Project 2 Cloud Data

Xiaoya Li(3033286791), Shiyun Huang(3033203799)

1. Data Collection and Exploration

1a) The paper proposes two new operational Arctic cloud detection algorithms using MISR imagery that can efficiently process the massive MISR data set one data unit at a time without requiring human intervention to identify cloud-free surface pixels. The data used in this study were collected from 10 MISR orbits of path 26 over the Arctic, northern Greenland, and Baff Bay. The repeat time between two consecutive orbits over the same path was 16 days, so the 10 orbits span approximately 144 days from April 28 through September 19, 2002. Six data units from each orbit are included in this study. The data investigated contained 57 data units with 7,114,248 1.1-km resolution pixels with 36 radiation measurements for each pixel. For each pixel, there are 36 dimensions (4 wavelengths at 9 angles). In each data unit, there are 196,608(384 rows x 512 columns) pixels at the 1.1-km resolution. To solve the arctic cloud detection problem, first start with constructing three features, CORR, SD, and NDAI based on EDA and domain knowledge, then build ELCM algorithm by setting thresholds on each feature and demonstrates that probability predictions can be obtained over the partly cloudy scenes by using ELCM labels to train Fisher's QDA. This study concludes that ELCM and ELCM-QDA provide the best performance to date among all available operational algorithms using MISR data. Three features provide sufficient separability and stability to separate clear regions from clouds that a classifier no more sophisticated than QDA provided performance comparable to that of much more sophisticated classifiers. These studies will eventually enable the scientific community to study how changing cloud properties may enhance or ameliorate any initial changes in the Arctic brought about by increasing concentrations of atmospheric carbon dioxide.

1b) For image 1 (FIGURE 1), 44% of pixels have a label -1 (cloud-free), 18% have a label 1 (cloud) and the rest 38% remain unlabeled; for image 2 (FIGURE 1), 37% are cloud-free, 34% are cloud and 28% are unlabeled; for image 3 (FIGURE 1), 29% are cloud-free, 18% are cloud and 52% are unlabeled.

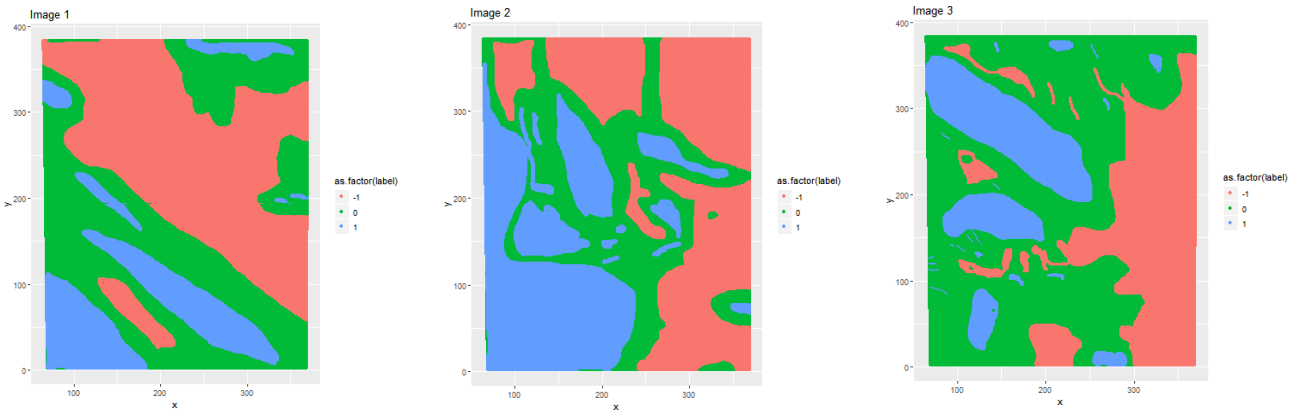

FIGURE 1

For those images, we can see that cloud appears in pieces rather than the random dots, which means if a pixel is clouded, its nearby pixels are more likely to be could. And the boundary between the could and could-free region are usually unlabeled. Therefore, i.i.d. assumptions for this dataset is invalid.

1c) (i) TABLE 1 shows the covariance between all the features. We can see that the radiances are highly correlated to each other except the relationship between AN and DF whose correlation is 0.489. It is more obvious from the left plot of FIGURE 2 that all the 5 radiances are positive correlated to each other. Also, the both NDAI, CORR and SD are somehow correlated to radiances negatively except DF, whose correlations are around 16%. Based on TABLE 1, NDAI, SD, and CORR are somehow positive correlated to each other; however, we can not see the obvious positive trend from the pairwise plot (right plot in FIGURE 2).

	NDAI	SD	CORR	DF	CF	BF	AF	AN
NDAI	1.0000000	0.6474474	0.5350207	-0.1639961	-0.4384729	-0.5710109	-0.6119935	-0.6085254
SD	0.6474474	1.0000000	0.4073057	-0.1965740	-0.4070290	-0.4912370	-0.5143340	-0.5068789
CORR	0.5350207	0.4073057	1.0000000	0.1477618	-0.2290945	-0.5182108	-0.6840183	-0.7460751
DF	-0.1639961	-0.1965740	0.1477618	1.0000000	0.8503037	0.6703445	0.5377937	0.4892642
CF	-0.4384729	-0.4070290	-0.2290945	0.8503037	1.0000000	0.9189584	0.8259473	0.7795202
BF	-0.5710109	-0.4912370	-0.5182108	0.6703445	0.9189584	1.0000000	0.9624793	0.9255600
AF	-0.6119935	-0.5143340	-0.6840183	0.5377937	0.8259473	0.9624793	1.0000000	0.9819174
AN	-0.6085254	-0.5068789	-0.7460751	0.4892642	0.7795202	0.9255600	0.9819174	1.0000000

TABLE 1

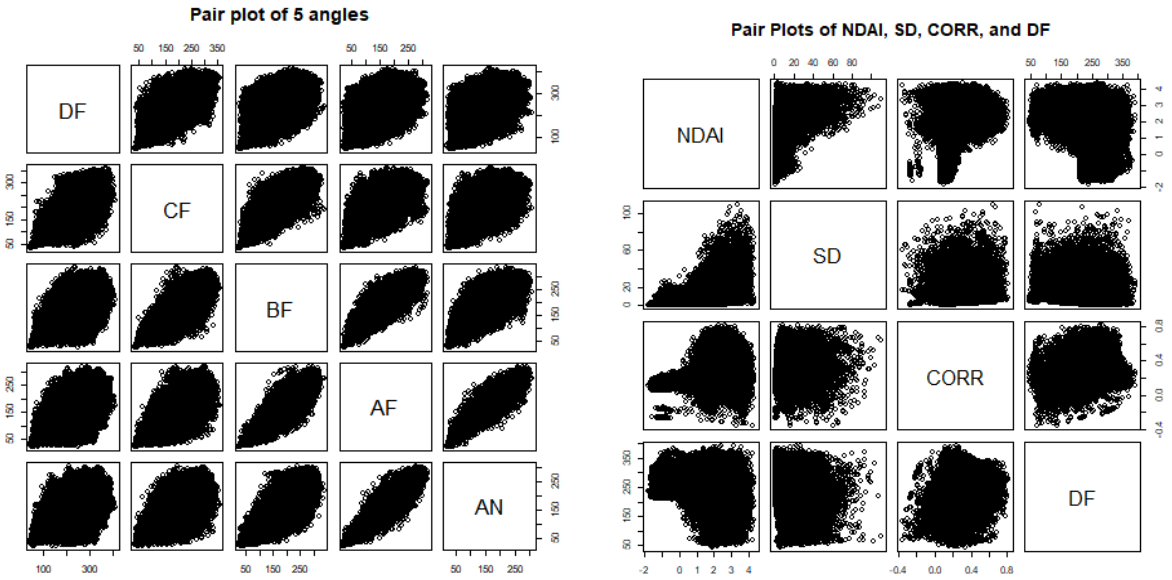

FIGURE 2

(ii) The relationship between the expert labels with the individual features is shown in FIGURE 3. The values of the 4 radiances (AF, AN, BF, CF) in the class 0 (no cloud) are larger than in the class 1 (cloud), and the values of CORR, SD, and NDAI in class 0 (no cloud) are less than in

the class 1 (cloud), which match the conclusion we made in (i) that CORR, SD, and NDAI are negatively correlated to radiances except DF.

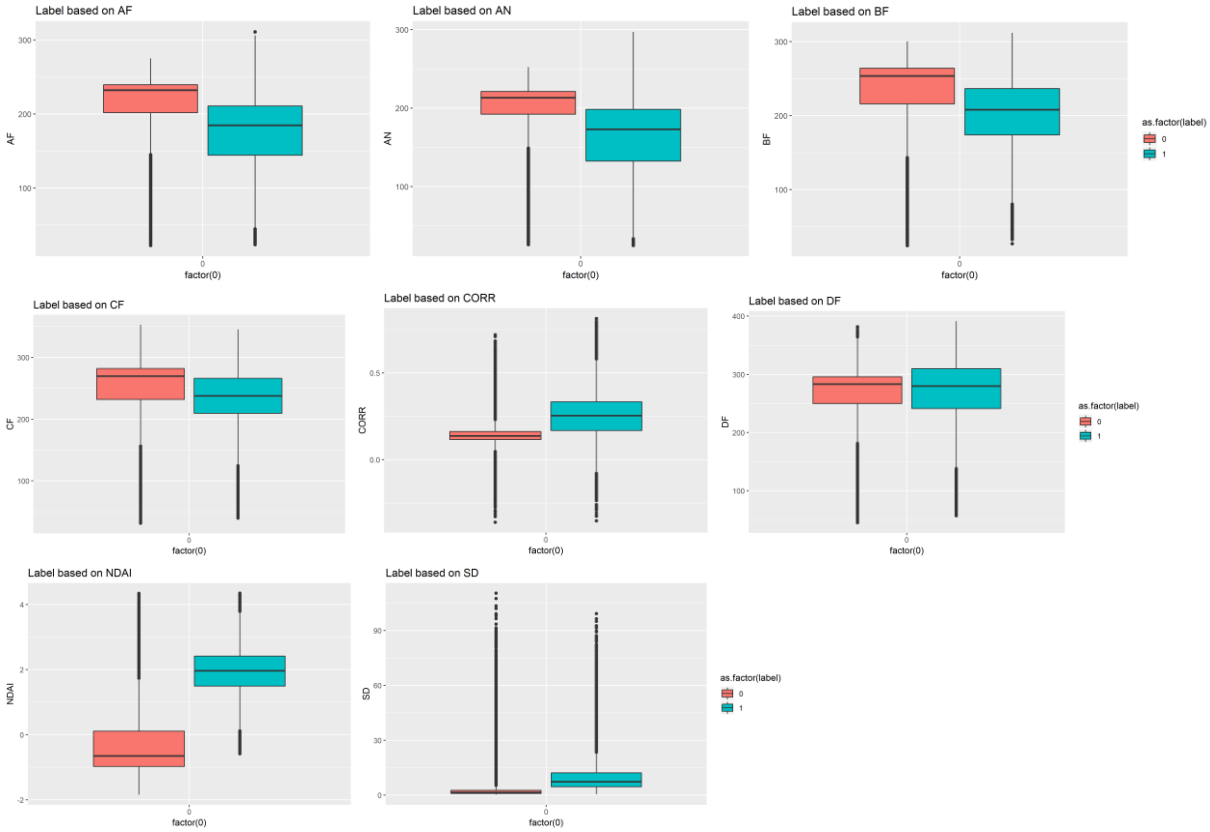

FIGURE 3

2. Preparation

2a) We combine the three images into one and split the data into three sets with two different ways. The first way is dividing the image into 4 by 4 grids based on x-coordinate and y-coordinate, and labeling it as 1 to 16. Then randomly drawing 10% of data be test set, which is randomly 2 whole grids, another 10%, which is another randomly 2 whole grids as the validation set, and the rest of them, which is 12 whole grids as the training set. The second way is that we can randomly choose 10% in each grid as the test set, another 10% as validation sets and rest of them as training set after separating the image into 4 by 4 grid based on the x coordinate and y-coordinate like the first method. These two methods are non-trivial because keeping the data as grids can make the data to be more independent of each other.

2b) For the data that splits using method 1, the trivial classifier scores 0.42 on the validation set and 0.14 on the test set. For the data that splits using method 2, the trivial classifier scores 0.61 on the validation set and 0.49 on the test set.
If the test set and validation set contain a great number of cloud-free observations, this trivial classifier will have high average accuracy. In general, this trivial classifier performs better on sunny days.

2c) We plot histograms of all features (y coordinate, x coordinate, NDAI, SD, CORR, Radiance angle DF, CF, BF, AF, AN), respectively by two groups (expert labels = -1 and expert labels = 1) and find the dip between two groups used as the threshold to identify two groups (shown in blue dashed line). We calculate the positive false and negative false for all histograms and realize that the best feature is NDAI, which has the smallest false positive 0.086441, and false negative 0.0151494. The second best feature is SD, which has false positive 0.03923369, and false negative 0.1201427. The third best feature is CORR, which has false positive 0.1284431, and false negative 0.03574913.

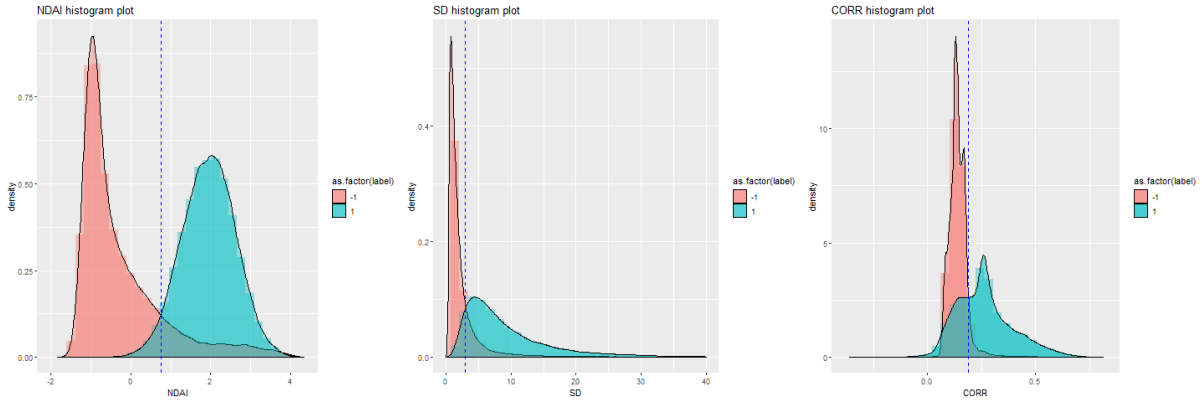

FIGURE 4

2d) The generic cross validation (CV) function is in github, and the file is named “utils.”

3. Modeling

3a) We try 4 classification methods including logistic regression,

i) Logistic regression

Logistic regression has the following assumptions.

First, the observations should be independent of each other. In fact, one place has cloud or not is correlated to its near places; however, the way we split the data can let the labels stay independent (the details of splitting data are explaining in question 2a). Second, the independent variables should not be too highly correlated with each other. In our case, we can see from the “Pair Plot of 5 Angles” that the radiances are highly correlated to each other. Also, SD, NDAI, and CORR, which are composed by some angles, are the best three features discussed in question 2c, so we are using SD, NDAI, and CORR in our model. Third, logistic regression requires a large sample size. The number of our training dataset is 191678, which is large enough. Hence, logistic regression is satisfied in this case. TABLE 2 shows the accuracies across five folds, the accuracies of average across folds and the test accuracy in two ways using binary logistic regression classification method.

	CV1	CV2	CV3	CV4	CV5	Mean	Test
Method1	0.9253259	0.7839734	0.8453375	0.9211426	0.8436098	0.8638778	0.9818584
Method2	0.8800638	0.8622484	0.9173530	0.8731481	0.8872394	0.8840105	0.8914245

TABLE 2

ii) LDA

LDA requires that the conditional probability distributions are normally distributed with the same covariance matrix, which means that the variables given labels should follow a normal distribution with the same covariance. In our case, three features(CORR, NDAI, and SD) approximately follows the normal distribution from the histogram plots showing above. Based on the histogram plots above, three features are not having the same variances. Hence, LDA is not really satisfied in this case. TABLE 3 shows the accuracies across five folds, the accuracies of average across folds and the test accuracy in two ways using LDA classification method.

	CV1	CV2	CV3	CV4	CV5	Mean	Test
Method1	0.9295066	0.7889326	0.8765761	0.9247349	0.8487718	0.8737044	0.9847515
Method2	0.9378743	0.8843581	0.9250360	0.8829635	0.8281272	0.8916718	0.8968532

TABLE 3

iii) QDA

The assumptions of QDA are similar to LDA except that QDA does not need to have identical covariance matrices for every class. Hence, QDA is satisfied in this case. TABLE 4 shows the accuracies across five folds, the accuracies of average across folds and the test accuracy in two ways using QDA classification method.

	CV1	CV2	CV3	CV4	CV5	Mean	Test
Method1	0.9293691	0.8145147	0.8587445	0.9195187	0.8509247	0.8746143	0.9880531
Method2	0.8887429	0.8845565	0.9160654	0.8933147	0.8767042	0.8918768	0.8978141

TABLE 4

iv) KNN

KNN is a non-parametric lazy learning algorithm, which means that it does not make any assumptions on the underlying data distribution. Hence, KNN is satisfied in this case. TABLE 5 shows the accuracies across five folds, the accuracies of average across folds and the test accuracy in two ways using KNN classification method with k = 11 neighbors. We choose the number k = 11 for split method 1 and k = 19 for split method 2 by cross validation, both k give the lowest mean cv-error.

	CV1	CV2	CV3	CV4	CV5	Mean	Test
Method1	0.9723761	0.8972691	0.8205107	0.4783853	0.6986774	0.7734437	0.0752272
Method2	0.8746156	0.8872312	0.8985975	0.6008396	0.9549714	0.8432511	0.02382897

TABLE 5

3b) ROC curves (left plot in FIGURE 5) shows the four different methods we use to fit the model. We can use ROC to evaluate the classifiers. The black curve is the ROC curve of logistic regression; the red curve is the ROC curve of the LDA; the green curve is the ROC curve of the QDA; the blue curve is the ROC curve of the KNN, and the grey dashed line is the trivial line representing random classifier. The x-axis of ROC represents the false positive rate and the Y-axis of ROC represents the true positive rate. Our goal is getting the smallest false positive rate and the largest true positive rate, which is the points closest to point (0,1). We can see that the curve nearest to point (0,1) is QDA. We choose the cut off is around 0.4866 where the false positive rate is 0.0111 because it is close to point (0,1) on QDA curve.

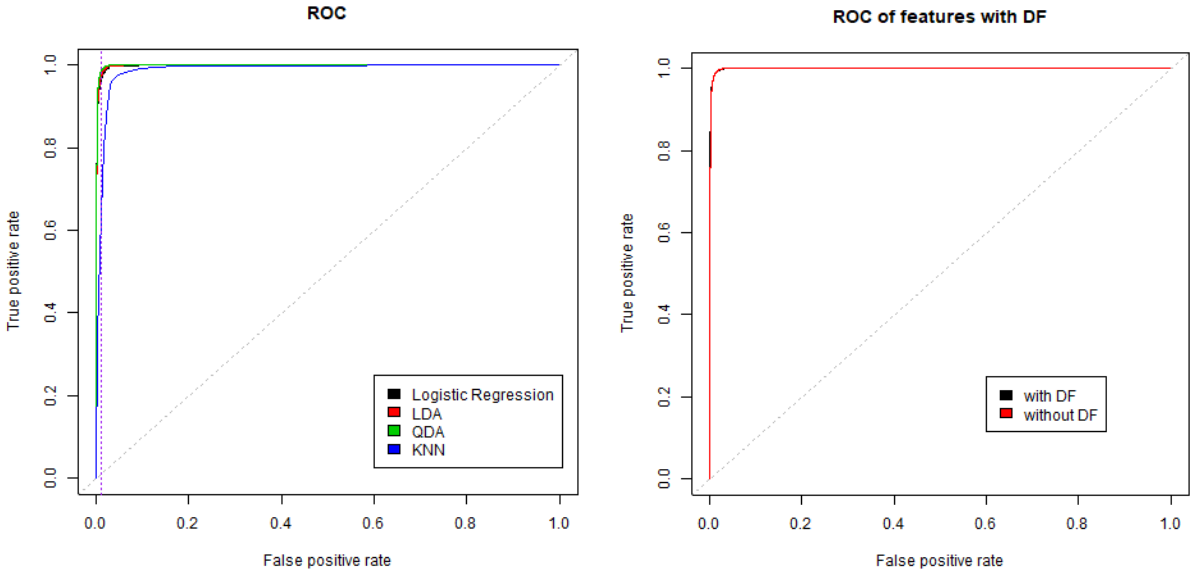

FIGURE 5

3c) Based on part 1(c), we know that DF does not have many relationships between other features, so we try to add DF feature to fit our QDA model, which is the best model compared to other methods. In TABLE 6 we can see that the model with or without DF doesn't change much. The model without DF has higher accuracy for predicting test than the model without DF, 0.9880 vs 0.9876, and higher accuracy for average across folds, 0.8746 vs 0.8627. Hence, the model with three features is still better.

	CV1	CV2	CV3	CV4	CV5	Mean	Test
Without DF	0.9293691	0.8145147	0.8587445	0.9195187	0.8509247	0.8746143	0.9880531
With DF	0.9260960	0.8043544	0.8270069	0.9077578	0.8482011	0.8626832	0.9875766

TABLE 6

4. Diagnostics

4a) QDA gives the best prediction among our classification model. However, after further investigation, the good performance of QDA relies on a good quality training dataset. We first

train our QDA on training data with 57% data in class 0. On the validation set with 90% class 0 observations, the model has 99% accuracy; on the test set with 70% class 0 observations, the model still has 98% accuracy. However, we shuffle the data again and form a training dataset that has 70% data in class 0. The model performance is not as good as the previous one. On the validation set with 65% of class 0 observations, the model has 91% accuracy; on the test set with 13% of class 0 observation, the model has only 86% accuracy. Moreover, as the figure shows below, the model predicts almost all observation on the test set to class 1, which isn't a meaningful prediction.

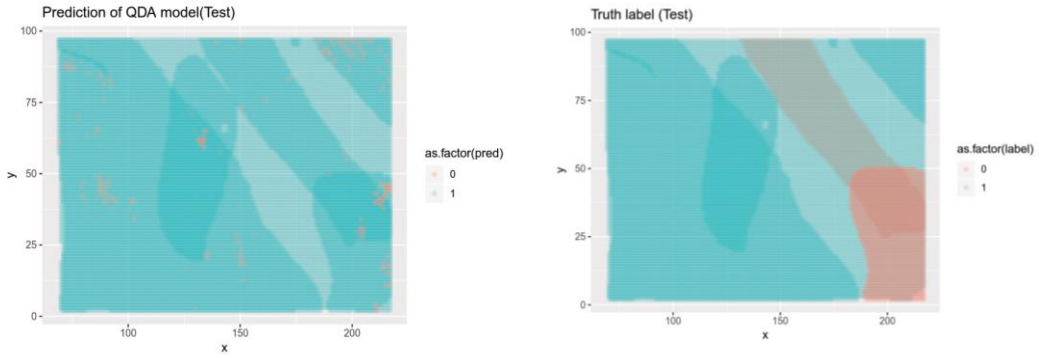

FIGURE 6

4b) For our best classification model, we notice that the features in our QDA model have patterns in the misclassification errors. The boxplots below show the distributions of SD, CORR, and NDAI, respectively grouped by all original data and misclassification data predicted by our training data model. In the first boxplot, all data labeled 0 has a smaller SD than all data labeled 1, and the misclassification data labeled 0 has larger SD than misclassification data labeled 1, which means that some no cloud points (labeled 0) with large SD are misclassified as cloud and some cloud points (labeled 1) with small SD are misclassified as no cloud. CORR and NDAI have the same situation as SD. We can see more details in the tables below. TABLE 7, TABLE 8 and TABLE 9 are the summary of the NDAI, SD, and CORR respectively grouped by original data labeled 0, original data labeled 1, classification data labeled 0 and classification data labeled 1. The numbers in the third row are closer to the second row and the numbers in the fourth row are closer to the first row in all three tables, so the misclassification data labeled 0, misclassified as labeled 1 is closer to original data labeled 1 and the misclassification data labeled 1, misclassified as labeled 0 is closer to original data labeled 0. Our model would predict the data with large SD, large CORR, and large NDAI as cloud even though they are no cloud because, in our original dataset, data labeled 0 has less SD, less CORR, and less NDAI than data labeled 1, and our model is composed by this three features.

FIGURE 7

		1st quantitle	median	mean	3rd quantitle
No cloud (label=0)	Original	-0.9756	-0.6555	-0.2627	0.1081
	Classification	1.790	2.359	2.409	2.978
Cloud (label=1)	Original	1.4936	1.9613	1.9496	2.4120
	Classification	0.7073	1.0077	0.9538	1.2415

TABLE 7

		1st quantitle	median	mean	3rd quantitle
No cloud (label=0)	Original	0.8608	1.4351	2.9785	2.6157
	Classification	4.5864	7.3006	9.8448	2.978
Cloud (label=1)	Original	4.9270	9.7757	15.0149	20.3174
	Classification	3.2343	4.6128	5.3372	6.5775

TABLE 8

		1st quantitle	median	mean	3rd quantitle
No cloud (label=0)	Original	0.1155	0.1359	0.1401	0.1617
	Classification	0.1677	0.2531	0.2630	0.3331

Cloud (label=1)	Original	0.1252	0.1688	0.1916	0.2464
	Classification	0.12819	0.15808	0.15828	0.18912

TABLE 9

4c) As shown in the previous part, our model performs poor when the observation falls in between the boundary of two classes. To account for this situation, we could try to fit new models like random forest or ada boosting. However, we could also try to add new features and hope it could capture some difference in the misclassification group. In this project, we goes for the second approach. In addition to our original features NDAI, SD and CORR, we add the label from EM algorithm using those 3 features. Using cross validation, we could see the best improvement comes from EM with 4 groups clustering as TABLE 10 shown.

	CV Fold1	CV Fold2	CV Fold3	CV Fold4	CV Fold5	Mean CV	Test
Original 3 features	0.757890	0.857833	0.948711	0.866910	0.906674	0.86760	0.98791
EM with #group = 3	0.748762	0.856897	0.950899	0.862327	0.932547	0.87029	0.986794
EM with #group = 4	0.794648	0.858959	0.963582	0.880661	0.935820	0.88673	0.988496

TABLE 10

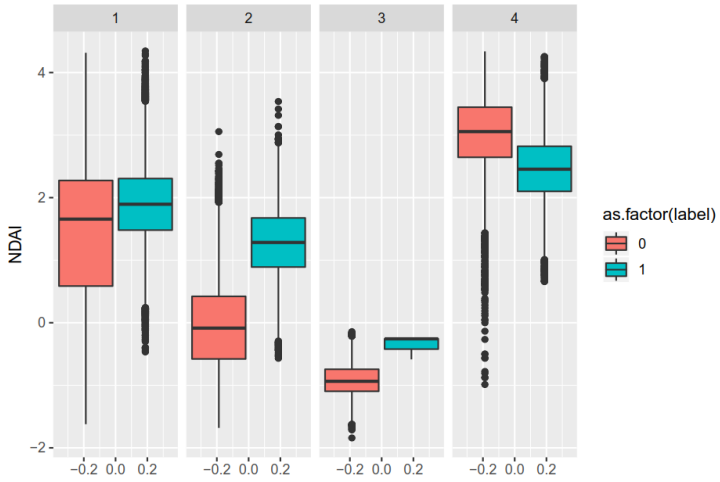

Figure 8 on the left shows the distribution of NDAI for EM with 4 groups clustering. We could see that except group 1, the two classes are separated in the other 3 groups.

FIGURE 8

4d) Our results in part 4(a) do not change as we modify the way of splitting the data. Actually, by the nature of the second split method, it will always generate training dataset with class 0

proportion closed to the class 0 proportion of all the data. Using training data with around 61% class 0 observations to fit QDA, the model doesn't perform better.
Our results in part 4(b) do not change as we modify the way of splitting the data. We use the second method mentioned in part 2(a) to split the data and fit the QDA model. Shown as FIGURE 9, the distribution of misclassification error using method 2 on CORR, SD, and NDAI are almost the same as FIGURE 7 above. Our model relies on these three features, and those features would predict the data with large SD, large CORR, and large NDAI as cloud even though they are no cloud, and predict the data with small SD, small CORR, and small NDAI as no cloud even though they are cloud.

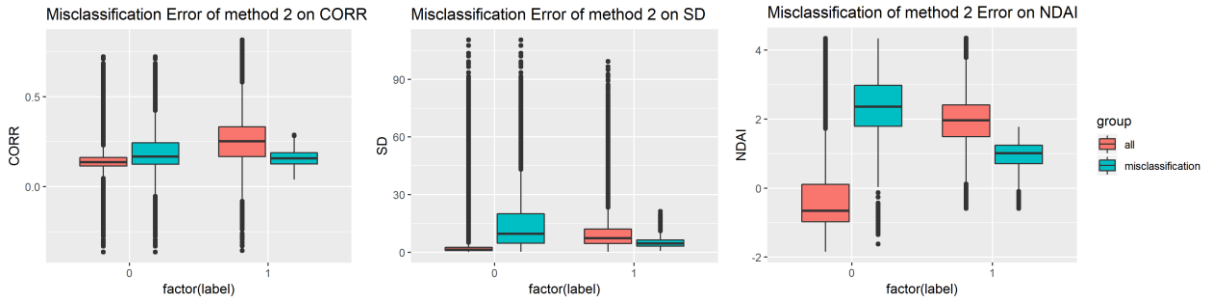

FIGURE 9

4e) In conclusion, QDA with four features (SD, NADI, CORR, and label created by EM) gives the best prediction (98.8500% of accuracy for test prediction) based on our research. First, our data has 11 features, which are x, y, CORR, NDAI, SD, AF, BF, DF, AN, CF and expert label. Expert label is the output, and x, y are the latitude and longitude of earth where the images are taken. AF, BF, DF, AN, and CF are the radiance angles. CORR, NDAI, and SD are calculated by some of these radiance angles. Predicting cloud data, we need to find some good and independent features to fit our model, which are CORR, NDAI, and SD; however, we realize that the model made up with these three features performs poor when the observation falls in between the boundary of two classes. Hence, we start to think whether the model can be improved if we use EM to compute the Maximum Likelihood estimate in the presence of missing or hidden data. We believe that set EM with 4 groups might be a good choice because we can separate the data into cloud data is predicted as cloud, cloud data is predicted as clear, no cloud data is predicted as cloud, and no cloud data is predicted as no cloud. In reality, adding the label from EM algorithm using SD, CORR, and SD features can predict a better model. Thus, QDA with four features (SD, NADI, CORR, and label created by EM) truly gives the best prediction that NASA can use to predict cloud data.

	set.seed(1)	set.seed(2)	set.seed(3)	set.seed(4)	set.seed(5)	set.seed(6)
Without EM	0.7935622	0.8145147	0.8587445	0.9195187	0.8509247	0.8746143
With EM	0.8960846	0.8043544	0.8270069	0.9077578	0.8482011	0.8626832

5. Reproducibility

Github link: https://github.com/xiaoyali97/stat154-proj2-sp19

Acknowledgment
In this project, we discuss and write the report together. Shiyun Huang focuses on 1(a), 1(c), 2(c), 3(a), 3(b), 3(c), 4(b), 4(d), and 4(e). Xiaoya Li focuses on 1(b), 2(a), 2(b), 2(d), 4(a), 4(d) and 5. When we have some questions or something that we are not sure, we would search online and ask GSI in office hour. Basically we follow the instruction to finish our project. First, we read the article carefully. Then we follow the instruction to understand the requirement and the data in 3 images. Next, we discuss, write coding, and analyze from the code together. Finally we finish the report together.