CRANFIELD UNIVERSITY

ANDREW SPOONER

MULTIVARIATE ANALYSIS METHODS FOR VETERINARY DIAGNOSTICS USING SIFT-MS

CRANFIELD HEALTH

PHD THESIS

SUPERVISOR: DR. C. BESSANT

2010

CRANFIELD UNIVERSITY

CRANFIELD HEALTH

PHD THESIS

2010

ANDREW SPOONER

MULTIVARIATE ANALYSIS METHODS FOR VETERINARY DIAGNOSTICS USING SIFT-MS

SUPERVISOR: DR C. BESSANT

© Cranfield University 2010. All rights reserved. No part of this publication may be reproduced without the written permission of the copyright owner.

Abstract

Selected ion flow tube mass spectroscopy (SIFT-MS) is an analytical method for the investigation of volatile organic compounds (VOCs). It produces mass to charge (m/z) ratio ion counts with a range of 10-200 m/z. Current data analysis involves sifting through the spectra files one at a time looking for peaks of interest. This is time consuming and requires expert knowledge. This thesis proposes, implements and demonstrates a novel approach to the analysis of SIFT-MS data using multivariate techniques similar to those employed to analyse electronic nose and gas chromatography mass spectroscopy (GCMS) data. The methodology was developed using a set of samples created in the laboratory that belonged to two groups which contained different VOCs found in biological samples. The methodology requires the removal of the m/z peaks associated with the precursors, then principal component analysis (PCA) and partial least squares discriminant analysis (PLSDA) methods were evaluated for biomarker discovery and sample classification. Both methods produced excellent results, identifying the volatiles in the mixtures and being able to classify samples with 100% accuracy. This methodology was then tested using a variety of samples. Ammonia was found as a possible marker for bovine TB (*Mycobacterium bovis)* infection using serum samples taken from wild badgers. Discrimination results of an accuracy of 67%±6% were acquired. The number of sample needed to build the best performing model from this dataset was empirically shown to be 120. It was shown to be effective for the discrimination of serum samples from cattle taken before and after introduction of bovine TB (*Mycobacterium bovis*) bacteria in a clinical trial (accuracy of 85% achieved). A similar dataset pertaining to infection by *Mannheimia haemolytica* failed to produce models that performed as well as the others - this is suspect to be due to a poor experimental design. Finally, discrimination accuracies of 88% for urine samples collected from cattle from herds infected with *Mycobacterium paratuberculosis* and 90% for urine samples collected in the same bovine TB trial as above were achieved. The novel multivariate approach to SIFT-MS data analysis has been shown to be effective with a number of datasets but it is sensitive to the experimental design. Recommendation for the consideration required for analysis using this method have been made.

Acknowledgements

I would like to thank my supervisors Dr. Conrad Bessant and Dr. Claire Turner for their help and understanding throughout the project. I would also like to thank Prof. Tony Turner for his initial interest in the project.

A big thank you has to go to Henri Knobloch who prepared and analysed most of the samples described in this thesis. A big part of a data analysis project is the data and Henri supplied a great quantity making analysis worthwhile.

I would also like to thank my family for their continued support and patience during my PhD project.

I would like to thank DEFRA, the VLA and the projects German partners at FLI for supply all of the biological samples for analysis.

This thesis forms part of project SE3221 which was funded by the UK's Department for Environment, Food and Rural Affairs and the author gratefully acknowledges DEFRA's funding.

Table of Contents

Abstract i

Acknowledgements i

List of Figures vi

List of Tables xx

Abbreviations xxi

1. Introduction and Literature Review 1

1.1. Introduction 1

1.2. Literature Review 2

1.2.1. Diseases Studied 3

1.2.2. Analytical Tools 6

1.2.3. Alternative technologies 9

1.2.4. Data Analysis Methods 11

1.2.5. Disease detection using multivariate techniques and volatile analysis 18

1.3. Aims & Objectives 19

2. Evaluation of Multivariate Analysis Methods for SIFT-MS Data 21

2.1. Introduction and Motivation 21

2.2. Data Collection & Measurement 22

2.3. Analysis with Trans SIFT mk. 2 23

2.3.1. Univariate & Classical Approach 23

2.3.2. PCA 34

2.3.3. PLSDA 45

2.3.4. Bootstrapped PLSDA 49

2.4. Analysis with Trans SIFT mk. 3 50

2.4.1. Univariate & Classical Approach 50

2.4.2. PCA 53

2.4.3. PLSDA 55
2.4.4. Combined dataset 56
2.5. Comparison and Conclusions 59
2.5.1. Comparison between SIFT machines 59
2.5.2. Conclusions 60
3. Application of methods to determination of tuberculosis status from serum samples taken from Wild Badgers 62
3.1. Introduction 62
3.2. Data Collection and Measurement 63
3.3. Data Analysis 63
3.3.1. Univariate & Classical Approach 63
3.3.2. PCA analysis 68
3.3.3. PLSDA analysis 88
3.3.4. Bootstrapped PLSDA analysis of data 94
3.4. Conclusions 97
4. Application of method to the determination of disease status from serum samples taken from Cattle 100
4.1. Introduction 100
4.2. Data Collection and Measurement 100
4.2.1. *Mannheimia haemolytica* data 102
4.2.2. *Mycobacterium paratuberculosis* data 104
4.2.3. Brucella data 104
4.2.4. *Mycoplasma bovis* data 104
4.2.5. *Mycobacterium bovis* (bovine TB) data 106
4.3. Results 107
4.3.1. *Mannheimia Haemolytica* Results 107

4.3.2. *Mycobacterium paratuberculosis* Results 112
4.3.3. Brucella results 117
4.3.4. Mycoplasma Bovis Results 123
4.3.5. Bovine TB (*Mycobacterium bovis*) Results 125
4.4. Discussion and Conclusion 132
5. Application of the methods to the classification of disease status from Urine collected from Cattle 136
5.1. Introduction 136
5.2. Data Collection & Measurement 136
5.2.1. Mycobacterium paratuberculosis 137
5.2.2. Bovine TB (*Mycobacterium* bovis) data 137
5.3. Data Analysis 138
5.3.1. Mycobacterium paratuberculosis Results 138
5.3.2. Bovine TB (*Mycobacterium* bovis) results 145
5.4. Conclusions 152
6. Discussion and Conclusions 154
6.1. Summary of Work Conducted 154
6.2. Proposed methodology 157
6.3. Overall Conclusions 158
6.4. Discussion of the performance in context of other detection methods 159
6.5. Further Work & Recommendations 162
References 165
Appendix 1 – Evaluation of a combination of SIFT-MS and multivariate data analysis for the diagnosis of Mycobacterium bovis in wild badgers i
Appendix 2 - Methodological variation in headspace analysis of liquid samples using electronic nose ix

List of Figures

Figure 1.1 - A schematic of the internal workings of a SIFT-MS machine (Smith and Spanel, 2005). The ions written in the tube are the precursor ions (upstream) and possible precursor-sample interactions (downstream). Different sample preparation techniques are also suggested. 7

Figure 1.2 - An example of a SIFT-MS FSM spectrum as produced by the standard software shipped with the machine. 12

Figure 2.1 - The H_3O^+ spectrum straight from the SIFT-MS machine for a sample of mixture A (sample no. 16). The peaks produced by the principal precursor and its water clusters have been numbered. 24

Figure 2.2 - The spectrum as captured using the H_3O^+ precursors for a sample of mixture A (sample number 16) after the m/z counts for the precursors (as outlined in Table 2.2) have been removed. 26

Figure 2.3 - The spectrum as captured using the H_3O^+ precursor for a sample of mixture B (sample number 1) after the m/z counts for the precursors have been removed. 28

Figure 2.4 - The intensities (counts) of m/z 18 (ammonia) as recorded on the spectra produced using the H_3O^+ precursor for all samples. Samples of mixture A have been coloured red and are on the left, samples of mixture B have been coloured blue and are on the right. Ammonia was one of the volatiles used to create mixture A but was not supposed to be present in mixture B. 29

Figure 2.5 - The intensities (counts) of m/z 33 (methanol) as recorded on the spectra produced using the H_3O^+ precursor for all samples. Samples of mixture A have been coloured red and are on the left, samples of mixture B have been coloured blue and are on the right. 30

Figure 2.6 - The intensities (counts) of m/z 47 (ethanol) as recorded on the spectra produced using the H_3O^+ precursor for all samples. Samples of mixture A have been coloured red and are on the left, samples of mixture B have been coloured blue and are on the right. 31

Figure 2.7 - The mean average spectrum (all thirty spectra of each sample type summed and divided by thirty) for the two mixtures. Mixture A is coloured red and

mixture B is coloured blue. A large number of extra volatiles not expected (see Table 2.1) can be seen. 32
Figure 2.8 - PCA scores plot of the H_3O^+ spectra for all the samples, red dots are individual samples of mixture A and blue crosses are samples of mixture B. Clusters have been highlighted using the corresponding colours showing that despite the first component containing 99.4% of the variance little discrimination can be seen on this axis alone. Partial discrimination can be seen on the secondary axis although the clusters overlap. 35
Figure 2.9 - The loadings plot of the first principal component for the above PCA scores plot (Figure 2.8) using the H_3O^+ precursor spectra only. The main contributors to the variance above are linked to the H_3O^+ precursor ion and its water clusters which are added to the sample as part of the analysis process i.e. they are not part of the original sample. 36
Figure 2.10 - PCA scores plot of the samples after the m/z associated with the precursors have been zeroed. The red dots are samples of mixture A and the blue crosses are samples of mixture B. Clear discrimination between the two groups can be seen, as highlighted by coloured eliplises (added manually). 38
Figure 2.11 - The loadings of the first principal component corresponding to the previous PCA scores plot (Figure 2.10) using the H_3O^+ data with the precursors removed. The largest peak (m/z 18) is linked to ammonia. 39
Figure 2.12 - The loadings of the second principal component of the previous PCA score plot (Figure 2.10) using the H_3O^+ data with the precursors removed. The largest peak (m/z 93) is linked to toluene. 40
Figure 2.13 - The percentage variance captured for the first 30 principal components (bars) and the cumulative sum of the components (line) for the PCA analysis (shown in Figure 2.10) created using the H_3O^+ data with the precursors removed. 41
Figure 2.14 - PCA scores plot of the mixtures data of all three precursor spectra combined into a single vector. m/z linked to the precursors ions have been zeroed. Clear separation can be seen between the two outlined clusters. 42
Figure 2.15 - The loadings of PC1 from the PCA analysis performed using data created by combining the spectra recorded using all three precursor ions (Figure 2.14). The m/z ratios from the largest peaks have been highlighted and the m/z ratio

prefixed with the letter corresponding to the precursor channel used in the sampling process. 43
Figure 2.16 - The loadings of PC2 from the PCA analysis performed using data created by combining the spectra recorded using all three precursor ions (Figure 2.14). The m/z from the largest peaks have been highlighted and the m/z numbers prefixed with the letter corresponding to the precursor channel used in the sampling process. 44
Figure 2.17 - The percentage of variance captured by each of the first 30 principal components (bars) and the cumulative sum (line) for the above PCA analysis (Figure 2.14) created using a combined vector of all three precursor channels. 45
Figure 2.18 - The results from a PLSDA leave one out (LOO) validation for the H_3O^+ precursor spectra. The four lines show the model accuracy (blue o), the LOO accuracy (green x). The LOO is then broken down into true positive (red +) and true negative (light blue dots). The m/z ratios associated with the precursors have been removed but no scaling methods were applied. 47
Figure 2.19 - The results from a PLSDA LOO validation for the H_3O^+ precursor spectra. The four lines show the model accuracy (blue o), the leave one out (LOO) accuracy (green x). The LOO is then broken down into true positive (red +) and true negative (light blue dots). The m/z ratios associated with the precursors have been removed and the data has been pre-treated using a mean centre algorithm (Eq. 1.2). 48
Figure 2.20 - The LOO accuracy of the PLSDA models built using only a specific number of samples (accuracy from s.d. of 20 bootstraps is shown as error bars). The data was mean centred (Eq. 1.2) before analysis. 49
Figure 2.21 - The spectrum from a sample of mixture A (sample no. 22) recorded using the H_3O^+ precursor and the newer mk.3 SIFT-MS machine. Peaks associated with m/z ratios caused by the precursors have been removed. 51
Figure 2.22 - The spectrum as recorded by the newer SIFT-MS machine using the H_3O^+ precursor for a sample (sample number 1) of mixture B. Peaks associated with m/z ratios caused by the precursors have been removed. 52
Figure 2.23 - The average intensity for the two mixtures as recorded using the H_3O^+ precursor and the newer mk.3 SIFT-MS machine. Red bars are the average counts of mixture A and blue bars are the average counts of mixture B. 53

Figure 2.24 - The PCA scores plot from the dataset collected using the newer SIFT-MS for the two mixtures. Red dots are mixture A and blue crosses are mixture B. The precursor m/z have been removed but no scaling techniques have been used. The data was collected using only the H_3O^+ precursor. Clear and separate clusters have been drawn around the samples in the corresponding colours. 54
Figure 2.25 - The variance captured by each of the first 30 principal components (bars) and the cumulative sum (line) for the above PCA analysis (Figure 2.24) created using the H_3O^+ spectra data recorded using the newer mk.3 SIFT-MS machine with the m/z ratios associated with the precursors removed 55
Figure 2.26 - The PCA scores plot from the data collected from both SIFT-MS machines. Red dots are mixture A and blue crosses are mixture B. The precursor m/z have been removed but no scaling techniques have been used. The data was collected using only the H_3O^+ precursor. The cluster of red and blue samples near the zero of both axis are samples recorded using the older mk.2 SIFT-MS machine and the two clusters further from the origin of both axises are samples recorded using the newer mk.3 SIFT-MS machine. The samples recorded using the mk.2 SIFT-MS machine have been outlined using the green oval and the samples recorded using the newer mk.3 SIFT-MS machine have been outlined using the purple oval. 57
Figure 2.27 - The percentage of variance captured for the first 30 principal components (bars) and the cumulative sum (line) for the above PCA analysis (Figure 2.26) 58
Figure 3.1 - The spectrum of sample 1 as recorded using the H_3O^+ precursor of the mk.2 SIFT-MS machine. This serum sample comes from a badger that was culture negative for Tuberculosis. 64
Figure 3.2 - The spectrum as recorded using the H_3O^+ precursor using the mk.2 SIFT-MS machine for sample 201. This serum sample comes from a badger that was confirmed as having Tuberculosis using the culture test. 65
Figure 3.3 - The average spectra as recorded using the H_3O^+ precursor channel using the mk.2 SIFT-MS for each sample group. Blue are samples as confirmed positive using the culture test and red are the ones that were not (negative). 66

Figure 3.4 - The percentage of non zero m/z counts for the spectra as recorded using the H_3O^+ precursor of the mk.2 SIFT-MS for serum samples from badgers that were not confirmed as tuberculosis positive (i.e. negatives) 67
Figure 3.5 - The percentage of non zero m/z counts from the spectra recorded using the H_3O^+ precursor of the mk.2 SIFT-MS machine for serum samples from badgers confirmed as tuberculosis positive 68
Figure 3.6 - PCA score plot of the H_3O^+ precursor spectra as recorded using the mk.2 SIFT-MS machine for the serum samples collected from the wild badgers. Blue samples are tuberculosis culture positive and red samples are tuberculosis culture negative. m/z peaks associated with the precursors have been removed. 69
Figure 3.7 - The loadings plot of the first principal component of the above PCA analysis (Figure 3.6). m/z 77 is the dominant cause for the first dimension structure. 70
Figure 3.8 - The loadings of the second principal component from the above PCA analysis. m/z 83 is the largest contributor to the structure seen on the second dimension. 71
Figure 3.9 - The percentage of variance captured by the first 30 principal components (bars) and the cumulative sum (line) of the above PCA (Figure 3.6) 72
Figure 3.10 - PCA scores plot of the H_3O^+ precursor spectra recorded using the mk.2 SIFT-MS machine for the serum samples collected from wild badgers. m/z peaks associated with the precursors as well as m/z 59, 77, 95 (acetone) and m/z 47, 65, 83 (ethanol) have been removed. Red samples are tb negative and blue are positive 73
Figure 3.11 - The loading of the second principal component from the above PCA analysis (Figure 3.10) showing that the slight discrimination is caused by m/z ratios 18, 54, 69 and 111 74
Figure 3.12 - The percentage of variance captured by the first 30 principal components (bars) and the cumlative sum (line) for the above PCA analysis (Figure 3.10) 75
Figure 3.13 - The counts for ammonia (m/z 18) for all samples of infected (blue) and uninfected (red) as recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. 76

Figure 3.14 - The m/z intensity of m/z 111 as recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. There is clearly a significant structure formed and under further investigation appears to be linked with the service time of the SIFT-MS machine. The x axis above is firstly samples separated by their classification group however within each group the samples are ordered by date of analysis by the SIFT-MS machine. The large block of uninfected samples and much smaller block of infected samples towards the right of each group were recorded after the service and they appear to have higher recorded values..... 77
Figure 3.15 - The percentage of non-zero counts for samples recorded before the mk.2 SIFT-MS machine service (Red) and after the machine service (blue) on the spectra produced using the H_3O^+ precursor..... 78
Figure 3.16 - The average counts for the m/z peak on spectra recorded before the mk.2 SIFT-MS machine service (red) and after the service (blue) using the H_3O^+ precursor channel..... 79
Figure 3.17 - PCA score plots of the serum samples collected from wild badgers recorded using the H_3O^+ precursor of the mk.2 SIFT-MS machine. The m/z associated with ethanol, acetone and the precursors have been removed as have m/z ratios above 109..... 80
Figure 3.18 - The percentage of variance captured by the first 30 principal components (bars) and the cumulative sum (line) for the above PCA analysis (Figure 3.17)..... 81
Figure 3.19 - PCA scores plot of the serum samples collected from wild badgers recorded by the mk. 2 SIFT-MS machine using the H_3O^+ precursor. The m/z associated with ethanol, acetone and the precursors have been removed as have m/z above 109. The data was pre-treated using a mean centring algorithm (Eq. 1.2)..... 82
Figure 3.20 - The percentage of the variance captured by the first 30 principal components (bars) and the cumulative sum (lines) for the above PCA analysis (Figure 3.19)..... 83
Figure 3.21 - PCA scores plot of the spectra of serum samples collected from wild badgers recorded using the mk.2 SIFT-MS machine and the H_3O^+ precursor ion. The m/z ratios associated with ethanol, acetone and the precursors have been removed as

have m/z ratios above 109. The data was pre-treated using an auto-scaling algorithm (Eq. 1.3). 84
Figure 3.22 - The percentage of variance captured by the first 30 principal components (bars) and the cumulative sum (line) for the above PCA analysis (Figure 3.21) 85
Figure 3.23 - The loadings of PC2 for the above PCA scores plot of the spectra of serum samples collected from wild badgers recorded using the mk.2 SIFT-MS machine and the H_3O^+ precursor ion where signs of discrimination between the class could be seen. The m/z ratios associated with ethanol, acetone and the precursors have been removed as have m/z ratios above 109. The data was pre-treated using an auto-scaling algorithm. 86
Figure 3.24 - PCA scores plot of the badger serum samples are recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. m/z peaks associated with acetone, ethanol and the precursors as well as m/z greater than 109 have been removed. The data was pre-treated using Log_{10} and then mean centred (Eq. 1.5). 87
Figure 3.25 - The percentage of variance captured by the first 30 principal components (bars) and the cumulative sum (line) for the above PCA analysis (Figure 3.24) 88
Figure 3.26 - The PLSDA plot for the model accuracy, LOO accuracy, true positives and true negatives associated with the LOO accuracies for the PLSDA model built using the H_3O^+ precursor spectra recorded using the mk.2 SIFT-MS machine from the headspace analysis of the serum samples from the wild badgers using tuberculosis culture status as the classification vector. The spectra data was mean centred and m/z ratios associated with the precursors were zeroed. 90
Figure 3.27 - The loadings plots of the first four latent variables of the PLSDA model built using the samples recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. The data was mean centred and m/z ratios associated with the precursors were removed..... 91
Figure 3.28 - The PLSDA curves for the model accuracy, LOO accuracy, true positives and true negatives associated with the LOO accuracies for the PLSDA model built using all three precursor spectra recorded using the mk.2 SIFT-MS machine from the headspace analysis of the serum samples from the wild badgers

using tuberculosis culture status as the classification vector. The spectra data was mean centred and m/z ratios associated with the precursors were zeroed. 92
Figure 3.29 - The PLSDA plots for the model accuracy, LOO accuracy, true positives and true negatives associated with the LOO accuracies for the PLSDA model built using the H_3O^+ precursor spectra recorded using the mk.2 SIFT-MS machine from the headspace analysis of the serum samples from the wild badgers using tuberculosis culture status as the classification vector. The spectra data was mean centred and m/z ratios associated ethanol, acetone and the precursors were zeroed. 93
Figure 3.30 - PLSDA LOO accuracies for models built using a range of sample subset sizes using the spectra recorded by the H_3O^+ precursor channel of the mk.2 SIFT machine from serum samples taken from wild badgers. Tuberculosis culture status was used as the classification vector. The data was mean centred and m/z ratios associated with the precursors were removed. 95
Figure 3.31 - A graph showing the PLSDA accuracies of models built and tested using four ratios of sample (60:40, 80:20, 90:10 and 95:5) from serum samples from wild badgers recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine and using the tuberculosis culture status as the classification vector. 10000 bootstraps were performed and the error bars on the graph show one standard deviation. 96
Figure 3.32 - A graph showing the PLSDA accuracies of models built and tested using four ratios of sample (60:40, 80:20, 90:10 and 95:5) from serum samples from wild badgers recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine and using a random permutation of the tuberculosis culture status as the classification vector. 10000 bootstraps were performed and the error bars on the graph show one standard deviation. 97
Figure 4.1 - The number of samples recorded for each time point for the *Mannheimia haemolytica* dataset showing the decrease in the numbers towards the end of the trial caused by fatalities 102
Figure 4.2 - The number of samples recorded at each sampling time point, showing how the trial was wound down after the clinicians found the introduction of the *Mycoplasma bovis* pathogen had failed to cause visible infection responses. 105

Figure 4.3 - The average m/z values for each of the groups of *Mannheamia Haemolytica* serum samples as recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. m/z ratios associated with the precursors have been removed. Blue is before introduction of the pathogen, turquise is post introduction group 1, yellow is post introduction group 2 and red if post introduction group 3 .. 108
Figure 4.4 - PCA scores plot of the *Mannheamia haemolytica* serum samples as recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. The m/z associated with the precursors and ammonia (18, 36, 54) have been removed. Blue dots are pre-introduction, black circles are post group 1, green pluses are post group 2 and red x are post group 3 as described in Table 4.2. 109
Figure 4.5 - The counts of m/z 65 for all the *Mannheamia haemolytica* serum samples as recorded using the H_3O^+ precursor spectra using the mk.2 SIFT-MS machine. The horizontal axis groups are those as described in Table 4.2. 110
Figure 4.6 - The counts of m/z 83 for all the *Mannheamia haemolytica* serum samples as recorded using the H_3O^+ precursor spectra using the mk.2 SIFT-MS machine. The horizontal axis groups are those described in Table 4.2. 111
Figure 4.7 - PLSDA plot for positive (0.5, 1 and 2 days after pathogen was introduced) verses negative samples (14, 7 and just before pathogen was introduced) of *Mannhemia haemolytica* serum samples recorded using all three precursor spectra of the mk.2 SIFT-MS machine. The blue line shows the accuracy of the model tested using samples used in its creation, the green shows the accuracy of a LOO validation which is then further broken down into the percentage of LOO true negatives and LOO true positives 112
Figure 4.8 - PCA of the *Mycobacterium paratuberculosis* serum sample spectra as recorded using the H_3O^+ precursor spectra of the mk.2 SIFT-MS machine. Samples numbered greater than 800 were recorded after the trial had concluded hence were not part of the randomisation procedure. m/z associated with the precursors have been removed and the data was pretreated using a mean centring algorithm. 113
Figure 4.9 - The average counts for the spectra of *Mycobacterium paratuberculosis* serum samples as recorded using the H_3O^+ precursor spectra of the mk.2 SIFT-MS machine. Blue bars are negatives and red bars are positives. 114

Figure 4.10 - PCA scores plot of the H_3O^+ precursor spectra from the samples of Mycobacterium paratuberculosis recorded using the mk.2 SIFT-MS machine. Samples numbered above 800 that were not recorded in a randomised order have been removed. m/z ratios associated with the precursors and ammonia (18, 36 and 54) have been removed and the data was pretreated using a mean centring algorithm. Approximate clusters have been draw outlining the slight discrimination. 115
Figure 4.11 - The loadings of PC1 for the PCA analysis of the *Mycobacterium paratuberculosis* cattle serum sample recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. Samples numbered above 800 that were not recorded in a randomised order were removed. m/z ratios associated with the precursors and ammonia (18, 36 and 54) were removed and the data was pretreated using a mean centring algorithm. 116
Figure 4.12 - PLSDA plot for the *Mycobacterium paratuberculosis* cattle serum samples recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. Samples number above 800 that were not recorded in a randomised order have been removed. m/z ratios associated with the precursors have been zeroed. The blue line indicates the accuracy of the model when tested against samples used in its construction, the green line indicates the accuracy of the LOO validation which is then further broken down into the LOO true negatives (red) and true positives (turquoise)..... 117
Figure 4.13 - The average spectra for the positive (red) and negative (blue) brucella in cattle samples as recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. m/z ratios associated with the precursors have been removed. 118
Figure 4.14 - PCA scores plot of the Brucella data collected by using the H_3O^+ channel of the mk.2 SIFT-MS machine showing one sample (no. 779) which stands out from the rest. m/z ratios associated with the precursors have been zeroed and the data was pre-treated using a mean center algorithm 119
Figure 4.15 – PCA scores plot of the Brucella serum sample data collected using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. m/z ratios associated with the precursors have been removed. Red samples are brucella positives and the blue samples are brucella negatives. The data was pre-treated using a mean center algorithm. An approximate cluster around the positive sample has been marked on the

graph. Approximately half the negative samples are within this cluster but they are more spread around the graph..... 120
Figure 4.16 - The levels of methanol (m/z 69) as recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine for serum samples of brucella..... 121
Figure 4.17 - PLSDA curves for the brucella data using data from all three precursor channels of the mk.2 SIFT-MS machine. The dark blue line shows the accuracy of the PLSDA model when tested with samples used in its production, the green curve shows the accuracy of the PLSDA as produced by a LOO validation and these results are further broken down into the LOO true negatives (red) and true positives (turquoise)..... 122
Figure 4.18 - The average counts for the Mycoplasma bovis data groups spectra as recorded using the H3O+ precursor channel of the mk.2 SIFT-MS machine. Before introduction of the pathogen are blue, group 1 are turquiose, group 2 are yellow and group 3 are red with the samples belonging to each group expained in Table 4.3. The m/z ratios associated with the precursors have been removed 123
Figure 4.19 - PCA score plot of the *Mycoplasma bovis* dataset using the H_3O^+ precursor spectra of the mk.2 SIFT-MS machine. Blue dot are before the pathogen was introduced, black o are post introduction group 1, green + are post introduction group2 and red x are post introduction group3. No discrimination can be seen apart from the single outlier. 124
Figure 4.20 - The intensities of m/z 18 as recorded using the H_3O^+ precursor of the mk.2 SIFT-MS machine for all the samples of serum for the *Mycoplasma bovis* dataset. The horizontal axis represents the groups that the samples belong to as descibed by Table 4.3..... 125
Figure 4.21 - PCA score plot for the bovine TB (*Mycobacterium bovis*) samples as recorded by the H_3O^+ precursor channel of the mk.3 SIFT-MS machine. m/z ratios associated with the precursors and ammonia (18, 36 and 54) have been removed. The data was pre-treated using a mean centring algorithm. Blue dot are time point 1, black o are time point 2, green + are time point 3 and red x are time point 4. A dashed line has been drawn across the graph at the origin of the first principal component axis. 126

Figure 4.22 - Plot of the average m/z counts for each sample group of the bovine TB (*Mycobacterium bovis*) serum control samples as recorded using the H_3O^+ precursor channel of the mk.3 SIFT-MS machine. m/z ratios associated with the precursors have been removed. The blue bars represent the m/z levels before the pathogen was introduced and the red bars represent the m/z levels after the pathogen was introduced. 128
Figure 4.23 - PCA scores plot of the control groups serum samples from the bovine TB (*Mycobacterium* bovis) clinical trial measured using the newer mk.3 SIFT-MS machine using only the H3O+ precursor spectra. Blue dot are negative samples and red x are positive samples. m/z associated with ammonia (m/z 18, 36, 54 and 72) and acetone (m/z 59 and 77) have been zeroed. The origin of PC1 has again be drawn with a dashed line..... 129
Figure 4.24 - The levels of m/z 69 which is associated with methanol for the serum samples from the Tuberclosis bovis control group with negative samples being before infection and the positive samples being after infection 130
Figure 4.25 - PLSDA plot for the data collected tuberculosis serum cattle control group samples showing a quickly overtrained model suggesting that there are too few samples. The dark blue line shows the accuracy of the model when tested with samples used in its creation, the green line shows a LOO accuracy which is broken down into the number of LOO true negatives (red) and LOO true positives (turquiose)..... 131
Figure 4.26 - PLSDA plot for the bovine TB serum from the control samples only using only H_3O^+ precursor channel of the mk.3 SIFT-MS machine and after m/z associated with ammonia (m/z 18, 36, 54 and 72), acetone (m/z 59 and 77) and propanol (m/z) have been removed. The dark blue line shows the accuracy of the model when tested with samples used in its creation, the green line shows a LOO accuracy which is broken down into the number of LOO true negatives (red) and LOO true positives (turquiose). 132
Figure 5.1 – The average H_3O^+ spectra for *Mycobacterium paratuberculosis* urine samples as recorded using the mk.3 SIFT-MS machine. Counts associated with the precursors have been zeroed. The blue spectrum is the average of the negative samples and the red is the average of the positive samples. 139

Figure 5.2 - The m/z counts for ammonia (18) for all infected (red) and all healthy (blue) *Mycobacterium paratuberculosis* urine samples as recorded using the H_3O^+ precursor of the mk.3 SIFT-MS machine showing the large variance 140
Figure 5.3 - PCA scores plot showing that no discrimination can be seen between the sample types when all the data from the H_3O^+ precursor spectra of the mk.3 SIFT-MS machine is used. No pre-treatment methods were performed but m/z ratios associated with the precursors have been removed. Blue crosses are *Mycobacterium paratuberculosis* negative samples and red dots are *Mycobacterium paratuberculosis* positive samples 141
Figure 5.4 - The loadings of PC1 for the above PCA analysis using the H_3O^+ precursor spectra from the mk.3 SIFT-MS machine. No pre-treatment methods were used but the m/z ratios associated with the precursor ions were zeroed. m/z peaks associated with ammonia (18, 36 and 54) clearly dominate the variance. 142
Figure 5.5 - PCA scores plot of the H_3O^+ urine spectra from cattle taken from herds in three farms. Red dots are *Mycobacterium paratuberculosis* positive samples and the blue crosses are the negative control group. Peaks with m/z associated with ammonia (18,36,54,72), acetone (59,77,95), propanol (79) and some others (35,46,53,64) have been removed. m/z peaks associated with the precursors have been removed and the data was pre-treated using a mean centre algorithm. 143
Figure 5.6 - PLSDA curves for the *Mycobacterium paratuberculosis* urine samples as recorded using the newer mk.3 SIFT-MS. Only data from the H_3O^+ precursor channel is being used. m/z associated with ammonia (18,36,54,72), acetone (59,77,95), propanol (79) and some others (35,46,53,64) have been removed. The blue line shows the percentage of correctly identified samples when all samples are used to build the model, the green line shows the percentage of correctly identified samples using a LOO validation process and these results are broken down into true positives (light blue) and true negatives (red). 144
Figure 5.7 - The average counts for bovine TB (*Mycobacterium* bovis) urine samples for before infection (blue) and after infection (red) as recorded using the H_3O^+ precursor spectra of the mk.3 SIFT-MS machine. m/z peaks associated with the precursors have been removed. 146

Figure 5.8 - PCA scores plot of the Bovine TB urine samples as analysed using the H_3O^+ precursor channel of the mk.3 SIFT-MS machine. Blue dots are control samples and red crosses are BCG samples. 147
Figure 5.9 - The counts of m/z 77 (acetone) for urine samples from the bovine TB (*Mycobacterium bovis*) trial as recorded by the H_3O^+ precursor of the mk.3 SIFT-MS machine. Samples are across the horizontal axis separated into the different time points of the trial 148
Figure 5.10 - PCA scores plot of the cattle urine bovine TB (*Mycobacterium bovis*) samples as recorded using the H_3O^+ precursor channel with m/z associated with precursors and ammonia (17, 18, 35, 36, 53, 54, 72), acetone (77), and some others (52, 58, 60, 78) removed. Red x are post-infection and blue dot are pre-infection. The data was pre-treated using a mean centring algorithm. 149
Figure 5.11 - The loadings of PC1 for the above PCA analysis of the cattle urine bovine TB samples as recorded using the H3O+ precursor channel with m/z associated with the precursors, ammonia (17, 18, 35, 36, 53, 54, 72), acetone (77) and some others (52, 58, 60, 78) removed. 150
Figure 5.12 - PLSDA LOO analysis of the cattle urine bovine TB (*Mycobacterium bovis*) samples as recorded using the H_3O^+ precursor channel with m/z associated with precursors and ammonia (17, 18, 35, 36, 53, 54, 72), acetone (77), and some others (52, 58, 60, 78) removed. The blue line is the model accuracy and the green line is the LOO validation accuracy which is then further broken down into the LOO true negatives (red line) and true positives (light blue line) 151

List of Tables

Table 2.1 - The H_3O^+ spectra m/z values expected for the compounds used to create the mixtures. Mixture one (A) contains ammonia, methanol and acetone. Mixture two (B) contains ethanol, acetonitrile and toluene. 23

Table 2.2 - The m/z peaks caused by the precursors on each of the three spectra recorded by the SIFT-MS machine when analysing aqueous samples. There is precursor crossover between the channels (e.g. m/z 19 is H_3O^+ and this is visible on all three channels) (Smith and Spanel, 2005). 25

Table 4.1 - Overview of the cattle serum datasets including the analysis machine, trial type and basic info about the number of samples 101

Table 4.2 - The grouping of the *Mannheimia haemolytica* samples for analysis with multivariate tools 103

Table 4.3 - A table showing the groupings of the *Mycoplasma bovis* data that formed the classification vector to be used in the display of the unsupervised multivariate analysis 106

Abbreviations

Here is a list of the abbreviations and notations used in this thesis.

Anova	Analysis of variance
Badger	*Meles meles*
BCG	Bacillus Calmette-Guérin
Bovine TB	*Mycobacterium bovis*
DEFRA	Department for the environment, food and rural affairs
ELISA	Enzyme linked immunosorbent assay
e-nose	Electronic nose
EU	European union
FLI	Friedrich-Loeffler-Institut
FSM	Full spectra scan mode
GC-MS	Gas chromatography mass spectroscopy
IAH	Institute of animal health
Log	Logarithm
Log_{10}	Base 10 logarithm
LOO	Leave one out
LV(s)	Latent variable(s)
m/z	Mass to charge ratio
MIM	Multiple ion monitoring mode
Mk.	Mark (version)
NIPALS	Non-linear iterative partial least squares
NN	Artificial neural network
paraTB	*Mycobacterium paratuberculosis*
PC(s)	Principal component(s)
PCA	Principal component analysis
PCR	Polymerase chain reaction
PLSDA	Partial least squares discriminant analysis
Possum	*Trichosurus vulpecular*
s.d.	Standard deviation
SIFT-MS	Selected ion flow tube mass spectroscopy
SVD	Singular value decomposition
SVM	Support vector machine
TB	Tuberculosis or *Mycobacterium tuberculosis*

UK	United Kingdom
VLA	Veterinary laboratories agency
VOC(s)	Volatile organic compound(s)

1. Introduction and Literature Review

1.1. Introduction

Selected ion flow tube mass spectroscopy (SIFT-MS) has been around for many years and was originally used for studying the ion-neutral reactions of high altitude plasmas (Adams and Smith, 1976). It has been used for biological analysis for approximately 15 years (Smith and Spanel, 1996). The technique allows the semi-quantitative investigation of the concentrations of volatiles in gas samples.

The data received from the machine is in one of two formats. Firstly, the most accurate method of investigation is by focusing on a single m/z ratio. At near real-time, levels can be monitored using this technique (Smith et al., 2008). This technique is useful when the investigator has a clear idea about the specific volatile or m/z ratio that is of interest.

The second method uses a switching method to scan through a range of m/z ratios to build up a picture of the whole range of volatiles in the sample. This method is less accurate than the first and also does not produce results in real time. It does however allow the investigator to find previously unknown markers for disease and look for unexpected volatiles.

Classically, the second method is used to explore volatiles within samples and it produces three spectra, each corresponding to one of the precursors used as a carrier for the sample under investigation. These spectra are then painstakingly analysed by hand looking at the levels of the trace volatiles compared to the levels normally expected. The problem, however, is that the levels of volatiles in biological samples can vary by several orders of magnitude (Turner et al., 2008). Finding markers is therefore a non-trivial task.

Multivariate analysis techniques have been used before in similar cases such as electronic nose (e-nose) (Scott et al., 2006), gas chromatography mass spectroscopy (GC-MS) and other types of spectroscopy (Goodacre et al., 2004; Dettmer et al., 2007) so it seems odd that no clear method has been proposed for SIFT-MS. This thesis aims to fill this void by proposing multivariate methods for SIFT-MS data analysis and evaluating them firstly using synthetic mixtures of compounds known to be present in biological samples and then applying them to real biological examples (badger serum, cattle serum and cattle urine) so that their effectiveness can be evaluated and discussed.

1.2. Literature Review

A wide range of multi-disciplinary ideas and topics have been used throughout the progression of this project. The literature review has been split into three sections covering the basics of the three topics required for the understanding of the results presented. Firstly, a background knowledge of the diseases is required to validate the need for this research. This includes the standard methods of detection and where applicable previous applications of volatile organic compound detection methods in the determination of disease status.

Secondly, a basic understanding of the SIFT-MS procedure is outlined. This is the analytical tool that creates the data and makes the project possible. An understanding of how the data is created is key to understanding how and why the data needs to be processed in order to reveal the required biological knowledge.

Finally, there is a section on the statistical tools and procedures that will be investigated to produce a methodology for the analysis of the data. This section covers the data preparation, exploration and classification tools that could be applied.

1.2.1. Diseases Studied

During the early planning stages of the project, one of the first diseases suggested that would be suitable to investigate the possibility of SIFT-MS linked with multivariate techniques as a technique to investigate veterinary diseases was *Mycobaterium bovis* (commonly known as bovine TB). Bovine TB shares many similarities with the human form of the disease *Mycobacterium tuberculosis*. Tuberculosis is one of the important diseases in the creation of modern medicine, with more than one Nobel prize being awarded for research of the bacteria and possible treatments/vaccines. Knowledge of the disease has been around since Hippocrates and yet a cure or accurate method for detection has yet to be discovered and both are desperately required (Perkins and Cunningham, 2007).

The bacteria have hard waxy outer layers that protect them against the body's natural defences. The bacteria are very slow growing (multiplying approximately once every 16 hours compared with 40 minutes for *E.coli* (Cox, 2004)) and hence the disease takes a long time to progress. In general, the bacteria infect the host's lungs and the only defence is for the macrophages to consume and immobilise the bacteria creating legions called granulomas. These legions form the basis for the gold standard test in cattle (Bourne, 2007; White et al., 2008). They are dissected (post mortem) and bacteria are cultured for up to 16 weeks and then a visual inspection (microscopy) will confirm or deny infection.

Alternative tests are needed that are both rapid and accurate (Perkins and Cunningham, 2007). In agriculture these tests will save farmers (and the government) money (Krebs, 1997; Bourne, 2007). The goal is to decrease the quarantine time that a herd is subjected to between suspected infection and confirmation. There is also a large possibility that any method of detection of the bovine version of the disease can also be used to detect the other strains, which also cause massive problems worldwide. Two possible options being considered are using tests that look for the specific gamma interferon immunity marker (Dalley et al., 2008) and using the scent of the bacteria to detect them (Pavlou et al., 2004; Fend et al., 2005; Fend et al., 2006;

Syhre and Chambers, 2008; Syhre et al., 2009). Until recently, detection using the second method has focused on the application of MOSFET and conducting polymer electronic noses, which despite the early promise have not fully delivered the expected results (Persaud and Dodd, 1982; Dickinson et al., 1998; Fend, 2004). One possible reason might be the inherent variability in the results obtained by different electronic noses and also the sensor drift (Knobloch et al., 2009c). Recently, some discrimination between infected and non-infected samples has been seen when investigating sputum samples collected from patients with tuberculosis using a separation technique and then analysis on GC-MS (Kaal et al., 2009).

Bovine TB generally infects cattle, however, other species can be cross-infected. These species include badgers (*Meles meles*) and deer in the UK and brush-tailed possums (*Trichosurus vulpecular*) and ferrets in New Zealand (White et al., 2008).

Recently, some biomarkers for tuberculosis were detected using GC-MS (Syhre and Chambers, 2008; Syhre et al., 2009). They were found by looking at the headspace above concentrated cultures. However, there was not one standard biomarker found, but several with a range of levels depending on the media that the cultures were grown on.

TB is not the only respiratory diseases in cattle, there are others that cause pneumonia which can be very costly to both the farmer and the state (Snowder et al., 2006). The Friedrich Loeffler Institute (FLI) have been investigating *Mannheimia haemolytica* and *Mycoplasma bovis* which are two known causes (Angen et al., 2009).

Mycoplasma bovis can cause many other visible effects in cattle such as mastitis and arthritis and due to the long incubation time (3 months) it is known as a chronic disease (Nicholas and Ayling, 2003). Like tuberculosis, the disease is spread by animals that are visually healthy but are shedding the bacteria via aerosol like coughs, sneezes and mucus discharges. *Mycoplasma bovis* shares many clinical and pathological symptoms with other cattle respiratory diseases and hence laboratory diagnostics are necessary for correct identification. Enzyme-linked immunosorbent

assays (ELISA), antibodies and polymerase chain reaction (PCR) (Angen et al., 2009) have all been used to detect the disease each with their own drawbacks. Antibody detection only detects after the disease has become clinical and the host has had an immune response and hence is likely to miss positive cases of infection whilst they are in the early stages of the disease. The antibodies are also present in high levels for months after the disease. PCR requires bacterial presence within the sample and is hence very sample dependent. For example, samples that could possibly contain *Mycobacterium bovis* are usually heat treated to ensure that infection from the samples is no longer possible. (Angen et al., 2009) used a catheter tube to collect sputum from the cattle directly (an invasive procedure).

Mannheimia haemolytica is not a chronic disease. Before sequencing the bacteria was classified as part of the pasteurella group of bacteria (Ackermann and Brogden, 2000), as the bacteria can be found in the milk of the infected cattle. It is a very rapidly progressing infection that causes pneumonia like symptoms only hours after infection (Ackermann and Brogden, 2000). The toxins build up in the lungs and these are usually associated with the cause of the high mortality rate. *Mannheimia haemolytica* is an opportunist infection relying on the weakened state of the host animal and generally does not cause symptoms in healthy animals (Duff and Galyean, 2007).

Brucellosis is another disease with historical references dating back to Hippocrates. Brucellosis in humans is generally linked to the drinking of unpasteurised milk from infected cattle or goats (Cutler et al., 2005). The disease has been eradicated from the UK but not from mainland Europe and hence tight controls on animal import and export are required (England et al., 2004; Cutler et al., 2005). The main strain of disease affecting cattle and causing problems for the veterinary laboratories agency (VLA) is *Brucella abortus*, which causes abortions of pregnancies in the cattle. Diagnostic tests are available that are highly effective but with so few cases, the cost of the disease is linked to the false positive rate (England et al., 2004; Cutler et al., 2005). Any method that can discriminate between the real infected samples and false positives could save the government money.

Mycobacterium paratuberculosis is an intestinal disease that affects many animals, both domestic and wild (Cocito et al., 1994). It has been linked with Crohn's syndrome in humans (Chamberlin et al., 2001). In cattle, the animals are usually infected at an early age, but the disease is slow progressing and chronic with clinical signs (such as diarrhoea) not present until age 2 to 5 years (Stabel, 1998). Bacteria are shed from the host at very high levels and hence once a herd is infected the disease spreads rapidly.

1.2.2. Analytical Tools

All the analytical data created for this project has been created by one of two selected ion flow tube mass spectrometers (SIFT-MS) machines. The SIFT-MS technology was originally devised to investigate high altitude terrestrial plasmas (Adams and Smith, 1976; Smith and Adams, 1980). The technology has more recently been adapted to investigate biological volatile gases (Smith and Spanel, 1996). Breath is generally studied, but volatiles from other bodily fluids such as sweat (Turner et al., 2008) and urine (Smith et al., 1999) have been studied.

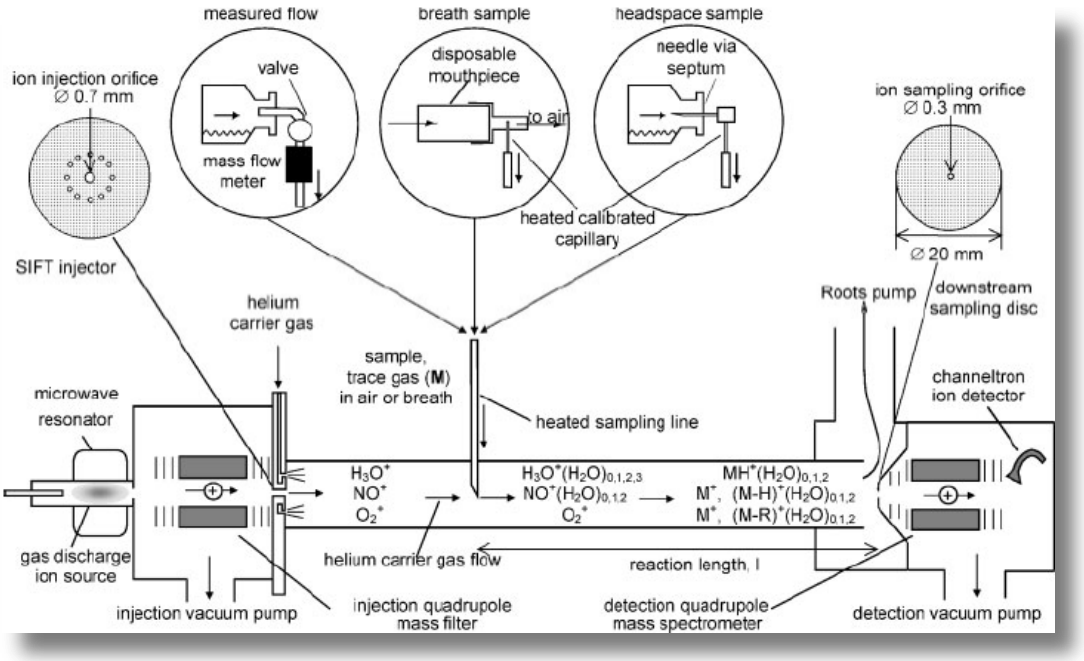


Figure 1.1 - A schematic of the internal workings of a SIFT-MS machine (Smith and Spanel, 2005). The ions written in the tube are the precursor ions (upstream) and possible precursor-sample interactions (downstream). Different sample preparation techniques are also suggested.

The SIFT-MS machine (Figure 1.1) has two operation modes. The first called full spectra scan mode (FSM), where the machine switches the final quadrupole detector through the full range of m/z values. This is the most useful mode for working with samples that have unknown composition. The other mode is called multiple ion monitoring mode (MIM), where the final quadrupole is given a few specific m/z values to monitor and real time recordings are made. The second mode has been used to detect how the volatiles change during the cycle of breathing (Smith et al., 2008). It has also been used to detect changes in acetaldehyde levels in the headspace above cancer cell lines (Sule-Suso et al., 2009). However, the basic process of detection is the same no matter which method is being used.

Positively charged ions are created in an ion source. A quadrupole mass filter is then used to select the required ion which once selected are called precursors. The SIFT-MS machine allows the user to choose one of three precursors, H_3O^+ (m/z 19), NO^+

(m/z 30) and O_2^+ (m/z 32). These precursors are then fed into a fast flow tube that has a helium (inert) carrier gas. A neutral sample of volatiles is fed into the flow tube at a measured rate via a heated (40°C) capillary tube. The samples can be fed into the tube using a variety of methods including a pass-through breathing tube (as used by the police breathalyser), dynamically sampled from the headspace above a vial or taken from a deflating bag.

The volatile organic compounds (VOCs) interact with the precursors and form reaction products which are carried down the flow tube towards another quadrupole mass filter. This time however, the products are sent to an ion detector. Using the measurement of the size of the flow tube, the speed of the fast flow and a few equations (Smith and Spanel, 2005) a rate constant can be calculated for every measurable volatile which can then be used to calculate the ratio between the measured number of precursor ions with no sample present and the combination of the precursor ions and sample ions. This ratio (with a few assumptions being made) can then be used to calculate the exact concentration of a specific VOC in the sample under investigation.

This makes the SIFT-MS a potentially quantitative mass spectrometer. However, due to the complexities and assumptions made it is very rarely used as such and is generally assumed to be semi-quantitative. Whilst this would not be suitable for a detailed analysis of a single sample, when coupled with multivariate techniques semi-quantitative analysis been shown to be suitable for disease diagnosis using data from other analytical methods (Zlatkis et al., 1981; Di Natale et al., 2003; Wang et al., 2003; Fend, 2004; Chen et al., 2005; Fend et al., 2005; Fend et al., 2006; Perez Pavon et al., 2006; Gaspar et al., 2009; Kischkel et al., 2010).

As all of the samples investigated in this thesis have unknown volatile compositions only the full spectra scan mode (FSM) is used.

1.2.3. Alternative technologies

The two most comparable technologies to SIFT-MS are electronic nose and GC-MS. Electronic noses have shown promise in this area but repeatability is questionable due to sensor drift. The main benefits of electronic noses are the speed that samples can be measured and the relatively small costs involved. Sample preparation is minimal and depending on the sensors used a wide range of volatiles can be monitored. GC-MS techniques are much more repeatable. Indeed, standard repositories of GC-MS responses are stored (such as the NIST database www.nist.gov). However, obtaining results is much more time consuming therefore obtaining datasets with are large number of samples is costly. This is further compounded by the sample preparation required.

Electronic noses were created to mimic human olfaction (Persaud and Dodd, 1982), and have since been used to detect differences in wines and beers (Di Natale et al., 1996; Ciosek et al., 2006; Cozzolino et al., 2006), detect disease (Di Natale et al., 2003; Fend, 2004; Pavlou et al., 2004; Fend et al., 2005; Fend et al., 2006), however, repeatability with electronic noses has been questioned (Knobloch et al., 2009c). The main problems come from the inherent instability of the sensors involved. Most of the above analysis has been performed using the original FET/MOSFET style electronic noses as developed by (Persaud and Dodd, 1982) or the newer conducting polymer based sensors (De Lacy Costello et al., 1996; Stussi et al., 1996) both of which have drawbacks.

FET/MOSFET sensors are potentiometric and are prone to drift and poisoning. Manufacturing allows for discrepancies in the conductive changes that make them useful olfactory sensors (Gopel, 1995; Albert et al., 2000). Conducting polymer sensors are supposedly less prone to drift, however the manufacturing process utilises random perturbation to give the sensors different responses (Stussi et al., 1996). This makes manufacturing two identical sensors near impossible (James et al., 2005). Both of these electronic nose technologies are not quantitative and rely on the successful identification of patterns alone.

Of course alternative electronic nose technologies exist that are not affected by the repeatability issues discussed above. Technologies based on QCM (quartz crystal microbalance) and SAW (surface acoustic wave) use mass based detection methods (Hauptmann et al., 1993; D'Amico et al., 1998; Chen et al., 2005). These methods are described as semi-quantitative. These have shown promise in detecting volatiles, however, the data returned from the machines is more complex and analysis is difficult. These technologies are also hampered by a larger problem; the analysis time per sample can be from several minutes to a couple of hours. This means that these methods are more akin to GC-MS than electronic noses which might be why the technologies are not widely implemented as they have no clear advantage.

GC-MS is also suitable for the detection of volatiles (Pasikanti et al., 2008; Issaq et al., 2009). Sampling times are dependent on column type and MS detector type. Fast GC-MS sample times can be performed in the five minute range (comparable to SIFT-MS) however, the resolutions and separations are poor at these retention lengths. The largest benefit to using GC-MS over SIFT-MS would be the superior data quality from 2D chromatography (GC-GC-MS). However, this is only obtained with much longer analysis time approaching 60 minutes (Shellie et al., 2005). GC-MS methods are generally quantitative however standards are sometimes required.

Breath samples are more difficult to analyse with GC-MS. They need to be concentrated and stored in tubes for example using solid phase micro-extraction (SPME) (Grote and Pawliszyn, 1997) or similar methods (Harper, 2000). The gaseous samples are forced through tubes lined with a sorbent material. These tubes are then attached to the GC-MS and the sample is desorbed by heating and the analysis recorded.

Finally, NMR and HPLC are suitable methods for the analysis of the liquid samples themselves (D'Imperio et al.; Cimato et al., 2006). These methods would however, remove the possibility of analysing aerosol based samples such as breath. From an

analytical perspective this is not a problem, however, it limits the possibility of developing a non-invasive detection method.

1.2.4. Data Analysis Methods

The standard approach to analysing the data is to use the SIFT-MS software provided with the instrument. Indeed, most published work utilises the MIM of the SIFT-MS (Turner et al., 2006a; Turner et al., 2006b; Turner et al., 2006c; Sule-Suso et al., 2009) and the software is ideally suited to this analysis.

As the volatiles linked to the diseases being studied are not known this method is not suitable for analysis. Instead the SIFT-MS machine will be used in FSM, to give as much information about the sample as possible. This produces standard count verses m/z graphs (Figure 1.2) similar to those achieved using other mass spectroscopy techniques.

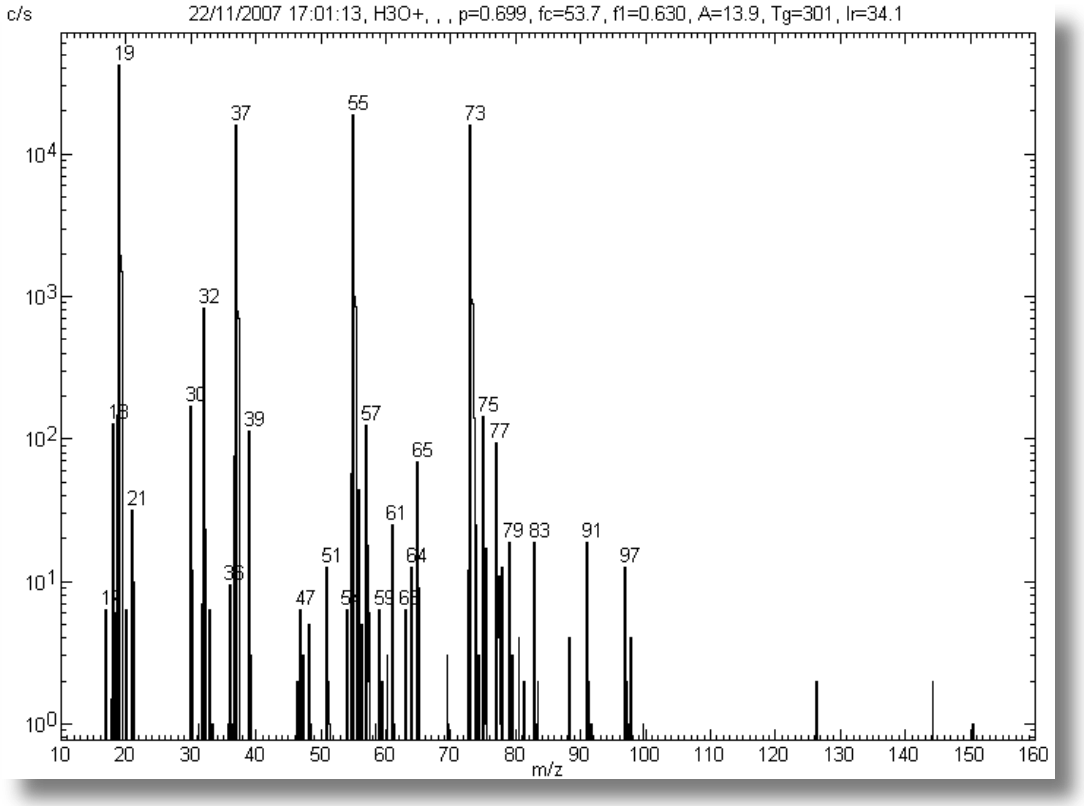


Figure 1.2 - An example of a SIFT-MS FSM spectrum as produced by the standard software shipped with the machine.

The software allows various method of plotting the data and allows the user to change the way the peaks are calculated by adding correction values. For example, if the user believes that the final detector is out of alignment they can move the point where the peak heights are calculated to a fractional m/z value and new peak heights will be created. Alternatively, if the user wants to compare a spectrum from a short sampling run with a longer run they can use the software to estimate the counts that would have been recorded over the longer scan time. It, however, only allows the viewing of one spectrum using one precursor at any one time (Figure 1.2). However, a useful feature is that the software will open all the files in a folder and export them to an Excel style spreadsheet. This allows the analysis of the data in alternative software.

The software chosen depends on the data format and the analysis to be performed. Simple line graphs can be drawn in Excel or any other spreadsheet software.

However, the analysis for the SIFT-MS is potentially more complicated than can easily be handled in Excel. There are software tools available for multivariate analysis such as SPSS, however, the methods are usually implemented by wizards and there are very few ways to firstly optimise and tweak the method and secondly script and automate some of the more repetitive stages of the analysis.

Matlab (Mathworks, USA), has for a long time been the choice for many scientists and engineers for working with matrices and spreadsheet-like data formats. There are many toolboxes that have been created to help with multivariate data analysis. One such toolbox is the PLS_Toolbox (Eigenvector Research, USA), which has tools for performing unsupervised analysis in the form of principal component analysis (PCA) and supervised predictive analysis in the form of partial least squares discriminant analysis (PLSDA). Other toolboxes exist for neural networks (ANN), support vector machines (SVM) and clustering methods.

The SIFT-MS data, as shown in the above figure, is in the form of counts for the range of m/z values exported from the software. Three tables of these counts can be collected, each table has a single vector for every sample and there are three precursors, one table for each. Other information is available as well, including the correction values for the flow rate.

When there are a group of samples under investigation (instead of just a single sample) one way of looking for information to use is to separate the samples into groups and then look for m/z that have counts for only one group of samples. This can be taken further by averaging the counts for each m/z across one group and then looking for any large differences between the groups.

In proteomics and transcriptomics, large numbers of m/z or genes remove the possibility of looking at the differences by hand and statistical tests such as students t-test (t-test) are used and then the genes or peptides that have a significant difference are ranked and displayed (Dewe, 2009). At some point, the usefulness of further univariate testing (even automated scripted testing such as ANOVA (Otto, 1999a))

becomes limited. Univariate approaches only consider a single feature (gene, peptide or m/z) at a time.

Many data sources contain variables that are linked and indeed the biological factors under investigation might be linked. In the case of the SIFT-MS machine a volatile can have multiple peaks on a single spectrum due to the interactions with the precursor ions (Smith and Spanel, 2005). There are also many cases in biology where we have qualitative traits that are caused by multiple factors. In these cases, there might not be a single marker for the disease or condition. A subtle change in a combination of factors might be the key to detecting the disease. In this situation the univariate approach may not be suitable and multivariate techniques may hold the key.

One well used multivariate method is principal component analysis (PCA) (Otto, 1999a; Jurs et al., 2000). PCA analysis takes the data and decomposes it into eigenvectors and eigenvalues. The data starts in the form of a matrix, $\boldsymbol{X}_{[n,m]}$, that has *n* rows and *m* columns corresponding to *n* samples and *m* variables. Using a method such as non-linear iterative partial least squares (NIPALS) or singular value decomposition (SVD) (Wu et al., 1997) a loadings matrix is obtained, $\boldsymbol{L}_{[m,p]}$, and a score matrix, $\boldsymbol{S}_{[n,p]}$ (where p is the number of components retained) and a residuals matrix, $\boldsymbol{E}_{[n,m]}$. The scores matrix is the important part in a data visualisation procedure as the plotting of two (or three) components gives an idea of the spread of the data. Factors contributing the most variance to the dataset are found in the earliest components. The higher the component number, the more noise that is likely to be associated. The loadings, however, are also useful as they provide insight into the factors that are most affected by the causes of the largest variance. If desired, the original data can be reconstructed (Eq. 1.1).

$$\boldsymbol{X}_{[n,m]} = \boldsymbol{S}_{[n,p]} \times \boldsymbol{L}_{[p,m]}{}' + \boldsymbol{E}_{[n,m]}$$

Eq. 1.1

where the loadings matrix has been transposed.

Of course, one of the assumptions of PCA is that the important factors are the ones causing the largest variance. In simple situations this may hold true. However, in many biological datasets there can be many factors that contribute to the variance such as age, diet, sex that have nothing to do with the factor under investigation (disease state for example). PCA is unsupervised, which means that any prior knowledge about these factors is not used by the algorithm.

Partial least squares discriminant analysis (PLSDA) (Barker and Rayens, 2003), takes the idea of PCA but adds information of the required class discrimination. It uses calculations of among-groups sum of squares to maximise the co-variance between the sample groups. The maths is complex and beyond the scope of this literature review but proof of the theorem can be found in (Barker and Rayens, 2003).

The problem comes with analysing the results. In PCA, components from the score matrix can be plotted to produce the typical PCA plots. PLSDA produces latent variables which are similar to the principal components of a PCA analysis, however, the plots produced from them are almost meaningless (Westerhuis et al., 2008). In fact, many papers were published damning the results of many PLSDA analyses (Brereton, 2006; Broadhurst and Kell, 2006). The problem comes from the ease of over-training the models. Indeed, it has been shown that good visible discrimination on the PLSDA plot can be obtained from random numbers (Brereton, 2006). The number of samples required to build a model increases exponentially with the number of variables and hence the majority of datasets analysed have just too few samples. The PLSDA analysis problem becomes one of validation.

Many ideas about validating PLSDA models are given, however, no group consensus has been agreed. Some validation methods use strict principles of division of the data into groups of samples for training, optimisation and testing (Westerhuis et al., 2008), others use the idea of repetitions (also known as bootstrapping) to get a statistical average result (Brereton, 2006).

Another problem that arises is deciding which metric to use to determine the performance of the validated model. In a simple two class problem there are four possible outcomes. A sample was positive and it was correctly classified (true positive), a sample was positive and was incorrectly classified (false negative), a sample was negative and correctly classified (true negative) and finally a sample was negative but incorrectly classified (false positive). The approaches are to look at the number of incorrectly classified samples (or the inverse), the sensitivity, the selectivity, the specificity and many more combinations (Broadhurst and Kell, 2006). There are arguments for the use of each but also flaws.

PLSDA is not the only type of supervised method available. Indeed artificial neural networks (NN) have been used with electronic nose discrimination (Gutierrez-Osuna, 2002; Ali et al., 2003; Aleixandre et al., 2004). Support vector machines (SVM) offer another possibility (Cortes and Vapnik, 1995) and indeed show better results when combined with a large number of samples and a large number of variables with efficient feature selection (Mahadevan et al., 2008). It is important to remember that the points made about validation and sample size are just as valid for NN and SVM (Brereton, 2006; Broadhurst and Kell, 2006).

One feature of all the above methods is the requirement that the data used is properly prepared for analysis. Generally, many simple pre-treatment algorithms are commonly used (Dickinson et al., 1998; Otto, 1999a; Pardo and Sberveglieri, 2002; Scott et al., 2006), including but not limited to mean centring which sets the mean of each variable to zero to remove any directional bias (Eq. 1.2).

$$x'_{ij} = x_{ij} - \bar{x}_j$$

Eq. 1.2

where x' is a new matrix of i rows and j columns, x is the original matrix of the same dimensions and $\bar{x}_j$ is the mean of each column. Auto-scaling is alternative method which performs mean centring (Eq. 1.2) but also divides by the standard deviation removing the magnitudes of the different variables (Eq. 1.3).

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_{x_j}}$$

Eq. 1.3

where x' is a new matrix of i rows and j columns, x is the original matrix of the same dimensions, $\bar{x}_j$ is the mean of each column and σ_{x_j} is the standard deviation of each column. And finally, range scaling exists in many forms but all of which scale each variable so that the maximum and minimum values are the same for all variables (it can also be used as a normalisation technique where every sample is scaled). In the below equation (Eq. 1.4), the limits are 0 and 1.

$$x'_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)}$$

Eq. 1.4

where x' is a new matrix of i rows and j columns, x is the original matrix of the same dimensions, $\min(x_j)$ is the minimum value of each column and $\max(x_j)$ is the maximum value of each column.

Another possibility is the taking of log values of the counts. In human olfaction the nerve responses have a logarithmic relationship to the concentration of VOCs reaching the nerve endings (Gopel et al., 1998). As one of our applications is looking at VOCs that are causing the specific odour of tuberculosis (Pavlou et al., 2004; Syhre and Chambers, 2008) there is the possible modification of the mean centred equation to include taking the logs. Indeed, the software for the SIFT-MS plots the peaks on a log graph (see Figure 1.2). With this in mind the mean centring equation (Eq. 1.2) can be modified to work with the logarithmic values instead (Eq. 1.5).

$$x'_{ij} = \log(x_{ij} + 1) - \overline{\log(x_j + 1)}$$

Eq. 1.5

The addition of 1 to the variables is to avoid the problem of taking the log of zero.

1.2.5. Disease detection using multivariate techniques and volatile analysis

No previous combination of SIFT-MS and multivariate techniques has been reported looking at disease detection and discrimination of samples. However, successful discrimination has been seen with both electronic noses and GC-MS.

The most applicable body of work is the detection of *Mycobacterium tuberculosis* using electronic nose technology (Pavlou et al., 2004; Fend et al., 2006). In this work, electronic nose technology was used to detect *M. tuberculosis* cultures grown from human patients. The electronic nose was used to measure all the samples and then a neural network was built that accurately classified the samples. Unfortunately, this is only part of the process. Growing tuberculosis cultures is a time consuming method of detection and obtaining the bacteria is difficult. In cattle and bovines, obtaining the granuloma from the lungs to perform this test is generally done post mortem (Krebs, 1997; Bourne, 2007; White et al., 2008).

Work has also been completed that shows that disease detection is possible when bacterial evidence is not present. Markers for lung and breast cancer have been found in breath volatiles using GC-MS (Phillips et al., 2003; Kischkel et al., 2010; Phillips et al., 2010). Breast cancer causes certain over stressed enzymes to produce known volatiles that have been detected in the breath of patients. Monte Carlo simulations were then used to detect them in GC-MS time slices. A sensitivity of 78.5% and specificity of 88.3% were recorded on the build dataset which dropped to a sensitivity of 75.3% and specificity of 84.8% on the validation dataset (Phillips et al., 2010). Lung cancer has also shown similar potential for a breath based detection method, however, due to the complexities and multiple causes the markers are less completely documented (Kischkel et al., 2010). Multivariate methods have shown great potential in this area (Gaspar et al., 2009).

Due to the complexities of breath analysis (Phillips et al., 1999) alternative samples have been investigated for the detection of lung cancer including the use of urine samples. Models involving mice have shown that detection of lung cancer is possible

from the analysis of volatiles above urine headspaces, hence a possible alternative to breath analysis (Matsumura et al., 2010).

Finally, a large body of research has used serum and blood samples in the search of biomarkers and classification methods. Some of this work has been criticised because of poor sample preparation and experimental design (Villar-Garea et al., 2007) and also by poor analysis (Brereton, 2006). A lot of the work was performed using HPLC and MS-MS looking specifically at protein markers such as looking for ovarian cancer (Petricoin Iii et al., 2002) or prostate cancer (Adam et al., 2002).

Most of the work performed on the headspace analysis of serum samples has been performed for drugs testing (Alizadeh et al., 2008; Schulz et al., 2009) and the detection of toxins (Hernandez et al., 2002). These methods were chosen as the analysis procedures allow rapid determination at point of care which is important in the initial screening. Other than the work performed at Cranfield (Pavlou et al., 2004; Fend et al., 2005; Knobloch et al., 2009a; Knobloch et al., 2009b; Knobloch, 2010), little work has be performed specifically for the detection of disease using the analysis of headspaces above serum with most work focusing on the analysis of the fluids themselves.

1.3. Aims & Objectives

As stated in the introduction, the overall aim of this thesis was to investigate the application of SIFT-MS analysis linked with multivariate analysis techniques to diagnose veterinary diseases in a range of sample matrices. The specific objectives followed to achieve this are listed below.

- Evaluate multivariate data analysis methods when applied to SIFT-MS data through the use of a proof of concept dataset created by analysing the headspace of two mixtures created in a lab. Specifically, investigate the possibilities of using the techniques for disease detection and biomarker discovery.

- Evaluate scaling and data pre-treatment required for the multivariate techniques used and propose a standardised methodology for analysis.
- Evaluate the methodology by application to badger serum tuberculosis detection and make modifications if necessary.
- Evaluate the methodology by application to cattle serum samples from a range of diseases and make modifications if necessary.
- Evaluate if the methodology is applicable to other samples such as urine and make modification if necessary.
- Based on the results achieved above, propose a methodology for the pre-treatment and analysis procedures taking into account the data source that can be the basis of further detailed analysis of applications of the method for disease investigation e.g. biomarker discovery and disease detection.
- Discuss the findings in respect to the impact, possible uses and further investigations that maybe possible.

2. Evaluation of Multivariate Analysis Methods for SIFT-MS Data

2.1. Introduction and Motivation

The selected ion flow tube mass spectroscopy (SIFT-MS) method enables rapid analysis of gaseous samples and headspaces. However, one of the bottlenecks in analysis is the length of time that it takes to manually analyse the results received from the system. Multivariate techniques have been used to speed up the process when gas chromatography mass spectroscopy (GCMS) and electronic nose (e-nose) methods have been used to collect the data (Scott et al., 2006; Pasikanti et al., 2008).

Another problem with the analysis of SIFT-MS data is the amount of data recorded. Expert knowledge is required to understand what features are worthwhile analysing. As multivariate techniques investigate the datasets as a whole, taking into account all available data, they allow the analysis of the data to be performed without this detailed expert knowledge of the system, although the final interpretations of this analysis may still require a basic understanding.

Since chemometrics-based multivariate techniques of analysis have not previously been performed on SIFT-MS data, a proof of concept trial has been conducted and the results are reported and discussed below. This chapter begins the creation of a methodological framework for the analysis of SIFT-MS data using multivariate techniques. The efficacy of the developed methodology is evaluated using data from simple synthetic mixtures of volatile organic compounds (VOCs). Problems may occur when using real biological samples but by having a well understood sample, theories and possible solutions to the problems will be easier to produce subsequently.

2.2. Data Collection & Measurement

The data for this experiment comes from samples prepared in the lab. Two mixtures of volatile liquids were made and the headspace above the mixtures analysed (Table 2.1). Fifteen samples of the one mixture were analysed and then fifteen samples of the second mixture were analysed. New batches of the mixtures were created and the process repeated creating a total of 60 samples, 30 of each mixture.

The use of manufactured samples ensures that the data collected will have known features and markers before analysis. The analysis methods can then be evaluated by their proficiency in detecting these features. The VOCs that the mixtures were created from were specifically chosen because they are found in biological samples (Turner et al., 2006a; Turner et al., 2006c).

Mixture one contained ammonia, methanol and acetone. Mixture two contained ethanol, acetonitrile and toluene. Liquid samples were created and then 1mL of sample pipetted into a 700 ml Nalophan bag. The bags were then filled with zero grade air (BOC, UK) and allowed to equilibrate in an incubator at 40°C. Each bag contains enough headspace to be analysed three times. The mixtures were created so that each one of these VOCs was at 1ppm in the bag when equilibrium was reached. The volumes of volatiles required were determined using Henri's law which can be corrected for temperature changes and enthalpy as described in (Knobloch, 2010). Then using known constants (Sander, 1999) and application of the modified form of Henry's law, 1ppm gaseous concentrations can be achieved from volume of the compounds liquid sample. For more detailed information on these calculations please refer to (Knobloch, 2010).

A 1ppm concentration was chosen as it was known to be of the same order of magnitude for many trace VOCs found in biological samples (Turner et al., 2006a; Turner et al., 2006b; Turner et al., 2006c; Turner et al., 2008). These compounds should produce the following peaks (Table 2.1).

Table 2.1 - The H_3O^+ spectra m/z values expected for the compounds used to create the mixtures. Mixture one (A) contains ammonia, methanol and acetone. Mixture two (B) contains ethanol, acetonitrile and toluene.

Compound	Mixture	H_3O^+ spectra peaks
Acetone	A	59, 77, 95
Acetonitrile	B	42, 60, 78
Ammonia	A	18, 36, 54
Ethanol	B	47, 65, 83
Methanol	A	33, 51, 69
Toluene	B	93

Once the mixtures had reached equilibrium (determined to be 15 minutes (Knobloch, 2010)), the bags were first connected to an electronic nose and two thirds of the volume used to record e-nose traces (Knobloch, 2010). The final third of the volatile headspace was sent through the SIFT-MS machine using a 5 minute full spectra scan, alternating through all three precursors.

2.3. Analysis with Trans SIFT mk. 2

2.3.1. Univariate & Classical Approach

As the contents of the samples was known and understood the univariate analysis should be simple. The VOCs have known peaks and analysis by comparing these known peaks would yield excellent results. However, the purpose of analysing these samples was to find methods for the analysis for unknown samples. Hence, the first assumption in this section was that the m/z ratios associated with the VOCs are unknown.

The first step in a multivariate analysis is to look at the raw data. This helps to gain an understanding of which techniques are likely to work. In the case of SIFT-MS we have spectra like the one in Figure 2.1.

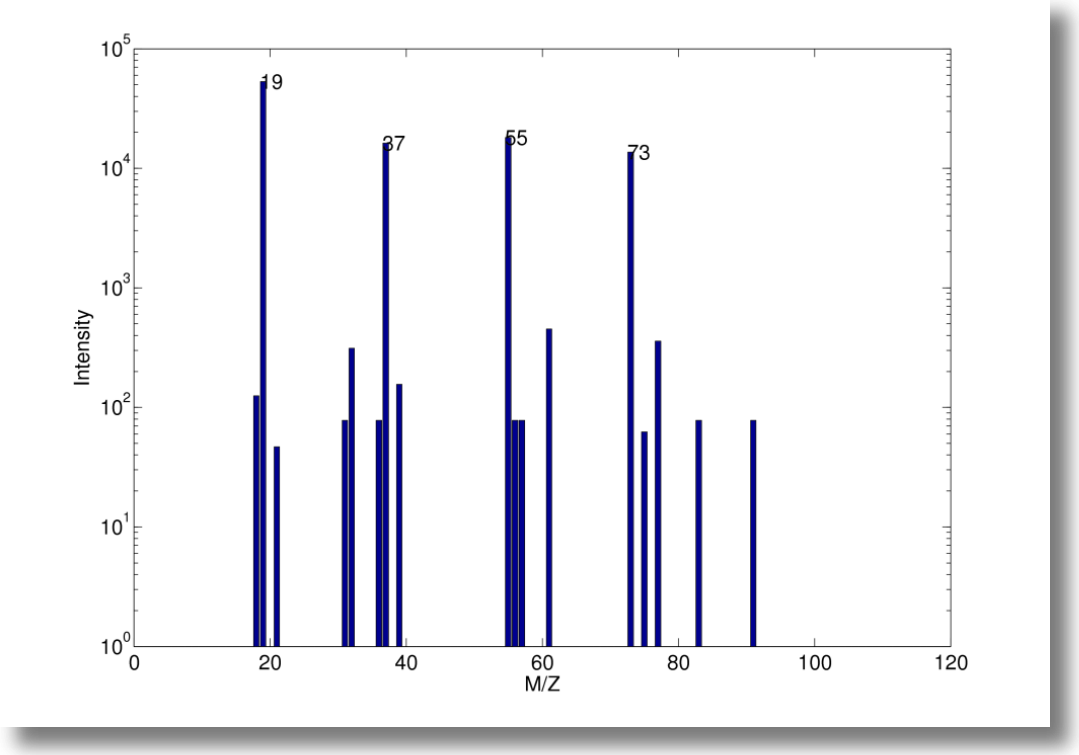


Figure 2.1 - The H_3O^+ spectrum straight from the SIFT-MS machine for a sample of mixture A (sample no. 16). The peaks produced by the principal precursor and its water clusters have been numbered.

This spectrum is dominated by peaks caused by the precursors that are added to the sample during analysis by the SIFT-MS machine. These carrier ions, that have the largest peaks, have very little to do with the sample under investigation. They do contain some information about the presence of water in the samples (if there was little water present in the samples fewer water clusters would be formed); however, as these samples were all aqueous, water clusters were expected. The biological samples that this analysis was prepared for were all aqueous as well, and hence the same peaks will be visible in their spectra. The spectra obtained from the other two precursors showed similar features except that the m/z of the precursors themselves were different (spectra not shown). There were crossovers between the different channels,

for example H_3O^+ at m/z 19 was present on all three spectra even if it was at much lower levels on its non-native channels. There is the possibility that VOCs present in the samples may have m/z ratios that are listed in all three rows of the table below. If this is the case it simply means that the SIFT-MS technique will not be able to accurately monitor the specific VOC.

The peaks identified as precursors (Smith and Spanel, 2005) have been tabulated below.

Table 2.2 - The m/z peaks caused by the precursors on each of the three spectra recorded by the SIFT-MS machine when analysing aqueous samples. There is precursor crossover between the channels (e.g. m/z 19 is H_3O^+ and this is visible on all three channels) (Smith and Spanel, 2005).

H_3O^+	19, 20, 21, 30, 32, 37, 38, 39, 55, 56, 57, 73, 74, 75, 91
NO^+	19, 30, 32, 37, 48, 50, 55, 57, 66, 73, 91
O_2^+	19, 30, 32, 33, 34, 37, 50, 55, 56, 57, 73

The precursor peaks need to be removed from the three spectra and then the spectra plotted again. This should reveal the peaks caused by the samples and not those contributed to by the additions necessary for the analysis process.

After the precursors have been removed the m/z associated with the samples become apparent. The below (Figure 2.2) is a spectrum from mixture A, which is only representative of half of the dataset. As the mixtures contain different volatiles (and not just the same volatiles at different levels) the spectrum for a sample of mixture B should be significantly different (see Figure 2.3). In the above example m/z 18 and 36 are caused by ammonia, m/z 77 is caused by acetone and m/z 83 is caused by ethanol. As this is a sample of mixture A (which should contain only ammonia, acetone and methanol), the presence of ethanol and the absence of methanol is a cause for concern.

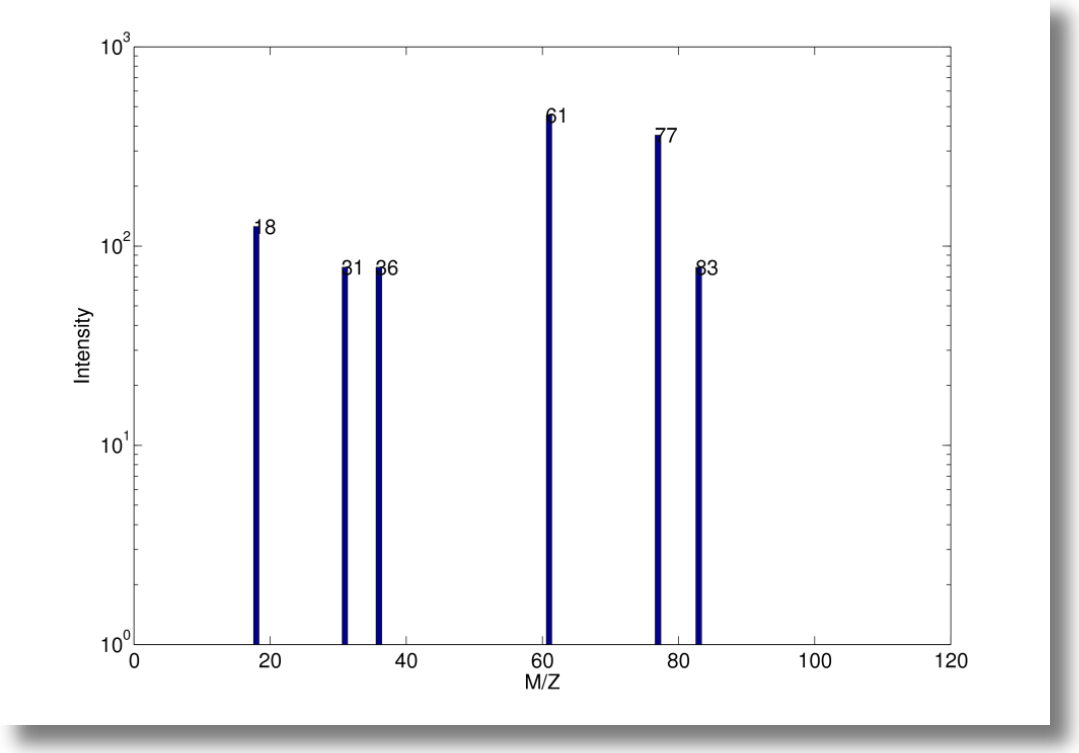


Figure 2.2 - The spectrum as captured using the H_3O^+ precursors for a sample of mixture A (sample number 16) after the m/z counts for the precursors (as outlined in Table 2.2) have been removed.

One possible reason for the contamination of the sample by ethanol is the known presence of this VOC in laboratory air (Smith and Spanel, 2005). Another possibility is contamination during the production of the samples. The samples were prepared by mixing the liquid forms together to create an aqueous solution. These were then transferred to Nalophan bags by pipette. Contamination could have been caused at this stage.

If either of these two causes above are applicable in this situation, there is nothing that can be done from a data analysis point of view. These problems will be apparent in all samples recorded. The SIFT-MS machine is situated in a lab and the air is subject to change, also sampling techniques and methods are performed by humans and hence subject to human error. Once the errors have been spotted, it is usually too

late to go back and resample. However, only a few VOCs will be affected and the multivariate methods will be able to make use of the remaining unaffected VOCs.

Finally, the detection method does rely on statistical sampling methods. The final quadrupole in the SIFT-MS machine cycles through m/z ratios and hence it is possible that VOCs in the sample were not recorded, which might explain the lack of a m/z peak associated with methanol. If this is the case then the only possible solution is to perform replications of sampling. This might not be possible due to limited sample size for instance but if enough samples are recorded this problem should just be captured by uncorrelated variance (i.e. noise).

The spectrum of sample 1 which is of mixture B (Figure 2.3) shows a large number of volatiles. m/z 18 is ammonia, m/z 60, 78, 96 are acetronitrile, m/z 83 is ethanol, m/z 93 is toluene and 61, 81 and 103 are not linked with the VOCs added to either mixture. As with the spectrum of mixture A (Figure 2.2), VOCs that were not part of the mixture in question were being detected. This is one of the reasons that analysing SIFT-MS spectra from biological samples one at a time would be a time consuming activity. Finding an unusual m/z doesn't necessarily lead to identification of a biomarker or indeed a definite link to the sample itself. Only when that m/z appears repeatedly can it be used as a marker, so a spectrum needs to be examined and then compared to every other spectrum in the dataset. One method of doing just this is to plot the heights of just one m/z value on a bar graph for all the samples. As ammonia has shown up in both the spectra examined so far, further investigation of m/z 18 would be interesting.

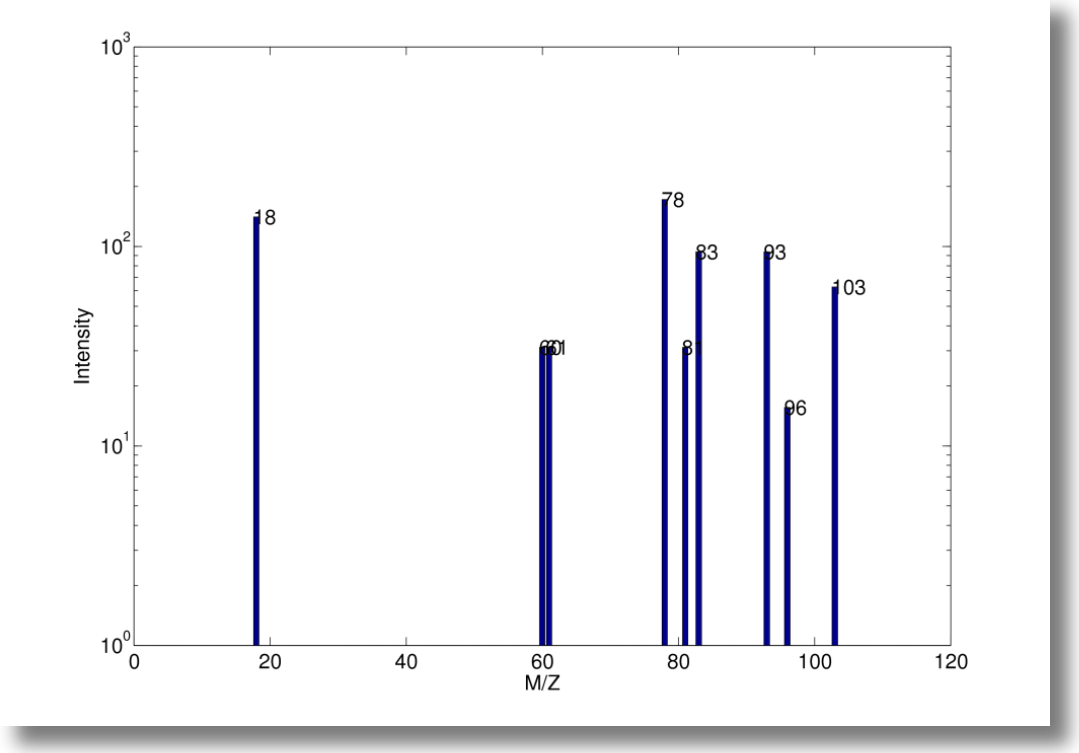


Figure 2.3 - The spectrum as captured using the H_3O^+ precursor for a sample of mixture B (sample number 1) after the m/z counts for the precursors have been removed.

Figure 2.4 shows the peaks for ammonia (m/z 18) over the sixty samples. It was created by extracting the values recorded for just m/z 18, sorting them by order of analysis and then separating them into two groups dependent on mixture. It is clear that mixture A, which had ammonia added, has much higher levels of ammonia. The first sample of each of the mixtures (starting from the left of the horizontal axis) was chosen to produce Figure 2.2 and Figure 2.3 respectively. The sample of mixture A has the lowest recorded level of ammonia (m/z 18) of all the samples of mixture A. Indeed, the levels are very similar to the first sample of mixture B and any conclusions formed from the analysis of Figure 2.2 and Figure 2.3 about ammonia levels being similar can now be seen as flawed. This highlights the need to analyse more than just one spectrum at a time.

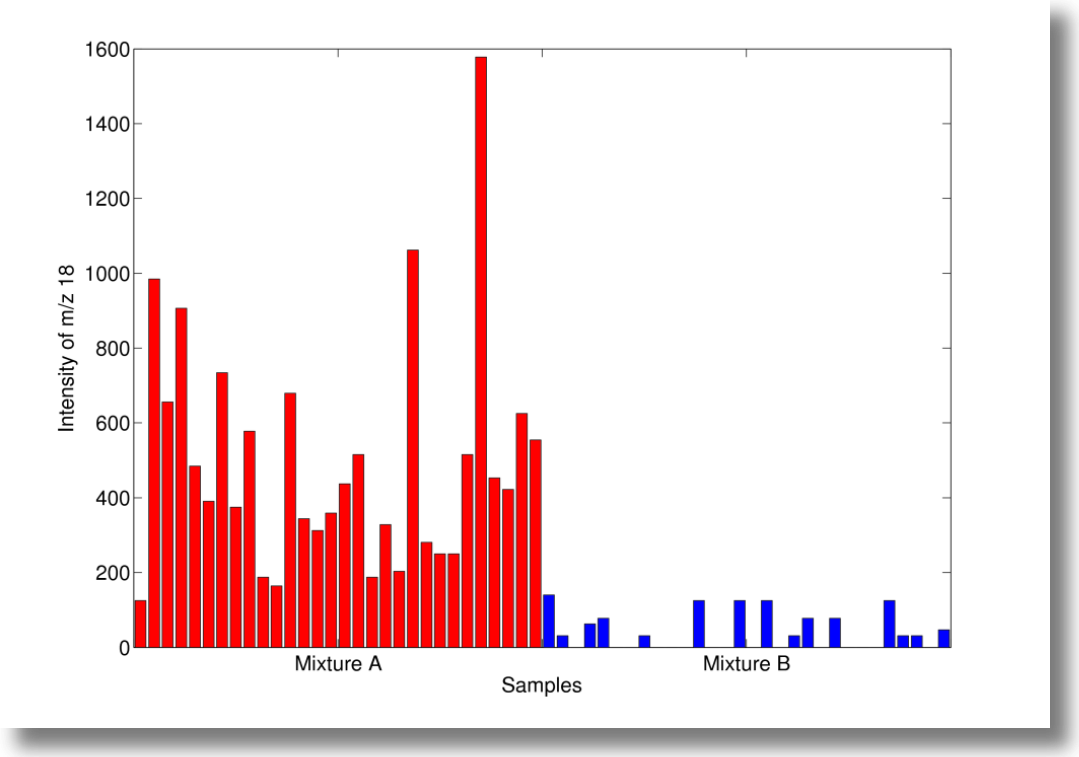


Figure 2.4 - The intensities (counts) of m/z 18 (ammonia) as recorded on the spectra produced using the H_3O^+ precursor for all samples. Samples of mixture A have been coloured red and are on the left, samples of mixture B have been coloured blue and are on the right. Ammonia was one of the volatiles used to create mixture A but was not supposed to be present in mixture B.

Figure 2.5 shows that methanol is present in all but five of the samples of mixture A. It is also present in three samples of mixture B. It would therefore appear that by chance the choice of sample 16 in Figure 2.2 was one of the few without methanol recorded.

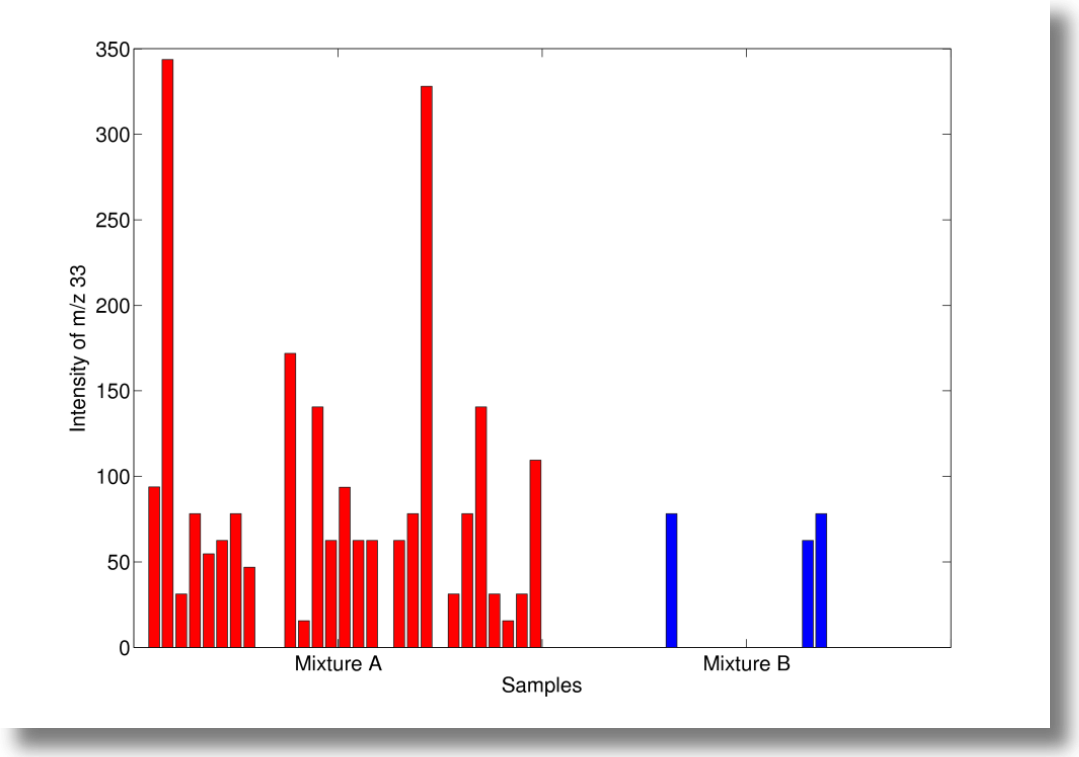


Figure 2.5 - The intensities (counts) of m/z 33 (methanol) as recorded on the spectra produced using the H_3O^+ precursor for all samples. Samples of mixture A have been coloured red and are on the left, samples of mixture B have been coloured blue and are on the right.

Figure 2.6 appears to show ethanol to be an unreliable marker. It is present in only nine of the samples of mixture B and is present in six of the samples of mixture A. Ethanol as recorded at m/z 47 would, in this situation, not be an effective marker.

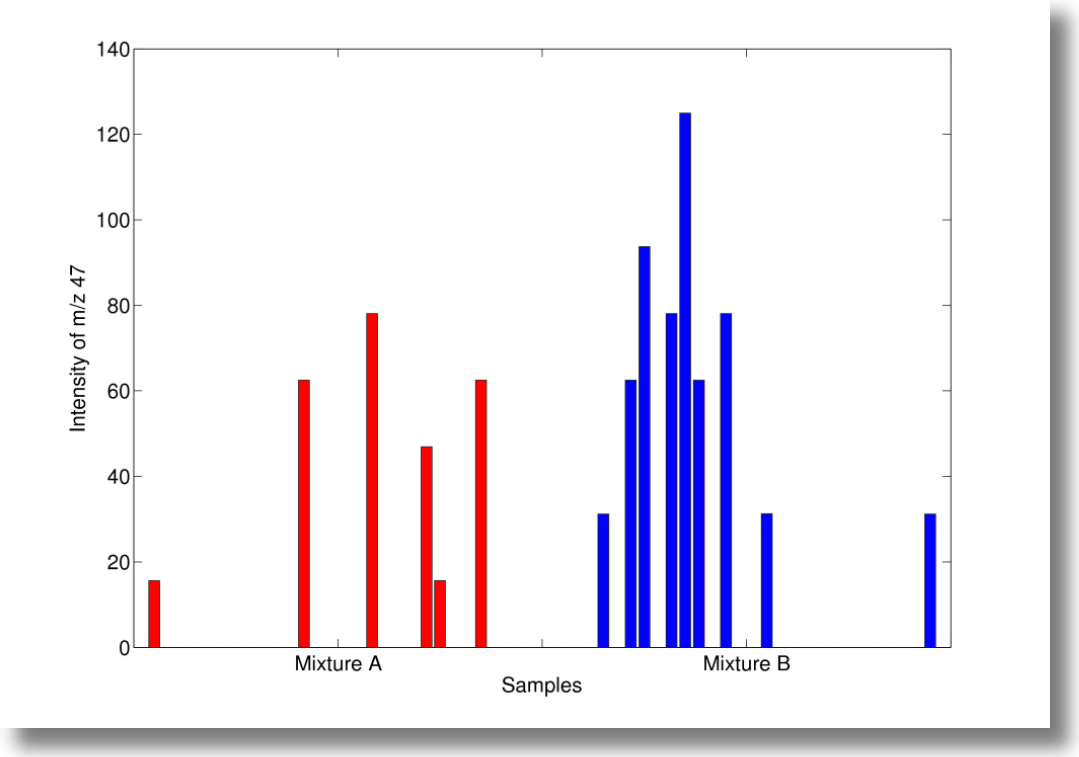


Figure 2.6 - The intensities (counts) of m/z 47 (ethanol) as recorded on the spectra produced using the H_3O^+ precursor for all samples. Samples of mixture A have been coloured red and are on the left, samples of mixture B have been coloured blue and are on the right.

The problem that the above data highlights is the need to look at specific m/z counts across all the samples to get the best picture of whether the specific m/z in question has an impact on the discrimination between the mixtures. This requires a prerequisite knowledge of the m/z peaks that are important, which is counter-intuitive to the search for unknown/new biomarkers. The alternative is to look at every single m/z ratio - the problem is then instead of having 60 spectra to investigate, there are anywhere up to 600 m/z graphs which would make this a very time consuming and subjective approach to the analysis.

One of the simplest solutions to the above problem is to calculate the mean peak height for all the samples of a particular group of samples, i.e., to create an average spectrum for each mixture. This reduces the number of spectra to be investigated to

just two: the average spectrum for a sample of mixture A and the average spectrum for a sample of mixture B. Once these averages have been calculated it is quite simple to plot them both on the same graph to investigate the differences between the sample groups (Figure 2.7).

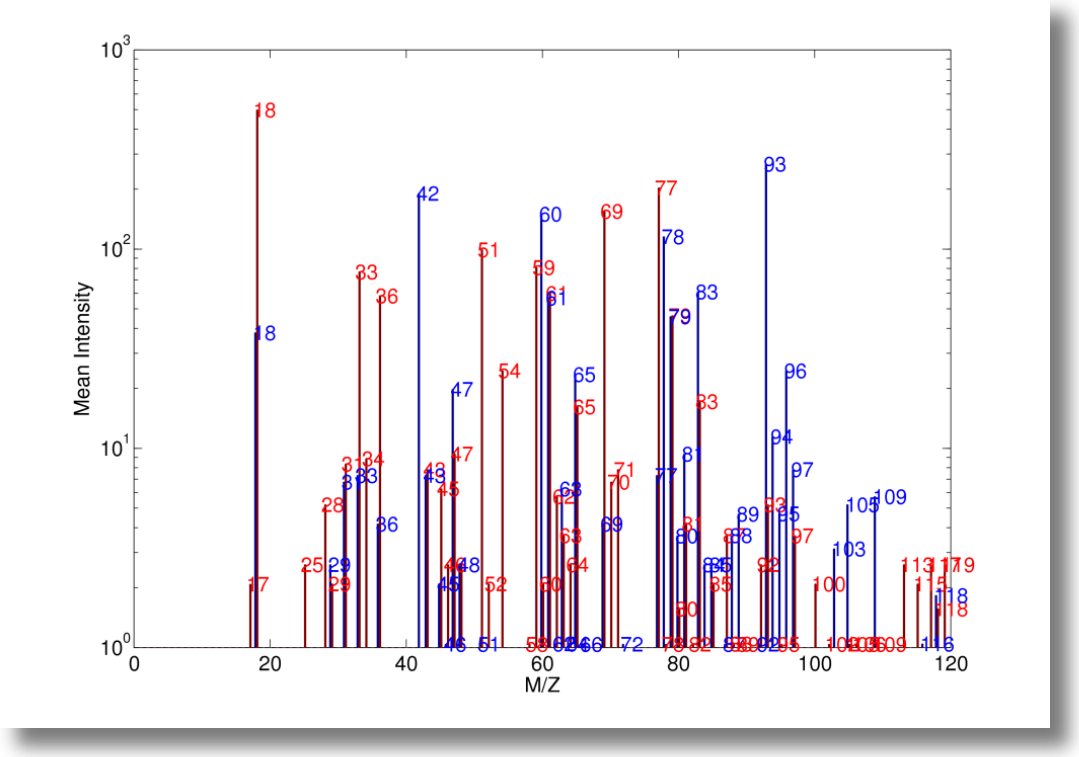


Figure 2.7 - The mean average spectrum (all thirty spectra of each sample type summed and divided by thirty) for the two mixtures. Mixture A is coloured red and mixture B is coloured blue. A large number of extra volatiles not expected (see Table 2.1) can be seen.

Looking at the averages in Figure 2.7, the more important m/z peaks that are linked to our mixtures (Table 2.1) can clearly be seen to have large average differences between the sample spectra. In particular, m/z 42 (acetronitrile) and m/z 93 (toluene) have clearly large counts only on the blue spectrum which is associated with mixture B. m/z 51 and 69 (methanol) have a large presence only on the red spectrum which is associated with mixture A.

There are many more differences between the two mean spectra that are not so clear. Some of these differences are significant (i.e. that the difference seen are not caused by natural variations and purely down to chance) and caused by substantive chemical differences between the mixtures. However, some of these differences will not be significant. They may be caused by background contamination from VOCs present in the laboratory or by a single very high value either caused by machine fault of contamination of a sample. The problem becomes one of classifying this significance. Adding the standard deviations to the above graph in the form of error bars would highlight the natural variance and allow a comparison between the difference in means and the natural variance. One problem is that it would make the graph considerably more difficult to interpret.

The alternative way to be sure that neither of the above conditions was contributing to the difference in the means would be to plot the m/z for all samples with differences in the means. An investigation into the most significant m/z would lead to the investigation of perhaps 20-30 graphs for the H_3O^+ precursor alone. Looking at the average spectra for the other precursor graphs (not shown here) there are similar numbers of differences involved. Using this method, there would only need to be perhaps 100 graphs of individual m/z across all samples. This has reduced the volume of manual work required, however 100 graphs is more than could be feasibly analysed manually.

In the analysis of microarray data, a similar process is automated and used as the initial data preparation step. Several thousand gene probes are narrowed down to 20-30 important ones that are then analysed (Goodacre et al., 2004). Indeed, a reduction in the ranges of 1000:1 is a powerful technique analysing 20-30 genes instead of 20,000-30,000 is a significant decrease in labour. With SIFT-MS data this reduction is much less absolute, analysing 100 m/z ratios instead of 600 m/z ratios is a starting point but a more powerful method would be advantageous.

Another problem that none of the above univariate methods is capable of solving is the problem of multiple peaks contributing to a single feature. Indeed, Figure 2.6

suggests that ethanol, as described by m/z 47 on the H_3O^+ channel, is not suitable as a marker. Ethanol can form water clusters and so peaks could be detected at m/z 53 or m/z 71 on the H_3O^+ channel alone, without considering the other spectra. To properly analyse the levels of ethanol, an expert would have to look at all three potential peaks on the H_3O^+ spectra and use expert knowledge in the determination of ethanol concentration in the samples. This, in a way, is a multivariate approach.

It also appears that the very simple experiment on known VOCs has produced a problem much more complicated than would have been expected. The mixtures have been created from volatiles that are known to cause 16 peaks but from the above figure it appears that more than 30 m/z peaks are being detected. The exact cause of these extra peaks is unknown but it could be spurious results from the SIFT-MS machine (low level spikes that the software has not identified and filtered out) or volatiles from the laboratory contaminating the results or valid peaks of volatiles in our samples that have not been directly associated with the VOCs.

2.3.2. PCA

Given the complexity of the data observed, the next step was to examine multivariate techniques. As mentioned in the literature review, multivariate techniques aim to reduce the dimensionality of the problem. If nothing else, the univariate approach has highlighted the large dimensionality of the data.

An important step in the analysis of complex data is the pre-treatment step where the data is modified in preparation for analysis using multivariate techniques (Otto, 1999a). There are many methods of data preparation available and the exact combination is dependent on the analytical tool used to create the data and the sample types being analysed. An investigation is required to determine the best approach for SIFT-MS headspace analysis.

The first PCA scores plot of the SIFT-MS spectra data (Figure 2.8) was not very revealing. There was little discrimination between the samples of mixture A (red

dots) and of mixture B (blue crosses). So far, no scaling has been performed and many other tools that are available that have yet to be trialled. However, before these tools are used, it is important to investigate the cause of the very large percentage variance on principal component one (PC1). There might be a simpler method to deal with the lack of discrimination depending on the cause.

An outlier can also be seen on the figure. Before dealing with it, the loadings plot will needed to be investigated to understand what m/z ratio holds the underlying data.

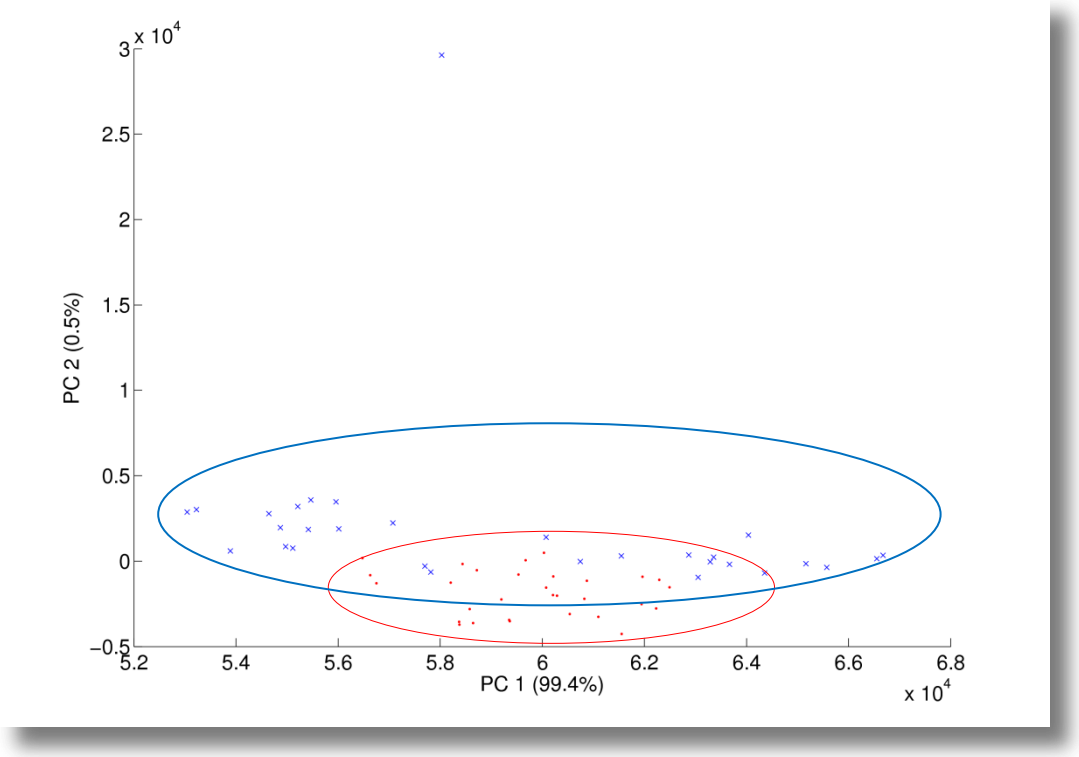


Figure 2.8 - PCA scores plot of the H_3O^+ spectra for all the samples, red dots are individual samples of mixture A and blue crosses are samples of mixture B. Clusters have been highlighted using the corresponding colours showing that despite the first component containing 99.4% of the variance little discrimination can be seen on this axis alone. Partial discrimination can be seen on the secondary axis although the clusters overlap.

The loadings of the first principal component (Figure 2.9) show that a large amount of the variance (99.4%) is attributed to the changing levels of the precursor ions themselves. These ions are added to the sample as part of the analysis process. They contain little information about the original sample contents. As in the univariate approach it appears that they need to be removed before analysis.

The loadings of the second component (not shown) display the same m/z ratios. This means that the outlier in (Figure 2.8) should be handled by the removal of the m/z ratio associated with the precursors.

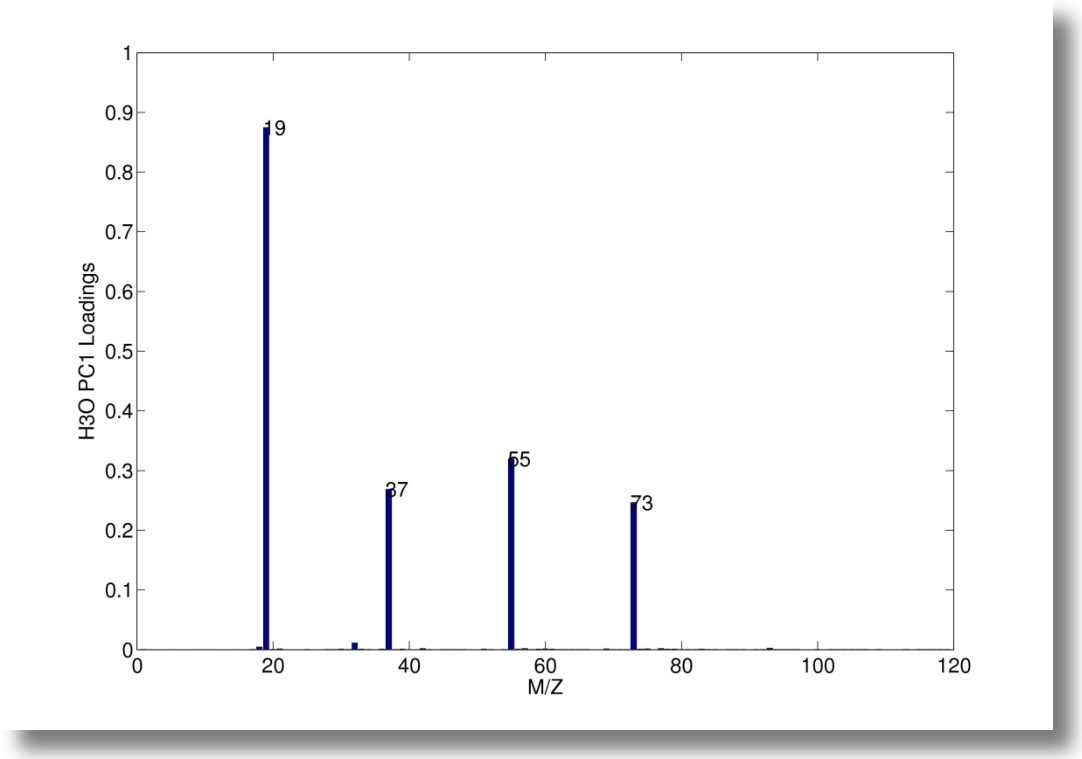


Figure 2.9 - The loadings plot of the first principal component for the above PCA scores plot (Figure 2.8) using the H_3O^+ precursor spectra only. The main contributors to the variance above are linked to the H_3O^+ precursor ion and its water clusters which are added to the sample as part of the analysis process i.e. they are not part of the original sample.

The precursor ions are added to the flow tube at a constant rate so the total quantities measured by the detector should be consistent and the variance contribution should be low. As a percentage of the total count, the variance contributed is small. However, as the magnitude of the peaks is 3-4 times greater than that of the sample VOCs, the actual contribution to the variance is large enough to cause problems, as shown in Figure 2.8.

The software supplied with the SIFT-MS machine can be used to normalise the traces using the recorded precursor levels. There are however two assumptions. First, that the total precursor level can be calculated from the three primary precursor peaks and second, that the samples interact with the precursors in the same way. Realistically, neither of the assumptions can be made if the samples being investigated change, and this method is only really used for on-line monitoring (Smith and Spanel, 2005).

Once the precursors have been removed the data separates nicely into two clusters (Figure 2.10). Investigating the loadings will show the cause(s) of this separation (Figure 2.11 and Figure 2.12).

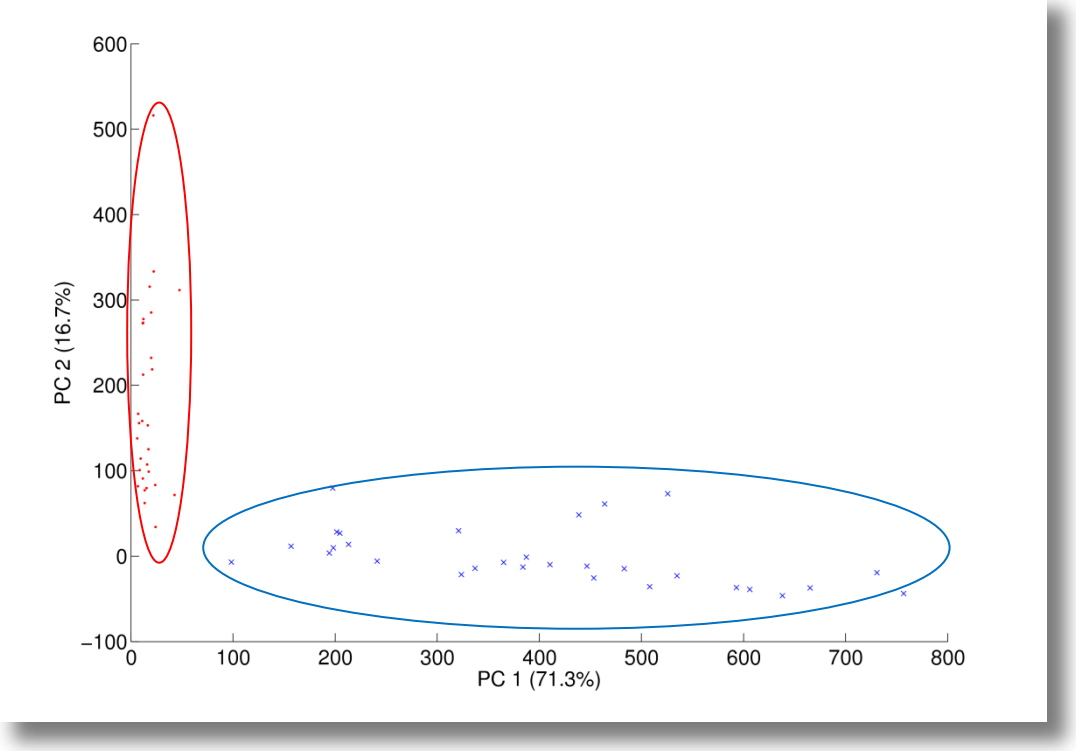


Figure 2.10 - PCA scores plot of the samples after the m/z associated with the precursors have been zeroed. The red dots are samples of mixture A and the blue crosses are samples of mixture B. Clear discrimination between the two groups can be seen, as highlighted by coloured eliplises (added manually).

The first component of Figure 2.10 is mostly influenced by m/z 18 according to the loadings (Figure 2.11). The peak at m/z 18 is linked to ammonia levels.

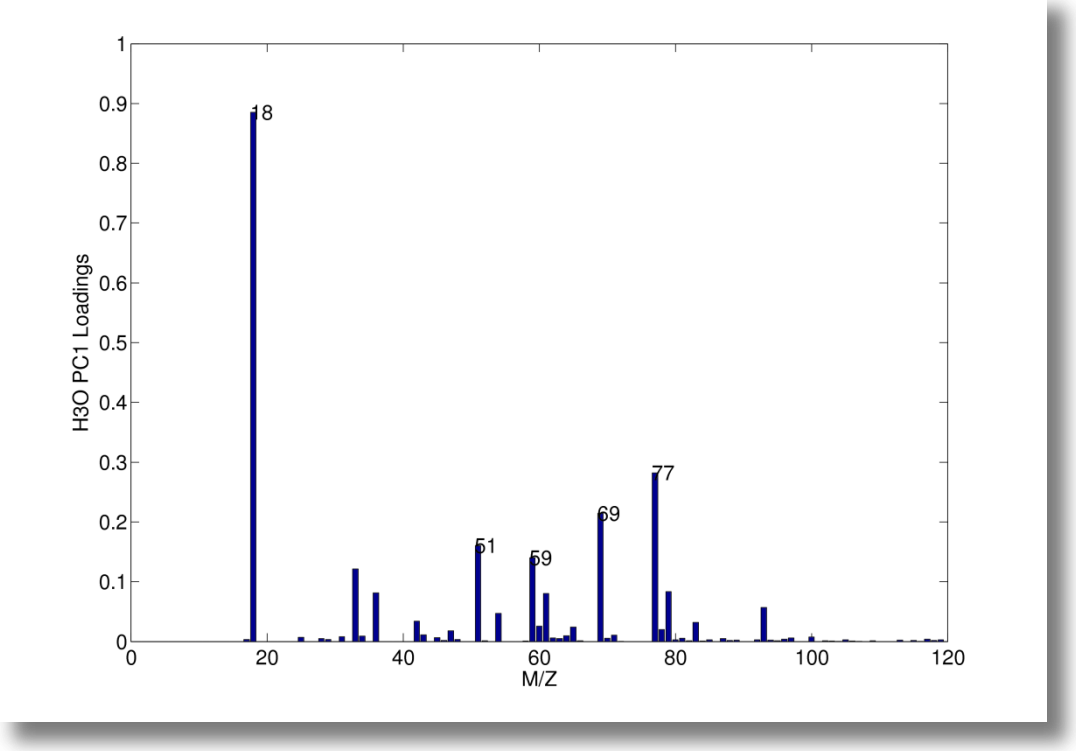


Figure 2.11 - The loadings of the first principal component corresponding to the previous PCA scores plot (Figure 2.10) using the H_3O^+ data with the precursors removed. The largest peak (m/z 18) is linked to ammonia.

The second component (Figure 2.12) of the PCA scores plot (Figure 2.10) is mostly linked to m/z 93 which is indicative of toluene. This links back to exactly what was seen in the PCA scores plot (Figure 2.10). Samples of mixture A (red dots) are all clustered around the zero mark of PC2. These samples have very little toluene. They do however, contain large quantities of ammonia. The observations made of the levels of ammonia observed in the samples of mixture A fluctuate, which causes the spread along the PC1 axis.

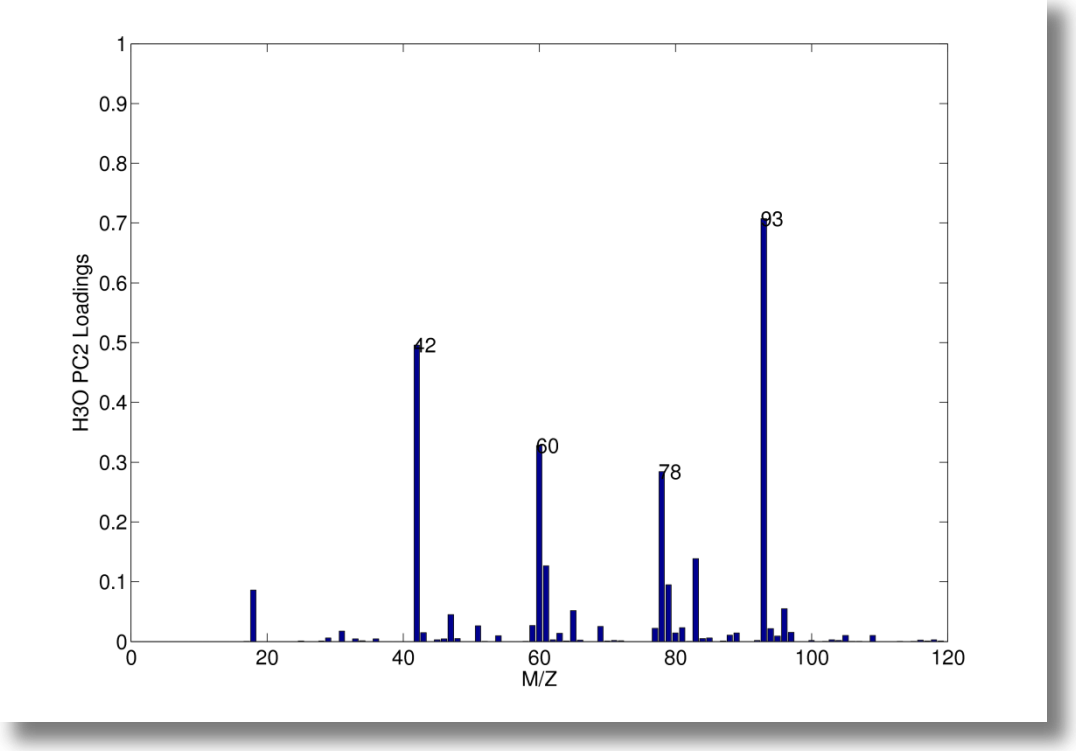


Figure 2.12 - The loadings of the second principal component of the previous PCA score plot (Figure 2.10) using the H_3O^+ data with the precursors removed. The largest peak (m/z 93) is linked to toluene.

A similar argument can be made but in reverse for the samples of mixture B (blue crosses). They are all clustered around the zero mark of PC1 suggesting that the samples contain little ammonia (m/z 18). They are spread out along the axis of PC2 suggesting that they have varying observations of the levels of toluene (m/z 93).

The variance is very quickly captured (Figure 2.13), with over 90% of the information captured by the first three principal components.

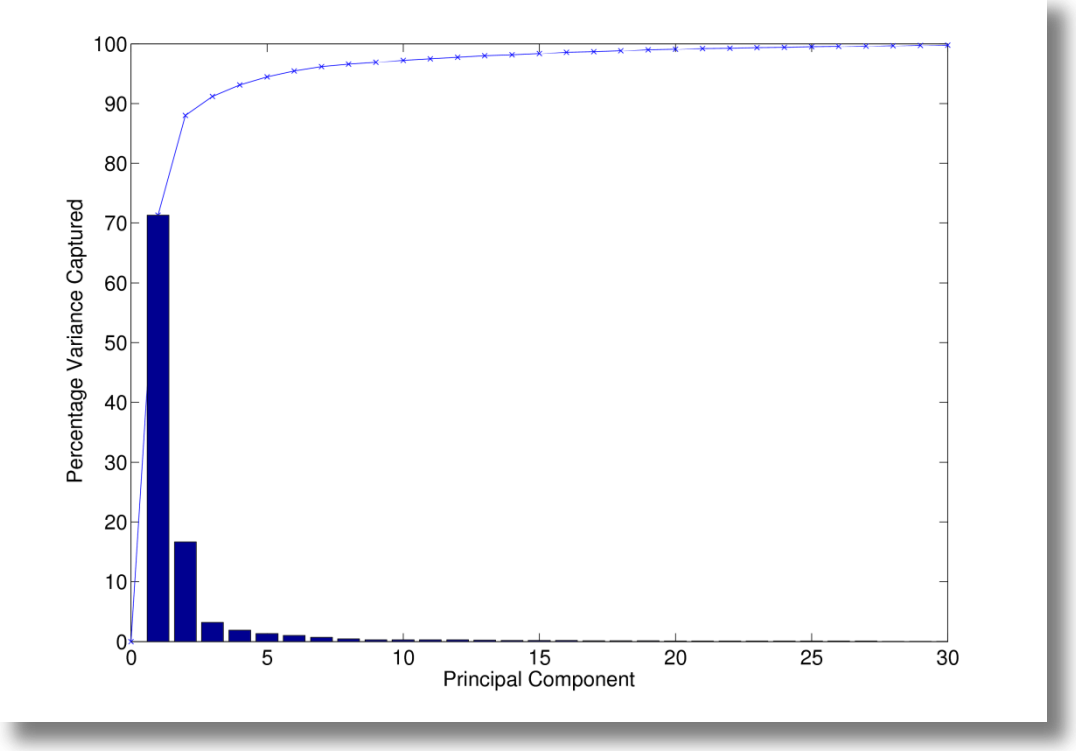


Figure 2.13 - The percentage variance captured for the first 30 principal components (bars) and the cumulative sum of the components (line) for the PCA analysis (shown in Figure 2.10) created using the H_3O^+ data with the precursors removed.

Next, an investigation in to whether combining the three spectra from the precursors (H_3O^+, NO^+, O_2^+) into a single dataset 600 variables wide will give better results. For every sample, there are three spectra (H_3O^+, NO^+, O_2^+) each with a maximum range of 1-200 m/z depicting the contents of the sample in a slightly different way. The H_3O^+ precursor is probably the most useful (Smith and Spanel, 2005). There will be a lot of the information replicated on the other two spectra but there will also be extra information about VOCs that cannot be detected on the H_3O^+ channel or have multiple causes on the H_3O^+ channel but only a single cause on either the NO^+ or O_2^+ channel. There is currently no reason to perform PCA using only a set of spectra from one precursor at a time. Combining the datasets and repeating the analysis gives the result shown in Figure 2.14.

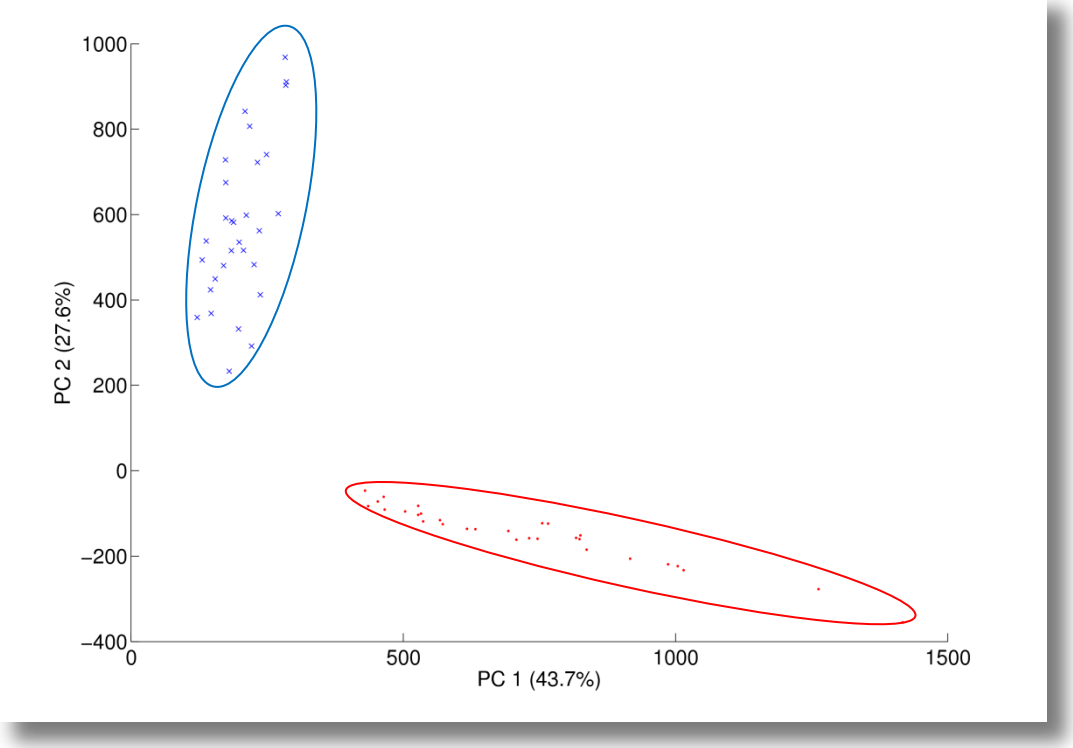


Figure 2.14 - PCA scores plot of the mixtures data of all three precursor spectra combined into a single vector. m/z linked to the precursors ions have been zeroed. Clear separation can be seen between the two outlined clusters.

The PCA on the dataset created from all three spectra combined (Figure 2.14) has produced similar results to the PCA of just the H_3O^+ spectra (Figure 2.10). After investigating the PCA scores and loadings plots of the other two spectra (results not shown) it can be seen that the PCA of all three spectra together combines features and visual patterns from the PCAs of the datasets from the separate precursors. At this stage, and with this dataset, it would only be a personal preference trying to decide which method gives the best results (however if there are no other reasons then fewer graphs is preferable). Investigating the loadings of the above PCA scores plot is not as simple as the loadings of a single precursor's spectra (Figure 2.11 and Figure 2.12). However, by prefixing the m/z with either H, N or O allows for the discrimination between the spectra that the m/z having an impact is from (Figure 2.15 and Figure 2.16).

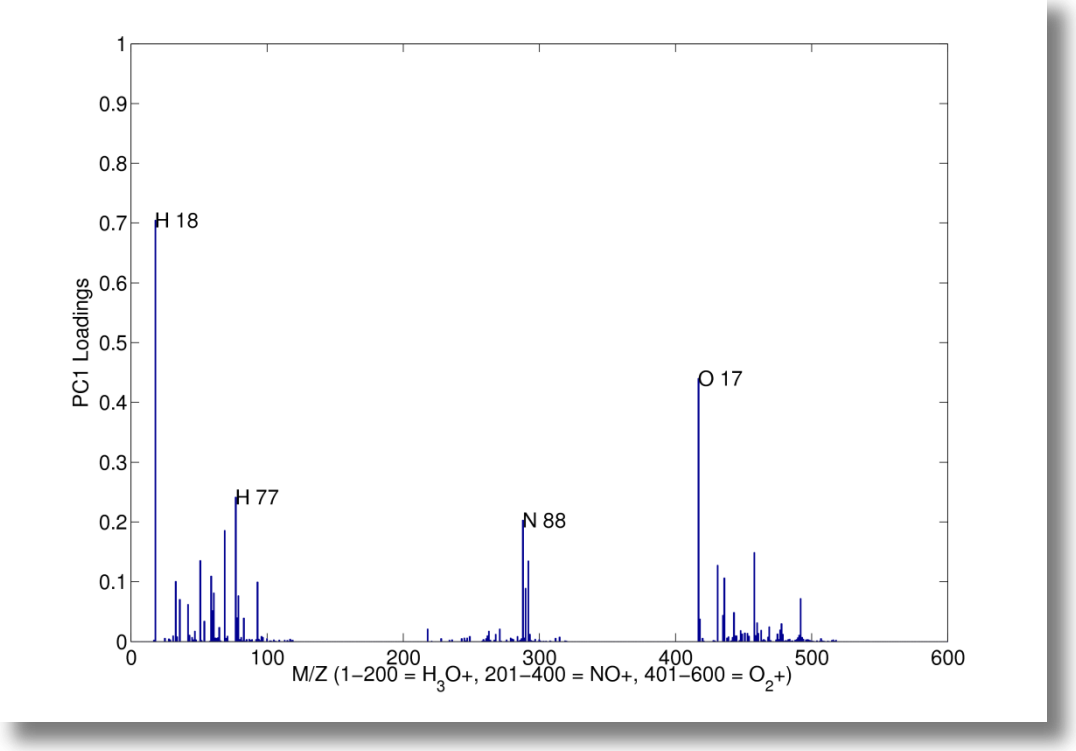


Figure 2.15 - The loadings of PC1 from the PCA analysis performed using data created by combining the spectra recorded using all three precursor ions (Figure 2.14). The m/z ratios from the largest peaks have been highlighted and the m/z ratio prefixed with the letter corresponding to the precursor channel used in the sampling process.

Ammonia once again has a large impact on the first principal component. The largest impact comes from the H_3O^+ spectra (m/z 18) but this is also accompanied by ammonia on the O_2^+ spectra (m/z 17). Ammonia does not appear to have the same order of variance on the NO^+ spectra. Instead we see a peak at 88 which could be associated with acetone. This theory would appear to be backed up by the presence of a peak at m/z 77 on the H_3O^+ spectra.

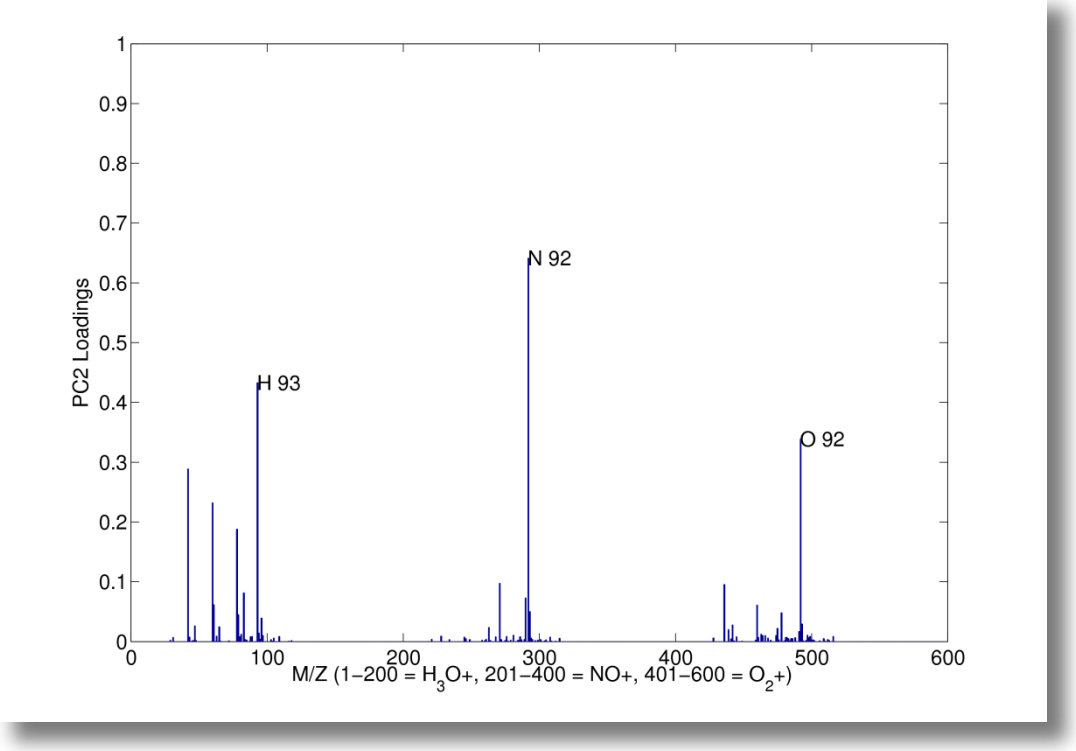


Figure 2.16 - The loadings of PC2 from the PCA analysis performed using data created by combining the spectra recorded using all three precursor ions (Figure 2.14). The m/z from the largest peaks have been highlighted and the m/z numbers prefixed with the letter corresponding to the precursor channel used in the sampling process.

The loadings of PC2 contain fewer dominant features (Figure 2.16) than those of PC1 (Figure 2.15). All the peaks are caused by toluene (93 on H_3O^+ and 92 on the others). On H_3O^+ it has been protonated causing the increased m/z. The loadings plots are more difficult to interpret and relate back to the original data when all three spectra are combined (Figure 2.15 and Figure 2.16) compared with the loadings graphs of H_3O^+ PCA analysis alone (Figure 2.11 and Figure 2.12). There are clearly arguments for and against the combination of the three precursors into a single dataset for analysis.

The number of principal components required to capture the same proportion of variance is much higher for the analysis of the three precursors combined (Figure 2.17) when compared to with analysis of the H3O+ precursor alone (Figure 2.13).

This agrees with the conclusion that there is more information available when the data from the three precursor channels is combined. It also shows that more complexity is included in the data as more principal components are required to capture the variance.

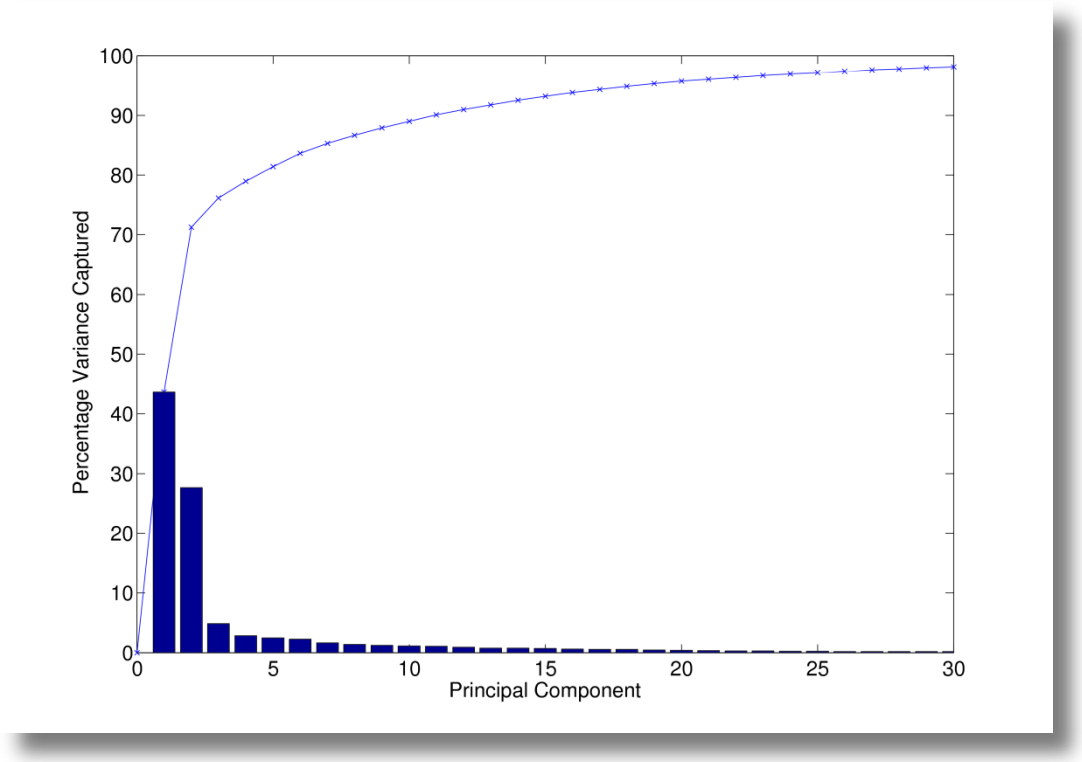


Figure 2.17 - The percentage of variance captured by each of the first 30 principal components (bars) and the cumulative sum (line) for the above PCA analysis (Figure 2.14) created using a combined vector of all three precursor channels.

2.3.3. PLSDA

The next step is to take the information about the samples and which group they belong to and try to discover more m/z ratios that may be having an impact on the differentiation. PCA showed that ammonia and toluene are important and by investigating more of the PCs, more m/z peaks that have a positive impact on the discrimination between the samples types would be found. The problem becomes, that with every new PC included, the data potentially contains less information and

more noise (Otto, 1999a). Supervised methods such as PLSDA hold the solution. PLSDA specifically rearranges the data into new dimensions that maximise the co-variance between the sample groups.

As introduced in the data analysis section of the literature review (Section 1.2.4), PLSDA uses a similar process to PCA in that the data is broken down into eigenvectors and then recombined with new dimensions. The difference is that PCA maximises the variance captured whereas PLSDA maximises the co-variance. As both of these are variance based analysis methods the m/z ratios associated with the precursor ions will need to be removed, for the reasons described earlier. The co-variance between the groups of samples should theoretically be small for the m/z ratios associated with the precursors. However, as previously observed, the magnitude of the counts is 3-4 times higher than the average m/z ratios associated with the VOCs that we are investigating; therefore the co-variance has a much higher chance to be larger than the co-variance from m/z ratios linked to the differences between the samples themselves.

PLSDA also supplies quantitative information about the accuracy of the separation. This can be used to determine the optimal data pre-treatment methods and if analysing the spectra separately or as one returns the best results.

Figure 2.18 shows that the model fails to identify the differences between known samples. PLSDA is more sensitive to bias in the data (all the m/z peaks are positive) than PCA so the lack of data pre-treatment is potentially the cause (Otto, 1999a; Wongravee et al., 2009). The important lines in the above figure for determining the accuracy of the model are the green leave one out (LOO) line, which shows the number of samples correctly identified using a model created using the rest of the samples, and the dark blue model line. This blue line contains information about the production of the model. It shows how well the model has built and if enough information is available it should always be able to reach 100%. The problem with this line is that it does not show how well the model performs as a predictor of

unknown samples. It only shows how well it can discriminate between the samples that were used to build the model.

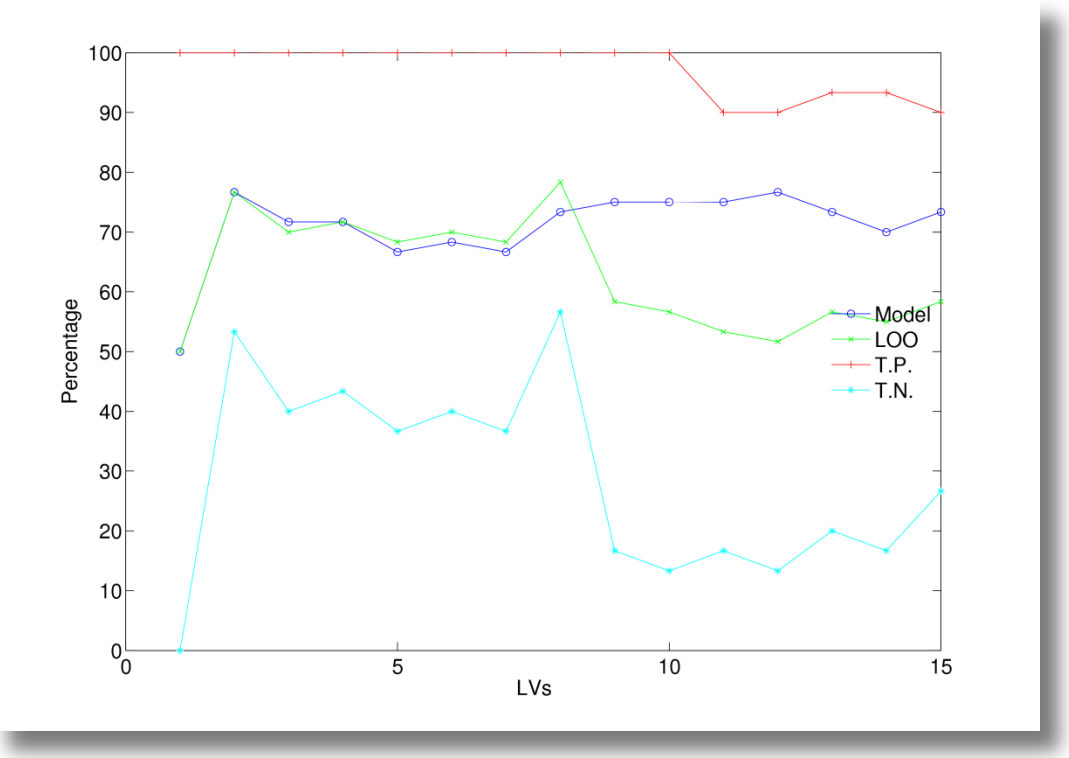


Figure 2.18 - The results from a PLSDA leave one out (LOO) validation for the H_3O^+ precursor spectra. The four lines show the model accuracy (blue o), the LOO accuracy (green x). The LOO is then broken down into true positive (red +) and true negative (light blue dots). The m/z ratios associated with the precursors have been removed but no scaling methods were applied.

After a mean centre algorithm (Eq. 1.2) has been used to pre-treat the data before analysis the PLSDA results are much more accurate at both predicting the known samples and the unknown LOO validation samples (Figure 2.19). The model achieves 100% accuracy (Figure 2.19) when tested against samples used in the creation of the model (blue line) and 100% LOO validation (green line) with a very low complexity model using only 1 or 2 LVs. As this is a non-complex case consisting of two mixture with different VOCs (specifically chosen this way to develop the method) the data reflects this (very few degrees of freedom) and hence the model becomes over-trained

very quickly. This can be seen in Figure 2.19; the model line (blue line with o) stays at 100% but the LOO line (green x) decreases towards 50% (even chance).

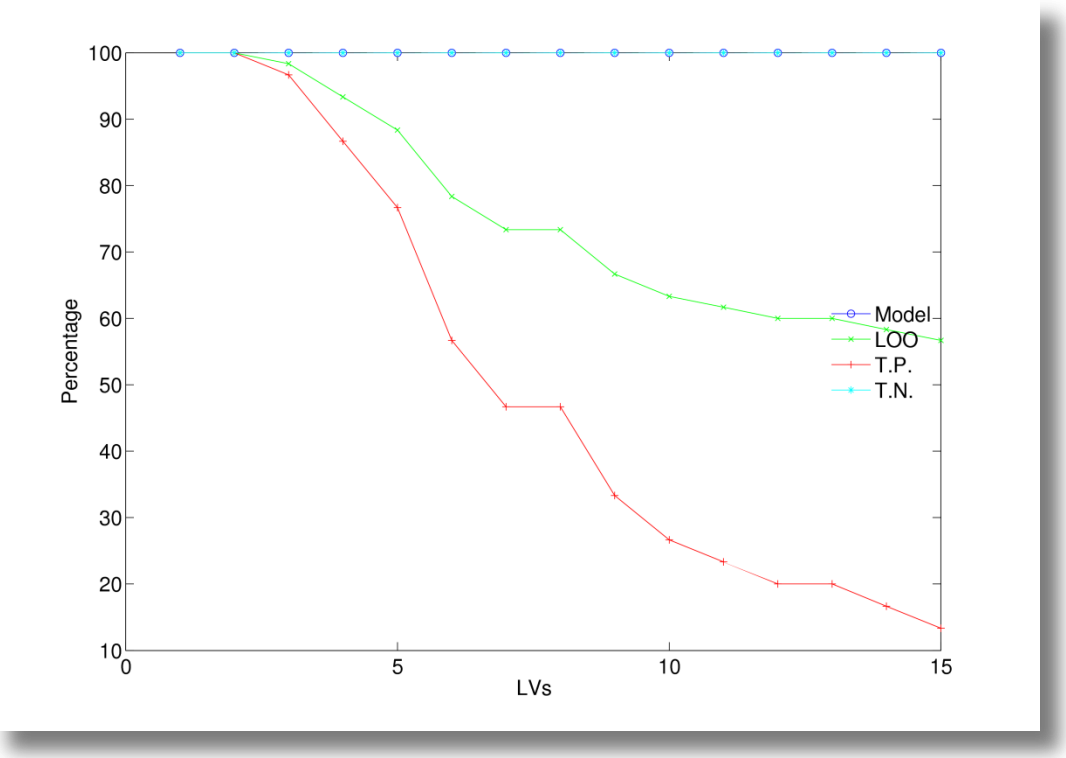


Figure 2.19 - The results from a PLSDA LOO validation for the H_3O^+ precursor spectra. The four lines show the model accuracy (blue o), the leave one out (LOO) accuracy (green x). The LOO is then broken down into true positive (red +) and true negative (light blue dots). The m/z ratios associated with the precursors have been removed and the data has been pre-treated using a mean centre algorithm (Eq. 1.2).

Mean centre is only one possible scaling method (Eq. 1.2). Other scaling methods that were used included auto-scale (Eq. 1.3) and Log_{10} followed by mean centre (Eq. 1.5), both these alternative scaling methods produced 100% accurate models, so no decisions about which is best can be made.

2.3.4. Bootstrapped PLSDA

Just building one model with all the samples in has confirmed that PLSDA is a possible method for discriminating between sample types. However, it hasn't given any information about the number of samples required to build the model. By only using a fraction of the samples randomly selected from the group, a model can be build that can be evaluated to find valuable information about the quality of the data and pre-treatment methods. This process can then be bootstrapped (repeated with a different set of randomly selected samples). Taking only the LOO accuracy from the PLSDA analysis the following graphs were created from PLSDA models built from various numbers of samples and 20 bootstraps.

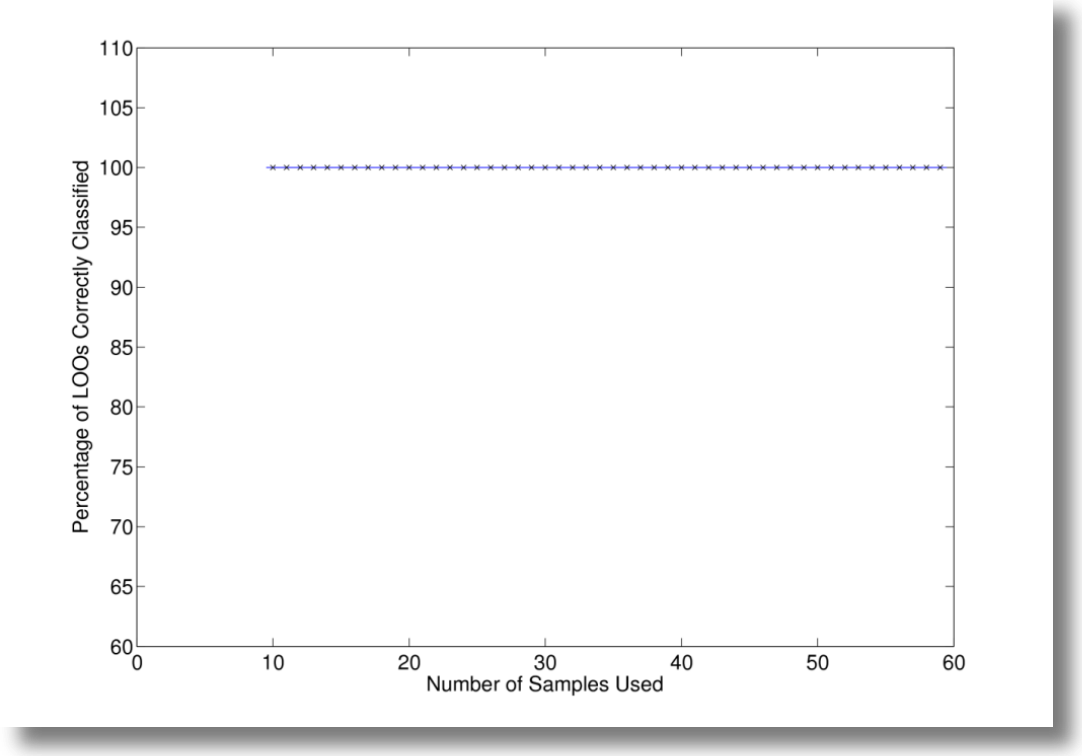


Figure 2.20 - The LOO accuracy of the PLSDA models built using only a specific number of samples (accuracy from s.d. of 20 bootstraps is shown as error bars). The data was mean centred (Eq. 1.2) before analysis.

Looking at the bootstrapped PLSDA LOO results (Figure 2.20) the PLSDA appears to be able to discriminate between the two mixtures involved with very few samples

(10). As bootstrapping requires a random selection of samples using less samples than this would increase the risk that only one sample would be chosen for one of the mixtures, if this happens the LOO algorithm will fail when it builds a model with the samples of only one group. The fact that the model is still successful at this level shows the simplicity of the dataset but it does validate the process. Looking at the other pre-treatment methods (results not shown) reveals that Log_{10} followed by mean centre pre-treatment performs equally flawlessly, but the auto-scaling requires more samples (greater than 40).

2.4. Analysis with Trans SIFT mk. 3

During the process of the project, a newer SIFT-MS machine was acquired. To evaluate the possibility that the multivariate protocol is machine independent the process outlined above was repeated using mixtures created using the same volatiles as above so that the responses could be compared. New mixtures had to be created as there was only enough of them made in the first instance for the measurement using the mk.2 SIFT-MS machine. The new SIFT-MS machine (mk. 3) has an apparently improved detector, so it should be able to detect (with a better accuracy) volatiles up to and above 160 m/z.

2.4.1. Univariate & Classical Approach

There are far more peaks in the spectrum recorded by the newer mk.3 SIFT-MS machine (Figure 2.21) when compared with a spectrum of a similar sample (new mixtures were created so any sample would be slightly different) recorded using the mk.2 SIFT-MS machine (Figure 2.2). It appears that the machine is more likely to detect trace volatiles present by either contamination during sample preparation or just introduced to the SIFT-MS analysis process from the lab air. The biggest peaks are generally the ones associated with mixture A (mixture A contains ammonia, methanol and acetone and the peaks associated with these VOCs are detailed in Table 2.1.

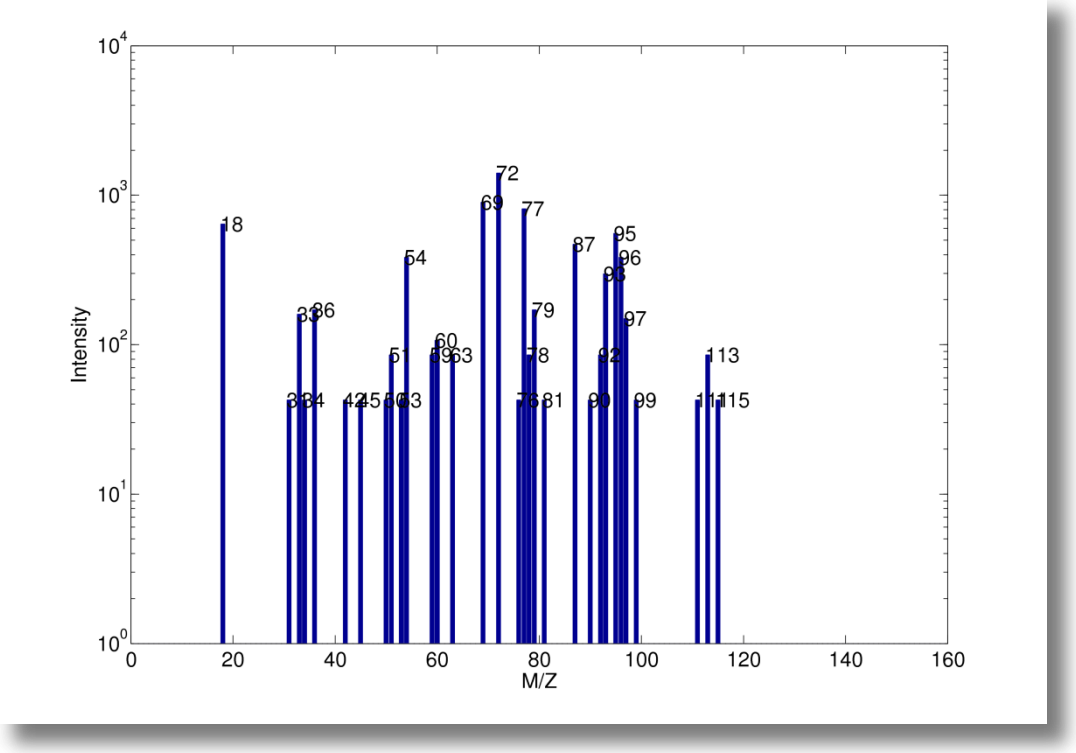


Figure 2.21 - The spectrum from a sample of mixture A (sample no. 22) recorded using the H_3O^+ precursor and the newer mk.3 SIFT-MS machine. Peaks associated with m/z ratios caused by the precursors have been removed.

The spectrum recorded for a sample of mixture B (Figure 2.22) shows a similarly high number of peaks. The larger peaks are again associated with the expected m/z.

One of the down sides of the newer mk.3 SIFT-MS machines more powerful detector is the floating baseline of detection. This is the level below which the m/z counts cannot be accurately determined but they are above the level where they would not be detected (a non-zero but unquantifiable amount). This can be seen by the fixed values of the lower intensity peaks on Figure 2.21 and Figure 2.22. These samples have been detected in quantities below the currently determined measurable level of the detector (which is determined by voltage, temperature and some other factors) i.e. there are trace amounts detected in the sample. As the detector is used, the temperature builds up and this level increases. The manufactures recommendation is to increase the rest

time between samples being recorded to keep this level at a minimum which will decrease the impact on the analysis.

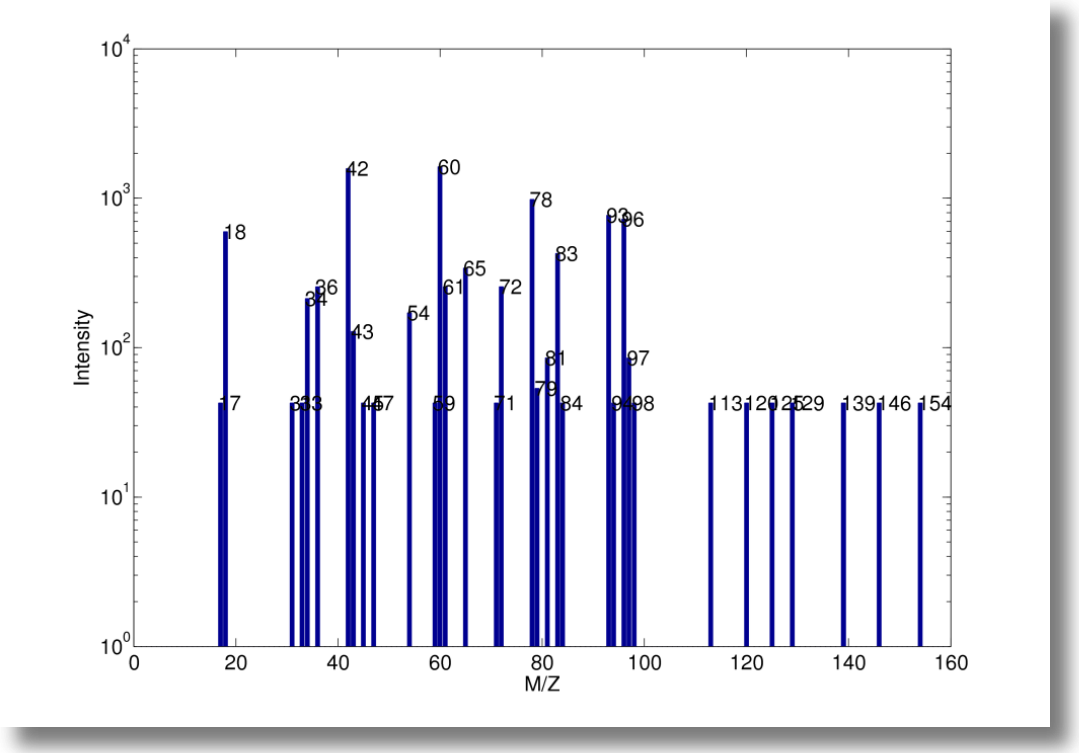


Figure 2.22 - The spectrum as recorded by the newer SIFT-MS machine using the H_3O^+ precursor for a sample (sample number 1) of mixture B. Peaks associated with m/z ratios caused by the precursors have been removed.

From a chemometrics point of view, there are methods for baseline corrections (Otto, 1999a), however, these work with data whose main peaks are affected by the changing baseline which is not the case here.

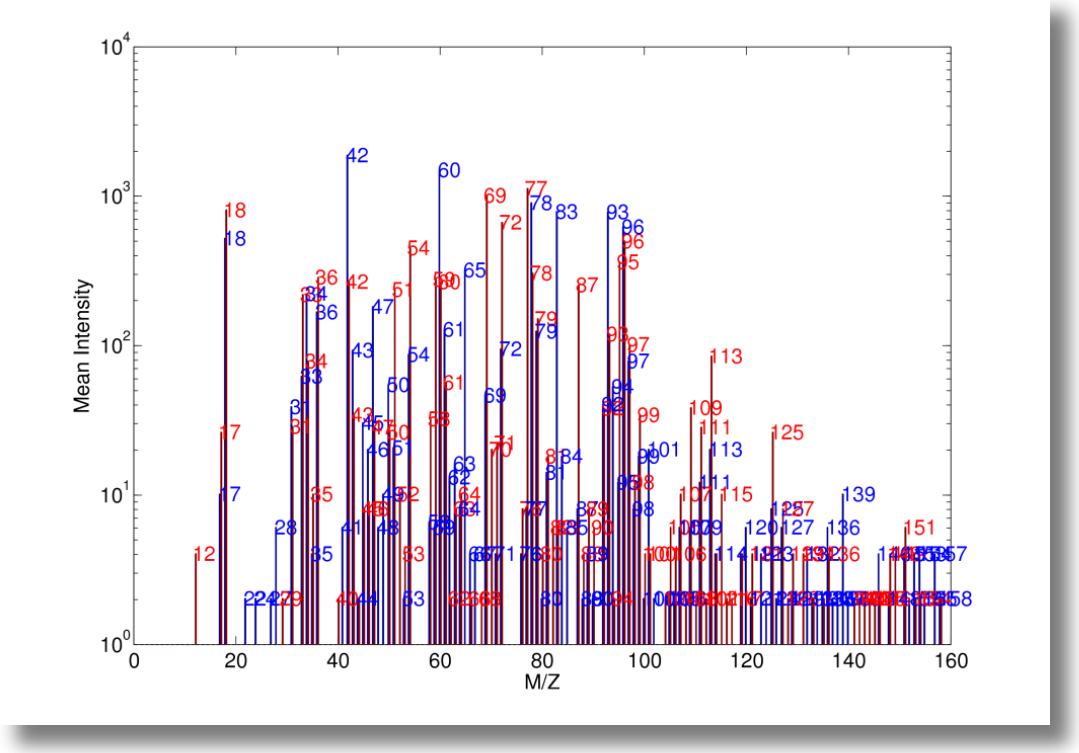


Figure 2.23 - The average intensity for the two mixtures as recorded using the H_3O^+ precursor and the newer mk.3 SIFT-MS machine. Red bars are the average counts of mixture A and blue bars are the average counts of mixture B.

The average intensity graph goes even further to show the number of trace VOCs that the machine is detecting. There are low levels of many m/z that have nothing to do with the VOCs used to create the two mixtures. The m/z associated with the VOCs from the mixtures appear to have some considerable differences between the spectra of the two groups which is reassuring.

2.4.2. PCA

So the data is reasonably similar apart from slightly more noise created by the more sensitive detector. Does PCA still show the differences between the sample types?

The PCA score plot (Figure 2.24) shows clear discrimination between the samples with separate clusters for each of the mixtures. Looking at the loadings of the two

components used above, PC1 is influenced by m/z 42, 60, 78, 96 (acetronitrile), 83 (ethanol) and 93 (toluene). PC2 is influenced by m/z 18, 54, 72 (ammonia), 69 (methanol) and 77 (acetone).

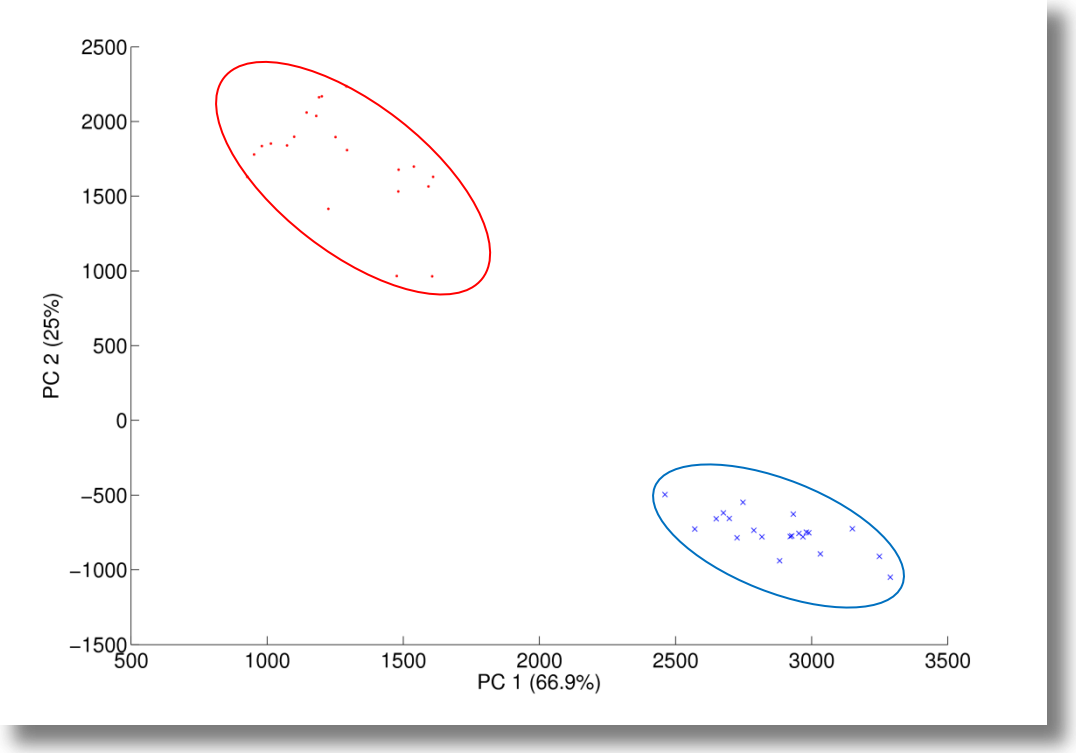


Figure 2.24 - The PCA scores plot from the dataset collected using the newer SIFT-MS for the two mixtures. Red dots are mixture A and blue crosses are mixture B. The precursor m/z have been removed but no scaling techniques have been used. The data was collected using only the H_3O^+ precursor. Clear and separate clusters have been drawn around the samples in the corresponding colours.

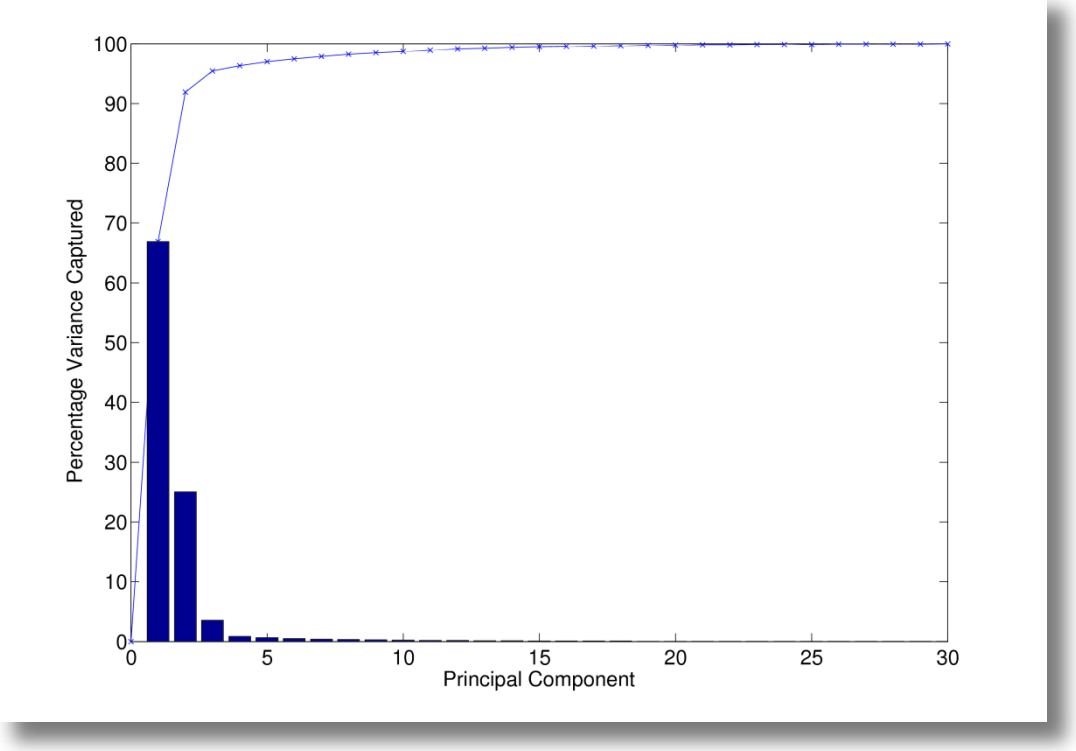


Figure 2.25 - The variance captured by each of the first 30 principal components (bars) and the cumulative sum (line) for the above PCA analysis (Figure 2.24) created using the H_3O^+ spectra data recorded using the newer mk.3 SIFT-MS machine with the m/z ratios associated with the precursors removed

This suggests that the new SIFT-MS is still compatible with the data analysis methodology and can also be used to search for biomarkers. Indeed, when comparing the PCA scores plots of the original machine Figure 2.10 with the similar PCA scores plot of the new machine Figure 2.24, it could be argued that the clusters formed by the new machine are more tightly packed.

2.4.3. PLSDA

PLSDA was performed on the data as before (with the older SIFT-MS). All three scaling methods produced 100% accurate models (results not shown). This either suggests that the newer SIFT-MS machine produces data that can achieve better

results than the older machine or data that is different enough that the auto-scaling algorithm (Eq. 1.3) is more compatible.

2.4.4. Combined dataset

A dataset was created from the spectra of the two machines to directly investigate the differences between the machines. The spectra between the two machines do indeed look different (Figure 2.2 and Figure 2.3 compared with Figure 2.21 and Figure 2.22) as the univariate methods used in previous sections have shown. However, the multivariate method can be applied to the data set as a whole and these methods showed some interesting results.

The score plot (Figure 2.26) shows two separate clusters for the different samples. It also shows two separate clusters for the two machines. When comparing this combined score plot with the two separate ones (Figure 2.10 and Figure 2.24) we can now see that the newer SIFT-MS machine appears to show better discrimination from the increased sensitivity of the detector. This, however, might be simply down to a magnitude difference between the responses of the two machines. A full investigation into a method for combining the dataset from the two machines is beyond the scope of this thesis but this initial analysis shows that it may be possible.

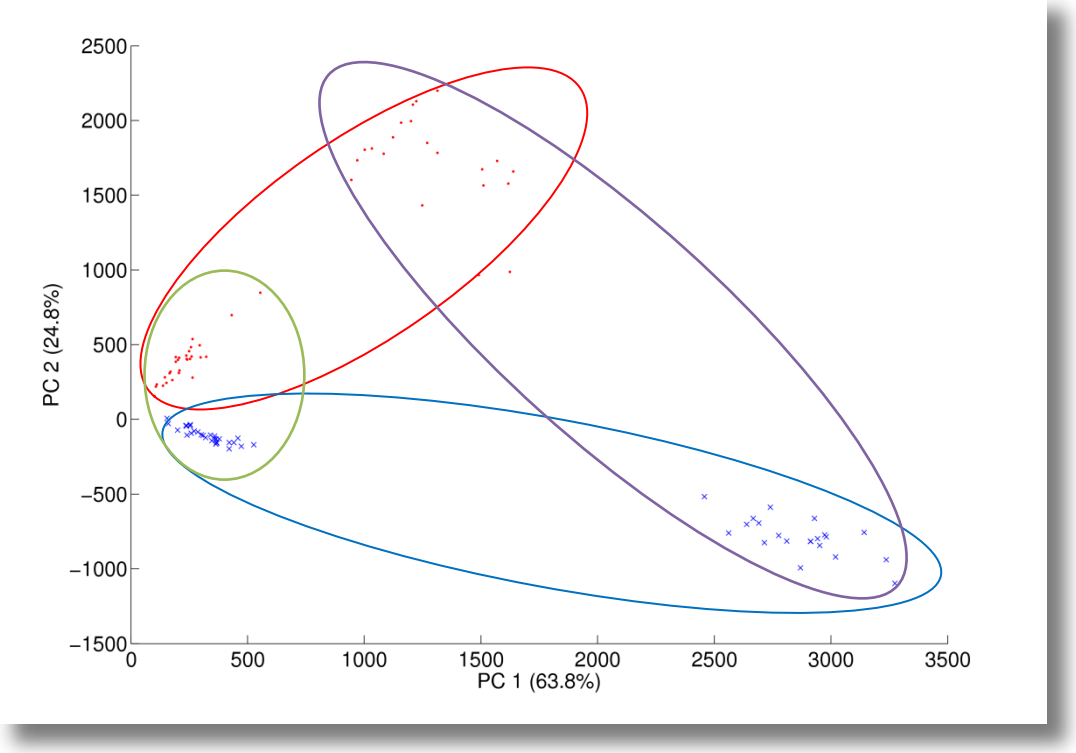


Figure 2.26 - The PCA scores plot from the data collected from both SIFT-MS machines. Red dots are mixture A and blue crosses are mixture B. The precursor m/z have been removed but no scaling techniques have been used. The data was collected using only the H_3O^+ precursor. The cluster of red and blue samples near the zero of both axis are samples recorded using the older mk.2 SIFT-MS machine and the two clusters further from the origin of both axises are samples recorded using the newer mk.3 SIFT-MS machine. The samples recorded using the mk.2 SIFT-MS machine have been outlined using the green oval and the samples recorded using the newer mk.3 SIFT-MS machine have been outlined using the purple oval.

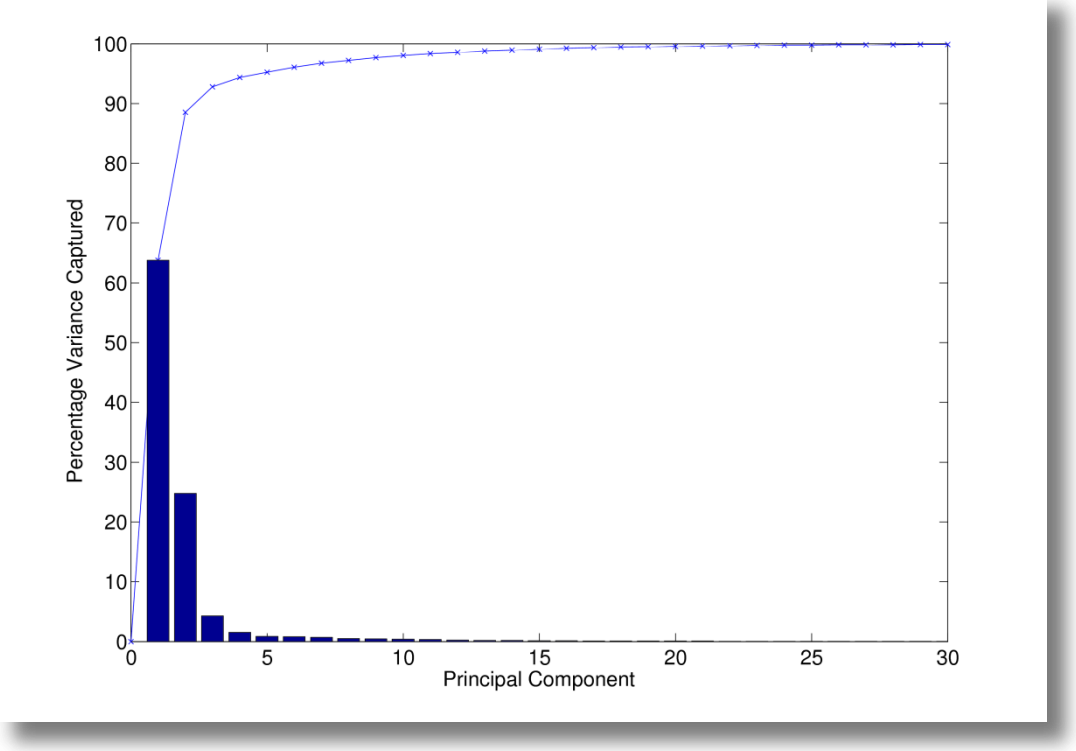


Figure 2.27 - The percentage of variance captured for the first 30 principal components (bars) and the cumulative sum (line) for the above PCA analysis (Figure 2.26)

The PCA scores plot (Figure 2.26) not only shows a clear discrimination between the two SIFT-MS machines but it also shows a clear discrimination between the sample types. This was a little unexpected. Machine drift is usually a major problem with other techniques such as e-nose (Dewe, 2009) when data is compared using only one machine which results in a complete incompatibility when comparing data from different machines (Knobloch et al., 2009c). PC1 captures the discrimination between the machines and this is associated with m/z 18 (ammonia), 42, 60 and 78 (acetonitrile). Then the discrimination seen between the mixtures on PC2 is associated with the same m/z peaks plus 69 (methanol), 72 (ammonia) and 77 (acetone). Taking this further the PLSDA results (not shown here) show 100% accuracy in prediction of the LOO when the data is treated with Log_{10} and then mean centred using the H_3O^+ spectra alone. With the combined dataset (all three precursors) both the mean centre pre-treatment and the Log_{10} then mean centre pre-treatment gives 100% LOO accuracy.

2.5. Comparison and Conclusions

2.5.1. Comparison between SIFT machines

Multivariate analysis methods on the data collected by both SIFT-MS machines produced excellent results, obtaining high classification rates and identifying the expected markers in this proof of concept trial. The methods achieved very good results, with both the PCA showing clear discrimination between the samples of the two mixtures and PLSDA classifying between the two sample types with 100% accuracy. This is the best possible result and is better than some similar tests with electronic nose (Knobloch et al., 2009c). One would expect that GC-MS could achieve similar results, however sampling times are considerably longer (generally measured in hours rather than minutes) (Kaal et al., 2009; Schulz et al., 2009).

The data might have different needs when problems arise in the future as there are differences between raw data collected by the two machines. For example, the new SIFT-MS is prone to background noise across all m/z values whereas the older SIFT-MS machine is prone to random spikes, however, little evidence for either situation has been shown in the above experiment.

The discrimination on the PCA scores plot was confirmed to be caused by the expected molecules by the high weight values on the variable loadings plots. This suggests that if discrimination can be seen on the scores plot the markers contributing to the discrimination will be determined on the loadings plots suggesting that this method is a suitable aid in the search for biomarkers. Indeed this method has been documented with other technologies such as MALDI-MS (Harrington et al., 2005).

With the above manufactured problem, the data from the two machines could be combined and the analysis still produced excellent results. It would not be expected that the datasets collected from biological samples to be able to be combined as it was above. The differences between biological samples are more complex and hence the differences between the machines may make the results harder to spot. To maximise

the difference between the samples and minimise all other variance any future work should be collected using a single SIFT-MS machine where possible (or at least machines of the same version).

2.5.2. Conclusions

The general data analysis methodology can now begin to be finalised.

Firstly before any analysis can be performed the m/z peaks caused by precursors need to be removed. They are much larger than the peaks caused by the sample and they have large variance when compared with the variance of the VOCs under investigation. The precursors also have little to do with the sample being investigated.

Second, univariate processes are not completely without benefit. The PCA analysis leads to m/z peaks that have high variance but the presentation of this single m/z as a bar graph for all samples immediately shows information about repeatability (i.e. is the peak always present in at least one of the sample groups?), variation within the sample groups (is the difference between the means of the two groups significant enough to be used as a marker?). A good knowledge of the univariate analysis methods is definitely required for thorough analysis. Of course, more complex univariate techniques could be used such as t-tests to confirm the significance between the classes. If this was scripted and looped, the results ranked and tabulated, it would likely yield some similar markers to that of the multivariate techniques. However, multivariate techniques take into account the relationships and correlations between different m/z ratios. The difference between the sample groups of a single m/z might not be significant on its own and discrimination might be achieved by the combination of several m/z ratios. This possibility cannot be investigated using univariate techniques alone.

Third, in this manufactured problem, the PCA analysis above showed clear discrimination. No further pre-treatment was needed for the discrimination after the m/z caused by the precursors had been removed. However, it is expected this is more

due to manufactured volatiles mixtures being considerably different making classification easier than the power of PCA. An investigation into the impacts of pre-treatment methods is still required on samples where the differences are likely to be more subtle.

For the data to be analysed using PLSDA the spectra need to be mean centred and possibly Log_{10} scaled before mean centring (both methods gave equally perfect results in this case). Auto-scaling the data proved to be ineffective with data from the older mk.2 SIFT-MS machine. One possible reason could be that the SIFT-MS is semi-quantitative and this information is removed when the standard deviation is normalised. Many small m/z peaks from VOCs that were not present in either of the mixtures but that are being detected in trace amount were having their impact elevated.

Finally, the results from the combined dataset were equal to the best discrimination from the analysis of the individual precursors. In the search for biomarkers where the loadings from PCA are important it is probably easiest to analyse only the H_3O^+ precursor information as the spectra from this source is most easily understood. However, when it comes to the analysis by PLSDA there is no benefit from the analysis of each precursor channel separately. The scores and loadings matrices produced during the production of the model have questionable value. Firstly, the scores plots from PLSDA can show clear discrimination even when there is a completely random dataset used which should not produce any discrimination (Brereton, 2006) and secondly, if this is indeed the case knowing the features causing this discrimination would appear to have little value.

Now that the multivariate techniques have been shown as a possible method for the analysis of large quantities of SIFT-MS data, proof that the methodology will work with biological samples is required. The next chapter will look at applying the methodology developed here to a data set collected from the analysis of real biological samples and what changes, if any, need to be made to account for the extra complexity of the data.

3. Application of methods to determination of tuberculosis status from serum samples taken from Wild Badgers

3.1. Introduction

The first case study is an investigation into the possibility of using SIFT-MS and multivariate techniques to differentiate between the headspace VOCs of serum samples from wild badgers, with a view to determining whether the methodology developed in the previous chapter can determine which badgers are infected with bovine tuberculosis (*Mycobacterium bovis*). The badgers (*Meles meles*) were captured as part of a veterinary laboratories agency (VLA) and department for the environment, food and rural affairs (DEFRA) trial looking at the prevalence of tuberculosis in their population. The samples obtained from the badgers have no information about sex, age, diet or general health and hence are likely to exhibit unexplained differences caused by factors other than tuberculosis status. These challenges are, however, normal when working with wild animals and are mostly unavoidable. This makes the task of disease discrimination a very difficult challenge but as this is one of the simpler datasets (samples have a single classifying status - they either have tuberculosis or they do not) and as a relatively large dataset this forms an ideal first challenge for the methodology developed in Chapter 2.

Secondly, the status of tuberculosis is a very important case study for both human health and agriculture and the project was funded for this reason by DEFRA and the VLA specifically to investigate this dataset. The samples were obtained over a few years as part of the Randomised Badger Culling Trial (RBCT) organised by an independent scientific group created to investigate the findings of the Krebs report produced previously (Krebs, 1997; Bourne, 2007).

3.2. Data Collection and Measurement

The data was collected using the SIFT-MS mk.2 over a period of three months. Small vials of frozen serum were shipped to Cranfield University from the VLA. These samples were thawed and 1.6ml of the sample was injected into 31 cm length (700 ml volume) Nalophan bags which were then filled with zero grade air (BOC) and incubated at a temperature of 40°C. The first two thirds of each sample was used to record e-nose traces (Knobloch, 2010) and the final third of the volatile headspace was sent through the SIFT-MS machine using a 5 minute full spectra scan, alternating through all three precursors. For a more detailed investigation into the sample preparation please see (Knobloch, 2010).

A total of 245 samples were analysed blind and in the order they were delivered. Later the samples were identified, with 51 being positives and 194 being negatives. The disease status was determined by post-mortem using a culture test of lung tissues. The culture test takes approximately 12 weeks. A positive result identifies the animal as TB positive but a negative result only means that TB was not detected and not that it was TB free. The exact accuracy of the culture test is unknown but generally seen as the gold standard test for TB (de la Rua-Domenech et al., 2006).

3.3. Data Analysis

3.3.1. Univariate & Classical Approach

The spectra are from biological samples and hence contain considerably more volatiles than the mixtures analysed in Chapter 2. There are also large variations between all the samples with volatiles being present in spectra from one sample and absent from another.

As found in the previous chapter the m/z peaks caused by the precursors (Table 2.2) needed to be removed first and then the above spectra produced. Once these had been removed the above figure (Figure 3.1) was created. There are familiar peaks (m/z 18 for ammonia as an example) that were seen in the mixtures created for chapter 2

(Table 2.2) but there are also many more peaks associated with VOCs that were not part of the mixtures trial. This is indeed expected as the synthetic mixtures were designed to contain VOCs that were known to exist in biological samples but in a much simplified form.

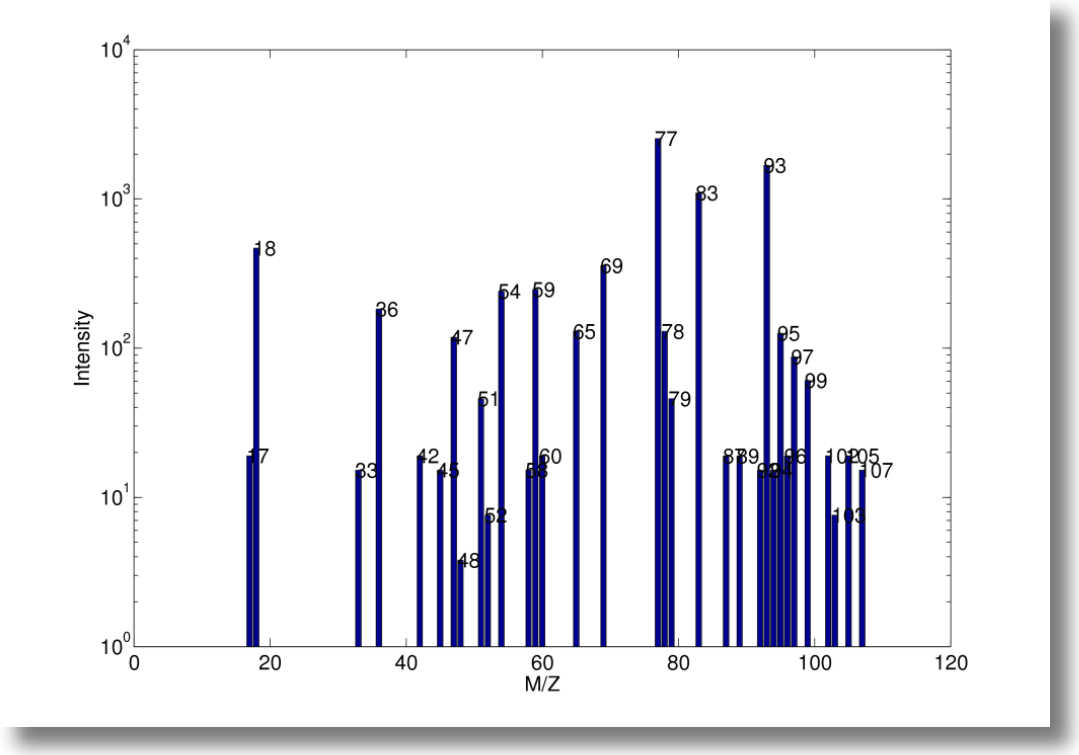


Figure 3.1 - The spectrum of sample 1 as recorded using the H_3O^+ precursor of the mk.2 SIFT-MS machine. This serum sample comes from a badger that was culture negative for Tuberculosis.

The two spectra (Figure 3.1 and Figure 3.2) initially look quite similar (they have the same basic shape) but after further analysis and a closer look at the trace VOCs, significant differences become apparent. The problem is in knowing which differences are caused by sex, diet, age or a vast number of other possibilities that have nothing to do with tuberculosis and which changes are the ones linked to tuberculosis. Another possibility is that some of the badgers might be suffering from other diseases.

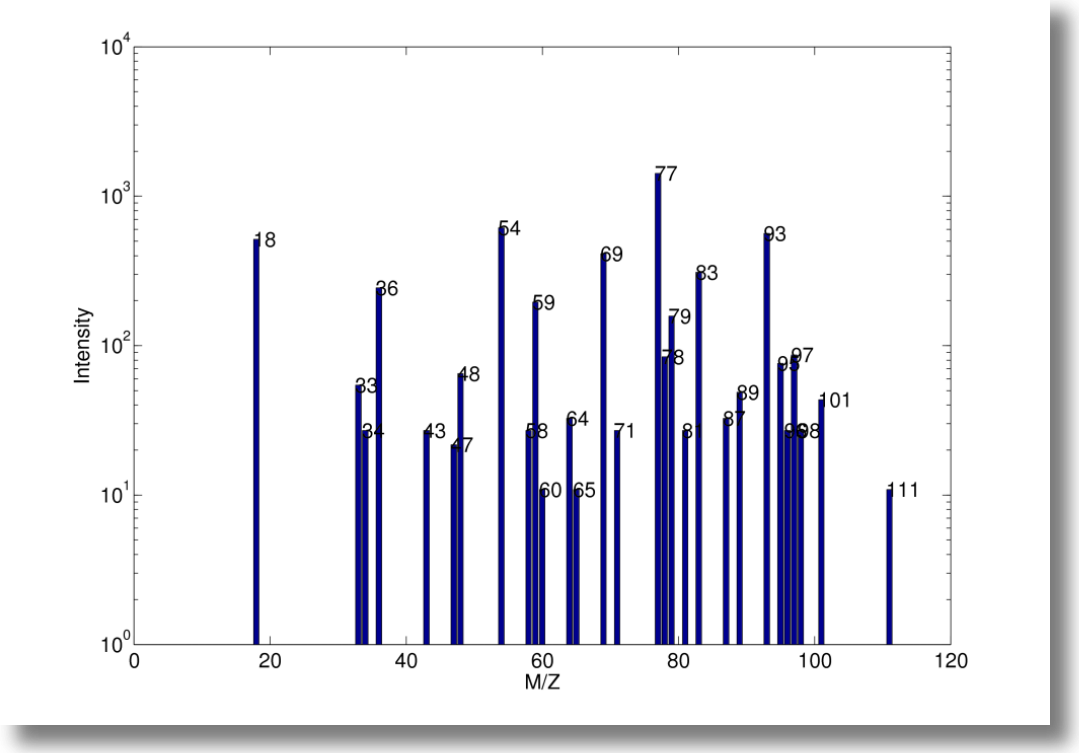


Figure 3.2 - The spectrum as recorded using the H_3O^+ precursor using the mk.2 SIFT-MS machine for sample 201. This serum sample comes from a badger that was confirmed as having Tuberculosis using the culture test.

Looking at the average peaks of each sample group (Figure 3.3) was obtained.

Some slight differences in the means can be seen, but which ones are significant? One way to check is to look at the number of samples that have counts (i.e. non zeros) at each m/z ratio on the spectrum. In the synthetic mixtures dataset of chapter 2 it would have been done with a single graph however, with this dataset the sample groups have different numbers of samples in each group (51 positive to 194 negative) and the graph would look skewed towards the negative hence two graphs are required.

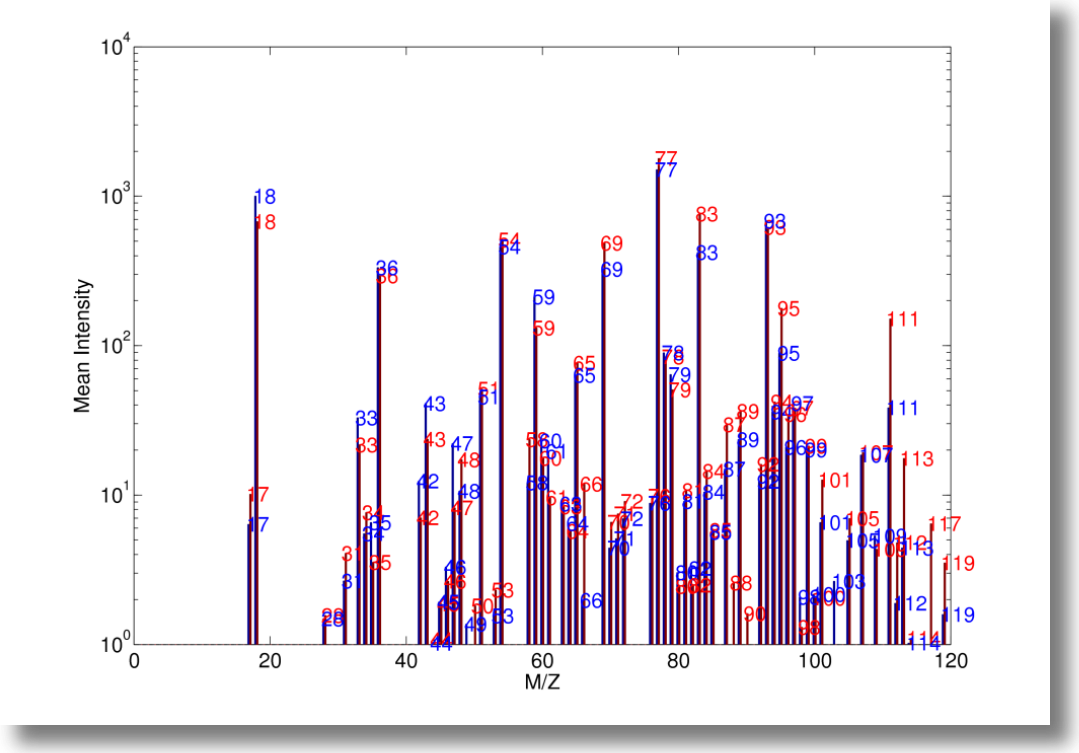


Figure 3.3 - The average spectra as recorded using the H_3O^+ precursor channel using the mk.2 SIFT-MS for each sample group. Blue are samples as confirmed positive using the culture test and red are the ones that were not (negative).

Figure 3.4 and Figure 3.5 show the coverage and repeatability of the range of m/z recorded using the SIFT-MS machine. It should be noted that although the SIFT-MS machine is technically able to record m/z well above 120 m/z the number of counts above 115 is low. Therefore the chance of getting reliable results is reduced. It also shows that some of the differences seen in the trace VOCs are caused by a very small fraction of the samples.

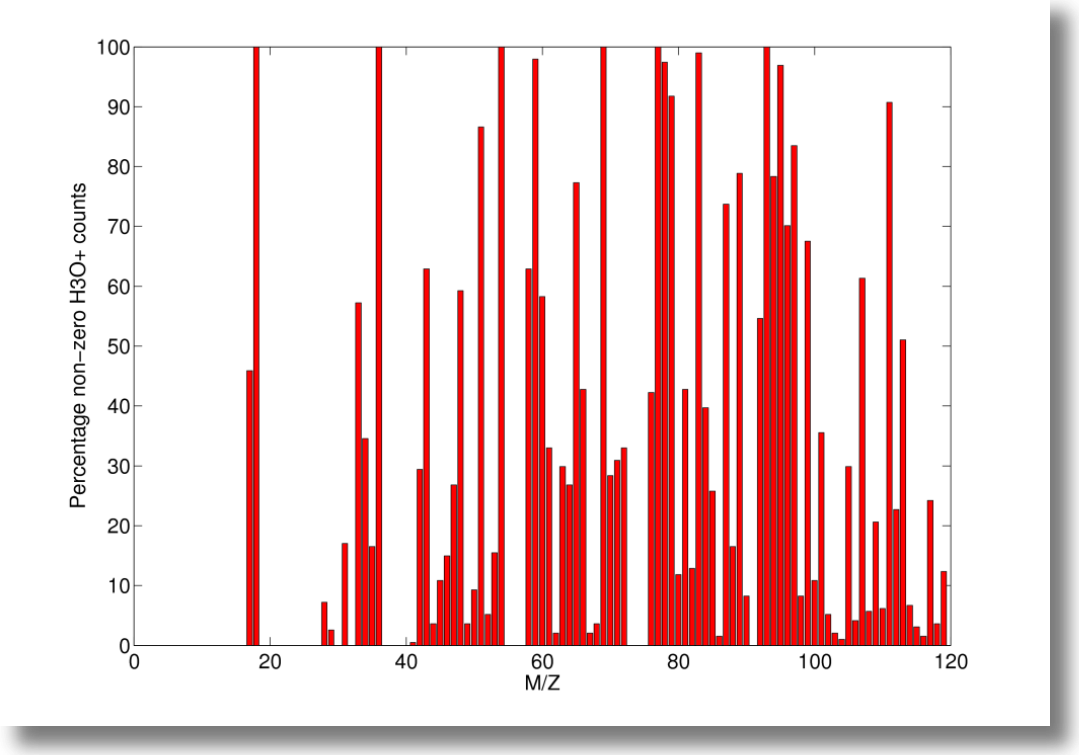


Figure 3.4 - The percentage of non zero m/z counts for the spectra as recorded using the H_3O^+ precursor of the mk.2 SIFT-MS for serum samples from badgers that were not confirmed as tuberculosis positive (i.e. negatives)

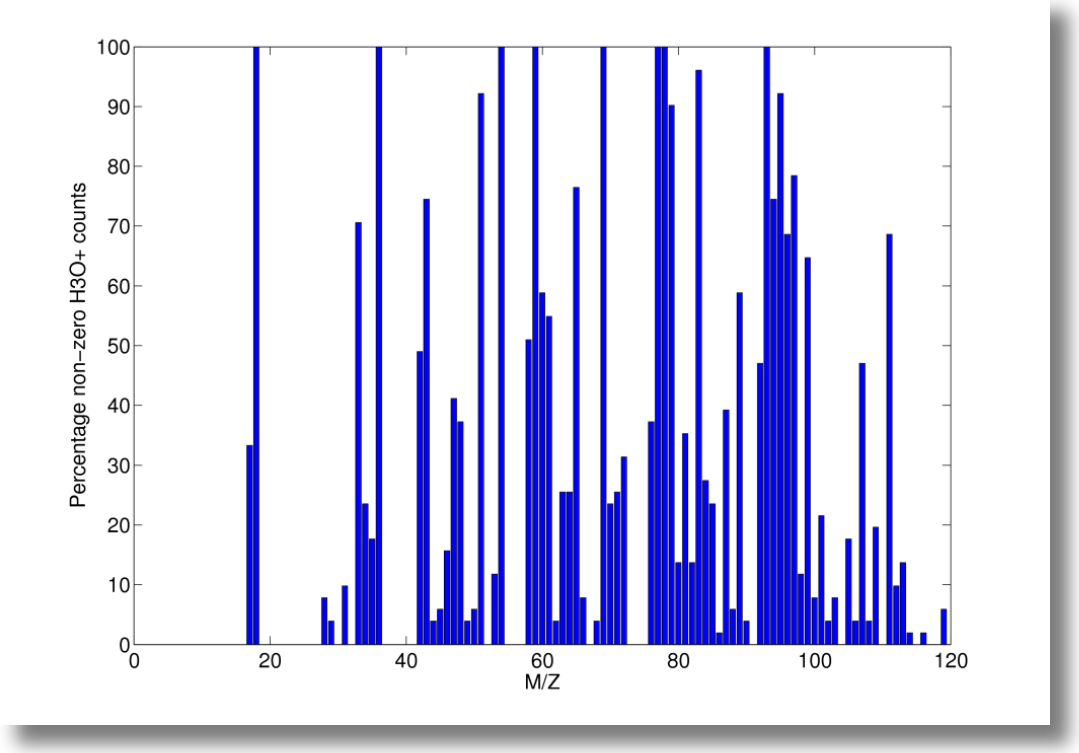


Figure 3.5 - The percentage of non zero m/z counts from the spectra recorded using the H_3O^+ precursor of the mk.2 SIFT-MS machine for serum samples from badgers confirmed as tuberculosis positive

3.3.2. PCA analysis

The initial PCA analysis (Figure 3.6) is not as promising as the results for the previous chapter. There is very little discrimination between the sample types. There does however appear to be some structure. Looking at the loadings plot the structure comes from acetone (m/z 77) for PC1 and ethanol (m/z 88) for PC2. The levels of acetone are far higher than would be expected if these were human samples (Turner et al., 2006a) and the vets postulated that these elevated levels might be caused by a breakdown product of the anaesthetic given to the badgers before the blood sample was taken. Badgers are also a foraging animal and the varying levels of ethanol are likely to be diet related (natural fermentation of fallen fruit for example) more than anything else.

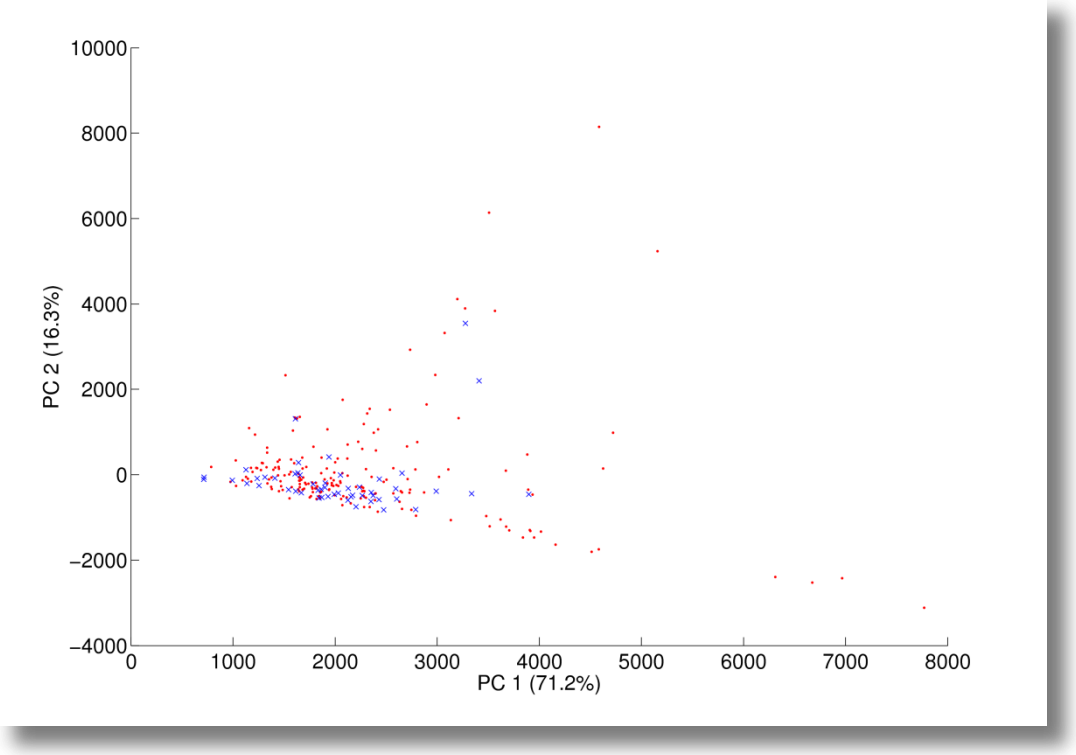


Figure 3.6 - PCA score plot of the H_3O^+ precursor spectra as recorded using the mk.2 SIFT-MS machine for the serum samples collected from the wild badgers. Blue samples are tuberculosis culture positive and red samples are tuberculosis culture negative. m/z peaks associated with the precursors have been removed.

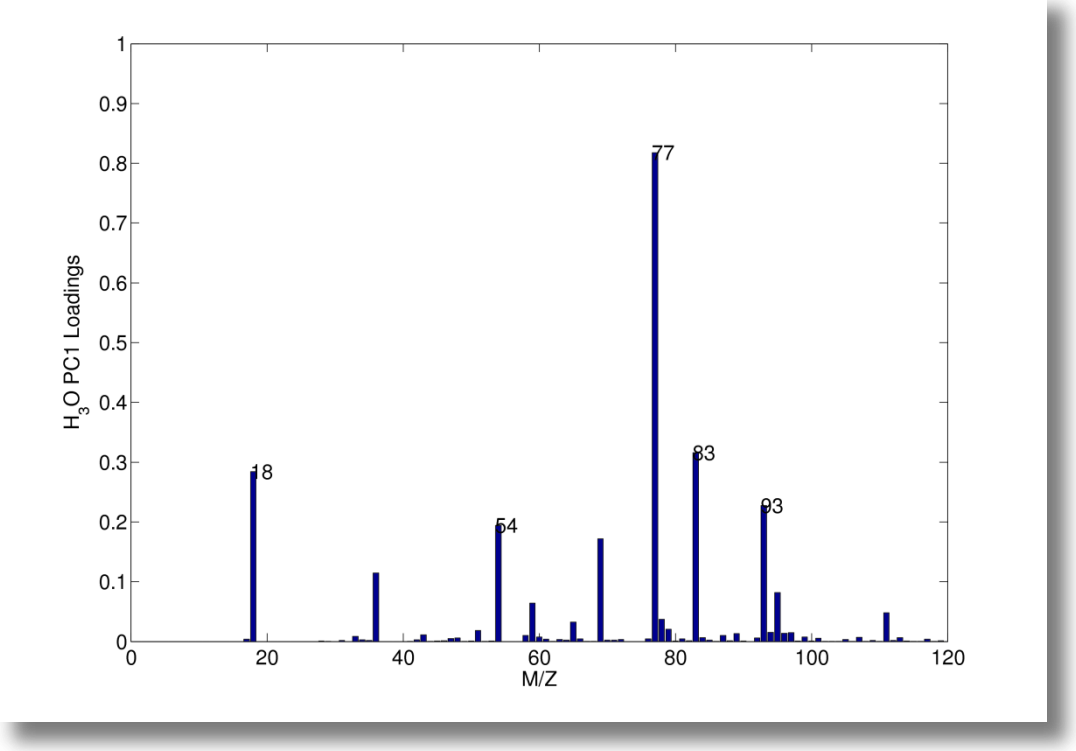


Figure 3.7 - The loadings plot of the first principal component of the above PCA analysis (Figure 3.6). m/z 77 is the dominant cause for the first dimension structure.

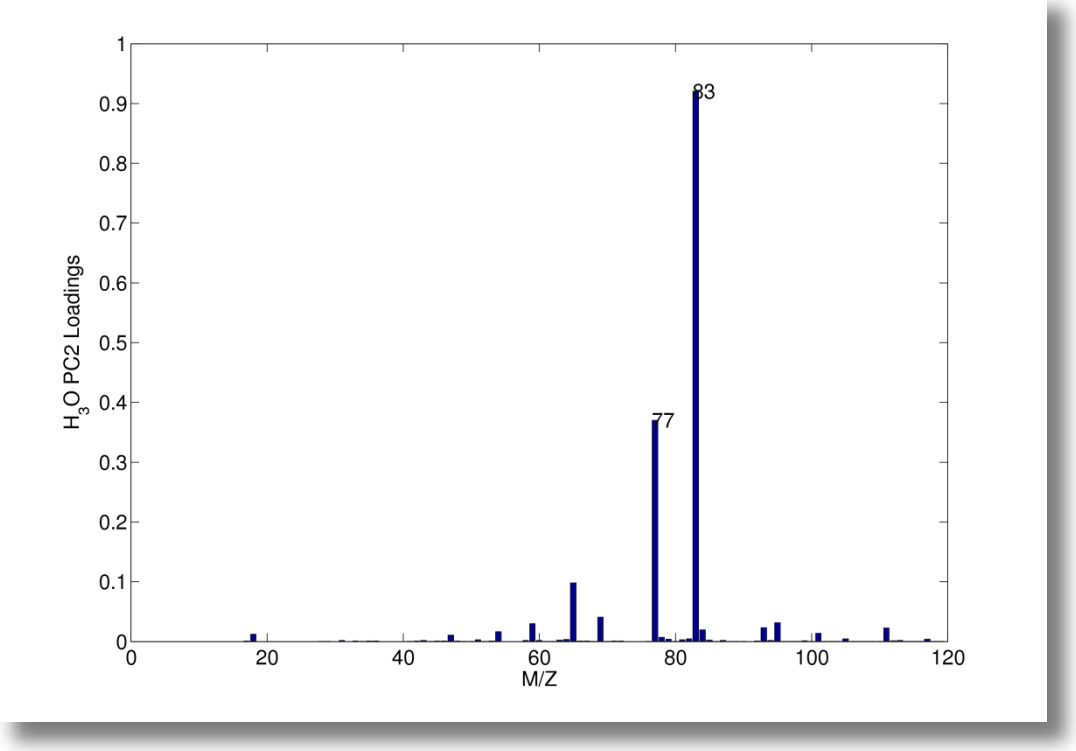


Figure 3.8 - The loadings of the second principal component from the above PCA analysis. m/z 83 is the largest contributor to the structure seen on the second dimension.

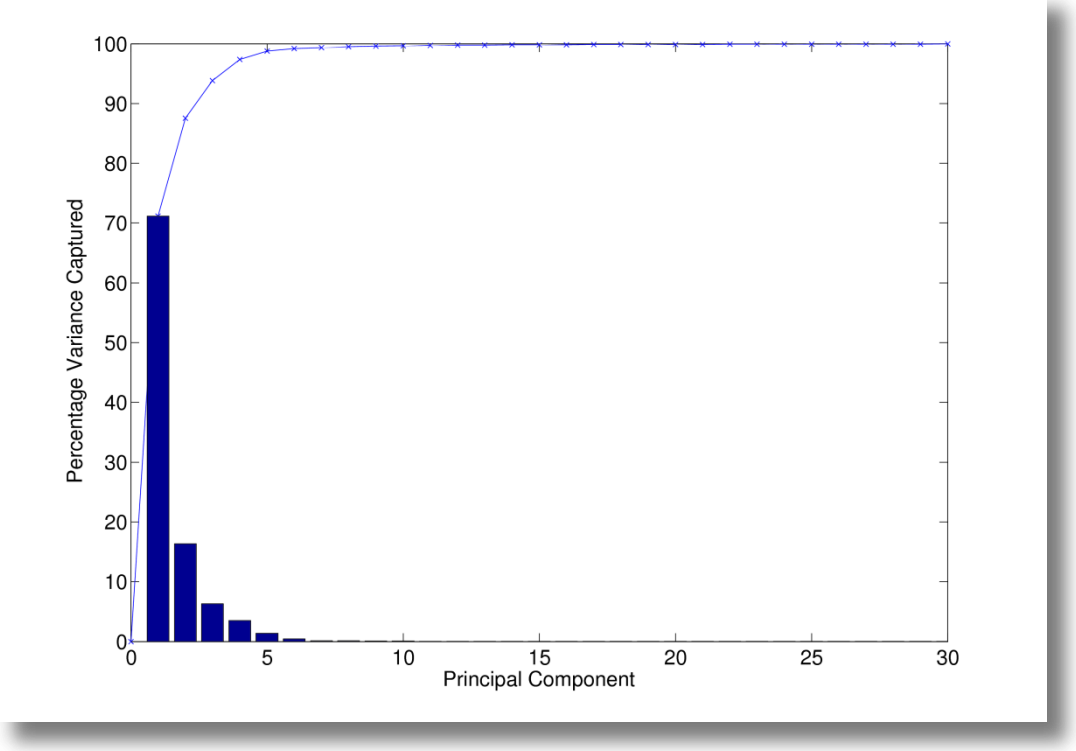


Figure 3.9 - The percentage of variance captured by the first 30 principal components (bars) and the cumulative sum (line) of the above PCA (Figure 3.6)

After acetone and ethanol have been zeroed, the PCA plots begin to show some discrimination, see Figure 3.10. The discrimination is only slight, no positive samples occur above 350 on the second component axis. The causes for this are ammonia (m/z 18, 36 and 54), methanol (m/z 69) and an unknown (m/z 111). Looking at m/z 18 (ammonia), it appears that the infected generally have a higher count than the uninfected.

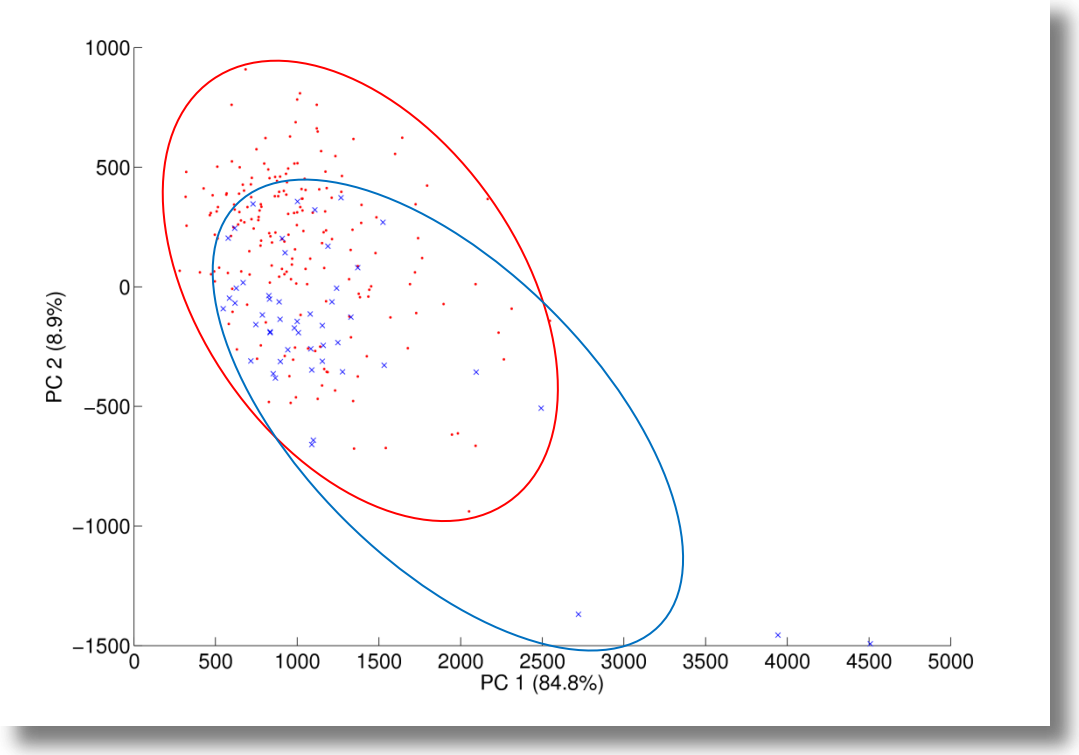


Figure 3.10 - PCA scores plot of the H_3O^+ precursor spectra recorded using the mk.2 SIFT-MS machine for the serum samples collected from wild badgers. m/z peaks associated with the precursors as well as m/z 59, 77, 95 (acetone) and m/z 47, 65, 83 (ethanol) have been removed. Red samples are tb negative and blue are positive

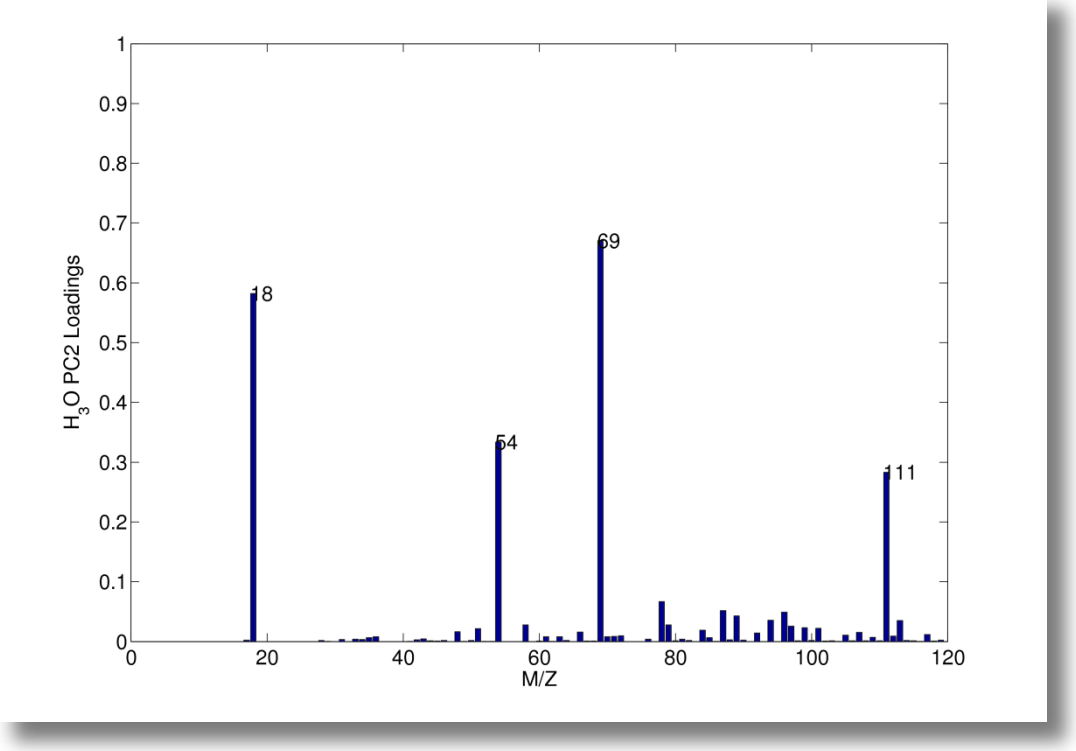


Figure 3.11 - The loading of the second principal component from the above PCA analysis (Figure 3.10) showing that the slight discrimination is caused by m/z ratios 18, 54, 69 and 111

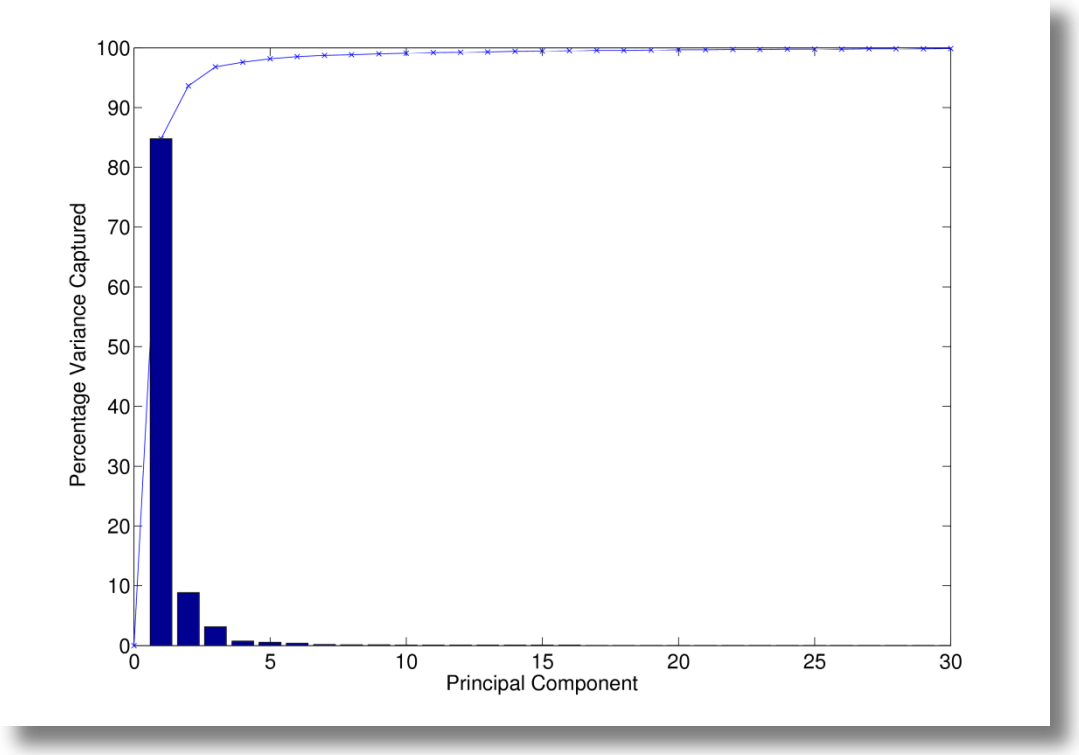


Figure 3.12 - The percentage of variance captured by the first 30 principal components (bars) and the cumlative sum (line) for the above PCA analysis (Figure 3.10)

As a single value, the difference in the levels of ammonia (Figure 3.13) would not be sufficient as a marker due to the large variance seen in both groups. The unknown volatile (m/z 111) has some interesting changes, as shown in Figure 3.14.

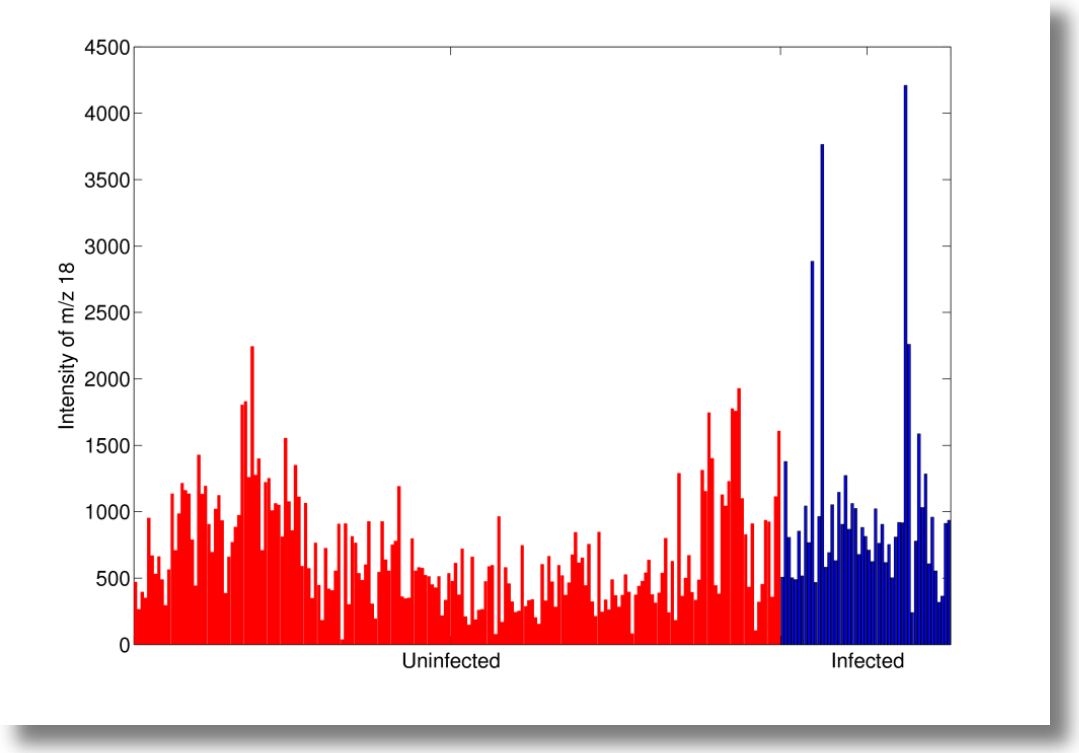


Figure 3.13 - The counts for ammonia (m/z 18) for all samples of infected (blue) and uninfected (red) as recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine.

As it should be clear from Figure 3.14, the counts of m/z 111 are not associated directly with positive or negative samples. A clear large block of negative samples and a smaller block of positive samples have a higher recorded level of m/z 111 than the other samples. The SIFT-MS mk.2 machine was serviced midway through the cycle of recording the spectra and this appears to coincide with the timestamps of the higher levels of m/z 111. The samples were recorded blind as they were sent from the VLA and apparently were sent in roughly even numbers (of positives and negatives) early within the trial and then much larger proportion of negative samples later in the trial. This has had the unfortunate side effect meaning that the service had a bigger impact on the levels of the uninfected than the infected.

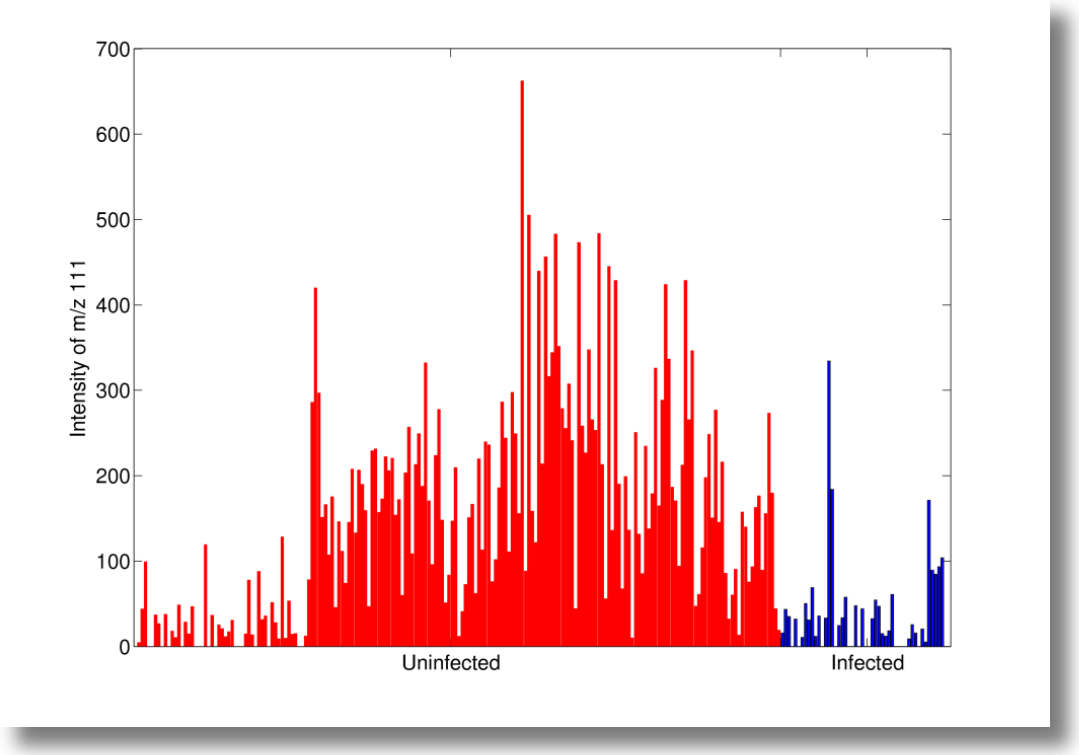


Figure 3.14 - The m/z intensity of m/z 111 as recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. There is clearly a significant structure formed and under further investigation appears to be linked with the service time of the SIFT-MS machine. The x axis above is firstly samples separated by their classification group however within each group the samples are ordered by date of analysis by the SIFT-MS machine. The large block of uninfected samples and much smaller block of infected samples towards the right of each group were recorded after the service and they appear to have higher recorded values.

Looking at range of raw spectra of samples recorded before the SIFT-MS machine service and after the service it appears that m/z 110/111 was about the largest VOCs that the pre-service SIFT-MS could manage and afterward the range was increased to approximately m/z 140. This can also be seen if the percentage of non zero counts for spectra recorded before the service and after the service are plotted (Figure 3.15).

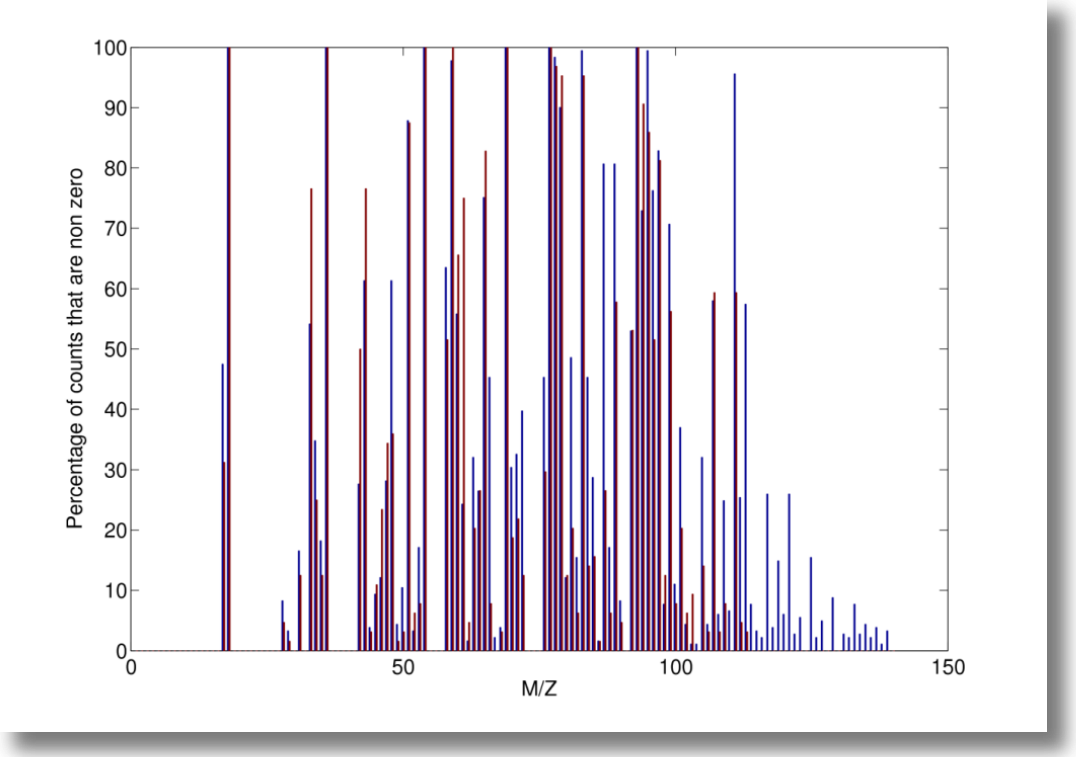


Figure 3.15 - The percentage of non-zero counts for samples recorded before the mk.2 SIFT-MS machine service (Red) and after the machine service (blue) on the spectra produced using the H_3O^+ precursor

It is now apparent that between the machine services the upper range of the SIFT-MS machine decreases. It also appears that there are other features in the spectra that have been affected. Strangely, there are some lower m/z ratio counts that are more favoured to detection with the pre-service machine and some that are less favoured. There is no logical reason for this and one can only assume that there are other causes than the machine service for the differences in the lower than 110 m/z ratio range of the spectra. Indeed, when the averages are plotted instead (Figure 3.16) a different set of differences appear.

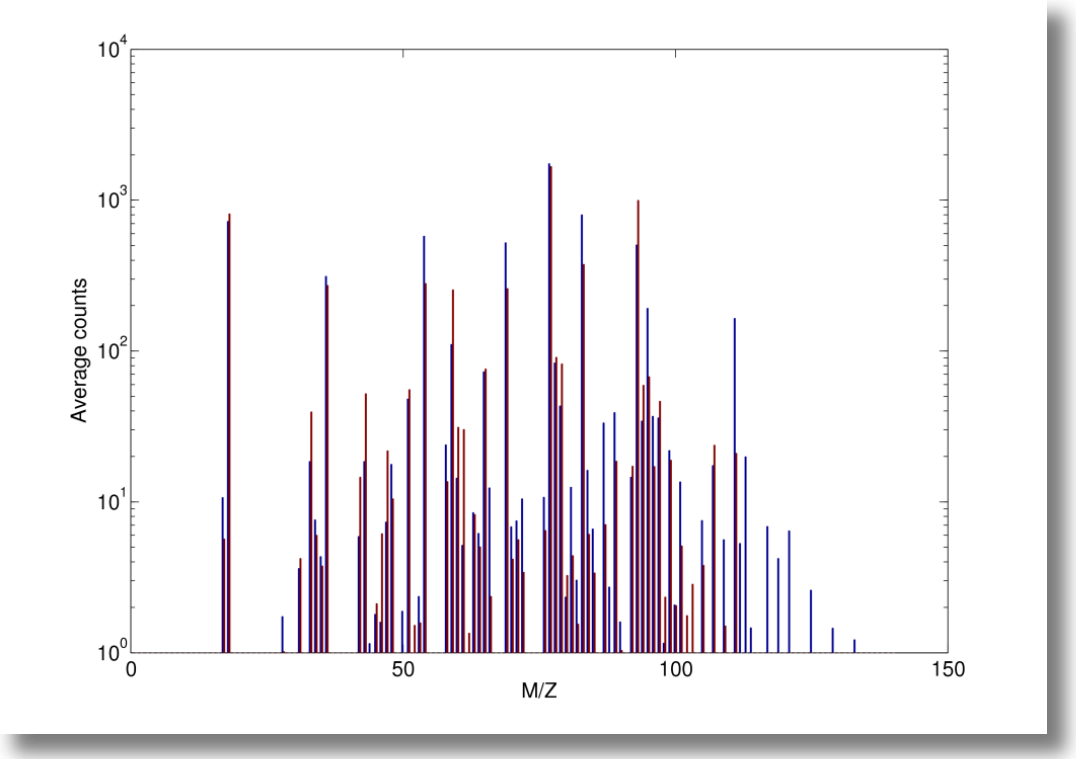


Figure 3.16 - The average counts for the m/z peak on spectra recorded before the mk.2 SIFT-MS machine service (red) and after the service (blue) using the H_3O^+ precursor channel

There are two possible solutions to this problem, each one has drawbacks. The first is to analyse the spectra from before and after the service separately which would seriously reduce the dataset size. The second is to remove the counts for all m/z above the minimum range that the SIFT-MS was able to measure and therefore removing the differences caused by the service and assume the differences between the counts before service and after service are limited to m/z ratios 110 and above.

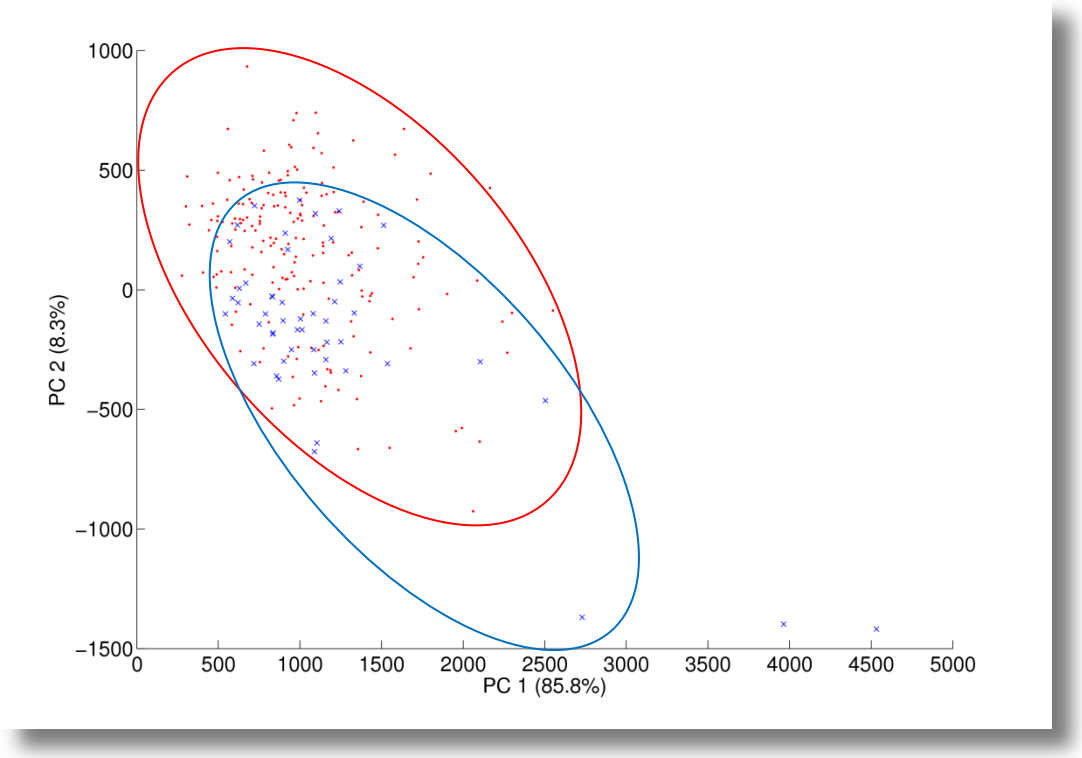


Figure 3.17 - PCA score plots of the serum samples collected from wild badgers recorded using the H_3O^+ precursor of the mk.2 SIFT-MS machine. The m/z associated with ethanol, acetone and the precursors have been removed as have m/z ratios above 109

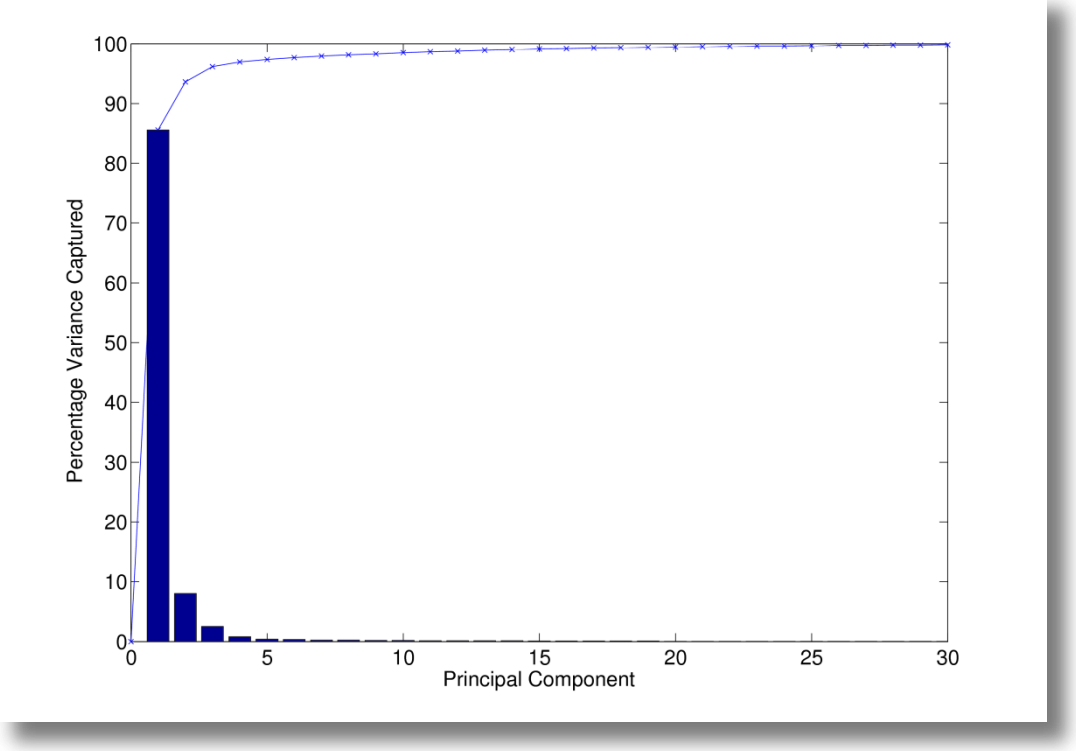


Figure 3.18 - The percentage of variance captured by the first 30 principal components (bars) and the cumulative sum (line) for the above PCA analysis (Figure 3.17)

Removing the m/z above 109 does not appear to have affected the discrimination seen with PCA (Figure 3.10 compared to Figure 3.17) which indicates that m/z 111 was probably not the contributing factor in the slight discrimination in the above PCA. In the previous chapter, no scaling methods were used on the data when analysing using PCA it was only introduced when the PLSDA analysis was performed. This was mainly due to the clear discrimination between the groups without scaling. As this is not the case there is now the possibility to investigate the effects of different pre-treatment techniques and draw conclusions about the effectiveness of each one.

Mean centring the data has very little effect on the PCA (see Figure 3.19) when compared with the PCA on the raw data, Figure 3.17. The data is clearly distributed around the graph in a similar fashion to the PCA on the raw data. The largest effect is the removal of the bias on the first component. The data is now centred on zero. This change was expected. Mean centring removes any large bias in either the positive or

negative direction, Eq. 1.2. The first component no longer contains this bias and hence the data becomes centred on zero. Examining the loadings (not shown) this similarity between the raw data and the mean centred data is confirmed.

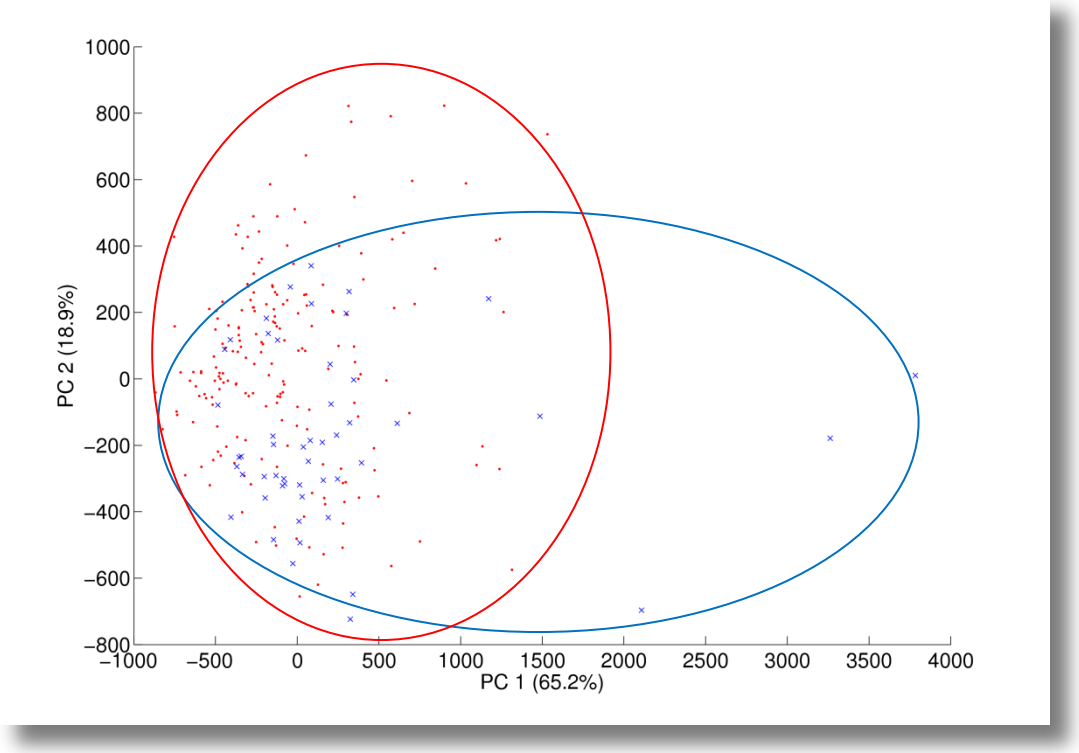


Figure 3.19 - PCA scores plot of the serum samples collected from wild badgers recorded by the mk. 2 SIFT-MS machine using the H_3O^+ precursor. The m/z associated with ethanol, acetone and the precursors have been removed as have m/z above 109. The data was pre-treated using a mean centring algorithm (Eq. 1.2).

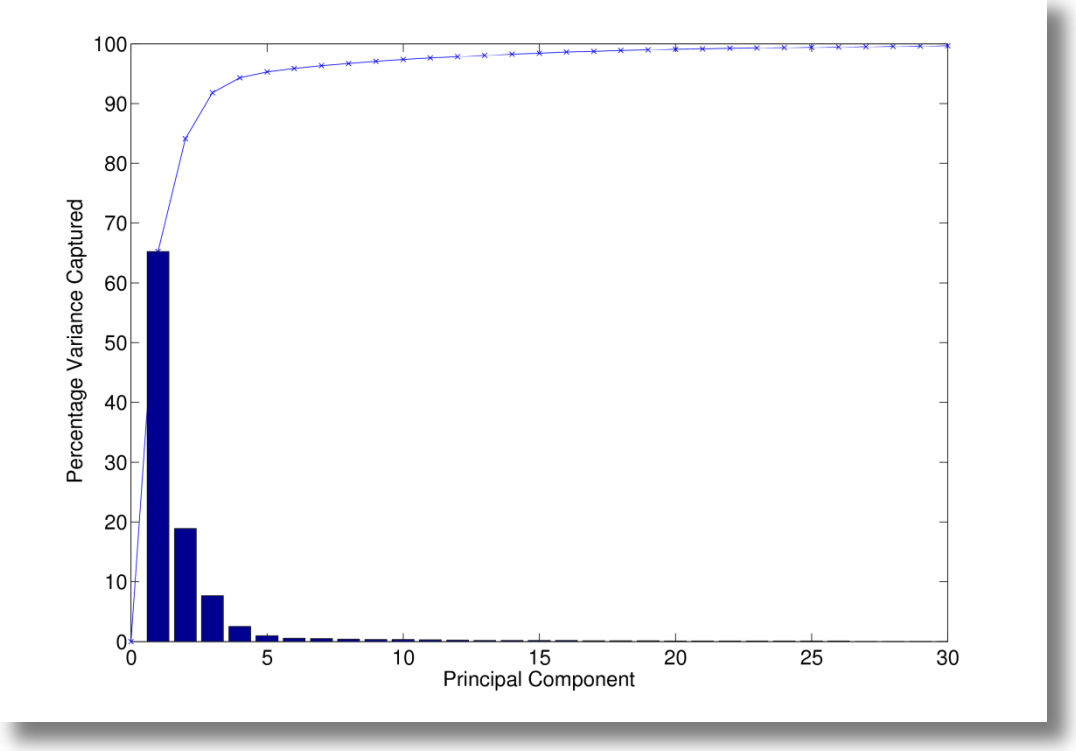


Figure 3.20 - The percentage of the variance captured by the first 30 principal components (bars) and the cumulative sum (lines) for the above PCA analysis (Figure 3.19)

The auto-scaling algorithm (Eq. 1.3) appears to show slightly better discrimination, Figure 3.21. The positive samples are mostly in the positive direction of the second principal component. There has been a significant change in the arrangement of the samples compared with both the mean centred (Figure 3.19) and raw PCA (Figure 3.17). This shows that the auto-scaling method changes the data considerably. In fact, it removes most of the quantitative information and instead the PCA is built using the patterns of volatiles available instead of the magnitudes of each. Investigating this further appears to confirm this. The variance captured by the first two components is very small (~11%). Indeed, the variance capture curve (Figure 3.22) shows that even using 30 latent variables less than 75% of the total variance is captured.

One possible cause for this is that every non zero m/z ratio now has almost equal weightings. It is no longer the case that if a m/z ratio has a large count it is likely to be more important than one with only a trace count level.

Secondly, there is a problem when inspecting the loadings plot. With the above PCA score plot discrimination appears to be best on the second component axis and the problem is highlighted in Figure 3.23 below.

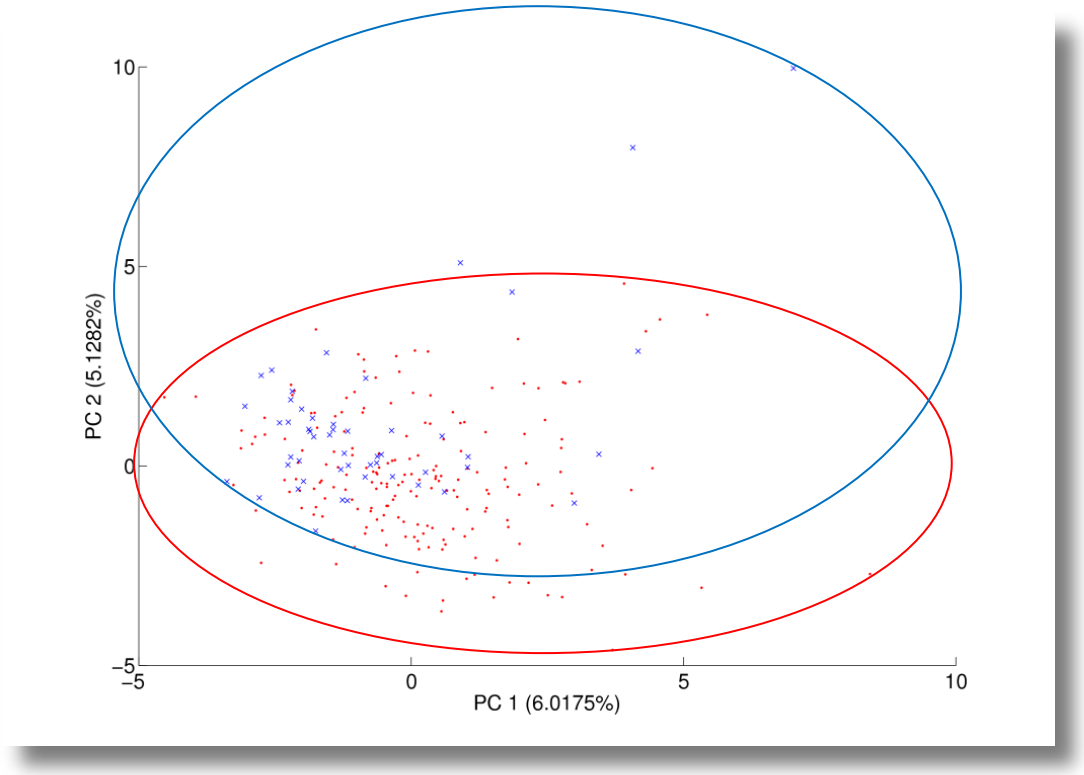


Figure 3.21 - PCA scores plot of the spectra of serum samples collected from wild badgers recorded using the mk.2 SIFT-MS machine and the H_3O^+ precursor ion. The m/z ratios associated with ethanol, acetone and the precursors have been removed as have m/z ratios above 109. The data was pre-treated using an auto-scaling algorithm (Eq. 1.3).

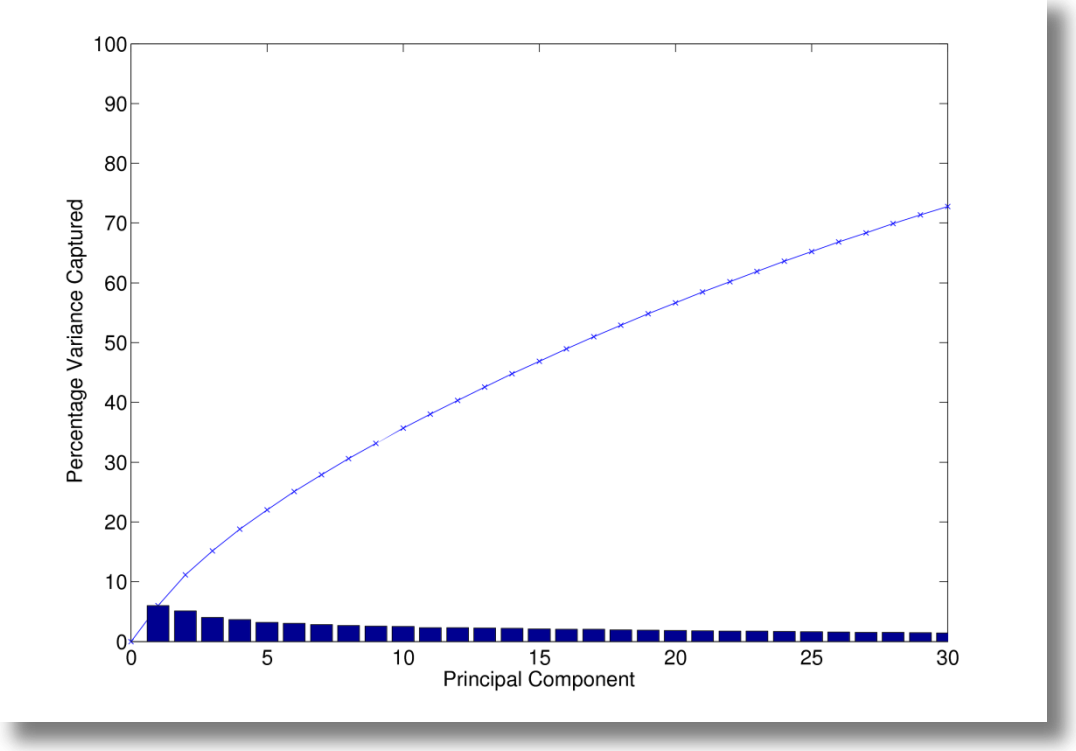


Figure 3.22 - The percentage of variance captured by the first 30 principal components (bars) and the cumulative sum (line) for the above PCA analysis (Figure 3.21)

The loadings (Figure 3.23) show that many of the m/z peaks are contributing to the covariance captured by the second component. The loadings for many volatiles that have very low levels and have a sparse coverage are included in the first component (not shown) and are the likely cause for the poor variance capture. Many of these volatiles are then present again on the second component. With all the noise from rare volatiles contributing to every PC it is probably only random chance that m/z 18 has the biggest impact on PC2, it could have been pushed into a lower PC. Clearly, the magnitude of the m/z peaks is important.

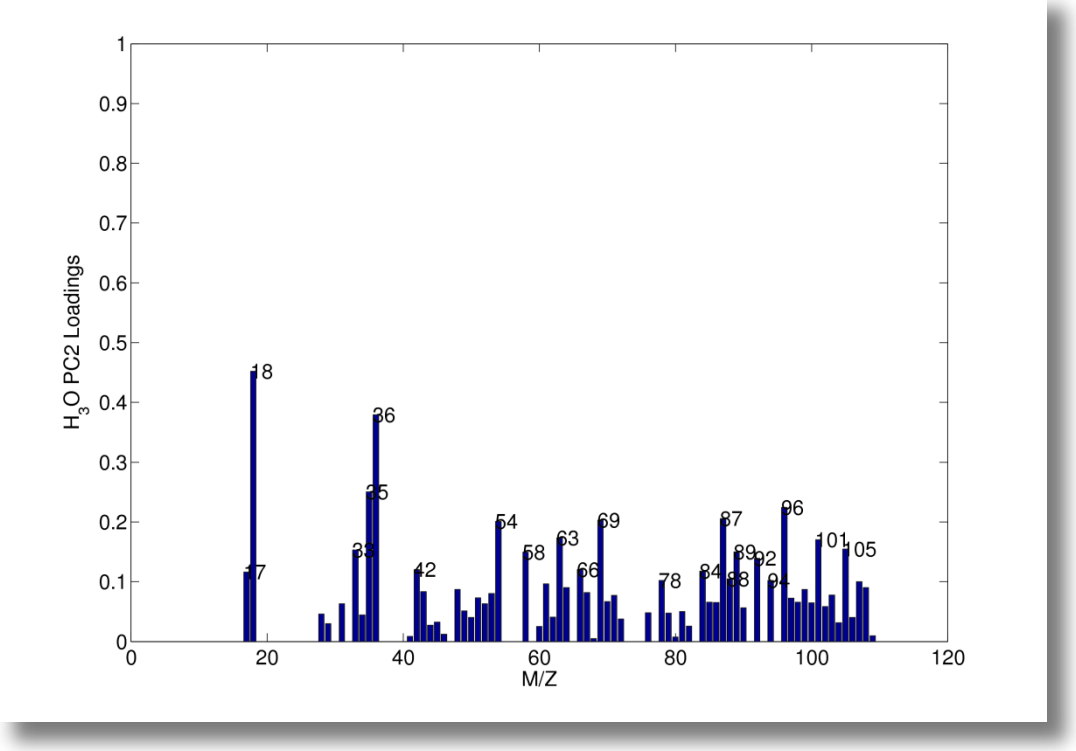


Figure 3.23 - The loadings of PC2 for the above PCA scores plot of the spectra of serum samples collected from wild badgers recorded using the mk.2 SIFT-MS machine and the H_3O^+ precursor ion where signs of discrimination between the class could be seen. The m/z ratios associated with ethanol, acetone and the precursors have been removed as have m/z ratios above 109. The data was pre-treated using an auto-scaling algorithm.

The Log_{10} and mean centre pre-treatment (Eq. 1.5) gives a similar separation between the clusters (Figure 3.24) with most of the discrimination being seen on one axis only. Most of the positive samples are limited to the negative side of PC1. However, looking at the loadings many m/z are impacting every PC in a similar manner to that seen for the auto-scale pre-treatment. This was not expected, as the data is most likely to be log normal.

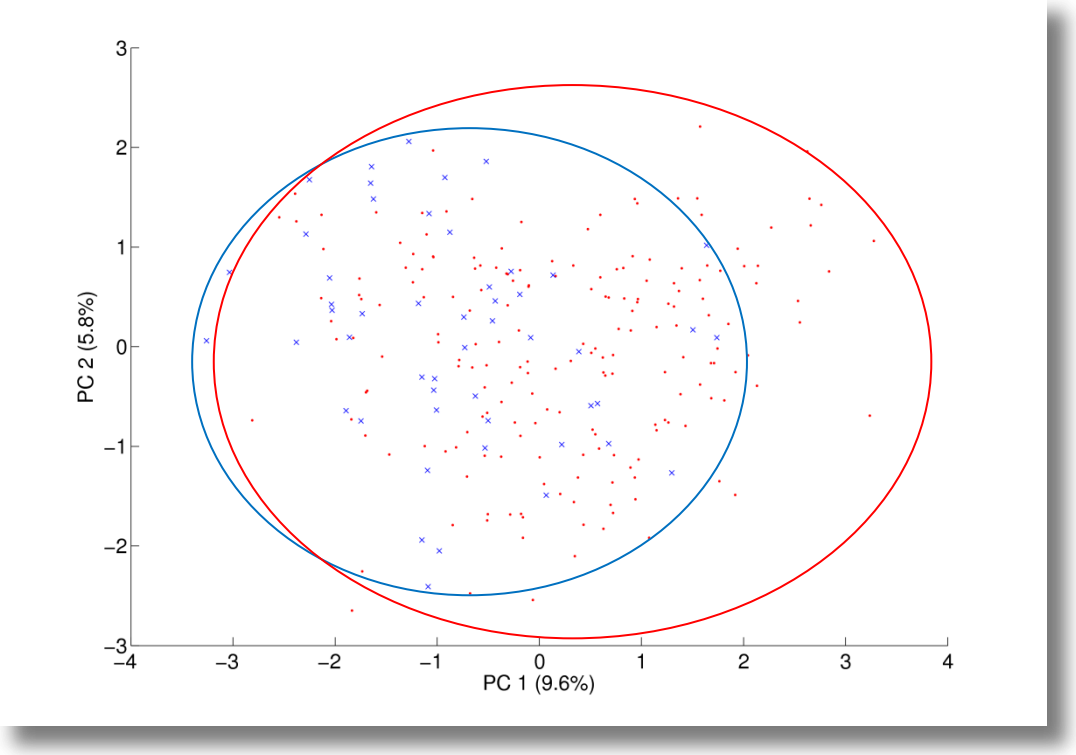


Figure 3.24 - PCA scores plot of the badger serum samples are recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. m/z peaks associated with acetone, ethanol and the precursors as well as m/z greater than 109 have been removed. The data was pre-treated using Log_{10} and then mean centred (Eq. 1.5).

The problem of the lack of variance being captured is almost as bad as the auto-scaled data. Figure 3.25 shows that even with 30 principal components only 85% of the total variance is captured. This is not as bad as the auto-scaled data where only 75% was captured but suggests that this pre-treatment is not suitable for SIFT-MS data.

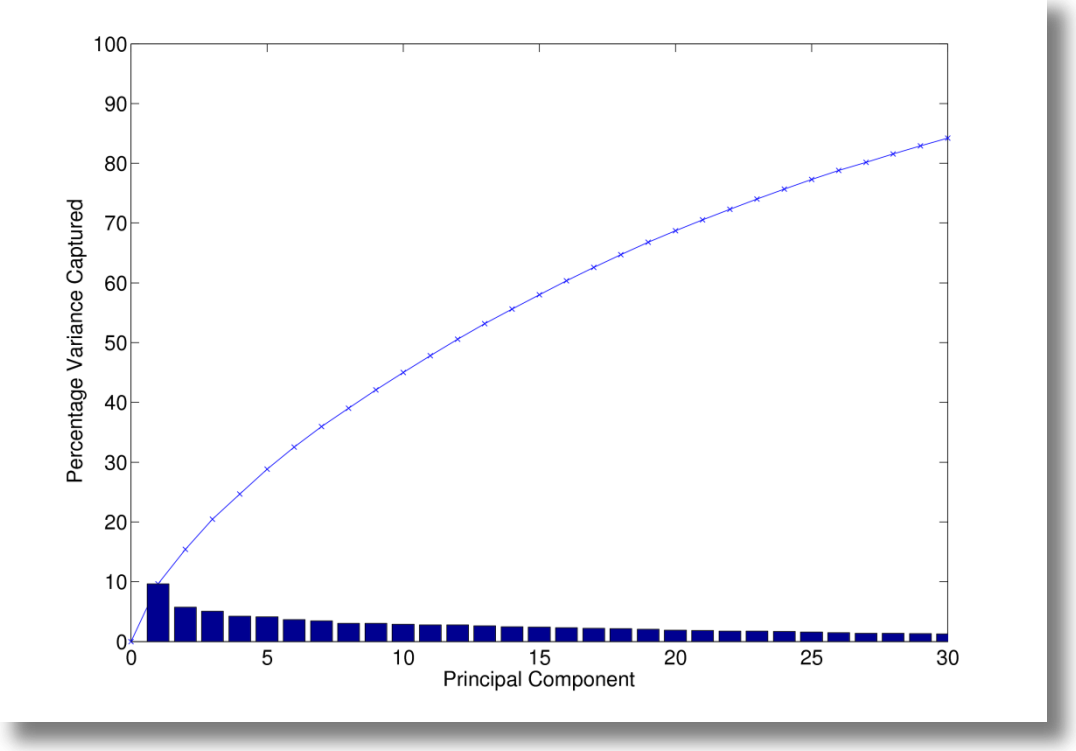


Figure 3.25 - The percentage of variance captured by the first 30 principal components (bars) and the cumulative sum (line) for the above PCA analysis (Figure 3.24)

With PCA and SIFT-MS data it appears that the best pre-treatment for the SIFT-MS is to leave as much information as untouched as possible. Mean centring the data gives the best visual discrimination and also allows for the breakdown and investigation of the causes for the clustering or structures seen in the scores plot.

3.3.3. PLSDA analysis

The PCA analysis did not perform as well as the analysis on the mixtures in Chapter 2 as there was considerable overlap between the clusters of samples. This might be due to the complex nature of the samples being investigated. They are made up of a much larger number of VOCs than the mixtures in Chapter 2 and they also have much subtler differences in the concentration differences between sample classes. There are also differences caused by a wide range of other factors that will have little correlation with tuberculosis status. One possibility could be that instead of a single

marker we could be looking for a subtle change in a range of volatiles that on their own are not large enough to overcome the large natural variance seen due to the biological nature of the samples. Previously, PLSDA gave perfect discrimination between the samples (100% leave one out accuracy).

As before, the three precursors were used to build three classification models and a LOO optimisation was performed. The m/z values associated with the precursors and their isotopologues have been removed. The maximum m/z ratio has been limited to 109 but no other removals (acetone and ethanol for example) as suggested in the above PCA analysis have been removed.

PLSDA returned 67% accuracy for the LOO optimisation for the H_3O^+ precursor spectra (shown below), 62% for the NO^+ precursor spectra and 64% for the O_2^+ precursor spectra (not shown). As the results for the H_3O^+ spectra are better than the others, this confirms that it is indeed the most useful precursor to be analysing samples with.

The PLSDA plot (Figure 3.26) show that the model required 4 LVs to reach the highest classification performance. With fewer LVs, the model performs badly which suggests that there is a large amount of non-discriminating information obscuring the classification. Looking at the curves shows a high true positive (78%) which is promising.

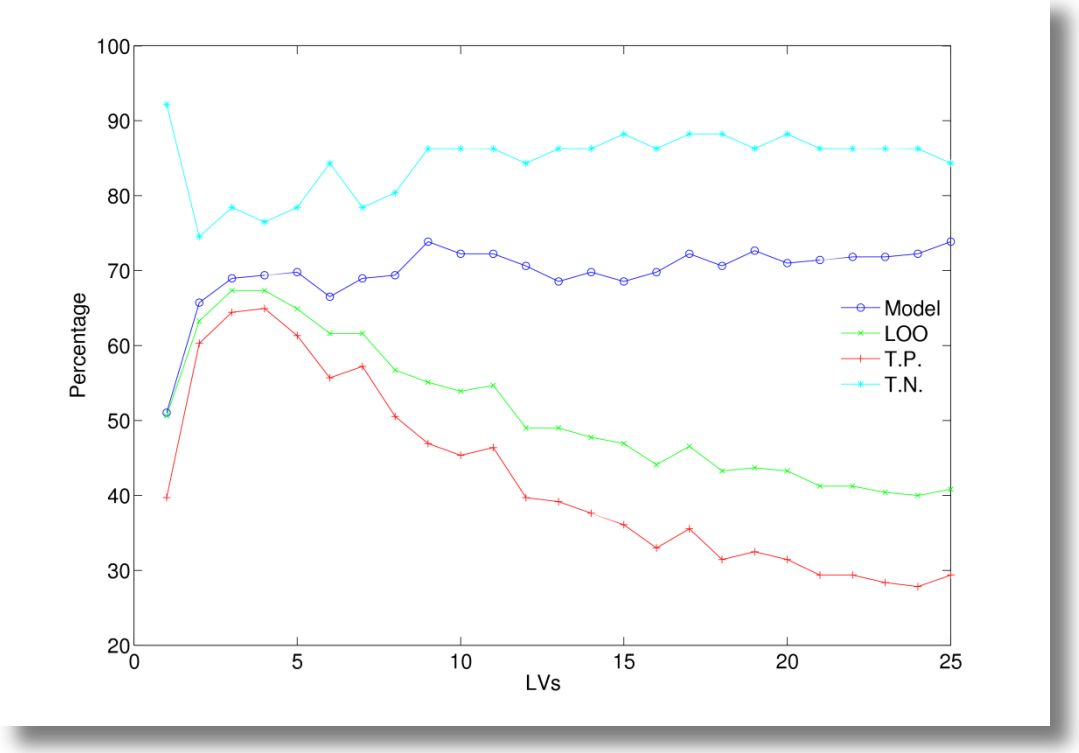


Figure 3.26 - The PLSDA plot for the model accuracy, LOO accuracy, true positives and true negatives associated with the LOO accuracies for the PLSDA model built using the H_3O^+ precursor spectra recorded using the mk.2 SIFT-MS machine from the headspace analysis of the serum samples from the wild badgers using tuberculosis culture status as the classification vector. The spectra data was mean centred and m/z ratios associated with the precursors were zeroed.

Investigating the loadings (Figure 3.27) shows this information could be linked to ammonia levels (m/z 18, 36 and 54). The loadings from the PLSDA model do not necessarily link to the LOO results as different samples are used to build the model every time but it is probably safe to assume that they are at least similar.

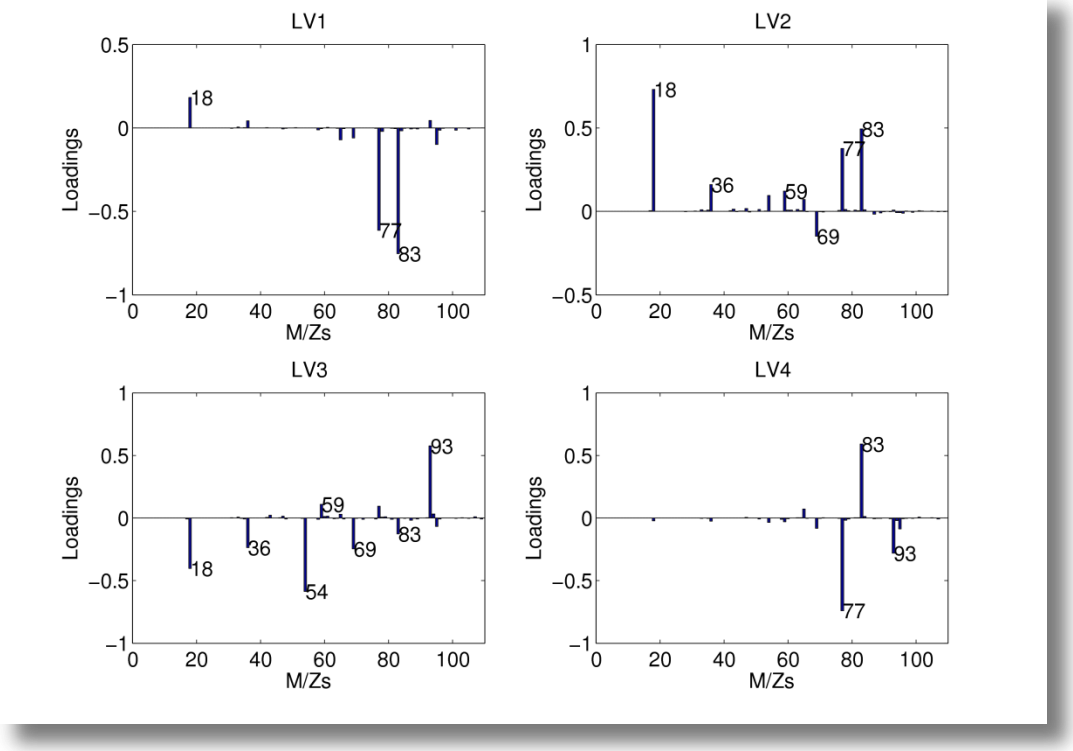


Figure 3.27 - The loadings plots of the first four latent variables of the PLSDA model built using the samples recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. The data was mean centred and m/z ratios associated with the precursors were removed

Combining all three precursor spectra into a single matrix and repeating the analysis improves the results fractionally giving 68% accuracy (Figure 3.28). This is mostly due to a large increase in the number of true positives identified using all three precursors together. The results also appear 1 LV further into the complexity.

Linking back to the PCA, the unhelpful information early in the complexity level of the PLSDA model might be linked to ethanol and acetone which were discussed as not being linked to tuberculosis by the clinicians. Removing them by zeroing the m/z peaks (ethanol m/z 83 and acetone m/z 77) gives the result shown in Figure 3.29.

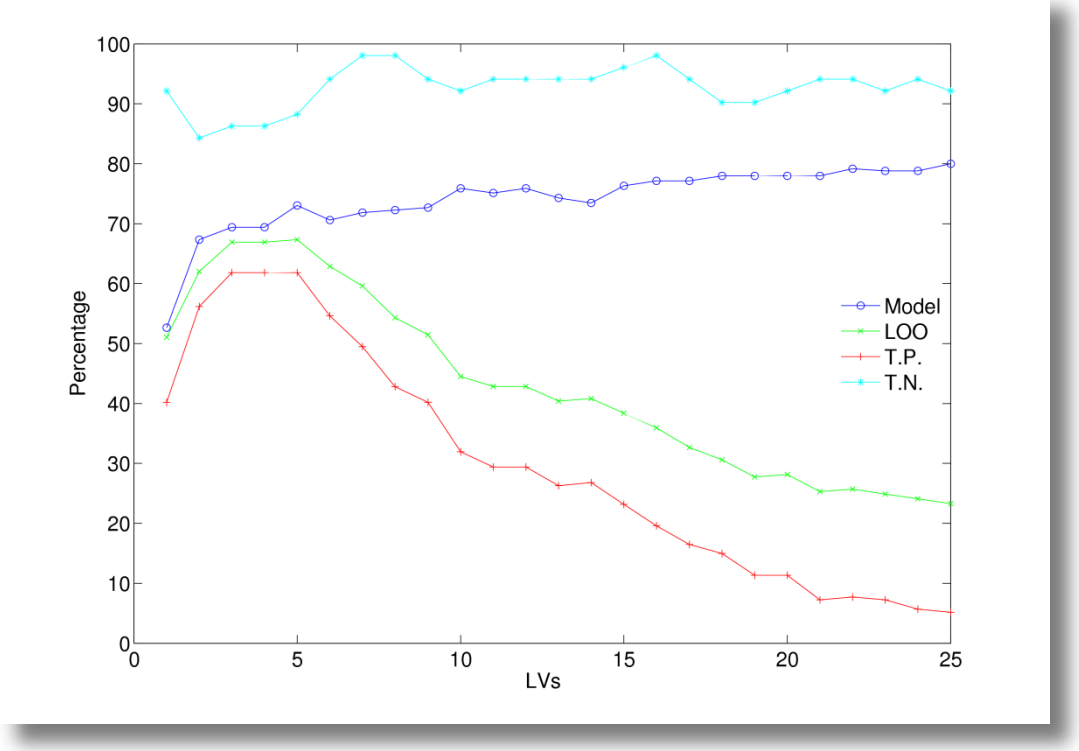


Figure 3.28 - The PLSDA curves for the model accuracy, LOO accuracy, true positives and true negatives associated with the LOO accuracies for the PLSDA model built using all three precursor spectra recorded using the mk.2 SIFT-MS machine from the headspace analysis of the serum samples from the wild badgers using tuberculosis culture status as the classification vector. The spectra data was mean centred and m/z ratios associated with the precursors were zeroed.

The accuracy of the model is still 67% however this now appears with just a single LV. This appears to agree with the earlier findings that ethanol and acetone are not linked to tuberculosis. The number of true positives has dropped from 78% to 72%. This is however balanced by an increase in the true negatives from 64% to 66%. There are however more negative samples than positives so the overall accuracy is always closer to that of the true negatives.

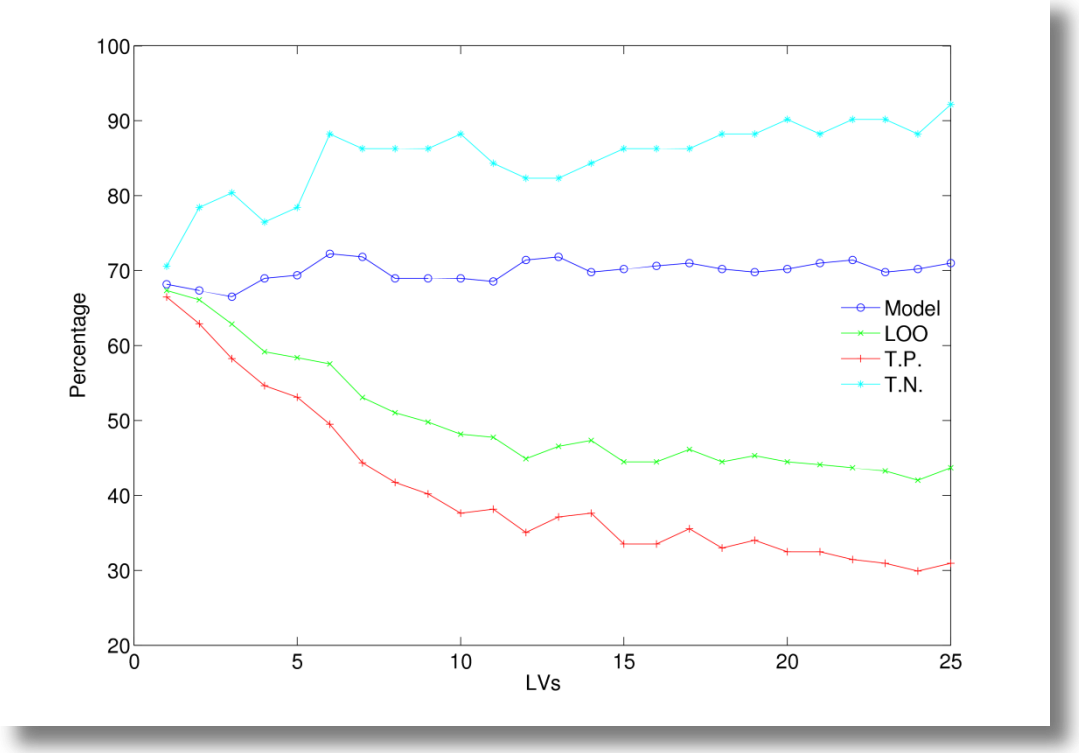


Figure 3.29 - The PLSDA plots for the model accuracy, LOO accuracy, true positives and true negatives associated with the LOO accuracies for the PLSDA model built using the H_3O^+ precursor spectra recorded using the mk.2 SIFT-MS machine from the headspace analysis of the serum samples from the wild badgers using tuberculosis culture status as the classification vector. The spectra data was mean centred and m/z ratios associated ethanol, acetone and the precursors were zeroed.

Repeating the analysis with all three precursors and the m/z linked to acetone and ethanol removed from the H_3O^+ spectra gives a similar response in that the accuracies stay approximately the same the true positive and true negative percentages migrate towards that of the LOO accuracy. However, the effects are not as complete as with the H_3O^+ data alone. This is likely to be caused by the inclusion of the acetone and ethanol m/z on the NO^+ and O_2^+ precursor channels where removals are more difficult as they are less well understood.

From the PLSDA analysis it is clear that the modelling techniques can cope (to a degree) with m/z values with high variance that are not linked to the classification

status of the samples. Removing them does decrease the complexity of the model but it does not affect the overall accuracy of the model in this situation.

Previously, it was found that mean centring the data and combination of Log_{10} and mean centring were the two best data pre-treatment methods when PLSDA results are required. PLSDA was performed on the H_3O^+ precursor spectra with and without the m/z associated with ethanol, acetone and toluene and in both cases the results were worse (64%) than the equivalent PLSDA models where mean centring only was used. However, when the three precursors were combined the results, although not quite as good (67%), were comparable to the alternative of only doing mean centring.

Unexpectedly, the Log_{10} plus mean centring pre-treatment causes the models produced to become overtrained with fewer latent variables than mean centring pre-treatment alone. This suggests that even though the range of intensities of volatiles is generally expected to be log normal, the differences caused by disease might not be.

3.3.4. Bootstrapped PLSDA analysis of data

Before moving on to investigate the predictive capability of the PLSDA model, the number of samples required to accurately build a model needs to be determined. This can be achieved using the bootstrapping method used on the test data set in chapter 2. This time it should give more interesting information as the model does not achieve 100% accuracy.

The results in Figure 3.30 show that a large number of samples are required to accurately build the model. Above 200 samples the model appears to be stable with increasing number of samples having little effect on the accuracy of the LOO validation. Below this there is a region of increasing accuracy with increasing numbers of samples used, as would be expected. Below this (approximately 125 samples) the accuracy of the model appears to stabilise again however the error associated with the model balloons out exponentially which suggests that the accuracy of the model is highly affected by the random sample selection.

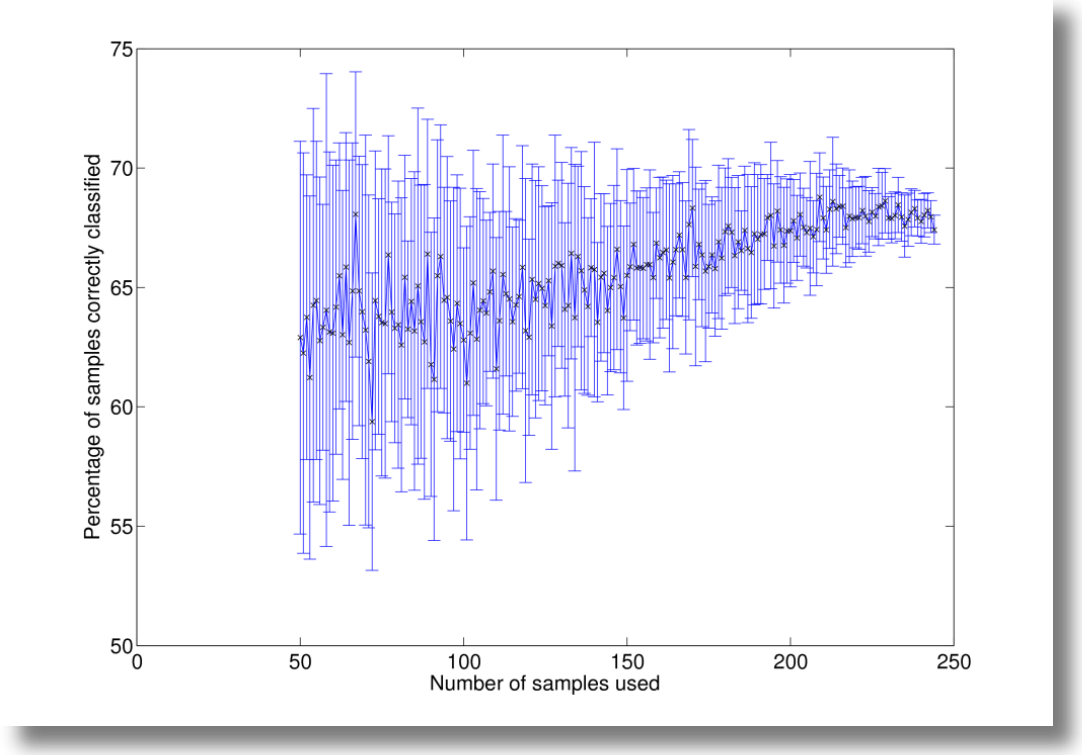


Figure 3.30 - PLSDA LOO accuracies for models built using a range of sample subset sizes using the spectra recorded by the H_3O^+ precursor channel of the mk.2 SIFT machine from serum samples taken from wild badgers. Tuberculosis culture status was used as the classification vector. The data was mean centred and m/z ratios associated with the precursors were removed.

A similar graph was produced for the H_3O^+ dataset with acetone and ethanol removed. The results were strikingly similar; however, a few subtle differences were apparent. The accuracy of the higher sample number models was slightly lower. This suggests that by including all three precursor spectra there is the possibility of a better prediction. However, this was balanced by an earlier decrease in the errors associated suggesting that the number of samples required to build a model (albeit not as accurate) was lower.

Using Figure 3.30 it is clear that the best range to build a model for prediction would be between 80-90% (196-220) of the samples. Above this range there was no increase

in the accuracy and the few samples left for testing will require more bootstraps to be performed and below this the model will not be as accurate.

The best results are shown (Figure 3.31) to be 67% ± 6% using a model created using 90% of the samples and tested using 10%. This is similar to the results from the original LOO analysis and the bootstrapped sample number investigations. Finally, the modelling was reformed but with a randomised class (Figure 3.32). This was to check that the accuracies seen above were purely due to chance. The randomised class vector was made by randomly ordering the original class vector. In this method the same numbers of positive and negative samples are retained but they will no longer be associated with tuberculosis. When the randomised class vector is used in the supervised method of PLSDA the discrimination will be limited to simple chance.

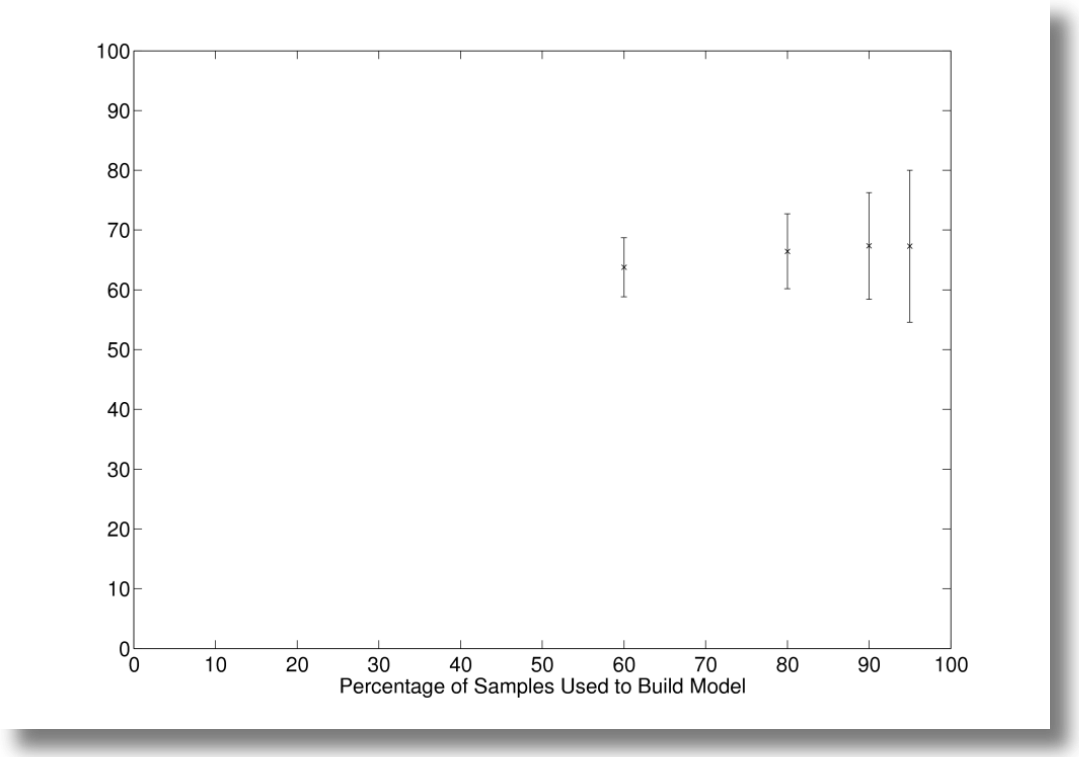


Figure 3.31 - A graph showing the PLSDA accuracies of models built and tested using four ratios of sample (60:40, 80:20, 90:10 and 95:5) from serum samples from wild badgers recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine and using the tuberculosis culture status as the classification vector. 10000 bootstraps were performed and the error bars on the graph show one standard deviation.

The accuracy of the model is approximately 46% for all the model/test ratios tested with the randomised class list order (Figure 3.32). This means that the accuracy seen in the above models is unlikely to be down to chance alone.

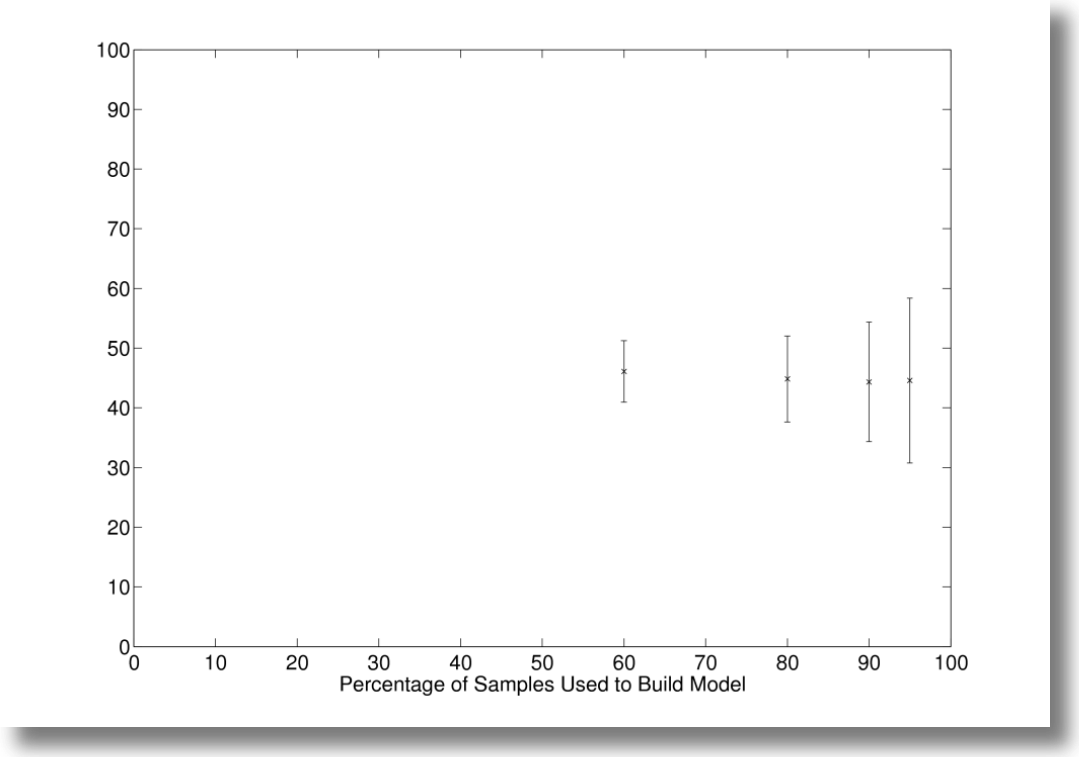


Figure 3.32 - A graph showing the PLSDA accuracies of models built and tested using four ratios of sample (60:40, 80:20, 90:10 and 95:5) from serum samples from wild badgers recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine and using a random permutation of the tuberculosis culture status as the classification vector. 10000 bootstraps were performed and the error bars on the graph show one standard deviation.

3.4. Conclusions

In this chapter, it has been shown that multivariate analysis methods can be used on SIFT-MS data of the headspace analysis of serum samples from wild badgers in the search for diagnostic tests for *Mycobacterium bovis* discrimination. The result of an accuracy of prediction of 67%±6% (Figure 3.31) is better than random chance (Figure

3.32) but offers plenty of room for improvement. One of the possible reasons for not getting a higher accuracy is the biological challenges associated with detecting tuberculosis in badgers. Alternative methods such as the gamma interferon test are only slightly better (Dalley et al., 2008) and hence the gold standard method of culture testing biopsies post mortem is still the only reliable method.

A second possible reason for the problems is the poor quality of samples. Serum samples may be easy to take and work with but they may not be the optimal samples (Perbellini et al., 2003). Serum samples were acceptable samples as they can be sterilised so that no *Mycobacterium bovis* bacteria survive. This allows for the measurement of samples outside of CAT3 laboratory conditions which was necessary for analysis with the SIFT-MS machine. Sputum samples are more likely to contain *Mycobacterium bovis* bacteria but it was not possible to collect such samples from badgers or cattle. Breath samples, although difficult to obtain, are also likely to contain the bacteria. Unfortunately, breath samples proved too difficult to sterilise and hence failed health and safety protocols.

Original work at detecting tuberculosis using volatiles avoided these problems by spiking uninfected serum samples with large doses of immobilised tuberculosis bacterium (Pavlou et al., 2004; Fend et al., 2005). Using this method, accurate classification can be achieved and the samples can be use safely in normal laboratory conditions. As a proof that detecting tuberculosis bacteria using volatiles is possible this was a good first step, however, it does not easily translate into real world applications. The most challenging part of the current process is the collection and growth of tuberculosis bacteria (Bourne, 2007; Cho and Brennan, 2007; Chaisson and Harrington, 2009). With the method proposed here, this challenging step is avoided by instead looking for markers left in the serum by the bacteria instead of the bacteria themselves.

A better understanding of the pre-processing has also been achieved. In Chapter 2 the simplicity of data created the problem that both mean centring (Eq. 1.2) and Log_{10} plus mean centring (Eq. 1.5) pre-treatments produced models of 100% accuracy.

Auto-scaling did not only perform worse than the other two it caused problems sorting through the loadings meaning that an investigation into the causes of the discrimination was not possible (e.g. identification of biomarkers). This is still the case with the badger serum sample but mean centring alone produced higher accuracies than the Log_{10} plus mean centring when working with data that had unhelpful m/z ratio removed with the H_3O^+ spectra alone.

With electronic nose, in general, auto-scaling methods produce the best results (Jurs et al., 2000; Gutierrez-Osuna, 2002; Scott et al., 2006). It is generally assumed that due to the difficulties with the equipment, pattern recognition is the best approach for discrimination (Pardo and Sberveglieri, 2002) as any quantitative information is likely linked to the machine drift. It is important to note, however, that the scaling method is not always the same for different electronic noses or different sample types and so the result here that mean centring alone might only be true for the combination of SIFT-MS and headspace analysis above bovine serum samples.

4. Application of method to the determination of disease status from serum samples taken from Cattle

4.1. Introduction

So far, multivariate techniques and SIFT-MS have been proven as a concept and also shown promising results when used as a disease detection method with serum samples taken from badgers during a bovine tuberculosis field study. This chapter expands the scope of the investigation to include cattle serum samples pertaining to four diseases, collected during both field studies and clinical trials. Also, the use of multivariate techniques as a method for identifying possible m/z biomarkers is evaluated.

This is the largest of the three biological dataset groups tackled in this thesis. The samples came from various sources and various diseases. The largest portion of the samples came from the Freidrich-Leoffler-Institut (FLI), Germany and contained both clinical trial samples (*Mannheimia haemolytica* and *Mycoplasma bovis*) and field samples (*Mycobacterium paratuberculosis*). The other samples came from the UK Veterinary Laboratories Agency (VLA) and cover historical field samples (Brucella) and new clinical trial bovine tuberculosis (*Mycobacterium bovis*) samples.

As a whole, it would be a very large dataset but it is more appropriate to separate them into the separate diseases, each of which pose a different challenge for analysis. It was initially hoped that the dataset could be combined and then used to show that the SIFT-MS plus multivariate techniques methodology could be used as a single detection method for a range of diseases, but the disparate sources of data precluded this.

4.2. Data Collection and Measurement

All the data, except that of the bovine TB samples, was recorded using the mk.2 SIFT-MS over a period of approximately nine months. The bovine TB data was

recorded after this time using the newer mk.3 SIFT-MS using the same 5 minute switching scan protocol. The standard data collection protocol was used as discussed in Chapters 2 and 3 and also (Knobloch, 2010). 1.6ml of serum thawed and then injected into 700ml bags. As before the headspace created was analysed with electronic nose (2/3rd volume) and then SIFT-MS (1/3rd volume).

The frozen serum samples were shipped to Cranfield and stored (-80°C) until the data collection could be performed in a single batch (in a randomised and blinded order). This process was chosen as it allowed the randomisation to be performed at Cranfield. A large database of the sample types was created and then each sample assigned a random number. The dataset was then sorted using these random numbers and analysed using the SIFT-MS machine using the sorted order.

Unfortunately, due to an error, the bovine TB samples were delivered months after and hence they were analysed using the newer SIFT-MS machine. As the multivariate analysis is also being performed separately, this is not an issue.

Table 4.1 summarises the samples.

Table 4.1 - Overview of the cattle serum datasets including the analysis machine, trial type and basic info about the number of samples

Dataset	**Type**	**SIFT-MS**	**Summary**
Mannheimia haemolytica	Trial	Mk. 2	232 samples over 13 time points
Mycobacterium paratuberculosis	Field	Mk. 2	43 Positives and 23 Negatives
Brucella	Field	Mk. 2	23 Positives and 25 Negatives
Mycoplasma bovis	Trial	Mk. 2	325 samples over 16 time points
Bovine TB	Trial	Mk. 3	60 samples over 4 time points

4.2.1. *Mannheimia haemolytica* data

This data set contains 232 samples. They came from 20 animals that were injected with very high doses of the bacteria. All the animals showed clinical signs 2 days after the pathogen had been introduced and there were fatalities before the end of the trial (Figure 4.1). There trial was run for a long time before the pathogen was introduced (28 days) but due to the rapid progression of the disease only lasted 5 days after pathogen introduction. Also, feedback from the clinicians involved suggested that the animals which survived to day 5 were all beginning to show signs of recovery.

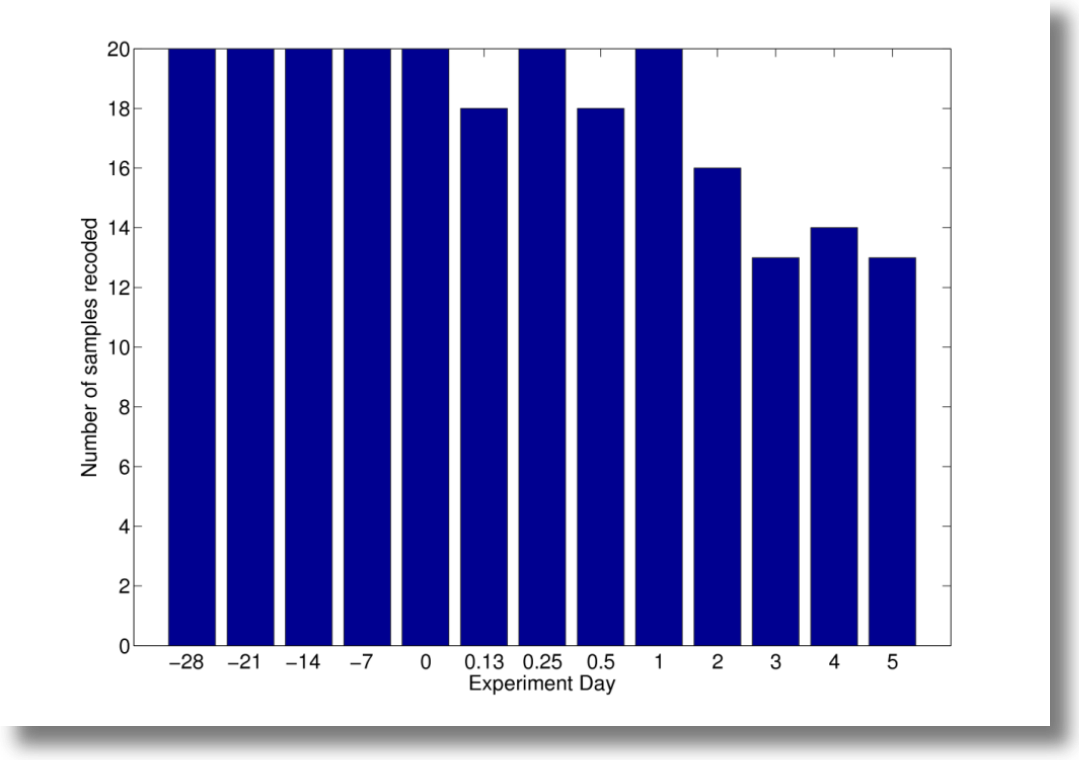


Figure 4.1 - The number of samples recorded for each time point for the *Mannheimia haemolytica* dataset showing the decrease in the numbers towards the end of the trial caused by fatalities

After a discussion with the clinicians, they suggested that the best way to analyse the data would be to look at the middle time points after the introduction of the pathogen (i.e. 1, 2 and 3 days on Figure 4.1). These time points should give the most difference

from the norm as the infection has had long enough to become established but not long enough for the animals immune system to fight back and the animals begin to recover. They suggested that by time point five the animals were either well on the way to recovery or had died.

Creating groups for this trial is important as there are no clear positive/negative boundaries. The animals were injected with the pathogen at time point zero, however, the early samples are unlikely to show as the animal did not show physical symptoms in the first day. The clinicians also noticed that by day 5 the remaining animals were showing signs that had recovered from the infection. The peak of the infection was estimated to be between 1 day and 3 days after introduction of the pathogen. The reasons behind the grouping are mostly from these clinical recommendations from the project collaborators, hence the following will be used (Table 4.2).

Table 4.2 - The grouping of the *Mannheimia haemolytica* samples for analysis with multivariate tools

Group	Time points
Uninfected	All time points before the pathogen was introduced (100 samples)
Post Introduction Group 1	0.13, 0.25 and 0.5 days after introduction (56 samples)
Post Introduction Group 2	1, 2 and 3 days after introduction (49 samples)
Post Introduction Group 3	4 and 5 days after introduction (27 samples)

As mentioned above, post-introduction group 2 is most likely to be useful as the positive class for supervised classification methods. During the unsupervised methods and the univariate approach the ordering will go one step further and order the samples by time point (although colour highlighting will still be based on these groups). In this way the clinicians will be able to compare the results with the expected peak disease levels.

4.2.2. *Mycobacterium paratuberculosis* data

This data set contained 66 samples. The samples came from three farms, two of which had infections of *Mycobacterium paratuberculosis* in the herds - the third did not. Initially, only 46 samples were sent to Cranfield. They comprised 23 positive samples and 23 negative samples. Subsequently another 20 positive samples were sent in a separate batch and hence they were not part of the randomisation process which means that some instrumental drift may be detectable and they might need to be removed from the analysis. Urine samples were also collected from these animals (although less successfully) and their analysis is discussed in Chapter 5.

4.2.3. Brucella data

This dataset came from samples sent by the VLA. It contains serum samples from a historical database collected to improve brucella detection. The samples come from all over Europe and therefore are likely to exhibit considerable variance in sample composition. As the number of cases of the disease is also small, they have also been collected over a considerable length of time. The current tests for Brucella is very accurate (with both exceptionally high numbers of true positives and true negative). The test is very heavily used and the main problem is the number of false positives (England et al., 2004) especially since the disease has been eradicated from the UK (Cutler et al., 2005). The UK has been officially brucella free since 1993 but due to EU legislation approximatly 50% of UK herds are routinely tested (England et al., 2004) which costs the government money.

4.2.4. *Mycoplasma bovis* data

This dataset contained 325 samples. They came from 30 cows, 24 of which had the pathogen introduced and 6 controls. Blood samples (and hence serum) were collected at 16 time points (t-11 days, t-5, t+0, t+0.04, t+0.25, t+1, t+2, t+3, t+5, t+7, t+10, t+14, t+17, t+21, t+28 and t+35) where t was the infection day. The samples from t+0 was taken just before introduction of the *Mycoplasma bovis* pathogen.

The animals in the trial showed no physical signs of the disease. Normally, no clinical symptoms become visible for up to 3 months after infection (Nicholas and Ayling, 2003) however, the trial was run to evaluate a new diagnostic test that promised much earlier diagnosis and to evaluate whether *Mycoplasma bovis* trials are possible.

As this was one of the first animal trials run at the new FLI animal centre not all samples were successfully taken. Also, when it became apparent that the animals were not showing any signs of infection and the evaluation of the new diagnostic test was compromised the trial was wound down. Figure 4.2 shows how the number of samples acquired per day decreased as the trial progressed.

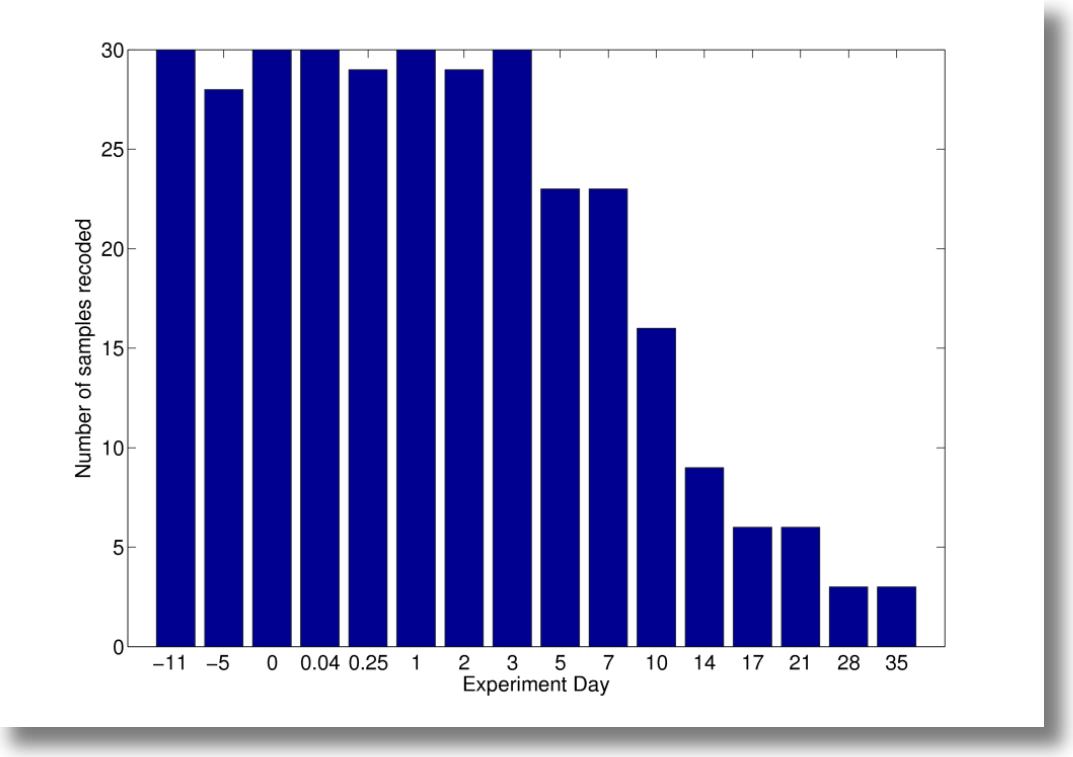


Figure 4.2 - The number of samples recorded at each sampling time point, showing how the trial was wound down after the clinicians found the introduction of the *Mycoplasma bovis* pathogen had failed to cause visible infection responses.

If a two class classification model is to be built, the problem in this case is choosing what classifies as a positive and what as a negative. We have control animals, however, once the introduction of the pathogen proved ineffective at infecting the animals these animals in the control group were the first to be removed from the trial

and hence inclusion will only confuse the problem, so they were removed from further analysis. Finally, as there are too few samples in each time point to stand on their own a grouping method is required. If 11 days before, 5 days before and just before the pathogen was introduced are grouped together as the negative class this group will contain approximately 75 samples. Attempting to keep the numbers of animals in each group even the grouping method in Table 4.3 was adopted.

Table 4.3 - A table showing the groupings of the *Mycoplasma bovis* data that formed the classification vector to be used in the display of the unsupervised multivariate analysis

Group	Experiment Days
Uninfected	11 days before, 5 days before and just before pathogen introduced
Post Introduction Group 1	0.04 days, 0.25 days and 1 day after pathogen introduced
Post Introduction Group 2	2 days, 3 days, 5 days and 7 days after pathogen introduced
Post Introduction Group 3	10 days or more after pathogen introduced

The clinicians suggested that it was unlikely that classification would be possible but the best chance would be to use samples from as late in the trial as possible (e.g. Post Introduction Group 3 in Table 4.3), but this will need to be explored further in the first stage of analysis. Even though it is likely that nothing will be found that can be linked to the disease it will be useful to investigate what would happen in the case of SIFT-MS and multivariate techniques not being able to detect or discriminate between disease states.

4.2.5. *Mycobacterium bovis* (bovine TB) data

The samples for this dataset came later than the others in this Chapter and were analysed separately using the newer mk.3 SIFT-MS machine. The data came from two trial houses, one from the VLA and one from a partner laboratory (Institute of Animal Health (IAH), Compton, UK). Both trials were run in parallel and were designed to investigate the effect of different immunisations against TB. Samples

were taken from the animals and they were then immunised and 2 weeks later a second round of samples were taken before very high doses of pathogen were introduced into the animals, making infection highly likely. Two more rounds of samples were taken 1 month and 2 months after the pathogen were introduced.

The immunisations under investigation were BGC, Ada85A and VPM. Of these only Ada85A and VPM showed any signs of a reduction in the infection status but the clinicians claim that none of the immunisations were truly effective and all animals showed signs of infection (granulomas confirmed as being caused by TB via culture test) whether slight or severe during the visual inspection of lungs post mortem. As TB does not have a cure and the body's only natural defence is to create these granulomas all animals will have TB bacteria present. Also the doses of the pathogen were high enough that visible signs of TB infection were extremely likely.

4.3. Results

As this chapter contains several datasets, instead of splitting this section into the different analysis methods as in previous chapters, it has been split into the different diseases that were analysed. They follow the same order as shown in the previous section and outlined in Table 4.1.

4.3.1. *Mannheimia Haemolytica* Results

As with the analyses performed previously, the investigation begins with the average m/z peaks for the groups (Figure 4.3).

Again, the largest peaks on the average spectra are associated with ammonia and there was very little difference between the groups. However, some of the other m/z ratios showed quite large differences. With the synthetic mixtures trial (Chapter 2) the graphs of the average spectra of each sample group were quite effective. Unfortunately, a combination of more volatiles and more groups has made the average plot confusing and hard to interpret. Several changes were attempted,

including separate plots (which make comparison of the individual peaks difficult) and negative axis swapping (which causes problems with the log scale) but nothing had the desired impact.

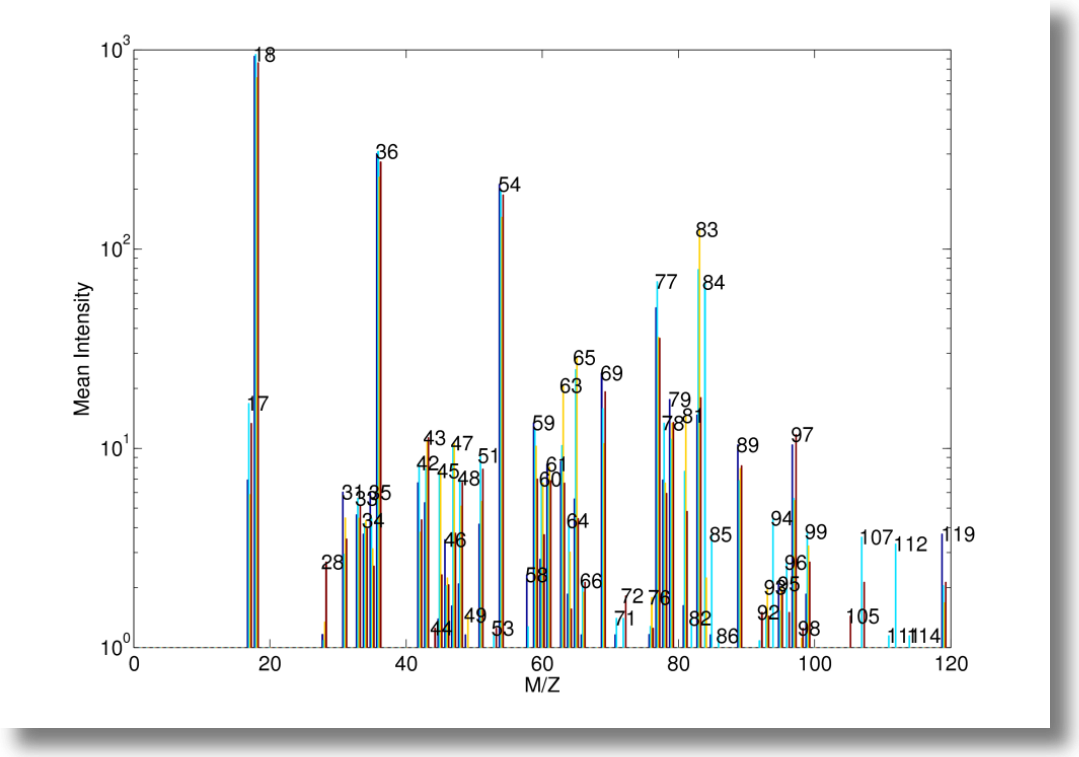


Figure 4.3 - The average m/z values for each of the groups of *Mannheamia Haemolytica* serum samples as recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. m/z ratios associated with the precursors have been removed. Blue is before introduction of the pathogen, turquise is post introduction group 1, yellow is post introduction group 2 and red if post introduction group 3

The initial PCA confirms that ammonia is the chief contributor to the variance. However, there is very little discrimination visible on the scores plot. The first step is to remove this volatile and then the PCA scores plot (Figure 4.4) shows some interesting features.

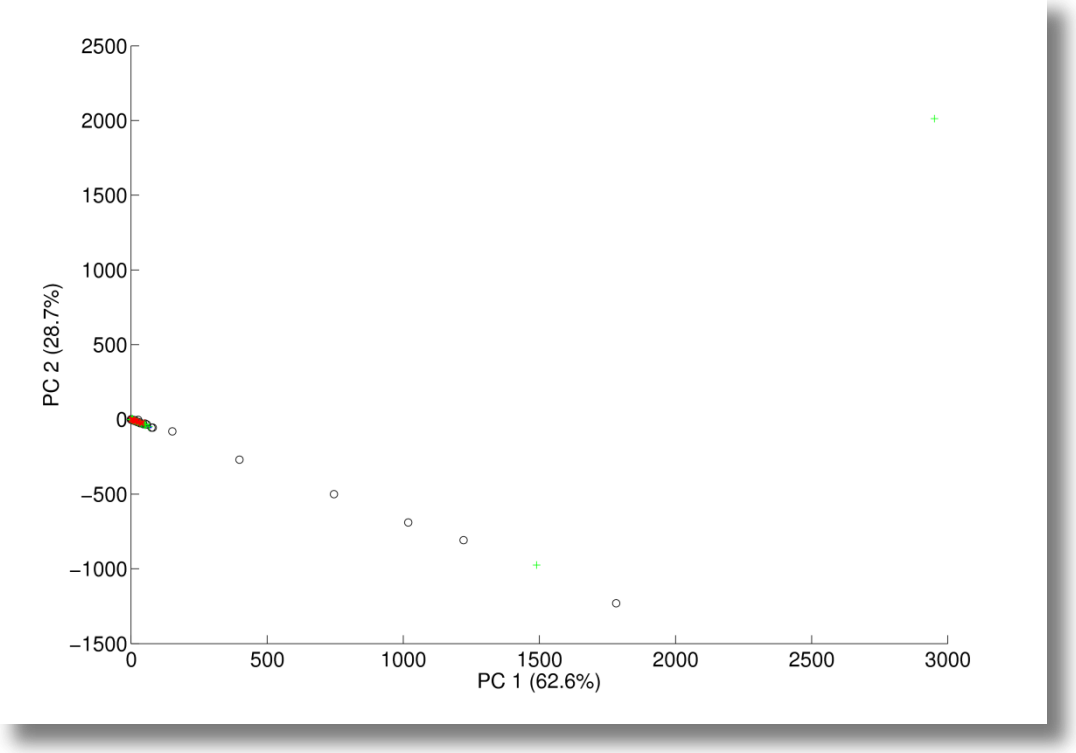


Figure 4.4 - PCA scores plot of the *Mannheamia haemolytica* serum samples as recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. The m/z associated with the precursors and ammonia (18, 36, 54) have been removed. Blue dots are pre-introduction, black circles are post group 1, green pluses are post group 2 and red x are post group 3 as described in Table 4.2.

There is some structure here and using the loadings it can be seen the single value out on PC2 at +2000 is caused by a large spike on m/z 84. As this is only a single value it is likely to be either a contamination or a random event recorded by the mk.2 SIFT-MS detector. The other feature is caused by a combination of m/z 65 and 83 which are associated with ethanol.

Both m/z 65 (Figure 4.5) and m/z 83 (Figure 4.6) have similar features. A handful of samples have elevated levels compared with the rest of the samples. The elevated samples come from the later stages of post-infection group 1 and all of post infection group 2. This occurs in too few of the samples to be used as a singular marker from this dataset but a further investigation into the causes would be of value. Removing

m/z 84 does not improve the separation as much as may be expected because m/z 65 and 83 then separate into two different PCs.

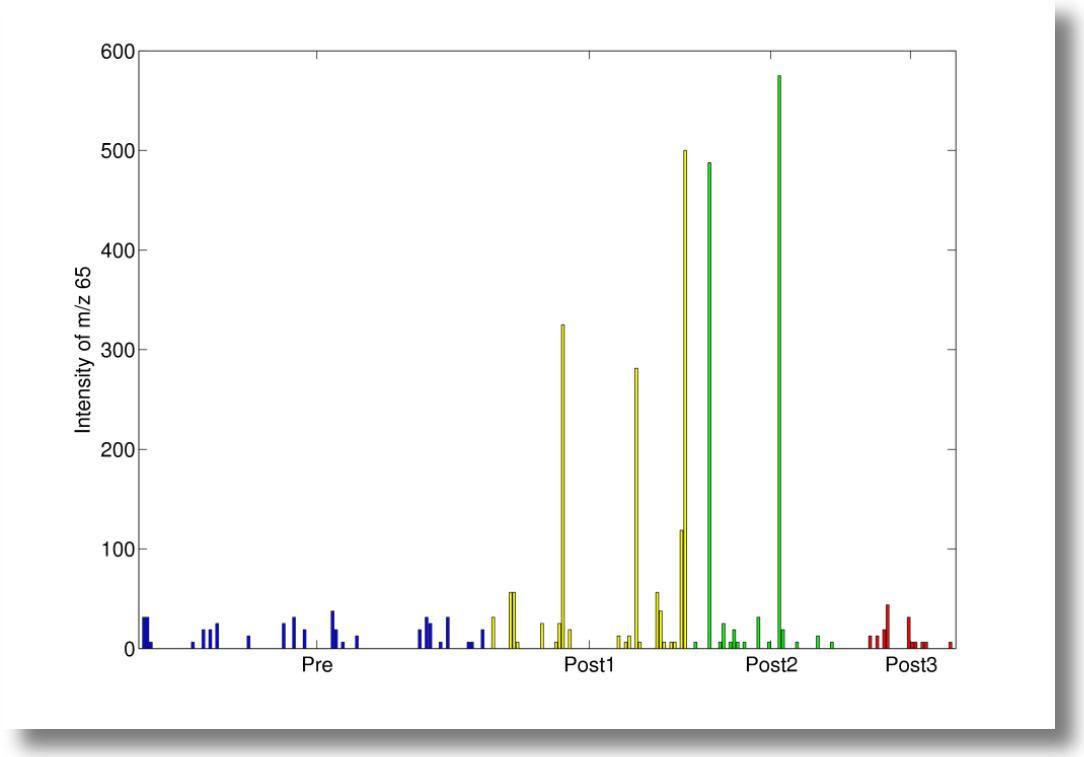


Figure 4.5 - The counts of m/z 65 for all the *Mannheamia haemolytica* serum samples as recorded using the H_3O^+ precursor spectra using the mk.2 SIFT-MS machine. The horizontal axis groups are those as described in Table 4.2.

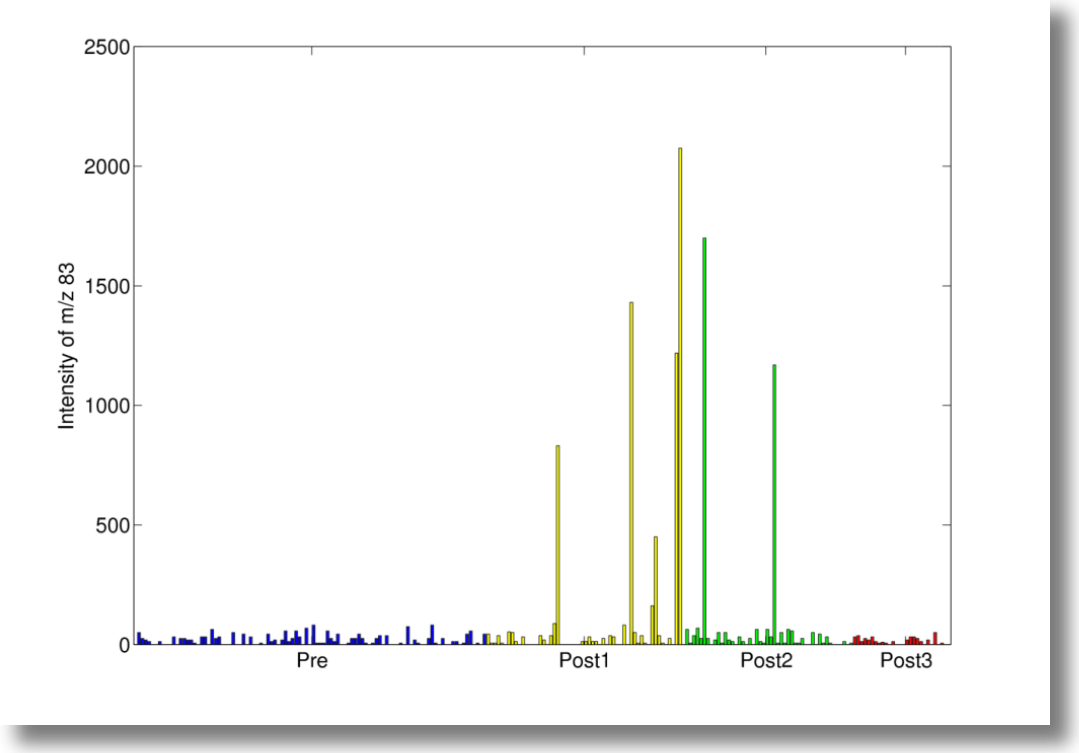


Figure 4.6 - The counts of m/z 83 for all the *Mannheamia haemolytica* serum samples as recorded using the H_3O^+ precursor spectra using the mk.2 SIFT-MS machine. The horizontal axis groups are those described in Table 4.2.

Further analysis showed that m/z 63, 47, 45, 81 all show a similar pattern to those of 65 and 83. Using this new information it should be possible to build a model of positive and negative samples using 14 days, 7 days and just before infection as negative samples and 0.5 days, 1 day and 2 days after infection as positive samples.

PLSDA managed to predict 69% of leave one out samples correctly using 7 latent variables (Figure 4.7). This shows the high complexity of the data and is a very good result especially due to the poor reliability of the possible biomarkers seen in the univariate analysis. As not all time points were used in the PLSDA analysis a total of only 114 samples were used in the model (60 negatives versus 54 positives).

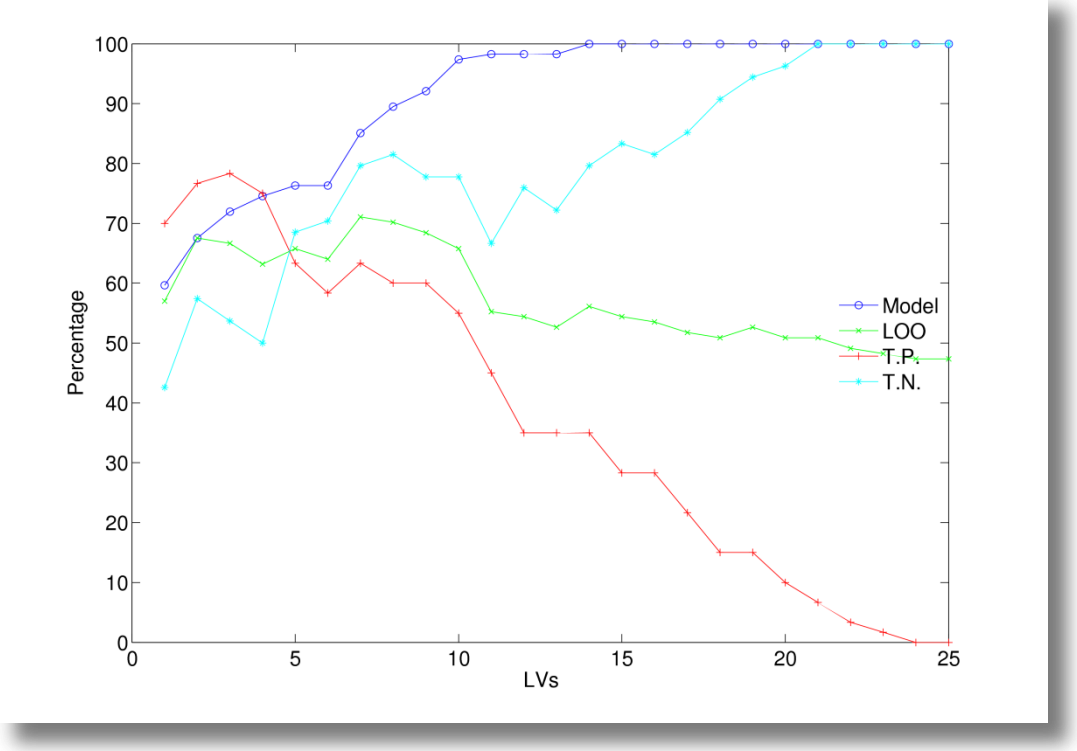


Figure 4.7 - PLSDA plot for positive (0.5, 1 and 2 days after pathogen was introduced) verses negative samples (14, 7 and just before pathogen was introduced) of *Mannhemia haemolytica* serum samples recorded using all three precursor spectra of the mk.2 SIFT-MS machine. The blue line shows the accuracy of the model tested using samples used in its creation, the green shows the accuracy of a LOO validation which is then further broken down into the percentage of LOO true negatives and LOO true positives

4.3.2. *Mycobacterium paratuberculosis* Results

As explained in the methodology part of this chapter not all of the *Mycobacterium paratuberculosis* samples were shipped to Cranfield with the main bulk of the samples. A small group of samples were therefore not part of the randomisation process and were all sampled together after the others.

The 20 samples recorded later were clearly different from the samples recorded during the original analysis stage. They can be seen to separate off from the main cluster of samples in Figure 4.8. Whether this difference is caused by the sample

being from a different farm, or because of a difference in the way they were stored and sampled, or if the difference was just down to machine drift is not clear but this confirms the benefits of a randomised sampling and measuring order.

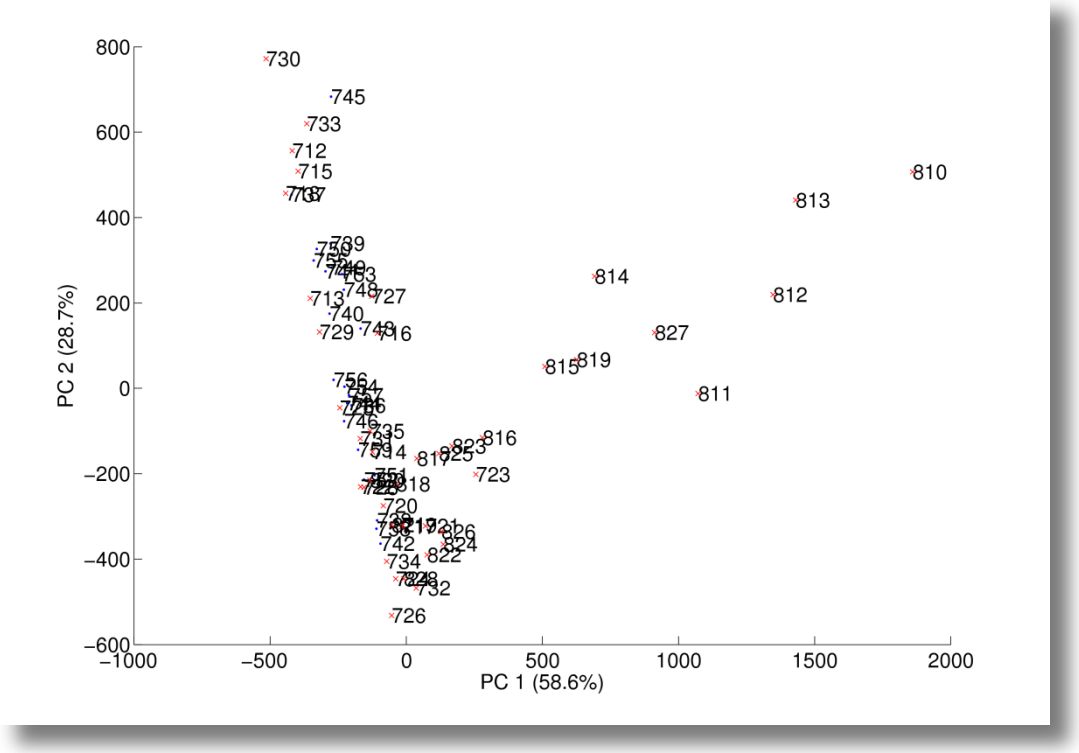


Figure 4.8 - PCA of the *Mycobacterium paratuberculosis* serum sample spectra as recorded using the H_3O^+ precursor spectra of the mk.2 SIFT-MS machine. Samples numbered greater than 800 were recorded after the trial had concluded hence were not part of the randomisation procedure. m/z associated with the precursors have been removed and the data was pretreated using a mean centring algorithm.

These 20 samples were removed from the dataset prior to further analysis.

In the average spectra plot in Figure 4.9 ammonia has the largest peaks but the differences between the means are some of the smallest of all the m/z ratios. There are much larger differences in some of the higher mass peaks. The initial PCA analysis shows that ammonia (m/z 18, 36 and 54) is again the largest contributor to

the variance but with no difference between the means of the sample groups, discrimination is once again poor.

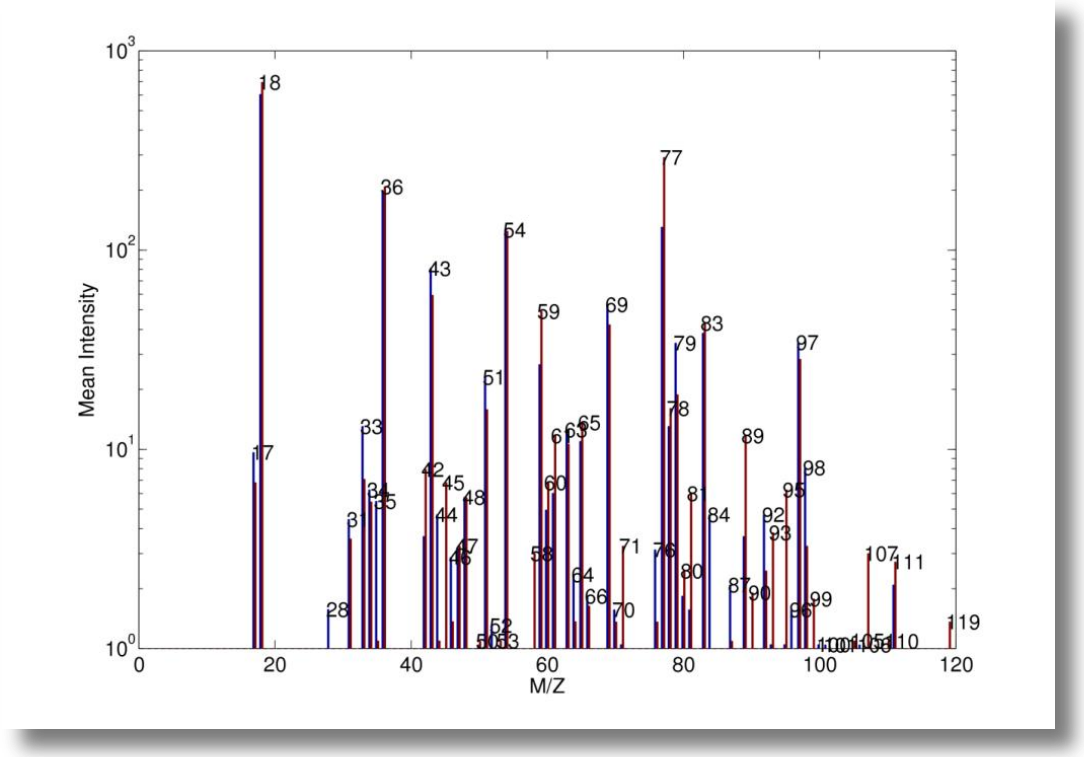


Figure 4.9 - The average counts for the spectra of *Mycobacterium paratuberculosis* serum samples as recorded using the H_3O^+ precursor spectra of the mk.2 SIFT-MS machine. Blue bars are negatives and red bars are positives.

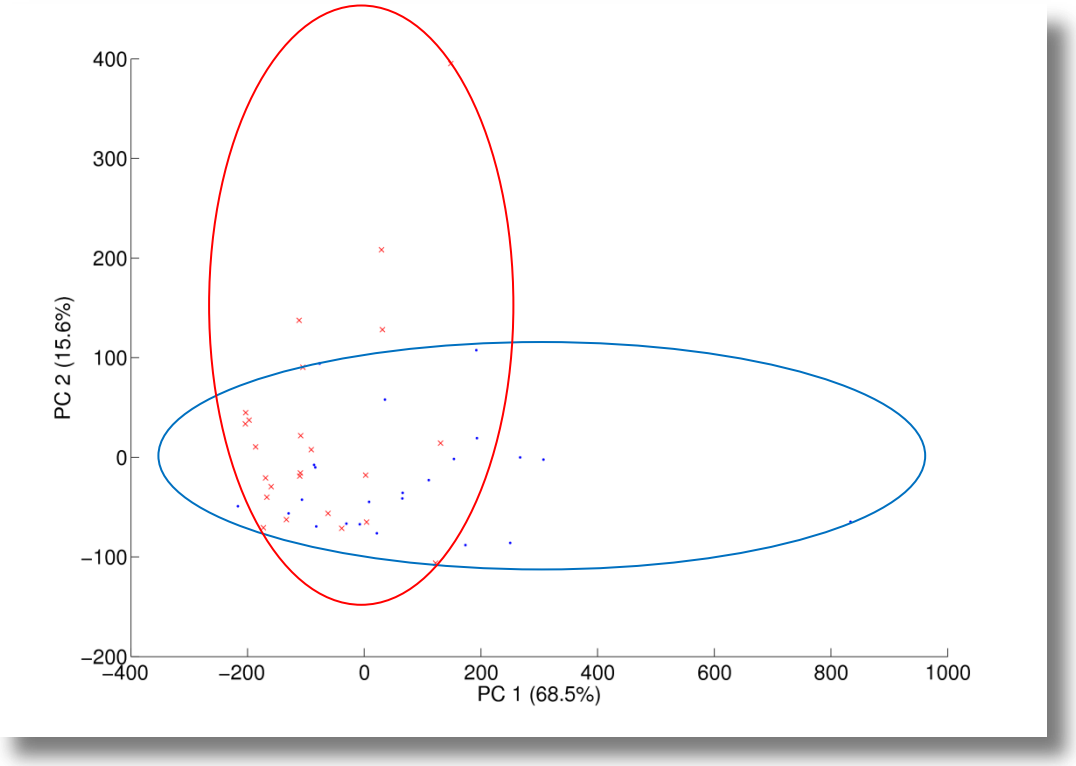


Figure 4.10 - PCA scores plot of the H_3O^+ precursor spectra from the samples of Mycobacterium paratuberculosis recorded using the mk.2 SIFT-MS machine. Samples numbered above 800 that were not recorded in a randomised order have been removed. m/z ratios associated with the precursors and ammonia (18, 36 and 54) have been removed and the data was pretreated using a mean centring algorithm. Approximate clusters have been draw outlining the slight discrimination.

After the removal of m/z ratios associated with ammonia, a partial discrimination can be seen between the positive and negative *Mycobacterium paratuberculosis* sample on the first PC axis of the above scores plot (Figure 4.10). This discrimination is likely to be caused by m/z 59 and 77, ascertained from an investigation of the loadings (Figure 4.11) and plots of all peaks of the m/z ratios (not shown). These m/z ratios (59 and 77) are associated with acetone levels. Further analysis does not reveal any other significant discriminators.

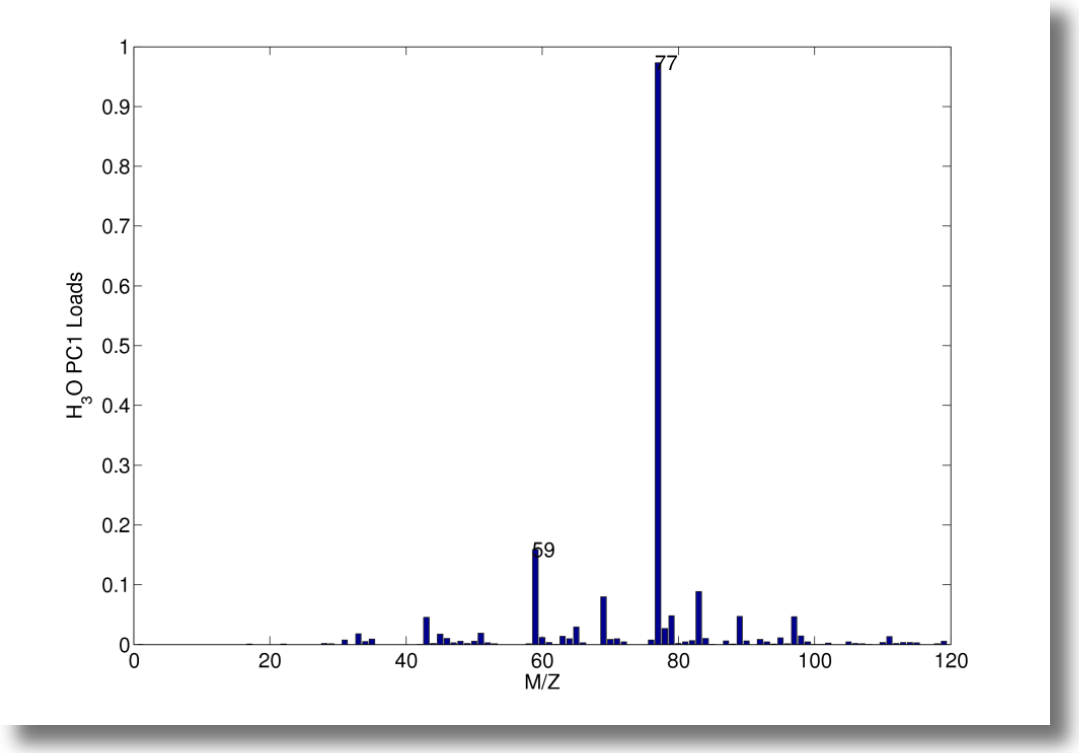


Figure 4.11 - The loadings of PC1 for the PCA analysis of the *Mycobacterium paratuberculosis* cattle serum sample recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. Samples numbered above 800 that were not recorded in a randomised order were removed. m/z ratios associated with the precursors and ammonia (18, 36 and 54) were removed and the data was pretreated using a mean centring algorithm.

The PLSDA results (Figure 4.12) show a maximum LOO accuracy of 72% using 3 LVs. Also, after this point the LOO accuracy falls back to even chance which suggests that the only discriminators are the ones already found.

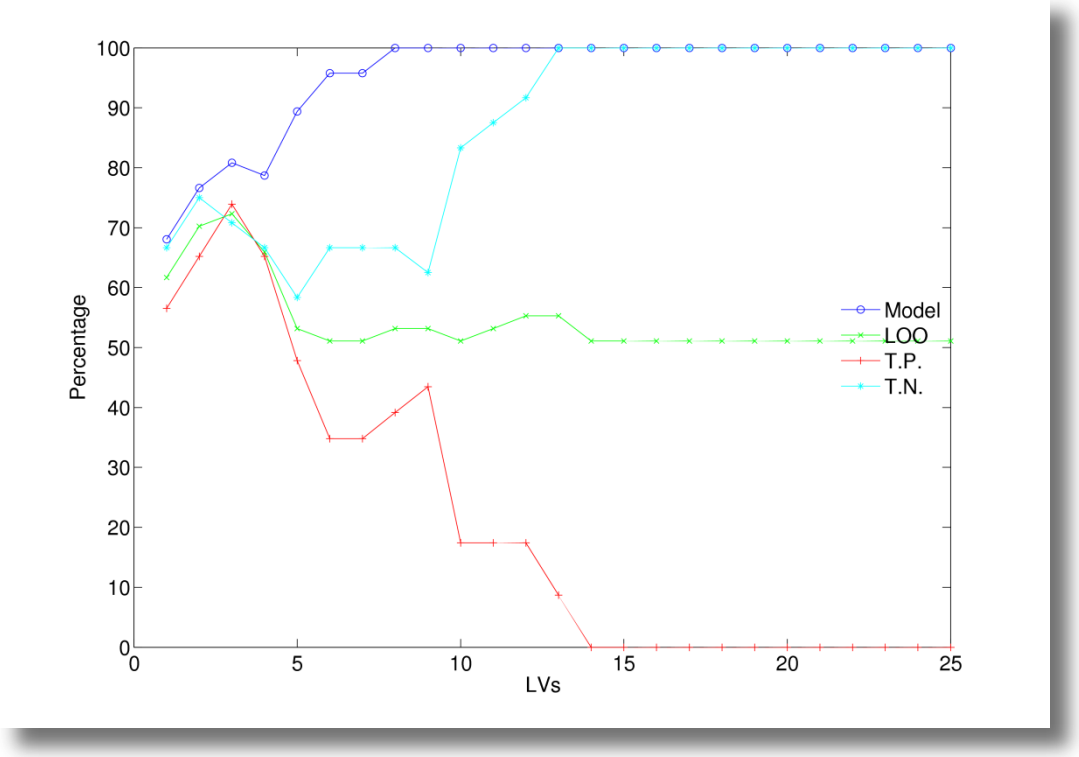


Figure 4.12 - PLSDA plot for the *Mycobacterium paratuberculosis* cattle serum samples recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. Samples number above 800 that were not recorded in a randomised order have been removed. m/z ratios associated with the precursors have been zeroed. The blue line indicates the accuracy of the model when tested against samples used in its construction, the green line indicates the accuracy of the LOO validation which is then further broken down into the LOO true negatives (red) and true positives (turquoise).

4.3.3. Brucella results

The average spectra from the brucella samples (Figure 4.13) show a large number of volatiles that have large differences between the means of their associated peaks.

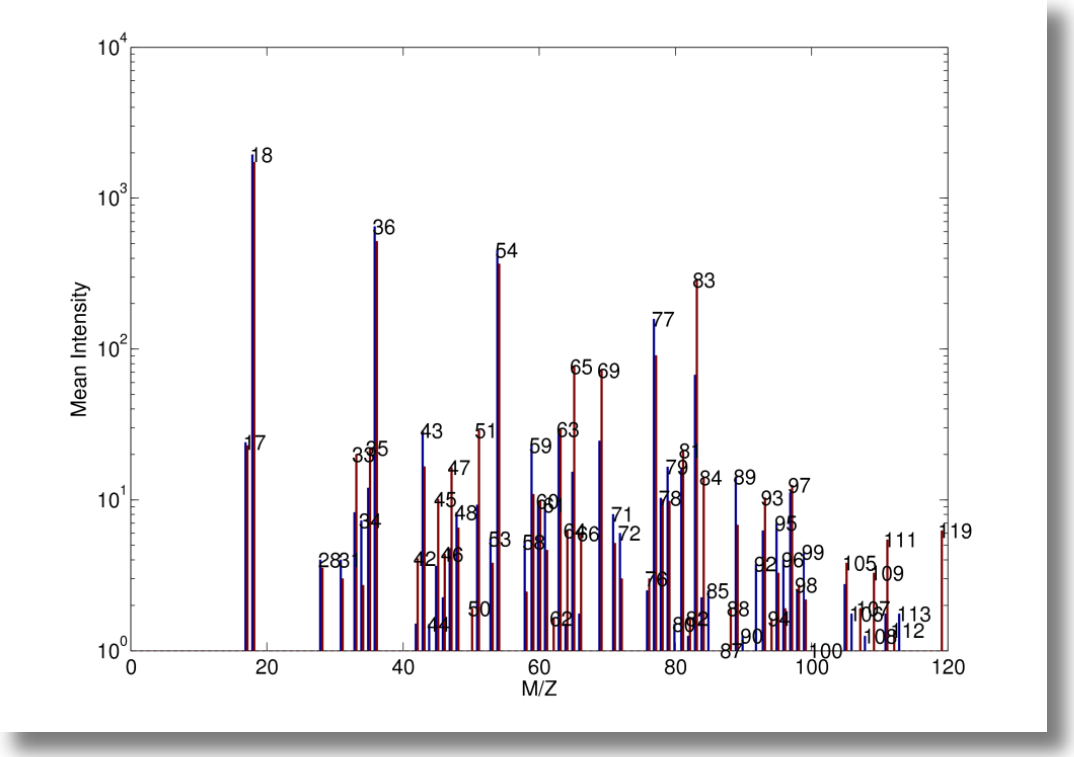


Figure 4.13 - The average spectra for the positive (red) and negative (blue) brucella in cattle samples as recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. m/z ratios associated with the precursors have been removed.

The average plot shows some differences between the means of some higher mass peaks (e.g. 83). The peaks of ammonia (18, 36 and 54) are once again the largest and have smaller differences between them than many of the other m/z ratios.

Figure 4.14 shows a sample (no. 779) that is significantly different for all the others. Investigating the loadings and confirming with visual comparisons between the peak heights, sample 779 has much higher levels of m/z 83 and also m/z 65 when directly compared with the other samples. Both of these m/z ratios relate to the levels of ethanol present in the samples. If only one m/z ratio had a significant increase over the other samples it would be hard to confirm whether it was a machine spike or real VOC levels but with two it is more likely that sample 779 has much higher levels of ethanol than the other samples. Knowing that the SIFT-MS machine is not the culprit and it is more likely to be caused by volatiles in the sample itself is not overly useful

information. The sample is still an outlier in regards to disease detection or biomarker discovery as one sample that has a high ethanol level does not give any information about the disease status. It simply means that this one sample is considerably different from the others but the reason for this difference is unknown. After further investigation it was found that sample 779 had further differences (but less significant) on other m/z ratios so before further analysis it was considered to be an outlier and removed from the dataset.

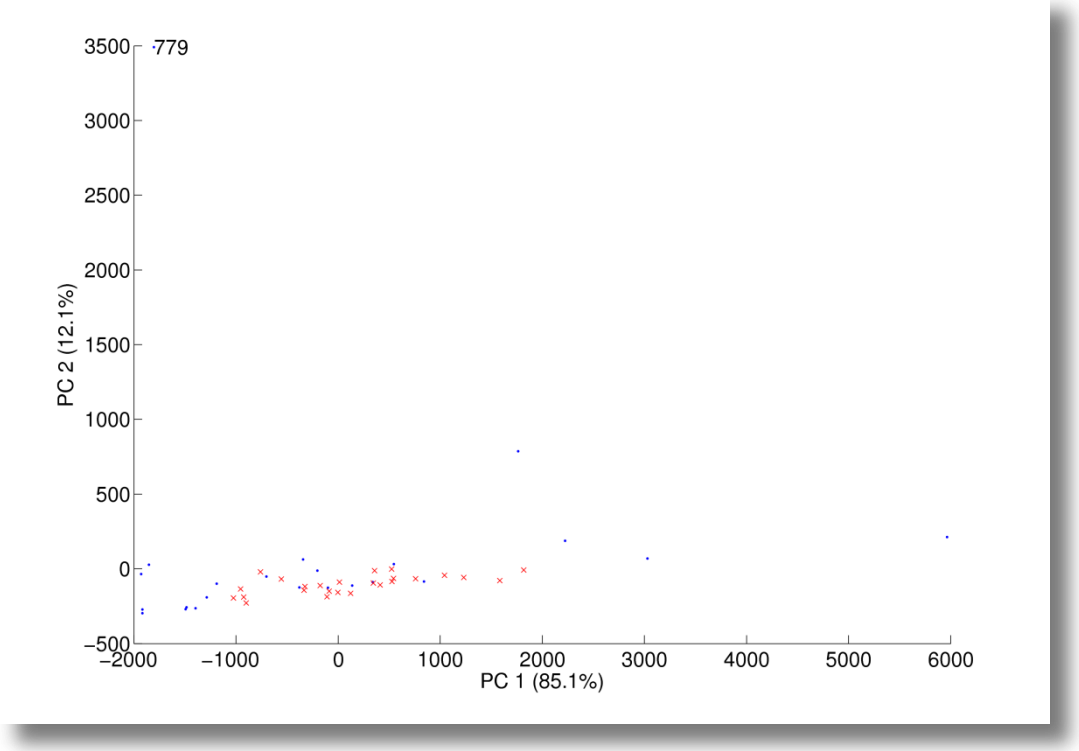


Figure 4.14 - PCA scores plot of the Brucella data collected by using the H_3O^+ channel of the mk.2 SIFT-MS machine showing one sample (no. 779) which stands out from the rest. m/z ratios associated with the precursors have been zeroed and the data was pre-treated using a mean center algorithm

After sample 779 had been removed, the PCA analysis on the dataset produced characteristics similar to the others previously discussed in this chapter (Figure 4.15). m/z ratios associated with ammonia dominated the loadings plot of PC1 for the

above scores graph but produce little discrimination on this axis. The loadings of PC2 are dominated by the variance caused by m/z 69 which is associated with methanol.

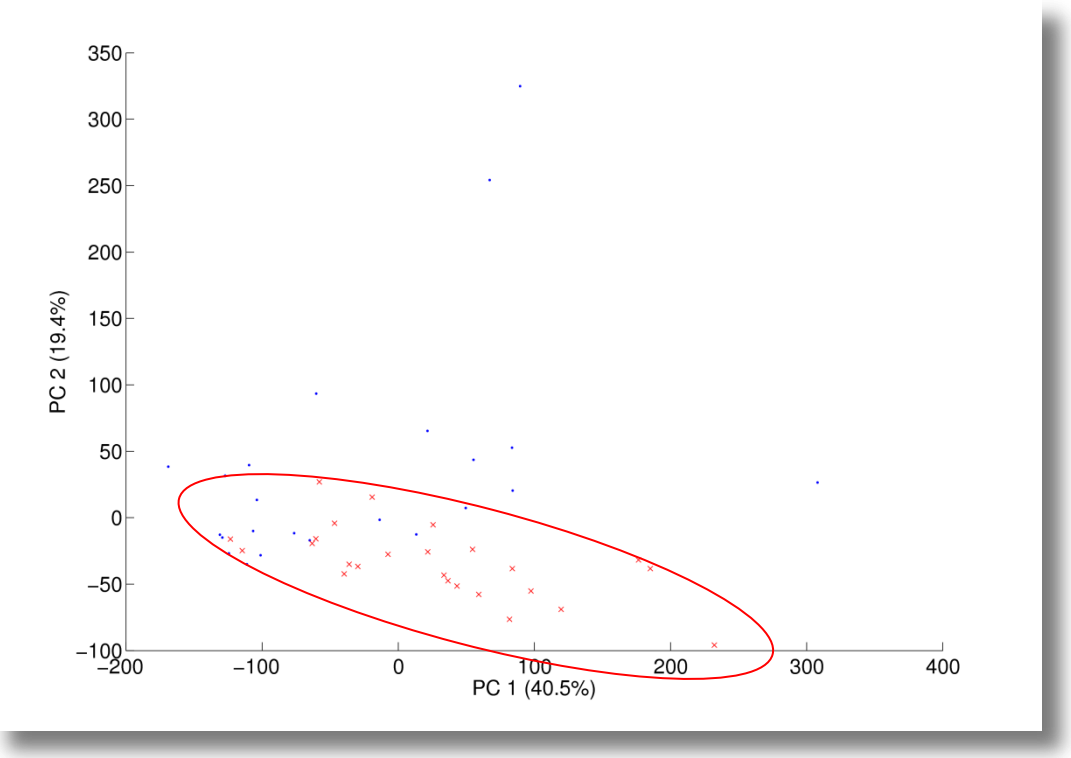


Figure 4.15 – PCA scores plot of the Brucella serum sample data collected using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. m/z ratios associated with the precursors have been removed. Red samples are brucella positives and the blue samples are brucella negatives. The data was pre-treated using a mean center algorithm. An approximate cluster around the positive sample has been marked on the graph. Approximately half the negative samples are within this cluster but they are more spread around the graph.

Figure 4.16, shows that approximately half of the negative samples have higher levels of m/z 69 than the highest positive sample.

Going further into the variance of the data, more examples of a single sample (like 779) with much higher values that all the other are found (Figure 4.14). Many different samples are affected, each with m/z peaks with higher values than the rest. It

is almost as if every sample has at least a single VOC with a large significant difference from all the others. The results of building classification models using PLDSA are shown in Figure 4.17.

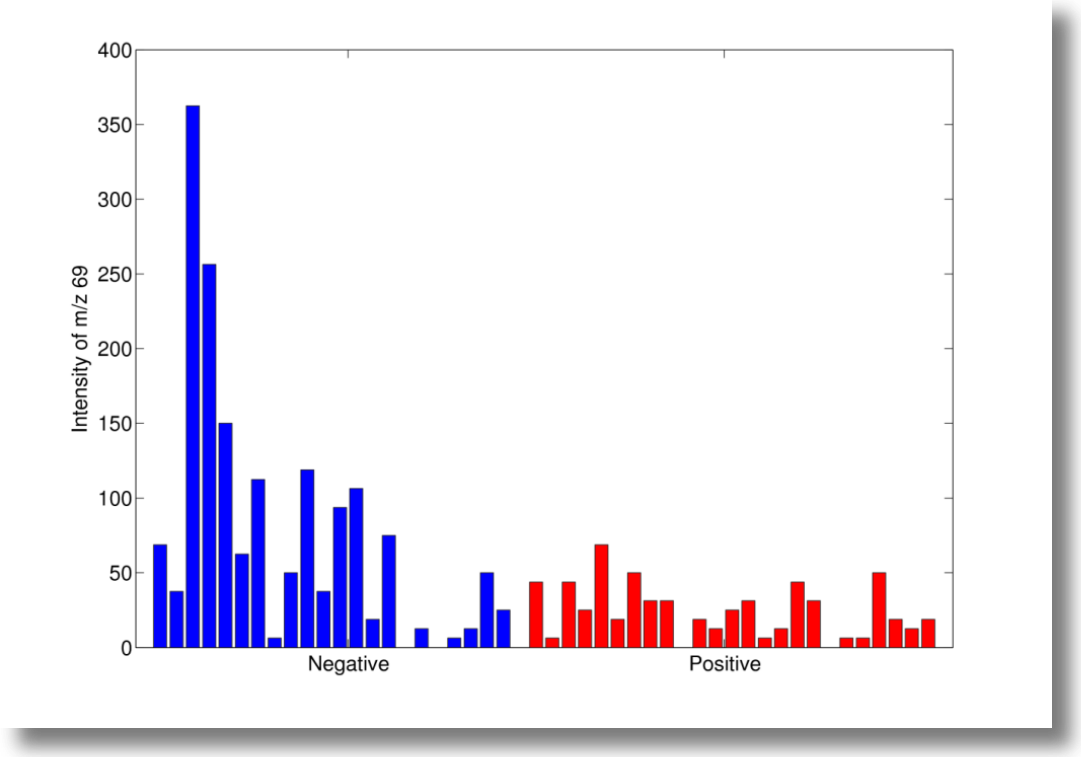


Figure 4.16 - The levels of methanol (m/z 69) as recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine for serum samples of brucella

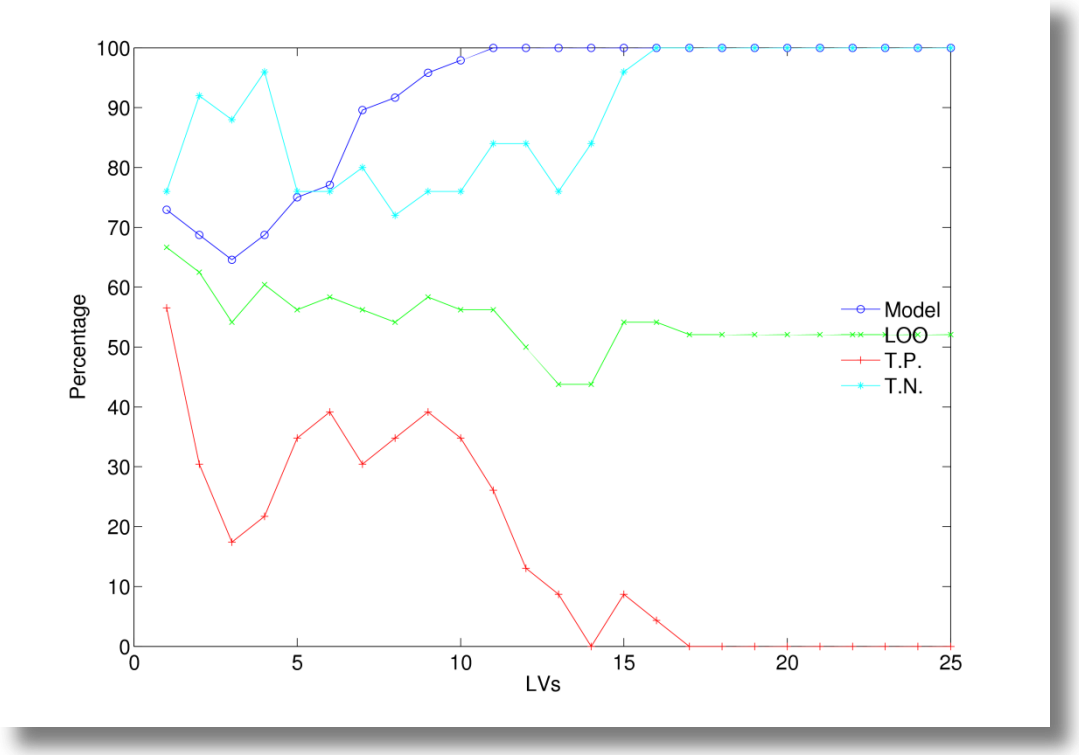


Figure 4.17 - PLSDA curves for the brucella data using data from all three precursor channels of the mk.2 SIFT-MS machine. The dark blue line shows the accuracy of the PLSDA model when tested with samples used in its production, the green curve shows the accuracy of the PLSDA as produced by a LOO validation and these results are further broken down into the LOO true negatives (red) and true positives (turquoise).

The PLSDA model fails to produce a LOO accuracy much above even chance with a best LOO accuracy of only 66%. Also, it shows that a large number of latent variables (11) are required to build a model that has good accuracy when tested against itself. This suggests there is high dimensionality to the data which agrees with the hypothesis that every sample has one or more individually high volatiles like ethanol for sample 779. The biggest problem with this dataset is that there is no control on the samples - they come from a range of decades and a range of European countries. With such a high complexity and a small sample size (48 samples) it is unsurprising that PLSDA validation failed (Brereton, 2006). More samples or more control is required (ideally both) for further meaningful investigation.

4.3.4. Mycoplasma Bovis Results

Figure 4.18 shows the average m/z counts for the four sample groups.

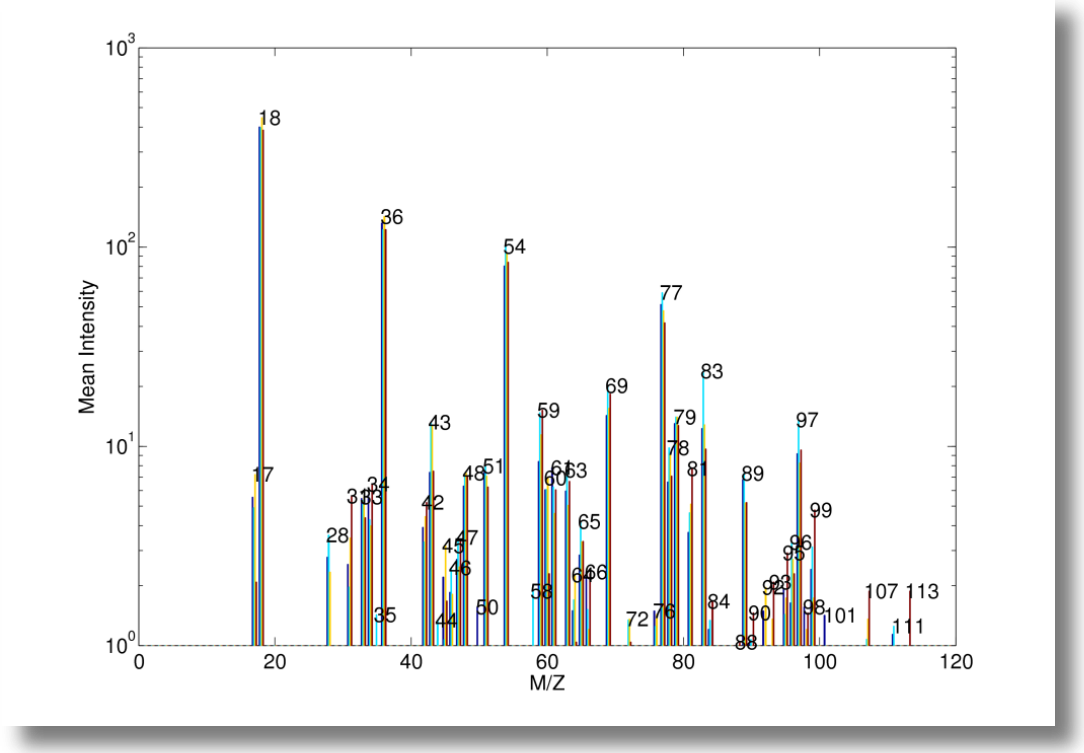


Figure 4.18 - The average counts for the Mycoplasma bovis data groups spectra as recorded using the H3O+ precursor channel of the mk.2 SIFT-MS machine. Before introduction of the pathogen are blue, group 1 are turquiose, group 2 are yellow and group 3 are red with the samples belonging to each group expained in Table 4.3. The m/z ratios associated with the precursors have been removed

The highest peaks (after the precursors have been removed) are associated with ammonia. The mean peaks for each m/z ratio appear to be closer together for these groups than those associated with other dataset. This in itself is not conclusive of anything and significant difference may still be possible if the data within each group is consistent. As with Figure 4.3, Figure 4.18 is difficult to interpret due to the complexity of the data and the limits of this simple univariate technique.

Looking at the loadings of the above PCA score plot (Figure 4.19) showed that ammonia (m/z 18, 36 and 54) was contributing the most variance to the model.

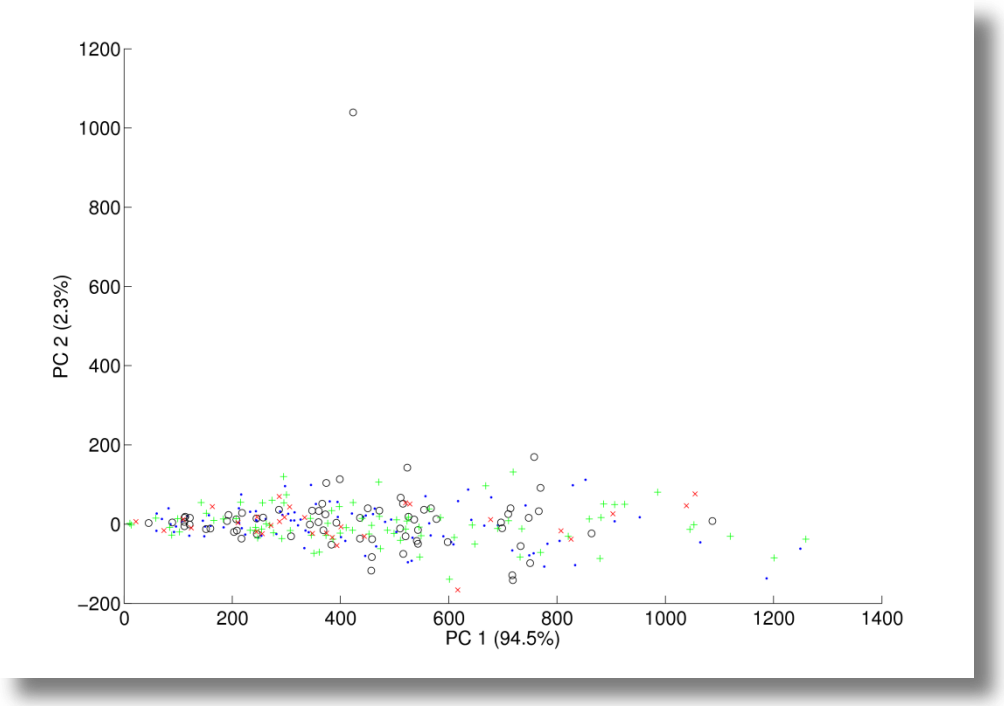


Figure 4.19 - PCA score plot of the *Mycoplasma bovis* dataset using the H_3O^+ precursor spectra of the mk.2 SIFT-MS machine. Blue dot are before the pathogen was introduced, black o are post introduction group 1, green + are post introduction group2 and red x are post introduction group3. No discrimination can be seen apart from the single outlier.

Figure 4.20 shows the peak height of m/z 18 as recorded using the H_3O^+ precursor channel of the mk.2 SIFT-MS machine. The variance spreads for all three ammonia peaks are have very similar patterns to these shown in Figure 4.20, i.e. there is a large difference between the samples but these differences are spread evenly across the sample groups and hence it is unlikely to be linked to *Mycoplasma bovis*. The outlier on the PCA graph (Figure 4.19) is caused by a single sample with very high ethanol levels.

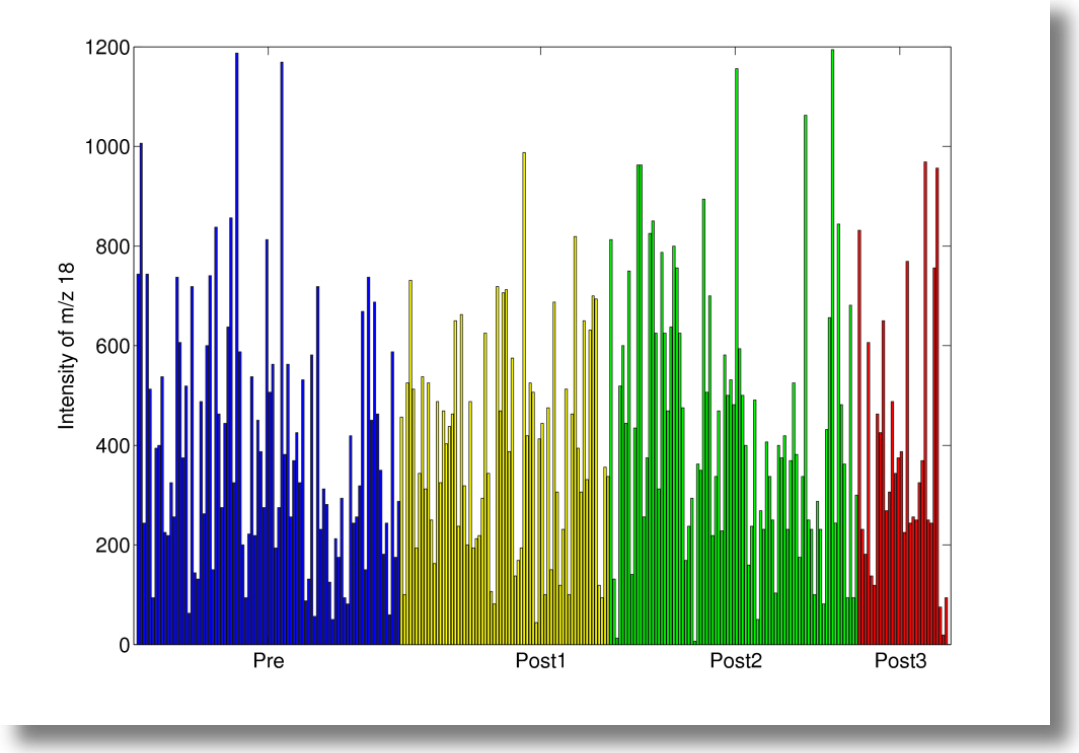


Figure 4.20 - The intensities of m/z 18 as recorded using the H_3O^+ precursor of the mk.2 SIFT-MS machine for all the samples of serum for the *Mycoplasma bovis* dataset. The horizontal axis represents the groups that the samples belong to as descibed by Table 4.3

Working through the next features that have high variance reveals similar graphs for acetone, propanol, methanol, acetaldehyde before finally reaching the volatiles at the limit of detection of the SIFT-MS mk.2 machine using FSM and an analysis time of 5 minutes (i.e. there are only tens of counts). One possibility is that the pathogen introduced was not strong enough to infect the animals causing a host response as hypothesised by the clinicians.

4.3.5. Bovine TB (*Mycobacterium bovis*) Results

With the larger number of factors in the meta data, choosing how to display the results of the analysis of this dataset is quite a challenge. Equally, this extends to the choice of classification vector for PLSDA.

Initially, looking at both the discrimination associated with either the time point or the immunisation type showed very little. However, after ammonia (m/z 18, 36, 54 and 72) had been removed, the following PCA scores plot was achieved.

The discrimination seen on the PCA scores plot (Figure 4.21) is that all of the samples of time point 1 (blue dots) are on the negative portion of PC1 (to the left of the dashed line) whereas the majority of the rest of the samples are on the positive portion of PC1. There are some crossover samples on the negative side of PC1.

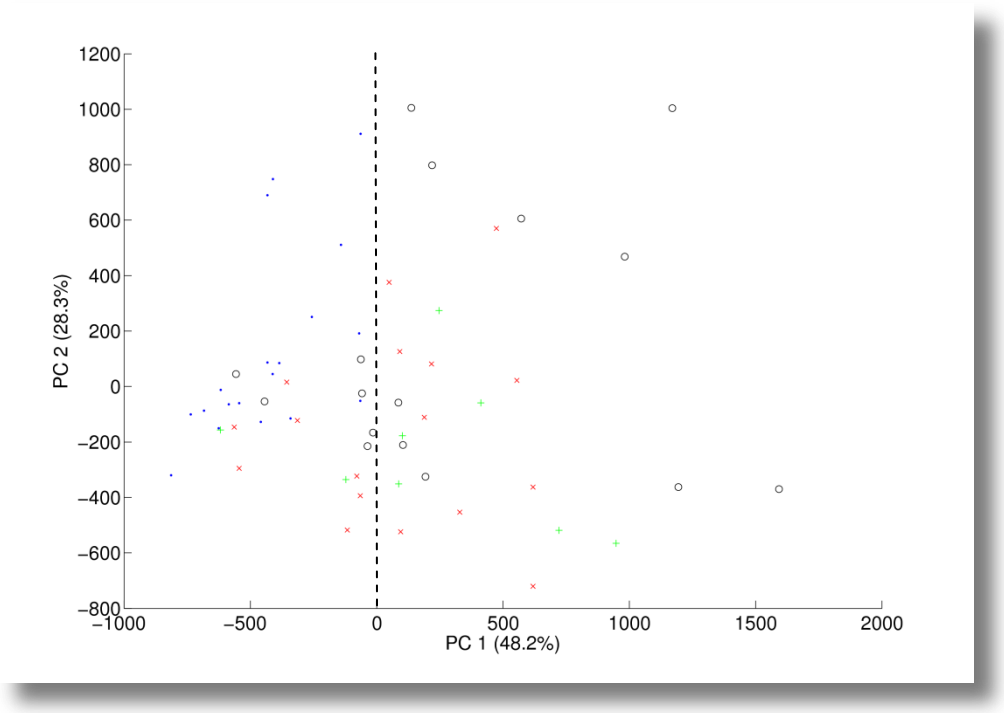


Figure 4.21 - PCA score plot for the bovine TB (*Mycobacterium bovis*) samples as recorded by the H_3O^+ precursor channel of the mk.3 SIFT-MS machine. m/z ratios associated with the precursors and ammonia (18, 36 and 54) have been removed. The data was pre-treated using a mean centring algorithm. Blue dot are time point 1, black o are time point 2, green + are time point 3 and red x are time point 4. A dashed line has been drawn across the graph at the origin of the first principal component axis.

This suggests that the SIFT-MS provided sufficient data to allow partial discrimination between samples taken before the vaccination step (samples taken at time point 1) and after the vaccination step (samples taken over the other three time points) when PCA analysis was performed. Indeed, closer investigation of the samples show misclassified consisted of a significant proportion of the samples appearing with the pre-vaccination time point 1. These are samples that are part of the control group which were not vaccinated. Therefore, to remove this problem, the best way forward is to look at only the control group animals. Now, as this was a vaccination trial, the control group still had the bovine TB (*Mycobacterium bovis*) pathogen introduced, so by looking at the serum sample taken before introduction and after introduction of the pathogen it should be possible to see if there are any differences.

As these samples were measured using the newer mk.3 SIFT-MS machine there are more m/z counts that have non-zero values (Figure 4.22). As such, there are a lot more variables to consider during the analysis. However, the values of the main volatiles are still generally high such as those for ammonia (m/z 18, 36, 54 and 72), etc.

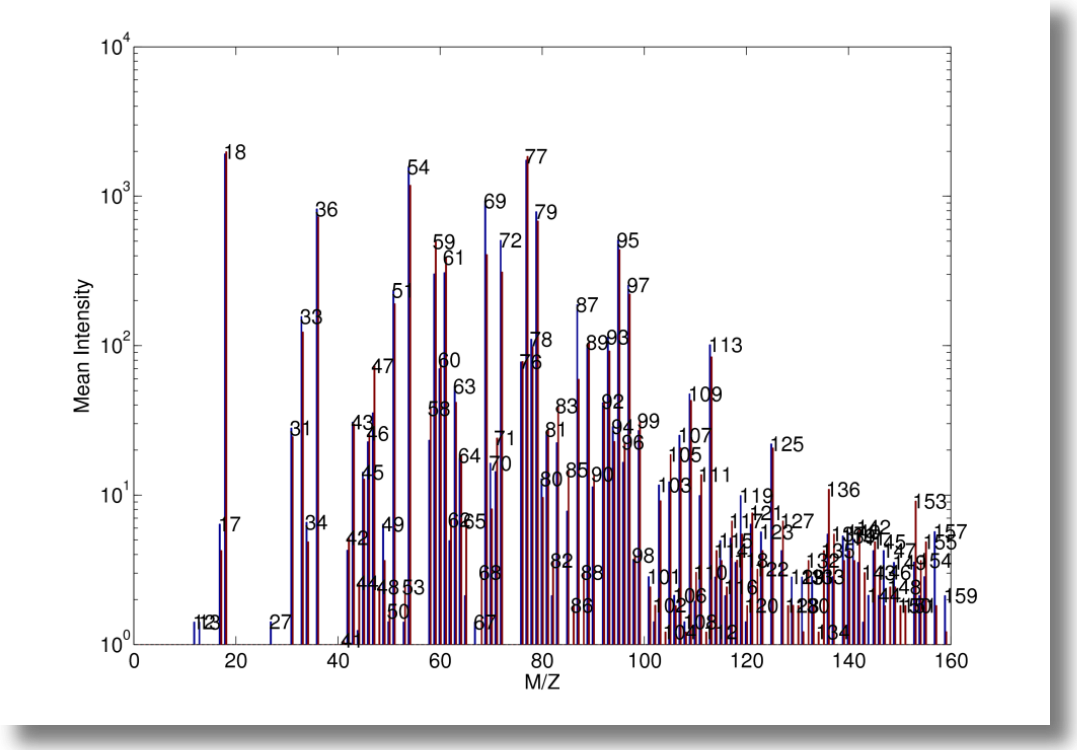


Figure 4.22 - Plot of the average m/z counts for each sample group of the bovine TB (*Mycobacterium bovis*) serum control samples as recorded using the H_3O^+ precursor channel of the mk.3 SIFT-MS machine. m/z ratios associated with the precursors have been removed. The blue bars represent the m/z levels before the pathogen was introduced and the red bars represent the m/z levels after the pathogen was introduced.

Investigating the variance using PCA showed that ammonia was the largest contributor to variance but it again appears to have nothing to do with the disease status. This is a slightly surprising result as ammonia appeared to be a good discriminator for tuberculosis in the badger serum dataset discussed in chapter 3. After ammonia, acetone is the next biggest contributor to the variance. It also appears to have very little to do with the disease status. After the m/z ratios associated with these two volatiles have been zeroed the PCA score plot in Figure 4.23 was achieved.

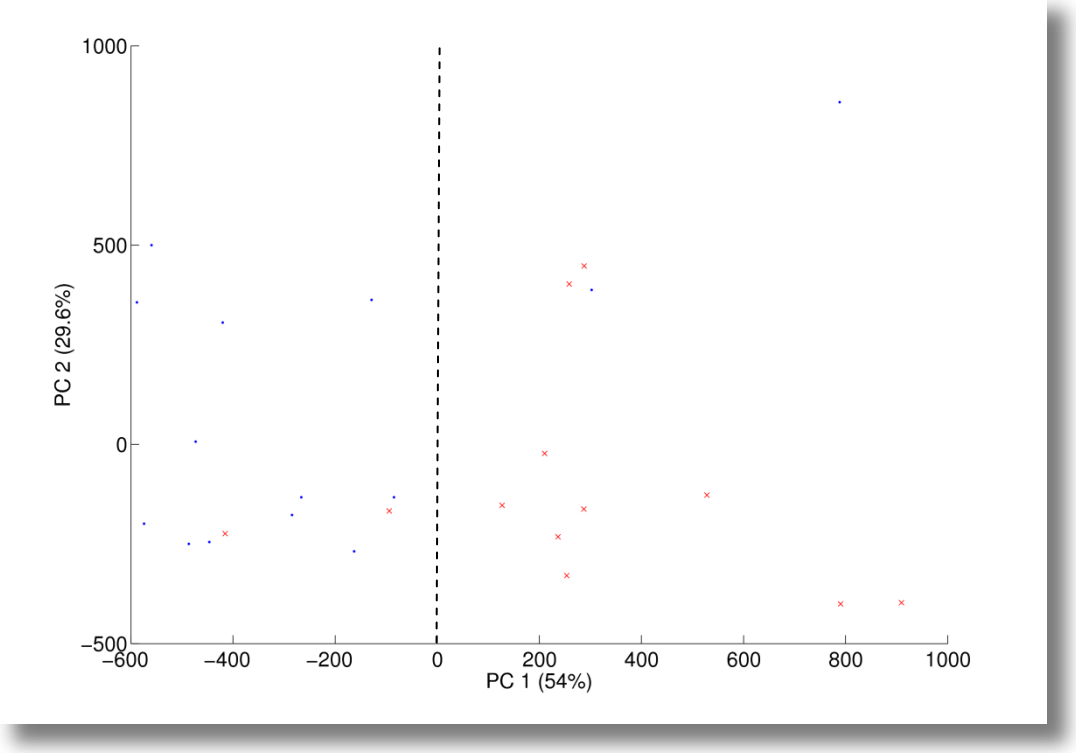


Figure 4.23 - PCA scores plot of the control groups serum samples from the bovine TB (*Mycobacterium* bovis) clinical trial measured using the newer mk.3 SIFT-MS machine using only the H3O+ precursor spectra. Blue dot are negative samples and red x are positive samples. m/z associated with ammonia (m/z 18, 36, 54 and 72) and acetone (m/z 59 and 77) have been zeroed. The origin of PC1 has again be drawn with a dashed line

The above discrimination seen on the first component of the above PCA plot is caused by differences seen in the levels of methanol, especially m/z 69. The same features are seen on all the m/z associated with methanol (33, 51, 69 and 87) but the differences are clearer on the two heavier m/z ratios (69 and 87). Due to this, both PCs are linked to the four methanol m/z peaks but the variance linked to PC1 is mostly being caused by the heavier two peaks (69 and 87) and PC2 is mostly being caused by the lighter two peaks (33 and 51).

With such good linear separation on one axis in the PCA scores plot (Figure 4.23), confirmed with the individual levels of m/z 69 (Figure 4.24), PLSDA should give some good results.

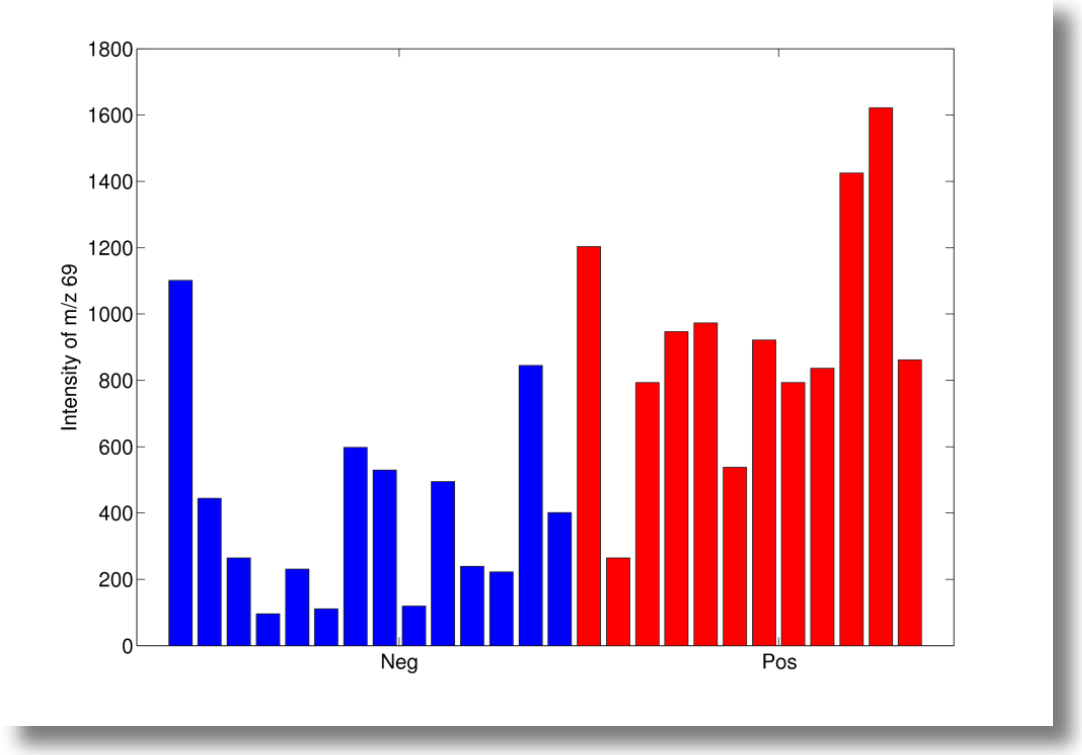


Figure 4.24 - The levels of m/z 69 which is associated with methanol for the serum samples from the Tuberclosis bovis control group with negative samples being before infection and the positive samples being after infection

Figure 4.25, shows the LOO accuracy as 72% with 2 LVs of complexity. It is surprising that the PLSDA results are not far better, given the clear separation seen in the PCA (Figure 4.23) and the counts of m/z 69 (Figure 4.24). The model does become quickly overtrained, it takes only 9 LVs for the LOO to level at even chance, which suggests that there may be too few samples. In the previous tuberculosis PLSDA results, using only the H_3O^+ spectra with selected variables did not improve the results (see chapter 2 and 3). However, in those examples, there were more samples and the data was collected using the older mk.2 SIFT-MS machine.

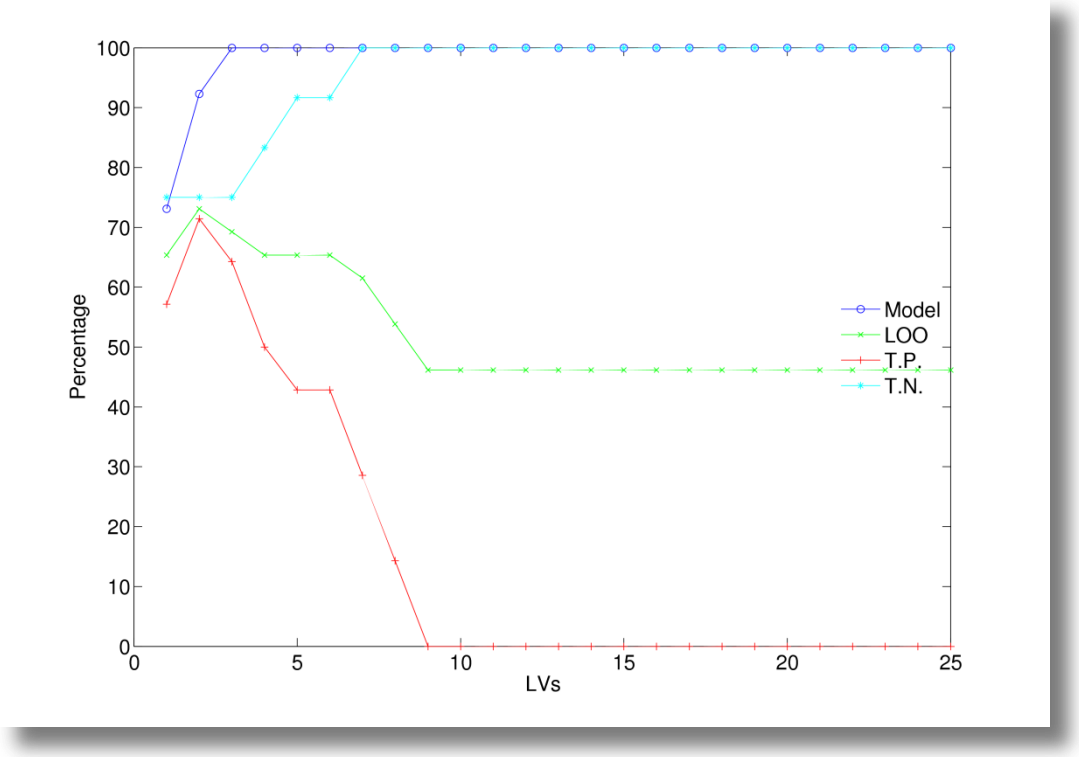


Figure 4.25 - PLSDA plot for the data collected tuberculosis serum cattle control group samples showing a quickly overtrained model suggesting that there are too few samples. The dark blue line shows the accuracy of the model when tested with samples used in its creation, the green line shows a LOO accuracy which is broken down into the number of LOO true negatives (red) and LOO true positives (turquiose).

Figure 4.26 shows that a much higher accuracy (85% LOO accuracy) can be achieved than was first found (Figure 4.25). There can be a wealth of information within the SIFT-MS data, but this result highlights the complexity of finding it and extracting it to allow meaningful interpretation. With the previous mk.2 SIFT-MS machine, removing variables never improved the results of the PLSDA analysis. This is clearly not the case with the mk.3 SIFT-MS. This goes to further highlight that it is not necessarily the m/z with the highest variance, or co-variance that are the most important. Yes, generally, higher variance gives better results but sometimes they are just noise or have a large natural variance.

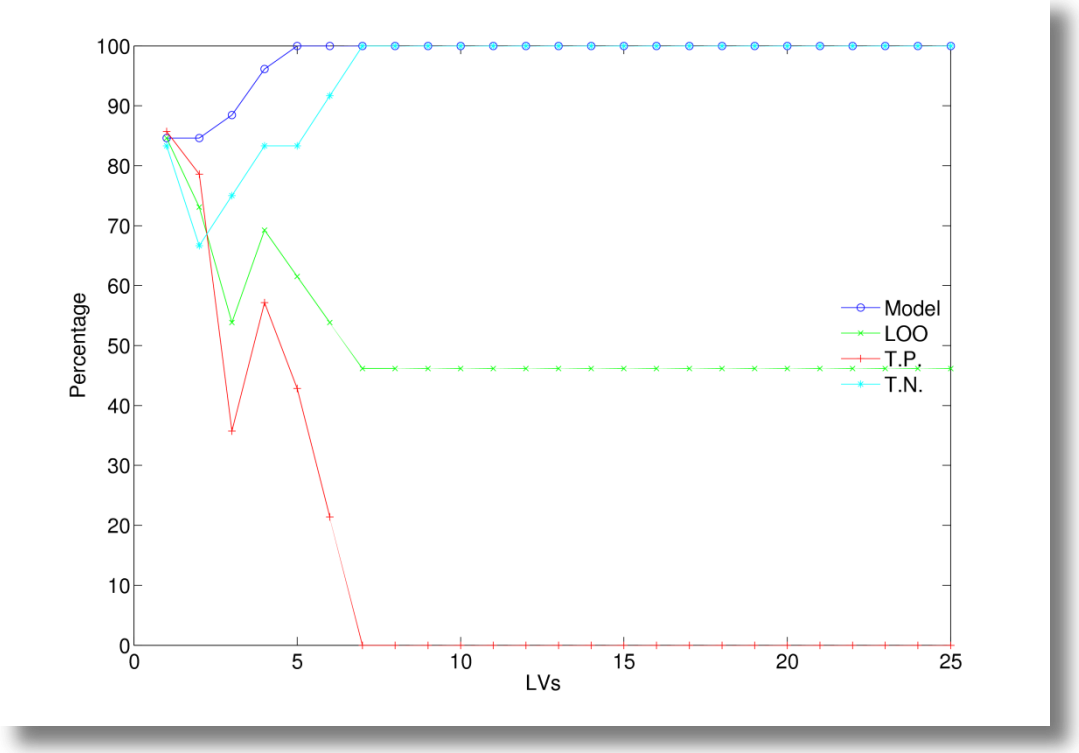


Figure 4.26 - PLSDA plot for the bovine TB serum from the control samples only using only H_3O^+ precursor channel of the mk.3 SIFT-MS machine and after m/z associated with ammonia (m/z 18, 36, 54 and 72), acetone (m/z 59 and 77) and propanol (m/z) have been removed. The dark blue line shows the accuracy of the model when tested with samples used in its creation, the green line shows a LOO accuracy which is broken down into the number of LOO true negatives (red) and LOO true positives (turquiose).

It also highlights the need for a large number of samples. The higher the number of samples the better the capture of variance and the more likely that the co-variance linked to the disease being studied could be spotted.

4.4. Discussion and Conclusion

The results from the cattle serum trial were varied. The best results were obtained from the analysis of the bovine TB trial which was the only dataset recorded using the newer mk.3 SIFT-MS machine. When looking at only the control group, an 85% LOO accuracy (86% true negative and 84% true positive) was achieved, which was

associated with methanol as a possible marker. This is better than the accuracy achieved with the samples taken from the wild badgers (74% accuracy) and is comparable to the results achieved by gamma interferon methods (93.6% specific and 80.9% sensitive) (Dalley et al., 2008).

The *Mycobacterium paratuberculosis* dataset performed the best out of the other datasets, which were all recorded using the mk.2 SIFT-MS machine. The LOO accuracy reached 72% and found acetone as a possible biomarker.

The *Mannheimia haemolytica* dataset did not produce models that performed as well as the others (LOO accuracy of 69%) but a number of m/z ratios (including 65 and 83) showed a marked increase during the middle of the trial when the disease would have been most severe.

The brucella samples produced the weakest models, only achieving a LOO accuracy of 66%. One possibility that may account for this was the large number of unique differences between each sample as they came from the widest area over the longest period of time of all the datasets. Brucella also has range of tests available, the most powerful diagnostic being an ELISA test, having greater than 98% accuracy (England et al., 2004).

Finally, the *Mycoplasma bovis* dataset failed to produce a useful model and also didn't lead to the identification of any markers. This however, was expected due to the lack of visible symptoms reported by the clinicians. On the one hand this was a disappointing result, but it shows that if that if there is nothing to find, multivariate analysis can confirm that there is indeed nothing to be found.

The method of data analysis employed works best with plenty of samples but having a clear structure to those samples that allows analysis has also been shown to be important. The design of the experiment should, ideally, be chosen with the data analysis method in mind. Unfortunately, most of the samples analysed here were collected for purposes other than the investigation of multivariate SIFT-MS

techniques and so the datasets were large enough to give good statistically viable conclusions but they are not large enough to divide up and still have the same statistically sound conclusions formed about the 20% of the samples that would have made up a testing set.

The two best results must be the Bovine TB (*Mycobacterium bovis*) results where a high discriminatory model was found and the *Mannheimia heamolytica* where several features were found that are possible biomarkers that require further investigation.

A simple investigation into the effects of combining the data was investigated. First, the best results came from the bovine TB group. Unfortunately, these are the only samples that were recorded using the mk.3 SIFT-MS machine and therefore could not be included in the combined dataset. The *Mycoplasma bovis* samples also could not be included due to the lack of any discriminating feature between any of the classes available. This left the possible samples coming from the brucella, the *Mannheamia haemolytica* and the *Mycobacterium paratuberculosis* dataset. When the initial analysis was performed the main feature visible was the massive variance in ammonia levels.

It was a little surprising that ammonia, which was found to be helpful in the detection of Bovine TB in badgers in chapter 3, was no longer helpful in the cattle investigations. This is probably linked to the digestive processes occurring in cattle contributing a much larger ammonia variance than the disease in question.

Serum samples, whilst one of the easiest to obtain and handle, are probably not the best suited to the investigation of some of these diseases (Perbellini et al., 2003). A good biological reason is required in the choice of samples (e.g. serum is probably the best for studying blood infections and leukaemia) but intestinal infections such as *Mycobacterium paratuberculosis* might be better studied with more appropriate samples. Breath samples have been found to give good results despite transient levels (Smith et al., 2008) of volatiles in human studies but unfortunately are near

impossible to collect from animals. The next chapter looks at whether urine samples would be a good alternative.

5. Application of the methods to the classification of disease status from Urine collected from Cattle

5.1. Introduction

The main purpose of the work described in this chapter was to investigate whether multivariate methods applied to SIFT-MS can cope with different types of biological samples. The previous sections have all used data from serum samples. This section looks at SIFT-MS data from urine samples. The concentrations of volatiles in urine are less often studied than the volatiles in the headspace above serum and also that of breath (Smith et al., 1999; Perbellini et al., 2003; Pasikanti et al., 2008; Matsumura et al., 2010). Urine is also likely to have a very different composition to both breath and serum. The lungs create a large breathe-to-blood interface where VOC and other chemicals are allowed to transfer from one to the other via diffusion which creates a biological reason for their similarity. The kidneys are the blood to urine interface, however, the urine filtration process is selective and driven so the chemical composition of urine is likely to be different.

5.2. Data Collection & Measurement

These are the final datasets to be analysed and contained two sample sets with two different diseases. The first one analysed was the *Mycobacterium paratuberculosis* cattle urine samples from the FLI. The second is the *Mycobacterium bovis* cattle urine samples from the VLA. Both sets are from different types of trials (field samples and a clinical trial) and were collected for very different reasons.

Both sets of samples were subjected to the same SIFT-MS analysis processes as discussed in previous chapters. Both sets were measured using the newer mk.3 SIFT-MS machine. Urine samples are more difficult to obtain than serum. The process involved clinicians waiting with a bucket and hoping that the animal performed.

The urine samples were received frozen. Slightly higher volumes were received so instead of 1.6ml being injected 2ml was injected into the 700ml Nalophan bags. Again, the headspace was analysed with the electronic nose (2/3rd) and then the SIFT-MS (1/3rd). A five minute switching scan was performed. For a more detailed analysis of the methodological choices please see (Knobloch, 2010).

5.2.1. Mycobacterium paratuberculosis

The *Mycobacterium paratuberculosis* urine samples were collected in parallel with the serum sample from the same animals as in Chapter 4. Unfortunately, they were recorded using a different SIFT-MS machine so any improvement in accuracies could be due to either the improved machine or the samples. They belonged to animals from three farms in which the animals from two of the farms were confirmed as being infected with the disease and the final farm was confirmed as negatives and hence could be used as control samples. A total of 29 samples of urine were successfully recorded, 9 from farm 1 (positives), 10 from farm 2 (positives) and 10 from farm 3 (negatives). Unlike previously all samples were delivered at the same time and they were all measured in a random and blinded order.

5.2.2. Bovine TB (*Mycobacterium* bovis) data

The bovine TB (*Mycobacterium bovis*) urine samples were collected in parallel with the serum samples from the previous chapter however only one team collected the extra urine samples so the data set is much smaller than that of the serum samples. Further to this only the animals belonging to the control and the BCG immunisation (which is known to be an ineffective vaccination) group had their urine collected as it was specifically collected for this project.

The data comes from 10 animals of which 5 are control and 5 are BCG vaccinated. Each animal has four samples taken: 1 before vaccination, 1 before infection and then 2 after infection and the samples are taken at two week periods. Unfortunately one sample was lost and hence there are 39 samples.

5.3. Data Analysis

5.3.1. Mycobacterium paratuberculosis Results

Following the protocol established in the previous chapters, the first step of analysis was to remove the m/z associated with the precursors by zeroing the whole channel. Then the simplest method of displaying this two class dataset is to show the mean counts for each sample group.

The first thing to note from the above Figure 5.1 is that these samples were recorded using the newer mk.3 SIFT-MS machine so there is a larger range of m/z peaks and they are generally higher in count rate than that of the older mk.2 instrument. The second thing to note is that the height of the peaks is greater than that of the serum samples recorded using the older machine. Both of these can be associated with the improved detector or the change in sample source.

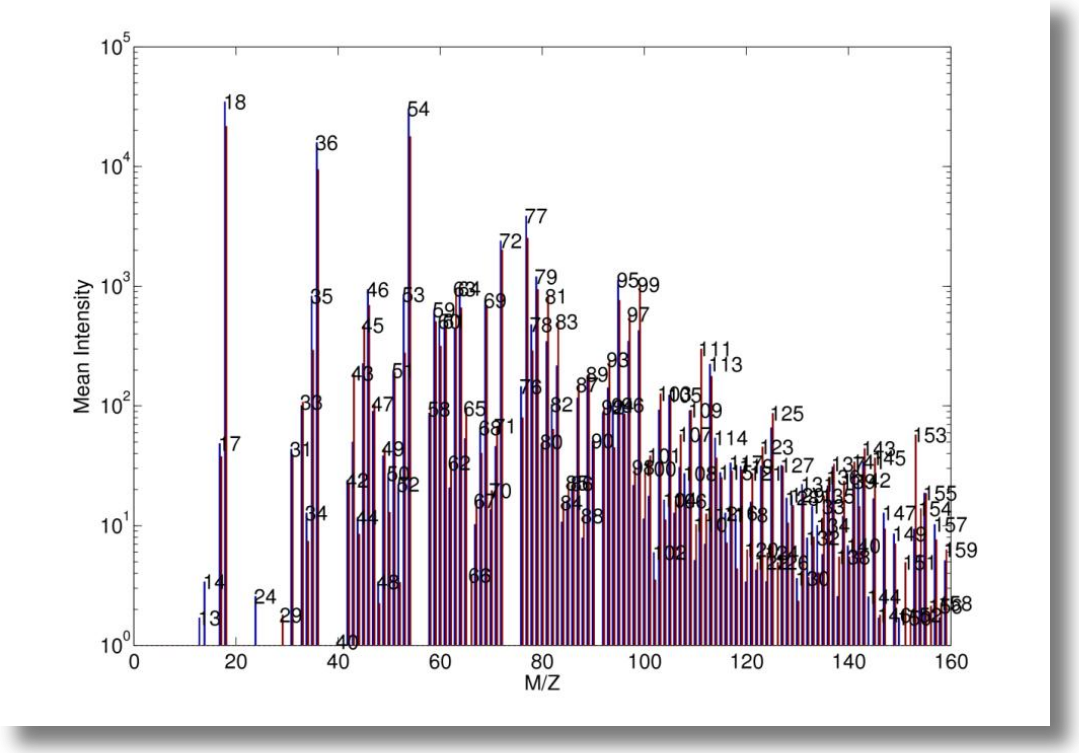


Figure 5.1 – The average H_3O^+ spectra for *Mycobacterium paratuberculosis* urine samples as recorded using the mk.3 SIFT-MS machine. Counts associated with the precursors have been zeroed. The blue spectrum is the average of the negative samples and the red is the average of the positive samples.

There are some large differences between the means of each sample class. Whether these are significant or not is still to be determined. The other spectra show similar features to the H_3O^+ spectra hence are not shown. The largest peaks (on the H_3O^+ spectra 18, 36 and 54) are all associated with ammonia. Previously this has been an unreliable marker in cattle due to the large random variance (probably caused by naturally occurring variations that would be seen in healthy animals). The usefulness of these average plots has decreased as the number of peaks being shown has increased.

Figure 5.2 shows the peaks for ammonia (m/z 18) for all the *Mycobacterium paratuberculosis* samples. The large variance in ammonia levels is most likely to be associated with intestinal bacteria levels and general animal hydration. This is a large

natural variance (i.e. a variance that would be present in samples from a group of healthy animals). This means that any variance caused by the disease state alone is completely hidden by this variance. The other m/z peaks associated with ammonia (36, 54 and 72) show similar patterns and the same large variance with no apparent link to disease classification.

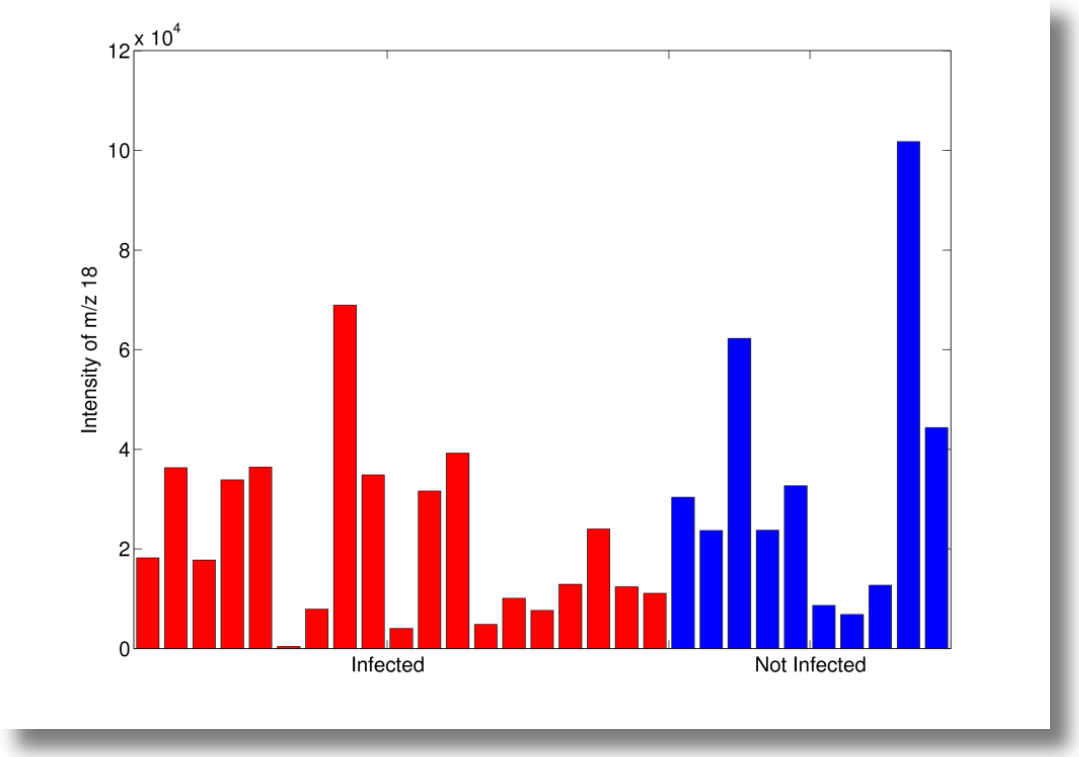


Figure 5.2 - The m/z counts for ammonia (18) for all infected (red) and all healthy (blue) *Mycobacterium paratuberculosis* urine samples as recorded using the H_3O^+ precursor of the mk.3 SIFT-MS machine showing the large variance

Moving on to PCA analysis, ammonia is the largest contributor to the top three principal components and the corresponding scores plot shows no discrimination between the sample types.

Figure 5.3 and Figure 5.4 show that the variance for ammonia clearly contributed one of the largest portions of variance after that of the precursors. The PCA score plot

does not show any discrimination which would appear to link with the hypothesis that the variance is not linked to the disease state.

Using this same combination of PCA analysis and investigating m/z ratios individually the following m/z are found to have large variances that are clouded by unlinked large variance; ammonia (m/z 18,36,54,72), acetone (59,77,95), propanol (79) and some others (35,46,53,64). When these have been removed the following PCA scores plot (Figure 5.5) can be produced.

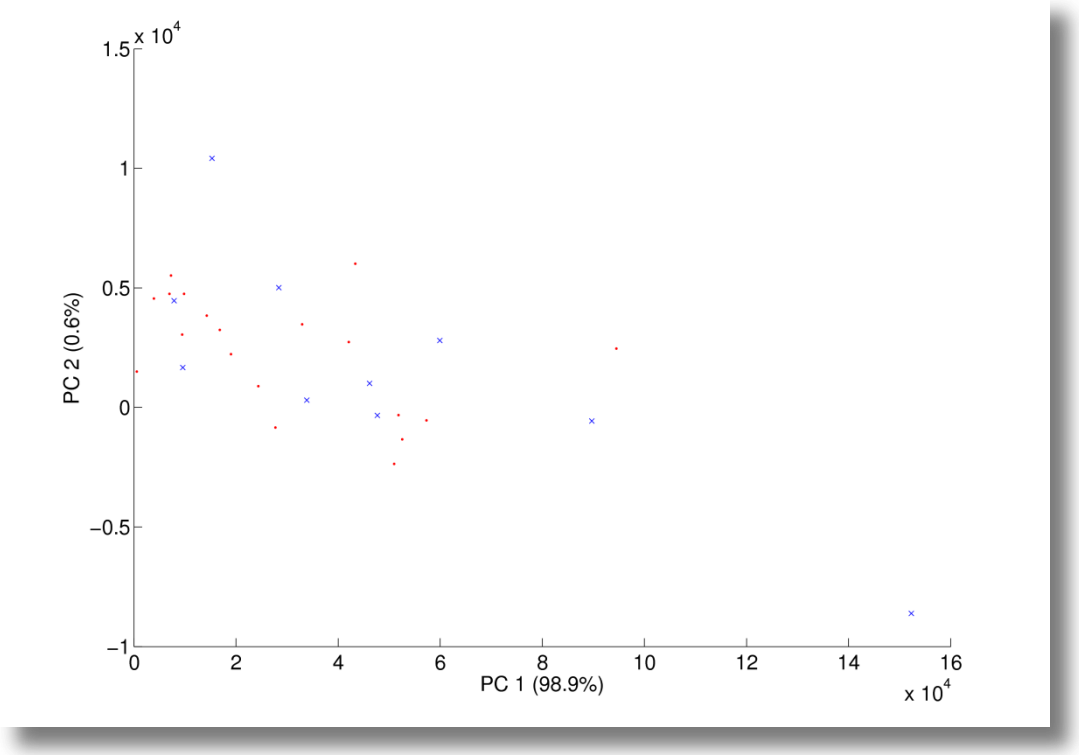


Figure 5.3 - PCA scores plot showing that no discrimination can be seen between the sample types when all the data from the H_3O^+ precursor spectra of the mk.3 SIFT-MS machine is used. No pre-treatment methods were performed but m/z ratios associated with the precursors have been removed. Blue crosses are *Mycobacterium paratuberculosis* negative samples and red dots are *Mycobacterium paratuberculosis* positive samples

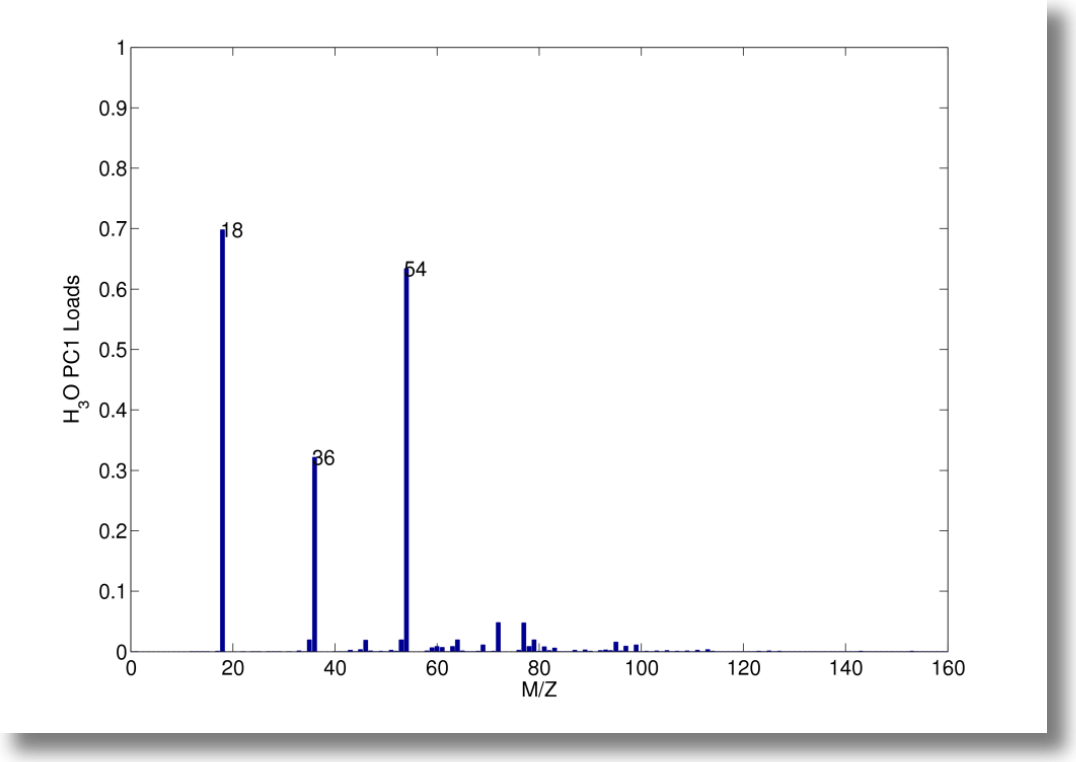


Figure 5.4 - The loadings of PC1 for the above PCA analysis using the H_3O^+ precursor spectra from the mk.3 SIFT-MS machine. No pre-treatment methods were used but the m/z ratios associated with the precursor ions were zeroed. m/z peaks associated with ammonia (18, 36 and 54) clearly dominate the variance.

In the below PCA scores plot (Figure 5.5) reasonable discrimination can be seen between the sample classes and this is associated with m/z ratios 45, 63, 81 and 99 which are all linked to acetaldehyde.

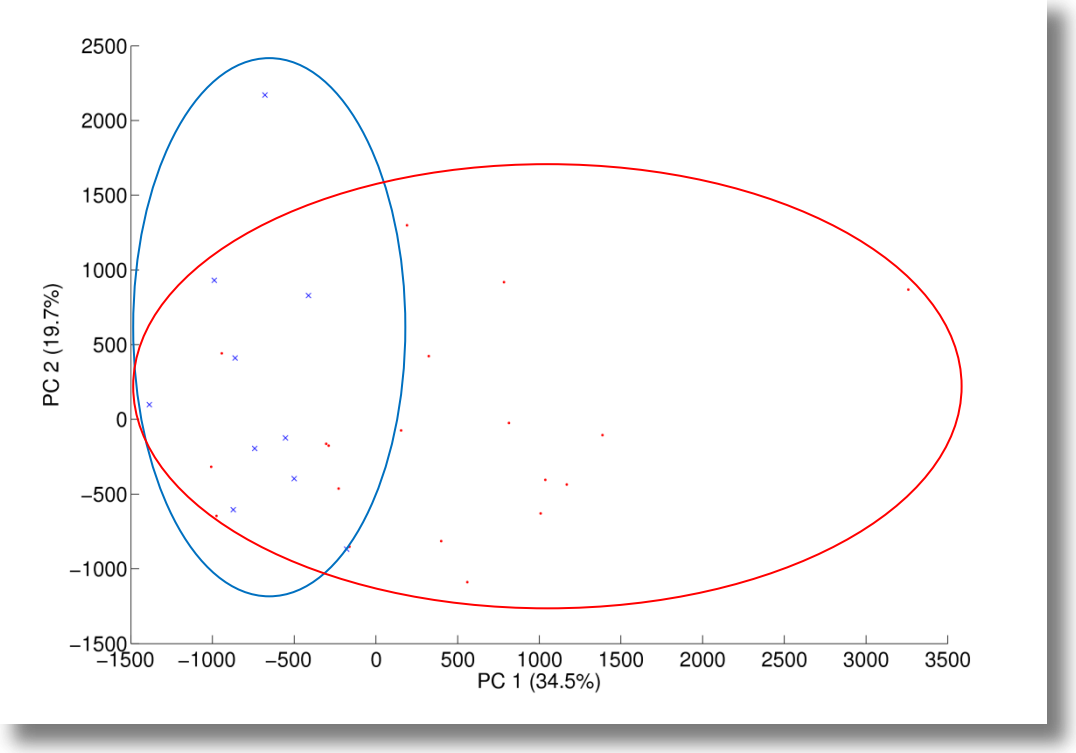


Figure 5.5 - PCA scores plot of the H_3O^+ urine spectra from cattle taken from herds in three farms. Red dots are *Mycobacterium paratuberculosis* positive samples and the blue crosses are the negative control group. Peaks with m/z associated with ammonia (18,36,54,72), acetone (59,77,95), propanol (79) and some others (35,46,53,64) have been removed. m/z peaks associated with the precursors have been removed and the data was pre-treated using a mean centre algorithm.

PLSDA doesn't perform well using all three precursors together. A similar lack of discrimination is seen with the just H_3O^+ data and the LOO score doesn't rise much above random chance until 12 LV's have been used. It is not until the sample removals as for the PCA above does PLSDA give good results (Figure 5.6).

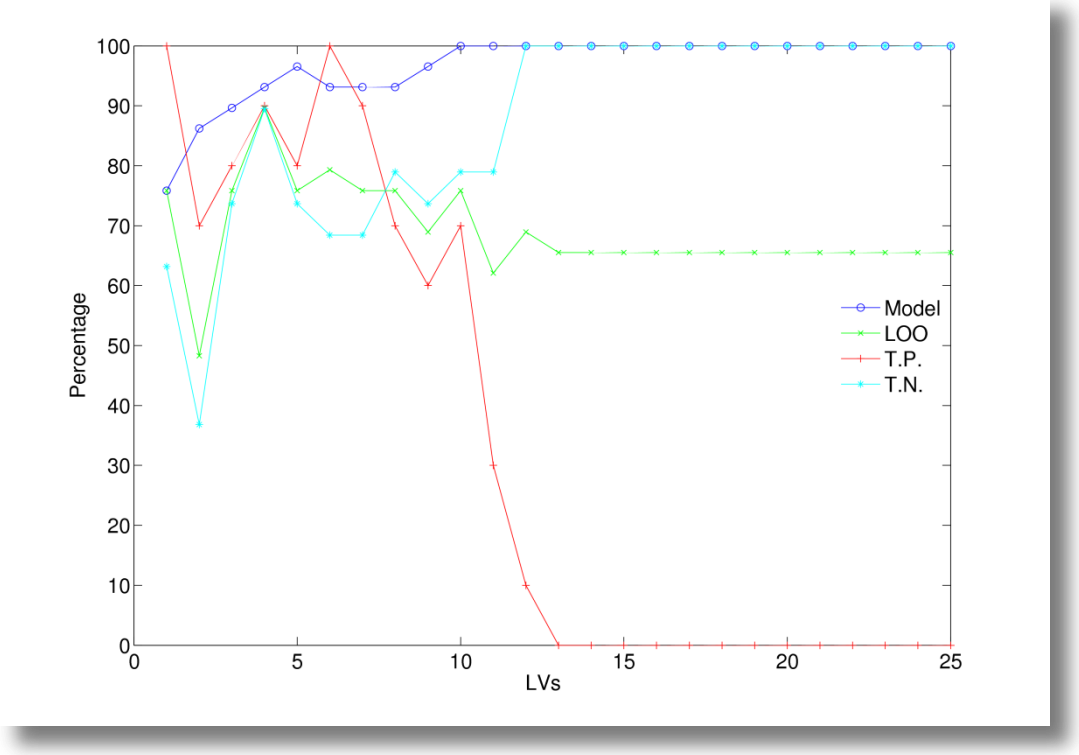


Figure 5.6 - PLSDA curves for the *Mycobacterium paratuberculosis* urine samples as recorded using the newer mk.3 SIFT-MS. Only data from the H_3O^+ precursor channel is being used. m/z associated with ammonia (18,36,54,72), acetone (59,77,95), propanol (79) and some others (35,46,53,64) have been removed. The blue line shows the percentage of correctly identified samples when all samples are used to build the model, the green line shows the percentage of correctly identified samples using a LOO validation process and these results are broken down into true positives (light blue) and true negatives (red).

The initial high LOO accuracy (75% with 1 latent variable) is probably linked to high levels of acetaldehyde as found using PCA. The model LOO accuracy then suddenly drops (51% accuracy with 2 latent variables) before recovering (88% accuracy with 4 latent variables). A 88% accuracy is a surprisingly good result for such a small dataset and indeed questionable (Brereton, 2006). Datasets of this size are known to cause overtraining in PLSDA models (Goodacre et al., 2004; Brereton, 2006). Unfortunately, as there are so few samples it would be hard to expand the analysis to investigate the accuracy of prediction.

Considering this problem further, the samples came from three separate farms. It would not take much for the animals in one farm (where the control samples are from) to have a significant difference to the others because of diet or location. Therefore, it is impossible to say with any certainty that this method is detecting the disease. The method of SIFT-MS and multivariate techniques appears to be working as it is detecting differences between the sample groups even though the differences might not be disease related. Better experiment design is essential and a model built that can detect infection created before testing on field samples is a likely first step.

5.3.2. Bovine TB (*Mycobacterium* bovis) results

The second disease under investigation in this chapter was bovine TB (*Mycobacterium bovis*). The data was from a clinical trial where all the animals were treated in exactly the same manner so the location issue seen with the *Mycobacterium paratuberculosis* samples did not apply. Urine samples were only taken at the VLA clinic and not the second clinic (IAH) as they were with the cattle serum samples for this experiment. Also, as the trial was looking at the effectiveness of some new immunisations (because the BCG vaccine that is useful in the fight against human form TB is ineffective in cattle) only the control samples and the BCG were sent. The immunised samples would only confuse the issue and were not analysed.

As with the *Mycobacterium paratuberculosis* urine samples, the *Mycobacterium bovis* urine samples have a much larger number of volatiles than their serum counterparts. This can be seen in the average of their peaks (Figure 5.7). With the bovine TB data, both sets were recorded using the newer mk.3 SIFT-MS machine so this increase indicates more volatiles present in the headspace above the sample.

There appears to be some subtle differences in the means of the count of the two groups of samples in Figure 5.7.

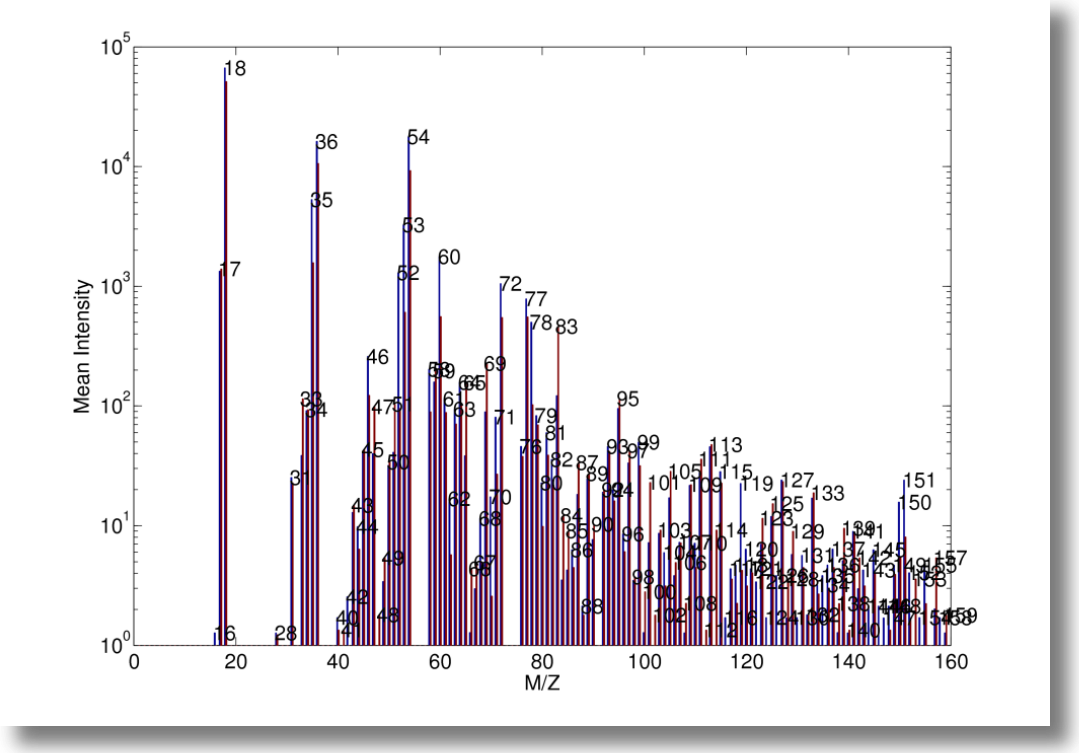


Figure 5.7 - The average counts for bovine TB (*Mycobacterium* bovis) urine samples for before infection (blue) and after infection (red) as recorded using the H_3O^+ precursor spectra of the mk.3 SIFT-MS machine. m/z peaks associated with the precursors have been removed.

As with the *Mycobacterium paratuberculosis* samples the initial PCA analysis results show little promise (Figure 5.8). The variance contributed by ammonia is again very large but has no contribution to separating the sample types. Again, as was spotted with the serum samples from this same dataset (Chapter 4) there are differences between the animals that are vaccinated and the control group. This is a discrimination that is not useful to the aim of this thesis, which is to detect infection status.

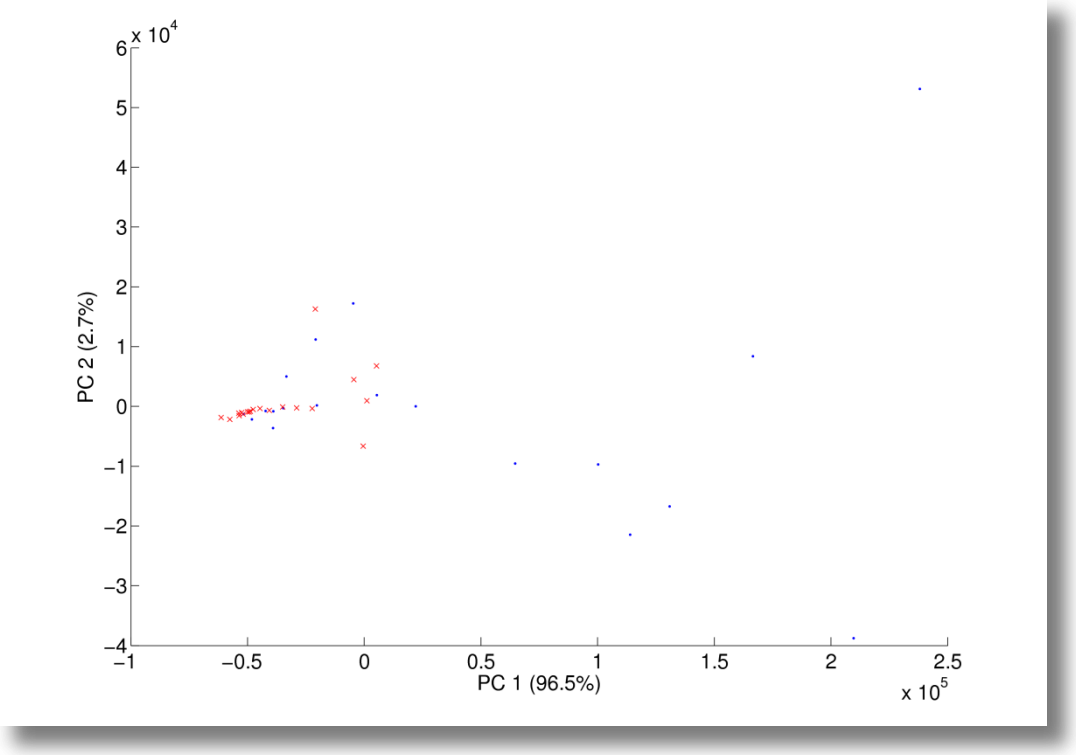


Figure 5.8 - PCA scores plot of the Bovine TB urine samples as analysed using the H_3O^+ precursor channel of the mk.3 SIFT-MS machine. Blue dots are control samples and red crosses are BCG samples.

Another problem with this dataset is the apparent number of spikes as shown by the example of acetone in the above figure (Figure 5.9). Sometimes these spikes are genuine levels of VOCs as not only the primary m/z ratio is affected but also those of water clusters. This was seen with the Brucella serum dataset where large peaks were appearing on only one sample and this was causing large variance to be detected that is not linked to the disease status and is unrepeatable and hence unusable as a discriminatory variable. Unfortunately, this problem is not limited to only one sample with this data set. There are other m/z ratios associated with other VOCs that have similar features to the above but for different samples.

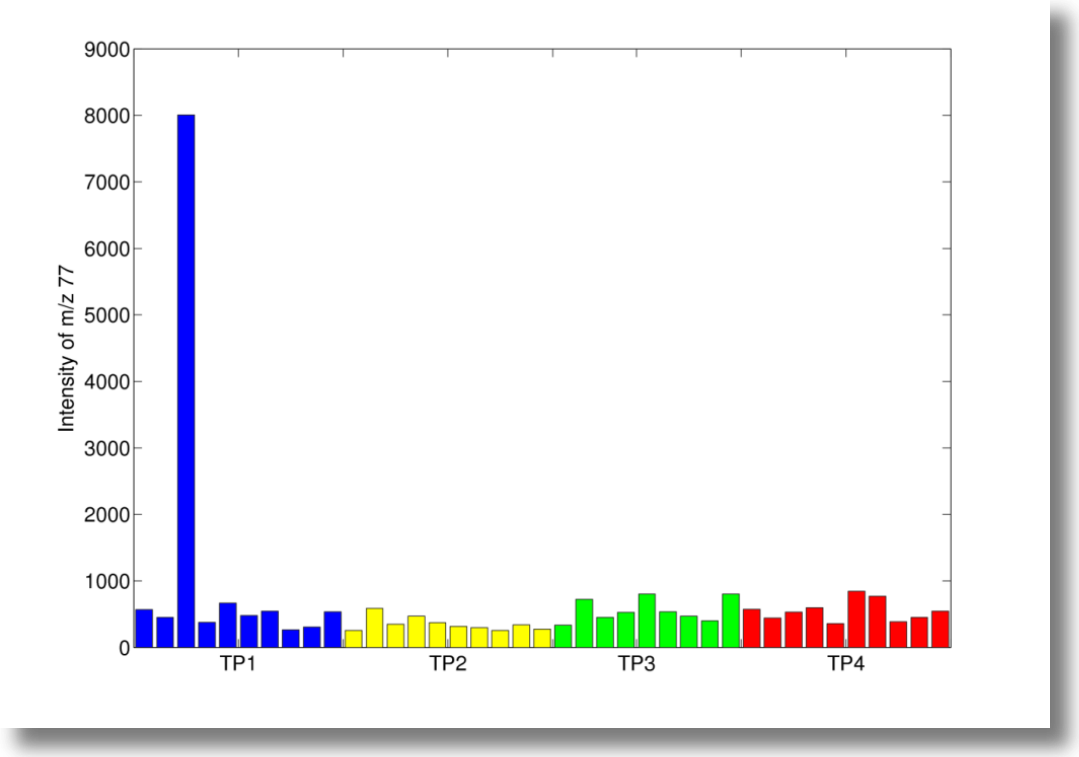


Figure 5.9 - The counts of m/z 77 (acetone) for urine samples from the bovine TB (*Mycobacterium bovis*) trial as recorded by the H_3O^+ precursor of the mk.3 SIFT-MS machine. Samples are across the horizontal axis separated into the different time points of the trial

It was not until m/z associated with ammonia (17, 18, 35, 36, 53, 54, 72), acetone (77), and some unknowns (52, 58, 60, 78) had been removed that discrimination on any scale was seen on the PCA scores plot.

Some discrimination can been seen on the PC1 axis of the PCA scores plot (Figure 5.10). Most of the negative samples are clustered to the very left of the axis with the rest of the samples (the positive and five negatives) being spread out along the rest of the axis.

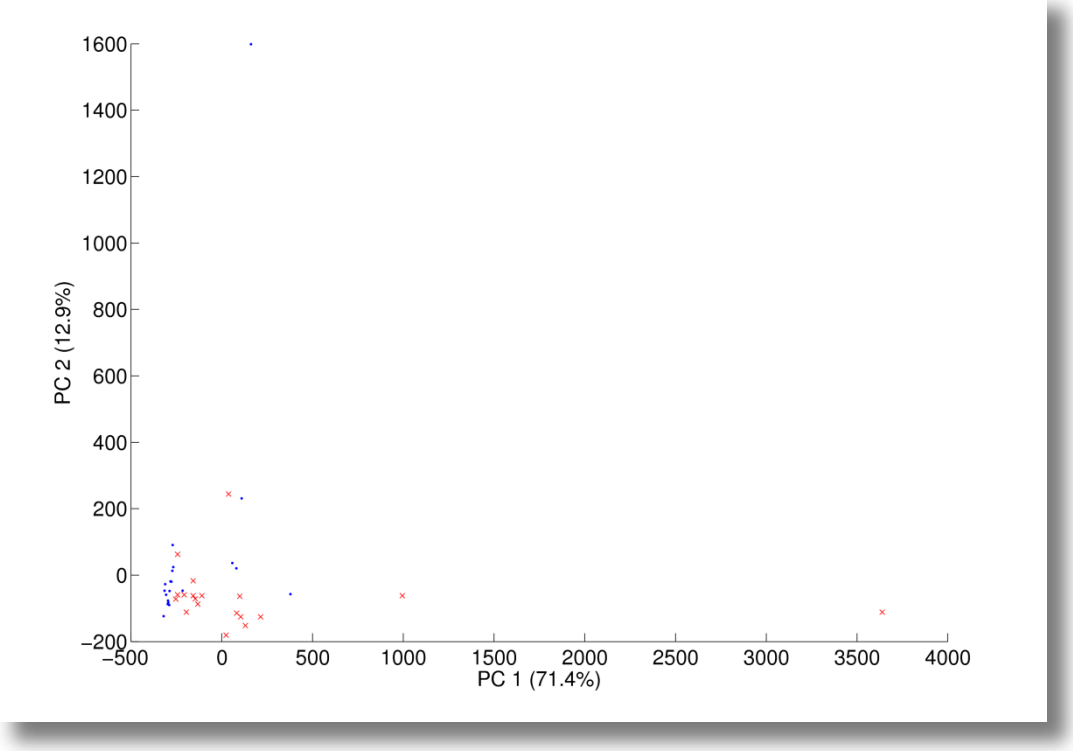


Figure 5.10 - PCA scores plot of the cattle urine bovine TB (*Mycobacterium bovis*) samples as recorded using the H_3O^+ precursor channel with m/z associated with precursors and ammonia (17, 18, 35, 36, 53, 54, 72), acetone (77), and some others (52, 58, 60, 78) removed. Red x are post-infection and blue dot are pre-infection. The data was pre-treated using a mean centring algorithm.

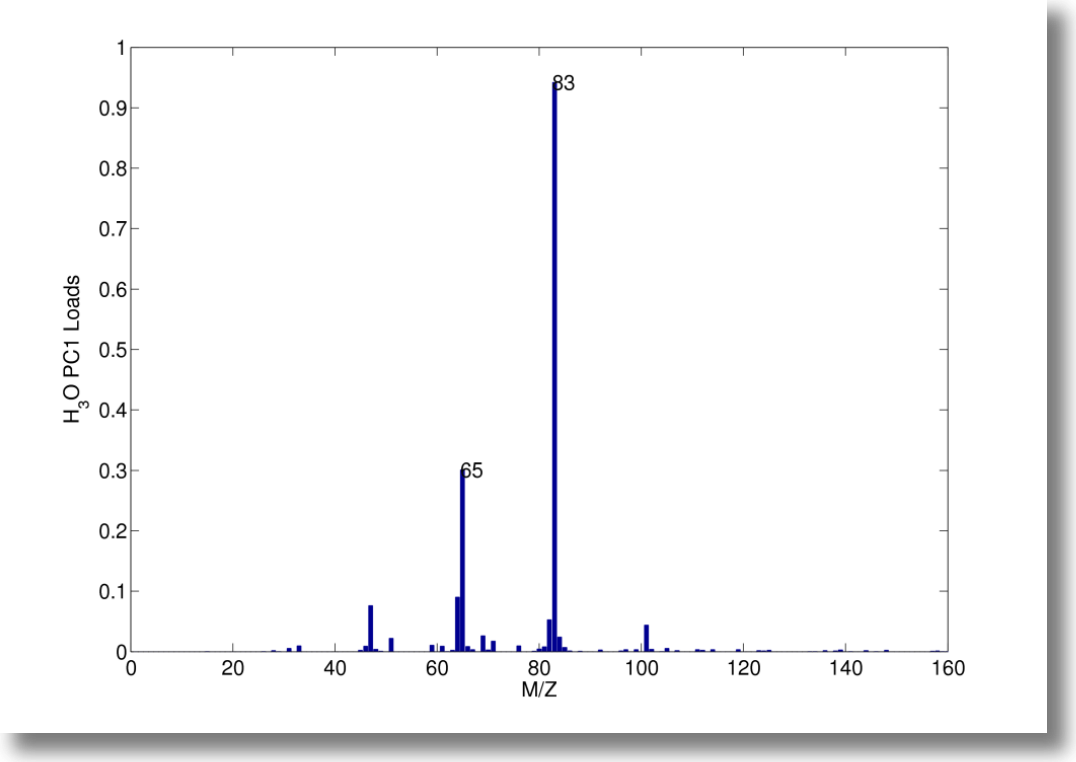


Figure 5.11 - The loadings of PC1 for the above PCA analysis of the cattle urine bovine TB samples as recorded using the H3O+ precursor channel with m/z associated with the precursors, ammonia (17, 18, 35, 36, 53, 54, 72), acetone (77) and some others (52, 58, 60, 78) removed.

Investigating the loadings (Figure 5.11), it can be seen that the discrimination seen on the scores plot (Figure 5.10) is caused by a combination of m/z 65 and 83 which are both associated with ethanol.

Ethanol was previously discounted in the badger TB trial as it was highly varied, possibly due to diet. In this instance, as the animals were all kept in an animal house during the trial the diets of all animals will be the same so ethanol may be a possible marker. Whether it would be an effective marker in field samples is another matter.

The PLSDA results (Figure 5.12) show that ethanol is an efficient discriminator (the first two LV's are associated with m/z 83 and 65 respectively). m/z 83 doesn't appear to be as useful as m/z 65 and indeed when looking at the univariate plots this is very

apparent (not shown). Removing m/z 83 however brings the highest accuracy to LV 1 but it also decreases the LOO accuracy to 85%. This suggests that the combination of them both is more effective than either one on their own. Looking further into the noise there is also another improvement when the sixth LV is added. This suggests that there may possibly be some further discriminators hidden in the data.

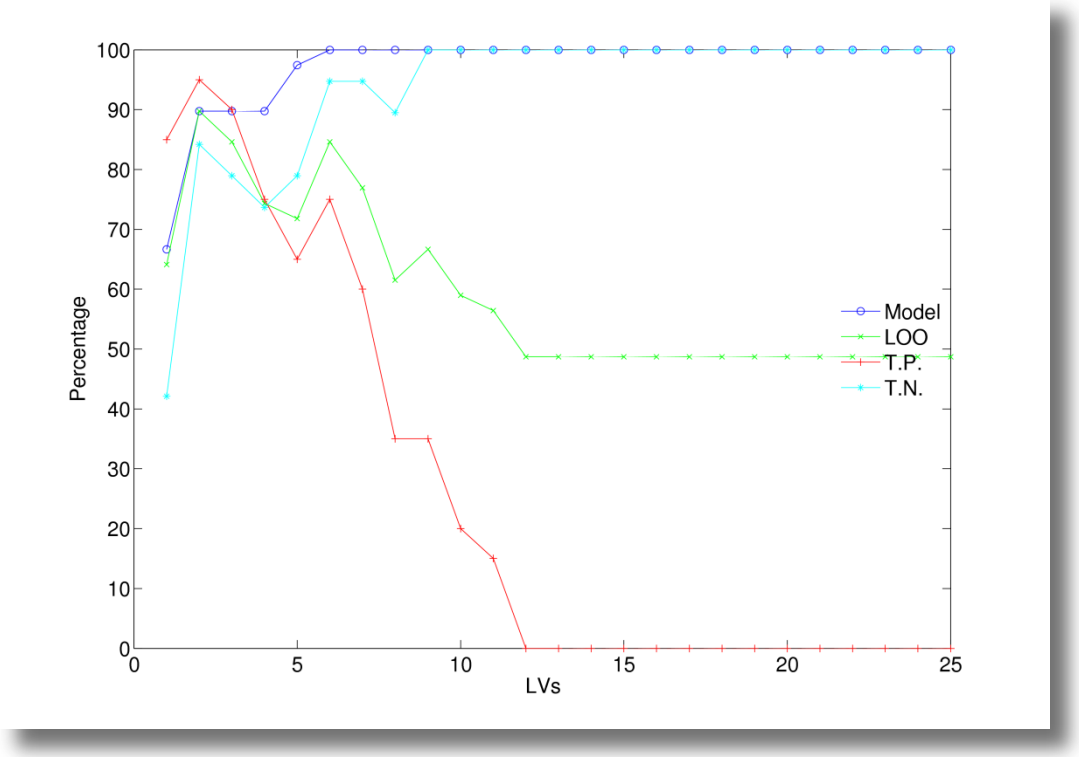


Figure 5.12 - PLSDA LOO analysis of the cattle urine bovine TB (*Mycobacterium bovis*) samples as recorded using the H_3O^+ precursor channel with m/z associated with precursors and ammonia (17, 18, 35, 36, 53, 54, 72), acetone (77), and some others (52, 58, 60, 78) removed. The blue line is the model accuracy and the green line is the LOO validation accuracy which is then further broken down into the LOO true negatives (red line) and true positives (light blue line)

Further analysis could be interesting; however the limited size of the data set restricts a thorough analysis and validation. The bootstrapping sample investigation in Chapter 3 found that more than 120 samples were required before a model's LOO accuracy

increased when dealing with badger serum samples. The similar complexity of this data suggests similar numbers would be required for a repeatable result.

5.4. Conclusions

Whilst this chapter has not conclusively proved that analysis of urine samples using a combination of SIFT-MS and multivariate analysis method is effective for disease diagnosis it has at least suggested that it may be so. The results of PLSDA show high LOO accuracies but due to the limited dataset sizes they cannot be properly validated. Both datasets have m/z peaks that might be linked to biomarkers and are worth investigating further. Also LOO accuracies of 88% for the *Mycobacterium paratuberculosis* and 90% for the TB are promising results (de la Rua-Domenech et al., 2006). Whilst the LOO validation step creates a quotable result, prediction accuracies are necessary for the proof. The complexity of urine headspace is such that it has a higher dimensionality than that of the samples created in the lab or from serum and it is such that more samples would be needed to effectively build models and validate them (Brereton, 2006).

Also, both diseases showed promise when being subjected to PLSDA optimisation, achieving high LOO accuracies. This suggests that there might be detectable changes in the composition of the urine headspace that could be used as a method for disease detection as opposed to just biomarker discovery. Unfortunately, there are also other features, other than the disease status, that could be associated.

Using urine instead of breath is not a novel idea (Perbellini et al., 2003; Matsumura et al., 2010). It is an area of increased research and has shown promising results. Lung cancer can be detected using breath volatiles (Kischkel et al., 2010), but problems can arise when collecting samples which stops animal trials. (Matsumura et al., 2010) showed that urine is an alternative sample using mice trials. Cancers could be detected in mice using GC-MS. Collecting urine is easier than breath but possibly not easier than serum, however, the increased LOO accuracies suggest that it is an area for more investigation.

With the *Mycobacterium paratuberculosis* samples, the positive and negative samples came from different farms. There are, therefore, a large number of factors that could be contributing to the differences caused by completely different environmental conditions. It is generally not good practice to have datasets that have this restriction and a range of sources for both the disease group and control group would have been better.

With the bovine TB (*Mycobacterium bovis*) samples, there is no such environmental change. All the animals were kept in controlled conditions and fed the same foods. But it was a time based trial and so there is four weeks of difference between the first set of pre-infection samples and the final set of post-infection samples. The samples could be exhibiting natural changes in body chemistry instead of effects of the disease in question. Also, the PLSDA results were found to be linked to ethanol which was discounted in the badger field study discussed in chapter 3. It was suggested that ethanol levels were too heavily linked with diet to be used as a marker. As this was a clinical trial the animals' diet would have been the same so the detection protocol developed might not be transferable to field animals.

This chapter has highlighted the need for a well designed trial that takes into account the final analysis procedures. Problems should not arise such as the one for the *Mycobacterium paratuberculosis* samples where all the negative samples were from one farm and the positives from two other farms. Any conclusions formed from this dataset could be attributed to environmental factors other than the disease status itself. These issues should be taken into account when planning any future studies.

6. Discussion and Conclusions

Until now, analysis of data from SIFT-MS has generally been performed by hand by an expert. In this thesis it has been shown that this is not the only way to analyse the results of SIFT-MS analysis, and that superior results can be obtained via other methods for certain applications. The aim of this chapter is to summarise these results and make recommendations to anyone with an interest in taking the work further or utilising the procedures for their own research.

6.1. Summary of Work Conducted

In Chapter 2, a proof of concept trial showed that multivariate analysis procedures are, in ideal conditions, suited to help in the analysis of SIFT-MS full spectra scan data. Two different mixtures were analysed using the SIFT-MS machines. They were created from VOCs that have been previously found during SIFT-MS headspace analysis of biological samples (Smith and Spanel, 2005; Turner et al., 2006a; Turner et al., 2006b; Turner et al., 2006c). Using PCA analysis and investigating both the scores and loadings plots showed that not only were the multivariate techniques able to identify the m/z ratios that contribute to significant differences between the two mixtures, the m/z ratios were themselves linked to the VOCs that were combined in the creation of the mixtures. This is the first suggestion that SIFT-MS analysis and multivariate techniques are suited to biomarker discovery.

The multivariate process was extended by the use of the supervised analysis procedure PLSDA. This produced a classification model that could with 100% accuracy detect which of the two mixtures any unknown sample belonged to. This trial was only a proof of concept but it suggested that multivariate techniques and SIFT-MS could be used as a possible rapid disease detection method.

In Chapter 3, the knowledge obtained from the proof of concept trial was refined during application of the method to a group of real biological samples. The samples in question were 245 serum samples collected from wild badgers in a study to

investigate bovine TB (*Mycobacterium bovis*) infection within their population. PCA analysis identified m/z 18 (ammonia) as a possible marker for infection. Further investigation would be required to investigate the causes of this increase in ammonia to confirm the link to TB infection status. PLSDA produced models that were able to classify TB status with an accuracy of 67% ± 6% using 90% of the samples to build the PLSDA model and 10% to validate the model. The process was bootstrapped 10,000 times to give accurate results and to remove the possibility that chance was impacting the final result a null hypothesis model was built and this had an accuracy of 46% ± 6%.

The number of samples required to build PLSDA models was investigated by comparing the LOO accuracies of models built using only a fraction of the available samples. 120 samples were required to build a model that was stable i.e. with fewer samples than this threshold models were built with similar accuracy but the bootstrapping showed increased variance between the accuracy levels. Increasing the number of samples above 120 increased the accuracy recorded by the LOO validation of the model built until approximately 220 samples were being included to build the models. After this point, the model accuracy did not increase further with the addition of samples.

Chapter 4 built on the biomarker discovery ideas devolved in chapter 2 and investigated in chapter 3. Five diseases were studied from a range of sources including field studies similar to the bovine TB in badgers and also clinical trial samples where the animals are kept in controlled environments and have large doses of the pathogen associated with the disease being investigated introduced.

The best classification results came from the bovine TB serum samples taken as part of an immunisation trial, achieving an 85% LOO accuracy (86% true negative and 84% true positive) and also suggested that methanol was a possible marker. The *Mycobacterium paratuberculosis* dataset performs second best producing a LOO accuracy of 72%. The *Mannheimia haemolytica* dataset didn't produce models that performed as well as expected (LOO accuracy of 69% for a clinical trial where

fatalities occurred). This is most likely linked to a difficulty in creating a classification vector. There were not clear positive and negative samples as this was a disease progression trial. This was not a failing of the trial however, as a number of m/z ratios (including 65 and 83) showed a marked increase during the middle of the trial when the disease would have been most severe. These are possible markers that require further investigation.

The brucella dataset produced a model with a LOO accuracy of 66% and failed to identify any possible markers. The results are likely to be poor as this dataset contains samples from the least controlled demographic (Europe wide) and they were collected over the largest time frame.

Finally, the *Mycoplasma bovis* dataset failed to produce a model and also failed to identify any markers. This was in part expected, due to the failings of the clinical trial, but confirmed that the method is unlikely to find possible markers if none are to be found.

Finally, two of the above diseases were investigated again but using data collected from urine samples instead of serum. LOO accuracies of 88% for the *Mycobacterium paratuberculosis* and 90% for the bovine TB were achieved. The cattle urine samples have higher levels and a higher number of volatiles when compared with their serum counterparts. Whilst this may be in part due to the change in analysis machine from the mk.2 to the mk.3, with the *Mycobacterium paratuberculosis* dataset both serum and urine bovine TB sample sets were recorded using the mk.3 SIFT-MS machine. Also, the higher LOO accuracies suggest that urine is a better sample for discrimination. However, as both the datasets contained very few samples (29 samples in the *Mycobacterium paratuberculosis* dataset and 19 samples in the bovine TB dataset), neither of the model accuracies were properly validated.

SVMs have not been used as the number of variables involved is limited. A maximum of 600 variables are possible however, after initial investigations with the mk.2 SIFT-MS machine the number of actual variables is closer to 100. When we

start looking for distinct variables we are left with approximately 20, as many variables are linked either across the three precursor channels or are simply with the addition of water. However, if a significantly larger dataset was created using urine samples recorded on the mk.3 SIFT-MS the increase in the number of variables by the two changes would be an ideal chance to test the accuracy of SVMs compared to PLSDA.

Data recorded using the newer mk.3 SIFT-MS machine appears to produce much higher LOO accuracies than that of the older mk.2 SIFT-MS machine. None of the results are directly comparable between the machines other than the simple two mixture proof of concept trial detailed in chapter 2. Since the accuracies for both machines were 100% in this trial there is little to be gained by comparing the results.

6.2. Proposed methodology

Combining the knowledge gained from all the investigations in this thesis an analysis procedure for SIFT-MS data from veterinary diagnostic samples can be formulated. Based on the results of these investigations it is recommended that the following generic analysis procedure has been formulated.

i. The proprietary software supplied with the machines needs to be used to export the data into a "database" format. This produces a table with a single row for every spectra recorded. The easiest method of manipulation is to use a spreadsheet program to combine the three spectra into one row vector and remove repeating SIFT-MS machine settings and adding a column contain the discrimination factors (these are usually sent in Excel files). Data is then imported into Matlab either via "xlsread" or "csvread". Before analysis the m/z ratios associated with the precursors (Table 2.2) need to be removed.
ii. For simple spectrum plots a log scale is advisable.
iii. For comparison of peak high for the sample m/z but different samples the log scale is not required.

iv. The data should then be treated with a mean centre algorithm (Eq. 2) in preparation for multivariate analysis.

v. PCA analysis should begin with analysis of just the H_3O^+ precursor data. PCA analysis works with all three precursors combined but to effectively and easily analyse the loading values to search for markers it is much easier with just the H_3O^+ precursor data.

vi. PLSDA LOO analysis should then be performed to investigate the possibilities of disease discrimination and classification.

vii. If PLSDA LOO gives good results and there are enough samples (greater than 100) and investigation into the numbers of samples required to build a model should be investigated by modifying the number of samples used in the LOO PLSDA optimisation.

viii. PLSDA can then be validated by separating into model and test sets (randomly selected samples to allow for bootstrapping) and accuracies achieved.

ix. Finally, the null hypothesis (classification is due to chance) needs to be rejected using a randomised classification vector and rebuilding of the PLSDA models using the same training testing ratios.

6.3. Overall Conclusions

The aims and objectives outlined in chapter 1 have all been achieved in the course of this work. Specifically:

- Multivariate methods including unsupervised (PCA) and supervised (PLSDA) techniques have been evaluated and shown to be an effective tool to help with biomarker discovery and they also show the possibly of predictive modelling for disease discrimination.
- A range of applicable scaling techniques were trialled and the mean centre algorithm was found to be the most generally effective method with the biological dataset being used with both the mk.2 and mk.3 SIFT-MS

machines. The m/z ratios associated with the precursors always need to be removed and other m/z ratios might need removing depending on data source.

- The methodology was applicable to the analysis of badger serum samples and identified ammonia as a possible marker. The technique is not yet advanced enough to be a competitive disease detection tool, however an accuracy of 67%±6% is better than random chance and also has the possibility for near real time detection once an accurate model has been developed.
- The results for cattle serum were similar to that of the badger serum, however, the clinical trial data produced much higher LOO accuracies. PCA analysis also showed possible markers that appeared to follow the expected disease progression.
- The methodology also performed excellently when applied to cattle urine samples.
- A general methodology has been outlined. This is a general procedure and other steps can be added removed. Different analysis tools could be used (for example SVM instead of PLSDA). Also, pre-treatment using mean centring works in this instance but another method might be applicable in other circumstances depending on the data source and the differences expected between sample groups.
- The findings have been discussed in each chapter. Some overall discussions between the different sample types and comparisons of samples type have been made.

6.4. Discussion of the performance in context of other detection methods

TB discrimination of 67% was achieved with serum samples collected from wild badgers, 85% from serum samples collected from a clinical trial and 90% from urine samples from the same trial. Whilst these methods are not as good as the experimental gamma interferon method, which had a 93.6% specificity and 80.9% sensitivity (Dalley et al., 2008), they are an improvement over the doped detection

that proved the concept of using volatile to detect TB using electronic nose (Fend et al., 2005).

Comparing the other diseases is a little more difficult, both *Mannhemia haemolytica* and *Mycoplasma bovis* do not have comparable techniques (Ackermann and Brogden, 2000; Nicholas and Ayling, 2003). The main problem is that they are part of a group of diseases that have similar physical symptoms and hence determining the cause is usually done with a genetic analysis to determine the exact strain (Snowder et al., 2006; Angen et al., 2009). However, the technique did uncover some volatiles that could be associated with the diseases.

Brucella has highly determined accuracies (Cutler et al., 2005). Indeed, multiple tests with accuracies in the high 90% range are usually combined (England et al., 2004). The problem is more one of improving this accuracy to minimise the cost of the few cases that are detected each year in import animals. It appears unlikely that SIFT-MS and multivariate techniques could replace the current tests for this disease.

Looking at the SIFT-MS technique of measuring and detecting volatiles and comparing it with similar technologies and methodologies, SIFT-MS is arguably better than electronic nose technology. There are many examples of electronic nose sensor problems in the literature (Vito et al., 2008; Knobloch et al., 2009c). Indeed many electronic nose technologies and sensor arrays are developed specifically to target the expected volatiles (D'Amico et al., 1998; Ciosek et al., 2006; Lebrun et al., 2008). Other examples used a short measuring time (i.e. the samples were recorded in the same day) which avoids some of the problems of sensor drift (Ali et al., 2003; Aleixandre et al., 2004).

The other comparable technology is GCMS. This is a well used and stable method for volatile detection. Headspace analysis is less common than sampling of other sample types (e.g. direct liquid sampling) but methods have been developed which are stable and accurate (Grote and Pawliszyn, 1997; Pasikanti et al., 2008). Indeed, GCMS is used as a validation tool for some of the results published using electronic noses that

have been developed to detect certain volatile features (Berna et al., 2008; Lebrun et al., 2008). One of the big downsides of GCMS is the length of time it takes from the point where the headspace is available for sampling to the final data being recorded. The sample preparation for headspace generally involves adsorption onto a tube, then desorption through the GCMS with a high definition sampling run (Grote and Pawliszyn, 1997; Harper, 2000; Pasikanti et al., 2008). Then there is the decision to make whether the sample is doped to allow quantitative measuring (Birkemeyer et al., 2005) or whether alternative methods are used to get quantifiable results (Wang et al., 2003). The sampling process is measured in tens of minutes and can take up to 3 hours. This is one of the reasons why electronic nose technologies are favoured over the more stable GCMS analysis. SIFT-MS sampling times (even for the most detailed analysis) do not take more than 15 minutes. This can be further realised as SIFT-MS monitoring of a few select m/z can perform real time analysis with sufficient accuracy to detect distinct differences between mouth, airway and lung breath samples (Smith et al., 2008).

The other problems come when analysing the data produced. Conducting polymer and MOSFET electronic noses can consist of a dozen or so sensors with five or more responses characteristics per sensor (Dickinson et al., 1998; Gopel et al., 1998; Walt et al., 1998; Eklov et al., 1999; Jurs et al., 2000; Gutierrez-Osuna, 2002; Pardo and Sberveglieri, 2002; James et al., 2005). The data analysis methods have been developed to make the best possible use of the data. GCMS hold more challenging data requirements. They continuously record similar mass spectra to the SIFT-MS (m/z ratios verse counts) but the time since the injection is also important (retention time). Most of the pattern detection methods use snapshot analysis. For example, PCA is considered an image compression technique (Otto, 1999a). GCMS data, because of the time element, is more comparable to a video. SIFT-MS is more comparable with the data analysis requirement of electronic nose.

6.5. Further Work & Recommendations

Most of the analysis in this thesis was performed on data collected from experiments that were not designed to study the possibilities of multivariate analysis using SIFT-MS data. A second large dataset of clearly defined samples (positive verses negatives) would be a good second validation for the bootstrapping procedures used in chapter 3 on the bovine TB in badger serum and is an ideal first step for any future work.

When designing experiments for SIFT-MS multivariate analysis, large datasets are a must; 200 samples with roughly equal numbers of positive and negatives are needed for through investigations using PLSDA validations and bootstrapping of the models. A single treatment method for the animals will remove the problems seen with a discriminating feature for the badger set. Samples from all animals must be taken using the same methods including the same anaesthetics.

Wild field samples are always going to have a large variance associated with them as there is no control on age, diet or environmental conditions. Metadata can be collected but is only useful for reducing the dataset size (removal of samples that have an adverse uncontrollable feature set for instance). The only way around this problem is to ensure that enough samples are collected so the sub-setting of the dataset does not reduce the number of samples below the threshold to produce a model that can be validated. Clinical trial samples do not have the problem of this large variation. Diet, age and environmental conditions are usually controlled. Unfortunately, these datasets are usually much smaller. Repeatedly sampling one animal does not solve this problem, it just creates more data that will be eventually cut down. One solution is to collect data from a range of clinical trials. As long as there are enough trials so that if one or two have considerably different features that the samples obtained from them can be removed modelling a validation might be possible.

It is also important that clinical trials have a clear uninfected control group. Some clinical trial (especially drug trials) have a control group that are given a placebo immunisation or treatment and hence all samples have the possibility of infection. For modelling, you need either positive or negatives. Positive verses positive but hopefully immune do not allow for model building.

Also, extending the number of multivariate techniques used would be advantageous. In this thesis a single unsupervised method and a single supervised method have been used to prove the concept and show that it is possible to analyse SIFT-MS data using multivariate techniques. SVMs would be a useful first step as a comparison to PLSDA. Equally cluster analysis on the PCA results would be informative next step for the unsupervised modelling. The data for the SVM trial will ideally be collected from samples containing a high number of VOCs (urine possibly) and recorded using the newer (mk.3) SIFT-MS machine.

It would be intriguing to investigate the ability to discriminate between different diseases which would indicate the possibility that the SIFT-MS machine could become a generic analysis tool. It would also suggest if the SIFT-MS and multivariate analysis is detecting general animal health or the specific diseases in question. This however would require a much larger dataset. The estimated number of samples required from the badger TB dataset was 200 for a single disease so multiplying this by the number of diseases to be investigated would be a good starting point.

Finally, the main body of samples investigated in this thesis are serum samples. These are probably not the most ideal samples for this type of investigation even though they are one of the easiest to capture and store. Urine samples showed more promising results, have similar storage requirements to serum samples and only a slightly more “tricky” collection process. There they would be ideal for further investigations. With human patients, breath samples are easily taken and non-invasive so despite the transient VOC levels seen during the collection process are still better than serum samples, unfortunately, breath samples from animals are almost impossible to collect.

References

ACKERMANN, M. R. & BROGDEN, K. A. (2000) Response of the ruminant respiratory tract to Mannheimia (Pasteurella) haemolytica. *Microbes and Infection,* 2**,** 1079-1088.

ADAM, B. L., QU, Y., DAVIS, J. W., WARD, M. D., CLEMENTS, M. A., CAZARES, L. H., SEMMES, O. J., SCHELLHAMMER, P. F., YASUI, Y., FENG, Z. & WRIGHT JR, G. L. (2002) Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research,* 62**,** 3609-3614.

ADAMS, N. G. & SMITH, D. (1976) The selected ion flow tube (SIFT); A technique for studying ion-neutral reactions. *International Journal of Mass Spectrometry and Ion Physics,* 21**,** 349-359.

ALBERT, K. J., LEWIS, N. S., SCHAUER, C. L., SOTZING, G. A., STITZEL, S. E., VAID, T. P. & WALT, D. R. (2000) Cross-reactive chemical sensor arrays. *Chemical Reviews,* 100**,** 2595-2626.

ALEIXANDRE, M., SAYAGO, I., HORRILLO, M. C., FERNAÌ • NDEZ, M. J., AREÌ • S, L., GARCIÌ • A, M., SANTOS, J. P. & GUTIEÌ • RREZ, J. (2004) Analysis of neural networks and analysis of feature selection with genetic algorithm to discriminate among pollutant gas. *Sensors and Actuators, B: Chemical,* 103**,** 122-128.

ALI, Z., JAMES, D., O'HARE, W. T., ROWELL, F. J. & SCOTT, S. M. (2003) Radial basis neural network for the classification of fresh edible oils using an electronic nose. *Journal of Thermal Analysis and Calorimetry,* 71**,** 147-154.

ALIZADEH, N., MOHAMMADI, A. & TABRIZCHI, M. (2008) Rapid screening of methamphetamines in human serum by headspace solid-phase microextraction using a dodecylsulfate-doped polypyrrole film coupled to ion mobility spectrometry. *Journal of Chromatography A,* 1183**,** 21-28.

ANGEN, Ã., THOMSEN, J., LARSEN, L. E., LARSEN, J., KOKOTOVIC, B., HEEGAARD, P. M. H. & ENEMARK, J. M. D. (2009) Respiratory disease in calves: Microbiological investigations on trans-tracheally aspirated bronchoalveolar fluid and acute phase protein response. *Veterinary Microbiology,* 137**,** 165-171.

BARKER, M. & RAYENS, W. (2003) Partial least squares for discrimination. *Journal of Chemometrics,* 17**,** 166-173.

BERNA, A. Z., TROWELL, S., CYNKAR, W. & COZZOLINO, D. (2008) Comparison of metal oxide-based electronic nose and mass spectrometry-based electronic nose for the prediction of red wine spoilage. *Journal of Agricultural and Food Chemistry,* 56**,** 3238-3244.

BIRKEMEYER, C., LUEDEMANN, A., WAGNER, C., ERBAN, A. & KOPKA, J. (2005) Metabolome analysis: The potential of in vivo labeling with stable isotopes for metabolite profiling. *Trends in Biotechnology,* 23**,** 28-33.

BOURNE, F. J. (2007) Bovine TB: The Scientific Evidence. IN TB, I. S. G. O. C. (Ed.

BRERETON, R. G. (2006) Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data. *TrAC - Trends in Analytical Chemistry,* 25**,** 1103-1111.

BROADHURST, D. I. & KELL, D. B. (2006) Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics,* 2**,** 171-196.

CHAISSON, R. E. & HARRINGTON, M. (2009) How research can help control tuberculosis. *International Journal of Tuberculosis and Lung Disease,* 13**,** 558-568.

CHAMBERLIN, W., GRAHAM, D. Y., HULTEN, K., EL-ZIMAITY, H. M. T., SCHWARTZ, M. R., NASER, S., SHAFRAN, I. & EL-ZAATARI, F. A. K. (2001) Review article: Mycobacterium avium subsp. paratuberculosis as one cause of Crohn's disease. *Alimentary Pharmacology and Therapeutics,* 15**,** 337-346.

CHEN, X., CAO, M., LI, Y., HU, W., WANG, P., YING, K. & PAN, H. (2005) A study of an electronic nose for detection of lung cancer based on a virtual SAW gas sensors array and imaging recognition method. *Measurement Science and Technology,* 16**,** 1535-1546.

CHO, S. N. & BRENNAN, P. J. (2007) Tuberculosis: Diagnostics. *Tuberculosis,* 87.

CIMATO, A., DELLO MONACO, D., DISTANTE, C., EPIFANI, M., SICILIANO, P., TAURINO, A. M., ZUPPA, M. & SANI, G. (2006) Analysis of single-cultivar extra virgin olive oils by means of an Electronic Nose and HS-SPME/GC/MS methods. *Sensors and Actuators, B: Chemical,* 114**,** 674-680.

CIOSEK, P., BRZOÌ • ZKA, Z. & WROÌ • BLEWSKI, W. (2006) Electronic tongue for flow-through analysis of beverages. *Sensors and Actuators, B: Chemical,* 118**,** 454-460.

COCITO, C., GILOT, P., COENE, M., DE KESEL, M., POUPART, P. & VANNUFFEL, P. (1994) Paratuberculosis. *Clinical Microbiology Reviews,* 7**,** 328-345.

CORTES, C. & VAPNIK, V. (1995) Support-vector networks. *Machine Learning,* 20**,** 273-297.

COX, R. A. (2004) Quantitative relationships for specific growth rates and micromolecular compositions of Mycobacterium tuberculosis, Streptomyces coelicolor A3(2) and Escherichia coli B/r: An integrative theoretical approach. *Microbiology,* 150**,** 1413-1426.

COZZOLINO, D., SMYTH, H. E., LATTEY, K. A., CYNKAR, W., JANIK, L., DAMBERGS, R. G., FRANCIS, I. L. & GISHEN, M. (2006) Combining mass spectrometry based electronic nose, visible-near infrared spectroscopy and chemometrics to assess the sensory properties of Australian Riesling wines. *Analytica Chimica Acta,* 563**,** 319-324.

CUTLER, S. J., WHATMORE, A. M. & COMMANDER, N. J. (2005) Brucellosis - New aspects of an old disease. *Journal of Applied Microbiology,* 98**,** 1270-1281.

D'AMICO, A., DI NATALE, C., MACAGNANO, A., DAVIDE, F., MANTINI, A., TARIZZO, E., PAOLESSE, R. & BOSCHI, T. (1998) Technologies and tools for mimicking olfaction: Status of the Rome 'Tor Vergata' electronic nose. *Biosensors and Bioelectronics,* 13**,** 711-721.

D'IMPERIO, M., GOBBINO, M., PICANZA, A., COSTANZO, S., CORTE, A. D. & MANNINA, L. Influence of harvest method and period on olive oil composition: An NMR and statistical study. *Journal of Agricultural and Food Chemistry,* 58**,** 11043-11051.

DALLEY, D., DAVE, D., LESELLIER, S., PALMER, S., CRAWSHAW, T., HEWINSON, R. G. & CHAMBERS, M. (2008) Development and evaluation of a gamma-interferon assay for tuberculosis in badgers (Meles meles). *Tuberculosis,* 88**,** 235-243.

DE LA RUA-DOMENECH, R., GOODCHILD, A. T., VORDERMEIER, H. M., HEWINSON, R. G., CHRISTIANSEN, K. H. & CLIFTON-HADLEY, R. S. (2006) Ante mortem diagnosis of tuberculosis in cattle: A review of the tuberculin tests, Î³-interferon assay and other ancillary diagnostic techniques. *Research in Veterinary Science,* 81**,** 190-210.

DE LACY COSTELLO, B. P. J., EVANS, P., EWEN, R. J., HONEYBOURNE, C. L. & RATCLIFFE, N. M. (1996) Novel composite organic-inorganic semiconductor sensors for the quantitative detection of target organic vapours. *Journal of Materials Chemistry,* 6**,** 289-294.

DETTMER, K., ARONOV, P. A. & HAMMOCK, B. D. (2007) Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews,* 26**,** 51-78.

DEWE, W. (2009) Review of statistical methodologies used to compare (bio)assays. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences,* 877**,** 2208-2213.

DI NATALE, C., DAVIDE, F. A. M., D'AMICO, A., NELLI, P., GROPPELLI, S. & SBERVEGLIERI, G. (1996) An electronic nose for the recognition of the vineyard of a red wine. *Sensors and Actuators, B: Chemical,* 33**,** 83-88.

DI NATALE, C., MACAGNANO, A., MARTINELLI, E., PAOLESSE, R., D'ARCANGELO, G., ROSCIONI, C., FINAZZI-AGROÌ€, A. & D'AMICO, A. (2003) Lung cancer identification by the analysis of breath by means of an array of non-selective gas sensors. *Biosensors and Bioelectronics,* 18**,** 1209-1218.

DICKINSON, T. A., WHITE, J., KAUER, J. S. & WALT, D. R. (1998) Current trends in 'artificial-nose' technology. *Trends in Biotechnology,* 16**,** 250-258.

DUFF, G. C. & GALYEAN, M. L. (2007) Board-invited review: Recent advances in management of highly stressed, newly received feedlot cattle. *Journal of Animal Science,* 85**,** 823-840.

EKLOV, T., MARTENSSON, P. & LUNDSTROM, I. (1999) Selection of variables for interpreting multivariate gas sensor data. *Analytica Chimica Acta,* 381**,** 221-232.

ENGLAND, T., KELLY, L., JONES, R. D., MACMILLAN, A. & WOOLDRIDGE, M. (2004) A simulation model of brucellosis spread in British cattle under several testing regimes. *Preventive Veterinary Medicine,* 63**,** 63-73.

FEND, R. (2004) Development of medical point-of-care applications for renal medicine and tuberculosis based on electronic nose technology. Silsoe, Cranfield University.

FEND, R., GEDDES, R., LESELLIER, S., VORDERMEIER, H. M., CORNER, L. A. L., GORMLEY, E., COSTELLO, E., HEWINSON, R. G., MARLIN, D. J., WOODMAN, A. C. & CHAMBERS, M. A. (2005) Use of an electronic nose

to diagnose Mycobacterium bovis infection in badgers and cattle. *Journal of Clinical Microbiology,* 43**,** 1745-1751.

FEND, R., KOLK, A. H. J., BESSANT, C., BUIJTELS, P., KLATSER, P. R. & WOODMAN, A. C. (2006) Prospects for clinical application of electronic-nose technology to early detection of Mycobacterium tuberculosis in culture and sputum. *Journal of Clinical Microbiology,* 44**,** 2039-2045.

GASPAR, E. M., LUCENA, A. F., DURO DA COSTA, J. & CHAVES DAS NEVES, H. (2009) Organic metabolites in exhaled human breath-A multivariate approach for identification of biomarkers in lung disorders. *Journal of Chromatography A,* 1216**,** 2749-2756.

GOODACRE, R., VAIDYANATHAN, S., DUNN, W. B., HARRIGAN, G. G. & KELL, D. B. (2004) Metabolomics by numbers: Acquiring and understanding global metabolite data. *Trends in Biotechnology,* 22**,** 245-252.

GOPEL, W. (1995) Controlled signal transduction across interfaces of 'intelligent' molecular systems. *Proceedings of the 2nd CEC Workshop on Bioelectronics: Interfacing Biology with Electronics. Part 2,* 10**,** 35-59.

GOPEL, W., ZIEGLER, C., BREER, H., SCHILD, D., APFELBACH, R., JOERGES, J. & MALAKA, R. (1998) Bioelectronic noses: a status report. Part I. *Biosensors and Bioelectronics,* 13.

GROTE, C. & PAWLISZYN, J. (1997) Solid-Phase Microextraction for the Analysis of Human Breath. *Analytical Chemistry,* 69**,** 587-596.

GUTIERREZ-OSUNA, R. (2002) Pattern analysis for machine olfaction: A review. *IEEE Sensors Journal,* 2**,** 189-202.

HARPER, M. (2000) Sorbent trapping of volatile organic compounds from air. *Journal of Chromatography A,* 885**,** 129-151.

HARRINGTON, P. D. B., VIEIRA, N. E., ESPINOZA, J., NIEN, J. K., ROMERO, R. & YERGEY, A. L. (2005) Analysis of variance-principal component analysis: A soft tool for proteomic discovery. *Analytica Chimica Acta,* 544**,** 118-127.

HAUPTMANN, P., LUCKLUM, R., HARTMANN, J., AUGE, J. & ADLER, B. (1993) Using the quartz microbalance principle for sensing mass changes and damping properties. *Proceedings of EUROSENSORS VI,* 37-38**,** 309-316.

HERNANDEZ, F., PITARCH, E., BELTRAN, J. & LAPEZ, F. J. (2002) Headspace solid-phase microextraction in combination with gas chromatography and tandem mass spectrometry for the determination of organochlorine and organophosphorus pesticides in whole human blood. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences,* 769**,** 65-77.

ISSAQ, H. J., VAN, Q. N., WAYBRIGHT, T. J., MUSCHIK, G. M. & VEENSTRA, T. D. (2009) Analytical and statistical approaches to metabolomics research. *Journal of Separation Science,* 32**,** 2183-2199.

JAMES, D., SCOTT, S. M., ALI, Z. & O'HARE, W. T. (2005) Chemical sensors for electronic nose systems. *Microchimica Acta,* 149**,** 1-17.

JONSSON, P., STENLUND, H., MORITZ, T., TRYGG, J., SJOSTROM, M., VERHEIJ, E. R., LINDBERG, J., SCHUPPE-KOISTINEN, I. & ANTTI, H. (2006) A strategy for modelling dynamic responses in metabolic samples characterized by GC/MS. *Metabolomics,* 2**,** 135-143.

JURS, P. C., BAKKEN, G. A. & MCCLELLAND, H. E. (2000) Computational methods for the analysis of chemical sensor array data from volatile analytes. *Chemical Reviews,* 100**,** 2649-2678.

KAAL, E., KOLK, A. H. J., KUIJPER, S. & JANSSEN, H. G. (2009) A fast method for the identification of Mycobacterium tuberculosis in sputum and cultures based on thermally assisted hydrolysis and methylation followed by gas chromatography-mass spectrometry. *Journal of Chromatography A,* 1216**,** 6319-6325.

KISCHKEL, S., MIEKISCH, W., SAWACKI, A., STRAKER, E. M., TREFZ, P., AMANN, A. & SCHUBERT, J. K. (2010) Breath biomarkers for lung cancer detection and assessment of smoking related effects -- confounding variables, influence of normalization and statistical algorithms. *Clinica Chimica Acta,* 411**,** 1637-1644.

KNOBLOCH, H. (2010) Serum and urine headspace analysis using electronic nose (e-nose) technology. Cranfield University.

KNOBLOCH, H., KAHLER, H., COMMANDER, N., REINHOLD, P., TURNER, C. & CHAMBERS, M. (2009a) Volatile organic compound (VOC) analysis for disease detection: Proof of principle for field studies detecting paratuberculosis and brucellosis. *AIP Conference Proceedings.* Brescia.

KNOBLOCH, H., TURNER, C., CHAMBERS, M. & REINHOLD, P. (2009b) Serum headspace analysis with an electronic nose and comparison with clinical signs following experimental infection of cattle with mannheimia haemolytica. *AIP Conference Proceedings.* Brescia.

KNOBLOCH, H., TURNER, C., SPOONER, A. & CHAMBERS, M. (2009c) Methodological variation in headspace analysis of liquid samples using electronic nose. *Sensors and Actuators, B: Chemical,* 139**,** 353-360.

KREBS, J. R., ANDERSON, R., CLUTTON-BROCK, T., MORRISON, I., YOUNG, D., DONNELLY, C.A., FROST, S., (...), WOODROFFE, R (1997) Bovine Tuberculosis in Cattle and Badgers. London, HMSO.

LEBRUN, M., PLOTTO, A., GOODNER, K., DUCAMP, M. N. & BALDWIN, E. (2008) Discrimination of mango fruit maturity by volatiles using the electronic nose and gas chromatography. *Postharvest Biology and Technology,* 48**,** 122-131.

MAHADEVAN, S., SHAH, S. L., MARRIE, T. J. & SLUPSKY, C. M. (2008) Analysis of metabolomic data using support vector machines. *Analytical Chemistry,* 80**,** 7562-7570.

MATSUMURA, K., OPIEKUN, M., OKA, H., VACHANI, A., ALBELDA, S. M., YAMAZAKI, K. & BEAUCHAMP, G. K. (2010) Urinary volatile compounds as biomarkers for lung cancer: A proof of principle study using odor signatures in mouse models of lung cancer. *PLoS ONE,* 5.

NICHOLAS, R. A. J. & AYLING, R. D. (2003) Mycoplasma bovis: Disease, diagnosis, and control. *Research in Veterinary Science,* 74**,** 105-112.

ONYEBUJOH, P. & ROOK, G. A. W. (2004) Tuberculosis. *Nature Reviews Microbiology,* 2**,** 930-932.

OTTO, M. (1999a) *Chemometrics*, Wiley-VCH.

OTTO, M. (1999b) *Chemometrics: statistics and computer application in analytical chemistry*, WILEY-VCH.

PARDO, M. & SBERVEGLIERI, G. (2002) Learning from data: A tutorial with emphasis on modern pattern recognition methods. *IEEE Sensors Journal,* 2**,** 203-217.

PASIKANTI, K. K., HO, P. C. & CHAN, E. C. Y. (2008) Gas chromatography/mass spectrometry in metabolic profiling of biological fluids. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences,* 871**,** 202-211.

PAVLOU, A. K., MAGAN, N., JONES, J. M., BROWN, J., KLATSER, P. & TURNER, A. P. F. (2004) Detection of Mycobacterium tuberculosis (TB) in vitro and in situ using an electronic nose in combination with a neural network system. *Biosensors and Bioelectronics,* 20**,** 538-544.

PERBELLINI, L., PRINCIVALLE, A., CERPELLONI, M., PASINI, F. & BRUGNONE, F. (2003) Comparison of breath, blood and urine concentrations in the biomonitoring of environmental exposure to 1,3-butadiene, 2,5-dimethylfuran, and benzene. *International Archives of Occupational and Environmental Health,* 76**,** 461-466.

PEREZ PAVON, J. L., DEL NOGAL SANCHEZ, M., PINTO, C. G., FERNANDEZ LAESPADA, M. E., CORDERO, B. M. & PENA, A. G. (2006) Strategies for qualitative and quantitative analyses with mass spectrometry-based electronic noses. *TrAC - Trends in Analytical Chemistry,* 25**,** 257-266.

PERKINS, M. D. & CUNNINGHAM, J. (2007) Facing the crisis: Improving the diagnosis of tuberculosis in the HIV era. *Journal of Infectious Diseases,* 196.

PERSAUD, K. C. & DODD, G. (1982) Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose. *Nature,* 299**,** 352-355.

PETRICOIN III, E. F., ARDEKANI, A. M., HITT, B. A., LEVINE, P. J., FUSARO, V. A., STEINBERG, S. M., MILLS, G. B., SIMONE, C., FISHMAN, D. A., KOHN, E. C. & LIOTTA, L. A. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet,* 359**,** 572-577.

PHILLIPS, M. (1997) Method for the collection and assay of volatile organic compounds in breath. *Analytical Biochemistry,* 247**,** 272-278.

PHILLIPS, M., CATANEO, R. N., DITKOFF, B. A., FISHER, P., GREENBERG, J., GUNAWARDENA, R., KWON, C. S., RAHBARI-OSKOUI, F. & WONG, C. (2003) Volatile markers of breast cancer in the breath. *Breast Journal,* 9**,** 184-191.

PHILLIPS, M., CATANEO, R. N., SAUNDERS, C., HOPE, P., SCHMITT, P. & WAI, J. (2010) Volatile biomarkers in the breath of women with breast cancer. *Journal of Breath Research,* 4.

PHILLIPS, M., HERRERA, J., KRISHNAN, S., ZAIN, M., GREENBERG, J. & CATANEO, R. N. (1999) Variation in volatile organic compounds in the breath of normal humans. *Journal of Chromatography B: Biomedical Sciences and Applications,* 729**,** 75-88.

RUA-DOMENECH, R. D. L., GOODCHILD, A. T., VORDERMEIER, H. M., HEWINSON, R. G., CHRISTIANSEN, K. H. & CLIFTON-HADLEY, R. S. (2006) Ante mortem diagnosis of tuberculosis in cattle: A review of the tuberculin tests, Î³-interferon assay and other ancillary diagnostic techniques. *Research in Veterinary Science,* 81**,** 190-210.

SANDER, R. (1999) *Compilation of Henry's Law Constants for Inorganic and Organic Species of Potential Importance in Environmental Chemistry (Version 3)*.

SCHULZ, K., BERTAU, M., SCHLENZ, K., MALT, S., DREÃŸLER, J. & LACHENMEIER, D. W. (2009) Headspace solid-phase microextraction-gas chromatography-mass spectrometry determination of the characteristic flavourings menthone, isomenthone, neomenthol and menthol in serum samples with and without enzymatic cleavage to validate post-offence alcohol drinking claims. *Analytica Chimica Acta,* 646**,** 128-140.

SCOTT, S. M., JAMES, D. & ALI, Z. (2006) Data analysis for electronic nose systems. *Microchimica Acta,* 156**,** 183-207.

SHELLIE, R. A., WELTHAGEN, W., ZROSTLIKOVÃ¡, J., SPRANGER, J., RISTOW, M., FIEHN, O. & ZIMMERMANN, R. (2005) Statistical methods for comparing comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry results: Metabolomic analysis of mouse tissue extracts. *Journal of Chromatography A,* 1086**,** 83-90.

SMITH, D. & ADAMS, N. G. (1980) ION/ION AFTERGLOW PLASMA AS A SOURCE OF SIMPLE AND CLUSTERED POSITIVE AND NEGATIVE IONS. *Journal of Physics D: Applied Physics,* 13**,** 1267-1273.

SMITH, D. & ADAMS, N. G. (1987) The selected ion flowtube (SIFT): Studies of ion neutral reactions. *Advances in Atomic and Molecular Physics,* 24**,** 49.

SMITH, D. & SPANEL, P. (1996) The novel selected-ion flow tube approach to trace gas analysis of air and breath. *Rapid Communications in Mass Spectrometry,* 10**,** 1183-1198.

SMITH, D. & SPANEL, P. (2005) Selected ion flow tube mass spectrometry (SIFT-MS) for on-line trace gas analysis. *Mass Spectrometry Reviews,* 24**,** 661-700.

SMITH, D., SPANEL, P., HOLLAND, T. A., AL SINGARI, W. & ELDER, J. B. (1999) Selected ion flow tube mass spectrometry of urine headspace. *Rapid Communications in Mass Spectrometry,* 13**,** 724-729.

SMITH, D., WANG, T., PYSANENKO, A. & SPANEL, P. (2008) A selected ion flow tube mass spectrometry study of ammonia in mouth- And nose-exhaled breath and in the oral cavity. *Rapid Communications in Mass Spectrometry,* 22**,** 783-789.

SNOWDER, G. D., VAN VLECK, L. D., CUNDIFF, L. V. & BENNETT, G. L. (2006) Bovine respiratory disease in feedlot cattle: Environmental, genetic, and economic factors. *Journal of Animal Science,* 84**,** 1999-2008.

STABEL, J. R. (1998) Johne's Disease: A Hidden Threat. *Journal of Dairy Science,* 81**,** 283-288.

STUSSI, E., CELLA, S., SERRA, G. & VENIER, G. S. (1996) Fabrication of conducting polymer patterns for gas sensing by a dry technique. *Materials Science and Engineering C,* 4**,** 27-33.

SULE-SUSO, J., PYSANENKO, A., PANL, P. & SMITH, D. (2009) Quantification of acetaldehyde and carbon dioxide in the headspace of malignant and non-malignant lung cells in vitro by SIFT-MS. *Analyst,* 134**,** 2419-2425.

SYHRE, M. & CHAMBERS, S. T. (2008) The scent of Mycobacterium tuberculosis. *Tuberculosis,* 88**,** 317-323.

SYHRE, M., MANNING, L., PHUANUKOONNON, S., HARINO, P. & CHAMBERS, S. T. (2009) The scent of Mycobacterium tuberculosis - Part II breath. *Tuberculosis,* 89**,** 263-266.

TURNER, C., PAREKH, B., WALTON, C., SPANEL, P., SMITH, D. & EVANS, M. (2008) An exploratory comparative study of volatile compounds in exhaled breath and emitted by skin using selected ion flow tube mass spectrometry. *Rapid Communications in Mass Spectrometry,* 22**,** 526-532.

TURNER, C., SPANEL, P. & SMITH, D. (2006a) A longitudinal study of ammonia, acetone and propanol in the exhaled breath of 30 subjects using selected ion flow tube mass spectrometry, SIFT-MS. *Physiological Measurement,* 27**,** 321-337.

TURNER, C., SPANEL, P. & SMITH, D. (2006b) A longitudinal study of breath isoprene in healthy volunteers using selected ion flow tube mass spectrometry (SIFT-MS). *Physiological Measurement,* 27**,** 13-22.

TURNER, C., SPANEL, P. & SMITH, D. (2006c) A longitudinal study of ethanol and acetaldehyde in the exhaled breath of healthy volunteers using selected-ion flow-tube mass spectrometry. *Rapid Communications in Mass Spectrometry,* 20**,** 61-68.

VILLAR-GAREA, A., GRIESE, M. & IMHOF, A. (2007) Biomarker discovery from body fluids using mass spectrometry. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences,* 849**,** 105-114.

VITO, S. D., FRANCIA, G. D., PIGA, M. & MARTINOTTO, L. (2008) Performance analysis of e-nose on-field calibration for city air pollution quantitative monitoring. *Proceedings of IEEE Sensors.* Lecce.

WALT, D. R., DICKINSON, T., WHITE, J., KAUER, J., JOHNSON, S., ENGELHARDT, H., SUTTER, J. & JURS, P. (1998) Optical sensor arrays for odor recognition. *Biosensors and Bioelectronics,* 13**,** 697-699.

WANG, W., ZHOU, H., LIN, H., ROY, S., SHALER, T. A., HILL, L. R., NORTON, S., KUMAR, P., ANDERLE, M. & BECKER, C. H. (2003) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Analytical Chemistry,* 75**,** 4818-4826.

WESTERHUIS, J. A., HOEFSLOOT, H. C. J., SMIT, S., VIS, D. J., SMILDE, A. K., VELZEN, E. J. J., DUIJNHOVEN, J. P. M. & DORSTEN, F. A. (2008) Assessment of PLSDA cross validation. *Metabolomics,* 4**,** 81-89.

WHITE, P. C. L., BOHM, M., MARION, G. & HUTCHINGS, M. R. (2008) Control of bovine tuberculosis in British livestock: there is no 'silver bullet'. *Trends in Microbiology,* 16**,** 420-427.

WONGRAVEE, K., HEINRICH, N., HOLMBOE, M., SCHAEFER, M. L., REED, R. R., TREVEJO, J. & BRERETON, R. G. (2009) Variable selection using iterative reformulation of training set models for discrimination of samples: Application to gas chromatography/mass spectrometry of mouse urinary metabolites. *Analytical Chemistry,* 81**,** 5204-5217.

WU, W., MASSART, D. L. & DE JONG, S. (1997) The kernel PCA algorithms for wide data. Part I: Theory and algorithms. *Chemometrics and Intelligent Laboratory Systems,* 36**,** 165-172.

ZLATKIS, A., BRAZELL, R. S. & POOLE, C. F. (1981) The role of organic volatile profiles in clinical diagnosis. *Clinical Chemistry,* 27**,** 789-797.

Appendix 1 – Evaluation of a combination of SIFT-MS and multivariate data analysis for the diagnosis of Mycobacterium bovis in wild badgers

Andrew D. Spooner,[a] Conrad Bessant,[a] Claire Turner,*[a] Henri Knobloch[a] and Mark Chambers[b]

The currently accepted "gold standard" tuberculosis (TB) detection method for veterinary applications is culturing from a tissue sample *post mortem*. The test is accurate but growing *Mycobacterium bovis* is difficult and the process can take up to 12 weeks to return a diagnosis. In this paper we evaluate a much faster screening approach based on serum headspace analysis using selected ion flow tube mass spectrometry (SIFT-MS). SIFT-MS is a rapid, quantitative gas analysis technique, with sample analysis times of as little as a few seconds. Headspace from above serum samples from wild badgers, captured as part of a randomised trial, was analysed. Multivariate classification algorithms were then employed to extract a simple TB diagnosis from the complex multivariate response provided by the SIFT-MS instrument. This is the first time that such multivariate analysis has been applied to SIFT-MS data. An accuracy of TB discrimination of approximately 88% true positive was achieved which shows promise, but the corresponding false positive rate of 38% indicates that there is more work to do before this approach could replace the culture test. Recommendations for future work that could increase the performance are therefore proposed

1 Introduction

As well as being a major health problem in the human population(Onyebujoh and Rook, 2004), tuberculosis (TB) is also a problem in agriculture, as highlighted by the media in the UK with the publication of the Krebs report (1997)(Krebs, 1997) and the Independent Scientific Group (ISG) on Cattle TB report (2007)(Bourne, 2007). The bovine TB situation is exacerbated in the UK and Ireland by a reservoir of *Mycobacterium bovis* infection in badgers (*Meles meles*). One of the problems identified is the need for a rapid detection method to decrease the time between test and treatment in humans and to decrease the time cattle are quarantined waiting for diagnosis. The ISG was commissioned by the UK government to look into the impact of badger culling on the prevalence and transmission of tuberculosis in cattle. As part of this, a large number of samples were collected from culled badgers, providing the basis for evaluating the novel diagnostic method described in this paper.

Historically (and long since abandoned as a diagnostic technique), tuberculosis was known for a characteristic smell on the breath of the infected. This smell must be associated either with the production of specific "marker" volatiles by the bacteria or an increase in the levels of particular volatiles produced by the host in response to the infection. This has led researchers to investigate the use of gas analysis techniques for rapid TB detection.

1.1 Choice of sample

TB is generally found in the lungs of the host be it human or animal and therefore the "gold standard" technique for detection of the disease is a culture test from lung tissues obtained *post mortem(Rua-Domenech et al., 2006)*. Sputum or tracheal aspirate samples from infected individuals can contain high levels of the bacteria but only when the host respiratory tract is infected and are either difficult or impossible to obtain from animals, and from children. As volatile organic compounds (VOCs) are known to be emitted in breath this would be the next biologically preferred sample, but the samples are difficult to obtain, store and transport safely. As the lungs provide a large interface between the body and the atmosphere, many VOCs found in breath can also be found in blood serum samples(Phillips, 1997). Such samples are relatively easy to obtain, store and distribute safely and hence are the sample of choice for this

study. In addition, previous work has shown that electronic nose technology has been able to discriminate between the serum of cattle infected with *M. bovis* and uninfected control animals(Fend et al., 2005).

1.2 SIFT-MS

SIFT-MS was originally developed to study ion neutral reactions at thermal interaction energies before being adapted to perform volatile gas analysis(Smith and Adams, 1987). More recently it was further developed for life science applications(Smith and Spanel, 2005). The technique relies on trace gases reacting with precursor ions (H_3O^+, NO^+ and O_2^+) generated in a microwave discharge through moist air. The precursor ions are selected using a quadrupole mass filter and then injected into a helium carrier gas which then passes down a flow tube. The sample of headspace gas to be analysed is introduced into the flow tube by means of a heated capillary at a known flow rate, and the trace gas components rapidly react with the precursor ions to generate product ions. These are then detected downstream. The SIFT-MS instrument may be operated in one of two modes: the first is the full scan mode where full spectra are obtained over a range of values of

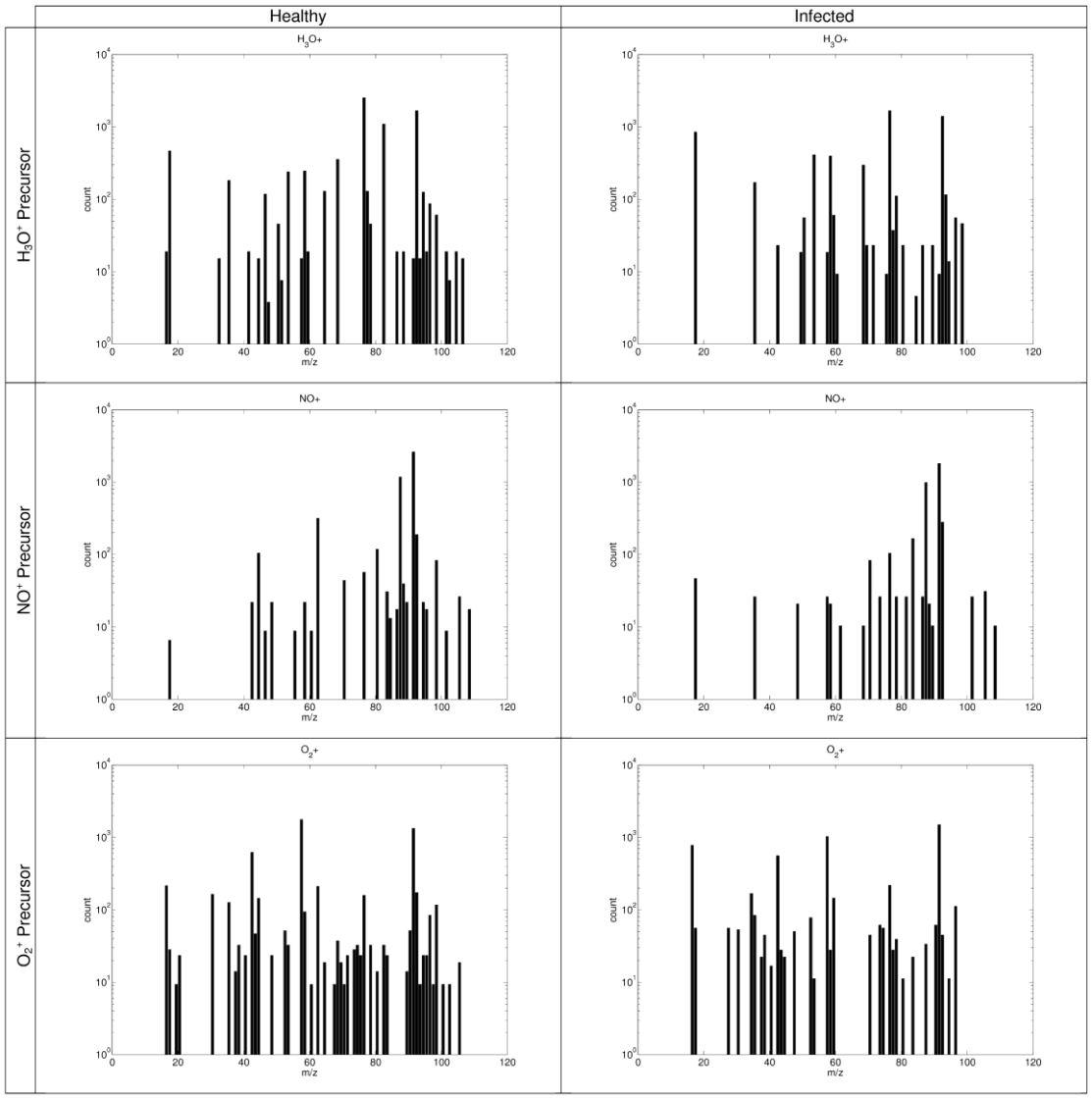


Figure 1 A table of the spectra using all three precursor ions (H3O+, NO+ and O2+) for the headspace above two serum samples (one from a badger later classified as tuberculosis negative, and one from a badger classified as tuberculosis positive using the culture test). The m/z values linked to the precursors have been zeroed (see text for explanation)

mass to charge ratio (m/z) – this is the mode used in this study. The alternative, which is good for

quantitative analysis of individual compounds may be carried out by rapidly switching the detector between preselected ions; the ratio of product ions to precursor ions enables the quantification of the analyte of interest(Smith and Spanel, 2005). The concentration of compounds may be readily determined through these data and knowledge of the reaction rate co-efficient between precursor and product ion. However, due to the relatively small user community, public databases of these reaction coefficients of specific VOCs are only available for a small number of compounds compared to the database of more than 200,000 GC-MS signatures contained in the NIST database (www.nist.gov). Therefore, alternative data analysis approaches are required. Some of the benefits of SIFT-MS over GC-MS are rapid analysis time, lower mass range and the relative simplicity of the spectrum returned. However, complex biological samples still yield highly multivariate responses which are difficult to analyse without computational assistance.

Previous studies involving SIFT-MS analysis of human or animal samples utilised prior knowledge about the spectra and compounds present(Turner et al., 2006a). Usually, markers are known and the effects of a trial are determined by the change in a small number of selected m/z peaks. The problem investigating diseases using volatile analysis are well known and two fold. Firstly, biological samples contain hundreds of volatiles that are naturally occurring and have a naturally high variance. Finding the two or three reliable markers can be difficult and require much manual work. Secondly, it is not guaranteed that these markers exist. The characteristic smell of the disease may be from the combination of ten or more individual compounds, each with an associated m/z peak, each of which, as a single entity, is not suitable as a marker. Indeed, they may all be present naturally, but in different relative concentrations.

Chemometrics techniques such as principal components analysis (PCA) and Partial Least Squares Discriminant Analysis (PLSDA(Barker and Rayens, 2003)) have been applied in similar situations to data from other sources such as electronic nose(Scott et al., 2006) and GC-MS(Jonsson et al., 2006). These are the approaches applied here.

2 Methodology

2.1 Sample collection and measurement

The samples were all collected from the Defra funded Randomised Badger Culling Trial managed by the ISG on Cattle TB(Bourne, 2007). Wild badgers from different regions across the UK, where TB has been repeatedly found in cattle, were cage trapped. The serum samples for this study were obtained from a subset of these badgers during trapping operations ending in 2005 that were first anaesthetised, then blood samples taken before the animals were killed by lethal injection. Samples from an extensive range of tissues were removed post-mortem and plated for diagnosis of *Mycobacterium bovis* infection status. These blood samples have previously been used to evaluate the potential for gamma interferon as a diagnostic tool for TB in badgers(Dalley et al., 2008). Individual blood samples were processed to produce sera and frozen for analysis at a later date.

The 245 sera samples from the above trial were shipped to Cranfield University where they were thawed. 700 µL of sample was pipetted into a sample bag made from Nalophan NA tubing (Kalle, UK) with diameter 65mm and length 31cm. 700 mL of zero grade (hydrocarbon free) air (BOC) was added and the sample bags incubated at 40°C. Two thirds (467ml) of headspace was used for other studies and the remaining third (233mL) analysed using SIFT-MS. Because the serum samples are aqueous, and incubated at 40^{O}C, the water vapour pressure was inherently high (about 7% water).

The SIFT-MS analysis was performed using 10 s scans per precursor which were repeated ten times across the three precursor ions (total analysis time of five minutes). Full scan mode was employed, over a mass to charge ratio of 10-109 (the limit of the SIFT-MS instrument used in this study at that time). The SIFT-MS instrument used was a Mk2 model from PDZ Europa (UK). The sample was incubated at 40°C for 10 minutes prior to and during connection to the SIFT-MS inlet capillary.

Once all the spectra were recorded and saved the raw data were imported into a standard spreadsheet format for analysis.

2.2 Data preparation and pre-processing

As the precursors are carrier ions added to the sample, they need to be removed from the dataset as they are clearly unrelated to the infection status of the sample. This is performed by zeroing the count for that precursor's ions across all samples. The ions removed are: m/z of 19, 20, 21, 30, 32, 37, 38, 39, 55, 56, 57, 73, 74, 75, 91. These ions comprise the 19 precursor plus isotopologues at m/z 20 and 21, plus water clusters (due to the aqueous nature of biological samples) and their isotopologues at 37, 38 and 39; 55, 56, 57; 74, 75 and 76 and 91. Very small amounts of 30 and 32 (NO^+ and O_2^+) are also present in the H_3O^+ spectrum, so

are also excluded. Based on similar reasoning, 19, 30, 32, 37, 48, 50, 55, 57, 66, 73, 91, are removed from the NO+ precursor channel and 19, 30, 32, 33, 34, 37, 50, 55, 56, 57, 73, from the O_2^+ channel. After removal of these precursors, univariate analysis is possible by investigating the spectra by hand. Figure 1 shows three spectra (one for each precursor) from the analysis of a serum sample from a wild badger classified as infected by the culture test and three spectra from the analysis of serum from an animal classified as uninfected by the culture test.

After removal of the peaks caused by the precursors, the three spectra from each sample are concatenated into a single vector of length 300 and all the samples combined into a matrix, ***X***. Hence, vector positions 1-100 correspond to the H_3O^+ spectra from 10 m/z to 109 m/z, vector positions 101-200 correspond to the NO^+ spectra from 10 m/z to 109 m/z and the last 100 vector positions correspond to the O_2^+ spectra from 10 m/z to 109 m/z. This dataset is then used for all subsequent multivariate analysis.

The dataset contained 194 samples that were culture negative and 51 samples that were culture positive (245 samples in total). It should be noted that although culture is a highly respected diagnostic tool, it is not 100% accurate(Dalley et al., 2008) and a small number of the negative tuberculosis classified badgers may have had an undiagnosed low level infection.

Heatmaps (plotting the intensities of each element of the matrix as a colour) were then used to verify that the data has been correctly imported and combined. The effect of data scaling was also investigated, with the methods considered being mean centring (Equation 1), auto scaling (Equation 2) and taking logs (Equation 3).

$$x_{ij} = x_{ij} - \bar{x}_j \qquad (1)$$

$$x_{ij} = \frac{\left(x_{ij} - \min\left(x_j\right)\right)}{\left(\max\left(x_j\right) - \min\left(x_j\right)\right)} \qquad (2)$$

$$x_{ij} = \log_{10}\left(x_{ij} + 1\right) \qquad (3)$$

2.3 Data Analysis

Having pre-processed and checked the data, the first analysis method employed was PCA. This is a common data exploration method(Otto, 1999b) which extracts and displays the characteristics that caused the greatest variance in the dataset. It breaks the data into Eigenvectors and Eigenvalues. The two matrices are then re-arranged and become scores - which indicate the relationships between the samples - and loadings, which reveal how much each measured variable contributes to the scores. The components required are selected and the remaining information ends up as a discarded error matrix. The scores matrix contains dimensions of decreasing variance content that can be plotted against each other to produce the figures.

Next, PLSDA(Barker and Rayens, 2003) was used to build classification models using the tuberculosis status of the samples. PLSDA is a supervised method, so information about the parameter of interest (in this case TB infection status) is required to train the algorithm to determine which parts of the acquired spectra capture the maximum difference between the class states. PLSDA first separates the data into two groups, one with negatives (or class 1) and the other with positives (class 2). It then continues in a similar fashion to PCA, in that Eigenvectors and Eigenvalues are obtained. However, with PLSDA the difference between the variance of the two groups (the co-variance), is maximised. Theoretically, the information returned should contain any information that could be caused by the property under investigation.

It is known that PLSDA models can over estimate the accuracy of classification if not properly validated(Westerhuis et al., 2008). Indeed, given sufficient data, any training algorithm should be able to correctly classify all the samples used to build the model. To cope with this problem, the model was optimised using a Leave One Out (LOO) method. In this method, a sample is classified against a model built using the rest of the samples. This process is then repeated with each sample until all the samples have been classified. Information about the number of true positive and true negative identifications was also extracted.

An important question in studies of this type is how many samples are required to build a classification model of suitable diagnostic power(Brereton, 2006). This is important to minimise both the number of samples collected and the number of analyses required. To answer this question, a random selection of 50 samples were selected and leave one out cross validation performed. The number of correctly classified samples was extracted and stored. The sample size was then increased by one (and randomly re-sampled)

and the process repeated until all but one sample had been included. The whole process was then bootstrapped (repeated ensuring different sets of random samples were used) 50 times to achieve a statistical distribution at each point.

3 Results and Disscussion

The raw spectra for any single positive sample and any single negative sample show a large number of similarities and a large number of differences. Comparing two positive or two negative spectra also shows a number of similarities and differences. The problem is therefore to track down reliable diagnostic differences between the groups of spectra. Clearly, separating any differences between the 194 negative spectra and the 51 positive spectra without using multivariate techniques would be both time consuming and subject to

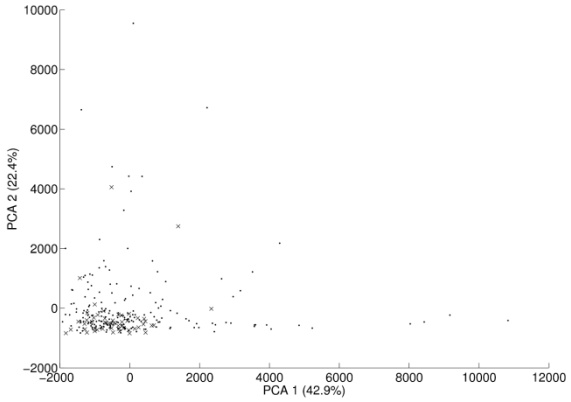


Figure 2 The PCA scores plot does not appear to show any discrimination between TB (×) and control samples (.). Investigating the loadings shows that the information on the plot relates to the concentrations of m/z 77 and m/z 83 which are likely to be caused by acetone (PC1 positive direction) and ethanol (PC2 positive direction) respectively. This information is unlikely to be related to TB but another influence such as diet

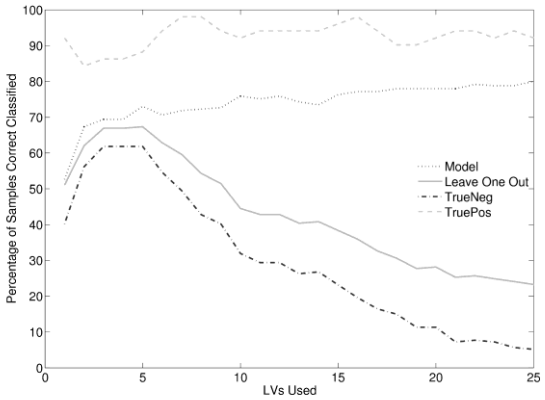


Figure 3 Partial Least Squares Discriminant Analysis (PLSDA) model (dotted line) and leave one out optimisation curves, correctly classified (line), true positives (dashed line) and true negatives (dash-dot line) for the first 25 latent variables (LV's) included. The accuracy of the PLSDA models (dotted line) shows the number of samples correctly classfied as a percentage. It is well known that this over estimates the accuracy of the model and therefore the leave one out results are more informative. The total percentage of samples correctly classified by the leave one out algorithm is then broken down into the percentage of true positives and true negatives. The optimal number of latent variables is five and this returns an accuracy (as estimated by LOO) of 67%, with 88% true positive and 62% true negative

potential error.

The PCA scores plot derived from the experimental data is shown in Figure 2. Although the figure does not show any discrimination between infected and uninfected, there does appear to be some structure. Analysis of the PCA loadings (the weight given to each m/z peak) shows that the structure seen is caused by acetone (PC1) and ethanol (PC2). The main ions implicated are 77 for acetone using H_3O^+, m/z 77 is the ion

resulting from acetone.H^+.H_2O i.e. one water cluster, which arises because of the high water vapour pressure in the samples. The water cluster of the isomer, propanal, will also produce ions at the same m/z, however, other work with GC-

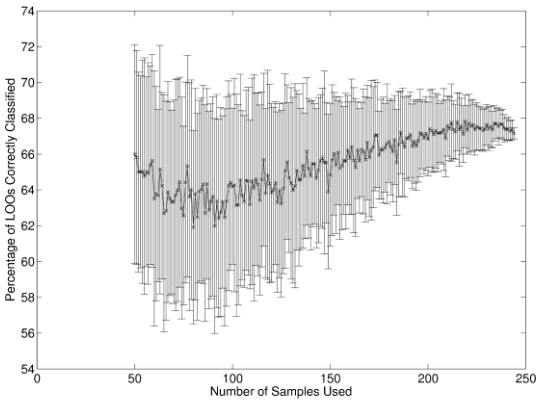


Figure 4 The mean and standard deviation of Partial Least Square Discriminant Analysis (PLSDA) leave one out (LOO) corrected classified results over a range of dataset sample sizes using 64 bootstraps. Error bars indicate one standard deviation, capturing 66% of the models produced.

MS (data not shown) indicates that it is likely to be largely, if not exclusively, due to acetone. m/z 88 was also found with NO^+, indicating acetone. Similarly, m/z 83 is the main ion generated for ethanol, which corresponds to ethanol.H^+.$2H_2O$, also using H_2O^+. These two compounds are associated with diet - ethanol is a natural product of decay in anaerobic conditions and acetone is a naturally occurring systemic compound that varies widely in concentration, depending on diet, blood sugar and individual metabolism(Turner et al., 2006a). This implies that the natural difference between the animals is larger than the difference caused by TB. This doesn't mean that there is not information in the data pertaining to TB, just that it does not cause the majority (65%) of the variance captured by the first two principal components. However, investigating PCA components of lower variance did not reveal clear disease-related discrimination either.

PLSDA has the potential to distinguish between samples that could not be separated using PCA because it maximises the co-variance between the acquired data and the sample classifications (control *versus* TB). This gives the possibility of distinguishing between samples. In Figure 3, an estimate of this ability has been produced (dotted line). More important is the result produced using the leave one out algorithm (solid line) in which models are built using all but one sample and then the sample is classified. The number of correctly classified positives (dashed line) and correctly classified negatives (dot-dashed line) is also shown as a percentage. The leave one out validation on the complete dataset shows that an overall accuracy of 67% can be achieved with 88% correctly classified positives and 62% correctly classified negatives using the currently available dataset. Analysing the loadings of the PLSDA analysis is more complicated than analysing the loadings of PCA. The discrimination seen is the result of seven latent variables and these all need to be analysed to find which m/z are the most likely to be responsible for the discrimination. 18, 36, 54(ammonia), 51(methanol), 77, 95(acetone), 83(ethanol), 93(toluene from the anaesthetic), 95 on H_3O^+; 18, 45(ethanol), 88(acetone), 92(toluene) 63 (unknown), on NO^+; 17, 43, 45, 58(acetone), 75, 77, 92(toluene), 109 (unknown) on O_2^+ are the m/z that have a high impact on the discrimination. These compounds are responsible for a high proportion of the co-variance used during classification, yet the list is in no way complete. One implication is that the differences between the TB positive and TB negative samples may be subtle and not due to just one or two biomarkers.

The relationship between the accuracy of the PLSDA LOO performance and the size of the training dataset is shown in Figure 4. It can be see that the maximum accuracy of the PLSDA has yet to be reached, although the average improvement in accuracy reduces as the number of samples increases, suggesting that provision of additional samples might not substantially improve diagnostic performance. It can also be seen that in this specific example a minimum number of samples required to produce a repeatable model is 150. Below this, the accuracy of the model (65%) is stable but the errors associated are very large and increase with fewer samples. It appears that using more than 250 samples is unlikely to substantially increase the accuracy of diagnosis.

4 Conclusions

It has been shown that by linking SIFT-MS analysis with chemometrics techniques we have made some progress towards a new method of disease detection provided that the dataset sizes are large enough to enable training of statistical models. PLSDA accuracy estimated using a leave one out (LOO) algorithm was 67%, with 88% true positives and 62% true negatives. A data set size of 150 to 250 samples was required for the samples under investigation.

This work has demonstrated for the first time a way of analysing large SIFT-MS datasets without trying to analyse individual compounds in individual spectra. Although the approach proposed is much faster than a traditional culture test, the accuracy achieved makes it unsuitable as a replacement for the culture test. The gamma interferon test is the most accurate test for TB in the live badger and achieved a 93.6% true negative and 80.9% true positive rate using the same subset of animals(Dalley et al., 2008). It is encouraging that the headspace analysis method detected even more badgers with culture-confirmed TB but at the expense of more apparent false positive results. However, the sensitivity of the gamma interferon test could also be increased by changing the test positive cut-off point if a higher false positive rate was accepted. It is believed that the overall accuracy achieved with the SIFT-MS method was limited due to the sample set being subject to large quantities of unknown variables. Factors such as sex, age, diet, location, general health of animal and anaesthetic injection would have all had an impact on the volatiles within the samples, although some of the samples classified as false positive may actually represent badgers with TB undiagnosed by culture. The anaesthetic (a cocktail of ketamine and medetomidine) in particular, appeared to lead to very high levels of toluene in the blood (seen as peaks at m/z 93 in H_3O^+ spectrum and 92 in NO^+ and O_2^+ spectra).

In terms of further work we recommend looking at a better controlled group of animals as this would introduce the possibility of looking for a group of markers specific to TB. Once these have been identified, it should be easier to remove from the data variance caused by factors other than tuberculosis. In addition, new developments in SIFT-MS instruments mean that better sensitivity and mass ranges are now available. Such an instrument would be able to analyse more compounds and thus would most likely enable better discrimination.

Acknowledgement

This paper forms part of project SE3221 which was funded by the UK's Department for Environment, Food and Rural Affairs and the authors gratefully acknowledge DEFRA's funding.

Notes and references

[a] *Cranfield University, College Road, Cranfield, Bedfordshire, MK43 0AL, UK.; E-mail: c.turner@cranfield.ac.uk*
[b] *TB Research Group, Department of Statutory and Exotic Bacteria, Veterinary Laboratories Agency Weybridge, New Haw, Addlestone, Surrey KT15 3NB, UK. Fax: 01932 357260; Tel: 01932 357494; E-mail: m.a.chambers@vla.defra.gsi.gov.uk.*

1. P. Onyebujoh and G. A. W. Rook, Nature Reviews Microbiology, 2004, 2, 930-932.
2. J. R. Krebs, ed. F. a. F. P. Ministry of Agriculture, 1997.
3. F. J. Bourne, ed. I. S. G. o. C. TB, 2007.
4. R. d. l. Rua-Domenech, A. T. Goodchild, H. M. Vordermeier, R. G. Hewinson, K. H. Christiansen and R. S. Clifton-Hadley, Research in Veterinary Science, 2006, 81, 190-210.
5. M. Phillips, Analytical Biochemistry, 1997, 247, 272-278.
6. R. Fend, R. Geddes, S. Lesellier, H. M. Vordermeier, L. A. L. Corner, E. Gormley, E. Costello, R. G. Hewinson, D. J. Marlin, A. C. Woodman and M. A. Chambers, Journal of Clinical Microbiology, 2005, 43, 1745-1751.
7. D. Smith and N. G. Adams, Advances in Atomic and Molecular Physics, 1987, 24, 49.
8. D. Smith and P. Spanel, Mass Spectrometry Reviews, 2005, 24, 661-700.
9. C. Turner, P. Spanel and D. Smith, Physiological Measurement, 2006, 27, 321-337.
10. M. Barker and W. Rayens, Journal of Chemometrics, 2003, 17, 166-173.
11. S. M. Scott, D. James and Z. Ali, Microchimica Acta, 2006, 156, 183-207.
12. P. Jonsson, H. Stenlund, T. Moritz, J. Trygg, M. Sjostrom, E. R. Verheij, J. Lindberg, I. Schuppe-Koistinen and H. Antti, Metabolomics, 2006, 2, 135-143.
13. D. Dalley, D. Dave, S. Lesellier, S. Palmer, T. Crawshaw, R. G. Hewinson and M. Chambers, Tuberculosis, 2008, 88, 235-243.
14. M. Otto, Chemometrics: statistics and computer application in analytical chemistry, WILEY-VCH, 1999.

Appedix 1

15. J. A. Westerhuis, H. C. J. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. J. Velzen, J. P. M. Duijnhoven and F. A. Dorsten, Metabolomics, 2008, 4, 81-89.
16. R. G. Brereton, TrAC - Trends in Analytical Chemistry, 2006, 25, 1103-1111.

Appendix 2 - Methodological variation in headspace analysis of liquid samples using electronic nose

Henri Knobloch1§, Claire Turner1, Andrew Spooner1, Mark Chambers2

1Cranfield Health, Cranfield University Cranfield, Bedfordshire, MK43 0AL, UK

2 TB Research Group,Department of Statutory and Exotic Bacterial Diseases Veterinary Laboratories

Agency Weybridge Woodham Lane New Haw, Addlestone Surrey KT15 3NB UK

§Corresponding author

Abstract

In past years, numerous electronic nose (e-nose) developments have been published describing analyses of solid-, liquid- or gaseous media in microbiological-, environmental-, agricultural- or medical applications. However, little has been reported about complex methodological pitfalls that might be associated with commercially available e-nose technology. In this paper, some of these pitfalls such as *temperature*, the *use of filters* and *mass flow* using different sampling methods (static- and dynamic sampling) are described for two generations of conducting polymer e-noses (ST114/214, CPs, both Scensive Tech. Ltd). A comparison with metal oxide semi-conducting field effect transistor/ metal oxide semiconductor (MOSFET/MOS) e-noses regarding *stability across replicates* and *over time* was made. Changes in temperature were found to give larger sensor responses, whereas the application of filters led to quantitative and qualitative changes in sensor responses due to a change in mass flow which was also affected by the sampling method. Static sampling provided more stable flows across replicates. Variation was investigated for CPs and MOSFET/MOS e-noses that gave different responses over time and across replicates. These methodological factors cause a lack of stability and reproducibility, demonstrating the pitfalls of e-nose technology and therefore limit their utility for discriminating between samples.

Keywords: Electronic nose, temperature effect, headspace, filter, sampling method

1. Introduction

Appedix 2

For more than 20 years, so-called electronic noses (e-noses) have been widely applied for headspace- and trace gas analysis from solid-, liquid- and gaseous samples. During this time, different sensing methods have been developed but the e-nose principle remained the same [1-5]. Generally, e-noses consist of an odour delivery system, an array of sensors, a data acquisition- and a data analysis unit. In all types of e-nose, the sensor array exploits the conversion of changes in electrical-, thermal-, mass- or optical properties into an analysable signal [6]. As for human noses, sensors show an overlapping specificity and therefore provide a non-quantitative sensor response, a so-called fingerprint, rather than a specific qualitative or quantitative answer. The lack of specificity is due to the fact that sensors react with functional groups and structures of analytes rather than specifically with the molecule itself [7, 8]. Sensor responses are then analysed using multivariate data analysis techniques in order to classify or discriminate a group of samples. But data analysis methods are complicated by the fact that there are multiple causes of variation –biological differences between samples or differences arising from the analysis itself (methodological variation). Methodological variation, including sensor drifts, may lead to unclear or contradictory results because the lack of repeatability or "disturbance" can be expressed as variation which may obscure biological variation. Hence the elimination of methodological variation is one of the biggest challenges associated with e-nose technology [9]. In this study, methodological influences associated with headspace analysis from liquids were investigated using two typical but different types of electronic nose systems; two portable conducting polymer sensor based e-noses (CPs) and two non-portable MOSFET/ MOS based e-noses. For each type, devices were compared.

Temperature is the main influencing factor in generating and analysing headspace from liquids. As predicted by Henry's law, temperature affects the vaporisation of molecules. Even minor increases in temperature lead to a higher concentration of molecules in the gas phase. For the same reason, it also affects relative humidity. Variation of e-nose responses due to these parameters has been reported by various authors [10-13]. CP e-noses are particularly sensitive to humidity [14]. The electrical conductivity for CP e-noses is increased especially for electrophobic and nucleophilic substances [15]. Both monomer structure and inter-chain behaviour change [16] due

to water vapour adsorption on the sensor surface. Filters have been applied to prevent sensor surface damage due to moisture and, in case of infectious material, to prevent contamination [17, 18].

Memory effects were also reported for CPs, in particular with gaseous compounds which have a strong proton affinity. The interaction with monomers leads to a further sensitivity loss [19]. Unlike MOSFET/ MOS e-noses, conducting polymer-based e-noses were also reported to be almost resistant to poisoning and to have a higher selectivity in comparison to semiconductors [10, 11]. Furthermore, the growth process for CP e-nose polymers is variable leading to different sensor arrays and responses [11, 20].

Regarding sampling methods, two different types can be distinguished: *static* and *dynamic*. The main difference between both methods is the constant pressure in the sampling apparatus for static sampling while pressure varies using dynamic sampling. Changes in pressure change the flow rate of molecules and therefore dynamic sampling may give more inconsistent results. Since dynamic sampling requires aerating the sample in order to compensate the pressure difference (e.g. lab air), sample dilution occurs leading to less sensitive analyses and temporal variability due to a nonequilibrium state (variation over replicates and runs) [12, 13, 21].

For MOSFET and MOS sensor-based e-noses, the effect of the temperature of the sensors and temperature of the samples analysed strongly influence the responses [22]. However, since sensors were strongly temperature and humidity controlled these parameters were not included in this study. MOSFET and MOS sensors were only considered in terms of repeatability over replicates and time.

The purpose of this study was to examine methodological variation during sampling process which may influence sensor responses and therefore sample classification and discrimination. For two CP e-noses, the variability in sensor response due to changes in temperature, flow rates and experimental setups (usage of filters, different sampling techniques) was assessed. Furthermore, the repeatability of analyses (replicates, over time) was investigated for CP and MOS/MOSFET e-noses. E-noses of the same type were compared and variation during the sampling process was assessed using both types of e-nose.

2. Materials and Methods

Appedix 2

2.1 Apparatus

Two types of e-nose were used: conducting polymer- (CP) and metal oxide semi-conducting field effect transistor/ metal oxide semiconductor (MOSFET/MOS) e-noses.

The Bloodhound ST114 and the ST214 CP e-noses (both Scensive Tech. Ltd.) contained 14 conducting polymer sensors. The headspace is conveyed from the sampling container via a sampling port to the sensor array using an internal pump. The sensor array is not in a temperature and flow controlled environment. The time for adsorption and desorption during analyses was 20 seconds and seven replicates were analysed unless otherwise stated. The total time per replicate was 55 seconds. Two MOSFET/ MOS e-noses (both NST 3220 Lab Emission Analyser; Applied Sensors, Linköping, Sweden) were assessed. The NST lab emission analyser houses ten MOSFET sensors (FE101A- FE105A, and FE101B- FE105B), twelve MOS sensors (MO101, MO102, MO104, MO110- MO118) and a humidity sensor. The principle of sampling is different from that of the ST214 e-noses. Samples are analysed from 30mL vials sealed with a septum and a lid which are placed in a rotating carousel and which maintains a constant temperature (e.g. 25°C) for sample equilibration. A double needle penetrates the septum and pumps headspace to the sensor panel where reversible adsorption processes take place at 140°C (MOSFET) and 170°C (MOS). The second needle maintains atmospheric pressure in the vial.

2.2 Sampling methods

The effect of two types of sampling on the sensor response was investigated.

Dynamic sampling: Using this method, a liquid sample (5 mL) was pipetted into a 30 mL vial and directly connected to the e-nose sampling port. When the e-nose started sniffing the headspace, a pressure difference ($\cdot p$) was generated which was compensated for by an incoming air stream (e.g. lab air).

Static sampling: An inflatable Nalophan bag (KALLE, UK) was attached via a polypropylene tube with a Swagelok fitting (Swagelok, England) to a Luer counterpart.

The Luer fitting could be directly connected to the e-nose. An internal pump delivered the headspace to the sensor array and the bag gradually collapsed due to the volume reduction. Nalophan is available in sleeves of different diameters and its

length may be varied according to sample volume required. The material does not emit volatiles detectable by e-nose (data not shown). To compare different sampling methods, the ratio of the liquid sample volume and headspace air volume was calculated (VSAMPLE/ VHEADSPACE).

2.3 Assessing the effect of temperature on e-nose response

The Bloodhound ST214 e-nose used in this study lacks a temperature controlled environment. In order to demonstrate the effect of temperature on the e-nose, reverse osmosis water (ROW) and bovine serum were analysed at different temperatures. 0.5 mL of serum and ROW were pipetted into small Nalophan bags with a volume of 0.7 L to give the ratio of sample- to bag volume of 7E-04. The serum was obtained from one clinically healthy bovine animal (breed "Holstein") held at the FLI (Friedrich Loeffler Institute, Jena, Germany). Both samples were incubated for 30 minutes and analysed at room temperature (21°C) and in a water bath (Type JB1, Grant Instruments, Ltd.) at 30°C and 40°C, respectively; ambient air was also analysed as a standard.

2.4 Mass flow and filters

Mass flows using different setups

The Bloodhound ST214 and ST114 e-noses used in this study lack a flow controller for adjusting mass flow. Hence, flow stability and rates were monitored under different sampling configurations using a mass flow sensor (Honeywell, AMW3300V). First, a Luer fitting was attached to the e-nose sampling port and to the mass flow sensor followed by a 0.45 µm pore sized Sartorious Minisart filter which was attached to the bag or sampling container. The mass flow rates for all combinations and both CPs were measured. The bag volume for static sampling was 0.7 L and the container volume for dynamic sampling was 30 mL. 1 mL of ROW or 5 mL of ROW were pipetted into the bag and the container as a standard solution, respectively. Flow rates were measured during the adsorption phase, performed in duplicate and given as averaged values. Temperature was kept constant at 25°C.

Mass flow stability for static sampling

The mass flow over eight successive replicates was monitored using static sampling. The experimental setup (calibration and analysis using two ST214) was similar to that described above with an increase in bag volume to 0.8L while the volume of a 10ppm

2-butanol standard solution was kept constant at 1mL (ratio: 1.25E-3). Time for adsorption was increased to 30 seconds.

Mass flows using different types of filter

The impact of two different filters, a 0.45 μm pore sized Sartorious Minisart filter and a 0.20 μm Whatman Acrodisc LC13 filter on the analysis of a 2% (v/v) 2-butanol solution was investigated. The changes to signal intensity (quantitative changes) and signal pattern (qualitative changes) of a CP e-nose (ST214) were studied. For comparison, ROW was also analysed either with no filter or the 0.45 μm filter. Eight replicates per experimental setup were taken. The temperature was kept constant at 25°C.

Selected ion flow tube mass spectrometry (SIFT-MS) was also used to investigate quantitative and qualitative changes in samples after passage through a filter. SIFT-MS is a real-time mass spectrometry method that provides quantitative data on compounds present. Full details are given in [23]. Duplicate sample bags containing 0.5 mL of bovine serum and filled with 0.7 L hydrocarbon-free air were analysed with and without a 0.20 μm PVDF filter (Whatman, Acrodisc LC13). The concentrations of key marker compounds such as water, acetone, methanol and ammonia were measured.

2.5 E-nose comparison and characterisation

Conducting polymer e-nose characterisation

Variability of the two CPs from the same production batch (both ST214 e-noses manufactured under the same conditions within the same period of time) across four time points over a day was assessed using bovine serum sampled statically. Small bags (0.8 L) were used for each time point and device. The serum sample volume was 0.9 mL which leads to a corresponding sample to headspace volume ratio of 1.2E-3.The temperature was maintained at 25°C. After incubating each bag for 15 mins at 25°C to establish an equilibrium, the bags were subsequently analysed (7 replicates) every two hours on four occasions beginning at 10:00 to assess the following factors:

- Device (describing variance caused by the e-noses)
- Time point (describing variance caused across time)
- Size of bag (describing differences according to bag size)

- Replicates (describing differences due to repetitive sampling across one analysis)

MOSFET/ MOS e-nose characterisation

In order to characterise the MOSFET/MOS e-noses, the variation caused by different sample volumes and position around the carousel was investigated. Twelve vials containing six different sample volumes (50 µL, 100 µL, 150 µL 200 µL, 250 µL and 500 µL) of bovine serum were sampled twice to test reproducibility. After 15 minutes equilibration samples were analysed randomly with an air flow of 60 mL per minute. The adsorption time was 30 seconds.

The two MOS/MOSFET e-noses were also used to investigate differences between the devices and across time to compare characteristics with the CP e-noses. 200 µL of bovine serum was pipetted into six vials per time point (sample volume to vial volume ratio: 6.7E-3). Three vials were analysed randomly at four different time points starting at 10:30 in 2h intervals. The adsorption phase time was 30 seconds. Differences caused by devices, over time, and replicates were investigated.

2.6 Statistical analysis

In order to determine the influence of a parameter on the sensor response (divergence, maximum amplitude of signal in sample analysis) univariate data analysis techniques were used. Linear regression and multifactor ANOVA were performed using the statistical package SPSS (version 11.5, 2002). The level of significance for these analyses was $P \leq 0.05$ and the impact of each parameter on the result was given as t-value. For e-nose comparison, Principle Component Analysis (PCA) was performed using Matlab on the raw datasets and on auto-scaled as well as mean-centred data.

Auto-scaling scales all values of a column of the data matrix by its standard deviation which eliminates different ranges or magnitudes of data. *Mean-centring* eliminates constant offsets by subtracting the column mean from the single values.

3. Results

3.1 Influence of temperature

Multifactor ANOVA and linear regression were performed on data obtained from the ST214 e-nose. The influence of temperature on signal traces of ROW and bovine serum was investigated.

Sensors 6 and 8 revealed significant differences between ROW and bovine serum (Figure 1). Serum samples led to more positive values in comparison to ROW; similar tendencies were observed for the other sensors but were not statistically significant.

These sensors were also influenced by *temperature*, hence a combination of both factors affected the sensor signals. Sensors 10 and 11 were neither influenced by the *substance* nor by *temperature*. The remaining sensors were significantly influenced by *temperature*. An increase in temperature led to more negative values (negative t-factor) so to a more negative change in divergence. Figure 1 illustrates the increase in signal intensity with increasing temperature. Ambient air data are also shown.

[INSERT figure 1 about here]

3.2 Mass flow and filters

Mass flows using different setups

The mass flow rates under different experimental conditions were 180 and 200mL/min respectively when the Luer fitting, flow sensor and the bag fitting were attached to both e-noses (ST114, ST214). The flow rates for static sampling were 98% (ST114) and 97% (ST214) of the original mass flows rates, while for dynamic sampling, rates decreased to 66% for both devices. For dynamic sampling, during the adsorption phase, mass flow varied due to the bubbling effect. The addition of the 0.45µm filter led to a significant drop in flow to 65% and 68% for ST114 and ST214 using static sampling and to 52% and 57% for dynamic sampling.

Mass flow stability

Changes in mass flow were observed testing different *sampling methods* and *devices* for both CP e-noses when analysing eight replicates of a 10ppm butanol solution (ANOVA, $P \leq 0.05$). Constant flow rates were found for replicates 2 to 5 for e-nose 1 (set as 100%) and e-nose 2 but the second e-nose had a 17% higher flow rate. For replicates 1, a minor decrease in flow rates of 4% was found for both e-noses and approximately 72% of the headspace volume was analysed (573 mL e-nose 1 and 576 mL for e-nose 2). Flow rates decreased significantly by 15 and 13%, respectively.

The decrease in mass flow occurred with progressive collapse of the bag.

Mass flows using different types of filter

Appedix 2

Comparing divergences obtained analysing the 2-butanol solution with no filter, a 0.45μm (Sartorious) and a 0.20μm filter (Whatman), significant differences were found for all sensors (ANOVA, $P \leq 0.001$, raw data not shown). The highest sensor responses, expressed as negative divergences, were seen when there was no filter (group 1, replicates 1 to 8). Attaching the 0.45 μm filter, the values became less negative (group 2, replicates 9 to 16); use of the 0.20 μm filter (group 3, replicates 17 to 24) led to lowest changes (see figure 2).

[INSERT figure 2 about here]

Qualitative changes in signal pattern with the use of filters were investigated by comparing the results of the 2-butanol solution with those obtained analysing ROW (figure 3). Using no filter, both solutions could be distinguished while no discrimination was found using the 0.45 μm filter. Furthermore, the biggest discrimination was found using the 0.20 μm filter and the unfiltered butanol solution. Principle Component 1, covering 99.67% of all variance, mainly represented changes due to filter application rather than differences between the solutions.

[INSERT figure 3 about here]

Quantitative data from SIFT-MS analyses of duplicate cattle serum headspace samples showed changes in some key marker compounds. The concentration of water measured changed from a mean of 5.15 % for unfiltered samples to a mean of 4.26% for filtered (0.20 μm Whatman filter) samples (reduction of 17%); ammonia from a mean of 771 parts-per-billion (ppb) for unfiltered and 547 ppb for filtered (reduction of 29%); methanol from a mean of 424 ppb for unfiltered and 962 ppb filtered (increase of 126%). Acetone remained unchanged, at a mean of 656 ppb unfiltered and 680 ppb unfiltered. Clearly, the addition of a filter changed the concentration of many volatile compounds.

3.3 E-nose comparison and characterisation

Conducting polymer e-nose characterisation

Sensor responses from all four time points obtained from similar CPs of the same type and manufacturer were analysed in Matlab using PCA. Principal Components 1 and 2 covered 99.10% of all variance. A clear discrimination between both e-noses was found using auto-scaled data. Without pre-treatment, data could not be interpreted.

Appedix 2

The difference between the devices covered 84.03% of all variance, and 15.07% of all changes were due to changed divergences over time using static sampling (see figure 4).

[INSERT figure 4 about here]

Considering PCA plots, time points 1 and 3 as well as 2 and 4 were grouped together for CP e-nose 1. Similar results were found for e-nose 2 however, responses of time point 4 were more scattered.

SPSS analysis using linear regression and multifactor ANOVA ($P \leq 0.05$) showed that the *device*, the *time points* and the *replicates* were potential factors influencing the result.

The factors *device* and *time point* had the biggest influence while factor *replicates* was almost not statistically relevant, except for two sensors.

Considering the responses over time for each device separately, divergences changed throughout the day for both e-noses (shown for one e-nose in figure 5). Sensor responses were highest at time point 1, decreased at time point 2 increased and decreased again for time points 3 and 4. Responses obtained at time points 1 and 3 as well as 2 and 4 were closest, respectively. Results for the other e-nose showed a similar trend. This result confirmed PCA findings.

[INSERT figure 5 about here]

For almost all sensors and both devices the temporal variation appeared sinusoidal, and there was a tendency for responses to decline across all time points and for both e-noses; demonstrated using SPSS multifactor ANOVA ($P \leq 0.05$)..

MOSFET/ MOS e-nose characterisation

It was found that the factor *replicate* had a statistically significant influence on some sensor responses only ($P \leq 0.05$). A few MOSFET sensors (FE103A, 104A, 105A, 104B) showed a significant difference between the first and the second reading of the same vial. Most of the sensors showed a similar tendency of a decrease from *replicate 1* to *replicate 2* although these results were not significant.

The *sample volume* had no influence on the results within the volume range evaluated (50 to 500µL).

Appedix 2

Analysing the averaged values across all *time points* and *replicates* for the two MOSFET/ MOS e-noses, there was a significant difference between the two. Except for the sensor FE104B, which was a MOSFET sensor, all other sensors demonstrated significantly different responses between the two devices. The first e-nose showed higher responses for MOSFET and MOS sensors (see figure 6).

[INSERT figure 6 about here]

The standard deviation from the base line for the first e-nose varied between 0.43 for the sensor MO111 (3.8% change in sensor response) and 35.22 for MO115 (16.4% change in sensor response). For the second e-nose, standard deviation was between 2.3 for the sensor MO112 and 19.0 for MO116 (54.0% change in sensor response). However, sensors with small responses varied up to 201%. The averaged standard deviation from the baseline was 9.9 for the first e-nose and 7.8 for the second device.

With the first e-nose, most of the MOSFET-generated values showed significant variation between the *replicates* using different vials, while samples analysed at different *time points* led to almost no changes. In contrast, the MOS sensors showed almost no changes between the *replicates* but data analysis indicated a significant decline in sensor responses over *time* for most sensors (multifactor ANOVA, $P \leq 0.05$). Only two sensor responses increased across time (MO111, MO112).

Analysing the same factors for the second e-nose, it was found that while the responses for all sensors were not affected by any variation in *replicates*, the traces of almost all MOS sensors were influenced by *time*. Generally, the responses decreased over the course of the day. In three cases the responses remained constant (MO101, MO102, MO104) and in another two cases the values increased across time (MO115, MO116).

The MOSFET sensors generally exhibited a decrease of signal intensity across time but this tendency was only statistically significant in two cases (FE105A, FE103B).

4. Discussion.

4.1 The effect of temperature

Temperature changes had a major effect on sensor response. Even the two sensors able to discriminate between ROW and serum samples were significantly influenced by temperature. Changing temperature led to both qualitative and quantitative changes since a higher temperature leads to a greater concentration of volatiles in the

headspace, which will generate different sensor responses. Such large changes in response due to temperature may mask the differences between samples, preventing discrimination or classification using multivariate techniques.

Changes in the physico-chemical properties of conducting polymers or their monomers when exposed to different temperatures have been described in literature. Impedance was decreased with an increase in humidity and temperature [10-13]. The temperature optimum was found to be 25°C [7, 24].

A temperature controller would add weight and bulk to the device. Some work has been carried out to construct a flow cell for temperature- and flow controlled analysis. Attempts were made to optimise the positions of the sensors on a heating block and the incoming gas so that headspace reacted in the same way with optimised sensors [25, 26]. With these changes in design, the most significant e-nose problems might be resolved to produce improved signal responses in terms of stability, reproducibility, response time and amplitude, all without losing e-nose portability.

4.2. Mass flow and filters and stability

Results of this study indicate that the previous assumption of an estimated mass flow rate of 200 ml/min using similar CP noses is not valid since flow rates change depending on sampling methods and filter usage. The ST114 had a flow rate of 180 mL/min and the ST214 of 200 and 235mL/min, respectively. A drop of 75% compared to the stable flow rates over replicates 2 to 5 was observed using dynamic sampling and the additional application of a 0.45 μm pore sized filter which was previously used in other studies [27, 28]. Bubbling air through the liquid sample led to variation and a decrease in pressure in the vial which was also found by others [29].

Dilution of the headspace and a non-equilibrium state led to variation between replicates also reported by others [30], while some authors obtained the best discrimination with dynamic sampling but better sensitivity using static sampling [6]. The filter was an additional resistance in flow rate for CP e-noses. In contrast, static sampling provided stable mass flow rates and equilibrium states because of the collapse of a sealed bag.

Therefore, static sampling can significantly improve e-nose responses by minimising variation between and over replicates. In trace gas analysis (lower ppm to ppb), stability is crucial as even small changes may lead to non-discrimination.

Qualitative and quantitative changes in signals were found when filters were used. The change was dependent on the pore size of the filters, which protect the conducting polymer sensing surface of the e-nose from being changed by a high extent of water vapour. Mass flow rates decreased with a decline in pore size. However removal of water vapour will also remove other volatile compounds and leads to a change in headspace composition. This was confirmed using CP e-noses combined with PCA data analysis (clustering according to filter application) and analysing quantitative SIFT-MS data. Since most biological samples contain water (e.g. blood/serum, urine), filters have been widely used for protection of the sensors (0.45 μm PTFE, Whatman/Hepavent) [26-29]. However, these results indicate that using a filter may not a valid approach since discrimination between 2-butantol and ROW was not possible once a 0.45 μm filter was introduced and 98% of all variation was due to different filter application.

Headspace was significantly changed.

Results of SIFT-MS analysis showed qualitative and quantitative changes in sample composition with the addition of a filter. The claimed advantage of using a filter for sensor protection [21, 30] is highly questionable since sensor response patterns change with the use of a filter showing changes in methanol or ammonia and therefore discrimination is made difficult.

4.3 E-nose characterisation and comparison

Conducting polymer e-nose characterisation

Temporal changes in sensor response have been reported elsewhere [12, 13]. The sine wave-like changes described here across the day are possibly due to semi-reversible changes on the sensor surface. Since this effect was observed for both CP e-noses using static bag sampling and confirmed using multifactor ANOVA and PCA it seems to be a systematic problem which is not associated with dilution of samples nor a nonequilibrium state. It appears to be the result of semi-reversible desorption. Certain molecules might be desorbed before the software determines a new baseline (offset) for acquisition, meaning that in later analyses, responses cannot develop the

same way since some binding sites are still occupied. So the changes in comparison to the baseline are different from the preceding analysis. Occupation of binding sites may come from water molecules of ambient air interacting with dopants or where headspace molecules were not desorbed properly. Purging with an inert gas such as nitrogen has been used by some authors to improve desorption and repeatability [7, 31]. The reasons why responses in this study were in the form of a sine wave, however, remain unclear. This may be evidence of memory effects which have been described by several authors [10, 19].

Further analyses are necessary to elucidate the cause of this effect. A comparison of sensor responses of similar CP e-noses produced by the same manufacturer was not possible due to different flow rates (180 mL/min to 235 ml/min) and qualitatively different responses, although the same substances were analysed under identical conditions (see section 2.5 and 3.3 CP e-nose comparison).

MOSFET/ MOS e-nose characterisation

The sensor response patterns obtained from two MOSFET/ MOS e-noses were different when sampling the same headspaces. However, the variation was random and not significant; high standard deviation was found across various sample volumes (50 to 500 μL) but not across vials in the carousel. This was also confirmed in the literature [32, 33]. Almost all MOSFET sensors of e-nose 1 were found to be replicate dependent and almost all MOS sensors varied over time. The fact that different values were obtained (independent to the sample volume) when vials were sampled twice contradicts the findings of others that dual needle sampling was a reliable sampling method [34].

The reason for this could be the simple fact that after sampling, the liquid had been "outgassed" and lost a proportion of its volatile compounds.

Similar to the CP e-noses, temporal changes in MOS sensors indicate an insufficient desorption process; the new baseline being determined with molecules still attached to sensor surface resulting in a different response.

Comparison between e-nose devices

Comparing e-noses of the same type under different experimental setups, a lack of comparability was a surprising and worrying observation in this study and points to important manufacturing and engineering inconsistencies. Lack of reproducibility of

sensor responses was exacerbated by the sampling principle, which mainly affected the portable CP e-noses.

Despite the fact CP e-nose and the MOSFET/ MOS e-nose were not comparable due to different sensing principles, there were changing sensor responses for both types of e-noses over time which were due to semi-reversible adsorption. Standardisation and improvement of production methods towards reliable and comparable sensors is urgently required, particularly for conducting polymer devices but also for MOSFET/ MOS sensors.

5. Conclusions

- Static sampling using the conducting polymer e-nose is better than dynamic sampling because of stable mass flow which leads to better repeatability.
- Filters are not recommended as they quantitatively change the mass flow rate and qualitatively change the sensor pattern which may lead to nondiscrimination.
- Temperature and mass flow rate controlled sensor chambers could enhance the repeatability of e-noses and therefore improve discrimination and classification using multivariate data analyses.
- Incomplete desorption from sensors affects both conducting polymers and MOS sensors. Longer purging with inert gases may solve this problem and lead to more consistent responses over time.
- Because of the sources of variation discussed in this paper, multivariate data analysis techniques are only suitable for data obtained with stable e-nose systems; otherwise methodological variation may mask biological differences between samples. To overcome this problem, some authors suggest an absolute calibration for e-noses [3, 35] or a compensation at the data analysis level [9] but we believe in improving methodology first, since the calibration or reference points also shift but not necessarily in a linear or predictable way (due to changes in temperature, humidity, sampling methodology and memory effects). Improvements in devices are advisable such as static sampling in a controlled ambient environment, the introduction of controlled sensor chambers, and automated purging. This should minimise methodological variation and enhance the classification and discrimination of samples based on the biological variation which is present in all types of sample

[36, 37]. Such improved and optimised e-noses may then be capable of fulfilling the considerable promise of this technology.

Acknowledgements

The authors acknowledge the UK *Department for Environment, Food and Rural Affairs* (DEFRA) for funding this project and the *Institute of Molecular Pathogenesis* at the 'Friedrich–Loeffler–Institut' (FLI, Jena, Germany) for providing bovine serum samples.

References

[1] P. Mielle, Electronic noses: Towards the objective instrumental characterization of food aroma, Trends in Food Science and Technology 7 (1996) 432-438.

[2] J.W. Gardner, Hyun Woo Shin, E.L. Hines, An electronic nose system to diagnose illness, Sens. Actuators, B, Chem 70 (2000) 19-24.

[3] A.P.F. Turner, N. Magan, Electronic noses and disease diagnostics, Nature Reviews Microbiology 2 (2004) 161-163.

[4] S. Nimmermark, Use of electronic noses for detection of odour from animal production facilities: a review, Water Science and Technology 44 (2001) 33–41.

[5] I.A. Casalinuovo, D. Di Pierro, M. Coletta, P. Di Francesco, Application of Electronic Noses for Disease Diagnosis and Food Spoilage Detection, Sensors 6 (2006) 1428- 1439.

[6] J. Lozano, J.P. Santos, J. Gutierrez, J. , M.C. Horrillo, Comparative study of sampling systems combined with gas sensors for wine discrimination , Sens. Actuators, B, Chem 126 (2007) 616-623.

[7] M. Chimenti, D. De Rossi, F. Di Francesco, C. Domenici, G. Pieri, G. Pioggia, O. Salvetti, A neural approach for improving the measurement capability of an electronic nose, Measurement Science and Technology 14 (2003), 815-821.

[8] C. DiNatale, A. Macagnano, E. Martinelli, R. Paolesse, G. D'Arcangelo, C. Roscioni, A. Finazzi-Agrò, A. D'Amico, Lung cancer identification by the analysis of breath by means of an array of non-selective gas sensors, Biosensors and Bioelectronics 18 (2003), 1209-1218.

[9] C. Di Natale, E. Martinelli, A. D'Amico, Counteraction of environmental disturbances of electronic nose data by independent component analysis , Sens. Actuators, B, Chem 82 (2002), 158-165.

[10] S. Ampuero, J.O. Bosset, The electronic nose applied to dairy products: a review, Sens. Actuators, B, Chem 94 (2003), 1-12.

[11] D.J. Strike, M.G.H. Meijerink, M. Koudelka-Hep, Electronic noses – A mini-review, Fresenius J Anal Chem 364 (1999) 499–505.

[12] A., Nake, B. Dubreuil, C. Raynaud, T. Talou, Outdoor in situ monitoring of volatile emissions from treatment plants with two portable technologies of electronic noses, Sens. Actuators, B, Chem 106 (2005) 36-39.

[13] K.R. Kashwan, M. Bhuyan, Robust Electronic-Nose System with Temperature and Humidity Drift Compensation for Tea and Spice Flavour Discrimination, IEEE Asian Conference on Sensors and International Conference on New Techniques and Biomeldical Research 2005

[14] D. Hodgins, The development of an electronic 'nose' for industrial and environmental applications, Sens. Actuators, B, Chem 27 (1995) 255-258.

[15] O.N. Timofeeva, B.Z. Lubentsov, Y. Z. Sudakova, D.N. Chernyshov, M.L. Klddekel, Conducting polymer interaction with gaseous substances: I Water, Synthetic Metals 40 (1991) 111-116.

[16] B.Z. Lubentsov, O.N. Timofeeva, S. Saratovskikh, V. Krinichnyi, A. Pelekh, V. Dmitrenko, M.L. Khidekel, The study of conducting polymer interaction with gaseous substances IV. The water content influence on polyaniline crystal structure and conductivity, Synthetic Metals 47 (1992) 187-192.

[17] R. Fend, C. Bessant, A.J. Williams, A.C.Woodman, Monitoring haemodialysis using electronic nose and chemometrics, Biosensors and Bioelectronics 19 (2004) 1581- 1590.

[18] A.K. Pavlou, N. Magan, C. McNulty, J.M. Jones, D. Sharp, J. Brown, A.P.F. Turner, Use of an electronic nose system for diagnoses of urinary tract infections, Biosensors and Bioelectronics 17 (2002) 893-899.

[19] A.L. Kukla, Y.M. Shirshov, S.A. Piletsky, Ammonia sensors based on sensitive polyaniline films, Sens. Actuators, B, Chem 37 (1996) 135-140.

[20] A.C. Partridge, Y.P. Harris, M.K. Andrews, High sensitivity conducting polymer sensors, Analyst 121 (1996) 1349-1353.

[21] A.D. Wilson, D.G. Lester, C.S. Oberle, Development of Conductive Polymer Analysis for the Rapid Detection and Identification of Phytopathogenic Microbes, Techniques 94 (2004) 419-431.

[22] E.L. Kalman, A. Löfvendahl, F. Winquist, I. Lundström, Classification of complex gas mixtures from automotive leather using an electronic nose, Anal Chimica Acta 403 (2000) 31-38.

[23] D. Smith, D. P. Španěl, P. Selected ion flow tube mass spectrometry (SIFT-MS) for on-line trace gas analysis, Mass Spectrometry Reviews 24 (2005) 661-700.

[24] J.H. Cho, J.B. Yu, J.S. Kim, S.O. Sohn, D.D. Lee,J.S. Huha, Sensing behaviors of polypyrrole sensor under humidity condition, Sens. Actuators, B, Chem 108 (2005) 389-392.

[25] M. Facitelli, A. Benassi, F. Di Francesco, C. Domenici, L. Marano, L., G. Pioggia, Fluid dynamic for simulation of a measurement chamber for electronic noses, Sens. Actuators, B, Chem 85 (2002) 166-174.

[26] F. Di Francesco, M. Falcitelli, L. Marano, G. Pioggia, A radially symmetric measurement chamber for electronic noses Sens. Actuators, B, Chem 105 (2005) 295-303.

[27] R. Fend, R. Geddes, S. Lesellier, H.-M. Vordermeier, L.A.L. Corner, E. Gormley, E. Costello, R.G. Hewinson, D.J. Marlin, A.C. Woodman, M.A. Chambers, Use of an Electronic Nose To Diagnose Mycobacterium bovis Infection in Badgers and cattle, J Clin Microbiology 43 (2005) 1745-1751.

[28] N. Magan, A.K. Pavlou, I. Chrysanthakisa, Milk-sense: a volatile sensing system recognises spoilage bacteria and yeasts in milk. Sens. Actuators, B, Chem 72 (2001) 28-34.

[29] S. Ampuero, S. Bogdanov, J.-O. Bosset, Classification of unifloral honeys with an MS-based electronic nose using different sampling modes: SHS, SPME and INDEX, Eur Food Res Technology 218 (2004) 198-207.

[30] T.H.H. Misselbrook, P.J. Hobbs. K.C. Persaud, Use of an Electronic Nose to Measure Odour Concentration Following Application of Cattle Slurry to Grassland, J Agr Eng Res 66 (1997) 213-220.

[31] W. Bourgeois, R.M. Stuetz, Use of a chemical sensor array for detecting pollutants in domestic wastewater, Water Research 36 (2002) 4505-4512.

Appedix 2

[32] S. Benedetti, C. Pompei, S. Mannino, Comparison of an Electronic Nose with the Sensory Evaluation of Food Products by "Triangle Test", Electroanalysis 16 (2004) 1801-1805.

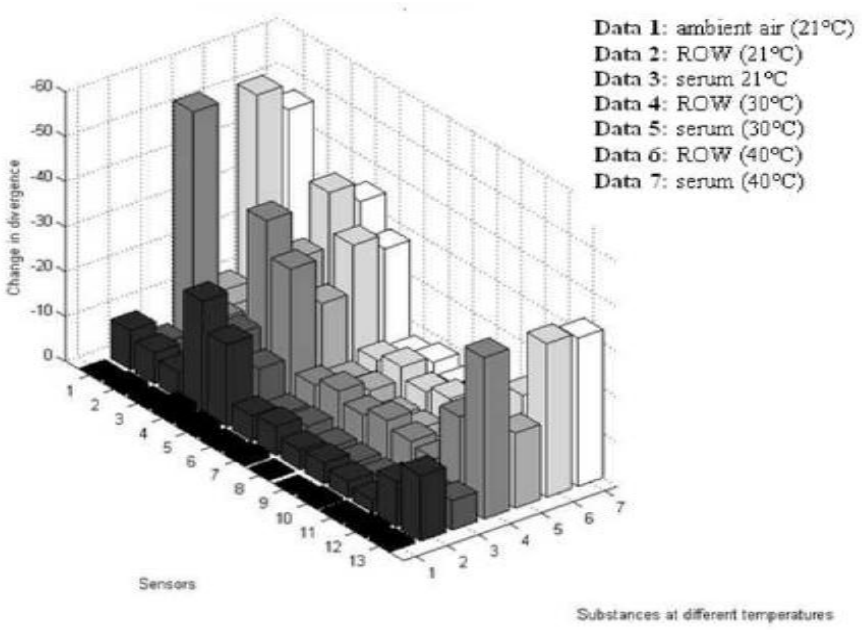
[33] M. Morvan, T. Talou, J.-F. Beziaub, MOS–MOSFET gas sensors array measurements versus sensory and chemical characterisation of VOC's emissions from car seat foams, Sens. Actuators, B, Chem 95 (2003) 212-223.

[34] M. Markelov, O.A. Bershevits, Methodologies of quantitative headspace analysis using vapor phase sweeping, Anal Chimica Acta 432 (2001) 213-227.

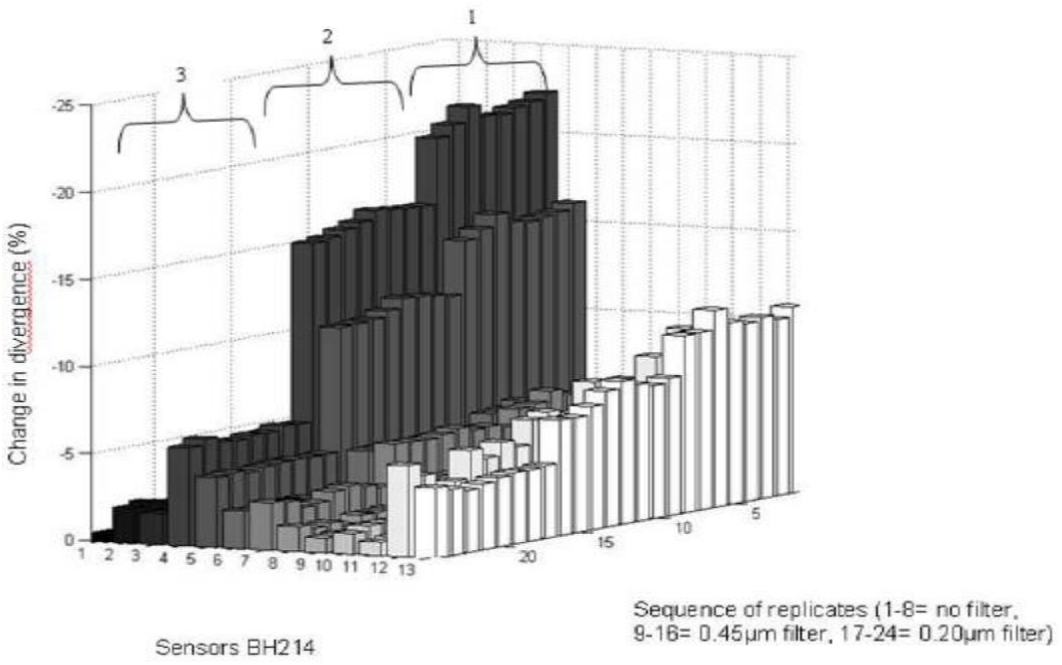
[35] T.D. Gibson, O. Prosser, J.N. Hulbert, R.W. Marshall, P. Corcoran, P. Lowery, E.A. Ruck-Keene, S. Herond, Detection and simultaneous identification of microorganisms from headspace samples using an electronic nose, Sens. Actuators, B, Chem 44 (1997) 413-422.

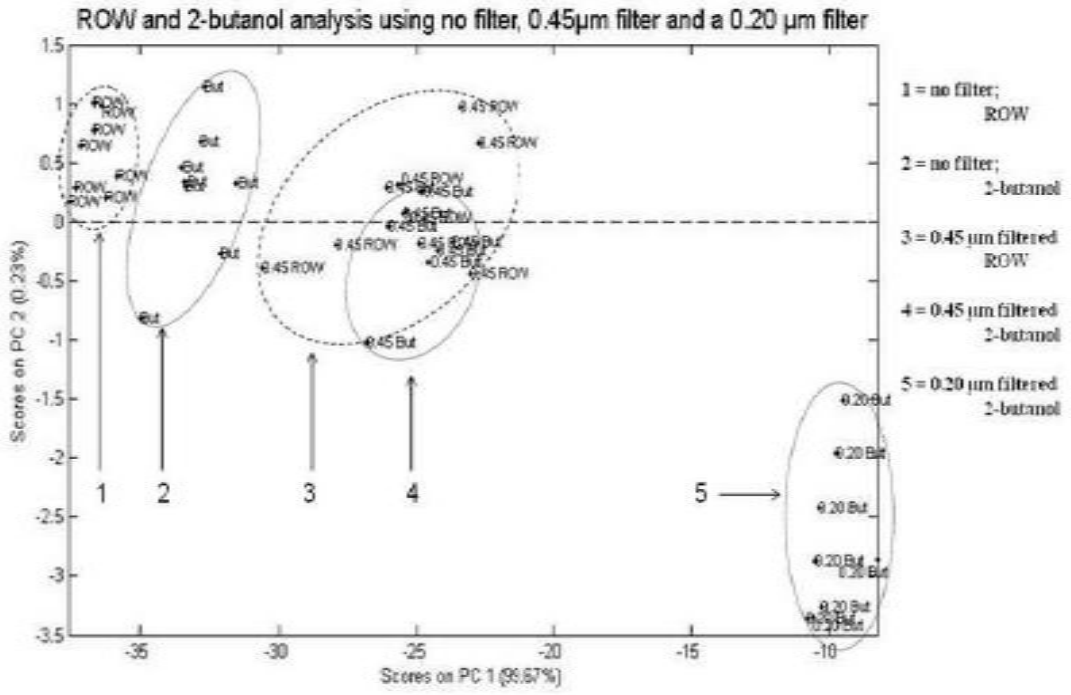
[36] W.J. Harper, W. J., The strengths and weaknesses of the electronic nose Advances in Exp Medicine and Biology 488 (2001) 59-71.

[37] P. Van Geloven, M. Honore, J. Roggen, S. Leppavuori, T. Rantala, The influence of relative humidity on the response of tin oxide gas sensors to carbon monoxide, Sens. and Actuators, B, Chem 4 (1991) 185-188.

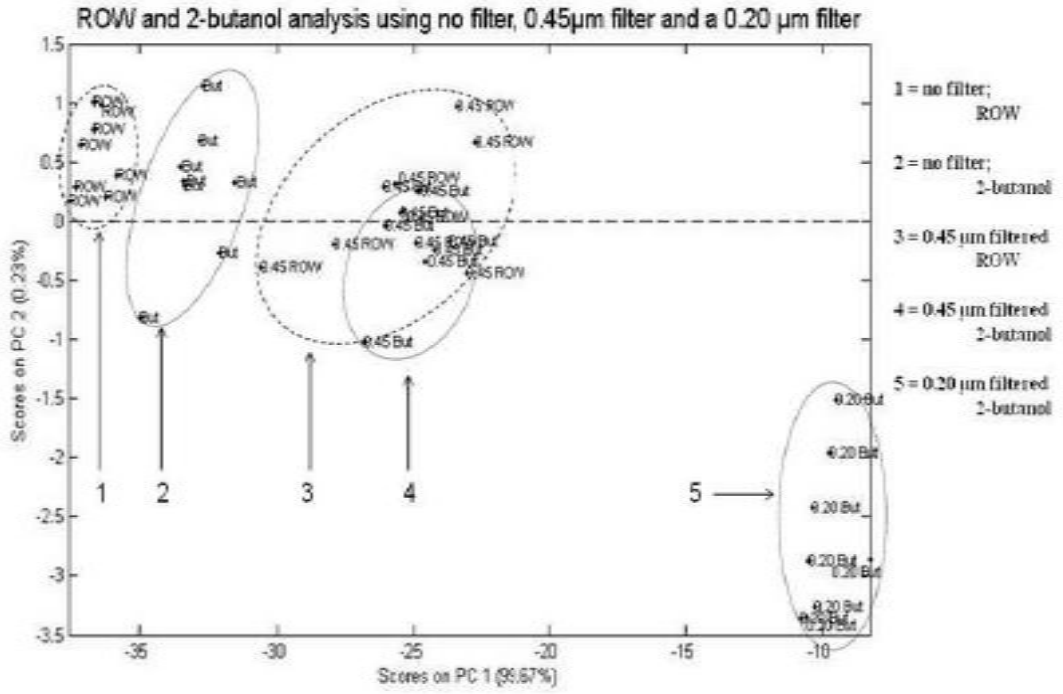


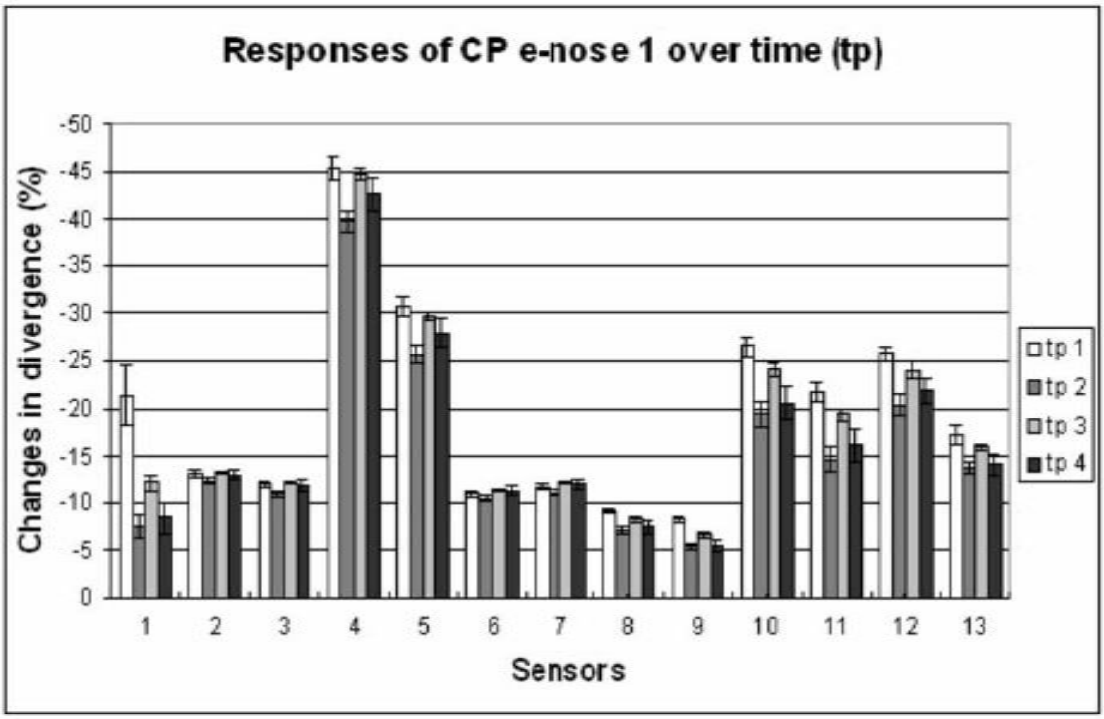
Appedix 2


Change in divergence (%)
Sensors BH214
Sequence of replicates (1-8= no filter,
9-16= 0.45µm filter, 17-24= 0.20µm filter)

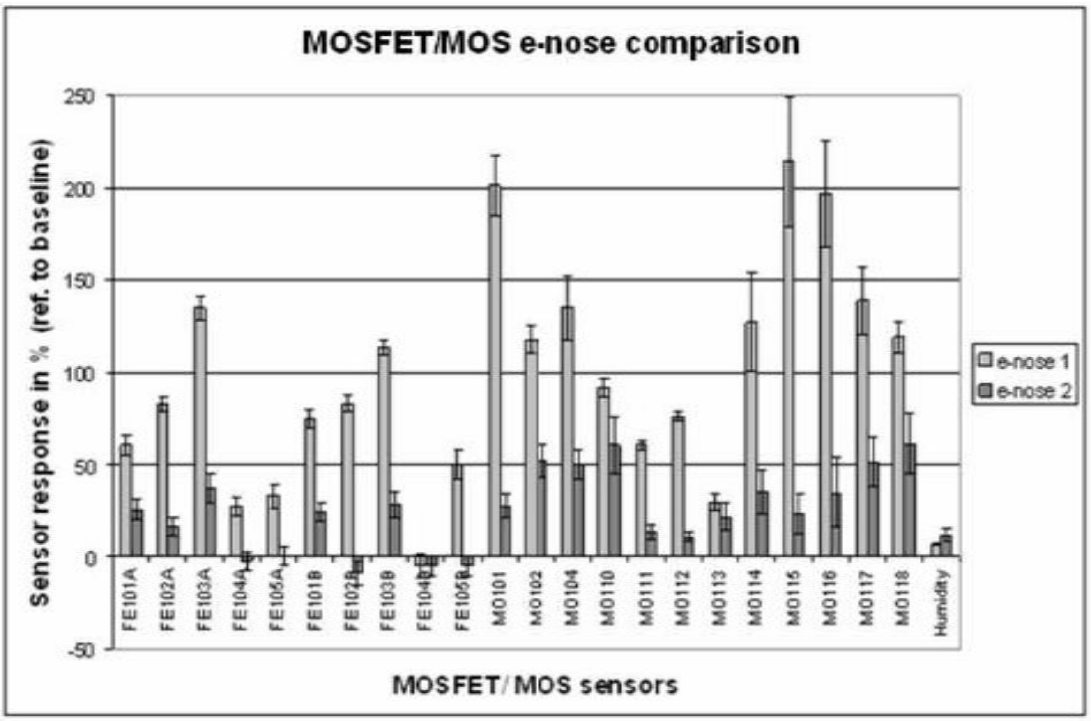

ROW and 2-butanol analysis using no filter, 0.45µm filter and a 0.20 µm filter
Scores on PC 2 (0.23%)
Scores on PC 1 (99.67%)
1 = no filter; ROW
2 = no filter; 2-butanol
3 = 0.45 µm filtered ROW
4 = 0.45 µm filtered 2-butanol
5 = 0.20 µm filtered 2-butanol

Appedix 2


ROW and 2-butanol analysis using no filter, 0.45µm filter and a 0.20 µm filter
Scores on PC 2 (0.23%)
Scores on PC 1 (99.67%)
1 = no filter; ROW
2 = no filter; 2-butanol
3 = 0.45 µm filtered ROW
4 = 0.45 µm filtered 2-butanol
5 = 0.20 µm filtered 2-butanol
1
2
3
4
5


Responses of CP e-nose 1 over time (tp)
Changes in divergence (%)
Sensors
tp 1
tp 2
tp 3
tp 4

Appedix 2


MOSFET/MOS e-nose comparison
Sensor response in % (ref. to baseline)
250
200
150
100
50
0
-50
FE101A
FE102A
FE103A
FE104A
FE105A
FE101B
FE102B
FE103B
FE104B
FE105B
MO101
MO102
MO104
MO110
MO111
MO112
MO113
MO114
MO115
MO116
MO117
MO118
Humidity
MOSFET/ MOS sensors
e-nose 1
e-nose 2