SUPERVISED AND UNSUPERVISED MACHINE-LEARNING APPROACHES TO MINERAL SEGMENTATION. Matthew R. M. Izawa[1], Brendon J. Hall[2], [1]Institute for Planetary Materials, Okayama University, Misasa, Tottori, 682-0193, Japan (matthew.izawa@gmail.com), [2]Enthought Inc., 200 W Cesar Chavez, Suite 202, Austin TX 78701 USA (bhall@enthought.com).

Introduction: Many analytical methods produce spatial maps of some sample property, such as the relative or absolute concentration of an element, such as the data produced electron beam mapping (SEM-EDS, EPMA), using synchrotron X-ray sources, and by particle-induced X-ray emission. These datasets can be very large, and difficult to completely characterize manually. Image stacks provide a rich set of features that can be exploited by machine learning algorithms to augment traditional analysis. Both supervised and unsupervised classification methods can be applied to the analysis of microanalytical image stack data sets. Here, we compare a supervised approach, Support Vector Machine/Markov Random Field (SVM-MRF), and an unsupervised approach, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBScan) and K-means clustering. We compare the results of each with classical optical and SEM petrography and point-counting in the relatively simple case of determining the modal mineralogy of thin section of a shergottite meteorite, NWA 7257.

Sample description and data acquisition: Northwest Africa 7257 is an enriched mafic shergottite consisting dominantly of elongate euhedral to subhedral pyroxene laths (consisting of a mixture of pigeonite, augite, and ferropigeonite) with interstitial plagioclase. Accessory phases include ilmenite, Fe-Cr-Ti spinel, pyrrhotite, chlorapatite, merrillite, K-rich silicate melt pockets, rare baddeleyite and olivine [1]. Backscattered electron (BSE) imagery and EDS X-ray spectra were acquired for a polished thin section of NWA 7257 using a JEOL JSM-7001F scanning electron microscope with Oxford X-max detector. The X-ray spectra were sampled to produce images which map the intensity of individual elements (Al, C, Ca, Fe, K, Mg, Na, O, P, S, Si, Ti, Zr). To assess modal mineralogy, a point count of 961 points covering ~1.2 cm^2 of the section was conducted.

Supervised classification: In a supervised learning approach, the classes into which the data are to be sorted are assigned by a human user. In this case, multiple training sets were created by manually assigning representative pixels from each class, both from spatially contiguous zones and from isolated subregions representative of particular phases. The SVM-MRF approach [2] consists of two steps: 1) Spectral Classification and 2) Spatial Regularization. The spectral classification step uses a support vector machine (SVM) to perform a probabilistic classification for every pixel in the image [3]. The SVM learns from the training data by dividing the N dimensional feature space into k sub-regions by maximizing the margins between the classes. Each pixel of the image is mapped to one of the k sub-regions to obtain the spectral classification map. Class probabilities are determined using Platt scaling [4]. The spectral classification considers only the information available at each pixel location. Most structures in the image are larger than one pixel. The SVM does not take the spatial structure of the data into account, which can result in a noisy classification map. A Markov Random Field (MRF) regularization step is used to smooth the result. This approach is based on the assumption that a pixel belonging to a given class is likely to be surrounded by pixels having the same class.The regularization is formulated as the minimization of a suitable energy function, which is minimized using a version of the Metropolis algorithm (also sometimes termed simulated annealing) [5].

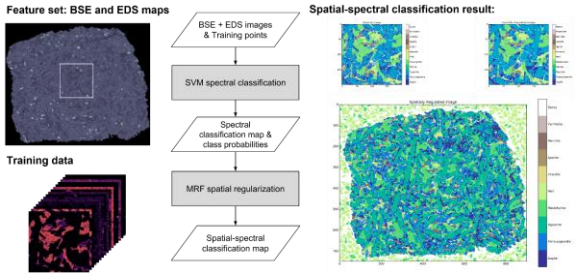

Figure 1: Schematic of the Support Vector Machine – Markov Random Field classification algorithm. An initial training data set was selected and pixel locations representative of different minerals (classes) were selected manually. This training data set was then used to train the SVM algorithm, which was then applied to the entire mapped area. The map and class probability output of the SVM was then used for spatial regularization to produce a spatial and spectral classification map and a mineral abundance map.

Unsupervised classification: In an unsupervised approach, there is no *a priori* definition of target classes. The assembled images are preprocessed to enhance contrast which normalizes intensity values within [0, 1]. The image stack is smoothed by applying a median filter to reduce noise. The images are stacked, so each pixel is associated with a 13-dimensional feature vector associated with the different EDS components.

The preprocessed data is passed to the HDBScan algorithm [6]. This projects each point into data space, and searches for significant regions of higher density (clusters) and separates them from lower density regions (noise). This density based approach enables rare minerals with relatively few points to be located. This approach allows for clusters of variable density to be identified, allowing for minerals with different component distributions. HDBScan also requires only one parameter, `min_cluster_size`, that indicates the smallest number of pixels that could belong to a relevant cluster.

The initial clustering can group together minerals of similar character. In particular, the pigeonite grains have distinct rims and cores that vary in relative amounts of Fe and Mg. These groups are further separated using the KMeans [5] partitioning algorithm.

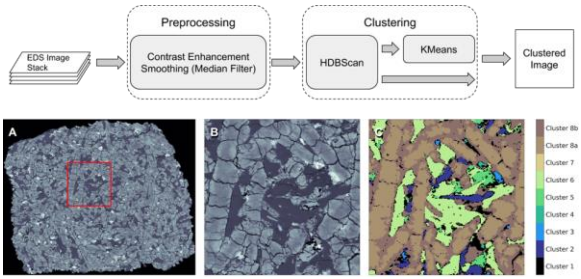

Figure 2: Schematic of the hierarchical clustering workflow. A) Contrast-enhanced BSE image of the NWA 7257 thin section. The 200×200 px subregion used in this study is indicated by the red square. B) Subregion used for this study. C) Results of hierarchical clustering applied to the test region. Pixels are grouped into clusters based on their density in data space

Results: The supervised classification approach clearly distinguishes many key petrographic features including the pigeonite cores of pyroxene laths, complex augite and ferropigeonite overgrowths, two phosphates (merrillite and chlorapatite) and interstitial maskelynite and K-rich melt.

The final unsupervised clustering results are shown for a sub-region of the entire dataset in Figure 1c. The assigned clusters align with mineral grain boundaries, and the cluster assignment has low noise. In order to interpret minerals, the distribution of the EDS components belonging to each assigned cluster was extracted, and manually assigned to a particular phase.

Determining modal mineralogy by point counting is a time-honoured but laborious process. The modal mineralogy as determined by point counting, supervised classification, and unsupervised classification are compared in Table 1.

	SVM+MRF	HDBScan	PC
Phase	Vol. %	Vol. %	Vol. %
Augite	6.5	n.d.	n.d.
Ferro-pigeonite	18.5	39.9	n.d.
Pigeonite	37.0	35.4	n.d.
Total Clinopyroxene	*62.0*	*75.3*	77.9
Maskelynite	22.2	15.9	17.6
High-K Melt	7.1	3.2	3.1
Low-K Melt	n.d.	0.6	n.d.
Ilmenite	1.2	1.6	1.3
Merrillite	4.2	3.3	3.8
Chlorapatite	0.1	n.d.	0.0
Pyrrhotite	0.5	0.1	0.0

Table 1: Modal mineral results using supervised, unsupervised, and point counting methods.

Discussion: Supervised classification using predetermined classes was able to identify the very minor phase chlorapatite, which was not identified by the unsupervised classifier. Of the three methods, only the supervised classification was able to reliably distinguish between the three forms of clinopyroxene present. On the other hand, the unsupervised classification approach was able to distinguish a K-poor melt phase which had not been identified in the training data sets or by manual analysis of the thin section. This illustrates a potential pitfall in SVM-based classification, which is that a class which is not included in the training data (e.g., a mineral that it is not expected to be present) may be missed, introducing a source of bias.

References: [1] Irving A. J. et al., (2012) 75th Meteoritical Society Meeting. [2] Tarabalka, Y. et al. (2010) IEEE Geoscience and Remote Sensing Letters, 7(4), 736–740. [3] Pedregosa, F. et al. (2011) JMLR 12, 2825–2830. [4] Lin H. T. et al. (2002) Mach. Learn., 68(3), 267–276. [5] Metropolis, A. et al. (1953) J. Chem. Phys., 21(6) 1087–1092. [6] Campello R.J.G.B. et al., (2013) *Advances in Knowledge Discovery and Data Mining*. PAKDD 2013. Lecture Notes in Computer Science, vol 7819. [7] Lloyd, Stuart P. (1982) *Information Theory*, IEEE Transactions 28.2, 129-137.