AUTOMATED CLASSIFICATION OF STRATOSPHERIC DUST. S. W. Bell,[1] J. Lasue,[2] and T. Stepinski.[2] [1]Amherst College, Amherst, MA 01002 (swbell11@amherst.edu), [2]Lunar and Planetary Institute, Houston, TX 77058 (lasue@lpi.usra.edu, tstepinski@lpi.usra.edu).

Introduction: Over the past several decades, NASA has been able to collect thousands of Interplanetary Dust Particles (IDPs) from the stratosphere using high-flying aircraft. Unfortunately, even in the high atmosphere there are a large number of particles of terrestrial origin, either from natural sources such as volcanoes or from artificial sources such as rocket fuel. To distinguish the origins of the particles, NASA's Johnson Space Center (JSC) defined a preliminary classification system and published the results in the multi-volume "Cosmic Dust Catalog." [1] Each page of the catalog contains a scanning electron microscope (SEM) image of the particle, an X-ray energy spectrometry (EDS) spectrum, and an origin classification. There were seven origin classifications: C (cosmic), C? (probably C), TCN (natural terrestrial contaminant), TCN? (probably TCN), TCA (artificial terrestrial contaminant), TCA? (probably TCA), and AOS (aluminum oxide sphere) [1]. This classification system used several criteria to assess origin: The main indicators of cosmic origin used were similar compositions to meteorite bulk compositions (chrondritic or Fe-rich) and evidence of atmospheric entry heating (partial melting and the existence of a black fusion crust). The main indicator of terrestrial origin was the existence of significant amounts of cadmium, lead, potassium, sodium, or chlorine. Particles were also classified as terrestrial contaminants if they exhibited any number of unusual EDS bulk compositions, such as pure Si or pure Al [1].

High noble gas content from prolonged exposure to solar wind is a strong indicator of cosmic origin, and subsequent studies with sample-destructive noble gas content analyses in general confirmed that the particles JSC classified as C were indeed cosmic [2]. However, new results from the Stardust mission, which collected IDPs from the coma of comet Wild 2, indicate that a significant fraction of the particles previously classified as terrestrial contaminants may actually be of cosmic origin [3].

In previous work to improve the classification system, we applied data mining techniques to cluster the EDS spectra of the 464 particles in Volume 15 (the largest volume) of the IDP Catalog into groups of similar particles [4]. We found that these groups largely reproduced previously identified classes of IDPs.

In this current study, we examined the cluster particles in Volume 16. These are particles that split apart when they hit the collection plates and were actually collected as clusters of particles.

Methods: Unfortunately, the digital versions of the original EDS spectra were unavailable, so it was necessary to redigitize the spectra from PDF scans of the actual pages of the catalog. Due to the large number of spectra to be digitized, we built a semi-automated digitization algorithm.

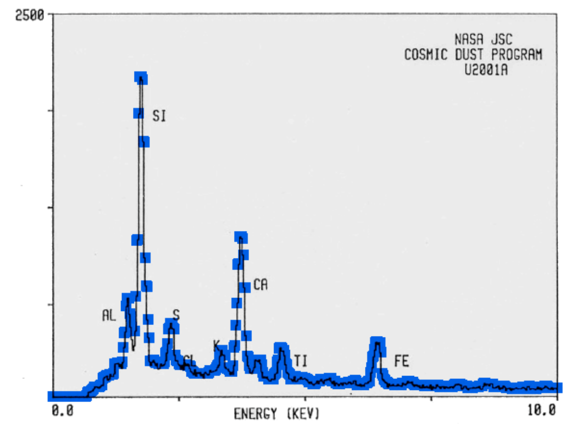

Figure 1: A sample EDS spectrum image with the .

Our first step was to separate the image into connected components and extracted the component representing the graph and frame. After using the frame to find the locations of the axes and determine the scale of the image, we removed the top and side axes and retrieved the maximum value in each column. These values were then scaled and recorded as the count values for the appropriate energy level. Several issues, such as letters "sticking" to the curve (see Fig. 1) or chunks of the curve missing, simply had to be dealt with manually. However, we succeeded in digitizing the spectra of 96 of the 117 the particles in Volume 16. (The rest of the spectra could not be digitized with our algorithm because they were displayed in a different format.) Because not all of the images had the same width along the energy axis, the digitized spectra were of differing lengths and needed to be rescaled, so we used linear interpolation to regenerate each spectrum with 1535 energy values. This allowed us to take each energy channel as a dimension, take the corresponding count value as the value in that dimension, and consider the spectra as vectors in a 1535-dimensional space.

Once we had all of the spectra plotted in the multidimensional space, we were able to determine similarity by simply calculating the Euclidean interpoint distances. Using this measure of similarity, we placed the spectra in groups using Ward's agglomerative clustering algorithm.

To help visualize these groupings, we used a visualization algorithm known as Sammon's Map [5] that represents points in a multidimensional space in two dimensions. The Sammon's Map, while of course not a perfect representation of the multidimensional space, does attempt to have the interpoint distances in two-dimensional space reflect the interpoint distances in multidimensional space.

To assess how pure the groups were with respect to the JSC classification labels, we used a standard data mining measure called "entropy." The entropy E_c of a group c is calculated by first finding n_{lc}, the number of particles of a particular label l in group c, and N_c, the total number of particles in the group, and then using Equation (1), where the summation is over all labels found in that particular group:

$$E_c = -\sum_l \frac{n_{lc}}{N_c} \log_2\left(\frac{n_{lc}}{N_c}\right). \tag{1}$$

An entropy of zero corresponds to a group with only one kind of label.

Results and Discussion: We determined that seven groupings best represented the data. With this number, the groupings were relatively pure with respect to the JSC classifications. (See Table 1.)

Cluster	All	Entropy	C	C?	TCN	TCN?	TCA	TCA?	AOS	?
1	47	0.342	44	3	0	0	0	0	0	0
2	9	0.000	9	0	0	0	0	0	0	0
3	9	1.352	5	3	1	0	0	0	0	0
4	9	1.880	1	0	1	0	5	1	0	1
5	8	0.811	6	0	0	2	0	0	0	0
6	9	0.986	7	1	0	0	0	0	0	1
7	5	0.722	0	0	0	0	4	0	0	1
Total	96	0.668	72	7	2	2	9	1	0	3

Table 1: The groupings with JSC classifications and entropies.

In the Sammon's Map (see Figure 2), the dense cluster (groups 1 and 2) in the center left corresponds to IDPs with compositions similar to chrondritic meteorites. There is a diffuse halo of outlier cosmic particles in surrounding them (group 3), and the particles at the bottom right of this diffuse halo correspond to IDPs with Fe-Ni-rich compositions. In the upper right, there is a group of contaminant "Low-Z" particles—particles with large amounts of elements with atomic numbers below 10—and in the far upper right there are two outlier high-Al contaminant particles. Overall, there is a trend from small amounts of low-Z material in the lower left to large amounts of low-Z material in the upper right and from low-Fe in the upper left to high-Fe in the lower right.

Most importantly, there is a clear separation between the cosmic and the contaminant particles. Only one contaminant particle was found in the cosmic region, and no cosmic particles were found in the contaminant region.

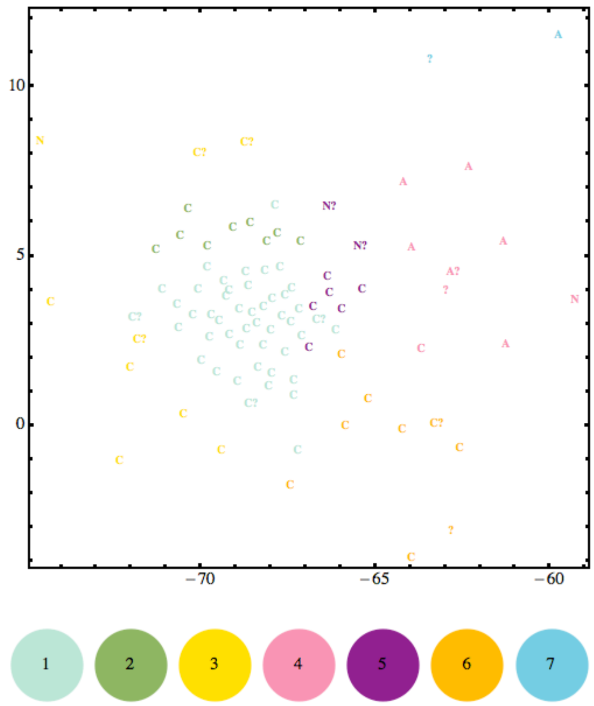

Figure 2: The Sammon's Map. Some classifications are represented in shortened form (TNC as "N", TNC? as "N?", TCA as "A", TCA? as "A?") and the unclassified particles are represented as "?".

Conclusion: We have demonstrated a technique that reproduces the separation between cosmic and contaminant particles without any human input or bias. This result helps validate both the JSC classification and our method, a technique that could be useful in classifying other extraterrestrial samples. These results will be compared to the previous analysis of the volume 15 of the 'Cosmic Dust Catalog'.

Acknowledgements: We would like to thank M. Zolensky for his assistance in understanding the catalog and fruitful discussion of our results. We would also like to thank the LPI internship program for making this project possible.

References: [1] Clanton, U. S. et al. *Cosmic Dust Catalog*. [2] Kehm k. et al. (2002) *Meteoritics & Planet. Sci.*, 37, 1323-1335. [3] Joswiak D. J. et al. (2008) *LPS XXXIX*, Abstract # 2177. [4] Lasue et al. (2009) *Meteoritics & Planet. Sci.* (submitted) [5] Sammon J. W. (1969) *IEEE Transactions on Computers*.