Multivariate data analysis using spectroscopic data of fluorocarbon alcohol mixtures

C Nothnagel

Dissertation submitted in partial fulfilment of the requirements for the degree Master of Science in Chemistry at the Potchefstroom campus of the North-West University

Supervisor:

Prof HM Krieg

Co-supervisor:

Prof SO Paul:

Assistant supervisor:

Dr LD Kock

May 2012

Declaration:

I, the undersigned, hereby declare that the work presented in this dissertation has not already been submitted to this or any other University and represents independent work by myself.

Signature: 

Date: 18 November, 2011

Abstract

Pelchem, a commercial subsidiary of Necsa (South African Nuclear Energy Corporation), produces a range of commercial fluorocarbon products while driving research and development initiatives to support the fluorine product portfolio. One such initiative is to develop improved analytical techniques to analyse product composition during development and to quality assure produce.

Generally the C-F type products produced by Necsa are in a solution of anhydrous HF, and cannot be directly analyzed with traditional techniques without derivatisation. A technique such as vibrational spectroscopy, that can analyze these products directly without further preparation, will have a distinct advantage. However, spectra of mixtures of similar compounds are complex and not suitable for traditional quantitative regression analysis. Multivariate data analysis (MVA) can be used in such instances to exploit the complex nature of spectra to extract quantitative information on the composition of mixtures.

A selection of fluorocarbon alcohols was made to act as representatives for fluorocarbon compounds. Experimental design theory was used to create a calibration range of mixtures of these compounds. Raman and infrared (NIR and ATR-IR) spectroscopy were used to generate spectral data of the mixtures and this data was analyzed with MVA techniques by the construction of regression and prediction models. Selected samples from the mixture range were chosen to test the predictive ability of the models.

Analysis and regression models (PCR, PLS2 and PLS1) gave good model fits (R^2 values larger than 0.9). Raman spectroscopy was the most efficient technique and gave a high prediction accuracy (at 10% accepted standard deviation), provided the minimum mass of a component exceeded 16% of the total sample.

The infrared techniques also performed well in terms of fit and prediction. The NIR spectra were subjected to signal saturation as a result of using long path length sample cells. This

was shown to be the main reason for the loss in efficiency of this technique compared to Raman and ATR-IR spectroscopy.

It was shown that multivariate data analysis of spectroscopic data of the selected fluorocarbon compounds could be used to quantitatively analyse mixtures with the possibility of further optimization of the method. The study was a representative study indicating that the combination of MVA and spectroscopy can be used successfully in the quantitative analysis of other fluorocarbon compound mixtures.

Key terms

Chemometrics, Multivariate data analysis, Partial least squares regression, Principal component regression, Raman spectroscopy, near infrared spectroscopy (NIR), attenuated total reflectance infrared spectroscopy (ATR-IR), Fourier transform spectroscopy, Fluorocarbon alcohols.

Abbreviations and acronyms

Abbreviation/Acronym	**Meaning/Definition**
PC/PC's	Principal component/Principal components
PCA	Principal component analysis
PCR	Principal component regression
PLS	Partial least squares regression
RMSE	Root mean square error
RMSEC	Root mean square error of calibration
RMSEP	Root mean square error of prediction
IR / NIR	Infrared / near infrared
ATR-IR	Attenuated total reflection infrared
[X]	The concentration of a species X
SSR	Regression sum of squares
SSE	Error sum of squares
SSTO	Total sum of squares of deviation
R^2	Coefficient of multiple determination.
r^2	Coefficient of simple determination.
CLS	Classical least squares regression
MVA	Multivariate analysis/ Multivariate data analysis
SDev	Standard deviation

List of symbols of physical quantities

Symbol	Meaning	Unit (SI)
E	**Energy**	**Joule**
E	**Electric field vector***	**$V.m^{-1}$**
λ	**Wave length**	**m**
I	**Intensity**	**$J.sr^{-1}$**
C	**Concentration**	**$kg.m^{-3}$**
a	**Specific absorptivity coefficient**	**$m^2.kg^{-1}$**
A	**Absorbance**	**-**
P	**Polarisation**	**C.m**
α	**Molecular polarisability**	**$C.m^2.V^{-1}$**
ν	**Frequency**	**s^{-1}**
ω	**Angular frequency**	**s^{-1}**
n	**Refractive index**	**-**
h	**Planck's constant**	**J.s**
$\tilde{\nu}$	**Wave number**	**m^{-1}**

*The use of the symbol E for both energy and electric field shall be distinguishable through context.

Table of contents

Chapter 1. Introduction

Index

1.1. Background and problem statement

The research conducted and presented for this dissertation is of an applied nature attempting to address a specific industrial challenge. It forms part of an ongoing collaborative research program between North-West University and the South African Nuclear Energy Corporation Pty (Ltd) (Necsa, 2010).

Necsa has significant expertise in fluorine chemistry due to its past and current interest in the nuclear fuel cycle. In this regard fluorine is very important as it reacts with uranium to form the compound UF_6, which is a gas above 56^0C as required for, amongst other things, isotopic enrichment purposes. The expertise in fluorine chemistry resulted in the development of spin-off commercial products, which are developed, manufactured and marketed by Pelchem, a commercial subsidiary of the Necsa Group - and includes a range of specialty inorganic and organic fluoride gases and liquids that are produced on-site with large-scale in-house designed fluorine cell technology and purification systems. The products are supplied to various local and international (70%) markets such as the refrigeration, solvent, detergent and semiconductor industries (Necsa, 2010).

New product development forms an important part of effective commercialization. Therefore research and development receives constant attention. As the manufacturing processes often produce mixtures of products, accurate analysis is essential for product development, optimization and quality control. Generally the C-F type products are in a

solution of anhydrous HF, and cannot be analyzed directly with traditional techniques such as gas chromatography (Siegemund, 2005) without derivatisation. The derivatisation however, is not only time consuming and complex, but can introduce unwanted side reactions of the analytes of interest. A technique such as vibrational spectroscopy, that can analyze these products directly without further preparation, would have a distinct advantage.

North-West University thus embarked on a collaborative study with Necsa R&D to assess the viability of accurate, quantitative product component identification in mixtures of fluorocarbon compounds. Due to the complexity of the spectra, simple calibration techniques would be inadequate. It was therefore decided, to investigate as an MSc project, the use of advanced multivariate data analysis techniques (i.e. Chemometrics), such as principle component analysis and multivariate regression to address this problem.

For the study, fluorinated alcohols (C2-C8) were chosen as representative chemical compounds. Their spectra include vibration peaks of carbon-fluorine bonds, which are representative of spectra of similar compounds that are of interest in the fluorocarbon industry. In addition they are (i) reasonably affordable, (ii) in liquid form that makes sample preparation accurate and convenient and (iii) they are safe to handle, which is not the case with many other fluorine compounds.

Only limited information was found in the literature on the analysis of fluorocarbon compounds such as the ones used in this study. The vibrational spectra of one of the components (2,2,3,3,3-Pentafluoro-1-propanol), used in this study, was analyzed by Badawi & Forner (2008), where it was shown that IR and Raman spectra of such fluorocarbon compounds are complex and suitable for application in multivariate analysis, thus providing evidence that this project might be feasible.

The Fourier Transform Raman spectrometers at NWU and Necsa are suitable to perform both Raman and IR analysis, whereas an Attenuated Total Reflection Infrared (ATR-IR) instrument is available at Necsa. Based on the practical availability of these instruments and the information obtained from the literature study, it was decided to use Raman

spectroscopy as the primary technique but to extent the study to include the use of infrared spectroscopy as well.

Screening experiments revealed that the near infrared part of the spectrum of the selected fluorinated alcohols was rich in detail, providing a potential set of suitable variables for accurate multivariate regression. Consequently, IR spectroscopy in the NIR region and ATR-IR spectroscopy was selected as the alternative or secondary techniques of investigation.

1.2. Aim and objectives

The aim of this study was to explore multivariate data analysis and spectroscopic techniques (Raman, NIR and ATR-IR) as a combined technique to quantitatively analyse fluorocarbon mixtures.

One of the objectives of the study was to assess the suitability of the selected spectroscopic techniques for the analysis of fluorocarbon alcohols. For a secondary objective, spectra that are complex, rich in information and representative of spectra of fluorocarbon compounds in general had to be obtained.

An experimental design had to be constructed that could be used to obtain a calibration range of samples containing mixtures of fluorocarbon alcohols, the spectral data of which could be used in multivariate data analysis. These designs were used to construct regression and calibration models from spectral data and to compare different multivariate data analysis techniques (for example PCR/PLS). The last objective was to test the predictive abilities of multivariate calibrations models, by predicting the known values of selected test samples of fluorocarbon alcohols.

1.3. Outline of thesis

The theoretical background of spectroscopy and multivariate data analysis is described in Chapter 2 to introduce important basic concepts that are needed to understand the method used, and the results obtained in this study.

The theoretical background is followed by the method chapter, Chapter 3, in which the methods of the experimental design, spectroscopic techniques and multivariate analysis, calibration and predictions are discussed. Some details were omitted from the method chapter, but included and discussed in the results chapter (Chapter 4) to obtain a more integrated perspective. The results of the experiments are presented and explained at hand of selected examples that are representative of the bulk of results.

The conclusions derived from this study are summarized in Chapter 5, while the appendixes, containing the comprehensive results generated in this study, are listed at the back of the thesis.

1.4. Bibliography

1. BADAWI, H.M. & FORNER, W. 2008. Solvent dependence of conformational stability and analysis of vibration spectra of 2,2,3,3,3-pentafluoro-1-propanol, *Spectrochimica Acta Part A.* 71: 388–397.
2. NECSA, http://www.necsa.co.za, Date of access: 2010.
3. SIEGEMUND, G., SCHWERTFEGER, W., FEIRING, A., SMART, B., BEHR, F., VOGEL, H. & MCKUSICK, B. 2005. Fluorine compounds, Organic. *Ullman's Enctclopedia of Industrial Chemistry*, Weinheim: Wiley-VCH.

Chapter 2. Theoretical Background

Index

This chapter provides a theoretical background and is divided into two parts. The first part is dedicated to spectroscopy as a subject and the second part to multivariate data analysis and experimental design (Chemometrics). Key concepts and theoretical considerations are discussed in these two sections with the purpose to include theory that is needed to understand the purpose, scope and result of this study.

2.1. Spectroscopy

2.1.1. Introduction

In this Section, background will be supplied on aspects of the spectroscopic methods that have been used in conjunction with Chemometrics to assess the applicability of the combined techniques to quantitatively distinguish between closely related species with similar complex spectra. Only details concerning the current work are discussed since the subject of spectroscopy is extensive and many detailed text books are available (Schutte, 1968; Skoog, 1971).

To distinguish quantitatively between compounds with similar spectral responses and complicated spectra is inherently difficult as it is highly unlikely that single spectral features that depend exclusively on the individual compound will be resolvable. For this reason, multivariate data analysis techniques must be employed. On the other hand, the chosen spectroscopic technique must be able to produce the complex spectra with sufficient resolution to provide good data for analysis.

In spectroscopic techniques, a molecule is subjected to radiation. The bond vibrations in such a molecule can interact in a number of different ways with the radiation depending on the specific spectroscopic technique (Colthup, 1975). The molecule can absorb, scatter or emit radiation. For different spectroscopic techniques the mechanism of the interaction varies.

Infrared spectra depend on coupling of radiation with oscillating electric dipoles in molecules in order to affect transitions in the molecular vibration and rotation spectra.

Inherently non-polar molecules, such as diatomic molecules, as well as strong polar solvents, such as water, are not suited for IR analysis since non-polar molecules are infrared inactive while strong polar solvents strongly shield infrared radiation from interacting with dissolved compounds even though they may be infrared active.

Raman spectroscopy has a coupling mechanism that does not require molecules to have oscillating dipoles but only to have a polarisability that changes as the molecule vibrates. Raman spectroscopy further utilises a light source that only has to polarize the molecule and does not necessarily have to resonate with any existing quantum levels in the sample. For this reason the probing light source can be chosen to have excellent transmission properties, also in polar solvents such as water.

It is therefore clear that different spectroscopic techniques have their own advantages and disadvantages that influence the choice of method depending on the application. For the purpose of this study, three spectroscopic techniques were explored: Raman, ATR-IR and NIR.

2.1.2. Fourier transform spectroscopy

Since many modern spectroscopic instruments (and all the instruments used in this study) utilise Fourier transform techniques, the basic principle thereof is discussed.

In Fourier transform spectrometers, a Michelson interferometer (Figure 2.1) divides a beam of light into two mutually coherent beams (Grant, 1968). One beam is directed towards a stationary mirror and the other one to a moveable mirror. The two beams are re-united and the intensity of the united beam is described by:

$$I = |\boldsymbol{E}|^2 = |\boldsymbol{E}_1|^2 + |\boldsymbol{E}_2|^2 + 2\boldsymbol{E}_1.\boldsymbol{E}_2 \cos kx = I_1 + I_2 + 2\boldsymbol{E}_1.\boldsymbol{E}_2 \cos kx \qquad (2.1)$$

Where: I is the intensity,

$\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ the electric fields of the two waves respectively,

and *k* the wave number of the radiation, $k = 2\pi/\lambda$.

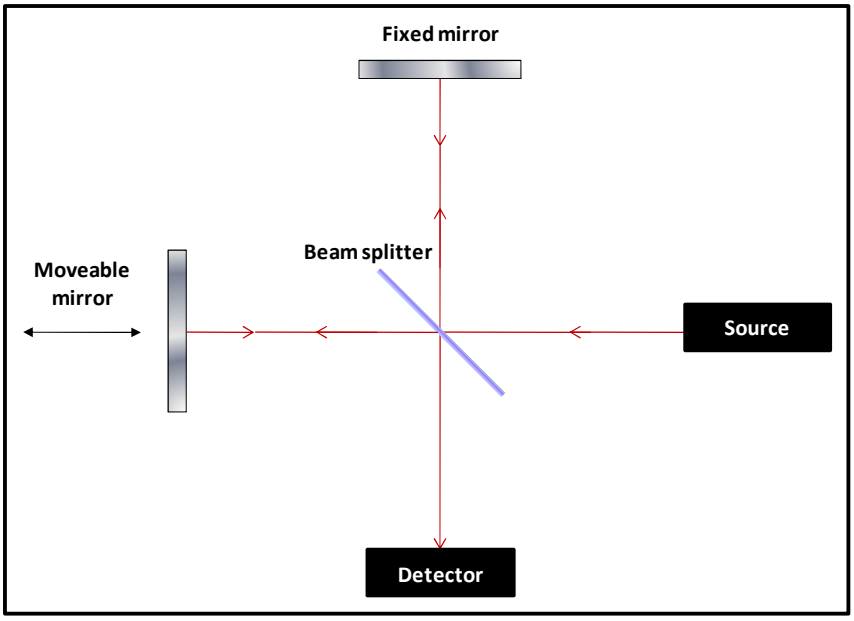


Figure 2.1: Michelson interferometer arrangement for Fourier transform spectroscopy.

When a sample is introduced before the united beam reaches the detector, the interaction of the sample with the light gives a spectral distribution, $S(k)$. The final intensity of the two re-united beams will depend on the spectrum. By recording the intensity as a function of the path difference between the two beams, the spectrum can be deduced. This method of obtaining a spectrum, as opposed to making use of a diffraction element such as a prism or grating, is known as Fourier transform spectroscopy.

Since the re-united intensity will have the same rule of composition but will be a summation over the whole spectrum (Grant, 1968), using equation 2.1 we can mathematically describe the final reunited intensity as:

$$I(x) = \int_0^\infty (1 + \cos kx) S(k)\, dk$$

$$= \int_0^\infty S(k) dk + \int_0^\infty S(k) \frac{e^{ikx} + e^{-ikx}}{2} dk$$

$$= \frac{1}{2} I(0) + \frac{1}{2} \int_{-\infty}^{\infty} S(k) e^{ikx}\, dk \qquad (2.2)$$

Where the intensity of the light that has zero path difference is denoted by $I(0)$. This part will provide background light with no spectral information, and can in principle be subtracted from the spectral part. Rearrangement of equation 2.2 reveals:

$$W(x) = 2I(x) - I(0) = \int_{-\infty}^{\infty} e^{ikx} S(k)\, dk \tag{2.3}$$

From equation 2.3 and Fourier transform theory it is clear that the functions $W(x)$ and $S(k)$ constitute a Fourier transform pair, from there the name Fourier transform spectroscopy. Accordingly we have:

$$S(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} W(x) e^{-ikx}\, dk \tag{2.4}$$

This shows how the spectrum of wave numbers can be deduced from $W(x)$ through use of a Fourier transform routine. Note that $W(x)$ is simply the recording of the intensity by a detector, at the point where the two beams re-join, as a function of the distance, x, of the movable mirror. No grating is required. The actual calculation of the Fourier transform is done by a high speed computer for which very effective fast Fourier transform routines exist. In Figure 2.2 it is shown how a sinusoidal signal is transformed by a Fourier transform routine from the time to the frequency domain (Atkins & Paula, 2002).

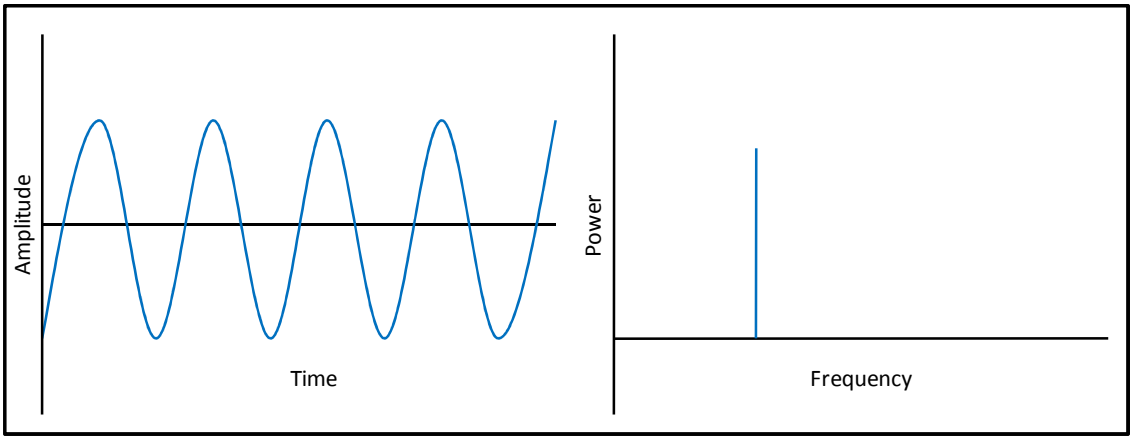


Figure 2.2: Transformation from time to frequency domain.

In many Fourier transform instruments, such as the Bruker Optics FT-Raman/IR, Vertex 70 series, used in this study (see Chapter 3), a He-Ne laser is used to accurately calibrate the distance x. The precisely known wavelength of the He-Ne laser is exploited to deduce distance from the corresponding interference intensity.

The Fourier transform technique is particularly useful for analysing the infrared absorption of gases where the spectrum is complicated, in which case it is known as Fourier Transform Infrared spectroscopy (FTIR). It has the further advantage over monochromator techniques in that all available light is utilised with high efficiency. This makes Fourier transform spectroscopy invaluable for the spectral analysis of weak sources. In Raman spectroscopy, the Stokes and anti-Stokes lines comprise only a very small part of the total scattered intensity. The wide applicability of Raman spectroscopy, coupled to the ability of the Fourier transform technique to exploit low intensity signals, forms a powerful combination.

2.1.3. The Beer-Lambert law

A parallel beam of light traversing an absorbing medium decays exponentially with distance into the medium (Willard, 1981):

$$I(x) = I_0 e^{-kx} \tag{2.5}$$

Where: I_0 is the intensity of the radiation before absorption,

x is the distance travelled into the medium,

$I(x)$ is the intensity at a specific position, x

and k is a characteristic constant that depends on the frequency and absorbing species.

The constant, k is best established through experimentation. For a fixed path length absorber cell and for a given solvent containing small quantities of the absorber molecules, equation 2.5, after rearrangement, becomes the Beer-Lambert law:

$$\mathrm{A} = \log\frac{I_0}{I} = 2.303\ a\ l\ C \qquad (2.6)$$

Where: The factor 2.303 is the base-ten conversion,

A is the absorbance,

l the length of the absorber cell,

C the concentration of the absorbing species

and a the specific absorptivity or specific absorption coefficient.

When the intensity is reduced to 10% of the original I_0 value, the absorbance value is equal to one. This is equivalent to only 10% transmission through the cell (Figure 2.3).

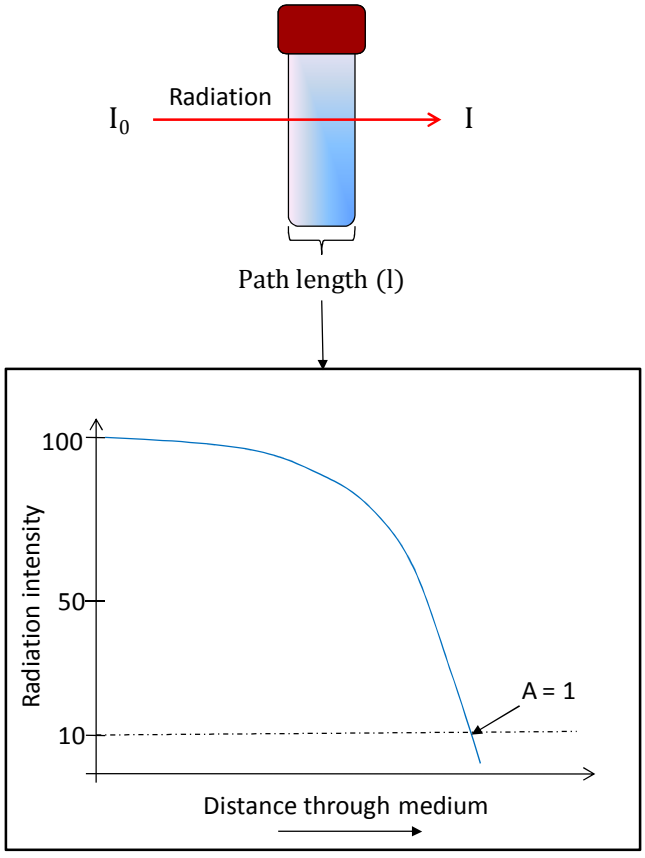


Figure 2.3: Extinction of radiation through the medium.

Multivariate regression is inherently a linear method. It is thus important to make sure that spectral data are expressed in the proper form to ensure a linear response with

concentration. Saturation of the spectrum occurs when the optical path length through the medium becomes so large that an insufficient amount of light passes on to the detector. Consideration should therefore be given to the optimization of the path length for the concentration range that will be used in experimental design to prevent spectral distortion due to extreme absorbance values.

It may be required to design a specific absorber cell for a given application to prevent saturation of the signal. For this purpose, A = 1 is a reasonable first choice, to be verified experimentally. After the specific absorptivity has been determined empirically and the concentration range of interest is established, equation 2.6 can be used to calculate a maximum cell length. The optimisation of the absorber cell length fell beyond the scope of this study.

2.1.4. Raman spectroscopy

The Raman effect is caused by the electronic polarisation of a molecule being radiated by Ultra-violet or visible light. The German, Adolf Smekal (Colthup, 1975) previously predicted the Raman effect but it was the Indian physicist, Sir C.V. Raman who noticed in 1928 that a small fraction of radiation is scattered and researched and described this phenomena, now known as the Raman effect.

In Raman scattering, the incoming radiation couples to molecular quantum states of the molecule via the polarizability of the molecular charge distribution. As the molecule vibrates, the bond length varies periodically. As a result the ability of the molecule to be polarized by an external electric field varies with the molecular frequency. The radiation energy polarizes the molecule and acts as a second oscillator at different frequencies, as shown in Figure 2.4. There are thus two oscillation frequencies coupled to each other via the polarizability of the molecule. The net result of two frequencies superimposing on the same system is well known (Grant, 1968), and result in the production of new frequencies equal to the sum and difference of the original frequencies.

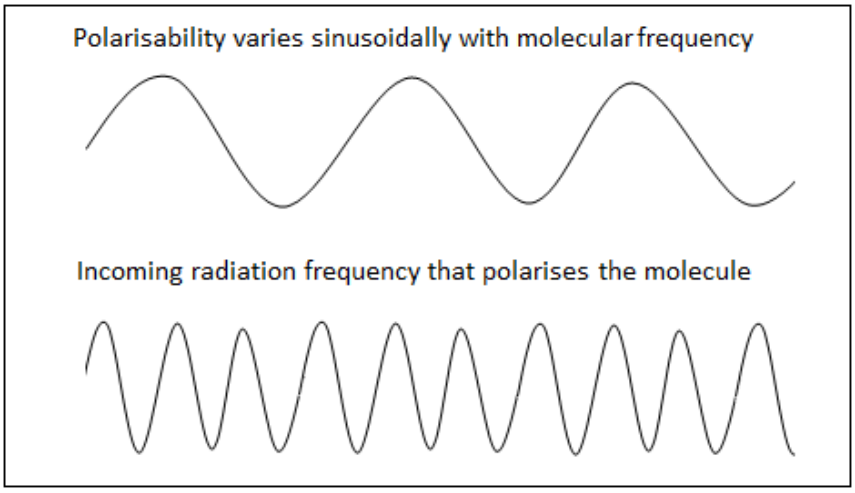


Figure 2.4: Two frequencies superimposed on the molecule (not to scale).

The coupling of the two oscillations then causes the polarisation of the molecule, P, to vary as follows (Colthup, 1975; Skoog, 1971):

$$P = \alpha E_0 cos(\omega_0 t) + \varepsilon q_0 E_0 \{cos(\omega_0 - \omega)t + cos(\omega_0 + \omega)t\} \quad (2.7)$$

Where: α is the molecular polarisability,

ω_0 is the angular frequency of the incoming laser light,

ω is the angular frequency of the sinusoidal polarisability change,

q is the amplitude of molecular vibration,

and ε represents the measure of change in α as q changes (strength of the coupling).

The first term describes the normal polarization of the molecule by the electric field leading to elastic scattering. The wave passes through the charge distribution and leaves the cloud in the same energy state as before the perturbation. This is known as Rayleigh scattering and the majority of photon scattering events are of this nature.

The second term is dependent on the change in polarisability (if $\varepsilon = 0$ then the second term falls away). This term is the part of the radiation that is scattered non-elastically. So far, the classical analogue of Raman scattering was discussed. Only about one in every million

scattering events is non-elastic and the energies of these scattered Raman photons can be described as:

$$h\nu = h\nu_0 - \Delta E_v \quad (2.8)$$

$$h\nu = h\nu_0 + \Delta E_v \quad (2.9)$$

Through the introduction of equations 2.8 and 2.9 the quantum concept has been introduced. In the first case, energy of the photon is decreased while the molecule is exited. The scattered photons are diminished in energy and appear as a spectral line at higher wave number, known as the Stokes line (Figure 2.5). The second case gives rise to the anti-Stokes line in the opposite direction. The Stokes line has the higher intensity and is most often used in spectroscopy.

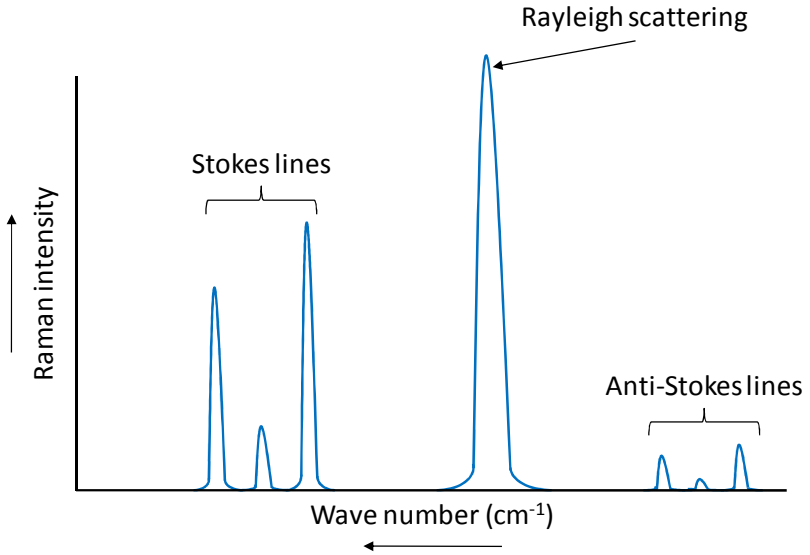


Figure 2.5: Elastic and non-elastic scattering at different wave numbers.

Figure 2.6 presents an energy level diagram that summarises the nature of the transitions due to the different scattering processes described so far. In Raleigh scattering excitation of the charge distribution occurs. The energy of the photon is converted to a charge oscillation that in turn emits a photon. The whole absorption-emission process takes only about 10^{-12} s and starts and ends in the same state. The molecule neither gains nor loses energy (elastic scattering). Raman scattering, on the other hand, is non-elastic and leaves the molecule in an excited state (Stokes) after the process, or starts from an excited state (anti-Stokes).

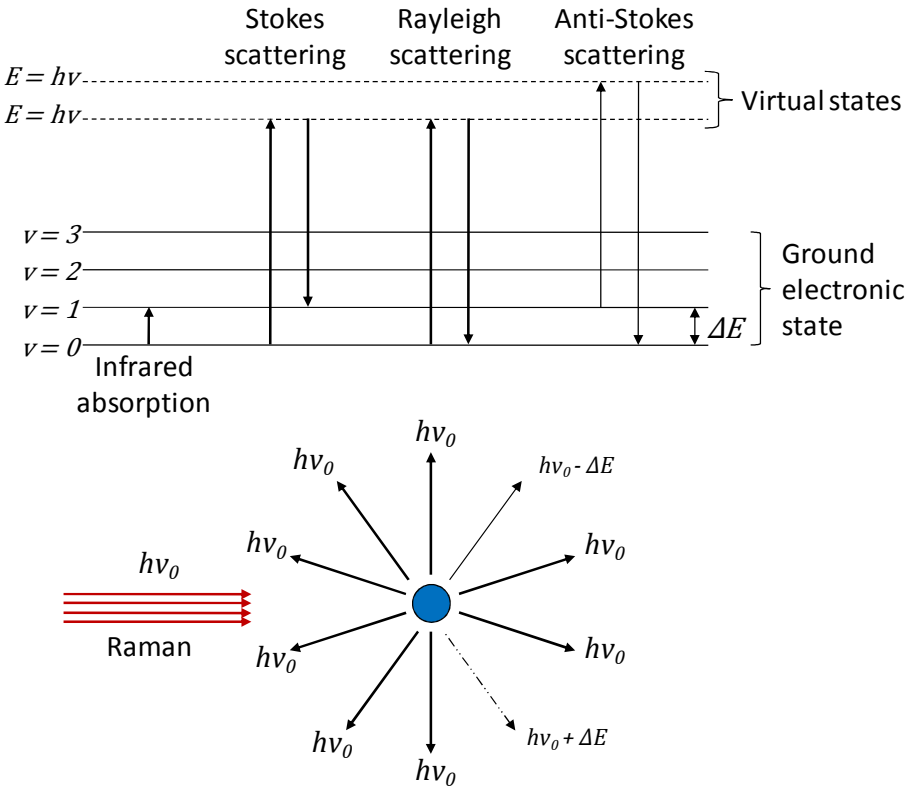


Figure 2.6: Elastic (Raleigh) and Raman scattering.

2.1.5. Infrared spectroscopy

Infrared (IR) spectroscopy is a well known and widely applied analytical technique and utilises the resonant quantum transitions due to vibrations and rotations in molecules in the infrared part of the electromagnetic spectrum (14000 – 20 cm^{-1}). The infrared light, at characteristic frequencies, is absorbed by molecules of which the principle is as follows:

For a molecule to be IR active it has to have a dipole moment. Energy transfer from the radiation field to the molecule happens via coupling of the oscillating electric field with the dipole moment of the molecule. When the energy of the radiation photon is equal to a characteristic quantum jump between discrete vibration energy levels of the molecule the photon is absorbed while the molecule makes a transition to an exited vibration state as shown in Figure 2.7 (Willard, 1981). Upon return to a lower (or ground) state the molecule

emits IR radiation of a characteristic frequency. Rotational transitions also cause IR absorption but this is only observed in gas samples and not in liquid or solid samples.

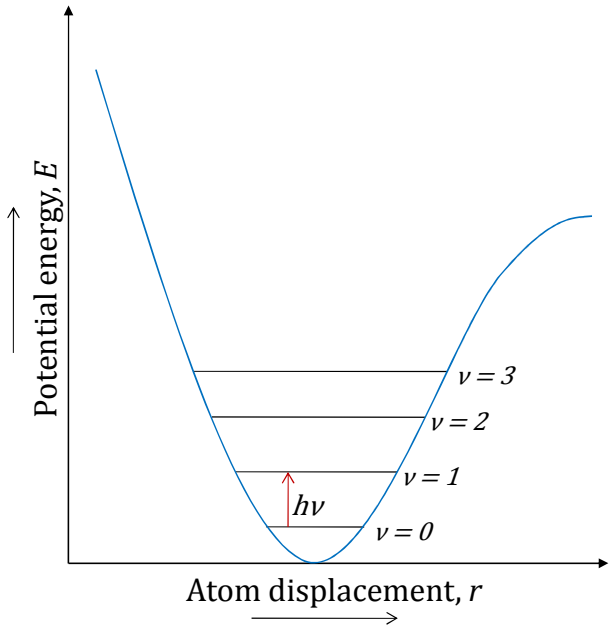


Figure 2.7: Vibration states of the anharmonic oscillator

Homo-nuclear molecules such as Cl_2 do not absorb IR radiation (they are not infrared active) because they do not have an electric dipole moment. Such molecules can still be studied by techniques such as Raman spectroscopy. This is the reason why Raman and IR spectroscopy can be viewed as complementary in some respects. The main reason however for IR to be such a well established technique compared to Raman, is that the instrumental development for IR preceded that of Raman (Skoog, 1971).

From an instrumental perspective it is convenient to divide the infrared into three sections (Willard, 1981) namely near-infrared (13000 – 4000 cm^{-1}), mid-infrared (4000 – 650 cm^{-1}) and far-infrared (650 – 10 cm^{-1}). These sections are not rigid and the definition may vary from one literature source to another. For this study, recordings of infrared spectra were made in the spectral region 6996 – 3946 cm^{-1}, which can best be classified as extending into the NIR region, and in the 3726 – 417 cm^{-1} region, which is classified as the mid infrared region.

A typical infrared spectrometer consists of a source that emits a broad band (quasi-continuous) infrared spectrum, a dispersion element such as a grating or prism, or a Fourier transform mechanism and a detector. As IR sources, Blackbody radiation is commonly used, such as from a tungsten filament lamp in the NIR region or a coil of Nichrome wire for the MIR section (Willard, 1981). A typical instrumental setup is shown schematically in Figure 2.8. When FTIR spectroscopy is performed, a Michelson interferometer, such as shown in Figure 2.1, is employed instead of a scanning grating. The theory and principals are discussed in Section 2.1.2.

An IR light source is collimated onto a dispersion element where a particular wave number is selected and passed on to a beam splitter. In most infrared spectrometer instruments two equivalent beams of radiant energy are taken from the source. By means of the beam splitter one beam passes along a reference path that contains all the optical elements of the other path, except for the sample and is not shown in the simplified diagram below. The other beam (shown) passes through the sample where the selected wave number interacts with the molecules and may be attenuated via absorption or passed through un-attenuated depending on the quantum conditions of the radiation-molecular system. The instruments used in this study however, all utilise a single-beam Fourier transform system. The transmitted energy interacts with the detector to produce a signal proportional to the intensity of the transmitted beam. By scanning over the wave number region an absorption spectrum is produced.

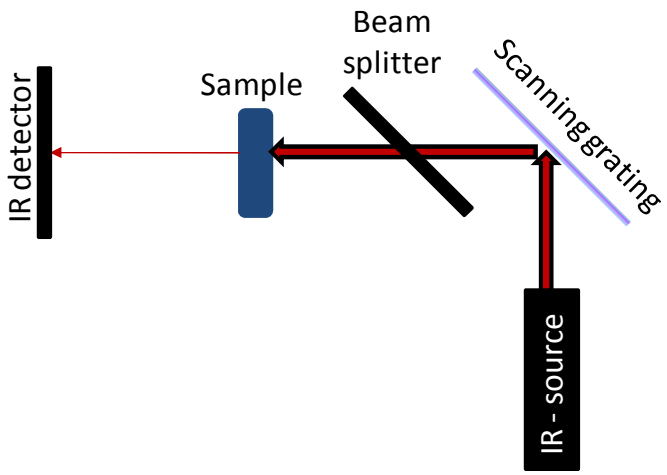


Figure 2.8: Simplified schematic of instrumental setup used in infrared spectroscopy.

Because there is such a large number of degrees of freedom for vibration and rotation bands, and also interference between them, infrared spectra are usually quite complex. This makes infrared techniques difficult for quantitative assessment of components when simple data analysis techniques and calibrations are attempted. Multivariate data analysis, however, not only makes it possible to keep track of many spectral features and their mutual interactions simultaneously, but in fact exploits the complexity of the spectrum to find rich calibration clues due to the availability of a much larger data set (Shao, 2010). This is especially advantageous in mid-infrared methods (such as ATR-IR) because the fingerprint region (1200-600 cm^{-1}) is rich in spectral lines that are unique to each component.

In the near-infrared region (13000-4000 cm^{-1}), the absorption bands are overtones or combinations of stretching vibrations (Willard, 1981) - usually C-H and O-H vibrations. Near infrared (NIR) is generally used for quantitative determinations of species such as water or some hydrocarbons such as alcohols (Skoog, 1971).

Narrow peaks, stray radiation, regular non-adherence to the Beer's Law and the small path lengths required are further disadvantages of IR as a quantitative analytical technique. The small path length of sample cells in IR makes it hard to duplicate and significant errors could occur. This can be remedied in part by adding repeat measurements to the multivariate data set or by increasing the sample set. The statistical error will be lower but the time and cost of the analysis will increase.

Normal infrared spectra can be performed readily on liquid samples, provided the solvent does not shield the radiation from reaching the analyte. For analysis of liquids that are not in solution (neat liquids), very thin sample cells or reflectance techniques (ATR-IR) are used.

2.1.6. Attenuated Total Reflection

The scope of infrared spectroscopy as a qualitative analytical tool has been boosted substantially by the technique of multiple internal reflections (Willard, 1981) which is known as attenuated total reflection (ATR). Light enters a crystal at an angle sufficient for total internal reflectance as depicted in Figure 2.9. Where the light reflects at the surface there is

an evanescent wave that ventures beyond the boundary. The evanescent wave penetrates the medium (with a depth of the same order as the wave length of the radiation) beyond the crystal's surface. As long as there are no species in this region that absorb light, the evanescent wave remains virtual (undetectable) and total internal reflection without loss of energy occurs. When an absorbing species is present within the penetration region of the evanescent wave, normal absorption takes place.

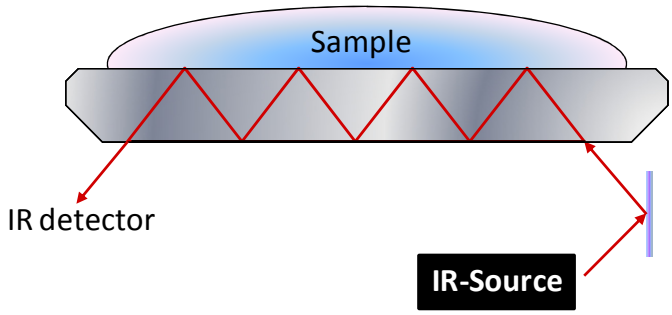


Figure 2.9: Total internal reflections in a crystal as utilised in ATR.

Sensitivity is improved through multiple passes while the short path length into the material prevents saturation effects. Thus almost anything that can be pressed up against the surface of the crystal can be analysed without further dilution or sample preparation. This includes liquids, powders and solid surfaces.

ATR-IR spectra are complex with multiple narrow peaks in the infrared fingerprint region and are therefore usually used for qualitative analysis rather than quantitative analysis. With multivariate analysis, the complex spectra do not constitute a fundamental problem for quantitative analysis.

Shifts in band intensities and frequencies often occur in ATR-IR spectroscopy. For qualitative analysis it is essential that corrections for these shifts are done. For quantitative multivariate analysis these shifts will not necessarily have a significant effect. The corrections for ATR-IR spectra are done by considering the effects of the penetration depth of the evanescent wave into the sample and the refractive index of the sample. The penetration depth is given as (Pike Technologies, 2011; Nishikida, 2010):

$$d_p = \frac{\lambda}{2\pi\sqrt{n_1^2 \sin^2 \theta - n_2^2}} \tag{2.10}$$

Where: λ is the wave length (nm) of the incident radiation,

n_1 is the refraction index of the crystal,

and n_2 is the refraction index of the sample/component

θ the crystal angle of incidence (45°),

The ATR absorbance can be expressed as (Pike Technologies, 2011):

$$A = (\ln e)\frac{n_2}{n_1}\frac{I_0}{\cos\theta}\frac{d_p}{2}a$$

$$= (\ln e)\frac{n_2}{n_1}\frac{I_0}{\cos\theta}\frac{\lambda}{4\pi\sqrt{n_1^2 \sin^2 \theta - n_2^2}}a \tag{2.11}$$

Where: $n_1$is the refractive index of the crystal,

n_2 the refractive index of the sample,

a the specific absorptivity,

I_0 the incident light intensity.

Preparation of samples using different concentrations of components, will lead to different values of n_2. The question then arises whether a preparation of samples in this way will introduce errors in the absorbance values that may negatively influence linear data analysis, and what the magnitude of such effects will be. In what follows, it will be shown how the effect on A of varying refractive indexes over the different compounds can be calculated and an estimate applicable for the experimental study will be made. For this purpose only the partial derivative with respect to n_2 needs be considered while the other variables shall be swept up into constants to yield the much simplified expression:

$$A = \frac{kn_2}{\sqrt{g - n_2^2}} \tag{2.12}$$

Where $g = n_1^2 \sin^2 \theta$ and *k* are constants of the system. Partial differentiation with respect to the sample refractive index yields:

$$\frac{\partial A}{\partial n_2} = \frac{k(g + n_2 - n_2^2)}{(g - n_2^2)^{3/2}} \tag{2.13}$$

The differential change in the absorbance due to change in refractive index is:

$$\Delta A = \frac{\partial A}{\partial n_2} \Delta n_2 \tag{2.14}$$

Then combining equations 2.12 and 2.13 in equation 2.14 yields:

$$\frac{\Delta A}{A} = \frac{(g + n_2 - n_2^2)}{(g - n_2^2)} \frac{\Delta n_2}{n_2} \tag{2.15}$$

For a diamond crystal ($n_1 = 2.417$) and an angle of incidence of 45^0, the value of $g = 2.9209$. Subsequently:

$$\frac{\Delta A}{A} = \frac{(2.9209 + n_2 - n_2^2)}{(2.9209 - n_2^2)} \frac{\Delta n_2}{n_2} \tag{2.16}$$

Using equation 2.16, we can now determine the sensitivity of absorbance due to a variation in the refractive index. The average refractive index of compounds used in this study is approximately 1.3. An approximate formula for estimating the effect of refractive index on absorbance in this case is:

$$\frac{\Delta A}{A} \cong 2 \frac{\Delta n_2}{n_2} \tag{2.17}$$

The maximum possible variation in refractive index is 2.4% (as derived from maximum difference in refractive index of the pure compounds) and thus we have a maximum of 4.8%

variation in absorbance. In practise this extreme scenario of using a single pure compound as a sample will not be encountered as many compounds are mixed into the sample. The variation in refractive index will thus be much less and its effect on absorbance will be negligible compared to random errors (shown in Chapter 4).

2.2 Chemometrics: Multivariate data analysis and experimental design for chemical applications.

2.2.1. Introduction

To minimize experimental, analytical and statistical errors, the scientific method relies heavily on careful and controlled experimentation followed by systematic analysis of the data in order to find relationships between variables or to perform calibrations that can be used as a basis for quantitative evaluation and prediction. Experimental design and data analysis thus constitute the two main pillars of the scientific method.

Simple methods such as graphical analysis and visual inspection of data have been, and still are, used with success but have practical limitations. With advancement of computing power, practical implementation of advanced numerical data analysis techniques became possible and was quickly shown to be powerful aids in finding and calibrating relationships in complicated data sets (Geladi, 1990). Chemometrics is one such method.

Consider a hypothetical example that will be used in this section to aid explanation. In an experiment a mixture exists that is made up of two components – A and B. In this experiment the viscosity of this mixture changes as the concentrations of A an B ([A] and [B]) or the temperature, T, change. The purpose of the experiment is to determine in what way the changes of [A], [B] and T impact the viscosity of the mixture.

In generalised terms, a variable *Y* (viscosity) depends on variables, X_i , $i = 1, \cdots, n$ ([A], [B] and T). Classical analytical approaches then suggest that the dependent variable *Y* depends on the values of the X_i while the latter are mutually independent. Mathematically this can be expressed as follows:

$$Y = F(X_1, X_2, \cdots X_n) \tag{2.18}$$

If the X variables are known to be independent, the experiment to uncover the relationship (equation 2.18) can be done by keeping all but one independent variable, X_i constant at a

time. This classical approach is often adopted without prior knowledge of the real relationship between the chosen X variables.

This is however not the usual situation that one encounters in chemistry. It is often not possible to pre-determine variables that will affect Y or to know whether those variables are independent or not. In terms of the viscosity example: if we keep the temperature and the concentration of compound B at constant levels, say 25° and 5g/L respectively, while changing the concentration of A, we assume that the change in viscosity is independently associated with the change in [A]. In other words, viscosity would not have been affected if we would have changed [B] to 1g/L. This is not necessarily true since there could be some interaction between A and B that influences how each of them affects the response.

In some literature, reference is made to X as *independent variables* but since X_i is not always independent of other X variables, the term *determining variable* shall be used. In calibration models that are constructed for predictions, the *determining variables* are also sometimes called *predictors*. Similarly, dependent variables shall be called *response variables*.

In chemistry it is quite often not possible to select independent determining variables. Changing one determining variable becomes equivalent to a change in the other and vice versa. To address this problem multivariate analysis is required. In chemical applications multivariate data analysis and experimental design form the core of what is known as Chemometrics.

In this section a background on Chemometrics will be presented to provide the reader with background information on the experimental design methodology, data processing, analyses and model construction.

2.2.2. Development of Chemometrics

Chemometrics had its start in the late 1960's to 1970's with the founding of the International Chemometric Society in 1974. In an article about the early development of

Chemometrics by P. Geladi and K. Esbensen (1990), interviews with some key persons in the field revealed important aspects of Chemometrics. Amongst these are the significant importance of experimental design and the influence of modern computers and technological developments on the development of Chemometrics. The advantage a student would gain from the introduction of experimental design in the early education in science was also seen as important. The fundamental roles of both mathematics and statistics in Chemometrics were emphasized.

2.2.3. Principles of multivariate data analysis

The basis for the mechanism used in multivariate data analysis and Chemometrics lies in statistical regression. Concepts from simple linear regression can be extended to a multivariate problem. Many books are available on this subject: For this brief introduction "Applied Statistics" by J. Neter et al. (Neter, 1982) was used as main reference. Other sources are listed in the bibliography at the end of the chapter.

Multivariate data often have an internal functional relation between the response variable and determining variable(s). It is the aim of regression models to uncover this relationship. In the case of spectroscopic data, the relation originates from the Beer-Lambert law and can be expressed as a linear relationship following appropriate logarithmic transformations.

2.2.3.1. Simple linear regression

Simple linear regression (SLR), also known as univariate linear regression (ULR), should be discussed in some detail as it provides a means to understand the principles of least squares regression while introducing the statistical concepts that also apply to multiple regression models. In SLR, a straight line relationship between determining and dependent variables is either known to exist from theory or is hypothesized to exist and subsequently confirmed.

Consider again the example where the response variable is the viscosity of a solution consisting of two substances A and B and the determining variables are the concentration of two substances [A] and [B] and the temperature, T. Figure 2.10 is a graphical representation of viscosity versus [A], while [B] and the temperature were kept constant. For each value of

[A] three hypothetical measurements of viscosity were taken. The three values differ due to natural random variation, also known as normal variation. The data thus scatters around the straight line fitted through it. The line can be drawn by hand or alternatively a mathematical means of finding a best fit in a reproducible and consistent manner can be used. Simple linear regression is an example of the latter approach.

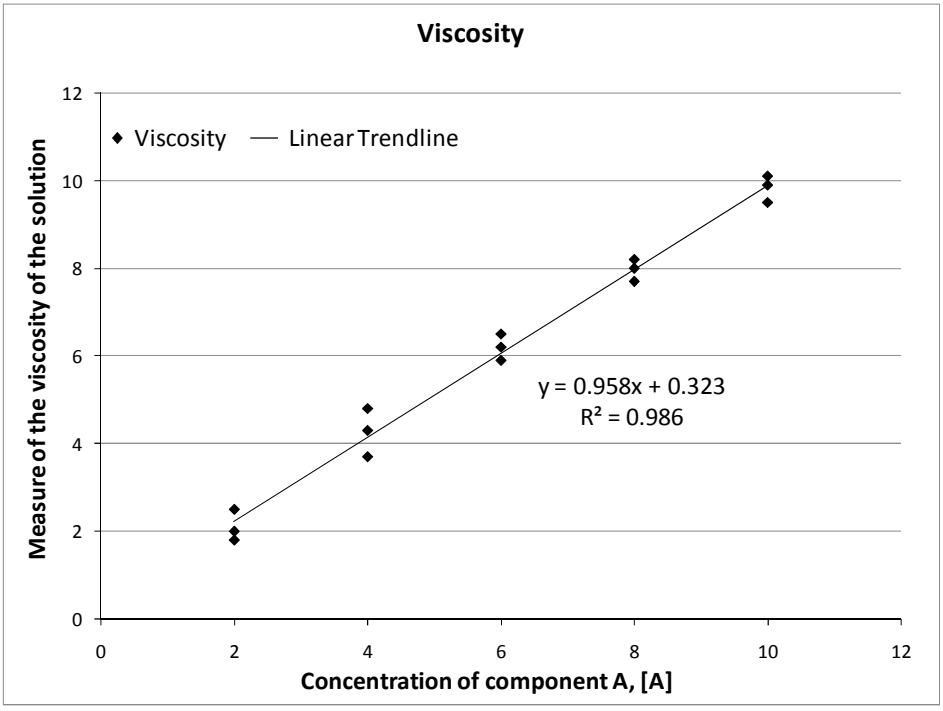


Figure 2.10: Viscosity variation as function of [A].

For the purpose of this discussion the concentration [A] shall be replaced with the usual determining variable X and the viscosity with the response variable Y. A linear relationship, that represents the line, can be written as follows:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{2.19}$$

Where: y_i is the measured response value at observation *i*,

β_0 and β_1 are linear coefficients which are unknown,

x_i is the determining variable value at observation *i*,

ε_i is the scatter component or error.

The quantity ε_i can be positive or negative and therefore, its average value over many data points will tend to zero. In linear regression, the line where $\sum\{\varepsilon_i\}^2$ is at a minimum is defined as the estimated regression function, expressed as:

$$\hat{Y} = b_0 + b_1 X \tag{2.20}$$

In terms of equation 2.20, the sum of all the squared deviations (errors) is:

$$S = \sum_{i=1}^{n} (y_i - b_1 x_i - b_0)^2 \tag{2.21}$$

The slope b_1 and offset b_0 for which *S* will be a minimum is found by setting the partial derivative of *S* equal to zero with respect to b_1 and b_0 respectively. The result is:

$$b_1 \sum_{i=1}^{n} x_i^2 + b_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \tag{2.22}$$

$$b_1 \sum_{i=1}^{n} x_i + b_0 n = \sum_{i=1}^{n} y_i \tag{2.23}$$

Solving equations 2.22 and 2.23 simultaneously yields the linear regression formulas which can be solved using a computer:

$$b_1 = \frac{\sum xy - \frac{1}{n}\sum x \sum y}{\sum x^2 - \frac{1}{n}(\sum x)^2} \tag{2.24}$$

$$b_0 = \frac{\sum y \sum y^2 - \sum x \sum xy}{n \sum x^2 - (\sum x)^2} \tag{2.25}$$

In equation 2.24 and 2.25 the subscripts have been dropped for simplicity as the context is clear from the preceding discussion.

Having calculated the linear regression, some further standard statistical formulas can be introduced.

Standard deviation and variance

The standard deviation can be thought of as the average distance from the mean of a data set to a point, which can be defined as:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{2.26}$$

Variance, which is the square of the standard deviation, is another convenient measure of the spread of the data set around the mean value, which has the added advantage of being positive definitive:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = var(X) \tag{2.27}$$

Covariance

Covariance is mathematically closely related to variance as can be seen from the forms of equations 2.28 and 2.29 below. Whereas variance provides a measure of variation with respect to the data mean, covariance provides a measure of variation of two variables *X* and *Y* with respect to each other.

$$var(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{n-1} \tag{2.28}$$

$$cov(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{2.29}$$

Variance is always positive as it is a true square. Covariance, on the other hand, is the product of two different quantities and can be positive or negative. Note that when X and Y

both deviate in the same direction with respect to their averages, the covariance is positive and vice versa. The sign of the covariance thus reveals if the two variables are in phase or out of phase with respect to their variation while the magnitude reveals the strength of the correlation. When two variables are varying in a completely uncorrelated way (no pattern linking the one to the other) the covariance is zero. When they are completely correlated and X deviates with the same sign as Y, the covariance is +1. For complete correlation but with X deviating with a different sign with respect to Y, the covariance is -1. Values in between describe a lesser degree of correlation with the sign indicating the phase.

With some algebraic transformation it can be seen that the numerator of equation 2.24 is the same as the covariance and that the denominator is the same as the variance. Another, more general and compact expression for the slope of a linear regression line then becomes:

$$m = \frac{Covariance(X, Y)}{Variance(X)} \tag{2.30}$$

This form is more general in that it can be easily extended to multivariate cases. It has the explicit form of slope because as X varies, Covariance(X, Y) provides a measure of dY and Variance(X) a measure of dX.

The Covariance Matrix

The covariance matrix for a 3-dimensional data set can be written as follows:

$$C = \begin{bmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{bmatrix} \tag{2.31}$$

Note that the diagonal elements of the matrix contain the variance of the data set while the off-diagonal elements contain the covariance. Off-diagonal terms can thus be positive or negative. Because of the definition of covariance, and the fact that multiplication is commutative, it follows that $cov(a,b) = cov(b,a)$ and consequently that the covariance

matrix of a data set is always symmetric. It is clear that this can be generalised to higher dimensions.

Linear correlation coefficient

The linear correlation coefficient, measures the strength and the direction of the linear relationship between *X* and Y. The linear correlation coefficient is sometimes referred to as the *Pearson product moment correlation coefficient* in honour of its developer Karl Pearson. (Paul et al., 1971) The mathematical formula for the correlation coefficient is:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\ \sqrt{n\sum y^2 - (\sum y)^2}} \tag{2.32}$$

The range of *r* is between -1 and 1 ($-1 \leq r \leq +1$). A perfect correlation exists when $r = \pm 1$, which implies that all the data points lie exactly on a straight line, where the sign of the slope is given by the sign of *r*. If the two variables are not correlated at all the value of *r* is zero, which means that no distinguishable mathematical relationship exists between the two variables. As a rule of thumb a correlation greater than 0.8 can be viewed as *strong*, whereas a correlation less than 0.5 is generally described as *weak*.

Coefficient of determination

Another useful quantity is the *coefficient of determination,* r^2. It measures the proportion of variance of one variable derived from the other through the linear relationship. For example, if $r = 0.922$, then $r^2 = 0.850$, which means that 85% of the total variation in *y* can be explained by the regression equation. The remaining 15% of the variation in *y* remains unexplained and represents the random scatter away from the line. From this explanation it follows that we can express r^2 as follows:

$$r^2 = \frac{Var(Y) - S}{Var(Y)} \tag{2.33}$$

Where the variability in Y, which is due to the linear relationship, is isolated from the total variability by subtracting S, the sum of the square of the error (equation 2.21), and expressing it as a fraction of the total variance.

2.2.3.2. Multiple Regression

In multiple linear regression (MLR) the methodology of simple linear regression is generalised to fit more than one determining variable (Neter, 1982; Benjamin, 1970). The extended model has the formula:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \qquad i = 1,2,3 \ldots, \mathrm{n} \tag{2.34}$$

In explicit matrix form this becomes:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1,p-1} \\ x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \beta_0 \\ \beta_0 \\ \vdots \\ \beta_0 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{2.35}$$

Provided we work with a centred data set, the constants can be eliminated and the coefficients' column vector can be calculated from a matrix equation of the form:

$$\mathrm{B} = (X^T X)^{-1} X^T Y \tag{2.36}$$

Note the analogy in the mathematical structure of equation 2.36 with 2.30. Equation 2.36 reveals some important potential pitfalls of multiple linear regression. Whenever independent variables are collinear (i.e. there exists a linear relationship between them), the matrix inverse $(X^T X)^{-1}$ is singular and B cannot be estimated. When two independent variables are nearly collinear, $(X^T X)^{-1}$ is a small quantity. This small number is very sensitive to noise in the measured X variables leading to a very unstable model. A large instability in the model may on the other hand, be indicative of near collinearity of independent variables.

MLR, in summary, cannot handle collinearity as the model becomes unstable near collinearity. For these cases, PCA provides a means to examine the structure in data sets that will allow identification and elimination of collinearity.

The coefficient of multiple determination for MLR is defined in close analogy with equation 2.33. In this case, however, the degrees of freedom of the extended problem must be accounted for. In order to define and discuss the coefficient of multiple determination, the sum of square terms commonly used in analyses of variance (ANOVA), is introduced in Table 2.1. Also listed are their equivalent quantities as introduced in Section 3.1.

Table 2.1: Sum of square terms of regression

Terms	Alternative notation	Formula	Meaning
SSTO	$(n-1)Var(Y)$	$\sum(y_i - \bar{y})^2$	Sum of squares of deviations of y_i from $\bar{Y}$
SSE	$(n-p)S$	$\sum(y_i - \hat{y})^2 = \sum(e_i)^2$	Sum of square of error
SSR	$(p-1)(Var(Y)-S)$	$\sum(\hat{y_i} - \bar{y})^2$	Measure of reduction of variability in Y by the regression line.

SSTO has $(n-1)$ degrees of freedom because there are *n* values and one constraint $\sum(y_i - \bar{y}) = 0$. SSE has $(n-p)$ degrees of freedom because there are *n* residuals and *p* constraints in the form of the estimating parameters (β_0, β_1..... β_{p-1}). The term SSR has $(p-1)$ degrees of freedom because there are *p* parameters in the regression function and one constraint $\sum(\hat{y}_i - \bar{y}) = 0$. The coefficient of determination for the multiple regression shall be indicated by R^2 to distinguish it from its SLR equivalent, r^2. Taking the degrees of freedom and equation 2.33, R^2 can be defined as follows:

$$R^2 = \frac{(p-1)(Var(Y)-S)}{(n-1)Var(Y)} = \frac{SSR}{SSTO}$$

$$= 1 - \frac{(n-p)S}{(n-1)Var(Y)} = 1 - \frac{SSE}{SSTO} \quad (2.37)$$

Notice that R^2 increases as additional determining variables (*p*) are added to the model for a fixed number of dependent variables (*n*). To counter the *p*-sensitivity on R^2 a new adjusted coefficient, R_a^2 is defined as follows:

$$R_a^2 = 1 - \frac{(n-1)SSE}{(n-p)SSTO} \quad (2.38)$$

Comparison of equations 2.37 and 2.38 shows that R_a^2 is now insensitive to the number of independent variables.

Although both SLR and MLR are strictly linear models, it can of course be used to analyse non-linear relationships provided appropriate linearization transformations are applied to the raw data first. As an example, consider the function:

$$Y = \frac{X^c}{Z} \quad (2.39)$$

Taking logarithms of equation 2.39, it becomes $\ln Y = clnX - lnZ$, which now has the required linear form $y = mx + c$, that makes the new data amenable to SLR and MLR.

2.2.3.3. Principal components

For multivariable data structures, it is not easy, or even possible in many cases, to do calculations by hand or to represent data graphically in a simple way to spot structure or correlations between variables. We need more general computational methods to perform this task and to calculate regression functions. Principle component analysis (PCA) provides such a method. It is a powerful method to identify patterns or relationships between variables in complex data sets.

As an illustration, consider again the three variable viscosity example introduced in Section 2.2.1. Let [A] and [B] vary while the temperature is kept constant, and evaluate the effect on viscosity. We may get a data swarm such as shown in Figure 2.11A. Figures 2.11B and 2.11C represent two projections of the data on the [A] and [B] axes respectively to cast more light on the actual data distribution. Even so, it is hard to picture proper mutual relations between variables or to recognise dominant structures within the data set. A large number of experiments would have to be done to deduce possible interactions between [A] and [B] or to determine an optimal mixture of A and B.

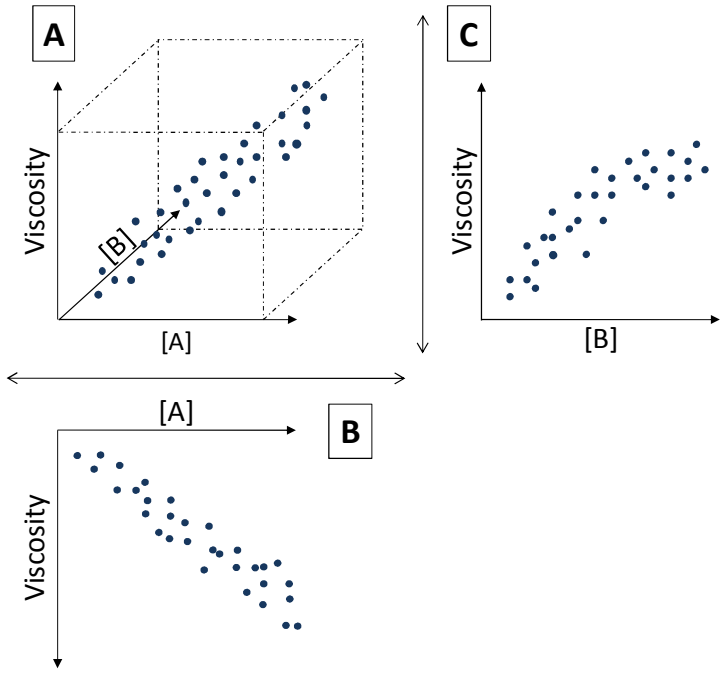


Figure 2.11: Example of 3D system

With principal components, however, it is not required to look at the data structure as shown in Figure 2.11. Instead a new dimension is created for the data that consists of coordinates of a magnitude of sample variance/structural information. PCA does this by relying on special properties of the covariance matrix (equation 2.31), which is a $n \times n$ matrix for n data dimensions. It is a symmetric square matrix that contains all the correlation information between different variables (off-diagonal terms) as well as

information on the scatter within the data set of a single variable (diagonal terms). This comprises all the information required to uncover relationships between variables.

In Figure 2.12, PCA reveals the line of the strongest correlation, PC1, for our previously presented scenario. The scatter/error component around PC1 reveals further, weaker correlations in itself. This correlation is PC2 and subsequent weaker correlation can be found as successive principal components. All principal components are always orthogonal to each other.

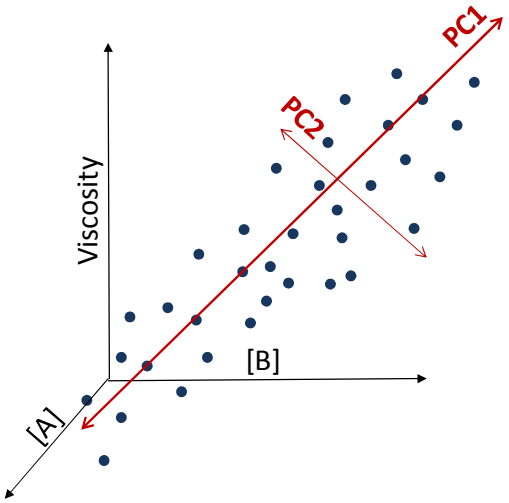


Figure 2.12: Principal components

The principal components are used to create a new data dimension called the principal component space, which often closely resembles the original data structure. However, assumptions and restrictions made on the original structure do not necessarily apply to the principal component space. In Figure 2.11, for example, it was only possible to evaluate the effect of [A] and [B] on viscosity separately. Although one intuitively knows that other factors might play a role, it is not possible to find these factors because the data space is made up only of the three known, measured variables. In a principal component space, the use of more than three vectors becomes possible. One of the vectors could possibly represent interactions of [A] and [B]. Choosing the amount of vectors/components to be used and analysing the source of the variation for each component becomes a very important task. Some information is lost when reducing the data dimensionality (by

eliminating PC's of minor importance), but what remains will contain the real structures while weaker structures, such as noise, will be removed.

2.2.4. The Chemometric method

Figure 2.13 is a flow chart representation of a typical Chemometrics process.

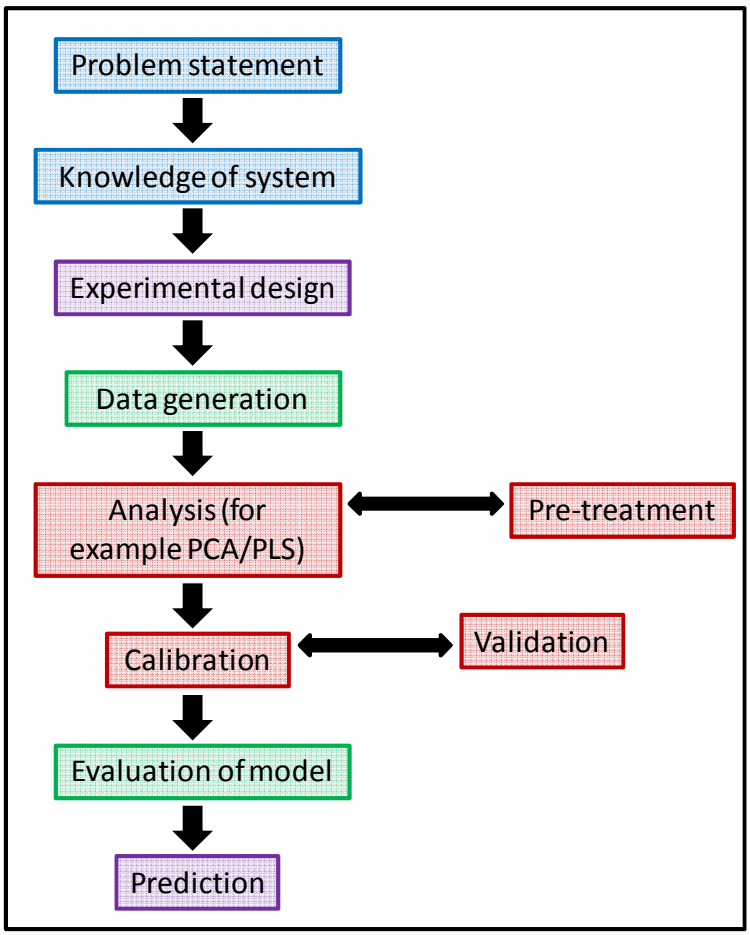


Figure 2.13: Flow diagram of chemometrics.

As with the normal scientific method, the chemometrician first needs to have a thorough understanding of the system and of what a specific problem entails.

The next step is crucial in Chemometrics – the experimental design. With proper experimental design the chemometrician minimizes the amount of work that needs to be done and maximises the amount of useful information to be gained from the experiments.

Once the data have been generated, the next step is to analyse the data *via* analysis methods such as PCA or PLS. In the process, insight is gained into the over-all data structure, outliers or groupings can be identified, decisions can be made regarding the number of components to use, appropriate variables can be selected, etc. The outcome of the initial analysis may suggest a change in the data to make it more suitable for regression, such as spectroscopic transformations, standardisation, weighing or mathematical transformations. After each change to the data set the analysis is repeated to see how well the data meet good regression criteria.

The next step is to construct the calibration model (PCR/ PLS for this study). During this process a number of different validations can be performed to test the credibility, stability and self-consistency of the model. Typical model evaluation criteria include the differences between the root mean squares of prediction and calibration and the R^2 values of the model.

These are the basic concepts used in this study and will be discussed further where required for the understanding of this thesis. Only aspects that are applicable on this study are discussed in detail. Many books are available on the subjects of multivariate analysis, Chemometrics and experimental design for further reading. (Esbensen, 2009; Sharaf, 1986)

2.2.4.1. Experimental design

The purpose of experimental design is to get the most information with the least amount of effort and to focus on important information and correlations between different variables. Sometimes a limited amount of sample is available or it is expensive and time-consuming to do experiments. Each type of experiment will require a different design but certain rules and guidelines in the theory of experimental design make it easier to decide on the design to be used. It is a very extensive topic in itself and for the purpose of this study only key concepts will be looked at (Esbensen, 2001).

In the viscosity example there are three determining variables - [A], [B] and temperature. One experimental design suitable to such a system is a factorial design as illustrated in Figure 2.14.

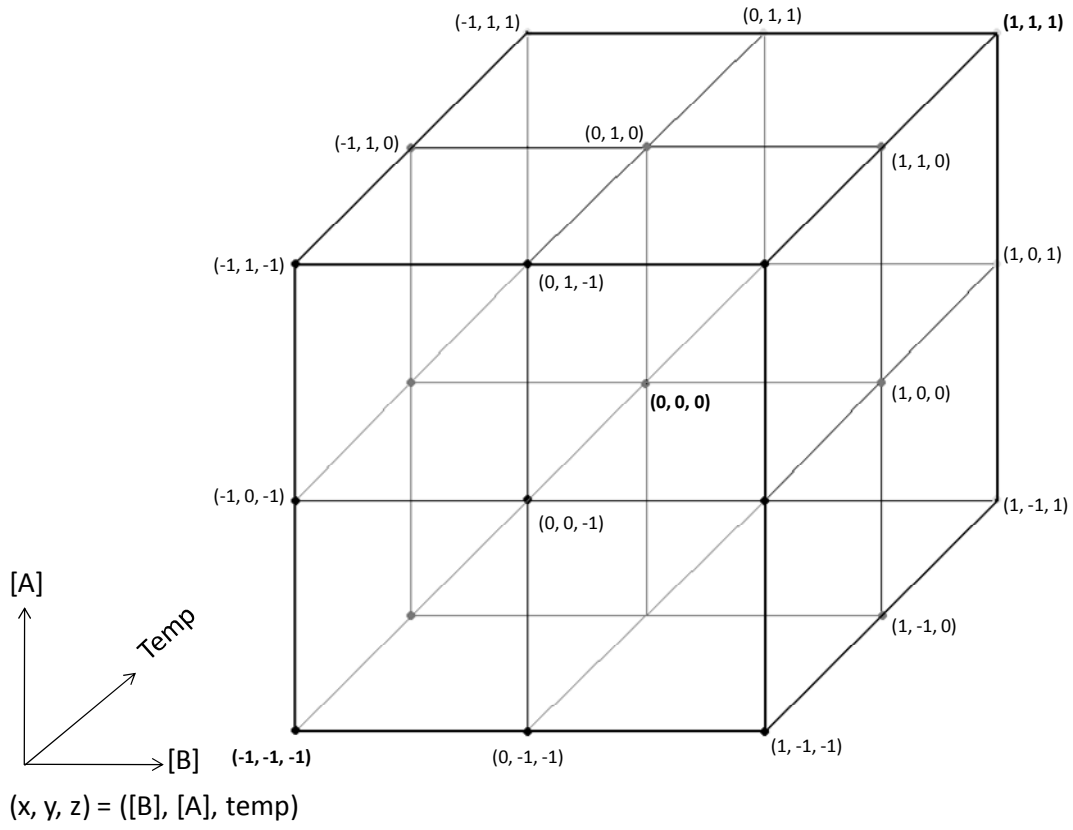


Figure 2.14: Graphical depiction of a factorial design as a lattice structure.

If all the nodes/coordinates shown in Figure 2.14, are used, the result is a full factorial design. A full factorial design with three variables, each with three levels (-1 = low; 0 = medium; 1 = high), requires 27 samples. To reduce this large number, we could choose only certain nodes. This is called a fractional factorial design and is used for screening and experimental procedures where a smaller sample set is required.

The function for a full factorial design with three variables is (Figure 2.15):

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \beta_5 X_{i1} X_{i3} + \beta_6 X_{i2} X_{i3} + \beta_7 X_{i1} X_{i2} X_{i3} \qquad (2.40)$$

Where: X_{i1} is [A] at observation *i*,

X_{i2} is [B] at observation *i*,

X_{i3} is temperature at observation *i*.

From the function and Figure 2.15 it follows that from only eight samples with three variables and just two levels per variable (low and high), a constant term, three linear terms and four interaction terms (three two-term and one three-term) can be estimated (Leardi, 2009).

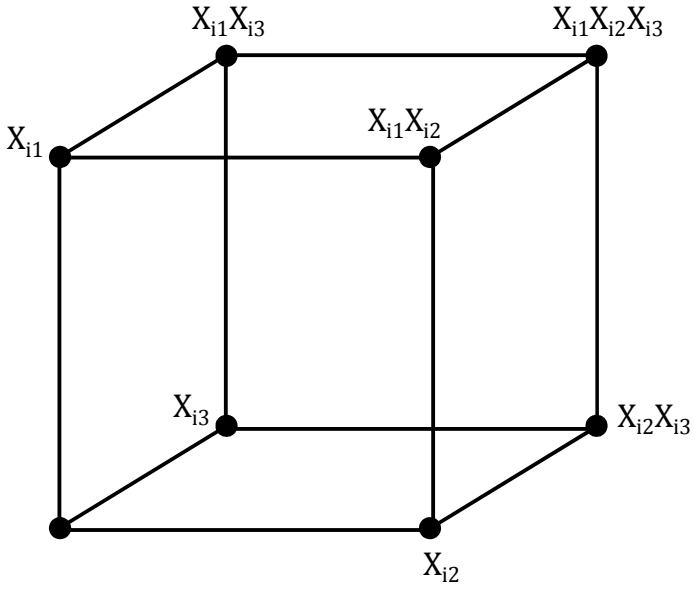


Figure 2.15: Information gained from full factorial design

The type of experiment and the design details depend on the design: designs exist for optimising a process, making inferences, analysis of effects, etc.

In this study, a mixture design was used. Mixture designs have one very important constraint that makes them different from factorial designs: the components of the mixture must always add up to unity (100%). This means that you cannot increase one component without decreasing another. It also implies that you cannot study interaction and square terms separately because these interactions are now linked. A three component mixture design is illustrated in Figure 2.16. The corners of the triangle are the three components and samples can be taken at any point within the triangle. The node in the middle of the triangle

represents a sample in which X_1, X_2 and X_3 are in equal amounts: each component contributes a third of the total mixture.

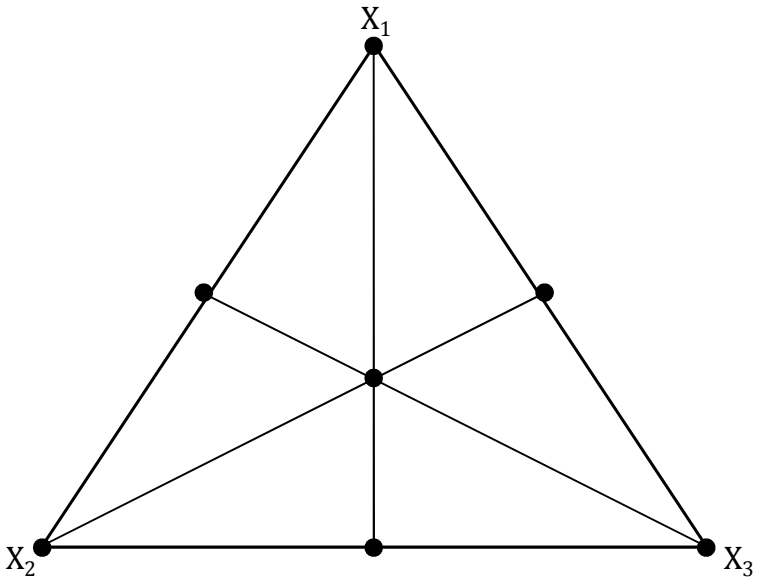


Figure 2.16: Graphical representation of three component mixture design

Software packages that have multivariable capabilities sometimes include experimental design tools. This greatly reduces the time and effort spent on planning, design and making/collecting of samples, especially in an experiment with a large number of variables.

2.2.4.2. Methods for analysis and regression (PCR and PLS)

Principal component regression (PCR) and partial least squares regression (PLS) are both factor analysis methods with full-spectrum capabilities. In spectral data, the bond vibrations that give rise to the variables, influence each other. For this reason it is advantageous to be able to analyse the entire spectral range or large wave number sections of the spectra. PCA and PLS are both similar in mechanism to classical least squares regression (CLS) (Haaland, 1988; Thomas, 1990). The classical least squares regression (CLS) mechanism, as applied to a problem typical of what is encountered in this study, is demonstrated:

A number of samples (m) are mixtures consisting of l components and their spectral data is taken. The data lay-out is shown in Table 2.2.

Table 2.2: Data structure of the experiment

Samples	Components	Spectral values
1, 2, ... m (rows)	1, 2, ... l (columns)	1, 2, ... n (columns)
1 2 3 . . . m	Y-matrix Concentration values for each component in a sample	X-matrix Spectral intensities

In order to construct a linear model, the CLS procedure is based directly on the known form of the Beer-Lambert law of spectral intensity response and the corresponding linearization transformation. The relationship reflected in Table 2.2 can be expressed as a CLS model in matrix form as follows:

$$\begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} = \begin{bmatrix} y_{11} & \cdots & y_{1l} \\ \vdots & \ddots & \vdots \\ y_{m1} & \cdots & y_{ml} \end{bmatrix} \begin{bmatrix} k_{11} & \cdots & k_{1n} \\ \vdots & \ddots & \vdots \\ k_{l1} & \cdots & k_{ln} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} & \cdots & \varepsilon_{1n} \\ \vdots & \ddots & \vdots \\ \varepsilon_{m1} & \cdots & \varepsilon_{mn} \end{bmatrix} \tag{2.41}$$

$$(n \times m) \quad = \quad (m \times l) \qquad (l \times n) \quad + \quad (n \times m)$$

This reflects the data structure explicitly. In less explicit but compact matrix notation, the equation becomes:

$$X = YK + E \tag{2.42}$$

Where: X is the $n \times m$ matrix of spectral intensities;

Y is the $m \times l$ matrix of factors/scores;

K is the $l \times n$ matrix of pure component spectra; and

E is the $n \times m$ matrix of spectral residuals/errors, not fitted by the PLS/PCR model.

For PLS/PCR the factoring remains similar as for the classical regression method but the principal component identification can now be used to construct a new data set, representing the original spectra. In PLS the original data sets, $X_i, (i = 1, \cdots, n)$ and

$Y_j, (j = 1, \cdots, l)$ are decomposed into their latent structures, which are the principle component structures inherent in the data sets. The principle component representing the most variation in X is extracted and explained by a principle component of Y that describes this variation the best. The latent variables showing most variation are known as the loading vectors and are used instead of the pure component spectra. The intensities of the loading vectors are known as the scores and are used to reconstruct calibration spectra. A linear relation of these intensities to the component concentration is modelled. The obtained residuals are the spectral residuals not fitted by the PLS/PCR model. This data transformation compresses data and reduces noise.

PCR is a PCA followed by a regression step. As explained previously, PCA extracts relationships between variables directly from the variance and covariance information. PCA do not utilise information of the concentrations of the mixture components.

PLS regression can be subdivided into PLS1 and PLS2. PLS1 regression only uses the concentration of one Y-variable to calculate the model. PLS1 can calculate the optimum amount of loading vectors for each component separately. PLS2 uses the information from all the Y-variables (component concentrations) together and calculates the same number of vector loadings to be used for all of the variables (Esbensen, 2001). Because PLS1 only utilises the optimum number of loading vectors for a given response calibration it has better predictive capabilities in general.

Each method has its advantages and disadvantages and each experiment has different factors influencing the performance of these models. It is therefore not a guarantee that one method will be the best and the choice of method should be based on knowledge of the experiment and testing/screening of models.

2.2.4.3. Data pre-treatment

Sometimes a data set might need some pre-treatment before a regression model can be constructed. Different pre-treatments and transformations are available. There is no single best approach and the choice will depend strongly on the type of data to be analysed.

Spectroscopic data are usually not standardized, normalised or weighted. This is because dividing the intensities (normally very small values) by the standard deviation amplifies noise which results in the loss of important information of variance in intensities. Where the intensities are influenced by external factors, such as varying path lengths, which negatively influence the quality of peak intensities, vector normalisation can be used (Conzen, 2006). The vector normalised data set does not contain intensity information explicitly, but still has the variability information as a function of wave number.

Multiplicative scatter correction (MSC) is a technique that can be used to correct multiplicative and additive effects and is usually applied to data of solid samples (such as powders) where light scattering is evident. Raman spectra often contain intensity contribution due to scattering and can benefit from MSC application. This transformation can also be applied to infrared spectra where background effects, varying optical path lengths, temperature and pressure changes can influence the spectra. (Esbensen, 2009; Conzen, 2006)

Where spectral data is dominated by broad peaks, and small features do not contribute much to the model, the first derivative of the spectrum can be calculated to enhance steep peaks that may contain important calibration information. This method is most often used on near infrared (NIR) data but caution should be exercised since the derivative will also have the effect of enhancing noise (Conzen, 2006).

Certain commonplace spectral transformations, such as absorbance to transmittance conversions or conversion to Kubelka-Munk units might be beneficial for analysis (Esbensen, 2009). Kubelka-Munk transformations are always used in diffuse reflectance data.

Spectral data must always be linearised before data analysis is attempted. The spectra obtained from modern instruments usually directly give absorbance values, A, which are already linearly dependent on concentration (Beer-Lambert law). When a spectroscopic technique is used that give an inherent non-linear output, appropriate linearization transformation should be applied prior to analysis.

Depending on the data set, different mathematical transformations can also be used. For example, a logarithmic transformation could be done on skewed data (log-normal distributed data). A data set should always be tested first before any pre-treatment is done to evaluate if the treatment had the wanted effect on the model. With all pre-treatment methods it should be noted that there is a possibility of loss of information, increased noise and over-treated data could give an unrealistic picture of the real situation.

2.2.4.4. Validation

When a regression/prediction model is constructed, it needs to have an estimate of how well the model will perform in future predictions. This step is known as validation. Computer programs usually do the validation and calibration step at the same time.

There are two types of validations used: Internal validation (cross-validation) for small sample sets (usually less than 50 samples) and external validation (test set validation) for large sample sets (Esbensen, 2009; Conzen, 2006).

In test set validation or test set switching, two sets of calibration samples are used. One is used as the calibration set and the other as the test set for validation. The two sets are then switched for a successive calibration and validation runs.

Cross-validation uses the same principle but is used where the sample set is too small to divide it into two sets. In this case several segments of the sample set are defined and used for validation in turn. The extreme case of this is known as full cross-validation. In full cross-validation just one sample at a time is omitted, leading to a validation process with as many steps as there are samples.

For evaluation of preliminary models, a simplified and fast technique can be used, while a comprehensive validation should be performed for the final model. It is important to note that when a validation gives a low error of prediction, it could be that the validation method

is calculating an over-optimistic value. For this reason it is good practise (although not always possible) to use more than one method to validate the final model.

2.3. Conclusions

Spectroscopic techniques exploit the quantum nature of the interaction of radiation with matter to study properties of atoms, molecules and bond vibrations. Firstly, the nature of these interactions provides different kinds of access to study molecules and compounds and secondly, spectroscopy is a fast method that requires little or no sample preparation while providing complex data sets that uniquely describes the species that are investigated.

Multivariate data analysis (MVA) is a powerful tool that can be implemented in situations where data are too complex to be analysed by conventional data analysis techniques such as the “one-variable-at-a-time” approach and simple linear regressions. MVA exploits complex data structures, as is the case in this study, where complex spectra of fluorocarbon mixtures were used to construct calibration and prediction models.

Experimental design forms an integral part of MVA and enables the analyst to construct experimental designs that are suitable for MVA while taking time and cost constraints into consideration.

The theoretical knowledge of techniques that were described in this chapter was used in constructing and executing the method (Chapter 3) used in this study and to process and interpret the subsequent results obtained (Chapter 4).

2.4. Bibliography

1. ATKINS, P. & DE PAULA, J. 2002. Atkin's physical chemistry. 7th ed. Oxford university press.
2. BEMJAMIN, J.R. & CORNELL, C.A. 1970. Probability, statistics and decisions for civil engineers. McGraw-Hill.
3. CAMO Software. The Unscrambler appendices: Method references. http://www.camo.com/downloads/user-manuals.html Date of access: 16 Sept. 2011.
4. COLTHUP, N.B., DALY, L.H. & WIBERLEY S.E. 1975. Introduction to infrared and Raman spectroscopy. 2nd ed. NY Academic Press.
5. CONZEN, J. 2006. Multivariate calibration: a practical guide for developing methods in the quantitative analytical chemistry. Bruker Optik.
6. ESBENSEN, K.H. 2001. Multivariate data analysis in practice. 5th ed. CAMO Software AS.
7. GELADI, P. & EBENSEN, K. 1990. The start and early history of Chemometrics: selected interviews Part 1. *J. Chemometrics*, 4:337-354.
8. GRANT, R.F. 1968. Introduction to Modern Optics. Published HOLT, RINEHART and WINSTON COMPANY. ISBN 0-03-065365-7. pp 84 – 86.
9. HAALAND, D.M. & THOMAS, E.V. 1988. Partial least-square methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Anal. Chem.* 60: 1193-1202.
10. LEARDI, R. 2009. Experimental design in chemistry: a tutorial. *Anal. Chim. Acta.* 652: 161-172.
11. NETER, J., WASSERMAN, W. & WHITMORE, G.A. 1982. Applied statistics. 2nd ed. Allyn and Bacon inc.
12. NISHIKIDA, K. & KEMOFERT K.D. Advanced ATR correction algorithm for infrared spectroscopy. Thermo electron corporation. Wisconsin. http://www.thermo.com/eThermo/CMA/PDFs/Various/File_21270.pdf. Date of access: Aug 2011.
13. PAUL, G.H., PORT, S.C. & STONE, C.J. 1971. Introduction to probability theory. Houghton Mifflin.

14. PIKE TECHNOLOGIES. ATR theory and applications, Application Note. http://www.piketech.com/files/pdfs/ATRAN611.pdf. Date of access: 10 Oct 2011.
15. SCHUTTE, C.J.H. 1968. The wave mechanics of atoms, molecules and ions. Edward Arnold publishers. London.
16. SHAO, X., BIAN, X., LIU, J., ZHANG, M. & CAI, W. 2010. Multivariate calibration method in near infrared spectroscopic analysis. *Anal. Methods.* 2: 1662 – 1666.
17. SHARAF, M.A., ILLMAN, D.L. & KOWALSKI, B.R. 1986. Chemometrics. Wiley.
18. SKOOG, D.A., HOLLER, F.J. & NIEMAN, T.A. 1971. Principles of instrumental analysis. 5th ed. Brooks/Cole.
19. THOMAS, E.V. & HAALAND, D.M. 1990. Comparison of multivariate calibration methods for quantitative spectral analysis. *Anal. Chem.* 62:1091-1099.
20. WILLARD, H.H., MERRIT, L.L., JOHN, A. D. & FRANK A.S. 1981. Instrumental Methods of Analysis. 6th ed. Published by D. VAN NOSTRAND COMPANY. ISBN 0-442-24503-5.

Chapter 3. Method, Planning and Design

Index

3.1. Introduction

In this chapter, the essential aspects of the experimental method, planning and design as well as the tools and procedures used for multivariate analysis, are presented and motivated. Some additional details of the multivariate method are explained in Chapter 4 in context of the experimental results. While always essential, the experimental planning and design is particularly important when multivariate regression methods are employed. As was pointed out in Section 2.2, special care should be taken to avoid accidental introduction of exact or near co-linearity between determining variables.

3.2. Materials, methods and design

3.2.1 Software tools

The multivariate analysis (MVA) tool used in this study was “The Unscrambler” software of the CAMO Software Company (CAMO Process AS, Nedre Vollgate 8, N-0158 Oslo, Norway). The basic mathematical background of some of the MVA processes executed by the software was presented in Section 2.2 in a general mathematical manner (software independent). In addition “The Unscrambler” has a number of other convenient and useful modalities that aid with experimental design (as was selectively discussed in Section 2.2) and data visualization modalities, which have not been discussed but simply employed for aided data analysis and data presentation.

3.2.2 Choice of chemicals

The primary aim of this study was to assess the ability of Chemometric techniques to quantitatively distinguish between fluorocarbon compounds in a mixture when overlapping spectral features make identification of pure peaks for use in simple calibration methods impossible. The choice of representative chemicals selected for this study was based on this requirement in combination with aspects of convenience and safety of handling and preparation. Fluorocarbon alcohols, listed in Table 3.1, were selected as representative chemicals because they all have similar spectral lines in the same wave number region, they all have the same functional single alcohol group and they are all liquids that can be handled with equal ease and safety.

3.2.3 Experimental design

The study was planned in two phases. The first phase entailed a short survey study, using a selection of only four fluorocarbon alcohols to assess the viability of the method. The second phase comprised of the much more comprehensive study involving six compounds. For convenience of reference, alphabetic notations for the compounds in each group were used as listed in Table 3.1. Note that in Table 3.1 the third component of the phase one and phase two experiments differ. To avoid confusion, the notations C' and C are used respectively and discussions are done in group context.

Table 3.1: Compound groupings for the two experimental phases

Group 1: Four compound selection for viability study design (Phase one)	
Short notation	**Compound name**
A	2,2,3,3,4,4,5,5-Octafluoro-1-pentanol
B	2,2,3,3,3-Pentafluoro-1-propanol
C'	2,2,3,4,4,4-Hexafluoro-1-butanol
D	2,2,2-Trifluoroethanol
Group 2: Six compound selection for a final experimental design (Phase two)	
Short notation	**Compound name**
A	2,2,3,3,4,4,5,5-Octafluoro-1-pentanol
B	2,2,3,3,3-Pentafluoro-1-propanol
C	1H,1H,2H,2H-Perfluoro-1-octanol
D	2,2,2-Trifluoroethanol
E	2,2,3,3-Tetrafluoro-1-propanol
F	1,1,1-Trifluoro-2-propanol

"The Unscrambler" software was used to perform mixture designs free of design-related problems, such as co-linearity and systematic errors through random construction of the mixture range. The mixture design for the screening experiment (Phase one) is shown in Table 3.2. The values in the table are in fractions of the maximum volume of 0.25 ml as explained in Section 3.3.

The six compounds of Group 2 were used in two experimental mixture designs for further investigation into the use of multivariate data analysis on the spectral data of fluorocarbon

compounds. The first is a normal mixture design spanning the mixture range uniformly (Table 3.3) while the second is a mixture design (Table 3.4) with interaction and square terms added to test for possible interactions between mixture components. For these Phase 2 samples the values in the Tables represent mass percentage.

Table 3.2: Experimental design for Phase one in volume units of 0.25 *ml*.

Sample	A	B	C'	D	Sample	A	B	C'	D
1a	1	0	0	0	30abd	0.2	0.4	0	0.4
2b	0	1	0	0	31abc	0.2	0.2	0.6	0
3c	0	0	1	0	32abcd	0.2	0.2	0.4	0.2
4d	0	0	0	1	33abcd	0.2	0.2	0.2	0.4
5abcd	0.25	0.25	0.25	0.25	34abd	0.2	0.2	0	0.6
6ab	0.8	0.2	0	0	35ac	0.2	0	0.8	0
7ac	0.8	0	0.2	0	36acd	0.2	0	0.6	0.2
8ad	0.8	0	0	0.2	37acd	0.2	0	0.4	0.4
9ab	0.6	0.4	0	0	38acd	0.2	0	0.2	0.6
10abc	0.6	0.2	0.2	0	39ad	0.2	0	0	0.8
11abd	0.6	0.2	0	0.2	40bc	0	0.8	0.2	0
12ac	0.6	0	0.4	0	41bd	0	0.8	0	0.2
13acd	0.6	0	0.2	0.2	42bc	0	0.6	0.4	0
14ad	0.6	0	0	0.4	43bcd	0	0.6	0.2	0.2
15ab	0.4	0.6	0	0	44bd	0	0.6	0	0.4
16abc	0.4	0.4	0.2	0	45bc	0	0.4	0.6	0
17abd	0.4	0.4	0	0.2	46bcd	0	0.4	0.4	0.2
18abc	0.4	0.2	0.4	0	47bcd	0	0.4	0.2	0.4
19abcd	0.4	0.2	0.2	0.2	48bd	0	0.4	0	0.6
20abd	0.4	0.2	0	0.4	49bc	0	0.2	0.8	0
21ac	0.4	0	0.6	0	50bcd	0	0.2	0.6	0.2
22acd	0.4	0	0.4	0.2	51bcd	0	0.2	0.4	0.4
23acd	0.4	0	0.2	0.4	52bcd	0	0.2	0.2	0.6
24ad	0.4	0	0	0.6	53bd	0	0.2	0	0.8
25ab	0.2	0.8	0	0	54cd	0	0	0.8	0.2
26abc	0.2	0.6	0.2	0	55cd	0	0	0.6	0.4
27abd	0.2	0.6	0	0.2	56cd	0	0	0.4	0.6
28abc	0.2	0.4	0.4	0	57cd	0	0	0.2	0.8
29abcd	0.2	0.4	0.2	0.2					

Table 3.3: Mixture design one for Phase two in mass percentage units

Sample	A	B	C	D	E	F	Sample	A	B	C	D	E	F
1	100	0	0	0	0	0	**31**	0	33	0	67	0	0
2	67	33	0	0	0	0	**32**	0	33	0	33	33	0
3	67	0	33	0	0	0	**33**	0	33	0	33	0	33
4	67	0	0	33	0	0	**34**	0	33	0	0	67	0
5	67	0	0	0	33	0	**35**	0	33	0	0	33	33
6	67	0	0	0	0	33	**36**	0	33	0	0	0	67
7	33	67	0	0	0	0	**37**	0	0	100	0	0	0
8	33	33	33	0	0	0	**38**	0	0	67	33	0	0
9	33	33	0	33	0	0	**39**	0	0	67	0	33	0
10	33	33	0	0	33	0	**40**	0	0	67	0	0	33
11	33	33	0	0	0	33	**41**	0	0	33	67	0	0
12	33	0	67	0	0	0	**42**	0	0	33	33	33	0
13	33	0	33	33	0	0	**43**	0	0	33	33	0	33
14	33	0	33	0	33	0	**44**	0	0	33	0	67	0
15	33	0	33	0	0	33	**45**	0	0	33	0	33	33
16	33	0	0	67	0	0	**46**	0	0	33	0	0	67
17	33	0	0	33	33	0	**47**	0	0	0	100	0	0
18	33	0	0	33	0	33	**48**	0	0	0	67	33	0
19	33	0	0	0	67	0	**49**	0	0	0	67	0	33
20	33	0	0	0	33	33	**50**	0	0	0	33	67	0
21	33	0	0	0	0	67	**51**	0	0	0	33	33	33
22	0	100	0	0	0	0	**52**	0	0	0	33	0	67
23	0	67	33	0	0	0	**53**	0	0	0	0	100	0
24	0	67	0	33	0	0	**54**	0	0	0	0	67	33
25	0	67	0	0	33	0	**55**	0	0	0	0	33	67
26	0	67	0	0	0	33	**56**	0	0	0	0	0	100
27	0	33	67	0	0	0	**57**	17	17	17	17	17	17
28	0	33	33	33	0	0	**58**	17	17	17	17	17	17
29	0	33	33	0	33	0	**59**	17	17	17	17	17	17
30	0	33	33	0	0	33							

Table 3.4: Mixture design two for Phase two in mass percentage units.

Sample	A	B	C	D	E	F	Sample	A	B	C	D	E	F
1	100	0	0	0	0	0	**37**	0	33	0	0	33	33
2	0	100	0	0	0	0	**38**	0	0	33	33	33	0
3	0	0	100	0	0	0	**39**	0	0	33	33	0	33
4	0	0	0	100	0	0	**40**	0	0	33	0	33	33
5	0	0	0	0	100	0	**41**	0	0	0	33	33	33
6	0	0	0	0	0	100	**42**	25	25	25	25	0	0
7	50	50	0	0	0	0	**43**	25	25	25	0	25	0
8	50	0	50	0	0	0	**44**	25	25	25	0	0	25
9	50	0	0	50	0	0	**45**	25	25	0	25	25	0
10	50	0	0	0	50	0	**46**	25	25	0	25	0	25
11	50	0	0	0	0	50	**47**	25	25	0	0	25	25
12	0	50	50	0	0	0	**48**	25	0	25	25	25	0
13	0	50	0	50	0	0	**49**	25	0	25	25	0	25
14	0	50	0	0	50	0	**50**	25	0	25	0	25	25
15	0	50	0	0	0	50	**51**	25	0	0	25	25	25
16	0	0	50	50	0	0	**52**	0	25	25	25	25	0
17	0	0	50	0	50	0	**53**	0	25	25	25	0	25
18	0	0	50	0	0	50	**54**	0	25	25	0	25	25
19	0	0	0	50	50	0	**55**	0	25	0	25	25	25
20	0	0	0	50	0	50	**56**	0	0	25	25	25	25
21	0	0	0	0	50	50	**57**	20	20	20	20	20	0
22	33	33	33	0	0	0	**58**	20	20	20	20	0	20
23	33	33	0	33	0	0	**59**	20	20	20	0	20	20
24	33	33	0	0	33	0	**60**	20	20	0	20	20	20
25	33	33	0	0	0	33	**61**	20	0	20	20	20	20
26	33	0	33	33	0	0	**62**	0	20	20	20	20	20
27	33	0	33	0	33	0	**63**	58	8	8	8	8	8
28	33	0	33	0	0	33	**64**	8	58	8	8	8	8
29	33	0	0	33	33	0	**65**	8	8	58	8	8	8
30	33	0	0	33	0	33	**66**	8	8	8	58	8	8
31	33	0	0	0	33	33	**67**	8	8	8	8	58	8
32	0	33	33	33	0	0	**68**	8	8	8	8	8	58
33	0	33	33	0	33	0	**69**	17	17	17	17	17	17
34	0	33	33	0	0	33	**70**	17	17	17	17	17	17
35	0	33	0	33	33	0	**71**	17	17	17	17	17	17
36	0	33	0	33	0	33							

3.3. Sample preparation

Nuclear magnetic resonance (NMR) insert vials with a maximum capacity of 0.25ml were used as sample holders (purchased from Separations). Note that in the Phase one design, volumes were used to measure components. Density differences and consequently minor uncertainties in concentrations was not a concern for this screening experiment that focussed on convenience and speed of sample preparation. The values in Table 3.2 are in units of 0.25 ml. The number 1 in the table thus represents 0.25 ml in the NMR vial whereas the number 0.25 in the table represents 0.0265 ml in the vial.

For the Phase two experiment, the volumes of components were transferred to the NMR vials through the use of a micropipette. Thereafter, the actual masses in gram in the NMR vials were measured to three decimal places on an electronic balance (as difference between initial and final vial mass). “Softgrip” Hamilton pipettes of 1-10µl (0.02inc) and 10-100µl (0.2inc) and a Shimadzu, Uni Bloo (UW420H) weight balance were used. The average mass per NMR vial was 0.3 g in total.

Three centred samples were prepared, one of which was repeated three times in spectral measurements, to assess the magnitude of the statistical error introduced by sample preparation. The difference between the sample preparation variance and instrumental variance can be assessed by comparing the three different samples to three measurements on a single standard and then using equation 3.1.

$$\sigma_{total}^2 = \sigma_{sample}^2 + \sigma_{instrument}^2 \qquad (3.1)$$

Where σ is the standard deviation as was determined over the sets of three samples or measurements.

An extra error term was determined, namely, the variance caused by the NMR tube in NIR. This was done by rotating a vial and taking a measurement at each 90° rotation (Figure 3.1).

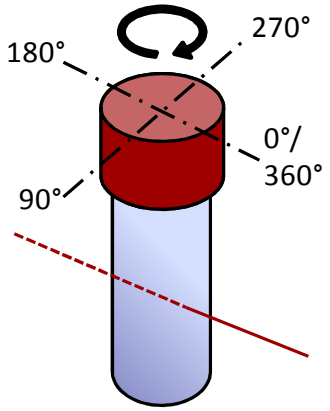


Figure 3.1: Rotation of NMR tube to test for variability.

Raman, Near Infrared (NIR) and Attenuated Total Reflection Infrared (ATR-IR) spectra were collected for each sample. While the instrumentation is discussed in the next section, the following sample holders were used.

For Raman and NIR measurements, the NMR vials were used as samples cells. The Raman modality of the instrument has a sample holder stage that fits the NMR vial. For the NIR modality, a make-shift sample holder (Figure 3.2) was built to fit into the standard sample holder slit of the instrument. For the ATR-IR measurements, a liquid flow-through cell was closed off with stoppers and used in a stationary mode. Ethanol was used as cleaning agent and N_2 gas was used to dry sample cells and pipettes.

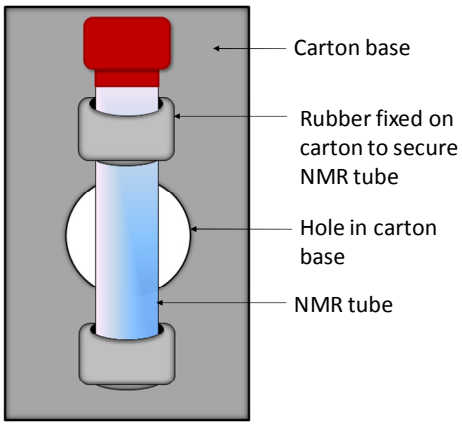


Figure 3.2: Make-shift NIR sample-holder

3.4. Spectroscopic methods

The pure component spectra are given in Appendix A. Instruments from Bruker Optics were used with an OPUS software interface. Further detail on the instrumentation used for the respective spectrometric methods were as follows:

3.4.1. Raman

The Raman spectrometer was a Bruker Optics FT-RAMAN/IR, Vertex 70 series instrument. This instrument has two application modalities, one for Raman and the other for infrared. Each modality has its own sample stage with sample holders as discussed above. As the name indicates, the instrument utilises the Fourier transform technique (rather than grating or prism dispersion elements) to generate the spectra from the incoming radiation, as discussed in Section 2.1.

The Raman modality of the instrument was fitted with optical components that provide access to the wave number range 96.2 cm^{-1} – 4838.9 cm^{-1} with an absolute spectral resolution of 8 cm^{-1}. A 500 mW NdYAG laser provided the interaction radiation at a wave length of 1064 nm while a LN-Ge Diode detector was used.

The output used in the X-matrix for the analysis (Section 2.2) was simply the direct Raman spectral intensities in the selected spectral range. To record one spectrum an average over 256 scans were selected for each sample. The full spectrum over the range was recorded without omitting any wave numbers.

3.4.2. NIR

For the NIR recordings, an IR modality of the Vertex 70 instrument was used. This modality has a He-Ne laser for accurate calibration of the interferometer (Section 2.1) and is equipped with a TE-DLaTGS detector. It was operated in the near infrared spectral range of 6995.873 - 366.781cm^{-1} with resolution 8 cm^{-1}. Spectra were recorded as absorbance measurements, averaged over 32 scans.

3.4.3. ATR-IR

The ATR-IR instrument used was a Bruker Optics FT-IR, Tensor 27, which is equipped with a He-Ne laser source that allows accurate calibration of the total reflection conditions and geometry. A RT-DLaTGS detector was used to record the spectral range of 5994.416 cm^{-1} - 397.313 cm^{-1} with an absolute resolution of 8 cm^{-1}. Reflectance spectra were recorded averaged over 32 scans.

Reflectance values should normally be linearised through an appropriate logarithmic transformation before the linear multivariate techniques are applied. However, it was found that the use of raw reflectance data resulted in a better PLS2 fit than could be achieved for the logarithmically transformed data set. This is most probably due to the fact that the ATR spectra are very close to absorbance spectra already (Pike technologies, 2011). As application of the PLS method to the raw data was empirically found to give the best fit, and the resulting model was able to predict concentrations with good accuracy, it was decided to perform analyses on the raw ATR-IR data.

3.5. Multivariate data analysis methods

Spectral data were imported into the ‘‘The Unscrambler’’ Software as the X-matrix for calibration and the volumes (Phase one experiment) or mass (Phase two experiment) were imported as Y-matrix variables.

Line plots of the spectra were drawn for pure compounds and mixtures and the wave number regions rich in spectral information were selected as the X-variable set as discussed in Chapter 4.

For the Phase one experiment a principal component analysis (PCA) and partial least squares regression (PLS2) was done on the data sets and three random samples were used as test samples for prediction.

For the Phase two experiment, a much more systematic and detailed approach was adopted. This is described in more detail in Chapter 4 where the data can provide the clarifying context. The general approach was as follows:

A first selection of spectral regions were analysed using all the samples with PCR, PLS1 and PLS2 to search for principle components, outliers, important variables, and to assess goodness of fit. These were repeated for the three sample design options listed below to find the over-all best combination of parameters for the final Phase two experiment:

1) Normal mixture design (Table 3.3).
2) Design with interaction and square terms added (Table 3.4).
3) Design with interaction and square terms (Table 3.4) with regression calculations that included these selected terms in "The Unscrambler".

R^2 and root mean square error terms were evaluated to decide on the best model parameters and outliers were detected and removed.

Two sets of test samples were used throughout, namely internal and external sets. The internal set comprised of a selection of samples that were part of the design matrix, whereas the external set was separately prepared (i.e. was not used in calibration). Some external test samples were prepared to fall outside the calibration range to allow the assessment of the extrapolation capability of the model. The compositions of test samples are given in Table 3.5.

Table 3.5: Internal (S) and external (T) test sample composition in terms of the Phase two compound group (units in gram).

	T1	T2	T3	T4	T5	S1	S2	S3	S4	S5
A	0.025	0.008	0.092	0	0	0	0.097	0.063	0	0.097
B	0	0.015	0.144	0	0.087	0	0	0.055	0.097	0.097
C	0.266	0.008	0	0.072	0	0	0.098	0.06	0	0
D	0.02	0.134	0	0.066	0	0.089	0	0	0.086	0
E	0	0.056	0.06	0.029	0	0.092	0	0.059	0	0.095
F	0	0.048	0	0.097	0.175	0.081	0.082	0.047	0.079	0

Vector normalisation (Section 2.2) as well as taking the first derivative of the NIR data were explored as possible pre-treatment options.

Once the set of parameters best suited for each spectroscopic method had been found, a PCR and PLS2 regression model was developed using these parameters. The net result of this process was a considerable reduction of complexity of the final Phase two experiment. Predictions were done on the external and internal test sets.

The pre-analysis for the Phase two experiment is schematically summarised in Figure 3.3.

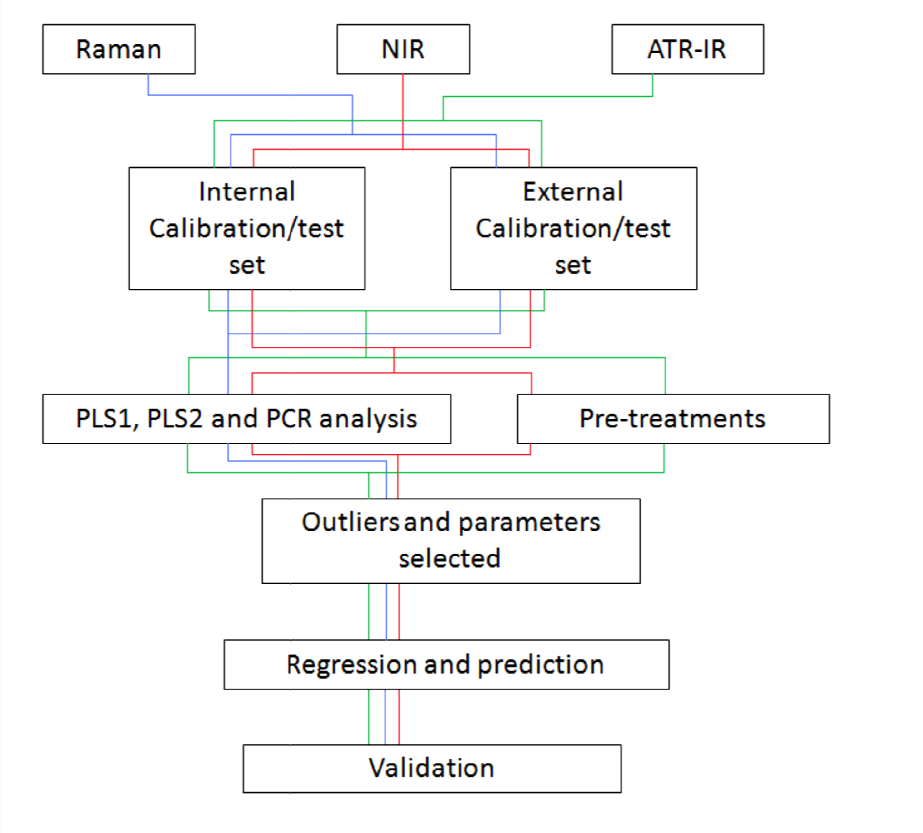


Figure 3.3: Schematic process used to choose a set of model parameters.

Validation was done using cross-validation with 13 segments (10 - 12 samples per segment) for analysis and regression results as described in Section 2.2. Final validation was achieved by comparing a two-segment cross-validation, a thirteen- segment cross-validation and a full cross-validation.

Other graphical representations of the regression data were used to evaluate possible further information that could be gained from multivariate analysis. As mentioned previously some details surrounding the methods are left for further discussion in the results chapter (Chapter 4) where it can be introduced in the context of the data and the analysis thereof.

3.6. Bibliography

1. PIKE TECHNOLOGIES. ATR theory and applications, Application Note. http://www.piketech.com/files/pdfs/ATRAN611.pdf. Date of access: 10 Oct 2011.

Chapter 4. Results and discussion

Index

4.1. Introduction

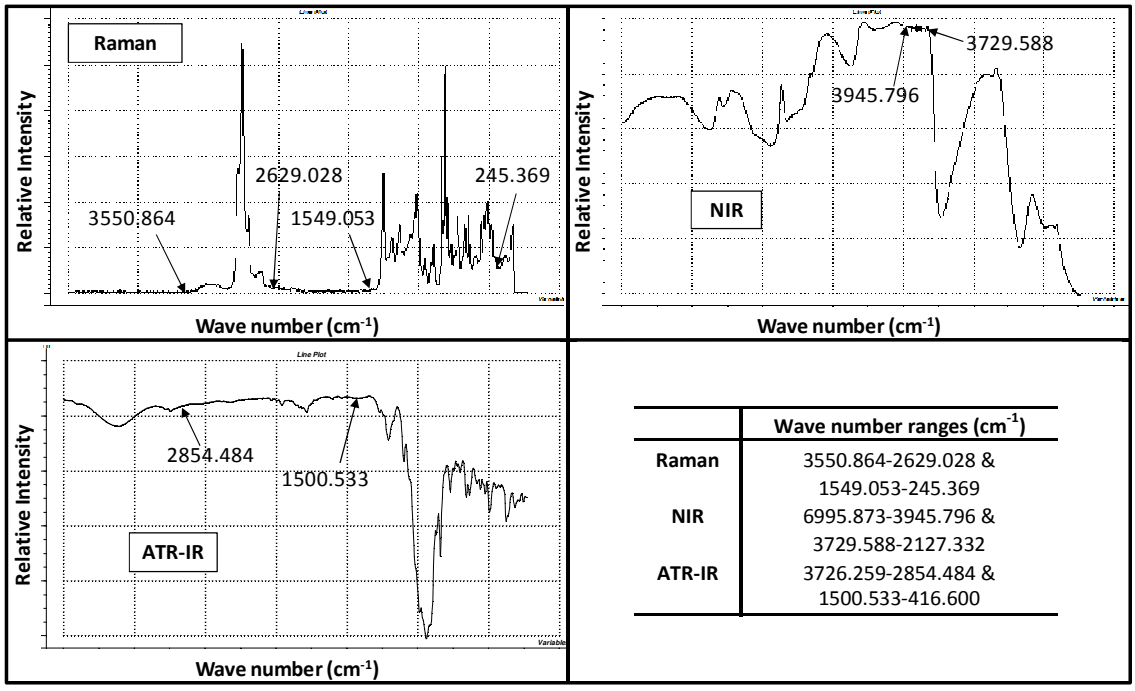
In this chapter the spectra generated from three spectroscopic techniques (Raman, NIR and ATR-IR) are represented. The results for a screening experiment that involves mixtures of four fluorocarbon alcohols (Phase one) as well as a subsequent, more detailed experiment that involves mixtures of six fluorocarbon alcohols (Phase two) are shown. Principal component regression (PCR) and partial least squares regression models were used to analyse data, to construct calibrations and to predict test samples.

The spectra, regression results and prediction results of all the experiments are listed in full in the appendixes and only selected tables and figures are shown in this chapter to serve as representative results.

The results showed that PCR and PLS2 regression could be used to predict the composition of mixtures of six fluorocarbon alcohols. It also served to illustrate the efficiency and accuracy of multivariate calibration as well as the ability of the method to identify outliers and to reveal important correlations and data structures.

4.2. Spectra

Raman, NIR and ATR-IR spectra for a typical centre sample that contains all components are shown in Figure 4.1. The respective pure component spectra are given in Appendix A. For each of the spectra there are areas with little or no information or areas where there is an irregular noise contribution (Such as the 3945.796 - 3729.588 cm^{-1} region in the NIR spectra). These areas, as selected by inspection, were omitted from the spectral data set before multivariate analysis. Also shown in Figure 4.1 are the wave number ranges used in this study for each of the spectroscopic methods.



	Wave number ranges (cm^{-1})
Raman	3550.864-2629.028 & 1549.053-245.369
NIR	6995.873-3945.796 & 3729.588-2127.332
ATR-IR	3726.259-2854.484 & 1500.533-416.600

Figure 4.1: Raman, NIR and ATR-IR spectra of a centre sample and the wave number regions selected.

The absorption peaks of the near infrared spectra are signal saturated due to the long path length of the nuclear magnetic resonance (NMR) tube that was used as sample cell. For qualitative spectral analysis these spectra would not be useful but since there are still differences in the data structure between samples, it might still be possible to do a multivariate analysis. This is indeed one of the great advantages of multivariate analysis above other more simple calibration methods.

The reason for using the NMR tubes despite the obvious disadvantage is two-fold. Firstly, for this study the objective was to assess the possible use of each spectroscopic technique in multivariate analysis and not to optimise the technique. By using the NMR tubes the power of multivariate analysis as a useful quantitative method can be investigated, even under non-ideal conditions. Secondly, a large number of samples were used in this study and using the liquid cells, with small path lengths, that are currently commercially available would have been time consuming to the point of being impractical. If the NIR spectroscopic technique is to be used for future developments, where emphasis is on optimisation of accuracy, a sample cell with a smaller path length should be chosen.

4.3. Four components (Phase one) viability experiment

To ascertain the overall viability of the method a four component experiment was conducted. Overall viability includes assessment of linearity of calibration, size of the error as well as considering whether the method is doable in practise. For this experiment a small sample set (51 samples) was used in a mixture design that covers the mixture range uniformly and only Raman and ATR-IR spectra were used (Chapter 3). A principle component regression (PCR) and partial least squares regression (PLS) were done on both spectroscopic data sets. The two regression models gave comparable results and only PLS2 results are shown in Table 4.2 and Table 4.3 as an example. The layout and interpretation of values in these tables (and other regression -result tables given later in this chapter) are explained in Table 4.1.

The R^2 (R-square) values in the PLS2 regression models for both Raman and ATR-IR spectra are close to 1 which indicates a good linear relation between the concentration of components in a sample and the spectral data set. For both spectroscopic techniques the adjusted R^2 values are also close to 1 which means that a good fit for future prediction can be expected.

The measurement units of the root mean square error of calibration and the root mean square error of prediction (RMSEC and RMSEP) are the same as the measurement unit of

the components (volume percent). When RMSEC and RMSEP have similar values it is an indication that the model will perform well for future predictions.

The lowest concentration value for any component in the 51 samples is 20% and amongst all the RSME values given, the highest value is 3.53% (Amongst ATR-IR results, Table 4.3). The maximum observed relative error therefore is around 18%.

Predictions were done with the PLS2 regression model on selected samples (S1-S6) and the results are listed in Table 4.4 and Table 4.5. The range of standard deviations is between 0.8 and 3.5 (in percent volume units) and the component's concentrations range from 0% to 100%. The predictions and the linearity of fit are sufficiently good to justify further experimentation for testing, in more detail, the use of multivariate analysis in the quantitative analysis of fluorocarbon alcohols.

Table 4.1: The layout and interpretation of tables showing regression results.

Sample components

	A B C D E F
R-sqr model	The raw R-square value of the model. Should be between 0.9 and 1 for a good model
R-sqr adjusted	Tells us how good fit can we expect for future predictions.
RMSEC	The root mean square error of calibration
RMSEP	The root mean square error of prediction
RMSEP-RMSEC	The smaller this value is, the better the calibration model is for future predictions.
PC's	The suggested amount of PC's that should be used

Root mean square error terms are in the same unit as the measurement unit of the components

Table 4.2: Regression result of PLS2 regression on Raman data of Phase one experiment.

Ramam PLS2	A	B	C'	D
R-sqr model	0.99	1.00	0.99	1.00
R-sqr adjusted	0.99	1.00	0.99	1.00
RMSEC	2.37	1.17	2.26	1.22
RMSEP	2.78	1.36	2.69	1.41
RMSEP-RMSEC	0.40	0.19	0.43	0.18

Table 4.3: Regression results of PLS2 regression on ATR-IR data Phase one experiment.

ATR-IR PLS2	A	B	C'	D
R-sqr model	0.99	0.99	0.99	0.98
R-sqr adjusted	0.99	0.99	0.99	0.98
RMSEC	3.01	2.53	2.02	3.31
RMSEP	3.23	3.02	2.37	3.53
RMSEP-RMSEC	0.22	0.49	0.35	0.22

Table 4.4: Prediction on selected samples (S1-S6) of the Phase one experiment using the PLS2 regression model for Raman data.

Raman PLS2		S1	S2	S3	S4	S5	S6
A	PV	59.0	40.2	40.6	20.9	18.5	0.9
	MV(%)	60.0	40.0	40.0	20.0	20.0	0.0
	SDevM	2.4	1.5	2.2	3.0	2.0	2.5
	PV-MV	-1.0	0.2	0.6	0.9	-1.5	0.9
B	PV	41.1	19.1	-0.1	81.0	-0.4	41.1
	MV(%)	40.0	20.0	0.0	80.0	0.0	40.0
	SDevM	1.2	0.7	1.1	1.4	0.9	1.2
	PV-MV	1.1	-0.9	-0.1	1.0	-0.4	1.1
C'	PV	0.1	20.9	1.0	-1.6	21.2	58.7
	MV(%)	0.0	20.0	0.0	0.0	20.0	60.0
	SDevM	2.4	1.5	2.2	2.9	1.9	2.5
	PV-MV	0.1	0.9	1.0	-1.6	1.2	-1.3
D	PV	-0.1	19.9	58.6	-0.4	60.7	-0.7
	MV(%)	0.0	20.0	60.0	0.0	60.0	0.0
	SDevM	1.3	0.8	1.1	1.5	1.0	1.3
	PV-MV	-0.1	-0.1	-1.4	-0.4	0.7	-0.7

PV: Value predicted by model for the component.

MV(%): The measured value of the component indicated as percent volume of 0.25ml sample.

SDevM: Deviation in prediction values within the model.

Table 4.5: Prediction on selected samples (S1-S6) of the Phase one experiment using the PLS2 regression model for ATR-IR data.

ATR-IR PLS2		S1	S2	S3	S4	S5	S6
A	PV	62.7	42.0	42.8	27.4	23.1	1.9
	MV(%)	60.0	40.0	40.0	20.0	20.0	0.0
	SDevM	2.6	2.2	2.5	3.5	2.7	3.5
	PV-MV	2.7	2.0	2.8	7.4	3.1	1.9
B	PV	42.5	19.9	-1.9	74.8	-1.4	42.6
	MV(%)	40.0	20.0	0.0	80.0	0.0	40.0
	SDevM	2.4	2.1	2.3	3.2	2.5	3.2
	PV-MV	2.5	-0.1	-1.9	-5.2	-1.4	2.6
C'	PV	-2.8	19.1	0.8	-2.2	20.7	59.0
	MV(%)	0.0	20.0	0.0	0.0	20.0	60.0
	SDevM	1.9	1.6	1.8	2.6	2.0	2.6
	PV-MV	-2.8	-0.9	0.8	-2.2	0.7	-1.0
D	PV	-2.4	19.0	58.3	0.0	57.5	-3.5
	MV(%)	0.0	20.0	60.0	0.0	60.0	0.0
	SDevM	2.8	2.4	2.6	3.8	2.9	3.8
	PV-MV	-2.4	-1.0	-1.7	0.0	-2.5	-3.5

PV: Value predicted by model for the component.

MV(%): The measured value of the component indicated as percent volume of 0.25ml sample.

SDevM: Deviation in prediction values within the model.

The conclusions drawn from Tables 4.4 and 4.5 can be depicted visually through the use of simple linear regression on the multivariate regression data of Tables 4.4 and 4.5 for the Raman and ATR-IR data respectively. In Figure 4.2 the results are shown for both methods and for two types of plots. The first is a plot of predicted values versus measured values and the second is a plot of the error versus measured values. Here the error refers to the third row (PV-MV) values in the respective tables. If the PLS2 predictions are accurate over all samples and over the whole concentration range, the first plot should have a slope of 45^0 (slope = 1), and an intercept of zero. The linear regression results confirm a good accuracy for both methods.

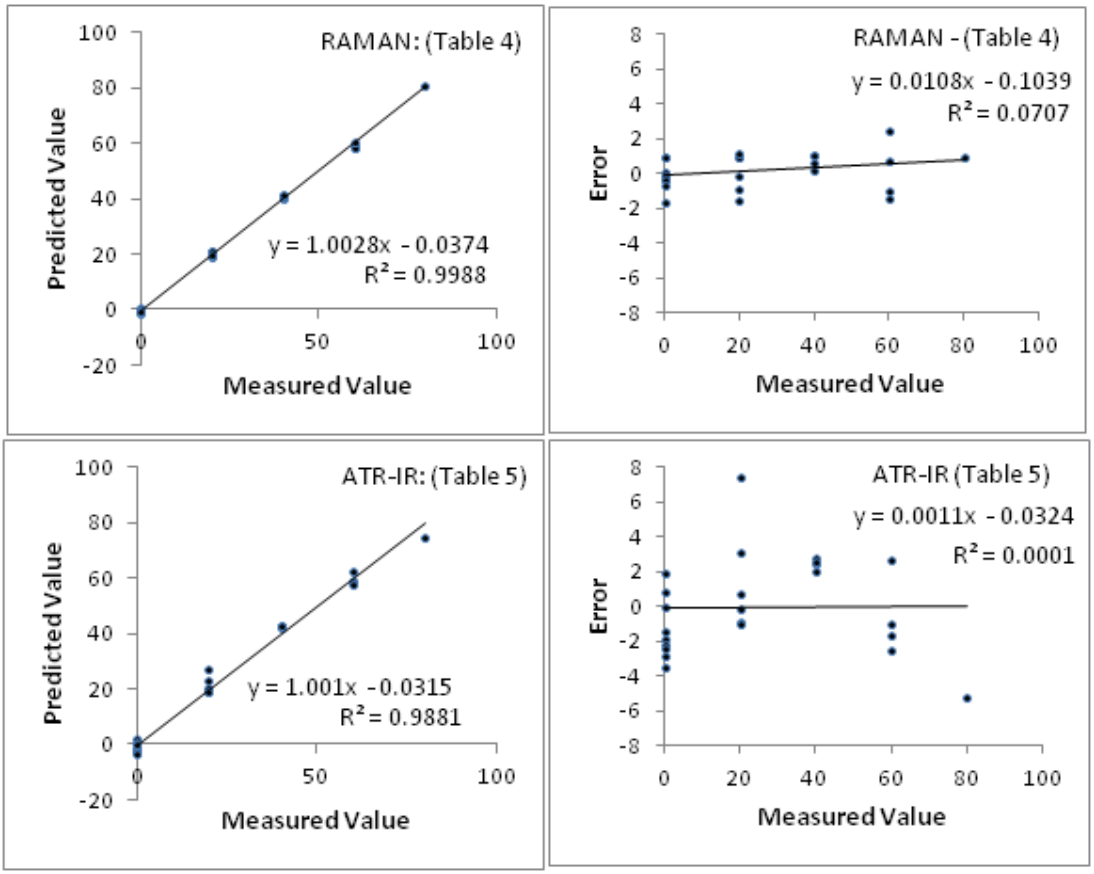


Figure 4.2: Plots of predicted values and error values, versus the measured values for both Raman and ATR-IR Phase one results.

The second plot type provides insight into error behaviour and further supports accuracy over the full range of samples and concentration values of Tables 4.4 and 4.5 for both methods. Simple linear regression reveals a slope and intercept of practically zero and, $R^2 \cong 0$. The values of the first two parameters support nearly equal distribution of data points above and below the zero line with no discernable concentration dependence on the error whereas the latter supports a completely random relation between concentration and error. The linear regression results of Figure 4.2 strongly support good prediction with inherently random prediction errors (no systematic effects) and with an average value independent of concentration for the PLS2 fit.

4.4. Six components (Phase two) experiment - Analysis

As described in the Chapter 3, the Phase two experiment was performed, starting with an analysis step that involved making a number of regression models (variations of PLS1, PLS2 and PCR models) to find a suitable set of parameters (for example the number of PC's to be used, pre-treatments that should be done and outliers identification). Appendix B contains tables of regression results that were used to find suitable parameters and are listed in the format described in Table 4.1.

The best model is defined as the one that yields R^2 values closest to 1 and the lowest RMSEP, RMSEC and RMSEP-RMSEC values.

The values for all the analysis models were similar and good (all yielding $R^2 > 0.9$). Slightly better values were achieved when interaction and square terms were added to the experimental design and the regression calculation (Particularly for NIR an ATR-IR data). The differences between the models with interaction and square terms and those without are very small though (<0.02), which means that the interactions do not play a significant role in the multivariate analysis of these mixtures.

4.5. Six components (Phase two) experiment - Regression and prediction

The experimental design with interaction and square terms added was used for compiling all subsequent regression models. A PLS2 and PCR regression model was constructed for all three spectroscopic techniques with parameters laid out as in Table 4.6, as determined by the analysis results. Although the PLS2 model gave slightly better regression results than the PCR model, the results were comparable and the PCR model was also retained for predictions because of the difference in mechanism between PCR and PLS2 and the subsequent effect it could have on predictions (Chapter 2). The regression and prediction results for both model types are listed in Appendix C. In this chapter only PLS2 results are shown and discussed further.

The outlier samples were detected using graphical representations such as residual sample variance plots and an X-Y influence plot (Figure 4.3) as produced by the ''The Unscrambler''

software. In the influence plot sample B2 can be seen to lie far removed from the other samples, which cluster together closely. In the residual sample variance plot, sample B2 has a residual sample variance much higher than the other samples. Since the sample sets are very large (close to a hundred), outliers, after identification, can safely be omitted from the data set without significant influence on the accuracy of the model.

Table 4.6: Amount of principle components to be included in regression models and outliers to be left out of calibrations.

	Raman	NIR	ATR-IR
PLS2, PC's	5	8	7
PCR, PC's	5	9	7
Outliers	S3, B2	S52, S54, S61, S72, S92	S24, S37

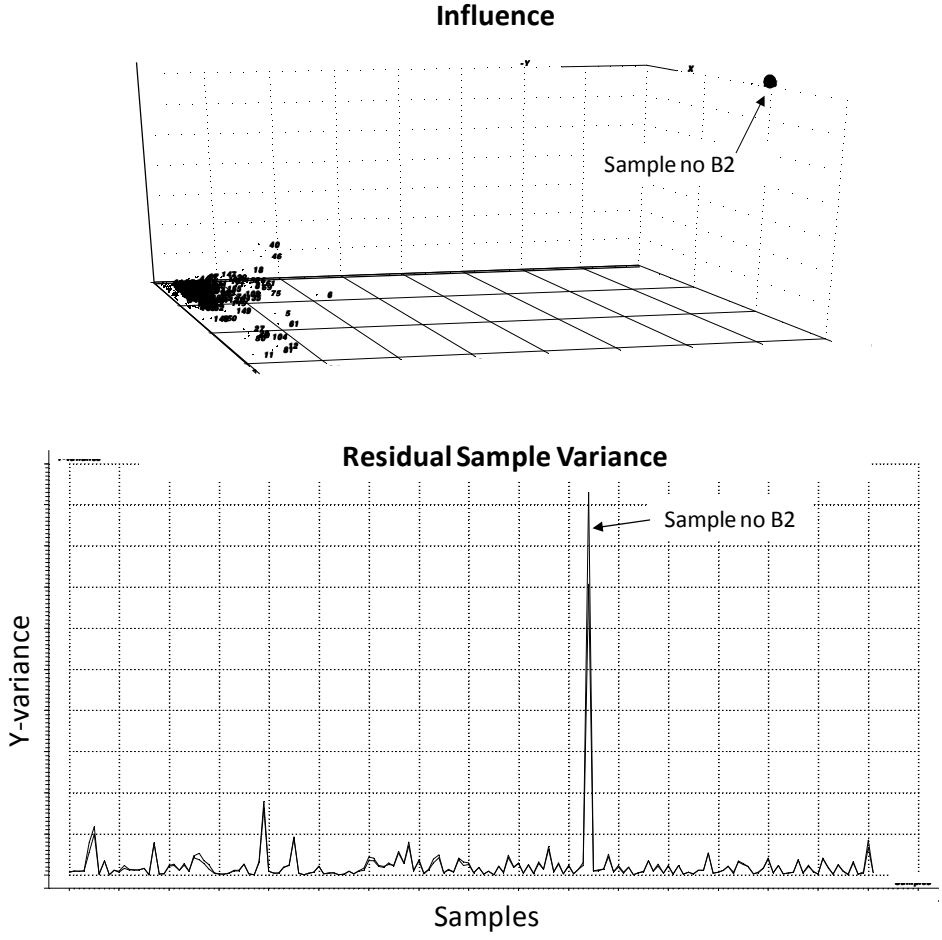


Figure 4.3: Outlier detection in "The Unscrambler".

As a representative example, the regression results for the PLS2 model done on the Raman spectral data are shown in Table 4.7. Both Internal and external data sets are shown (See Chapter 3 for explanation of differences between internal and external data sets).

The R^2 and error terms of the regressions done on the external data sets are very close to the R^2 and error terms of regressions done on the internal sets. This allows direct comparison of predicted results of the external and internal test sets. All the R^2 values are greater than 0.90, which indicates a good linear fit with good future prediction capability.

Table 4.7: Result for a PLS2 model on Raman data (One of the result tables listed in Appendix C).

	Ramam PLS2	A	B	C	D	E	F
External	R-sqr model	0.9972	0.9951	0.9949	0.9956	0.9887	0.9939
	R-sqr adjusted	0.9970	0.9944	0.9945	0.9950	0.9872	0.9930
	RMSEC	0.0031	0.0036	0.0042	0.0034	0.0057	0.0037
	RMSEP	0.0032	0.0039	0.0044	0.0036	0.0061	0.0039
	RMSEP-RMSEC	0.0001	0.0003	0.0002	0.0002	0.0004	0.0003
Internal	R-sqr model	0.9972	0.9951	0.9949	0.9955	0.9887	0.9939
	R-sqr adjusted	0.9970	0.9945	0.9941	0.9949	0.9875	0.9927
	RMSEC	0.0031	0.0037	0.0042	0.0034	0.0057	0.0037
	RMSEP	0.0033	0.0039	0.0045	0.0037	0.0061	0.0040
	RMSEP-RMSEC	0.0002	0.0002	0.0003	0.0002	0.0003	0.0003
	PC's	5	5	5	5	5	5

Comparing the values of the different regression models it can be seen that the models performed similarly (R^2 values range from 0.9383 to 0.9772). The model dependent differences (between PLS2 and PCR) are not as significant as differences due to adoption of different spectroscopic techniques or due to the difference in component composition. Raman spectroscopy performed the best (average R^2 value of 0.9938 and error terms ranged from 0.0031 to 0.0132). ATR-IR performed second best (average R^2 value of 0.9719 with error terms ranging from 0.0061 to 0.0132). The NIR (average R^2 value of 0.9715 and error terms ranging from 0.0062 to 0.0137) models gave less accurate predictions, which is most probably due to the signal saturation caused by the long path lengths of the NMR

sample cells. A direct comparison of NIR with the other two techniques is thus not possible with the current data. Despite the saturation of the spectra, the NIR nevertheless gave remarkably good results, indicating the great advantage of multivariate analysis to extract data structure even from non-optimal data sets.

The error terms (RMSE's) are in the same units as the Y-variables. In this experiment it is the amount of component given in grams. In the test samples, the components range from zero gram to a maximum amount of between 0.0970 and 0.1750 grams, depending on the component. The error terms range from 0.0031g to 0.0137g depending on the model, spectroscopic technique and choice of component. When a component is present in low concentrations the percentage error is therefore high and for higher concentrations the percentage error decrease, which is expected to result in better predictability. This can be seen reflected in charts of the prediction results (given for all predictions in Appendix C). Figure 4.4 and Figure 4.5 are the charts with prediction result for components A and C respectively shown here as representative examples. In Figures 4.4 and 4.5, T1-T5 are the external test samples averaged over three measurements and S1-S5 are the internal test samples (Table 3.5).

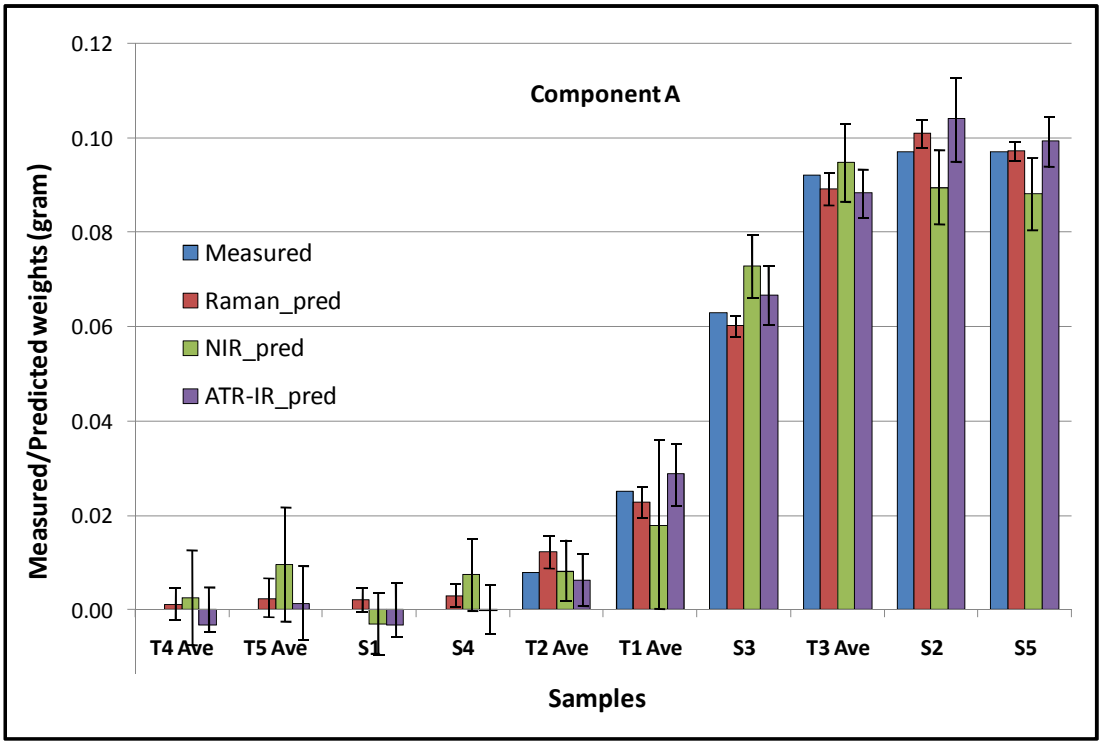


Figure 4.4: Prediction results for component A.

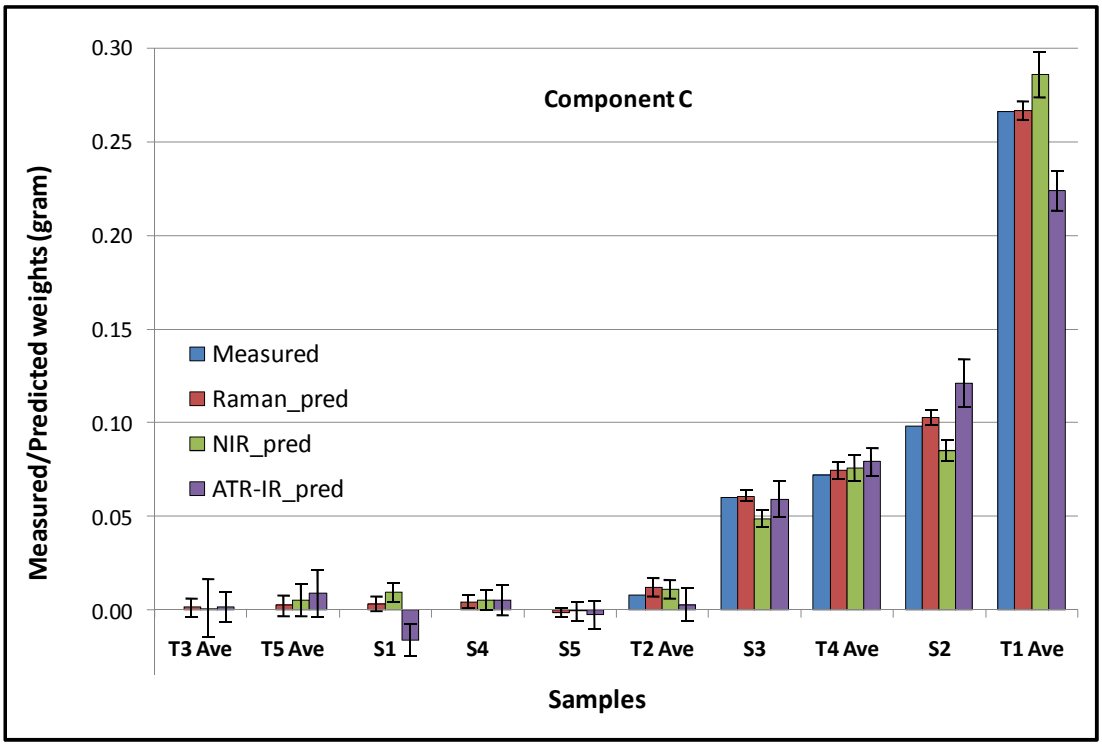


Figure 4.5: Prediction results for component C.

The standard deviation increases as a particular component's concentration decreases. There is no observable association between the standard deviation and the amount of components in the sample or the amounts in gram of other components in the sample. For this reason, using larger sample volumes and keeping the relative concentration of a component high will be an advantage for predicting a sample's composition. In general the internal test set was better predicted than the external test set. This is expected because the internal set falls into the exact scope of the regression model (interpolation) whereas the external set has some values that fall outside the scope of the calibration range (extrapolation). The values outside the calibration range are also amongst the smallest values (in gram) which further ads to the loss of predictive ability.

Expressing the standard deviation as percentage standard deviation makes it easier to interpret the precision of the prediction. The percentage standard deviation is calculated as:

$$\%SDev = \frac{SDev}{Predicted\ value} \times 100$$

The choice of an acceptable standard deviation depends on the application. In this study the composition of each sample needs to be predicted. For illustrative purposes an acceptable limit of 10% is set for the percentage standard deviation.

The smallest value (gram) amongst the predicted components in a sample, for which the percentage standard deviation is under 10%, is 0.047g (For Raman data). This represents 16% of an average sample size of 0.3g. Therefore to obtain predictions using Raman data, with a percentage standard deviation of less than 10%, the component mass must be at least 16% of the total sample size. The minimum concentration for individual components could be calculated for each spectroscopic technique. This allows assignment of a ''prediction limit'' analogous to a detection limit, above which prediction becomes reliable.

4.6. Validation of the models and error contributions

The validation results are given in full in Appendix D1 (Tables D-1 to D-7). Three validation techniques (Cross-validation with 13 and 2 segments and Full cross-validation) were used in each regression model. The root mean square error terms, for Raman and ATR-IR spectral data sets, generally do not differ greatly between the three validation techniques used, which indicate good stable calibrations. The average percentage standard deviation (where percentage standard deviation in this case is taken as the standard deviation divided by the average error value multiplied by 100) for Raman is 1% and for ATR-IR it is 3%. For NIR spectral data the average percentage standard deviation is 5%, which is still good but the variability in relative standard deviation is much larger with a maximum of 37%. This indicates that the models done on NIR spectral data are not as reliable as those for Raman and ATR-IR.

Errors that are introduced during sample preparation and measurement are listed in Appendix D2. Table 4.8, Table 4.9 and Table 4.10 are summarised versions of the error measurements. Table 4.8 shows the average value and standard deviation over three centre samples that were made separately (In gram weighed per volume measured from the pipette). The largest standard deviation is measured for component A and is less than 9% of the corresponding average value.

In Table 4.9 the error introduced by the variation of the spectroscopic method is measured by calculating the average and standard deviation of the minimum and maximum peak intensity, the mean value over intensities and the standard deviation over the spectral range. For the Raman spectra the error is larger for the minimum peak intensities and for NIR the error is lowest for the minimum peak intensities. This is to be expected since the very low intensities in Raman constitute the noise contribution while for NIR the larger intensities are unreliable due to the signal saturation. This is a further indication that the lower predictive performance of the NIR method is most probably due only to the fact that non-optimal absorption cell lengths were used for the method. The method variation is generally less significant than the variation due to weight differences.

In Table 4.10 the error introduced by varying structural interference from the glass NMR tubes was explored by turning a NMR tube filled with ethanol by 90° and taking a measurement at each turn. The variance due to these orientation differences is less than the variance due to repeat measurements (Table 4.9). Therefore the over-all greatest source of variance is in the weighing of components to make the samples. This can be decreased by increasing sample size and keeping the lowest concentration relatively high.

Table 4.8: Average value and standard deviation of each component over the three centre samples.

g/µl	A	B	C	D	E	F
Sdev	1.4E-04	3.5E-05	9.0E-05	1.7E-05	4.6E-05	3.0E-05
Ave	1.6E-03	1.4E-03	1.5E-03	1.3E-03	1.5E-03	1.2E-03

Table 4.9: Average value and standard deviation in peak intensity units of each component over the three repeat measurements of one of the centre samples.

Centre Samples	Min	Max	Mean	Sdev
Sdev Raman	2.8E-07	3.4E-06	3.4E-07	5.5E-07
Average Raman	1.8E-06	5.4E-04	8.4E-05	8.2E-05
Sdev NIR	5.1E-05	1.5E-04	1.0E-04	2.9E-05
Average NIR	1.3E-03	2.3E-03	1.8E-03	3.2E-04
Sdev ATR-IR	3.4E-04	7.9E-04	6.8E-04	1.1E-04
Average ATR-IR	4.2E-03	9.4E-03	8.4E-03	1.2E-03

Table 4.10: The average value and standard deviation of spectral intensities over different orientations of the NMR tube (NIR).

NIR	Min	Max	Mean	Sdev
Sdev	0.0024	0.0040	0.0036	0.0004
Average	0.1926	0.3279	0.2750	0.0391

4.7. Further information gained from multivariate analysis

Some wave number regions sometimes associates with a specific component. Knowing which wave numbers are closely associated with which component can help to interpret the prediction results better. In Figure 4.6 the score plot with principal components 1 and 2 and the corresponding 2D scatter plot of the loadings of a PLS2 regression model is presented (performed on Raman data set). Component F is circled in green on the score plot. On the loading plot the green circle contains two wave numbers that are associated with component F, namely 793.070 cm^{-1} and 796.927 cm^{-1}. It is not always easy to find wave numbers that specifically associates with a component from loading plots. Regression coefficient plots provide another means to find important variables (Figure 4.7).

One of the wave numbers identified as being important for a specific component can now be used to group the samples in the score plot. If the wave number 793.070 cm^{-1} is truly associated with component F, then the colour grouping should agree with the component's relative composition in the samples. The score plot is also set to show the samples as the value of component F from the Y-matrix rather than the sample name to test the theory. Figure 4.8 shows the resulting score plot. It can indeed be seen that the unit representing the intensity of the wave number 793.070 cm^{-1} decreases (red to blue), as the amount of component F decreases (from 0.249g in red to 0g in blue).

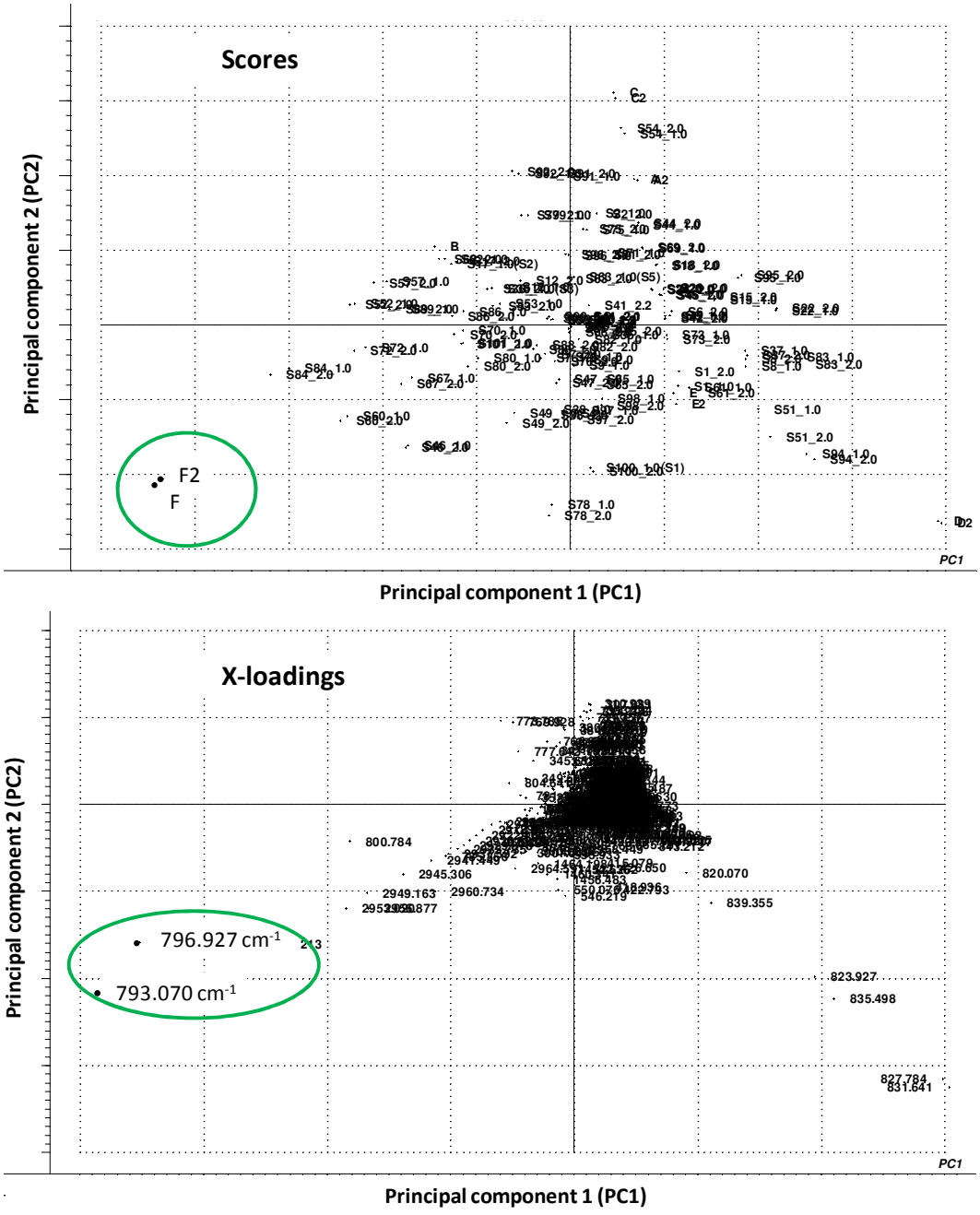


Figure 4.6: Correlation between a component and specific wave numbers

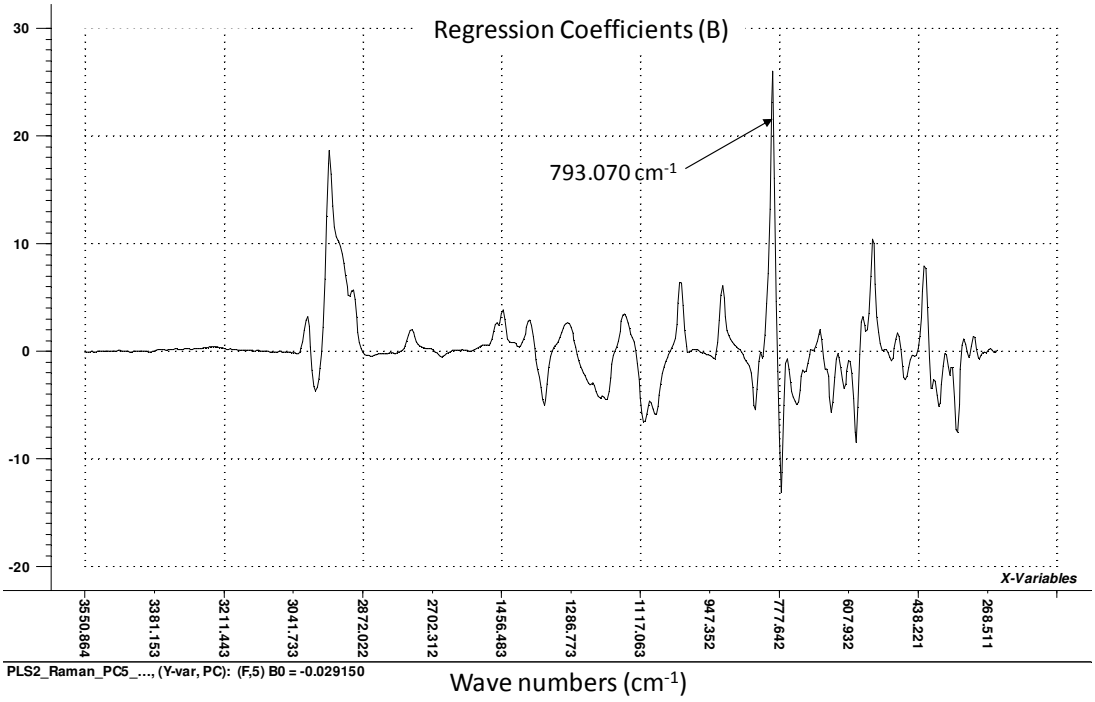


Figure 4.7: Regression Coefficient plot for component F

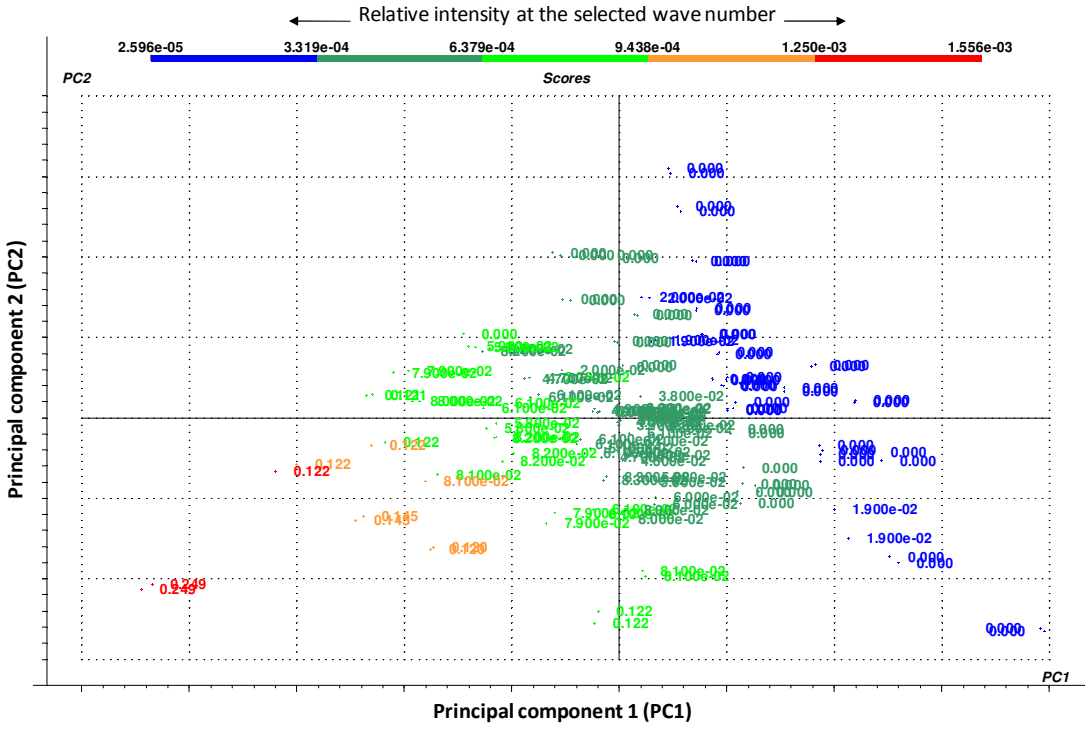


Figure 4.8: Colour grouping done in ''The Unscrambler'' for component F

Other associations that could be found were for component C (735.214 cm^{-1}) and component A (804.641 cm^{-1}), which were used to show the prediction results earlier in this chapter (Figure 4.4 and Figure 4.5). Figures 4.9 and 4.10 show the score plots for components A and C respectively, done in the same manner as for component F in Figure 4.8. It can be seen that component C is strongly correlated to PC2 (strong red to blue trend from high PC2 unit to low PC2 unit). Component A on the other hand is weakly correlated to both PC1 and PC2 (no strong red to blue trend). Components that are strongly correlated to a principle component are expected to have better predictability. A component strongly correlated to one of the early PC's (PC1 for example) should have better predictability than a component strongly correlated to one of the later PC's (PC5 for example). Using this technique some insight into which components will be better predicted can be achieved. The prediction results are indeed, as shown in Figures 4.4 and 4.5, better for component C than for component A.

The wave numbers around the 600-900 cm^{-1} region for Raman are due to C-F, C-H and carbon chain vibrations in 2,2,3,3,3-pentafluoro-1-propanol (component B) and this will be true for similar fluorocarbon alcohols like the ones used in this study. The predictions are predominantly influenced by the fluorocarbon vibrations and not by the alcohol functional group. The study can therefore be seen as a good representative study for the use of multivariate analysis and spectroscopic techniques in the prediction of composition of fluorocarbon compounds.

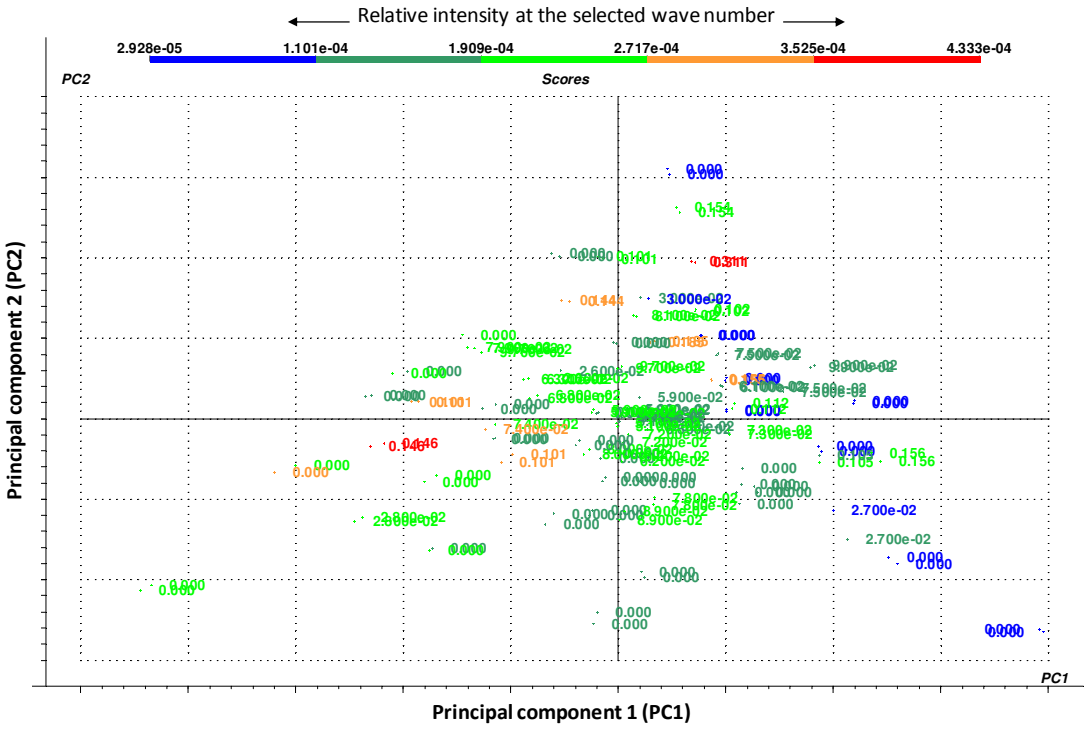


Figure 4.9: Colour grouping done in "The Unscrambler" for component A

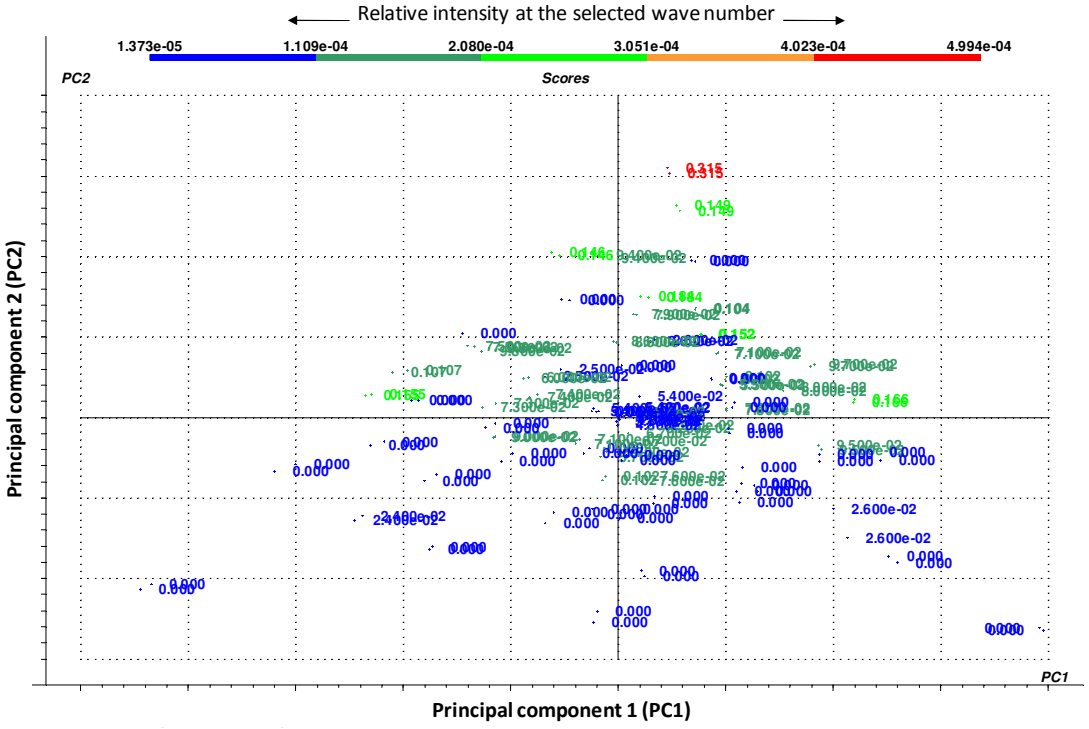


Figure 4.10: Colour grouping done in "The Unscrambler" for component C

Chapter 5. Conclusion and Recommendations

5.1. Conclusions

The main conclusion of this study is that the method, which involves use of multivariate analysis and development of a multivariate regression models was successful in predicting the composition of mixtures of complex fluoro alcohols from Raman, NIR and ATR-IR spectroscopic data.

Analysis and regression models (PCR, PLS2 and PLS1) gave good model fits (R^2 values larger than 0.9). Raman spectroscopy was the most efficient technique and gave good prediction (at 10% accepted standard deviation), provided the minimum mass of a component exceeded 16% of the total sample. The dominant source of error was found to be in the sample preparation step due to the difficulty of measuring small amounts of material.

Although ATR-IR did not perform as well as Raman, it outperformed NIR as applied in this study. However the main factor leading to the decrease in efficiency of using NIR data was shown to be the signal saturation resulting from too long an optical path through the sample. Despite this, NIR still gave remarkably good results, confirming the ability of multivariate analysis to extract information, even from non-optimal data sets.

The study can be seen as a good representative study for the use of multivariate analysis and spectroscopic techniques in the prediction of composition of fluorocarbon compounds. The reason for this is that the wave numbers around the 600 - 900 cm^{-1} region for Raman are expected to also yield good calibrations for other similar fluorocarbons as the predictions are predominantly influenced by the fluorocarbon vibrations and not by the alcohol functional group (Badawi, 2008).

Considering industrial uses this technique could be useful for quantitative screening of fluorocarbon products but identification and quantification of a particularly by-product or

impurities in small concentrations should be done with complementary analytical techniques.

5.2. Recommendations

In this study, the focus was not on improving accuracy of the method but to demonstrate overall feasibility. As a result, there is room for improvement in a number of areas. For overall performance the following improvements are recommended:

The error due to sample preparation can be reduced by using larger samples, from which vials can be filled, instead of preparing small samples directly in the vials as was done in this study. The smallest sample mass can be estimated using the arguments developed in Section 4.6.

The accuracy of the method on NIR data can be improved through optimisation of the absorption path length. When required to design a specific absorber cell to prevent saturation of the signal, the choice of an absorbance value $A \leq 1$ is a reasonable first choice, as it ensures transmission of ≥ 10 %. After the specific absorptivity has been determined (by separate test calibrations) and the concentration range of interest established equation 2.6 can be used to calculate the corresponding cell length. In the current work, the cell length has not been optimised.

The spectral regions selected for the respective spectroscopic techniques, have been chosen based on a visual survey. Exploratory experiments were then performed to find principle components, important variables and outliers. Thereafter a multivariate regression model was developed and the goodness of fit and prediction capability of the model was tested. This was the only quantitative attempt made at selecting a spectral region or making decisions regarding important parameters to use. It is therefore entirely possible that improved models can be constructed for different spectral regions or by using the same regions with more exhaustive variable analyses and subsequent omission of portions of the spectrum from the data set. Recommended techniques to aid such a selection can be found in Section 4.7.

5.3. Bibliography

1. BADAWI, H.M. & FORNER, W., 2008. Solvent dependence of conformational stability and analysis of vibrational spectra of 2,2,3,3,3-pentafluoro-1-propanol. *Spectrochim. Act, Part A*. 71: 388-397.

Acknowledgements

I would like to thank, for their individual contribution to the completion of this M.Sc study and dissertation:

- My Creator to whom I give all praise.
- The North-West University (NWU) and the Department of Science and Technology (DST) as supporting bodies.
- Necsa for financial aid and the use of their facilities.
- My promoters - Henning Krieg, Sylvia Paul and David Kock - for their guidance, support and contributions.
- Mr M. Lekgoathi (Necsa) for technical assistance.
- My family and friends for continuous support and encouragement.

.

Appendix A. Individual spectra of components

Index

Table A-1: Assignment of symbols to components in the experiments.

Component name	Symbol used in Phase one experiment	Symbol used in Phase two experiment
2,2,2-Trifluoroethanol	D	D
1,1,1-Trifluoro-2-propanol		F
2,2,3,3-Tetrafluoro-1-propanol		E
2,2,3,3,3-Pentafluoro-1-propanol	B	B
2,2,3,4,4,4-Hexafluoro-1-butanol	C'	
2,2,3,3,4,4,5,5-Octafluoro-1-pentanol	A	A
1H,1H,2H,2H-Perfluoro-1-octanol		C

A1. Raman Spectra of pure components

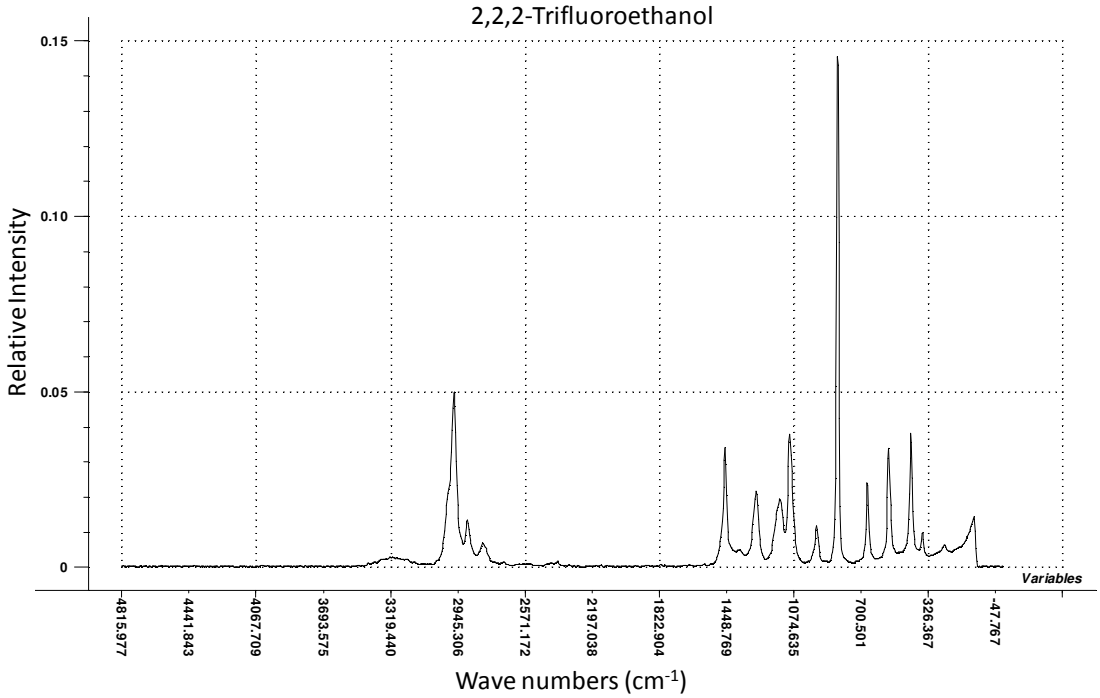


Figure A-1: Raman spectrum of 2,2,2-Trifluoroethanol

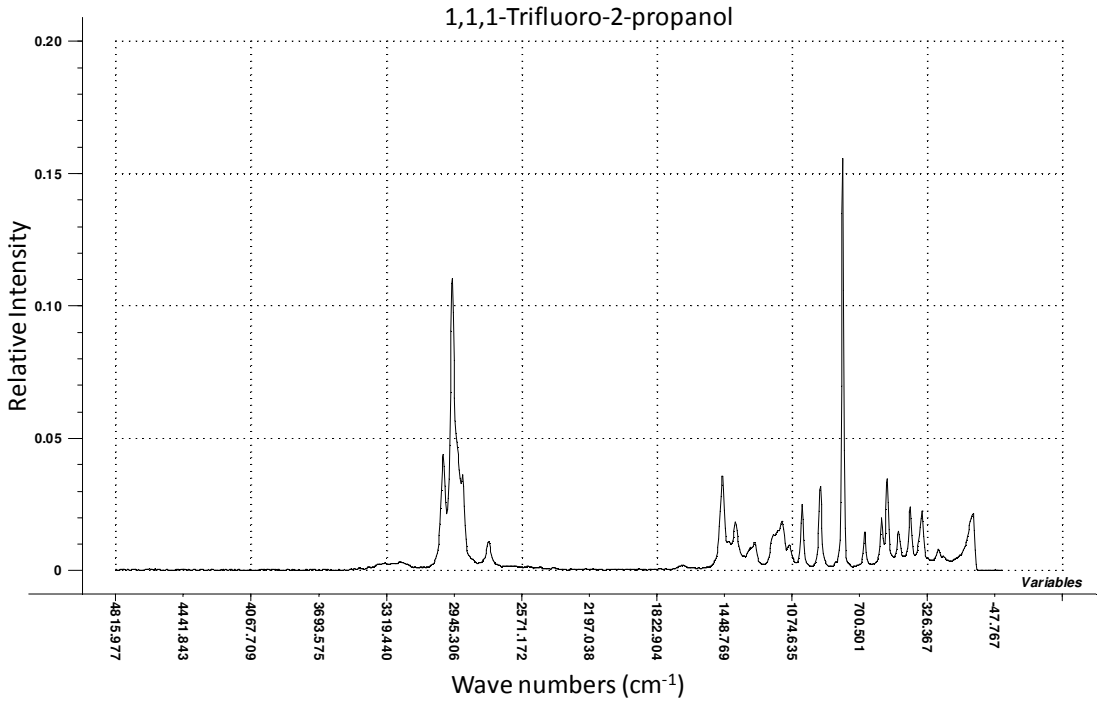


Figure A-2: Raman spectrum of 1,1,1-Trifluoro-2-propanol

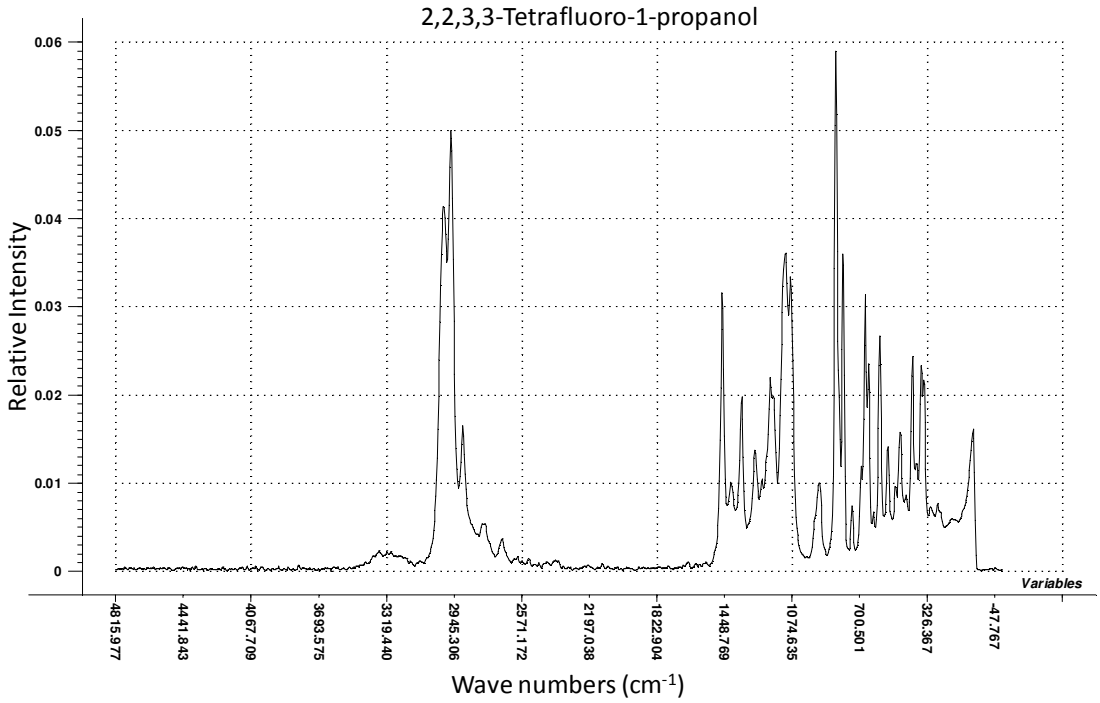


Figure A-3: Raman spectrum of 2,2,3,3-Tetrafluoro-1-propanol

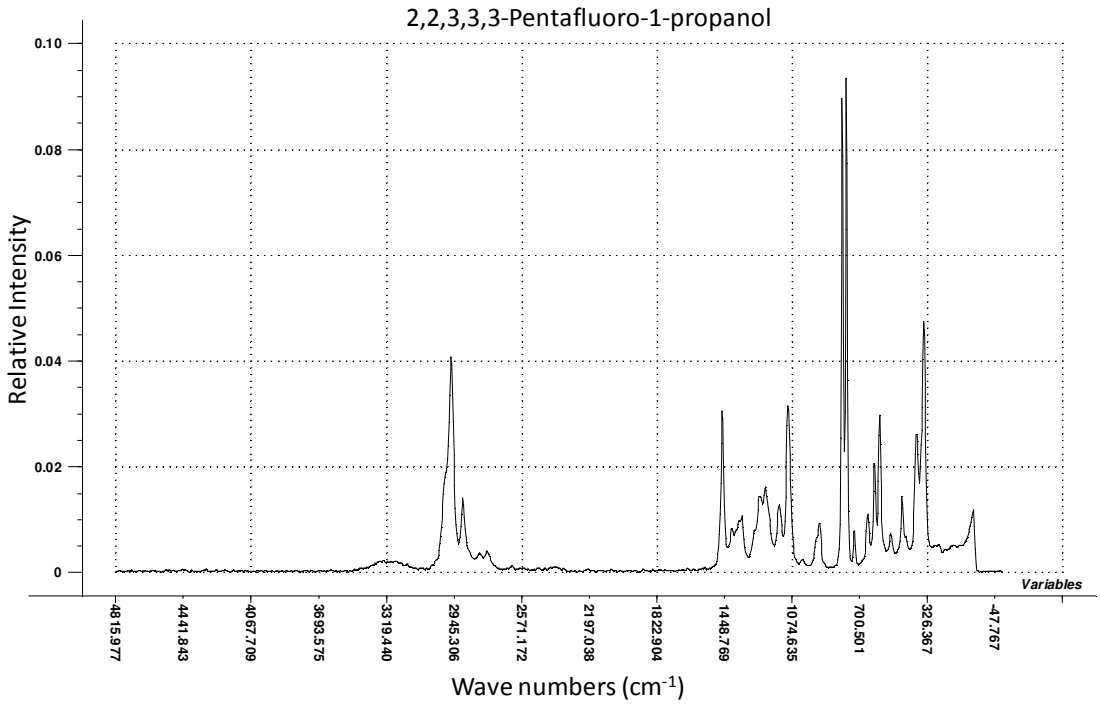


Figure A-4: Raman spectrum of 2,2,3,3,3-Pentafluoro-1-propanol

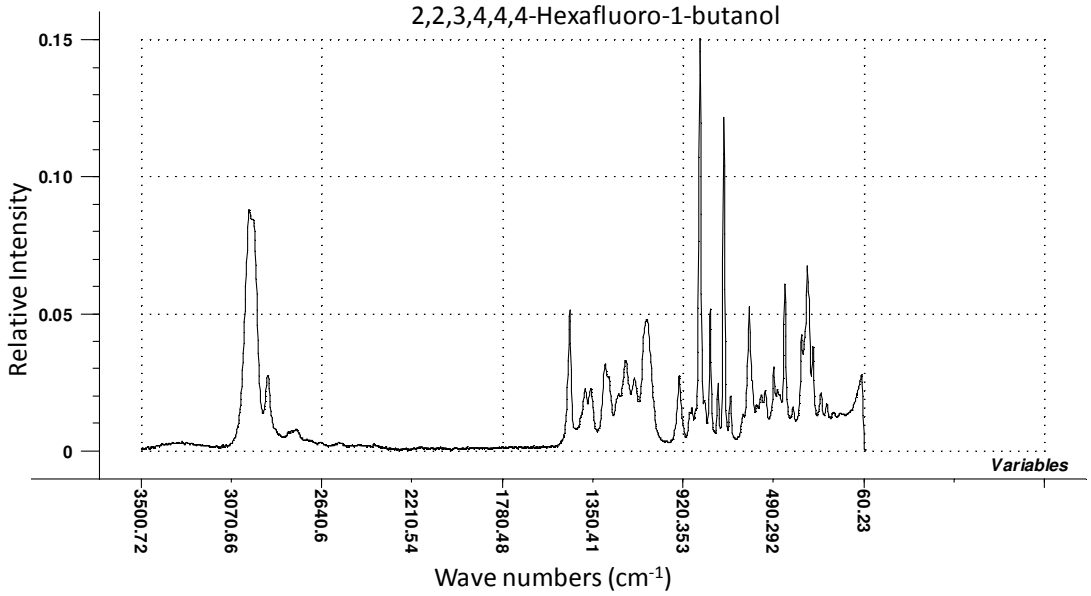


Figure A-5: Raman spectrum of 2,2,3,4,4,4-Hexafluoro-1-butanol

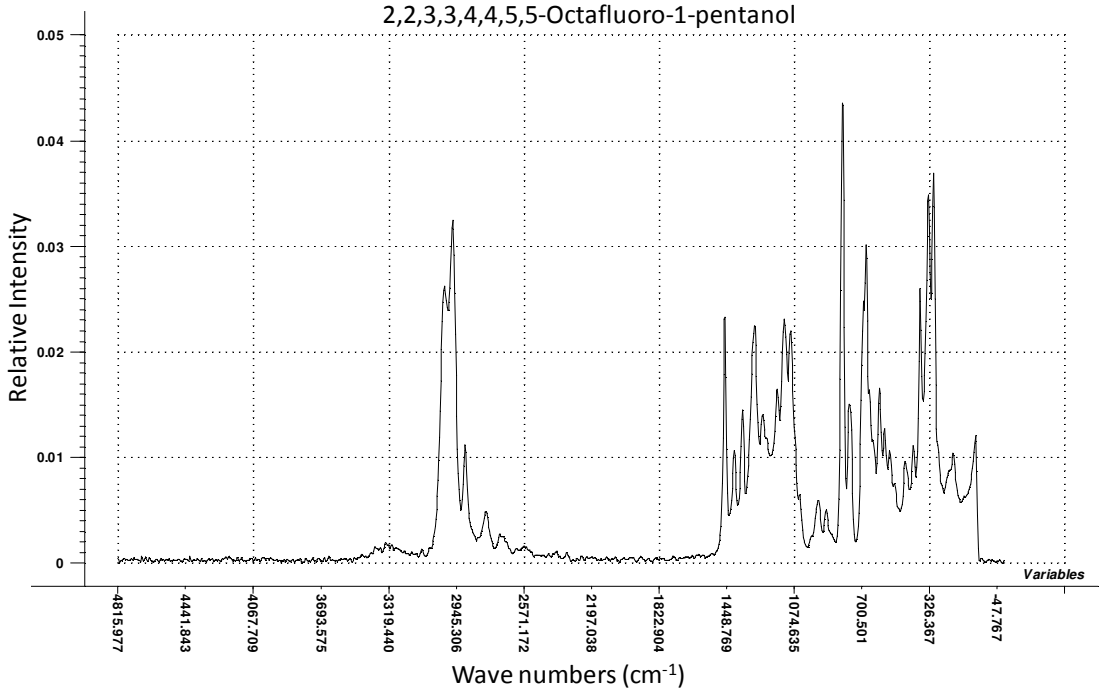


Figure A-6: Raman spectrum of 2,2,3,3,4,4,5,5-Octafluoro-1-pentanol

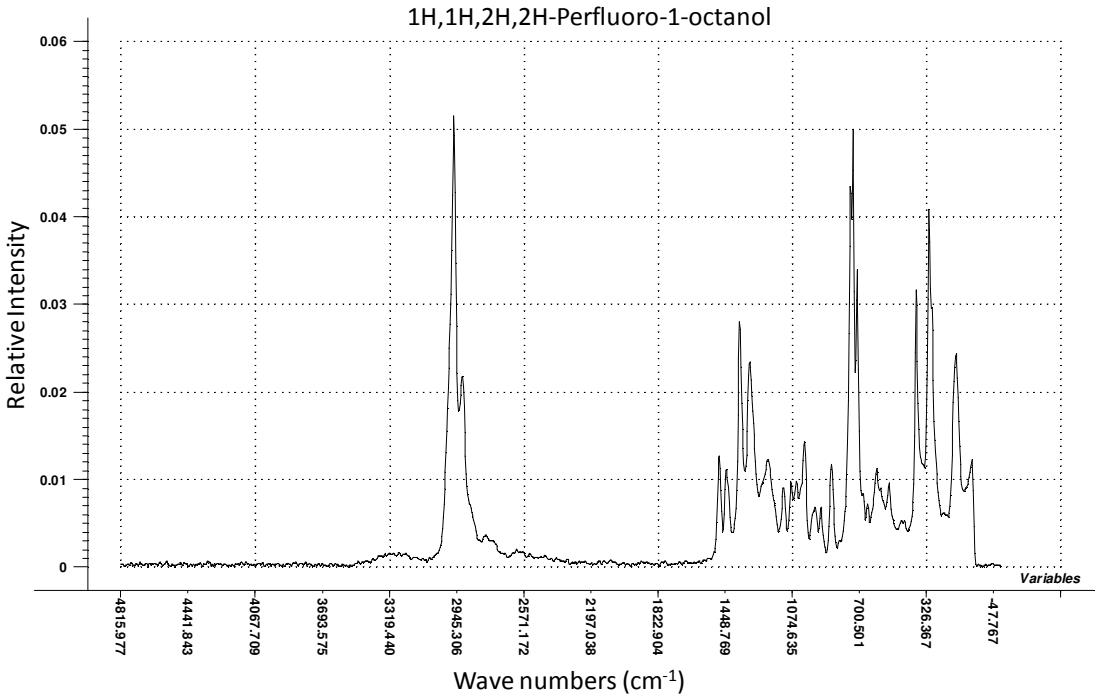


Figure A-7: Raman spectrum of 1H,1H,2H,2H-Perfluoro-1-octanol

A2. NIR Spectra

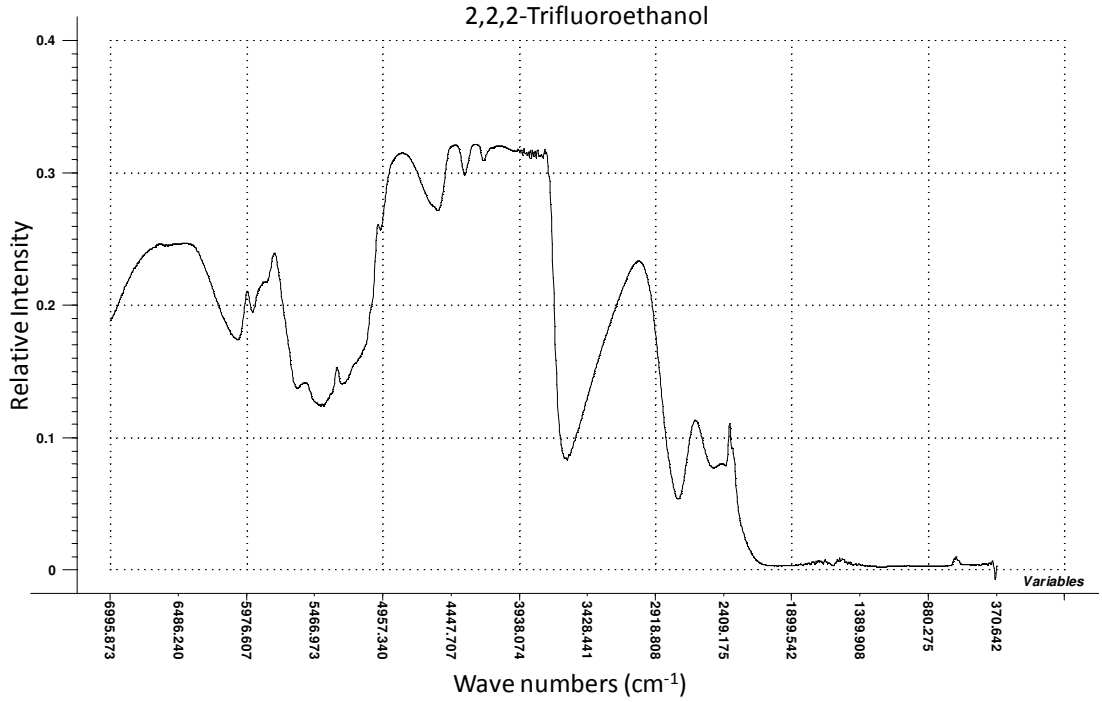


Figure A-8: NIR spectrum of 2,2,2-Trifluoroethanol

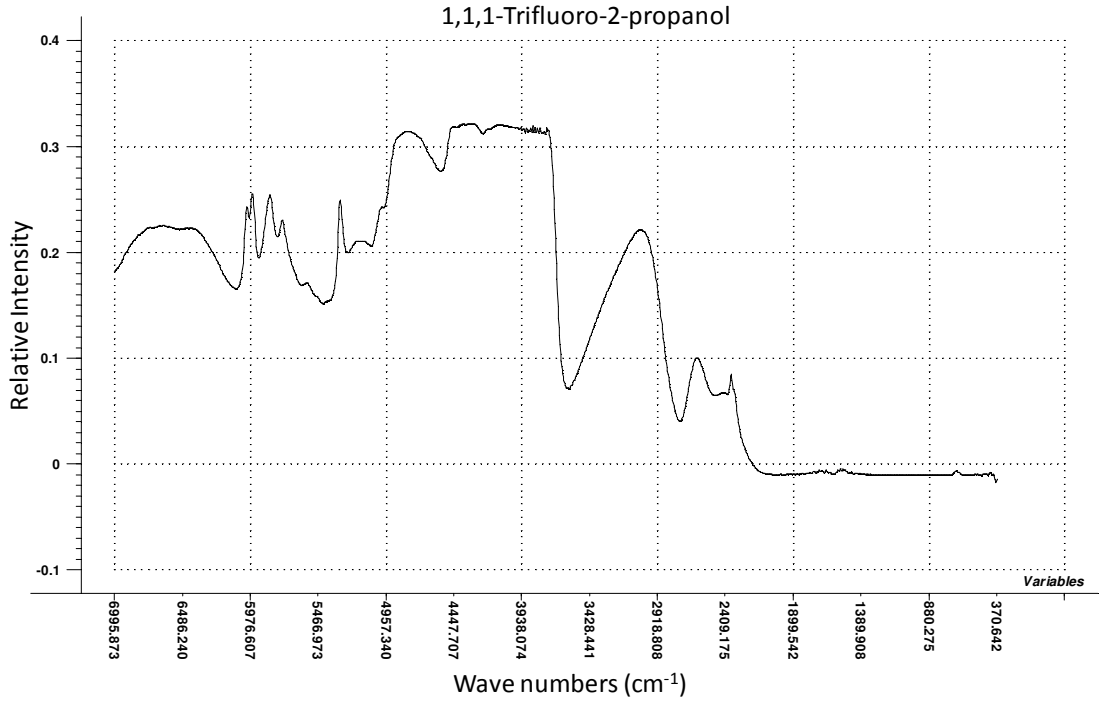


Figure A-9: NIR spectrum of 1,1,1-Trifluoro-2-propanol

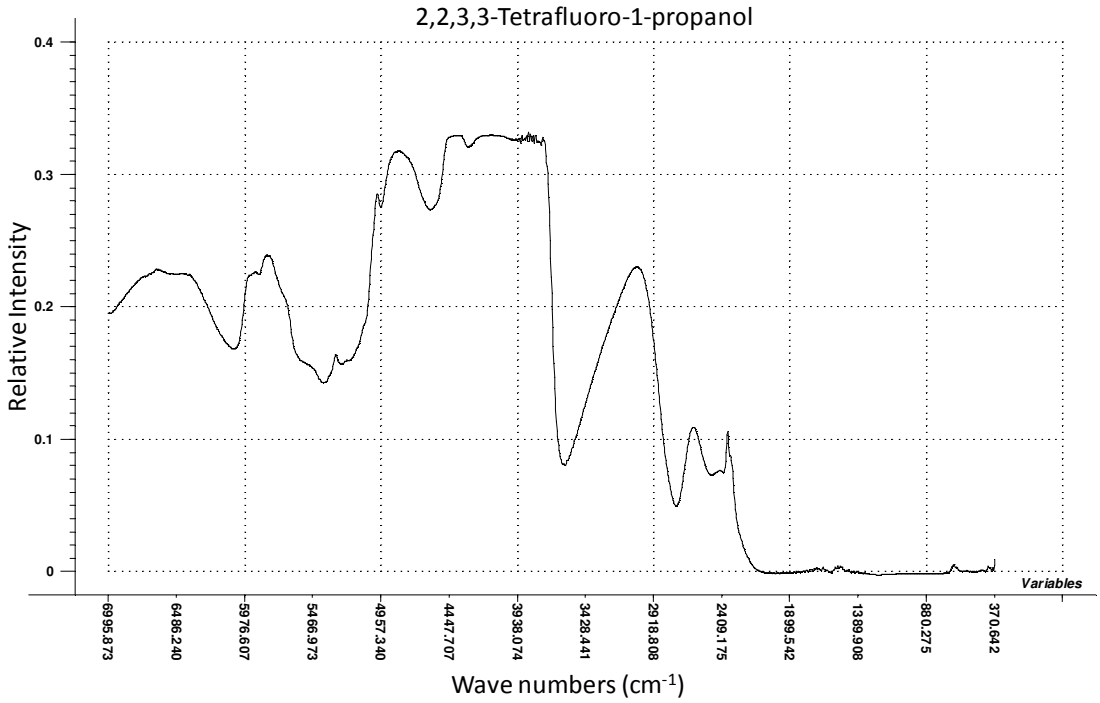


Figure A-10: NIR spectrum of 2,2,3,3-Tetrafluoro-1-propanol

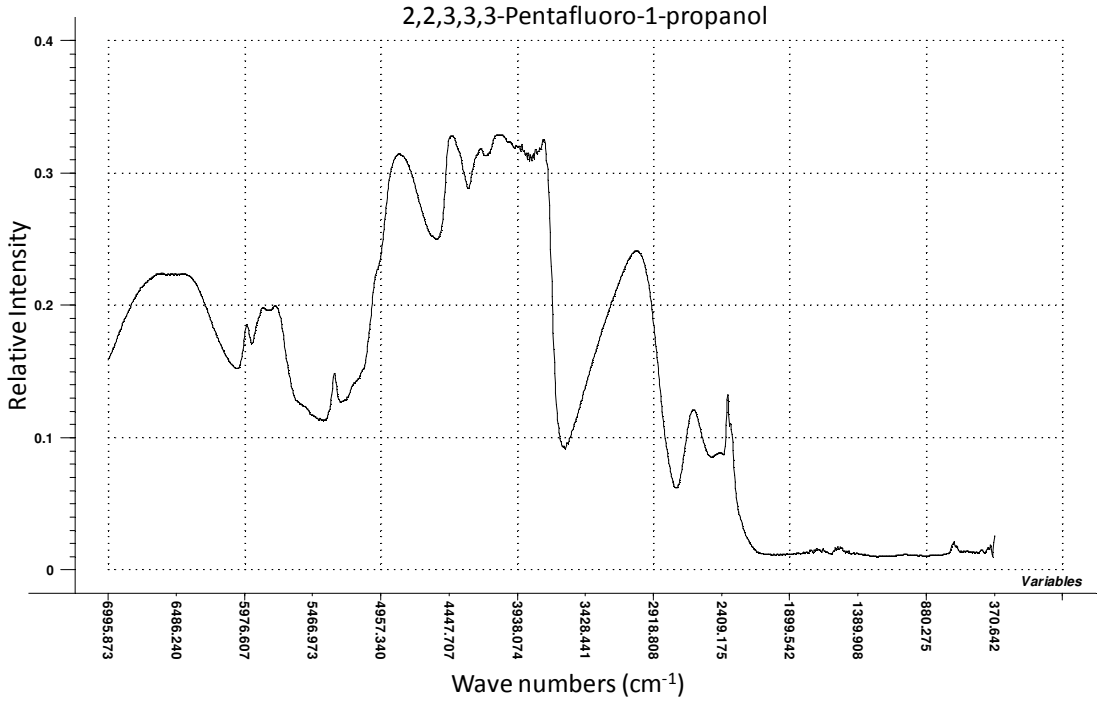


Figure A-11: NIR spectrum of 2,2,3,3,3-Pentafluoro-1-propanol

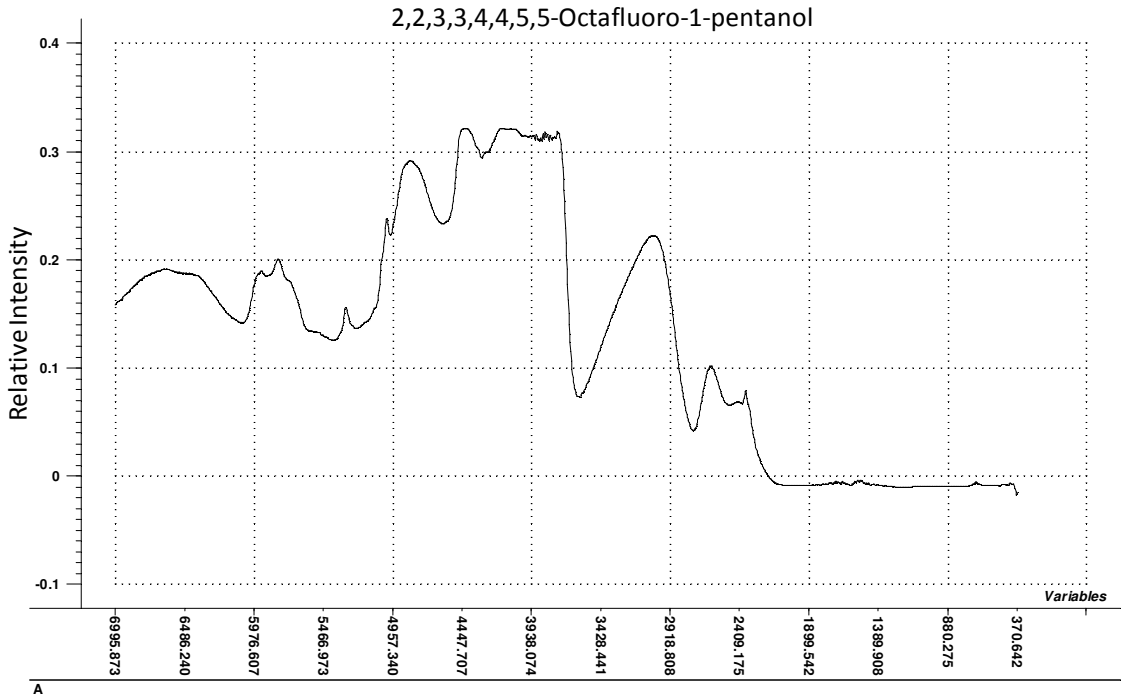


Figure A-12: NIR spectrum of 2,2,3,3,4,4,5,5-Octafluoro-1-pentanol

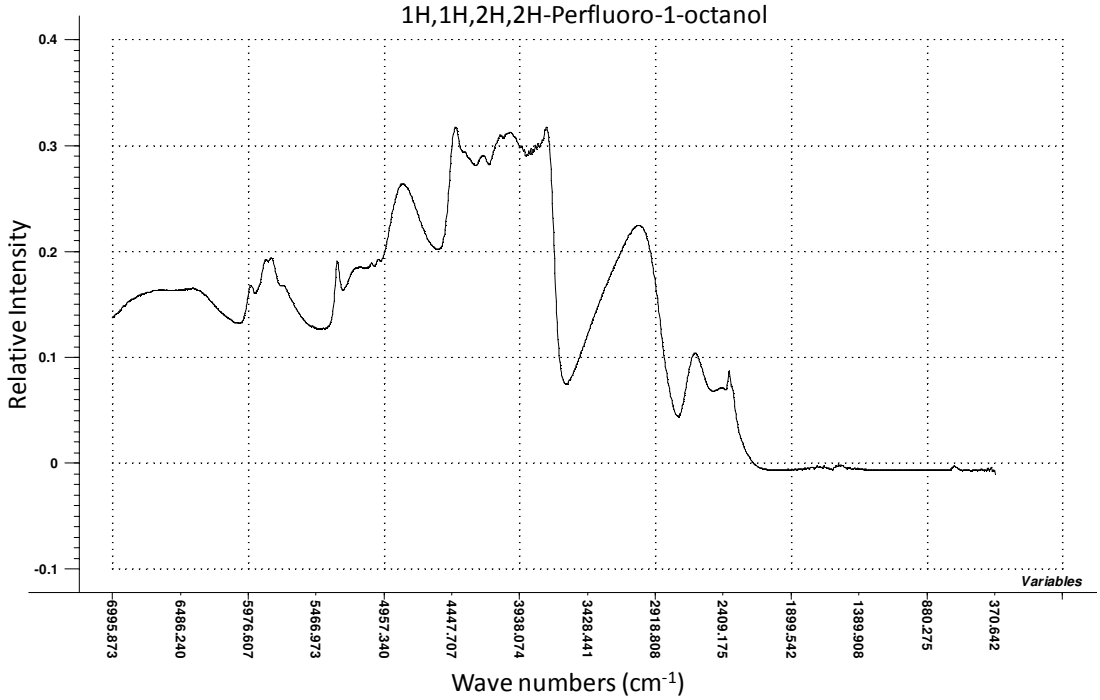


Figure A-13: NIR spectrum of 1H,1H,2H,2H-Perfluoro-1-octanol

A3. ATR-IR Spectra

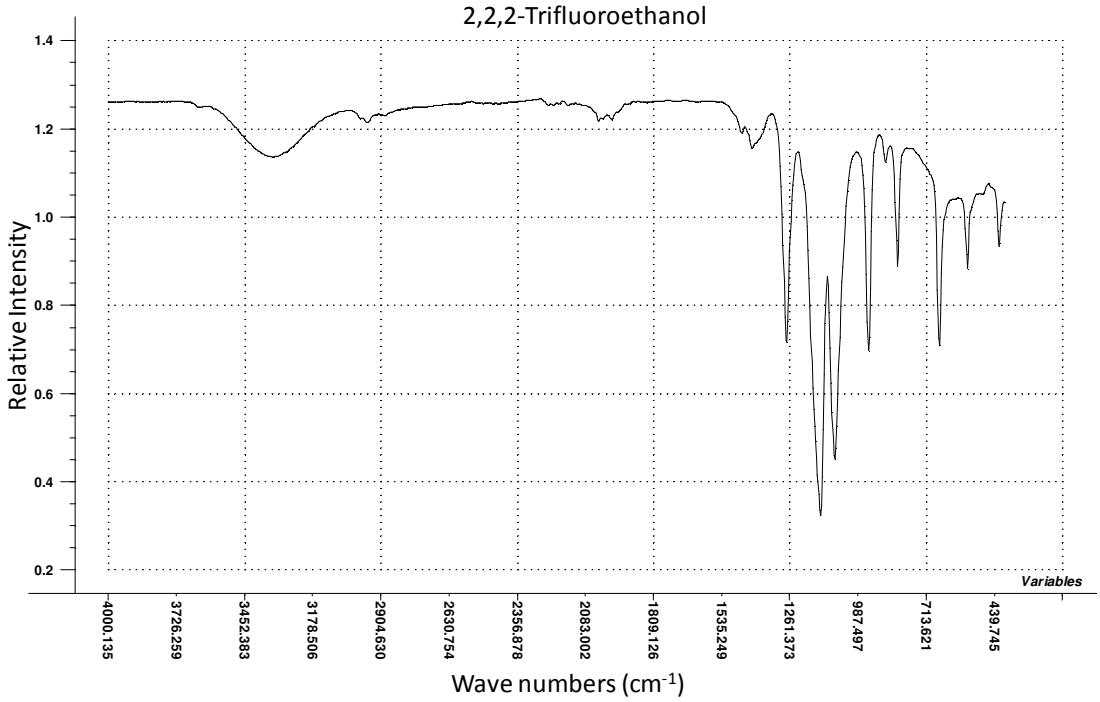


Figure A-14: ATR-IR spectrum of 2,2,2-Trifluoroethanol

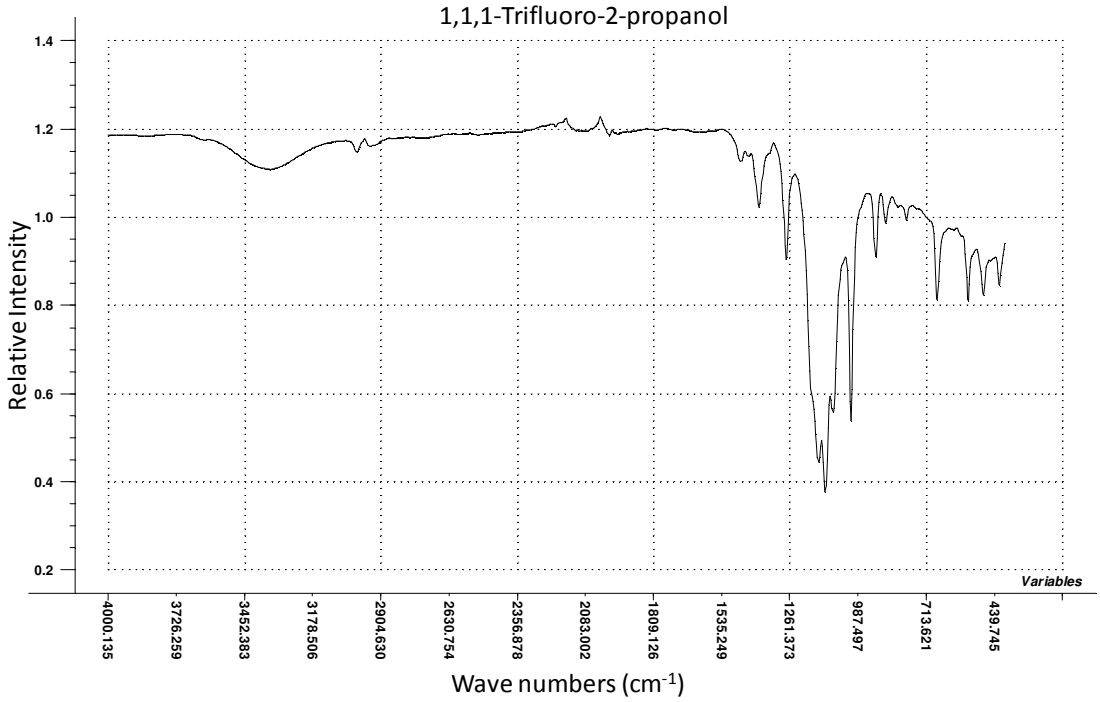


Figure A-15: ATR-IR spectrum 1,1,1-Trifluoro-2-propanol

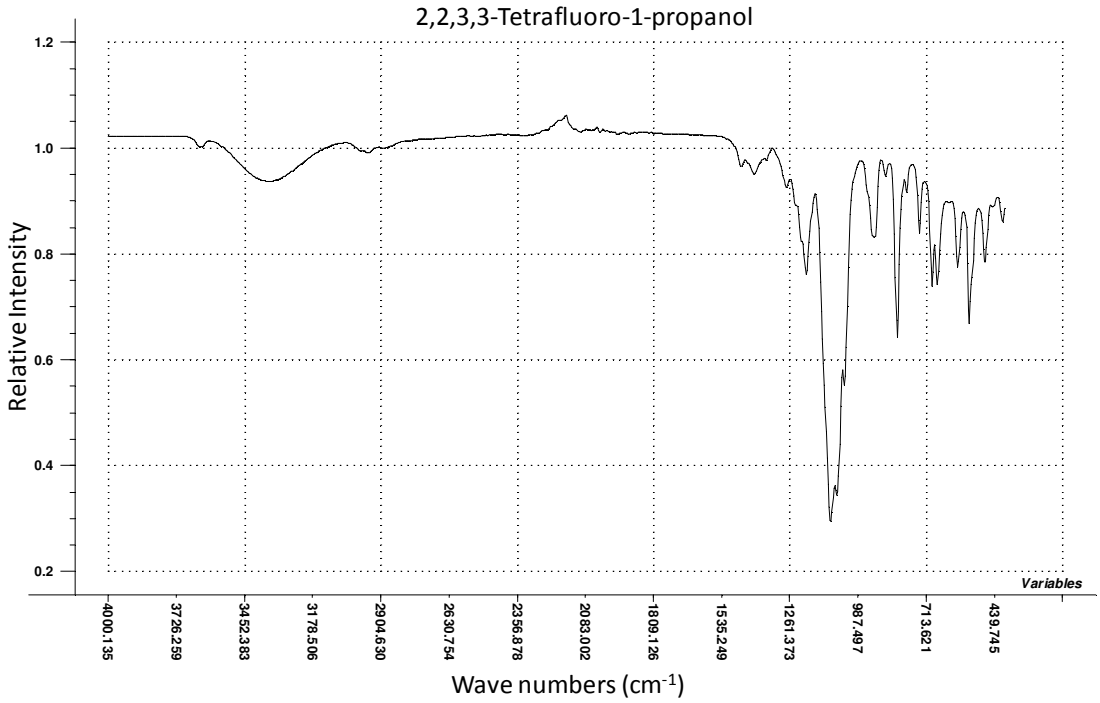


Figure A-16: ATR-IR spectrum 2,2,3,3-Tetrafluoro-1-propanol

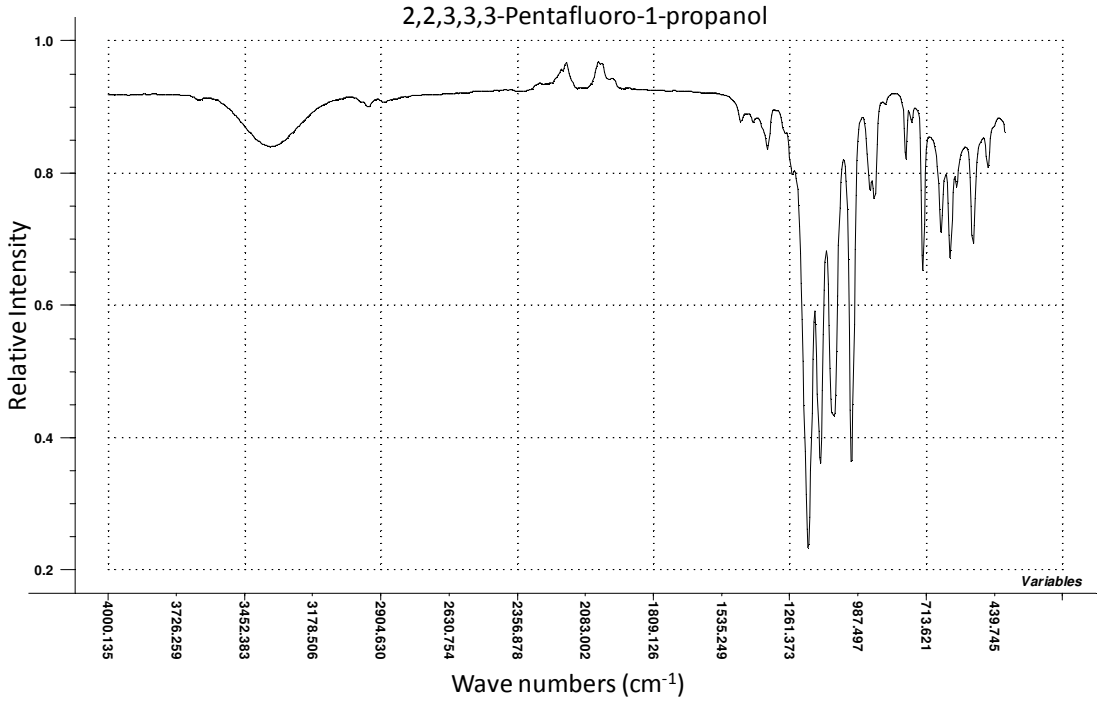


Figure A-17: ATR-IR spectrum 2,2,3,3,3-Pentafluoro-1-propanol

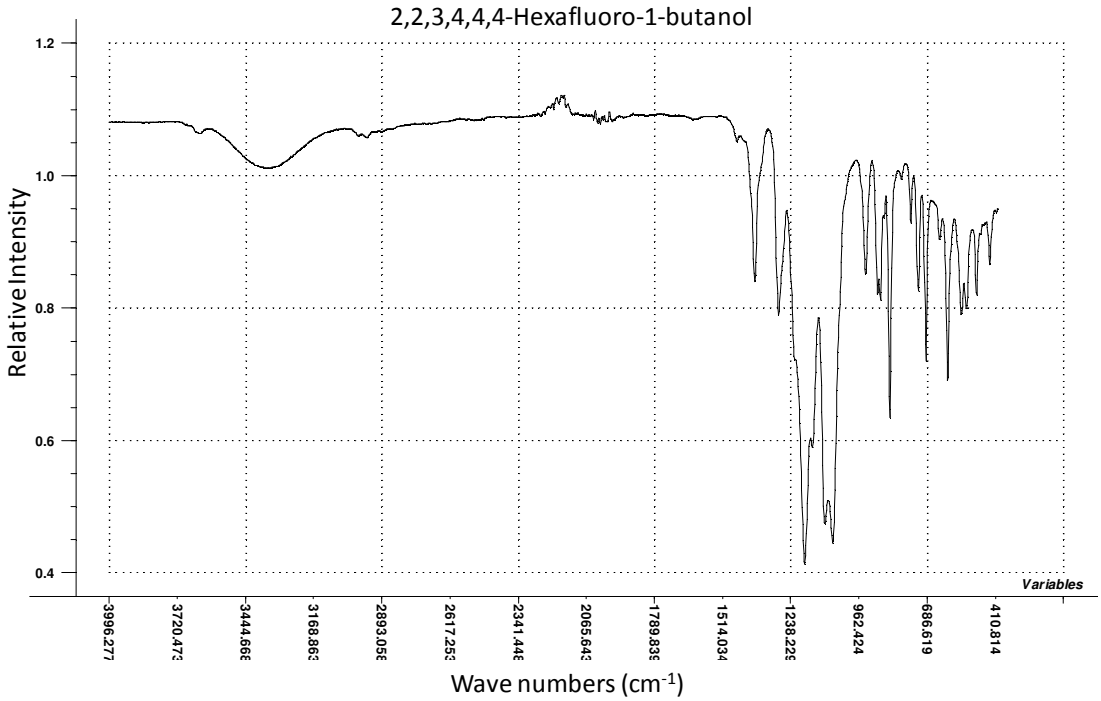


Figure A-18: ATR-IR spectrum 2,2,3,4,4,4-Hexafluoro-1-butanol

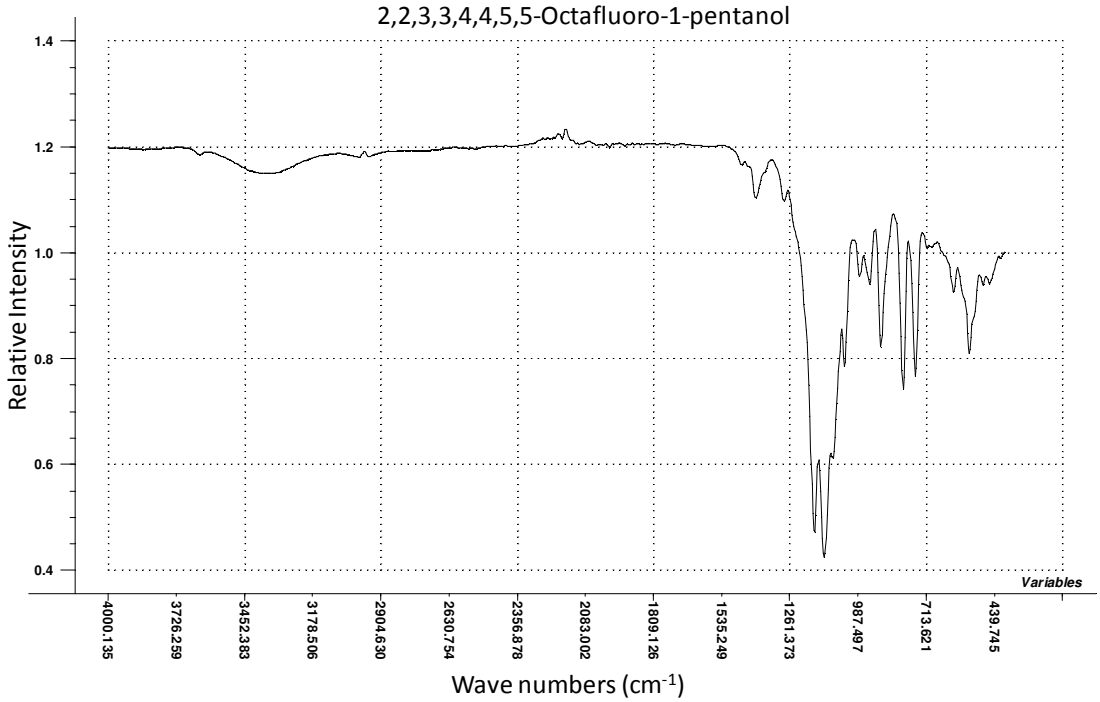


Figure A-19: ATR-IR spectrum 2,2,3,3,4,4,5,5-Octafluoro-1-pentanol

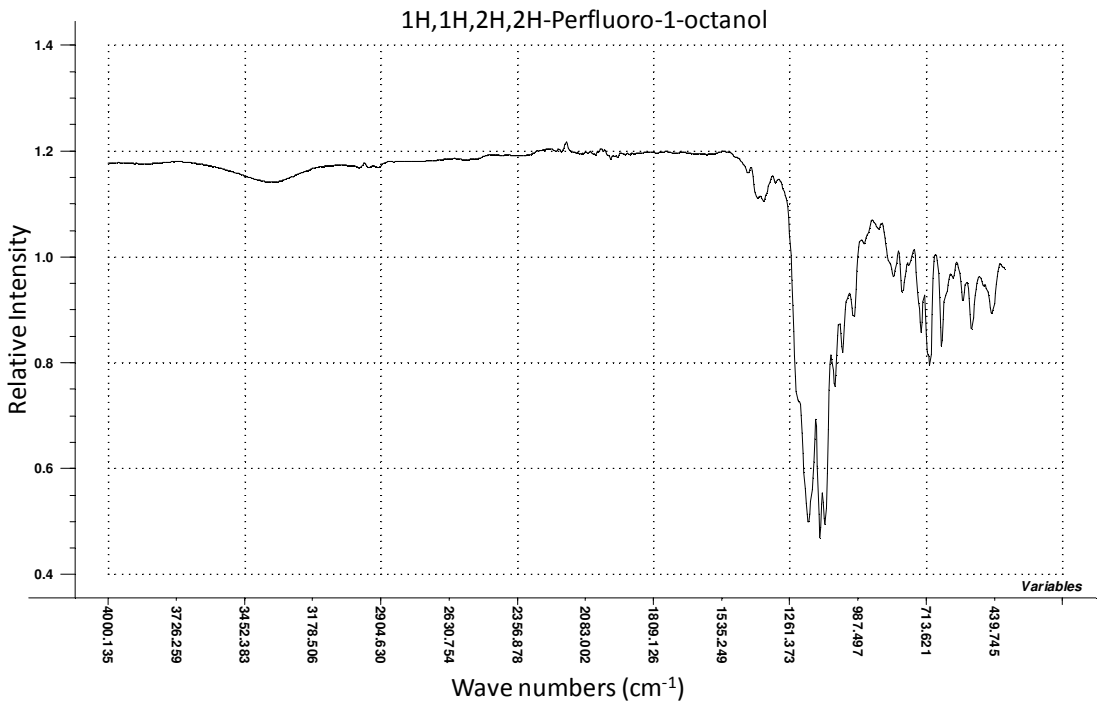


Figure A-20: ATR-IR spectrum 1H,1H,2H,2H-Perfluoro-1-octanol

Appendix B. Analysis results

Index

B1. Analysis of six components (Phase two) experiment

Table B-1: Analysis result table layout.

	A	B	C	D	E	F
R-sqr model	The raw R-square value of the model. Should be between 0.9 and 1 for a good model					
R-sqr adjusted	Tells us how good fit can we expect for future predictions.					
RMSEC	The calibration error					
RMSEP	The expected prediction error					
RMSEP-RMSEC	The smaller this value is, the better the calibration model is for future predictions.					
PC's	The suggested amount of PC's that should be used					

B1.1. Raman

B1.1.1. PLS2 analysis

Table B-2: PLS2 analysis on Raman spectra of the normal mixture design.

	A	B	C	D	E	F
R-sqr model	0.9949	0.9931	0.9969	0.9959	0.9881	0.9522
R-sqr adjusted	0.9941	0.9916	0.9966	0.9954	0.9863	0.9493
RMSEC	0.0050	0.0056	0.0042	0.0039	0.0071	0.0124
RMSEP	0.0053	0.0061	0.0045	0.0042	0.0076	0.0131
RMSEP-RMSEC	0.0004	0.0005	0.0003	0.0003	0.0005	0.0006
PC's	5	5	5	5	5	5

Table B-3: PLS2 analysis on Raman spectra of the design with interaction and square terms added to the design but not used in the regression.

	A	B	C	D	E	F
R-sqr model	0.9958	0.9909	0.9948	0.9944	0.9858	0.9937
R-sqr adjusted	0.9949	0.9889	0.9943	0.9937	0.9845	0.9930
RMSEC	0.0037	0.0053	0.0046	0.0038	0.0064	0.0039
RMSEP	0.0041	0.0058	0.0049	0.0040	0.0067	0.0041
RMSEP-RMSEC	0.0004	0.0006	0.0002	0.0002	0.0003	0.0002
PC's	5	5	5	5	5	5

Table B-4: PLS2 analysis on Raman spectra of the design with interaction and square terms added to the design and the regression calculations.

	A	B	C	D	E	F
R-sqr model	0.9958	0.9909	0.9948	0.9944	0.9858	0.9937
R-sqr adjusted	0.9950	0.9892	0.9943	0.9937	0.9840	0.9930
RMSEC	0.0037	0.0053	0.0046	0.0038	0.0064	0.0039
RMSEP	0.0041	0.0058	0.0049	0.0040	0.0069	0.0042
RMSEP-RMSEC	0.0004	0.0005	0.0002	0.0002	0.0005	0.0002
PC's	5	5	5	5	5	5

B1.1.2. PLS1 analysis

Table B-5: PLS1 analysis on Raman spectra of the normal mixture design.

	A	B	C	D	E	F
R-sqr model	0.9950	0.9915	0.9933	0.9946	0.9905	0.9386
R-sqr adjusted	0.9940	0.9900	0.9927	0.9941	0.9893	0.9335
RMSEC	0.0049	0.0062	0.0062	0.0045	0.0064	0.0141
RMSEP	0.0054	0.0068	0.0066	0.0047	0.0068	0.0148
RMSEP-RMSEC	0.0005	0.0006	0.0004	0.0003	0.0004	0.0007
PC's	5	4	4	4	5	2

Table B-6: PLS1 analysis on Raman spectra of the design with interaction and square terms added to the design but not used in the regression.

	A	B	C	D	E	F
R-sqr model	0.9958	0.9890	0.9911	0.9835	0.9896	0.9867
R-sqr adjusted	0.9950	0.9862	0.9904	0.9825	0.9885	0.9848
RMSEC	0.0037	0.0058	0.0060	0.0065	0.0055	0.0057
RMSEP	0.0041	0.0066	0.0064	0.0068	0.0058	0.0061
RMSEP-RMSEC	0.0004	0.0008	0.0003	0.0003	0.0003	0.0004
PC's	5	4	4	3	5	3

Table B-7: PLS1 analysis on Raman spectra of the design with interaction and square terms added to the design and the regression calculations.

	A	B	C	D	E	F
R-sqr model	0.9958	0.9890	0.9911	0.9835	0.9896	0.9767
R-sqr adjusted	0.9951	0.9865	0.9902	0.9821	0.9887	0.9760
RMSEC	0.0037	0.0058	0.0060	0.0065	0.0055	0.0075
RMSEP	0.0041	0.0065	0.0064	0.0069	0.0057	0.0077
RMSEP-RMSEC	0.0004	0.0006	0.0004	0.0004	0.0003	0.0002
PC's	5	4	4	3	5	2

B1.1.3. PCR analysis

Table B-8: PCR analysis on Raman spectra of the normal mixture design.

	A	B	C	D	E	F
R-sqr model	0.9945	0.9930	0.9970	0.9956	0.9862	0.9523
R-sqr adjusted	0.9936	0.9918	0.9965	0.9952	0.9846	0.9482
RMSEC	0.0051	0.0056	0.0042	0.0040	0.0077	0.0124
RMSEP	0.0055	0.0060	0.0044	0.0042	0.0081	0.0132
RMSEP-RMSEC	0.0003	0.0005	0.0003	0.0002	0.0004	0.0008
PC's	5	5	5	5	5	5

Table B-9: PCR analysis on Raman spectra of the design with interaction and square terms added to the design but not used in the regression.

	A	B	C	D	E	F
R-sqr model	0.9956	0.9909	0.9947	0.9938	0.9830	0.9938
R-sqr adjusted	0.9952	0.9895	0.9941	0.9927	0.9814	0.9934
RMSEC	0.0038	0.0053	0.0047	0.0040	0.0070	0.0039
RMSEP	0.0040	0.0058	0.0050	0.0044	0.0074	0.0040
RMSEP-RMSEC	0.0002	0.0005	0.0003	0.0004	0.0004	0.0002
PC's	5	5	5	5	5	5

Table B-10: PCR analysis on Raman spectra of the design with interaction and square terms added to the design and the regression calculations.

	A	B	C	D	E	F
R-sqr model	0.9956	0.9909	0.9947	0.9938	0.9830	0.9938
R-sqr adjusted	0.9951	0.9893	0.9941	0.9931	0.9811	0.9927
RMSEC	0.0038	0.0053	0.0047	0.0040	0.0070	0.0039
RMSEP	0.0041	0.0058	0.0050	0.0043	0.0075	0.0043
RMSEP-RMSEC	0.0003	0.0005	0.0003	0.0003	0.0005	0.0004
PC's	5	5	5	5	5	5

B1.2. NIR

B1.2.1. PLS2 analysis

Table B-11: PLS2 analysis on NIR spectra of the normal mixture design, spanning entire spectral region.

	A	B	C	D	E	F
R-sqr model	0.9946	0.9267	0.9816	0.9715	0.9714	0.9268
R-sqr adjusted	0.9338	0.9197	0.9787	0.9636	0.9635	0.9168
RMSEC	0.0163	0.0179	0.0103	0.0103	0.0110	0.0154
RMSEP	0.0181	0.0190	0.0112	0.0118	0.0125	0.0165
RMSEP-RMSEC	0.0017	0.0012	0.0009	0.0015	0.0015	0.0011
PC's	9	9	9	9	9	9

Table B-12: PLS2 analysis on NIR spectra of the normal mixture design, using the 6995.873-3945.796cm^{-1} wave number region.

	A	B	C	D	E	F
R-sqr model	0.9484	0.9242	0.9828	0.9700	0.9743	0.9279
R-sqr adjusted	0.9441	0.9121	0.9809	0.9664	0.9717	0.9189
RMSEC	0.0158	0.0182	0.0099	0.0105	0.0105	0.0153
RMSEP	0.0165	0.0197	0.0106	0.0112	0.0112	0.0163
RMSEP-RMSEC	0.0007	0.0015	0.0006	0.0007	0.0008	0.0011
PC's	7	7	7	7	7	7

Table B-13: PLS2 analysis on the first derivative of NIR spectra of the normal mixture design, using the 6076.989-3945.796cm^{-1} wave number region.

	A	B	C	D	E	F
R-sqr model	0.9493	0.9215	0.9759	0.9661	0.9689	0.9424
R-sqr adjusted	0.9397	0.9078	0.9715	0.9623	0.9621	0.9302
RMSEC	0.0156	0.0185	0.0118	0.0112	0.0115	0.0137
RMSEP	0.0172	0.0204	0.0130	0.0119	0.0128	0.0154
RMSEP-RMSEC	0.0015	0.0019	0.0012	0.0007	0.0013	0.0017
PC's	7	7	7	7	7	7

Table B-14: PLS2 analysis on NIR spectra of the normal mixture design after factor normalisations, using the 6995.873-3945.796cm^{-1} wave number region.

	A	B	C	D	E	F
R-sqr model	0.9224	0.9129	0.9638	0.9578	0.9636	0.9121
R-sqr adjusted	0.9108	0.8976	0.9591	0.9467	0.9550	0.8927
RMSEC	0.0193	0.0195	0.0144	0.0125	0.0124	0.0169
RMSEP	0.0210	0.0210	0.0155	0.0141	0.0139	0.0187
RMSEP-RMSEC	0.0016	0.0015	0.0010	0.0016	0.0015	0.0018
PC's	7	7	7	7	7	7

Table B-15: PLS2 analysis on NIR spectra on the design with interaction and square terms added to the design but not used in the regression.

	A	B	C	D	E	F
R-sqr model	0.9489	0.9481	0.9862	0.9724	0.9797	0.9664
R-sqr adjusted	0.9463	0.9300	0.9834	0.9684	0.9738	0.9488
RMSEC	0.0130	0.0127	0.0076	0.0085	0.0076	0.0091
RMSEP	0.0134	0.0148	0.0083	0.0092	0.0087	0.0112
RMSEP-RMSEC	0.0004	0.0021	0.0008	0.0007	0.0011	0.0022
PC's	8	8	8	8	8	8

Table B-16: PLS2 analysis on NIR spectra of the design with interaction and square terms added to the design and the regression calculations.

	A	B	C	D	E	F
R-sqr model	0.9616	0.9557	0.9869	0.9745	0.9803	0.9805
R-sqr adjusted	0.9558	0.9495	0.9842	0.9696	0.9745	0.9755
RMSEC	0.0112	0.0117	0.0073	0.0081	0.0075	0.0069
RMSEP	0.0121	0.0126	0.0081	0.0089	0.0086	0.0078
RMSEP-RMSEC	0.0009	0.0008	0.0008	0.0008	0.0011	0.0009
PC's	9	9	9	9	9	9

B1.2.2. PLS1 analysis

Table B-17: PLS1 analysis on NIR spectra of the normal mixture design.

	A	B	C	D	E	F
R-sqr model	0.9633	0.9308	0.9830	0.9765	0.9769	0.9389
R-sqr adjusted	0.9590	0.9202	0.9804	0.9732	0.9716	0.9299
RMSEC	0.0133	0.0174	0.0099	0.0093	0.0099	0.0141
RMSEP	0.0143	0.0188	0.0106	0.0100	0.0110	0.0152
RMSEP-RMSEC	0.0010	0.0014	0.0007	0.0007	0.0010	0.0011
PC's	7	7	6	7	7	6

Table B-18: PLS1 analysis on NIR spectra on the design with interaction and square terms added to the design but not used in the regression.

	A	B	C	D	E	F
R-sqr model	0.9640	0.9784	0.9860	0.9766	0.9814	0.9806
R-sqr adjusted	0.9523	0.9681	0.9774	0.9715	0.9774	0.9739
RMSEC	0.0109	0.0082	0.0076	0.0078	0.0073	0.0069
RMSEP	0.0125	0.0100	0.0096	0.0086	0.0081	0.0081
RMSEP-RMSEC	0.0016	0.0019	0.0020	0.0008	0.0008	0.0012
PC's	7	9	6	7	7	6

Table B-19: PLS1 analysis on NIR spectra of the design with interaction and square terms added to the design and the regression calculations.

	A	B	C	D	E	F
R-sqr model	0.9845	0.9784	0.9886	0.9766	0.9814	0.9806
R-sqr adjusted	0.9795	0.9709	0.9824	0.9720	0.9754	0.9734
RMSEC	0.0071	0.0082	0.0069	0.0078	0.0073	0.0069
RMSEP	0.0083	0.0095	0.0086	0.0086	0.0085	0.0081
RMSEP-RMSEC	0.0011	0.0013	0.0017	0.0008	0.0011	0.0012
PC's	9	9	7	7	7	6

B1.2.3. PCR analysis

Table B-20: PCR analysis on NIR spectra of the normal mixture design, using the 6995.873-3945.796cm^{-1} wave number region.

	A	B	C	D	E	F
R-sqr model	0.9516	0.9221	0.9826	0.9730	0.9733	0.9336
R-sqr adjusted	0.9459	0.9122	0.9808	0.9705	0.9698	0.9265
RMSEC	0.0153	0.0184	0.0100	0.0100	0.0107	0.0147
RMSEP	0.0163	0.0197	0.0106	0.0107	0.0115	0.0157
RMSEP-RMSEC	0.0011	0.0013	0.0006	0.0007	0.0009	0.0010
PC's	8	8	8	8	8	8

Table B-21: PCR analysis on NIR spectra on the design with interaction and square terms added to the design but not used in the regression.

	A	B	C	D	E	F
R-sqr model	0.9547	0.9505	0.9864	0.9734	0.9794	0.9768
R-sqr adjusted	0.9499	0.9427	0.9841	0.9693	0.9754	0.9697
RMSEC	0.0122	0.0124	0.0075	0.0083	0.0077	0.0075
RMSEP	0.0129	0.0133	0.0082	0.0090	0.0085	0.0087
RMSEP-RMSEC	0.0007	0.0009	0.0007	0.0007	0.0007	0.0012
PC's	9	9	9	9	9	9

Table B-22: PCR analysis on NIR spectra of the design with interaction and square terms added to the design and the regression calculations.

	A	B	C	D	E	F
R-sqr model	0.9547	0.9505	0.9864	0.9734	0.9794	0.9768
R-sqr adjusted	0.9497	0.9437	0.9839	0.9698	0.9748	0.9730
RMSEC	0.0122	0.0124	0.0075	0.0083	0.0077	0.0075
RMSEP	0.0129	0.0133	0.0083	0.0090	0.0086	0.0082
RMSEP-RMSEC	0.0007	0.0009	0.0008	0.0007	0.0009	0.0007
PC's	9	9	9	9	9	9

Table B-23: PCR analysis on the first derivative of NIR spectra of the normal mixture design, using the 6076.989-3945.796cm-1 wave number region.

	A	B	C	D	E	F
R-sqr model	0.9554	0.9108	0.9807	0.9653	0.9711	0.9434
R-sqr adjusted	0.9496	0.9001	0.9783	0.9614	0.9664	0.9389
RMSEC	0.0147	0.0197	0.0105	0.0114	0.0111	0.0136
RMSEP	0.0158	0.0211	0.0113	0.0121	0.0120	0.0143
RMSEP-RMSEC	0.0011	0.0014	0.0007	0.0007	0.0010	0.0008
PC's	8	8	8	8	8	8

Table B-24: PCR analysis on NIR spectra of the normal mixture design after factor normalisations, using the 6995.873-3945.796cm-1 wave number region.

	A	B	C	D	E	F
R-sqr model	0.9052	0.9086	0.9611	0.9621	0.9570	0.9209
R-sqr adjusted	0.8928	0.9003	0.9545	0.9576	0.9521	0.9106
RMSEC	0.0214	0.0199	0.0150	0.0119	0.0135	0.0160
RMSEP	0.0230	0.0211	0.0164	0.0127	0.0145	0.0173
RMSEP-RMSEC	0.0016	0.0012	0.0014	0.0008	0.0010	0.0013
PC's	8	8	8	8	8	8

B1.3. ATR-IR analysis

B1.3.1. PLS2 analysis

Table B-25: PLS2 analysis on ATR-IR spectra of the normal mixture design.

	A	B	C	D	E	F
R-sqr model	0.9867	0.9718	0.9714	0.9847	0.9854	0.9240
R-sqr adjusted	0.9810	0.9630	0.9641	0.9811	0.9790	0.8978
RMSEC	0.0078	0.0111	0.0129	0.0074	0.0077	0.0157
RMSEP	0.0097	0.0128	0.0147	0.0083	0.0094	0.0189
RMSEP-RMSEC	0.0018	0.0017	0.0017	0.0009	0.0017	0.0032
PC's	7	7	7	7	7	7

Table B-26: PLS2 analysis on ATR-IR spectra on the design with interaction and square terms added to the design but not used in the regression.

	A	B	C	D	E	F
R-sqr model	0.9847	0.9660	0.9742	0.9831	0.9836	0.9788
R-sqr adjusted	0.9762	0.9515	0.9660	0.9770	0.9766	0.9704
RMSEC	0.0070	0.0102	0.0105	0.0066	0.0068	0.0073
RMSEP	0.0087	0.0125	0.0122	0.0077	0.0083	0.0087
RMSEP-RMSEC	0.0017	0.0023	0.0017	0.0011	0.0015	0.0014
PC's	7	7	7	7	7	7

Table B-27: PLS2 analysis on ATR-IR spectra of the design with interaction and square terms added to the design and the regression calculations.

	A	B	C	D	E	F
R-sqr model	0.9847	0.9660	0.9742	0.9831	0.9836	0.9788
R-sqr adjusted	0.9753	0.9496	0.9660	0.9783	0.9754	0.9688
RMSEC	0.0070	0.0120	0.0105	0.0066	0.0068	0.0073
RMSEP	0.0089	0.0127	0.0121	0.0076	0.0084	0.0089
RMSEP-RMSEC	0.0019	0.0006	0.0016	0.0011	0.0016	0.0017
PC's	7	7	7	7	7	7

B1.3.2. PLS1 analysis

Table B-28: PLS1 analysis on ATR-IR spectra of the normal mixture design.

	A	B	C	D	E	F
R-sqr model	0.9869	0.9695	0.9704	0.9837	0.9840	0.9253
R-sqr adjusted	0.9811	0.9574	0.9540	0.9792	0.9774	0.8976
RMSEC	0.0078	0.0115	0.0131	0.0077	0.0081	0.0156
RMSEP	0.0095	0.0138	0.0164	0.0088	0.0095	0.0188
RMSEP-RMSEC	0.0017	0.0022	0.0032	0.0011	0.0014	0.0032
PC's	5	4	5	4	3	5

Table B-29: PLS1 analysis on ATR-IR spectra on the design with interaction and square terms added to the design but not used in the regression.

	A	B	C	D	E	F
R-sqr model	0.9866	0.9643	0.9761	0.9828	0.9784	0.9779
R-sqr adjusted	0.9758	0.9440	0.9664	0.9756	0.9614	0.9709
RMSEC	0.0065	0.0105	0.0101	0.0066	0.0078	0.0074
RMSEP	0.0090	0.0133	0.0122	0.0079	0.0104	0.0087
RMSEP-RMSEC	0.0025	0.0029	0.0021	0.0012	0.0026	0.0013
PC's	5	4	5	4	3	5

Table B-30: PLS1 analysis on ATR-IR spectra of the design with interaction and square terms added to the design and the regression calculations.

	A	B	C	D	E	F
R-sqr model	0.9866	0.9643	0.9761	0.9828	0.9784	0.9779
R-sqr adjusted	0.9769	0.9394	0.9691	0.9779	0.9699	0.9676
RMSEC	0.0065	0.0105	0.0101	0.0066	0.0078	0.0074
RMSEP	0.0087	0.0137	0.0116	0.0078	0.0092	0.0091
RMSEP-RMSEC	0.0022	0.0033	0.0015	0.0012	0.0015	0.0016
PC's	5	4	5	4	3	5

B1.3.3. PCR analysis

Table B-31: PCR analysis on ATR-IR spectra of the normal mixture design.

	A	B	C	D	E	F
R-sqr model	0.9852	0.9714	0.9705	0.9843	0.9852	0.9218
R-sqr adjusted	0.9785	0.9624	0.9622	0.9800	0.9793	0.8956
RMSEC	0.0083	0.0112	0.0131	0.0075	0.0078	0.0159
RMSEP	0.0099	0.0132	0.0152	0.0086	0.0093	0.0194
RMSEP-RMSEC	0.0017	0.0020	0.0021	0.0011	0.0016	0.0034
PC's	7	7	7	7	7	7

Table B-32: PCR analysis on ATR-IR spectra on the design with interaction and square terms added to the design but not used in the regression.

	A	B	C	D	E	F
R-sqr model	0.9805	0.9655	0.9728	0.9826	0.9833	0.9755
R-sqr adjusted	0.9718	0.9489	0.9587	0.9779	0.9763	0.9661
RMSEC	0.0079	0.0103	0.0108	0.0067	0.0068	0.0078
RMSEP	0.0097	0.0125	0.0135	0.0077	0.0083	0.0094
RMSEP-RMSEC	0.0018	0.0022	0.0027	0.0010	0.0014	0.0016
PC's	7	7	7	7	7	7

Table B-33: PCR analysis on ATR-IR spectra of the design with interaction and square terms added to the design and the regression calculations.

	A	B	C	D	E	F
R-sqr model	0.9805	0.9655	0.9728	0.9826	0.9833	0.9755
R-sqr adjusted	0.9718	0.9504	0.9653	0.9762	0.9756	0.9644
RMSEC	0.0079	0.0103	0.0108	0.0067	0.0068	0.0078
RMSEP	0.0096	0.0124	0.0121	0.0078	0.0084	0.0095
RMSEP-RMSEC	0.0017	0.0021	0.0013	0.0012	0.0015	0.0017
PC's	7	7	7	7	7	7

Appendix C. Regression en prediction results

Index

C1. Tables with R^2 values and error terms of regressions

All regression models are done on the design with interaction and square terms added to the regression calculation.

Table C-1: Layout and meaning of values in regression result tables

Regression model used

Components

Spectroscopic technique

		Ramam PLS2	A	B	C	D	E	F
Values obtained using the external calibration and test set samples	External	**R-sqr model**	The raw R-square value of the model. Should be between 0.9 and 1 for a good model					
		R-sqr adjusted	Tells us how good fit can we expect for future predictions.					
		RMSEC	The calibration error					
		RMSEP	The expected prediction error					
		RMSEP-RMSEC	The smaller this value is, the better the calibration model is for future predictions.					
Values obtained using the internal calibration and test set samples	Internal	**R-sqr model**						
		R-sqr adjusted						
		RMSEC						
		RMSEP						
		RMSEP-RMSEC						
		PC's	The suggested amount of PC's that should be used					

Table C-2: Regression values of PLS2 regression model done on Raman spectra

	Ramam PLS2	A	B	C	D	E	F
External	R-sqr model	0.9972	0.9951	0.9949	0.9956	0.9887	0.9939
	R-sqr adjusted	0.9970	0.9944	0.9945	0.9950	0.9872	0.9930
	RMSEC	0.0031	0.0036	0.0042	0.0034	0.0057	0.0037
	RMSEP	0.0032	0.0039	0.0044	0.0036	0.0061	0.0039
	RMSEP-RMSEC	0.0001	0.0003	0.0002	0.0002	0.0004	0.0003
Internal	R-sqr model	0.9972	0.9951	0.9949	0.9955	0.9887	0.9939
	R-sqr adjusted	0.9970	0.9945	0.9941	0.9949	0.9875	0.9927
	RMSEC	0.0031	0.0037	0.0042	0.0034	0.0057	0.0037
	RMSEP	0.0033	0.0039	0.0045	0.0037	0.0061	0.0040
	RMSEP-RMSEC	0.0002	0.0002	0.0003	0.0002	0.0003	0.0003
	PC's	5	5	5	5	5	5

Table C-3: Regression values of PCR regression model done on Raman spectra

	Ramam PCR	A	B	C	D	E	F
External	R-sqr model	0.9972	0.9951	0.9948	0.9954	0.9874	0.9939
	R-sqr adjusted	0.9970	0.9946	0.9944	0.9949	0.9866	0.9932
	RMSEC	0.0031	0.0036	0.0042	0.0035	0.0060	0.0037
	RMSEP	0.0032	0.0038	0.0045	0.0037	0.0063	0.0039
	RMSEP-RMSEC	0.0001	0.0002	0.0002	0.0002	0.0002	0.0002
Internal	R-sqr model	0.9972	0.9950	0.9948	0.9953	0.9873	0.9939
	R-sqr adjusted	0.9970	0.9944	0.9945	0.9947	0.9860	0.9931
	RMSEC	0.0031	0.0037	0.0043	0.0035	0.0061	0.0037
	RMSEP	0.0033	0.0039	0.0045	0.0038	0.0065	0.0040
	RMSEP-RMSEC	0.0001	0.0003	0.0002	0.0002	0.0004	0.0003
	PC's	5	5	5	5	5	5

Table C-4: Regression values of PLS2 regression model done on NIR spectra

	NIR PLS2	A	B	C	D	E	F
External	R-sqr model	0.9643	0.9537	0.9877	0.9800	0.9846	0.9678
	R-sqr adjusted	0.9616	0.9418	0.9832	0.9769	0.9796	0.9620
	RMSEC	0.0108	0.0117	0.0063	0.0071	0.0069	0.0082
	RMSEP	0.0113	0.0131	0.0073	0.0077	0.0080	0.0091
	RMSEP-RMSEC	0.0004	0.0015	0.0011	0.0006	0.0011	0.0009
Internal	R-sqr model	0.9647	0.9550	0.9881	0.9801	0.9860	0.9688
	R-sqr adjusted	0.9585	0.9383	0.9816	0.9773	0.9811	0.9603
	RMSEC	0.0108	0.0116	0.0062	0.0072	0.0067	0.0081
	RMSEP	0.0119	0.0137	0.0077	0.0077	0.0078	0.0093
	RMSEP-RMSEC	0.0011	0.0021	0.0015	0.0005	0.0011	0.0012
	PC's	8	8	8	8	8	8

Table C-5: Regression values of PCR regression model done on NIR spectra

	NIR PCR	A	B	C	D	E	F
External	R-sqr model	0.9699	0.9605	0.9877	0.9797	0.9846	0.9770
	R-sqr adjusted	0.9641	0.9442	0.9831	0.9750	0.9751	0.9702
	RMSEC	0.0099	0.0108	0.0063	0.0072	0.0069	0.0070
	RMSEP	0.0109	0.0130	0.0074	0.0081	0.0090	0.0080
	RMSEP-RMSEC	0.0009	0.0022	0.0011	0.0009	0.0020	0.0010
Internal	R-sqr model	0.9697	0.9604	0.9883	0.9798	0.9859	0.9767
	R-sqr adjusted	0.9617	0.9451	0.9831	0.9751	0.9807	0.9703
	RMSEC	0.0101	0.0108	0.0062	0.0072	0.0067	0.0070
	RMSEP	0.0114	0.0128	0.0074	0.0081	0.0079	0.0080
	RMSEP-RMSEC	0.0013	0.0020	0.0013	0.0009	0.0012	0.0010
	PC's	9	9	9	9	9	9

Table C-6: Regression values of PLS2 regression model done on ATR-IR spectra

	ATR-IR PLS2	A	B	C	D	E	F
External	R-sqr model	0.9887	0.9667	0.9773	0.9815	0.9830	0.9772
	R-sqr adjusted	0.9824	0.9475	0.9605	0.9724	0.9753	0.9635
	RMSEC	0.0061	0.0101	0.0088	0.0069	0.0071	0.0070
	RMSEP	0.0077	0.0128	0.0118	0.0084	0.0086	0.0091
	RMSEP-RMSEC	0.0016	0.0027	0.0030	0.0016	0.0015	0.0021
Internal	R-sqr model	0.9883	0.9684	0.9798	0.9816	0.9825	0.9786
	R-sqr adjusted	0.9813	0.9468	0.9669	0.9736	0.9711	0.9650
	RMSEC	0.0063	0.0100	0.0084	0.0069	0.0073	0.0069
	RMSEP	0.0079	0.0130	0.0109	0.0084	0.0094	0.0090
	RMSEP-RMSEC	0.0016	0.0031	0.0025	0.0015	0.0022	0.0020
	PC's	7	7	7	7	7	7

Table C-7: Regression values of PCR regression model done on ATR-IR spectra

	ATR-IR PCR	A	B	C	D	E	F
External	R-sqr model	0.9859	0.9658	0.9760	0.9807	0.9827	0.9740
	R-sqr adjusted	0.9785	0.9422	0.9589	0.9733	0.9730	0.9632
	RMSEC	0.0068	0.0102	0.0091	0.0070	0.0071	0.0075
	RMSEP	0.0084	0.0132	0.0118	0.0083	0.0090	0.0091
	RMSEP-RMSEC	0.0016	0.0030	0.0027	0.0012	0.0019	0.0016
Internal	R-sqr model	0.9841	0.9682	0.9782	0.9810	0.9822	0.9743
	R-sqr adjusted	0.9707	0.9462	0.9613	0.9711	0.9686	0.9524
	RMSEC	0.0073	0.0100	0.0088	0.0070	0.0073	0.0076
	RMSEP	0.0101	0.0130	0.0119	0.0088	0.0099	0.0105
	RMSEP-RMSEC	0.0027	0.0030	0.0031	0.0017	0.0026	0.0029
	PC's	7	7	7	7	7	7

C2. Tables with prediction results

For tables C-8 to C-13, PV is the value predicted by the regression model, MV is the value in grams measured, SDevM is the standard deviation of prediction of the model used and PV-MV is the difference between the predicted value and the real measured values. T1-T5 AVE are the external test samples averaged over three measurements and S1-S5 are the internal test samples.

Table C-8: Prediction results done with a PCR model of the Raman data

Raman PCR		T1 Ave	T2 Ave	T3 Ave	T4 Ave	T5 Ave	S1	S2	S3	S4	S5
A	PV	0.0228	0.0123	0.0893	0.0013	0.0025	0.0022	0.1010	0.0601	0.0030	0.0971
	MV	0.0250	0.0080	0.0920	0.0000	0.0000	0.0000	0.0970	0.0630	0.0000	0.0970
	SDevM	0.0031	0.0032	0.0030	0.0030	0.0038	0.0021	0.0026	0.0017	0.0023	0.0016
	PV-MV	-0.0022	0.0043	-0.0027	0.0013	0.0025	0.0022	0.0040	-0.0029	0.0030	0.0001
B	PV	0.0004	0.0188	0.1413	0.0031	0.0865	0.0031	-0.0039	0.0539	0.0958	0.0983
	MV	0.0000	0.0150	0.1440	0.0000	0.0870	0.0000	0.0000	0.0550	0.0970	0.0970
	SDevM	0.0037	0.0038	0.0036	0.0036	0.0046	0.0026	0.0031	0.0021	0.0027	0.0019
	PV-MV	0.0004	0.0038	-0.0027	0.0031	-0.0005	0.0031	-0.0039	-0.0011	-0.0012	0.0013
C	PV	0.2667	0.0122	0.0016	0.0745	0.0027	0.0035	0.1030	0.0607	0.0046	-0.0014
	MV	0.2660	0.0080	0.0000	0.0720	0.0000	0.0000	0.0980	0.0600	0.0000	0.0000
	SDevM	0.0043	0.0044	0.0041	0.0042	0.0053	0.0029	0.0036	0.0024	0.0031	0.0022
	PV-MV	0.0007	0.0042	0.0016	0.0025	0.0027	0.0035	0.0050	0.0007	0.0046	-0.0014
D	PV	0.0229	0.1320	0.0017	0.0685	0.0021	0.0906	0.0000	-0.0004	0.0869	-0.0016
	MV	0.0200	0.1340	0.0000	0.0660	0.0000	0.0890	0.0000	0.0000	0.0860	0.0000
	SDevM	0.0035	0.0036	0.0034	0.0034	0.0043	0.0025	0.0030	0.0020	0.0026	0.0018
	PV-MV	0.0029	-0.0020	0.0017	0.0025	0.0021	0.0016	0.0000	-0.0004	0.0009	-0.0016
E	PV	-0.0065	0.0503	0.0572	0.0246	-0.0045	0.0874	-0.0014	0.0588	-0.0004	0.0989
	MV	0.0000	0.0560	0.0600	0.0290	0.0000	0.0920	0.0000	0.0590	0.0000	0.0950
	SDevM	0.0060	0.0062	0.0058	0.0059	0.0074	0.0042	0.0051	0.0034	0.0045	0.0032
	PV-MV	-0.0065	-0.0057	-0.0028	-0.0044	-0.0045	-0.0046	-0.0014	-0.0002	-0.0004	0.0039
F	PV	-0.0026	0.0441	-0.0004	0.0961	0.1657	0.0784	0.0843	0.0506	0.0759	-0.0003
	MV	0.0000	0.0480	0.0000	0.0970	0.1750	0.0810	0.0820	0.0470	0.0790	0.0000
	SDevM	0.0037	0.0039	0.0036	0.0037	0.0047	0.0026	0.0031	0.0021	0.0027	0.0019
	PV-MV	-0.0026	-0.0039	-0.0004	-0.0009	-0.0093	-0.0026	0.0023	0.0036	-0.0031	-0.0003

Table C-9: Prediction results done with a PLS2 model of the Raman data

Raman PLS2		T1 Ave	T2 Ave	T3 Ave	T4 Ave	T5 Ave	S1	S2	S3	S4	S5
A	PV	0.0227	0.0122	0.0892	0.0012	0.0024	0.0022	0.1010	0.0601	0.0030	0.0972
	MV	0.0250	0.0080	0.0920	0.0000	0.0000	0.0000	0.0970	0.0630	0.0000	0.0970
	SDevM	0.0031	0.0032	0.0030	0.0030	0.0038	0.0021	0.0026	0.0017	0.0023	0.0016
	PV-MV	-0.0023	0.0042	-0.0028	0.0012	0.0024	0.0022	0.0040	-0.0029	0.0030	0.0001
B	PV	0.0004	0.0188	0.1410	0.0030	0.0865	0.0030	-0.0039	0.0539	0.0958	0.0983
	MV	0.0000	0.0150	0.1440	0.0000	0.0870	0.0000	0.0000	0.0550	0.0970	0.0970
	SDevM	0.0037	0.0038	0.0036	0.0036	0.0046	0.0025	0.0031	0.0021	0.0027	0.0019
	PV-MV	0.0004	0.0038	-0.0030	0.0030	-0.0005	0.0030	-0.0039	-0.0011	-0.0012	0.0013
C	PV	0.2667	0.0122	0.0016	0.0745	0.0027	0.0035	0.1030	0.0607	0.0045	-0.0014
	MV	0.2660	0.0080	0.0000	0.0720	0.0000	0.0000	0.0980	0.0600	0.0000	0.0000
	SDevM	0.0042	0.0044	0.0041	0.0041	0.0052	0.0030	0.0036	0.0024	0.0032	0.0022
	PV-MV	0.0007	0.0042	0.0016	0.0025	0.0027	0.0035	0.0050	0.0007	0.0045	-0.0014
D	PV	0.0228	0.1320	0.0016	0.0684	0.0019	0.0905	0.0000	-0.0005	0.0868	-0.0016
	MV	0.0200	0.1340	0.0000	0.0660	0.0000	0.0890	0.0000	0.0000	0.0860	0.0000
	SDevM	0.0034	0.0036	0.0033	0.0034	0.0043	0.0024	0.0029	0.0019	0.0026	0.0018
	PV-MV	0.0028	-0.0020	0.0016	0.0024	0.0019	0.0015	0.0000	-0.0005	0.0008	-0.0016
E	PV	-0.0064	0.0507	0.0575	0.0250	-0.0041	0.0876	-0.0013	0.0589	-0.0002	0.0989
	MV	0.0000	0.0560	0.0600	0.0290	0.0000	0.0920	0.0000	0.0590	0.0000	0.0950
	SDevM	0.0057	0.0060	0.0056	0.0056	0.0071	0.0040	0.0049	0.0032	0.0042	0.0030
	PV-MV	-0.0064	-0.0053	-0.0025	-0.0040	-0.0041	-0.0044	-0.0013	-0.0001	-0.0002	0.0039
F	PV	-0.0026	0.0440	-0.0004	0.0961	0.1657	0.0784	0.0842	0.0506	0.0759	-0.0003
	MV	0.0000	0.0480	0.0000	0.0970	0.1750	0.0810	0.0820	0.0470	0.0790	0.0000
	SDevM	0.0037	0.0039	0.0036	0.0037	0.0047	0.0026	0.0032	0.0021	0.0028	0.0020
	PV-MV	-0.0026	-0.0040	-0.0004	-0.0010	-0.0093	-0.0026	0.0022	0.0036	-0.0032	-0.0003

Table C-10: Prediction results done with a PCR model of the NIR data

NIR PCR		T1 Ave	T2 Ave	T3 Ave	T4 Ave	T5 Ave	S1	S2	S3	S4	S5
A	PV	0.0288	0.0090	0.0969	0.0031	0.0152	-0.0028	0.0888	0.0720	0.0091	0.0915
	MV	0.0250	0.0080	0.0920	0.0000	0.0000	0.0000	0.0970	0.0630	0.0000	0.0970
	SDevM	0.0101	0.0080	0.0086	0.0124	0.0108	0.0084	0.0093	0.0078	0.0089	0.0055
	PV-MV	0.0038	0.0010	0.0049	0.0031	0.0152	-0.0028	-0.0082	0.0090	0.0091	-0.0055
B	PV	-0.0137	0.0228	0.1353	-0.0008	0.0597	0.0139	-0.0009	0.0554	0.0832	0.0938
	MV	0.0000	0.0150	0.1440	0.0000	0.0870	0.0000	0.0000	0.0550	0.0970	0.0970
	SDevM	0.0119	0.0095	0.0102	0.0146	0.0128	0.0095	0.0105	0.0088	0.0101	0.0063
	PV-MV	-0.0137	0.0078	-0.0087	-0.0008	-0.0273	0.0139	-0.0009	0.0004	-0.0139	-0.0032
C	PV	0.2823	0.0109	0.0005	0.0760	0.0036	0.0089	0.0856	0.0492	0.0050	-0.0015
	MV	0.2660	0.0080	0.0000	0.0720	0.0000	0.0000	0.0980	0.0600	0.0000	0.0000
	SDevM	0.0068	0.0054	0.0058	0.0084	0.0073	0.0055	0.0061	0.0051	0.0058	0.0036
	PV-MV	0.0163	0.0029	0.0005	0.0040	0.0036	0.0089	-0.0124	-0.0108	0.0050	-0.0015
D	PV	0.0255	0.1247	-0.0054	0.0580	0.0023	0.0838	0.0101	-0.0058	0.0963	0.0005
	MV	0.0200	0.1340	0.0000	0.0660	0.0000	0.0890	0.0000	0.0000	0.0860	0.0000
	SDevM	0.0075	0.0060	0.0064	0.0092	0.0080	0.0059	0.0066	0.0055	0.0063	0.0039
	PV-MV	0.0055	-0.0093	-0.0054	-0.0080	0.0023	-0.0052	0.0101	-0.0058	0.0103	0.0005
E	PV	-0.0070	0.0561	0.0681	0.0320	0.0032	0.0873	0.0198	0.0614	-0.0076	0.1110
	MV	0.0000	0.0560	0.0600	0.0290	0.0000	0.0920	0.0000	0.0590	0.0000	0.0950
	SDevM	0.0082	0.0065	0.0070	0.0101	0.0088	0.0058	0.0064	0.0054	0.0062	0.0038
	PV-MV	-0.0070	0.0000	0.0081	0.0030	0.0032	-0.0047	0.0198	0.0024	-0.0076	0.0160
F	PV	-0.0060	0.0463	-0.0012	0.0990	0.1713	0.0752	0.0809	0.0505	0.0795	-0.0017
	MV	0.0000	0.0480	0.0000	0.0970	0.1750	0.0810	0.0820	0.0470	0.0790	0.0000
	SDevM	0.0074	0.0059	0.0063	0.0091	0.0079	0.0059	0.0065	0.0055	0.0063	0.0039
	PV-MV	-0.0060	-0.0017	-0.0012	0.0020	-0.0037	-0.0058	-0.0011	0.0035	0.0005	-0.0017

Table C-11: Prediction results done with a PLS2 model of the NIR data

NIR PLS2		T1 Ave	T2 Ave	T3 Ave	T4 Ave	T5 Ave	S1	S2	S3	S4	S5
A	PV	0.0179	0.0081	0.0948	0.0025	0.0095	-0.0031	0.0895	0.0729	0.0074	0.0881
	MV	0.0250	0.0080	0.0920	0.0000	0.0000	0.0000	0.0970	0.0630	0.0000	0.0970
	SDevM	0.0178	0.0064	0.0078	0.0098	0.0119	0.0065	0.0075	0.0063	0.0072	0.0075
	PV-MV	-0.0071	0.0001	0.0028	0.0025	0.0095	-0.0031	-0.0075	0.0099	0.0074	-0.0089
B	PV	-0.0029	0.0235	0.1383	0.0009	0.0658	0.0134	-0.0009	0.0546	0.0849	0.0980
	MV	0.0000	0.0150	0.1440	0.0000	0.0870	0.0000	0.0000	0.0550	0.0970	0.0970
	SDevM	0.0207	0.0074	0.0091	0.0114	0.0139	0.0075	0.0086	0.0073	0.0083	0.0086
	PV-MV	-0.0029	0.0085	-0.0057	0.0009	-0.0212	0.0134	-0.0009	-0.0004	-0.0121	0.0010
C	PV	0.2857	0.0112	0.0010	0.0758	0.0052	0.0094	0.0849	0.0489	0.0054	-0.0005
	MV	0.2660	0.0080	0.0000	0.0720	0.0000	0.0000	0.0980	0.0600	0.0000	0.0000
	SDevM	0.0116	0.0041	0.0051	0.0064	0.0078	0.0042	0.0049	0.0041	0.0047	0.0048
	PV-MV	0.0197	0.0032	0.0010	0.0038	0.0052	0.0094	-0.0131	-0.0111	0.0054	-0.0005
D	PV	0.0279	0.1247	-0.0049	0.0583	0.0034	0.0844	0.0098	-0.0064	0.0965	0.0013
	MV	0.0200	0.1340	0.0000	0.0660	0.0000	0.0890	0.0000	0.0000	0.0860	0.0000
	SDevM	0.0121	0.0043	0.0053	0.0067	0.0082	0.0042	0.0049	0.0041	0.0047	0.0048
	PV-MV	0.0079	-0.0093	-0.0049	-0.0077	0.0034	-0.0046	0.0098	-0.0064	0.0105	0.0013
E	PV	-0.0062	0.0562	0.0682	0.0320	0.0036	0.0870	0.0201	0.0617	-0.0073	0.1110
	MV	0.0000	0.0560	0.0600	0.0290	0.0000	0.0920	0.0000	0.0590	0.0000	0.0950
	SDevM	0.0126	0.0045	0.0056	0.0070	0.0085	0.0043	0.0049	0.0041	0.0047	0.0049
	PV-MV	-0.0062	0.0002	0.0082	0.0030	0.0036	-0.0050	0.0201	0.0027	-0.0073	0.0160
F	PV	-0.0158	0.0453	-0.0039	0.0972	0.1663	0.0750	0.0813	0.0514	0.0781	-0.0056
	MV	0.0000	0.0480	0.0000	0.0970	0.1750	0.0810	0.0820	0.0470	0.0790	0.0000
	SDevM	0.0141	0.0051	0.0062	0.0078	0.0095	0.0051	0.0059	0.0049	0.0056	0.0058
	PV-MV	-0.0158	-0.0027	-0.0039	0.0002	-0.0087	-0.0060	-0.0008	0.0044	-0.0009	-0.0056

Table C-12: Prediction results done with a PCR model of the ATR-IR data

ATR-IR PCR		T1 Ave	T2 Ave	T3 Ave	T4 Ave	T5 Ave	S1	S2	S3	S4	S5
A	PV	0.0299	0.0055	0.0877	-0.0039	0.0019	-0.0031	0.1060	0.0643	-0.0002	0.0982
	MV	0.0250	0.0080	0.0920	0.0000	0.0000	0.0000	0.0970	0.0630	0.0000	0.0970
	SDevM	0.0068	0.0055	0.0051	0.0047	0.0083	0.0072	0.0110	0.0077	0.0063	0.0066
	PV-MV	0.0049	-0.0025	-0.0043	-0.0039	0.0019	-0.0031	0.0090	0.0013	-0.0002	0.0012
B	PV	0.0182	0.0098	0.1380	-0.0042	0.0824	-0.0051	-0.0141	0.0607	0.1050	0.1080
	MV	0.0000	0.0150	0.1440	0.0000	0.0870	0.0000	0.0000	0.0550	0.0970	0.0970
	SDevM	0.0110	0.0089	0.0082	0.0077	0.0134	0.0092	0.0141	0.0099	0.0081	0.0085
	PV-MV	0.0182	-0.0052	-0.0060	-0.0042	-0.0046	-0.0051	-0.0141	0.0057	0.0080	0.0110
C	PV	0.2230	0.0028	0.0017	0.0796	0.0087	-0.0164	0.1200	0.0599	0.0058	-0.0021
	MV	0.2660	0.0080	0.0000	0.0720	0.0000	0.0000	0.0980	0.0600	0.0000	0.0000
	SDevM	0.0098	0.0079	0.0073	0.0068	0.0119	0.0085	0.0129	0.0091	0.0074	0.0077
	PV-MV	-0.0430	-0.0052	0.0017	0.0076	0.0087	-0.0164	0.0220	-0.0001	0.0058	-0.0021
D	PV	0.0276	0.1275	-0.0008	0.0603	0.0018	0.0937	-0.0014	-0.0135	0.0881	-0.0078
	MV	0.0200	0.1340	0.0000	0.0660	0.0000	0.0890	0.0000	0.0000	0.0860	0.0000
	SDevM	0.0066	0.0054	0.0049	0.0046	0.0081	0.0062	0.0096	0.0067	0.0055	0.0057
	PV-MV	0.0076	-0.0065	-0.0008	-0.0057	0.0018	0.0047	-0.0014	-0.0135	0.0021	-0.0078
E	PV	0.0012	0.0595	0.0578	0.0322	-0.0025	0.0923	-0.0056	0.0574	-0.0014	0.0968
	MV	0.0000	0.0560	0.0600	0.0290	0.0000	0.0920	0.0000	0.0590	0.0000	0.0950
	SDevM	0.0072	0.0058	0.0053	0.0050	0.0087	0.0070	0.0106	0.0075	0.0061	0.0064
	PV-MV	0.0012	0.0035	-0.0022	0.0032	-0.0025	0.0003	-0.0056	-0.0016	-0.0014	0.0018
F	PV	-0.0027	0.0604	0.0036	0.1014	0.1625	0.0982	0.0811	0.0566	0.0682	-0.0014
	MV	0.0000	0.0480	0.0000	0.0970	0.1750	0.0810	0.0820	0.0470	0.0790	0.0000
	SDevM	0.0072	0.0059	0.0054	0.0050	0.0088	0.0074	0.0114	0.0080	0.0065	0.0068
	PV-MV	-0.0027	0.0124	0.0036	0.0044	-0.0125	0.0172	-0.0009	0.0096	-0.0108	-0.0014

Table C-13: Prediction results done with a PLS2 model of the ATR-IR data

ATR-IR PLS2		T1 Ave	T2 Ave	T3 Ave	T4 Ave	T5 Ave	S1	S2	S3	S4	S5
A	PV	0.0288	0.0062	0.0883	-0.0032	0.0013	-0.0032	0.1040	0.0667	0.0001	0.0994
	MV	0.0250	0.0080	0.0920	0.0000	0.0000	0.0000	0.0970	0.0630	0.0000	0.0970
	SDevM	0.0062	0.0051	0.0046	0.0044	0.0076	0.0057	0.0087	0.0059	0.0050	0.0051
	PV-MV	0.0038	-0.0018	-0.0037	-0.0032	0.0013	-0.0032	0.0070	0.0037	0.0001	0.0024
B	PV	0.0178	0.0097	0.1380	-0.0038	0.0824	-0.0050	-0.0146	0.0609	0.1050	0.1080
	MV	0.0000	0.0150	0.1440	0.0000	0.0870	0.0000	0.0000	0.0550	0.0970	0.0970
	SDevM	0.0105	0.0086	0.0078	0.0075	0.0129	0.0095	0.0147	0.0099	0.0083	0.0085
	PV-MV	0.0178	-0.0053	-0.0060	-0.0038	-0.0046	-0.0050	-0.0146	0.0059	0.0080	0.0110
C	PV	0.2235	0.0030	0.0017	0.0793	0.0088	-0.0162	0.1210	0.0589	0.0056	-0.0024
	MV	0.2660	0.0080	0.0000	0.0720	0.0000	0.0000	0.0980	0.0600	0.0000	0.0000
	SDevM	0.0097	0.0079	0.0072	0.0069	0.0119	0.0079	0.0122	0.0082	0.0069	0.0070
	PV-MV	-0.0425	-0.0050	0.0017	0.0073	0.0088	-0.0162	0.0230	-0.0011	0.0056	-0.0024
D	PV	0.0276	0.1280	-0.0006	0.0602	0.0016	0.0939	-0.0017	-0.0130	0.0880	-0.0074
	MV	0.0200	0.1340	0.0000	0.0660	0.0000	0.0890	0.0000	0.0000	0.0860	0.0000
	SDevM	0.0068	0.0056	0.0051	0.0048	0.0084	0.0061	0.0094	0.0064	0.0054	0.0055
	PV-MV	0.0076	-0.0060	-0.0006	-0.0058	0.0016	0.0049	-0.0017	-0.0130	0.0020	-0.0074
E	PV	0.0012	0.0594	0.0577	0.0321	-0.0025	0.0923	-0.0054	0.0573	-0.0014	0.0968
	MV	0.0000	0.0560	0.0600	0.0290	0.0000	0.0920	0.0000	0.0590	0.0000	0.0950
	SDevM	0.0069	0.0056	0.0051	0.0049	0.0084	0.0068	0.0105	0.0070	0.0059	0.0060
	PV-MV	0.0012	0.0034	-0.0023	0.0031	-0.0025	0.0003	-0.0054	-0.0017	-0.0014	0.0018
F	PV	-0.0021	0.0596	0.0030	0.1007	0.1630	0.0978	0.0828	0.0549	0.0682	-0.0024
	MV	0.0000	0.0480	0.0000	0.0970	0.1750	0.0810	0.0820	0.0470	0.0790	0.0000
	SDevM	0.0073	0.0060	0.0054	0.0052	0.0090	0.0064	0.0098	0.0066	0.0055	0.0056
	PV-MV	-0.0021	0.0116	0.0030	0.0037	-0.0120	0.0168	0.0008	0.0079	-0.0108	-0.0024

C3. Charts that show the prediction results and the standard deviations of predictions

T1-T5 AVE are the external test samples averaged over three measurements and S1-S5 are the internal test samples.

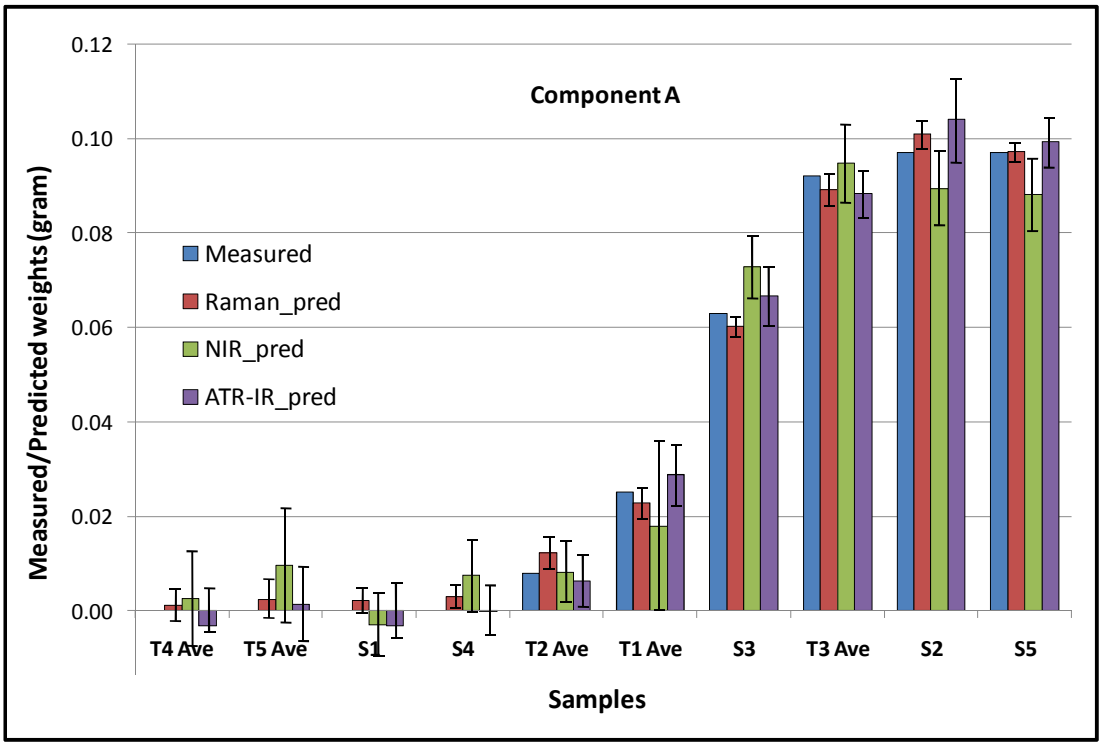


Figure C-1: Prediction results for component A

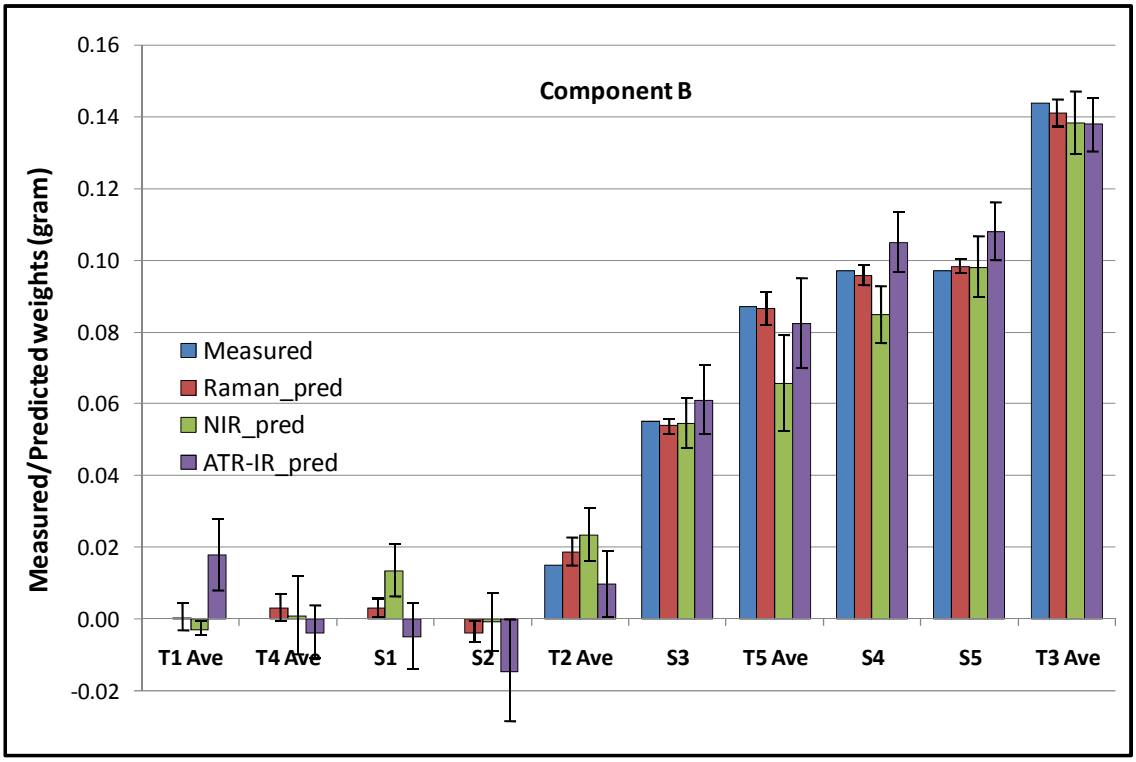


Figure C-2: Prediction results for component B

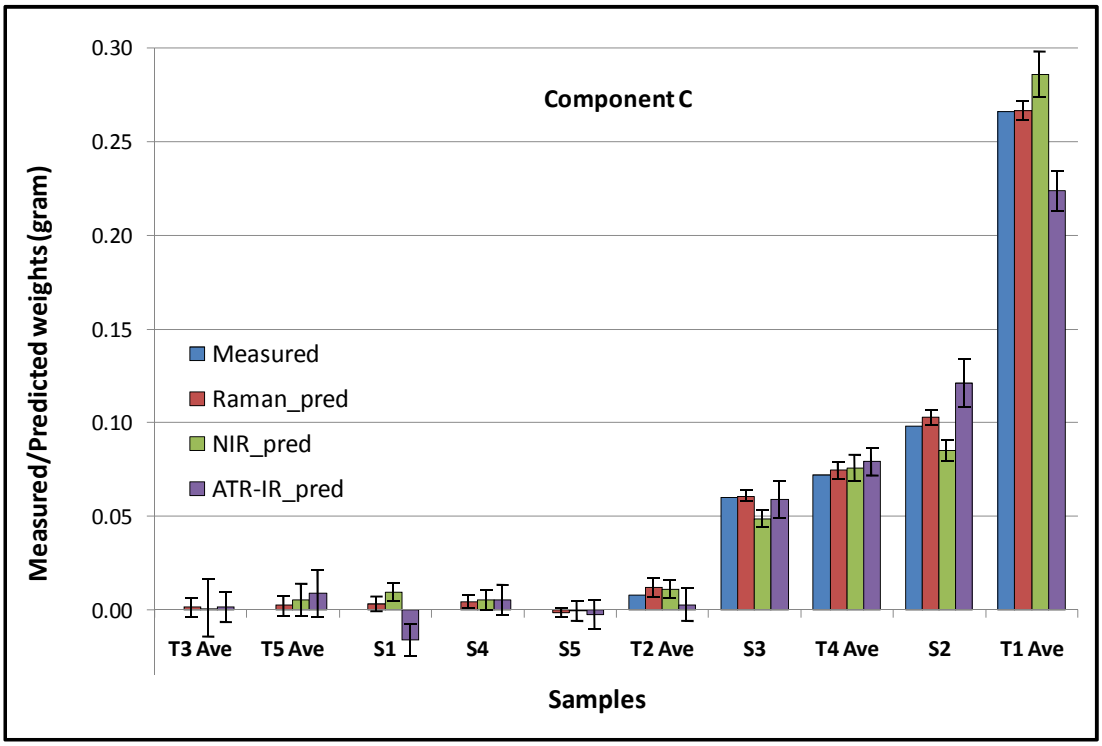


Figure C-3: Prediction results for component C

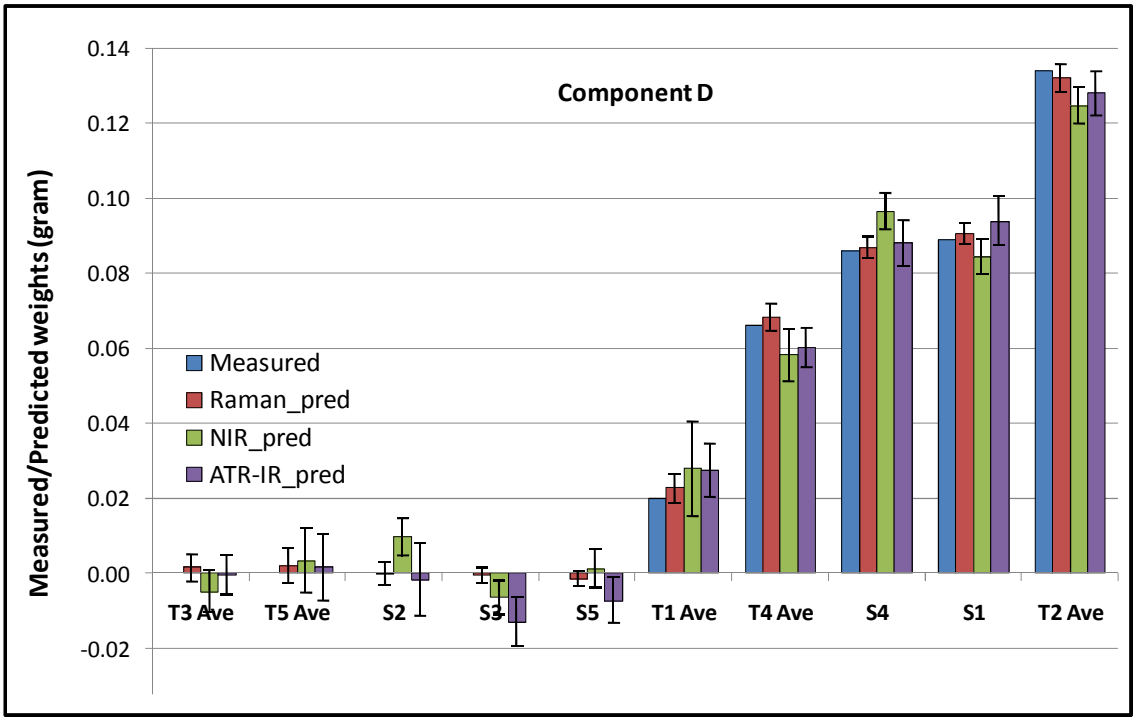


Figure C-4: Prediction results for component D

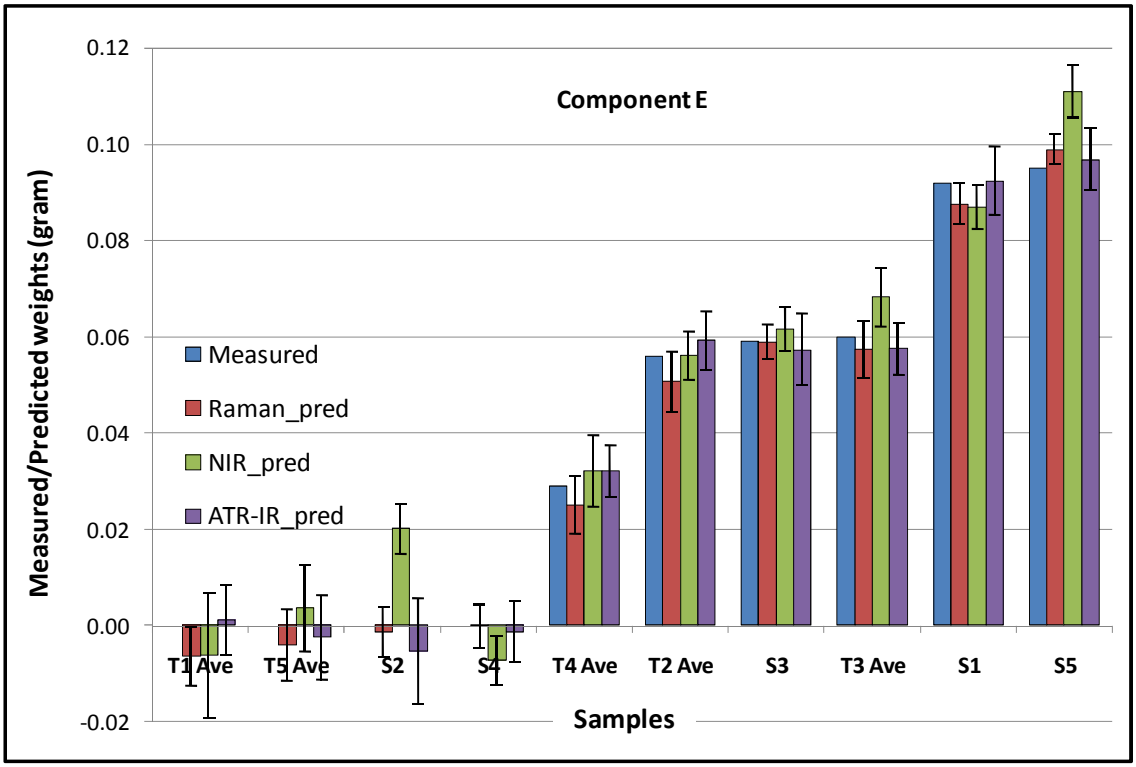


Figure C-5: Prediction results for component E

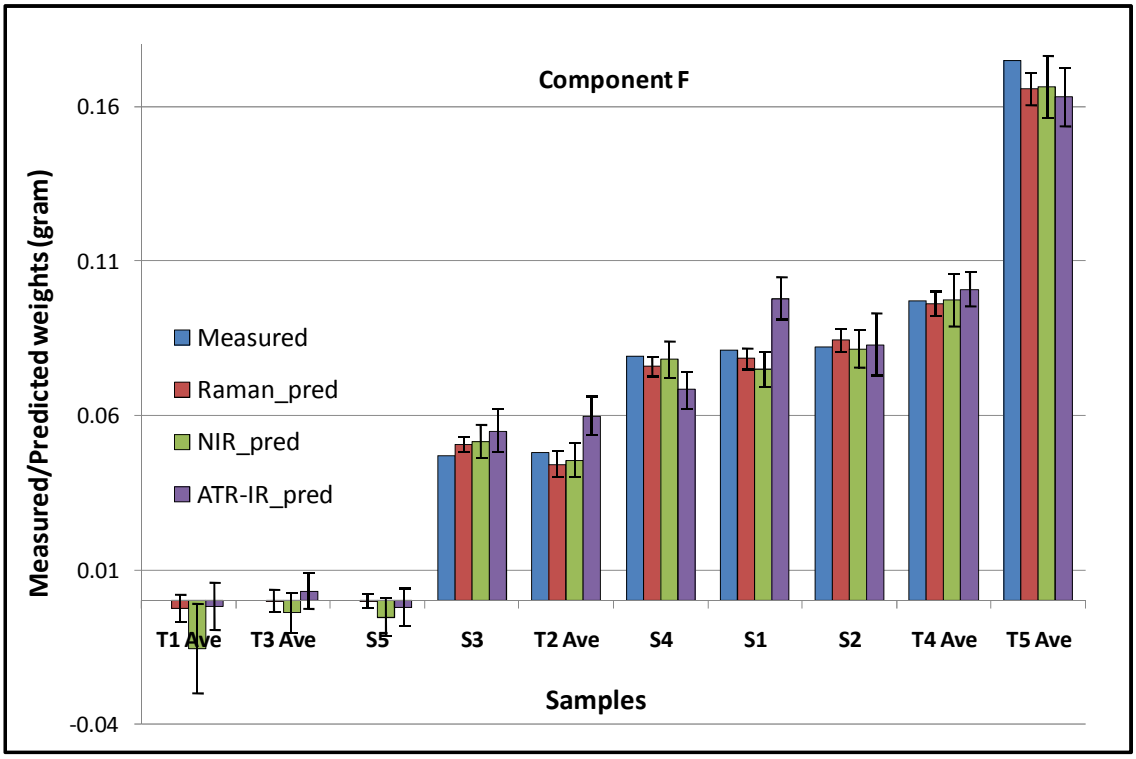


Figure C-6: Prediction results for component F

Appendix D. Validation and errors

Index

D1. Validation results

Each regression was repeated three times, each time with another validation technique: Full cross-validation; cross-validation with 13 segments and a cross-validation with 2 segments. The average error values of the regression models were taken over these three techniques and the standard deviation over the three values was calculated.

Table D-1: Layout and meaning or terms used in validation result tables

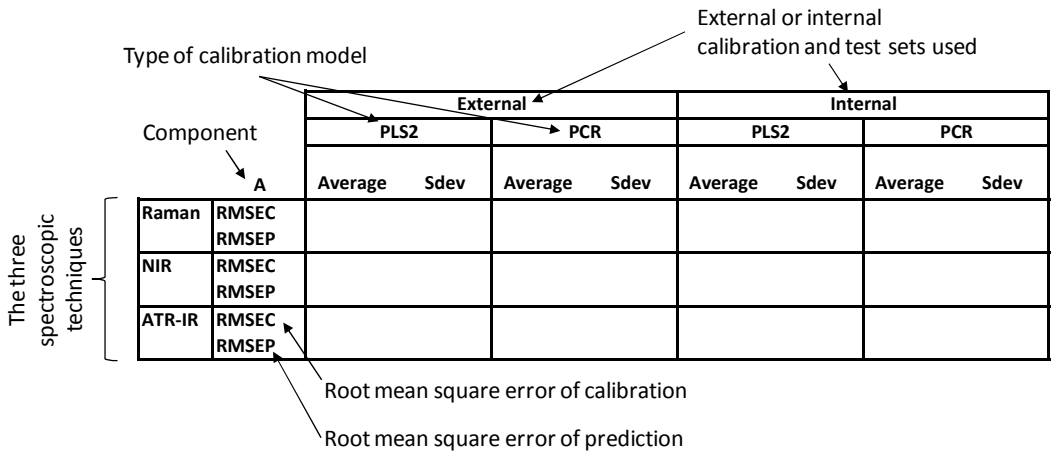


Table D-2: Average of root mean square errors and the standard deviation thereof between three validation techniques as calculated for component A

		External				Internal			
		PLS2		PCR		PLS2		PCR	
	A	Average	Sdev	Average	Sdev	Average	Sdev	Average	Sdev
Raman	RMSEC	0.0031	0.0000	0.0031	0.0000	0.0031	0.0000	0.0031	0.0000
	RMSEP	0.0033	0.0001	0.0032	0.0000	0.0033	0.0001	0.0033	0.0000
NIR	RMSEC	0.0108	0.0000	0.0099	0.0000	0.0108	0.0000	0.0101	0.0000
	RMSEP	0.0126	0.0022	0.0111	0.0003	0.0137	0.0034	0.0113	0.0001
ATR-IR	RMSEC	0.0061	0.0000	0.0068	0.0000	0.0063	0.0000	0.0073	0.0000
	RMSEP	0.0081	0.0007	0.0085	0.0001	0.0085	0.0008	0.0094	0.0007

Table D-3: Average of root mean square errors and the standard deviation thereof between three validation techniques as calculated for component B

		External				Internal			
		PLS2		PCR		PLS2		PCR	
	B	Average	Sdev	Average	Sdev	Average	Sdev	Average	Sdev
Raman	RMSEC	0.0036	0.0000	0.0036	0.0000	0.0037	0.0000	0.0037	0.0000
	RMSEP	0.0039	0.0001	0.0039	0.0001	0.0040	0.0001	0.0039	0.0001
NIR	RMSEC	0.0117	0.0000	0.0108	0.0000	0.0116	0.0000	0.0108	0.0000
	RMSEP	0.0138	0.0008	0.0134	0.0017	0.0134	0.0003	0.0124	0.0004
ATR-IR	RMSEC	0.0101	0.0000	0.0102	0.0000	0.0100	0.0000	0.0100	0.0000
	RMSEP	0.0134	0.0008	0.0136	0.0008	0.0135	0.0008	0.0140	0.0017

Table D-4: Average of root mean square errors and the standard deviation thereof between three validation techniques as calculated for component C

		External				Internal			
		PLS2		PCR		PLS2		PCR	
	C	Average	Sdev	Average	Sdev	Average	Sdev	Average	Sdev
Raman	RMSEC	0.0042	0.0000	0.0042	0.0000	0.0042	0.0000	0.0043	0.0000
	RMSEP	0.0044	0.0001	0.0045	0.0000	0.0046	0.0001	0.0045	0.0001
NIR	RMSEC	0.0063	0.0000	0.0063	0.0000	0.0062	0.0000	0.0062	0.0000
	RMSEP	0.0096	0.0036	0.0079	0.0013	0.0086	0.0015	0.0071	0.0003
ATR-IR	RMSEC	0.0088	0.0000	0.0091	0.0000	0.0084	0.0000	0.0088	0.0000
	RMSEP	0.0112	0.0006	0.0116	0.0002	0.0117	0.0011	0.0127	0.0017

Table D-5: Average of root mean square errors and the standard deviation thereof between three validation techniques as calculated for component D

		External				Internal			
		PLS2		PCR		PLS2		PCR	
	D	Average	Sdev	Average	Sdev	Average	Sdev	Average	Sdev
Raman	RMSEC	0.0034	0.0000	0.0035	0.0000	0.0034	0.0000	0.0035	0.0000
	RMSEP	0.0036	0.0000	0.0037	0.0000	0.0037	0.0000	0.0037	0.0001
NIR	RMSEC	0.0071	0.0000	0.0072	0.0000	0.0072	0.0000	0.0072	0.0000
	RMSEP	0.0077	0.0002	0.0086	0.0010	0.0078	0.0001	0.0080	0.0001
ATR-IR	RMSEC	0.0069	0.0000	0.0070	0.0000	0.0069	0.0000	0.0070	0.0000
	RMSEP	0.0090	0.0011	0.0086	0.0005	0.0086	0.0003	0.0082	0.0007

Table D-6: Average of root mean square errors and the standard deviation thereof between three validation techniques as calculated for component E

		External				Internal			
		PLS2		PCR		PLS2		PCR	
	E	Average	Sdev	Average	Sdev	Average	Sdev	Average	Sdev
Raman	RMSEC	0.0057	0.0000	0.0060	0.0000	0.0057	0.0000	0.0061	0.0000
	RMSEP	0.0062	0.0003	0.0064	0.0001	0.0061	0.0000	0.0063	0.0002
NIR	RMSEC	0.0069	0.0000	0.0069	0.0000	0.0067	0.0000	0.0067	0.0000
	RMSEP	0.0079	0.0001	0.0094	0.0018	0.0090	0.0022	0.0078	0.0001
ATR-IR	RMSEC	0.0071	0.0000	0.0071	0.0000	0.0073	0.0000	0.0073	0.0000
	RMSEP	0.0091	0.0007	0.0088	0.0004	0.0092	0.0003	0.0100	0.0007

Table D-7: Average of root mean square errors and the standard deviation thereof between three validation techniques as calculated for component F

		External				Internal			
		PLS2		PCR		PLS2		PCR	
	F	Average	Sdev	Average	Sdev	Average	Sdev	Average	Sdev
Raman	RMSEC	0.0037	0.0000	0.0037	0.0000	0.0037	0.0000	0.0037	0.0000
	RMSEP	0.0040	0.0000	0.0040	0.0001	0.0043	0.0004	0.0039	0.0001
NIR	RMSEC	0.0082	0.0000	0.0070	0.0000	0.0081	0.0000	0.0070	0.0000
	RMSEP	0.0107	0.0026	0.0087	0.0013	0.0093	0.0001	0.0078	0.0002
ATR-IR	RMSEC	0.0070	0.0000	0.0075	0.0000	0.0069	0.0000	0.0076	0.0000
	RMSEP	0.0087	0.0004	0.0094	0.0002	0.0090	0.0001	0.0101	0.0003

D2. Measurement and instrumental errors

Table D-8: The error introduced by making up the samples by transferring components with micropipette to a NMR tube and weighing the component in gram

g/µl	A	B	C	D	E	F
Centre sample no1	1.8E-03	1.3E-03	1.6E-03	1.3E-03	1.5E-03	1.1E-03
Centre sample no2	1.5E-03	1.4E-03	1.4E-03	1.3E-03	1.4E-03	1.2E-03
Centre sample no3	1.5E-03	1.4E-03	1.5E-03	1.3E-03	1.5E-03	1.2E-03
Sdev	1.4E-04	3.5E-05	9.0E-05	1.7E-05	4.6E-05	3.0E-05
Ave	1.6E-03	1.4E-03	1.5E-03	1.3E-03	1.5E-03	1.2E-03

Table D-9: The instrumental error found when repeat measurements are done on the same centre sample

	Centre Samples	Min	Max	Mean	Sdev
Raman	S41 no1	2.0E-06	5.4E-04	8.5E-05	8.3E-05
	S41 no2	1.5E-06	5.4E-04	8.4E-05	8.2E-05
	S41 no3	1.8E-06	5.4E-04	8.4E-05	8.2E-05
	Sdev	2.8E-07	3.4E-06	3.4E-07	5.5E-07
NIR	S41 no1	1.3E-03	2.5E-03	1.9E-03	3.5E-04
	S41 no2	1.4E-03	2.4E-03	1.8E-03	3.1E-04
	S41 no3	1.3E-03	2.2E-03	1.7E-03	2.9E-04
	Sdev	5.1E-05	1.5E-04	1.0E-04	2.9E-05
ATR-IR	S66 no1	4.4E-03	9.7E-03	8.7E-03	1.2E-03
	S66 no2	3.8E-03	8.5E-03	7.6E-03	1.1E-03
	S66 no3	4.4E-03	1.0E-02	8.9E-03	1.3E-03
	Sdev	3.4E-04	7.9E-04	6.8E-04	1.1E-04

Table D-10: The error introduced by variations in the glass structure of the NMR tube in NIR spectroscopy

NIR	Min	Max	Mean	Sdev
0deg	0.1894	0.3225	0.2702	0.0385
90deg	0.1924	0.3274	0.2746	0.0390
180deg	0.1957	0.3326	0.2793	0.0396
270deg	0.1941	0.3310	0.2777	0.0395
360deg	0.1914	0.3259	0.2732	0.0388
Sdev	0.0024	0.0040	0.0036	0.0004