Using Multivariate Data Analysis and ATR-FTIR Spectroscopy for Modeling Components Present During CO_2 Capture with Amines

Master Thesis in Process Technology

Helene Irgens Sjo

Department of Chemistry, University of Bergen

June 1. 2018

Acknowledgments

I want to thank my supervisor Professor Bjørn Grung for excellent help and supervision during the work on this thesis. Thank you for helping me with all my questions and problems. Also, thank you Egil Nodland for the help and guidance at the lab.

I would also thank TCM for the opportunity to write about this subject, and for providing the samples.

Most of all I would like to thank all my fellow students for making the last five years genuinely memorable. A special thanks to Henning Kvaløy and Vilde Loug Pedersen for all the coffee-breaks, advice and discussions on and of topic.

Furthermore, a big thank you to my family and my boyfriend for your support. You inspire me, and always encourage me to do my best.

Thank you,

Helene Irgens Sjo

Abstract

This master thesis is a collaboration between Technology Centre Mongstad (TCM) and University of Bergen. The project is to develop a method to accurately predict total inorganic carbon, total alkalinity and density using spectroscopy and multivariate data analysis. These variables can be used to determine the CO_2-loading and MEA concentration.

The CO_2 concentrations in the atmosphere have been increasing since the 19th century; the increase has been affected by anthropogenic CO_2 emissions. The most significant source of anthropogenic CO_2 is the combustion of fossil fuels, especially in large power plants. The use of post-combustion CO_2 capture at large power plants can decrease the amount of emissions drastically. Monoethanolamine (MEA) has been extensively studied as an aqueous solvent to use in CO_2 capture and is a good choice for this purpose. Other solvents have not been this extensively studied. Therefore it is not sure which solvent that is the best choice yet.

This thesis aims to use multivariate data analysis to create models that can be used for prediction of the compounds in the MEA-solution at different times in the process. Three response variables are chosen, total inorganic carbon (TIC), total alkalinity (TOT_ALK) and density. TIC can be used to find the CO_2 concentration, TOT_ALK for finding the amine concentration and the density is correlated to the CO_2-loading.

The three response variables are predicted using partial least squares (PLS) models, preprocessing of the data is done with extended multiplicative signal correction (EMSC) or Savitzky-Golay differentiation. Outlier detection has been performed with principal component analysis (PCA). The achieved models have good predictive abilities, with small prediction errors and residuals.

List of Abbreviations

ATR-IR	Attenuated total reflectance infrared
ATR-FTIR	Attenuated total reflectance Fourier transform infrared
EMSC	Extended multiplicative signal correction
IR	Infrared radiation
MEA	Monoethanolamine
MLR	Multiple linear regression
MSC	Multiplicative signal correction
OPD	Optical path difference
PCA	Principal component analysis
PC	Principal component
PLS	Partial least squares
PRESS	Predicted residual error sum of squares
RMSEP	Root mean square error of prediction
RMSECV	Root mean square of cross-validation
RSD	Residual standard deviation
SR	Selectivity ratio
TIC	Total inorganic carbon
TOT_ALK	Total alkalinity
VIP	Variable importance prediction

Table of Contents

1 Introduction

The buildup of CO_2 and other greenhouses gases in the atmosphere has led to a rise in the temperature of the earth, causing what is known as the greenhouse effect [1, p. 165-166]. This increase of CO_2 is mainly due to anthropogenic emissions, where the most significant sources are burning of fossil fuels and large power plants. To decrease the emissions, the CO_2 can be captured, stored and used for other purposes.

The capture of CO_2 from flue gas can be performed by using a chemical solvent that absorbs the CO_2 [1, p. 253-254]. The CO_2 is captured by this chemical agent and later released to be compressed, transported or stored. The chemical solvent being used must be regeneratable so that it can be used multiple times. The most commonly used agent for CO_2 capture is Monoethanolamine (MEA) and diethanolamine; they have a high water solubility and the ability to absorb high amounts of gas. The required heating of the solvent to release the CO_2 is relatively little compared to other chemical solvents, which makes the regeneration process more accessible and can be done at lower temperatures than the alternatives. Because of their high absorption capability, over 95% recovery of CO_2 can be accomplished using MEA.

Amine scrubbing is a technology used in the capture of CO_2 [2]. In a typical system design for a power plant, the flue gas is passed through the absorber, containing the aqueous amine. The amine absorbs the CO_2 and travels to the desorber where the solution is heated using water vapor until the amine releases the CO_2. The process is then repeated, using the regenerated amine.

The use of multivariate methods and analysis, while the process is happening, can be a significant advantage, especially if new mixtures or methods are tested for the process, that way the process can be stopped if problems are detected [3]. The way things are now these problems are not detected until the problem has occurred and may already have caused huge problems.

Several spectroscopic methods have been studied, determining if they are suited for this use. ATR-FTIR has shown to have good applicability and is applied as spectroscopic method in this thesis. ATR-FTIR used together with regression analysis, and partial least squares (PLS) have given satisfying predictions with small deviations.

2 Spectroscopy

2.1 The Electromagnetic Spectrum

The term light refers to all wavelengths or frequencies of electromagnetic radiation, both the visible light and the radiation that cannot be seen, but whose effect can be measured [4]. The entire range of wavelengths and frequencies comprises the electromagnetic spectrum. The spectrum is divided into regions according to increasing frequency, which is displayed in **figure 2-1**.

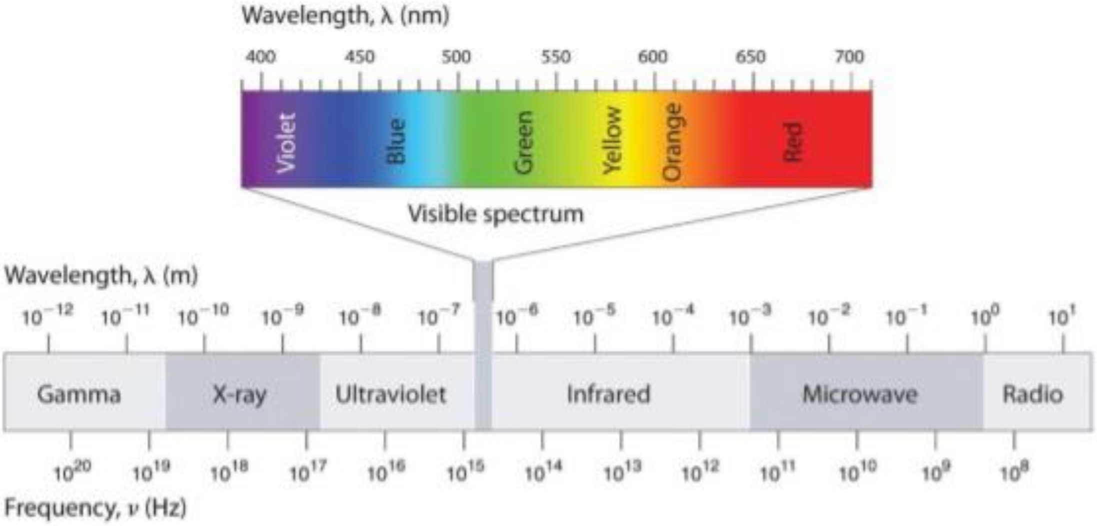


Figure 2-1 The electromagnetic spectrum [5]. The highlighted area is the visible part of the spectrum, and the grey area is the part of the electromagnetic radiation that cannot be seen

The infrared region is between the microwaves and the visible light, whereas the near-infrared region is located closest to the visible light in wavelength and is defined as the region from 780 – 2500 nm (nanometers) [5, p. 364]. The highlighted region is the visible light, reaching from 400 nm – 700 nm.

The light behaves both as a particle and a wave and can be seen as either, depending on the property being measured [6]. As a wave, light has the properties wavelength and frequency. Frequency ν is the number of waves that passes a given point per second given in s^{-1}, or more commonly called Hertz (Hz). Wavelength λ is the distance between two corresponding positions of head-to-head waves.

The velocity of light in a vacuum is a constant and is the product of frequency and wavelength:

$$c = v \cdot \lambda \qquad \textbf{Equation 2-1}$$

where c is the speed of light in vacuum (3 x 10^8 m/s).

Wavelength and frequency are inversely proportional, meaning that long wavelengths give low frequencies and opposite. In vibrational spectroscopy the use of wavenumber units is more commonly used, wavenumber is linear with energy and defined as follows [14]:

$$\bar{v} = \frac{1}{\lambda} = \frac{v}{c} \qquad \textbf{Equation 2-2}$$

where $\bar{v}$ is the wavenumber.

As a particle, the light behaves like a particle of energy, called a photon. This represents the energy in the electromagnetic spectrum, and is given by the Bohr equation:

$$E = h \cdot v = \frac{hc}{\lambda} \qquad \textbf{Equation 2-3}$$

where h is Planck's constant (6.626 x 10^{-34} Js).

2.2 Vibrational Spectroscopy

The study of the interaction between electromagnetic radiation (light) and matter is called spectroscopy. The intensities of absorption or emission of radiation at one or more specific wavelengths are measured [6].

When a molecular, or atomic, system absorbs or emits light, the system moves from one quantized energy level to another. The difference in energy levels must, according to the Bohr frequency condition, be equal to the light emitted or absorbed. This principle is used in spectroscopy to review the energy levels of matter and displays how matter and energy interact.

When the radiation interacts with the sample, the electrons are excited to a higher energy state than before the interaction. If the excited electrons emit photons and thus relaxing to a lower state, emission is observed.

Reflection or transmission can be observed depending on the specifications of the sample surface. Medium difference is the most trivial and effective reason for reflection, while a sample that is transparent at the given wavelength will lead to the radiation passing through the sample. The different interactions are displayed in **figure 2-2**.

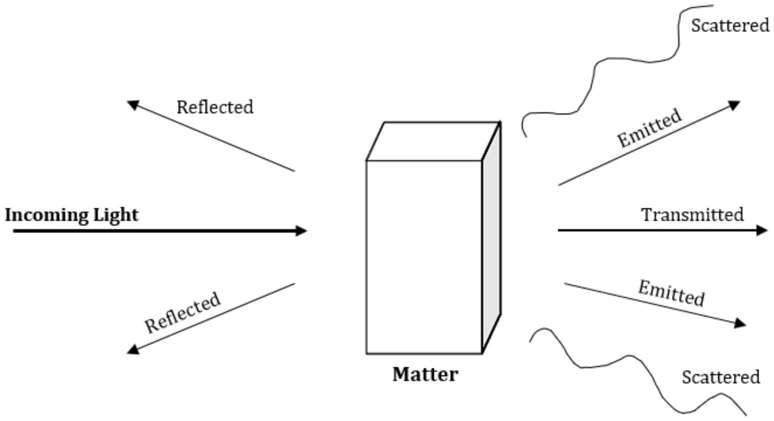


Figure 2-2 Interactions of light and matter, adapted from [5]. Displaying reflected, transmitted, emitted and scattered radiation

A photon striking the outer layer of an atom can only be absorbed if the amount of energy is precisely equal to the energy differences between the high and low quantum energy levels.

The different spectroscopic techniques consider the different interactions between the photons and the electrons in molecules and atoms, and spectroscopy can thus give valuable, complex and definite information about the interactions in the substance being studied.

Preprocessing of vibrations spectroscopy spectra are essential; the spectra can contain multiplicative or additive effect that will make the interpretation and analysis difficult and may lead to wrong conclusions [7]. The goal of the preprocessing is to remove these unwanted effects. In a quantitative chemical analysis of complex samples from spectroscopy, unwanted light scattering effects are often present and complicate the analysis [8]. These unwanted effects occur due to physical variation in the samples, differences in particle shape and size, sample surface and packing and so on.

In some cases, a simple multiplicative effect of the light scattering occurs, like a change of optical path length, which will change the scale of the whole spectrum by a given factor. Additive effects in the form of a simple baseline shift can under some conditions be observed. These simple cases of pure multiplicative and additive effect of light scattering are rare, and an oversimplification. A complex system will most likely have both effects present, which needs to be corrected before further analysis.

2.3 Infrared Spectroscopy

The chemical bonds that hold molecules together are never at rest but vibrate continuously [6]. These vibrational movements induce absorption in the infrared region. The infrared radiation can also excite rotational movements of molecules, giving rotational bands, which are overlaid on the vibration bands.

Samples analyzed by IR spectroscopy must have a dipole moment, so solids, liquids, and gases can be studied by this technique. The molecules in the samples are identified by determination of the chemical structure according to the frequencies of the absorbed IR radiation.

Vibrations in the molecule that lead to a change in the dipole moment can be recorded; there are two types of molecular vibrations that can lead to this: bending and stretching. The different types of bending and stretching are represented in the **figure 2-3**.

In IR spectroscopy the relationship between the incident and transmitted radiation and the concentration of the sample is given by Beer-Lambert law. Empirically it has been found that the transmitted intensity varies with the length and molar concentration of the sample [10, p. 479-480]:

$$A = \varepsilon c L$$ **Equation 2-4**

Where A is the absorbance, ε is the molar absorption coefficient, and L is the length. The molar absorption coefficient depends on the frequency of the incident radiation and has the greatest value where the most intense absorption is.

The spectrum of the sample being analyzed is found by plotting the absorbance or transmittance versus the wavenumber [11, p. 13-15]. The energy differences between the excited and ground state are proportional to the wavenumber.

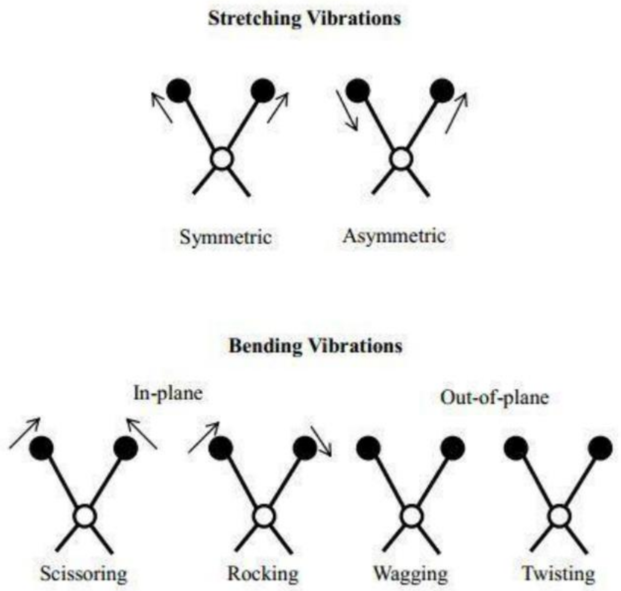


Figure 2-3 Molecular vibrations, displaying both stretching and bending [9]

2.3.1 Fourier Transform Infrared Spectroscopy

The interferometer is an optical device consisting of an IR-source, a beam splitter, and two mirrors, one moving and one fixed [12]. The incident light from the source is split by the beam splitter, and one half is sent to each of the two mirrors. The beam is then reflected by the mirrors and recombine at the beam splitter before it goes through the sample and travels to the detector. This principle is displayed in **figure 2-4**.

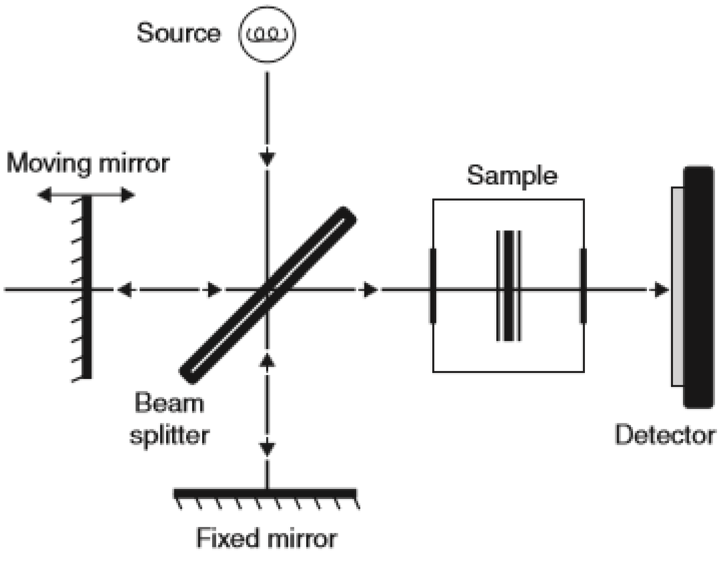


Figure 2-4 Principle of the interferometer [12]

Since one of the mirrors are moving, the distance the beam has to travel varies, this difference in travel is called optical path difference (OPD). The intensity of the beam is the highest at zero path difference when the mirrors are at the same distance from the beam splitter, the beams now interfere constructively. A low-intensity beam occurs when the light beams from the mirrors are out of phase and interfere destructively. Complete destructive interference is achieved when the path difference is an integer + ½ multiple of the wavelength. The interferogram is a plot over the intensity of the beams over the OPD.

The interferogram obtained using an interferometer must be transformed to become a readable spectrum. The spectrum is a representation of the intensity over wavenumber or frequency, and to do this conversion Fourier transform (FT) can be used [12].

The Fourier Transformation is a mathematical method that transforms the interferogram to a spectrum. The conventional spectrum is produced by breaking down the interferogram into sine waves for each wavelength of the light. The method involves integration of the original data between zero and maximum path difference; the signal is converted from intensity over a given time to intensity with respect to frequency.

2.3.2 Attenuated Total Reflectance Infrared Spectroscopy

Attenuated total reflectance Fourier transform infrared spectroscopy (ATR-FTIR) is a favorable choice for examining samples containing water; this is a surface characterization technique where the IR light enters a specific ATR crystal [6]. This crystal provides attenuated total reflectance (ATR) of the IR beam inside the crystal. The crystal is in contact with the sample and provides internal reflections of the IR beam. From these internal reflections, evanescent waves that penetrate the sample are formed. These waves enter the sample at a depth of 0.5-2 μm, causing IR radiation and interaction with the sample, resulting in an ATR-FTIR spectrum of the sample. The principle of the ATR-FTIR spectrometer is shown in **figure 2-5**, displaying the reflection angle and the evanescent waves. The sample is place directly on the crystal. ATR-FTIR can be used to analyze the surface of a sample, and the penetration depth can be changed by changing the reflection angle.

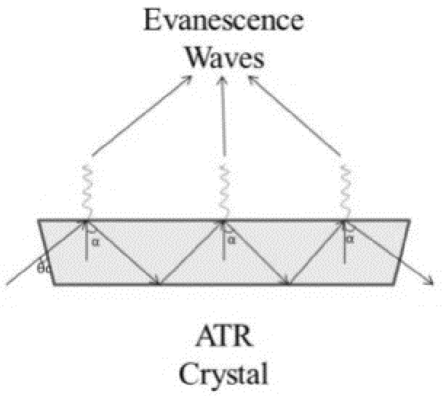


Figure 2-5 Principle of ATR-IR, where α is the reflection angle [6]

In ATR-IR one usually uses mid-range IR beams because of the fingerprint region. This region contains molecule-specific molecular vibrations and can be used to identify compounds in the sample.

Another advantage with ATR-FTIR is that the sample preparation is more straightforward than for IR, the sample is placed directly on the crystal. The amount of sample needed for this spectroscopic method is minimal, and there is no need to destroy a sample.

Since ATR-FTIR is based on the use of evanescent waves the sensitivity is not as good as for the transmission methods. This is because in the transmission technique the waves pass through the sample while for the evanescent wave technique the waves only touch the surface of the sample. As a result, the evanescent wave method only examines a small part of the sample, while the transmission method allows for analyzing the whole sample, both bulk and surface.

2.3.3 Interpretation of Spectra

A disadvantage of IR-spectroscopy is the large absorption of O-H from water, which potentially can bury signals from other compounds in the aqueous solution. If the samples contain water, ATR-FTIR spectroscopy is a better choice.

The most significant region of the IR-spectrum is from 4000-665 cm^{-1} [13]. In the high-frequency region, the stretching vibrations for the most important functional groups are found.

Amines show broad, moderate absorption in the low-frequency region, especially in the region ensuing 950 cm^{-1}, which can be seen in **figure 2-6.**, in the form of C-NH_2 absorption. The fingerprint portion of the spectrum, from 1300 – 900 cm^{-1} is often complex, with interacting vibration modes. This region has unique absorption for every molecular specie and can be used to identify which compounds are present in the samples.

O-H is one for the most important functional groups, and the stretching of O-H produces a broad band in the region 3700 to 3600 cm^{-1} [14]. The stretching of the inorganic compounds in the aqueous amine solutions (see section 3.2), are found in the region from 2000 to 900 cm^{-1}. The organic carbons, in form of C-H stretching are identified in the region enclosing 2900 cm^{-1}, these are often buried by the O-H peak, making it hard to identify them.

The stretching vibrations described in **figure 2-6** are the most important when studying solutions of aqueous MEA present during the CO_2 capture process (more in section 3.2).

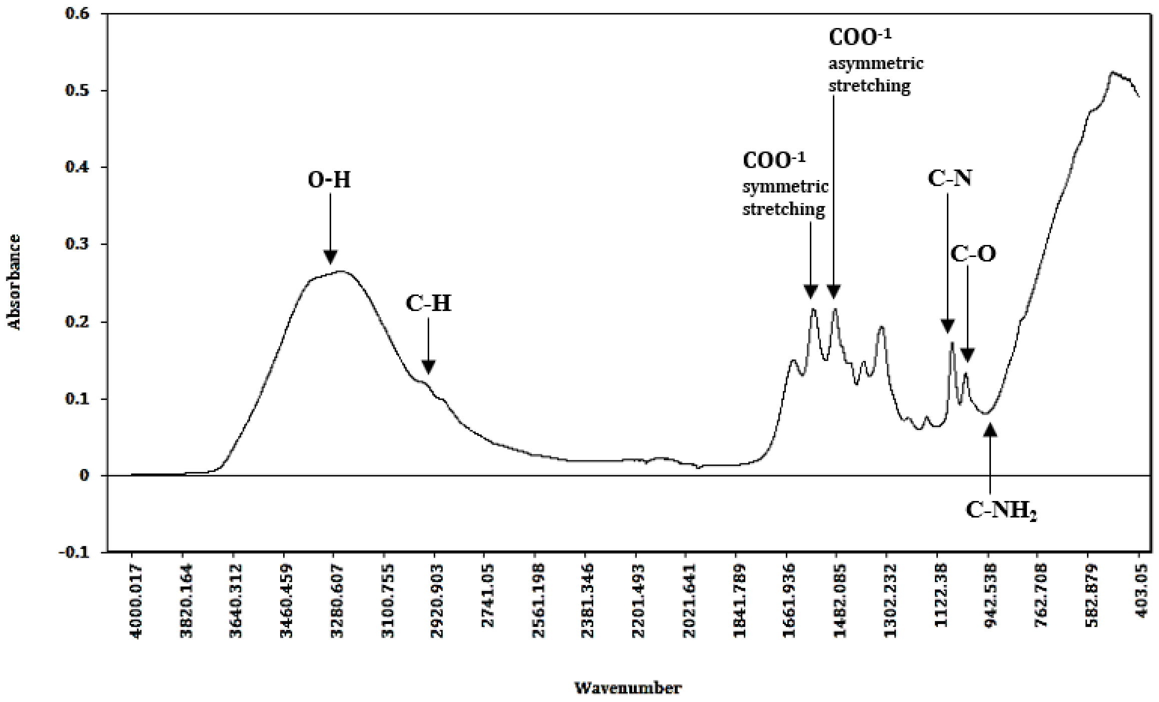


Figure 2-6 ATR-FTIR spectrum of a sample of aqueous MEA solution, fully loaded with CO_2, displaying the most important stretching vibrations. Based on [13, 14]

3 CO_2 Capture

One of the major causes of global warming is the increase in the global CO_2 concentration in the atmosphere, alongside the increase of other greenhouse gases [15]. This increase in the CO_2 concentration is due to human activities, such as the use of fossil fuels, and has been found to be the leading cause of climate change. The chief source of anthropogenic CO_2 emissions is through the burning of natural gases and coal.

Numerous techniques for CO_2 capture exist and are classified into four groups: pre-combustion, post-combustion, oxy-fuel combustion and electrochemical separation [16]. Post-combustion CO_2 capture is the technique with the highest potential since it can be retrofitted to already existing power plants.

In post-combustion capture, the technology is retrofitted to the existing power plants, and the CO_2 is captured after the fossil is burnt. The process uses chemical absorption, physical adsorption or membranes to capture the CO_2. High energy requirements are related to the chemical absorption, due to the regeneration of the solvents and loss during the absorption. Monoethanolamine (MEA) has been used for decades to capture CO_2 and is extensively studied [17]. Aqueous MEA solutions can absorb CO_2 at low pressure with acceptable absorption/desorption kinetics.

3.1 Amines

Amines are compounds where one or more of the hydrogen atoms in the ammonia molecule has been replaced with an organic group [18]. Amines are divided into primary, secondary or tertiary amines, depending on the number of organic groups bonded to the nitrogen atom. Amines with organic groups that are not too large are soluble in water.

MEA is a primary amine containing one organic group and is soluble in water. MEA has a hydroxyl group, also making it a primary alcohol. MEA is a weak base and can be used directly in industrial processes, e.g. for removal of acidic gases like CO_2 [19].

3.2 CO2 Capture Using Amines

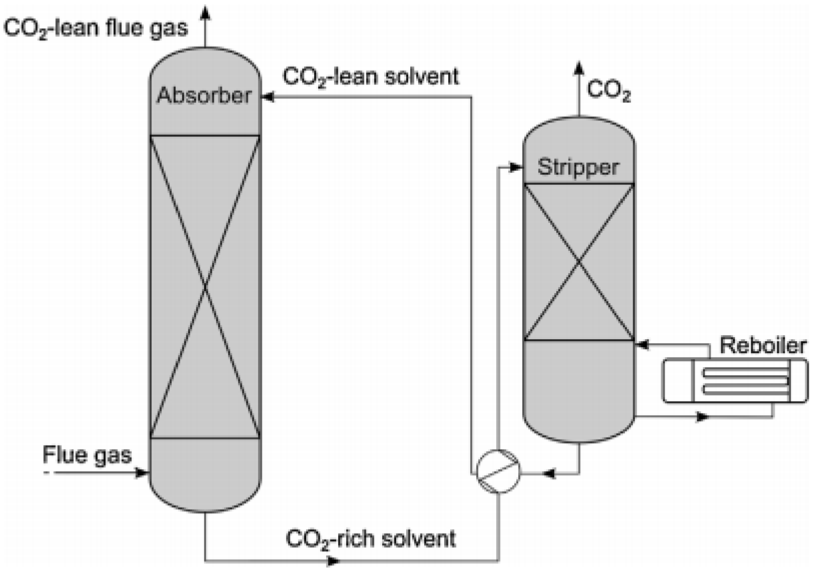


Figure 3-1 Schematic presentation of an amine scrubbing unit [3]

The principle of CO_2 capture using amine scrubbing is demonstrated in **figure 3-1**. The flue gas flows through the absorber, countercurrent to the amine solution, where the CO_2 is absorbed via chemical reactions with the amine solution [17]. The CO_2 is now absorbed by the amines, as a part of the liquid solution. The liquid stream of amine solution and CO_2 then moves over to the stripper (desorber); here the solvent is regenerated, and the CO_2 is released into a gas stream, containing water vapor and CO_2. The amines are regenerated through heating, when the amines are heated they release the CO_2, and the amines can then be used again. During the regeneration in the stripping column degradation products can form, which will affect the amines ability to absorb CO_2 [2]. The compounds can also affect the equipment, in the form of for example corrosion. The gas stream containing CO_2 and water vapor is cooled down, and the water condensed [17]. Pure CO_2 can now be produced for storage and transport.

During the CO_2 capture in aqueous amines, two amine molecules react with one CO_2 molecule, which forms a carbamate ion and a protonated amine. This limits the loading capacity to 0.5 mole CO_2 per mole amine. The overall reaction is as follows:

$$2\ \mathrm{MEA} + \mathrm{CO_2} \rightarrow \mathrm{MEACOO^-} + \mathrm{MEAH^+}$$ **Equation 3-1**

The process can happen in a single-step direct mechanism or a two-step zwitterion mechanism. Recent studies have determined that this happen in a zwitterion mechanism [20]. This zwitterion mechanism includes the formation of a zwitterion as an intermediate, undergoing deprotonation by another MEA which then forms carbamate and protonated MEA. The underlying reactions to achieve the overall reaction are presented in Appendix A.

When studying the CO_2 capture process and the compounds present at different stages of the process, samples are taken [21]. One set of samples are taken after the solvent has passed through the absorber, these samples contain MEA and absorbed CO_2, and are called CO_2 rich samples. These samples are fully loaded with CO_2, and also contain products formed in the absorption process. Another set of samples are taken after the regeneration of the solution in the desorber and contains mostly MEA and are called CO_2 lean samples; these samples also contain the same byproducts as the rich samples. The only difference between lean and rich samples is the amount of CO_2, the difference in CO_2 concentration gives the samples different properties, like viscosity and density. Because of this difference analysis has to be performed separately for lean and rich samples.

Figure 3-2 displays the spectra of one lean and one rich sample. The difference between the samples is largest in the fingerprint region to the right, where the rich sample has a higher intensity of absorption. This is the region where the inorganic carbons absorb radiation (see section 2.3.3). The rich samples are fully loaded with CO_2 and byproducts from the absorption process, these compounds are inorganic carbons, and is the reason for the higher intensity.

The region around 950 cm^{-1} display a big difference between the two samples and is where the C-NH_2 absorption appear [13]. The C-NH_2 is MEA, and there is more pure MEA in the lean samples, resulting in a higher intensity for the lean samples here.

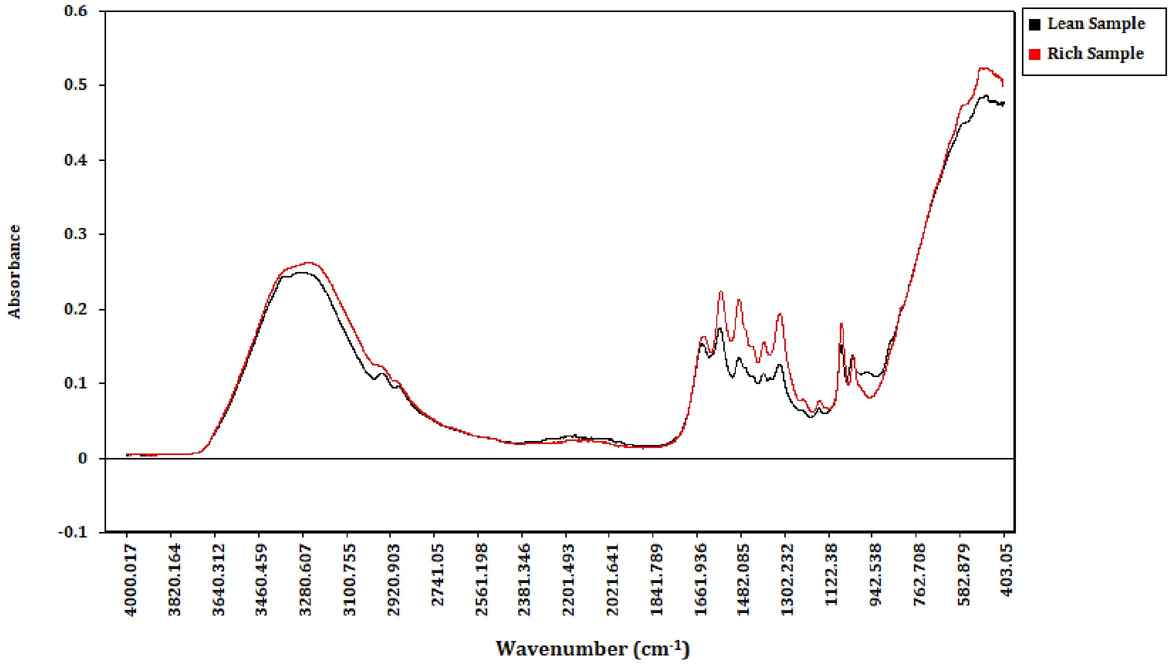


Figure 3-2 ATR-FTIR spectra for one lean (in black) and one rich (in red) sample, presenting the difference between the two

Information about the concentration of MEA and the CO_2-loading is vital when analyzing the CO_2 capture process [22]. The use of different approaches can be implemented to achieve this information, online process monitoring using ATR-FTIR is one of them.

The analysis of the total alkalinity (TOT_ALK) is used to find the concentration of MEA in the solution, while the total inorganic carbon (TIC) in the solution is used for determining the CO_2 concentration. The difference in density between the lean and rich samples says something about the amount of absorbed CO_2; the density is correlated to the CO_2-loading (mole CO_2 / mole MEA). Information about these compounds during the process will give a better understanding of the process, and also make it possible to monitor how much CO_2 is removed from the flue gas. The monitoring of the solution after it has passed through the desorber is vital in order to avoid degradation products, which will reduce the ability to capture CO_2 [2].

4 Multivariate Data Analysis

Multivariate data analysis is the use of different methods for analysis of data containing many variables [23]. In large datasets it can be difficult to extract the important information, spectroscopy is one example of where the datasets are massive. Chemometrics uses advanced mathematical and statistical methods for planning and optimizing processes and to extract relevant information from the data.

In the ensuing chapters the following notation is used:

Bold font and upper-case letter: **X** = matrix

Bold font and lower-case letter: **x** = vector

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

4.1 Decomposition of the Matrix

Data can be presented in a matrix, called the observation matrix [23]. This matrix contains all the information and can be decomposed into rows, containing I objects, and columns containing J variables. This makes up the matrix **X** with the size I × J.

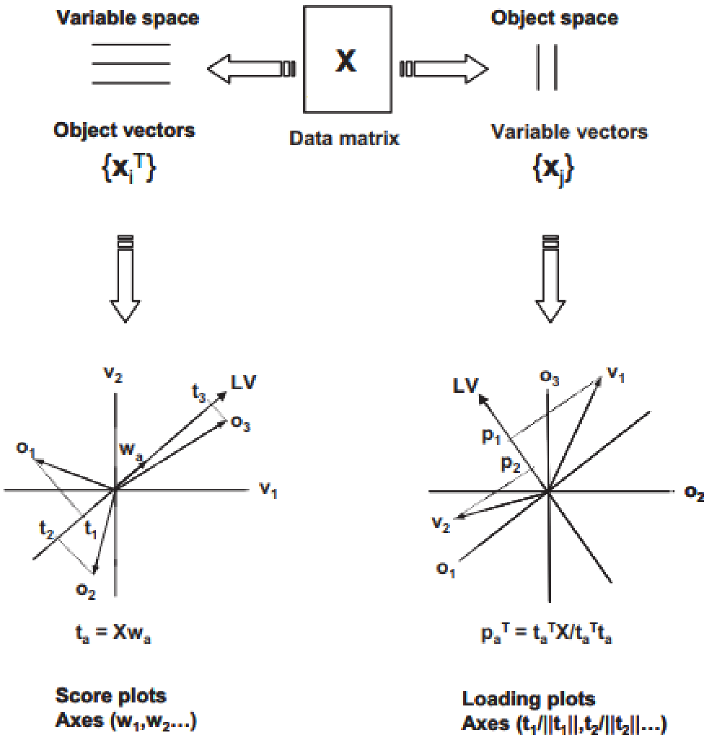


Figure 4-1 Decomposition of the original data matrix into variable and objects space [24]

This matrix can be looked at in two ways: either from the objects' point of view or the variables' [24]. By plotting all object vectors in a J dimensional space where the axes are the J variables we get the variable space. By plotting the variable vectors in an I dimensional space where the axes represent the I objects we get the object space. The two spaces are displayed in **figure 4-1**.

All the information in the dataset can be found in the object and variable space. The two spaces can be used to find correlations and similarities in the objects and variables; this can be done by studying the correlation between two objects m and n:

$$\cos \varphi = \frac{\mathbf{x}_{\mathrm{m}}^{\mathrm{T}} \times \mathbf{x}_{\mathrm{n}}}{\|\mathbf{x}_{\mathrm{m}}\| \times \|\mathbf{x}_{\mathrm{n}}\|}$$ **Equation 4-1**

If this angle equals 0°, the objects are perfectly positively correlated, if the angle is 90° the objects do not contain any overlapping information. If the angle is 180°, the objects are perfectly negatively correlated. The angle and distance between objects in this plot can be used to find information about correlation and similarity between objects. The same applies to variables. When the number of objects, and axes, exceeds two, this graphical display is no longer possible.

4.2 Preprocessing

Preprocessing of data is performed to remove effects that do not represent physical, chemical or biological aspects of the data [7]. Data coming straight from the instrument can contain noise, baseline differences, scattering or other features clouding the significant information.

When examining data using multivariate calibration models, preprocessing is crucial. When preprocessing is not performed the unwanted parts of the data, like noise, will be mixed with the important information [25].

Which preprocessing needed depends on the data, the instrument and which mathematic model is to be used in further examinations. For vibrational spectroscopy preprocessing can be divided into two groups: filtering methods and model-based methods. The filtering methods transform the spectra into a presumably better version, by for instance differentiation or normalization. The model-based methods allow for evaluation and separation of the physical and chemical variations in the spectra.

4.2.1 Baseline Correction

A spectrum can be expressed as a function of the concentrations, the pure spectra and the baseline [7]:

$$\mathbf{x}^{\mathrm{T}} = \sum \mathrm{y}_{\mathrm{k}} \mathbf{c}_{\mathrm{k}} + \mathbf{g}(\bar{\mathrm{v}}) \qquad \textbf{Equation 4-2}$$

where $\mathbf{x}^{\mathrm{T}}$ is the spectrum, $y_k$is the concentration for component k, $\mathbf{c}_{\mathrm{k}}$is the pure spectra for component k and $\mathbf{g}(\bar{\mathrm{v}})$ is the baseline.

A correct correction of the baseline will reduce the number of significant variables needed in the decomposition and will make the interpretation of data more accessible. The baseline can be expressed as:

$$\mathbf{g}(\bar{\mathrm{v}}) = \mathrm{b}_0 + \mathrm{b}_1\bar{\mathrm{v}} + \mathrm{b}_2\bar{\mathrm{v}}^2 + \cdots \qquad \textbf{Equation 4-3}$$

By differentiation, one can remove additive baselines, and by double differentiation, a sliding baseline can be removed if it is linear. By additive baseline, or offset, one refers to the spectra being move either up or down in relation to each other.

4.2.2 Normalization

When examining chemical data, one is often only interested in relative amount. In that case, normalization can be used to remove the effect of the total sample size, this a multiplicative correction [7]. Normalization transforms the data in such a manner that they can be compared, by for instance giving them the same size or length.

4.2.3 Differentiation and Smoothing

As mentioned before differentiation can be used to remove baselines, but it can also be used to smooth the data to reduce the noise [26]. Savitzky-Golay is a numerical method which performs both differentiation and smoothing of the data. Numerical differentiation refers to differentiation of each point of the data and can only be used on continuous data.

The method is based on the principle of least squares, which states: a set of points are to be fitted to a curve:

$$\mathrm{h}(\bar{\mathrm{v}})=\mathrm{b}_0+\mathrm{b}_1\bar{\mathrm{v}}+ \mathrm{b}_2\bar{\mathrm{v}}+ \mathrm{b}_3\bar{\mathrm{v}}^3 \qquad \textbf{Equation 4-4}$$

The coefficients, b, are to be chosen so that when the v̄s are put into the equation the square of the differences between calculated numbers h and the actual numbers are as small as possible.

When using Savitzky-Golay, a window size is chosen, with size from 5-25 using only odd numbers. The points in the window are then fitted to the curve in **equation 4-4**, and when the values for the coefficients are found the derivative of the midpoint of the window is calculated; this will be the point in the smoothed version of the data. The window is then moved one step at the time until it has covered all the points. This procedure will reduce the noise in the data with approximately the square root of the window size.

The plots will be more challenging to interpret after performing smoothing and differentiation, as can be seen in **figure 4-2**.

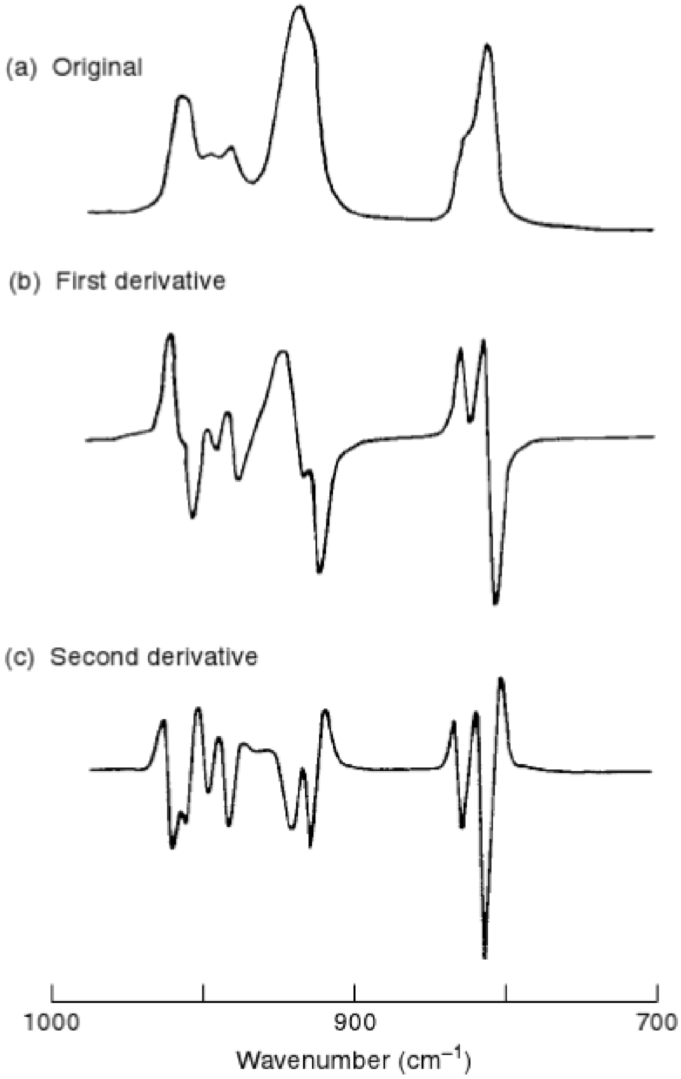


Figure 4-2 Display of spectra before and after differentiation, of first and second degree [14].

4.2.4 Multiplicative Signal Correction

Multiplicative signal correction (MSC) is an algorithm which can be used to remove additive and multiplicative contributions from an interfering signal [27]. The method can also be used to remove offsets and baseline effects and to normalize the data. When using FTIR spectroscopy to create spectra of a sample, one can get an undulating background. This background can be removed using extended multiplicative signal correction extended multiplicative signal correction (EMSC).

The basic idea of MSC has been extended to EMSC, which can be used for correction of, e.g., water vapor, sample thickness, CO_2, temperature and salt concentrations. EMSC has extended flexibility which takes into consideration more selective correction for different kinds of unwanted effects which cannot be corrected with other conventional preprocessing techniques. It is also a reliable tool for correction of additive baseline effects, multiplicative scaling effects, and interference [28].

Every spectrum in **X** can be written as:

$$\mathbf{x}_0 = \mathrm{b}_0 + \mathrm{b}_1\mathbf{x}_\mathrm{r} + \mathbf{e} \qquad \textbf{Equation 4-5}$$

Where $\boldsymbol{x}_r$ is a reference spectrum, usually the average spectrum. The coefficients are found by linear regression. The original spectrum is now corrected using these coefficients, giving the corrected spectrum $\mathbf{x}_\mathrm{c}$:

$$\mathbf{x}_\mathrm{c} = \frac{\mathrm{x}_0 - \mathrm{b}_0}{\mathrm{b}_1} \qquad \textbf{Equation 4-6}$$

this must be done individually for each spectrum.

In the extended version (EMSC) the reference spectrum is fitted to a polynomial:

$$\mathbf{x}_0 = \mathrm{b}_0 + \mathrm{b}_1\mathbf{x}_\mathrm{r} + \mathrm{b}_2\bar{\mathrm{v}} + \mathrm{b}_3\bar{\mathrm{v}}^2 + \mathbf{e} \qquad \textbf{Equation 4-7}$$

where $\bar{\nu}$ is the specific wavenumber.

The corrected spectrum is now given by:

$$\mathbf{x}_\mathrm{c} = \frac{\mathrm{x}_0 - \mathrm{b}_0 - \mathrm{b}_2\bar{\mathrm{v}} - \mathrm{b}_3\bar{\mathrm{v}}^2}{\mathrm{b}_1} \qquad \textbf{Equation 4-8}$$

4.3 Principal Component Analysis

In exploratory analysis, the primary goal is interpretation of the data [29]. The intention is to describe the system using as few variables (called principal components in PCA) as possible and make graphs that are easy to interpret but contain much information. Principal component analysis (PCA) is a useful tool for doing this.

The first principal component (PC1) is the linear combination of the original variables that explain most of the variation in the original observation matrix [30]. The columns (variables) of **X** can be denoted $\boldsymbol{x}_j$ (j=1, 2, ..., J) and are vectors in the I-dimensional space. These $\boldsymbol{x}$-variables can be written as a linear combination:

$$\mathbf{t} = \mathbf{w_1}\,\mathbf{x_1} + \cdots + \mathbf{w_j}\,\mathbf{x_j} \qquad \textbf{Equation 4-9}$$

t is a linear combination of the **x**-vectors and are called score vectors. **w** is the unit vector with elements, called the weights, which has the same direction as the **x**-vectors.

Since the goal of PCA is to create a model with fewer variables than the original matrix, the target is to find a vector **t** containing as much as possible of the variation relevant to the problem. When the first PC has been found, the information explained by this PC is removed from the observation matrix and PC2 can be extracted. PC1 and PC2 are orthogonal, meaning that the scalar product is zero and they do not contain any overlapping information.

After extracting PC2, the same procedure can be performed, and PC3 can be extracted. This process can be carried out until the total rank is equal to the rank of **X**, but since the goal is to reduce the number of variables, we stop before this. The number of necessary PCs to explain all the variation can be determined by using a simple rule of thumb: when a component explains less than $\frac{100\%}{J}$ it should not be included.

Principal component analysis on a data matrix can be expressed like this:

$$\mathbf{X} = \mathbf{TP^T} + \mathbf{E} = \hat{\mathbf{X}} + \mathbf{E} \qquad \textbf{Equation 4-10}$$

Where **T** is a (I × A) matrix containing orthogonal score vectors **t** and **P** is a (J × A) matrix containing the orthogonal loadings. **E** is the residual matrix, containing the information not explained by **T** and **P**. A is the total number of principal components. This decomposition of the PCA is displayed in **figure 4-3**.

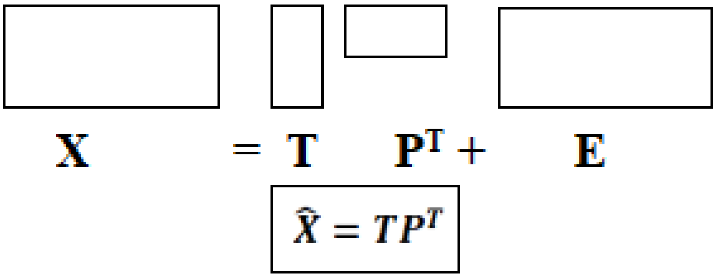


Figure 4-3 Decomposition of the matrix in PCA. The residual matrix and the model approximation has the same dimension as the original data matrix. Adapted from [30]

4.3.1 Scores and Loadings

When exploring the data, the terms scores and loading are very helpful. The scores can be found by projecting the objects in **X** on the PCs: $\mathbf{t}_j = \mathbf{X}\mathbf{w}_j$. For PCA, the loadings ($\mathbf{p}_j$) are equal to the weights ($\mathbf{w}_j$), meaning that the scores, loadings and weights are orthogonal. The scores have different lengths, which are proportional with how much of the variance in **X** they explain. Hence the score plot can be used to interpret how good a set of PCs are for the given data set [30]. The score plot is commonly made as a scatter plot, with the PCs on the axes.

When visualizing and interpreting the data, the score and loading plots are essential [6]. The scores can be plotted in many ways; the scatter plot is one. The scatter plot is made with the principal components on the axes. In a scatter plot of the scores, the distance and angle between objects can quickly be evaluated and based on the spread of the scores the importance of the principal component can be assessed. This plot can also be used to identify outliers in the data set (section 4.3.3) [30].

By plotting the scores and loadings in the same plot, the biplot is attained. The biplot can be used to examine the model as a whole and gives an overview of which variables explain which objects.

When determining if a PC is important, we look at the explained variance by the component. This is because the amount of information in a component is strongly associated with the amount of explained variance.

The importance of a component in PCA can be evaluated using the eigenvalues. The loadings are the eigenvectors of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, which is the cross-product matrix. The normalized scores are equal to the eigenvectors of the $\mathbf{XX}^{\mathbf{T}}$matrix. The matrix $\mathbf{T}^{\mathrm{T}}\mathbf{T}$ gives a diagonal matrix, with the eigenvalues of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ along the diagonal. The eigenvalues are proportional with the explained variance by the PC. The bigger the eigenvalues, the bigger the importance of the PC.

4.3.2 Residual Standard Deviation

The decomposition of the data matrix gives scores, loadings and a residual matrix. The residual matrix contains the noise. The residuals can be found by extracting the part of the data explained by the principal components from the original data [30]:

$$\mathbf{e}_{\mathrm{i}}^{\mathrm{T}} = \mathbf{x}_{\mathrm{i}}^{\mathrm{T}} - \textstyle\sum_{\mathrm{a}=1}^{\mathrm{A}} \mathbf{t}_{\mathrm{ia}}\mathbf{p}_{\mathrm{ia}}^{\mathrm{T}} \qquad \textbf{Equation 4-11}$$

Where $\boldsymbol{e}_i$ is the residual vector for sample i after being fitted to the model. Residual standard deviation (RSD) is a measure of how good a model is, and can be calculated in the following way [31]:

$$\mathrm{RSD} = \sqrt{\frac{\mathbf{e}_{\mathrm{i}}^{\mathrm{T}}\mathbf{e}_{\mathrm{i}}}{\mathrm{I-A}}} \qquad \textbf{Equation 4-12}$$

Where I is the total number of variables, and A is the total number of principal components. Residual standard deviation gives a measure of the accuracy of the variable in consideration to the model, meaning how far from the model the variable is.

4.3.3 Outliers

Outliers are samples that for some reason behave differently than the other samples [30]. There can be different reasons for this; the sample can for example simply be wrong or mismeasured. The outliers can disturb further investigations and should be corrected or removed. Outlier detection is an essential part of the multivariate analysis. The outliers can be detected by using, for instance, the score plot or RSD vs. leverage plot [31].

Leverage is calculated for each sample and is a measure of how much influence an object has on the model. Leverage is calculated by the following equation [29]:

$$h_{iA} = \frac{1}{I} + \sum_{a=1}^{A} \frac{t_{ia}^2}{\lambda_a}$$ **Equation 4-13**

Where t_{ia} is the score value for component i for principal component a, and λ_a is the corresponding eigenvalue. The leverage is always between zero and one, where a high value means that the sample is far from the average sample and thus a possible outlier. By studying the RSD vs. Leverage plot, one can examine how well the principal components describes the system [31]. Small outliers will have low values for leverage and high values for RSD. High RSD values means that the sample is different from the modeled samples. High values for both leverage and RSD arguments for the sample being an outlier, and it should be removed. From the RSD vs. Leverage plot in **figure 4-4** samples 31 and 32 can be identified as outliers. 23 is identified as a possible small outlier and should be examined further.

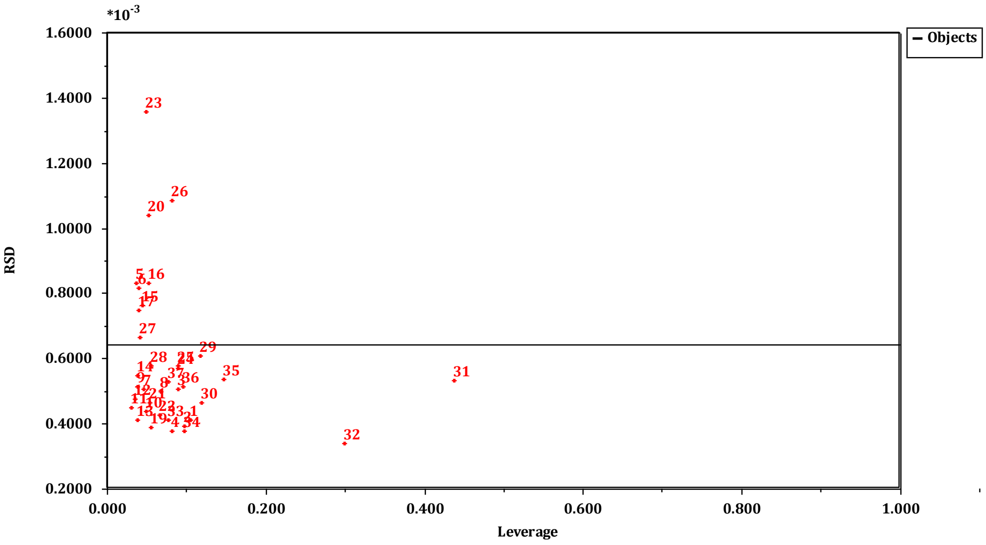


Figure 4-4 RSD vs. Leverage plot used to detect outliers

When performing outlier detection, one should not only use the RSD vs. Leverage plot, the RSD limit for rejecting samples may be to narrow [31]. This can occur when the number of variables is large, the collinearity between the variables is strong (which is the case for spectral data) or replication of samples in the calibration set [33].

These effects will exaggerate the degrees of freedom, resulting in to narrow RSD limits. The narrow limits lead to samples being classified as outliers, even though they are not.

In the score plot, one will usually see a cluster of the samples, and the outliers will lie outside of this collection [30]. Before removing the outlier, it is essential to check if this is, in fact, an outlier. This can be done by examining the original data set and the preprocessed data set, the preprocessing may correct the outlying nature of the sample, so this is important before removing samples identified as possible outliers. If the sample that has been detected as an outlier has a considerable influence on the data, the sample should be removed so that the model is not affected by it; this may lead to the wrong model and poor prediction.

From the score plot in **figure 4-5** two objects are identified as outliers, 21970 and 21972. After these outliers are removed a new score plot should be made to investigate if more outliers appear now that the two objects with large deviation, and thus considerable influence, on the model, is removed.

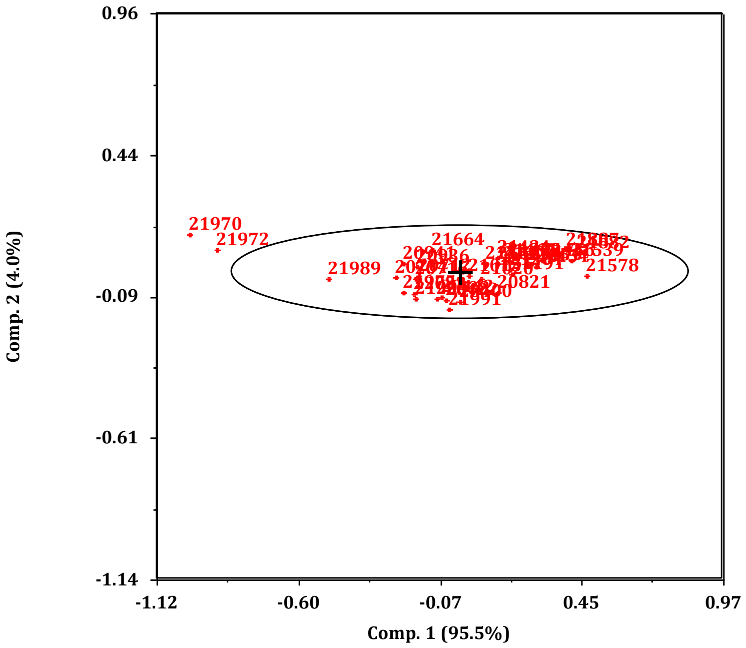


Figure 4-5 Score plot of component 1 versus component 2 after performing PCA. The ellipse is provided by Sirius 11.0, and objects outside this are identified as outliers

A normal plot function of the scores can be used to indicate outlying objects, as these will deviate from the straight line expected in such a plot. **Figure 4-6** presents such a normal plot, and objects 31, 32 and 34 deviate from the straight line and are therefore identified as outliers.

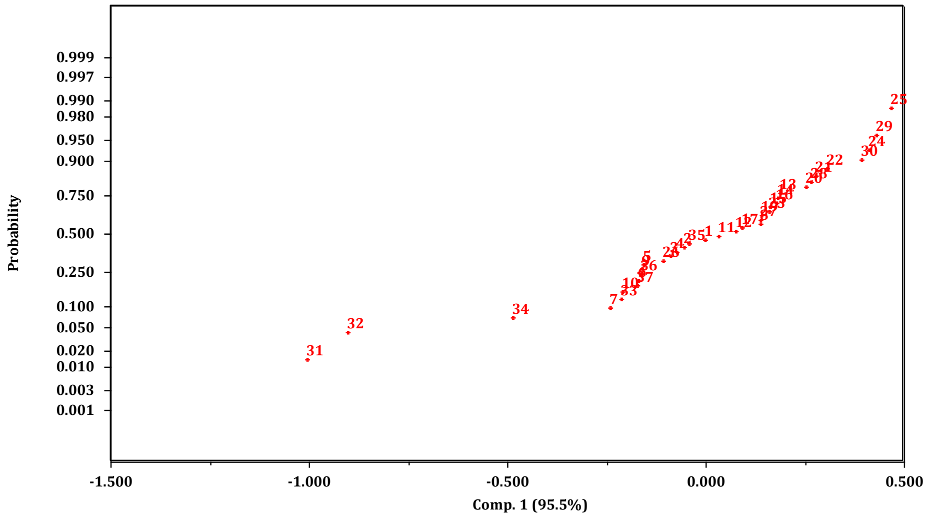


Figure 4-6 Scores vs. Objects normal plot for the first component, describing most of the variation (95.5%) in the data

4.3.4 Multiple Linear Regression

The relationship between a set of *x*-variables and one or more response variables *y* can be determined using predictive modeling, e.g., by using a model where the *y*-variables are described by the *x*-variables and the noise is left in the residuals [24]:

$$y = f(x_1, x_2, \dots, x_J) + e_y \qquad \textbf{Equation 4-14}$$

The function can be explained using a linear polynomial:

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_Jx_J + e_y \qquad \textbf{Equation 4-15}$$

Where b_j(j=0, 1, 2, ..., J) are the regression coefficients which describe the effect of the corresponding term. e_y is the residual in y. This can also be expressed in matrix form:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}_y \qquad \textbf{Equation 4-16}$$

Multiple linear regression can now be used to calculate the regression coefficients from the following equation:

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{X}^{+}\mathbf{y}$$ **Equation 4-17**

4.4 Partial Least Squares

Partial least squares (PLS) is a regression approach with the aim to find a model that shows that one or more response variables can be explained by a set of predictor variables [24].

PLS deals with both **X** and **Y** data, where **X** is the raw data, in the case of spectroscopy it contains the spectra, and **Y** contains the responses. The data in the matrices are related to each other, and the objects are presented in both **X** and **Y** space. The PLS model then gives the relation between the two, based on the covariance.

While PCA is based on the criterion of maximum variance for the decomposition step, PLS uses another criterion. PLS expresses the response as a function of a given set of variables based on the collinearity between them by calculating a normalized weight vector based on the response **y** and the data matrix **X**:

$$\mathbf{w}_{\mathrm{PLS},1}^{\mathrm{T}} = \frac{\mathbf{y}^{\mathrm{T}}\mathbf{X}}{\|\mathbf{y}^{\mathrm{T}}\mathbf{X}\|}$$ **Equation 4-18**

The scores for the PLS components are calculated successively by projecting the variables **X** on $\mathbf{w}_{\mathrm{PLS},1}$, the loadings are found by projecting **X** on the resulting scores. The predictive power of each component is checked using cross-validation (section 4.4.2). After the first PLS component has been found, the part of **X** explained by this is removed and the process is carried out again, calculating the second PLS component.

4.4.1 Model Validation

Root mean square error of prediction (RMSEP) can be used for assessing model quality. The value can be found with the following equation [33]:

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{I}}$$ **Equation 4-19**

Where $\widehat{y_1}$ is the predicted values and y_i is the original values, I is the total number of samples. A model is concluded to be good if the RMSEP value is low compared to the measured values y_i.

4.4.2 Cross-Validation

Cross-validation can be used to determine the number of significant components to include in the model and says something about the predictive ability of the model [33]. The method leaves out one or more samples, and a PLS model with 1, 2, …, A components are made calculating the predictive ability for the left-out samples. The method is then repeated leaving out another set of samples, the predicted residual error sum of squares (PRESS) is calculated, and the lowest value of PRESS establishes how many components to include in the model. PRESS can be calculated with the following formula [34]:

$$\mathrm{PRESS_a} = \sum_{i=1}^{I} \frac{(y_i - (\hat{y}_{(i),a})_i)^2}{I} \qquad \textbf{Equation 4-20}$$

where y_i is the ith element of **y** and $\hat{y}_{(i),a}$ is the estimate of **y** from PLS with a components when the ith observation has been eliminated.

From this the root mean squared error of cross-validation (RMSECV) can be calculated with:

$$\mathrm{RMSECV_a} = \sqrt{\frac{\mathrm{PRESS_a}}{I}} \qquad \textbf{Equation 4-21}$$

The RMSECV value can be evaluated in the same way as the RMSEP, a low value compared to the measured values indicate that the model is going to perform well. The RMSECV values are calculated for each component a, and by plotting these the RMSECV plot is obtained. This plot can be used for determination of how many components to include in the model, when the value decreases by including component, the component should be included. When the RMSECV value ceases to decrease by including it, the component should not be included.

4.4.3 Double Cross-Validation

Double cross-validation separates the data into two nested loops, one outer loop and one inner loop [35]. The objects in the outer loop are split randomly into a calibration set and a test set which is then used to estimate the prediction performance of the model made from the calibration set on the test set. The inner loop consists of a calibration set like the outer loop, and cross-validation is used to find the number of components to include in the PLS model.

In a repeated double cross-validation this procedure is repeated many times, classically 100 times, in a repetition loop with different random splits of the objects every time.

The available predicted y-values and test sets are increased using double cross-validation, and thus the prediction ability of the model can be better evaluated.

The double cross-validation gives the RMSEP value for each component and an overall value for the model with the optimal number of components. Double cross-validation is a good way to evaluate models when there all the available samples have been used to build the model, and no validation set is accessible.

4.4.4 Coefficient of Multiple Determination

The coefficient of multiple determination can be used to determine how good a model is. It determines the proportion of explained variation by the model [36, p. 686]. The coefficient is given by:

$$\mathrm{R}^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$ **Equation 4-22**

SSE is a measure of how much variation in y is left unexplained by the model. SST is the total sum of squares, capturing the sum of squares about the horizontal line. SSE and SST can be calculated with the following formulas [36, p. 631-633]:

$$\mathrm{SSE} = \sum(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2$$ **Equation 4-23**

$$\mathrm{SST} = \sum(\mathrm{y_i} - \bar{\mathrm{y}})^2$$ **Equation 4-24**

where $\hat{y}_i$ is the predicted value for y_i and $\bar{y}$ is the average of all y_i.

The value of R^2 is between zero and one, where a value of one means that the fitted model explains all the observed variation [36, p. 686]. The value of R^2 can be inflated by including to many components in the model, this will give a high value even though the model is not necessarily very good.

To avoid this problem the adjusted coefficient of multiple determination (R_a^2) can be used in addition to R^2. The adjusted coefficient considers that its value may be high just because the number of predictors is high relative to the amount of data. R_a^2 will decrease if the number of predictors included in the model is large relative to the amount of data. R_a^2 is given by:

$$\mathrm{R}_\mathrm{a}^2 = 1 - \frac{\mathrm{I}-1}{\mathrm{I}-(\mathrm{A}+1)}\frac{\mathrm{SSE}}{\mathrm{SST}}$$ **Equation 4-25**

The value of R_a^2 can be equal to R^2, but never higher. A similar value of the two coefficients points to the model being good, while a big difference between the two is a red flag and the chosen model probably has too many predictors relative to the amount of data.

4.4.5 Variable Selection

When predicting a response from a data set, multivariate calibration models are commonly applied [37]. These models can handle large data sets where the number of variables exceeds the number of samples. Even though the models can handle this, it can be an advantage to reduce the number of variables to make interpretation simpler and the predictions better. Variable selection is a method that can be used to do this, and it can also improve the statistical properties of the data. Variable selection can also be an advantage for computational reasons.

Variable selection is very sensitive to outliers because it is based on assessing minor differences in the model. Variable selection is an iterative method, meaning the analysis is performed stepwise until satisfactory results are obtained.

Variable selection can remove variables with little variation or variables that are similar. When there are variables that are similar the removal of some of these is a simple way to reduce the total number of variables.

After performing the variable selection, the predictions will in most cases not improve, but since the number of variables is reduced the further analysis using multivariate calibration models will be more straightforward and give better results. Variable selection can be affected by the application of preprocessing, so preprocessing should always be performed before the variable selection to avoid this.

After variable selection, a new model will be built with the variables chosen. This model is then compared to the model from before variable selection, to check if the variable selection improved the model. In the case where the model does not improve, one should go back and check if it was due to bad choices of variables to remove.

4.4.5.1 Variable Importance for Projection

Variable importance for projection (VIP) gives a measure of how much a variable contributes to describing two sets of data: the dependent and the independent variables [38]. The VIP values are given as follows:

$$\mathrm{VIP_j}=\sqrt{\frac{\sum_{\mathrm{a=1}}^{\mathrm{A}} \mathrm{W_{ja}^2\ SSY_a{\cdot}J}}{\mathrm{SSY_{total}{\cdot}A}}} \qquad \textbf{Equation 4-26}$$

This gives the VIP value for variable j, $\mathrm{W_{ja}}$ is the weight for variable j and component a. SSY_a is the sum of squares of explained variance for component a and J is the total number of variables. SSY_{total} is the total sum of squares of explained variance and A is the total number of components.

Given a one-dimensional **Y** space, **y**, the total sum of squares is given as:

$$\mathrm{SSY_a} = \mathbf{b}_\mathrm{a}^2 \mathbf{t}_\mathrm{a}^\mathrm{T} \mathbf{t}_\mathrm{a} \qquad \textbf{Equation 4-27}$$

$$\mathrm{SSY_{total}} = \mathbf{b}^2 \mathbf{T}^\mathrm{T} \mathbf{T} \qquad \textbf{Equation 4-28}$$

where **b** is the vector of coefficients from the PLS inner relation, **T** is the scores matrix for **X**.

The weights from PLS reflects on the covariance between two variables, in this case, the dependent and independent [37]. By using these weights in the calculation of VIP, the importance of the information in respects to the modeling of the dependent variables can be assessed.

As a rule of thumb, a VIP value below one indicates a non-important variable. Variables with a VIP less than one should however not just be removed, the data set must be examined to check whether removing variables based on this criterion is a good idea. Usually, one should set a lower value for VIP, and start by removing only these, and then asses the model quality. This method can be repeated until the model is satisfying.

4.4.5.2 Selectivity Ratio

Another technique of variable selection that can be used is selectivity ratio (SR), which is the ratio between the explained variance of each variable and residual variance. A high SR-value indicates that the variable has good predictive performance.

The SR-value can be derived with the following formula [39]:

$$\mathrm{SR_i} = \frac{\mathrm{v_{explained,i}}}{\mathrm{v_{residual,i}}}\ (\mathrm{i} = 1,2,3 \dots) \qquad \textbf{Equation 4-29}$$

where $\mathrm{v_{explained,i}}$ is the explained variance for variable i and $\mathrm{v_{residual,i}}$ is the residual variance for the same variable i.

Since both SR and VIP are calculated individually, the values can be represented in the same way as the spectra. The SR-plot can be used to identify the most important regions in the spectra, meaning the regions with the highest SR-values [40]. A high SR-value indicates that there is a strong correlation between the given variable and the dependent variable.

5 Method

5.1 Software

The multivariate data analysis is done in Sirius version 11.0 (Pattern Recognition Systems AS, Bergen, Norway).

MATLAB R2017b (The MathWorks, Natick, Massachusetts, USA) is used for sorting the samples.

OMNIC 9.8 Spectra Software (Thermo Scientific, Waltham, Massachusetts, USA) is used to examine the spectra during the measurements, and transformation of the spectra after the measurements are done.

5.2 Measurements

The infrared measurements in this thesis were done with a Nicolet iS 50 FTIR Spectrometer with an ATR diamond. The spectral range for the spectrometer is 15 – 27.000 cm^{-1} and a spectral resolution better than 0.09 cm^{-1}. The measurements were completed in the laboratory for spectrometry at the Department of Chemistry, University of Bergen.

The samples are provided by TCM, 279 samples consisting of lean and rich samples. Differences in the individual samples can be observed based on color-differences, the color ranges from light yellow to dark brown. The darker samples contain more degradation products than the ones with a lighter shade. Since the instrument used to measure the absorption of the samples was ATR-FTIR no preprocessing of the samples was necessary.

The measurements are made in the range from 400 to 4000 cm^{-1}, with 32 scans and a collection length of 47 seconds. The samples are collected with format %Reflectance, which is transformed to log(1/R) (absorbance) using OMNIC Spectra Software.

The measurements are done on the samples straight out of the fridge, a drop of the sample is placed on the crystal, and the measurements are made. Only a small number of samples, about 10, were removed from the fridge at a time.

The temperature was not monitored during the measurements, and no replicated measurements are made to examine how different temperatures of the samples would affect the results. Between every sample, the area is cleaned with water and dried, the background spectra are then checked and confirmed to be equal before starting the measurements on a new sample. Every 30. minutes a new background spectrum is taken.

Response variable for the samples, in the form of measurements of different characteristics of the samples, was provided by TMC, and the measurements were performed by them. The responses were the organized and matched with the belonging samples, which showed that not all the responses had enough measurements to be used in the modeling. Only tree responses are modeled in this thesis; total inorganic carbon (TIC), total alkalinity (TOT_ALK) and density. These responses are explained in detail in chapter 3.2 and the modeling is described the ensuing sections.

5.3 Multivariate Data Analysis

The multivariate data analysis is performed in Sirius. The data with the chosen response variable is imported into Sirius and analysis is carried out. The analysis is done for lean and rich separately, and with one response variable at the time.

TCM provided responses for the samples and using an in-house MATLAB code the samples are sorted, and the responses are organized with the correct sample. The samples are also sorted into lean and rich samples, depending on where in the process the sample is collected (chapter 3.2).

5.3.1 Building Models and Predicting Responses

The first step of multivariate data analysis is exploratory analysis and outlier detection, using PCA. A model is built including all the objects and variables in the dataset except the response variable. The outliers are now detected using the score plot, RSD vs. Leverage and normal plot of the scores. If the score plot indicated one or more objects being large outliers, i.e., showing large deviations from the rest of the objects, they should be removed first. After removing the most substantial outliers, another model is made to identify smaller outliers.

In the exploratory analysis, the raw spectra of the samples are investigated before and after outlier detection, making sure that there are no more deviating samples.

If the spectra show a spread after the outlier detection, preprocessing is needed before building a model. After the preprocessing is done, outlier detection is performed as described over, and the number of outliers is compared. The number of outliers can be reduced after preprocessing because the preprocessing can correct the outlying behavior of the object.

After the outliers are detected, and if the data contains enough samples, the data are split into two sets. One training set and one validation set, the training set is used to build the model, and the validation set is used to test the prediction abilities of the model. When the dataset does not contain enough data to make two separate sets, the model is built using all the data and is validated based on the information from the regression analysis and double cross-validation. If and when more samples become available, the model can be validated using these.

The models are built using PLS regression analysis. For each model ten components are extracted, with the given response as the dependent variable, using 100 iterations for the cross-validation. The model is then analyzed to find the optimal number of components to include in the model. This analysis is done based on the model dimensions plot from Sirius, which shows the RMSECV values for each component. A component should be included if the RMSECV value for the model decreases by including it. The cross-validation value is also important when determining how many components to include in the model; the value should be below one.

After determining how many components to use, the model performance can be investigated. This is done using the Predicted vs. Measured plot, which show how well the model predicts the measured values for the response, this plot should give a close to linear plot with little spread. This plot gives the RMSECV value for the model.

The response residuals are also used when investigating the model; this plot shows the residuals of the responses. Using the normal plot function, this plot should be a straight line passing through y = 0.5, which means that the data are normally distributed, and the model is good.

The normal plot of the scores is investigated, and it should display a straight line. If this line is not straight, but show, e.g. polynomial behavior, another preprocessing technique should be investigated. There should not by any objects lying away from the straight line, these will be identified as outliers.

If the plots mentioned above point to the model performing well, the objects in the validation set can be fitted to the model. When fitting the objects, the identified outliers are omitted, these are identified as objects not behaving like the rest of the data, and the model is not expected to be able to predict these accurately.

How well the model performs is now evaluated using the same plots as used for investigating the model in the regression analysis, as described above. When fitting the objects in the validation set to the model, the Predicted vs. Measured plot is the most interesting. This shows how well the model predicts the objects that were not included when building the model. The plot also gives the RMSEP value, which should be similar to the RMSECV value, and small compared to the measured value, to conclude that the model is good.

Sirius provides a report, giving the predicted values and the prediction error for each object, if the results are satisfying the model is concluded to be good and able to predict the given response. If the results are not satisfying, one can go back and try, e.g. variable selection to see if this can improve the model. If the variable selection does not improve the model, one can also try and narrow the window of wavelengths used; the window should be chosen based on where the compounds connected to the response variable has the highest intensity of absorption.

A model is built for each preprocessing technique, and then all the models are compared to find the model with the best predictive abilities. The conclusion is made based on the number of outliers and components, the cross-validation values, the coefficient of multiple determination and the RMSEP values.

6 Results and Discussion

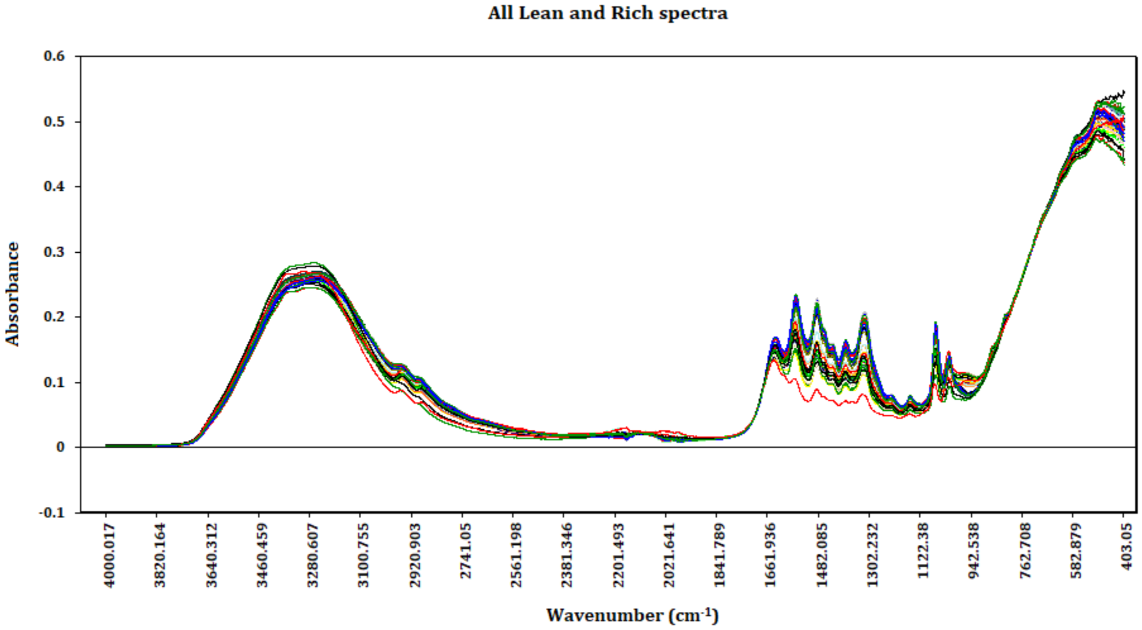


Figure 6-1 All lean and rich samples

Figure 6-1 shows a plot of all the spectra from all the samples. The most substantial differences in the samples are observed in the fingerprint region from 1700 to 800 cm^{-1}. The broad peak to the left is due to the content of water in the samples and is O-H stretching, which has a strong and broad absorption in the region 3400 to 3200 cm^{-1} [13].

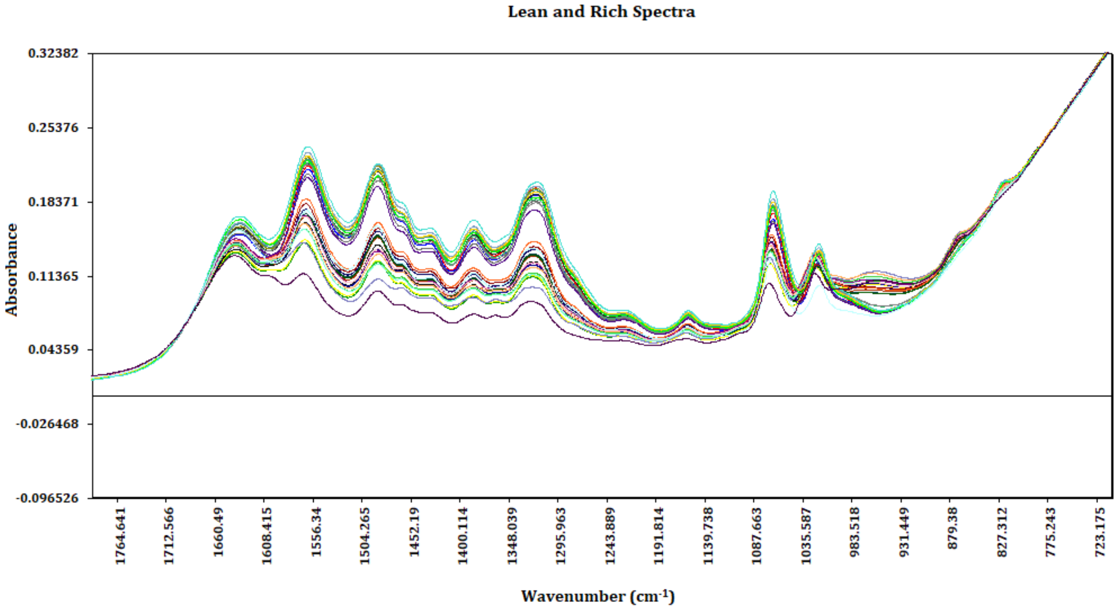


Figure 6-2 Fragment of lean and rich spectra

The close up of the fingerprint region in **figure 6-2** displays the most complex part of the spectra, and where most variation is found. Here, a clear difference between the lean and reach samples can be observed, especially in the region enclosing 950 cm^{-1}. This is where the C-NH_2 absorption appear, which is MEA [13]. The lean samples have a higher intensity in this range because it contains more pure MEA than the rich samples (see section 3.2).

The rich samples are collected after the solution has passed through the absorber, and thus contain CO_2. The inorganic carbons absorb radiation in the region from 1600 to 1300 cm^{-1}, leading to higher intensity for the rich samples in this range (see section 2.3.3).

6.1 Lean Samples

In the following description, the models presented for each response is the best one. Several approaches were tried out and investigated as well, but these models are not presented here. The models are made using a small window of the spectrum, the part of the spectrum showing most variation and containing most of the information about the samples. The window chosen varies for each response variable, depending on what the response describes.

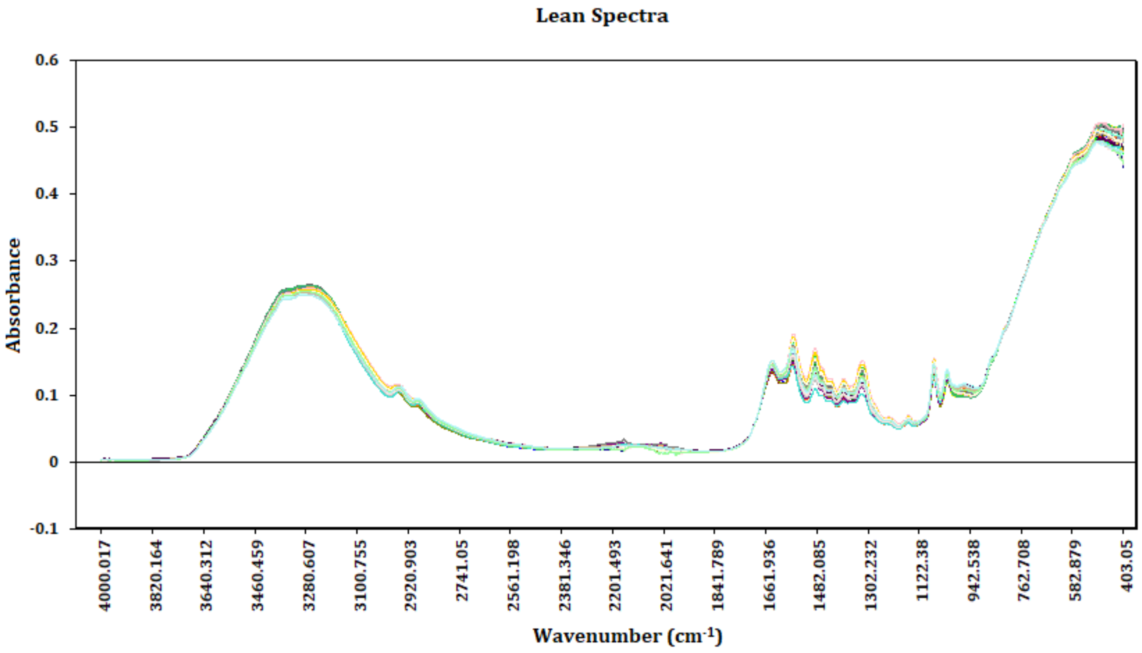


Figure 6-3 Spectrum of all lean samples

From the spectrum (**figure 6-3**), it is observed that the region from 2500 to 1800 cm^{-1} and 700 to 400 cm^{-1} contain noise. These regions were excluded when building all the models. The preprocessing methods used requires continuous data; the preprocessing is therefore performed before removing these regions.

6.1.1 Total Inorganic Carbon

Total Inorganic Carbon (TIC) is be used to determine the concentration of CO_2 and is, therefore, an interesting response to explore, and a useful variable to be able to predict. The dataset contains 70 samples with responses, with an average TIC value of 1.23 mole/kg.

The spectral region used when building a model for this response where chosen based on where the inorganic carbons have high intensity in the spectrum, the selected window is from 1670 to 1000 cm^{-1} [13].

The raw spectra of all the samples in **figure 6-4** show a spread, especially in the largest wavenumbers. This indicates that some sort of preprocessing is needed to obtain a good model.

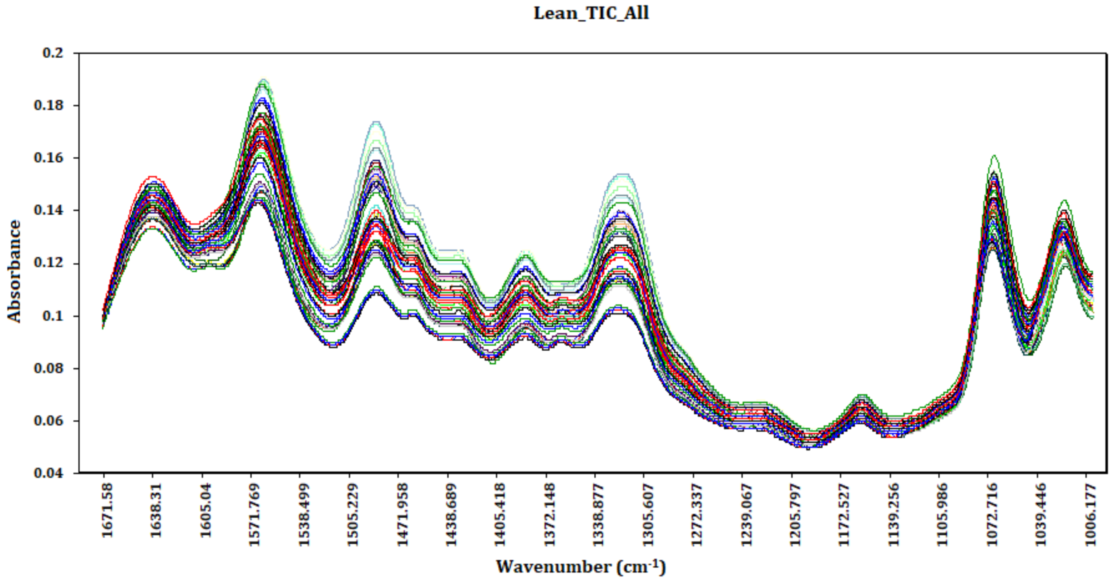
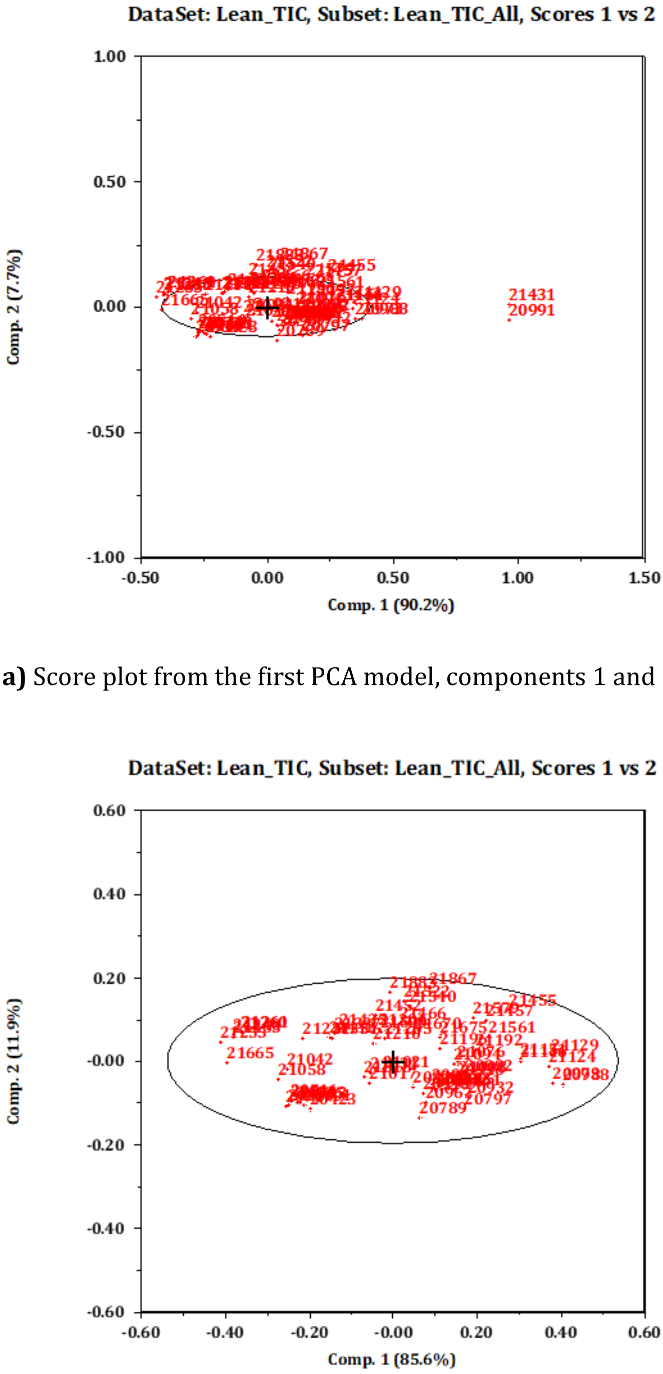


Figure 6-4 Fragment of raw spectra for the lean samples used to build a model with TIC as response variable

PCA has successfully been employed for identifying outliers; the model is built based on the entire dataset, excluding the response variable. The results are given below.



a) Score plot from the first PCA model, components 1 and 2

b) Score plot from the second PCA model, after removal of extreme outliers

Figure 6-5 Score plots using component one and two from the two PCA models, used for outlier detection. The ellipse is provided by Sirius, and objects outside this are identified as outliers

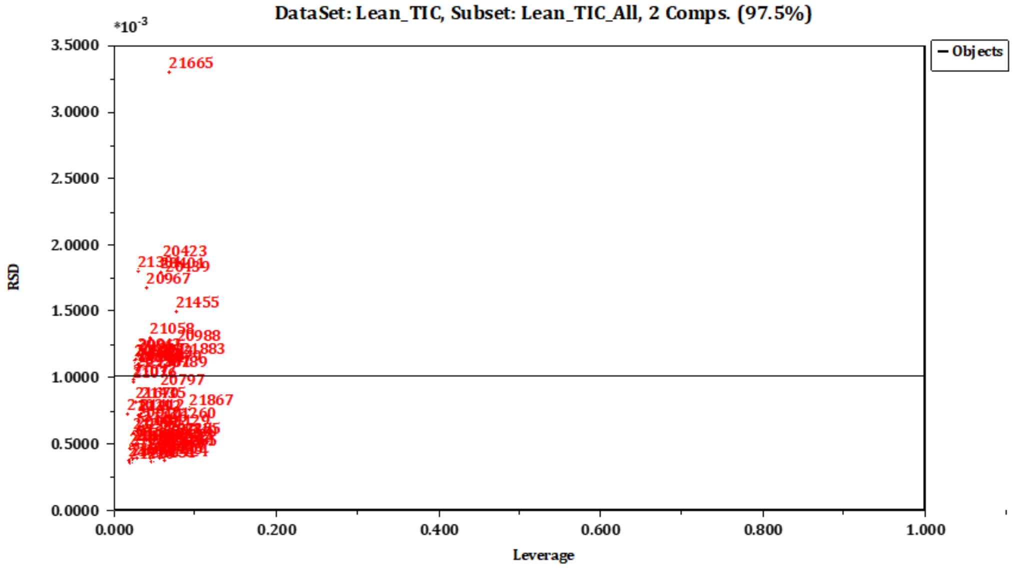


Figure 6-6 RSD vs. Leverage with two components obtained from PCA, used for outlier detection. This plot is made after the most extreme outliers are removed, to identify smaller outliers

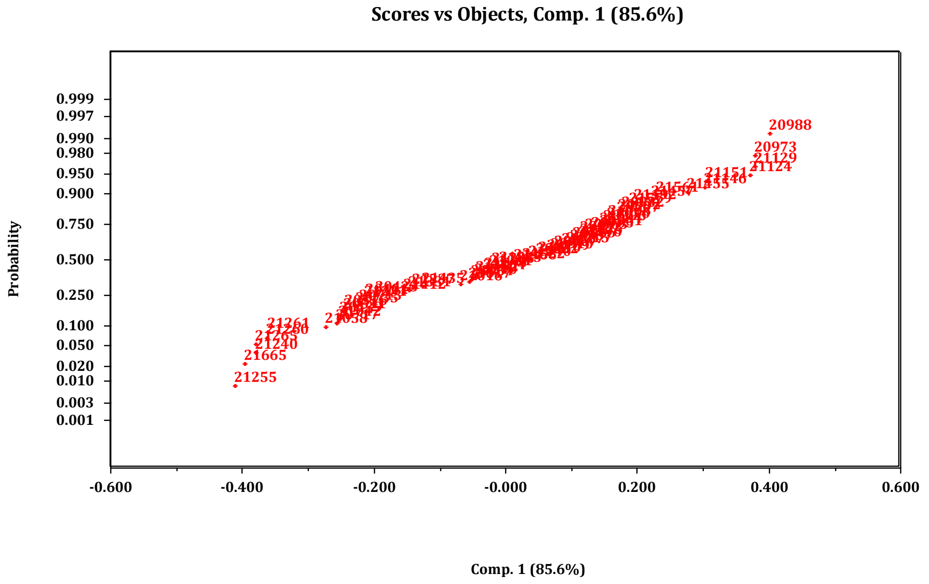


Figure 6-7 Normal plot of the scores for component one, used for outlier detection. Objects deviating from the straight line are identified as outliers. This plot is made after the most extreme outliers are removed

The score plot in **figure 6-5 a** shows two samples with a significant deviation from the others (object 21431 and 20991), meaning that they have a big influence on the model.

The two outliers are removed, and a new model is made to identify smaller outliers. Score plot from the second PCA, displayed in **figure 6-5 b**, shows no outliers.

The RSD vs. Leverage plot in **figure 6-6** is from the model after the two extreme outliers are removed, and shows one outlier, sample 21665, having high values for RSD and low values for leverage, indicating it is a small outlier. The normal plot of the scores for component 1 in **figure 6-7** shows some objects deviating from the straight line; these are outliers. In total thirteen outliers has been identified and are presented in **table 6-1**.

Table 6-1 Outliers identified using PCA

Identified outliers
20797, 20988, 20991, 21042, 21129, 21240, 21255, 21260, 21261, 21265, 21431, 21455, 21665

After the outlier detection, the dataset is divided into a training set and a validation set. The models are built based on the training set, and the validation set is used to validate the model. The outliers are not included in either data sets; the model is not expected to model these since they deviate from the rest of the objects.

Two of the preprocessing techniques used both gave satisfactory results: second order EMSC and Savitzky-Golay with second order differentiation, a window size of 25 and a third-degree polynomial. Based on the number of outliers, cross-validation values, and prediction abilities it was concluded that the use of EMSC gave the best results, the validation parameters for both models are presented in **table 6-2**.

When performing Savitzky-Golay as preprocessing the outlier detection has to be completed after the preprocessing, the outlying nature of the objects may be corrected by the preprocessing. The outlier detection after preprocessing resulted in more outliers than before, which supports the conclusion of EMSC being the best model. These outliers are not presented here.

The results for the model after EMSC preprocessing is presented here, the results for the model based on Savitzky-Golay preprocessing is attached in Appendix B1.

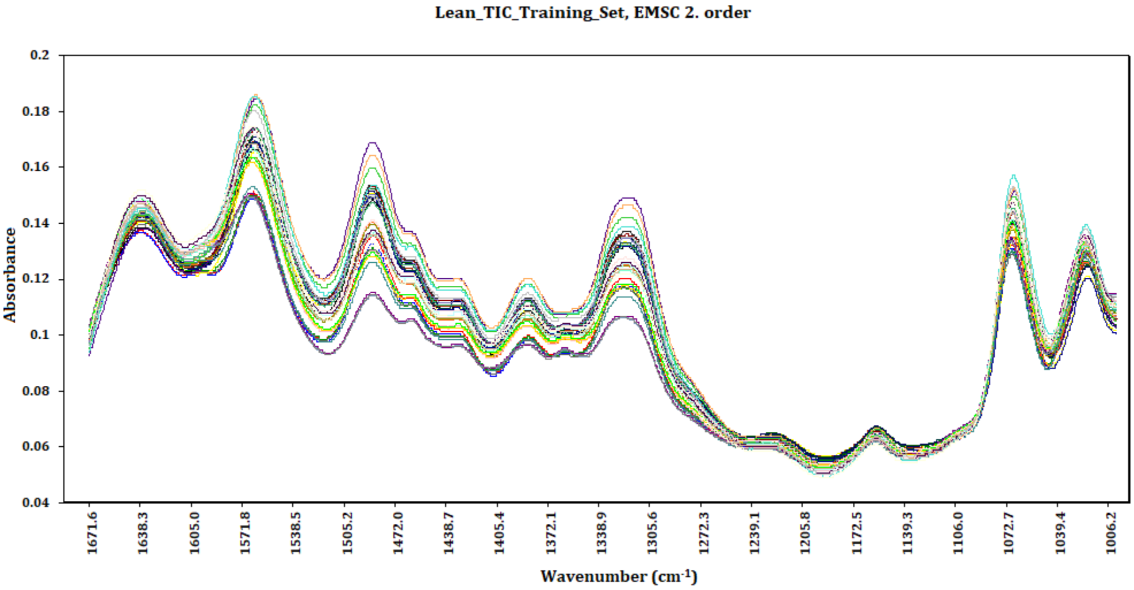


Figure 6-8 Spectra for the chosen region of wavenumbers after performing second-order EMSC as preprocessing

From the plot in **figure 6-8**, it can be observed that the spread of the spectra has decreased after preprocessing with second-order EMSC. EMSC is a good technique for removing these physical effects in the samples. The plot displays the training set after removal of the outliers.

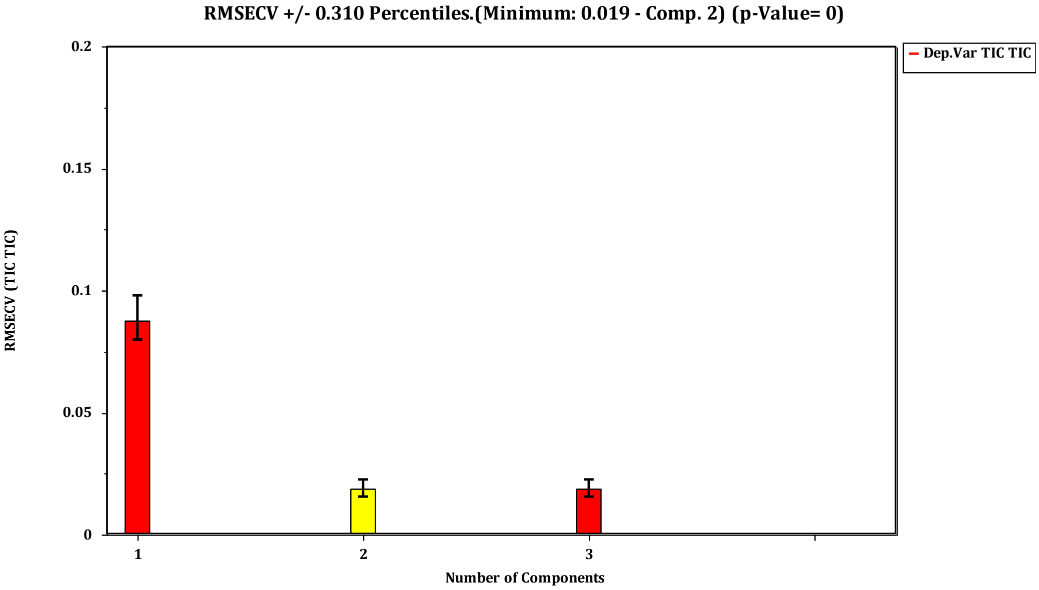


Figure 6-9 RMSECV-plot for the first three components in the PLS model. Used to determine the number of components to include in the model, the yellow bar indicates that two components should be included

The model dimensions plot (**figure 6-9**) presents the RMSECV values for the first three components extracted and suggests that two components should be included in the model. The weighted regression coefficients in **figure 6-10** show the RMSECV values for the two components, and also display that the components describe the data and not noise. The Predicted vs. Measured plot from the regression analysis showed good linearity, and a promising RMSECV value when two components are included, this plot is not presented here. The standard deviation of the cross-validation values (CsvSD), displayed in **table 6-2**, indicate that the two components should be included, and further analysis is carried out using two components.

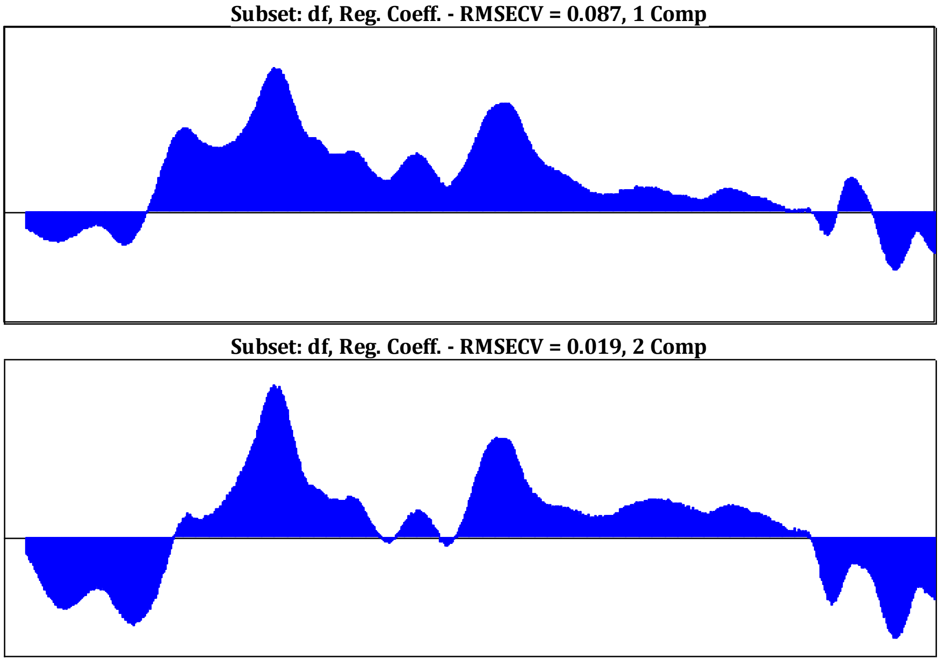


Figure 6-10 Weighted regression coefficients for the two components included in the PLS model

The model dimensions plot for the Savitzky-Golay model suggested that three components should be included, and the weighted regression coefficients showed that the components described mostly the data. These plots are can be found in Appendix B1. The RMSECV value for the last component included in this model was 0.016, which is lower than for the EMSC model. However, the cross-validation value for the last component was much higher, at a value of 0.78 in comparison to 0.20 for the EMSC model.

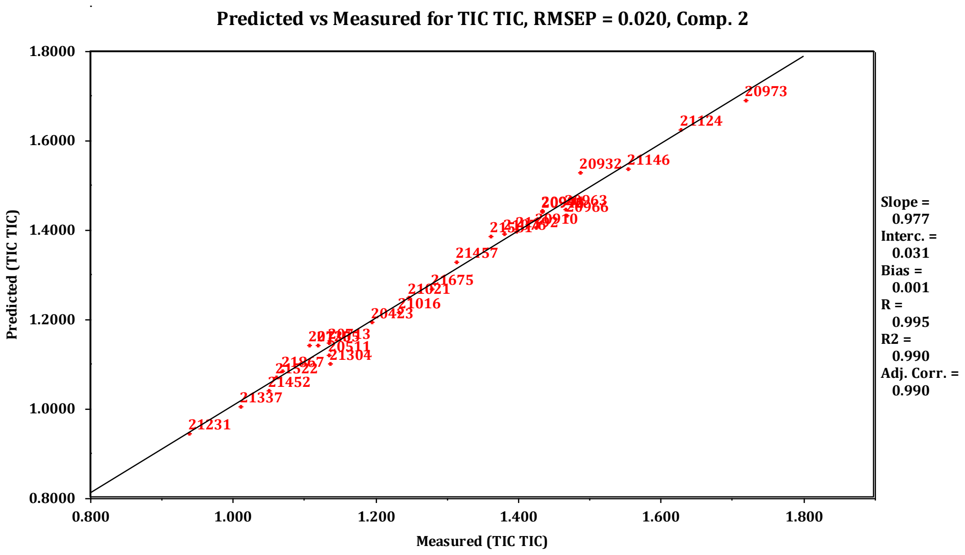


Figure 6-11 Predicted vs. Measured for the validation set, displaying the RMSEP value, R^2 and R_a^2

The Predicted vs. Measured plot in **figure 6-11** shows that the model accurately predicts the objects in the validation set. The coefficient of multiple determination is almost one, which means that the model describes virtually all the variation. The adjusted coefficient has the same value, which confirms that the number of components is correct, and indicate that the model can be trusted.

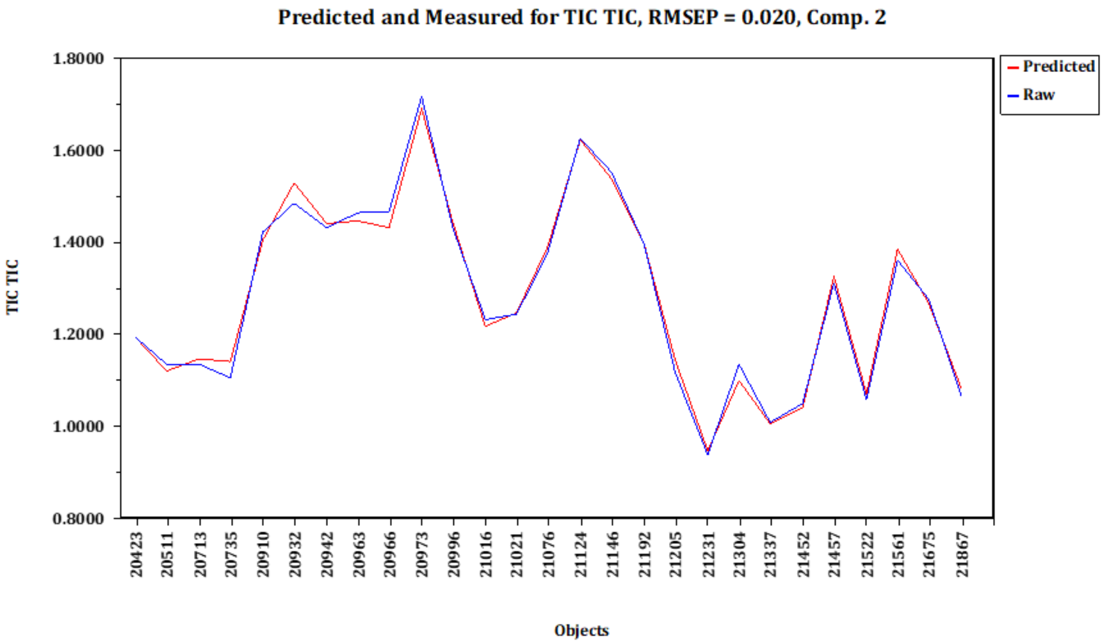


Figure 6-12 Predicted and Measured for the validation set, displaying good compliance between the predicted and measured values

The Predicted and Measured plot (**figure 6-12**) show that the predicted and measured values are approximately the same. The model has an average prediction error of 1.2% and small residuals, which is good. The RMSECV value from the regression analysis is 0.019, the fact that this value is very similar to the RMSEP value is a good sign, and points to the model being good. The RMSEP value for the model is 0.020, which is about 1-2% of the TIC values, verifying that the model is making acceptable predictions.

The RMSEP value for the Savitzky-Golay model was the same as for the EMSC model but based on the overall analysis of the number of components, cross-validation values and number of outliers the EMSC model is concluded to be the best model. Validation parameters for both models are presented in the table below.

Table 6-2 Validation parameters for the model with TIC as dependent variable

Validation parameters	**EMSC**	**Savitzky-Golay**
Explained information, dependent variable	99.21%	99.89%
Cross-validation on last component, CsvSD	0.20	0.78
Number of components in the model	2	3
RMSECV	0.019	0.016
RMSEP	0.020	0.020
R^2	0.990	0.992
Adjusted R^2	0.990	0.991
Average residual (absolute value)	0.016	0.019
Average prediction error	1.24%	1.22%

The average prediction error and residuals are calculated based on the report of the regression analysis after the model has been used to predict the data in the validation set.

6.1.2 Density

The density is correlated to the CO_2-loading, making it an interesting response variable to study. The CO_2-loading is given by moles CO_2/moles amine so that the density will increase as the amount of CO_2 in the sample increases. The density is given in kg/m^3, and the average value for density in this dataset is 1073.5 kg/m^3.

When using density as response the first model was built using all the wavenumbers, except the noisy parts. Different preprocessing techniques were used, and several models were made. To achieve the best possible model, the window of wavenumbers where narrowed down, ending up with a window from 1670 to 1000 cm^{-1}. When narrowing down this window the prediction abilities of the models built for different windows and the SR-plot was used. The SR-plot is attached in Appendix B2. The final window of wavenumbers is presented in the **figure 6-13** below. The spectra show quite a lot of spread, which is reduced by preprocessing and removal of outliers.

The dataset contains 129 samples with responses, which is enough samples to make two big datasets. The samples are split into a training set and a validation set after outlier detection; the outliers are not included in either set.

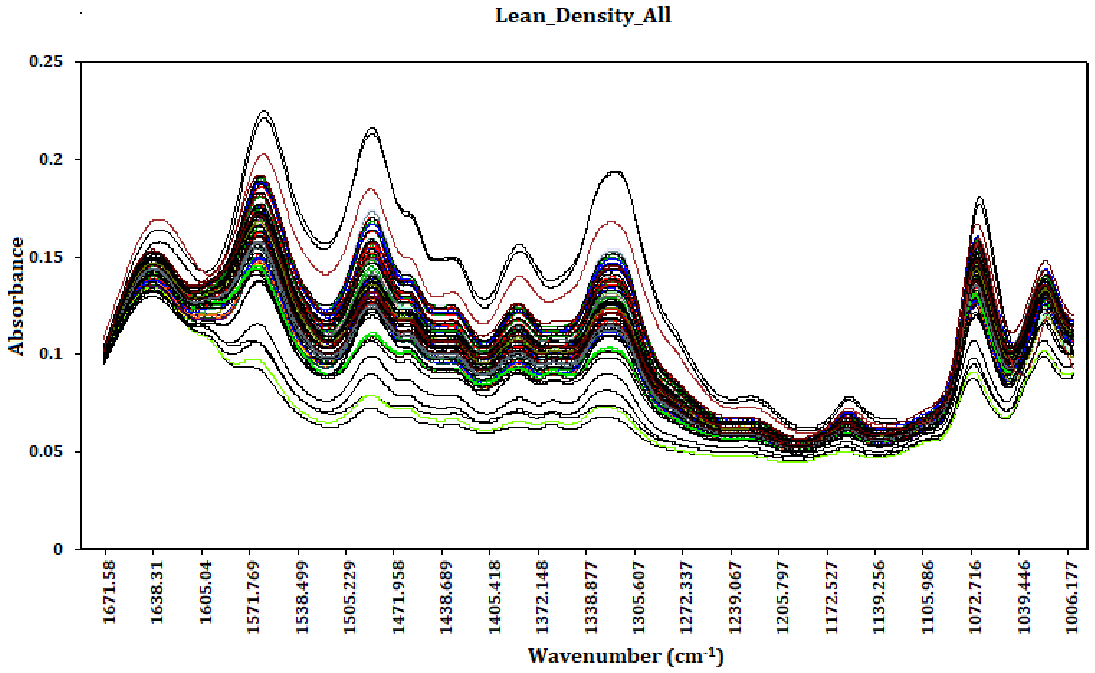


Figure 6-13 The fragment of the raw spectra for the lean samples used to build a model with density as response variable

Different methods of preprocessing were tried, but the one that gave the best result was second order Savitzky-Golay with a window size of 25 fitting the objects to a third-degree polynomial. The conclusion of which model was the best one was based on the number of components included in the model, the cross-validation values, the number of outliers, the RMSEP value and the ability to predict the samples in the validation set.

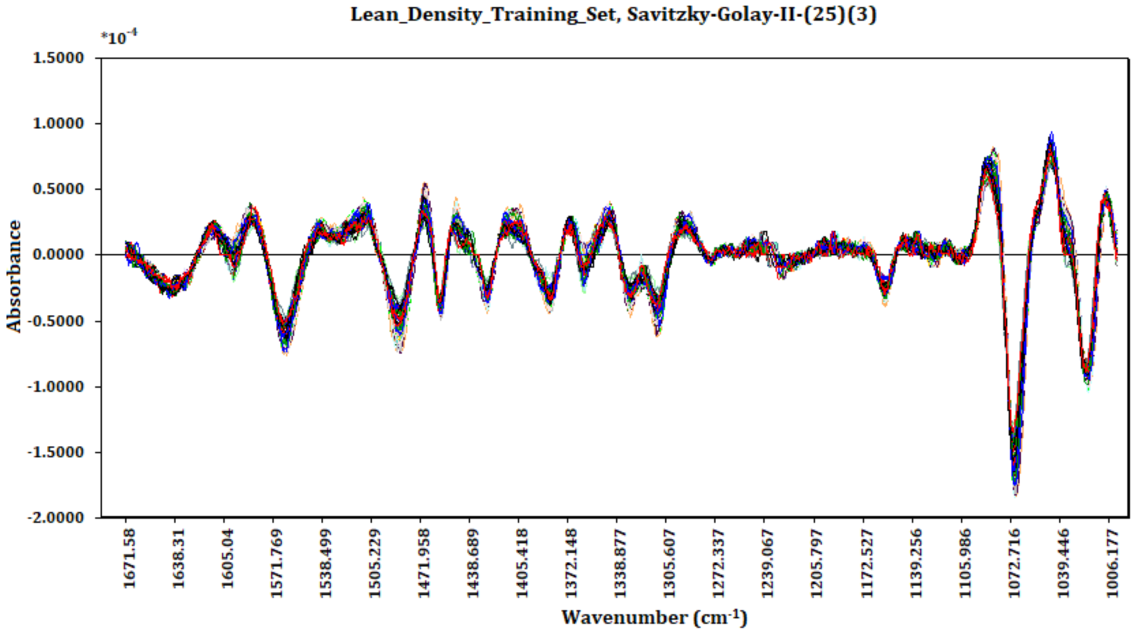


Figure 6-14 Spectra for the chosen region of wavenumbers, for the samples in the training set, after performing second-order Savitzky-Golay with a window of 25 and a third-degree polynomial as preprocessing

The spectra are after preprocessing are displayed in **figure 6-14**, and as can be observed in the plot, the interpretation becomes harder after performing differentiation by second-order Savitzky-Golay.

Since Savitzky-Golay is a preprocessing technique that can correct the outlying nature of an object, the outlier detection was performed after preprocessing; the identified outliers are presented in **table 6-3**. Outlier detection was performed using the score-plot, RSD vs. Leverage and Scores vs. Objects obtained from PCA. The most substantial outliers are removed first, to make sure that smaller outliers can be identified, the large outliers will influence the model in such a way that the smaller outliers might be buried. The plots used for the outlier detection can be found in Appendix B2.

Table 6-3 Outliers identified using PCA

Identified outliers
20991, 21431, 21665, 21786, 21830, 21916, 21936, 21969, 21971, 21990

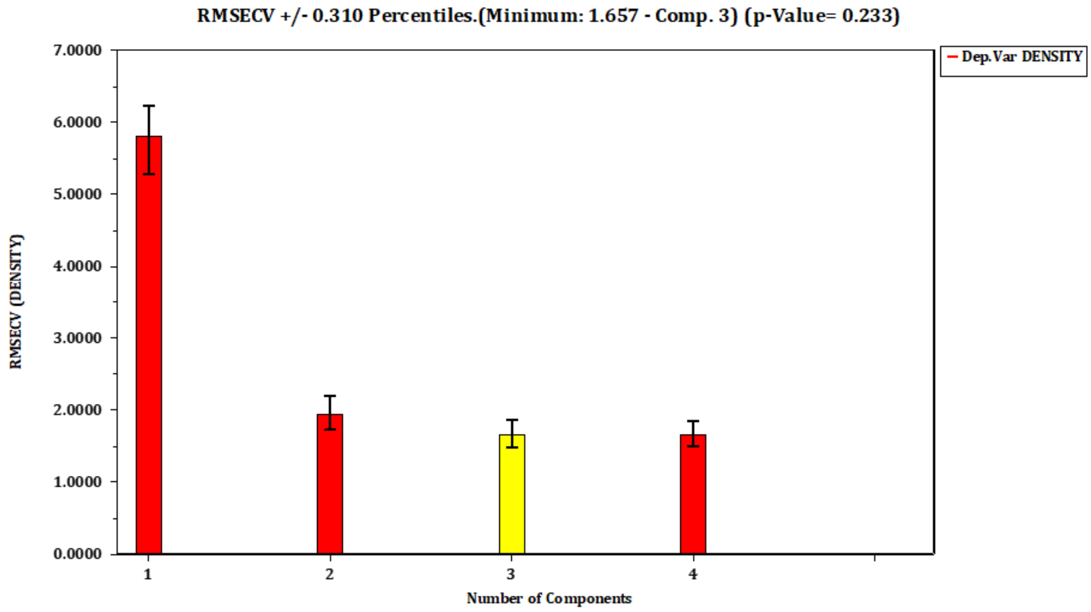


Figure 6-15 RMSECV-plot for the first four components extracted in the PLS model. Used to determine how many components to include in the model, the yellow bar indicate that three components should be included

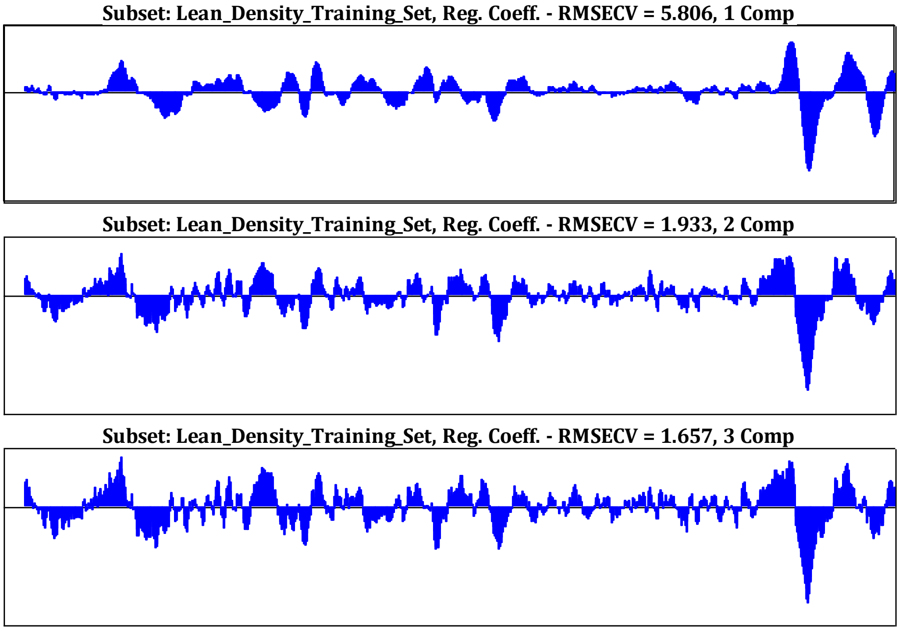


Figure 6-16 Weighted regression coefficients for the three components included in the model

As can be seen from the model dimensions plot in **figure 6-15** the suggested number of components to include in the model is three. The weighted regression coefficients in **figure 6-16** shows the RMSECV values for the three components, and also display that the components describe mostly data and not only noise. The RMSECV value decreases for the first three components, which leads to the conclusion of including three components in the model. The cross-validation values for the components are also taken into consideration when deciding how many components to include, and confirms that three components should be used, the value of the third component is presented in **table 6-4**. The calibration plot from the regression analysis shows god linearity when including three components, and a promising RMSECV value. This calibration plot is not presented here.

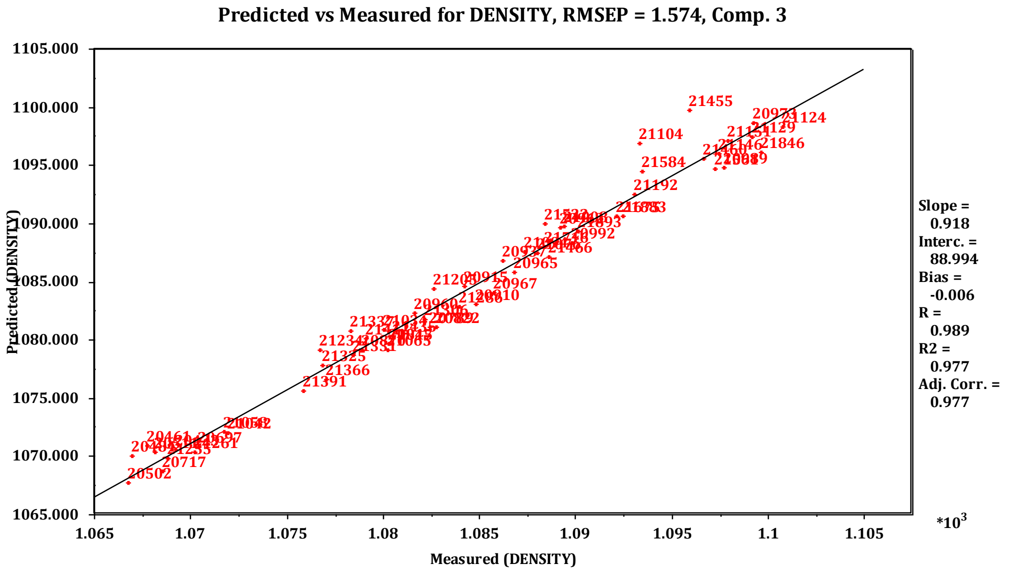


Figure 6-17 Predicted vs. Measured for the validation set, displaying the RMSEP value, R^2 and R_a^2

The Predicted vs. Measured plot (**figure 6-17**) show a good correlation between the predicted and measured values, with a coefficient of multiple determination of 0.977 and an adjusted coefficient with the same value. This means that 97.7% of the variation is described by the model, and since the coefficients have the same value as the number of components included in the model is correct.

The RMSEP value for this model is higher than for the TIC model, but still small compoare to the measured values, this is expected considering that the values for density are higher than the values for TIC.

The average prediction error for the model is 0.11%, which means that the model predicts the density with very satisfyingly accuracy. From the Predicted vs. Measured plot one can observe two samples (21104 and 21455) that lie a little further from the straight line, these are not identified as outliers, but they have deviating behavior from the rest of the samples. The deviation is more likely because of a mistake during the measurements, or a mistype when the value is typed in than the model being wrong.

Table 6-4 Validation parameters for the model with density as dependent variable

Validation parameters	
Explained information, dependent variable	99.83%
Cross-validation on last component, CsvSD	0.86
Number of components in the model	3
RMSECV	1.657
RMSEP	1.574
R^2	0.977
Adjusted R^2	0.977
Average residual (absolute value)	1.243
Average prediction error	0.115%

The average prediction error and residuals are calculated based on the report of the regression analysis after the model has been used to predict the data in the validation set.

The RSD plot in **figure 6-18** confirms that the model is decent, the plot should be a straight line and pass through 0.5, which it does.

In summary, the model can accurately predict the density, with small residuals and prediction error. The RMSECV and RMSEP value are similar, and both calibration and prediction curves show good linearity. The RMSEP value is in the 0-1% range of the density values, leading to the conclusion that the model is excellent.

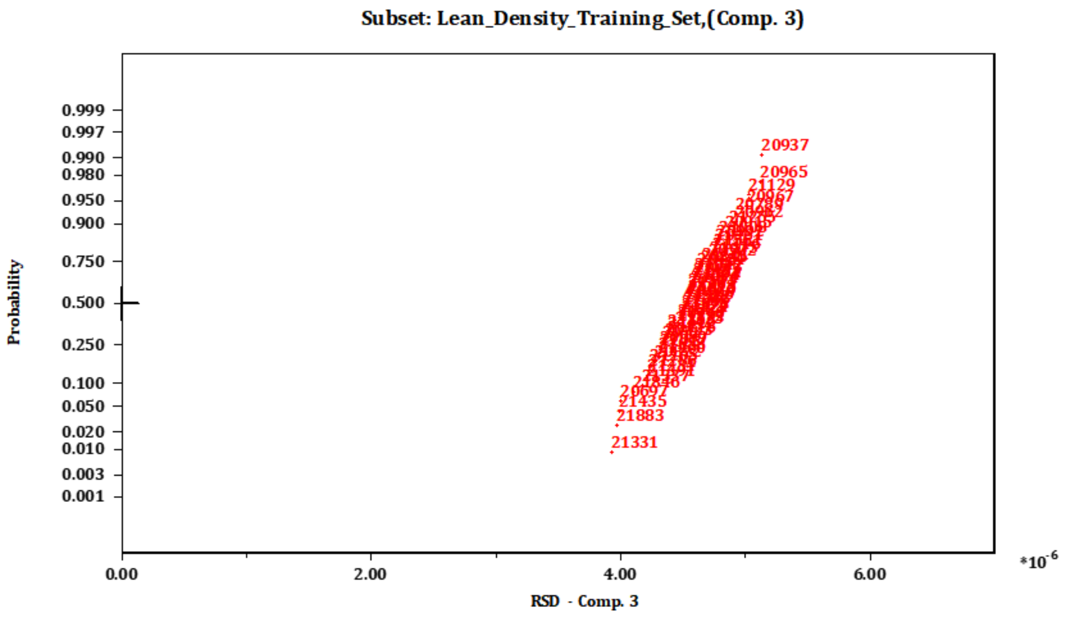


Figure 6-18 Normal plot of the residual standard deviation after for the objects in the validation set

6.1.3 Total Alkalinity

Total alkalinity (TOT_ALK) is determined by titration with HCl. It is used to determine the concentration of amines in the solution and is expected to have a higher value in the lean samples than in the rich samples (chapter 3.2). The dataset contains 64 samples and responses, with an average TOT_ALK value of 4.8 mole/kg.

Several approaches were tested for this dataset and several windows of wavenumbers. The regions of wavenumbers that gave the best results were: 3000 to 2800 cm^{-1}, 1680 to 1280 cm^{-1} and 1100 to 1000 cm^{-1}. These regions of wavenumbers were found based on where the organic carbons absorb IR; the C-H stretch is in the region of 2927 – 2864 cm^{-1} and the other regions were detected using the SR-plot [14]. Predictions of amine concentration in an aqueous sample have been determined using PLS before, using approximately the same spectral regions [42].

The spectra of the chosen wavenumbers are displayed in **figure 6-19**, the spectra have a spread in the mid-region, which is removed with preprocessing. The preprocessing techniques used demands that the data is continuous, the preprocessing was therefore performed before the spectra were split into these regions.

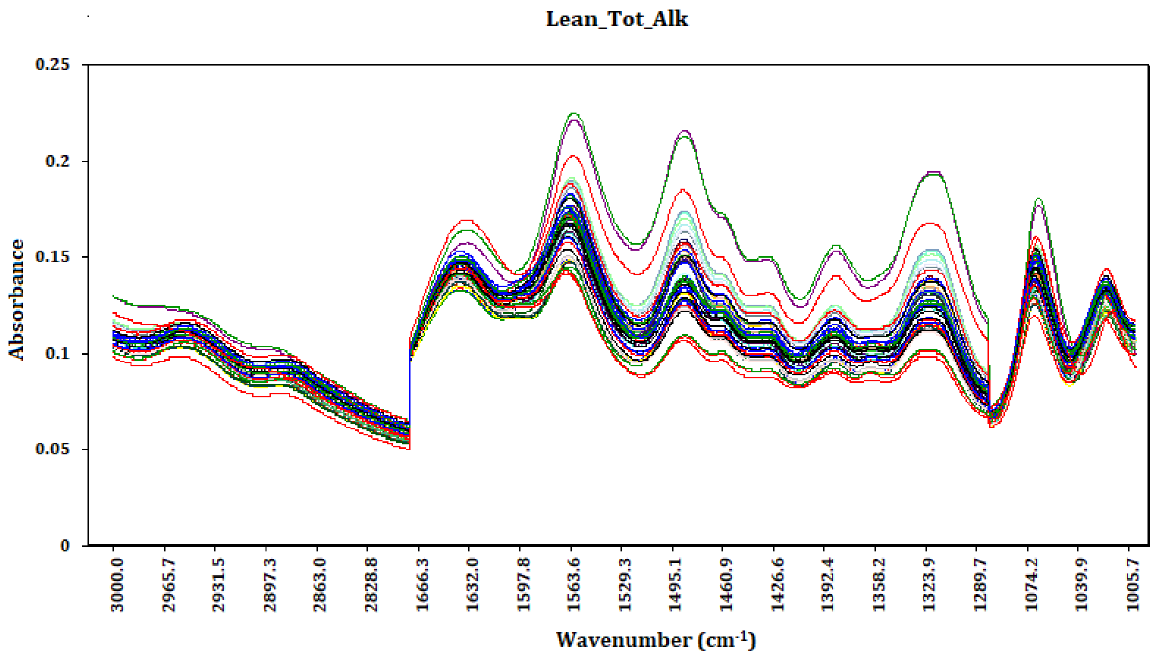


Figure 6-19 The fragment of raw spectra for the lean samples used to build a model with TOT_ALK as response variable

Outlier detection was performed using the score plots, RSD vs. Leverage and normal plot of the scores from PCA. Three sizable outliers are detected in the score plot, these are removed first, and a new model is made to identify smaller outliers. The outliers are presented in **table 6-5,** and the plots used for outlier detection is attached in Appendix B3.

Table 6-5 Outliers identified using PCA

Identified outliers
20961, 20966, 20967, 20988, 20991, 21129, 21260, 21431, 21462, 21665, 21699, 21786

The first models that were built described very little of the dependent variable, TOT_ALK, so the response variable was investigated closer. The normal plot shows that the response variable is not normally distributed, as can be seen in **figure 6-20**, which explains why it is so difficult to model.

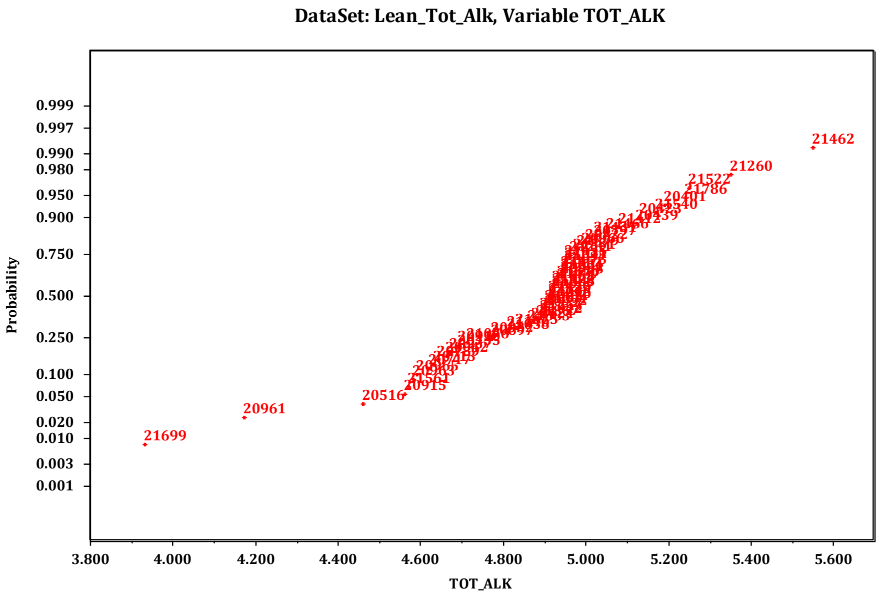


Figure 6-20 Normal plot of the response variable TOT_ALK

Preprocessing in the form of ½ root transform was performed on the response, which made it normally distributed, as seen in **figure 6-21**.

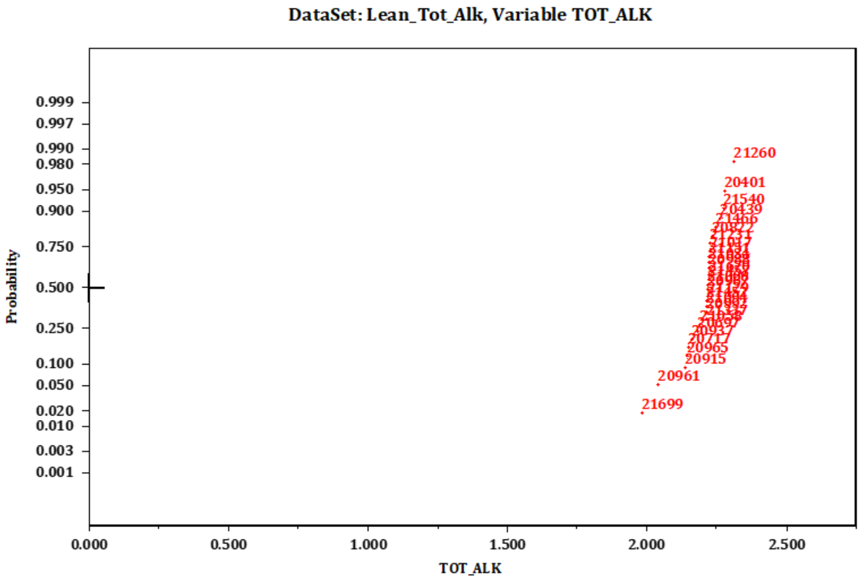


Figure 6-21 Normal plot of the response variable after preprocessing with root ½

New models are now made, and the results have improved, the description of the dependent variable has increased by more than 50%.

After analysis of the models made, it is concluded that the dataset is not large enough to split it into a training set and a validation set. The prediction abilities of the models made with the training set are not good, and especially not the coefficients of multiple determination, with values around 0.30. Which means that only about 30% of the variation in the samples can be described by the model, as can be seen on the plots as well, which are scattered (attached in Appendix B3). The model did, however, have a decent RMSEP value, but the model was not able to predict the values deviating from the average, which can be observed in the Predicted and Measured plot (attached in Appendix B3).

Therefore, the entire dataset is used to build the model, and the model is validated using double cross-validation and the results from the regression analysis. Proper validation of the model can be performed as soon as more data becomes available.

The best models were obtained using second-order Savitzky-Golay and second order EMSC, and the best model is concluded to by with second-order Savitzky-Golay with a window of 21 and a third-degree polynomial. The conclusion was made based on the number of components, RMSECV values, and cross-validation values. Only the best model is presented here.

The spectra after preprocessing are presented in **figure 6-22**. The part of the spectra from 3000 to 2800 cm^{-1} seems to be filled with noise, but a model was built where this region was omitted, but the prediction abilities did not improve compared to the model were this area was included. The area is therefore included in the model, and the results are presented below.

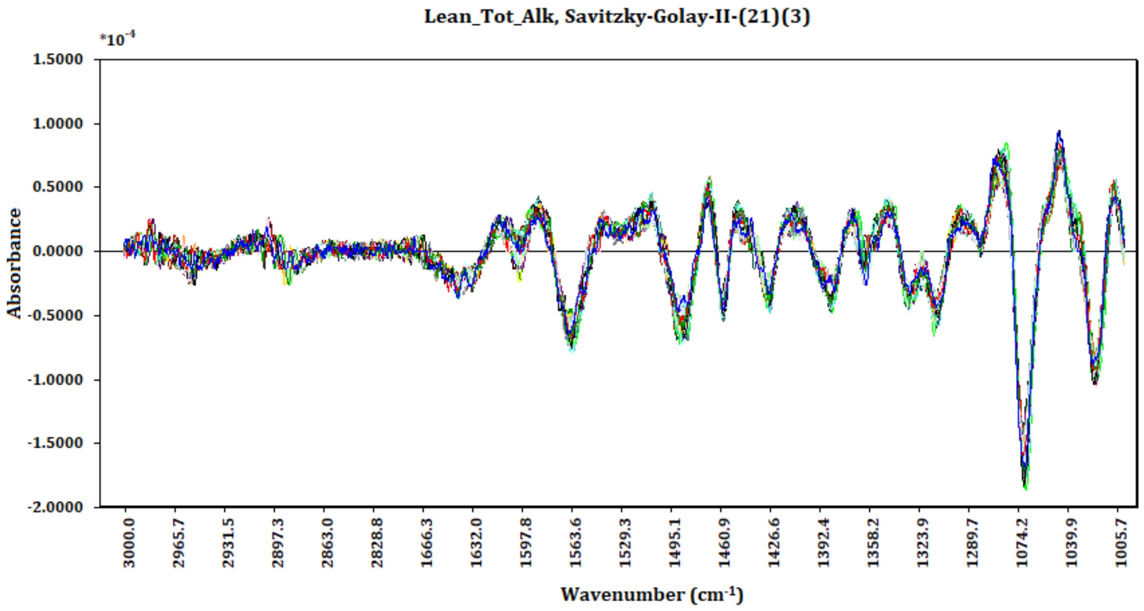


Figure 6-22 Spectra of the chosen region of wavenumbers after performing second-order Savitzky-Golay with a window of 21 and a third-degree polynomial as preprocessing

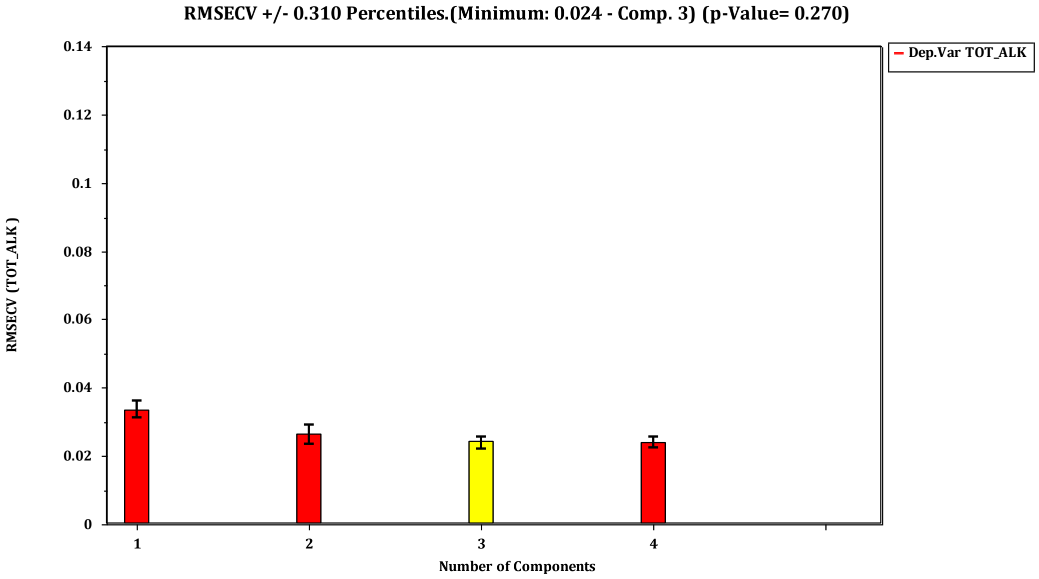


Figure 6-23 RMSECV-plot for the first four components extracted. Used to determine how many components to include in the model, the yellow bar indicates that three components should be included

The model dimensions plot (**figure 6-23**) suggest that three components should be included in the model, this is confirmed by the weighted regression coefficients in **figure 6-24**. The weighted regression coefficients show that the first component mainly describes the data, component two and three describe some noise, but they are included in the model based on the RMSECV values and the cross-validation values.

A component should be included in the model as long as the RMSECV value decreases by including it, which it does for the third component.

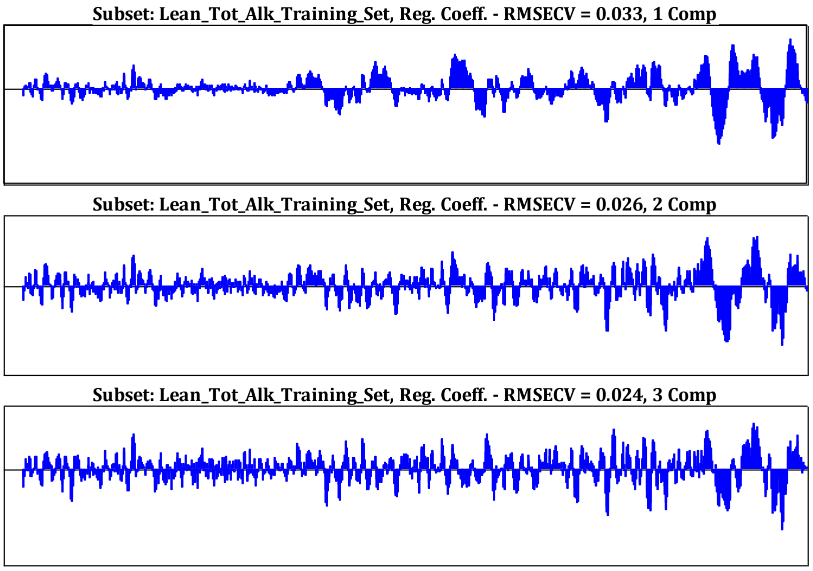


Figure 6-24 Weighted regression coefficients for the three components included in the model

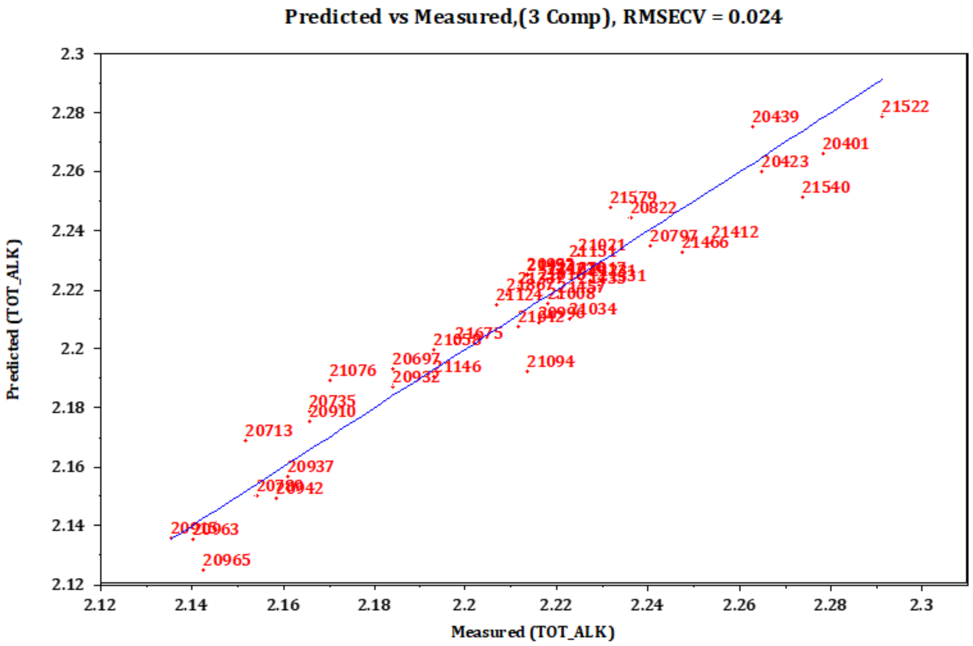


Figure 6-25 Predicted vs. Measured from the regression analysis, displaying the RMSEC-value

The calibration curve in **figure 6-25** has good linearity when including three components, and the RMSECV value is 0.024, which means that the model should give satisfactory predictions.

The cross-validation values are below one, which is one of the demands for a component to be included in the model. The amount of described independent and dependent information, presented in **table 6-6**, by the components is not as high as desired, but the performance of the model is still decent.

The plot in **figure 6-25** is for the model after the response variable has been preprocessed with root ½, and the RMSECV improved a lot after this was done. Before the preprocessing of the response, the RMSECV value was 0.111, and the Predicted vs. Measured plot is more scattered, as can be seen in **figure 6-26**. The cross-validation values for the model with no preprocessing of the response had higher values as well, which lead to the conclusion that the model with the preprocessed response is favorable. The validation parameters for the model after preprocessing of the response is presented in **table 6-6**.

Both models had the same number of outliers, and even though the model with no preprocessing of the response only has two components the RMSECV value improved in such a magnitude that the model after preprocessing is assumed to have much better prediction abilities and is the favorable one.

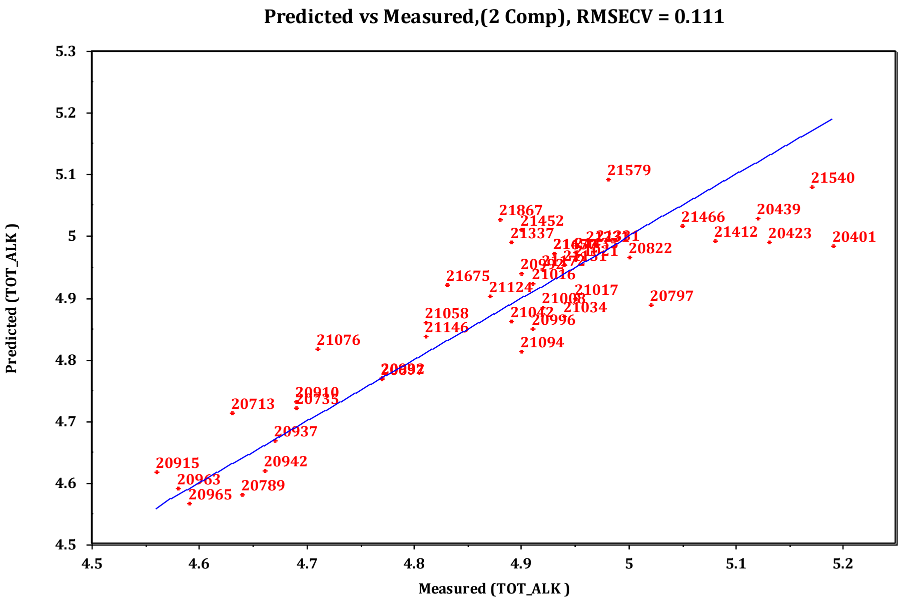


Figure 6-26 Predicted vs. Measured from the regression analysis when the response variable is not preprocessed. Only two components are included in this model

Table 6-6 Validation parameters for the model with TOT_ALK as dependent variable

Validation parameters	
Explained information, dependent variable	78.66%
Cross-validation on last component, CsvSD	0.77
Number of components in the model	3
RMSECV	0.024
RMSEP, *from double cross-validation*	0.024

Double cross-validation was used as a model evaluation and resulted in a RMSEP value of 0.024, which is the same as the RMSECV value from the regression analysis. The Predicted vs. Measured plot from the double cross-validation is presented in the **figure 6-27** and looks very similar the one from the regression analysis in **figure 6-25**.

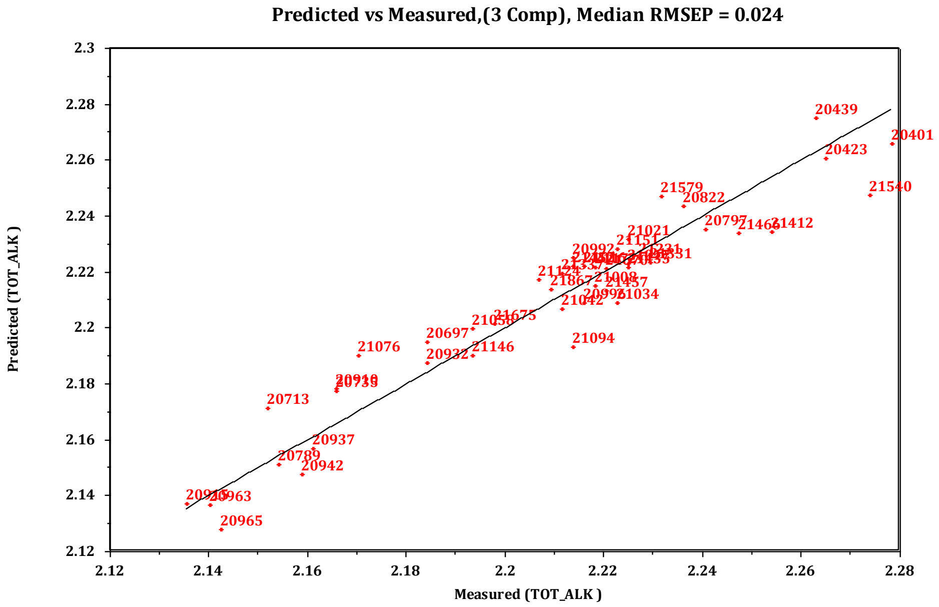


Figure 6-27 Predicted vs. Measured from double cross-validation, displaying the RMSEP-value

Based on the analysis of the models presented here, it is concluded that a model built based on the preprocessed response, and all the objects give the best model. With a RMSEP value of about 0.5% of the measurement ranges from double cross-validation the model should be capable of making satisfactory predictions.

Good linearity can be observed in the calibration plots when the entire dataset was used to build the model; this linearity was not present when the data was split in two, confirming the choice to include the entire dataset in the regression analysis.

6.2 Rich Samples

For each response several models have been made and analyzed, only the best ones are presented here. From **figure 6-28,** the same observation as for the lean samples can be made: there are two regions with noise. These regions are not used in the modeling; the areas are 2500 to 1800 cm^{-1} and 700 to 400 cm^{-1}.

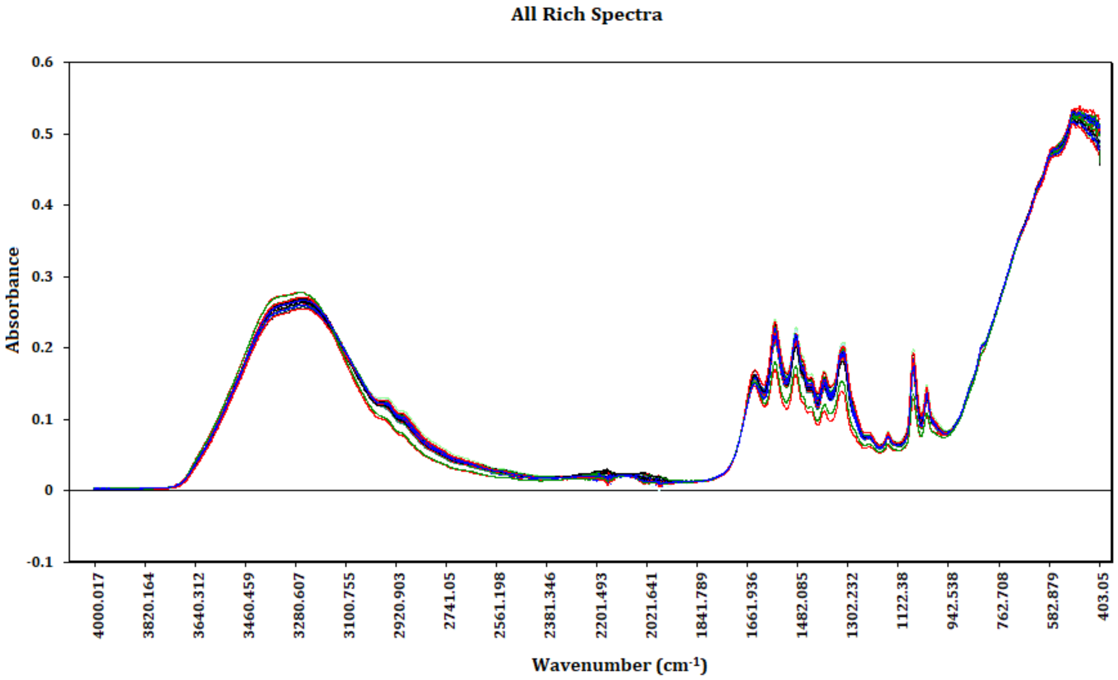


Figure 6-28 All the rich spectra

6.2.1 Total Inorganic Carbon

The dataset contains 37 samples with response values, the response values are given in moles/kg and has an average value of 2.23 mole/kg for this dataset. This is, as expected, a higher value than for the lean samples because the rich samples contain more CO_2 from the absorption process (see section 3.2).

The total inorganic carbon is expected to be modeled best in the areas where the inorganic carbon has the highest intensity of absorption. The window of wavenumbers was chosen based on this and the spectral region used for modeling is from 1670 to 1000 cm^{-1} [13]. The chosen window is displayed in **figure 6-29** below, based on this it can be observed that there are some deviating spectra, these are probably outliers and will be removed in the exploratory analysis or by preprocessing.

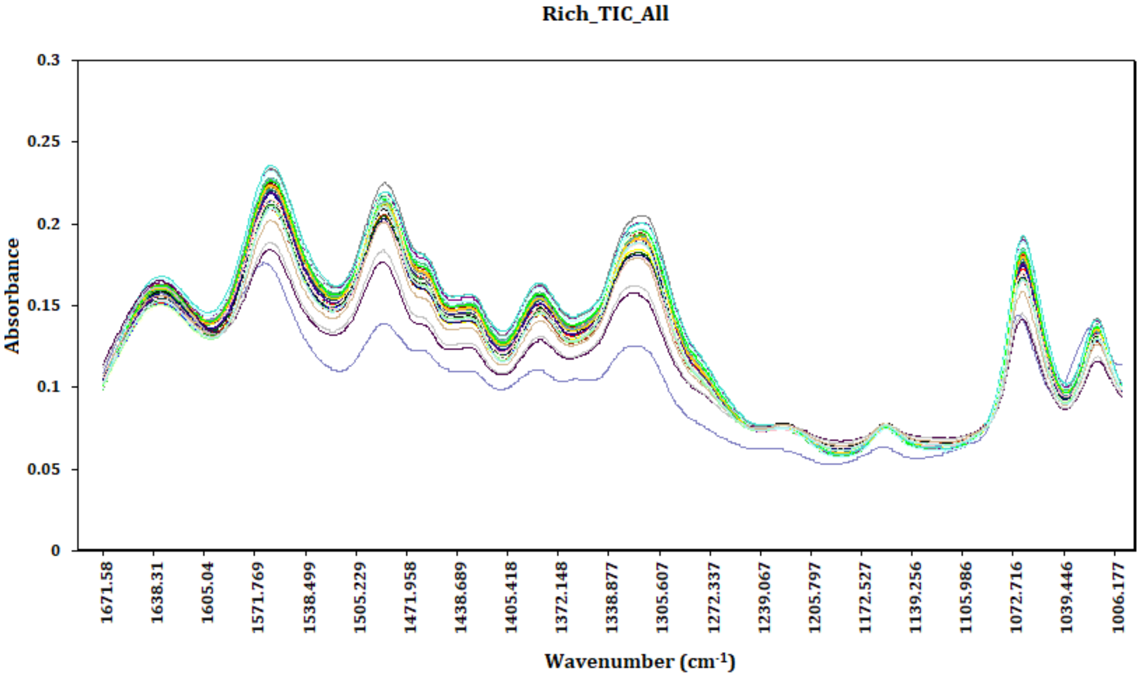


Figure 6-29 Fragment of raw spectra for the rich samples used to build a model with TIC as response variable

The score plots, RSD vs. Leverage and Scores vs. Objects plots obtained from PCA is used for outlier detection. The score plots show three large outliers, these are removed first, and a new model is made. The model changes dramatically after removal of these, and smaller outliers can now be detected. The plots used in this analysis is attached in Appendix C1. The detected outliers are presented in **table 6-7**.

Table 6-7 Outliers identified using PCA

Identified outliers
20400, 20941, 21029, 21431, 21578, 21664, 21970, 21972, 21975, 21989, 21991, 22023

After trying different preprocessing techniques, second order EMSC resulted in the best model. This conclusion was made based on several factors: the RMSEP values, cross-validation values, number of components included in the models and how much dependent information was explained by the components. The results from the other models are not presented here.

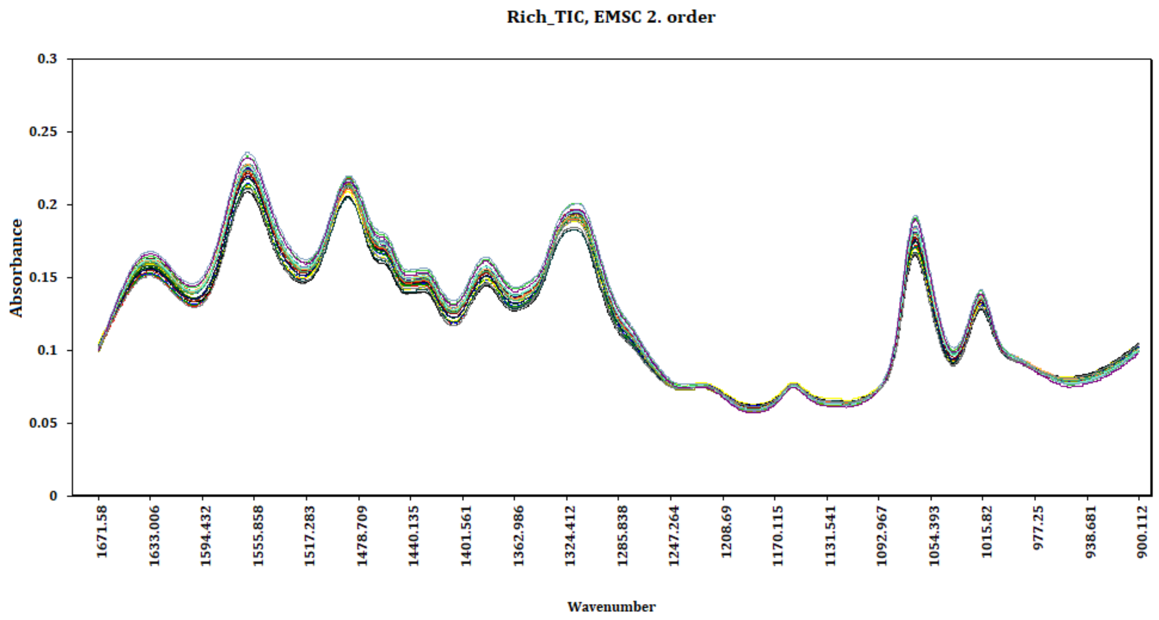


Figure 6-30 Spectra of the chosen region of wavenumbers after performing second-order EMSC as preprocessing

The spectra after the preprocessing and removal of outliers are presented in **figure 6-30**, which displays that the spread of the data is severely decreased, which is a good base for building a model.

The dataset with the samples corresponding to the response TIC is tiny, only containing 37 samples, and the entire dataset was therefore used to build the model, which later can be tested on a validation set once more samples are available.

The first model described 64.71% of the dependent variable and gave a decent RMSECV value, but to try and get the amount of described information higher variable selection was performed. The variable selection performed was VIP, removing variables with VIP less than 0.5.

The model was then evaluated to make sure that the variable selection improved the model. The cross-validation value for the first component dropped drastically and a decrease of the cross-validation value for the second component led to the conclusion that the variable selection improved the model. A reduction in the RMSECV value confirmed this conclusion.

Variable selection with a higher VIP limit was also tested, but this did not improve the model further. The validation parameters for the model from before and after variable selection are presented in **table 6-8**. The results presented below are for the model after performing variable selection.

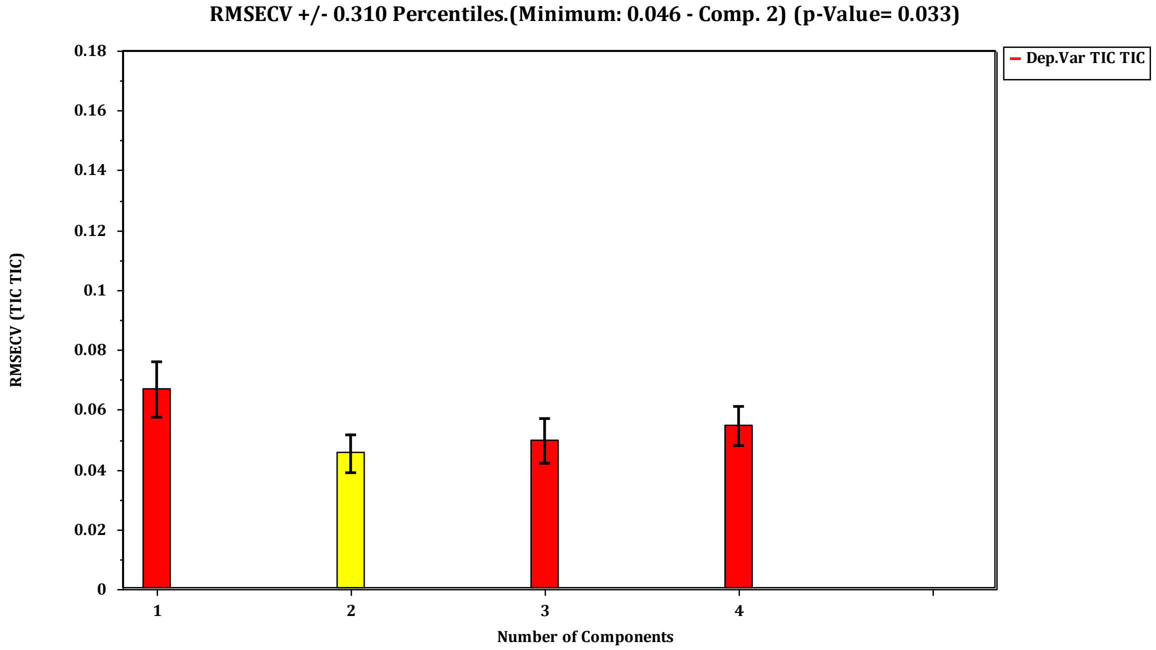


Figure 6-31 RMSECV-plot for the first four components extracted. Used to determine the number of components to include in the model, the yellow bar indicates that two components should be included

The model dimensions plot in **figure 6-31** advice to include two components in the model, the cross-validation values for the components support this. Two components are therefore included in the model, resulting in decent linearity in the calibration plots and a decent RMSECV value.

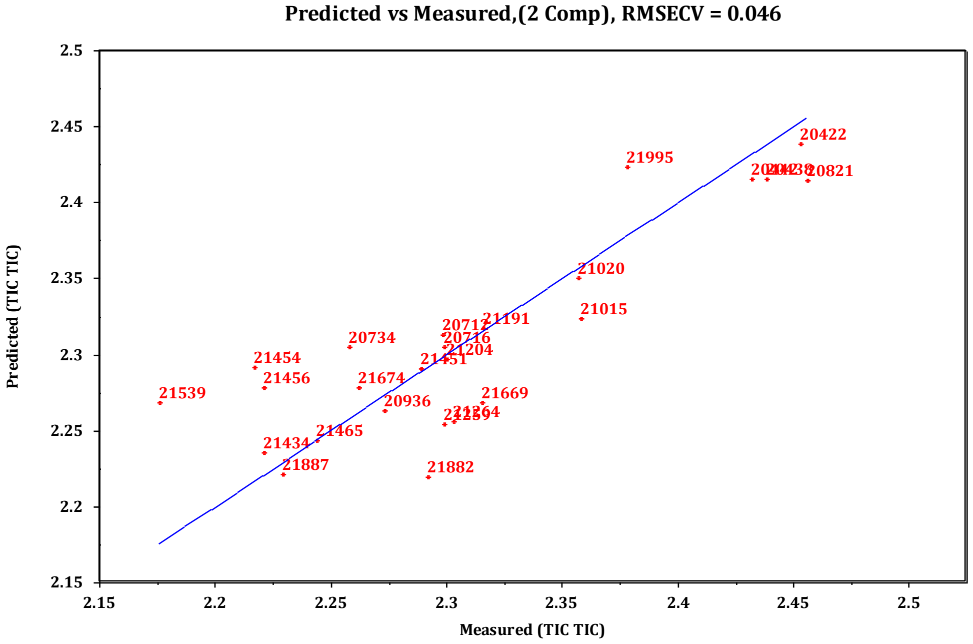


Figure 6-32 Predicted vs. Measured from the regression analysis, displaying the RMSECV value

The Predicted vs. Measured plot from the regression analysis is a measure of the models' ability to predict the objects that are used to make the model. Because of the lack of a validation set for this model, the validation is based on the results of the regression analysis. The RMSECV value is small compared to the measured values, which is a good sign. The calibration plot in figure 6-32 does not show as good linearity as desired, but other models and techniques did not improve this. The model is therefore concluded to be tolerable and is expected to make adequate predictions.

The score-plot in **figure 6-33** shows that there are no outliers, even though the Predicted vs. Measured show some spread in the objects. The TIC values are very small, which explains why the objects look to be spread, a small deviation from the straight line will look more extreme when the values are of such small magnitude.

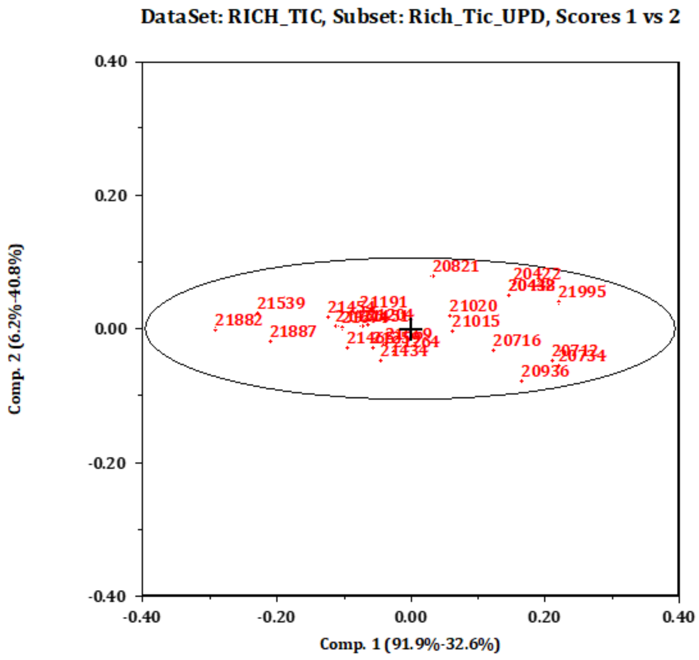


Figure 6-33 Score plot for component 1 and 2, obtained from the regression analysis

Figure 6-34 displays the Response residuals; this plot should be a straight line and pass through y = 0.5 for a model to be good. As can be observed, the objects have lined up an approximately linear line. The graph shows some objects lying further to the left than the rest, even though they are on the line straight line. The same type of clustering can be observed in the Predicted vs. Measured plot above; the objects seem to be in two groups, one to the left and one to the right. The same clusters can be seen in the score plot. This grouping of the objects may be due to the small number of samples to build the model and, and a model containing more samples would probably smooth out the spread.

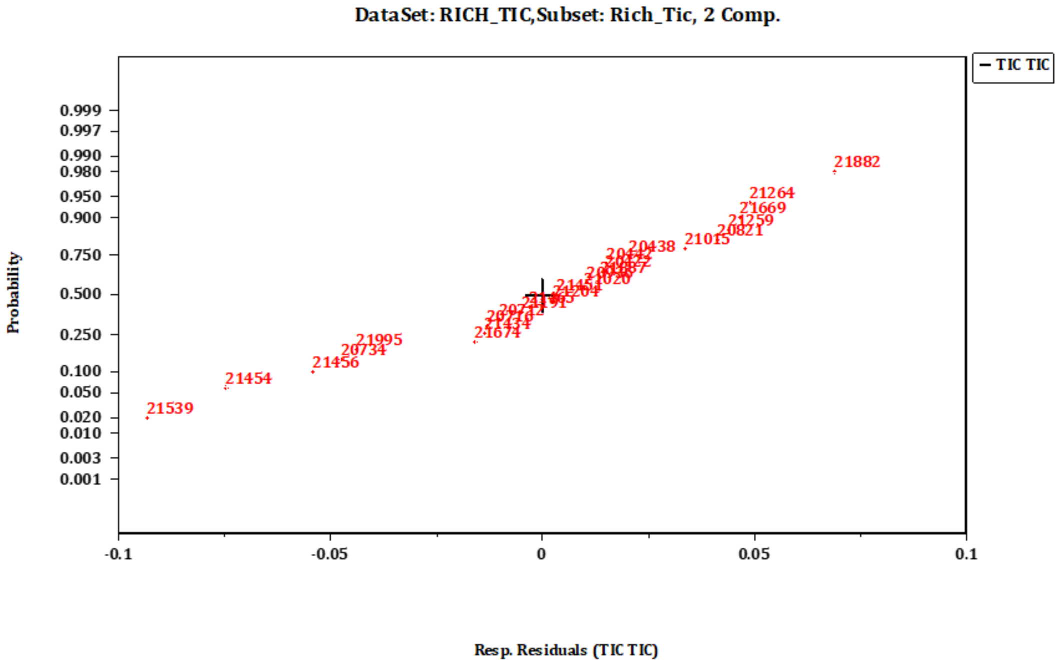


Figure 6-34 Normal plot of the response residuals from the regression analysis

Table 6-8 Validation parameters for the model with TIC as dependent variable

Validation parameters	
Explained information, dependent variable	64.71%
Cross-validation on last component, CsvSD	0.70
Number of components in the model	2
RMSECV	0.049
Explained information, dependent variable *after variable selection*	73.42%
Cross-validation on last component, CsvSD *after variable selection*	0.68
Number of components in the model *after variable selection*	2
RMSECV *after variable selection*	0.046
RMSEP, double cross-validation	0.046

A model evaluation using double cross-validation was carried out to validate the model. The obtained results confirm the conclusion made about the performance of the model. The model dimensions plot from the double cross-validation is presented in **figure 6-35** and suggest two components should be included in the model and gives the minimum RMSECV value at 0.0457 which is the same result as for the regression analysis.

From the double cross-validation, a Predicted vs. Measured plot is also obtained, presented in **figure 6-36**, this plot gives a RMSEP value of 0.046, which also confirms the conclusions made about the predictive abilities from the regression analysis.

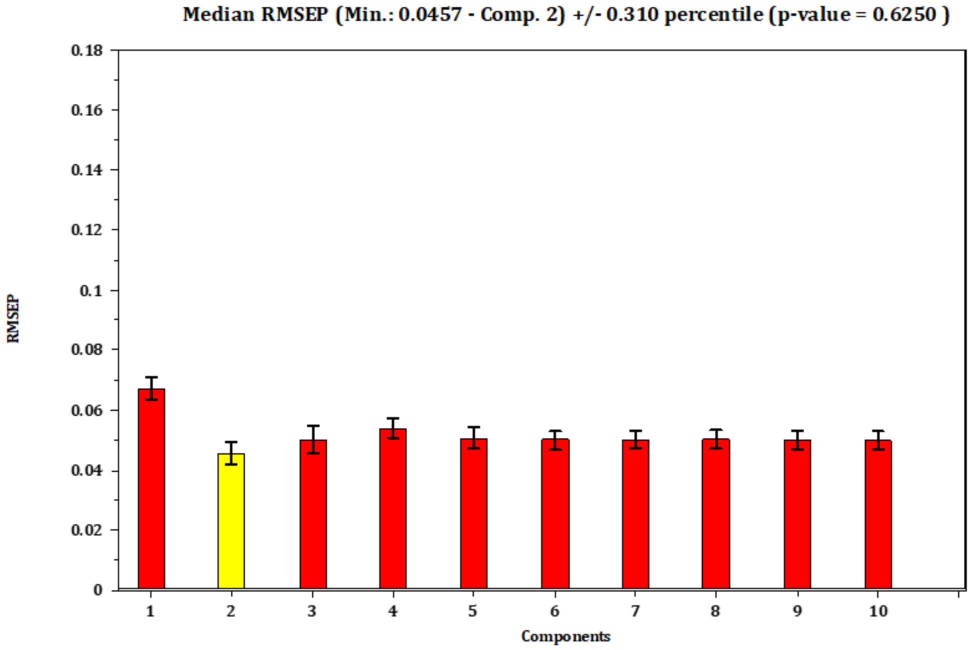


Figure 6-35 RMSECV-plot obtained from the double cross-validation, indicating that two components should be included in the model

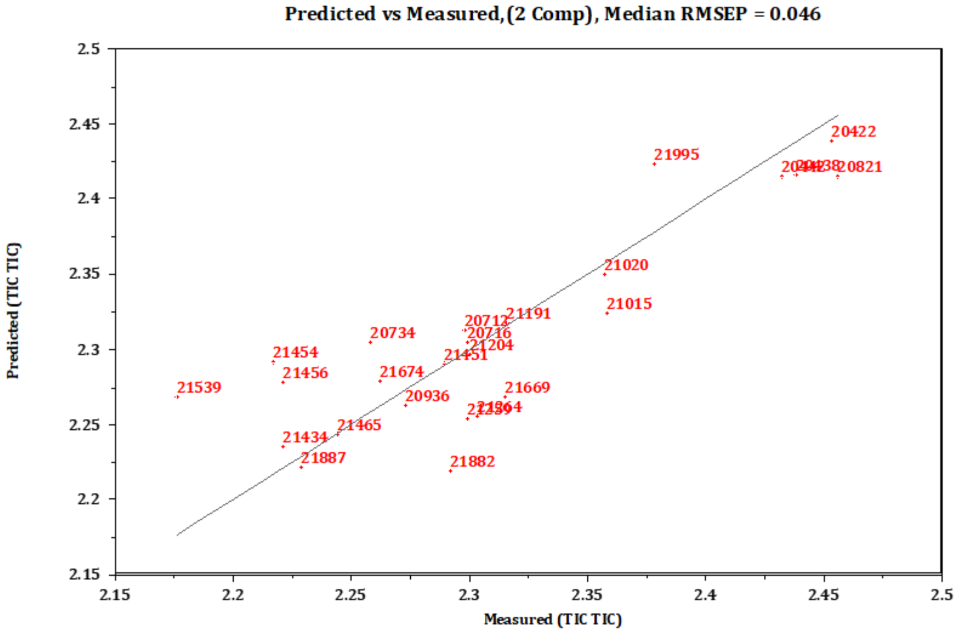


Figure 6-36 Predicted vs. Measured from double cross-validation, displaying the RMSEP-value

In summary, the amount of described variation in the dependent variable is not as high as desired, but the validation parameters are still adequate, and the predictions are expected to be tolerable. The RMSEP value from the double cross-validation is in the 1-2% range of the TIC values, meaning that the model is expected to have small prediction errors and residuals.

6.2.2 Density

The dataset consists of 99 CO_2 rich samples and responses, which gives a good base for building a model and is enough to split the data into a training set and a validation set. The density is given in kg/m^3, and the average value for density in this dataset is 1128 kg/m^3, which is higher than for the lean samples. This is because the rich samples contain more CO_2 than the lean samples, and this affects the density of the samples (more in section 3.2).

The density model was first built based on the entire range of wavenumbers, excluding the ones identified as noisy. Since the density models for the lean samples had the best prediction ability in the region from 1670 to 1000 cm^{-1}, this area was also investigated for the rich samples. This region resulted in the best model here as well, based on the

RMSEP values, cross-validation values, and overall prediction abilities. The model based on this region is the only one presented here.

From the spectra in **figure 6-37** below it is observed some spread and some deviations, which is expected to be removed using outlier detection and preprocessing.

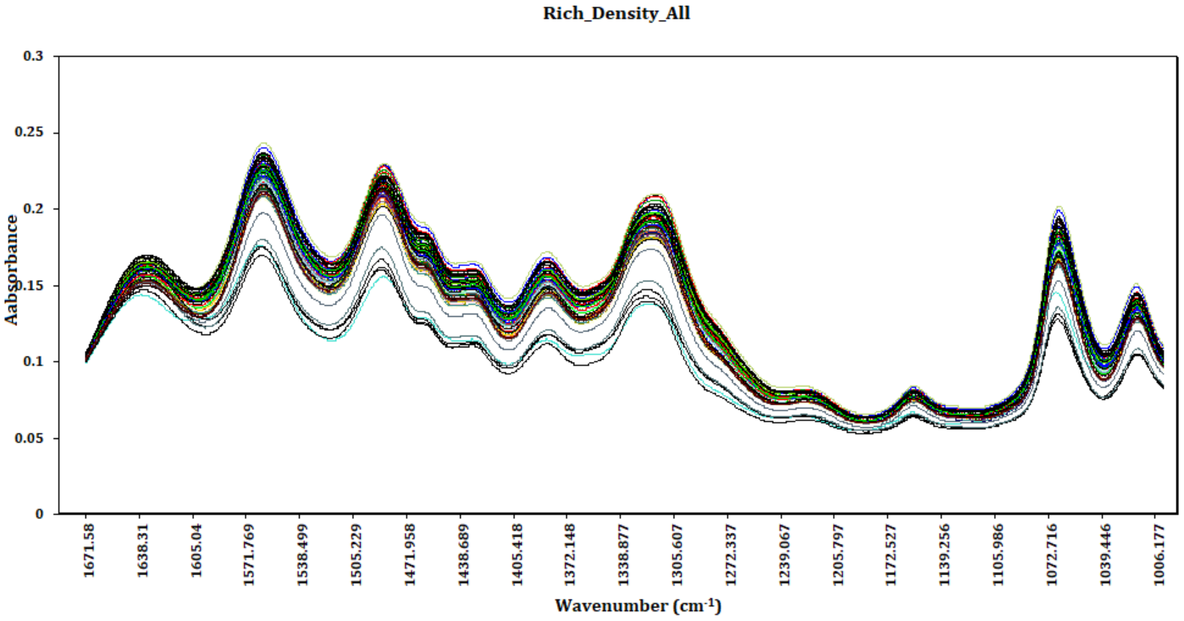


Figure 6-37 Fragment of raw spectra for the rich samples used to build a model with density as response variable

Outlier detection using Score plots, RSD vs. Leverage and Scores vs. Objects from PCA showed that the dataset contains 12 outliers, presented in **table 6-9**. The most extreme outliers were removed first, and then new models were made to detect smaller outliers. The plots used to identify the outliers are attached in Appendix C2.

Table 6-9 Outliers identified using PCA

Identified Outliers
20914, 20915, 20990, 21430, 21664, 21829, 21845, 21915, 21952, 21970, 21972, 2189

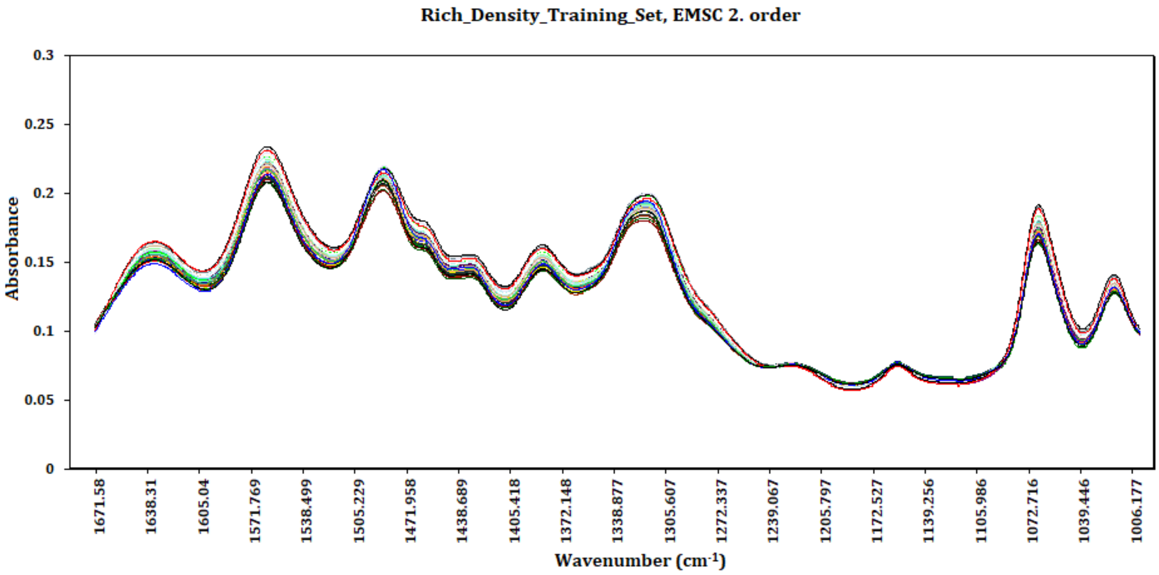


Figure 6-38 Spectra of the chosen region of wavenumbers after performing second-order EMSC as preprocessing

Different preprocessing techniques were tested, and the models evaluated. Based on the results the most promising model was when using EMSC of the second order, the conclusion was made based on the RMSEP, cross-validation values and the number of outliers. The results from this model are presented here; the other models are not presented here. The plot in **figure 6-38** shows the samples in the training set after preprocessing and removal of outliers and show that the spread in the spectra decreased after preprocessing.

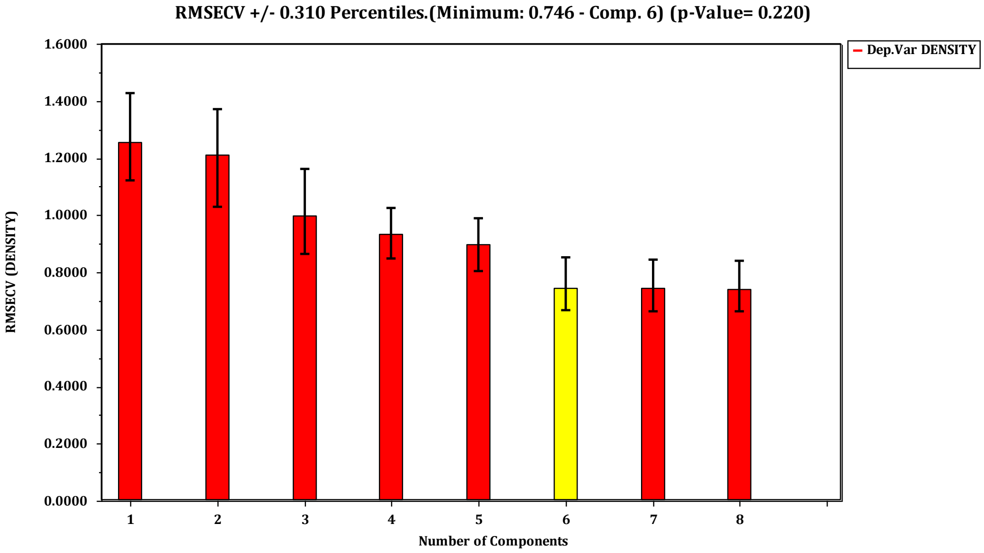


Figure 6-39 RMSECV-plot for the first eight components extracted, the yellow bar indicates that six components should be included

The model dimensions plot in **figure 6-39** suggests that six components should be included in the model. The weighted regression coefficients in **figure 6-40** show that some noise is included in the last two components but considering the RMSECV value and cross-validation values, the fifth and sixth components were included in the model. The inclusion of these two components also improved the RMSEP value for the model.

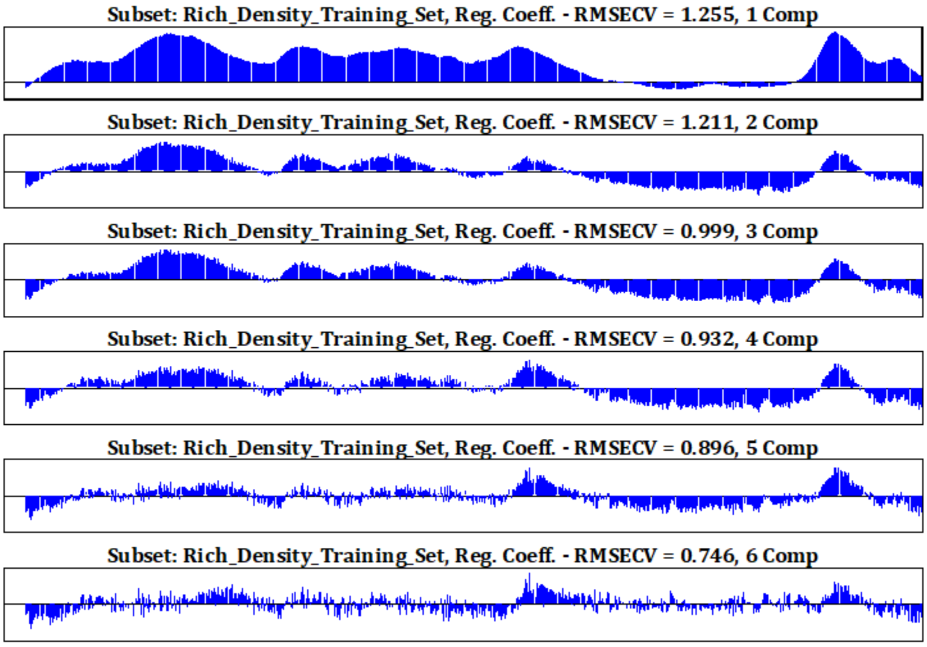


Figure 6-40 Weighted regression coefficient for the six components included in the PLS model

The validation set was then used to evaluate the model; the model gave good predictions of the objects in the validation set. The RMSEP value is 0.649 which is an excellent value considering that the density values are big. The Predicted vs. Measured is displayed in **figure 6-41** and presents an approximately straight line with small deviations. The average prediction error for the model is 0.046%, and the residuals are small.

The coefficient of multiple determination and the adjusted coefficient has the same value, which confirms that the choice of including six components in the model was right. With a value of 0.991, the coefficient of multiple determination states that 99.1% of the variation in the data has been described by the model.

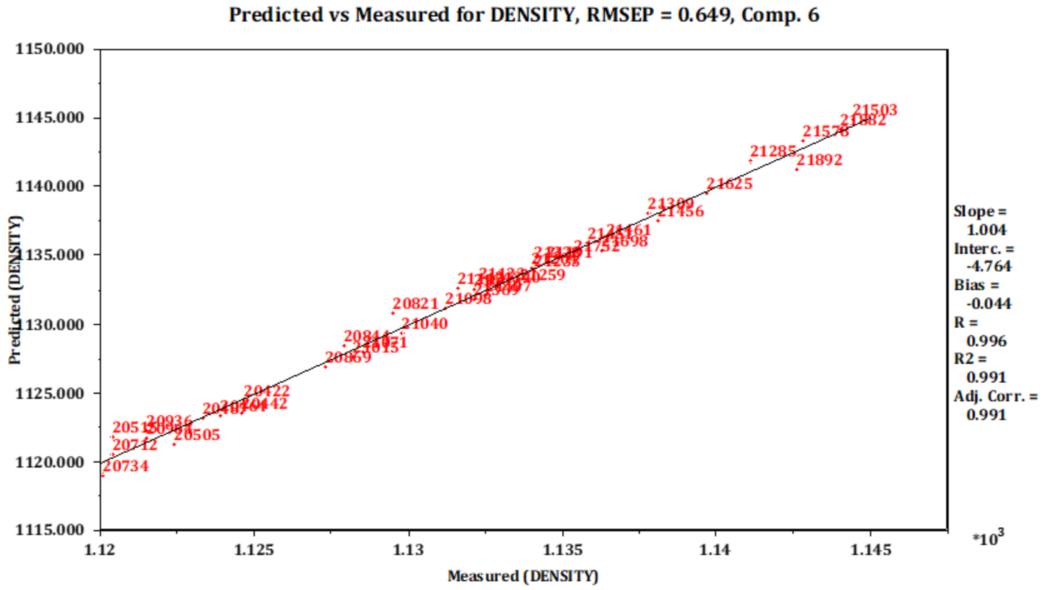


Figure 6-41 Predicted vs. Measured for the validation set, displaying the RMSEP value, R^2 and R_a^2

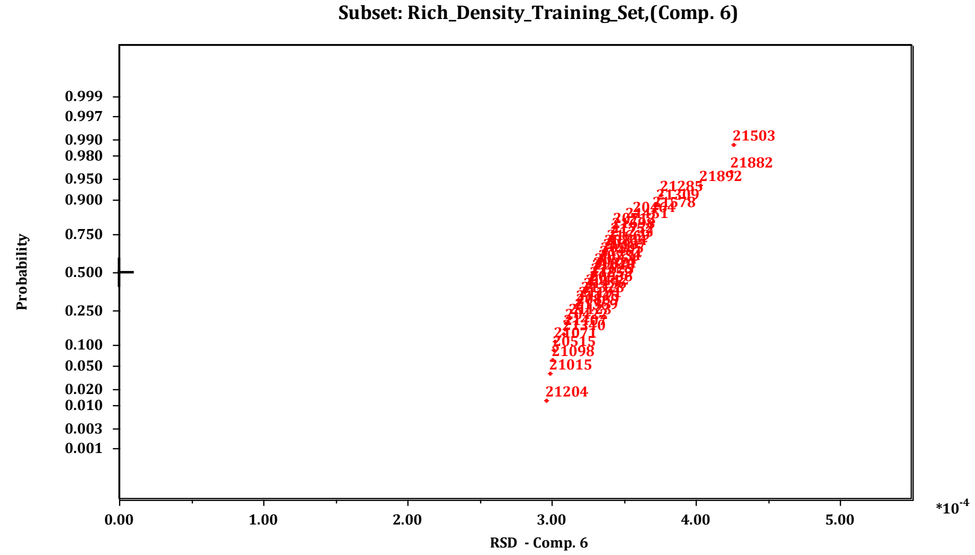


Figure 6-42 Normal plot of the residual standard deviation

The normal plot of the RSD (**figure 6-42**) for the validation set show some deviation from the straight line, these objects were examined further, using the score plot and the normal plot of the scores, but were not found to be outliers (these plots are not presented here). The model was also able to predict these with small errors, but of all the samples, these have the most significant deviations which explain why they deviate from the straight line. The overall validation parameters for the model is presented in the table below.

Table 6-10 Validation parameters for the model with density as dependent variable

Validation parameters for the PLS model	
Explained information, dependent variable	99.76%
Cross-validation on last component, CsvSD	0.83
Number of components in the model	6
RMSECV	0.749
RMSEP	0.649
R^2	0.991
Adjusted R^2	0.991
Average prediction error	0.046%
Average residual (absolute value)	0.525

The model is concluded to be good based on the overall results presented above. A small prediction error is achieved, and very small residuals considering the values are of such magnitude. The RMSECV value from regression analysis and the RMSEP value based on the prediction of the objects in the validation set is very similar, which also confirms that the model has good predictive performance. The RMSEP value is about 0.5% of the density values, which is an excellent result.

6.2.3 Total Alkalinity

The window chosen to be used for building this model is where the organic compounds absorb radiation. The C-H stretch for organic compounds is in the region 2927 – 2864 cm^{-1} [13]. The region from 1680-1280 cm^{-1} and 1100-1000 cm^{-1} are also included based on the SR-plots which showed that these are the essential wavenumbers. The chosen spectral regions are displayed in **figure 6-43**.

The dataset for TOT_ALK was meager; it contains only 32 samples, so the entire dataset was used to build a model. There are no data left to use as a validation set; validation was therefore performed using the results from the regression analysis and double cross-validation. The average value of TOT_ALK is 4.6 mole/kg for the rich samples, which is a slightly lower than for the lean samples (more in section 3.2).

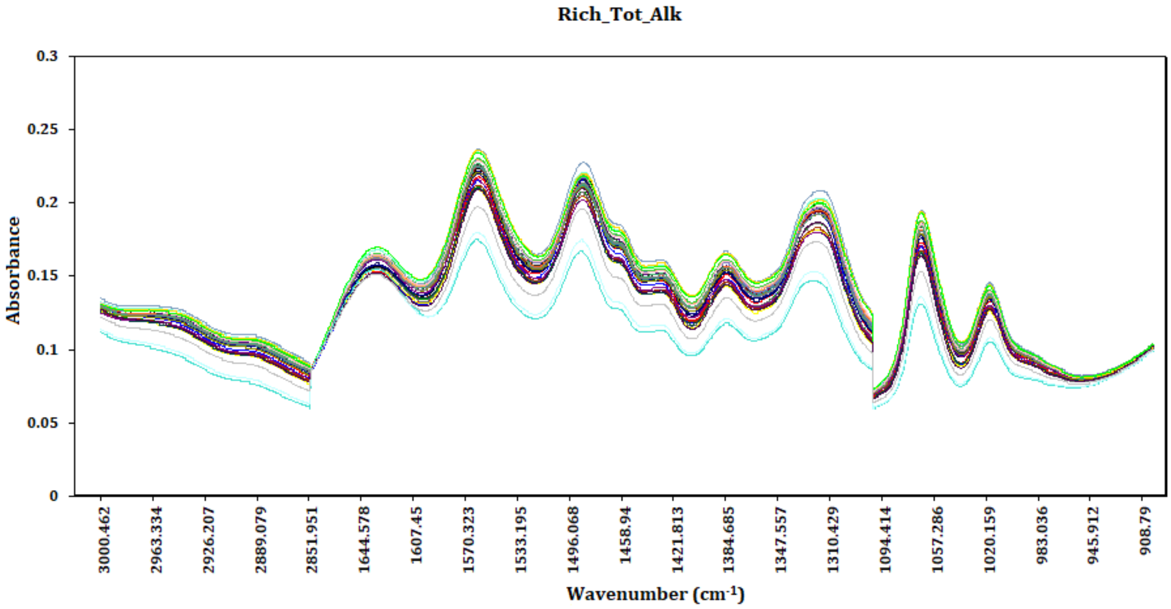


Figure 6-43 Fragment of raw spectra for the lean samples used to build a model with TOT_ALK as response variable

Several different preprocessing techniques have been tested and evaluated, second order Savitzky-Golay with a window of 21 and a third-degree polynomial gave the best results. This conclusion was made based on the RMSECV values, cross-validation values, the number of outliers and the number of components included in the models.

Savitzky-Golay needs continuous spectra, so the preprocessing was performed before the data was narrowed to the chosen regions. The outlier detection is performed after the preprocessing, which resulted in fewer outliers than when performed before. The outliers were identified using score plots, RSD vs. Leverage and Scores vs. Objects obtained from PCA (attached in Appendix C3). The most extreme outliers are removed first, as this will dramatically change the model. The twelve identified outliers are presented in the **table 6-11**.

Table 6-11 Outliers identified using PCA

Identified outliers
20400, 20438, 20712, 20821, 21539, 21578, 21664, 21882, 21887, 21970, 21972, 21989

After preprocessing with Savitzky-Golay, the spectra look very different, as can be seen in **figure 6-44**. The region in the highest wavenumbers show some noise, same as for the lean samples, and a model without this region was investigated. The model with this area included resulted in the best model and is the one displayed here.

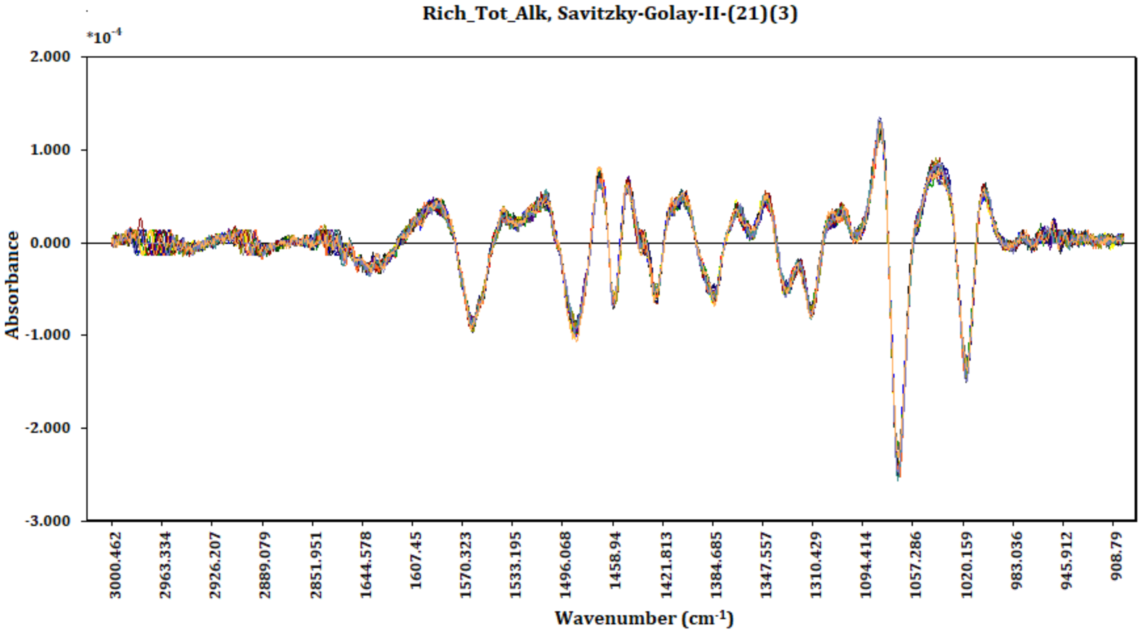


Figure 6-44 Spectra for the chosen region of wavenumbers after performing second-order Savitzky-Golay with a window of 21 and a third-degree polynomial as preprocessing

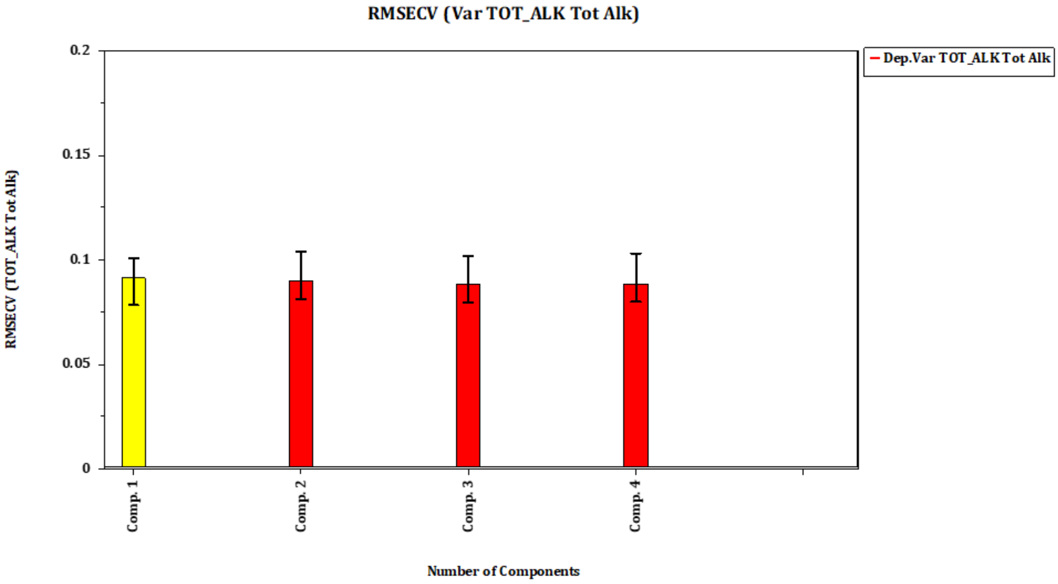


Figure 6-45 RMSECV-plot for the first four components extracted, the yellow bar indicates that one component should be included

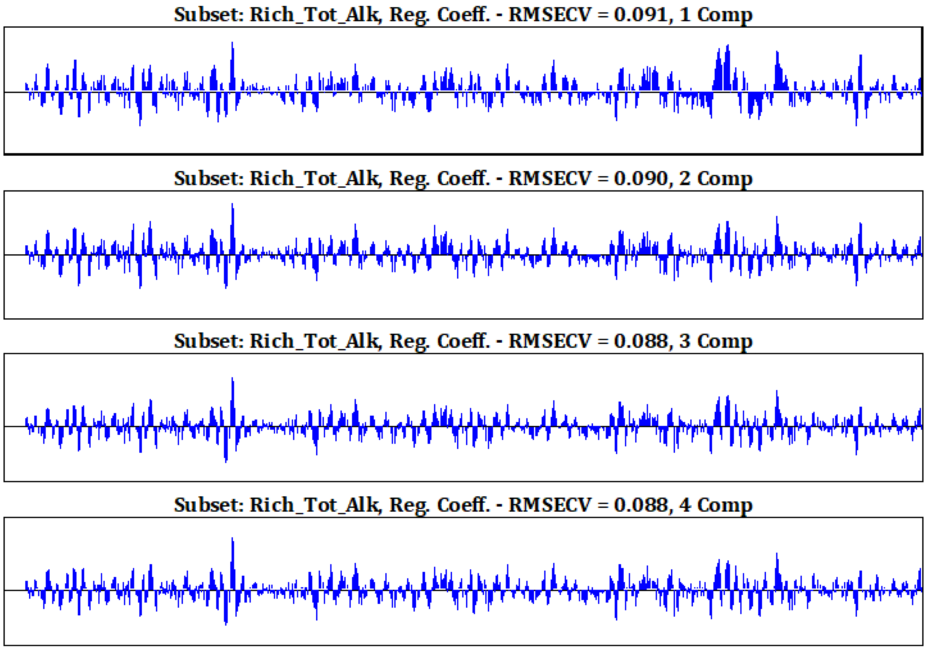


Figure 6-46 Weighted regression coefficients for the first four components extracted

The model dimensions plot in **figure 6-45** suggest that one component should be included in the model, but after investigating the weighted regression coefficients, the RMSECV values and cross-validation values for each component, it was concluded to include three components in the model. Investigation of the Scores vs. Objects, score plots and Predicted vs. Measured plots also suggested that three components should be included in the model.

The weighted regression coefficient plots in **figure 6-46** show that the components describe some noise, and the amount of described independent information by the components is only about 50% in total when using three components but using other preprocessing techniques and variable selection did not improve the model. Other windows of wavenumbers were also tested, but this did not result in better models. Because the dataset contains so few variables, the number of variables after removing the outliers is even smaller, and it is probably not enough data to create a better model.

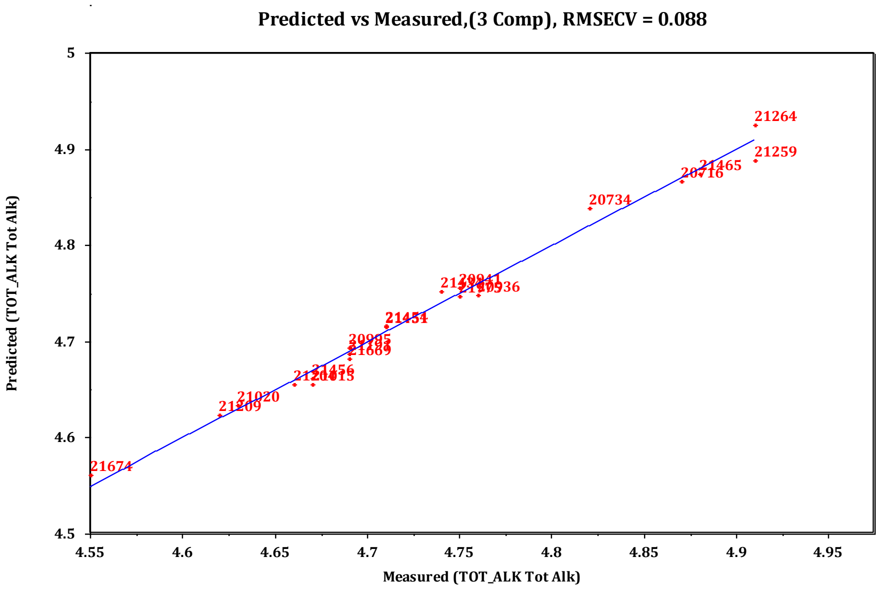


Figure 6-47 Predicted vs. Measured from regression analysis, displaying the RMSECV value

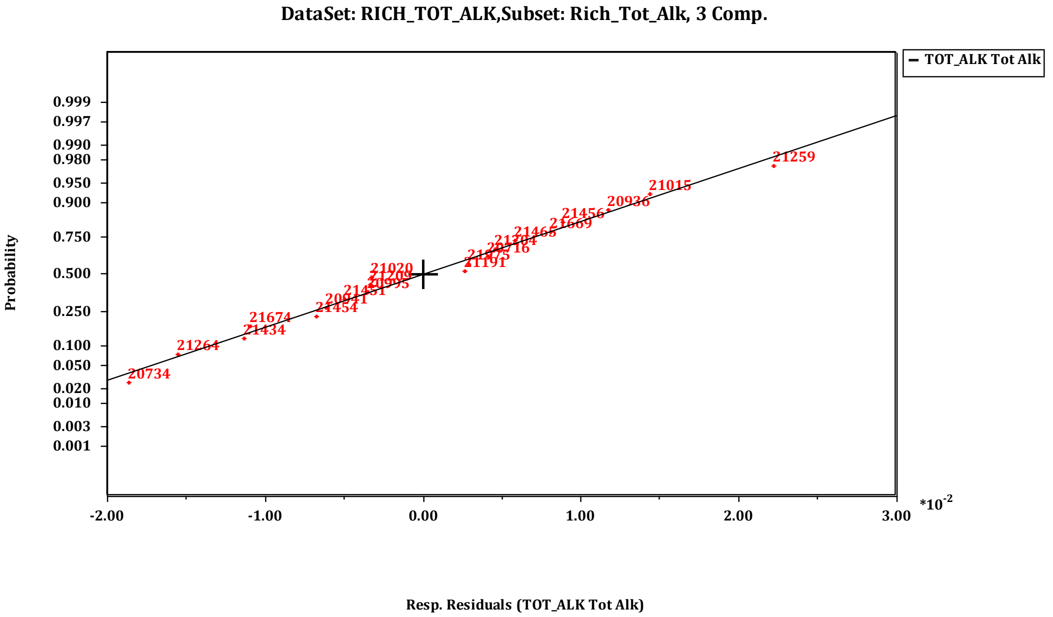


Figure 6-48 Normal plot of the response residuals from the regression analysis

The Predicted vs. Measured plot (**figure 6-47**) from the regression analysis looks good and display a good correlation between the predicted and measured values with a decent RMSECV value. The response residuals are investigated as well, displayed in **figure 6-48**, and shows good linearity which together with the Predicted vs. Measured plot leads to the conclusion that the model is performing satisfyingly. The validation parameters for the model is presented in **table 6-12**.

Table 6-12 Validation parameters for the model with TOT_ALK as dependent variable

Validation parameters	
Explained information, dependent variable	94.52%
Cross-validation on last component, CsvSD	0.98
Number of components in the model	3
RMSECV	0.088
RMSEP, *from double cross-validation*	0.088
R^2	0.995
Adjusted R^2	0.994

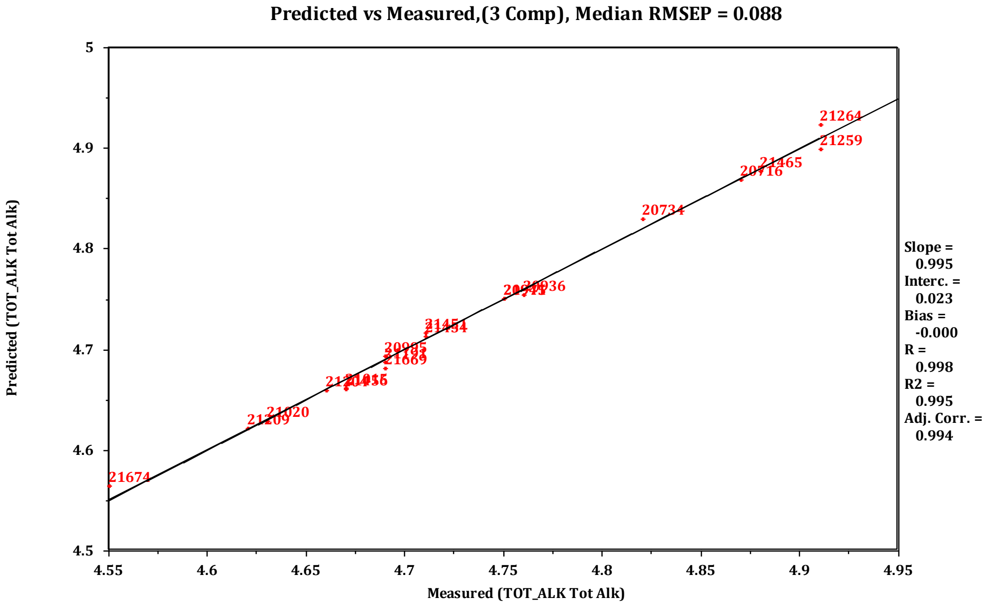


Figure 6-49 Predicted vs. Measured from double cross-validation, displaying the RMSEP value, R^2 and R_a^2

From the double cross-validation, a Predicted vs. Measured plot (**figure 6-49**) is achieved and confirms the conclusions from the regression analysis. The coefficient of multiple determination and the adjusted coefficient is approximately the same, which also argues for a decent model with the right number of components. The RMSEP value from the double cross-validation and the RMSECV value from the regression analysis is equal and is about 1% of the measured values of TOT_ALK, which also argues for a good model.

7 Conclusion

This thesis aimed to build models that were able to predict total inorganic carbon, total alkalinity, and density. Multivariate data analysis methods have been utilized to examine the data and create models that can accurately predict the responses. Preliminary analysis of all the samples showed a significant difference in the lean and rich samples, which lead to the conclusion to build separate models for the lean and rich samples.

The results from PCA of all the samples in the dataset showed that the score plots effectively can be employed to identify objects with deviating behavior. The models were built using regression PLS and optimized using variable section in the form of SR and VIP.

Analysis of the different models showed that the models were better on narrow windows of wavenumbers, based on where the compounds being analyzed had the most intense absorption. For each response, the same region gave good results for the lean and rich samples. The region used for TIC is 1670 to 1000 cm^{-1}, which is chosen based on the stretching vibrations of the inorganic carbons. TOT_ALK is modeled using three regions, based on the absorption of organic carbons; 3000 to 2800 cm^{-1}, 1680 to 1280 cm^{-1} and 1100 to 1000 cm^{-1}. The models for density gave the best results in the fingerprint region of the spectrum; 1670 to 1000 cm^{-1}.

Chemical absorption is extensively studied, especially using amines. Most researchers have focused on the MEA-concentration and CO_2 concentration, while few models exist for the responses explored in this thesis. The use of ATR-FTIR spectrometry and multivariate analysis to study the MEA- and CO_2 concentration has been explored before and has provided good results [42]. The PLS models are built using mean centering and normalization as preprocessing, using two regions of the spectra. The spectral regions used was 2730 to 3760 cm^{-1} and 770 to 1760 cm^{-1}, which corresponds to the regions used in this thesis. The results from the PCA showed groupings of the samples, depending on the concentration of CO_2 and MEA. The same groupings are found on the samples used in this thesis when studying the score plots for both lean and rich samples.

Preprocessing of the spectra was needed to avoid effects that do not represent the chemical variation, EMSC and Savitzky-Golay were found to be methods that improved the model performance significantly.

The models for TIC, for both lean and rich samples, was made with data preprocessed with second-order EMSC. The model for the lean samples was made with a training set and validated using the objects omitted from the training set. The resulting model had a prediction error of 1.2% and small residuals, and with a RMSEP value of 1-2% of the measurement ranges, the model is concluded to be good.

The dataset for the rich samples was meager, and all objects were included in the calibrations. This model was optimized using variable selection, in the form of VIP, which improved the results. Validation was done with double cross-validation, which gave a RMSEP value equal to the RMSECV from the regression analysis, ranging from 1-2% of the TIC values. The model is expected to perform well on new samples.

For the rich samples density was predicted with a prediction error of 0.05%, resulting in great predictions and small residuals. Second order EMSC was applied as preprocessing. The density predictions for the lean samples was not as good, but still satisfying, with a prediction error of 1.2%. The preprocessing used for this model was second order Savitzky-Golay.

The two datasets for TOT_ALK were too small to be divided into training sets and validation sets, so the models are made with all the objects available. Both lean and rich samples are preprocessed using second-order Savitzky-Golay. The model for the rich samples gave equal RMSECV and RMSEP values, about 2% of the measured values, obtained from regression analysis and double cross-validation, respectively.

For the lean samples, the response variable was transformed with root ½ to achieve normal distribution. This resulted in equal RMSECV and RMSEP values, of 0.024, which is about 0.5-1% of the measured values, which is a satisfying result. Both models look promising and are expected to perform well on other samples, but without validations set, it is difficult to conclude that the models will perform well.

In summary, the use of ATR-FTIR spectroscopy equipped with multivariate data analysis has proved to be satisfactory techniques to monitor the compounds present during the CO_2 capture process using amines. This approach is capable of predicting the responses with reasonably good accuracy and will be a good tool in the online reaction monitoring, improving the knowledge and monitoring of the process while it is happening.

7.1 Further Work

Three of the datasets did not have enough samples to be divided into a training set and a validation set, so these models were validated using the results from regression analysis and double cross-validation. When more samples become available, these models should be tested to see if they have as good prediction abilities as assumed. Moreover, if not, more samples should be included, and new models should be made, which hopefully will give models capable of making satisfactory predictions.

The degradation products from the absorption process have not been studied in this thesis because there were not enough data to do this. Continued work should include studying these and making models for predicting how much degradation product is in the solvent after using it several times, the presence of degradation products can reduce the solvents abilities to absorb CO_2 and should, therefore, be monitored during the process.

8 References

[1] Baird, C. & Cann, M. C. (2012). *Environmental chemistry* (5th ed). New York: W.H. Freeman and Co.

[2] Dutcher, B., Fan, M. & Russell, A.G. (2015). Amine-based CO_2 capture technology development from the beginning of 2013 - A Review. *Acs Applied Materials & Interfaces*, *7*(4), 2137–2148.

[3] Kachko, A., Ham, L. V., Bardow, A., Vlugt, T. J. & Goetheer, E. L. (2016). Comparison of Raman, NIR and ATR-FTIR spectroscopy as analytical tools for in-line monitoring of CO_2 concentration in an amine gas treating process. *International Journal of Greenhouse Gas Control*, *47*, 17-24.

[4] Ball, D. W. (2006). The electromagnetic spectrum. In *Field guide to spectroscopy* (Vol. FG08). Bellingham, Washington: SPIE Press, p. 6.

[5] Nortvedt, R. & Kvaal, K. (1996). Vurdering av næringsmiddelkvalitet. In: R., Nortvedt, F., Brakstad, O. M., Kvalheim & T., Lundstedt, *Anvendelse av kjemometri innen forskning og industri* (pp. 363-379). Oslo: Tidsskriftforlaget Kjemi.

[6] Koç, M. & Karabudak, E. (2017) History of spectroscopy and modern micromachined disposable Si ATR-IR spectroscopy. *Applied Spectroscopy Reviews*, *53*(5), 420-438.

[7] Karstang, T. V. (1996). Forbehandling av Data. In: R., Nortvedt, F., Brakstad, O. M., Kvalheim & T., Lundstedt, *Anvendelse av kjemometri innen forskning og industri* (pp. 129-144). Oslo: Tidsskriftforlaget Kjemi.

[8] Martens, H., Nielsen, J.P. & Engelsen, S.B. (2003). Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. *Analytical chemistry*, *75*(3), 394–404.

[9] Olawumi, T. T. (n.d.). Schematic representation of the different molecular vibration modes showing bending and stretching vibrations. Retrieved April 23, 2018, from https://www.researchgate.net/figure/Schematic-representation-of-the-different-molecular-vibration-modes-showing-bending-and_fig9_275583514.

[10] Atkins, P. W. & Paula, J. D. (2014). *Atkins physical chemistry* (10th ed.), Oxford: Oxford University Press.

[11] Larkin, P. (2011). *Infrared and Raman spectroscopy: principles and spectral interpretation*. Amsterdam Netherlands: Elsevier.

[12] Sun, D. (2009). Fourier Transform Infrared (FTIR) Spectroscopy. In *Infrared spectroscopy for food quality analysis and control.* Amsterdam: Academic Press/Elsevier.

[13] Silverstein, R. M., Webster, F. X. & Kiemle, D. J. (2005). Infrared spectrometry. In *Spectrometric identification of organic compounds* (7th ed.). Hoboken, NJ: John Wiley & Sons.

[14] Stuart, B. H. (2004). *Infrared spectroscopy: Fundamentals and applications*. Chichester, West Sussex: Wiley.

[15] Creamer, A. E. & Gao, B. (2015). *Carbon dioxide capture: An effective way to combat global warming*. Cham, Switzerland: Springer International Publishing.

[16] Kenarsari, S. D., Yang, D., Jiang, G., Zhang, S., Wang, J., Russell, A. G., . . ., Fan, M. (2013). Review of recent advances in carbon dioxide separation and capture. *The Royal Society of Chemistry*, *3*, 22739-22773.

[17] Stowe, H. M. & 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), 6887–6899.

[18] Eğe, S. N. (2004). *Organic chemistry: Structure and reactivity* (5th ed.). Boston, MA: Houghton Mifflin.

[19] Weissermel, K. & Arpe, H-J. (2003). *Industrial Organic Chemistry* (4th ed.). Weinheim, Germany: Wiley-VCH.

[20] Hwang, G. S., Stowe, H. M., Paek, E., & Manogaran, D. (2014). Reaction mechanism of aqueous Monoethanolamine with carbon dioxide: A combined quantum chemical and molecular dynamics study. *Physical Chemistry Chemical Physics*, *17*(2), 831-839.

[21] Tati, P., Buschle, B., Milkowski, K., Akram, M., Pourkashanian, M. & Lucquiaud, M. (2018). Flexible operation of post-combustion CO_2 capture at pilot scale with demonstration of capture-efficiency control using online solvent measurements. *International Journal of Greenhouse Gas Control*, *71*, 253-277.

[22] Einbu, A., Citfja, 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*, 55–63.

[23] Kvalheim, O. M. (1996). Fra data til informasjon. In: R., Nortvedt, F., Brakstad, O. M., Kvalheim & T., Lundstedt, *Anvendelse av kjemometri innen forskning og industri* (pp. 53-65). Oslo: Tidsskriftforlaget Kjemi.

[24] Rajalahti, T. & Kvalheim, O. M. (2011). Multivariate data analysis in pharmaceutics: A tutorial review. *International Journal of Pharmaceutics*, *417*(1-2), pp.280–290.

[25] Rinnan, Å. (2014). Pre-processing in vibrational spectroscopy when, why and how. *Analytical Methods*, *6*(18), 7124–7129.

[26] Steinier, J., Termonia, Y. & Deltour, J. (1972). Smoothing and differentiation of data by simplified least square procedure. *Analytical Chemistry*, *44*(11), 1906-1909.

[27] Byrne, H. J., Knief, P., Keating, M. E. & Bonnier, F. (2016). Spectral pre and post processing for infrared and Raman spectroscopy of biological tissues and cells. *Chemical Society Reviews*, *45*(7), 1865–1878.

[28] Afseth, N. K. & Kohler, A. (2012). Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemometrics and Intelligent Laboratory Systems*, *117*, 92–99.

[29] Isaksson, T. & Næs, T. (1996). Prinsipal komponent analyse. In: R., Nortvedt, F., Brakstad, O. M., Kvalheim & T., Lundstedt, *Anvendelse av kjemometri innen forskning og industri* (pp. 145-151). Oslo: Tidsskriftforlaget Kjemi.

[30] Bro, R. & Smilde, A.K. (2014). Principal component analysis. *Analytical Methods*, *6*(9), 2812–2831.

[31] Tranter, R.L. (2000). *Design and analysis in chemical research.* Sheffield: Sheffield Academic Press.

[32] Grung, B. & Kvalheim, O. M. (1994). Rank determination of spectroscopic profiles by means of cross validation. *Chemometrics and Intelligent Laboratory Systems*, *22*(1), 115-125.

[33] 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), 529–541.

[34] Gil, J.A. & Romera, R. (1998). On robust partial least squares (PLS) methods. *Journal of Chemometrics*, *12*(6), pp.365–378.

[35] Filzmoser, P. et al., 2009. Repeated double cross validation. *Journal of Chemometrics*, *23*(4), 160–171.

[36] Devore, J. L. & Berk, K. N. (2012). *Modern Mathematical Statistics with Applications*, New York, NY: Springer.

[37] Andersen, C.M. & Bro, R. (2010). Variable selection in regression—a tutorial. *Journal of Chemometrics*, *24*(11-12), 728–737.

[38] Farrés, M., Platikanov, S., Tsakovski, S. & Tauler, R. (2015). Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. *Journal of Chemometrics*, *29*(10), 528–536.

[39] Kvalheim, O. M., Chan, H., Benzie, I. F., Szeto, Y., Tzang, A. H., Mok, D. K. & Chau, F. (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), 98–105.

[40] Rajalahti, T., Arneberg, R., Kroksveen, A. C., Berle, M., Myhr, K. & Kvalheim, O. M. (2009). Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. *Analytical Chemistry*, *81*(7), 2581–90.

[41] 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), 985–994.

9 Appendices

A. CO_2 Capture Reactions

The steps of CO_2 capture using MEA, with the zwitterion mechanism [15]:

a) CO_2 binding with MEA, and formation of the zwitterionic adduct:

$\mathrm{MEA} + \mathrm{CO_2} \rightarrow \mathrm{MEA^+COO^-}$

b) Formation of carbamate and solvated proton, by deprotonation:

$\mathrm{MEA^+COO^-} + \mathrm{H_2O} \rightarrow \mathrm{MEACOO^-} + \mathrm{H_3O^+}$

c) Formation of protonated MEA:

$\mathrm{MEA} + \mathrm{H_3O^+} + \mathrm{MEACOO^-} \rightarrow \mathrm{MEAH^+} + \mathrm{H_2O} + \mathrm{MEACOO^-}$

d) Formation of carbamic acid:

$\mathrm{MEACOO^-} + \mathrm{H_3O^+} \rightarrow \mathrm{MEACOOH} + \mathrm{H_2O}$

B. Lean Samples

1. Total Inorganic Carbon

Results from the regression analysis using Savitzky-Golay as preprocessing:

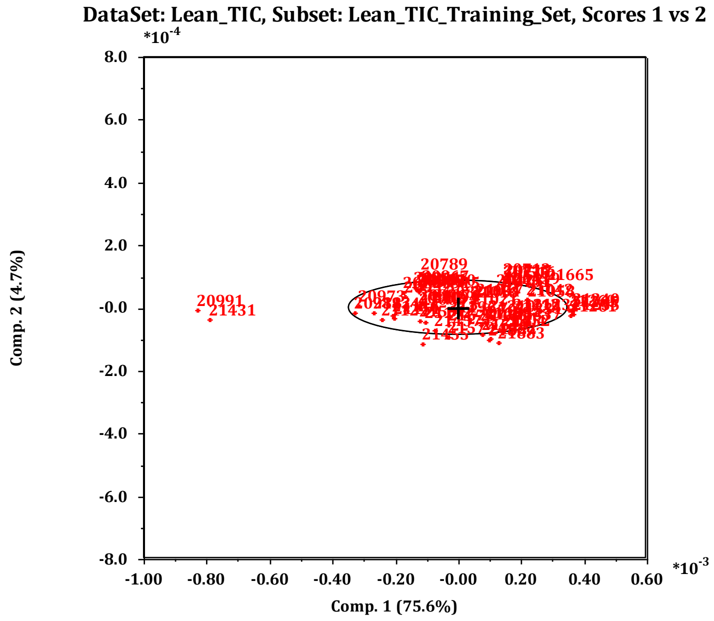


Figure 9-1 Score plot for component 1 and 2, obtained from PCA. The ellipse is provided by Sirius and is used to identify outliers, objects lying outside is identified as outliers. Objects 20991 and 21431 are outliers

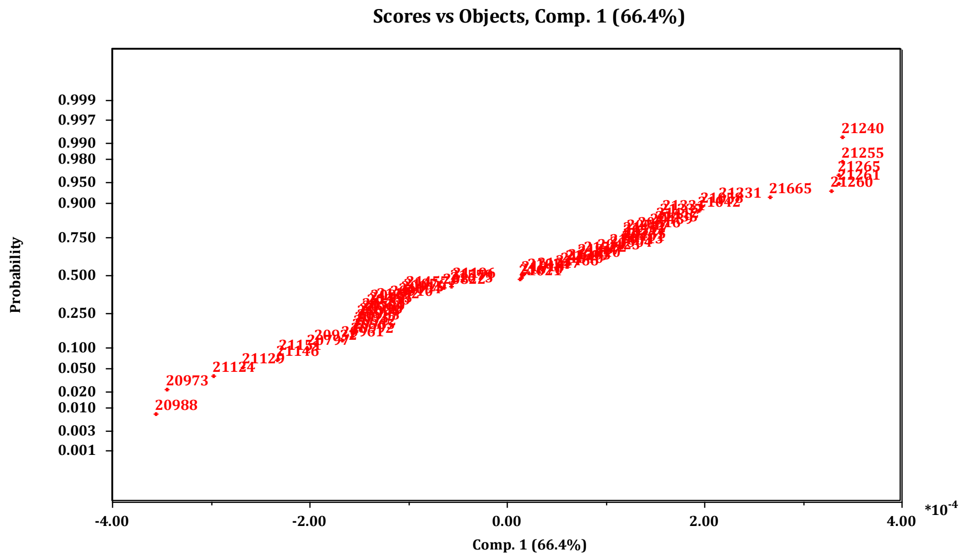


Figure 9-2 Scores vs. Objects for component 1, objects deviating the straight line are identified as outliers

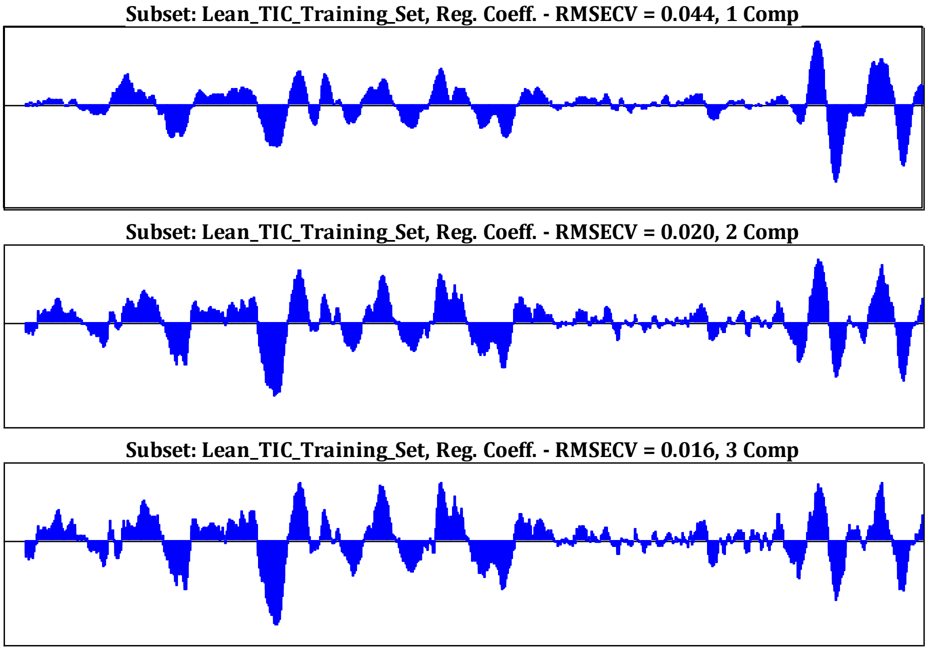


Figure 9-3 Weighted regression coefficients for the three components included in the model. Displaying that the coefficients describe data and not just noise

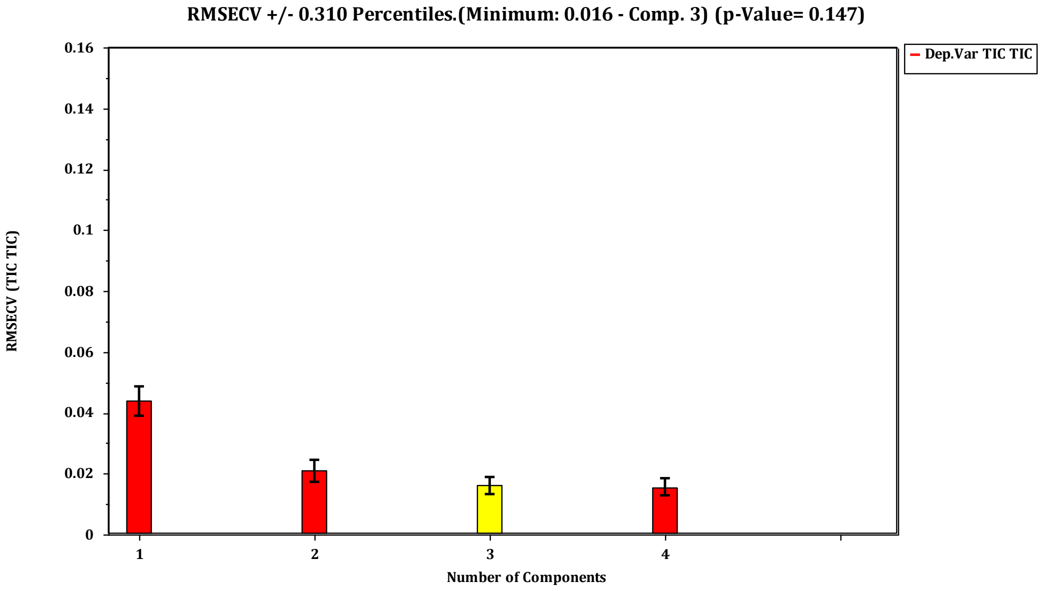


Figure 9-4 RMSECV-plot for the first four component extracted. Used to determine the number of components to include in the model, the yellow bar indicate that three components should be included

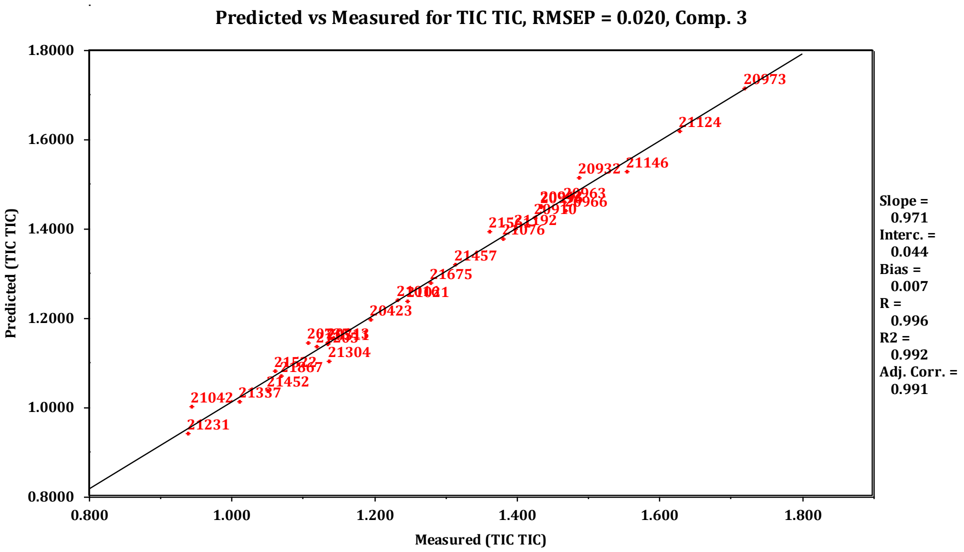


Figure 9-5 Predicted vs. Measured for the validation set, displaying the RMSEP value, R^2 and R_a^2

2. Density

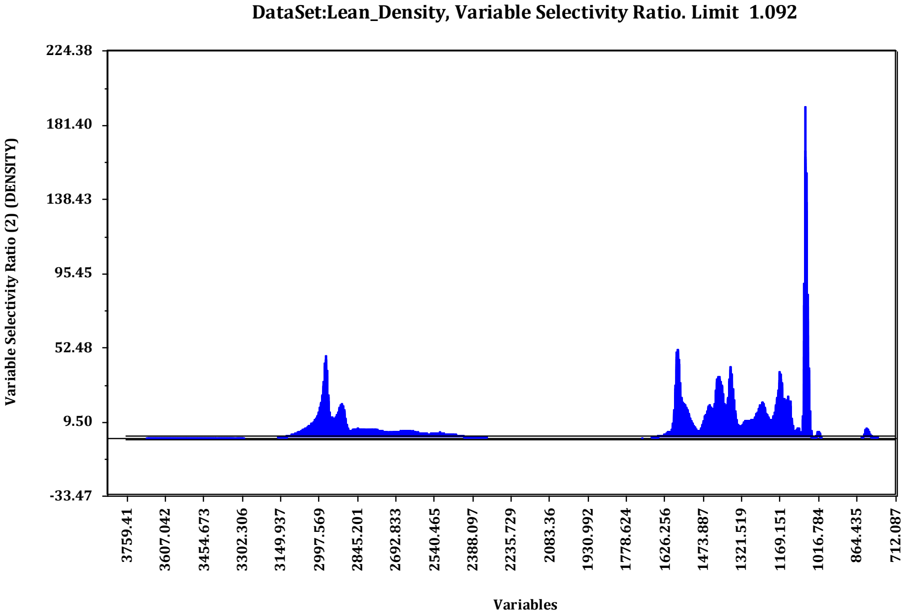


Figure 9-6 SR-plot used to determine which window of wavenumbers to use when building the model

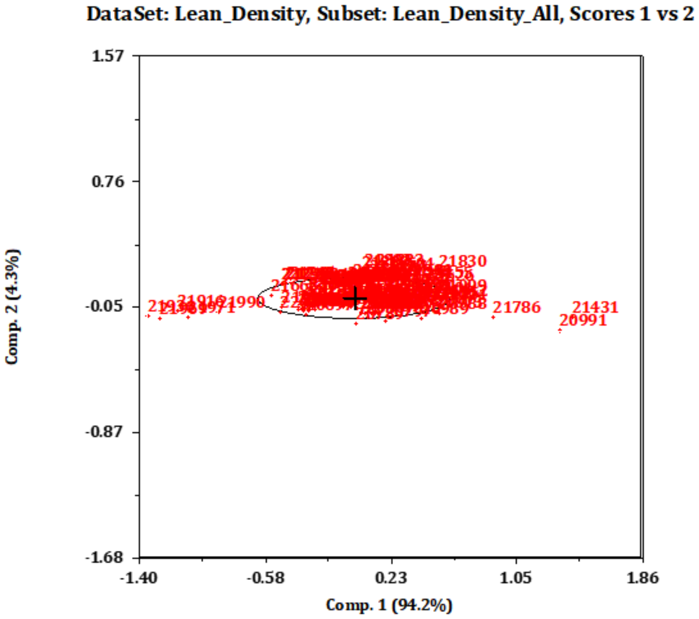


Figure 9-7 Score plot for component 1 and 2 from PCA, the ellipse is provided by Sirius, and objects lying outside this is identified as outliers

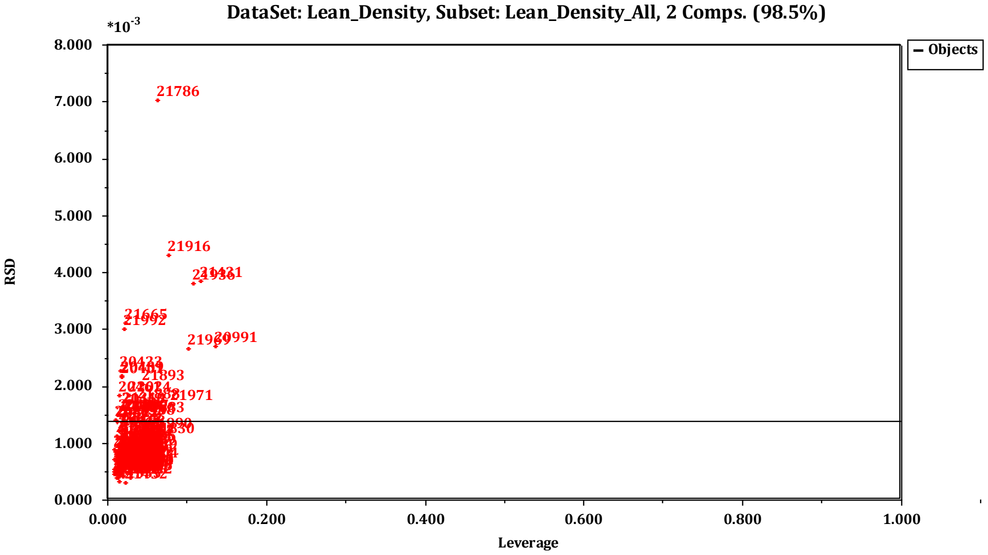


Figure 9-8 RSD vs. Leverage using two components, obtained from PCA. Used for outlier detection

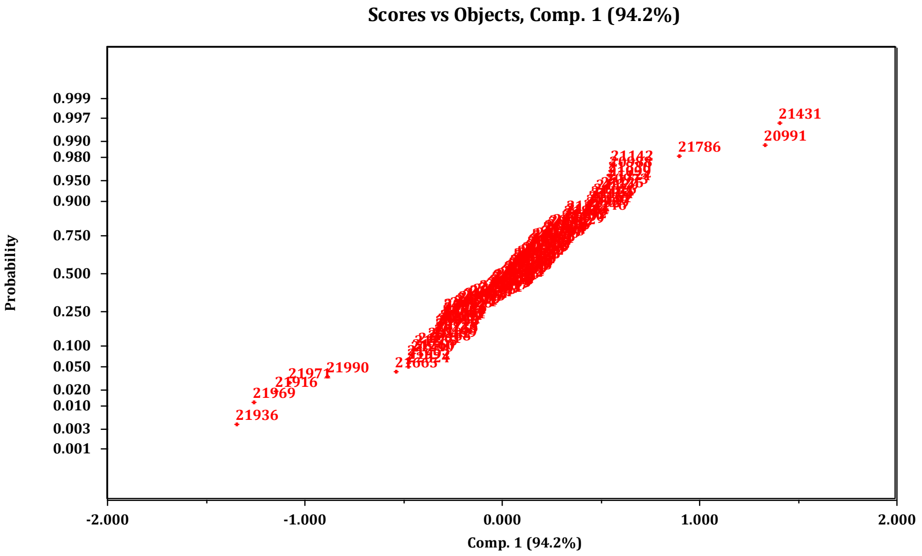


Figure 9-9 Scores vs. Objects for the first component, used for outlier detection. Objects disobeying the straight line are identified as outliers

3. Total Alkalinity

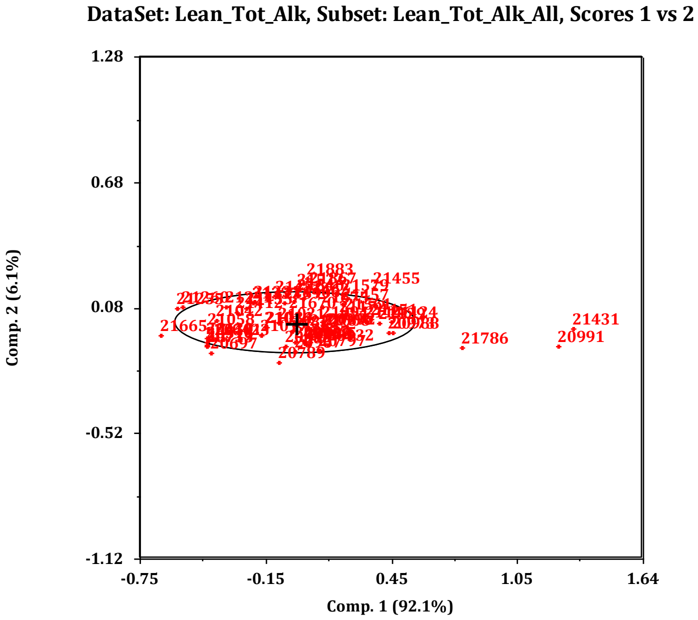


Figure 9-10 Score plot for component 1 and 2, used for outlier detection. The objects outside the ellipse is identified as outliers

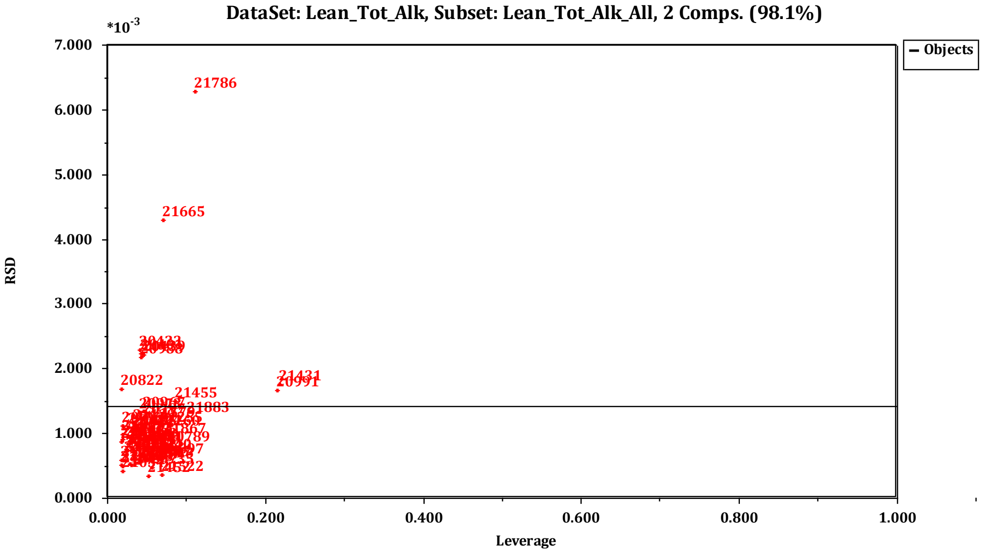


Figure 9-11 RSD vs. Leverage using two components, used for outlier detection

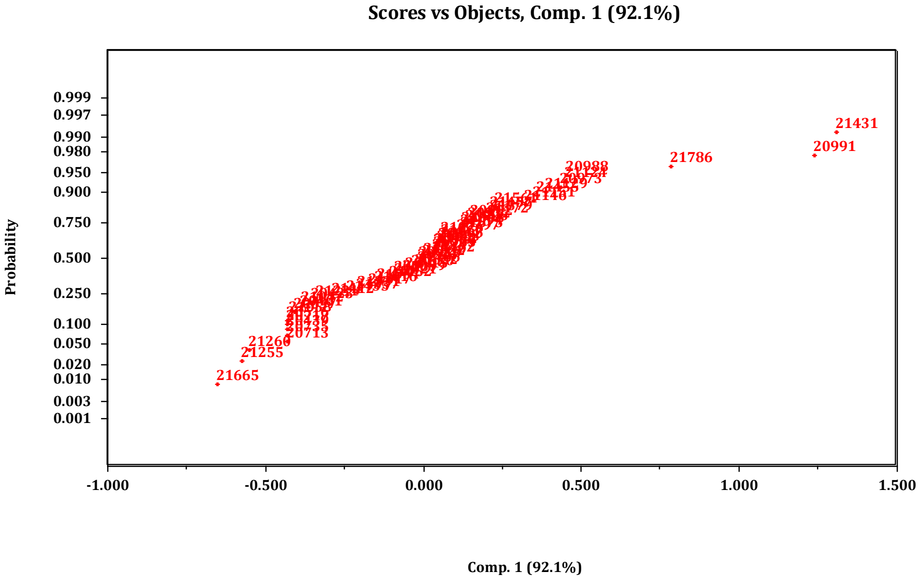


Figure 9-12 Scores vs. Objects for component 1, objects deviating the straight line are identified as outliers

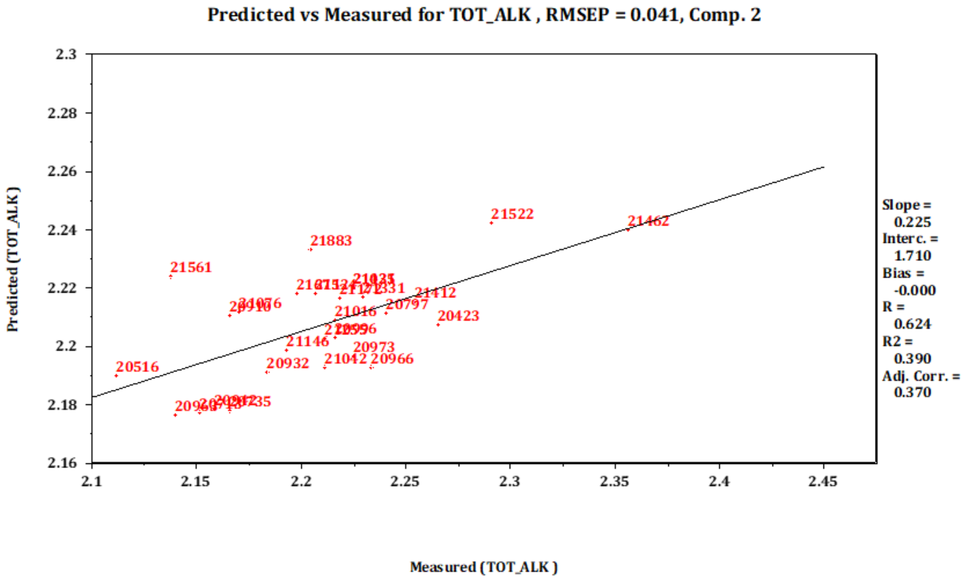


Figure 9-13 Predicted vs. Measured for the model built when the data was split into training set and validation set, displaying the RMSEP value, R^2 and R_a^2

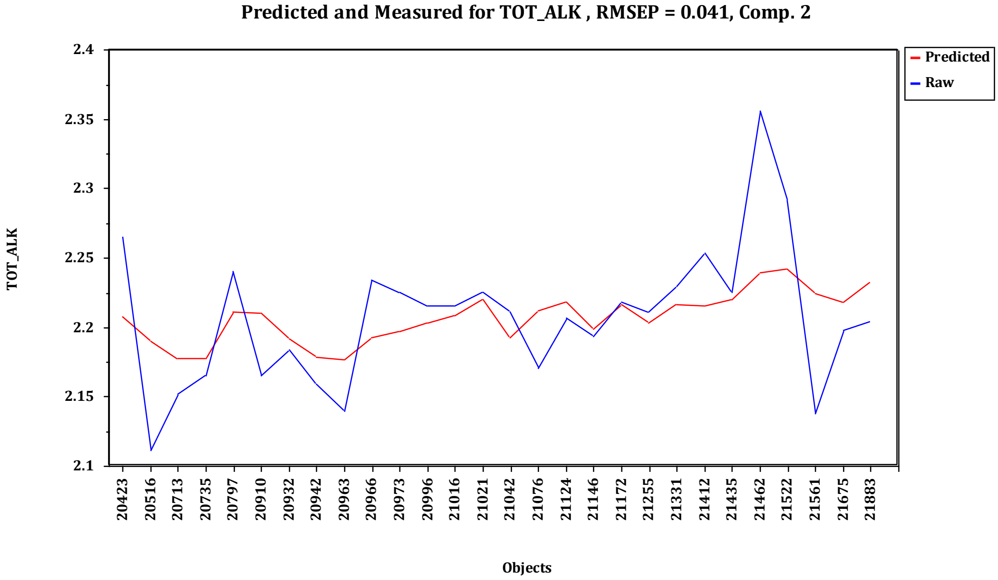


Figure 9-14 Predicted and Measured for the model built when the data was split into training set and validation set. Displaying that the predictions only is able to predict the average values of the measured values

C. Rich Samples

1. Total Inorganic Carbon

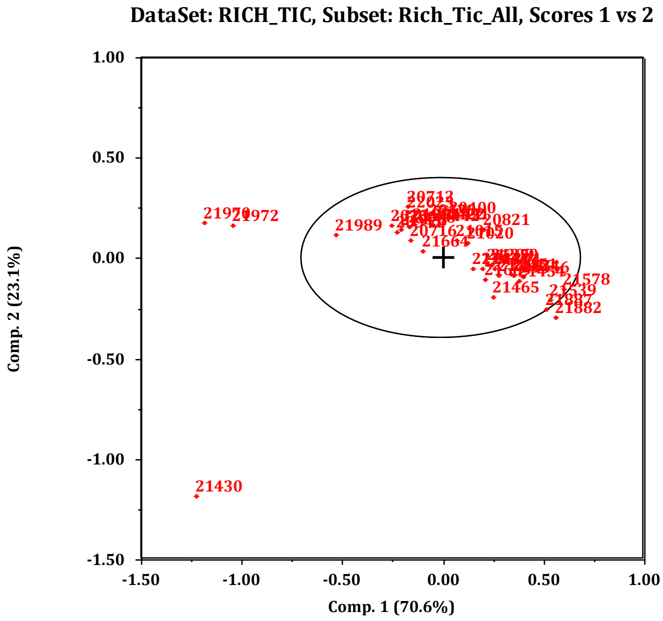


Figure 9-15 Score plot for component one and two, used for outlier detection. The objects outside the ellipse are identified as outliers

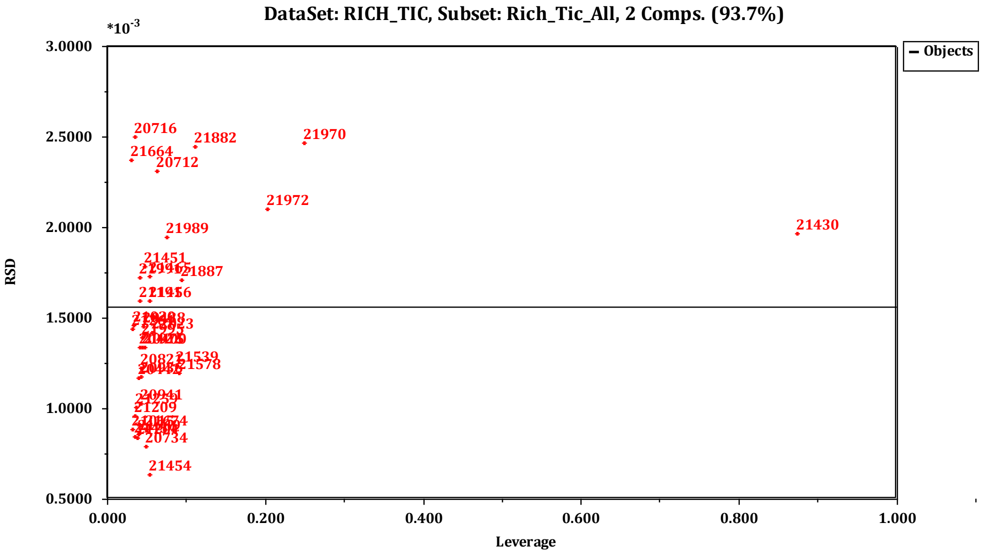


Figure 9-16 RSD vs. Leverage for two components, used for outlier detection

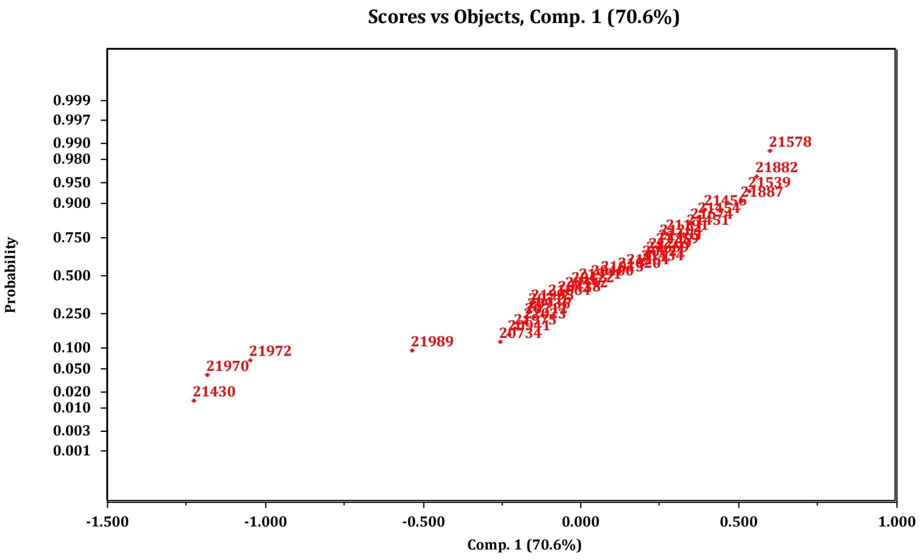


Figure 9-17 Scores vs. Objects component one, objects deviating the straight line are identified as outliers

2. Density

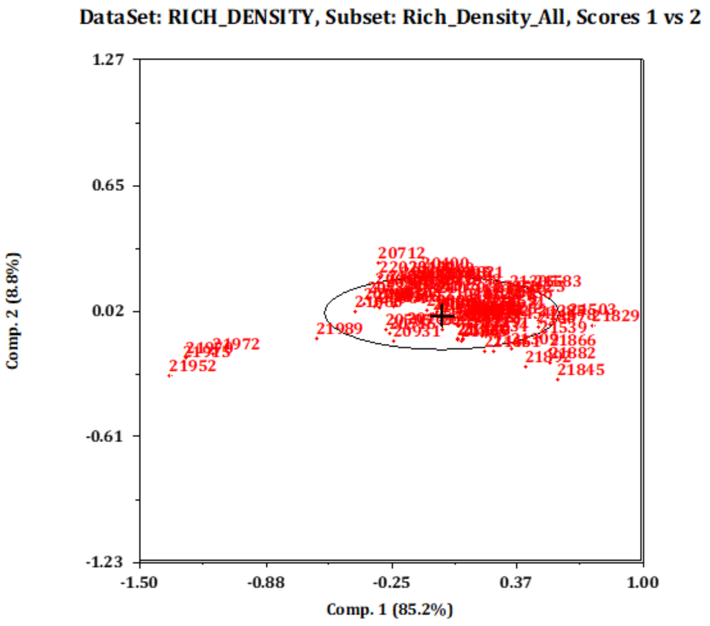


Figure 9-18 Score plot for component one and two, used for outlier detection. Objects outside the ellipse are identified as outliers

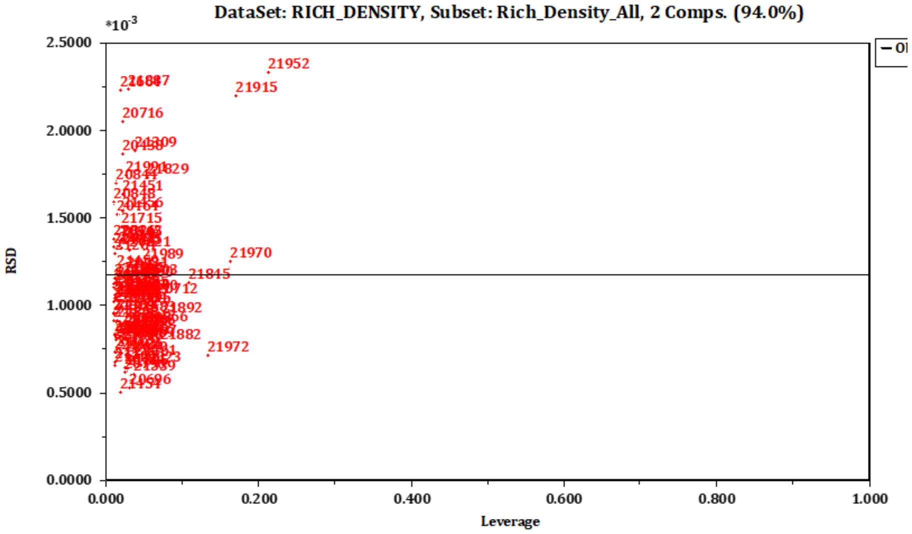


Figure 9-19 RSD vs. Leverage using two components, used for outlier detection

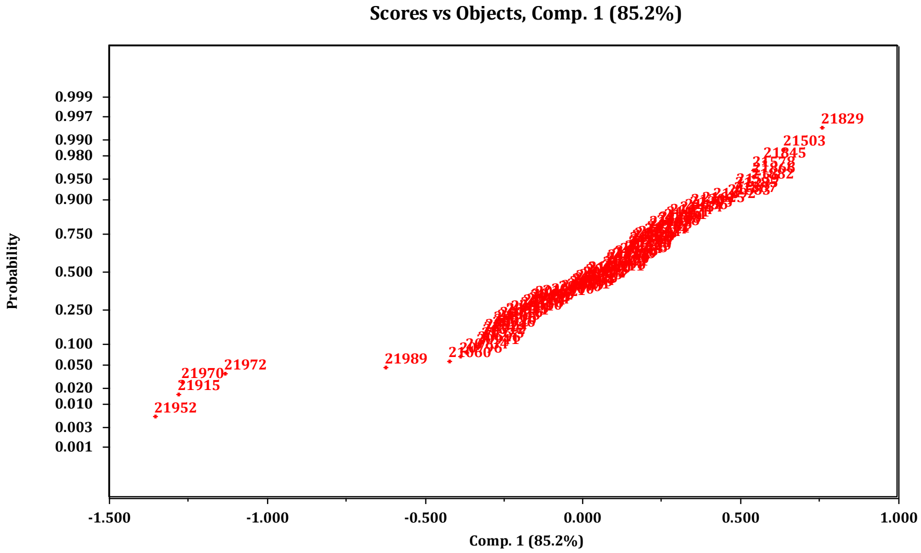


Figure 9-20 Scores vs. Objects normal plot component one. Objects disobeying the straight line are identified as ouliers

3. Total Alkalinity

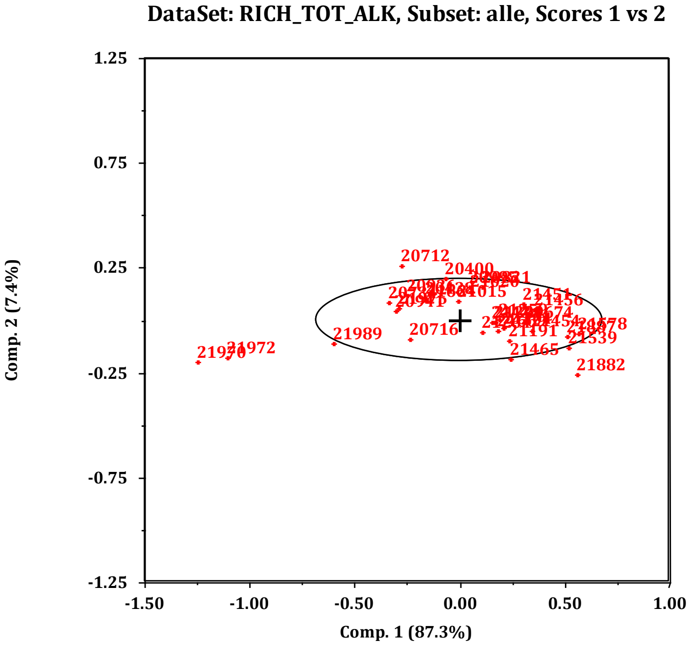


Figure 9-21 Score plot for component one and two, used for outlier detection. Objects outside the ellipse are identified as outliers

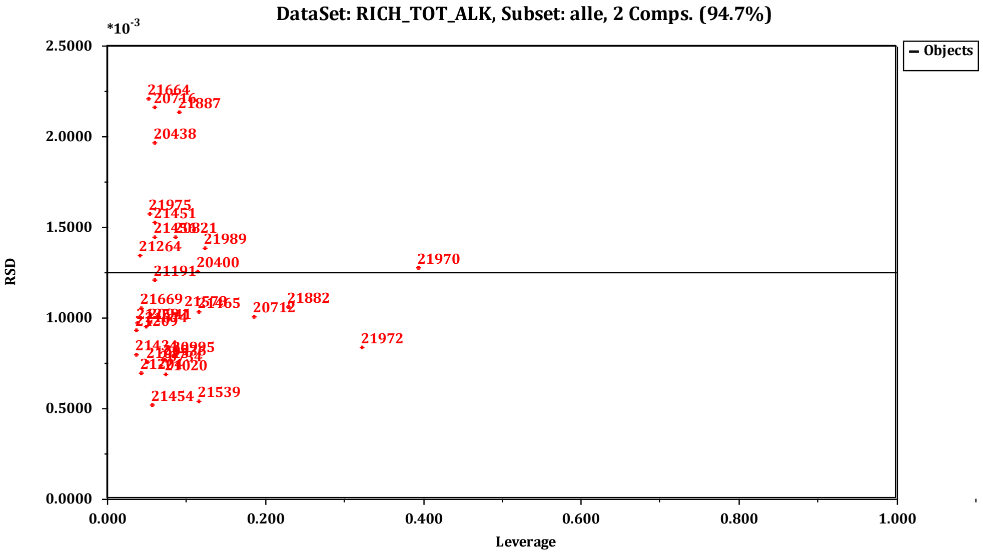


Figure 9-22 RSD vs. Leverage for two components, used for outlier detection

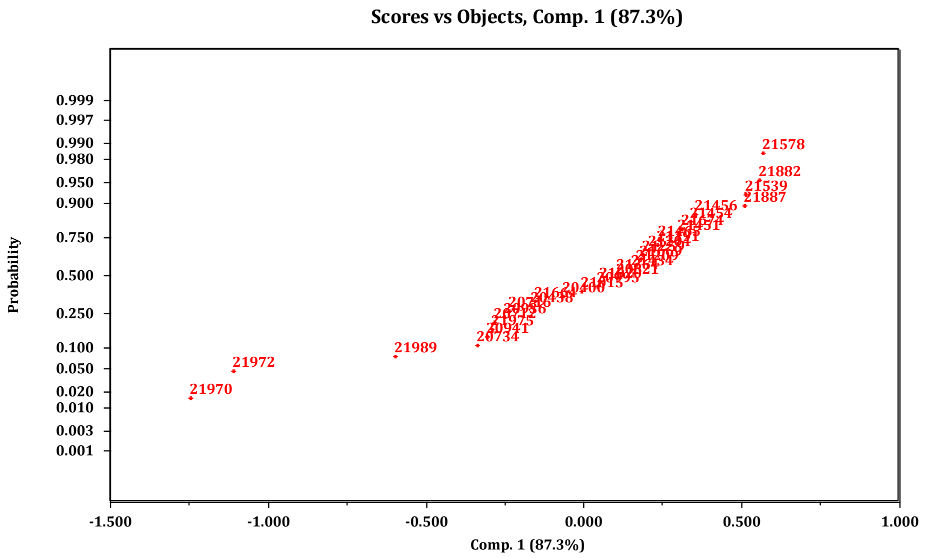


Figure 9-23 Scores vs. Objects