Project 2 - CREATIVE NAME

Johnny Antoun (0679537) Jose Pliego (2716768)

November 2021

Contents

1 Data Collection and Exploration

1.1 (a)

The impact of increasing amounts of atmospheric carbon dioxide on Earth's climate is an important issue today. Prevalence of carbon dioxide leads to higher surface air temperatures with the strongest dependencies being in the Arctic region. In the polar regions, cloud coverage is important in modulating the sensitivity of the Arctic to increasing surface air temperatures but existing algorithms to detect clouds do not perform well in these regions due to the similarity between visible and infrared electromagnetic radiation emited from clouds and snow/ice-covered surfaces.

This study describes NASA scientists and statisticians attempt to devise improved cloud detection algorithms that work well in polar regions. They rely on measurements from the Multiangle Imaging SpectroRadiometer (MISR) that differs from traditional multispectral sensors that take measurements in a single view. The MISR sensor comprises nine cameras at different angles (4 forward, 4 backward and one nadir) in four spectral bands (blue, red, green and near-infrared). The MISC cameras cover 360-km-width swath of Earth's surface that extend across daylight side of the Earth from Arctic down to Antarctica in approximately 45 minutes with a total 233 distinct, but overlapping, such swaths. MISR completes accumulation of data from all 233 paths in around 15 days with each path subdivided into 180 blocks (block number increasing from the North Pole to the South Pole).

It is clear from the MISR data collection process that the resulting dataset is massive, posing computational constraints. Standard classification frameworks are not readily applicable given the size of the data and thus the difficulty of obtaining expert labels for training. Clustering is not ideal either because data units (three consecutive blocks) could be entirely cloud-covered or cloud-free. Consequently, the challenge is to combine clustering and classification in a computationally efficient manner.

The data collected in this study consists of 6 data units from consecutive 10 MISR orbits of path 26 (rich in surface features). Out of the total of 60 data units, three are excluded because the surfaces were open water, with the total included corresponding to around 7.1 million 1.1-kn resolution pixels. Experts label 71.5% of valid pixels

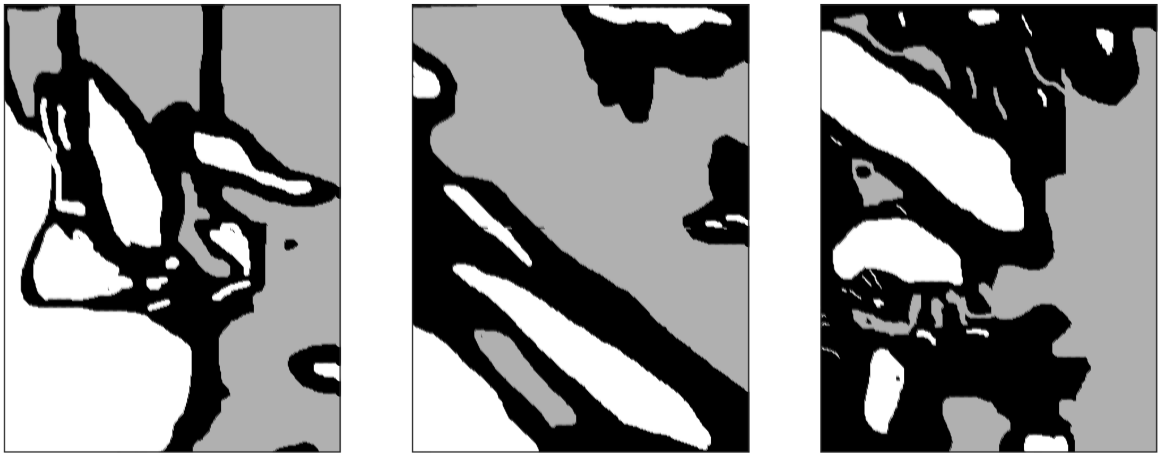
An existing cloud detection algorithm for MISR data exist, (L2TC), but it do not work well in the polar regions. the algorithm generally works well with the exeption of polar regions because low cloud heights lead to lower accuracy. This algorithm look for cloud pixels whereas the NASA scientists and statisticians found a better approach is to model the surface because it doesn't change materially from different different views. The proposed algorithm, enhanced linear correlation matching (ELCM), based on threholding three features with values that are either fixed or data-adaptive: correlation of MISR images of the same scene from a different angle (CORR), standard deviation of MISR nadir camera pixel values across a scene (SDan) and normalized difference angular index (NDAI) which relates to changes in a scene with changes in view direction. The CORR and SDan cutoff values are set for fixed values during operational processing whereas the NDAI threshold is either kept the same or updated at a new data unit based on a data-adaptive algorithm. Labels resulting from the ELCM algorithm are then used to train Fisher's quadratic discriminant analysis (QDA) to produce probability labels i.e. probability of cloudiness.

In conclusion, the study shows that the three physical features (CORR, SDan, NDAI) contain enough information to separate clouds from ice- and snow-covered surfaces. The ELCM algorithm combines classification and clustering in a way that makes it suitable for real-time, operational MISR data processing and is more accurate than existing MISR operations algorithms. Statisticians are involving in all the steps of data processing unlike other projects were they come in after the fact to develop methodologies.

1.2 (b)

Label	Img1	Img2	Img3	Total
No cloud	37.54%	43.98%	29.35%	36.96%
Unlabeled	28.68%	38.24%	52.17%	39.7%
Cloud	33.77%	17.78%	18.48%	23.34%

Table 1: Label distribution in the data.



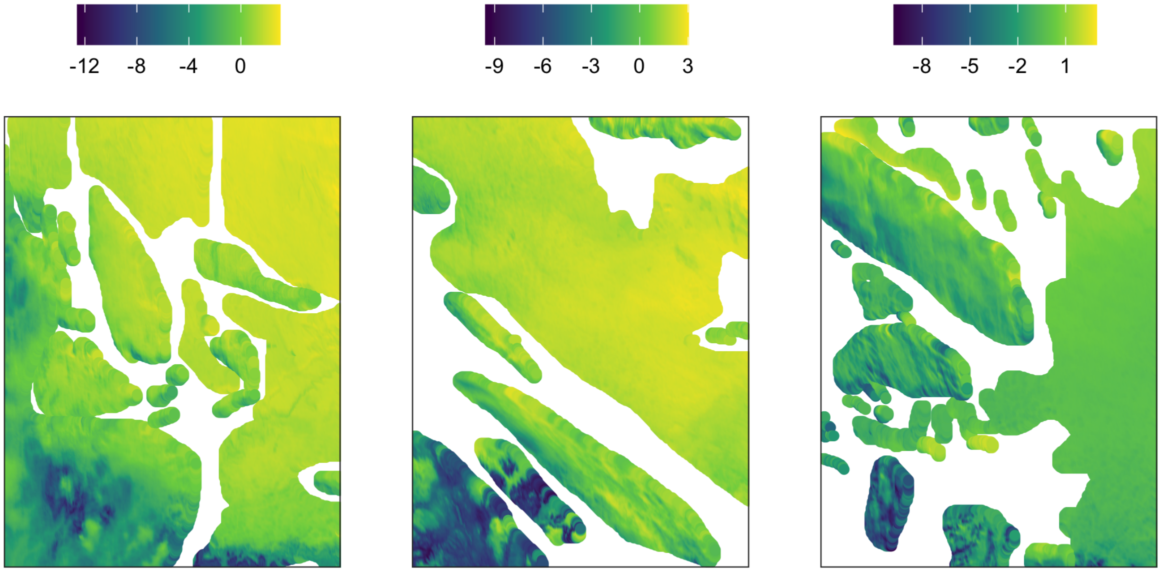
(a) Labels for image 1.

(b) Labels for image 2.

(c) Labels for image 3.

Figure 1: Image labels plotted in the (x, y) plane. White pixels correspond to clouds, gray pixels to land surface, and black pixels are unlabeled.

1.3 (c)



(a) PC1 representation for image 1.

(b) PC1 representation for image 2.

(c) PC1 representation for image 3.

Figure 2: First principal component scores plotted in the (x, y) plane. White regions correspond to missing (unlabeled) sections.