BRUKER

DIFFRAC.SUITE

● User Manual

DIFFRAC.EVALUATION PACKAGE
DIFFRAC.EVA Cluster Analysis

Original Instructions

XRD

Innovation with Integrity

We have checked the contents of this manual for agreement with the hardware and software described. Since deviations cannot be precluded entirely, we cannot guarantee full agreement. However, the data in this manual are reviewed regularly and any necessary corrections are included in subsequent editions. Suggestions for improvement are welcome.

All configurations and specifications are subject to change without notice.

Order no. DOC-M88-EXX205 V1 Dez 03, 2014.

Printed in the Federal Republic of Germany.

Bruker AXS GmbH
Östliche Rheinbrückenstr. 49
76187 Karlsruhe, Germany
Tel.   +49 721 50997-0
Fax +49 721 50997-5654
info@bruker-axs.de
www.bruker.com

Table of Contents

1 Preamble 9

2 Quick Start Guide 11

2.1 User Interface Elements 11
2.2 Preparing for Cluster Analysis 11
2.2.1 Multiple Data Sets 12
2.2.2 Data Loading 13
2.2.3 Data Types and Data Formats 13
2.3 Performing Cluster Analysis 14
2.3.1 Scan Pre-Processing Options 15
2.3.2 Start the Analysis 16
2.3.3 Evaluating the Results 17
2.3.4 Repeating the Cluster Analysis 18
2.3.5 Publish the Results 19

3 Clustering and Pre-Processing Options 21

3.1 Run Cluster Analysis Dialog 21
3.1.1 Per dataset options 21
3.1.2 Common Options 22
3.1.3 Label extracting options 22
3.2 Scan Pre-Processing Options Window 23

4 Results Display 27

4.1 Cluster Analysis Node Properties 27
4.2 General View Features 28
4.2.1 Single and Multiple Selection 28
4.2.2 Zooming and Panning 30
4.2.3 Mouse and Keyboard Actions in 3D Views 30
4.2.4 Graphical Context Menu 30
4.2.5 View Properties 30
4.3 Graphical Views 31
4.3.1 Cell Display View 31
4.3.2 Dendrogram View 35
4.3.3 3D MMDS View 40
4.3.4 3D PCA View 45
4.3.5 6D Plot View 46
4.4 Validation Views 50
4.4.1 Scree Plot View 50
4.4.2 Minimum Spanning Tree (MST) View 51
4.4.3 Silhouettes View 55
4.4.4 Fuzzy Clustering View 58
4.4.5 Parallel Coordinates Plot View 60
4.4.6 Space Explorer View 62
4.5 Text Output Views 64
4.5.1 Numerical Results View 64
4.5.2 Log File View 67
4.5.3 Report Writer 68

4.6 References 70

5 Pre-Screening 73

5.1 Introduction 73

5.2 The Pre-Screening Process 73

6 Quality Control 77

6.1 Introduction 77

6.2 Example 77

7 Tutorial 81

7.1 Analysis of a Single Dataset 81

7.1.1 Import the Data to Be Analyzed 81

7.1.2 Running Cluster Analysis 82

7.1.3 Examine the Analysis Results 83

7.1.4 Reporting 91

7.1.5 Cluster Analysis Using Reference Samples 92

7.2 Amorphous Content and Pattern Shifts 95

7.2.1 Import Data and Run the Cluster Analysis 95

7.2.2 Analysis of the Results 96

7.2.3 Reprocessing the Data Allowing for an X-Shift 98

7.2.4 Remove Amorphous Samples from the Analysis 101

7.3 6D Plot 102

Index 111

List of figures

Figure 2.1: Exploring the screen ... 11

Figure 2.2: Data loaded into data set 1 ... 12

Figure 2.3: The Cluster Analysis command in the context menu ... 14

Figure 2.4: The Run Cluster Analysis dialog ... 15

Figure 2.5: The Scan Pre-Processing Options window ... 15

Figure 2.6: A typical dendrogram view containing also a silhouette and a scan graphics ... 17

Figure 2.7: The Clean Results command ... 18

Figure 2.8: The Report Writer view as an example for a text output view ... 19

Figure 3.1: The per dataset options for four datasets ... 21

Figure 3.2: The per dataset options for four datasets with manually changed weights ... 21

Figure 3.3: The Common and Label extracting options ... 22

Figure 3.4: The Scan Pre-Processing window displaying the graphical scan overview with a masked region and the property grid ... 23

Figure 4.1: Example of Result views in tabs ... 27

Figure 4.2: The Create context menu for the Cluster Analysis node ... 27

Figure 4.3: 3D PCA view with a single selection for sample FORM D1 ... 29

Figure 4.4: Multiple selection in a Cell Display view ... 29

Figure 4.5: Cell Display view in the default pie chart mode ... 32

Figure 4.6: Cell Display view in stacked mode ... 32

Figure 4.7: Cell Display with selected cluster FORMA, known references present ... 33

Figure 4.8: A cell which represents a mixture, tooltip with percentage displayed ... 33

Figure 4.9: Cell Display when Hide results similar to references has been selected ... 34

Figure 4.10: Cell Display with selected cluster FORMA, no known references present ... 34

Figure 4.11: Dendrogram view displaying the dendrogram control, cluster list, silhouette and scan control ... 36

Figure 4.12: Part of the dendrogram with changed cut-level ... 37

Figure 4.13: The Move Cut-Line command marked with an activity check ... 38

Figure 4.14: Dendrogram after manually moving the cut-line to 0.4 ... 38

Figure 4.15: Selecting the default Clustering method in the Cluster Analysis settings ... 39

Figure 4.16: 3D MMDS view ... 41

Figure 4.17: Default coordinate system orientation for 3D views ... 42

Figure 4.18: Changed coordinate system orientation in 3D views ... 42

Figure 4.19: Most representative member (MRM) of a cluster ... 43

Figure 4.20: Least representative member (LRM) of a cluster ... 43

Figure 4.21: 3D PCA view ... 45

Figure 4.22: 6D Plot View displaying different sample sizes and shapes for reaction time and solvent ... 47

Figure 4.23: Scree Plot view ... 50

Figure 4.24: Minimum Spanning Tree (MST) view ... 52

Figure 4.25: Minimum Spanning Tree after clicking the Link Decrement command ... 53

Figure 4.26: Silhouettes view ... 56

Figure 4.27: Displaying the histogram membership in a tooltip and referencing it to the Dendrogram view 57
Figure 4.28: Fuzzy Clustering view 59
Figure 4.29: Parallel Coordinates Plot view 61
Figure 4.30: Space Explorer view 62
Figure 4.31: Numerical results view 65
Figure 4.32: Numerical Results view with a selection below the diagonal 66
Figure 4.33: The Log File view 67
Figure 4.34: The Report Writer view 69
Figure 4.35: Report Writer View settings 70
Figure 5.1: Pre-Screening dialog before starting the analysis 74
Figure 5.2: The Matching options in the Settings dialog 74
Figure 5.3: Pre-Screening dialog during analysis with matching progress displayed 75
Figure 5.4: Pre-Screening result with a part of the result grid marked for clipboard copy 76
Figure 6.1: Figure 52: Run Cluster Analysis dialog in quality control mode 78
Figure 6.2: 3D MMDS view in Quality Control mode with reference envelope drawn 79
Figure 6.3: 3D MMDS view in Quality Control mode with reference samples drawn 80

List of tables

Table 2.1: Number of individual result sets depending on the number of datasets 12
Table 2.2: Data formats which can be used for the cluster analysis........ 13
Table 2.3: The parameters which are read from the first line of a XY or XYE file 14
Table 4.1: Cluster Analysis node properties 27
Table 4.2: Generally available mouse and keyboard commands in 3D views........ 30
Table 4.3: Command in the graphical context menu 30
Table 4.4: General view properties for all views 30
Table 4.5: Generally available mouse and keyboard commands in the Cell Display 35
Table 4.6: Cell Display view properties........ 35
Table 4.7: Generally available mouse and keyboard commands in the Dendrogram 39
Table 4.8: Dendrogram view properties........ 40
Table 4.9: 3D MMDS view properties 43
Table 4.10: Mouse and keyboard commands in the MMDS view........ 44
Table 4.11: Mouse and keyboard commands........ 48
Table 4.12: 6D Plot View properties 48
Table 4.13: Scree Plot view properties 51
Table 4.14: Mouse and keyboard commands........ 53
Table 4.15: Minimum Spanning Tree View Properties 54
Table 4.16: Mouse and keyboard commands........ 57
Table 4.17: Silhouettes View Properties........ 57
Table 4.18: Fuzzy Clustering view properties........ 60
Table 4.19: Parallel Coordinates Plot view properties 61
Table 4.20: Mouse and keyboard commands........ 62
Table 4.21: Space Explorer View properties 63
Table 4.22: Mouse and keyboard commands........ 64
Table 4.23: Numerical Results view properties 66
Table 4.24: Description of Toolbar Buttons 67
Table 4.25: Log File view properties........ 68
Table 4.26: Description of Toolbar Buttons 69
Table 5.1: Pre-screening options........ 76
Table 6.1: Special Quality Control mode properties 79

1 Preamble

DIFFRAC.EVA Cluster Analysis is a part of the DIFFRAC.EVA software package designed to cluster patterns using full profile X-ray data. Cluster analysis partitions the data into sets based on their similarity – patterns that are very similar will end up in the same cluster whilst dissimilar patterns will be in different ones. This allows the user to simplify collections of patterns and reduce the effort involved in data analysis. The use of the full profile allows for more flexible and accurate identification of samples, even when data quality is low or preferred orientation effects are significant. Cluster analysis has an extensive literature; we can recommend: *’Data Clustering: Theory, Algorithms and Applications’ by Gan, Ma and Wu published by SIAM, 2007*; *‘Cluster Analysis’ by Everitt, Landau and Leese, published by Arnold, 2001* and *‘Modern Multivariate Statistical Techniques’ by Izenman published by Springer, 2008* as useful texts.

The software provides an easy to use interface to several powerful and novel statistical methods to rank samples in order of their similarity to any other selected sample, allowing unknowns to be quickly identified. In quantitative mode, given a mixture pattern and potential pure phase patterns, it can identify which patterns are in the mixture, and quantify their proportions quickly and easily using a non-Rietveld based approach.

The matching procedure can be automated for computer-controlled high-throughput analysis. A nearly unlimited number of patterns (> 10000) can be prescreened, and the clustering algorithm allows for datasets of up to 2000 patterns with four different data sets to then be analyzed in a single run. The cluster analysis provides highly flexible graphical output to summarize and visualize the results. This highlights any unusual data, and means that time is not wasted looking at the many patterns that behave exactly as expected. It can work with or without the provision of reference patterns.

The software is an adaption of the POLYSNAP software developed at Glasgow University with significant upgrading including the use of multiple CPU cores. References are given at the end.

DIFFRAC.EVA Cluster Analysis is an integral part of DIFFRAC.EVA starting with version 4.0.

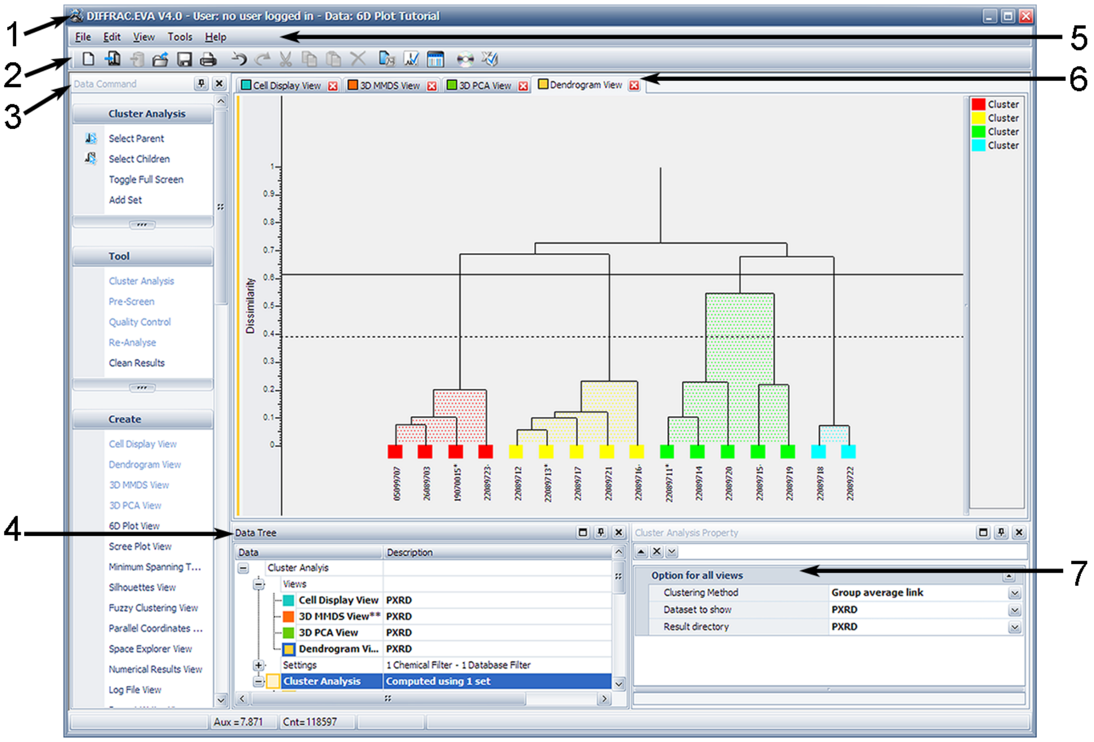
Only the cluster analysis module will be described in the following chapters.

The term **pattern** is used differently in cluster analysis compared to the X-ray phase analysis in DIFFRAC.EVA. The **pattern** describes the input data in general for the cluster analysis. However, the p**attern** in the phase analysis describes the specific d-I lists which characterize a crystallographic phase.

2 Quick Start Guide

2.1 User Interface Elements

After DIFFRAC.EVA has been started and the cluster analysis is available according to the license level and a cluster analysis has been performed, the main window will appear as follows:



Data	Description
Cluster Analyis	
Views	
Cell Display View	PXRD
3D MMDS View**	PXRD
3D PCA View	PXRD
Dendrogram Vi...	PXRD
Settings	1 Chemical Filter - 1 Database Filter
Cluster Analysis	Computed using 1 set

Option for all views

Clustering Method	Group average link
Dataset to show	PXRD
Result directory	PXRD

Figure 2.1: Exploring the screen

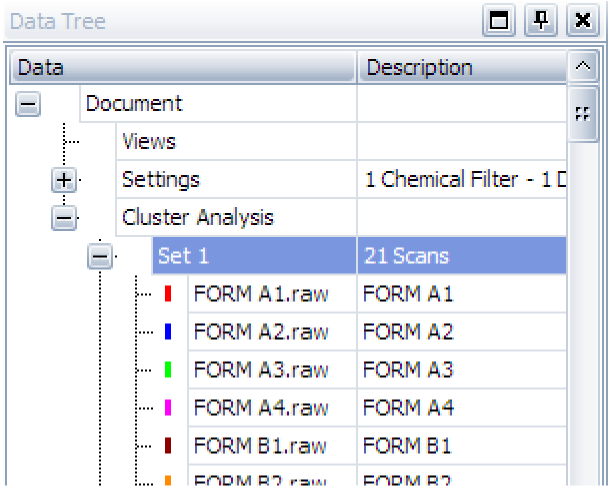
1	Title bar	5	Menu bar
2	Toolbar	6	View window
3	Data command panel	7	Data property panel
4	Data tree panel with Cluster Analysis node		

A node **Cluster Analysis** is created automatically in the data tree which is the starting point for all subsequent actions related to the cluster analysis.

2.2 Preparing for Cluster Analysis

A cluster analysis is carried out by using the **Cluster Analysis** node in DIFFRAC.EVA's data tree.

By default, the **Cluster Analysis** node has a child node **Set 1** which is used to hold the data which are to be analyzed:



Data Tree

Data	Description
⊟ Document	
Views	
⊞ Settings	1 Chemical Filter - 1 D
⊟ Cluster Analysis	
⊟ Set 1	21 Scans
🟥 FORM A1.raw	FORM A1
🟦 FORM A2.raw	FORM A2
🟩 FORM A3.raw	FORM A3
🟪 FORM A4.raw	FORM A4
🟫 FORM B1.raw	FORM B1
🟧 FORM B2.raw	FORM B2

Figure 2.2: Data loaded into data set 1

2.2.1 Multiple Data Sets

Additional data sets can be added by using the command **Add Set** which is available in the **Cluster Analysis** node. Up to four data sets can be used. Multiple data sets have extensive uses but the following are especially relevant to X-ray diffraction:

Using data and derivative data or Fourier transform data.

Using data collected on different instruments or under different conditions.

When multiple data sets are input, each data set is analyzed individually and all the possible combinations of the data sets are used. Data are combined using the individual differences scaling method (INDSCAL) of Carroll and Chang (*Carroll, J.D. & Chang, J.J. (1970). Psychometria 35, 283-319.*) by scaling the differences between individual distance matrices which are derived from the correlations between patterns.

When more than one dataset is used, the cluster analysis will create the following number of individual result sets:

Table 2.1: Number of individual result sets depending on the number of datasets

Number of datasets used	1	2	3	4
Number of individual result sets	1	3	7	15

Potential Problems

Running automatic analysis on multiple datasets at the same time requires coordination of all the individual datasets in order for the results to make sense. For this reason there are problems that can occur when there are inconsistencies between the datasets:

- Each dataset should have the same number of files, which should all have the same filenames. Should there be an inconsistent number of files between datasets then an error message will be displayed and the cluster analysis will be unable to continue the analysis.
- Should there be an equal number of files, but they have been inconsistently named, then a warning message will be displayed.

The user can choose to continue with the analysis, or to abort. If analysis is carried out then the user is warned to treat the combined results with caution as the program can no longer be certain if the correct files have been matched together when creating the combined results. The individual results should be unaffected.

2.2.2 Data Loading

To load data into a data set

1. select the **set node** and
2. use the **Import from Files** command.

A set node has also one child node called **Reference** which may hold known reference samples.

1. Select the **Reference** node and
2. use the **Import from Files** command to load the reference samples.

2.2.3 Data Types and Data Formats

The cluster analysis can be carried out for XRD and XRF scan data.

It is possible to load data in the following formats:

Table 2.2: Data formats which can be used for the cluster analysis

File Extension	Data Format
BRML	DIFFRAC.SUITE measurement files
RAW	DIFFRAC*plus* measurement files
SSD	SPECTRA*plus* measurement files
Y, XY, XYE	Text files with y, x-y, and x-y-error values
TXT	Text files with a leading description line which were used in POLYSNAP 3

BRML

This is the measurement data format of the Bruker AXS DIFFRAC.SUITE measurement software. Files containing single and multiple scans can be read.

RAW

This is the measurement data format of the Bruker AXS DIFFRAC*plus* measurement software for XRD instruments. Files containing single and multiple scans can be read.

SSD

This is the measurement data format of the Bruker AXS SPECTRA*plus* measurement software for XRF instruments. Files containing single and multiple scans can be read.

Y

This is a text data format with one y value per line. The first line must contain some additional information:

_x1_dx 20.02 0.02

The text “_x1_dx” is the tag for the first line. The second value is the start position. The third value is the step size. The values are separated by spaces.

Files containing single scans can be read.

XY, XYE

This is a text data format containing x and y values (XY) and additionally an error value (XYE) per line. If the first line does not start with a number it may contain additional information. The values belonging to the same step must be separated either by space or tab characters.

Files containing single scans can be read.

If the first line does not start with a number, the software assumes that parameter names and values are given in pairs. The following parameters are read from the first line if available:

Table 2.3: The parameters which are read from the first line of a XY or XYE file

Parameter	Description
id	Sample identifier
comment	User comment
operator	Instrument operator
anode	X-ray tube anode
scantype	Scan type, e.g. step scan or continuous scan
Timeperstep	Measurement time per step

TXT

This is a text data format containing x and y values per line. If the first line does not start with a number it may contain additional information. The values belonging to the same step must be separated either by space or tab characters. The step sizes must be constant.

Files containing single scans can be read.

2.3 Performing Cluster Analysis

As soon as at least three samples are in a data set, the **Tool** command group for the **Cluster Analysis** node becomes active.

To carry out a cluster analysis,

1. select the **Cluster Analysis** command in the **Tool** command group.

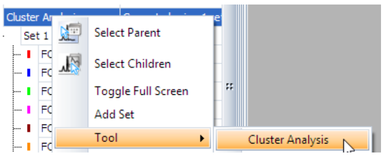
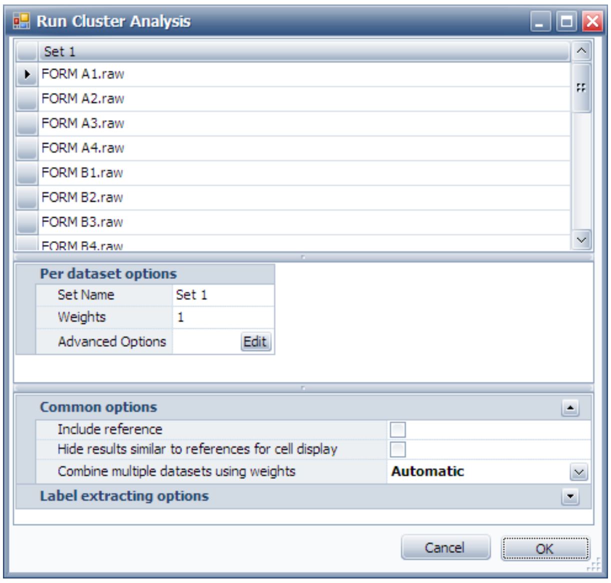


Figure 2.3: The Cluster Analysis command in the context menu

⇨ The **Run Cluster Analysis** dialog opens:



Run Cluster Analysis

Set 1

FORM A1.raw

FORM A2.raw

FORM A3.raw

FORM A4.raw

FORM B1.raw

FORM B2.raw

FORM B3.raw

FORM B4.raw

Per dataset options

Set Name	Set 1
Weights	1
Advanced Options	Edit

Common options

Include reference	□
Hide results similar to references for cell display	□
Combine multiple datasets using weights	Automatic

Label extracting options

Figure 2.4: The Run Cluster Analysis dialog

The **Run Cluster Analysis** dialog allows changing some basic parameters like the weights for a set and some common and labelling options. In general, the default options work best.

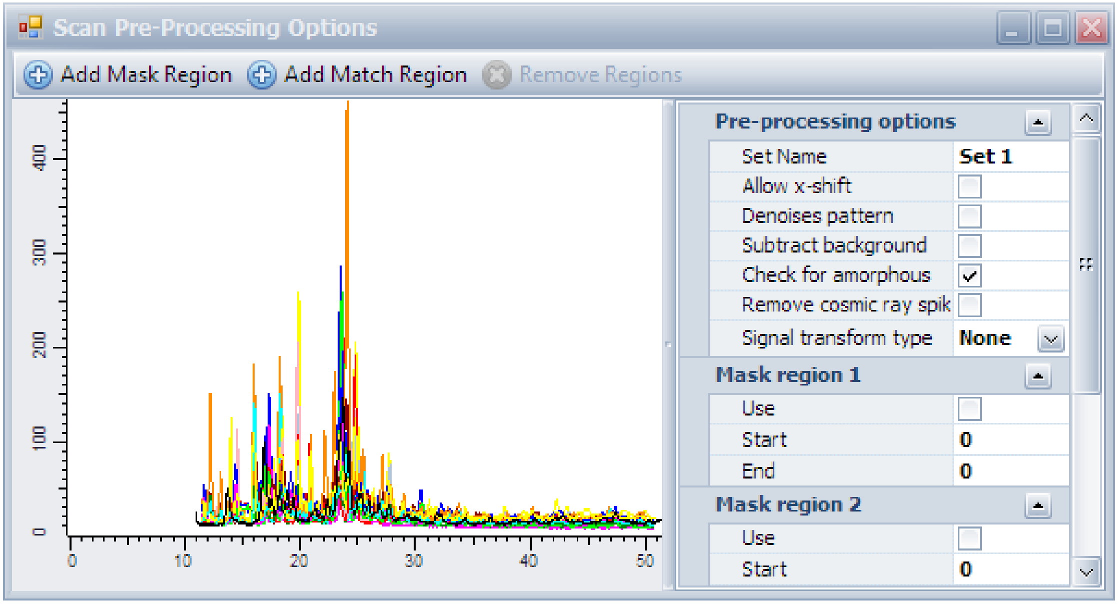
At this point every one of the *n* patterns is correlated with every other pattern using a combination of Pearson and Spearman correlation coefficients to generate a *(n×n)* symmetric correlation matrix ρ**.** This forms the basis of all subsequent calculations.

2.3.1 Scan Pre-Processing Options

There are other parameters for the cluster analysis which are preset with meaningful values. To display these parameters and possibly change them

1. click on the **Edit** button for the **Advanced Options**.

⇨ The **Scan Pre-Processing Options** window opens:



Pre-processing options

Set Name	Set 1
Allow x-shift	□
Denoises pattern	□
Subtract background	□
Check for amorphous	☑
Remove cosmic ray spik	□
Signal transform type	None

Mask region 1

Use	□
Start	0
End	0

Mask region 2

Use	□
Start	0

Figure 2.5: The Scan Pre-Processing Options window

The scan pre-processing options can be changed in this window. Up to three regions can be excluded from the analysis (Mask Region) or one region can be selected for the analysis if only a certain range of the data is of interest (Match Region).

Any change in the pre-processing option is carried out immediately. The window must be closed using the red close button in the upper right corner.

2.3.2 Start the Analysis

After setting all parameters the cluster analysis is ready to be started.

To run the cluster analysis,

1. click the **OK** button in the **Run Cluster Analysis** dialog.

⇨ A progress bar is displayed during the analysis. Depending on the number of samples the analysis can take from seconds to minutes to be finished.

2.3.3 Evaluating the Results

When the cluster analysis is finished, the results are displayed in one or more views. The most important view for a cluster analysis is the **Dendrogram** view:

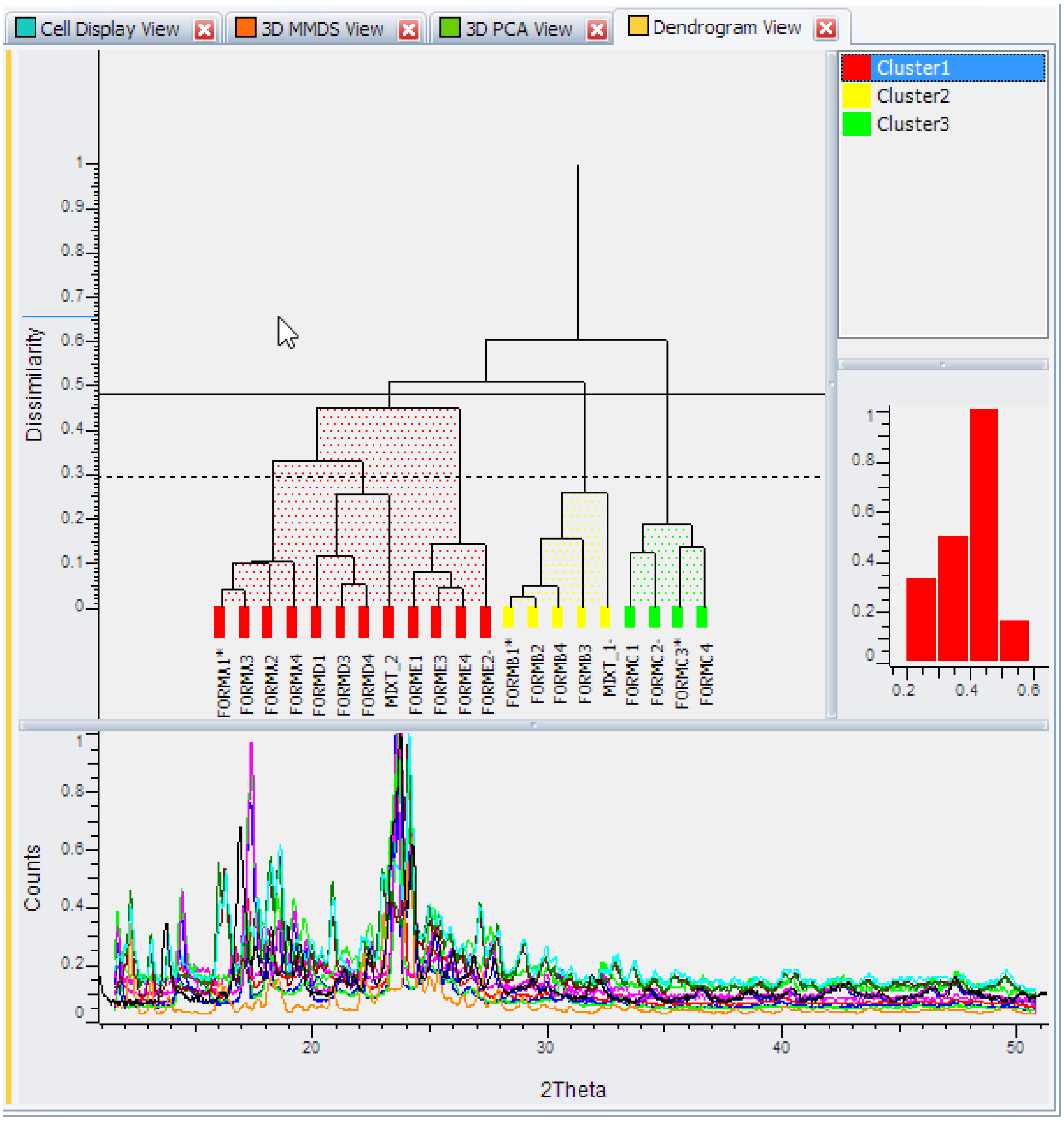


Figure 2.6: A typical dendrogram view containing also a silhouette and a scan graphics

The DIFFRAC.EVA cluster analysis present not just the dendrogram in this view (upper left area), but also a cluster selection in the upper right corner, a silhouette view below and a scan view for the selected samples in the lower part of the view.

Other important views like the **Cell Display**, the **3D MMDS** and the **3D PCA** views can be configured to be displayed automatically after each cluster analysis. These are described in later sections of this manual.

2.3.4 Repeating the Cluster Analysis

The analysis can be repeated if required, e.g. if more samples have been added to a set or a new data set was created.

To do so, the cached analysis results must be deleted with the **Clean Results** command, which is available in the **Tool** command group:

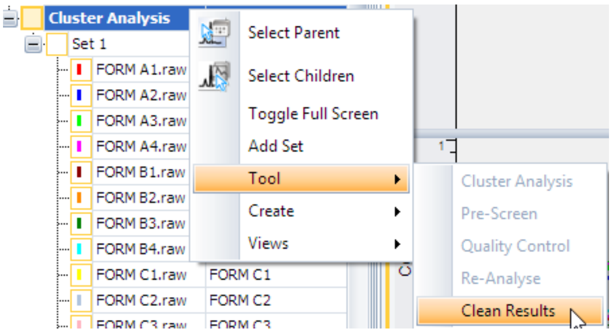
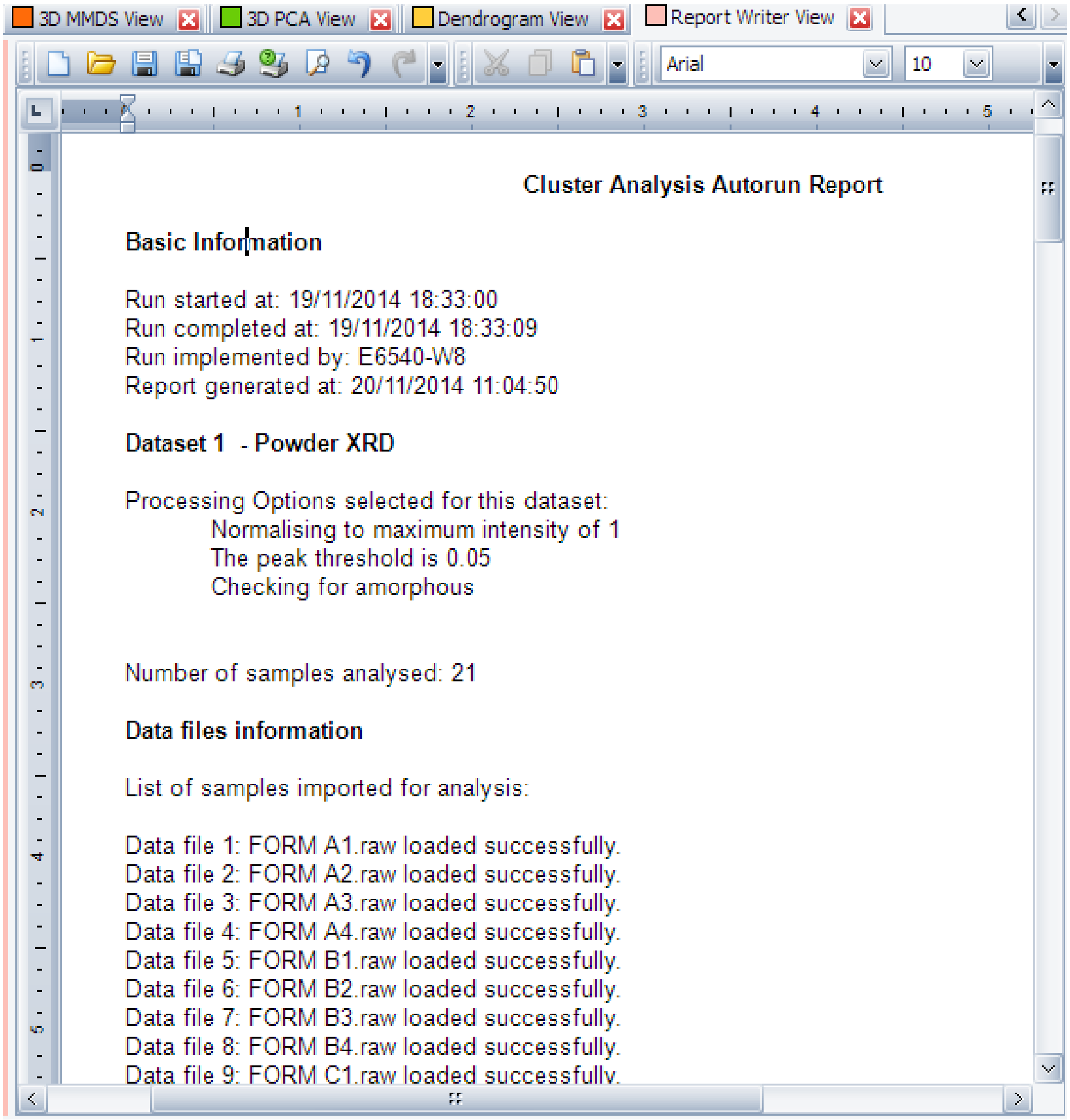


Figure 2.7: The Clean Results command

After cleaning the results all cached data and the corresponding views are deleted. The analysis can be started again using the **Cluster Analysis** command.

2.3.5 Publish the Results

When the cluster analysis results has been inspected and possibly modified, the results can be displayed in specialized output views such as the **Report Writer** view:



Cluster Analysis Autorun Report

Basic Information

Run started at: 19/11/2014 18:33:00
Run completed at: 19/11/2014 18:33:09
Run implemented by: E6540-W8
Report generated at: 20/11/2014 11:04:50

Dataset 1 - Powder XRD

Processing Options selected for this dataset:
Normalising to maximum intensity of 1
The peak threshold is 0.05
Checking for amorphous

Number of samples analysed: 21

Data files information

List of samples imported for analysis:

Data file 1: FORM A1.raw loaded successfully.
Data file 2: FORM A2.raw loaded successfully.
Data file 3: FORM A3.raw loaded successfully.
Data file 4: FORM A4.raw loaded successfully.
Data file 5: FORM B1.raw loaded successfully.
Data file 6: FORM B2.raw loaded successfully.
Data file 7: FORM B3.raw loaded successfully.
Data file 8: FORM B4.raw loaded successfully.
Data file 9: FORM C1.raw loaded successfully.

*Figure 2.8: The **Report Writer** view as an example for a text output view*

All cluster analysis views can be printed by using the **Print Preview** command.

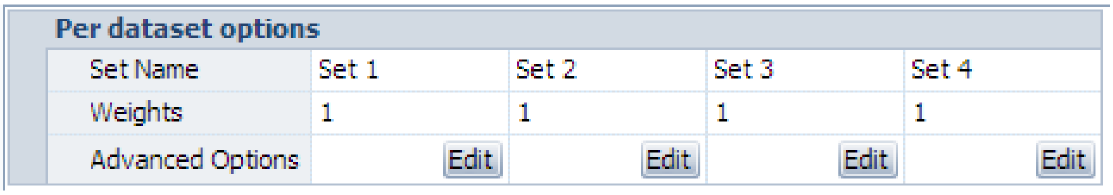
3 Clustering and Pre-Processing Options

3.1 Run Cluster Analysis Dialog

The **Run Cluster Analysis** dialog allows setting some options which are described below.

3.1.1 Per dataset options

Per dataset options are displayed in a table with values for every active dataset individually.



Per dataset options

Set Name	Set 1	Set 2	Set 3	Set 4
Weights	1	1	1	1
Advanced Options	Edit	Edit	Edit	Edit

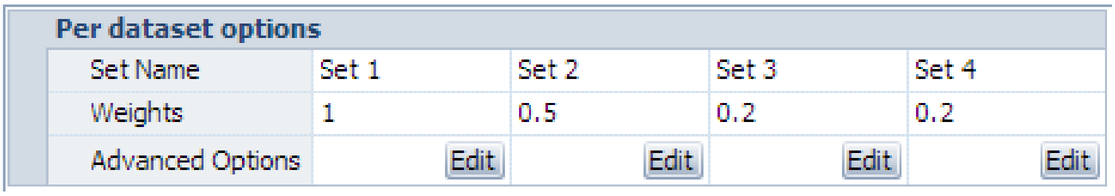
Figure 3.1: The per dataset options for four datasets

Set Name:

This is by default Set *x (x = 1 .. 4)* and cannot be changed.

Weights:

The default is one and the range is from zero to one. If several sets are used they can be weighted differently. This is possible only if the common option **Combine multiple datasets using weights** is set to **Manual**.



Per dataset options

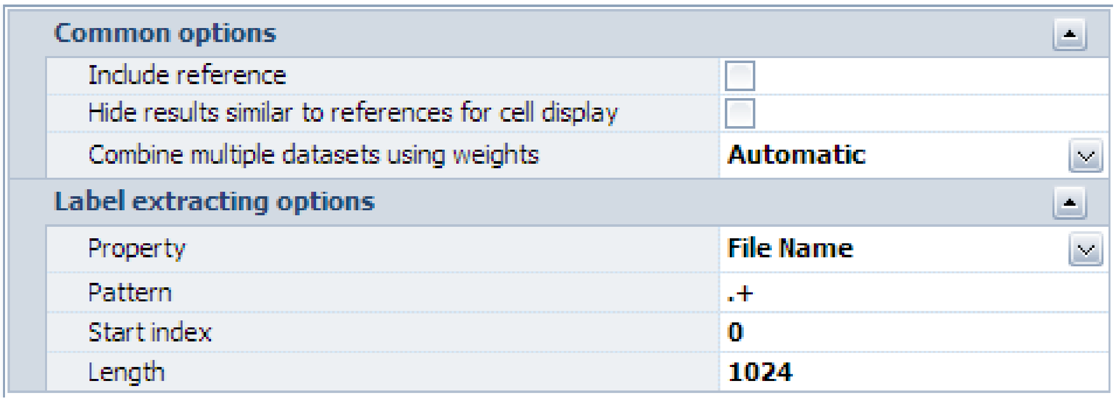
Set Name	Set 1	Set 2	Set 3	Set 4
Weights	1	0.5	0.2	0.2
Advanced Options	Edit	Edit	Edit	Edit

Figure 3.2: The per dataset options for four datasets with manually changed weights

Advanced Options:

An **Edit** button opens the **Scan Pre-Processing options** window.

3.1.2 Common Options



Common options

Include reference	□
Hide results similar to references for cell display	□
Combine multiple datasets using weights	Automatic

Label extracting options

Property	File Name
Pattern	.+
Start index	0
Length	1024

Figure 3.3: The Common and Label extracting options

Include reference:

This property applies to all datasets where reference files have been provided. This allows the reference files to be plotted in all the graphical displays along with the samples, for a visual comparison of how close the references files are to the sample files. When this option is deselected the reference files are still used to analyze the samples, however are not included in any of the displays.

Hide results similar to reference for cell display:

If this option has been chosen, any samples which are good matches to the provided reference patterns (or appear to be composed purely of a mixture of those references) are greyed out on the results display, so that only new samples unlike anything seen previously are highlighted. Such patterns are also hidden by default on the 3D plots, and taken off to one side on the dendrogram.

Combine multiple datasets using weights:

This option is valid for multiple datasets only. If this option is set to **Automatic** (the default), the cluster analysis combines the results using the **INDSCAL** method. This is strongly recommended but it can be overridden by setting this option to **Manual** and entering numerical values for the weights manually in the **Per dataset options** (see figure: *The per dataset options for four datasets with manually changed weights [▸ 21]*).

Selecting **None** means that the datasets will not be combined and the results will only be presented as individual cases.

3.1.3 Label extracting options

These are advanced options. They control the behavior of the filter applied to the scan input data for the cluster analysis. It works by filtering through the label string of each scan and keeping only those scans which have been validated for cluster analysis and which are shown in the above grid.

There are four options in this category:

Property:

Specifies which property of the scans is to be used as an incoming label; either **File name** or *S***ample name**.

Pattern:

A RE (Regular Expression) pattern to be used as a filter to be applied to the sub-string of the incoming label.

Start index:

A zero based start index of the sub-string in the incoming label.

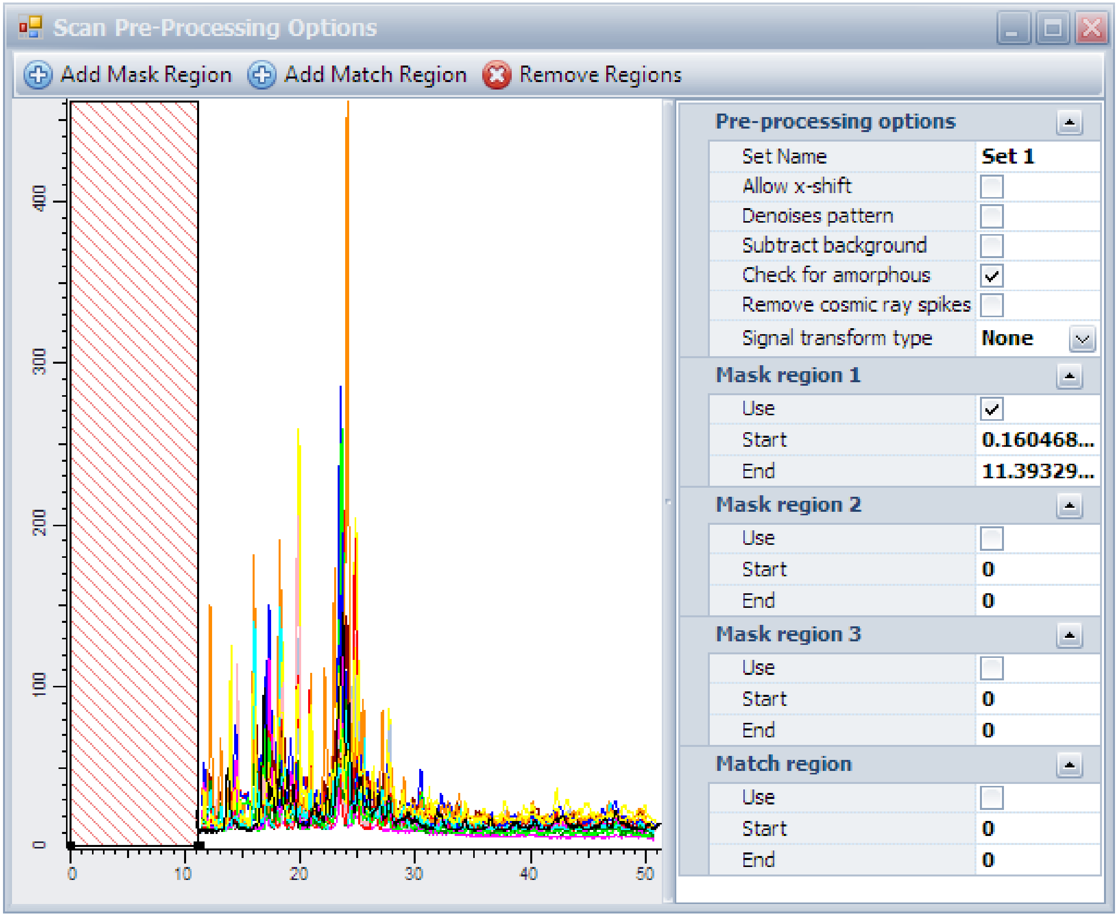
Length:

The length of the sub-string of the incoming label to be used.

The default values for these options are set in a way that entire labels up to 1024 characters are used.

3.2 Scan Pre-Processing Options Window

The **Scan Pre-Processing Options** window displays an overview of the scans which belong to the selected set on the left and a property grid on the right. **Toolbar** buttons are available on top of the window.



Pre-processing options	
Set Name	Set 1
Allow x-shift	□
Denoises pattern	□
Subtract background	□
Check for amorphous	☑
Remove cosmic ray spikes	□
Signal transform type	None
Mask region 1	
Use	☑
Start	0.160468...
End	11.39329...
Mask region 2	
Use	□
Start	0
End	0
Mask region 3	
Use	□
Start	0
End	0
Match region	
Use	□
Start	0
End	0

Figure 3.4: The Scan Pre-Processing window displaying the graphical scan overview with a masked region and the property grid

The properties have the following functionality:

Set name:

Displays the set name and cannot be changed.

Allow x shift:

This option controls how patterns will be shifted in an attempt to maximize the correlations between them. An optimal shift in 2θ between patterns is often required arising from equipment settings, especially due to variation in the sample height, and data collection protocols. We use the forms:

$$\Delta(2\theta) = a_0 + a_1 * \cos(\theta)$$

which corrects for varying sample heights in reflection mode, or,

$$\Delta(2\theta) = a_0 + a_1 * \sin(\theta)$$

which corrects for transparency errors or, for example, transmission geometry with constant specimen-detector distance, and

$$\Delta(2\theta) = a_0 + a_1 * \sin(2\theta)$$

which provides transparency and thick specimen error corrections. The parameters a_0 and a_1 are constants refined automatically to maximize pattern correlations. The a_0 and a_1 parameters are determined by the program. The default value is a *sinθ* shift. Other methods can be selected in the **Cluster Analysis** page of the **Settings** dialog, **Matching & References** area.

Denoises pattern:

Removes noise from a signal using wavelet smoothing. The smoothing factor can be set in the **Settings** dialog, on the **Cluster Analysis** page, in the **Data Input** and **Processing** section.

Subtract background:

Removes background signal where present. It is important that the correct data type is specified before using this option, as the type of background subtraction applied is different for each type of dataset.

Check for amorphous:

Applicable only to powder X-ray diffraction datasets. When selected this option monitors the patterns for any samples that appear to be amorphous or non-crystalline. These samples are then labelled as such in the cell display, and can optionally be discarded.

Remove cosmic ray spikes:

When selected this option monitors the Raman pattern for cosmic ray spikes (characterized by an extreme intensity peak with a very narrow range) and removes them from the pattern. This should not be used with XRD data, and is provided for old Raman instruments where cosmic ray signals were not filtered out.

Signal transform type:

Selecting this option brings up a dialog box allowing the selection of either a Fourier transformation, or first or second order derivative, to be applied to the selected spectrum.

Mask region x (x=1..3):

This option allows the user to set a sub-region of the pattern to a value of zero. This tool is useful when the pattern contains a spurious peak or unwanted standard, or if for some reason the user wishes to only compare a certain region of the pattern. Up to three separate regions of a pattern can be masked out. Clustering can be improved by judicious use of masking. The software suggests suitable regions after each cluster analysis and this can be examined in the log file.

The **Mask** regions can be added by using the **Add Mask Region** button on top of the window or the edit fields in the property grid.

Match region:

This option allows the user to select the sub-range within the pattern that will be used for the matching process, disregarding the other sections of the pattern. This is used when there is only a specific region of the pattern that is of interest.

The **Match** region can be added by using the **Add Match Region** button on top of the window or the edit fields in the property grid.

4 Results Display

An integral part of cluster analysis is the associated set of visualization tools. The cluster analysis results are presented in a comprehensive set of different views which are described below. All views are grouped in the view window which presents every view in a stacked tab.

1. Click a tab title to display the corresponding view.

To remove a view from the display,

1. click the **Close** button ☒ in the corresponding tab heading.

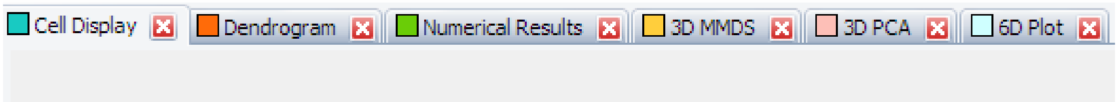


*Figure 4.1: Example of **Result views** in tabs*

There are eleven types of graphical views available:

Cell Display, **Dendrogram**, **3D MMDS**, **3D PCA**, **6D Plot**, **Scree Plot**, **Minimum Spanning Tree**, **Silhouettes**, **Fuzzy Clustering**, **Parallel Coordinates Plot** and **Space Explorer**.

The views can be opened by using the commands which are available on the **Cluster Analysis** node:

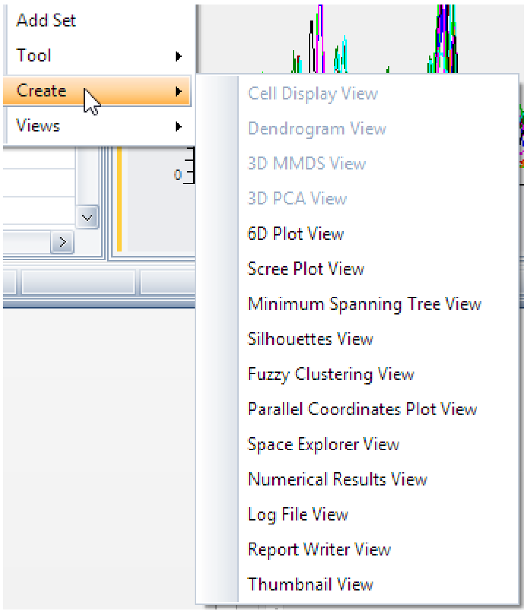


Figure 4.2: The Create context menu for the Cluster Analysis node

Every individual view can be opened only once at a time. If a view is already displayed, its command entry is greyed out in the menu. If a view is closed with the red button on its tab the create command becomes active and the view can be opened again.

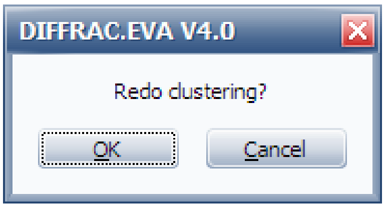
4.1 Cluster Analysis Node Properties

The Cluster Analysis node has some properties to control the results display for all views.

Table 4.1: Cluster Analysis node properties

Property	Description
Option for all views	
Clustering Method	Select the clustering method: Single link Complete link Weighted average Link Centroid Group average link (default) Automatically select best
Dataset to show	The current version allows only “PXRD” data.
Results directory	The current version allows only “PXRD” as directory.

If the clustering method has been changed, the programs displays the following message:



Redo clustering?

If **OK** is clicked, the cluster calculation is repeated according to the selected method. If **No** is chosen, the property is reset to its former value.

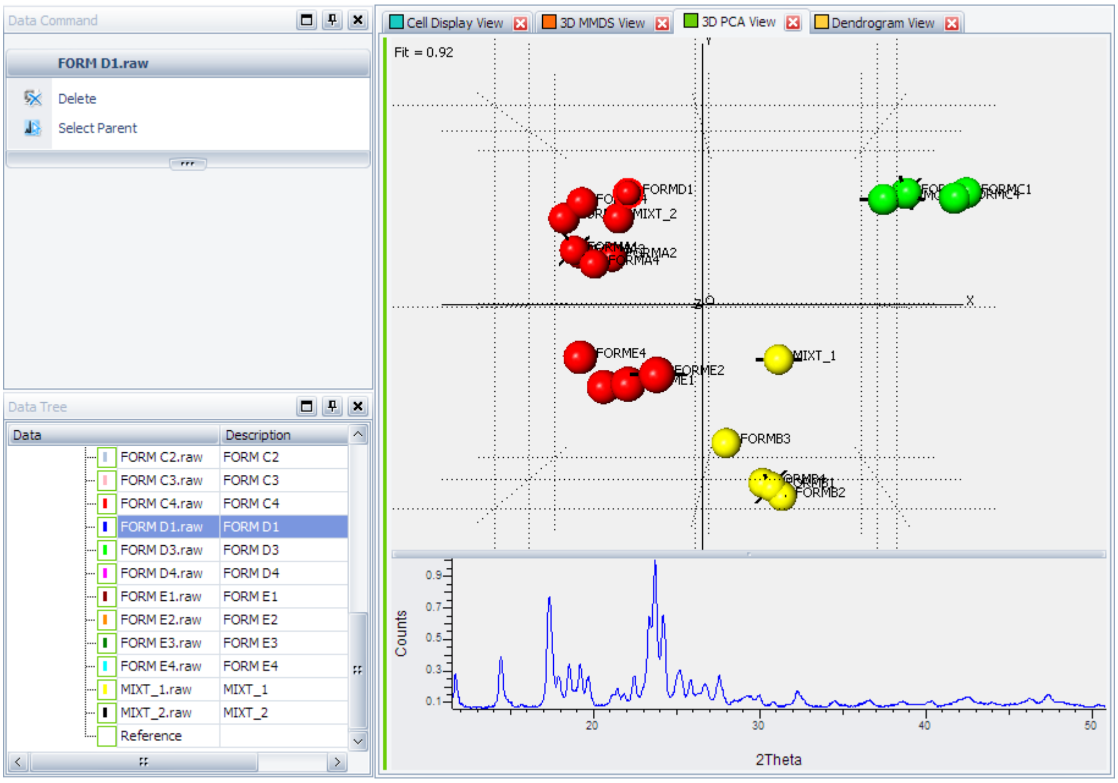
4.2 General View Features

The various graphical display panes - the **Cell Display**, **Dendrogram**, **Screen** and the **3D MMDS** and **PCA** plots - all share many similarities in their options and controls, and are hence initially described together here. They are all important aids to visualizing the data.

4.2.1 Single and Multiple Selection

The samples in most of the views respond to clicks with the left mouse button. As a result, they become selected and change their appearance. Selected samples in the dendrogram are drawn as a longer rectangle and selected samples in **3D** views are drawn with a highlighted border. A selection in the **Cell Display** draws the sample in a different angle than the other samples and draws a thin squared selection area around the sample.

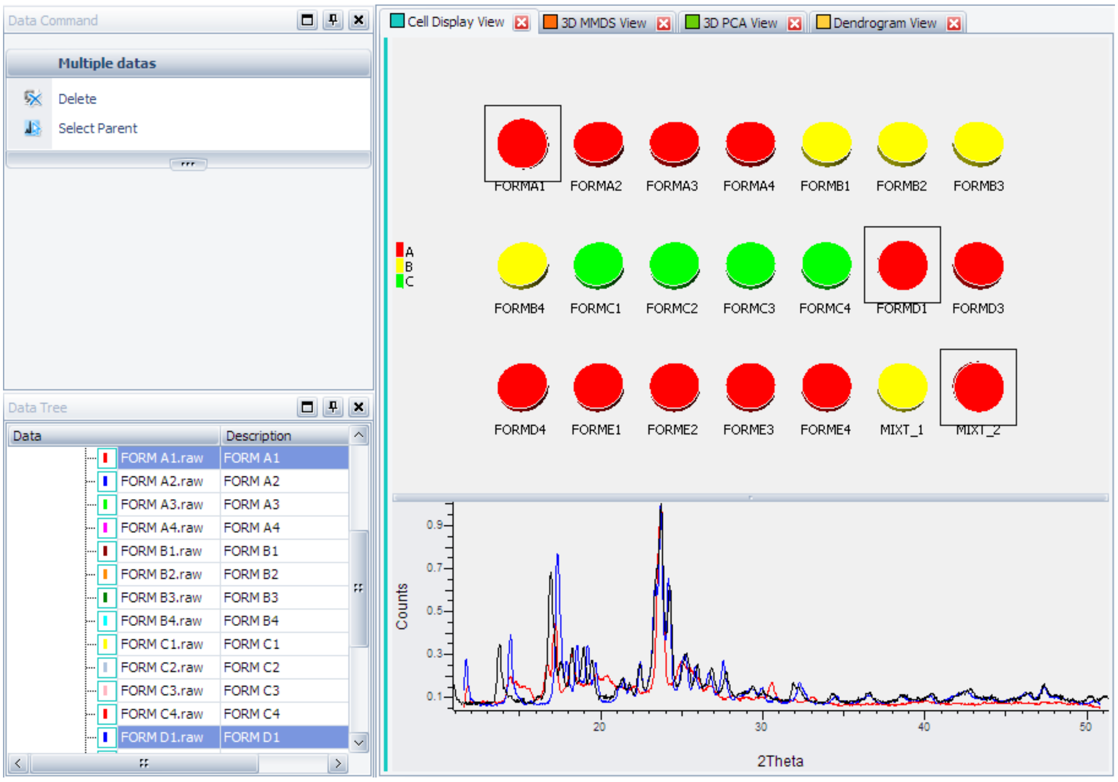
The selection may trigger additional actions: the samples will be selected in the data tree and their measured data may be displayed in a graphical control below the view:



Data	Description
FORM C2.raw	FORM C2
FORM C3.raw	FORM C3
FORM C4.raw	FORM C4
FORM D1.raw	FORM D1
FORM D3.raw	FORM D3
FORM D4.raw	FORM D4
FORM E1.raw	FORM E1
FORM E2.raw	FORM E2
FORM E3.raw	FORM E3
FORM E4.raw	FORM E4
MIXT_1.raw	MIXT_1
MIXT_2.raw	MIXT_2
Reference	

Figure 4.3: 3D PCA view with a single selection for sample FORM D1

Multiple non-contiguous selection is achieved by clicking on multiple samples with the **Control** key held down on the keyboard, for example:



Data	Description
FORM A1.raw	FORM A1
FORM A2.raw	FORM A2
FORM A3.raw	FORM A3
FORM A4.raw	FORM A4
FORM B1.raw	FORM B1
FORM B2.raw	FORM B2
FORM B3.raw	FORM B3
FORM B4.raw	FORM B4
FORM C1.raw	FORM C1
FORM C2.raw	FORM C2
FORM C3.raw	FORM C3
FORM C4.raw	FORM C4
FORM D1.raw	FORM D1

Figure 4.4: Multiple selection in a Cell Display view

Individual patterns can be de-selected in a similar manner, and their profiles will be removed from the graph.

Alternatively, a continuous number of consecutively displayed patterns may be selected by holding the **Shift** key down and clicking on the first and then last pattern in the desired range.

4.2.2 Zooming and Panning

▶ To zoom in to a region of a graphical display:

1. with the left mouse button held down, drag a rectangle over the region you wish to zoom.
   - ➨ The screen will then redraw with the contents of the rectangle filling the display area.

▶ To move the contents of the display window (panning) (for example to move the contents up to see more results than will fit in the window by default)

1. hold down the **Alt** key on the keyboard, and drag the mouse in the desired direction of movement.

4.2.3 Mouse and Keyboard Actions in 3D Views

The generally available actions in all **3D** views are described in the following table.

Table 4.2: Generally available mouse and keyboard commands in 3D views

Mouse/Keyboard Action	Description
Mouse Drag	Zoom In/Out
Shift+Mouse Drag	Rotate
Alt+Mouse Drag	Translate
Click Object	Select
Ctrl+C	Copy Graph
I	Show ID
H	Show help window

More actions may be available in the individual views and are described in the respective sections.

4.2.4 Graphical Context Menu

By clicking the **right mouse** button the **graphical** context menu is displayed. It contains the commands:

Table 4.3: Command in the graphical context menu

Command	Description
Reset	Resets the zoom area
Deselect all	Resets the selection

4.2.5 View Properties

Views have a set of properties which control the view’s appearance or how it is to be printed.

The following properties are common to all views:

Table 4.4: General view properties for all views

Property	Description
Printing	
Printable	Select the check box to print the selected view
Paper orientation	Paper orientation: portrait, landscape or default. If the default option is selected, the paper orientation chosen in the print preview will be applied to the view.
DPI Multiple	For graphical views only: factor for result bitmap resolution over screen resolution
View	
Name	Name chosen for the view
Description	View description. Can be edited
Scan Options	
Original Scans	Display the original scans instead of the pre-processed scans
Auto Reset Zoom	Reset zoom if the content of the scan control is changing
Y Shift	Y shift in pixels to visually separate the scans in y direction

The view specific properties are described in the view specific sections.

4.3 Graphical Views

4.3.1 Cell Display View

The cell display is one of the default views for the cluster analysis.

Each individual scan is represented by a pie chart which is colored according to the cluster it belongs to. The cluster colors come from the dendrogram. If one or more scans have been selected their 1D representation is displayed below the cells.

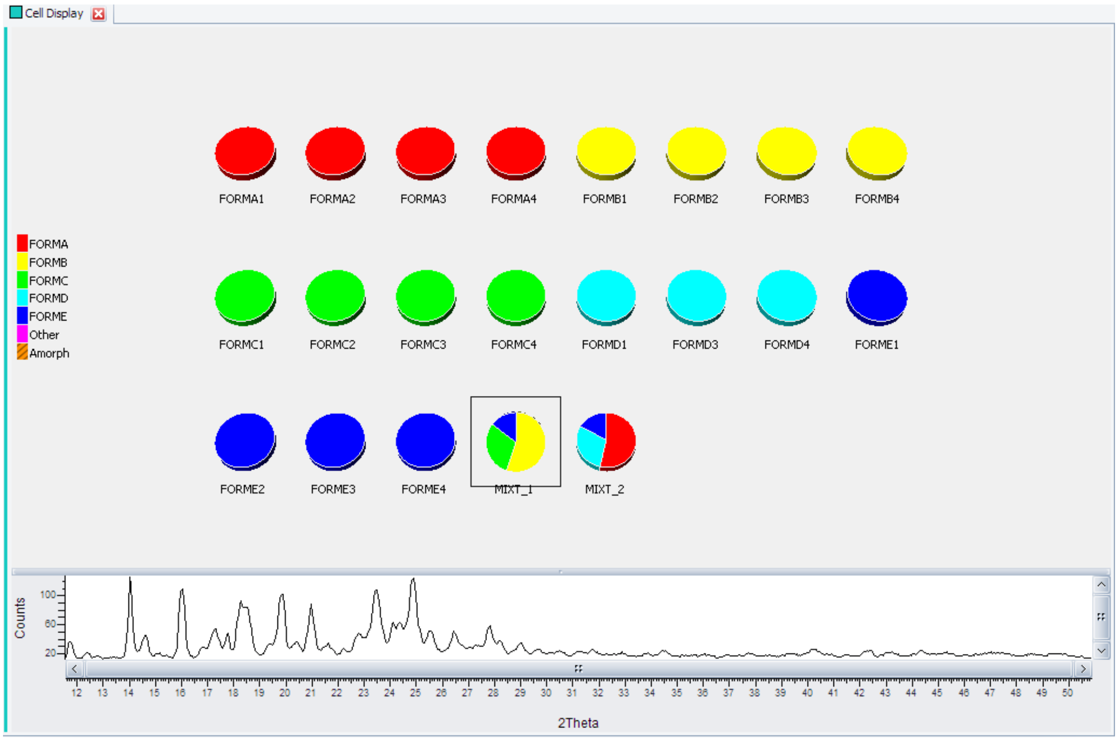


Figure 4.5: Cell Display view in the default pie chart mode

A list of the detected clusters with their color markers is displayed on the left side of the graphics.

The cell display can also be displayed in the stacked mode:

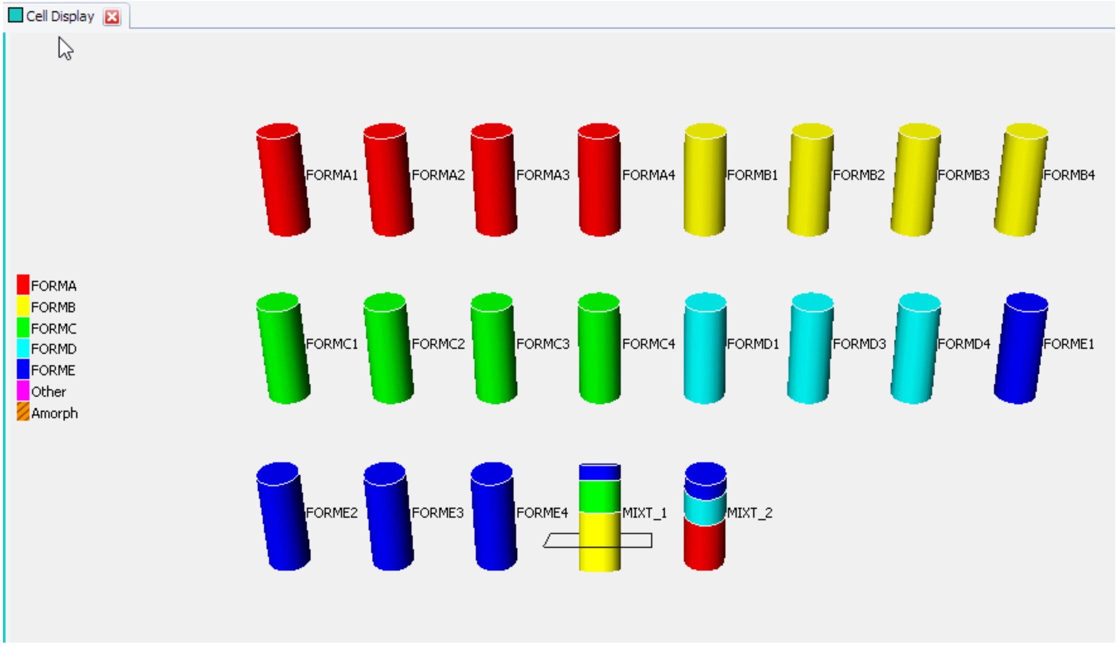


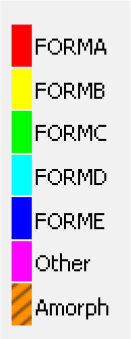
Figure 4.6: Cell Display view in stacked mode

To switch between modes, use the **Cell Display's** property **Stack Mode**.

Depending on the presence or absence of known phases for a given run, the color-coding is obtained from two different sources.

If known phases are available:

If a database of known phases was provided, semi-quantitative analysis is carried out using the reference pattern profiles giving an estimate of the composition of each sample. The key on the left hand side of the display shows a list of those known phases, with the labels shown generated from the relevant pattern filenames:



Each known phase has a unique color assignment; in addition colors are shown for patterns considered to be either non-crystalline/amorphous (Amorph) or unlike any other known pattern provided (Other).

Each individual pattern that corresponds to a known phase (*i.e.* one that gave good matching statistics when being compared to a known phase) is given the same color as that known phase - for example, in the screenshot above, the patterns that matched well to Form B are all 38 the same color of yellow. All of the patterns that match to Form B can be selected at once by clicking on the color in the key:

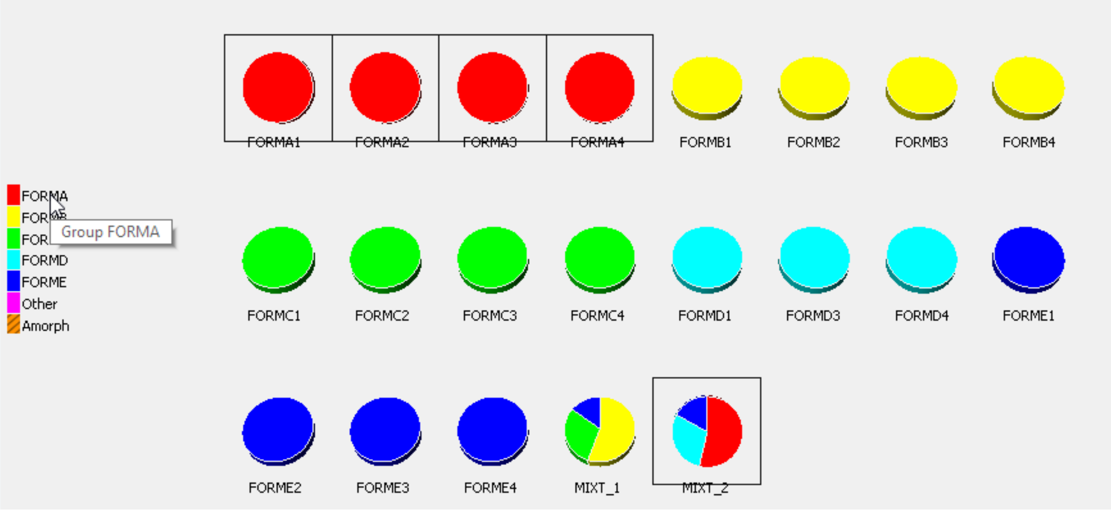


*Figure 4.7: **Cell Display** with selected cluster **FORMA**, known references present*

A pie that contains multiple colors (*e.g.* sample MIXTURE above) represents a pattern which is thought to be a mixture of two or more of the known patterns, and the colors within the pie-chart again correspond to the phases thought to comprise it. Allowing the cursor to hover over a particular component of a mixture brings up a tooltip describing which phase is present and in what amount:

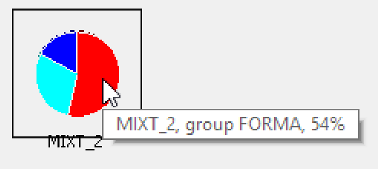


Figure 4.8: A cell which represents a mixture, tooltip with percentage displayed

Hide results similar to references for cell display

If this option has been chosen, any samples which are good matches to the provided reference patterns are greyed out on the display, so that only new samples which unlike anything seen previously are highlighted. Such patterns are also hidden by default on the 3D plots, and also placed on one side on the dendrogram.

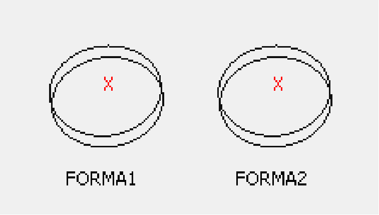


Figure 4.9: ***Cell Display*** *when* ***Hide results similar to references*** *has been selected*

If no known phases are available:

On the other hand, if no known phases are available what *appear* to be similar results are displayed:

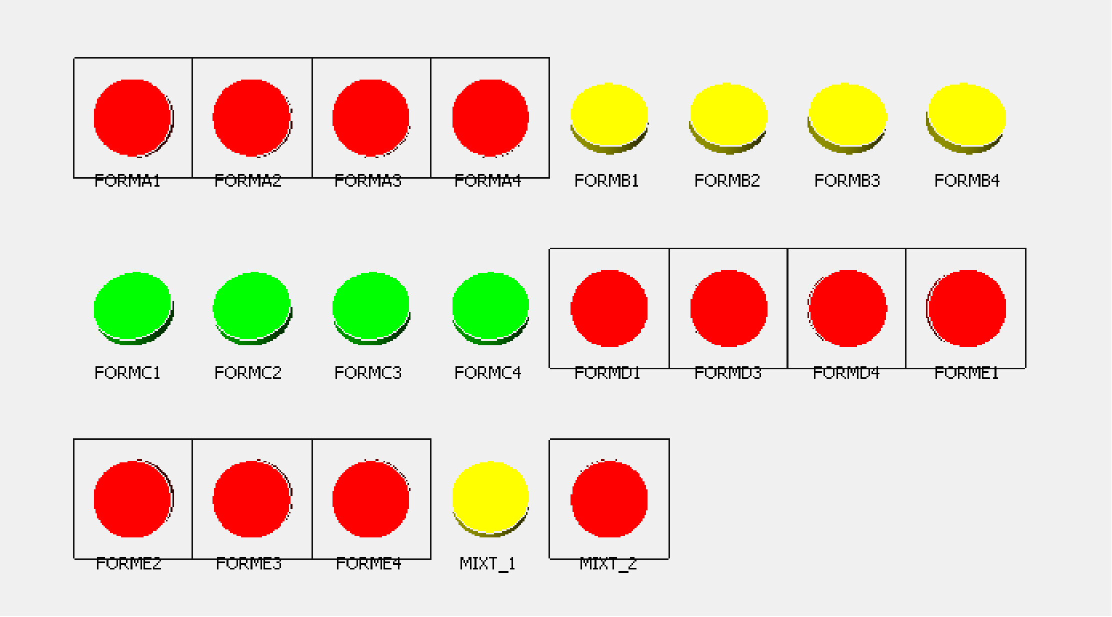


Figure 4.10: ***Cell Display*** *with selected cluster* ***FORMA****, no known references present*

However, there are subtle and very important differences. The colors of the pies obviously no longer correspond to those of known phases - the colors are now merely representative of patterns which are similar to each other. In the example above Pies 1, 2, 3 and 4 are similar to each other as are 5, 6, 7 and 8. Pattern 21 is dissimilar to all the other samples, as it is a color not shared by any other samples. The colors of the pies match those of the dendrogram.

Because clustering results are being used, the groups shown are entirely dependent on the cut-level used on the dendrogram display. The cut-level is discussed in more detail in the Dendrogram section, but it is important to note that when no known phases are present, altering the cut-level or otherwise editing the dendrogram display will cause the cell display colors to be altered and updated accordingly.

This display mode can also be accessed when known phases are present, by means of the Pseudo Cell Only option in the **Cell Display's** properties.

Mouse Actions and Keyboard Shortcuts

Table 4.5: Generally available mouse and keyboard commands in the Cell Display

Mouse Action	Description
Mouse Drag	Zoom In/Out
Shift+Mouse Drag	Rotate
Alt+Mouse Drag	Translate
Click Object	Select
Ctrl+C	Copy Graph
I	Show ID
H	Show help window

Cell Display View Properties

Table 4.6: Cell Display view properties

Property	Description
Action	
Stack Mode	The samples are displayed as rods instead of tablets if checked. If samples were detected as mixtures the rods are colored stacks.
Pseudo Cell Only	Ignore reference phases if present and display as if no known phases were available
Attributes	
Show Label	Show labels for all samples
Show Selected Label	Show labels for selected samples
Max Label Length	Maximum label length defined as the number of characters
Line Width	Outer line width for the cell drawing
Background Color	View’s background color
Foreground Color	Text color for label display
Label Font	Font for label display
Scale Factor	Scale factor for cell size
3D Attributes	
Rendering Quality	The factor to control rendering quality

4.3.2 Dendrogram View

The dendrogram provides a visual means to display the results of the hierarchical method of data classification using cluster analysis. It is the most important technique for visualizing the results of cluster analysis and should always be examined first. The dendrogram itself takes the form of a tree-diagram in which each single terminal branch is representative of a single object (in this case an individual pattern from the data input).

The initial cut point is set by the program, and is shown by the solid horizontal line. Upper and lower confidence limits on this cut-level are shown with dotted lines where appropriate.

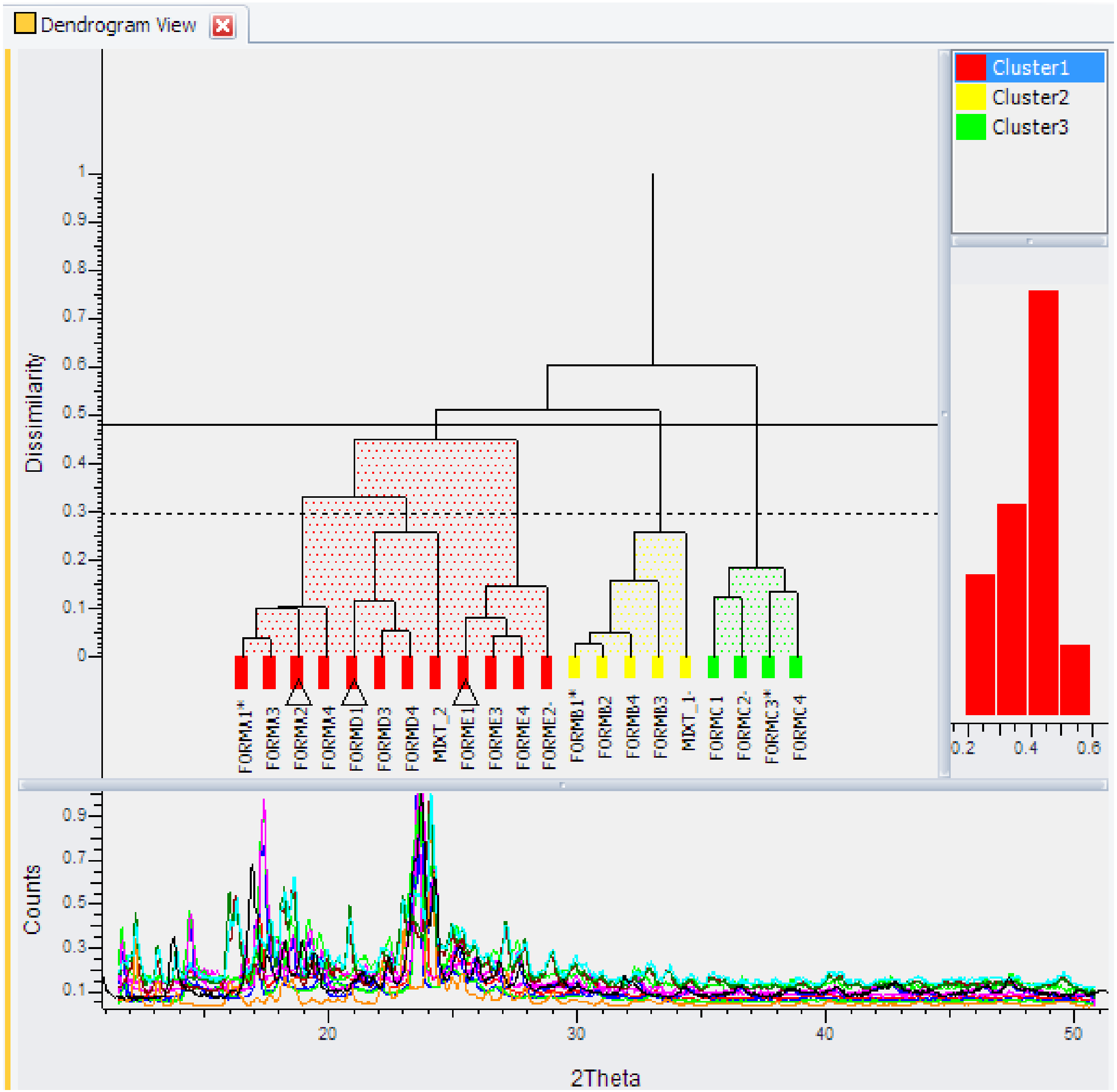


Figure 4.11: Dendrogram view displaying the dendrogram control, cluster list, silhouette and scan control

Every scan is represented by a rectangular box which has the color of the cluster it belongs to. If one or more scans have been selected the box height is increased and the scan's 1D representation is displayed below the diagram.

Each pattern is marked with an ID along the bottom axis of the dendrogram. Each ID is the same as in the other displays - for example, FORMA1 on the dendrogram is the same pattern as FORMA1 in the cell display.

Each pattern can be selected by clicking on the box above its number. When a pattern is selected, the sample information along with its pattern profile is displayed in the bottom half of the display window.

Multiple patterns can be selected by holding down the Control key and clicking on different patterns. A series of consecutive patterns can be quickly selected by clicking on the first pattern then holding the Shift key and clicking on the last pattern. This selects all the patterns in the range in one step.

Note that once a particular pattern is selected, the patterns on either side can be selected in turn by means of the left and right arrow keys, thus allowing for quickly scanning through multiple patterns.

The view of the dendrogram can be zoomed in on by dragging a rectangle over the relevant area with the left-hand mouse button down as with the graph and cell displays.

The position of the dendrogram on the horizontal axis can be altered by holding down the Alt key and moving the mouse left or right. This can be useful if the zoom has been used and the whole tree no longer fits on the screen at one time.

Patterns are joined together by a series of lines. The further up the similarity axis (y-axis) the patterns are joined, the less similar they are. Therefore, in the screenshot above, patterns 1 and 3 are joined at a high level of similarity (nearly 1.0), and are therefore very similar, whereas patterns 1 and 12 are not joined until a similarity of less than 0.4, indicating a large difference between them.

Given the calculated similarity between patterns, it is then possible to categorize similar patterns as belonging to the same cluster. This is done by drawing a horizontal line across the display at a given similarity level - this is called the cut-level.

The optimum cut-level is determined by the dust analysis using a combination of several different techniques in order to determine the number of clusters that statistically best represents the data given. These techniques include principle component analysis, metric multidimensional scaling, the C-H test, gamma statistics, etc. Further details are given in the references listed at the end of this chapter.

The cut-level is then drawn on the dendrogram, and different patterns which are grouped together below this line are considered to be similar enough to be thought of as being in the same cluster:

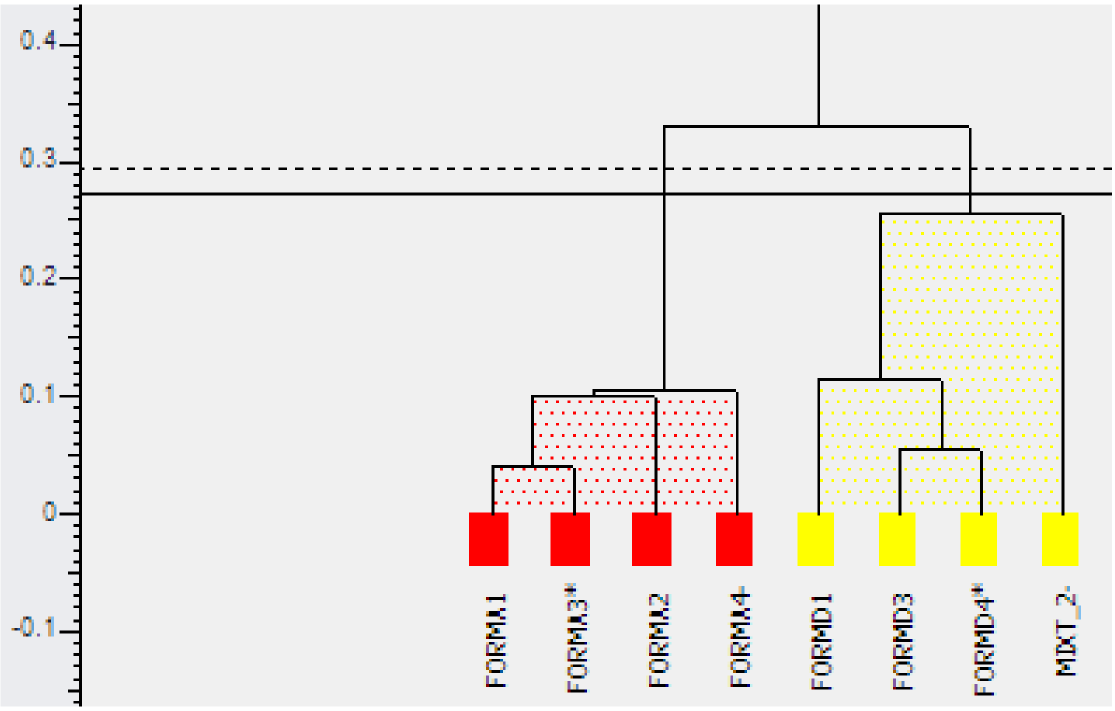


Figure 4.12: Part of the dendrogram with changed cut-level

In the screenshot above, the horizontal cut-point is set at around 0.28, and therefore patterns FORMA1, FORMA3, FORMA2 and FORMA4 are considered to be in one cluster, whereas FORMD1, FORMD3, FORMD4 and MIXT_2 are in a separate and distinct one.

The different clusters are color coded, and if no known phases are present, these dendrogram results are used to generate a pseudo-cell display (see Section *Cell Display View [▶ 35]*).

In this case the colors used here will correspond to the colors in that display. Optionally, the same color-coding can be used to help interpret the results in the various 3D plots.

The **Collapse All Clusters** command, accessed through the right-click menu on the dendrogram node, redraws the dendrogram with only the three most representative patterns (marked with a star) of each cluster shown. This allows easier interpretation of a crowded display when many patterns are being analyzed.

Changing the Cut-level

Cluster analysis is not an exact science, and allocating patterns to clusters can depend on the level of detail you consider relevant. This means that the dendrogram cut line may not be appropriate for the use required. If the program-calculated cut-level is not considered to be optimal, it can be overridden by the user. This is done by activating the **Move Cut-Line** command in the **Dendrogram view's** context menu. The mouse cursor changes into a circle in this mode and the command is marked by a check:

Figure 4.13: The Move Cut-Line command marked with an activity check

By dragging the cut-line with the left mouse button pressed the cut-line can be moved to a new position. The cluster coloring will be updated in the dendrogram as soon as the mouse button has been released. This update is only temporary and does not affect other views.

For example, moving the cut-level up to around 0.4 results in the following:

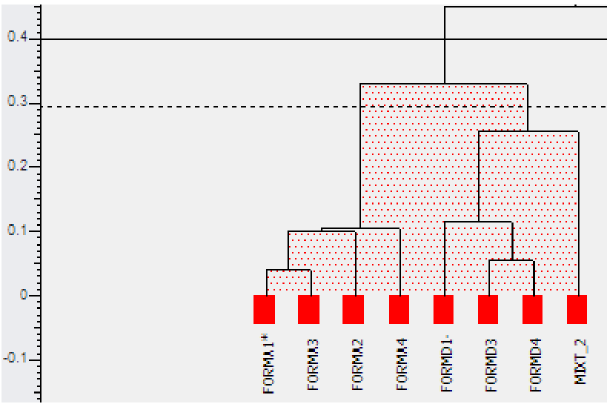
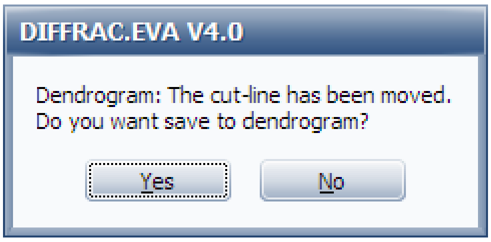


Figure 4.14: Dendrogram after manually moving the cut-line to 0.4

When the **Move Cut-Line** command is selected again, the cut-line will be changed permanently. This may be a lengthy process with large data sets because it involves re-clustering and re-drawing views. For that reason the following message is displayed before the calculation starts:



Dendrogram: The cut-line has been moved.
Do you want save to dendrogram?

If the question is answered with **Yes**, the calculation will start and the cut-line will stay at its new position. Otherwise the old cut-line position is restored and no calculations are carried out.

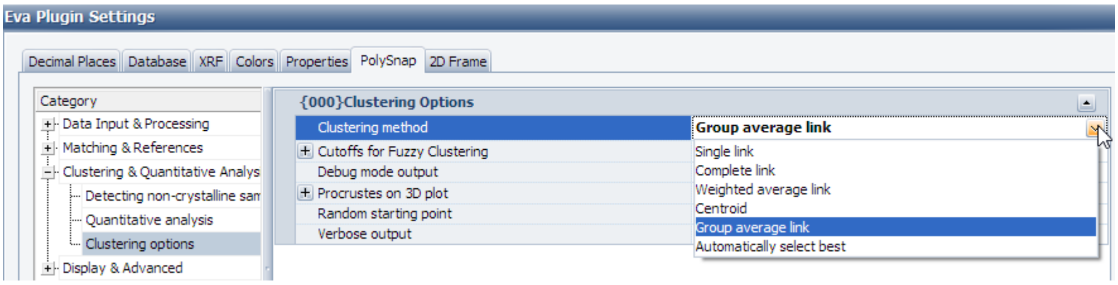
Note that the color assignations have updated accordingly. If the user chooses to retain this change when closing the dendrogram display or switching to another display pane, then the pseudo-cell display will be updated accordingly, and the modification noted in the program log file. If the change is not retained, the previous value is kept.

Changing the Clustering Method

There are numerous methods available for generating dendrograms and these correspond to different clustering methods. The cluster analysis can be set to select what it calculates to be the most appropriate clustering method for a given problem. The user can also choose to re-run just the clustering part of the analysis to experiment with how different individual clustering methods would affect the results.

1. Select **Settings** from **EVA's Help** menu.

⇨ The **EVA Plugin Settings** dialog box will appear:



{000}Clustering Options

Clustering method: Group average link

*Figure 4.15: Selecting the default **Clustering method** in the **Cluster Analysis settings***

1. Select the **Cluster Analysis** tab to display all possible options.

To select an individual cluster method,

1. choose **Clustering options** from the tree on the left and
2. change the **Clustering Method** form the drop down list.
3. Click OK to save the selection.

To The recalculation process may take several minutes on larger data sets.

For further details see reference 2 in the list of references at the end of this section.

Reverting to a previous dendrogram

Because the program retains earlier saved versions of the dendrogram, it is possible to revert to them if required at a later stage. This function is accessed via the **Edit|Undo** menu option.

Selecting the **Undo** command reverts to the most recent version before saving the cut-line. This process can be continued until the original cut-line if the number of undo operations allows it.

Mouse Actions and Keyboard Shortcuts

Table 4.7: Generally available mouse and keyboard commands in the Dendrogram

Mouse Action	Description
Mouse Drag	Zoom In/Out
Shift+Mouse Drag	Rotate
Alt+Mouse Drag	Translate
Click Object	Select
Ctrl+C	Copy Graph
I	Show ID
H	Show help window

Dendrogram View Properties

Table 4.8: Dendrogram view properties

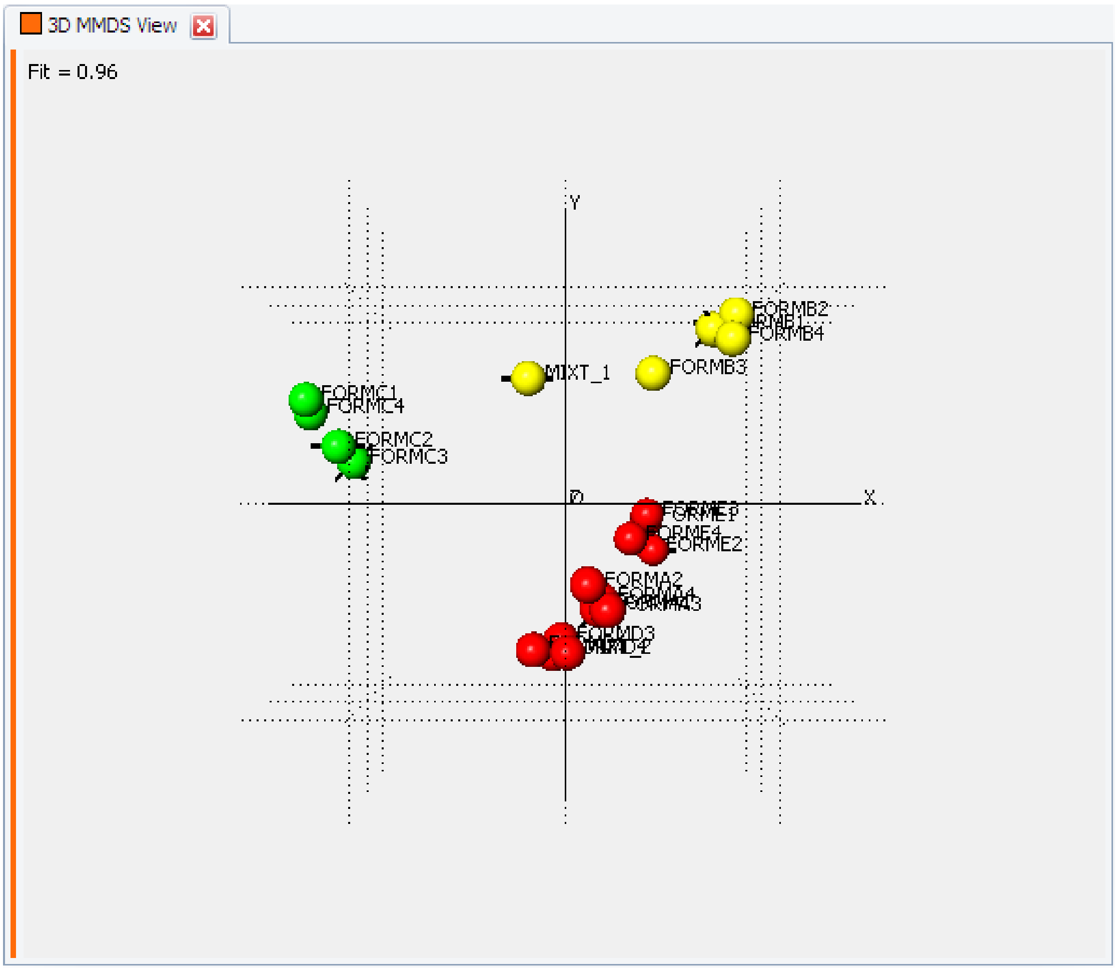
Property	Description
Attributes	
Show Label	Show the labels for all samples
Show Selected Label	Show the labels for selected samples
Max Label Length	Maximum label length defined as the number of characters
Line Width	Outer line width for the cell drawing
Background Color	View's background color
Foreground Color	Text color for label display
Label Font	Font for label display
Auto Save Cut-Line	Whether to auto save (including saving the cut-line position, re-clustering, reload views and exiting moving mode *etc.*), immediately after the cut-line has been moved

4.3.3 3D MMDS View

The **3D MMDS** view displays the results of metric multidimensional scaling which makes use of distances between objects calculated from the correlation matrix generated by the matching process to produce a three-dimensional spatial representation of the samples. Each point which appears on this spatial representation corresponds to one of the patterns. The closer two points appear on the plot, the more similar the patterns are, and the more different an object is to another the further apart they will appear. Therefore groups of similar patterns appear to cluster together. The samples are placed in a box of unit dimensions; the orientation of the spheres in this box is arbitrary.

The multidimensional scaling performed is based on calculated proximities rather than observations. First for n patterns, we generate an (*n x n*) distance matrix **D** based on dissimilarities, *δrs= 1, 2,...., n*, computed from the correlation matrix ρ, of the individual patterns. Each object is compared against itself and every other object. The result of an object being paired against itself gives a dissimilarity of zero, which corresponds to the diagonal of the matrix. The goal of this method is to derive a set of underlying dimensions, with co-ordinates that should create a Euclidean distance matrix, which in turn should be the same or very close to the *δrs* of the original **D.**

The initial view of the **3D** plot shows only the X and Y axis - the Z axis lies in projection.



Fit = 0.96

*Figure 4.16: **3D MMDS** view*

The individual points plotted on either of the **3D** plots are colored to correspond to the color groupings shown in the **Dendrogram** plot.

Changing the dendrogram cut-level causes the colors in the **3D** plots to be updated. A small numerical label in the top-left corner of the display gives an indication as to the goodness of fit of these results. This parameter is computed as a correlation coefficient (based on both the Pearson and Spearman correlation coefficients) between **t**he observed and calculated **D** matrices. Normally one sees values > 0.8; low values can indicate an inability to match the two matrices using three dimensions. This is rare, even with 1000 patterns, although there is a reduction in the overall correlation coefficient as *n* increases.

The orientation of the **3D** plot can be altered by holding down the **Shift** key and dragging the mouse in any direction as desired; the plot rotates as shown below:

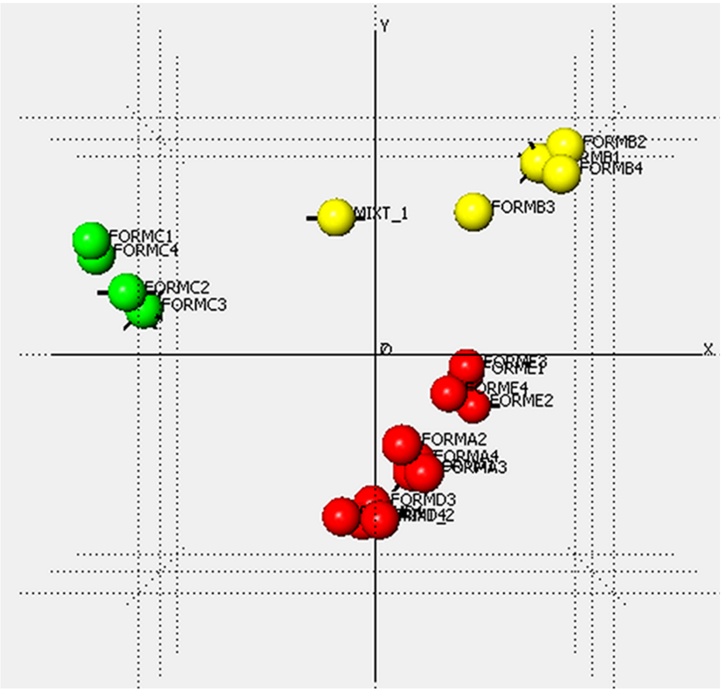


Figure 4.17: Default coordinate system orientation for 3D views

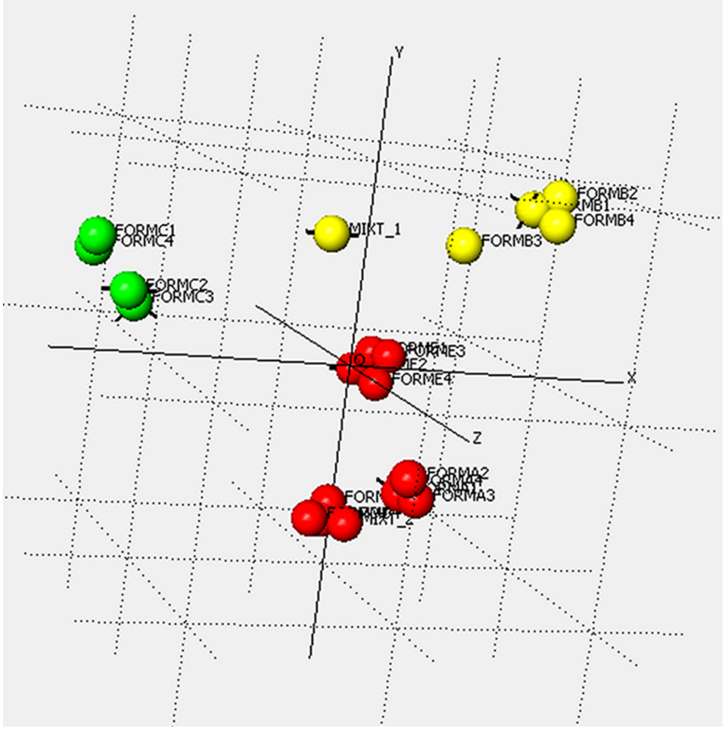


Figure 4.18: Changed coordinate system orientation in 3D views

In the above plots it can be seen that there are 4 or 5 clusters depending on P16 and whether it can be considered to belong to an adjacent cluster or not.

A variety of views can be used to gain a better understanding of the distribution of the pattern data points in the three-dimensional space.

As in the other graphics screens, any particular area of the **3D** view can be zoomed in by dragging a rectangle over the relevant region.

The pattern corresponding to each display point can be selected by clicking on it, this in turn updates the pattern information display in the lower portion of the window.

The points representing the patterns can be enlarged or shrunk to suit any zoom level by holding down the **Ctrl** key and moving the mouse either up or down. An upward movement will reduce the size of the spheres, a downward movement will increase the size.

The **3D** plot position itself can be translated by holding down the Alt key and then moving the mouse in any direction as required.

The drawing quality of the spheres can be altered if needed – with many points plotted, working with the display can be much faster if the rendering quality is reduced. This is especially the case with lower-powered graphics cards. To do this,

1. click on the **3D MMDS** view in the tree and
2. change the **Rendering Quality** value in the properties.

The most representative member (MRM) of a cluster is made by four spikes protruding from the sphere:

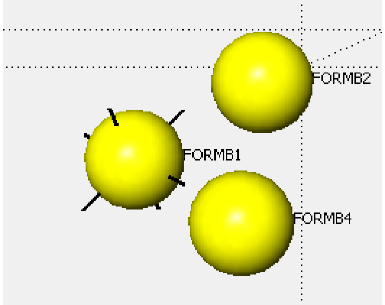


Figure 4.19: Most representative member (MRM) of a cluster

The least representative member of a cluster (LRM) is marked by a couple of spikes:

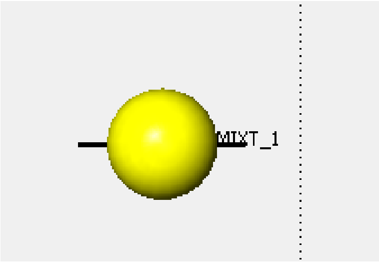


Figure 4.20: Least representative member (LRM) of a cluster

Further options can be accessed using the view’s property grid.

3D MMDS View Properties

Table 4.9: 3D MMDS view properties

Property	Description
Attributes	
Show Label	The labels which appear next to the plotted points can be turned on or off with this option. This may aid in seeing an overall clustering pattern when the display is crowded with many spheres.
Show Selected Label	Show the labels for selected samples

Property	Description
Max Label Length	Maximum label length defined by the number of characters. Useful if many samples are displayed.
Show Axes	The three axes x, y, and z are drawn in the graph
Show Grid	The grid which appears in the 3D plot can be hidden or displayed
Line Width	The axis line width
Background Color	View's background color
Foreground Color	Text color for label display
Axis Color	The color for axes and axis labels
Label Font	The font for label display
3D Attributes	
Show Top View	This option brings up a small simplified overview of the plot in the lower right hand corner. It can be useful for orientating yourself when zoomed into the display.
Rendering Quality	Controls the graphics drawing quality. Zero shows no spheres, one is the lowest quality and three is the default.
Light Position	Controls where the spheres are lighted from. It can be changed to achieve a different illumination effect.
Render As Dots and Show As Transparent	Alters the way the spheres are plotted as shown in the diagrams below. This can be useful to identify if for example, a single pattern of one color is hidden within a group of another patterns.
Transparency	The value controls the transparency of the balls. Zero means no transparency while larger numbers up to nine increase the transparency.
Grand Tour	
Grand Tour	The display can be animated for easier interpretation.

Mouse Actions and Keyboard Shortcuts

Table 4.10: Mouse and keyboard commands in the MMDS view

Mouse Action	Description
Mouse Drag	Zoom In/Out
Shift+Mouse Drag	Rotate
Alt+Mouse Drag	Translate
Click Object	Select

Mouse Action	Description
Ctrl+C	Copy Graph
I	Show ID
H	Show help window

4.3.4 3D PCA View

The **3D PCA** view is based on the results from principle component analysis of the modified correlation matrix. The sphere coordinates are derived from the eigenvectors of the correlation matrix ρ that correlates all the input patterns. The use and interaction options available for this plot are identical to those for the **3D** Plot (MMDS) described in the previous Section.

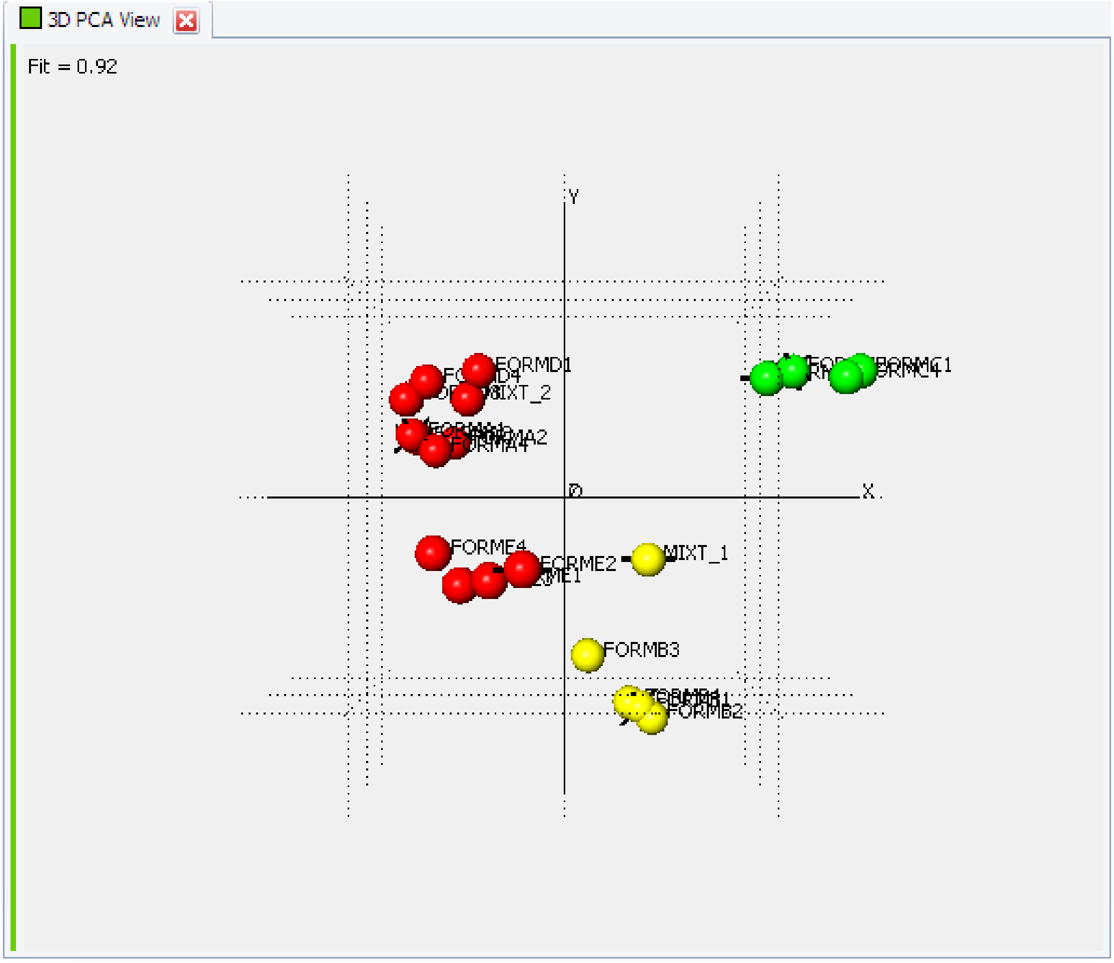

Fit = 0.92

Figure 4.21: 3D PCA view

The properties of the **3D PCA** view are exactly the same as for the **3D MMDS** view and are described there. The same is valid for the available mouse and keyboard actions.

In general the **PCA** method is less powerful than the **MMDS** method and it will have a lower fit than the **MMDS** results, but it can be useful with certain data.

4.3.5 6D Plot View

The cluster analysis provides the facility to allow either of the two standard three-dimensional plots discussed earlier to be augmented by up to three additional user-specified dimensions, as described here.

These additional dimensions are used to represent information recorded about each sample regarding its method of preparation. Available information that can be plotted is for example (assuming it is available for the measurements):

- Mass
- Total Volume
- Counterion
- Stirrer Rate
- Sample Presentation
- Solvent
- Antisolvent
- Initial Temperature
- Isolation Temperature
- Cooling Rate
- Heating Rate
- Reaction Time
- Antisolvent Volume

The information to be used in the **6D** plot is not limited to the above mentioned properties, but can be configured freely in an external file.

The 3 additional plotting dimensions that are available to represent these data are:

- Point Size
- Point Shape
- Point Color

By using this approach, it can be possible to discover if there is a connection between a particular combination of sample preparation conditions and the resulting clusters (note however that there are some restrictions as to which fields may be plotted as which dimension).

To take advantage of this feature, use the **6D Plot** view. The following special properties will be available.

Special Properties of the 6D Plot

Use MMDS

The MMDS plot will be used if checked, otherwise the **PCA** plot will be employed.

Size

The default possible information types available to be plotted as the size dimension are as mentioned above, *e.g.*: Mass, Total Volume, *etc.* The **size** dimension is useful for visualizing numeric values.

When each pattern was imported, any sample information fields were parsed, and the maximum and minimum values of each field stored.

This data is then used to scale each data point to an appropriate size.

Color

Any of the above mentioned properties can be visualized as color. A maximum of 15 different colors can be displayed.

Additionally, the selection **Color** allows the color scheme of the dendrogram at the currently saved cut-level to be applied to the plot.

Shape

This is provided as a means of plotting the more textually orientated information fields, such as solvent and counterion. Up to six different shapes are available (sphere, cone, cylinder, tetrahedron, octahedron and cube), allowing for up to six different solvents, for example, to be represented on the plot.

Once the settings are complete, the **Apply** button must be clicked. The generated plot is then shown:

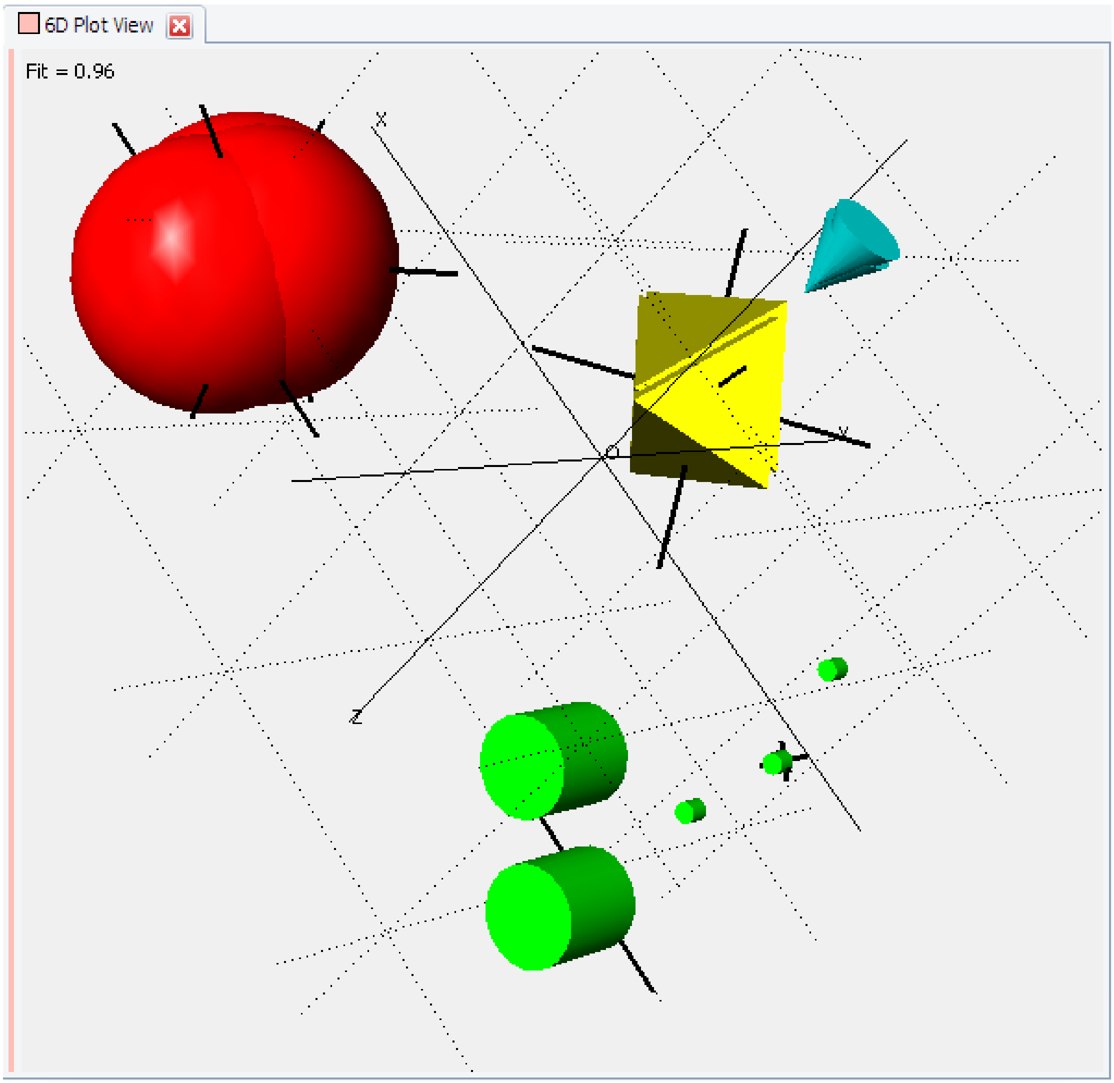


*Figure 4.22: **6D Plot** View displaying different sample sizes and shapes for reaction time and solvent*

The standard graphics controls for moving, zooming and rotating the display etc. all apply as normal, as do the controls for selecting individual patterns of interest, and displaying their details in the pattern information pane.

The settings used to produce the plot are saved to the log file for future reference.

Editing Sample Preparation Information Details and Location

The sample preparation details must be provided in a text file. There is an example file located after installation in the tutorial folder (C:\ProgramData\Bruker AXS\Tutorial\Cluster Analysis\6d-Plot) called *sample_6dinfo_format.txt*.

The location of this file can be altered by changing the option in the program **Settings** dialog, **Cluster Analysis** tab, category **Data Input and Processing**.

Editing this file enables the user to change the field names, types and order expected for information relevant to the **6D** plot.

The user must also provide the additional **6D** plot information for all samples in this file (1 line per pattern). These should be in the same order as the patterns are loaded into the program; line 1 should correspond to Sample 1, and so on.

The following example illustrates how to define the properties:

<HEADER_NAME>; SampleID; Mass; Reactor Identity; Total Volume; Solvent

<HEADER_TYPE>; TEXT; NUMBER; TEXT; NUMBER; TEXT

The corresponding lines per sample with the respective property values look like:

05099707;60;11A;1;Ethanol;

Mouse Actions and Keyboard Shortcuts in the 6D Plot View

Table 4.11: Mouse and keyboard commands

Mouse Action	Description
Mouse Drag	Zoom In/Out
Shift+Mouse Drag	Rotate
Alt+Mouse Drag	Translate
Click Object	Select
Ctrl+C	Copy Graph
I	Show ID
H	Show help window

6D Plot View Properties

Table 4.12: 6D Plot View properties

Property	Description
6D Extra Information	
Use MMDS	The MMDS plot will be used if checked, otherwise the PCA plot
Size	Information type to be plotted as the size dimension.
Color	Information type to be plotted as the color dimension.
Shape	Information type to be plotted as the shape dimension.
Action	

Property	Description
Simplified Mode	Whether to present data in a simplified way. Only the most representative members of a cluster are displayed.
Mask Color	Choose the color for masking.
Mask Color Mode	Whether and how to mask objects with selected color.
Attributes	
Show Label	The labels which appear next to the plotted points can be turned on or off with this option. This may aid in seeing an overall clustering pattern when the display is crowded with many data sets.
Show Selected Label	Show label for selected samples.
Max Label Length	Maximum label length as the number of characters. Useful if many samples are displayed.
Show Axes	The three axes x, y, and z are drawn in the graph.
Show Grid	The grid which appears in the 3D plot can be hidden or displayed.
Line Width	The axis line width.
Background Color	View’s background color.
Foreground Color	Text color for label display.
Axis Color	The color for axes and axis labels.
Label Font	The font for label display.
3D Attributes	
Show Top View	This option brings up a small simplified overview of the plot in the lower right hand corner. It can be useful for orientating yourself when zoomed into the display.
Rendering Quality	Controls the graphics drawing quality. Zero shows no spheres, one is the lowest quality and three is the default.
Light Position	Controls where the spheres are lighted from. It can be changed to achieve a different illumination effect.
Render As Dots and Show As Transparent	Alter the way the spheres are plotted as shown in the diagrams below. This can be useful to identify if for example, a single pattern of one color is hidden within a group of another patterns.
Transparency	The value controls the transparency of the balls. Zero means no transparency while larger numbers up to nine increase the transparency.
Grand Tour	

Property	Description
Grand Tour	The display can be animated for easier interpretation.

4.4 Validation Views

Cluster analysis in EVA provides several techniques for validating the clustering which can be very useful with difficult data sets. The validation views include the following six types, each with a different method of validating the data.

4.4.1 Scree Plot View

The **Scree plot** is a 2 dimensional graph. Along the x-axis is the Eigenvalue Number and the y-axis is made up from the eigenvalue itself.

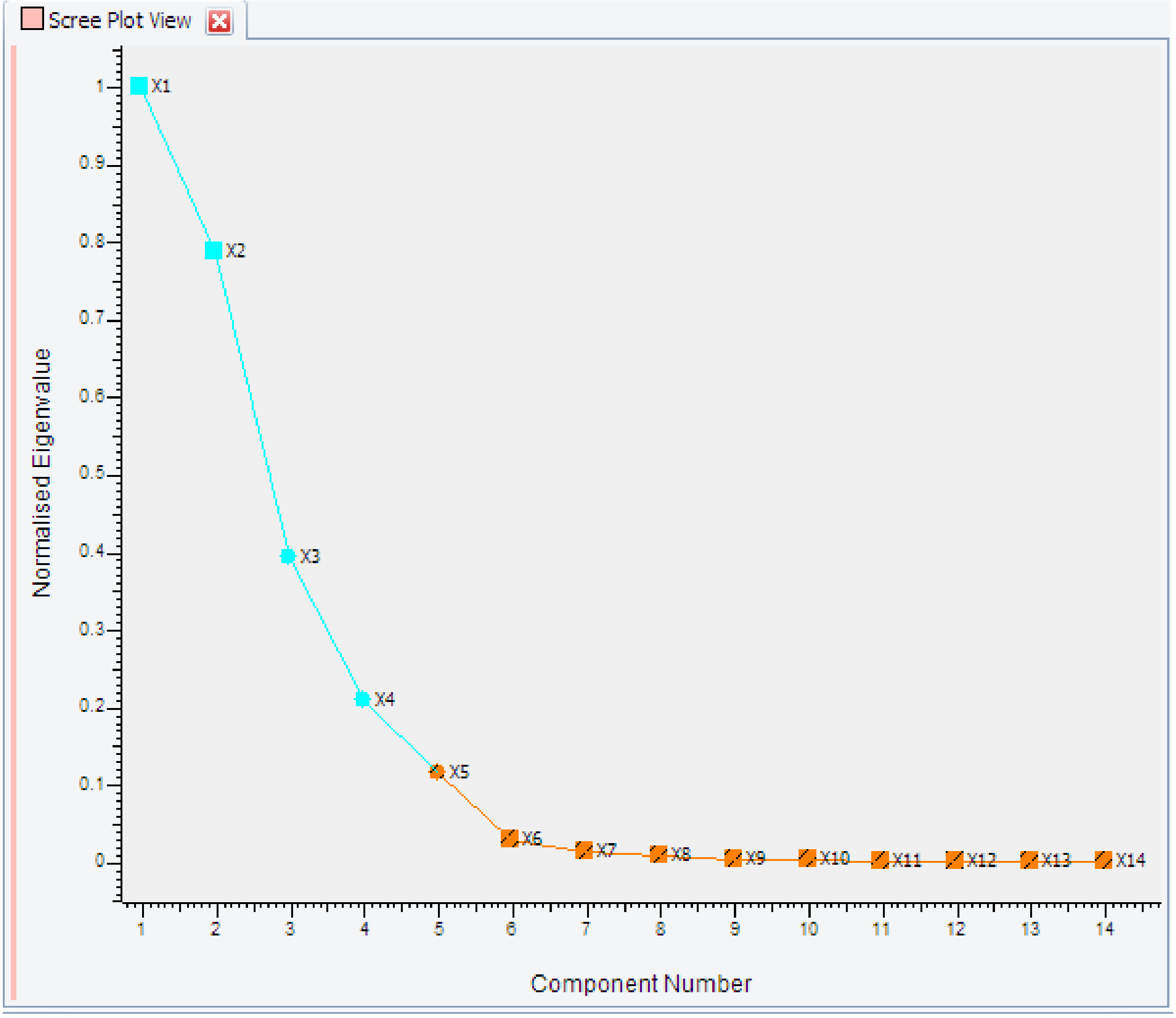


Figure 4.23: Scree Plot view

Eigenvalues are derived from the standardized correlation matrix. The eigenvalues are sorted in descending order.

What this represents is the minimum number of clusters that can be used to describe the entire data set being examined. The point where the plotted line changes color - e.g. between 5 and 6 in the example above, suggests that just 5 clusters are needed to explain over 95% of the variation in the data. A well behaved scree plot should have a reasonably steep initial descent. A gradual, sloping descent indicates difficulty in establishing the number of clusters required, so the program-generated dendrogram cut-level should be examined especially closely.

The plot can be zoomed in by dragging a rectangle over the relevant area.

Scree Plot View Properties

Table 4.13: Scree Plot view properties

Property	Description
Attributes	
Show Label	The labels which appear next to the plotted points can be turned on or off with this option. This may aid in seeing an overall clustering pattern when the display is crowded with many points.
Show Selected Label	Show label for selected samples.
Max Label Length	Maximum label length defined as the number of characters. Useful if many samples are displayed.
Line Width	The axis line width.
Background Color	View’s background color.
Foreground Color	Text color for label display.
Label Font	The font for label display.

4.4.2 Minimum Spanning Tree (MST) View

The **Minimum Spanning Tree** (MST) view presents a different way of partitioning the patterns into different clusters. All of the patterns are initially joined together by a single line, in the order of increasing distance between them.

The initial view shows the samples connected in this manner:

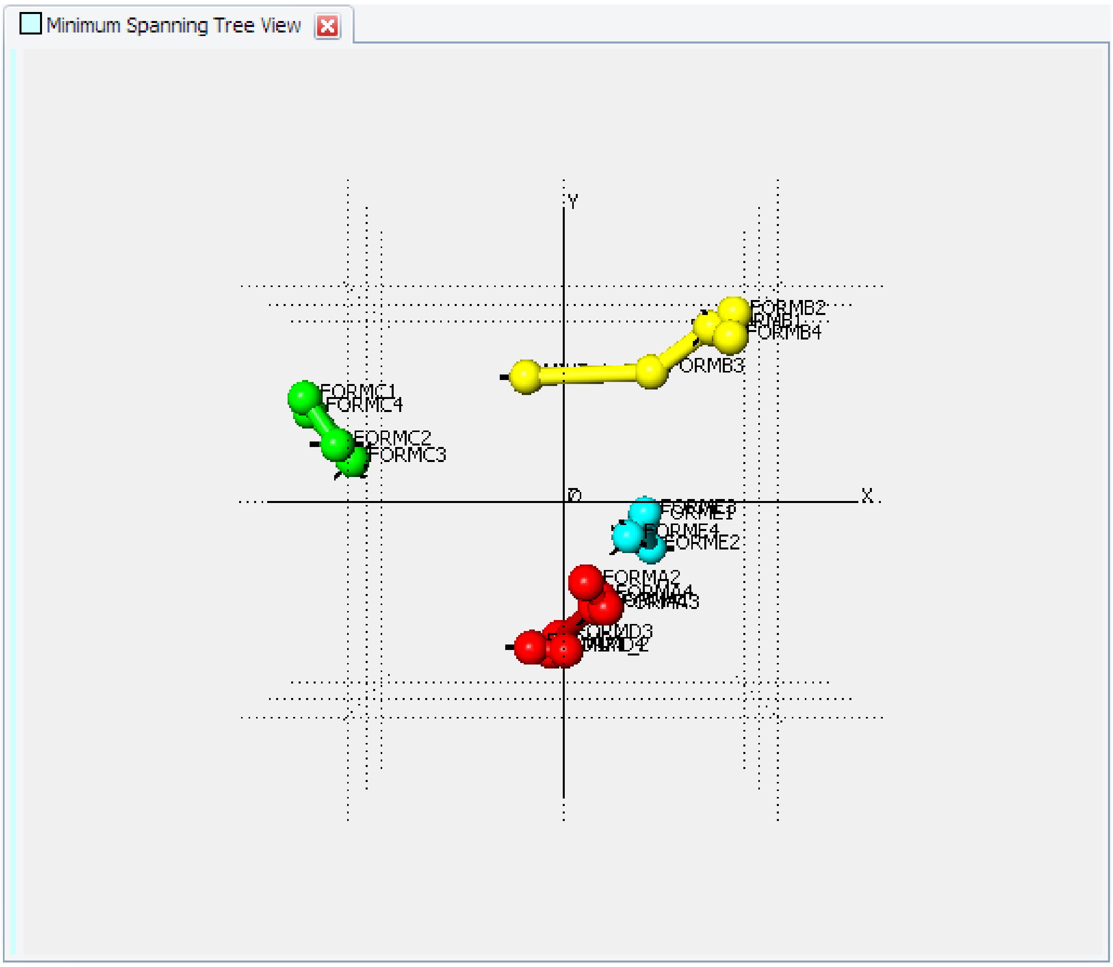


*Figure 4.24: **Minimum Spanning Tree** (MST) view*

The program then cuts links between the patterns in order of decreasing maximum distance until the estimated number of clusters is reached. The user can then manually adjust this by clicking the **Link Decrement** command to reduce the number of links by one, or by clicking the Link Increment command to increase the number of links by one:

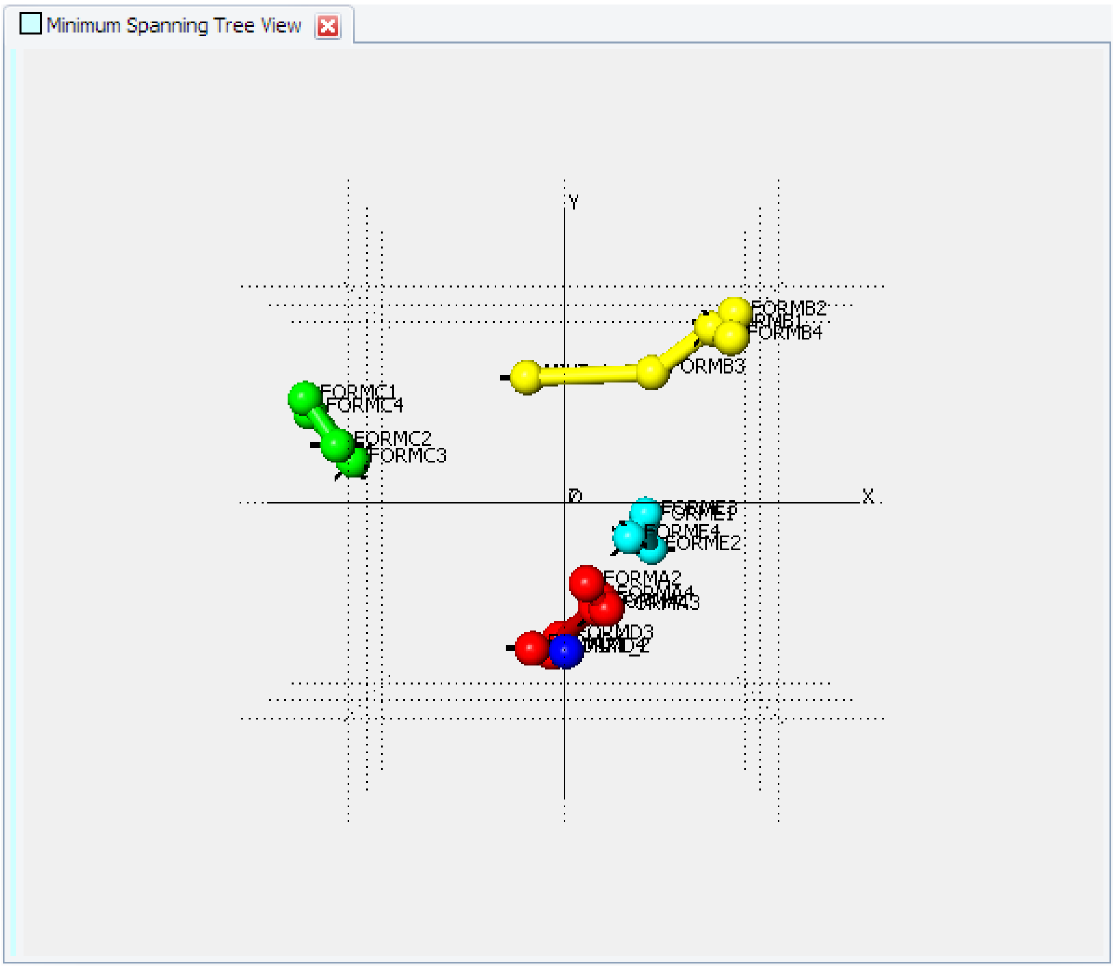


*Figure 4.25: **Minimum Spanning Tree** after clicking the **Link Decrement** command*

This process can be repeated, cutting the next smallest distance, and creating an additional cluster each time. The reverse is also true: links can be added until all patterns are linked.

The display can be zoomed, rotated and otherwise manipulated in a similar manner to the other **3D** views.

Note that each time the number of clusters changes the most representative patterns in each are recalculated.

Note that the colors here do **not** correspond to those on the dendrogram.

Mouse Actions and Keyboard Shortcuts in the MST view

Table 4.14: Mouse and keyboard commands

Mouse Action	Description
Mouse Drag	Zoom In/Out
Shift+Mouse Drag	Rotate
Alt+Mouse Drag	Translate
Click Object	Select
Ctrl+C	Copy Graph
I	Show ID

Mouse Action	Description
H	Show help window

Minimum Spanning Tree View Properties

Table 4.15: Minimum Spanning Tree View Properties

Property	Description
Action	
Links	
Attributes	
Show Label	The labels which appear next to the plotted points can be turned on or off with this option. This may aid in seeing an overall clustering pattern when the display is crowded with many points.
Show Selected Label	Show label for selected samples.
Show Axes	The three axes x, y, and z are drawn in the graph.
Show Grid	The grid which appears in the 3D plot can be hidden or displayed.
Max Label Length	Maximum label length defined as the number of characters. Useful if many samples are displayed.
Line Width	The axis line width.
Background Color	View’s background color.
Foreground Color	Text color for label display.
Axis Color	The color for axes and axis labels.
Label Font	The font for label display.
3D Attributes	
Show Top View	This option brings up a small simplified overview of the plot in the lower right hand corner. It can be useful for orientating yourself when zoomed into the display.
Rendering Quality	Controls the graphics drawing quality. Zero shows no spheres, one is the lowest quality and three is the default.
Light Position	Controls where the spheres are lighted from. It can be changed to achieve a different illumination effect.
Render As Dots and Show As Transparent	Alter the way the spheres are plotted as shown in the diagrams below. This can be useful to identify if for example, a single pattern of one color is hidden within a group of another patterns.

Property	Description
Transparency	The value controls the transparency of the balls. Zero means no transparency while larger numbers up to nine increase the transparency.
Grand Tour	
Grand Tour	The display can be animated for easier interpretation.

4.4.3 Silhouettes View

For each of the current clusters as defined by the dendrogram cut-level, this display shows a histogram. Silhouettes (Rousseeuw, P.J. (1987). *J. Computation & Appl. Math.*, **20**, 53-65.) provide an alternative formalism for assessing the compactness and isolation of clusters, and also for identifying those members of a given cluster which are well established members of the core cluster or outlying, and thus potentially problematic. If the *i-th* pattern belongs to cluster *Cr* then we define the silhouette, *s(i)* as follows:

$$a(i) = \frac{\sum'_{j \in Cr} \mathrm{d_{ij}}}{n_r - 1}$$

$$b(i) = min_{s \neq 0}\left(\frac{\sum_{j \in C_s} d_{ij}}{n_s}\right)$$

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where there are n_r patterns in cluster *r,* n_s patterns in cluster *s*, and d_{ij} is the distance between patterns *i* and *j*. The values of s(i) lie between –1 and +1. The lower the value of the silhouette, the more likely it is that the pattern either:

Belongs to a different cluster and you should try altering the dendrogram cut level, or

The pattern is a mixture. The **MMDS** and **PCA plots** will be useful here to identify the closest neighbors of pattern *i*. One histogram is shown for each cluster.

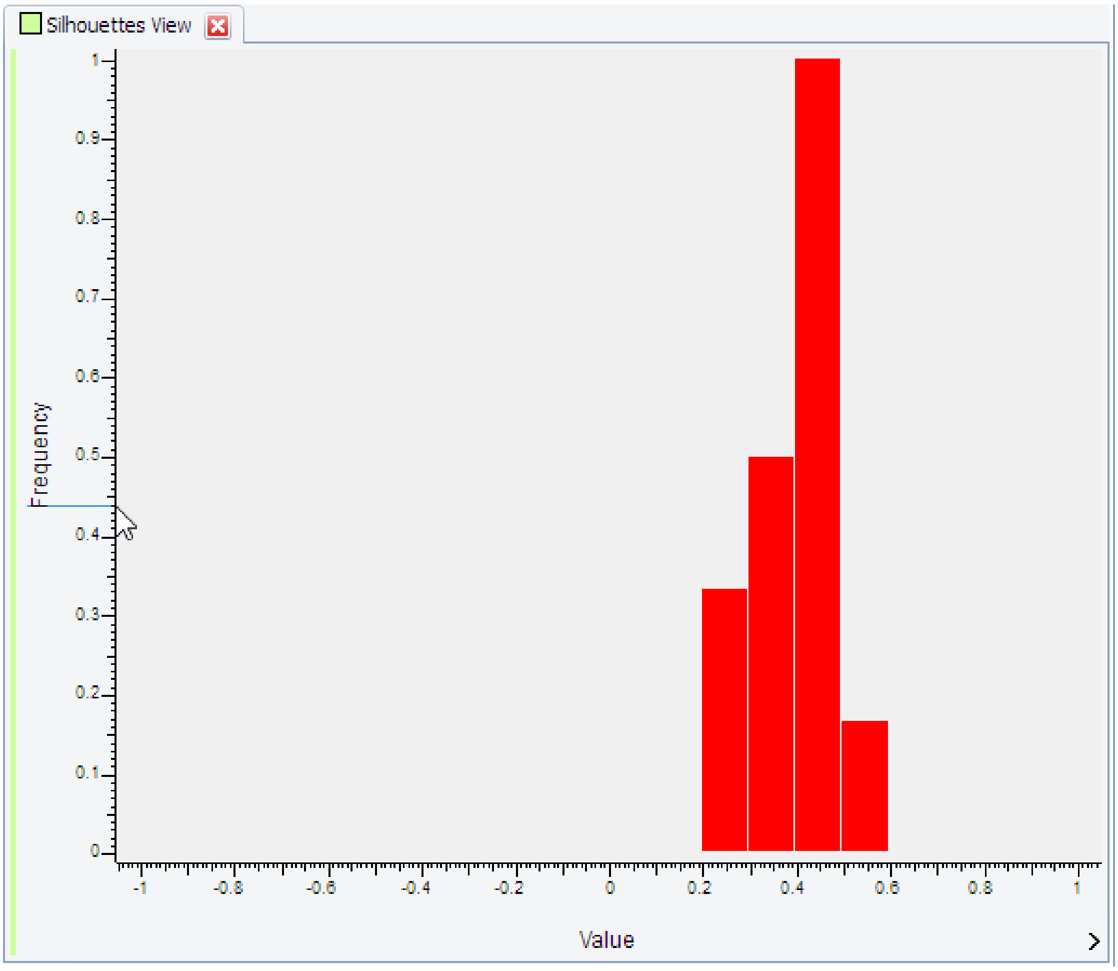


Figure 4.26: Silhouettes view

To move to the next or previous histogram,

1. click on the view's **Next Cluster** or **Previous Cluster** command, or use the scroll wheel on the mouse while pressing the **Alt** key.

Allowing the mouse to hover over a particular column shows which pattern numbers correspond to that score. For example, the yellow cluster appears to have a member that has a reasonably low membership score - moving the mouse over it shows that this is member MIXT_1, which makes sense when looking at the dendrogram, where it is the least tightly linked member:

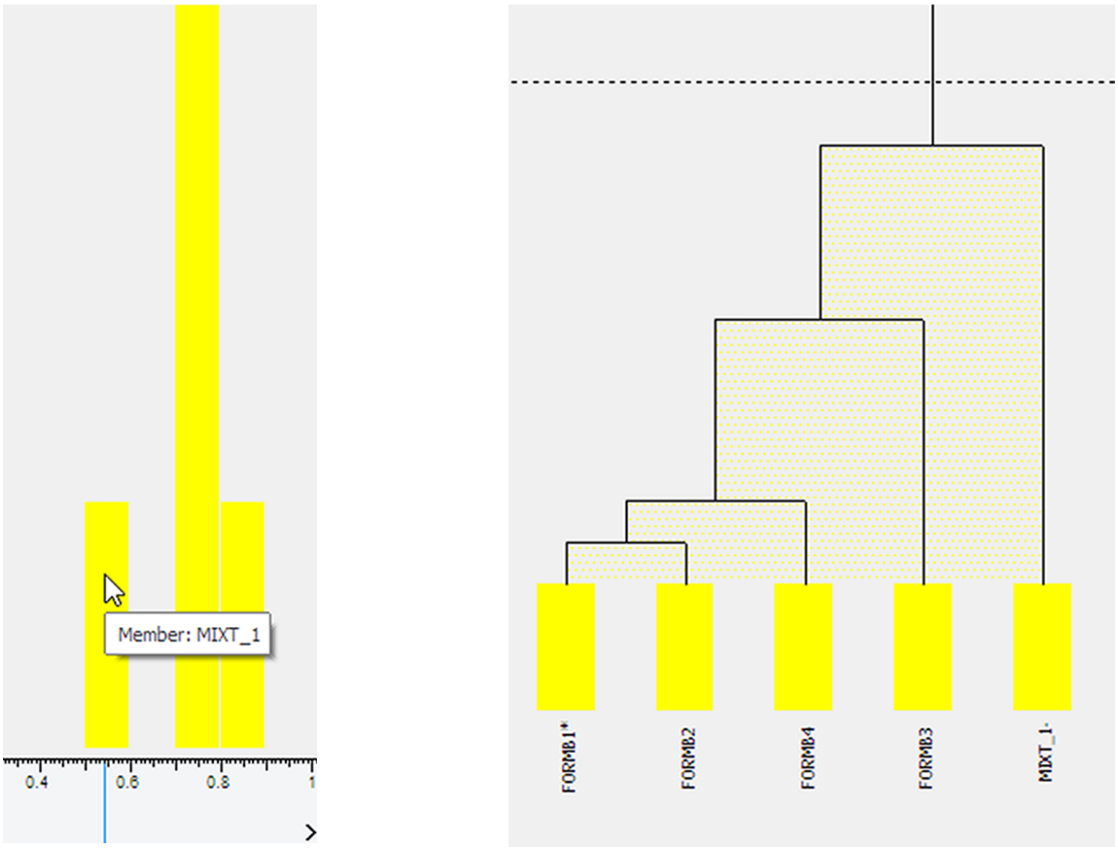


Figure 4.27: Displaying the histogram membership in a tooltip and referencing it to the ***Dendrogram*** *view*

Note that if the dendrogram cut-level has been manually adjusted since the previous time this display was accessed, there may be a short delay as new silhouettes are calculated for the revised cluster membership list. Clicking on a column selects the patterns presented in that column, allowing them to later be located in a **3D** or the **dendrogram** view.

Mouse Actions and Keyboard Shortcuts in the Silhouette view

Table 4.16: Mouse and keyboard commands

Mouse Action	Description
Mouse Drag	Zoom In/Out
Shift+Mouse Drag	Rotate
Alt+Mouse Drag	Translate
Click Object	Select
Ctrl+C	Copy Graph
I	Show ID
H	Show help window

Silhouettes View Properties

Table 4.17: Silhouettes View Properties

Property	Description
Attributes	
Background Color	View’s background color

4.4.4 Fuzzy Clustering View

This view is superficially similar to the silhouettes, and is navigated in the same manner. With fuzzy clustering, and unlike the standard clustering methods used elsewhere in the program, a single pattern can be assigned to more than one cluster. Using the concept of membership, a value can be calculated for how well a given pattern fits in a given cluster. A low membership score may suggest that that pattern either does not belong in that cluster, or that it is a mixture. A single pattern having reasonably high scores in more than one cluster may indicate a mixture. Only patterns which fall outside a defined threshold range, and therefore are samples that may need to be examined manually in more detail, are shown in the fuzzy clustering output. If this is not the case for any of the patterns, then no output is shown - this is the ideal case, as all patterns clearly belong to only one cluster.

Checking the output from this calculation should therefore help to highlight borderline or unusual cases that do not fit neatly into one cluster or another. Such examples may possibly be mixtures, or evidence of some other problem such as a different background or a *2θ*-shift.

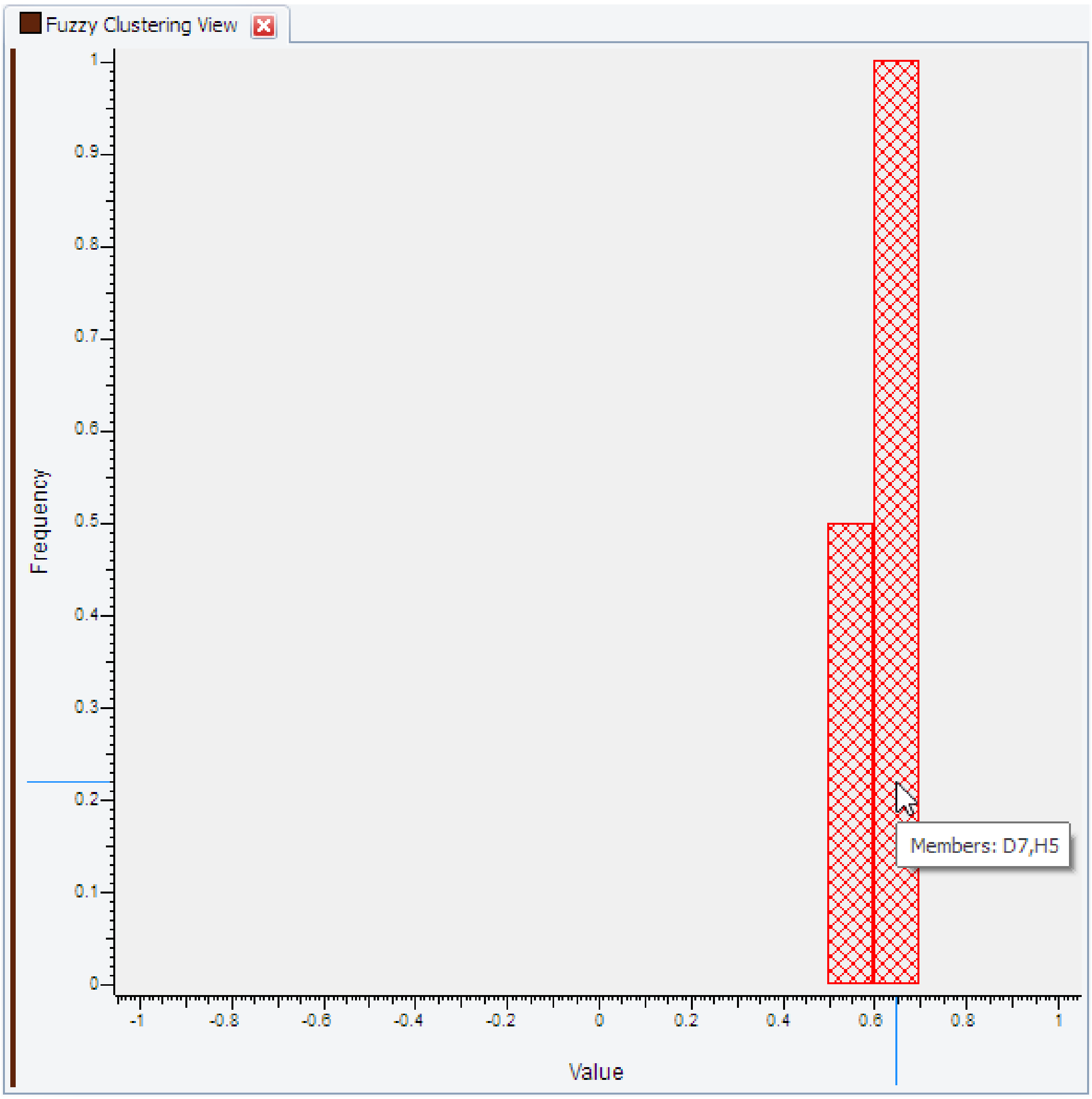


Figure 4.28: ***Fuzzy Clustering*** *view*

Every pattern with a problematic fuzzy clustering membership is represented by a rectangular box which has the color of the cluster it belongs to. Patterns which are non-problematic are not displayed. If one or more scans have been selected the box height is increased and the scan's 1D representation is displayed below the diagram.

Detailed numeric output from this method is available in the log file view, e.g.:

121 0.37 0.15	Pattern belongs to cluster 1 but has small memberships for all clusters.
122 0.05 0.76	Pattern belongs to cluster 2 and is OK.
123 0.00 0.76	Pattern belongs to cluster 2 and is OK.
124 0.13 0.16	Pattern belongs to cluster 1 but has small memberships for all clusters.
125 0.38 0.00	Pattern belongs to cluster 1 but has small memberships for all clusters.
126 0.00 0.00	Pattern belongs to cluster 1 but has small memberships for all clusters.

127 0.43 0.12	Pattern belongs to cluster 1 but has small memberships for all clusters.
128 0.39 0.02	Pattern belongs to cluster 1 but has small memberships for all clusters.
129 0.14 0.14	Pattern belongs to cluster 1 but has small memberships for all clusters.
130 0.11 0.12	Pattern belongs to cluster 1 but has small memberships for all clusters.
131 0.28 0.65	Pattern belongs to cluster 1 but membership of that cluster is < 0.5 and membership of another cluster is > 0.5.
132 0.29 0.65	Pattern belongs to cluster 1 but membership of that cluster is < 0.5 and membership of another cluster is > 0.5.
133 0.09 0.18	Pattern belongs to cluster 1 but has small memberships for all clusters.
134 0.19 0.15	Pattern belongs to cluster 1 but has small memberships for all clusters.

Pattern selection features as described above for the **Silhouettes** view are also available in this view.

The method becomes less useful as the number of patterns increases, and when there are more than 100 patterns the technique is not used.

Fuzzy Clustering View Properties

Table 4.18: Fuzzy Clustering view properties

Property	Description
Attributes	
Line Width	Line width in pixels
Background Color	View’s background color

4.4.5 Parallel Coordinates Plot View

So far we have made the assumption that we can work in 3 dimensions with the **MMDS** and **PCA** plots. EVA provides techniques for testing this hypothesis in the form of **Parallel Coordinate Plots**, the **Grand Tour** and the **Space Explorer**. These will now be described in detail.

Rather than plotting the patterns in a 3D space, here the first 6 dimensions of the clusters are drawn in a linear fashion, allowing the user to see if the cluster separations evident in the first 3 dimensions still hold together in higher dimensional space. Whereas the first three dimensions are plotted as x, y and z axes on a 3D plot here they are plotted as the first, second and third vertical axes on the plot, with the fourth, fifth and sixth dimensions plotted above. (For more information about this invaluable method of visualization in many dimensions see *‘Parallel Coordinates’ by A. Inselberg, published by Springer, 2008*.)

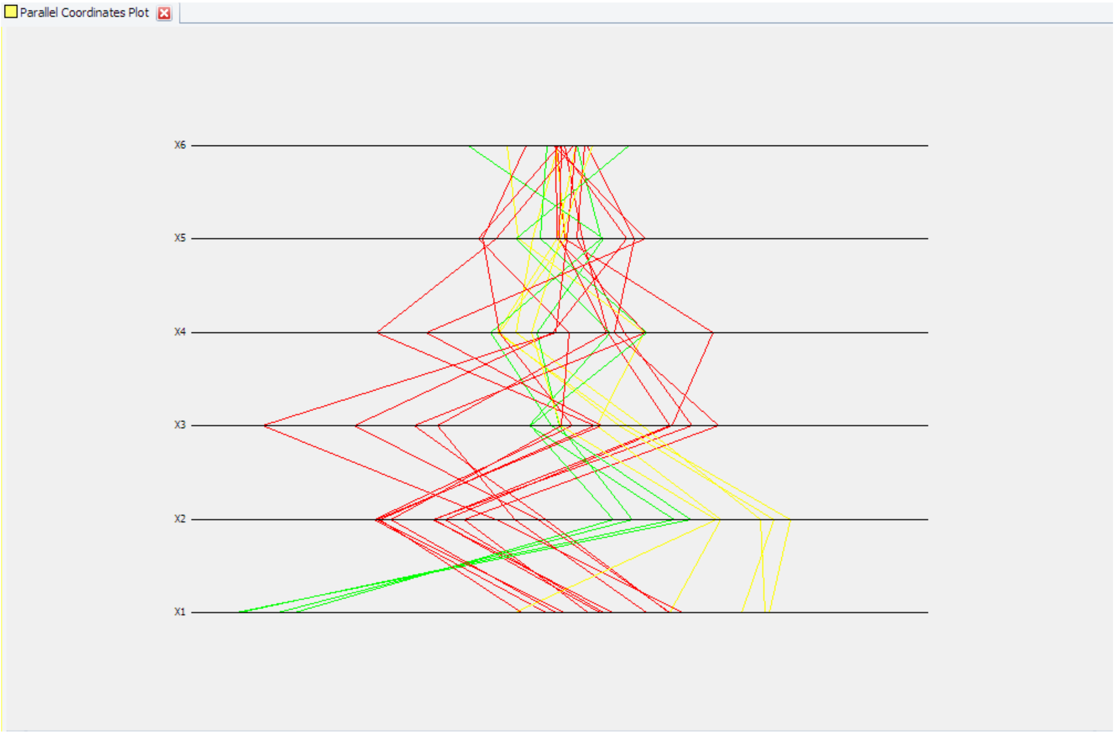


*Figure 4.29: **Parallel Coordinates Plot** view*

While the three axes on the **3D** plot are usually displayed orthogonal to each other, they are arranged horizontally parallel to each other in this plot. As in the **3D** plot each pattern is given a value (between one and zero) for each dimension representing the associated coordinate from the MMDS calculation extended to 6 dimensions, and this is plotted for each axis. While in 3D space this becomes a data point represented by a sphere, in the parallel coordinates plot this becomes a line joining up the different values for an object as the value varies from dimension to dimension.

The colors are taken from the Dendrogram so it is easy to identify the separate clusters. By selecting the **Grand Tour** in the properties grid, the display can be animated by rotating round the 6 different axes simultaneously.

Parallel Coordinates Plot View Properties

Table 4.19: Parallel Coordinates Plot view properties

Property	Description
Attributes	
Line Width	The line width in pixels.
Background Color	View’s background color.
Foreground Color	Text color for label display.
Axis Color	The color for axes and axis labels.
Label Font	The font for label display.
Grand Tour	
Grand Tour	The display can be animated for easier interpretation.

Mouse Actions and Keyboard Shortcuts in the Parallel Coordinates Plot

Table 4.20: Mouse and keyboard commands

Mouse Action	Description
Mouse Drag	Zoom In/Out
Click Object	Select
Ctrl+C	Copy Graph
I	Show ID

4.4.6 Space Explorer View

This view resembles the standard **3D** plot, with the exception of some additional commands.

Like the parallel coordinates view it allows analysis of higher dimensions of data to be viewed. However, here they can be visualized on orthogonal axes. For this reason only three dimensions can be displayed at the one time. The display opens with the dimensions 1, 2 and 3 plotted as in a normal 3D plot. The extra commands **Next Axis Combination and Last Axis Combination** allow the user to change the axes being drawn.

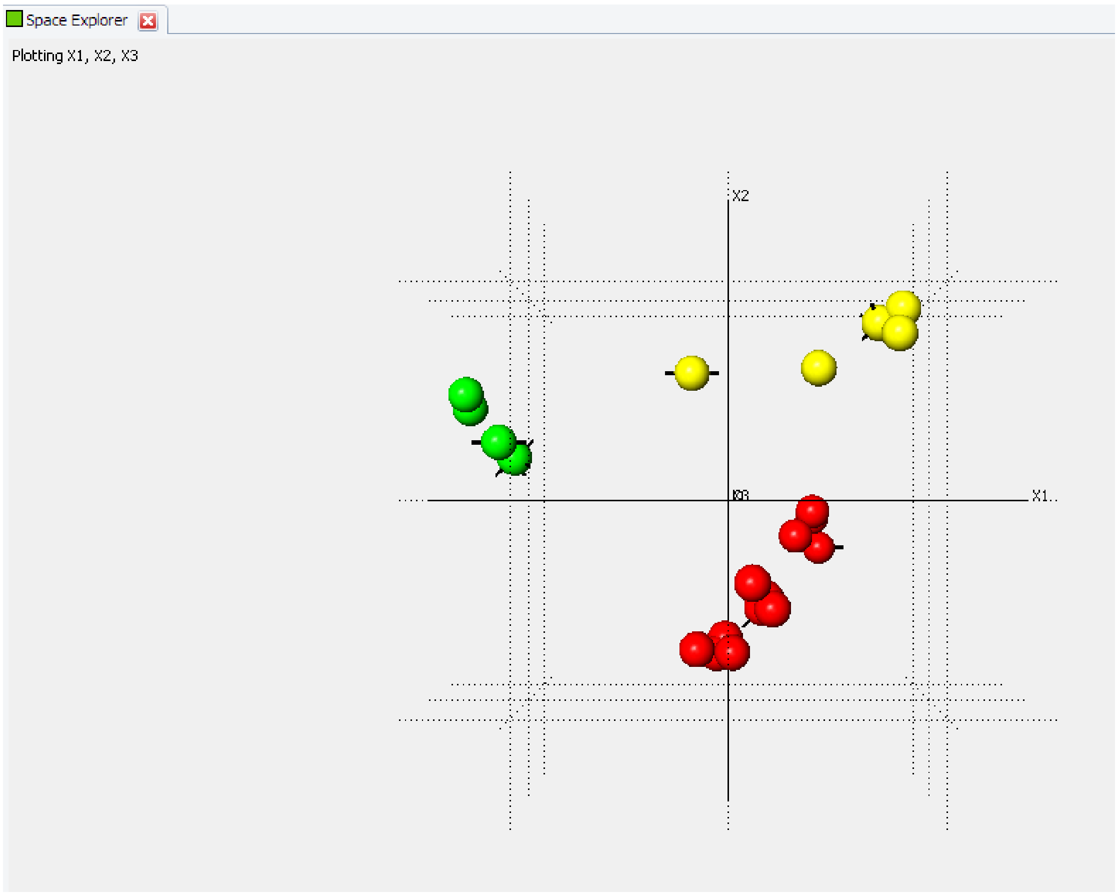


Figure 4.30: Space Explorer view

By iterating through the various combinations of the first six dimensions by means of the **Next** and **Previous** commands all of the other combinations can be accessed. The legend in the upper left corner updates to show which dimensions are currently being drawn:

This allows the user to see if the cluster separations evident in the first three dimensions still hold in higher dimensional space (up to 6).

Space Explorer View Properties

Table 4.21: Space Explorer View properties

Property	Description
Action	
Simplified Mode	Whether to present data in a simplified way. Only the most representative members of a cluster are displayed.
Attributes	
Show Label	The labels which appear next to the plotted points can be turned on or off with this option. This may aid in seeing an overall clustering pattern when the display is crowded with many points.
Show Selected Label	Show labels for selected samples.
Show Axes	The three axes x, y, and z are drawn in the graph.
Show Grid	The grid which appears in the 3D plot can be hidden or displayed.
Max Label Length	Maximum label length defined as the number of characters. Useful if many samples are displayed.
Line Width	The axis line width.
Background Color	View’s background color.
Foreground Color	Text color for label display.
Axis Color	The color for axes and axis labels.
Label Font	The font for label display.
3D Attributes	
Show Top View	This option brings up a small simplified overview of the plot in the lower right hand corner. It can be useful for orientating yourself when zoomed into the display.
Rendering Quality	Controls the graphics drawing quality. Zero shows no spheres, one is the lowest quality and three is the default.
Light Position	Controls where the spheres are lighted from. It can be changed to achieve a different illumination effect.
Render As Dots and Show As Transparent	Alter the way the spheres are plotted as shown in the diagrams below. This can be useful to identify if for example, a single pattern of one color is hidden within a group of another patterns.
Transparency	The value controls the transparency of the balls. Zero means no transparency while larger numbers up to nine increase the transparency.

Property	Description
Grand Tour	
Grand Tour	The display can be animated for easier interpretation.

Mouse Actions and Keyboard Shortcuts in the Space Explorer view

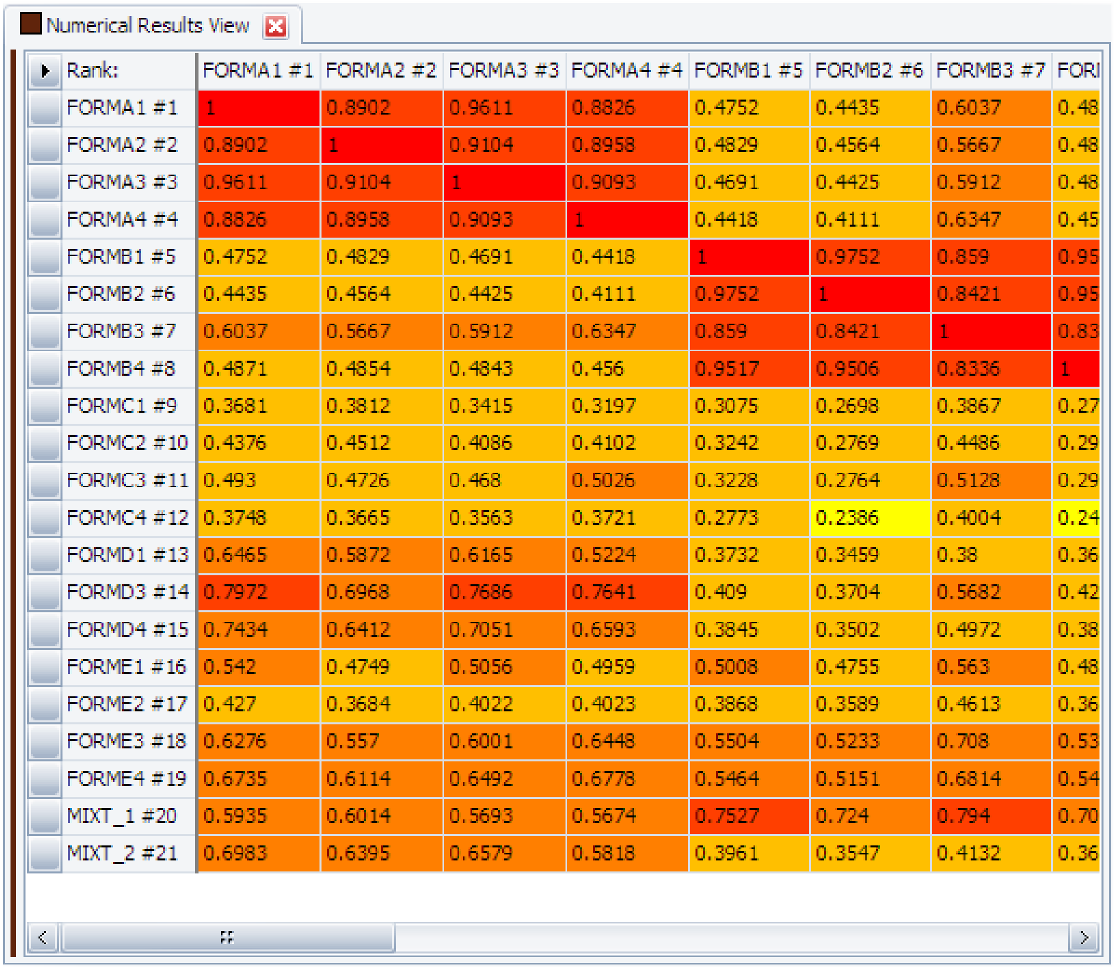
Table 4.22: Mouse and keyboard commands

Mouse Action	Description
Mouse Drag	Zoom In/Out
Shift+Mouse Drag	Rotate
Alt+Mouse Drag	Translate
Click Object	Select
Ctrl+C	Copy Graph
I	Show ID
H	Show help window

4.5 Text Output Views

4.5.1 Numerical Results View

This view is used to display the pattern correlation matrix, ρ**.** If there are a large number of patterns then the entire matrix may not fit within the display, and scroll bars will be displayed along the side and bottom of the table.



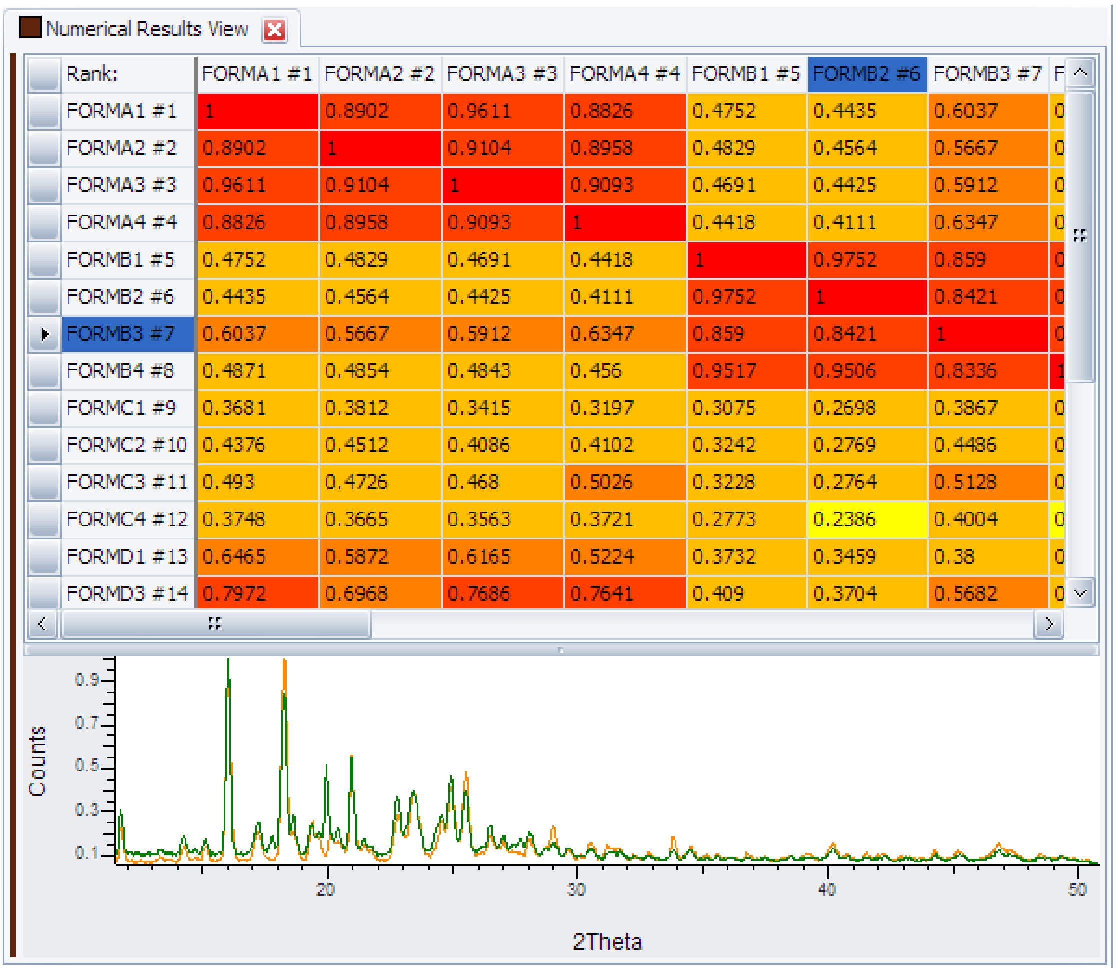
Rank:	FORMA1 #1	FORMA2 #2	FORMA3 #3	FORMA4 #4	FORMB1 #5	FORMB2 #6	FORMB3 #7	FORI
FORMA1 #1	1	0.8902	0.9611	0.8826	0.4752	0.4435	0.6037	0.48
FORMA2 #2	0.8902	1	0.9104	0.8958	0.4829	0.4564	0.5667	0.48
FORMA3 #3	0.9611	0.9104	1	0.9093	0.4691	0.4425	0.5912	0.48
FORMA4 #4	0.8826	0.8958	0.9093	1	0.4418	0.4111	0.6347	0.45
FORMB1 #5	0.4752	0.4829	0.4691	0.4418	1	0.9752	0.859	0.95
FORMB2 #6	0.4435	0.4564	0.4425	0.4111	0.9752	1	0.8421	0.95
FORMB3 #7	0.6037	0.5667	0.5912	0.6347	0.859	0.8421	1	0.83
FORMB4 #8	0.4871	0.4854	0.4843	0.456	0.9517	0.9506	0.8336	1
FORMC1 #9	0.3681	0.3812	0.3415	0.3197	0.3075	0.2698	0.3867	0.27
FORMC2 #10	0.4376	0.4512	0.4086	0.4102	0.3242	0.2769	0.4486	0.29
FORMC3 #11	0.493	0.4726	0.468	0.5026	0.3228	0.2764	0.5128	0.29
FORMC4 #12	0.3748	0.3665	0.3563	0.3721	0.2773	0.2386	0.4004	0.24
FORMD1 #13	0.6465	0.5872	0.6165	0.5224	0.3732	0.3459	0.38	0.36
FORMD3 #14	0.7972	0.6968	0.7686	0.7641	0.409	0.3704	0.5682	0.42
FORMD4 #15	0.7434	0.6412	0.7051	0.6593	0.3845	0.3502	0.4972	0.38
FORME1 #16	0.542	0.4749	0.5056	0.4959	0.5008	0.4755	0.563	0.48
FORME2 #17	0.427	0.3684	0.4022	0.4023	0.3868	0.3589	0.4613	0.36
FORME3 #18	0.6276	0.557	0.6001	0.6448	0.5504	0.5233	0.708	0.53
FORME4 #19	0.6735	0.6114	0.6492	0.6778	0.5464	0.5151	0.6814	0.54
MIXT_1 #20	0.5935	0.6014	0.5693	0.5674	0.7527	0.724	0.794	0.70
MIXT_2 #21	0.6983	0.6395	0.6579	0.5818	0.3961	0.3547	0.4132	0.36

Figure 4.31: Numerical results view

When more than 500 entries are in the matrix, it is coloured using a colour scheme set up in the **Settings** dialog, **Cluster Analysis** tab, **Display & Advanced** category. This is intended to make it easier to spot trends or outliers in the results, but can be turned off to improve legibility.

If the program was run with the **Allow Offsets** option turned on, then the results from this can be seen by selecting a particular cell, and then allowing the mouse to hover over that cell for a second, until a tooltip appears. This contains two numbers in the form ($a_{0,}a_1$), where a_0 is the amount of linear offset applied, and a_1 the amount of non-linear offset applied. A diagonal line of 1 should be present to show the result of each pattern matched against itself. Clicking on the *1* for each pattern will produce the profile and information for that pattern in the relevant area below.

For comparison purposes two patterns can be overlaid on each other by clicking on a number above or below the line of *1*.



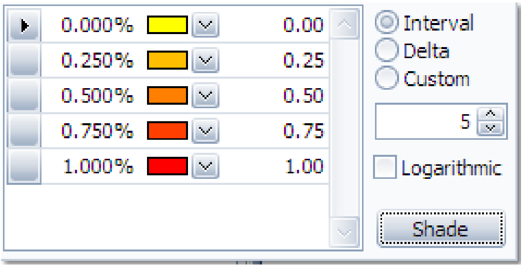
Rank:	FORMA1 #1	FORMA2 #2	FORMA3 #3	FORMA4 #4	FORMB1 #5	FORMB2 #6	FORMB3 #7	F
FORMA1 #1	1	0.8902	0.9611	0.8826	0.4752	0.4435	0.6037	0
FORMA2 #2	0.8902	1	0.9104	0.8958	0.4829	0.4564	0.5667	0
FORMA3 #3	0.9611	0.9104	1	0.9093	0.4691	0.4425	0.5912	0
FORMA4 #4	0.8826	0.8958	0.9093	1	0.4418	0.4111	0.6347	0
FORMB1 #5	0.4752	0.4829	0.4691	0.4418	1	0.9752	0.859	0
FORMB2 #6	0.4435	0.4564	0.4425	0.4111	0.9752	1	0.8421	0
FORMB3 #7	0.6037	0.5667	0.5912	0.6347	0.859	0.8421	1	0
FORMB4 #8	0.4871	0.4854	0.4843	0.456	0.9517	0.9506	0.8336	1
FORMC1 #9	0.3681	0.3812	0.3415	0.3197	0.3075	0.2698	0.3867	0
FORMC2 #10	0.4376	0.4512	0.4086	0.4102	0.3242	0.2769	0.4486	0
FORMC3 #11	0.493	0.4726	0.468	0.5026	0.3228	0.2764	0.5128	0
FORMC4 #12	0.3748	0.3665	0.3563	0.3721	0.2773	0.2386	0.4004	0
FORMD1 #13	0.6465	0.5872	0.6165	0.5224	0.3732	0.3459	0.38	0
FORMD3 #14	0.7972	0.6968	0.7686	0.7641	0.409	0.3704	0.5682	0

Figure 4.32: Numerical Results view with a selection below the diagonal

This allows a visual comparison to help decide if the matching results displayed are sensible or not.

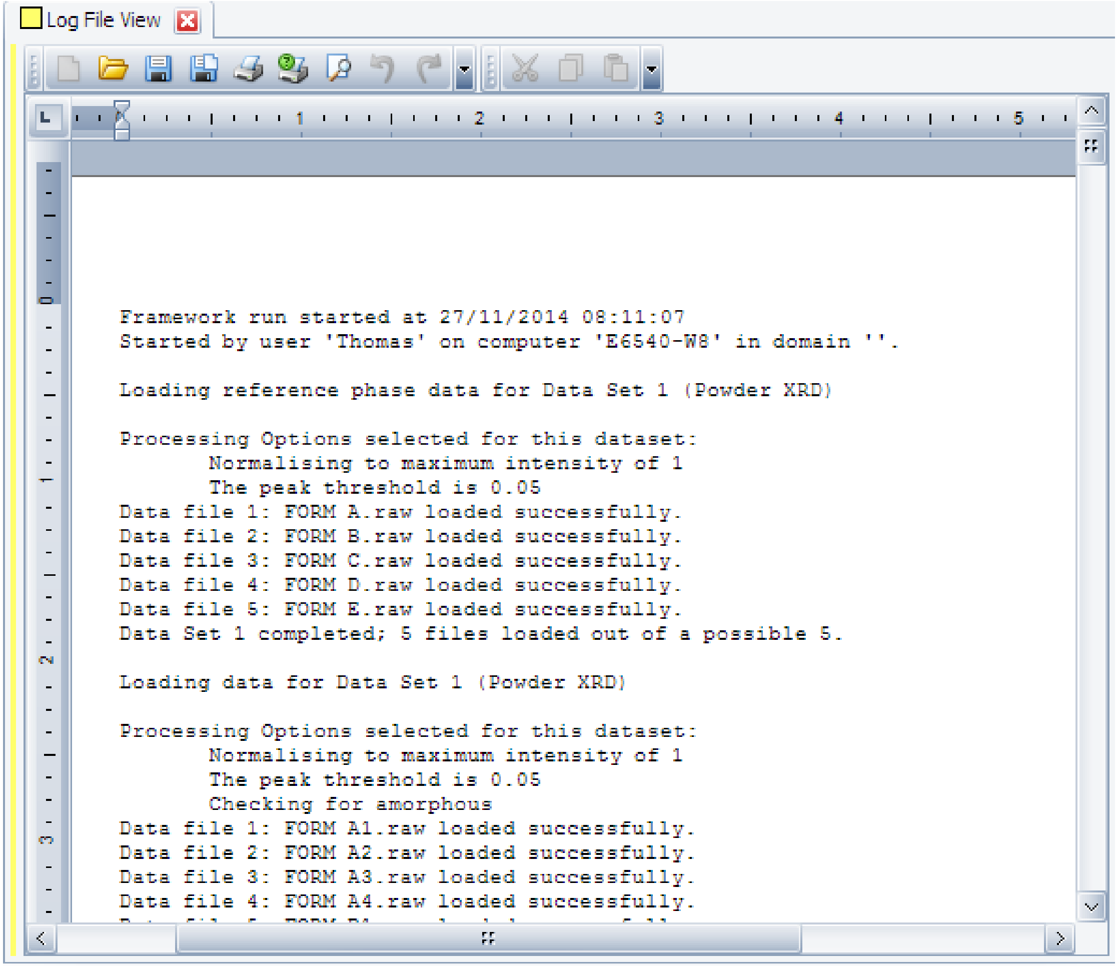
Numerical Results View Properties

Table 4.23: Numerical Results view properties

Property	Description
Attributes	
Cell Color	Display the cells colored
Cell Color Palette	Choose how the cells are colored:  0.000% 🟨 0.00 0.250% 🟧 0.25 0.500% 🟧 0.50 0.750% 🟥 0.75 1.000% 🟥 1.00 ● Interval ○ Delta ○ Custom 5 □ Logarithmic Shade
X Shift	To show scans with optimized x shift

4.5.2 Log File View

The results of the file import, processing, pattern matching, clustering etc., and any subsequent changes to the results are displayed here in the **Log File** view:



```
Framework run started at 27/11/2014 08:11:07
Started by user 'Thomas' on computer 'E6540-W8' in domain ''.

Loading reference phase data for Data Set 1 (Powder XRD)

Processing Options selected for this dataset:
        Normalising to maximum intensity of 1
        The peak threshold is 0.05
Data file 1: FORM A.raw loaded successfully.
Data file 2: FORM B.raw loaded successfully.
Data file 3: FORM C.raw loaded successfully.
Data file 4: FORM D.raw loaded successfully.
Data file 5: FORM E.raw loaded successfully.
Data Set 1 completed; 5 files loaded out of a possible 5.

Loading data for Data Set 1 (Powder XRD)

Processing Options selected for this dataset:
        Normalising to maximum intensity of 1
        The peak threshold is 0.05
        Checking for amorphous
Data file 1: FORM A1.raw loaded successfully.
Data file 2: FORM A2.raw loaded successfully.
Data file 3: FORM A3.raw loaded successfully.
Data file 4: FORM A4.raw loaded successfully.
```

*Figure 4.33: The **Log File** view*

A scroll bar appear on the right hand side and of the bottom of the view to allow the user to view all of the text, as the output is normally quite long.

The text can be selected with the mouse, copied to the clipboard (**toolbar** button or **Ctrl-C**) and pasted to other applications.

Log File View Toolbar

Table 4.24: Description of Toolbar Buttons

Toolbar Button	Description
💾	**Save as…** - Opens a standard Save As dialog box. The log file can be saved in several file formats: RTF Files (*.rtf) - Rich Text Format HTML Files (*.html) Text Files (*.txt) - ASCII Text Format
🖨	**Print -** Prints the log file to the default printer immediately.
🖨	**Print… -** Opens the Print dialog allowing printer selection and other configurations before printing.

Toolbar Button	Description
✂	**Cut** - Highlight an area by holding the left mouse button down and dragging over a relevant area. Selecting Cut causes the selection to be removed from the screen and copied into the clipboard.
⧉	**Copy** - A selected region of the report pane is copied in to the clipboard, and can then be pasted into another application.
📋	**Paste** - An object or text which has previously been copied into the clipboard is pasted at the current insertion point.

Note: in 21 CFR Part 11 mode commands which modify the log file (cut, paste) are not available.

Log File View Properties

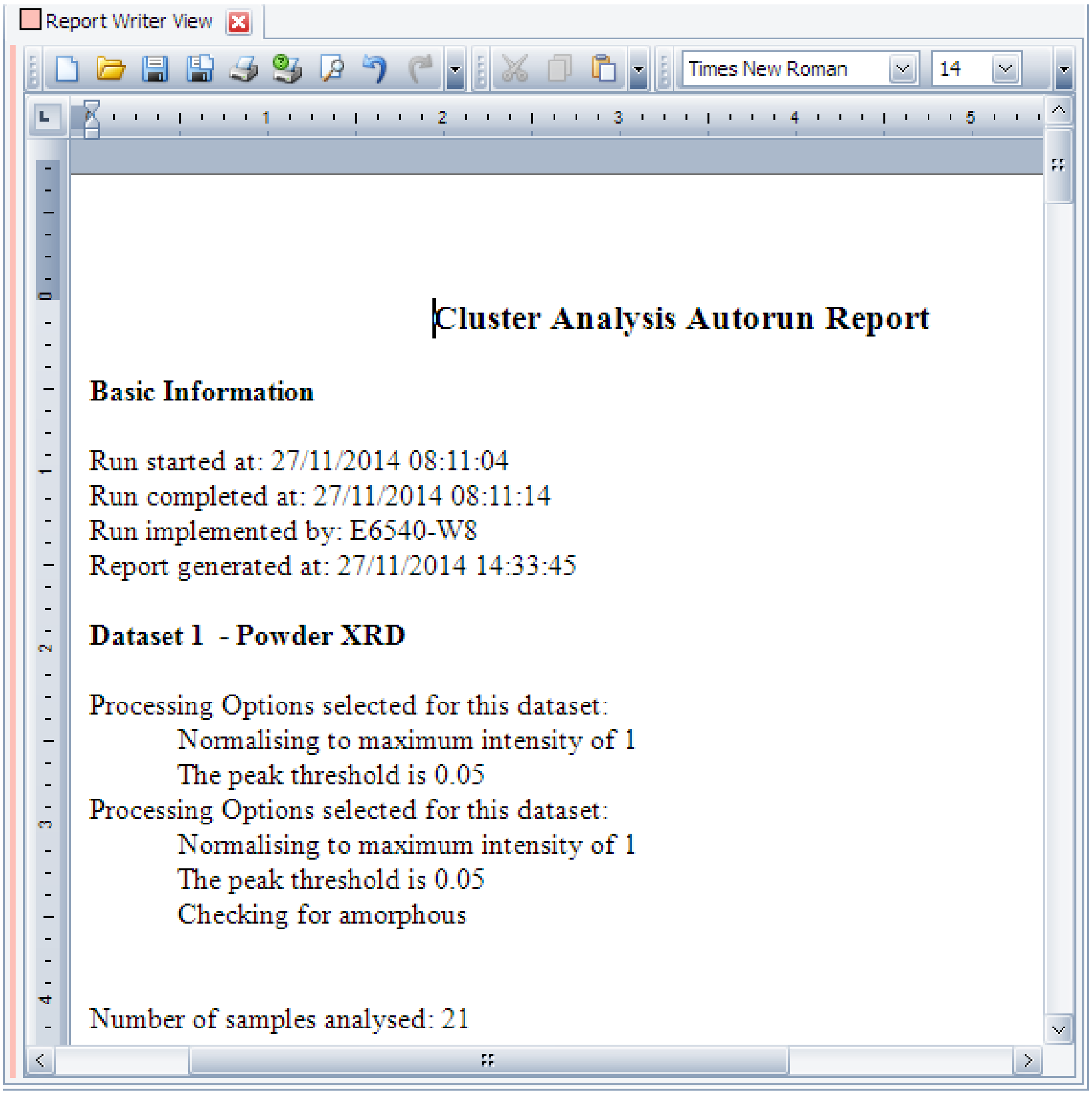
Table 4.25: Log File view properties

Property	Description
Attributes	
Font	Select the text font for the log file display.

4.5.3 Report Writer

The report writer is an area that functions as a basic word-processor, in which the user can produce reports containing information from any of the output displays, or any other text or bitmapped graphics that can be pasted from the clipboard as desired. The basic report is generated automatically by the program and can then be edited as needed.

The pane itself initially consists of a white screen with a standard text editing toolbar is provided at the top of the window, allowing choice of font, font size, text style and text alignment in the usual manner:



Cluster Analysis Autorun Report

Basic Information

Run started at: 27/11/2014 08:11:04
Run completed at: 27/11/2014 08:11:14
Run implemented by: E6540-W8
Report generated at: 27/11/2014 14:33:45

Dataset 1 - Powder XRD

Processing Options selected for this dataset:
Normalising to maximum intensity of 1
The peak threshold is 0.05
Processing Options selected for this dataset:
Normalising to maximum intensity of 1
The peak threshold is 0.05
Checking for amorphous

Number of samples analysed: 21

*Figure 4.34: The **Report Writer** view*

The report can be printed using the view's toolbar commands or using the traditional **Print Preview** command in the main toolbar or menu.

Report Writer Toolbar

Table 4.26: Description of Toolbar Buttons

Toolbar Button	Description
💾	**Save as…** - Opens a standard Save As dialog box. The log file can be saved in several file formats: RTF Files (*.rtf) - Rich Text Format HTML Files (*.html) Text Files (*.txt) - ASCII Text Format
🖨	**Print -** Prints the log file to the default printer immediately.
🖨	**Print… -** Opens the Print dialog allowing printer selection and other configurations before printing.
✂	**Cut** - Highlight an area by holding the left mouse button down and dragging over a relevant area. Selecting Cut causes the selection to be removed from the screen and copied into the clipboard.

Toolbar Button	Description
⧉	**Copy** - A selected region of the report pane is copied in to the clipboard, and can then be pasted into another application.
📋	**Paste** - An object or text which has previously been copied into the clipboard is pasted at the current insertion point.

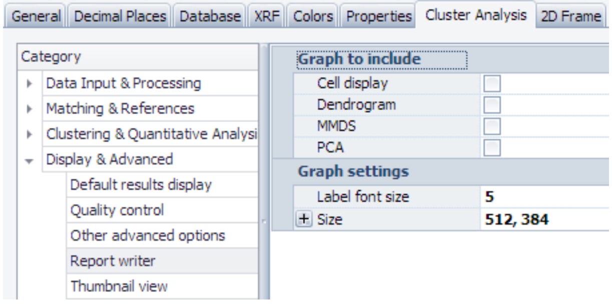
Note: in 21 CFR Part 11 mode commands which modify the report (cut, paste) are not available.

Report Writer View Properties

The **Report Writer** view has no specific properties beside the general name and description attributes.

Report Writer View Settings

The following settings are available in the Settings dialog, Cluster Analysis tab:



Graph to include

Cell display	□
Dendrogram	□
MMDS	□
PCA	□

Graph settings

Label font size	5
+ Size	512, 384

*Figure 4.35: **Report Writer View** settings*

The **Report Writer** View can include miniaturized graphics of certain views. The views which should be included must be selected in the settings. It is also possible to set the size of the graphics.

4.6 References

The following references may be useful:

‘High Throughput Powder Diffraction: I Full-profile Qualitative and Quantitative Powder Diffraction Pattern Analysis’ C.J. Gilmore, G. Barr, and J. Paisley, *J. Appl. Cryst.* (2004). 37, 231-242.

‘High Throughput Powder Diffraction: II Applications of Clustering Methods and Multivariate Data Analysis’ G. Barr, W. Dong and C.J. Gilmore, *J. Appl. Cryst.* (2004). 37, 243-252.

High Throughput Powder Diffraction III: The Application of Full Profile Pattern Matching and Multivariate Statistical Analysis to Round-Robin and Related Powder Diffraction Data Gordon Barr, Wei Dong, Christopher Gilmore *and* John Faber *J. Appl. Cryst.* (2004). 37, 243-252.

SNAP-1D: A Computer Program for Qualitative and Quantitative Powder Diffraction Pattern Analysis Using the Full Pattern Profile Gordon Barr, Christopher J. Gilmore, and Jonathan Paisley *J. Appl. Cryst.* (2004). 37, 665-668.

PolySNAP: A Computer Program for Analyzing High Throughput Powder Diffraction Data Gordon Barr, Wei Dong, and Christopher J. Gilmore *J. Appl. Cryst.* (2004). 37, 635-642.

‘Automation of Solid Form Screening Procedures in The Pharmaceutical Industry and How to Avoid the Bottlenecks.’ R. Storey, R. Docherty, P. Higginson, C. Dallman, C. Gilmore, G. Barr, W. Dong, *Crystallography Reviews* (2004). 10, 45-56.

‘High Throughput Powder Diffraction: IV Cluster Validation using Silhouettes and Fuzzy Clustering.' G. Barr, W. Dong, and C.J. Gilmore, *J. Appl. Cryst.* (2004). 37, 874-882.

‘A quick method for the quantitative analysis of mixtures. 1. Powder X-ray diffraction.’ W Dong, C J Gilmore, G Barr, C Dallman, N Feeder and S Terry (2007). *J. Pharm. Sci* DOI 10.1002/jps.

‘High Throughput Powder Diffraction V: The Use of Raman Spectroscopy with and without X-ray Powder Diffraction data’. G. Barr, G. Cunningham, W. Dong, C.J. Gilmore and T. Kojima. (2009). *J. Appl. Cryst.* 42, 706-714.

‘PolySNAP 3: A Computer Program for Analyzing High Throughput Data from Diffraction and Spectroscopic Sources’. G. Barr, W. Dong and C.J. Gilmore, *J. Appl. Cryst.* (2009). 42, 965-974.

5 Pre-Screening

5.1 Introduction

Pre-screening mode assumes you have a single new unknown sample file, which you wish to compare to a large number of existing patterns (e.g. > 10,000). On a reasonably modern PC, 10,000 patterns should take around 10 minutes to process. This allows an effectively unlimited number of patterns to be narrowed down to a size suitable for full cluster analysis.

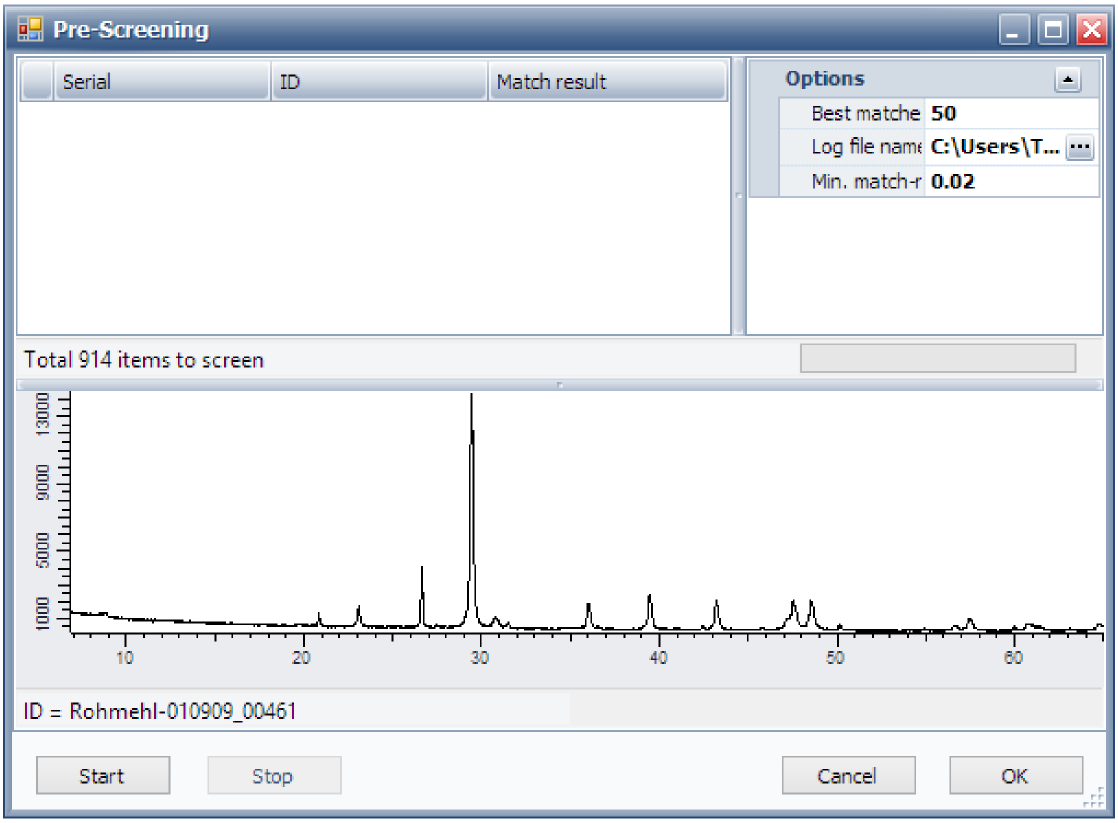
The existing library patterns correspond to samples already collected; the user wishes to know which of the many library patterns the new sample is most similar to (if any).

The cluster analysis is limited by default to performing cluster analysis on up to 2000 patterns per single dataset. It doesn't make sense however to perform a match everything-against-everything for the large library of data, so pre-screening it to identify the most relevant patterns is the obvious solution.

Once the most similar patterns have been identified in this manner, they can then be examined in the cluster analysis.

5.2 The Pre-Screening Process

1. Import the unknown data into **Set 1**.
2. Import all reference data into the **Set 1 | Reference** node.
3. Start the **Pre-Screen** command on the **Cluster Analysis** node.
4. The **Pre-Screening** dialog is displayed:



Serial	ID	Match result

Options

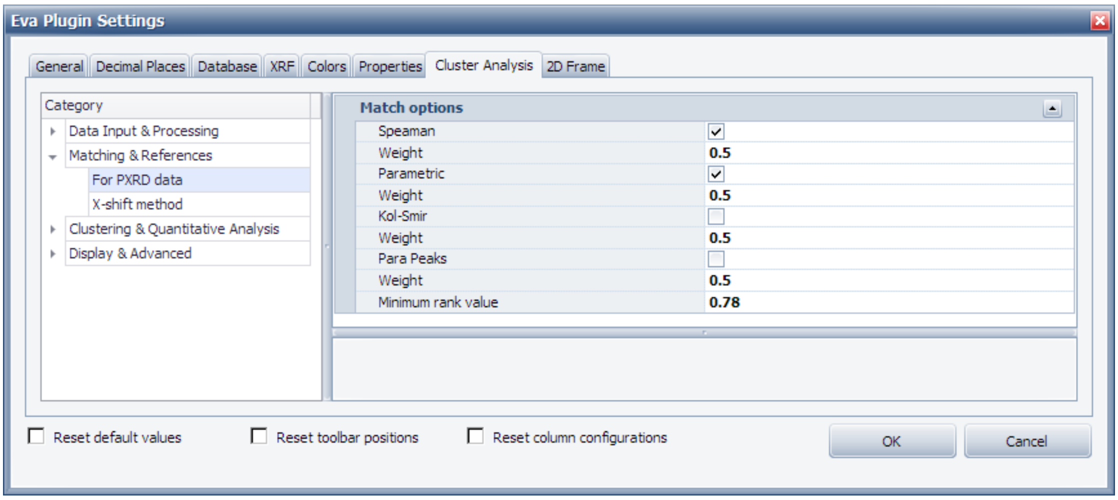
Best matche	50
Log file name	C:\Users\T...
Min. match-r	0.02

Total 914 items to screen

ID = Rohmehl-010909_00461

Figure 5.1: Pre-Screening dialog before starting the analysis

The properties ion the top right property grid allow the user to control the matching process. **Best matches to keep** controls how many of the top matches to the new sample should be returned. This can be any number between 1 and 2000 (default of 50). To limit poor matches, a default minimum rank threshold (default of 0.02) can also be set here. The matching rank is calculated using the match tests selected in the **Settings** dialog box. The default is 50%-50% Spearman-Parametric full profile matching statistic:



Category
- Data Input & Processing
- Matching & References
  - For PXRD data
  - X-shift method
- Clustering & Quantitative Analysis
- Display & Advanced

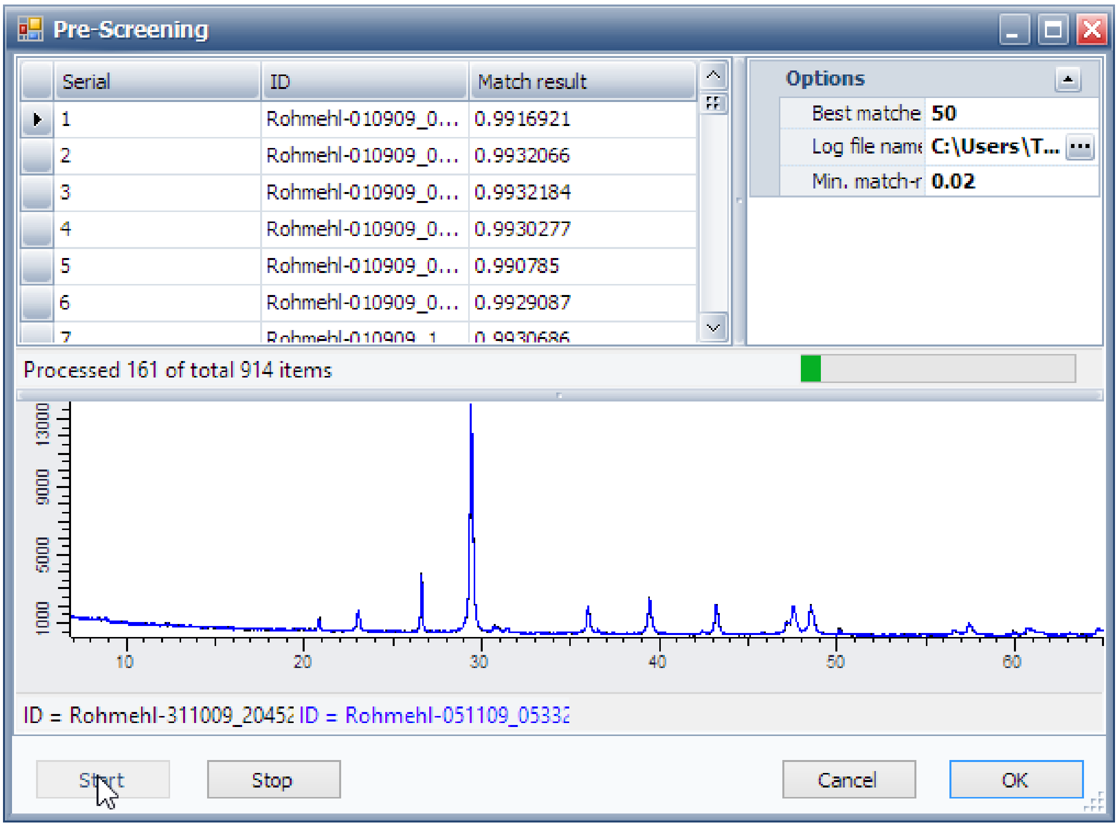
Match options

Speaman	☑
Weight	0.5
Parametric	☑
Weight	0.5
Kol-Smir	□
Weight	0.5
Para Peaks	□
Weight	0.5
Minimum rank value	0.78

□ Reset default values □ Reset toolbar positions □ Reset column configurations

*Figure 5.2: The **Matching** options in the **Settings** dialog*

After clicking the **Start** button the pre-screening process is executed. A progress bar and text output in the dialog informs the user about current progress:



	Serial	ID	Match result
▸	1	Rohmehl-010909_0...	0.9916921
	2	Rohmehl-010909_0...	0.9932066
	3	Rohmehl-010909_0...	0.9932184
	4	Rohmehl-010909_0...	0.9930277
	5	Rohmehl-010909_0...	0.990785
	6	Rohmehl-010909_0...	0.9929087
	7	Rohmehl-010909_1	0.9930686

Best matche	50
Log file name	C:\Users\T...
Min. match-r	0.02

Processed 161 of total 914 items

ID = Rohmehl-311009_20452 ID = Rohmehl-051109_05332

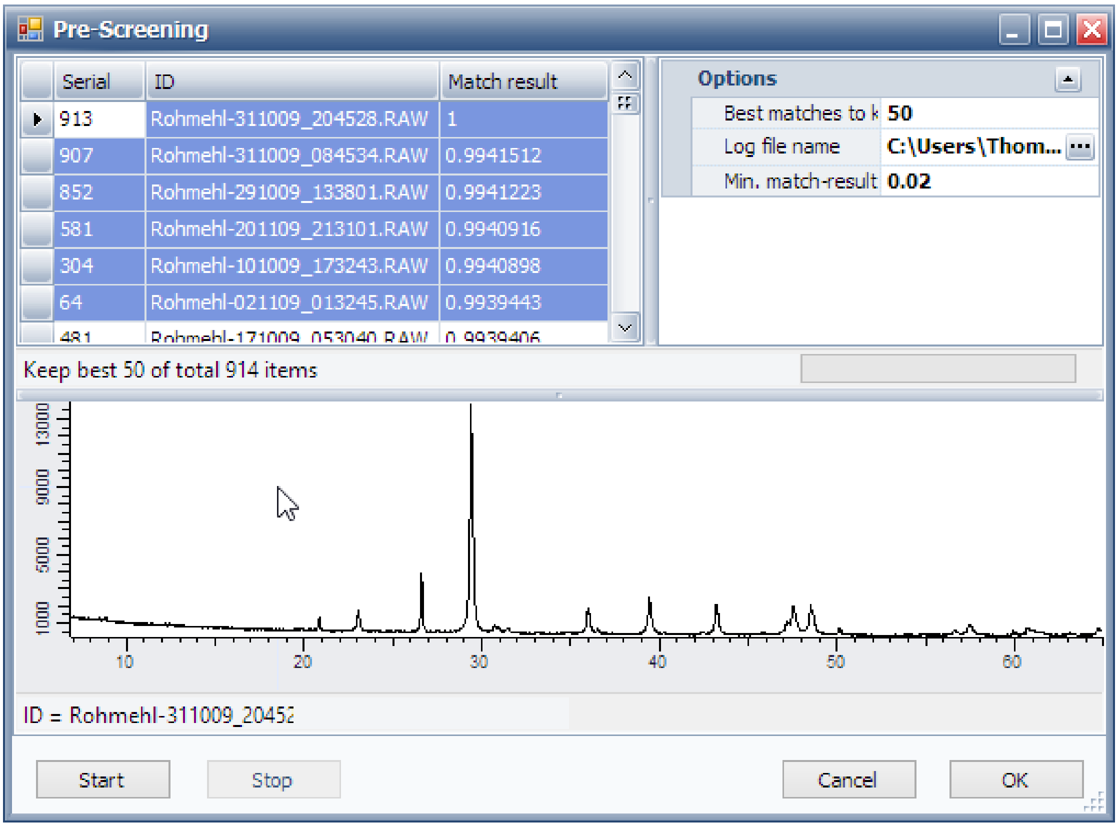
*Figure 5.3: **Pre-Screening** dialog during analysis with matching progress displayed*

Once matching begins, the total files found and the number examined is shown along with the progress bar. Matching continues until all the files found have been examined, or until the **Stop** button is clicked. **Stop** can be clicked at any time to end the process.

While executing the calculation the intermediate results are displayed in the result grid in the top left area of the dialog. The data which are currently being compared are displayed in the graphic in the lower part of the dialog.

When the pre-screening is finished, the results are displayed in the results grid in order of the similarity (column Match result).

The results can be highlighted using the mouse and copied to the clipboard using the usual **Ctrl-C** keyboard command.



Serial	ID	Match result
913	Rohmehl-311009_204528.RAW	1
907	Rohmehl-311009_084534.RAW	0.9941512
852	Rohmehl-291009_133801.RAW	0.9941223
581	Rohmehl-201109_213101.RAW	0.9940916
304	Rohmehl-101009_173243.RAW	0.9940898
64	Rohmehl-021109_013245.RAW	0.9939443
481	Rohmehl-171009_053040.RAW	0.9939406

Options

Best matches to k	50
Log file name	C:\Users\Thom...
Min. match-result	0.02

Keep best 50 of total 914 items

ID = Rohmehl-311009_2045Ƨ

Figure 5.4: Pre-Screening result with a part of the result grid marked for clipboard copy

When pre-screening is finished, the results can be copied into **Set 1** for further cluster analysis by clicking the **OK** button. Clicking **Cancel** closes the dialog without further calculations.

Pre-Screening Options

Table 5.1: Pre-screening options

Option	Description
Best matches to keep	The maximum number of matches to be shown in the results grid and the log file.
Log file name	The path of the log file.
Min. match-result	Minimum rank threshold (default: 0.02) to reject poor matches.

6 Quality Control

6.1 Introduction

Quality Control is designed for situations where the stability of a material is being monitored over time; for example as part of a production line system, or for periodic checks of equipment alignment or other issues comparing the current setup to the one the last time a standard sample was measured.

A certain number of **Reference** patterns must be available; these are patterns that for the purposes of the analysis are considered to be a good representation of what is expected to be seen.

Various **Sample** patterns are then imported and compared to those reference patterns, and any that vary significantly from the ideal are noted and highlighted.

The results are displayed graphically with 'good' sample patterns shown within the surface of reference patterns, and 'bad' sample patterns appearing outside it. The tolerance for what is considered ‘good’ and ‘bad’ and hence what warnings need to be issued can be adjusted to the user’s needs.

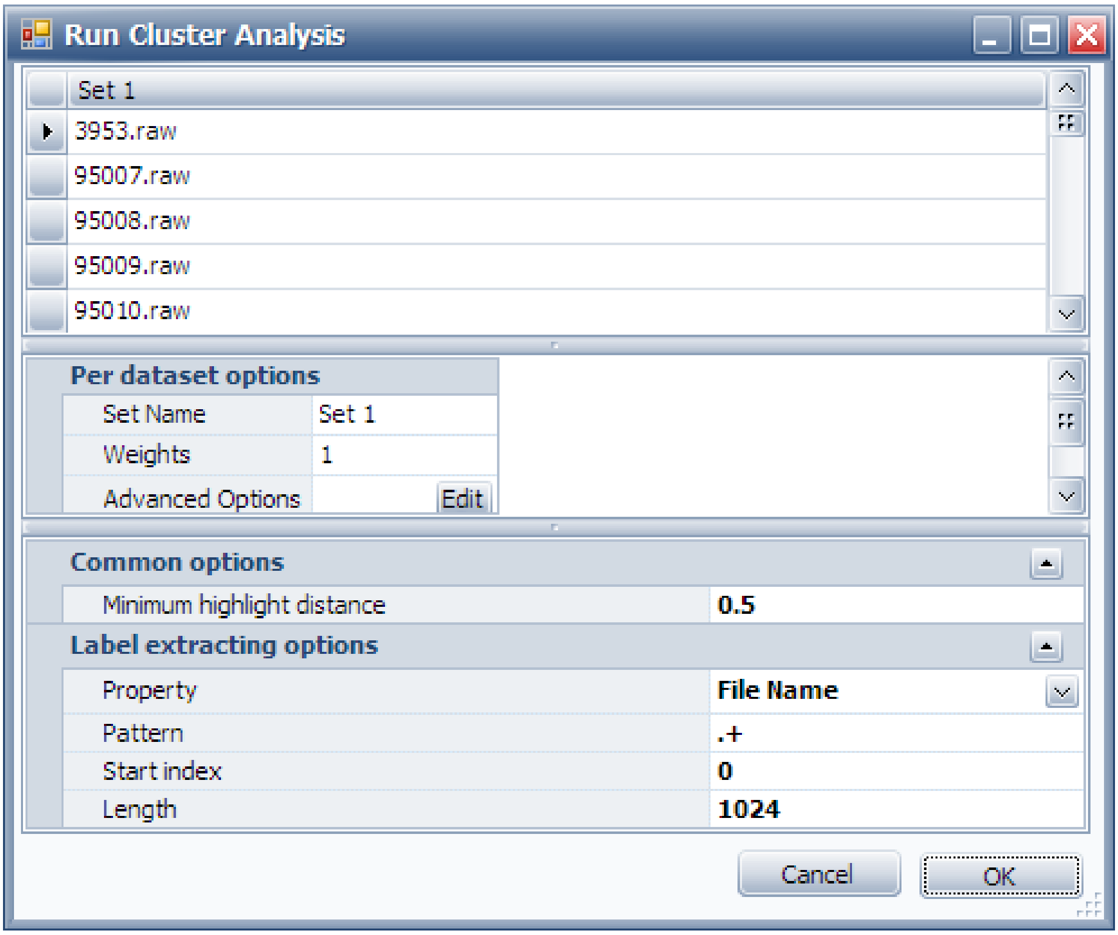
6.2 Example

The best way to illustrate the value of this mode of operation is through a worked example. The data for this example can be found in the **Quality** folder, within the **Tutorial** folder installed along with the program.

To run a quality control the reference sample measurements must be imported into the **Reference** node of **Set 1**. The samples to be checked are imported into the **Set 1** node.

When the samples are loaded,

1. The **Cluster Analysis** node must be selected.
2. The **Quality Control** command will be active and can be clicked.



Run Cluster Analysis

Set 1
3953.raw
95007.raw
95008.raw
95009.raw
95010.raw

Per dataset options

Set Name	Set 1
Weights	1
Advanced Options	Edit

Common options

Minimum highlight distance	0.5

Label extracting options

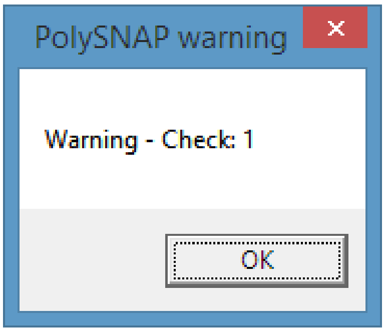
Property	File Name
Pattern	.+
Start index	0
Length	1024

*Figure 6.1: Figure 52: Run **Cluster Analysis** dialog in **quality control** mode*

⇨ The **Run Cluster Analysis** dialog is opened.

The **Common Options** contain a setting important for the Quality Control mode. The **Minimum highlight distance** can be set (default is 0.5). This acts a 'strictness' control. The lower the number (ranging from 0 to 1.0), the less far the sample patterns are allowed to vary from the centroid position of the reference patterns before they are highlighted as potentially a problem.

Clicking **OK** starts the calculation. During the analysis phase of the calculation, a warning dialog is displayed if any of the sample patterns are considered to be outside the set threshold; in this case:



PolySNAP warning

Warning - Check: 1

A simplified version of the cluster analysis result display is available after the calculation: the **3D MMDS** view and the **Log File** view:

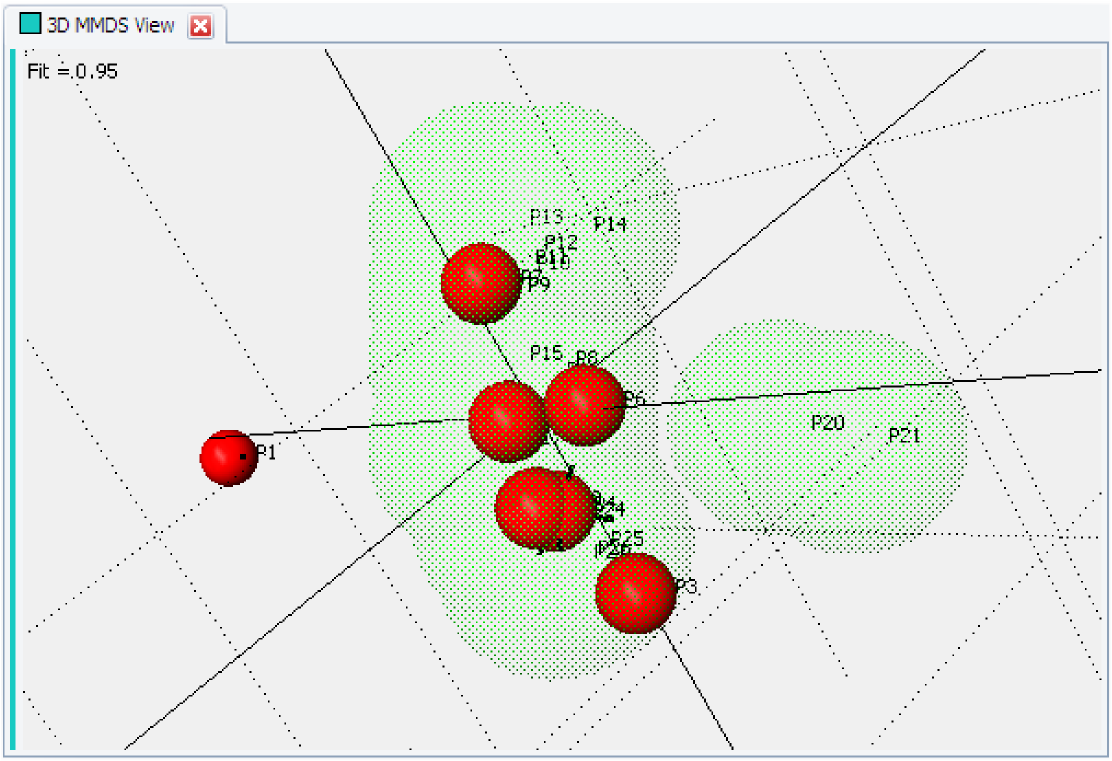


*Figure 6.2: **3D MMDS** view in **Quality Control** mode with reference envelope drawn*

The **3D MMDS** view can be interacted with like the standard **3D** plots. The red spheres are the new sample patterns, and the reference samples are represented by the shaded green area (envelope). Sample patterns within this shaded zone are considered to have passed the analysis, whereas samples outside it, such as P1 in the above example, have failed.

To display the references and the shaded green area (envelope) two properties are available in the **Quality Control** version of the **MMDS** view only:

Quality Control Mode Properties of the 3D MMDS View

Table 6.1: Special Quality Control mode properties

Property	**Description**
Show Envelope	A greed shaded area is drawn around the area covered by the reference samples.
Show Reference	The reference samples are drawn as green balls.

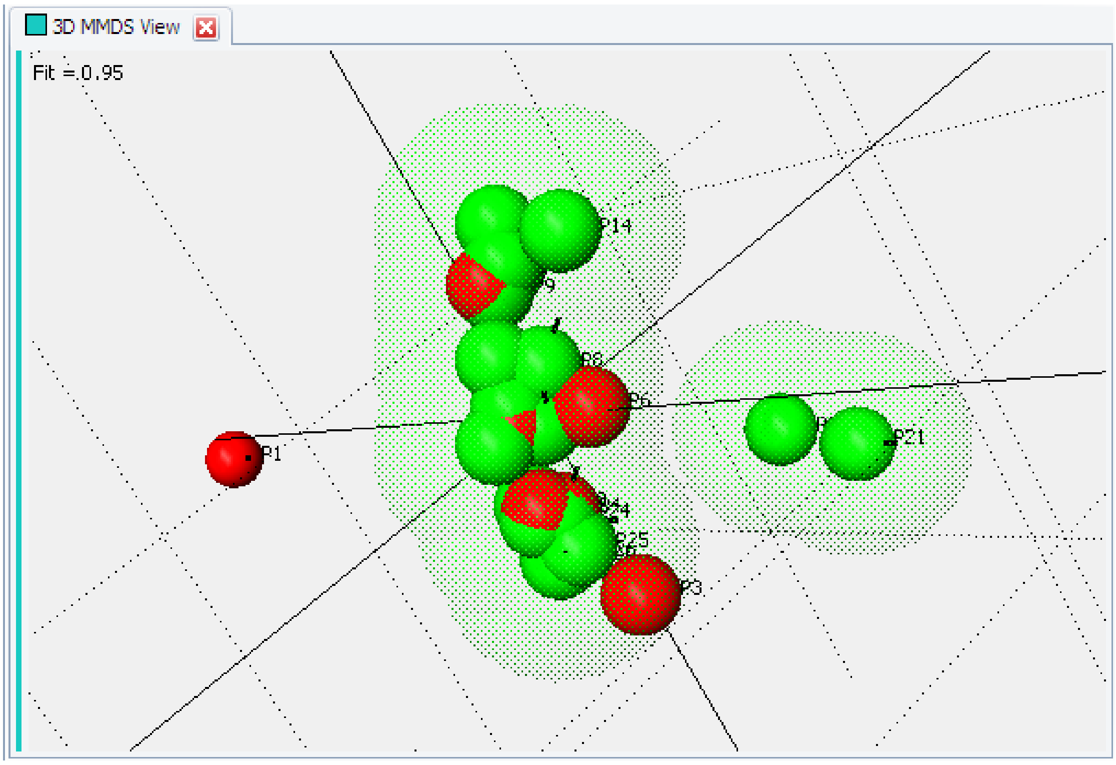


Figure 6.3: 3D MMDS view in Quality Control mode with reference samples drawn

Clicking on the **Log File** view brings up a cluster analysis audit trail log, which notes details of the patterns imported, and highlights any thought to be suspect.

7 Tutorial

This tutorial has been designed to guide the user through a few examples using cluster analysis with typical data that might be encountered in general use. It is intended as a basic introduction to using the program. It should therefore be read in conjunction with the manual itself for a more detailed explanation where necessary.

The tutorial requires the user to have already installed DIFFRAC.EVA, and be familiar with Windows-based interfaces. The data files used in the tutorial are installed along with the software, and can usually be found in the **Tutorial | Cluster Analysis** folder in *C:\ProgramData\BrukerAXS.*

There are three main sections to the tutorial: getting used to working with a single dataset, working with advanced features like detecting amorphous content and pattern shift, and using the **6D** Plot.

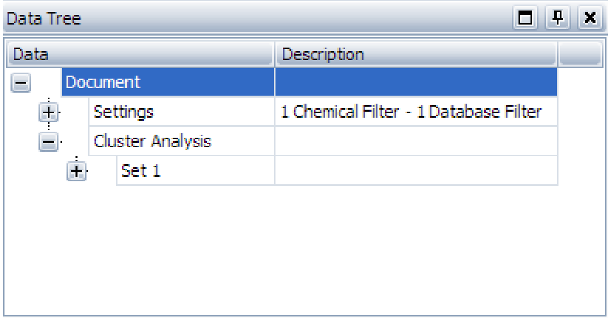
7.1 Analysis of a Single Dataset

To gain experience with the Cluster Analysis, a walkthrough of a simple run using 22 X-ray powder diffraction patterns is presented. The example assumes the program defaults are used.

To begin, launch **DIFFRAC.EVA** from the shortcut in the Windows **Start** menu.

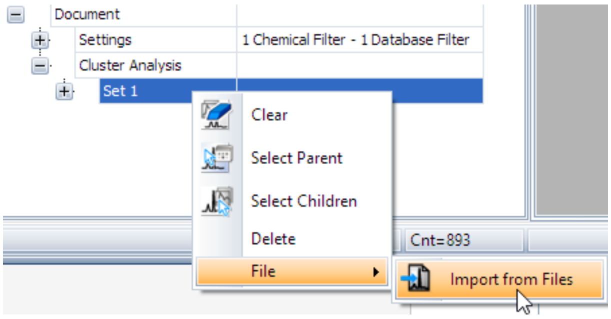
7.1.1 Import the Data to Be Analyzed

Once the program has been launched, it presents an empty document tree with a **Cluster Analysis** node and **Set 1** as child node:



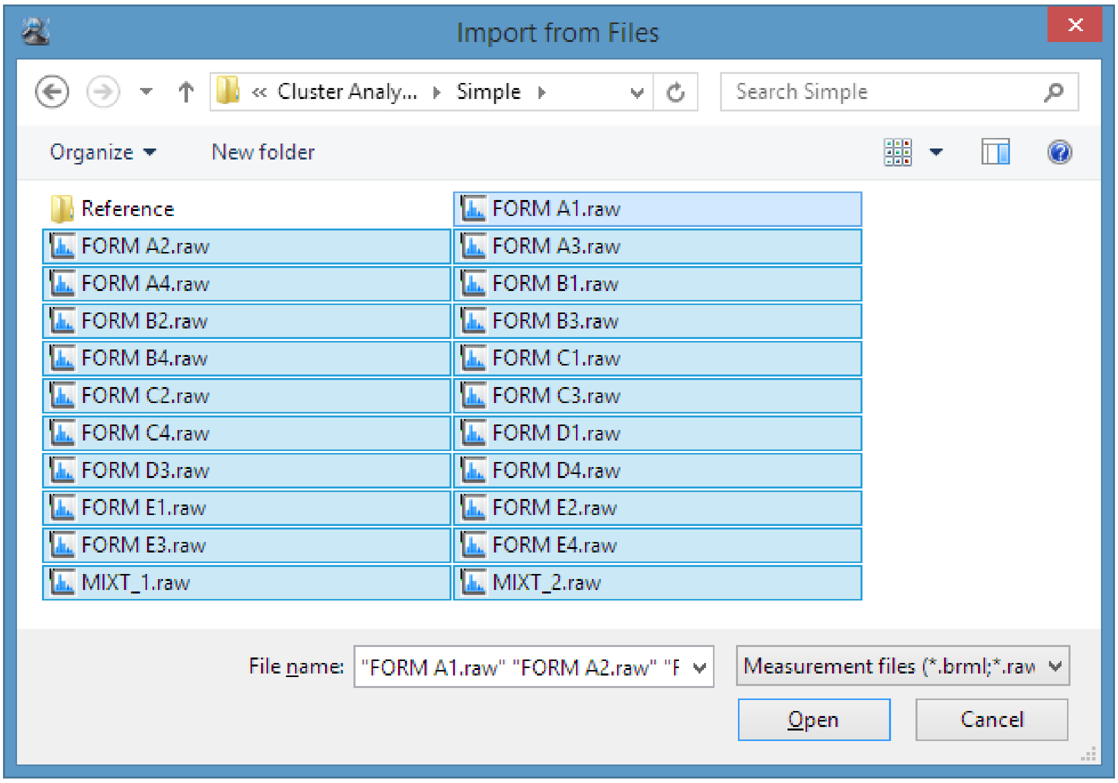
To load the data to be analyzed,

1. select the **Set 1** node,
2. open the context menu and click the **Import from Files…** command in the context menu’s **File** sub-menu:



➨ The **Import from Files** dialog opens and allows selecting the data.

3. Navigate to the folder *C:\ProgramData\Bruker AXS\Tutorial\Cluster Analysis\Simple* and
4. select all the RAW files in the folder:



File name: "FORM A1.raw" "FORM A2.raw" "F

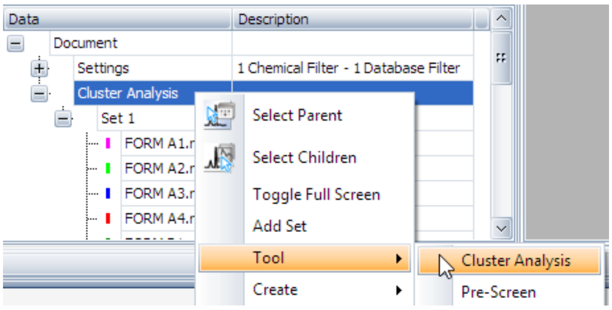
Measurement files (*.brml;*.raw

5. Click **OK** to import the data into **Set 1**.

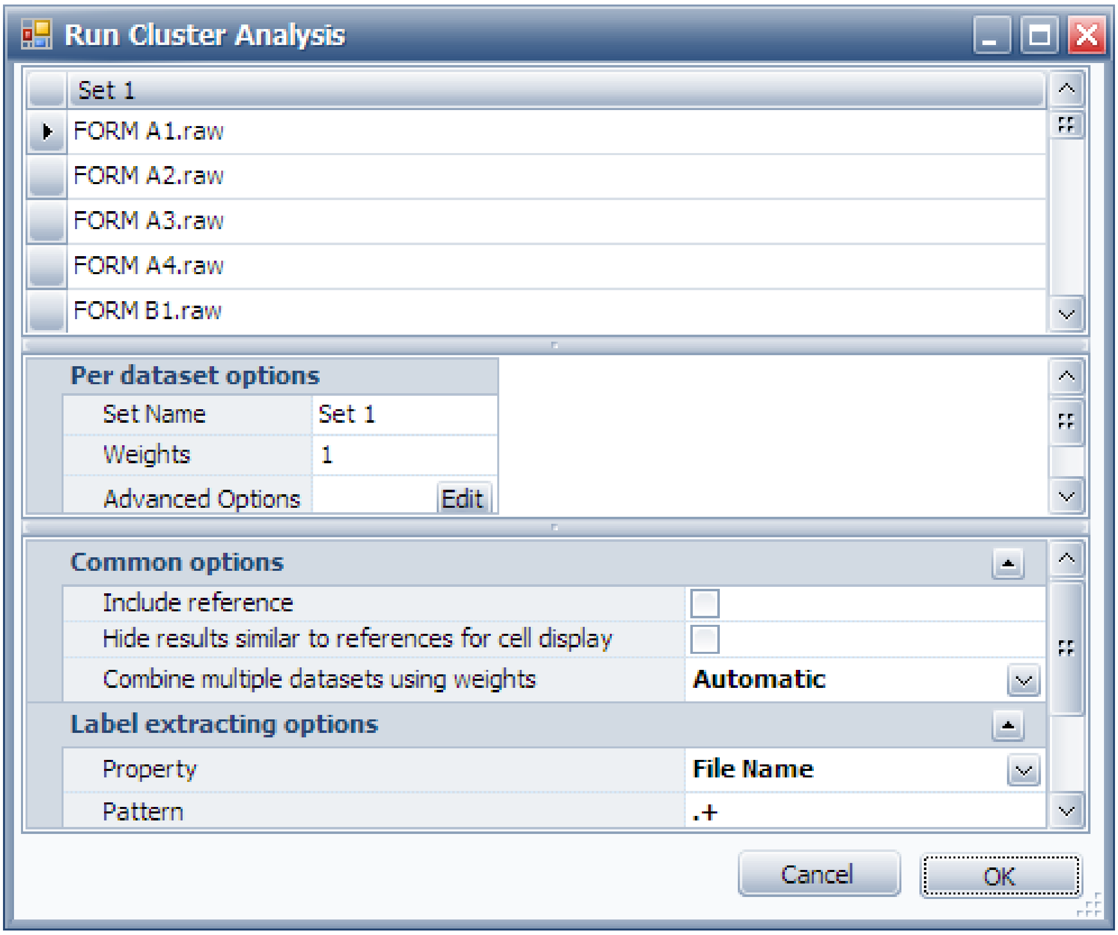
⇨ The program is now ready for running the cluster analysis.

7.1.2 Running Cluster Analysis

1. Cluster analysis will be started by selecting the **Cluster Analysis** node and
2. clicking the **Cluster Analysis** command in the **Tool** context menu:



➨ The **Run Cluster Analysis** dialog opens and presents the loaded data and available options:



Run Cluster Analysis

Set 1

- FORM A1.raw
- FORM A2.raw
- FORM A3.raw
- FORM A4.raw
- FORM B1.raw

Per dataset options	
Set Name	Set 1
Weights	1
Advanced Options	Edit

Common options	
Include reference	□
Hide results similar to references for cell display	□
Combine multiple datasets using weights	**Automatic**

Label extracting options	
Property	**File Name**
Pattern	.+

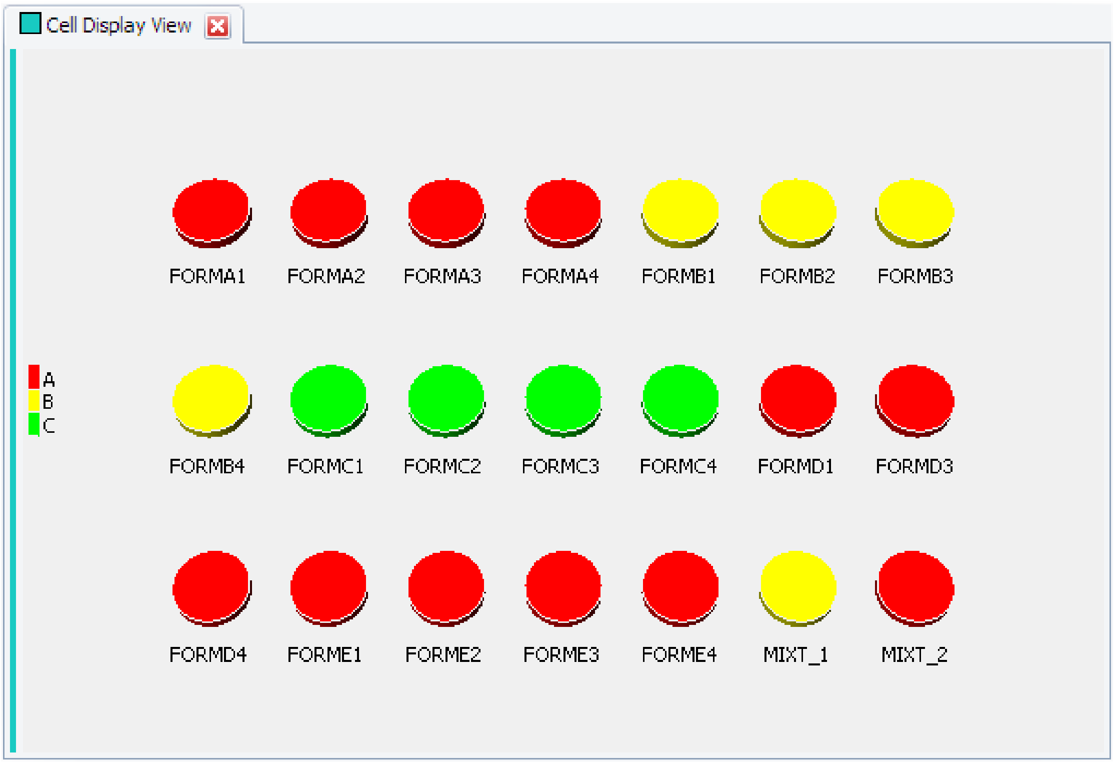
3. The options have sensible default values and the analysis can be started by clicking the **OK** button.

⇨ After displaying a progress bar briefly the analysis results are presented in various views.

Note: depending on the screen size, the views may be presented slightly different than in this tutorial.

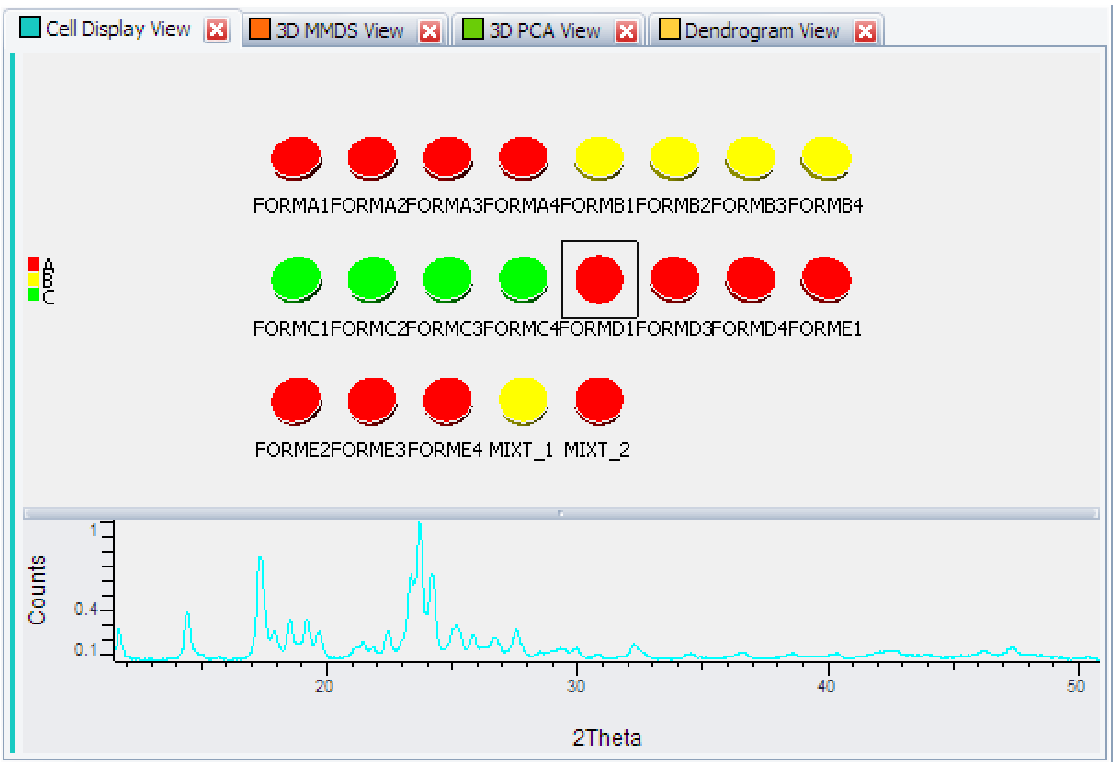
7.1.3 Examine the Analysis Results

There are a number of ways to view and examine the cluster analysis results. The first of these - the default view - is the **Cell Display View**, which visually represents the contents of each sample. Each cell (shown as a disc) represents a different pattern, with color being used to denote the suggested grouping of compounds. In other words, similar samples are given the same color:



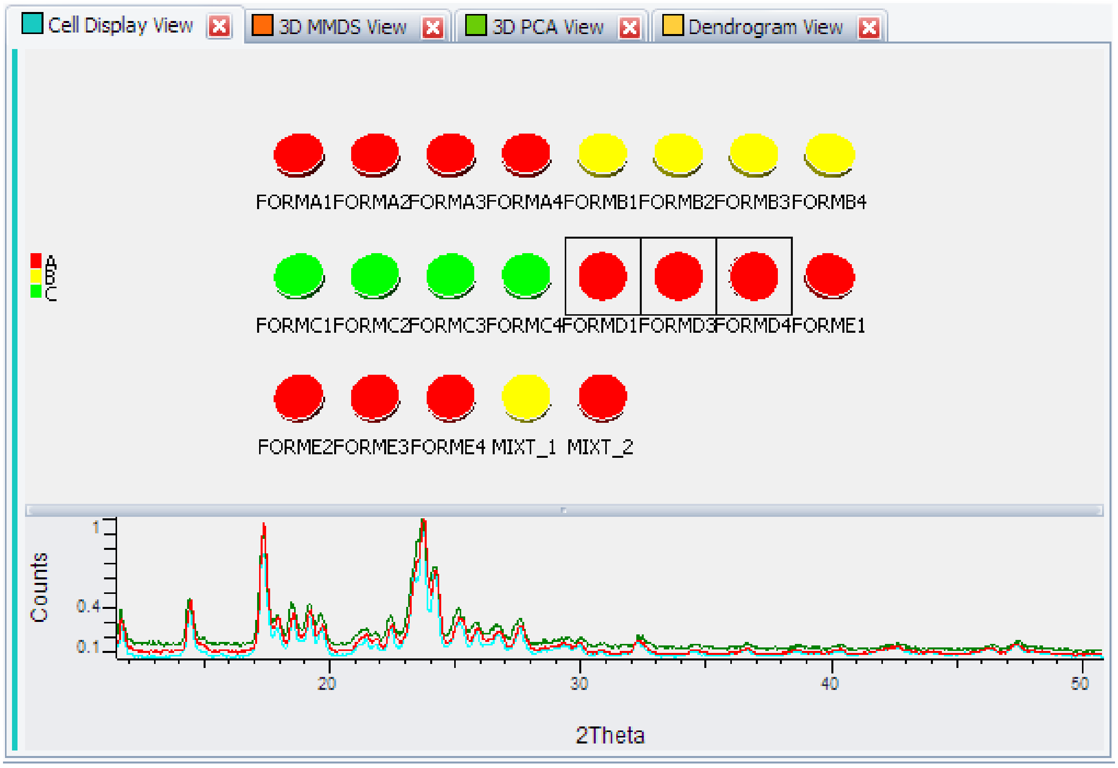
1. In the **Cell Display View** click on cell **FORMD1**.

   ➨ The associated full pattern profile of pattern D1 is now displayed in the lower region of the view.

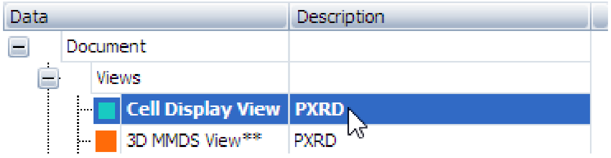


2. Hold down the **shift** key and click on cell **FORMD4**.

   ➨ The cells from D1 to D4 are now selected, and the profiles of patterns D1, D3 and D4 are overlaid to allow a visual comparison. They obviously represent the same compound.

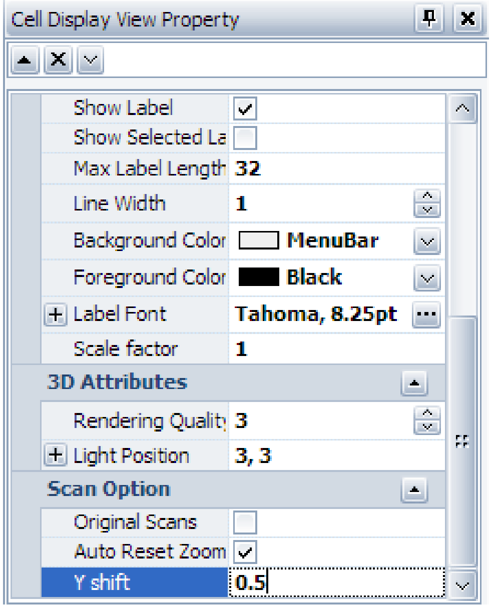

Cell Display View
3D MMDS View
3D PCA View
Dendrogram View
A
B
C
FORMA1FORMA2FORMA3FORMA4FORMB1FORMB2FORMB3FORMB4
FORMC1FORMC2FORMC3FORMC4FORMD1FORMD3FORMD4FORME1
FORME2FORME3FORME4 MIXT_1 MIXT_2
Counts
1
0.4
0.1
20
30
40
50
2Theta

3. Click in the **Data Tree** on the **Cell Display View** node:



Data	Description
Document	
Views	
Cell Display View	PXRD
3D MMDS View**	PXRD

➨ The content of the **Properties** panel shows the **Cell display View's Properties**:



Cell Display View Property

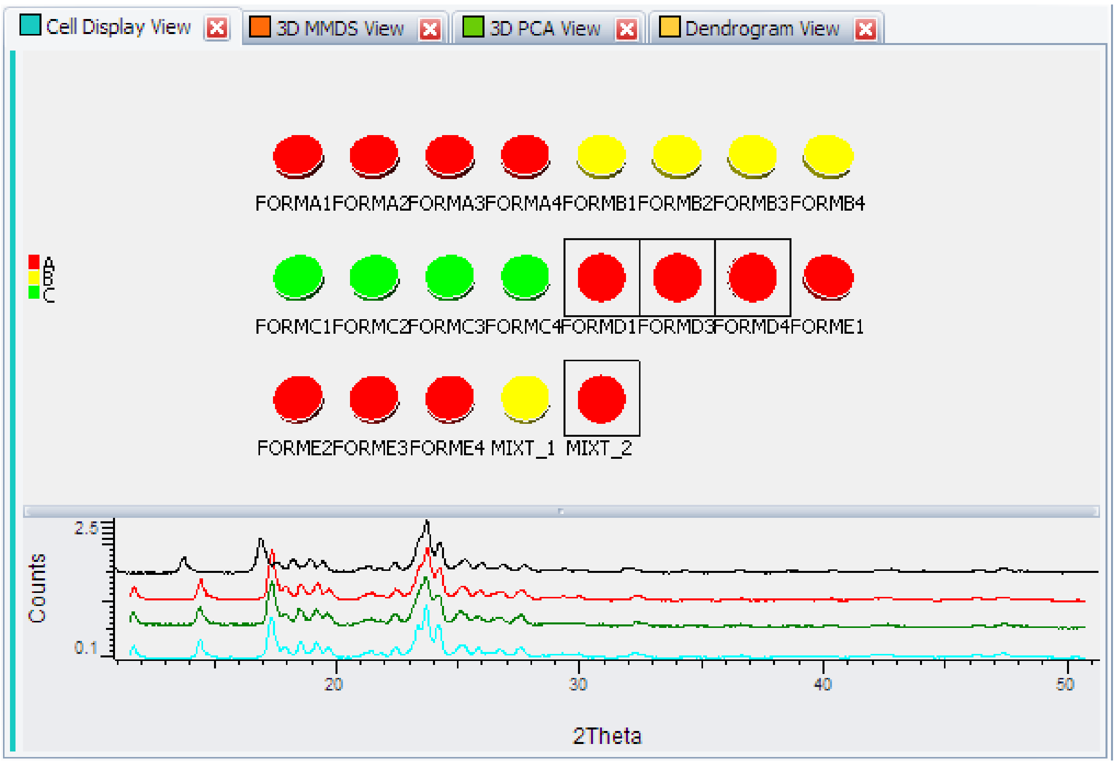
Show Label	☑
Show Selected La	□
Max Label Length	32
Line Width	1
Background Color	MenuBar
Foreground Color	Black
Label Font	Tahoma, 8.25pt
Scale factor	1
3D Attributes	
Rendering Quality	3
Light Position	3, 3
Scan Option	
Original Scans	□
Auto Reset Zoom	☑
Y shift	0.5

4. Click on the **Y Shift** property field and enter 0.5.

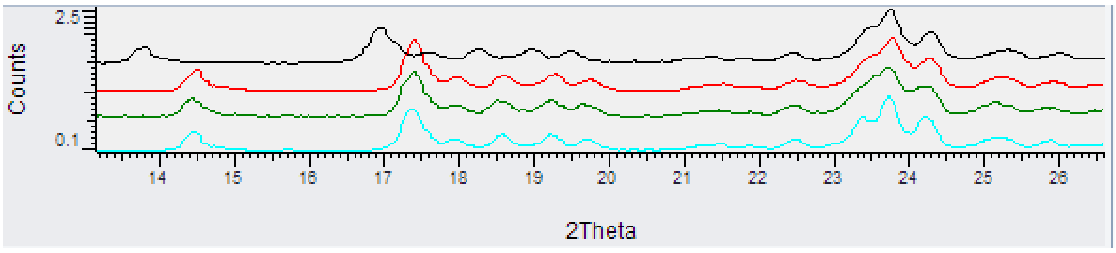
➨ This will display the multiple patterns with an offset along the y-axis of the plot, when any further cells are selected.

5. Hold down the **Control** key and click on the other cell colored red, **MIXT2**.

- ➨ The display now includes this new fourth pattern, and to allow an easy comparison are displayed overlaid with an offset along the y-axis:


Cell Display View
3D MMDS View
3D PCA View
Dendrogram View
FORMA1 FORMA2 FORMA3 FORMA4 FORMB1 FORMB2 FORMB3 FORMB4
FORMC1 FORMC2 FORMC3 FORMC4 FORMD1 FORMD3 FORMD4 FORME1
FORME2 FORME3 FORME4 MIXT_1 MIXT_2
A
B
C
Counts
2.5
0.1
20
30
40
50
2Theta

6. Zoom into the area between around 13° and 25° by holding down the **left mouse** button and dragging a box over the desired area.
   - ➨ When the mouse button is released, the graph region is redrawn to show just the selected area.
7. Stretching in x direction in and out smoothly can be accomplished by clicking in the scan display and then holding down the **control** key while moving the scroll wheel of the mouse (if available).
8. Stretching in y direction works if the **control** and **shift** keys are held down while scrolling the mouse wheel.

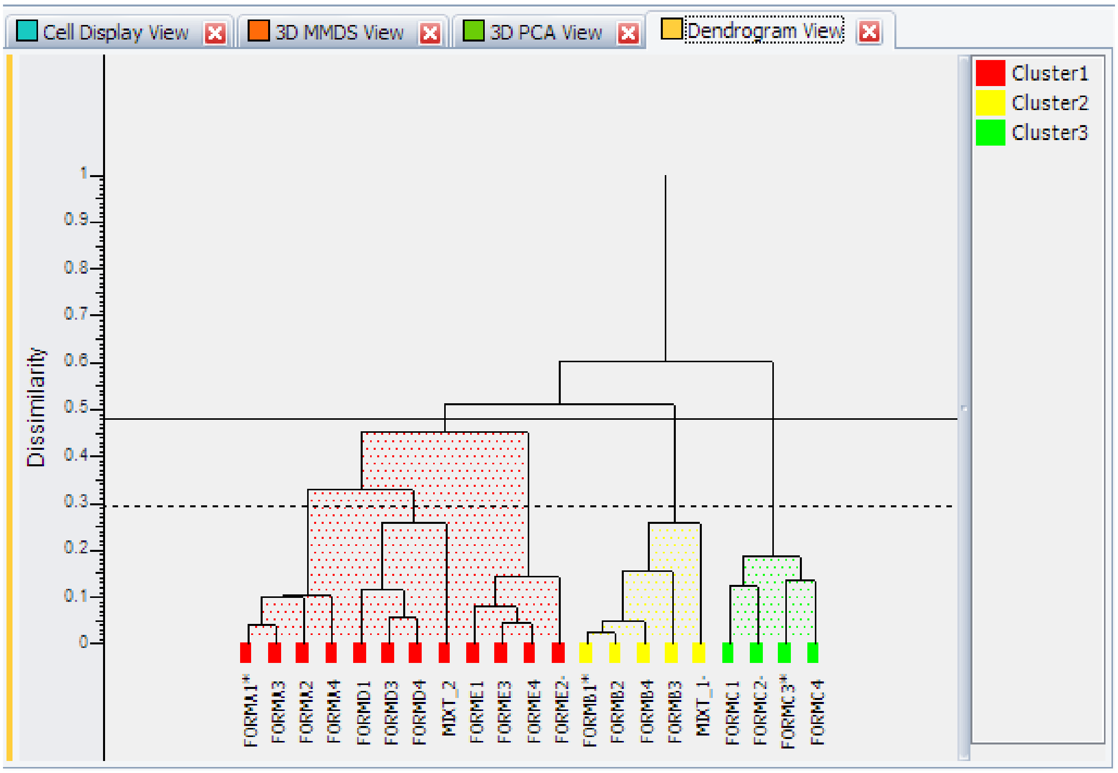


9. To move around the cell area and position the view more accurately, hold down the **Alt** key and **left mouse** button simultaneously while dragging with the mouse.
   - ➨ The display updates in real time.
   - ➨ These zoom and movement functions are the same for all graphical views within the program.
10. To check which of the plotted profiles correspond to which sample, hover the mouse pointer over part of the plotted line; a tooltip appears with the sample filename in it.
    - ➨ This closer view of the overlaid patterns makes it easier to see that the top pattern (MIXT2) is noticeably different from the other patterns - note the peak below 14° for example - and, despite being colored the same, may not actually belong to this group of compounds.

11. Reset the view with a **right click** of the mouse in the graph pane and selecting **Zoom Reset** from the pop-up menu.
    - ➨ There are a series of tabs on top of the **Cell Display View**, each of which display a different view of the data.
12. Select the tab labelled **Dendrogram View**.

The partitioning of the data into groups that were displayed in the colored cells is carried out by the cluster analysis of the sample data. In **Cluster Analysis** there are five different methods of clustering available, each of which tend to give slightly different results. The program computes the best, in the sense of the most internally consistent, dendrogram method and displays the results from that.

Note: The user can choose to view the results from the other methods and overrule the selected one if required. The default method is **Group Average Link**; selecting the best method automatically (**Automatically select best**) is not the default and can be slow so the use of the **Group Average Link** is recommended.



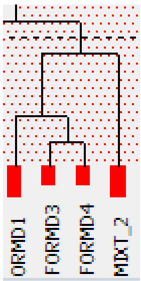
A dendrogram provides a visual display of the results from the hierarchical method of data classification using cluster analysis. The dendrogram itself takes the form of a tree-diagram in which each terminal branch (colored box) is representative of a single pattern sample.

The higher up the similarity scale two samples are connected by a horizontal line (called a 'tie-bar'), the less similar they are. Therefore, samples FORMA1 and A2, with a dissimilarity value of around 0.1 are very similar, whereas samples FORMA1 and FORMC1, which are only joined much further up the tree by a horizontal line with a dissimilarity value of approximately 0.6, are quite different.

In this dendrogram there are six separate clusters, each distinguished by its own color. These are the same colors displayed earlier in the cell display and throughout most of the other **Cluster Analysis Views**. The number of clusters are defined by the solid horizontal cut-line which in this case was initially set to 0.48. The calculation of this level is carried out *via* a number of statistics. The confidence level on this choice of cut position is shown by the dotted line. Selecting a cut-line for a dendrogram is a difficult procedure, and the results must be treated with caution. The program-calculated level therefore should always be carefully examined by the user to see if looks sensible.

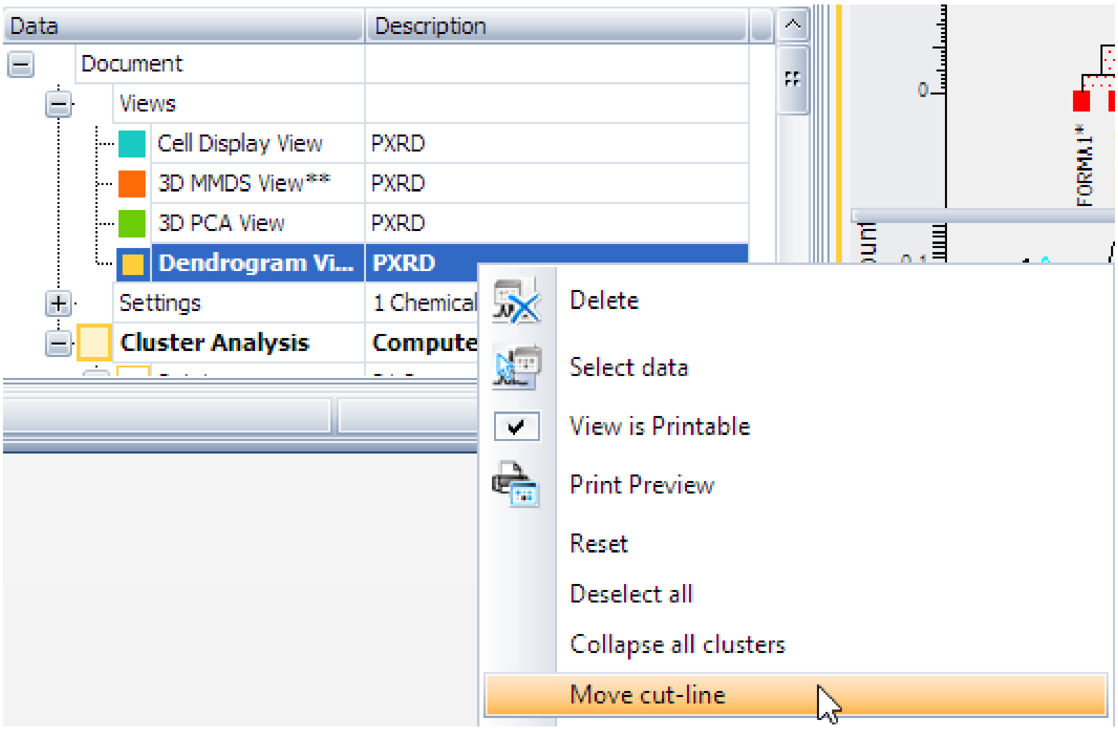
Adjusting the cut level upwards creates fewer separate clusters, and effectively reduces the discrimination between differences; adjusting the cut-level downwards creates more separate clusters.

1. Click on red square cell **FORMD1** in the **dendrogram graph**.
   - ➨ The pattern profile of the sample is displayed.
2. Using the **control** key and the left mouse button, also select cell **MIXT2**.
   - ➨ These are the two patterns that appeared different from one another when they were overlaid in the cell display.



- ➨ Looking at the position of the cut-line it is seen to be only very slightly above the similarity line between cell **MIXT2** and the other red cells. With the confidence levels indicated by the yellow lines being between 4 and 8 clusters, it is possible that the present level is not ideal.

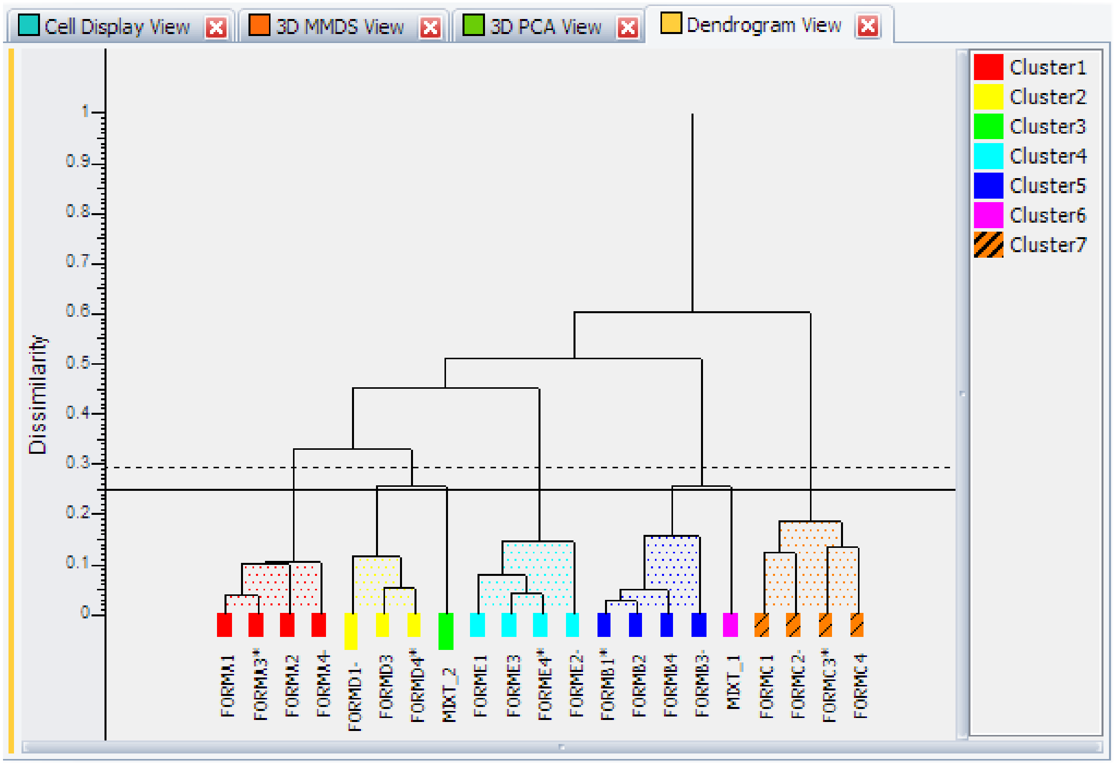
3. To manually adjust the cut-line, select the **Dendrogram View** in the **Data Tree**,
4. open the context menu with the **right mouse** button and
5. click on **Move cut-line**.
   - ➨ The dendrogram is now in **Move cut-line** mode. The mouse cursor in the **Dendrogram view** becomes a ring to indicate this mode.



Data	Description
Document	
Views	
Cell Display View	PXRD
3D MMDS View**	PXRD
3D PCA View	PXRD
Dendrogram Vi...	PXRD
Settings	1 Chemical
Cluster Analysis	Compute

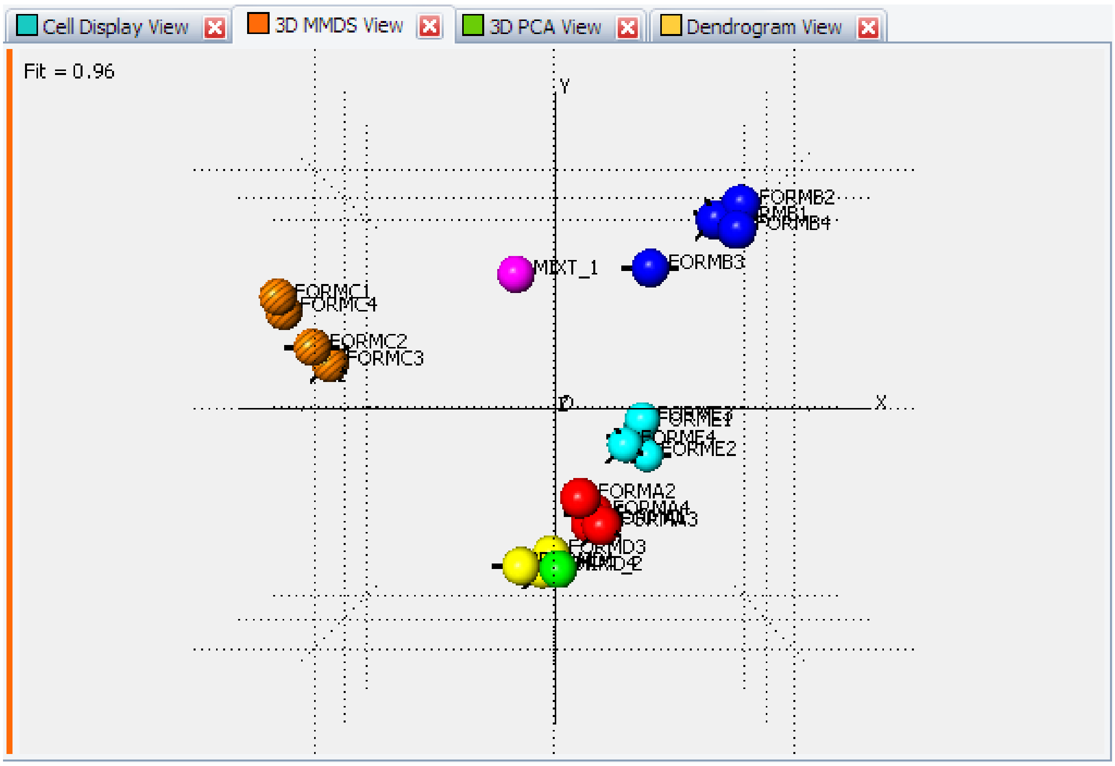
6. A left mouse button click moves the cut line to the new position.

7. Move the cut-line down slightly below the confidence line so there are now 7 different clusters.
   - ➨ Notice that the assigned colors change, and that **MIXT2** is now in a cluster of its own.
8. Click on the **Move cut-line** command in the **View's context** menu again to leave the **Move cut-line** mode.
   - ➨ A message box is displayed with the question if the cut-line should be saved. After answering yes the change in clustering is made permanent.
   - ➨ From the file names used in this demonstration example it can be seen that the samples are now all properly grouped, with the two mixtures separate from the rest. Notice in the dendrogram display that these two mixture patterns are quite dissimilar to anything else, having a low similarity connection value to their neighbors:



The differences between samples can also be viewed by making use of the distances derived from the correlation matrix to give a representation of the data in three dimensions. There are two methods used for calculating the resulting 3-D plots: these are **Metric Multi-Dimensional Scaling** (MMDS) and **Principle Component Analysis** (PCA). They both give different views of the samples because of the differences in calculation and will therefore give different results. Sometimes these differences can be substantial. The control of each view is the same so only the **3D MMDS View** will be described here in detail.

1. The **3D MMDS View** is displayed by clicking on its tab on top of all views:



- ➨ The initial view shows only the x and y axis, while the z axis lies in projection. Each sphere represents a sample. The position on the plot is taken from the MMDS calculation. The color of each sample is taken from the dendrogram display to allow easy comparison of the results from these different methods. Allowing the mouse to hover over a sample displays the sample label in a tooltip popup.

Samples that are similar are drawn close to each other, so they are seen to clump together in groups. Note that differently colored samples can also be close, this shows that they also have similarities. This can be seen by the yellow group being very close to the green sample, which is pattern MIXT2. When comparing with the dendrogram display, the MIXT2 sample is only separated from the yellow group by the current level of the cut-line.

Also notice the number plotted at the top-left of the display, in this case, 0.96. This is a correlation coefficient called the goodness of fit that measures the quality of the 3D representation to the original data. The closer to 1.0 it is, the greater the reliability of the results. This value tends to decrease when larger data sets are used. Use the following methods for exploring the 3-D plot:

Action	Description
Rotate the 3D plot	Drag while holding the **Shift** key and the left mouse button.
Move plot laterally	Drag while holding the **Alt** key and the left mouse button.
Alter the size of the spheres	Drag up or down while holding the **Control** key and the left mouse button.
Zoom	Hold left mouse button and draw a rectangle over area from top left to bottom right.
Unzoom	Hold left mouse button and draw a rectangle over area from bottom right to top left.
Select a sample	Click on the sphere with left mouse button.
Select multiple samples	Hold **Control** key and select spheres with the **left mouse** button.

Action	Description
Clear selection	Right click mouse and choose **Deselect all**.
Reset zoom and selection	Right click mouse and choose **Reset.**

The **3D views** are useful to spot patterns that are quite different from the others, as they tend to stand out on their own and are not easily grouped. Also, cases where the colors (from the dendrogram) and the positions (from the MMDS) of the samples appear to contradict each other are samples that should be looked at manually in more detail.

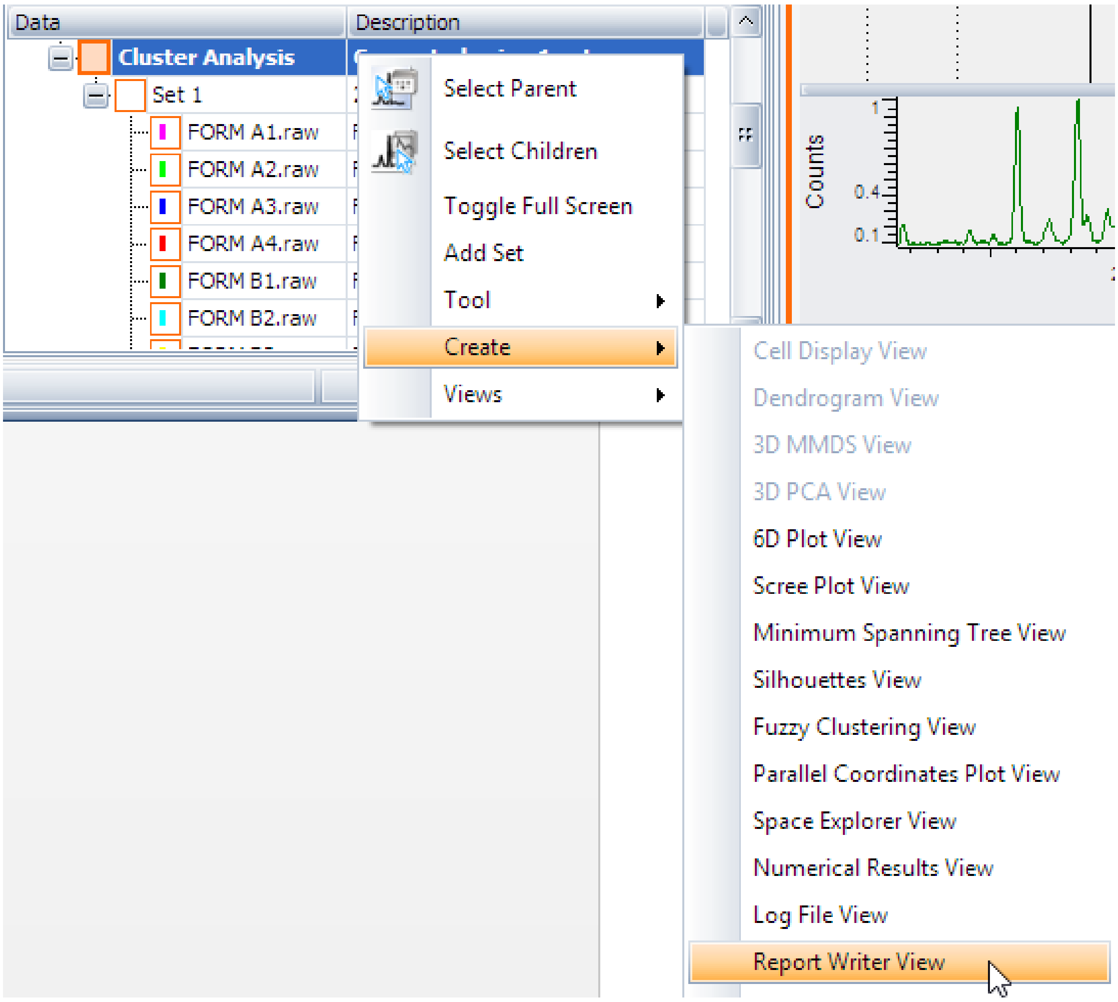
Finally, notice that some of the spheres have four spikes protruding from them; these show which samples are the **Most Representative Members** of that cluster. The smaller the distance, the tighter the cluster, and the more similar the samples are within it.

7.1.4 Reporting

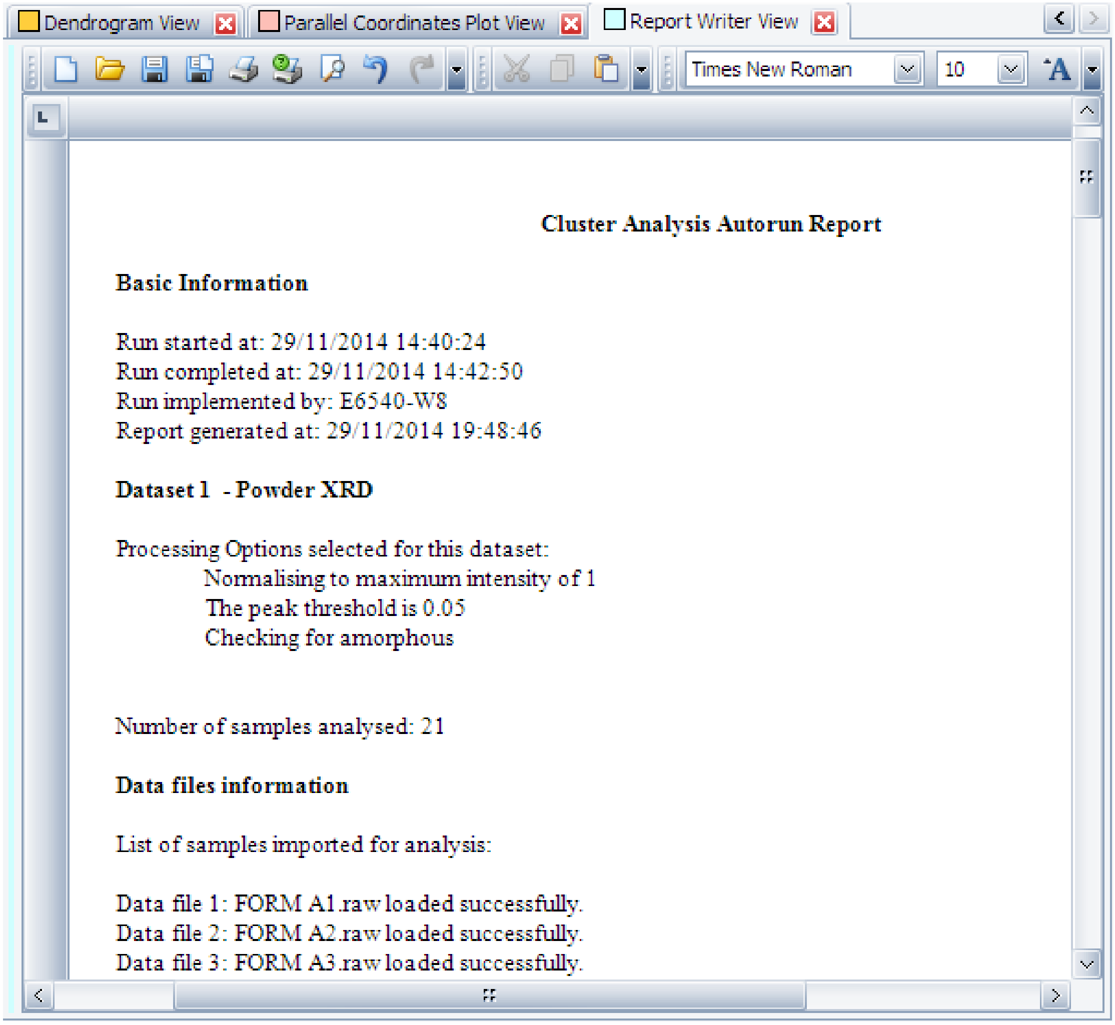
A report of the calculated results is automatically generated. This contains the information gained from the data and optionally some of the graphical views of the results that have been obtained.

To open the report,

1. select the **Cluster Analysis** node in the **Data Tree** and
2. click on **Create Report Writer View**:



➨ The Report Writer View is opened:



Cluster Analysis Autorun Report

Basic Information

Run started at: 29/11/2014 14:40:24
Run completed at: 29/11/2014 14:42:50
Run implemented by: E6540-W8
Report generated at: 29/11/2014 19:48:46

Dataset 1 - Powder XRD

Processing Options selected for this dataset:
Normalising to maximum intensity of 1
The peak threshold is 0.05
Checking for amorphous

Number of samples analysed: 21

Data files information

List of samples imported for analysis:

Data file 1: FORM A1.raw loaded successfully.
Data file 2: FORM A2.raw loaded successfully.
Data file 3: FORM A3.raw loaded successfully.

The report begins by detailing when the calculations were done and on which computer. It then continues with the settings used for the analysis and a summary of the results.

If selected in the **Cluster Analysis** settings in the **Settings dialog**, the output from all of the main graphic displays can be automatically included in this report.

The report may be edited, or additional information can also be added to the report manually, as in a standard word processor, by using the simple formatting tools supplied. For example, to add one of the other views to the report,

1. switch to the required view (e.g. **Parallel Coordinates Plot View**), and
2. click in the graphics area.
3. Push the **Control** and **C** key together to copy the graphics to the clipboard.
4. In the **Report Writer View**, click in the report at the point where you want the image to be inserted,
5. right-click and use the **Control-V** key combination to insert the graphics into the report.

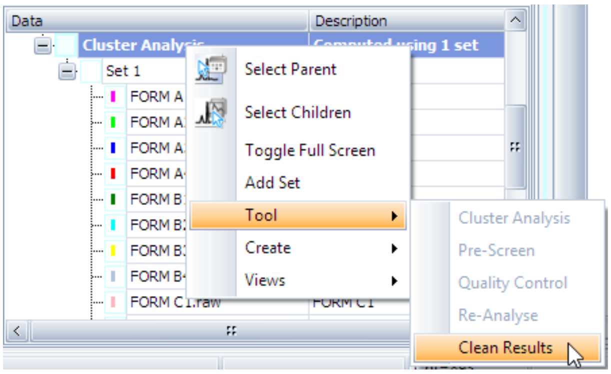
⇨ The report can be saved by using the **Save as....** toolbar button. The report can be saved in various formats, e.g. as an RTF file and can be opened and edited in any standard word processing package.

7.1.5 Cluster Analysis Using Reference Samples

The Cluster Analysis can use extra information to give a more advanced understanding of the dataset.

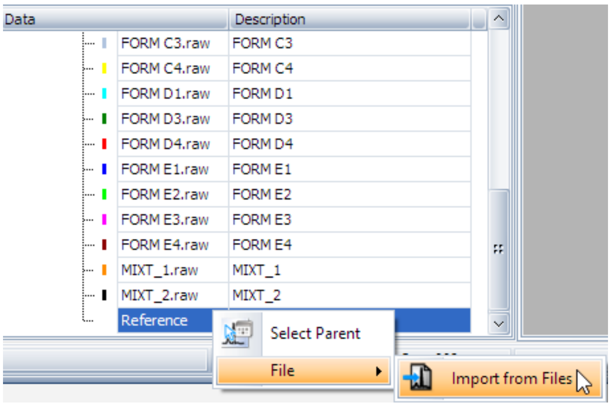
The current data set can be re-used but the analytical results must be deleted.

1. Use the **Clean Results** command in the **Cluster Analysis** node's context menu:



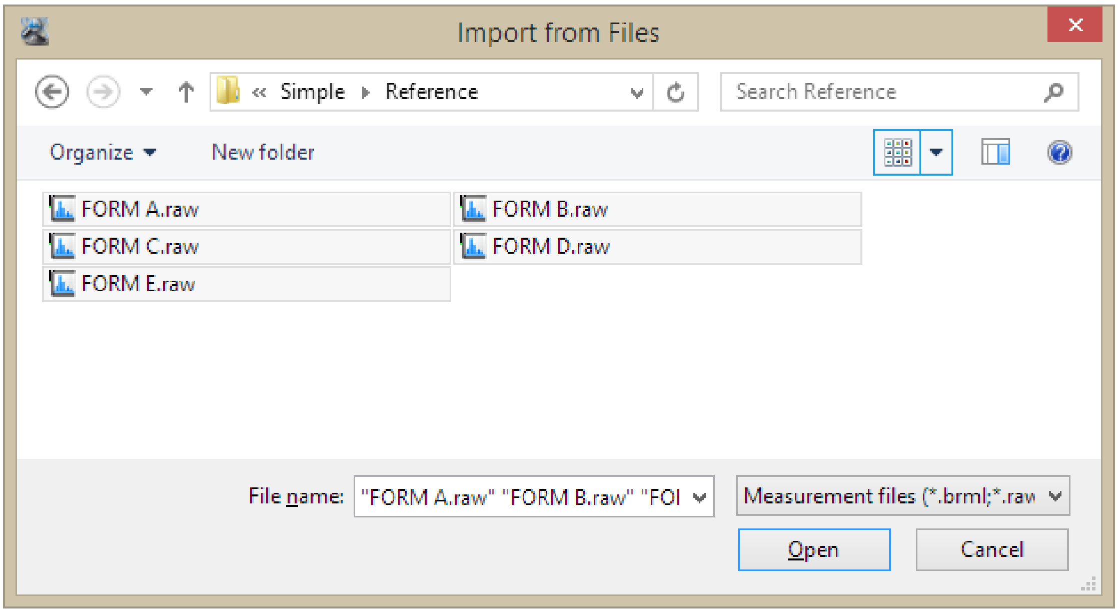
➨ **Clean Results** deletes all calculation results and removes all views which belong to the cluster analysis. The data remain intact.

2. To import reference data the **Reference** node at the end of the data list must be selected and the **/mport from Files** command must be clicked:



Data	Description
FORM C3.raw	FORM C3
FORM C4.raw	FORM C4
FORM D1.raw	FORM D1
FORM D3.raw	FORM D3
FORM D4.raw	FORM D4
FORM E1.raw	FORM E1
FORM E2.raw	FORM E2
FORM E3.raw	FORM E3
FORM E4.raw	FORM E4
MIXT_1.raw	MIXT_1
MIXT_2.raw	MIXT_2
Reference	

3. Go to the **Reference** sub-folder of the **Simple** tutorial folder in the **Import from Files** dialog and select all data:

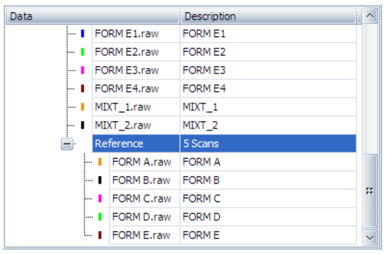


File name: "FORM A.raw" "FORM B.raw" "FOI

Measurement files (*.brml;*.raw

4. Click **Open** to import the data into the **Reference** node.

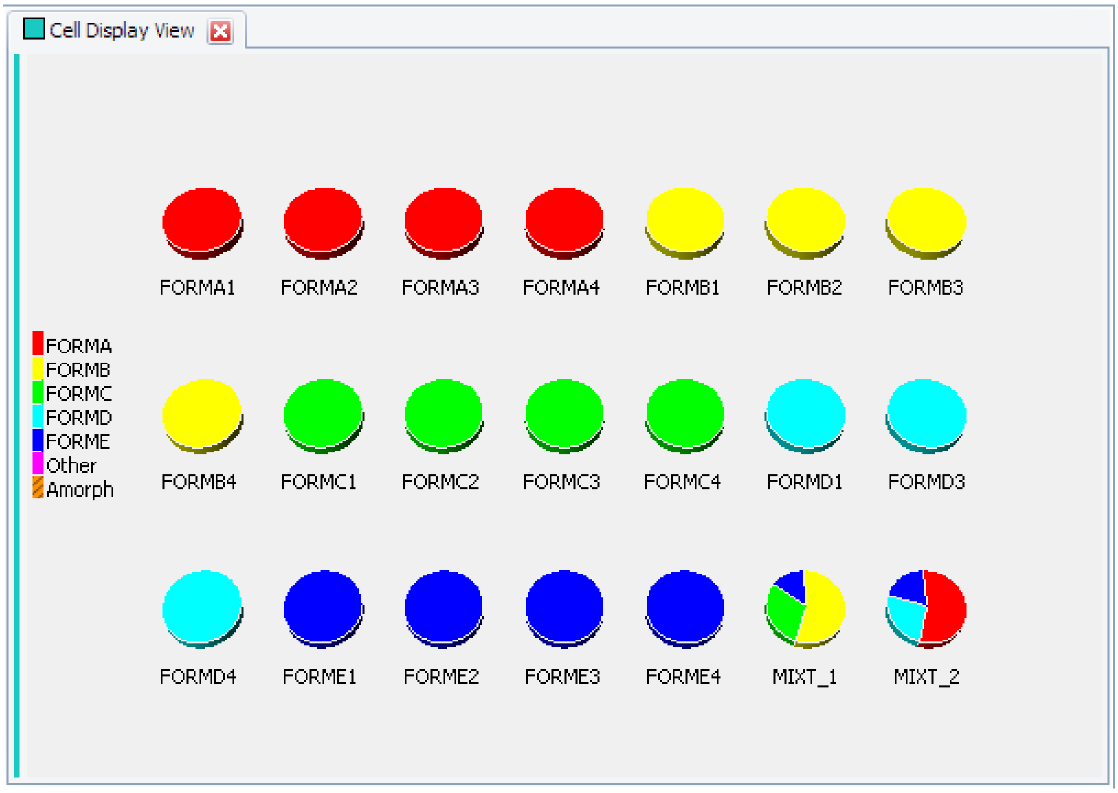
➨ The reference data are displayed in the tree:



Data	Description
FORM E1.raw	FORM E1
FORM E2.raw	FORM E2
FORM E3.raw	FORM E3
FORM E4.raw	FORM E4
MIXT_1.raw	MIXT_1
MIXT_2.raw	MIXT_2
Reference	5 Scans
FORM A.raw	FORM A
FORM B.raw	FORM B
FORM C.raw	FORM C
FORM D.raw	FORM D
FORM E.raw	FORM E

5. The cluster analysis must be repeated using the same steps as described above in the **Run the Cluster Analysis** section.

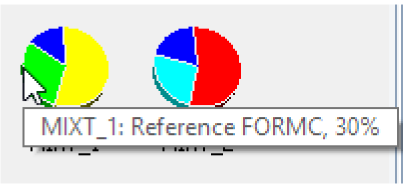
⇨ After the analysis has been performed the results open with a new **Cell Display View**:



By including known reference samples three important changes have been made to the cell display.

The reference samples that were provided to the program are known patterns of pure phases that the samples in the dataset are to be compared to. Because of this the cells in the cell display are now colored according to the known phases that they match and not according to the results from the dendrogram.

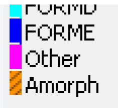
The presence of reference files has a more profound effect on the mixture samples:



MIXT_1: Reference FORMC, 30%

Instead of being assigned to a particular cluster they are now analyzed using the known phases as a reference, and how much of each known, pure phase present in each mixture is calculated. The cell is now displayed as a pie chart, divided up and colored according to the percentage composition estimated using the full pattern profiles without *Rietveld refinement*. Holding the mouse cursor over a slice of the pie chart will open a tooltip box displaying the phase, and percentage that the slice represents.

Finally, there are two extra keys that have been added onto the cell display legend:



FORMD
FORME
Other
Amorph

Each known phase has its own entry, however, there is now also a key called **Other**. Any pattern that does not match any of the known phases provided above a specified similarity cut-off will be assigned here.

Note: The **Other** group is not a group of similar patterns, but a collection of patterns that do not match the standard reference phases. However none of the samples in our dataset are in this group as they all match one of the phases, apart from the mixtures.

1. Open the **Log File View** to inspect the quantitative results.
2. Scroll down through the information to the section headed **Quantitative Analysis**.

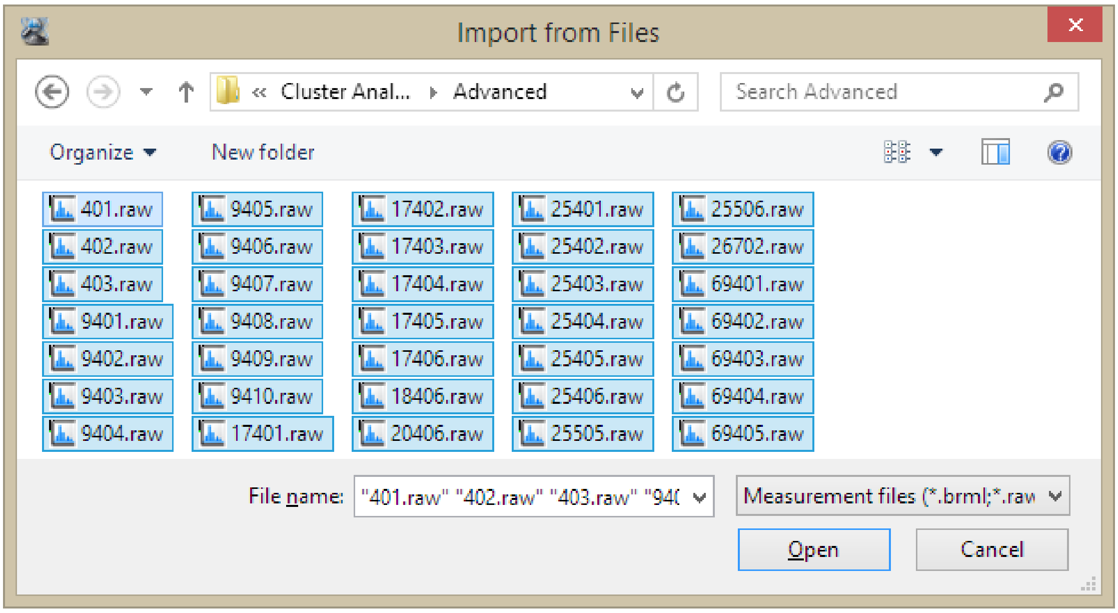
⇨ The numeric output from the analysis can be found here.

7.2 Amorphous Content and Pattern Shifts

This part will cover some of the more advanced options available that include pattern matching while allowing for a 2θ-shift, and automatic identification of non-crystalline samples.

7.2.1 Import Data and Run the Cluster Analysis

1. Start a new analysis run by using the **Clear** command on the **Data Tree's Document** node.
2. Select **Set 1** and import the data files from the Tutorial's **Advanced** sub-folder:



401.raw
402.raw
403.raw
9401.raw
9402.raw
9403.raw
9404.raw
9405.raw
9406.raw
9407.raw
9408.raw
9409.raw
9410.raw
17401.raw
17402.raw
17403.raw
17404.raw
17405.raw
17406.raw
18406.raw
20406.raw
25401.raw
25402.raw
25403.raw
25404.raw
25405.raw
25406.raw
25505.raw
25506.raw
26702.raw
69401.raw
69402.raw
69403.raw
69404.raw
69405.raw

File name: "401.raw" "402.raw" "403.raw" "94(

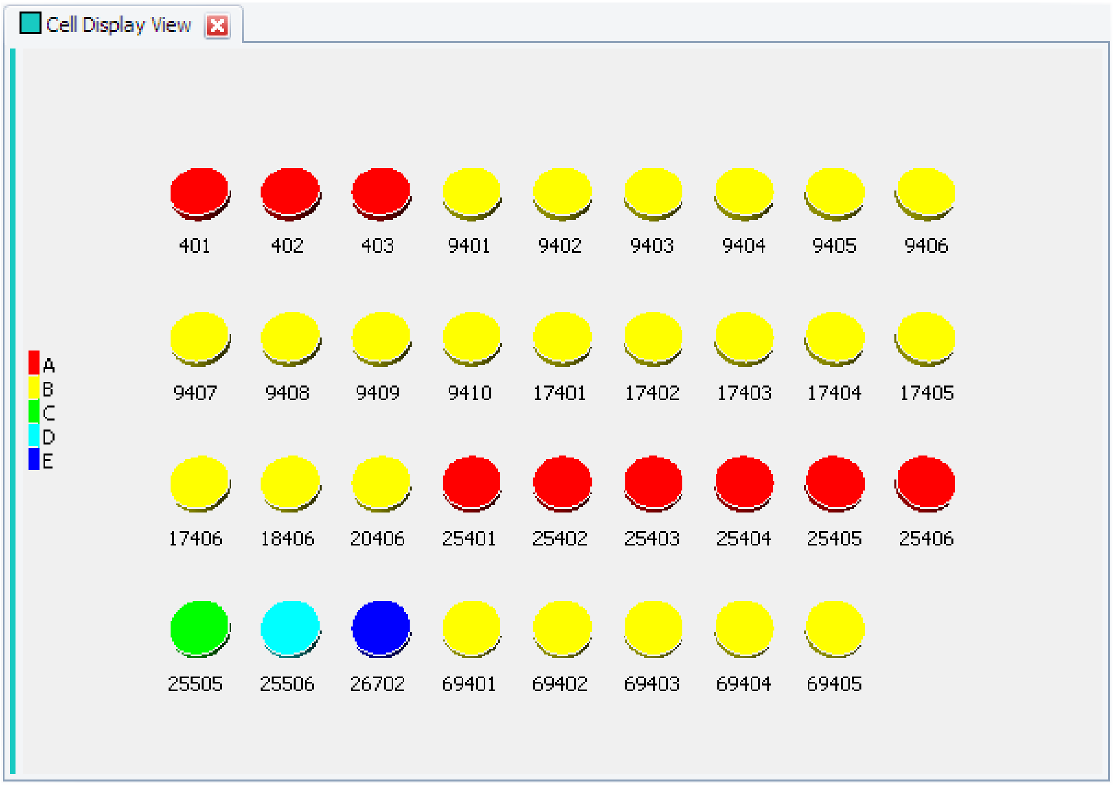
Measurement files (*.brml;*.raw

3. After importing the data start the analysis by using the **Cluster Analysis** command on the **Cluster Analysis** node.
4. Leave the **Advanced Options** settings at their default values and click **OK**.

⇨ After the pattern files are loaded, they are checked for amorphous content, matched with one another, and after that undergo a cluster analysis. This may take a little time to complete, but once finished the results views will appear with an initial view of the **Dendrogram** and **Cell Display** if configured.

7.2.2 Analysis of the Results

The 35 samples contained within the database are now presented in the **Cell Display View:**



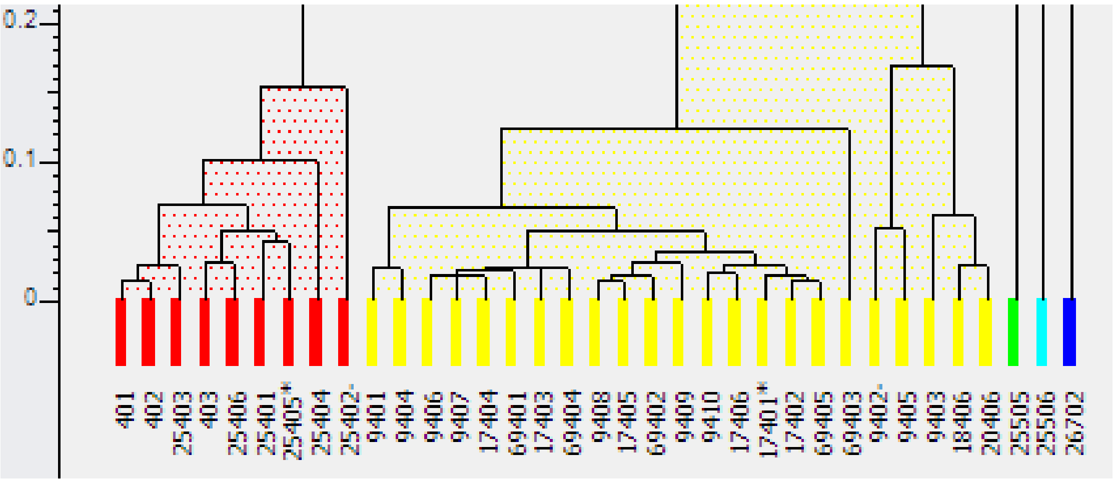
The program is using the last few digits of the filename of each sample to label them. There are three patterns, 25505, 25506 and 26702, which do not belong to any cluster.

Opening and checking the Log File View reveals:

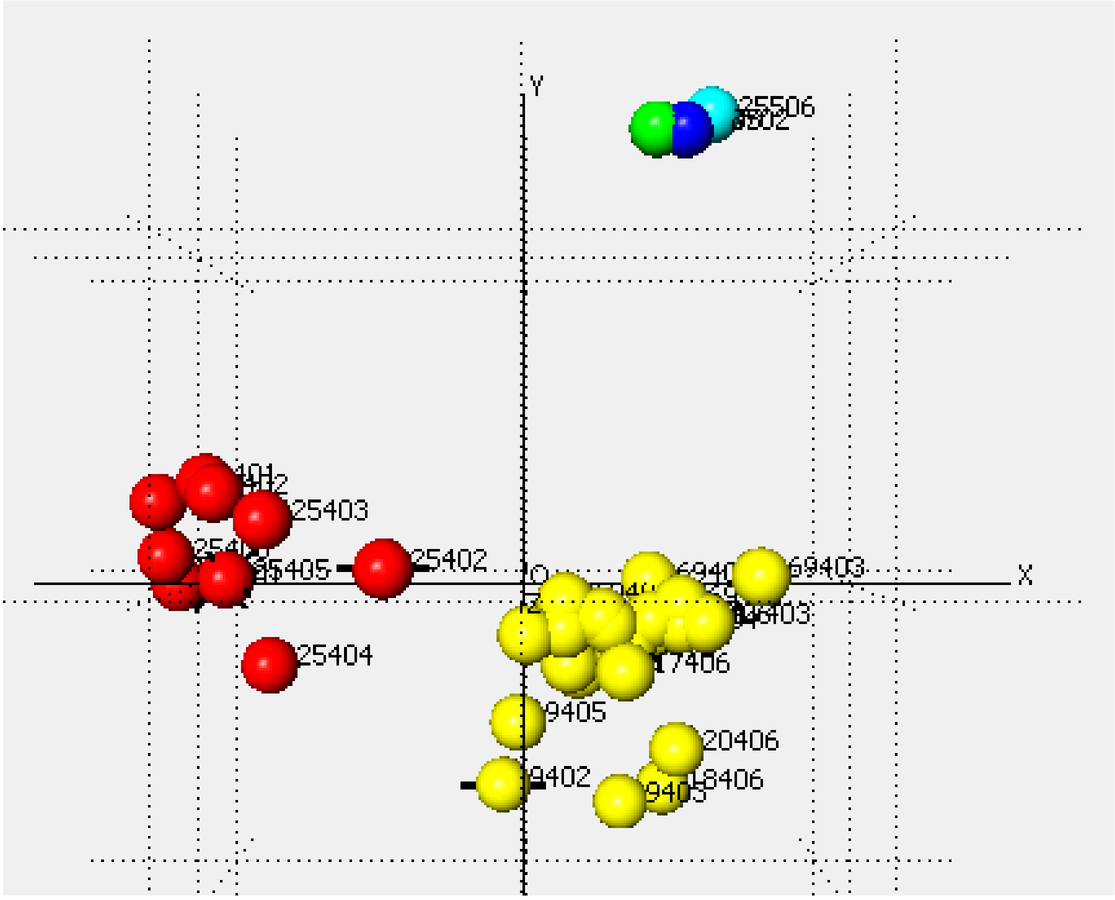
```
Sample          has only 0 marked peaks, and an amorphous
25505.raw       indicator of 99.2 - Probably amorphous.

Sample          has only 0 marked peaks, and an amorphous
25506.raw       indicator of 98.8 - Probably amorphous.

Sample          has only 0 marked peaks, and an amorphous
26702.raw       indicator of 100.0 - Probably amorphous.

3 samples were flagged as possibly non-crystalline.
```

Identification of such amorphous samples is done on the basis of checking to see if any signal (corresponding to peaks) would be left after subtraction of the entire amorphous hump. The method tends to take a conservative approach.

Looking at the Dendrogram View tab:



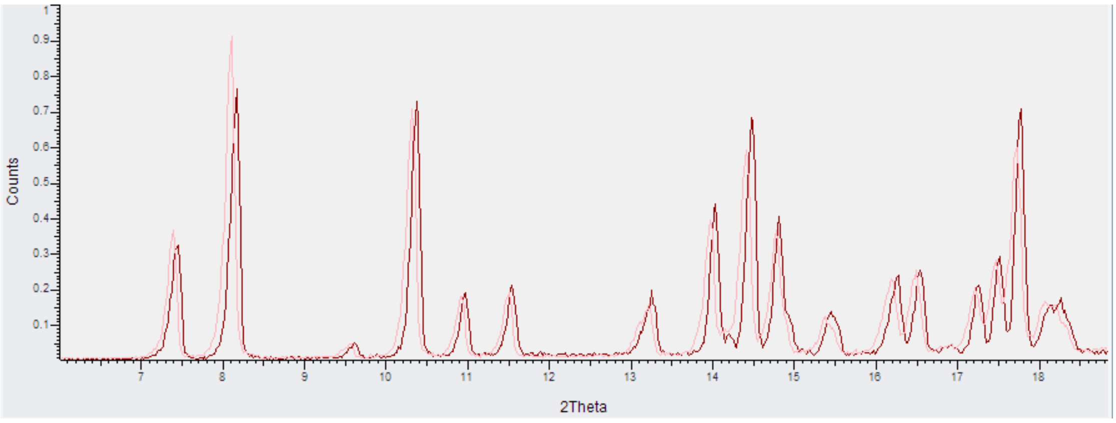
It is seen that the three non-crystalline patterns are placed on the far right of the diagram with a zero similarity to the rest. This is deliberate, in order to remove them from the main clusters. Compare this to the 3D (MMDS) plot:



Again, the three non-crystalline samples are quite separate from the rest of the patterns. Both the dendrogram and 3D plots suggest a loose grouping of the rest of the patterns, suggesting there may be some differences between them.

1. Using the **Control** key, click two patterns which are on opposite sides of the dendrogram, for example patterns 17401 and 25402.

   ➨ Examining their profiles, it appears that in addition to some preferred orientation issues, there seems to be a noticeable 2θ-shift between the otherwise relatively similar profiles:



- ➨ This completes the initial analysis of the data, but the 2θ-shift in some of the samples could be examined in more detail.

2. Prepare the re-analysis run by using the **Clean Results** command on the **Cluster Analysis** node.

7.2.3 Reprocessing the Data Allowing for an X-Shift

When collecting powder diffraction data from a diffractometer the sample or instrument alignment can result in linear or non-linear shifts along the x-axis of the resulting pattern. This can especially be a problem if the sample height varies from sample to sample, giving rise to systematic errors in the pattern matching unless it is accounted for. However, to allow for this is a time consuming process and should therefore not be used unless such a shift is suspected - it is switched off by default.

A general expression for the shift is:

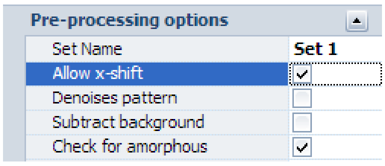
$$\Delta(2\theta) = \cos\theta\, a_0 + a_1 \sin\theta$$

where the a_0 coefficient corresponds to a linear (zero-point) shift described earlier, and the a_1 coefficient a non-linear component. The requirement is to find values of a_0 and a_1 that result in a maximum matching correlation result between two patterns.

The same data as in the previous run will now be examined again. Unlike the last time, the program will vary the x-offset parameters to attempt to maximize the match result.

1. Keep the same input data (**Advanced** folder) as used before in the previous run of the Cluster Analysis.

2. After importing the data start the analysis by using the **Cluster Analysis** command on the **Cluster Analysis** node.
3. Leave the **Advanced Options** settings at their default values and click **OK**.
4. Make sure that the **/nclude reference** per dataset option is not selected.
5. Open the **Advanced Options** window and turn on the **Allow x-shift calculation** (**sin theta**) checkbox.



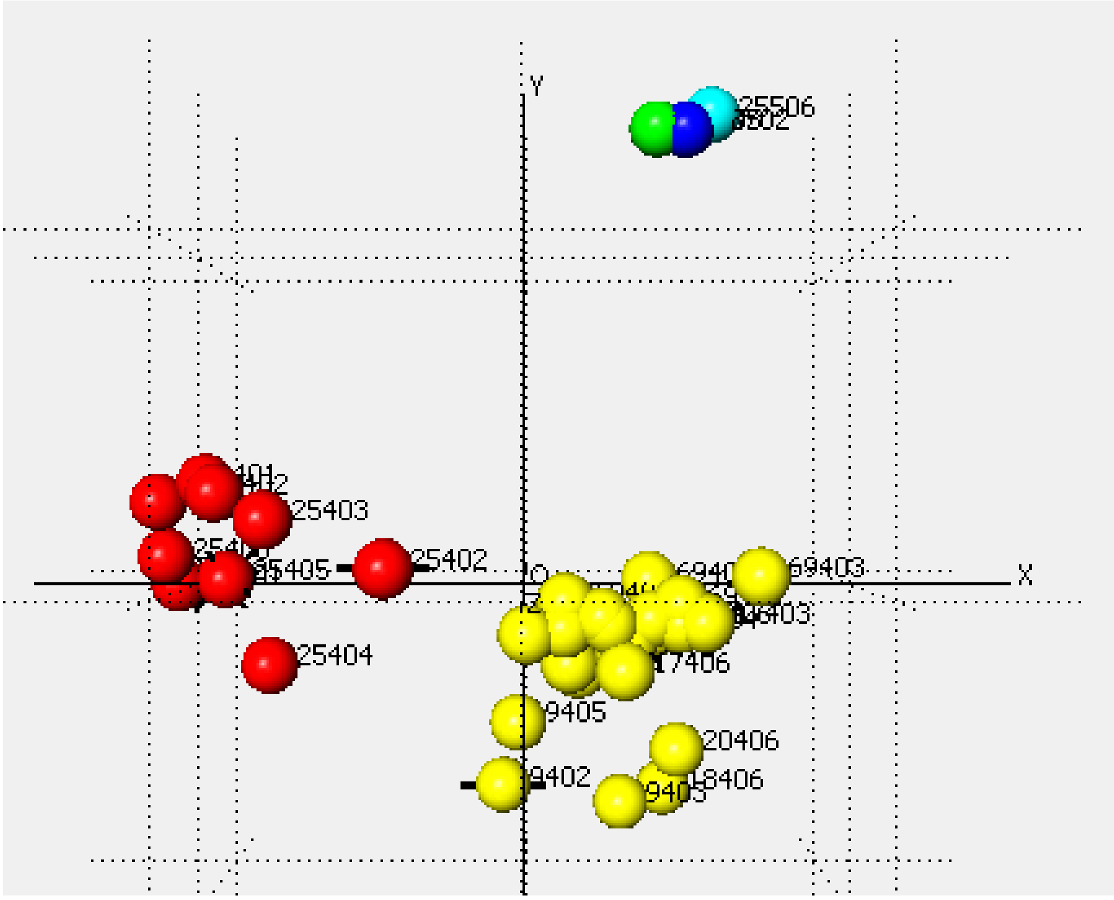
Pre-processing options

Set Name	Set 1
Allow x-shift	☑
Denoises pattern	□
Subtract background	□
Check for amorphous	☑

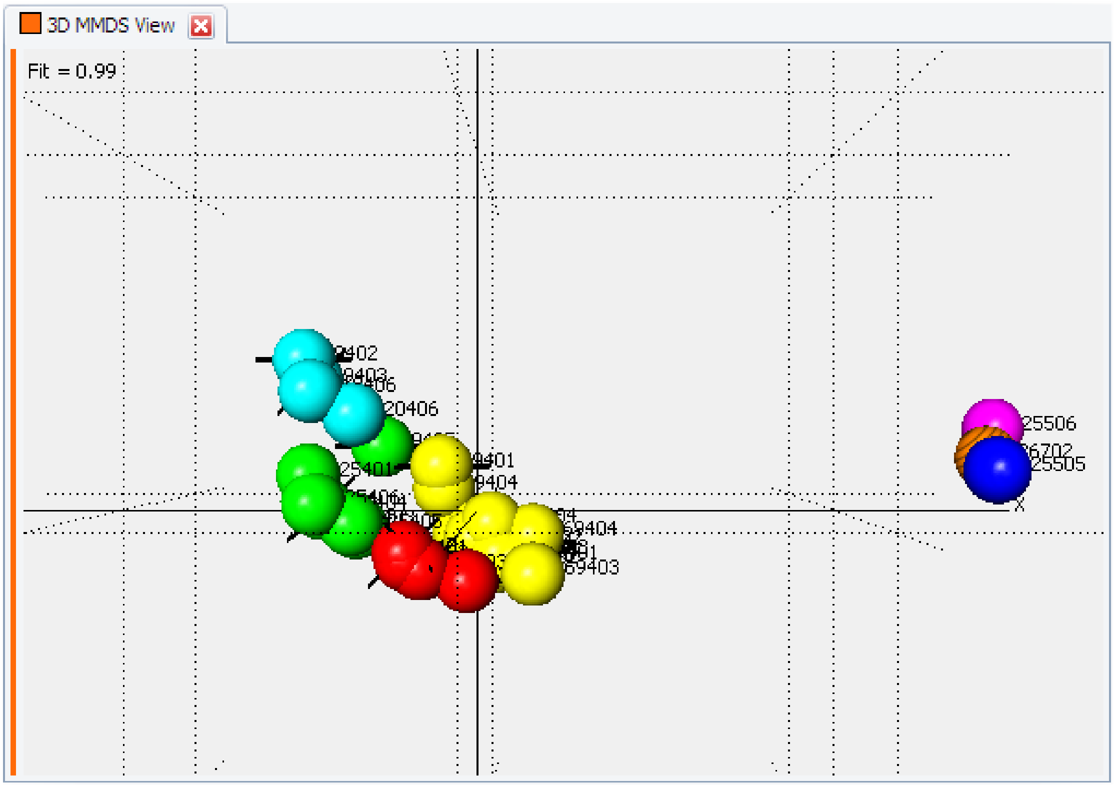
6. Close the **Advanced Options** windows and click **OK** in the **Run Cluster Analysis** dialog.

- ➨ The same data as before is now analyzed allowing x-shifts on the patterns. The time required to process this will take longer than a normal **Cluster Analysis** run.

Now look at the 3D MMDS plot. Previously, it looked like this:

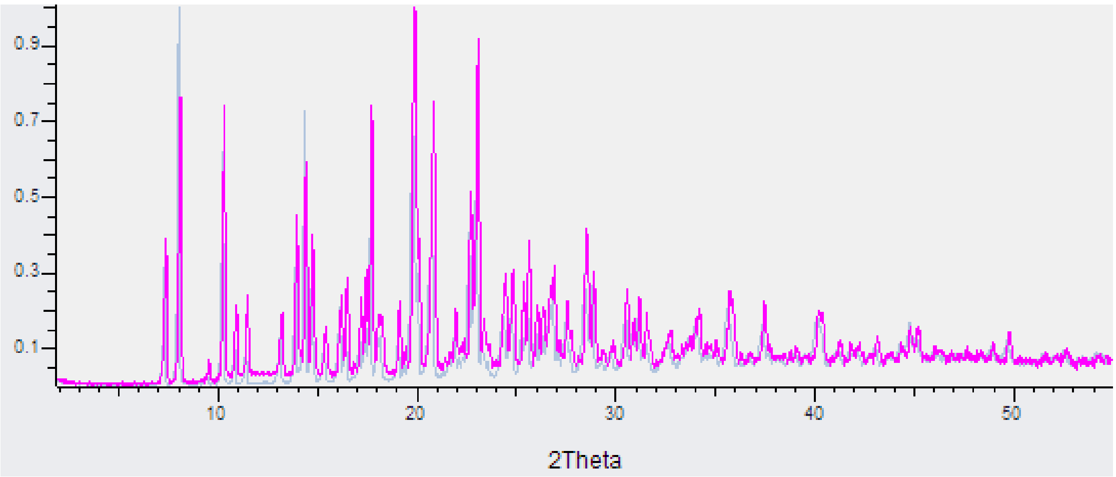


Now, with the option to calculate the best-offset value for each pattern turned on, it looks like this:

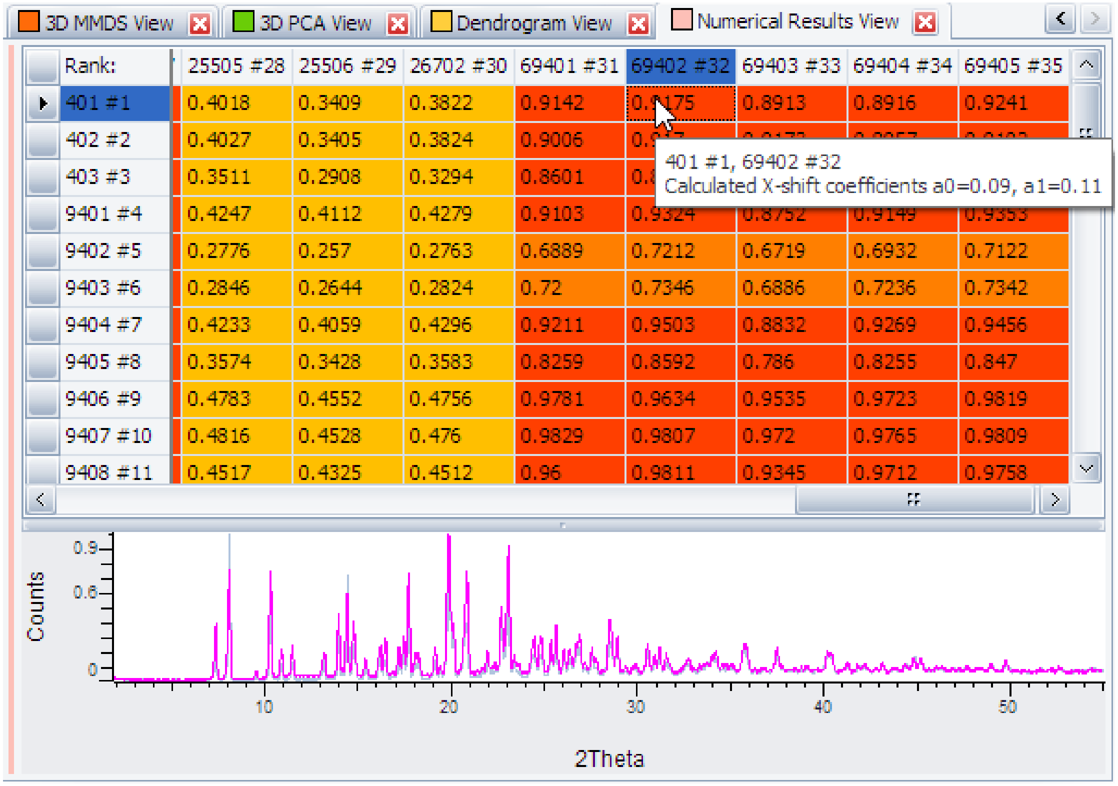


The three non-crystalline patterns are still quite separate, but the rest of the patterns have condensed together as a result of allowing for the 2θ-shift, showing that a large part of the differences between the pattern profiles was due to variation in sample heights during data collection. The program still separates them out within this grouping due to the preferred orientation issues.

Similarly with the dendrogram display, the similarity values between the patterns are much improved. The remaining differences appear to arise from preferred orientation effects, which are quite noticeable in some cases. For example, overlay the profiles of samples 401 and 69402:



To see how much the patterns have been shifted as part of the calculation, open the **Numerical Results View.** Locate 401.txt in the list on the left hand side, and read along to find where this row crosses the 69402.txt column:



Rank:	25505 #28	25506 #29	26702 #30	69401 #31	69402 #32	69403 #33	69404 #34	69405 #35
401 #1	0.4018	0.3409	0.3822	0.9142	0.9175	0.8913	0.8916	0.9241
402 #2	0.4027	0.3405	0.3824	0.9006	[illegible]	[illegible]	[illegible]	[illegible]
403 #3	0.3511	0.2908	0.3294	0.8601	[illegible]	[illegible]	[illegible]	[illegible]
9401 #4	0.4247	0.4112	0.4279	0.9103	0.9324	0.8752	0.9149	0.9353
9402 #5	0.2776	0.257	0.2763	0.6889	0.7212	0.6719	0.6932	0.7122
9403 #6	0.2846	0.2644	0.2824	0.72	0.7346	0.6886	0.7236	0.7342
9404 #7	0.4233	0.4059	0.4296	0.9211	0.9503	0.8832	0.9269	0.9456
9405 #8	0.3574	0.3428	0.3583	0.8259	0.8592	0.786	0.8255	0.847
9406 #9	0.4783	0.4552	0.4756	0.9781	0.9634	0.9535	0.9723	0.9819
9407 #10	0.4816	0.4528	0.476	0.9829	0.9807	0.972	0.9765	0.9809
9408 #11	0.4517	0.4325	0.4512	0.96	0.9811	0.9345	0.9712	0.9758

401 #1, 69402 #32
Calculated X-shift coefficients a0=0.09, a1=0.11

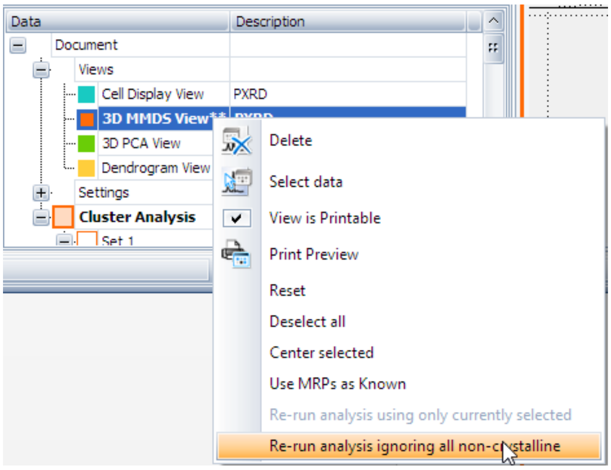
Clicking on the rank value shows the two profiles overlaid in the bottom pane; hovering the mouse over this value shows a tooltip with the calculated values of a_0 and a_1 for this pair of patterns.

The **Report Writer View** will note the use of an x-offset calculation in the output, as the identification of the non-crystalline samples.

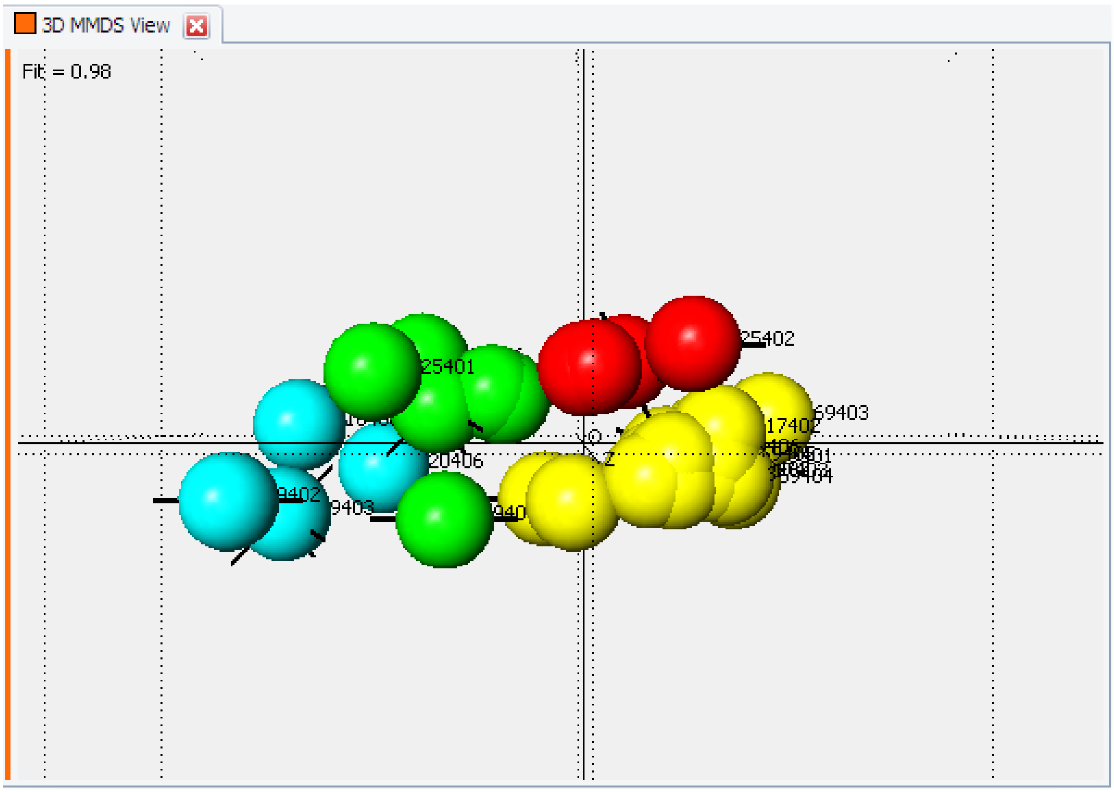
7.2.4 Remove Amorphous Samples from the Analysis

In cases where there is a larger proportion of amorphous samples in a dataset, it may be helpful to remove these samples and re-run the analysis without them to give a clearer idea of what is going on.

This can be done automatically by using the **Re-Run Analysis Ignoring Non-Crystalline** command which is available in the view node's context menu in the data tree:



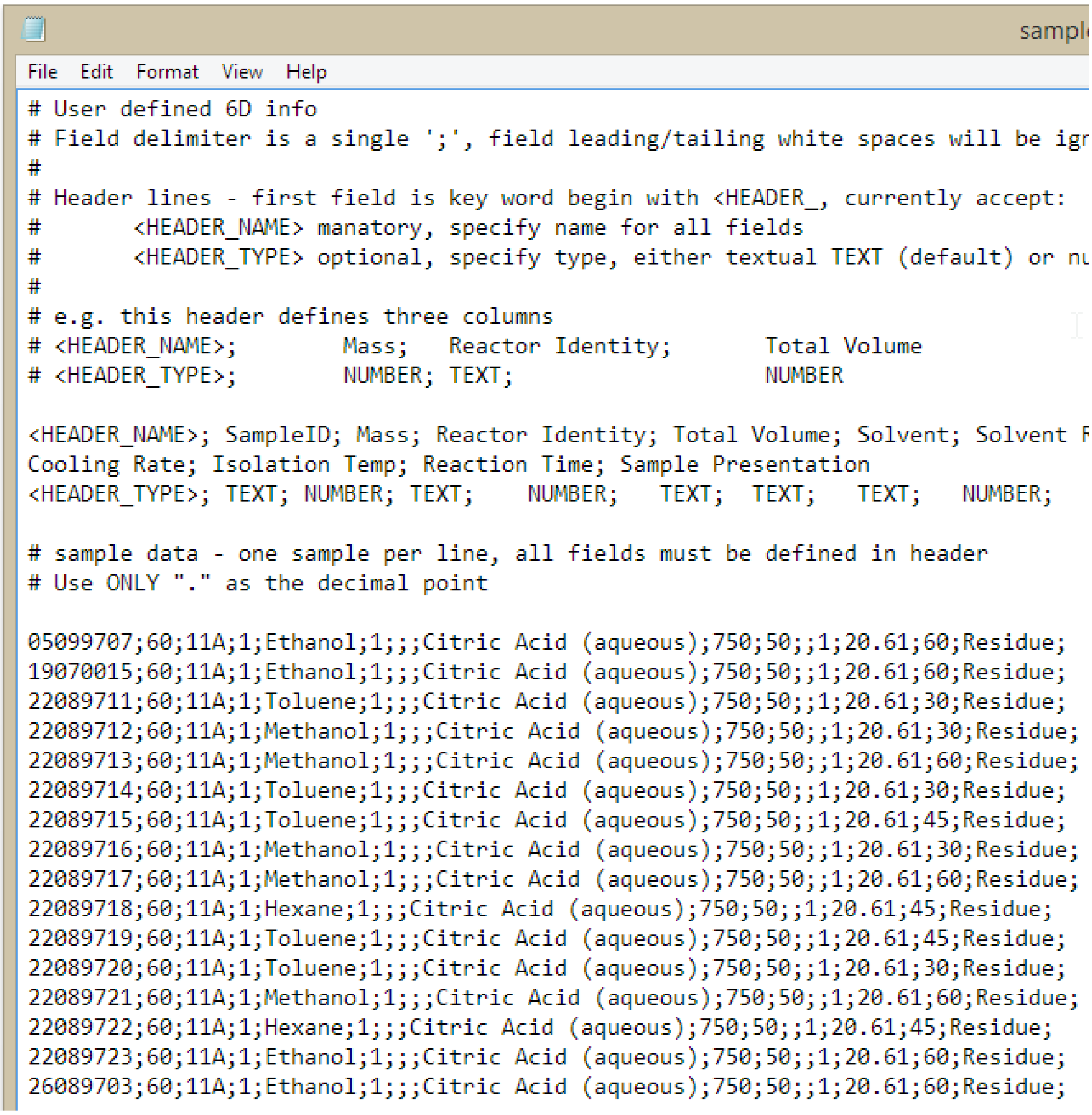
After re-calculation there are no more amorphous samples visible in the 3D plots and the dendrogram:



7.3 6D Plot

There are a number of methods for analyzing groups of patterns, some of which were detailed in the previous examples. However, these methods only concerned the powder diffraction pattern itself and not any other associated data that may be available. For example, this additional information can include sample preparation details such as the solvent used, concentration, pH, volume, temperature, etc., and up to three of these can be incorporated into the 6D plot at one time.

1. To load the data to be analyzed, select the **Set 1** and
2. click the **Import from Files…** command in the context menu.
   ➨ The **Import from Files** dialog opens and allows selecting the data.
3. Navigate to the folder *C:\ProgramData\Bruker AXS\Tutorial\Cluster Analysis\6D-Plot,*
4. select all RAW files in the folder and
5. click the **Open** button.
6. Open the file *C:\ProgramData\Bruker AXS\Tutorial\Cluster Analysis\6D-Plot\sample_6dinfo_tutorial.txt* in **Notepad** to display the additional sample information:

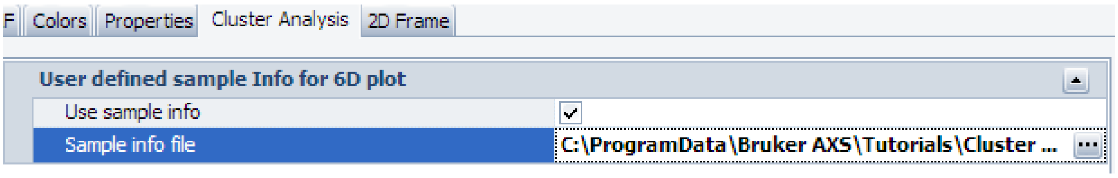


```
# User defined 6D info
# Field delimiter is a single ';', field leading/tailing white spaces will be igr
#
# Header lines - first field is key word begin with <HEADER_, currently accept:
#       <HEADER_NAME> manatory, specify name for all fields
#       <HEADER_TYPE> optional, specify type, either textual TEXT (default) or nu
#
# e.g. this header defines three columns
# <HEADER_NAME>;        Mass;    Reactor Identity;       Total Volume
# <HEADER_TYPE>;        NUMBER; TEXT;                    NUMBER

<HEADER_NAME>; SampleID; Mass; Reactor Identity; Total Volume; Solvent; Solvent F
Cooling Rate; Isolation Temp; Reaction Time; Sample Presentation
<HEADER_TYPE>; TEXT; NUMBER; TEXT;    NUMBER;    TEXT;   TEXT;    TEXT;    NUMBER;

# sample data - one sample per line, all fields must be defined in header
# Use ONLY "." as the decimal point

05099707;60;11A;1;Ethanol;1;;;Citric Acid (aqueous);750;50;;1;20.61;60;Residue;
19070015;60;11A;1;Ethanol;1;;;Citric Acid (aqueous);750;50;;1;20.61;60;Residue;
22089711;60;11A;1;Toluene;1;;;Citric Acid (aqueous);750;50;;1;20.61;30;Residue;
22089712;60;11A;1;Methanol;1;;;Citric Acid (aqueous);750;50;;1;20.61;30;Residue;
22089713;60;11A;1;Methanol;1;;;Citric Acid (aqueous);750;50;;1;20.61;60;Residue;
22089714;60;11A;1;Toluene;1;;;Citric Acid (aqueous);750;50;;1;20.61;30;Residue;
22089715;60;11A;1;Toluene;1;;;Citric Acid (aqueous);750;50;;1;20.61;45;Residue;
22089716;60;11A;1;Methanol;1;;;Citric Acid (aqueous);750;50;;1;20.61;30;Residue;
22089717;60;11A;1;Methanol;1;;;Citric Acid (aqueous);750;50;;1;20.61;60;Residue;
22089718;60;11A;1;Hexane;1;;;Citric Acid (aqueous);750;50;;1;20.61;45;Residue;
22089719;60;11A;1;Toluene;1;;;Citric Acid (aqueous);750;50;;1;20.61;45;Residue;
22089720;60;11A;1;Toluene;1;;;Citric Acid (aqueous);750;50;;1;20.61;30;Residue;
22089721;60;11A;1;Methanol;1;;;Citric Acid (aqueous);750;50;;1;20.61;60;Residue;
22089722;60;11A;1;Hexane;1;;;Citric Acid (aqueous);750;50;;1;20.61;45;Residue;
22089723;60;11A;1;Ethanol;1;;;Citric Acid (aqueous);750;50;;1;20.61;60;Residue;
26089703;60;11A;1;Ethanol;1;;;Citric Acid (aqueous);750;50;;1;20.61;60;Residue;
```

The information for the individual samples are given as defined in the header in one line per sample.

This file must be altered for other samples and processing parameters according to your requirements.

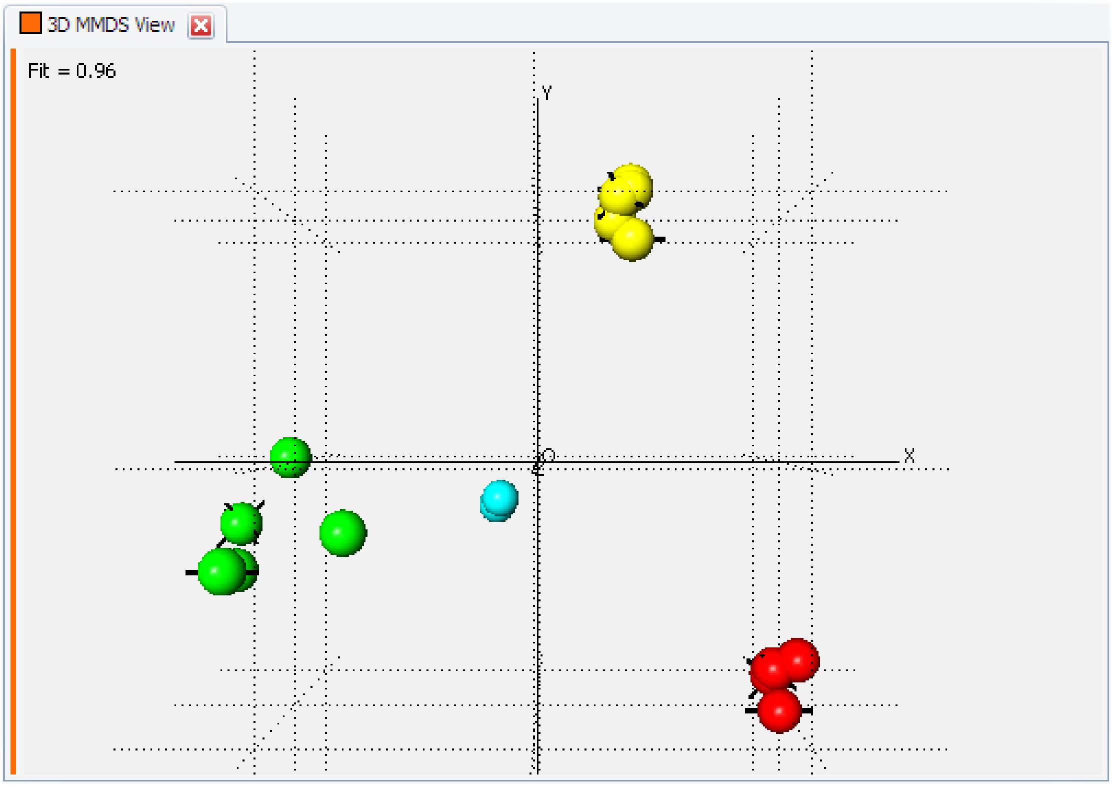
1. Make sure that the file *C:\ProgramData\Bruker AXS\Tutorial\Cluster Analysis\6D-Plot\sample_6dinfo_tutorial.txt* is configured in the **Settings** dialog, **Cluster Analysis** tab, group **User defined sample info for 6D plot**:



User defined sample Info for 6D plot

Use sample info	☑
Sample info file	C:\ProgramData\Bruker AXS\Tutorials\Cluster ...

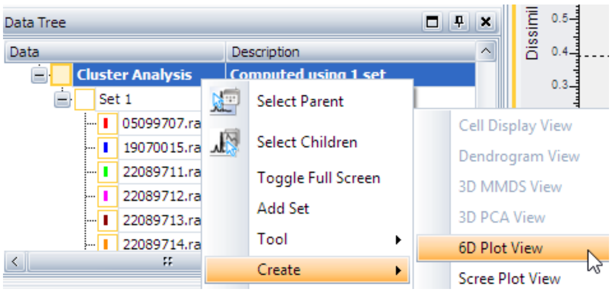
2. Select the **Cluster Analysis** node in the data tree and click on the **Cluster Analysis** command in the context menu. The **Run Cluster Analysis** dialog opens.
3. Accept all default values and click the **OK** button.
   - ➨ The analysis is carried out and the default views are opened.
4. Select the 3D MMDS view:



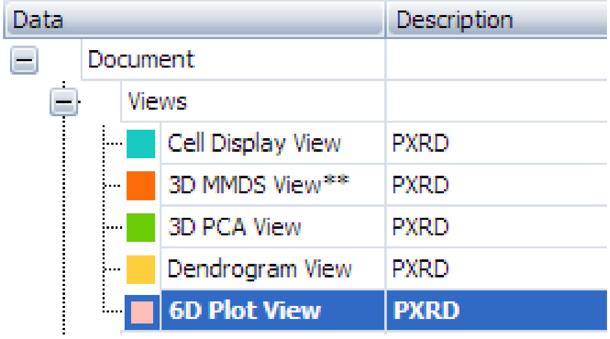
This view displays several distinct main clusters with the cells spread out along the x-axis. By selecting the cells individually, and examining the additional sample information, it can be seen that there appears to be a correlation between the solvent used in preparing the samples, and the different groupings of patterns. For example, all of the red samples are ones that were prepared with ethanol as solvent.

This information can be viewed more easily in a graphical manner, by enhancing the 3D-plot by plotting additional information on it, in the form of extra dimensions. For example, we can plot the different types of solvent used in terms of different shapes, and different reaction times in terms of different sizes of each plotted point.

1. Select the **Cluster Analysis** node and use the **Create 6D Plot View** command to open the 6D plot.

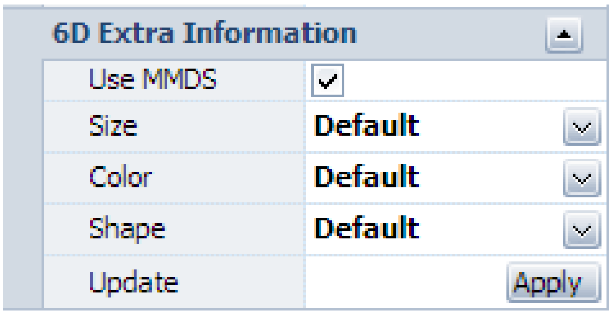


2. Select the **6D Plot View** node in the **Views** list:



Data	Description
Document	
Views	
🟦 Cell Display View	PXRD
🟧 3D MMDS View**	PXRD
🟩 3D PCA View	PXRD
🟨 Dendrogram View	PXRD
🟥 6D Plot View	PXRD

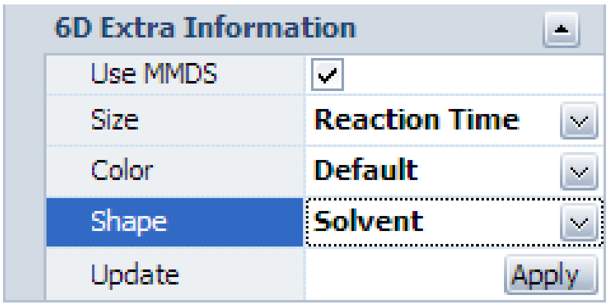
➨ The **6D Plot View's** properties are displayed in the property grid:



6D Extra Information

Use MMDS	☑
Size	Default
Color	Default
Shape	Default
Update	Apply

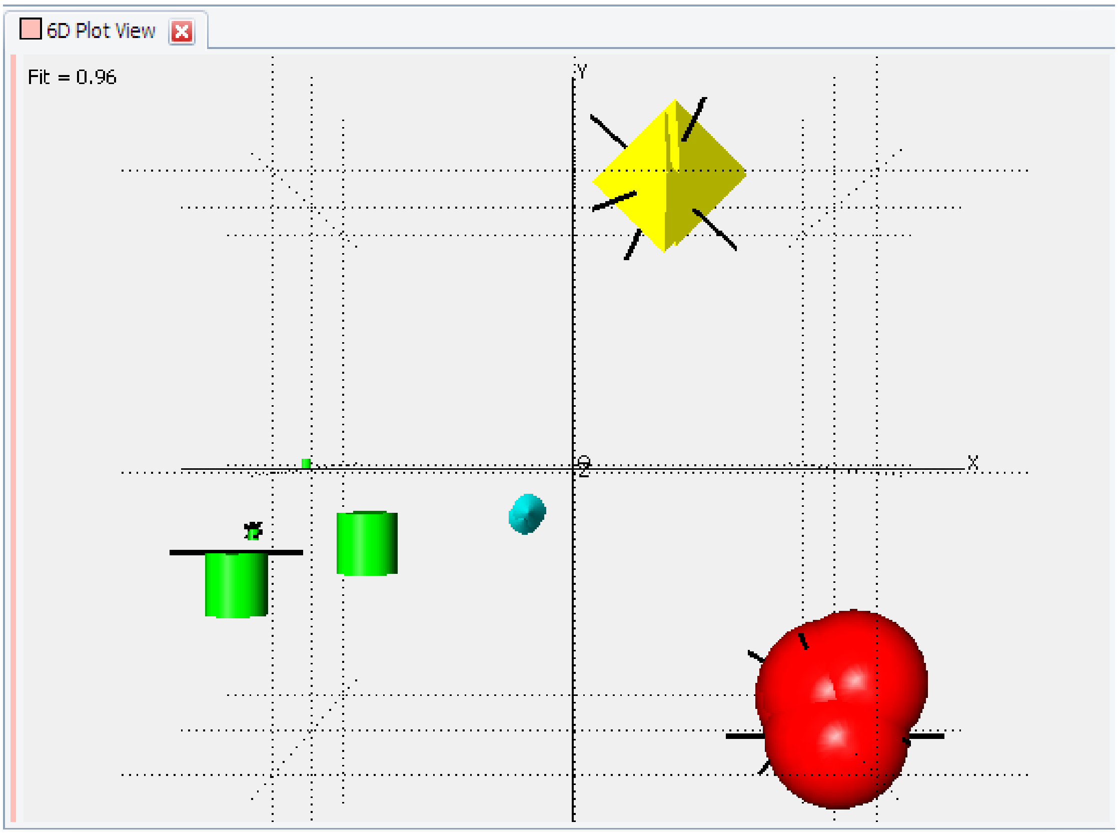
3. Click on the **Size** field's selector button and choose **Reaction Time**.
4. Leave the Color field on Default.
5. Click on the **Shape** field's selector and choose **Solvent**.



6D Extra Information

Use MMDS	☑
Size	Reaction Time
Color	Default
Shape	Solvent
Update	Apply

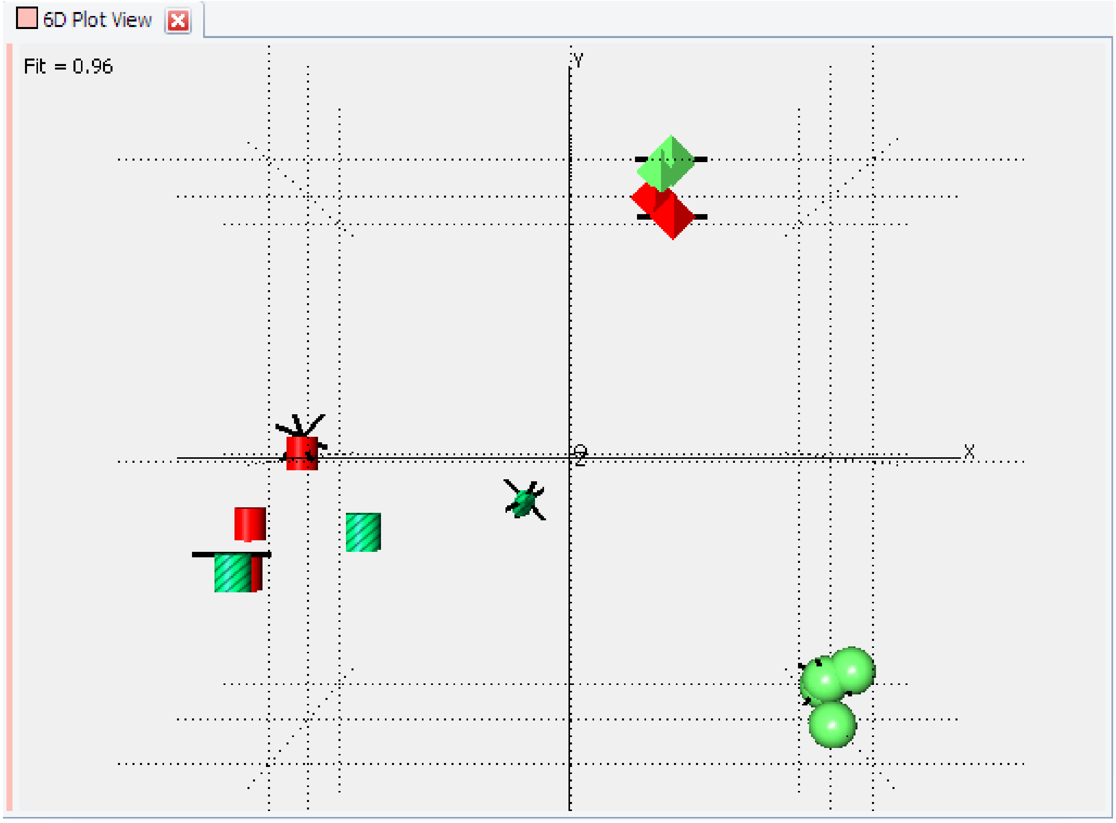
6. Click the **Apply** button.



It can be seen that while the plotted points are in the same positions and colors as before, there are now several different sizes and shapes on the display. Note that the different shapes are clustered together - spheres are next to spheres, cylinders are next to cylinders, etc. Select one of the sphere-shaped patterns by clicking on it. Note that the solvent used in all of the patterns plotted as red spheres is ethanol; whereas all the patterns plotted as cylinders, for example, used methanol. While this information was available before, it is easier to spot trends in the data when viewed in this manner.

From the different sizes of shapes plotted, it can be seen that there were three different reaction times used in this set of experiments. This may be seen more clearly by re-plotting the display, using color instead of size to represent the different amounts.

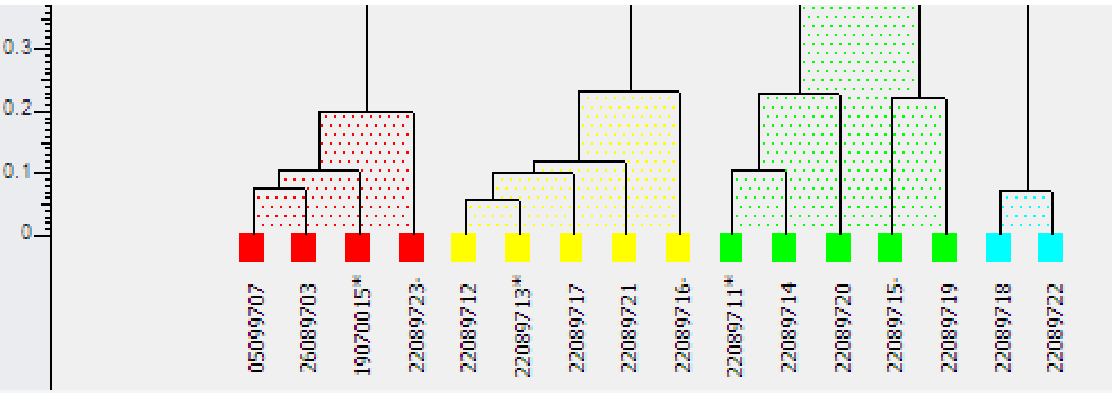
1. Set the **Size** to default,
2. set **Color** to **Reaction Time**, and ensure that **Shape** is still set to **Solvent**.
3. Click **Apply**.



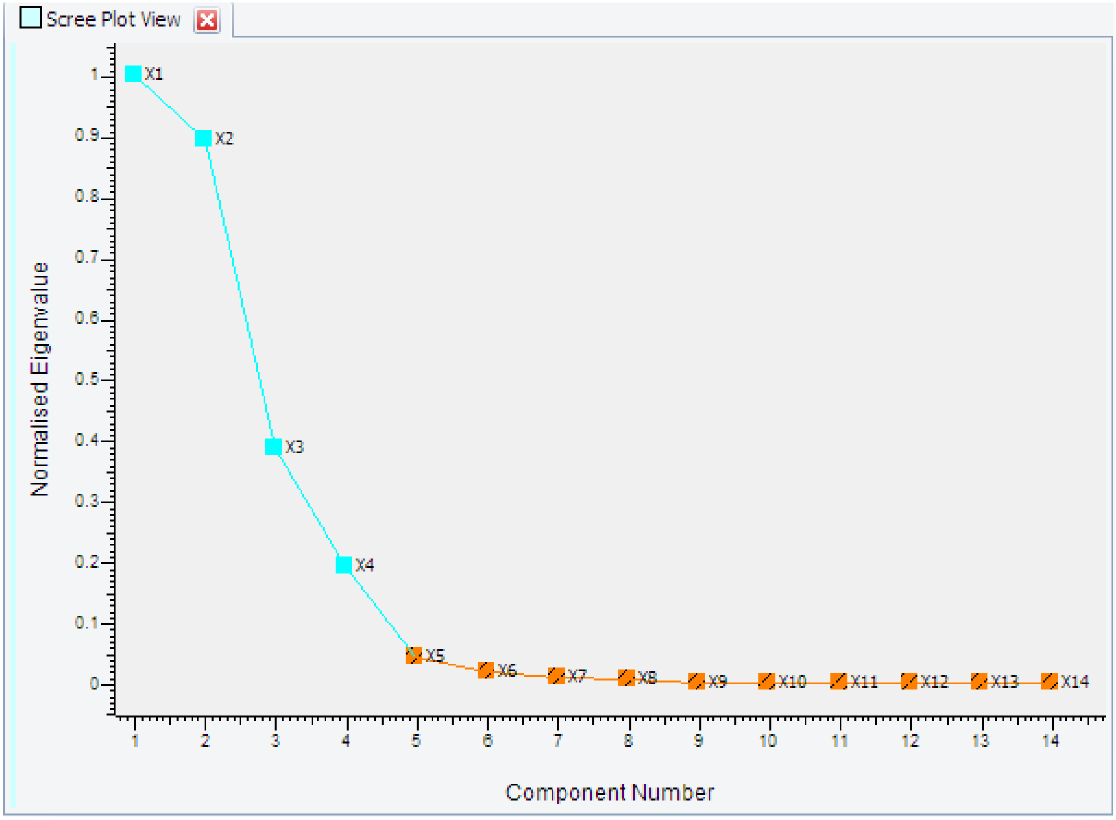
- ➨ The different colors - green, red and striped, represent the differing reaction times used. It is interesting to note that the two patterns 4 and 8 come up as separate colors - and hence times - from the rest of the patterns they are similar to.

4. Switch to the **Dendrogram** tab.

- ➨ Notice that pattern 22089716, in yellow, seems more separated than any of the rest of the yellow patterns:



5. We need to consider if the cut-line is positioned correctly. Create a **Scree Plot View from the Cluster Analysis** context menu.



The scree plot is one of the many methods used for estimating the cut-level of the dendrogram, and can be a useful visual aid for the user. It is derived from the eigenvalues of the pattern correlation matrix. To interpret the scree plot the gradient between the values (number of clusters) need to be analyzed. Where the eigenvalues are very different from one another there will be a steep gradient between clusters. This is one of the methods used to select the number of clusters.

For this dataset the number of clusters suggested from this is between 4 and 5, because it is here that the plotted line changes color, and the gradient starts to level out. Five clusters corresponds to the current cut level; 4 would correspond to raising the cut line such that the green and blue patterns were considered part of a single cluster.

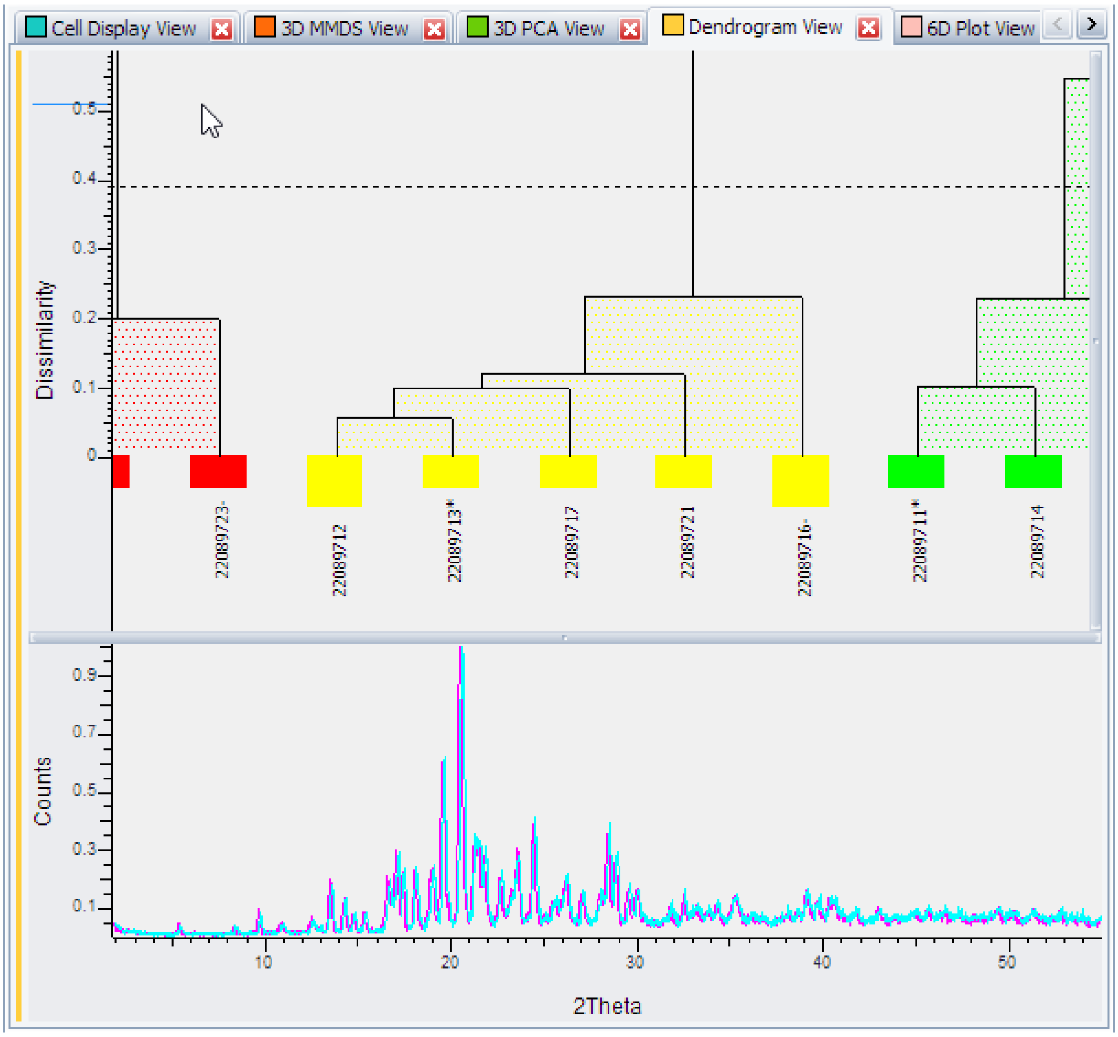
In cases such as this where the graphic displays may not give a clear cut solution, it can often be useful to examine the detailed statistical output from the analysis. This is shown in the **Log File View**.

1. Create the **Log File View** from the context menu of the **Cluster Analysis** node, and
2. scroll to the very bottom of the output.
3. Scroll back up slowly, until you see the output from the cluster analysis.
   - ➨ One section of it reads as follows:

```
                Estimation of the number of clusters
                ------------------------------------

The median value is 5
From principal components analysis (non transformed matrix):   5
From principal components analysis (transformed matrix):       4
From multidimensional metric scaling:                          4
From the gamma statistic using single linkage:                 1
From the Calinski-Harabasz statistic using single linkage:    5
From the C-statistic using single linkage:                    5
From the gamma statistic using group averages:                1
From the Calinski-Harabasz statistic using group averages:    5
From the C-statistic using group averages:                    5
From the gamma statistic using the Ward method:               1
From the Calinski-Harabasz statistic using the Ward method:   5
From the C-statistic using the Ward method:                   5
From the gamma statistic using complete linkage:              1
From the Calinski-Harabasz statistic using complete linkage:  5
From the C-statistic using complete linkage:                  5
The median value is 5

Combined weighted estimate of the number of clusters is       4


Maximum estimate is        5
Minimum estimate is        1
```

This suggests that 5 is indeed the correct number of clusters.

1. Go to the **Dendrogram view**, and
2. select patterns 22089716 and another of the yellow patterns. Consider their overlaid profiles - the similarity between them suggests that the program was correct in its initial placement of the cut line, and the reason 22089716 is separated somewhat from the rest appears to be down to an x-shift.


Cell Display View
3D MMDS View
3D PCA View
Dendrogram View
6D Plot View
Dissimilarity
0.5
0.4
0.3
0.2
0.1
0
22089723·
22089712
22089713*
22089717
22089721
22089716·
22089711*
22089714
Counts
0.9
0.7
0.5
0.3
0.1
10
20
30
40
50
2Theta

Index

Numerics

3D MMDS view 40
3D MMDS View Properties 43
3D PCA View 45
6D Plot View 46
  Special Properties of the 6D Plot 46

A

Advanced Options 21
Allow x shift 24

B

BRML 13

C

Cell Display view 32
  Hide results similar to references for cell display 34
  If known phases are available 33
  If no known phases are available 34
Cell Display view properties 35
Check for amorphous 24
Cluster Analysis node properties 27
Combine multiple datasets using weights 22
Common Options 22

D

Data command panel 11
Data Formats 13
Data property panel 11
Data tree pane 11
Data Types 13
Dendrogram View 17, 35
  Changing the Clustering Method 39
  Changing the Cut-level 38
Dendrogram view properties 40
Denoises pattern 24

F

Fuzzy Clustering View 58
Fuzzy Clustering View Properties 60

G

General View Features 28
Graphical Context Menu 30
Graphical Views 31

H

Hide results similar to reference for cell display 22

I

Include reference 22
INDSCAL 12, 22

L

Label extracting options 22
Length 23
Log File View 67
Log File View Properties 68
Log File View Toolbar 67

M

Mask Region 16, 25
Match Region 16, 25
Minimum Spanning Tree 51
Minimum Spanning Tree View Properties 54
Mouse Actions and Keyboard Shortcuts
  in 3D MMDS View 44
  in 6D Plot View 48
  in cell display 35
  in Dendrogram View 39
  in the Minimum Spanning Tree 53
  in the Parallel Coordinates Plot 62
  in the Silhouettes View 57
  in the Space Explorer View 64
Mouse and Keyboard Actions
  in 3 D Views 30
Multiple Data Sets 12

N

Numerical Results View 64
Numerical Results View Properties 66

P

Paper orientation 28, 31
Parallel Coordinates Plot View 60
Parallel Coordinates Plot View Properties 61
Pattern 23
Per dataset options 21
Pre-Screening 73
Property 22

Q

Quality Control 77

R

RAW ... 13
Remove cosmic ray spikes ... 24
Report Writer ... 68
Report Writer Toolbar ... 69
Report Writer view ... 19
Report Writer View Properties ... 70
Report Writer View Settings ... 70
Run Cluster Analysis dialog ... 21

S

Scan Pre-Processing ... 15
Scan Pre-Processing Options Window ... 23
Scree Plot View ... 50
Scree Plot View Properties ... 51
Set name ... 21, 24
Signal transform type ... 24
Silhouettes View ... 55
Silhouettes View Properties ... 57
Single and Multiple Selection ... 28
Space Explorer View ... 62
Space Explorer View Properties ... 63
SSD ... 13
Start index ... 23
Subtract background ... 24

T

Title bar ... 27
TXT ... 14

U

User Interface Elements ... 11

V

Validation Views ... 50
View
    printable 28, 31
View Properties ... 30

W

Weights ... 21

X

XY, XYE ... 14

Y

Y ... 13

Z

Zooming and Panning ... 30

Bruker Corporation

info@bruker.com
www.bruker.com

Order No: DOC-M88-EXX205