Project 2 - CREATIVE NAME

Johnny Antoun (0679537) Jose Pliego (2716768)

November 2021

Contents

1 Data Collection and Exploration

1.1 Paper summary

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 Data Summary

From this section and throughout our work, we "trimmed" the images so that they are perfectly rectangular. To do this, we filtered each image to have x-coordinate values between 70 and 368. By doing this cleaning step, we lost 2902 pixels in total (0.84%), with each of the three images loosing a similar proportion of pixels.

We decided to do this trimming because some features like Gabor filters require a rectangular matrix as an input (Chen (2006), Haghighat (2015), Mouselimis (2021)).

As a first step in the exploratory data analysis, let's take a look at the distribution of the expert labels in each image and in aggregate. The distribution is shown as a percentage frequency table (table 1). We can see that the distribution varies greatly from one image to another, with image 1 having a similar amount of cloud and non-cloud pixels, image 2 having a majority of non-cloud pixels, and image 3 having most of its pixels unlabeled.

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.

To get a sense of how labels look like in space, we can plot them in the (x, y) plane. In figure 1 we can see that cloudy regions and surface regions are separated by unlabeled regions, where the expert could not say for sure whether the pixels correspond to clouds or ice/snow.

(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.

Figure 1 is very interesting because it shows the strong spatial dependence present in the images. Cloudy pixels are generally surrounded by more cloudy pixels, and the same pattern repeats for surface (non-cloudy) pixels. This phenomenon presents a significant challenge for pixel classification because we cannot treat the different pixels as independent and identically distributed. We must account for the spatial dependence in our classification models so they can generalize well to future unobserved images.

1.3 Exploratory Data Analysis

To get a broad sense of how the predictors are related between them and with the labels, we include a correlation plot of the relevant variables in figure 2. The label column is encoded as 1 if the pixel is a cloud and -1 if it is not, so a positive correlation between label and a predictor can be roughly interpreted as an increase in the predictor being associated with an increase in "cloudiness." For this figure, we remove the pixels that are unlabeled to retain the numerical interpretation of the labels and the correlation with the predictors.

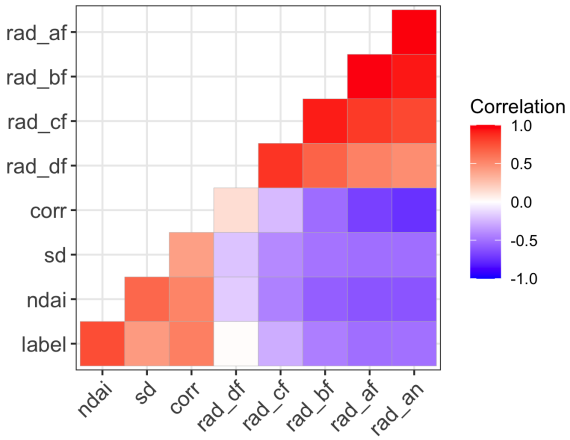


Figure 2: Visual representation of Pearson's correlation matrix.

Figure 2 shows a positive correlation between label and the author-defined variables `ndai`, `sd`, and `corr`, and also shows a negative correlation between label and most radiance measurements. It is also interesting to note that the author-defined variables are negatively correlated with the radiance measurements while positively correlated among themselves. Similarly, radiance measurements are positively correlated. This was expected since these variables are all measuring light reflection at different angles.

To take a closer look at some of these relationships, figure 3 shows the distribution of the author-defined variables and `rad_an` for the two pixel classes. The figure includes density estimates and the median for each class. All variables were centered and scaled. For the variable `sd`, we use a logarithmic transformation because the original variable is heavily right-skewed.

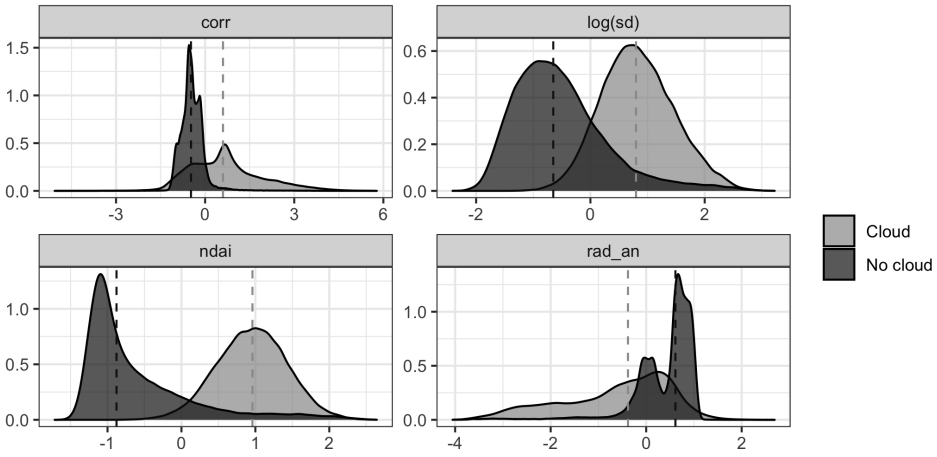
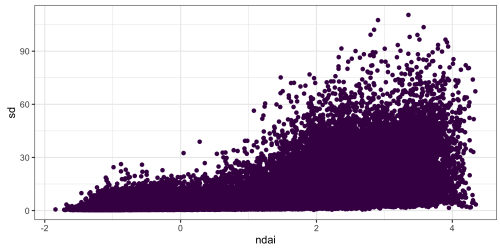


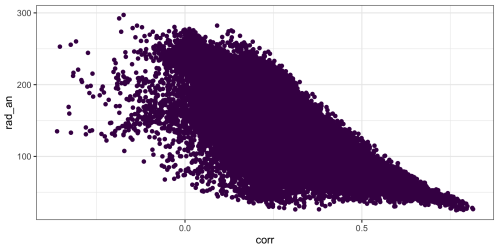
Figure 3: Distribution of author-defined variables and nadir radiance (rad_an) across pixel labels (median values highlighted).

In figure 3 we can see that the distributions are visibly different for both classes. The differences and the distance between the medians are greater in the plots of `log(sd)` and `ndai`. It is interesting to note that Shi et al. (2008) use fixed cutoff thresholds for `sd` and `corr` but not for `ndai` since the "appropriate thresholds vary from one data unit to another." In our case, it seems like `ndai` is the most discriminating feature. We formalize this notion of "best features" on section 2.

We include two scatterplots between the predictors that convey interesting properties in figure 4. The relationship between `ndai` and `sd` is shown in panel 4a. We see that the two variables are positively correlated (as was shown in figure 2) but also see that the spread of `sd` increases as `ndai` increases. Similarly, panel 4b shows the negative correlation between `corr` and `rad_an`, while also showing that the variance in `rad_an` increases as `corr` decreases. We chose to include these two scatterplots because other correlated variables show a more straightforward linear relationship without the heteroscedasticity property.



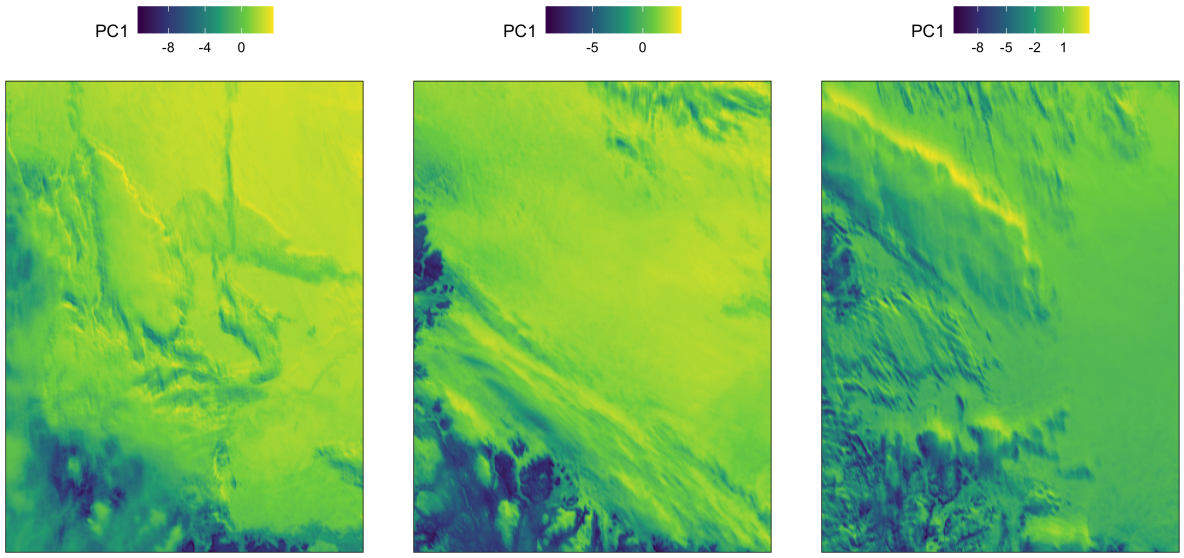
(a) Scatterplot between `ndai` and `sd`.



(b) Scatterplot between `corr` and `rad_an`.

Figure 4: Scatterplots for two pairs of highly correlated features.

To conclude the exploratory data analysis, we use Principal Component Analysis to reconstruct the three images using information from all the predictors. Since the principal components do not use label information, we obtain the scores and loadings using all the pixels. In figure 5 we can see the first principal component scores plotted in space. Note that the first principal component shows spatial dependence between neighboring pixels. This detail is interesting because the (x, y) coordinates were not used to calculate the principal components, so it supports the authors' claim that some features convey implicit spatial structure when defined using local patches of pixels.



(a) PC1 representation for image 1. (b) PC1 representation for image 2. (c) PC1 representation for image 3.

Figure 5: First principal component scores plotted in the (x, y) plane.

The first two principal components account for 80.7% of the variance in the dataset. Using the principal components as predictors instead of the original features may be useful in the modeling stage. As we saw in this analysis, some of the original predictors are highly correlated and using a set of orthogonal predictors (principal components) may yield better model results.

Finally, to support the claim that the principal components contain some of the spatial structure present in the data, we calculate the median and interquartile range (IQR) of the first two principal components by label. We choose IQR as a measure of spread and the median as representing the center because the distributions are not symmetric or unimodal.

Label	PC	Median	IQR
No cloud	PC1	1.69	1.94
No cloud	PC2	-0.50	0.88
Cloud	PC1	-0.86	3.06
Cloud	PC2	0.98	2.12

Table 2: Summary statistics for principal components.

In table 2 we see that the first two principal components have more spread in cloudy pixels, and the medians are different for both types of labels. This suggests that principal components may be useful in classification models. In section 3 we formalize these claims and assess model performance.

Bibliography

Chen, C. H. 2006. *Signal and Image Processing for Remote Sensing*. 1st ed. CRC Press.

Haghighat, Mohammad. 2015. *gabor: Gabor Feature Extraction*. https://github.com/mhaghighat/gabor.

Mouselimis, Lampros. 2021. *OpenImageR: An Image Processing Toolkit*. https://CRAN.R-project.org/package=OpenImageR.

Shi, Tao, Bin Yu, Eugene E Clothiaux, and Amy J Braverman. 2008. "Daytime Arctic Cloud Detection Based on Multi-Angle Satellite Data with Case Studies." *Journal of the American Statistical Association* 103 (482): 584–93. https://doi.org/10.1198/016214507000001283.