

## **An automated hybrid clustering technique applied to spectral data sets**

N.C. Wilson and C. M. MacRae  
CSIRO Minerals, Bayview Avenue, Clayton 3168, Australia

Spectral imaging allows users to collect data without prior knowledge of the sample composition. The downside of spectral imaging is that large data sets are produced and extracting the important information can be difficult. One approach for reducing the data is to use principal component analysis [1] which extracts the underlying chemical components. An alternate approach is to use automatic clustering algorithms [2] which classifies the data into groups.

Our typical map size is 1000 x 1000 steps, giving one million pixels, and up to 2048 channels or energy slices at each pixel. Our aim is to reduce this multi dimensional data set down to a single picture. This is achieved by classifying each pixel into a particular phase. A simple approach is to manually identify clusters of pixels from scatter and ternary plots. Pixels of a similar composition will congregate in a scatter plot. Manual selection is a difficult task on phases with multidimensional chemistry, introduces human bias, and leads to pixels not being defined in a cluster. To produce a better defined phase map based on all of the multidimensional information, we turn to automated clustering algorithms.

There are two main classes of clustering algorithms, hierarchical and non-hierarchical. In hierarchical clustering a series of associations are performed. Initially each of the  $m$  pixels are assigned to their own cluster, and then repeatedly the two closest clusters are merged, until all of the pixels are in one cluster. All the mergers are recorded, and this information gives a hierarchical tree. Algorithms of this type generally require memory storage of order  $m \times m$  bytes which becomes impractical for a million pixel map. In comparison, non-hierarchical algorithms are much better suited to large data sets, and are often used in remote sensing applications. The most commonly used non-hierarchical clustering method is the  $k$ -means algorithm [3], which is an iterative method that searches for the best set of  $k$  cluster centroids.

The  $k$ -means algorithm requires a choice of  $k$ , where  $k$  is the number of clusters to be found. The centroids are given an initial position in the  $n$  dimensional data set, then each pixel in the map is assigned to the closest centroid. The centroids are then moved in  $n$  dimensional space to the mean of the data points assigned to them, then with these new centroid positions, the data points are then reassigned to their closest centroid. This process is repeated until some convergence criterion is met, such as no further movement of the centroids.

There are various implementations of the  $k$ -means algorithm, using different definitions of distances between pixel and centroid and different choices for the initial positioning of clusters. For producing a satisfactory phase map the most important parameter is  $k$ , which sets an upper bound on the number of phases that can be found. The  $k$ -means algorithm moves the centroids to phases that contain the largest number of points, and so choosing a small number for  $k$  can lead to phases that occupy only a small number of pixels in the map being to be missed in complex samples. In response to this, the value of  $k$  can be increased, but this leads to an overwhelming number of phases which makes a phase map difficult to interpret.

To solve this problem, we have taken an approach that uses both hierarchical and non-hierarchical clustering techniques. The first step is to use a  $k$ -means type algorithm with a large value for  $k$ , for example 1000. Having reduced our original data set from the order of one million pixels to 1000 centroids, it is then possible to process the centroids using a hierarchical clustering method. The result of the hierarchical clustering is a tree, with the most significant chemical differences in the top branch and the least significant chemical differences at the bottom. This tree can then be used to interactively control the displayed phase map. Working down a branch of the tree, the broad top level groupings can be split to reveal more subtle classifications. Thus the user can easily visualize the different phases which are important to them, while ignoring other slight variations down other tree branches.

To illustrate this technique, clustering has been applied to a map taken on a cross section of a paint flake (Figure 1). This is from a detailed examination on the Mark Rothko painting called “no. 37 (Red)” [4]. The purpose of this study was to understand the chemistry of the paint layers used by Rothko in order to aid the conservation work on the painting. Samples like this are challenging to study as the artist has deliberately mixed many mineral pigments to achieve the desired effect. The clustering is able to resolve the bottom layer of the artists paint from the restorer’s layers above.

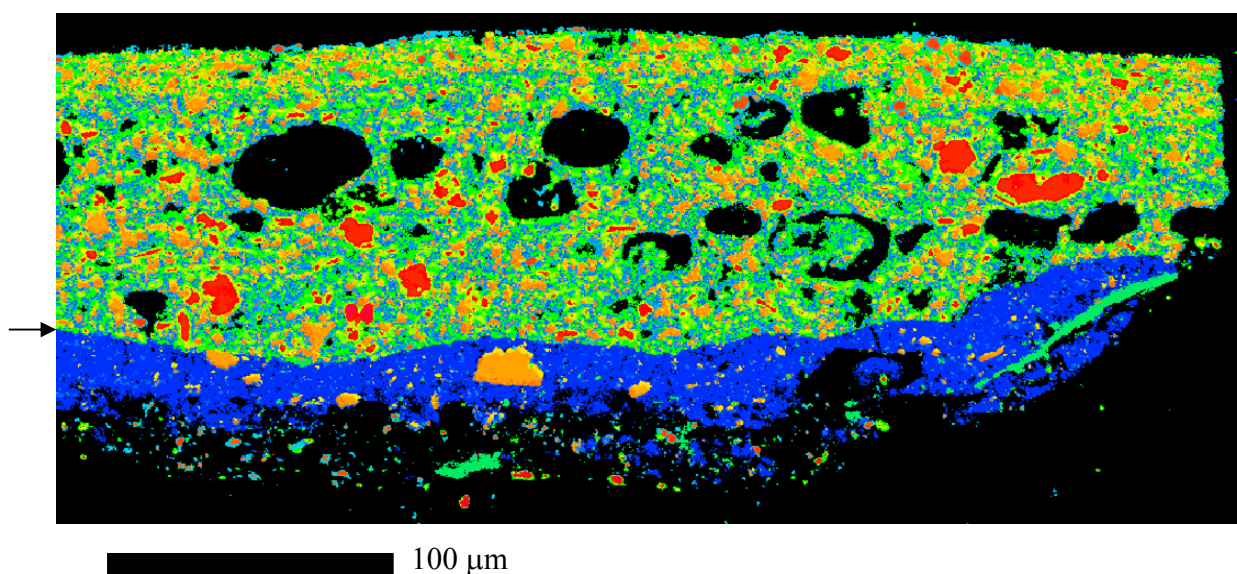


Figure 1. Clustered Rothko paint flake from “no. 37 (Red)” showing the 30 most chemically dissimilar clusters. The layer above the arrow is the restorer’s paint layer, while the layer below is the artist’s original painting.

#### References

- [1] P. Kotula et al., *Microsc. Microanaly.* (2003) 9, 1
- [2] N. C. Wilson, et al., *The 16th Australian conference on Electron Microscopy*, Canberra, 2000, 46
- [3] J. B. MacQueen, *Proc. Symp. Math. Statist. and Probability*, 5th Berkeley, 1, (1967) 281
- [4] The authors acknowledge the Conservation Department at the National Gallery of Victoria for making available the Rothko paint flake. The Rothko no.37 (Red) was painted in 1956 and is now owned by the National Gallery of Victoria, Australia.