AUTOMATED ANALYSIS OF MARS MULTISPECTRAL OBSERVATIONS. K. Wagstaff, *Applied Physics Laboratory, Johns Hopkins University, Laurel MD 20723, USA, (kiri.wagstaff@jhuapl.edu)*, J. F. Bell III, *Department of Astronomy, Cornell University, Ithaca NY 14853, USA, (jfb8@cornell.edu)*.

Multispectral and hyperspectral imagers are now commonly used to obtain remote sensing measurements for the study of Mars, and many more such measurements are planned for the future. These techniques present a number of data collection, processing, and analysis challenges for planetary scientists. For example, CRISM, the spectrometer that will fly on Mars Reconnaissance Orbiter in 2005, is expected to return about 9 terabytes of data over the mission duration. Each multispectral map will be 5120x5120 pixels in size (25 MB). It is not obvious how to easily browse this data, much less perform detailed analyses of it. More data means more information and the opportunity for new insights about Mars, but it carries with it a heavier and heavier burden for the analysis process.

Conventional methods for analyzing multispectral data include techniques such as examining absorption band depths at specific wavelengths or plotting two-dimensional histograms of radiance at different wavelengths. Selecting wavelengths and band depths that will yield the most compositional or mineralogical insight requires a significant amount of expertise about the object being observed. More critically, the process can be very time-consuming, with each histogram providing a single two-dimensional slice of the data for interpretation. For spectral data with only 10 wavelengths, there are 900 such histograms. Instruments such as the HST Space Telescope Imaging Spectrograph (STIS) have observed Mars at 1024 wavelengths; there are 1,047,552 corresponding possible histograms. This is not an upper limit, however; there are several other histograms that can provide insights, such as plotting radiance at one wavelength against the radiance ratio at two other wavelengths, or plotting radiance against a band depth or slope feature.

Of course, not every pair of features will yield interesting results when plotted against each other. Often areas of interest are already known, such as the 900 nm band depth or the radiance at 440 or 750 nm for Mars observations. However, our knowledge of Mars is certainly not comprehensive. The ability to discover additional informative relationships is critical.

In this abstract, we describe the result of applying an automated clustering algorithm to two data sets composed of Mars observations. One data set was collected by STIS on the Hubble Space Telescope; the other was obtained by the Mars Pathfinder Lander. We find that the results are comprehensible and, when a manual analysis is available for comparison, there is a good amount of agreement between the two sets of results. The automated analysis, however, requires significantly less time to produce results.

Hubble Space Telescope STIS Data. We first introduce the STIS data set. These observations were collected by a spectrograph on HST at several different wavelengths (Bell III et al., 2001). The observing period was from April to May of 1999, corresponding to the middle of northern summer on Mars. Consequently, the quality of the observations of the northern part of the planet is high, while the southern latitudes have poorer coverage. Each map is composed of 360×180
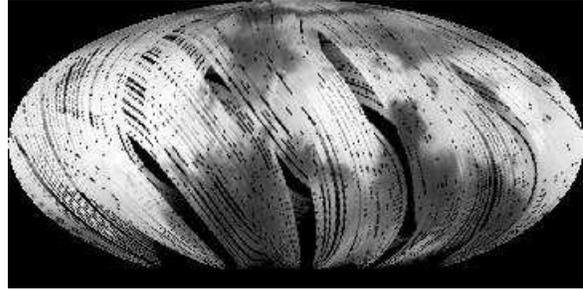

Figure 1: Global map of Mars, observed at 907 nm from Earth orbit by an imaging spectrograph on the Hubble Space Telescope. The image is in a Mollweide equal-area map projection.

pixels; the goal is to partition the pixels into coherent regional clusters on Mars. Figure 1 shows all of the pixels at a specific wavelength, 907 nm. Brighter areas correspond to higher intensity. The black striping and missing data regions are artifacts of the scanning method used by the spectrograph. Excluding missing data and pixels that fall outside the map, each image is composed of 47,221 pixels. Each pixel is described by 26 features, corresponding to the pixel's intensity at each of 26 different wavelengths, 528–1016 nm, selected from the original 1024.

We applied the k-means clustering algorithm to this data. K-means (MacQueen, 1967) takes as input a data set and the number of desired clusters, k. It produces a partition of the data set into k clusters such that items inside a cluster are very similar and distinct from items in other clusters. Each cluster is represented by its mean, or centroid. For multispectral analysis, the center of a cluster is a spectrum composed of the mean values for each of the wavelengths.

Figure 2 shows the result of partitioning the STIS data into seven clusters. Most of the clusters vary only in their albedo; their general spectral shapes are similar. Two noteworthy exceptions are the green and grey clusters. The green cluster is
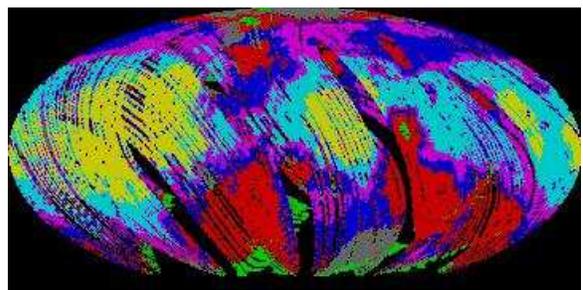

Figure 2: Results of k-means clustering with $k = 7$. Each color represents a spectral class that was determined by the algorithm to be spectrally unique. Many of the classes follow the classical albedo boundaries.

composed of very dark and flat spectra. It occurs in the southern regions of the data set, where observing conditions were poor (near the limb, and thus artificially darkened), near the north polar cap, and in an isolated spot that turns out to correspond to Syrtis Major. The southern regions can be considered an artifact of the data collection process. The north polar spots are likely to be dark soils and/or bedrock exposed by the retreat of the north polar cap, due to the time of year. Syrtis Major has been consistently identified as an unusually dark region on Mars; see e.g., McCord et al. (1971).

The other exceptional cluster is represented by grey in Figure 2. It occurs at the north pole and in a round region near the southern part of the planet. The average spectrum for this cluster is also unusually flat spectrum, but it differs from the green cluster in that it is much brighter (reflectance about 0.12 versus 0.08 for the green cluster). It does not exhibit the typical Martian spectral qualities of low blue and high red reflectance. We hypothesize that this cluster represents ice and/or clouds observed on the planet, and the physical location of the grey pixels bears this out. The north pole is a likely place to find ice, and the southern region corresponds directly to Hellas Crater, where thin clouds may form.

In both of these cases, the areas of interest highlighted by the analysis do not necessarily provide new knowledge about Mars. However, even at the coarse resolution provided by this data set, the results we obtain are consistent with previous knowledge about the planet. This provides support for using such methods on other Mars data sets, where less existing knowledge may be available.

MPF Lander Data. The Imager for the Mars Pathfinder Lander obtained a series of 12-color multispectral images during the lander's mission in 1997. A subset of these images, focusing primarily on soil deposits, has been analyzed by Bell et al. (2000) using manual and histogram-based classification methods. Figure 3 shows the 58 multispectral spot images at 600 nm. Bell et al. identified eight distinct soil and dust units: "Bright I-IV", "Surface dust", "Atmospheric dust", "Dark", and "Disturbed" (disturbed soil occurs in regions where the rover's wheels traveled). We also applied the k-means clustering algorithm to this data set, specifying eight clusters, as suggested by Bell et al. The resulting classification of the spot data is shown in Figure 4, where each cluster is represented by a different color. For the following discussion, we will refer to the k-means clusters by their color, and the clusters obtained by Bell et al. by the above descriptions.

The average cluster spectrum for each class is shown in Figure 5. They are color-coded to match the colors used in Figure 4. There are several interesting similarities to the clusters manually identified by Bell et al. The red and purple clusters correspond well to the "Dark" and "Disturbed" soil units. For both clusters, R_{750} is less than 0.2, matching the condition specified by Bell et al. The two soil types are distinguished in Bell et al.'s analysis by their R_{440} and 900 nm band depth values. "Dark" soil units have $R_{440} > 0.035$, while "Disturbed" soil has $R_{440} < 0.035$. We do not observe these exact values, but the 900 nm band depth for the red cluster is twice as large (0.06) as that for the purple cluster.

The orange cluster obtained by k-means corresponds well to the unit identified as atmospheric dust by Bell et al. In both analyses, it occurs in sky areas of the spot images. In our results, it also occurs on the top of a rock (spot 16). It is characterized by a high blue reflectance and a low red/blue ratio (i.e., the spectrum is relatively flat compared to other areas). In our analysis, this cluster is also distinguished from the other clusters due to its negative band depth at 900 nm. Every other cluster has either a negligible or a positive band depth at this wavelength. We also found that our cluster covers 7% of the pixels, while the atmospheric dust identified by Bell et al. only covers 4%. We hypothesize that some "Bright III" units may be included in this cluster, as they are generally characterized by a negative 900 nm band depth.

The remaining clusters are all composed of relatively bright material. They probably correspond to a collection of the "Bright I", "Bright II", "Bright IV", and "Surface Dust" clusters. We have not yet matched them more closely with each type. The distinctions identified by Bell et al. between these types depend on the 800–1000 nm slope and the 900 nm absorption band depth. Calculating these values over the cluster representative spectra did not sufficiently distinguish them; for example, all of the calculated 800–1000 nm slopes were between -0.00011 and -0.00010. We believe that calculating these values on individual spectra and then looking at the range of values for each cluster may yield more insight.
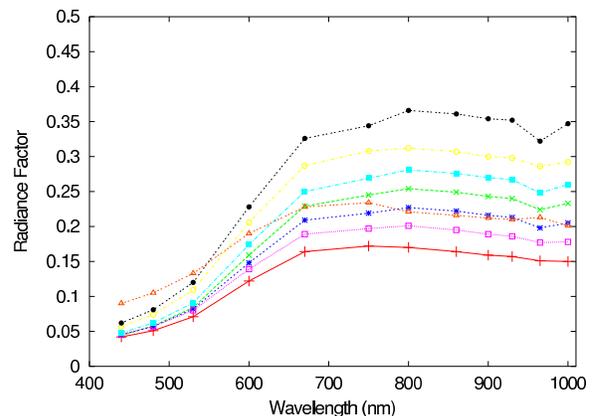

Figure 5: Average spectra for each of the eight clusters identified in Figure 4.

The analysis results we have obtained on this data set via an automated method are largely consistent to the results obtained by a manual analysis. This correspondence lends support to using automated methods to provide a "first pass" analysis of a data set.

Conclusions. Automated data analysis provide a means for handling the massive data volumes that planetary scientists increasingly face. By identifying distinct sub-groups in a data set, they highlight both overall trends and interesting exceptions. The speed of automated methods can save time devoted to preliminary investigations and make the analysis of terabyte-scale data sets feasible.

References. [1] Bell et al. (2000) JGR, 105, 1721–1755. [2] Bell et al. (2001) Bull. Am. Astro. Society, 33, 1127. [3] MacQueen (1967) Fifth Symp. on Math, Statistics, and Probability, 281–297. [4] McCord et al. (1971) Icarus, 14, 245–151.
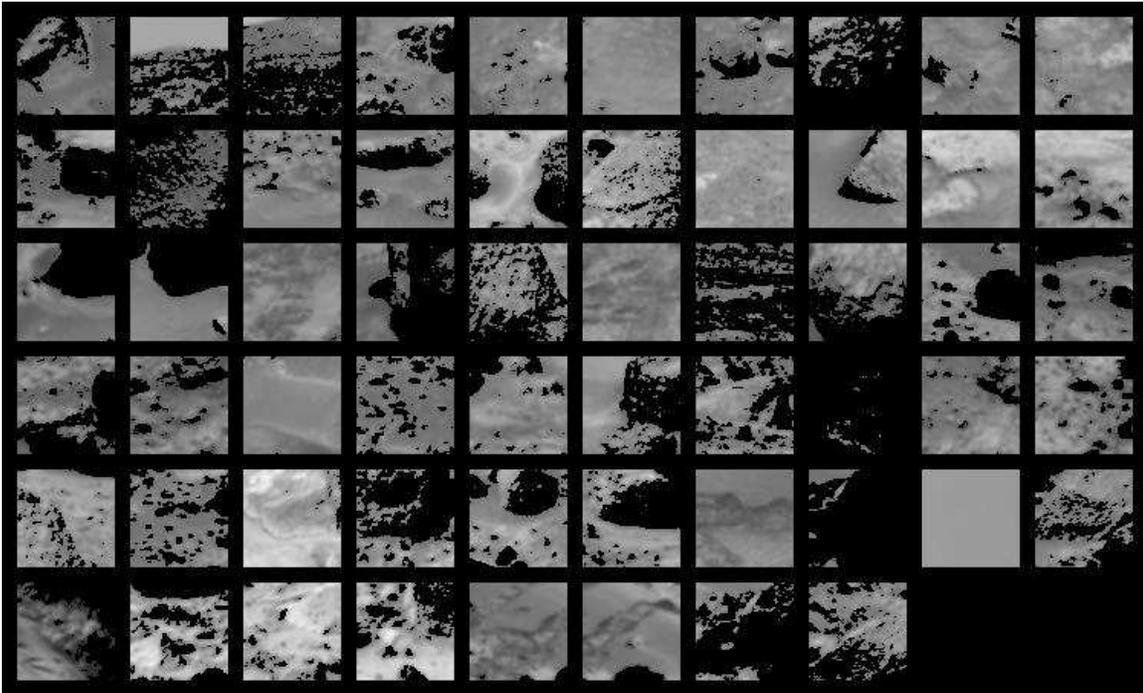
Figure 3: MSS soils and dust composite (combined left and right eye) mosaic of the 58 multispectral spots (each 64x64 pixels). The greyscale image shows intensity at 600 nm. Dark areas are regions where rocks, shade, and parts of the spacecraft have been masked out. Spots are numbered 1–58, starting at the top left and proceeding left to right.
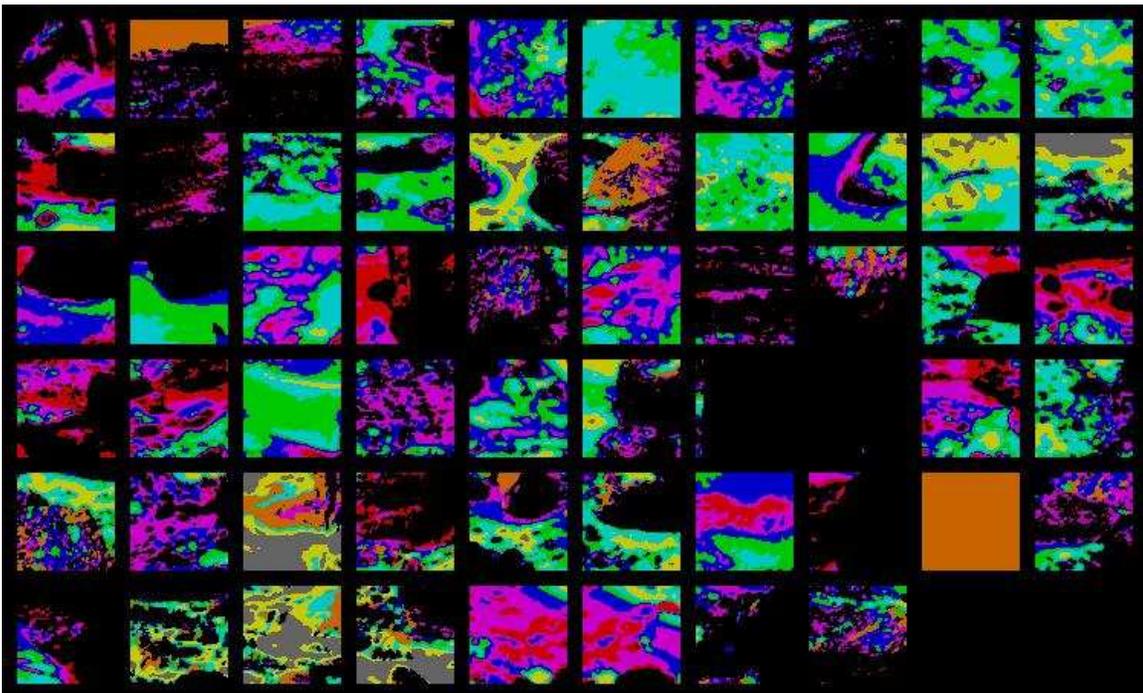

Figure 4: Automated classification of the MSS data into eight distinct spectral types.