Proceedings of SPIE , the International Society for Optical Engineering, Vol. 5826, 2005.

Qualitative and quantitative analysis of chlorinated solvents using Raman spectroscopy and machine learning.

Jennifer Conroy,[a] Alan G. Ryder,[*a] Marc N. Leger,[a] Kenneth Hennessey,[b] and Michael G. Madden.[b]
[a] Department of Chemistry, and National Centre for Biomedical Engineering Sciences, National University of Ireland, Galway, Galway, Ireland;
[b] Department of Information Technology, National University of Ireland, Galway, Galway, Ireland.

ABSTRACT

The unambiguous identification and quantification of hazardous materials is of increasing importance in many sectors such as waste disposal, pharmaceutical manufacturing, and environmental protection. One particular problem in waste disposal and chemical manufacturing is the identification of solvents into chlorinated or non-chlorinated. In this work we have used Raman spectroscopy as the basis for a discrimination and quantification method for chlorinated solvents. Raman spectra of an extensive collection of solvent mixtures (200+) were collected using a JY-Horiba LabRam, infinity with a 488 nm excitation source. The solvent mixtures comprised of several chlorinated solvents: dichloromethane, chloroform, and 1,1,1-trichloroethane, mixed with solvents such as toluene, cyclohexane and/or acetone. The spectra were then analysed using a variety of chemometric techniques (Principal Component Analysis and Principal Component Regression) and machine learning (Neural Networks and Genetic Programming). In each case models were developed to identify the presence of chlorinated solvents in mixtures at levels of ~5%, to identify the type of chlorinated solvent and then to accurately quantify the amount of chlorinated solvent.

Keywords: Raman spectroscopy, hazardous materials, chlorinated solvents, non-chlorinated solvents, chemometrics, machine learning.

1. INTRODUCTION

Differentiating between chlorinated and non-chlorinated solvents is important in waste disposal, pharmaceutical manufacturing, and environmental protection. In this study particular attention is paid to that part applicable to waste disposal. Depending on whether solvent waste is chlorinated or not will determine firstly how it is transported for disposal and secondly and more importantly the disposal method to be employed. To enable this, solvent waste must be identified and labelled correctly. In a research laboratory environment solvent waste is usually separated into chlorinated and non-chlorinated streams and disposal is usually straightforward. However, the possibility of accidental contamination, or mislabelling can occur, and in these situations discrimination between the two waste streams is necessary. In many other situations unknown waste must be correctly identified before transport and disposal. In such circumstances it is desirable to have a reliable analysis method, which can first identify the unknown contents as being chlorinated or not and second to quantify the concentration present. This study investigates the possibility of using Raman spectroscopy, in combination with chemometric and machine learning techniques, to carry out these tasks.

Raman spectroscopy is based on the inelastic scattering of monochromatic light from a material.[1] When monochromatic light illuminates a material, most of the light is elastically scattered at the same wavelength (Rayleigh scattering) as the incident light. A very small portion (1 in 10^{10}) of the light, however, is inelastically scattered at a different wavelength to the incident light. This inelastic scattering is known as Raman scattering and is due to the interaction of light with the vibrational and rotational motions of the molecules within the material. Raman spectroscopy and FT-IR are related techniques with their respective spectra complimentary to one another. This is due to the different selection rules. IR

* alan.ryder@nuigalway.ie; phone 353-91-492943; fax 353-91-494596; www.nuigalway.ie/chem/AlanR/

spectroscopy is due to the changes in dipole moment during molecular vibrations, whereas Raman spectroscopy involves a change in polarizability.[1] Since the Raman spectrum depends on the chemical structure of a molecule, any given Raman spectrum is unique to the molecular structure of the material in question and can be used as a molecular fingerprint for unambiguous identification purposes. Raman spectroscopy has a number of advantages over IR. In many cases, sample preparation is often simpler or not required for Raman. Samples can also be analysed through glass or plastic containers and aqueous samples are also easily analysed because water has is a weak Raman scatterer.

The use of Raman spectroscopy as a possible forensic tool has been widely reported since the early 1990's.[2,3,4] It has proven successful in such applications as narcotic identification,[5,6] fibre identification,[7] paints/pigments analysis,[8] gunshot residue analysis,[9] gemstone identification,[10] and finally explosive identification,[11] and detection.[12,13,14] Hazardous materials such as asbestos have also been successfully analysed using Raman spectroscopy.[15]

In this study we are investigating the suitability of Raman spectroscopy for the identification and quantification of chlorinated solvents in mixtures. To do this we are using a combination of chemometric and machine learning techniques. Quantitative and qualitative analysis using Raman spectroscopy has struggled greatly in the past to gain acceptance due to a number of difficulties such as fluorescence, low signal/noise ratios and irreproducible spectral intensities. However with the use of multivariate techniques (Chemometrics), these spectral limitations may be overcome.[16,17,18,19,20] Chemometrics can be defined as the science of extracting relevant information from chemical measurements using mathematical methods.[21]

For multivariate regression, the most common and effective method to reduce the dimensionality of data is principal component analysis (PCA). PCA is used to reduce the dimensionality of a data set while retaining as much of the variation as possible in the original data. PCA factorises the Raman spectral matrix, **X**, into the form **RL**, where **L**, contains the loading vectors which are also known as factors, eigenvectors, Principal Components (PC), or latent variables. These give a new set of orthogonal basis vectors used to describe the data set. **R** is known as the scores matrix, which gives the co-ordinates of each point in **X** within the new co-ordinate system **L**. In principle, the first few principal components contain mostly information relevant to the system, while the later principal components contain mostly noise. These later PC's can be safely eliminated with minimal loss of information. Principal Component regression (PCR) is a multivariate regression method, which is done on the scores of the retained principal components.

2. MATERIALS AND METHODS

2.1 Instrumentation

Raman spectra were recorded on a Labram Infinity (J-Y Horiba) spectrometer, equipped with a liquid nitrogen cooled CCD detector and a 488 nm Argon ion laser excitation source which had a typical power output of ~7 mW at the sample. The liquid samples were held in a 1 cm pathlength quartz cuvette, the cuvette was held in a macro sample holder (J-Y Horiba). The macro lens has a focal length of 40 mm, which focuses through the cuvette to the centre of the liquid. Each recorded spectrum consisted of the average of 3 x 10 s exposures. All spectra were recorded at a set interval of 350-3500 cm^{-1}, with the confocal hole set at 200 µm and using a 1800 line/mm grating to give a resolution of ~11 cm^{-1}.

2.2 Materials

The concentrations chosen and the different mixtures have been made-up to try and replicate possible industrial and laboratory scenarios. Twenty-five chlorinated and non-chlorinated solvents, of various grades were used, as listed in table 1. These solvents were then mixed in various concentrations to give 225 samples of both chlorinated and non-chlorinated mixtures. A summary of these combinations is shown in Table 2.

Each sample was made up to 100 ml. The smallest quantity measured was 5 ml, which was done using a 5 ml pipette. Any quantity above 10 ml was measured using a graduated cylinder. All samples were mixed in separate conical flasks. Samples were always prepared and run on the Raman spectrometer on the same day. Over several months of spectral

data acquisition, cyclohexane was used as an external standard to monitor the varying Raman intensity from day to day. A Raman spectrum of cyclohexane was taken after every 6 samples (see figure 1).

Solvents	Grade	Solvents	Grade
Acetone	HPLC	Acetophenol *	Analytical
Toluene	Spectroscopic	n-Pentane	Analytical
Cyclohexane	Analytical & Spectroscopic	Xylene	Analytical
Acetonitrile	Spectroscopic	Nitromethane	Analytical
2-Propanol	Spectroscopic	Dimethylformamide	Analytical
1,4-Dioxane	Analytical & Spectroscopic	Nitrobenzene*	Analytical
Hexane	Analytical	Tetrahydrofuran	Analytical
1-Butanol	Analytical & Spectroscopic	Diethyl Ether	Analytical
Methyl Alcohol	Analytical	Petroleum Acetate	Analytical
Benzene	Analytical	Chloroform	Analytical & Spectroscopic
Ethyl Acetate	Analytical	Dichloromethane	Analytical & Spectroscopic
Ethanol	Analytical	1,1,1-Trichloroethane	Analytical & Spectroscopic
Cyclopentane	Analytical		

Table 1: List of chlorinated and non-chlorinated solvents and the various grades used. *Solvents containing fluorescent impurities.

	Pure Solvents	Binary Mix.	Ternary Mix.	Quaternary Mix.	Quintary Mix.	Total
Chlorinated	3	96	40	12	0	151
Non-Chlorinated	22	23	12	10	7	74
Total Number	25	119	52	22	7	225

Table 2: Summary of various chlorinated and non-chlorinated solvent mixtures

2.3 Data Analysis

Multivariate analysis was performed using the Unscrambler chemometrics software package (V8.0, CAMO, Trondheim, Norway). These results were compared with four popular machine-learning techniques: Support Vector Machine,[22] Ripper,[23] k-Nearest Neighbour,[24] and C4.5 Decision Tree.[25] The machine learning analysis was carried out using the WEKA machine learning package with default settings.[24]

Pre-processing is an integral part of the chemometric process as it can enable the removal (or minimisation) of unwanted features such as cosmic rays, noise, background scatter, and fluorescence. One of the most commonly used methods is derivative pre-processing where the derivative of the spectral curve is taken along a "window" of points. The 1st derivative is expected to remove constant offset (baseline), while 2nd derivative is expected to remove constant offset and sloping baselines. Here Savitzky Golay derivitisation was used with 7-point averaging and a 2nd degree polynomial order are used.

Cross validation is used to counter the problem of "losing" data, which arises from separating data into calibration and validation subsets which is an issue with datasets such as these where there are large numbers of components. Here a "leave-23-out" cross-validation was used as the data set was essentially divided into 10 subsets. In an iterative process, one subset of samples is left out and its concentrations are predicted with a model built from the remaining data. The root-mean-squared-error (RMSE) is a value used to quantify the errors in predicting concentrations of selected samples:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m} (c_i - c'_i)^2}{df}} \qquad \textbf{Equation 1}.$$

Here, c_i and c'_i are the true and estimated concentrations, respectively for the analyte of interest in the predicted sample *i*. The denominator *df* is the degrees of freedom used in the calculation. In this work the root-mean-squared-error-cross-validation (RMSECV) was calculated for each model and used to compare the effect of different pre-processing methods. The minimum RMSECV value is expected to occur when the optimal number of principal components is retained.

3. RESULTS AND DISCUSSION

The aim is to first utilise Raman spectroscopy in conjunction with chemometric techniques to identify the presence of chlorinated substances within mixtures of various chlorinated and non- chlorinated solvents and second to quantify these chlorinated substances. Our intermediate goal is to get a robust prediction error of less than 5%. As mentioned earlier, cyclohexane was used as an external standard to monitor the day-to-day inherent variances of the Raman intensity. The standard deviation of four main cyclohexane peaks and the total integral area of the spectra were calculated (table 3). The percentage variances show the degree of the inherent variance from day-to-day amongst the Raman intensities. This underlying error must be taken into account when analysing the results from PCA and PCR.

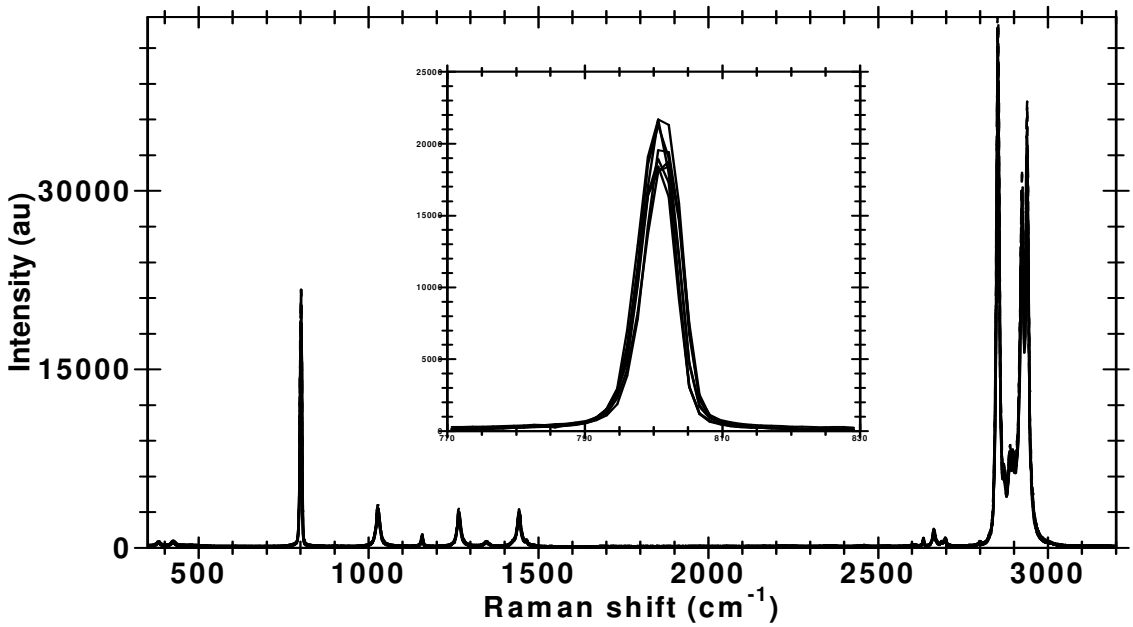

Figure 1: Raman spectra of cyclohexane, collected over 3 months (representative sample). The inset shows the intensity variation in the 801 cm^{-1} peak (see Table 3 for additional details).

3.1 Raw data

PCA was carried out on the original raw Raman spectral data. It was not successful in discriminating the chlorinated solvents from the rest, as can be seen in the PCA score plot in figure 2. PC1 (29%) and PC2 (24%) are dominated by baseline effects are related to two pure solvents acetophenol and nitrobenzene (figure 3). These two pure solvents were the only two samples in the dataset that contained appreciable baseline effects, which were probably due to traces of fluorescent impurities (figure 4). Investigation of the leverage[26] and loadings plots of all the samples revealed acetophenol and nitrobenzene to have very high leverages compared to that of all other samples. This is due to the fluorescent background of both samples having too much of an influence on the PCA model. This obviously hinders the effective identification of solvent type and therefore needs to be corrected for by data pre-processing. Normalising data reduces the inherent weighing present within spectra so this is applied to the data set first.

3.2 Normalised Data

PCA was then carried out on all the Raman spectra after max-normalisation and this gave a much-improved score plot of PC1 against PC2 (figure 5). However, the chlorinated solvents are not discriminated in these first few PC's. Figure 6 illustrates how the loadings of the first and second PC's are dominated by the strong C-H stretches that are common to all of the solvents. This region of the Raman spectrum (2700 cm^{-1} – 3300 cm^{-1}) contains a great deal of overlap among all samples. The different characteristic peaks in all samples are more apparent in the "fingerprint" region (200 cm^{-1}-1800 cm^{-1}). One exception to the above, in this dataset, is the Raman spectrum of acetonitrile, it contains a peak at ~ 2225 cm^{-1}, which is due to C≡N. This peak is present in the loadings of PC 1 and shows that acetonitrile is one of the solvents discriminated by this PC. The score plot of PC1 against PC2 for the normalised data show various groupings. These groupings are due to various non-chlorinated solvents that the chlorinated were mixed with, and marked in figure 5.

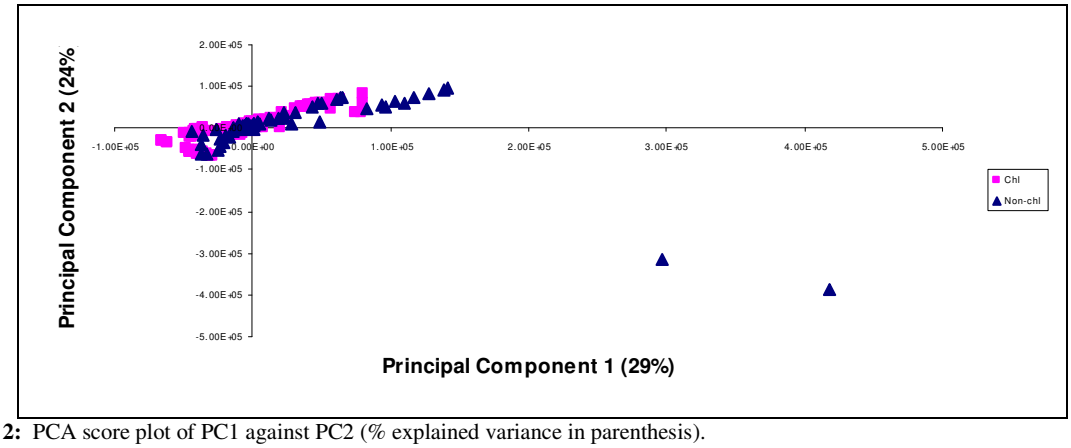

Figure 2: PCA score plot of PC1 against PC2 (% explained variance in parenthesis).

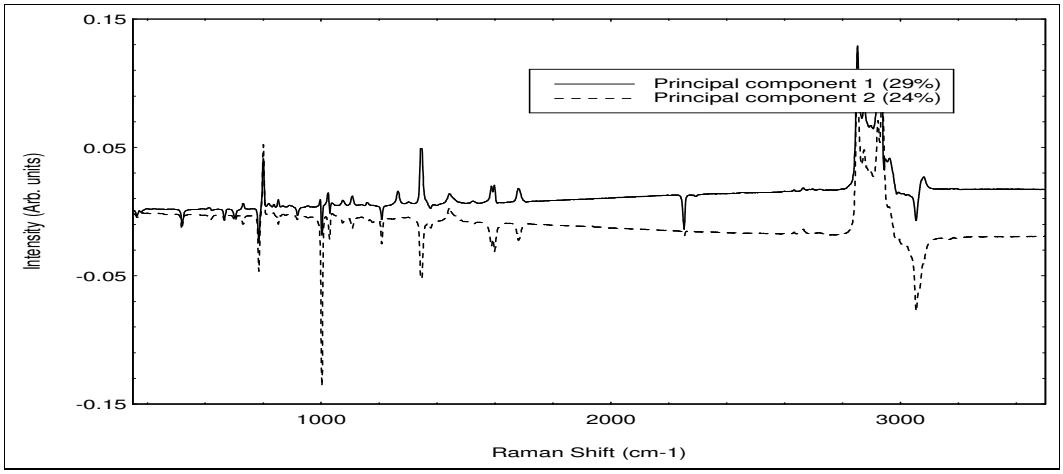

Figure 3: Loading of PC 1 and PC 2 of PCA on raw Raman spectra.

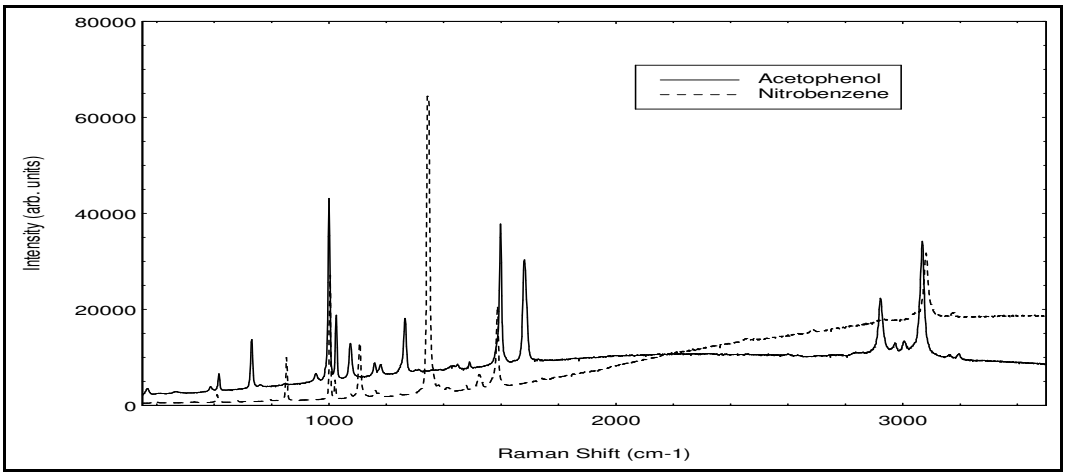

Figure 4: Raman spectra of the two solvents with fluorescing interference.

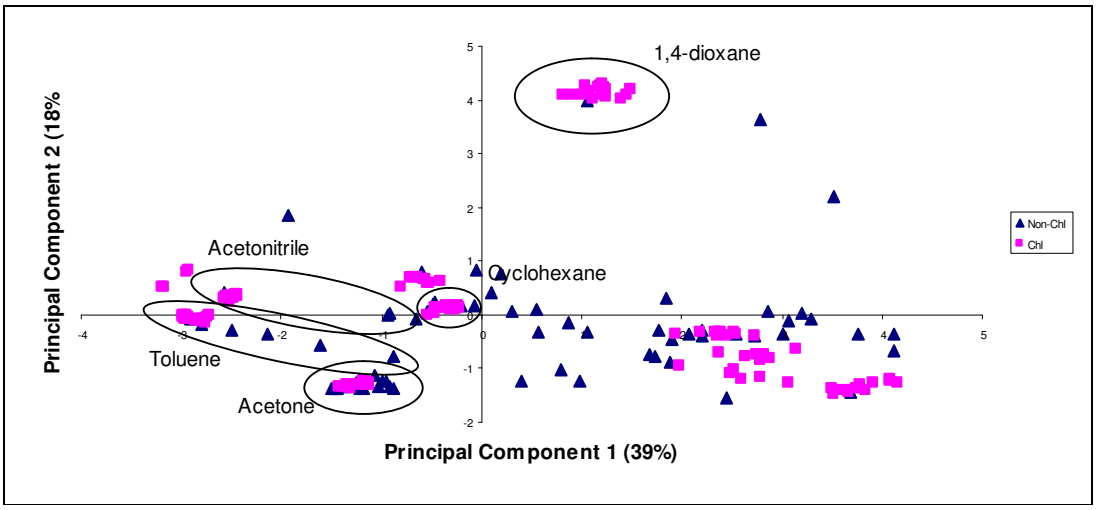

Figure 5: PCA score plot of PC1 against PC2 of Raw normalised Raman dataset (% explained variance in parenthesis).

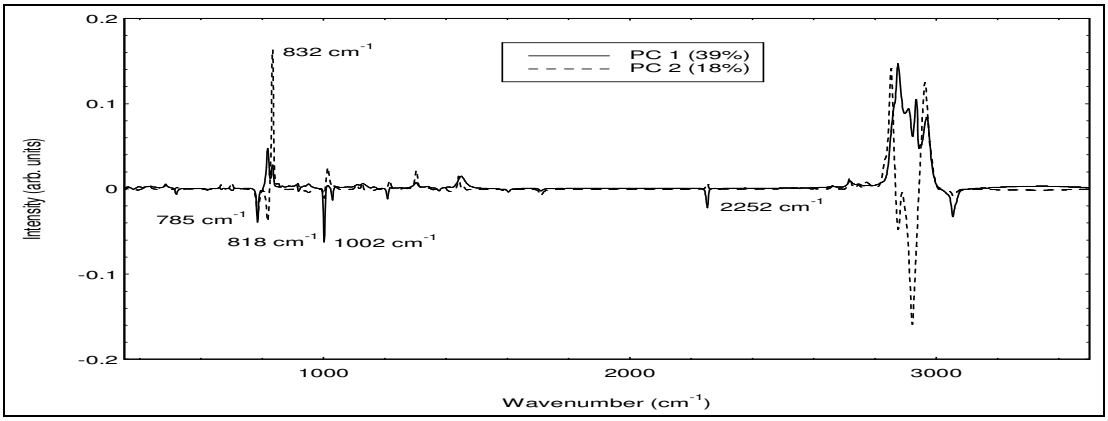

Figure 6: Loading of PC 1 and PC 2 of all raw normalised Raman spectra (% explained variance in parenthesis).

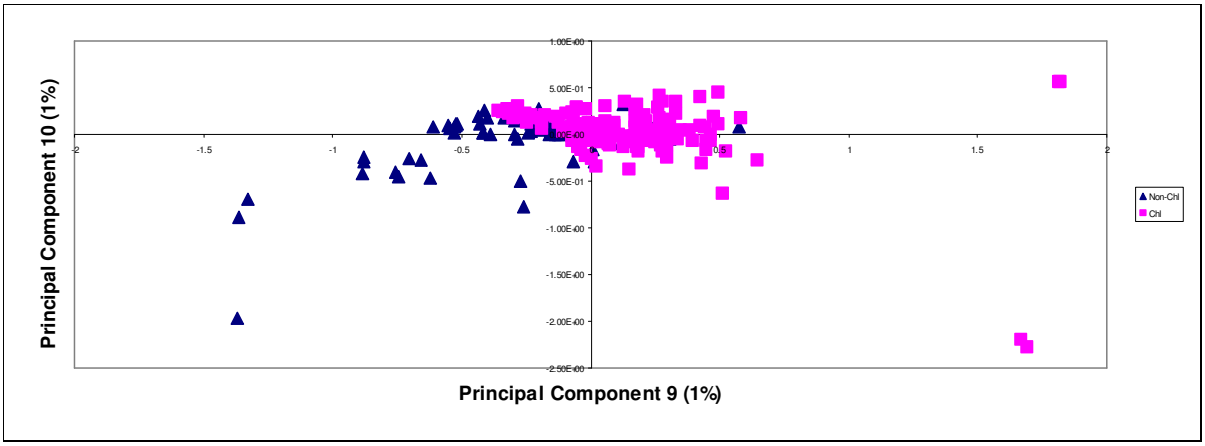

Figure 7: Score plot of PC 9 against PC 10 of normalised Raman dataset (% explained variance in parenthesis).

Acetone, 2-propanol, 1,4-dioxane, and toluene are discriminated by PC 1 and PC2 (figure 6), as the peaks at ~ 785 cm^{-1}, ~818 cm^{-1}, ~832 cm^{-1}, and ~ 1002 cm^{-1} respectively show. Given the chlorinated solvents are the minor components, occurring at no more than 20% in a mixture; the non-chlorinated solvents are discriminated for by the higher PC's. PC 9 and PC 10 of the normalised data show some discrimination of the chlorinated solvents (figure 7).

3.3 Derivatives

The derivative pre-processing method was then applied the raw Raman spectral dataset. A first derivative followed by max-normalisation was applied to the data set as was a second derivative again followed by max-normalisation. Both methods gave similar results. The chlorinated components were again better discriminated in the higher PC's. PC 9 and 10 in both cases, gave similar score plots to those observed in figure 7 above. A great deal of overlap amongst the scores for the chlorinated and non-chlorinated samples was observed. This can be directly related to the overlap of Raman peaks observed amongst the spectra of all the different samples, which makes it difficult to discriminate the chlorinated from the rest.

3.4 Reduced spectral area

One method to increase discrimination of chlorinated samples would be to focus in on the region of the Raman spectrum where the chlorinated components show the most intense peaks (the "chlorinated" region) and carry out PCA on this region of the spectrum alone. Figure 8 shows the Raman spectra of the chlorinated solvents with the chlorinated region (350 cm^{-1} – 1168 cm^{-1}) indicated. PCA was carried out on the Raman spectra in the chlorinated region. The chlorinated solvents show better discriminated in lower PC's as shown in figure 9. This is due to the fact that this part of the spectrum has a lot of information attributed to the chlorinated components. By analysing the chlorinated region the overlap of the C-H stretches (2700 cm^{-1} – 3300 cm^{-1}) is eliminated. It also reduces the time taken to carry out the PCA, from ~ 50 minutes to ~ 10 minutes.

The same pre-processing methods applied to the entire dataset were also applied to the chlorinated region. None of these methods were of particular merit in further discriminating the chlorinated solvents. This is shown in figure 10 below. The discrimination of chlorinated solvents is very similar. Figure 9 is an illustration of the score plot obtained for PC 5 against PC 8 when PCA was applied to all the raw chlorinated section of the Raman dataset. It shows good separation of the chlorinated from the non-chlorinated solvents. Figure 10 shows the score plot obtained for PC 4 against PC 8 when the second derivative was applied to the chlorinated region of the Raman dataset. Figure 11 shows the same score plot as figure 10 but with the 100% chlorinated samples removed. Good separation of chlorinated from non-chlorinated is seen.

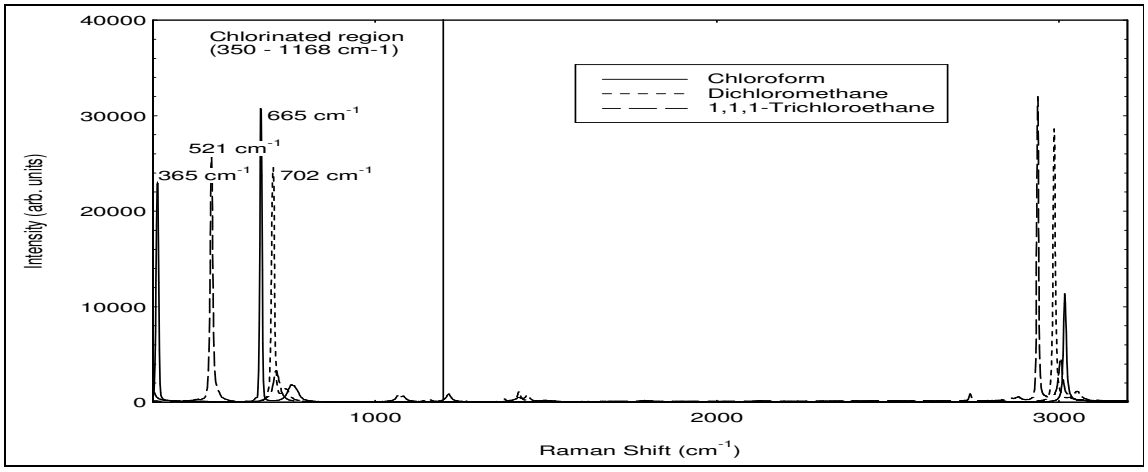

Figure 8: Raman spectra of chlorinated solvents showing the chlorinated region (350-1168 cm^{-1}).

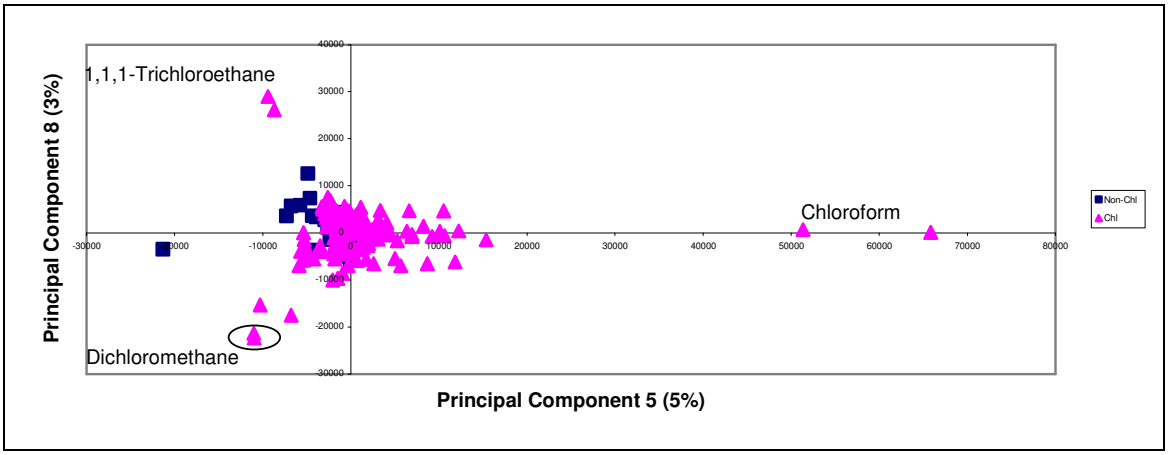

Figure 9: PCA score plot of PC 5 against PC 8 using only the raw chlorinated region of Raman spectra.

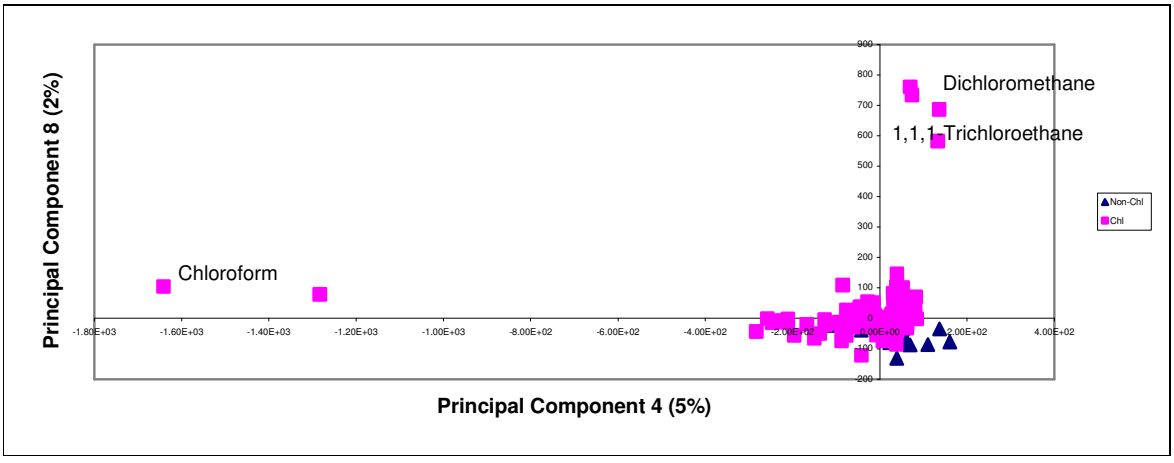

Figure 10: Score plot of PC4 against PC8 chlorinated region 2nd derivative

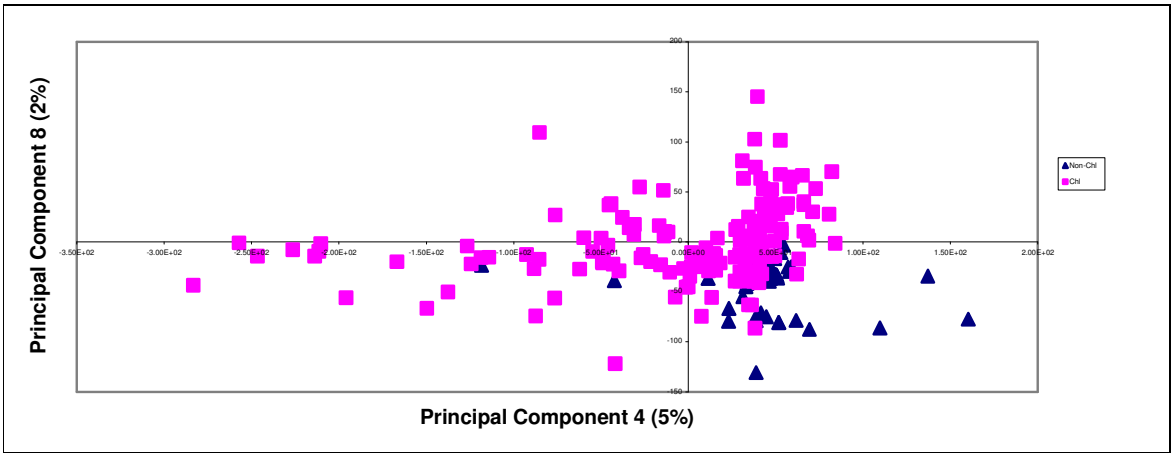

Figure 11: Score plot of PC 4 against PC 8 chlorinated region 2nd derivative without the 100% Chlorinated samples in plot

3.5 Machine Learning (ML)

The ML analysis was carried out using the WEKA machine learning package with default settings. Four experiments were undertaken using 10-fold (leave-23-out) cross validation and four ML techniques. The first three experiments involved the detection of (1) the presence/absence of 1,1,1-trichloroethane, (2) the presence/absence of chloroform, and (3) the presence/absence of dichloromethane from mixtures based on their Raman spectra. The fourth experiment involved the detection of the presence/absence of any of the three chlorinated compounds (either individually or in combination) in mixtures from their Raman spectra. The four ML classification techniques are widely used in machine learning research.[24]

3.6 Comparison of Machine Learning (ML) & Chemometrics

A comparison of the percentage errors achieved using the chemometric and ML techniques can be seen from tables 4 and 5 respectively. Table 4 shows the results when the predicted chlorinated concentrations are converted into a binary test, for comparison between PCR and ML results. If a sample has a predicted concentration below the threshold, it is said that the chlorinated compound is not present. If the predicted concentration is greater than the threshold then the sample is said to contain the chlorinated compound. Table 4 gives the errors associated with making this binary determination for each of the chlorinated compounds. All the ML analysis was carried on the raw Raman dataset. The chemometric analysis was on raw data and pre-processed data. It can be seen form table 4 that normalisation had an adverse effect on the errors. This increase in error due to max-normalisation is due to the fact that normalisation involves the division of each row by its maximum absolute value. Therefore the inherent variation of Raman intensity over the data collection period (table 3), is somewhat corrected for by the normalisation and the associated error is then included when PCR is carried out on the dataset. The best percentage error in predicting the presence of chlorinated compounds can be seen when PCR on 2nd derivative data is carried on the chlorinated region of the dataset.

Cyclohexane peak	Standard Deviation	% Variance
801 cm^{-1}	2045.8	10
1027 cm^{-1}	647.6	19
1444 cm^{-1}	602.5	20
2851 cm^{-1}	3646	8
Total Integrated Area	22929.4	11

Table 3: Standard deviation calculated for selected Raman bands of cyclohexane from data collected over a period of ~2 months.

% error in predicting presence of chlorinated compounds with various thresholds										
Solvents	Pre-Processing	CH_3CCl_3			$CHCl_3$			CH_2Cl_2		
	(Threshold)	**2%**	**3%**	**5%**	**2%**	**3%**	**5%**	**2%**	**3%**	**5%**
All	PCR-Raw	27.0	14.3	12.6	17.8	7.4	**3.0**	14.8	5.2	**3.0**
All	PCR-1st derivative	20.9	9.1	7.4	19.1	6.5	**2.6**	10.9	4.3	**3.5**
All	PCR-2nd derivative	17.4	9.6	7.0	18.3	5.7	**2.6**	15.7	7.8	**5.7**
All	PCR-max-norm	17.8	10.9	10.4	17.8	7.4	**5.2**	21.3	11.7	**4.8**
All	PCR-1st Der-max norm	17.8	13.9	10.9	19.1	9.1	**4.3**	21.7	11.7	**3.9**
All	PCR-2nd Der-max norm	22.2	15.2	10.9	18.3	13.0	**2.2**	20.0	14.3	**3.0**
Chl-section	PCR-Raw	11.3	**3.5**	**4.3**	10.9	5.7	**3.0**	3.9	3.0	**2.6**
Chl section	PCR-1st derivative	9.1	**2.6**	**3.0**	12.2	8.3	**2.6**	4.8	3.5	**3.5**
Chl section	PCR-2nd derivative	13.5	**2.2**	**3.0**	13.0	5.2	**3.9**	6.1	4.3	**3.0**
Chl-section	PCR-max-norm	13.5	4.3	4.3	12.2	6.5	**3.0**	8.7	3.9	**2.2**
Chl-section	PCR-1st Der-max norm	16.1	**3.5**	**5.7**	18.3	7.0	5.2	10.4	8.3	**3.0**
Chl-section	PCR-2nd Der-max norm	15.7	7.4	6.1	14.3	10.0	7.4	9.6	8.7	**3.0**

Table 4: Percentage error in predicting presence of chlorinated compounds with various thresholds.

Classifier	**% Classification Error**			
	111-trichloroethane CH_3CCl_3	**Chloroform $CHCl_3$**	**Dichloromethane CH_2Cl_2**	**Any Chlorinated Compound**
Support Vector Machine	**4.8***	8.3	**4.4**	**4.8***
Ripper	7.9	**4.8**	**3.1**	**7.4**
K Nearest Neighbour	24.9	22.3	9.6	**6.1**
C4.5 Decision Tree	**7.0**	**3.5***	**2.6***	10.5

Table 5: Comparison of machine learning methods for the classification of chlorinated solvents.

Table 5 details the classification results using the four ML techniques on the Raman spectra (raw, no preprocessing) of the solvent mixtures. The best result for classification of each solvent or group of solvents is denoted using an asterisk and results in bold show no significant difference from the best result. Significance was tested using a corrected resampled t-test[27] at the 95% confidence level. The ML technique that showed the greatest classification accuracy overall was the Ripper rule-learning algorithm, though the results of the SVM classifier also look promising.

		RMSECV (#PC's)		
# of solvents	**Pre-processing**	**1,1,1-Trichloroethane**	**Chloroform**	**Dichloroethane**
All	PCR-Raw	4.8 (18)	3.4 (15)	2.7 (15)
All	PCR-Max Normalised	5.7 (17)	3.5 (15)	3.9 (14)
All	PCR-1stDeriv-Max norm	7.6 (17)	4.8 (16)	4.9 (12)
All	PCR-2ndDeriv-Max norm	7.8 (18)	5.9 (16)	4.5 (15)
All-Chl. section	PCR-Raw	3.3 (13)	3.3 (7)	2.0 (11)
All-Chl. section	PCR-Max Normalised	3.5 (12)	3.8 (7)	2.7 (10)
All-Chl. section	PCR-1stDeriv.-Max norm	4.8 (12)	4.5 (8)	3.8 (10)
All-Chl. section	PCR-2ndDeriv.-Max norm	6.8 (11)	6.5 (5)	3.7 (11)
All	PCR-1stDeriv.	3.8 (20)	3.4 (13)	3.2 (12)
All	PCR-2ndDeriv.	2.8 (20)	2.9 (12)	2.7 (13)
All-Chl. section	PCR-1stDeriv.	**2.9 (12)**	**2.8 (7)**	**1.9 (9)**
All-Chl. section	PCR-2ndDeriv.	**2.8 (14)**	**2.8 (5)**	**1.9 (10)**

Table 6: Summary of RMSECV values for various spectral regions and processing methods (# of PC's retained in parenthesis).

It is interesting to note that while the k-Nearest Neighbour algorithm shows poor accuracy in classifying the presence/absence of the individual chlorinated compounds, it has one of the best results when classifying whether 'any' chlorinated compound is present. Conversely, C4.5 achieves good results on each of the individual chlorinated

compounds but performs badly on detection of the presence/absence of any chlorinated compound. This may be due to the ratio of positive to negative examples of the target analyte in the different experiments (approx 1:2 positive:negative in Experiments 1 to 3 and approx 2:1 positive:negative in Experiment 4). Different machine learning algorithms have differing abilities to deal with datasets that are 'balanced' differently, i.e. have fewer or more positive examples.

3.7 Quantification of chlorinated compounds

Table 6 is a summary of all the various pre-processing methods that were applied to the Raman spectral dataset along with the RMSECV values. From these it can be seen that better RMSECV values are achieved using 1st or 2nd derivative pre-processing without normalisation. This adverse effect of normalisation can also be seen when the chlorinated section of the Raman spectra are analysed. As mentioned in section 3.6 above normalisation incorporates the error associated with the day-to-day fluctuations of the Raman intensity. There is a clear increase in error from PCR on raw data to PCR on 2nd derivative-max normalised data. The worst errors are obtained for 1,1,1-trichloroethane in all cases of PCR. The best results for each of the chlorinated solvents were achieved using PCR on either the 1st or 2nd derivatives of all the chlorinated section of the spectra. RMSECV of less than 3% is achieved which is promising as our intermediate goal was to achieve errors of less than 5%.

4. CONCLUSIONS

This study investigated the use of Raman spectroscopy for differentiating between chlorinated and non-chlorinated solvent mixtures by employing chemometric and machine learning techniques. Quantification of the level of chlorinated solvent in low to medium concentration (less than 20%) using the same methods was also studied. Traditional chemometrics (e.g. PCA/PCR) illustrated the difficulty in discriminating the solvent mixtures with low chlorinated concentration when the full spectral range was employed. This is largely due to spectral overlap in the C-H region. However, restricting the spectral range to the 350-1300 cm^{-1} (C-Cl stretch region of the spectrum) allowed for a better discrimination using fewer PC's. Identification of the presence or absence of chlorinated solvents was achieved with an error of approximately 5% using the 2nd derivative of the restricted spectral range.

To compare the results obtained using chemometric to the machine learning methods various threshold levels were set (2, 3 and 5%). The machine learning methods showed the best results for classifying the solvents. 1,1,1-trichloroethane was best discriminated for using support vector machines. Both chloroform and dichloromethane were best discriminated for using C4.5 decision tree. When all three solvents were discriminated together support vector machines gave the best result. The machine learning technique that showed the greatest classification accuracy was the Ripper rule-learning algorithm, which had a best result in each of the solvent classification experiments. Support Vector Machines and C4.5 Decision Tree also did well with best results in three out of four classification experiments. This indicates that these ML methods seem to be very similar in performance to traditional chemometrics, although we are investigating this further to determine exact error levels.

Quantification of the chlorinated solvents was achieved with an error of less than 3%. Again the best error here was achieved using PCR on the chlorinated region of the Raman spectra that had been treated with a 2nd derivative. These results are promising, however, there is scope for improvement. One issue that is being addressed is the inherent variation of Raman intensity from day to day and an intensity correction method using the external cyclohexane standard is being explored. This should further reduce predictive errors leading to a robust analytical method.

5. ACKNOWLEDGEMENTS

This work was funded via a grant from Enterprise Ireland's Commercialisation Fund Technology Development Programme (grant no: TD/03/212). The Raman instrumentation was provided by the National Centre for Biomedical Engineering Science as part of the Irish Higher Education Authority's Programme for Research in Third Level Institutions. AR is supported by Science Foundation Ireland grant no. 02/IN.1/M231.

6. REFERENCES

[1] R.L. McCreery. *Raman spectroscopy for chemical analysis*, Wiley Chemical Analysis Series, Vol. 157, J. Winefordner, ed., John Wiley & Sons, New York, 2000.

[2] S.D. Harvey, M.E. Vucelick, R.N. Lee, and B.W. Wright. "Blind field test evaluation of Raman spectroscopy as a forensic tool," *Forensic Sci. Int.*, **125**(1), 12-21, 2002.

[3] C.M. Hodges and J. Akhavan. "The use of Fourier Transform Raman spectroscopy in the Forensic identification of illicit drugs and explosives," *Spectrochim. Acta*, **46**(2), 303-307, 1990.

[4] A.H. Kuptsov. "Applications of Fourier transform Raman spectroscopy in forensic science," *J. Forensic Sci.*, **39**, 305-318, 1994.

[5] S.E.J. Bell, D.T. Burns, A.C. Dennis, and J.S. Speers. "Rapid analysis of ecstasy and related phenethylamines in seized tablets by Raman Spectroscopy," *Analyst*, **125**, 541-544, 2000.

[6] A.G. Ryder. "Classification of narcotics in solid mixtures using Principal Component Analysis," *J. Forensic Sci.*, **47**(2), 275-284, 2002.

[7] I.P. Keen, G.W. White, and P.M. Fredericks. "Characterization of Fibres by Raman Microprobe Spectroscopy," *J. Forensic Sci.*, **43**(1), 82-89, 1998.

[8] E.M. Suzuki and M. Carrabba. "In Situ Identification and Analysis of Automotive Paint Pigments using Line Segment Excitation Raman Spectroscopy: Inorganic Topcoat Pigments," *J. Forensic Sci.*,. **46**(5), 1053-1069, 2001.

[9] S. Stich, D. Bard, L. Gros, H.W. Wenz, J. Yarwood, and K. Williams. "Raman Microscopic Identification of Gunshot Residue," *J. Raman Spectrosc*., **29**, 787-790, 1998.

[10] T. Ostertag, L. Kiefert, and H.A. Hanni. "Raman spectroscopic applications to Gemmology," in *Handbook of Raman Spectroscopy*, I.R. Lewis, Editor. 2001.

[11] N. Gupta and R. Dahmani. "AOTF Raman spectrometer for the remote detection of explosives," *Spectrochim. Acta (A)*, **56**, 1453-1456, 2000.

[12] I.R. Lewis, N.W. Daniel Jr, N.C. Chaffin, P.R. Griffiths, and M.W. Tungol. "Raman spectroscopic studies of explosive materials: towards a fieldable explosive detector," *Spectrochim. Acta (A)*, **51**, 1985-2000, 1995.

[13] N.F. Fell, J.A. Vanderhoff, R.A. Pesce-Rodriguez, and K.L. McNesby. "Characterization of Raman spectral changes in Energetic Materials and Propellants during heating," *J. Raman Spectrosc.*, **29**, 165-172, 1998.

[14] I.P. Hayward, T.E. Kirkbride, D.N. Batchelder, and R.J. Lacey. "Use of Fibre Optic Probe for the Detection and Identification of Explosive materials by Raman Spectroscopy," *J. Forensic Sci.*, **40**(5), 883-884, 1995.

[15] P.R. Griffiths, I.R. Lewis, and N.C. Chaffin. "Infrared and Raman spectroscopic studies of asbestos, transite and concrete," *Mikrochim. Acta*, **14**, 191-192, 1997.

[16] C.J. Strachan, D. Pratiwi, K.C. Gordon, and T. Rades. "Quantitative analysis of polymorphic mixtures of carbamazepine by Raman spectroscopy and principal component analysis," *J. Raman Spectrosc.*, **35**, 347-352, 2004.

[17] F. Estienne, D.L. Massart, N. Zanier-Szydlowski, and P. Marteau. "Multivariate calibration with Raman spectroscopic data: a case study," *Anal. Chim. Acta*, **424**, 185-201, 2000.

[18] F. Estienne and D.L. Massart. "Multivariate calibration with Raman data using fast principal component regression and partial least squares methods," *Anal. Chim. Acta*, **450**, 123-129, 2001.

[19] F. Estienne. "A comparison of multivariate calibration techniques applied to experimental NIR data sets Part III: Robustness against instrumental perturbation conditions," *Chemometr. Intell. Lab.*, **73**, 207-218, 2004.

[20] R.G. Brereton. "Introduction to multivariate calibration in analytical chemistry," *Analyst*, **125**, 2125-2154, 2000.

[21] H. Martens and T. Naes. *Multivariate Calibration*. 1989, Chichester, England: John Wiley & Sons. 419.

[22] V. Vapnik, *The nature of statistical learning theory*, New York:Springer-Verlag.

[23] W.W. Cohen, Fast effective rule induction, *Proc. Twelfth International Conference on Machine Learning*, Tahoe City, CA. San Francisco: Morgan Kaufmann, pp. 115-123, 1995.

[24] I.H. Witten and E. Frank. *Data Mining: Practical machine learning tools with Java implementations*, ed. M. Kaufmann, San Francisco. 2000.

[25] J.R. Quinlan, *C4.5: Programs for machine learning*, San Francisco: Morgan Kaufmann, 1993.

[26] T. Naes. "Leverage and influence measures for Principal Component Regression," *Chemometr. Intell. Lab.,* **5**, 155-168, 1989.

[27] C. Nadeau and Y. Bengio. Inference for the generalization error, *Advances in Neural Information Processing Systems* **12**, MIT Press, 2000.