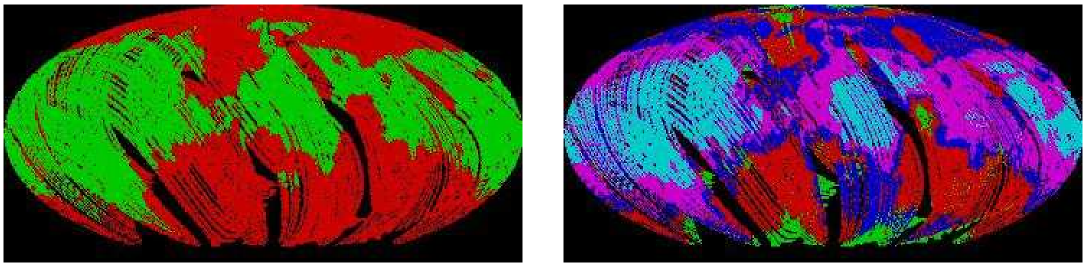


(a) $k = 2$ (b) $k = 5$

Figure 6.2: Two classifications of Mars by the k-means algorithm

Finally, we also know that the spectral slope between 800 and 1000 nm is important for distinguishing Martian spectra (see McSween Jr. et al., 1999; Bell et al., 2000). We calculate this value by subtracting the radiance at 801 nm from the radiance at 1016 nm. Technically, this slope (SL) should be divided by the wavelength separation (215 nm), but since this value is constant for all spectra, we omitted that additional step.

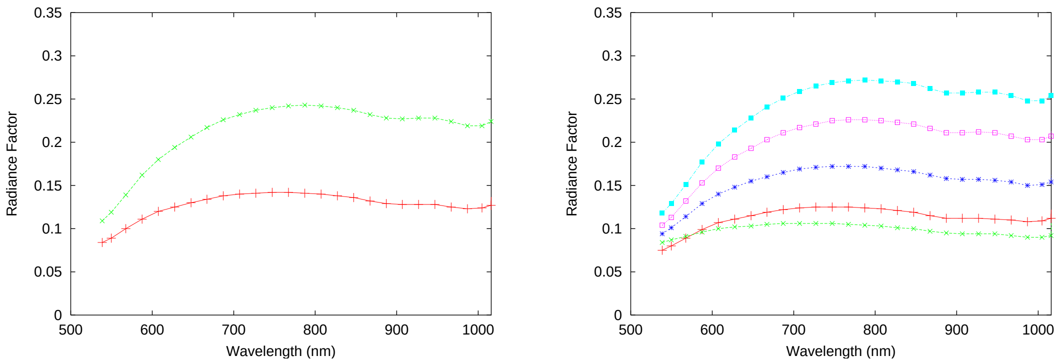
6.3.2 K-means Results

To provide a basis for later comparisons, we ran the regular k-means algorithm on the STIS data set. The output provided by this algorithm indicates what can be learned from an automated analysis of the data set without any specific domain knowledge. The input data includes the 26 spectral observations and five derived features for each pixel.

Before describing the results we obtained with k-means, we will first justify our choice of k-means for this problem. We do not claim that k-means is the only useful clustering algorithm for this problem. It may also be instructive to analyze the data using, for example, an algorithm which attempts to find a good value for k on its own, rather than having it specified as input. However, k-means is useful as an exploratory tool for spectral analysis precisely because we can control the number of clusters it generates. This allows us to examine Mars at different levels of generality. As we will show, the output at different k values provides us with different insights about the planet.

Figure 6.2 shows the output of the k-means algorithm for $k = 2$ and $k = 5$. In these images, each pixel is colored according to its classification by the clustering algorithm (the color values themselves are not meaningful). Black regions correspond to non-planet regions or to missing data. For $k = 2$, the planet is divided into two regions that correspond directly to its characteristic light and dark areas (cf. Figure 6.1). With larger values of k, such as $k = 5$, the planet splits into finer distinctions.

An examination of the average cluster spectra reveals that the cluster separations are dominated by simple albedo differences. Figure 6.3 plots, for each cluster identified by k-means, the average spectrum for that cluster. The x-axis is wavelength in nm, and the y-axis is the radiance factor, I/F. This value is defined as the observed radiance (I) divided by the radiance of sunlight reflected off a "perfect" diffuse reflector at the same distance as Mars (F). The colors used for each spectrum correspond to the colors used in Figure 6.2. Visually,



(a) $k = 2$

(b) $k = 5$

Figure 6.3: Average cluster spectra for the partitions shown in Figure 6.2

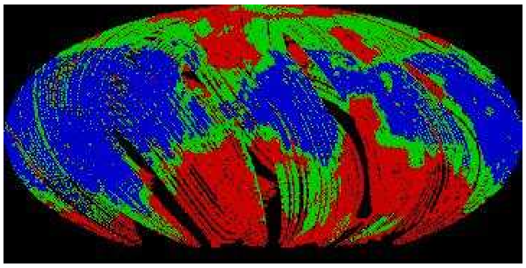
the average spectrum in each cluster is very similar to the other cluster spectra in shape, and brightness is the major distinguishing factor. This is not surprising; a principal components analysis (PCA) of this data set (and other Mars observations in general) confirms that the first principal component is albedo and that this component typically explains 90–95% of the variance in spectral observations of Mars (Bell, 1992). An exception to this general rule is the cluster identified by green x's in Figure 6.3b; its shape differs in important ways from the rest of the average cluster spectra. We will return to this interesting exception in Section 6.6.

6.4 Constrained Clustering With Soft Constraints

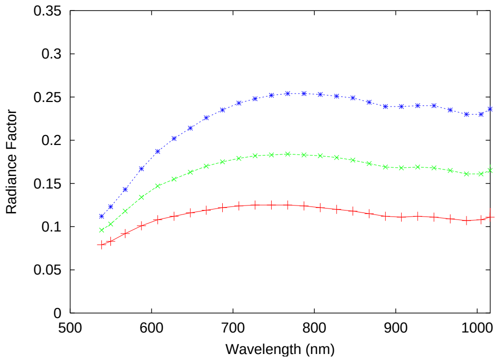
We have noted in previous chapters that some kinds of domain knowledge are perhaps better encoded as soft constraints (or *preferences*) rather than as hard constraints. This is particularly true for our spectral analysis of the Mars data. We do not yet know precisely what can be found in this data set, but we have some general knowledge about the problem domain that should be of use to a clustering algorithm. Therefore, before proceeding to describe the constraints that we used for this problem, we first describe our formulation for soft constraints and present an algorithm that can use them, SCOP-KMEANS.

6.4.1 Soft Constraints

Previously, we defined two kinds of hard constraints, must-link and cannot-link (see Section 3.1.1). Each constraint was specified as a relationship between two data items. For soft constraints (preferences), we will augment each relationship with an additional *strength* factor, s, that indicates how reliable the constraint is. This single constraint formulation ($\langle d_i, d_j, s \rangle$) subsumes both soft and hard constraints. The absolute value of s ranges from 0 to 1, with higher values indicating a stronger constraint. We also attach a sign to the strength; negative values indicate a preference against being grouped together, while positive values indicate a preference towards being grouped together. A constraint such as $\langle d_i, d_j, 1 \rangle$ is



(a) Classification



(b) Average cluster spectra

Figure 6.5: K-means output for $k = 3$

Comparison to B1. Our observations do not include wavelengths as low as 410 nm, but we can examine the average values for FL for each cluster (this is the $\frac{R_{528}}{R_{762}}$ ratio). Running with $k = 3$ (and no constraints) produces clusters that correspond to one cluster covering the "dark" areas and two clusters collectively covering the "bright" areas (see Figure 6.5a). In the absence of constraints, we find that the two bright clusters have FL values of 0.523 and 0.446, which indicates that one has a ratio that is 17% higher than the other. The average cluster spectra are shown in Figure 6.5b. This agrees well with **B1**. It is important to note that these three clusters were identified based on their homogeneity; we did not identify "bright" and "dark" regions explicitly for the algorithm to analyze.

When performing the same analysis but incorporating a spatial contiguity constraint with strength 0.5, this result is unchanged; we still find two bright clusters with the same relationship. However, the clusters are slightly more contiguous (per-pixel contiguity rises from 0.802 to 0.814).

Comparison to B2. We observe less of a distinction between the bright and dark regions in terms of BD_{860} than was reported by Bell et al. Rather than the dark areas having a 3–5% deeper absorption band, we instead see a difference of only about 2%. However, the trend is the same.

Comparison to B3. Interestingly, our results directly contradict **B3**. Recall that the deepest band depth at 950 nm in our data set is only 6%. We observe a very minimal average 950-nm absorption band for bright regions and an absorption of 1–2% for dark regions.

There are two possible reasons for the discrepancy between our results and **B3**. First, we are using a very coarse division of the planet into two or three regions and labeling those as "bright" and "dark." Although the averages show very small absorption features, the constituent pixels in each cluster possess a variety of band depths. Pixels in the bright cluster have a BD_{950} that ranges from 4% to −7% (a negative band depth indicates a local reflectivity *maximum* rather than an absorption feature). Pixels in the dark cluster range from having a BD_{950} of 6% to −6%.

Second, we are probably not calculating the continuum ("expected") radiance value at 950 nm in the same way that Bell et al. did. This is important, because it directly affects the band depth calculation. This may explain why the maximum band depth we observed differs so greatly from their observations.

Regardless, we are encouraged by the degree of agreement we find between our independent analyses of the same planet. We find agreement even though the observations are separated by five years and some of the derived features have not been calculated in exactly the same way. In terms of observing conditions, Bell et al. noted that at the time they were collecting their data, Mars was experiencing very cloudy weather. In contrast, our data set was created during extraordinarily clear Martian conditions. This allows us to see more of the planetary mineralogy, due to the reduction in obscuring ice and clouds. As we will see in the next section, our automated analysis was also able to detect and isolate areas where ice and clouds are present.

6.6.3 Novel Scientific Findings

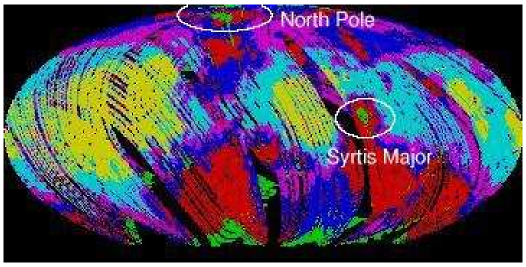
In addition to results that confirm or suggest alternatives to existing knowledge about Mars, our analysis also suggest several other interesting features of our particular data set. These observations add to the sum of our knowledge about the planet and are therefore valuable from a scientific viewpoint. In this section, we detail five interesting observations that came out of our automated analysis of the data set.

950-nm band depth. We found that this absorption band was a very distinct feature for regions on Mars. When clustering with $k = 2$, we find that the bright cluster has an average band depth of 0.3%. This is minimal, but it clearly contrasts with the average result for the dark clusters. Rather than finding an absorption feature at 950 nm, we find that the average value for BD_{950} is −2% (a local reflectivity maximum). This result suggests that the bright regions may be more likely to have a Fe^{3+} absorption band at this location.

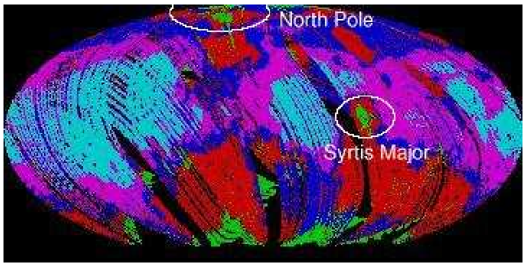
800 to 1000 nm slope. This feature also turns out to be important for distinguishing regions on Mars, as we hypothesized. The overall average slope for this data set is negative from 800 to 1000 nm. Our results indicate that bright regions, on average, have a slope that is twice as strongly negative as the dark regions.

Unusually dark low-red regions. There are two regions on Mars that show up as having abnormally *dark and flat* spectra, in comparison with the rest of the planet. Most of Mars is covered with spectra that have low FL values; the average for our data set is 0.530. However, these two regions have an unusually high average FL value of 0.820 (at $k = 6$). In fact, every pixel in the cluster has an above-average FL value (from 0.642 to 1.018). These spectra, because they are so relatively flat, can also be considered "low-red" regions; they will appear much less red than most of the planet does.

Besides being flat, these spectra are also very dark, with very low overall radiance values. The spatial location of these pixels is relevant; they occur in two places, Syrtis Major and



(a) With contiguity constraints, $k = 6$, $s_{contig} = 0.5$



(b) Without constraints, $k = 5$

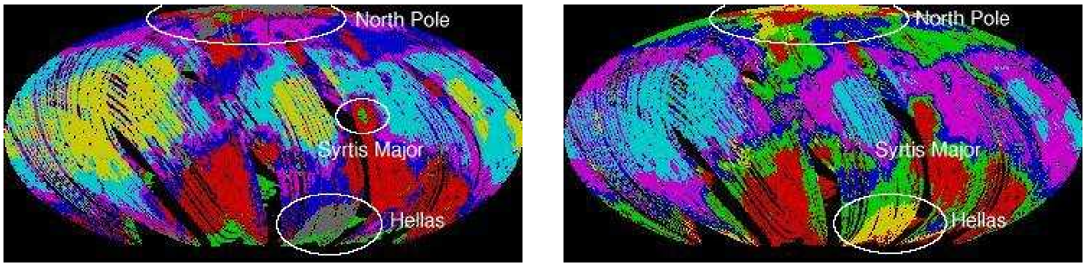
Figure 6.6: Dark low-red regions on Mars

two patches just southwest of the north polar cap. These are two places previously known to be very dark (McCord et al., 1971). Figure 6.6a shows the output of SCOP-KMEANS when using contiguity constraints with $k = 6$. The two regions of interest are labeled. (There are other pixels also assigned to this cluster, near the south edge of the planet. However, they are mostly artifacts of the data gathering process; they were observed under very poor conditions due to the viewing geometry between the Earth and Mars.)

These results were obtained when clustering with contiguity constraints. If we do not use it, a similar cluster appears at $k = 5$, but it is much less distinguished; its average FL value is only 0.799. This, too, is above average, but it is less spatially (and spectrally) coherent. The variance for the cluster obtained when using the regular k-means algorithm is 0.068, while the variance of the corresponding cluster when using COP-KMEANSis only 0.057. Lower variance means that the average cluster spectrum is a better fit to the items in the cluster, spectrally. The average spectrum for the unconstrained cluster is shown in Figure 6.3b. It is the flattest cluster in the graph, clearly distinguished from the other clusters.

Ice/cloud regions. There are also two regions on Mars that are distinguished by *bright and flat* spectra. They appear in the north polar region and over the Hellas Basin. If we increase k to 7, this cluster appears with a FL of 0.745, which is higher than average yet distinguished from the dark, flat regions above. In fact, the higher value of k also permits a tighter fit to the dark flat cluster; it now has an average FL of 0.825. Because these regions are significantly brighter than dark flat cluster, and because of their spatial locations, we hypothesize that they represent ice (water ice or CO_2 ice) present at the pole and thin clouds present over the Hellas Basin. Both of these hypotheses are physically plausible and in accord with what we believe to be possible on Mars.

Figure 6.7a shows the output of SCOP-KMEANS when using contiguity constraints with $k = 7$. The north polar region is almost entirely covered by the ice/cloud cluster, as is the Hellas basin. The dark flat cluster is still visible (cf. Figure 6.6a). Once again, these are the results obtained when clustering with contiguity constraints. Without that information, the results are again less satisfactory. At $k = 6$, a cluster similar to the ice/cloud cluster appears (Figure 6.7b), but its FL value is at 0.751. We also observe that the ice/cloud



(a) With contiguity constraints, $k = 7$, $s_{contig} = 0.5$

(b) Without constraints, $k = 6$

Figure 6.7: Ice/cloud regions on Mars

cluster obtained when using contiguity constraints is a better fit to the data; that cluster's average variance is only 0.149, compared to an average variance of 0.178 obtained when clustering without constraints. In addition, without contiguity constraints the dark flat cluster disappears; some of its pixels appear to have been subsumed by the ice/cloud cluster, which explains why that cluster's FL value is higher.

Slope constraint has little effect. We experimented with different values for s_{slope} and different ways of calculating these constraints, but found that the constraints have little effect on the final results. This appears to be because the clusters found in the absence of constraints are already fairly well separated with respect to this feature; darker areas tend to have lower slope values than brighter regions, as mentioned above. It is possible that calculating this feature as a relative value rather than an absolute one might yield values that correlate less with albedo. We have not tested this hypothesis yet. Although the slope constraint did not significantly alter the results in this case, it is still a useful piece of knowledge to include. For exploratory analyses in particular, blind analysis of the data can produce misleading results. By including additional information such as the slope constraint, we can *guarantee* that the output will be reasonable in terms of this particular domain knowledge. Without such constraints, we cannot be certain that the results will be physically meaningful.

6.7 Summary

In this chapter, we have achieved two major goals. First, we have developed a formulation for soft constraints and a clustering algorithm that can make effective use of those constraints. Second, we applied this algorithm to a very large database of Mars observations and found that the results are of value scientifically both for confirming the results of independent analyses of the planet and for highlighting points of disagreement. These results indicate that constrained clustering can be a useful tool for scientific discovery. We believe that scientists from a variety of disciplines can benefit from adding constrained clustering to their set of frequently used analysis tools.