Automated classification of visible and infrared spectra using cluster analysis

G. A. Marzo,[1] T. L. Roush,[1] and R. C. Hogan[2]

Received 10 August 2008; revised 14 January 2009; accepted 5 May 2009; published 11 August 2009.

[1] Planetary space experiments collect large volumes of data whose scientific content requires understanding. Marzo et al. (2006) presented an unsupervised cluster analysis scheme that is able to reduce a spectral data set to a few clusters, allowing for more focused and rapid evaluation of their scientific meaning. Here, we extend the original approach to account for the measurement uncertainty and build a classification scheme. We apply the clustering technique to the ASTER and RELAB libraries of visible and infrared spectral reflectance. These spectral libraries are documented, allowing assignment of a label to each spectrum reflecting its physical and chemical properties. We assess the ability of the original and extended approaches to identify natural clusters of the library spectra and estimate associated uncertainties of the results. We evaluate the scientific meaning of the derived clusters based on the labels contained within each cluster. Once the cluster meanings are defined, we test our classification scheme using a training-testing approach and evaluate the accuracy of assigning the unknown spectra to the correct cluster.

Citation: Marzo, G. A., T. L. Roush, and R. C. Hogan (2009), Automated classification of visible and infrared spectra using cluster analysis, *J. Geophys. Res.*, *114*, E08001, doi:10.1029/2008JE003250.

1. Introduction

[2] Planetary space experiments are currently collecting a huge amount of information which needs to be explored in order to understand its scientific content. In particular, the analysis of spectral measurements, with the advent of planetary spectrometers able to entirely map a planet, constitutes a challenge to researchers. Among the others, TES (Thermal Emission Spectrometer) on board Mars Global Surveyor [*Christensen et al.*, 2001] and, more recently, OMEGA (Observatoire pour la Minéralogie, l'Eau, les Glaces et l'Activité) on board Mars Express [*Bibring et al.*, 2006] and CRISM (Compact Reconnaissance Imaging Spectrometer for Mars) on board Mars Reconnaissance Orbiter [*Murchie et al.*, 2007] are currently completing the spatial and temporal coverage of Mars, providing the scientific community with high-resolution spectral information at a spatial scale up to 18 m/pixel [*Murchie et al.*, 2007]. In the case of CRISM, such a level of detail translates into ~34 Terabits of visible and infrared spectra.

[3] Compositional interpretative efforts can be greatly facilitated by an automated classification scheme that is robust and scientifically meaningful. Such a scheme would also be an essential element of intelligent remote sensing systems that will operate on, or near, distant planets, moons or asteroids [*Castaño et al.*, 2003; *Roush et al.*, 2004], e.g., next generation Mars rovers [*Vago et al.*, 2006; *Crisp et al.*, 2008]. Ultimately such systems are intended to operate for weeks without human intervention. Therefore such systems must be autonomously capable of responding to the scientific content of sensor data within a science driven mission scenario.

[4] Techniques which seek to separate data into constituent groups are commonly referred to as cluster analysis, and are extensively applied in diverse research fields [*Everitt*, 1980]. In general, clustering techniques can be divided into two groups: supervised and unsupervised. Supervised clustering splits the data set into predefined classes based on training data which define class boundaries in a multivariate space. By contrast, unsupervised clustering splits the data set into classes based on the natural distribution of the data in the multivariate space. These techniques produce entirely arbitrary output classes and require a posteriori interpretation. *Marzo et al.* [2006, 2008] developed and evaluated an unsupervised statistical clustering scheme able to reduce a spectral data set to a few clusters allowing for more focused and rapid evaluation of their scientific meaning. The technique adopts a partitioning clustering algorithm based upon the *Calinski and Harabasz* [1974] criterion for identifying the natural number of clusters.

[5] As an extension to the previous clustering approach of *Marzo et al.* [2006] we build and evaluate a classification scheme following the approach of *Roush and Hogan* [2007]. Here we also introduce a refinement to the original criterion for detecting the number of clusters to account for the measurement uncertainty. We report the application of

[1]Space Science and Astrobiology Division, NASA Ames Research Center, Moffett Field, California, USA.
[2]Bay Area Environmental Research Institute, NASA Ames Research Center, Moffett Field, California, USA.

0148-0227/09/2008JE003250

Table 1. Hierarchical Expert Labels and Their Occurrence in Libraries[a]

			ASTER			RELAB		
Class	Subclass	Group	C	M	F	C	M	F
Element	unspecified	unspecified	2	2	2			9
Metal	sulfide	unspecified	10	10	10		4	6
Oxide-hydroxide	hydroxide	Al			1			1
		Fe				2	3	2
		Mg	1	1	1			
		unspecified					2	
	oxide	hematite	2	2	3	2	1	5
		rutile	5	5	6			2
		spinel	2	2	2		2	51
		$X_2O + XO$			1			1
		unspecified						10
Salt	arsenate	unspecified	2	2	2			
	carbonate	anhydrous	12	12	12		17	20
		hydrous	3	3	3		8	13
	halide	chlorides	3	3	2			1
		fluorides	2	2	2			
		unspecified	2	1				
	phosphate	unspecified	4	4	4			5
	sulfate	anhydrous	6	6	6		1	3
		hydrous	7	7	7		4	31
		unspecified						1
	tungstate	unspecified	1	1	1			

[a]Abbreviations: C, coarse; M, medium; F, fine (see text).

the classification technique to ASTER (http://speclib.jpl.nasa.gov/) and RELAB (http://www.planetary.brown.edu/relab/ [*Pieters and Hiroi*, 2004]) reflectance spectral libraries of minerals, focusing our analyses on the visible and infrared, and on three different particle size ranges. These spectral libraries, discussed in more detail in next section, are documented databases that allow assignment of a label to each spectrum reflecting its physical and chemical properties, resulting in an appropriate data set for training and evaluating our classifier. In this work, in particular, we assess the ability of the original and extended criteria to identify natural clusters of the library spectra, estimate associated uncertainties of the results, and evaluate the scientific meaning of the derived clusters, based upon labels selected to represent the spectra contained within each cluster. Our classification scheme is tested by (1) determining clusters from a randomly selected subset of the spectral libraries; (2) associating a label with each cluster; (3) assigning new spectra not included in the randomly selected set to a cluster: and (4) evaluating the accuracy of assigning the new spectra to the correct cluster.

[6] In the next two sections a description of the library data sets and of the clustering scheme is provided. This is followed by a discussion of the training of the classifier and associated results. We then present the results of testing a trained classifier and the associated confusions and accuracies. Finally, we discuss our results and present our conclusions.

2. Data Selection

[7] In this work we use data from ASTER and RELAB libraries containing two independent sets of reflectance spectral measurements of various minerals. Spectral libraries include extensively studied materials which are therefore appropriate for training our classifier and evaluating the classifications obtained in the subsequent testing phase. We limited spectral information to the 0.4–6 μm wavelength region and, hereafter, we refer to this broad spectral region as VIS-IR. This wavelength region is representative of the range chiefly dominated by reflected solar energy. In this limited range all samples from the two libraries have overlapping wavelength coverage even if they are obtained with different spectral sampling. In the spectral region studied, ASTER library contains spectra with different spectral sampling from 0.001 to 0.007 μm and RELAB data present spectral sampling from 0.0009 to 0.007 μm. In order to provide more samples for the analysis, we used a linear interpolation to resample all the spectral data to a common set of wavelengths at the lowest spectral resolution available for the two libraries. The resulting number of bands of each spectrum is 2429 and has a spectral sampling of 0.001 to 0.007 μm at the shortest and longest wavelengths, respectively.

[8] The spectral libraries also contain descriptions of the sample composition and information regarding their particle size. The compositional information is used to associate a set of hierarchical labels with each sample; here we use the labels described by *Roush and Hogan* [2007] and reported in Tables 1 and 2. The label hierarchy describes an increasing level of mineralogical information which breaks down into mineral class, subclass, and group names. The ASTER library includes specific and consistent definitions for three particle size ranges of the samples; fine, medium, and coarse grain having grains diameters of <45, 45–125, and >250 μm, respectively. These names are also used to provide a grain size label for the RELAB data. However, the RELAB data exhibit a broader range of individual grain sizes within each of these categories. For each RELAB sample we identify the mean of its grain size distribution and use this value to associate it with the ASTER grain size (e.g., RELAB Hematite grain size 0–100 micron is labeled as medium grain size using the consistent ASTER defini-

Table 2. Hierarchical Expert Labels and Their Occurrence in the Data Set[a]

			ASTER			RELAB		
Class	Subclass	Group	C	M	F	C	M	F
Silicate	cyclosilicate	cyclosilic	4	4	4			
	inosilicate	amphibole	6	6	6			
		Ca-amphi						1
		clinopyro	3	3	3	1	4	14
		ortopyro	1	1	2	1	3	2
		pyroxene						1
		pyroxenoid	2	2	2			2
		unspecified						1
	nesosilicate	Al_2SiO_5	1	1	1			
		garnet	2	2	2			7
		humite	1	1	1			
		olivine	1	1	2	3	11	10
		zircon	1	1	1			2
	phyllosilicate	chlorite	7	6	8		2	
		clay	5	5	13	2	30	21
		mica	7	7	7		2	8
		serpentine	1	1	1		7	10
		unspecified					1	11
	sorosilicate	epidote	2	2	2			
		hemimorph	1	1	1			
		idocrase	1	1	1			
		melilite						7
		unspecified						3
	tectosilicate	alcali-feld	2	2	3			
		felspathoid	1	1	1			1
		nephelene						1
		plagio glass				2		
		plagio feld	6	6	7	2	1	24
		SiO_2	4	5	6	3	6	9
		scapolite			1			
		zeolite	2	2	3			28
		unspecified					1	
	unspecified	unspecified						6

[a]Abbreviations: C, coarse; M, medium; F, fine (see text).

tions). Table 1 shows the number of spectra associated with each label contained within each spectral library for several different classes and Table 2 shows the information for the silicate class. In the column at the left of both Tables 1 and 2 the coarsest level of hierarchy label is class, where the largest difference between the chemistry and structures of the materials exists, e.g., silicates versus salts. In the next column the subclass label provides more detailed chemical and structural information, e.g., structures formed by combining silicate tetrahedra into mineral structures. Finally, the group labels provide a higher level of distinction between chemistry and structural types, e.g., clinopyroxenes versus orthopyroxenes. Inspection of Tables 1 and 2 reveals that the ASTER spectral library has a quasi-homogeneous distribution of spectra over different grain sizes, while, in the RELAB library, the finest grain size has the most number of samples and several categories do not have corresponding samples in the medium and coarse grain sizes. This will make comparison of the results for different grain sizes difficult, because particle size has an effect on the spectral properties [*Gaffey et al.*, 1993]. In general, as particle size decreases the reflectance of samples increases and the amplitude of spectral features decreases. Thus one might anticipate that the ability to distinguish between different minerals might be diminished as the grain size decreases. To partly address this issue we use both raw and normalized reflectance data. To create normalized data we scale all spectra to set their maximum reflectance to unity. Such a scaling removes differences in reflectance levels without affecting the amplitude of spectral features as a function of grain size.

[9] There are two steps in the classification process: training and testing. In order to retain data for testing we require a particular sample set to have at least three spectra at the group level. Two spectra are randomly selected for training so at least one spectrum is reserved for testing. This process implies that only a fraction of the original mineral groups are represented in the training phase. In Table 3 we summarize the number of groups actually represented during training, and the corresponding number of subclasses and classes. For completeness, in the last row of Table 3, we also report the total number of spectra used in the training phase. For each grain size we independently train and test for ten cases (realizations) of cluster analyses of both raw

Table 3. Training Population in Terms of Number of Spectra Representing Group, Subclass, and Class Labels[a]

	A + R			ASTER		
	C	M	F	C	M	F
Classes	4	4	5	4	4	4
Subclasses	11	12	14	10	10	10
Groups	18	22	32	16	16	18
Training spectra	36	44	64	32	32	36

[a]Abbreviation: A + R, ASTER + RELAB.

and normalized data. We repeat these operations using the combination of ASTER and RELAB libraries and the ASTER library by itself. This provides the ability to evaluate and generalize the results. In all of our analyses we do not include spectra with the label 'Unspecified' as there is typically insufficient information to provide any more detailed labels, or from a scientific perspective they represent minerals that only occur in minor abundances in rocks [*Roush and Hogan*, 2007].

3. Clustering Scheme

[10] We use the clustering approach developed by *Marzo et al.* [2006]. It is based on the K-means cluster algorithm [*MacQueen*, 1967; *Hartigan and Wong*, 1979] but eliminates the need to specify the number of clusters a priori, by including a criterion able to identify the natural number of clusters present in the data set [*Calinski and Harabasz*, 1974]. Our cluster analysis is a statistical multivariate framework that iteratively group spectra by minimizing the dimension of the clusters and, simultaneously, maximizing their distance in the data space, whose dimensions are defined by the number of spectral bands in the data set. We refer the reader to *Marzo et al.* [2006] for a detailed description of the technique, while here some details about the criterion, and its extension, are reported.

[11] To find the natural number of clusters in a data set, *Calinski and Harabasz* [1974] propose the use of the ratio CH given by

$$CH = \frac{\overline{\mathbf{B}}}{g-1} \Big/ \frac{\overline{\mathbf{W}}}{N-g} \tag{1}$$

where N is the number of the samples and g the number of clusters considered. Once the g clusters are selected, it is possible to define the between group, $\mathbf{B}_c$, and within group, $\mathbf{W}_c$, covariance matrices for the cth cluster. In this context $\mathbf{B} = \sum_{c=1}^{g} \mathbf{B}_c$ and $\mathbf{W} = \sum_{c=1}^{g} \mathbf{W}_c$ are, respectively, the between and within group dispersion matrix [*Marzo et al.*, 2006], and the overline sign indicates the trace of the matrix. $\overline{\mathbf{B}}$ and $\overline{\mathbf{W}}$ represent the average distance between clusters and the average radius of each cluster, respectively. In this expression a value of CH increasing monotonically with g suggests no cluster structure, CH decreasing monotonically with g suggests a hierarchical structure, and CH rising to a maximum at g suggests the natural presence of g clusters. Generally, some relative maxima arise, providing different solutions for the natural number of clusters included in the data set. The strongest interpretation of the criterion implies using the absolute maximum as the natural number of clusters. However, a degenerate absolute maximum will occur when the number of clusters is equal to the number of samples because of the singularity in the definition of CH at this limit. The cases examined in the present work generally show a maximum at this limit and thus suggests that the spectra included in the reflectance VIS-IR libraries are distinct from one another enough to consider each one a different cluster. This is not surprising because, assuming no measurement uncertainty, each spectrum represents a different material characterized by a unique chemical and physical structure. Therefore we adopt the weaker interpretation of the criterion, which considers each maximum as a possible candidate for the natural number of clusters present in the data set. In particular, here we decide to consider the next to last maximum. This choice assures the most possible diversity in the realization of clusters and removes the degenerate case $g = N$.

[12] The definition of the CH criterion can be extended to include the effect of an estimated uncertainty for a single measurement. From the geometrical point of view, a set of spectral measurements of the same mineral and grain size forms a spheroid in the multidimensional space whose radius relates to the measurement uncertainty whatever its cause. Here we are assuming that the measurement uncertainty is Gaussian distributed and this assumption is valid across the entire spectral range for each spectra. Overlapping clusters whose radius is comparable to the uncertainties involved are not distinguishable from each other. If $\mathbf{W}_{err}$ is the lower limit of $\overline{\mathbf{W}_c}$, the two quantities representing respectively the measurement uncertainty and the radius of the cth cluster, for a given cluster we propose to define a new cluster radius as following:

$$\overline{\mathbf{W}'_c} = \max\left(\overline{\mathbf{W}_c}, \mathbf{W}_{err}\right). \tag{2}$$

Such a modification only penalizes small overlapping clusters. Large clusters should also be penalized. That can be achieved by replacing $\overline{\mathbf{W}}$ in equation (1) with the following weighted sum:

$$\overline{\mathbf{W}'} = \sum_{c=1}^{g} \left(\frac{\overline{\mathbf{W}'_c}}{\mathbf{W}_{err}} \right) \overline{\mathbf{W}'_c}, \tag{3}$$

obtaining

$$CH = \frac{\overline{\mathbf{B}}}{g-1} \Big/ \frac{\overline{\mathbf{W}'}}{N-g} \tag{4}$$

[13] Here we report results obtained by using both the original definition (OD, equation (1)) and the extended definition (ED, equation (4)). When using the OD the next to last maximum is considered and, when using the ED, we assume a 2% measurement error and the strongest interpretation of the criterion, i.e., considering the CH absolute maximum. An example of both cases is reported in Figure 1, left. We evaluate our assumptions below.

[14] Before their application, we evaluate both definitions using ten different realizations of the clustering for two different cases: choosing randomly the cluster initialization, i.e., the cluster centroids which are iteratively relocated in the clustering process [*MacQueen*, 1967], for a fixed training data set (Figure 1, left) and choosing randomly the training data set (Figure 1, right). In addition, for the extended definition, we vary the assumed measurement error comparing 1%, 2%, and 3% keeping fixed both the data set and the cluster initialization (Figure 1, middle). The random selection of the data set (Figure 1, right), even if performed at the group level, implies that different data sets can be constituted by very distinct spectra, e.g., in the clay

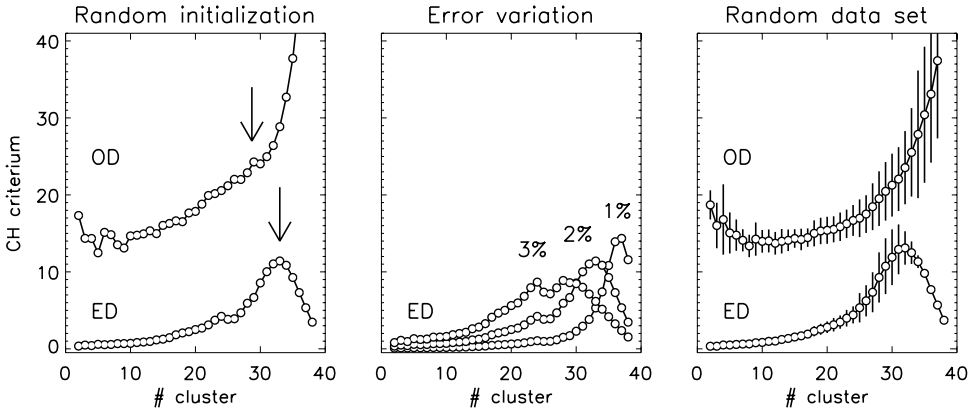

Figure 1. The average CH criterion curves using original (OD, equation (1)) and extended definition (ED, equation (4)) of clustering criterion in different conditions: (left) choosing randomly the cluster initialization for a fixed training data set, (middle) varying the assumed measurement error keeping fixed both the data set and the cluster initialization, and (right) choosing randomly the training data set. In the left panel, the upper arrow indicates the next to last maximum of the OD curve and the lower arrow points to the absolute maximum of the ED curve. The curves reported here are obtained using a subset of the combination of the two spectral libraries and coarse grain size.

group montmorillonite versus nontronite [*Bishop et al.*, 2008]. Such a case strongly influences the OD curve, which presents a much higher variance than the ED curve. The latter, on the contrary, does not appear associated with a high variance and therefore we expect the ED to provide a consistent number of clusters over different training sessions. An appropriate measure of the robustness of the criteria is to perform many clustering sessions by initializing randomly the clusters but keeping fixed the data set (Figure 1, left). In this case both the definitions appear to be very stable showing negligible uncertainties associated with the curves. We also investigate the influence of the measurement error when the ED is used. From the middle panel of Figure 1 is evident that the error adopted can strongly affect the number of clusters by shifting the absolute maximum to the right as the error is decreased. This suggests that a thorough investigation of the measurement uncertainties (random and systematic) should by done before applying the ED.

4. Classification and Evaluation

[15] Evaluating the scientific meaningfulness of a classifier designed to operate in an unsupervised and automated fashion is crucial to build confidence that the classification of an unknown data set will yield meaningful results. An appropriate test must consider the type of data that will be acquired and the type of information that is being sought. The spectral data we use in this study were measured in the VIS-IR where features due to vibrational and electronic transitions associated with minerals are prominent. The characteristics of these features are sensitive to the composition, crystalline structure, and grain size of the samples. We rely upon the following criteria to evaluate the meaningfulness of the results: number of clusters derived during training, purity of these clusters, confusion matrices illustrating how test spectra are correctly and incorrectly associated with training clusters, and finally the overall accuracy of the classification.

4.1. Number of Clusters

[16] The number of clusters provides a measure of the repeatability of the clustering results, and insight into the training process using the initial data. We evaluate both segregated, i.e., ASTER only, and combined, i.e., ASTER and RELAB, libraries for all grain sizes independently for ten realizations of the cluster analysis using the ED of the criterion. For comparison we also evaluate the number of clusters for the combined libraries using the OD of the criterion. Using different libraries allows evaluation of how the information content of the initial training impacts the number of clusters. Evaluation as a function of grain size within a single library allows us to evaluate the influence of having more samples within a particular grain size. If the variance of the number of clusters is small, then this implies that the clustering is robust and relatively insensitive to the random selection of the initial training data. Alternatively, if the variance is large, then the cluster analysis is highly sensitive to the perturbations of the input training data. Figure 2 shows the number of clusters determined for the raw and normalized reflectances for all the cases previously described.

[17] The number of clusters for the combined libraries increases with decreasing grain size, while it is almost constant when only one library is considered. This is expected because there are more finer-grained than coarser-grained samples, especially in the RELAB library (Tables 1 and 2). Such a trend reflects what is illustrated in Table 3 where it is evident that, for combined libraries, the number of groups represented in the training data set increases with finer grain sizes. Even for the case of the ASTER library there is a correspondence between the trend reported in Table 3 and Figure 2. In both cases the number of clusters obtained is generally close to the actual number of groups represented in

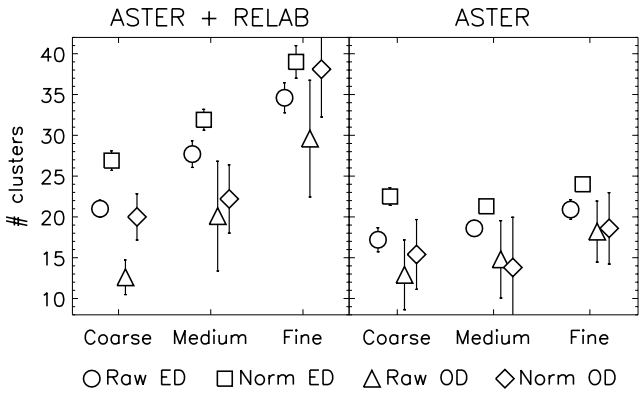

Figure 2. Averages of the obtained number of clusters for ten realizations of the clustering and their associated standard deviations.

the training data set. However, it can be noticed that the ED produces more clusters than OD and, within them, normalized data sets yield a larger number of clusters with respect to the raw ones, even if we do not see differences statistically relevant. This suggests the clustering results are more sensitive to the spectral features and not to reflectance levels. The standard deviation of the number of clusters retrieved is generally low, $\sim$1, when the ED of the criterion is used, and much higher, $\sim$5, when the OD is considered. The latter is expected because of the large variability of the OD when a random data set is used, as described in section 3. The small associated variances suggest that the clustering approach using the ED is relatively robust.

4.2. Cluster Purity and Label Assignment

[18] Once the number of clusters is defined, and the clustering performed, each cluster is associated with a label via three cluster identification matrices (CID), one for each label type (class, subclass, and group), which define the compositional meaning of each cluster. The uniqueness of the association between clusters and labels is measured using the purity index (PI). If matrix rows represent the obtained clusters and columns the labels, the purity (PI_c) for the cth row, i.e., cluster, is defined as follows:

$$PI_c = \begin{cases} 1 - \dfrac{1}{M-1} \sum_{i \neq i\max} \left(\dfrac{N_i}{N_{i\max}} \right) & M > 1 \\ 1 & M = 1 \end{cases} \quad (5)$$

where M is the number of labels represented in the cluster, N_i is the frequency for label i, $N_{i\max}$ is the maximum frequency, and $N_i/N_{i\max}$ is the relative frequency of label i with respect to label $i$max. This measure is designed to be unity when only one label is represented and zero when all labels are represented with equal frequency. In the case $M = 2$ this definition is more intuitive: it varies linearly with the relative frequency $N_{i\neq i\max}/N_{i\max}$. The overall purity for the clustering is the average of the cluster purities. An example of CID matrix is reported in Table 4 and the corresponding spectra are shown in Figure 3.

[19] The cluster purities are reported in Figure 4. We see that purities increase from group to class and ED generally provides a higher purity than OD. No statistically significant differences are observed between raw and normalized data set or among different grain sizes. However, in general, normalized data sets provide higher purities and lower uncertainties, in particular for coarse grain size (see Figure 4, top).

[20] At this stage, we assign the labels to the clusters via CID matrices by adopting a majority rule, i.e., the label with the highest occurrence is associated with the cluster. When PI_c is zero there is complete ambiguity and all the labels associated with the cluster occur with equal probability. In such a case the Euclidean distance between the cluster centroid and each sample included in it is computed, then the label of the nearest neighbor is selected. In case of an ambiguous cluster formed by only two spectra, the selection is random because they are equidistant from their centroid. In the example provided here (Table 4 and Figure 3), those cases where the assignment of a label to a cluster is ambiguous (CIDs 6, 8, and 21) a random selection is used to assign the label (e.g., CID 8, inosilicate is selected). Figure 3 illustrates a typical case of the clustering result. Each cluster represents unique spectral characteristics regardless to their subclass labels. This implies that different clusters can arise from the same subclass (e.g., CIDs 17 and 27 are both associated with the sulfate subclass label but they represent anhydrous sulfate and hydrous sulfate at the group level label, respectively) or different subclass can be condensed to a single cluster because of their spectral similarities (e.g., CID 8 where a phyllosilicate, mica, is clustered together with an inosilicate, amphibole).

[21] With the cluster labels defined, we have trained our classifier to potentially recognize unknown spectra. The CID matrix allows the prediction for the label of an unknown spectrum, presented to the trained algorithm. A sample can activate only a single cluster. This cluster is then mapped to a label via the CID matrix. This is the predicted label distinguishing it from the actual label assigned by the human expert based upon information independent of the spectral data.

4.3. Classification and Accuracy

[22] Here we test our classification scheme. For each realization of the classification, the spectra not used for the training phase are considered unknowns, and used for testing. We assign the unknowns to the previously defined clusters, then evaluate the accuracy of the assignment. Each unknown spectrum is either associated with a single cluster or it is rejected. The rejection means it does not belong to a predefined cluster and, potentially, it would constitute a new cluster if included in the training data set at a later stage.

[23] *Rousseeuw* [1987] suggests that the silhouette width (SW) is a confidence indicator for the membership of a sample to a given cluster. It is based upon the comparison between the average Euclidean distance between the sample to all the other samples contained in the given cluster (D) and the minimum average distance between the sample and all the samples belonging to other clusters (E):

$$SW = \frac{E - D}{\max(D, E)} \quad (6)$$

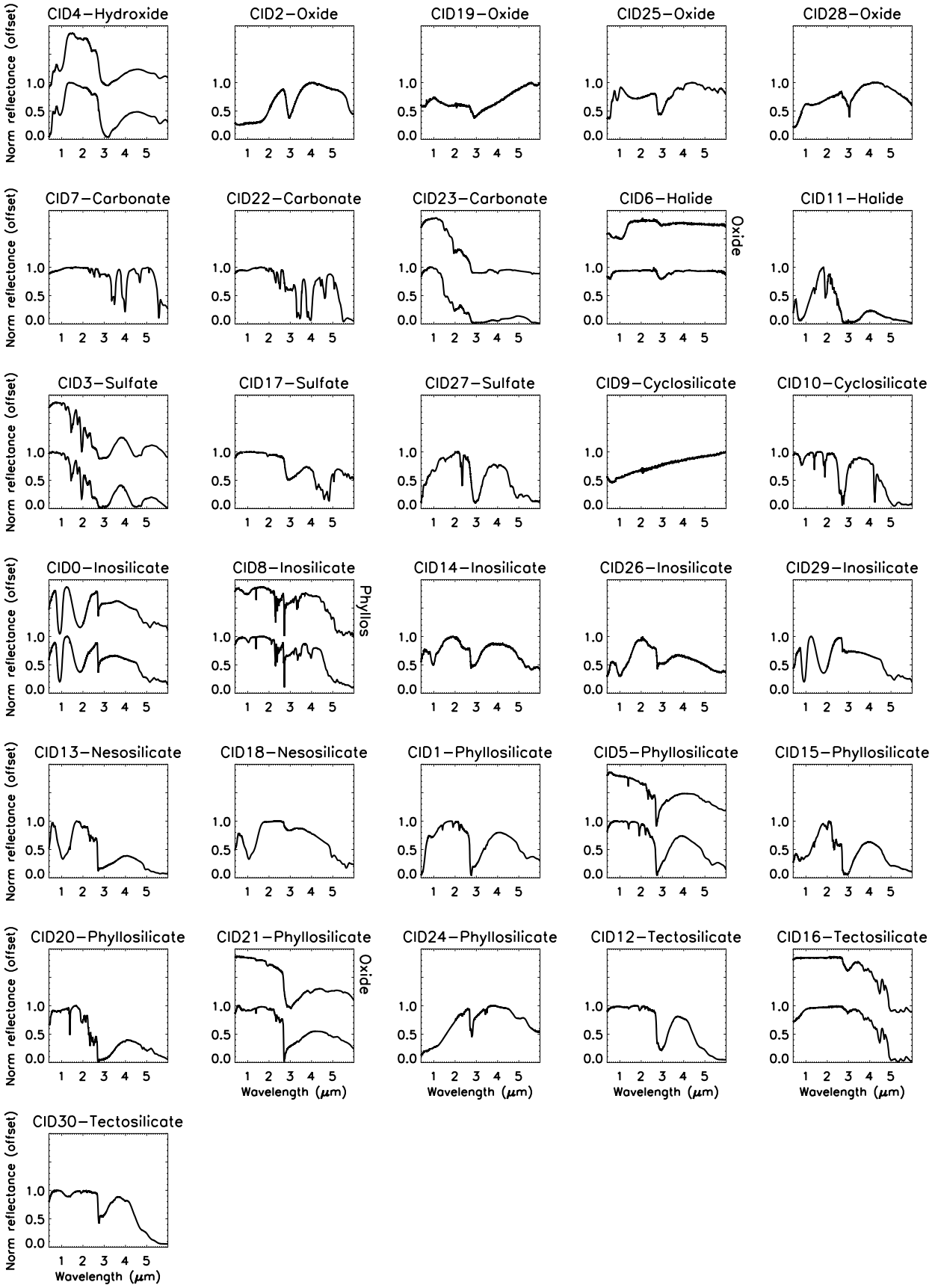

Figure 3. Spectra in the clusters corresponding to the CID matrix in Table 4. When two spectra are present, the top spectrum is offset at +1. In case of ambiguous clusters (CIDs 6, 8, and 21), the alternative label is reported on the right side, corresponding to the appropriate spectrum (see text).

Table 4. Example of CID Matrix Relative to Combined Libraries, Medium Grain Size, and Normalized Data[a]

Cluster	Hydro	Oxide	Carbo	Halide	Sulfate	Cyclo	Ino	Neso	Phyllo	Tecto	Label
4	2										hydroxide
2		1									oxide
19		1									oxide
25		1									oxide
28		1									oxide
7			1								carbonate
22			1								carbonate
23			2								carbonate
6		1		1							halide
11				1							halide
3					2						sulfate
17					1						sulfate
27					1						sulfate
9						1					cyclosilicate
10						1					cyclosilicate
0							2				inosilicate
8							1		1		inosilicate
14							1				inosilicate
26							1				inosilicate
29							1				inosilicate
13								1			nesosilicate
18								1			nesosilicate
1									1		phyllosilicate
5									2		phyllosilicate
15									1		phyllosilicate
20									1		phyllosilicate
21		1							1		phyllosilicate
24									1		phyllosilicate
12										1	tectosilicate
16										2	tectosilicate
30										1	tectosilicate

[a]The CID purity is 0.903. The labels assigned to each cluster are reported in the right column.

This measure is defined between -1 and 1, the latter providing evidence that the sample belongs to the given cluster and the former that they are unrelated. We decide to reject a vector when:

$$SW < \mathbf{W}_{err} \cdot \max(D, E) \tag{7}$$

where $\mathbf{W}_{err}$ has been defined in equation (2). From a geometrical point of view the previous definition implies that a sample is rejected when the difference of its distances from the two closest cluster centroids is equal to or less than its measurement uncertainty. This potential ambiguity in determining the membership of a sample to a cluster, which arises when the measurement uncertainty is considered, leads to the introduction of a criterion for its rejection. Later, we compare the quality of our classifications considering and ignoring this rejection criterion.

[24] A general method for comparing predictor and expert classification makes use of a confusion matrix (CM) where the expert labels of the training samples are used to label the rows and columns [*Kohavi and Provost*, 1998]. The matrix contains the cumulated tally of the true/predictor label pairs of a set of test samples not used to train the classifier. Test samples correctly classified populate the diagonal elements of the CM because predicted and true labels are identical. We adopt the normalized sum of the CM diagonal elements, i.e., $\overline{CM}/N$, as a standard measure of classification accuracy [*Kohavi and Provost*, 1998]. Accuracy can also be defined in a similar way for each column. Unit accuracy would indicate that all test vectors were classified correctly.

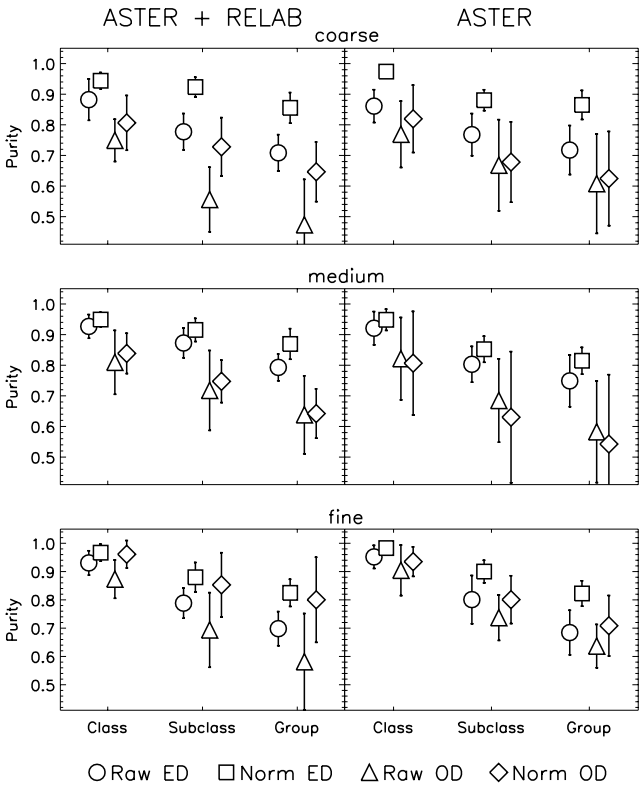

Figure 4. Summary of purities obtained for different combinations of libraries, grain sizes, and label levels.

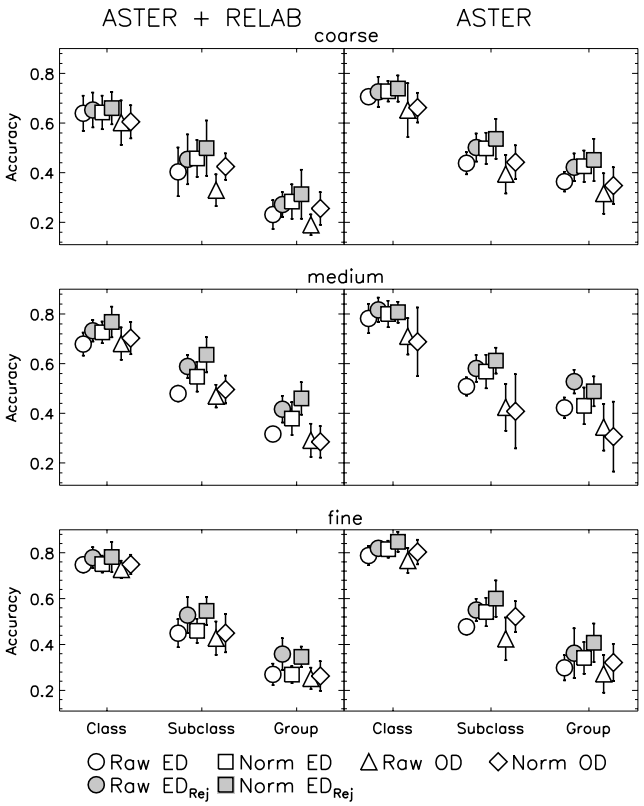

Figure 5. Summary of accuracies obtained for different combinations of libraries, grain sizes, and label levels. *Rej* means with rejection.

Example CMs are provided in Tables 5 and 6. They refer to the same test case without and with the rejection criterion, respectively. A summary of accuracies derived from all training/testing of all libraries are shown in Figure 5. It is clearly visible that results at class level are much more accurate than subclass and group level, while there are no significant differences between different grain sizes and data sets. The introduction of rejection improves the accuracy of the classification although not at a statistically significant level. Nonetheless, this improvement motivates us to incorporate the rejection criterion into the classifier scheme. The trends depicted in Figures 4 and 5 are consistent with similar results presented by *Roush and Hogan* [2007] using an entirely different clustering technique based on self-organizing maps.

5. Discussion and Conclusions

[25] We extended the clustering technique by *Marzo et al.* [2006] to build a classification scheme able to potentially recognize unknown spectra. We showed that both criteria used to find the natural number of clusters are robust, being insensitive to different initializations of the clustering process (Figure 1, left). Moreover, the extended definition of the criterion is able to indicate a natural number of clusters also for many randomly generated data sets (Figure 1, right).

[26] The number of clusters for the combined libraries increases with respect to the ASTER library alone, as expected, because in the combination more materials are represented, as summarized in Table 3. The number of clusters detected using both OD and ED is, on average, close to the number of groups actually represented by the training sets, and the trend observed across different grain sizes is consistent with that reported for the group row in Table 3. This provides confidence that the clustering approach is able to distinguish different minerals that comprise a group (e.g., for carbonates various cations such as Fe, Mg, Ca, etc.) at this level of detail. An example of this ability is reported in Figure 3 where carbonates (CIDs 7, 22, and 23) are separated in anhydrous (CIDs 7 and 22) and hydrous carbonates (CID 23). The former is further separated on the base of the spectral influence of the Mg cation, providing the ability to distinguish between a calcite ($CaCO_3$, CID 7) and a dolomite ($CaMg(CO_3)_2$, CID 22) even if this level of detail is not represented by our class, subclass, and group labels. Moreover, the number of clusters for the combined libraries increases with decreasing grain size, while it is almost constant when only one library is considered. This is also expected because there are more finer-grained than coarser-grained samples in RELAB library (see Tables 1 and 2). When the ED is adopted, the standard deviation of the retrieved number of clusters is small, further confirming that the classification approach is robust to the initial random selection of training data. We see no statistical differences in the resulting number of clusters between raw and normalized spectral data, suggesting that the clustering results are mainly sensitive to the spectral features and not to reflectance levels.

Table 5. Example of a Confusion Matrix Relative to Combined Libraries, Medium Grain Size, Normalized Data, and No Rejection[a]

Predicted / Actual	Hydro	Oxide	Carbo	Halide	Sulfate	Cyclo	Ino	Neso	Phyllo	Tecto
Hydroxide	1		2		4				1	
Oxide		2					2		1	
Carbonate			15		3				1	
Halide		1	2					1		
Sulfate					5	1		1		1
Cyclosilicate		2					1			
Inosilicate			2	1		1	4		3	1
Nesosilicate			1		1			7		
Phyllosilicate		1	13		1		3	1	45	1
Tectosilicate			1				1			11
Accuracy	1.00	0.33	0.42	0.00	0.36	0.00	0.36	0.70	0.88	0.79

[a]The average accuracy is 0.616.

Table 6. Example of a Confusion Matrix Relative to Combined Libraries, Medium Grain Size, Normalized Data, and With Rejection[a]

Predicted / Actual	Hydro	Oxide	Carbo	Halide	Sulfate	Cyclo	Ino	Neso	Phyllo	Tecto
Hydroxide	1		1		3					
Oxide		2							1	
Carbonate			11		2				1	
Halide		1	2							
Sulfate					4					1
Cyclosilicate		1								
Inosilicate			2	1			3			1
Nesosilicate					1			6		
Phyllosilicate			8		1		1	1	38	
Tectosilicate			1							11
Accuracy	1.00	0.50	0.44	0.00	0.36	0.00	0.75	0.86	0.95	0.85
Rejected		2	11		3	2	7	3	11	1

[a]The average accuracy is 0.717.

[27] The number of natural clusters detected adopting the ED generally provides a higher number of clusters (Figure 2). The greater number of clusters corresponds to a higher cluster purity with respect to the OD cases (Figure 4), and even higher when normalized data are considered. If we assume that the association between library spectra and labels is correct, then a higher purity provides higher confidence in classification schemes. Generally, it appears that the purity index decreases as the labels become associated with more detailed chemical and structural information. This suggests that it is easier to distinguish between distinctly different types of materials, but difficult to distinguish the more subtle differences between similar materials. We also observe an increase of variance from class to group label. We speculate this is associated with the lack of sufficient detail connoted in particular by the group labels. For example, at the group level the carbonates are only separated into hydrous and anhydrous carbonates. Within each of these labels remains structural (e.g., calcite versus argonite) and chemical variability (e.g., Mg- versus Ca-carbonate) that a more detailed label might capture. However, this is difficult to confirm because a more detailed level of labeling would result in even fewer spectral data available for training and testing. This is clearly where additional analyses would benefit from more extensive libraries.

[28] Accuracies do not present significant variations due to grain size (Figure 5). There is a clear significant overall decrease in the accuracies as the labels become more detailed. This is similar to the results for the purities and again may reflect the inability of the labels at the more detailed level to connote the complete chemical and structural variability that the sample spectra reflect. This accuracy decrease is not surprising because each class has a larger number of representatives than the subclasses and groups have. For example, the class of silicates would incorporate all the individual labels of the various different structural types of silicates (e.g., neso versus ino, etc.) into a single label. This provides evidence that more extensive training data sets lead to more accurate results. While the overall accuracies associated with the more detailed labels (subclass and group) may be low, the accuracy of individual labels may be significantly higher in spite of the very limited training set used, as shown in Tables 5 and 6. In these examples nesosilicates, tectosilicates, and phyllosilicates, which have several representatives, have a relatively high accuracy (>85% using rejection) compared to other subclass labels.

[29] We observe that for the class level and fine grain size, accuracies computed with the rejection criterion are higher than those obtained without it. However, this increase is not statistically significant and it might suggest that the adopted rejection criterion requires a higher rejection threshold. The adopted rejection criterion is a conservative choice which follows the introduction of the measurement uncertainty. The marginal improvement in the accuracies obtained using the current rejection criterion suggests some residual cluster overlap. This will be address in a future study that will fine tune the rejection threshold.

[30] We conclude that the classifier described herein is robust and sensitive to the reflectance spectral features. The scientific meaningfulness of the clusters increases when going from a detailed to coarser level of labeling where a larger data set is used for training, even if individual mineral labels can be predicted with relatively high accuracies. We expect that the classifier would be improved with the addition of a more extensive library for training and a more complex label structure. In the meantime, the technique is appropriate for a wide range of applications including analysis of different terrestrial and planetary data sets and enabling pattern recognition capabilities on next generation planetary rovers. The application of such an automatic technique, as preliminary level of investigation, would be helpful in the systematic analysis of very large spectral data set. Future efforts will focus on applying this approach to MRO/CRISM spectroscopic data [*Murchie et al.*, 2007].

[31] **Acknowledgments.** This research was supported by an appointment to the NASA Postdoctoral Program at the Ames Research Center, administered by Oak Ridge Associated Universities through a contract with NASA. T.L.R. acknowledges research support from NASA's Planetary Geology and Geophysics Program.

References

Bibring, J.-P., et al. (2006), Global mineralogical and aqueous mars history derived from OMEGA/Mars express data, *Science*, *312*, 400–404, doi:10.1126/science.1122659.

Bishop, J. L., M. D. Lane, M. D. Dyar, and A. J. Brown (2008), Reflectance and emission spectroscopy study of four groups of phyllosilicates: Smectites, kaolinite-serpentines, chlorites, and micas, *Clay Miner.*, *43*, 35–54.

Calinski, T., and J. Harabasz (1974), A dendrite method for cluster analysis, *Commun. Stat.*, *3*, 1–27.

Castaño, R. L., R. C. Judd, R. C. Anderson, and T. Estlin (2003), Machine learning challenges in Mars rover traverse science, in *Proceedings of the*

20th International Conference on Machine Learning 1615, edited by T. Fawcett and N. Mishra, AAAI Press, Menlo Park, Calif.

Christensen, P. R., et al. (2001), Mars global surveyor thermal emission spectrometer experiment: Investigation description and surface science results, *J. Geophys. Res.*, *106*, 23,823–23,871.

Crisp, J. A., J. P. Grotzinger, A. R. Vasavada, J. S. Karcz, and MSL Science Team (2008), Mars science laboratory: Science overview, *LPI Contrib. 1401*, 24.

Everitt, B. S. (1980), *Cluster Analysis*, John Wiley, New York.

Gaffey, S. J., L. A. McFadden, D. Nash, and C. M. Pieters (1993), Ultraviolet, visible, and near-infrared reflectance spectroscopy: Laboratory spectra of geologic materials, in *Remote Geochimical Analysis: Elemental and Mineralogical Composition*, edited by C. M. Pieters and P. A. J. Englert, pp. 43–77, Cambridge Univ. Press, Cambridge, U. K.

Hartigan, J. A., and M. A. Wong (1979), A k-means clustering algorithm, *Appl. Stat.*, *28*, 100–108.

Kohavi, R., and F. Provost (1998), Glossary of terms, *Mach. Learn.*, *30*, 271–274.

MacQueen, J. (1967), Some methods for classification and analysis of multivariate observations, in *5th Berkeley Symp. Math. Statistic and Probability*, vol. 1, edited by L. M. Le Cam and J. Neyman, pp. 281–296, Univ. of Calif. Press, Berkeley, Calif.

Marzo, G. A., T. L. Roush, A. Blanco, S. Fonti, and V. Orofino (2006), Cluster analysis of planetary remote sensing spectral data, *J. Geophys. Res.*, *111*, E03002, doi:10.1029/2005JE002532.

Marzo, G. A., T. L. Roush, A. Blanco, S. Fonti, and V. Orofino (2008), Statistical exploration and volume reduction of planetary remote sensing spectral data, *J. Geophys. Res.*, *113*, E12009, doi:10.1029/2008JE003219.

Murchie, S., et al. (2007), Compact Reconnaissance Imaging Spectrometer for Mars (CRISM) on Mars Reconnaissance Orbiter (MRO), *J. Geophys. Res.*, *112*, E05S03, doi:10.1029/2006JE002682.

Pieters, C. M., and T. Hiroi (2004), RELAB (Reflectance Experiment Laboratory): A NASA multiuser spectroscopy facility, in *Lunar and Planetary Institute Conference Abstracts, Lunar and Planetary Institute Conference Abstracts 1720*, vol. 35, edited by S. Mackwell and E. Stansbery, Lunar and Planet. Inst., Houston, Tex.

Roush, T. L., and R. Hogan (2007), Automated classification of visible and near-infrared spectra using self-organizing maps, in *IEEE 2007 Aerospace Conference*, edited by D. Williamson, E. Bryan, and E. Cardoza, p. 1456, IEEE, Piscataway, N. J.

Roush, T. L., R. Shipman, P. Morris, P. Gazis, and L. Pedersen (2004), Essential autonomous science inference on rovers (EASIR), in *IEEE 2004 Aerospace Conference*, edited by C. Scott, IEEE, Piscataway, N. J.

Rousseeuw, P. J. (1987), Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *J. Comp. Appl. Math.*, *108*, 53–65.

Vago, J. L., B. Gardini, P. Baglioni, G. Kminek, G. Gianfiglio, and Exomars Project Team (2006), Science objectives of ESA's ExoMars mission, in *European Planetary Science Congress 2006*, p. 76, Copernicus, Germany.

R. C. Hogan, Bay Area Environmental Research Institute, NASA Ames Research Center, MS 245-3, Moffett Field, CA 94035-1000, USA.

G. A. Marzo and T. L. Roush, NASA Ames Research Center, MS 245-3, Moffett Field, CA 94035-1000, USA. (giuseppe.a.marzo@nasa.gov)