UNIVERSITY OF SASKATCHEWAN

Canadian Light Source | Centre canadien de rayonnement synchrotron

CLS STXM Data Analysis

James (Jay) Dynes
Spectro-Microscopy Beamline
April 25th, 2019

Principal Component Analysis & Cluster Analysis

Data for this Section in Al-stack-analysis

Principal Component Analysis-Cluster Analysis

- **References**

  (1) Lerotic M. 2005. Cluster analysis in soft X-ray spectromicroscopy: Finding the patterns in complex specimens. J. Electron Spectroscopy and Related Phenomena 144-147-1137-1143.
  (2) Letrotic M. 2004. Cluster analysis of soft X-ray spectromicroscopy data. Ultramicroscopy 100, 35-57.

Why use PCA-CA?

- Complexity of the sample precludes the use of linear combination fitting of stacks using reference spectra of 'pure compounds' and/or to find new "unexpected" compounds
  - Goal is to identify representative spectra and produce quantitative component maps

- **Principal Component Analysis** (PCA) used orthogonalize spectroscopy data and discards much of the noise present in the data
  - Goal is to describe the specimen by a set of abstract components

$$S = 1 \ldots S_{abstract}, \text{ where } S_{abstract} \leq N$$

  - Abstract components describe main spectroscopic signatures in the data
  - Each signature arises from linear combination of several different chemical species, so that there is not a simple relationship between one abstract component and one particular chemical component

Matrix, eigenvectors, eigenspectra….

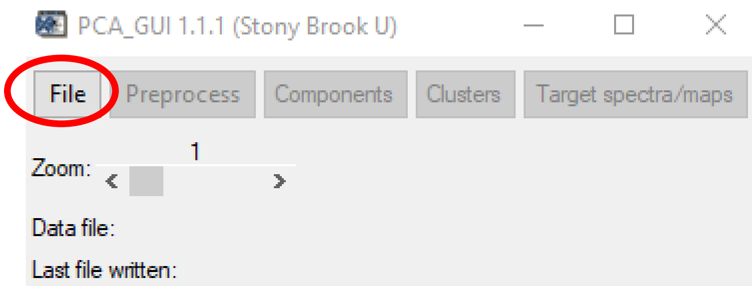
- **Cluster Analysis** (CA) used to find natural groupings of spectra, which then calculates the average spectra and displays the thickness maps associated with these spectra

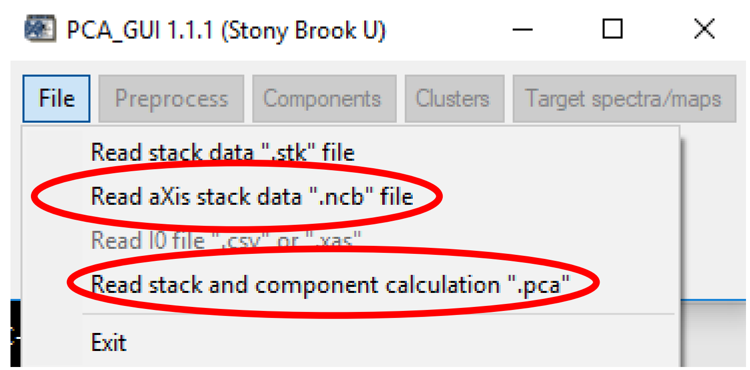
Principal Component Analysis

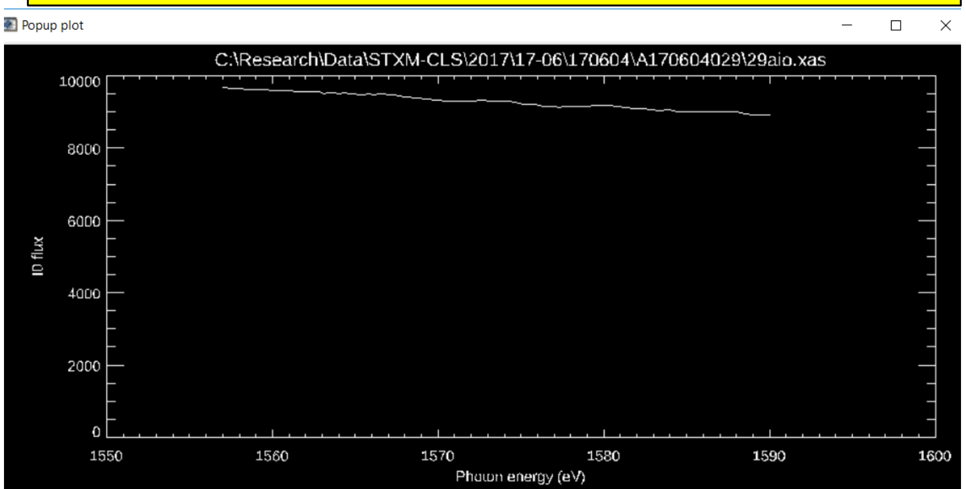
aXis2000 → Stacks → Statistical analysis → PCA_GUI (CJJ Dec 2005) which opens PCA_GUI 1.1.1 (Stony Brook U) GUI

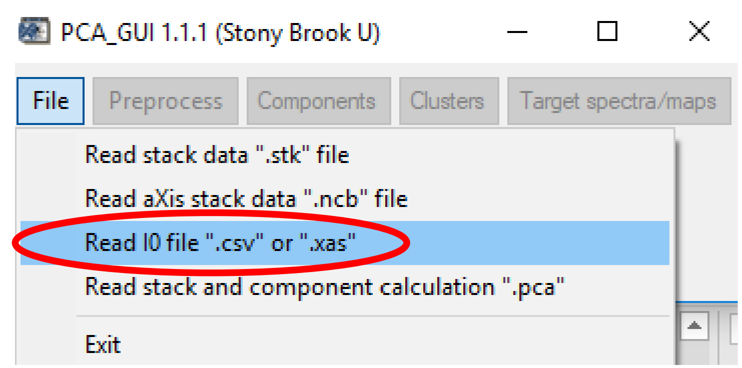
File → Read aXis stack .ncb file → Browse to aligned stack file that has not been changed to optical density as going to change it to optical density using the Io ".xas" file generated from the stack analyze

Read Io file ".csv" or ".xas" which now becomes active after selecting the .ncb file → Browse to the Io "xas" file created using stack analyze → Displays Io spectrum →**Dismiss** to go back to PCA_GUI



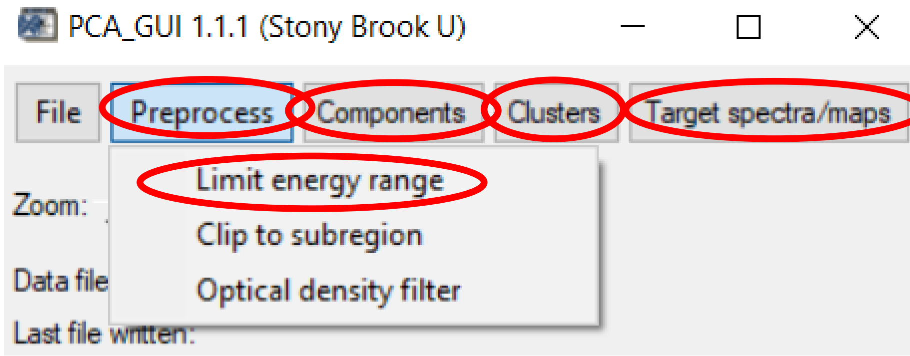




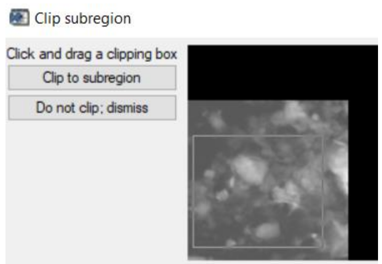


Principal Component Analysis

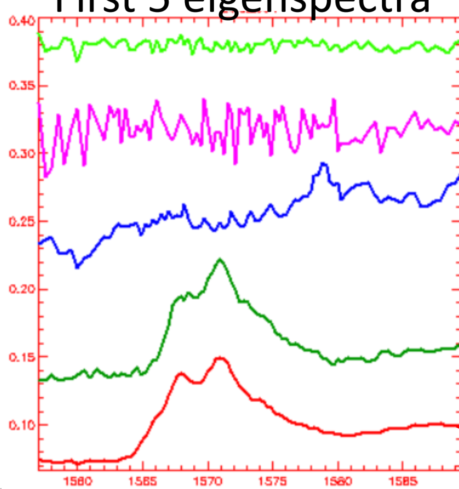
Preprocess → (1) **Limit energy range,** make energy range smaller to exclude certain features, such as K L-edge peaks in C K-edge spectrum → (2) **Clip to subregion,** can select a subregion by dragging box over area→ (3) **Optical density filter**

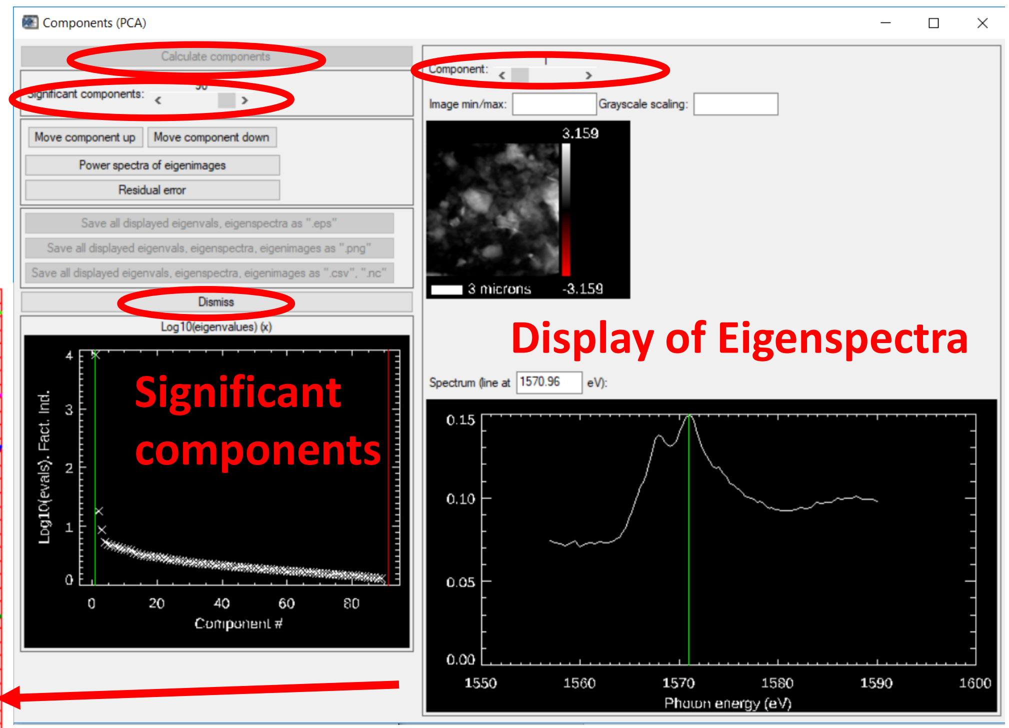


Components → Calculate components → Save PCA file which shows the eigenvalues versus component number- can be reloaded → Reduce the number of components and the Save options buttons become active → **Dismiss** to move to cluster analysis



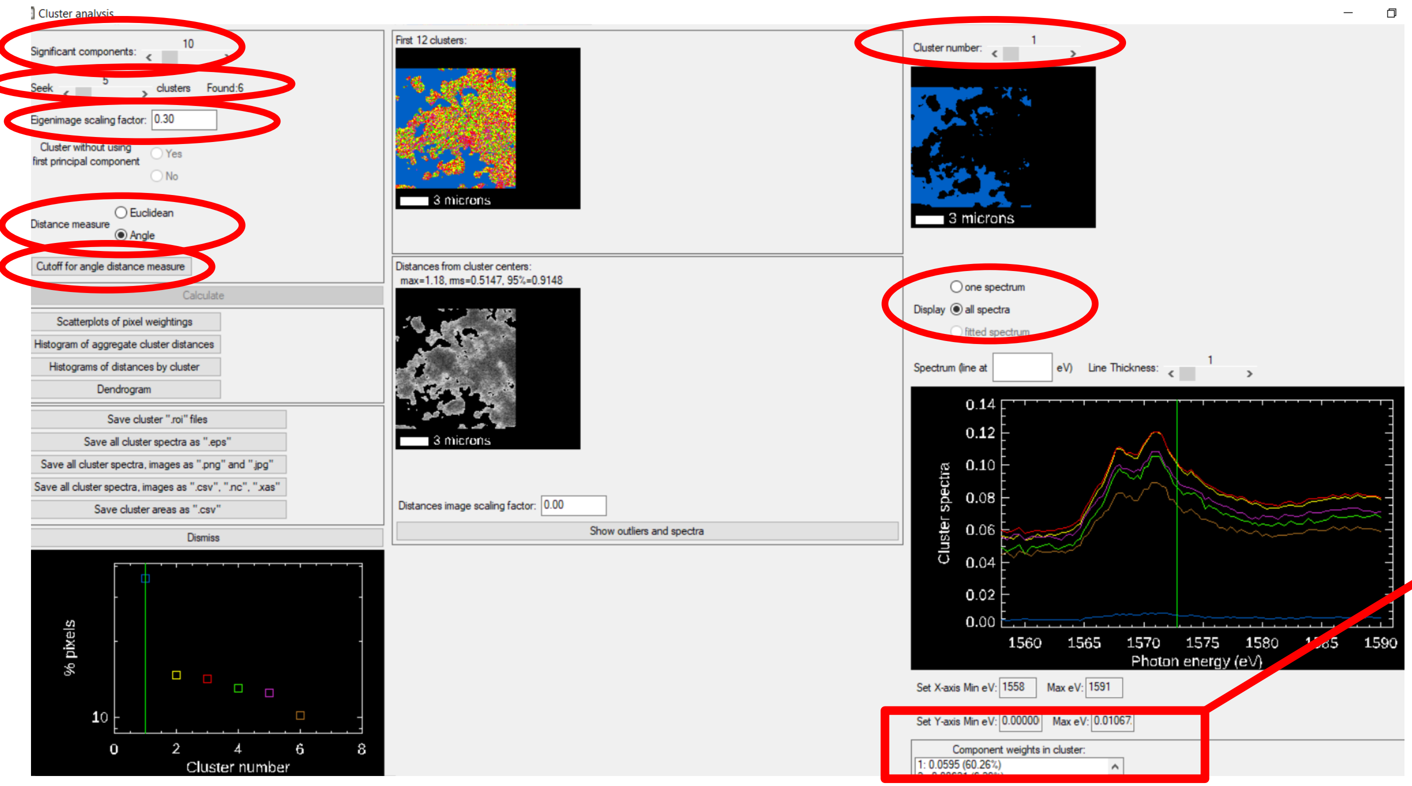
First 5 eigenspectra





Cluster Analysis

Cluster, opens to cluster GUI → **Angle → Cutoff for angle distance measure,** opens to GUI, typically use a cutoff of 0.02 → **Significant components**, move slider bar → **Seek**, move slider bar → **Calculate**



Component weights in cluster
1: 0.0595 (60.26%)
2: -0.00621 (6.29%)
3: -0.00427 (4.33%)
4: 0.00434 (4.40%)
5: 0.00 (0.06%)
6: 0.0009 (0.95%)
7: 0.0137 (13.89%)
8: 0.00419 (4.25%)
9: (0.00228 (2.31%)
10: -0.00322 (3.26%)

Cluster Analysis

Angle versus Euclidean

- Variations in thickness are often classified as distinct regions which is undesirable
- By using angle distance measure rather than a conventional Euclidian distance measure we can suppress thickness variations, thereby cluster data based more completely on chemical speciation alone
  - Uses ratio of eigenvalues instead of difference

Number of Clusters to seek

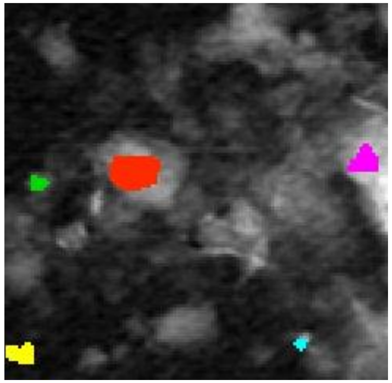
- Observed that more is better than less - try high number such as 20, then reduce to see if difference

Number of Significant components to select

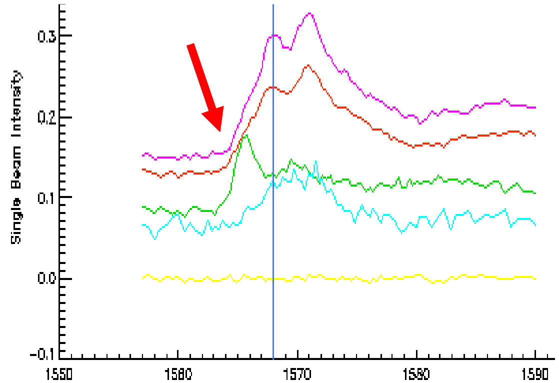
- Again try higher number, then reduce to see if difference

Cluster Analysis

Manual Extraction of Spectra from Stack

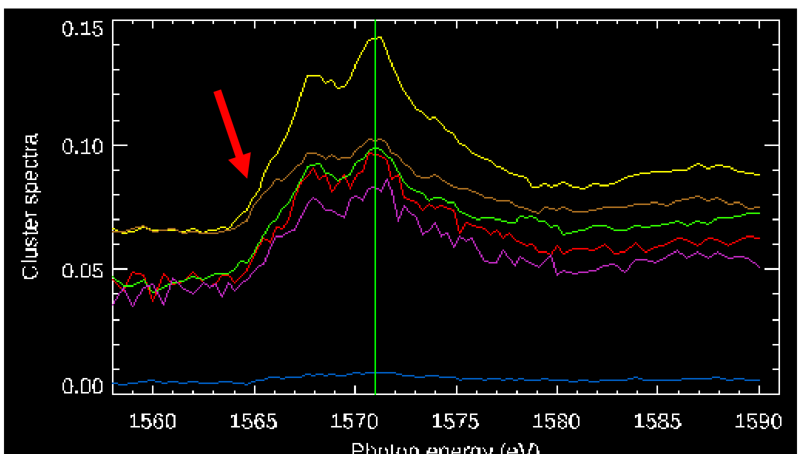


1568 eV

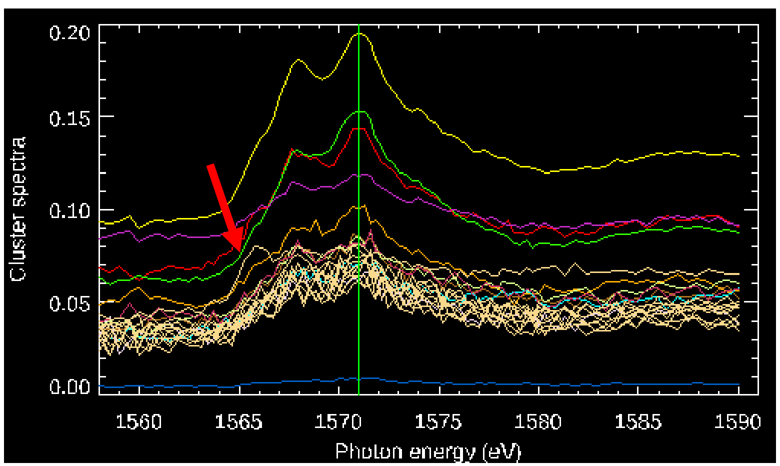


Problem is if do not use enough number of clusters may not identify all representative spectra

PCA Significant Components = 5,
CA number of clusters = 5



PCA Significant Components = 10,
CA number of clusters = 20

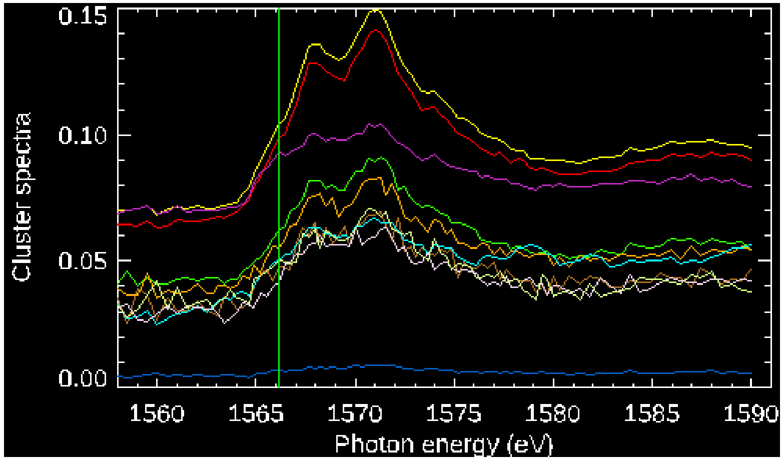


Scaling factor 0.3

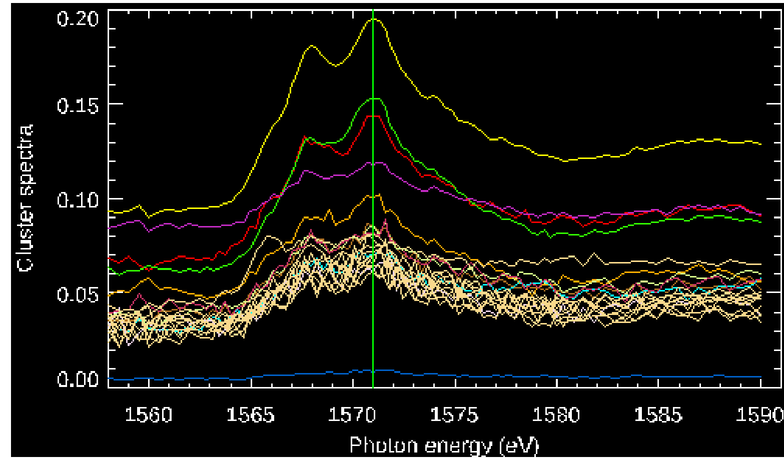
Cluster Analysis, Scaling factor

PCA = 10, CA = 20,

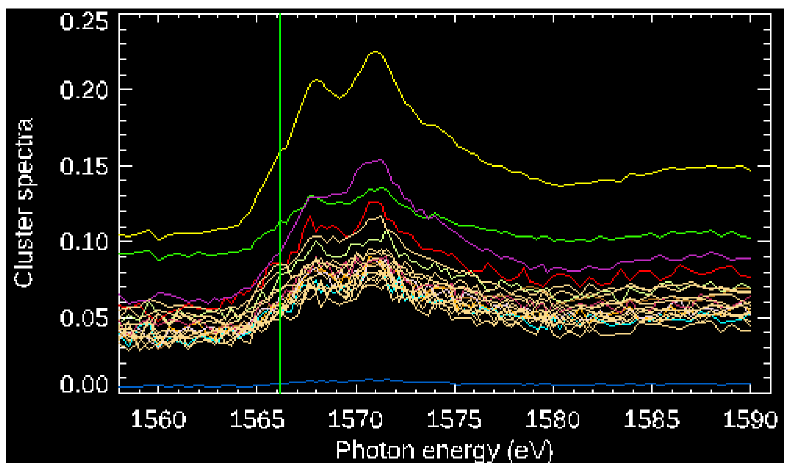
Scaling factor 0.2, 10 spectra



Scaling factor 0.3, 20 spectra



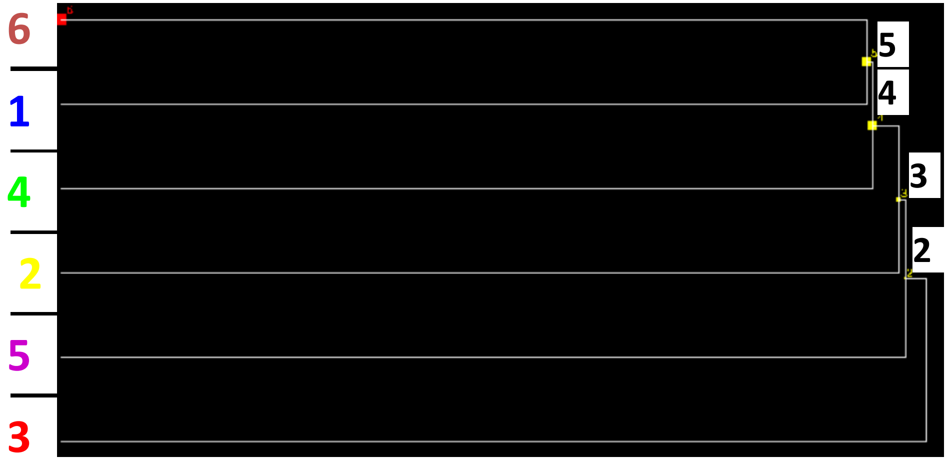
Scaling factor 0.4, 21 spectra



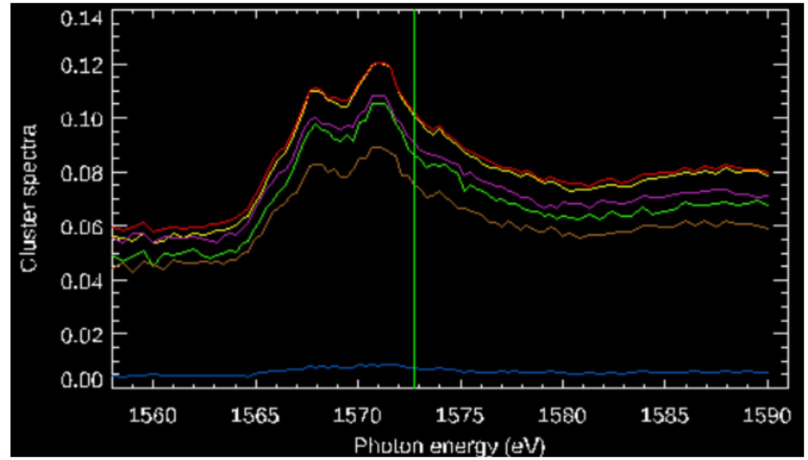
Scaling factor 0.3 perhaps best for identifying representative spectra

Cluster Analysis, Dendrogram

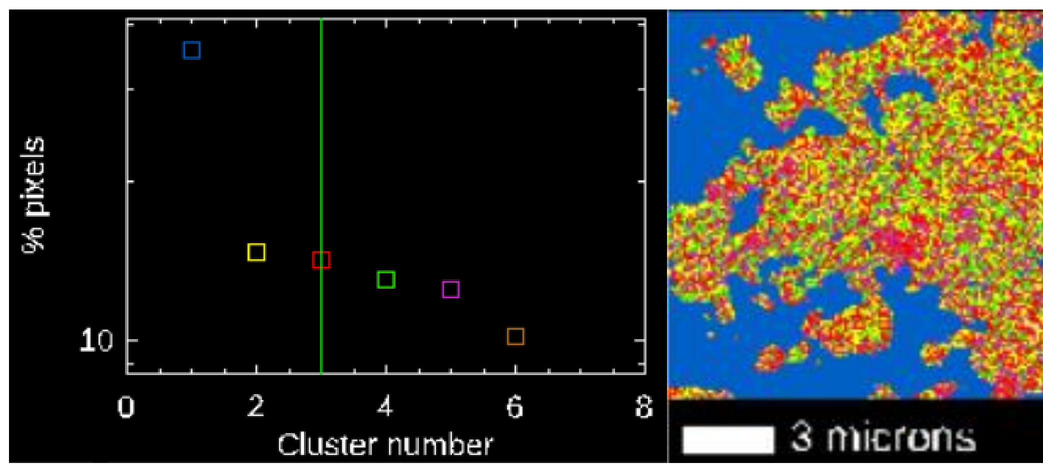
Component
Number



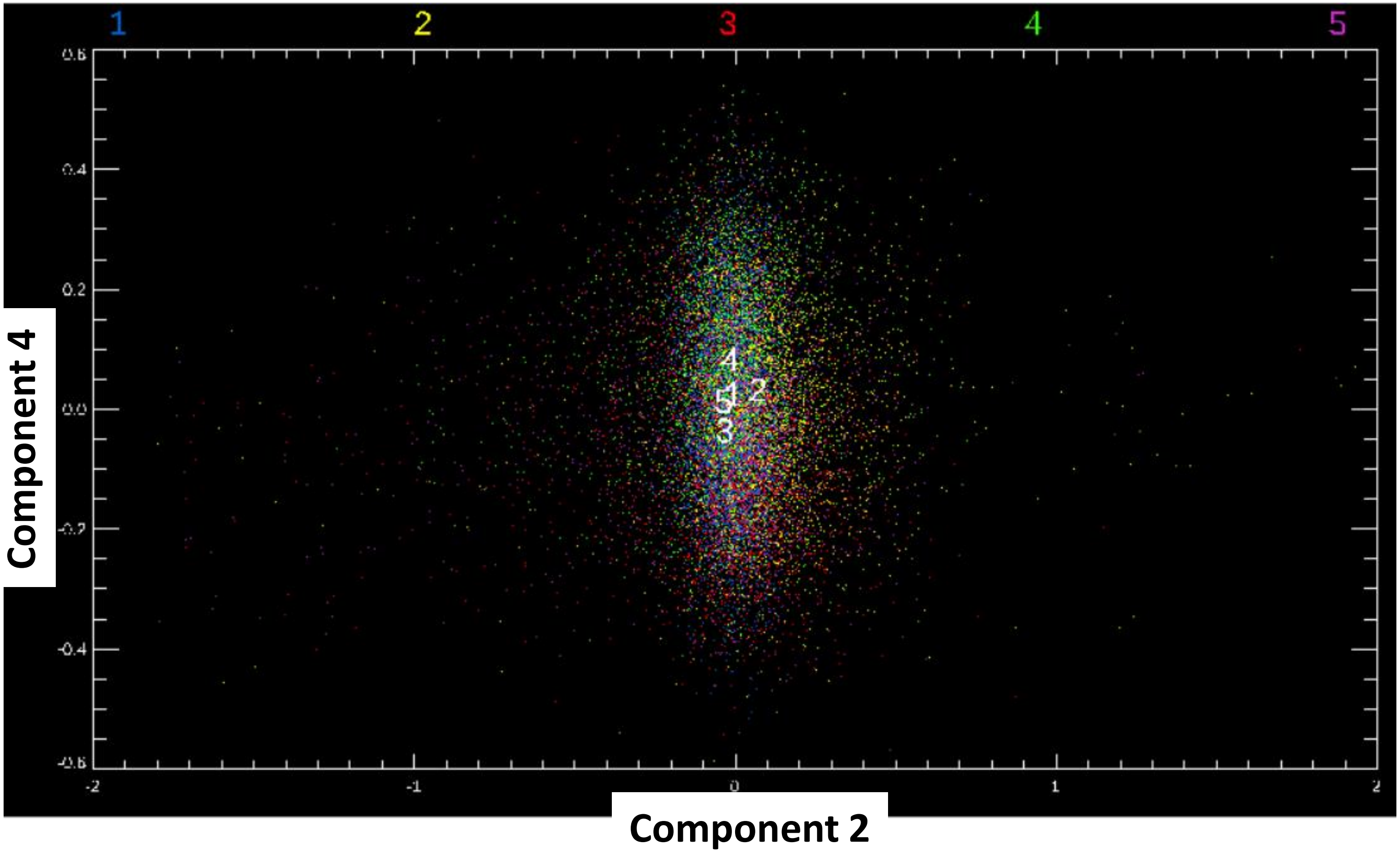
PCA Significant Components = 5,
CA number of clusters = 5, 6 Clusters



The key to interpreting a dendrogram is to focus on the height at which any two objects are joined together.

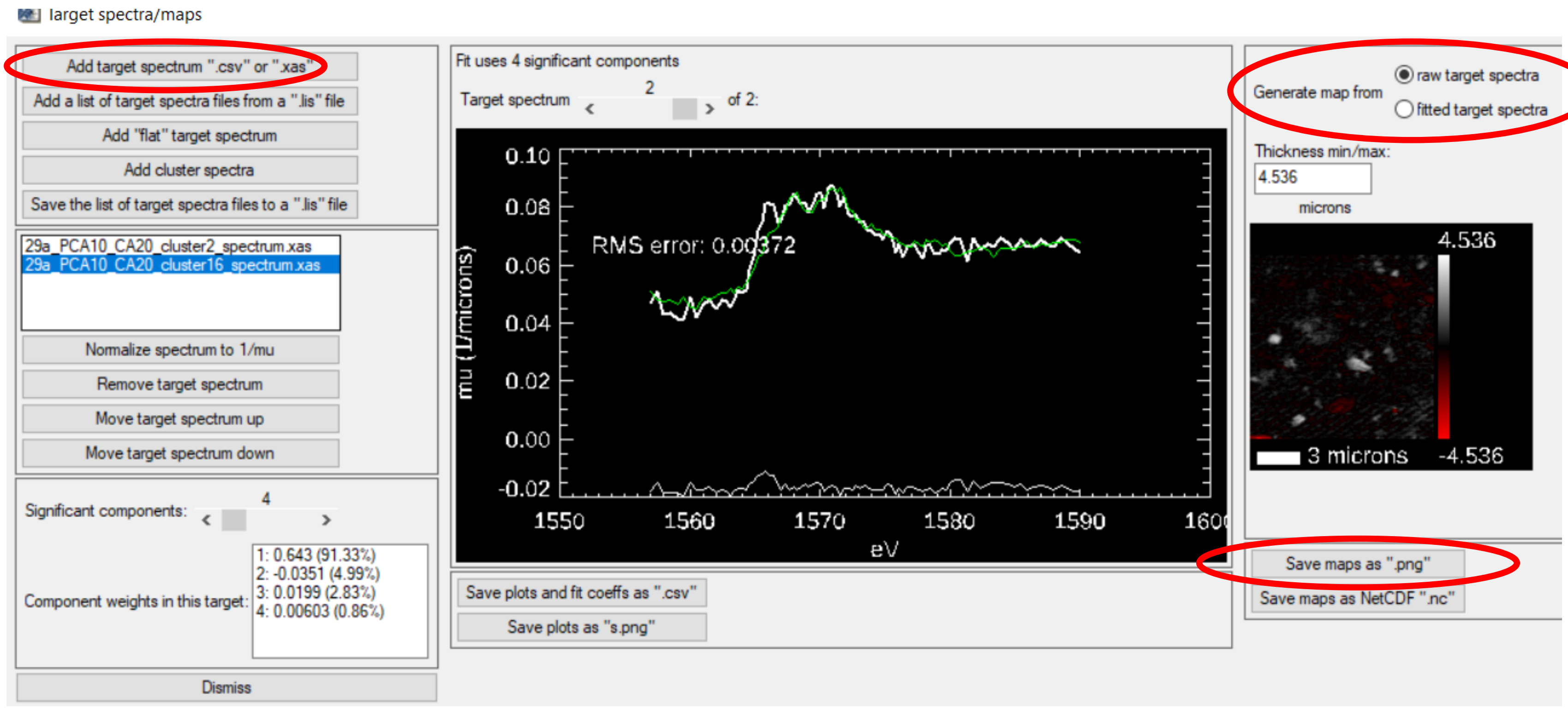


Cluster Analysis, Scatter plots



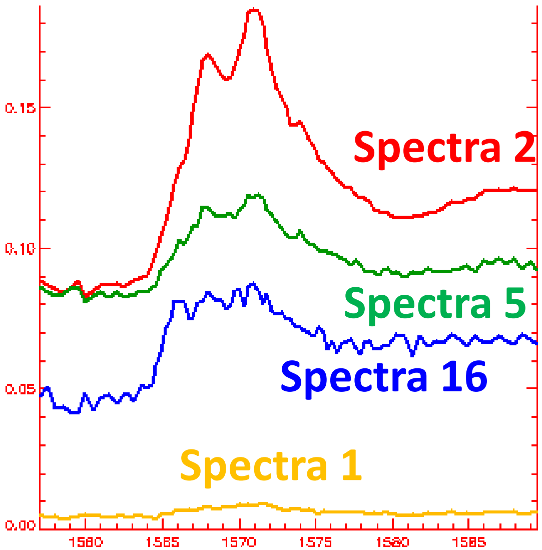
Target Spectra/maps

Target spectra/maps → Add target spectrum ".csv" or ".xas" → Save maps as ".png"→

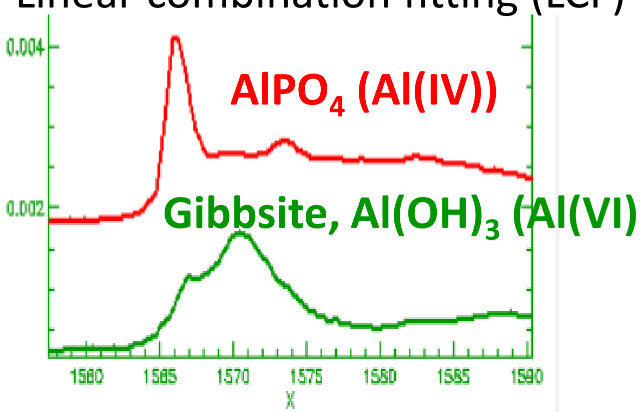


Target Spectra/maps: Comparing PCA-CA & LCF

Used spectra (2, 5 and 16) from the PCA 10, CA 20 cluster fitting

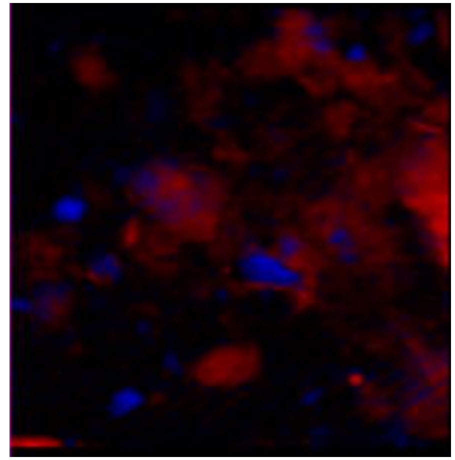


Linear combination fitting (LCF)

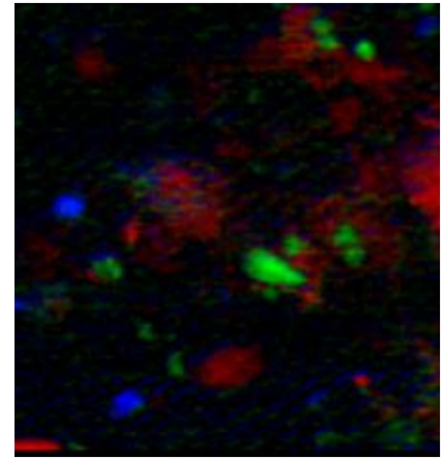


PCA-CA

Spectra 2 Spectra 16

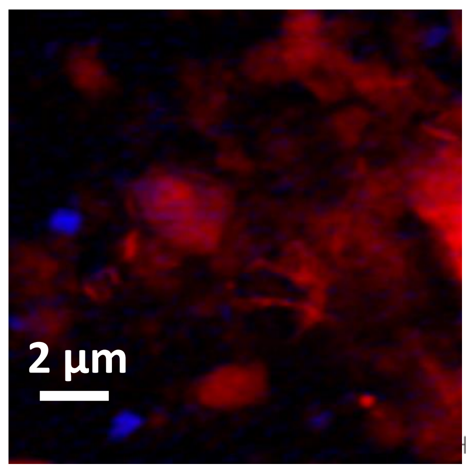


Spec 2 Spec 5 Spec 16



LCF

Gibbsite (Al(VI)) $AlPO_4$ (Al(IV))



Similar results between PCA-CA and LCF when use similar spectra in the fittings

Mantis PCA-CA

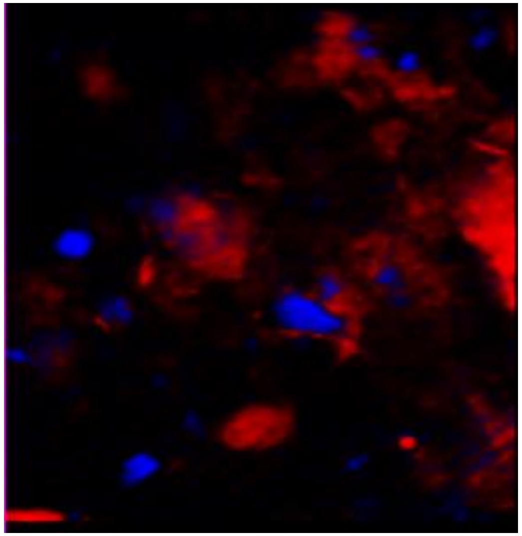
Mantis: Non-negative Matrix PCA-CA

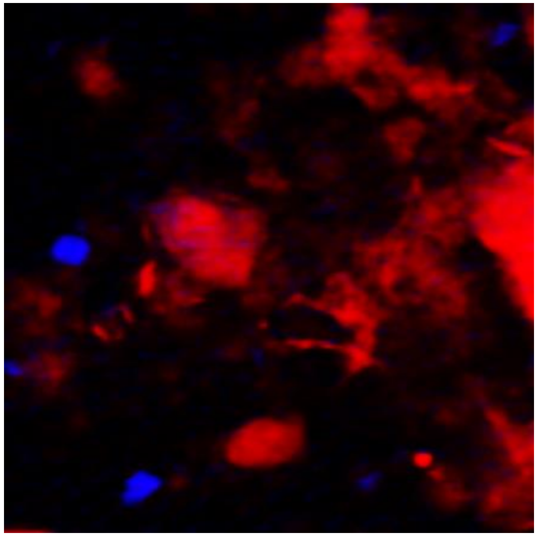
(1) Mak, R., 2014. Non-negative matrix analysis for effective feature extraction in X-ray spectromicroscopy. Faraday Discussions 17, 357-371.
(2) Lerotic, M. 2014. MANTiS: a program for the analysis of X-ray spectromicroscopy data. J. Synchrotron Rad. 21, 1206-1212.

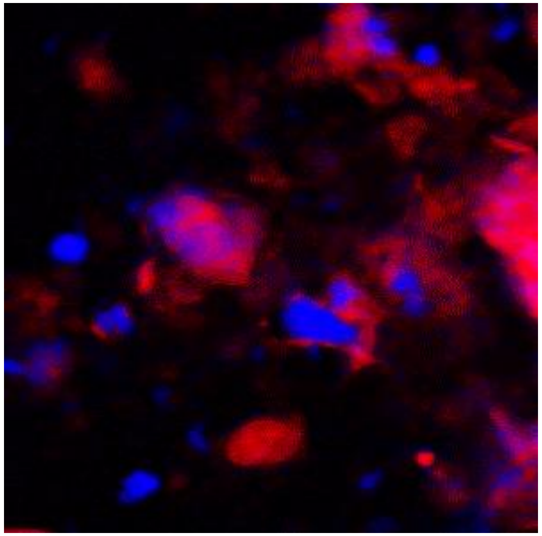
Download Mantis at https://bitbucket.org/mlerotic/spectromicroscopy

- Sometimes the component maps produced by cluster analysis using the PCA-CA (axis2000) code can yield some regions with slightly negative values, which are unphysical, and thus represent limitations in the analysis
- Mantis uses non-negative matrix analysis (NNMA) to constrain the analysis to eliminate negative values

PCA-CA(axis2000)	LCF	Mantis (NNMA)
Spectra 2(Al(VI))	**$AlPO_4$ (Al(IV))**	**Spectra 16(Al(VI))**
Spectra 16 (Al(IV)))	**Gibbsite (Al(VI))**	**Spectra 19 (Al(IV)))**



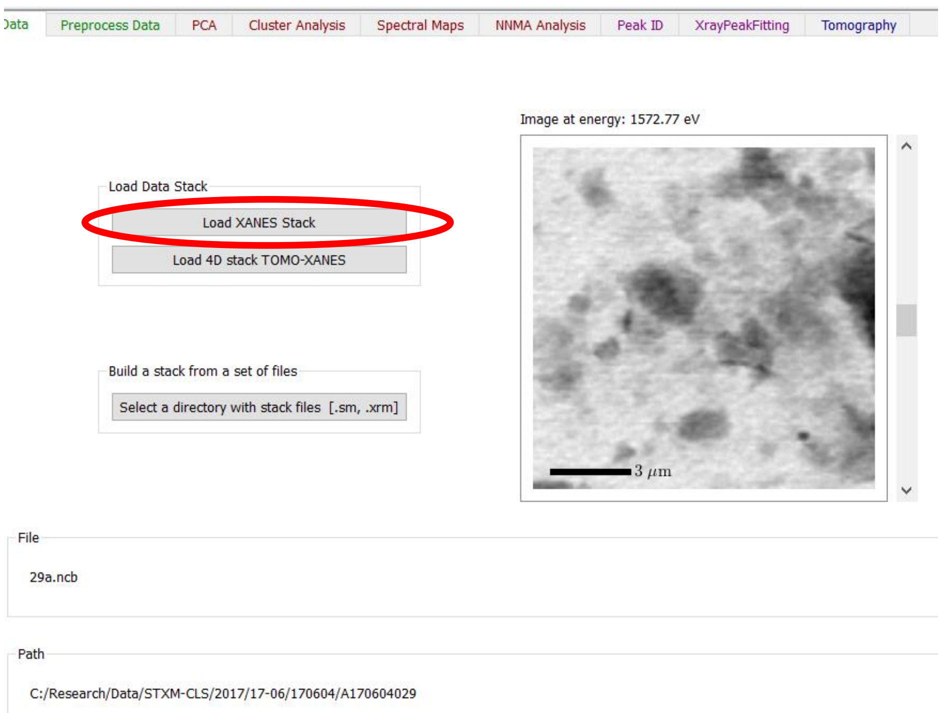




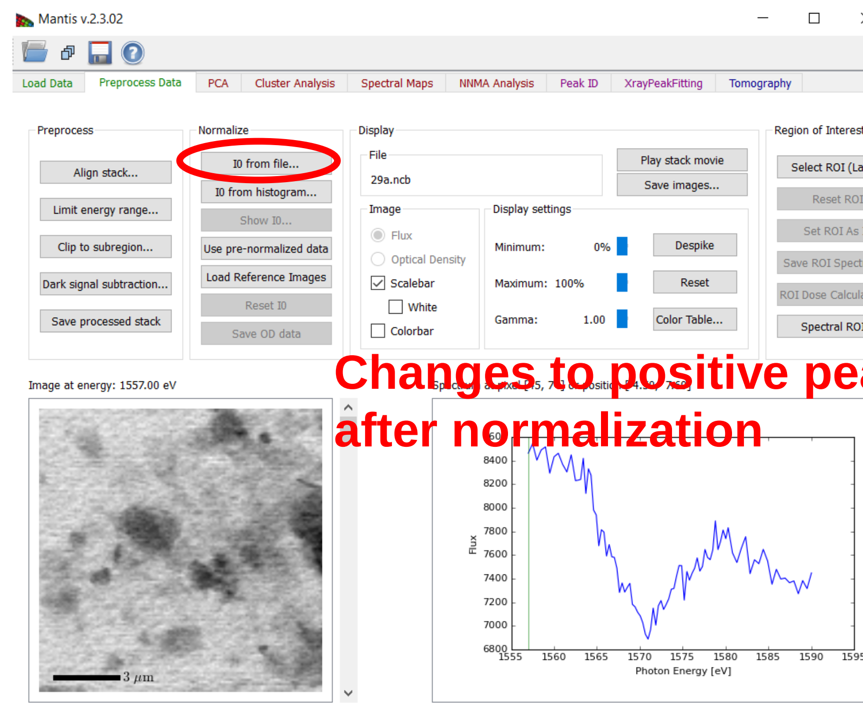
Mantis: Normalizing Stack

Load XANES Stack→ Preprocess Data – need to normalize using Io, opens preprocessing GUI → **Io from file**, browse to file, should get a positive peak

Data GUI



Preprocess Data GUI

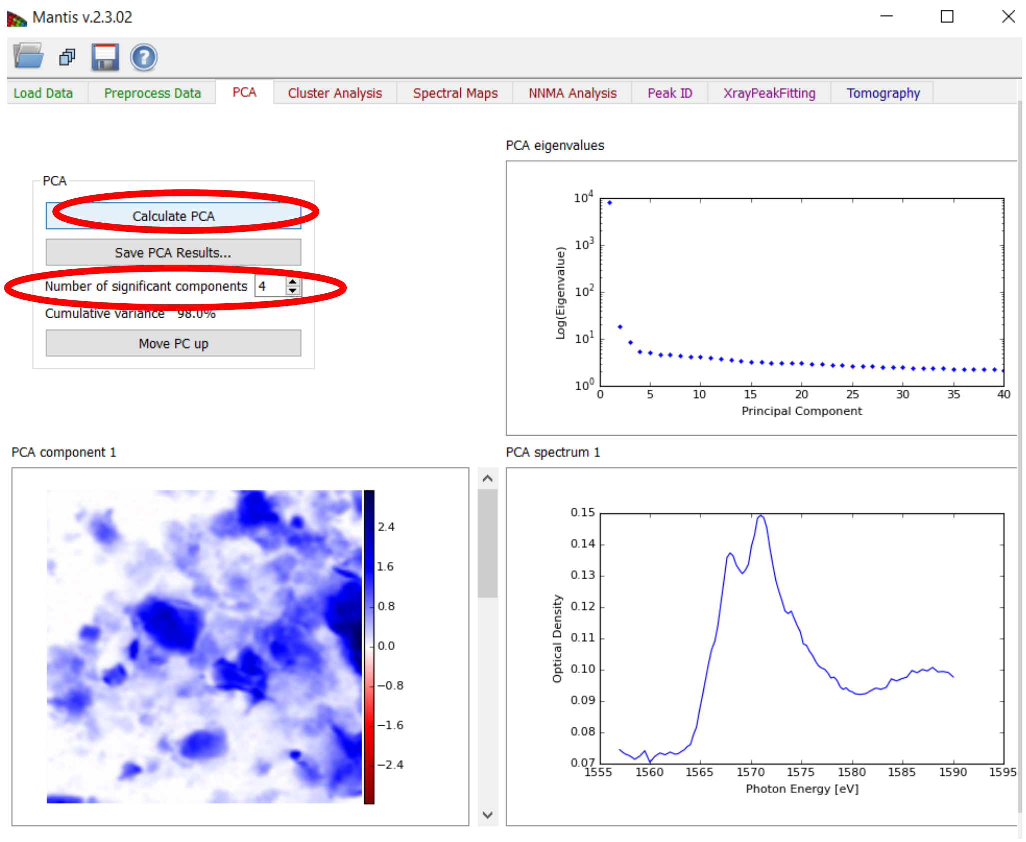


Changes to positive peak after normalization

Mantis: PCA

PCA → Calculate PCA

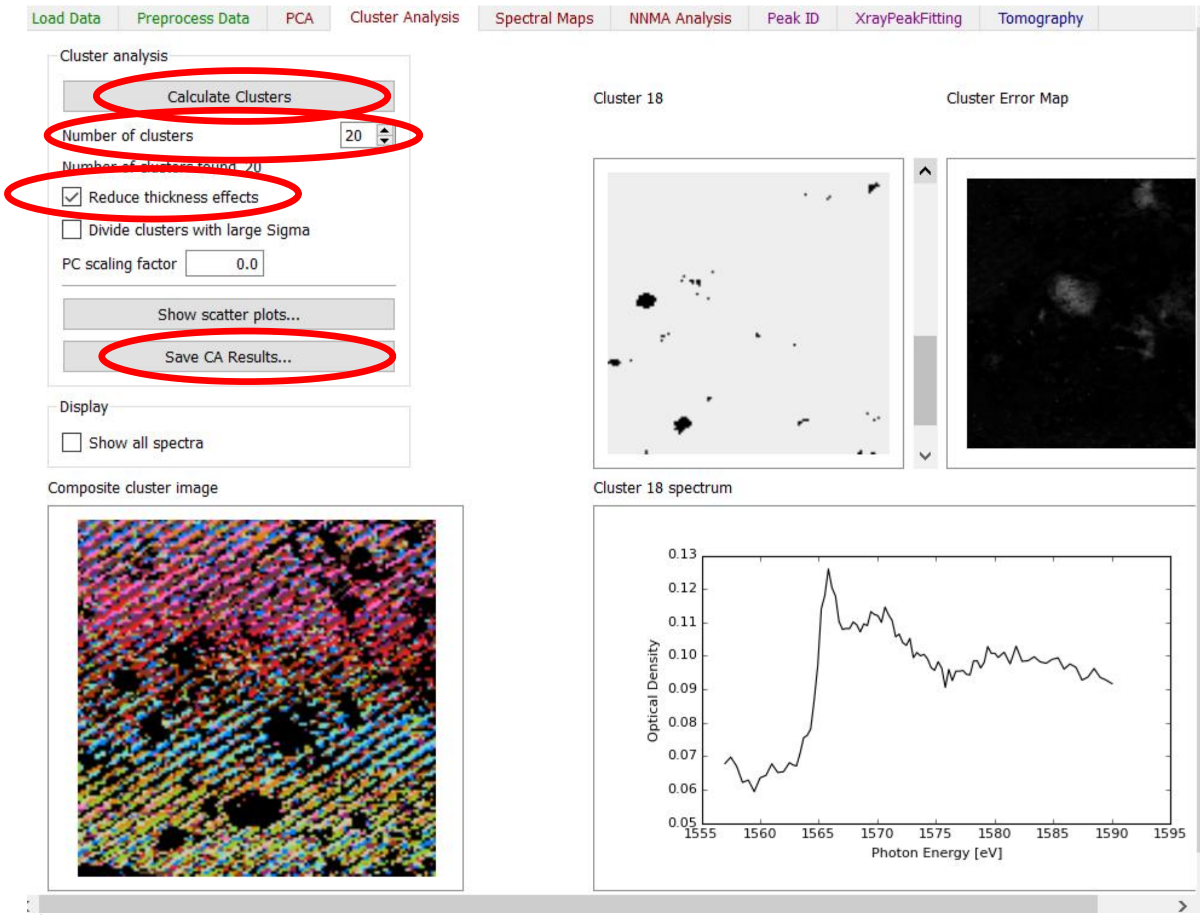
PCA GUI



Mantis: Cluster Analysis

Cluster Analysis → Calculate Clusters

Cluster Analysis GUI

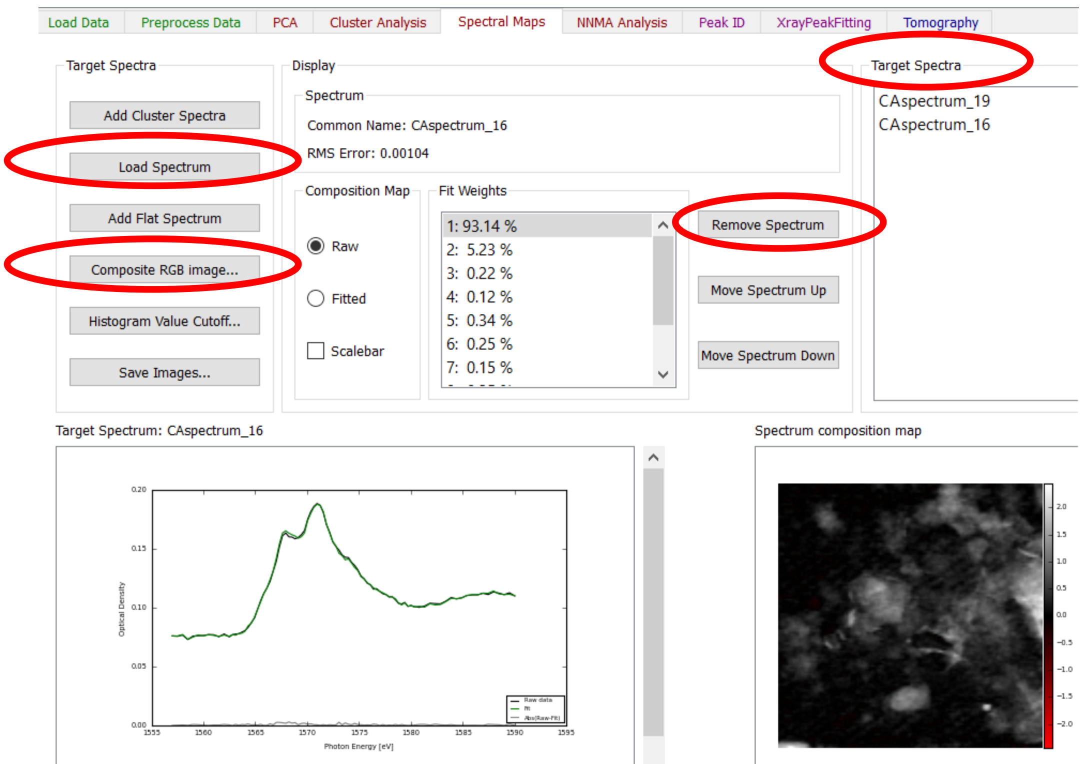


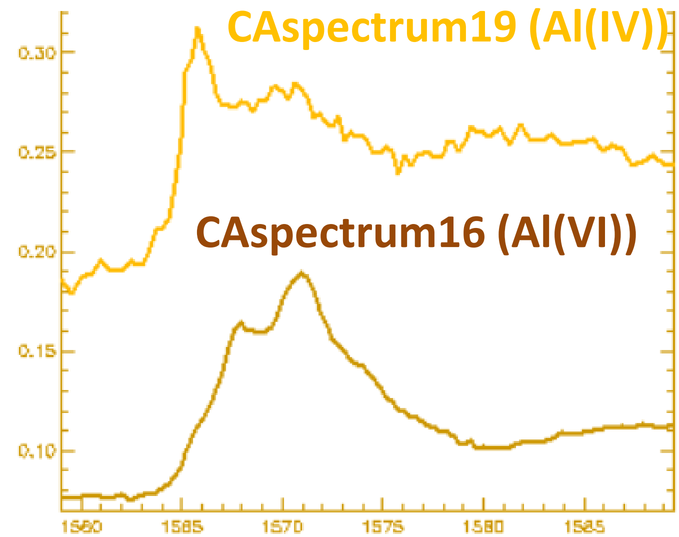
- ➢ To derive the Al (IV) spectrum needed to use clusters = 20 and check-off the reduce thickness effects
- ➢ Noticed that could get different results when doing the calculation a 2nd time using the same settings, particularly when changing the parameters
- ➢ Save results for spectral mapping

Mantis: Spectral maps

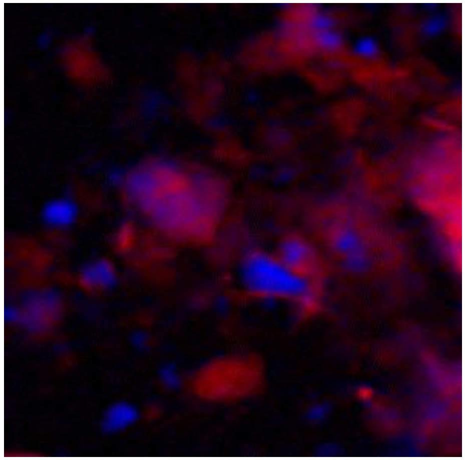
Spectral Mapping → Target Spectra loaded are what was last used in the cluster analysis. Remove and add spectra as appropriate.

PCA GUI





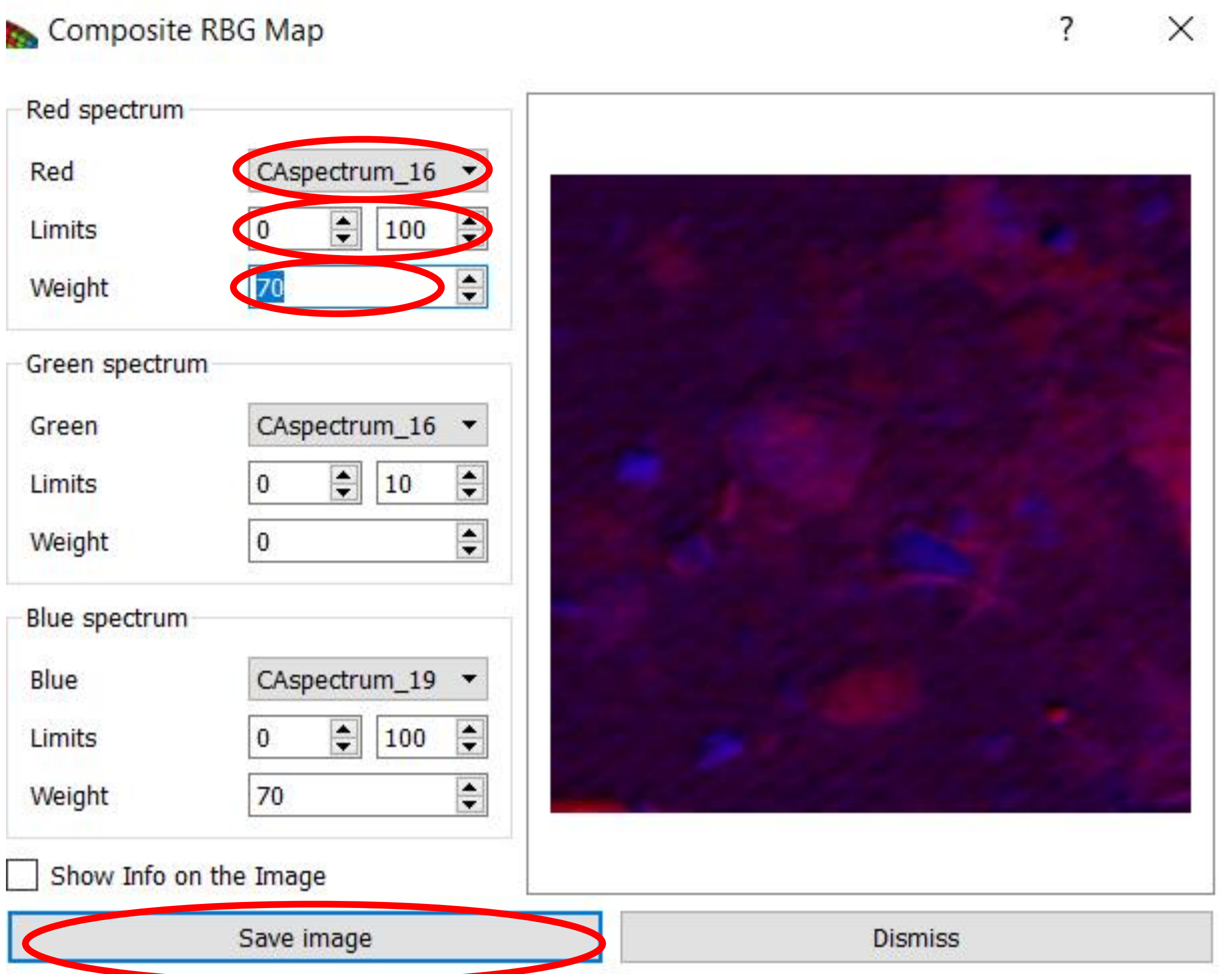
Spectra 16(Al(VI))
Spectra 19 (Al(IV)))



Saved images as ".png" and opened in axis2000 as composite RGB image requires 3 spectra to be opened

Mantis: RGB Composite Image

Spectral Mapping→ RGB Composite Image – 3 spectra must be loaded before can use this GUI

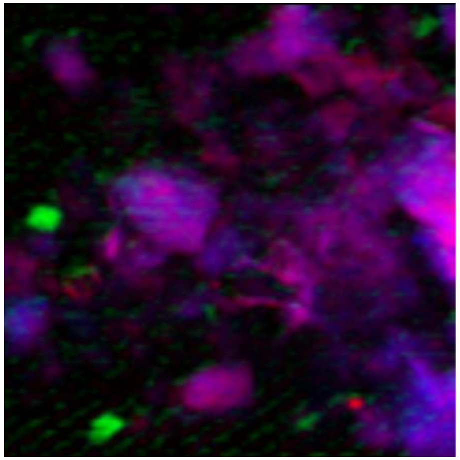


- Can change the lower and upper limits and the weight for each image

Correlation Analysis

Correlation Analysis

- We commonly produce RGB image overlays to visually determine whether two variables, such as two elements, are associated with each other (i.e., colocalized)



- Appears **Al(VI)** and Fe(III) associated with each other based on visual assessment. **Al(IV)** occurs as separate phase from Al(VI) and Fe(III)

- Correlation analysis is performed to mathematically quantify the colocalization of two variables, by generating a binary map of pixels for which the signal intensities meet certain criteria of correlation of colocalization

- Scatterplots or so-called 2D histograms are used to visualize the relationship between two variables

- Pearson's coefficient or Manders' coefficient are commonly calculated as a measure of the colocalization of two variables. Pearson's coefficient, unlike Manders' coefficient takes into account intensity in addition to location, so is more accepted to represent the correlation of the colocalization. (Ref. Adler, J. Cytometry Part A 2010,77A, 733-742)

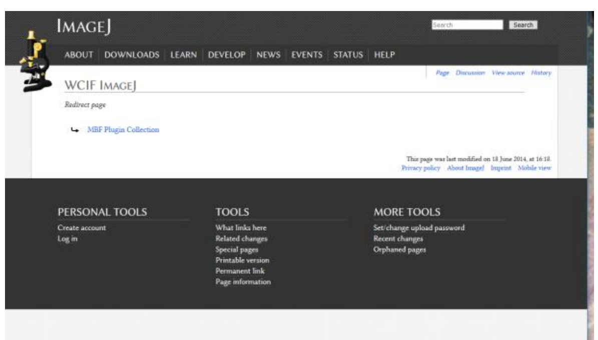
Correlation Analysis: ImageJ

- ImageJ is a public domain Java image processing program inspired by National Institutes of Health (NIH) https://imagej.nih.gov/

- There are many options for downloading ImageJ. You can download from the NIH site, which has 100's of plugins available. I have used WCIF (now MBF) and ImageJ (NIH) in the past. I currently use the Fiji package which has the advantage to the load tif files that require no further manipulation, which was not true for the other ImageJ packages. The difference is the number of plugins each initially comes with. You can easily install the plugins you want on any ImageJ platform.

https://fiji.sc/

https://imagej.net/index.php?title=WCIF_ImageJ&redirect=no



https://imagej.nih.gov/ij/

Correlation Analysis: ImageJ - ScatterJ

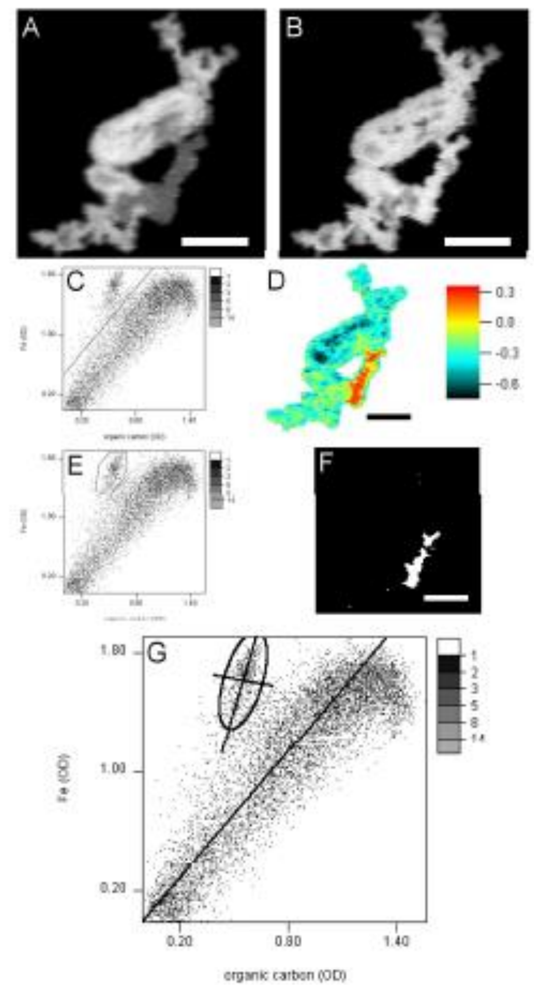
Reference

(1) Zeitvogel, F.. 2016. ScatterJ: An ImageJ plugin for the evaluation of analytical microscopy datasets. J. Microscopy 261, 148-156.

- Localization plugins come in most ImageJ packages. I have currently been using ScatterJ which in addition to determining Pearson's coefficient from the scatterplots also allows you to analyze different areas of the scatterplot (i.e., backmapping).

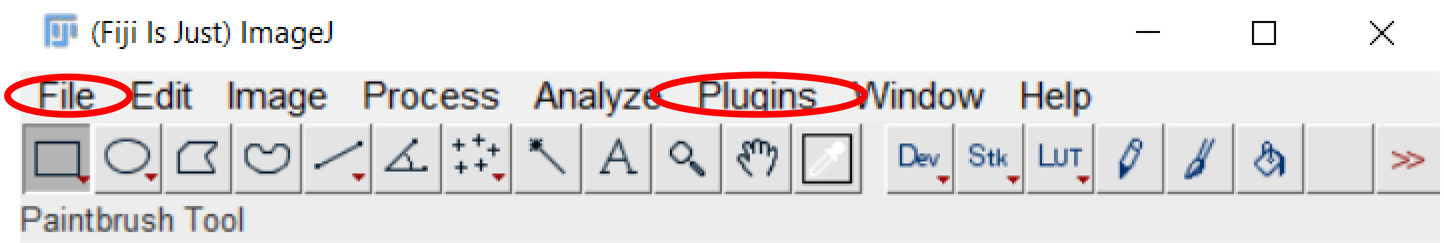
Download the plugin from http://download.savannah.gnu.org/releases/scatterj/

The plugins (ScatterJ_.class and ScatterJn_.class) are just put into the plugin folder in ImageJ and then they will appear in ImageJ plugin pulldown.



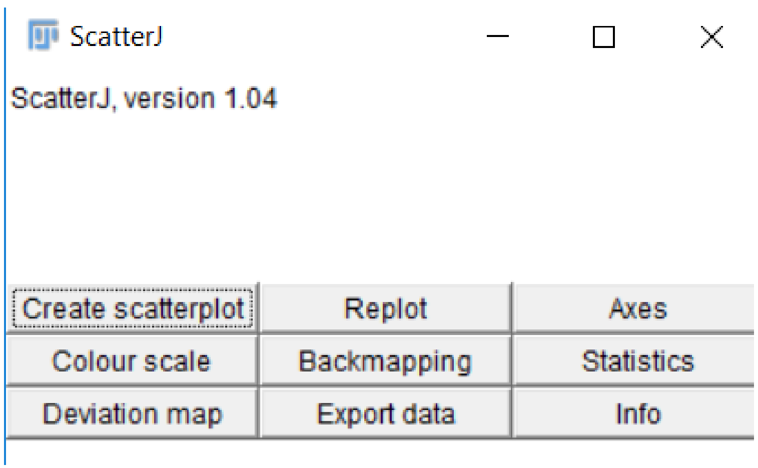
Correlation Analysis: ImageJ - ScatterJ

ImageJ GUI

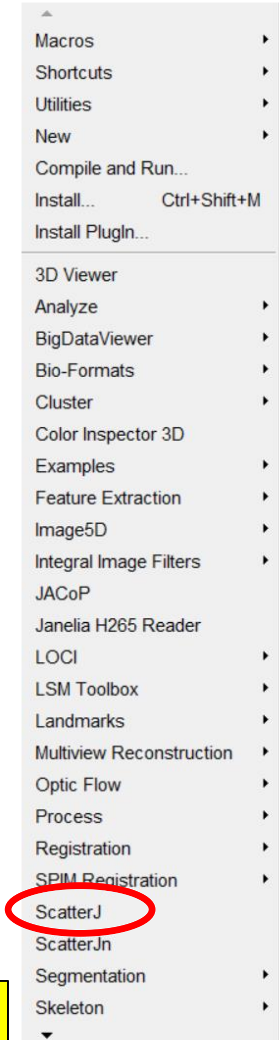


To open scatterJ, **Plugins → ScatterJ**

ScatterJ GUI



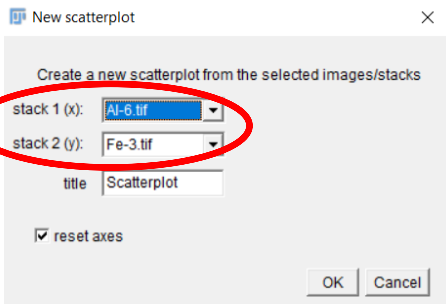
Note that a colocalization macro is located in the **Analyze** pulldown which gives the same information as ScatterJ, except the backmapping capabilities.

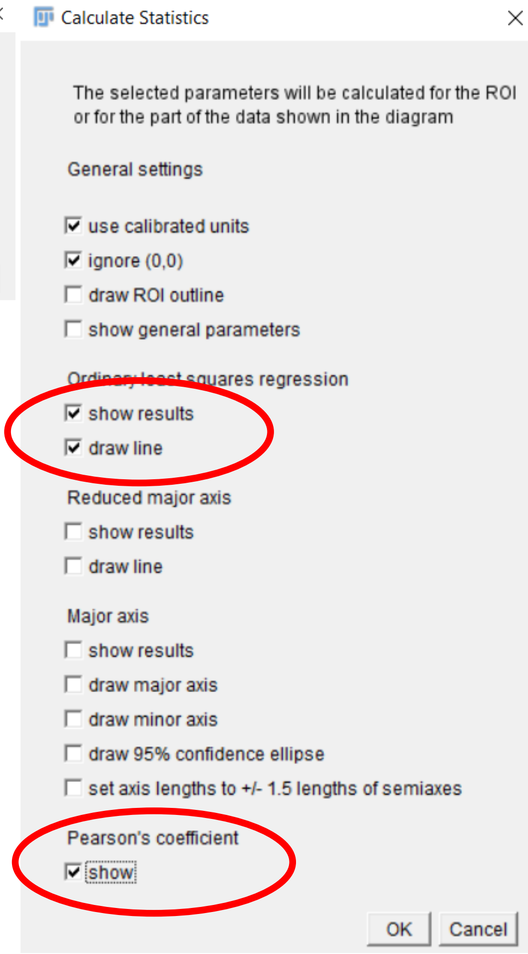


To open files from ImageJ **File → Open**, browse to files. Tif files work. From axis2000 save the ".axb" files as Tif (data) files by **Write → Graphics → Tif → data** Note that you need to unclick the scalebar button, otherwise the scalebar will become part of the tif image.

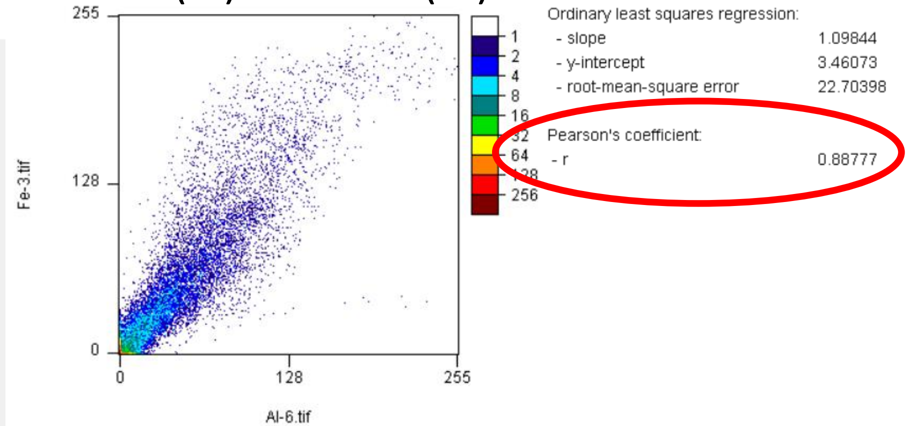
Correlation Analysis: ScatterJ

From scatterJ **create scatterplot**, select the files → **Statistics**

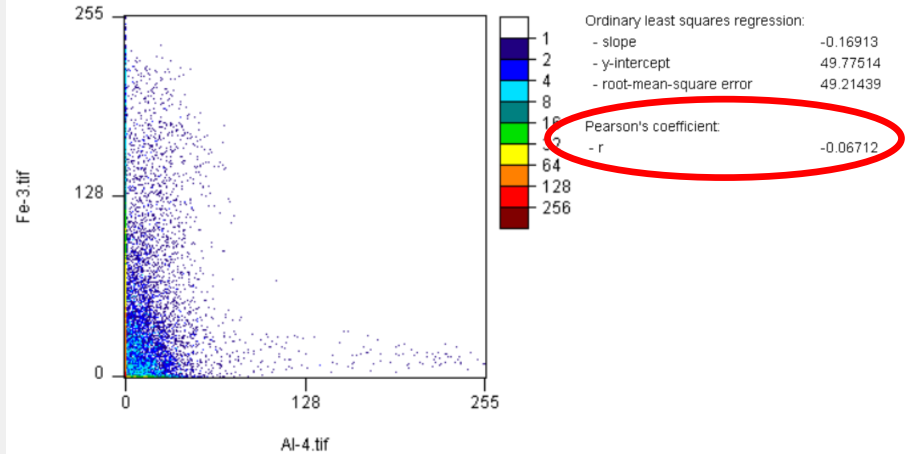




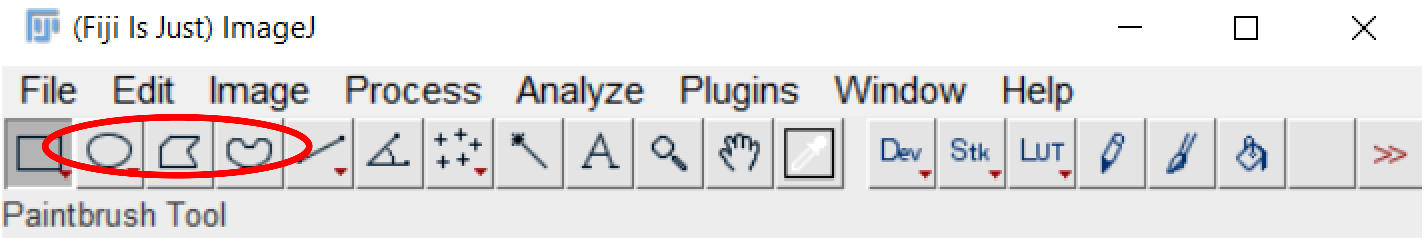
Fe(III) versus Al(VI)



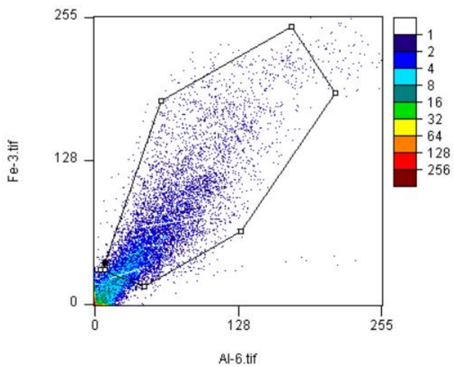
Fe(III) versus Al(IV)

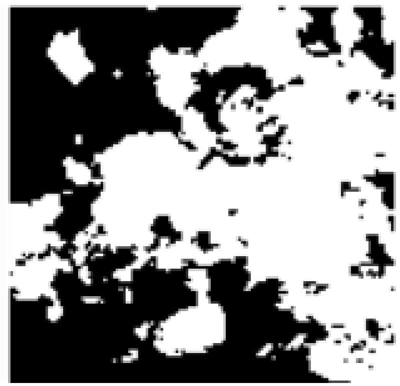


Correlation Analysis: ScatterJ



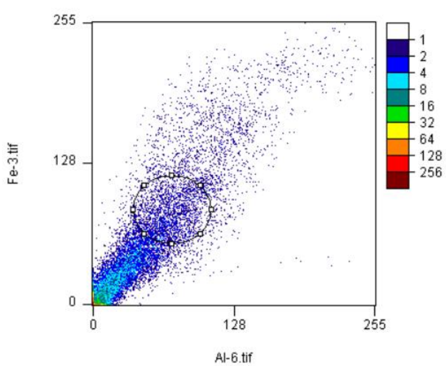
First select the area using the paintbrush tools → **Backmapping**, will use the area selected → **Statistics**

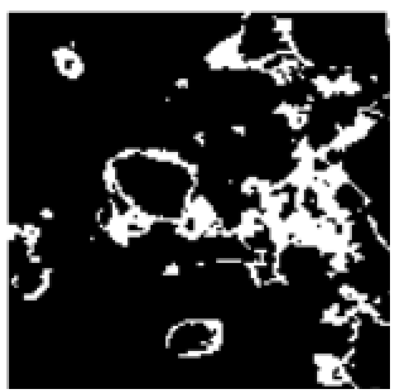




Ordinary least squares regression:	
- slope	1.05961
- y-intercept	8.95509
- root-mean-square error	25.91899

Pearson's coefficient:	
- r	0.81874

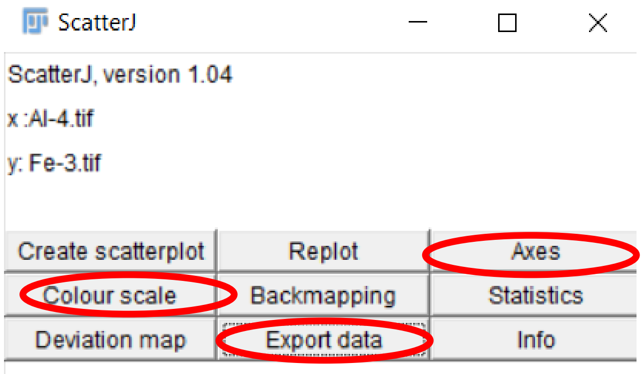


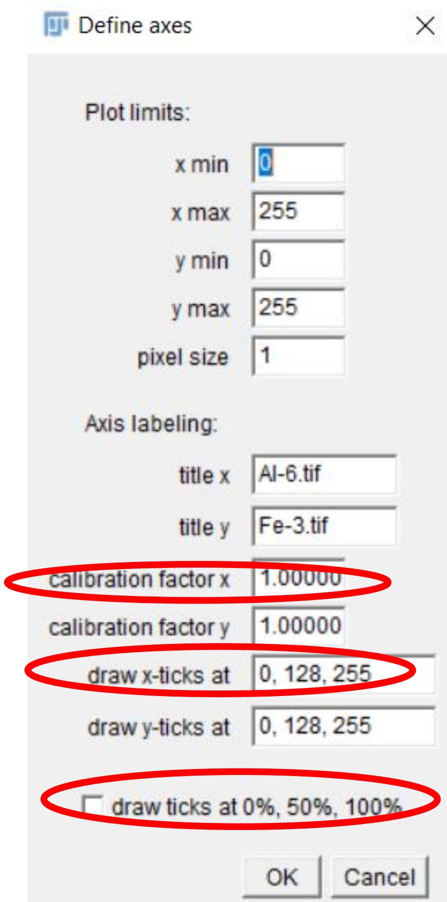


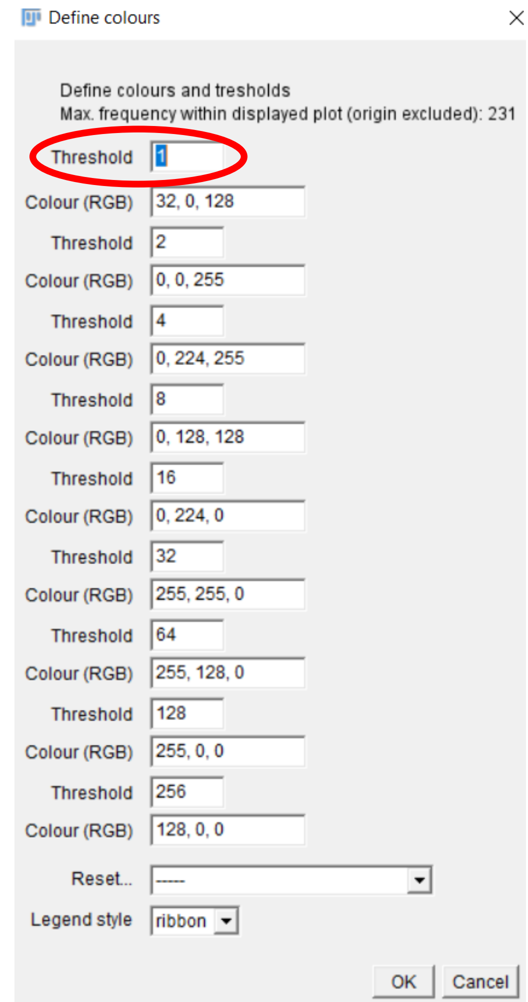
Ordinary least squares regression:	
- slope	0.15136
- y-intercept	62.74773
- root-mean-square error	11.49279

Pearson's coefficient:	
- r	0.17470

Correlation Analysis: ScatterJ







- Can change the axes to the "real" values by changing the calibration factor
- The values in the colours can also be changed to reflect the "real" values
- Data can be exported to use in another program. The data will be either save as 256 values or the "real" values, depending on what was used last