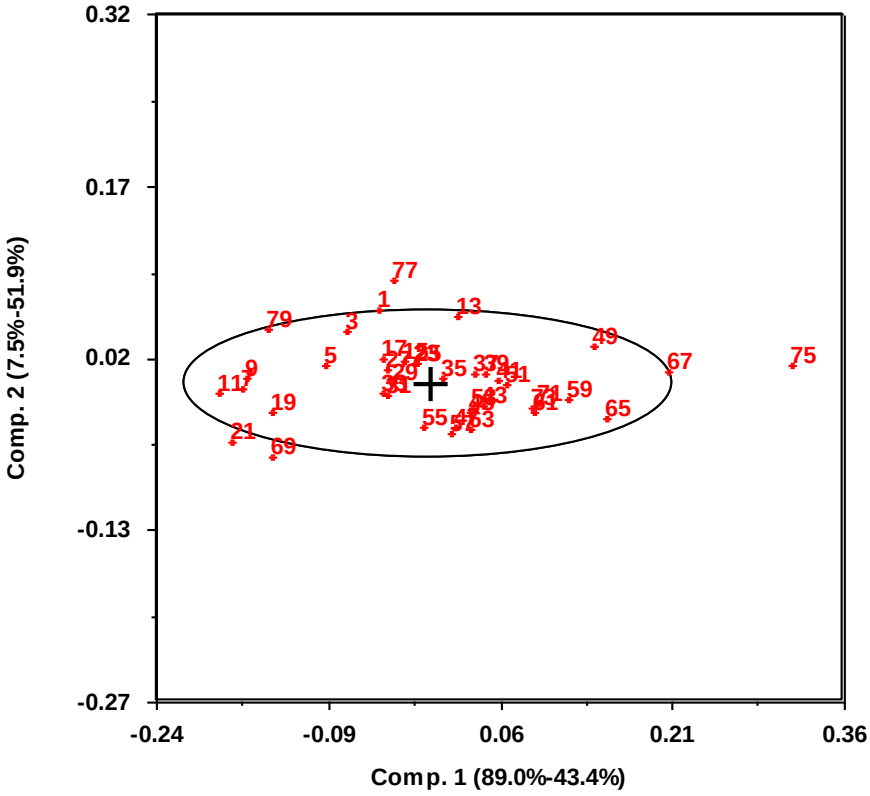
DataSet: mea2_rich_tot_alk_v2, Subset: Training set, Scores 1 vs 2



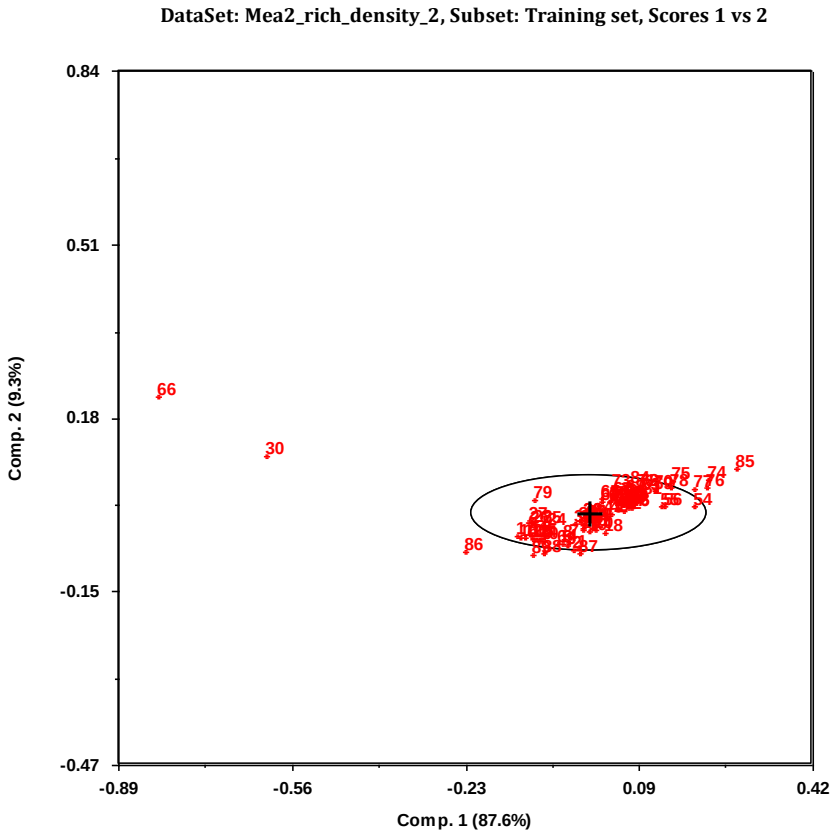


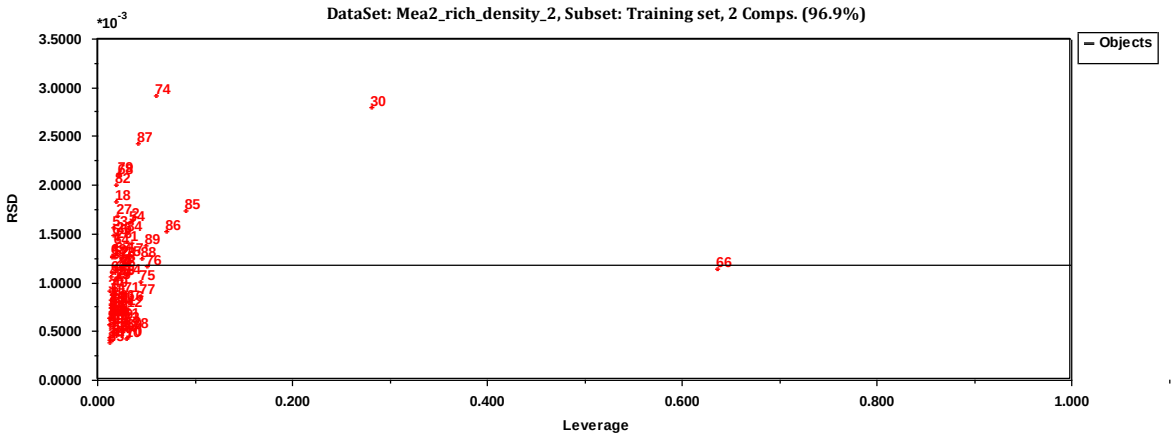


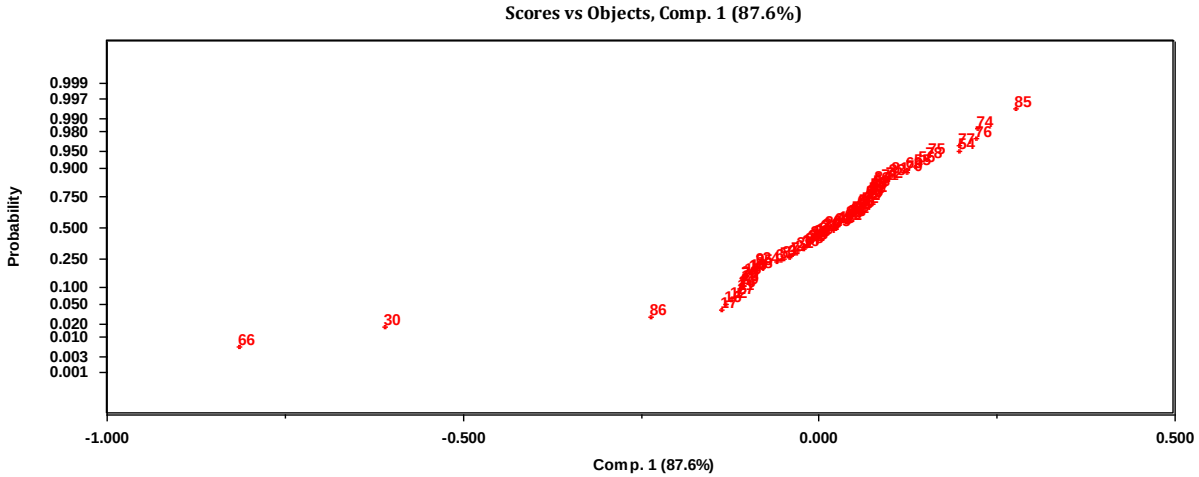
DataSet: mea2_rich_tot_alk_v2, Subset: Training set, Scores 1 vs 2



D-5. Rich Density






Scores vs Objects, Comp. 1 (87.6%)
Probability
0.999
0.997
0.990
0.980
0.950
0.900
0.750
0.500
0.250
0.100
0.050
0.020
0.010
0.003
0.001
-1.000
-0.500
0.000
0.500
Comp. 1 (87.6%)
85
74
76
77
54
86
30
66