The University of Hull

The Application of Chemometrics to Spectroscopic and Process Analytical Data

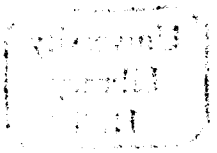
Being a Thesis Submitted for the Degree of Doctor of Philosophy

in the University of Hull

By

Victoria Catherine Loades, MChem

March 2003



Acknowledgements

During the entire course of this work I have been supervised by Dr A. D. Walmsley whose extensive help, guidance and enthusiasm provided a wealth of opportunities for me to carry out my research, many thanks are owed for both this and his continued support.

This work has been funded by the Engineering and Physical Sciences Research Council (EPSRC) with assistance also from the Centre for Process Analytics and Control Technology (CPACT). I would like to thank those that I have collaborated with and for the opportunities it has given me.

I would like to thank everyone one who has helped and supported my in any and everyway throughout this work especially including colleagues in the Chemometrics and Analytical Science Group at Hull University, especially Della who has helped keep me on the writing up rails.

My greatest thanks and appreciation goes to Steve who has been there for me throughout.

Glossary of terms

At-line	A manual sample is taken from the process that is then analysed close to the process or within the manufacturing plant.
ANN	Artificial neural network
ANOVA	Analysis of variance
Auto	Autoscaling
Correlation	Product moment correlation coefficient
CPAC	Center for Process Analytical Chemistry (US)
CPACT	Centre for Process Analytics and Control Technology
CV	Cross validation
Dx6	Design Expert v6.0
GC	Gas Chromatography
GMS	Guided microwave spectrometer
HPLC	High pressure liquid chromatography
In-line	The sample to analyser interface is located directly on the process stream, removing the need to re-circulation loops.
LV	Latent variable
LR	Linear regression
MLR	Multiple linear regression
MCEC	Measurement & Control Engineering Center (US)
ML	Matlab
MNCN	Mean centring
MSC	Multiplicative scatter correction
MWS	Microwaves
NIR	Near infrared spectroscopy
NIST	National Institute of Standards and Technology (US)

Off-line	A sample is taken manually then transported to the central laboratory for analysis by skilled technicians.
On-line	An automated system takes the sample from the process stream the transports the sample to the process analyser.
OSC	Orthogonal signal correction
PAC	Process analytical chemistry
PC	Principal component
PCA	Principal component analysis
PLS	Partial least squares
Poly-PLS	Polynomial partial least squares
PLST	PLS_Toolbox
PRESS	Predicted error sum of squares
RI	Refractive index
RR	Ridge regression
RMSPE	Root mean square value of prediction error
SG/Savgol	Savitsky Golay filtering
Spl-PLS	Spline inner relationships partial least squares
SNV	Standard normal variate
UV/Vis	Ultraviolet/visible spectroscopy
Vis	Visible spectroscopy
VS-MLR	Variable-selection multiple linear regression
WRR	Weighted ridge regression
XRF	X-ray fluorescence
XRD	X-ray diffraction

Abstract

The research has included collaboration with number of different companies and consortiums involving spectroscopic measurements with the application of chemometric techniques.

For the 'European Framework 5', Standards Measurements and Testing (SMT) chemometrics network consortium a certified reference dataset based on visible metals complex spectra was developed. An inter-laboratory study was carried out which demonstrated the between subject significant difference for chemometric data analysis.

An industrial collaboration with BNFL, Springfield's, this work consisted of producing a PLS regression model which could be used to predict levels of uranyl and nitrate in uranyl nitrate liquors samples, which were analysed by Raman spectroscopy which was insensitive to temperature.

A substantial amount of work has been in the development of GMS with multivariate calibration for process analysis. The GMS is designed for the analysis of flowing mixtures, slurries and moisture content. The method is currently hindered by the existing calibration method; here PCA, PLS and weighted ridge regression (WRR) have been applied to the broadband, complex spectra to successfully allow measurement of a range of samples including; aqueous, organic, fermentation and non-homogeneous samples.

Table of Contents

Chapter 1: Introduction

Chapter 2: Development of a Standard Reference Dataset Based on Visible Spectra of Metal Complexes

Chapter 3: Analysis of Uranyl Nitrate Liquors by UV/Vis and Raman Spectroscopies

Chapter 4: Application of GMS for Chemical Measurement

Chapter 5: Monitoring the Beer Fermentation Process by GMS

Chapter 1

Introduction

1 Background

The use of process analytical chemistry (PAC) to monitor and control industrial chemical processes is becoming more wide spread. The main centres involved in research in this area are the American MCEC[1] and CPAC[2] and their British counterpart CPACT[3].

The motivation to research into process analysis is to develop new techniques and applications to allow real-time monitoring of industrial processes. Traditionally a sample would be taken from the process then transported to the central site laboratory for analysis; the results would be reported back to the plant control personnel. The procedure could take anything from an hour to days to achieve the analysis, during which time the process is continuing blind to the potential knowledge of a process upset, which could lead to lost batches, lead times or unnecessary over processing of a completed stage of the process. All of these scenarios would lead to direct financial loss for the company. The logical step was to move the analysis near to or directly within the process, producing on- or in-line measurement systems.

In recent years a substantial proportion of analytical methods of measurement have been successfully introduced as on/in/at line systems during a chemical manufacturing process. The types of process analysis include[4]; Off-line, At-line, in-line and on-line.

The analysers required for process analytical chemistry need to be simple to operate and maintain, substantially more robust to stand up to the harsh environments (extreme temperatures, flammable atmospheres, etc.), that can be encountered on a process plant (i.e., they must be intrinsically safe). Ideally process analysers should be easily operated and require minimal maintenance.

In theory, all analytical methods should be able to be modified for process analysis. Where possible, simple measurements are favoured; for example determination of analyte physical properties such as pH or RI where one response is given for the measurement. Where these are not sufficient to give the required information, spectroscopic or chromatographic methods are often utilised.

Some fixed wavelength spectroscopic methods are relatively simple. Where multiple wavelength spectroscopies are employed, interpretation of the measurement responses is not always straight forward. These often require the use of chemometric techniques to extract the desired information from a given signal.

The research undertaken in this thesis covers several areas grouped together with the common theme of developing measurements for process analytical chemistry, in particular, spectroscopic methods of analysis.

2 Process analysis

When choosing a system for process analysis, there are several factors that require consideration: knowledge of what chemical and physical properties are required for the process, (e.g. temperature, pressure of the process stream); how fast the process is progressing; time available for analysis (immediate or would every few minutes suffice?); and the accuracy and precision of measurement required.

The measurements for process analytical chemistry fall into three main categories, 'Wet Chemistry' measurements are typically based on physical parameters (e.g. pH, RI, density, calorimetry), but also include titrations and flow injection analysis.

The remaining two categories are spectroscopy and chromatography. GC is often used on-line for analysis in the petroleum industry for the measurement of octane number[5]. The application of HPLC for on-line process analysis has been discussed[6], it can offer significant advantages over spectroscopic or flow injection methods as it can analyse complex mixtures of a number of components over a wide concentration range using a fairly simple calibration. Both types of chromatographic method are fairly laborious to install as an on-line system. A major problem with chromatographic methods is that they require regular maintenance and for this the instrument is often required to be taken off line meaning that the plant either has to be shut down or operate without analysis. Also of concern are the flammable gases required for GC and the cost and hazardous nature of many of the solvents required as mobile phases for HPLC.

Whilst these methods are suitable for complex samples, on-line spectroscopy has been taken up more readily due to its simpler instrumentation in comparison.

Applications of spectroscopy in process analysis include; UV/Vis, fluorescence, chemiluminescence, NIR, MIR, Raman, atomic and NMR techniques[7]. On-line MS has been shown to be suitable for monitoring bioprocesses[8]. Raman spectroscopy is a relatively new method for process analysis. The main advantage of this method is that it is virtually transparent to water, and therefore it is an ideal solution for processes with a large amount of background water. The esterification of ethanol by acetic acid has been successfully analysed by Raman spectroscopy[9]. The disadvantage of this method is the reduced sensitivity in comparison to NIR and MIR.

Many of the problems with process analysers are due to sampling; correct sampling procedures to give true representation of the process are essential for any type of process analysis. The necessary consideration include, is the sample reaching the analyser, what is the sample homogeneity, is it separating or different from the process stream, i.e., is the sample representative.

An industry-wide problem has been identified when implementing the chosen process analyser system; these are conflicts in fixtures and fittings, operating software, etc. A group under the CPAC umbrella has been commissioned to investigate further into this area, 'NeSSI' (New Sampling/Sensor Initiative)[10], part of the focus of this group is to simplify and streamline the sampling process.

3 Experimental design

Any measurements and subsequent calibration model is only as good as the samples it is based on, therefore the design and planning of experiments is essential. Introduction to the principles and practices of experimental design can be found in the books 'Design and Analysis of Experiments' By Montgomery[11] and 'Response Surface Methodology' by Myers and Montgomery[12]. Software is available to calculate optimal experimental designs (essential for more complex designs) and include 'Design Expert™' by Stat-Ease Inc [13] and 'MODDE™'[14] by Umetrics. The computer-based designs are useful, reducing the number of experiments required for complex designs. These designs are appropriate for unusual design shapes that can occur when there are constraints on the levels of the components.

The term 'Experimental Design' is described by Miller and Miller[15] to define the stages that; (i) identify factors which will influence the result of the experiment, (ii) design the experiment so that the effects of uncontrolled factors are minimised, (iii) use of statistical analysis to separate and evaluate the effects of the various factors involved.

Experimental design is used in this work to plan the number and concentrations of samples for experiments, especially where there are a several components. A number of different design types have been used, for example, full / partial factorial, mixture and optimal designs.

3.1 Factorial design

A tradition approach to planning an experiment is to vary one of the factors at a time, e.g. pH, concentration, time, temperature. When many factors are involved, this can be a lengthy process. The major disadvantage of this method is that it can not account for any potential interactions between the factors.

To overcome this issue, the principle of factorial experimental design, where factors are varied together, is established. For most designs the factors are varied at different levels; these are known as factorial design and have the notation of N^k where N is the number of levels and k is the number of factors. For a two level, two factor design, the number of experiments required is $2^2 = 4$, this increases to $2^3 = 8$ and $2^4 = 16$ as the number of factors in the two level experiment increases.

3.2 Optimal experimental design

Where several factors are varied, as opposed to two or three in factorial designs, a response surface, more sophisticated than that of a cube of the factorial designs is produced. Optimal designs are complex computer generated designs of which there are several types; D, G, and F. D-optimal design is used in the scope of this work as it will produce a maximum variance in the samples with a minimum number of samples. It will work by producing maximum variation in the response surface.

4 Statistical analysis

4.1 Determination of outliers

An outlier is a value in a set of results which appears to be very different from the remaining values in the dataset. There are two main tests, Dixon's[16] and Grubbs[17], to evaluate whether the suspected value is statistically an outlier. Dixon's test works by comparing the suspected value to the nearest value in the dataset to it. Grubbs test is often used in preference to Dixon's and is recommended for use by ISO. These methods are only relevant for samples from a normal distribution.

4.1.1 Grubbs test

The test for outliers followed by Grubbs compares the deviation between the suspected values and the mean and standard deviation of the samples in the dataset. The null hypothesis (H_0) is all the values come from the same population. If G < Gcrit, then the null hypothesis is true and the sample is not an outlier; if G > Gcrit, the null hypothesis is untrue and the sample is an outlier. The critical values for G (Gcrit) can be found in statistical tables.

$$G = \frac{\left|value - \bar{x}\right|}{s.d}$$

Equation 1 Grubbs test for outliers

Where:
G = statistic G
value = suspected outlier value
$\bar{x}$ = mean
s.d = standard deviation

4.2 Analysis of variance

Analysis of variance (ANOVA) is a statistical technique that can be used to identify variations in data and compare whether data is statistically the same. Miller and Miller describe the technique well[18].

For this work, the most relevant use of ANOVA is the comparison of several means within samples and between samples. For comparing between sample variations, a one-tailed significance F-test is used because there is only one source of independent variation. For the ANOVA test, the null hypothesis is that there is not a significant variation between the samples. Similar to Grubbs test, if this is true, then the F value calculated will be less than Fcrit and if there is a significance difference in the Fcrit > Fcalc.

The parameters used to calculate the F critical values in the F tables are based on degrees of freedom. These are the number of independent pieces of information that go into the estimate of a parameter[19].

For a one-way ANOVA test, comparing the difference between two datasets there is between sample variation with degrees of freedom = h – 1, where h = number of samples. The within sample variation has degrees of freedom = h(n-1), where n is the number of members. The total degrees of freedom is N - 1 where N = nh = total number of measurements.

5 Chemometrics

Generally, chemometrics concerns the application of statistical analysis to chemical data to gain knowledge of the process/system under scrutiny.

The progress of the field of chemometrics has been well documented in a range of tutorial and review articles (see Barry Lavine in 'Analytical Chemistry' every 2 years discussing the currents trends and major new developments[20],[21],[22],[23]).

There are two main areas of chemometrics; supervised and unsupervised. Unsupervised methods, such as PCA and cluster analysis, are used to distinguish trends in the data without the benefit of reference information. The second area is known as supervised modelling, where the reference information is known and can be used for calibration. Such methods include PLS, PCR, MLR etc.

Examples of general introduction to multivariate calibration, including MLR, PCR and PLS regressions can be found in the book[24] and tutorial text[25]. A development in PLS has discussed by Wold et al[26], with proposal of applying OSC to the data to remove effects which have no correlation to the reference information before PLS calibration.

Generally, standard methods of MLR and PLS with the modified RR method of WRR are used for calibration in this work as the type of measurements calibrated are often complex, requiring calibrations which will not raise as many questions as the measurement.

One of the issues of using chemometrics in industry has been the lack consistency between software from different companies that is used manipulate the data. An initiative, 'Chemometrics for On-line Process Analysis' (COPA)[27], has been formed as a partnership between the analyzer vendors, chemometrics software vendors, and users to streamline the application of chemometric techniques to process analytical methods. Part of the aim is to gain consistency between various companies' software packages to produce spectral data in a standard format, which could then easily be transferred to any chemometric software for data processing.

5.1 Unsupervised modelling

In this research unsupervised modelling in the form of principal component analysis (PCA)[28] has been used to track trends in spectral data. Alternative methods of unsupervised modelling, such as multivariate curve resolution (MCR), can also be used. An application of MCR to on-line spectroscopy can be found in the paper by Miller[29].

5.1.1 PCA

PCA is a technique for reducing the amount of data when there is correlation present within the data. The decomposition of PCA is detailed in frame 5.1.1-1[30]. The data matrix is reduced to a scores matrix and a loadings matrix. The scores give the information regarding any trends between samples and the loadings the variation and importance of the variables. The principal components are ordered

such that the first accounts for the majority of the variation within the data leading to the last which has least variation within the data. The maximum number of principal components that can be extracted is the same as the number of samples. However, this is often capped so that the total variation captured within the principal components reaches a certain level, e.g. 95 % of the total variation in the data or the level of the noise in the data, if it was known.

Frame 5.1.1-1 PCA Decomposition

For a matrix (X) with m rows and n columns the covariance matrix of X is;

$$Cov(\mathbf{X}) = \frac{\mathbf{X}^T\mathbf{X}}{m-1} \qquad \text{Equation 2}$$

The data matrix (X) is decomposed as a sum of the outer product vectors t_i and p_i and a residual matrix, E:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \ldots + \mathbf{t}_K\mathbf{p}_k^T + \mathbf{E} \qquad \text{Equation 3}$$

The t_i vectors are the scores; these describe how the samples relate to each other; the p_i vectors are the loadings which describe how the variables relate to each other. These are related to eigenvectors in the covariance matrix:

$$Cov(\mathbf{X})\mathbf{p}_i = \lambda_i\mathbf{p}_j \qquad \text{Equation 4}$$

Where λ_i is the eigenvalue associated to the eigenvector p_i (loadings). The original data matrix is related to pairs of scores (t_i) and loadings in the following equation;

$$\mathbf{X}\mathbf{p}_i = \mathbf{t}_i \qquad \text{Equation 5}$$

The original data matrix multiplied by the loadings (p_i) will result in the scores matrix. The eigenvalue shows the amount of variance between the t_i and p_i pairs. The loadings plots from PCA can be used to identify noise in the data. If the data hasn't been mean centred then the first loading plot will be the average spectrum of the data.

5.2 Supervised modelling

Multivariate calibration enables prediction of analyte concentration in the presence of varying amounts of spectrally active interferences (their contribution is modelled). In applications of multivariate calibration the aim is to predict a property of interest from a multivariate measurement by using a model[31].

The objective of multivariate calibration is to build a model that describes the relationship between the dependent variables (concentrations) and independent variables (spectra). Validation of the calibration model is essential to ensure that it is able to predict independent samples.

The method of MLR searches for a single factor of correlation between the predictor variables (e.g. spectra) and the predicted variables (e.g. concentrations). It is the simplest method which is most often successful in fairly simple situations. Variable selection procedures, when used in conjunction with MLR (VS-MLR), have been shown to improve prediction[32]. This method is seriously affected by collinearity within the X-block (spectral) data producing unreliable model coefficients which cannot successfully be employed to predict test samples.

If MLR is unsuccessful for calibration then PCR can be tried, this is a factor analysis method that maximises the variance in the data to improve prediction. This method is an extension to PCA. As for PCA the data is separated into a number of factors (LVs) the calibration model regression is then performed based on these factors. If PCR is not chosen, then PLS can be employed, this finds factors as for PCR but seeks to achieve maximum variance and correlation in the data, maximising the covariance in the data. As PLS will find the correlation and variance in the spectra it is often considered to be superior to PCR and subsequent models regularly require fewer latent variables to capture the information in the data.

The main drawback of the factor analysis methods of PCR and PLS is the number of factors to be included has to be decided. By using too few LVs, features in the original data maybe excluded. If too many LVs are used, information maybe included that does not relate to the reference data, this can cause interference and instability in the calibration model.

Another regression technique is ridge regression[33]. This is calculated in a similar way to MLR but the inversion matrix is stabilised by adding a constant to the diagonal. The benefit of this method is that the data is not reduced into factors which then have to be decided for inclusion.

For the methods described in this section it is assumed that the error is only within the dependent variables, but this is can limit the ability to produce a realistic

calibration model in practice. Faber and Kowalski[34] suggest an alternative expression which also takes into account the measurement error propagation of the independent variables. For this it is assumed that there is an exact linear relationship between the dependent and independent variables, the resulting expressions produced are validation for classical errors in variables models.

5.2.1 Partial least squares

A tutorial describing PLS regression was written by Geladi[35]. The PLS algorithm can take the form of either SIMPLS or NIPALS. Also, it can follow PLS1 or 2 procedures. For PLS1 the calculation is for only 1 component at a time, whereas application of PLS2 will allow calculation of multiple components.

The number of latent variables chosen to model the data is often decided based on how many are required to give the lowest error of prediction with a sensible number of variables in the samples considered, i.e., if there are 4 components in the samples, the use of 10 LVs is likely to over fit the data. The number of LVs used is generally where the error in prediction of the calibration and test dataset change very little (often when there is less than 2 % improvement) with the addition of more latent variables. The RMSPE is the error in prediction of samples, and the equation can be seen in section 5.2.3 where procedures for validating models are discussed.

In addition to the RMSPE, the correlation coefficient of the line of best fit of predictions and prediction residuals should also be taken into account.

Often it can be better to choose too few LVs than too many because this can lead to over fitting of the data. This may lead to a situation where the model works perfectly for the calibration data but when used for validation or unknown data results in very poor errors of prediction.

5.2.1.1 PLS SIMPLS

This is a later development of PLS to that of the traditional NIPALS algorithm. SIMPLS is often much quicker to compute, but this is becoming less of a factor as computers processing capability increases. A difference between SIMPLS and NIPALS is that the x-variance (spectral) information is contained within the loads and instead of the scores. For univariate calculations the results are the same for both types of PLS algorithm, but for multivariate applications there can be differences in results.

5.2.1.2 PLS NIPALS

For the standard form of the PLS algorithm, non-iterative Partial Least Squares (NIPALS)[36] has been used. This method is useful when there is more than one predictor variable. The NIPALS algorithm (described in frame 5.2.1.2-1) calculates the scores (t), loadings (p), weights (w). It can work for more than one predictor variable, and the (y) scores (u) and loadings (q) are predicted for the Y-block. Also calculated are the 'Inner-relationship' vector coefficients (b) which relates to the X- and Y-block scores.

Frame 5.2.1.2-1 PLS decomposition

The column of Y, y_j with most variance is the starting estimate of u_j.

In the X data block:

$$\mathbf{w}_1 = \frac{\mathbf{X}^T\mathbf{u}_1}{\left\|\mathbf{X}^T\mathbf{u}_1\right\|}$$

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$$

In the Y data block:

$$\mathbf{q}_1 = \frac{\mathbf{u}_1^T\mathbf{t}_1}{\left\|\mathbf{u}_1^T\mathbf{t}_1\right\|}$$

$$\mathbf{u}_1 = \mathbf{Y}\mathbf{q}_1$$

The X data block loadings are calculated and scores and weights re-scaled accordingly;

$$\mathbf{p}_1 = \frac{\mathbf{X}_1^T\mathbf{t}_1}{\left\|\mathbf{t}_1^T\mathbf{t}_1\right\|}$$

$$\mathbf{p}_{1new} = \frac{\mathbf{p}_{1old}}{\left\|\mathbf{p}_{1old}\right\|}$$

$$\mathbf{t}_{1new} = \mathbf{t}_{1old}\left\|\mathbf{p}_{1old}\right\|$$

$$\mathbf{w}_{1new} = \mathbf{w}_{1old}\left\|\mathbf{p}_{1old}\right\|$$

The regression coefficients b is determined for the inner relationships

$$b_1 = \frac{\mathbf{u}_1^T\mathbf{t}_1}{\mathbf{t}_1^T\mathbf{t}_1}$$

Once the scores and loadings for the first latent variable are calculated the residuals are determined;

$$\mathbf{E}_1 = \mathbf{X} - \mathbf{t}_1\mathbf{t}_1^T$$

$$\mathbf{F}_1 = \mathbf{Y} - \mathbf{b}_1\mathbf{u}_1\mathbf{q}_1^T$$

The method is repeated but this time for the next latent variable. Where X and Y are seen they are replaced by residuals E_1 and F_1.

PLS forms the following inverse matrix;

$$\mathbf{X}^+ = \mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1}\left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T$$

5.2.1.3 Polynomial PLS

The simplest approach for fitting of data proving to be non-linear is that of polynomial PLS. For this the inner relationships are fitted to a polynomial function of desired order[37].

5.2.1.4 Spline PLS

Nonlinear partial least squares with spline inner relationships are described by Wold[38] and in a tutorial by Frank[39]. Spline PLS (SPL-PLS), is an extension of PLS for non-linear inner relationships.

Splines are function estimates that are obtained by fitting piecewise polynomials. The x-range is split into intervals; these intervals are separated by the knot locations. A spline function is defined by the number of knots, their position and location and the coefficients of the polynomial fitted at each location. The degree of the spline ranges from zero upwards, but 1st or 2nd order is most commonly used to prevent overfitting.

5.2.1.5 Orthogonal signal correction

The method of orthogonal signal correction (OSC) was introduced by Wold[40]. This is an alternative pre-processing method that aims to remove systematic noise whilst leaving as much information relating to the concentrations of sample spectra. Pre-processing methods used to remove baseline drift and systematic noise, such as derivatisation, can also remove information relating to the Y matrix (concentrations) of the spectra. In this method the concentrations are included in the calculation and the aim is to remove factors which are orthogonal to Y, i.e. totally unrelated to the concentrations.

Since this first publication several modifications have been made to the approach; two examples of theses are by Fearn[41] and Brown[42]. The Fearn method applies the

same approach but using a modified algorithm with the aim of improving prediction of subsequent calibration models (e.g. PLS). Brown's method is 'Piecewise OSC' (POSC) based on Fearn's algorithm but local features are selected in the spectra and OSC performed over regions instead of the entire spectra at once. Brown has compared the method to the original and Fearn's and found that PLS models based on POSC data required fewer latent variables and with better predictive power for the given (NIR) data.

5.2.2 Ridge regression

Ridge Regression (RR) is method that is based on correlations within the data. It works by adding a value (θ) to the ridge (or diagonal) of the correlation matrix[43].

$$b_F(\theta) = (F'F + \theta I_r)^{-1} F'Y \qquad \text{Equation 6}$$

Where;

b_F	regression coefficients
θ	Ridge constant (positive value between 0-1)
F	mean centred X matrix
$F'F$	correlation matrix
Y	predicted variable (e.g. concentration)
I	identity matrix, size r x r (r = number of data points/wavelengths)

This has the effect of maximizing the variation and orthogonality of the data. A benefit is that the procedure can improve the signal to noise ratio of the spectra. RR is ideal for ill-conditioned/collinear data where $X'X$ (inverse matrix) is near

to or actually singular. Under these conditions problems are incurred when calculating PLS models. Another advantage of the RR over PLS is that it does not decompose the data into latent variables and hence removes the issue of which latent variables to include when modeling data.

5.2.2.1 Weighted Ridge Regression

An adapted version of this method is Weighted Ridge Regression (WRR). For this method it is not necessary to calculate the values of θ allowing the regression coefficients to be computed more quickly.

$$b_F = (F'F + \theta \times diag(F'F))^{-1} F'Y \qquad \text{Equation 7}$$

In this work, calibration models have been calculated using this method instead of the standard RR.

5.2.3 Validation of calibration models

Once a calibration model has been produced it can be used to predict the levels of unknown reference information in samples.

To produce a calibration model the samples are separated into training sets which are used to build the model, and validation sets which are used for validation of the model. The test set can either be randomly selected from the samples available or from additional samples from an experimental design. A common method is to

randomly separate the samples and 2/3's make the training dataset and 1/3 the validation dataset.

The cross validation method of 'leave-one-out' has been used where the number of samples is small or as an initial investigation for calibration. The leave one out method (CV) is where a sample is removed from the original dataset and used as the validation set and the remaining samples are the test set, the model is built and validated and this repeated until all the samples have been removed.

The leave-one-out method can be applied to the training samples to identify the model conditions with lowest error of prediction and then the test samples predicted based on this model. From the predicted values the prediction error sum of squares (PRESS) and root mean square value of prediction error (RMSPE)[44] are determined.

$$PRESS = \sum(y_i - \hat{y}_I)^2 \qquad \text{Equation 8}$$

$$RMSPE = (PRESS/n_t)^{1/2} \qquad \text{Equation 9}$$

Where; n_t Number of samples in test/validation set

y_i Actual value

$\hat{y}_i$ Predicted value

When evaluating a model it is also necessary to plot the residual prediction error for each sample as this can identify rogue samples with unusually high error in comparison to the remainder of the dataset.

$$residual = (y_i - \hat{y}_I) \qquad \text{Equation 10}$$

5.2.3.1 Determination of correlation

The correlation coefficient is a value between -1 and 1 that is used to give an indication of the closeness of the relationship between dependent (measurement) and independent (reference information) values. The product moment correlation coefficient (r) is calculated as in equation 11. For strongly correlated data the value of r is close to +/- 1[45]. Values closer to 0, indicate no correlation in the data.

$$r = \frac{\sum_i \{(x_i - \bar{x})(y_i - \bar{y})\}}{\{[\sum_i (x_i - \bar{x})^2][\sum_i (y_i - \bar{y})^2]\}^{\frac{1}{2}}}$$

Equation 11 Product moment correlation coefficient

Where;

x_i = actual value

y_i = predicted value

$\bar{x}$ = mean of actual values

$\bar{y}$ = mean of predicted values

The correlation coefficient is often calculate from the line of best fit of a straight line graph (equation 12) of actual vs predicted information.

$$y = bx + a$$

Equation 12 Equation of a straight line

Where b is the slope and a is the intercept on the y-axis.

This is used for two situations in this work (i) to give the goodness of fit between actual and predicted values from calibration models; (ii) the correlation coefficient is calculated between reference information and each variable in the spectra to determine the correlation before calibration.

5.2.4 Improvement of calibration models

It can be beneficial to apply scaling, derivatisation, smoothing or combinations of these to the data. These procedures can maximise the variation in the data, reduce spectral noise, improve the correlation between the dependent and independent variables and reduce baseline variation and drift.

5.2.4.1 Scaling methods

When applying any scaling method to data used for calibration models great care should be taken to ensure that the average mean of any training data is consistent with that of the validation and prediction data. If these are different the re-scaling of prediction information can be affected by propagation of errors due to inaccurate factors in the scaling.

Background subtraction

Background subtraction removes the spectra of zero time or zero component from the remaining spectra. This allows the variance within the spectra to be maximised by removing the magnitude of the initial spectra.

Mean centering

A common scaling technique is that of mean centering, where the means of the response variables, (spectra), and the dependent variables (concentrations) are subtracted, removing the magnitude from the data. For the majority of chemometric algorithms, e.g. PCA and PLS it is assumed that the data is mean centred prior to analysis. In a short communication by Seasholtz and Kowalski[46],

it was identified that in certain situations it is better not to mean centre the data i.e. where the response data (i) vary linearly with concentration, (ii) have no baseline (when there is a component with a zero response that does not change in concentration) or (iii) have no closure in the concentrations (for each sample the concentration of all the components add to a constant, e.g. 100 %).

Autoscaling

To autoscale, the mean is subtracted from each variable and then divided by the standard deviation. The data is weighted due to variance and not magnitude; this can be useful when the components are different e.g. concentration, temperature.

Range scaling

This method is also known as normalisation; it converts the measurement between its maximum and minimum value such that the scaled value lies between 0 and 1. The drawback of this method is that it can be supersensitive to outliers[47].

Standard normal variate

Standard normal variate (SNV) can be used to correct for a drifting baseline. It works by autoscaling along the samples.

5.2.4.2 Derivatisation and smoothing

Smoothing is a filtering method that can be used to remove noise from a data set, two example methods are; 'Moving Average' and 'Savitsky Golay'[48]. For moving average the spectra is split evenly into groups, the average of each group is determined then the spectrum is reformed using only the average points. The

Savitsky Golay filtering method is similar; however, this time each point in the group has a weighting with the points furthest away from the central point being weighted the least and the central point being weighted the most. This method is more selective than the simple moving average. The window size for the smoothing filters is often taken to be approximately the square root of the number of data points in spectroscopic data. A drawback of the use of spectral smoothing is that the spectral resolution can be reduced. Derivatisation is used to remove offset or curvature from the data, second derivative is often taken in UV/Vis and NIR[49] to sharpen and resolve overlapping peaks in spectra. This can result in spurious satellite peaks emerging in the spectra. With the Savitsky Golay algorithm, it is possible to apply smoothing and derivatisation simultaneously to reduce this problem.

5.3 Application of chemometrics to spectroscopic data

The commercial benefits of process analysis have been discussed in section 2, including the particular benefits of spectroscopy for measurement. The drawback of spectroscopic methods is that a single response at one wavelength is often not sufficient for the determination of analyte, and an entire spectrum at multiple wavelengths is required. It is then necessary to use chemometric methods to extract the required information from the spectral measurements.

Chemometrics can also be applied to remove artefacts from spectra which were a result of transferring the measurement to an on-line process environment, e.g.

temperature variation, effect of cables used to transfer the measurement signal from the spectrometer to the place of process measurement.

A paper by Smola[50] discusses the qualitative and quantitative analysis of oxytetracycline by NIR. The aim of the work was to replace time consuming analysis of raw materials, which required sample preparation, with analysis by NIR spectrometry. Reference analysis was by the Karl Fisher method (water content) and a colorimetric assay. The results found that with the aid of derivatisation (2^{nd} order) to remove baseline shifts and handling scattering effects, PCA was used to develop a cluster model that could then be used for sample identification. Derivatisation was also used to pre-treat the samples prior to PLS regression to remove between sample variations. The PLS models could predict the water content with test samples with a standard error of +/- 0.0708. PCR was also applied, but gave higher errors of prediction. The work showed that NIR spectroscopy was suitable for the desired application and once implemented will result in cost reductions being achieved.

Another application of chemometrics to NIR spectroscopy has been the analysis of a pharmaceutical process, including a study of different preprocessing techniques[51]. The aim of this work was to investigate the feasibility of replacing an off-line HPLC analysis of a pharmaceutical process with on-line NIR. The analyte solution is chemically complex and greater knowledge was desired. A range of pre-processing techniques have been investigated to maximise the response to the analytes of interest and reduce unwanted variability due to

physical parameters such as temperature and scattering effects. The preprocessing methods included standard techniques, normalization, derivatisation and some recent advances in preprocessing methods such as multiplicative scatter correction (MSC), OSC and optimised scaling. PLS calibration models were used to predict the analytes of interest, except for where optimised scaling was applied, which used PCR. The best model was PCR based on preprocessing by first order differentiation and optimised scaling. Optimised scaling has not had a wide uptake since its introduction in 1992. Briefly the method introduces a scaling vector for each sample, for least squares the intensity of one sample is zero and another 1. A search should be performed to optimise according to the reference sample chosen for an optimal model. This sample dependence is a significant drawback of the method.

The use of Savitsky Golay derivatisation and smoothing to enhance chemical signals is discussed in reference 48. This involved the investigation of Raman spectra which included some that had a higher response due to fluorescence, which varied between samples, than Raman peaks of interest. The spectral correlation coefficient and PCA analysis demonstrated the superiority of 2^{nd} order Savitsky Golay smoothing and derivatisation to suppress background noise and background signals of high intensity and variability.

The application of chemometric techniques to spectroscopic data is liberal, and as shown in the previous examples, often more than one method is applied to the data to gain the most appropriate technique, usually that with lowest error of

prediction. This approach where dependency on a certain chemometric method is potentially problematic, two scenarios are that the data is over-processed to produce unstable calibration or that a measurement method is deemed unsuitable when really insufficient time or expertise has been allocated to the data.

The drawback of apply different techniques to data is that it could become unclear if a measurement method is/is not suitable depending time period allocated by the user has for the chemometric data processing. It could be felt that if enough pre-processing methods and calibration methods are applied to data then eventually there will be correlation between the dependant and independent measurements. During this work the correlation between the raw measurements and the reference information is investigated to establish if there is real correlation in the data in first instance without the need for extensive data processing, and where this has been applied to keep a realistic view on whether the measurements really are of any use.

A collaborative study has found that the between user variation for the type of data processing/pre-treatment methods applied and the evaluation of outliers within the data varies considerably between those performing the data analysis (this area is considered further in chapter 2). These results highlight a drawback of chemometrics in that it is individual-dependent, which could lead to false negatives and positives depending on who performed the analysis.

The methods of PLS, PCR and NN used for calibration in much of the examples discussed in this section all have a number of variables which need to be chosen for the calibration model. Part of this work will involve the investigation of WRR for calibration which has substantially fewer input variables, this should result in a more stable model and if applied, less between-user variation.

The application of WRR to data is the only non-widely recognised method used in this work. This was to give maximum confidence in the new measurement techniques developed and not have the type of data processing being controversial.

5.4 Chemometric software and calculations

All Chemometric techniques have been performed using Matlab™ and the PLS_Toolbox™ v2.1 (Eigenvector research, Inc., Manson, WA 1998) running under MatlabR11 or 12 (The MathWorks Ltd, Matrix House, Cowley Park, Cambridge, CB4 0HH). The PCs used for the data analyses were either a PIII 650 MHz with 258 RAM under win2000 or a PIV 1500 MHz with 512 RAM operating with Windows^{XP}.

Matlab is a programming language and data visualization tool. The programme works in a desk top environment similar to Windows. The following desktop windows were used in this work;

Command window:	Issues commands for Matlab processing, normally one line functions at a time.
Command history:	A running history of previous commands typed into the command window
Workspace window:	This is a GUI that is used to view, load or save files which are currently in use.

Matlab script files are used when performing a number of repetitive commands or calculations. Examples of these for calculation of correlation coefficient at a range of wavelengths and performing WRR are appendixed.

5.4.1 Procedures for PCA analysis

The PCA results are given as plots of the sample loadings for each spectral variable for each principal component, the sample scores for a range of principal component are plotted against each other. The percentage variance in the spectral data captured by each principal component is reported and the total percentage variance for a number of principal components is given.

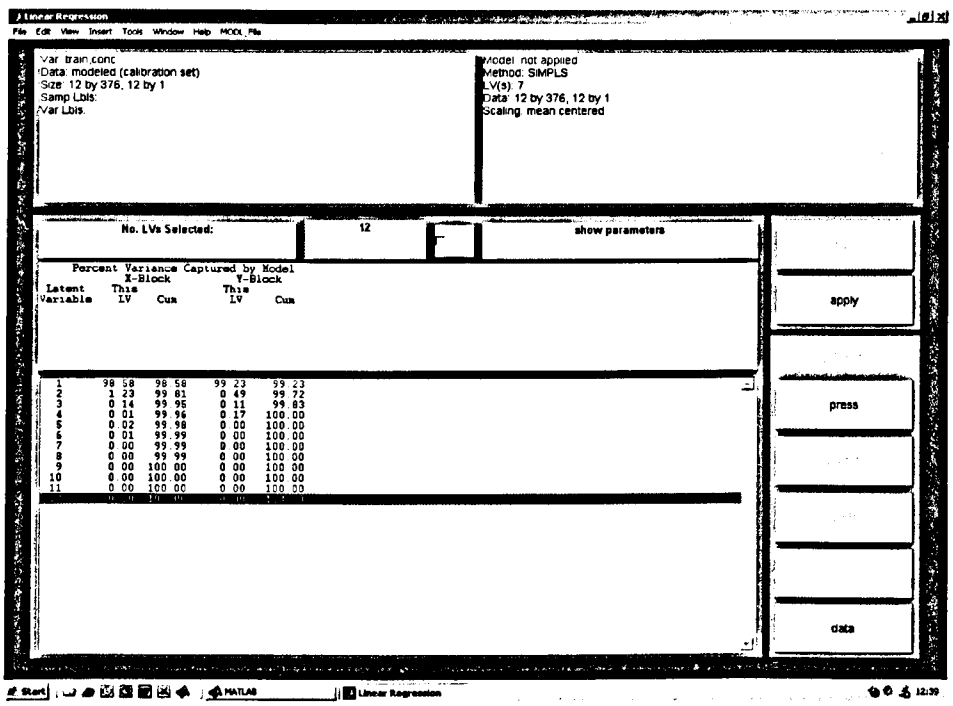
5.4.2 Procedures for PLS analysis

PLS is carried out by two methods. Both can be used for PLS1 type or PLS2 type. The first employs the 'MODLGUI' function of the PLS_Toolbox, a graphic user interface useful for initial calculations. For the second method script files are written to perform the calculations giving greater freedom and control over the types of calibration models produced with an expanded variation in the sample selection, pre-treatment and filtering methods.

5.4.2.1 Method 1: Employing a graphical user interface

The graphical user interface is launched by typing 'MODLGUI' at the Matlab command line. The spectra and reference information can be loaded then the function used to perform limited pre-treatment methods and set cross validation methods. The main output of use is a plot of the RMSEC and RMSECV for each set of reference information data at a number of latent variables. The screen for the MODLGUI can be seen in figure 5.4.2.1-1.

Figure 5.4.2.1-1 Screen for the PLS 'MODLGUI' function



5.4.2.2 Method 2: With Matlab command line

PLS_Toolbox files are called from the Matlab command line and are used to perform NIPALS PLS, polynomial PLS and Spline PLS. The sample spectra and reference data are loaded and any scaling or filtering to the data applied, and the

model calculated. The calibration model is then used this to predict the reference information of unknown spectra at a range of latent variables. The command line I/O (input/output) function for the various types of PLS used in this work is described in section 6.1.

6 Appendix

6.1 PLS_Toolbox Matlab functions

6.1.1 PLS NIPALS

Frame 6.1.1-1 Calculation of PLS NIPALS

I/O function: [b,ssq,p,q,w,t,u,bin] = pls(x,y,maxlv);	
Input information;	**Output information;**
x = training spectra	b = regression vectors
y = reference information	ssq = the fraction of variance used in the x and y matrices
maxlv = max no. of LVs	p = spectra loadings
	q = reference information loadings
	w = spectral weights
	t = spectral scores
	u = reference information scores
	bin = inner relation coefficients

Frame 6.1.1-2 Prediction based on PLS NIPALS

I/O function: [yprdn,resn,scoresn] = modlpred(newx,bin,p,q,w,lv,plots)	
	Output information;
	yprdn = predicted reference information
	resn = residuals
	scoresn = scores

6.1.2 Polynomial PLS

Frame 6.1.2-3 Calculation of poly-PLS

I/O function: [p,q,w,t,u,b,ssqdif] = polypls(x,y,lv,n);	
Input information;	**Output information;**
x = training spectra	p = spectra loadings
y = reference information	q = y-block loadings
lv = no. of LVs	w = x-block weights
n = no. inner relationships	t = x-block scores
	u = y-block scores
	b = spectra inner- relation coefficients
	ssqdif = variance in the data explained

Frame 6.1.2-4 Prediction of poly-PLS

I/O function: ypred = polypred(x,b,p,q,w,lv)	
	Output information; ypred = predicted reference information

6.1.3 Spline PLS

Frame 6.1.3-5 Calculation of Spline-PLS

I/O function: ; [P,W,T,U,C,cfs,ks,ssq] = spl_pls(x,y,knots,deg,lv,plots);	
Input information; x = training spectra y = reference information lv = no. of LVs knots = no. knots in the spline deg = degree of spline	**Output information;** P = spectra loadings W = x-block weights T = x-block scores U = y-block scores C = inner coefficients cfs = spline coefficients ls = knot locations ssq = variance captured by model

Frame 6.1.3-6 Prediction of Spline-PLS

I/O function: ypred = splspred(newx,P,W,C,cfs,ks,lvs,plots)	
	Output information; ypred = predicted reference information

6.1.4 OSC

Frame 6.1.4-7 Calculation of OSC spectra

I/O function: ; [nx,nw,np,nt] = osccalc(x,y,nocomp);	
Input information; x = training spectra y = reference information nocomp = no. OSC components	**Output information;** nx = OSC spectra nw = weights np = loadings nt = scores

PLS Nipals is the carried out on the OSC spectra.

Frame 6.1.4-8 Calculation of OSC to unknown spectra

I/O function: [newx] = oscapp(x,nw,np,nofact);	
	Output information; newx = OSC corrected new spectra

6.2 PLS_Toolbox scaling functions

Frame 6.2-1 Scaling methods

Pre-treatment	Function name	Input	Output
Median	med(x)	x = spectra	
Mean centre	mncn(x)	x = spectra	
Autoscaling	auto(x)	x = spectra	
SNV	auto(x')	x = spectra	

6.3 PLS_Toolbox Savitsky Golay filter

Frame 6.3-1 Savitsky Golay smoothing and derivatisation

I/O Function: [y_hat,cm] = savgol(y,width,order,deriv);	
Inputs	**Outputs**
x = spectra width = no. of points in filter order = polynomial order deriv = derivative order	[y_hat] = smoothed and differentiated matrix cm = matrix coefficients

6.4 Matlab script files

Frame 6.4-1 Calculation of correlation coefficient throughout spectra

```
% Calculation of correlation coefficient where no. variables = 2046

clear all
load spectra.asc
load conc.asc
for i = 1:2046;
   result = corrcoef(spectra(:,i),concA;
   finalresult(i) = result(2,1);
end
```

7 References

[1] http://mcec.engr.utk.edu/. Viewed: 13/1/03.

[2] http://www.cpac.washington.edu/. Viewed: 13/1/03.

[3] http://www.strath.ac.uk/Other/cpact/index.html. Viewed: 13/1/03.

[4] Hassell D. C and Bowman E. M, 'Process Analytical Chemistry for Spectroscopists', Applied Spectroscopy, 1998, 52 (1), 18A – 29A.

[5] McLennan F. M and Kowalski B. R, 'Process Analytical Chemistry', Blackie, Glasgow, UK, 1995, 91.

[6] http://www.cpact.com/Project7/project7sum.htm. Viewed: 28/5/2003.

[7] Chalmers J. M, 'Spectroscopy in Process Analysis', CRC Press, Sheffield, UK, 2000.

[8] Wright R.G. 'Monitoring Bioprocesses by On-line Process Mass Spectrometry', IFPAC 2002, San Diego, USA.

[9] Amphiah – Bonney R. J and Walmsley A. D, 'Monitoring of the Acid Catalysed Esterification of Ethanol by Acetic Acid Using Raman Spectroscopy', Analyst, 1999, 124, 1817 – 1821.

[10] http://www.cpac.washington.edu/NeSSI/NeSSI.htm. Viewed: 13/1/03.

[11] Montgomery D. C, 'Design and Analysis of Experiments', 5th Edition, Wiley, New York, 2001.

[12] Myers R.H and Montgomery D. C, 'Response Surface Methodology', Wiley, New York, 1995.

[13] Design-Expert 6, Users guide, Stat-Ease Inc., Minneapolis, 2000.

[14] http://www.umetrics.com/software_modde.asp?section=products. Viewed: 13/1/03.

[15] Miller J.N and Miller J.C, 'Statistics and Chemometrics for Analytical Chemistry', 4th Edition, Prentice Hall, Harlow, UK, 2000, 183.

[16] Miller J.N and Miller J.C, 'Statistics and Chemometrics for Analytical Chemistry', 4th Edition, Prentice Hall, Harlow, UK, 2000, 54.

[17] Miller J.N and Miller J.C, 'Statistics and Chemometrics for Analytical Chemistry', 4th Edition, Prentice Hall, Harlow, UK, 2000, 55.

[18] Miller J.N and Miller J.C, 'Statistics and Chemometrics for Analytical Chemistry', 4th Edition, Prentice Hall, Harlow, UK, 2000, 57 - 64.

[19] http://davidmlane.com/hyperstat/A42408.html. Viewed: 28/2/2003.

[20] Lavine BK, Workman J, 'Chemometrics', Analytical Chemistry, 2002, 74 (12), 2763R - 2769R.

[21] Lavine BK, 'Chemometrics', Analytical Chemistry, 2000, 72 (12), 91R – 97R.

[22] Lavine BK, 'Chemometrics', Analytical Chemistry, 1998, 70 (12), 209R – 228R.

[23] Brown SD, Sum ST, Despagne F, Lavine BK, 'Chemometrics' Analytical Chemistry, 1996, 68 (12), 21R – 61R.

[24] Martens H, Naes T, 'Multivariate Calibration', Wiley, Guildford, UK, 1998.

[25] Brereton R.G, 'Introduction to Multivariate Calibration in Analytical Chemistry', Analyst, 2000, 125 (11), 2125 – 2154.

[26] Wold S, Trygg J, Berglund A and Antti H, 'Some Recent Developments In PLS Modeling', Chemometrics and Intelligent Laboratory Systems, 2001, 58, 131 – 150.

[27] http://www.cpac.washington.edu/interlink/4-03-interlink.htm. The CPAC InterLink: The Center for Process Analytical Chemistry Internal Newsletter Vol. 4:3, October 2002. Viewed: 13/1/02.

[28] http://www.chemweb.com/alchem/articles/985883672250.html. Viewed 20/1/03.

[29] Miller C.E, 'Chemometrics for On-line Spectroscopy Applications – Theory and Practice' Journal of Chemometrics, 2000, 14, 513 – 528.

[30] Wise B.M and Gallagher N. B, 'PLS_Toolbox V2.0 for use with Matlab™', User Manual, Eigenvector Research Ltd, Manson, WA, USA, 1998, 32 - 34.

[31] Faber N.M, 'Estimating the Uncertainty in Estimates of Root Mean Square Error of Prediction: Application to Determining the Size of an Adequate Test Set in Multivariate Calibration', Chemometrics and Intelligent Laboratory Systems, 1999, 49, 79 – 89.

[32] Walmsley A.D, 'Improved Variable Selection Procedure for Multivariate Linear Regression', Analytica Chimica Acta, 1997, 354, 225 – 232.

[33] Wise B.M and Gallagher N.B, 'PLS_Toolbox 2.0 for use with Matlab™', User Manual, Eigenvector Research Ltd, Manson, WA, USA, 1998.

[34] Faber K, Kowalski B.R, 'Propagation of Measurement Errors for the Validation of Predictions Obtained by Principal Component Regression and Partial Least Squares', Journal of Chemometrics, 1997, 11, 181 – 238.

[35] Geladi P and Kowalski B. R, 'Partial least squares regression a tutorial', Analytica Chimica Acta, 1986, 185, 1 – 17.

[36] Wise B.M and Gallagher N.B, 'PLS_Toolbox 2.0 for use with Matlab™', User Manual, Eigenvector Research Ltd, Manson, WA, USA, 1998, 81 - 84.

[37] Massart D.L, Vandeginste B.G.M, Buydens L.M.C, DeJong S, Lewi P.J and Smeyers-Verbeke J, 'Handbook of Chemometrics and Qualimetrics: Part A', 1997, Elsevier, Amsterdam, 322.

[38] Wold S, 'Nonlinear Partial Least Squares Modelling II, Spline Inner Relation', Chemometrics and Intelligent Laboratory Systems, 1992, 14, 71 – 84.

[39] Frank I. E, Tutorial: 'Modern Nonlinear Regression Methods', Chemometrics and Intelligent Laboratory Systems, 1995, 27, 1 – 9.

[40] Wold S, Antti H, Lindgren F, Ohman J, 'Orthogonal Signal Correction of Near Infrared Spectra', Chemometrics and Intelligent Laboratory Systems, 1998, 44, 175 – 185.

[41] Fearn T, 'On Orthogonal Signal Correction', Chemometrics and Intelligent Laboratory Systems, 2000, 50, 47 – 52.

[42] Feudale R.N, Huwei Tan, Brown S.D, 'Piecewise Orthogonal Signal Correction', Chemometrics and Intelligent Laboratory Systems, 2002, 36, 129 – 138.

[43] Draper N.R and Smith H, 'Applied Regression Analysis', 3rd edn., Wiley, New York, 1998,ch. 17, 387-396.

[44] Vandeginste B.G.M, Massart D.L, Buydens L.M.C, DeJong S, Lewi P.J and Smeyers-Verbeke J, 'Handbook of Chemometrics and Qualimetrics: Part B', Elsevier, Amsterdam, 1997, 369.

[45] Miller J.N and Miller J.C, 'Statistics and Chemometrics for Analytical Chemistry', Fourth Edition, Prentice Hall, Harlow, UK, 2000, 111.

[46] Seasholtz M.B and Kowalski B.R, 'The Effect of Mean Centering on Prediction in Multivariate Calibration', Journal of Chemometrics, 1992, 6, 103 – 111.

[47] Malinowski E.R, 'Factor Analysis in Chemistry', Second Edtion, Wiley, New York, USA, 1991, 41.

[48] Savitsky A, Golay M.J.E, 'Smoothing and Differentiation of Data by Simplified Least Squares Procedures', Analytical Chemistry, 1964, 36 (8), 1627 – 1638.

[49] Roggo Y, Duponchel L, Noe B, Huvenne JP, 'Sucrose Content Determination of Sugar Beets by Near Infrared Reflectance Spectroscopy. Comparison of calibration methods and calibration transfer', Journal of Near Infrared Spectroscopy, 2002, 10, 137-150.

[50] Smola N and Urleb U, 'Qualitative and Quantitative Analysis of Oxytetracycline by Near Infra Red Spectroscopy', Analytica Chimica Acta, 2000, (410) 203 – 210.

[51] Stordrange L, Libnau F. O, Malthe-Sorenssen D, Kvalheim O. M, 'Feasibility Study of NIR for Surveillance of a Pharmaceutical Process, Including a Study of Different Preprocessing Techniques', Journal of Chemometrics, 2002, 16, 529 – 541.

Chapter 2

Development of a Reference Dataset Based on Visible Spectra of Metal Complexes

1 Introduction

This work was for the European Framework 5, 'The Standards Measurements and Testing' programme of the European Network Consortium. The aim was to produce a spectroscopic reference dataset with an absolute minimum of experimental and hence spectral errors. The dataset will then be used for a 'round robin study' to act as an inter-laboratory comparison, but for chemometric data analysis instead of chemical analysis. The bench mark dataset will then be universally available for downloading from a website. This could then be used in the development of new applications and algorithm design. In addition to this, an investigation into the effect of measurement based on two different types of spectrometer will be undertaken. Samples will be measured using a standard analytical spectrometer and industrial spectrometer designed for use on industrial process plants.

The use of inter-laboratory studies for chemical analysis, to ensure that there is no bias in results between different laboratories, is now commonplace. Standard reference materials, with known composition and confidence intervals, are available from most chemical suppliers. The aim is to extend this principle to results obtained by different users for multivariate calibration of the same data.

Reference datasets have already been produced by NIST, with the objective that: 'The purpose of this project is to improve the accuracy of statistical software by providing reference datasets with certified computational results that enable the objective evaluation of statistical software'[1].

There are several datasets which can be downloaded from the NIST website[2], each of these have certified mean and standard deviation estimates for the results of linear regression and non-linear regression. The datasets are not spectroscopic and the certification is based on a pre-determined mathematical method of analysis.

The European framework has carried out a similar study to that pursued here; they have published a discussion paper of the preliminary results[3]. This study was based on NIR spectra of natural forage samples collected over 3 years, which were used to predict the levels of moisture and crude protein content. The study found that for the six participants the RMSEPs were acceptable but the actual predictions varied considerably between them. During the development of a reference dataset based on visible metals spectra a larger number of subjects will participate. The dataset produced will follow an experimental design strategy and will be measured in one day to minimise variation in the spectra from instrumental drift or environment conditions.

For the study, a set of accurate, reproducible spectra was required. Metals complexes following a design of experiments strategy were chosen to be analysed by visible spectroscopy. Solutions of cobalt, nickel and copper salts are traditionally used as standard solutions for UV/Vis spectroscopy[4]. The French Standards organisation 'Laboratoire National d'Essais' (LNE) use nitrates of these metals. The four transition metals ions chosen were cobalt (Co^{2+}), chromium (Cr^{3+}), copper (Cu^{2+}) and nickel (Ni^{2+}). Each has a different absorption spectrum in the visible region of the electromagnetic spectrum with maximum absorption at

different wavelengths[5]. The metal ion solutions were in nitric acid to prevent oxidation.

The solutions were prepared in nitric acid to due good stability and minimum interference in the spectra samples spectra in the visible region.

Three main datasets were produced. The first was the pre-study that consisted of spectra simulated from those of the pure metals spectra; the second was recorded on a low-resolution spectrometer that was designed for industrial process monitoring. The third was recorded using a standard analytical spectrometer. For both of the second and third datasets duplicate samples sets were produced where the entire procedure was carried out on two consecutive days.

Initial investigations (PCA and PLS) were performed on all the datasets. The dataset which did not highlight any spurious samples after PCA, and with minimal errors of prediction after PLS regression, was chosen for use in the study. The subjects were provided with two predefined datasets. The first was to be used for training of the regression model, using this model, the subjects were to predict the volume of metal ions in a second dataset of test samples. The trial consisted of 20 participants with a range of experience in multivariate calibration. The instruction given to the participants was to produce the best calibration model in their opinion from the training samples and then to use this to predict the test samples. The user's selection of software, algorithm and sample pre-treatment methods were

unrestricted. The predicted values were reported along with the model details and comments of outlier samples.

1.1 Beers law

For UV/Vis spectroscopy the absorbance spectrum is related to sample concentration using Beers law[6], equation 13. When considering a mixture of multiple absorbing analytes the absorbance at a given wavelength is the sum of the absorbance of each analyte (equation 14). Based on this, the absorbance spectrum of a mixture can be estimated using the absorbance spectrum of each analyte at known concentration and pathlength.

$$A = \varepsilon bc$$ Equation 13 Beers law

$$A=\varepsilon_1 bc_1+\varepsilon_2 bc_2+\varepsilon_3 bc_3+\varepsilon_4 bc_4+.....$$ Equation 14 Application of Beers law to mixtures

Where;

A = absorbance

ε = molar absorptivity ($cm^{-1}mol^{-1}l$)

b = pathlength of radiation (cm)

c = concentration of absorbing analyte (mol/l)

2 Experimental

2.1 Initial measurements of 1000 µg/ml metal ions

The UV/Vis spectra of nitric acid, 5 % (v/v), and solutions of cobalt, chromium, copper and nickel (1000 µg/ml) in nitric acid 4 % (v/v) were recorded. At the maximum absorbance for each metal ion, the wavelength and amount of absorbance was noted.

2.2 Experimental design

The maximum absorbance of the 1000 µg/ml metals ions was less than 0.3 for each of the components. 10,000 µg/ml solutions are used for the actual samples for higher absorbances. The samples were constrained so that the sum of the absorbance was less than 1 absorbance. A full factorial two level design was used to calculate the concentrations of a set of training samples in table 2.1-1. The test sample concentrations (table 2.1-2) were randomly generated within the constrained concentration limits of the experimental design. In addition to this, for samples 23 – 26 at least one of the metals was excluded from the sample, the remaining concentrations were randomly generated within the limits of the designs concentration range.

Table 2.1-1 Table of training samples for analysis by visible spectroscopy, metal ion composition: Volume (ml) of metal ion standard, 10,000 µg/l, made up to 10 ml with nitric acid, 5 % (v/v)

Sample number	Co^{2+}	Cr^{3+}	Cu^{2+}	Ni^{2+}
1	3.2	1	1.2	2.4
2	4	1	1.2	2.4
3	3.2	1.4	1.8	2.4
4	4	1.4	1.8	2.4
5	3.2	1	1.2	2.4
6	4	1	1.2	2.4
7	3.2	1.4	1.8	2.4
8	4	1.4	1.8	2.4
9	3.2	1	1.2	2.8
10	4	1	1.2	2.8
11	3.2	1.4	1.2	2.8
12	4	1.4	1.2	2.8
13	3.2	1	1.8	2.8
14	4	1	1.8	2.8
15	3.2	1.4	1.8	2.8
16	4	1.4	1.8	2.8

Table 2.1-2 Table of test samples for analysis by visible spectroscopy, metal ion composition: Volume (ml) of metal ion standard, 10,000 µg/l, made up to 10 ml with nitric acid, 5 % (v/v)

Sample number	Co^{2+}	Cr^{3+}	Cu^{2+}	Ni^{2+}
17	3.2	1	1.7	2.6
18	3.4	1.4	1.6	2.8
19	3.7	1.1	1.7	2.7
20	3.2	1	1.3	2.7
21	3.7	1.2	1.8	2.5
22	3.8	1.3	1.4	2.4
23	0	1.1	1.3	0
24	3.8	0	1.6	2.6
25	3.8	1.3	0	2.7
26	0	1.4	1.5	0

2.3 Preparation of solutions

The solutions were prepared in 'A' grade 10 ml volumetric flasks, these were washed and rinsed with 5 % (v/v) nitric acid prior to preparation.

2.3.1 Initial measurements of 1000 µg/ml metal ions

The spectra of 1000 µg/ml of Co^{2+}, Cu^{2+}, Cr^{3+} and Ni^{2+} single element standards from QMx laboratories ltd, Thaxted, CM6 2PY, UK were recorded.

2.3.2 Sample measurement based on 10,000 µg/ml metal ions

The metal solutions were added to the flasks in a random order. For the 26 samples and 4 different metal types there were 104 additions to the flasks. These additions, for each sample, were numbered 1-26 for cobalt, 27 – 52 for chromium, 53 - 78 copper and 79 – 104 for nickel. The numbers 1-104 were randomised and the resulting order used for the sample preparation.

Reagents

For the two experiments measured on different spectrometers, two sets of reagents were purchased from the same batches from QMx.

Table 2.3.2-1 Table of reagents used for visible metals spectra

Reagent
Nitric acid reagent blank (500 ml), HNO_3, 5 % (v/v)
Single element cobalt std (250 ml), Co^{2+} in 4 % HNO_3, 10,000 ± 30 µg/ml
Single element chromium std (250 ml), Cr^{3+} in 4 % HNO_3, 10,000 ± 30 µg/ml
Single element copper std (250 ml), Cu^{2+} in 4 % HNO_3, 10,000 ± 30 µg/ml
Single element nickel std (250 ml), Ni^{2+} in 4 % HNO_3, 10,000 ± 30 µg/ml

2.4 Sample analysis

The samples were analysed in a random order. The duplicate sets for each spectrometer were prepared and measured on consecutive days.

2.4.1 Analysis by an analytical spectrometer

Spectrometer:	Perkin Elmer UV/Vis lambda Bio 10*
Control PC:	Pentium (I) 166 MHz processor, 16 MB RAM
Operating software:	UV WinLab v2.80.03
Scan Range:	1000 – 350 nm (900 – 350nm for pre-study)
Scan Speed:	960 nm/min
Resolution:	0.5 nm
Smoothing:	None
Number of data points:	1300 (1100 for pre-study)

* Chalfont Road, Seer Green, Beaconsfield, Bucks, HP9 2FX, UK.

2.4.2 Analysis by a process type spectrometer

Spectrometer:	Carl Zeiss double beam UV/Vis spectrometer*
System:	MCS 500
Control PC:	Pentium (II) 300 MHz processor, 64 MB RAM
Operating software:	Aspect plus
Cycle mode:	Single
Scan Range:	738.3614 – 349.8884 nm
Resolution:	2.1167 nm

Integration time:	350.0 ms
Number of flashes:	7
Accumulation:	50
Number of data points:	184

* Clairet Scientific, 17 Scirocco Close, Moulton Park Industrial Estate, Northampton, NN3 6AP, UK.

3 Results

3.1 Pre-study

The spectra of the pure metal ions of Co^{2+}, Cr^{3+}, Cu^{2+} and Ni^{2+} in nitric acid are plotted in figure 3.1-1. The absorbance spectra of each of the metals are distinctly different. When overlaid in figure 3.1-2, regions where the maximum absorbance has the potential to be masked by other components are apparent.

Figure 3.1-1 Visible spectra of pure metal ions (1000 µg/l)

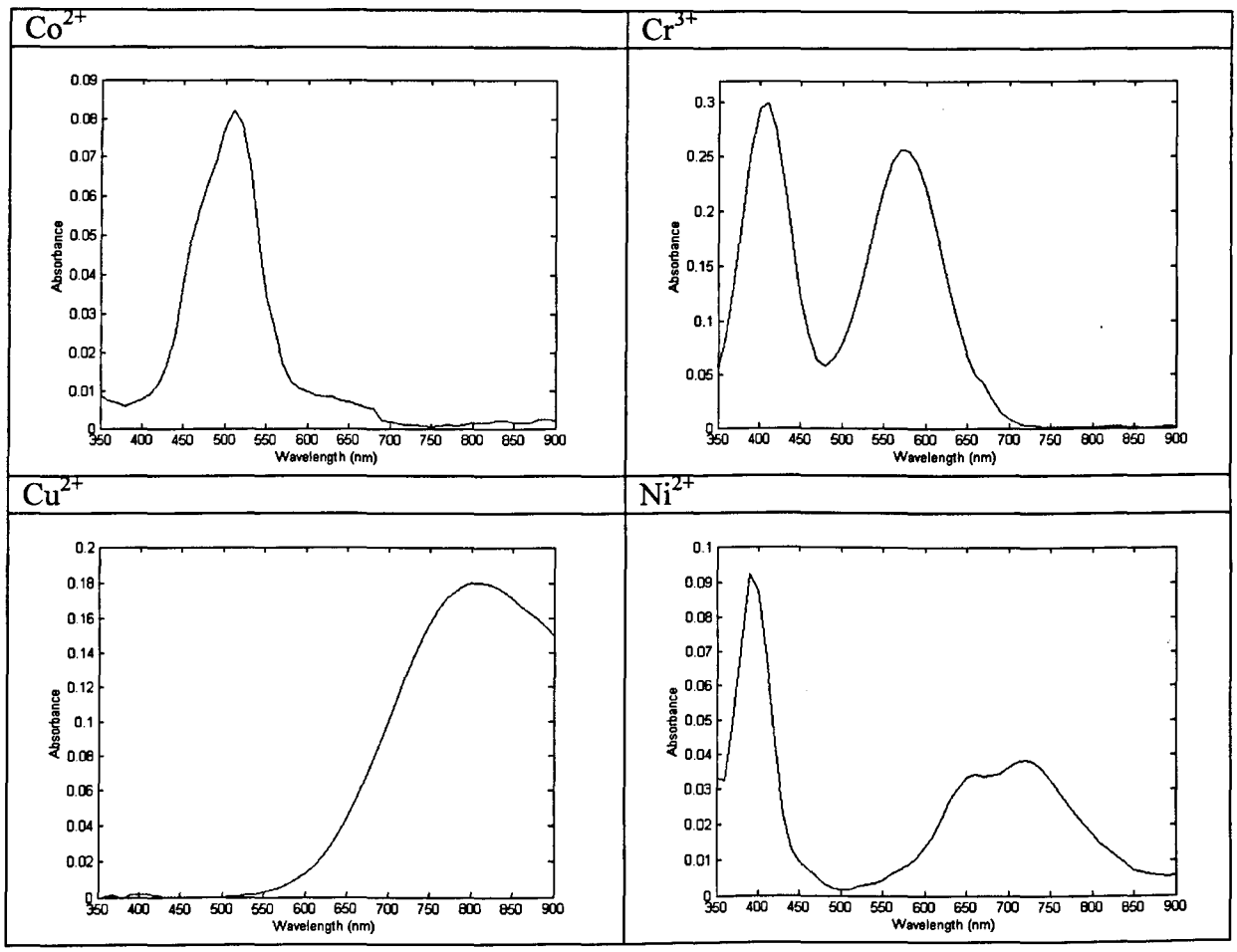
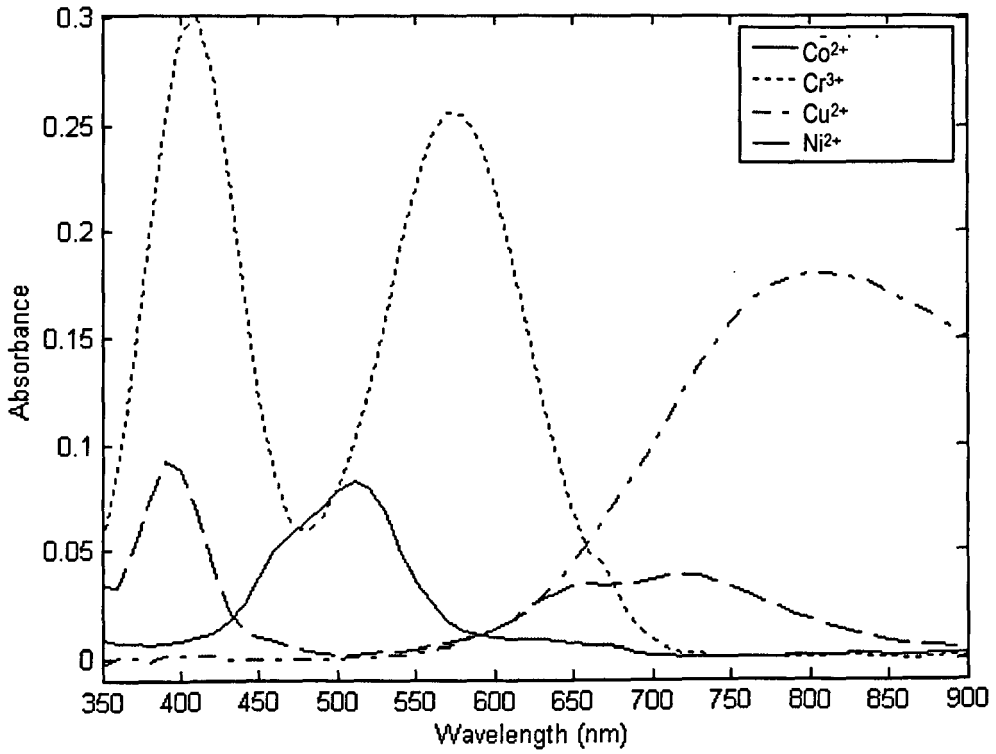


Figure 3.1-2 Visible spectra of 1000 µg/l metal ions



From the pure spectra an experimental design was produced so that the maximum sum of all the metals absorbance was not greater than 0.8 to ensure that the actual sample used in the trials were within the linear range for visible spectra. An additional test set of 10 samples was included for validation and testing of future models. The test samples were made up of random concentrations within the limits of the design, in some cases, deliberately containing only 2 or 3 of the metals components, to investigate if some users or software in the study would consider these outliers.

The visible spectra of the metal ions solutions have been plotted in two groups, figure 3.1-3, those including and those excluding samples at zero concentration. Visually there is a clear difference between the two sets.

Figure 3.1-3 Visible spectra of metal ion solutions; samples with all metal ions present

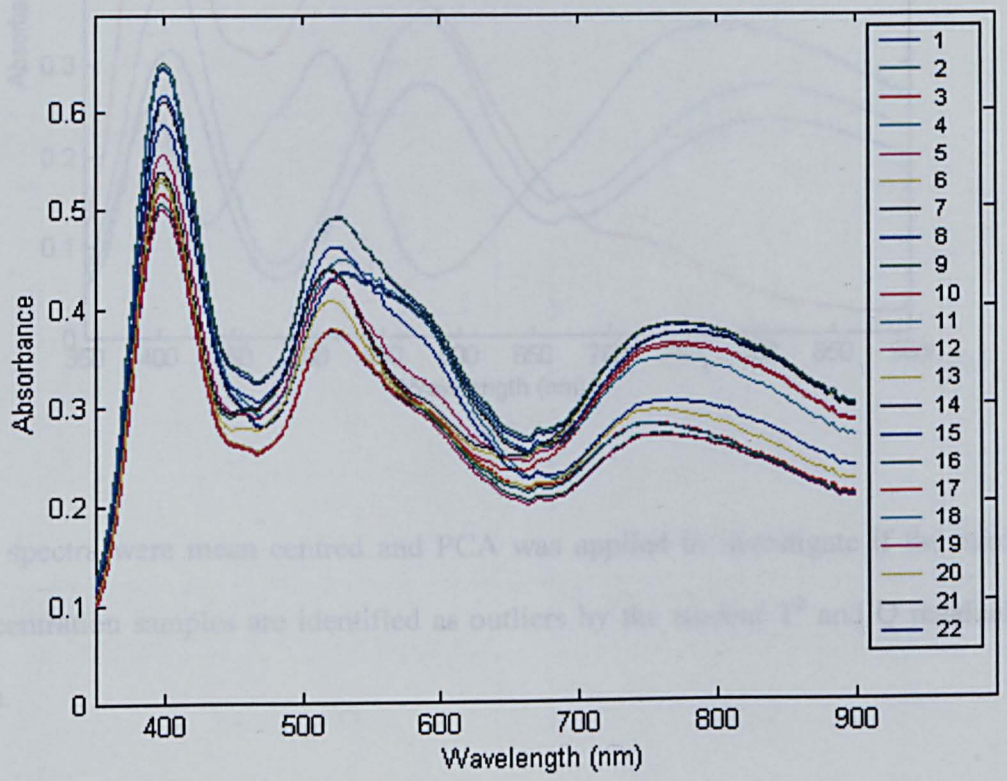
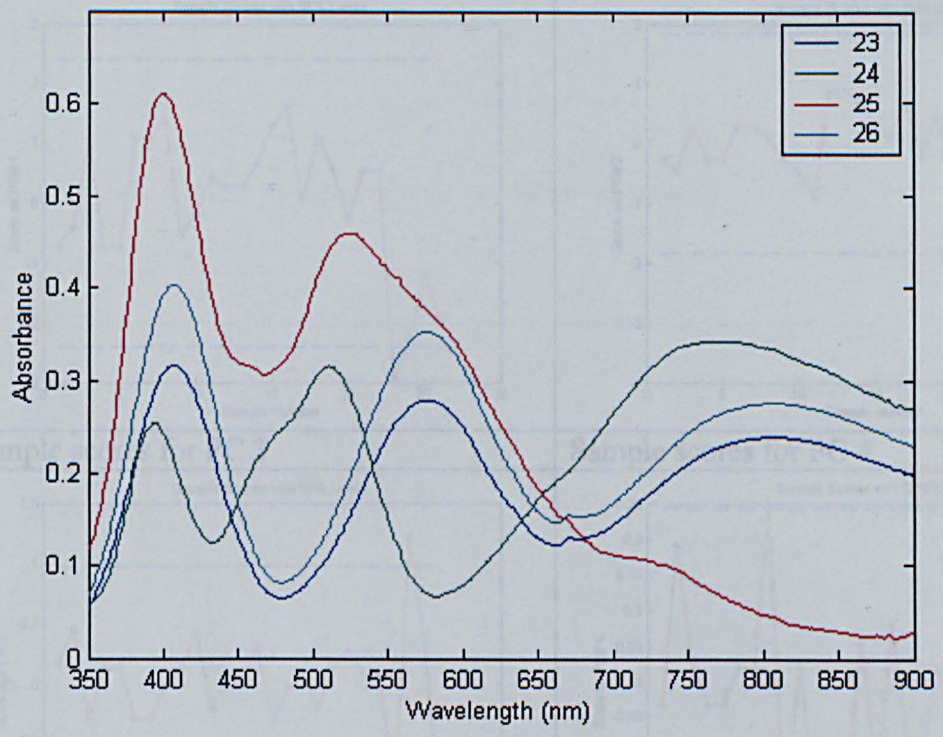


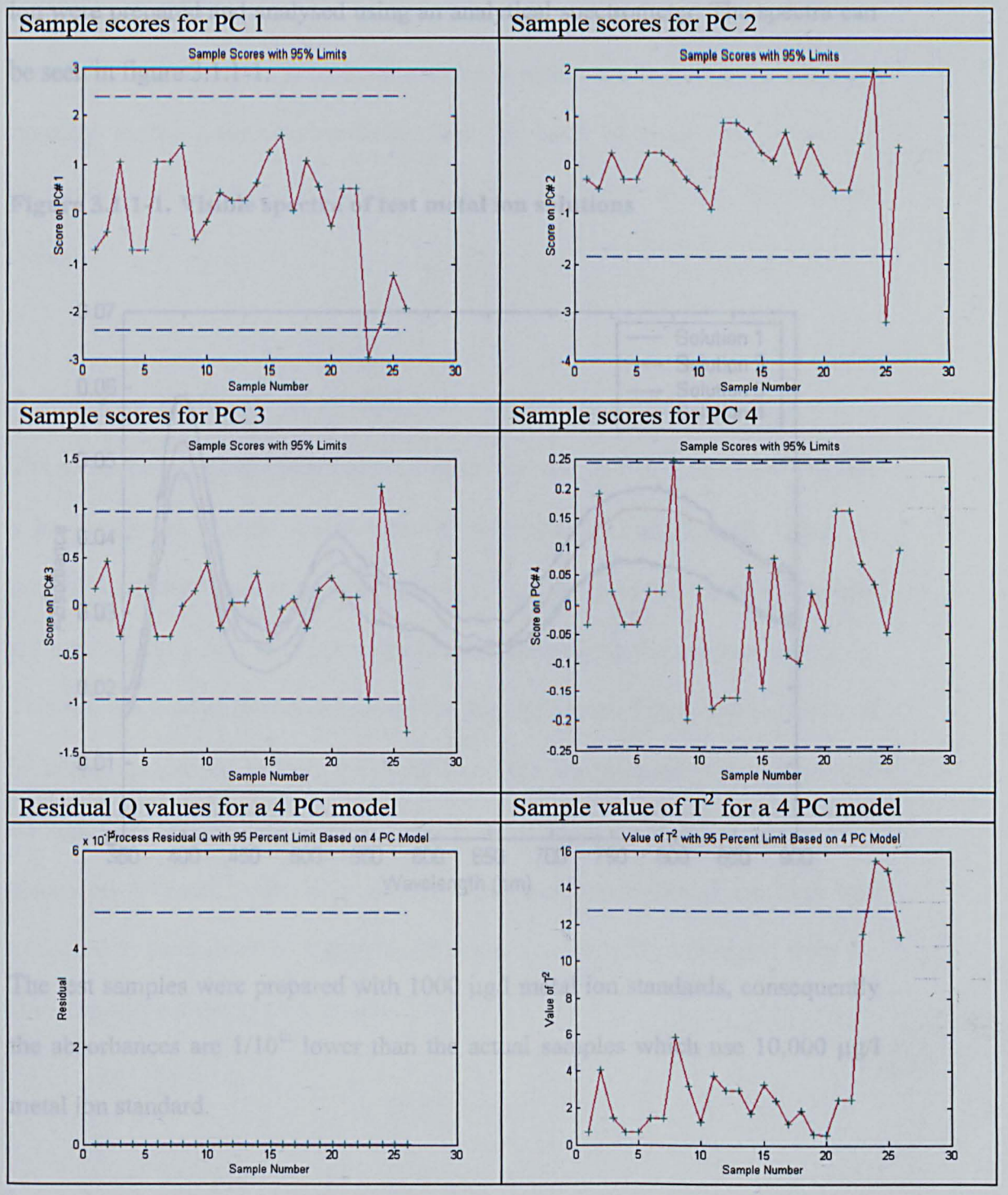
Figure 3.1-3 Visible spectra of metal ion solutions; samples with zero concentration for some metal ions



The spectra were mean centred and PCA was applied to investigate if the zero concentration samples are identified as outliers by the student T^2 and Q residual tests.

The samples scores plots, in figure 3.1-4, show that those containing all four components are within the 95 % limit of the model. The samples with one or more missing components, numbers 22 – 26, are often outside the 95 % limit. These could be mistaken for outliers and removed from the dataset when in fact they are part of the experimental design.

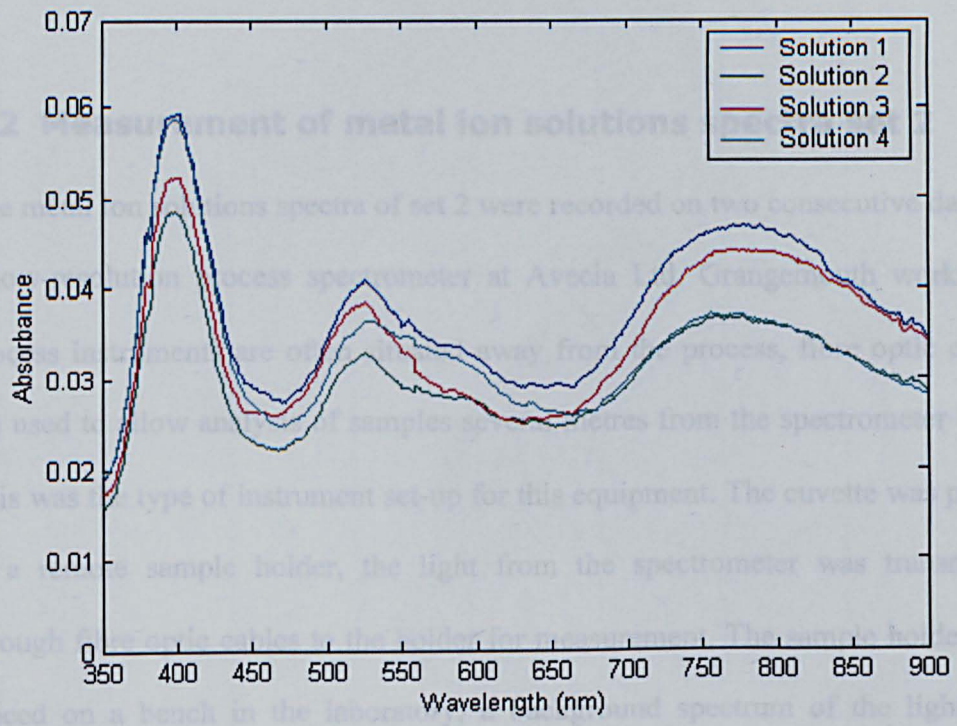
Figure 3.1-4 PCA samples scores plots, residual Q values and T^2 value results for the simulated metal ion solutions spectra based on a 4 PC model



3.1.1 Measurement of four test metal ion solutions spectra

Four samples including those of lowest and highest concentration of each metal ion were prepared and analysed using an analytical spectrometer. The spectra can be seen in figure 3.1.1-1.

Figure 3.1.1-1. Visible spectra of test metal ion solutions



The test samples were prepared with 1000 µg/l metal ion standards, consequently the absorbances are $1/10^{th}$ lower than the actual samples which use 10,000 µg/l metal ion standard.

The correlation between the actual test samples measured and the simulated spectra in dataset 1 was calculated to be above 0.998 for all four samples

measured. This was a good indication that the measured spectra will be similar to those simulated from Beers law. The lowest and highest sample absorbances were in the range of 0.18 and 0.06 absorbance which when multiplied by 10 to correct for the low concentration is 0.18 – 0.6 which is within the desired range to ensure linearity in the samples absorbance. On the basis of these results, sample preparation and analysis continued without amendment to the experimental design.

3.2 Measurement of metal ion solutions spectra set 2

The metal ion solutions spectra of set 2 were recorded on two consecutive days on a low-resolution process spectrometer at Avecia Ltd. Grangemouth works. As process instruments are often situated away from the process, fibre optic cables are used to allow analysis of samples several metres from the spectrometer body. This was the type of instrument set-up for this equipment. The cuvette was placed in a remote sample holder, the light from the spectrometer was transmitted through fibre optic cables to the holder for measurement. The sample holder was placed on a bench in the laboratory, a background spectrum of the light was measured at the start of each day, and this was automatically subtracted from the absorbance spectrum of each sample.

Figure 3.2-2 Metal ion solutions of dataset 2.2, visible spectra measured with an industrial process spectrometer

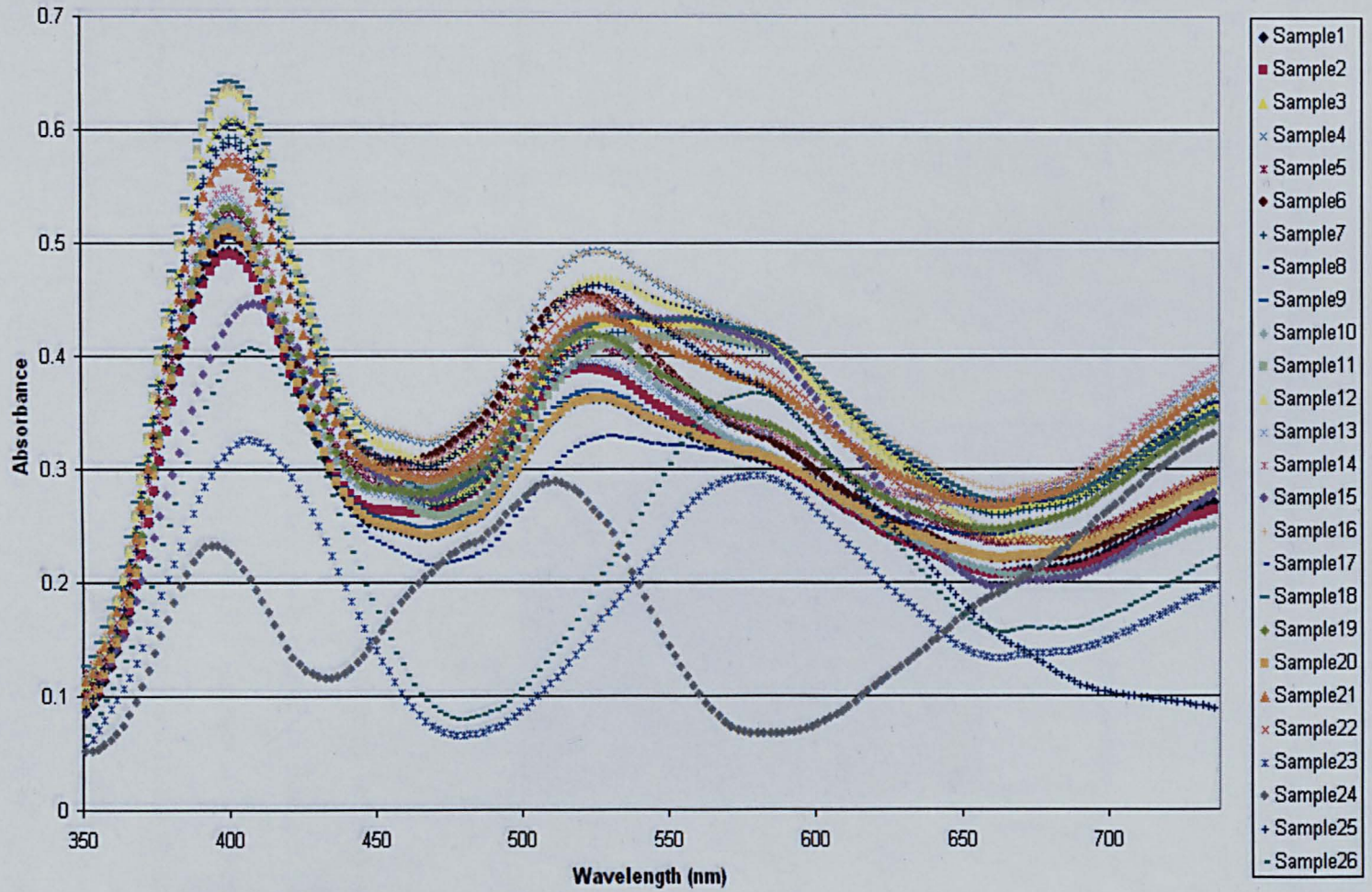
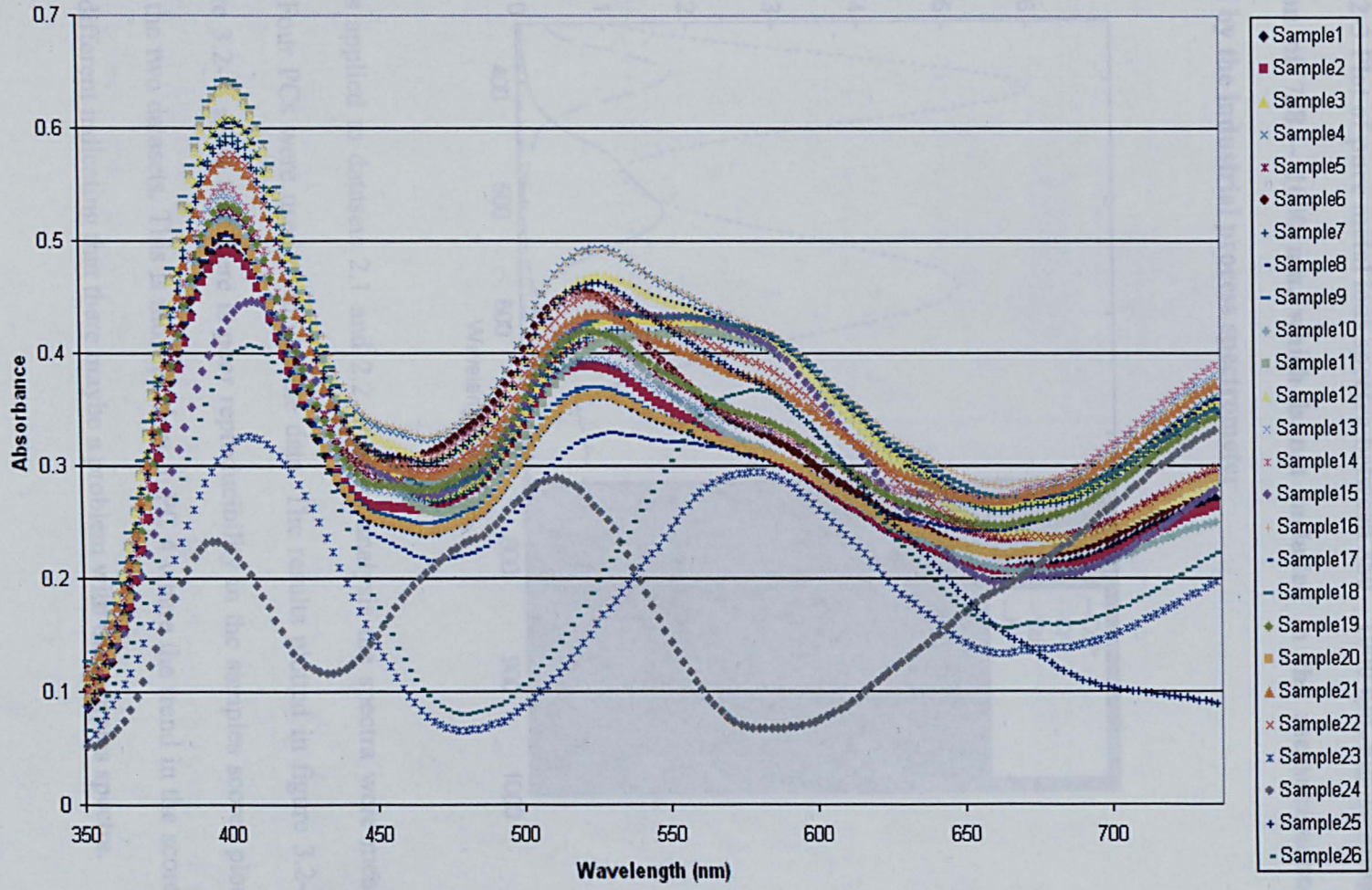
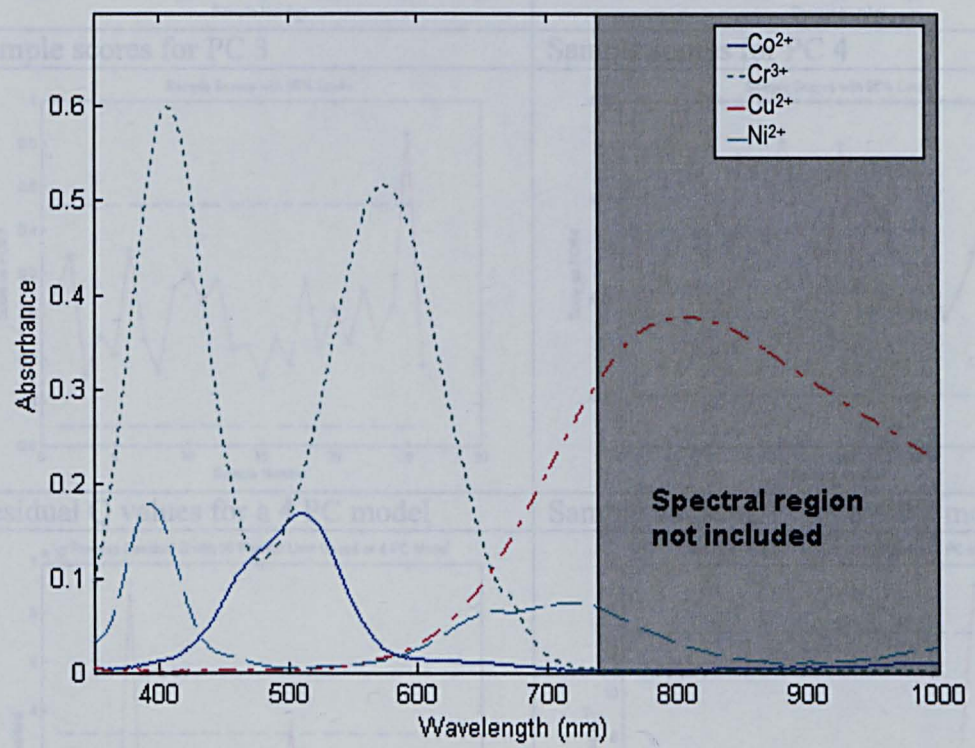


Figure 3.2-2 Metal ion solutions of dataset 2.2, visible spectra measured with an industrial process spectrometer



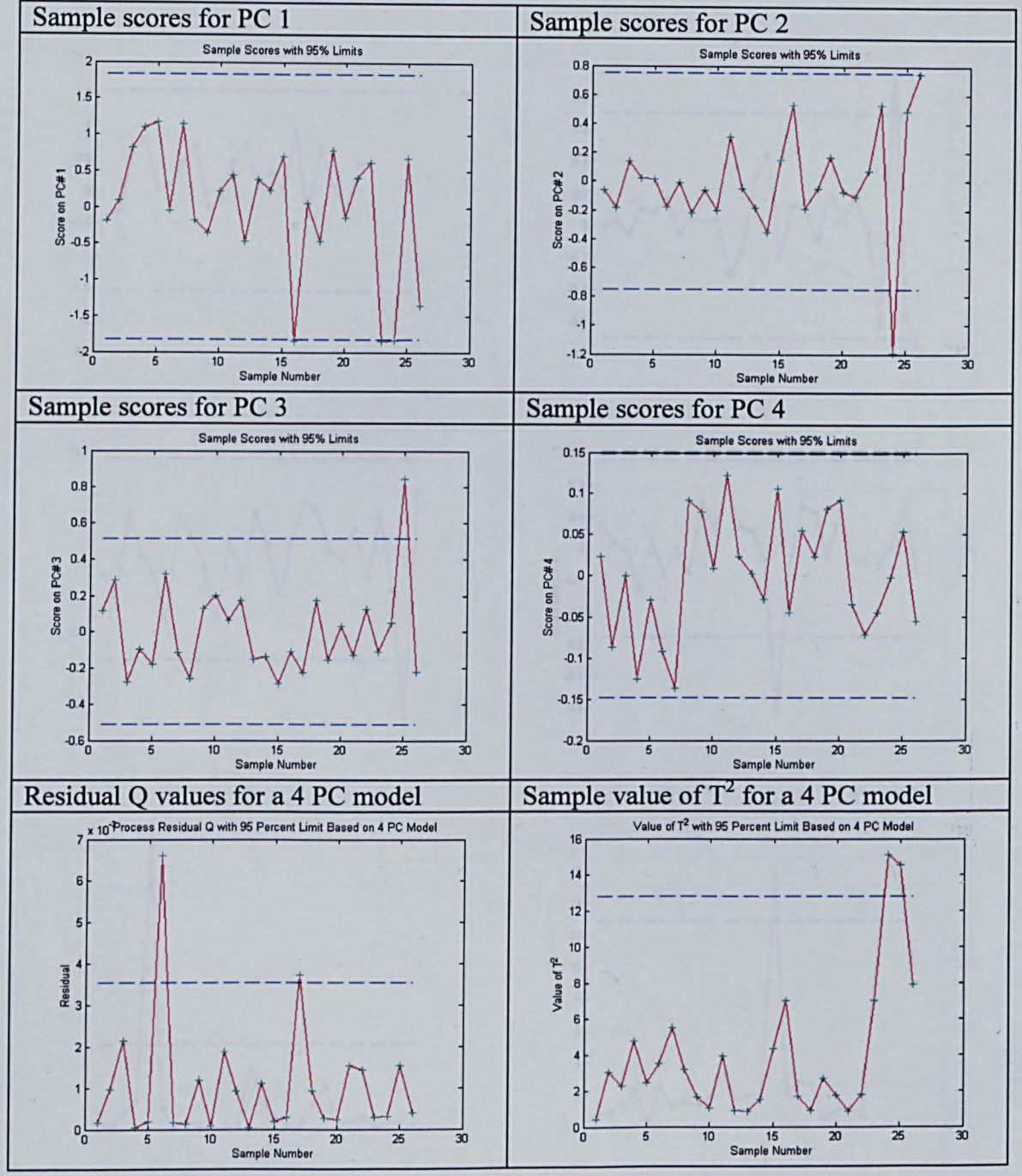
The industrial spectrometer only measured spectra to 738 nm, and not to 1000 nm. This significantly affects the samples' spectra. The maximum absorbance for Cu^{2+} is at 813 nm. In addition to this, Ni^{2+} absorbs in this region.

Figure 3.2-3 Plot of pure metal ions spectra between 350 – 1000 nm, highlighting the region of 738 – 1000 nm, which is not included in the measurements recorded by the industrial process spectrometer



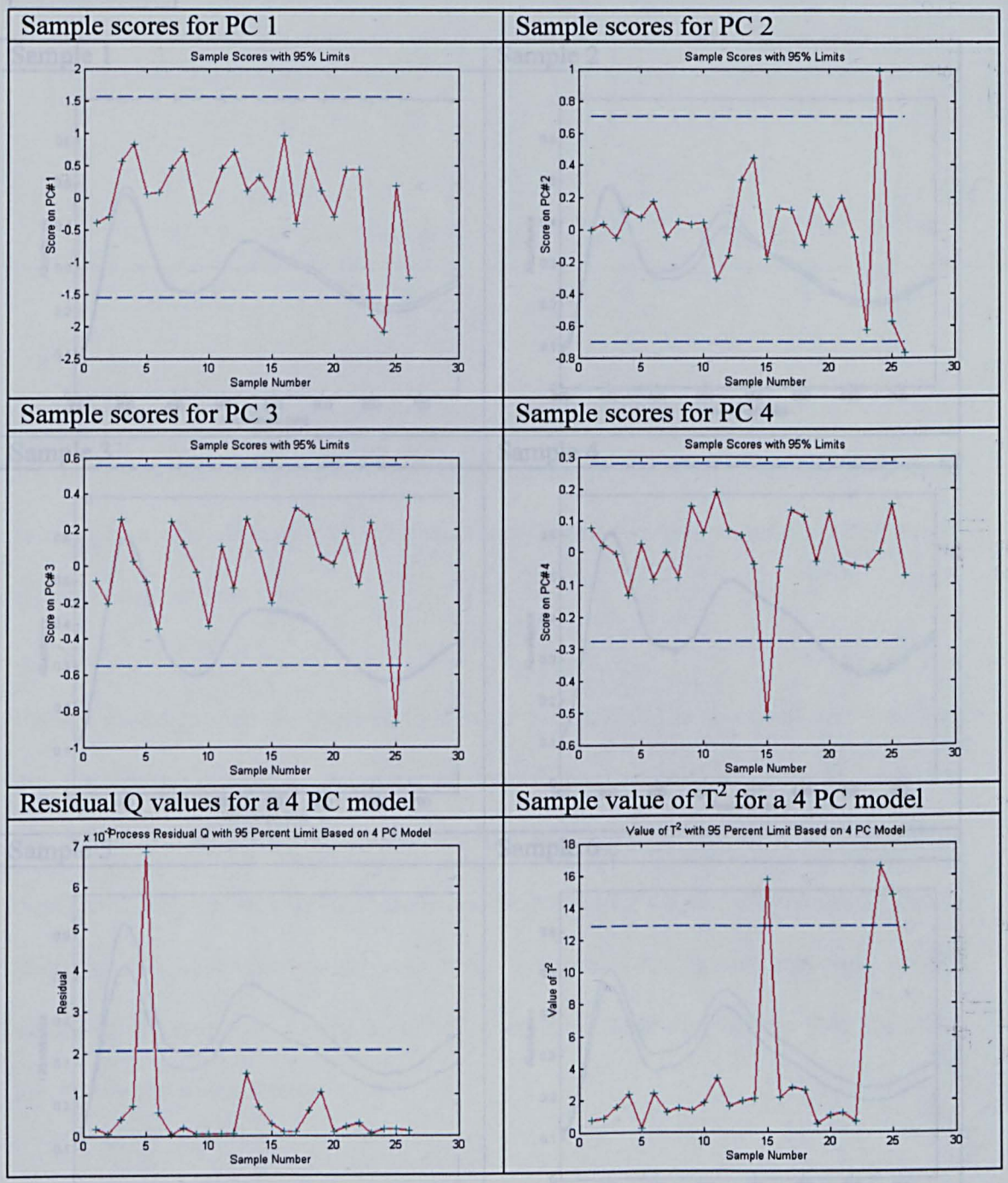
PCA was applied to datasets 2.1 and 2.2. Prior to analysis the spectra were mean centred. Four PCs were used to model the data. The results plotted in figure 3.2-4 and figure 3.2-5, show that there is poor reproducibility in the samples scores plots between the two datasets. This is most evident in PC 4 where the trend in the scores are very different indicating that there maybe a problem with the samples spectra.

Figure 3.2-4 Dataset 2.1 PCA results; samples scores plots, residual Q values and T^2 value results based on a 4 PC model



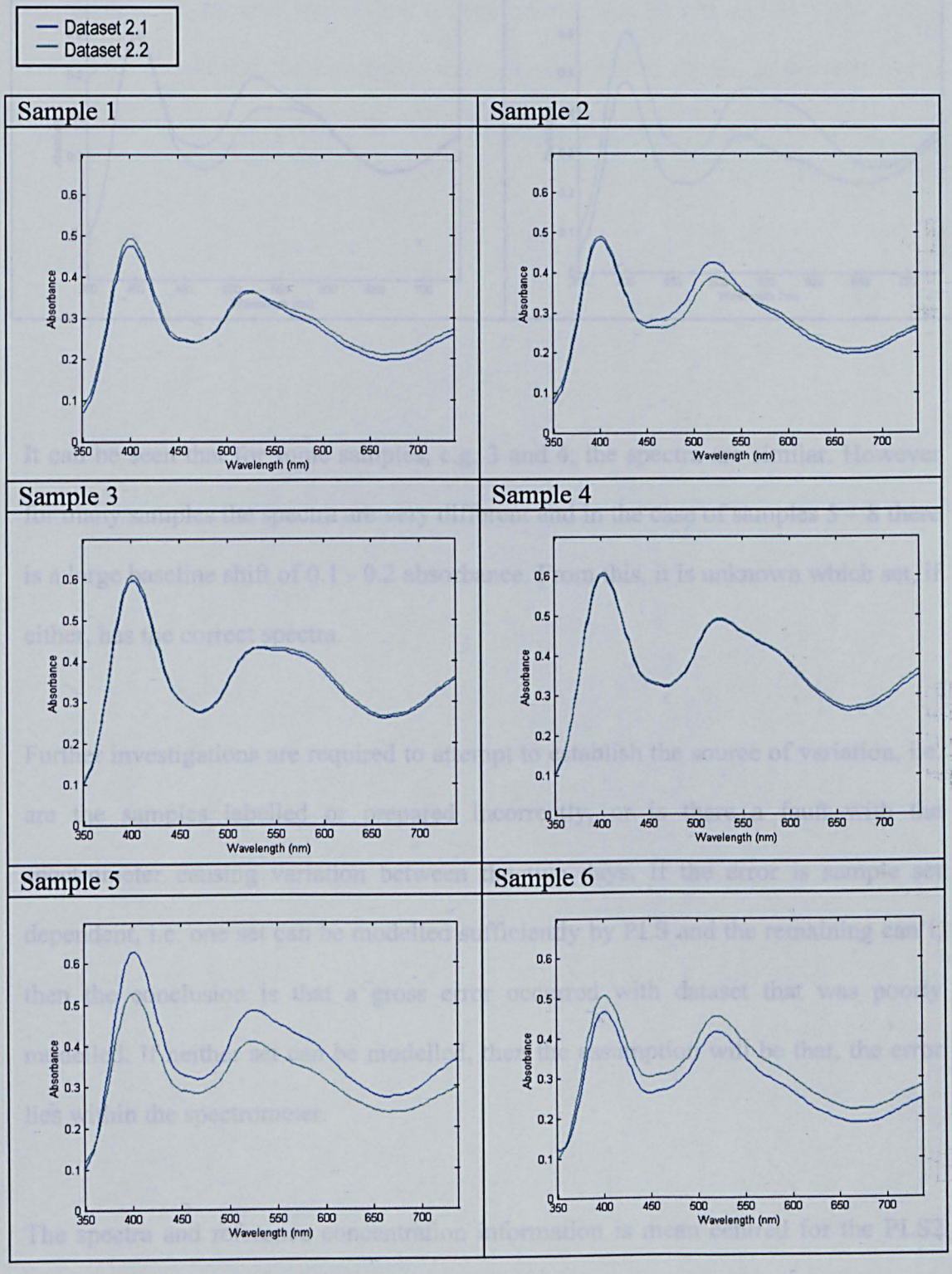
The T^2 test highlights samples 23 – 26, i.e., those missing at least one metal ion, to be of high leverage and possible outliers. This is also seen in dataset 2.2.

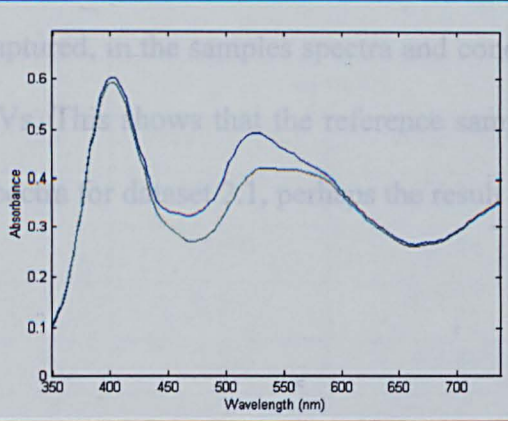
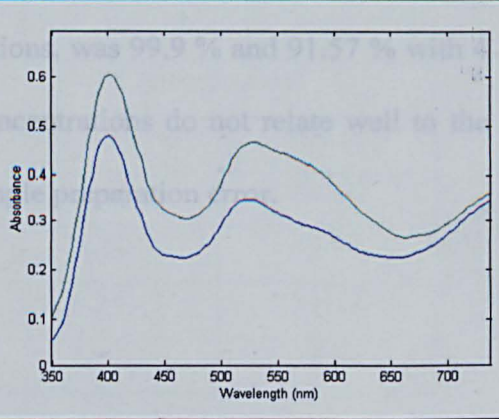
Figure 3.2-5 Dataset 2.2 PCA results; samples scores plots, residual Q values and T^2 value results based on a 4 PC model



For set 2.2, the T^2 shows sample 15 to be an outlier in addition samples 23 – 26. The samples scores plots are not the same between the two datasets. The spectra of each sample 1 – 8 for each dataset are plotted in figure 3.2-5.

Figure 3.2-6 Plots of samples 1 – 8 datasets 2.1 and 2.2 metal ions spectra



Sample 7	Sample 8
	

It can be seen that for some samples, e.g. 3 and 4, the spectra are similar. However for many samples the spectra are very different and in the case of samples 5 – 8 there is a large baseline shift of 0.1 - 0.2 absorbance. From this, it is unknown which set, if either, has the correct spectra.

Further investigations are required to attempt to establish the source of variation, i.e. are the samples labelled or prepared incorrectly, or is there a fault with the spectrometer causing variation between the two days. If the error is sample set dependent, i.e. one set can be modelled sufficiently by PLS and the remaining can't, then the conclusion is that a gross error occurred with dataset that was poorly modelled. If neither set can be modelled, then the assumption will be that, the error lies within the spectrometer.

The spectra and reference concentration information is mean centred for the PLS2 CV calibration models. The percentage variance in the spectra and concentrations are tabulated in table 3.2-1 for dataset 2.1 and table 3.2-2 for dataset 2.2. For dataset 2.1, 99.89 % of the spectral and only 58.04 % of the variance in the samples metal ion

content was captured with 4 LVs. In contrast, for dataset 2.2 the % variance captured, in the samples spectra and concentrations, was 99.9 % and 91.57 % with 4 LVs. This shows that the reference sample concentrations do not relate well to the spectra for dataset 2.1, perhaps the result of sample preparation error.

Table 3.2-1 Dataset 2.1: Percent variance captured by PLS2 model

	Spectra		Concentration	
LV	This LV (%)	Total (%)	This LV (%)	Total (%)
1	77.38	77.38	24.36	24.36
2	15.46	92.84	21.62	45.98
3	6.53	99.37	8.74	54.72
4	0.52	99.89	3.32	58.04

Table 3.2-2 Dataset 2.2: Percent variance captured by PLS2 model

	Spectra		Concentration	
LV	This LV (%)	Total (%)	This LV (%)	Total (%)
1	71.02	71.02	48.01	48.01
2	16.57	87.59	32.65	80.66
3	9.91	97.50	9.24	89.89
4	2.42	99.93	1.68	91.57

The actual metal concentration is plotted against predicted concentration for both datasets in figures 3.2-7-10 and figures 3.2-11-14. Generally the goodness of fit is better for the spectra in dataset 2.2 than for 2.1. The predictions for Cr^{3+} and Cu^{2+} have the most scatter about the line of best fit for the samples with all the components present. Conversely Co^{2+} and Ni^{2+} had very poor prediction for sample 16 and samples 23 – 26 where at least one component was excluded from the dataset. PCA did not indicate sample 16 to be an outlier whereas 23 – 26 were.

The poor predictions for dataset 2.1 are in agreement with the poor capture of the concentration information. Even when the number of LVs is extended to 10 (not tabulated) less than 70 % of the concentration variation is captured. It is concluded

that there is a gross error with dataset 2.1 and it will be excluded from further work. Because of this, a dataset of industrial based measurement was not available for the intercomparison study and the development of a standard dataset.

Figure 3.2-7 Dataset 2.1: PLS2 Co^{2+} actual Vs predicted results based on 4 LVs

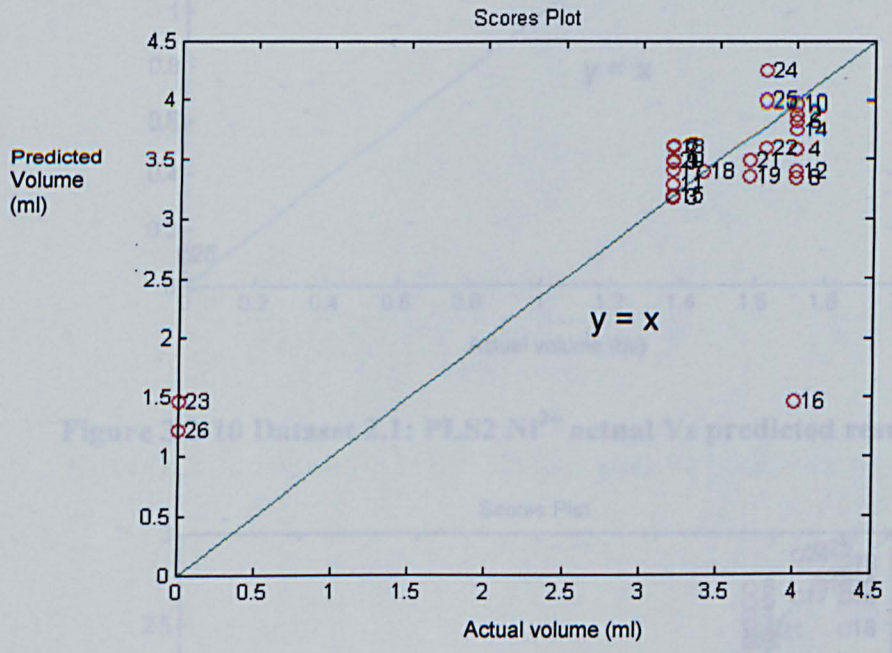


Figure 3.2-8 Dataset 2.1: PLS2 Cr^{3+} actual Vs predicted results based on 4 LVs

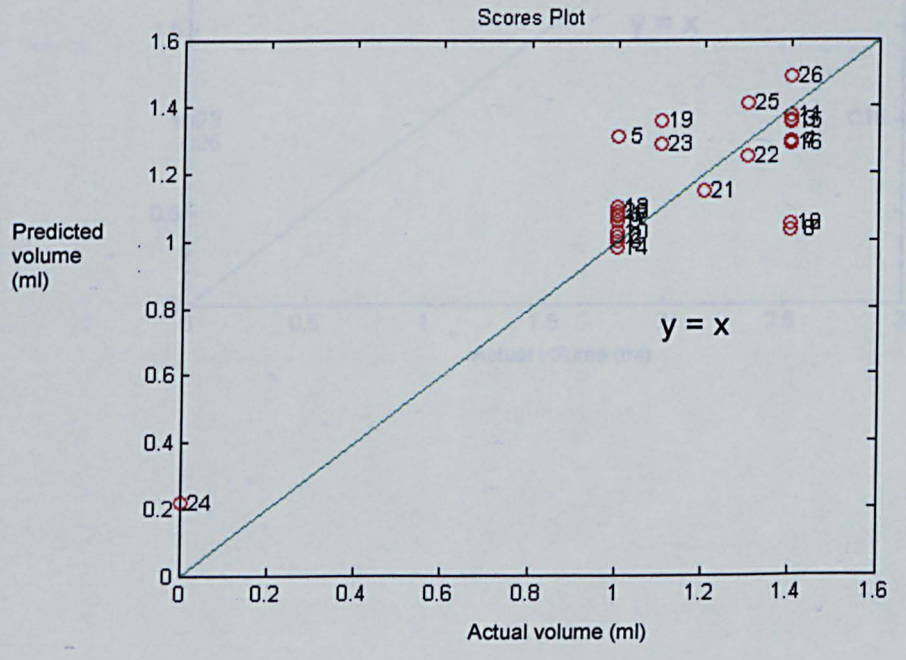


Figure 3.2-9 Dataset 2.1: PLS2 Cu^{2+} actual Vs predicted results based on 4 LVs

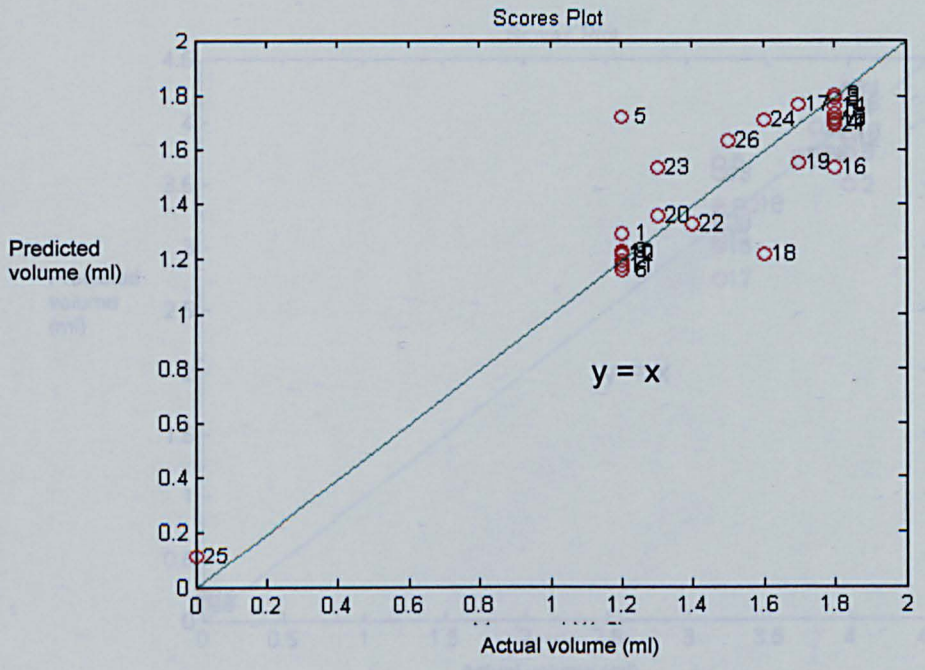


Figure 3.2-10 Dataset 2.1: PLS2 Ni^{2+} actual Vs predicted results based on 4 LVs

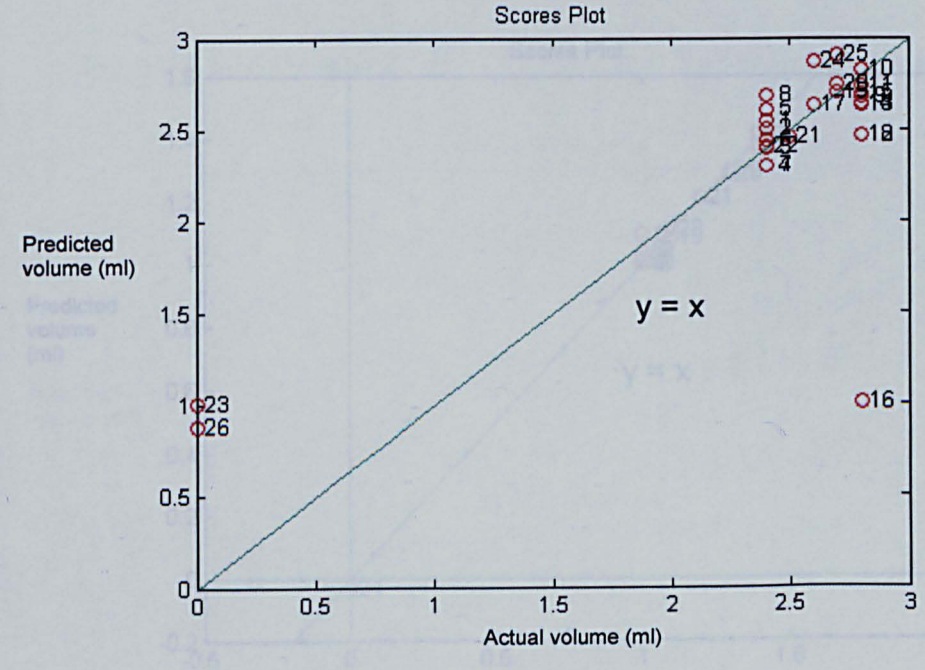


Figure 3.2-11 Dataset 2.2: Co^{2+}PLS2 actual Vs predicted results based on 4 LVs

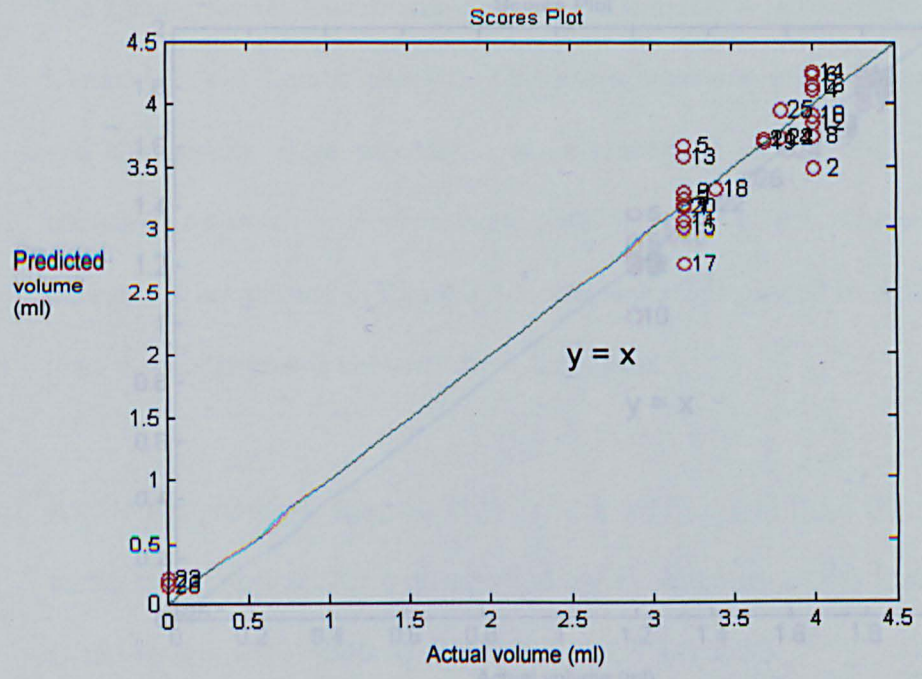


Figure 3.2-12 Dataset 2.2: Cr^{3+}PLS2 actual Vs predicted results based on 4 LVs

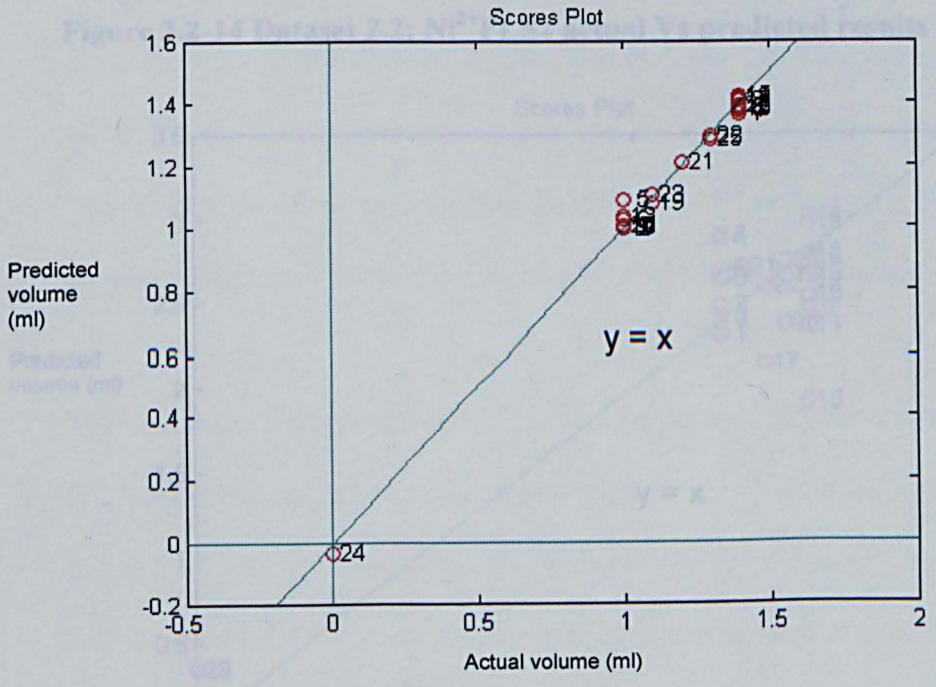


Figure 3.2-13 Dataset 2.2: Cu^{2+}PLS2 actual Vs predicted results based on 4 LVs

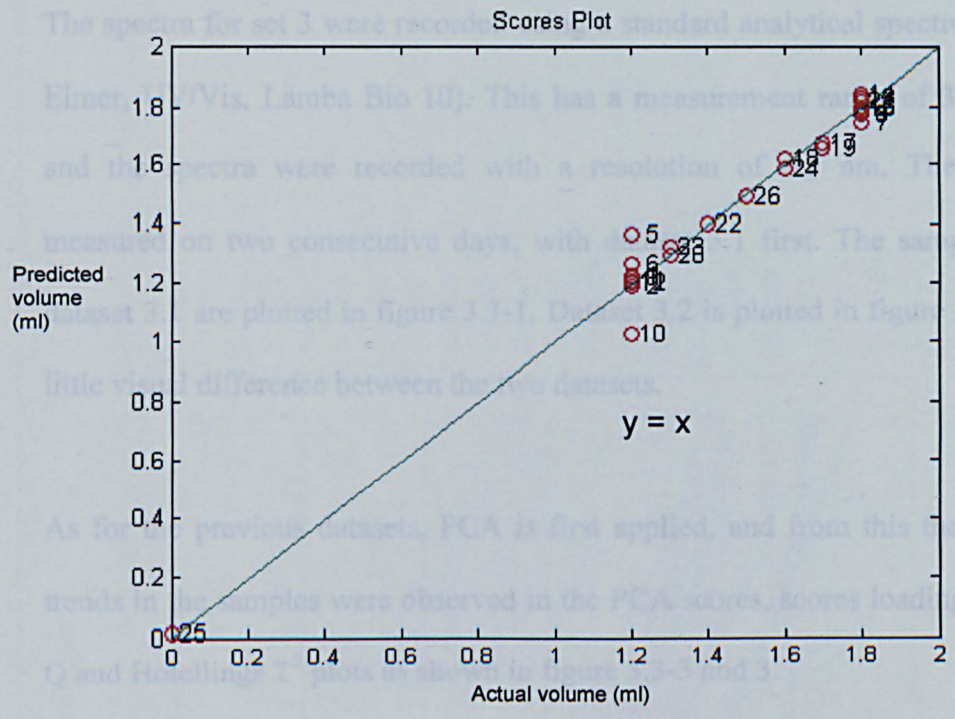
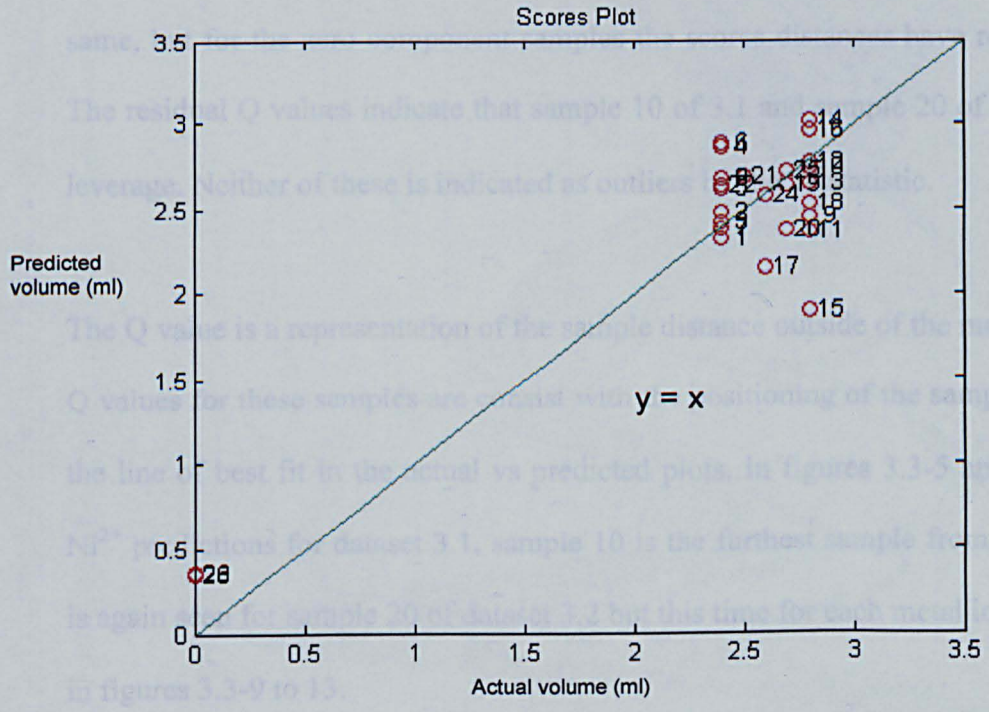


Figure 3.2-14 Dataset 2.2: Ni^{2+}PLS2 actual Vs predicted results based on 4 LVs



3.3 Measurement of metals mixtures spectra set 3

The spectra for set 3 were recorded using a standard analytical spectrometer (Perkin Elmer, UV/Vis, Lamba Bio 10). This has a measurement range of 350 – 1000 nm and the spectra were recorded with a resolution of 0.5 nm. The spectra were measured on two consecutive days, with dataset 3.1 first. The samples spectra of dataset 3.1 are plotted in figure 3.3-1. Dataset 3.2 is plotted in figure 3.3-2. There is little visual difference between the two datasets.

As for the previous datasets, PCA is first applied, and from this the variation and trends in the samples were observed in the PCA scores, scores loadings and residual Q and Hotellings T^2 plots as shown in figure 3.3-3 and 3.

Comparison of the PCA results between datasets 3.1 and 3.2 finds that the scores plots for PC's 1, 2 and 4 are almost identical, for PC 3 the scores patterns are the same, but for the zero component samples the scores distances have reversed signs. The residual Q values indicate that sample 10 of 3.1 and sample 20 of 3.2 have high leverage. Neither of these is indicated as outliers by the T^2 statistic.

The Q value is a representation of the sample distance outside of the model. The high Q values for these samples are consist with the positioning of the sample away from the line of best fit in the actual vs predicted plots. In figures 3.3-5 and 8, Co^{2+} and Ni^{2+} predictions for dataset 3.1, sample 10 is the furthest sample from the line. This is again seen for sample 20 of dataset 3.2 but this time for each metal ions prediction, in figures 3.3-9 to 13.

Figure 3.3-2 Dataset 3.2 metal ions spectra recorded using an analytical spectrometer

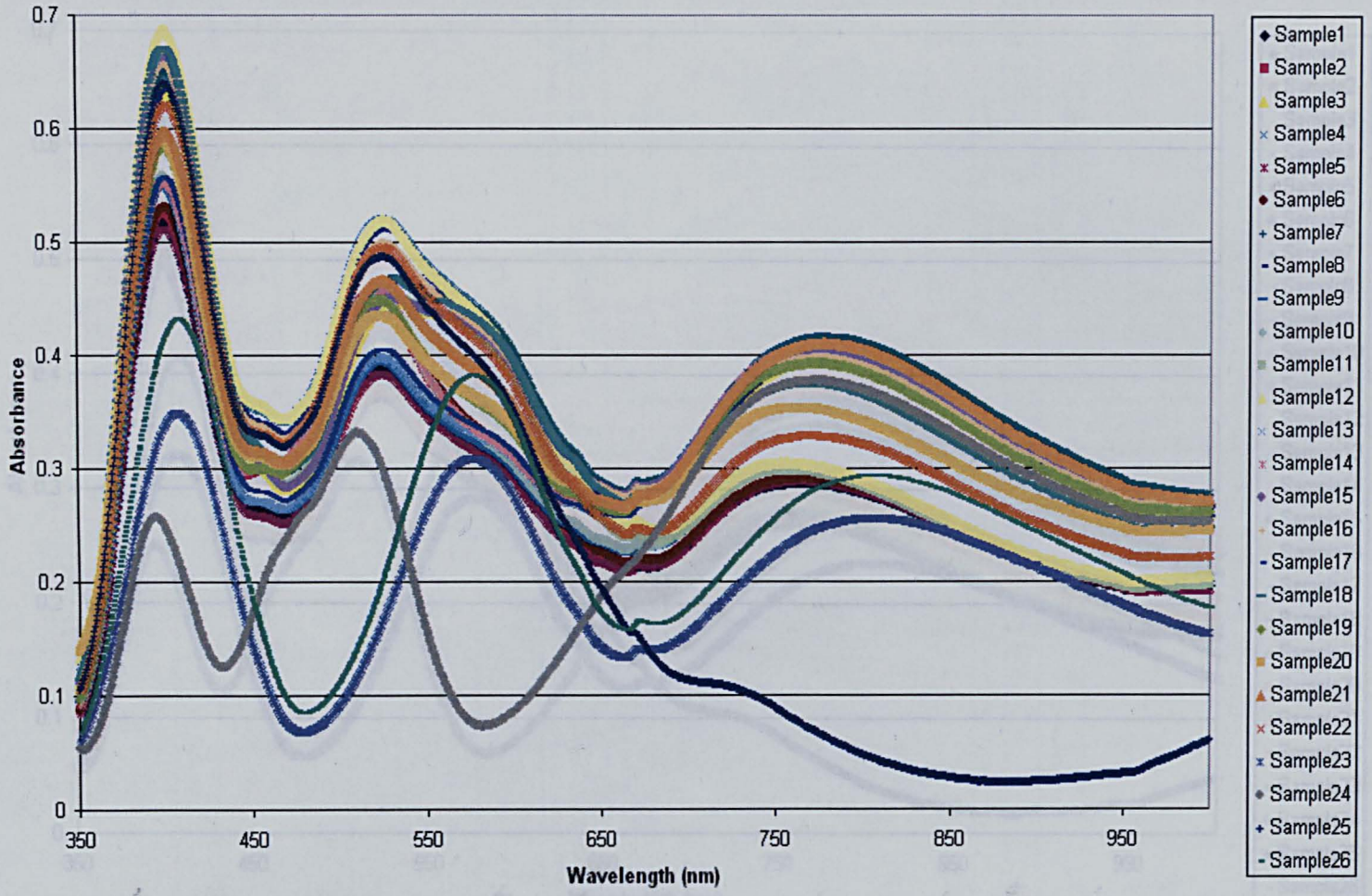


Figure 3.3-1 Metals spectra recorded using an analytical spectrometer

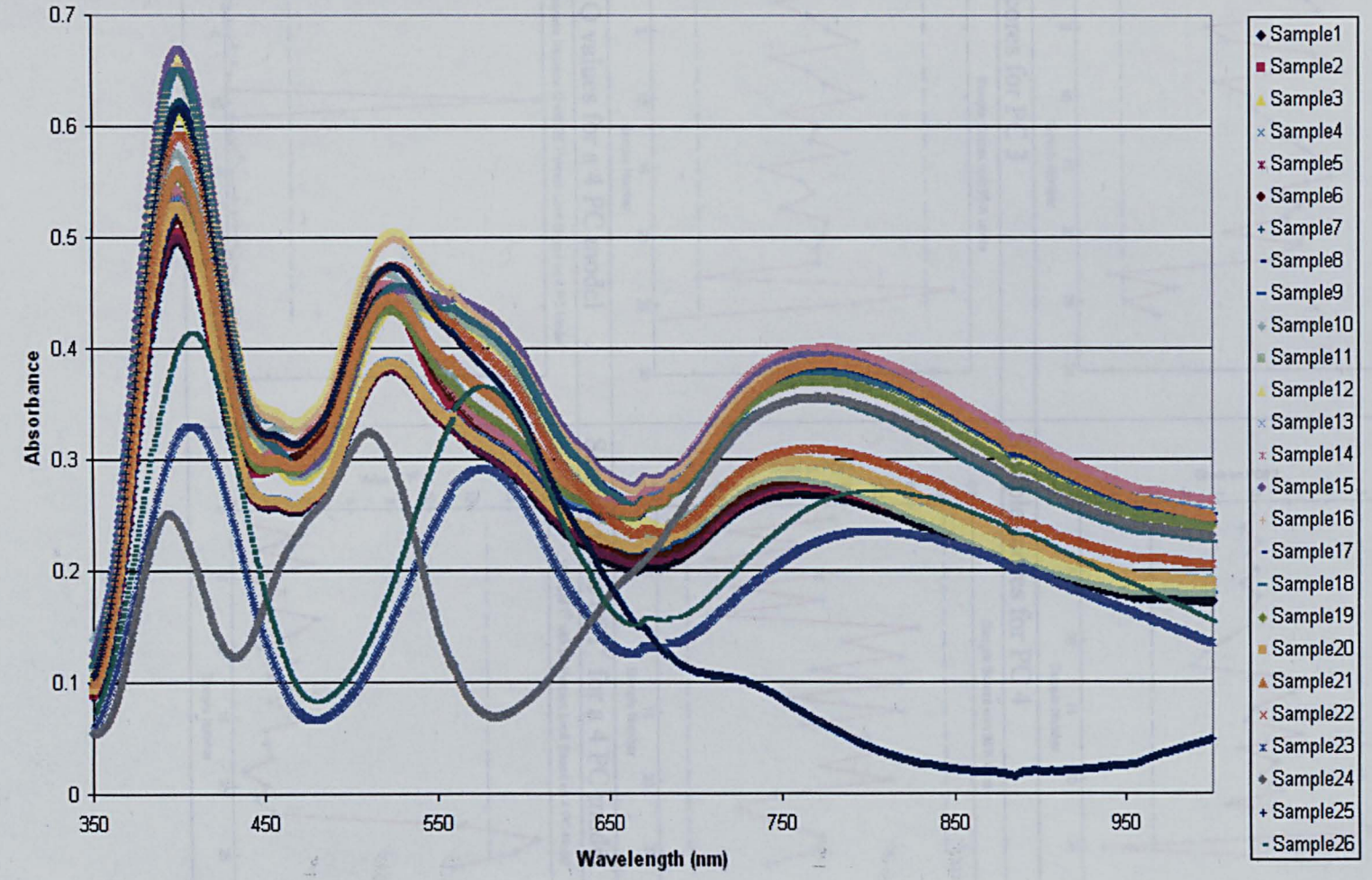


Figure 3.3-3 Dataset 3.1: PCA samples scores plots, residual Q values and T^2 value results based on a 4 PC model

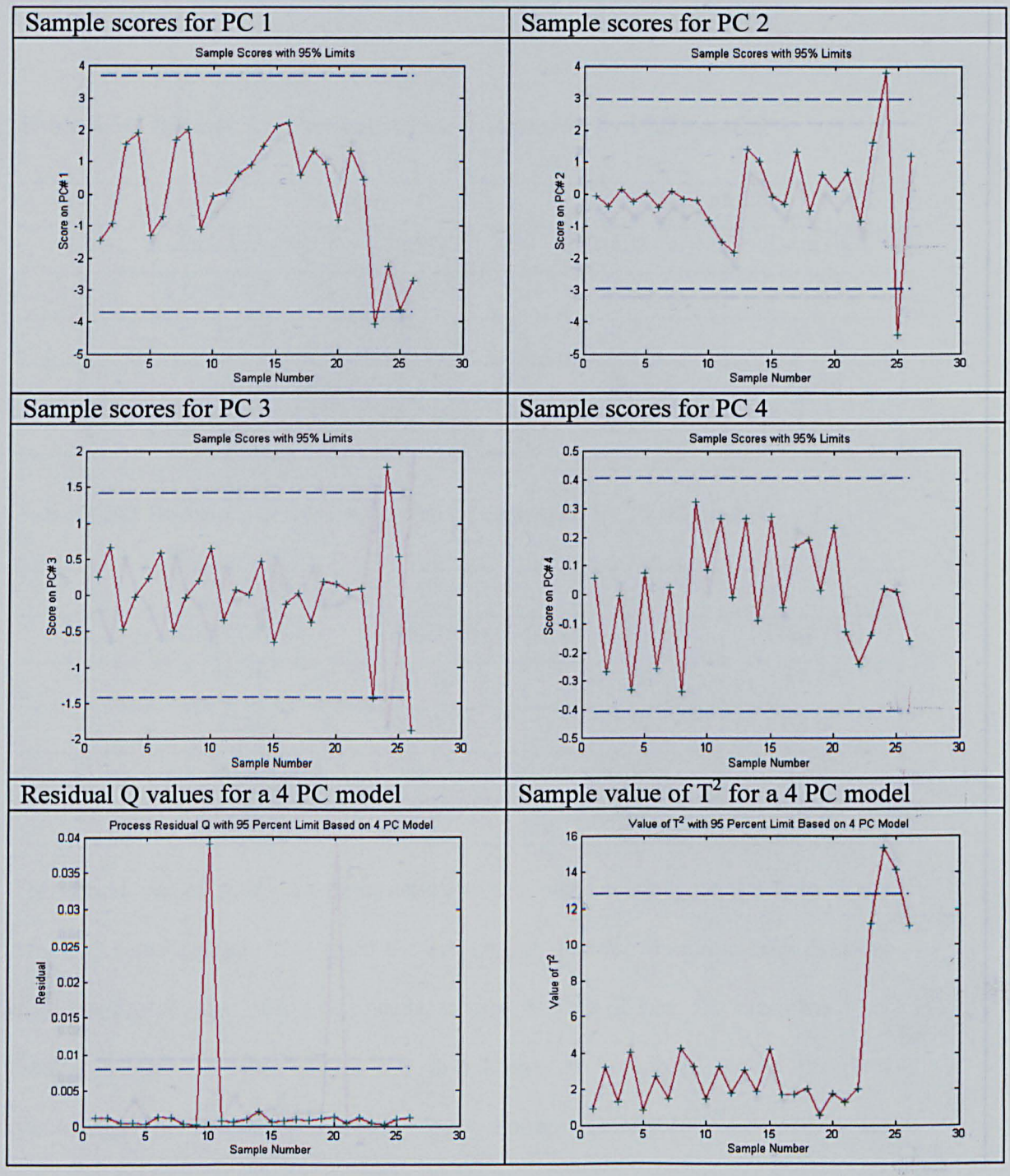
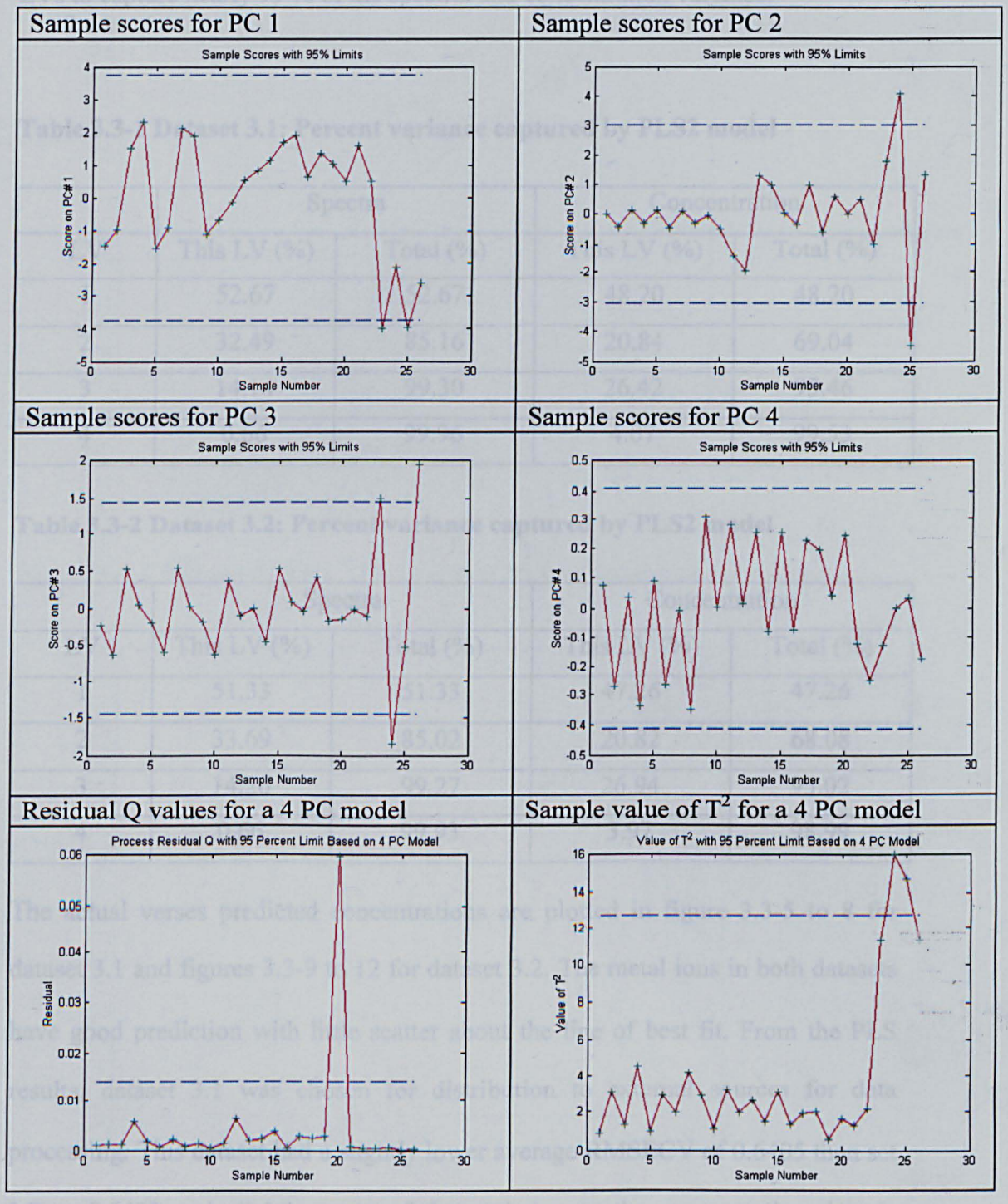


Figure 3.3-4 Dataset 3.2: PCA samples scores plots, residual Q values and T^2 value results based on a 4 PC model



For both datasets PLS CV models are calculated; mean centering scaled the data. The percentage variance captured by the models is good with both sets requiring 4 LVs to capture nearly 99 % of the spectral and concentration variance.

Table 3.3-1 Dataset 3.1: Percent variance captured by PLS2 model

	Spectra		Concentration	
LV	This LV (%)	Total (%)	This LV (%)	Total (%)
1	52.67	52.67	48.20	48.20
2	32.49	85.16	20.84	69.04
3	14.14	99.30	26.42	95.46
4	0.66	99.96	4.07	99.53

Table 3.3-2 Dataset 3.2: Percent variance captured by PLS2 model

	Spectra		Concentration	
LV	This LV (%)	Total (%)	This LV (%)	Total (%)
1	51.33	51.33	47.26	47.26
2	33.69	85.02	20.82	68.08
3	14.26	99.27	26.94	95.02
4	0.66	99.93	3.97	98.99

The actual verses predicted concentrations are plotted in figure 3.3-5 to 8 for dataset 3.1 and figures 3.3-9 to 12 for dataset 3.2. The metal ions in both datasets have good prediction with little scatter about the line of best fit. From the PLS results, dataset 3.1 was chosen for distribution to external sources for data processing. This dataset had a slightly lower average RMSECV of 0.6405 than set 3.2 at 0.6492 and slightly more of the variation in the concentration data is captured within 4 latent variables.

Figure 3.3-5 Dataset 3.1: Co^{2+} PLS2 actual Vs predicted plots

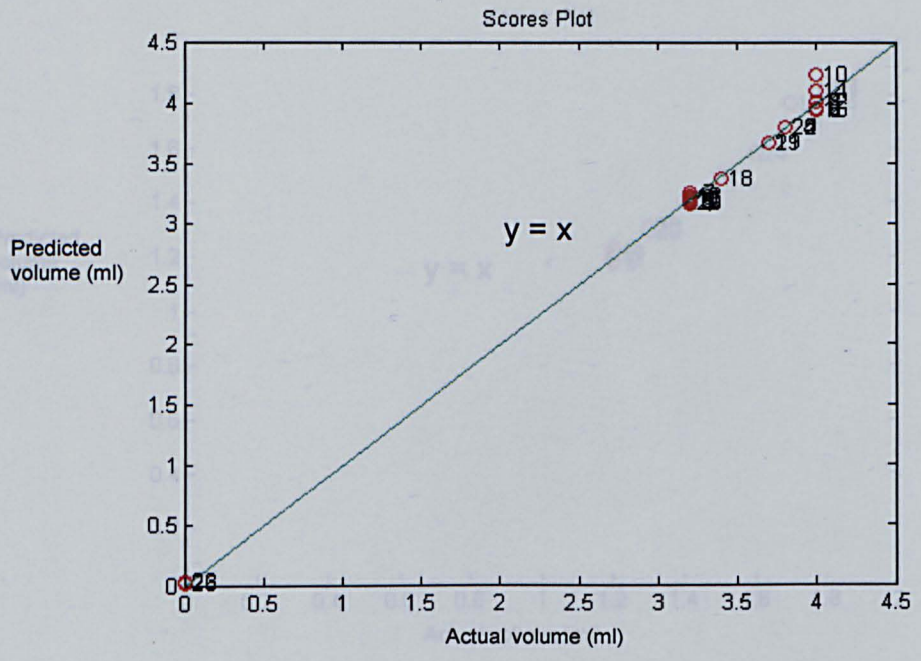


Figure 3.3-6 Dataset 3.1: Cr^{3+} PLS2 actual Vs predicted plots

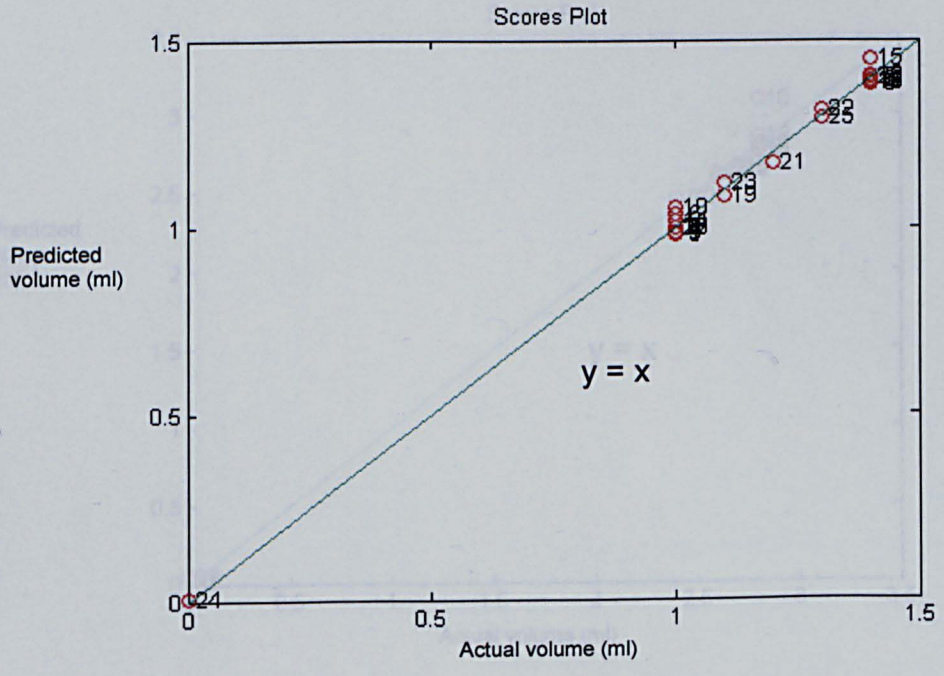


Figure 3.3-7 Dataset 3.1: Cu^{2+} PLS2 actual Vs predicted plots

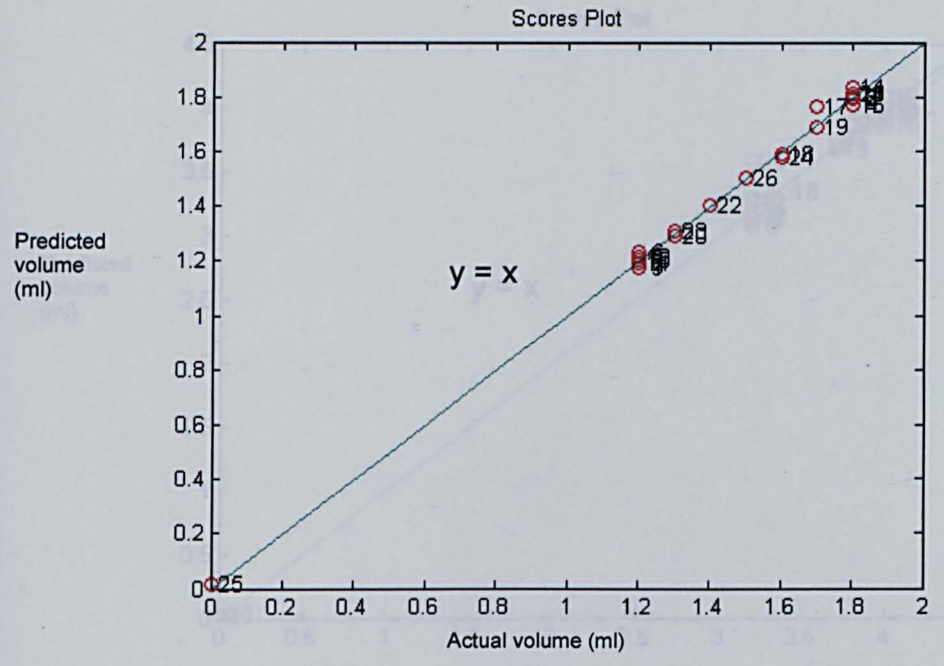


Figure 3.3-8 Dataset 3.1: Ni^{2+} PLS2 actual Vs predicted plots

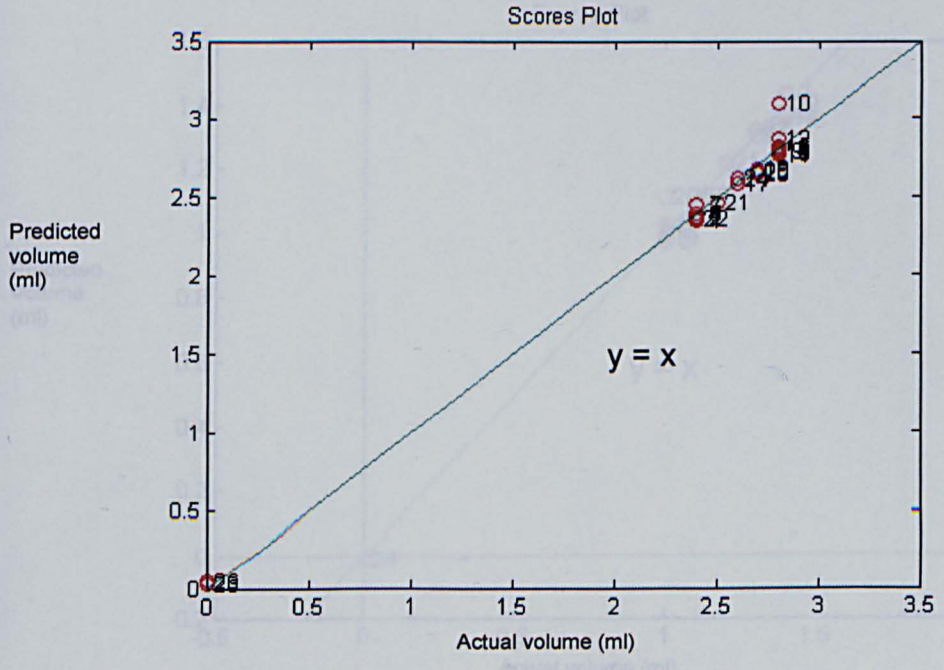


Figure 3.3-9 Dataset 3.2: Co^{2+} PLS2 actual Vs predicted plots

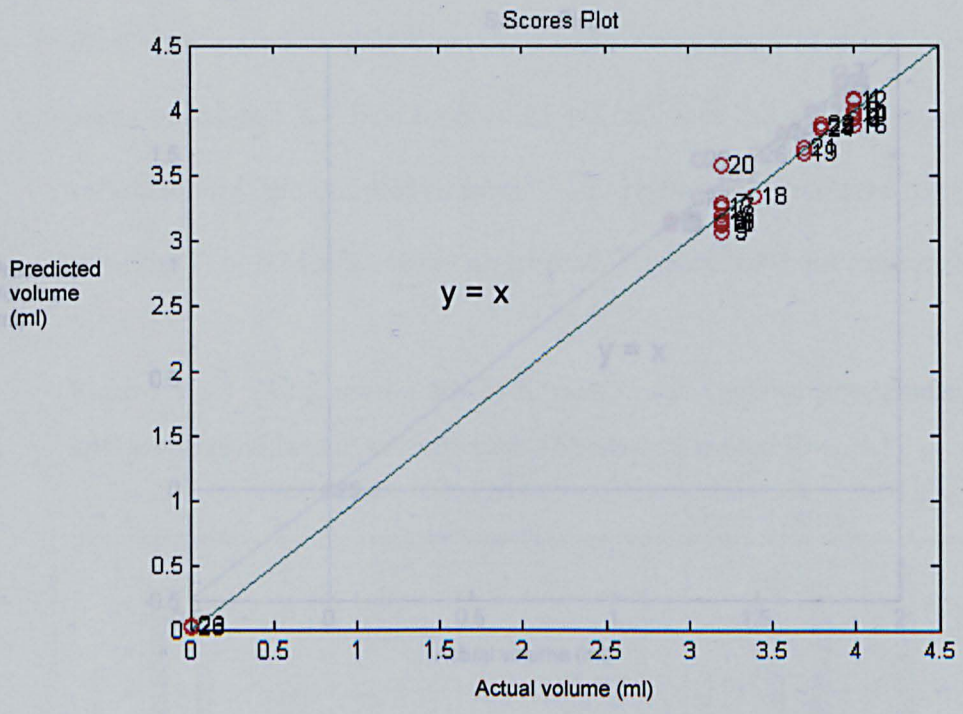


Figure 3.3-10 Dataset 3.2: Cr^{3+} PLS2 actual Vs predicted plots

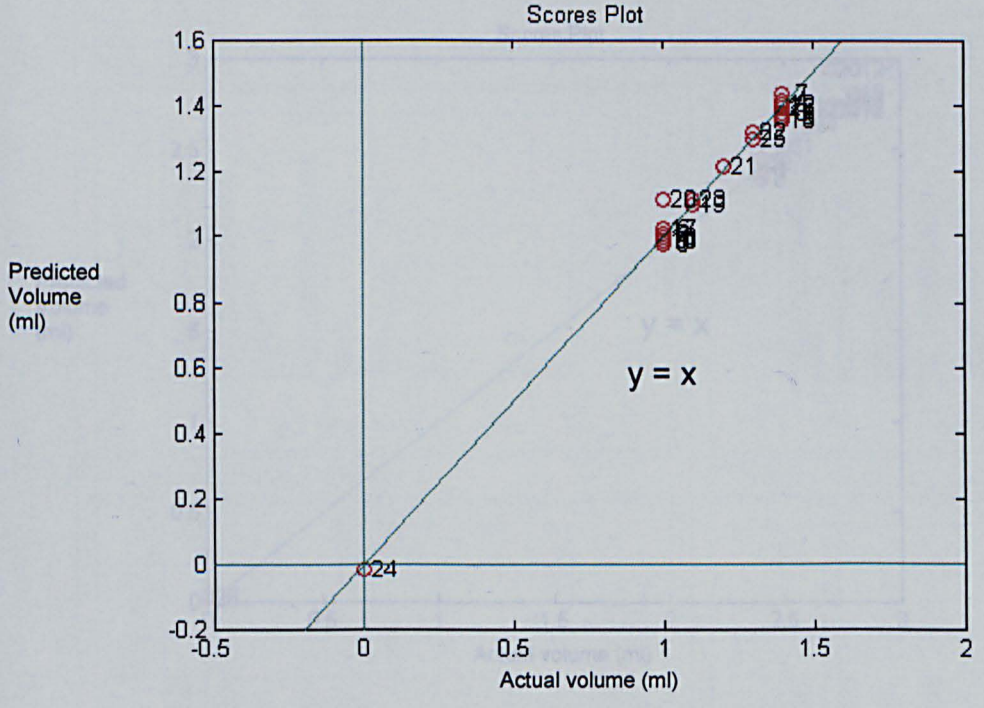


Figure 3.3-11 Dataset 3.2: Cu^{2+} PLS2 actual Vs predicted plots

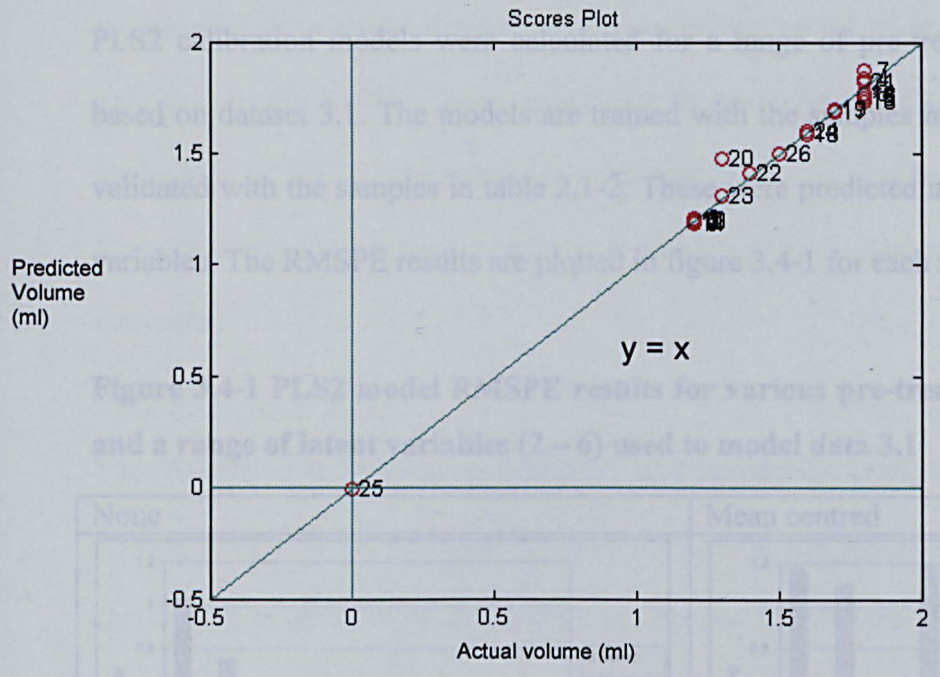
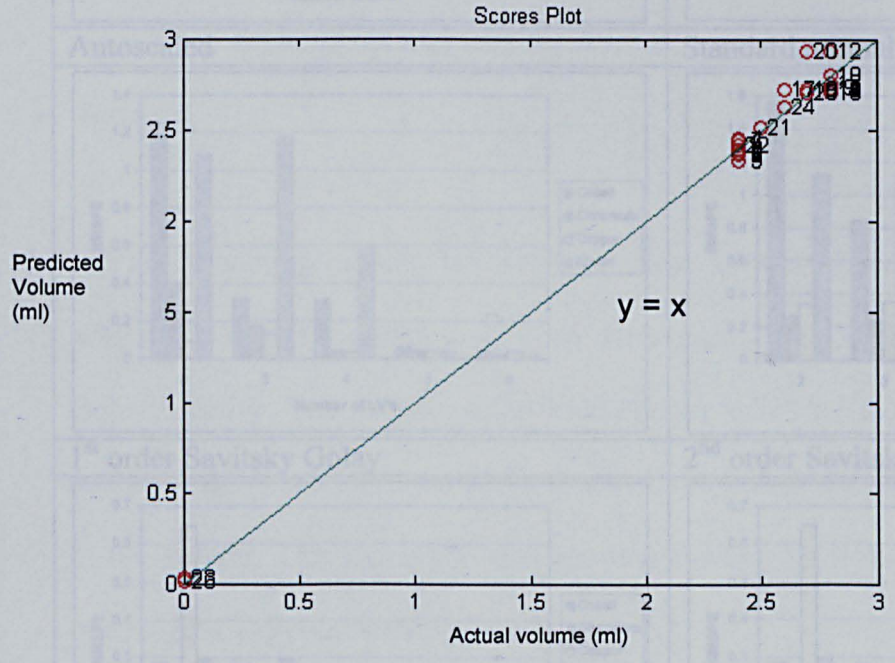


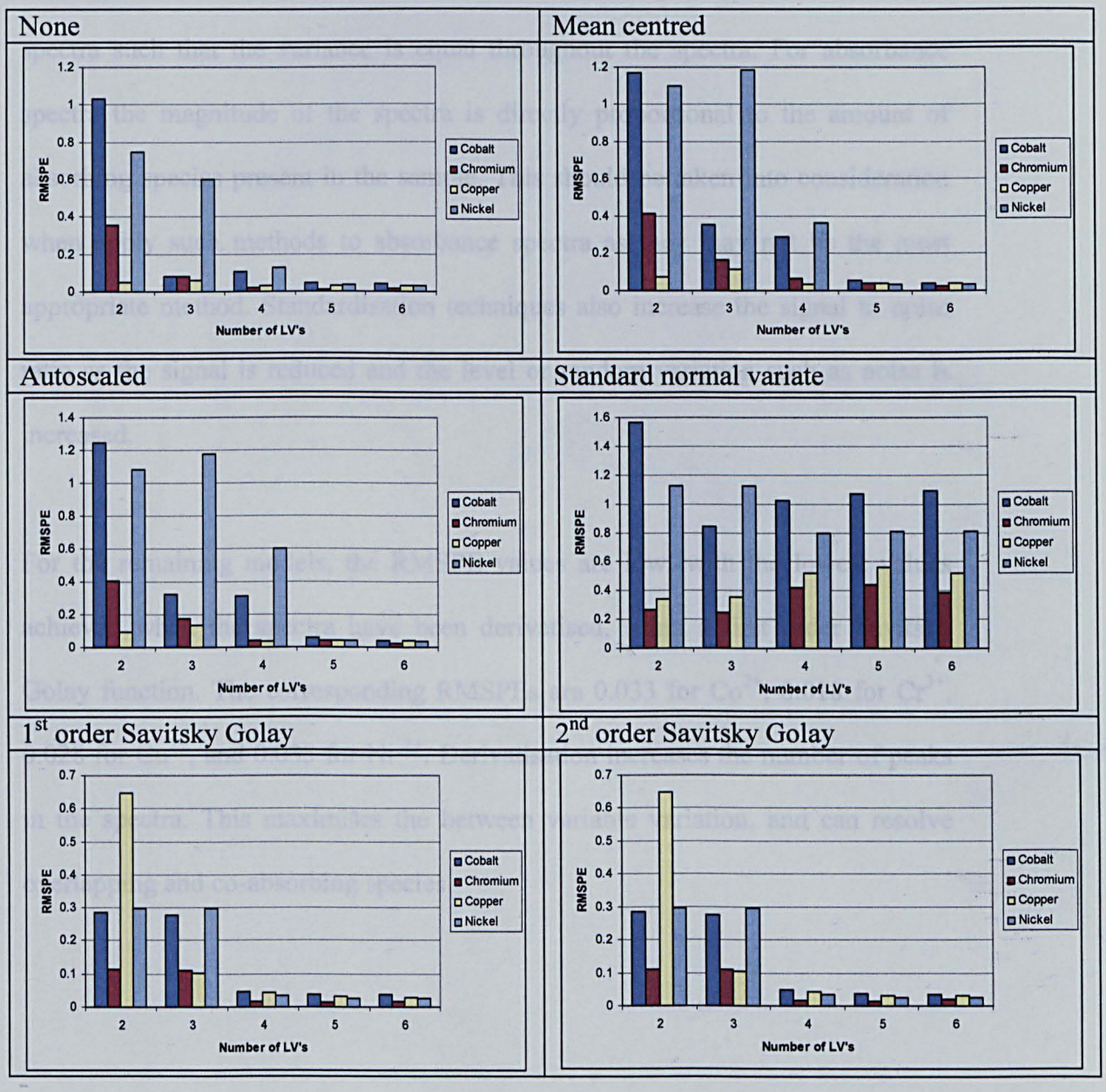
Figure 3.3-12 Dataset 3.2: Ni^{2+} PLS2 actual Vs predicted plots



3.4 Effect of pre-treatment method on PLS prediction

PLS2 calibration models were calculated for a range of pre-treatment methods based on dataset 3.1. The models are trained with the samples in table 2.1-1 and validated with the samples in table 2.1-2. These were predicted using 2 – 6 latent variables. The RMSPE results are plotted in figure 3.4-1 for each metal ion.

Figure 3.4-1 PLS2 model RMSPE results for various pre-treatment methods and a range of latent variables (2 – 6) used to model data 3.1



The plots show that models based on 2 or 3 LVs have high RMSPE values, and are insufficient to model the data well. Where 4 or more LVs are used, there is a reduction in the RMSPE value, especially for prediction of Co^{2+} and Ni^{2+}. The application of SNV to the spectra is an exception to this, and the RMSPE values for each metal ion are similar when the number of LVs used to model the data increases.

Standardisation techniques such as SNV and autoscaling work to standardise the spectra such that the variance is equal throughout the spectra. For absorbance spectra the magnitude of the spectra is directly proportional to the amount of absorbing species present in the sample. This should be taken into consideration when apply such methods to absorbance spectra as they may not be the most appropriate method. Standardisation techniques also increase the signal to noise ratio as the signal is reduced and the level or random variation such as noise is increased.

For the remaining models, the RMSPE values are low, with the lowest values achieved when the spectra have been derivatised, using a first order Savitsky Golay function. The corresponding RMSPEs are 0.033 for Co^{2+}, 0.018 for Cr^{3+}, 0.028 for Cu^{2+}, and 0.033 for Ni^{2+}. Derivatisation increases the number of peaks in the spectra. This maximises the between variable variation, and can resolve overlapping and co-absorbing species.

The SNV and derivatised spectra of the training and test samples are plotted in figures 3.4-2 to 5. These show that SNV has altered some of the variation in the spectra to be consistent across the absorbance range. Conversely, and as expected, derivatisation has increased the number of peaks in the spectra.

Generally the PLS models of this data are not dependent on a specific pre-treatment method with only small variations between methods.

Figure 3.4-2 Plot of SNV standardised training samples spectra of dataset 3.1

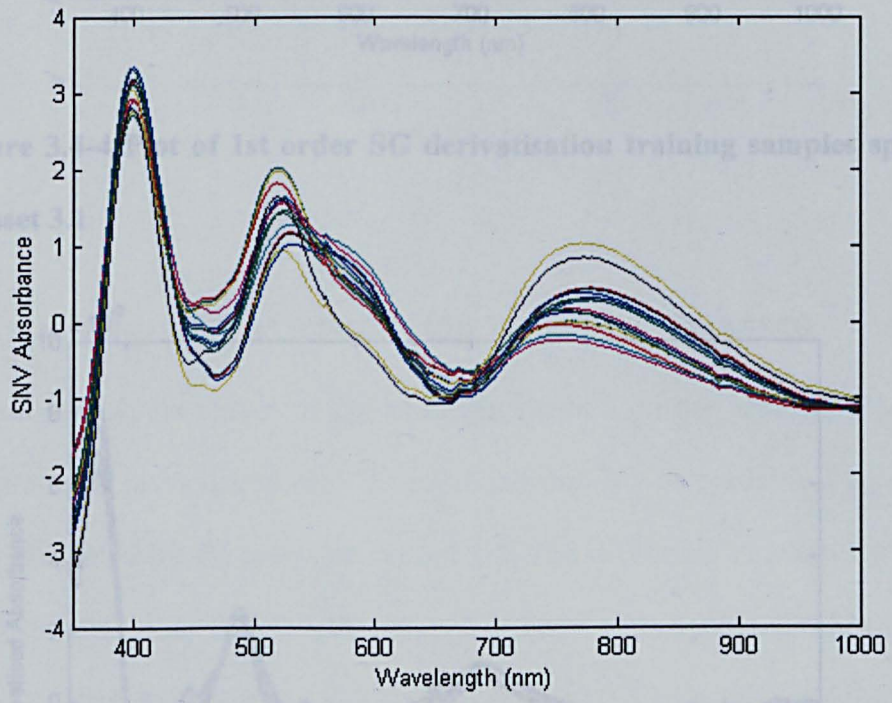


Figure 3.4-3 Plot of SNV standardised test samples spectra of dataset 3.1

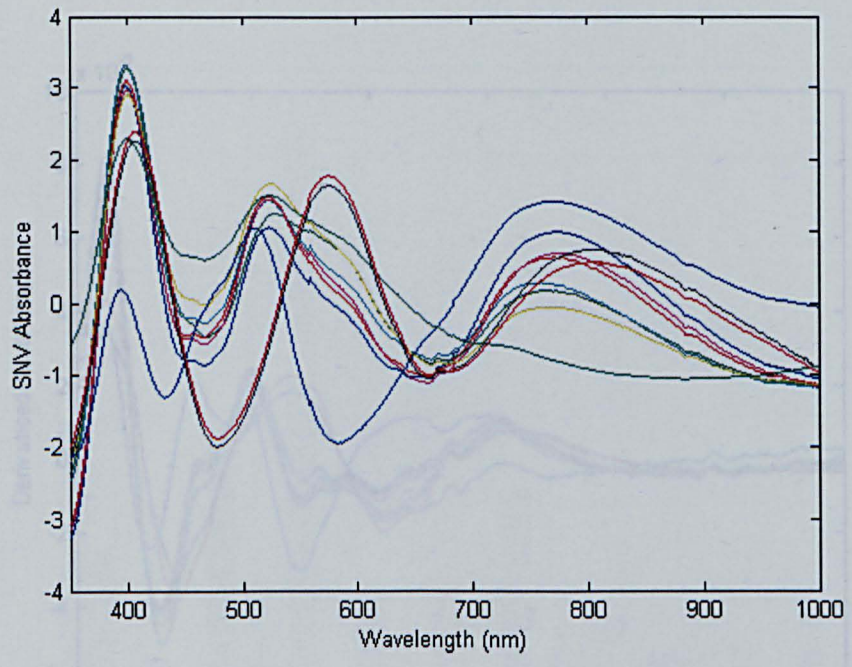


Figure 3.4-4 Plot of 1st order SG derivatisation training samples spectra of dataset 3.1

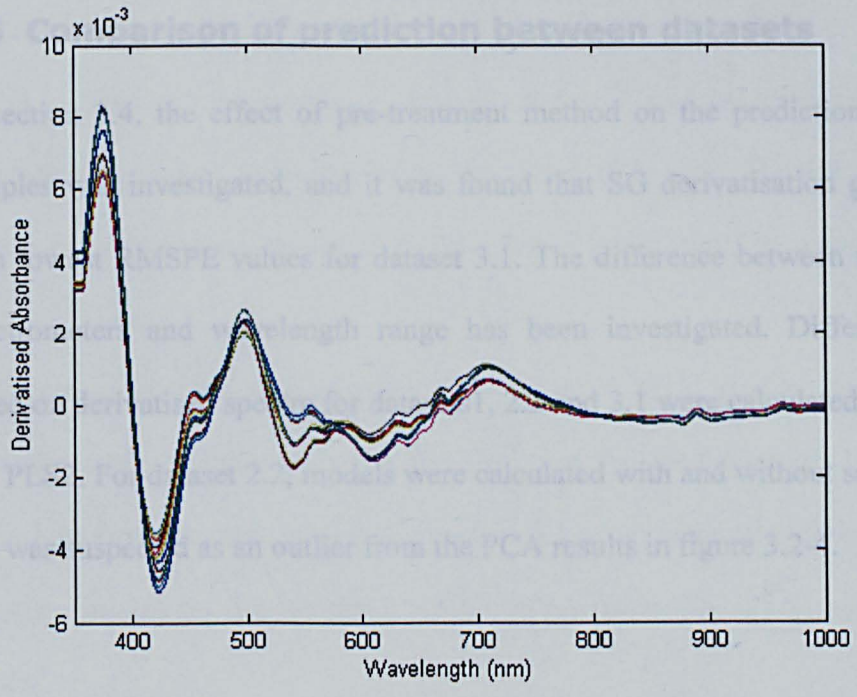
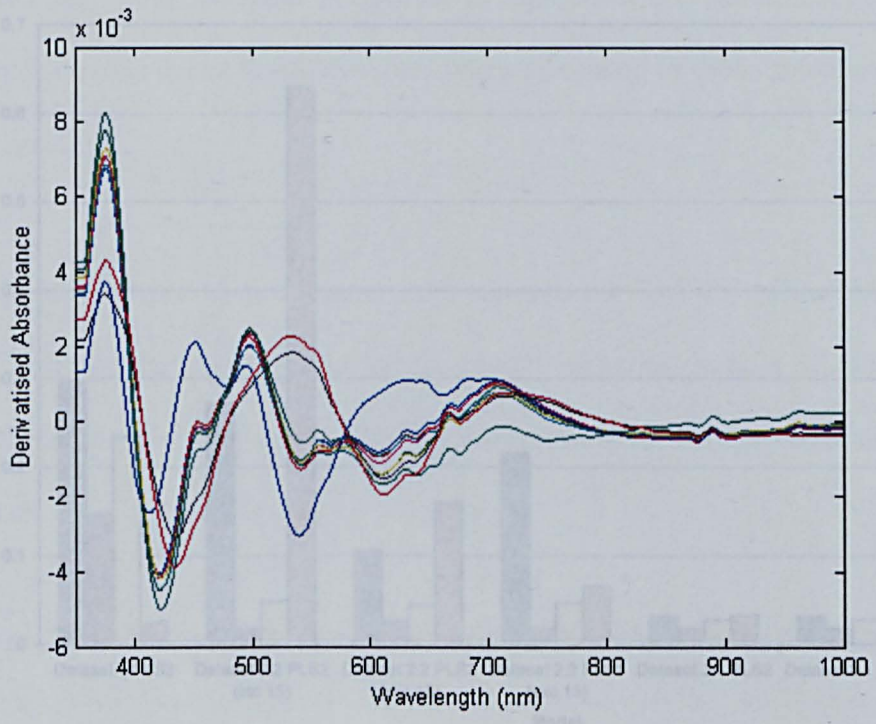


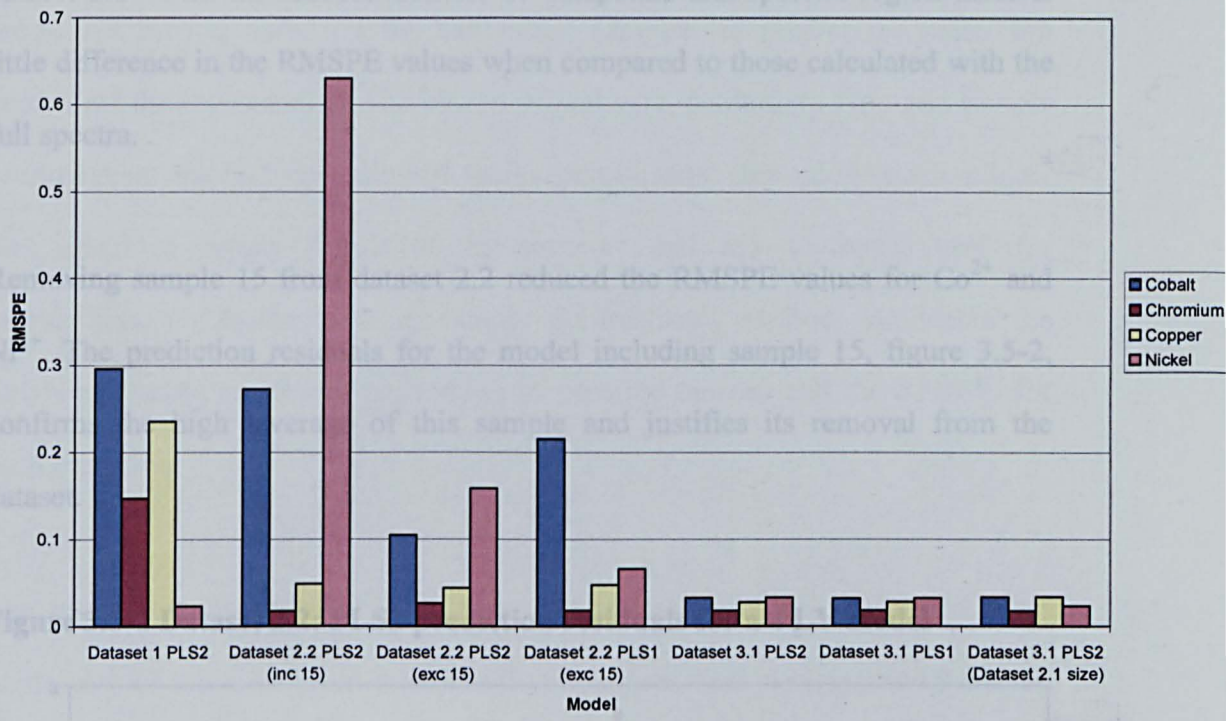
Figure 3.4-5 Plot of 1st order SG derivatisation test samples spectra of dataset 3.1



3.5 Comparison of prediction between datasets

In section 3.4, the effect of pre-treatment method on the prediction of the test samples was investigated, and it was found that SG derivatisation gave models with lowest RMSPE values for dataset 3.1. The difference between the datasets, spectrometers and wavelength range has been investigated. Different models based on derivatised spectra for datasets1, 2.2 and 3.1 were calculated using PLS1 and PLS2. For dataset 2.2, models were calculated with and without sample 15, as this was suspected as an outlier from the PCA results in figure 3.2-4.

Figure 3.5-1 Summary of PLS models for the metal ion solutions spectral datasets



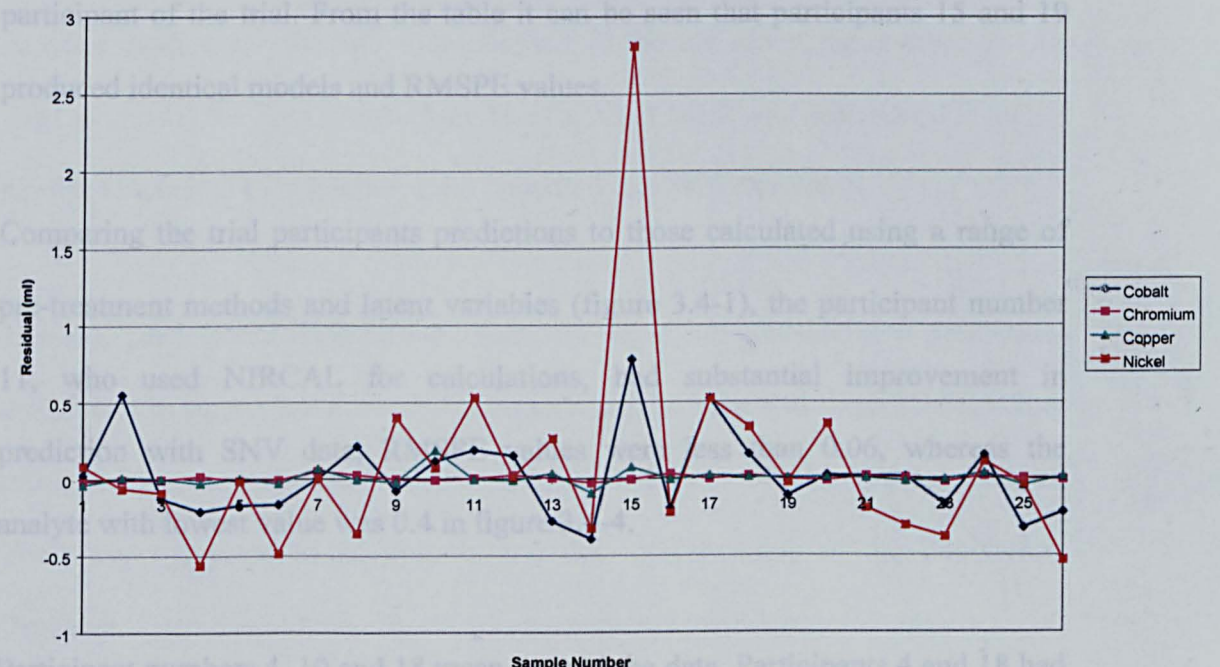
Dataset 1 has significantly higher errors of prediction than those of datasets 2.2 and 3.1. This is the simulated dataset, and the spectra were based on the measurement of pure metal ions of concentration 1000 µg/l. These had a very low absorbance of less than 0.1 which could result in lower signal to noise ratio.

The increase in prediction error for Co^{2+} when PLS1 is used to model the data indicates there is dependence on the Co^{2+} ion for prediction of the remaining metal ions in the sample, and this situation is known as serial correlation. This is not seen in dataset 3.1 as the RMSPE values from The PLS1 and PLS2 are similar.

The difference in quality of the spectra is also seen when dataset 3.1 is reduced to the same wavelength range and number of data points as those of spectra in dataset 2.2. With the reduced number of datapoints and spectral region there is little difference in the RMSPE values when compared to those calculated with the full spectra.

Removing sample 15 from dataset 2.2 reduced the RMSPE values for Co^{2+} and Ni^{2+}. The prediction residuals for the model including sample 15, figure 3.5-2, confirms the high leverage of this sample and justifies its removal from the dataset.

Figure 3.5-2 Dataset 2.2: PLS2 prediction residuals for a 4 LV model



3.6 Collaborative trial results

In the trial 20 people took part. They were given dataset 3.1 with details of how to produce a model, based on the calibration samples, to predict the metal ion content of the test samples. The choice of software, calibration type and sample pre-treatment method was allowed to the participants' discretion. Each subject was asked to supply details of the software and any toolboxes used for calculations, regression method, sample pre-treatment method, applicable the number of latent variables, the number of possible outliers and the RMSPE for each metal ion.

The model details and RMSPE values are tabulated in table 3.6-1 for each participant of the trial. From the table it can be seen that participants 15 and 19 produced identical models and RMSPE values.

Comparing the trial participants predictions to those calculated using a range of pre-treatment methods and latent variables (figure 3.4-1), the participant number 11, who used NIRCAL for calculations, had substantial improvement in prediction with SNV data; RMSPE values were less than 0.06, whereas the analyte with lowest value was 0.4 in figure 3.4-4.

Participant numbers 4, 10 and 18 mean centred the data. Participants 4 and 18 had similar RMSPE values; number 10 used GRAMMS software that produced a model that gave much lower RMSPE values of less than 0.036, for the same number of LVs.

Derivatisation gave the most consistent results between participants, for numbers 1, 8, 12, 13 and 20 the RMSPE values are similar to those in figure 3.4-1. The participants 2 and 3 autoscaled the data and reported much lower RMSPE values. As both used PLS1 instead of PLS2 used the lower errors in prediction are anticipated. There is a difference in results for the participant numbers 7 and 9 who also autoscaled, these participants used PLS2 but had much higher error prediction for some metal ions nickel e.g., an RMSPE of 0.4 instead of less than 0.05 as expected.

The largest variations are where the data has not been scaled. For participants 5 and 17 the RMSPE values are much higher than calculated, however true comparison is not possible as number 5 used MLR, indicating PLS to be a more superior method and the individual number 17 did not report the number of LVs used to model the data. Participant 14 also used MLR and had RMSPE values similar to where 2 LVs modelled the untreated data in figure 3.4-1.

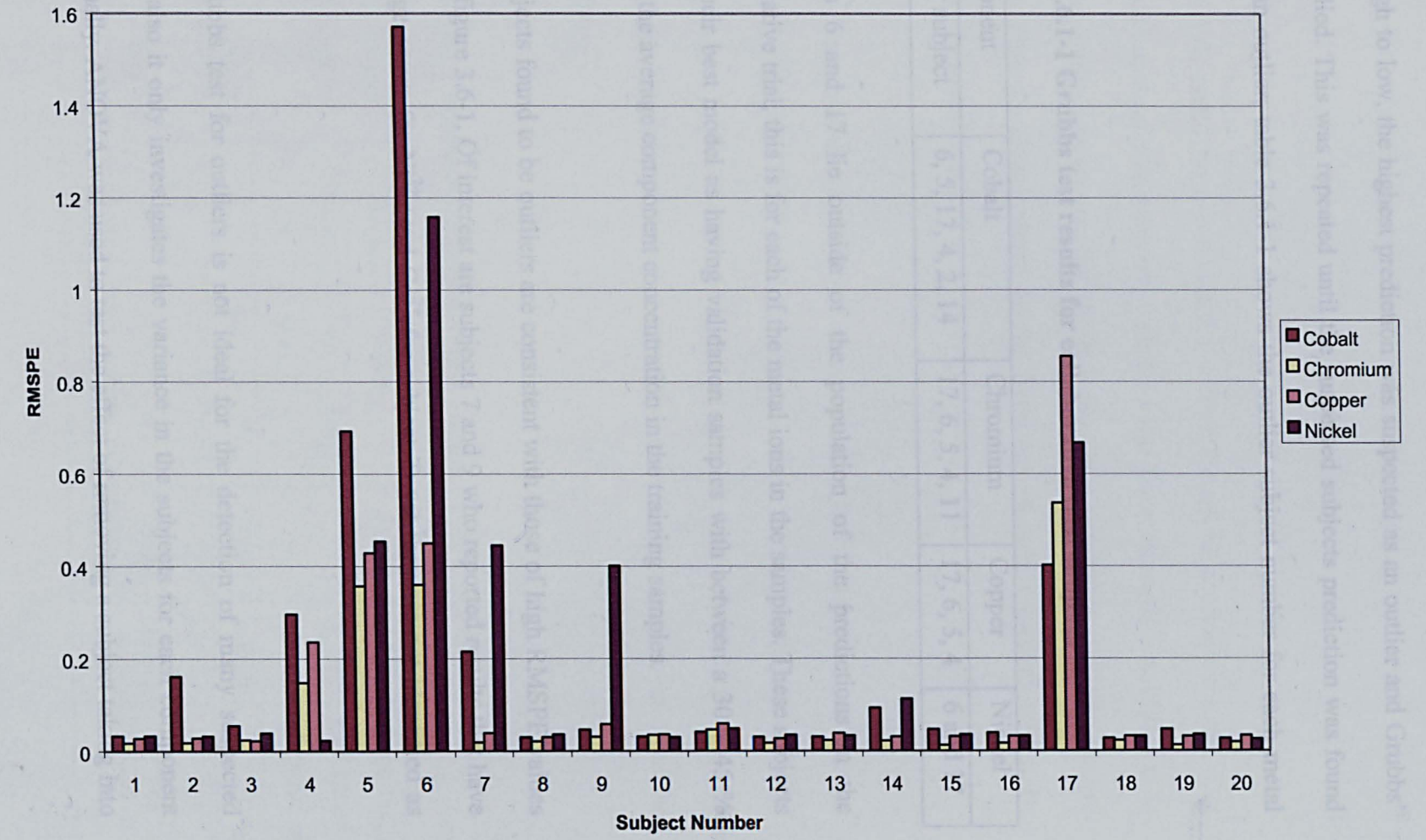
The trial participants RMSPE values are for metal ion concentration predictions are plotted in figure 3.6-1. From the graph it can be seen that the RMSPE values reported by some individuals are very high, the reason for which is not obvious. These are suspected to be outliers and not representative of the population. Grubbs test was used to identify if these participants predictions are outliers.

Table 3.6-1 Results reported by each individual, including the type of calibration model, pre-treatment method, number of latent variables used, number of possible outliers and the RMSPE for each metal ion in the samples

Subject Number	Software	Method	Pre-treatment	No. of LVs	No. of outliers	RMSPE			
						Co^{2+}	Cr^{3+}	Cu^{2+}	Ni^{2+}
1	ML,PLST	PLS1	1st Savisky Golay	4	0	0.0328	0.0176	0.0276	0.0330
2	ML,PLST	PLS1	Autoscale	4	0	0.1610	0.0180	0.0282	0.0323
3	ML,PLST	PLS1	Autoscale	5	0	0.0552	0.0241	0.0220	0.0391
4	ML	PLS2	Mean centre	4	0	0.2958	0.1478	0.2354	0.0233
5	ML	MLR	none	-	0	0.6922	0.3567	0.4289	0.4539
6	SIMCA-P	PLS1	OSC	4	6	1.5714	0.3599	0.4501	1.1577
7	#	PLS2	Autoscale	4	2	0.2151	0.0183	0.0393	0.4453
8	ML,PLST	PLS1	1st order derivatisation	4	0	0.0296	0.0211	0.0300	0.0358
9	ML	PLS2	Autoscale	6	4	0.0457	0.0311	0.0583	0.4012
10	GRAMM	PLS2	Mean centre	4	0	0.0312	0.0343	0.0359	0.0298
11	NIRCAL	PLS1	SNV	6	2	0.0410	0.0459	0.0595	0.0483
12	ML,PLST	PLS2	1st order derivatisation	4	0	0.0311	0.0174	0.0270	0.0349
13	ML	PLS1	1st derv.	4	0	0.0310	0.0225	0.0383	0.0329
14	ML	MLR	none	-	#	0.0925	0.0219	0.0307	0.1123
15	ML	PLS	none	5	#	0.0469	0.0121	0.0314	0.0355
16	ML	PLS	none	6	#	0.0392	0.0165	0.0319	0.0327
17	ML	PLS	none	#	#	0.4022	0.5373	0.8550	0.6673
18	ML	PLS	Mean centre	7	#	0.0273	0.0221	0.0317	0.0318
19	ML	PLS	none	5	#	0.0469	0.0121	0.0314	0.0355
20	ML PLST	MLR	1st order derivatisation	-	0	0.0526	0.0265	0.0183	0.0332

ML = Matlab, PLST = PLS_Toolbox, # = not provided

Figure 3.6-1 Calibration model RMSPE values for each analyte prediction as reported by different individuals



3.6.1 Investigation of subjects predictions in the collaborative trial

For each component each subject's prediction was arranged in numerical order from high to low, the highest prediction was suspected as an outlier and Grubbs' test applied. This was repeated until the suspected subjects prediction was found not be an outlier; table 3.6.1-1 shows the outlier subject number for each metal ion.

Table 3.6.1-1 Grubbs test results for collaborative trial subjects

Component	Cobalt	Chromium	Copper	Nickel
Outlier subject	6, 5, 17, 4, 2, 14	17, 6, 5, 4, 11	17, 6, 5, 4	6 and 17

Subjects 6 and 17 lie outside of the population of the predictions in the collaborative trial, this is for each of the metal ions in the samples. These subjects chose their best model as having validation samples with between a 30 – 45 % error of the average component concentration in the training samples.

The subjects found to be outliers are consistent with those of high RMSPE values seen in figure 3.6-1. Of interest are subjects 7 and 9 who reported results that have very large errors for Ni^{2+} (and Co^{2+} for subject 7) yet they are not identified as outliers.

The Grubbs test for outliers is not ideal for the detection of many suspected values; also it only investigates the variance in the subjects for each component individually. ANOVA was used to test the effect of removing a subject taking into

account the RMSPE value of all of the metal ions present in the samples. This will determine if the subjects with all components are considered to come from the same population based on prediction of all the components.

3.6.2 Analysis of variance

An ANOVA, based on a two tailed F-test at the 95 % confidence level, was calculated based on each subjects' RMSPE value for each metal ion. This was to test if the predictions are statistically from the same population. The subjects are removed from the dataset based on the frequency of outlier predictions for the metal ions in table 3.6.1-1.

First the F-test was calculated based on the all of the individuals who participated in the trial reported RMSPE values (test 1). This found that the individuals' predictions are representative of one population. Table 3.6.2-1 reports the ANOVA results where participants' RMSPE values have been removed from the population.

Table 3.6.2-1 Collaborative trial ANOVA results

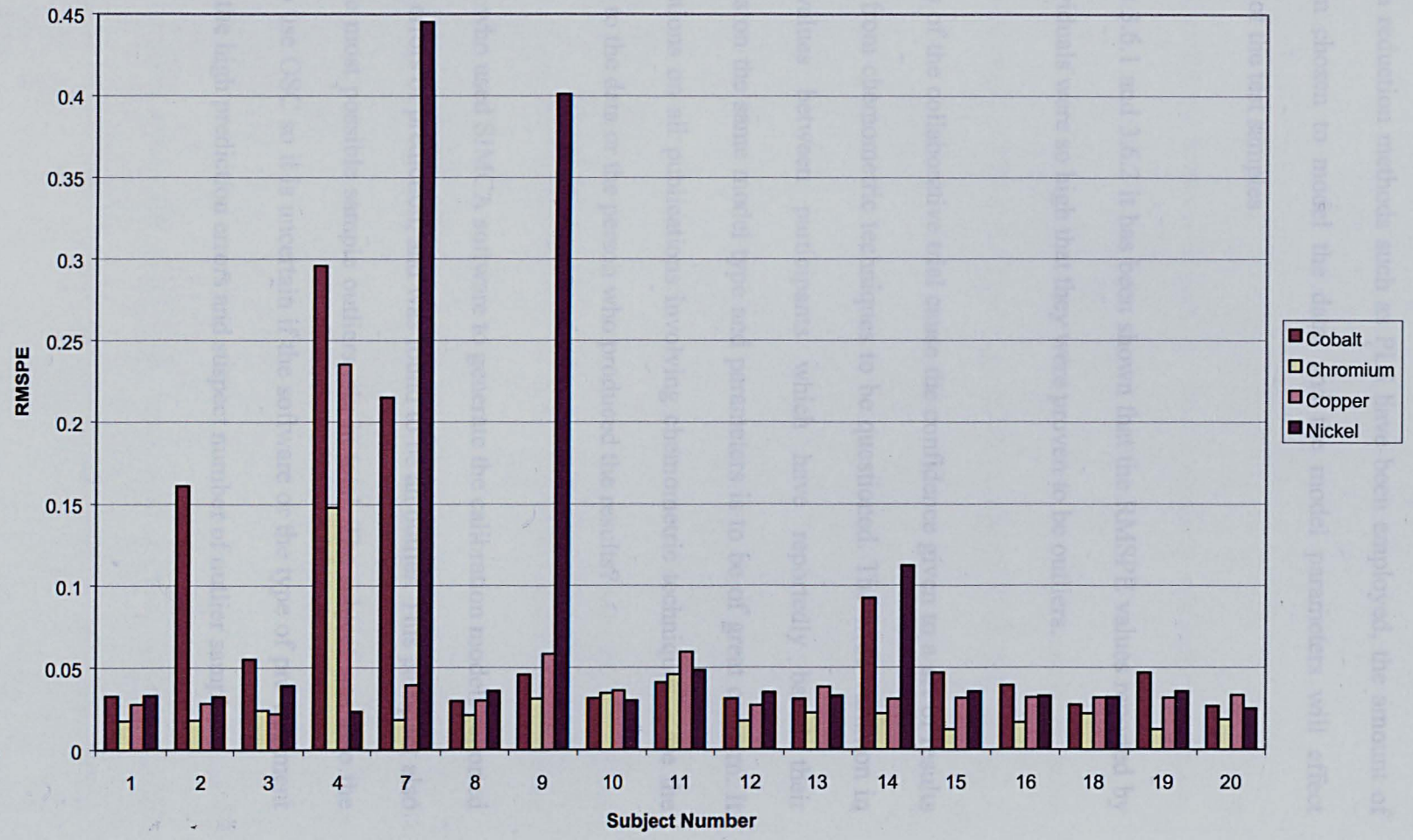
Test No.	Subjects removed	Degrees of freedom			Fcalc	Fcrit	Significant difference in population?
		Between groups	Within groups	Total			
1	None	19	60	79	9.0893	1.9636	Yes
2	6	18	57	75	13.4997	1.9973	Yes
3	6 and 17	17	54	71	7.9386	2.0343	Yes
4	6, 17 and 5	16	51	67	1.9801	2.0753	No

The results of the F-test found that after removing the RMSPE values for subjects 6, 17 and 5 the remaining subjects are statistically from the same population at the 95 % confidence level. The new dataset with outlying subject prediction removed is plotted in figure 3.6.2-1.

The removal of these subjects predictions are in accordance to those found to be outliers in section 3.6.1 from Grubbs test. Subjects 6 and 17 were outliers for all of the metal ions and subject 5 for three of the four of the metal ions. This was also the case for the results of subject 4, but this subject reported a slightly lower average RMSPE.

The effect of removing subject 4's RMSPE values was not investigated by ANOVA, as there was not a significant difference in the population after subject 5 was removed. The new trial dataset and corresponding individuals RMSPE results based on the new population can be seen in figure 3.6.2-1.

Figure 3.6.2-1 Reported RMSPE values for the new trial dataset after outlier subjects predictions have been removed.



3.6.3 Discussion of variation in prediction ability between subjects

Direct comparison of the RMSPE values reported by the individuals who participated in the trial is difficult for several of reasons. The regression method, where data reduction methods such as PLS have been employed, the amount of information chosen to model the data, and the model parameters will effect prediction of the test samples.

In sections 3.6.1 and 3.6.2 it has been shown that the RMSPE values reported by some individuals were so high that they were proven to be outliers.

The results of the collaborative trial cause the confidence given to a set of results calculated from chemometric techniques to be questioned. The wide variation in RMSPE values between participants which have reportedly based their calculations on the same model type and parameters is to be of great concern. It raises questions on all publications involving chemometric techniques – are the results true to the data or the person who produced the results?

The person who used SIMCA software to generate the calibration models reported the highest errors of prediction, and was found to be an outlier. This partipant also detected the most possible sample outliers, six in total. The subject was also the only one to use OSC so it is uncertain if the software or the type of pre-treatment has caused the high prediction errors and suspect number of outlier samples.

The second subject found to be outlying used PLS without scaling to model the data, much of the requested information such as type of PLS and number of latent variables was not provided. The high errors are reasoned to be a result of too few LVs or as software or human error, as a number of other subjects used various PLS methods with successful prediction.

In the trial results the between-subject variation is apparent in a number of cases. There are some inconsistencies in the results between users who have seemingly used the same conditions and software and yet achieve significantly different results. An example of this is the final subject with outlying RMSPEs, number 5. This subject used the same procedure as subject 14, (MLR, no pre-treatment, no outliers identified) yet resulted in substantially higher RMSPE values.

The reported results based on PLS regression method also included irregularities in the differences in prediction for the same model. The subjects who autoscaled the data and calibrated by PLS2, found outlier samples in the data, indicating that this pre-treatment may effect the data in some way. However those that used PLS1 and autoscaled did not report any outlier samples. As expected, for models with the same type of regression and pre-treatment method, as the number of LVs increased the RMSPE decreased.

The effect of the use of different software may be seen in the results of subjects 4 and 10, who have both used PLS2 with mean centred data. Subject 4 with calculations performed using Matlab has predictions several times higher than

those of subject 10 who used GRAMMS. This was for prediction of Co^{2+}, Cr^{3+} and Cu^{2+}, and there was little difference in the prediction of Ni^{2+} ion concentration.

As for the results obtained from optimising the type of pre-treatment (section 3.4), derivatisation improved the condition of the data and provided the lowest errors of prediction of the test samples. This is seen for the results of subjects 1, 8, 12, 13 and 20, which all have the lowest RMSPE values based on MLR or PLS regression. Some subjects' results have similarly low RMSPEs to those obtained through use of derivatisation, but these models are compensated by requiring a higher number of LVs, for example subject 18 with 7 LVs of mean centred data.

This collaboration has found that there is a greater variance between users than there is between regression and pre-treatment methods when producing calibration models. This is true for this data, which are very simple absorbance spectra obtained through experimental design and with very low levels of noise.

One of the aims of this trial was to produce certified results for the prediction of the validation samples. There is a problem in that the RMSPE alone is insufficient for this. The subjects were re-contacted and asked to provide the actual numeric predictions of each metal concentration for each sample which certification could be based on, a limited number obliged and these were used.

3.6.4 Estimations of test sample prediction

To produce a reference dataset, the predicted values of the test samples from 6 people were used. There were two new subjects and four from the original trial (subjects 1, 16, 19 and 20 are reported in table 3.6.4-1). All subjects had several years experience of chemometric techniques. They were asked to provide only the raw predictions of the test samples and model parameters; subjects 1 and 5 used the same type of regression model and reported identical predictions.

Table 3.6.4-1 Regression model parameters and RMSPE values for results of subjects participating in the trial to produce a reference dataset

Subject	Regression	Pre-treatment	No. LVs	RMSPE			
				Cobalt	Chromium	Copper	Nickel
1	PLS		5	0.0469	0.0121	0.0314	0.0354
2	VS-MLR*[8]	none	n/a	0.0499	0.0105	0.0303	0.0426
3	PLS2	1st savgol	4	0.0478	0.0161	0.0431	0.0354
4	MLR	1st savgol	n/a	0.0265	0.0183	0.0332	0.0251
5	PLS	none	5	0.0469	0.0121	0.0314	0.0355
6	PLS	none	6	0.0393	0.0165	0.0319	0.0326

* Conditions: selected data points: 632, 371, 887, 276, 322, squash 1 = 0.8, squash 2 = 1.35, number of iterations = 10

For all subjects, estimates are plotted against the actual concentration in figure 3.6.4-1. There is good prediction and the R^2 values are all above 0.99.

Figure 3.6.4-1 Actual Vs mean estimates of Co^{2+} prediction

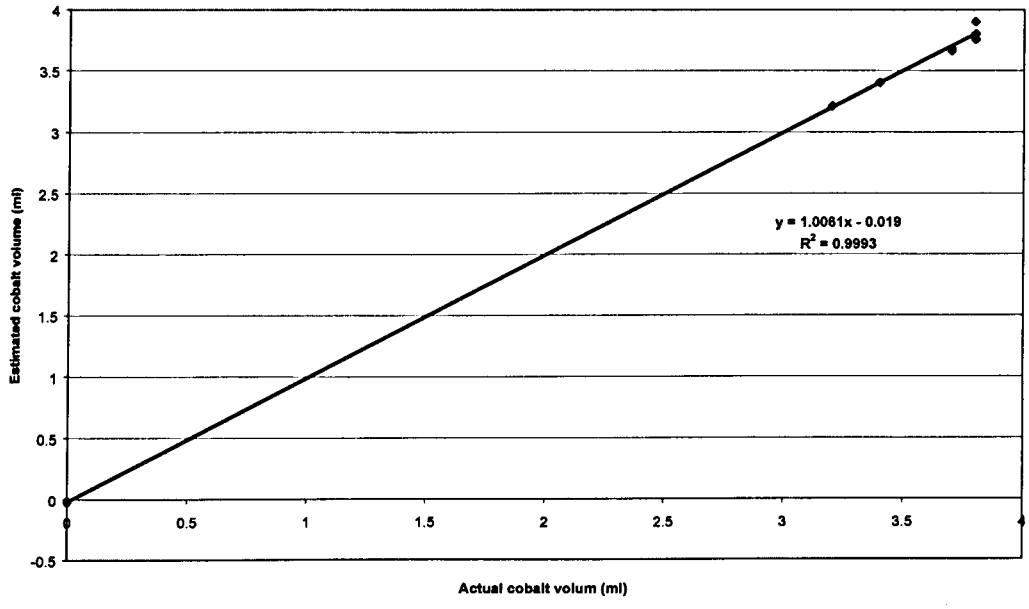


Figure 3.6.4-2 Actual Vs mean estimates of Cr^{3+} prediction

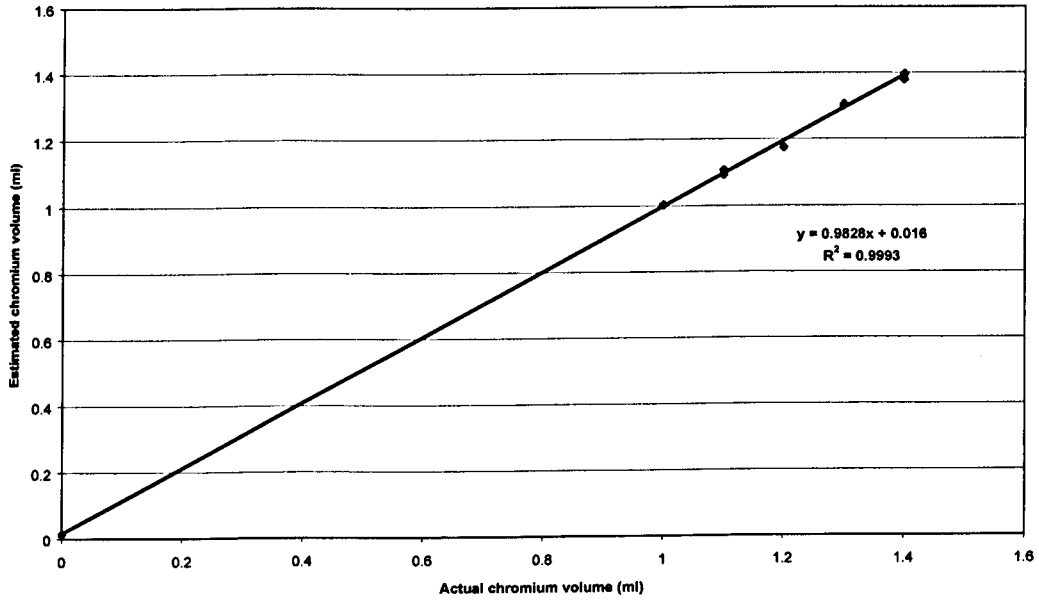


Figure 3.6.4-3 Actual Vs mean estimates of Cu^{2+} prediction

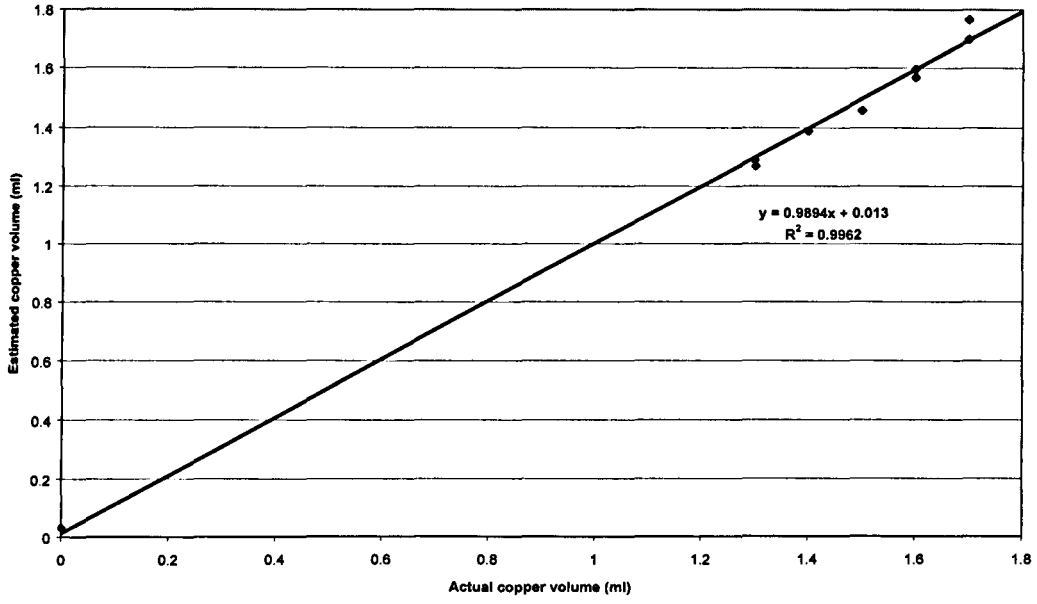
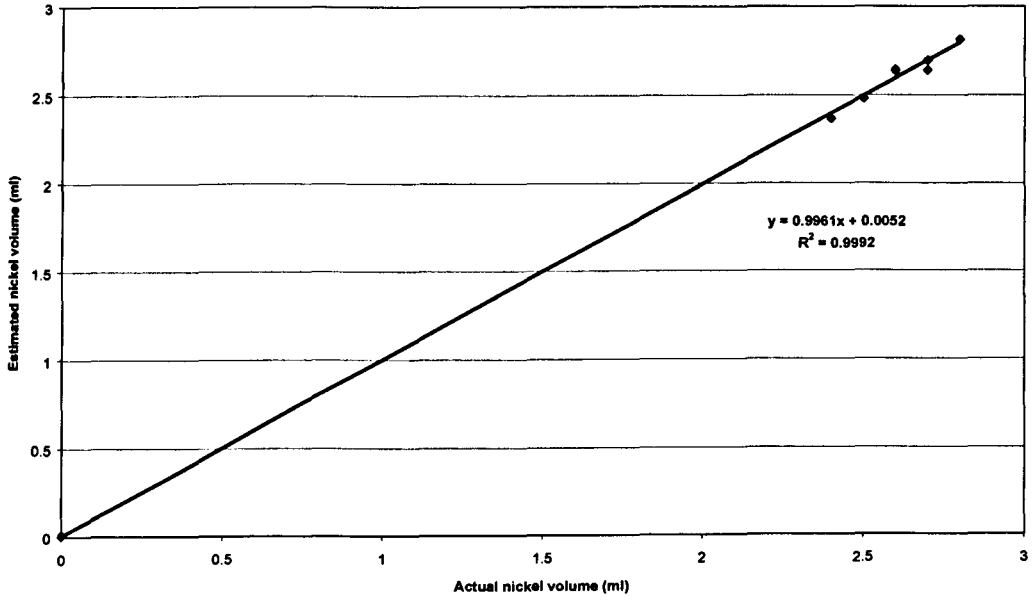


Figure 3.6.4-4 Actual Vs mean estimates of Ni^{2+} prediction



Based on the predictions the certified mean estimate and standard deviation of each component of each sample is calculated, these can be seen in table 3.6.4-2.

Table 3.6.4-2 Volume estimation (ml) and standard deviation of test sample predictions for Co^{2+}, Cr^{3+}, Cu^{2+} and Ni^{2+}

Sample No.	Co^{2+}		Cr^{3+}		Cu^{2+}		Ni^{2+}	
	Est	std	Est	std	Est	std	Est	std
17	3.2092	0.0156	1.0029	0.0042	1.7650	0.0081	2.6342	0.0066
18	3.4034	0.0059	1.3957	0.0017	1.5969	0.0037	2.8119	0.0031
19	3.6791	0.0152	1.0930	0.0045	1.6984	0.0085	2.6926	0.0340
20	3.2128	0.0144	1.0029	0.0030	1.2893	0.0059	2.6959	0.0132
21	3.6605	0.0150	1.1748	0.0041	1.8167	0.0083	2.4798	0.0216
22	3.8022	0.0093	1.3046	0.0091	1.3880	0.0162	2.3636	0.0187
23	-0.0122	0.0191	1.1070	0.0139	1.2695	0.0138	0.0048	0.0167
24	3.7549	0.0186	0.0141	0.0157	1.5695	0.0291	2.6414	0.0254
25	3.9015	0.0344	1.3009	0.0046	0.0316	0.0080	2.6394	0.0077
26	-0.0275	0.0245	1.3785	0.0077	1.4580	0.0206	0.0079	0.0188

4 Conclusions

The simulated spectra of the mixtures of metal ions in aqueous solution, where the absorbances were calculated according to Beer's law from the spectra of the individual metal ions, were found to be comparable to spectra of the samples measured. This confirmed that all the samples would have absorbances within the linear range of response for visible spectroscopy, prior to preparation and measurement of the samples. The spectra of the example and theoretical samples were measured up to 900 nm; for the actual samples, this range was extended to 1000 nm.

The industrial spectrometer did not cover the entire visible range; the samples were recorded between 350 – 738 nm, cutting off the region between 738 and 1000 nm where the maximum absorbance for copper and some absorbance for nickel occur. A gross error was identified in set 2 recorded with the industrial spectrometer. When the same samples from datasets 2.1 and 2.2 were plotted they were found to be different. This is thought to be a result of preparation or mislabelling error. Initial PLS models gave accurate prediction of dataset 2.2. For dataset 2.1, the percentage variance captured by the model and prediction were very poor. Based on this information, dataset 2.1 was discarded and dataset 2.2 used for the remaining work involving the spectra recorded with the industrial spectrometer.

The datasets recorded using the industrial process spectrometer gave higher errors of prediction than those from 'like-for-like' calibration models based on the

analytical spectrometers' datasets. Unlike for a standard laboratory spectrometer, the samples recorded with the process spectrometer were not isolated from background variation of light by placing the cuvette within the spectrometer for analysis. Instead the samples are analysed in cuvettes on the bench, with a background spectrum recorded on commencement of analysis. The assumption that the influence of background light does not vary during this time could be incorrect and may attribute to the higher errors of prediction for these samples, given that the resolution was found not to be a significant factor when set 3 was reduced to the same resolution without causing a substantial increase in prediction error.

The collaborative trial results have shown that prediction between users varies significantly, consistent with the results of the forage data trial[3]. The increased number of participants for the metal ions data, 20 compared to 6, for the forages did not reduce the spread in the results. Both of the studies have confirmed the need for uniformity between users for confidence in any calibration models.

For the trial, all of the participants used multivariate analysis for calibration as opposed to univariate calibration. As the samples contain 4 analytes at least 4 wavelengths would be required for univariate analysis, the choice of which wavelengths to select is not obvious as there is co-absorbance between the species through the spectra. It is preferable to apply multivariate methods as these will select the most appropriate spectral regions for significant contribution for the model.

A second trial was also carried out, this produced estimates and standard deviations for prediction of the validation samples. This trial consisted of 6 subjects whose predictions of the metal ions were consistent.

5 Further Work

The objective of producing a dataset with minimal experimental errors has been achieved with the samples measured with the analytical spectrometer. A dataset based on measurement from an industrial spectrometer has not been achieved. For this to be possible the reason why poor measurements were recorded using the industrial spectrometer needs to be established. Initially it was thought that the resolution of the spectrometer was the fault. This was proven untrue when the dataset measured on the analytical spectrometer did not have such high errors when the resolution was simulated. General repeatability and reproducibility measurements should be performed on the industrial spectrometer to ensure its correct operation for measurement.

The collaborative study should be repeated using the same dataset, but with more consideration of the subjects chosen with the following points:

- Subjects should be chosen from preference of software, this will confirm if certain software produces different predictions.
- Gaining exact information of the software used: in the study, subjects stated Matlab as the calculation tool, but this does not automatically perform PLS regression; for some subjects it is unknown if bespoke algorithms or a toolbox was used for the calculations.

6 References

[1] http://www.itl.nist.gov/div898/strd/general/bkground.html Viewed: 7/2/03.

[2] http://www.itl.nist.gov/div898/strd/general/dataarchive.html Viewed 20/2/03.

[3] Ruisanchez I, Rius F.X, Maspoch S, Coello J, Azzouz T, Tauler R, Sarabia L, Ortiz M.C, Fernandez J.A, Massart D, Puigdomenech A, Garcia C, 'Preliminary results of an interlaboratory study of chemometric software and methods on NIR data. 'Predicting the content of crude protein and water in forages', Chemometrics and Intelligent Laboratory Systems, 2002, (63), 93 – 105.

[4] UV Spectrometry Group, 'UV Spectroscopy: Techniques, Instrumentation and Data Handling', Volume 4, 1st Edn., Chapman and Hall, London, 1993.

[5] Skoog D.A, Holler F.J and Niemen T.A, 'Principles of Instrumental Analysis', 5th Edition, Saunders, Philadelphia, 1998, 338.

[6] Skoog D.A, Holler F.J and Niemen T.A, 'Principles of Instrumental Analysis', 5th Edition, Saunders, Philadelphia, 1998, 301 - 303.

[7] http://www.galactic.com/ Viewed: 7/2/03.

[8] Walmsley A. D, 'Improved variable selection procedure for multivariate linear regression', Analytica Chimica Acta, 1997, 354 (1-3), 225 – 232.

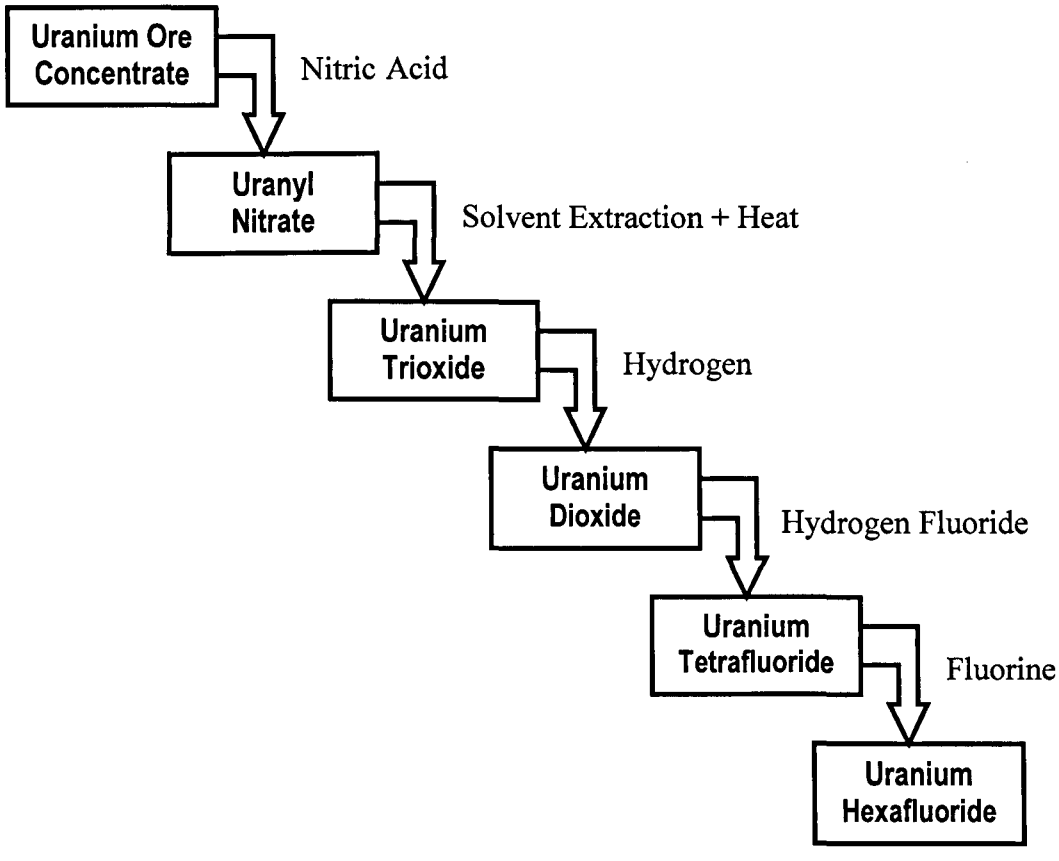
Chapter 3

Analysis of Uranyl Nitrate Liquors by UV/Vis and Raman Spectroscopies

1 Introduction

Uranium ore concentrate is converted to uranium hexafluoride in the nuclear industry by the process shown in figure 1-1, the uranium ore concentrate conversion method[1].

Figure 1-1 Uranium ore concentrate conversion



During the first step of the process, uranyl nitrate liquor ($UO_2(NO_3)_2$) is produced by dissolving uranium ore concentrate (U_3O_8) in nitric acid. In uranyl nitrate the oxidation state of uranium is U^{6+} and the uranium ore concentrate a mixture of U^{6+} and U^{4+}.

The temperature of this process is uncontrolled and can vary between ambient and 90 °C. The uranium ore concentrate is impure and the exact level of uranium content unknown. Currently the uranyl nitrate liquors are analysed by X-ray fluorescence spectrometry (XRF)[2].

The aim of this work is to investigate alternative methods of analysis that could be implemented for on-line monitoring of this process for nitrate (NO_3^-) and uranyl ion (UO_2^{2+}) concentration[3]. In addition to the benefits of improved process control, there is a clear safety advantage of performing this analysis on-line as the process stream is radioactive.

UV/Vis and Raman spectroscopies have been identified by BNFL as suitable analysis methods. Both have been utilised for on-line monitoring commercially and have been employed for similar analysis. UV/Vis spectroscopy is the more desirable method as the instrumentation and implementation for an on-line analysis system is considerably cheaper and less troublesome than for Raman spectroscopy.

The UV/Vis absorption spectrum of uranyl nitrate solution was documented in 1964[4]. The absorption spectrum was plotted in the wavelength region of 340 – 500 nm, the maximum molar extinction coefficient around 430 nm.

Second derivative spectrophotometry has been used for the determination of uranium in the presence of iron in 1M nitric acid and also in yellow cake,

magnesium diuranate in this case[5]. 'Yellow cake' is a term used loosely in the nuclear industry to describe impure uranium ores ranging from ammonium, magnesium, potassium and calcium diuranates through to nearly pure uranium concentrate of U_3O_8[6]. The spectra were recorded between 360 – 480 nm, and as the pre-study samples contained iron, a wavelength region where iron did not absorb but uranium did was sought. This was found to occur at 408.2 nm on the second derivative spectra, by visually comparing the spectra of samples with and without iron. This gave a relative error of 1 %. Yellow cakes samples in the range of 7 – 25 mg/g were analysed and at 408.2 nm gave an error of 6 – 8 %, which was reduced to 1.6 % when the yellow cake was purified prior to spectrophotometric analysis.

Micro-Raman spectroscopy was used to characterize uranium oxides in-situ[7]. Micro-Raman spectroscopy uses an optical microscope interfaced to a Raman spectrometer; the powder samples were placed on a sample slide under the microscope that focused on a spot size of 3 μm.

The spectrum of uranyl nitrate was recorded between 600 – 3500 cm^{-1} and the spectrum described to contain peaks that are characteristic of the uranyl ion and the nitrate ion. The uranyl ion was shown to dominate at 880 cm^{-1} and the paper considers the effect of temperature of the uranyl nitrate spectra and discusses changes in the peak intensity between 25 - 100 °C. The elevated temperature was achieved by flowing hot air over the sample. This work is a good reference to demonstrate the feasibility of Raman spectroscopy for analysis of uranyl nitrate,

but precise methods such as this requiring the use of a microscope cannot easily be transferred to the on-line analysis of solutions.

The purpose of this work is to build multivariate calibration models from the UV/Vis and Raman spectra of uranyl nitrate liquor in order to predict uranyl and nitrate concentration, which are insensitive to temperature. The samples' spectra were recorded in the temperature range of 25 – 90 °C, in 5 °C increments. To do this calibration models were first produced at one fixed temperature, 50 °C (the typical operating temperature), and then additional models including temperature as a prediction variable was investigated.

The experiments were designed and performed by personnel from the BNFL, Springfields site, and only the author undertook the data analysis.

2 Experimental

Due to the hazardous nature of the samples, BNFL personnel carried out all of the experimental work at the BNFL Springfields site. Once the measurements were recorded, the UV/Vis and Raman spectral files, and sample compositions were forwarded for data processing. A summary of the sample spectra and composition can be seen in table 1-1.

The immersion probe for each spectrometer was placed within a reactor vessel (3 l). The samples were prepared by adding nitric acid to uranium oxide powders until all the powder had dissolved to make uranyl nitrate liquor. The liquor was heated to 95 °C. Once the liquor had cooled to 90 °C, 10 repeat spectra were recorded, and this procedure was continued at 5 °C intervals until the temperature of the mixture reached 25 °C. After spectroscopic analysis, the % uranyl and nitrate concentration of each sample was determined by XRF spectrometry.

2.1 Sample analysis

2.1.1 Analysis by UV/Vis spectroscopy

UV Spectrometer:	Ocean Optics S2000, dual halogen source.
Probe:	Hellma ATR probe with a 3 bounce crystal.
Operating software:	not provided
Scan range:	not provided
Spectral files:	*.txt
Number of data points:	2047

2.1.2 Analysis by Raman spectroscopy

Raman Spectrometer:	Kaiser Raman instrument.
Probe:	Kaiser immersion probe (type; IMO-0.1-10.5)
Operating software:	Hologram
Scan range:	50 – 3000 cm^{-1}
Spectral files:	*.spc
Number of data points:	9834

2.2 Preparation of spectra files for data processing

The uranyl nitrate samples spectra provided in *.txt and *.spc formats were opened and converted to *.mat files using SIMCA-P 9.0 (Umetrics Ltd, Woodside Road, Winkfield, Windsor, Berkshire, SL4 2DX). The *.mat files can then be used in Matlab for data processing.

The UV spectra were supplied as raw incident radiation (air background). This was transferred to absorbance spectra using equation 15. Where *Po* and *P* are the intensity of the incident and transmitted radiation, respectively. In this case *Po* was the incident measurement through air.

$$A = \log\left(\frac{Po}{P}\right)$$

Equation 15 Calculation of absorbance spectra[8]

Table 2.2-1 Description of uranyl nitrate samples and % uranyl and nitrate determined by XRF spectrometry

Sample	UV/Vis Spectra ID	Raman Spectra ID	Nitrate (% w/w)	Uranyl (% w/w)
1	SCD98_01	NCOD 01	10.75	16.59
2	SCD98_02	NCOD 02	20.35	17.65
3	SCD98_03	NCOD 03	0.35	17.20
4	SCD98_04	NCOD 04	8.55	21.54
5	SCD98_05	NCOD 05	11.21	20.03
6	SCD98_06	NCOD 06	0.28	20.89
7	SCD98_07	NCOD 07	14.92	20.54
8	SCD98_08	NCOD 08	15.12	22.96
9	SCD98_09	NCOD 09	8.05	18.86
10	SCD98_10	NCOD 10	2.44	23.73
11	SCD98_11	NCOD 11	5.25	25.72
12	SCD98_12	NCOD 12	1.71	26.62
13	SCD98_13	NCOD 13	5.31	28.35
14	SCD98_14	NCOD 14	1.25	28.90
15	SCD98_15	NCOD 15	5.19	29.57
16	SCD98_16	NCOD 16	0.38	31.58
17	SCD98_17	NCOD 17	14.57	11.26
18	SCD98_18	NCOD 18	9.40	11.24
19	SCD98_19	NCOD 19	4.71	13.79
20	SCD98_20	NCOD 20	9.74	16.99
21	SCD98_21	NCOD 21	0.46	12.48
22	SCD98_22	NCOD 22	22.83	12.86
23	SCD98_23	NCOD 23	14.58	15.26
24	SCD98_24	NCOD 24	19.18	14.38
25	SCD98_25	NCOD 25	20.89	20.78
26	SCD98_26	NCOD 26	20.35	23.15
27	SCD98_27	NCOD 27	22.20	26.66
28	SCD98_28	NCOD 28	24.34	21.64
29	SCD98_29	NCOD 29	unknown	unknown
30	SCD98_30	NCOD 30	16.08	26.46
31	SCD98_31	NCOD 31	18.20	25.89
32	SCD98_32	NCOD 32	0.07	31.89
33	SCD98_33	NCOD 33	17.06	27.85
34	SCD98_34	NCOD 34	12.22	23.75
35	SCD98_35	NCOD 35	4.94	24.12
36	SCD98_36	NCOD 36	11.12	27.09
37	SCD98_37	NCOD 37	16.49	18.82
38	SCD98_38	NCOD 38	2.98	19.37
39	SCD98_39	NCOD 39	11.06	24.85
40	SCD98_40	NCOD 40	5.25	17.71
41	SCD98_41	NCOD 41	5.04	21.86

3 Results

3.1 Investigation of Samples

Of the original 41 samples only 31 were found to contain all of the correct number of spectra and reference information, these are summarised in first row of tables 3.1-1 and 2. The second row identifies the samples that were incomplete; these were discarded from the dataset. The 31 samples were different for the UV/Vis and Raman datasets.

Table 3.1-1 Summary of samples measured by UV/Vis spectroscopy used for data processing

Sample included for data processing	1, 2, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23, 24, 26, 27, 28, 30, 34, 35, 36, 37, 38, 40
Sample removed	3, 9, 17, 25, 29, 31, 32, 33, 39, 41

Table 3.1-2 Summary of samples measured by Raman spectroscopy used for data processing

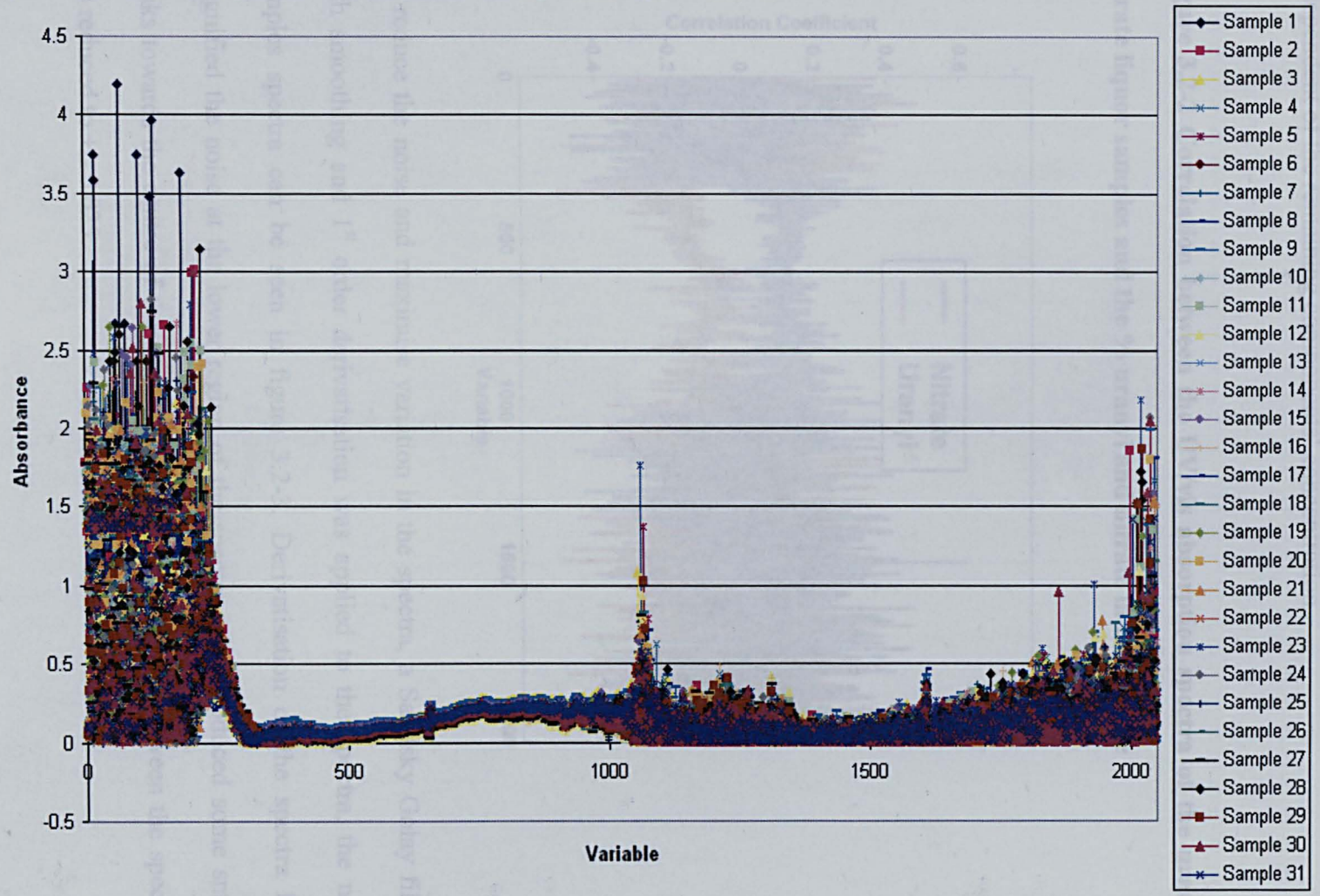
Sample included for data processing	1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 24, 26, 30, 33, 35, 36, 37, 38, 39, 40, 41
Sample removed	7, 17, 23, 25, 27, 28, 29, 31, 32, 34

3.2 Analysis by UV/Vis Spectroscopy

The sample spectra recorded by UV/Vis spectroscopy are investigated first. The spectra recorded when the sample temperature was 50 °C are plotted in figure 3.2-1. The exact wavelength region of the UV/Vis spectrum recorded is unavailable therefore the spectra are plotted in terms of variables instead of wavelength.

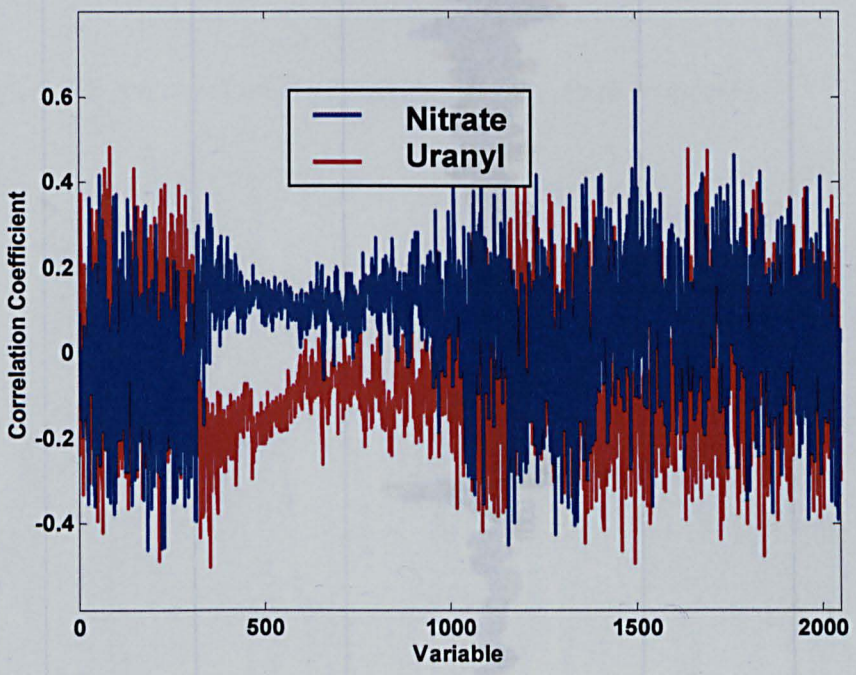
There are high levels of noise at both ends with of the spectra with little absorbance that looks to vary with component concentration.

Figure 3.2-1 UV/Vis spectra of uranyl nitrate liquor samples measured at 50 °C



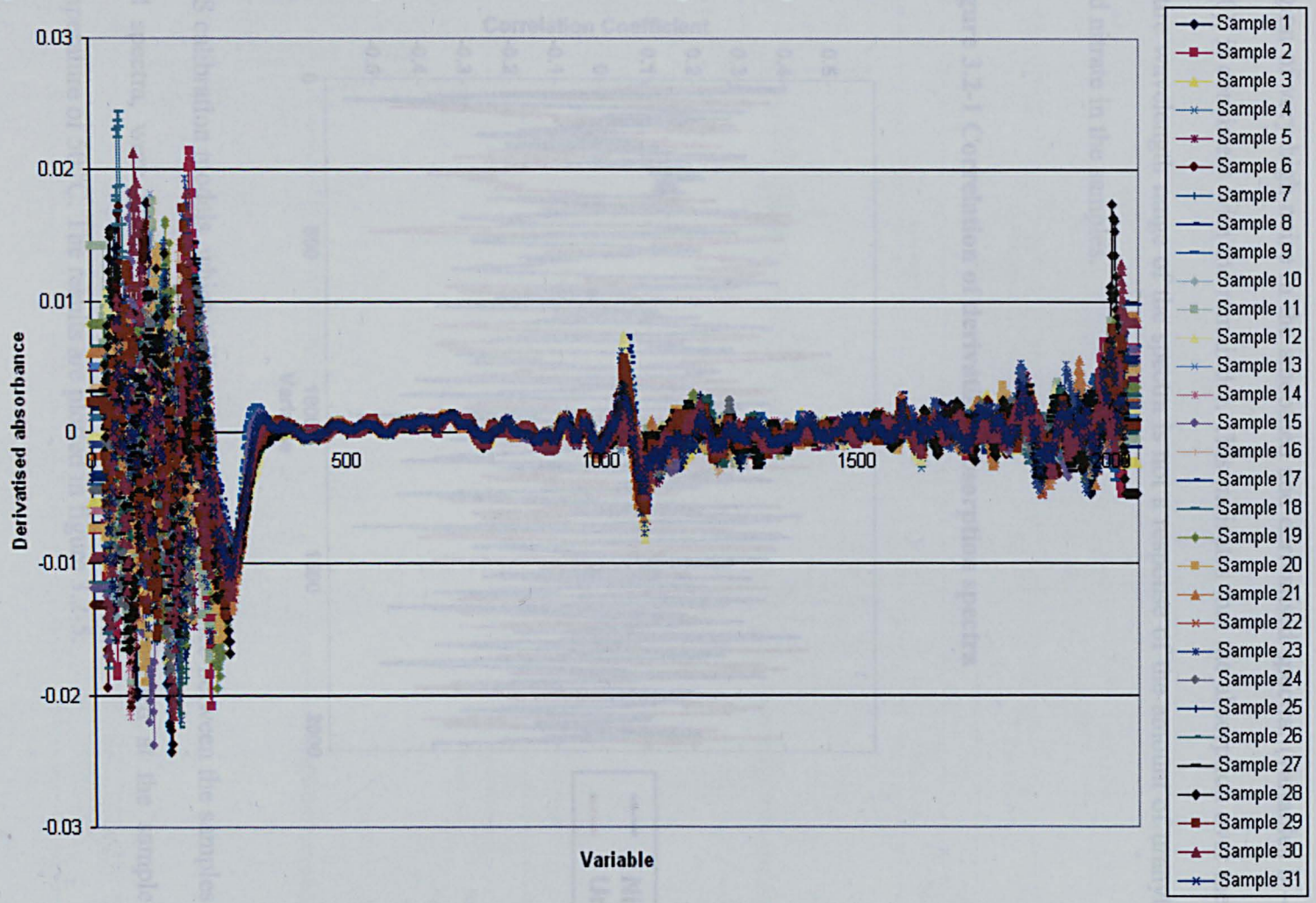
The typical correlation level between the sample concentrations and the absorption spectra is +/- 0.2 for both components (see figure 3.2-2). The shape of the correlation coefficient graph for the uranyl and nitrate levels in the samples reflects each other. This indicates that the components are serially correlated and predictions may not be independent of the remaining components' concentration.

Figure 3.2-2 Correlation between the UV/vis absorption spectra of the uranyl nitrate liquor samples and the % uranyl and nitrate in the sample



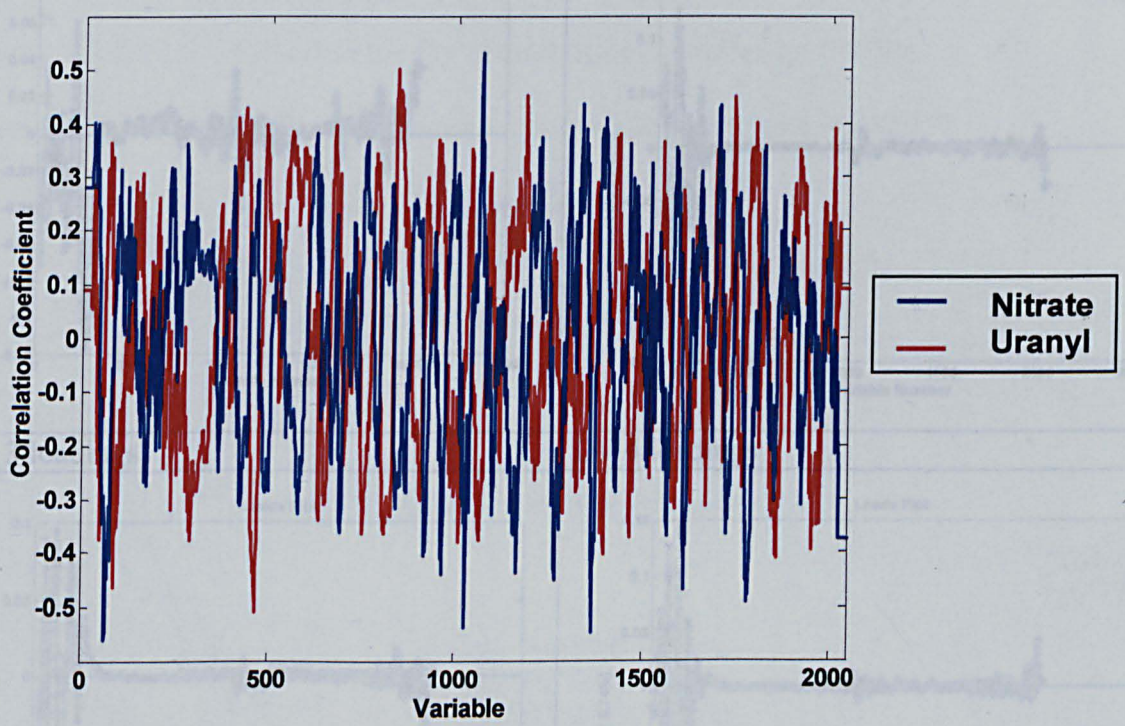
To reduce the noise and maximise variation in the spectra, a Savitsky Golay filter with smoothing and 1st order derivatisation was applied to the spectra, the new samples spectra can be seen in figure 3.2-3. Derivatisation of the spectra has magnified the noise at the lower region of the spectra and introduced some small peaks towards the centre of the spectra. The scale of variation between the spectra has reduced to +/- 0.025.

Figure 3.2-3 Derivatised UV/Vis spectra of uranyl nitrate liquor samples measured at 50 °C



Based on the derivatised spectra in figure 3.2-3, the maximum correlation between the spectra and concentration of uranyl and nitrate in the samples, increases to +/- 0.5. This correlation is only for certain variables, but generally the correlation for both of the components resembles random noise. Correlation in the region of very high noise in the smoothed and derivatised spectra, (variables 0 – 250), is consistent with the remainder. This indicates that the absorption over the entire wavelength range of the spectra is not a response of the amount of uranyl and nitrate in the samples.

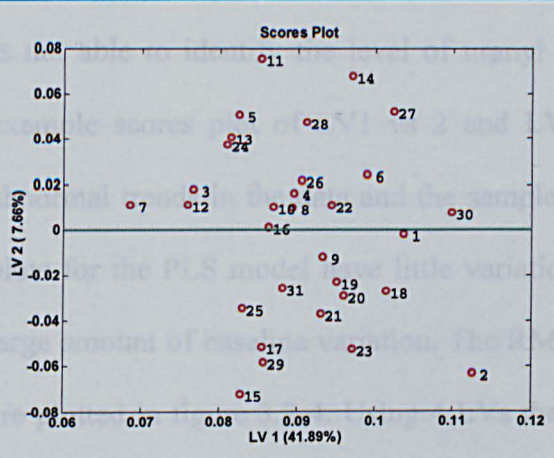
Figure 3.2-1 Correlation of derivatised absorption spectra



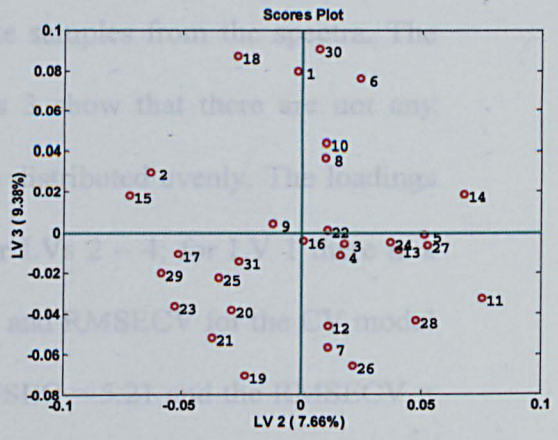
PLS calibration models, which will maximise the covariance between the samples and spectra, were produced with the filtered spectra recorded at the sample temperature of 50 °C. The results are plotted in figure 3.2-5.

Figure 3.2-5 PLS results based on 1st derivative spectra, of uranyl liquor samples by UV/Vis spectroscopy at 50 ^{0}C; scores and variable loadings plots

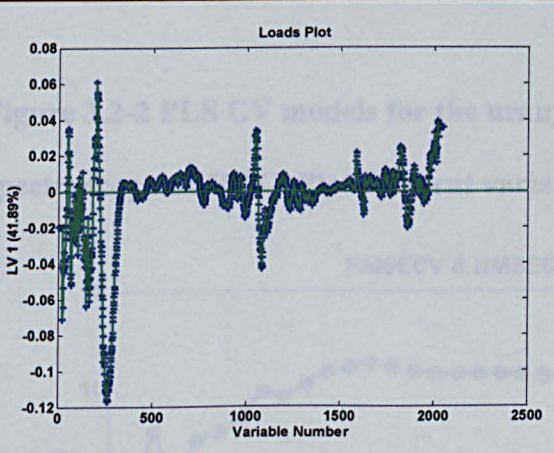
Scores plot of LV1 against LV 2



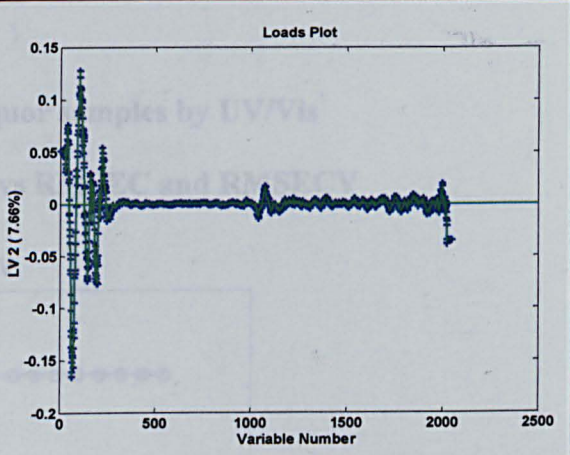
Scores plot of LV2 against LV 3



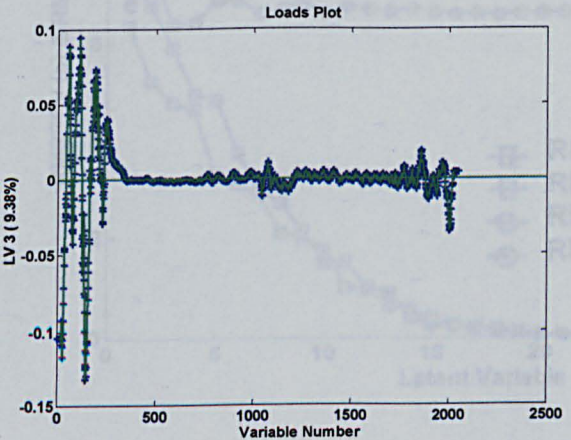
LV 1 loadings



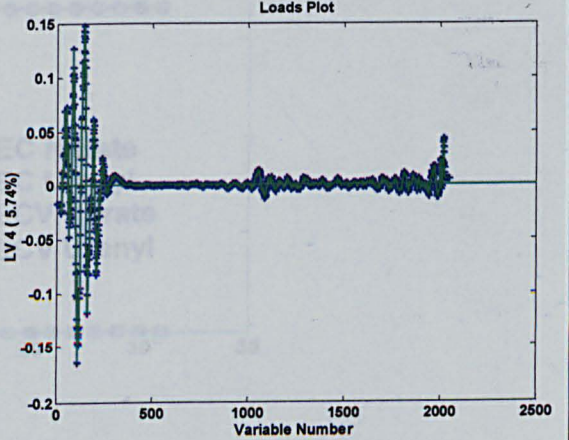
LV 2 loadings



LV 3 loadings

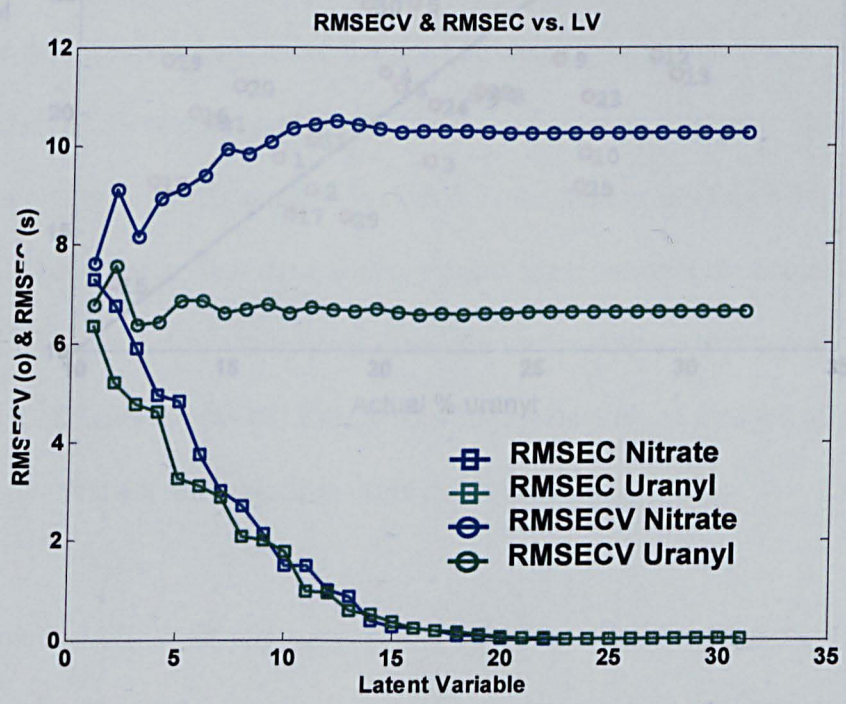


LV 4 loadings



The PLS model was calculated with the derivatised spectra. The sample number in the scores plots is in order from low to high % concentration of uranyl. As the sample scores are not in numerical order for any of the first 3 LVs the CV model is not able to identify the level of uranyl in the samples from the spectra. The example scores plot of LV1 vs 2 and LV2 vs 3 show that there are not any abnormal trends in the data and the samples are distributed evenly. The loadings plots for the PLS model have little variation for LVs 2 – 4; for LV 1 there is a large amount of baseline variation. The RMSEC and RMSECV for the CV model are plotted in figure 3.2-4. Using 4 LVs the RMSEC = 5.21 and the RMSECV = 7.56.

Figure 3.2-2 PLS CV models for the uranyl liquor samples by UV/Vis spectroscopy at 50 °C; Plot of latent variables vs RMSEC and RMSECV



The actual against predicted % concentration for uranyl and nitrate in the samples are plotted in figure 3.2-3. These show significant spread in the predictions with only a few concentrations for each component being accurately predicted. The predictions for nitrate levels seem to have an 'S' shaped structure indicating that there is some sort of serial correlation in the predictions. The actual Vs predicted plot show a general trend in the ability to predict an unknown sample.

Figure 3.2-3 PLS model (4 LVs) results for the uranyl liquor samples by UV/Vis spectroscopy at 50 °C; plots of actual Vs predicted % uranyl

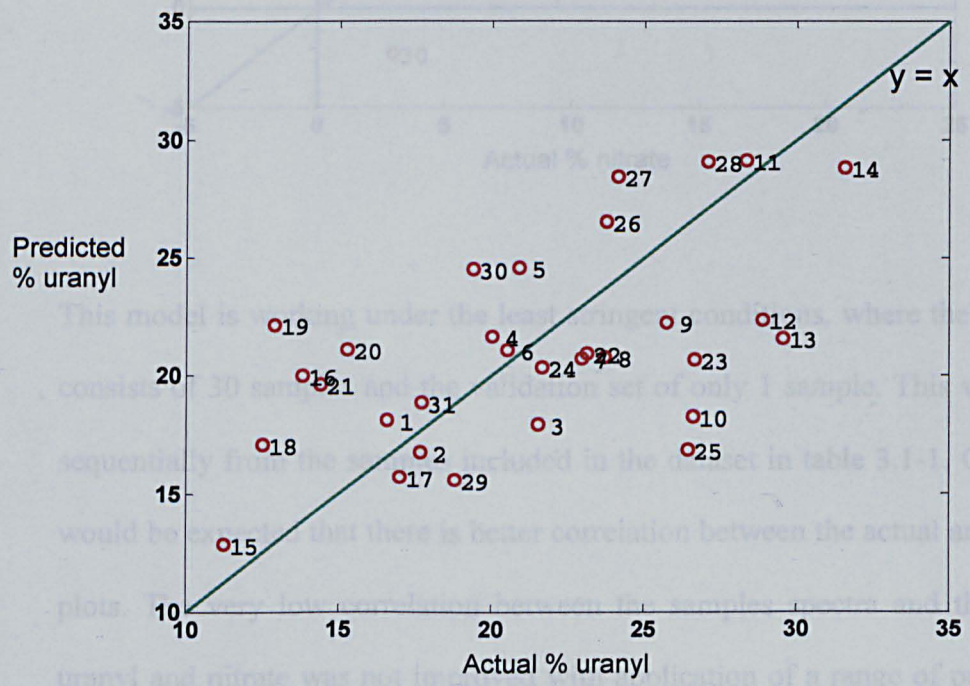
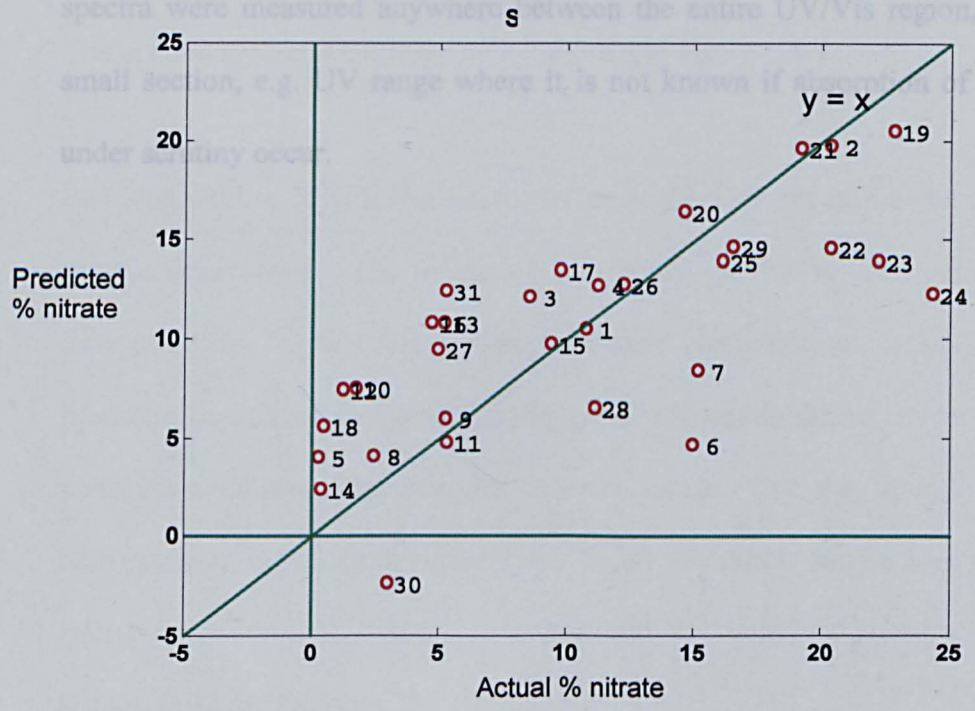


Figure 3.2-8 PLS model (4 LVs) results for the uranyl liquor samples by UV/Vis spectroscopy at 50 °C; plots of actual Vs predicted % nitrate



This model is working under the least stringent conditions, where the training set consists of 30 samples and the validation set of only 1 sample. This was selected sequentially from the samples included in the dataset in table 3.1-1. Given this it would be expected that there is better correlation between the actual and predicted plots. The very low correlation between the samples spectra and the levels of uranyl and nitrate was not improved with application of a range of pre-treatment methods that are not included in this document.

Mechanical failure of the spectrometer could have lead to the poor measurements of the samples. The literature has shown measurement between 360 – 480 nm allowed for the determination of uranyl in samples[5]. It is difficult to compare the

spectra recorded to those seen in literature because the wavelength range and resolution was not documented at the time of analysis. It is possible that, the spectra were measured anywhere between the entire UV/Vis region, and just a small section, e.g. UV range where it is not known if absorption of the species under scrutiny occur.

3.3 Analysis by Raman Spectroscopy

The uranyl nitrate liquor samples were also monitored by Raman spectroscopy. The spectra are good quality, and are plotted in figure 3.3-1.

Previous work at BNFL has shown the region below 500cm^{-1} to be attributed to sample fluorescence. The remaining two peaks are due to each component, the first peak has highest correlation to uranyl concentration, as expected from literature, and the second peak has highest correlation to nitrate concentration. The correlation between the samples Raman spectra and the uranyl and nitrate concentration is plotted in figure 3.3-2. From this plot it can be seen that there is strong correlation, R = 0.8 – 0.9, between the sample concentration and the Raman intensity between 750 - 1250cm^{-1}.

Figure 3.3-1 Uranyl nitrate liquor samples Raman spectra measured at 50 °C

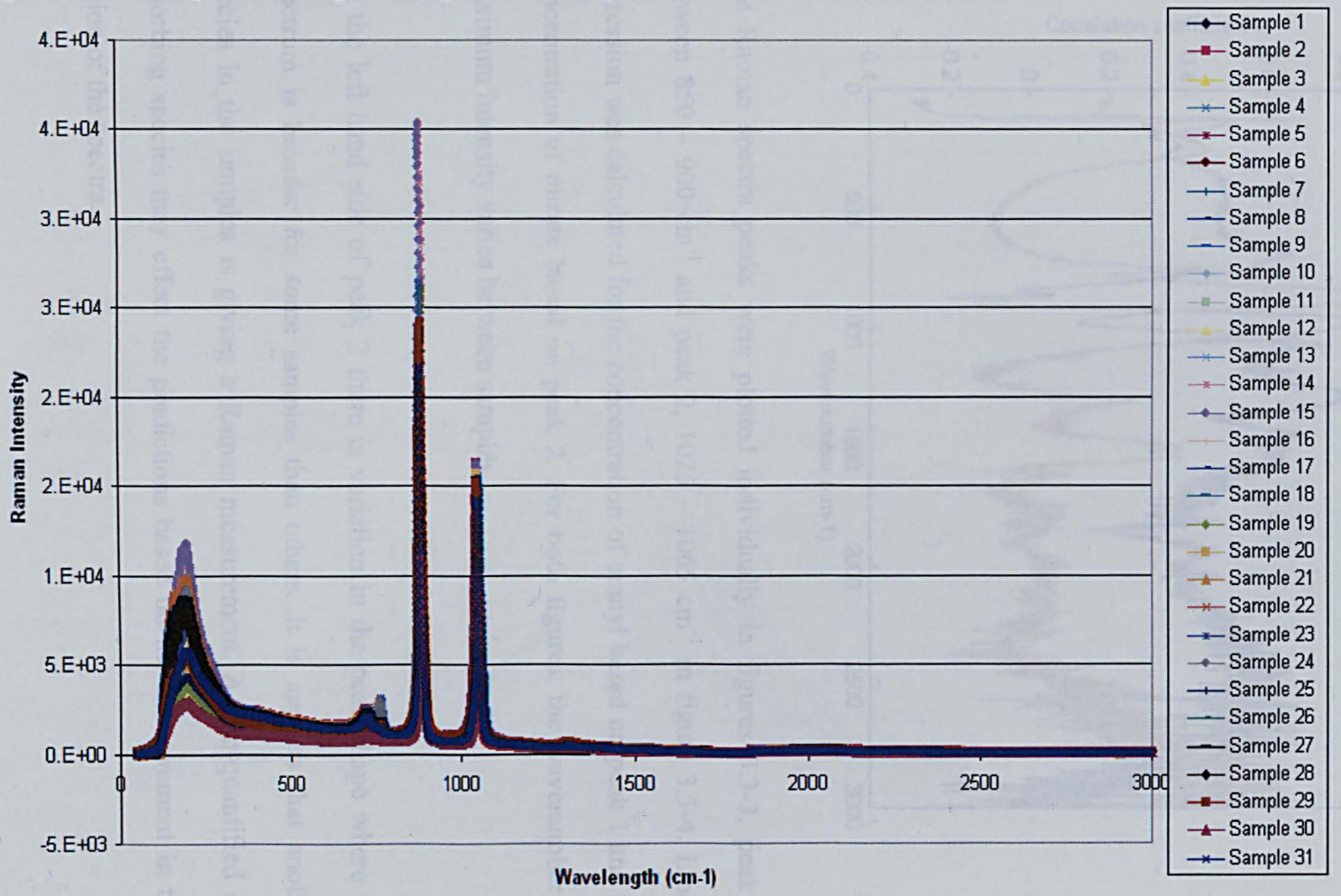
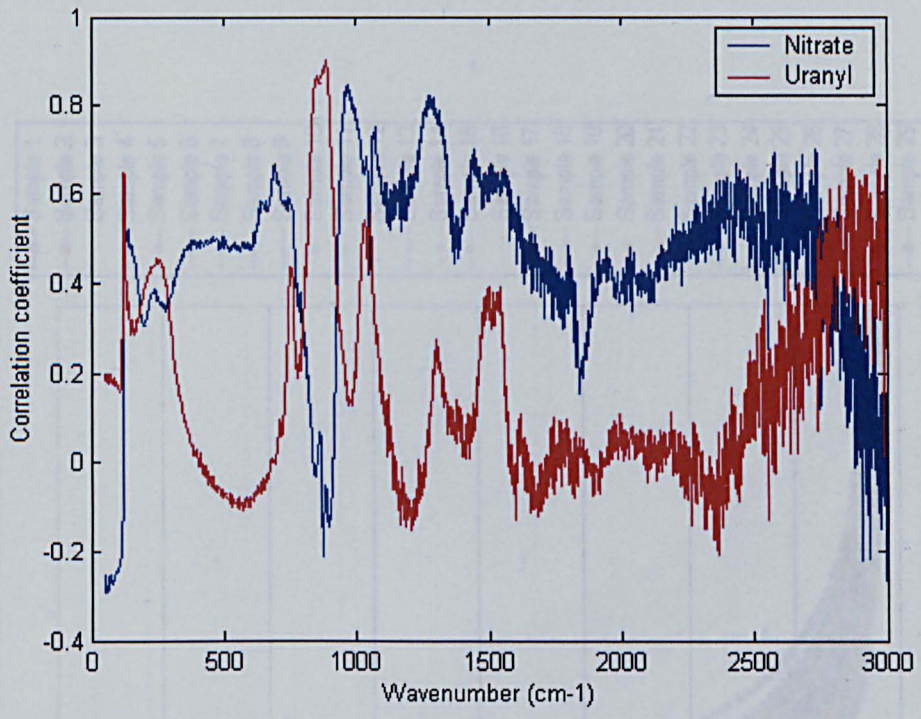


Figure 3.3-2 Correlation coefficients between the Raman spectra at 50 °C of uranyl nitrate liquor samples and sample uranyl and nitrate concentration



The Raman spectra peaks were plotted individually in figures 3.3-3, peak 1, between 850 – 900 cm^{-1} and peak 2, 1025 – 1065 cm^{-1} in figure 3.3-4. Linear regression was calculated for the concentration of uranyl based on peak 1 and the concentration of nitrate based on peak 2. For both figures, the wavenumber of maximum intensity varies between samples.

On the left hand side of peak 2 there is variation in the peak shape where the spectrum is broader for some samples than others. It is appears that another species in the samples is giving a Raman measurement. A non-quantified co-absorbing species may effect the predictions based on the measurement in this region of the spectra.

Figure 3.3-3 Raman spectra of uranyl nitrate liquors, 'Peak 1', 850 – 900 cm^{-1} measured at 50 °C

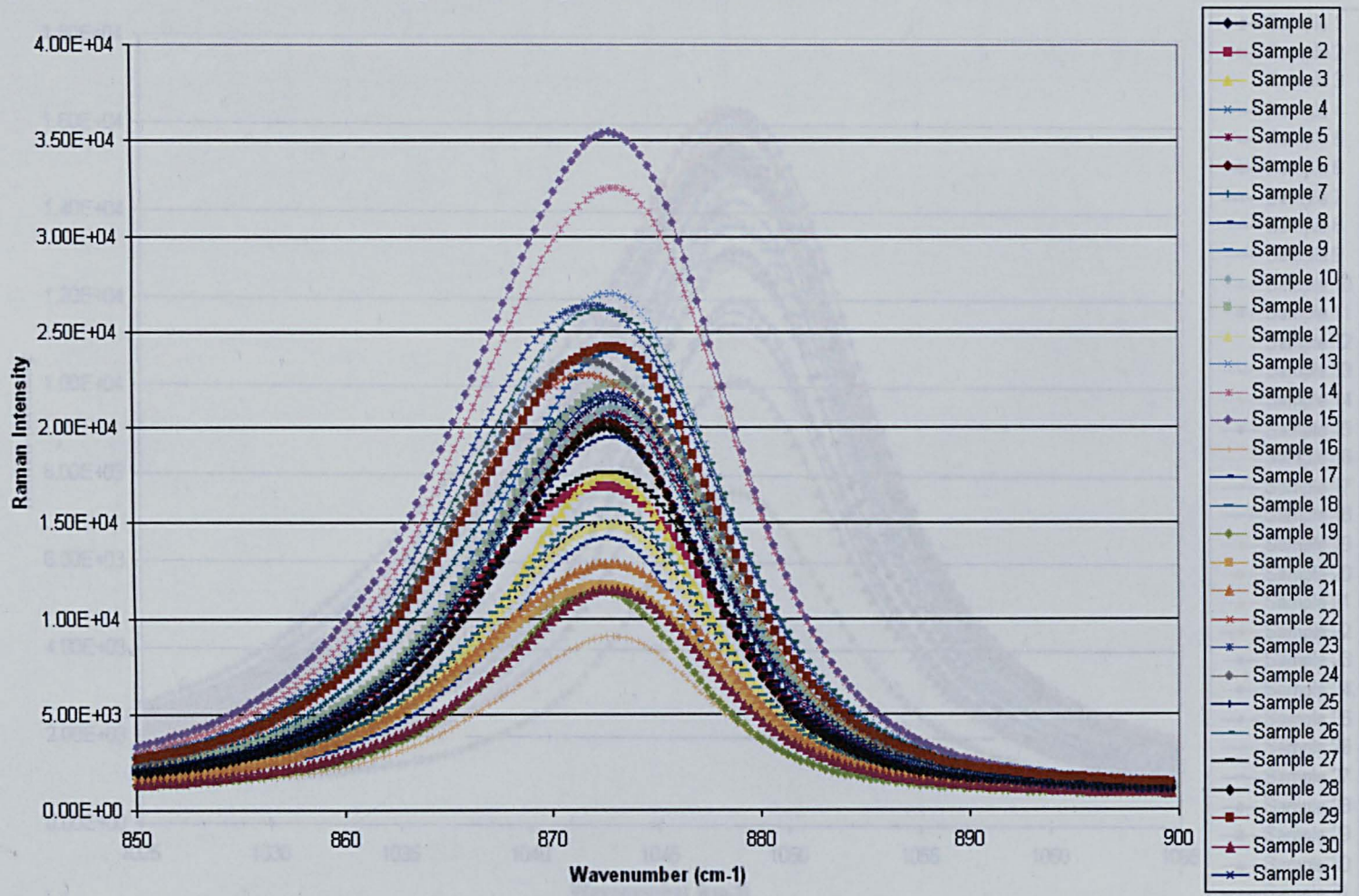


Figure 3.3-4 Raman spectra of uranyl nitrate liquors, 'Peak 2', 850 – 900 cm^{-1} measured at 50 °C

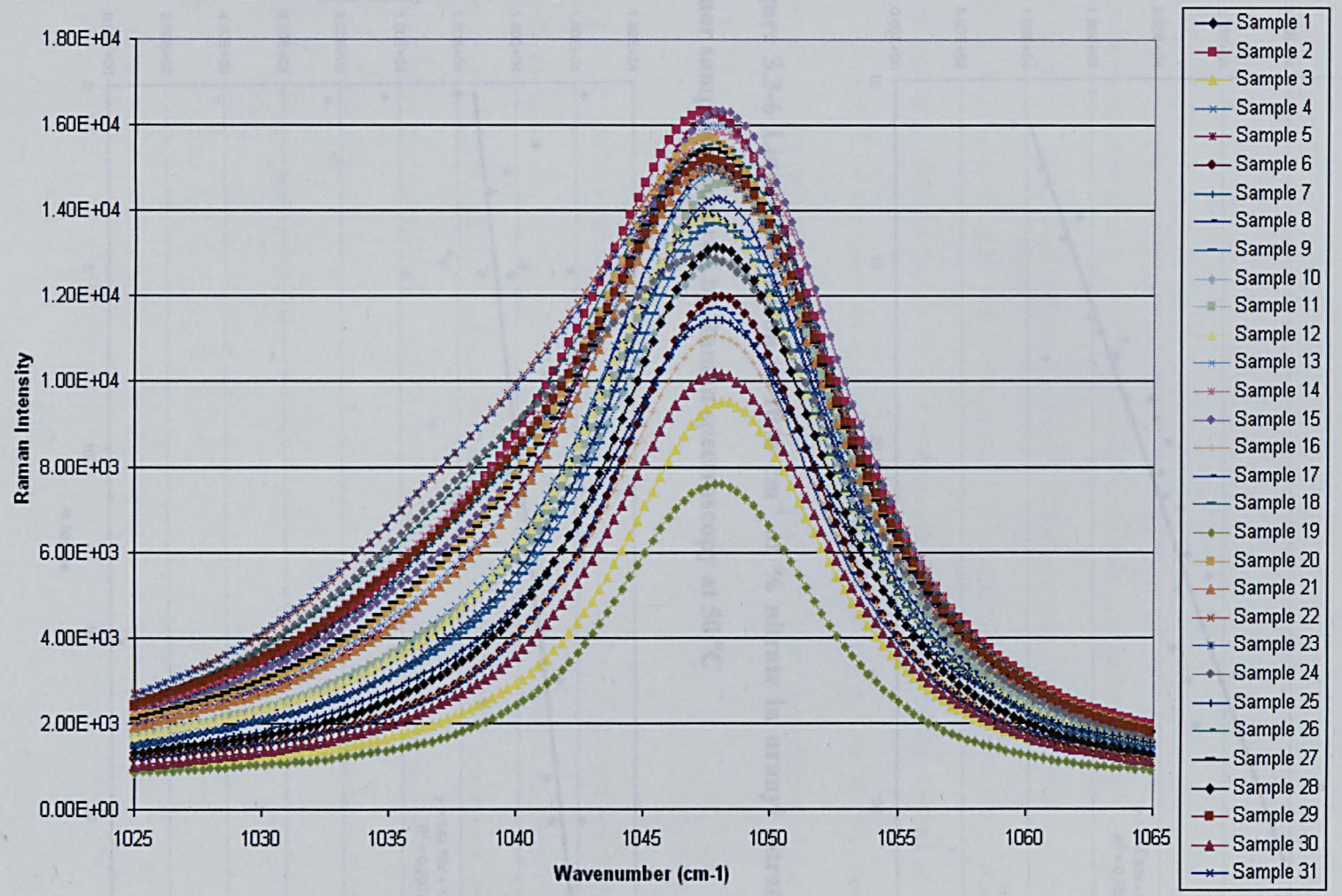


Figure 3.3-5 Linear regression at 873 cm^{-1} of % uranyl in uranyl nitrate liquor samples measured by Raman spectroscopy at 50 °C

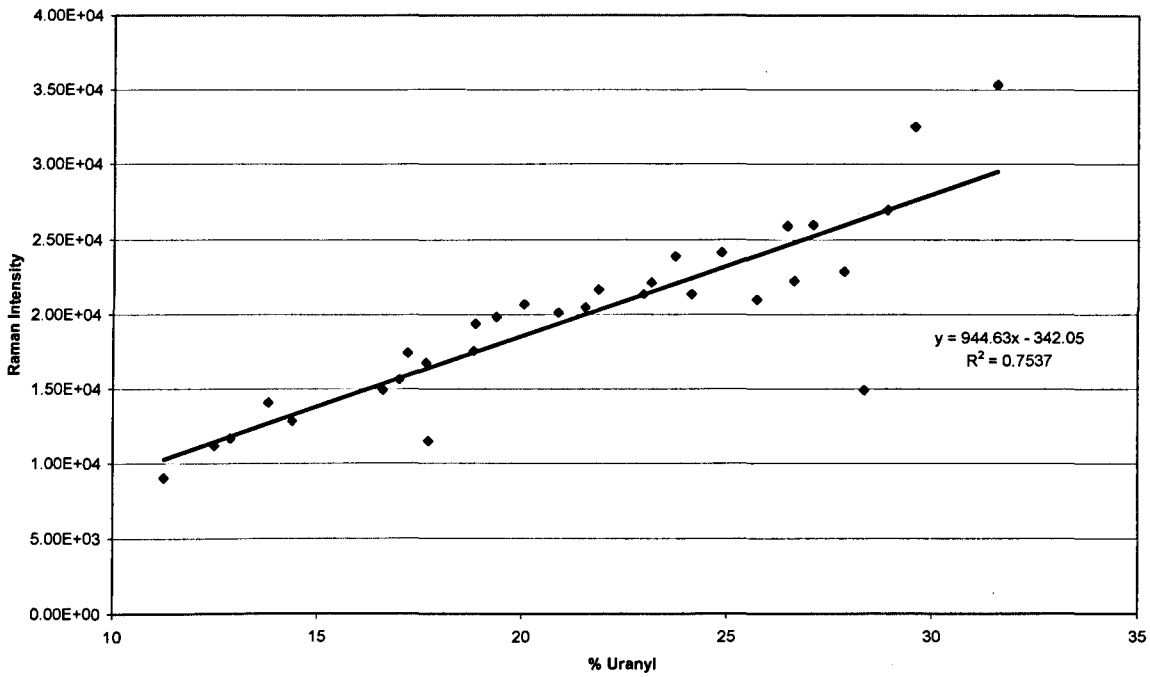
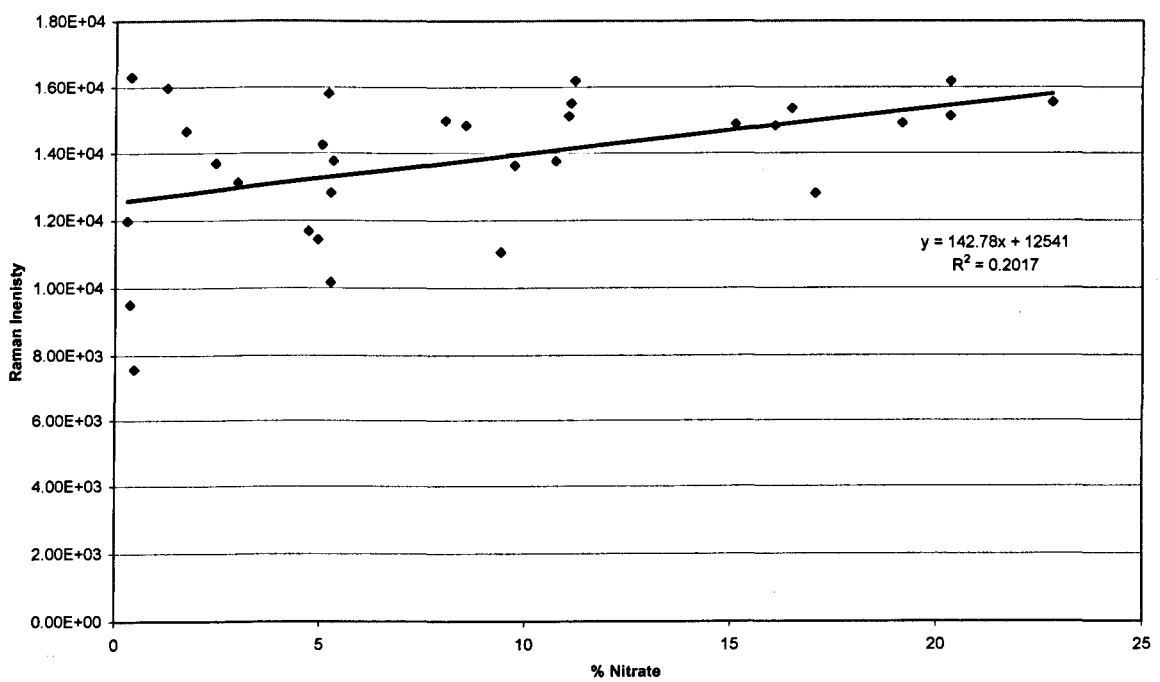


Figure 3.3-6 Linear regression at 1048 cm^{-1} of % nitrate in uranyl nitrate liquor samples measured by Raman spectroscopy at 50 °C



Univariate linear regression was applied to the spectra of fixed temperature at 50 °C. The actual vs predicted concentrations are plotted in figures 3.3-5 and 3.3-6 for uranyl and nitrate respectively. The method gave reasonable agreement for uranyl with an R^2 of 0.7537, but poor agreement with the nitrate concentration with R^2 of 0.2017.

The univariate regression where a single set of Raman intensities are used to model the data has shown to be unsuitable for nitrate prediction and has a higher than desirable correlation for the % uranyl in the samples. Magnification of the peak's, highlighted that the maximum peak intensity was varying with wavenumber between samples. Broadening on the first portion of the second peak was also noted which could be a result of impurities giving a Raman response. Multivariate calibration is used to overcome and minimise the effects of these problems.

The Raman spectra were reduced to include only the region of analyte response, which reduced the number of variables and also the computation time. To maximise variation in the samples spectra and remove baseline variations, the spectra were derivatised using the Savitsky Golay filter. The derivatised spectra are plotted in figure 3.3-3. The correlation based on the derivatised spectra in the region of sample response, 350 – 1250 cm^{-1} is plotted in figure 3.3-4.

Figure 3.3-7 1st Derivative spectra of uranyl nitrate liquor samples by Raman spectroscopy

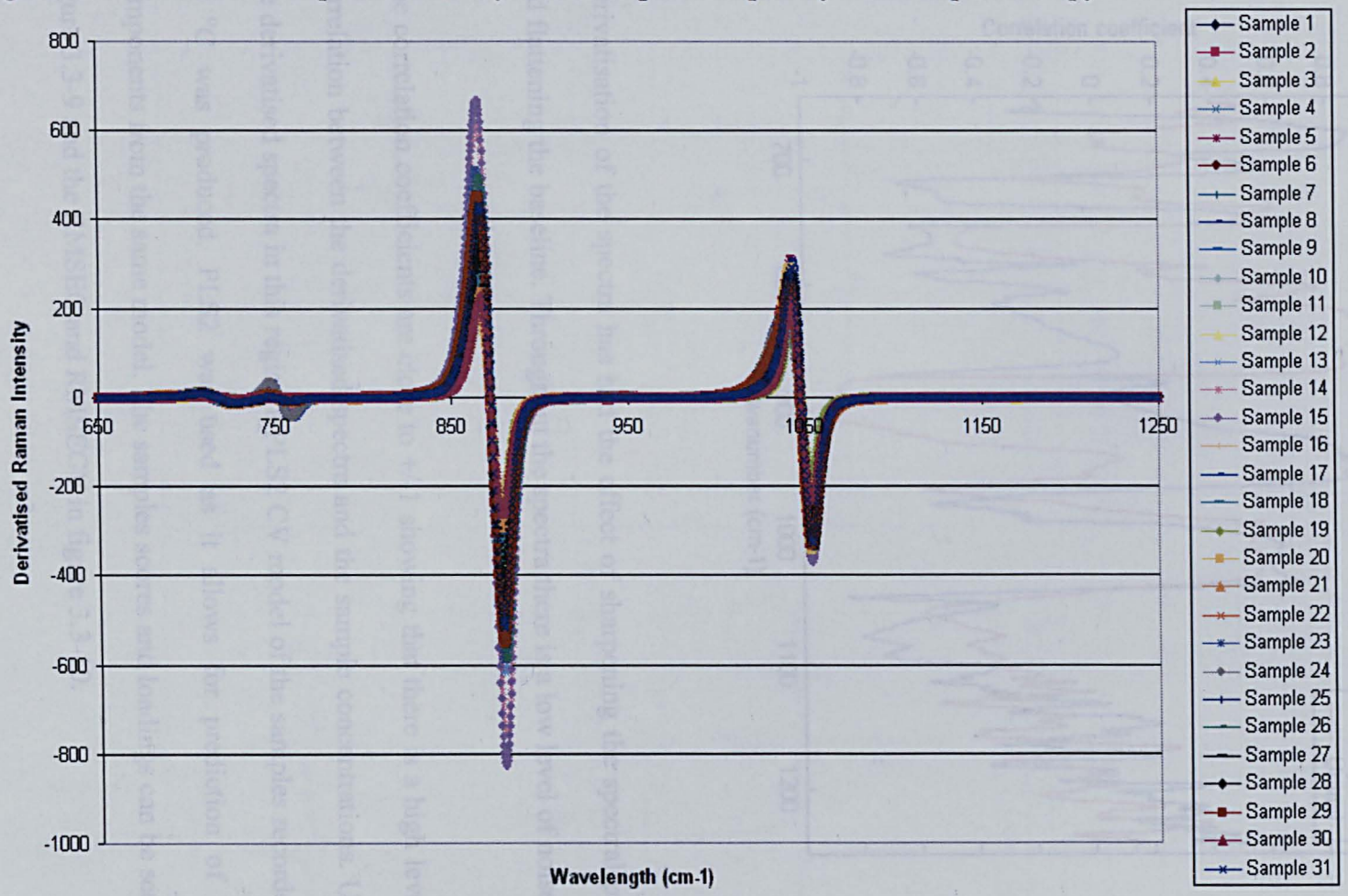
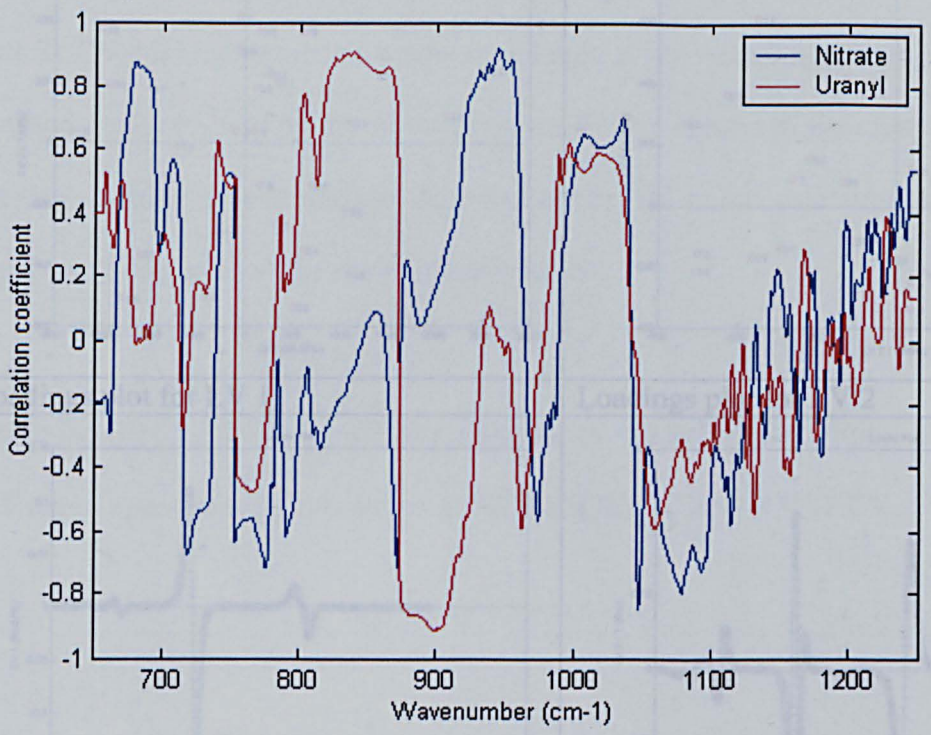


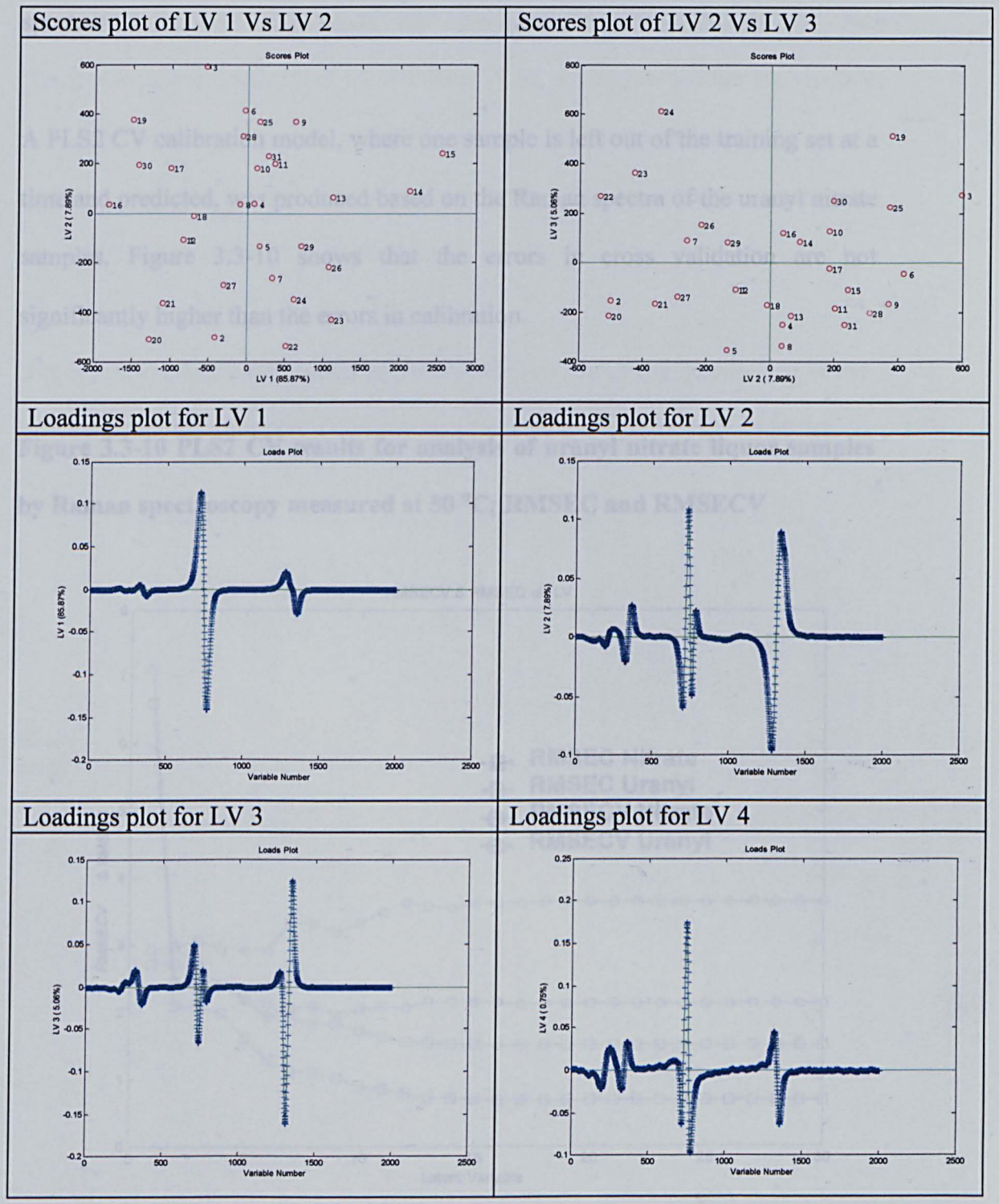
Figure 3.3-8 Correlation coefficients between the 1st derivative Raman spectra of uranyl nitrate liquor samples in the region of 650 – 1250 cm^{-1} and sample uranyl and nitrate concentration



Derivatisation of the spectra has had the effect of sharpening the spectral peaks and flattening the baseline. Throughout the spectra there is a low level of noise.

The correlation coefficients are close to +/-1 showing that there is a high level of correlation between the derivatised spectra and the sample concentrations. Using the derivatised spectra in this region a PLS2 CV model of the samples recorded at 50 °C was produced. PLS2 was used as it allows for prediction of both components from the same model. The samples scores and loadings can be seen in figure 3.3-9 and the RMSEC and RMSECV in figure 3.3-10.

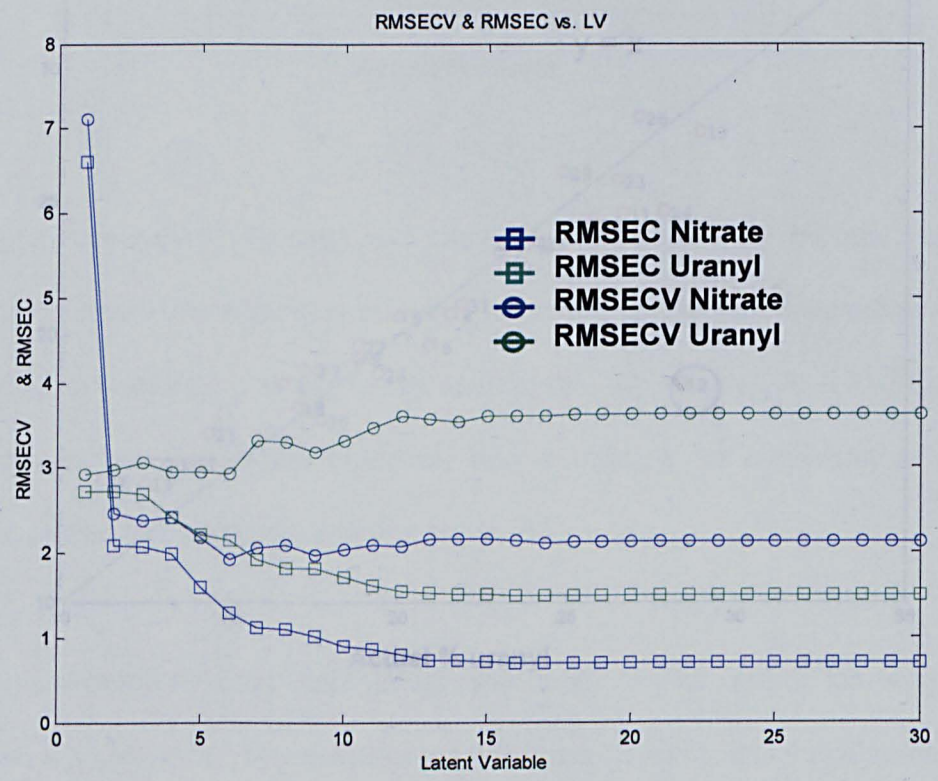
Figure 3.3-9 PLS2 CV Results of uranyl nitrate liquors by Raman spectroscopy; Samples scores plots and variable loadings plots



The PCA samples scores plots do not show trending in the spectra and the loadings plots have responses in the regions where peaks are seen in the original spectra.

A PLS2 CV calibration model, where one sample is left out of the training set at a time and predicted, was produced based on the Raman spectra of the uranyl nitrate samples. Figure 3.3-10 shows that the errors in cross validation are not significantly higher than the errors in calibration.

Figure 3.3-10 PLS2 CV results for analysis of uranyl nitrate liquor samples by Raman spectroscopy measured at 50 °C; RMSEC and RMSECV



Using 4 LVs the RMSECV is 2.62 and the RMSEC is 1.98. In figures 3.3-11 and 12 the actual vs predicted levels of uranyl and nitrate are plotted. There is good agreement between the actual and predicted % concentrations for both components, confirming that the standard PLS2 model can predict the levels of both components very well at 50 °C. Sample 12 (circled) is the most poorly predicted sample for both components indicating that this sample is a possible outlier; this was removed from the dataset.

Figure 3.3-11 PLS2 CV actual vs predicted plots for uranyl prediction of the uranyl nitrate liquor samples measured by Raman spectroscopy at 50 °C

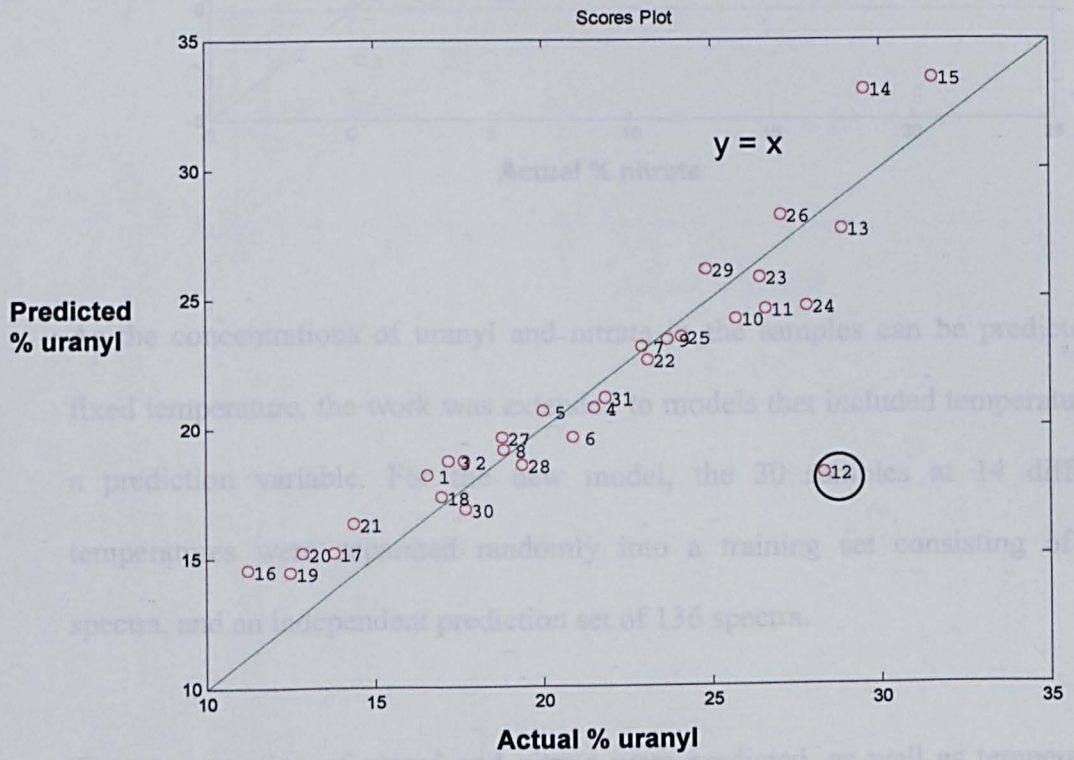
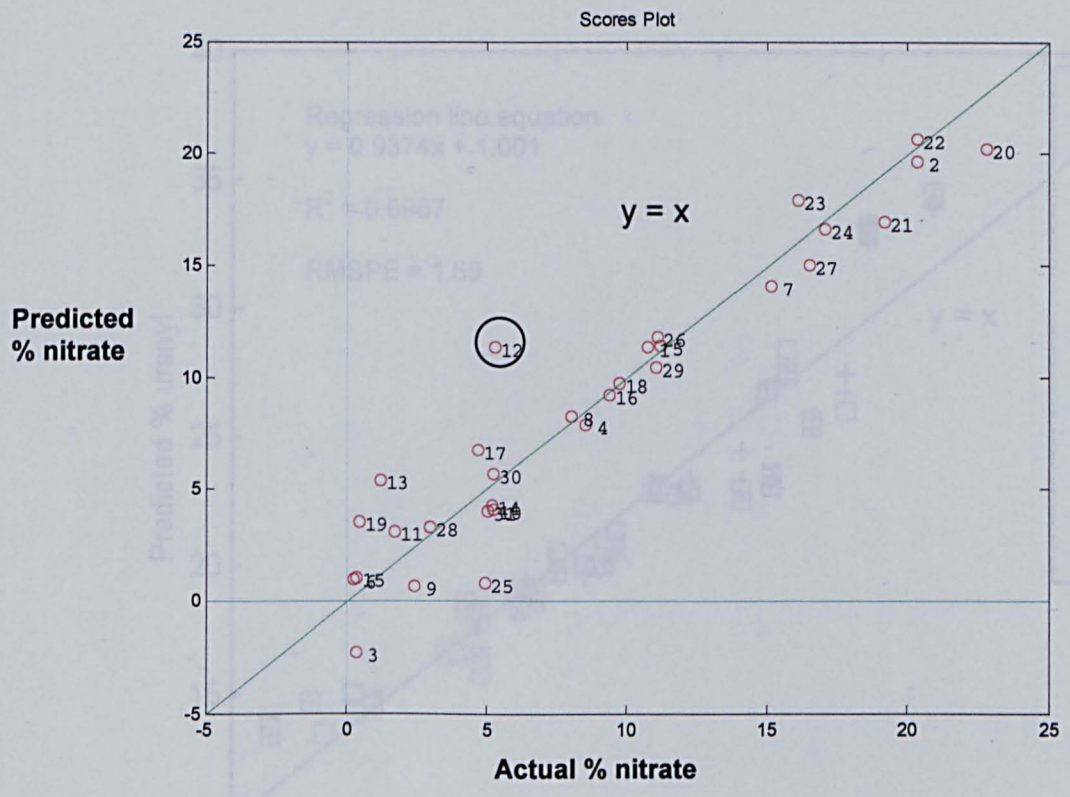


Figure 3.3-12 PLS2 CV actual vs predicted plots for nitrate prediction of the uranyl nitrate liquor samples measured by Raman spectroscopy at 50 °C



As the concentrations of uranyl and nitrate in the samples can be predicted at fixed temperature, the work was extended to models that included temperature as a prediction variable. For the new model, the 30 samples at 14 different temperatures were separated randomly into a training set consisting of 284 spectra, and an independent prediction set of 136 spectra.

The concentration of uranyl and nitrate were predicted, as well as temperature, based on a 6 LV model. The actual vs predicted results and residual prediction are plotted for uranyl and nitrate concentrations and temperature in figures 3.3 – 13 to 18.

Figure 3.3-13 PLS2 actual vs predicted plot for uranyl concentration in the uranyl nitrate liquors measured by Raman spectroscopy between 25 – 90 °C

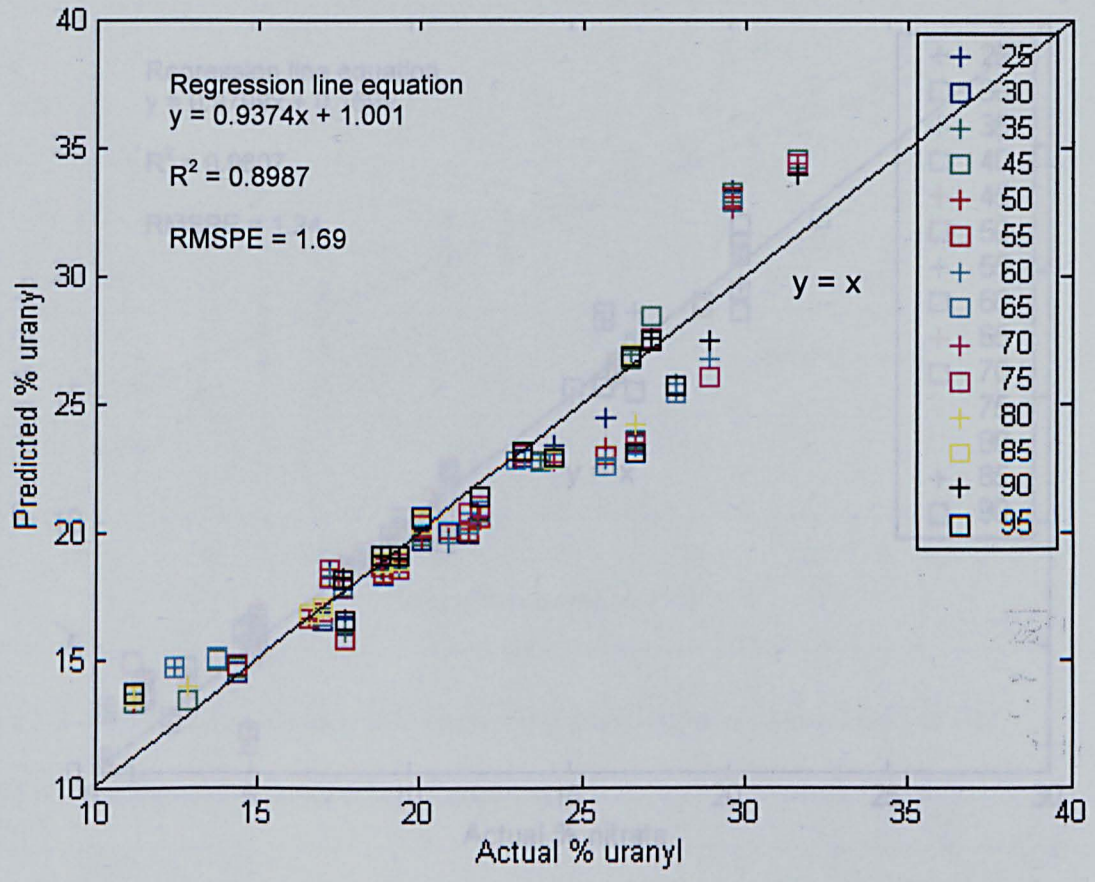


Figure 3.3-14 PLS2 actual vs predicted plot for nitrate concentration in the uranyl nitrate liquors measured by Raman spectroscopy between 25 – 90 °C

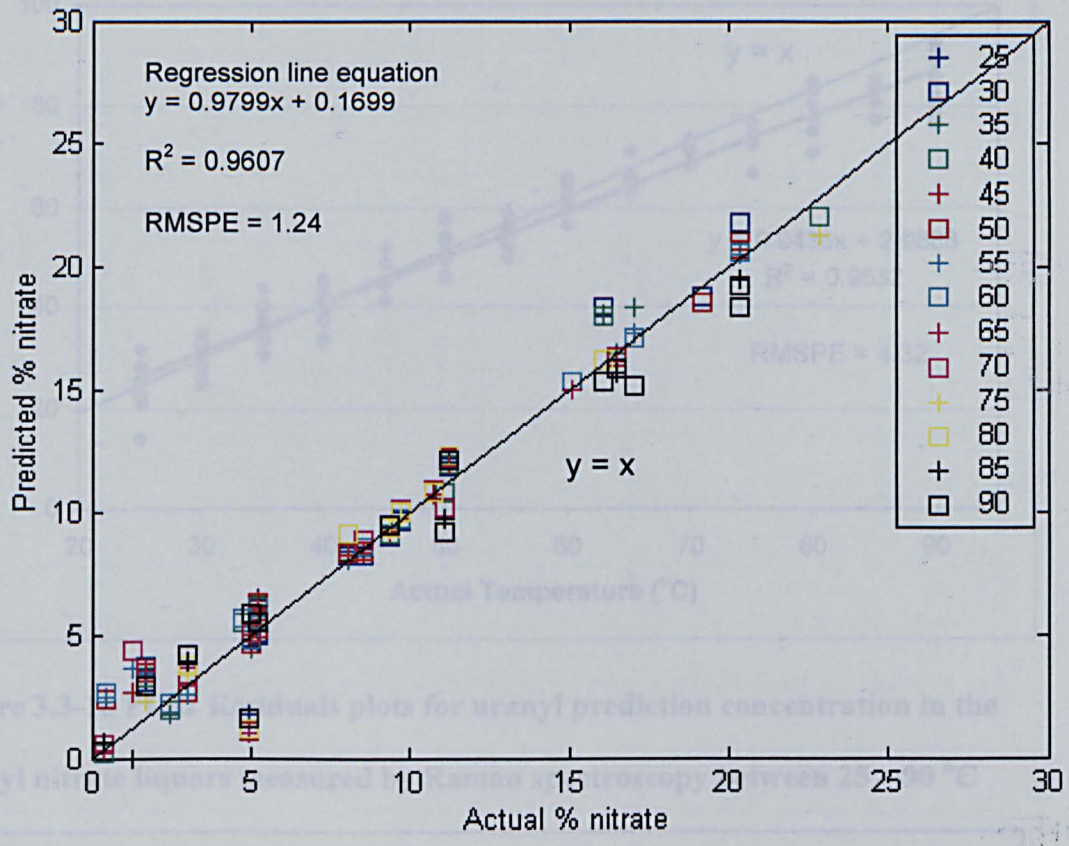


Figure 3.3-15 PLS2 actual vs predicted plot for Temperature of the uranyl nitrate liquors measured by Raman spectroscopy between 25 – 90 °C

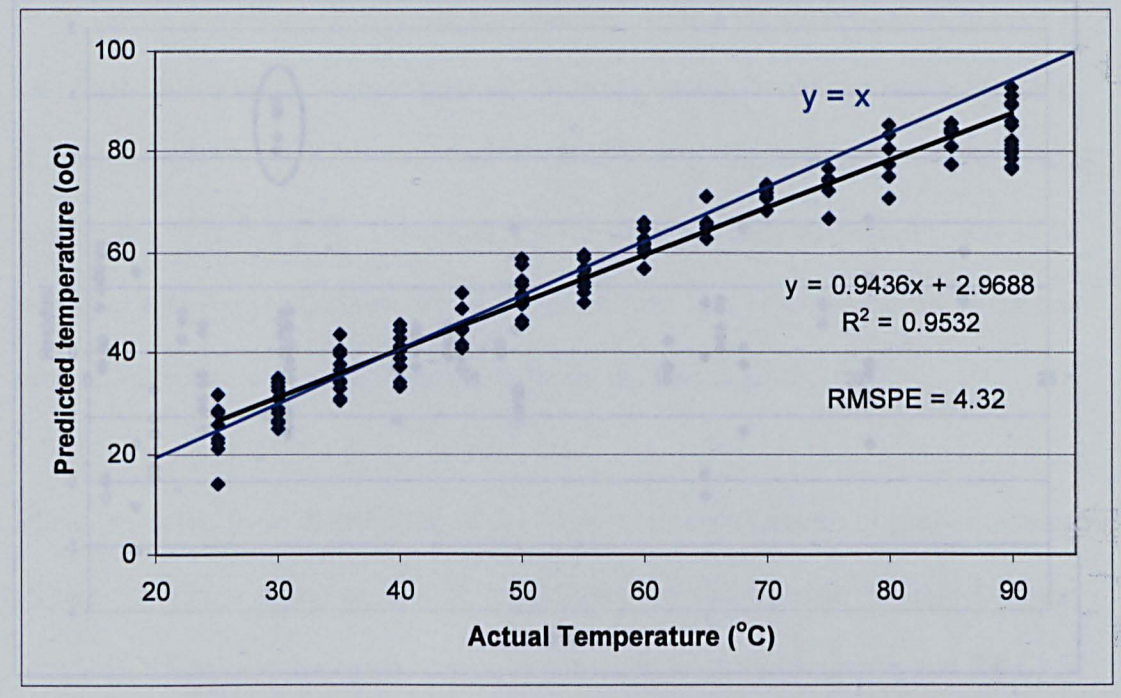


Figure 3.3-16 PLS2 Residuals plots for uranyl prediction concentration in the uranyl nitrate liquors measured by Raman spectroscopy between 25 – 90 °C

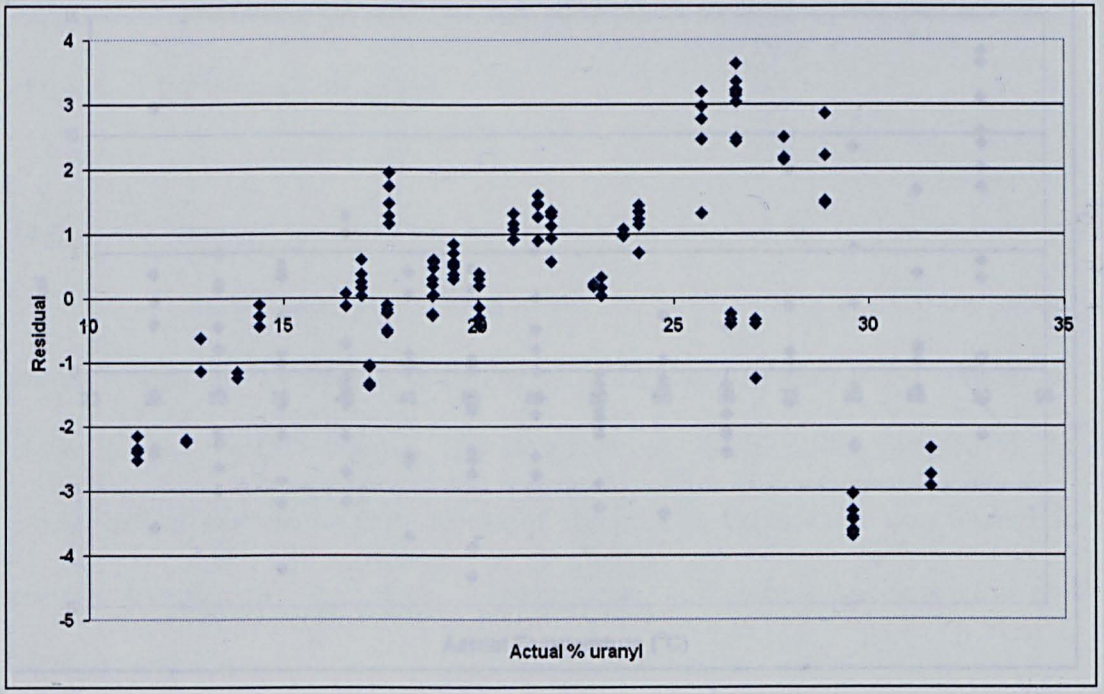


Figure 3.3-17 PLS2 Residuals plots for nitrate prediction concentration in the uranyl nitrate liquors measured by Raman spectroscopy between 25 – 90 °C

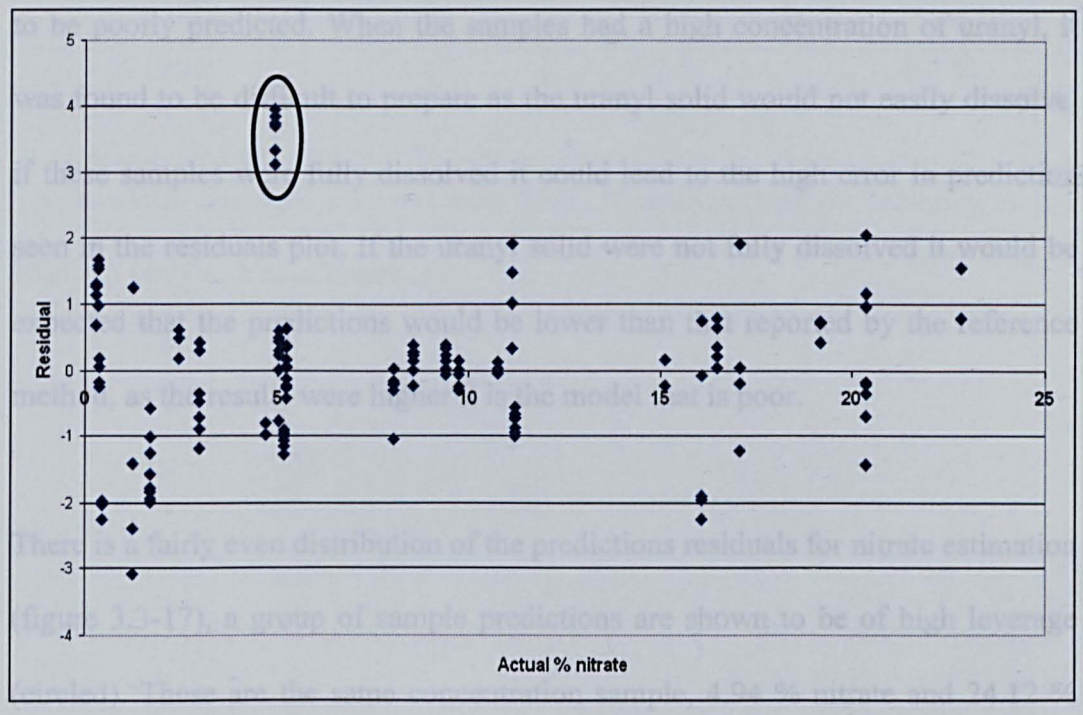
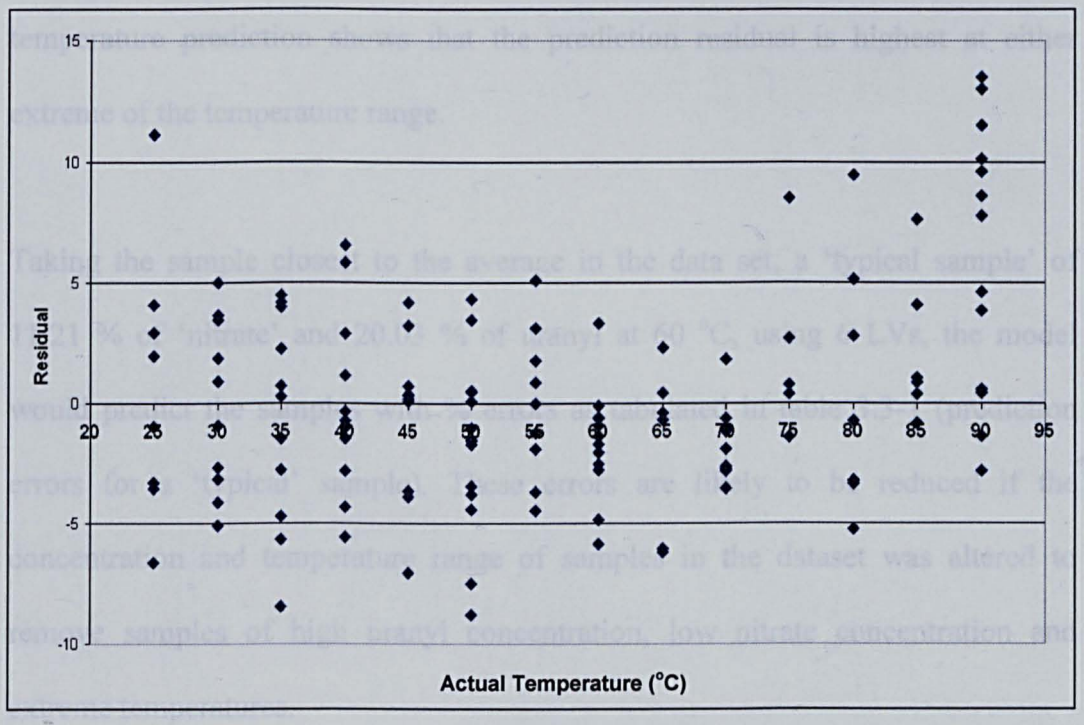


Figure 3.3-18 PLS2 Residuals plots for temperature prediction of the uranyl nitrate liquor samples measured by Raman spectroscopy between 25 – 90 °C



For all components there is some spread in the predictions. The residuals plot for uranyl prediction (figure 3.3-16) identifies the high uranyl concentration sample to be poorly predicted. When the samples had a high concentration of uranyl, it was found to be difficult to prepare as the uranyl solid would not easily dissolve, if these samples were fully dissolved it could lead to the high error in prediction seen in the residuals plot. If the uranyl solid were not fully dissolved it would be expected that the predictions would be lower than that reported by the reference method, as the results were higher it is the model that is poor.

There is a fairly even distribution of the predictions residuals for nitrate estimation (figure 3.3-17), a group of sample predictions are shown to be of high leverage (circled). These are the same concentration sample, 4.94 % nitrate and 24.12 % uranyl, sample 25 in table 6-3. This sample does not have high leverage for the prediction of uranyl concentration. Figure 3.3-18, the sample residuals for sample temperature prediction shows that the prediction residual is highest at either extreme of the temperature range.

Taking the sample closest to the average in the data set, a 'typical sample' of 11.21 % of 'nitrate' and 20.03 % of uranyl at 60 °C, using 6 LVs, the model would predict the samples with % errors as tabulated in table 3.3-1 (prediction errors for a 'typical' sample). These errors are likely to be reduced if the concentration and temperature range of samples in the dataset was altered to remove samples of high uranyl concentration, low nitrate concentration and extreme temperatures.

Table 3.3-1 Prediction errors for a 'typical' sample

	Nitrate (%)	Uranyl (%)	Temperature (°C)
RMSPE	1.69	1.24	4.32
% Error	15.1 %	6.2 %	7.2 %

The errors in the table may not meet BNFL's requirements for accuracy as this likely to be 10 – 5 % or lower.

4 Conclusions

The spectra recorded by UV/Vis were very poor quality as they were noisy and featureless. The correlation between the sample spectra and the reference information was calculated across the spectra and found to be +/- 0.2, for highly correlated data this figure should be close to +/- 1. A range of calibration models was calculated. The results show that at a fixed temperature, levels of uranyl and nitrate could only be predicted with greater than 25 % error. The work was not extended to include varied temperatures as it was deemed that the correlation between the spectra and the reference information was too low at +/- 0.2, at this level there is not considered to be any correlation between the samples spectra and reference concentrations.

The difference in the quality of the spectra between that recorded by UV/Vis and Raman spectroscopy is apparent on visually comparing the two corresponding datasets. When the correlation between the UV/Vis and the sample concentrations (figure 3.2-2), is compared to that seen with the Raman spectra (figure 3.3-2), it can be seen that the correlation throughout the UV/Vis spectra is similar to that seen in the regions of the Raman spectra which correspond to the spectrum's baseline, above 1500 cm^{-1}, and not sample analyte response. This can be taken as evidence that the UV/Vis spectra are not a result of variations in the sample analyte concentrations. Either the samples are not absorbing in the region analysed or there is a problem with the instrumentation used for measurement.

In contrast to this, the results obtained using the Raman spectra were successful. A first order Savitsky Golay derivatisation was applied which flattened the baseline and magnify the Raman intensity peaks. A PLS2 calibration model was produced based on the derivatised spectra, which could predict the levels of uranyl with 15.1 % error, nitrate with 6.2 % error and the sample temperature with 7.2 % error.

The residuals plots show that there were two groups of samples that were poorly predicted. These were samples with greater than 25 % uranyl or less than 5 % nitrate. From the information given by BNFL, it is known that samples with high uranyl content effects the solubility in nitric acid, and low nitrate levels will also effect the solubility of the uranyl.

The aim of this work was to not produce a calibration model within a certain criteria, but to give an indication of the expected error in prediction for uranyl, nitrate and temperature in uranyl nitrate liquors over the broadest range likely to be seen in the process environment. BNFL did not have any specific requirements for the accuracy or precision of the methods under scrutiny, generally a method of process analysis is considered feasible if the errors are less than 10 %. Using this approach the UV/Vis measurements documented here are not suitable for the analysis of uranyl nitrate samples but Raman spectroscopy would be suitable. This is in contrast to the results reported in the peer-reviewed literature that clearly document the use of UV/Vis spectrometry for measurement of uranyl nitrate.

A case study of multivariate calibration with Raman spectroscopic data has been reported by Estienne[9] *et al.* The aim was to investigate the replacement of classical calibration with inverse calibration of xylene mixtures measured by Raman spectra. The most efficient calibration method was to be established before transferring the measurement to an on-line process analytical environment. The calibration methods included MLR (stepwise and GA-MLR), PCR with variable selection and PLS. Presenting the results as a % of the classical calibration method i.e. an RMSECV of less than 100 % implied that the new method has lower errors in prediction compares the methods. The conclusion of Estienne *et al*[9] study found, as in this chapter, that multivariate calibration gave improved prediction over univariate or classical calibration. Estienne *et al* suggested this was because the multivariate methods could overcome the effect of impurities. The results found that for one component the relative RMSECV was 70 % for stepwise variable selection based on Fourier domain denoised spectra, the highest values were seen with the use of PCR. Such involved methods were not applied to uranyl nitrate spectra as the purpose was to prove the principle that Raman spectroscopy was a suitable method of analysis and not to search extensively for an optimal calibration method.

McGill[10] et al carried out a comparative study, for this study an esterification reaction was monitored in-line by NIR, Raman and UV-visible spectrometries and also at-line by NMR. For the esterification reaction 4 runs were performed and it was found that, out of the in-line methods, Raman spectrometry had the highest between-run precision and the UV-visible spectrometry the poorest. As for the

uranyl and nitrate prediction in this work the esterification product, 2-butyl crotonate, was predicted based on the derivatised spectra to remove any slight baseline variations. Univariate calibration was found to be sufficient for predictions based on the Raman and UV-visible spectra, this is in contrast for the results of the uranyl nitrate study which found multivariate analysis necessary for prediction. For prediction the in-line NIR and at-line NMR spectrometries were to the superior to Raman and UV-visible spectrometries.

5 Further Work

In order to determine the feasibility of using UV/Vis spectroscopy for the analysis of uranyl nitrate liquors the exact region of the spectra where the measurements were recorded should be established. Once the wavelength region is known, calibration should be performed in the region of 400 – 450 nm. If these fail to produce accurate results, which is expected given the lack of correlation between the samples spectra and the concentrations of uranyl and nitrate, the experiments should be repeated once the instrumentation has been checked for faults.

The Raman measurements were successful for measurement of uranyl, nitrate and prediction of the temperature of the samples. The next stage of this work would be to install an on-line analysis system to control the process. This would lead to financial cost benefits from elimination of over-engineering.

6 Appendix

Table 3.3-1 UV/Vis samples used for data processing

Solution ID	Sample	Nitrate %w/w	Uranyl %w/w
NCOD 01	1	10.75	16.59
NCOD 02	2	20.35	17.65
NCOD 04	3	8.55	21.54
NCOD 05	4	11.21	20.03
NCOD 06	5	0.28	20.89
NCOD 07	6	14.92	20.54
NCOD 08	7	15.12	22.96
NCOD 10	8	2.44	23.73
NCOD 11	9	5.25	25.72
NCOD 12	10	1.71	26.62
NCOD 13	11	5.31	28.35
NCOD 14	12	1.25	28.90
NCOD 15	13	5.19	29.57
NCOD 16	14	0.38	31.58
NCOD 18	15	9.40	11.24
NCOD 19	16	4.71	13.79
NCOD 20	17	9.74	17.00
NCOD 21	18	0.46	12.48
NCOD 22	19	22.83	12.86
NCOD 23	20	14.58	15.26
NCOD 24	21	19.18	14.38
NCOD 26	22	20.35	23.15
NCOD 27	23	22.20	26.66
NCOD 28	24	24.34	21.64
NCOD 30	25	16.08	26.46
NCOD 34	26	12.22	23.75
NCOD 35	27	4.94	24.12
NCOD 36	28	11.12	27.09
NCOD 37	29	16.49	18.82
NCOD 38	30	2.98	19.37
NCOD 40	31	5.25	17.71

Table 3.3-2 Raman samples used for data processing

Solution ID	Sample	Nitrate % w/w	Uranyl % w/w
NCOD 01	1	10.75	16.56
NCOD 02	2	20.35	17.65
NCOD 03	3	0.35	17.20
NCOD 04	4	8.55	21.54
NCOD 05	5	11.21	20.03
NCOD 06	6	0.28	20.89
NCOD 08	7	15.12	22.96
NCOD 09	8	8.05	18.86
NCOD 10	9	2.44	23.73
NCOD 11	10	5.25	25.72
NCOD 12	11	1.71	26.62
NCOD 13	12	5.31	28.35
NCOD 14	13	1.25	28.90
NCOD 15	14	5.19	29.57
NCOD 16	15	0.38	31.56
NCOD 18	16	9.40	11.24
NCOD 19	17	4.71	13.79
NCOD 20	18	9.74	16.99
NCOD 21	19	0.46	12.48
NCOD 22	20	22.83	12.86
NCOD 24	21	19.18	14.38
NCOD 26	22	20.35	23.15
NCOD 30	23	16.08	26.46
NCOD 33	24	17.06	27.85
NCOD 35	25	4.94	24.12
NCOD 36	26	11.12	27.09
NCOD 37	27	16.49	18.82
NCOD 38	28	2.98	19.37
NCOD 39	29	11.06	24.85
NCOD 40	30	5.25	17.71
NCOD 41	31	5.04	21.86

7 References

[1] Personal communication (Post): Clarke C. G, BNFL, Springfields, 28/11/02.

[2] Personal communication (E-mail): Clarke C. G, BNFL, Springfields, BNFL internal document: 'Analytical Method for the Rapid Determination of Nitric Acid Soluble Uranium in Miscellaneous Solids and Liquors Using X-ray Fluorescence Spectrometry', 27/11/02.

[3] Meeting minutes BNFL Springfields, 21/5/02: 'Data Analysis of Uranyl Nitrate Product Dissolution Process', Loades V. C, 22/5/02.

[4] Rabinowich E and Belford R. L, 'Spectroscopy and Photochemistry of Uranyl Compounds', Pergammon, Oxford, UK, 1964.

[5] Relan G.R, Dubey A.N, Bhanu A.U and Vaidyanathan S, 'Application of Second Derivative Spectrophotometry for the Determination of Uranium in Yellow Cake', Journal of Radioanalytical and Nuclear Chemistry, 1994, 182 (2), 437 – 443.

[6] Personal communication (E-mail): Clarke C. G, BNFL, Springfields, 12/11/02.

[7] Palacios M, L, Taylor S, H, 'Characterization of Uranium Oxides Using in-situ Micro-Raman Spectroscopy', Applied Spectroscopy, 54, 9, 2000, 1372 – 1378.

[8] Skoog D.A, Holler F.J and Niemen T.A, 'Principles of Instrumental Analysis', , 5th Editions, Saunders, 1998, 139.

[9] Estienne F, Massart D.L, Zanier-Szydlowski N, Marteau Ph, 'Multivariate Calibration with Raman Spectroscopic Data: A Case Study', Analytica Chimica Acta, 424, 2000, 185 – 201.

[10] McGill C. A, Nordon A and Littlejohn D, 'Comparison of in-line NIR, Raman and UV-visible spectrometries, and at-line NMR spectrometry for the monitoring of an esterification reaction', Analyst, 127, 2002, 287 – 292.

Chapter 4

Application of Guided Microwave Spectroscopy for Chemical Measurement

1 Introduction

The benefits of microwave spectroscopy for process analysis are detailed for some microwave analysers[1]. The signal magnitude changes in microwaves spectra due to variation in sample composition. Large signal variations are especially seen with highly polar molecules such as water and alcohols. The microwave spectrometer generator and detector are solid state circuitry reducing many of the problems associated with spectrometers such as NIR spectrometers which may suffer from wear, vibration, stray light and can be more sensitive to ambient temperature variation and condensation. Microwaves will penetrate most materials except for metals, meaning that the whole sample is analysed[2].

Many on-line analysis methods are invasive, as they require a measurement probes to be placed in the reactor or connecting piping. One goal is to develop new methods that are non-invasive, and microwave spectroscopy is ideal for this because it is capable of analysis with pathlengths of many cm compared with many on-line spectroscopy techniques that use mm pathlengths. The advantage of long pathlengths is that the measurement cavity can potentially form part of a feeding pipe, benefiting from very simple engineering for sample analysis and reduced issues of representative sampling during analysis.

A commercially available guided microwave spectrometer[3] designed for on-line analysis and used for applications of food analysis such as moisture content of grain and fat in meat processing is investigated here for use as a process analyser for chemical analysis.

1.1 Theory of microwave spectroscopy

The microwave region lies between the infrared and radio frequencies of the electromagnetic spectrum at frequencies between 30 GHz to 300 MHz relating to a wavelength range of 1 cm to 1 m.

Traditionally microwave spectroscopy has been used for the analysis of gases as these yield sharp response bands that can be used to fingerprint samples. From gaseous samples, microwave spectroscopy has also led to the accurate determination of bond lengths, and with the aid of isotopic dilution, precise relative atomic masses, bond lengths and angles have been obtained.

Microwave spectra are due to the ability of the analyte to rotate within the microwave field, as the sample phase varies from gases to liquids to solids the rotation ability is reduced resulting in a change of spectra from well resolved sharp peaks to broadband often collinear spectra. This spectroscopy is suited to polar molecules that can easily produce an alternating polarization and rotate in the microwave field. The bands are due to the quantised rotational energy of the analyte; defined in equation 16.

$E_j = J(J + 1)\, h^2 / 8\pi I^2$ Equation 16 Rotation energy in a microwave field

Where;

E_j = Rotational energy

J = Rotational number (integer values 0, 1.... etc.)

I = Moment of inertia

h = Planck's constant

Microwaves interact with matter by coupling energy from the electromagnetic field by electric and magnetic dipole interactions[4]. For microwave absorption to occur the materials must have a low dielectric constant otherwise they will be reflected. Thus they are reflected by metallic surfaces but can be absorbed through plastics and glass[5].

The two main physical factors that produce a response to the microwave energy is the dielectric constant and dielectric loss. The dielectric constant is similar to the refractive index of a sample.

As an electromagnetic wave passes through a mixture it induces alternating polarization within the mixture. As this process stores some of the wave's energy, in effect it slows the velocity of the wave as it travels through the sample. The ability of a mixture to store energy and slow a wave's travel is the refractive index.

1.1.1 Dielectric Constant

The dielectric constant (ε') is used to describe the velocity loss factor of the wave as it passes through a sample. At the microwave frequency, the dielectric constant,

is determined according to equation 17. The dielectric constant is a dimensionless relative measurement.

$$\varepsilon' = \frac{(V_{VAC})^2}{(V_{MIX})^2}$$

Equation 17 Calculation of dielectric constant

Where;

Vvac = Velocity through vacuum

Vmix = velocity through analyte mixture

1.1.2 Dielectric Loss

The second important factor is the dielectric loss (ε"). This relates to the amplitude loss factor of the wave as it passes through a sample. As the molecule is polarized and de-polarised by the electromagnetic wave passing through the sample some energy is converted to heat through friction. This 'heat' energy is not returned to the wave and as a result the amplitude of the wave reduces as it passes through the mixture. Thus ε" is the ability to attenuate the wave.

The frequency of the wave affects the ability to store energy, at lower frequencies the wave has most time to the rotate the molecules (windup to store and unwind to release the energy). As the frequency increases the speed at which the wave attempts to rotate the molecules increases, for some molecules, especially larger ones or those with high moments of inertia and spring rates, this is too fast and rotation is difficult. As a result the energy storage and loss effects are reduced with increasing frequency.

1.2 Applications of GMS

The GMS spectrometer utilises the decimetre waves between 0.25 – 3.2 GHz, (250 - 3200 MHz). The background and principles of the GMS analyser was first published in 1992[6]. The paper describes how the system can be used for on-line determination of a broad range of analysis methods including; 0 – 100 % moisture content, total and dissolved solids, percentage additives, methanol in fuel and conductivity in the range of 10^{-14} – 10^{3} μs/cm.

Using an at-line setup this spectrometer has been successfully used to determine the moisture content in tobacco samples[7]. Five different types of tobacco were analysed with moisture levels between 10 – 50 %. Several multivariate calibration models were produced to determine which method gave best prediction. It was found that, accurate prediction was achieved when the weight of the sample was included in the regression. Using PLS the moisture levels were predicted with 2 % error. This was an order of magnitude higher than the existing error for the NIR method of 0.2 %, however, the NIR models did have substantially more samples, approx 1500, compared to approx 50 for the GMS method.

Samples are analysed in a sample chamber that is also a waveguide. The effect of this is a cut-off frequency in the sample spectra (equation 18).

$$f_c = \upsilon_{vacuum} \Big/ 2a\sqrt{\varepsilon'}$$

Equation 18 GMS Cut-off frequency

This frequency (f_c) is a function of the dielectric constant (ε') and the waveguide's geometry (a = distance between the two plates).

The GMS instrument reports the signal response in terms of normalised chamber signals. This is to cancel out any variations that could result from the microwave power output, receiver sensitivity and coaxial cable loss[3]. The normalised response is equal to the measurement through the chamber, minus an internal calibration plus and external calibration.

The amplitude of the response is in the logarithmic units of hexibels (hb's)[8]. This is a non-conventional unit that is utilised to allow the additions and subtractions to be equivalent to multiplications and divisions in normal units, to give a normalised response. The hexibel is related to the decibel in that 1db = 340.16 hb. A decibel is a unit which can describe the difference in power levels, voltage or current.

1.3 Application of GMS for Chemical Measurements

The work cited for applications of the GMS spectrometer mainly involves the measurement of moisture in solids (grains, tobacco). The aim of this work is to investigate the GMS system for chemical measurement. Chemometric techniques will be used to resolve the measurements. Initially the measurements will be based on simple liquid solutions, this is then extended to more complex systems of beer fermentation in Chapter 5 and measurement of complex industrial process samples in Chapter 6.

In keeping with the measurement of molecules which have high dielectric constants and are hydrogen bonded, a series of alcohol solutions will be analysed. The first will be simply ethanol in water samples; in addition to creating calibration models, these samples will be used to establish the repeatability of the system for measurement. GMS will then be used for the analysis of alcohol mixtures containing 2, 3 and 4 alcohols. Analysis of homologous alcohols is spectroscopically challenging, as it requires the resolution of spectral features of molecules with similar function groups.

NIR spectroscopy has been shown to be suitable for the identification of a mixture of 24 straight and branched chain alcohols[9]. PCA was applied to the NIR spectra of the mixture which was recorded between 1000 – 2500 nm. The PCA scores plots for PC1 vs PC2 showed the samples to be separated according to branched and straight chain, and positions in order of carbon chain length.

Another study determined the concentrations of methanol and ethanol in carbon tetrachloride solutions by NIR[10]. Three sets of 11 samples were analysed with carbon tetrachloride as the solvent. These were in the range of (i) 0 -1 % alcohol, (ii) 0 - 10 % alcohol and (iii) 0 - 100 % alcohol. The spectra were recorded at 28.5 ^{o}C in a thermostated sample holder. The authors reported non-linearity in absorbance measurement when plotted against the sample's concentration at several example wavelengths. The paper attributes this to nonlinear temperature effects of hydrogen bonding within the samples. An example spectrum of pure methanol is plotted when measured between 23 – 28.5 ^{o}C and the absorbance

variation is non-linear. To model the data, variable selection of up to 5 wavelengths was used for calibration of each set with MLR, PCR, PLS and ANN. All performed well with less than 0.02 % relative standard error. Generally, the predictions were comparable between methods. The paper concluded that non-linearity in the samples was caused by effects of concentration and temperature. Given that the samples were recorded at fixed temperature, the between-sample variation of this should have been minimal and the non-linearity should have been due to the fluctuation of hydrogen bonding with concentration.

2 Experimental

2.1 Procedure for GMS Analysis

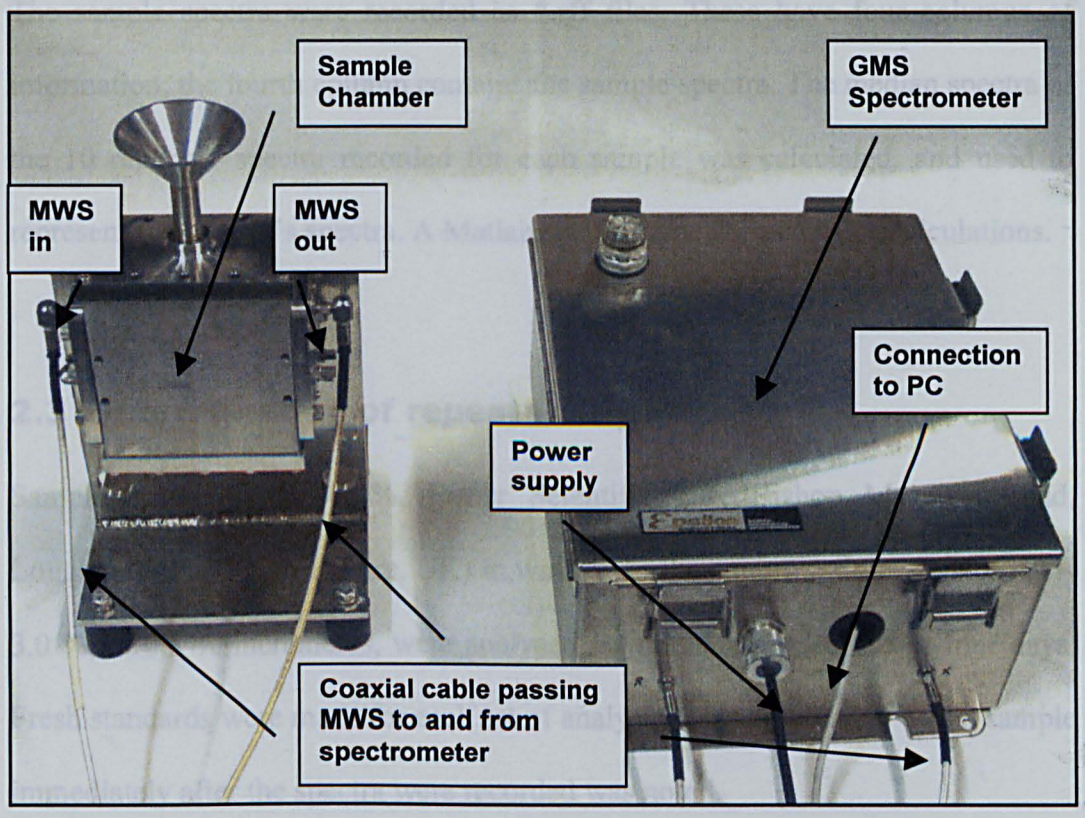
The setup of the GMS spectrometer can be seen in figure 2.1.1-1. The samples were measured in a remote sample chamber made of stainless steel, which has internal dimensions of 10.0×4.7×11.5 cm, with a total internal volume of 540 cm^3. Samples were generally 500 ml in volume. The sample chamber was connected to the GMS via two coaxial cables (450 cm in length) which transmitted the microwaves to and from the spectrometer to the sample chamber. A PC connected to the instrument recorded the spectra using Linefit™ v1.43, Epsilon Industrial Inc*. For each sample, 10 scans were recorded consecutively and the data transferred to Matlab. The median of the 10 spectra were used for data analysis.

2.1.1 Instrumentation

Instrument:	Guided microwave spectrometer*
Control PC:	Pentium (I) 166 MHz processor, 64 MB RAM
Power:	5 mW
Dielectric range	1 - 85
Frequency range:	200 – 3200 MHz
Resolution:	8 MHz
Number of data points:	375
Operating software:	LineFit v1.43

* Epsilon Ind. Inc, 2215 Grand Avenue Parkway, Austin, Texas, 78728, USA.

Figure 2.1.1-1 Epsilon industrial guided microwave spectrometer



2.2 General procedure for GMS analysis

The spectrometer was switched on for a minimum of 2 hours prior to analysis. Throughout this work the spectrometer's sample chamber was used in an at-line setup. The spectrum of 500 ml of distilled water was analysed prior to those of the samples. Generally where standard samples are analysed these are prepared in 500 ml 'A' grade volumetric flasks. The samples were poured into the top of the sample chamber, left for a minute to settle then the spectra recorded 10 times. Collection of the 10 spectra took approximately 1 minute. Between samples, the chamber was rinsed with 500 ml of distilled water.

2.2.1 Preparation of spectral files for data processing

The sample spectra were recorded as *.rff files. These have four columns of information; the fourth column contains the sample spectra. The median spectra of the 10 replicate spectra recorded for each sample was calculated, and used to represent the sample's spectra. A Matlab script file performed these calculations.

2.3 Determination of repeatability

Samples of ethanol (99 %, Fisher Scientific UK, Bishop Meadow Road, Loughborough, Leicestershire, UK) in water over the concentration range of 0.5 – 3.0 %, in 0.5 % increments, were analysed six times over a course of four days. Fresh standards were made for each set of analysis. The temperature of the sample immediately after the spectra were recorded was noted.

2.4 Analysis of alcohol solutions

Guided microwave spectroscopy should respond well to aqueous and alcohol solutions. The aim of this work is to measure a series of homologous alcohol solutions with 1 – 4 alcohols present then to apply linear and, where necessary, nonlinear multivariate calibration methods to allow for prediction of each alcohol.

2.4.1 Ethanol in water solutions

Samples of ethanol, in water (distilled) in the following range; 1, 3, 5, 8, 10, 12, 15, 20, 23, 25, 28 and 30 % v/v were prepared in duplicate in 500 ml volumetric flasks and analysed by GMS.

2.4.2 Binary mixtures of alcohols

Three sets of binary mixtures of alcohols were analysed. The samples were constrained to have a total sample volume of 500 ml, so the minimum and maximum volume of alcohol was 150 and 350 ml. There was one replicate of the middle concentration sample. The sets are described in table 2.4-1 and samples in table 2.4-2.

Table 2.4-1 Description of binary mixtures of alcohol sample sets

Dataset	Alcohol 'A'	Alcohol 'B'
1	Methanol	Ethanol
2	Methanol	Propanol
3	Ethanol	Propanol

Table 2.4-2 Sample volumes of binary mixtures of alcohol samples

Sample	Alcohol A (ml)	Alcohol B (ml)
1	150	350
2	175	325
3	200	300
4	225	275
5	250	250
6	275	225
7	300	200
8	325	175
9	350	150
10	250	250

The alcohols were dispensed into the flasks using a bottle top dispenser (Fortuna® Optifix® bottle top dispenser for solvents, Fisher Scientific). The methanol, ethanol and butanol were Absolute, 99 % strength, Fisher Scientific. The propanol was Rectapur™, ALR grade (Prolabo, Fisher Scientific). For the odd numbered

samples alcohol A was placed in the flasks first and for the even number samples alcohol B was dispensed into the flasks first.

2.4.3 Tertiary mixtures of alcohol samples

The samples were calculated by a D-optimal experimental design, with 2 replicates constrained within the volume range of 100 – 200 ml with a total volume of 500 ml. The samples were rounded to the nearest 5 ml, which resulted in several replicate samples (see table 2.4.3-1).

Table 2.4.3-1 Description of methanol, ethanol and propanol alcohol samples

Sample	Methanol (ml)	Ethanol (ml)	Propanol (ml)
1	150	150	200
2	200	100	200
3	200	200	100
4	200	150	150
5	100	200	200
6	100	200	200
7	165	165	170
8	150	200	150
9	200	100	200
10	180	180	140
11	200	150	150
12	140	180	180
13	180	140	180
14	200	200	100
15	165	165	170
16	160	180	160

2.4.4 Quaternary mixtures of alcohol samples

The samples consisted of methanol, ethanol, propanol and butanol. The concentrations were determined by a partial two level D-optimal experimental design, linear with centre point replicates. The minimum and maximum volumes

were 50 ml and 250 ml, and the sum of all four volumes was equal to 500 ml as calculated using Design Expert (samples 1 -14). The volumes of samples (15 – 24) were randomly generated within the constraints of the design.

Table 2.4.4-1 Description of quaternary mixtures of alcohol samples

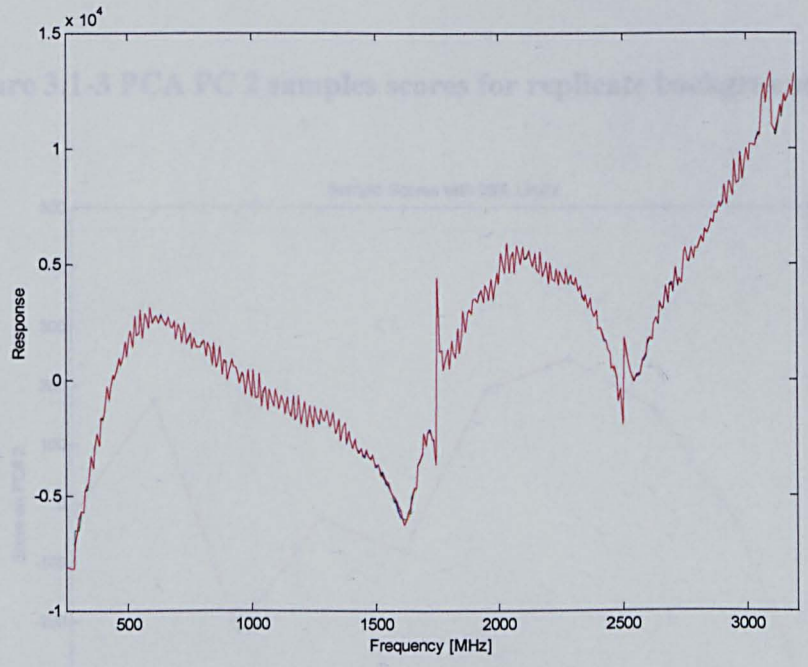
Sample	Methanol (ml)	Ethanol (ml)	Propan-1-ol (ml)	Butan-1-ol (ml)
1	250	150	50	50
2	250	150	50	50
3	50	150	50	250
4	50	50	250	150
5	50	250	150	50
6	150	50	250	50
7	125	125	125	125
8	50	250	150	50
9	50	150	50	250
10	50	50	250	150
11	50	50	150	250
12	50	250	150	50
13	250	50	150	50
14	250	50	50	150
15	120	245	80	55
16	170	130	100	100
17	210	210	80	0
18	150	65	50	235
19	95	95	100	210
20	135	60	60	245
21	90	70	180	160
22	60	240	110	90
23	210	70	150	70
24	110	60	220	110

3 Results

3.1 Initial measurements: GMS background spectra

Prior to all analysis by GMS the background spectra of the empty sample chamber was recorded. Each analyte spectrum was recorded 10 times. From these 10 replicates the median spectra were calculated and used for data processing. The aim of this section is to demonstrate that there is no systematic variation in the 'replicate' spectra. For this, the replicate spectra of background measurements are used (see figure 3.1-1)

Figure 3.1-1 Replicate GMS background spectrum of air



The sample scores in figure 3.1-2 to 5 for PCs 1 – 4, show random variation between the scores for each PC. The procedure of taking the median spectrum to be representative of each measurement should result in unbiased sample spectra.

Figure 3.1-2 PCA PC 1 samples scores for replicate background spectra

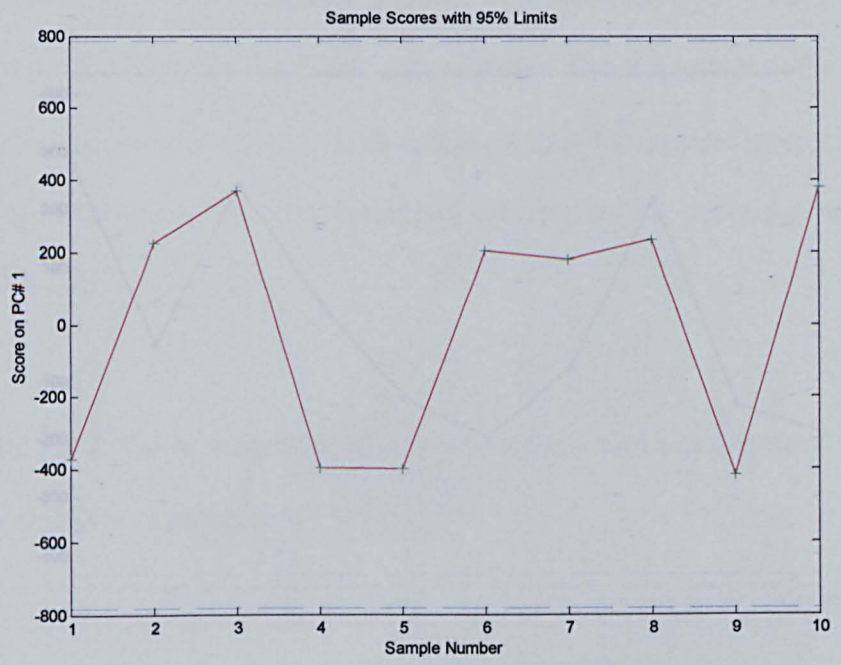


Figure 3.1-3 PCA PC 2 samples scores for replicate background spectra

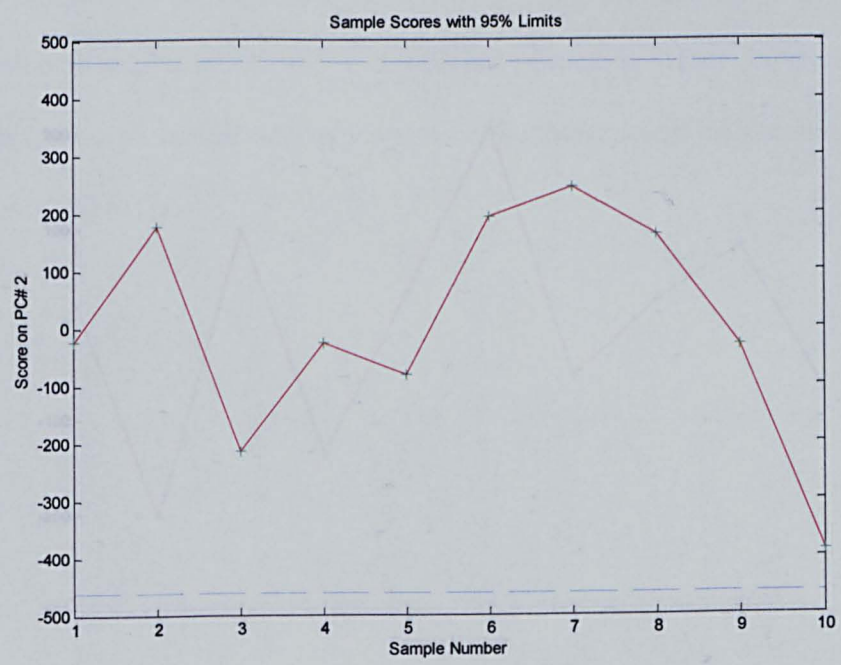


Figure 3.1-4 PCA PC 3 samples scores for replicate background spectra

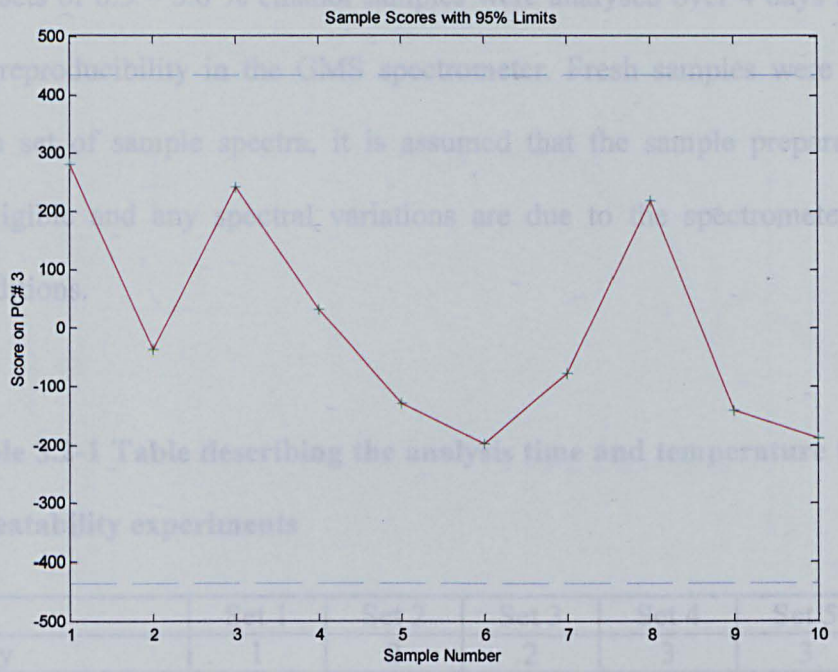
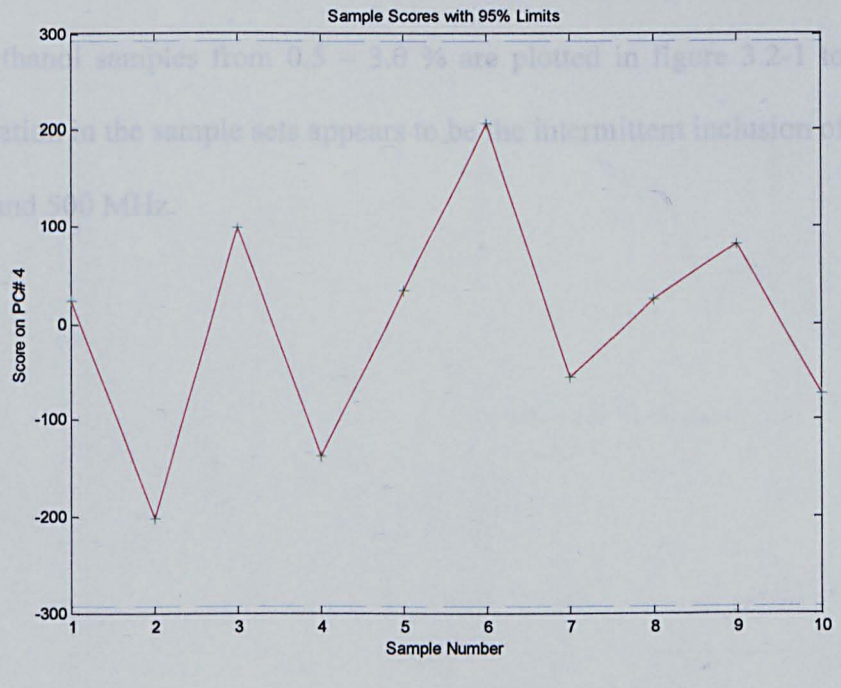


Figure 3.1-5 PCA PC 4 samples scores for replicate background spectra



3.2 Determination of GMS repeatability

Six sets of 0.5 – 3.0 % ethanol samples were analysed over 4 days to investigate the reproducibility in the GMS spectrometer. Fresh samples were prepared for each set of sample spectra, it is assumed that the sample preparation error is negligible and any spectral variations are due to the spectrometer or ambient conditions.

Table 3.2-1 Table describing the analysis time and temperature for the GMS repeatability experiments

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6
Day	1	2	2	3	3	4
Av. temp (°C)	20	20	21	20	21	16
Time	16:00	11:20	14:20	10:30	15:50	09:30

The temperature for sample sets 1 – 5 was between 20 or 21 °C; only for set 6 was there a variation in temperature at 16 °C. The median spectra for each sample sets of ethanol samples from 0.5 – 3.0 % are plotted in figure 3.2-1 to 6. The only variation in the sample sets appears to be the intermittent inclusion of a sharp peak around 500 MHz.

Figure 3.2-1 Plots of GMS repeatability experiments spectra set 1

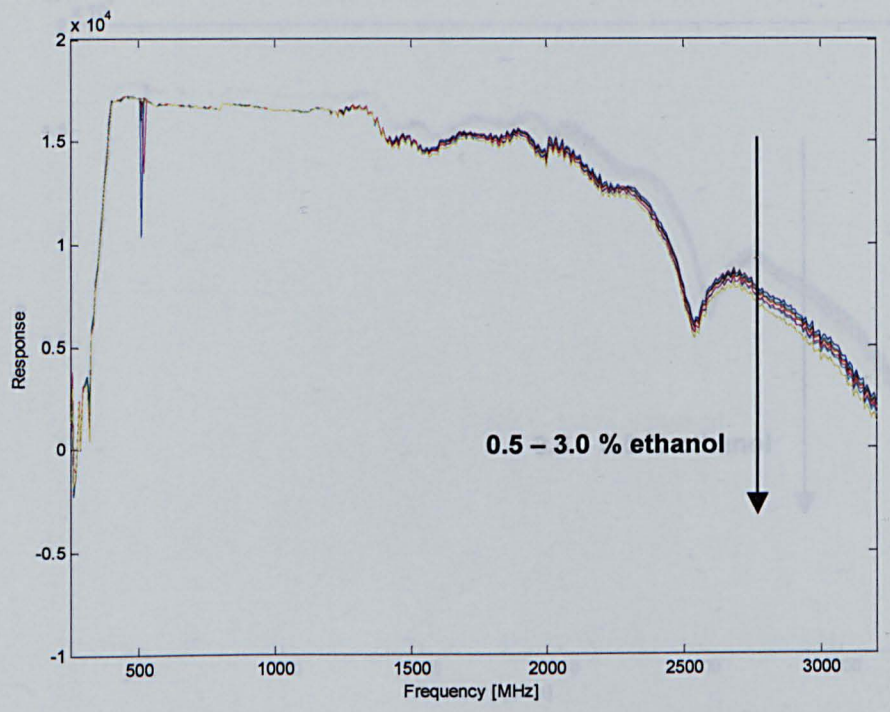


Figure 3.2-2 Plots of GMS repeatability experiments spectra set 2

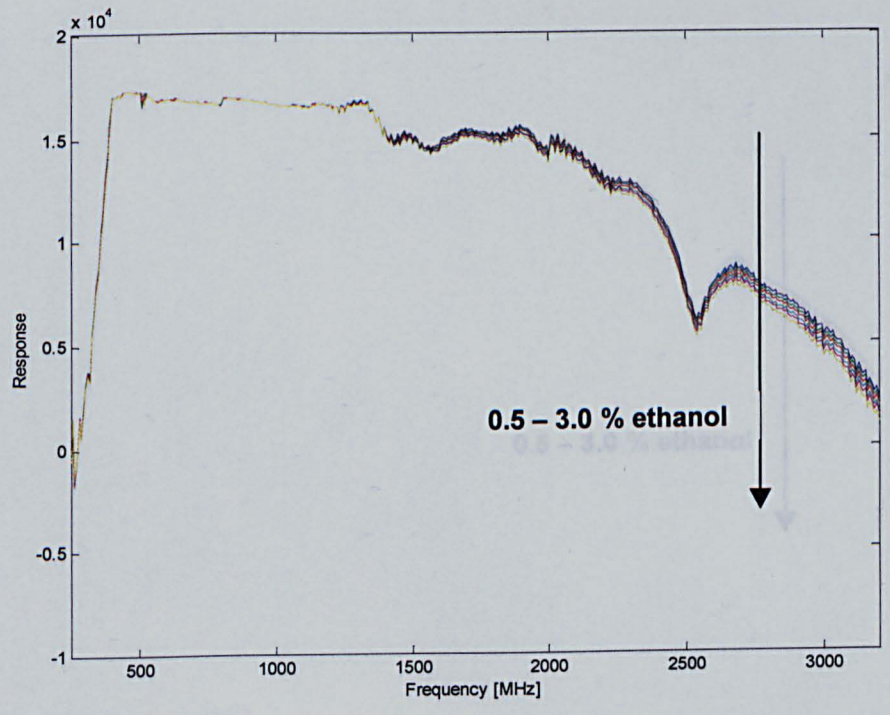


Figure 3.2-3 Plots of GMS repeatability experiments spectra set 3

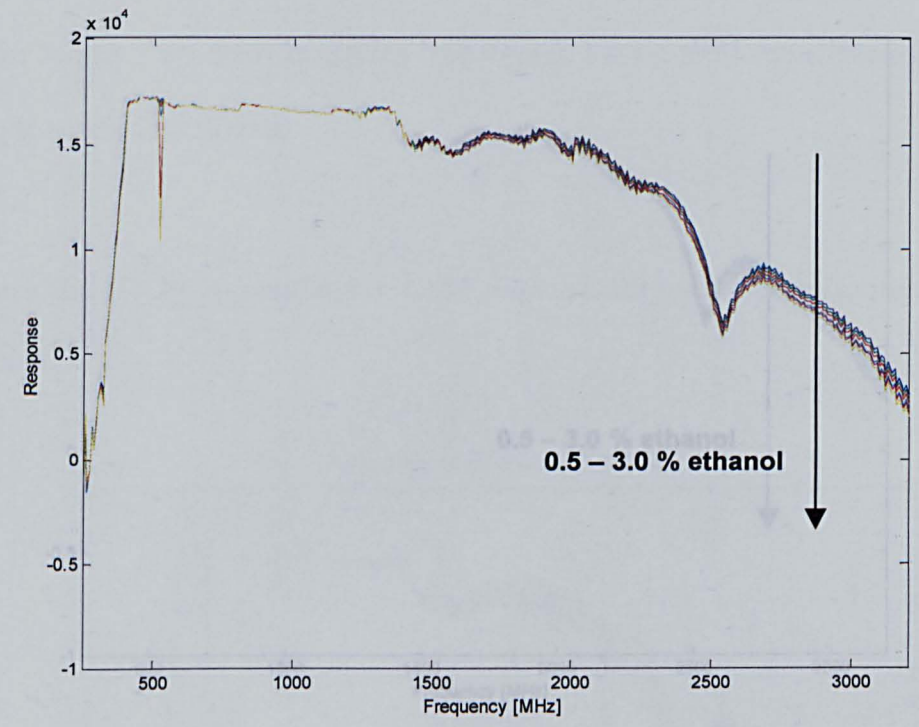


Figure 3.2-4 Plots of GMS repeatability experiments spectra set 4

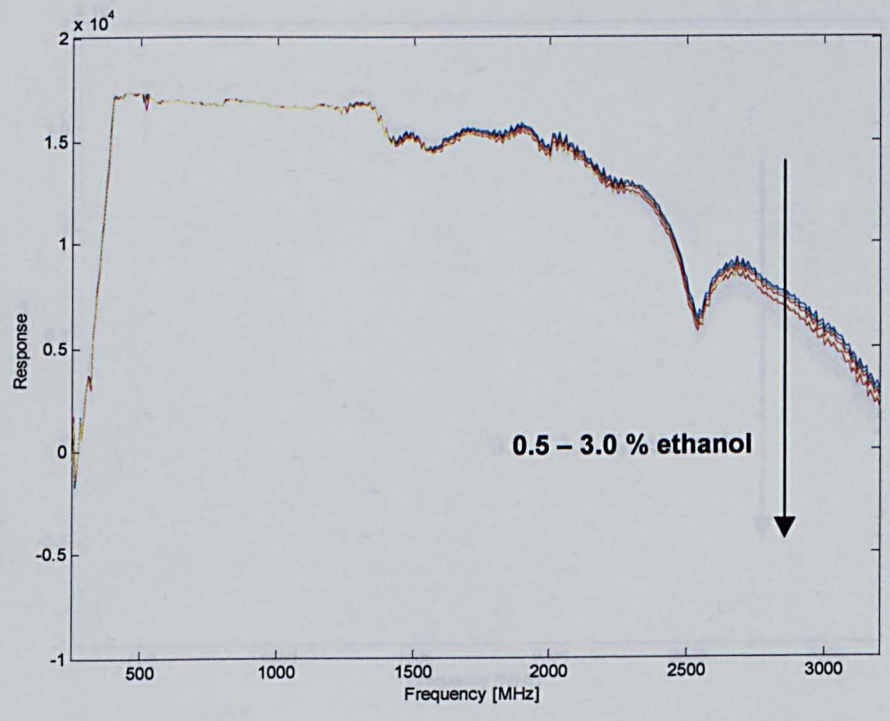


Figure 3.2-5 Plots of GMS repeatability experiments spectra set 5

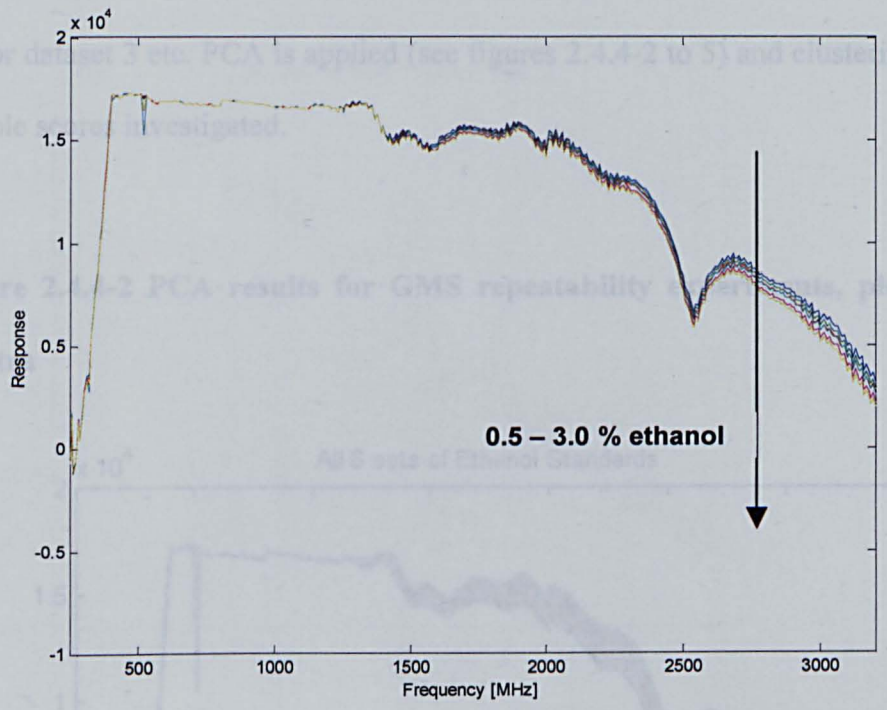
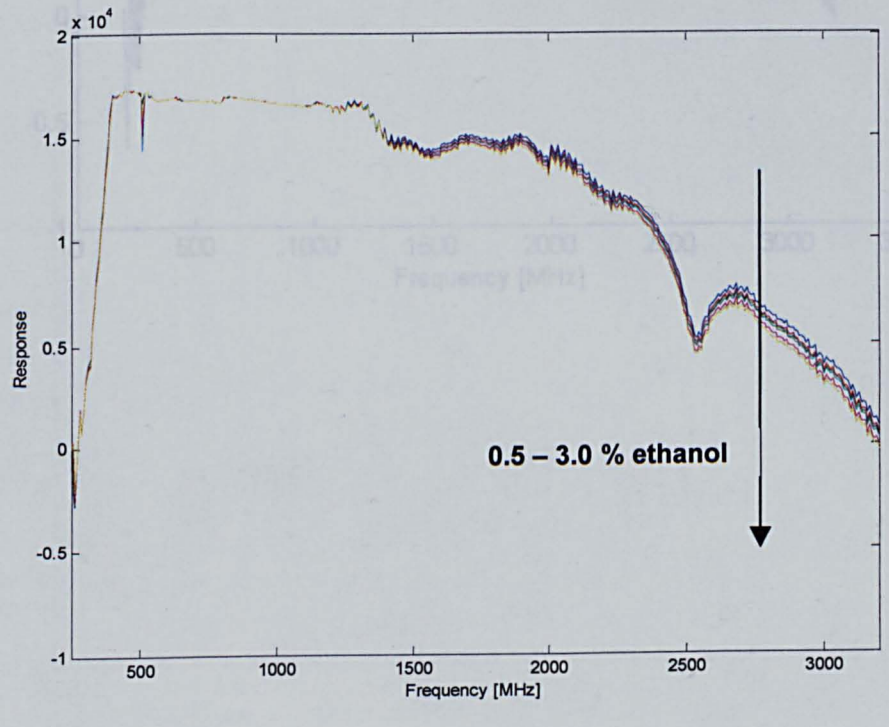


Figure 3.2-6 Plots of GMS repeatability experiments spectra set 6



The sample sets were grouped together; the six datasets contain 6 samples each of 0.5 – 3.0 % ethanol. The samples are numbered from 1 – 6 for dataset 1, then 7 - 12 for dataset 3 etc. PCA is applied (see figures 2.4.4-2 to 5) and clustering in the sample scores investigated.

Figure 2.4.4-2 PCA results for GMS repeatability experiments, plot of all spectra

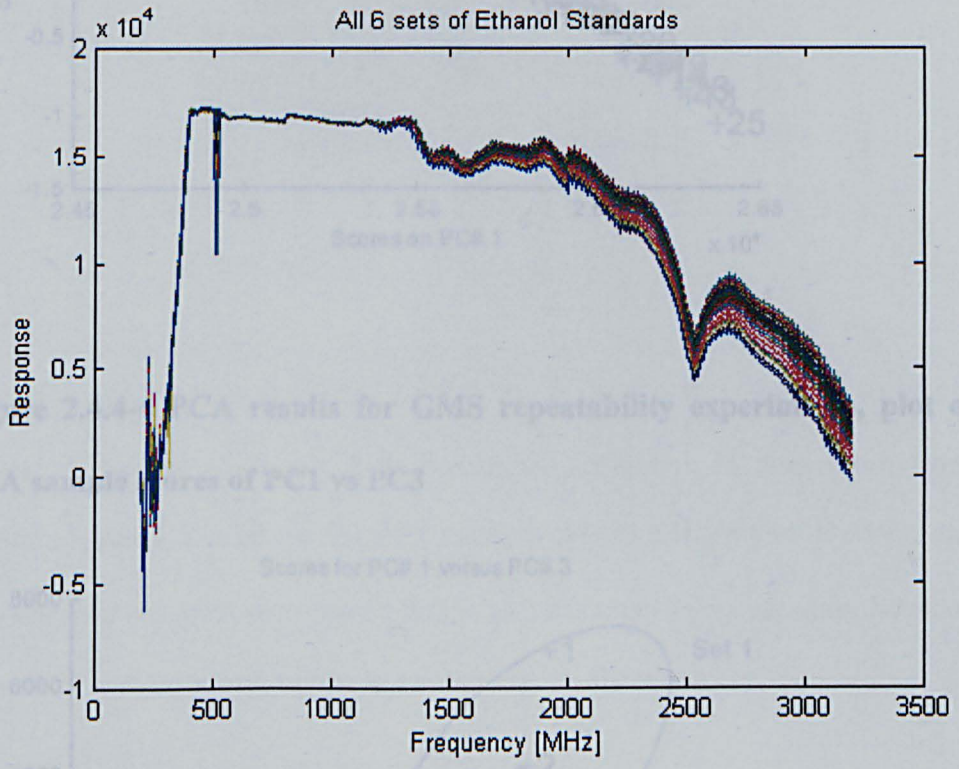


Figure 2.4.4-3 PCA results for GMS repeatability experiments, plot of all PCA sample scores of PC1 vs PC2

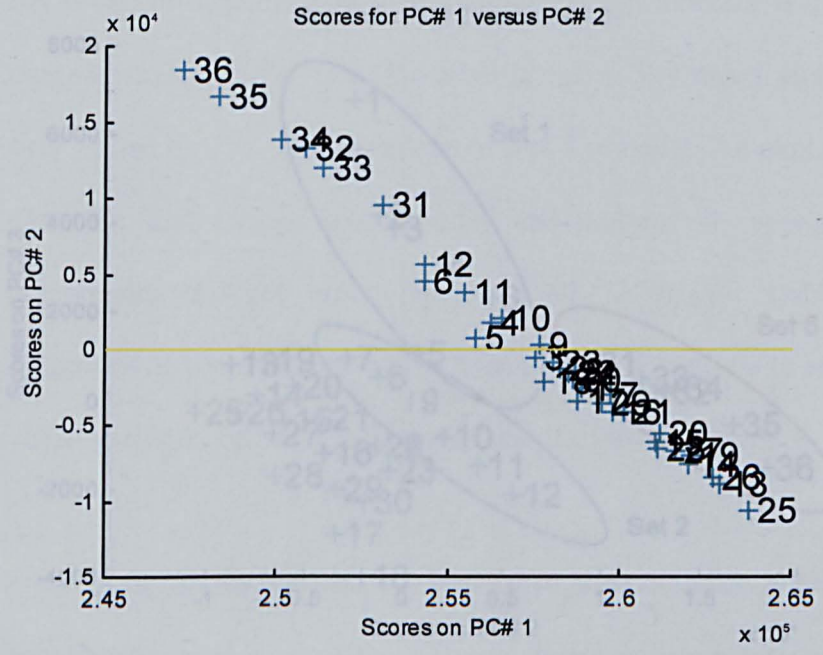


Figure 2.4.4-4 PCA results for GMS repeatability experiments, plot of all PCA sample scores of PC1 vs PC3

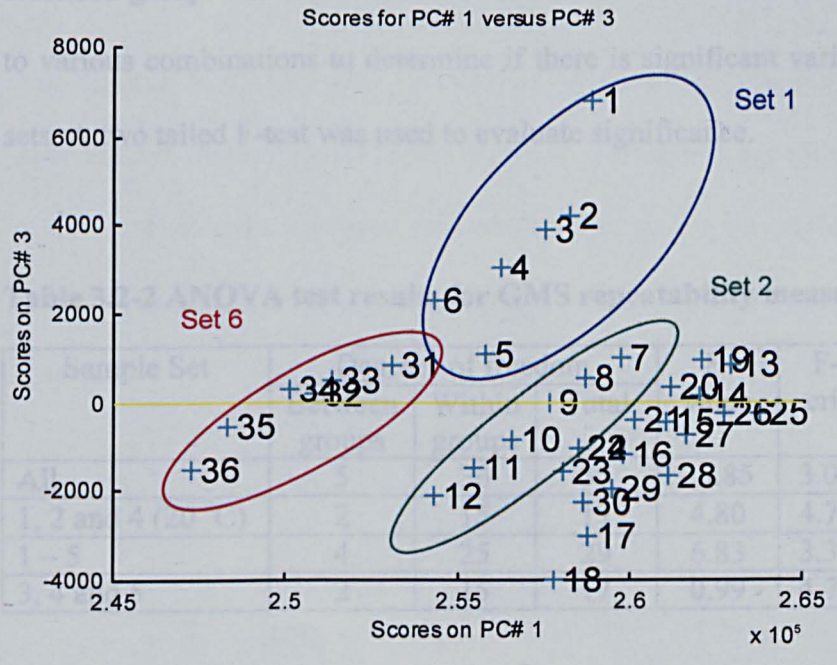
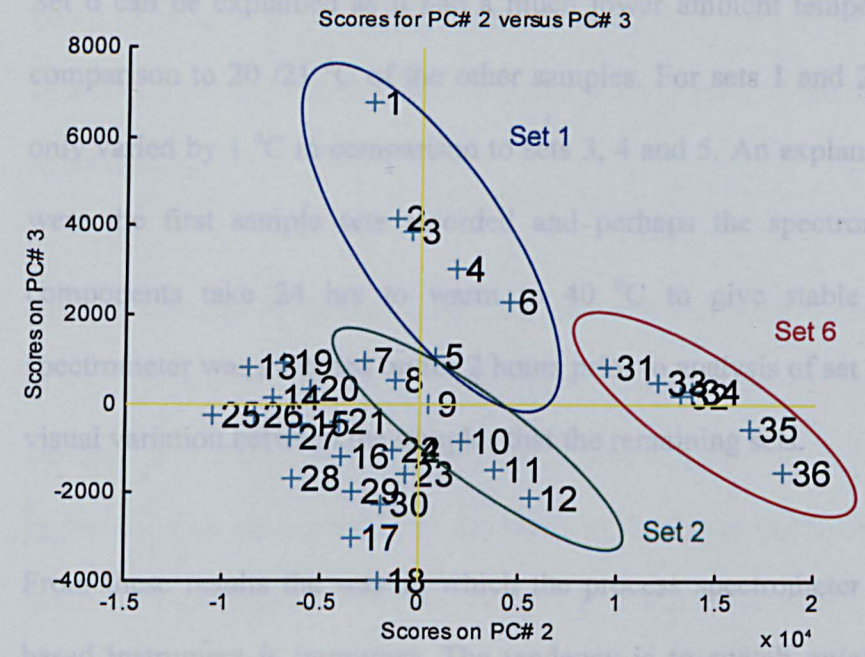


Figure 2.4.4-5 PCA results for GMS repeatability experiments, plot of all PCA sample scores of PC2 vs PC3



The PCA scores plots is figures 2.4.4-4 and 5 show sets 3, 4 and 5 to be closely clustered and sets 6, 1 and 2 to be further separated, in that order, from the clustered group. The scores for PC1 for each set were taken and ANOVA applied to various combinations to determine if there is significant variation between the sets. A two tailed F-test was used to evaluate significance.

Table 3.2-2 ANOVA test results for GMS repeatability measurements

Sample Set	Degrees of freedom			F-value	F-crit	Significant difference between sets?
	Between groups	Within groups	Total			
All	5	30	35	25.85	3.03	Yes
1, 2 and 4 (20 °C)	2	15	17	4.80	4.77	Yes
1 – 5	4	25	29	6.83	3.35	Yes
3, 4 and 5	2	15	17	0.99	4.77	No

The ANOVA results are consistent with the PCA scores plots showing that sample sets 3, 4 and 5 are closely related and 1, 2 and 6 are significantly different. Set 6 can be explained as it had a much lower ambient temperature (16 °C) in comparison to 20 /21 °C of the other samples. For sets 1 and 2, the temperature only varied by 1 °C in comparison to sets 3, 4 and 5. An explanation is that these were the first sample sets recorded and perhaps the spectrometer's electrical components take 24 hrs to warm to 40 °C to give stable recordings. The spectrometer was switched on for 2 hours prior to analysis of set 1. Set 2 has more visual variation between the samples that the remaining sets.

From these results the way in which the process spectrometer is used as a lab based instrument is important. The tendency is to switch on/off the instrument daily as for other laboratory based equipment, which could affect the spectrometers performance if a lengthy stabilisation time is required. In contrast, a process instrument would be in continuous use and rarely switched off.

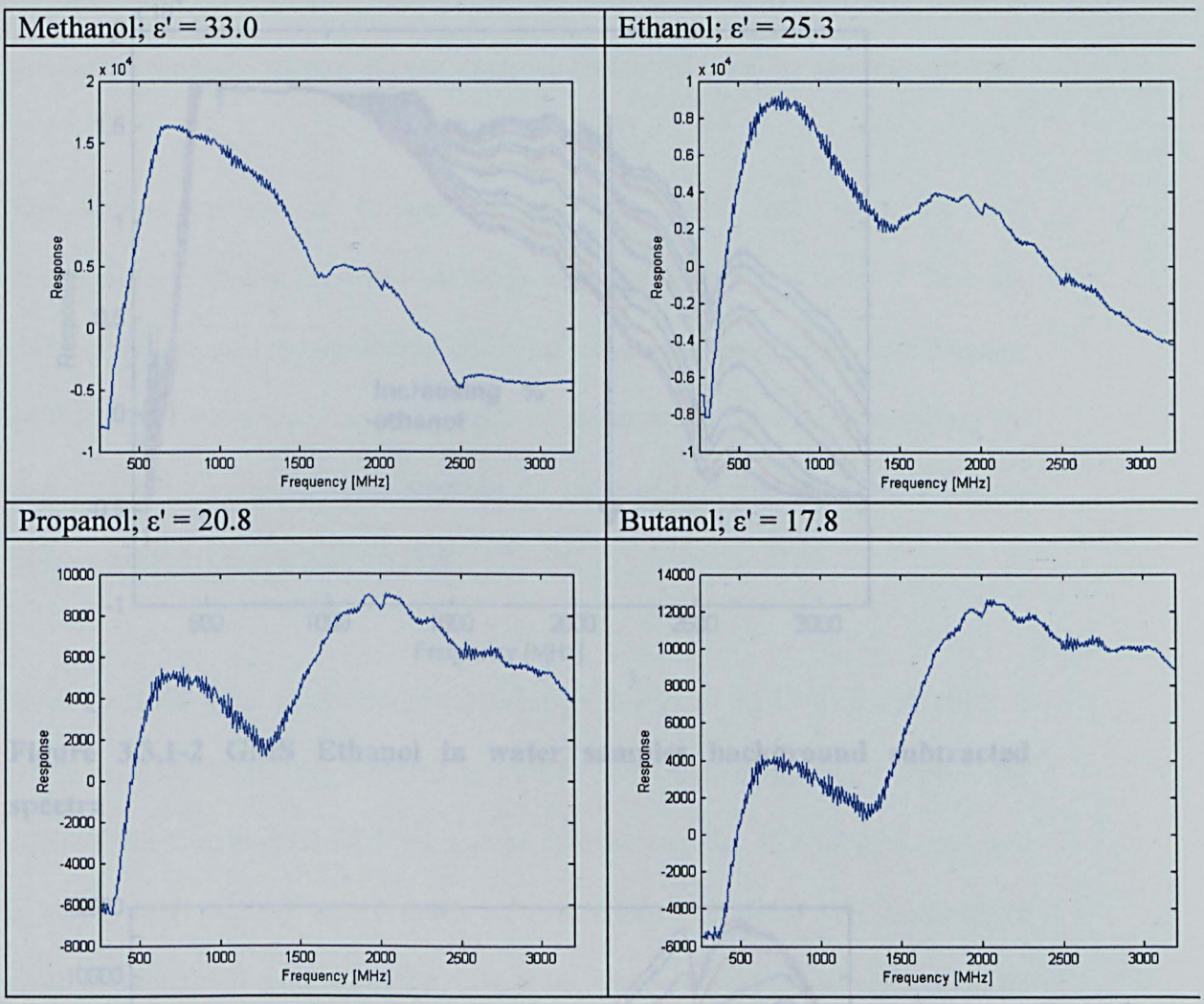
The sample temperature is also shown to be important. For experiments over several days the sample and sample chamber temperatures should be constant, for a series of samples analysed in a short time, where the ambient temperature is constant to a couple of degrees, control of temperature would be ideal but not essential.

3.3 Alcohol mixtures analysis

In this section various alcohol mixtures are analysed by GMS, mixtures compromising of 2 – 4 alcohols were analysed. The analysis of a homologous set of alcohols (or similar analytes), will give an indication of the limitations of GMS measurements.

The pure spectra of methanol, ethanol, propanol and butanol are plotted in figure 3.3-1. These pure spectra of each alcohol demonstrate the effect of dielectric constant and losses on spectral shape. From equation 18 it is known that the higher the dielectric constant of the material the lower the cut-off frequency will occur. Comparing the spectrum cut-offs in the measurements of methanol and butanol can see this. The general shapes of the spectra of the two compounds are also quite different. The cut-off frequency is sharp and the spectra well attenuated in the case of methanol. As the carbon chain length and size of the molecules increase the magnitude of the spectra is reduced. This is because as the number of carbons in the molecule increases it becomes more difficult for the wave to rotate the molecule. In effect the dielectric loss factors are such that the waves attenuation is reduced.

Figure 3.3-1 GMS spectra of pure methanol, ethanol, propanol and butanol



3.3.1 Analysis of ethanol in water

Samples of ethanol in water in the concentration range of 1 – 30 % ethanol were analysed by GMS with the aim to calibrate for ethanol concentration. The spectra and background subtracted spectra can be seen in figure 3.3.1-1.

Figure 3.3.1-1 GMS Ethanol in water samples spectra

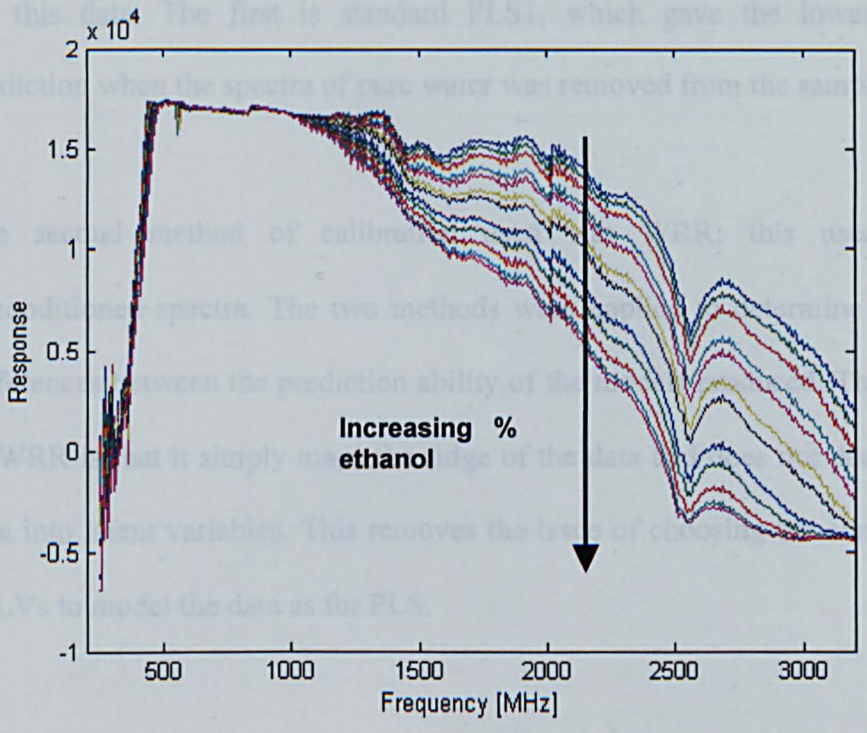
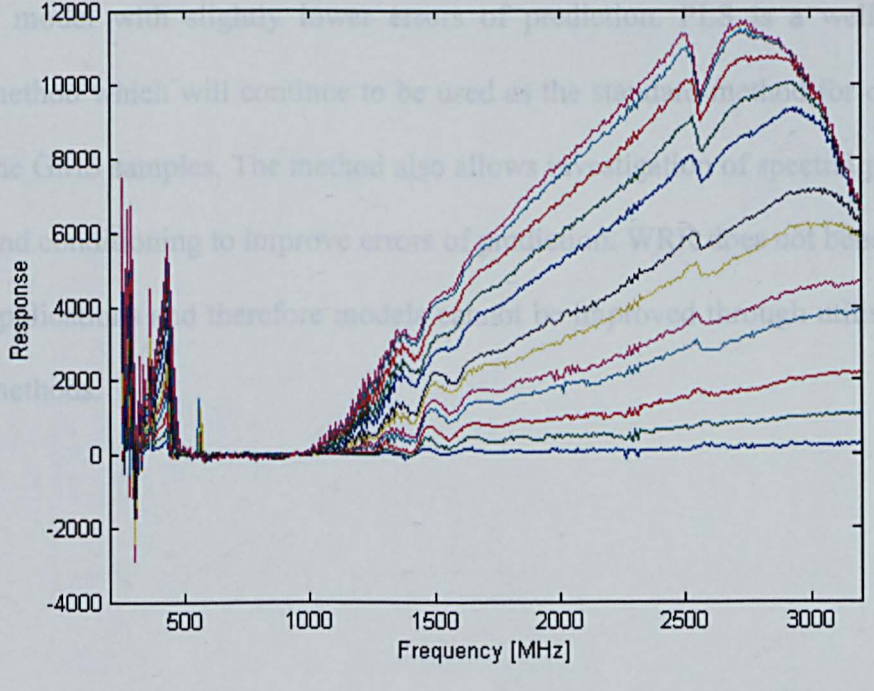


Figure 3.3.1-2 GMS Ethanol in water samples background subtracted spectra



Calibration models for ethanol prediction have been calculated by two methods for this data. The first is standard PLS1, which gave the lowest errors of prediction when the spectra of pure water was removed from the samples spectra.

The second method of calibration used was WRR; this used the raw, unconditioned spectra. The two methods were applied to determine if there are differences between the prediction ability of the models produced. The advantage of WRR is that it simply maps the ridge of the data and does not decompose the data into latent variables. This removes the issue of choosing the correct number of LVs to model the data as for PLS.

Both methods gave good errors of prediction; for PLS1 (2 LVs) the RMSPE is 0.266 and for WRR the RMSPE is 0.284. The small difference between the RMSPE for each method show them to be comparable. The PLS method produced a model with slightly lower errors of prediction. PLS is a well documented method which will continue to be used as the standard method for calibration for the GMS samples. The method also allows investigation of spectral pre-treatments and conditioning to improve errors of prediction. WRR does not benefit from such applications and therefore models cannot be improved through utilisation of such methods.

Figure 3.3.1-2 PLS1 calibration model, plot of actual % ethanol vs predicted % ethanol for the analysis of ethanol in water by GMS; RMSPE = 0.28

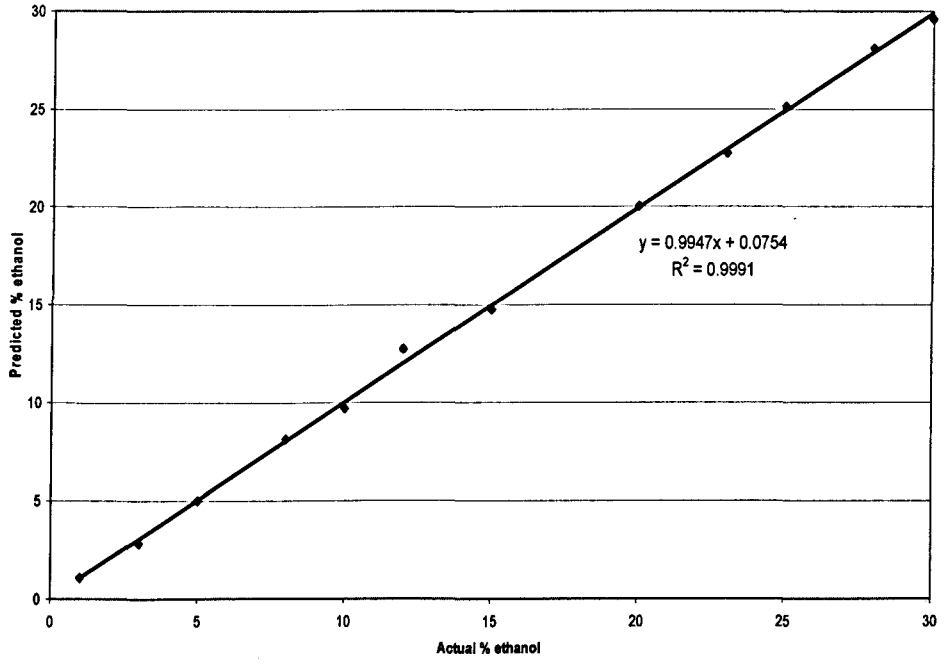
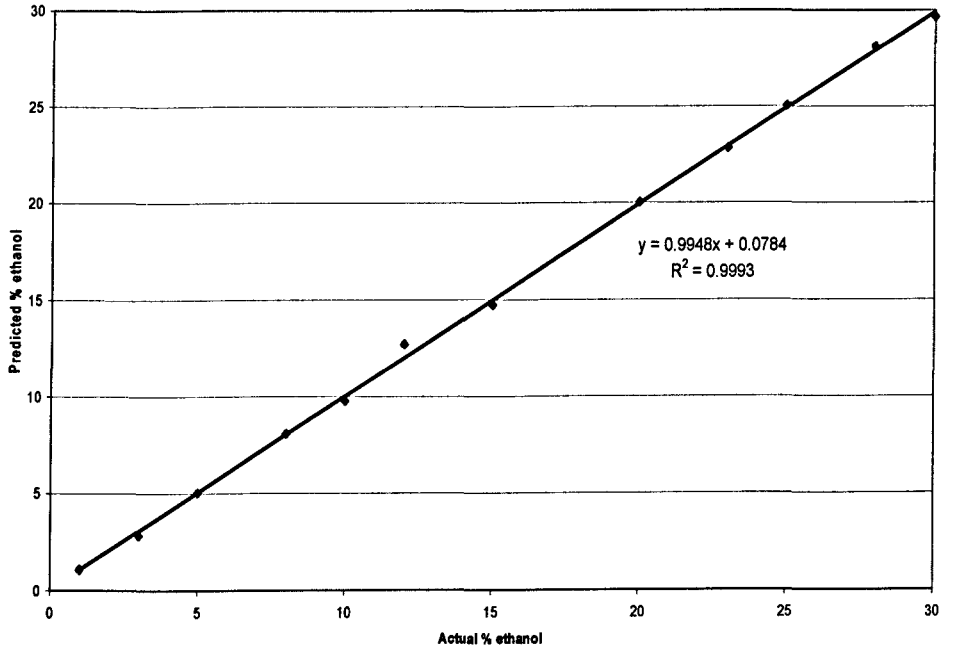


Figure 3.3.1-3 WRR calibration model, plot of actual vs predicted % ethanol for the analysis of ethanol in water by GMS; RMSPE = 0.27



3.3.2 Analysis of binary mixtures of alcohols

Alcohol mixtures; set 1: methanol and ethanol, set 2: methanol and propanol, and set 3: ethanol and propanol were analysed by GMS. The samples' spectra are plotted for each mixture set in figures 3.3.2-1, 3 and 5. The PLS CV results are also plotted in figures 3.3.2-2, 4 and 6.

All the spectral sets have good variation in response as the sample concentrations vary. As expected from the pure spectra, samples containing methanol had the highest spectral response. Dataset 2 was modelled with the lowest errors. The spectra of this dataset had the most variation in magnitude response between the samples throughout the entire spectral region. Set 2 consisted of methanol and propanol, these have the least similar pure spectrum out of those in the binary alcohol sample mixtures. For datasets 1 and 3 there is only good variation for approximately half of the frequency range.

To model the data, the spectra were first mean centred and then PLS2 applied.

Table 3.3.2-1 Binary mixtures of alcohols PLS2 RMSEC and RMSECV

Set	Average RMSEC	Average RMSECV
1	0.62	1.04
2	0.18	0.33
3	0.48	1.06

The sample concentrations are in the range of 30 – 70 % alcohol so CV errors of approximately 1 % are acceptable for prediction for a 2 LV model, demonstrating that GMS is suitable for this analysis of simple 2 component alcohol mixtures.

Figure 3.3.2-1 Binary mixtures of alcohols set 1, GMS spectra

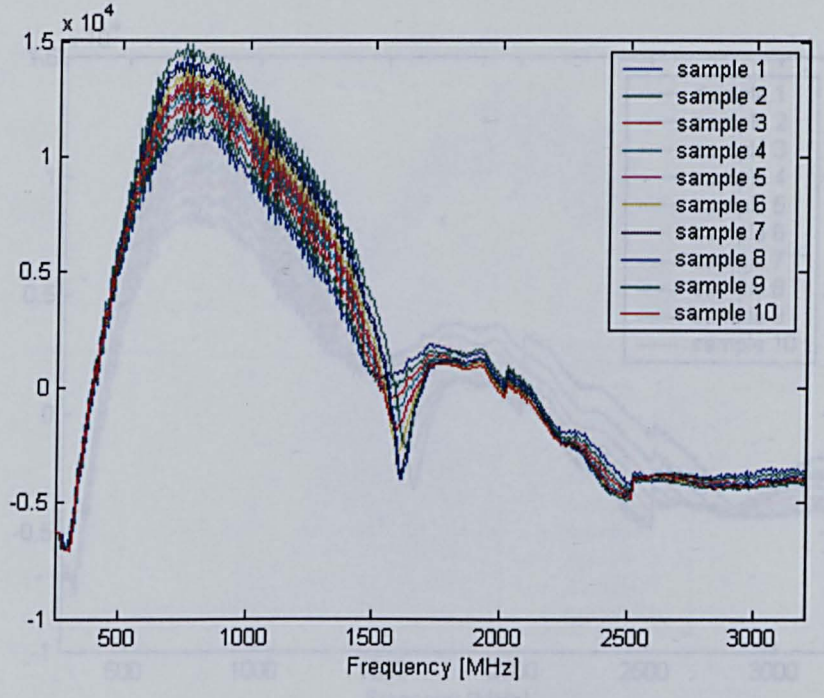


Figure 3.3.2-2 Binary mixtures of alcohol set 1, PLS2 CV model plot of RMSEC and RMSECV

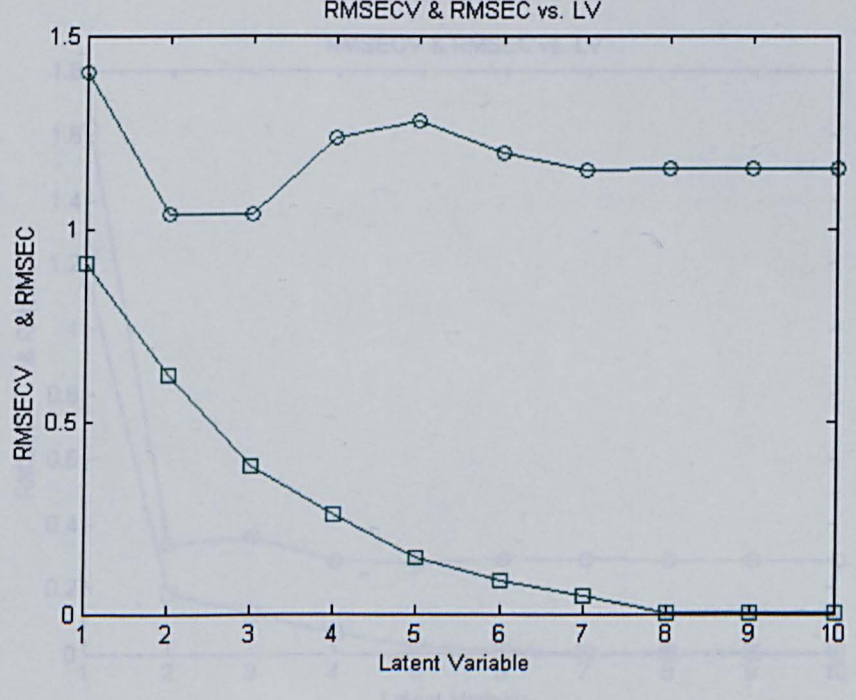


Figure 3.3.2-3 Binary mixtures of alcohols set 2, GMS spectra

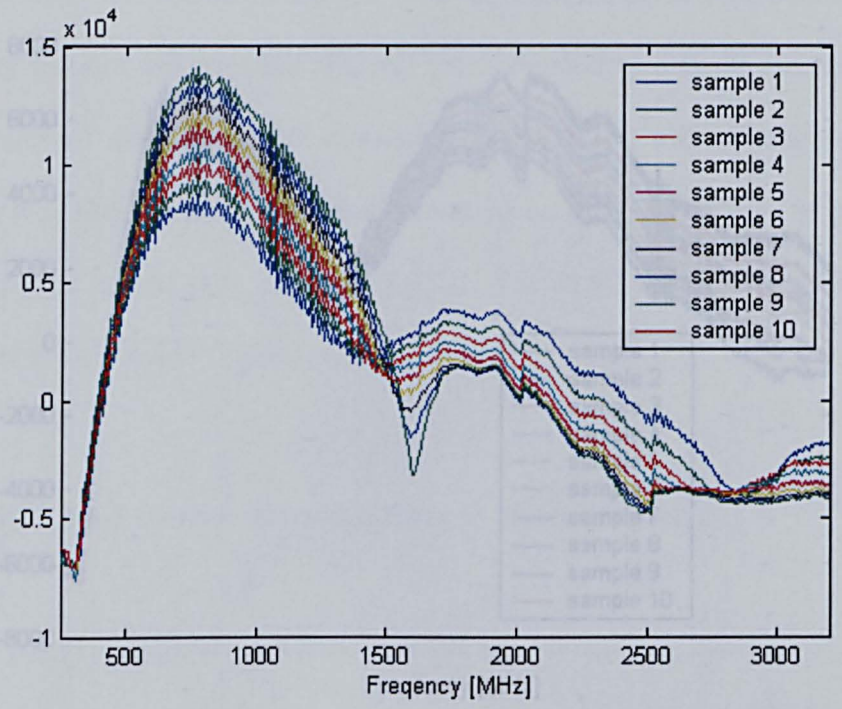


Figure 3.3.2-4 Binary mixtures of alcohol set 2, PLS2 CV model plot of RMSEC and RMSECV

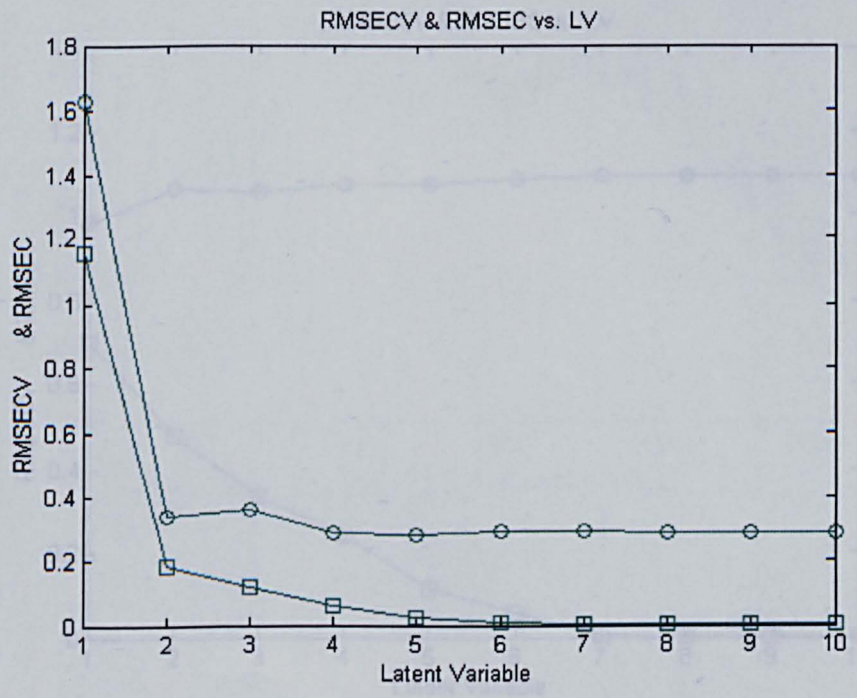


Figure 3.3.2-5 Binary mixtures of alcohol set 3, GMS spectra

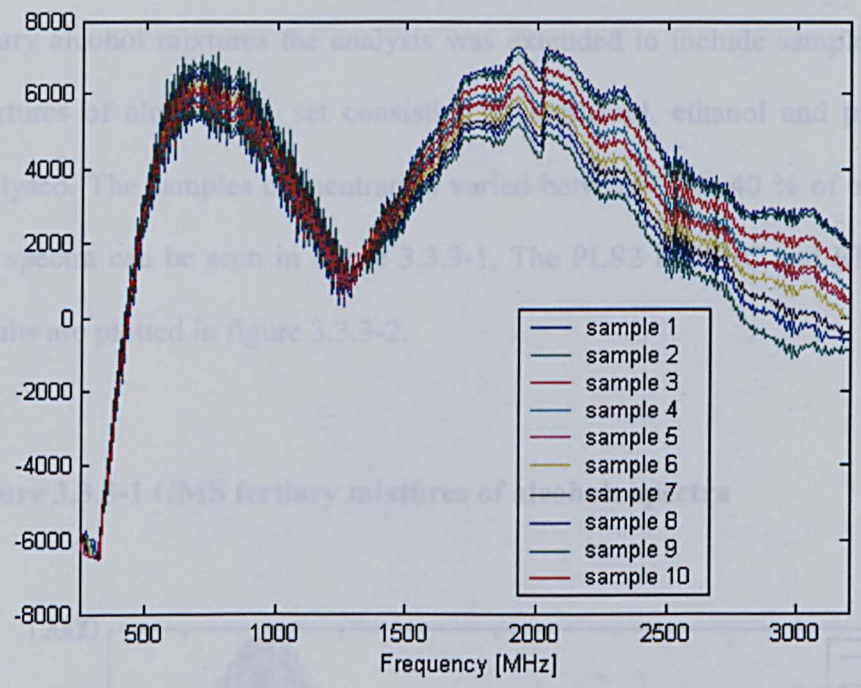
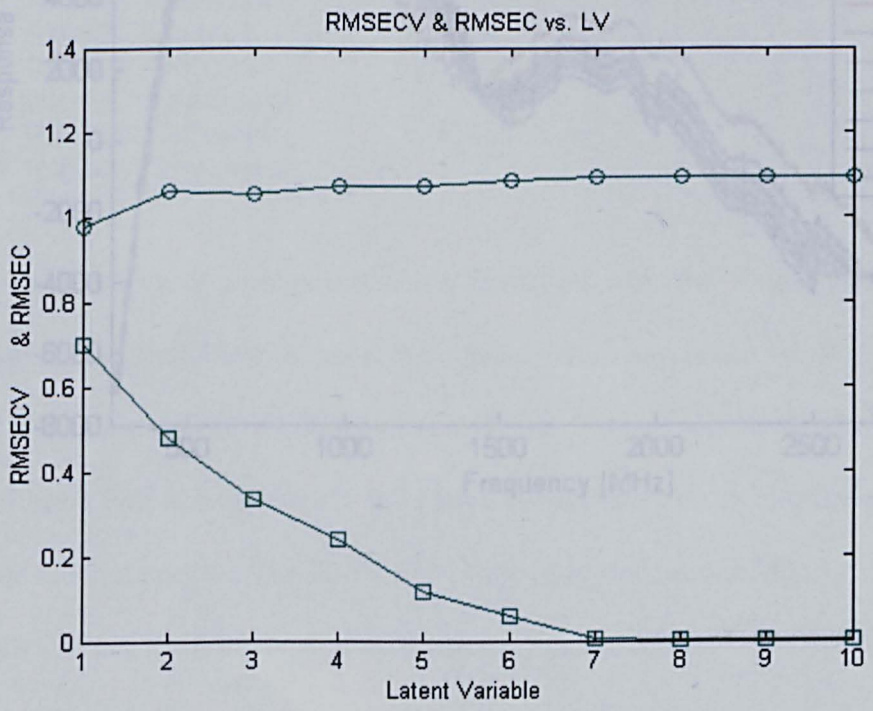


Figure 3.3.2-6 Binary mixtures of alcohol set 3, PLS2 CV model plot of RMSEC and RMSECV



3.3.3 Analysis of tertiary mixtures of alcohols

Since GMS has shown in the previous experiments to be suitable for analysis of binary alcohol mixtures the analysis was extended to include samples of tertiary mixtures of alcohols. A set consisting of methanol, ethanol and propanol was analysed. The samples concentration varied between 20 – 40 % of each alcohol, the spectra can be seen in figure 3.3.3-1. The PLS2 RMSEC and RMSECV CV results are plotted in figure 3.3.3-2.

Figure 3.3.3-1 GMS tertiary mixtures of alcohols spectra

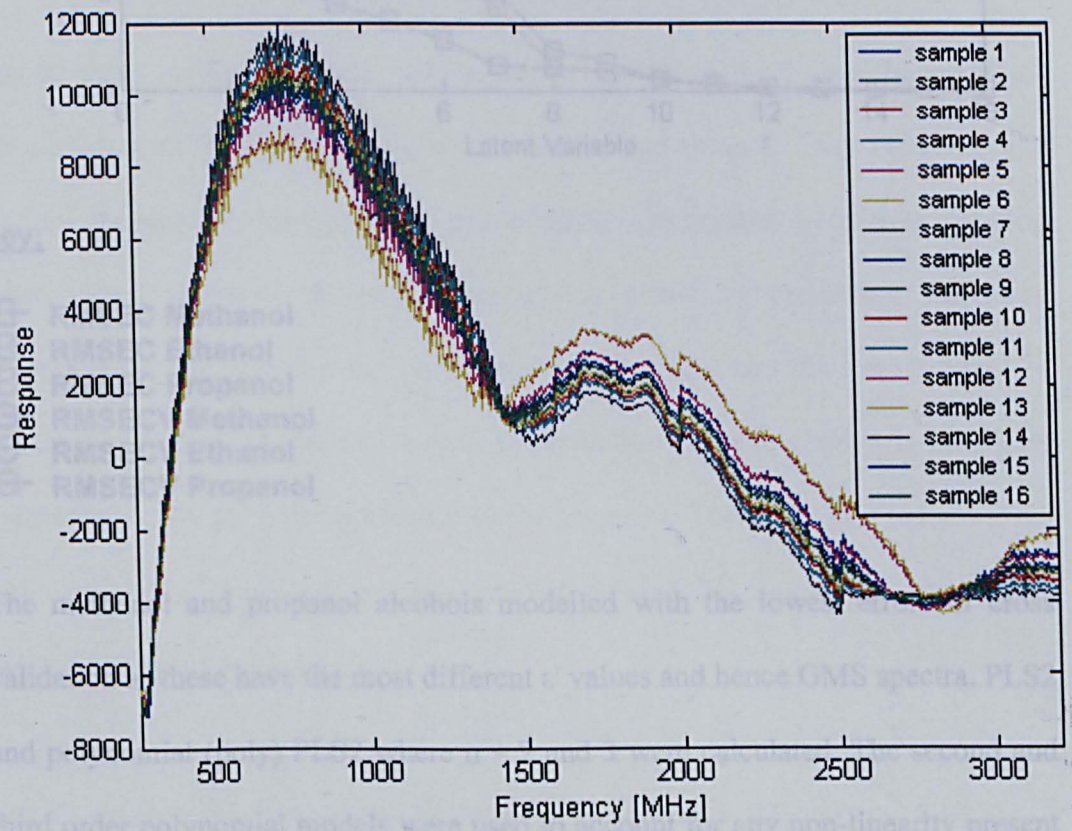
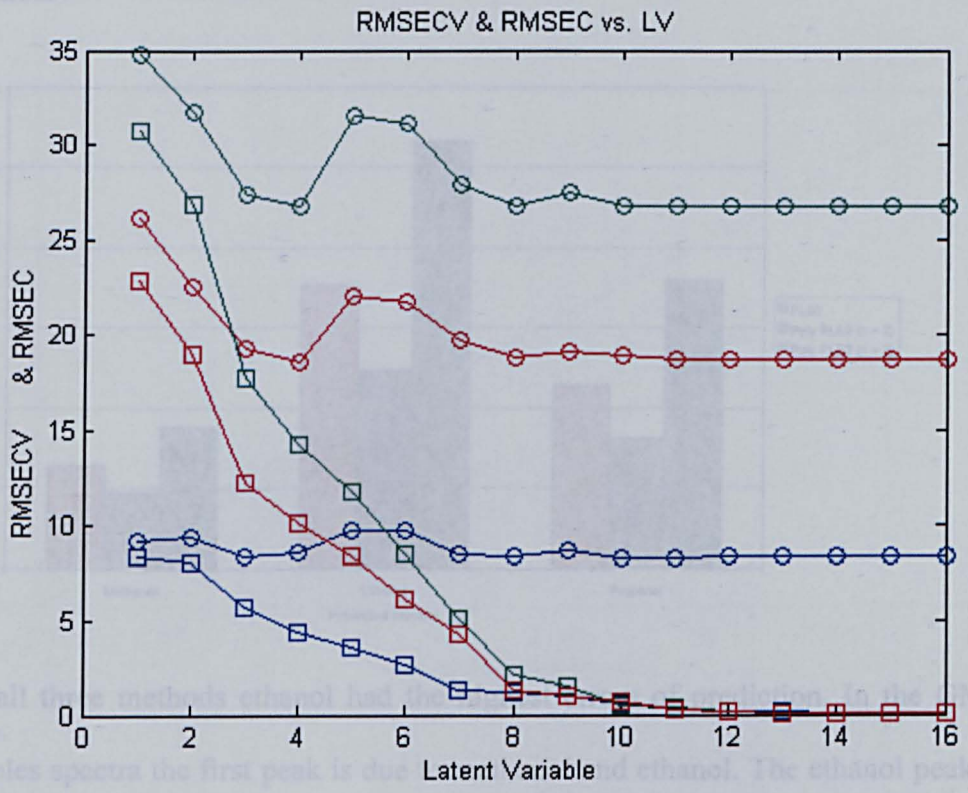


Figure 3.3.3-2 PLS2 CV model results of GMS tertiary mixtures of alcohols

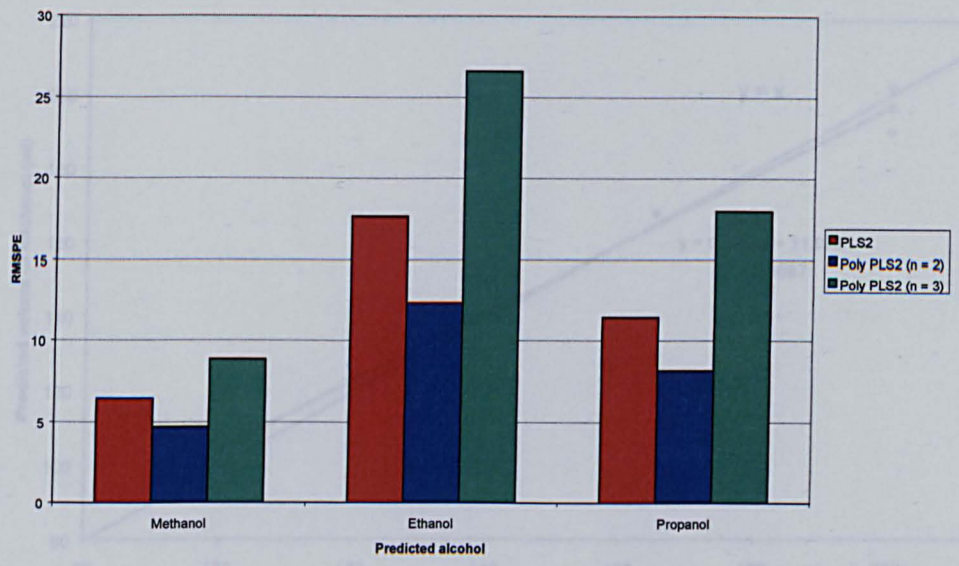


Key;

RMSEC Methanol
RMSEC Ethanol
RMSEC Propanol
RMSECV Methanol
RMSECV Ethanol
RMSECV Propanol

The methanol and propanol alcohols modelled with the lowest errors of cross validation as these have the most different ε' values and hence GMS spectra. PLS2 and polynomial (poly) PLS2 where n = 2 and 3 were calculated. The second and third order polynomial models were used to account for any non-linearity present in the alcohol spectra. The RMSPE's of the independent validation samples based on a 4 LV model plotted for each dataset in figure 3.3.3-.3.

Figure 3.3.3-3 PLS2 and Poly PLS2 RMSPE results for tertiary mixtures of alcohols



For all three methods ethanol had the highest errors of prediction. In the GMS samples spectra the first peak is due to methanol and ethanol. The ethanol peak is half the magnitude of the response of methanol. The methanol could be masking ethanol and propanol peaks, increasing the errors of prediction for these components. The second order polynomial PLS2 model had the lowest errors of prediction. For this model the % error in prediction for each alcohol is 2.9 % for methanol, 7.5 % for ethanol and 4.9 % for propanol. The PLS2 and second order polynomial PLS2 results of actual vs predicted % alcohol for each component are plotted in figures 3.3.3-4 to 9.

For all alcohols the lines of best fit have a very high intercept showing a large off-set in the line and hence a skew (also seen in the deviation from the y = x line) in the calibration predictions. The magnitude of the prediction offset is consistent with the RMSPE values.

Figure 3.3.3-4 PLS2 results for set 4 alcohol sample mixtures; plots of actual vs predicted % methanol (ml)

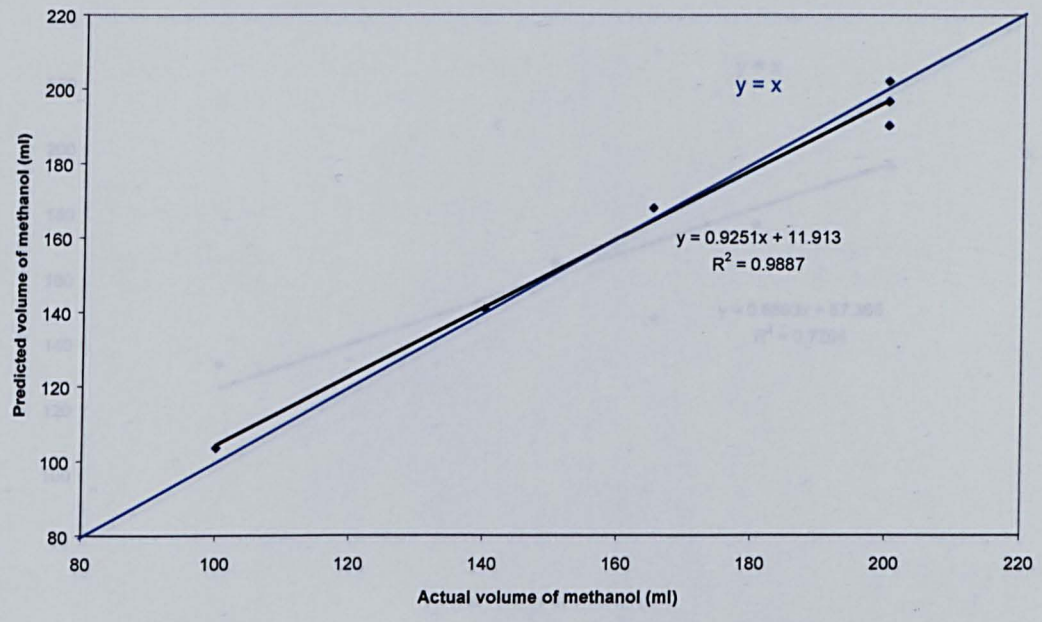


Figure 3.3.3-5 Poly PLS2 results for set 4 alcohol sample mixtures; plots of actual vs predicted % methanol (ml)

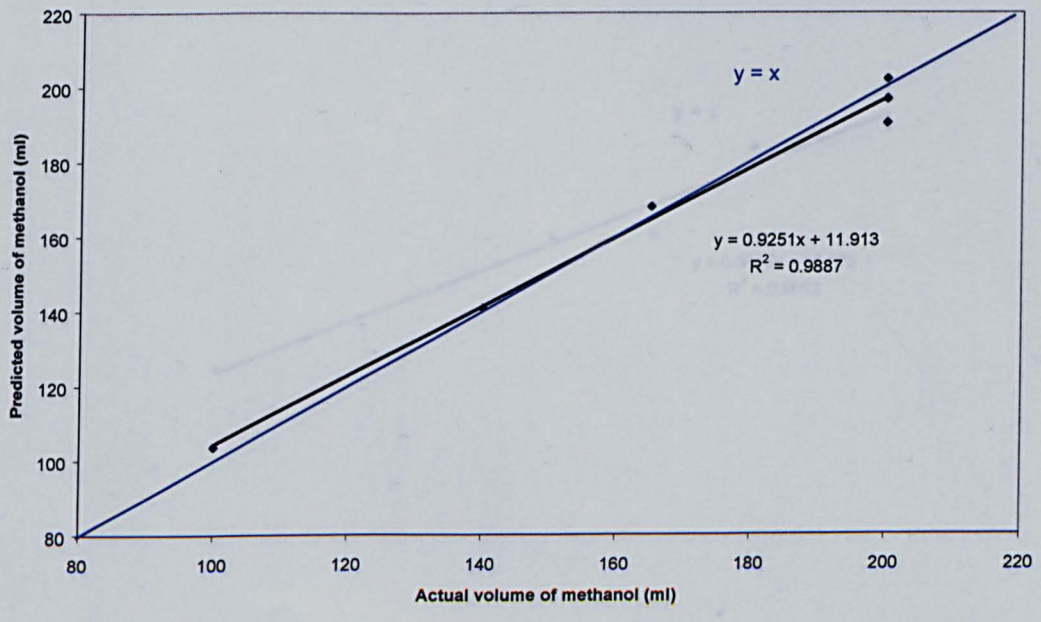


Figure 3.3.3-6 PLS2 results for set 4 alcohol sample mixtures; plots of actual vs predicted % ethanol (ml)

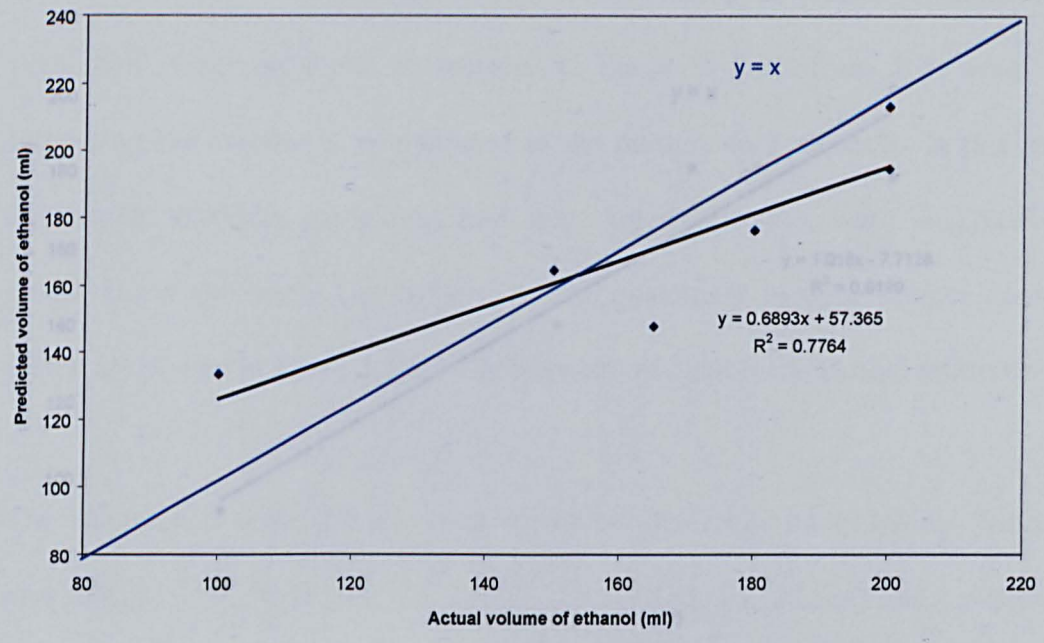


Figure 3.3.3-7 Poly PLS2 results for set 4 alcohol sample mixtures; plots of actual vs predicted % ethanol (ml)

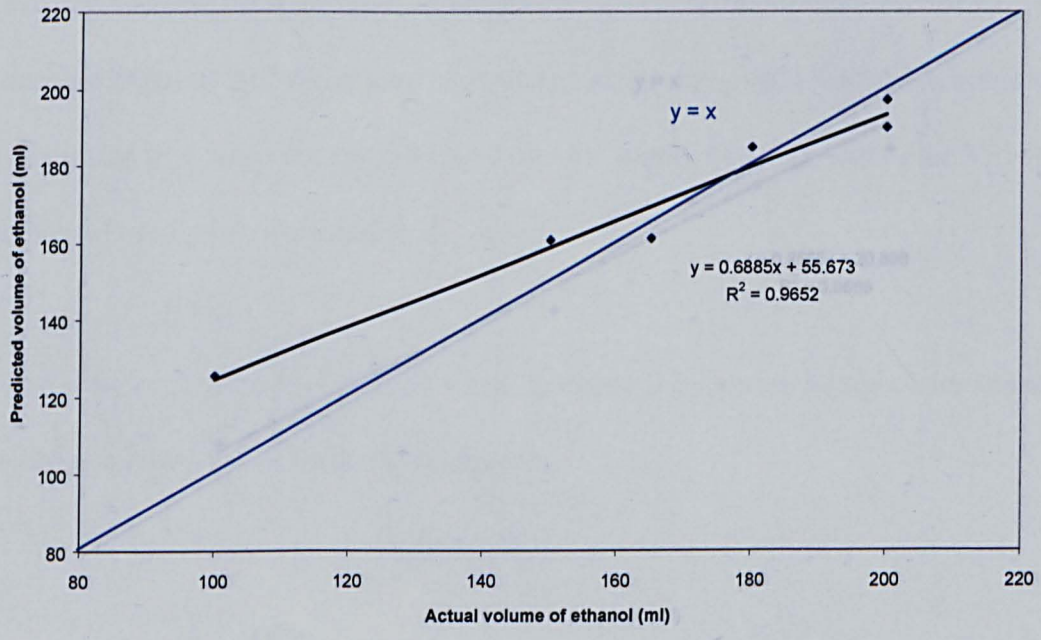


Figure 3.3.3-8 PLS2 results for set 4 alcohol sample mixtures; plots of actual vs predicted % propanol (ml)

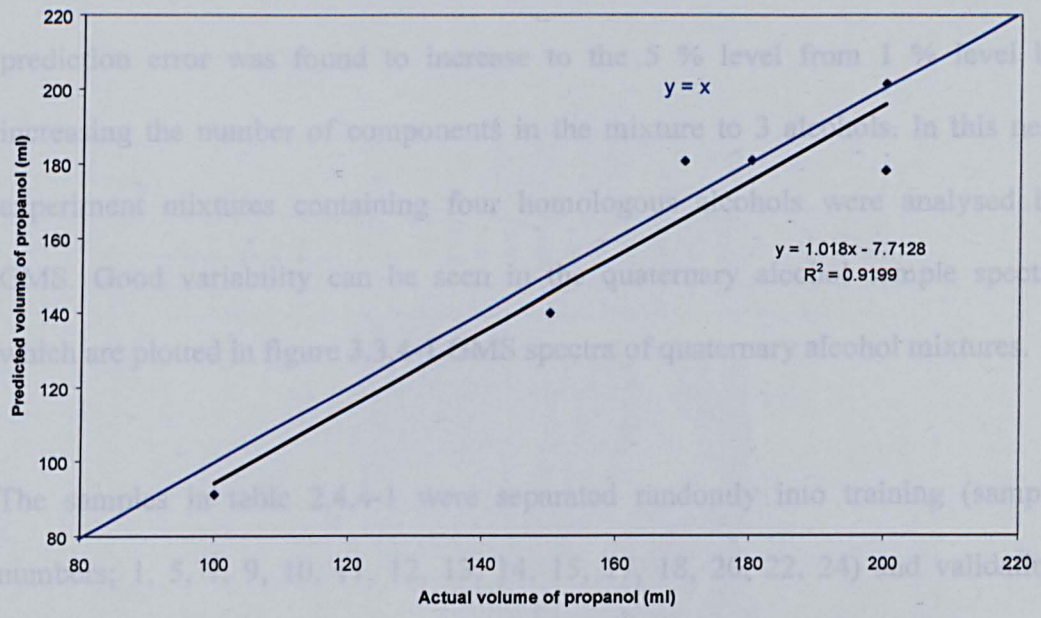
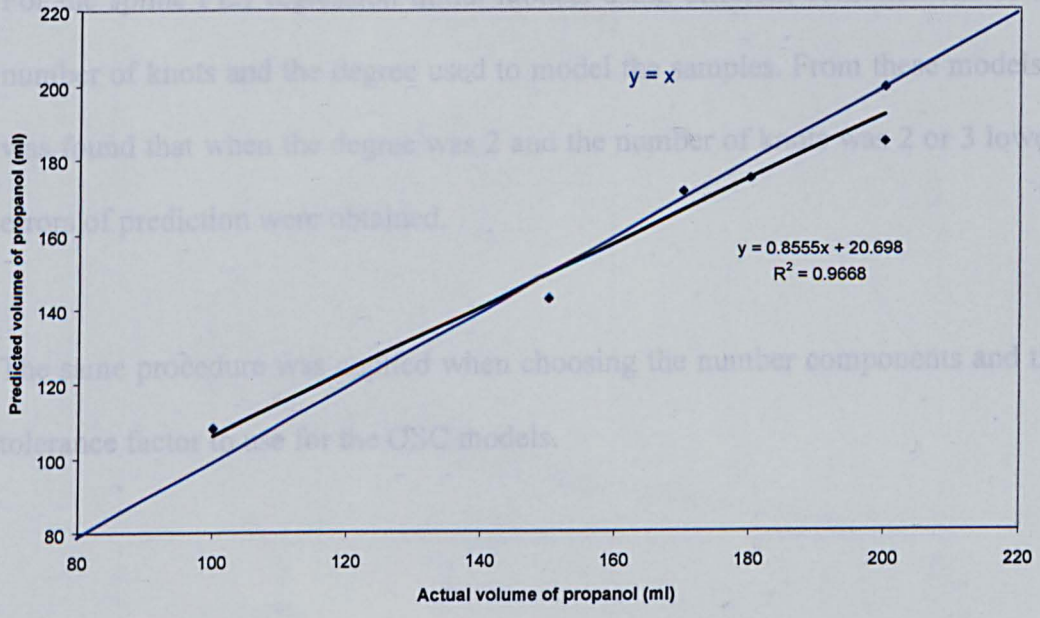


Figure 3.3.3-9 Poly PLS2 results for set 4 alcohol sample mixtures; plots of actual vs predicted % propanol (ml)



3.3.4 Analysis of quaternary mixtures of alcohols

The previous two experiments have shown GMS spectra of binary and tertiary alcohol mixtures to be suitable for the determination of alcohol content. The prediction error was found to increase to the 5 % level from 1 % level by increasing the number of components in the mixture to 3 alcohols. In this next experiment mixtures containing four homologous alcohols were analysed by GMS. Good variability can be seen in the quaternary alcohol sample spectra which are plotted in figure 3.3.4-1 GMS spectra of quaternary alcohol mixtures.

The samples in table 2.4.4-1 were separated randomly into training (sample numbers; 1, 5, 7, 9, 10, 11, 12, 13, 14, 15, 17, 18, 20, 22, 24) and validation datasets (sample numbers; 2, 3, 4, 6, 8, 16, 19, 21, 23). A range of models were produced and the predictions of the validation samples based on models using 1 – 6 LVs plotted in figure 3.3.4-2 for the PLS models.

For the spline PLS regression initial models using different combinations of the number of knots and the degree used to model the samples. From these models it was found that when the degree was 2 and the number of knots was 2 or 3 lowest errors of prediction were obtained.

The same procedure was applied when choosing the number components and the tolerance factor to use for the OSC models.

Figure 3.3.4-1 GMS spectra of quaternary alcohol mixtures

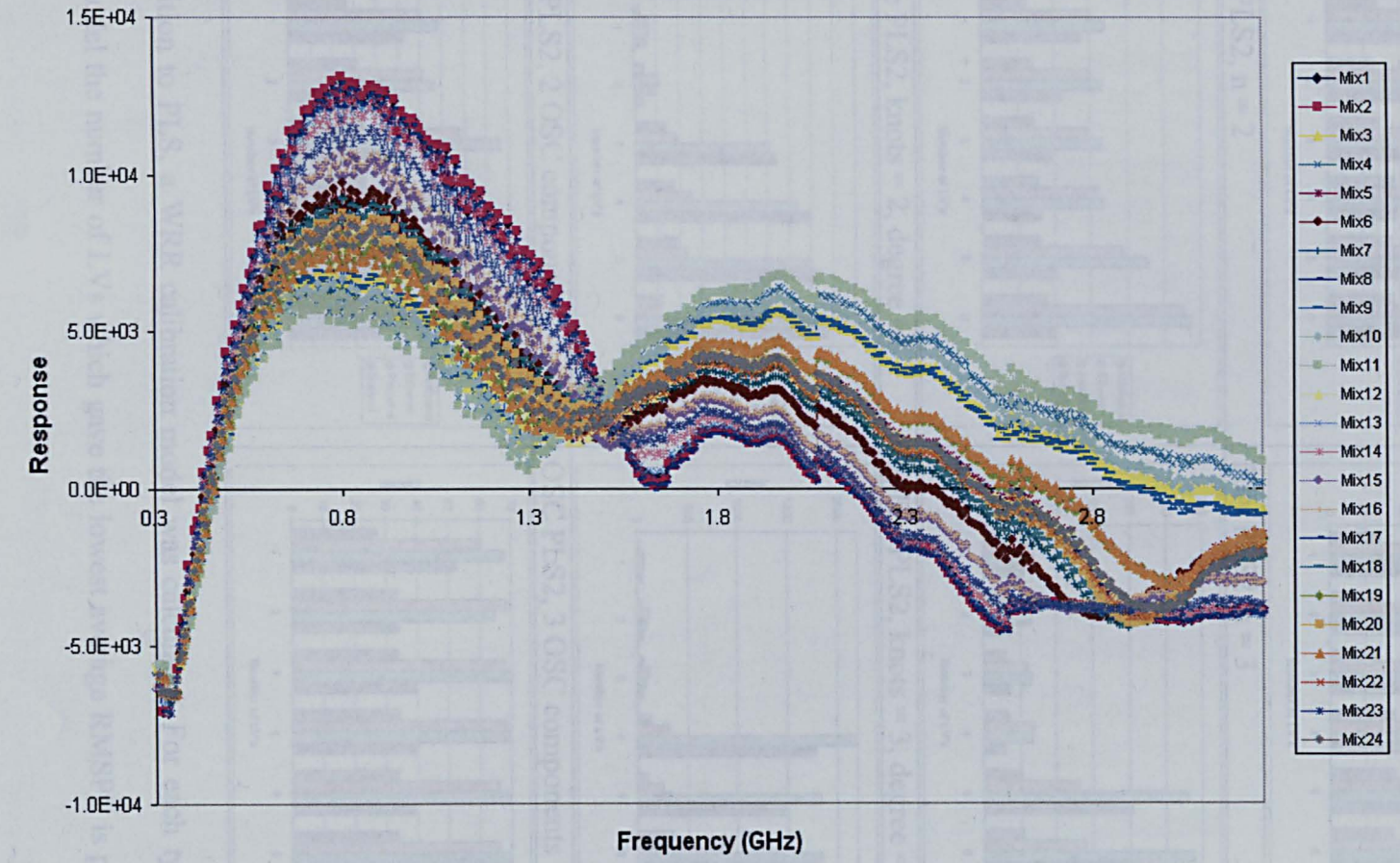
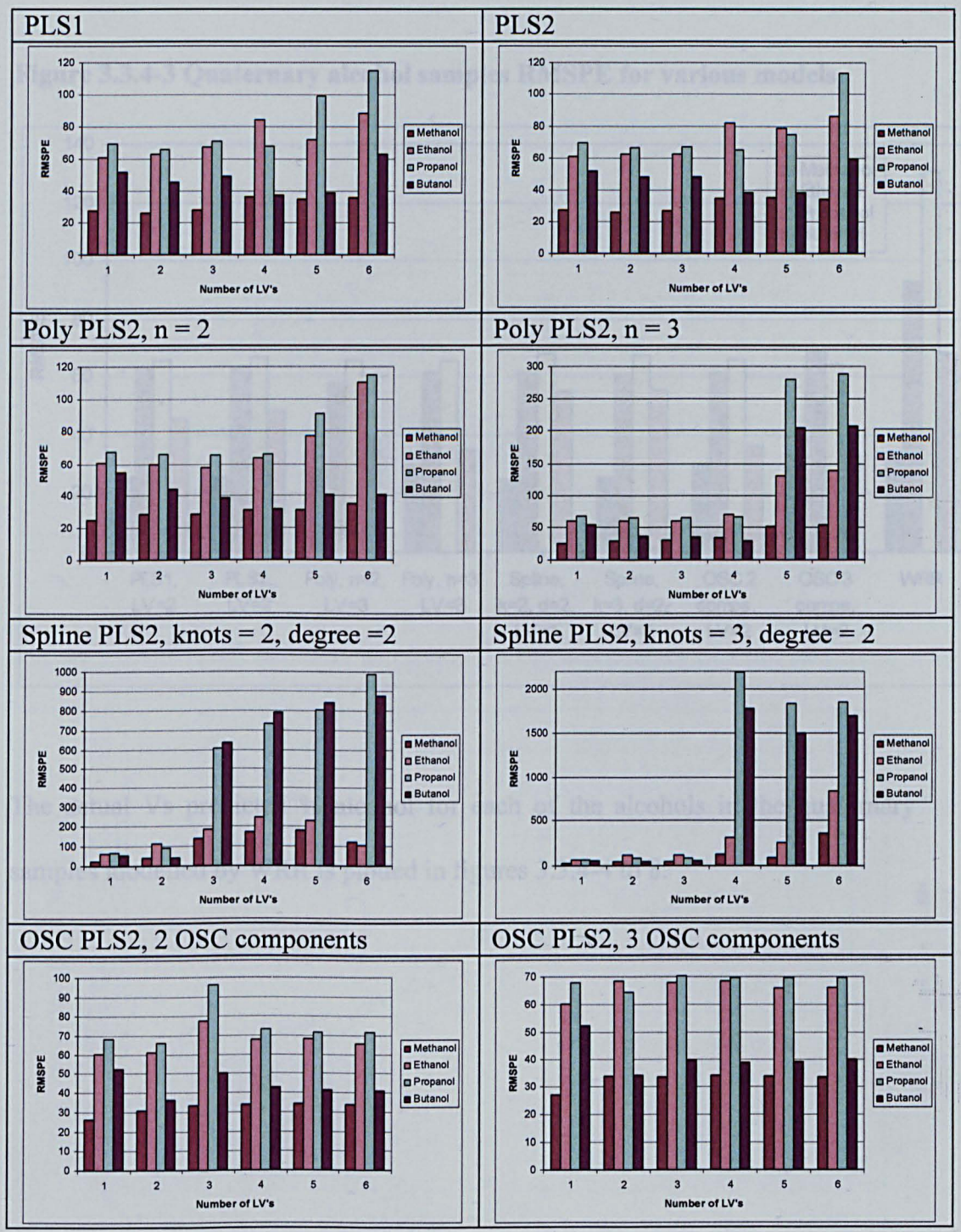


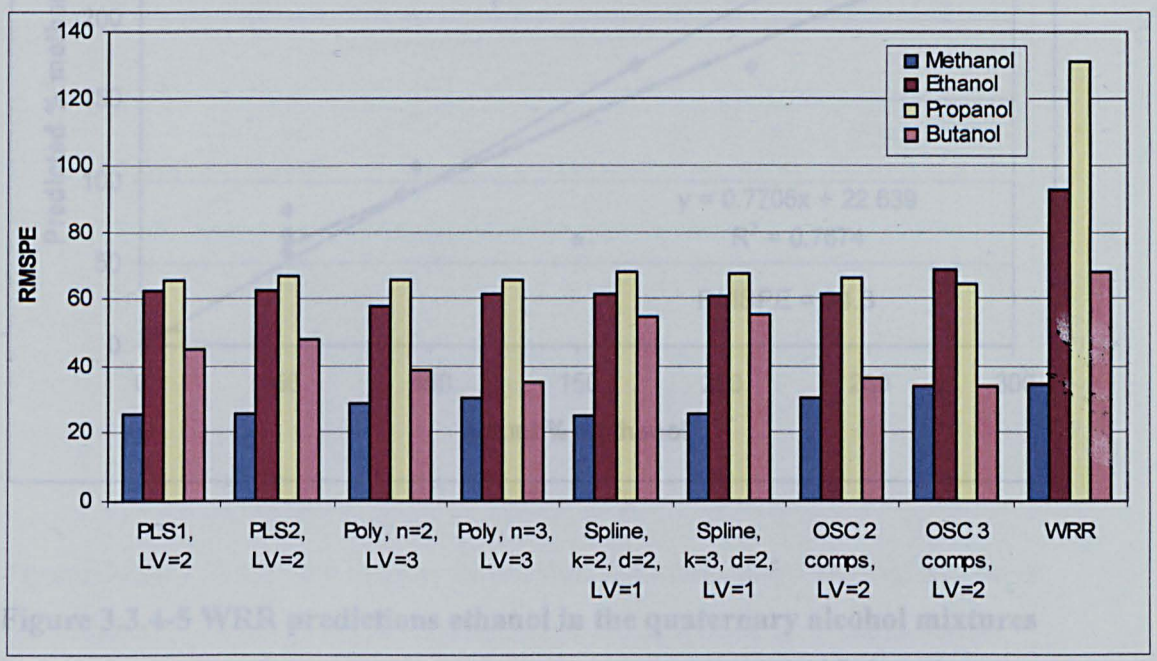
Figure 3.3.4-2 Quaternary alcohol mixtures, various PLS model RMSPE predictions for models calibrated with 1 – 6 LVs



In addition to PLS, a WRR calibration model was calculated. For each type of PLS model the number of LVs which gave the lowest average RMSPE is plotted

in figure 3.3.4-3 with the WRR results. WRR has the highest errors of prediction; there is little obvious difference between the different PLS models.

Figure 3.3.4-3 Quaternary alcohol samples RMSPE for various models



The actual Vs predicted % alcohol for each of the alcohols in the quaternary samples modelled by WRR is plotted in figures 3.3.4-4 to 8.

Figure 3.3.4-4 WRR predictions methanol in the quaternary alcohol mixtures

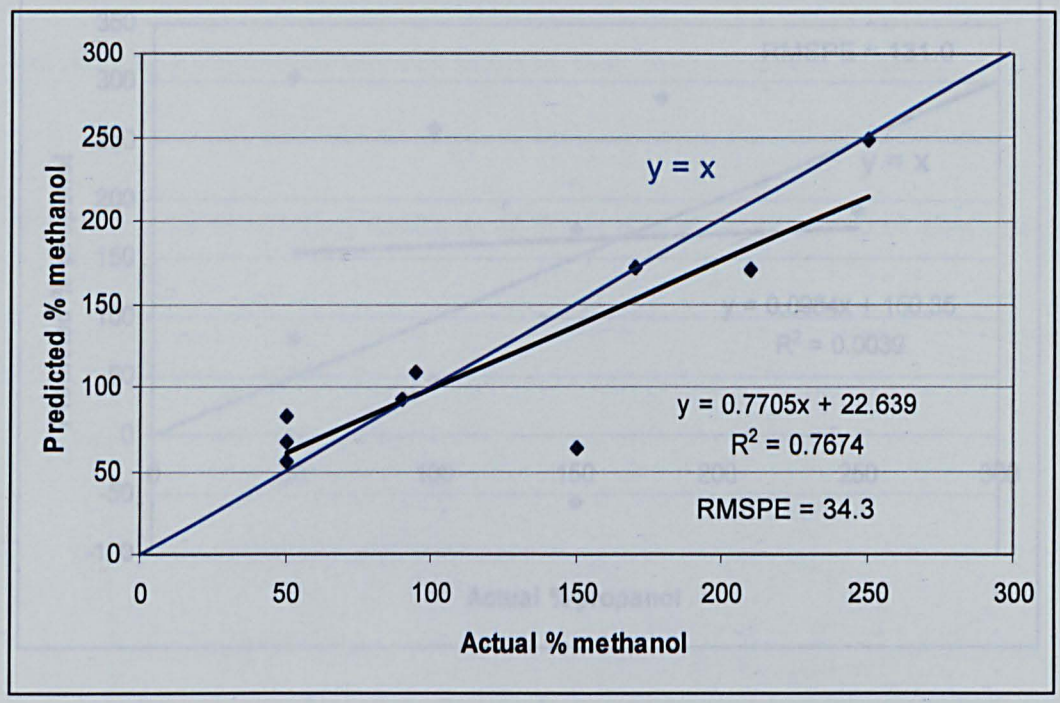


Figure 3.3.4-5 WRR predictions ethanol in the quaternary alcohol mixtures

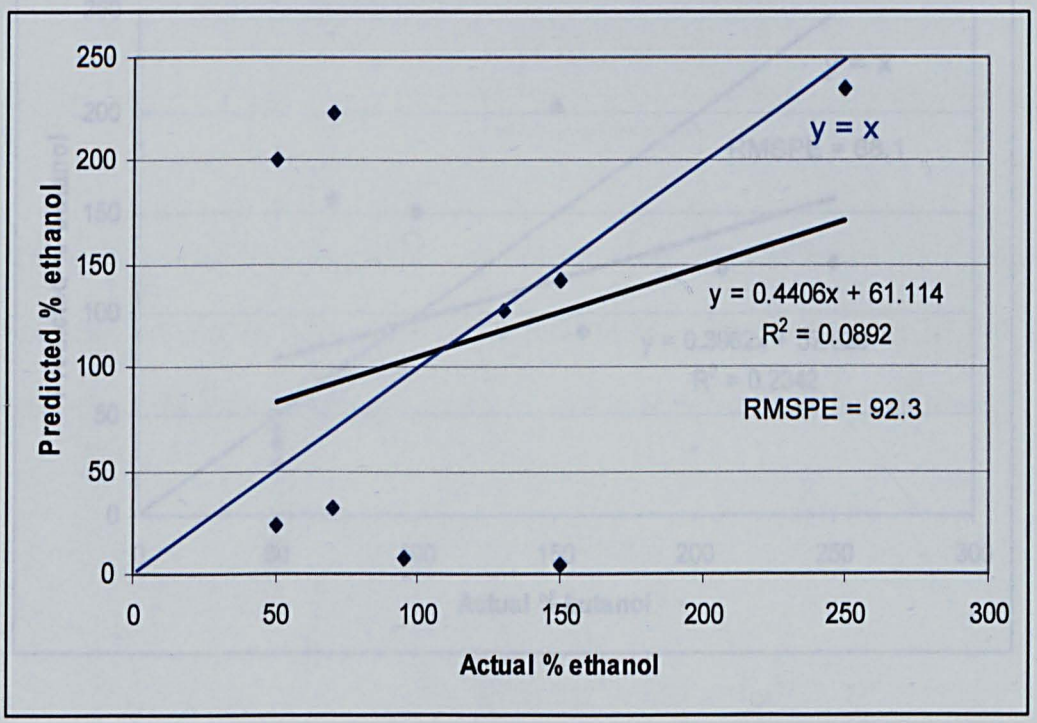


Figure 3.3.4-6 WRR predictions propanol in the quaternary alcohol mixtures

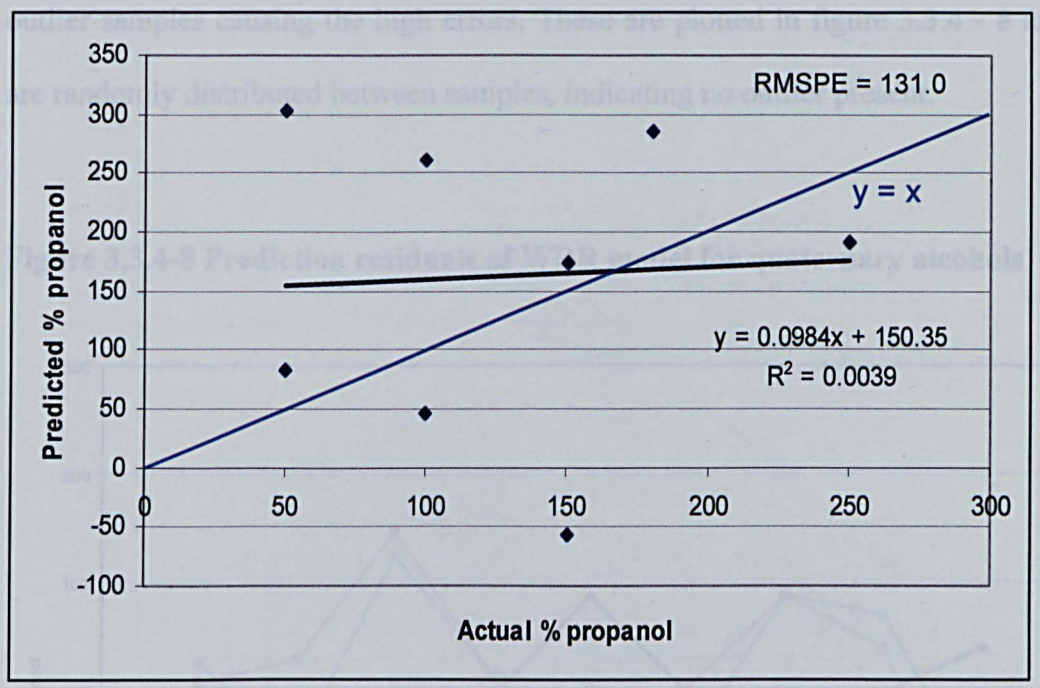
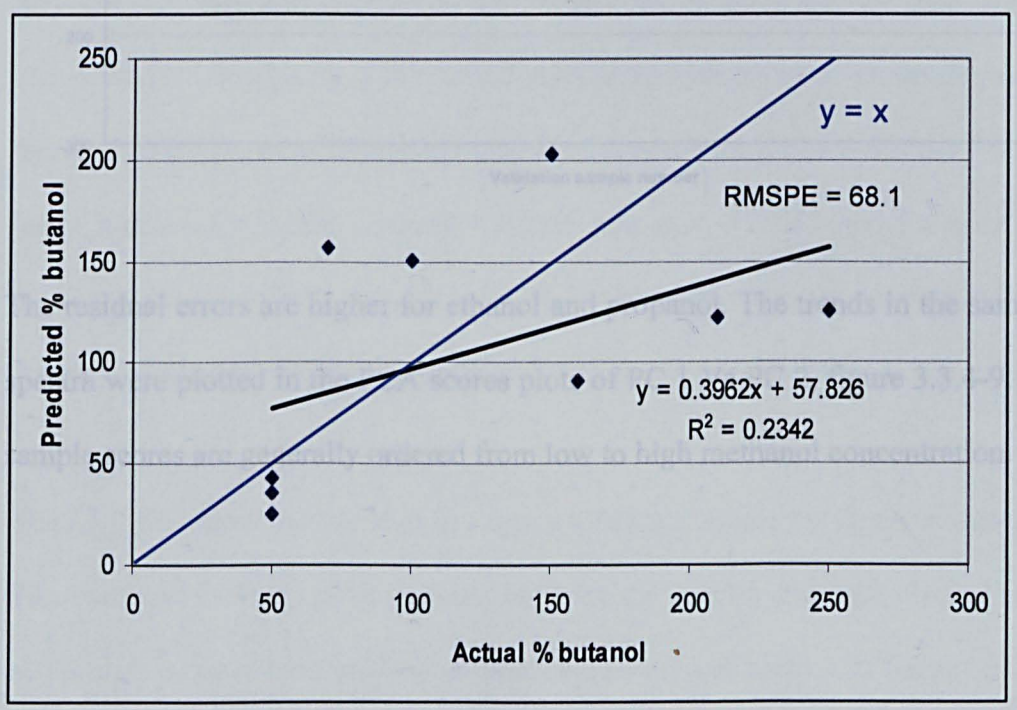
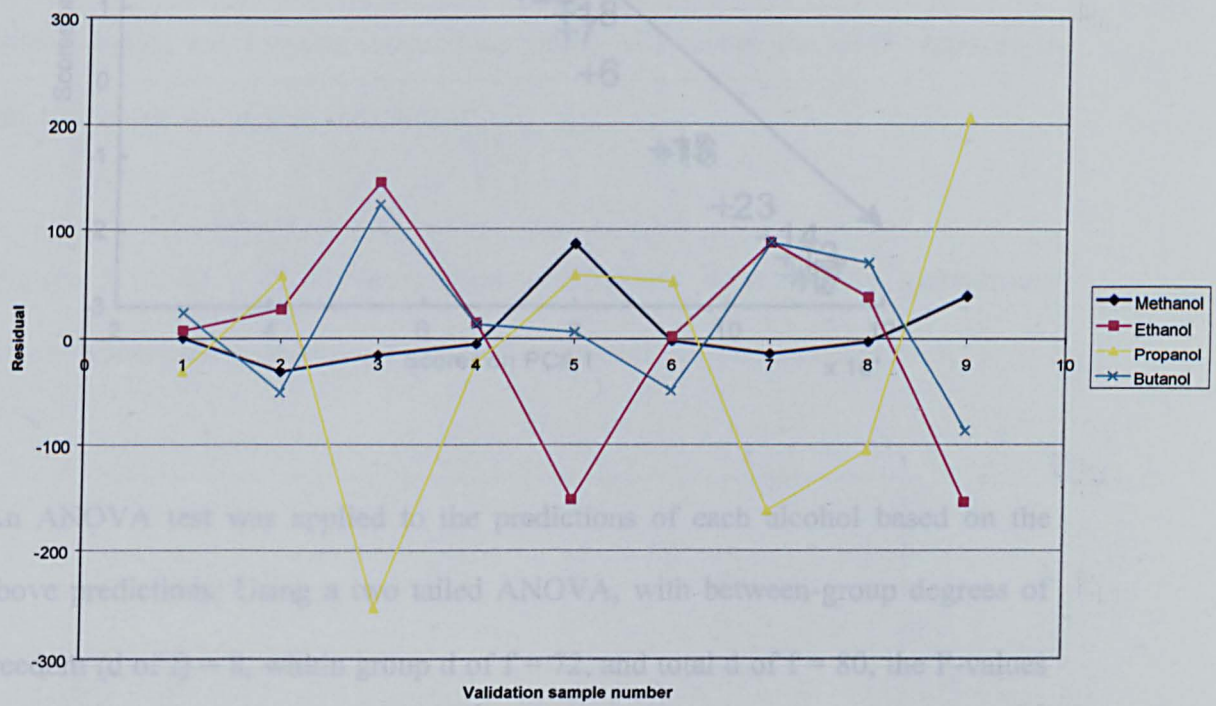


Figure 3.3.4-7 WRR predictions butanol in the quaternary alcohol mixtures



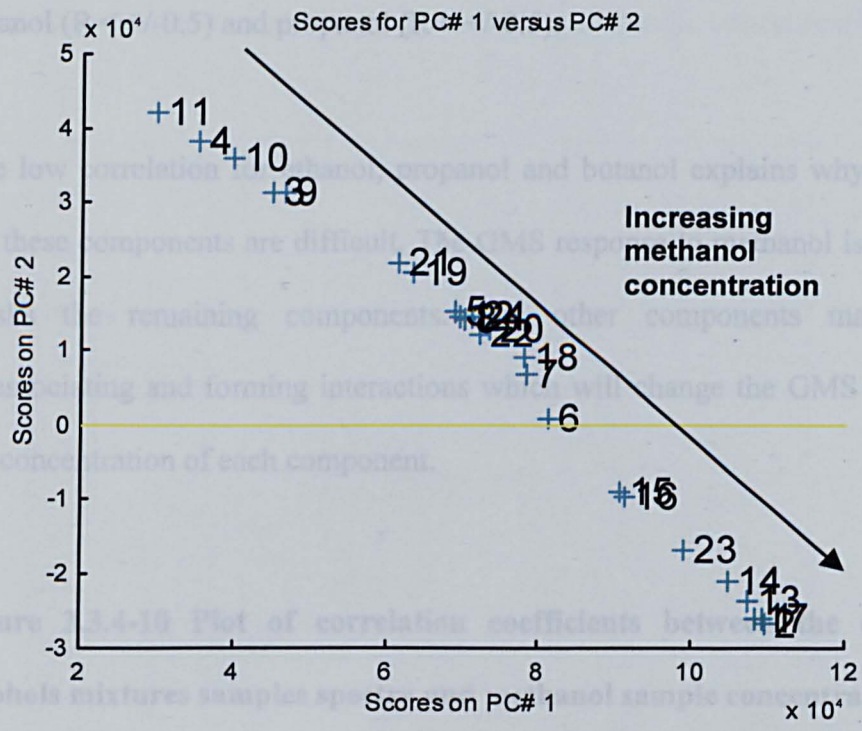
The prediction residuals are calculated for each sample to identify if there are outlier samples causing the high errors. These are plotted in figure 3.3.4 - 8 and are randomly distributed between samples, indicating no outlier present.

Figure 3.3.4-8 Prediction residuals of WRR model for quaternary alcohols



The residual errors are higher for ethanol and propanol. The trends in the samples spectra were plotted in the PCA scores plots of PC 1 Vs PC 2, figure 3.3.4-9. The sample scores are generally ordered from low to high methanol concentration.

Figure 3.3.4-9 Quaternary alcohol mixtures PCA scores plot of PC 1 V PC 2



An ANOVA test was applied to the predictions of each alcohol based on the above predictions. Using a two tailed ANOVA, with between-group degrees of freedom (d of f) = 8, within group d of f = 72, and total d of f = 80, the F-values were; methanol = 0.039, ethanol = 0.229, propanol = 0.539 and for butanol = 0.347. As the F-crit level value was 2.374, the test found there to be no significant difference between the each model's ability to predict each alcohol in the samples.

The RMSPE values are too high to suggest GMS is suitable for this measurement. To determine the level of correlation between the spectra and each alcohol in the samples, the correlation coefficient was calculated and plotted in figures 3.3.4 – 10 to 13. The results show that there is very good correlation between methanol

(R > +/-0.99) and butanol (R > +/-0.75) but considerably less correlation for ethanol (R < +/-0.5) and propanol (R < +/-0.5).

The low correlation for ethanol, propanol and butanol explains why predictions for these components are difficult. The GMS response to methanol is so strong it masks the remaining components. The other components may also be disassociating and forming interactions which will change the GMS response to the concentration of each component.

Figure 3.3.4-10 Plot of correlation coefficients between the quaternary alcohols mixtures samples spectra and methanol sample concentration

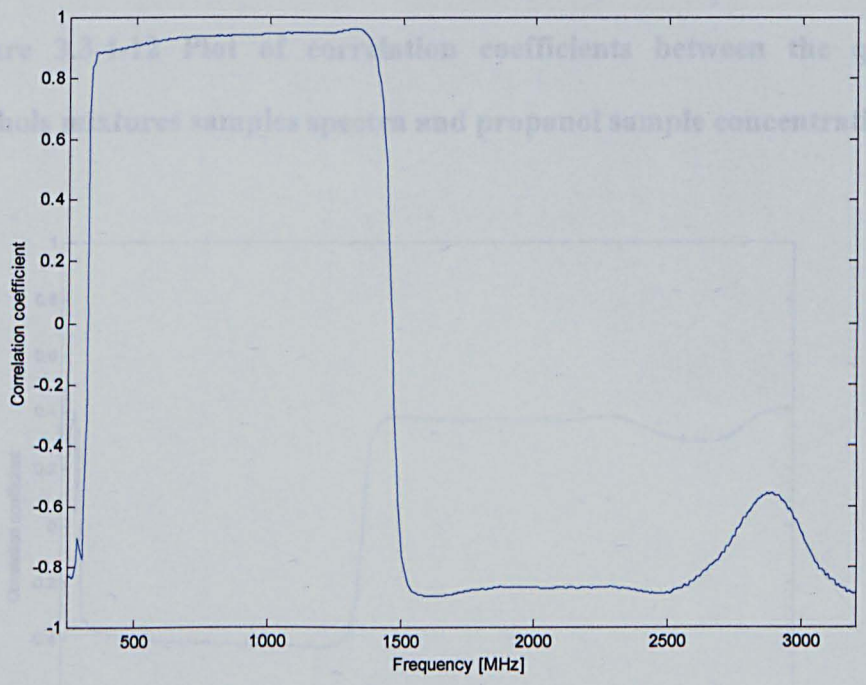


Figure 3.3.4-11 Plot of correlation coefficients between the quaternary alcohols mixtures samples spectra and ethanol sample concentration

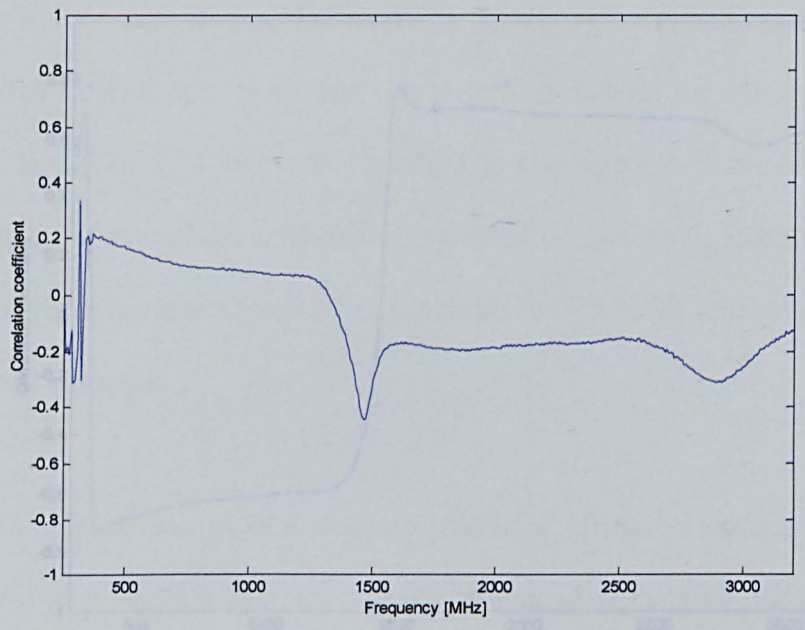


Figure 3.3.4-12 Plot of correlation coefficients between the quaternary alcohols mixtures samples spectra and propanol sample concentration

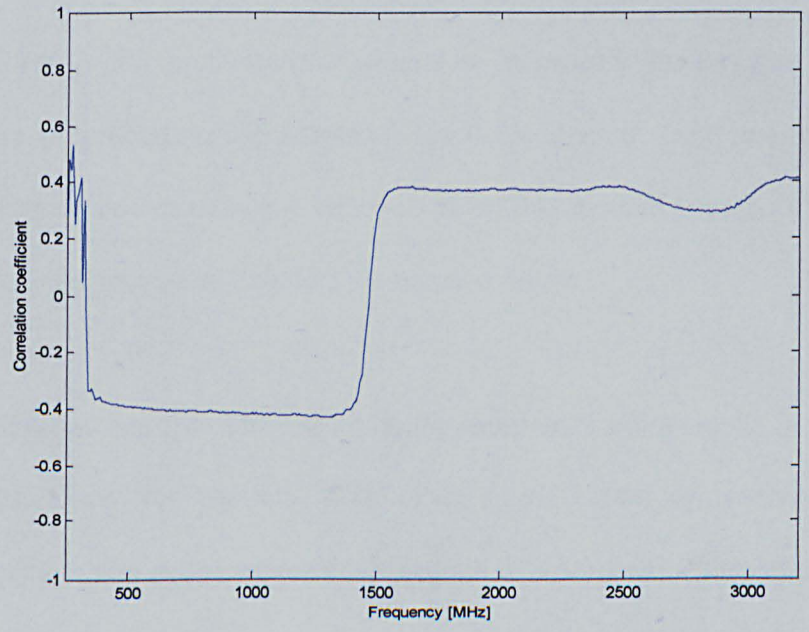
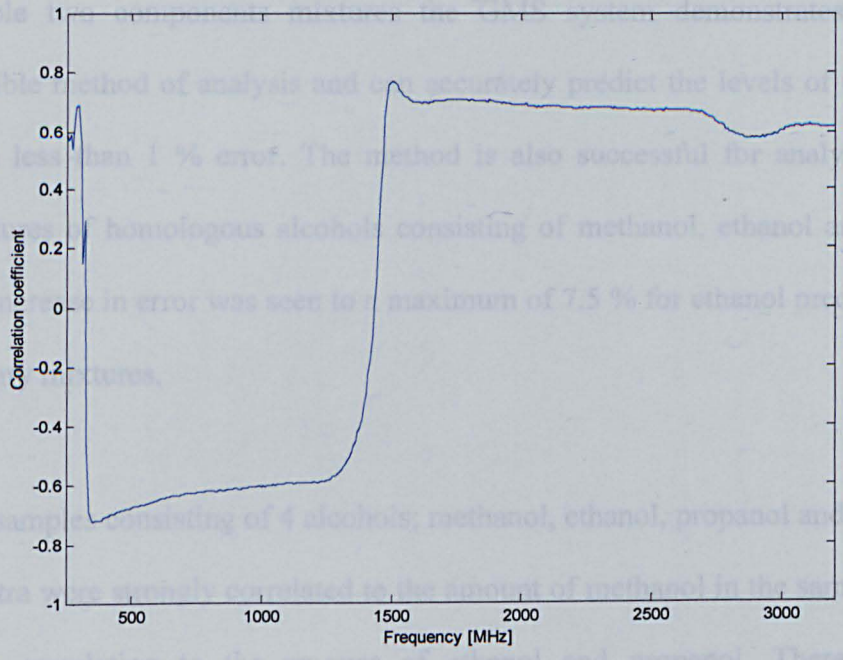


Figure 3.3.4-13 Plot of correlation coefficients between the quaternary alcohols mixtures samples spectra and butanol sample concentration



4 Conclusions

This section of work based on the analysis of alcohols by GMS has shown that for simple two components mixtures the GMS system demonstrates itself as a feasible method of analysis and can accurately predict the levels of each alcohol with less than 1 % error. The method is also successful for analysing tertiary mixtures of homologous alcohols consisting of methanol, ethanol and propanol. An increase in error was seen to a maximum of 7.5 % for ethanol prediction in the tertiary mixtures.

For samples consisting of 4 alcohols; methanol, ethanol, propanol and butanol, the spectra were strongly correlated to the amount of methanol in the samples but had little correlation to the amount of ethanol and propanol. There was some correlation for butanol. The samples scores plots for PCs 1 and 2 confirmed that the major variation in the samples spectra was due to methanol concentration. The GMS spectra of quaternary homologous alcohol samples could not be resolved to give prediction for ethanol or propanol in the samples and have higher errors of prediction for methanol (20.0 %) than in the previous datasets. The individual alcohols can not be resolved by this method because of the similarities in the pure spectra and similar dielectric constants.

Limitations begin to be seen when the number of components in the samples are raised. Once this happens it becomes much harder to resolve the spectra of samples which are composed of properties with similar dielectric constants and spectra.

The conclusion is drawn that GMS is an optimal method with smaller numbers of components in samples and components with the large differences in dielectric constant and pure spectra.

5 Further Work

The aim of this section of work was to demonstrate the feasibility of GMS for the measurement of chemical species. It has solely looked at measurement of alcohol solutions. The further work to be carried out in this area of chemical measurement is almost endless. Some suggestions of where this should begin are listed below.

- Measure other molecules with different functional groups that would rotate in the microwave field, i.e. carboxylic acids, nitriles, amines.
- Look at aromatic compounds.
- Investigate the effect of stereochemistry on the GMS response.
- Measure inorganic molecules which are easily polarised e.g. sodium chloride.
- Measure more multiple component samples to establish if the limitation is the number of components in a sample or interaction within the sample.
- Gain more information about microwave spectra and look at developing algorithms that can predict the shape of multicomponent spectra.

6 References

[1] http://www.rhinoanalytics.com/Advantages.htm Viewed: 20/1/03.

[2] Nyfors E and Vainikainen P, 'Industrial Microwave Sensors', 1989, Artech House.

[3] Daniewicz J. L, 'Improved On-line Measurement of Water Content and other Product Mixture Ratios Using Microwave Spectrum Analysis', Advances in Instrumentation and Control, 1992, 47, 617 – 632.

[4] McLennan F and Kowalski B, 'Process Analytical Chemistry', 1995, Blackie, Glasgow, UK.

[5] Kingston H, Haswell S, 'Microwave-Enhanced Chemistry, Fundamentals, Sample Preparation, and Applications', 1997, American Chemical Society.

[6] Daniewicz J. L, 'Improved On-line Measurement of Water Content and other Product Mixture Ratios Using Microwave Spectrum Analysis', Advances in Instrumentation and Control, 1992, 47, 617 – 632.

[7] Dane A. D, Rea G. J, Walmsley A. D, Haswell S. J, 'The Determination of Moisture in Tobacco By Guided Microwave Spectroscopy and Multivariate Calibration', Analytica Chimica Acta, 2001, 429, 185 – 194.

[8] Personal communication (E-mail), Walker C, Onix VG, 28/7/99.

[9] Batten G. D, Flinn P.C, Welsh L. A, Blakeney A. B, 'Leaping Ahead With Near Infrared Spectroscopy', 1995, NIR Spectroscopy Group, Royal Australian Chemical Institute, 75 – 78.

[10] Li Y, Brown C. W, Lo S. C, 'Near Infrared Spectroscopic determination of Alcohols – Solving Non-linearity with Linear and Non-linear Methods', Journal of Near Infrared Spectroscopy, 1999, 7, 55 – 62.

Chapter 5

Monitoring of the Beer Fermentation Process by Guided Microwave Spectroscopy

1 Introduction

Previous work has shown that GMS can be accurately used to predict levels of ethanol in water mixtures (see chapter 4). To further develop this type of application, GMS will be used to monitor the rate of formation of ethanol in beer fermentation processes. The fermentation wort is a complex mixture of many components. The main major constituent will be water along with sugar which will be converted to ethanol during the fermentation[1]. The polarity of water and ethanol will ensure these will cause the greatest response in the microwave field.

$$(C_6H_{10}O_5)_x \longrightarrow C_6H_{10}O_6 \longrightarrow CH_3CH_2OH + 2CO_2$$

For beer fermentation, this is a slow reaction which will allow investigation of the system over several days and generate a large number of spectra over this period. There are many properties of beer than can be measured. Official methods developed by the association of official analytical chemists are summarised in table 1-1[2].

Table 1-1 Official methods of beer analysis

Analyte	Method
Alcohol in beer (by weight)	Refractometer
Colour of beer	Spectrophotometric and photometric
Ethanol in beer	Specific gravity
Ethanol in beer	Gas chromatography
Glycerol in beer	Dichromate oxidation
Haze of beer after chilling	Visual and nephelometric
pH of beer	Potentiometric
Total acidity of beer	Indicator and potentiometric titration
Viscosity of beer	Viscometer

There are several methods for the analysis of alcoholic beverages or beer. The first standard method was to determine the alcohol content of beverages by specific gravity.

GC is the normal method for alcohol analysis as it is the most accurate method of alcohol determination in beer, it is also recommended by the Association of Official Analytical Chemists as the method for alcohol content of beer. Typically this method can give results to the nearest 0.1 % v/v.

A classic method for determining the end of the fermentation period of beer is specific gravity. The specific gravity of brew is expected to reduce from 1.036 to 1.005 units from the start to end of fermentation. The fermentation period is complete when the S.G has been 1.005 for two days. The end measurement is called the final gravity; the lower this number the more alcohol is present in the beer. The specific gravity of pure water is 1.000 and pure ethanol is 0.791[3]. The specific gravity is not a very accurate method of analysis as it is subject to fluctuations according to temperature. However the main benefits of this method are the speed of analysis and that it does not require any expensive equipment or maintenance.

Due to the complexity of the sample, dark colour, and the presence of particulates and dissolved gases, beer fermentation has previously been difficult to analyse on-line. In 1994 a group of authors published two spectroscopic methods for the determination of ethanol in beer. The first was the use of derivative Fourier-

transform infrared spectrometry[4] and this was followed by a stopped flow near-infrared method[5] which had improved sensitivity and a higher throughput rate of samples. Both methods allowed direct prediction of ethanol content without the use of multivariate calibration. The drawback is the sample preparation requirement of de-gassing of the beer prior to analysis combined with a lengthy calibration procedure that requires the use of ethanol and maltose standards and two cells with different pathlengths.

A commercial analyser, the SCABA™ 5611 beer analyser is produced by Foss. This instrument can perform at-line analysis for alcohol in the range of 0 – 7 % within 3 minutes with accuracy of 0.02 %. Alcohol is measured with a ceramic sensor by oxidation of evaporated vapours from the samples[6]. This method requires 60 ml of degassed beer sample for analysis. Another Foss beer analyser is the Rapidtec™ 5665, a NIR spectrophometer. This does not require full filtration, but bubbles have to be removed by shaking or stirring the sample. It is reported to analyse samples at-line within 1 minute; however in addition to this, the sampling time is to be considered.

The major advantage of using microwave spectroscopy as an alternative method of analysis is that it will analyse the whole sample including any dissolved gases and therefore the sample does not need to be degassed prior to analysis, removing this bottleneck for analysis. The methods can also be in-line allowing rapid analysis.

Another advantage is that the measurement cavity for a microwave spectrometer can be located completely in-line and thus is non-intrusive. Typically, insertion probes are required for NIR. These can be problematic because the pathlengths for NIR are very small which can be blocked with the biomass produced during the fermentation.

The aim of the study was to perform the fermentation in two different scenarios; 1) As a large scale batch process, 30 l and 2) a small scale, 0.5 l fermentation.

In the first part of the study, 30 l of beer was fermented in a 70 l glass batch reactor. To prevent fouling of the microwave cavity a flow cell was employed for analysis of the large scale batch fermentation. The flow cell was made of Teflon, which gives minimum absorption in the microwave region. It was fixed in place for the fermentation period as movement could result in small variations in the spectra recorded. The cell was cylindrical in shape with a sample volume of 100 ml. A re-circulation loop continuously fed the fermentation broth through a Teflon cylinder in the GMS sample chamber for analysis. Fermentation samples were taken from the loop to be analysed by GC for ethanol content.

The second scenario is much simpler. Instead of brewing on a large scale and re-circulating the fermentation broth, the beer was brewed directly inside the GMS sample chamber. This reduced the overall brew volume to 0.5 l, but has the advantage of simplicity. The sample chamber was adapted to allow temperature control throughout the fermentation.

Preliminary studies were carried out to verify the suitability of the Teflon cylinder as a flowcell for the sample chamber in the large scale batch fermentation. Two experiments were performed where ethanol in water samples and beer spiked with additional ethanol were analysed in the Teflon cylinder within the GMS sample chamber. The cell was made of Teflon as this material has the lowest response to microwave radiation for common materials used to construct flowcells[7], see table 1-2.

Table 1-2 Dielectric constants of some common materials

Material	Dielectric constant
Teflon	2.0
Glass	3.7 – 10.0
Glass silica	3.8
Quartz	4.2

2 Experimental

2.1 Feasibility of GMS monitoring using a Teflon cylinder

As a pre-study to the analysis of beer, ethanol in water samples were analysed to show that small variations in alcohol can be monitored by GMS. The samples were analysed in a Teflon cylinder which in the future could be modified to be used as a flow cell for analysis of continuously flowing systems for the large scale fermentation. The cylinder is approximately 2.8 cm wide and 10 cm tall with an internal volume of 100 ml. For analysis it was placed in the centre of the GMS sample chamber and held loosely in place by some metal plates fixed on top of the sample chamber.

2.1.1 Analysis of water spiked with ethanol

97 ml of water was placed in the flowcell and the spectrum was recorded. Then 0.3 ml of pure ethanol was pipetted directly into the water, stirred and the spectrum recorded. This was repeated until 10 additions of ethanol in 0.3 ml aliquots had been made.

2.1.2 Analysis of spiked beer samples

97 ml of beer sample was spiked with 10 aliquots of 0.3 ml ethanol (Fisher Scientific) and the GMS spectrum recorded after each addition.

2.2 GC analysis for ethanol in beer samples

The alcohol content of the beer fermentation samples was accurately determined by gas chromatography (GC). An internal standard (propanol, Fisher Scientific) was used in the calibration for ethanol content.

2.2.1 Preparation of solutions

Internal standard stock solution

25 % propan-1-ol in water (100 ml)

Calibration standards

2 % ethanol: 0.5 ml ethanol to 10 ml stock solution, diluted to 25 ml.

4 % ethanol: 1.0 ml ethanol to 10 ml stock solution, diluted to 25 ml.

6 % ethanol: 1.5 ml ethanol to 10 ml stock solution, diluted to 25 ml.

8 % ethanol: 2.0 ml ethanol to 10 ml stock solution, diluted to 25 ml.

Beer Sample

15 ml of filter beer sample was added to 10 ml of stock solution.

Each solution was analysed in duplicate with an injection volume of 1 μl.

2.2.2 GC instrumental method

Gas Chromatograph:	Perkin Elmer Autosystem XL
Column:	PorOpak Q (Packed, 2m)
Detector type:	Flame Ionisation (FID)
Detector temperature:	250 °C
Injector temperature:	250 °C
Oven temperature:	190 °C
Head Pressure:	30 Psi
Carrier Gas:	Nitrogen
Run time:	5 minutes

From the duplicate analysis the average peak area ratio for the ethanol and propanol peaks was calculated. The % ethanol of each standard was plotted against the peak area ratio. The % ethanol of the unknown beer sample was determined by interpolation from the graph.

2.3 On-line analysis of 30 l batch fermentation

30 l of beer was fermented in a 70 l batch reactor for 3 weeks. The laboratory set-up of the batch reactor can be seen in figure 2.3-1, the sample chamber during analysis in figure 2.3-2 and the Teflon cylinder used in figure 2.3-3.

Using two homebrew beer kits, Geordies Bitter™ (Viking Brews, A division of Wander Ltd, Kings Langley, Herts, WD4 8LJ), 30 l of beer wort was added with yeast and sugar to the reactor.

Figure 2.3-1 Laboratory set up of the batch reactor and recirculation system used to monitor the fermentation of beer (30 l) by GMS

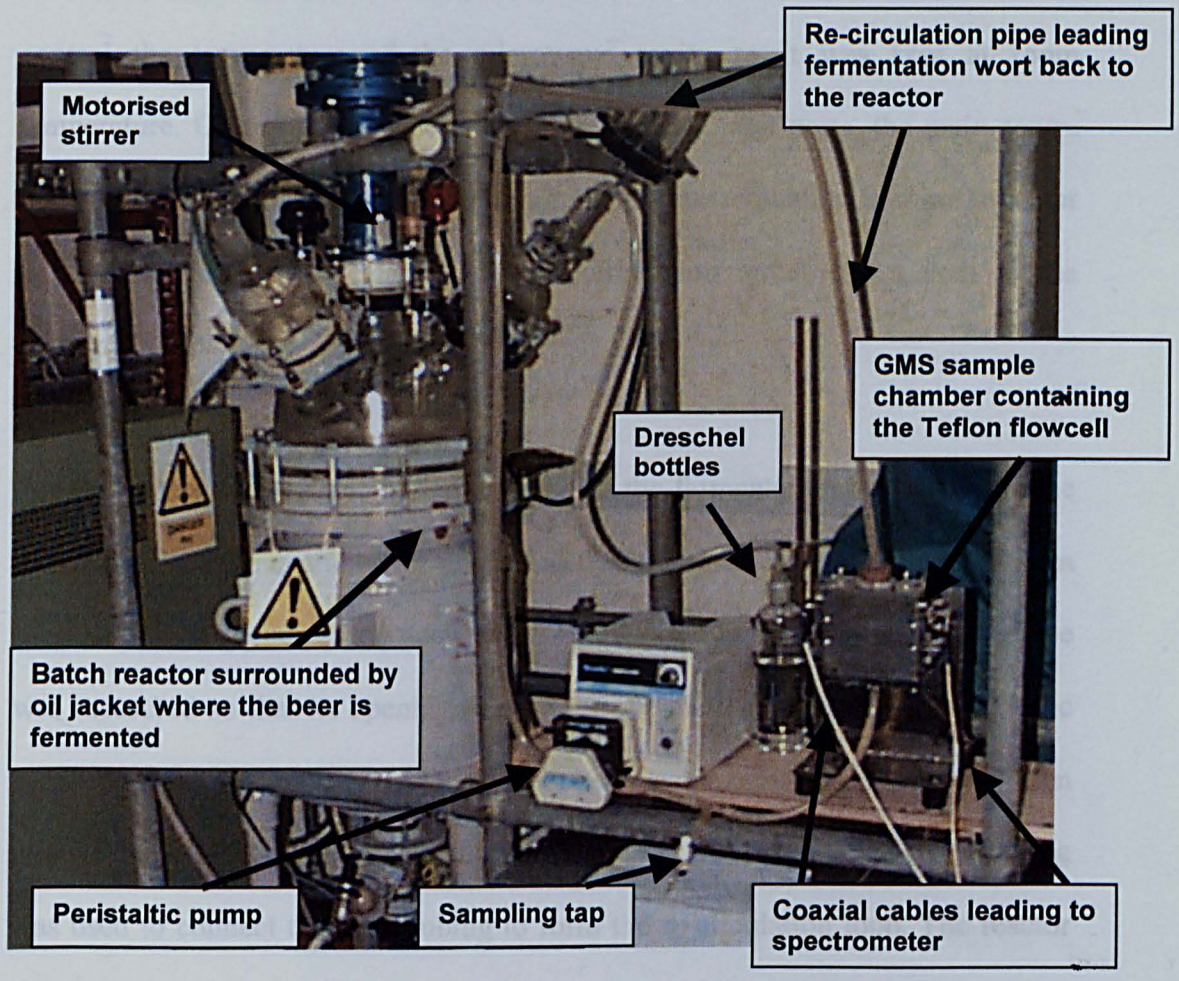


Figure 2.3-2 GMS sample chamber during fermentation

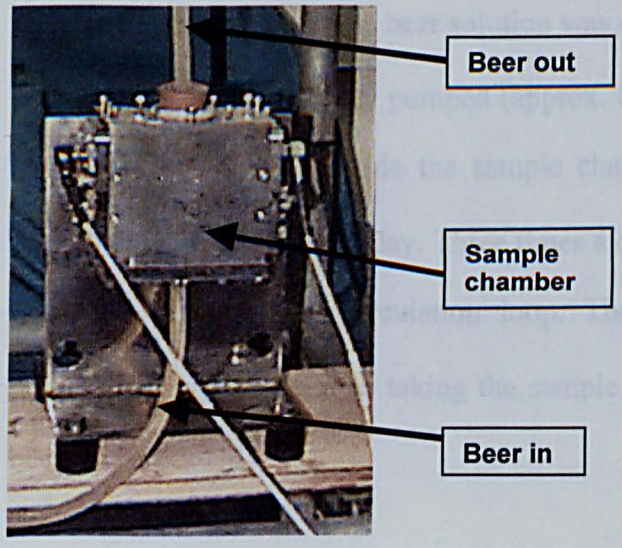


Figure 2.3-3 Teflon cylinder as flowcell

The oil jacket surrounding the reactor was set to 24 °C. The temperature of the beer was measured by a thermocouple inside the reactor. It was not possible to control the temperature of the microwave cavity so it was kept at room temperature. On top of the reactor there are 5 inlet/outlet ports, the main centre one contains the stirrer paddle with motor above. The remaining four are used for the thermocouple, re-circulation inlet, re-circulation outlet and a vent to the dreschel bottles.

For the experiments, a re-circulation loop fed the fermentation brew through the microwave sample chamber for on-line analysis. A glass tube was made (80 cm long and 1 cm diameter) to attach to the fittings of a reactor inlet port. The tube was measured so that its opening was just above the stirrer paddle. A peristaltic pump was used to pull beer through from the reactor into the flowcell and then pump it back into the reactor via an inlet port at the front. Peristaltic pump tubing was used to connect the glass tubing to form the re-circulation loop. The reactor was vented through 2 Dreschel bottles (1 l).

Between 9 am and 5 pm the beer solution was continuously stirred by a motorised stirrer. The beer was slowly pumped (approx. 0.5 l/min) from the reactor through to the Teflon cylinder inside the sample chamber. Sample spectra were taken periodically throughout the day. Three times a day, a 50 ml beer sample was taken from the tap in the re-circulation loop. The % alcohol of the sample was determined by GC. Prior to taking the sample the spectrum of the re-circulating beer was recorded.

2.4 In-line analysis of 0.5 l batch fermentation

A simpler way to monitor the fermentation of beer is on a small scale by containing the fermentation wort within the microwave sample chamber. The experimental setup can be seen in figure 2.4-1. During this experiment the temperature of the sample chamber was controlled.

A different homebrew kit to the first experiment was used ('Extra strong bitter', Unican Foods Ltd, P.O. Box 171, Reepham, Norwich, NR10 4BJ). The fermentation wort was 110g/l instead of the recommended 83g/l, this was to produce a high alcohol content to maximise variation in the spectra (the sugar and yeast concentrations were scaled accordingly).

The base of the sample chamber was sealed. Using the homebrew beer kit a beer wort solution was made up with 60.3 g of wort syrup and 33.1 g of sugar. 550 ml of wort solution was placed in the sample chamber. 0.2 g of yeast was sprinkled onto the solution. A ventilation pipe was added to the lid of the sample chamber leading to two 250 ml Dreschel bottles in series. A radiator was wrapped around the sample chamber; see figure 2.4-2, through this water was continuously pumped whilst controlled at 22 °C using a water bath, Grant 3G LTD (Fisher Scientific) waterbath and recirculator, precision of +/- 0.1 °C.

The beer wort was fermented for two weeks, the microwave spectra were recorded during this period every 30 minutes between the hours of 9 am and 5 pm. At the end of this period the final strength of the beer was determined by GC.

Figure 2.4-1 Laboratory set up for the in-line fermentation of beer

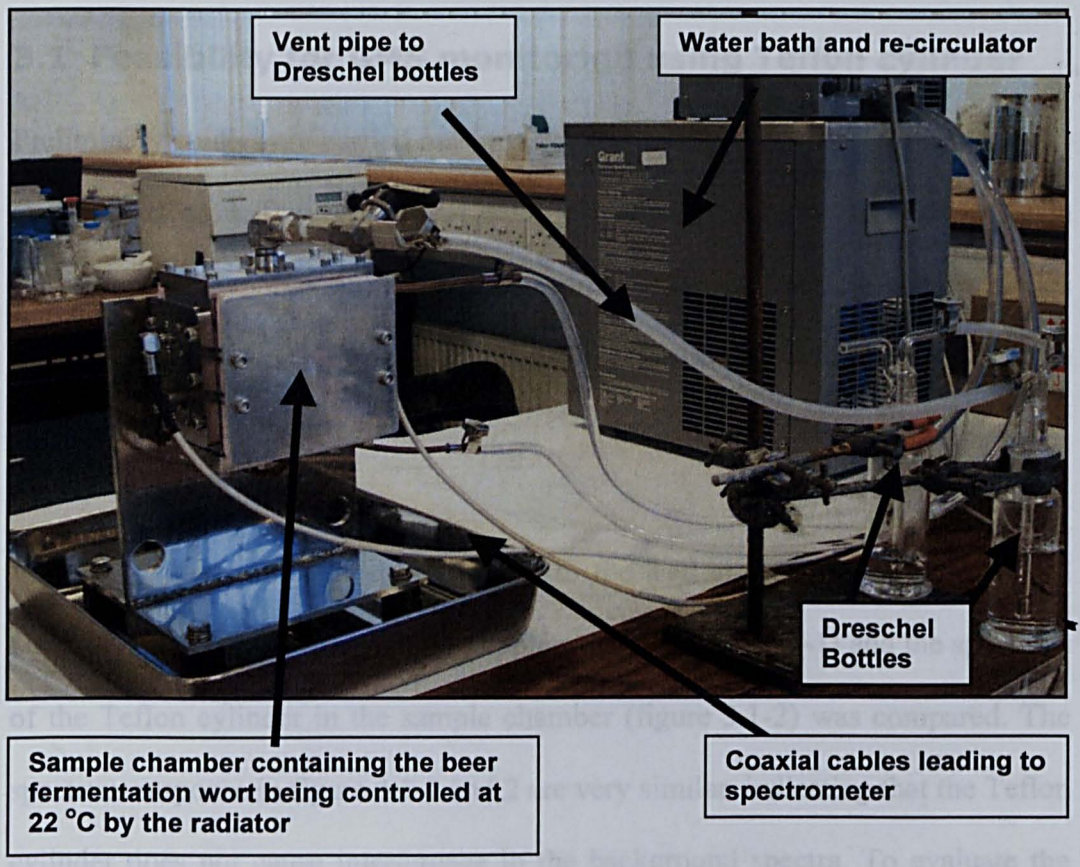
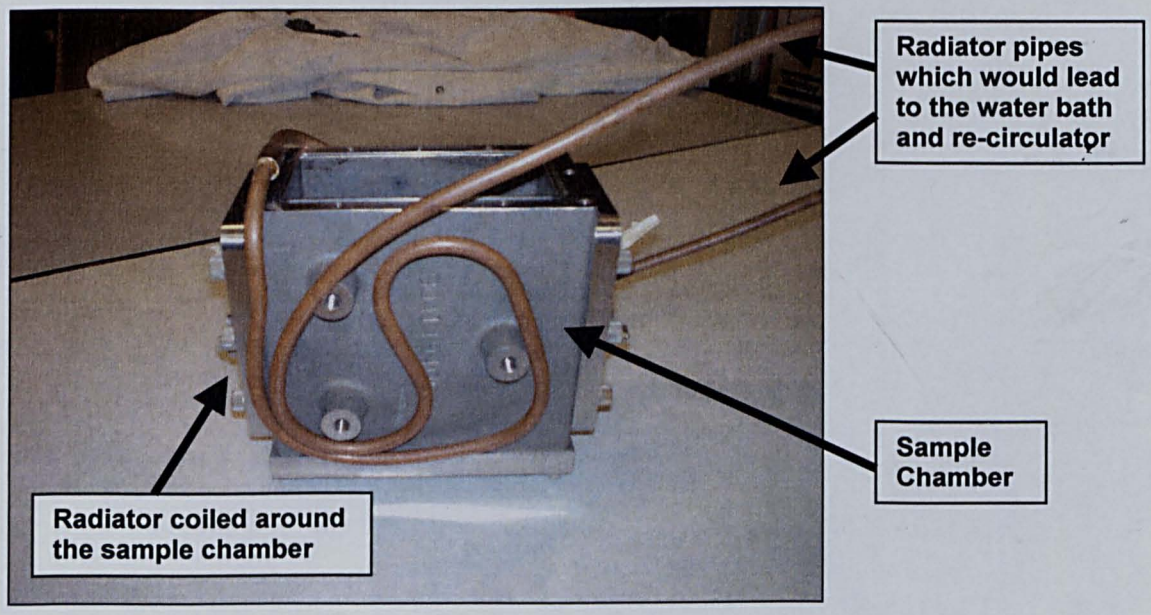


Figure 2.4-2 GMS sample chamber with radiator attachment



3 Results

3.1 Feasibility for GMS monitoring using Teflon cylinder

Preliminary studies are carried out to verify the suitability for the Teflon cylinder to be used as a flowcell in the sample chamber for the large scale batch fermentation. Two experiments were performed where ethanol in water samples and beer spiked with additional ethanol are analysed in the Teflon cylinder within the GMS sample chamber.

In order to determine the response due to the Teflon cylinder the background GMS spectrum of air in the empty sample chamber (figure3.1-1) and the spectrum of the Teflon cylinder in the sample chamber (figure 3.1-2) was compared. The spectrum response in figure 3.1-1 and 2 are very similar, indicating that the Teflon cylinder does not cause interference in the background spectra. To evaluate the real difference in the background spectra with and without the Teflon cylinder within the GMS cavity, the residuals between the two spectrums were plotted.

Figure 3.1-1 GMS background spectrum of air

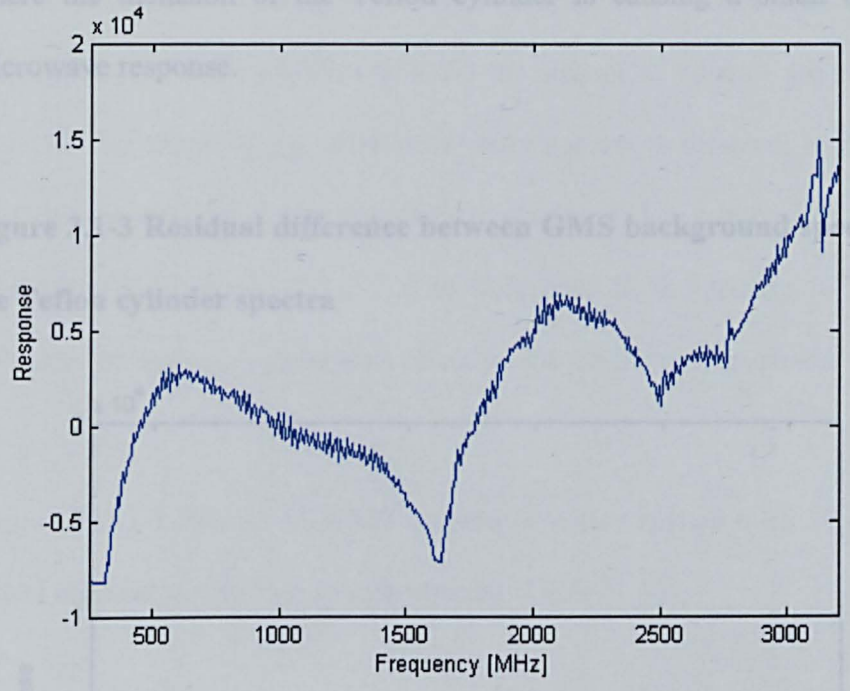
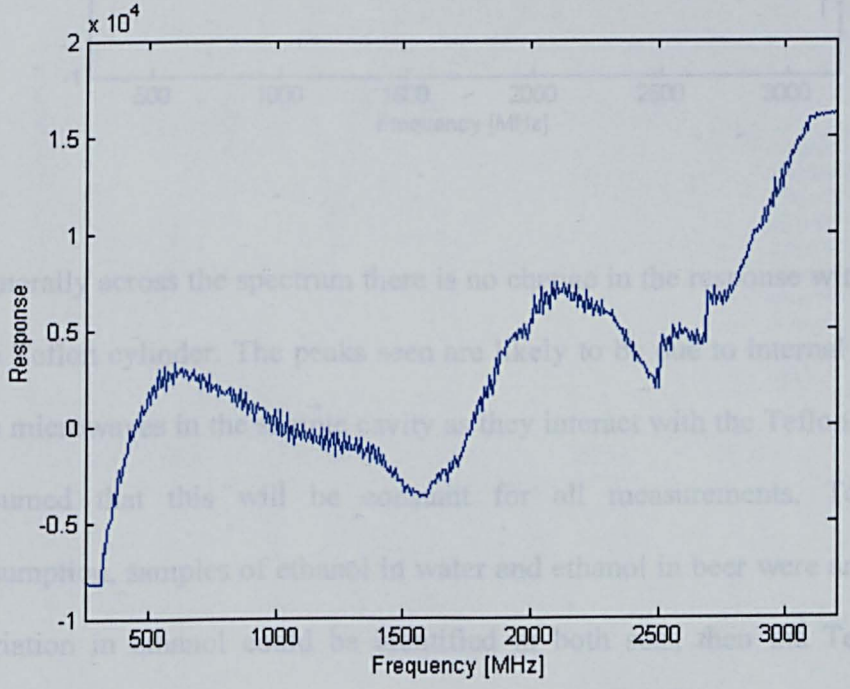
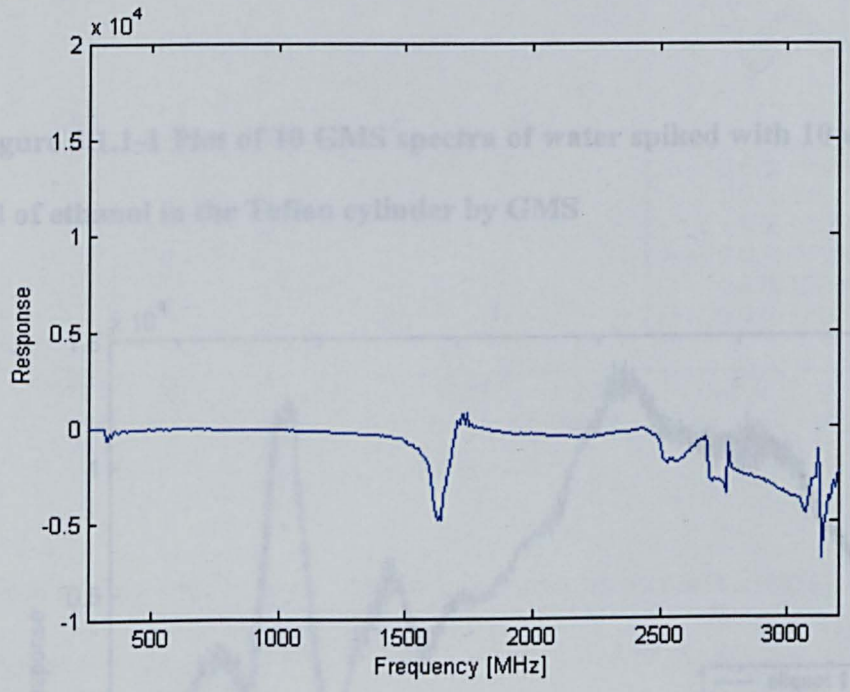


Figure 3.1-2 GMS spectrum of the Teflon cylinder in air



The residual differences between the two spectra, figure 3.1-3, show the regions where the inclusion of the Teflon cylinder is causing a small change in the microwave response.

Figure 3.1-3 Residual difference between GMS background spectra and with the Teflon cylinder spectra



Generally across the spectrum there is no change in the response with and without the Teflon cylinder. The peaks seen are likely to be due to internal reflections of the microwaves in the sample cavity as they interact with the Teflon cylinder. It is assumed that this will be constant for all measurements. To verify this assumption, samples of ethanol in water and ethanol in beer were analysed. If the variation in ethanol could be identified in both sets, then the Teflon cylinder would be used for measurements of the beer fermentation in the batch reactor.

3.1.1 Analysis of spiked water in the Teflon cylinder by GMS

Water held in the Teflon cylinder was spiked with ethanol. The 10 spectra measured after each addition of a 0.3 ml aliquot of ethanol are shown in figure 3.1.1-1 The shape of the ethanol in water spectra is different to those measured without the cylinder during the work in chapter 4. For comparison the spectra are plotted again in figure 3.1.1-2. The difference in the spectra in figures 3.1.1-1 indicate the Teflon cylinder was effecting the GMS measurements.

Figure 3.1.1-1 Plot of 10 GMS spectra of water spiked with 10 aliquots of 0.3 ml of ethanol in the Teflon cylinder by GMS

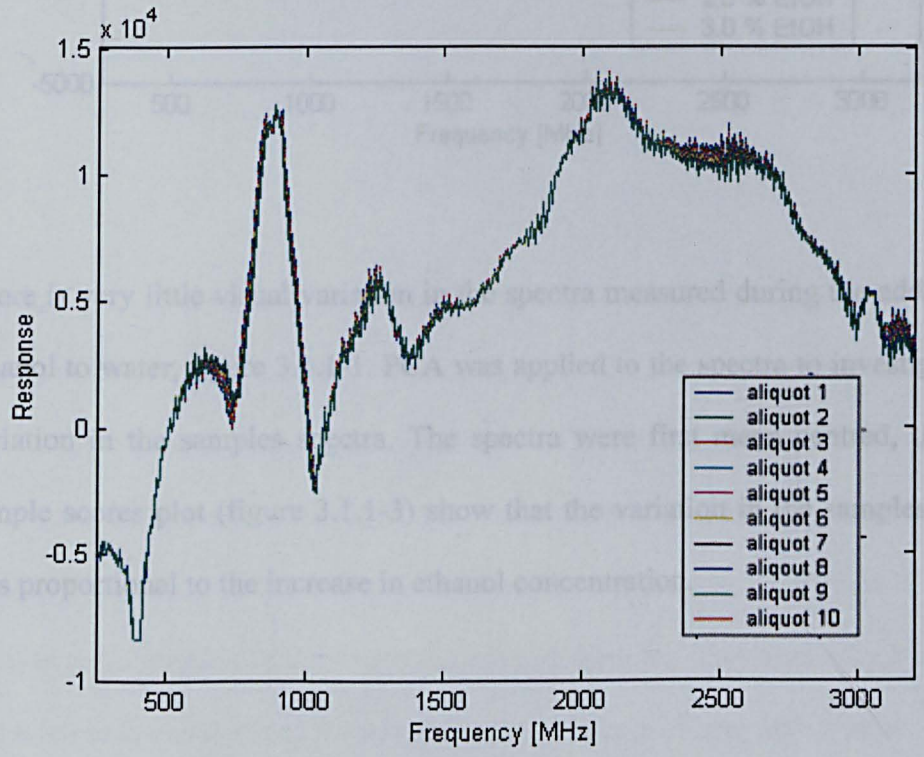
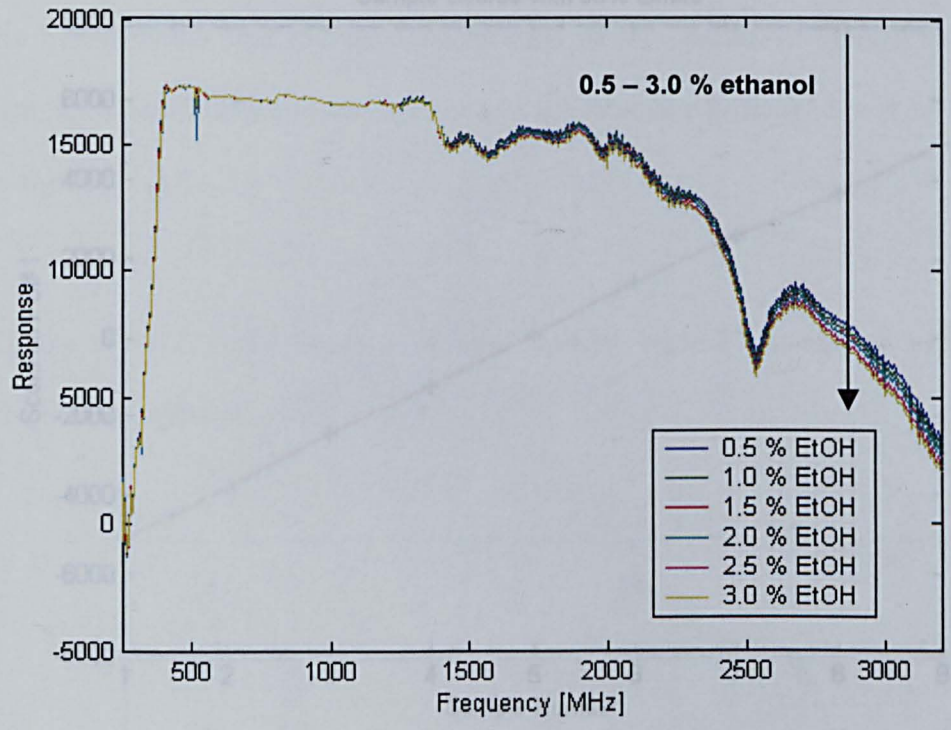
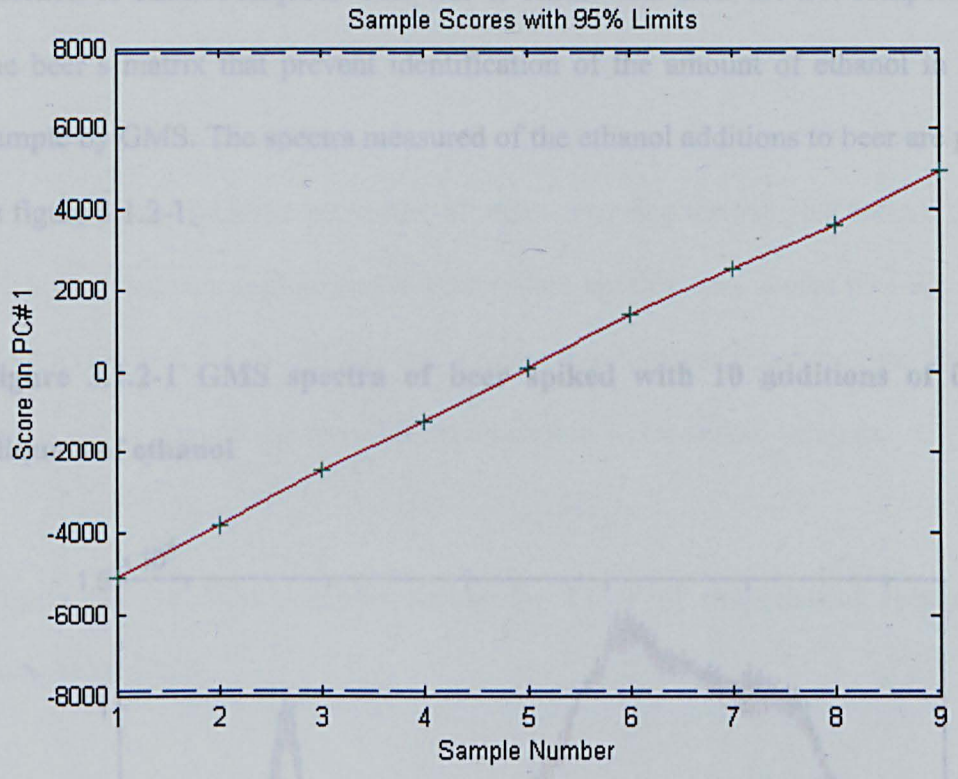


Figure 3.1.1-2 Plot of 6 ethanol in water (0.5 – 3.0 %) samples measured by GMS



There is very little visual variation in the spectra measured during the addition of ethanol to water, figure 3.1.1-1. PCA was applied to the spectra to investigate the variation in the samples spectra. The spectra were first mean centred, the PC1 sample scores plot (figure 3.1.1-3) show that the variation in the samples scores was proportional to the increase in ethanol concentration.

Figure 3.1.1-3 PCA sample scores for PC1 of ethanol additions to water in the Teflon cylinder by GMS

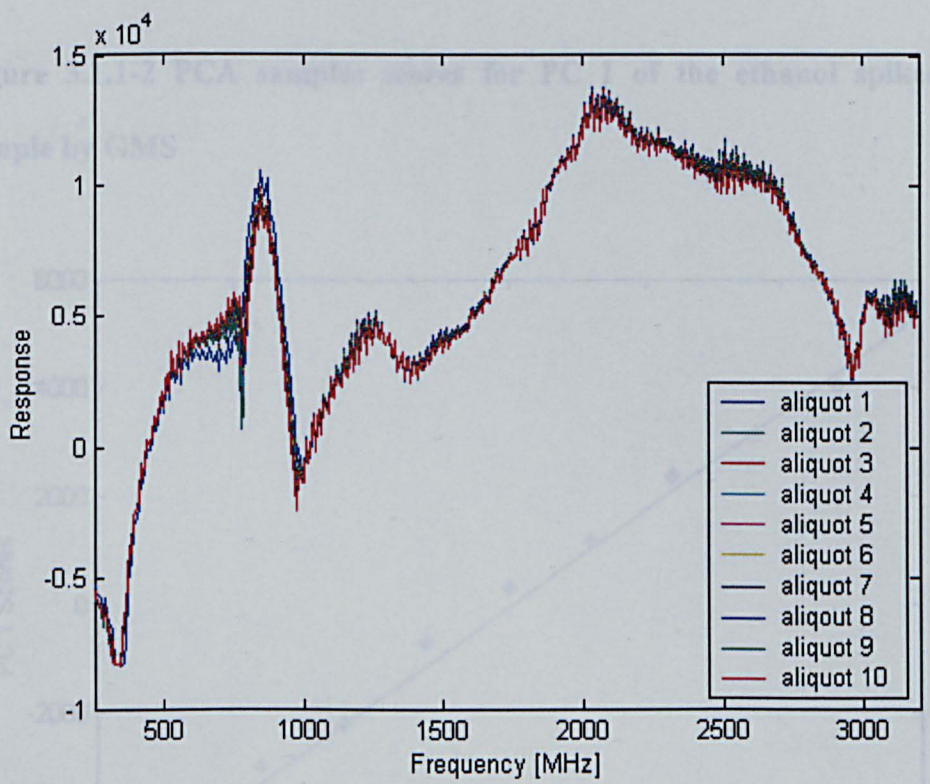


The samples spectra are very different to those seen in Chapter 4, where low concentration ethanol samples were analysed in the sample chamber without the use of the Teflon cylinder (see figure 3.1.1-2). The Teflon cylinder reduces the severity of the cut-off in the GMS spectrum and, artefacts are also apparent in the region of 750 – 1250 MHz. These artefacts do not mask the variation in sample analyte concentration in the water and ethanol samples. The feasibility study was extended to the analysis of the more complex samples of beer and ethanol.

3.1.2 Analysis of spiked beer in Teflon cylinder by GMS

The previous experiment was repeated but involved the analysis of beer with the addition of ethanol aliquots. This was to ensure that there are not components in the beer's matrix that prevent identification of the amount of ethanol in a beer sample by GMS. The spectra measured of the ethanol additions to beer are plotted in figure 3.1.2-1.

Figure 3.1.2-1 GMS spectra of beer spiked with 10 additions of 0.3 ml aliquots of ethanol

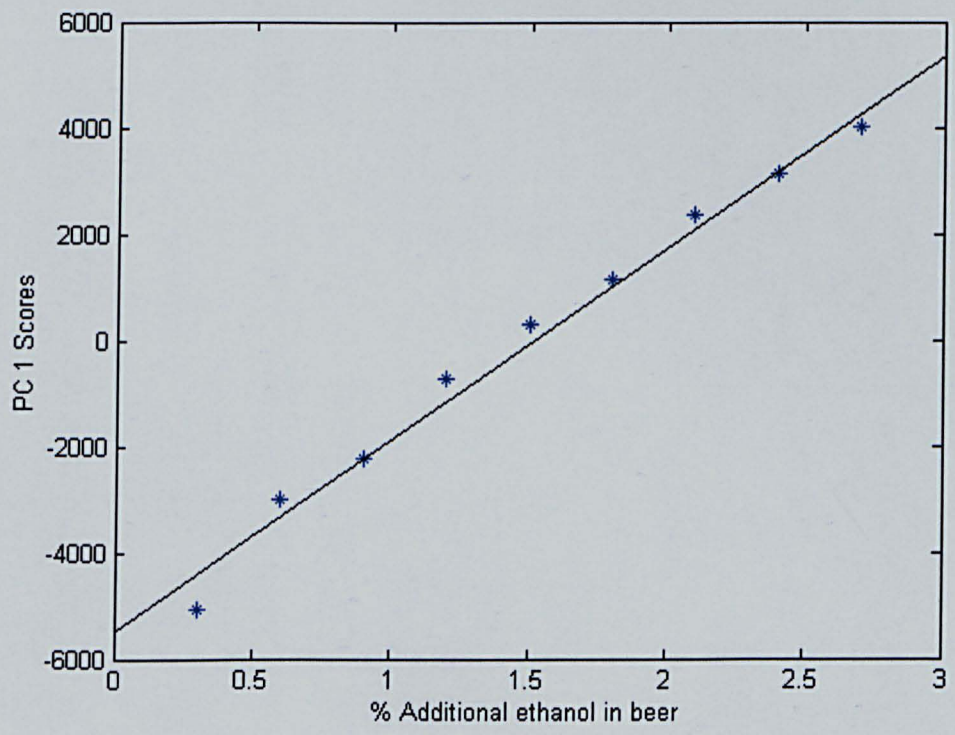


The spectra of ethanol additions to beer (figure 3.1.2-1) are a similar shape to those of water (figure 3.1.1-1), but have more noise. The magnitude of the response in the region of 750 – 1250 MHz has reduced. The GMS response to

beer would be expected to be lower than water, as it will contain components which have a lower dielectric constant (e.g. carbohydrates, sugar, and yeast) and resulting in a reduced spectral response.

The ethanol spiked beer spectra have more visual variation than the ethanol spiked water spectra. The same procedure of mean centering the samples spectra prior to PCA analysis was applied to the beer spiked spectra. The scores plot for PC1 is shown (figure 3.1.1-2) and displays a linear trend between the samples. The trend line is not as straight as seen with the ethanol in water-spiked samples.

Figure 3.1.1-2 PCA samples scores for PC 1 of the ethanol spiked beer sample by GMS



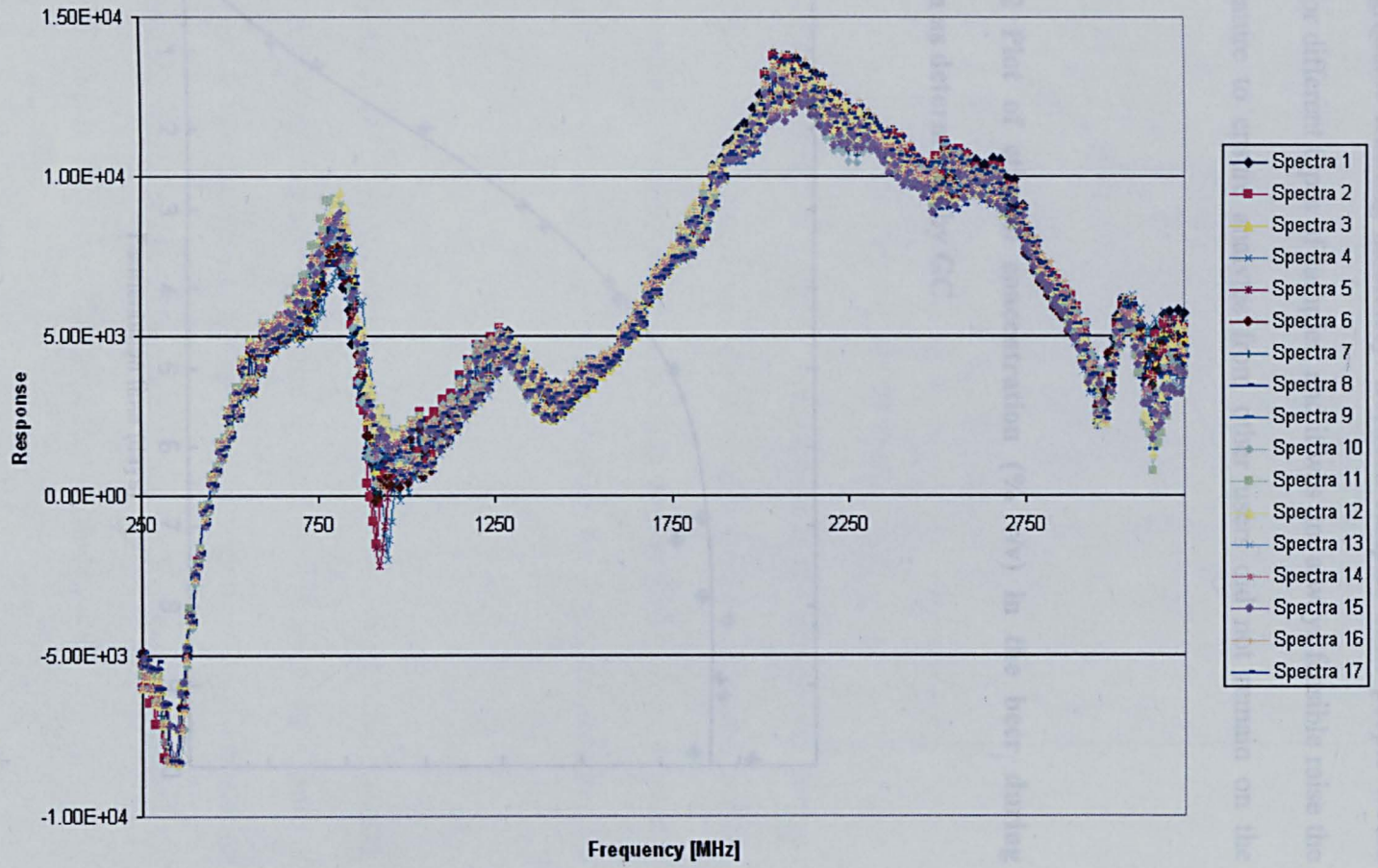
These two experiments have shown that changing the sample analysis set-up to use a Teflon cylinder as a flowcell produces artefacts in the samples spectra. By analysing samples in this cylinder the microwaves are being passed from the source through air to Teflon, to sample, to Teflon and back through air to the detector. Reflections will occur at each of these interfaces. These reflections will ultimately change the amount of microwave energy received by the detector and hence the response from the GMS instrument for a given sample.

However, taking this into consideration and that the desired trend in ethanol concentration could be clearly seen in the PCA samples scores plots, the Teflon cylinder is considered suitable for use as a flowcell for monitoring of the beer fermentation process.

3.2 Analysis of beer fermentation by GMS: 30 l batch

In this experiment 30 l of beer was brewed inside a batch reactor. The fermentation wort was circulated throughout and the GMS spectra recorded. Samples of the wort were collected and the % ethanol determined by GC. The fermentation spectra are plotted in figure 3.2-1. These are fairly noisey but are a good likeness to those of the spiked beer samples as recorded during the pre-study investigations to this experiment.

Figure 3.2-1 GMS Beer fermentation spectra recorded over 10 days



The rate of formation of ethanol, as determined by GC, is plotted in figure 3.2-2. This shows ethanol to be formed linearly during the first 5 days of fermentation (figure 3.2-3). This then plateaus and the ethanol content varies with quite a large error between samples. This could be a result of oxidisation of the ethanol or and error in the GC analysis. The GC instrument used for these experiments belonged to the undergraduate teaching laboratory. Between samples other people would use the GC for different types of samples, and it was not always feasible raise the oven temperature to ensure analytes from other users did not remain on the column.

Figure 3.2-2 Plot of ethanol concentration (% v/v) in the beer during fermentation as determined by GC

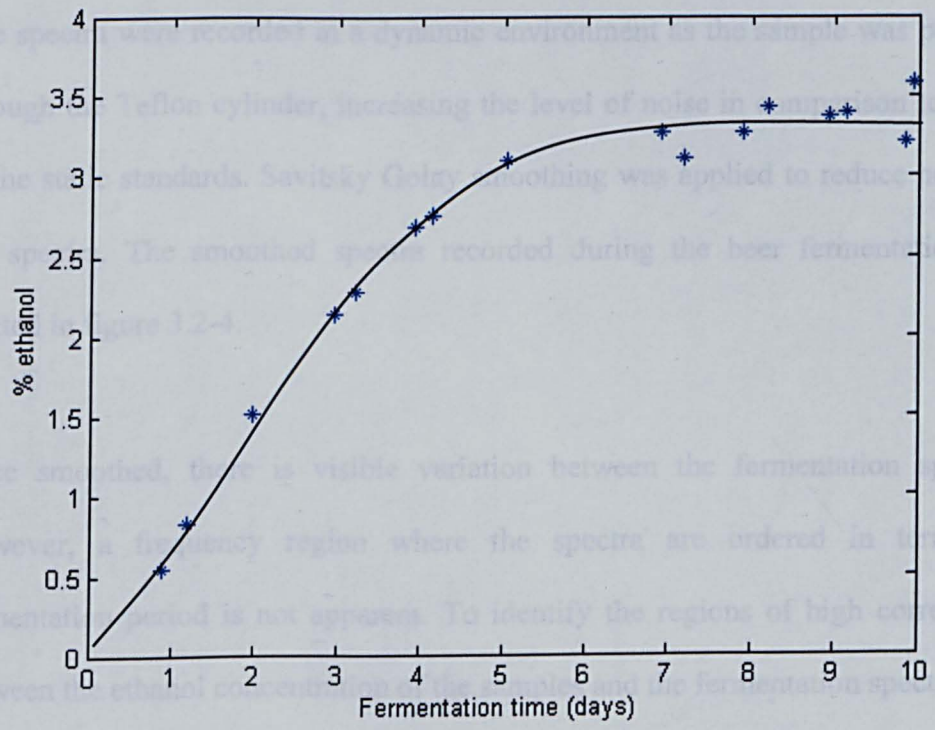
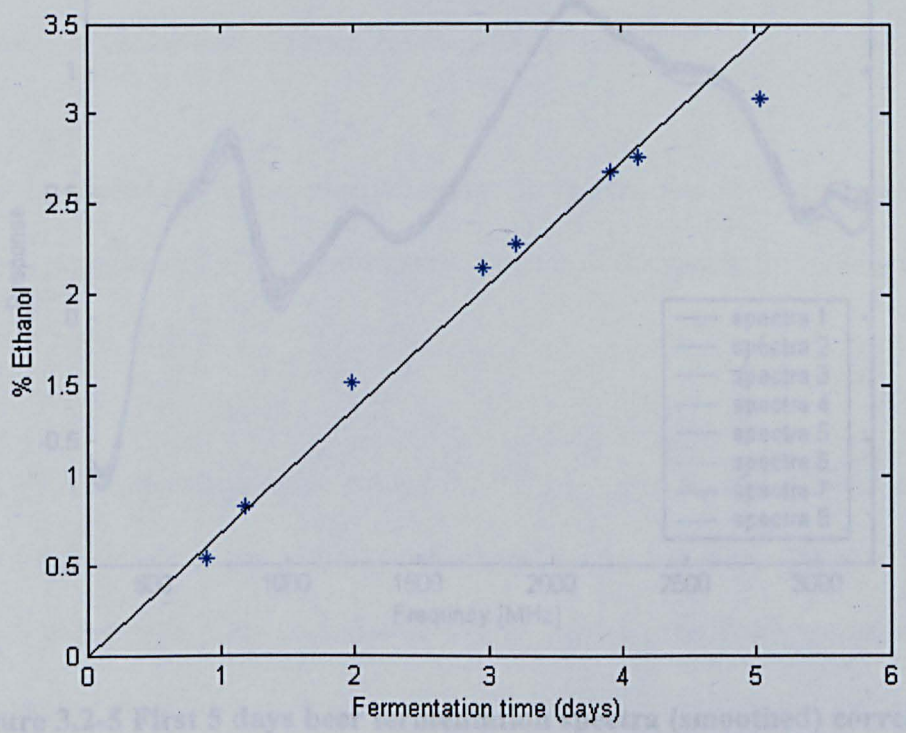


Figure 3.2-3 Plot of ethanol concentration (% v/v) in the beer during the first 5 days of fermentation as determined by GC



The spectra were recorded in a dynamic environment as the sample was pumped through the Teflon cylinder, increasing the level of noise in comparison to those of the static standards. Savitsky Golay smoothing was applied to reduce noise in the spectra. The smoothed spectra recorded during the beer fermentation are plotted in figure 3.2-4.

Once smoothed, there is visible variation between the fermentation spectra. However, a frequency region where the spectra are ordered in terms of fermentation period is not apparent. To identify the regions of high correlation between the ethanol concentration of the samples and the fermentation spectra, the correlation coefficient is calculated. This is plotted in figure 3.2-5.

Figure 3.2-4 First 5 days beer fermentation spectra Savitsky Golay smoothed

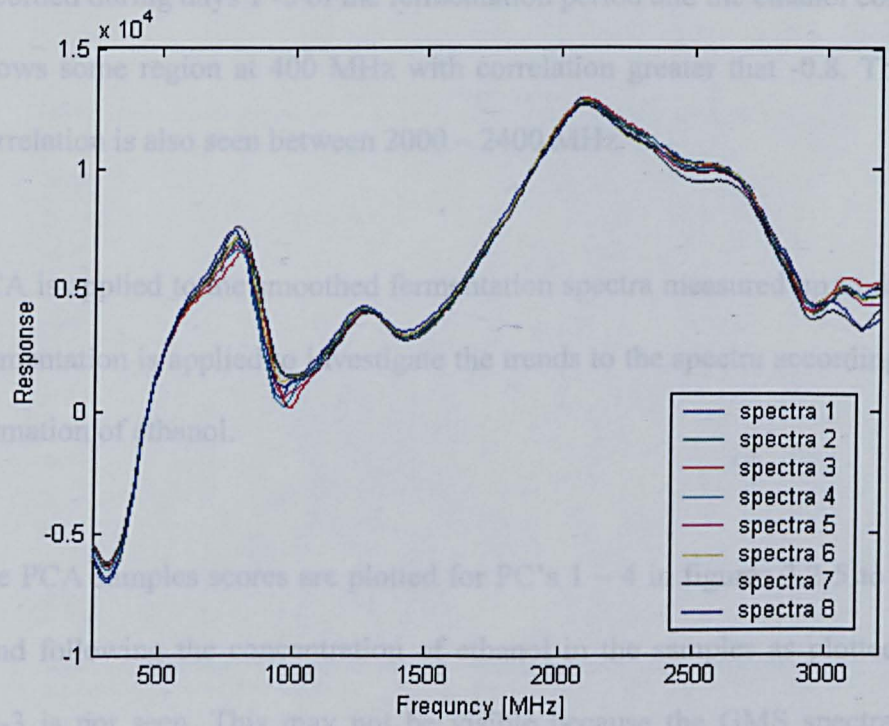
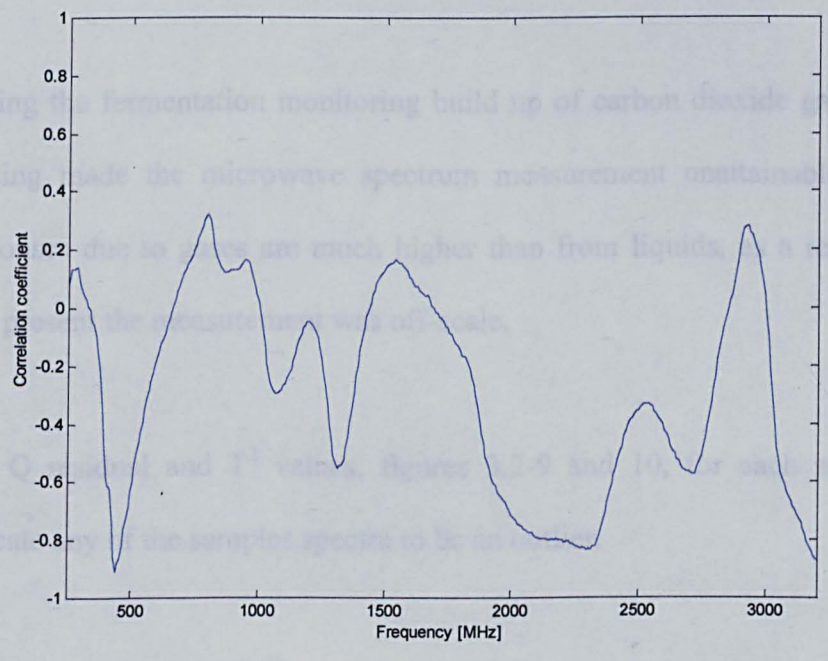


Figure 3.2-5 First 5 days beer fermentation spectra (smoothed) correlation to % ethanol in the 30 l batch beer samples



The correlation coefficient between the smoothed beer fermentation spectra recorded during days 1 -5 of the fermentation period and the ethanol concentration shows some region at 400 MHz with correlation greater that -0.8. This level of correlation is also seen between 2000 – 2400 MHz.

PCA is applied to the smoothed fermentation spectra measured up to day 5 of the fermentation is applied to investigate the trends to the spectra according to rate of formation of ethanol.

The PCA samples scores are plotted for PC's 1 – 4 in figures 3.2-5 to 8. A clear trend following the concentration of ethanol in the samples as plotted in figure 3.2-3 is not seen. This may not be visible because the GMS spectra could be masked by other components within the fermentation wort, such as ethanoic acid and dissolved carbon dioxide both of which will also give a response to microwaves.

During the fermentation monitoring build up of carbon dioxide gas just prior to venting made the microwave spectrum measurement unattainable. Microwave responses due to gases are much higher than from liquids, as a result when gas was present the measurement was off-scale.

The Q residual and T^2 values, figures 3.2-9 and 10, for each sample do not indicate any of the samples spectra to be an outlier.

Figure 3.2-5 PCA scores for PC 1 of smoothed spectra samples for the first five days of beer fermentation

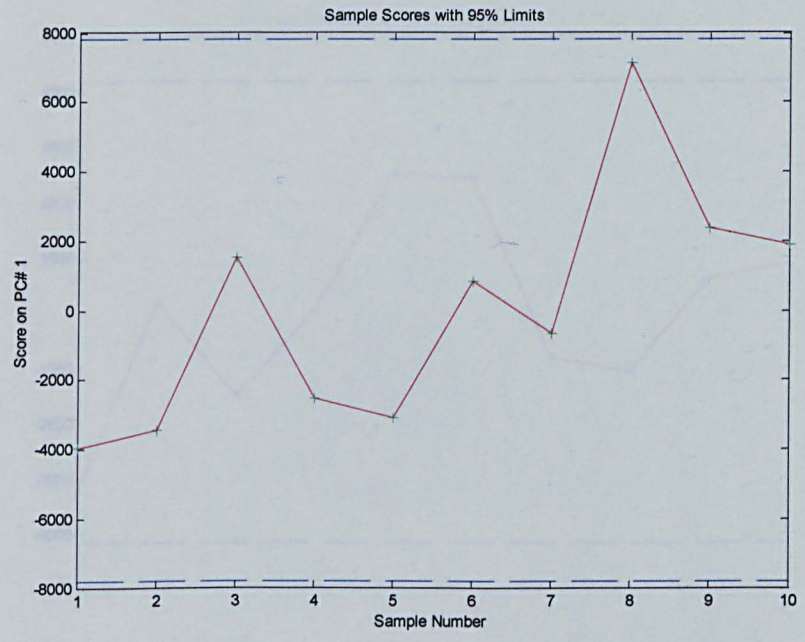


Figure 3.2-6 PCA scores for PC 2 of smoothed spectra samples for the first five days of beer fermentation

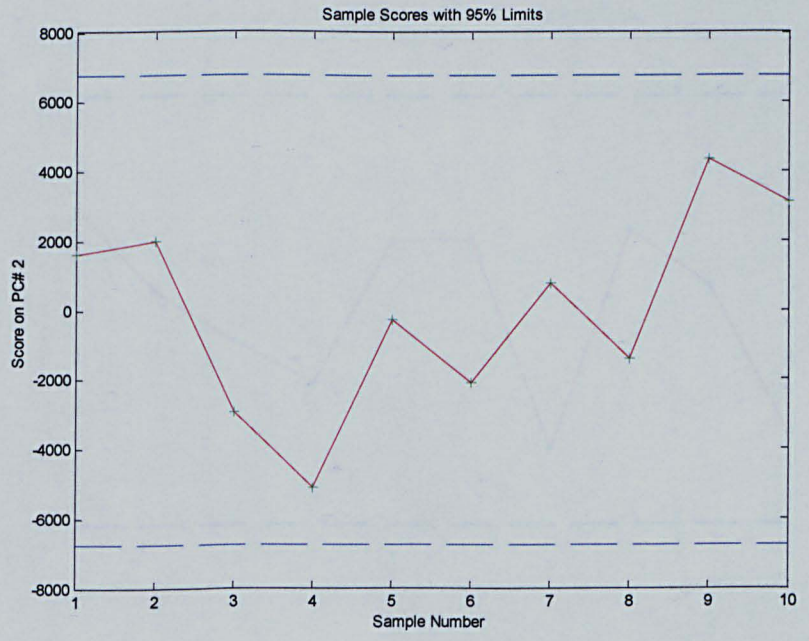


Figure 3.2-7 PCA scores for PC 3 of smoothed spectra samples for the first five days of beer fermentation

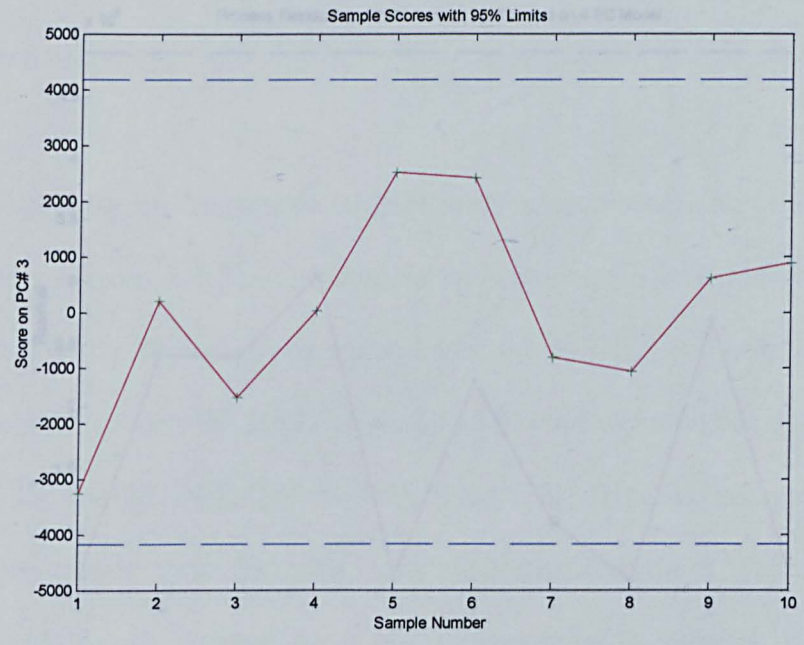


Figure 3.2-8 PCA scores for PC 4 of smoothed spectra samples for the first five days of beer fermentation

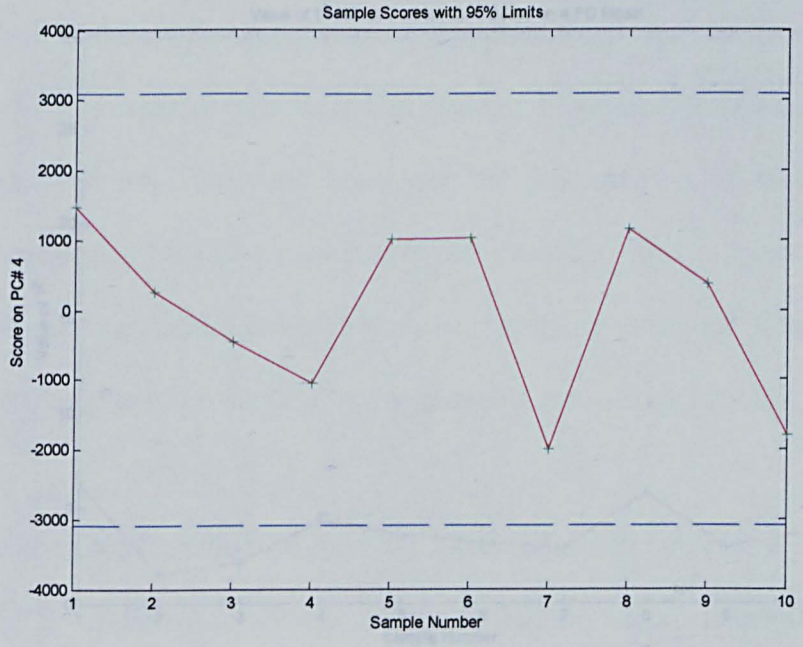


Figure 3.2-9 PCA Q residual for a 4 PC model of smoothed spectra samples for the first five days of beer fermentation

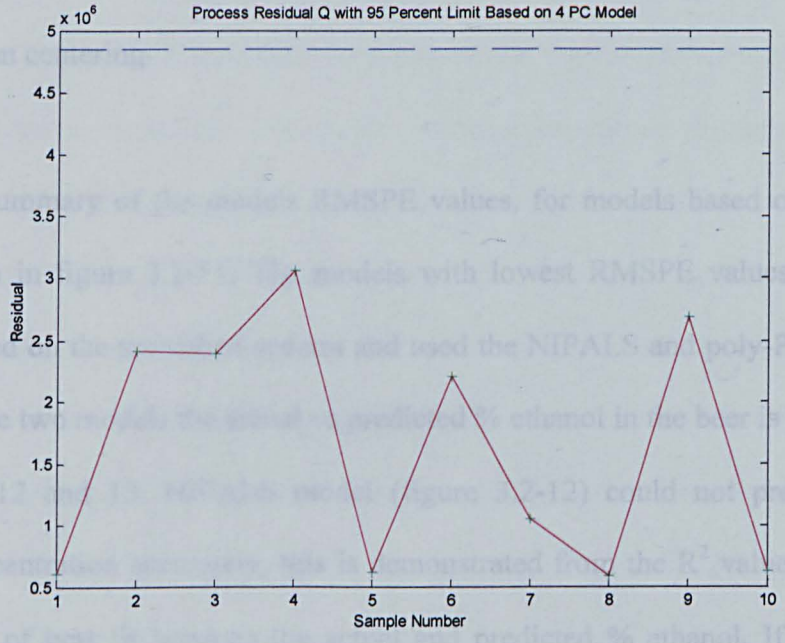
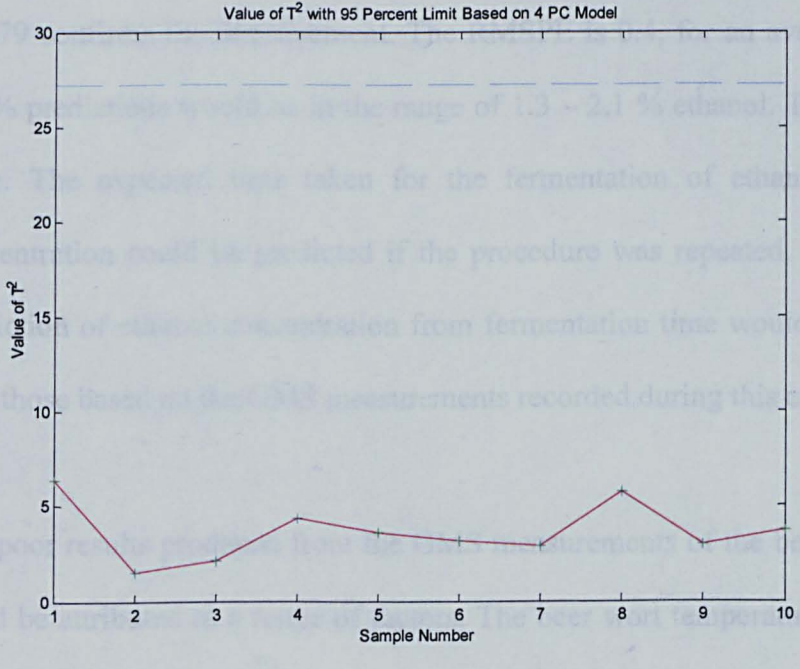


Figure 3.2-6 PCA T^2 residual for a 4 PC model of smoothed spectra samples for the first five days of beer fermentation



To model the ethanol concentration in the fermentation spectra a series of PLS models were calculated, these used PLS-NIPALS, poly-PLS and OSC followed by PLS. The models were calculated with smoothed spectra and with/without mean centering.

A summary of the models RMSPE values, for models based on 3 LVs, can be seen in figure 3.2-11. The models with lowest RMSPE values were calculated based on the smoothed spectra and used the NIPALS and poly-PLS methods. For these two models the actual vs predicted % ethanol in the beer is plotted in figures 3.2-12 and 13. NIPALS model (figure 3.2-12) could not predict the ethanol concentration accurately, this is demonstrated from the R^2 value of 0.024 for the line of best fit between the actual and predicted % ethanol. If the values were predicted accurately R^2 would be close to 1.

The poly-PLS model can predict the concentration of ethanol, the R^2 value is 0.7979 confirms the improvement. The RMSPE is 0.4, for an average sample of 1.7 % predictions would be in the range of 1.3 – 2.1 % ethanol. This range is too large. The expected time taken for the fermentation of ethanol to a certain concentration could be predicted if the procedure was repeated. It is likely that prediction of ethanol concentration from fermentation time would more accurate than those based on the GMS measurements recorded during this experiment.

The poor results produced from the GMS measurements of the beer fermentation could be attributed to a range of factors. The beer wort temperature was constant

when inside the reactor, but the recirculation loop and GMS sample chamber were not thermostated and the sample would have fluctuated in temperature as the ambient room temperature varied as it passed through the re-circulation system. The GMS system analyses everything within the sample chamber and is sensitive to volume variations. During the fermentation carbon dioxide gas is given off would bubble through the recirculation loop and into the flowcell during analysis, this could result in variations in the spectra due to changes in the liquid volume effecting the GMS spectra. The level of the analyte, ethanol, within the fermentation wort varied by less than 4 %, so the external factors could influence the GMS spectra to a greater extent interfering with the small spectral change due to ethanol.

To establish if this system can be used for the measurement of ethanol production in beer the experimental set-up should be modified to thermostat the entire system, control the analyte volume and minimise the effects of dissolved gases. Consideration should also be given to establish if the moving sample affects the total volume of analyte measured in a given time and consequently affects the GMS spectra.

Figure 3.2-11 Bar chart of RMSPE values for various PLS models based on the 30 l beer fermentation spectra recorded during days 1 – 5 of the fermentation

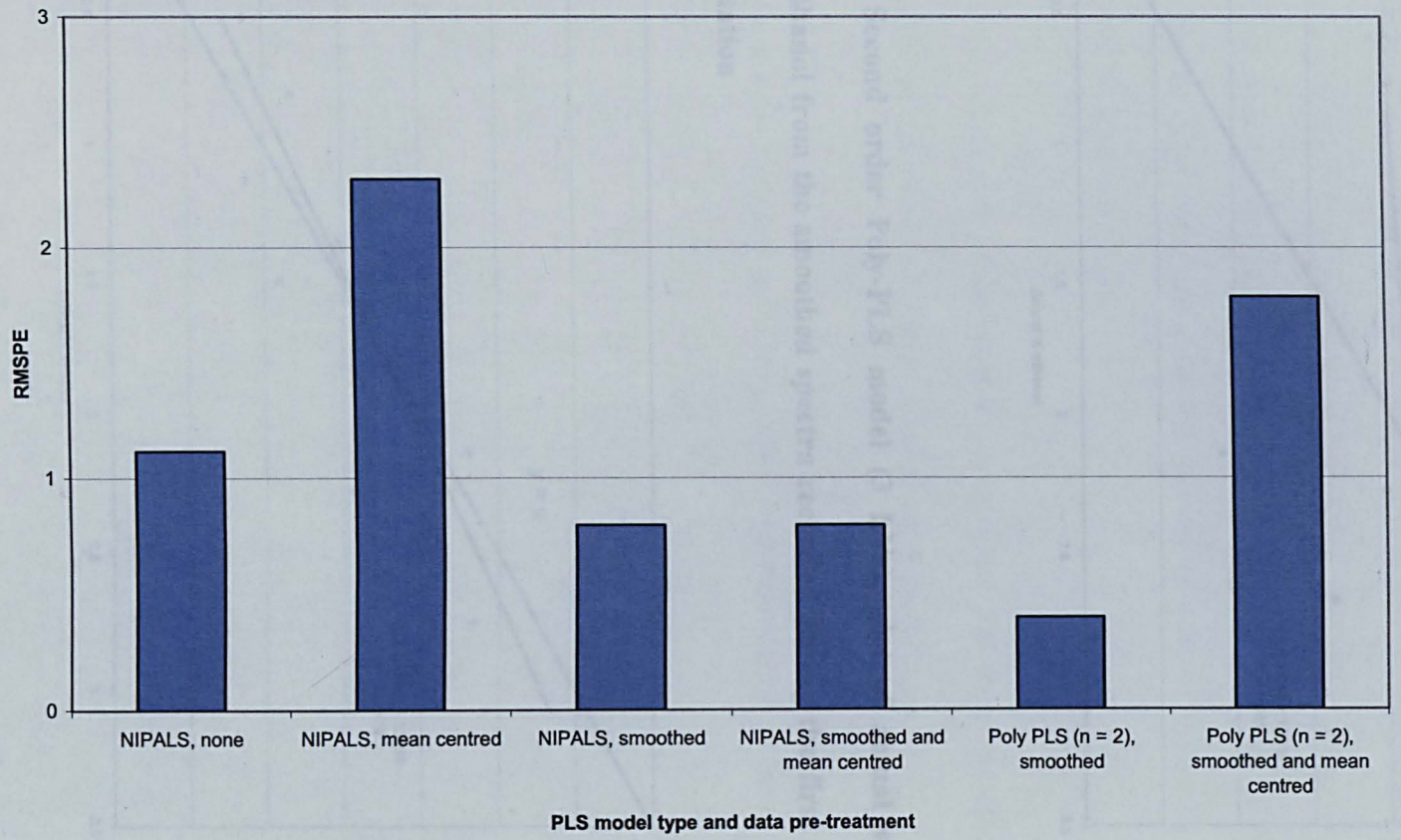


Figure 3.2-12 PLS model using 3 LVs, plot of actual vs predicted % ethanol from the smoothed spectra recorded during the first 5 days of fermentation

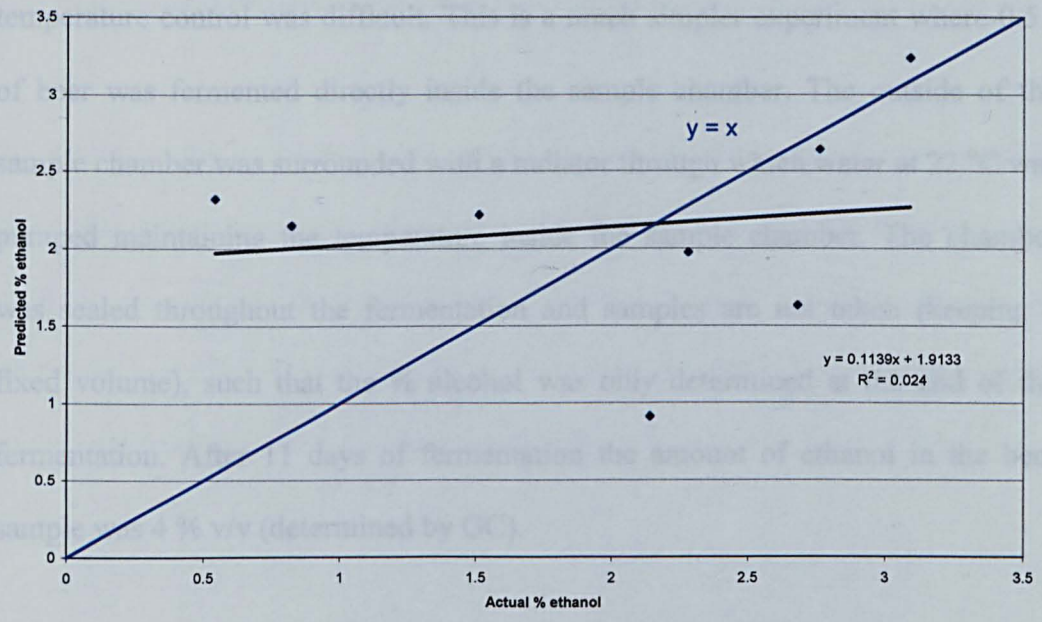
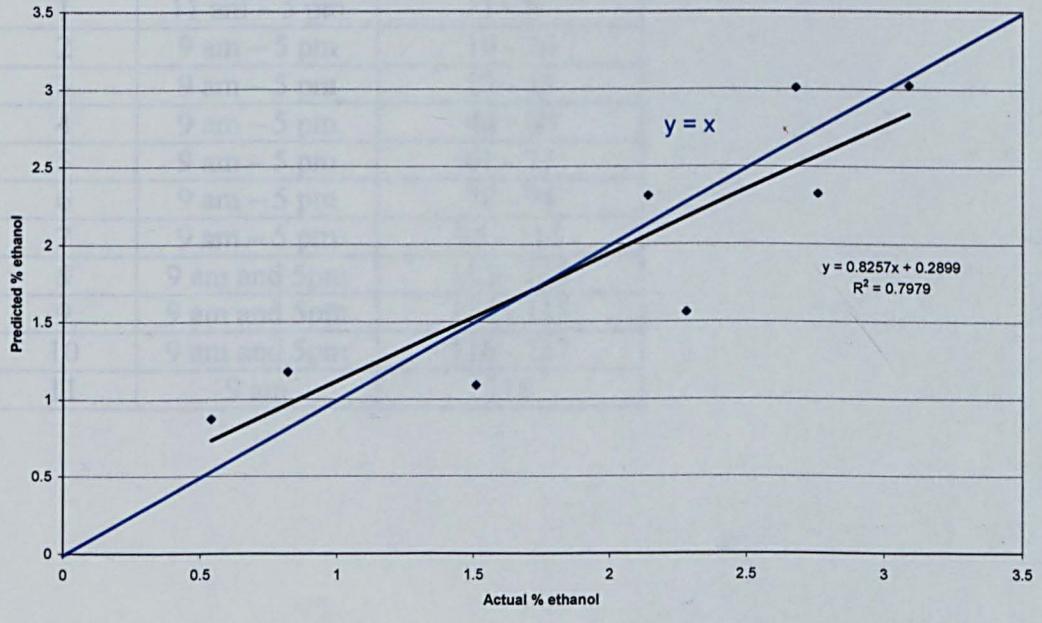


Figure 3.2-13 Second order Poly-PLS model (3 LVs), plot of actual vs predicted % ethanol from the smoothed spectra recorded during the first 5 days of fermentation



3.3 The monitoring of beer fermentation by GMS: 0.5 l

The previous experiment involved the large scale fermentation process where temperature control was difficult. This is a much simpler experiment where 0.5 l of beer was fermented directly inside the sample chamber. The outside of the sample chamber was surrounded with a radiator through which water at 22 °C was pumped maintaining the temperature inside the sample chamber. The chamber was sealed throughout the fermentation and samples are not taken (keeping a fixed volume), such that the % alcohol was only determined at the end of the fermentation. After 11 days of fermentation the amount of ethanol in the beer sample was 4 % v/v (determined by GC).

Table 3.3-1 describes when each of the spectra was recorded. For those collected between 9 am and 5 pm the measurements were taken every 30 minutes.

Table 3.3-1 Time of spectra collection during the 0.5 l fermentation of beer

Day	Time recorded	Spectra numbers
1	11 am – 5 pm	1 - 9
2	9 am – 5 pm	10 - 26
3	9 am – 5 pm	27 - 43
4	9 am – 5 pm	44 - 60
5	9 am – 5 pm	61 - 77
6	9 am – 5 pm	78 - 94
7	9 am – 5 pm	95 - 111
8	9 am and 5pm	112 - 113
9	9 am and 5pm	114 - 115
10	9 am and 5pm	116 - 117
11	9 am	118

In figure 3.3-1 the spectra measured at 4pm of each day is plotted. During the fermentation there are very little visible variations in the spectra, and the spectra resemble those of the ethanol in water standards (see figure 3.1.1-2).

To maximise the variation in the spectra the spectra are background subtracted, where the first spectrum recorded during the fermentation was subtracted from the each of the spectra recorded at 4pm. The background subtracted spectra are plotted in figure 3.3-2. Savitsky Golay smoothing was applied to reduce the noise in the background spectra, the smoothed spectra are plotted in figure 3.3-3.

Figure 3.3 1 GMS 0.5 l beer fermentation spectra recorded over 11 days

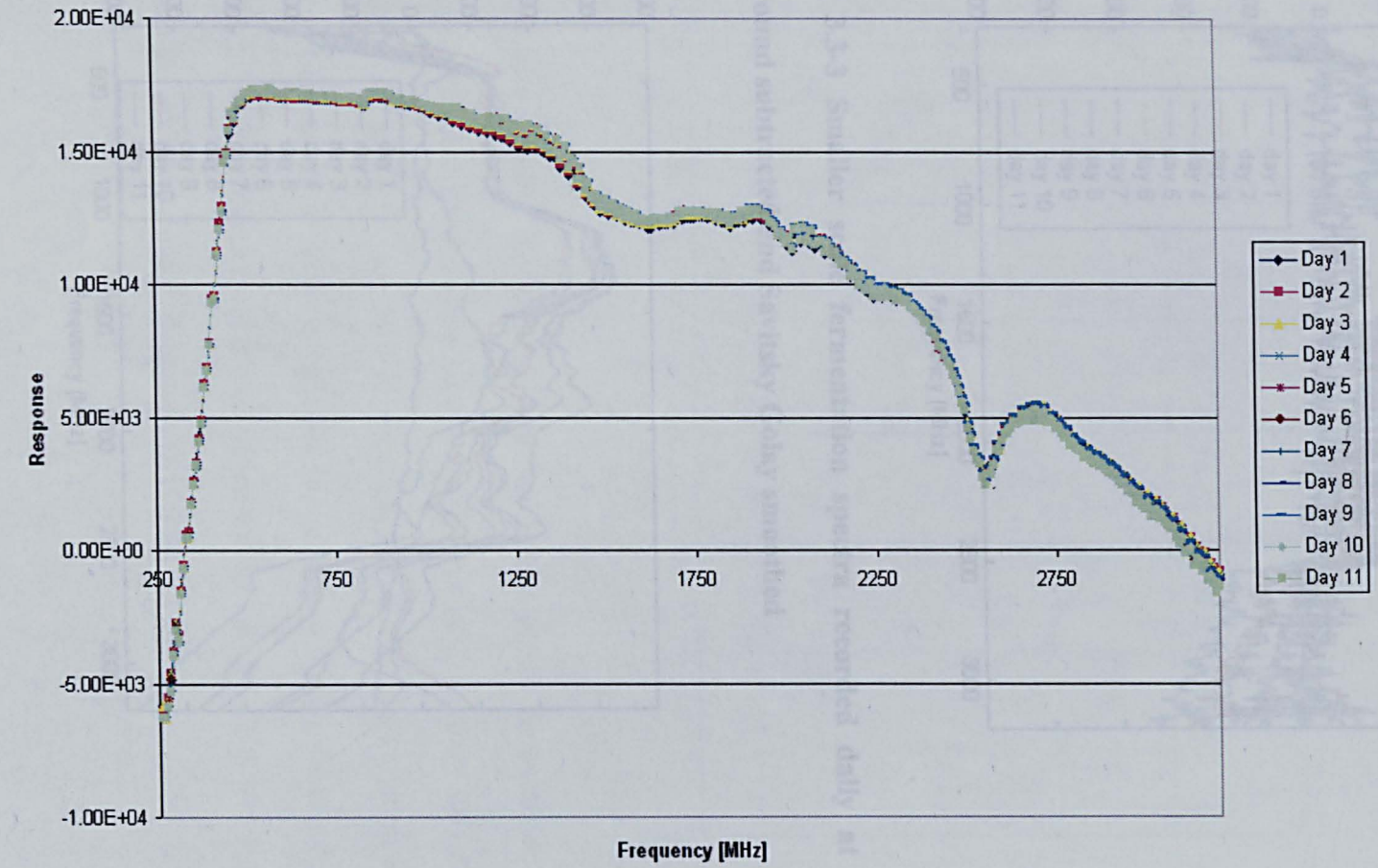


Figure 3.3-2 Smaller scale fermentation spectra recorded daily at 4pm background subtracted

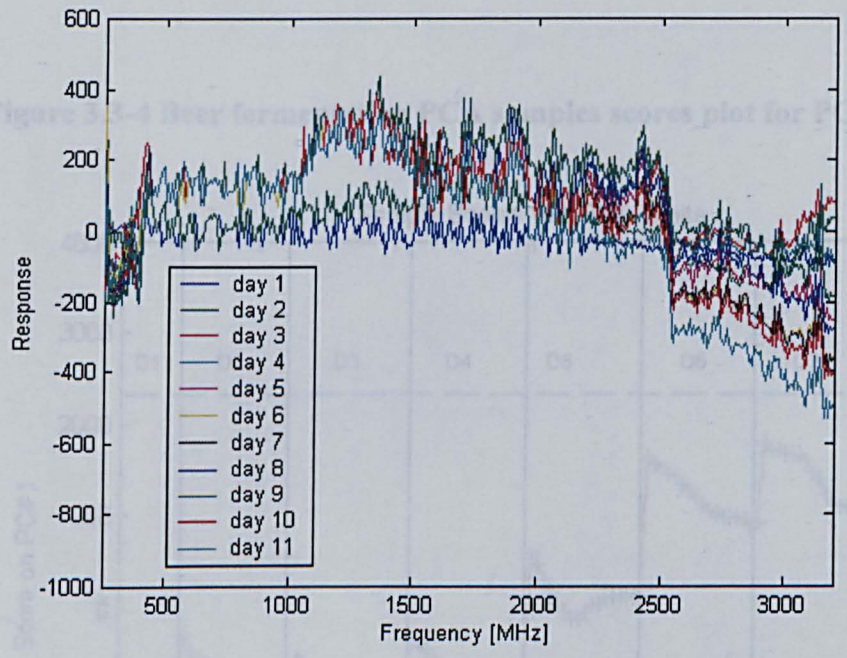
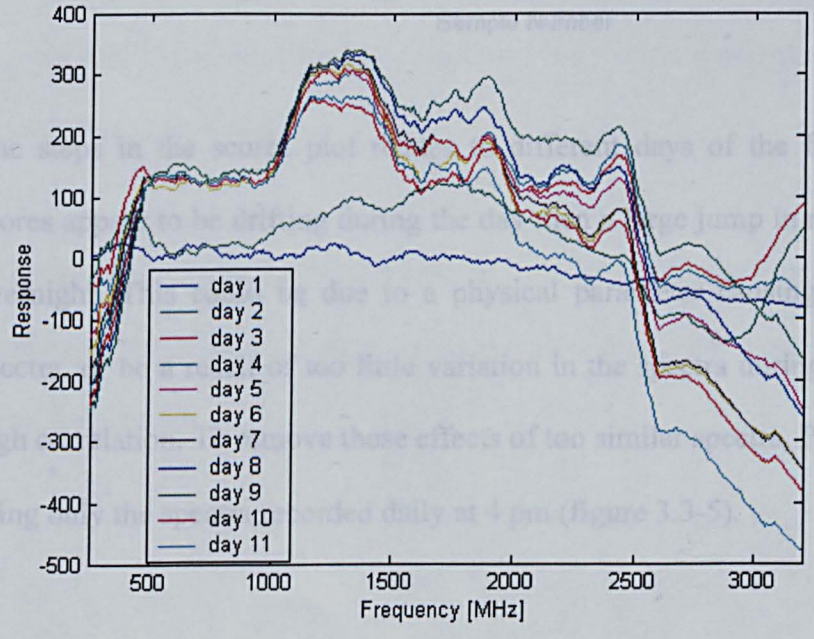
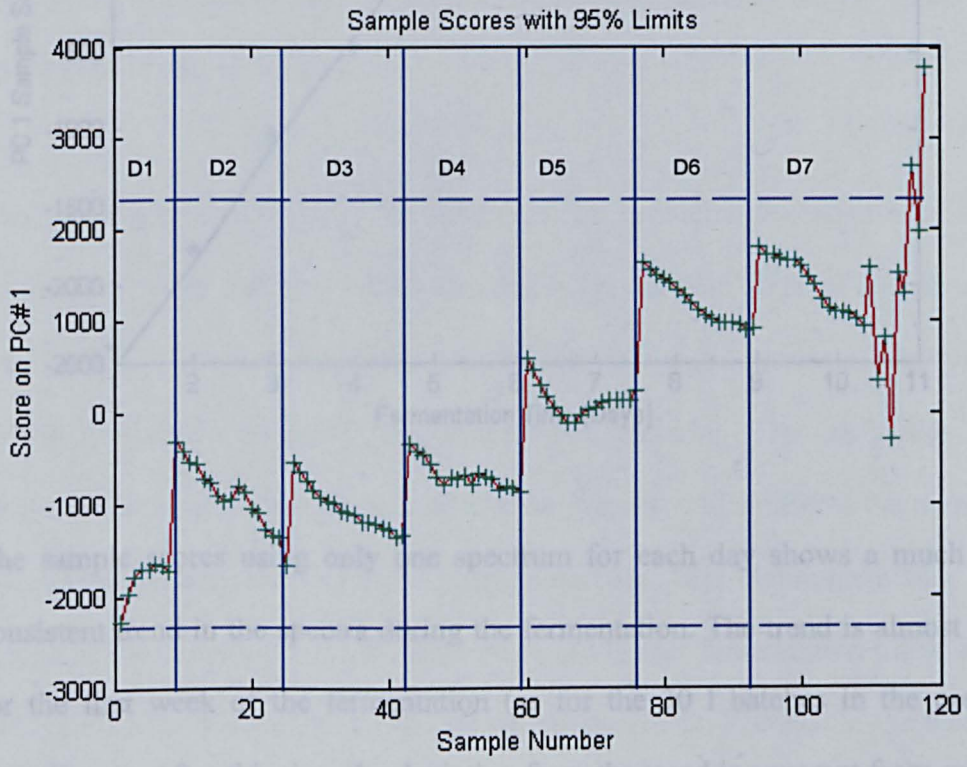


Figure 3.3-3 Smaller scale fermentation spectra recorded daily at 4pm background subtracted and Savitsky Golay smoothed



PCA was applied to all of the spectra recorded throughout during the fermentation (as described in table 3.3-1). The samples scores are plotted for PC 1 in figure 3.3-4.

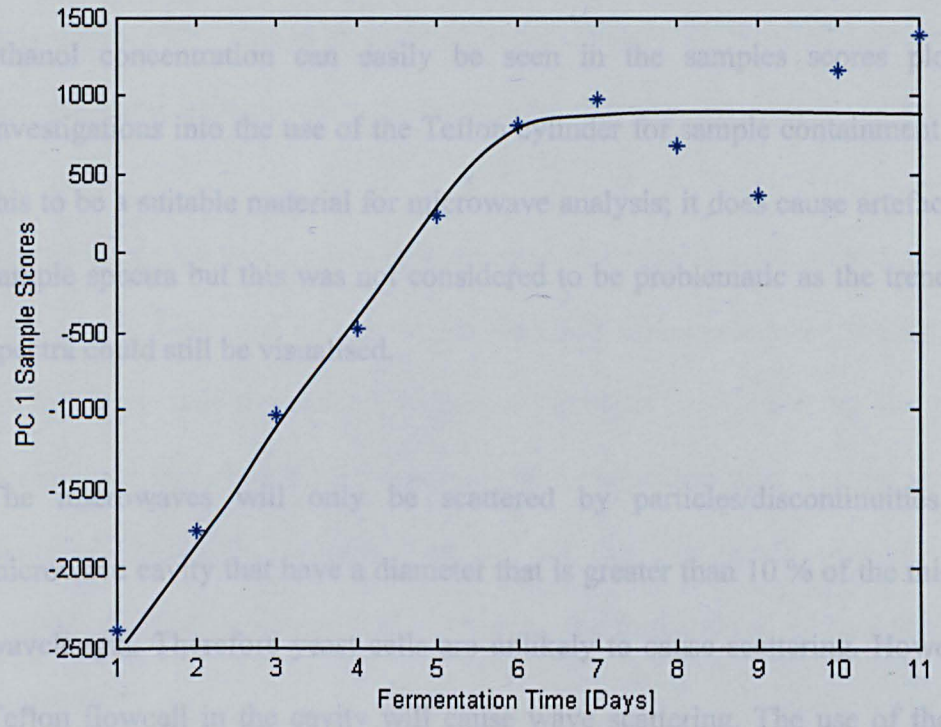
Figure 3.3-4 Beer fermentation PCA samples scores plot for PC 1



D = Day

The steps in the scores plot relates to different days of the fermentation. The scores appear to be drifting during the day then a large jump in magnitude is seen overnight. This could be due to a physical parameter causing variation in the spectra, or be a result of too little variation in the spectra during the day causing high correlation. To remove these effects of too similar spectra, PCA was repeated using only the spectra recorded daily at 4 pm (figure 3.3-5).

Figure 3.3-5 Beer fermentation spectra at 4 pm daily; PCA scores for PC 1



The sample scores using only one spectrum for each day shows a much more consistent trend in the spectra during the fermentation. The trend is almost linear for the first week of the fermentation (as for the 30 l batches in the previous experiment). After this time the deviation from the trend is apparent from samples recorded on days 8 and 9.

From this experiment the trend in the spectra during the fermentation process was more apparent in the PCA samples scores plots than from those of the large scale batch fermentation.

4 Conclusions

This work has shown that for standard samples the effect of the increase in ethanol concentration can easily be seen in the samples scores plots. The investigations into the use of the Teflon cylinder for sample containment showed this to be a suitable material for microwave analysis; it does cause artefacts in the sample spectra but this was not considered to be problematic as the trends in the spectra could still be visualised.

The microwaves will only be scattered by particles/discontinuities in the microwave cavity that have a diameter that is greater than 10 % of the microwave wavelength. Therefore yeast cells are unlikely to cause scattering. However, the Teflon flowcell in the cavity will cause wave scattering. The use of the Teflon cylinder also changes the amount of analyte present, which affects the number of encounters between the wave and polar molecules. This highlights a flaw in the design of the experiment for the large scale (30 l) beer fermentation for which the Teflon cylinder was used as a flowcell for measurement. The aim of the work was to look at changes that occur in the brew during fermentation, from the microwave spectra collected. As the spectra were taken under the same conditions, scattering caused by the Teflon cylinder should not effect the between-spectra variation.

The spectra from the large scale fermentation process recorded on-line could not be used to predict accurately the amount of ethanol produced throughout the process. Samples recorded during the first 5 days of the fermentation were most suitable for calibration models as they contained the most variation due to the

rapid production of ethanol which slowed after this period. The lowest RMSPE value was 0.4, achieved when the spectra were smoothed and modelled by polynomial PLS. This is an error of approximately 25 % of an average sample concentration, indicating that the method requires improvement for analysis.

The fermentation process was then repeated in a much simpler experiment. 500ml of beer was fermented directly inside the GMS sample chamber, the temperature of the brew was maintained with a re-circulator bath reducing the effect of temperature variations on the spectra. The fermentation wort was not sampled or disturbed during the process. The aim for this experiment was to use trend analysis methods, such as PCA to visualise the formation of ethanol via the PCA scores plots.

The results were consistent with the previous experiment and the greatest amount of variation in the spectra occurred in the first 5 days. From visual interrogation of the samples it was difficult to see the change to due to formation of ethanol. The PCA scores plots show the rise in ethanol content. There is a drifting of the samples scores over the course of each day, although it was anticipated that the sample temperature control was sufficient to maintain accurately the sample temperature.

These experiments show that guided microwave spectroscopy can be used to monitor the rate of formation of ethanol in beer but development in the

experimental set-up is required to ensure the samples are not affected by variations in ambient temperature.

5 Further Work

The fermentation should be repeated under controlled temperature conditions. This should be arranged for the large scale batch reaction from which sampling is possible without significantly affecting the volume of the fermentation wort.

The reference analysis for the determination of ethanol in the beer requires improvement. It is not known if the GC signal was drifting causing the variation in the beer % ethanol towards the end of the batch or whether the fermentation wort was oxidising. A way to improve the situation would be use a dedicate beer analyser such as the Scaba™ mentioned in the introduction which has proven reliability for reference analysis.

The Teflon cylinder should not be used for further experiments, and should either be replaced with an alternative that fits the dimensionality of the sample chamber or the sample chamber should be adapted to function as a flowcell itself.

Most importantly, the GMS system requires modification to allow the sample spectra to be recorded automatically. This would prevent the daily jumps in the samples spectra and scores plots especially visible in the results of the small scale fermentation experiment. After consultation with the manufacturers this should be possible with addition of new circuitry to the spectrometer[8].

6 References

[1] Townshend A, Ed., 'Encyclopaedia of Analytical Science', Vol. 2, Academic Press, London, 1995, pg. 1226.

[2] Helrich K, 'Official Methods of Analysis', fifteenth edition, Association of Official Analytical Chemists, Arlington, Virginia, USA, 1990, pg. 710 - 711.

[3] Hornsey I. S, 'Brewing', RSC Paperbacks, London, 1999, 183.

[4] Gallignani M, Garrigues S, de la Guardia M, 'Derivative Fourier-Transform Infrared Spectrometric Determination of Ethanol in Beers', Analyst, 1994, 119 (8), 1773 – 1778.

[5] Gallignani M, Garrigues S, de la Guardia M, 'Stopped-flow Near-Infrared Spectrometric Determination of Ethanol and Maltose in Beers', Analytica Chimica Acta, 1994, 276, 155 – 161.

[6] www.foss.dk/c/p/solutions/products/showprodfamily.asp?prodfamilypkid=103. Viewed: 20/1/03.

[7] http://www.asiinstr.com/dc1.html#List, Viewed 20/2/03

[8] Butler D, Thermo Moisture Systems, Personal Communication (E-mail), 10/5/02.

Chapter 6

Analysis of Industrial Process Samples by Guided Microwave Spectroscopy

1 Introduction

The aim of this work was to investigate the feasibility of using microwave spectroscopy for monitoring of an industrial process. Under investigation is Avecia's BIT process[1]. This has multiple stages during the manufacturing process. One of the mid-stages requires the oxidation of an intermediate product to reach 100 % to gain maximum yield of the final product. An oxidation level of less than 98 % will affect the end product morphology which results in plant system blockages and loss of product.

This oxidation stage has become a bottleneck in the manufacture, as the intermediate stage is processed for several hours more than strictly necessary to ensure sufficient oxidation. The time taken to manufacture the product could be reduced with the implementation of an on-line monitoring system to determine the oxidation level of the intermediate product.

The oxidation stage product is a darkly coloured mixture of organic, immiscible aqueous phases and solid particulates (see figure 2.1-1). The exact details of the process were not released by Avecia for confidentiality reasons. Only the following information was provided about the samples under scrutiny, the aqueous phase contained bisamide and thiosalicylamide, the solid phase contained the oxidation product which was dithiodibenzamide. The oxidation level of the dithiodibenzamide is inferred by HPLC analysis of the aqueous layer components of bisamide and thiosalicylamide. This is a time consuming method which

requires the sample to be separated into the different layers prior to the HPLC analysis which would then take between 40 – 60 minutes.

There are few references of measurement systems for multiphase samples; to date it has been much simpler to produce an engineering answer for such samples so that they are separated into forms that can be analysed separately. A feasibility study performed by Avecia personnel has already ruled out the use of NIR spectroscopy to monitor this process.

The GMS method of analysis should be suitable for the analysis of this process. The microwave response will be a measurement of the composition of the entire sample, including all phases. The oxidation process samples contain polar molecules which are suitable for microwave measurement. The particulates should not interfere with the GMS measurement, as the particulates are 200 - 300 μm in diameter, for interference the particulates diameter must be greater than 1/10th of the GMS sample cavity pathlength.

To investigate the effect of partially soluble solutions, a short preliminary study was carried out. For this, a series of acetonitrile (ACN) in water samples (5 – 30 %) were analysed and the amount ACN was predicted by PLS and WRR with approximately 1 % error[2].

Figure 2.1-1 Avecia BIT process sample

2 Experimental

2.1 Analysis of acetonitrile in water

Acetonitrile (HPLC Grade, 99 %, Fisher Scientific) samples, of concentrations 5, 8, 10, 12, 15, 20, 23, 25, and 30 % v/v in water, were prepared in 500 ml volumetric flasks and analysed by GMS.

2.2 Analysis of industrial process samples

10 samples were taken from the process steam, and an additional sample was prepared in the laboratory to give a more even distribution of oxidation levels.

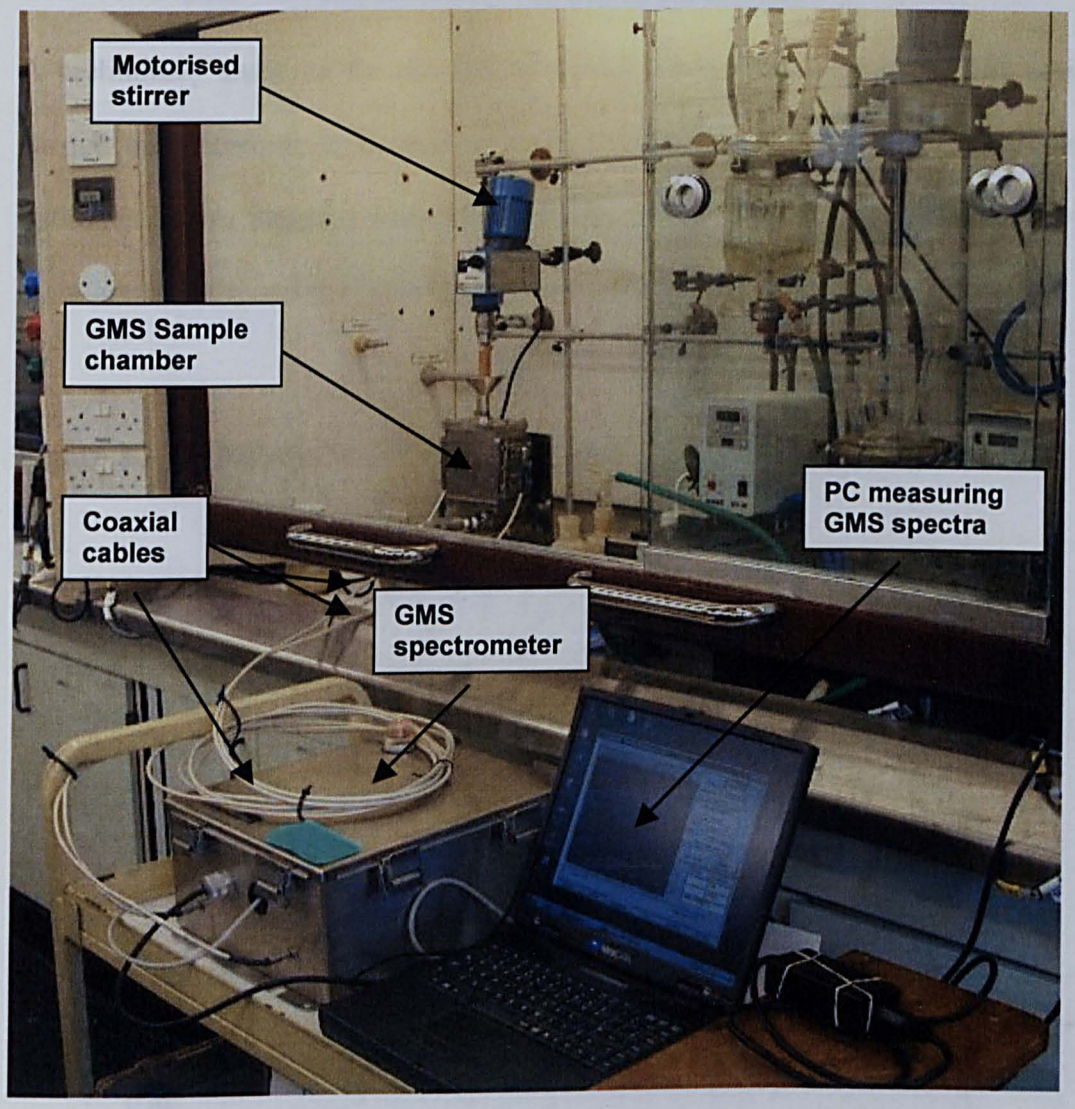
2.2.1 Reference analysis for % oxidation by HPLC

The samples were separated in the laboratory and the aqueous layer analysed by Avecia personnel using HPLC for indirect determination of the oxidation level of the solid within the slurry.

2.2.2 Analysis by GMS

The analysis was carried out at Avecia's Huddersfield works, and the laboratory set-up can be seen in figure 2.2.2-1. The samples are highly dermatitic[3] so were only handled by trained Avecia personnel. Analysis was in a random order, each sample was shaken then charged into the sample chamber, this was sealed and the solution agitated with a motorized stirrer to ensure homogeneity and to represent the process stream. After the first minute of stirring, the spectrum of each sample was recorded 10 times. The samples were pumped out of the sample chamber which was then rinsed with distilled water.

Figure 2.2.2-1 Laboratory set-up for analysis of industrial process samples spectra by GMS



2.2.2.1 Effect of phase variation in industrial process sample

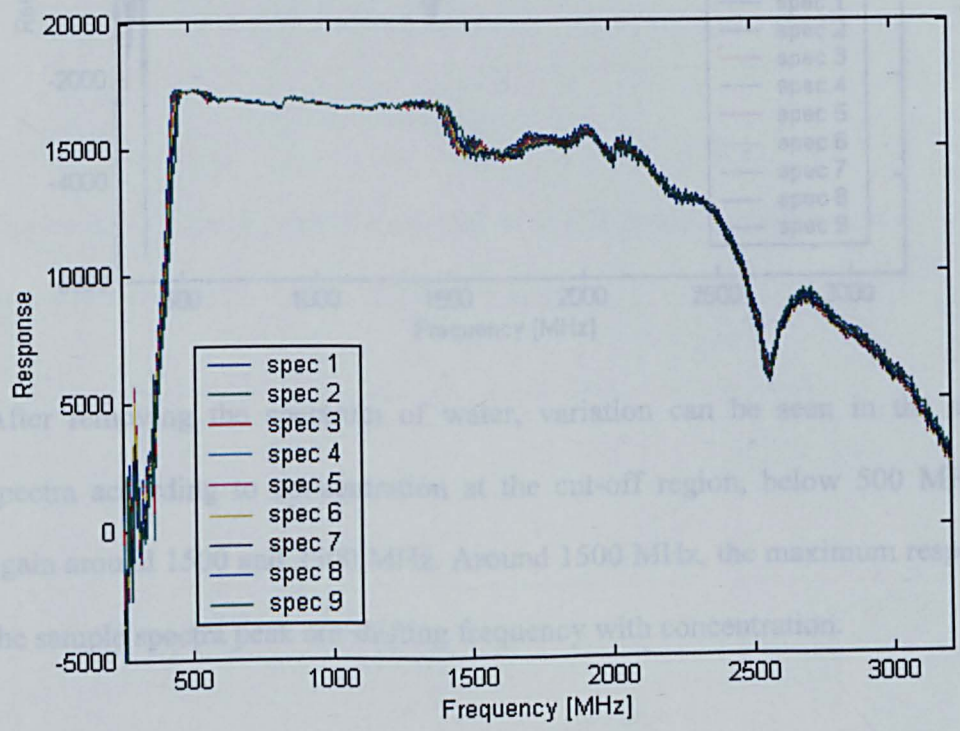
The aim was to determine if the level of mixing or phase variation effected the microwave spectrum of a sample. Sample 6, of 92.0 % conversion (see table 3.2.1-1) was placed into the sample chamber, the solution was agitated with a motorised stirrer for 2 minutes, the stirrer was then switched off and the spectrum recorded continuously for a further 3 minutes.

3 Results

3.1 Analysis of acetonitrile in water by GMS

A preliminary study on the analysis of partially soluble solutions by GMS was carried out. Acetonitrile is only partially soluble in water and does not form hydrogen bonds. Samples were analysed between the concentration range of 5 – 30 % acetonitrile, and the GMS spectra can be seen in figure 3.1-1.

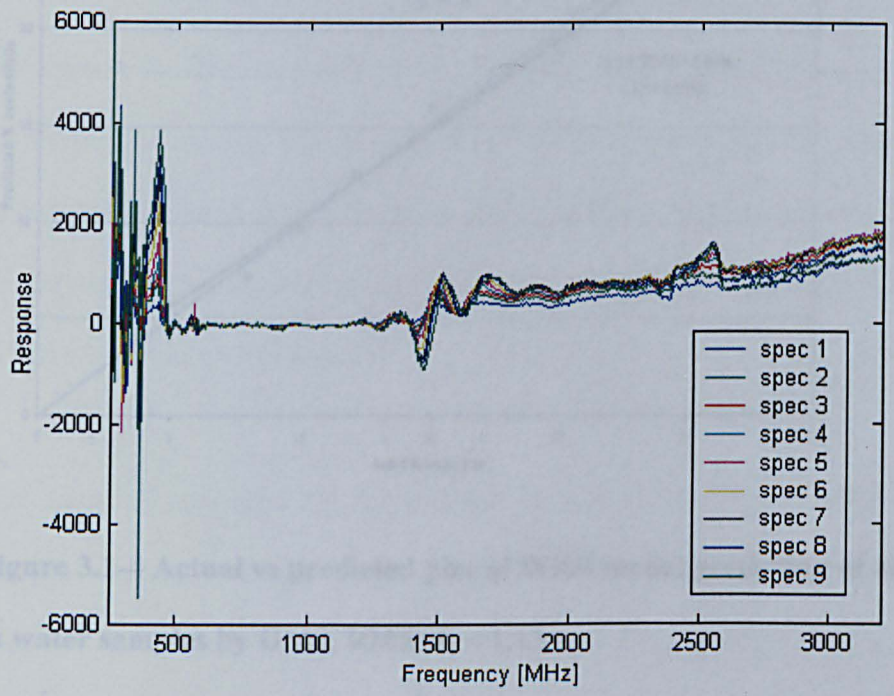
Figure 3.1-1 GMS spectra of 9 acetonitrile in water samples (5 – 30 %)



The samples spectra look very similar to that of pure water and there is very little variation as the level of acetonitrile increases to 30 %. To increase the variation in the spectra and minimise the response to water, the acetonitrile samples are background subtracted by subtracting the spectrum of water, see figure 3.1-1. The

GMS spectrum of water is very reproducible provided that the same volume of water (and analyte samples where background subtracted) is measured.

Figure 3.1-2 GMS background spectra of acetonitrile in water samples



After removing the spectrum of water, variation can be seen in the samples spectra according to concentration at the cut-off region, below 500 MHz, and again around 1500 and 2500 MHz. Around 1500 MHz, the maximum response of the sample spectra peak are shifting frequency with concentration.

This is a common feature of the guided microwave system previously seen in chapter 5 with the alcohols mixtures. Using the background subtracted spectra, PLS (2 LVs) and WRR CV calibration models were calculated. In figure 3.1-3 the actual vs predicted % acetonitrile is plotted for the PLS and WRR models.

Figure 3.1-3 Actual Vs predicted plots of PLS model prediction of acetonitrile in water samples by GMS, RMSPE = 1.24

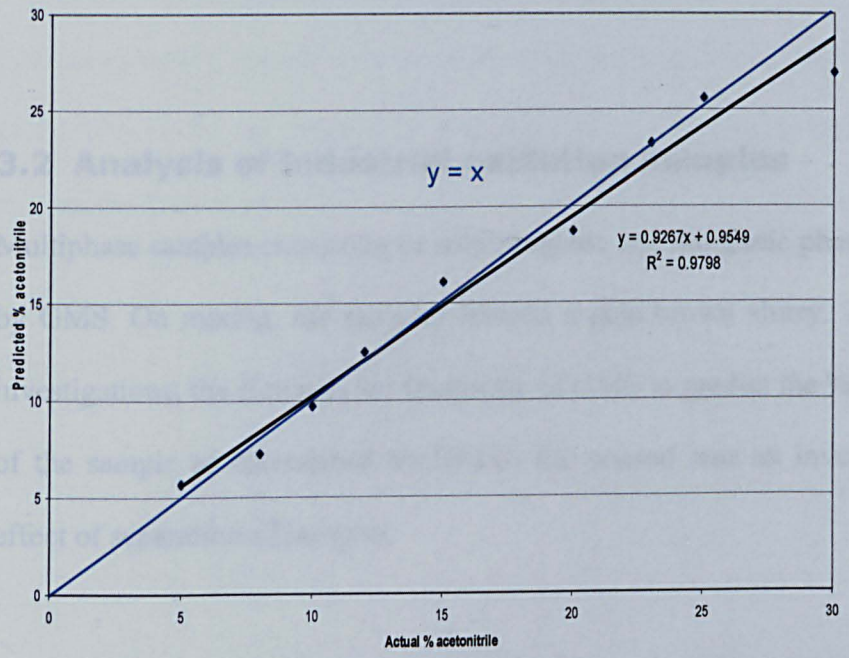
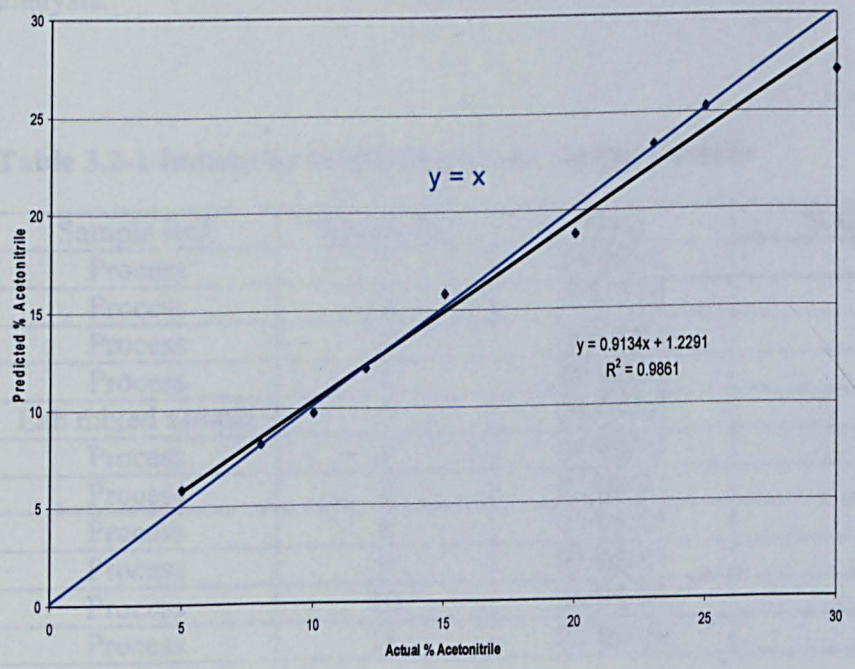


Figure 3.1-4 Actual vs predicted plot of WRR model prediction of acetonitrile in water samples by GMS, RMSPE = 1.13



Both methods have good agreement when the actual % acetonitrile is plotted against the predicted % acetonitrile. The methods were found to be comparable with the RMSPE value for both methods being just over 1 %.

3.2 Analysis of industrial oxidation samples

Multiphase samples consisting or solid, organic and inorganic phases are analysed by GMS. On mixing, the samples formed a pale brown slurry. There were two investigations; the first was the feasibility of GMS to predict the % oxidation level of the sample as determined by HPLC; the second was an investigation of the effect of separation of samples.

The samples measured for these experiments are detailed in table 3.2-1. The sample reference details sample origin, i.e. it was taken from the process or prepared in the laboratory. The % conversion is the oxidation % from HPLC analysis.

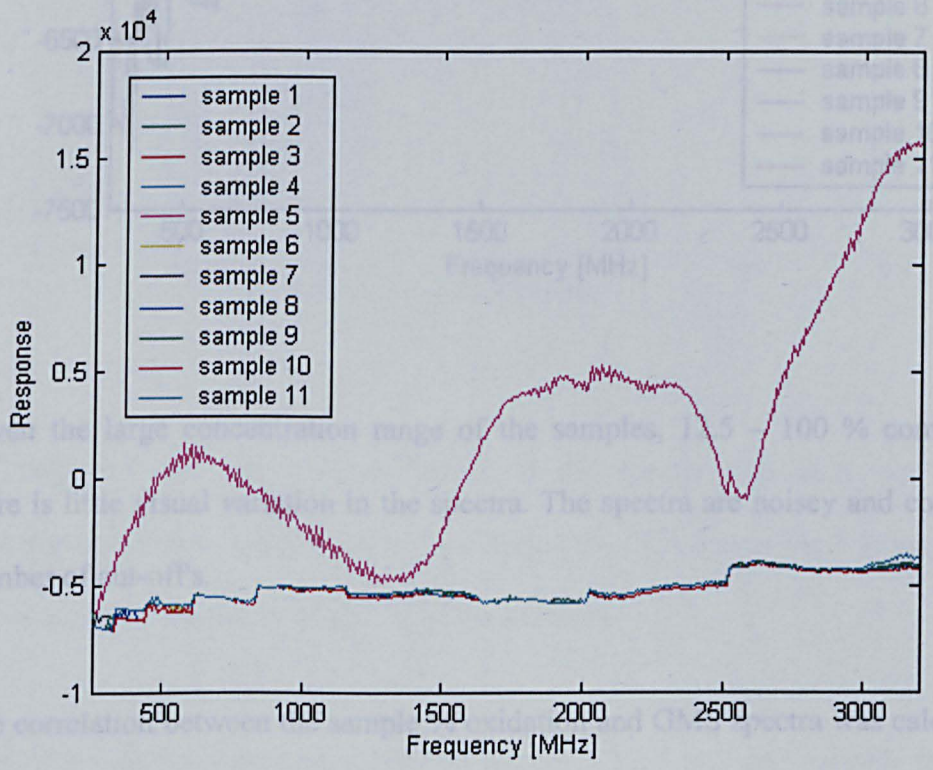
Table 3.2-1 Industrial oxidation process samples details

Sample Ref:	Sample No.	Date	% Conversion
Process	1	02/08/01	12.5
Process	2	02/08/01	45.9
Process	3	06/08/01	50.0
Process	4	02/08/01	52.4
Lab mixed sample	5	08/08/01	84.3
Process	6	06/08/01	92.0
Process	7	02/08/01	93.0
Process	8	02/08/01	98.4
Process	9	02/08/01	99.8
Process	10	06/08/01	99.9
Process	11	02/08/01	100

3.2.1 Analysis of oxidation process samples by GMS

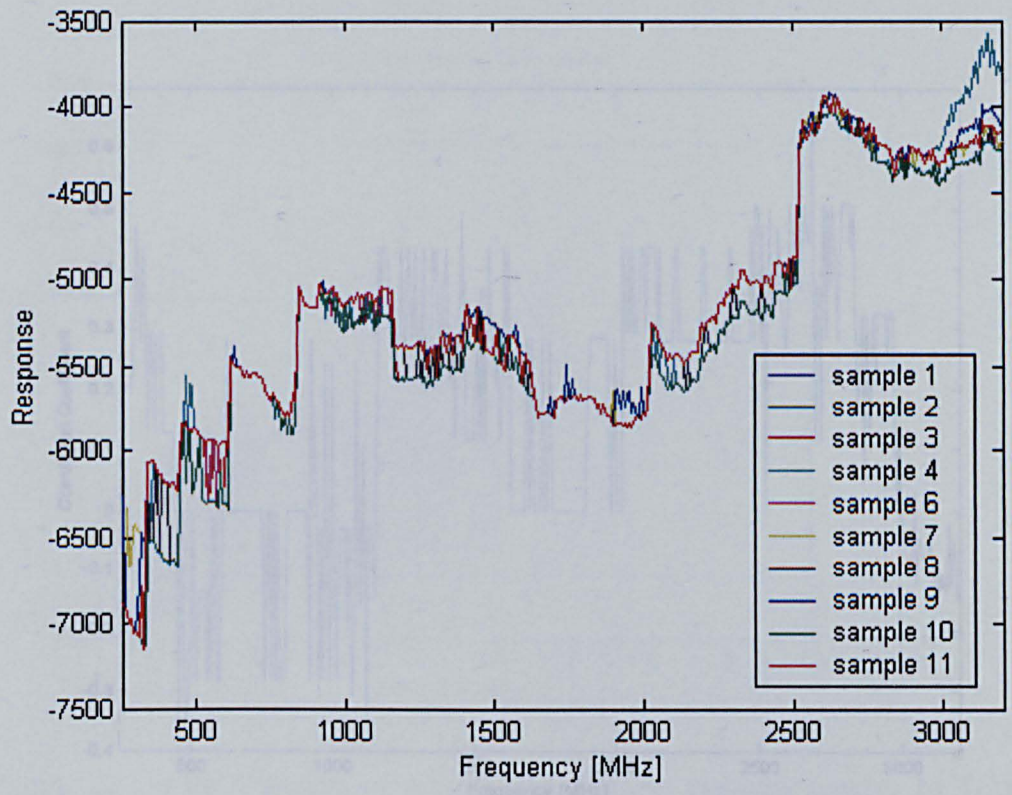
The GMS spectra of the oxidation process samples are plotted in figure 3.2.1-1. The spectrum of sample 5 was easily identified as an outlier. This sample was the laboratory mixed sample, where a low oxidation sample (12 %), was oxidised up to 84.6 % in the laboratory[4]. The sample was consequently excluded from the dataset for further analysis.

Figure 3.2.1-1 Avecia oxidation samples GMS spectra



The spectra of samples from the process stream are in figure 3.2.1-2.

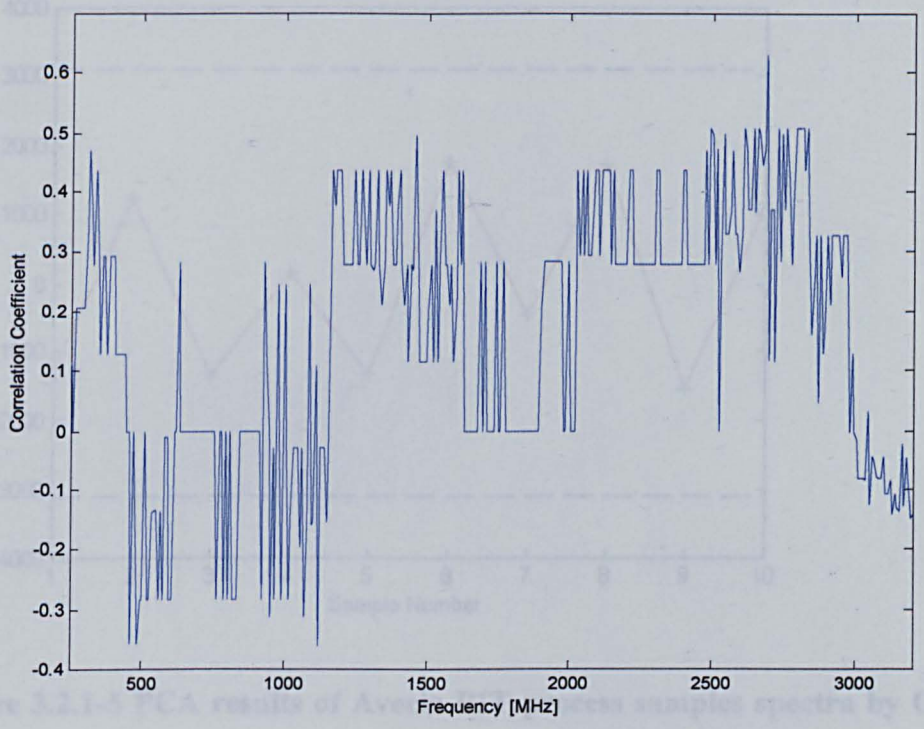
Figure 3.2.1-2 Avecia oxidation samples GMS spectra excluding sample 5



Given the large concentration range of the samples, 12.5 – 100 % conversion there is little visual variation in the spectra. The spectra are noisey and contain a number of cut-off's.

The correlation between the sample % oxidation and GMS spectra was calculated and plotted in figure 3.2.1-3. The maximum correlation was less than 0.6, indicating that the spectra measured are not highly related to the samples oxidation.

Figure 3.2.1-3 Avecia oxidation samples correlation to % oxidation and spectra



As there are not any obvious trends in the spectra, the next step in the investigation was to apply PCA and scrutinize the PCA scores plots to see if they are consistent with the % oxidation of the samples.

The PCA samples scores are plotted in figures 3.2.2-3 to 7 and variable loadings in figure 3.2.2-8 to 11. The scores plots do not show a trend with the % oxidation of the samples. The loadings PC plots are noisy and the shapes reflect each other, suggesting each may be cancelling the previous information out. The Q residual (figure 3.2.2-12) highlights samples 2 and 10 to be of high leverage. Hotellings T^2 test (figure 3.2.2-13) does not identify any sample to be an outlier.

Figure 3.2.1-4 PCA results of Avecia BIT process samples spectra by GMS; PC 1 scores

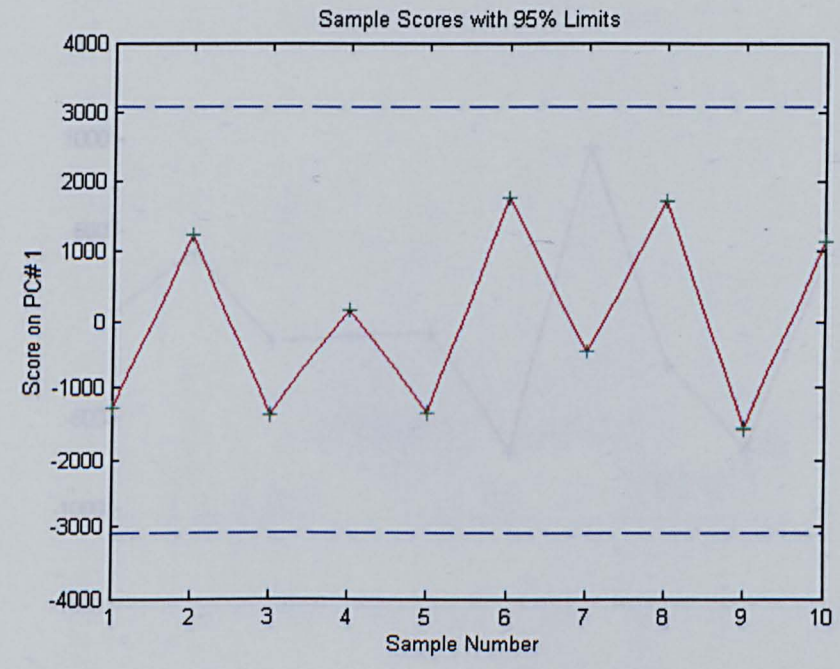


Figure 3.2.1-5 PCA results of Avecia BIT process samples spectra by GMS; PC 2 scores

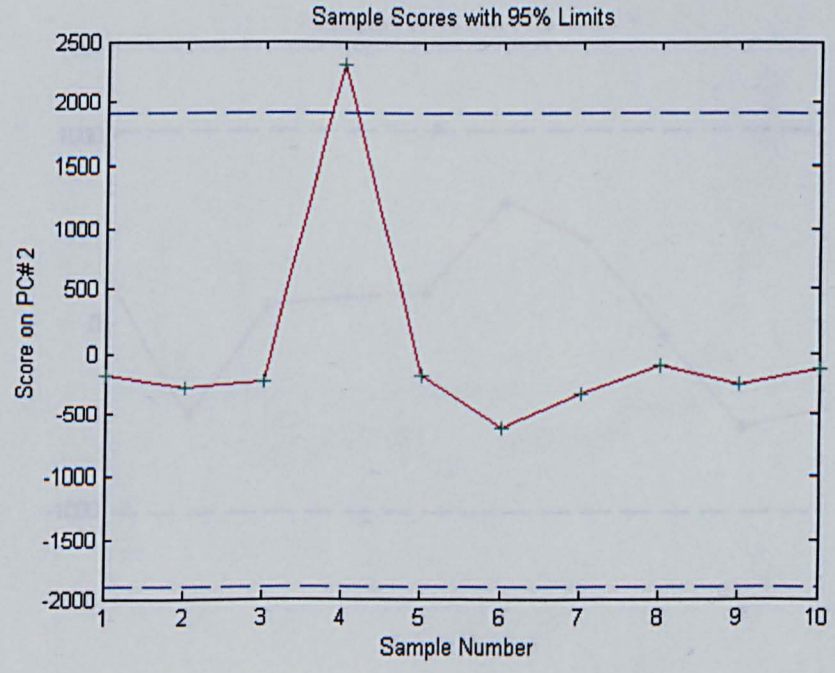


Figure 3.2.1-6 PCA results of Avecia BIT process samples spectra by GMS; PC 3 scores

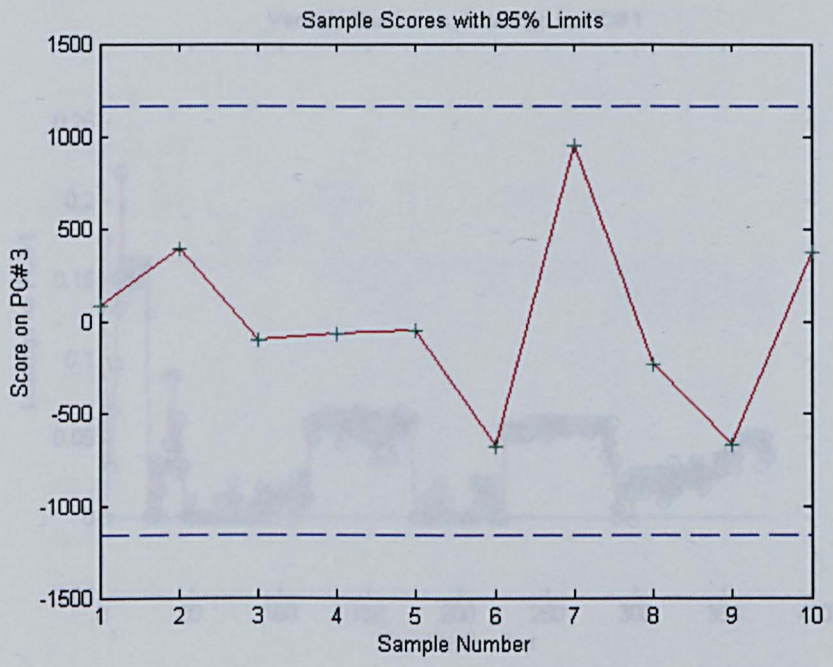


Figure 3.2.1-7 PCA results of Avecia BIT process samples spectra by GMS; PC 4 scores

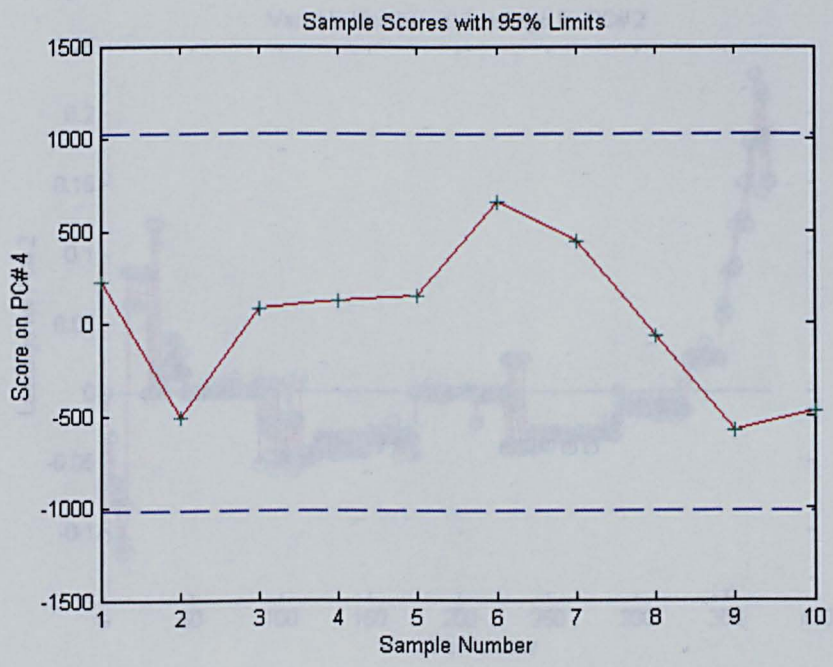


Figure 3.2.1-8 PCA results of Avecia BIT process samples spectra by GMS; PC 1 loadings

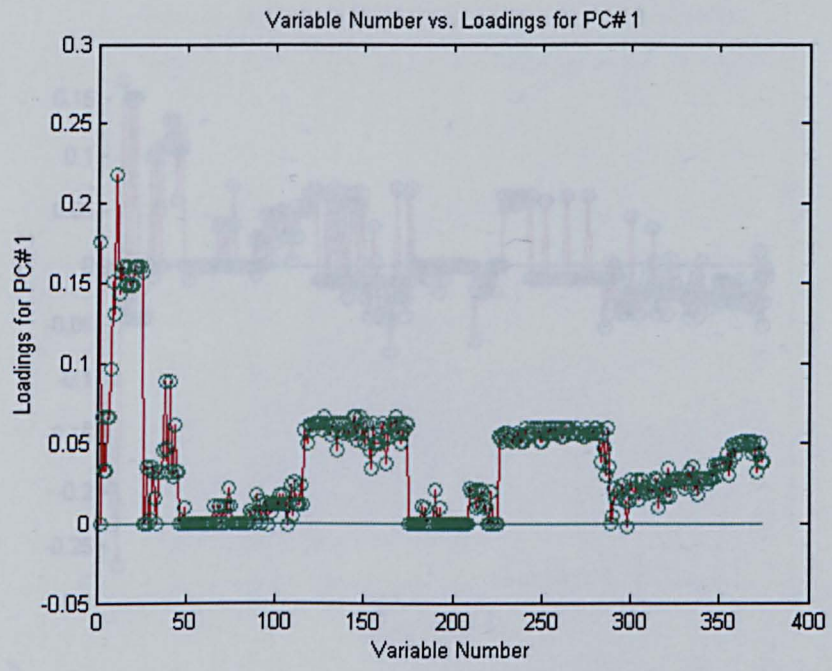


Figure 3.2.1-9 PCA results of Avecia BIT process samples spectra by GMS; PC 2 loadings

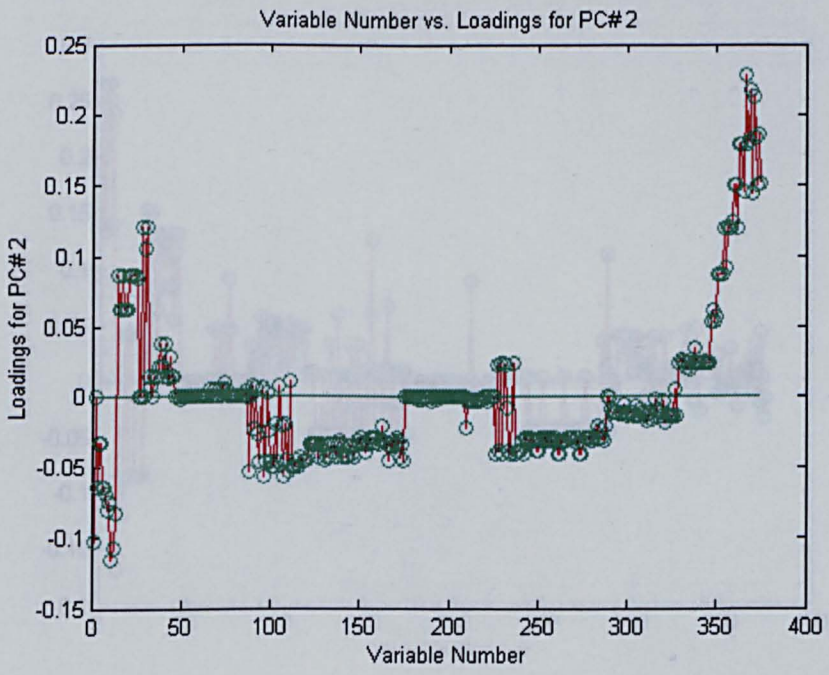


Figure 3.2.1-10 PCA results of Avecia BIT process samples spectra by GMS; PC 3 loadings

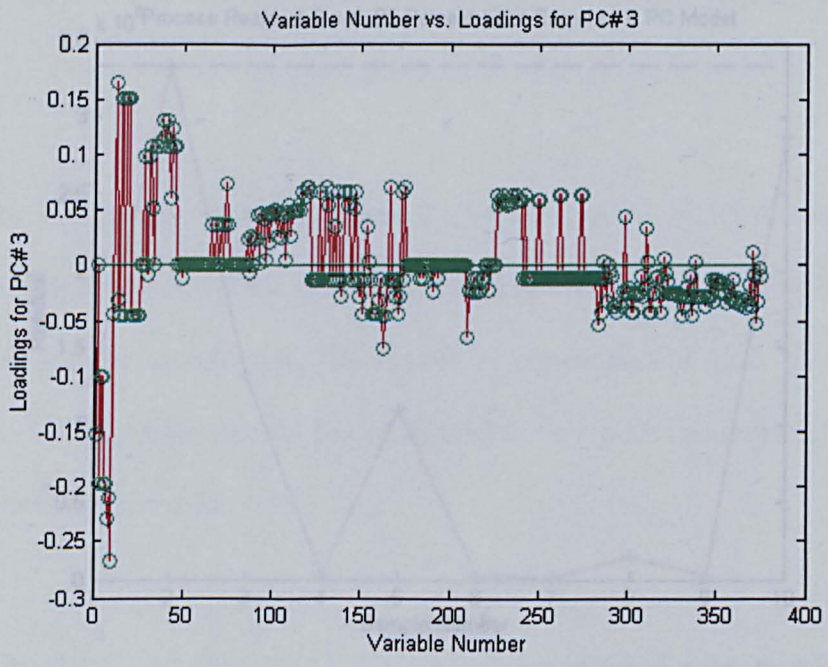


Figure 3.2.1-11 PCA results of Avecia BIT process samples spectra by GMS; PC 4 loadings

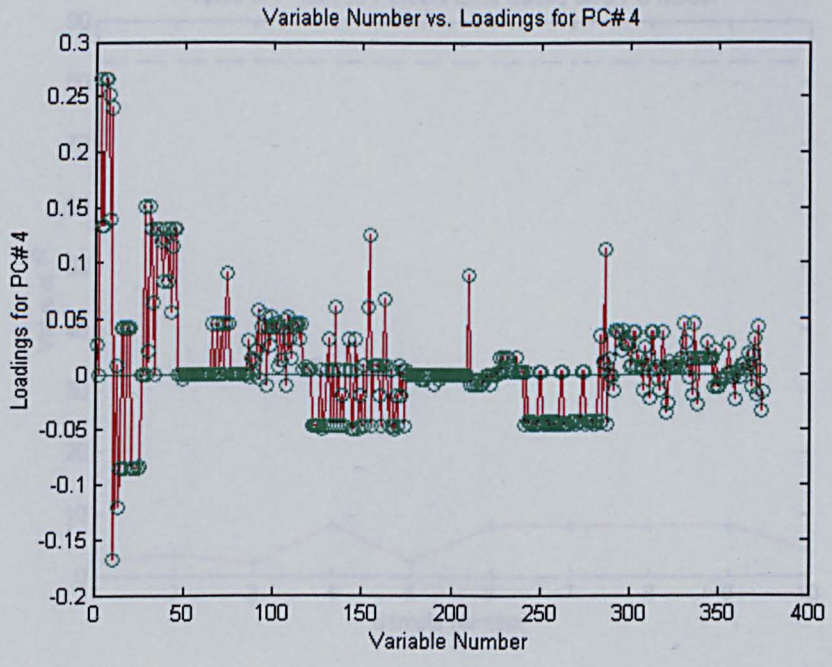


Figure 3.2.1-12 PCA results of Avecia BIT process samples spectra by GMS; Q residuals

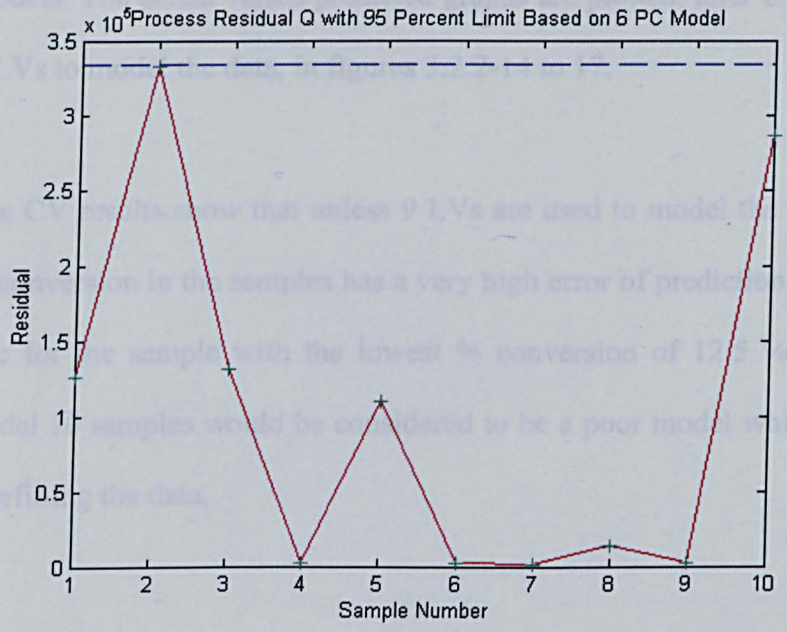
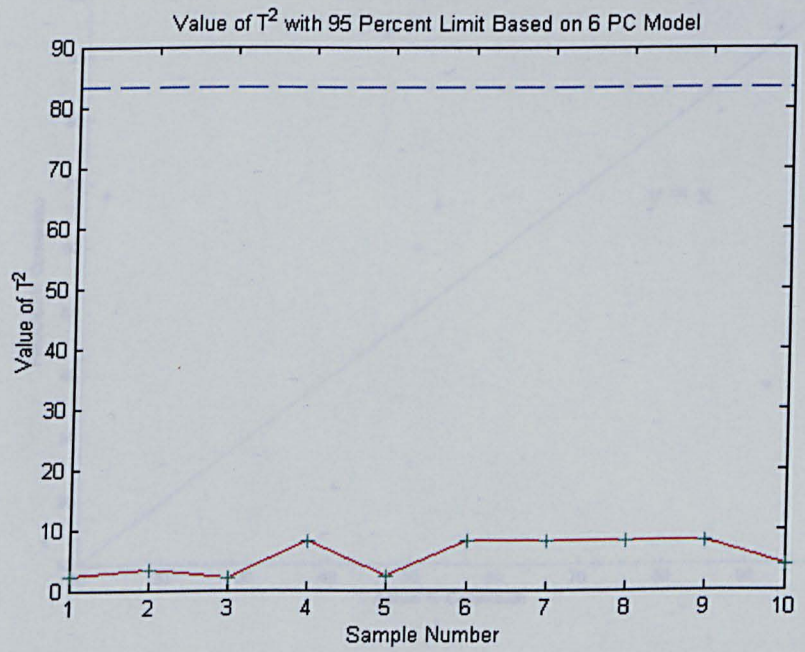


Figure 3.2.1-13 PCA results of Avecia BIT process samples spectra by GMS; T^2 values



Given the lack of correlation between the sample oxidation and the GMS spectra PLS was used to maximize the covariance in the data to produce calibration models. The actual verses predicted graphs are plotted, after CV using 3, 5, 7 and 9 LVs to model the data, in figures 3.2.2-14 to 17.

The CV results show that unless 9 LVs are used to model the data, the predicted % conversion in the samples has a very high error of prediction. This is especially true for the sample with the lowest % conversion of 12.5 %. Using 9 LVs to model 10 samples would be considered to be a poor model which is suspected of overfitting the data.

Figure 3.2.1-14 Plot of PLS CV actual vs predicted oxidation in samples using 3 LVs to model the data.

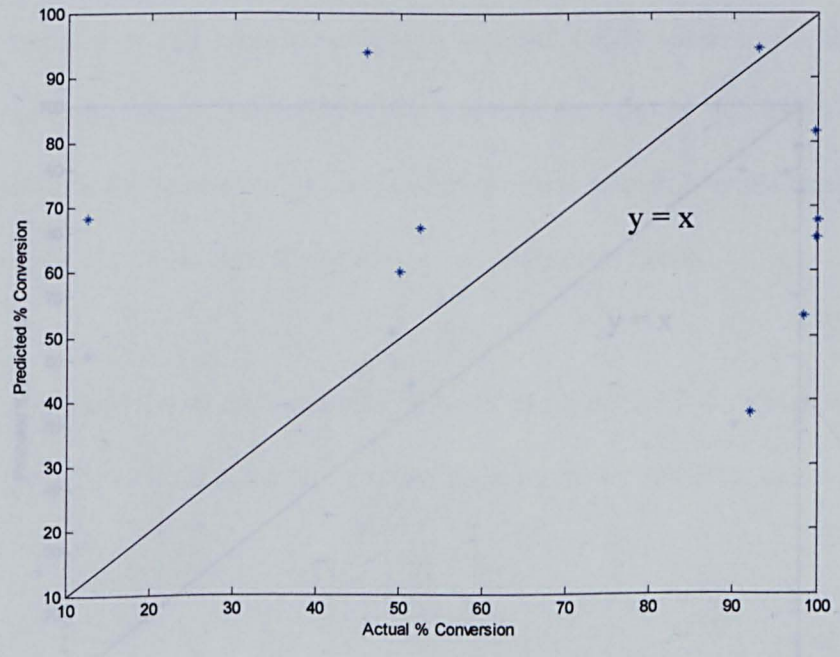


Figure 3.2.1-15 Plot of PLS CV actual vs predicted oxidation in samples using 5 LVs to model the data.

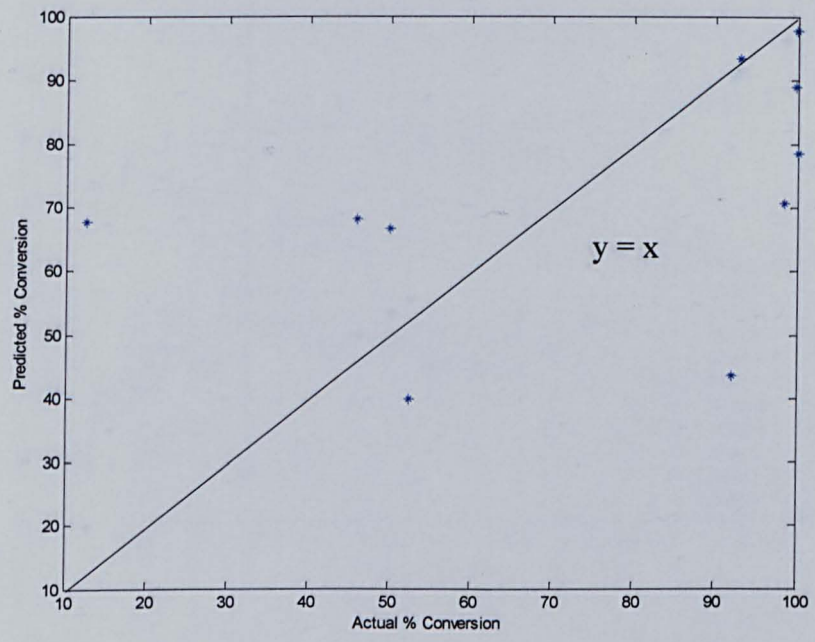


Figure 3.2.1-16 Plot of PLS CV actual vs predicted oxidation in samples using 7 LVs to model the data.

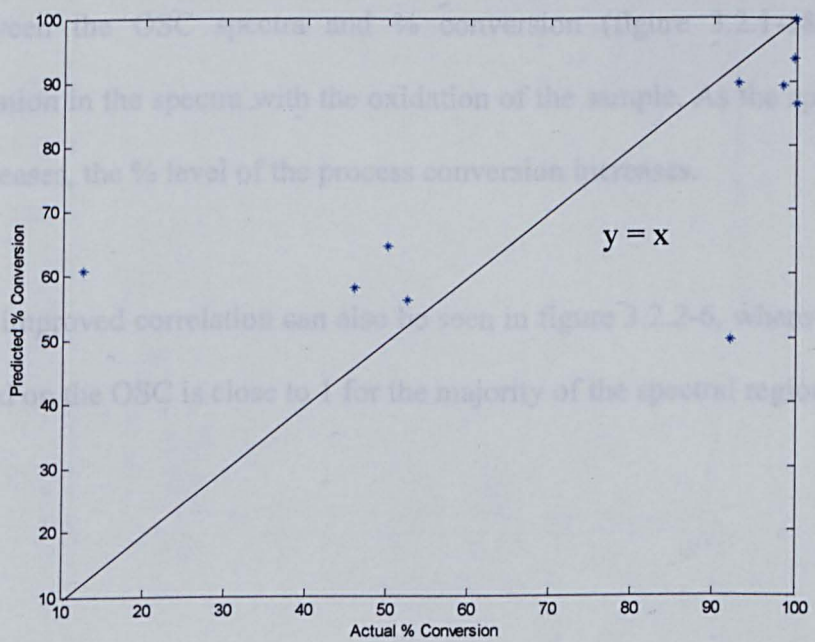
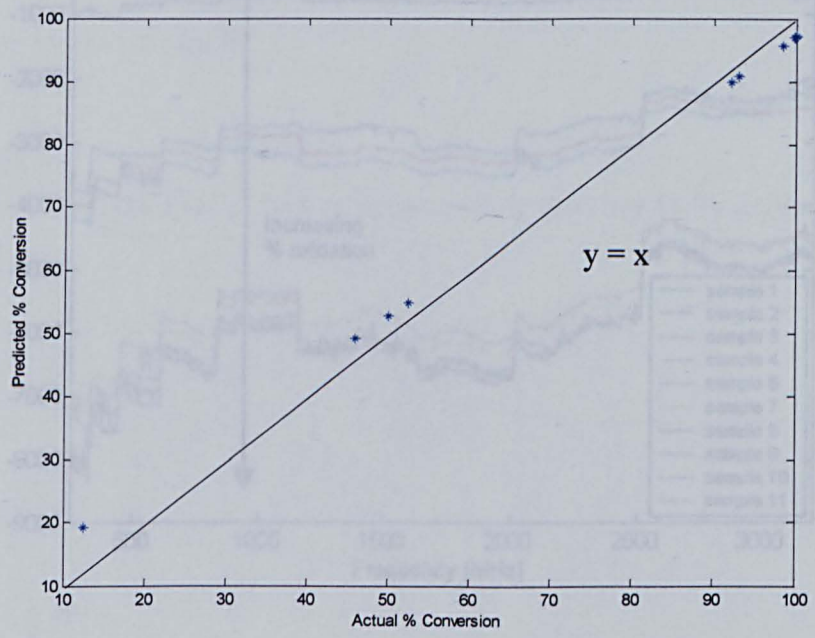


Figure 3.2.1-17 Plot of PLS CV actual vs predicted oxidation in samples using 9 LVs to model the data.



To improve the calibration model the method orthogonal signal correction (OSC) was applied to the sample spectra. The OSC GMS spectra and the correlation between the OSC spectra and % conversion (figure 3.2.1-18) shows clear variation in the spectra with the oxidation of the sample. As the spectral response decreases, the % level of the process conversion increases.

The improved correlation can also be seen in figure 3.2.2-6, where the correlation based on the OSC is close to 1 for the majority of the spectral region.

Figure 3.2.1-18 OSC oxidation samples spectra

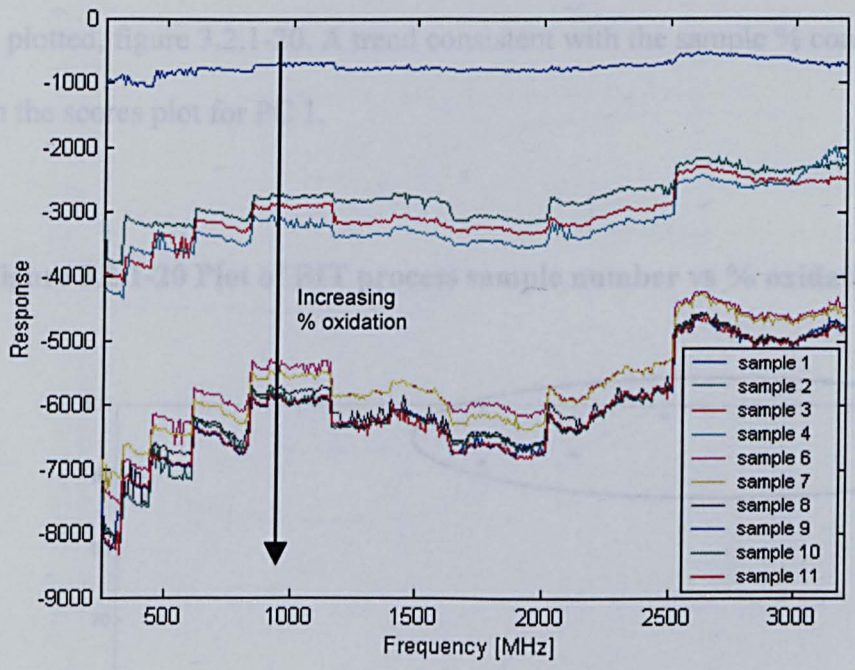
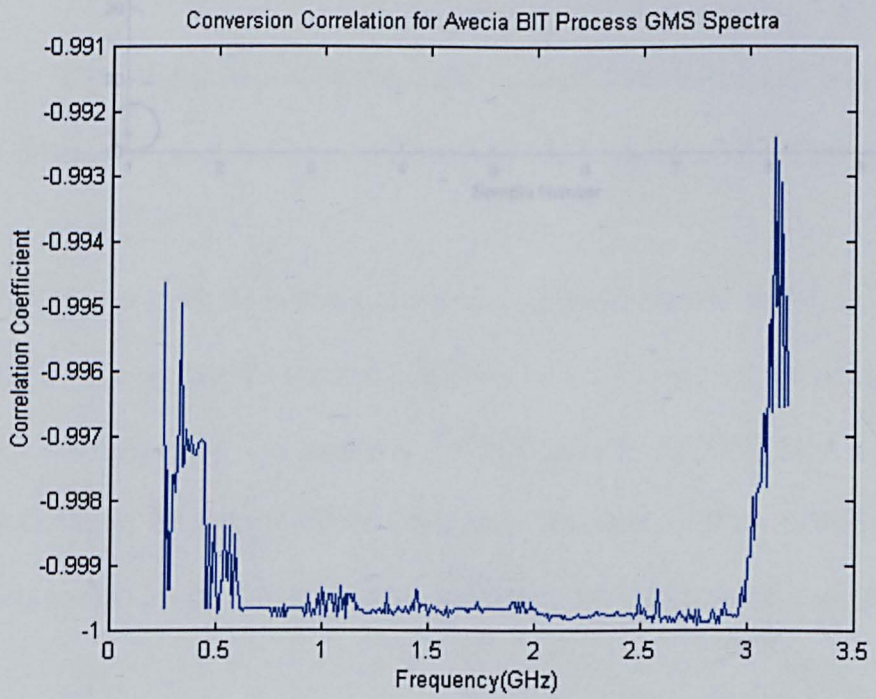


Figure 3.2.1-19 Correlation between OSC spectra and oxidation % in samples



The application of OSC has increased the correlation between the sample oxidation and spectra. PCA was applied to the OSC spectra and the scores for PC 1 plotted, figure 3.2.1-20. A trend consistent with the sample % conversion is seen in the scores plot for PC 1.

Figure 3.2.1-20 Plot of BIT process sample number vs % oxidation

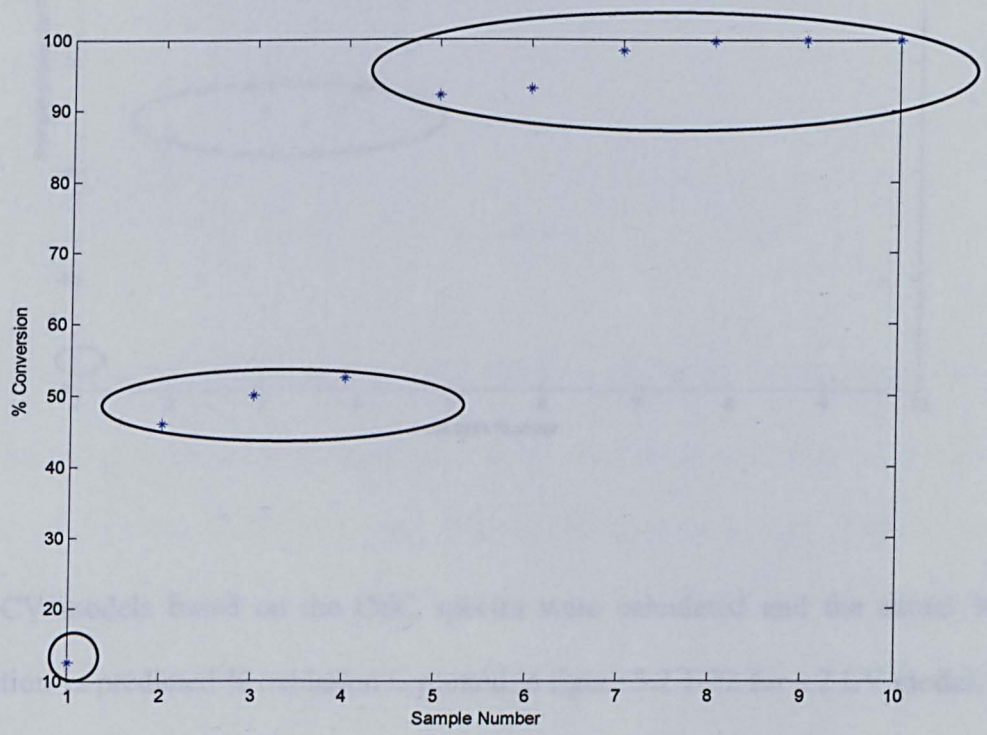
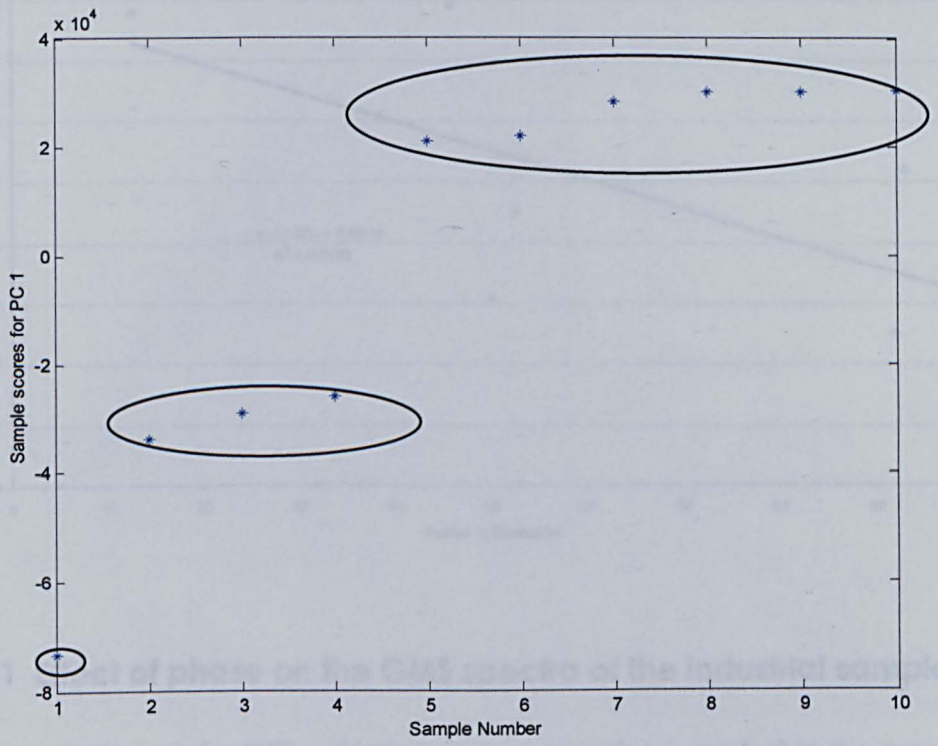


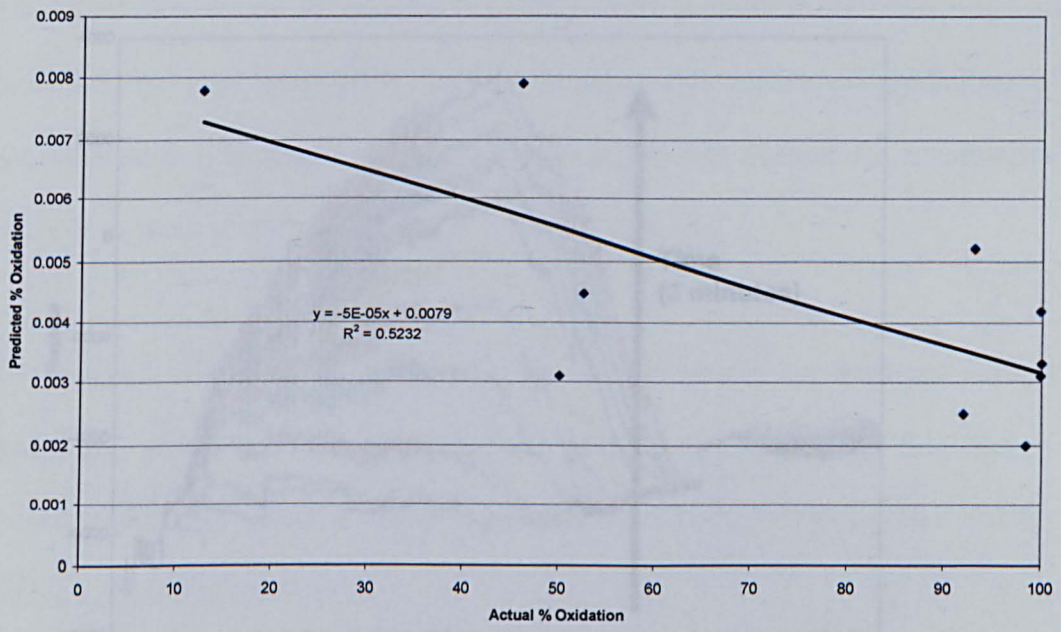
Figure 3.2.1-21 Plots of PCA scores plot for PC1 of OSC oxidation samples spectra



PLS CV models based on the OSC spectra were calculated and the actual % oxidation vs predicted % oxidation is plotted in figure 3.2.1-22 for a 2 LV model.

The poor correlation between the actual and predicted % oxidation, seen in figure 3.2.2-7 shows that the method could not be used to predict the oxidation levels of unknown samples. The trends in the OSC spectra and OSC PCA scores plots are over fitting the data by using OSC and consequently PLS models based on this data can not be used to predict the oxidation levels of unknown samples.

Figure 3.2.1-22 Actual Vs predicted % oxidation of OSC industrial samples



3.2.1.1 Effect of phase on the GMS spectra of the industrial samples

The ground state of the BIT oxidation process samples is such that the samples consist of 3 phases, solid precipitate, organic and aqueous phases. When mixed this forms a pale brown slurry. An investigation of the effect of the sample phase on the GMS spectra was carried out. The sample was thoroughly mixed then left to separate back into its various phases for 3 minutes whilst the GMS spectra were recorded. The spectra and PC 1 scores are plotted in figure 3.2.1.1-1.

As the sample separates back into its original phases the spectra change from an almost flat line to increase in response with a peak maxima of 3500 units The scores on PC1 show the sample is separating linearly with time, demonstrating that the effect of dielectric constant and sample phase is important and that GMS can be used to determine sample homogeneity.

Figure 3.2.1.1-1 Effect of sample homogeneity on oxidation samples by GMS

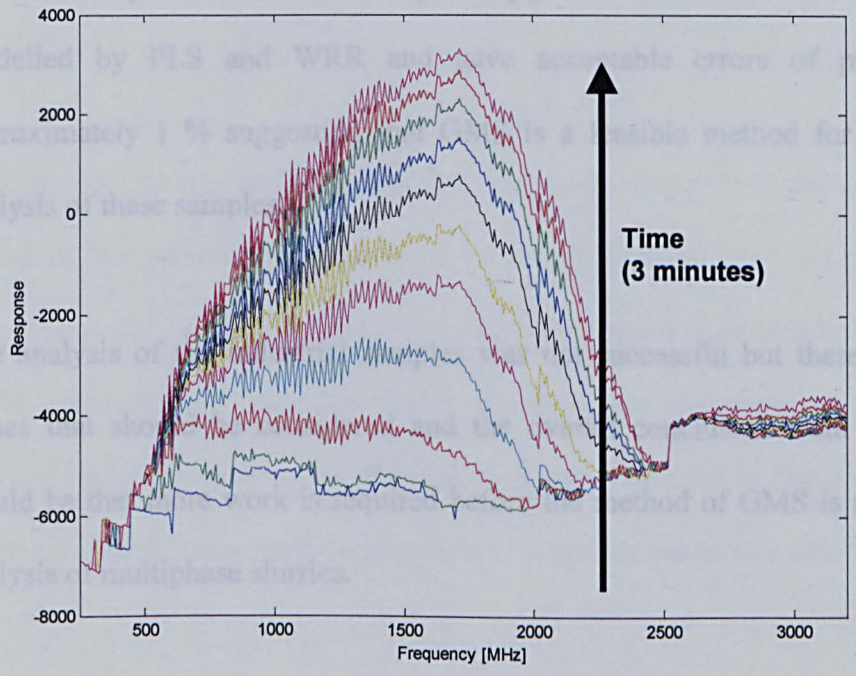
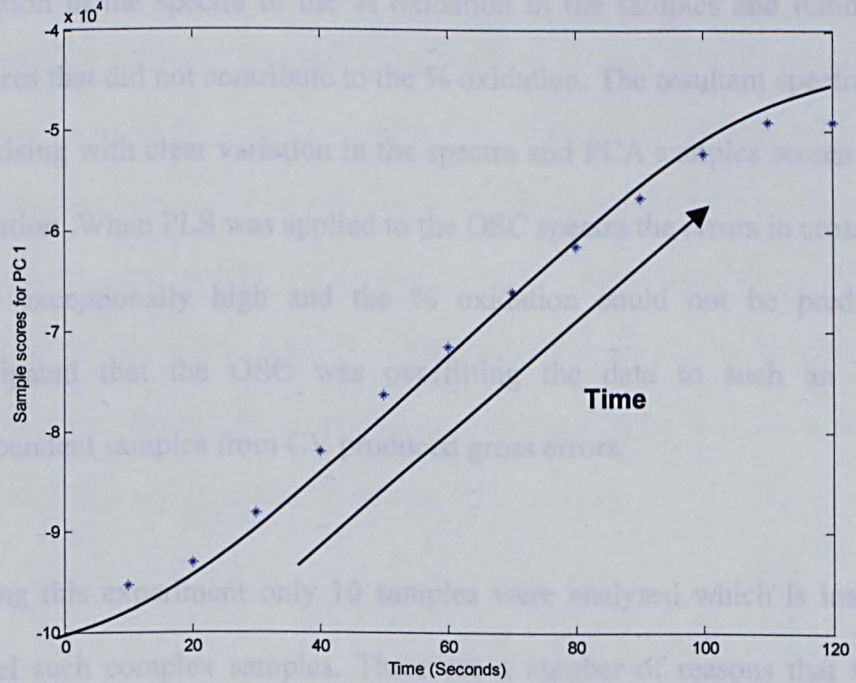


Figure 3.2.1.1-1 PCA scores for PC 1 of sample homogeneity experiment spectra by GMS



4 Conclusions

The initial experiment showed that partially soluble acetonitrile in water could be modelled by PLS and WRR and gave acceptable errors of prediction of approximately 1 % suggesting that GMS is a feasible method for quantitative analysis of these samples.

The analysis of the industrial samples was not successful but there are several issues that should be considered and the overall conclusion from this section would be that more work is required before the method of GMS is ruled out for analysis of multiphase slurries.

When the industrial samples were measured the GMS response was very low and the variation between samples difficult to visualise in the spectra and PCA scores plots. Orthogonal signal correction was applied to the spectra to maximise the variation in the spectra to the % oxidation in the samples and remove spectral features that did not contribute to the % oxidation. The resultant spectra were very promising with clear variation in the spectra and PCA samples scores with the % oxidation. When PLS was applied to the OSC spectra the errors in cross validation were exceptionally high and the % oxidation could not be predicted. It is anticipated that the OSC was overfitting the data to such an extent that independent samples from CV produced gross errors.

During this experiment only 10 samples were analysed which is insufficient to model such complex samples. There are a number of reasons that would have

affected the quality of the measurements recorded during this experiment which could lead to greater variability between the spectra than the % oxidation of the samples.

The sample volume analysed was not accurately controlled, the samples were stored in jars prior to analysis, these were approximately the same volume but previous work[5] had shown that the weight of samples is an important factor for GMS measurements. Another factor which could contribute to uncontrolled spectral variation between the samples is the ratio of solid to organic to aqueous phases. As these ratios vary the sample's dielectric constant will also vary.

For the measurements it could be that the variation in the GMS spectra due to these factors is greater than that for the % oxidation leading to the inability to model this data.

The method did show that there was potential for this type of spectra to be used for the analysis of oxidation level but further experiments with controlled, designed samples would be required to confirm suitability for on-line application. The results were recorded as the sample phase varied from slurry to the original separate phases highlighted the GMS system's sensitivity to changes in the sample homogeneity. When analysing multiphase systems it is essential that the sample is thoroughly mixed for comparative analysis.

5 Further Work

The following experiments are proposed to establish if the industrial process samples from GMS may measure Avecia's BIT oxidation process and also to gain more information where the boundaries are for measurements of complex samples.

- To repeat the experiments of chapter 5 but use nitrile mixtures of multiple components to see if mixtures of 4 nitriles are possible where 4 alcohols were not

- Repeat the previous experiments but with much greater control from preparation and planning. Samples should be taken regularly from one batch during the process to gain a large number of samples and an even spread of % oxidation of the samples. Fixed volume portions should be transferred to cylinders and the % of each phase noted as accurately as possible. The fixed volume samples should then be weighed prior to GMS measurement as this is likely to vary with different ratios of solid, liquid and aqueous layers.

- If the data cannot be modelled in the above experiment a series of simple multiphase samples, following an experimental design approach should be analysed by GMS.

6 References

[1] Meeting minutes, Avecia Huddersfield works, 28/6/01: 'Kick-off feasibility study for on-line reaction monitoring of oxidation stage of BIT process by microwave spectroscopy', Avecia internal document, Personal communication (E-mail): E. Polwart, Avecia Grangemouth Works, 5/7/01.

[2] Walmsley A.D, Loades V.C, 'The determination of acetonitrile in water and ethanol in water by guided microwave spectroscopy with multivariate calibration', The Analyst, 2001, 417 - 420.

[3] 'Experimental Risk Assessment Proforma', Experiment: Investigate guided microwave spectroscopy of BIT process samples, Avecia internal document, Personal communication (E-mail): E. Polwart, Avecia Grangemouth Works, 8/8/01.

[4] Personal communication (E-mail), B. Heap, Avecia, Huddersfield Works, 6/4/02.

[5] Dane A. D, Rea G. J, Walmsley A. D, Haswell S. J, 'The Determination of Moisture in Tobacco By Guided Microwave Spectroscopy and Multivariate Calibration', Analytica Chimica Acta, 2001, 429, 185 – 194.

Chapter 7

Conclusions

The application of chemometrics to spectroscopic process analytical data has been investigated. There have been several interesting discoveries that will impact on the manufacturing environment for new and alternative methods of analysis and in the field of chemometrics were the urgent need for attention of application was identified.

The chemical manufacturing industry has a great interest in process analysis and when applied effectively can bring great financial and environmental benefits. The ideals of process analysis are to have a very simple measurement system for what is often a very complex mixture. The accuracy and precision of laboratory based measurements systems can rarely be reproduced in the process environment of simple systems. The uptake of process analysis universally has been restricted because analytical measurements are not always easy to transfer to a manufacturing environment. Equipment has to be able to stand up to the harsh manufacturing environment yet still make sensible measurements.

When applied effectively the benefits of process analysis can lead to increased knowledge of the process in real-time can lead to substantial cost benefits. The measurements taken are often complex with many interferrants, multivariate analysis methods are used to overcome these issues.

The first example of this was the multivariate calibration of spectra of uranyl nitrate liquors. The aim was to establish if either UV/Vis or Raman spectroscopies were suitable for the analysis of uranium ore conversion process. The company do

not have large funds for the project and any on-line analysis technique would have to be at minimal expense. This made the application of a Raman instrument highly unlikely in reality. The feasibility work was still undertaken as an instrument was easily borrowed from the manufacturer and if successful would imply that other similar processes which have more resources maybe analysed by this method.

The first stage of the uranium ore concentrate conversion process was simulated in the laboratory at BNFL. The Raman and UV/Vis spectra were used to monitor the process. The implementation of an on-line analytical solution would result in improved process control and downstream yield improvements. The aim of the work undertaken was to establish if either technique could be used to predict the amount of uranyl and nitrate in the samples and the temperature of the samples. The objective of using a single calibration model for simultaneous prediction of all components was achieved. A PLS2 calibration model based on first order derivatised spectra, predicted the levels of nitrate and uranyl with a 7 % error and the sample temperature with 5 % error. The work found that the inexpensive UV/Vis measurements were not sufficient for the process data even though literature had shown measurement to be possible in simple solutions. The Raman technique was suitable for measurement but, is too costly to install and maintain. Ideally a method between these in terms of sensitivity and cost would be beneficial.

The application of spectroscopic measurement to industrial process analysis is the main feature for the remaining work. Existing methods of process analysis cover a

wide range of applications satisfactorily. There is one area where there is little representation for potential methods of analysis. This is the analysis of samples which contain a high proportion of solid particulates. The problem of analysing this type of sample is currently solved by over engineering of the process or the samples to allow analysis of the solution without particulates present. An example of this is the NIR instruments developed by Foss for fermentation monitoring, these are effective instruments for measurement, but require sample pre-treatment. Here the method of guided microwave spectroscopy has been used to analysis particulate containing samples in entirety, without separation.

The of advantage microwaves over spectroscopies such as NIR is that they have long pathlengths of many cm. The result is large sample chambers of several hundred ml. There are several advantages of this technique that are a result of the increased sample volume, as it improves the representation of the sample from the process stream and are less likely to suffer fouling.

Guided microwave spectroscopy is a relatively new and under utilised method for process analysis. Its application has been limited due to the very complex, broadband spectra that are produced from measurement of non-gaseous samples. The complex spectra are the disadvantage of the method. To understand the signal chemometric techniques are essential. The spectra are also difficult to predict. This could limit the range of applications to on-line environments, especially in the field of pharmaceutical analysis. The pharmaceutical industry is heavily

regulated; a technique where the composition of the spectral response is difficult to prove is unlikely to be well received.

The microwave response from water is very strong, as the technique is sensitive to variations in hydrogen bonding as the dielectric constant of a material will alter with hydrogen bonding. The issue of non-linearity of hydrogen bonding is been demonstrated in the development stage of this method for chemical analysis. For the analysis of samples with 1 or 2 alcohols standard regression methods were sufficient for calibration, when this was extended to 3 polynomial methods were required to account for the non-linearities. These could not be overcome for accurate prediction when there were 4 alcohols in the mixtures. This is a low number of components for a measurement system and may indicate why the method has not been widely received in an industrial environment.

The GMS has been shown as suitable for the monitoring of the beer fermentation process. The advantages of the GMS system of were evident in that 1 pint of beer was fermented directly inside the GMS sample cavity, the entire fermentation wort was measured included dissolved gases and particulates and the fermentation could still be monitored.

The way this experiment was undertaken could be applied to an industrial environment. If the need for a guided wave cavity could be eliminated, and sufficient microwaves could be received for detection, then in principle, microwave spectroscopy could be used for analysis across process reactors,

eliminating the need for sampling or re-circulation loops. Methods of process analysis are extending to measurements through packaging. Examples include the analysis of pharmaceuticals raw materials through the packaging in which the material is transported by NIR. The packaging has been standardised by suppliers to make measurements possible. Until the microwave system can be used without the guided cavity the method will not be suitable to such applications.

Another example of where GMS has demonstrated feasibility for measurement was for the monitoring of Avecia's BIT oxidation process. The samples were non-homogeneous examples of the industrial process consisting of three separate phases; organic, aqueous and solid which formed slurry on mixing. This mixture could not be measured by NIR, the GMS spectral measurements were low in response and required an intensive form or pre-treatment for prediction. This work was inconclusive as to whether GMS was suitable for measurement. The samples contained a large number of components which could limit the GMS system response to the analyte of interest. It is in these circumstances the use of chemometrics is essential.

The measurements for the BIT process was of the entire sample, part of the problem for calibration of such samples is that the reference measurements were based only on one phase of the sample. It is expected that correlation should be straight forward between a single phase measurement and a multiphase measurement, but not surprising in reality when weaknesses occur.

To overcome some of the problems when relating the measurement of a single component to the spectra of a solution which has several factors or components varying new chemometric techniques have been developed. The relatively new technique of orthogonal signal correction can be used to clean-up spectral signals so that only information that directly relates to the analyte of interest is used in the calibration. This can be an excellent technique for data which has a poorly correlated signal. However the technique has received much criticism. There can be a tendency for too much information being removed from the signal and as a result calibration models are over-fitted and are not relevant for independent datasets.

Orthogonal signal correction has been used for poorly correlated data with exceptionally large improvements in the calibration data. This method should be used with caution as it has been shown in this thesis to overfit the calibration data to an extent that prediction is not possible after CV.

Much of the work has been involved the investigation of new methods of measurement with the view to be applied in an industrial environment for on-line process monitoring and analysis. A range chemometric techniques and pre-treatment methods have been applied to gain information, visualise trends or predict analyte properties within corresponding spectroscopic data. These have been restricted to the mainstream methods of mainly PCA and PLS to minimise queries over the chemometric technique and focus on development of the new measurement technique and application.

The calibration method of WRR has also been applied for model generation in some datasets during this work. The method has shown to be comparable to PLS models. This method has the advantage that it does not decompose the data into factors which then have to be recombined using a fewer number of factors.

The procedure of using a standard reference material for chemical analysis to ensure a measurement method is correct is widely used by analytical chemists, and is considered an integral part of 'Good Laboratory Practice'. Yet this approach has not been standardised for the application of chemometrics. For industrial analysis, there is little point of time and cost consuming instrumental calibration procedures, if this approach is not then undertaken for the computations applied to the measurements. This gap in analytical procedures was noticed by NIST, which lead to the development of a range of reference datasets, available for analysts to ensure their skills and software are in order.

A spectroscopic dataset of multiple wavelengths was not included in the work at NIST. For this thesis, a spectroscopic dataset was produced to which could provide the need for this type of dataset for reference analysis. The measurements were to be simple, easy to measure with low experimental and instrumental errors. To facilitate this, metal ion solutions of composition from an experimental design methodology were measured by visible spectroscopy. In order to gain typical errors of prediction of the samples a collaborative trial was undertaken. From this the mean and standard deviation of the samples predictions were estimated from

the results of a number of individuals. The spectra of the metal ions and the sample estimates of concentration will be available to those that wish to use it.

The procedure of chemometric data analysis is one that is heavily influenced by individuals' choice. There are not any rules which demonstrate that a certain technique should be used for a specific circumstance or type of spectroscopic data. This issue was highlighted in the large range of calibration methods and pre-processing techniques that was applied to the visible spectra of the metal ion complexes by participants of the collaborative trial. The trial found that even where the participants had reported using the same procedure the results were not the same. This between-user variation was also seen in a similar trial based on NIR forage data. The results of these two studies have found that further investigations are necessary to establish the cause and minimise the between-user variation. Based on the results of the collaborative study of the metal ions, an individual could conclude that the measurements were incorrect, when in reality the data processing was incorrect. This is a serious issue for the past and future application of chemometrics, as it is difficult for the subject to be embraced if there is doubt over the quality of the applications. Currently there is a large amount of resources given to the development of new chemometric techniques. For transfer of chemometrics to the industrial environment the focus should be re-direct from new techniques, sometimes with only marginal benefits, to ensuring the existing techniques are applied correctly.

Overall this work has successfully shown some new applications of standard measurement systems to produce a standard dataset and monitor uranyl nitrate samples. The initial developments have been undertaken for a new instrumental measurement method for process analysis which has promising results and in the future will have a large impact on measurement of industrial processes such as moisture measurement, drying of pharmaceutical products and analysis of heterogeneous samples.

Appendix

Published paper

'Determination of Acetonitrile and Ethanol in Water By Guided Microwave Spectroscopy With Multivariate Calibration'

Determination of Acetonitrile and Ethanol in Water By Guided Microwave Spectroscopy With Multivariate Calibration

Anthony D. Walmsley* and Victoria C. Loades

* Corresponding Author
E-mail address: a.d.walmsley@chem.hull.ac.uk
Tel: 01482 465470
Fax: 01482 466416

Department of Chemistry, Faculty of Science and the Environment, University of Hull, Cottingham Road, Hull, HU6 7RX, UK.

Keywords: Guided Microwave Spectroscopy, Multivariate Calibration, Chemometrics, Partial Least Squares, Weigthed Ridge Regression.

Justification

The current growing interest in process analysis is quickly moving away from traditional methods of analysis. As a result there is considerable interest in alternative methods, such as Microwave. Traditionally the problem with Microwaves has been in the data analyses. This paper investigates the application of chemometric modeling and demonstrates that it is feasible to use microwaves for calibration.

Abstract

The feasibility of using guided microwave spectroscopy (GMS) utilizing the frequency range 0.25 – 3.20 GHz, combined with multivariate calibration for the determination of acetonitrile or ethanol concentration in water. A wide range of different concentrations was used (up to 30 % v/v). Partial Least squares (PLS) and Weighted Ridge Regression (WRR) was applied to generate a model for prediction, based upon the microwave spectra.

A high level of collinearity was observed in both of the sample data sets and this was reduced by background subtraction. The prediction ability for the two types of regression models were found to be comparable with the percentage error of prediction (PEP) being approximately 2.5 % for the acetonitrile samples and 1.1 % for ethanol samples.

Introduction

The aim of this work is to test the feasibility of Guided Microwave Spectroscopy (GMS) using 0.25 – 3.20 GHz frequency range for the analysis of solvents in water [1]. The two solvents under investigation are acetonitrile and ethanol. These have significantly different dielectric constants (table 1) [2].

Classically solvent concentrations can be determined by gas chromatography [3]. This can be lengthy procedure requiring frequent instrument calibration. So far its application for in-line analysis has been problematic and unreliable.

The overall aim is to apply GMS for in-line analysis during production. Microwave spectroscopy has often been ignored as a suitable in-line analysis method because the spectra are often broadband without clear peaks. This can make understanding and using the spectra for the prediction of concentration difficult [4]. Here chemometric techniques have been applied to enable prediction of concentration from the solvent spectra.

Microwave spectroscopy will respond well to solutions with a large dipole moment and that are hydrogen bonded i.e. ideal for water. The form of the water to be analysed will affect the microwave absorption characteristics [4].

GMS

The GMS instrument will give a spectral response to the change of ε' and ε'' as microwave radiation passes through a sample. Changes in the dielectric constant, ε' are caused by a reduction of wave velocity across the chamber during analysis.

$$\varepsilon' = c^2/v^2 \qquad [1]$$

Where:

ε' = Dielectric Constant.
c = Velocity of light in a vacuum.
v = Velocity through sample.

Variations in ε'' are a result of energy loss to heat due to friction as molecules orientate in the microwave field [5].

Data Analysis

The data analysis was carried out using Matlab™ version 5.3.1[6] and the PLS_Toolbox version 2.0 [7].

The Condition Number is a measure of the quality of the data. A condition number above 30 is indicative of collinearity within the data [8]. Collinearity is a feature of the

data matrix and will significantly affect the least squares efficiency in a detrimental manner. It is determined by ratio of the maximum and minimum values after singular value decomposition.

$$ConditionNumber = \frac{MaxSingularValue}{MinSingularValue} \qquad [2]$$

Multivariate Calibration

Multivariate calibration can be described as the modeling of data that has multiple measurements for a number of samples. Here the commonly applied PLS-NIPALS algorithm and the less frequently referenced Weighted Ridge Regression methods are used to model the multivariate spectral data. The aim of the modeling is to produce the best predictive model.

Partial Least Squares (PLS)

The PLS algorithm works by decomposing the spectral data and relating concentrations into latent variables [9]. The number of latent variables is chosen which result in the lowest Percentage Error of Prediction of the validation data (equation 5).

Weighted Ridge Regression

Ridge Regression (RR) is another modeling technique. The method is based on correlations within the data. It works by adding a value (θ) to the ridge (or diagonal) of the correlation matrix [10].

$$b_F(\theta) = (F'F + \theta I_r)^{-1} F'Y \qquad [3]$$

Where:
b_F are the regression coefficients,
θ = Ridge Constant (positive value between 0-1),
F = mean centred X matrix,
$F'F$ is the correlation matrix,
I is the identity matrix, size r x r (r = number of data points/wavelengths).

This has the effect of maximizing the variation and orthogonality of the data. A benefit is that the procedure can improve the signal to noise ratio of the spectra. Ridge Regression is ideal for ill-conditioned/collinear data where $X'X$ (inverse matrix) is near to or actually singular. Under these conditions, problems are incurred when calculating models by PLS. Another advantage of the Ridge Regression over PLS is that it does not

decompose the data into latent variables and hence removes the issue of which latent variables to keep when modeling data.

In this paper calculations were made using the improved method of Weighted Ridge Regression (WRR), (equation 4) instead of the standard RR described above [11].

$$b_F = (F'F + \theta \times diag(F'F))^{-1} F'Y \qquad [4]$$

In WRR it is not necessary to calculate the values of θ allowing the regression coefficients to be computed more quickly.

Data Pre-treatment

For each sample the spectra is scanned and recorded 10 times whilst situated in the microwave cavity. Of these 10 spectra the median is taken and used to represent the microwave spectra of that sample.

For PLS the solvent spectra are background subtracted. To do this the spectra of pure water without any solvent added is subtracted from those with solvent. This maximizes the variation in the spectra according to solvent concentration. The spectra were not background subtracted before WRR.

Cross Validation

The calibration models are validated using the Leave-One-Out cross validation method. For this a sample is sequentially removed from the calibration data set, the model is generated using the reduced sized data set. Then the validation data spectra concentrations are predicted using the new model. This procedure is repeated until all of the calibration samples have been removed once.

In order to test the goodness of fit of the prediction model outputs the average Percentage Errors of Prediction (PEP) are calculated (equation 5).

$$PEP = \left\{ \frac{\sum \sqrt{\left(y_{actual} - y_{predicted}\right)^2 / y_{actual}}}{n} \right\} \times 100 \qquad [5]$$

Experimental

Instrumentation

The instrument used throughout this project was an Epsilon Industrial Guided Microwave Spectrometer (GMS), Epsilon Industrial Inc, 2215 Grand Avenue Parkway, Austin, Texas, 78728. The GMS has a Bandwidth of 0.25 – 3.20 GHz and dielectric range of 1 - 85.

The GMS has a remote cavity where the sample spectra are recorded. The cavity is of stainless steel construction and has internal dimensions of 10.0×4.7×11.5 cm, with a total internal volume of 540 cm^3. This is connected to the GMS via two coaxial cables, which were 450 cm. The microwaves are transmitted along a cable to the cavity and pass through the sample to the detector. Here the response is passed along the remaining cable back to the spectrometer. A PC connected to the instrument records the spectra using Linefit Software (Version 1.43), Epsilon Industrial Inc., and then the data is transferred to Matlab™.

Reagents

For each solvent 500 ml standards were made up volumetrically in duplicate. The acetonitrile (acetonitrile, HPLC Grade, 99 %, Fisher Chemicals, Loughborough, UK) standards were of the following concentrations; 5, 8, 10, 12, 15, 20, 23, 25, and 30 % v/v. The ethanol (ethanol, Absolute, 99 %, Fisher Chemicals, Loughborough, UK) samples were of a wider range of concentrations, these were; 1, 3, 5, 8, 10, 12, 15, 20, 23, 25, 28 and 30 % v/v.

Procedure

Background spectra of water were recorded prior to analysis. The microwave cavity was rinsed with water (500 ml) and then filled with the 500 ml sample standard. This was left for 1 minute to stand before 10 spectra were recorded consecutively. After which the cavity was emptied of standard. The cavity was then rinsed with water and the procedure repeated.

Results and Discussion

The microwave spectra of acetonitrile samples can be seen in figure 1a and ethanol samples, figure 2a. Changes in the ethanol spectra with concentration can be clearly seen, the variation in spectral response for acetonitrile samples is slight and only distinguishable after background subtraction, figure 1b. The differences in responses are due to the form of water after it is mixed with the different solvents.

The condition number of the acetonitrile and ethanol samples is 2400 and 2000 respectively. Pre-treatment techniques were applied to the spectra and the effect on condition number noted. It was found that background subtraction significantly reduces the condition numbers to below 1000 whilst other methods such as mean centering, autoscaling and smoothing had a detrimental effect and increased the condition number by several orders of magnitude. Thus prior to PLS the spectra are background subtracted.

From PLS the samples scores plots of PC1 against PC2 (figures 3a and 3b) show that for both sets of solvent standards the sample scores are ordered in terms of concentration and lie on a curve. When the training and validation spectra are combined the scores positions for each concentration are clustered together demonstrating reproducibility of the system.

The PEP from Leave-One-Out cross validation of the sample spectra using PLS and WRR modeling gave comparable predictions for each sample set (table 2). The ethanol samples had a lower PEP than the acetonitrile samples and a correlation coefficient (r^2) closer to 1.

As previously mentioned the GMS spectral variation for ethanol is significantly more than that of acetonitrile. The increased response would result in the ethanol samples having a reduced level of noise, improving condition and therefore having lower errors of prediction. The difference in signal to noise ratios for PLS modeling are distinguishable in the regression vectors plots. These show the spectral regions that contribute to the model. When compared the acetonitrile samples (figure 4a) model has fewer peaks in comparison to those of ethanol samples (figure 4b).

This work has shown that GMS over the 0.25 – 3.20 GHz region is a suitable for the analysis of acetonitrile and ethanol in water up to 30 % v/v of analyte when PLS and WRR are employed for modeling of the data to enable prediction of concentration. The procedure worked with errors of less than 2.5 % error of prediction. Typically a sample of 15 % v/v acetonitrile would be predicted as in the region of 14.6 – 15.4 % (taking 2.5 % PEP) and an ethanol sample between 14.8 – 15.2 % (1.1 % PEP).

This paper only describes the initial results for basic system of solutions of one analyte in water. It has been applied to solvents of very different structures. Due to its strong polarity water has the strongest response to microwaves. It is expected that the ethanol solutions gave a greater response to the microwave field because of the ability to hydrogen bond and therefore increased mobility in water allowing the molecules to easily rotate in the microwave field. The acetonitrile standards had a poor response over much

of the frequency range. This could be a result of the overall polarity of the solution decreasing as the acetonitrile levels increase.

A natural extension will be to apply GMS with multivariate calibration to solutions of different analytes or multiple components. Further work should be carried out to determine the limits of detection of analyte by this method.

Acknowledgements

This work was funded by the EPSRC. We would like to thank VG Gas for supplying GMS used through out this work.

References

1 Adrie D.Dane, Gerard J. Rea, Anthony D. Walmsley and Stephen J. Haswell, Anal. Chim. Acta, 2001, **429**, 187-196.

2 David R. Lide, Handbook of Chemistry and Physics, CRC Press, 77th edn., pp. 6-8 and 8-10.

3 Alan Townshend, Encyclopedia of Analytical Science, Academic Press, Vol. 3, 1995, pp. 1872.

4 F.McLennan and B.Kowalski, Process Analytical Chemistry, Wiley, 1st edn., 1995, ch. 10, pp. 349-350.

5 Epsilon Industrial Inc. Guided Microwave Spectroscopy Series Analyser Instruction manual. 1996.

6 The Mathworks Inc. Matlab® 5.3.1. User manuals. 1999.

7 B.M Wise and N.B.Gallagher, PLS_Toolbox for use with MATLAB™, Eigenvector Research, Inc., Manson, WA 1998.

8 David A. Belsley, Edwin Kuh, Roy E. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, Wiley, 1980, 1st edn., ch. 3, pp. 85-107.

9 P.Geladi, B.R. Kowalski, Anal. Chim. Acta, 1986, **185**, 1 - 17.

10 Norman R. Draper and Harry Smith, Applied Regression Analysis, Wiley-Interscience, 1998, 3rd edn., ch. 17, pp. 387-396.

11 S.J.Haswell and A.D. Walmsley, Anal. Chim. Acta, 1999, **400**, 399 – 412.

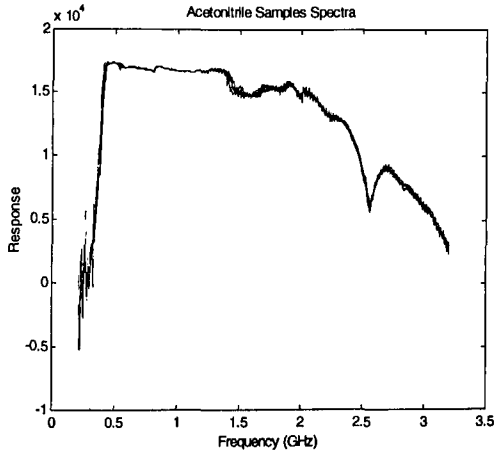
Table 1 Dielectric Constants

Substance	**Dielectric Constant**
Acetonitrile	37.5 (20°C)
Ethanol	24.3 (25°C)
Water	80.2 (20°C)

Table 2 Modeling Results After Leave-One-Out Cross Validation.

	Acetonitrile		Ethanol	
	PEP	r^2	PEP	r^2
PLS-NIPALS	2.51 %	0.9798	1.02 %	0.9991
WRR	2.41 %	0.9861	1.09 %	0.9993

Figure 1a Acetonitrile Samples Spectra.
Figure 1b Acetonitrile Samples Spectra Background Subtracted.



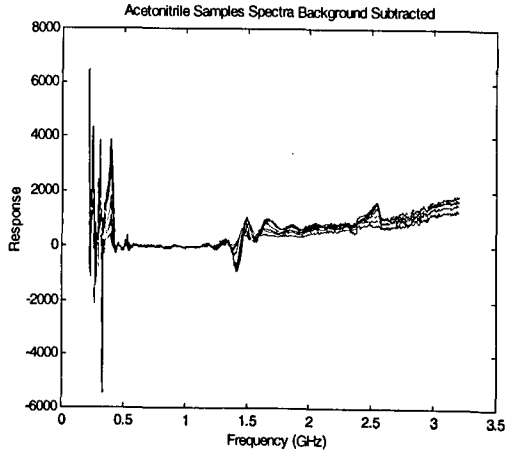
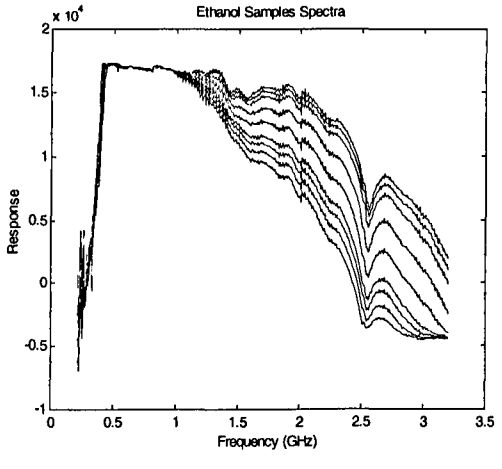


Figure 2a Ethanol Samples Spectra.
Figure 2b Ethanol Samples Spectra Background Subtracted.



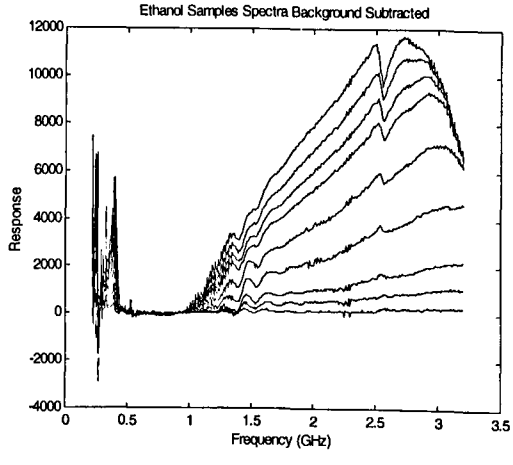
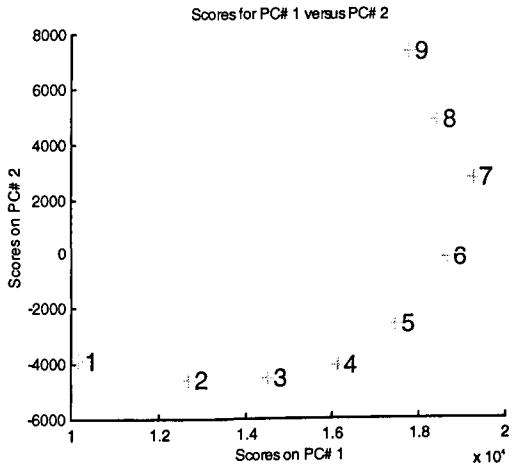


Figure 3a Acetonitrile Scores Plots of PC1 (92.62%) Verses PC2 (6.37%).
Figure 3b Ethanol Scores Plots of PC1 (98.58%) Verses PC2 (1.23%).



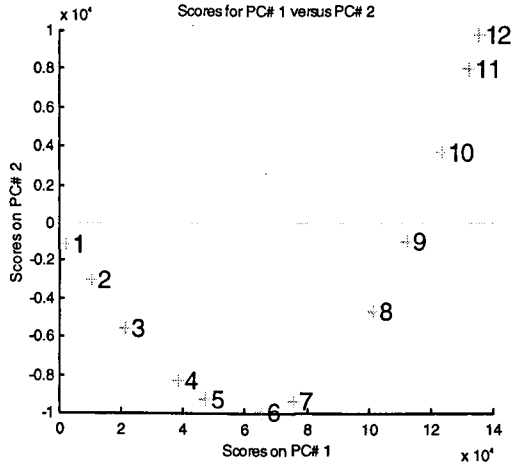


Figure 4a PLS Acetonitrile Regression Vectors.

Figure 4b PLS Ethanol Regression Vectors.