Application of Cluster Analysis to High-Throughput Multiple Data Types

Gordon John Cunningham

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

School of Chemistry
University of Glasgow

Supervisor: Professor Chris Gilmore

June 2011

University of Glasgow

WestCHEM

DECLARATION

The thesis has been written in accordance with the University regulations and all work presented is original and performed by the author unless otherwise stated and referenced in the text.

Gordon J Cunningham

ABSTRACT

PolySNAP is a program used for analysis of high-throughput powder diffraction data. The program matches diffraction patterns using Pearson and Spearman correlation coefficients to measure the similarity of the profiles of each pattern with every other pattern, which creates a correlation matrix. This correlation matrix is then used to partition the patterns into groups using a variety of cluster analysis methods. The original version could not handle any data types other than powder X-ray Diffraction. The aim of this project was to expand the methods used in PolySNAP to allow it to analyse other data types, in particular Raman spectroscopy, differential scanning calorimetry and infrared spectroscopy data. This involves the preparation of suitable compounds which can be analysed using these techniques. The main compounds studied are sulfathiazole, carbamazepine and piroxicam. Some additional studies have been carried out on other datasets, including a test on an unseen dataset to test the efficacy of the methods.

The optimal method for clustering any unknown dataset has also been determined.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

ACKNOWLEDGEMENTS

UNIVERSITY OF GLASGOW

Thanks to Professor Chris Gilmore for the opportunity to work on the PolySNAP project and Professor Chick Wilson for the use of his lab for the lab-based part of my research.
I would also like to thank Dr Gordon Barr for his assistance with the PolySNAP software, Dr Lynne Thomas for her assistance with the Departmental Diffractometers and with other lab issues, Mr Andy Monaghan for his assistance with DSC and TGA, Mrs Kim Wilson for assistance with IR and Mr Stuart Mackay for IT support.
Thanks to Dr Andy Parkin, who sadly passed away during my research, for all of his assistance and contributions to my research.
Lastly thanks to all the people in both the Gilmore group and Chicklets, that I shared an office and lab with during my research and to all the other people in the Chemistry Department.

PHARMORPHIX

Thanks to Professor Chris Frampton, Dr Suzanne Buttar and all others at Pharmorphix for the use of their instruments for data collection and all other assistance rendered.

UNIVERSITY OF STRATHCLYDE

Thanks to Professor Duncan Graham and Dr Aaron Hernandez-Santana for the use of their Raman instrument and to Dr Alistair Florence for his assistance in selecting appropriate materials for study.

CHAPTER 1 BACKGROUND AND PREVIOUS WORK

1.1 HIGH THROUGHPUT DATA COLLECTION

High throughput data collection techniques are capable of collecting large amounts of data, in excess of 500 powder diffraction patterns or spectra, in a single day. These techniques are useful in a variety of fields where large numbers of samples are produced and need to be analysed, for example the pharmaceutical industry where large numbers of potential drug precursors need to be analysed, in searches for polymorphs or in chemical manufacturing companies where samples are taken regularly to test their purity. This volume of data could not be practically analysed by hand in any reasonable timeframe so methods have been developed to allow rapid analysis of this data. These techniques were initially[1-7] developed for X-ray powder diffraction data, however the techniques can be extended to other types of data in which high-throughput data can be collected, for example Raman or infrared (IR) data. These datatypes can allow for further analysis of the materials being studied and can prove useful for identifying similarities between materials should the X-ray analysis prove inconclusive.

This thesis will study the extension of the X-ray analysis techniques to the additional datatypes of Raman spectroscopy, infrared spectroscopy, differential scanning calorimetry and thermal gravimetric analysis.

1.2 POLYSNAP

PolySNAP[1-7, 26] is a computer program used for the analysis of high-throughput data supplemented, if required, by numerical data. The program matches 1-dimensional patterns, using the Pearson and Spearman correlation coefficients, to measure the similarity of each pattern with every other pattern. This creates a correlation matrix which is then used to partition the patterns into groups of similar patterns using a variety of cluster analysis methods. The program has two main forms of functionality: automatic analysis mode and manual analysis mode.

1.2.1 AUTOMATIC ANALYSIS MODE FUNCTIONALITY

Automatic analysis mode is used to match a number of samples, minimum 3, maximum 2000, using up to four datatypes. The analysis method for automatic analysis mode works as outlined in the flowchart in Figure 1.

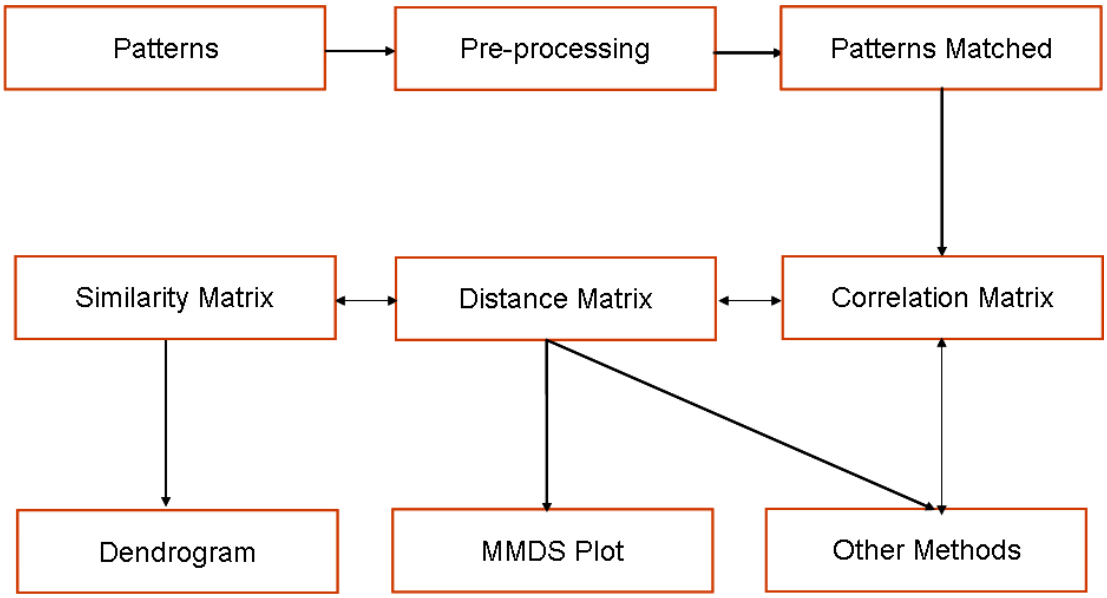

Figure 1 - Flow Chart of PolySNAP methods

All patterns are first read in. The patterns then have any requested pre-processing applied to them (see Section 1.25), and for powder patterns they are checked to see if they are crystalline. The check for crystallinity occurs as follows:

1. The background for each pattern is estimated and its intensity integrated.
2. Non-background intensity is estimated
3. Diffraction peaks are located
4. The ratio of background to non-background intensity is estimated. If it falls below 3% then the material is determined to be non-crystalline. Also if less than 3 peaks are located then the material is determined to be non-crystalline.

All samples are then matched against all other crystalline samples, using two matching methods.[1] The first of these methods is the Pearson correlation coefficient[30], Equation 1, where two diffraction patterns, each containing n measured points, $n[(x_1,y_1),\ldots(x_n,y_n)]$, are used, along with the mean of their intensities over the full diffraction pattern, to generate the correlation coefficient r_{xy}. $\bar{x}$ and $\bar{y}$ are the means of the intensities taken over the full range of the diffraction pattern

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\left[\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2\right]^{1/2}}$$

Equation 1- Pearson coefficient

The second method uses the Spearman correlation coefficient[20] which is shown in Equation 2, where two diffraction patterns, each containing *n* measured points, $n[(x_1,y_1),\ldots(x_n,y_n)]$, are transformed into ranks $R(x_i)$ and $R(y_i)$. These ranks are used to give a correlation coefficient ρ_{xy}.

Ranking is a numerical transformation that can be carried out on both numerical and non-numerical data. When numbers are ranked they are replaced by the position they would obtain if the numbers were to be sorted into numerical order. For example the numbers 2.8, 1.2, 4.5 and 2.9 would be replaced with 2, 1, 4 and 3. By applying ranking to a dataset, it is possible to convert a complex series of information into a regular series, thus allowing easier manipulations to be carried out.

$$\rho_{xy} = \frac{\sum_{i=1}^{n} R(x_i)R(y_i) - n\left(\frac{n+1}{2}\right)^2}{\left[\sum_{i=1}^{n} R(x_i)^2 - n\left(\frac{n+1}{2}\right)^2\right]^{1/2}\left[\sum_{i=1}^{n} R(y_i)^2 - n\left(\frac{n+1}{2}\right)^2\right]^{1/2}}$$

Equation 2 – Spearman rank order coefficient

A correlation matrix is a matrix containing all of the correlation coefficients from the pattern matching. Correlation coefficients always have a value between -1.0 and 1.0. A correlation of 1.0 means that the 2 patterns being matched are identical, a correlation of 0.0 means that the patterns do not match in any way while a correlation of -1.0 means that the peaks match but all have inverse intensity.

Figure 2[1] shows an example of each of these possibilities. The example for Figure 2C was created by inverting the data on a pattern and matching it against itself.

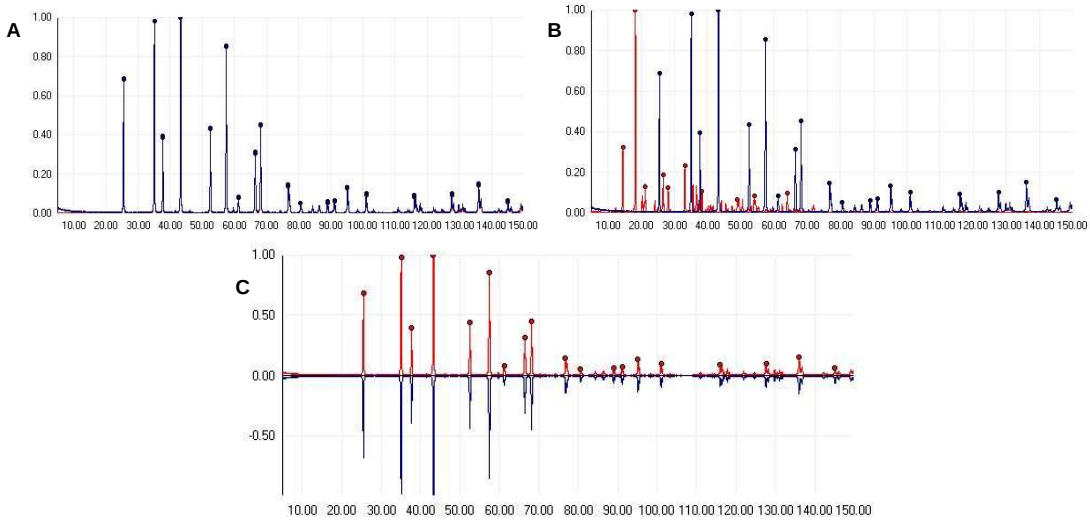

Figure 2 - A - Correlation coefficient of +1 Between Powder Patterns, B - Correlation coefficient of ~0 Between Powder Patterns, C - Correlation Coefficient of -1 Between Powder Patterns.

The correlation matrix, $\boldsymbol{\rho}$, is then converted to a distance matrix, **d**, using Equation 3.

$$\mathbf{d} = 0.5(1-\boldsymbol{\rho})$$

Equation 3 - Distance matrix equation

The elements of the distance matrix are always in the range of 0.0 to 1.0. A correlation of -1.0 will convert to a distance of 1.0 while a correlation of 1.0 will convert to a distance of 0.0. The smaller the distance value between two patterns the more similar they are.
If two patterns have a high positive correlation coefficient (towards 1.0) they are said to be similar. If two patterns have a high negative correlation coefficient (towards -1.0) they are said to be dissimilar. Equation 4 and 5 are used to calculate dissimilarity.

$$\boldsymbol{\delta}_{ij} = 1 - d_{ij} / d_{\max}$$

Equation 4 - Similarity Matrix

Where $\boldsymbol{d}_{\max}$ is the maximum distance in matrix $\boldsymbol{d}$ between two cluster i,j.

$$\delta_{ij} = d_{ij} / d_{ij}^{\max}$$

Equation 5 - Dissimilarity Matrix Calculation

The distance matrix and the similarity matrix are then used to derive a variety of different visualisation techniques.[2] The first such visualisation technique is the dendrogram as shown in Figure 3.

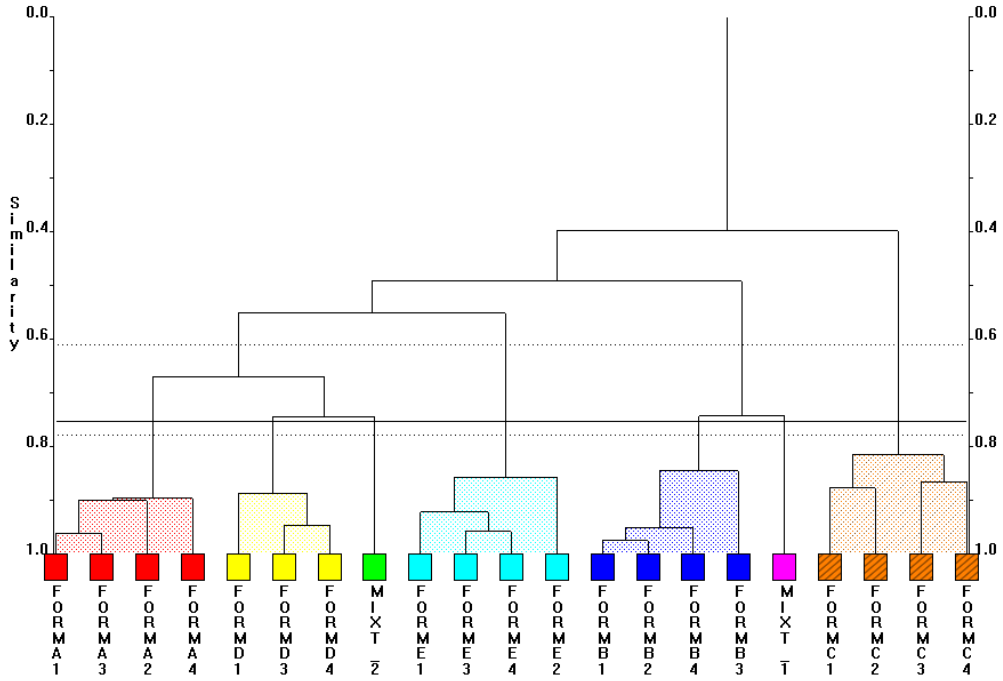

Figure 3 - PolySNAP Dendrogram

A dendrogram is a tree diagram, derived from similarities, which shows similarity between patterns based on the height of connecting lines between them. All patterns are placed at the bottom of the dendrogram, with each of the boxes representing a single pattern. Initially, all patterns are classified as being in different clusters; however they are linked together, stepwise, by horizontal bars which are known as tie bars. The height of these tie bars will reflect how similar two patterns are based on the distance matrix. Similar patterns have low (near to the bottom and the samples) tie bars while patterns which are dissimilar will have tie bars higher up the dendrogram. A single solid horizontal line, known as the cut-level, runs horizontally across the dendrogram. This distinguishes the different clusters present in the dendrogram. The dotted lines above and below the cut-level represent the confidence limits for the cut-level. The determination of the positions of these lines will be covered later. It is possible for the user to adjust the cut-level up or down.

If two samples are linked by tie bars below the cut-level, they are classed as being in the same cluster. Two samples linked by tie bars above the cut-level are classed as being in different clusters. In this dendrogram the samples are split into seven clusters, each assigned a different colour. If the cut-level were adjusted downwards, clusters would be split as the tie bars linking them would now be above the cut-level. Likewise if the cut-level were adjusted upwards, clusters would be merged as the tie bars linking them would

now lie below the cut-level. The new clustering would lead to an updating of the cluster colouring to reflect this.

When examining a dendrogram, it is sometimes possible to determine if it is a ‘good’ or a ‘bad’ one. Figure 3 showed an example of a good one while Figure 4 shows a bad one.

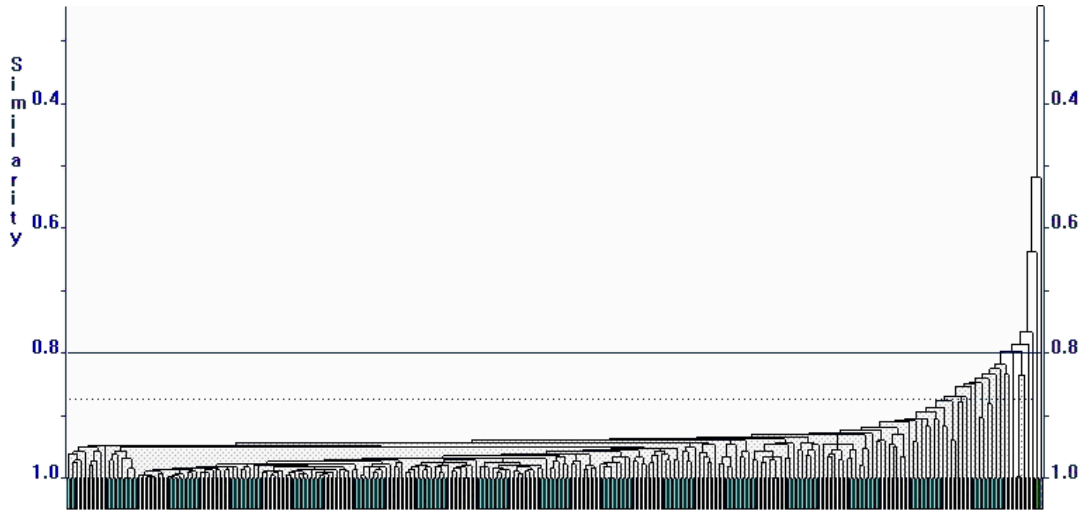

Figure 4 - Example of a 'bad' Dendrogram

This dendrogram clearly shows an example of chaining where a sample is linked to the succeeding sample by a slightly higher tie bar. This then progresses along the dendrogram, forming a stair like structure. A ‘bad’ dendrogram possess no distinct clusters and only shows gradual changes between samples.

The dendrogram can be generated in many different ways. PolySNAP offers the following methods

1. Single link
2. Complete link
3. Weighted average link
4. Centroid
5. Group-Average link

When two clusters *i, j* are combined, Equation 6 is used to calculate the new distance between the cluster and an existing cluster *k.*

$$d_{k(i,j)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma \left| d_{ki} i d_{kj} \right|$$

Equation 6 - Distance calculation when combining clusters

Parameters α_i, α_j, β and γ vary dependent on which clustering method is used. The different methods available are shown in Table 1.

Method	α_i	α_j	β	γ
Single Link	1/2		0	-1/2
Complete Link	-1/2		0	1/2
Weighted Average Link	1/2		0	0
Centroid	$\frac{n_i}{(n_i + n_j)}$		$\frac{-n_i n_j}{(n_i + n_j)^2}$	0
Group-Average Link	$\frac{n_i}{(n_i + n_j)}$	$\frac{n_j}{(n_i + n_j)}$	0	0

Table 1 - Variable parameters in each clustering method

Although not normally used, the PolySNAP software is capable of determining an optimal clustering method. There are several methods of carrying this out:

1. Minimum cluster overlap.
2. Mean intra-cluster distance.
3. Centroid cluster distances.
4. Combining all three.

The distance matrix can also be used to generate a metric multidimensional scaling (MMDS) plot (Figure 5).

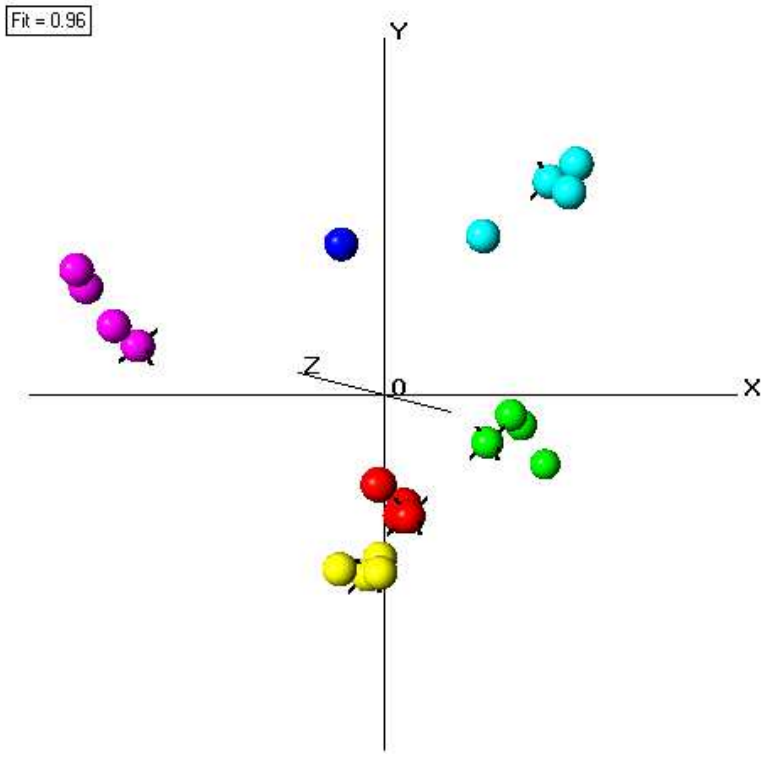

Figure 5 - PolySNAP MMDS Plot

The functionality of MMDS, as described by Gower[8], attempts to define a set of *p* dimensions that will produce a Euclidean distance matrix, $\mathbf{d}^{calc}$, which is equivalent to the distance matrix $\mathbf{d}^{obs}$. Matrix $\mathbf{d}^{obs}$ has zero diagonal elements and so is classed as a positive semidefinite. A positive definite, **A**, can be calculated using Equation 7.

$$A = -\frac{1}{2}\left(\mathbf{I}_n - \frac{1}{n}\mathbf{i}_n\mathbf{i}_n'\right)\mathbf{D}\left(\mathbf{I}_n - \frac{1}{n}\mathbf{i}_n\mathbf{i}_n'\right)$$

Equation 7 - Equation for Matrix A Calculation

Where $\mathbf{I}_n$ is an *(n x n)* identity matrix, $\mathbf{i}_n$ is an *(n x 1)* vector of unities and **D** is a distance squared matrix as defined in Equation 8.

$$\mathbf{D} = 0.25\left(1 - \boldsymbol{\rho}\right)^2$$

Equation 8 - Distance-squared matrix calculation

Where **ρ** is a correlation matrix as defined previously.

Matrix $\left(\mathbf{I}_n - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{i}_n\mathbf{i}_n'\right)$ is a centering matrix, so called as **A** has been derived by centering the rows and columns in **D**. The next step is to obtain the eigenvectors $v_1, v_2, \ldots v_n$ and their

corresponding eigenvalues $\lambda_1, \lambda_2, \ldots \lambda_n$. A total of p of the eigenvalues of **A** are positive, while $(n - p)$, the remainder, are zero. A set of co-ordinates are defined for the non-zero values using the matrix $\mathbf{X}(n \times p)$ using Equation 9.

$$\mathbf{X} = V\Lambda^{1/2}$$

Equation 9 - Matrix X calculation

Λ is the vector of the eigenvalues.

Since we are working in three dimensions $p = 3$ so, with **X**, each pattern can be plotted onto a three-dimensional graph, the MMDS plot. The Euclidian matrix produces a 3-dimensional matrix with every pattern assigned a set of (x, y, z) coordinates. These coordinates can then be plotted in a 3-dimensional plot. Each sample is represented in this plot by a sphere.

A computed distance matrix, $\mathbf{d}^{calc}$, can be produced using $\mathbf{X}(n \times 3)$ and compared with the observed matrix, $\mathbf{d}^{obs}$. The calculated and observed distance matrix are correlated, using both Pearson and Spearman correlation coefficients, and the mean correlation coefficient from this is displayed in the upper-left corner, labelled as 'Fit'. This serves as a check of the accuracy of the MMDS calculations. Ideally a value of greater than 0.95 is desired, but this figure reduces as n increases.

The MMDS plot also shows the most representative sample for each cluster. This only appears for clusters with three or more samples present and is represented by adding spikes to the sphere which represents this sample in the MMDS plot. The most representative sample is the sample with the lowest mean distance to all other samples in the cluster. The method of calculating the mean average distance is shown in Equation 10.

$$i = \min\left[\sum_{\substack{j=1 \\ i,j\in J}}^{m} d_{ij} / m\right]$$

Equation 10 - Most Representative Sample

Where i is the most representative sample in cluster J.

The principal components analysis (PCA) plot is another means of representing the data in 3 dimensions. A PCA plot is shown in Figure 6.

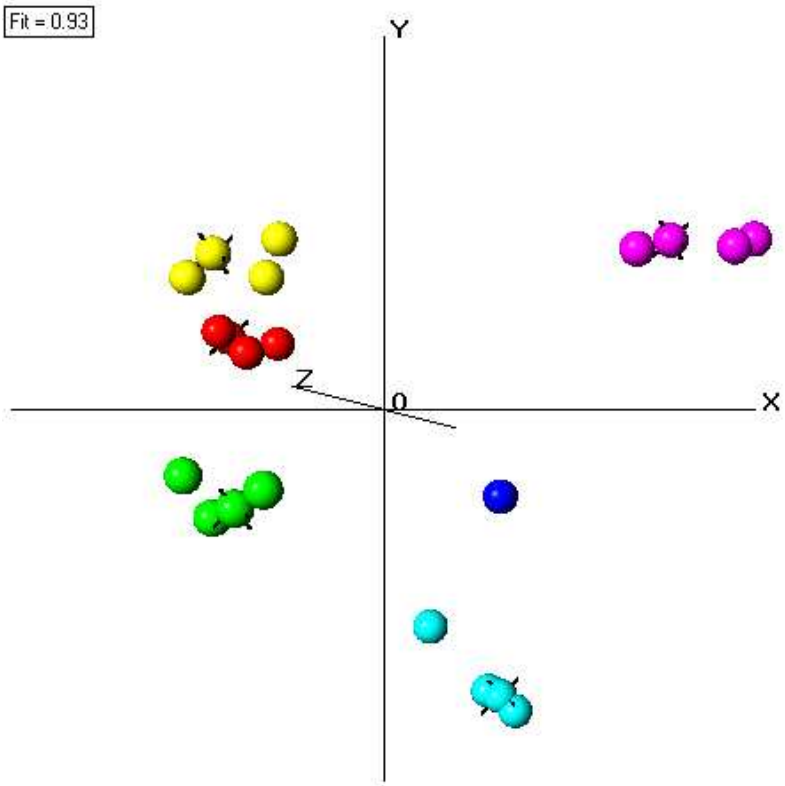

Figure 6 - PolySNAP PCA plot

A three-dimensional score plot is produced by carrying out principal components analysis on the correlation matrix. The PCA plot is then drawn using the score plot in a similar method to the MMDS plot. The PCA plot, like the MMDS plot, displays how good the score plot matches with the observed distance matrix and displays this in the upper left corner.

PolySNAP chooses whether the PCA plot or MMDS plot gives the optimal display for each dataset by comparing the fits from the two plots and choosing the one with the highest value.

A cell display (Figure 7) is generated using the results from the dendrogram.

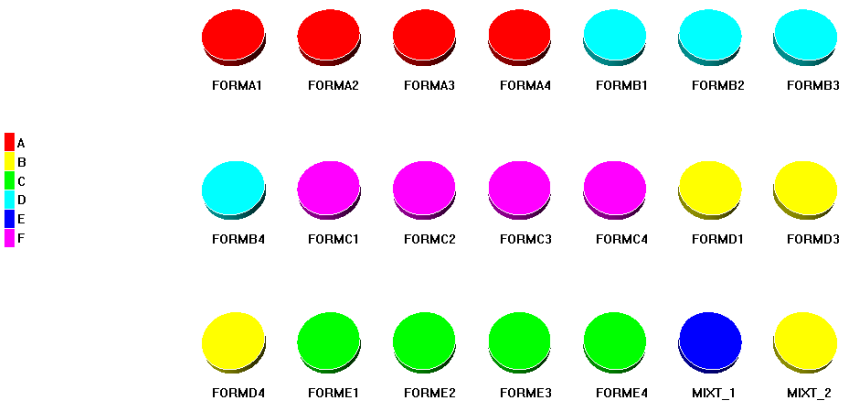

Figure 7 - PolySNAP Cell Display

Each pattern is represented by a cell. The colouring of the cells represents the cluster or set that a pattern is assigned to. Patterns with the same colour have been determined to be similar by the clustering methods. Cells are used as high-throughput powder diffraction experiments are typically run using well plates (see Chapter 2), therefore it is useful to have a display representing the layout of a well plate.

The number of clusters present, and hence the position of the cut-level, is determined using a variety of different methods. Initially the following two methods are used:

1. Eigenvalue analysis of $\boldsymbol{\rho}$, $\mathbf{A}$ and a transformed version of $\boldsymbol{\rho}$.
2. Cluster analysis methods

The transformed version of $\boldsymbol{\rho}$ uses a standardized version $\boldsymbol{\rho}_s$, where the rows and columns have a mean and unit variance of zero. Matrix $\rho_s\rho$ is computed and subjected to eigenanalysis.

Cluster analysis is less commonly used in crystallography than eigenvalue analysis. Based on literature[9,10], three possible methods are available.

1. Calinski and Harabasz (CH) test[11]
2. A variant of Goodman and Kruskal's[12] γ test[10]
3. The C test[9]

These methods can be carried out on several of the different dendrogram generation methods in order to minimise the bias towards any one classification schemes. The following list shows all of the possible methods for determining the number of clusters:

1) Eigenvalue analysis of **A**
2) Eigenvalue analysis of $\boldsymbol{\rho}$
3) Eigenvalue analysis of transformed $\boldsymbol{\rho}$
4) Calinski/Harabasz test on single linkage
5) γ test on single linkage
6) C test on single linkage
7) Calinski/Harabasz test on group average linkages
8) γ test on group average linkages
9) C test on group average linkages
10) Calinski/Harabasz test on centroid method
11) γ test on centroid method
12) C test on centroid method
13) Calinski/Harabasz on complete linkages
14) γ test on complete linkages
15) C test on complete linkages

Eigenvalue analysis is carried out by sorting the eigenvalues of $\boldsymbol{\rho}$, **A** and the transformed version of $\boldsymbol{\rho}$ in the correct matrix into descending order until a fixed percentage of the variables, usually 95%, have been accounted for. Figure 8 shows a plot, known as a scree plot, of the eigenvalues from a data run.

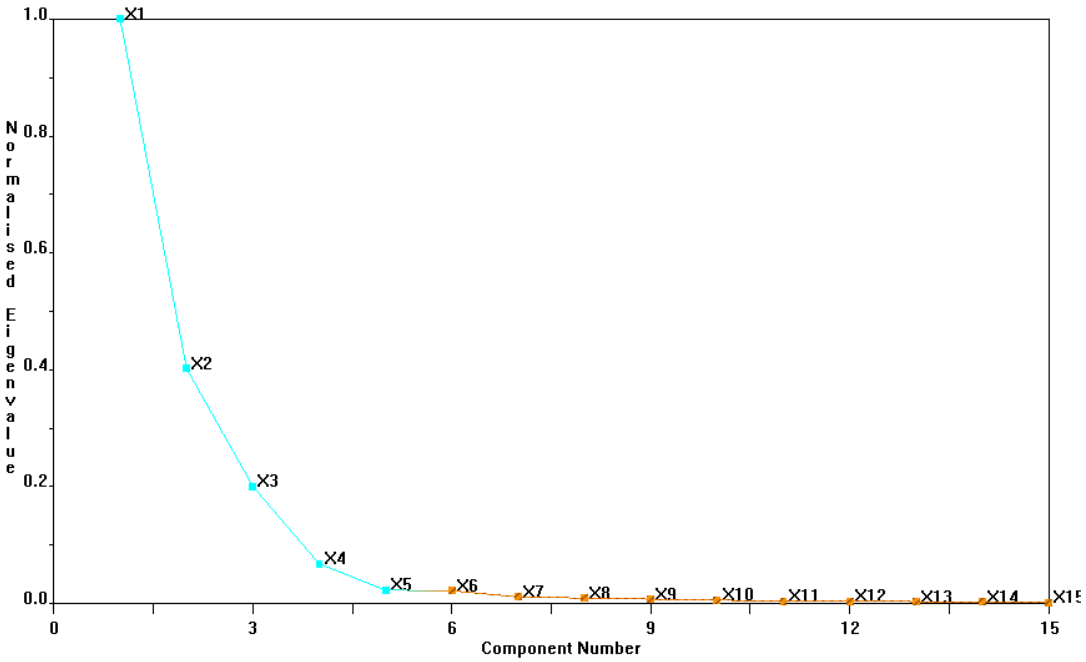

Figure 8 - Scree Plot

The scree plot uses eigenvalues, generated from the correlation matrix to determine how many clusters are present. The point in the plot where it changes colour is the point where 95% of the variability has been accounted for and shows the estimated number of clusters present in the dataset. As such the scree plot can be viewed as a one-dimensional representation of the eigenvalue methods for calculating the number of clusters present. A scree plot should ideally show a steep initial descent, changing rapidly to a shallow descent. One with a shallow or multiply stepped initial descent may indicate poor clustering.

Eigenvalue analysis of $\boldsymbol{\rho}$ and **A** uses matrices that have been previously defined, however the transformed $\boldsymbol{\rho}$ matrix has not yet been defined. For this matrix $\boldsymbol{\rho}$ is standardised to give $\boldsymbol{\rho}_{\mathbf{s}}$, where the rows and columns have zero mean and unit variance.

The CH test uses Equation 11.

$$CH(c) = \left[B/(c-1) \right] / \left[W/(n-c) \right]$$

Equation 11 - CH Test Equation

This method defines a centroid for each cluster. W is the total within-cluster sum of squared distances across the cluster centroids, *B* is the total between-cluster sum of squared distances and *c* is the number of clusters chosen to maximise the equation.

The Goodman and Kruskal γ test uses the dissimilarity matrix as defined earlier in Equation 4.

All within-cluster dissimilarities are compared with all between-cluster dissimilarities. If the within-cluster dissimilarity is less than the between-cluster dissimilarity the comparison is said to be concordant. If the opposite is true it is discrepant. If they are equal they are disregarded. These values are used with Equation 12.

$$\gamma(c) = (S_+ - S_-)/(S_+ + S_-)$$

Equation 12 - Goodman and Kruskal γ Test

Where S_+ is the number of concordant comparisons and S_- is the number of discordant comparisons.

In the C test we choose a value of *c* which minimises Equation 13.

$$C(c) = \left[D(c) - D_{\min} \right] / \left(D_{\max} - D_{\min} \right)$$

Equation 13 – The C Test

Where D(c) is the sum of all within-cluster dissimilarities. If r dissimilarities are present, $D_{\min}$ is the sum of the r smallest dissimilarities and $D_{\max}$ the sum of the r largest dissimilarities.

The maximum and minimum values produced from the fifteen methods appear as dotted lines on the dendrogram, denoting the confidence limit for the cluster estimation. A weighted mean value of all the estimates is taken, the result of which appears as the initial cut-level on the dendrogram.

A further method of constructing clusters is through the use of minimum spanning tree,[25, 26] as shown in Figure 9.

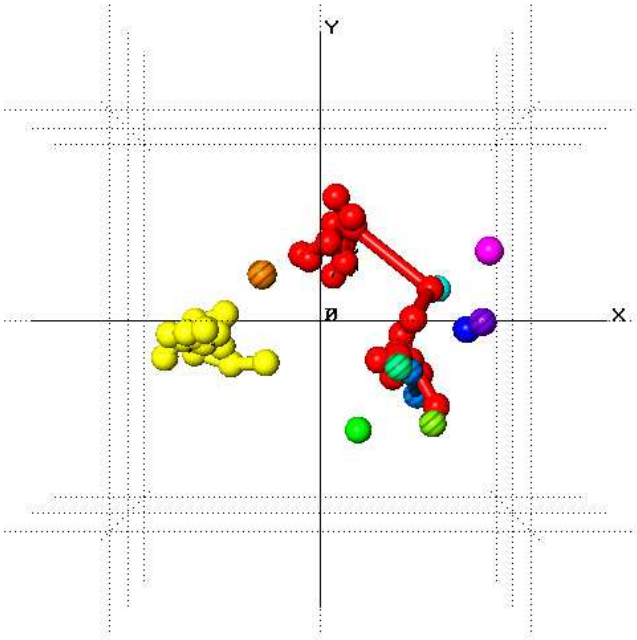

Figure 9 - Minimum Spanning trees

A spanning tree is graph where all of the vertices are connected by a set of lines. A vertex can have multiple lines connected to it, but at no point can the lines form a loop. In a minimum spanning tree, the lines are drawn so that so that the minimum distance of lines are used to connect all of the points. Multiple minimum spanning trees can exist in a single graph as, if a link is broken, two minimum spanning trees will now be present on either side of the broken link.

This plot initially has all the samples tied together with a series of lines, a single minimum spanning tree. Links are cut, in decreasing size order beginning with the longest line, until

the estimated number of clusters from the scree plot is reached, and therefore the number of minimum spanning trees is equal to the number of clusters. Once the estimated number of clusters has been reached, the user can choose to manually decrease or increase the number of links either creating new clusters or merging further clusters. Previously broken links can also be remade in increasing size order. When a link is broken the number of clusters present will increase by one. When a link is formed two clusters, the two which are closest to one another, will form into one larger cluster.

It is also possible to supply PolySNAP with a correlation matrix that has been generated from an external source. The software will not carry out the initial steps of pre-processing the data and correlating each pattern but will instead follow the flowchart as shown in Figure 10.

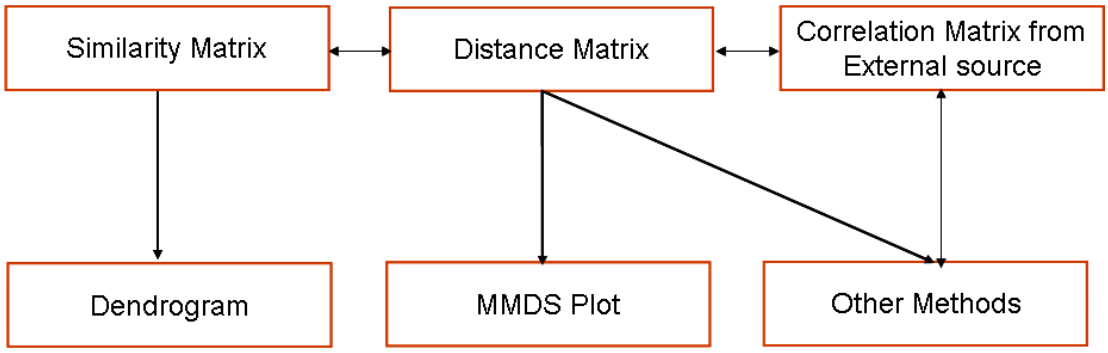

Figure 10 - Amended flowchart for correlation matrix input

All methods after the correlation matrix is read by the software are unchanged.

1.2.2 VALIDATION TECHNIQUES

1) Silhouettes

There are many methods built in which allow the clustering to be checked.[4, 26] The first of these methods is the silhouette. Figure 11 shows the silhouette[21, 22] plot for each cluster.

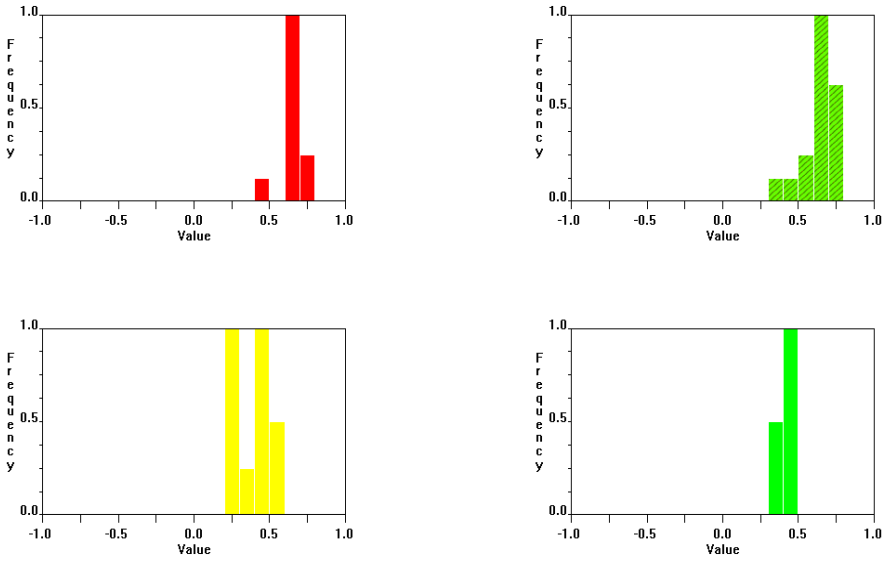

Figure 11 - Silhouettes

Silhouettes show, for each cluster as identified in the dendrogram, how well each pattern fits into the cluster. Silhouettes are only displayed for clusters with three or more patterns. The values on the silhouette x-axis run from -1 to 1. Each member of a cluster is assigned a silhouette. The first step in getting the silhouettes value is to generate a dissimilarity matrix. The equation for this has already been shown in Equation 5.

Two values, a_i and b_i (where i is the pattern being examined), must be defined before the next step can be carried out. These values are defined in equations 14 and 15.

$$a_i = \sum_{\substack{j \in C_r \\ j \neq 1}} \delta_{ij} / (n_r - 1)$$

Equation 14 - Definition for a_i

$$b_i = \min_{s \neq r} \left(\sum_{j \in C_r} \delta_{ij} / n_s \right)$$

Equation 15 - Definition for b_i

Both of these values are used to define the average dissimilarity of pattern *i* with respect to the cluster C_r. The silhouette for this pattern is defined using Equation 16.

$$h_i = (b_i - a_i) / max(a_i, b_i)$$

Equation 16 - Value of Silhouette

The value of h_i is in the range $-1.0 \leq h_i \leq 1.0$.

The higher the value, the more likely a pattern is to belong to this cluster. Lower values imply that either the sample belongs to another cluster or that it is a mixture. In general, samples with a value of greater than 0.5 can be said to be clustered correctly. Values outside of this range should be examined to see if their clustering has been determined correctly. Any gaps between the different regions are also of interest as they can further suggest that samples on either side of the gaps should be in different clusters.

2) Fuzzy Clustering

Figure 12 shows a representation of fuzzy clustering.[10, 14, 26]

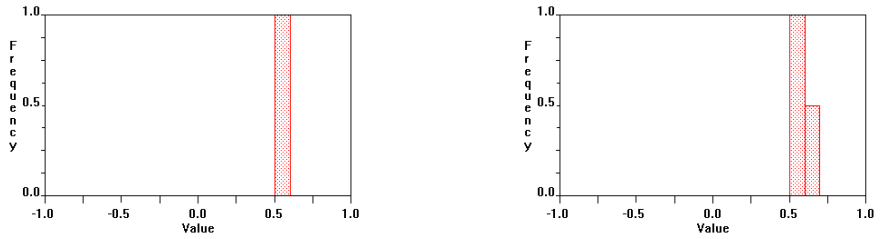

Figure 12 - Fuzzy Clustering

Fuzzy clustering allows a sample to potentially be included in more than one cluster. Hitherto, cluster membership is expressed by a membership matrix ***U(n x c)*** in which individual coefficients, u_{ik}, represent the membership of pattern *i* in cluster *k*. Coefficients are equal to unity if *i* belongs to *c* and is otherwise zero, as shown in Equation 17

$$u_{ik} \in [0,1] (i = 1,...,n; k = 1,...,c)$$

Equation 17 – Cluster Membership

These constraints can be relaxed as shown in Equations 18, 19 and 20.

$$0 \leq u_{ik} \leq 1 (i = 1,...,n; k = 1,...,c)$$

Equation 18 - Relaxed Cluster Memberships Constraints 1

$$0 < \sum_{i=1}^{n} u_{ik} < n\left(k = 1, \ldots, c\right)$$

Equation 19 - Relaxed Cluster Memberships Constraints 2

$$\sum_{k=1}^{c} u_{ik} = 1$$

Equation 20 - Relaxed Cluster Memberships Constraints 3

These relaxed constraints now allow for the possibility of fuzzy clusters[10, 13, 14] where a sample can belong to more than one cluster, for example in mixtures.
The Matrix **U** is calculated *via* two possible methods

1. Additive clustering where **U** is determined by minimising the difference between observed and calculated matrixes. The minimised function for this is shown in Equation 21.

$$\eta_1^2 = \sum_{i \neq j=1}^{n} \left(s_{ij} - \sum_{k=1}^{c} \min\left(u_{ik}, u_{jk}\right) \right) / \sum_{i \neq j=1}^{n} \left(s_{ij} - \bar{s} \right)$$

Equation 21 - First Fuzzy Clustering Method Minimised Function

Where s is defined as

$$\bar{s} = \left[1/n\left(n-1\right)\right] \sum_{i \neq j=1}^{n} \left(s_{ij}\right)$$

Equation 22 - Definition

Where α is a constant that scales **s** and **U**.

2. A general algorithm using aggregation operators as shown in Equation 23

$$J = \sum_{i \neq j=1}^{c} \left[s_{ij} - \sum_{k=1}^{c} \min\left(u_{ik}, u_{jk}\right) \right]$$

Equation 23 - Second Fuzzy Clustering Method Minimised Function

Both of these techniques need a starting value for **U**. In PolySNAP the initial cluster assignment from the dendrogram is used so if a pattern *i* belongs to a cluster *j*, $u_{ij} = 0.8$. The value is otherwise given a random value scaled in accordance with Equation 20.
The two methods minimise different functions and so give different results. The second method tends to give values with a wider range. Where $u_{ij} < 0.3$, the value is usually treated as zero.

1.2.3 VALIDATION EXAMPLE

The 'multiple 1' tutorial dataset (as distributed with the PolySNAP software) will be used as an example of each of the validation techniques in action. This dataset will be studied in more detail in Chapter 4. The dataset consists of three polymorphs of sulfathiazole. The dataset contains forty-eight samples, with each polymorph having multiple measurements. The dendrogram and MMDS plot are shown in Figure 13.

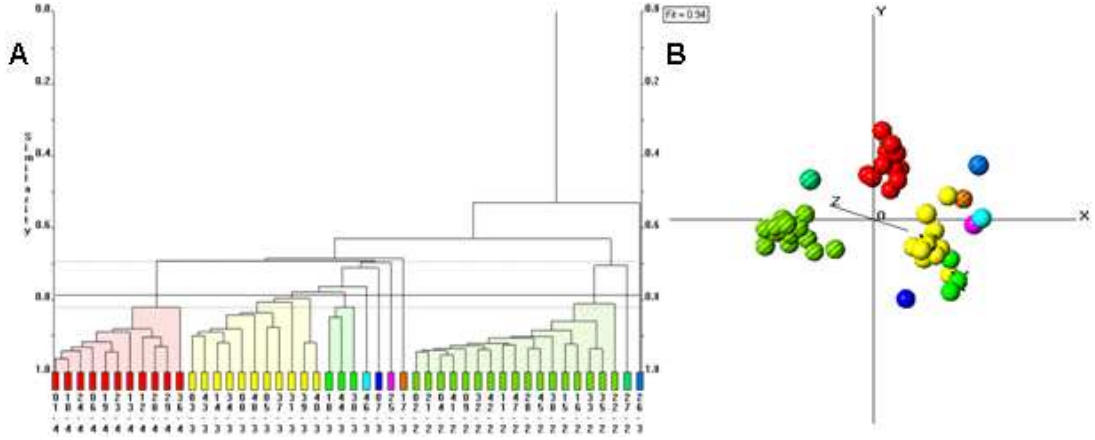

Figure 13 – A -Validation Example Dendrogram, B – Validation Example MMDS Plot

The dendrogram shows three large groups with a single outlier present (striped blue cluster) which would be expected to be with the yellow cluster. The MMDS plot shows the cluster to be clearly separated from one another, with many of the outliers lying close to the yellow cluster.
The scree plot is shown in Figure 14.

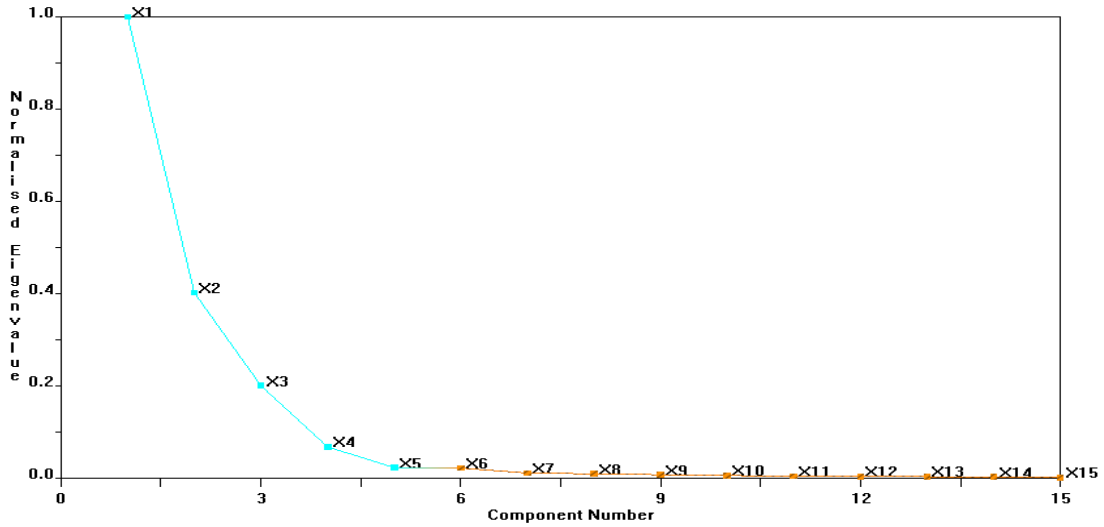

Figure 14 - Scree Plot Example

The scree plot shows the expected steep initial drop and suggests that there are five clusters present. The minimum spanning tree is shown in Figure 15, along with the effect of removing and adding one link. The changes to the minimum spanning trees are circled in green.

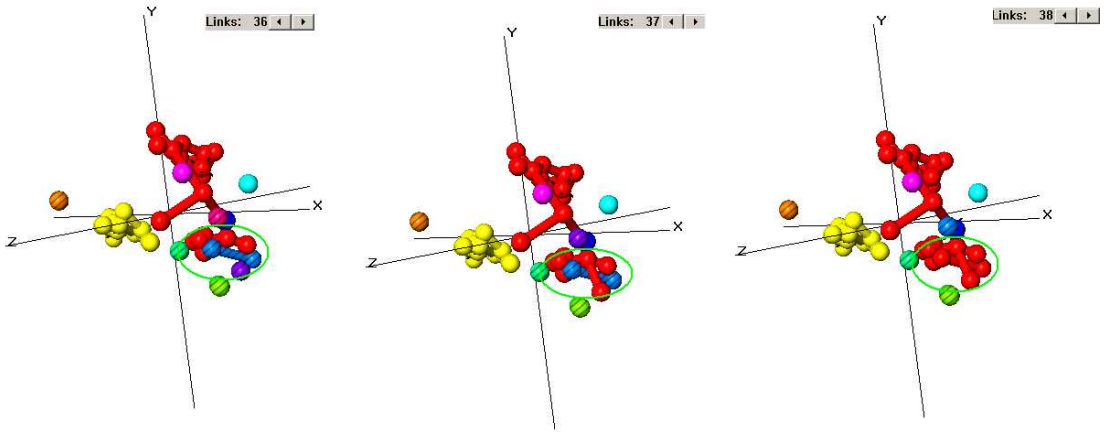

Figure 15 - Minimum Spanning Tree

The central example shows a large interlinked red cluster with an interlinked yellow cluster and interlinked blue cluster nearby. Breaking a single link (the left most example) will split a single sample off the bottom and form a purple cluster. Adding a single link (the right most example) will merge the blue and red tree.

Figure 16 shows the silhouettes and Figure 17 the fuzzy clusters

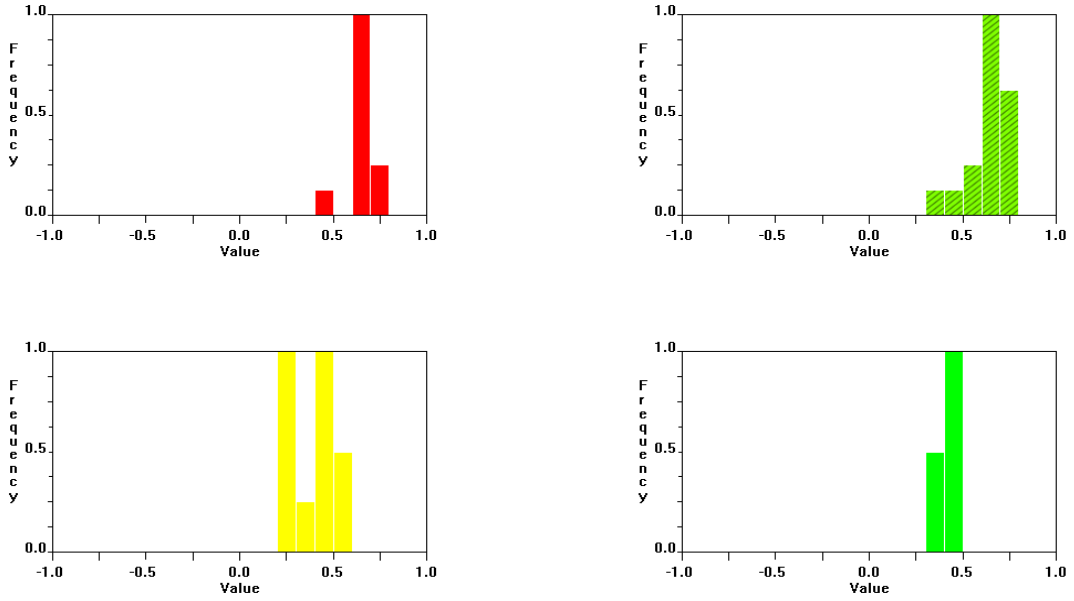

Figure 16 - Silhouettes for Example

In the red cluster, sample 36-4 is a potentially ambiguous sample, lying below 0.5. All of the remaining bars, and therefore all remaining samples, lie above 0.5. There is also a small gap between the region with sample 36-4 in it and the remaining regions, suggesting that it should be in a separate cluster.

For the striped green cluster, sample 22-2 and 35-2 are potentially ambiguous samples.

For the yellow cluster samples 31-3, 37-3, 39-3, 40-3 and 48-3 are potentially ambiguous samples. For the green cluster, all three samples lie below 0.5 and are therefore potentially ambiguous.

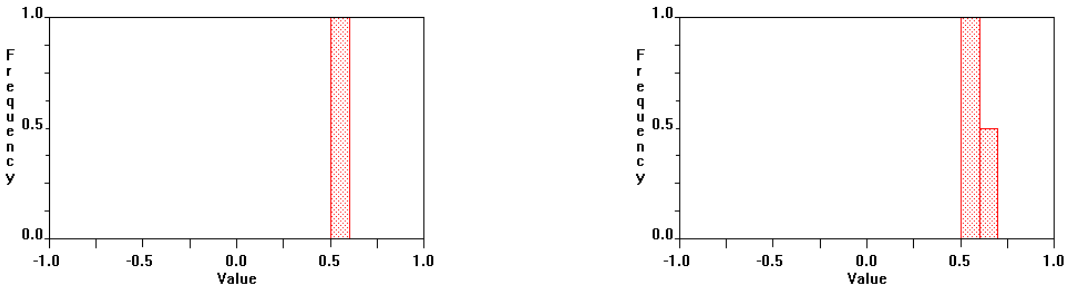

Figure 17 - Fuzzy Clustering for Example

The cluster memberships for the fuzzy clustering are shown in Table 2.

	1	2	3	4	5	6	7	8	9	10	
01-4	0.1	0.08	0.09	0.1	0.1	0.09	0.15	0.39	0.45	0.94	
02-2	0.08	0.11	0.08	0.09	0.08	0.09	0.12	0.29	1	0.35	
03-3	0.12	0.09	0.11	0.12	0.12	0.11	0.21	0.9	0.43	0.45	
04-2	0.08	0.1	0.08	0.08	0.09	0.1	0.13	0.33	1	0.38	
05-3	0.13	0.11	0.12	0.12	0.11	0.1	0.19	0.91	0.37	0.48	
06-4	0.1	0.09	0.1	0.1	0.11	0.1	0.16	0.43	0.47	0.91	
07-3	0.12	0.09	0.1	0.12	0.11	0.84	0.25	0.47	0.49	0.35	
08-3	0.11	0.1	0.1	0.1	0.11	0.11	0.19	0.9	0.48	0.39	
09-2	0.08	0.11	0.08	0.09	0.08	0.1	0.12	0.32	1	0.37	
10-4	0.1	0.09	0.09	0.1	0.11	0.1	0.15	0.38	0.46	0.94	
11-2	0.09	0.12	0.08	0.08	0.09	0.1	0.12	0.33	0.98	0.41	
12-4	0.1	0.1	0.1	0.09	0.1	0.09	0.14	0.37	0.46	0.93	
13-4	0.1	0.1	0.1	0.1	0.11	0.09	0.14	0.39	0.5	0.9	
14-3	0.12	0.1	0.11	0.12	0.11	0.12	0.23	0.92	0.46	0.38	
15-2	0.07	0.09	0.07	0.07	0.07	0.08	0.09	0.23	1	0.3	
16-2	0.07	0.09	0.07	0.07	0.07	0.08	0.1	0.25	1	0.31	
17-3	0.13	0.1	0.12	0.12	0.85	0.11	0.21	0.49	0.34	0.45	
18-3	0.13	0.09	0.11	0.13	0.13	0.14	0.86	0.55*	0.41	0.42	<==
19-4	0.09	0.09	0.09	0.1	0.11	0.09	0.14	0.37	0.55*	0.9	<==
20-4	0.11	0.12	0.12	0.11	0.11	0.09	0.14	0.41	0.45	0.91	
21-2	0.08	0.11	0.08	0.08	0.08	0.09	0.11	0.28	1	0.35	
22-2	0.06	0.08	0.06	0.06	0.06	0.08	0.09	0.21	0.97	0.27	
23-4	0.09	0.09	0.09	0.09	0.1	0.09	0.14	0.35	0.55*	0.91	<==
24-4	0.1	0.09	0.09	0.1	0.11	0.11	0.18	0.4	0.5	0.91	
25-3	0.14	0.1	0.12	0.85	0.12	0.11	0.22	0.51*	0.31	0.39	<==
26-3	0.12	0.1	0.86	0.12	0.11	0.09	0.18	0.45	0.26	0.38	
27-2	0.11	0.84	0.11	0.1	0.09	0.09	0.12	0.36	0.60*	0.38	<==
28-2	0.08	0.11	0.07	0.07	0.07	0.1	0.11	0.26	1	0.33	
29-4	0.11	0.11	0.11	0.1	0.11	0.09	0.15	0.42	0.45	0.92	
30-3	0.13	0.1	0.11	0.12	0.11	0.13	0.86	0.51*	0.38	0.33	<==
31-3	0.12	0.11	0.1	0.1	0.1	0.1	0.19	0.89	0.42	0.34	
32-2	0.08	0.11	0.07	0.08	0.08	0.09	0.11	0.29	1	0.34	
33-2	0.08	0.09	0.08	0.09	0.08	0.11	0.14	0.34	0.97	0.36	
34-3	0.12	0.09	0.11	0.11	0.11	0.11	0.2	0.91	0.44	0.4	
35-2	0.09	0.1	0.09	0.09	0.09	0.11	0.15	0.42	0.91	0.37	
36-4	0.1	0.11	0.1	0.09	0.09	0.09	0.12	0.33	0.42	0.9	
37-2	0.12	0.11	0.12	0.11	0.1	0.09	0.17	0.9	0.39	0.38	
38-2	0.08	0.1	0.07	0.08	0.08	0.09	0.11	0.29	0.99	0.34	
39-3	0.1	0.07	0.09	0.1	0.1	0.11	0.19	0.91	0.37	0.33	
40-3	0.1	0.09	0.1	0.1	0.1	0.1	0.17	0.91	0.39	0.34	
41-2	0.07	0.1	0.07	0.08	0.08	0.1	0.12	0.28	1	0.35	
42-2	0.08	0.12	0.08	0.08	0.08	0.09	0.11	0.3	1	0.35	
43-3	0.12	0.09	0.11	0.12	0.12	0.12	0.23	0.92	0.42	0.41	
44-3	0.13	0.09	0.11	0.12	0.12	0.13	0.86	0.53*	0.37	0.36	<==
45-2	0.08	0.11	0.08	0.08	0.08	0.1	0.11	0.27	1	0.36	
46-3	0.85	0.11	0.13	0.14	0.13	0.12	0.24	0.57*	0.31	0.41	<==
47-2	0.08	0.11	0.08	0.08	0.1	0.09	0.11	0.3	1	0.35	
48-3	0.11	0.09	0.1	0.11	0.1	0.1	0.18	0.9	0.42	0.37	

Table 2 – Fuzzy Clustering Numerical Values

In the table, column 1 shows the sample name while the subsequent columns, the number of these varies depending on how many clusters are present, shows the correlations that the sample has to each cluster. If a sample has no correlations greater than 0.50 or has more than one correlation greater than 0.50 then it will be marked with <= = in the final column of the table. For these samples, the cluster membership which is determined to be optimal is marked with an *. For this dataset no samples have all correlations less than 0.5.

The fuzzy clustering contains several samples. 18-3, 25-3, 30-3, 44-3 and 46-3 are present in the first fuzzy clustering plot. The second plot contains samples 19-4 and 25-3 in the first bar and 27-2 in the second bar.

All of the patterns in this example that are marked as fuzzy clusters have more than one membership of greater than 0.5. There are no samples present in this dataset with all correlations <0.5.

Figure 18 shows the parallel coordinate plot for the example.

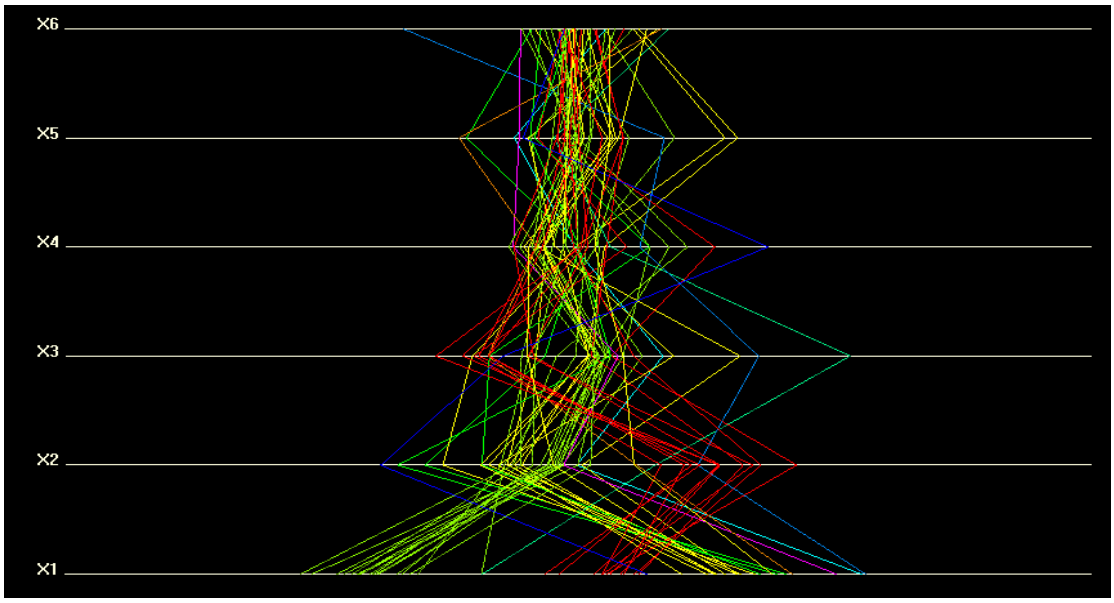

Figure 18 – Parallel Coordinate Plot for Example

In the currently shown orientation, the first three dimensions show that the clusters are clearly defined. Moving into dimension four, the yellow cluster spreads out but tightens up again in the fifth and sixth dimension.

The dataset was well clustered initially; however the validation methods assist in determining that the clustering is optimal. The red cluster in particular is well clustered with just one sample in it being potentially ambiguous according to the silhouettes. This sample is not present in the fuzzy clustering as it has a membership value of 0.9 with the red cluster. The entire green cluster is marked as potentially ambiguous by the silhouettes,

which is further reinforced by the fuzzy clustering also having all of the samples in this cluster marked as ambiguous. This implies that the green and yellow cluster should be further inspected and possibly merged.

Further discussion on this dataset and its cluster membership can be found in Chapter 4.

1.2.4 INDIVIDUAL DIFFERENCES SCALING METHOD (INDSCAL)

The individual differences scaling method, INDSCAL[15], can be used to combine the distance matrixes of two or more datasets into a single distance matrix. This new distance matrix can then be used to generate a new dendrogram combining the clustering of all of the input datasets. The method determines the optimal weights to assign to each pattern before combining them.

Combining two or more datasets is a useful technique as it allows for a further confirmation of the results. In the case of a dataset for which there is no expected result, this can serve as an additional confirmation if the clustering from the combined dataset is similar to the clustering of the individual datasets. In the case of a dataset for which a certain result was expected, it can serve as a useful means of confirming that result. If one of the datasets being combined is of poorer quality it may also serve as a check of the results if the results from combining this 'poorer' quality data with the second dataset gives the clustering which was anticipated.

The automatic combination of distance matrixes functions as follows:

A group-average matrix, **G,** is generated by either taking an average of the values of the input distance matrixes or by randomly generating a matrix. The distance matrixes are converted into inner product form (Equation 24)

$$\mathbf{B}_k = -1/2(\quad)\mathbf{D}_{\boldsymbol{k}}(\quad)$$

Equation 24 – Inner product matrix form

Where **I** is the identity matrix, $\mathbf{N} = \mathbf{11'}/N$ and $\mathbf{D}_k$ is an (n x n) squared-distance matrix for datatype k with K total datatypes.

The inner product form matrixes are now matched to the weighted form of the group average matrix.

$$S = \sum_{1}^{K} \left\| \mathbf{B}_k - \mathbf{GW}_k^2 \mathbf{G}' \right\|$$

Equation 25 – weighted group average function

where $\mathbf{W}_k$ is the weighted matrix. This function is minimised and $\mathbf{W}_k$ scaled so that

$$\sum_{k=1}^{k} \mathbf{W}_k^2 = K\mathbf{I}$$

Equation 26 - weight matrices function

This equation is repeatedly cycled through by keeping one of the parameters fixed while solving the other through least squares. The process is repeated until a minimum value for *S* is obtained. The group-average matrix can be used as a standard distance matrix to generate a dendrogram, MMDS plot, PCA plot etc.

A group-average method is preferable to simply averaging the two matrices as a simple average of the two is more likely to produce a matrix which does not represent either of the two initial matrices well.

The tutorial dataset ‘multiple 1’ will be used to demonstrate INDSCAL. This is the same dataset that was used for the validation methods example and will be looked at in greater detail in Chapter Four. Three clusters are expected to be present in this dataset. The dataset contains both PXRD and Raman data and all the dendrograms and MMDS plots are shown in Figure 19.

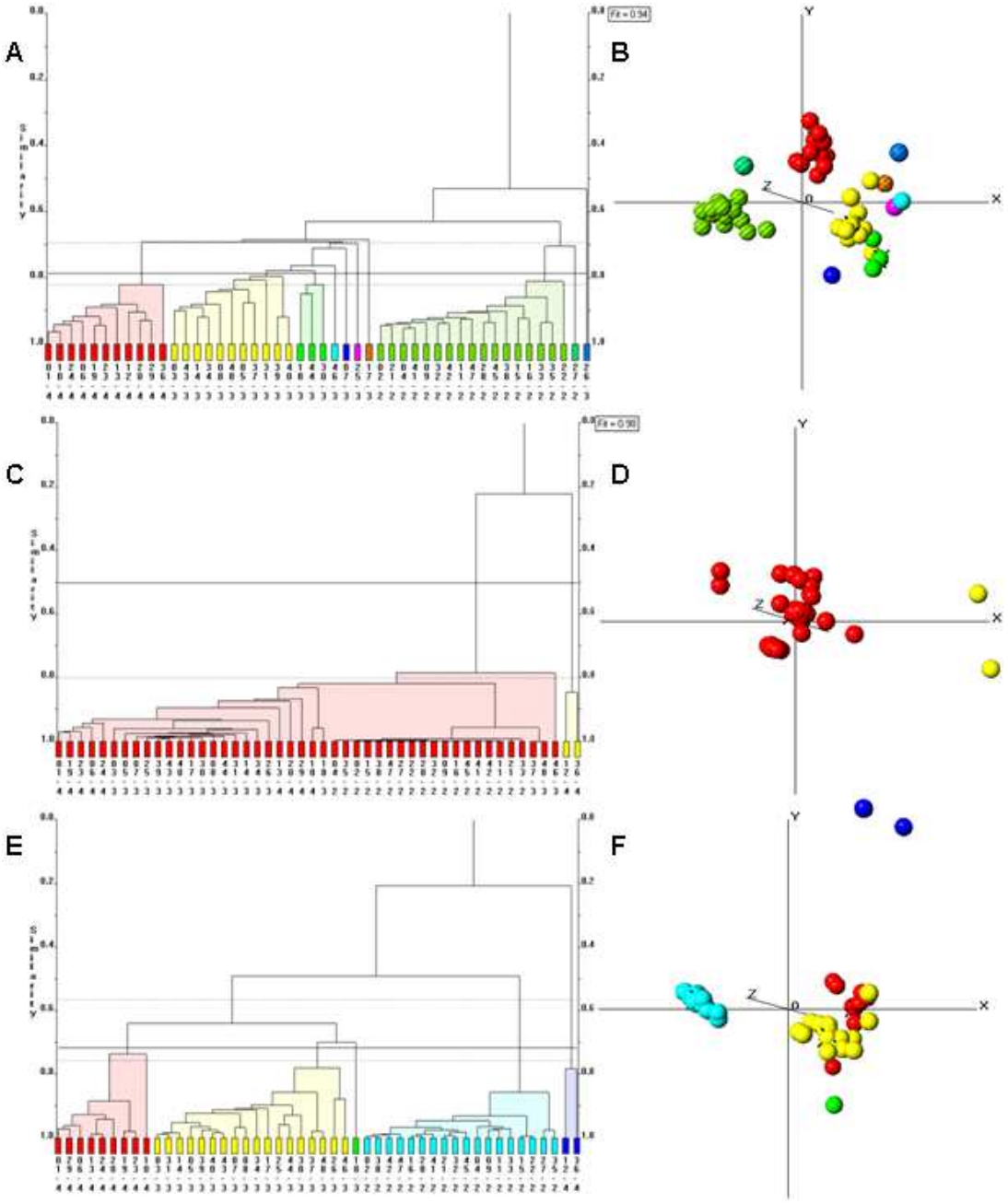

Figure 19 – INDSCAL Example A – PXRD Dendrogram; B – PXRD MMDS Plot; C – Raman Dendrogram; D – Raman MMDS Plot; E – INDSCAL Combined PXRD; F – INDSCAL Combined MMDS Plot

The X-ray data, Figure 19-A, shows two clearly defined clusters. An adjustment of the cut-level upwards would unite many of the separated samples in the centre into a third cluster. The Raman data, Figure 19-C, has nearly all the samples in a single cluster with little clear difference being seen between these samples.

The combined dataset, Figure 19-E, clearly shows three separate clusters, with two outliers at the right of the dendrogram which would otherwise be expected to be in the red cluster. This example clearly shows the benefits of using INDSCAL to combine two datasets. The

INDSCAL dendrogram retains the three clusters that appeared in the initial PXRD dendrogram however it has far fewer outliers present.

1.2.5 DATA PRE-PROCESSING OPTIONS

There are many advanced options available which can be applied to the data before pattern matching occurs. These options are as follows:

1. **Denoise patterns**

This option allows the user to denoise, or smooth the pattern using wavelets[17, 18] *via* Stein's unbiased risk estimate threshold (SURE).[19] A wavelet is a wave which begins at zero, rises to a maximum amplitude, then decreases to zero once more. For this method to function we must first select a threshold λ_j *via* the wavelet coefficients at each wavelet level *j*. This allows the coefficients to be shrunk at each level. This will allow an estimate of $\hat{f}$ and *f* that has a small mean square error, *i.e.* a an estimate of $\hat{f}$ that has a small risk, R($\hat{f}$,*f*). This is shown in Equation 27.

$$R\left(\hat{f}, f\right) = \mathrm{E}\left[\frac{1}{\mathrm{n}}\sum_{i=1}^{n}\left(\hat{f}\left(i/n\right) - f\left(i/n\right)\right)^2\right]$$

Equation 27 – Estimate of Risk for Data

This can also be expressed as wavelet coefficients as shown in Equation 28.

$$\left(\hat{f}, f\right) \propto \mathrm{E}\left[\sum_j \sum_k \left(\theta_{j,k} - \theta_{j,k}\right)^2\right]$$

Equation 28 – Estimate of Risk for Wavelets

For both of these equations E is the expected value which is taken as an integral of the estimator $\hat{f}$.

This allows us to transform the data in the original patterns or spectra into wavelet coefficients. If the risk is minimised in wavelets then it is also minimised in the original data.

The risk is calculated from the data using the SURE method. To minimise the risk, a threshold value is first chosen for each wavelet value. For any observed dataset $x = (x_1 \ldots x_n)$, the risk can be written as shown in Equation 29.

$$SURE(\lambda; x) = d - 2 \cdot \#\{k : |x_k| \leq \lambda\} + \sum_{k=1}^{d} \min^2(|x_k|, \lambda)$$

$$= -d - 2 \cdot \#\{k : |x_k| > \lambda\} + \sum_{k=1}^{d} \min^2(|x_k|, \lambda)$$

Equation 29 - SURE Risk Estimation

Where #S for set S gives the cardinality of that set, which is defined as the number of elements present in the set.

For large sample sizes the law of large numbers will guarantee that SURE gives close to the true risk. The law of large numbers states that, when an experiment with an expected value as an output is repeated a large number of times, the average of the results will be close to the expected value, hence if a large number of samples are used with the SURE method, the resulting output should also be close to the true risk.

2. Subtract background

Background subtraction removes the background from each pattern in the dataset. Background removal operates *via n*th order polynomial functions that are fitted to the data and then subtracted to produce a flat baseline. Three 2θ domains are defined for this, however more can be used if needed.

3. Check for amorphous samples

This option checks each sample in the dataset to determine if it is amorphous. An amorphous sample is said to be non-crystalline as it has only long range ordering of its atomic positions. Amorphous samples are marked as such in the dataset. A further option is available through the menu which will remove any samples listed as amorphous from the dataset.

The amorphous samples are identified as follows:

The background for each sample is calculated and intensity is integrated, the non-background intensity is then estimated. Diffraction peaks are located and the ratio of background intensity to non-background intensity is then compared. If the value is less than 3%, or a value set by the user, or if the number of peaks is less than 3, or a value set by the user, the sample is said to be amorphous.

Amorphous, or non-crystalline materials, are materials which lack long-range ordering of the crystal structure. An example amorphous PXRD pattern is shown in Figure 20.

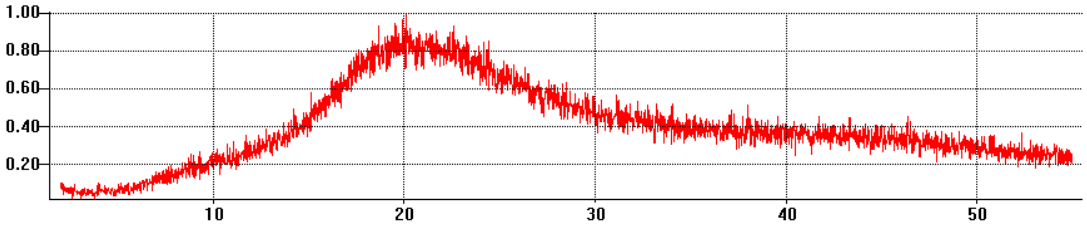

Figure 20 - Amorphous PXRD pattern

4. **Remove Cosmic Ray Spikes**

This option checks each sample for cosmic ray spikes and removes them if present. This option is only applicable to Raman data. This method is required as cosmic ray spikes, spikes on the spectra produced by cosmic rays hitting the detector, give peaks which should not otherwise be present in the spectra and which will interfere with pattern matching. These peaks are usually removed by the instrument's software, however as this cannot be guaranteed to be the case the option also exists here.

5. **Mask specified regions**

Up to three regions of the dataset can be specified by the user. These areas are ignored when the patterns are being matched. This is useful for removing unwanted peaks or areas of a pattern that may be due to a reference material that has been added or should the user decide that this area should be disregarded for any reason.

6. **Set matching range subset**

Allows the user to ignore all but a specified region of the dataset. Only data within this area is matched. All other points in the pattern are ignored. This is useful as it allows the user to determine which areas of a spectra or pattern are most important and only match that area.

1.2.6 SIGNAL TRANSFORMS

Signal transforms can be applied to the dataset before pattern matching is carried out. The available options are as follows:

I. **Fourier transform**

Applies a Fourier transform to all patterns in the dataset then matches the patterns produced from the Fourier transform.

II. Derivative

Calculates a first or derivative of all patterns in the dataset then matches the patterns produced from the derivative. A derivative is a measurement of the rate of change in y when compared to x. The user specifies the order of derivative applied to the dataset. The derivative can be either first or second order. For the first derivative

$$f'(x) = \lim_{h \to 0}[f(x+h) - f(x)]/h$$

Equation 30 - First Derivative Equation

Where f is the function being studied, f' is the derivative of this function. If a tangent line is taken through a curve on the graph so that the tangent does not meet transversally (Figure 21), the initial point the line is passed through is point x and the second point is x+*h*. For optimum results *h* should be close to zero.

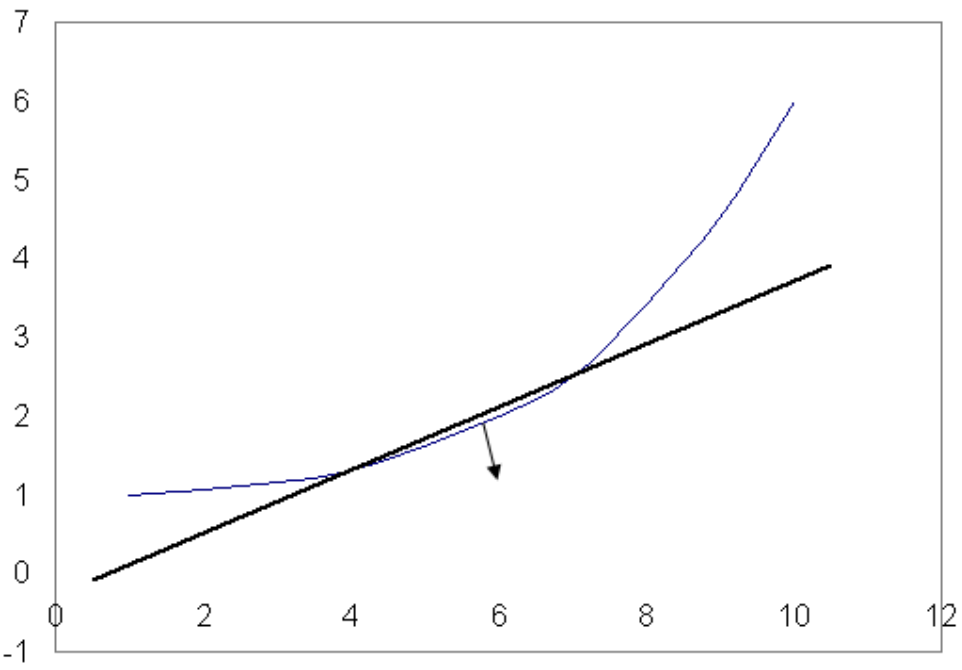

Figure 21 - First Derivative

1.2.7 QUANTITATIVE ANALYSIS MODE

Manual matching mode also matches patterns using the previously described methods; however it does not produce dendrograms and MMDS plots. It is most useful as a tool for quantitatively matching samples in a dataset, however is also useful when there are a small number of samples in a dataset that need to be examined in greater detail.

Quantitative analysis attempts to identify what individual components make up a mixture. Assuming that a sample pattern *S*, which is a mixture of *N* components with S consisting of data points $S_1 \ldots S_2 \ldots S_m$ and N database patterns making up fractions $P_1 \ldots P_2 \ldots P_N$ of the pattern, a series of equations will be built up with x_{11} being the first point in pattern 1:

$$\begin{aligned} &x_{11}P_1 + x_{12}P_2 + x_{13}P_3 + \ldots + x_{1N} + P_N = S_1, \\ &x_{21}P_1 + x_{22}P_2 + x_{23}P_3 + \ldots + x_{2N} + P_N = S_2, \\ &\vdots \\ &x_{m1}P_1 + x_{m2}P_2 + x_{m3}P_3 + \ldots + x_{mN} + P_N = S_m \end{aligned}$$

Equation 31 - Linear Equation for Quantitative Analysis

These can be written in matrix form as shown in Equation 32.

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1N} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & x_{m3} & \cdots & x_{mN} \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_N \end{pmatrix} = \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_N \end{pmatrix}$$

Equation 32 – Matrix form of Quantitative Analysis Linear Equations

in matrix notation:

$$\mathbf{x.p = S}$$

Equation 33 - Shortened form of Matrix Equation for Quantitative Analysis

A solution is sought where:

$$\chi^2 = \left|\mathbf{x.p} - \mathbf{s}\right|^2$$

Equation 34 - Quantitative Analysis Sought Minimum

As $N << m$, the method of least squares can be used.

Least squares can have problems with matrixes that are poorly conditioned. A matrix is conditioned by taking the ratio of the largest and smallest value in its corresponding diagonal matrix **W**. If this value is approaching infinity it is said to be singular, otherwise it is said to be poorly conditioned. Many powder patterns, especially if the full profile is taken, can be poorly conditioned. Poorly conditioned matrixes can be dealt with by using the singular value decomposition (SVD) method[27].

SVD decomposes the **x** matrix into constituent matrices to give:

$$\mathbf{p} = \mathbf{V}.\mathrm{diag}\left(1/w_j\right).\mathbf{U}^{\mathbf{T}}.\mathbf{S}.$$

Equation 35 - SVD Decomposed Matrix X Solution

W has elements that are either positive or zero. If most of these are small then matrix **p** can be approximated with a small number of terms of **S** (effectively producing the sample pattern from a combination of a few database patterns)**.**

A variance-covariance matrix can be obtained from matrix **V** and the diagonal of matrix **W**:

$$\mathrm{cov}(i,\ j) = \sum_{i=1}^{N}\left(\frac{\mathbf{V}_{ji}\mathbf{V}_{jk}}{w_i^2}\right)$$

Equation 36 - Variance-Covariance Matrix

This allows the calculation of variation in the component percentages. The fractional percentages in powder diffraction arise from the component mixtures scattering power, p_i-p_N.the values of p can be used to calculate the weight fraction for a particular phase provided that atomic absorption coefficents are known. This requires the unit-cell dimensions and cell contents[28]. The formula for the weight fraction of component n in a mixture of N is[29]:

$$c_n = p_n\mu^* / \mu_n$$

Equation 37 - General Formula for Weight Fraction

where

$$\mu^* = \sum_{j=1}^{N} c_j\mu_j^*$$

Equation 38 – Weight Fraction Calculations

and

$$\mu_j^* = \mu_j / \rho_j$$

Equation 39 - Weight Fraction Calculations

Where μ_j is the atomic X-ray absorption coefficient and ρ_j is the density of component *j*. The variance, or standard deviation, can be calculated using

$$\sigma^2(c_n) = \left[\frac{1}{(1-p_n)\mu_n^*}\right]\left[\frac{1}{(1-p_n)^2}\left(\sum_{\substack{j=1 \\ j\neq n}}^{N}\mu_j^* c_j\right)^2 \sigma^2(p_n) + p_n^2 \sum_{\substack{j=1 \\ j\neq n}}^{N}\left(\mu_j^*\right)\sigma^2(c_j)\right]$$

Equation 40 - Standard Deviation Calculation

A quantitative matching example is shown in Figure 22.

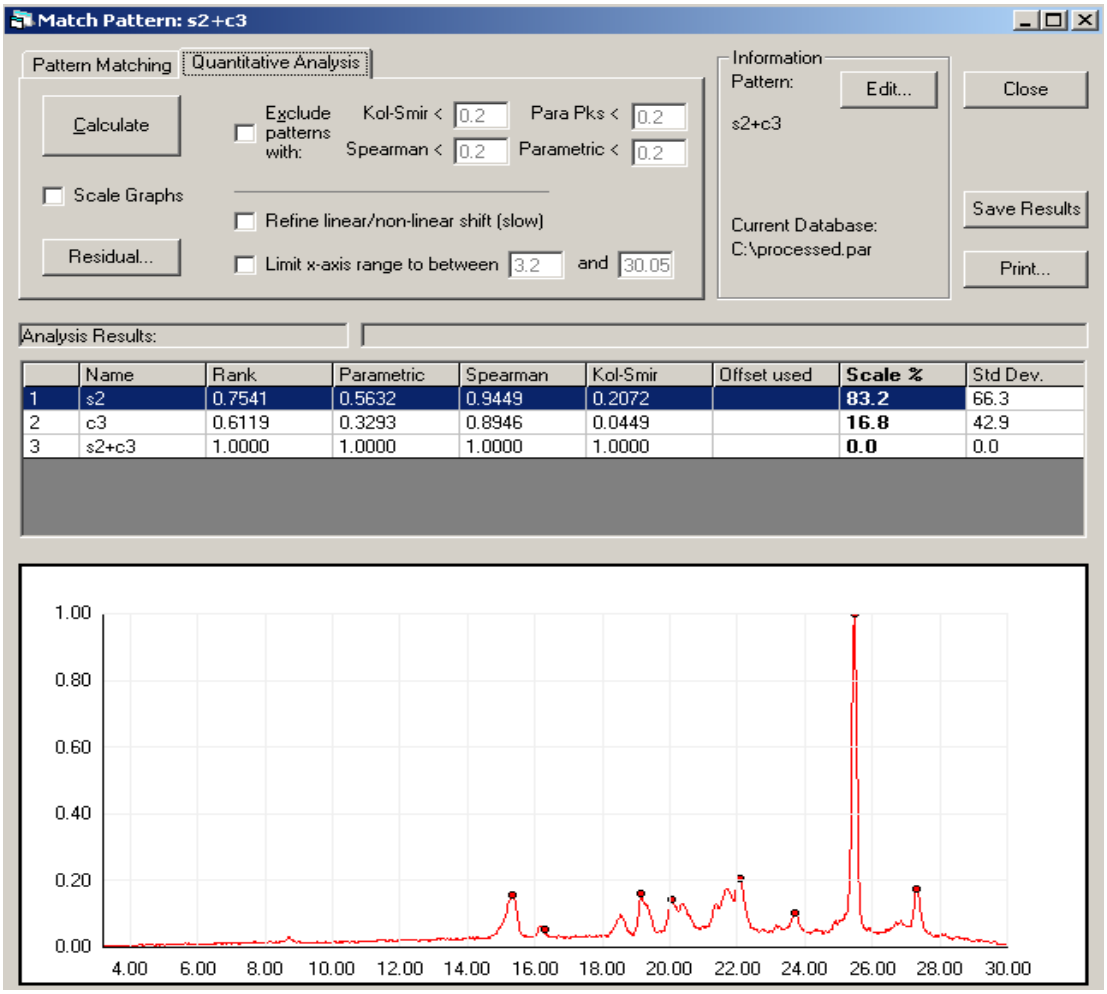

Figure 22 - Quantitative Matching in Manual Analysis Mode

The dots on the pattern represent the points at which the software has determined there to be a significant peak. This example shows a mixture of sulfathiazole form 2 and carbamazepine form 3 in an 80:20 composition. The predicted composition of 83.2:16.8 closely matches the actual composition.

1.3 OTHER PATTERN MATCHING SOFTWARE

At present the main alternative to PolySNAP is High-Score plus from PANalytical. This is a similar piece of software in that it allows pattern matching of PXRD patterns however it differs in a number of ways:

1. The software does not offer MMDS plots which have proven to be superior to PCA plots.
2. The software cannot analyse any data beyond PXRD data.

1.4 REFERENCES

1. C. J. Gilmore, G. Barr, J. Paisley (2004). "High-throughput powder diffraction. I. A new approach to qualitative and quantitative powder diffraction pattern analysis using full pattern profiles." *Journal of Applied Crystallography* **37**: 231-242.
2. G. Barr, W. Dong, C. J. Gilmore (2004). "High-throughput powder diffraction. II. Applications of clustering methods and multivariate data analysis." *Journal of Applied Crystallography* **37:** 243-252.
3. G. Barr, W. Dong, C. J. Gilmore, J. Faber (2004). "High-throughput powder diffraction. III. The application of full-profile pattern matching and multivariate statistical analysis to round-robin-type data sets." *Journal of Applied Crystallography* **37**: 635-642.
4. G. Barr, W. Dong, C. J. Gilmore (2004). "High-throughput powder diffraction. IV. Cluster validation using silhouettes and fuzzy clustering." *Journal of Applied Crystallography* **37**: 874-882.
5. G. Barr, G. Cunningham, W. Dong, C. J. Gilmore, T. Kojima (2009). "High-throughput powder diffraction V: The use of Raman spectroscopy with and without X-ray powder diffraction data." *Journal of Applied Crystallography* **42**: 706-714.
6. G. Barr, C. J. Gilmore, J. Paisley (2004). "SNAP-1D: a computer program for qualitative and quantitative powder diffraction pattern analysis using the full pattern profile." *Journal of Applied Crystallography* **37**: 665-668.
7. G. Barr, W. Dong, C. J. Gilmore (2004). "PolySNAP: a computer program for analysing high-throughput powder diffraction data." *Journal of Applied Crystallography* **37**: 658-664.
8. Gower, J. C. (1966). "Some Distance Properties of Latent Root and Vector Methods used in Multivariate Analysis." *Biometrika* **53**: 325-338.
9. G. W. Milligan, M. C. Cooper (1985). "An Examination of Procedures for Determining the Number of Clusters in a Data Set." *Psychometrika* **50**(2): 159-179.
10. Gordon, A. D. (1999). *Classification 2nd Edition*. Boca Raton, Chapman and Hall.
11. T. Calinski, J. Harabasz (1974). "A Dendrite Method for Cluster Analysis." *Communications in Statistics* **3**: 1-27.
12. L. A. Goodman, W. H. Kruskal (1954). "Measures of association for cross classifications." *Journal of the American Stastical Association* **49**: 732-764.
13. Everitt, B. S. (1993). *Cluster Analysis, 3rd Edition*. London, Arnold.
14. Sato, M. (1966). *Fuzzy Clustering Models and Applications*. New York, Physica Verlag.

15. J. D. Carroll, J. J. Chang (1970). "Analysis of individual differences in multidimensional scaling *via* an n-way generalization of "Eckart-Young" decomposition." *Psychometrika* **35(**3): 283-319.
16. J. A. Nelder, R. Mead (1965). "A simple method for function minimization." *The Computer Journal* **7**: 308-313.
17. D. L. Donoho, I. M. Johnstone (1994). "Adapting to unknown smoothness *via* wavelet shrinkage." *Journal of the American Statistical Association* **90**: 1200-1224.
18. Ogden, R. T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*. Boston, Birkhauser.
19. Stein, C. M. (1981). "Estimation of the mean of a multivariate normal distribution." *The Annals of Statistics* **9**: 1131-1151.
20. Spearman, C. (1904). "The proof and measurement of association between two things." *American Journal of Psychology* **15**: 72-101.
21. Rousseeuw, P. J. (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis " *Journal of Computational and Applied Mathematics* **20**: 53-65
22. L. Kaufman, P. J. Rousseeuw (1990). *Finding Groups in Data: an introduction to cluster analysis*, Wiley.
23. Inselberg, A. (1985). "The Plane with Parallel Coordinates." *The Visual Computer* **(1):** 69-91.
24. R. Moustafa, E. Wegman (2006). *Multivariate Continuous Data — Parallel Coordinates*, Springer.
25. R. G. Gallager, P. A. Humblet, P. M. Spira (1983). "A Distributed Algorithm for Minimum-Weight Spanning Trees." *ACM Transactions on Programming Languages and Systems* **5**(1): 66-77.
26. G. Barr, W. Dong, C. J. Gilmore (2009). "PolySNAP 3: a computer program for analysing and visualizing high-throughput data from diffraction and spectroscopic sources." *Journal of Applied Crystallography* **42**: 965-974.
27. W. H, Press, S. A. Teukolsky (1992). *Numerical Recipes in C: the art of scientific computing*. Cambridge, Cambridge University Press.
28. G. Cressey, P. F. Schofield (1996). "Rapid whole-pattern profile-stripping method for the quantification of multiphase samples." *Powder Diffraction* **11**(1): 35-39.
29. J. Leroux, D. H. Lennox, K. Kay (1953). "Direct quantitative X-ray analysis by diffraction-absorption technique." *Analytical Chemistry* **25(**5): 740-743.

30. Fisher, R. A. (1915). "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population." *Biometrika* **10**(4): 507-521.

CHAPTER 2 DATA MEASUREMENT TECHNIQUES USED

All of the following techniques include an example of the type of pattern or spectra that will be produced by the instrument collecting the data. For consistency all of these patterns are from sulfathiazole form 3. The structure of sulfathiazole is shown in Figure 23.

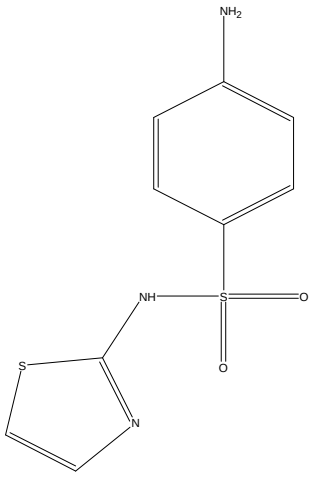

Figure 23 - Sulfathiazole structure

Sulfathiazole form 3 has the following crystallographic parameters[13]:
Cell lengths: **a** 17.570(9)Å **b** 8.574(4) Å **c** 15.583(8) Å
Cell angles: **α** 90 **β** 112.93(1) **γ** 90
Space group: $P2_1/c$
Z: 8, Z' = 2

2.1 POWDER X-RAY DIFFRACTION

2.1.1 X-RAY DIFFRACTION BACKGROUND

X-ray diffraction occurs when a beam of X-rays interacts with a crystalline material. X-ray diffraction can be described using the Bragg method, developed by W. L. Bragg,[8, 11, 12] which is shown in Figure 24.

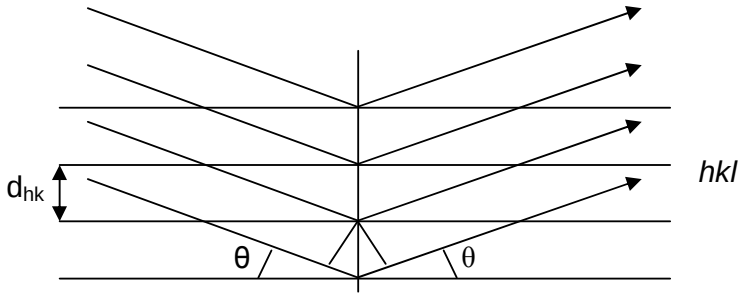

Figure 24 - Diffraction from Bragg lattice planes

The Bragg diffraction model is a geometric model in which X-rays are diffracted off of sets of parallel planes passing through the crystal. Each plane is defined by three integers, given the symbols *h, k* and *l*, which define the planes orientation with respect to the unit cell edges. The spacing between the planes is given the symbol d_{hkl} as its value is determined by the geometry of the crystal lattice. The angle of incidence and reflection for the X-ray beam are identical and are given the symbol θ. Reflection from adjacent planes gives interference as it is unlikely that the beams will still be in phase. The wavelength of the X-ray beam is given the symbol λ. The Bragg equation (Equation 41) shows the condition required for diffraction to occur:

.

$$n\lambda = 2d_{hkl} \sin\theta$$

Equation 41 - Bragg equation

All values have been previously defined except *n* which is an integer, usually 1.
The scattering of X-rays from a single crystal produces a diffraction pattern. Figure 25 is an example:

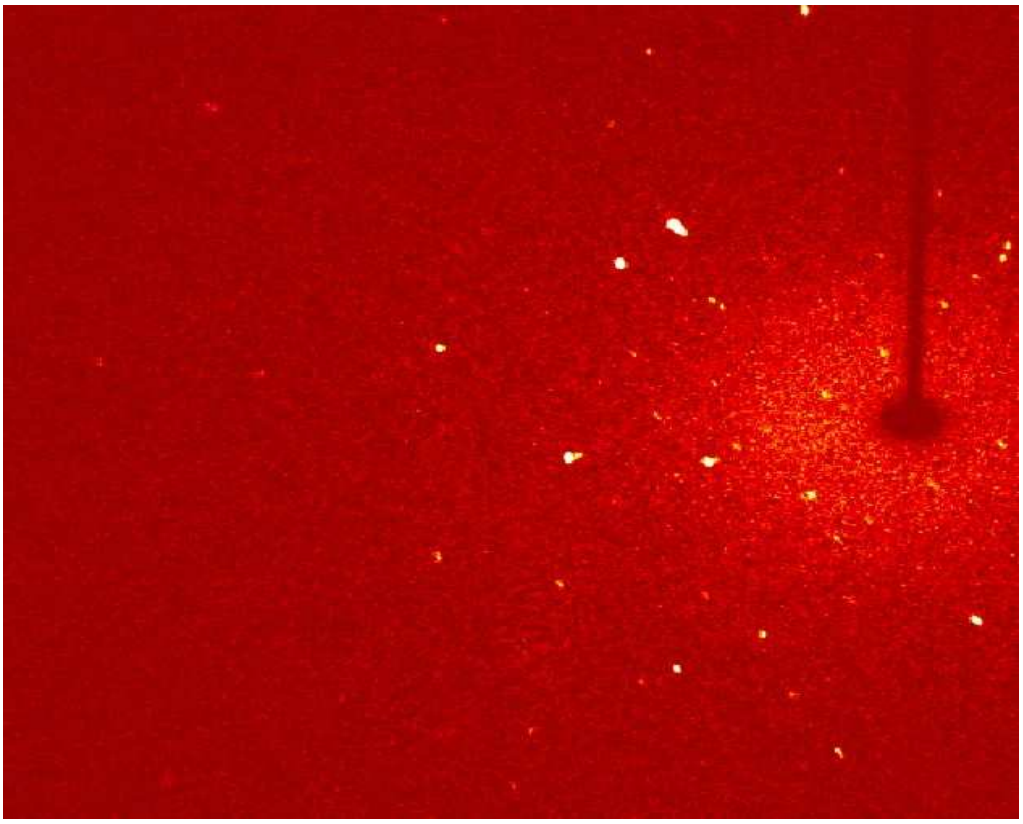

Figure 25 - Diffraction pattern

2.1.2 SINGLE CRYSTAL DIFFRACTION

Single crystal x-ray diffraction is a non-destructive technique which is used to provide information on a crystal structures internal lattice. This data includes the unit cell lengths, bond lengths and bond angles.

For a 2-dimensional array of molecules, the crystal lattice is defined by positioning a point on the same position on every molecule in the unit cell. The resulting array of regularly spaced points would give the lattice structure of the crystal.

The unit cell is defined by taking 4 of these points which form a parallelogram with 2 pairs of identical sides, a and b and 1 included angle γ. This can be repeated in all directions to build up a crystal. The ideal parameters for a unit cell are for the sides to be as short possible ($a \leq b$) and with γ as close to 90º as possible. A representation of a lattice structure and 2 potential unit cells are shown in Figure 26. A unit cell with the sides and angle marked are shown in Figure 27.

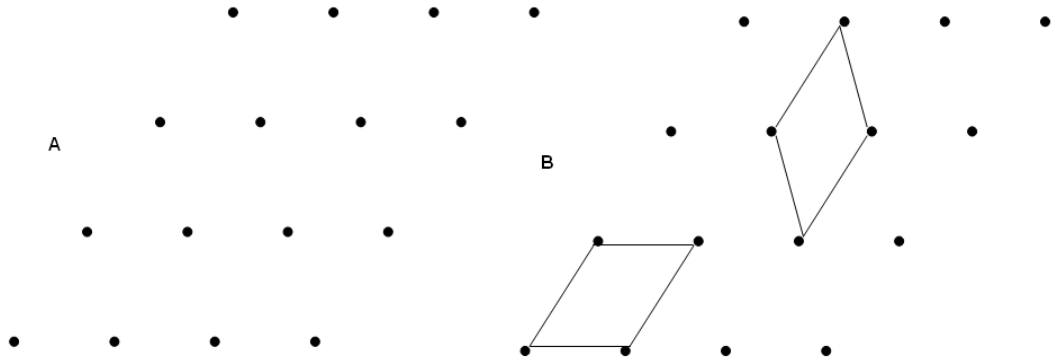

Figure 26 A – Lattice Structure Diagram. B – With potential unit cells drawn

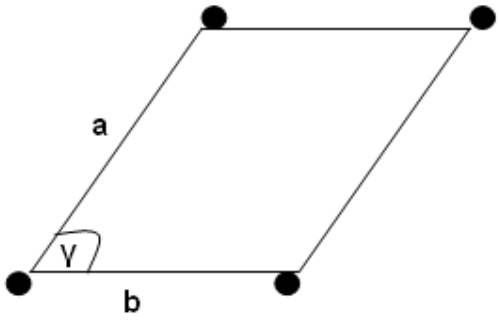

Figure 27 - Unit cell

When moving up to 3 dimensions an additional side, c, is added as are two additional angles α and β.

2.1.3 POWDER X-RAY DIFFRACTION

Powder X-ray diffraction (PXRD) data are measured on a powder containing, ideally, micro-crystals in every possible alignment.[1, 9, 10] As such, diffraction patterns should be produced in every possible orientation. These patterns will form rings, such as those in Figure 28.

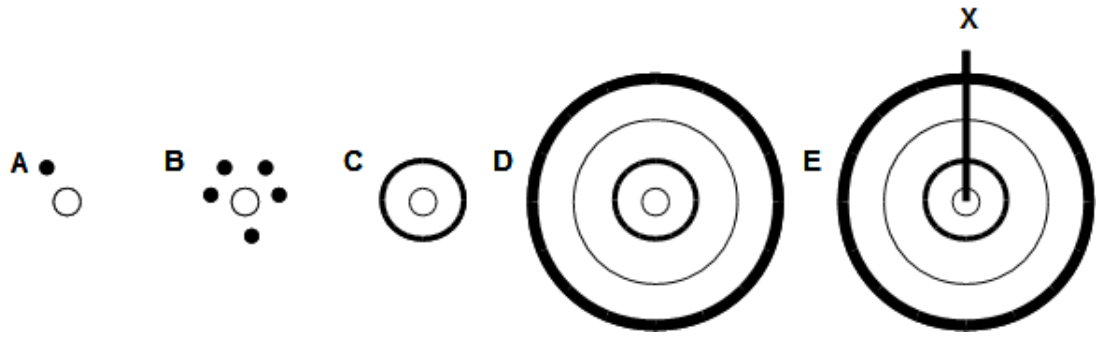

Figure 28 - Powder Diffraction diagram A – Reflection from a single crystal. B – Reflection from five crystals. Each crystal has a different orientation C – Reflection from crystals with all possible orientations. D – Complete powder diffraction pattern. E – Method of measurement of a diffraction pattern.

A single crystal with only one reflection would give the pattern shown in Figure 28–A. Five of these crystals in different orientations would give the pattern shown in Figure 28–B. If this crystal is present in all orientations then the reflections will form circles, shown in Figure 28–C. Full diffraction patterns will merge and form a series of concentric rings such as those in Figure 28–D. A point detector is passed outwards from the centre of the rings, cutting through each ring, as shown in Figure 28–E. As the detector passes over the rings (Line X in Figure 28-E), a PXRD pattern is generated. An example PXRD pattern is shown in Figure 29.

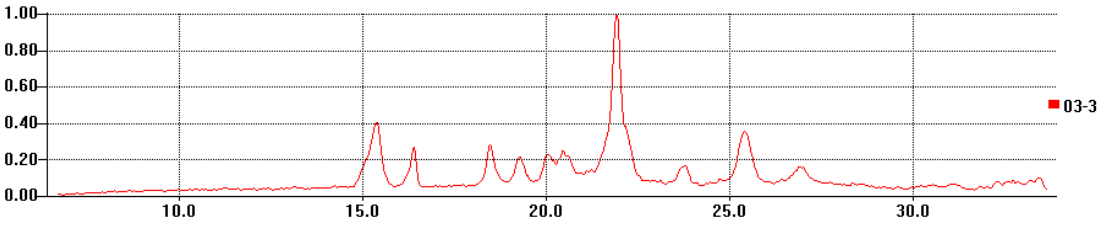

Figure 29 – Sulfathiazole Form 3 PXRD pattern

It is also possible to collect the data using an area detector. An area detector will capture part or all of the arc of rings produced during powder diffraction rather than collecting data along a single line across the rings. Each ring can then be integrated to produce a single data point which can then be plotted as a 1D powder pattern. This can give a more accurate result as it will minimise the effects of small amounts of preferred orientation. If the sample is suffering from large preferred orientation issues it will not, however, fully resolve them.

2.1.4 PREFERRED ORIENTATION

Preferred orientation is a major problem in powder diffraction. If the crystals in the powder are not lying in all possible geometries, but instead have arranged themselves into a regular configuration, certain peaks in the PXRD pattern will dominate and will swamp the pattern. Figure 30 shows an example of this with the red line showing a pattern without preferred orientation and the blue line showing a PXRD of the same material but with preferred orientation. The preferred orientation peak is circled.

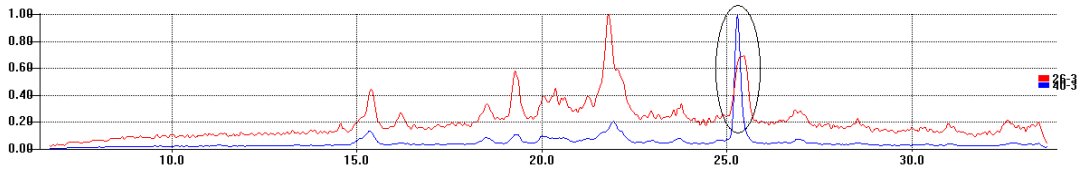

Figure 30 - Sulfathiazole Form 3 PXRD pattern with preferred orientation

This regular configuration of crystals commonly arises due to the shape of the crystals causing them to align in a common direction. For example, crystals that are flat plates will prefer to stack on top of one another while needles will prefer to lie side by side. In general the smaller the crystal, the less of a problem preferred orientation is. As such preferred orientation can be reduced by carefully grinding the crystals during preparation. There is however a danger in grinding in that the heat generated from it can induce phase changes in the samples, resulting in a different polymorph, or even a mixture of a different polymorph and the original polymorph being present in the ground samples than is present in the original material.

The material can be prepared in capillaries, flat plates or well plates.

Flat plate samples can be used for either reflection mode, where the material is placed at an angle such that the beams reflect from the powder to the detector or for transmission mode where the material passes through the sample. A flat plate is prepared by placing the material on a flat plate and smoothing the surface evenly. The plate is spun along the vertical axis during collection. Flat plate samples are easier to prepare than capillaries, however can suffer from severe preferred orientation problems if the materials are insufficiently ground as well as having poorer signal to noise ratios.

Well plates are similar to flat plates, however rather than loading a plate and smoothing its surface, a small well is filled with the material and its surface smoothed. Well plates are used for high-throughput collection as a well plate will contain many wells each of which can contain a different material. Due to their similarities to flat plate collection, well plates can suffer from the same preferred orientation problems.

Capillaries can only be used with transmission geometry. A capillary is loaded with the material to be studied, the end is sealed and the filled capillary mounted on the instrument. When loading a capillary, the material must be loaded to the correct height within the capillary so that the material fully interacts with the x-ray beam. The material is spun along the capillary axis during data collection to try to minimise the preferred orientation problem. Capillaries are less sensitive to preferred orientation problems than flat plates or well plates however they are much more time consuming to prepare.

2.2 RAMAN SPECTROSCOPY

2.2.1 RAMAN BACKGROUND

All Raman spectroscopy data was collected using a Witec Alpha 300 with a 300-785nm laser and x10 objective lens, 0.25mm aperture and 30g/mm grate.

Raman spectroscopy[2, 3, 4, 7] operates using inelastic scattering of light when it interacts with matter.

There are two possible outcomes for the interaction of light and matter

1. The light is absorbed
2. The light is scattered.

When absorption occurs, a photon is absorbed in order to promote a molecule to an excited state. This occurs when the energy of the photon matches the energy gap between the ground and excited state of the molecule.

There are two types of possible scattering - elastic scattering where the incident and reflected beam have the same energy, referred to as Rayleigh scattering and inelastic scattering where the beams have different energy.

Inelastic scattering can occur by two methods. Both methods begin with the absorption of a photon, promoting the molecule to an excited state. The electron can then fall back down to an excited state which is higher than the initial ground state, giving Stokes scattering, or it can be promoted from an excited state and fall back to the ground state giving anti-Stokes scattering. Both types of scattering are shown in Figure 31.

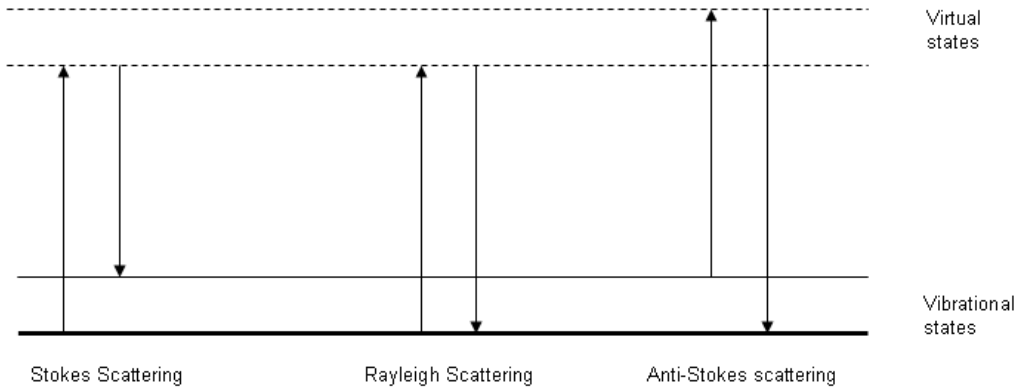

Figure 31 - Types of Raman Scattering

At room temperature, more electrons are likely to be found in a non-excited state than an excited state. This leads to Stokes scattering appearing much stronger than anti-Stokes scattering. In general, only Stokes scattering is recorded. An example Raman spectrum is shown in Figure 32.

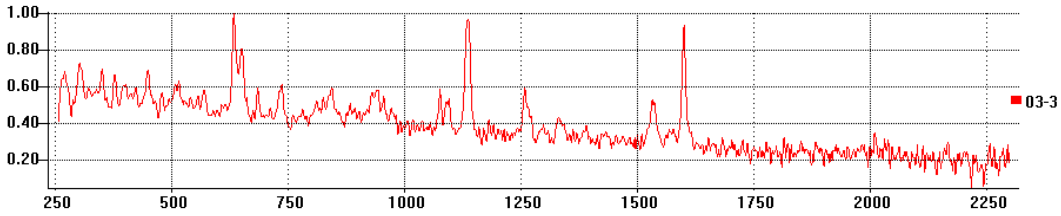

Figure 32 - Sulfathiazole Form 3 Raman spectrum

The x-axis is measured in reciprocal centimetres (cm^{-1}).

2.2.2 PROBLEMS WITH RAMAN DATA

A commonly encountered problem with Raman data in a high-throughput environment is to have substantial regions of the spectra which are highly similar, or identical for each sample.

Due to differences in the pattern background, the spectra in Figure 33–A look to be different, however, with background removal applied, as shown in Figure 33-B, the similarities between the two patterns can be clearly seen.

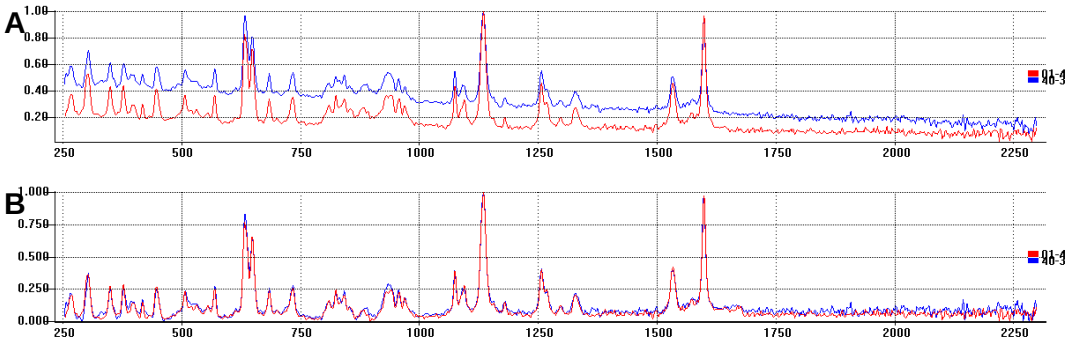

Figure 33 - A - Overlay off sulfathiazole forms 3 and 4 spectra, B - Overlay of sulfathiazole forms 3 and 4 spectra with background removed

This high similarity makes it very difficult to distinguish between individual spectra. As such it is not uncommon to have Raman datasets where all patterns show a similarity of greater than 90% as measured by correlation coefficients.

This problem can sometimes be resolved by processing the data into first or second derivative form before pattern matching. First derivative data shows much clearer separation of peaks; however this comes with a trade-off in that the clustering can be dramatically different and the spectra can be much noisier. An INDSCAL combination of the original and derivative data can resolve both problems by giving a dendrogram which shows much clearer separation and often maintains the clustering that is shown in the original dataset. An example of a Raman spectrum with the original data and first and second derivatives is shown in Figure 34.

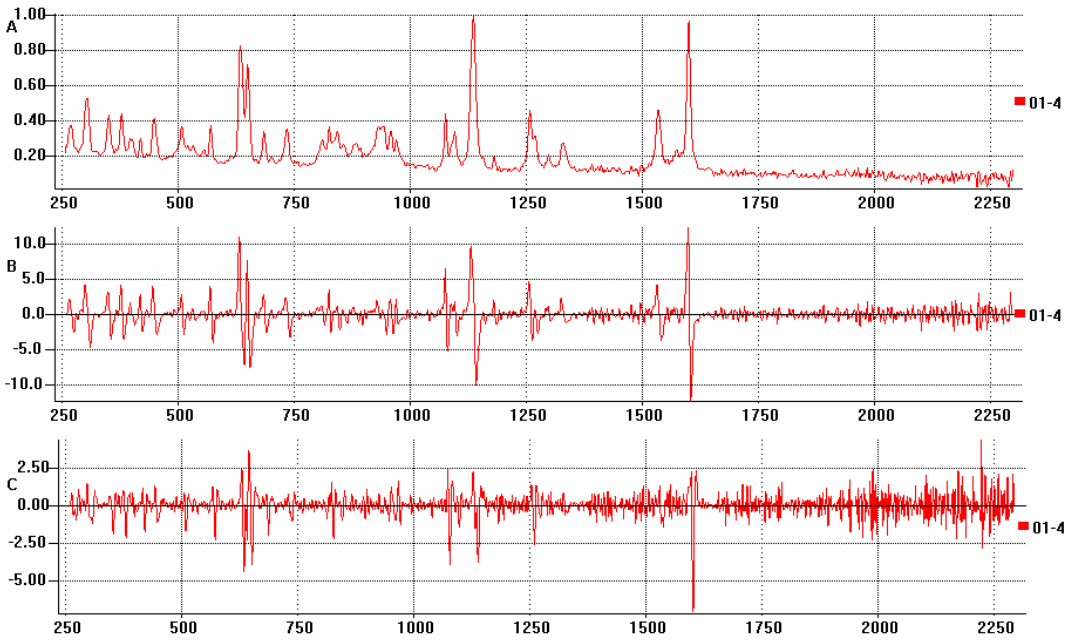

Figure 34 - A - Original Raman Spectra, B - 1st Derivative Raman Spectra, C – 2nd Derivative Raman Spectra

Applying a first derivative will give a noticeable increase in the tie bar heights in the corresponding dendrogram between materials where as a second derivative will give an even larger increase in tie bar heights. The second derivative samples however show much poorer clustering. Combining all three types of data using INDSCAL gives good clustering with clear separation between samples.

An example of Raman combination, using the ‘multiple 1’ dataset will now be shown. This dataset will be covered in more detail in Chapter 4. Figure 35 shows the original datasets dendrogram, Figure 36 shows the dendrogram for the first derivative data; Figure 37 shows the dendrogram for the second derivative data and Figure 38 shows the dendrogram for the combined dataset.

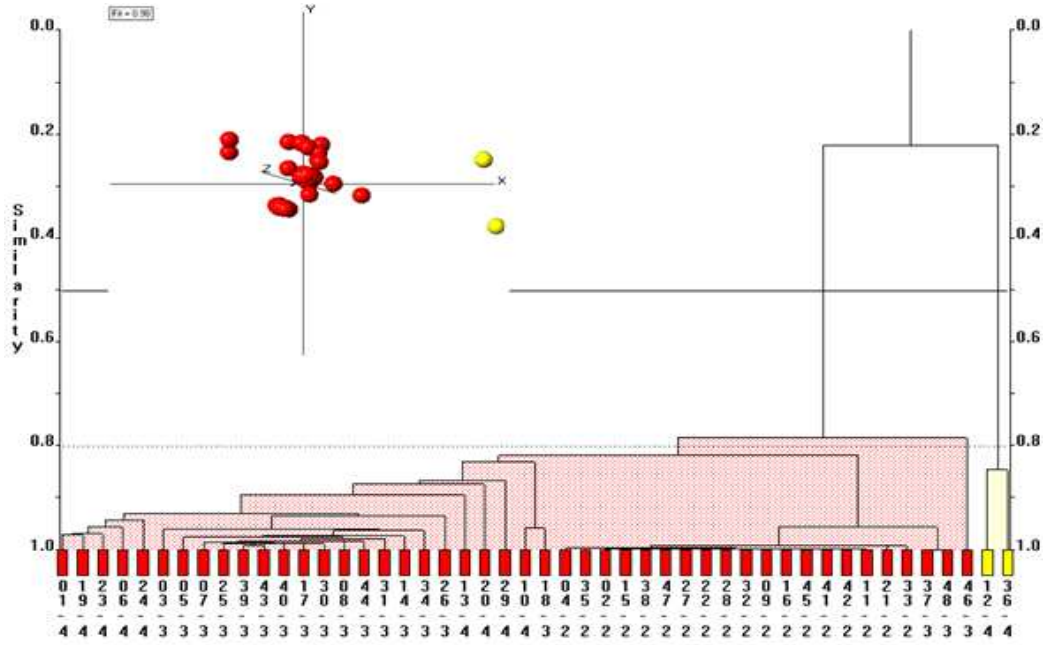

Figure 35 - Original Raman Data

The original data shows clear separation of the form 2 samples with intermixing of the form 3 and 4 samples. The form 2 samples cannot be clearly differentiated from one another as many tie bars are near 1.0.

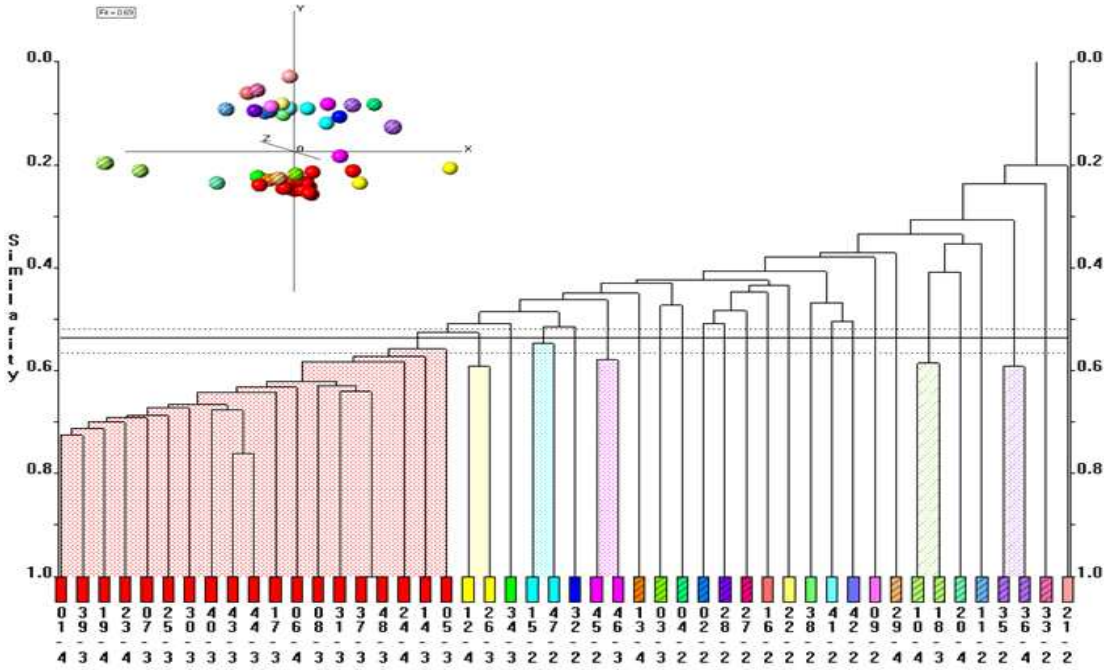

Figure 36 - First Derivative Raman Data

The first derivative data shows dendrogram chaining, implying that the dendrogram is not showing good clustering.

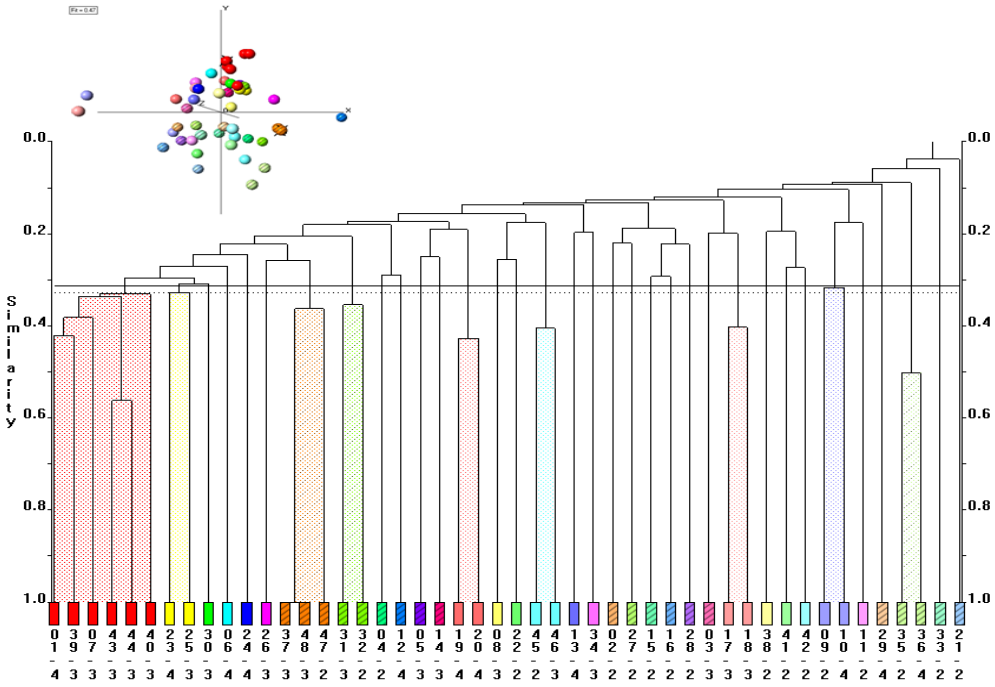

Figure 37 - Second Derivative Raman Data

The second derivative data also shows chaining, as well as not showing any clear clusters.

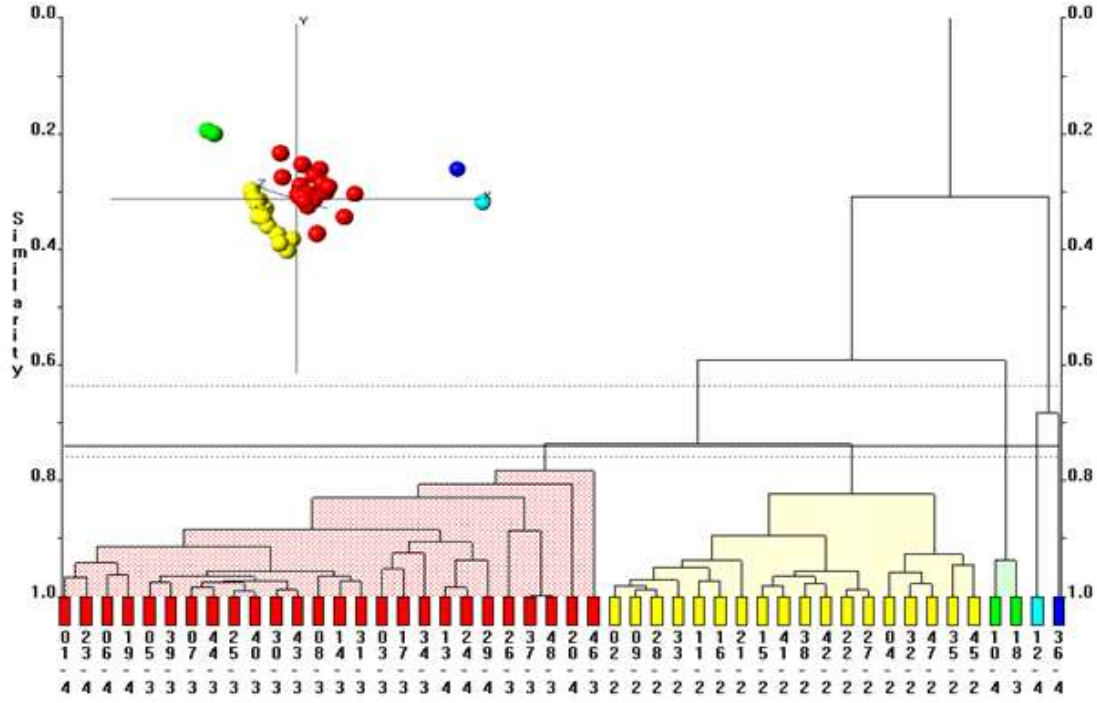

Figure 38 - INDSCAL Combined Raman

The combined data shows the form 2 samples clearly separated with the forms 3 and 4 sample intermixed. This dendrogram however shows clearer separation of the samples than that seen in the original dataset, especially for the form 2 samples.

2.2.3 RAMAN ANALYSIS

A new method for matching Raman spectra is being developed which looks for significant peaks in a pattern and gives these peaks priority over less significant peaks during pattern matching. Significant peaks are treated as being in three groups.

1. Very significant peaks – a peak present in 1 pattern but not in the other.
2. Significant peaks – A peak present in both patterns but with differing heights.
3. Insignificant peaks – A peak present in both patterns with the same height in both patterns.

The following diagram (Figure 39) shows 2 example patterns which show all 3 types of peak.

- The peaks marked as 1 in each pattern fall into the second category.
- The peaks marked as 2 and 3 in each pattern fall into the third category
- The peak marked as 4 falls into the first category

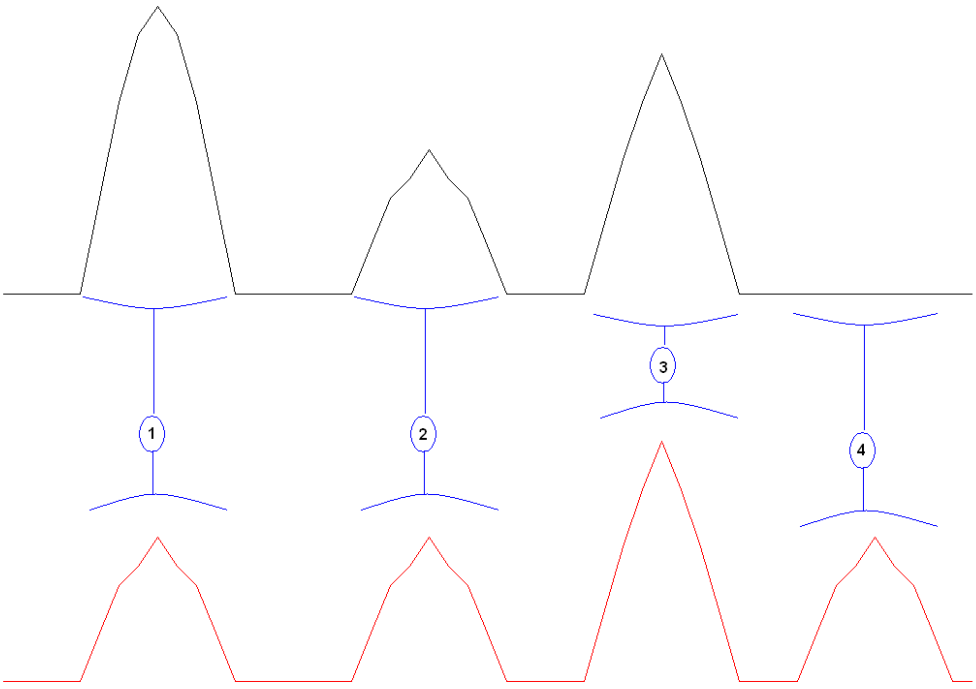

Figure 39 - Examples of Different Peak Types

This will allow for clearer separation of clusters as large areas of similarity, which would increase the overall similarity of the spectra, will be suppressed. This method is currently carried out by examining the spectra by eye, determining areas that are showing low similarity, therefore allowing a determination of the areas of high similarity, and setting the software to only match the high similarity areas.

2.3 DIFFERENTIAL SCANNING CALORIMETRY

2.3.1 HEAT FLUX DSC BACKGROUND

All differential scanning calorimetry (DSC) data was collected using a heat flux DSC instrument.[5] Two types of instruments have been used, a TA instruments Q100 and TA instruments Q1000.

Before measurement of sample data, the DSC instrument is calibrated using a known standard, for example, sapphire disks, which allows the instrument to determine how much energy is needed to raise the temperature of an empty pan by a known amount, for example 10ºC a minute. A schematic diagram of a DSC instrument is shown in Figure 40.

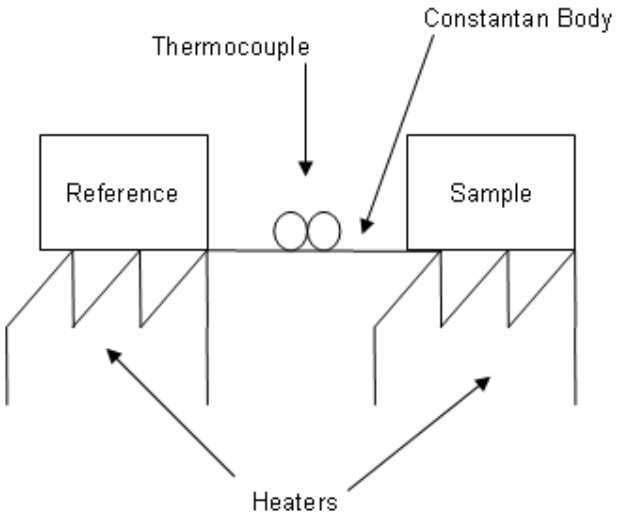

Figure 40 - Heat Flux DSC Schematic

The DSC experiment requires both a reference, which is an empty aluminium pan, and a sample, an aluminium pan containing a small amount of the material being studied. Typically 2-5mg is enough to produce a good DSC pattern.

The amount of energy required to heat the empty pan at the predetermined rate is supplied to both pans using the heaters. The pan containing the sample will heat at a different rate

from the empty reference pan. This difference in temperature will allow heat to flow through the body from the warmer pan to the cooler pan. This heat flow is detected by the thermocouple. Heat flow from the reference to the sample is indicative of the sample melting, a process which requires a large intake of energy with very little change in temperature. Heat flow from the sample to the reference is indicative of a crystallisation which produces a large amount of energy, resulting in more rapid heating of the sample pan.

The instrument has a large range of possible heating rates; however, 10 ºC min^{-1} is commonly used. A slower heating rate allows peaks to be more precisely positioned, however it decreases the height of all peaks. A rate of 10 ºC min^{-1} gives a good balance between peak position accuracy and peak height. An example DSC pattern is shown in Figure 41.

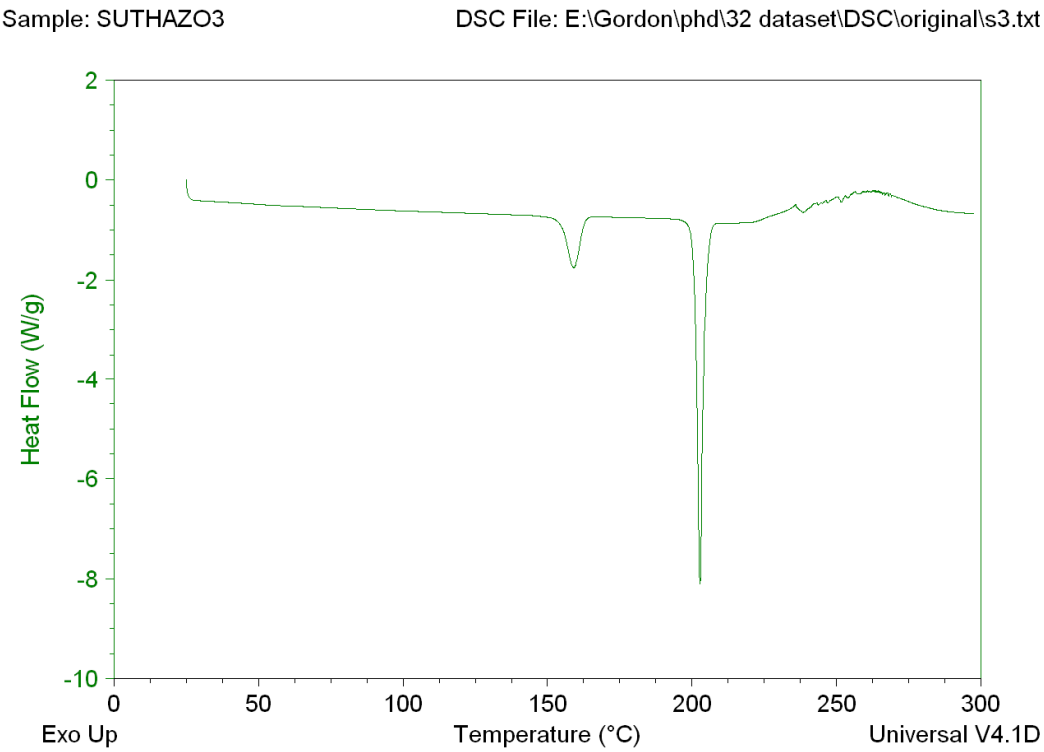

Figure 41 - Example DSC Pattern

The pattern in Figure 41 also shows that the material undergoes degradation after 100ºC. The example pattern does not show any crystallisation peaks. However the two peaks present at 144-153ºC and 193-202ºC are examples of melting peaks.

The melting points are measured from the point where the initial downwards slope of the peak begins to the tip of the peak. The return slope from peak tip to baseline does not contain any data on the samples melting. For this setup melting peaks always point

downwards while crystallisation peaks always point upwards. For melting, this is due to heat flowing from the hotter reference pan to the cooler sample pan as the sample takes in energy to melt, thus giving a negative heat flow. For the crystallisation peaks this is due to heat being produced by the material crystallising flowing from the hotter sample pan to the cooler reference pan.

2.4.2 PROBLEMS WITH DSC DATA

Dependent on the instrument, DSC data cannot always be directly read by the PolySNAP software due to issues with the file encoding on the instrument. The output file from both the Q100 and Q1000 instruments are encoded in such a way that ever character in the file is followed by a binary character. A program has been written which allows these extra characters to be stripped out. The code for this program can be found in Appendix II.
It is also possible for the DSC pattern to occasionally show a small loop where the temperature cools slightly during a heating ramp. This commonly occurs near the start of a heating cycle, usually within the first 2ºC of heating. The program for stripping out the excess binary characters also searches the data and strips out any data loops. An example of the sort of small loops that can occur is shown in Figure 42.

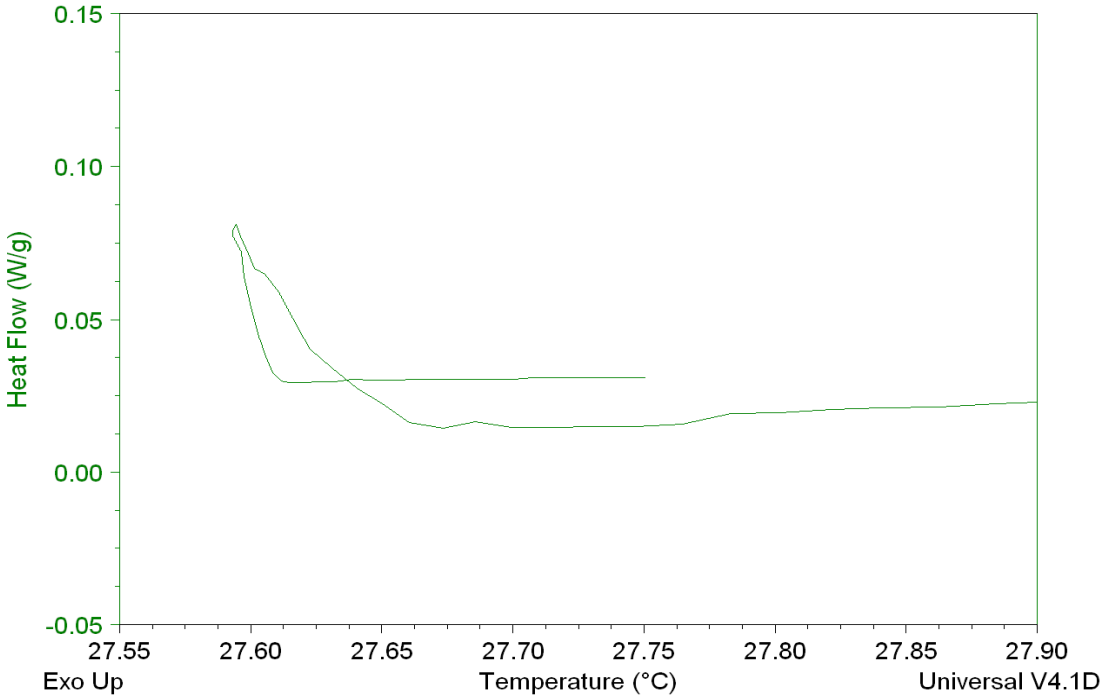

Figure 42 - DSC Starting Loop

The pattern in which this loop occurred is shown in Figure 43.

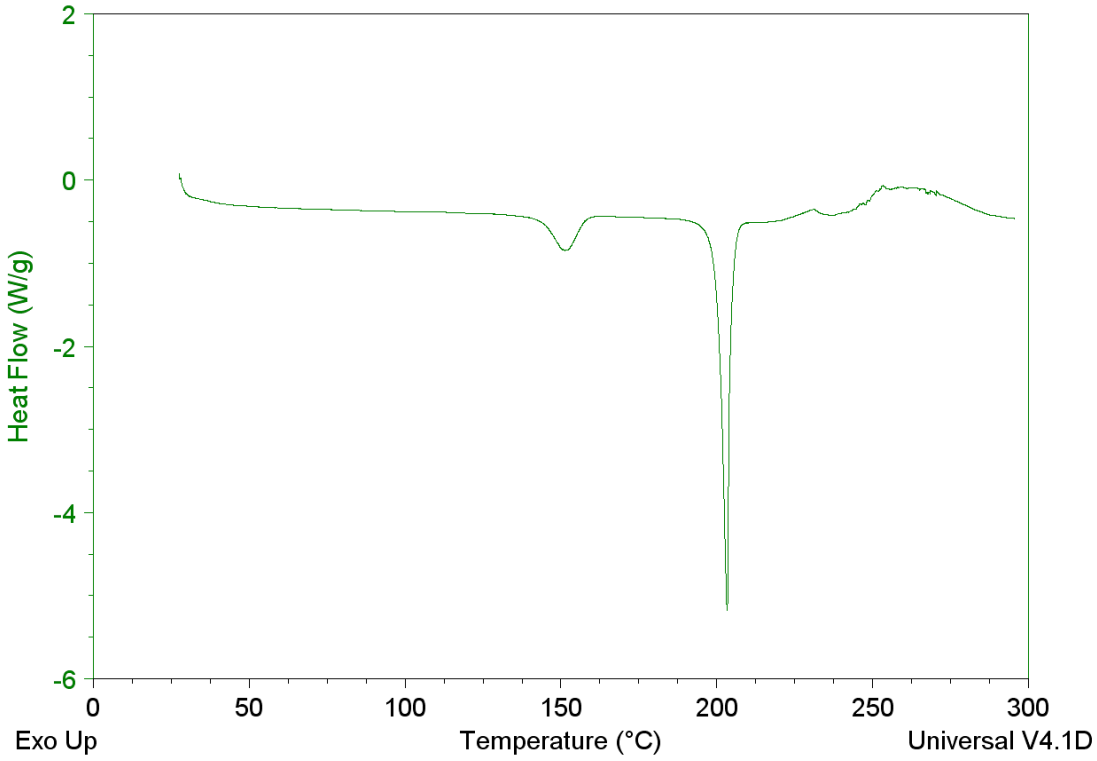

Figure 43 - DSC Pattern

In the full pattern the loop, occurring over a very small range at the very start of the pattern, is not visible.

2.4 THERMAL GRAVIMETRIC ANALYSIS

Thermal gravimetric analysis (TGA) is a technique which measures the change in the mass of a sample as it is heated.

2.4.1 TGA BACKGROUND

A TGA schematic is shown in Figure 44.

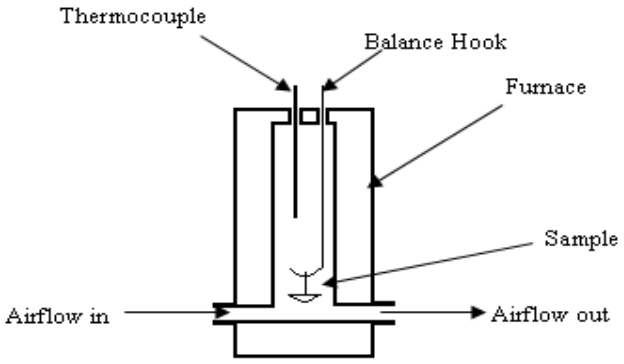

Figure 44 - TGA Schematic

An empty platinum-alumina pan is weighed in the instrument at room temperature and the pan has some of the material to be studied, typically 5-10mg, added. The furnace is heated at a pre-determined heating rate, typically 15ºC min^{-1}. As the material is heated it melts, resulting in a small decrease in mass due to vapour loss and material decomposition and eventually boils resulting in a large decrease of mass in the pan. The loss of mass is plotted against temperature, as shown in Figure 45.

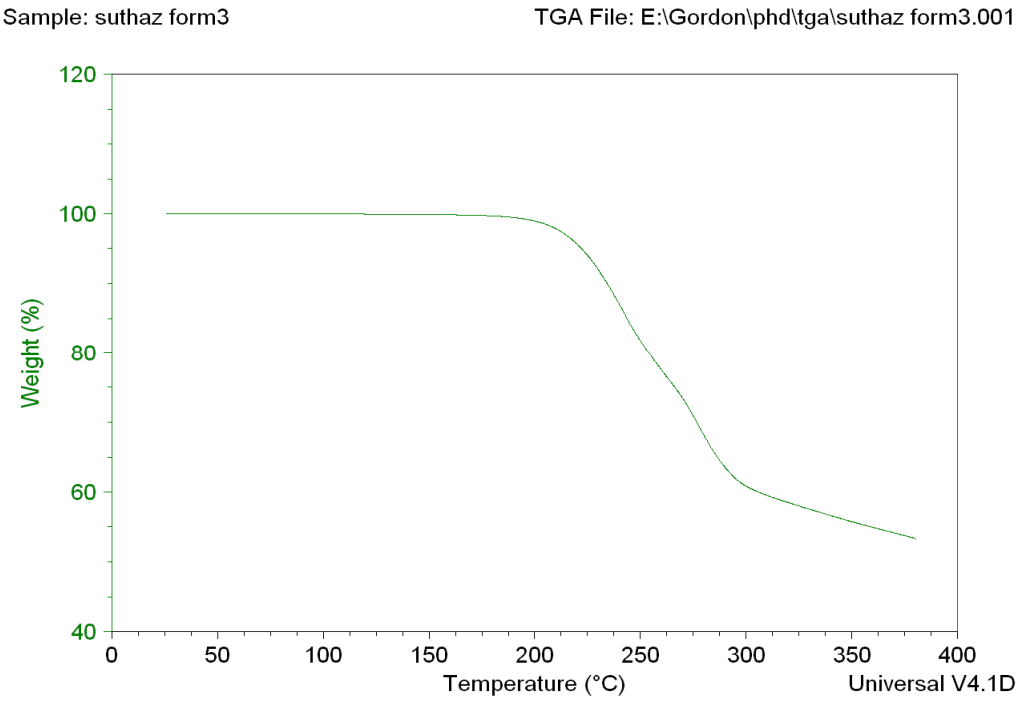

Figure 45 – Sulfathiazole Form 3 TGA pattern

This example does not show the result of the material completely boiling away; however the continual drop, starting around 199C, corresponds to the previously discussed degradation of the material.

2.5 INFRARED

2.5.1 IR BACKGROUND

All infrared data was collected using a Fourier transform infrared (FTIR) spectrometer.[6] Two types of instrument were used, the first is a Shimadzu FTIR-8400S which uses a technique called attenuated total reflectance (ATR) to collect the spectrum. The material is placed on top of a small diamond window and pressed down using an attached clamp. The IR beam then passes through the sample, reflects off the base of the clamp, and returns to down to the collector. The second type of instrument used was a JASCO FT/IR 4100. For this instrument the material is pressed into a disk with potassium bromide (KBr) used to give the material bulk. The IR beam is then passed through the disk. Unlike the ATR technique, the KBr one is destructive to the sample.

Figure 46 is a schematic of a Michelson interferometer, which was invented by Albert Abraham Michelson in 1880.

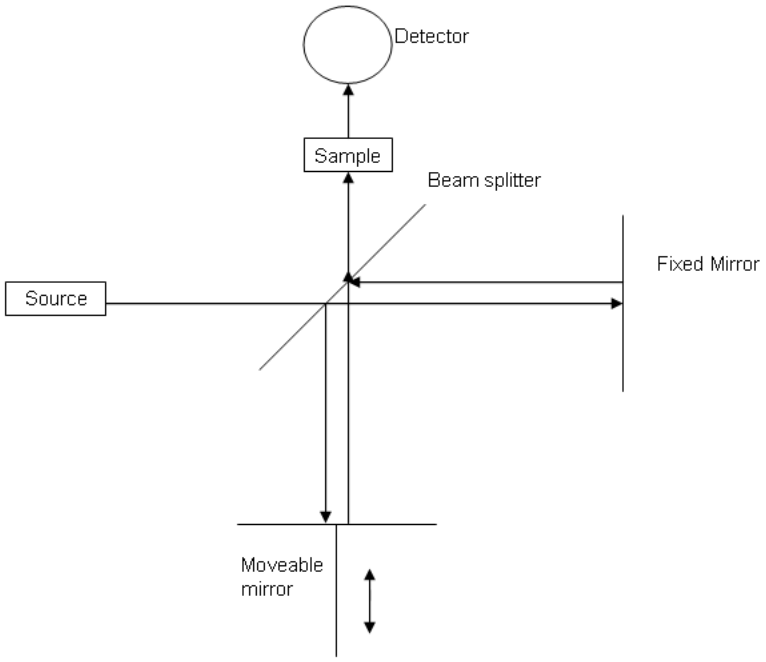

Figure 46 – Optical Diagram of a Michelson interferometer

The Michelson interferometer functions by taking a beam of light (the source) and splitting it into two beams using a beam splitter. The beam splitter consists of a half silvered mirror which lets half the light pass through towards a fixed mirror and reflects the other half towards a moveable mirror. The beams are reflected back towards the beam splitter, where they are recombined and proceed onto a detector.

If the two mirrors are equidistant from the beam splitter, the distance travelled by both beams will be the same. This is known as the zero path distance (ZPD). A mirror displacement is produced by moving the mirror away from the ZPD which is given the symbol Δ. Moving the mirror creates an optical path difference between the two beams which is related to the mirror displacement by the following equation

$$\delta = 2\Delta$$

Equation 42 - Optical path difference equation

If the amplitudes of the two beams are in phase they will combine constructively and have a high intensity. If they are out of phase they will combine destructively and have a low intensity. The detector measures the variation of light intensity with optical path difference. When plotted this is known as an interferogram. A complete interferogram is produced by moving the mirror back and forth once, a process known as a scan. The signal-to-noise ratio of a sample can be reduced by combing multiple interferograms. As such multiple scans are carried out on each sample measurement. The interferogram can be Fourier transformed into a spectrum.

The velocity of the movable mirror must be carefully controlled and monitored to allow the instrument to consistently produce a repeatable interferogram and allow experiments to be repeated.

An example FTIR spectrum is shown in Figure 47.

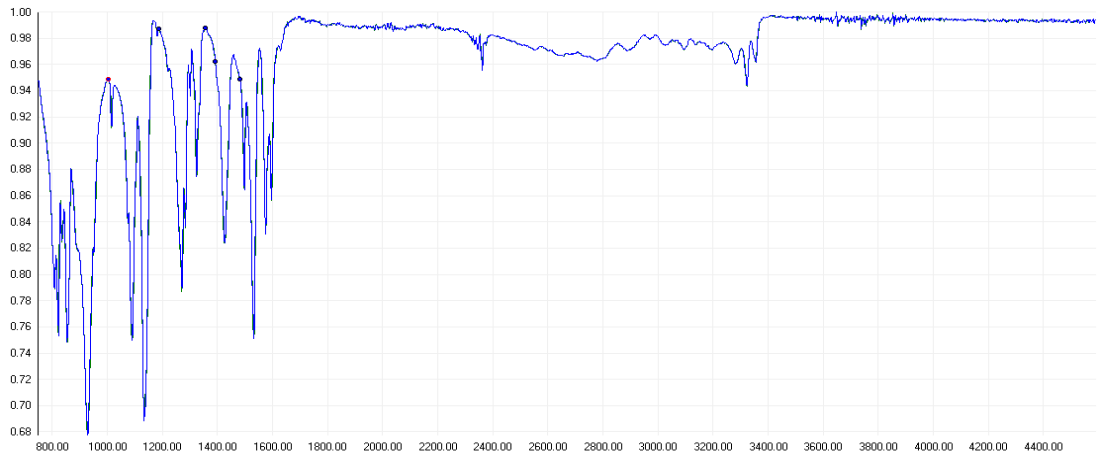

Figure 47 - Sulfathiazole form 3 IR spectra

An interpreted version of this IR spectrum can be seen in Figure 48.

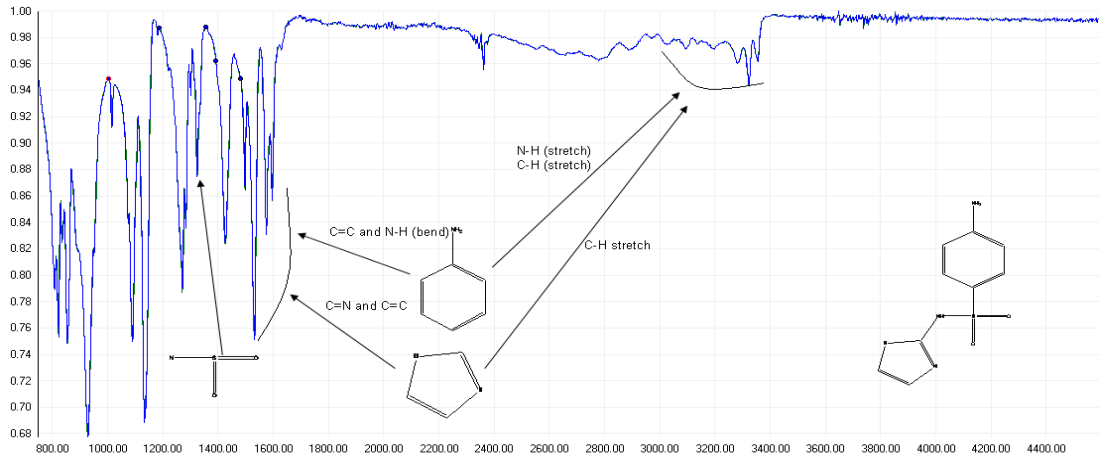

Figure 48 - Interpreted sulfathiazole form 3 IR

Both the 5 and 6-membered rings are aromatic and have peaks in the aromatic regions of the IR spectrum. These interpretations are neither used nor required by PolySNAP for analysis.

2.5.2 PROBLEMS WITH IR DATA

As with Raman data, a commonly encountered problem with IR is the occurrence of similar or identical regions found in all spectra in a dataset. The differences between these spectra can be subtle as shown in Figure 49-A.

These patterns show a 98% similarity making it difficult to differentiate between these two patterns. If the long 'tail' after 1750cm^{-1} is ignored, the patterns now show a similarity f 99%. This problem can be partly resolved by applying a 1st derivative to the each pattern before clustering them. The 1st derivative of these two patterns is shown in Figure 49-B.

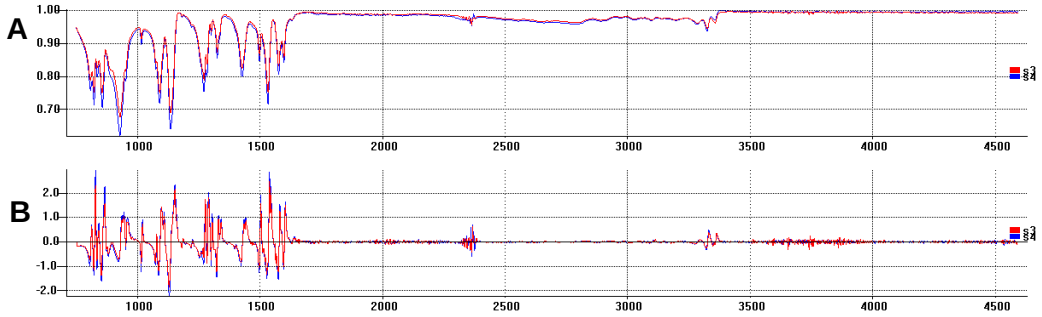

Figure 49 - A - Comparison of Sulfathiazole Forms 3 and 4 IR Spectra, B - Comparison of Sulfathiazole Forms 3 and 4 IR Spectra with 1st Derivative applied

The patterns now show an 85% similarity making them easier to differentiate from one another.

A second problem encountered with Infrared spectroscopy is to have peaks which appear to show the material having a transmittance of over 100%, Figure 50.

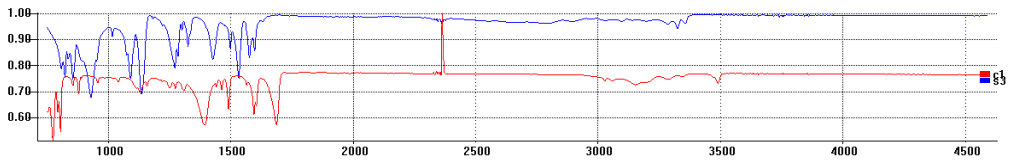

Figure 50 – Comparison of Carbamazepine Form 1 and Sulfathiazole Form 3 IR Spectra

The carbamazepine form 1 spectrum shows a large peak at approximately 2350cm^{-1}. This peak is actually representing a transmittance of 120%. This extra peak is not always present in samples, however its presence can cause issues with the clustering of spectra and so it must be removed. This peak is commonly caused due to a re-emittance of absorbed energy from other frequencies.

2.6 SAMPLE REUSABILITY

Depending on the quantity of sample possessed, some of these techniques are more useful than others. PXRD, Raman and IR using the ATR method are all non-destructive so the sample can be reused multiple times. DSC and TGA are completely destructive techniques in that the material has melted and will have formed a new polymorph or mixture of polymorphs on cooling or degraded entirely. If a further DSC or TGA pattern must be collected on the material a new sample must be used. IR using KBr is partially destructive in that the material cannot be used for anything else once pressed into a disk, however if the disk is retained an IR can be collected on it again at a later date.

2.7 REFERENCES

1. Evans, J. (2009). Powder Diffraction. 12th BCA/CCG Intensive Teaching School in X-ray Structure Analysis, Trevelyan College, University of Durham.
2. Ferraro, J. R. (1994). *Introductory Raman Spectroscopy*, Academic Press.
3. Laserna, J. J. (1996). *Modern Techniques in Raman Spectroscopy*, Wiley.
4. Smith, E. (2005). *Modern Raman Spectroscopy : A Practical Approach*, Wiley.
5. Hemminger, W. (1984). *Calorimetry: Fundamentals and Practice*, Verlag Chemie.
6. Smith, B. C. (1996). *Fundamentals of Fourier Transform Infrared Spectroscopy*, CRC Press.
7. B. P. Straughan, S. Walker (1976). *Spectroscopy Volume 2*, Chapman and Hall.
8. W. H. Bragg, W. L. Bragg (1913). "The Reflection of X-rays by Crystals." *Proceedings of the Royal Society Series A* **88**: 428-438.
9. V. K. Pecharsky, P. Y. Zavalij (2009). *Fundamentals of Powder Diffraction and Structural Characterization of Materials*. New York, Springer.(R. E. Dinnebier 2008)
10. R. E. Dinnebier, S. J. L. Billinge (2008). *Powder Diffraction: Theory and Practice*. Cambridge, Royal Society of Chemistry.
11. Cullity, B. D. (1978). *Elements of X-ray Diffraction*. London, Addison-Wesley.
12. N. Kasai, M. Kakudo (2005). *X-ray Diffraction by macromolecules*, Springer.
13. Allen, F. H. (2002). "The Cambridge Structural Database: a quarter of a million crystal structures and rising." *Acta Crystallographica Section B* **58**: 380-388.

CHAPTER 3 DATASETS USED

3.1 POLYMORPHISM

Polymorphism is a process by which a material can form several different crystal structures each of which is referred to as a polymorph. An example of two different crystal structures from a single material is shown in Figure 51.[11]

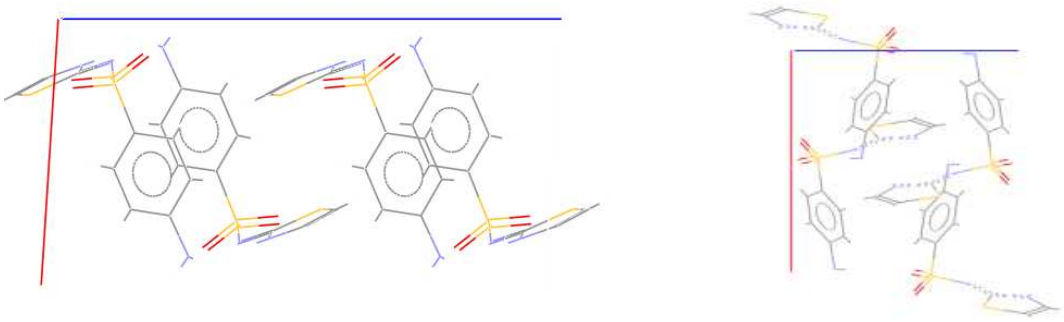

Figure 51 - Sulfathiazole Forms 2 and 4 unit cell

	Sulfathiazole Form 2		Sulfathiazole Form 4
Space group	$P2_1/c$	Space group	P 1 1 21/n
Cell lengths	**a** 8.235(4)Å **b** 8.550(4) Å **c** 15.558(8) Å	Cell lengths	**a** 10.867(3) Å **b** 11.456(3) Å **c** 8.543(2) Å
Cell angles	**α** 90 **β** 93.67(1) **γ** 90	Cell angles	**α** 90 **β** 90 **γ** 91.87(2)
Z	4	Z	4

Both of these polymorphs come from the crystallisation of sulfathiazole, which is a well studied antibiotic.

3.2 MATERIALS STUDIED

3.2.1 SULFATHIAZOLE

Sulfathiazole has 5 known polymorphs. Blagden,[4] describes the preparation of forms 1-4 while Hughes[1] describes the preparation of form 5. Due to the many different numbering techniques used across literature for these polymorphs, they are numbered based on the Cambridge Structural Database (CSD) refcodes.

- Form 1 - suthaz01
- Form 2 – suthaz
- Form 3 – suthaz02
- Form 4 – suthaz04
- Form 5 – suthaz05

The polymorphs were prepared as follows:
Form 1: Sulfathiazole (1g) was dissolved in iso-propanol. The solution was sealed in parafilm, some small holes were pierced into the parafilm and it was then left until the solvent had evaporated.
Form 2: sulfathiazole (1g) was dissolved in a hot 1:1 solution of nitro methane and ethanol. The solution was left to cool to room temperature, sealed in parafilm with some small holes were pierced into it and it was then left until the solvent had evaporated.
Form 3: sulfathiazole (1g) was dissolved in a hot solution of ammonia. The solution was left to cool to room temperature, sealed with parafilm with some small holes pierced into it and left until the solvent had evaporated.
Form 4: sulfathiazole (1g) was dissolved in boiling water. The solution was left to cool to room temperature, sealed with parafilm with some small holes pierced into it and left until the solvent had evaporated.
Form 5: sulfathiazole (1g) was dissolved in water. The solution was boiled dry and left in an oven at 110ºC to dry for 4 hours.

Forms 2, 3 and 4 were successfully prepared using these methods. The form 1 preparation resulted in crystals of form 2 while the form 5 preparation resulted in crystals of form 3. The crystals of the successfully prepared forms can be seen in Figure 52. The structure of sulfathiazole can be seen in Figure 53–A.

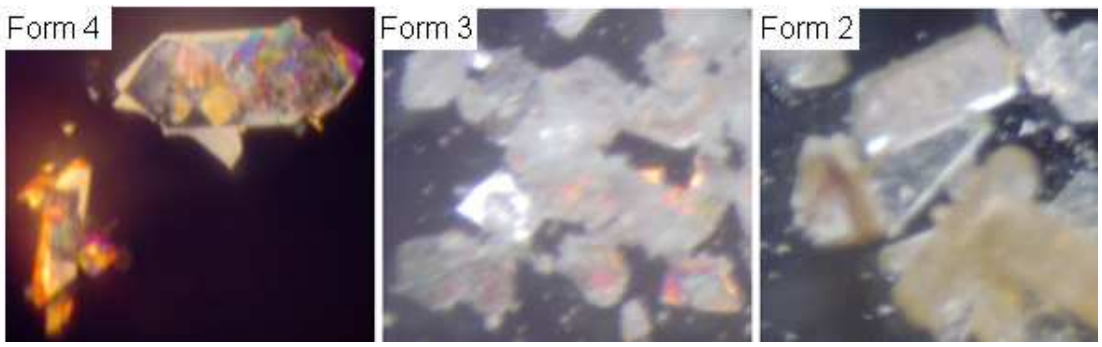

Figure 52 - Sulfathiazole crystals

3.2.2 CARBAMAZEPINE

The structure of carbamazepine can be seen in Figure 53-B

Carbamazepine has four known polymorphs. These polymorphs are prepared as described by Lang.[3] Three of these crystallisations were attempted.

Form 1: Heated to 150ºC with no solvent present and held at this temperature for 3 hours.

Form 2: Carbamazepine (1g) was dissolved in ethanol at 80ºC. The solution was left to cool to room temperature, sealed with parafilm with some small holes pierced into it and left until the solvent had evaporated.

Form 3: Carbamazepine (1g) was dissolved in ethanol. The solution was sealed with parafilm with some small holes pierced into it and left until the solvent had evaporated.

Forms 1 and 3 were successfully produced using these methods. The form 2 preparation consistently yielded form 3.

3.2.3 PIROXICAM

The structure of piroxicam can be seen in Figure 53-C

Piroxicam has three known polymorphs. These polymorphs are described by Vrečer.[2] Only the preparation of form 2 has been attempted.

Form 2: Piroxicam (1g) was dissolved in isopropyl alcohol. The solution was cooled to room temperature, sealed in parafilm, and left until the solvent had evaporated.

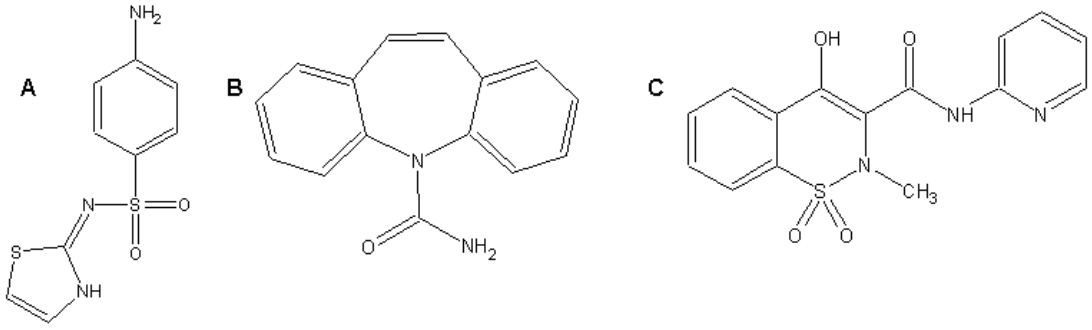

Figure 53 – Structure of A - Sulfathiazole, B - Carbamazepine, C – Piroxicam

Figure 54 shows the packing of each of the polymorphs studied and Table 3 shows their unit cell information[6-11].

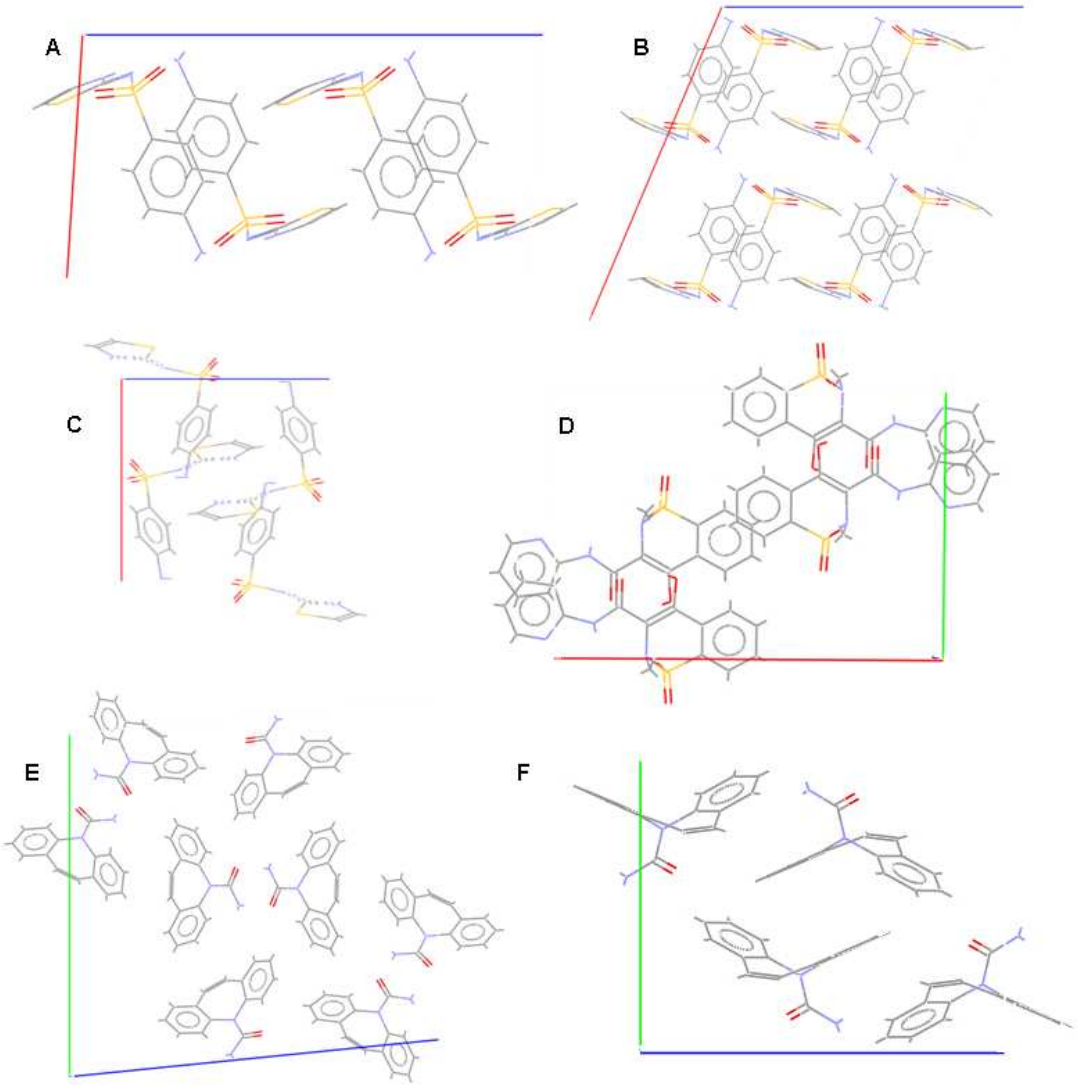

Figure 54 - Packing for A - Sulfathiazole Form 2, B - Sulfathiazole Form 3, C - Sulfathiazole Form 4, D - Piroxicam Form 2, E - Carbamazpeine Form 1, F - Carbamazepine Form 3

	Sulfathiazole Form 2		Sulfathiazole Form 3
Space group	$P2_1/c$	Space group	P 2_1/c
Cell lengths	**a** 8.235(4) Å **b** 8.550(4) Å **c** 15.558(8) Å	Cell lengths	**a** 17.570(9) Å **b** 8.574(4) Å **c** 15.583(8) Å
Cell angles	**α** 90 **β** 93.67(1) **γ** 90	Cell angles	**α** 90 **β** 112.93(1) **γ** 90
Z	4	Z	8

	Sulfathiazole Form 4		Piroxcam Form 2
Space group	P 2_1/n	Space group	P 2_1/c
Cell lengths	**a** 10.7740(10) Å **b** 8.4670(10) Å **c** 11.3670 (10) Å	Cell lengths	**a** 17.5877(4) Å **b** 11.8592(3) Å **c** 6.93840(10) Å
Cell angles	**α** 90 **β** 91.65(10) **γ** 90	Cell angles	**α** 90 **β** 97.5614(9) **γ** 91.87(2)
Z	4	Z	4

	Carbamazepine Form 1		Carbamazepine Form 3
Space group	P -1	Space group	P 2_1/n
Cell lengths	**a** 5.1705(6) Å **b** 20.574(2) Å **c** 22.245(2) Å	Cell lengths	**a** 7.537(1) Å **b** 11.156(2) Å **c** 13.912(3) Å
Cell angles	**α** 84.124(4) **β** 88.008(4) **γ** 85.157(4)	Cell angles	**α** 90 **β** 92.86(2) **γ** 91.87(2)
Z	8	Z	4

Table 3 – Crystallographic Information for Polymorphs of Sulfathiazole, Carbamazepine and Piroxicam

3.3 DATASETS

3.3.1 SULFATHIAZOLE DATASET

The three successfully prepared polymorphs of sulfathiazole were used in a sulfathiazole only dataset. Three 16-well plates were prepared by randomly choosing a material and filling each well in turn. This was repeated until all wells were filled. The first 3 wells were filled with one of each of the polymorphs to ensure that each sample was used at least once. For the remaining 45 wells, a sample was chosen at random and added to that well. In total the following samples were present in the dataset:

11 Form 4 samples
19 Form 3 samples
18 Form 2 samples

A sixteen well plate is shown in Figure 55.

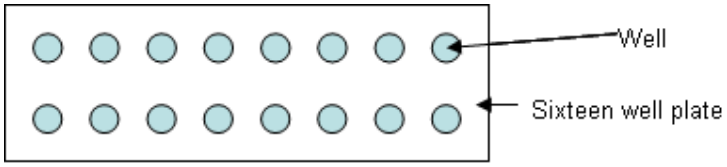

Figure 55 - Sixteen Well Plate

A sixteen-well plate is a glass slide with sixteen small hollows, or wells, in it. Each well can be filled with a different material. Well plates are not restricted to sixteen wells, but can contain any number of wells. For the dataset being studied, three such sixteen well plates were used to give forty-eight wells in total.

3.3.2 SULFATHIAZOLE/CARBAMAZEPINE DATASET

The sulfathiazole/carbamazepine dataset contains sixteen samples. The first five samples are the three polymorphs of sulfathiazole and the two polymorphs of carbamazepine. The remaining eleven samples consist of mixtures of these materials. The full dataset, including compositions of mixtures, by mass, is summarised in Table 4.

Sample Number	Sample ID	Composition	Sample Number	Sample ID	Composition
1	sulfathiazole form 4		9	carbamazepine forms 1 + 3	72:27
2	sulfathiazole form 3		10	sulfathiazole form 2 + 3 + 4	53:18:28
3	sulfathiazole form 2		11	sulfathiazole 2 + carbamazepine 1	50:50
4	carbamazepine form 1		12	sulfathiazole 3 + carbamazepine 1	50:50
5	carbamazepine form 3		13	sulfathiazole 4 + carbamazepine 1	61:38
6	sulfathiazole forms 4 + 3	42:57	14	sulfathiazole 2 + carbamazepine 3	80:19
7	sulfathiazole forms 3 + 2	37:62	15	sulfathiazole 3 + carbamazepine 3	83:16
8	sulfathiazole forms 4 + 2	68:31	16	sulfathiazole 4 + carbamazepine 3	82:17

Table 4 - Sulfathiazole/Carbamazepine Dataset Compositions

This experiment was set-up in order to provide a dataset that contains polymorphs of different materials as well as mixtures of different materials. The mixtures allow proper testing of the manual analysis mode to be carried out as well as allowing further testing of the automatic analysis mode in PolySNAP.

3.3.3 SULFATHIAZOLE/CARBAMAZEPINE/PIROXICAM DATASET

The sulfathiazole/carbamazepine/piroxicam dataset consisted of the sixteen samples from the sulfathiazole/carbamazepine dataset and the sixteen samples shown in Table 5. The pure piroxicam polymorph was included along with fifteen mixtures.

Sample Number	Sample ID	Composition	Sample Number	Sample ID	Composition
17	Piroxicam form 2		25	Carbamazepine forms 1 + 3 + sulfathiazole form 4	33:33:33
18	Piroxicam form 2 + carbamazepine form 1	12:88	26	Carbamazepine form 1 + sulfathiazole forms 2 + 3	26:32:42
19	Piroxicam form 2 + carbamazepine form 3	28:72	27	Carbamazepine form 1 + sulfathiazole forms 3 + 4	33:33:33
20	Piroxicam form 2 + sulfathiazole form 2	22:78	28	Carbamazepine form 1 + sulfathiazole forms 2 + 4	33:33:33
21	Piroxicam form 2 + sulfathiazole form 3	16:84	29	Carbamazepine form 3 + sulfathiazole forms 2 + 3	15:46:39
22	Piroxicam form 2 + sulfathiazole form 4	47:53	30	Carbamazepine form 3 + sulfathiazole forms 3 + 4	24:66:10
23	Carbamazepine forms 1 + 3 + sulfathiazole form 2	48:31:21	31	Carbamazepine form 3 + sulfathiazole forms 2 + 4	24:45:31
24	Carbamazepine forms 1 + 3 + sulfathiazole form 3	23:47:30	32	Piroxicam form 2 + sulfathiazole forms 1 + 3	12:66:22

Table 5 - Sulfathiazole/carbamazepine/piroxicam dataset compositions

3.3.4 BULK MATERIALS DATASET

The mixtures dataset includes the pure materials from six materials and various mixtures of these materials. The pure materials are summarised in Table 6.

1	Malonic acid
2	Methyl urea
3	Urea
4	salicylic acid
5	Oxalic acid dihydrate
6	Zinc nitrate hexahydrate

Table 6 - Mixtures dataset pure materials

These materials were all chosen as they were available in large quantities in the laboratory. Due to time constraints, materials which were already available in the laboratory were used for this dataset.

All possible 1:1 mixtures of these materials were prepared, however due to reactions occurring between the materials in some of these mixtures, data were only collected on certain mixtures. These mixtures are shown in Table 7.

7	Methyl urea + urea
8	Methyl urea + salicylic acid
9	Methyl urea + zinc nitrate
10	Urea + salicylic acid
11	Urea + oxalic acid
12	Salicylic acid + Oxalic acid
13	Salicylic acid + zinc nitrate
14	Oxalic acid + zinc nitrate

Table 7 - Mixtures of Materials in Mixtures Dataset

The remaining potential mixtures were abandoned due to reactions occurring after mixing.

3.4 REFERENCES

1. D. S. Hughes, M. B. Hursthouse, T. Threlfall, S. Tavener (1999). "A new polymorph of sulfathiazole." *Acta Cryst* **C55**: 1831-1833.
2. F. Vrečer, M. Vrbinc, A. Meden (2003). "Characterization of piroxicam crystal modification." *International Journal of Pharmaceutics* **256**: 3-15.
3. M. Lang, J. K. Kampf, A. J. Matzger (2002). "Form IV of Carbamazepine." *Journal of Pharmaceutical Sciences* **91**(4): 1186-1190.
4. N. Blagden, R. J. Davey, H. F. Lieberman, L. Williams, R. Payne, R. Roberts, R. Rowe (1998). "Crystal chemistry and solvent effects in polymorphic systems - sulfathiazole." *J. Chem. Soc., Faraday Trans.* **94**(8): 1035-1044.
5. G. J. Kruger, G. Gafner (1970). "The Crystal structure of Sulphathiazole II." *Acta Crystallographica B* **27**: 326-333.
6. G. J. Kruger, G. Gafner (1972). "The Crystal Structures of Polymorphs I and III of Sulphathiazole." *Acta Crystallographica B* **28**: 272-283.
7. T. Gelbrich, D. S. Hughes, M. B. Hursthouse, T. L. Threlfall (2008). "Packing similarity in Polymorphs of Sulfathiazole." *CrsytEngComm* **10**: 1328-1334.
8. J. N. Lisgarten, R. A. Palmer, J. W. Saldanha (1989). "Crystal and Molecular Structure of 5-carbamyl-5H-dibenzo [b,f]." *Journal of Chemical Crystallography* **19**(4): 641-649.
9. P. Fernandes, K. Shankland, A. J. Florence, N. Shankland, A. Johnston (2007). "Solving Molecular Crystal Structures from X-ray Powder Diffraction Data: The challenges posed by γ-carbamazepine and chlorothiazide N,N,-dimethylformamide (1/2) solvate." *Journal of Applied Crystallography* **96**(5): 1192-1202.
10. A. R. Sheth, S. Bates, F. X. Muller, D. J. W. Grant (2005). "Local Structure in Amorphous Phase of Piroxicam from Powder X-ray Diffractometry." *Crystal Growth and Design* **5**(2): 571-578.
11. Allen, F. H. (2002). "The Cambridge Structural Database: a quarter of a million crystal structures and rising." *Acta Crystallographica Section B* **58**: 380-388.

CHAPTER 4 THE 48 SAMPLE SULFATHIAZOLE DATASET

4.1 THE DATASET

The 48 sample sulfathiazole dataset contains polymorphs two, three and four of sulfathiazole. This dataset has appeared in Chapter 1 as an example to show the functionality of both the validation techniques and the INDSCAL method. Table 8 summarises the composition of this dataset.

Sample Number	Sample ID	Name in PolySNAP		Sample Number	Sample ID	Name in PolySNAP
1	form 4	01-4		25	form 3	25-3
2	form 2	02-2		26	form 3	26-3
3	form 3	03-3		27	form 2	27-2
4	form 2	04-2		28	form 2	28-2
5	form 3	05-3		29	form 4	29-4
6	form 4	06-4		30	form 3	30-3
7	form 3	07-3		31	form 3	31-3
8	form 3	08-3		32	form 2	32-2
9	form 2	09-2		33	form 2	33-2
10	form 4	10-4		34	form 3	34-3
11	form 2	11-2		35	form 2	35-2
12	form 4	12-4		36	form 4	36-4
13	form 4	13-4		37	form 3	37-3
14	form 3	14-3		38	form 2	38-2
15	form 2	15-2		39	form 3	39-3
16	form 2	16-2		40	form 3	40-3
17	form 3	17-3		41	form 2	41-2
18	form 3	18-3		42	form 2	42-2
19	form 4	19-4		43	form 3	43-3
20	form 4	20-4		44	form 3	44-3
21	form 2	21-2		45	form 2	45-2
22	form 2	22-2		46	form 3	46-3
23	form 4	23-4		47	form 2	47-2
24	form 4	24-4		48	form 3	48-3

Table 8 - 48 sample sulfathiazole dataset composition

The dataset contains eighteen form 2 samples, nineteen form 3 samples and eleven form 4 samples in total.

For this dataset, PXRD data was collected on a Bruker C2 GADDS. Each sample was run for two minutes over a 5-35° range in 2θ. Raman data was collected on a Bruker SENTINAL with a 532nm laser, integrated into the Bruker C2, over a range of 250cm^{-1} to 2500cm$^{-1.}$

4.2 DATASET CLUSTERING

It is expected that the eighteen form 2 samples will form a single cluster, the nineteen form 3 samples will form a second cluster and the eleven form 4 samples will form a third cluster.

For each dendrogram in this and all subsequent datasets a score will be given that is determined by dividing the number of samples that are incorrectly clustered by the total number of samples. This score will be on the scale of 0-1 with the smaller the number, the better the dendrogram matches the predicted clustering.

4.2.1 PXRD DATA

The dendrogram and MMDS plot for the PXRD data are shown in Figure 56.

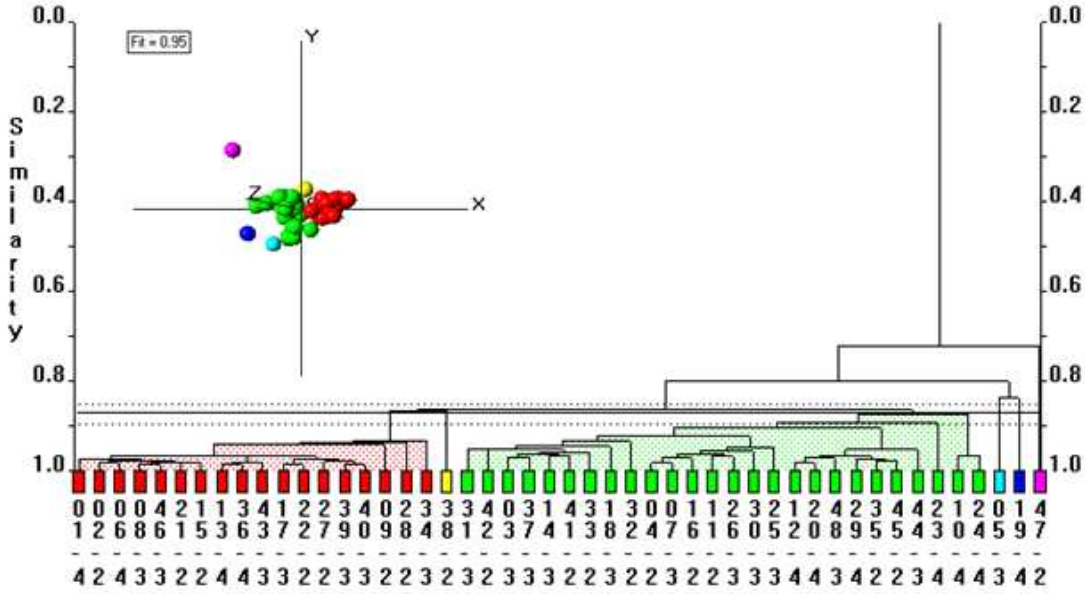

Figure 56 – PXRD dendrogram and MMDS Plot

The dataset does not give the expected clustering. All three materials are mixed together and no contiguous groups of samples are visible. This dendrogram has a score of 0.83.

The MMDS plot does not show clear separation of the clusters either. Many of the patterns have problems with a high background (Figure 57–A) or are suffering from preferred orientation (Figure 57–B).

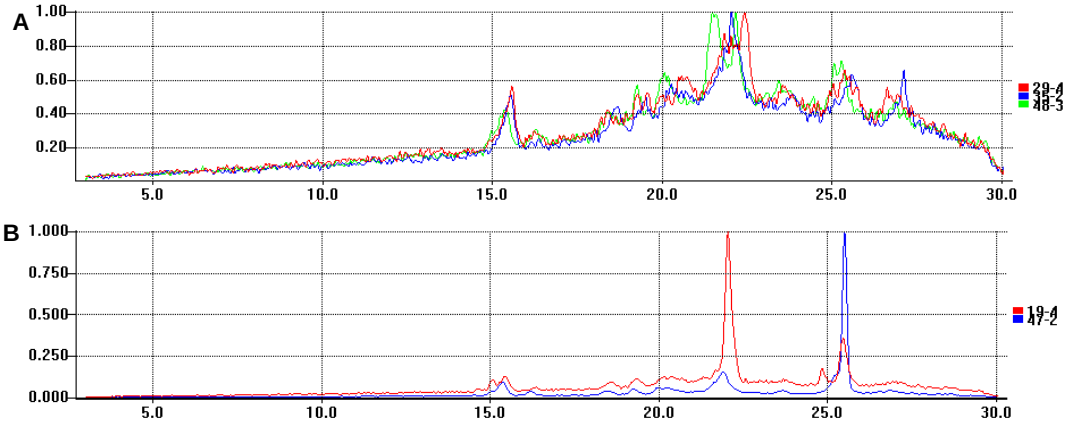

Figure 57 -A - overlay of samples showing poor background. B - Overlay of samples showing preferred orientation

Re-running the dataset with background subtraction applied gives a small improvement to the high background problem, however this is not a large enough improvement to make a major difference to the clustering. The dendrogram and MMDS plot for the removed background run are shown in Figure 58.

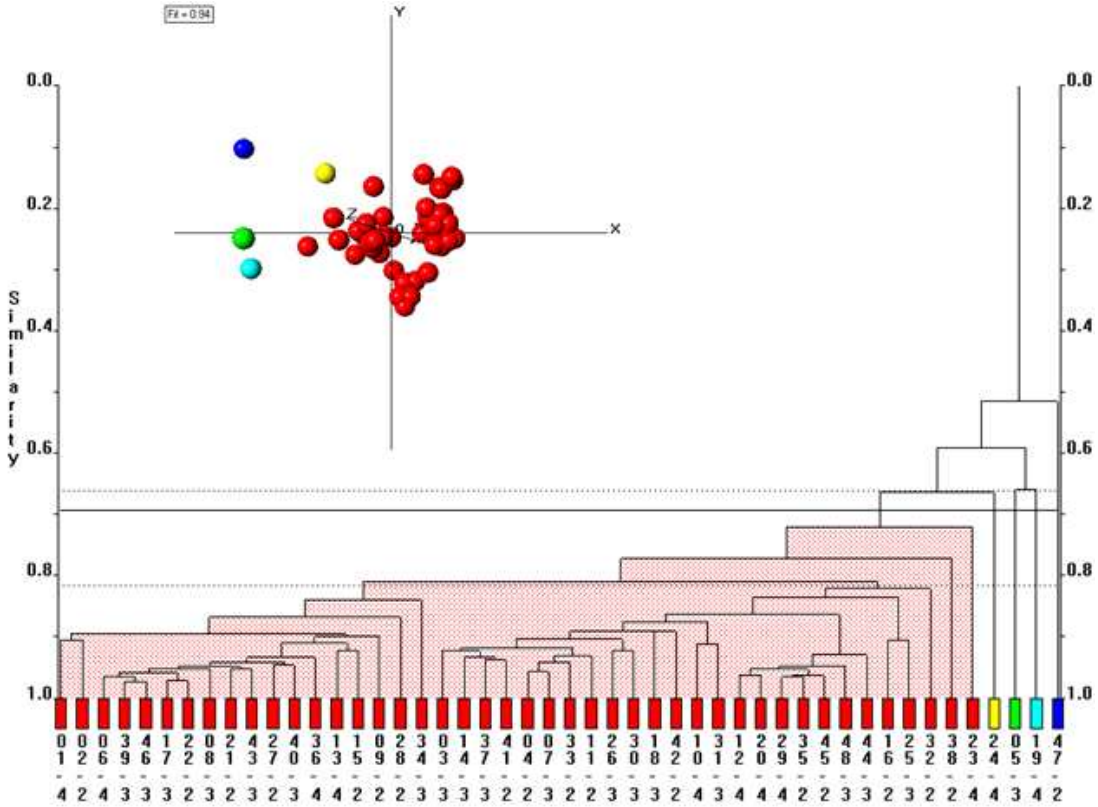

Figure 58 - Dendrogram for PXRD with Background Removed

When the data is examined, it can be seen that all the peaks in the PXRD patterns appear above 14º as shown in Figure 59. This dendrogram has a score of 0.81, giving a very small improvement over the non pre-processed dendrogram.

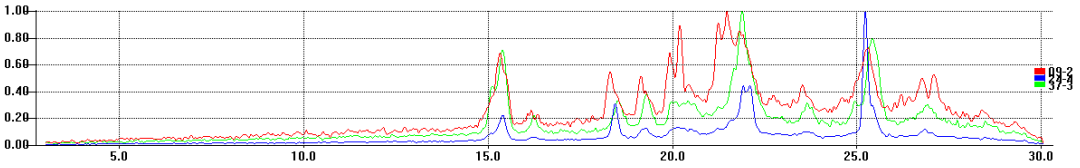

Figure 59 - PXRD Peak Positions

By only matching data between 14° and 30°, improved clustering may be produced. The area below this will have correlations nearing 100% which will skew the correlations for the remainder of the patterns. The resulting dendrogram and MMDS plot are shown in Figure 60.

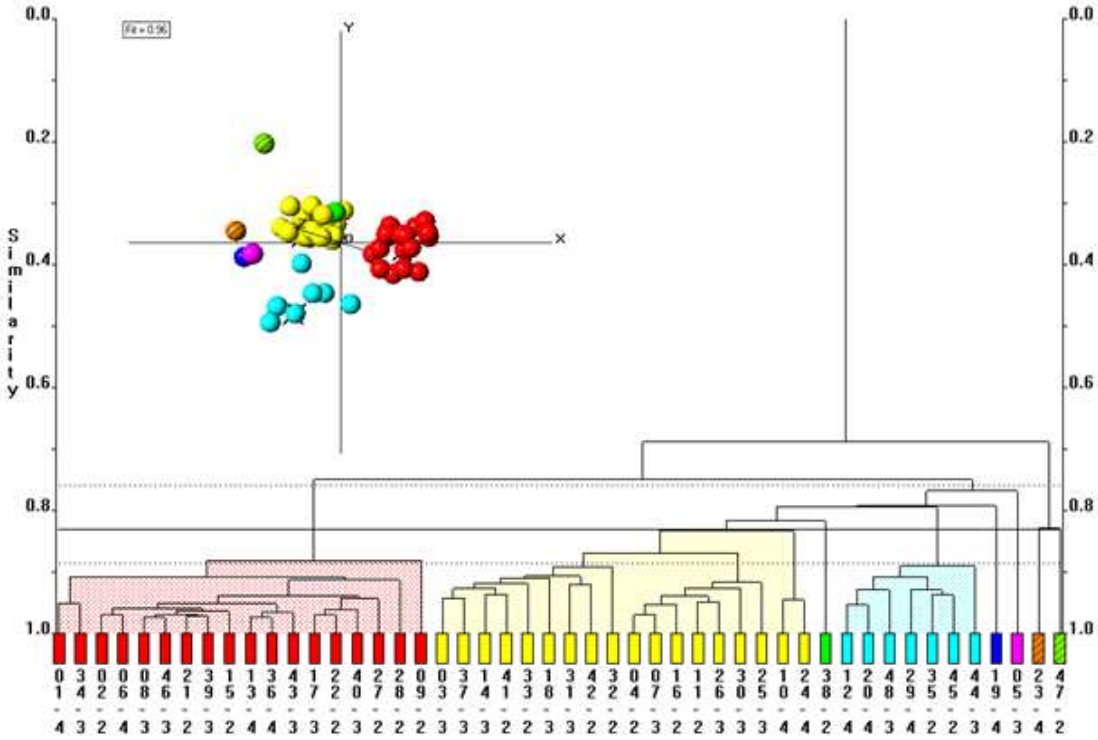

Figure 60 – 14° to 30° Data PXRD Dendrogram and MMDS Plot

The data still does not show any large contiguous groups of similar samples.
The dendrogram has a score of 0.83, identical to the score of the original PXRD run.

4.2.2 RAMAN DATA

The data was initially run with no pre-processing applied. As described in Chapter 2, the data has a high similarity between samples (Figure 61). This results in datasets where all the spectra are linked by low lying tie bars, making it difficult to distinguish between individual samples.

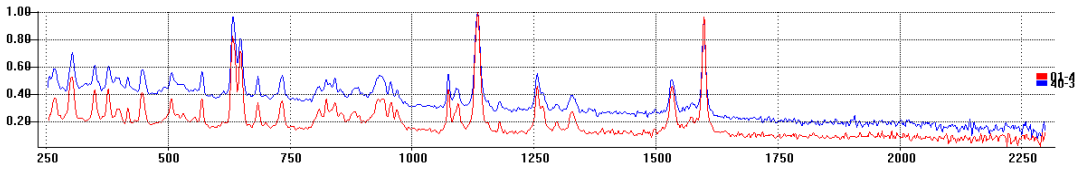

Figure 61 - Overlay of Sulfathiazole Forms 3 and 4

Figure 62 shows the dendrogram produced from the unprocessed Raman data.

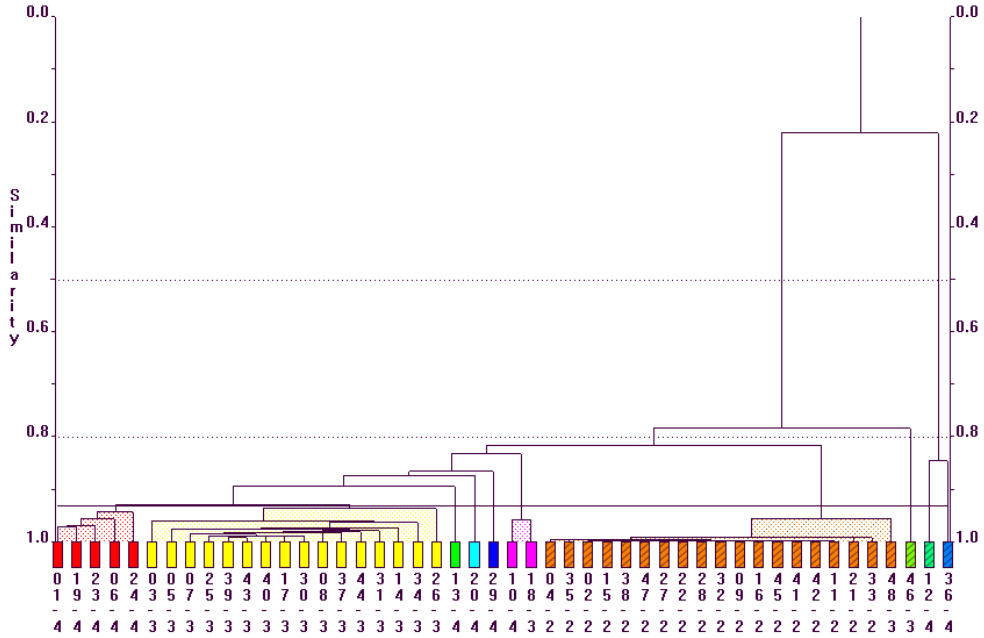

Figure 62 - Dendrogram for Unprocessed Raman Data

The dendrogram is fairly well clustered with the form 2 samples all clustered together in the striped brown cluster and the yellow cluster containing all but three of the form 3 samples. The form 4 samples are split into three groups, with two of the groups separated by the form 3 cluster and the other one separated by the form 2 sample.

This dataset is significantly better clustered than the PXRD ones with a score of 0.19.

By applying a first derivative to the dataset, the differences between individual patterns can be enhanced; however this comes at the cost of clear clustering, as shown in Figure 63.

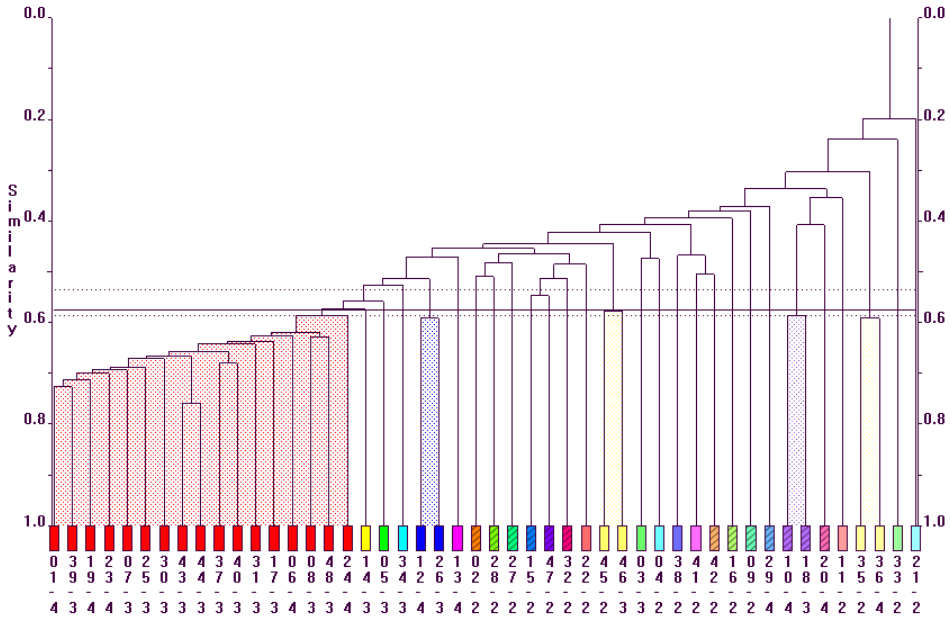

Figure 63 – Dendrogram for First Derivative Raman Data

This dendrogram has a score of 0.45, clearly much poorer than that of the original Raman dendrogram.

By combining the distance matrixes from both the standard and first derivative data using the INDSCAL method (Figure 64), the best features of both datasets can be retained.

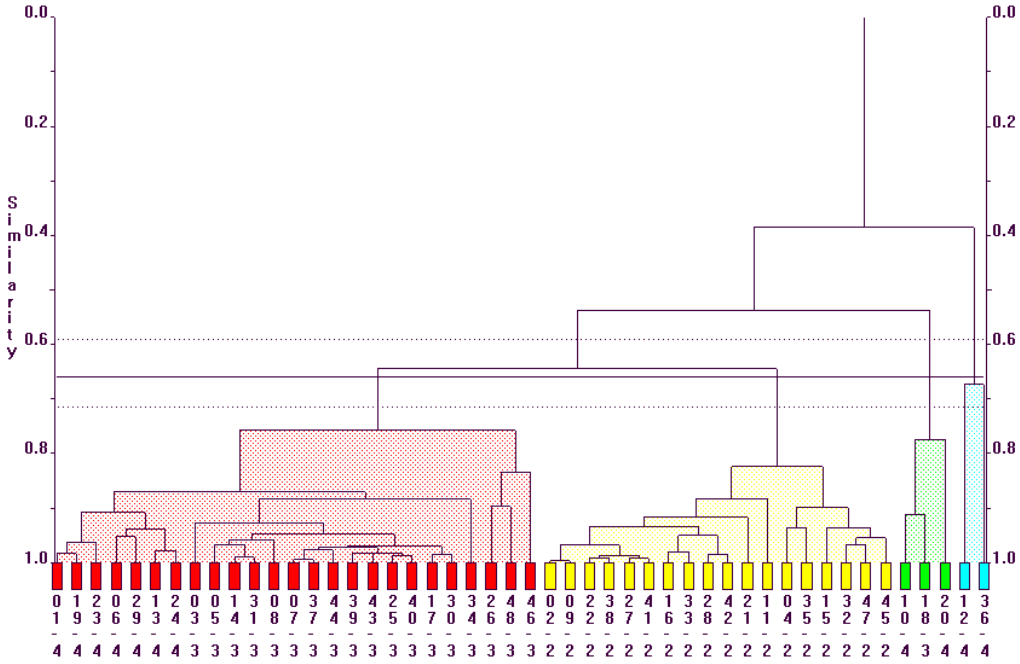

Figure 64 – Combination Raman data dendrogram

This combined dendrogram has clearly distinguished samples and good clustering. The only outliers belong to samples with poor quality spectra. The dendrogram has a score of 0.1, giving it a small improvement over the original Raman dataset and a substantial improvement over the first derivative dendrogram. Figure 65 shows a comparison of a poorer quality Raman spectra (blue) and a better quality Raman spectra (red) for sulfathiazole form 4.

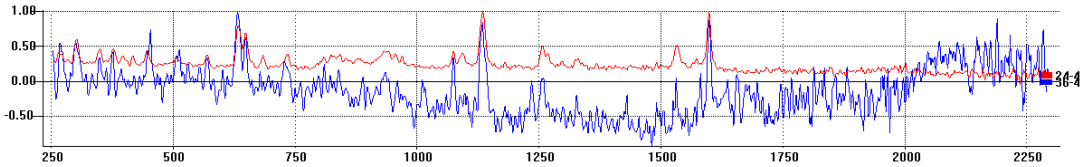

Figure 65 - Poorer Quality and Better Quality form 4 Raman Spectra

The background wave is not as smooth in the 'poorer' quality dataset, making it harder to see the smaller peaks. Larger peaks can still be seen easily however the smaller peaks are swamped by the background causing the patterns to not be matched with one another.

Figure 66 shows a comparison of a poorer quality Raman spectra (blue) and a better quality Raman spectra (red) for sulfathiazole form 3. These samples are not re-runs, but are actually samples present in different wells, the poorer one being in well 36 and the better one in well 24 in the well plate.

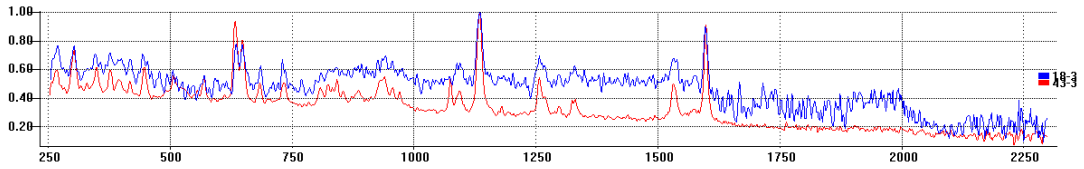

Figure 66 - Poorer Quality and Better Quality form 3 Raman Spectra

Once again the noisy background is rendering the smaller peaks difficult to see while the larger peaks are still clearly visible.

The cut-level can be adjusted, by hand, to more easily show the form 3 and form 4 clusters. The dendrogram and MMDS plot for this are shown in Figure 67.

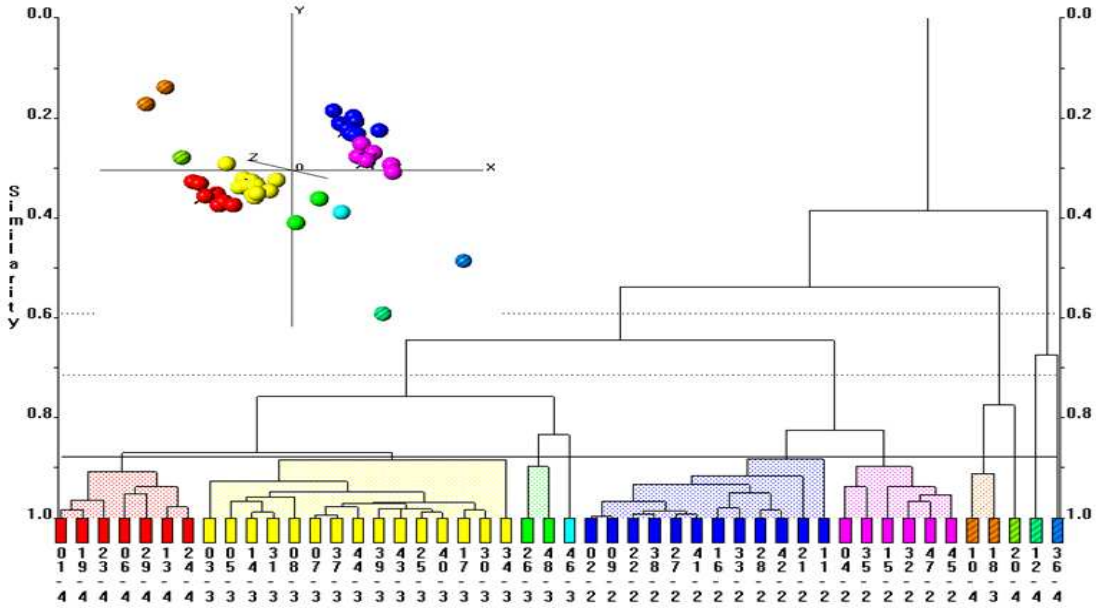

Figure 67 - Combined Dendrogram with Cut-level Adjusted

Adjusting the cut-level in the combined dendrogram splits the form 2 samples into two separate clusters but separates the form 4 and form 3 samples into two separate clusters. The MMDS plot shows that although separated the purple and blue clusters, which represent the previously combined form 2 cluster, lie close together. The form 3 samples, in the yellow cluster and the form 4 samples, in the red cluster, also lie close together as is expected due to the cut-level only having being lowered a small amount. The splitting of

the form 2 samples affects the score of this dendrogram when compared to the same dataset without a cut-level adjustment. IT now has a score of 0.29

It is also possible to run the dataset with second derivative pre-processing. The dendrogram for this is shown in Figure 68.

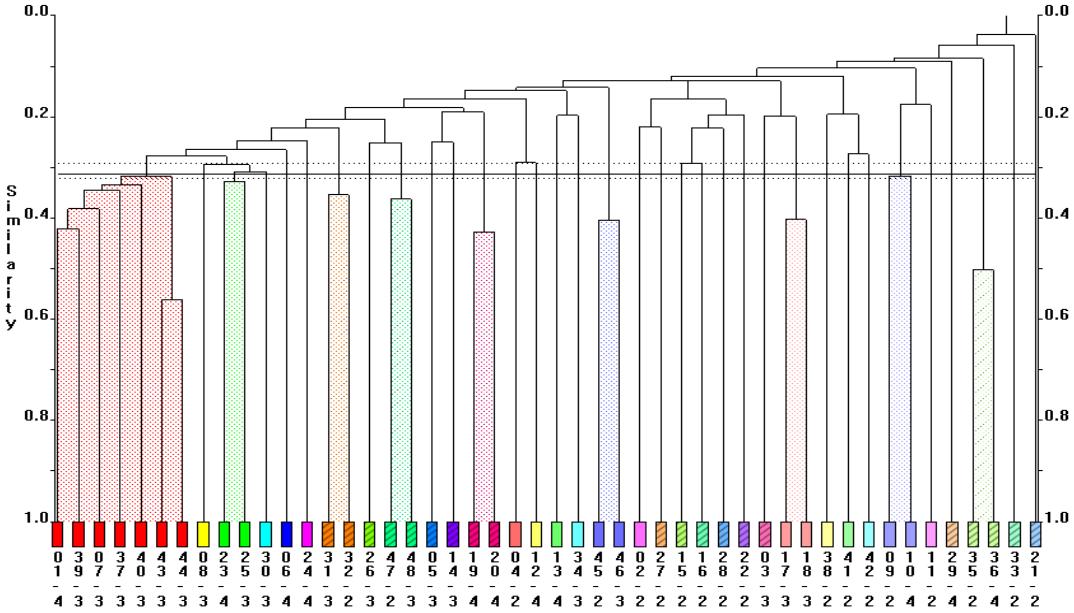

Figure 68 - Dendrogram for Second Derivative Raman Data

This gives an even more dramatic separation of the previously closely tied samples over the first derivative method; however this comes at the cost of even poorer clustering. The score has now risen to 0.69, clearly much poorer than the first derivative dendrogram.

A combination, using INDSCAL, of the second derivative and original dataset is shown in Figure 69.

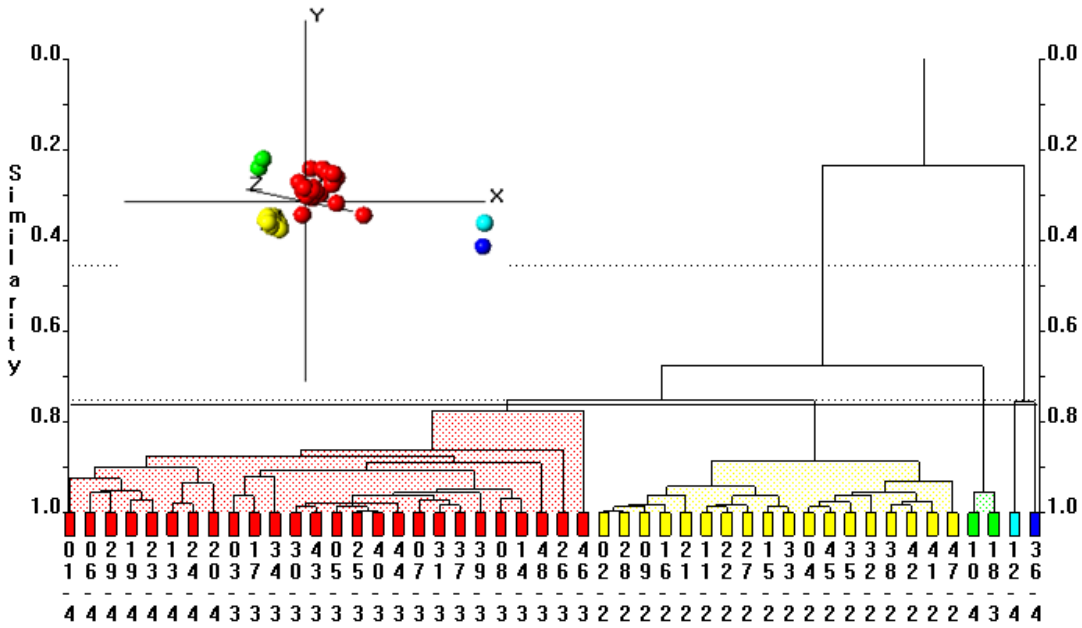

Figure 69 - Combined Second Derivative and Original Dataset Dendrogram and MMDS Plot

This dataset shows a small improvement over the combined first derivative and original dendrogram. One of the outliers is now clustered as expected, reducing the number of outliers from five to four. This reduction in the outliers has caused the score to fall slightly from the combined first derivative and original Raman dataset run to 0.08.

Despite the slightly improved clustering over the first derivative combined dataset, the MMDS plot does not show as clear a separation between the different clusters as the first derivative runs MMDS plot shows.

The resulting dendrogram from a combination of first derivative, second derivative and the original Raman data is shown in Figure 70.

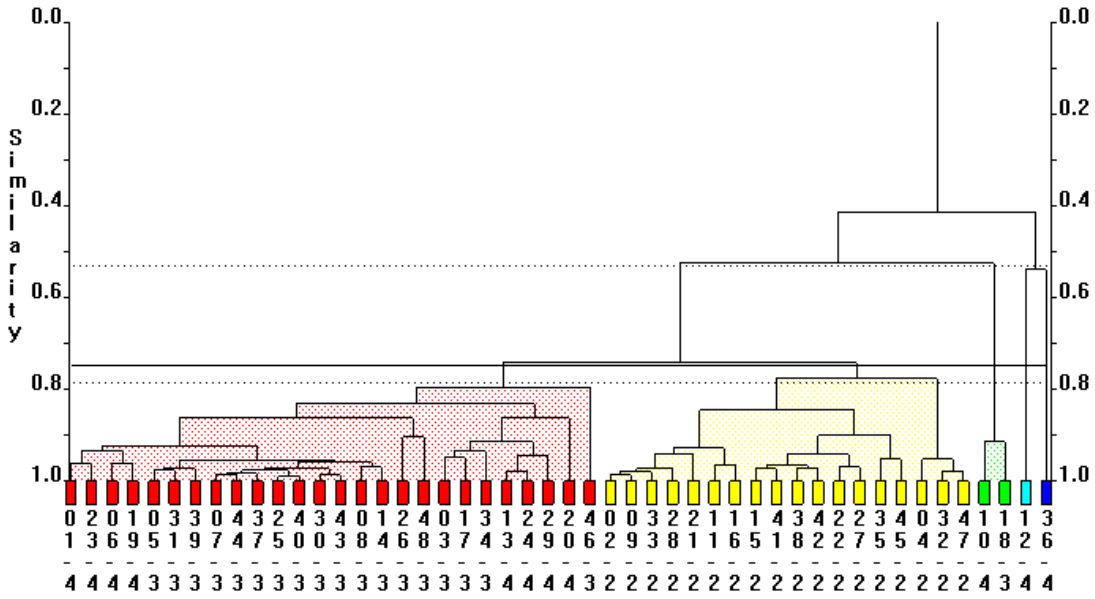

Figure 70 - Combined First Derivative, Second Derivative and Original Data Raman Dendrogram

This dendrogram shows poorer clustering than the previous combined dendrograms. Although only four outliers are present the form 4 samples that were previously well clustered are now split into two equal sized groups by the form 3 samples. The score for this dendrogram is 0.19.

4.2.3 TRIMMING RAMAN DATA

The dataset was examined for areas of high similarity as described in section 2.2.3. When the data is examined closely, it is revealed that all of the patterns in the dataset have high similarity in the area beyond 1750cm^{-1} (Figure 71).

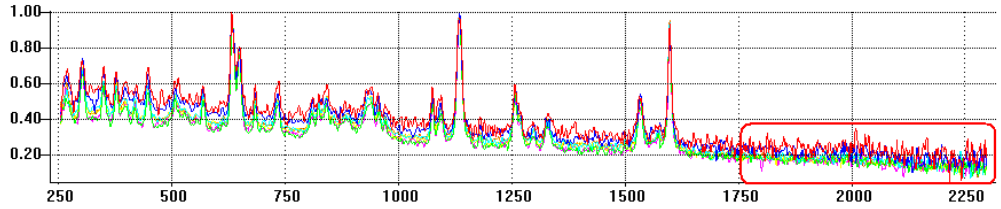

Figure 71 - Overlay of some Raman spectra showing long 'tail' of similar data from 1750cm^{-1}

These areas, when matching occurs, have correlation coefficients approaching 1.0. This will raise the overall correlation coefficient for the patterns towards 1.0.

By ignoring this region and only carrying out pattern matching on the area from 250cm^{-1} to 1750cm^{-1}, a dendrogram with improved clustering over the original unprocessed run is produced. Figure 72 shows the dendrogram and MMDS plot resulting from this run.

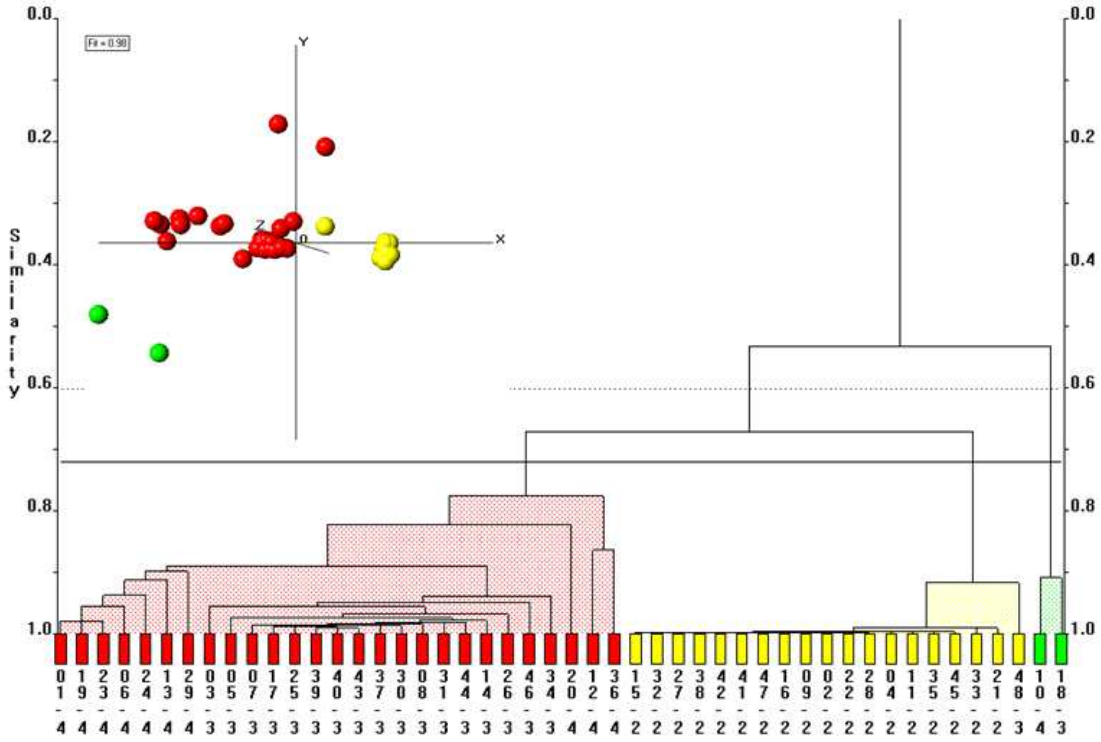

Figure 72 - Cut-off Raman data dendrogram and MMDS Plot

The dendrogram shows clear separation of the form 2 samples with all such samples grouped together in the yellow cluster. A single form 3 sample is also present in this cluster. The form 3 samples are, with two exceptions (48-3 in the yellow cluster and 18-3 in the green cluster), clustered together while the form 4 samples are split into two groups, one larger and one smaller, by the form 3 samples with a single outlier. Two of the form 4 samples in the smaller group are joined to the remainder of the red cluster with a much higher tie-bar. The dendrogram now has a score of 0.15, a small improvement over the non cut-off datasets score of 0.19.

The MMDS plot is much more clearly defined than in the PXRD example. With the exception of a single outlier close to the red cluster, the yellow cluster is clearly defined. This sample is the lone form 3 sample among the form 2 samples in the yellow cluster. The red cluster is split into two larger groups with two outliers. The group closest to the origin contains the form 3 samples while the other group contains the form 4 samples. The two outliers from the red cluster are the form 4 samples with the higher tie-bar.

The PXRD pattern and Raman spectra for the two outliers present in the green cluster are shown in Figure 73.

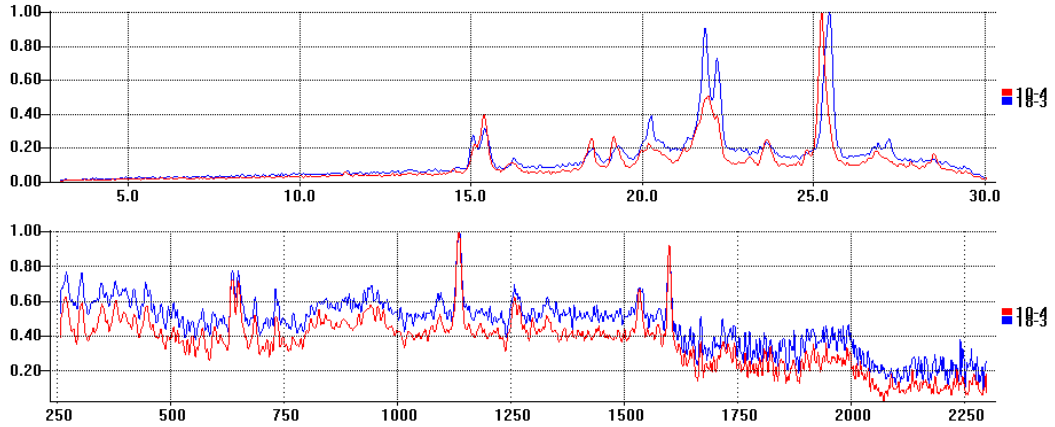

Figure 73 – Outliers PXRD Pattern and Raman Spectra

Figure 74 shows an overlay of the most representative samples of the red (sample 22-2) and yellow (sample 07-3) clusters.

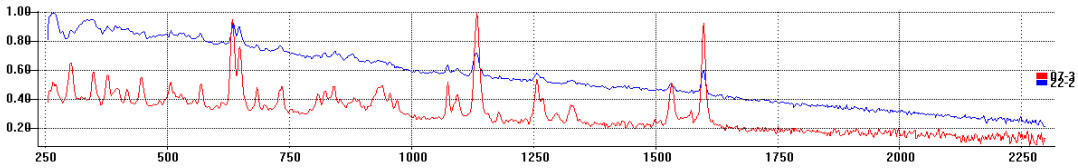

Figure 74 - Most Representative Samples

4.2.4 CLUSTER VALIDATION

The silhouettes for this dataset are shown in Figure 75 and the fuzzy clustering in Figure 76.

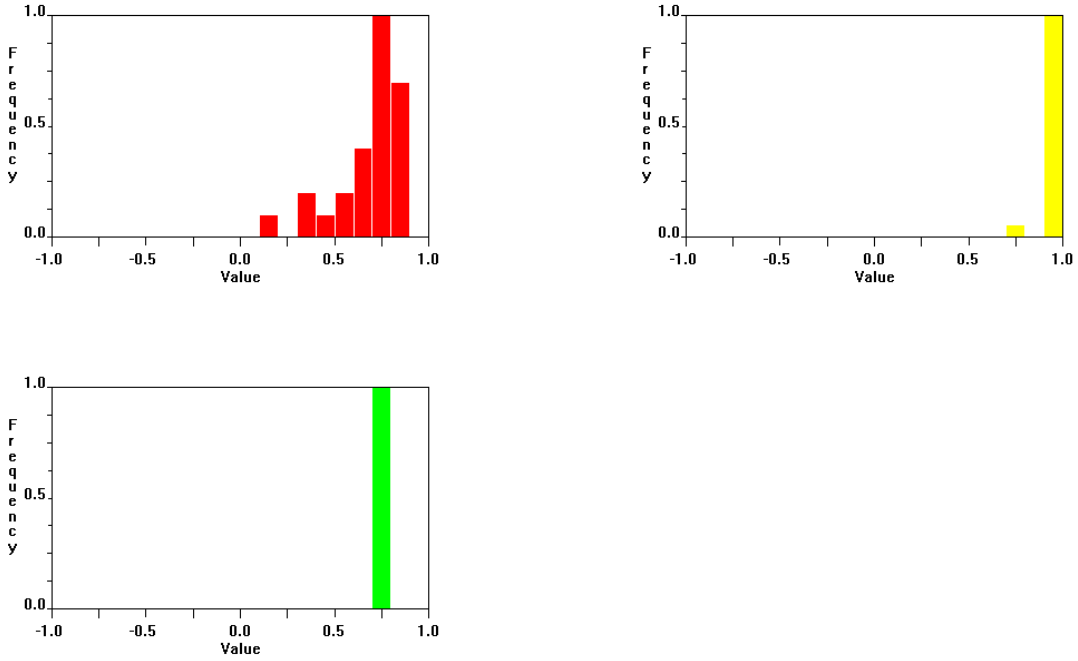

Figure 75 - Silhouettes

The red cluster silhouette has the first of the two form 4 outliers (36-4) in a distinct region just below 0.25. The bar above 0.25 contains the second one of these form 4 outliers (12-4) and another form 4 outlier (20-4) which is present in the more diffuse of the two red clusters groups. The remaining silhouettes, with the exception of the two upmost silhouettes, all refer to the samples in the more diffuse cluster.

The lone silhouette at 0.75 in the yellow cluster represents sample 48-3, which is the outlier from this cluster in the MMDS plot.

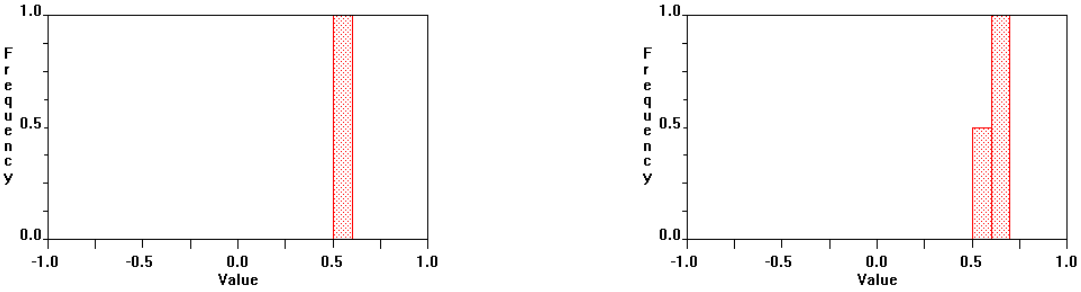

Figure 76 - Fuzzy Clustering

The numerical values for the fuzzy clustering are shown in Table 9. For the cluster assignments, cluster 1 is considered to be the green cluster, cluster 2 the red cluster and cluster 3 the yellow cluster.

	1	2	3				1	2	3	
01-4	0.24	1	0.16			25-3	0.23	0.93	0.45	
02-2	0.18	0.4	0.97			26-3	0.2	0.9	0.49	
03-3	0.21	0.86	0.54*	<==		27-2	0.18	0.42	0.97	
04-2	0.17	0.42	0.97			28-2	0.18	0.43	0.96	
05-3	0.22	0.89	0.52*	<==		29-4	0.24	0.95	0.13	
06-4	0.25	0.99	0.11			30-3	0.23	0.92	0.48	
07-3	0.24	0.95	0.41			31-3	0.23	0.95	0.41	
08-3	0.23	0.93	0.46			32-2	0.17	0.42	0.97	
09-2	0.18	0.42	0.96			33-2	0.19	0.43	0.95	
10-4	0.86	0.62*	0	<==		34-3	0.26	0.94	0.33	
11-2	0.18	0.41	0.96			35-2	0.17	0.43	0.96	
12-4	0.07	0.83	0.35			36-4	0.07	0.7	0.53*	<==
13-4	0.25	0.96	0.07			37-2	0.23	0.94	0.43	
14-3	0.23	0.94	0.43			38-2	0.18	0.41	0.97	
15-2	0.18	0.42	0.97			39-3	0.24	0.94	0.43	
16-2	0.18	0.42	0.96			40-3	0.23	0.93	0.46	
17-3	0.24	0.93	0.46			41-2	0.18	0.42	0.96	
18-3	0.88	0.59*	0.09	<==		42-2	0.18	0.41	0.97	
19-4	0.24	1	0.23			43-3	0.24	0.92	0.48	
20-4	0.26	0.92	0.07			44-3	0.25	0.95	0.39	
21-2	0.17	0.42	0.96			45-2	0.18	0.43	0.96	
22-2	0.18	0.42	0.96			46-3	0.17	0.86	0.54*	<==
23-4	0.23	0.99	0.25			47-2	0.18	0.42	0.97	
24-4	0.26	0.98	0.07			48-3	0.21	0.67*	0.78	<==

Table 9 – Fuzzy Clustering Numerical Values

The first fuzzy clustering group includes samples 03-3, 05-3, 36-4 and 46-3, all of which are present in the red cluster. 36-4 is one of the two outliers from the red groups while the remaining three samples are present in the more tightly grouped of the two groups.
Based on the numerical results:

- Sample 03-3 could potentially be present in either the red or yellow clusters.
- Sample 05-3 could potentially be present in either the red or yellow clusters.
- Sample 36-4 could potentially be present in either the red or yellow clusters.
- Sample 46-3 could potentially be present in either the red or yellow clusters.

The second of the fuzzy clustering groups shows two bars. The smaller of these contains sample 18-3, in the green cluster, while the larger one contains sample 48-3, the yellow cluster outlier, and sample 10-4, in the green cluster.
Based on the numerical results:

- Sample 10-4 could potentially be present in either the green or red clusters.

- Sample 18-3 could potentially be present in either the green or red clusters
- Sample 48-3 could potentially be present in either the red or yellow clusters.

The minimum spanning trees, with the initial number of links, one link removed and one added, are shown in Figure 77.

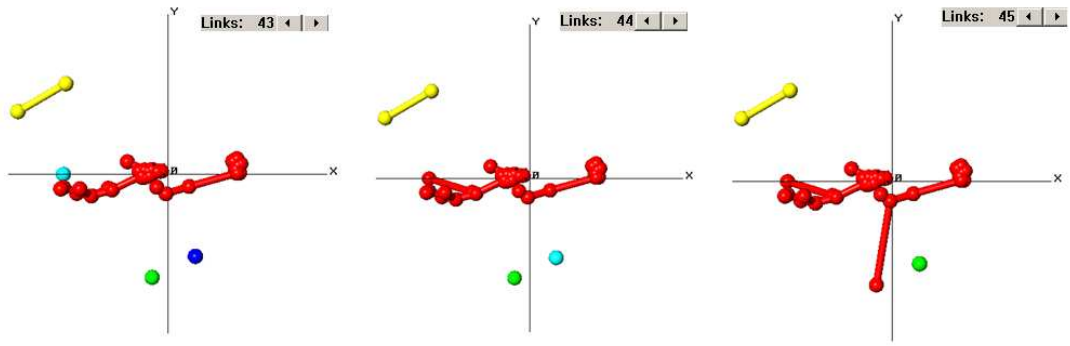

Figure 77 - Minimum Spanning Trees

Adding a link adds 12-4 to the red cluster from the green cluster while removing a link takes 20-4 out of the red cluster. Sample 20-4 is one of the outliers previously noticed in the silhouettes and fuzzy clustering.

The outlier from the yellow cluster, sample 48-3, is present in the fuzzy clustering' again suggesting it could be may be more similar than the dendrogram is showing. The appearance of this sample in a separate, lower, band in the silhouettes for the yellow cluster also implies that cluster membership could be changed easily. Overall the validation techniques support the expected clustering while also showing the samples that appear outside of the expected clusters as being potentially ambiguous.

4.3 OPTIMAL RAMAN PRE-PROCESSING

The effects of different combinations of pre-processing options on the data were investigated to determine what gave the optimal clustering. The results of this study are shown in Table 10.

Pre-processing Applied	Sulfathiazole Full Profile Score	Sulfathiazole Derivative Score	Sulfathiazole Full Profile Cut-off Score	Sulfathiazole Derivative Cut-off Score
No Pre-processing	0.19	0.48	0.15	0.91
Remove Cosmic Ray Spikes	0.19	0.5	0.19	0.54
Remove Cosmic Ray Spikes and Denoise	0.13	0.31	0.10	0.42
Remove Cosmic Ray Spikes, Denoise and Remove Background	0.44	0.4	0.17	0.21
Denoise	0.13	0.4	0.1	0.25
Remove Background	0.42	0.5	0.23	0.46
Denoise and Remove Background	0.35	0.35	0.19	0.42

Table 10 – Summary of Pre-Processing Options Applied to Raman Data

Removing cosmic ray spikes appears to make no difference on the original data with no processing applied or with peak smoothing being applied. In the original run, it is only with background removal applied that any difference begins to appear, in that case it degrades the clustering. In the derivative it also has a negative effect on the clustering. For the cut-off dataset it produces improved clustering when combined with background removal and has not effect when combined with the peak smoothing.

Two methods in both the full pattern match and the match of the data with a cut-off give optimal clustering. The preferred method is the one in which the smallest amount of pre-processing needs to be applied. Denoising the full profile pattern (dendrogram and MMDS plot in Figure 78 and denoising the cut-off pattern (dendrogram and MMDS plot in Figure 80) are chosen as the optimal method for this reason.

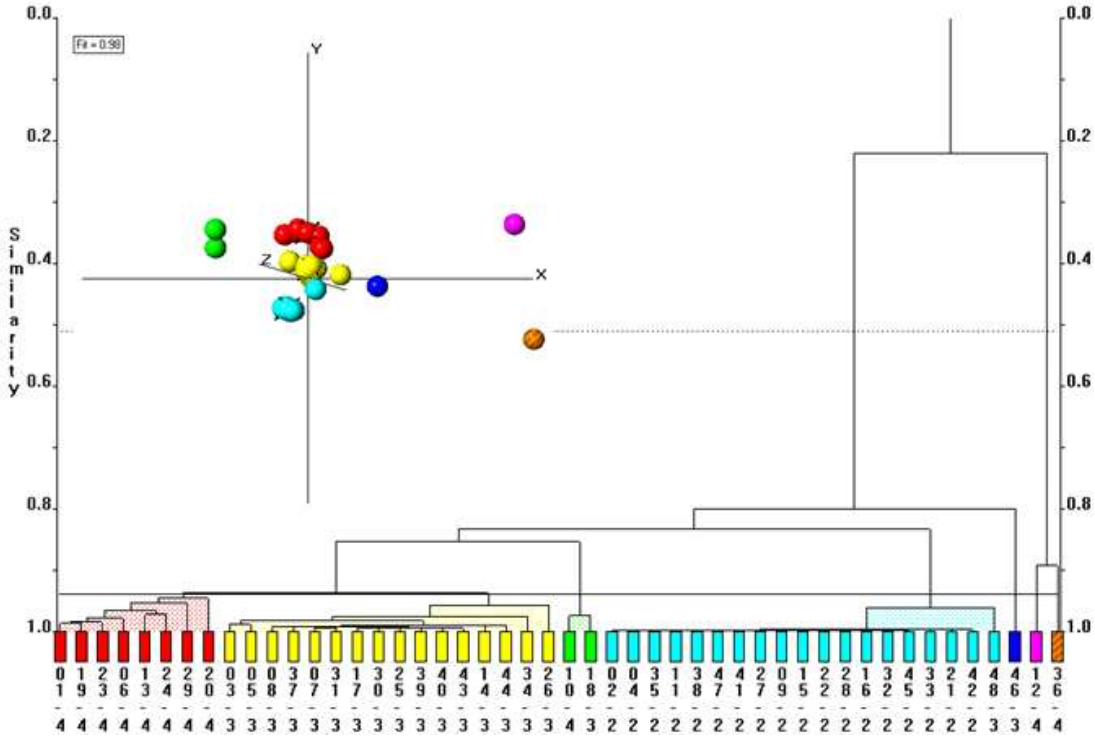

Figure 78 - Dendrogram and MMDS plot of Denoised Dataset

The aquamarine cluster contains all the form 2 samples and one form 3 sample. The yellow contains all but three of the form 3 samples, and the red cluster contains all but three of the form 4 samples.

The MMDS plot shows the aquamarine cluster, containing the form 2 samples, is tightly grouped with a single outlier. The outlier in this group is the single form 3 sample in this cluster. The yellow cluster, containing the form 3 samples, is tightly grouped. There is a single outlier from this group which has a slightly higher tie-bar height than the remainder of this cluster. The red cluster, containing the form 4 samples, is tightly grouped with no outliers. The remaining samples all contain 'poorer' quality data with noisier backgrounds. The spectra for these samples are shown in Figure 79.

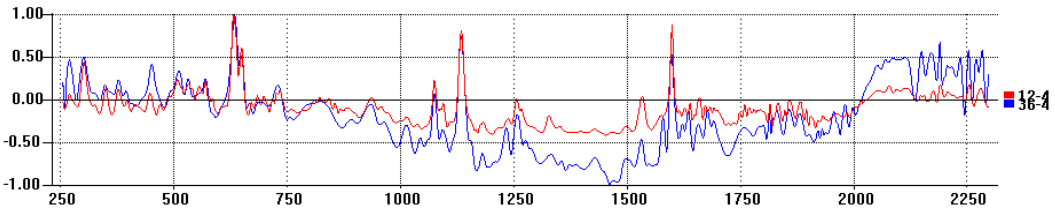

Figure 79 - Spectra for Outliers

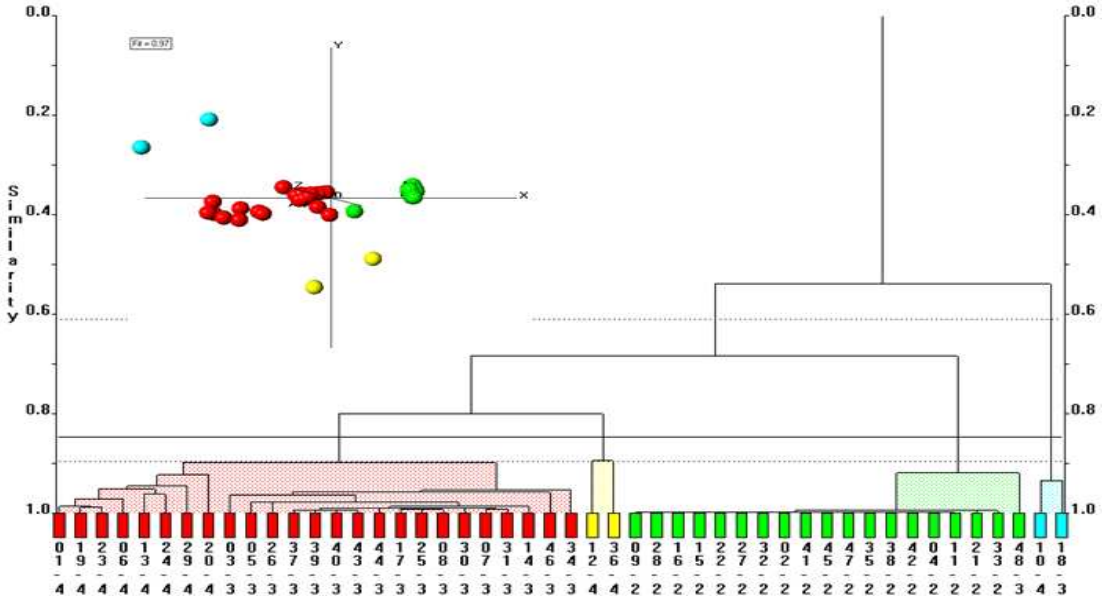

Figure 80 - Dendrogram and MMDS Plot of Denoised Dataset with Cut-off

This dendrogram has the form 2 samples once again all clustered together, this time in the green cluster, with a single form 3 sample again present in this cluster. The form 3 samples are clustered together with two outliers and the form 4 samples are clustered together with three outliers.

Lowering the cut-level further splits the form 3 and form 4 samples as shown in Figure 81.

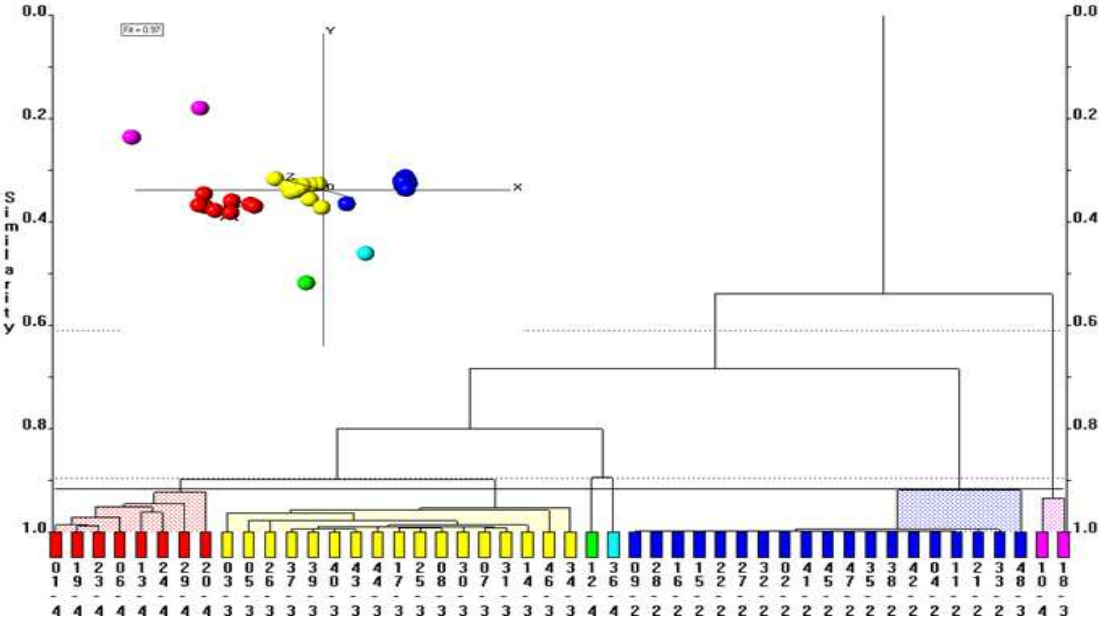

Figure 81 - Dendrogram and MMDS Plot of Denoised Dataset with Cut-off and Adjusted Cut-level

The form 2 samples are now all clustered together in the blue cluster, with a single form 3 sample again present in this cluster. The form 3 samples are clustered together in the

yellow cluster with the exception of two outliers and the form 4 samples are clustered together in the red cluster with the exception of three outliers.

The clustering in the MMDS plot is much more clearly defined than in the dataset which is not cut-off. The blue cluster is tightly clustered with a single outlier that is the lone form 3 sample in this cluster. Both the red and yellow clusters also show tight clustering. The remaining outliers all have problems with 'poorer' quality data.

Overall denoising the pattern coupled with matching data below 1700cm^{-1} is the optimal method of data pre-processing and so is preferred for all future clustering.

The overlays of the poorly clustered samples are shown in Figure 82 (form 4 samples) and Figure 83 (form 3 samples).

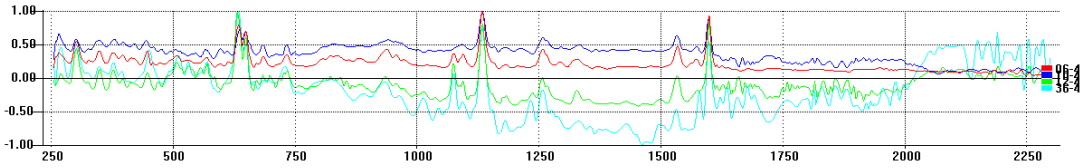

Figure 82 - Poorly Clustered Form 4 Samples

The poorly clustered form 4 samples, 10-4 in the purple cluster and 12-4 and 36-4 in the green and aquamarine clusters, when overlaid with the most representative sample from the form 4 cluster (06-4) shows that there are clear differences in the background between these patterns. The remaining form 4 samples do not suffer from this higher background problem.

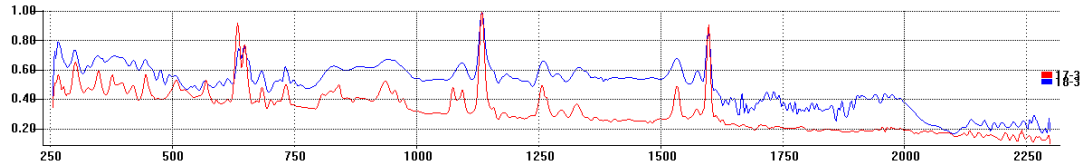

Figure 83 - Poorly Clustered Form 3 Samples

The poorly clustered form 3 sample (18-3), when overlaid with the most representative sample from the form 4 cluster shows that there are clear differences in the background between these patterns. The remaining form 3 samples do not suffer from this higher background problem.

The silhouettes are shown in Figure 84 and the fuzzy clustering in Figure 85. The numerical data for the fuzzy clustering is in Table 11

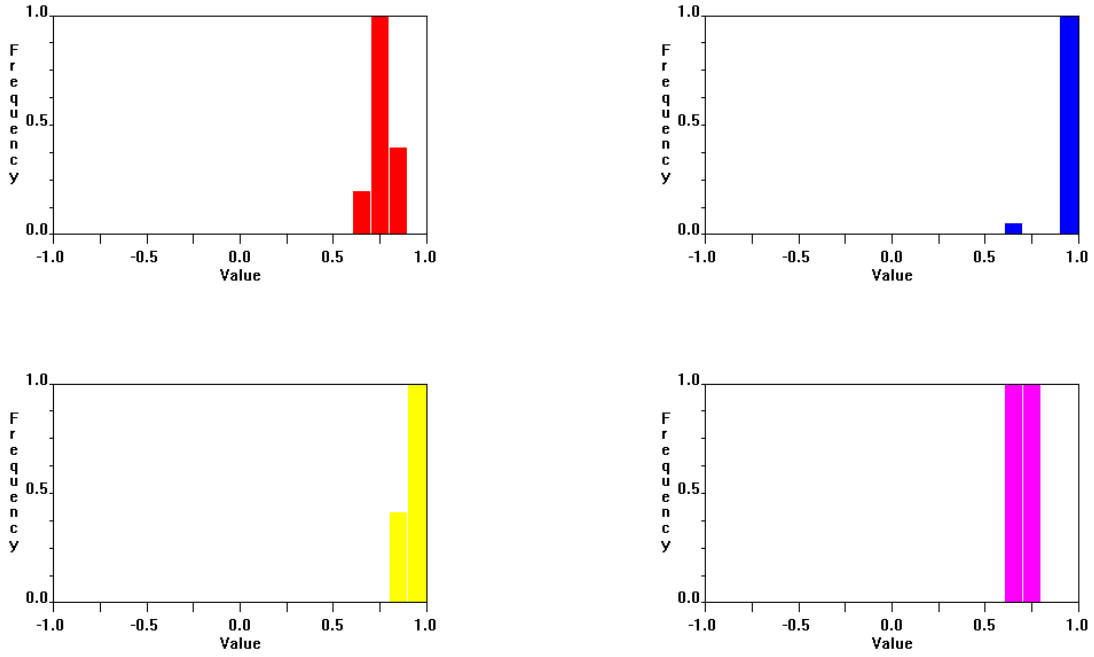

Figure 84 – Silhouettes

For the red cluster only sample 20-4 is found with a silhouette below 0.75.

For the blue silhouettes, the only sample that lies below 0.75 is 48-3, which is an outlier from this cluster in the MMDS plot.

The silhouettes for the yellow cluster are all found above 0.75 so no outliers are identified.

For the pink cluster sample 10-3 lies below 0.75. This was one of the poorly clustered form 3 samples previously examined.

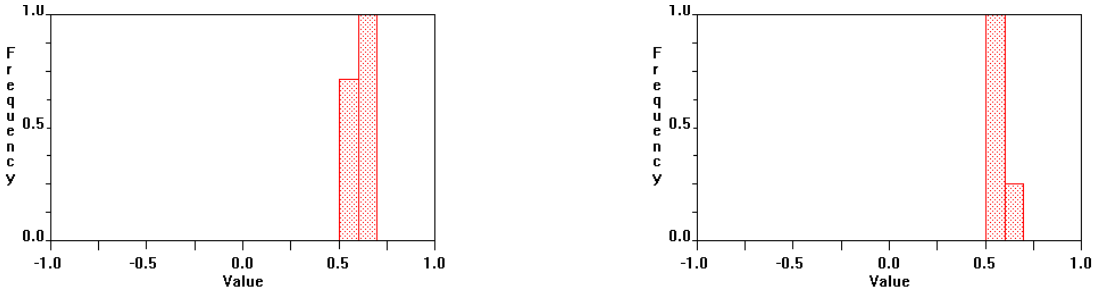

Figure 85 - Fuzzy Clustering

	1	2	3	4	5	6	
01-4	0.18	0.26	0.13	0.15	0.63*	0.86	<==
02-2	0.09	1	0.12	0.11	0.4	0.13	
03-3	0.15	0.58*	0.14	0.14	0.86	0.31	<==
04-2	0.09	1	0.13	0.11	0.41	0.13	
05-3	0.16	0.54*	0.14	0.14	0.88	0.34	<==
06-4	0.19	0.23	0.12	0.14	0.62*	0.87	<==
07-3	0.17	0.43	0.14	0.15	0.91	0.4	
08-3	0.16	0.48	0.14	0.15	0.9	0.37	
09-2	0.1	1	0.12	0.11	0.41	0.14	
10-4	0.87	0.04	0.07	0.08	0.52*	0.48	<==
11-2	0.1	1	0.12	0.11	0.4	0.13	
12-4	0.07	0.37	0.18	0.83	0.60*	0.35	<==
13-4	0.21	0.2	0.11	0.13	0.60*	0.88	<==
14-3	0.17	0.45	0.14	0.15	0.91	0.39	
15-2	0.1	1	0.13	0.11	0.41	0.13	
16-2	0.1	1	0.12	0.11	0.41	0.14	
17-3	0.17	0.48	0.14	0.14	0.9	0.38	
18-3	0.87	0.19	0.08	0.08	0.5	0.42	
19-4	0.17	0.33	0.14	0.15	0.63*	0.84	<==
20-4	0.22	0.21	0.1	0.12	0.58*	0.87	<==
21-2	0.09	1	0.13	0.11	0.4	0.13	
22-2	0.1	1	0.13	0.11	0.41	0.14	
23-4	0.17	0.35	0.14	0.15	0.64*	0.83	<==
24-4	0.21	0.19	0.11	0.13	0.61*	0.89	<==
25-3	0.17	0.46	0.14	0.15	0.9	0.39	
26-3	0.14	0.51*	0.15	0.15	0.89	0.34	<==
27-2	0.1	1	0.12	0.11	0.4	0.13	
28-2	0.1	1	0.13	0.11	0.41	0.14	
29-4	0.2	0.27	0.12	0.13	0.60*	0.86	<==
30-3	0.17	0.5	0.14	0.14	0.89	0.37	
31-3	0.17	0.43	0.14	0.15	0.91	0.4	
32-2	0.09	1	0.13	0.11	0.41	0.13	
33-2	0.1	1	0.12	0.11	0.4	0.14	
34-3	0.2	0.37	0.12	0.13	0.91	0.42	
35-2	0.09	1	0.13	0.11	0.41	0.14	
36-4	0.06	0.61*	0.82	0.18	0.53*	0.23	<==
37-2	0.17	0.45	0.14	0.15	0.91	0.39	
38-2	0.09	1	0.12	0.11	0.4	0.13	
39-3	0.18	0.45	0.14	0.14	0.91	0.39	
40-3	0.16	0.48	0.14	0.14	0.9	0.38	
41-2	0.1	1	0.12	0.11	0.41	0.14	
42-2	0.1	1	0.12	0.11	0.4	0.13	
43-3	0.17	0.5	0.14	0.14	0.9	0.37	
44-3	0.18	0.42	0.13	0.14	0.91	0.41	
45-2	0.09	1	0.13	0.11	0.41	0.14	
46-3	0.11	0.55*	0.15	0.15	0.86	0.29	<==
47-2	0.09	1	0.13	0.11	0.41	0.13	
48-3	0.13	0.84	0.15	0.15	0.59*	0.31	<==

Table 11 - Fuzzy Clustering Numerical Data

Cluster 1 is the purple cluster, cluster 2 is the blue cluster, cluster 3 is the green cluster, cluster 4 is the aquamarine cluster, cluster 5 is the yellow cluster and cluster 6 is the red cluster.

For the first fuzzy clustering plot, the lower of the two lines contains samples 10-4, 13-4, 20-4, 48-3 and 36-4 while the upper line refers to 01-4, 06-4, 12-4, 19-4, 23-4, 24-4 and 29-4.

For these samples:

- Sample 01-4 could potentially be in the yellow or red cluster
- Sample 06-4 could potentially be in the yellow or red cluster
- Sample 10-4 could potentially be in the purple or yellow cluster
- Sample 12-4 could potentially be in the aquamarine or yellow cluster
- Sample 13-4 could potentially be in the yellow or red cluster
- Sample 19-4 could potentially be in the yellow or red cluster
- Sample 20-4 could potentially be in the yellow or red cluster
- Sample 23-4 could potentially be in the yellow or red cluster
- Sample 24-4 could potentially be in the yellow or red cluster
- Sample 29-4 could potentially be in the yellow or red cluster
- Sample 36-4 could potentially be in the blue or yellow cluster
- Sample 48-3 could potentially be in the blue or yellow cluster

The second plot contains two lines, the lower of which represents samples 03-3, 05-3, 26-3 and 46-3 while the upper bar contains 36-4.

For these samples:

- Sample 03-3 could potentially be in the blue or yellow cluster
- Sample 05-3 could potentially be in the blue or yellow cluster
- Sample 26-3 could potentially be in the blue or yellow cluster
- Sample 46-3 could potentially be in the blue or yellow cluster

4.3.1 COMBINED PXRD AND RAMAN DATA

By combining the PXRD data and the unprocessed Raman data using the INDSCAL method, a new dendrogram and MMDS Plot (Figure 86) are generated which are a major improvement over the PXRD dataset.

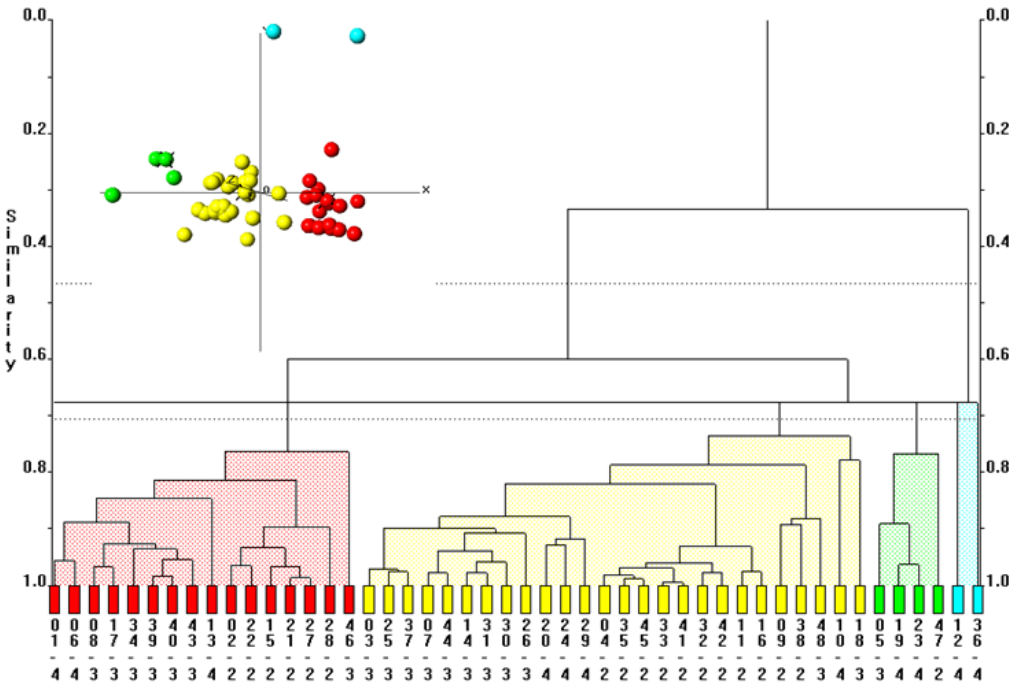

Figure 86 - Combined PXRD and Raman Dendrogram

This dendrogram loses the contiguous clustering of the form 2 samples that was present in the unprocessed Raman dendrogram. The samples are, however, easier to distinguish between than in the unprocessed dendrogram. The form 3 and 4 samples are not as well clustered as previously seen. This dendrogram has a score of 0.5.

The MMDS plot shows three separated clusters, however some of the samples between the red and yellow cluster could belong to either of these clusters.

The combination is re-run using the optimum Raman pre-processing methods (dendrogram and MMDS plot in Figure 87).

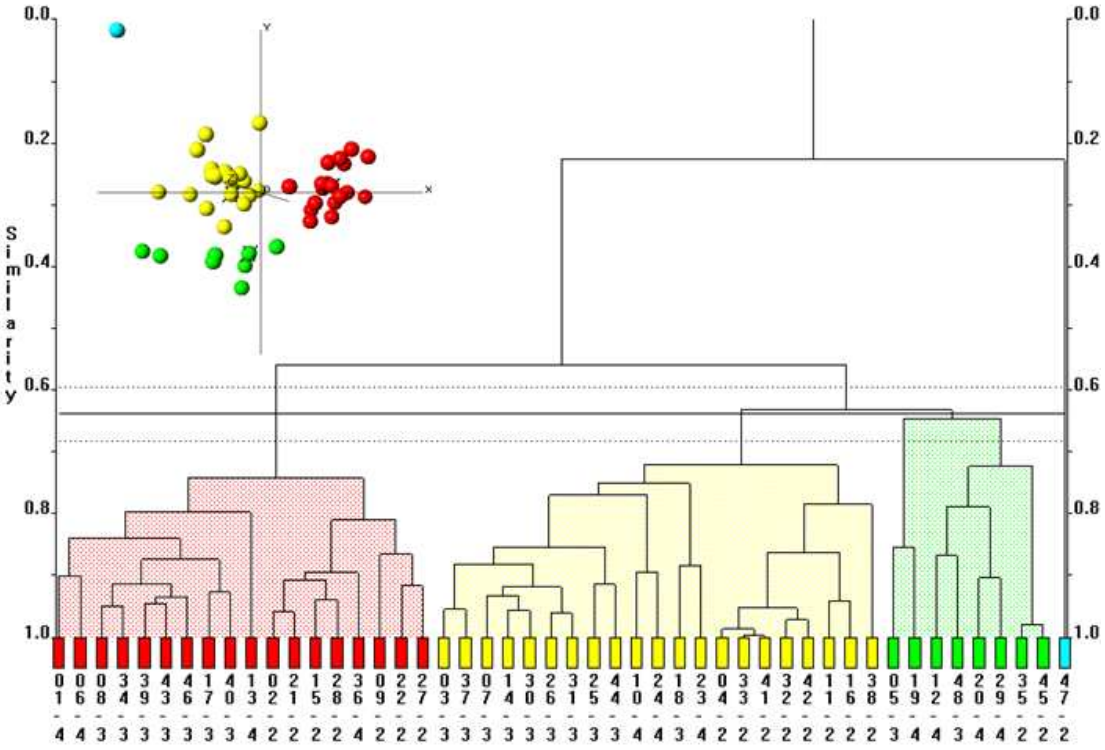

Figure 87 - Combined PXRD and Raman Dendrogram Using Optimum Raman Clustering

The combined dendrogram suffers from the same problem as the non-processed run. The good clustering of the form 2 samples is still missing, and the form 3 and 4 samples are still scattered.

The MMDS plot is not as tightly grouped as in the pure Raman run. The yellow cluster is very diffuse, however this can be explained as there are is a high tie bar linking the form 2 and form 3 groups found in this cluster.

Overall the clustering for the pre-processed INDSCAL dendrogram is not noticeably improved over the INDSCAL method which does not use pre-processing. With a score of 0.65 its clustering is actually slightly poorer.

4.3.2 RE-RUN X-RAY DATA

The PXRD data was re-run in order to attempt to resolve the dual problems of preferred orientation and poor background. The crystals were reground in the hopes of removing both problems. The data has been collected over the same range, the same timeframe and on the same instrument. The dendrogram for the re-run data is shown in Figure 88.

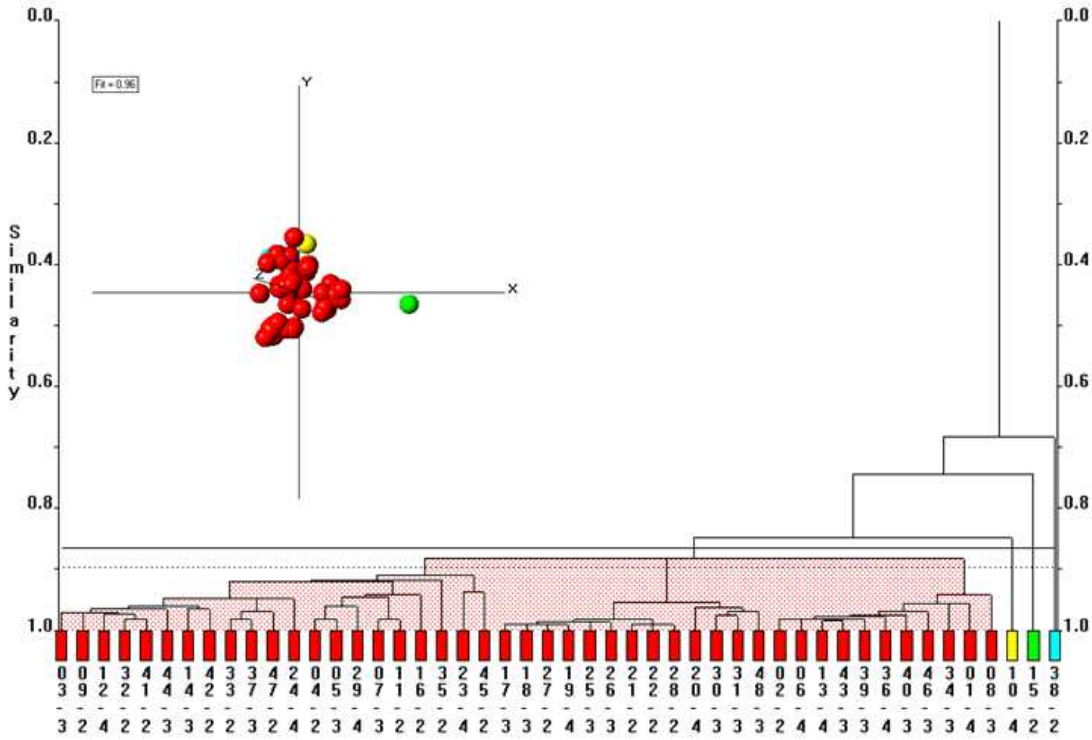

Figure 88 - Re-run PXRD Dendrogram and MMDS Plot

The re-run did not show improved clustering over the original data. No large contiguous groups of forms are visible. Due to the poor clustering when compared to previous runs, this dataset has a score of 0.79.

The MMDS plot does not show clear separation between samples.

The outliers that are not part of the large red cluster, when examined, are shown to have some preferred orientation issues. These patterns are shown in Figure 89.

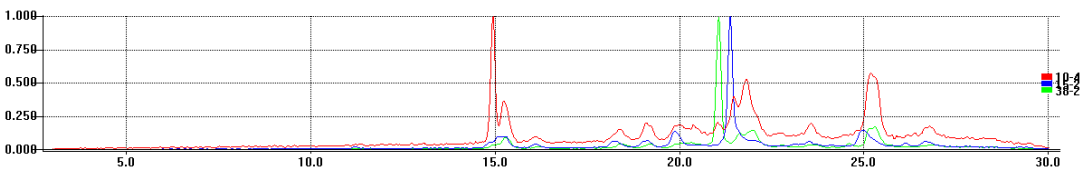

Figure 89 - re-run PXRD Patterns with Preferred Orientation

4.3.3 SECOND X-RAY DATA RE-RUN

The problem of poor sample quality was determined to be due to an alignment problem with the instrument's detector. The detector was realigned and the dataset collected a third time.

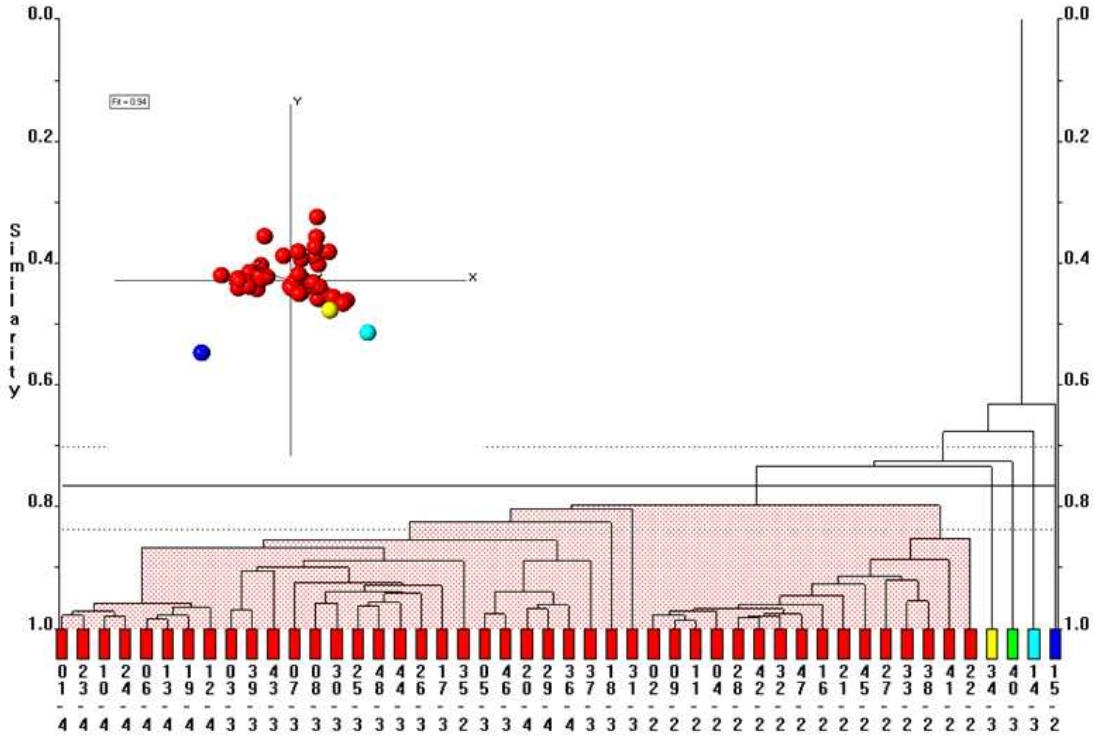

Figure 90 – Second Re-run PXRD Dendrogram and MMDS Plot

The second re-run shows improved clustering when compared to the previous two PXRD runs. All but two of the form 2 samples are clustered together. The form 3 samples are clustered together with eight outliers and the form 4 samples are clustered together with three outliers. The vastly improved clustering in this dataset when compared to the previous PXRD datasets results in a score of 0.27.

The MMDS plot does not show clear separation between samples.

There are a small number of outliers present in the dendrogram. When examined they are all revealed to have preferred orientation issues. These patterns are shown overlaid in Figure 91.

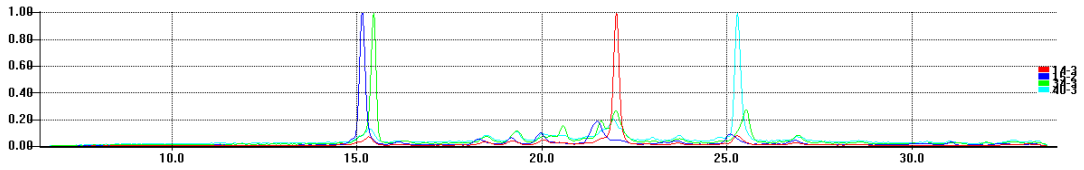

Figure 91 - Overlay of Preferred Orientation Samples from Second X-ray Re-run

4.3.4 HIGHER RANGE X-RAY DATASET

The X-ray data were re-run over a higher range. The dataset was still collected on a Bruker C2 GADDS, however was now collected over a 15-45º range as opposed to 5-35º. The dendrogram and MMDS plot for the higher range run are shown in Figure 92.

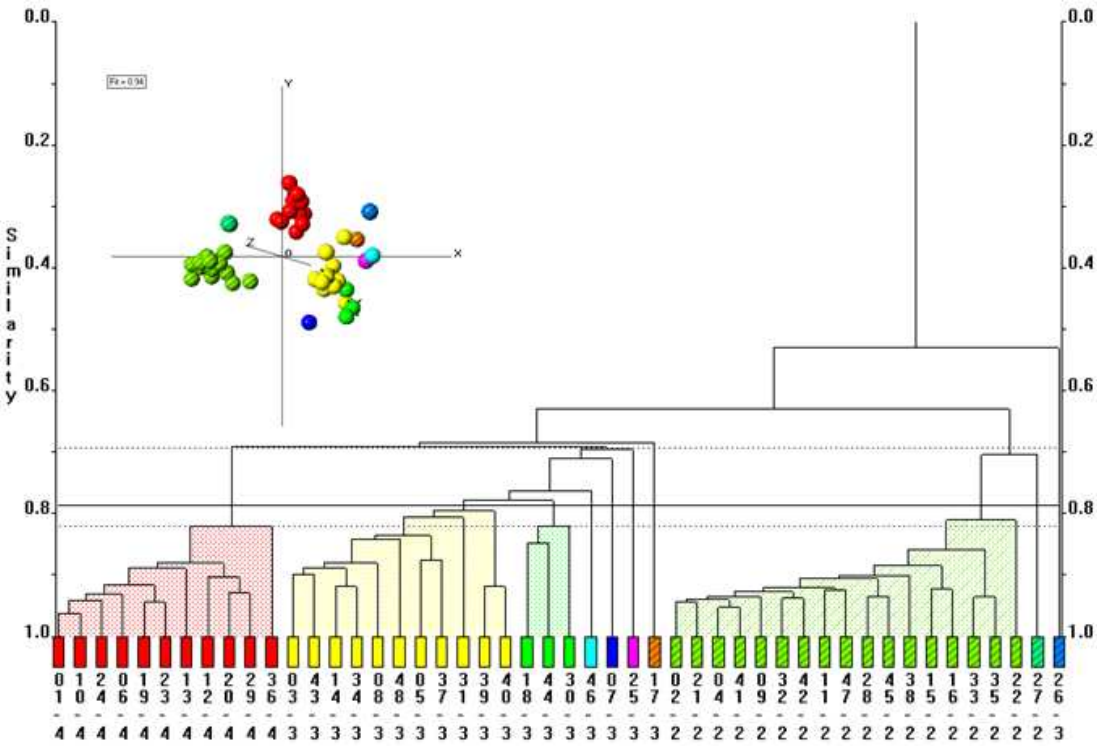

Figure 92 – Higher Range Run PXRD Dendrogram

The data is well clustered, with all three materials clearly separated. The red cluster contains all of the form 4 samples, the yellow cluster contains a large number of the form 3 samples, with the almost all of the remainder being spread across the adjoining green, aquamarine, blue, purple and striped brown clusters. The striped green cluster contains all but one of the form 2 samples with the additional one being in the adjoining striped dark green cluster. A small adjustment to the cut-level will merge the green, aquamarine, blue, purple and striped brown clusters into the yellow cluster and the striped dark green cluster into the striped green cluster. This will result in all but one of the form 3 samples, located

in the hashed blue cluster, being clustered as expected. In the current dendrogram, the large separation of the samples into different clusters results in a score of just 0.18 which is still a significant improvement over previously seen PXRD dendrogram. By raising the cut-level to the point where the red and yellow clusters are just about to merge this would improve to 0.06, a vast improvement over all previous clustering seen for PXRD data in this dataset.

The MMDS plot shows clearly separated clusters with the form 4 samples (red cluster) the form 2 samples (hashed green clusters) and the form 3 samples (remaining colours) all clearly separated. Despite the differences in the dendrogram, the misclustered form 3 (hashed blue) is still positioned close to the form 3 samples. The

An overlay of this poorly clustered sample with the most representative form 3 sample is shown in Figure 93.

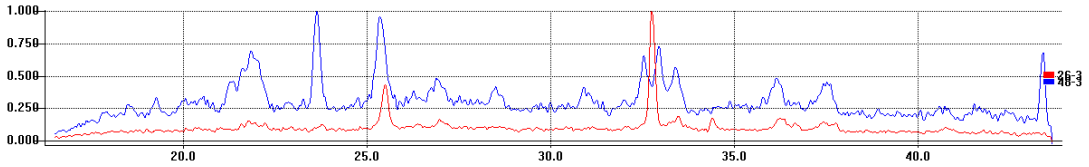

Figure 93 - Poorly clustered Form 3 sample overlay

The poorer quality sample is suffering from preferred orientation issues.

The dendrogram and MMDS plot produced from the combination of the PXRD and the Raman dataset with no pre-processing applied are shown in Figure 94.

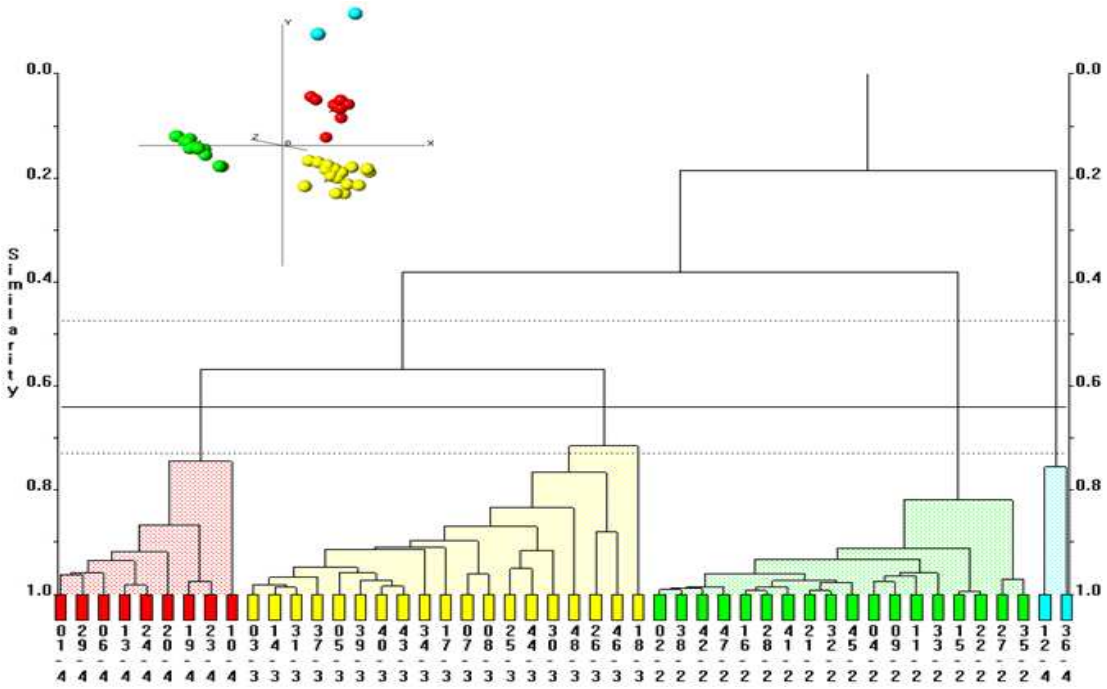

Figure 94 - Higher Range Run Combined Dendrogram and MMDS Plot

The combined dendrogram has two of the form 4 samples in an unexpected location. The form 2 samples (green cluster) and the form 3 samples (yellow cluster) are all clustered as expected.

The MMDS plot shows the clusters to be clearly defined with the two outliers (aquamarine) being positioned far from the remaining clusters. The score has now improved to 0.04, an improvement even over that seen in the good quality higher range X-ray dataset.

The poorly clustered samples are overlaid with the most representative form 4 sample (24-4). The X-ray data were a good match, however the Raman data showed some clear differences which are shown in Figure 95.

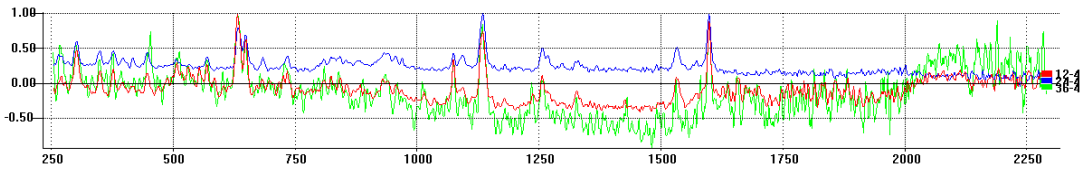

Figure 95 - Overlay of Poorly Clustered Samples

The large difference in background between these samples accounts for them not being clustered with the remainder of the form 4 samples.

The PXRD data was again combined with the Raman data. For this second combination, the Raman data was run with the previously determined optimal pre-processing methods. The dendrogram and MMDS plot are shown in Figure 96.

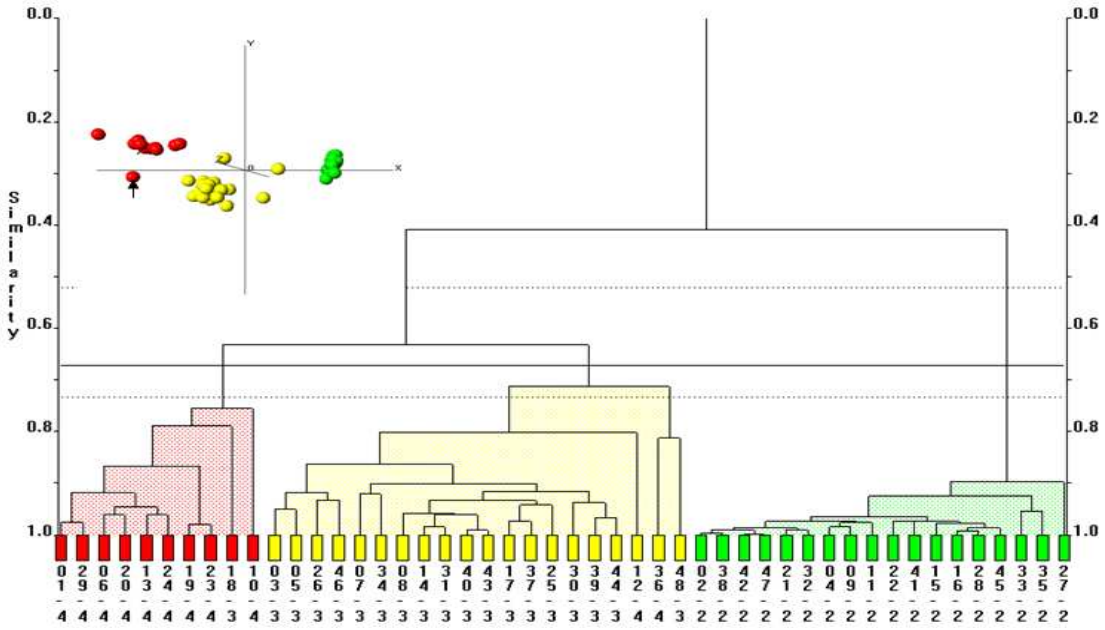

Figure 96 - Higher Range Run Combined Dendrogram

The clustering in this dataset is poorer than in the previous combination. Three form 4 samples and two form 3 samples are not clustered as expected. The form 2 samples are all grouped together as expected.

The MMDS plot shows a clearly defined green cluster, containing all the form 2 samples. The red cluster is clearly separated from the yellow cluster, however contains some outliers. The outlier marked with an arrow in the MMDS plot is the lone form 3 samples in this dataset.

The yellow cluster is clearly defined with three outliers. The two uppermost ones represent the two misplaced form 4 samples in this cluster, which are also the only two misplaced samples in the INDSCAL combination with no pre-processing applied. The remaining outlier is the lone form 3 sample in this cluster with the higher tie-bar than the remaining ones. The score has dropped slightly to 0.06, the same as that seen for the PXRD data alone.

4.3.5 HIGHER RANGE RUN PXRD AND DERIVATIVE RAMAN COMBINATION

The previously studied effects of combining derivative Raman data with the original Raman data produced good clustering in the combined dendrogram. Second derivative data combined with the original data produced the best results from these combinations. The effect of combining second derivative Raman data, the original Raman data and the higher range PXRD data will now be studied.

The dendrogram for this combination is shown in Figure 97.

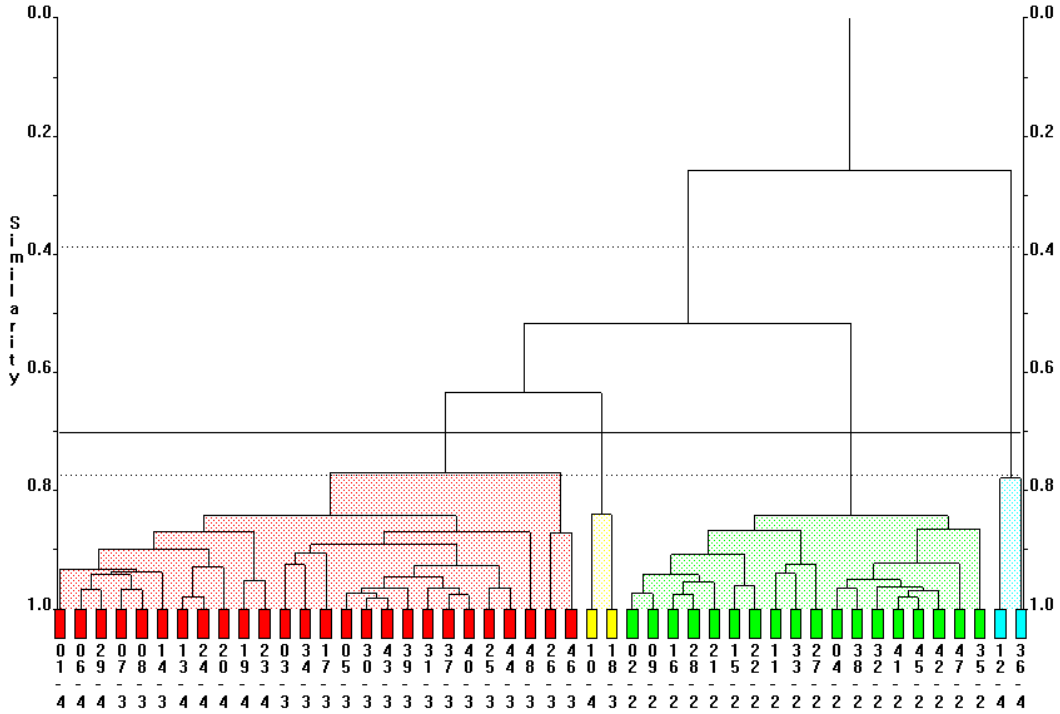

Figure 97 - Combined Second Derivative and Original Raman and Higher Range PXRD Dendrogram

This dataset shows clearly separated form 2 samples in the green cluster. The form 3 and 4 samples are both present in the red cluster. The form 3 samples are clearly defined within this cluster; however the form 4 samples are split in half by some form 3 samples. There are four outliers from these clearly defined clusters.

The two outliers in the aquamarine cluster have previously been revealed to be due to poorer quality Raman data. This dataset has a score of 0.15.

Figure 98 shows the resulting dendrogram when first derivative and the original Raman data are combined with the higher range PXRD data.

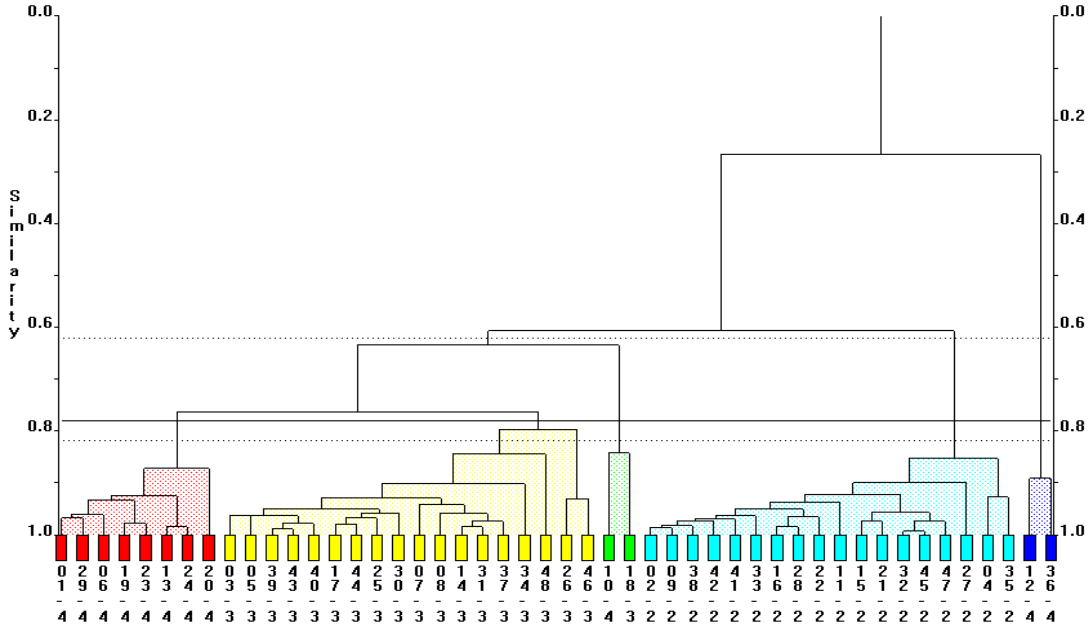

Figure 98 - Combined First Derivative and Original Raman and Higher Range PXRD Dendrogram

The resulting dendrogram has four outliers however unlike the second derivative combined dendrogram, the form 4 and form 3 samples are present in separate clusters. This dataset has a score of 0.08.

The result of combining first derivative, second derivative and the original Raman data, along with the higher range PXRD data, can be seen in Figure 99.

The two outliers in the blue cluster have previously been revealed to be due to poorer quality Raman data.

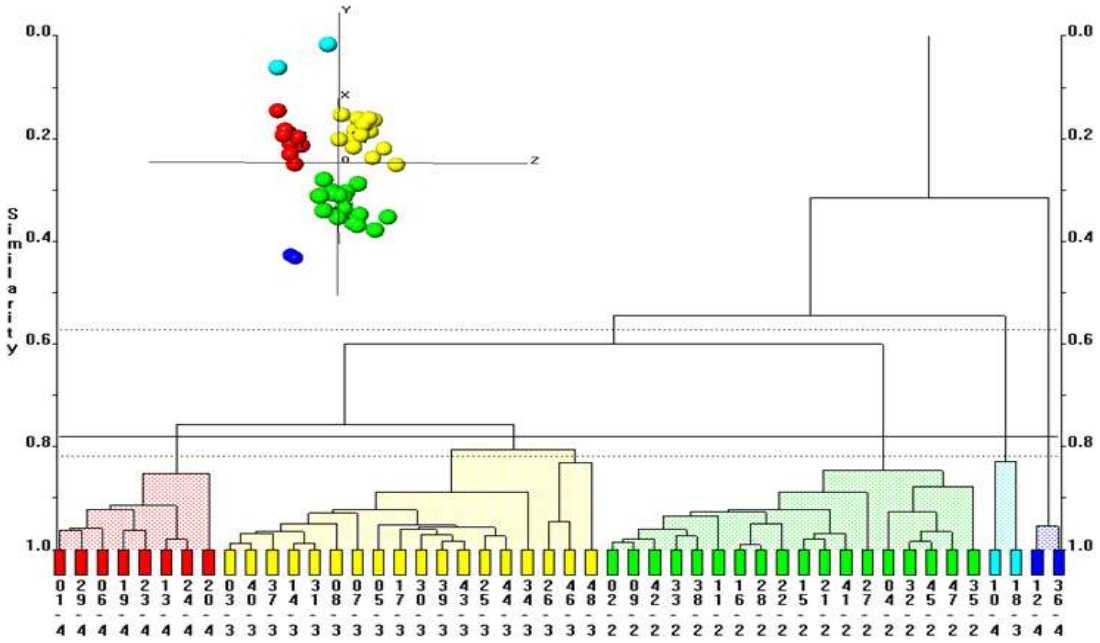

Figure 99 - All Raman Combined and Higher Range PXRD Dendrogram

This dataset shows very clear clustering. All form 2 samples are present in the green cluster. The yellow cluster contains all but one of the form 3 samples, the missing sample being an outlier from the main clusters. The red cluster contains all but three of the form 4 samples, with these three appearing as outliers elsewhere. In total four outliers are present, with each of the major clusters containing no mixtures of different polymorphs. This dendrogram has a score of 0.08.

All three of the major clusters are clearly separated from one another in the MMDS plot. The two outliers in the blue cluster have previously been discussed when it was revealed that they were due to poorer quality Raman data.

The other two misclustered samples are again due to Raman data rather than the data. The Raman spectra are shown, along with the most representative form 3 and form 4 samples, in Figure 100.

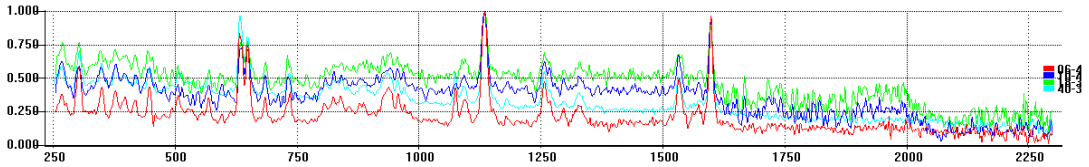

Figure 100 - Overlay of Poorly Clustered Samples

The most representative samples (red and aquamarine lines) are revealed to have much smoother baselines and lower backgrounds than the misclustered spectra.

If these outliers are removed and the dataset re-clustered, the resulting INDSCAL combination of all four datatypes gives the dendrogram and MMDS plot shown in Figure 101.

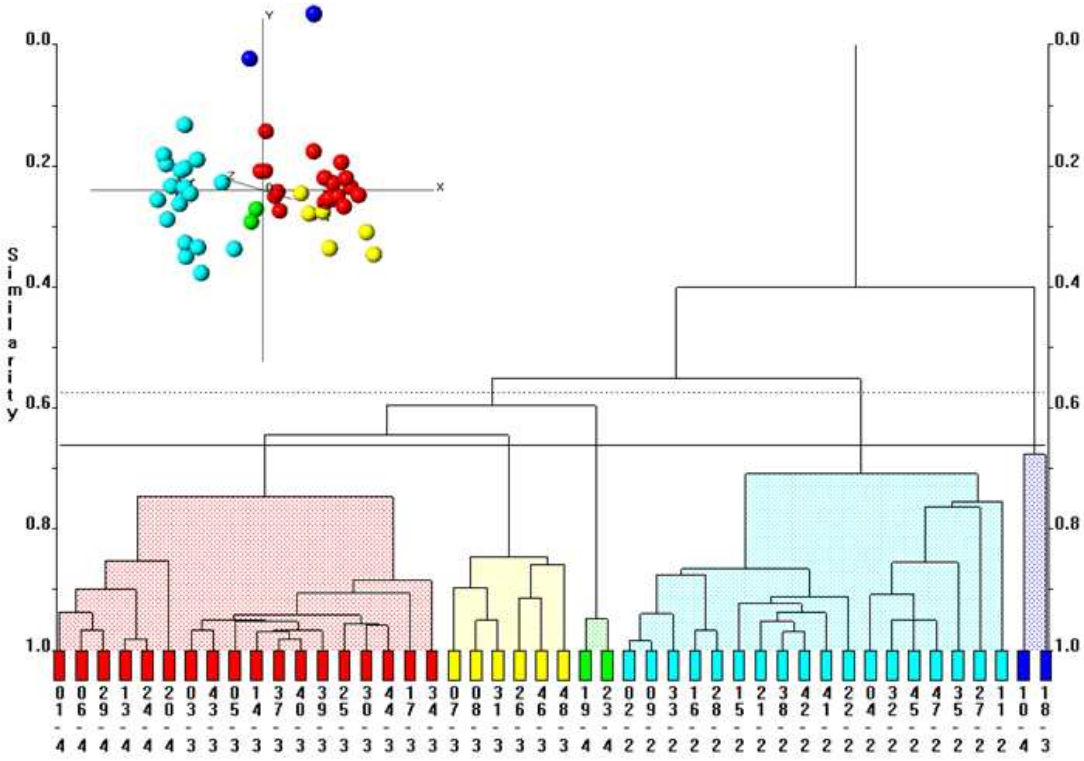

Figure 101 - All Raman Combined and Higher Range PXRD Dendrogram with Outliers Removed

The form 2 samples are all found in the aquamarine cluster. The form 3 samples are split across the yellow cluster and the red cluster with a single outlier in the blue cluster. The red cluster also contains the majority over the form 4 samples with a single outlier in the blue cluster and two outliers in the green cluster. It is possible to separate the form 4 and 3 samples in the red cluster by lowering the cut-level however this will also split the form 2 samples into two separate clusters. This dataset has a score of 0.08.

Figure 102 shows the effects of applying the optimal pre-processing to second derivative, first derivative and original Raman data and then combining all of these with the higher range PXRD data.

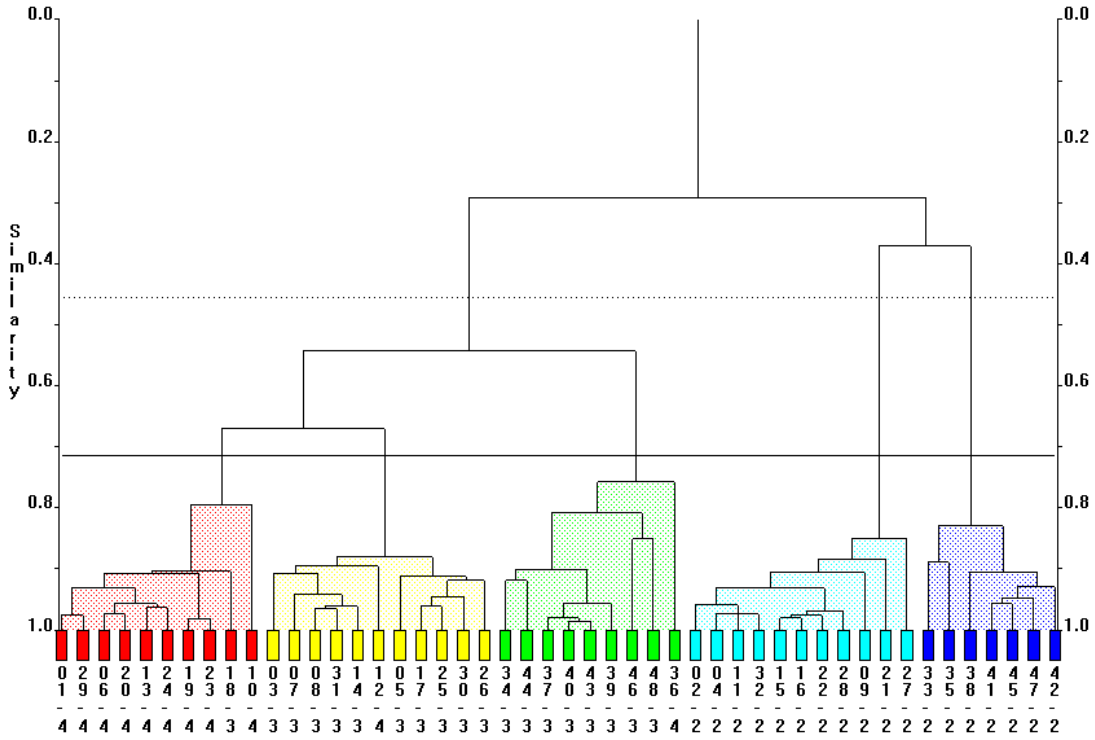

Figure 102 - All Raman Combined with Pre-processing and Higher Range PXRD MMDS

The dataset is not as well clustered as the equivalent run without applied pre-processing. There are still four outliers, however they are now combined into other clusters rather than lying separately from the main clusters. In addition to this the form 3 samples are no longer clustered together into a single cluster but are split into two clusters. The form 2 samples also have been split into two clusters. The dendrogram has a score of 0.19.

4.4 FLOWCHART

The PXRD data was run without pre-processing and with every possible combination of pre-processing available. The results are shown in Table 12.

	Score
no pre-processing	0.06
denoise	0.06
background	0.1
background and denoise	0.1

Table 12 - Misplaced samples

No pre-processing and denoised only give identical results. Background removal and background removal plus denoise also give identical results. As such the optimal methodology for clustering this dataset is shown in Figure 103.

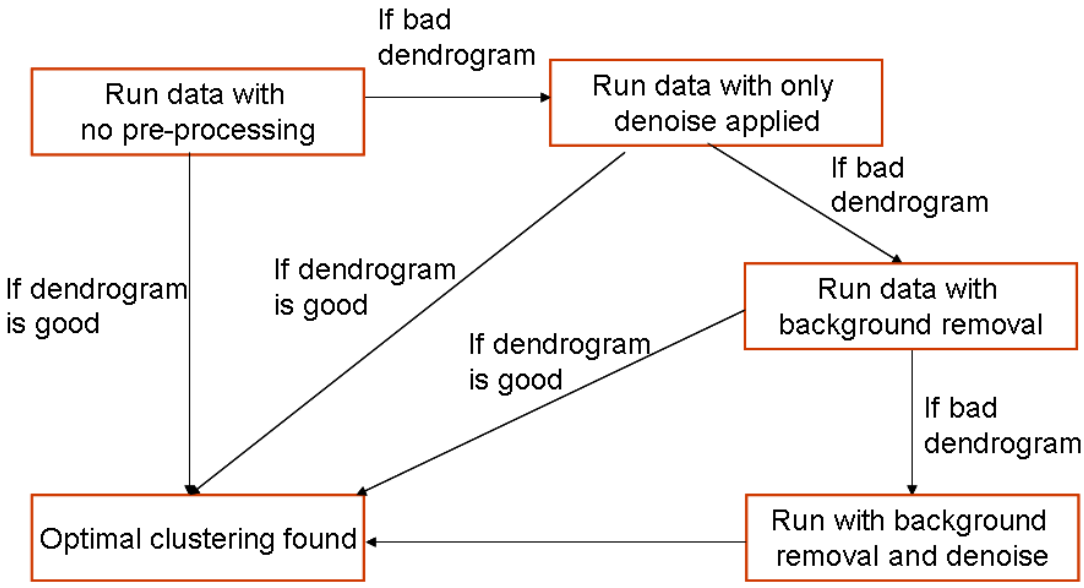

Figure 103 - Flowchart for optimum PolySNAP clustering

A dendrogram can be determined to be good or bad as follows.

1. The dendrogram shows ‘chaining’ of the samples or has no clear clustering
2. The scree plot does not show the characteristic steep initial drop before smoothing out
3. The maximum and minimum confidence on the dendrogram shows a large separation.

If any of the above problems exist in a dataset then the flowchart should be followed onto the next step.

4.5 CONCLUSION

- The 48 sample dataset has successfully tested the use of Raman data alongside PXRD data in PolySNAP. Raman data can, with careful selection of the appropriate pre-processing methods, give good results. The main problem with Raman data, that of high pattern similarity causing all patterns to have similarities of greater than 95%, can be overcome through the use of first or second derivative pre-processing of the data before pattern matching is performed. Second derivative pre-processing gives poorer clustering on its own, however when combined with the original data it can yield far better clustering that seen from matching the original data. A first derivative and original data combination also gives good clustering; however this is not as good as the second derivative combination.
- INDSCAL combinations of PXRD and Raman data also give good results. These results appear to be less dependent on pre-processing and can partially allow a poor quality dataset to give reasonable results if it is combined with a good quality one.
- Combining derivative Raman data, the original Raman data and PXRD data also gives good results. The best results, for this dataset, come from a combination of both first and second derivative Raman data with the original Raman data and PXRD data with no pre-processing applied.
- The various re-runs of the PXRD data show that, although the software can do a lot to try and resolve some of the problems of poorer quality data, there is still no substitute for having the best possible data quality to begin with.
- The optimal method for clustering, based on the scores assigned to each dendrogram from the mis-clustered samples, involves no pre-processing or denoising. No pre-processing is preferred as this involves applying no changes to the dataset.

CHAPTER 5 SULFATHIAZOLE/CARBAMAZEPINE DATASET

5.1 THE DATASET

The sulfathiazole/carbamazepine (SUTHAZ/CBZ) dataset contains polymorphs two, three and four of sulfathiazole and polymorphs one and three of carbamazepine. Table 13 summarises the composition of this dataset.

For this dataset, PXRD data were collected on a Bruker C2 GADDS with each sample being run for two minutes over a 3-30° range in 2θ. Raman data were collected on a Witec alpha 300 with a 785nm laser and an x10 objective lens with 0.25 aperture and 300g/mm grate. DSC data were collected on a TA Instruments Q100. IR data were collected on a JASCO FT/IR 4100.

Sample Number	Sample ID	Name in PolySNAP	Composition
1	Sulfathiazole Form 4	s4	
2	Sulfathiazole Form 3	s3	
3	Sulfathiazole Form 2	s2	
4	Carbamazepine Form 1	c1	
5	Carbamazepine Form 3	c3	
6	Sulfathiazole Forms 3 and 4	s4+3	58:42
7	Sulfathiazole Forms 2 and 3	s3+2	63:37
8	Sulfathiazole Forms 2 and 4	s4+2	32:68
9	Carbamazepine Forms 1 and 3	c1+3	72:28
10	Sulfathiazole Forms 2, 3 and 4	s2+3+4	53:18:29
11	Sulfathiazole Form 2 and Carbamazepine Form 1	s2+c1	50:50
12	Sulfathiazole Form 3 and Carbamazepine Form 1	s3+c1	50:50
13	Sulfathiazole Form 4 and Carbamazepine Form 1	s4+c1	61:39
14	Sulfathiazole Form 2 and Carbamazepine Form 3	s2+c3	80:20
15	Sulfathiazole Form 3 and Carbamazepine Form 3	s3+c3	83:17
16	Sulfathiazole Form 4 and Carbamazepine Form 3	s4+c3	82:18

Table 13 – Suthaz/Cbz dataset composition

The notation for PolySNAP is follows. A pure polymorph takes the first letter of the materials name and the number that has been assigned to that polymorph in previous

literature to give a two character name, for example sulfathiazole form 2 becomes s2. For mixtures of different polymorphs of the same material, the first letter of the materials name is used at the start followed by the numbers of each polymorph present in the mixture, with the numbers separated by pluses. For example a mixture of polymorphs 2 and 3 of sulfathiazole will be called s2+3. Finally for mixtures of polymorphs of different materials, the two character names for the relevant polymorphs are used, separated by pluses. For example mixtures of carbamazepine form 1 (c1) and sulfathiazole form 3 (s3) will be called c1+s3.

5.2 SIMULATED DATASET

5.2.1 SIMULATED DATA CLUSTERING

Simulated powder data for each of the pure materials was taken from the CSD and were combined to produce the predicted patterns for each mixture. For example the c1+3 pattern was produced by combining the patterns of c1 and c3 in a 72:28 ratio. The resulting dendrogram and MMDS plot are shown in Figure 104.

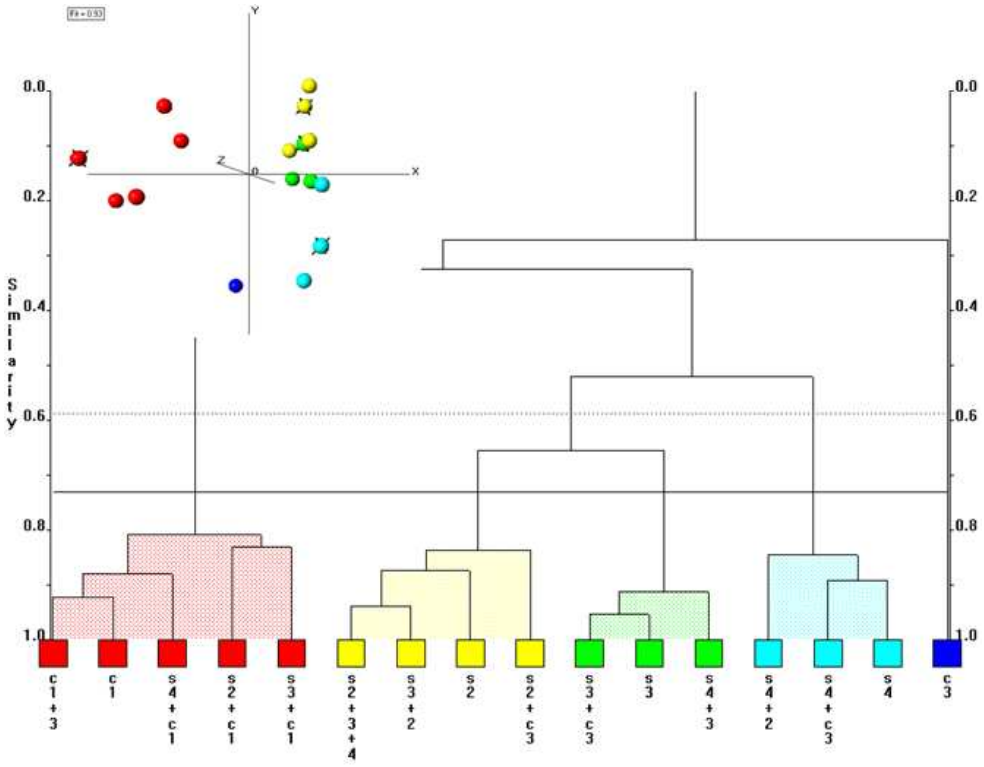

Figure 104 – Dendrogram and MMDS Plot for Simulated Dataset Clustering

The red cluster contains samples c1+3, c1, s4+c1, s2+c1 and s3+c1. The yellow cluster contains sample s2+3+4, s3+2, s2 and s2+c3. The green cluster contains samples s3+c3, s3 and s3. The aqua cluster contains samples s4+2, s4+c3 and s4 and the blue cluster contains samples c3. The silhouettes are shown in Figure 105.

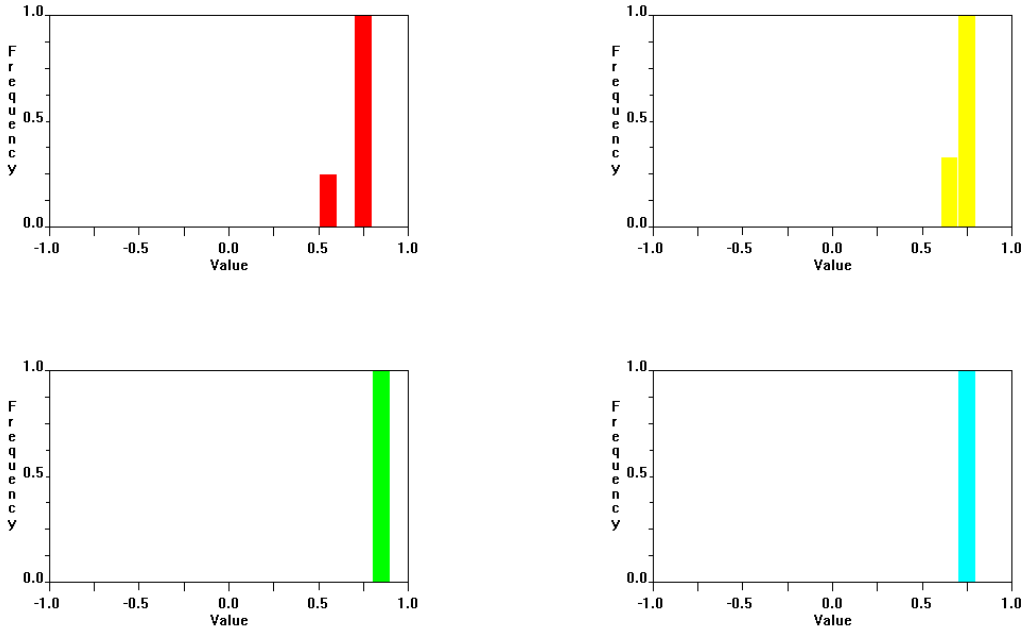

Figure 105 – Silhouettes for simulated dataset

All of the patterns in the silhouettes lie above 0.5. The lower lying bar in the red clusters silhouette represents sample s3+c1. The lower lying bar in the yellow clusters silhouette represents sample s2+c3.

5.3 FINDING THE OPTIMAL CLUSTERING

Minitab[1] has been used to generate optimal Pearson and Spearman correlation matrixes as a tool for the determination of optimal methods for analysing each dataset and to allow for the development of a flowchart showing the optimal clustering methods that can be used with an unknown dataset.

The Minitab method was carried out as follows:

1) The Cambridge structural database[11] was searched for the pure patterns of each material studied.
2) These pure patterns were manually combined, in the ratios shown above, to produce the mixture patterns.
3) All of these patterns were initially correlated using Minitab to produce a Pearson correlation matrix.

4) The patterns were ranked and again correlated to produce a Spearman correlation matrix.
5) Both correlation matrixes are imported into PolySNAP, as outlined in Chapter 1, along with the collected data.

Figure 106 shows the dendrogram and MMDS plot for the Pearson correlation matrix and Figure 109 the dendrogram and MMDS plot for the Spearman correlation matrix.

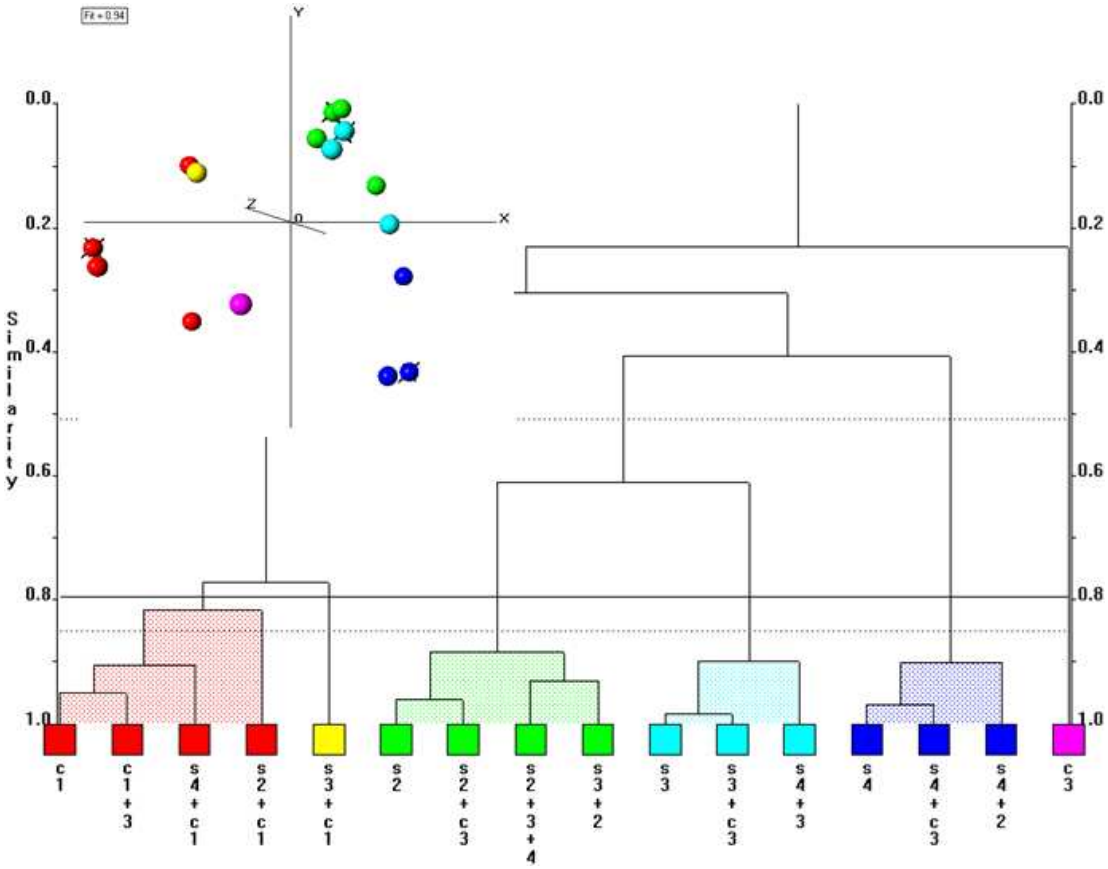

Figure 106 - Dendrogram and MMDS plot for Pearson correlation matrix using simulated data

The red cluster contains sample c1, c1+3, s4+c1 and s2+c1. The yellow cluster contains sample s3+c1. The green cluster contains samples s2, s2+c3, s2+3+4, and s3+2. The aquamarine cluster contains samples s3, s3+c3 and s4+3. The blue cluster contains samples s4, s4+c3 and s4+2 and the purple cluster contains sample c3.
The silhouettes are shown in Figure 107.

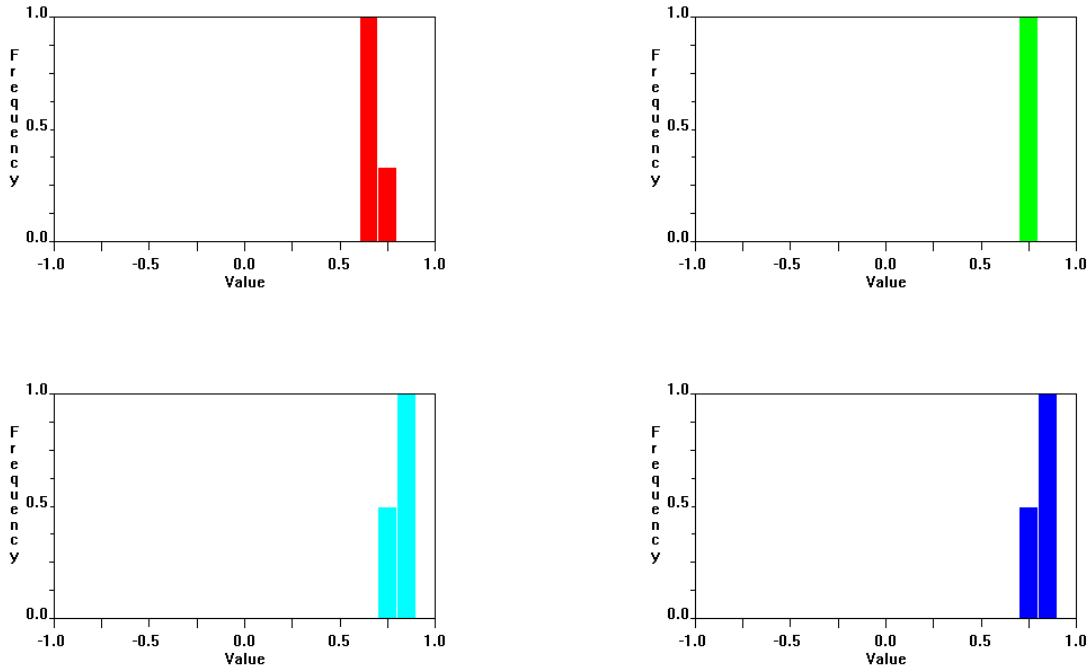

Figure 107 - Pearson Silhouettes

For the red cluster the lower bar, just below 0.75, contains samples c1+3, s2+c1 and s4+c1. All samples in the green cluster are present in a single bar. The aquamarine cluster has sample s4+3 in the lower bar, present at 0.75. The blue cluster has sample s4+2 present in the lower bar at 0.75.

Fuzzy clustering is not present for this dataset.

When the dendrogram for this dataset is compared to the dendrogram for the simulated dataset (Figure 102), it is revealed that the datasets are only differing by 1 sample. This can be corrected by raising the cut-level to combine the yellow and red cluster as shown in Figure 108.

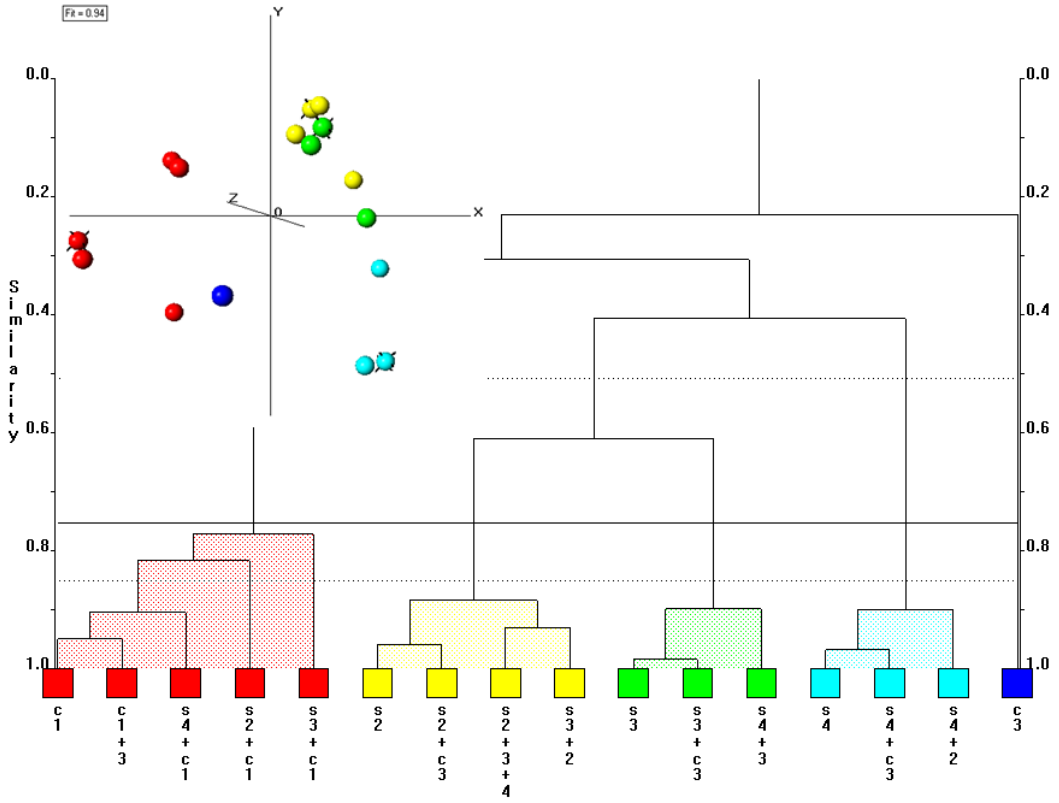

Figure 108 – Adjusted dendrogram and MMDS plot for Pearson correlation matrix

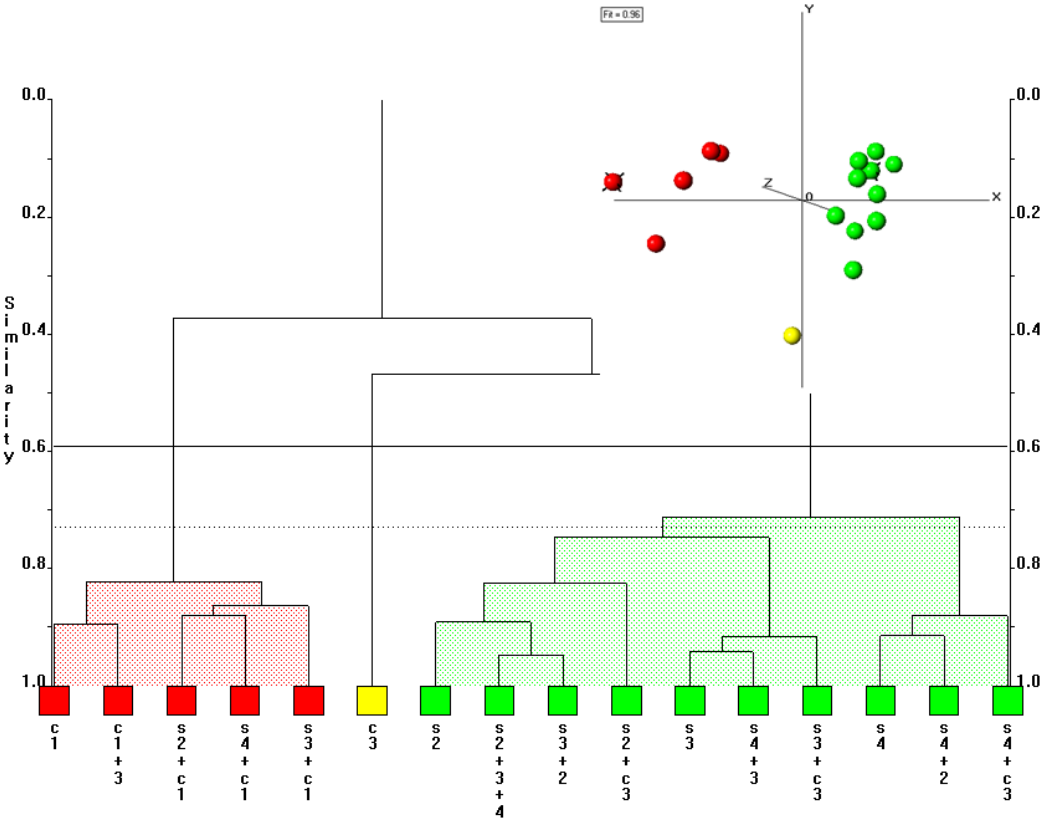

Figure 109 - Dendrogram and MMDS plot for Spearman correlation matrix

The red cluster contains samples c1, c1+3, s2+c1, s4+c1 and s3+c1. The yellow cluster contains sample c3. The green cluster contains samples s2, s2+3+4, s3+2, s2+c3, s3, s4+3, s3+c3, s4, s4+2 and s4+c3.

The silhouettes are shown in Figure 110.

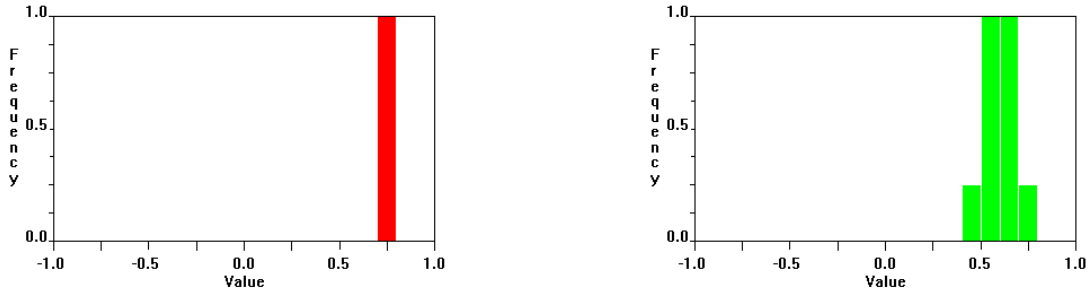

Figure 110 - Spearman Silhouettes

For the red cluster all samples are present in a single bar. The green cluster has sample s4+c3 in the lower bar just below 0.5. The bar just above 0.5 contains samples s2, s4, s2+c3 and s3+c3. The bar just below 0.75 contains samples s3, s3+2, s4+2 and s4+3. the upper most bar contains sample s2+3+4.

The fuzzy clustering is shown in Figure 111.

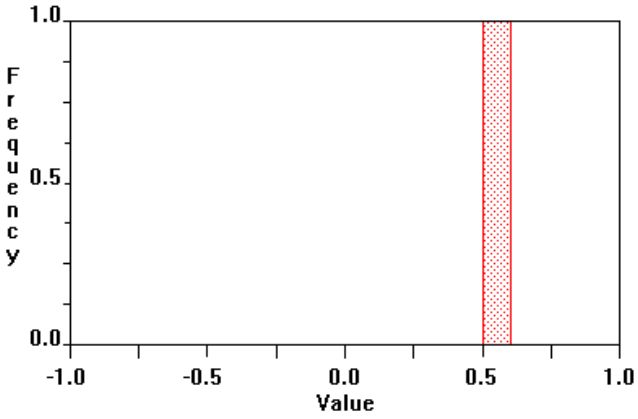

Figure 111 - Spearman Fuzzy Clustering

The table of fuzzy clustering results is shown in Table 14.

	1	2	3	
c1	0.01	0.01	0.77	
c3	0.04	0.75	0.17	
s2	0.27	0.27	0.17	
s3	0.34	0.25	0.22	
s4	0.22	0.34	0.19	
c1+3	0.04	0.06	0.77	
s2+3+4	0.33	0.32	0.22	
s2+c1	0.16	0.09	0.73	
s2+c3	0.24	0.29	0.22	
s3+2	0.33	0.27	0.22	
s3+c1	0.75	0.06	0.56*	<==
s3+c3	0.34	0.29	0.26	
s4+2	0.27	0.34	0.21	
s4+3	0.34	0.32	0.23	
s4+c1	0.13	0.07	0.77	
s4+c3	0.22	0.34	0.19	

Table 14 – Spearman Fuzzy clustering results

For the table of results, column 1 represents the green cluster, column 2 represents the yellow cluster and column 3 represents the red cluster. The fuzzy clustering peak corresponds to sample s3+c1 in the red cluster. This sample could appear in either the red or green cluster.

As the PXRD data for the simulated dataset exactly matches the data from the Pearson correlation coefficient, this will be considered the optimal method for analysing the dataset.

5.4 RAMAN AND IR DATASET ANALYSIS

The pure Raman data were compared to find areas of significant peaks. Overlays of these spectra are shown in Figure 112.

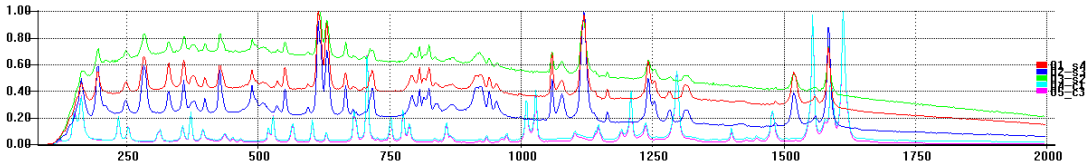

Figure 112- Overlay of pure Raman data

The red line shows sample s4, the blue line sample s3, the green line sample s2, the aquamarine line sample c1 and the purple line sample c3. The spectra between samples c1 and c3 can be seen to be nearly identical with a 98% similarity reported between the two. This comes entirely from differences in the background. The largest difference between the sulfathiazole samples comes from samples s3 and s2, with a reported similarity of 85%. The most notable differences between these two patterns can be seen in the difference between some of the peak heights. An overlay of these two patterns is shown in Figure 113.

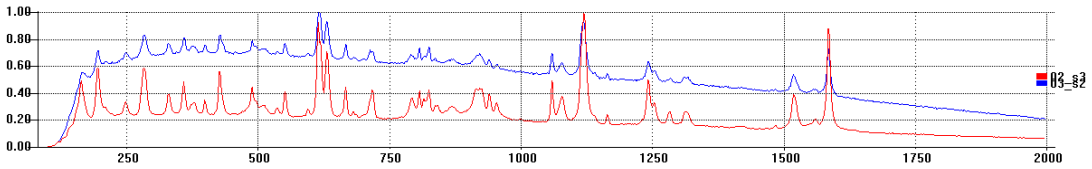

Figure 113 - Overlay of samples s2 and s3

Both spectra contain the same peaks however some of these peaks, in particular at 600-650cm^{-1}, 1100-1130cm^{-1} and 1560-1600cm^{-1} all have noticeable differences in peak height and so can be classed as being of lower significance between the spectra.

A comparison between the sulfathiazole and carbamazepine samples spectra reveals that almost all areas of the spectra have differences in peak positions. The only area that does not differ between the sulfathiazole and carbamazepine spectra is the region from 1750cm^{-1} to 2000cm^{-1}. As such this area will not be used during clustering.

The IR dataset was also examined to try and find areas of high similarity that can be ignored in all future clustering. An overlay of the spectra of the pure materials is shown in Figure 114.

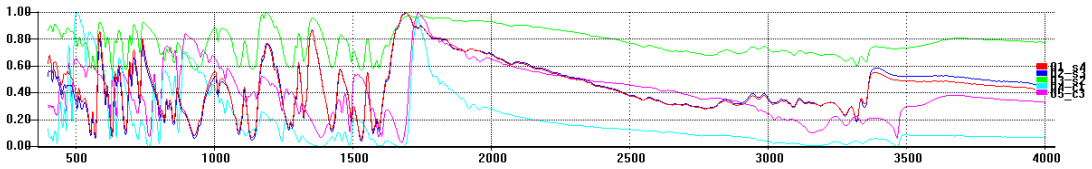

Figure 114 - Overlay of pure IR materials

All of the spectra show a long ‘tail’ from 2000 to 3000 cm^{-1} with no peaks present. The largest area of dissimilarity can be seen in the region before 1750cm^{-1} with small

differences occurring in the region from 3000-3500cm^{-1}. As such only the areas before 1750cm^{-1} and from 3000-3500cm^{-1} will be used in the cluster analysis.

5.5 DATASET CLUSTERING

5.5.1 EXPECTED CLUSTERING

The PXRD patterns for the pure polymorphs were combined to produce predicted patterns for each of the mixtures. Unlike for the predicted dataset from the CSD, this data was created using data specifically collected for this project. These predicted patterns were clustered and the dendrogram and MMDS plot are shown in Figure 115. This method of determining the expected clustering is different from that used in Chapter 4 as this dataset does not consist entirely of the pure polymorphs of a single material, making the expected clustering harder to determine.

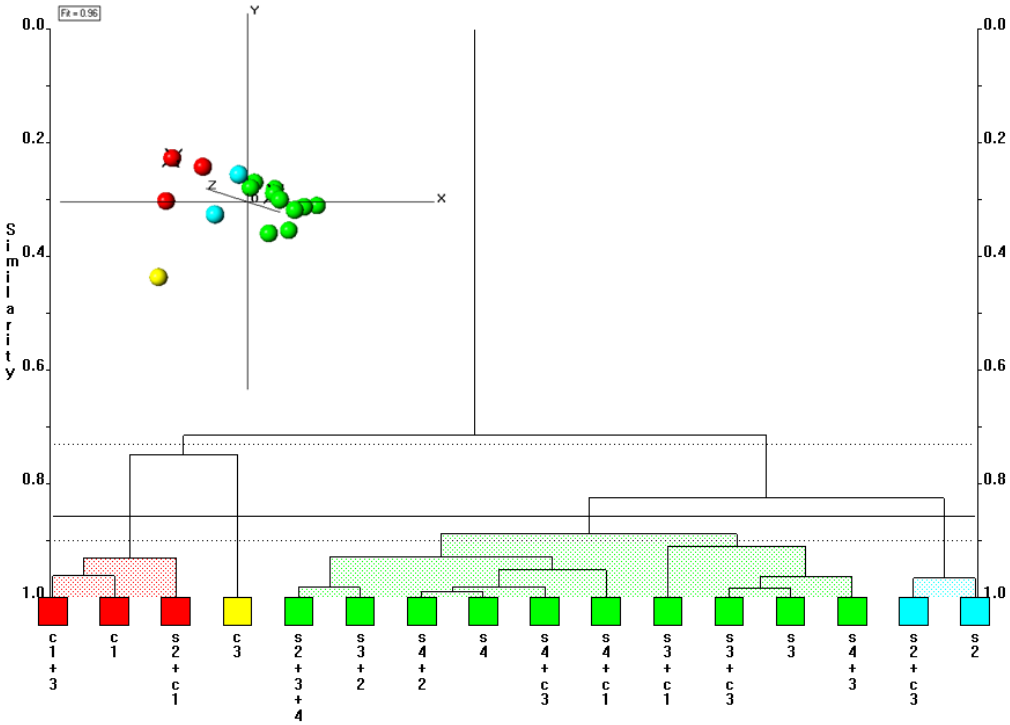

Figure 115 - Dendrogram and MMDS Plot for simulated data

The red cluster contains patterns c1, c1+3 and s2+c1. The yellow cluster contains the c3 patterns. The aquamarine cluster contains the s2+c3 and the s2 patterns and the green cluster contains all the remaining patterns. The green cluster could be easily split into two smaller clusters by adjusting the cut-level downwards. The silhouettes are shown in Figure 116 and fuzzy clustering in Figure 117 with the numerical results in Table 15.

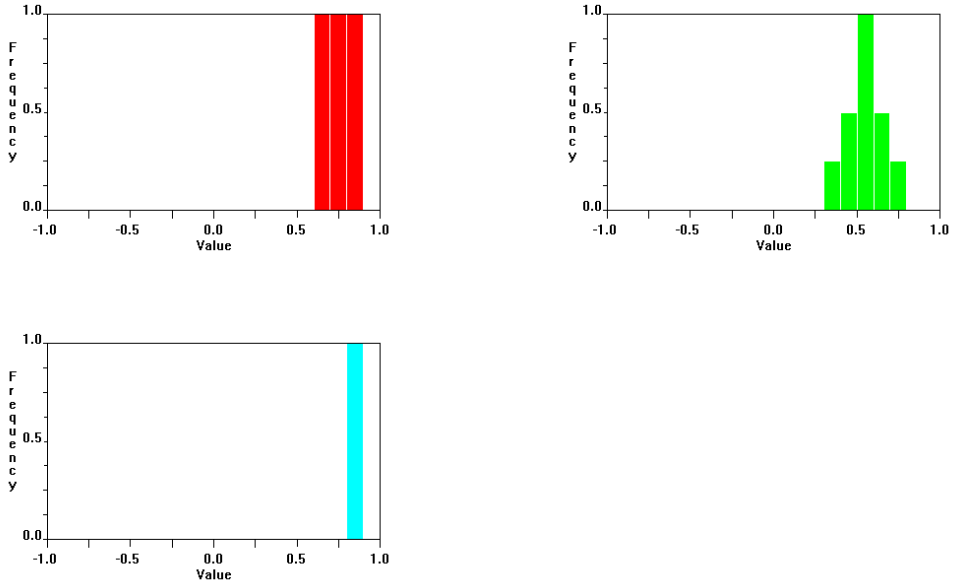

Figure 116 – Silhouettes

All of the patterns in the red and aquamarine cluster lie well above 0.5. In the green cluster, the bar just above 0.25 corresponds to pattern s3+c1 while the bar slightly below 0.5 corresponds to patterns s3+c3 and s3. All three of these patterns along with pattern s4+3, are present in the second of the potential clusters that could be produced by manually lowering the cut-level.

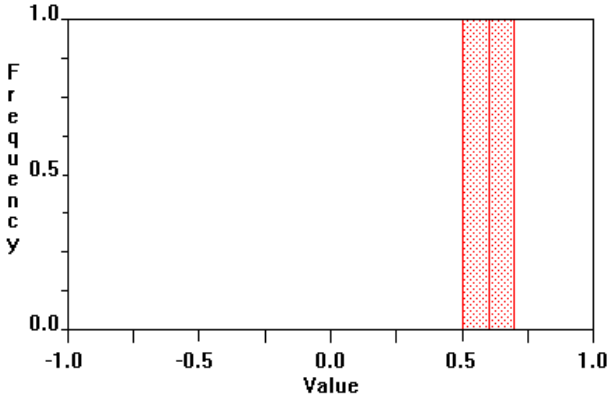

Figure 117 - Fuzzy Clustering

	1	2	3	4	
c1+3	0.24	0.96	0.27	0.54*	<==
c1	0.21	0.96	0.25	0.52*	<==
c3	0.89	0.34	0.27	0.51*	<==
s2+3+4	0.22	0.33	0.32	1	
s2+c1	0.22	0.96	0.31	0.63*	<==
s2+c3	0.25	0.36	0.93	0.67*	<==
s2	0.22	0.34	0.93	0.67*	<==
s3+2	0.22	0.32	0.31	1	
s3+c1	0.21	0.36	0.27	0.99	
s3+c3	0.21	0.28	0.25	1	
s3	0.18	0.25	0.24	1	
s4+2	0.21	0.31	0.3	1	
s4+3	0.21	0.28	0.27	1	
s4+c1	0.22	0.36	0.3	1	
s4+c3	0.22	0.3	0.29	1	
s4	0.2	0.28	0.27	1	

Table 15 - Fuzzy Clustering Numerical Values

Cluster 1 represents the yellow cluster, cluster 2 the red cluster, cluster 3 the aquamarine cluster and cluster 4 the green cluster.

The first bar in the fuzzy clustering plot, slightly above 0.5, contains patterns c1+3, c1 and c3. These patterns could all be potentially be grouped as follows:

- C1+3 could potentially be in either the red or green cluster
- C1 could potentially be in either the red or green cluster
- C3 could potentially be in either the yellow or green cluster.

The second bar corresponds to patterns s2+c1, s2+c3 and s2. These patterns could all potentially be clustered as follows:

- S2+c1 could potentially be in the red or green cluster
- S2+c3 could potentially be in the aquamarine or green cluster
- S2 could potentially be in the aquamarine or green cluster.

Due to the silhouette calculations, the cut-level was adjusted to split the green cluster into two separate clusters. The resulting dendrogram and MMDS for this are shown in Figure 118.

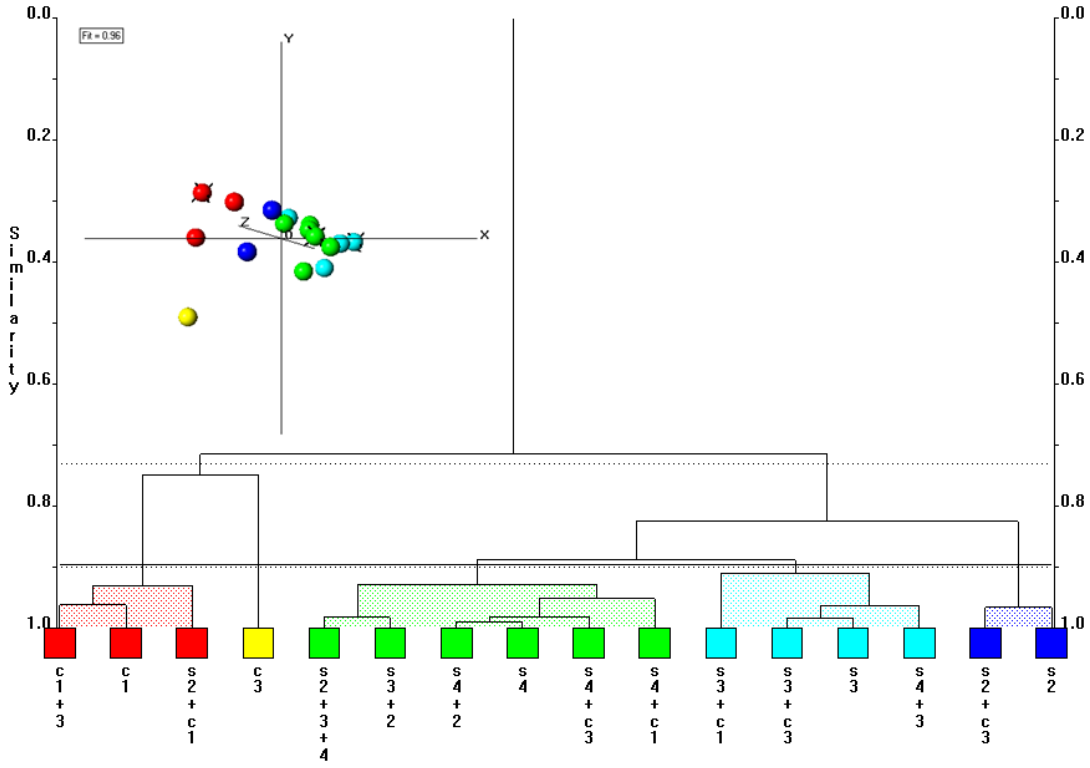

Figure 118 – Dendrogram and MMDS Plot for Expected Clustering using Simulated with Cut-level Adjusted

The simulated dataset had the following clusters present:

1) c1+3, c1, s4+c1, s2+c1 and s3+c1
2) s2+3+4, s3+2, s2, s2+c3
3) s3+c3, s3, s4+3
4) s4+2, s4+c3, s4
5) c3

The expected clustering dataset has the following clustering:

1) c1+3, c1 and s2+c1
2) s2+3+4, s3+2, s4+2, s4, s4+c3 and s4+c1
3) s3+c1, s3+c3, s3 and s4+3
4) s2+c3 and s2
5) c3

By comparing the predicted and measured dataset, it can be seen that four samples have moved to different cluster. As the simulated clustering gives a better match to the Pearson correlation coefficient, the expected clustering from the simulated dataset will be treated as optimal. As such the expected clustering is as follows

1) A cluster containing samples c1+3, c1, s4+c1, s2+c1 and s3+c1
2) A cluster containing samples s2+3+4, s3+2, s2, s2+c3
3) A cluster containing samples s3+c3, s3, s4+3
4) A cluster containing samples s4+2, s4+c3, s4
5) A cluster containing sample c3

5.5.2 PXRD DATA

The PXRD data was run with no pre-processing applied to it. The dendrogram and MMDS plot are shown in Figure 119.

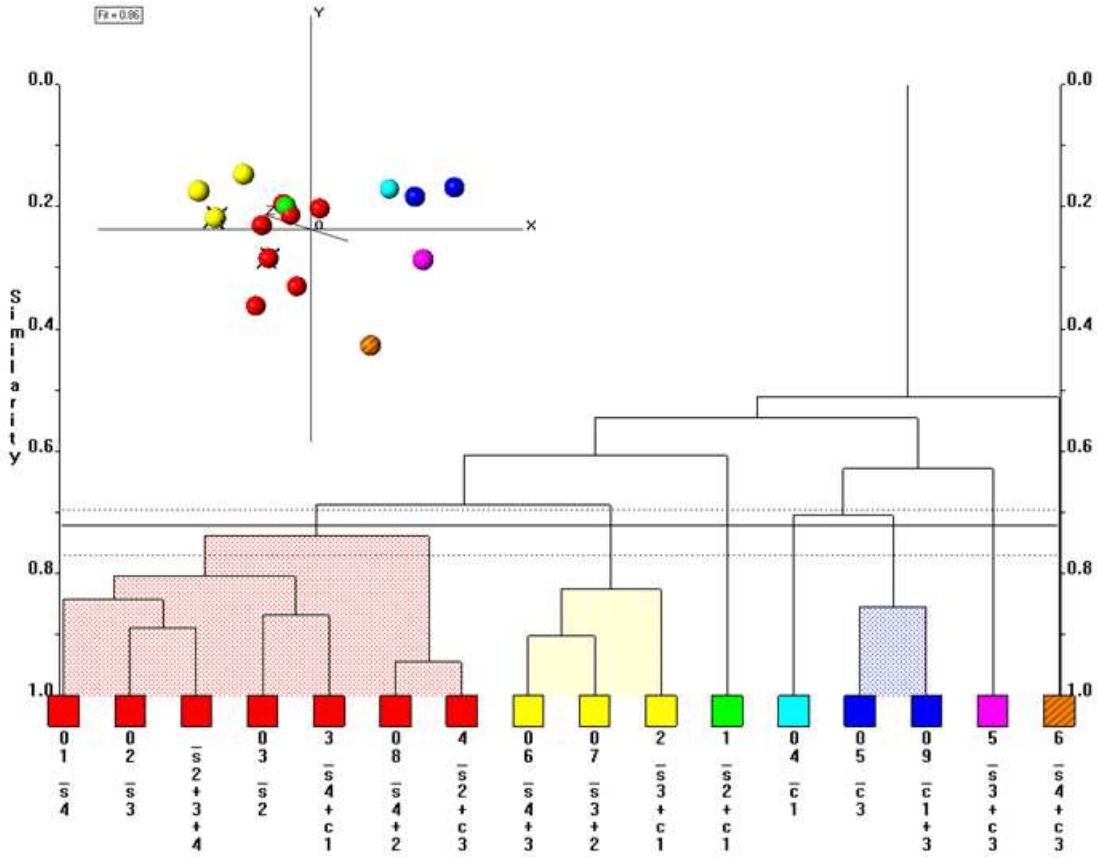

Figure 119 - Suthaz/Cbz PXRD Dendrogram and MMDS Plot

The PXRD data does not generate the expected clustering. The red cluster contains samples s2 and s2+c3, which form predicted cluster 2, as well as samples s4 and s4+2 which are part of expected cluster 4, s2+3+4 which is part of expected cluster 2, s4+c1 which is part of expected cluster 1 and s3 which is part of expected cluster 3. The yellow cluster contains sample s4+3 which is part of expected cluster 4, s3+c1 which is part of expected cluster 1 and sample s3+2 which is part of predicted cluster 2. The green cluster

contains sample s2+c1 and the aquamarine cluster contains sample c1 which are both part of predicted cluster 1. The blue cluster contains the sample c1+3, which is part of predicted cluster 1, as well as sample c3 which would be expected to be clustered separately from all other samples. The purple cluster contains sample s3+c3, part of predicted cluster 3 and the striped brown cluster contains sample s4+c3 which is part of predicted cluster 4. This dataset has a score of 0.56. As this equates to just over half of the samples being misclustered the dataset is not currently giving good results.

5.5.3 RAMAN DATA

The Raman data dendrogram and associated MMDS plot are shown in Figure 120. The dataset did not have the area beyond 1750cm^{-1} included when matching the spectra.

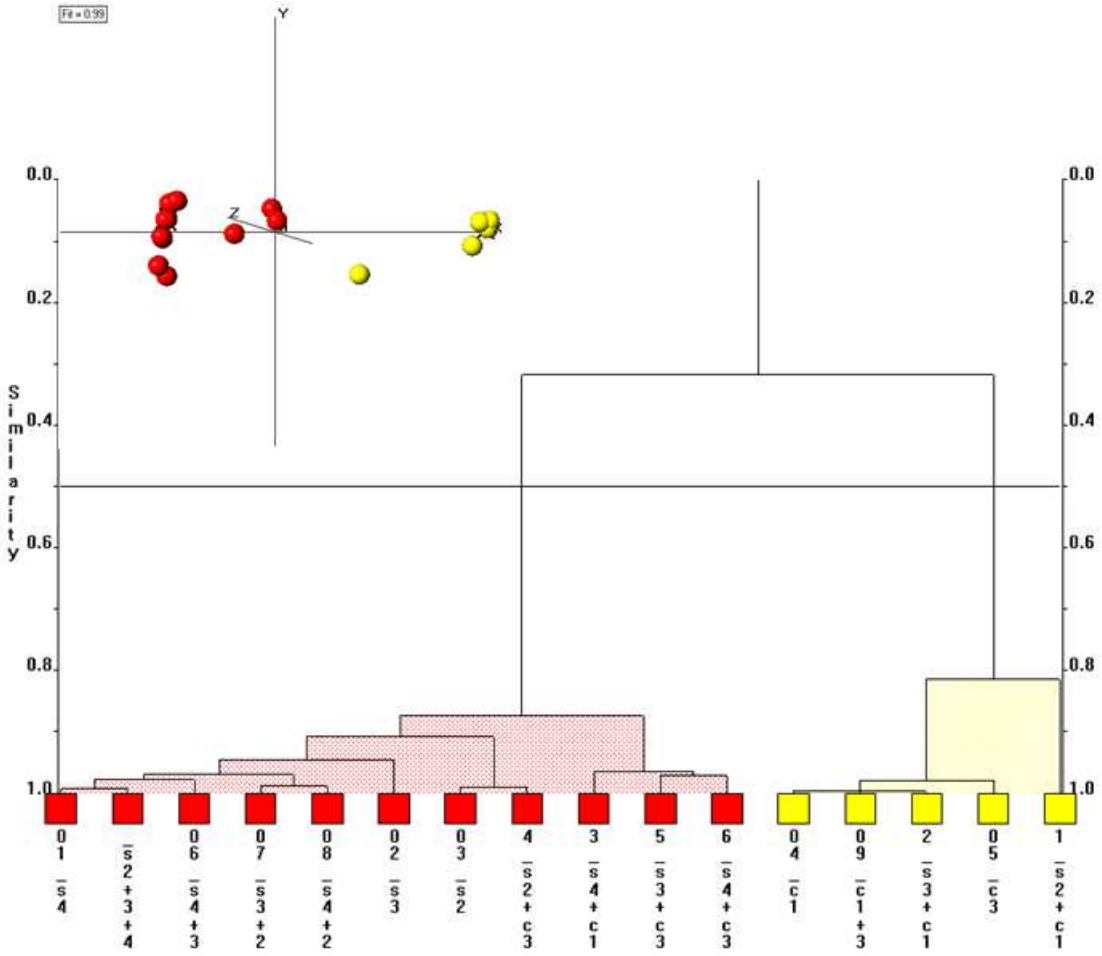

Figure 120 – Suthaz/Cbz Raman Dendrogram and MMDS Plot

The Raman dataset does not give the expected clustering. The red cluster contains samples s2+3+4, s3+2, s2 and s2+c3 which make up predicted cluster 2, s3+c3, s3 and s4+3 which

make up expected cluster 3, s4+2, s4+c3 and s4 which make up expected cluster 4 as well as sample s4+c1 which makes up expected cluster 1. The yellow cluster contains samples c1+3, c1, s2+c1 and s3+c1 which make up expected cluster 1 as well as sample c3 from expected cluster 5. The Raman dendrogram has a score of 0.5, equating to half of the samples being misclustered. This is a small improvement over the PXRD dataset.

5.5.4 DSC DATA

The DSC data dendrogram and MMDS plot are shown in Figure 121.

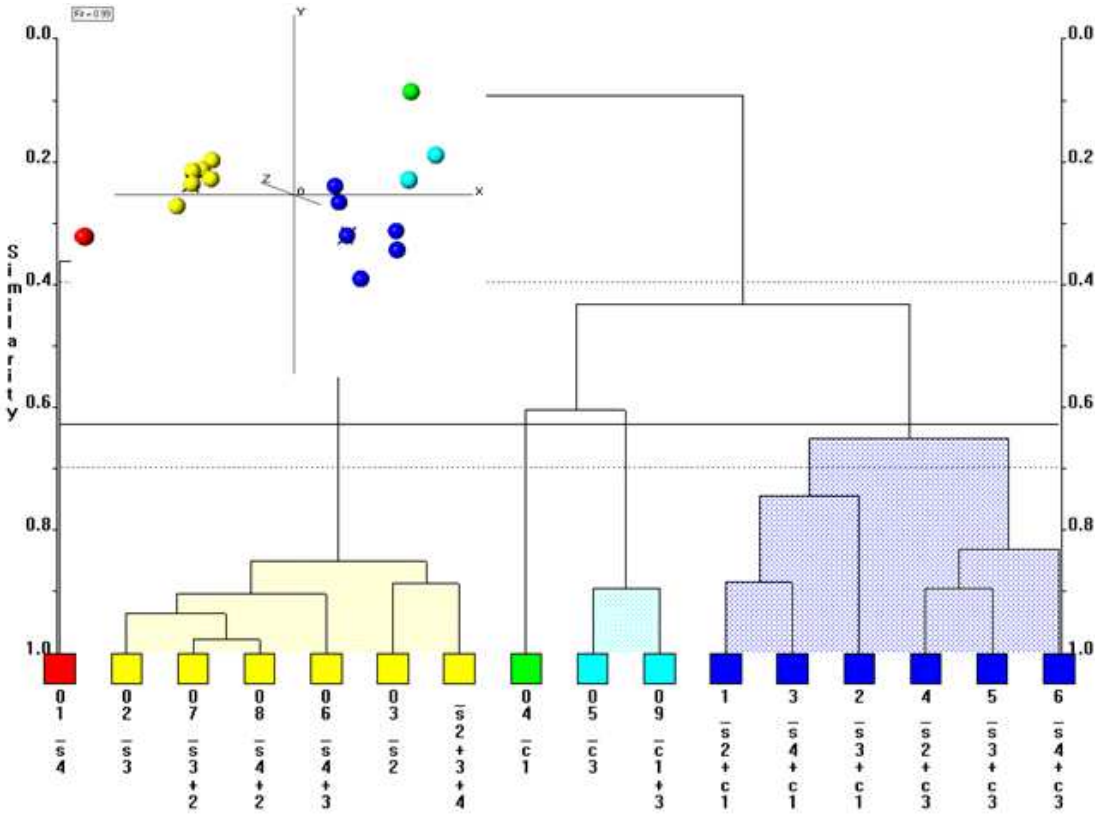

Figure 121 –Suthaz/Cbz DSC Dendrogram

The DSC data does not closely match the expected clustering. The red cluster contains sample s4 which is part of expected cluster 4. The yellow cluster contains samples s3+2, s2+3+4 and s2 which are part of predicted cluster 2, s4+2 which is part of expected cluster 3 and as well as samples s3 and s4+3 which are part of expected cluster 2. The green cluster contains sample c1 which was in predicted cluster 1. The aquamarine cluster contains samples c3 which is predicted to be in cluster 5 on its own and c1+3 which is predicted to be in cluster 1. The blue cluster contains samples s2+c1, s4+c1 and s3+c1 which were predicted to be in cluster 1 and s3+c3 which were predicted to be in cluster 3,

samples s4+c3 which are predicted to be in cluster 4 and sample s2+c3 which was predicted to be in cluster 2. The DSC data has a score of 0.5, again equating to half of the samples being misclustered.

It was discovered when examining the sulfathiazole form 4 sample that the peaks are all significantly smaller than in other DSC samples in this dataset. This is shown in Figure 122.

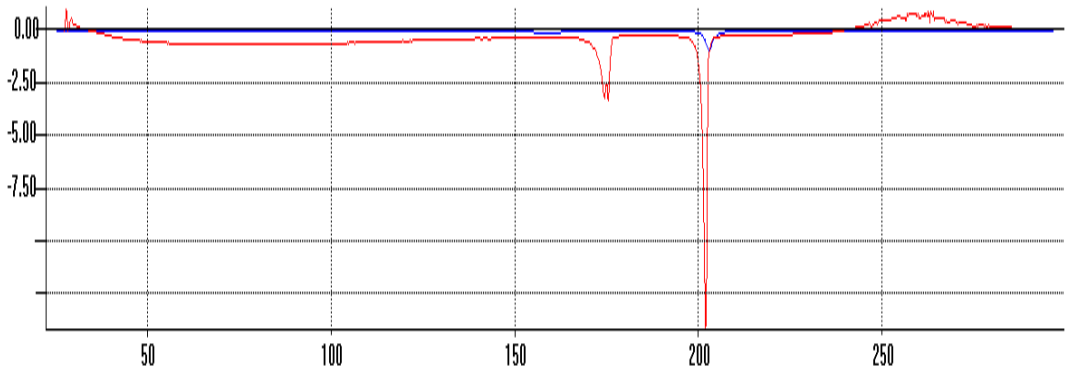

Figure 122 - Overlay of Sulfathiazole Form 3 (blue) and Sulfathiazole Form 4 (red) DSC

As can be seen, the form 3 pattern has larger and wider peaks than the form 4 pattern. The problem will lie in the extra width present as this will result in large correlation differences between the two patterns. This can occur when either a small amount of sample, or a slow heating rate, are used. For this dataset all samples are heated at a rate of 10°C a minute so the problem lay with too small a sample being collected. A small number of other samples also had the same issue. All of these samples were recollected and the dataset re-run.

Two additional features visible in these spectra are the large area in the blue spectra above 250Cº which corresponds to a degradation of the material and the initial drop at the beginning of the blue spectra which is not present in the red spectra. This initial drop is due to the instrument beginning at a slightly lower temperature than that specified so preheating itself to the correct temperature. The dendrogram and MMDS plot for this re-run are shown in Figure 123.

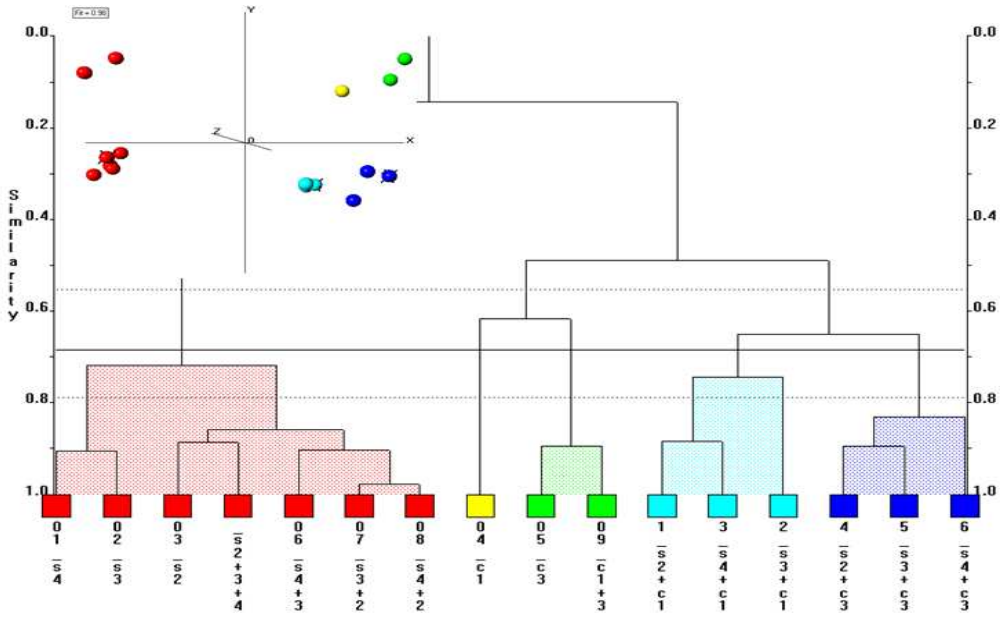

Figure 123 - Suthaz/Cbz re-run DSC Dendrogram and MMDS Plot

The red cluster contains all the pure sulfathiazole materials and sulfathiazole only mixtures. The yellow and green clusters contain all the carbamazepine samples and the carbamazepine only mixture. The aquamarine and blue clusters contain the sulfathiazole/carbamazepine mixtures. The expected clustering is not seen here. The dataset now has a score of 0.69.

5.5.5 IR DATA

The IR data is shown in Figure 124.

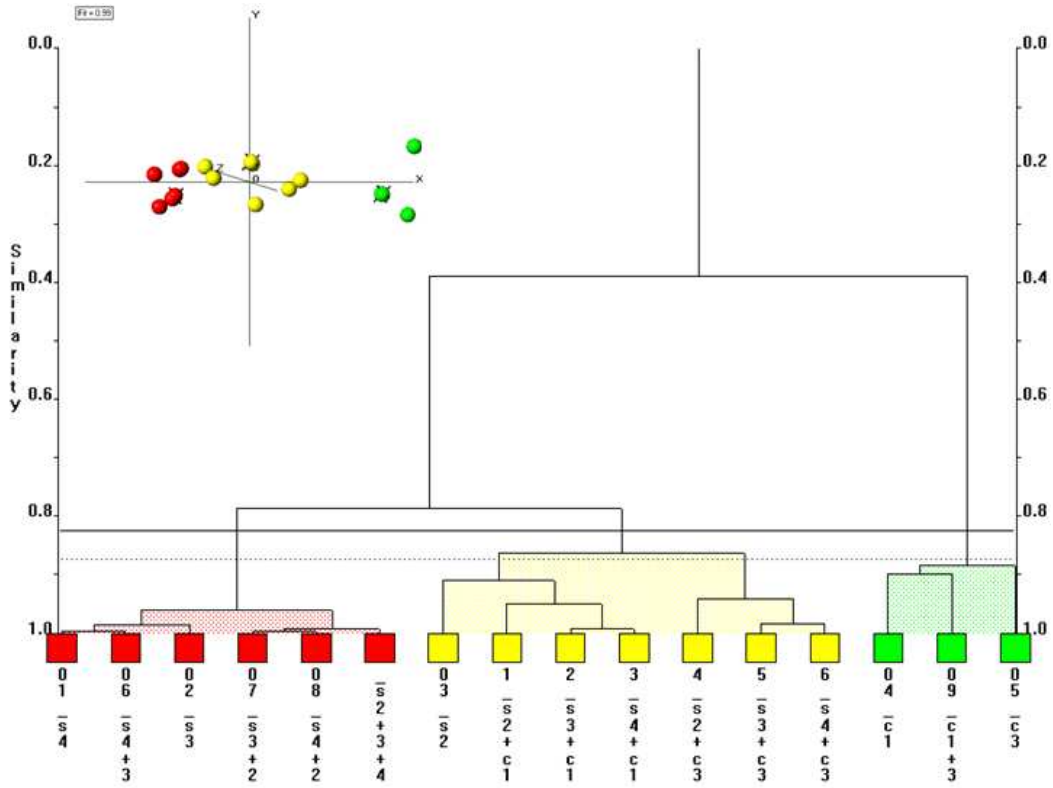

Figure 124 - Suthaz/Cbz IR Dendrogram and MMDS Plot

The IR data does not give the expected clustering. The red cluster contains samples s3, and s4+3, which are part of predicted cluster 4, samples s2+3+4 and s3+2 which are part of predicted cluster 2 and samples s4+2 and s4 which are part of expected cluster 4. The yellow cluster contains samples s2 and s2+c3 which are part of expected cluster 2, as well as samples s4+c1, s2+c1 and s3+c1, which are part of expected cluster 1, s4+c3, part of expected cluster 4 and s3+c3 which is part of expected cluster 3. The green cluster contains samples c1 and c1+3 which are part of expected cluster 1 as well as sample c3 which would be expected to be clustered on its own.

The MMDS plot shows three distinct clusters. The dataset has a score of 0.5, again equating to eight samples being misclustered.

5.5.6 COMBINED DATA

The combined dataset, combining all four data types using INDSCAL, is shown in Figure 125.

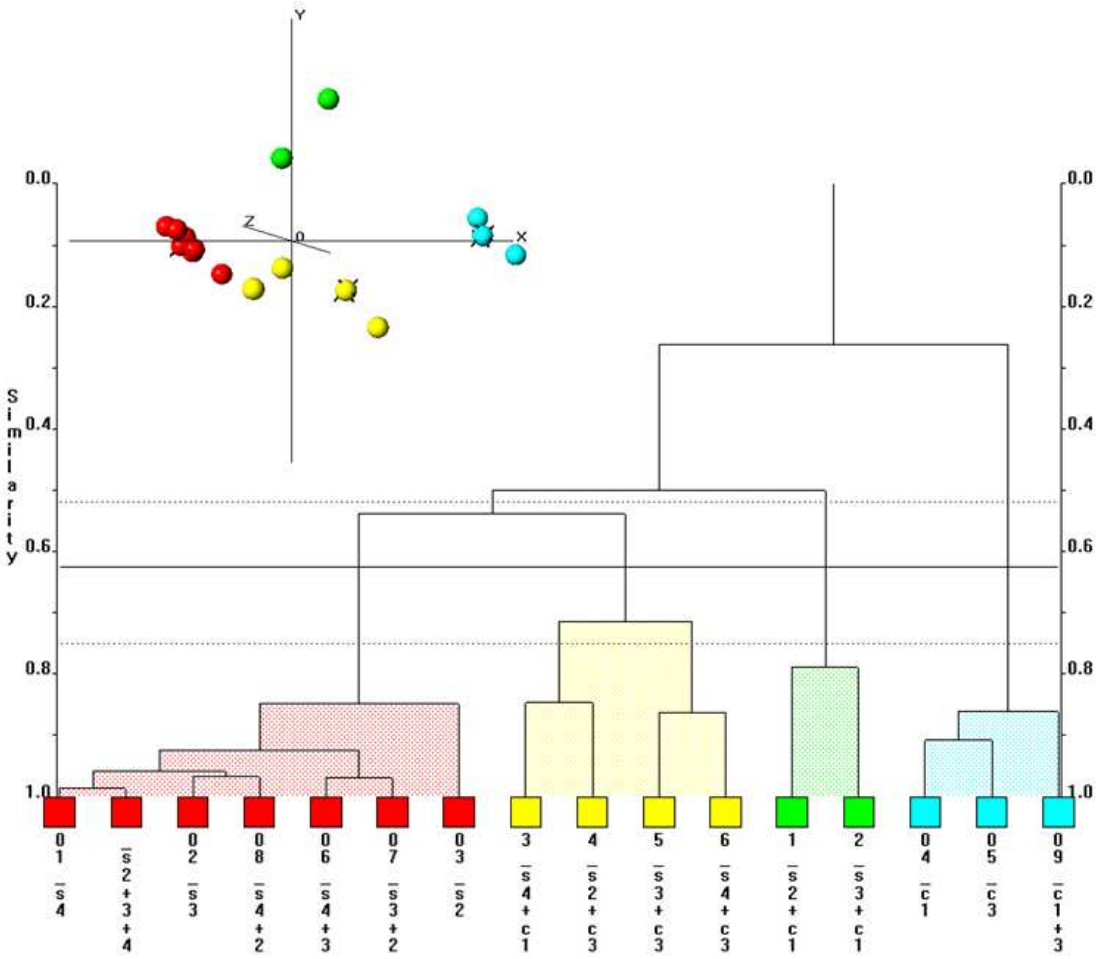

Figure 125 – Combined Suthaz/Cbz Dataset Dendrogram and MMDS Plot

The red cluster contains samples s2+3+4 and s3+2 which are part of expected cluster 2 and s4+2 and s4, which are part of predicted cluster 4. The yellow cluster contains samples s4+c1 which is part of expected cluster 1, s4+c3 which are part of predicted cluster 4, s2+c3 which is part of predicted cluster 2 and s3+c3 which is part of predicted cluster 3. The green cluster contains samples s2+c1 and s3+c1 which are part of predicted cluster 1. The aquamarine cluster contains samples c1 and c1+3 which are part of predicted cluster 1 and c3 which was predicted to be clustered on its own.

The MMDS plot shows a clearly separated red and aquamarine cluster. The yellow cluster is rather diffuse, as are the two outlying samples in the green cluster. One of those samples, s2+c1, lies a lot closer to the yellow cluster than the other, however both lie too far away to

be reasonably included. The dataset has a score of 0.56, equating to just over half the samples being misclustered.

5.6 DERIVATIVE DATA

The Raman dataset was re-run using both first and second derivative processing. The dendrogram and MMDS plot that results from applying a first derivative to the data are shown in Figure 126.

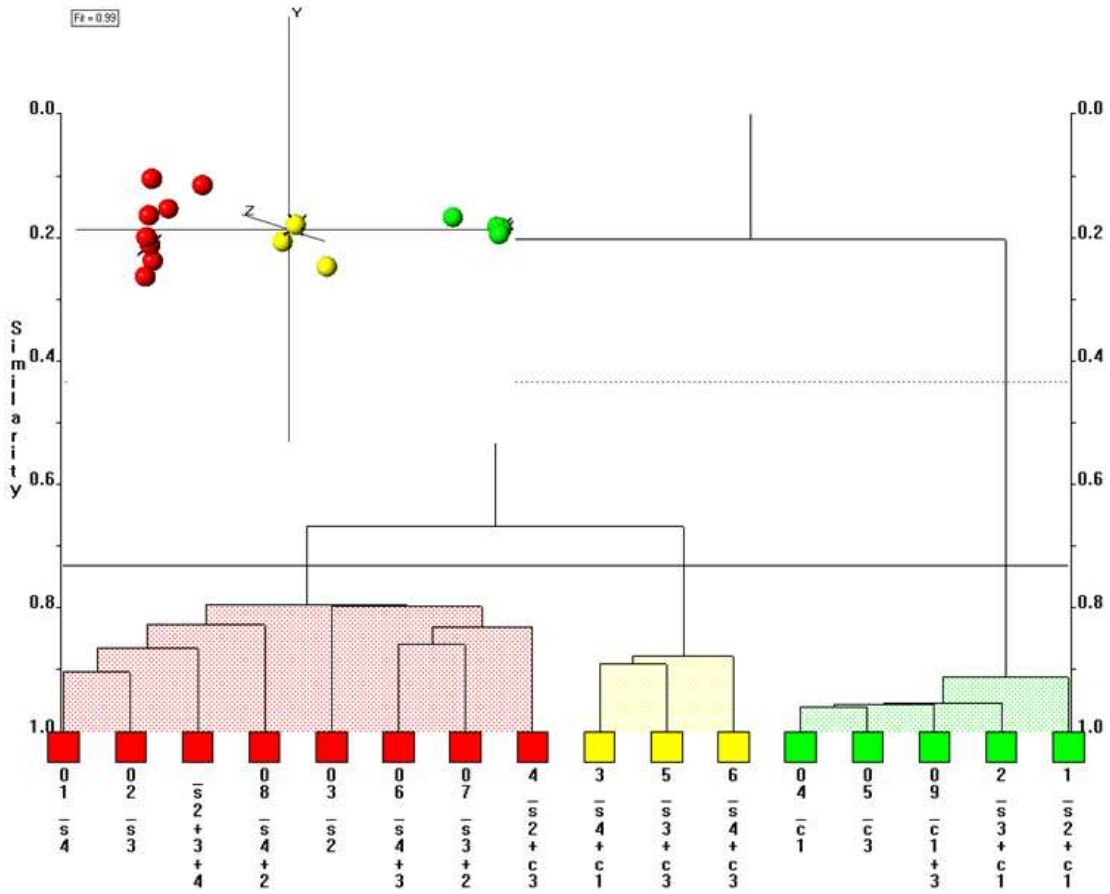

Figure 126 - First Derivative Raman Dendrogram and MMDS Plot

The red cluster contains samples s2+3+4, s3+2, s2 and s2+c3 which are part of expected cluster 2 samples s4 and s4+2, which were part of predicted cluster 4 and samples s3 and s4+3 which are part of expected cluster 3. The yellow cluster contains samples s4+c1 which is part of expected cluster 1, s4+c3 which are part of expected cluster 4 and sample s3+c3 which is part of expected cluster 2. The green cluster contains samples s2+c1, s3+c1, c1 and c1+3 which make up part of expected cluster 1 and sample c3 which makes up expected cluster 5.

The MMDS plot shows three clearly defined clusters. The red cluster is rather diffuse but is clearly separated from the other two clusters. The dendrogram has a score of 0.56.

The dendrogram and MMDS plot for the second derivative Raman data are shown in Figure 127.

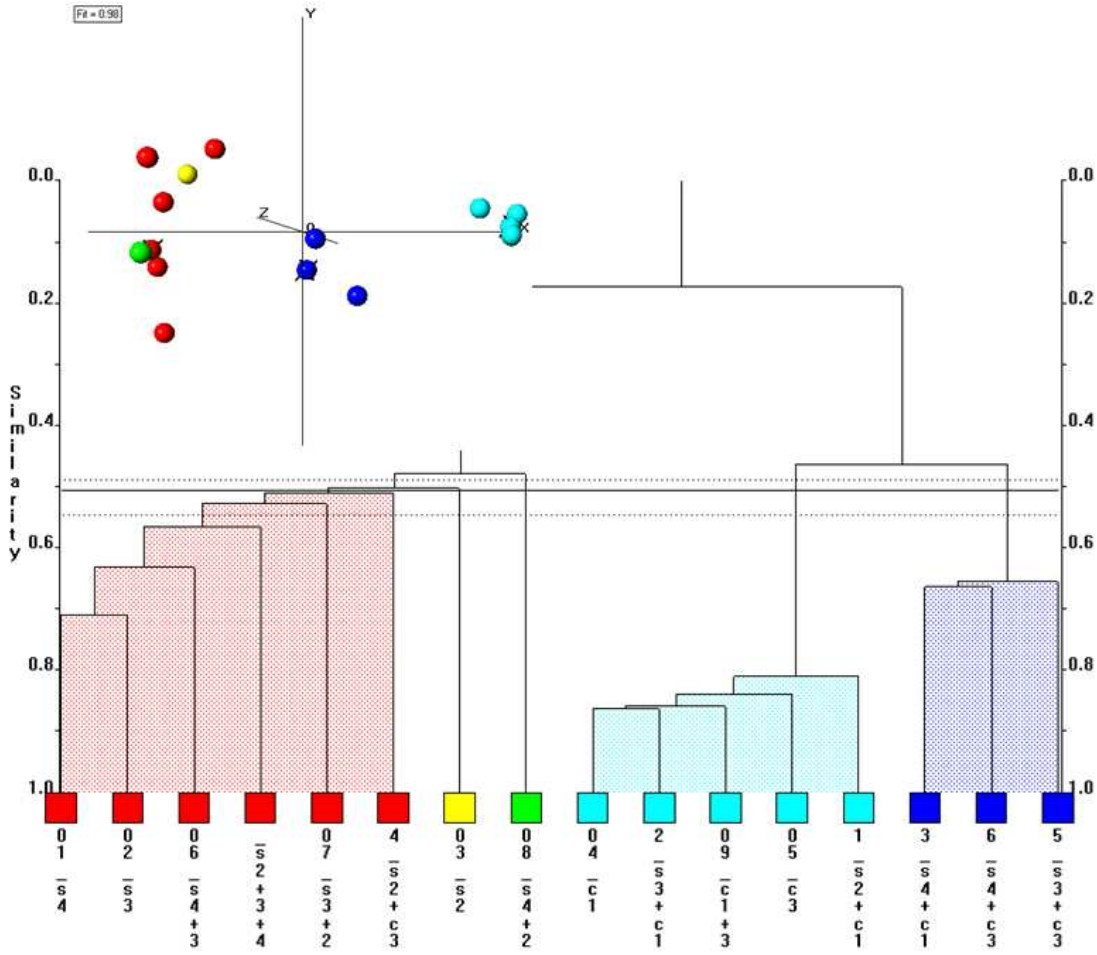

Figure 127 - Second Derivative Raman Dendrogram and MMDS Plot

The red cluster contains sample s4 which is part of expected cluster 4, s2+3+4, s3+2 and s2+c3 which are expected to be in cluster 2 and s3 and s4+3 which are expected to be in cluster 3. The yellow cluster contains s2 which is expected to be in cluster 2. The green cluster contains sample s4+2 which is expected to be in cluster 4. The aquamarine cluster contains samples c1, c1+3, s2+c1 and s3+c1which make up part of expected cluster 1 and c3 which is the lone sample in expected cluster 5. The blue cluster contains sample s4+c1 which is expected to be in cluster 1, sample s4+c3 which are expected to be in cluster 4 and sample s3+c3 which is expected to be in cluster 3. In the MMDS plot, the yellow and green samples appear to be intermixed with the red cluster, showing poor separation of the

clusters. The blue and aqua clusters are more clearly separated. The dendrogram has a score of 0.44.

The IR data was also re-run with both a first and second derivative applied. The dendrogram and MMDS plot for the first derivative run are shown in Figure 128.

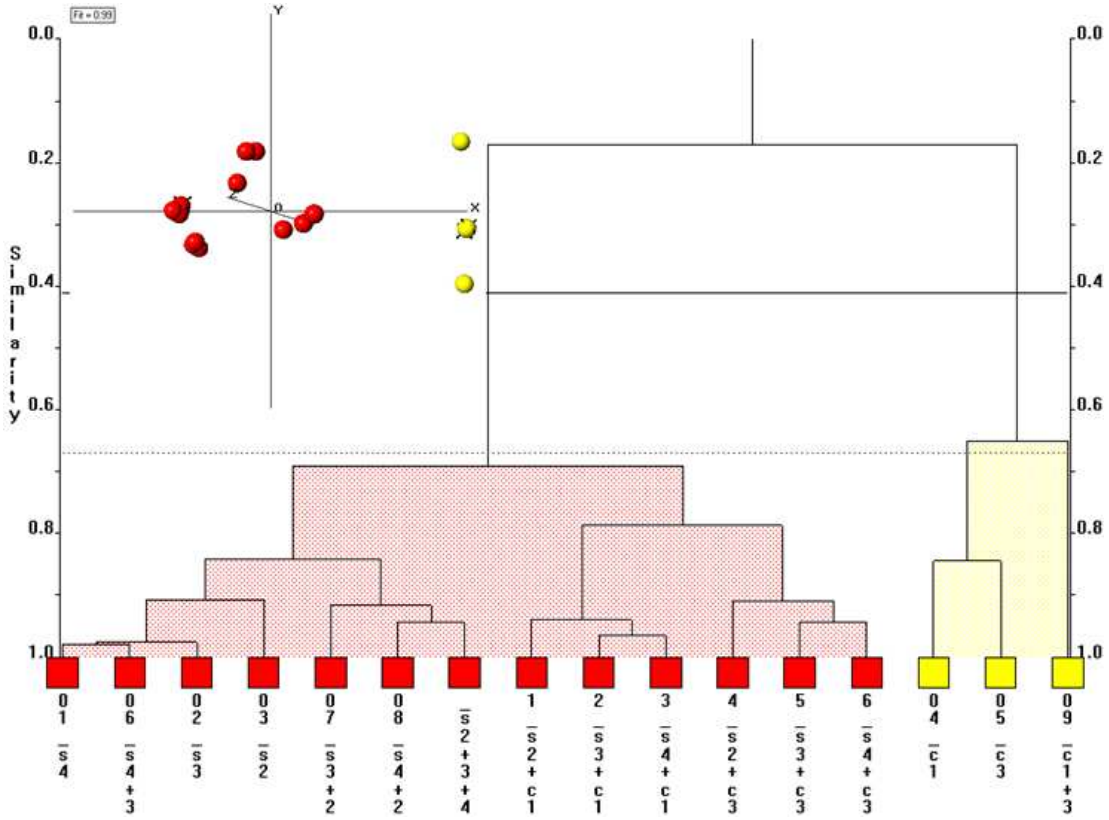

Figure 128 - First Derivative IR Dendrogram

The first derivative IR run is not as well clustered as the unprocessed Raman run. The red cluster contains samples s2+3+4, s3+2, s2+c3 and s2 which are expected to be in cluster 2, s4+2, s4 and s4+c3 which are expected to be in cluster 4, s4+c1, s2+c1 and s3+c1 which is expected to be in cluster 1 and samples s3, s3+c3 and s4+3 which were all expected to be in cluster 3. The yellow cluster contains samples c1 and c1+3 which are part of expected cluster 1 and c3 which is the lone sample in expected cluster 5. The MMDS plot shows that the red cluster is very diffuse. The six red samples are the six sulfathiazole/carbamazepine mixtures. The remaining red samples clustered towards the left of the plot are the pure sulfathiazole and sulfathiazole only mixtures. The dendrogram has a score of 0.8.

Figure 129 shows the second derivative dendrogram and MMDS plot.

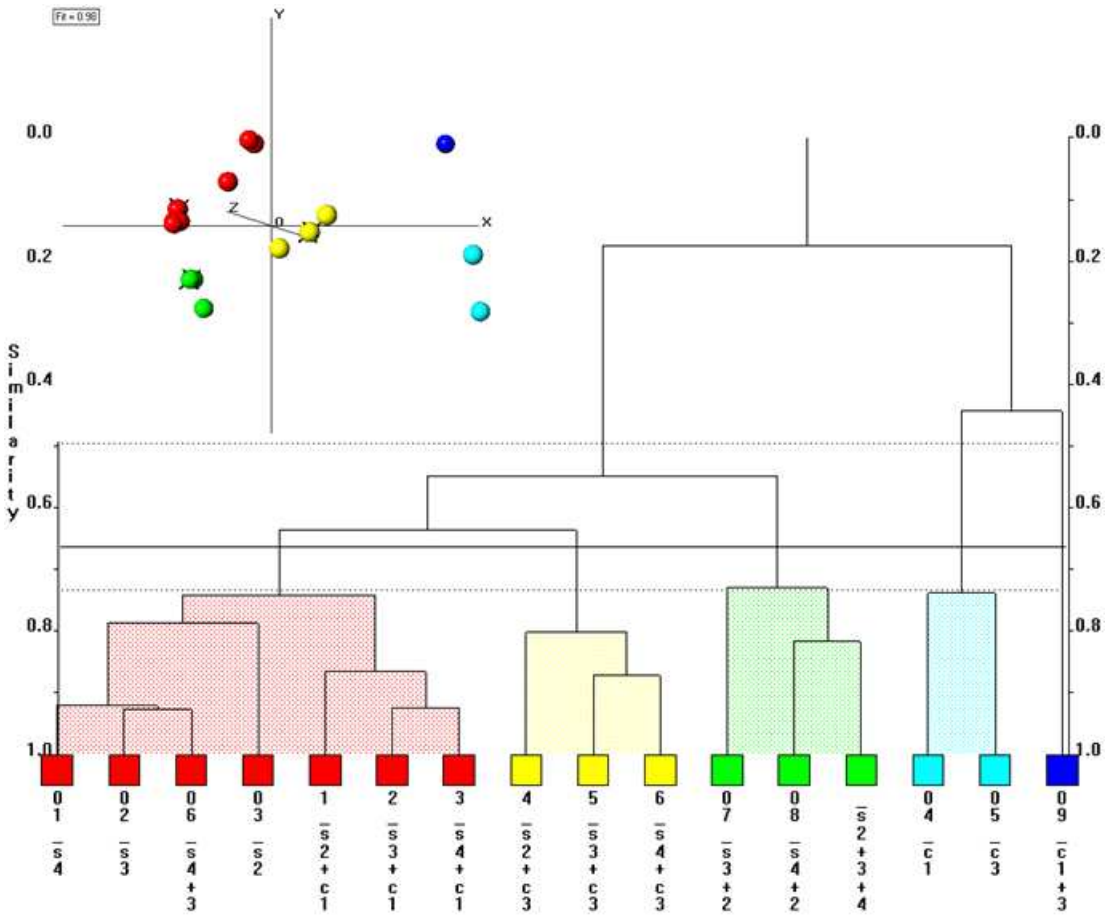

Figure 129 - Second Derivative IR Dendrogram and MMDS Plot

The red cluster contains samples s4 from expected cluster 4, s4+c1, s2+c1 and s3+c1 from expected cluster 1, s3 and s4+3 from expected cluster 3 and s2 from expected cluster 2. The yellow cluster contains samples s2+c3 from expected cluster 2, s3+c3 from expected cluster 3 and s4+c3 from expected cluster 4. The green cluster contains samples s3+2 from expected cluster 1, s4+2 from expected cluster 4 and s2+3+4 from expected cluster 3. The aquamarine cluster contains sample c1 from expected cluster 1 and c3 from expected cluster 5. The blue cluster contains sample c1+3 from expected cluster 1.

The MMDS plot is very similar in appearance to the plot from the first derivative run. The dendrogram still has a score of 0.8.

5.7 TGA DATA

TGA data was collected for some of the samples in this dataset. Three samples, s3+c1, sulfathiazole form 3 and s3+c3 and c3, are missing from the dataset. These samples were not collected due to a fault in the machine preventing further data being collected. As there are only thirteen samples for this dataset, the scores would appear different for what they would be for a sixteen sample dataset. As such the scores for the TGA will be scaled as if they were from a sixteen sample dataset so that they can be better compared to the scores for the other datatypes. The new dataset is summarised in Table 16.

Sample Number	Sample ID	Name in PolySNAP	Composition
1	Sulfathiazole Form 4	s4	
2	Sulfathiazole Form 3	s3	
3	Sulfathiazole Form 2	s2	
4	Carbamazepine Form 1	c1	
5	Carbamazepine Form 3	c3	
6	Sulfathiazole Forms 3 and 4	s4+3	58:42
7	Sulfathiazole Forms 2 and 3	s3+2	63:37
8	Sulfathiazole Forms 2 and 4	s4+2	32:68
9	Carbamazepine Forms 1 and 3	c1+3	72:28
10	Sulfathiazole Forms 2, 3 and 4	s2+3+4	53:18:29
11	Sulfathiazole Form 2 and Carbamazepine Form 1	s2+c1	50:50
12	Sulfathiazole Form 4 and Carbamazepine Form 1	s4+c1	61:39
13	Sulfathiazole Form 2 and Carbamazepine Form 3	s2+c3	80:20

Table 16 – TGA Dataset

The Dendrogram and MMDS plot for this run are shown in Figure 130.

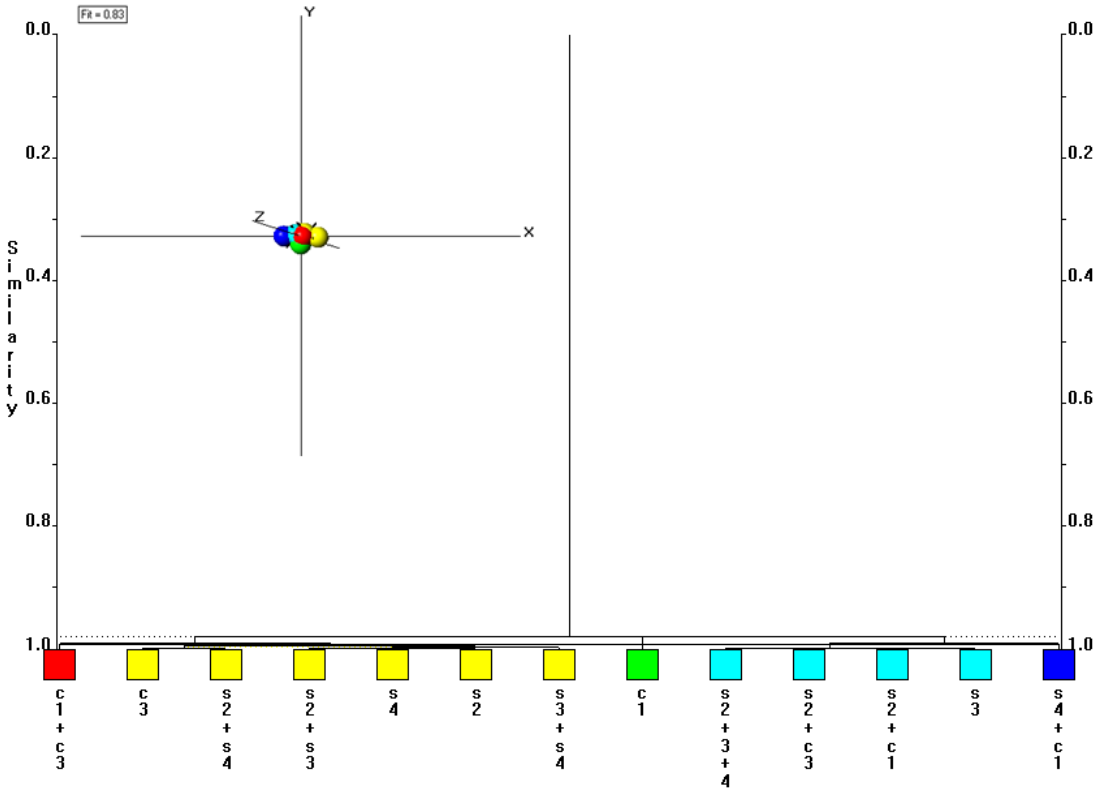

Figure 130 - TGA Dendrogram and MMDS Plot

The red cluster contains sample c1+3 which was predicted to be in cluster 1. The yellow cluster contains samples s2 and s3+2 which are predicted to be in cluster 2, s3+4 from expected cluster 3 and samples s4+2 and s4 from expected cluster 4 and c3 from expected cluster 5. The green cluster contains carbamazepine form 1 which is predicted to be in cluster 1. The aquamarine cluster contains s2+3+4 which is predicted to be in cluster 2, sample s3 which is predicted to be in cluster 3, sample s2+c3 which is predicted to be in cluster 2 and s2+c1 which is predicted to be in cluster 1. The blue cluster contains sample s4+c1 which are predicted to be in cluster 1.

The dataset shows all samples to have a very high similarity, even more so than was seen in the Raman dataset. The clustering is not as expected. The dendrogram has a score of 0.76. This is higher than that seen for any preceding data type in this dataset.

The dataset was re-run with both first and second derivatives applied, however this did not affect the clustering. The problem with distinguishing between TGA patterns lies in them all consisting of a single downwards slope. The temperature at which this slope begins varies between samples; however this variation is not large enough to allow samples to be clearly distinguished. Figure 131 shows an overlay of the three pure sulfathiazole samples TGA patterns.

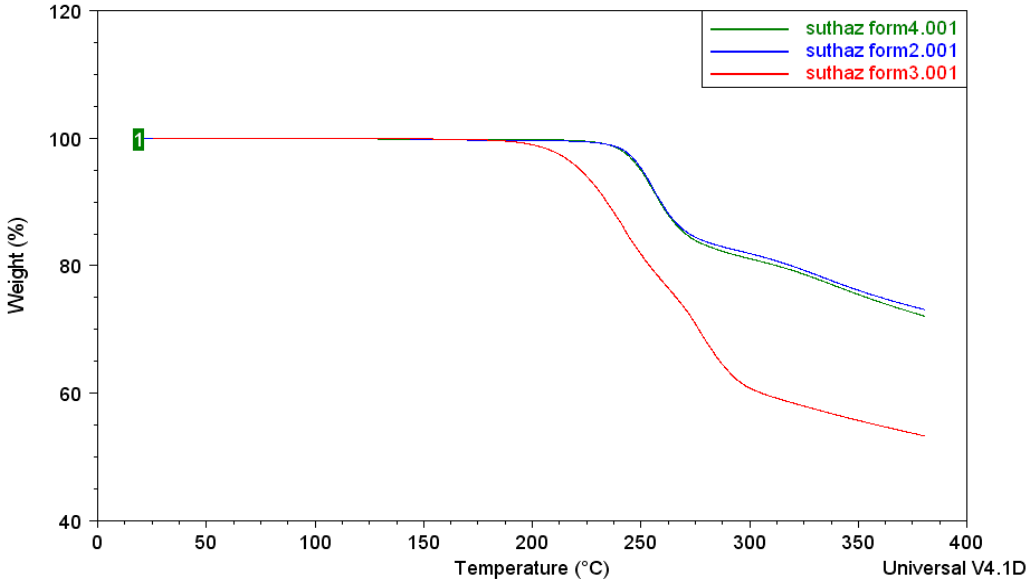

Figure 131 – Sulfathiazole TGA Pattern Overlay

The sulfathiazole form 2 and 4 patterns are clearly similar. PolySNAP reports a 99.8% correlation between these two patterns. The sulfathiazole form 3 pattern looks to be substantially different from the remaining, however this still equates to a 98.8% correlation between the sulfathiazole form 2 and sulfathiazole form 3 patterns and a 98.9% correlation between the sulfathiazole form 4 and sulfathiazole form 3 patterns. An overlay of sulfathiazole form 4 and carbamazepine form 1 is shown in Figure 132.

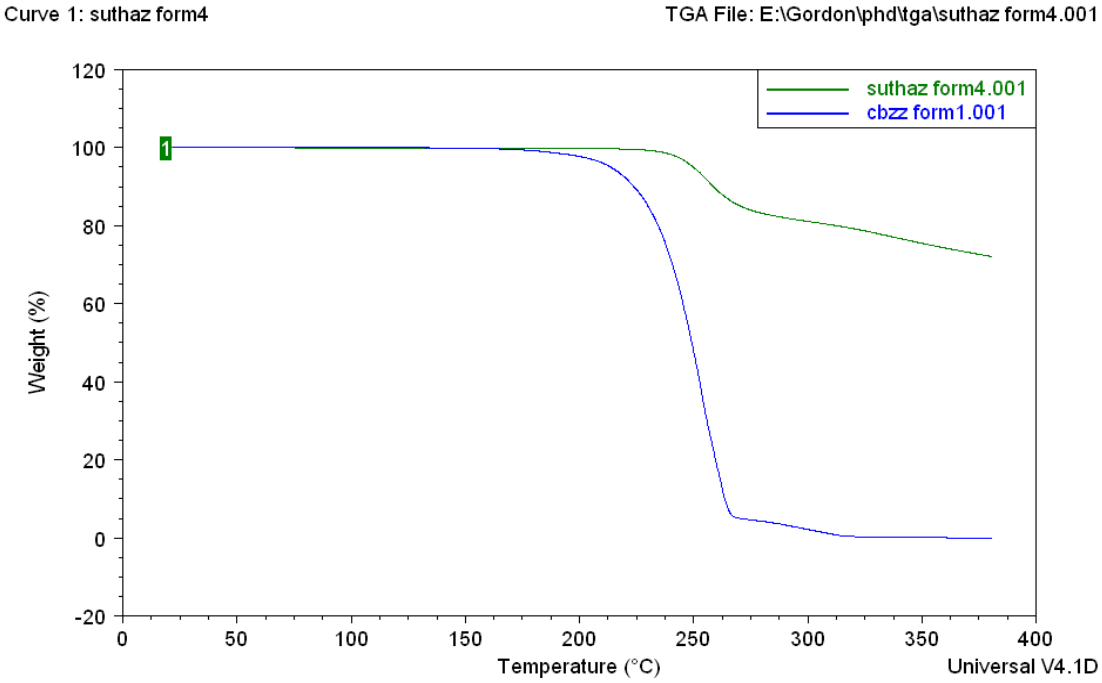

Figure 132 - Sulfathiazole and Carbamazepine TGA Pattern Overlay

These patterns appear to be highly different; however they still have a 97.5% similarity according to PolySNAP.

Figures 130 and 131 both show that the TGA pattern have no useful data below 175°C. The dataset was re-run without the data in this area to see if this improves the clustering. The resulting dendrogram and MMDS plot are shown in Figure 133.

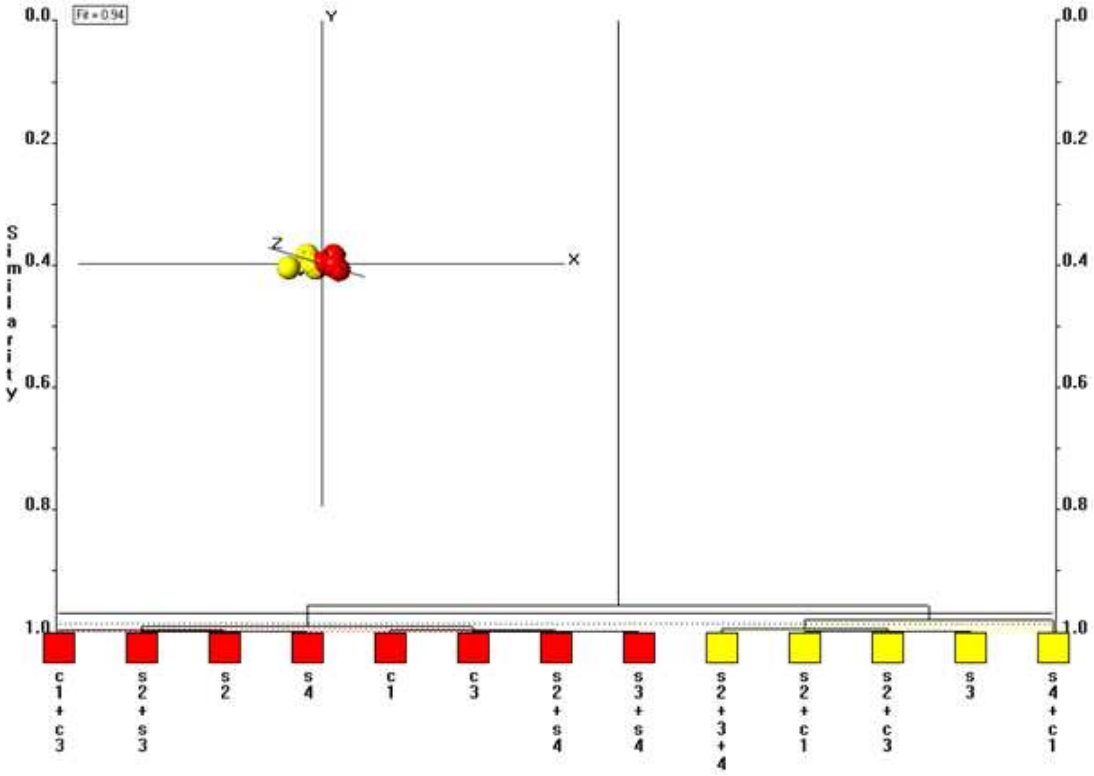

Figure 133 - TGA Dendrogram and MMDS Plot with Cut-off at 175°C

Re-running the dataset with a cut-off does not improve the clustering. The red cluster now contains samples s4 and s4+2 from expected cluster 4, s3+2 and s2 from predicted cluster 2, c1 and c1+3 from expected cluster 1, c3 from expected cluster 5 and s3+4 from expected cluster 3. The yellow cluster contains samples s2+3+4 from expected cluster 2, s4+c1 and s2+c1 from expected cluster 1, s2+c3 from expected cluster 2 and s3 from expected cluster 5. The dendrogram now has as score of 0.66, still higher than that seen for any data type studied so far for this dataset.

The PolySNAP clustering techniques are not useful for this type of data.

5.8 FLOWCHART

All of the possible combinations of pre-processing were applied to the PXRD dataset and the results compared to the optimal clustering. The scores for these datasets are shown in Table 17.

	Score
no pre-processing	0.56
denoise	0.56
background	0.81
background and denoise	0.81

Table 17 – Misplaced samples

As can be seen the no pre-processing and denoising only datasets give similar results. The background removal and background removal and denoise also give similar results. These results match that seen in section 4.4. The flowchart for this dataset is shown in Figure 134.

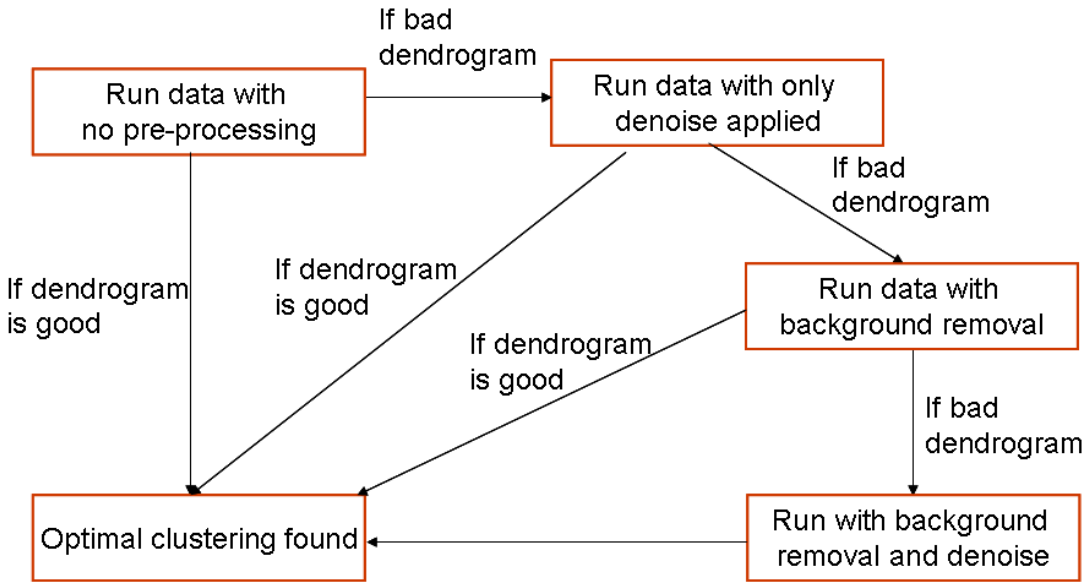

Figure 134 - Flowchart for optimal clustering determination

This flowchart is unchanged from that shown in section 4.4. The classification of a dendrogram as ‘good’ or ‘bad’ is also unchanged from this dataset.

5.9 QUANTITATIVE ANALYSIS

The materials were compared using the PolySNAP manual analysis mode. The results of this are shown in Table 18. The data was compared using the SVD method.

The results, with pre-processing applied to the data, are shown in Table 19 and Table 20.

PXRD	Samples	Actual	Predicted	Difference	Raman	Samples	Actual	Predicted	Difference	IR	Samples	Actual	Predicted	Difference
	s2+3	63:37	32.1:67.9	4.9		s2+3	63:37	52.2:47.8	10.8		s2+3	63:37	68.1:31.9	5.1
	s2+4	32:68	8.5:91.5	23.5		s2+4	32:68	23.8:76.2	8.2		s2+4	32:68	16.5:83.5	15.5
	s3+4	58:42	29.8:70.2	28.2		s3+4	58:42	61.7:38.3	3.7		s3+4	58:42	64.5:35.5	6.5
	s2+3+4	53:18:29	14.1:83.8:2.0	46.9		s2+3+4	53:18:29	1.1:26.1:72.8	34.6		s2+3+4	53:18:29	15.2:14.1:70.7	28.1
	c1+3	28:72	27:73	1		c1+3	72:28	81.4:18.6	9.4		c1+3	72:28	29.6:70.4	42.4
	s2+c1	50:50	66.7:33.3	16.7		s2+c1	50:50	19.6:80.4	30.4		s2+c1	50:50	52.4:47.6	2.4
	s2+c3	80:20	83.2:16.8	3.2		s2+c3	80:20	98.4:1.6	18.4		s2+c3	80:20	90.2:9.8	10.2
	s3+c1	50:50	85.2:14.8	35.2		s3+c1	50:50	2.7:97.3	47.3		s3+c1	50:50	25.3:74.7	24.7
	s3+c3	83:17	5.9:94.1	11.1		s3+c3	83:17	64.5:35.5	18.5		s3+c3	83:17	68:32	15
	s4+c1	61:39	22:78	39		s4+c1	61:39	72.9:27.1	11.9		s4+c1	61:39	60.1:39.9	0.9
	s4+c3	82:18	69.8:30.2	12.2		s4+c3	82:18	55.9:44.1	26.1		s4+c3	82:18	92.9:7.1	10.9
		Mean absolute difference		20.17			Mean absolute difference		19.94			Mean absolute difference		14.70
		RMS difference		6.08			RMS difference		6.01			RMS difference		4.43
		Max absolute difference		26.73			Max absolute difference		27.36			Max absolute difference		27.70
		Min absolute difference		19.17			Min absolute difference		16.24			Min absolute difference		13.80

Table 18 – Data from Mixtures in Manual Analysis Mode

PXRD	Samples	Actual	Processed Predicted 1	Difference 1	Raman	Samples	Actual	Processed Predicted 1	Difference 1	IR	Samples	Actual	Processed Predicted 1	Difference 1
	s2+3	63:37	33.4:66.6	3.60		s2+3	63:37	49.7:50.3	13.30		s2+3	63:37	21.3:78.7	41.70
	s2+4	32:68	10.2:89.8	21.80		s2+4	32:68	23.8:76.2	8.20		s2+4	32:68	22:78	10.00
	s3+4	58:42	30.9:69.1	27.10		s3+4	58:42	57:43	1.00		s3+4	58:42	11.6:88.4	46.40
	s2+3+4	53:18:29	13.9:84.3:1.7	36.03		s2+3+4	53:18:29	18.7:44.5:36.8	22.87		s2+3+4	53:18:29	19.6:6.5:73.9	29.93
	c1+3	28:72	25.3:74.7	2.70		c1+3	28:72	34.9:65.1	37.10		c1+3	28:72	15.5:84.5	56.50
	s2+c1	50:50	74.5:25.5	24.50		s2+c1	50:50	17.7:82.3	32.30		s2+c1	50:50	17.7:82.3	32.30
	s2+c3	80:20	98.8:1.2	18.80		s2+c3	80:20	94.6:5.4	14.60		s2+c3	80:20	15.8:84.2	64.20
	s3+c1	50:50	76.2:23.8	36.20		s3+c1	50:50	6.6:93.4	43.40		s3+c1	50:50	77.8:22.2	27.80
	s3+c3	83:17	12.6:87.4	4.40		s3+c3	83:17	62.7:37.3	15.70		s3+c3	83:17	32:68	51.00
	s4+c1	61:39	53.4:46.6	7.60		s4+c1	61:39	67.3:32.7	6.30		s4+c1	61:39	79.6:20.4	18.60
	s4+c3	82:18	65.8:34.2	16.20		s4+c3	82:18	53.2:46.8	28.80		s4+c3	82:18	38.2:61.8	43.80
		Mean absolute difference		18.08			Mean absolute difference		20.32			Mean absolute difference		38.38
		RMS difference		5.45			RMS difference		6.13			RMS difference		11.57
		Max absolute difference		18.12			Max absolute difference		23.08			Max absolute difference		28.38
		Min absolute difference		15.38			Min absolute difference		19.32			Min absolute difference		25.82

Processed Predicted 1 - background remove and smoothed

Table 19 - Data from Mixtures in Manual Analysis Mode with Pre-processing Option 1

PXRD	Samples	Actual	Processed Predicted 2	Difference 2	Raman	Samples	Actual	Processed Predicted 2	Difference 2	IR	Samples	Actual	Processed Predicted 2	Difference 2
	s2+3	37:63	32.1:67.9	4.90		s2+3	63:37	52.5:47.5	10.50		s2+3	63:37	25.2:74.8	37.80
	s2+4	32:68	8.4:91.6	23.60		s2+4	32:68	26.3:73.7	5.70		s2+4	32:68	26.9:73.1	5.10
	s3+4	58:42	29.7:70.3	28.30		s3+4	58:42	61.1:38.9	3.10		s3+4	58:42	10.9:88.1	47.10
	s2+3+4	53:18:29	25.2:70.6:4.1	35.10		s2+3+4	53:18:29	1.7:25.4:72.9	34.20		s2+3+4	53:18:29	26.5:57:16.5	26.00
	c1+3	28:72	27:73	1.00		c1+3	72:28	62.6:37.4	9.40		c1+3	72:28	0.3:99.7	71.70
	s2+c1	50:50	66.7:33.3	16.70		s2+c1	50:50	19.7:80.3	30.30		s2+c1	50:50	15.2:84.8	34.80
	s2+c3	80:20	83.2:16.8	-3.20		s2+c3	80:20	97.9:2.1	17.90		s2+c3	80:20	19.3:80.7	60.70
	s3+c1	50:50	85.2:14.8	35.20		s3+c1	50:50	2.7:97.3	47.30		s3+c1	50:50	75:25	25.00
	s3+c3	17:83	5.8:94.2	11.20		s3+c3	83:17	65.2:34.8	17.80		s3+c3	83:17	18.2:81.8	64.80
	s4+c1	61:39	56.9:43.1	4.10		s4+c1	61:39	73.7:26.3	12.70		s4+c1	61:39	81.2:18.8	20.20
	s4+c3	82:18	72.1:27.9	9.90		s4+c3	82:18	56.4:43.6	25.60		s4+c3	82:18	37.4:62.6	44.60
		Mean absolute difference		15.16			Mean absolute difference		19.50			Mean absolute difference		39.80
		RMS difference		4.57			RMS difference		5.88			RMS difference		12.00
		Max absolute difference		20.04			Max absolute difference		27.80			Max absolute difference		34.70
		Min absolute difference		18.36			Min absolute difference		16.40			Min absolute difference		31.90

Processed Predicted 2 - smoothed

Table 20 - Data from Mixtures in Manual Analysis Mode with Pre-processing Option 2

For PXRD, a predicted result is said to closely match the actual values if the values are within 10% of each other in either direction.

For the PXRD data, five of the non-processed predictions closely match the actual results. With background removal and smoothing applied, five samples still closely match. With just smoothing applied, five of the predictions still closely match the results.

For the Raman data, three of the non-processed predicted results match the actual values. Two samples match with smoothing and background removal applied. With smoothing applied and no background removal there are again three samples closely matching. By expanding the allowed variance to 15%, five samples match with the actual values when no processing is applied, four match for the background and smoothed data and five samples now match from the smoothed data with the actual result.

For the IR data, four of the samples closely match with the non-processed predicted results. Two of the results, with smoothing and background removal on, match the actual results. With just smoothing applied four samples match the actual results. By expanding the margins of error to 15%, five of the samples match in the unprocessed data, three of the samples in the background removal and smoothing applied run matches and five of the samples match in the smoothing applied run.

5.10 CONCLUSIONS

The conclusions are as follows.

- The sulfathiazole/carbamazepine dataset has successfully tested the use of both DSC and IR data with PolySNAP alongside Raman and PXRD data. It was initially expected that DSC data would be difficult to compare as the patterns only have a small number of peaks present within them. This did not turn out to be the case as all peaks represent a melt or phase change and so can have radically different positions between polymorphs.
- Despite this ease of comparison the DSC data does not give the expected clustering in any of the datasets. It does show clear separation between the clusters in the MMDS plot however the contents of the clusters do not match with the expected clustering.
- IR data was expected to have the same problems of high similarity between patterns that Raman data showed. This however turned out not to be the case and different IR samples can be readily distinguished from one another. When either first or second derivatives were applied to the IR data, this causes degradation in the clustering results. Overall the clustering from IR data needs further work before it can be used as readily as that of Raman and PXRD data.
- INDSCAL combinations of all of the results give improved clustering.
- The importance of choosing the optimal pre-processing options for composition determination is shown here. For Raman data denoising and removing the background seem to give the optimal result. Allowing a variance of 15% rather than the standard 10% gives an improved result. For IR data increasing the allowed variance to 15% does not readily improve the match. The optimal match for IR data is to not pre-process it at all.
- TGA data has been tested and found to be off little use with PolySNAP. The patterns are all too similar for the software to be able to distinguish any clear difference between them. Due to these problems no further studies of TGA data will be carried out.
- Running a dataset with no pre-processing initially is the optimal way to determine the clustering. Further runs with different pre-processing should be carried out depending on the quality of the dendrogram for the non pre-processed data.

5.11 REFERENCES

1. Minitab – software for statistics. *www.minitab.com*

CHAPTER 6 SULFATHIAZOLE/CARBAMAZEPINE/ PIROXICAM DATASET

6.1 THE DATASET

The sulfathiazole/carbamazepine/piroxicam (SUTHAZ/CBZ/PIROX) dataset contains polymorphs two, three and four of sulfathiazole, polymorphs one and three of carbamazepine and polymorph two of piroxicam. Table 21 summarises the composition of this dataset.

For this dataset, PXRD data were collected on a Bruker C2 GADDS. Each sample was run for two minutes over a 3-30° range in 2θ. Raman data were collected on a Witec alpha 300 with a 785nm laser and an x10 objective lens with 0.25 aperture and 300g/mm grate. DSC data were collected on a TA Instruments Q100. IR data were collected on a Shimadzu FTIR-8400S.

Sample Number	Sample ID	Name in PolySNAP	Composition
1	Piroxicam Form 2	p	
2	Piroxicam Form 2 and Carbamazepine Form 1	p2+c1	12:88
3	Piroxicam Form 2 and Carbamazepine Form 3	p2+c3	28:72
4	Piroxicam form 2 and sulfathiazole form 2	p2+s2	22:78
5	Piroxicam Form 2 and Sulfathiazole Form 3	p2+s3	16:84
6	Piroxicam Form 2 and Sulfathiazole Form 4	p2+s4	47:53
7	Carbamazepine Forms 1 and 3 and Sulfathiazole Form 2	c1+c3+s2	48:32:20
8	Carbamazepine Forms 1 and 3 and Sulfathiazole Form 3	c1+c3+s3	24:47:29
9	Carbamazepine Forms 1 and 3 and Sulfathiazole Form 4	c1+c3+s4	33:33:33
10	Carbamazepine Form 1 and Sulfathiazole Forms 2 and 3	c1+s2+s3	26:32:42
11	Carbamazepine Form 1 and Sulfathiazole Forms 3 and 4	c1+s3+s4	33:33:33
12	Carbamazepine Form 1 and Sulfathiazole Forms 2 and 4	c1+s2+s4	33:33:33
13	Carbamazepine Form 3 and Sulfathiazole Forms 2 and 3	c3+s2+s3	15:46:39
14	Carbamazepine Form 3 and Sulfathiazole Forms 3 and 4	c3+s3+s4	24:66:10
15	Carbamazepine Form 3 and Sulfathiazole Forms 2 and 4	c3+s2+s4	24:45:31
16	Piroxicam Form 2 and Carbamazepine Forms 1 and 3	p2+c1+c3	12:66:22

Table 21 - Sulfathiazole-carbamazepine-piroxicam Dataset

6.2 SIMULATED DATASET

6.2.1 SIMULATED DATA CLUSTERING

Simulated powder data for each of the pure materials was taken from the CSD and were combined to produce the predicted patterns for each mixture. For example the p2+c1 pattern was produced by combining the patterns of p and c1 in a 12:88 ratio. The resulting dendrogram and MMDS plot are shown in Figure 135.

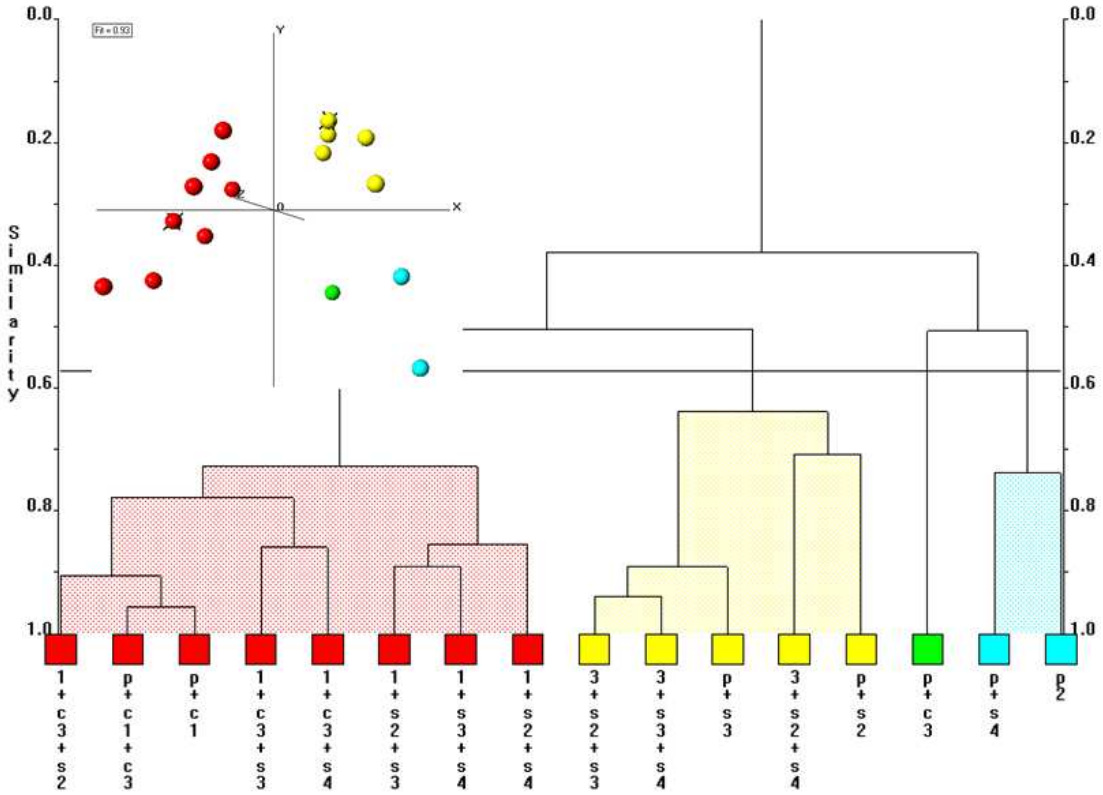

Figure 135 - Dendrogram and MMDS Plot for Simulated Dataset Clustering

The red cluster contains samples c1+c3+s2, p+c1+c3, p+c1, c1+c3+s3, c1+c3+s4, c1+s2+s3, c1+s3+s4 and c1+s2+s4. The yellow cluster contains sample c3+s2+s3, c3+s3+s4, p+s3, c3++s2+s4 and s2+c3. The green cluster contains sample p+c3 and the aqua cluster contains samples p+s4 and p. The silhouettes are shown in Figure 136.

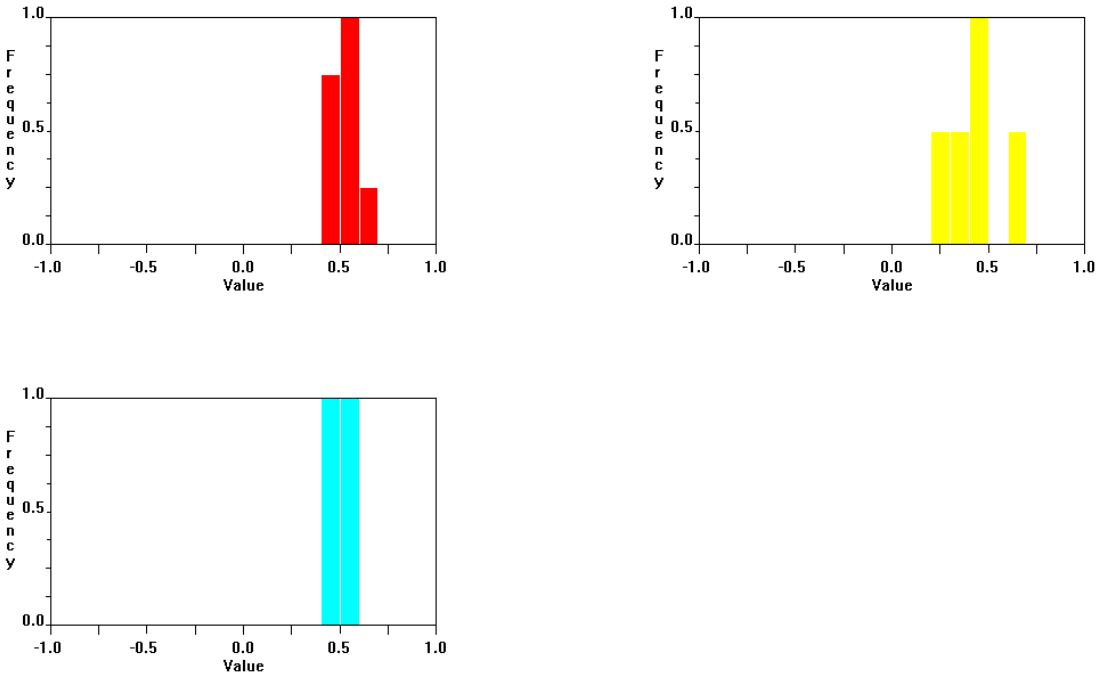

Figure 136 - Silhouettes for simulated dataset

For the red cluster the lower bar, lying below 0.5, contains samples c1+c3+s3, c1+s2+s3 and p+c1. The central bar, just above 0.5, contains samples c1+c3+s4, c1+s2+s4, c1+s3+s4 and p+c1+c3. The uppermost bar contains sample c1+c3+s2. For the yellow cluster the lower bar, lying just below 0.25 contains sample p+s2. The second bar lying just above 0.25 contains sample c3+s2+s4 while the third bar lying just below 0.5 contains samples c3+s3+s4 and p+s3. The uppermost bar contains sample c3+s2+s3. For the aqua cluster the lower bar, lying just below 0.5 contains p+s4 while the upper bar just above 0.5 contains p.

6.3 FINDING THE OPTIMAL CLUSTERING

The dataset was run through PolySNAP along with its ideal correlation coefficients from Minitab, as outlined in Section 5.3.

Figure 137 shows the dendrogram and MMDS plot for the Pearson correlation matrix and Figure 139 the dendrogram and MMDS plot for the Spearman correlation matrix.

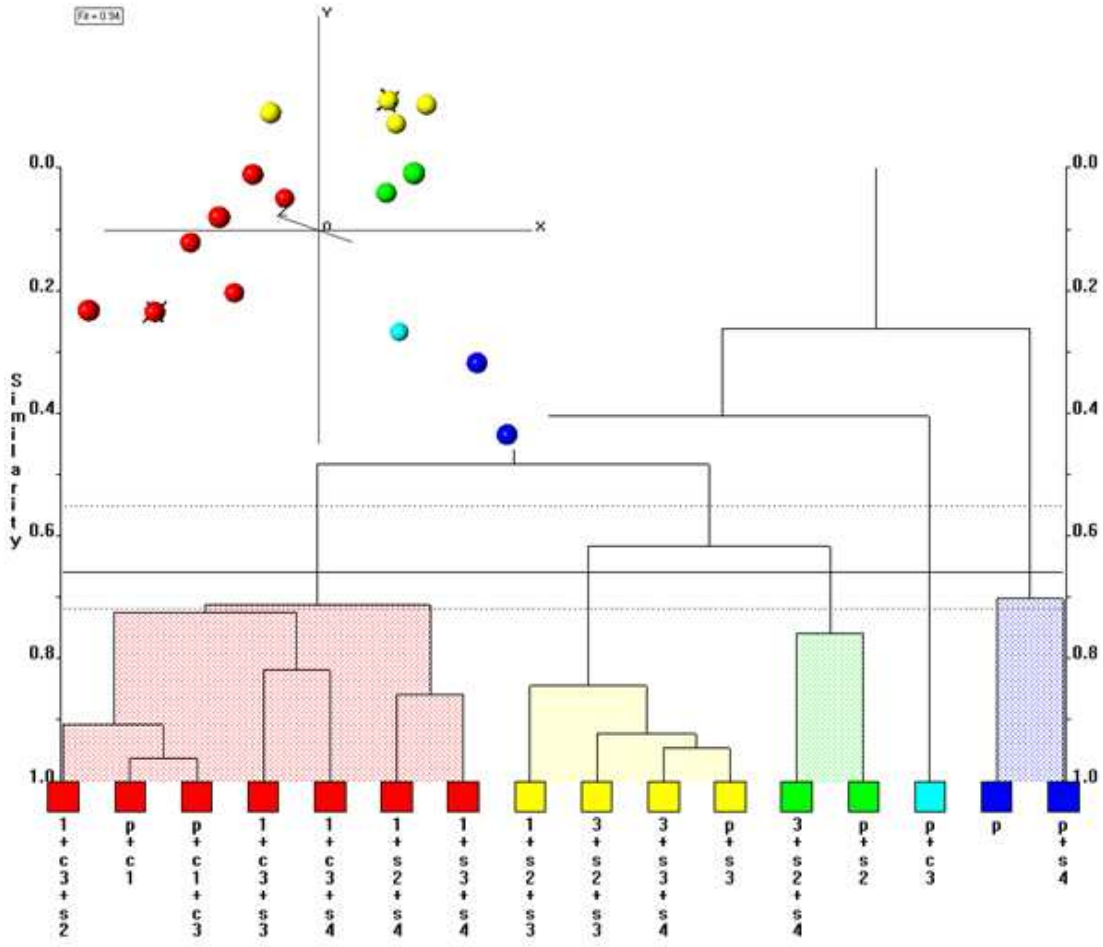

Figure 137 - Pearson correlation dendrogram and MMDS plot

For the Pearson correlation the red cluster contains samples c1+c3+s2, p+c1, p+c1+c3, c1+c3+s3, c1+c3+s4, c1+s2+s4 and c1+s3+s4. The yellow cluster contains samples c1+s2+s3, c3+s2+s3, c3+s3+s4 and p+s3. The green cluster contains sample c3+s2+s4 and p+s2. The aquamarine cluster contains samples p+c3 and the blue cluster contains samples p and p+s4.

The silhouettes are shown in Figure 138.

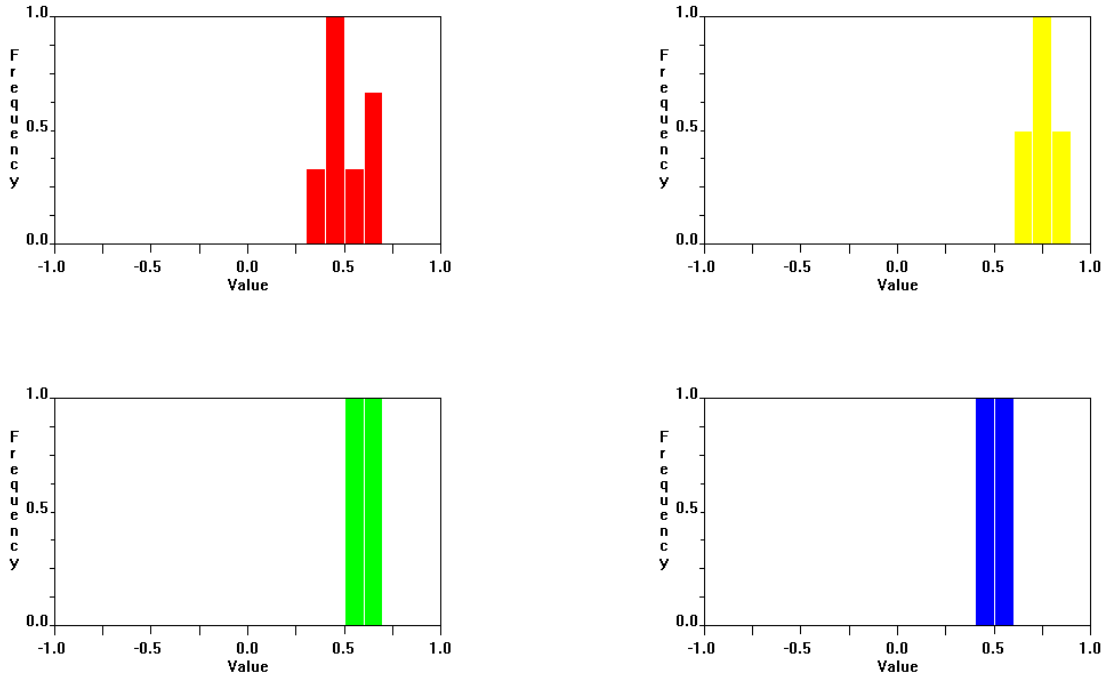

Figure 138 - Pearson Silhouettes

For the red cluster the lower bar, just above 0.25 corresponds to sample c1+c3+s3. The next bar, just below 0.5, corresponds to samples c1+s3+s4, c1+s2+s4 and p+c1. The next bar, just above 0.5 corresponds to sample c1+c3+s4. For the yellow cluster the lower bar, just below 0.75, corresponds to sample c1+s2+s3. The middle bar, at 0.75, corresponds to sample c3+s2+s4 and p+s3. For the green cluster the lower bar, just above 0.5, corresponds to sample c3+s2+s4. For the blue cluster the lower bar, just below 0.5, corresponds to sample p+s4. There is no fuzzy clustering for this dataset as no samples have samples that could potentially belong to more than one cluster.

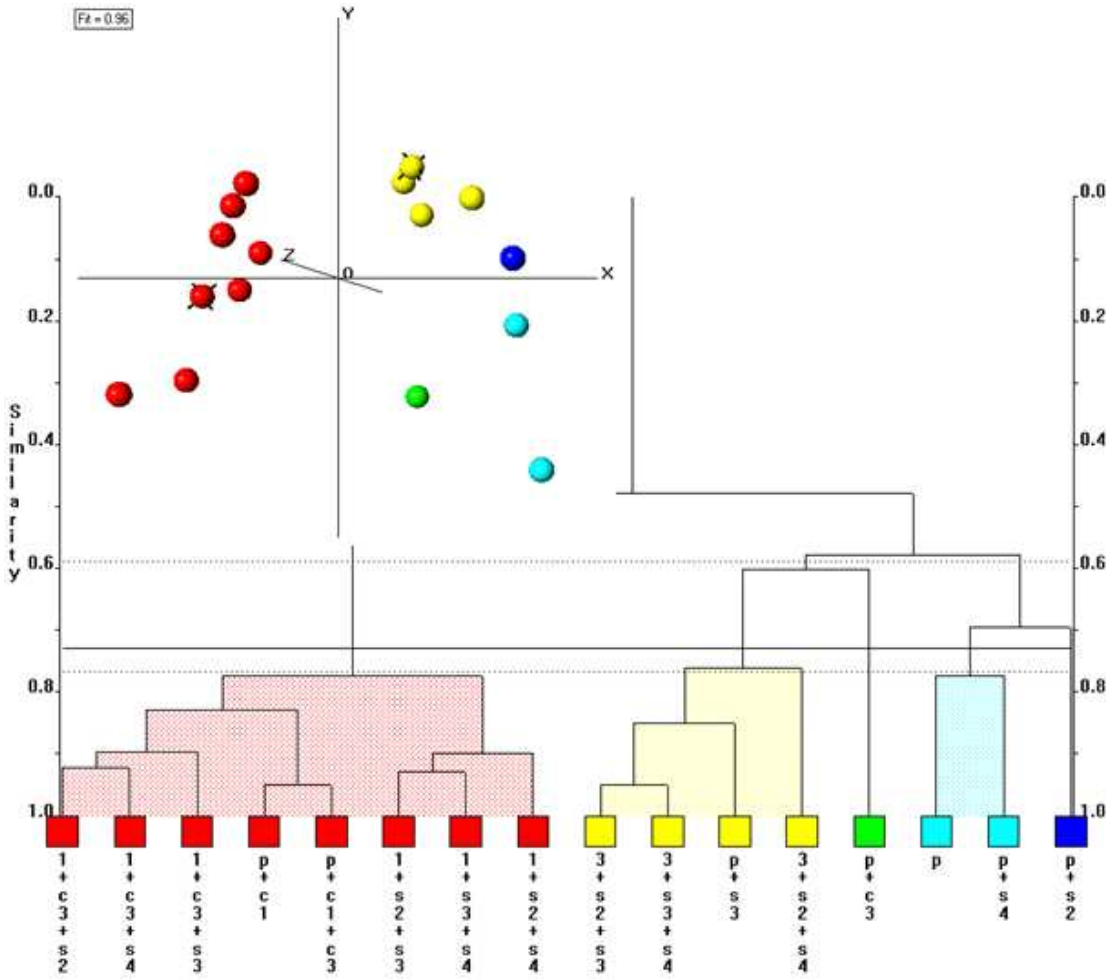

Figure 139 - Spearman correlation coefficient dendrogram and MMDS plot

The red cluster contains samples c1+s3+s2, c1+c3+s4, c1+c3+s3, p+c1, p+c1+c3, c1+s2+s3, c1+s3+s4 and c1+s2+s4. The yellow cluster contains samples c3+s2+s3, c3+s3+s4, p+s3 and c3+s2+s4. The green cluster contains sample p+c3. The aquamarine cluster contains samples p and p+s4 while the blue cluster contains sample p+s2.
The silhouettes are shown in Figure 140.

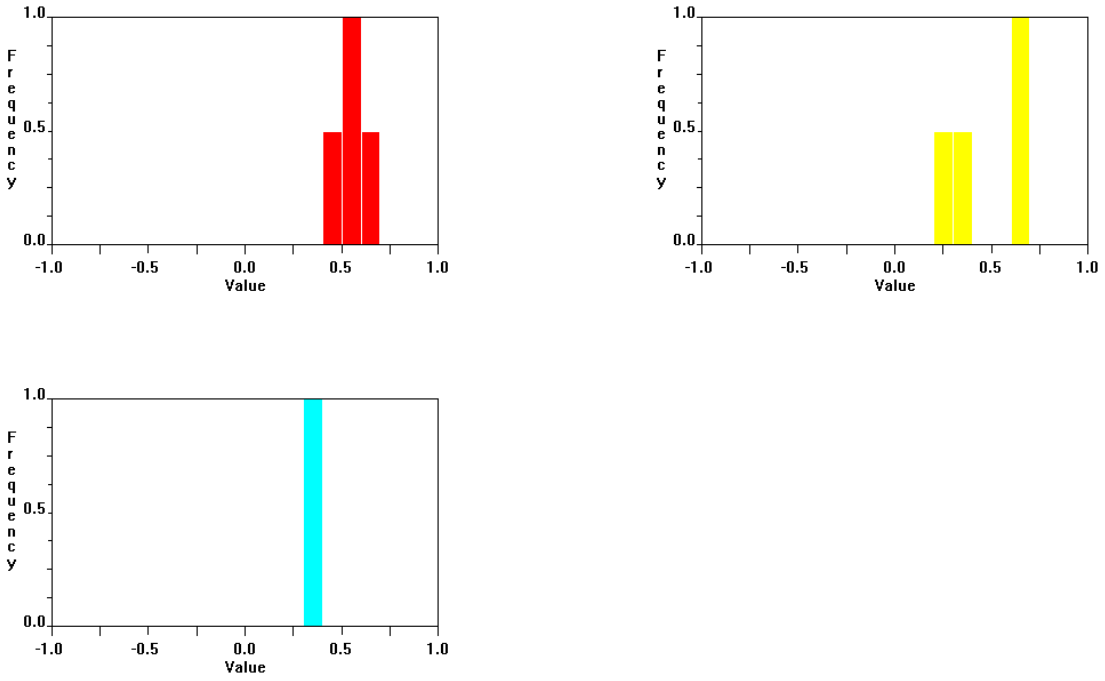

Figure 140 - Spearman correlation silhouettes

For the red cluster the lower bar, just below 0.5, contains sample c1+s2+s3 and p+c1. The middle bar contains samples c1+s3+s3, c1+s2+s4, c1+s3+s4 and p+c1+c3. For the yellow cluster the lower bar, just below 0.25, contains sample c3+s2+s4. The middle bar just above 0.25 contains sample p+s3. For the aquamarine cluster the samples all lie within a single bar. The fuzzy clustering is shown in Figure 141 with the numerical results in table 22.

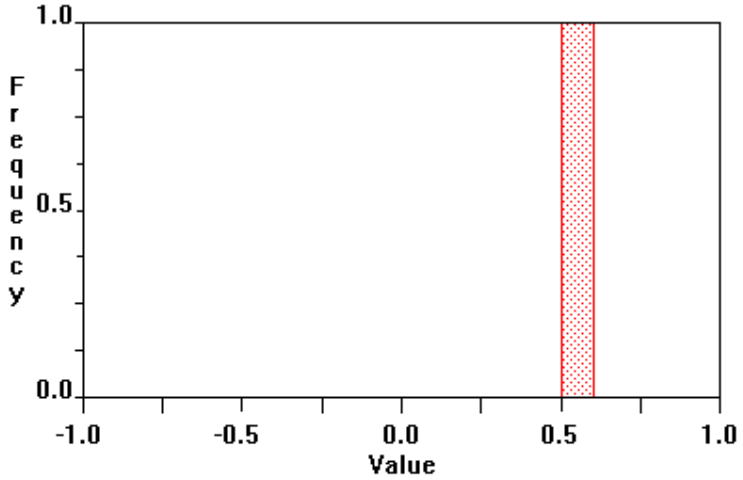

Figure 141 - Spearman fuzzy clustering

	1	2	3	4	5	
c1+c3+s2	0.17	0.13	0.18	0.27	1	
c1+c3+s3	0.19	0.15	0.2	0.38	0.99	
c1+c3+s4	0.18	0.17	0.2	0.3	1	
c1+s2+s3	0.13	0.14	0.2	0.9	0.60*	<==
c1+s2+s4	0.13	0.15	0.21	0.31	0.99	
c1+s3+s4	0.13	0.15	0.18	0.37	0.98	
c3+s2+s3	0.18	0.2	0.29	0.99	0.39	
c3+s2+s4	0.21	0.23	0.92	0.42	0.44	
c3+s3+s4	0.19	0.2	0.25	0.98	0.4	
p	0.17	0.92	0.21	0.21	0.18	
p+c1	0.11	0.09	0.07	0.12	0.99	
p+c1+c3	0.16	0.13	0.12	0.19	1	
p+c3	0.9	0.28	0.25	0.31	0.43	
p+s2	0.17	0.29	0.92	0.36	0.25	
p+s3	0.16	0.25	0.25	0.96	0.28	
p+s4	0.2	0.94	0.28	0.33	0.3	

Table 22 – Spearman fuzzy clustering results

The fuzzy clustering contains one sample, c1+s2+s3. For the numerical results, column 1 represents the green cluster, column 2 the aquamarine cluster, column 3 the yellow cluster, column 4 the red cluster and column 5 the blue cluster. Sample c1+s2+s3 could potentially appear in either the blue or red cluster.

Comparing the results for both the Pearson and Spearman correlations to the predicted results, it is again revealed that the Pearson result is highly similar. A small adjustment of the cut-level will allow the two dendrograms to match exactly. This adjusted cut-level is shown in Figure 142.

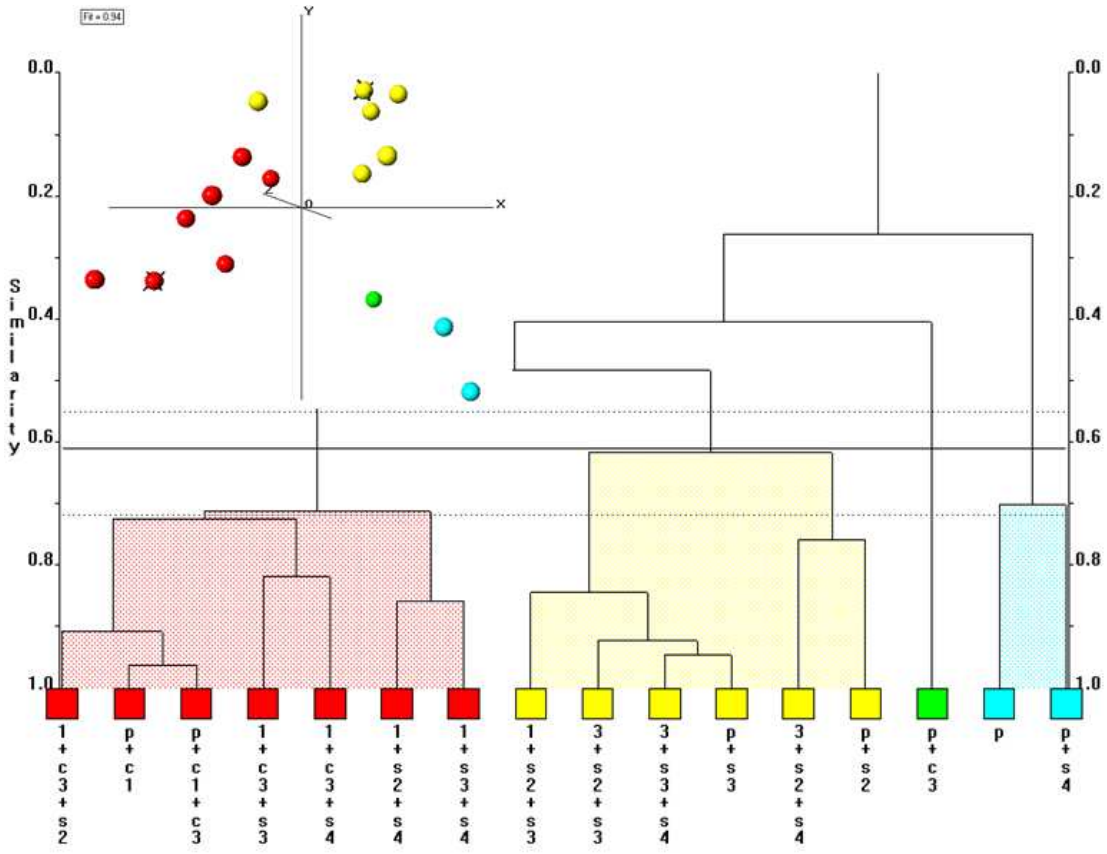

Figure 142 - Pearson dendrogram and MMDS plot with adjusted cut-level

As the Pearson correlation again matches exactly with the predicted result, the Pearson correlation dendrogram will be treated as showing the optimal clustering.

6.3 RAMAN AND IR REGIONS OF SIMILARITY

The Raman and IR datasets were examined to determine areas of similarity and therefore which areas should be given priority when matching the spectra. The Raman spectra for the pure materials of sulfathiazole, carbamazepine and piroxicam are shown in Figure 143.

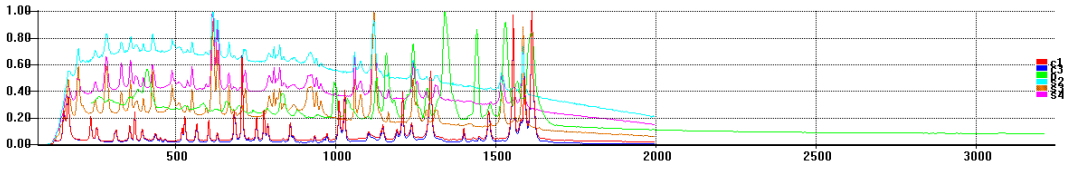

Figure 143 - Raman spectra for pure materials

As with the sulfathiazole and carbamazepine spectra, no peaks appear after 1750cm^{-1}. As such only the areas before 1750cm^{-1} will be used when comparing the Raman spectra.

The IR spectra for the pure materials are shown in Figure 144.

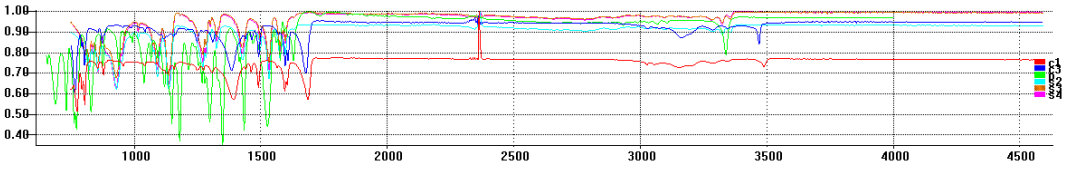

Figure 144 - IR spectra for pure materials

The areas of significance for IR spectra appear to lie in the region before 1750cm^{-1} and between 3000 and 3500cm^{-1}. As such these areas shall be used exclusively during matching of the IR spectra.

6.4 DATASET CLUSTERING

6.4.1 EXPECTED CLUSTERING

As for the dataset in Chapter 5, the PXRD patterns for the pure polymorphs were combined to produce predicted patterns for each of the mixtures. The dendrogram and MMDS from this are shown in Figure 145.

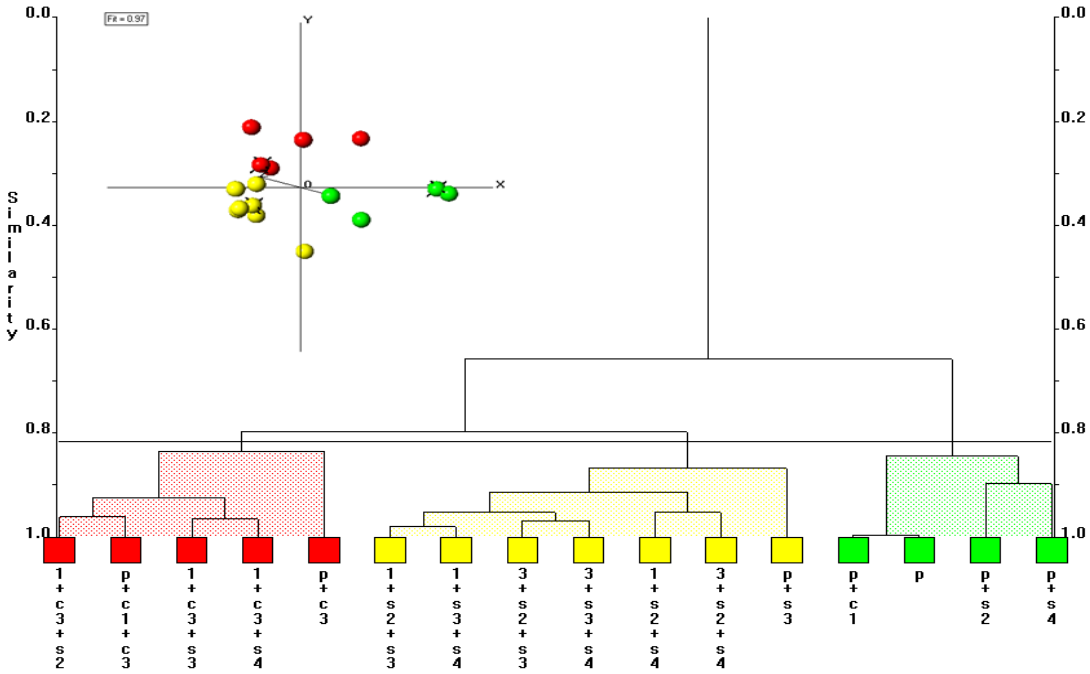

Figure 145 - Expected Clustering

The red cluster contains patterns c1+c3+s2, p+c1+c3, c1+c3+s3, c1+c3+s4 and p+c3. The yellow cluster contains patterns c1+s2+s3, c1+s3+s4, c3+s3+s4, c1+s2+s4 and p+s3. The green cluster contains patterns p+c1, p, p+s2 and p+s4. Some of the tie bars lie close to the cut-level so the silhouettes (Figure 146) and fuzzy cluster (Figure 147 and Table 23) will be analysed. For the fuzzy clustering cluster 1 is the red cluster, cluster 2 is the yellow cluster and cluster 3 is the green cluster.

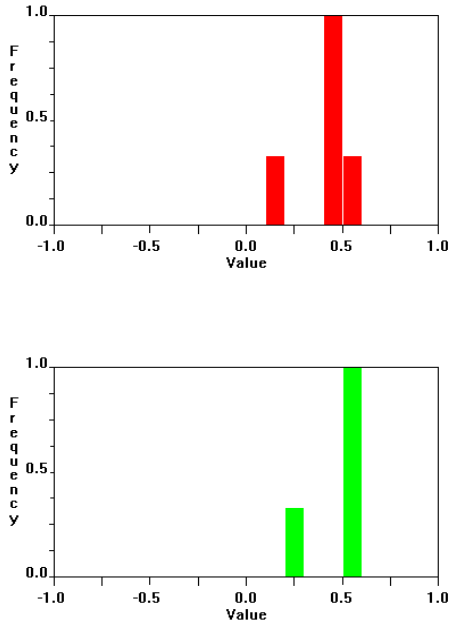

Figure 146 – Silhouettes

For the red cluster, all but one of the patterns lies below 0.5. The lowest region contains pattern p+c3, which also has the highest tie bar in the cluster, while the second region corresponds to patterns p+c1+c3, c1+c3+s3 and c1+c3+s4. For the yellow cluster, three patterns lie below 0.5. The lowest such region contains pattern p+s3 while the second, and last region below 0.5, contains patterns c1+s2+s4 and c3+s2+s4. The green cluster has one pattern below 0.5, sample p+s2 which also has a tie bar lying very close to the cut-level.

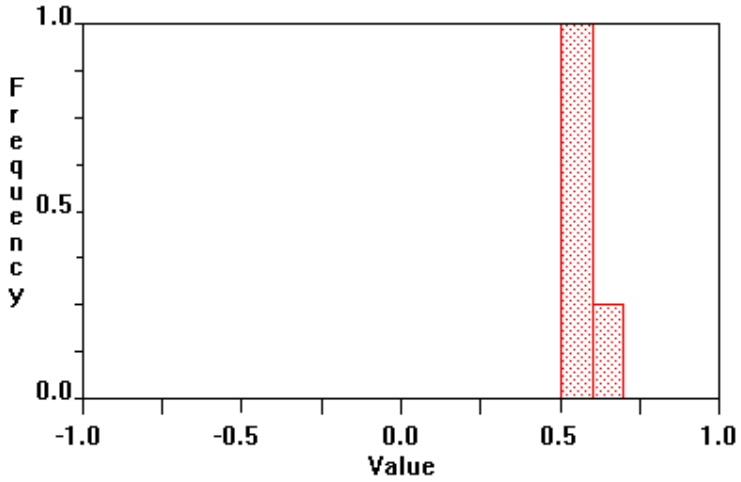

Figure 147 - Fuzzy Clustering

	1	2	3	
c1+c3+s2	1	0.52*	0.31	<==
c1+c3+s3	0.99	0.60*	0.35	<==
c1+c3+s4	0.99	0.60*	0.35	<==
c1+s2+s3	0.47	1	0.35	
c1+s2+s4	0.48	1	0.35	
c1+s3+s4	0.46	1	0.34	
c3+s2+s3	0.48	1	0.37	
c3+s2+s4	0.49	1	0.37	
c3+s3+s4	0.45	1	0.34	
p+c1+c3	0.99	0.5	0.39	
p+c1	0.34	0.31	0.99	
p+c3	0.96	0.41	0.43	
p+s2	0.45	0.55*	0.97	<==
p+s3	0.37	0.99	0.39	
p+s4	0.41	0.51*	1	<==
p	0.32	0.29	0.99	

Table 23 – Expected Fuzzy Clustering Numeric Data

The fuzzy clustering plot contains two regions of patterns. The lower of these two bars contains patterns c1+c3+s2, c1+c3+s3, p+s2 and p+s4. These patterns could be found in the following clusters

- c1+c3+s2 could potentially be in either the red or yellow cluster
- c1+c3+s3 could potentially be in either the red or yellow cluster
- p+s2 could potentially be in either the yellow or green cluster
- p+s4 could potentially be in either the yellow or green cluster.

The second region in the fuzzy clustering contains pattern c1+c3+s4. This pattern could potentially be in either the red or yellow cluster.

Due to patterns p+c3 in the red cluster, p+s3 in the yellow cluster and p+s2 in the green cluster having very low correlations to the remainder of their cluster, the cut-level shall be lowered slightly to separate these patterns. The resulting dendrogram and MMDS plot from this is shown in Figure 148.

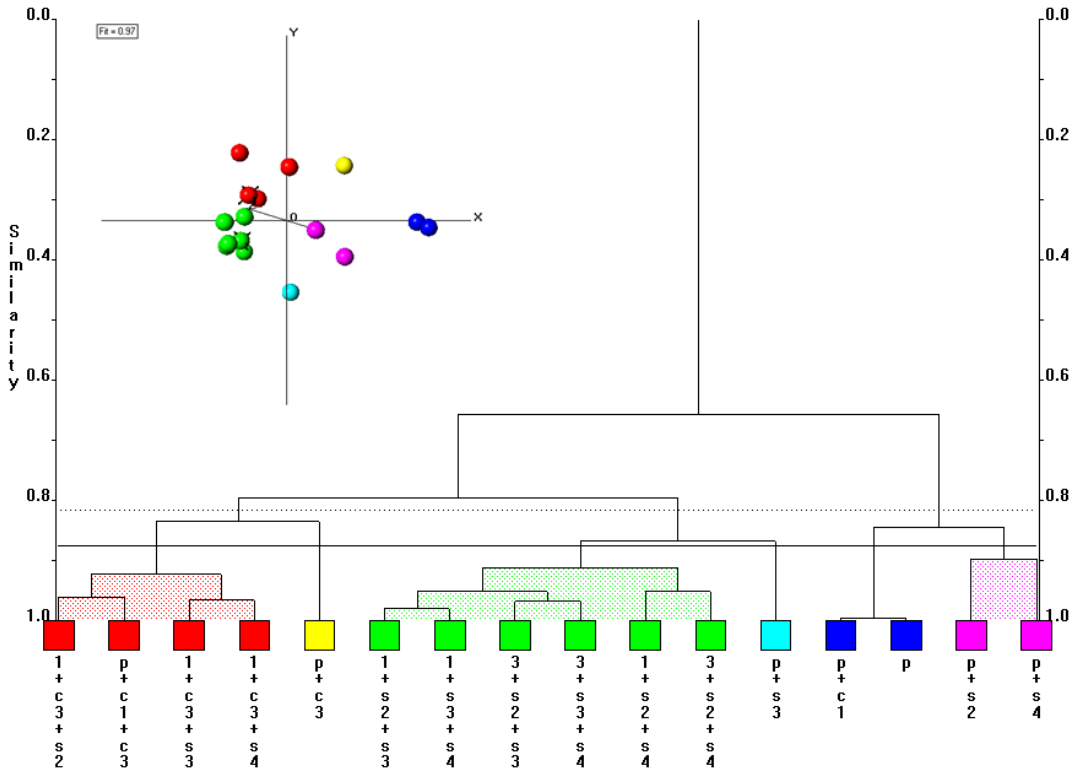

Figure 148 – Expected Clustering with Adjusted Cut-level

The simulated dataset had the following clusters present:

1) c1+c3+s2, p+c1+c3, p+c1, c1+c3+s3, c1+c3+s4, c1+s2+s3, c1+s3+s4, c1+s2+s4
2) c3+s2+s3, c3+s3+s4, p+s3, c3+s2+s4, p+s2
3) p+c3
4) p+s4, p

The expected clustering dataset has the following clustering:

1) c1+c3+s2, p+c1+c3, c1+c3+s3 c1+c3+s4
2) p+c3
3) c1+s2+s3, c1+s3+s4, c3+s2+s3, c3+s3+s4, c1+s2+s4 and c3+s2+s4
4) p+s3
5) p+c1 and p
6) p+s2 and p+s4

By comparing the predicted and expected dataset, it can be seen that several of the samples have moved to different clusters. As the predicted clustering matched with the manually calculated Pearson correlation coefficient, this will be treated as the optimal clustering. The optimal clustering is as follows:

1) A cluster containing samples c1+c3+s2, p+c1+c3, p+c1, c1+c3+s3, c1+c3+s4, c1+s2+s3, c1+s3+s4, c1+s2+s4
2) A cluster containing samples c3+s2+s3, c3+s3+s4, p+s3, c3+s2+s4, p+s2
3) A cluster containing samples p+c3
4) A cluster containing samples p+s4, p

6.4.2 PXRD DATA

The dendrogram and MMDS plot for the PXRD data are shown in Figure 149.

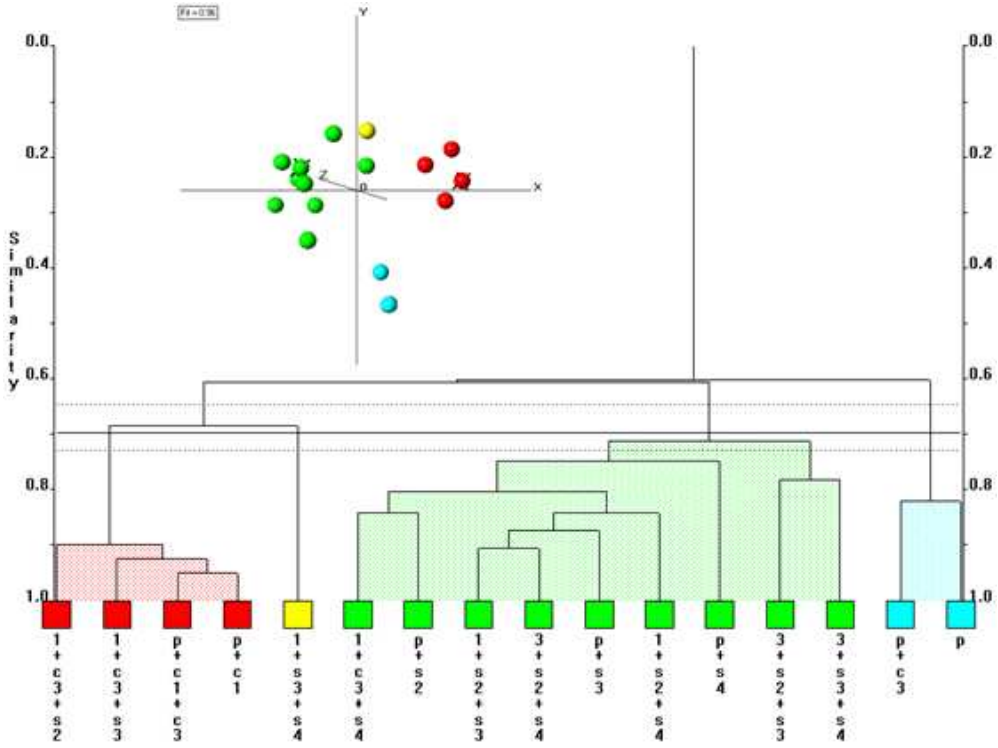

Figure 149 - PXRD Dendrogram and MMDS Plot

The red cluster contains samples c1+c3+s2, p+c1+c3, c1+c3+s3 and p+c1 from expected cluster 1. The yellow cluster contains sample c1+s3+s4 which is part of expected cluster 1. The green cluster contains samples p+s2, c3+s2+s4, c1+s2+s3, p+s3 and c3+s3+s4 from expected cluster 2, p+s4 from cluster 4 and c1+c3+s4, c3+s2+s3 and c1+s2+s4 from expected cluster 1.

The aquamarine cluster contains sample p+c3 from expected cluster 3 and p from expected cluster 4.

The PXRD data is not as poorly clustered as it first appears. Half of expected cluster 1 is in the red cluster with the other half in the green cluster, intermixed with other samples. The entirety of expected cluster 2 is in the green cluster, again intermixed with other samples. The score for this dendrogram is 0.44, implying that more than half of the samples are clustered as expected.

6.4.3 RAMAN DATA

The dendrogram and MMDS plot for the Raman data are shown in Figure 150. The Raman data was matched solely on data before 1750cm^{-1}.

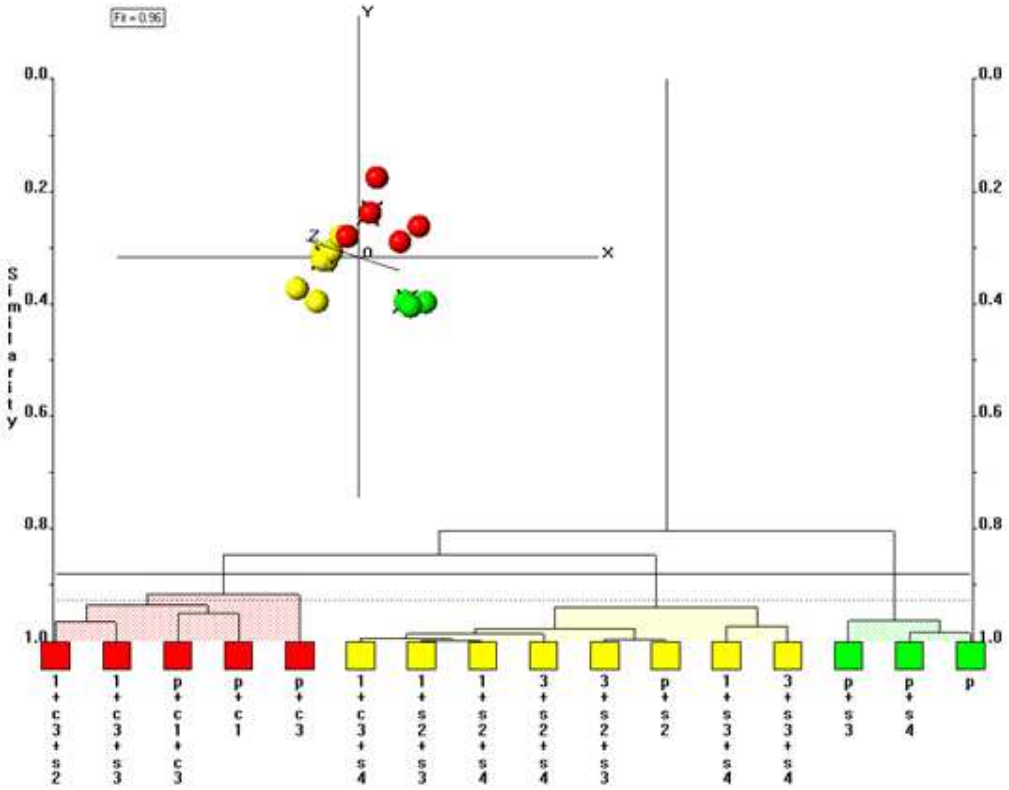

Figure 150 - Raman Dendrogram and MMDS plot

The red cluster contains samples c1+c3+s2, p+c1, c1+c3+s3 and p+c1+c3 from expected cluster 1 as well as sample p+c3 which is the lone sample in expected cluster 3. The yellow cluster contains sample c1+c3+s4, c1+s2+s3, c1+s2+s4 and c1+s3+s4 from expected cluster 1, c3+s2+s4, c3+s2+s3, c3+s3+s4 and p+s2 which are part of expected cluster 2. The green cluster contains samples p+s3, which is expected to be cluster on its own in cluster 4 and p+s4 and which make up expected cluster 2. This dendrogram has a score of

0.44, the same as that for the PXRD dataset implying that the dataset isn't as poorly clustered as initially appear.

6.4.4 DSC DATA

The dendrogram and MMDS plot for DSC data are shown in Figure 151.

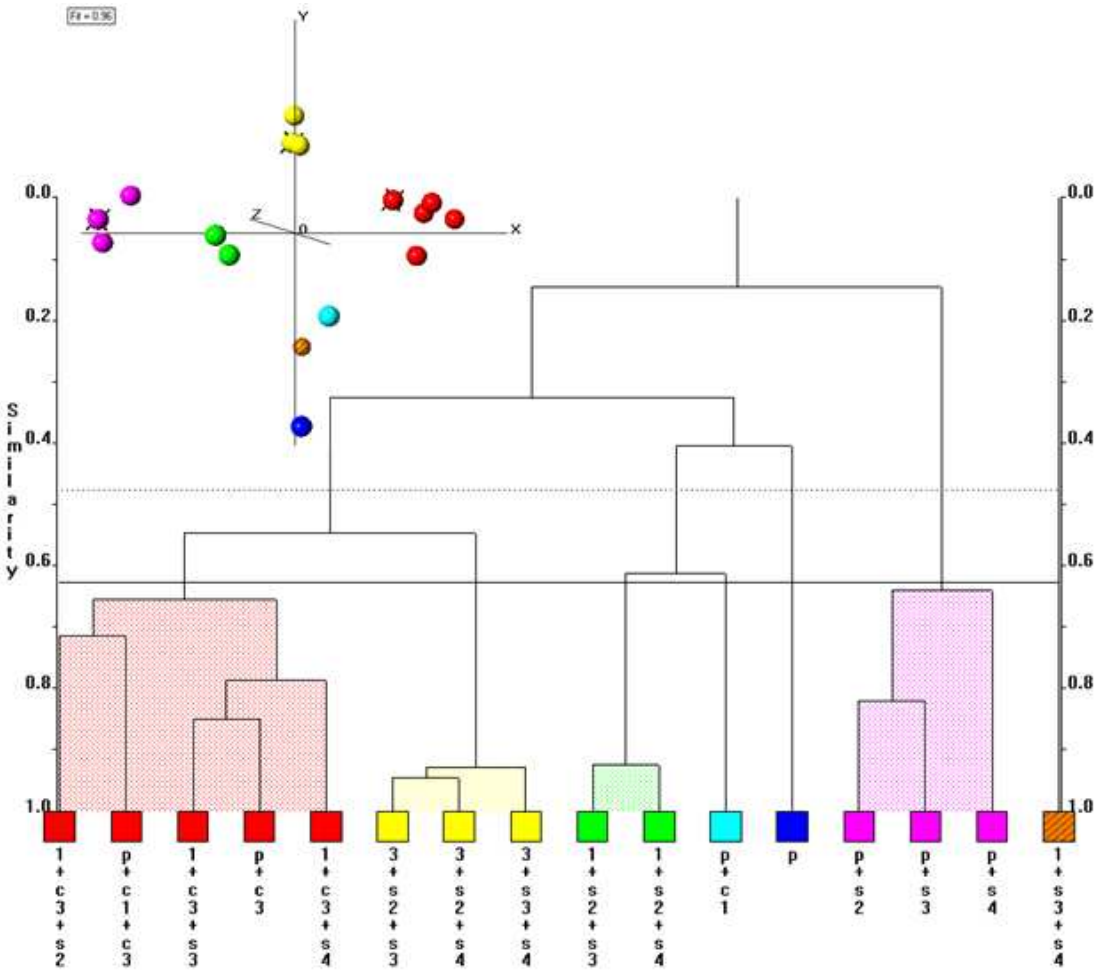

Figure 151 - DSC Dendrogram and MMDS Plot

The red cluster contains samples c1+c3+s2, c1+c3+s3, c1+c3+s4 and p+c1+c3, which make up part of predicted cluster 1, as well as sample p+c3, which makes up the entirety of predicted cluster 4. The yellow cluster contains samples c3+s2+s4, c3+s2+s4 and c3+s3+s4, which make up part of expected cluster 2. The green cluster contains samples c1+s2+s3 and c1+s2+s4 which makes up part of expected cluster 1. The aquamarine cluster contains p+c1 which is part of expected cluster 1. The blue cluster contains sample p which is part of excepted cluster 2. The purple cluster contains samples p+s2 which is

part of expected cluster 2, p+s4 which make up expected cluster 4 and p+s3 which makes up part of expected cluster 1. The striped brown cluster contains sample c1+s3+s4 which is part of expected cluster 1.

This does not give the expected clustering. The clustering is poorer for this dataset than in the X-ray or Raman datasets with a score for this dendrogram of 0.5. This score still equates to half of the samples being clustered as expected.

6.4.5 IR DATA

The dendrogram and MMDS plot for IR are shown in Figure 152.

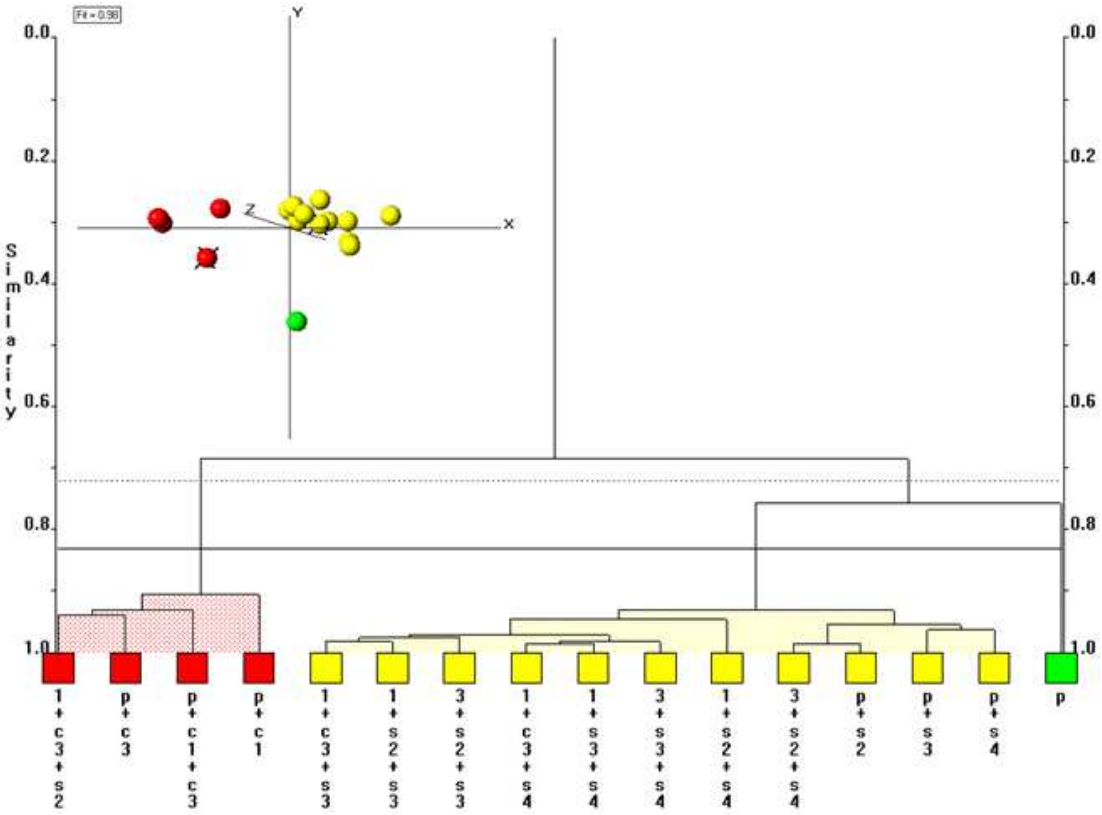

Figure 152 - IR Dendrogram and MMDS Plot

The red cluster contains samples c1+c3+s2, p+c1+c3 and p+c1 which are part of expected cluster 1 and p+c3 which makes up the entirety of expected cluster 2. The green cluster contains sample p which is part of expected cluster 2. The yellow cluster contains the remaining eleven samples in the dataset.

The dataset is not as well clustered as that derived from PXRD or Raman data. The dendrogram has a score of 0.56, slightly poorer than that of the DSC dendrogram and with just over half of the samples incorrectly clustered.

6.4.6 COMBINED DATA

The dendrogram and MMDS plot from the combined dataset, combining all four data types using INDSCAL, are shown in Figure 153.

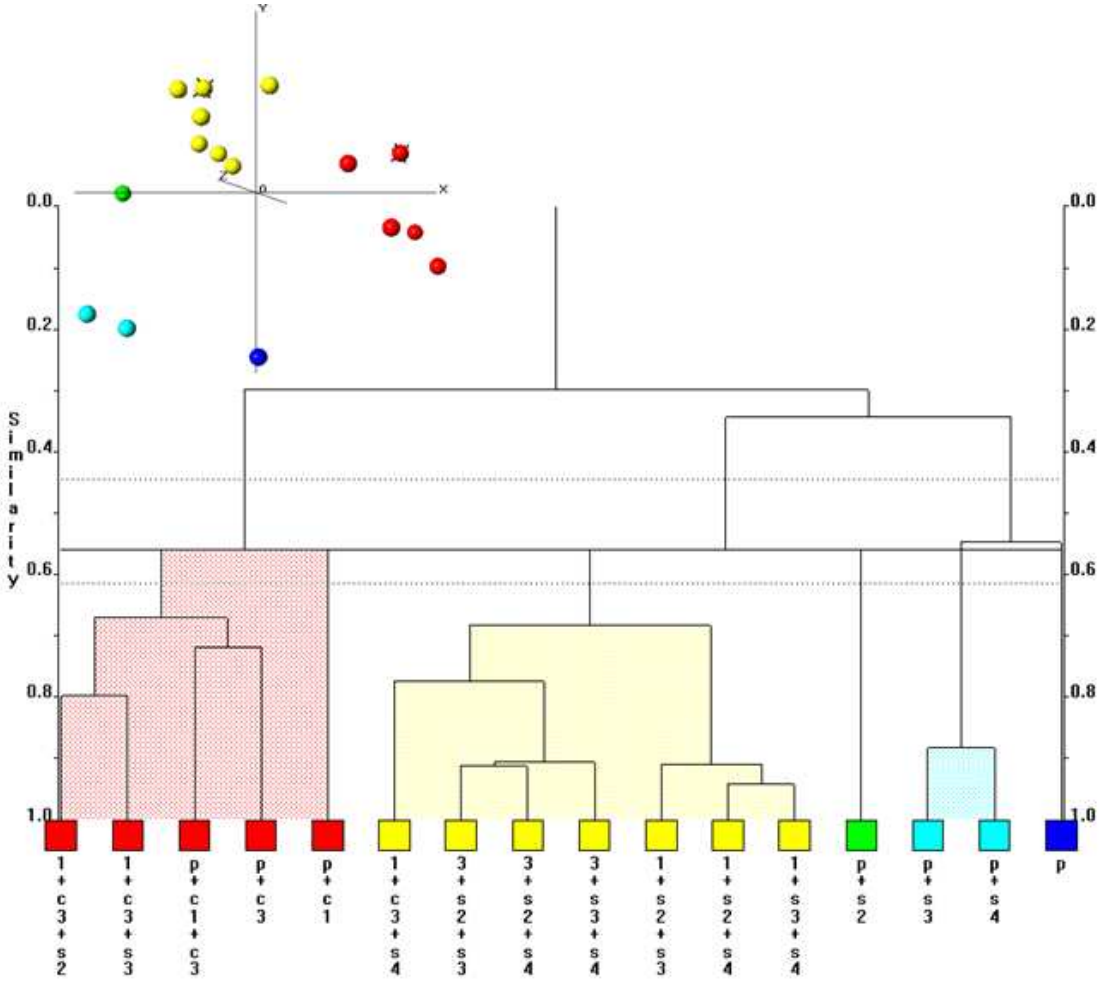

Figure 153 - Combined Dendrogram and MMDS Plot

The red cluster contains samples c1+c3+s2, c1+c3+s3, p+c1+c3 and p+c1 which are part of expected cluster 1 as well as sample p+c3 which makes up expected cluster 3.
The yellow cluster contains samples c1+c3+s4, c1+s2+s3, c1+s2+s4 and c1+s3+s4 which are part of expected cluster 1 and c3+s2+s3, c3+s2+s4 and c3+s3+s4, which make up part of expected cluster 4. The green cluster contains p+s2 which is part of expected cluster 1.

The aquamarine cluster contains p+s3 which make up predicated cluster 3 and p+s4 which is part of expected cluster 4. The blue cluster contains the pure piroxicam sample which is part of expected cluster 1. The combined dataset dendrogram has a score of 0.5, half of the samples are clustered as expected.

6.5 DERIVATIVE DATA

The Raman and IR data is re-run with both first and second derivatives applied.

6.5.1 RAMAN

The dendrogram and MMDS plot for the first derivative re-run are shown in Figure 154.

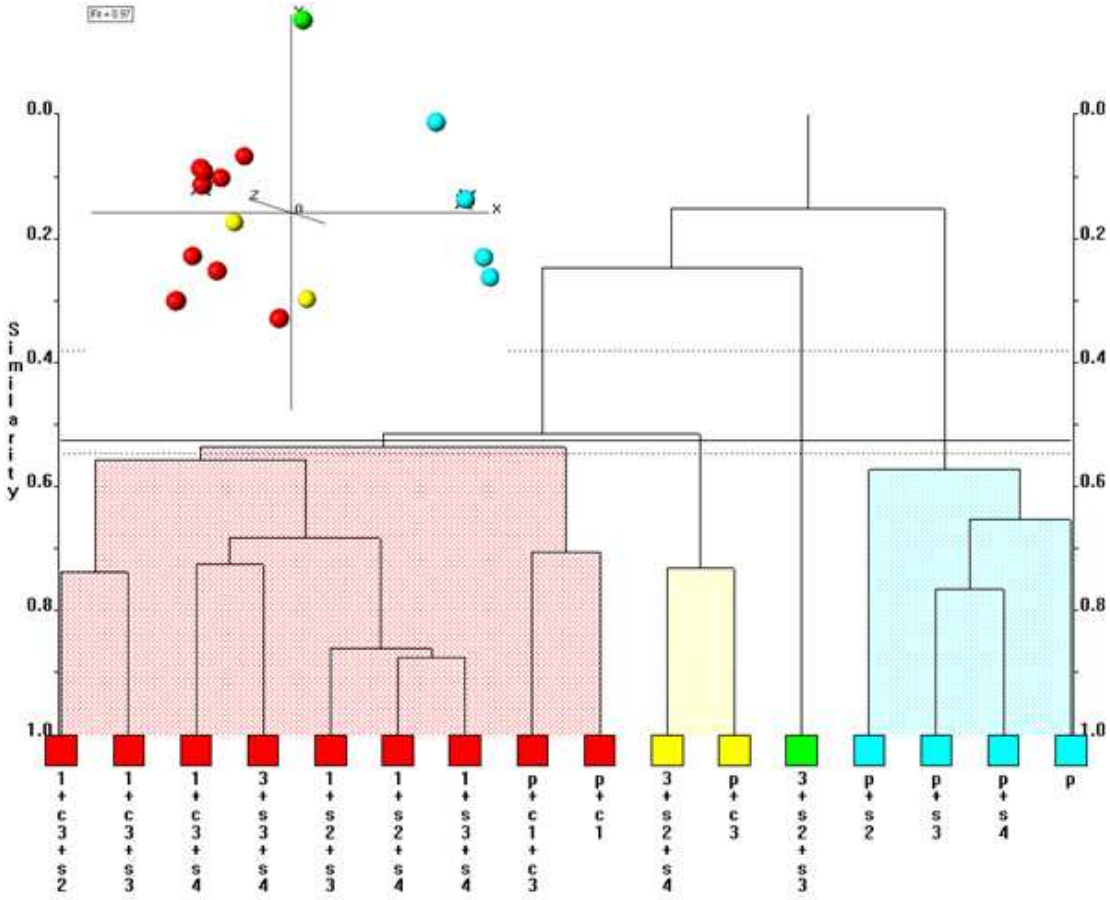

Figure 154 - First Derivative Raman Dendrogram and MMDS Plot

The red cluster contains samples c1+c3+s2, c1+c3+s3, c1+c3+s4, c1+s2+s3, c1+s2+s4, p+c1, c1+s3+s4 and p+c1+c3, which make up part of expected cluster 1 and c3+s2+s4,

which makes up part of expected cluster 2. The yellow cluster contains c3+s2+s4 which makes up part of expected cluster 2 and p+c3 which is the entirety of expected cluster 3. The green cluster contains sample c3+s2+s3 which is part of expected cluster 2. The aquamarine cluster contains sample p+s2 which is part of expected cluster 2, p+s4 and p which make up expected cluster 4 and p+s3 which makes up part of expected cluster 2. The dataset shows reasonable clustering, though not as good as that seen in the original Raman run. The first derivative dendrogram has a score of 0.38 showing better clustering than that seen in the original Raman dataset.

The second derivative Raman dendrogram and MMDS are shown in Figure 155.

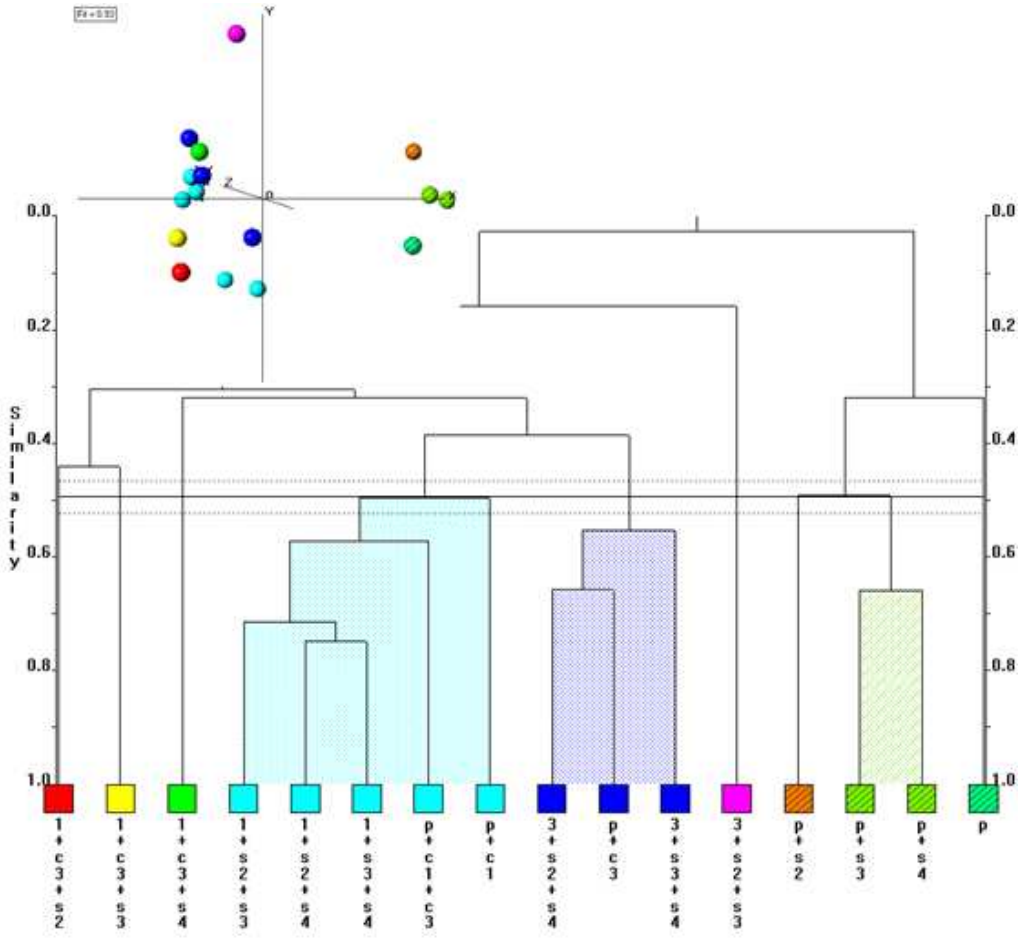

Figure 155 - Second Derivative Raman Dendrogram and MMDS Plot

The red cluster contains sample c1+c3+s2 which is part of expected cluster 1. The yellow cluster contains sample c1+c3+s3 which is part of expected cluster 1. The green cluster contains sample c1+c3+s4 which is part of expected cluster 1. The aquamarine cluster contains samples c1+s2+s3, c1+s2+s4, p+c1+c3, p+c1 and c1+s3+s4 which are part of

expected cluster 1. The blue cluster contains samples c3+s2+s4 and c3+s3+s4 which are part of expected cluster 2 and p+c3 which makes up expected cluster 3. The purple cluster contains sample c3+s2+s3 which is part of expected cluster 2. The striped brown cluster contains sample p+s2 which is part of expected cluster 2. The striped light green cluster contains sample p+s3 which make up cluster 2 and p+s4 which is part of cluster 4. The striped dark green cluster contains the pure piroxicam sample which is part of cluster 4. Overall the dataset does not give good clustering. The dendrogram has a score of 0.44, identical to that seen in the original Raman dataset.

6.5.2 INFRARED

The dendrogram and MMDS plot for the first derivative run are shown in Figure 156.

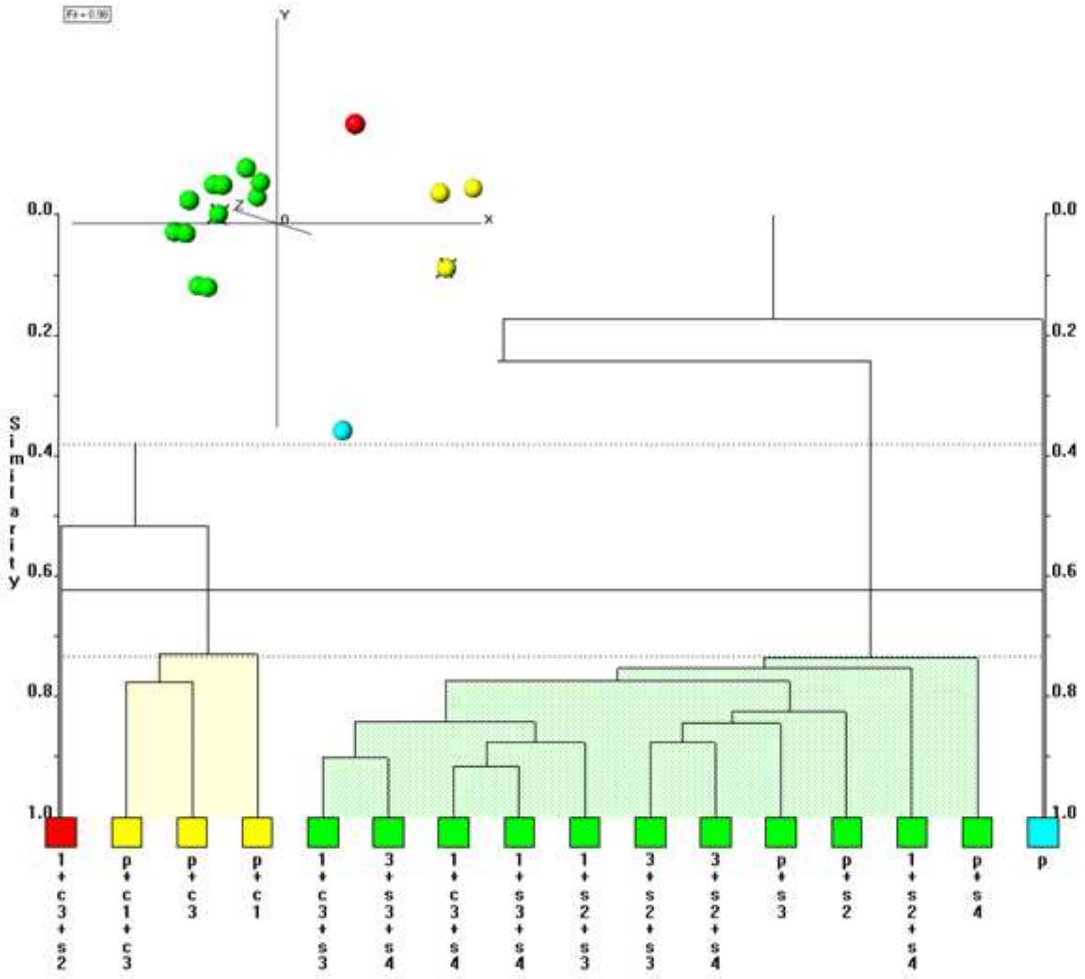

Figure 156 - First Derivative IR Dendrogram and MMDS Plot

The red cluster contains sample c1+c3+s2 which is part of expected cluster 1. The yellow cluster contains samples p+c1+c3 and p+c1which is part of expected cluster 1 and p+c3 which makes up expected cluster 3. The green cluster contains samples c3+s3+s4, p+s3, p+s2, c3+s2+s3 and c3+s2+s4 which make up expected cluster 2, c1+s3+s4, c1+c3+s3, c1+c3+s4 and c1+s2+s4 which make up part of expected cluster 1 and p+s4 which make up part of expected cluster 4. The aquamarine cluster contains the pure piroxicam sample which is part of expected cluster 4. The MMDS plot shows the clusters to be well defined. The dendrogram here has a score of 0.44, an improvement over the 0.56 seen in the original IR dataset.

The second derivative IR dendrogram and MMDS plot are shown in Figure 157.

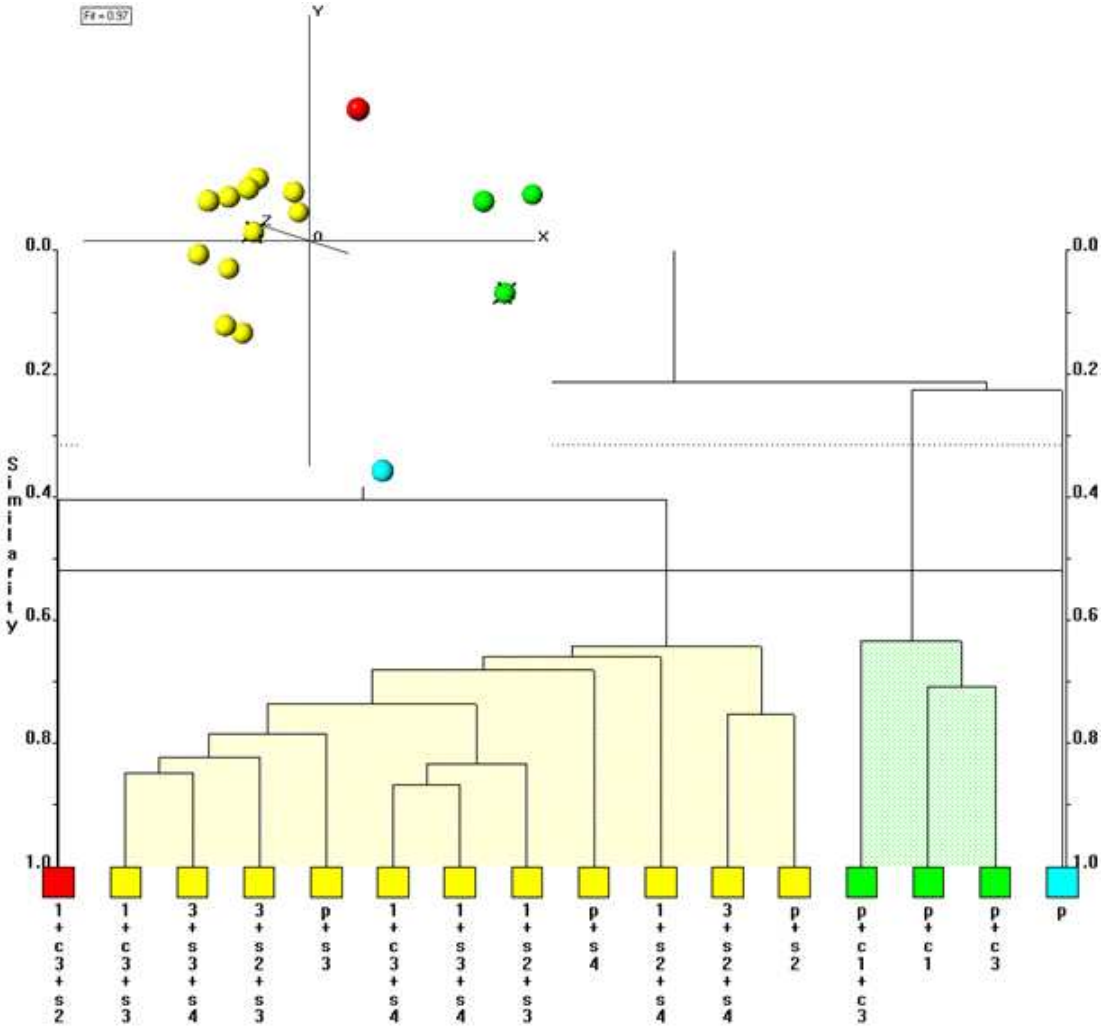

Figure 157 - Second Derivative Dendrogram and MMDS Plot

The red cluster contains sample c1+c3+s2 which is part of expected cluster 1. The yellow cluster contains samples c1+c3+s3, c1+s2+s3, c1+s2+s4 and c1+c3+s4 which are part of expected cluster 1, c3+s3+s4, p+s3, p+s2, c3+s2+s3, c3+s2+s4 and c3+s2+s4 which make up expected cluster 2 and p+s4 which make up part of expected cluster 4. The green cluster contains sample p+c1+c3 and p+c1 which makes up part of expected cluster 1 and p+c3 which makes up expected cluster 3. The aquamarine cluster contains the pure piroxicam sample which makes up part of expected cluster 4. The score for this dendrogram is 0.5.

6.6 FLOWCHART

All of the possible combinations of pre-processing were applied to the PXRD dataset and the results compared to the optimal clustering. The number of misplaced samples is shown in Table 24.

	Score
no pre-processing	0.43
denoise	0.43
background	0.5
background and denoise	0.5

Table 24 - Misplaced samples

As can be seen no pre-processing and denoising give optimal results. As no pre-processing gives best results with least processing applied it is the preferred method.
The flowchart for this result is shown in Figure 158.

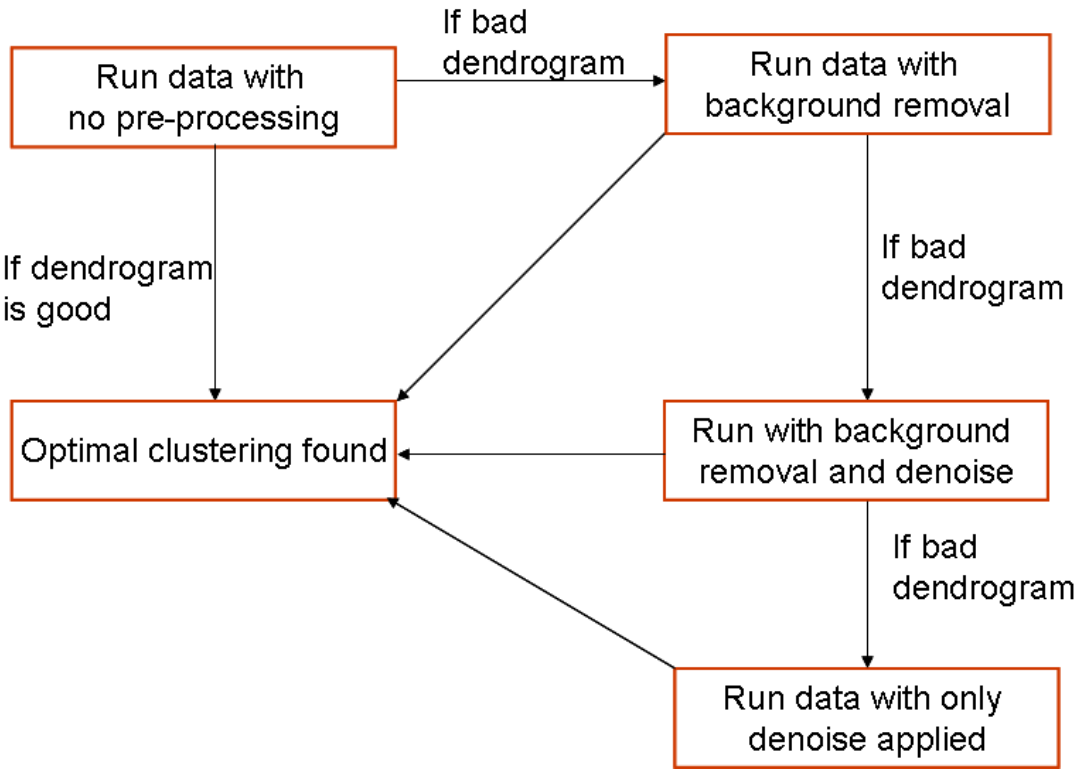

Figure 158 - Flowchart for initial 16 samples

This flowchart is again the same as that previously developed for the datasets in Chapters 4 and 5. A bad dendrogram is again defined as one that

1. Shows ‘chaining’ of the samples or has no clear clustering
2. The scree plot does not show the characteristic steep initial drop before smoothing out
3. The maximum and minimum confidences shows a large separation.

6.7 THIRTY-TWO SAMPLE DATASET

The sixteen samples from this dataset were combined with the sixteen from the sulfathiazole-piroxicam dataset, as discussed in Chapter 5. The 16 additional samples that are being added to this dataset, as originally studied in Chapter 5, are summarised in Table 25.

Sample Number	Sample ID	Name in PolySNAP	Composition
1	Sulfathiazole Form 4	s4	
2	Sulfathiazole Form 3	s3	
3	Sulfathiazole Form 2	s2	
4	Carbamazepine Form 1	c1	
5	Carbamazepine Form 3	c3	
6	Sulfathiazole Forms 3 and 4	s4+3	58:42
7	Sulfathiazole Forms 2 and 3	s3+2	63:37
8	Sulfathiazole Forms 2 and 4	s4+2	32:68
9	Carbamazepine Forms 1 and 3	c1+3	72:28
10	Sulfathiazole Forms 2, 3 and 4	s2+3+4	53:18:29
11	Sulfathiazole Form 2 and Carbamazepine Form 1	s2+c1	50:50
12	Sulfathiazole Form 3 and Carbamazepine Form 1	s3+c1	50:50
13	Sulfathiazole Form 4 and Carbamazepine Form 1	s4+c1	61:39
14	Sulfathiazole Form 2 and Carbamazepine Form 3	s2+c3	80:20
15	Sulfathiazole Form 3 and Carbamazepine Form 3	s3+c3	83:17
16	Sulfathiazole Form 4 and Carbamazepine Form 3	s4+c3	82:18

Table 25 – Additional Sixteen Samples

6.8 SIMULATED DATASET

6.8.1 SIMULATED DATA CLUSTERING

The simulated dataset from Chapter 5 was combined with the Chapter 6 simulated dataset to give a 32 sample simulated dataset. The dendrogram and MMDS plot for this are shown in Figure 159.

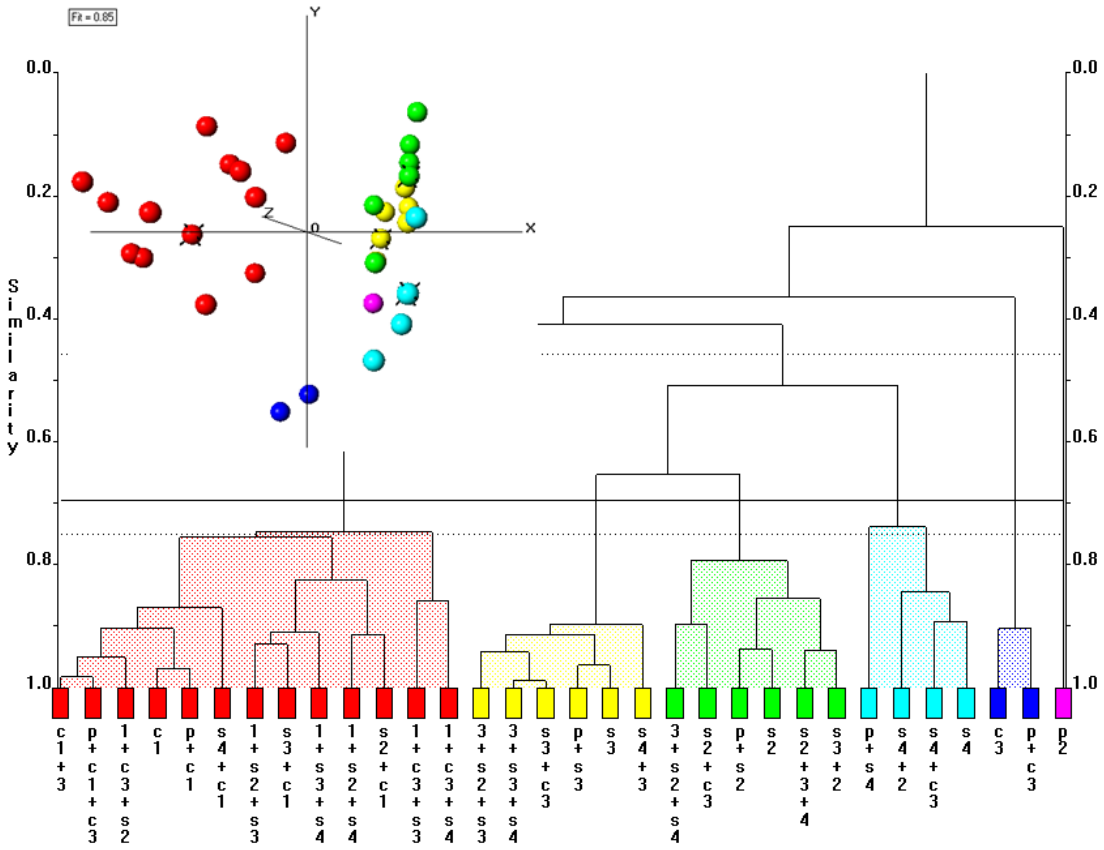

Figure 159 - Simulated 32 sample dataset

The red cluster contains sample c1+3, p+c1+c3, c1+c3+s2, c1, p+c1, s4+c1, c1+s2+s3, s3+c1, c1+s3+S4, c1+s2+s4, s2+c1, c1+c3+s3 and c1+c3+s4. The yellow cluster contains samples c3+s2+s3, c3+s3+s4, c3+c3, p+s3, s3 and s4+3. The green cluster contains samples c3+s2+s4, s2+c3, p+s2, s2, s2+3+4 and s3+2. The aquamarine cluster contains samples p+s4, s4+2, s4+c3 and s4. The blue cluster contains samples c3 and p+c3 while the purple cluster contains sample p2.

The silhouettes are shown in Figure 160.

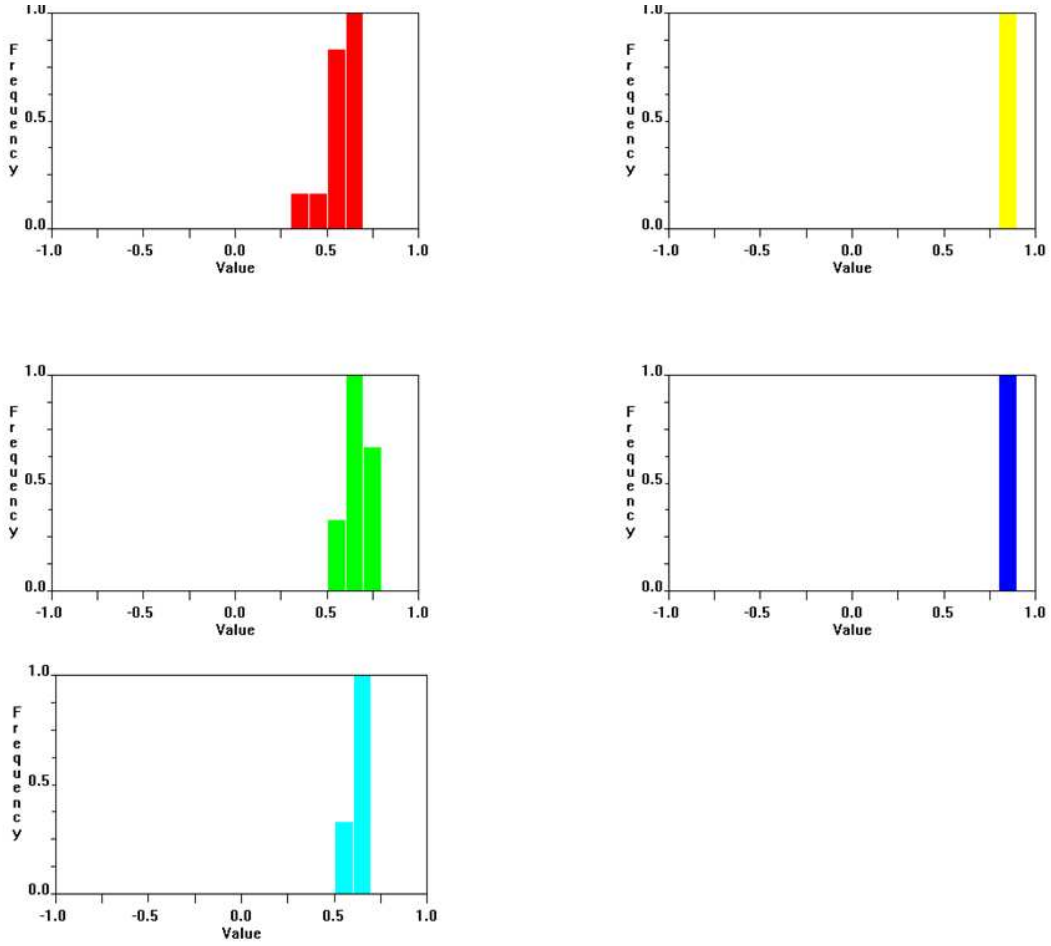

Figure 160 – 32 sample Predicted silhouettes

For the red cluster the lower bar, just above 0.25, contains sample c1+c3+s3. The second bar, just below 0.5, contains sample c1+s2+s3. The third bar, just above 0.5, contains samples c1+c3+s4, c3+s2+s4, c3+s3+s4, c1 and p+c1. For the yellow cluster all samples are present in one bar. For the green cluster the lower bar, just above 0.5, represents sample c3+s2+s4. The middle bar represents samples p+s2, s2+c3 and s3+2. The blue cluster has all samples in a single bar. For the aquamarine cluster the lower bar, just above 0.5, represents sample p+s4.

The following samples are predicted as clustering together

- c1+3, p+c1+c3, c1+c3+s2, c1, p+c1, s4+c1, c1+s2+s3, s3+c1, c1+s3+s4, c1+s2+s4, s2+c1, c1+c3+s3 and c1+c3+s4
- c3+s2+s3, c3+s3+s4, s3+c3, p+s3, s3 and s4+3
- c3+s2+s4, s2+c3, p+s2, s2, s2+3+4 and s3+2
- p+s4, s4+2, s4+c3 and s4
- c3 and p+c3
- p2

6.9 FINDING THE OPTIMAL CLUSTERING

The dataset was run through PolySNAP along with its ideal correlation coefficients, as outlined in Chapter 3.

Figure 161 shows the dendrogram and MMDS plot for the Pearson correlation matrix and Figure 164 the dendrogram and MMDS plot for the Spearman correlation matrix.

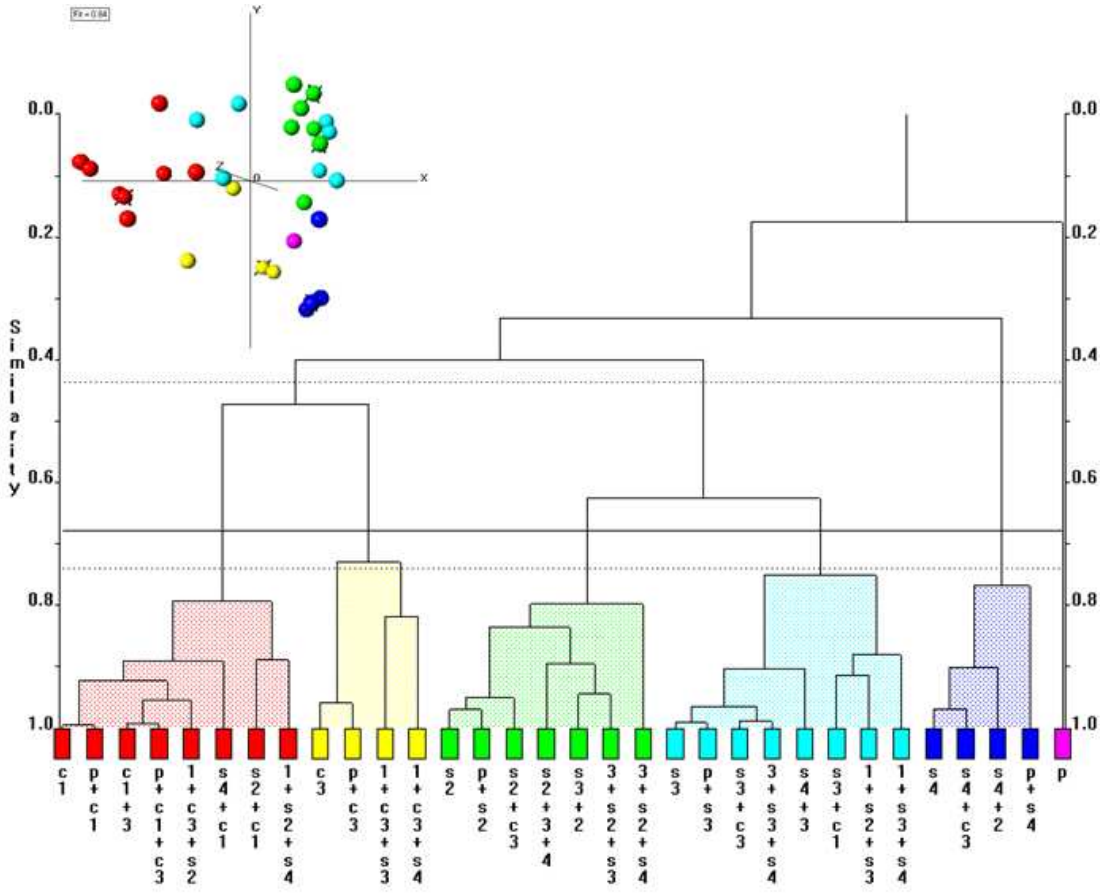

Figure 161 - Pearson correlation coefficient dendrogram and MMDS Plot

The red cluster contains samples c1, p+c1, c1+3, p+c1+c3, c1+c3+s2, s4+c1, s2+c1 and c1+s2+s4. The yellow cluster contains samples c1, p+c3, c1+c3+s3 and c1+c3+s4. The green cluster contains samples s2, p+s2, s2+c3, s2+3+4, s3+2, c3+s2+s3 and c3+s2+s4. The aquamarine cluster contains samples s2, p+s3, s3+c3, c3+s3+s4, s4+3, s3+c1, c1+s2+s3 and c1+s3+s4. The blue cluster contains samples s4, s4+c3, s4+2 and p+s4 and the purple cluster contains sample p.

The silhouettes for this dataset are shown in Figure 162.

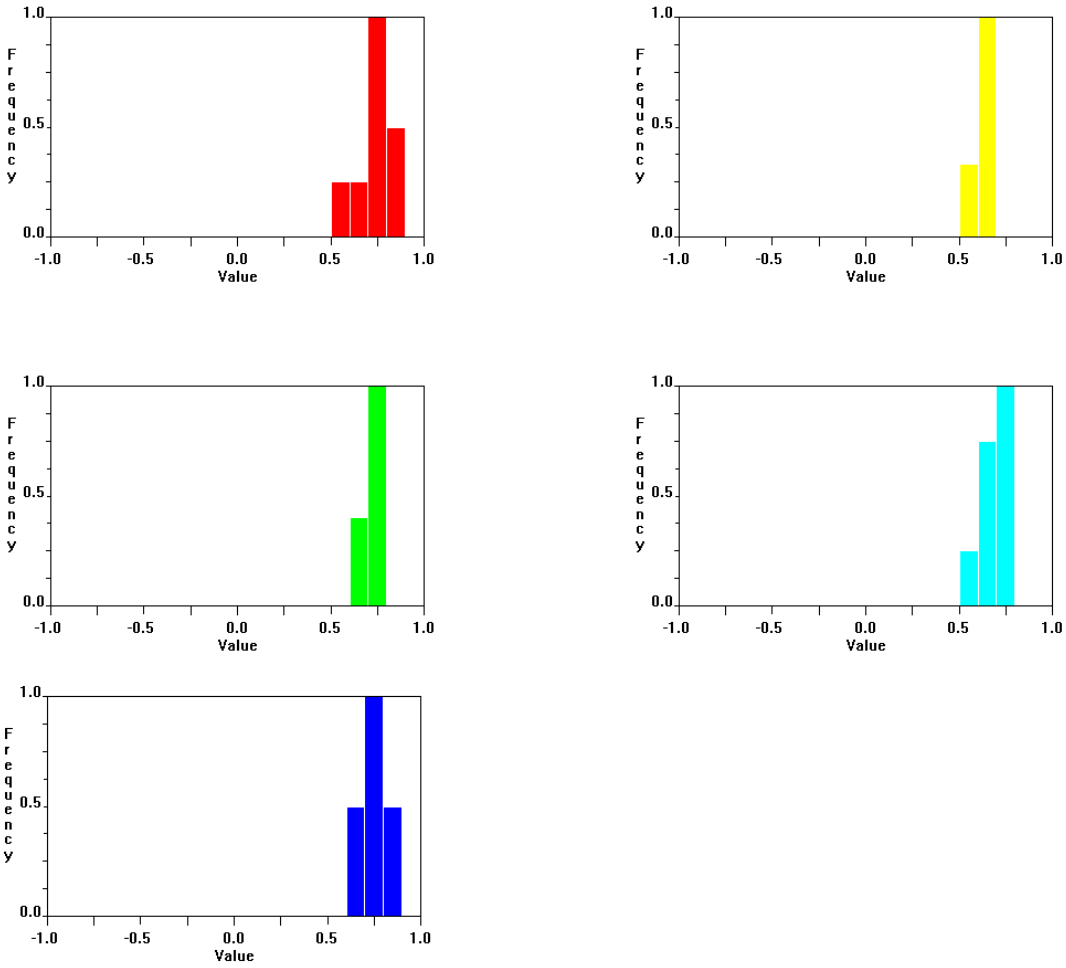

Figure 162 - Pearson correlation silhouettes

For the red cluster the lower bar, just above 0.5, represents sample c1+s2+s4. The next bar, just below 0.75, represents sample s2+c1. The third bar, at 0.75, represents samples c1, c1+3, s4+c1 and c1+c3+s2. The final bar represents samples p+c1 and p+c1+c3. For the yellow cluster the lower bar just above 0.5 represents sample c1+c3+s4. For the green cluster the lower bar, just below 0.75, represents samples c3+s2+s3 and c3+s2+s4. For the aquamarine cluster the lower bar, just above 0.5, represents sample s3+c1 while the middle bar, just below 0.75, represents samples s4+3, c1+s2+S3 and c1+s3+s4. For the blue

cluster the lower bar just below 0.75 represents sample p+s4 while the middle bar at 0.75 represents samples s4+2 and s4+c3.

The fuzzy clustering is shown in Figure 163 with the numerical results in Table 26.

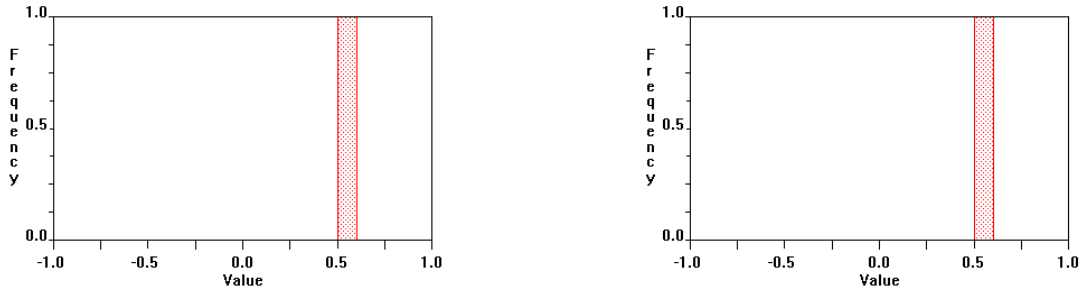

Figure 163 - Pearson Fuzzy Clustering derived from Correlation Coefficient

The first fuzzy clustering chart represents samples s3+c1 and c1+c3+s4. The second chart represents sample c3+s2+s3.

For the table of results, column 1 represents the purple cluster, column 2 the yellow cluster, column 3 the blue cluster, column 4 the green cluster, column 5 the aquamarine cluster and column 6 the red cluster. The first samples, s3+c1, could potentially appear in either the aquamarine or red cluster, the second sample, c1+c3+s3, could potentially appear in either the yellow or red cluster. The third sample, c3+s2+s3, could potentially appear in either the green or aquamarine cluster.

	1	2	3	4	5	6
c1	0.03	0.07	0	0	0.1	0.88
c3	0.05	0.85	0.06	0.08	0.11	0.05
s2	0.06	0	0.04	0.83	0.16	0.16
s3	0.06	0.09	0.04	0.28	0.83	0
s4	0.03	0.08	0.87	0.09	0.14	0.04
c1+3	0.05	0.26	0	0.01	0.15	0.84
s2+3+4	0.07	0.06	0.29	0.76	0.36	0.13
s2+c1	0.07	0.03	0.01	0.38	0.21	0.76
s2+c3	0.08	0.15	0.06	0.82	0.19	0.18
s3+2	0.07	0.04	0.05	0.77	0.44	0.11
s3+c1	0.06	0.13	0	0.15	0.67	0.55*
s3+c3	0.07	0.2	0.05	0.28	0.82	0
s4+2	0.05	0.05	0.79	0.37	0.21	0.11
s4+3	0.07	0.12	0.34	0.27	0.76	0
s4+c1	0.05	0.11	0.24	0.03	0.16	0.82
s4+c3	0.04	0.23	0.84	0.1	0.16	0.06
c1+c3+s2	0.07	0.32	0.03	0.2	0.2	0.77
c1+c3+s3	0.07	0.73	0.04	0.2	0.43	0.32
c1+c3+s4	0.06	0.67	0.3	0.09	0.18	0.52*
c1+s2+s3	0.08	0.08	0.03	0.41	0.69	0.38
c1+s2+s4	0.08	0.07	0.33	0.38	0.24	0.67
c1+s3+s4	0.07	0.15	0.29	0.17	0.64	0.48
c3+s2+s3	0.07	0.2	0.06	0.66	0.55*	0.08
c3+s2+s4	0.08	0.3	0.36	0.69	0.21	0.17
c3+s3+s4	0.07	0.27	0.13	0.27	0.79	0.02
p	0.85	0.07	0.12	0.1	0.01	0
p+c1	0.06	0.08	0	0	0.11	0.87
p+c1+c3	0.09	0.25	0	0.02	0.15	0.84
p+c3	0.15	0.83	0.1	0.11	0.1	0.04
p+s2	0.15	0	0.08	0.82	0.15	0.15
p+s3	0.11	0.1	0.06	0.29	0.82	0
p+s4	0.24	0.12	0.77	0.14	0.11	0.03

Table 26 - Pearson Fuzzy Clustering derived from Pearson coefficient

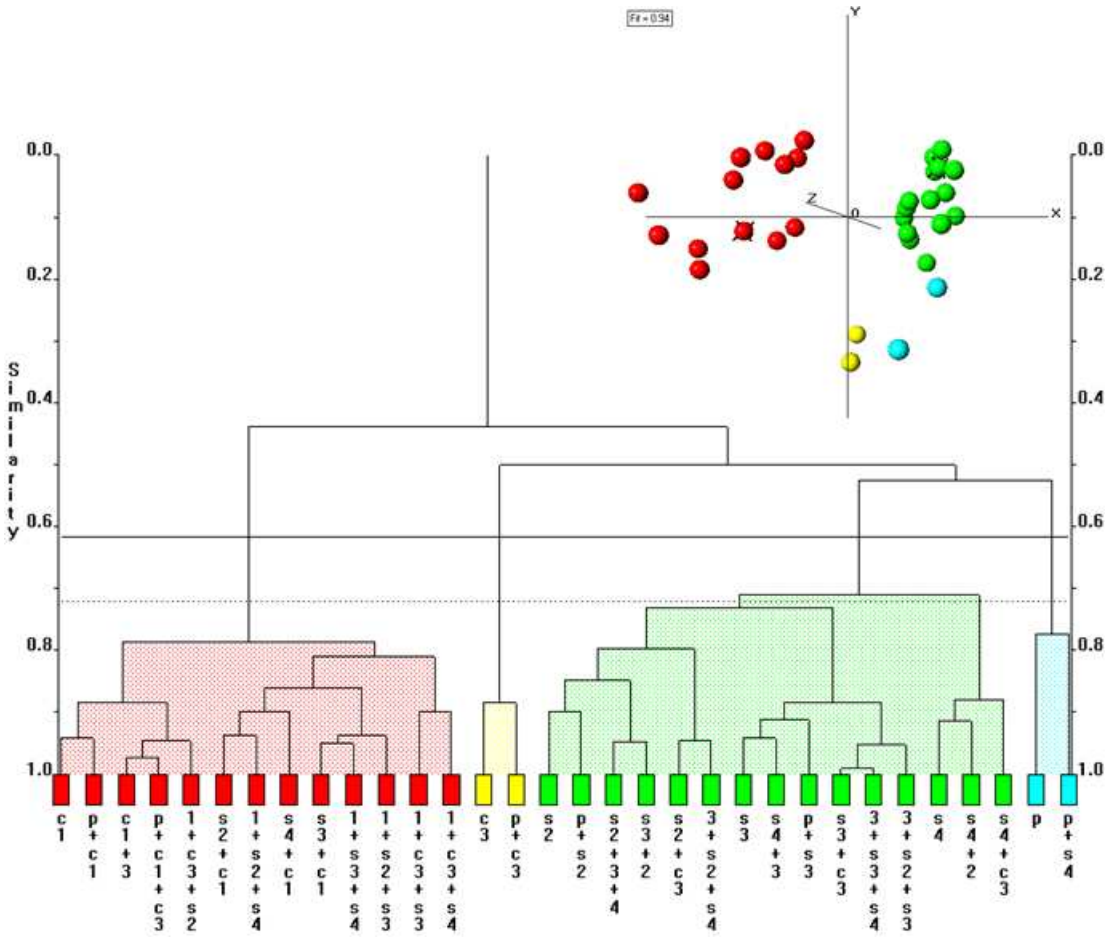

Figure 164 - Spearman correlation coefficient dendrogram and MMDS plot

For the spearman correlation, the red cluster contains samples c1, p+c1, c1+3, p+c1+c3, c1+c3+s2, s2+c1, c1+s2+s4, s4+c1, s3+c1, c1+s3+s4, c1+s2+s3, c1+c3+s3 and c1+c3+s4. The yellow cluster contains samples c3 and p+c3. The green cluster contains samples s2, p+s2, s2+3+4, s3+2, s2+c3, c3+s2+s4, s3, s4+3, p+s3, s3+c3, c3+s3+s4, c3+s2+s3, s4, s4+2 and s4+c3. The aquamarine cluster contains samples p and p+s4.

The silhouettes are shown in Figure 165.

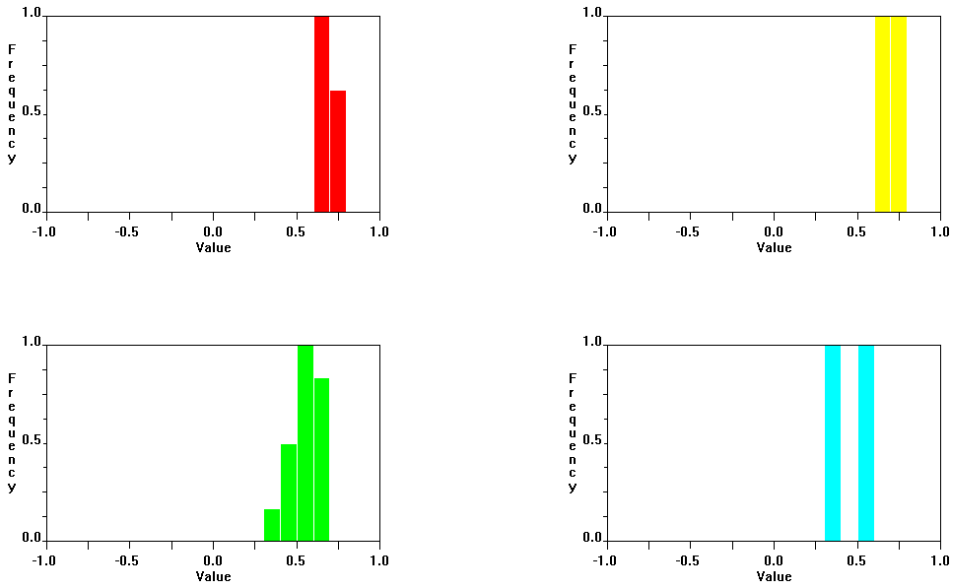

Figure 165 - Spearman correlation silhouettes derived from spearman correlation

For the red cluster the lower bar, just below 0.75, contains samples c1, c1+c3+s3, c1+c3+s4, c1+s2+s3, c1+s2+s4, c1+s3+s4, p+c1 and p+c1+c3. For the yellow cluster the lower bar, just below 0.75, contains sample p+c3. For the green cluster the lower bar just above 0.25 contains sample p+s2, the second bar, just below 0.5 contains samples s2, s4 and s4+c3. The third bar, just above 0.5, contains samples s3, s2+c3, s3+c3, c3+s2+s4, c3+s3+s4 and p+s3. For the aquamarine cluster the lower bar, just above 0.25, contains sample p+s4. The fuzzy clustering is shown in Figure 166 with the numerical results in Table 27.

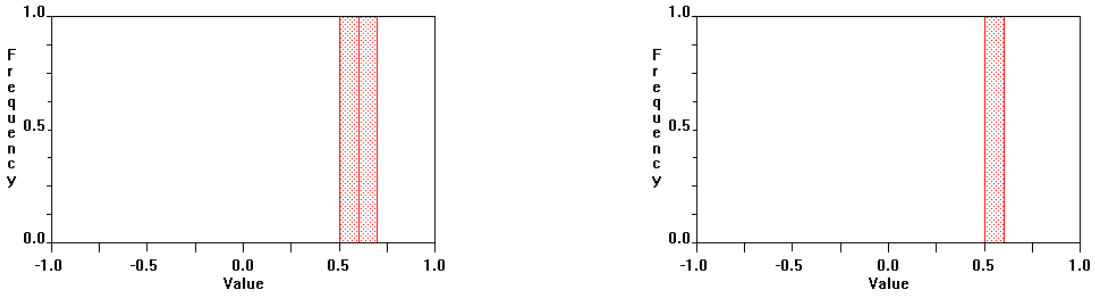

Figure 166 - Spearman fuzzy clustering derived from spearman correlation

	1	2	3	4	
c1	0.46	0.55*	0.1	0	<==
c3	0	0.59	0.19	0.26	
s2	0.07	0.02	0.19	0.77	
s3	0.22	0.11	0.37	0.56*	<==
s4	0.01	0.05	0.77	0.33	
c1+3	0.44	0.63*	0.16	0.07	<==
s2+3+4	0.1	0.09	0.32	0.78	
s2+c1	0.45	0.49	0.2	0.26	
s2+c3	0.07	0.22	0.22	0.73	
s3+2	0.12	0.06	0.23	0.79	
s3+c1	0.46	0.5	0.26	0.26	
s3+c3	0.21	0.25	0.39	0.54*	<==
s4+2	0.04	0.05	0.72	0.46	
s4+3	0.2	0.13	0.47	0.55*	<==
s4+c1	0.45	0.52*	0.28	0.13	<==
s4+c3	0.01	0.19	0.74	0.32	
c1+c3+s2	0.41	0.61*	0.21	0.2	<==
c1+c3+s3	0.19	0.75	0.23	0.3	
c1+c3+s4	0.2	0.76	0.27	0.21	
c1+s2+s3	0.41	0.41	0.29	0.42	
c1+s2+s4	0.41	0.45	0.33	0.32	
c1+s3+s4	0.42	0.45	0.36	0.31	
c3+s2+s3	0.13	0.23	0.27	0.75	
c3+s2+s4	0.06	0.24	0.31	0.69	
c3+s3+s4	0.2	0.28	0.41	0.54*	<==
p	0.55	0	0.17	0.23	
p+c1	0.49	0.57*	0.14	0.02	<==
p+c1+c3	0.45	0.61*	0.17	0.08	<==
p+c3	0.01	0.57	0.24	0.27	
p+s2	0.12	0.02	0.19	0.72	
p+s3	0.24	0.11	0.37	0.55*	<==
p+s4	0.1	0.04	0.7	0.28	

Table 27 - Spearman fuzzy clustering results

For the first plot the lower bar, just above 0.5, contains samples c1, s4+c1 and p+c1. The upper bar contains samples c1+3, c1+c3+s2 and p+c1+c3. For the second chart the single bar represents samples s3, s3+c3, s4+3, c3+s3+s4 and p+s3.

For the table of results, column 1 represents the aquamarine cluster, column 2 the red cluster, column 3 the green cluster and column 4 the yellow cluster.

Sample c1 and s4+c1 and p+c1 show a low correlation with the red cluster, just above 0.5 and a low correlation with the aquamarine cluster, just below 0.5. Samples c1+3, c1+c3+s2 and p+c1+c3 show a low correlation with the red cluster and a low correlation with the aqua cluster. Samples s3, s3+c3, s4+3, c3+s3+s4 and p+s3 show a low correlation with the yellow cluster.

The Pearson dendrogram again closely matches that of the predicted result.

6.10 THIRTY-TWO SAMPLE DATASET CLUSTERING

6.10.1 EXPECTED CLUSTERING

The ideal powder patterns were combined with the ideal powder patterns from Chapter 5 and run through PolySNAP to determine the expected clustering. The dendrogram and MMDS plot for the expected clustering are shown in Figure 167.

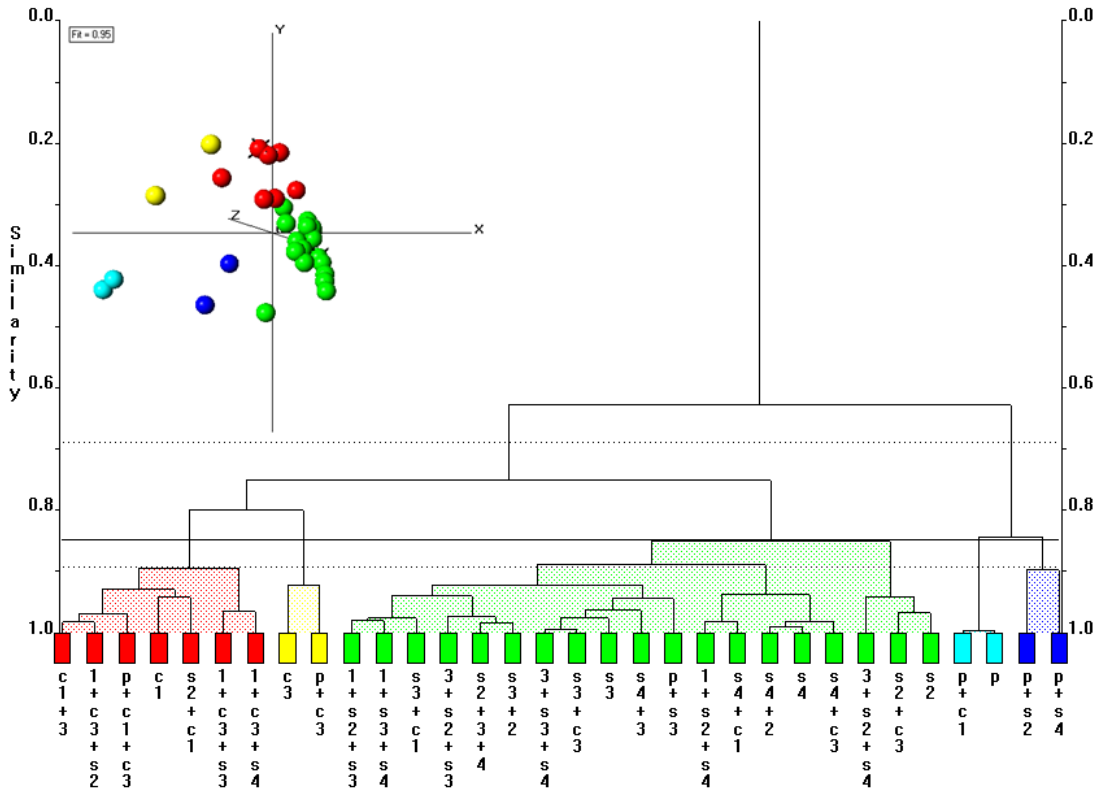

Figure 167 – Expected Clustering Dendrogram and MMDS Plot

The red cluster contains patterns c1+3, c1+c3+s2, p+c1+c3, c1, s2+c1, c1+c3+s3 and c1+c3+s4. The yellow cluster contains patterns c3 and p+c3. The green cluster contains patterns c1+s2+s3, c1+s3+s4, s3+c1, c3+s2+s3, s2+3+4, s3+2, c3+s3+s4, s3+c3, s3, s4+3,

p+s3, c1+s2+s4, s4+c1, s4+2, s4, s4+c3, c3+s2+s4, s2+c3 and s2. The aquamarine cluster contains patterns p+c1 and p. The blue cluster contains pattern p+s2 and p+s4. The large green cluster has several tie-bars lying close to the cut-level. As such the silhouettes and fuzzy clustering for this dataset will be analysed. The silhouettes are shown in Figure 168 and the fuzzy clustering in Figure 169 and Table 27.

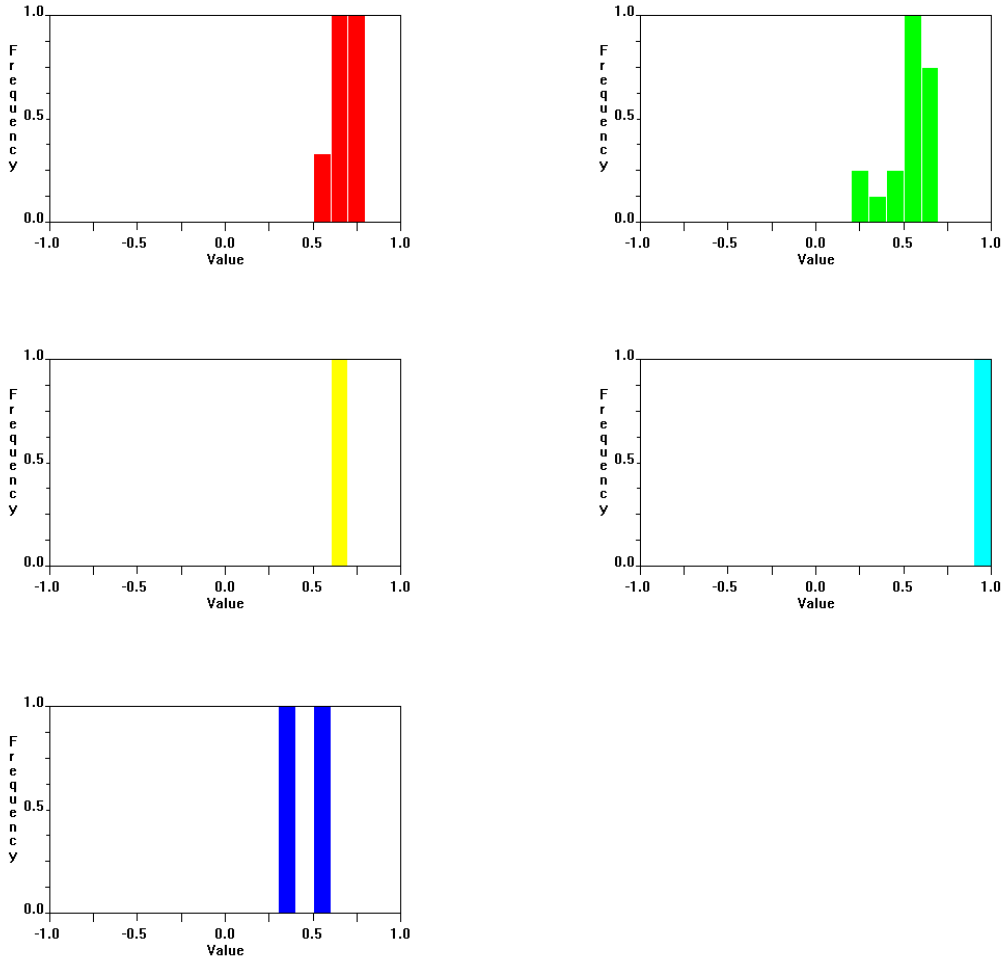

Figure 168 – Silhouettes

The red cluster has no samples below 0.5. The green cluster has three regions below 0.5. The lowest of these bands, at 0.25, contains patterns s2+c3 and s2. The next band contains pattern p+s3 and the third, at just below 0.5, contains patterns s3+c1 and s2. The yellow and aquamarine clusters have no patterns below 0.5. The blue cluster has one band below 0.5, containing pattern p+s4.

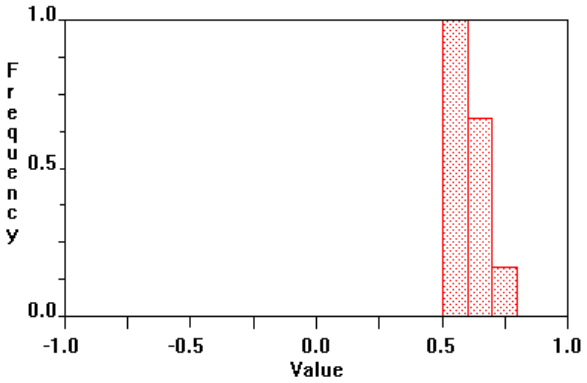

Figure 169 - Fuzzy Clustering

	1	2	3	4	5	
c1+3	0.16	0.22	0.94	0.54*	0.16	<==
c1+c3+s2	0.17	0.23	0.93	0.58*	0.15	<==
c1+c3+s3	0.18	0.24	0.87	0.68*	0.17	<==
c1+c3+s4	0.19	0.22	0.88	0.70*	0.16	<==
c1+s2+s3	0.2	0.18	0.4	0.99	0.16	
c1+s2+s4	0.2	0.18	0.44	0.97	0.16	
c1+s3+s4	0.2	0.18	0.4	1	0.16	
c1	0.15	0.17	0.92	0.52*	0.14	<==
c3+s2+s3	0.21	0.21	0.39	1	0.17	
c3+s2+s4	0.22	0.23	0.41	0.97	0.17	
c3+s3+s4	0.19	0.19	0.33	1	0.16	
c3	0.17	0.88	0.45	0.52*	0.18	<==
p+c1+c3	0.2	0.23	0.93	0.53*	0.22	<==
p+c1	0.26	0.22	0.26	0.38	0.93	
p+c3	0.24	0.89	0.41	0.50*	0.28	<==
p+s2	0.86	0.21	0.36	0.67*	0.27	<==
p+s3	0.21	0.16	0.24	0.97	0.21	
p+s4	0.87	0.22	0.29	0.62*	0.31	<==
p	0.26	0.21	0.24	0.35	0.93	
s2+3+4	0.22	0.17	0.35	1	0.16	
s2+c1	0.18	0.17	0.89	0.66*	0.15	<==
s2+c3	0.2	0.22	0.41	0.93	0.16	
s2	0.19	0.16	0.34	0.94	0.14	
s3+2	0.21	0.17	0.35	1	0.16	
s3+c1	0.17	0.17	0.41	0.97	0.15	
s3+c3	0.17	0.17	0.3	0.99	0.15	
s2	0.16	0.13	0.24	1	0.14	
s4+2	0.2	0.16	0.31	1	0.15	
s4+3	0.19	0.15	0.29	1	0.15	
s4+c1	0.19	0.18	0.41	0.98	0.15	
s4+c3	0.2	0.19	0.33	0.99	0.16	
s4	0.19	0.15	0.27	1	0.14	

Table 28 - Fuzzy Clustering Numerical Values

Cluster 1 is the blue cluster, cluster 2 is the yellow cluster, cluster 3 is the red cluster, cluster 4 is the green cluster and cluster 5 is the aquamarine cluster.
The fuzzy clustering plot contains three regions. The first of these regions corresponds to patterns c1+3, c1+c3+s2, c1, c3, p+c1+c3 and p+c3. These patterns could be clustered as follows:

- Pattern c1+3 could potentially be in the red or green cluster
- Pattern c1+c3+s2 could potentially be in the red or green cluster
- Pattern c1 could potentially be in the red or green cluster
- Pattern p+c1+c3 could potentially be in the red or green cluster
- Pattern p+c3 could potentially be in the yellow or green cluster.

The second region corresponds to patterns c1+c3+s3, p+s2, p+s4 and s2+c1. These patterns could be clustered as follows:

- Pattern c1+c3+s3 could potentially be in the red or green cluster
- Pattern p+s2 could potentially be in the blue or green cluster
- Pattern p+s4 could potentially be in the blue or green cluster
- Pattern s2+c1 could potentially be in the red or green cluster

The third region corresponds to pattern c1+c3+s4. This pattern could potentially be in either the red or green cluster.
The cut-level was adjusted downwards and produced the dendrogram and MMDS plot shown in Figure 170.

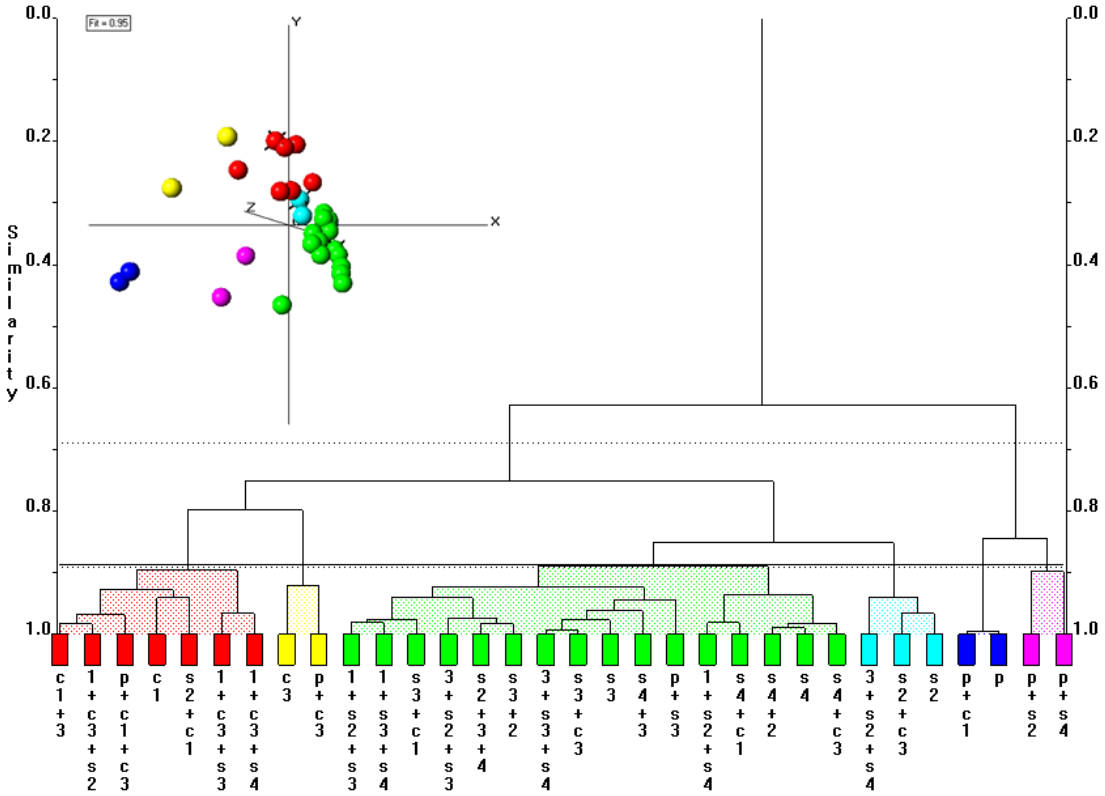

Figure 170 - Expected Clustering Dendrogram and MMDS Plot with Adjusted Cut-Level

The predicted clustering had the following samples clustered together.

- c1+3, p+c1+c3, c1+c3+s2, c1, p+c1, s4+c1, c1+s2+s3, s3+c1, c1+s3+s4, c1+s2+s4, s2+c1, c1+c3+s3 and c1+c3+s4
- c3+s2+s3, c3+s3+s4, s3+c3, p+s3, s3 and s4+3
- c3+s2+s4, s2+c3, p+s2, s2, s2+3+4 and s3+2
- p+s4, s4+2, s4+c3 and s4
- c3 and p+c3
- p

The expected clustering has the following samples clustered together

- c1+3, c1+c3+s2, p+c1+c3, c1, s2+c1, c1+c3+s3 and c1+c3+s4
- c3 and p+c3
- c1+s2+s3, c1+s3+s4, s3+c1, c3+s2+s3, s2+3+4, s3+2, c3+s3+s4, s3+c3, s3, s4+3, p+s3, c1+s2+s4, s4+c1, s4+2, s4, s4+c3
- c3+s2+s4, s2+c3 and s2
- p+c1 and p
- p+s2 and p+s4

As the predicted clustering closely matches that of the Pearson dendrogram it shall be used as the optimal method. The following clustering is expected.

1) A cluster containing samples c1+3, p+c1+c3, c1+c3+s2, c1, p+c1, s4+c1, c1+s2+s3, s3+c1, c1+s3+s4, c1+s2+s4, s2+c1, c1+c3+s3 and c1+c3+s4
2) A cluster containing samples c3+s2+s3, c3+s3+s4, s3+c3, p+s3, s3 and s4+3
3) A cluster containing samples c3+s2+s4, s2+c3, p+s2, s2, s2+3+4 and s3+2
4) A cluster containing samples p+s4, s4+2, s4+c3 and s4
5) A cluster containing samples c3 and p+c3
6) A cluster containing sample p

6.10.2 PXRD DATA

The PXRD data were run with no pre-processing applied. The dendrogram and MMDS plot are shown in Figure 171.

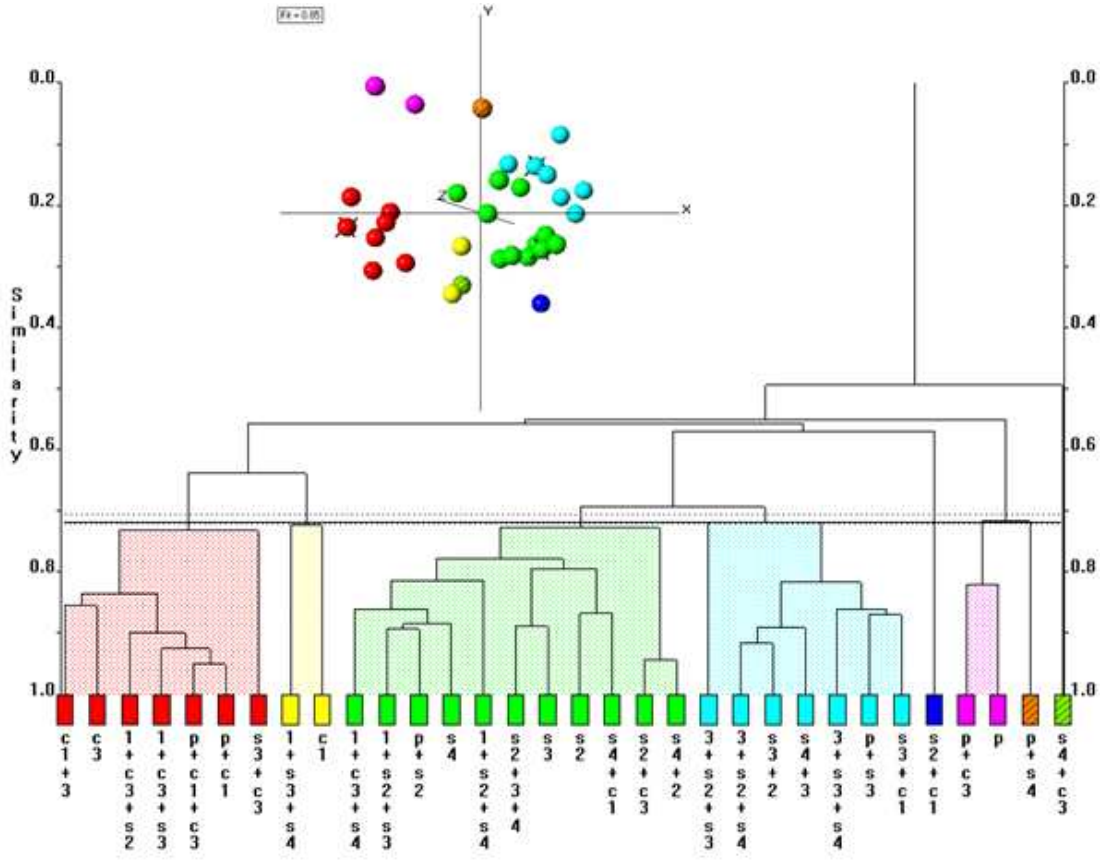

Figure 171 – Thirty-Two Sample PXRD Dendrogram and MMDS Plot

The dendrogram does not show the expected clustering. The red cluster contains patterns c1+3, c1+c3+s2, p+c1+c3 and c1+c3+s3 from expected cluster 1, c3 from expected cluster 5 and s3+c3 from expected cluster 2. The yellow cluster contains samples c1+s3+s4 and c1 from expected cluster 1. The green cluster contains samples c1+c3+s4, s4+c1, c1+s2+s3 and c1+s2+s4 from expected cluster 1, s4 and s4+2 from expected cluster 4 and s2+3+4, p+s2, s2 and s2+c3 from expected cluster 3 and s3 from expected cluster 2. The aquamarine cluster contains samples c3+s2+s3, s4+3, p+s3, c3+s3+s4 and c3+s2+s4 from expected cluster 2, s3+c1 from expected cluster 1 and s3+2 from expected cluster 3. The blue cluster contains samples s2+c1 from expected cluster 1. The purple cluster contains p+c3 from expected cluster 5 and p from expected cluster 6. The striped brown cluster contains p+s4 from expected cluster 4 and the striped green cluster contains s4+c3 from expected cluster 4.

The MMDS plot does not show clear separation between the clusters, in particular between the aquamarine and green clusters. The dendrogram has a score of 0.59, with more than half of the samples being misclustered.

The silhouettes for this dataset are shown in Figure 172 and the fuzzy clustering in Figure 173 with its numeric data in Table 29.

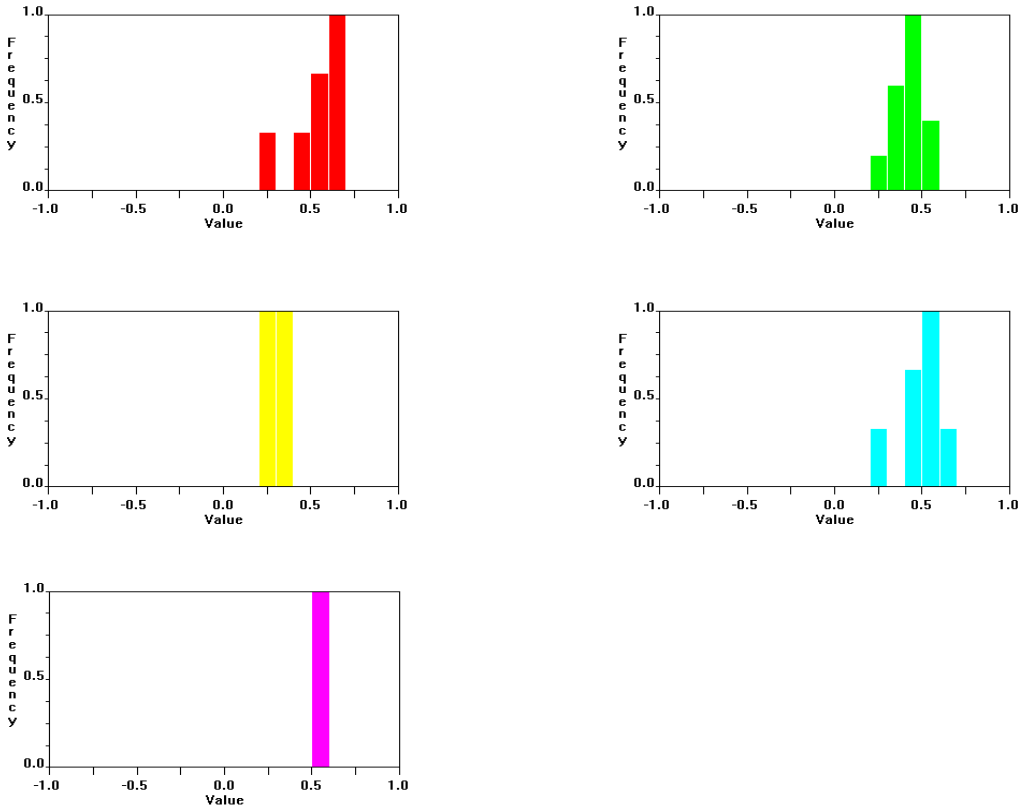

Figure 172 - Silhouettes for PXRD Data

None of the clusters have any patterns with silhouette values above 0.75.
The red cluster has two patterns with values below 0.5. The lowest one is sample s3+c3 and the upper one is c3. The first band above 0.5 contains c1+c3 and c1+c3+s2. The upper most band contains the remaining samples. The lowest band on the green cluster contains only the s4+2. The next band up contains three samples; s2+3+4, s2, c1+s2+s4. The uppermost band, the only band above 0.5, represents two samples; s4 and c1+s2+s3. The remaining band, just below 0.5, contains the remaining samples. For the yellow cluster, both of the samples are below 0.5. The lower one represents c1+s3+s4 while the upper band represents c1. The aquamarine cluster has three samples appearing below 0.5. The lowest of these bands contains just c3+s2+s3 with the next band containingc3+s3+s4 and s4+3. The uppermost band contains the p+s3. The remaining patterns lie in the large band below this one. The purple cluster contains both of its patterns in one band that lies just over 0.5. None of the remaining clusters contain enough patterns for silhouettes to be generated for them.

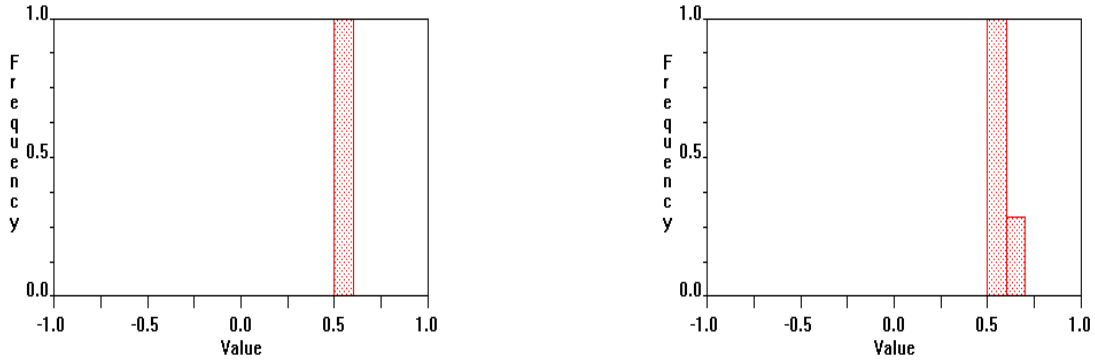

Figure 173 - Fuzzy Clustering for PXRD Data

	1	2	3	4	5	6	7	8	
c1+3	0.12	0.12	0.1	0.19	0.16	0.24	0.41	1	
c1+c3+s2	0.12	0.11	0.11	0.17	0.16	0.26	0.42	1	
c1+c3+s3	0.13	0.12	0.13	0.18	0.19	0.32	0.47	1	
c1+c3+s4	0.13	0.12	0.14	0.17	0.16	0.38	0.99	0.37	
c1+s2+s3	0.13	0.14	0.17	0.19	0.18	0.5	1	0.34	
c1+s2+s4	0.13	0.13	0.16	0.18	0.2	0.42	0.95	0.5	<==
c1+s3+s4	0.14	0.15	0.14	0.9	0.17	0.39	0.61*	0.44	<==
c1	0.13	0.14	0.11	0.9	0.15	0.3	0.52*	0.4	<==
c3+s2+s3	0.12	0.12	0.13	0.16	0.19	0.94	0.52*	0.32	<==
c3+s2+s4	0.12	0.13	0.15	0.17	0.16	0.99	0.61*	0.29	<==
c3+s3+s4	0.11	0.11	0.12	0.13	0.16	0.99	0.42	0.22	
c3	0.13	0.13	0.11	0.19	0.16	0.3	0.48	1	
p+c1+c3	0.13	0.11	0.14	0.18	0.21	0.27	0.43	1	
p+c1	0.13	0.11	0.12	0.19	0.19	0.25	0.41	1	
p+c3	0.12	0.11	0.16	0.16	0.92	0.38	0.44	0.42	
p+s2	0.12	0.12	0.16	0.17	0.18	0.39	0.99	0.31	
p+s3	0.12	0.13	0.15	0.17	0.18	0.99	0.58*	0.31	<==
p+s4	0.12	0.12	0.88	0.16	0.24	0.43	0.59*	0.34	<==
p	0.11	0.1	0.16	0.15	0.92	0.28	0.39	0.38	
s2+3+4	0.12	0.15	0.13	0.18	0.13	0.38	0.99	0.3	
s2+c1	0.13	0.88	0.11	0.17	0.11	0.35	0.52*	0.28	<==
s2+c3	0.13	0.13	0.12	0.17	0.13	0.34	1	0.3	
s2	0.12	0.13	0.12	0.19	0.14	0.4	0.98	0.28	
s3+2	0.11	0.13	0.13	0.14	0.12	0.99	0.5	0.22	
s3+c1	0.12	0.13	0.13	0.16	0.15	0.99	0.53*	0.27	<==
s3+c3	0.12	0.1	0.12	0.14	0.15	0.26	0.38	0.98	
s3	0.12	0.13	0.14	0.17	0.12	0.38	1	0.24	
s4+2	0.11	0.11	0.1	0.13	0.1	0.31	0.98	0.18	
s4+3	0.12	0.13	0.13	0.15	0.12	0.97	0.56*	0.24	<==
s4+c1	0.14	0.15	0.14	0.22	0.15	0.42	0.99	0.4	
s4+c3	0.88	0.12	0.11	0.16	0.13	0.29	0.46	0.35	
s4	0.14	0.15	0.14	0.18	0.13	0.41	1	0.28	

Table 29 – PXRD Fuzzy Clusters Numeric Data

For this dataset cluster 1 is the striped green cluster, cluster 2 is the blue cluster, cluster 3 is the striped brown cluster, cluster 4 is the yellow cluster, cluster 5 is the purple cluster, cluster 6 is the aquamarine cluster, cluster 7 is the green cluster and cluster 8 is the red cluster.

In the first plot the bar at 0.5 contains only sample 6. This could be present in either the green or red cluster.

In the second plot, the upper most band, lying just below 0.75, contains two samples, c1+s3+s4, c3+s2+s4. Both of these samples are present in the aquamarine cluster.

- Sample c1+s3+s4 could potentially be in either the yellow or green cluster
- Sample c3+s2+s4 could potentially be in either the aquamarine or green cluster

The lower band, lying just above 0.5, contains pattern c1 from the yellow cluster, patterns c3+s2+s3, p+s3, s3+c1 and s4+3 from the aquamarine cluster and p+s4 in the striped brown cluster.

- Sample c1 could potentially be in yellow or green cluster
- Sample c3+s2+s3 could potentially be in the aquamarine or green cluster
- Sample p+s3 could potentially be in the aquamarine or green cluster
- Sample s3+c1 could potentially be in the aquamarine or green cluster
- Sample s4+3 could potentially be in the aquamarine or green cluster
- Sample p+s4 could potentially be in the striped brown or green cluster

Overall the PXRD dataset was not well clustered.

6.10.3 RAMAN DATA

The dendrogram and MMDS plot for the Raman data are shown in Figure 174. The Raman data was collected over the ranges determined in section 6.3.

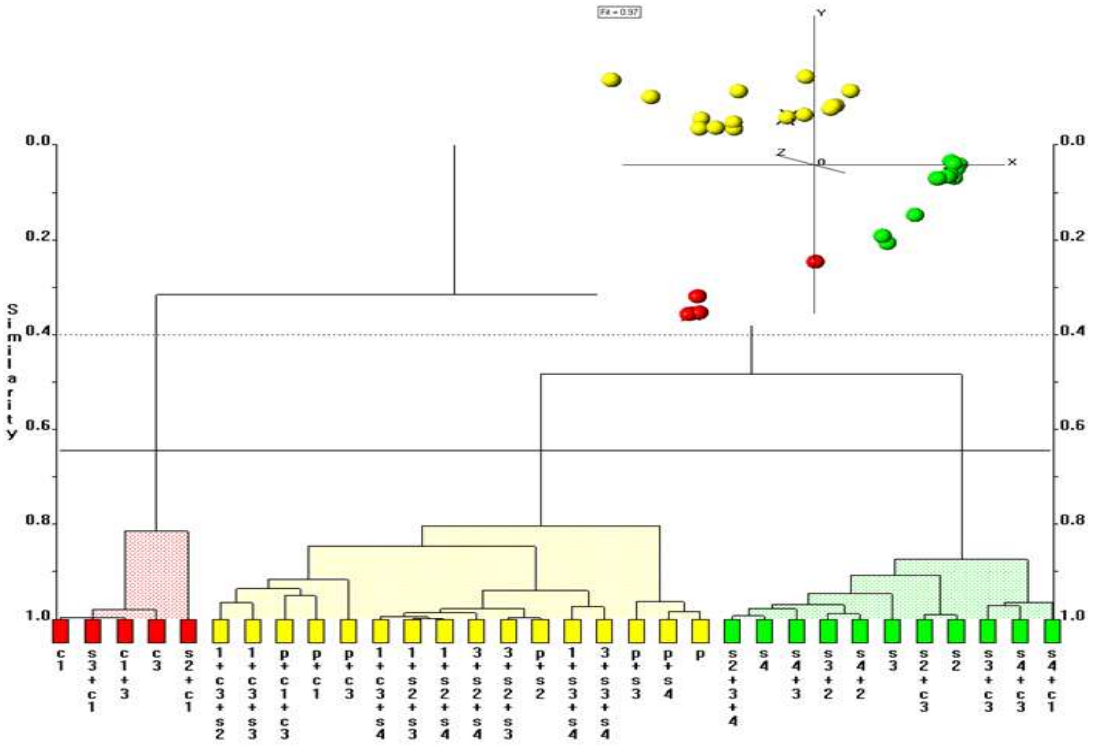

Figure 174 – Thirty-Two Sample Raman Dendrogram and MMDS Plot

The red cluster contains samples c1, s2+c1, s3+c1 and c1+3 from expected cluster 1 and sample s3 from expected cluster 2.
The yellow cluster contains samples c1+c3+s2, p+c1, c1+s2+s3, c1+s3+s4, c1+s2+s4, c1+c3+s3, p+c1+c3 and c1+c3+s4 from expected cluster 1, sample p which makes up expected cluster 6, sample p+s2 which makes up part of expected cluster 3 and p+s4 which makes up part of expected cluster 4 and samples c3+s2+s3, c3+s3+s4 and c3+s2+s4 from expected cluster 2.
The green cluster contains samples s2+3+4, s3 and s2+c3 from expected cluster 3, s4, s3+2, s4+2, s2 and s4+c3 from expected cluster 4, s4+3, s3+c3 from expected cluster 2 and s4+c1 from expected cluster 1.
The MMDS plot shows the clusters to be clearly separated from one another. The clusters are however rather diffuse. The Raman dendrogram has a score of 0.69, a particularly poor result with over two thirds of the samples being misclustered.

The silhouettes for the Raman dataset are shown in Figure 175.

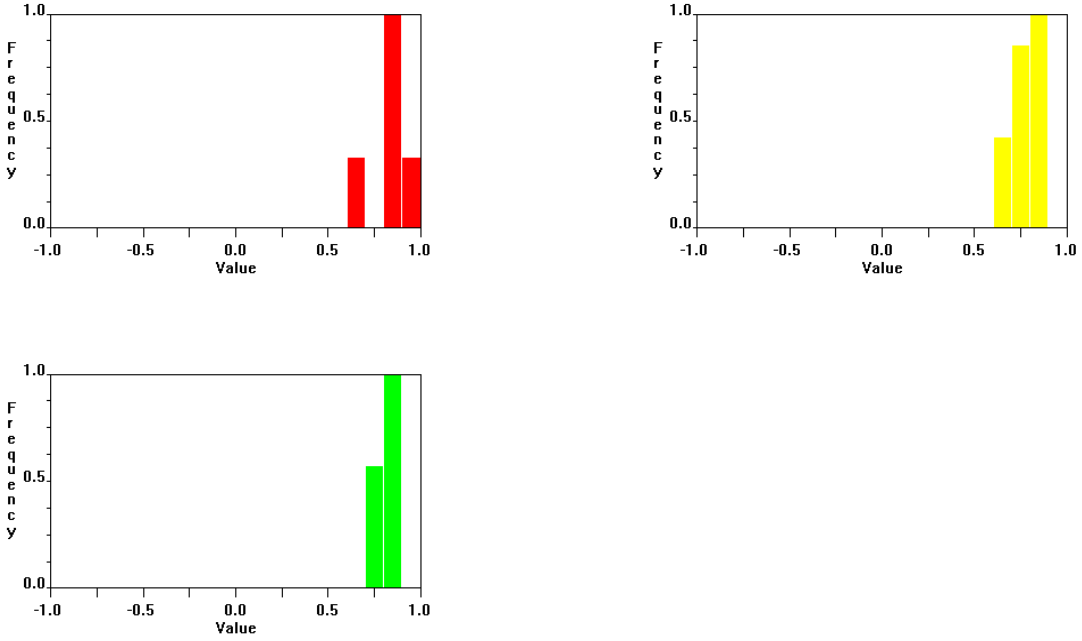

Figure 175 - Silhouettes for Raman Data

Compared to the PXRD data run, the silhouettes do not have any samples below 0.5. In the red cluster, the lowest samples, and only one below 0.75 is s2+c1, which is the pattern that appears a large distance from the remainder of the cluster.

For the yellow cluster, three of the patterns are below 0.75. These are p+s3, p and p+c1.

The green cluster has no samples below 0.75.

The fuzzy clustering is shown in Figure 176. The numeric data for fuzzy clustering is shown in Table 30.

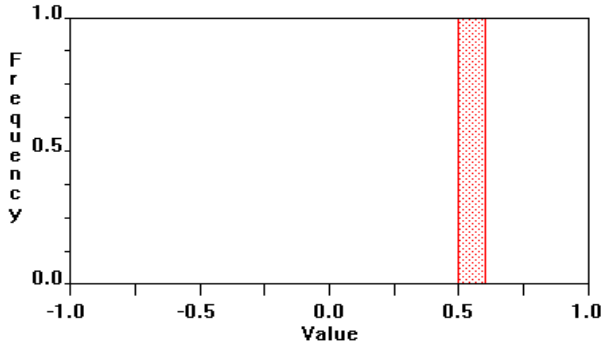

Figure 176 - Fuzzy Clustering for Raman Data

	1	2	3	
c1+3	0.02	0.83	0.09	
c1+c3+s2	0.09	0.1	0.8	
c1+c3+s3	0.08	0.19	0.76	
c1+c3+s4	0.44	0.09	0.71	
c1+s2+s3	0.44	0.08	0.71	
c1+s2+s4	0.44	0.08	0.71	
c1+s3+s4	0.25	0.14	0.75	
c1	0.02	0.82	0.09	
c3+s2+s3	0.51*	0.02	0.67	<==
c3+s2+s4	0.45	0.07	0.71	
c3+s3+s4	0.31	0.12	0.73	
c3	0.06	0.81	0.13	
p+c1+c3	0.02	0.24	0.76	
p+c1	0.09	0.21	0.72	
p+c3	0	0.21	0.78	
p+s2	0.28	0.02	0.74	
p+s3	0	0.12	0.73	
p+s4	0	0.16	0.75	
p	0.07	0.22	0.72	
s2+3+4	0.79	0.08	0.19	
s2+c1	0.38	0.67	0.21	
s2+c3	0.75	0.05	0.28	
s2	0.77	0.05	0.26	
s3+2	0.79	0.05	0.22	
s3+c1	0.04	0.82	0.09	
s3+c3	0.68	0.35	0.18	
s3	0.76	0.1	0.18	
s4+2	0.8	0.06	0.22	
s4+3	0.78	0.08	0.17	
s4+c1	0.76	0.24	0.19	
s4+c3	0.68	0.38	0.17	
s4	0.79	0.08	0.19	

Table 30– Raman Fuzzy Clusters Numeric Data

Cluster 1 is the green cluster, cluster 2 is the red cluster and cluster 3 is the yellow cluster. Samples c3+s2+s3 is the only sample represented in the fuzzy clustering. This sample could, according to the numerical data, potentially belong to either green or yellow cluster.

6.10.4 DSC DATA

The dendrogram and MMDS plot for the DSC data are shown in Figure 177.

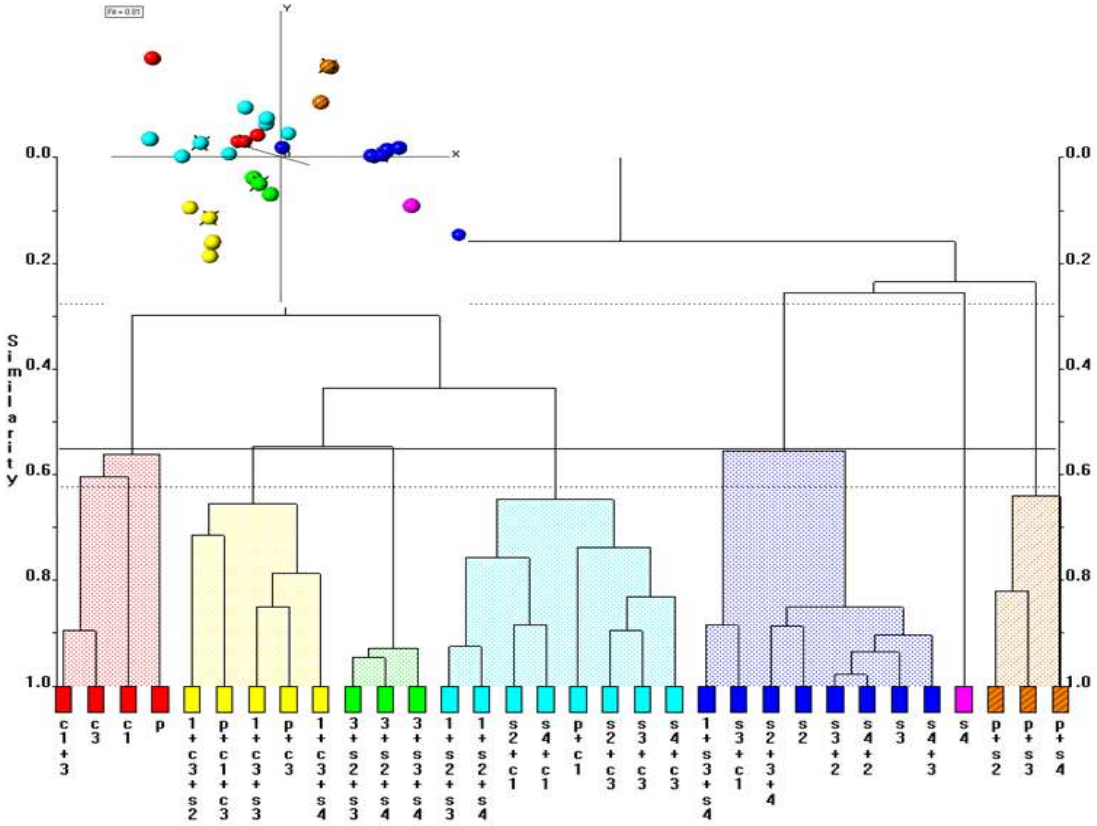

Figure 177 - Thirty-Two Sample DSC Dendrogram and MMDS Plot

The red cluster contains samples c1 and c1+3 from expected cluster 1, sample p from expected cluster 6 and c3 from expected cluster 5. The yellow cluster contains samples c1+c3+s2, c1+c3+s3 and c1+c3+s4 from expected cluster 1 and p+c3 from expected cluster 5. The green cluster contains samples c3+s2+s3, c3+s2+s4 and c3+s3+s4 from expected cluster 2. The aquamarine cluster contains samples c1+s2+s3, p+c1, c1+s2+s4, s2+c1 and s4+c1 from expected cluster 1, s3+c3 from expected cluster 2 and s4+c3 from expected cluster 4 and s2+c3 from expected cluster 3. The blue cluster contains samples c1+s3+s4 and s3+c1 from expected cluster 1, s2+3+4 and s3+2 from expected cluster 3, s4+2 and s2 from expected cluster 4 and s4+3 and s3 from expected cluster 2. The purple cluster contains sample s4 from expected cluster 4. The striped brown cluster contains samples p+s2 from expected cluster 3 and p+s4 from expected cluster 4 and p+s3 from expected cluster 2. The DSC dendrogram has a score of 0.66, showing that two thirds of the samples are not clustered as expected.

The MMDS plot does not show clear separation of the different clusters.

The silhouettes are shown in Figure 178.

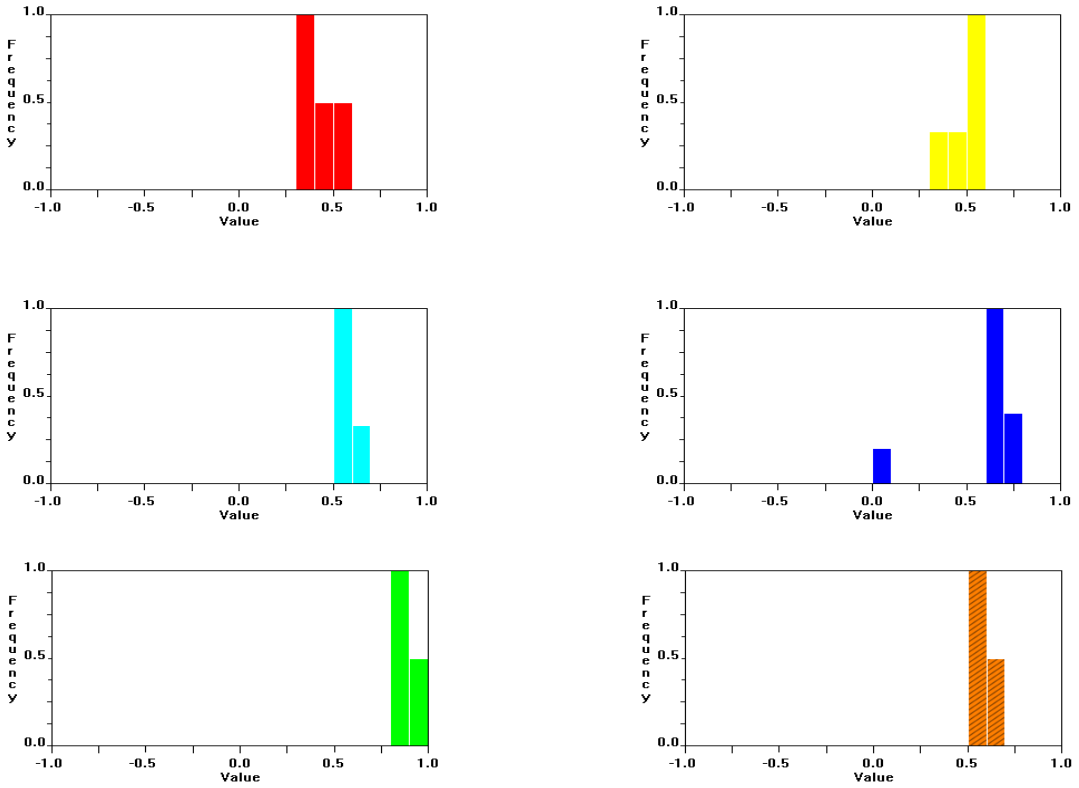

Figure 178 - DSC Silhouettes

The lowest band in the red cluster represents samples p and c1. The second band represents c1+c3 and the uppermost, the only one above 0.5, represents c3.

The lowest band in the yellow cluster represents p+c1+c3. The middle band represents c1+c3 +s2 while the uppermost band, again the only one above 0.5, represents the remaining samples.

The aquamarine cluster has all bands above 0.5. The uppermost band represents s3+c3 and c1+s2+s3. The lower band represents the remaining samples in the cluster.

The blue cluster has two bands above 0.5, with one being exactly on 0.75, and a third band barely above 0. The band at 0 represents s3+c1. This sample lies a long distance from any other samples in the MMDS plot. The uppermost band represents c1+s3+s4 and s3+2. The middle band represents all remaining blue samples.

The green cluster has all samples above 0.75 so its silhouettes will not be examined in any detail.

The striped brown cluster has two bands above 0.5. The lower valued of these contains the p+s2 and p+s4 samples with the upper valued one containing p+s3.

The fuzzy clustering is shown in Figure 179 and the numeric data in Table 31.

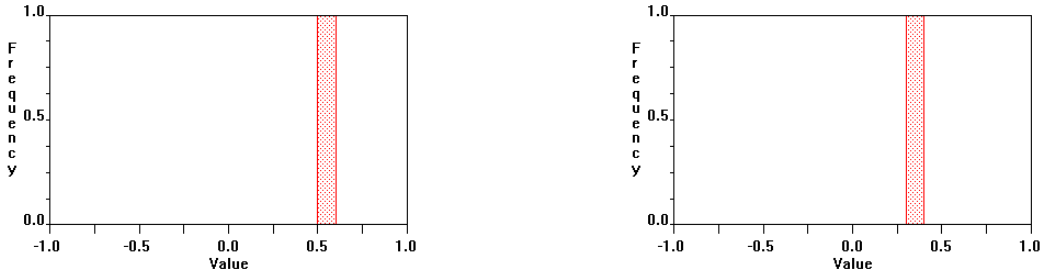

Figure 179 - DSC Fuzzy Clustering

	1	2	3	4	5	6	7	
c1+3	0	0.02	0.09	0.71	0.23	0	0.02	
c1+c3+s2	0.01	0	0.67	0	0.27	0.2	0.02	
c1+c3+s3	0	0	0.7	0.01	0.09	0.1	0	
c1+c3+s4	0	0	0.69	0.07	0.26	0.24	0.02	
c1+s2+s3	0	0.24	0.04	0.06	0.67	0.19	0.22	
c1+s2+s4	0	0.15	0.03	0.04	0.66	0.12	0.21	
c1+s3+s4	0	0	0	0.02	0	0	0.73	
c1	0	0.09	0.05	0.64	0.17	0.04	0	
c3+s2+s3	0.07	0.13	0.27	0.01	0.23	0.73	0.12	
c3+s2+s4	0.1	0.15	0.24	0	0.14	0.75	0.11	
c3+s3+s4	0.06	0.14	0.24	0.03	0.26	0.73	0.09	
c3	0	0.03	0.1	0.74	0.23	0	0.02	
p+c1+c3	0	0	0.63	0.24	0.27	0.09	0	
p+c1	0	0	0.23	0.25	0.62	0.01	0.12	
p+c3	0	0	0.75	0.03	0.04	0.11	0	
p+s2	0.05	0.73	0	0.13	0.08	0.06	0.13	
p+s3	0.05	0.77	0	0.1	0.14	0.13	0.17	
p+s4	0.06	0.7	0	0	0.09	0.1	0.05	
p	0	0	0.06	0.63	0.21	0	0.11	
s2+3+4	0.08	0.07	0	0	0.07	0.07	0.73	
s2+c1	0.01	0.21	0.07	0.18	0.62	0.13	0.17	
s2+c3	0	0	0.29	0.18	0.64	0.02	0	
s2	0.14	0.12	0	0	0.03	0.08	0.73	
s3+2	0.12	0.08	0	0	0.08	0.07	0.76	
s3+c1	0	0.14	0.14	0.2	0.54*	0.09	0.32*	<==
s3+c3	0	0.04	0.21	0.17	0.71	0.09	0	
s3	0.1	0.06	0.01	0	0.13	0.09	0.75	
s4+2	0.12	0.09	0	0	0.06	0.07	0.76	
s4+3	0.07	0.05	0.01	0	0.1	0.08	0.75	
s4+c1	0	0.24	0.05	0.19	0.62	0.14	0.28	
s4+c3	0	0	0.31	0.08	0.65	0.2	0	
s4	0.81	0	0	0	0	0	0.13	

Table 31 - DSC Fuzzy Clusters Numeric Data

For this dataset cluster 1 is the purple cluster, cluster 2 the striped brown cluster, cluster 3 the yellow cluster, cluster 4 the red cluster, cluster 5 the aquamarine cluster, cluster 6 the green cluster and cluster 7 the blue cluster

The only sample present in fuzzy clustering is the s3+c1 sample, which is present in the blue cluster. This is present in both bars in both plots.

The sample could potentially be present in either the aquamarine or blue cluster.

6.10.5 IR DATA

The dendrogram and MMDS plot for the IR data is shown in Figure 180. The dataset was collected over the ranges determined in section 6.3.

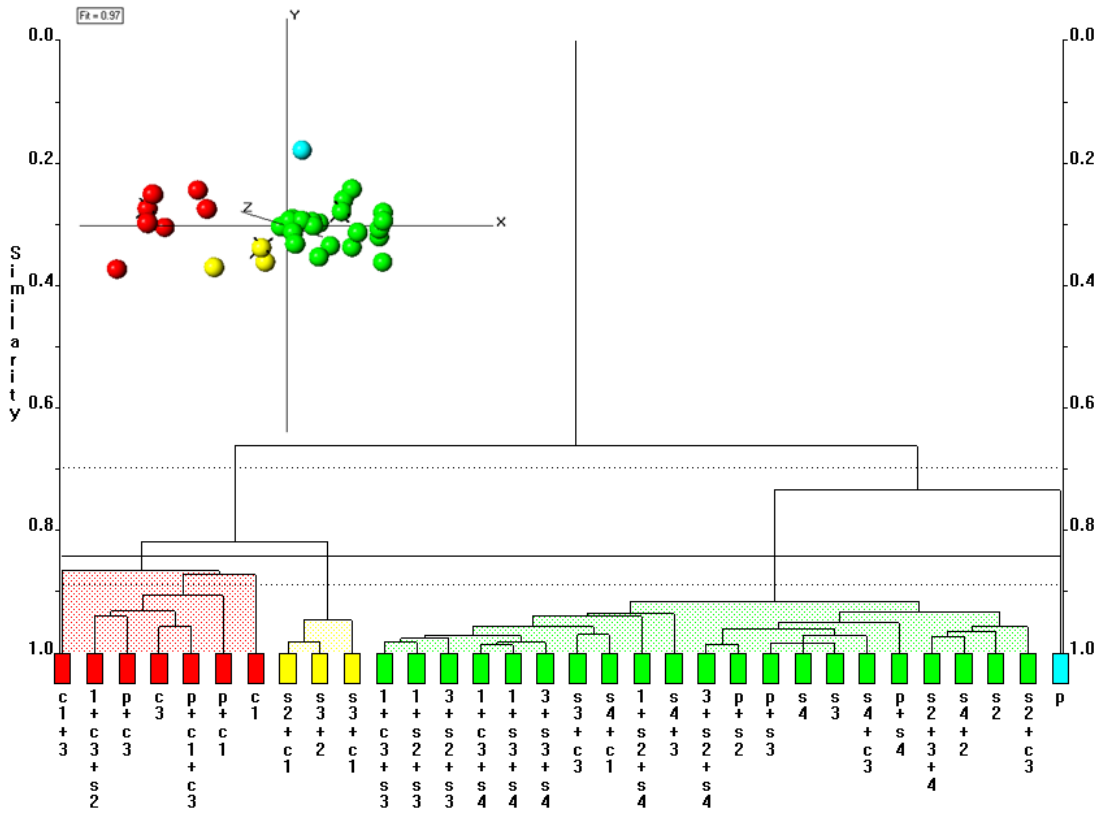

Figure 180 - Thirty-Two Sample IR Dendrogram and MMDS Plot

The red cluster contains samples c1+3, c1, c1+c3+s2, p+c1 and p+c1+c3 from expected cluster 1 and samples c3 and p+c3 which make up expected cluster 5.

The yellow cluster contains samples s2+c1 and s3+c1 from expected cluster 1 and s3+2 from expected cluster 3.

The aquamarine cluster contains the pure piroxicam sample from expected cluster 6.

All remaining samples are in the green cluster.

In the MMDS plot, the clusters are mostly diffuse. The yellow cluster lies close to the green cluster, however is still clearly separated. The dendrogram has a score of 0.62, again showing a particularly poor result.

The silhouettes for the IR dataset are shown in Figure 181.

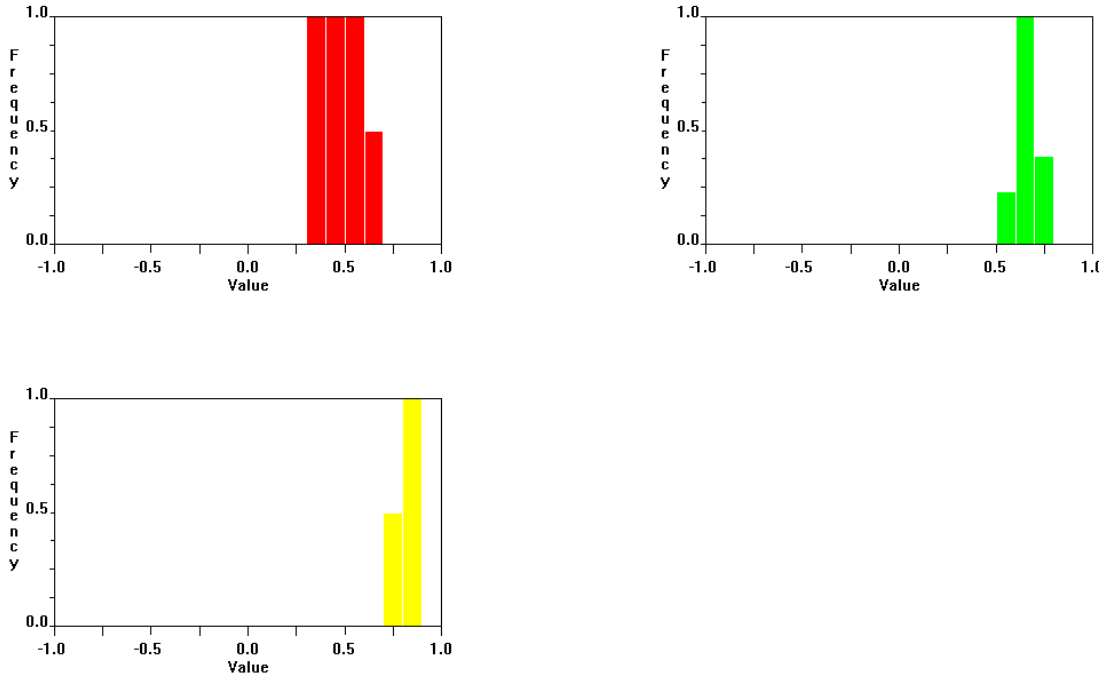

Figure 181 - IR Silhouettes

In the red cluster two of the bands lie below 0.5 and two above 0.5. The lowest band contains c1+3 and c1. The next band contains c3 and p+c1. The first band above 0.5 contains c1+c3+s2 and p+c3, while the last band contains p+c1+c3.

The green cluster has no bands below 0.5. One of the bands is exactly on 0.75. The lower band contains s3+c3, s4+2 and p+s4. The second band contains samples c1+c3+s4, c1+c3+s3, c1+s2+s4, p+s2, p+s3, s2+c3, s2, s3, s4+3, s4+c1, s4+c3 and s4. The remaining peak is over 0.75 so will not be discussed.

The yellow cluster has all samples over 0.75 and so will not be discussed.

No samples have been found for this dataset with cluster memberships less than 0.5 so no fuzzy clustering plot or information will be shown.

Overall the dataset is better clustered than the PXRD, Raman or DSC datasets.

6.10.6 COMBINED CLUSTERING

The combined dendrogram, combining all four data types using INDSCAL, is shown in Figure 182.

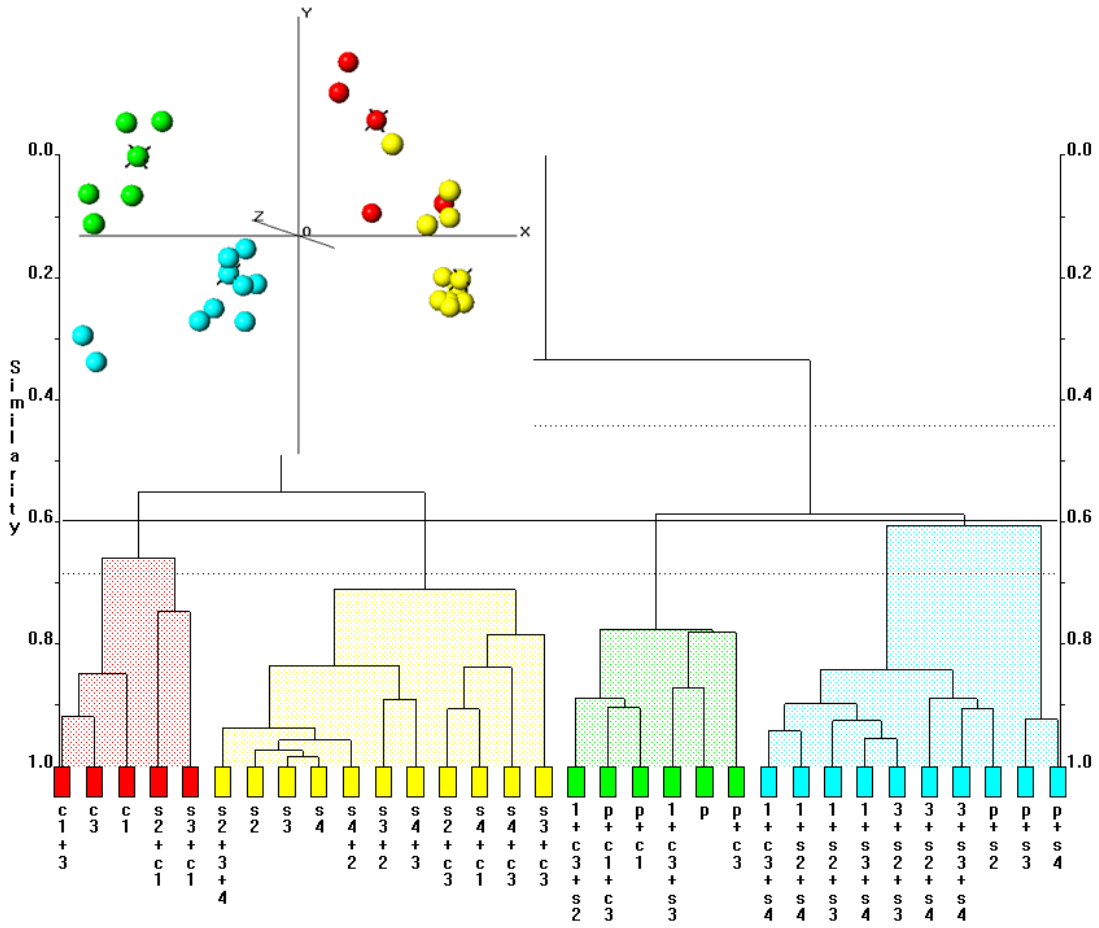

Figure 182 – Combined Dendrogram

The red cluster contains samples s2+c1, c1+3, s3+c1 and c1 from expected cluster 1 and c3 from expected cluster 5. The yellow cluster contains samples s2+3+4, s2, s2+c3 and s3+2 from expected cluster 3, s3, s3+c3 and s4+3 from expected cluster 2, s4, s4+c3 and s4+2 from expected cluster 4 and s4+c1 from expected cluster 1. The green cluster contains samples c1+c3+s2, c1+c3+s3, p+c1 and p+c1+c3 from expected cluster 1, p from expected cluster 6 and p+c3 from expected cluster 5. The aquamarine cluster contains sample c3+s2+s4, c3+s3+s4 and c3+s2+s3 from expected cluster 2, c1+c3+s4, c1+s2+s4, c1+s2+s3 and c1+s3+s4 from expected cluster1 and p+s3 and p+s2 from expected cluster 3.

The MMDS plot shows the green and aquamarine clusters as being clearly separated. The red and yellow cluster appear to be close together, however if the plot is reoriented they

can be seen to be clearly separate. The dendrogram has a score of 0.66, again showing two thirds of the samples to be poorly clustered.

The combined dataset shows improved clustering over the previous methods. It is a massive improvement over the PXRD and DSC methods and shows a small improvement over the Raman and IR dataset.

6.11 THIRTY-TWO SAMPLE DERIVATIVES

The first and second derivative runs for the full dataset were also studied.

6.11.1 FIRST DERIVATIVE RAMAN DATASET

The first derivative Raman dendrogram and MMDS plot is shown in Figure 183.

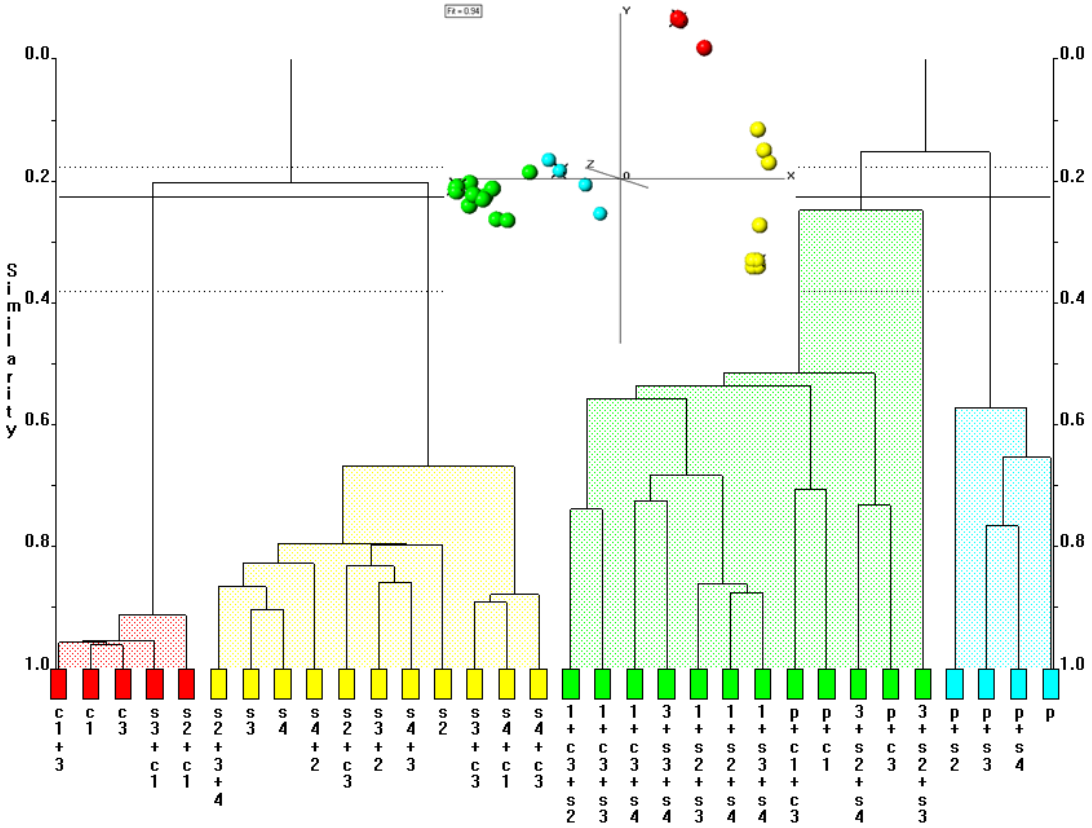

Figure 183 - Thirty-Two Sample Dataset Raman First Derivative Dendrogram and MMDS Plot

The red cluster contains samples c1+3, s3+c1, s2+c1 and c1 from expected cluster 1 and c3 from expected cluster 5. The yellow cluster contains samples s2+3+4, s2 and s3+2 from expected cluster 3, s3 and s3+c3 from expected cluster 2, s4, s4+3 and s4+2 from expected

cluster 4, s3+c3 and s4+c3 from expected cluster 3 and p+c1 from expected cluster 5. The green cluster contains samples c1+c3+s2, c1+s2+s4, c1+s2+s3, c1+s3+s4, c1+c3+s3, c1+c3+s4 and p+c1+c3 from expected cluster 1 and c3+s2+s4, c3+s2+s4 and c3+s2+s3 from expected cluster 2. The aquamarine cluster contains samples p+s2 from expected cluster 3, p+s4 from expected cluster 4, p from expected cluster 6 and p+s3 from expected cluster 2.

The MMDS plot shows clear separation of the different clusters.

The dendrogram has a score of 0.72. The score from this dataset shows poorer clustering that that seen in the original Raman dataset.

6.11.2 SECOND DERIVATIVE RAMAN DATASET

The second derivative Raman dendrogram and MMDS plot is shown in Figure 184.

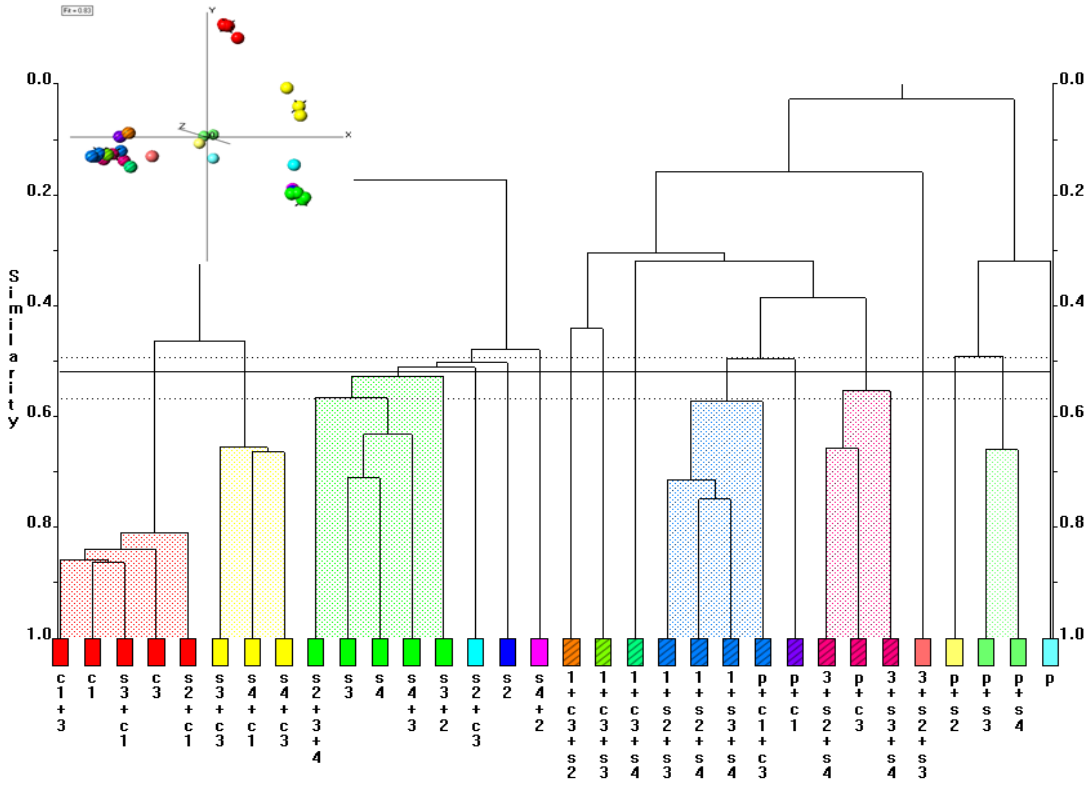

Figure 184 - Thirty-Two Sample Raman Second Derivative Dendrogram and MMDS Plot

At first glance, the second derivative Raman dataset appears to be much poorer clustered than any of those previously studied.

The red cluster contains samples c1, s2+c1, s3+c1 and c1+3 from expected cluster 1 and c3 from expected cluster 5. The yellow cluster samples s3+c3 from expected cluster 2, s4+c1

from expected cluster 1 and s4+c3 from expected cluster 4. The green cluster contains samples s2+3+4 and s3+2 from expected cluster 3, s3 and s4+3 from expected cluster 2 and s4 from expected cluster 4. The aquamarine cluster contains samples s2+c3 from expected cluster 3. The blue cluster contains samples s2 from expected cluster 3. The purple cluster contains samples 4+2 from expected cluster 4. The striped brown, striped light green, striped dark green cluster contains three samples that are part of expected cluster 1. The striped blue cluster contains samples c1+s2+s3, c1+s2+s4, c1+s3+s4 and p+c1+c3 from expected cluster 1. The striped purple cluster contains sample p+c1 from expected cluster 1. The striped red cluster contains sample c3+s2+s4 and c3+s3+s4 from expected cluster 2 and p+c3 from expected cluster 5. The orange cluster contains sample c3+s2+s3 from expected cluster 2. The pale yellow cluster contains sample p+s2 from expected cluster 3 while the pale green cluster contains sample p+s4 from expected cluster 4 and p+s3 from expected cluster 3. The pale aquamarine cluster contains the lone piroxicam sample from expected cluster 2.

The MMDS plot appears poorly resolved however, as already stated; raising the cut-level will merge many of these into larger clusters. The score for this dataset is 0.69, signifying a small improvement over the first derivative Raman dataset.

Although this dataset appears to be poorly clustered, it is not as bad as it initially appears. Raising the cut-level will merge many of the clusters and give a result similar to the original and first derivative Raman datasets. This is shown in Figure 185.

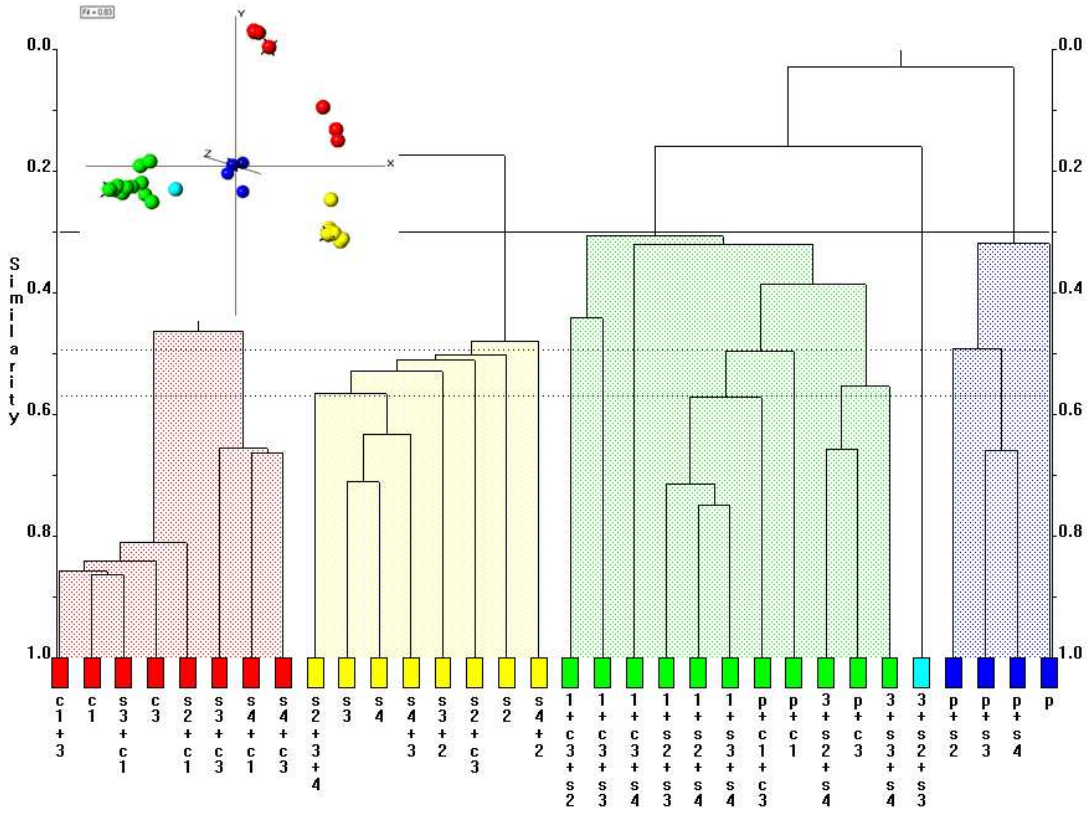

Figure 185 - Thirty-Two Sample Raman Second Derivative Dendrogram and MMDS Plot with Adjusted Cut-Level

The red cluster now contains samples c1+3, c1, s2+c1, s4+c1 and s3+c1 from expected cluster 1, sample s3+c3 from expected cluster 2 and s4+c3 from expected cluster 4. The yellow cluster now contains samples s2+3+4, s2, s2+c3 and s3+2 from expected cluster 3, s3 and s4+3 from expected cluster 2, s4 and s4+2 from expected cluster 4. The green cluster now contains samples c1+c3+s2, c1+c3+s3, c1+c3+s4, c1+s2+s3, c1+s2+s4, c1+s3+s4 and p+c1+c3 from expected cluster 1, c3+s3+s4 and c3+s2+s4 from expected cluster 2 and sample p+c3 from expected cluster 5. The aquamarine cluster contains sample c3+s2+s3 from expected cluster 2. The blue cluster contains samples p+s2 from expected cluster 3, p+s4 from expected cluster 4, p+s3 from expected cluster 2 and p from expected cluster 6.

The dendrogram now appears to be much more clearly defined with the newly merged green cluster in particular being much clearer. The combination of many of the smaller clusters results in an improvement in the dendrogram score which now lies at 0.53, a much improved results with almost half of the samples now being clustered as expected. Overall the clustering is not as good as that seen in the first derivative or original Raman data.

6.11.3 FIRST DERIVATIVE IR DATASET

The first derivative IR dendrogram and MMDS plot are shown in Figure 186.

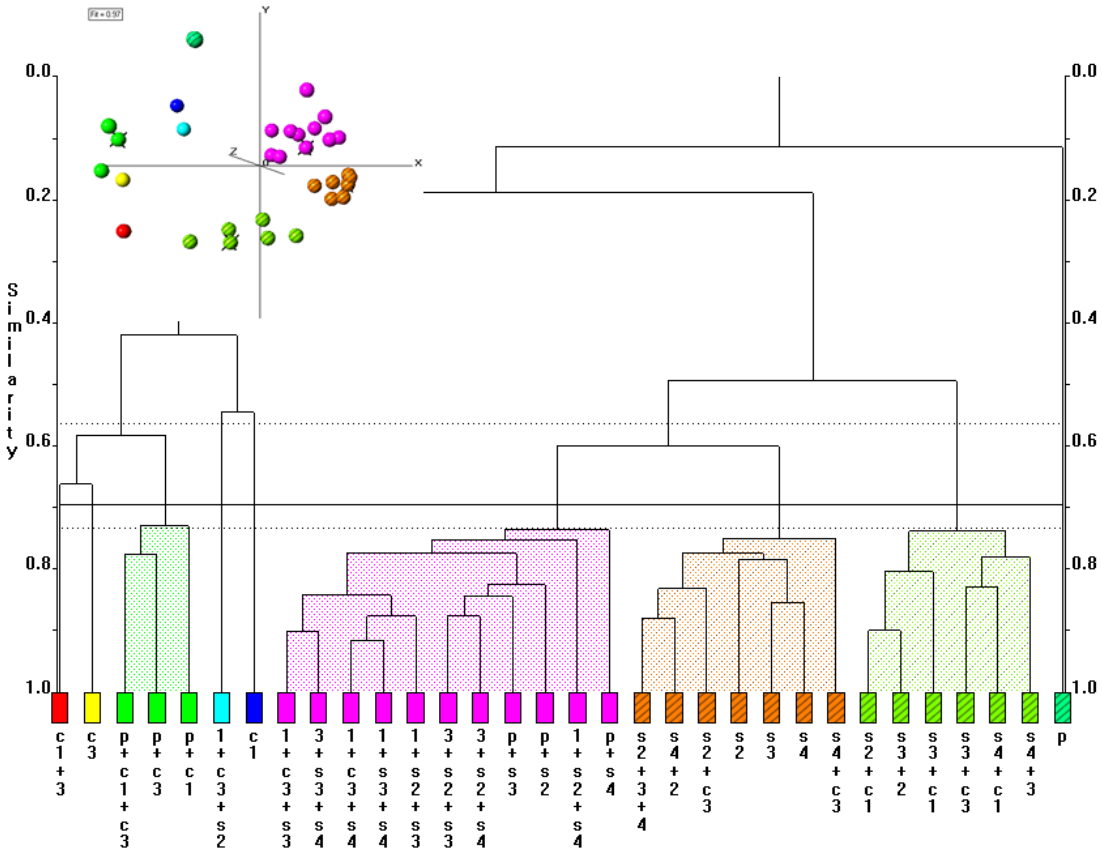

Figure 186 - Thirty-Two Sample First Derivative IR Dendrogram and MMDS Plot

The red cluster contains sample c1+3 from expected cluster 1 while the yellow cluster contains sample c3 from expected cluster 5. The green cluster contains sample p+c1+c3 and p+c1 from expected cluster 1 and p+c3 from expected cluster 5. The aquamarine cluster contains sample c1+c3+s2 from expected cluster 1 and the blue cluster contains sample c1 from expected cluster 1. The purple cluster contains samples c1+c3+s3, c1+s3+s4, c1+s2+s3, c1+s2+s4 and c1+c3+s4 from expected cluster 1, samples c3+s3+s4, p+s3 and c3+s2+s4 from expected cluster 2 and p+s2 from expected cluster 3. The striped brown cluster contains samples s2+3+4, s2+c3 and s2 from expected cluster 3, s4+2, s4+c3 and s4 from expected cluster 4 and s3 from expected cluster 2. The striped light green cluster contains samples s2+c1, s4+c1 and s3+c1 from expected cluster 1, s3+2 from expected cluster 3 and s3+c3 and s4+3 from expected cluster 2. The striped dark green cluster contains the pure piroxicam sample from expected cluster 6.

The MMDS plot shows clear separation of the clusters; however the striped light green cluster is rather diffuse.

The dendrogram now has a score of 0.69, worse that the 0.62 seen for the original IR result.

6.11.4 SECOND DERIVATIVE IR DATASET

The second derivative IR dendrogram and MMDS plot are shown in Figure 187.

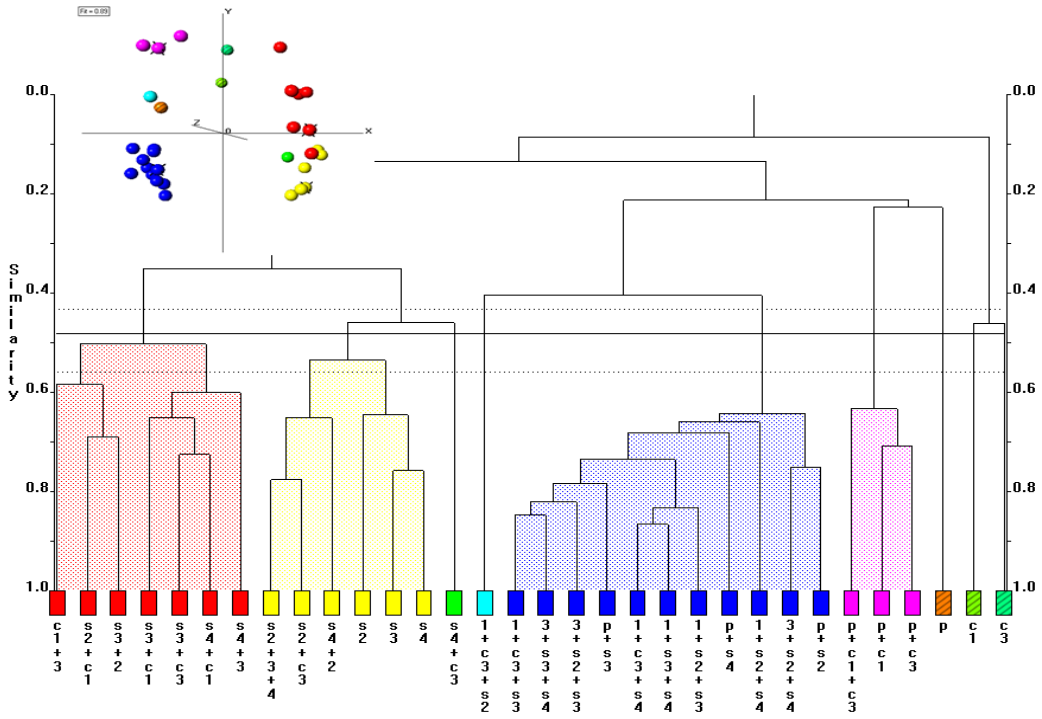

Figure 187 - Thirty-Two Sample Second Derivative IR Dendrogram and MMDS Plot

The red cluster contains samples c1+3, s3+c1, s4+c1 and s2+c1 from expected cluster 1 and s3+2 from expected cluster 3 and s3+c3 and s4+3 from expected cluster 2. The yellow cluster contains samples s2+3+4, s2+c3 and s2 from expected cluster 3, s4+2 and s3 from expected cluster 2 and s4 from expected cluster 4. The green cluster contains sample s4+c3 from expected cluster 4 while the aquamarine cluster contains sample c1+c3+s2 from expected cluster 1. The blue cluster contains samples c1+c3+s3, c1+c3+s4, c1+s2+s3, c1+c3+s4 and c1+s2+s4 from expected cluster 1, c3+s2+s3, p+s3 and c3+s2+s4 from expected cluster 2, p+s4 from expected cluster 4 and p+s2 from expected cluster 3. The purple cluster contains samples p+c1+c3 and p+c1 from expected cluster 1 and p+c3 from expected cluster 5. The striped brown cluster contains the pure piroxicam sample from

expected cluster 6, the striped light green cluster contains sample c1 from expected cluster 1 and the striped dark green cluster contains sample c3 from expected cluster 5.
The MMDS shows the blue and yellow clusters to be tightly grouped. The red cluster is much more diffuse. The dendrogram now has a score of 0.72, a worse result than that seen for either the first derivative or original IR datasets.

6.12 FLOWCHART

All of the possible combinations of pre-processing were applied to the PXRD dataset and the results compared to the optimal clustering. The number of misplaced samples is shown in Table 32.

	Score
no pre-processing	0.40
denoise	0.44
background	0.48
background and denoise	0.48

Table 32 - Misplaced samples for 32 samples dataset

As can be seen no pre-processing, gives the optimal result with denoise following with two more misplaced samples. Remove background and remove background and denoise both have equal number of misplaced samples.
The flowchart for this result is shown in Figure 188.

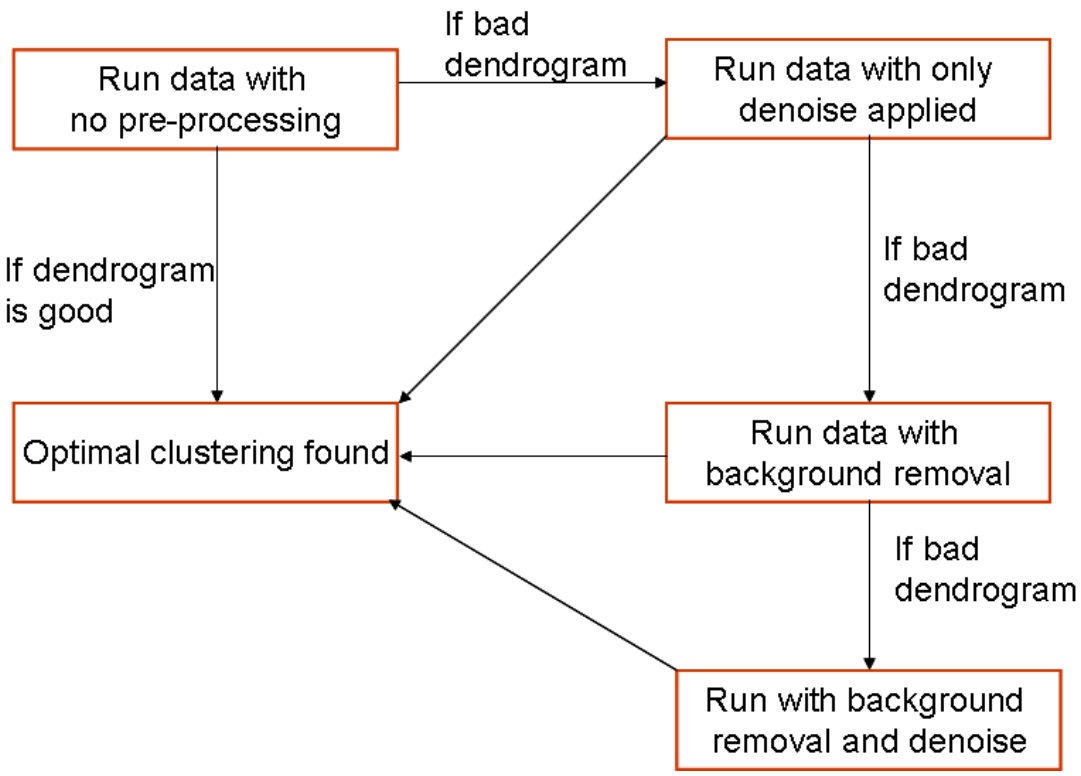

Figure 188- Flowchart for 32 sample dataset

This flowchart again shows that running a dataset with no pre-processing applied gives optimal results however if a 'bad' dendrogram, as already defined, is the result then the flowchart should again be preceded through.

6.13 QUANTITATIVE ANALYSIS

The materials were compared using the PolySNAP manual analysis mode. The results of this are shown in Table 33. The data was compared using the SVD method.
The results, with pre-processing applied to the data, is shown in Table 34 and 35.

PXRD	Samples	Actual	Predicted	Difference	Raman	Samples	Actual	Predicted	Difference	IR	Samples	Actual	Predicted	Difference
	p+c1	13:87	52:48	39.00		p+c1	13:87	63.9:36.1	50.90		p+c1	13:87	72.9:27.1	58.90
	p+c3	28:72	82.4:17.6	54.40		p+c3	28:72	85.8:14.2	56.80		p+c3	28:72	73.7:26.3	45.70
	p+s2	22:78	31.6:68.4	9.60		p+s2	22:78	16.8:83.2	5.20		p+s2	22:78	75.7:24.3	53.70
	p+s3	16:84	30.3:69.7	14.30		p+s3	16:84	57.4:42.6	41.40		p+s3	16:84	74.6:25.4	58.60
	p+s4	47:53	37.1:62.9	6.00		p+s4	47:53	63.2:36.8	16.20		p+s4	47:53	77.1:22.9	30.10
	c1+c3+s2	48:32:22	3.4:88.9:7.7	38.60		c1+c3+s2	48:32:22	40.9:47.5:11.7	14.30		c1+c3+s2	48:32:22	67:18.7:14.3	13.33
	c1+c3+s3	23:47:29	1.9:76.5:21.6	50.67		c1+c3+s3	23:47:30	50.3:47.4:2.2	18.97		c1+c3+s3	23:47:29	14.4:19.6:65.9	24.30
	c1+c3+s4	33:33:33	6:8.9:85.2	33.87		c1+c3+s4	33:33:33	45.7:43:11.3	14.80		c1+c3+s4	33:33:33	36.9:0.8:62.3	21.80
	c1+s2+s3	26:31:41	0.4:35.4:64.2	17.73		c1+s2+s3	26:31:41	4.7:67.6:27.8	25.37		c1+s2+s3	26:31:41	17.9:43.7:38.5	7.77
	c1+s3+s4	33:33:33	43.4:32.6:24	6.60		c1+s3+s4	33:33:33	7.8:36.5:55.6	17.10		c1+s3+s4	33:33:33	8.1:35.8:56.1	9.23
	c1+s2+s4	33:33:33	20.9:32.6:55.5	11.67		c1+s2+s4	33:33:33	5.6:60.4:34.1	18.97		c1+s2+s4	33:33:33	11.4:50.9:37.6	14.70
	c3+s2+s3	15:46:39	14.9:64.1:21	12.07		c3+s2+s3	15:46:39	3.6:67.5:28.9	14.33		c3+s2+s3	15:46:39	19.5:77:3.4	23.7
	c3+s3+s4	24:66:10	3.4:35.2:61.5	34.30		c3+s3+s4	24:66:10	5.3:38.9:55.8	24.53		c3+s3+s4	24:66:10	10.9:33.9:55.2	13.47
	c3+s2+s4	24:45:31	0.2:43.2:56.6	16.73		c3+s2+s4	24:45:31	4.6:36.7:58.8	18.50		c3+s2+s4	24:45:31	2.5:12.9:84.6	35.73
	p+c1+c3	12:66:22	33.9:9.4:56.8	39.77		p+c1+c3	12:66:22	10.8:46.3:42.9	13.93		p+c1+c3	12:66:22	48.7:33.7:17.6	24.47
		Mean absolute difference		25.69			Mean absolute difference		23.42			Mean absolute difference		29.03
		RMS difference		6.63			RMS difference		6.05			RMS difference		7.50
		Max absolute difference		28.71			Max absolute difference		33.38			Max absolute difference		29.87
		Min absolute difference		19.69			Min absolute difference		18.22			Min absolute difference		21.27

Table 33 – Data from Mixtures in Manual Analysis Mode

PXRD	Samples	Actual	Processed Predicted 1	Difference 1	Raman	Samples	Actual	Processed Predicted 1	Difference 1	IR	Samples	Actual	Processed Predicted 1	Difference 1
	p+c1	13:87	50.2:49.8	37.20		p+c1	13:87	34.3:65.7	21.30		p+c1	13:87	34.3:65.7	21.30
	p+c3	28:72	77.3:22.7	49.30		p+c3	28:72	62:38	34.00		p+c3	28:72	62:38	34.00
	p+s2	22:78	22.2:77.8	0.20		p+s2	22:78	0.6:99.4	21.40		p+s2	22:78	0.6:99.4	21.40
	p+s3	16:84	22.9:77.1	6.90		p+s3	16:84	53.2:46.8	37.20		p+s3	16:84	53.2:46.8	37.20
	p+s4	47:53	36:64	11.00		p+s4	47:53	57:43	10.00		p+s4	47:53	57:43	10.00
	c1+c3+s2	48:32:22	14.3:79.9:5.8	32.60		c1+c3+s2	48:32:22	6.7:44.6:48.2	26.70		c1+c3+s2	48:32:22	6.7:44.6:48.6	41.50
	c1+c3+s3	23:47:29	1.1:71.7:17.2	19.47		c1+c3+s3	23:47:29	51.4:48.6:0	17.67		c1+c3+s3	23:47:29	49.9:47.8:2.3	18.13
	c1+c3+s4	33:33:33	4.9:8.9:86.2	35.13		c1+c3+s4	33:33:33	50.9:48.2:0.9	21.73		c1+c3+s4	33:33:33	50.9:58.2:0.9	36.07
	c1+s2+s3	26:31:41	3.2:34.5:62.4	15.90		c1+s2+s3	26:31:41	64.9:26:9.1	25.27		c1+s2+s3	26:31:41	64.3:26:9.1	25.07
	c1+s3+s4	33:33:33	38.7:36.8:24.5	6.00		c1+s3+s4	33:33:33	57.2:4.4:38.4	19.40		c1+s3+s4	33:33:33	57.2:4.4:38.4	19.40
	c1+s2+s4	33:33:33	17.1:30.4:52.5	12.67		c1+s2+s4	33:33:33	41.8:34.6:23.6	6.60		c1+s2+s4	33:33:33	41.8:34.6:23.6	6.60
	c3+s2+s3	15:46:39	1.4:76.5:22	20.37		c3+s2+s3	15:46:39	39:20.6:40.4	16.93		c3+s2+s3	15:46:39	39:20.6:40.4	16.93
	c3+s3+s4	24:66:10	4.9:36.1:59	32.67		c3+s3+s4	24:66:10	41.8:13.8:44.4	34.80		c3+s3+s4	24:66:10	41.8:44.4:13.8	14.40
	c3+s2+s4	24:45:31	7:38.6:54.4	15.60		c3+s2+s4	24:45:31	45.9:35.3:18.9	14.57		c3+s2+s4	24:45:31	45.9:35.3:18.9	14.57
	p+c1+c3	12:66:22	33.9:9.4:56.8	37.77		p+c1+c3	12:66:22	5.2:48.8:46	16.00		p+c1+c3	12:66:22	5.2:48.8:46	12.03
		Mean absolute difference		22.18			Mean absolute difference		21.57			Mean absolute difference		21.91
		RMS difference		5.73			RMS difference		5.57			RMS difference		5.66
		Max absolute difference		27.12			Max absolute difference		15.63			Max absolute difference		19.59
		Min absolute difference		21.98			Min absolute difference		14.97			Min absolute difference		15.31

Processed Predicted 1 - background remove and smoothed

Table 34 - Data from Mixtures in Manual Analysis Mode with Pre-processing 1

PXRD	Samples	Actual	Processed Predicted 2	Difference 2	Raman	Samples	Actual	Processed Predicted 2	Difference 2	IR	Samples	Actual	Processed Predicted 2	Difference 2
	p+c1	13:87	52:48	39.00		p+c1	13:87	63.7:36.3	51.70		p+c1	13:87	36.3:63.7	23.30
	p+c3	28:72	82.4:17.6	54.40		p+c3	28:72	85.7:14.3	57.70		p+c3	28:72	73.8:26.2	45.80
	p+s2	22:78	31.6:68.4	9.60		p+s2	22:78	16.6:83.4	5.40		p+s2	22:78	75.8:24.2	53.80
	p+s3	16:84	30.3:69.7	14.30		p+s3	16:84	57.2:42.8	41.20		p+s3	16:84	74.6:25.4	58.60
	p+s4	47:53	37.1:62.9	9.90		p+s4	47:53	63.1:36.9	16.10		p+s4	47:53	77.1:22.9	30.10
	c1+c3+s2	48:32:22	9.8:68.8:3.4	31.20		c1+c3+s2	48:32:22	41.6:47.6:10.8	11.07		c1+c3+s2	48:32:22	67.3:18.4:14.3	13.53
	c1+c3+s3	23:47:29	1.9:76.5:21.6	19.33		c1+c3+s3	23:47:29	50.3:47.4:2.2	18.17		c1+c3+s3	23:47:29	14.6:19.4:66	24.33
	c1+c3+s4	33:33:33	4:8.9:85.2	34.80		c1+c3+s4	33:33:33	45.6:43:11.4	14.73		c1+c3+s4	33:33:33	37.1:0.5:62.4	22.15
	c1+s2+s3	26:31:41	0.4:35.4:64.2	17.73		c1+s2+s3	26:31:41	4.7:68.1:27.2	24.07		c1+s2+s3	26:31:41	21.2:39.5:39.3	5.00
	c1+s3+s4	33:33:33	43.4:32.6:24	6.60		c1+s3+s4	33:33:33	7.8:36.5:55.7	17.13		c1+s3+s4	33:33:33	10.2:34.6:55.1	15.50
	c1+s2+s4	33:33:33	20.9:23.6:55.5	14.67		c1+s2+s4	33:33:33	5.6:60.7:33.7	18.60		c1+s2+s4	33:33:33	11.3:52.5:36.2	14.80
	c3+s2+s3	15:46:39	14.9:64.1:21	12.07		c3+s2+s3	15:46:39	3.7:68:28.3	14.67		c3+s2+s3	15:46:39	18.7:80.6:0.7	25.53
	c3+s3+s4	24:66:10	3.4:35.1:61.5	34.33		c3+s3+s4	24:66:10	5.2:38.9:55.9	30.60		c3+s3+s4	24:66:10	10.8:34.1:55.1	30.07
	c3+s2+s4	24:45:31	0.2:43.1:56.7	17.13		c3+s2+s4	24:45:31	4.6:59.1:36.3	12.87		c3+s2+s4	24:45:31	4:6.2:89.8	59.87
	p+c1+c3	12:66:22	38.9:3.4:57.7	41.73		p+c1+c3	12:66:22	10.9:46.3:42.8	13.87		p+c1+c3	12:66:22	56.6:35:8.5	29.70
		Mean absolute difference		23.79			Mean absolute difference		23.19			Mean absolute difference		30.14
		RMS difference		6.14			RMS difference		5.99			RMS difference		7.78
		Max absolute difference		30.61			Max absolute difference		34.51			Max absolute difference		29.73
		Min absolute difference		17.19			Min absolute difference		17.79			Min absolute difference		25.14

Processed Predicted 2 - smoothed

Table 35 - Data from Mixtures in Manual Analysis Mode with Pre-processing 2

For PXRD, a predicted result is said to closely match the actual values if the values are within 10% of each other in either direction.

For the PXRD data, two samples (p+s2 and c1+s3+s4) closely match the actual results. By applying smoothing and removing the background, this increases to four samples (p+s2, p+s3, p+s4 and c1+s3+s4). With only smoothing applied, two samples (p+s2, p+s4) now match.

For the Raman data, one of the samples (p+s2) matches. If background removal and smoothing are applied, this increases to two samples (p+s4 and c1+s2+s4). Sample p+s2 no longer matches, and indeed is now reported as being almost entirely pure sulfathiazole form 2. With just smoothing applied, one sample (p+s2) matches. If the allowed variance is extended to 15%, five samples now match (p+s2, c1+c3+s2, c1+c3+s4, c3+s2+s3 and p+c1+c3) when no pre-processing is applied. With smoothing and background removal applied, samples p+s4, c1+s2+s4 and c3+s2+s4 match with the actual result. With only smoothing applied, sample p+s2 once again matches its actual composition along with samples c1+c3+s2, c1+c3+s4, c3+s2+s3, c3+s2+s4 and p+c1+c3. The two pre-processing options also highlight the massive difference that background removal can make to composition prediction. Sample p+s2 has already been discussed, however samples c1+s2+s3, c1+s3+s4, c1+s2+s4, c3+s2+s3, c3+s3+s4 and c3+s3+s4 all exhibit similar large shifts in their calculated compositions.

The IR data initially has two samples (c1+s2+s3 and c1+s3+s4) which match the actual composition. For the first pre-processing option samples p+s4 and c1+s2+s4 match. For the second pre-processing option, sample c1+s2+s3 now matches. If the allowed variance is extended to 15%, five samples (c1+c3+s2, c1+s2+s3, c1+s3+s4, c1+s2+s4 and c3+s3+s4) now match for the unprocessed data; five of the samples now match for the first pre-processing option (p+s4, c1+s2+s4, c3+s3+s4. c3+s2+s4 and p+c1+c3). For the second pre-processing option, c1+c3+s2, c1+c3+s3 and c1+s2+s4 are the only samples which closely match. As with the Raman data, the IR data clearly highlights the effect that background removal can have on the composition prediction.

6.14 CONCLUSIONS

- The sulfathiazole/carbamazepine/piroxicam dataset has given further confirmation that IR and DSC data can be successfully used with PolySNAP alongside PXRD data. The usefulness of Raman data has also been further confirmed as a valuable technique to be used in conjunction with PXRD data.
- The DSC data has again been shown to not match closely with the expected clustering.
- For this dataset the use of INDSCAL to combine datasets has also been proven to be effective as it yielded an improved result over that shown in the individual datasets, even when these combinations include the poorer PXRD and DSC datasets.
- The importance of choosing the optimal pre-processing options for composition determination is shown here. For Raman data removing the background appears, for this dataset, to give optimal results. For IR data a variance of 15% rather than the 10% as used for PXRD data is preferable as is using either background removal or smoothing on the dataset.
- For the smaller dataset background removal with or without denoising and applying no pre-processing both yield the optimal clustering.
- For the larger dataset no pre-processing gives the optimal clustering.

CHAPTER 7 BULK MATERIAL DATASET

7.1 THE DATASET

The bulk materials dataset contains six pure materials and a further eight mixtures of these materials. The dataset composition is summarised in Table 36.

For this dataset, PXRD data was collected on a Panalytical X'pert Pro, flat plate, over a range of 5-35°. A different instrument was used as the original was not accessible at that time. Raman data was collected on a Witec alpha 300 with a 785nm laser and an x10 objective lens with 0.25 aperture and 300g/mm grate. DSC data was collected on a TA Instruments Q100. IR data was collected on a JASCO FT/IR 4100.

Sample Number	Sample Name	Name In PolySNAP
1	Malonic acid	MA
2	Methyl urea	MU
3	Urea	U
4	salicylic acid	SA
5	Oxalic acid dihydrate	OA
6	Zinc nitrate hexahydrate	ZN
7	Methyl urea + urea	MU+U
8	Methyl urea + salicylic acid	MU+SA
9	Methyl urea + zinc nitrate	MU+ZN
10	Urea + salicylic acid	U+SA
11	Urea + oxalic acid dihydrate	U+OA
12	Salicylic acid + oxalic acid dihydrate	SA+OA
13	Salicylic acid + zinc nitrate	SA+ZN
14	Oxalic acid dihydrate + zinc nitrate	OA+ZN

Table 36 – Bulk Material Dataset

7.2 DATASET CLUSTERING

7.2.1 EXPECTED CLUSTERING

Ideal mixture patterns were again created by combining the pure patterns in the correct ratio. The dendrogram and MMDS plot for this are shown in Figure 189.

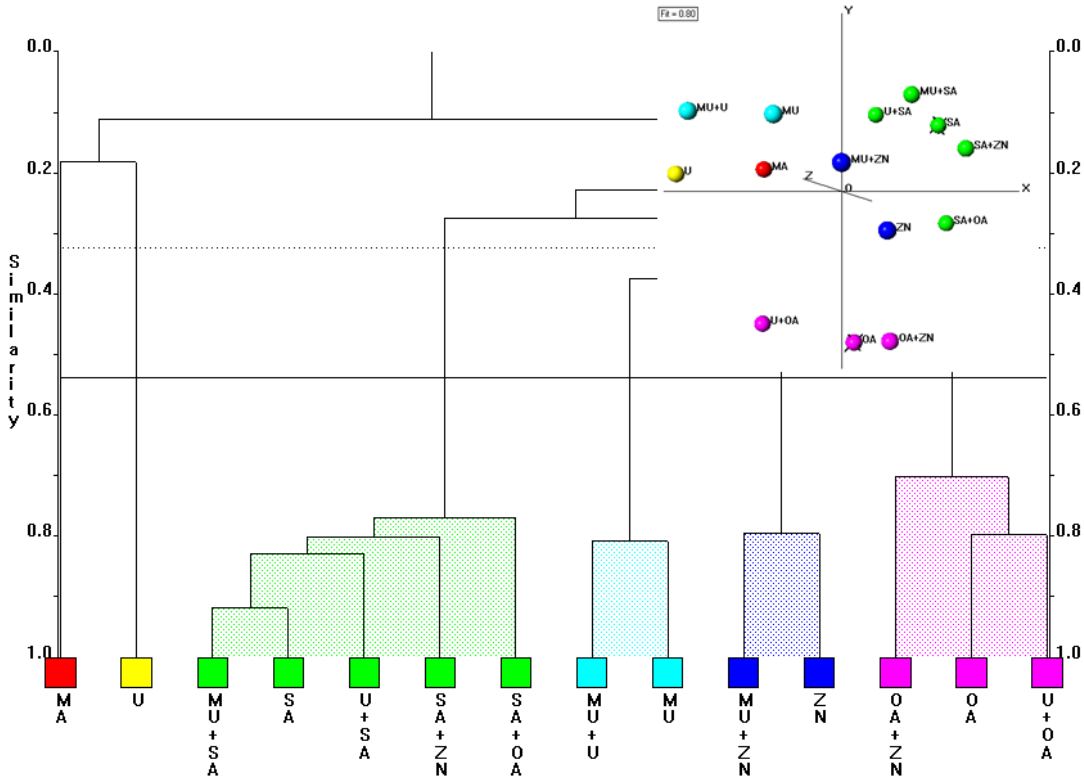

Figure 189 - Dendrogram and MMDS Plot for Expected Clustering

The red cluster contains sample MA while the yellow cluster contains sample U. The green cluster contains samples MU+SA, SA, U+SA, SA+ZN and SA+OA. The aquamarine cluster contains samples MU+U and MU. The blue cluster contains samples MU+ZN and ZN. The purple cluster contains samples OA+ZN, OA and U+OA.

The silhouettes for this dataset are shown in Figure 190.

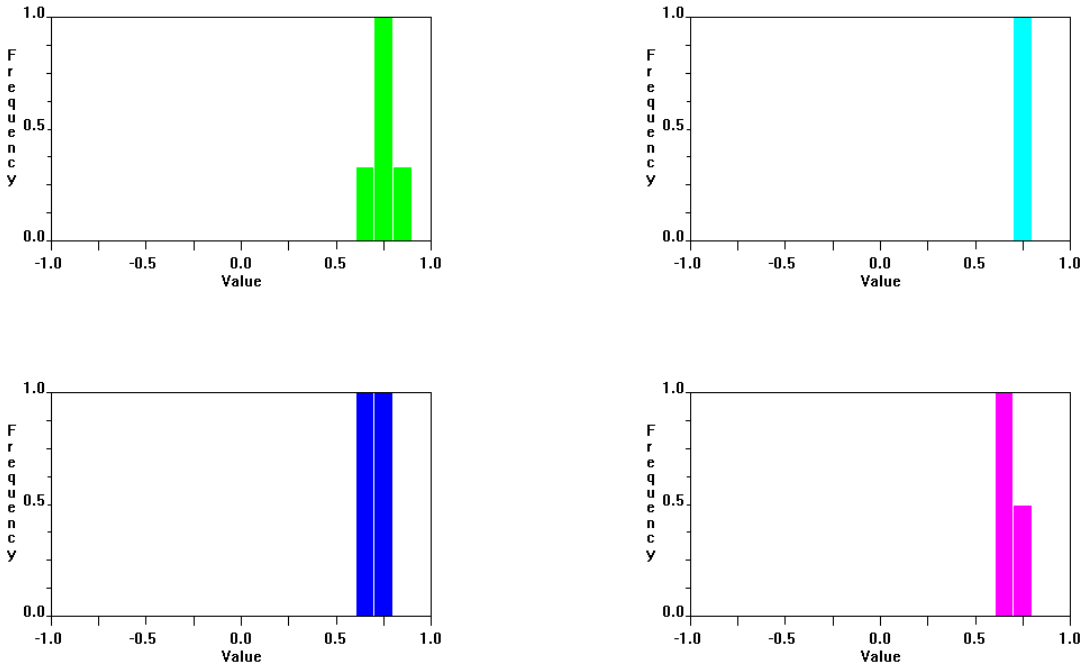

Figure 190 - Silhouettes

None of the regions in the silhouettes are below 0.5. There is no fuzzy clustering for the dataset so the initial clustering shall be used. The following clustering is therefore expected to be as follows:

1) A cluster containing sample MA
2) A cluster containing sample U
3) A cluster containing samples MU+SA, SA, U+SA, SA+ZN and SA+OA
4) A cluster containing samples MU+U and MU
5) A cluster containing samples MU+ZN and ZN
6) A cluster containing samples OA+ZN, OA and U+OA

7.2.2 PXRD DATA

The dendrogram and MMDS Plot are shown in Figure 191.

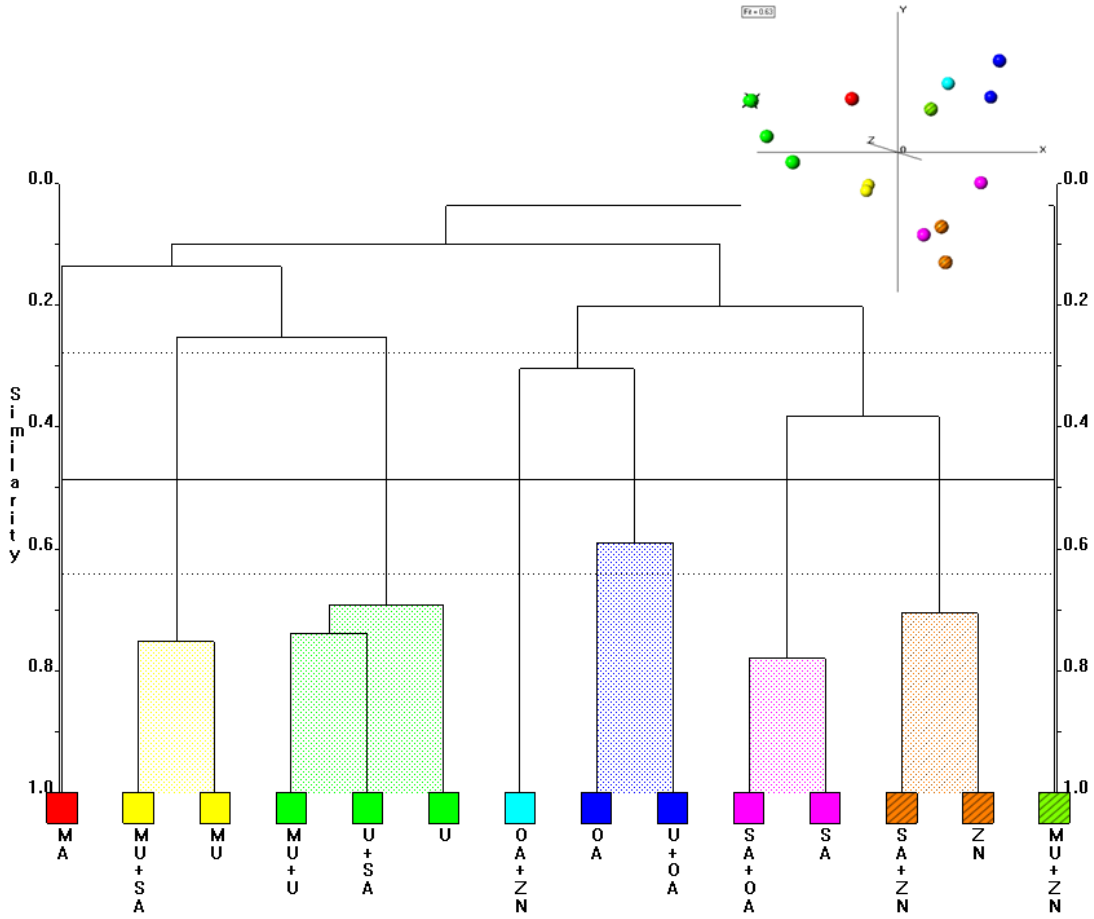

Figure 191 - Bulk Dataset PXRD Dendrogram and MMDS Plot

The red cluster contains the pure malonic acid sample which was expected to be alone in cluster 1. The yellow cluster contains sample MU which is in expected cluster 4 and MU+SA which is in expected cluster 3. The green cluster contains sample MU+U which is in expected cluster 4, U+SA which is in expected cluster 3 and U which is predicted to be alone in expected cluster 1. The aquamarine cluster contains sample OA+ZN which is predicated to be in expected cluster 6. The blue cluster contains samples OA and U+OA which are predicted to be in cluster 6. The purple cluster contains samples SA+OA and SA from expected cluster 3. The striped brown cluster contains sample SA+ZN from expected cluster 3 and ZN from expected cluster 5. The striped green cluster contains MU+ZN from expected cluster 5.

The MMDS plot shows the clusters to be very diffuse. The green cluster, despite all of its tie-bars being outside the limits of the number of estimated clusters, is particularly diffuse.

The score for this dendrogram is 0.43, showing that approximately two fifths of the dataset is incorrectly clustered.

The silhouettes for this dataset are shown in Figure 192.

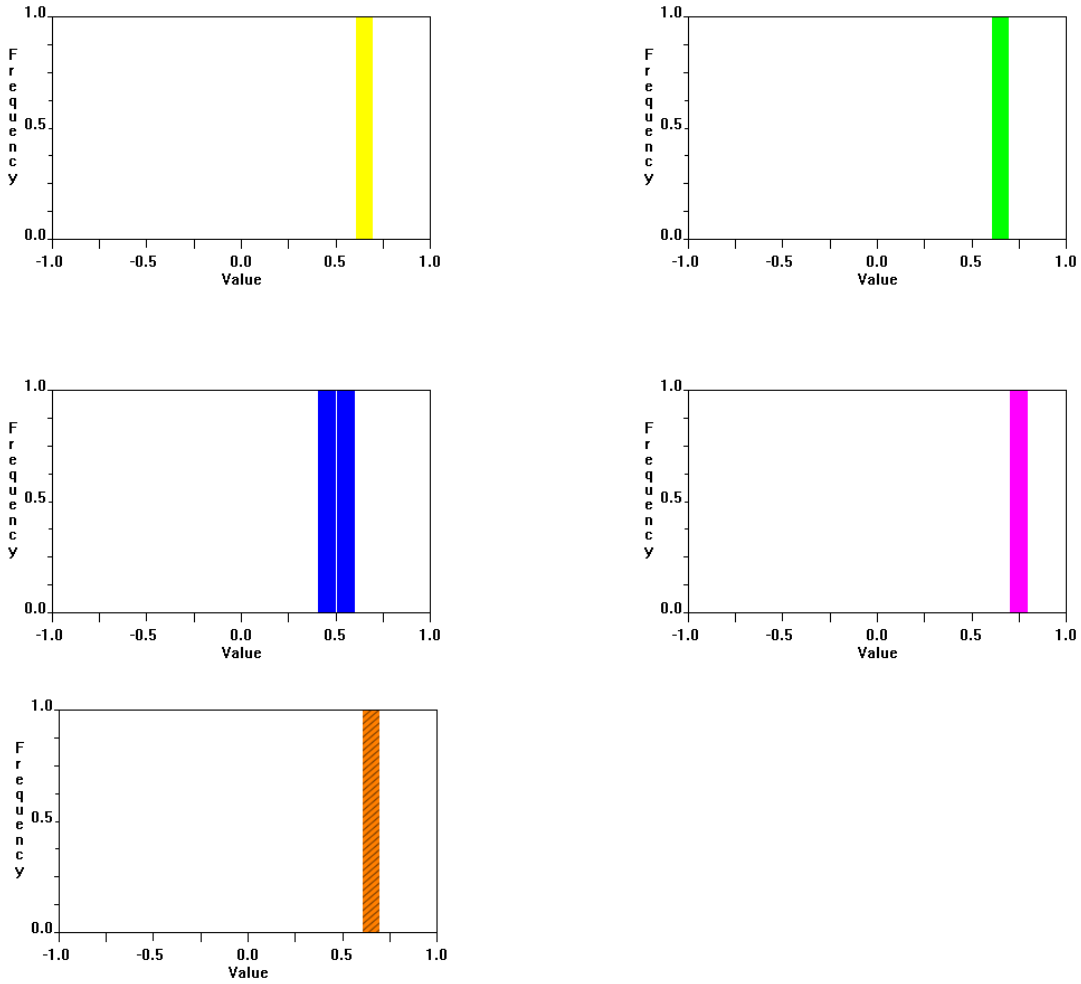

Figure 192 - PXRD Silhouettes

All but the blue clusters silhouettes have the samples all present in a single band. The blue cluster has the pure oxalic acid sample present in the band below 0.5 and the urea/oxalic acid mixture present in the band above 0.5. There are no fuzzy clusters for the dataset. This lack of fuzzy clusters, combined with only one sample being in a separate region in its silhouette, implies that the clustering is unambiguous for this dataset.

7.2.3 PXRD RE-RUN

Some of the samples have noticeable preferred orientation peaks present, for example samples MU, U+OA and U+SA as shown in Figure 193.

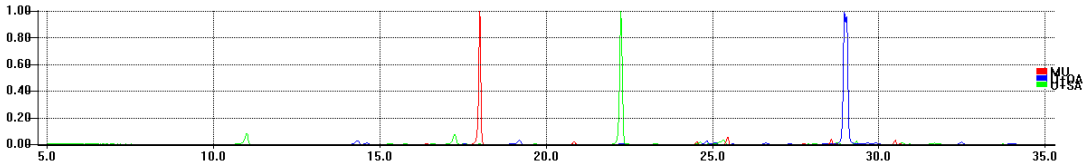

Figure 193 - PXRD Preferred Orientation

These peaks have been manually reduced in size so that the other peaks are not overwhelmed by them to see what effect this has on the datasets clustering.
The re-run PXRD dendrogram and MMDS plot, with preferred orientation peaks manually removed, are shown in Figure 194.

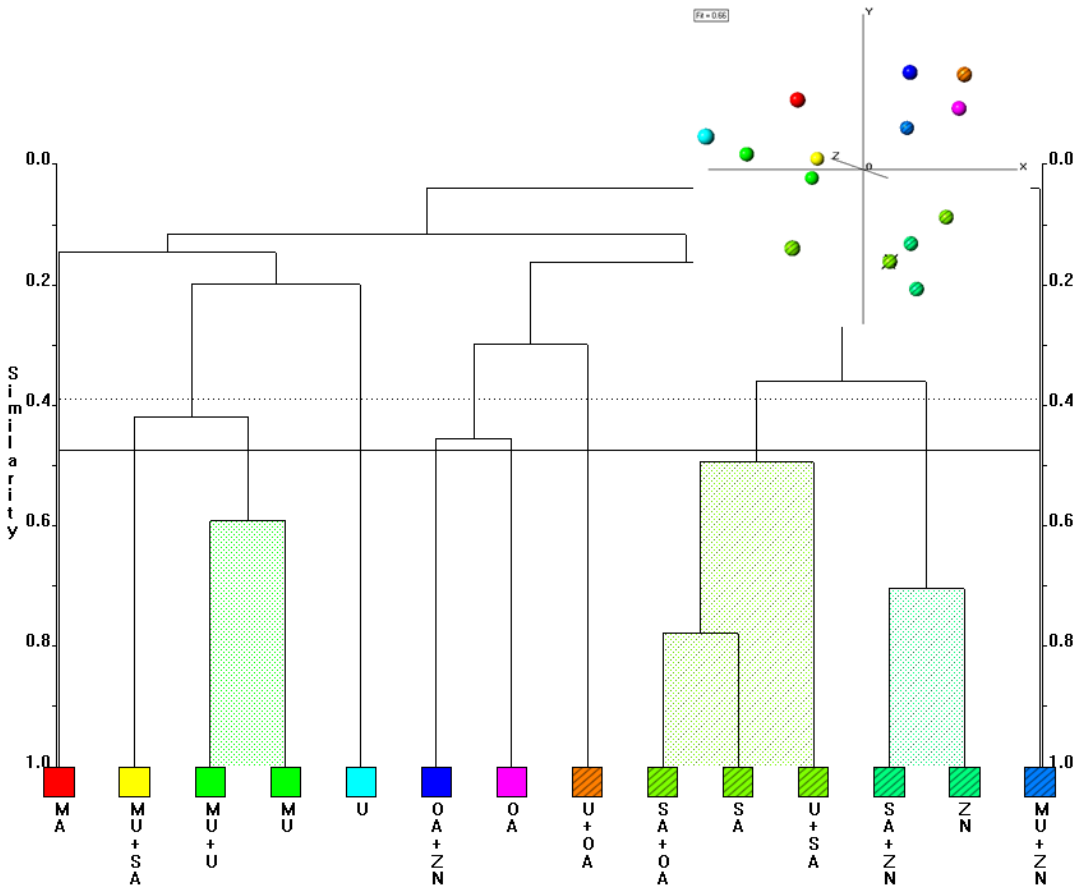

Figure 194 - Bulk Dataset PXRD Dendrogram and MMDS Plot re-run

The red cluster contains the pure malonic acid sample which is expected to be alone in predicted cluster 1. The yellow cluster contains sample MU+SA which was expected to appear in predicted cluster 3. The green cluster contains samples MU+U and MU which are predicted to make up cluster 4. The aquamarine cluster contains the pure urea sample which is expected to appear alone in cluster 2. The blue cluster contains sample OA+ZN which is expected to appear in cluster 6. The purple cluster contains sample OA which is expected to appear in cluster 6. The striped brown cluster contains sample U+OA from expected cluster 6. The striped green cluster contains samples SA+OA, SA and U+SA from expected cluster 3. The striped dark green cluster contains sample SA+ZN from expected cluster 3 and ZN from expected cluster 5. The striped blue cluster contains the methyl urea/zinc nitrate mixture from expected cluster 5.

The MMDS plot shows the clusters to again be diffuse with some intermixing between the light and dark striped green clusters. The rerun dendrogram has a score of 0.21, showing a substantial improvement in clustering over that seen in the original PXRD dendrogram.

7.2.4 RAMAN DATA

The dendrogram and MMDS Plot for the Raman data are shown in Figure 195.

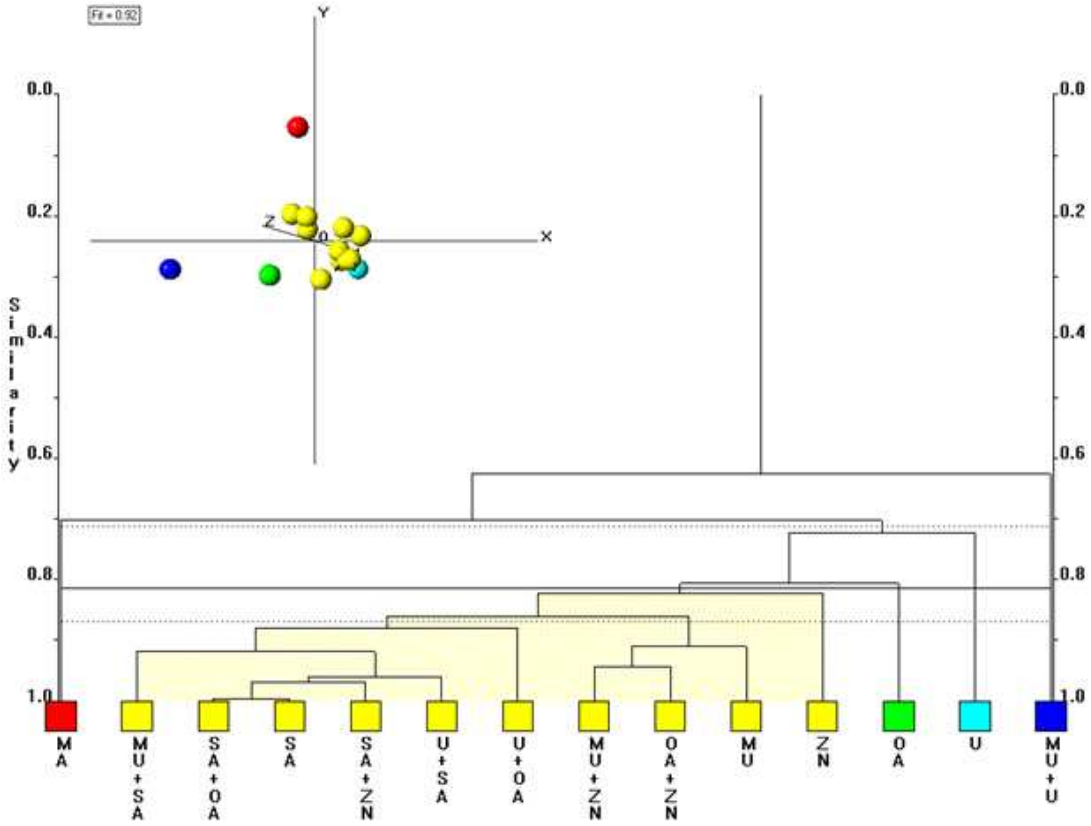

Figure 195 - Bulk Dataset Raman Dendrogram and MMDS Plot

The red cluster contains the pure malonic acid sample from expected cluster 1. The yellow cluster contains samples MU+SA, SA, SA+OA, SA+ZN and U+SA which all make up expected cluster 3, U+OA and OA+ZN from expected cluster 6, MU+ZN and ZN from expected cluster 5 and MU from expected cluster 4. The green cluster contains sample OA from expected cluster 6 while the aquamarine cluster contains the pure urea sample which was expected to be alone in expected cluster 2. The blue cluster contains sample MU+U from expected cluster 4.

The MMDS plot is much more tightly clustered than the one for the PXRD plot. This is, however, a common occurrence for Raman datasets. The Raman dendrogram has a score of 0.29, a small decline in clustering correctness from that seen in the rerun PXRD dendrogram.

The silhouettes for this dataset are shown in Figure 196.

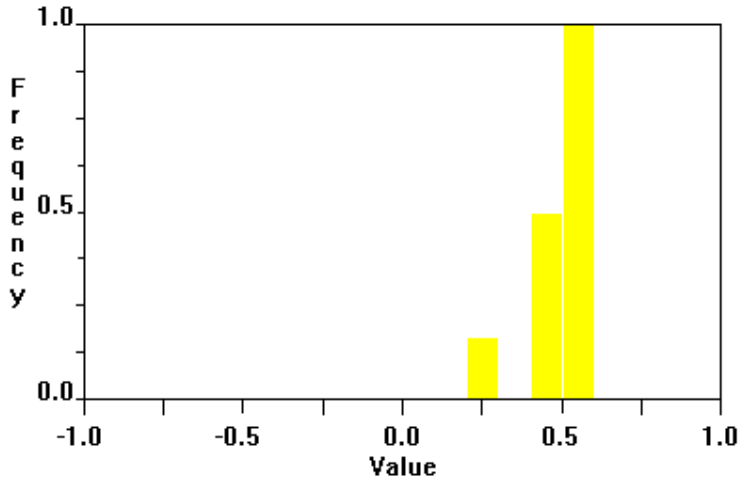

Figure 196 – Raman Silhouettes

As the yellow cluster is the only cluster to contain more than one sample, it is also the only cluster to have silhouettes.

The lowest band, at 0.25, contains the zinc nitrate sample. The second band, slightly lower than 0.5, contains the methyl urea sample, salicylic acid/zinc nitrate and the urea/oxalic acid sample. The uppermost band contains the remaining samples in the yellow cluster.

7.2.5 DSC DATA

The DSC dendrogram and MMDS plot are shown in Figure 197.

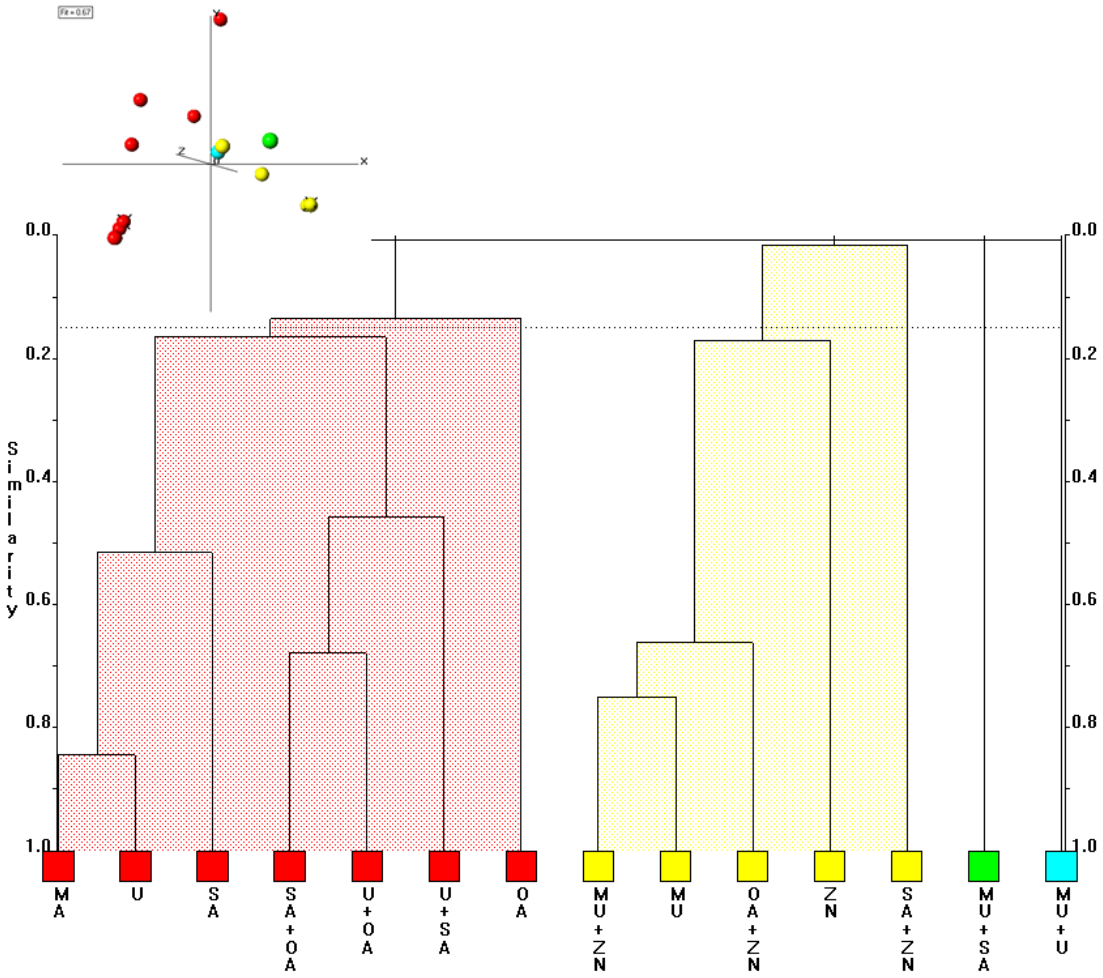

Figure 197 - Bulk Dataset DSC Dendrogram and MMDS Plot

The red cluster contains sample MA from expected cluster 1, U from expected cluster 2, SA, SA+OA and U+SA from expected cluster 3 and OA and U+OA from expected cluster 6. The yellow cluster contains samples MU+ZN and ZN from expected cluster 5, MU from expected cluster 4, OA+ZN from expected cluster 6 and SA+ZN from expected cluster 3. The green cluster contains sample MU+SA from expected cluster 3 while the aquamarine cluster contains sample MU+U from expected cluster 4.

The MMDS plot shows the first three samples (reading left to right in the MMDS plot) in the yellow cluster to be tightly grouped with the remainder diffuse. The red cluster is highly diffuse. The DSC dendrogram has a score of 0.5 showing half of the samples being misclustered.

The silhouettes are shown in Figure 198.

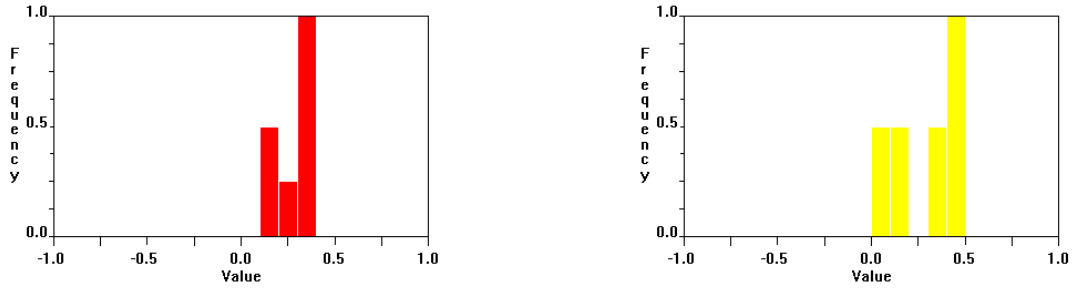

Figure 198 - DSC Silhouettes

For the red cluster the lowest band, which lies below 0.25, contains the pure oxalic acid and the urea/oxalic acid mixture. The band that lies on 0.25 contains the pure salicylic acid sample, with the remaining red cluster samples being in the final band.

For the yellow cluster, the lowest band, immediately above 0, contains the salicylic acid/zinc nitrate mixture. The next band contains the pure zinc nitrate sample while the band which lies just above 0.25 contains the pure methyl urea sample. The remaining yellow cluster samples are present in the final band.

No samples have been found for this dataset with cluster memberships less than 0.5 so no fuzzy clustering plot or information will be shown.

The cut-level for this dendrogram was adjusted downwards. The resulting dendrogram and MMDS plot can now be seen in Figure 199.

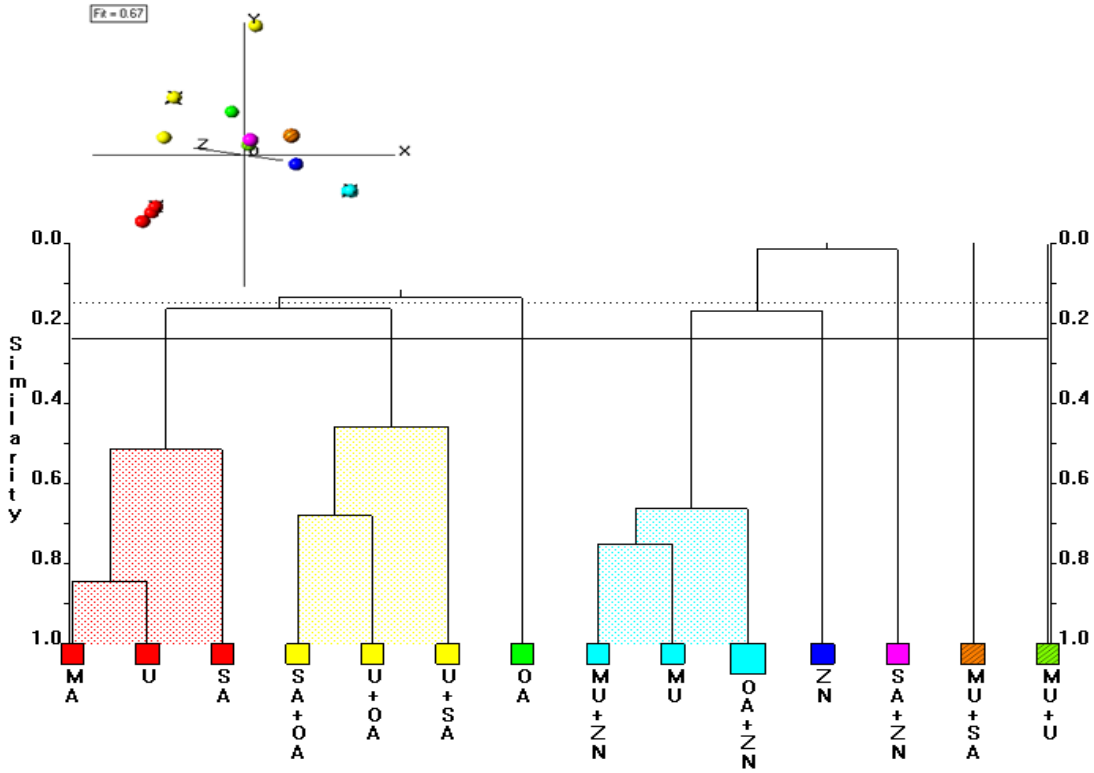

Figure 199 - Bulk Dataset DSC Dendrogram and MMDS Plot

The red cluster contains sample MA from expected cluster 1, U from expected cluster 2 and SA from expected cluster 3. The yellow cluster contains samples SA+OA and U+SA from expected cluster 3 and U+OA from expected cluster 6. The green cluster contains sample OA from expected cluster 6. The aquamarine cluster contains samples MU+ZN from expected cluster 5, MU from expected cluster 4 and OA+ZN from expected cluster 6. The blue cluster contains sample ZN from expected cluster 5. The purple cluster contains sample SA+ZN from expected cluster 3. The striped brown cluster contains sample MU+SA from expected cluster 3 while the striped green cluster contains sample MU+U from expected cluster 4. The adjustment of the cut-level did not cause any change in the score for this dendrogram, leaving it at 0.5.

7.2.6 IR DATA

The dendrogram and MMDS plot for the IR data are shown in Figure 200.

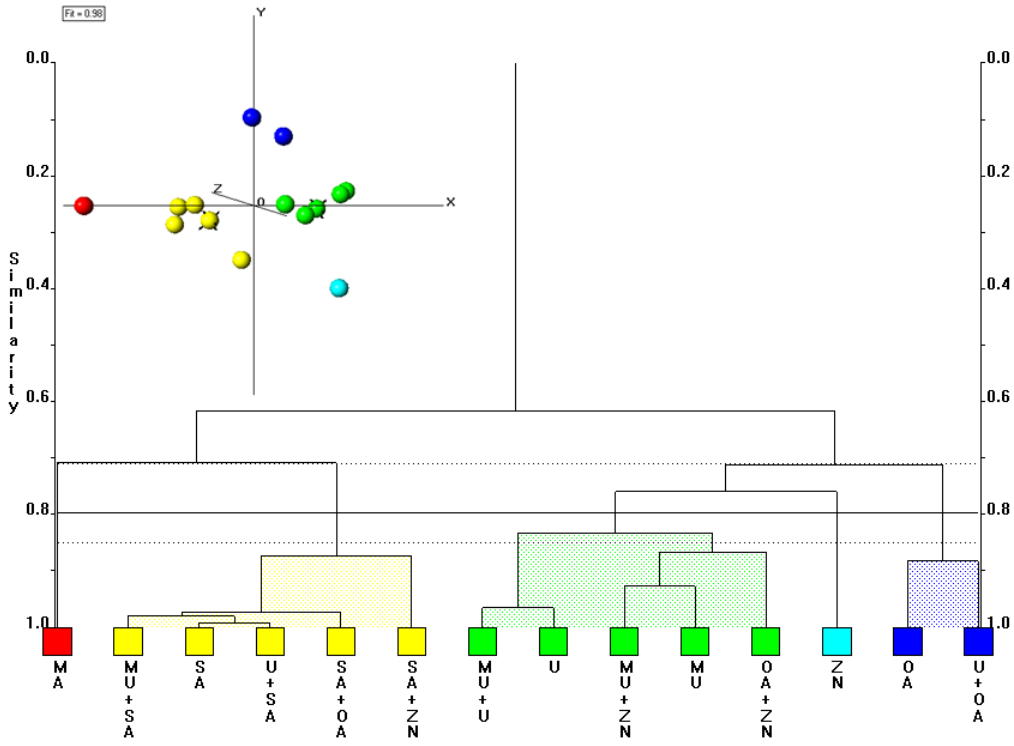

Figure 200 - Bulk Dataset IR Dendrogram and MMDS Plot

The red cluster contains the pure malonic acid sample which was expected to be alone in cluster 1. The yellow cluster contains samples MU+SA, SA, U+SA, SA+ZN and SA+OA

from expected cluster 3. The green cluster contains samples MU+U and MU from expected cluster 4, sample U from expected cluster 2, MU+ZN from expected cluster 5 and OA+ZN from expected cluster 6. The aquamarine cluster contains the pure zinc nitrate sample from expected cluster 5. The blue cluster contains samples OA and U+OA from expected cluster 6.

The MMDS plot shows each cluster to be clearly separated from one another, with the samples within the clusters being tightly grouped. The IR dendrogram has a score of 0.29, the same as that seen in the Raman dendrogram.

The silhouettes are shown in Figure 201.

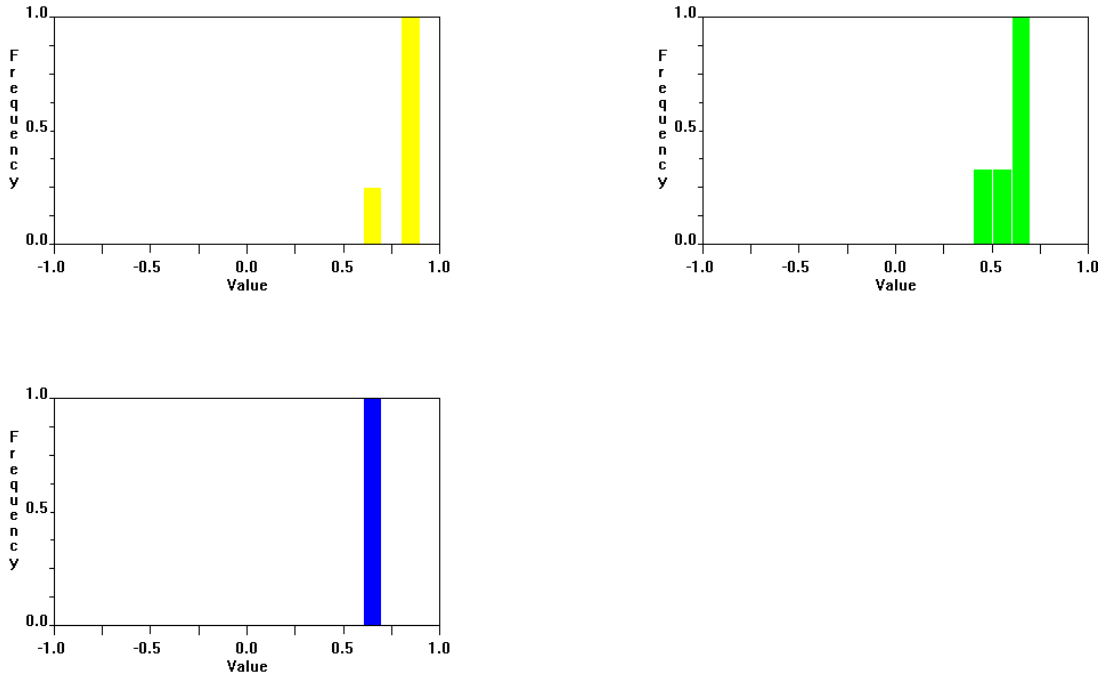

Figure 201 - IR Silhouettes

The yellow cluster has the salicylic acid/zinc nitrate sample in the lower band, present just below 0.75, with the upper band containing all remaining yellow cluster members. The green cluster has the oxalic acid/zinc nitrate sample in the lower band, below 0.5, with the band immediately on 0.5 contains the pure urea sample. The upper band contains the remaining green cluster sample. The blue cluster silhouette has one band with both samples present in it.

7.2.7 COMBINED DATASETS

The PXRD, Raman, DSC and IR datasets were combined using INDSCAL. The resulting dendrogram and MMDS plot are shown in Figure 202. The PXRD data with preferred orientation peaks removed is used for this combination.

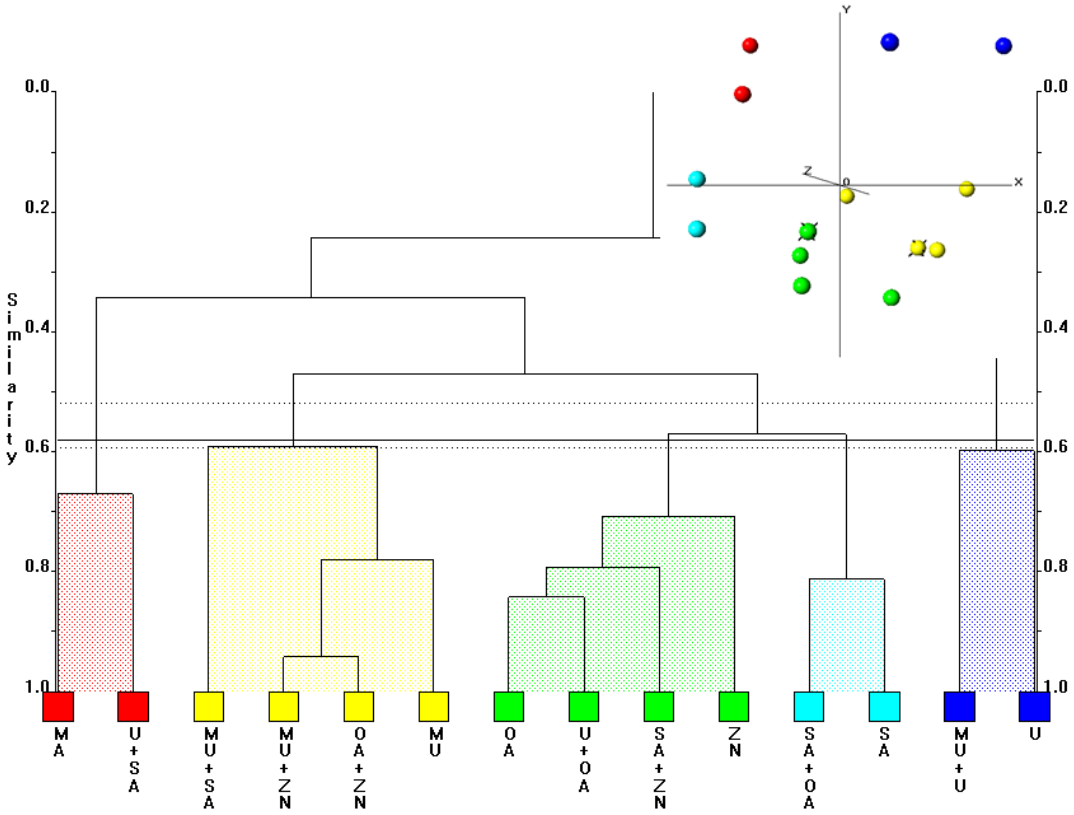

Figure 202 - Bulk Dataset Combined Dendrogram and MMDS Plot

The red cluster contains sample MA from expected cluster 1 and sample U+SA from expected cluster 3. The yellow cluster contains sample MU+SA from expected cluster 3, OA+ZN from cluster 6, MU+ZN from expected cluster 5 and MU from expected cluster 4. The green cluster contains samples OA and U+OA from expected cluster 6, SA+ZN from expected cluster 3 and ZN from expected cluster 5. The aquamarine cluster contains samples SA+OA and SA from expected cluster 3. The blue cluster contains samples U and MU+U from expected cluster 4.

The MMDS plot shows the clusters to be clearly separated from one another. Within the clusters, in particularly the yellow and green one, there is some highly noticeable spreading out of the samples. The combined dendrogram has a score of 0.43, a substantial decline from that seen in preceding datasets with the exception of the DSC dataset.

As a means of testing if this large decline in clustering correctness was due to the presence of the particularly poor DSC data the PXRD, Raman and IR datasets were combined

without the DSC dataset. The resulting dendrogram and MMDS plot for this are shown in Figure 203.

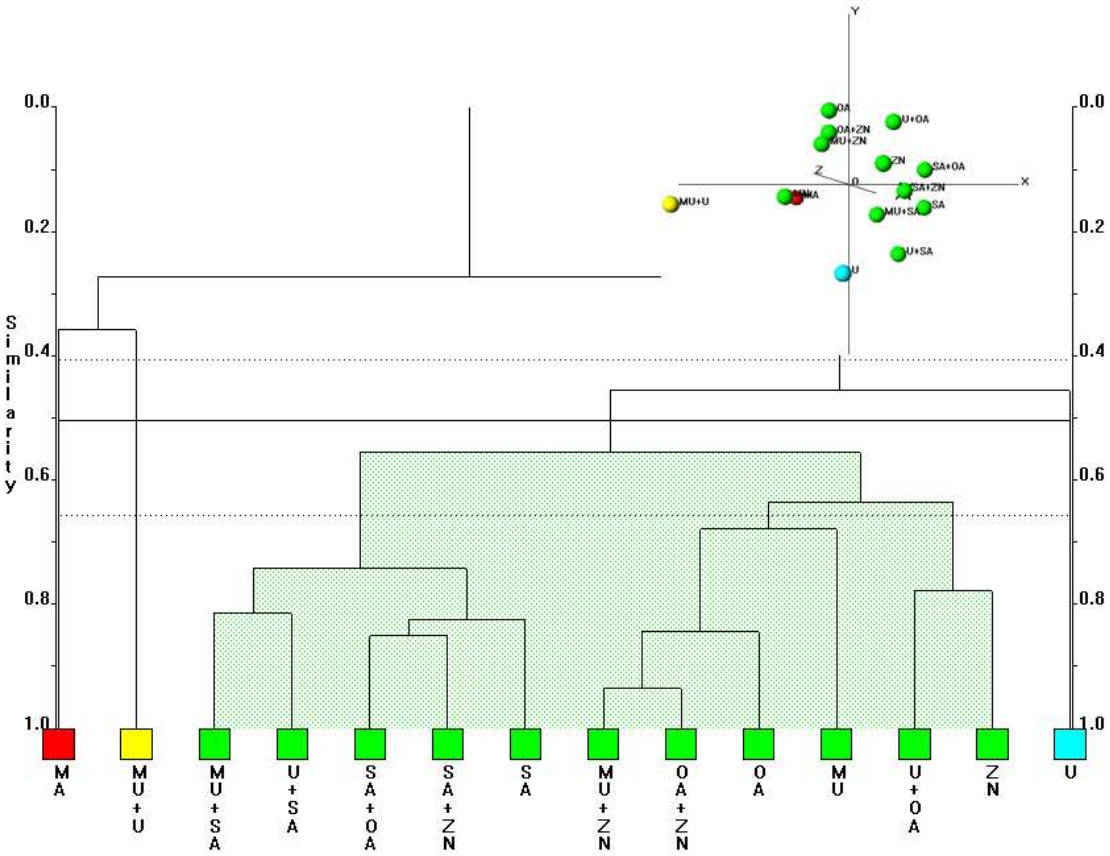

Figure 203 - Combined PXRD, Raman and IR Bulk dataset

This combined dendrogram has a score of 0.21, a sizeable improvement over the 0.43 seen in the previous combined dendrogram. This implies that, for this dataset, DSC data is a particularly poor choice of data type.

7.3 DERIVATIVE DATA

7.3.1 RAMAN

The first derivative Raman dendrogram and MMDS plot are shown in Figure 204.

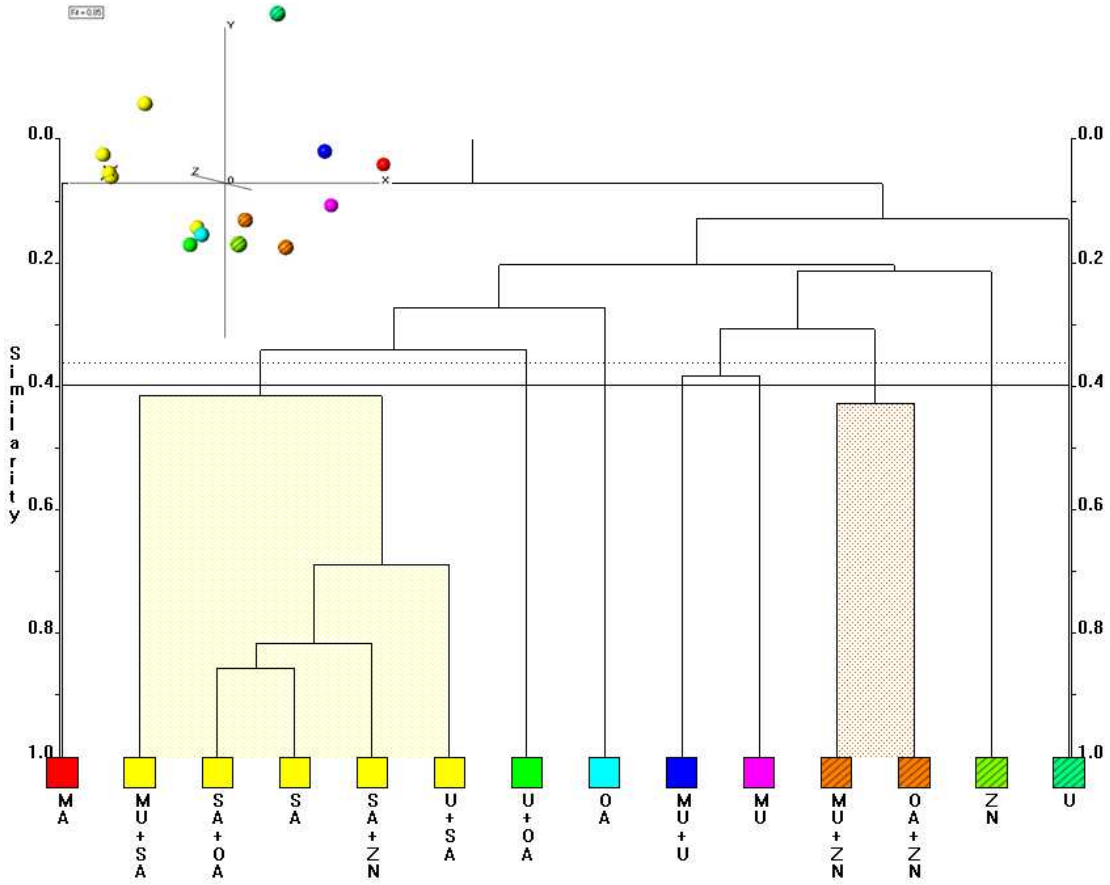

Figure 204 - First Derivative Raman Dendrogram and MMDS Plot

The red cluster contains the malonic acid sample which is expected to be in cluster 1 on its own. The yellow cluster contains samples MU+SA, SA+OA, SA, U+SA, SA+ZN and SA+OA which make up expected cluster 3. The green cluster contains sample U+OA from expected cluster 6. The aquamarine cluster contains the oxalic acid sample from expected cluster 6. The blue cluster contains sample MU+U from expected cluster 4. The purple cluster contains sample MU from expected cluster 4. The striped brown cluster contains sample MU+ZN from expected cluster 5 and sample OA+ZN from expected cluster 6. The light green cluster contains sample ZN from expected cluster 5 and the dark green cluster contains sample U from expected cluster 2.

The MMDS plot shows some intermixing of the clusters. The first derivative Raman dendrogram has a score of 0.29, showing reasonable clustering despite the poor appearance of the dendrogram.

The second derivative Raman dendrogram and MMDS plot are shown in Figure 205.

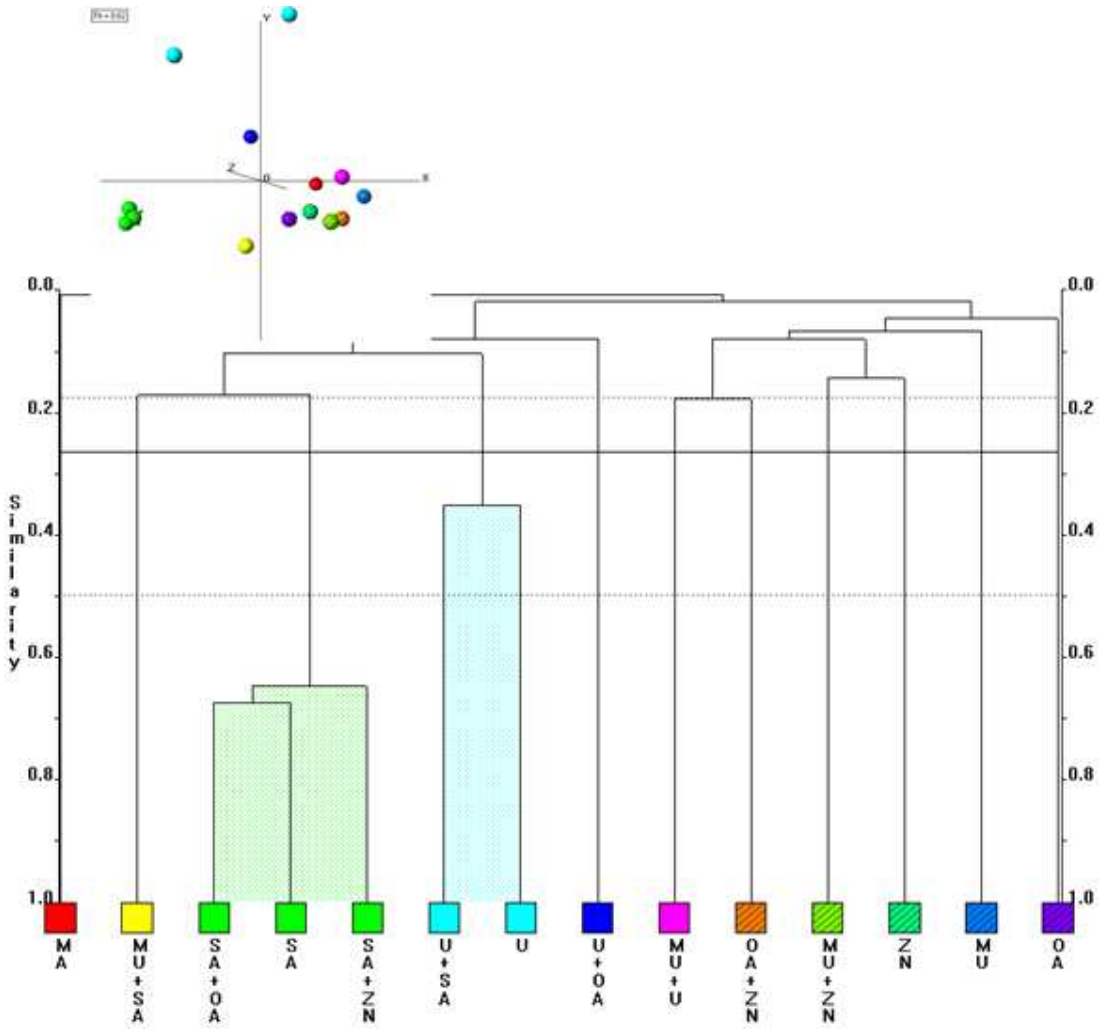

Figure 205 – Second Derivative Raman Dendrogram and MMDS Plot

The red cluster contains the malonic acid sample which is the lone sample in cluster 1. The yellow cluster contains sample MU+SA from expected cluster 3. The green cluster contains samples SA+OA, SA and SA+ZN from expected cluster 3. The aquamarine cluster contains sample U from expected cluster 2 and U+SA from expected cluster 3. The blue cluster contains sample U+OA from expected cluster 6 while the purple cluster contains sample MU+U from expected cluster 4. The striped brown cluster contains the sample OA+ZN from expected cluster 6. The striped light green cluster contains sample

MU+ZN from expected cluster 5 and the striped dark green cluster contains sample ZN from expected cluster 5. The striped blue sample contains sample MU from expected cluster 4 and the striped purple sample OA from expected cluster 6.
The MMDS plot shows the green, aquamarine and blue clusters to be well separated. The remaining clusters are closely grouped. The second derivative Raman dendrogram has a score of 0.5, a much poorer result than that for the first derivative dendrogram.

7.3.2 IR

The first derivative IR dendrogram and MMDS plot are shown in Figure 206.

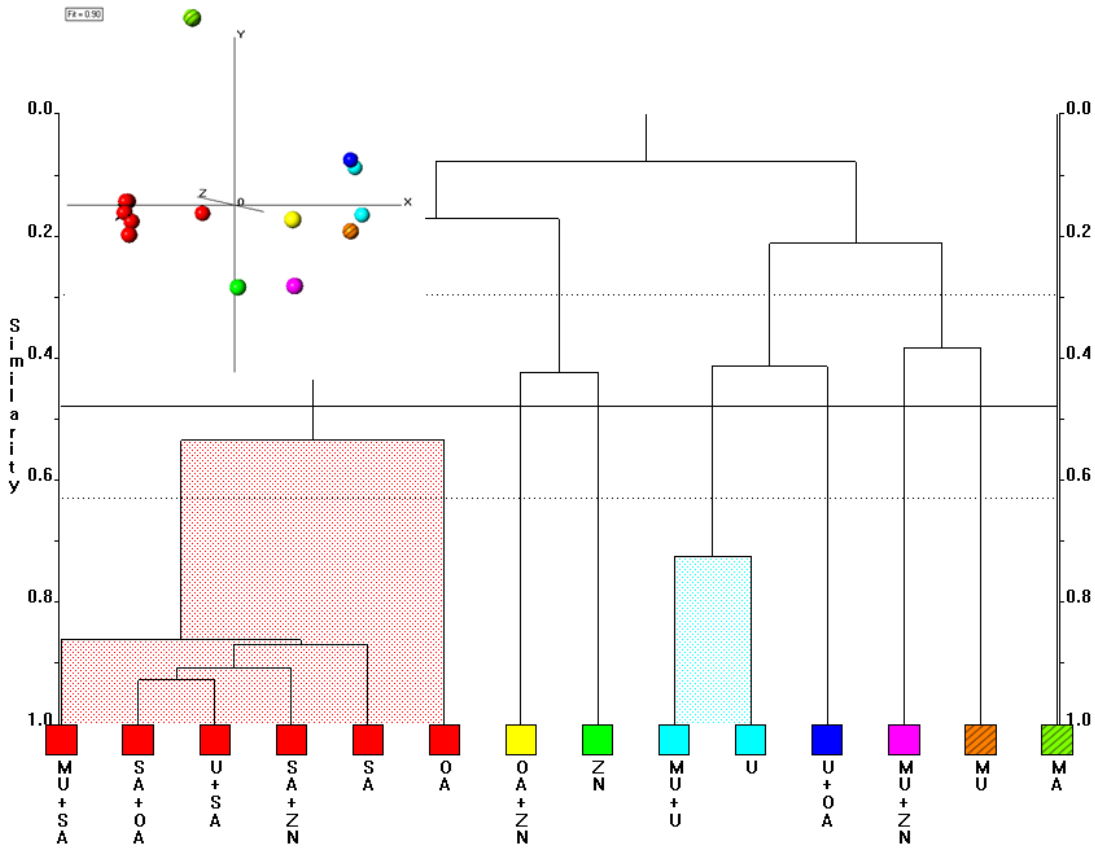

Figure 206 - First Derivative IR Dendrogram and MMDS Plot

The red cluster contains samples MU+SA, SA, U+SA, SA+ZN and SA+OA from expected cluster 3 and OA from expected cluster 6. The yellow cluster contains sample OA+ZN from expected cluster 6. The green cluster contains sample ZN from expected cluster 5. The aquamarine cluster contains sample MU+U from expected cluster 4 and U from expected cluster 2. The blue cluster contains samples U+OA from expected cluster 6. The purple cluster contains sample U+OA from expected cluster 6. The striped brown cluster

contains sample MU from expected cluster 4. The striped green cluster contains sample MA from expected cluster 1.

The MMDS plot shows the clustering to be highly diffuse. The lone blue sample and the aquamarine cluster, although appearing very close together, are actually diffuse when viewed along the x-axis. The first derivative IR dendrogram has a score of 0.21, an improvement on the 0.29 seen for the original IR dataset.

The second derivative IR dendrogram and MMDS plot are shown in Figure 207.

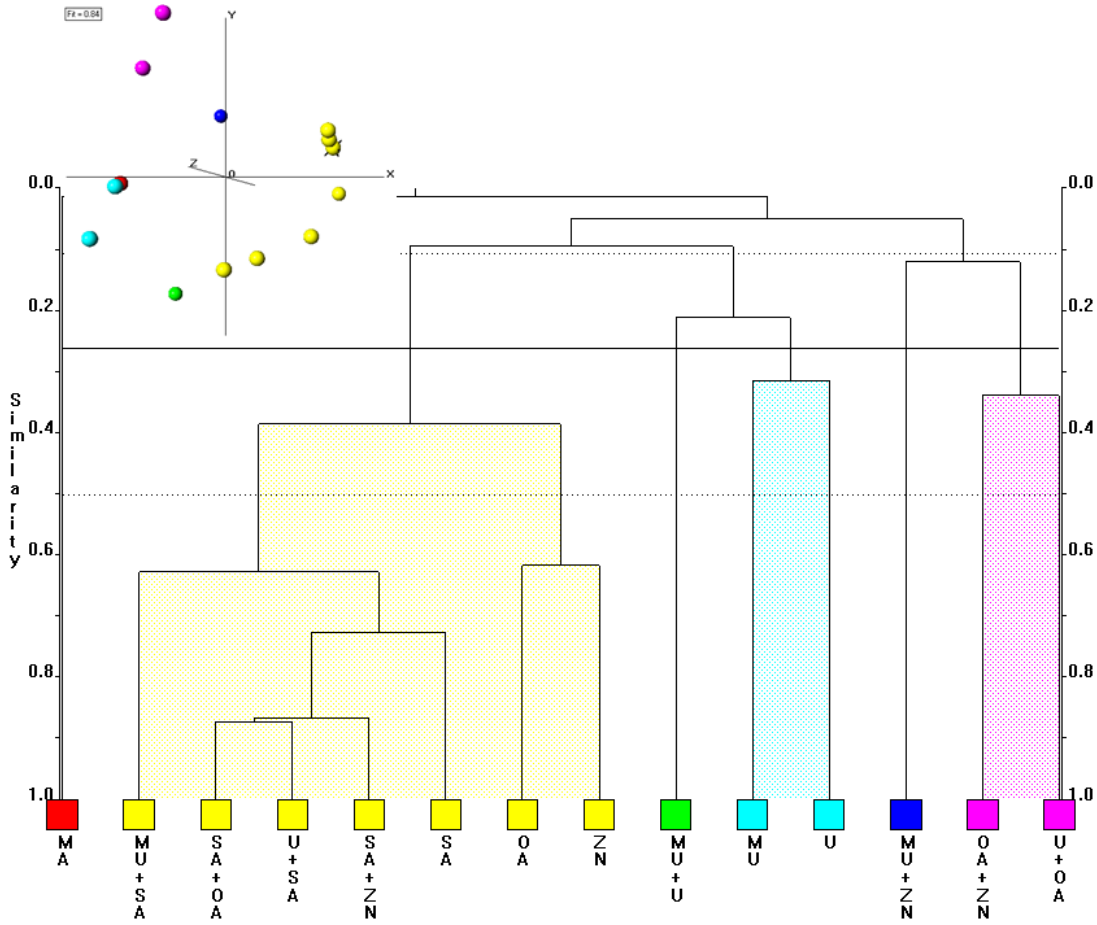

Figure 207 - Second Derivative IR Dendrogram and MMDS Plot

The red cluster contains the pure malonic acid sample which is expected to form expected cluster 1. The yellow cluster contains samples MU+SA, SA, U+SA, SA+ZN and SA+OA from expected cluster 3, sample ZN from expected cluster 5 and sample OA from expected cluster 6. The green cluster contains the methyl urea/urea sample from expected cluster 4. The aquamarine cluster contains sample MU from expected cluster 4 and U from expected cluster 2. The blue cluster contains MU+ZN from expected cluster 5. The purple cluster contains samples OA+ZN and U+OA from expected cluster 6.

The MMDS plot shows the clusters, especially the large yellow one, to be highly diffuse. The dendrogram has a score of 0.21, the same as that for the first derivative IR dendrogram and a small improvement over the result seen for the original IR dendrogram.

7.4 FLOWCHART

All of the possible combinations of pre-processing were applied to the PXRD dataset and the results compared to the optimal clustering. The number of misplaced samples is shown in Table 37.

	Score
no processing	0.21
denoise	0.24
background	0.29
background and denoise	0.21

Table 37 – Score for PXRD pre-processing options

As can be seen no pre-processing and background removal and denoise both give the optimal clustering with denoise being marginally poorer and background removal only slightly poorer than denoise.

The flowchart for this result is shown in Figure 208.

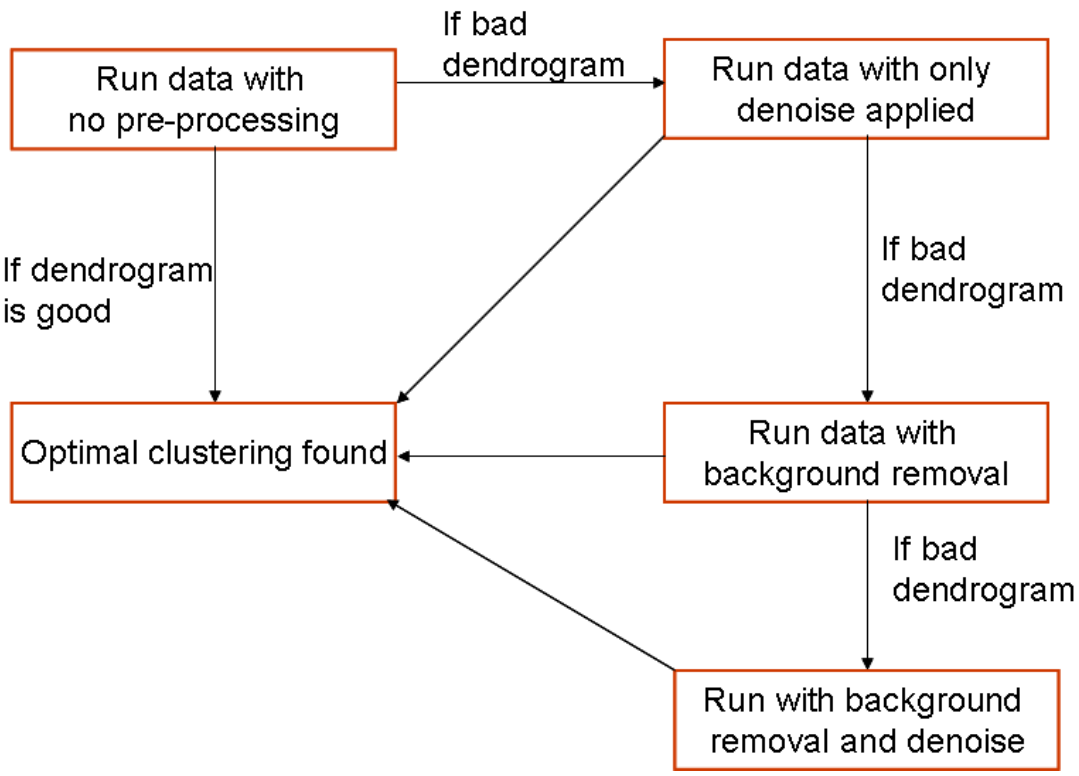

Figure 208 - Flowchart for 32 sample dataset

As the differences are small, only 2 samples difference between no pre-processing and denoise, denoise is given second priority due to it involving the least pre-processing. A 'bad' dendrogram is as previously defined.

7.5 QUANTITATIVE ANALYSIS

The materials are compared using the PolySNAP quantitative analysis mode. The results of this are shown in Table 37. For the PXRD data, the dataset with the preferred orientation peaks shortened is used. The data were compared using the SVD method. The results, with pre-processing applied to the data, are shown in Table 38 and 39.

PXRD	Samples	Actual	Predicted	Difference	Raman	Samples	Actual	Predicted	Difference	IR	Samples	Actual	Predicted	Difference
	MU+U	50:50	71.2:28.8	21.20		MU+U	50:50	23:77	27.00		MU+U	50:50	14.4:85.6	35.60
	MU+SA	50:50	85.2:14.8	35.20		MU+SA	50:50	23.1:76.9	26.90		MU+SA	50:50	10.1:89.9	39.90
	MU+ZN	50:50	57.4:42.6	7.40		MU+ZN	50:50	49.5:50.5	0.50		MU+ZN	50:50	53:47	3.00
	U+SA	50:50	51.2:48.8	1.20		U+SA	50:50	34:66	16.00		U+SA	50:50	5.6:94.4	44.40
	U+OA	50:50	68.6:31.4	18.60		U+OA	50:50	36.2:63.8	13.80		U+OA	50:50	56.4:43.6	6.40
	SA+OA	50:50	60.9:39.1	10.90		SA+OA	50:50	91.1:8.9	41.10		SA+OA	50:50	90.9:9.1	40.90
	SA+ZN	50:50	37.5:62.5	12.50		SA+ZN	50:50	94.5:5.5	44.50		SA+ZN	50:50	62.9:37.1	12.90
	OA+ZN	50:50	85.4:14.6	35.40		OA+ZN	50:50	80.2:19.8	30.20		OA+ZN	50:50	54.2:45.8	4.20
		Mean absolute difference		17.80			Mean absolute difference		25.00			Mean absolute difference		23.41
		RMS difference		6.29			RMS difference		8.84			RMS difference		8.28
		Max absolute difference		17.60			Max absolute difference		24.50			Max absolute difference		20.99
		Min absolute difference		16.60			Min absolute difference		19.50			Min absolute difference		20.41

Table 38 – Data from Mixtures in Manual Analysis Mode

PXRD	Samples	Actual	Processed Predicted 1	Difference 1	Raman	Samples	Actual	Processed Predicted 1	Difference 1	IR	Samples	Actual	Processed Predicted 1	Difference 1
	MU+U	50:50	76.4:23.6	26.40		MU+U	50:50	71.1:28.9	21.10		MU+U	50:50	20.1:79.9	29.90
	MU+SA	50:50	91.2:8.8	41.20		MU+SA	50:50	9.3:90.7	40.70		MU+SA	50:50	32.9:67.1	17.10
	MU+ZN	50:50	61.3:38.7	11.30		MU+ZN	50:50	38.4:61.6	11.60		MU+ZN	50:50	22.3:77.7	27.70
	U+SA	50:50	50.1:49.9	0.10		U+SA	50:50	31.6:68.4	18.40		U+SA	50:50	1.8:98.2	48.20
	U+OA	50:50	15.2:84.8	34.80		U+OA	50:50	12.9:87.1	48.67		U+OA	50:50	83.5:16.5	33.50
	SA+OA	50:50	61:39	11.00		SA+OA	50:50	99.7:0.3	49.70		SA+OA	50:50	98.6:1.4	48.60
	SA+ZN	50:50	37.6:62.4	12.40		SA+ZN	50:50	71:29	21.00		SA+ZN	50:50	62:38	12.00
	OA+ZN	50:50	95.2:4.8	45.20		OA+ZN	50:50	64.5:35.5	14.50		OA+ZN	50:50	28.4:78.6	21.60
	Mean absolute difference			22.80		Mean absolute difference			28.21		Mean absolute difference			29.83
	RMS difference			8.06		RMS difference			9.97		RMS difference			10.54
	Max absolute difference			22.70		Max absolute difference			21.49		Max absolute difference			18.78
	Min absolute difference			22.40		Min absolute difference			16.61		Min absolute difference			17.83

Processed Predicted 1 - background remove and smoothed

Table 39 - Data from Mixtures in Manual Analysis Mode with Pre-processing 1

PXRD	Samples	Actual	Processed Predicted 2	Difference 2	Raman	Samples	Actual	Processed Predicted 2	Difference 2	IR	Samples	Actual	Processed Predicted 2	Difference 2
	MU+U	50:50	71.2:28.8	21.20		MU+U	50:50	42.4:57.6	7.60		MU+U	50:50	14.4:85.6	35.60
	MU+SA	50:50	85.2:14.8	35.20		MU+SA	50:50	23.1:76.9	26.90		MU+SA	50:50	10.1:89.9	39.90
	MU+ZN	50:50	57.5:42.5	7.50		MU+ZN	50:50	49.5:50.5	0.50		MU+ZN	50:50	73.7:26.3	23.70
	U+SA	50:50	51.1:48.9	1.10		U+SA	50:50	34:66	16.00		U+SA	50:50	5.6:94.4	44.40
	U+OA	50:50	31.5:68.5	18.50		U+OA	50:50	36.2:63.8	13.80		U+OA	50:50	56.4:43.6	6.40
	SA+OA	50:50	61:39	11.00		SA+OA	50:50	91.1:8.9	41.10		SA+OA	50:50	90.9:9.1	40.90
	SA+ZN	50:50	37.5:62.5	12.50		SA+ZN	50:50	94.5:5.5	44.50		SA+ZN	50:50	62.9:37.1	12.90
	OA+ZN	50:50	85.1:14.9	35.10		OA+ZN	50:50	80.2:19.8	30.20		OA+ZN	50:50	45.8:54.2	4.20
		Mean absolute difference		5.92			Mean absolute difference		22.58			Mean absolute difference		26.00
		RMS difference		2.09			RMS difference		7.98			RMS difference		9.19
		Max absolute difference		29.28			Max absolute difference		22.08			Max absolute difference		21.80
		Min absolute difference		4.82			Min absolute difference		21.93			Min absolute difference		18.40

Processed Predicted 2 - smoothed

Table 40 - Data from Mixtures in Manual Analysis Mode with Pre-processing 2

For PXRD, a predicted result is said to closely match the actual values if the values are within 10% of each other in either direction.
For the PXRD data two of the samples (U+SA and MU+ZN) initially match the prediction. With the first pre-processing method applied, U+SA is the only sample to match and for the second pre-processing method, samples U+SA and MU+ZN both again match
For the Raman data sample MU+ZN is the only sample to match. No patterns match when the background is removed and peak smoothing is applied. With just smoothing applied sample MU+ZN again matches. If the allowed difference is increased to 15% MU+ZN and U+OA now match with no processing applied. With background removal applied sample MU+ZN again matches. With smoothing applied samples MU+ZN, U+OA and MU+U now match.
For the IR data samples MU+ZN, U+OA and OA+ZN match. With the first processing option applied no samples match and with the second OA+ZN and U+OA are the only samples to match. With the allowed difference increased to 15% MU+ZN, U+OA and OA+ZN match. With background removal and smoothing applied only sample SA+ZN matches and with just smoothing applied U+SA and OA+ZN are the only ones to match.

7.6 CONCLUSIONS

- The bulk materials dataset offers further evidence as to how the combined usage of Raman, DSC and IR data can offset the effects of a PXRD dataset with preferred orientation problems.
- The results from DSC are particularly poor for this dataset. This however is due to their being additional complexities as one of the materials is a hydrate. This can also be due to the presence of acids and bases which may react on heating the mixtures.
- The huge effects that pre-processing options can have on the determination of composition of datasets is also further demonstrated. For Raman data, allowing a variance of 15% rather than the 10% and collecting the data with smoothing applied gives the optimal results. For IR the data gives equivalent results at both 10 and 15% allowed variance and does best with no applied processing.

CHAPTER 8 AN UNKNOWN DATASET

8.1 THE DATA

The unknown dataset was a dataset, consisting of PXRD and Raman data, supplied by Professor Chris Frampton at Pharmorphix. This was a blind test with the data being supplied without any information as to what each pattern represented. The dataset contains 48 samples, of which the identity of each sample type was unknown. It was known that each sample is from a pure material, with no mixtures of these materials present. The number of different materials present was unknown; however it is known that each material can be present more than once. The X-ray data was collected on a Bruker C2 GADDS and was collected three times. The first collection, henceforth referred to as X-ray 1, was collected over a range of 7-35°, the second, henceforth referred to as X-ray 2, was collected over a range of 3-30° and the third, henceforth referred to as X-ray 3, was collected over a range of 16-44°. The Raman data was collected on a Bruker SENTINEL, integrated into the Bruker C2, and was collected over a range of 250-2300cm^{1-}.
Table 40 shows the file name that was assigned to each sample when the dataset was delivered, along with the new name that was assigned to each sample to allow ease of use with PolySNAP.

Sample Number	Sample Name	Name in PolySNAP	Sample Number	Sample Name	Name in PolySNAP
1	CSF-135-20-1-A1	01	25	CSF-135-20-2-B1	25
2	CSF-135-20-1-A2	02	26	CSF-135-20-2-B2	26
3	CSF-135-20-1-A3	03	27	CSF-135-20-2-B3	27
4	CSF-135-20-1-A4	04	28	CSF-135-20-2-B4	28
5	CSF-135-20-1-A5	05	29	CSF-135-20-2-B5	29
6	CSF-135-20-1-A6	06	30	CSF-135-20-2-B6	30
7	CSF-135-20-1-A7	07	31	CSF-135-20-2-B7	31
8	CSF-135-20-1-A8	08	32	CSF-135-20-2-B8	32
9	CSF-135-20-1-B1	09	33	CSF-135-20-3-A1	33
10	CSF-135-20-1-B2	10	34	CSF-135-20-3-A2	34
11	CSF-135-20-1-B3	11	35	CSF-135-20-3-A3	35
12	CSF-135-20-1-B4	12	36	CSF-135-20-3-A4	36
13	CSF-135-20-1-B5	13	37	CSF-135-20-3-A5	37
14	CSF-135-20-1-B6	14	38	CSF-135-20-3-A6	38
15	CSF-135-20-1-B7	15	39	CSF-135-20-3-A7	39
16	CSF-135-20-1-B8	16	40	CSF-135-20-3-A8	40
17	CSF-135-20-2-A1	17	41	CSF-135-20-3-B1	41
18	CSF-135-20-2-A2	18	42	CSF-135-20-3-B2	42
19	CSF-135-20-2-A3	19	43	CSF-135-20-3-B3	43
20	CSF-135-20-2-A4	20	44	CSF-135-20-3-B4	44
21	CSF-135-20-2-A5	21	45	CSF-135-20-3-B5	45
22	CSF-135-20-2-A6	22	46	CSF-135-20-3-B6	46
23	CSF-135-20-2-A7	23	47	CSF-135-20-3-B7	47
24	CSF-135-20-2-A8	24	48	CSF-135-20-3-B8	48

Table 41 - Unknown Dataset

8.2 DATASET CLUSTERING

As this is an unseen dataset there were no prior expectations for what clustering should appear.

8.2.1 PXRD DATASETS

The dendrogram and MMDS plot for the X-ray 1 dataset are shown in Figure 205.

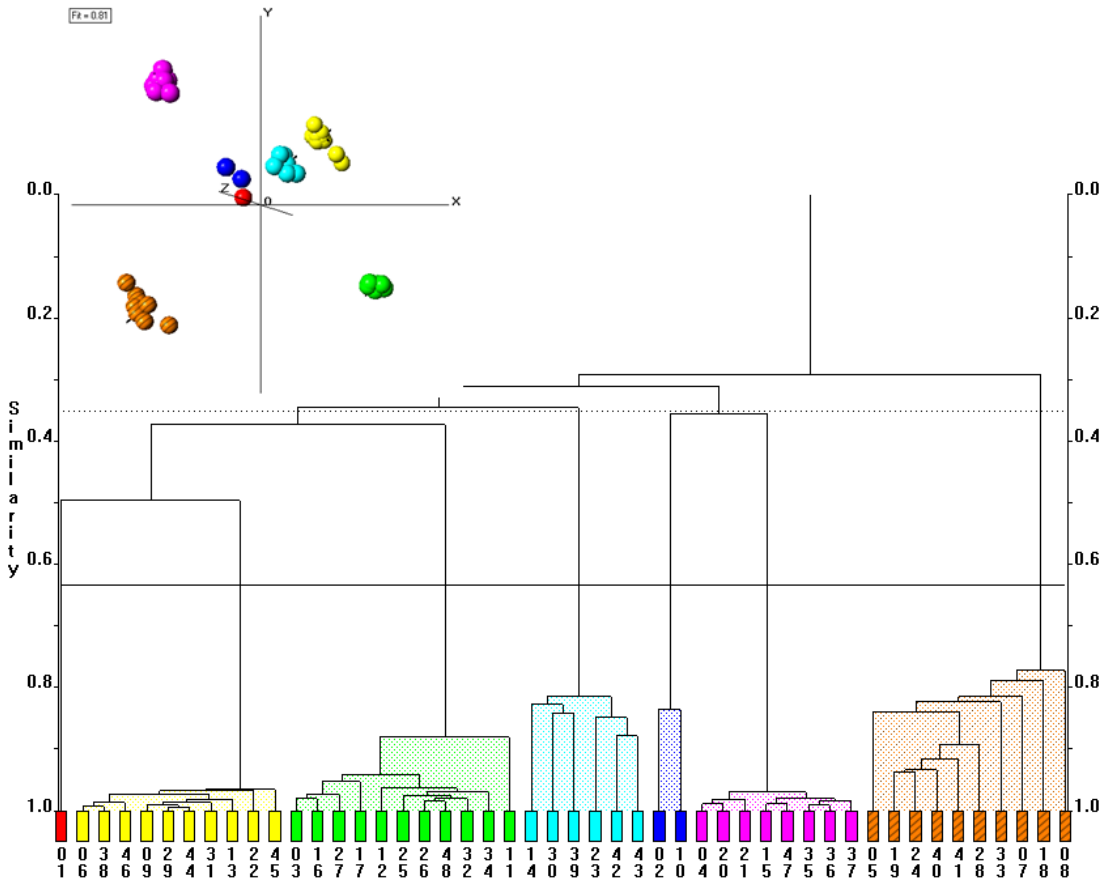

Figure 209 - X-ray 1 Dataset Dendrogram and MMDS Plot

The dendrogram suggests that there are seven different materials present in the dataset. Sample 01, appearing in the red cluster, appears to be the only instance of this material included within the dataset. Samples 06, 09, 13, 22, 29, 31, 38, 44, 45 and 46 are all present in the yellow cluster, suggesting that they may all be the same material. Samples 03, 11, 12, 16, 17, 25, 26, 27, 32, 34 and 48 are all present in the green cluster, suggesting that they are all the same material. Samples 14, 23, 30, 39, 42 and 43 are all present in the in the aquamarine cluster, suggesting that they may all be of the same material. Samples 02

and 10 are likely to be both of the same material as they are all present in the blue cluster. Samples 04, 15, 20, 21, 35, 36, 37 and 47 are all likely to be the same material as they are all present in the purple cluster. Finally samples 05, 07, 08, 18, 19, 24, 28, 33, 40 and 41 are present in the striped brown cluster suggesting they are all similar.

The MMDS plot shows the samples to be clearly separated. The lone sample in the red cluster appears to be lying close to the blue cluster; however rotating the MMDS plot reveals this to not be the case.

The dendrogram and MMDS plot for the X-ray 2 dataset are shown in Figure 206.

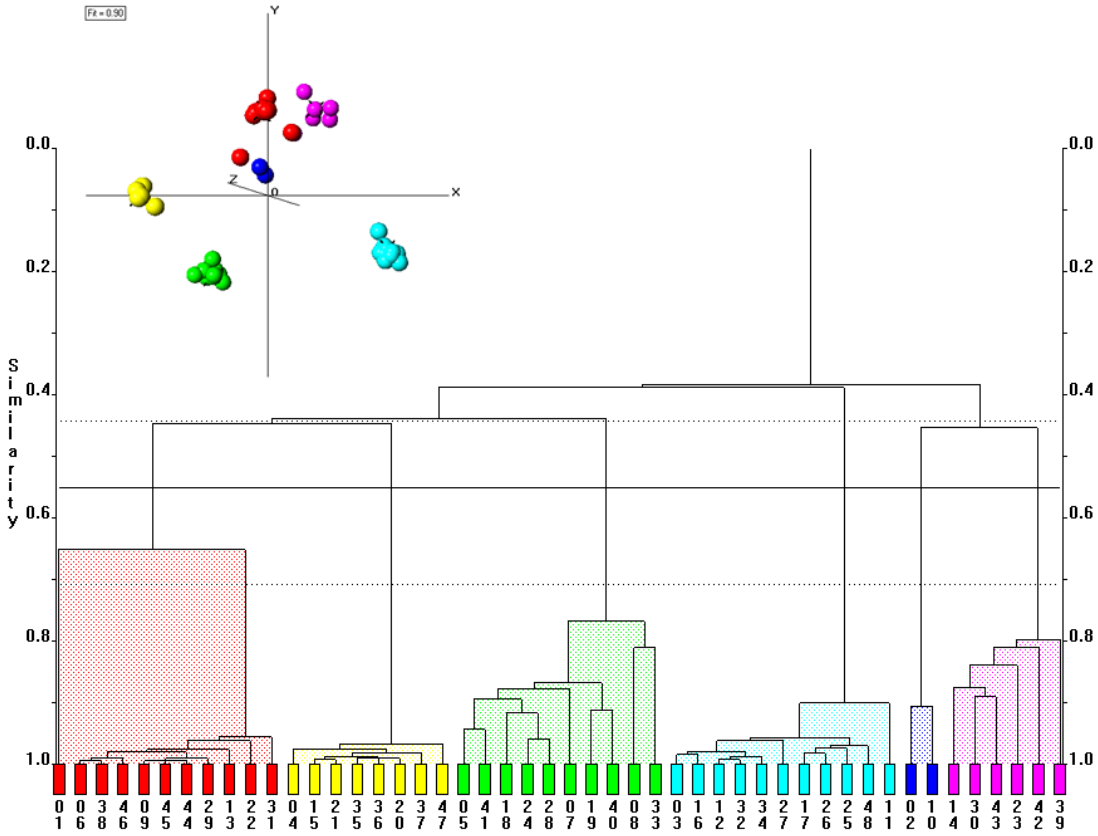

Figure 210 - X-ray 2 Dataset Dendrogram and MMDS Plot

The yellow cluster contains samples 04, 15, 20, 21, 35, 36, 37 and 47, which were all clustered together in the purple cluster in dataset X-ray 1. The green cluster contains samples 05, 07, 08, 18, 19, 24, 28, 33, 40 and 41 which were all clustered together in the striped brown cluster in dataset X-ray 1. The aquamarine cluster contains samples 03, 11, 12, 16, 17, 25, 26, 27, 32, 34 and 48 which were all clustered together in the green cluster in dataset X-ray 1. The blue cluster contains samples 02 and 10 which were also present in the blue cluster in dataset X-ray 1. The purple cluster contains samples 14, 23, 30, 39, 42 and 43, which were all clustered together in the aquamarine cluster in dataset X-ray 1. As

all of these samples are unchanged between datasets X-ray 1 and X-ray 2 it is highly probable that the materials present within each of these clusters will be the same as one another within each cluster.

The only change for this dataset comes in the red cluster, which contains a combination of the red cluster, containing sample 01 and the yellow cluster, containing samples 06, 09, 13, 22, 29, 31, 38, 44, 45 and 46. These samples can be separated again by lowering the cut-level, without affecting any of the other clusters and also while staying within the calculated upper and lower limits for the number of clusters. There are three theories for why this has occurred.

1) The first theory is that sample 01 is a poorer quality pattern in the first of the two datasets studied so far and is actually the same material as the second cluster.
2) The second theory is that sample 01 is of a poorer quality in the second dataset and is indeed different as suggested in dataset X-ray 2.
3) The third theory is that neither pattern is of a poorer quality, but that the patterns are highly similar due to sample 01 being a different polymorph of the material that is present in the second cluster.

Normally two different polymorphs of the same material will have different PXRD patterns, thus discounting theory 3. The earlier work on sulfathiazole however has shown that in some cases two polymorphs can have similar patterns. For the sulfathiazole dataset this was clearly demonstrated by polymorphs 2 and 3. In order to confirm which of these theories is correct, the sample 01 pattern will be compared to the most representative sample in the second cluster in each of the two datasets so far.

For the X-ray 1 dataset, sample 01 was compared with sample 44, which is the most representative sample in the second cluster. The overlay is shown in Figure 207.

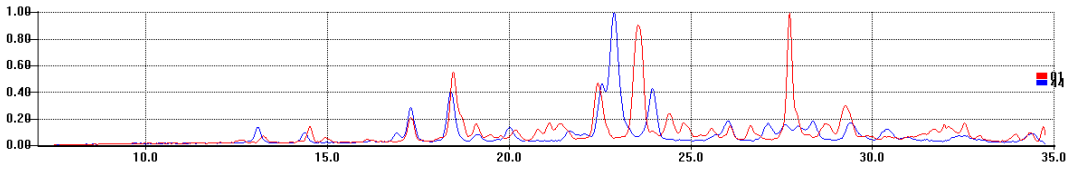

Figure 211 - Overlay of Samples 44 and 01 from Dataset X-ray 1

Neither of the two patterns appears to be of poor quality, suggesting that the first theory is incorrect.

For the X-ray 2 dataset, the cut-level is adjusted to separate these samples into two separate clusters; sample 44 is given as the most representative sample of the new cluster that

appears. As such it has been overlaid with sample 01 to allow a comparison of the two materials PXRD patterns. The overlay is shown in Figure 208.

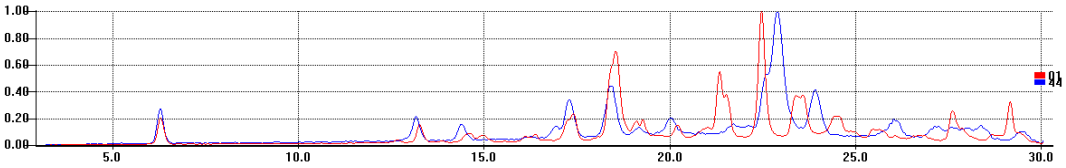

Figure 212 - Overlay of Samples 44 and 01 from Dataset X-ray 2

Neither of the patterns appears to be of poorer quality suggesting that the second theory is incorrect.

As both sets of overlays contain good quality data, and some noticeable differences are present between the two patterns in both datasets, theory 3 – where sample 01 is believed to be a different polymorph of the material present in the second cluster, is believed to be correct. This hypothesis will be further tested in the remaining datasets.

The X-ray 3 dataset dendrogram and MMDS plot are shown in Figure 209.

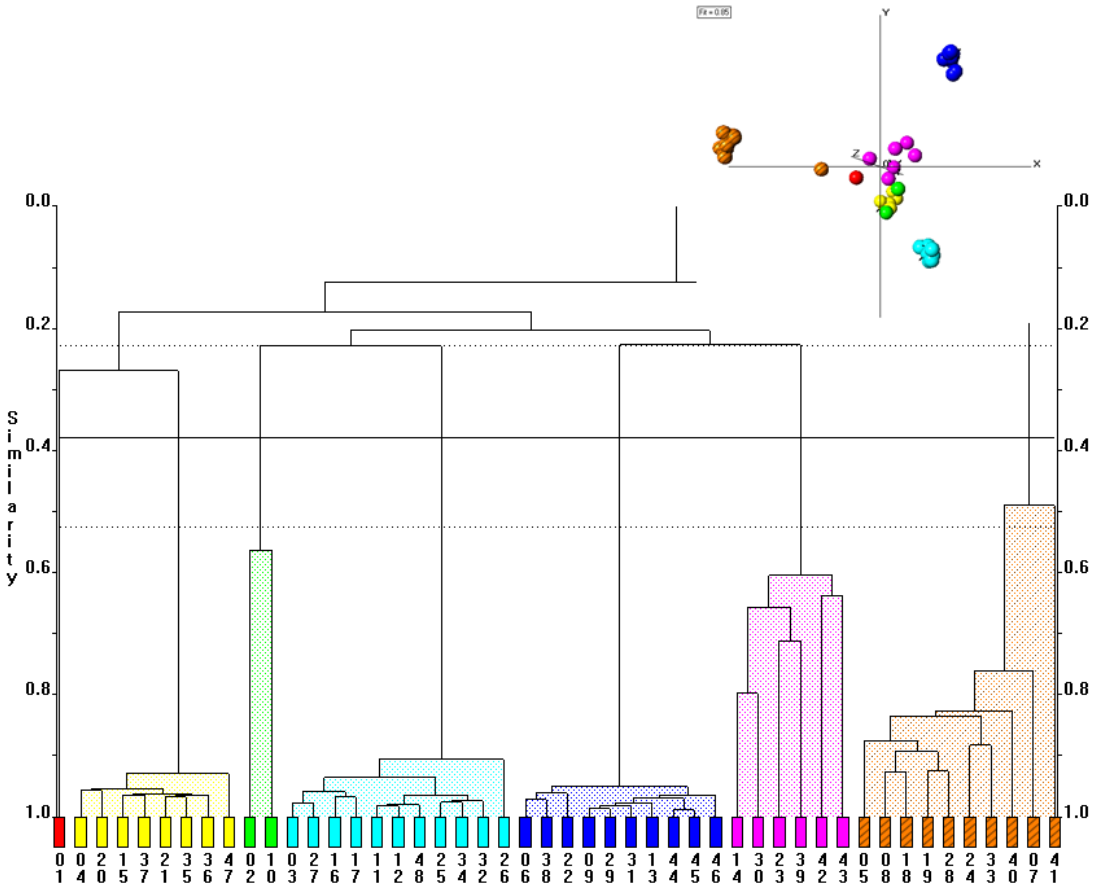

Figure 213 - X-ray 3 Dataset Dendrogram and MMDS Plot

The red cluster contains sample 01. This sample will be examined in further detail. The yellow cluster contains samples 04, 15, 20, 21, 35, 36, 37 and 47 which have previously been grouped together. The green cluster contains samples 02 and 10, which have been grouped together in both of the previous datasets. The aquamarine cluster contains samples 03, 11, 12, 16, 17, 25, 26, 27, 32, 34 and 48, which have been clustered together in both of the previous datasets. The blue cluster contains samples 06, 09, 13, 22, 29, 31, 38, 44, 45 and 46. These samples were clustered together in dataset X-ray 1 and with sample 01 in dataset X-ray 2 so further study of this dataset shall be carried out. The purple cluster contains samples 14, 23, 30, 39, 42 and 43 which have previously been clustered together. The striped brown cluster contains samples 05, 07, 08, 18, 19, 24, 28, 33, 40 and 41 which have been clustered together in the previous datasets.

As sample 01 is closely tied to a different cluster than in the previous dataset it shall be compared to the most representative sample in that cluster, which is sample 36. Figure 210 shows an overlay of these.

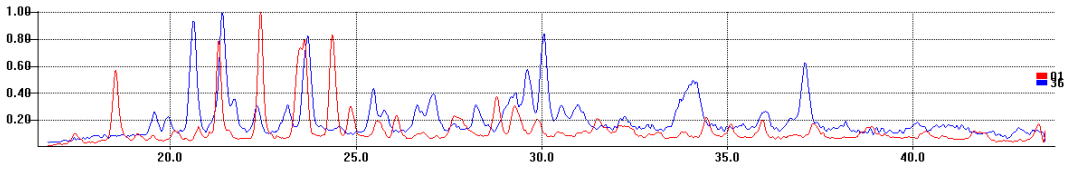

Figure 214 - Overlay of Samples 36 and 01 from Dataset X-ray 3

The overlay shows these two samples to be clearly different.

Figure 211 shows an overlay of samples 01 and the most representative sample from the cluster that it has previously been most closely linked to, sample 44.

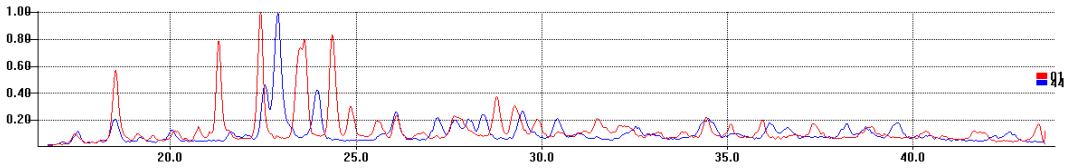

Figure 215 - Overlay of Samples 44 and 01 from Dataset X-ray 3

Once again sample 01 has a clearly different pattern from sample 44. Due to the clusters not being closely tied, the hypothesis that it is a polymorph off the materials present in the second cluster cannot be confirmed at this time.

The optimal dataset from all three x-ray datasets appears to be dataset X-ray 1 as this shows clear separation between the clusters while the samples within each cluster all show clear similarities. A large change in the cut-level would be required to either separate the samples within the clusters or merge any of the clusters.

8.2.2 RAMAN DATASET

The dendrogram and MMDS plot for the Raman data is shown in Figure 212.

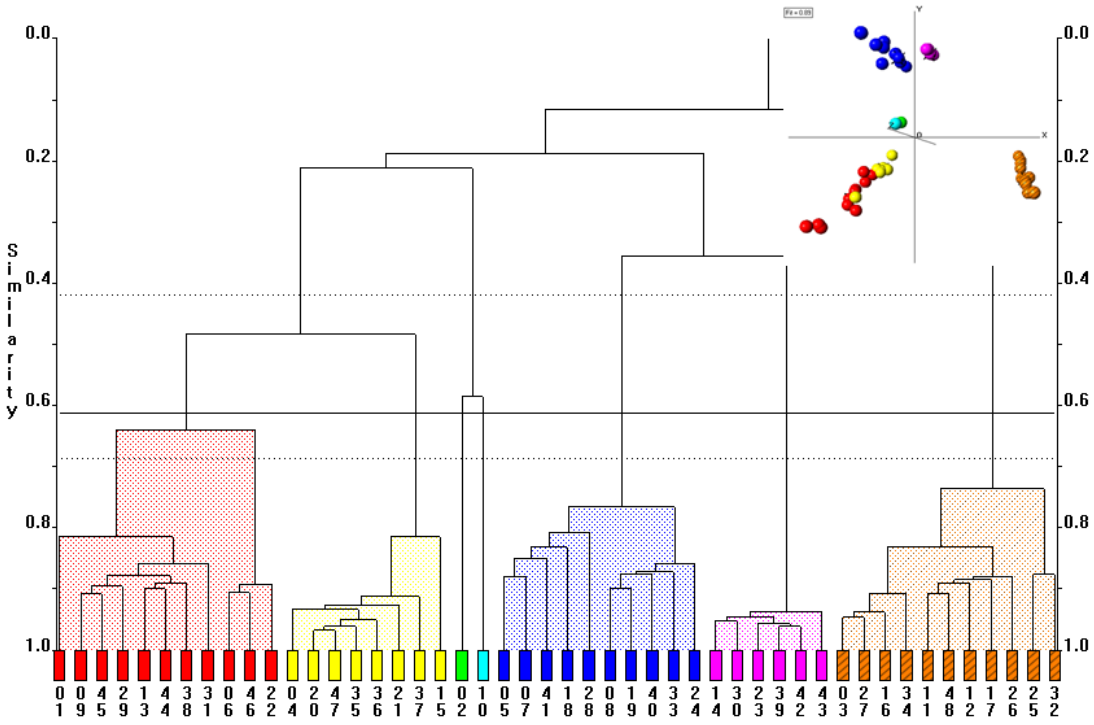

Figure 216 - Raman Data Dendrogram and MMDS Plot

The red cluster contains samples 09, 13, 22, 29, 31, 38, 44, 45 and 46 as well as sample 01. This time sample 01 cannot be separated by a cut-level adjustment. This will again be studied in more detail. The yellow cluster contains samples 04, 15, 20, 21, 35, 36, 37 and 47 which have previously been clustered together. Samples 02 and 10, which have previously been clustered together, are now in separate clusters. These clusters can be merged by raising the cut-level without affecting any other cluster. These samples will again be studied in further detail. The blue cluster contains samples 05, 07, 08, 18, 19, 24, 28, 33, 40 and 41 which have previously been grouped together. The purple cluster contains samples 14, 23, 30, 39, 42 and 43 which have previously been grouped together. The striped brown cluster contains samples 03, 11, 12, 16, 17, 25, 26, 27, 32, 34 and 48 which have previously been grouped together.

As it is not possible to separate sample 01 from the red cluster, sample 44, which has previously been the most representative sample, will be compared to sample 01. The overlay of this is shown in Figure 213.

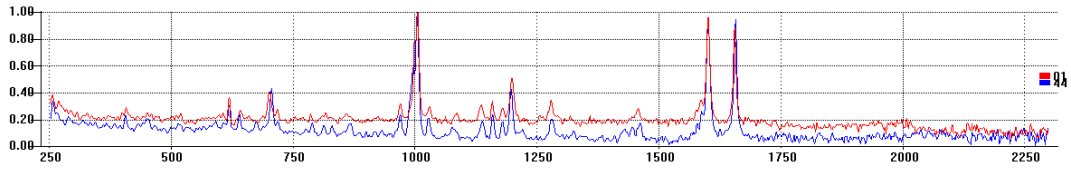

Figure 217 - Overlay of Samples 44 and 01 from Raman Dataset

All of the major bands match closely between these two samples. This suggests that the theory of sample 01 being a polymorph of the material present in the remainder of this cluster is correct.

The comparison of samples 02 and 10 is shown in Figure 214.

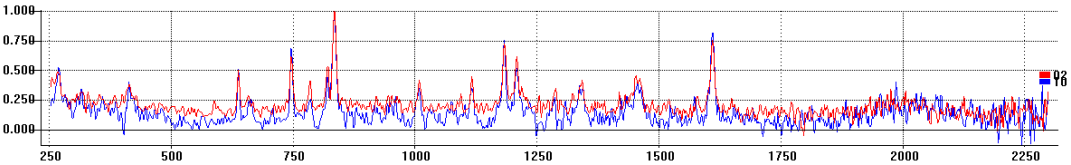

Figure 218 – Overlay of Samples 02 and 10 from Raman Dataset

These spectra again appear to be match closely. This, combined with these samples being clustered together in the previous three datasets, suggests that the materials are the same.

8.2.3 RAMAN DERIVATIVES

The Raman data has both first and second derivatives applied to it. The dendrogram for the first derivative is shown in Figure 215.

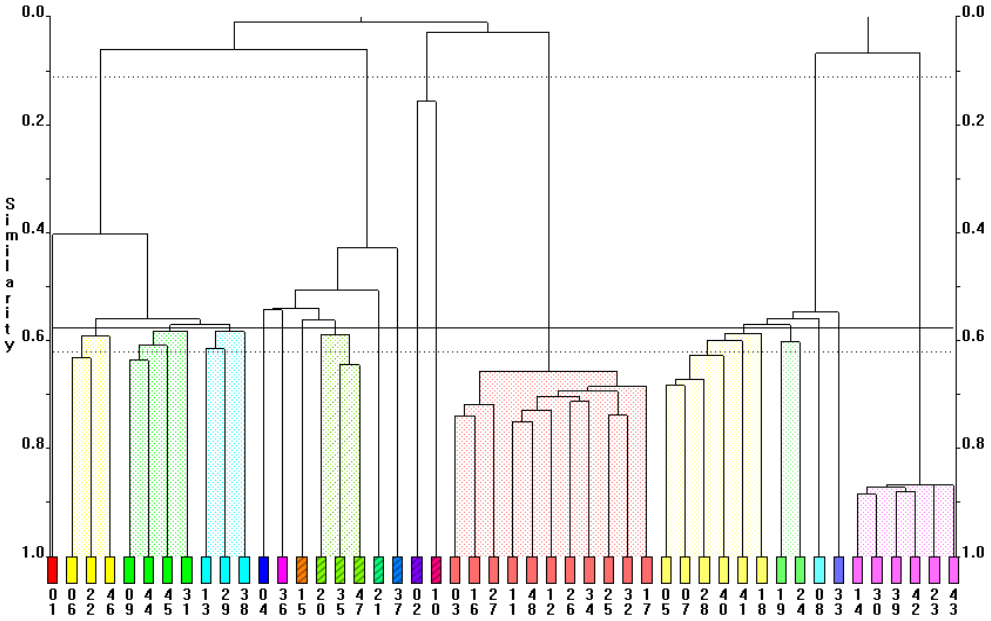

Figure 219 - First Derivative Raman Dendrogram

The clustering appears to be poor; however an adjustment of the cut-level upwards, while staying within the upper and lower estimate of number of clusters, yields a familiar result. This is shown in Figure 216.

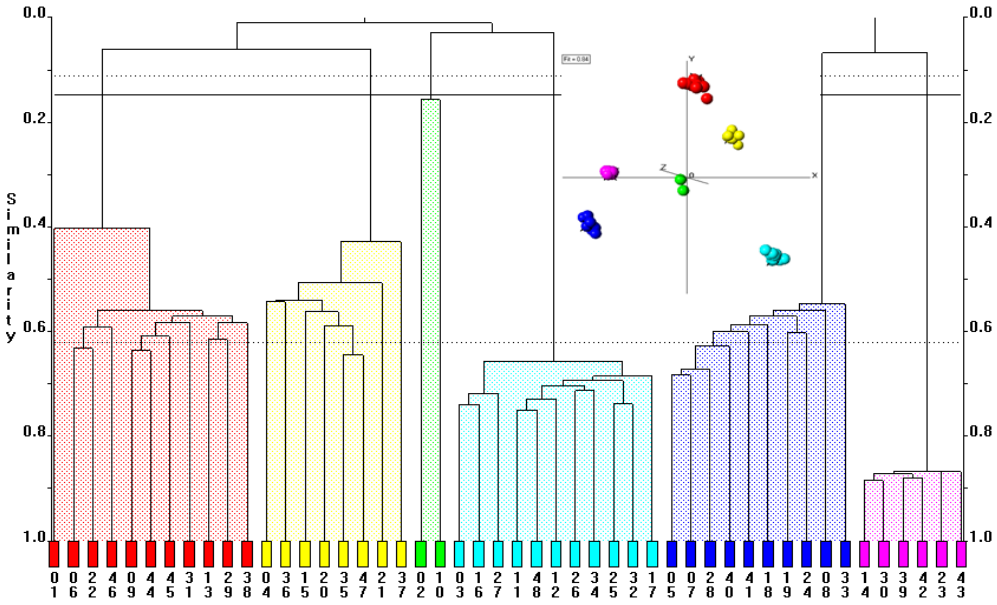

Figure 220 - Dendrogram and MMDS Plot for First Derivative Data with Adjusted Cut-level

The red cluster now contains samples 01, 06, 09, 13, 22, 29, 31, 38, 44, 45. Sample 01 is once again being grouped with the materials which it is believed to be a different polymorph of. The yellow cluster contains samples 04, 15, 20, 21, 35, 36, 37 and 47 which have previously been grouped together. The green cluster contains samples 02 and 10 which have previously been clustered together. The aquamarine cluster contains samples 03, 11, 12, 16, 17, 25, 26, 27, 32, 34 and 48 which have previously been grouped together. The blue cluster contains samples 05, 07, 08, 18, 19, 24, 28, 33, 40 and 41 which have previously been clustered together. The purple cluster contains samples 14, 23, 30, 39, 42 and 43 which have previously been grouped together.

The dendrogram and MMDS plot for the second derivative Raman data are shown in Figure 217.

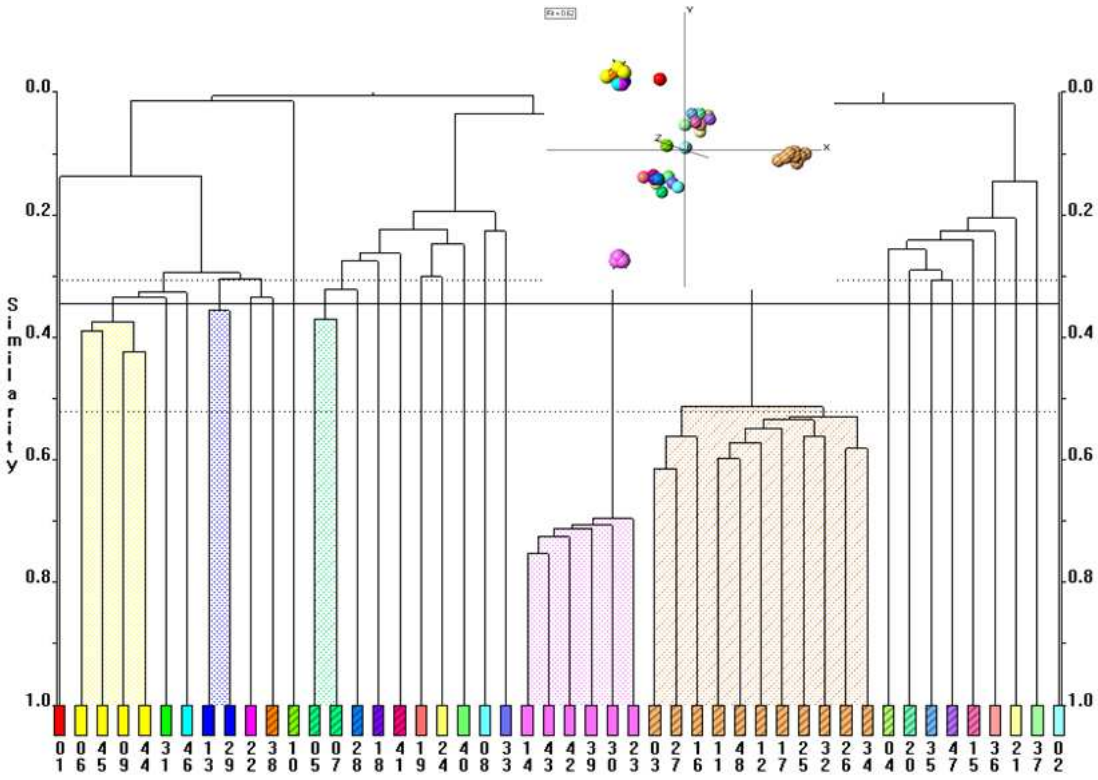

Figure 221 - Second Derivative Raman Dendrogram and MMDS Plot

This dendrogram is not as good as that previously seen. The only large clusters present are the pink cluster which contains samples 14, 23, 30, 39, 42 and 43 and the lighter striped brown cluster which contains samples 03, 11, 12, 16, 17, 25, 26, 27, 32, 34 and 48. These samples have been grouped in these respective clusters in all datasets previously examined. No information can be gleaned about the remainder of the datasets clustering due to the large number of separate clusters. It is not possible to combine these clusters while still remaining within the upper and lower cluster estimate.

8.2.4 COMBINED DATASET

The INDSCAL dendrogram, combining all three X-ray datasets and the Raman dataset, is shown, along with the corresponding MMDS plot, in Figure 218.

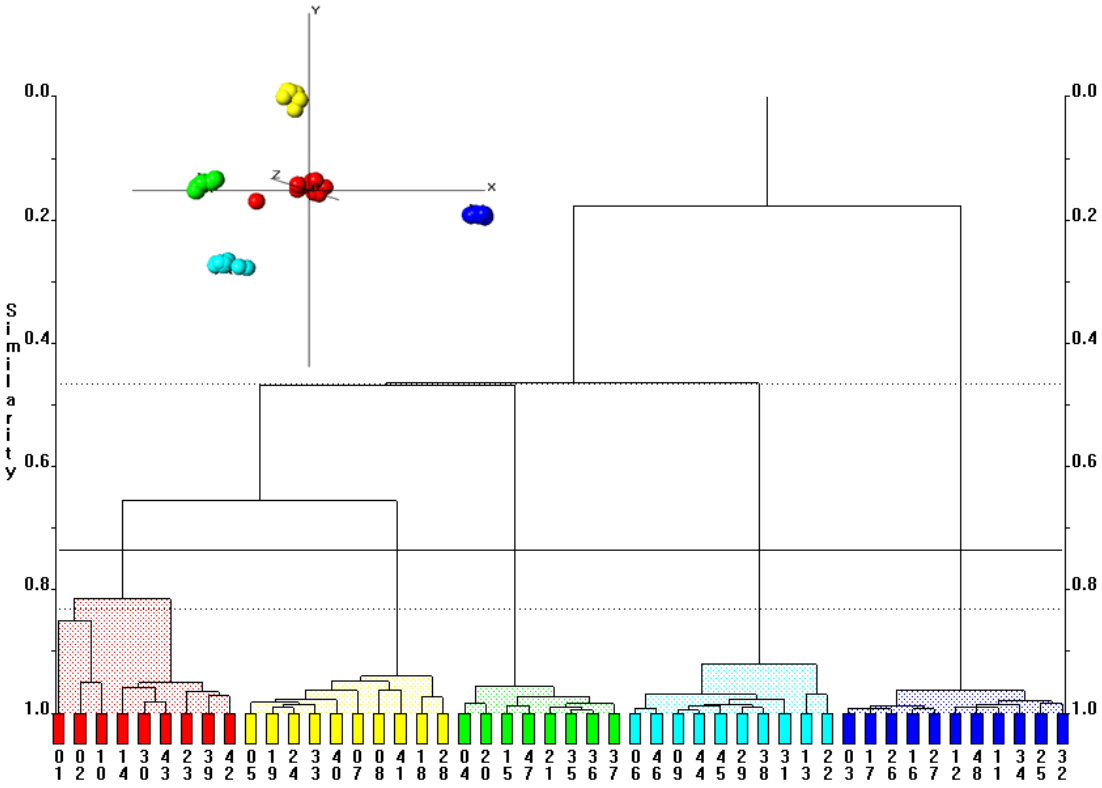

Figure 222 – INDSCAL Combined Dendrogram and MMDS Plot

The red cluster contains samples 01, 02, 10, 14, 23, 30, 39, 42 and 43. This is a mixture of three of the previously occurring clusters. The yellow cluster contains samples 05, 07, 08, 18, 19, 24, 28, 33, 40 and 41 which have previously been clustered together. The green cluster contains samples 04, 15, 20, 21, 35, 36, 37 and 47 which have previously been clustered together. The aquamarine cluster contains samples 06, 09, 13, 22, 29, 31, 38, 44, 45 which have previously been clustered together and the blue cluster contains samples 03, 11, 12, 16, 17, 25, 26, 27, 32, 34 and 48 which have previously been clustered together. The three combined clusters in the red cluster can be separated by lowering the cut-level as shown in Figure 219.

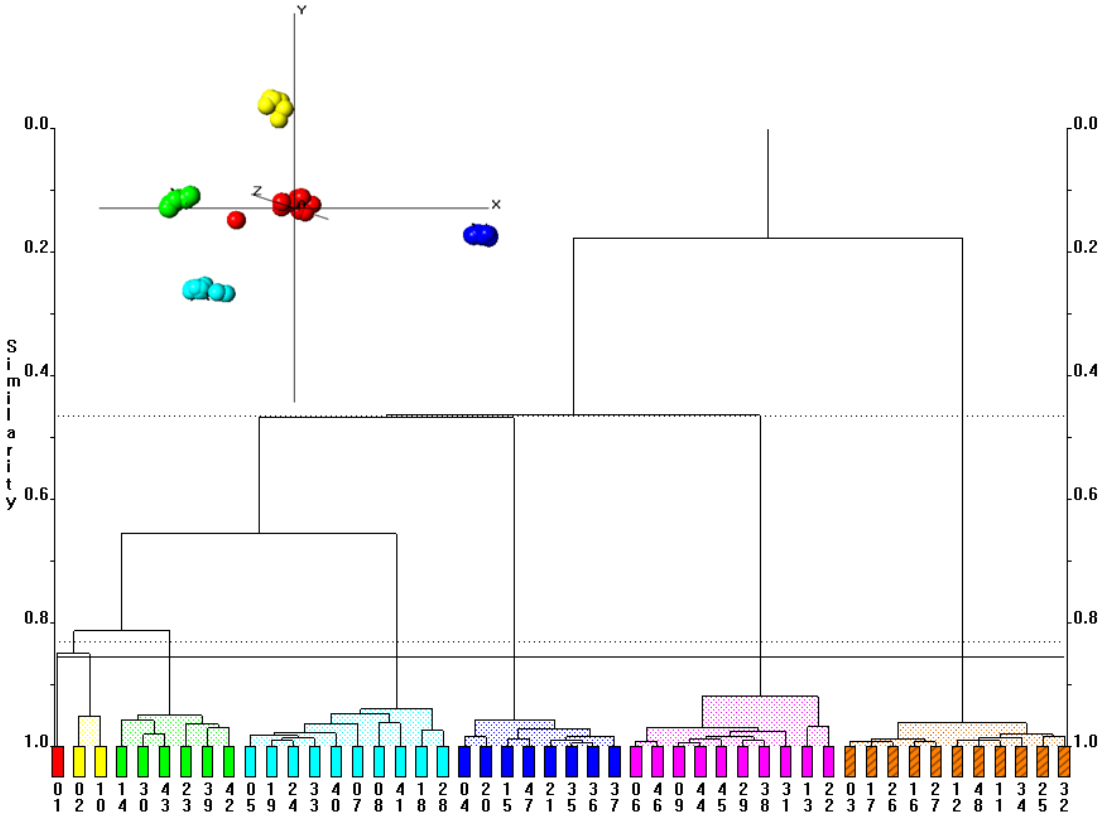

Figure 223 – INDSCAL Combined Dendrogram and MMDS Plot with Lowered Cut-level

With the new red cluster containing sample 01 and the new yellow cluster containing sample 02 and 10, the clustering exactly matches that shown in the initial X-ray runs.

8.3 DATASET COMPOSITION

It was reported to Professor Frampton that samples 14, 23, 30, 39, 42 and 43 are all of the same material, that samples 04, 15, 20, 21, 35, 26, 37 and 47 cluster together and are most probably of the same material, that samples 05, 07, 08, 18, 19, 24, 28, 33, 40 and 41 cluster together and are likely to be the same material, that samples 03, 11, 12, 16, 17, 25, 26, 27, 32, 34 and 48 cluster together and are off the same material and that samples 02 and 10 cluster together and are off the same material. It was also reported that samples 06, 09, 13, 22, 29, 31, 38, 44, 45 and 46 always cluster together and so are off the same materials and that, as sample 01 is sometimes clustered with them and sometimes without, that it is likely to be a different polymorph of the materials in this group.

He replied with the list of what materials were actually present in each well in the well plate. This is summarised in Table 41.

Sample Number	Sample Name	Actual Sample Name	Sample Number	Sample Name	Actual Sample Name
1	CSF-135-20-1-A1	Ketoprofen_2	25	CSF-135-20-2-B1	Allopurinol
2	CSF-135-20-1-A2	Ibuprofen	26	CSF-135-20-2-B2	Allopurinol
3	CSF-135-20-1-A3	Allopurinol	27	CSF-135-20-2-B3	Allopurinol
4	CSF-135-20-1-A4	Flurbiprofen	28	CSF-135-20-2-B4	Acetominophen
5	CSF-135-20-1-A5	Acetominophen	29	CSF-135-20-2-B5	Ketoprofen
6	CSF-135-20-1-A6	Ketoprofen	30	CSF-135-20-2-B6	Piroxicam
7	CSF-135-20-1-A7	Acetominophen	31	CSF-135-20-2-B7	Ketoprofen
8	CSF-135-20-1-A8	Acetominophen	32	CSF-135-20-2-B8	Allopurinol
9	CSF-135-20-1-B1	Ketoprofen	33	CSF-135-20-3-A1	Acetominophen
10	CSF-135-20-1-B2	Ibuprofen	34	CSF-135-20-3-A2	Allopurinol
11	CSF-135-20-1-B3	Allopurinol	35	CSF-135-20-3-A3	Flurbiprofen
12	CSF-135-20-1-B4	Allopurinol	36	CSF-135-20-3-A4	Flurbiprofen
13	CSF-135-20-1-B5	Ketoprofen	37	CSF-135-20-3-A5	Flurbiprofen
14	CSF-135-20-1-B6	Piroxicam	38	CSF-135-20-3-A6	Ketoprofen
15	CSF-135-20-1-B7	Flurbiprofen	39	CSF-135-20-3-A7	Piroxicam
16	CSF-135-20-1-B8	Allopurinol	40	CSF-135-20-3-A8	Acetominophen
17	CSF-135-20-2-A1	Allopurinol	41	CSF-135-20-3-B1	Acetominophen
18	CSF-135-20-2-A2	Acetominophen	42	CSF-135-20-3-B2	Piroxicam
19	CSF-135-20-2-A3	Acetominophen	43	CSF-135-20-3-B3	Piroxicam
20	CSF-135-20-2-A4	Flurbiprofen	44	CSF-135-20-3-B4	Ketoprofen
21	CSF-135-20-2-A5	Flurbiprofen	45	CSF-135-20-3-B5	Ketoprofen
22	CSF-135-20-2-A6	Ketoprofen	46	CSF-135-20-3-B6	Ketoprofen
23	CSF-135-20-2-A7	Piroxicam	47	CSF-135-20-3-B7	Flurbiprofen
24	CSF-135-20-2-A8	Acetominophen	48	CSF-135-20-3-B8	Allopurinol

Ketoprofen_2	1	1
Ketoprofen	6,9,13,22,29,31,38,44,45,46	10
Ibuprofen	2,10	2
Allopurinol	3,11,12,16,17,25,26,27,32,34,48	11
Acetominophen	5,7,8,18,19,24,28,33,40,41	10
Flurbiprofen	4,15,20,21,35,36,37,47	8
Piroxicam	14,23,30,39,42,43	6
	Total	48

Table 42 - Unseen Dataset Actual Composition

The predictions for the unseen dataset match 100% with the actual composition.

8.4 CONCLUSION

- PolySNAP has successfully analysed an unknown dataset, one which consists of both PXRD and Raman data and correctly estimated its composition. The software has successfully detected the more noticeable differences of two materials which are completely different as well as the smaller differences between two highly similar polymorphs.
- The blind test provides clear validation for the usefulness of the Raman techniques that have been included as well as for the INDSCAL methodology. The INDSCAL in particular has proven to be useful as it accurately determined the correct cluster membership.

CHAPTER 9 CONCLUSIONS AND FUTURE WORK

9.1 CONCLUSION

The effectiveness of pattern matching PXRD data has been further reinforced by datasets previously shown. When moderate or good quality PXRD data is supplied good clustering is normally produced. This is particularly noticeable with the higher range dataset in Chapter 4 and the X-ray data in Chapter 8. Poorer quality PXRD data can give good results however it is preferable to look at this sort of data alongside another datatype. This is most clearly demonstrated by the lesser quality X-ray data from Chapter 4.

Raman data can be used successfully alongside PXRD data in PolySNAP or on its own. The Raman data in Chapter 4 and in the unseen dataset in Chapter 8 show this particularly well. The pure Raman data has problems of high similarity, however this can be overcome, with a small loss of the correctness of the clustering, by taking a first or second derivative of the Raman data and matching these, as shown in Chapter 4. The higher the derivative taken, the larger the difference in similarities becomes, however this does result in larger losses of clustering correctness. For optimal Raman results, an INDSCAL combination of the original and derivative Raman data can be used. This gives clearer differences between pattern similarities without losing clustering correctness.
Good quality Raman data can be a very useful tool when used in conjunction with poorer quality PXRD data, as shown in Chapter 4.

DSC data can be used with PolySNAP however the results are not always good. The success of DSC pattern matching depends on many factors that arise in the collection of the patterns, for example mass of material analysed and heating rate used during collection. Further work would be required to perfect this technique with PolySNAP.
.
IR data can be used with other data types or on its own. This dataset has some similarity problems however these are not as high as they are in Raman data. This dataset is not currently as useful as Raman data however, with more work; it could prove to be equally as useful a tool as Raman data has been shown to be.

The INDSCAL methods themselves have proven to be a useful technique as they can allow datasets with problems such as preferred orientation to be combined with datasets that do not have these problems to produce a combination which gives the expected clustering.

This can also be used to combine two good quality datasets to check if the combination is still giving the same results.
INDSCAL can also successfully combine multiple datasets of different data types and still give produce good clustering. Currently INDSCAL works particularly well with Raman and PXRD data however could in the future be used equally well with DSC and IR data.

TGA is not a viable technique to use with PolySNAP at this time, was shown in Chapter 5. There are very few regions of significant difference between the patterns and so all patterns tend to have high similarities, even if they are from completely different materials. TGA could potentially be useful if two materials appear to be similar in the other data types but it is believed that they should not be. If they show clear differences in their melting point then a TGA pattern would pick up on this and could clearly separate them.

Four flowcharts have been produced over the course of the work in an attempt to produce a scheme that can be followed to derive the optimal clustering from any dataset. All of these flowcharts have been identical so the optimal flowchart has been determined to be as follows:

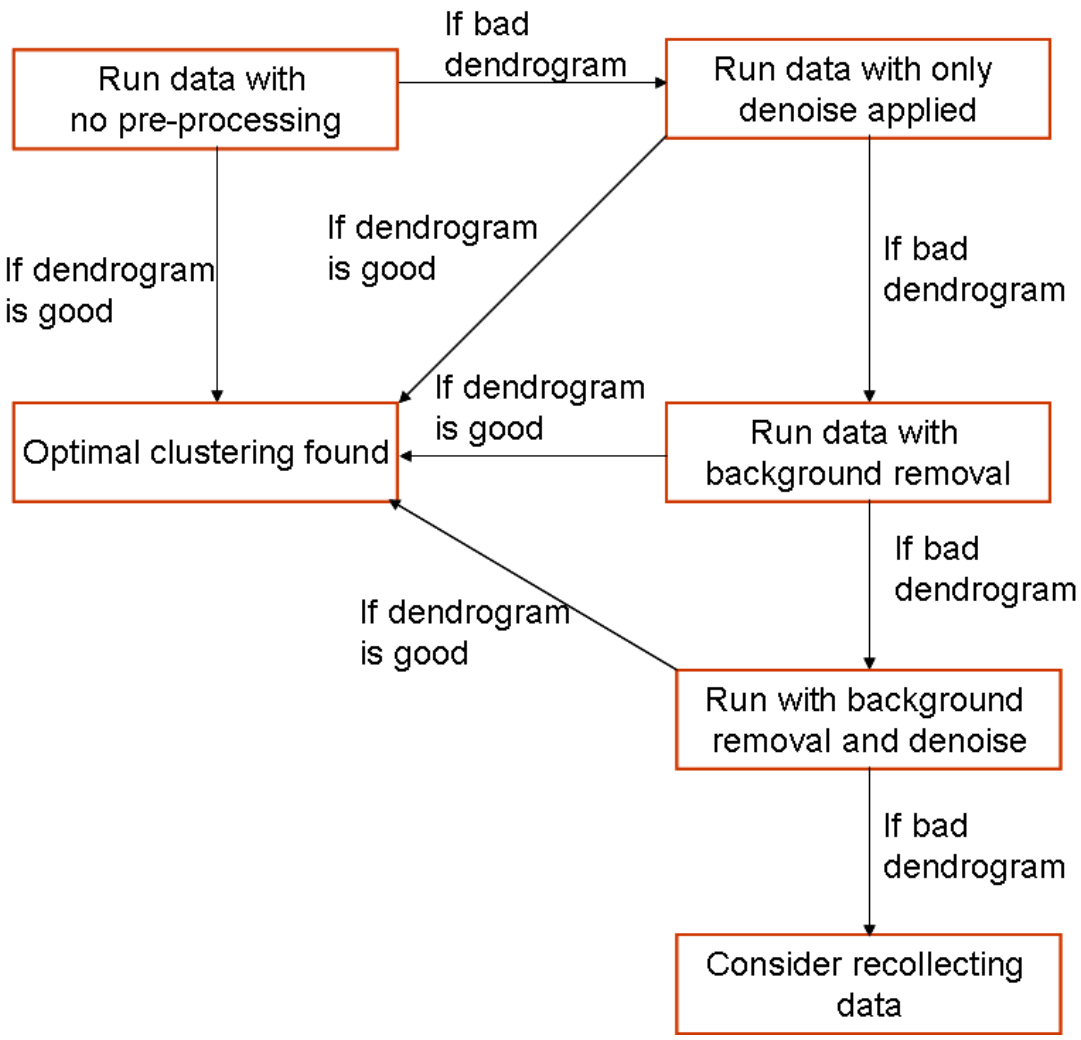

Figure 224 - Flowchart for devising optimal clustering

The recommended method is to begin with no pre-processing applied. If the dendrogram is good, showing no examples of chaining, then the optimal clustering has been found. If the dendrogram is not good then the dataset should be rerun with denoising applied. If the dendrogram is good with this result then the optimal result has been found otherwise the data should be rerun with background removal applied. If the dendrogram is still showing signs of chaining then it should be rerun again with background removal and denoising applied. If it is still showing chaining then INDSCAL combinations of the earlier methods can be attempted however it is best to recollect the dataset as there may be problems of poor quality data being present.

The methods for determining if a dendrogram is good or bad are as follows:

1. The dendrogram shows ‘chaining’ of the samples or has no clear clustering
2. The scree plot does not show the characteristic steep initial drop before smoothing out
3. The maximum and minimum confidence limits on the dendrogram have a large separation.

This finalised method was developed after the studies into the unseen dataset described in Chapter 9. It should be noted however that the optimal clustering from that dataset was obtained using a dataset with no pre-processing applied.

9.2 FUTURE WORK

The following areas could still be explored with the PolySNAP project:

- Although only Raman, IR and DSC were studied in this work, there is no reason that other solid state techniques couldn’t be used, for example solid state NMR. This is not a high-throughput technique however the PolySNAP methods should still work with the dataset.
- A quality assessment of PXRD could be carried out to try and determine what sort of quality of PXRD data is required in order to produce consistently good results.
- Further analysis can be carried out on TGA to see if there are any ways to resolve the high pattern correlation problems. Initial studies were carried out into using derivative of TGA data however this did not give any noticeable improvement to the clustering. This is an area which could be focused on further to see if any improvements can be made.
- DSC data does not give good results when compared to the other techniques. This is likely due to DSC data including much less data than PXRD, Raman and IR data. Further work can be carried out on this to attempt to improve the matching for DSC data.
- IR data can give good results on occasion however further work would be required to perfect this technique.
- Rather than carrying out an analysis of the Raman and IR spectra by hand to determine which areas are dissimilar and therefore would be best used in pattern matching, a program could be developed which could analyse each pattern in the dataset, determine what areas are showing large differences and only supply those areas to PolySNAP. Some initial work has been carried out to try and devise a method to automate this and the pseudo code is presented in appendix I.

- The methodology for determining if a dendrogram is ‘good’ or ‘bad’ could be automated to allow the software to calculate the difference between the upper and lower confidence limits in the dendrogram. If the same dataset was run four times with different pre-processing in each run, for example run 1 – no pre-processing, run 2 – denoise, run 3 – remove background and run 4 remove background and denoise, the software could calculate a score to determine which of these had the optimal gap between these limits. One possible method of calculating this score is shown in Equation 43.

$$1 - \frac{Avggap}{\max gap}$$

Equation 43 - Possible method for determining optimal Dendrogram gap

Further work would be required to refine the optimal method of calculating this score.

APPENDIX I RAMAN AND IR MATCHING PSEUDOCODE

The following pseudo code would work for a dataset containing two files. Additional arrays can be added for existing files.

OPEN FIRST FILE
STORE X-AXIS DATA INTO ARRAY 1.1
STORY Y-AXIS DATA INTO ARRAY 1.2
OPEN SECOND FILE
STORE X-AXIS DATA INTO ARRAY 2.1
STORY Y-AXIS DATA INTO ARRAY 2.2
CHECK THAT BOTH X-AXIS ARRAYS ARE ON THE SAME SCALE AND START POINT
STARTING FROM IDENTICAL START POINT ON EACH ARRAY COMPARE Y-AXIS RESULTS
IF RESULTS IDENTICAL OR HIGHLY SIMILAR – DISCARD
IF RESULTS LARGELY DISSIMILAR – ADD TO ARRAY OF SIGNIFICANT POINTS
REPEAT UNTIL FILES FULLY COMPARED

APPENDIX II DSC PRE-PROCESSING PROGRAM

The following program, written in Java, strips out the binary characters present within TA instruments DSC files and allows PolySNAP to read the files.

CLASS MAIN

```
public class Main {
    /** Creates a new instance of Main */
    public Main() {
    }
    /**
     * @param args the command line arguments
     */
    public static void main(String[] args) {
        dscframe d = new dscframe();
        d.setSize(600,300);
        d.setVisible(true);
    }
}
```

CLASS DSC FRAME

```
import java.awt.*;
import java.awt.event.*;
import javax.swing.*;
import javax.swing.filechooser.FileFilter;
import java.io.*;
import java.util.*;

public class dscframe extends JFrame implements ActionListener {
    //GUI components
    JButton load,process;
    JTextField text,num;
    JTextArea info;
    JLabel lab,sortlab, ramplab, typelab;
    JPanel l,p,te,nm,mid,bottom,upbot,lowbot,la,sla,s,rla,r,tyla,ty;
    JMenuBar mb;
```

```
JMenu file;
JMenuItem loadmenu,procmenu,exitmenu;
JComboBox sort, ramp, type;

//File read and write tools
String path;
JFileChooser ch;
FileReader f;
BufferedReader b;
FileWriter fw;
StringTokenizer st;
String comp, comp2, comp3, header, fin, fin2, fin3;
String dir = "Up";
String run = "Heat";
String inst = "Q100";
int sect =1;
int active=1;
int filename = 1;
String s1,e1,s2,e2,s3,e3;
boolean firstfound = false;
File folder,current;
File[] filelist;

//arrays for storing two lines of data for comparison
String [] row1 = new String[4];
String [] row2 = new String[4];
String [] altrow1 = new String[5];
String [] altrow2 = new String[5];

/** Creates a new instance of dscframe */
public dscframe() {
   this.setDefaultCloseOperation(EXIT_ON_CLOSE);
   this.buildFrame();  }

/**Builds the GUI for the program*/
```

```
private void buildFrame() {
    Container pane = this.getContentPane();

    //top panel contains menubar.
    mb = new JMenuBar();
    file = new JMenu("File");
    loadmenu = new JMenuItem("Load");
    loadmenu.addActionListener(this);
    procmenu = new JMenuItem("Process");
    procmenu.addActionListener(this);
    procmenu.setEnabled(false);
    exitmenu = new JMenuItem("Exit");
    exitmenu.addActionListener(this);
    file.add(loadmenu);
    file.add(procmenu);
    file.add(exitmenu);
    mb.add(file);

    //middle panel contains text area for info on file processing and text
    //field to display currently selected file.

    info = new JTextArea();
    info.setEditable(false);
    JScrollPane scroll = new JScrollPane(info);
    text = new JTextField(20);
    text.setEditable(false);
    te=new JPanel();
    te.add(text);
    mid = new JPanel();
    mid.setLayout(new BorderLayout());
    mid.add(te,BorderLayout.NORTH);
    mid.add(scroll,BorderLayout.CENTER);

    //bottom panel contains buttons and text field for column selection.
    load = new JButton("Load");
```

```
load.addActionListener(this);
l=new JPanel();
l.add(load);
process = new JButton("Process");
process.addActionListener(this);
process.setEnabled(false);
p=new JPanel();
p.add(process);
lab = new JLabel("Select Column(default is 2)");
la = new JPanel();
la.add(lab);
num = new JTextField(2);
nm = new JPanel();
nm.add(num);
typelab= new JLabel("Instrument Type");
type = new JComboBox();
type.addActionListener(this);
type.addItem("Q100");
type.addItem("Q2000");
tyla=new JPanel();
tyla.add(typelab);
ty=new JPanel();
ty.add(type);
sortlab = new JLabel("Data Direction");
sort = new JComboBox();
sort.addActionListener(this);
sort.addItem("Up");
sort.addItem("Down");
sla=new JPanel();
s=new JPanel();
sla.add(sortlab);
s.add(sort);
ramplab = new JLabel("Heat Method");
ramp = new JComboBox();
ramp.addActionListener(this);
```

```
    ramp.addItem("Heat");
    ramp.addItem("Heat-Cool");
    ramp.addItem("Heat-Cool-Heat");
    rla = new JPanel();
    rla.add(ramplab);
    r = new JPanel();
    r.add(ramp);
    upbot=new JPanel();
    upbot.setLayout(new GridLayout(1,4));
    upbot.add(l);
    upbot.add(p);
    upbot.add(la);
    upbot.add(nm);
    lowbot=new JPanel();
    lowbot.setLayout(new GridLayout(1,6));
    lowbot.add(tyla);
    lowbot.add(ty);
    lowbot.add(sla);
    lowbot.add(s);
    lowbot.add(rla);
    lowbot.add(r);
    bottom=new JPanel();
    bottom.setLayout(new GridLayout(2,1));
    bottom.add(upbot);
    bottom.add(lowbot);


    //assemble panels into overall gui.
    pane.add(mb,BorderLayout.NORTH);
    pane.add(mid,BorderLayout.CENTER);
    pane.add(bottom,BorderLayout.SOUTH);
}

/**
 * Handles button clicks. Clicking load calls the ld method. Clicking
```

```
 * process calls the process method.
 */

public void actionPerformed(ActionEvent e) {
   if(e.getSource()==load) {
      this.ld();
   }
   if(e.getSource()==process) {
      this.process();
   }
   if(e.getSource()==loadmenu) {
      this.ld();
   }
   if(e.getSource()==procmenu) {
      this.process();
   }
   if(e.getSource()==exitmenu) {
      System.exit(0);}}

/**
 *Called when the load button is clicked. Creates a FileChooser and allows
 *the user to select the file to be loaded and processed.
 */

private void ld() {
   firstfound = false;
   int sect =1;
   int active=1;
   int filename = 1;
   ch = new JFileChooser();
   ch.setFileSelectionMode(JFileChooser.DIRECTORIES_ONLY);
   int returnVal = ch.showOpenDialog(dscframe.this);
   if(returnVal == JFileChooser.APPROVE_OPTION) {
      path = ch.getSelectedFile().getName();
      text.setText(path);
```

```
        info.append("Opened "+path+"\n");
        process.setEnabled(true);
        procmenu.setEnabled(true);} }

/**
 * Called when the process button is clicked. Creates a file reader and
 * reads the specified file and stores the data into a temporary string
 * until "StartOfData" found. When this is found the following rows are
 * read two at a time and compared to each other. The value input into the
 * num text box allows the corresponding column in each of the two rows.
 * If the selected value in row 2 is higher than the value in row 1 then
 * row 2 is written to the output file. The row is dumped if this doesn not
 * hold true. If the column to be checked falls outside to the range of 1-4
 * then the vale us set to 2 by default. If no value is entered the value is
 * also set to 2.When the file has been fully read the temporary string
 * is written to a new file (old file name appended by _proc). Each line of
 * the file is read character at a time to remove any extra characters added
 * by the binary header.
 */

 private void getinput()
 {
    folder = ch.getSelectedFile();
    filelist = folder.listFiles();
    for(int i = 0;i<filelist.length;i++)
    {
        if(filelist[i].isFile()){
            current = filelist[i];
//call the data checking method
   this.typesort();
    //call the output method.
    this.output();} }}
 private void process() {
    dir = (String) sort.getSelectedItem();
    run = (String)ramp.getSelectedItem();
```

```
    inst = (String)type.getSelectedItem();
    this.getinput(); }
public void typesort()
{
    //check if data is of a type produced by a Q100
    if(inst.compareTo("Q100")==0)
    {
    comp=null;
    header=null;
    int value=2; //set column to be read to row 2 (default value)
    try {
        /*
         *check if a value has been input into the 'num' text box and if it
         *is on the correct range
         */
        try {
            value =  Integer.parseInt(num.getText());
            if(value>4) {
                value =2;}
            if(value<0) {
                value=2;  }
        } catch(NumberFormatException nfe) {}
        value--; //set column value to a corresponding array value
        //read marked file
        f = new FileReader(current);
        b = new BufferedReader(f);
    } catch(FileNotFoundException nfe) {
        info.append("Error with file\n");}
    boolean endOfFile = false;//end of file check
    boolean datafound = false;//start of data check
    boolean firstline = true;//first line processed check
    while (!endOfFile) {
        //read data from file.
        try {
            String data = b.readLine();
```

```
//check for end of file.
if (data==null) {
   endOfFile=true;
} else {
   //Following code removes extra binary info characters
   int dlen = data.length();//get length of file
   String data2 = new String();//setup new string for trimmed data
   for(int z=0;z<dlen;z++) {
      char c1 = data.charAt(z);//read string character at a time
      if((int)c1>0) {//check if character is above 0 in ASCII table
         data2+= c1;//if so add to list}}
   if(data2.length()>1) {//check if new string has length greater than 1
      data = data2;//if so allow to be output
   } else//otherwise repeat for new line of program
      {data = b.readLine();//get line
      dlen = data.length();//check length
      data2 = new String();//wipe data2
      for(int z=0;z<dlen;z++) {
         char c1 = data.charAt(z);//read amd check char
         if((int)c1>0) {//check if character is above 0 in ASCII table
            data2+= c1;//output char}}
      data = data2;//load trimmed line to data.
   }
   //if start of data not found then add straight to output file.
   if(datafound==false) {
      header=header+"\n"+data;}
   //check if current line is start of data.
   if(data.compareTo("StartOfData")==0) {
      System.out.println("found start");
      comp=header;
      fin=null;
      datafound=true;
   }if(datafound==true) {
      //check if firstline of data has been processed yet.
      if(firstline==true) {
```

```
firstline=false;
data = b.readLine();
//Following code removes extra binary info characters
dlen = data.length();//get length of file
data2 = new String();//setup new string for trimmed data
for(int z=0;z<dlen;z++) {
   char c1 = data.charAt(z);//read string character at a time
   if((int)c1>0) {//check if character is above 0 in ASCII table
      data2+= c1;//if so add to list} }

if(data2.length()>1) {//check if new string has length greater than 1
   data = data2;//if so allow to be output
} else//otherwise repeat for new line of program
{data = b.readLine();//get line
   dlen = data.length();//check length
   data2 = new String();//wipe data2
   for(int z=0;z<dlen;z++) {
      char c1 = data.charAt(z);//read amd check char
      if((int)c1>0) {//check if character is above 0 in ASCII table
         data2+= c1;//output char}}
   data = data2;//load trimmed line to data.
   s1=data;}
st=new StringTokenizer(data);
int arraynum=0;
while (st.hasMoreTokens()) {
   row1[arraynum] = st.nextToken();
   arraynum++;}
//read a second line of data to match against line 1
//saves a second loop through the cycle after the
//first line is loaded.
data = b.readLine();
//Following code removes extra binary info characters
dlen = data.length();//get length of file
data2 = new String();//setup new string for trimmed data
for(int z=0;z<dlen;z++) {
```

```
        char c1 = data.charAt(z);//read string character at a time
        if((int)c1>0) {//check if character is above 0 in ASCII table
            data2+= c1;//if so add to list} }
    if(data2.length()>1) {//check if new string has length greater than 1
        data = data2;//if so allow to be output
    } else//otherwise repeat for new line of program
    {
        data = b.readLine();//get line
        dlen = data.length();//check length
        data2 = new String();//wipe data2
        for(int z=0;z<dlen;z++) {
            char c1 = data.charAt(z);//read amd check char
            if((int)c1>0) {//check if character is above 0 in ASCII table

                data2+= c1;//output char}}
        data = data2;//load trimmed line to data.
    }}
st=new StringTokenizer(data);
int arraynum=0;

while (st.hasMoreTokens()) {

    String val = st.nextToken();
    if(val.compareTo("-2.00000")==0) {
        System.out.println("here");
        String val2 = st.nextToken();
        System.out.println("2nd val "+val2);
        String val3 = st.nextToken();
        System.out.println("3rd val "+val3);

        if(val3.compareTo("0.000000")==0) {
            String val4 = st.nextToken();
            System.out.println("4th val "+val4);
            if(val4.compareTo("0.000000")==0) {
                System.out.println(active);
```

```
                    if(active==1) {
                       comp2=header;
                       fin2=null;}
                    if(active==2) {
                       comp3=header;
                       fin3=null;}
                    active++;
                    sect++;}}
                } else {
                    row2[arraynum] = val;
                    arraynum++;  }}
               if(active==1)
               {e1=data;}
               if(active==2)
               {e2=data;}
               if(active==3)
               {e3=data;}
               /*
                *Check if the data type is set to heat, heat-cool or
                * heat-cool-heat and choose the appropriate method
                */
               /*
                *if heat - check if data is runnning in the up or down
                *direction (input tab) and then check the data is
                *running in the correct direction.
                */
               if(run.compareTo("Heat")==0)//heat run
               {  if(dir.compareTo("Up")==0) {
   if(Double.parseDouble(row1[value])<Double.parseDouble(row2[value]))
                       comp=comp+"\n"+data;}
                   if(dir.compareTo("Down")==0) {
if(Double.parseDouble(row1[value])>Double.parseDouble(row2[value]))
                       comp=comp+"\n"+data;
                   }System.arraycopy(row2,0,row1,0,4);}
```

```
                    /*
                     *if heat-cool - check data is running in the upward
                     *direction until the marker is encountered.
                     *When this encountered change to downward direction.
                     *Output both as 2 seperate files.
                     */
                    if(run.compareTo("Heat-Cool")==0) {//heat-cool run
                       if(active == 1) {
if(Double.parseDouble(row1[value])<Double.parseDouble(row2[value])) {
                           comp=comp+"\n"+data;
                            if(firstfound==false)
                          {  s1 = data;
                           firstfound = true;
                          }}}if(active == 2) {

if(Double.parseDouble(row1[value])>Double.parseDouble(row2[value])) {
                           comp2=comp2+"\n"+data;
                           if(firstfound==false)
                          {s1 = data;
                           firstfound = true;  }}  }
                       System.arraycopy(row2,0,row1,0,4);}

                    /*
                     *if heat-cool-heat - check data is running in the upward
                     *direction until the first marker is encountered.
                     *When this encountered change to downward direction
                     *until the second marker is reached. When this occurs
                     *switch back to upwards direction.
                     *Output all 3 seperate files.
                     */

                    if(run.compareTo("Heat-Cool-Heat")==0) {//heat-cool-heat run
                       if(active == 1) {
  if(Double.parseDouble(row1[value])<Double.parseDouble(row2[value])) {
```

```
                    comp=comp+"\n"+data;  }}
                if(active == 2) {
if(Double.parseDouble(row1[value])>Double.parseDouble(row2[value])) {
                    comp2=comp2+"\n"+data;}}
                if(active ==3) {
if(Double.parseDouble(row1[value])<Double.parseDouble(row2[value])) {
                    comp3=comp3+"\n"+data;  }}
                System.arraycopy(row2,0,row1,0,4);
              }}}} catch(IOException ioe) {}} }

    //check if data is of a type produced by a Q2000
    else if(inst.compareTo("Q2000")==0)
    {
    comp=null;
    header=null;
    int value=2; //set column to be read to row 2 (default value)
    try {
       /*
        *check if a value has been input into the 'num' text box and if it
        *is on the correct range
        */
       try {
          value =  Integer.parseInt(num.getText());
          if(value>5) {
             value =2;
          }
          if(value<0) {
             value=2;
          }
       } catch(NumberFormatException nfe) {}
       value--; //set column value to a corresponding array value
       //read marked file
       f = new FileReader(ch.getSelectedFile());
       b = new BufferedReader(f);
    } catch(FileNotFoundException nfe) {
```

```
   info.append("Error with file\n");
}
boolean endOfFile = false;//end of file check
boolean datafound = false;//start of data check
boolean firstline = true;//first line processed check
while (!endOfFile) {
   //read data from file.
   try {
      String data = b.readLine();
      //check for end of file.
      if (data==null) {
         endOfFile=true;
      } else {
         //Following code removes extra binary info characters
         int dlen = data.length();//get length of file
         String data2 = new String();//setup new string for trimmed data
         for(int z=0;z<dlen;z++) {
            char c1 = data.charAt(z);//read string character at a time
            if((int)c1>0) {//check if character is above 0 in ASCII table
               data2+= c1;//if so add to list}}
         if(data2.length()>1) {//check if new string has length greater than 1
            data = data2;//if so allow to be output
         } else//otherwise repeat for new line of program
         {data = b.readLine();//get line
            dlen = data.length();//check length
            data2 = new String();//wipe data2
            for(int z=0;z<dlen;z++) {
               char c1 = data.charAt(z);//read amd check char
               if((int)c1>0) {//check if character is above 0 in ASCII table
                  data2+= c1;//output char } }
            data = data2;//load trimmed line to data.}
         //if start of data not found then add straight to output file.
         if(datafound==false) {
            header=header+"\n"+data;}
         //check if current line is start of data.
```

```
if(data.compareTo("StartOfData")==0) {
   System.out.println("found start");
   comp=header;
   fin=null;
   datafound=true;}
if(datafound==true) {
   //check if firstline of data has been processed yet.
   if(firstline==true) {
      firstline=false;
      data = b.readLine();
      //Following code removes extra binary info characters
      dlen = data.length();//get length of file
      data2 = new String();//setup new string for trimmed data
      for(int z=0;z<dlen;z++) {
         char c1 = data.charAt(z);//read string character at a time
         if((int)c1>0) {//check if character is above 0 in ASCII table
            data2+= c1;//if so add to list
         } }
      if(data2.length()>1) {//check if new string has length greater than 1
         data = data2;//if so allow to be output
      } else//otherwise repeat for new line of program
      {
         data = b.readLine();//get line
         dlen = data.length();//check length
         data2 = new String();//wipe data2
         for(int z=0;z<dlen;z++) {
            char c1 = data.charAt(z);//read amd check char
            if((int)c1>0) {//check if character is above 0 in ASCII table
               data2+= c1;//output char
            }}
         data = data2;//load trimmed line to data.
         s1=data;}
      st=new StringTokenizer(data);
      int arraynum=0;
      while (st.hasMoreTokens()) {
```

```
                altrow1[arraynum] = st.nextToken();
                arraynum++;  }
            //read a second line of data to match against line 1
            //saves a second loop through the cycle after the
            //first line is loaded.
            data = b.readLine();
            //Following code removes extra binary info characters
            dlen = data.length();//get length of file
            data2 = new String();//setup new string for trimmed data
            for(int z=0;z<dlen;z++) {
                char c1 = data.charAt(z);//read string character at a time
                if((int)c1>0) {//check if character is above 0 in ASCII table
                    data2+= c1;//if so add to list} }
if(data2.length()>1) {//check if new string has length greater than 1
                data = data2;//if so allow to be output
            } else//otherwise repeat for new line of program
            {
                data = b.readLine();//get line
                dlen = data.length();//check length
                data2 = new String();//wipe data2
                for(int z=0;z<dlen;z++) {
                    char c1 = data.charAt(z);//read amd check char
                    if((int)c1>0) {//check if character is above 0 in ASCII table
                        data2+= c1;//output char}}
                data = data2;//load trimmed line to data.}}
            st=new StringTokenizer(data);
            int arraynum=0;
            while (st.hasMoreTokens()) {
                String val = st.nextToken();
                if(val.compareTo("-2.00000")==0) {
                    System.out.println("here");
                    String val2 = st.nextToken();
                    System.out.println("2nd val "+val2);
                    String val3 = st.nextToken();
                    System.out.println("3rd val "+val3);
```

```
        if(val3.compareTo("0.000000")==0) {
            String val4 = st.nextToken();
            System.out.println("4th val "+val4);
            if(val4.compareTo("0.000000")==0) {
                String val5 = st.nextToken();
                if(val5.compareTo("0.000000")==0)
                {
                System.out.println(active);
                if(active==1) {
                    comp2=header;
                    fin2=null;}
                if(active==2) {
                    comp3=header;
                    fin3=null;}
                active++;
                sect++;}}}} else {
        altrow2[arraynum] = val;
        arraynum++;}}
if(active==1)
{e1=data;}
if(active==2)
{e2=data;  }
if(active==3)
{e3=data;}
/*
 *Check if the data type is set to heat, heat-cool or
 * heat-cool-heat and choose the appropriate method
 */
/*
 *if heat - check if data is runnning in the up or down
 *direction (input tab) and then check the data is
 *running in the correct direction.
 */
if(run.compareTo("Heat")==0)//heat run
{if(dir.compareTo("Up")==0) {
```

```
if(Double.parseDouble(altrow1[value])<Double.parseDouble(altrow2[value]))
                    comp=comp+"\n"+data;
                }if(dir.compareTo("Down")==0) {
 if(Double.parseDouble(altrow1[value])>Double.parseDouble(altrow2[value]))
                    comp=comp+"\n"+data;
                }  System.arraycopy(altrow2,0,altrow1,0,4);   }
              /*
               *if heat-cool - check data is running in the upward
               *direction until the marker is encountered.
               *When this encountered change to downward direction.
               *Output both as 2 seperate files.
               */
              if(run.compareTo("Heat-Cool")==0) {//heat-cool run
                 if(active == 1) {
          if(Double.parseDouble(altrow1[value])<Double.parseDouble(altrow2[value])) {
                    comp=comp+"\n"+data;
                     if(firstfound==false)
                   {
                    s1 = data;
                    firstfound = true;
                   }}}if(active == 2) {

if(Double.parseDouble(altrow1[value])>Double.parseDouble(altrow2[value])) {
                    comp2=comp2+"\n"+data;
                    if(firstfound==false)
                   {  s1 = data;
                    firstfound = true;
                   }}}  System.arraycopy(row2,0,row1,0,4);}

              /*
               *if heat-cool-heat - check data is running in the upward
               *direction until the first marker is encountered.
               *When this encountered change to downward direction
               *until the second marker is reached. When this occurs
               *switch back to upwards direction.
```

```
                    *Output all 3 seperate files.
                    */
                   if(run.compareTo("Heat-Cool-Heat")==0) {//heat-cool-heat run
                      if(active == 1) {
if(Double.parseDouble(altrow1[value])<Double.parseDouble(altrow2[value])) {
                          comp=comp+"\n"+data;
                        }}
                      if(active == 2) {
if(Double.parseDouble(altrow1[value])>Double.parseDouble(altrow2[value])) {
                          comp2=comp2+"\n"+data;
                        }}if(active ==3) {

if(Double.parseDouble(altrow1[value])<Double.parseDouble(altrow2[value])) {
                          comp3=comp3+"\n"+data;
                        }}System.arraycopy(altrow2,0,altrow1,0,4);}}}
        } catch(IOException ioe) {
        } } }}

   /**
    *This method, called by the process method, outputs the data after the file has been checked.
    *Three different output methods are possible depending on the Method (Sort combo box) selected.
    */
   private void output()
   {
      try {
         filename=1;
         sect=1;
         if(active>3)
         {active=3;}
         String name = current.getPath();
         int namel = name.length();
         String namen = name.substring(namel-namel,namel-4);
         info.append("Number of sections = "+sect+"\n");
```

```
if(active==1) {
   name = namen+"_proc_"+filename+".txt";
   info.append("Processing completed.\nSaved as "+name+"\n");
   fw = new FileWriter(name);
   fw.write(comp);
   fw.close();
}
if(active == 2) {
   name = namen+"_proc_"+filename+".txt";
   info.append("Processing of part 1 completed.\nSaved as "+name+"\n");
   fw = new FileWriter(name);
   fw.write(comp);
   fw.close();
   filename++;
   name = namen+"_proc_"+filename+".txt";
   info.append("Processing of part 2 completed.\nSaved as "+name+"\n");
   fw = new FileWriter(name);
   fw.write(comp2);
   fw.close();
}
if(active == 3) {
   name = namen+"_proc_"+filename+".txt";
   info.append("Processing of part 1 completed.\nSaved as "+name+"\n");
   fw = new FileWriter(name);
   fw.write(comp);
   fw.close();
   filename++;
   name = namen+"_proc_"+filename+".txt";
   info.append("Processing of part 2 completed.\nSaved as "+name+"\n");
   fw = new FileWriter(name);
   fw.write(comp2);
   fw.close();
   filename++;
   name = namen+"_proc_"+filename+".txt";
   info.append("Processing of part 3 completed.\nSaved as "+name+"\n");
```

```
        fw = new FileWriter(name);
        fw.write(comp3);
        fw.close();
    }
} catch(IOException ioe) {
}}}
```