

# Analysing Rock Samples for the Mars Lander

Jonathan Oliver,  
Dept. EECS  
Uni. of California,  
Berkeley, CA 94720  
jono@eecs.berkeley.edu

Ted Roush, Paul Gazis  
NASA Ames Research Center,  
MS 245-3,  
Moffett Field, CA 94035  
{troush,gazis}@mail.arc.nasa.gov

Wray Buntine, Rohan Baxter  
& Steve Waterhouse,  
Ultimode Systems, 2560 Bancroft Way #213,  
Berkeley, CA 94704  
{wray,rohan,stevew}@ultimode.com

## Abstract

In the near future NASA intends to explore various regions of our solar system using robotic devices such as rovers, spacecraft, airplanes, and/or balloons. Such platforms will carry imaging devices, and a variety of analytical instruments intended to evaluate the chemical and mineralogical nature of the environment(s) that they encounter. The imaging and/or spectroscopic devices will acquire tremendous volumes of data. The communication band-widths are restrictive enough so that only a small portion of these data can actually be sent to Earth.

The aim of this research was to develop a system which analyses rock spectra to automatically determine which spectra are interesting, and to compress the spectral data for communication to Earth. In the research we report here we classify laboratory data using clustering techniques (ACPro an enhanced version of Auto-class) and provide the planetary scientists with a rapid, visually oriented method of evaluating the underlying chemical and mineralogical information contained within the clusters. We show how clustering can be used to identify interesting rock samples and estimate the compression that using such a system can achieve.

## Introduction

In the near future NASA intends to explore various regions of our solar system using robotic devices such as rovers, spacecraft, airplanes, and/or balloons. Such platforms will likely carry imaging devices, and a variety of analytical instruments intended to evaluate the chemical and mineralogical nature of the environment(s) that they encounter. Historically, mission operations have involved:

- return of scientific data from the craft;
- evaluation of the data by space scientists;
- recommendations of the scientists regarding future mission activity;
- transmission of commands to the craft; and
- activity by the craft in response to those commands.

This cycle is then repeated for the duration of the mission with command opportunities once or perhaps twice

per day. In a rapidly changing environment, such as might be encountered by a rover traversing hundreds of meters a day or a spacecraft encountering an asteroid, an operation cycle of this nature is not amenable to rapid long range traverses, discovery of novelty, or rapid response to any unexpected situations. In addition to real-time response issues, there are issues related to data volume. Modern imaging and/or spectroscopic devices can generate enormous amounts of data. Data volumes during a typical traverse can easily exceed on-board memory capabilities and communications bandwidth available for transmission back to Earth. This implies that some decisions regarding data selection and acquisition must be made on board the spacecraft. These decisions are distinct from electromechanical control, health, and navigation issues associated with robotic operations. They anticipate a long term goal of automating scientific discovery based upon data returned by sensors of the robot craft. Such an approach would eventually enable it to understand what is interesting because the data deviates from expectations generated by current theories/models of planetary processes that could have resulted in the observed data. Such interesting data and/or conclusions can then be selectively transmitted to Earth thus reducing memory and communications demands.

Here we report on one aspect of research that begins to address such on-board science understanding issues. We focus upon analysis and understanding of a data set intended to represent one which might be obtained by a robotic craft. This data consists of an extensive laboratory effort characterizing the amount of incident light that is reflected by the samples at visual and near-infrared (0.2-3 micrometer) wavelengths. From a geologic or planetary science perspective knowledge regarding the current rocks, minerals, and/or ices present on a surface, and their spatial and temporal distribution provide evidence regarding what evolutionary processes have been acting on a particular body and over what geological time scales.

Derivation of mineralogy or composition from reflectance spectra has involved a variety of qualitative, semi-quantitative, and quantitative approaches. One qualitative approach is spectral curve matching where

a comparison of the unknown spectrum to a catalog of some reference suite of spectra is performed (see (Gaffey *et al.* 1993) and references therein). This provides rapid identification of candidate minerals, but typically suffers from poor definitions of what a good match is, what a match actually implies, and incompleteness of comparison libraries. A semi-quantitative approach is spectral feature matching (Gaffey *et al.* 1993). This involves isolating and matching individual spectral features rather than the entire spectral curve. For spectra having well defined and isolated spectral features the interpretation is relatively unambiguous. However, for those spectra lacking clearly defined features, or mixtures where the features are no longer isolated, there is more considerable ambiguity. One quantitative approach relies upon empirical measurements of individual minerals and mixtures of these with other minerals (e.g. (Cloutis *et al.* 1986; 1990; Sunshine, Pieters, & Pratt 1990)). This approach relates diagnostic parameters (e.g. band area, band position, relative depths of bands, and relative spectral slopes) quantitatively extracted from measured spectra to the variables associated with a specific set of samples (e.g. grain size, compositional, or mineral structural variation). Interpretation of derived spectral parameters requires appropriate calibrations, the development of which requires a major laboratory effort. As a result only a limited number of calibrations exist. This empirical approach may be augmented by an analytical approach that relies upon calculations of the reflectance spectrum from the optical constants of candidate minerals (e.g. (Clark & Roush 1984; Nelson 1989; Cruikshank *et al.* 1993)). However, optical constants of candidate minerals are sparse. Neural network classification has been applied to asteroid spectra, but was not used to determine compositional information (Howell, Merenyi, & Lebofsky 1994; Merenyi *et al.* 1997). Thus an unsupervised mineral classifier would eliminate some of the subjectivity of the qualitative or semi-quantitative approaches while potentially eliminating the continued reliance on extensive laboratory work required for the more quantitative approaches. Mixture models have been successfully used to cluster spectra for a range of other applications (e.g., (Goebel *et al.* 1989; Adams, Smith, & Johnson 1986; Martin *et al.* 1996; Garcia-Haro, Gilabert, & Melia 1996)).

Our initial goals were to classify the laboratory data within the context of the KDD process (Fayyad, Piatetsky-Shapiro, & Smyth 1995). We applied clustering techniques and provided the planetary scientists with a rapid, visually oriented method of evaluating the underlying chemical and mineralogical information contained within the clusters.

## The Data Set

The laboratory data used was obtained from the US Geological Survey (USGS) and is described by (Clark *et al.* 1990; 1993) and at <http://speclab.cr.usgs.gov>. It consists of approximately 500 spectra of individual

minerals, plants, and elements. Many individual minerals provide a broad compositional sampling. In some cases, reflectance measurements of many different grain size separates of the same mineral are included within the data set. In addition to the measured reflectances of these samples, ancillary information such as chemical composition, specific mineral composition, grain size determination, and assessment of sample purity is provided.

The measurements consisted of the relative light intensity for 488 wavelengths (or channels where each channel is one wavelength of light). For example, the spectrometer readings for the mineral Acmite (Aegirine)(Pyroxene group) were:

Channel	0	1	2	3	...	487
Wavelength	0.205	0.213	0.221	0.229	...	3.232
Intensity	?	0.027	0.028	0.026	...	0.217

## An Initial Visualisation

The relative intensity readings for the 488 wavelengths is called a spectrum. We may plot a graph of the relative intensity readings for each wavelength. For example, the spectrum for Acmite (Aegirine)(Pyroxene group) is shown in Figure 1.

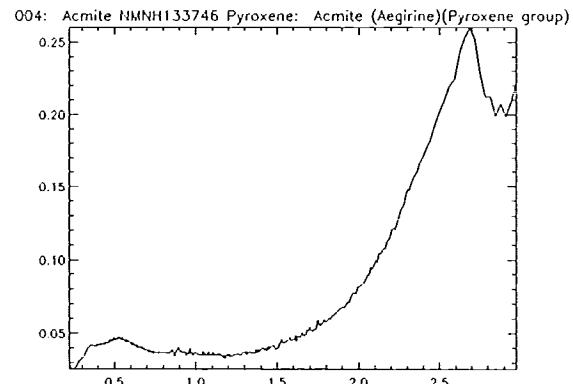


Figure 1: The spectrum for Acmite

The USGS data set consists of the spectra for 497 individual minerals, plants, and elements. These spectra range over such a broad range of shapes and albedos<sup>1</sup> that it is difficult, if not impossible, for even an expert to determine from the spectra into which classes different minerals might fall.

## Clustering the USGS Data Set

Since the USGS data set as a whole was complex and difficult to understand, we decided that a first step towards understanding the data set would be to cluster the spectra into groups where the spectra in each group would hopefully be similar.

<sup>1</sup>The albedo is the ratio of the flux reflected from a surface to the flux that is incident upon it.

## Clustering with ACPro

To cluster the 497 spectra, we used ACPro (<http://www.ultimode.com>). ACPro is based on the successful AutoClass (Cheeseman *et al.* 1988; Cheeseman & Stutz 1995) and Snob (Wallace & Boulton 1968; Wallace 1990; Oliver, Baxter, & Wallace 1996) research programs, and is being developed with the assistance of the NASA AutoClass team using a NASA commercialisation award. These use mixture models (Everitt & Hand 1981; McLachlan 1992) to represent clusters. The use of data-based prior distributions for parameter values allows automated selection of the number of clusters and for the means, variances and relative abundance of the clusters to have reasonable values.

## Problematic Issues for Clustering Spectra

A major consideration when clustering the USGS data set was that there was a large number of attributes (488), with a small number of records (497). This led to the problem that if we assumed a full covariance matrix there would have an extraordinary number of parameters to estimate (of the order  $\frac{488}{2} * 487$  parameters for each cluster). We therefore assumed a multivariate Gaussian distribution with a restricted (diagonal) covariance matrix.

An important feature of ACPro (available in Snob, but not in Autoclass) is the ability to make attributes of a cluster "irrelevant". This means that when an attribute's characteristics within a cluster do not differ significantly from the population attribute characteristics, the population characteristics serve as the "default" characteristics. This feature was important with the current data sets since they had 488 attributes and only 497 records. This provides a simple hierarchical clustering model, and greatly reduces the number of parameters to be estimated.

## An Initial Clustering

Our initial clustering was performed by using each of the 488 channels as attributes, and each of the 497 rock samples as records. ACPro found 30 clusters when applied to these data. Many of the clusters found corresponded to recognized mineral groups or sub-groups, but others were difficult to explain. For example, the samples in cluster 0 were all olivine's with associated concentrations of  $FeO$ ,  $MgO$  and  $SiO_2$ . On the other hand, six of the 30 clusters produced were default (or junk) clusters, whose samples lacked special properties which fit nicely into other clusters.

## Transforming the Data

To investigate alternatives to clustering the raw data, we considered transforms of the raw spectral data, and clustered the transformed data using ACPro. We considered two specific transformations:

- A difference operator — here we let  $x_{i,j}$  be the relative intensity reading for sample  $i$  ( $i \in [0, 496]$ ) at channel  $j$  ( $j \in [0, 487]$ ). We let  $y_{i,j} = x_{i,j+1} - x_{i,j}$

for  $j \in [0, 486]$  and assumed the  $y_{i,j}$  followed a multivariate Gaussian distribution with a diagonal covariance matrix.

- A convex hull operator — the background was removed to emphasize features in the spectra that corresponded to spectral lines. Such a procedure eliminates absolute reflectance information, but focuses on isolating specific sample absorption bands that are superposed upon this continuum. The background removal process involved a 3-channel running average to suppress noise spikes followed by a 'convex hull operator' (Grove, Hook, & Paylor II 1992). The convex hull operator estimates a series of straight line segments to local maxima in the reflectance spectrum. The values of the reflectance spectrum are subtracted from this mathematical estimate of a continuum (i.e. no absorption) thus eliminating albedo information yet identifying and isolating regions within the reflectance spectrum where absorptions are located.

**Clustering the Differenced Data** ACPro found 20 clusters when applied to the  $y_{i,j}$ . Some of these clearly are related to specific samples that share a common property in composition. However, approximately half of the clusters represent default groupings, mixing many samples of sharing little or no common properties.

**Clustering the Convex Hull Data** ACPro found 45 clusters when applied to data transformed by the convex hull operator of which 22 are shown in Figure 2. The spectral properties within the clusters were generally very good and only four clusters represented default groupings. Of the 16 plants contained in the data, 10 are contained in cluster 16 and 4 in cluster 17. The spectra in both clusters have very similar overall shapes but the distinction between the clusters appears to be related to the strength of the broad reflectance minimum near  $1.0 \mu m$ . Cluster 21 consists entirely of the sulfate mineral jarosite. Many clusters consist of materials with similar overall spectral shape but distinctively different albedo levels, e.g. cluster 4 and 5. This suggests that future analyses should also retain information regarding absolute albedo information.

## Visualising the Clusters

In addition to evaluating the spectral characteristics of each cluster, the ACPro Visualiser also allows the geologist to investigate the ancillary compositional information contained within the data. An example is provided in Figure 3 which is a histogram of the  $Fe_2O_3$  content of as a function of each cluster. One can rapidly determine that cluster 36 has the highest content.

Clusters with low concentrations of  $Fe_2O_3$  are not shown in the histogram. Alternatively the composition of each cluster can be investigated. An example is provided in Figure 4 which shows the elemental abundances determined for cluster 20. The dominance of  $FeO$ ,  $Al_2O_3$ , and  $SiO_2$  suggest that these are Almandine garnets with minor elemental substitutions of  $Mg$ ,

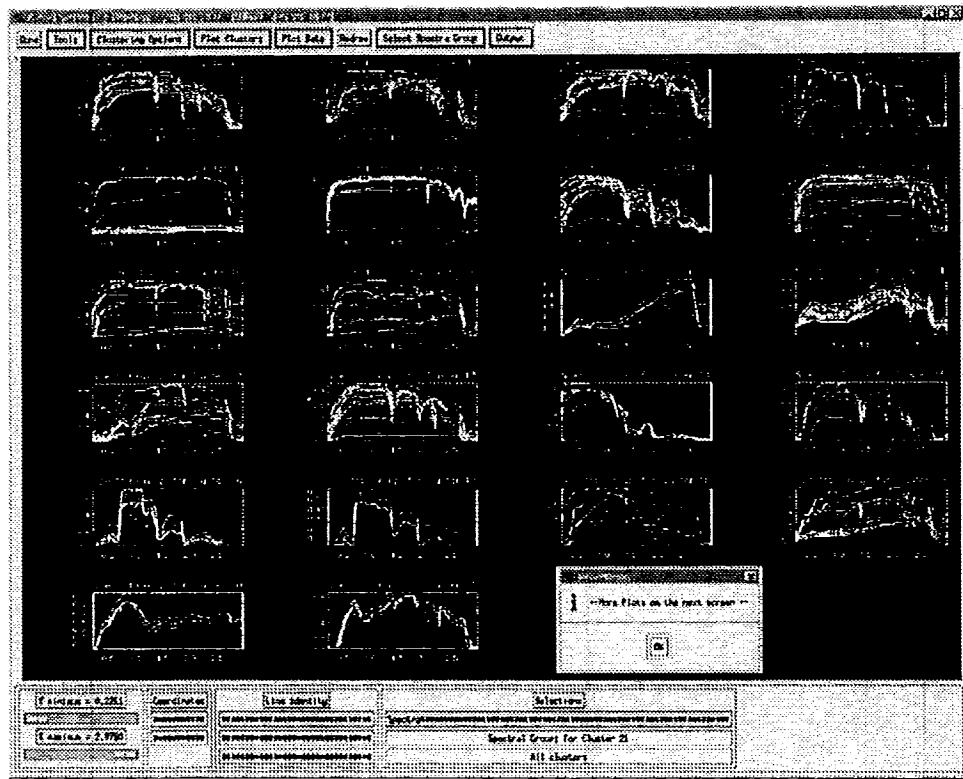


Figure 2: 22 Clusters from the Convex Hull Clustering

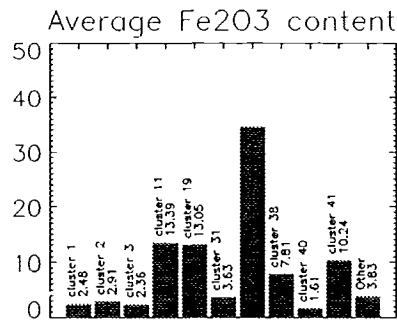


Figure 3: The Average  $Fe_2O_3$  Content for each Cluster Containing  $Fe_2O_3$

$Ca$ , and  $Mn$ . A more complete examination of the clusters is currently underway.

### Compression of the USGS Data Set

We now consider the problem of compressing the spectral data for communication to Earth. Clustering models (specifically mixture models used by ACPro) such as the ones described here may be used for data compression (Wallace & Boulton 1968; Cover & Thomas 1991). We assume that the relative light intensity for each of the 488 channels is measured to an accuracy of 0.0005.

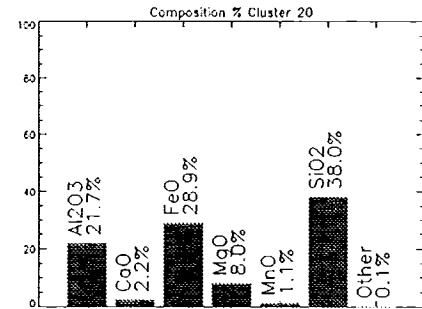


Figure 4: The Elemental Abundances for Cluster 20

Table 1 gives details for the compression rates on the USGS data set for the three data sets we describe. The original data set could be encoded in 2,616,343 bits, while the convex hull transformed data could be encoded in 1,814,786 bits.

### Discussion

The interpretation of the clusters found and the results in Table 1 lead us to the following conclusions:

- The difference operator was inappropriate for this application since it led to a clustering which was difficult to interpret, and it didn't compress the data.
- The convex hull operator was very useful for this ap-

	Raw Data	Differenced Data	Convex Hull Data
Bits with no clustering	2,616,343	2,865,597	2,155,346
# clusters	30	20	45
Bits using clustering	2,202,369	2,716,124	1,814,786

Table 1: Bits to Transmit the USGS Data Set.

plication since it could be explained by planetary scientists at NASA, and it led to compression of the data.

## Conclusion

We developed a system for the analysis of rock spectra using clustering techniques. We clustered the USGS laboratory data using ACPro and provided the planetary scientists at NASA Ames with a rapid, visually oriented method of evaluating the underlying chemical and mineralogical information contained within the clusters. In addition, we found that clustering was useful for the compression of spectral data for communication to Earth.

## Acknowledgments

We would like to thank the anonymous referees for their helpful comments.

## References

- Adams, J. B.; Smith, M. O.; and Johnson, P. E. 1986. Spectral mixture modeling - a new analysis of rock and soil types at the viking lander-1 site. *Journal of Geophysical Research-Solid Earth and Planets* 91(NB8):8098-8112.
- Cheeseman, P., and Stutz, J. 1995. Bayesian classification (AUTOCLASS): Theory and results. In Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R., eds., *Advances in Knowledge Discovery and Data Mining*. The AAAI Press, Menlo Park.
- Cheeseman, P.; Self, M.; Kelly, J.; Taylor, W.; Freeman, D.; and Stutz, J. 1988. Bayesian classification. In *Seventh National Conference on Artificial Intelligence*, 607-611.
- Clark, R., and Roush, T. 1984. Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications. *Geophysical Research* 80:6329-6340.
- Clark, R.; King, T.; Swayze, G.; and Vergo, N. 1990. High spectral resolution reflectance spectroscopy of minerals. *Geophysical Research* 95:12653-12680.
- Clark, R.; Swayze, G.; Gallagher, A.; King, T.; and Calvin, W. 1993. The U.S. geological survey, digital spectral library: Version 1: 0.2 to 3.0 microns. Technical Report Open File Report 93-592, The U.S. Geological Survey.
- Cloutis, E.; Gaffey, M.; Jakowski, T.; and Reed, K. 1986. Calibration of phase abundance, composition, and particle size distribution for olivine-orthopyroxene mixtures from reflectance spectra. *Geophysical Research* 91:11641-11653.
- Cloutis, E.; Gaffey, M.; Smith, D.; and Lambert, R. 1990. Metal silicate mixtures: Spectral properties and applications to asteroid taxonomy. *Geophysical Research* 95:8323-8338.
- Cover, T., and Thomas, J. 1991. *Elements of Information Theory*. New York: John Wiley and Sons, Inc.
- Cruikshank, D.; Roush, T.; Owen, T.; Geballe, T.R.; de Bergh, C.; Schmitt, B.; Brown, R.; and Bartholomew, M. 1993. Ices on the surface of triton. *Science* 261:742-745.
- Everitt, B., and Hand, D. 1981. *Finite Mixture Distributions*. London: Chapman and Hall.
- Fayyad, U. M.; Piatetsky-Shapiro, G.; and Smyth, P. 1995. From data-mining to knowledge discovery: An overview. In *Knowledge Discovery in Data Bases II*. chapter 1, 1-37.
- Gaffey, M.; Lebofsky, L.; Nelson, M.; and Jones, T. 1993. Asteroid surface compositions from earth-based reflectance spectroscopy. In *Remote Geochemical Analyses: Elemental and Mineralogical composition*. New York: Cambridge University Press.
- Garciaharo, F. J.; Gilabert, M. A.; and Melia, J. 1996. Linear spectral mixture modeling to estimate vegetation amount from optical spectral data. *International Journal of Remote Sensing* 17(17):3373-3400.
- Goebel, J.; Volk, K.; Walker, H.; Gerbault, F.; Cheeseman, P.; Self, M.; Stutz, J.; and Taylor, W. 1989. A bayesian classification of the IRAS LRS atlas. *Astronomical Astrophysics* 222:L5-L8.
- Grove, C.; Hook, S.; and Paylor II, E. 1992. Laboratory reflectance spectra of 160 minerals, 0.4 to 2.5 micrometers. Technical report, Jet Propulsion Laboratory, Pasadena California.
- Howell, E.; Merenyi, E.; and Lebofsky, L. 1994. Classification of asteroid spectra using a neural network. *Geophysical Research* 99:10847-10865.
- Martin, P. D.; Pinet, P. C.; Bacon, R.; Rousset, A.; and Bellagh, F. 1996. Martian surface mineralogy from 0.8 to 1.05 mu-m tiger spectroimager measurements in terra-sirenum and tharsis-montes formation. *Planetary and Space Science* 44(8):859-888.
- McLachlan, G. 1992. *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley.
- Merenyi, E.; Howell, E.; Rivkin, A.; and Lebofsky, L. 1997. Prediction of water in asteroids from spectral data shortward of 3 mm. *Icarus* 129:421-439.
- Nelson, M. 1989. *Determination of modal composition of intimate mineral mixtures using bidirectional reflectance theory*. Ph.D. Dissertation, Univ. of Hawaii.
- Oliver, J.; Baxter, R.; and Wallace, C. 1996. Unsupervised Learning using MML. In *Machine Learning: Proceedings of the Thirteenth International Conference (ICML 96)*, 364-372. Available on the WWW from <http://www.cs.monash.edu.au/~jono>.
- Sunshine, J.; Pieters, C.; and Pratt, S. 1990. Deconvolution of mineral absorption bands: An improved approach. *Geophysical Research* 95:6955-6966.
- Wallace, C., and Boulton, D. 1968. An information measure for classification. *Computer Journal* 11:185-194.
- Wallace, C. 1990. Classification by minimum-message-length inference. In Goos, G., and Hartmanis, J., eds., *Advances in Computing and Information - ICCI '90*. Berlin: Springer-Verlag. 72-81.