C.S. Guiang and R.Y. Levine, Cloud detection and characterization using topological data analysis, Proc. SPIE 8534, Remote Sensing of Clouds and the Atmosphere XVII; and Lidar Technologies, Techniques, and Measurements for Atmospheric Remote Sensing VIII, 85340E, 2012.

Copyright 2012 Society of Photo-Optical Instrumentation Engineers. One print or electronic copy may be made for personal use only. Systematic reproduction and distribution, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper are prohibited.

http://dx.doi.org/10.1117/12.978078

See next page.

# Cloud Detection and Characterization using Topological Data Analysis

Chona S. Guiang\* and Robert Y. Levine Spectral Sciences, Inc. 4 Fourth Ave, Burlington, MA USA 01803

#### ABSTRACT

The presence of cirrus clouds introduces complex heating and cooling effects on the atmosphere and can also interfere with remote sensing from satellite-based sensors or from high-altitude aircraft. Detection of cirrus clouds thus provides an opportunity for atmospheric correction to introduce accurate compensation to images of the earth's surface. Previous work on detection and characterization of cirrus clouds have been based on observing spectral signatures on a spectral channel with significant water absorption, or calculation of radiant intensity ratios over a water band to a reference spectral channel.

Our proposed approach is based on applying computational homology to characterize the topological properties of cirrus clouds. We utilize an application called JPLEX to study the persistent homology of multi-dimensional simplicial complexes built from available hyperspectral or multispectral data. The technique has been successfully applied to discriminate subtle features in high dimensional noisy data sets. Previous examples include anomaly detection in hyperspectral images. The analysis makes use of the entire multidimensional data set (not just one or a combination of spectral bands) which may offer advantages in discriminating among various cloud types in a scene, as well as determining other characteristics of cirrus clouds such as altitude and thickness. Our initial computational experiment with an AVIRIS scene has demonstrated that JPLEX is able to discriminate between cumulus and cirrus clouds.

Keywords: Cirrus, hyperspectral, multispectral, imaging, cumulus, topological data analysis

#### 1. INTRODUCTION

Hyper- and multi-spectral imaging and analysis of clouds from high altitude platforms has critical applications to weather prediction and climate modeling <sup>1</sup>. In this paper we consider automated detection and characterization of clouds based on discriminating spectral features. For example, cirrus clouds have light penetration to the ground, low uniform reflectivity, and less water absorption of light reflection to higher altitudes. Alternatively, cumulus clouds do not transmit light, are irregularly shaped, allow specular reflections to a high altitude sensor, and may have a higher water column density to the sensor. Reflected background features include strong water absorption, areas of uniform reflectivity, and identifiable reflecting materials (such as roads and vegetation). Reflection from the ground also allows for possible trace species absorption and particular specular properties such as the well-known restrahlen feature in the long-wave infrared (LWIR)<sup>2</sup>.

Spectrally-based cloud detection and assessment has an extensive literature. In a series of papers, Gao and coworkers proposed and demonstrated with AVIRIS images the technique of using a spectral channel near 1.37 µm to detect cirrus clouds in down-looking views from high altitudes <sup>3,4,5</sup>. This wavelength typically provides enough water vapor absorption to suppress both the surface reflectance and low-altitude clouds while at the same time adequately transmitting high-altitude cloud-scattered radiation. A drawback is that because the transmittance of the cloudscattered radiance is not 100% (it depends on the water vapor column density above the cloud and on the sensor viewing and solar angles), the 1.37 µm signal by itself provides insufficient information for deriving the cloud properties, including the optical depth, that are needed for modeling its radiance contributions at other wavelengths and hence for properly compensating the surface reflectance spectrum. In an attempt to obtain more discriminating information on cirrus clouds, we proposed processing spectral imagery data to allow the retrieval of a cirrus cloud radiance signal from three wavelength channels in the vicinity of a partially absorbing water vapor band, such as the 1.13  $\mu$ m band, over spatially structured terrain<sup>6</sup>. The method, which uses spatial filtering and linear regression to cancel the surface background, was applied to several rural and urban AVIRIS scenes. Because the 1.13 µm cloud signal is only weakly absorbed, it directly yields an approximate cloud reflectance. The ratio of the 1.13 µm and 1.37 um cloud signals indicates the water vapor column above the cloud, hence an effective cloud altitude, which in turn can be used to determine the cloud optical thickness. Alternative cloud measurements to characterize altitude, thickness, and particle distribution include LIDAR<sup>7</sup> and airborne sampling<sup>8</sup>.

In this paper we consider nonlinear approaches to processing hyper- and multispectral data sets, which also have the potential for extension to a large number of spectral bands. The challenge to analyzing hyperspectral/multispectral imagery lies in the high dimensionality of the data and the ubiquitous presence of noise, which can be introduced by the sensor, platform motion, shot noise or image misregistration. High dimensionality places a limit on visualization of the data and introduces additional complexity in finding patterns and/or performing cluster analysis. Clustering becomes more difficult because the increase in dimensionality introduces sparsity, which exaggerates the distance between all points to the point of decreasing the difference between far-off and adjacent points. Dimensionality reduction techniques are popular because they represent the data in a reduced dimensional space where the distance metric is more meaningful in differentiating (or finding similarities) among points. Spectral unmixing techniques such as Pixel Purity Index (PPI)<sup>9</sup>, Optical Real-Time Adaptive Spectral Identification System (ORASIS)<sup>10</sup> and Sequential Maximum Angle Convex Cone (SMACC)<sup>11</sup> endmember extraction attempt to resolve the hyperspectral data in terms of "pure" component spectra. In techniques based on decomposing each pixel in a scene to different components, using an accurate library of material spectra, unambiguous assignment of material(s) to a pixel may be difficult for spectrally similar materials or in the presence of noise. An alternative dimensional reduction technique is Stochastic Neighborhood Embedding (SNE) that maps high dimensional data to a lower dimension while preserving the local character of the data<sup>12,13</sup>. This is important for remotely sensed hyperspectral data through the atmosphere because, as discussed above, the "holes" in spectral data and unique features of specular reflections may indicate the type of cloud or background – and need to be retained in the lower dimensional representation.

Ultimately, however, even in a lower dimensional representation that retains local properties and provides robust separation of spectral data points, a decision on cloud type is necessary based on these properties. In this paper, we examine the use of computational topology for pixel classification. The starting point of such an analysis relies on discovering the underlying topological characteristics of the data for different length scales in the spectral space. Towards this end, we calculated the homology of multispectral data using PLEX<sup>14</sup>, a collection of MATLAB routines and Java classes developed at Stanford University as a research tool for building and studying simplicial complexes, generated from real or synthetic point-cloud data. The input to PLEX is a point cloud P taken from a multi-dimensional data set,  $\mathbf{P} \in \mathbb{R}^n$ , where *n* is the dimensionality of the data. From  $\mathbf{P}$ , PLEX calculates filtered simplicial complexes or streams over an interval<sup>14,15</sup>. Each simplicial complex  $\sigma(t)$  at filtration length scale *t* is characterized by a set of Betti numbers<sup>16</sup>; over a filtration interval, the stream is thus described by a set of Betti<sub>k</sub> intervals, k = 0, 1, ..., n-1. The Betti<sub>k</sub> intervals taken together quantitatively describe the homology of the data set, and are plotted as bar codes over the filtration interval  $[t_{min}, t_{max}]$ . To obtain the persistent homology, we count the number of Betti<sub>k</sub> intervals that remain beyond  $t_{max}$ . When calculating the persistent homology of a multi-dimensional data set, it is crucial that the filtration interval chosen represents meaningful length scales in the data. Our approach relies on running PLEX using different patches of a given multispectral scene as input, and then comparing the resulting bar codes to test for discrimination. Figure 1 demonstrates how unique properties of a data cloud would appear in JPLEX bar codes. At sufficiently small scales, the data exists as four clusters (Betti0=4) since a simplex limited in segment length cannot cross the cluster. At about the same scale in which the clusters are bridged, a triangle appears between the two clusters, indicating the topological feature of Betti0=1 and Betti1=1. This feature will disappear if the scale increases so that the gap between the clusters can be crossed.



Figure 1. Demonstration of distinctive local homologies in JPLEX bar codes.

Figure 2 demonstrates a combined SNE/JPLEX algorithm in which SNE dimensional reduction is employed to classify different sub-images from a hyperspectral data cube as belonging to cirrus, cumulus, and background – dominated areas. The separated regions are then used to derive JPLEX bar codes indicating the cloud type. As shown in the figure, for a class of SNE algorithms that parameterize the map between the higher and lower dimensional spaces, we have the option of running JPLEX directly on the mapped lower dimensional data sets. Because the mapping retain local topological properties, in principle the mapped regions will be as discerning of image classification as the original data.



Figure 2. SNE/JPLEX algorithm flowchart.

In this paper we consider the different components of the processing in Figure 2 applied to a hyperspectral AVIRIS data set in which cirrus, cumulus, and background subimages have been clearly identified. Section II contains a description of the AVIRIS data<sup>3</sup>, separation of subimages via scatterplots, and the application of SNE classification. Section III then shows the derivation of JPLEX bar codes for each category identified through the SNE processing (as well as via direct visualization). A conclusion follows in Section IV.

## 2. SNE PROCESSING OF HYPERSPECTRAL AVIRIS DATA

We derived our 6D and 3D point cloud input spectra from the Shelton Cube 1 data generated by AVIRIS (Airborne Visible and InfraRed Imaging Spectrometer), which acquires calibrated radiance images in 224 bands from 400 to 2500 nm. Figure 3 shows the true color composite image of Shelton Cube 1, along with the pixel regions where we obtained background, cirrus and cumulus point clouds.



Figure 3. True color composite images of the AVIRIS Shelton Cube 1 data, where both images have a resolution of 614 x 512 pixels. The figure on the right show the pixel areas from where background (red), cirrus (green) and cumulus (blue) pixels are sampled.

Because of the large number of bands available in hyperspectral data, it is often necessary to use a small subset of bands for further analysis. In this work, a method based on Stochastic Neighbor Embedding called t-SNE, that computes a lower dimensional embedding in multidimensional data that preserves the local structure of the data in the original space.

For our calculations, we made use of the parametric and nonparametric t-SNE software developed by Maaten and co-workers<sup>17</sup>. Both the parametric and nonparametric versions of t-SNE were run to calculate the lower-dimensional embedding on six-dimensional point clouds constructed by aggregating background, cirrus and cumulus pixels. In the parametric and non-parametric cases, we investigated different spectral regions by constructing the point clouds over two non-overlapping spectral ranges: 500-900 nm and 950-1350 nm, where the latter corresponds to a partially absorbing water band. Six band centers are picked randomly over either interval. Both the training and test data sets consist of 1050 points, corresponding to 350 background, cirrus and cumulus points randomly picked from the pixel areas shown in Figure 3. The lower-dimensional embedding varies from run to run, because optimization of the non-convex objective function leads to a different solution depending on the initialization of the optimization algorithm. As such, we conducted several runs and picked the embedding corresponding to the best separation of the data. Figures 4 and 5 depict, respectively, the t-SNE results for the 500-900 nm and 950-1350 nm data sets.



Figure 4. Parametric t-SNE (left) and nonparametric t-SNE scatter plots of background, cirrus and cumulus points from the 6D dataset, collected over the 500-900 nm spectral range, in the embedded lower dimensional space.



Figure 5. Nonparametric t-SNE scatter plots of background, cirrus and cumulus points from the 6D dataset, collected over the 950-1350 nm spectral range, in the embedded lower dimensional space.

There is very good separation of the background, cirrus and cumulus pixels in the 500-900 nm data set, but poorer quality clustering is obtained for the data set collected over the 950-1350 nm range, especially in the parametric t-SNE method. Although we expect variability among the different pixel types to be especially large in this spectral region due to differences in water column amounts, the t-SNE clustering consistently produced poor feature separation during our numerical experiments with different sets of six randomly selected bands. This is a counterintuitive result that will be investigated further in future work, alongside the determination of the optimal set of n bands over which to sample the n-dimensional data set.

To investigate the robustness of t-SNE against the specific wavelength centers chosen for the input, we conducted several t-SNE calculations where we chose different sets of six wavelengths for each run. Cluster separation remains good even when a different combination of six bands is chosen over the 500-900 nm spectral range. In each case, we were able to obtain good clustering of the point cloud data in the lower 2D or 3D space, based on quantitative measures such as classification error, in the case of nonparametric t-SNE, and the trustworthiness metric associated with parametric t-SNE.

While nonparametric t-SNE is applicable as an aid to visualizing high dimensional data sets, the parametric t-SNE algorithm enables the association of data with a particular category after training with a suitable data set that accurately represents the different classification. The lower dimensional embedding information is stored as weights and biases of a neural network, and will be used to generate well-separated point clouds which will serve as further input to JPLEX.

## 3. JPLEX PROCESSING OF HYPERSPECTRAL AVIRIS DATA

JPLEX provides a quantitative way of describing the shape of point cloud data in Euclidean space using the tools of computational homology. We ran JPLEX on point cloud data sets in the original spectral space, over the same two non-overlapping spectral ranges used in the t-SNE studies. Each input data set corresponded to one of the colored regions shown in Figure 3. Due to memory constraints, however, only 3D subsets of the original 6D data were input to JPLEX, where the three bands are chosen as alternating elements of the original six spectral bands.

JPLEX output takes the form of bar codes, which capture the shape of the filtration complexes that live and die over a suitable length scale. Each of the input radiance data is normalized to take on values from 0 to 1; the scale invariance of the topological analysis ensures that any choice of normalization does not affect the output.

To provide a clearer correspondence between the point cloud data input and the resulting JPLEX analysis, we present both the 3D scatter plots of the background, cirrus and cumulus pixels in Figure 6, and the corresponding bar codes in Figure 7. Both figures contain the results only for the data set collected over the 500-900 nm spectral range. JPLEX processing of the data set over the partially absorbing water band did not yield bar codes that allow for discrimination among background versus cirrus versus cumulus pixels. In Figure 6, there are clear qualitative differences in the cumulus scatter plots from those generated from the cirrus and background point clouds. In contrast, the difference between cirrus and background scatter plots is more subtle; also, there is no clear representative 3D scatter prototype for either cirrus or background. The intra-class variation in the scatter plots are of the same order as the interclass variation, which makes discrimination based on the 3D shape alone quite difficult. This is to be expected, however, because background pixels typically exhibit greater clutter than cloudy pixels. Similarly, the transparency of cirrus clouds over the 500-900 nm range allows the underlying ground pixels to give rise to the same clutter behavior observed for the background pixels.

The JPLEX bar codes shed some light on the structure of the 3D data. For example, two or more persistent clusters are observed for the cumulus data in the form of two to three long-lived Betti0 intervals. One such cluster can be attributed to specular reflections off the cumulus surface that leads to brighter radiance signatures, which can be seen clearly from the 3D scatter plots for the cumulus data. Conversely, a cluster corresponding to darker pixels can also be observed from the scatter plots and thus correspond to one of the longer lived Betti0 intervals. Although not obvious from the 3D scatter plot, the cumulus data is also characterized by loops that persist over larger filtration lengths (longer Betti1 intervals) than either cirrus or background data. As observed in the 3D scatter plots, the background and cirrus bar codes are similar in shape, although some differences are more apparent. There is in general a greater spread among the Betti1 intervals for the cumulus data versus background, and a greater number of smaller clusters (Betti0 intervals) are observed as well for the cirrus data. The cirrus bar codes, while displaying some similarity to those corresponding to background data, also seem to show characteristics of the cumulus bar

codes in the number of clusters and a greater spread among the Betti1 intervals. Because cirrus-containing pixels exhibit some of the reflective properties of cumulus while also allowing ground radiance to the sensor, the hybrid appearance of the bar codes is somewhat expected.

The similarity between background and cirrus bar codes raises the question of the "purity" of some of the background pixels used in the JPLEX processing. Contamination by unobserved cirrus cloud can create additional artificial similarity between cirrus and background that adversely affects discrimination. Such contamination issues underscores the need for an automated pre-sorting algorithm that facilitates collection of more homogeneous cirrus-only and background-only pixels that are then input to JPLEX to yield characteristic bar codes for each pixel class. Based on the t-SNE results, we are convinced that calculating the low dimensional embedding and performing the subsequent association should be considered prerequisites to JPLEX or any other implementation of topological analysis.

One unexpected result is the greater discrimination when extracting point cloud data from the near visible spectral range 500-900 nm, in contrast to the poor separability in lower dimensional embedded space and similar-looking bar codes obtained when the point cloud data is collected over the partially absorbing spectral band 950-1350 nm. This is especially surprising given that MODIS cloud determination is performed over a strong water absorption band, and background-subtraction approaches<sup>6</sup> utilize band ratios between shoulder and strongly absorbing bands for cirrus detection. Although this can be seen as a shortcoming, the utility of our approach is certainly applicable in scenes where very high humidity levels preclude the use of water bands.



Figure 6. 3D scatter plots of background (top row), cirrus (middle row) and cumulus points (bottom row) taken from the colored regions of the Shelton Cube 1 AVIRIS data as shown in Figure 3. The axes correspond to spectral bands over the 500-900 nm spectral range, and each point represents scaled radiance over [0,1].



Figure 7. JPLEX bar code output for each of the scatter plots in Figure 6. From top to bottom, the rows show the resulting Betti0 and Betti1 intervals for background, cirrus and cumulus point clouds, respectively.

Although the information extracted from topological data analysis is contained in the 3D scatter plots, applying TDA offers a few advantages: 1) TDA uncovers subtle differences that may not be directly visible to the human eye; 2) TDA provides a quantitative measure of the shape of point cloud data; 3) the use of the bar code representation allows for the use of statistics (e.g., use of distance measures, variance, averages) to compare TDA outputs.

### 4. CONCLUSIONS AND FUTURE WORK

By performing topological data analysis (TDA) using the JPLEX computational homology software, we were able to demonstrate a new approach for identifying cirrus and cumulus clouds in hyperspectral data. Our TDA-based method was applied to different 3D radiance data sets taken from the AVIRIS Shelton Cube 1 scene, where each data set consists solely of cirrus, cumulus or background pixels. The TDA output is a set of different bar codes, each set reflecting the distinctive shape of the underlying structure in 3D spectral space, which may be used for

subsequent pixel classification of other AVIRIS data. Preliminary TDA analysis of data collected over two nonoverlapping spectral ranges, at 500-900 nm and 950-1350 nm, shows that classification is possible when the spectral data is sampled from the first wavelength range. One potential pitfall with application of TDA as a classifier is the presence of contaminated pixels in the training set (e.g., background pixels with invisible cirrus, or pixels with very thin cirrus layers.) To forestall this problem, we require an independent method of ensuring that the JPLEX training data is labeled accurately, and that each class of pixels is free of impurities from other pixel classes. Towards this end, we have successfully applied the parametric and non-parametric t-SNE (Stochastic Neighbor Embedding) and produced well-separated clusters of points in a lower-dimensional 3D spectral space from the original 6D data. The training set for JPLEX can then be assembled from maximally separated subsets of the resulting t-SNE clusters.

In principle, TDA can be applied to data of any Euclidean dimensionality. Due to memory constraints, however, we were only able to run JPLEX on 3D data inputs. In future studies, we will consider other software implementations of TDA that will get around the memory issues and allow the analysis of higher dimensional data. In addition, we intend to explore other spectral ranges, specifically in the SWIR range, to assess the performance of TDA/JPLEX, and to find an optimum spectral subset over which to perform the classification and/or cloud detection. To lay the groundwork for automated pixel classification using TDA, we will consider alternate representations of persistent homology beyond bar codes; for example, Harer's persistence diagrams (PD)<sup>15</sup>. Recent work has yielded a theoretical framework for defining an abstract space for PDs, and the capability to perform statistics on these diagrams<sup>18</sup>. These developments make it possible to perform automatic pixel classification after training TDA with appropriately labeled data, and furthermore, to quantitatively evaluate the accuracy of such a classifier using ROC curves. Finally, we hope to construct a fully automated and accurate pixel classification algorithm that incorporates both t-SNE clustering and TDA classification using hyperspectral or multispectral data as input.

#### REFERENCES

- [1] <u>http://www.nytimes.com/2012/05/01/science/earth/clouds-effect-on-climate-change-is-last-bastion-for-dissenters.html?\_r=2&ref=science</u>
- [2] Pollock, J.B., Toon, O.B., and Khare, B.N., "Optical properties of some terrestrial rocks and glasses", Icarus, 19, 372-389 (1973).
- [3] http://aviris.jpl.nasa.gov/
- [4] http://modis-atmos.gsfc.nasa.gov/ docs/CMUSERSGUIDE.pdf
- [5] Gao, B.-C. and Kaufman, Y.J., "Correction of Thin Cirrus Effects in AVIRIS Images Using the Sensitive 1.375 micron Cirrus Detecting Channel," Proceedings of Fifth Annual JPL Earth Science Workshop, JPL, 1, 59-62, (1995).
- [6] Adler-Golden, S.M., Levine, R.Y., Berk, A., Bernstein, L.S., Anderson, G.P., and Pukall, "Detection of Cirrus Clouds at 1.13 microns in AVIRIS Scenes over Land, "Proc. SPIE, 3756, 368-373, (1999).
- [7] http://www-calipso.larc.nasa.gov/
- [8] <u>http://campaign.arm.gov/sparticus/</u>
- [9] Boardman, J.W., "Geometric mixture analysis of imaging spectrometry data," Proc. Int. Geoscience and Remote Sensing Symposium, 4, 2369-2371(1994).
- [10] Bowles, J., Antoniades, M., Baumback, M., Grossman, J., Haas, D., Palmadesso, P., and Stracka, J., "Real Time Analysis of Hyperspectral Data Sets Using NRL's ORASIS Algorithm," Proc. SPIE, 3118, 38-45, (1997).
- [11] Gruninger, J., Sundberg, R.L., Fox, M.J., Levine, R., Mundkowsky, W.F., Salisbury, M.S., and Ratcliff, A.H., "Automated Optimal Channel Selection for Spectral Imaging Sensors, *Proc SPIE*, 4381, [4381-07], (2001).
- [12] van der Maaten, L.J.P. and Hinton, G.E. "Visualizing High-Dimensional Data Using t-SNE," Journal of Machine Learning Research 9, 2579-2605 (2008).
- [13] van der Maaten, L.J.P.. "Learning a Parametric Embedding by Preserving Local Structure." In Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AI-STATS), JMLR W&CP 5, 384-391 (2009).
- [14] http://comptop.stanford.edu/u/programs/jplex/
- [15] Edelsbrunner, H. and Harer, J. "Computational Topology: An Introduction." American Mathematical Society (2010).
- [16] Munkres, J.R., Elements of Algebraic Topology (Perseus Books, New York, 1993), 24
- [17] http://homepage.tudelft.nl/19j49/Matlab\_Toolbox\_for\_Dimensionality\_Reduction.html
- [18] Mileyko, Y., Mukherjee, S. and Harer, J., "Probability measures on the space of persistence diagrams," Inverse Problems, 124007 (2011).