STAT 154: Project 2 Cloud Data

Anh Bui (ID# 3034323491) and Vincent Myers (ID# 3034325740)

April 26, 2019

Problem 1: Data Collection and Exploration

Part (a)

Introduction: Understanding of carbon disoxide levels in the Arctic requires accurate Arctic-wide measurements of cloud coverage, as clouds modulate the sensitivity of the Arctic to increasing surface air temperatures. Clouds are difficult to differentiate from ice and snow, as they have many of the same visual properties. Understanding cloud cover in the Arctic is critical to evaluating its impact on atmospheric radiation (and therefore warming) in the Arctic. The launch of MISR on NASA's Terra satellite in 1999 provides nine viewing angles and four spectral bands, covering a 360-km wide swath of Earth's curface. The cloud detection algorithm used by MISR was designed before MISR was launched. The algorithm does not work well over polar regions, and potential alternative algorithms are limted by the massive amount of data that must be processed. The solution proposed by the paper involves finding cloud-free pixels, instead of the prior approach of looking for cloudy pixels. The solution uses an ELCM algorithm to label the data as cloudy or not-cloudy, and then uses QDA to compute a probability of cloudiness for each pixel.

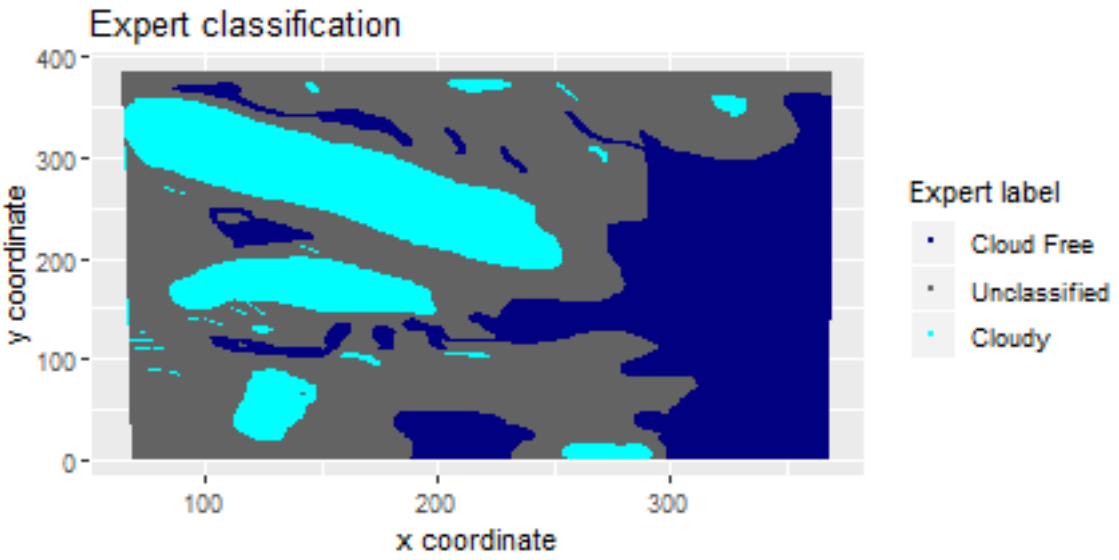
Data and Methodology: The data was collected from 10 MISR orbits of Path 26 over the Arctic, Northern Greenland and Baffin Bay. The problem was solved in three steps: (1) construct three features (CORR, SD and NDAI) using EDA and domain knowledge; (2) build the ELCM algorithm by setting thresholds on each of the three features; and (3) predicting the probability of cloudiness using QDA on the expert labels.

Results: The ELCM algorithm has an agreement rate (agreement with the expert labels) of 91.8%, which compares favorably to the agreement rates for MISR ASCM (83.23%), SDCM (80%), and the offline SVM (80.99%). The ELCM-QDA solution provides additional information in the form of cloudiness probabilities.

Conclusion: The paper shows the importance of the three chosen features (CORR, SD and NDAI) in determining cloudiness of each pixel. It also highlights the importance of having statisticians involved in a study from the start in order to help design the processes, and illustrates the usefulness of statistical methods. Finally, the paper contributes to an improved understanding of the relationship between cloud cover and changes in the Artic climate.

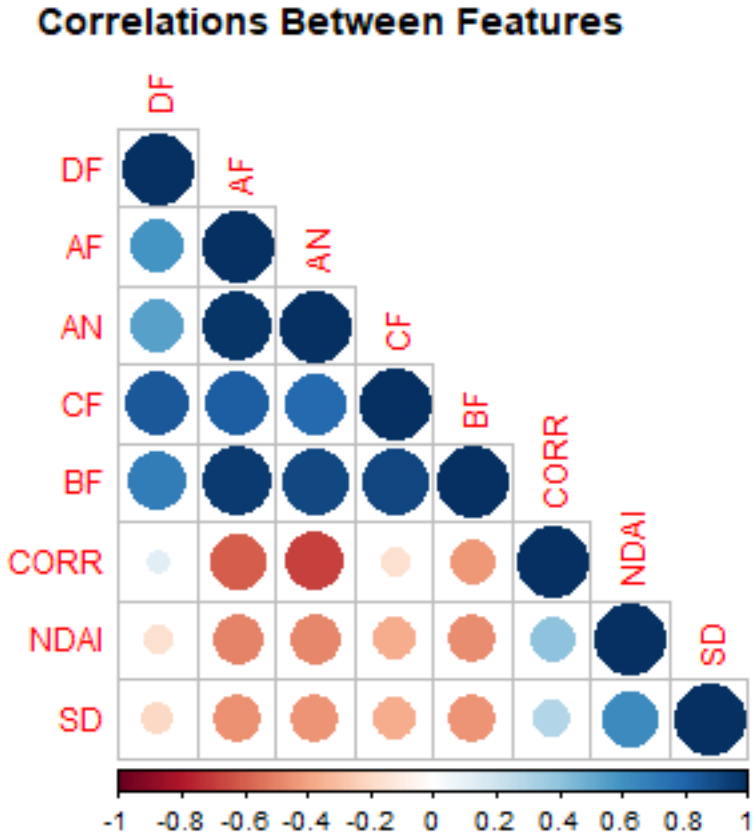
Part (b)

There are 36.8 percent of the pixels are labeled as -1 (cloud-free), 23.4 percent are labeled as 1 (cloudy), and 39.8 percent of the pixels are not labeled. Most of the pixels (72%) with an x-coordinate greater than 300 are labeled cloud-free. In general, the pixels are grouped by the expert labels; clouds pixels are clustered together, non-cloud pixels are clustered together, and unlabeled pixels are clustered together. Because of these clusters, we would not interpret the data to be independently distributed; for instance, if a given pixel is surrounded by cloud pixels, that pixel is very likely to be cloudy as well.



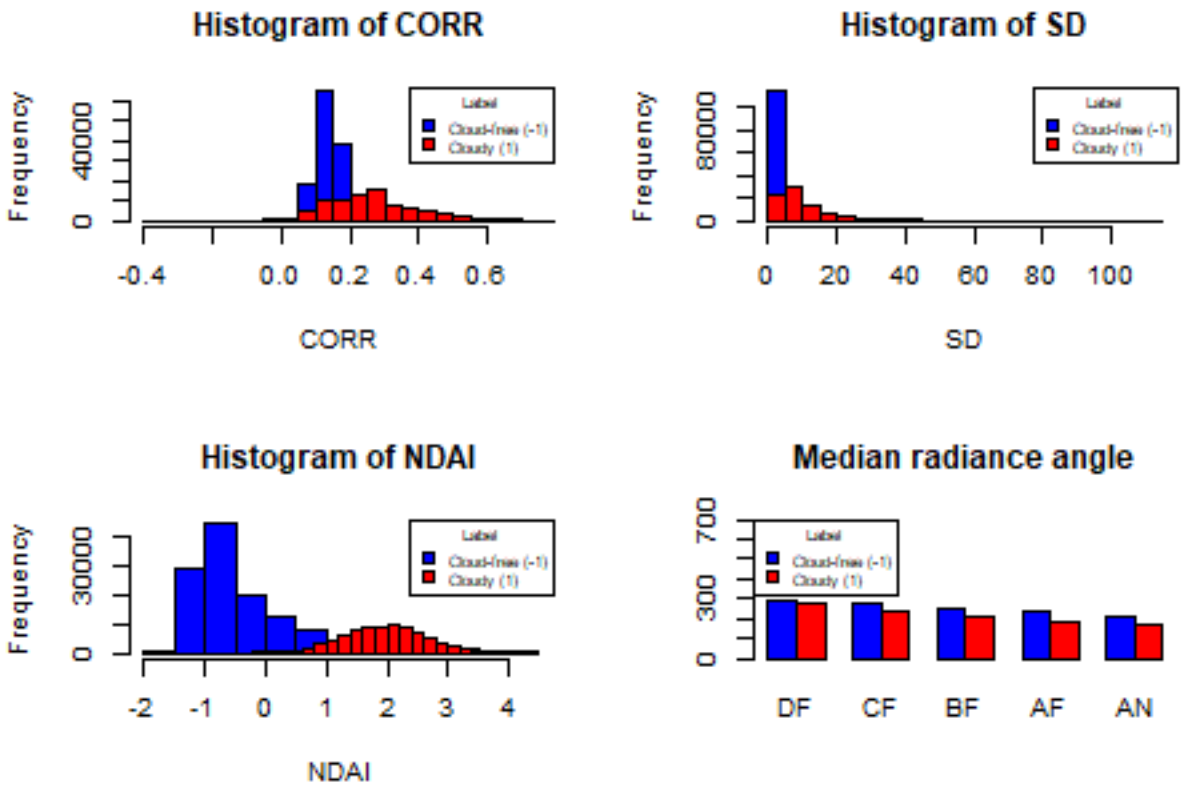
Part (c)

The correlation plot demonstrates that the values of the angular features are highly correlated with one another. The angles AF, AN and BF have strong negative correlations with CORR, NDAI and SD. NDAI and SD are positively correlated with one another.



The charts below demonstrate some of the differences between cloudy pixels and cloud-free pixels. In general,

cloudy pixels have higher CORR, SD, and NDAI values than cloud-free pixels. Cloudy pixels also have lower median radiance angles for AN, AF, BF and CF than cloud-free pixels.



Problem 2: Preparation

Part (a)

First Method: Divide the data by the expert labels (-1 and 1), and use 20 percent of the data in each category for testing and another 20 percent for validation (leaving 60 percent for training). This ensures that both cloudy and cloud-free data are well-represented in each of the training, validation, and test sets.

Second Method: Split the data into blocks based on the x and y-coordinates. For example, the first block includes observations that have x-coordinate and y-coordinate values that are less than or equal to 100; the second block includes data points with x-coordinate and y-coordinate values between 100 and 200; and so on. We can use 20 percent of the observations in each block for the test set and another 20 percent of each block for the validation set (again leaving 60 percent for training).

Part (b)

Given a trivial classifier which sets all labels to -1, the accuracy rates of validation and test set are 61 percent (only considering data with expert labels). The classifier would have a high accuracy rate if our random selection from the original data set had a higher-than-expected percentage of rows labeled as cloud-free.

Part (c)

To assess the relative importance of individual predictors in the model, we fit individual logistic regressions on each of the predictors, and chose the predictors that have smallest test error rates. The three best predictors based on this standard are NDAI, SD, and CORR.

```
##   error_rate Xnames
## 1     0.1011   NDAI
## 2     0.1553   CORR
## 3     0.1589     SD
## 4     0.2298     AN
## 5     0.2448     AF
## 6     0.2744     BF
## 7     0.3185     DF
## 8      0.333     CF
```

Part (d)

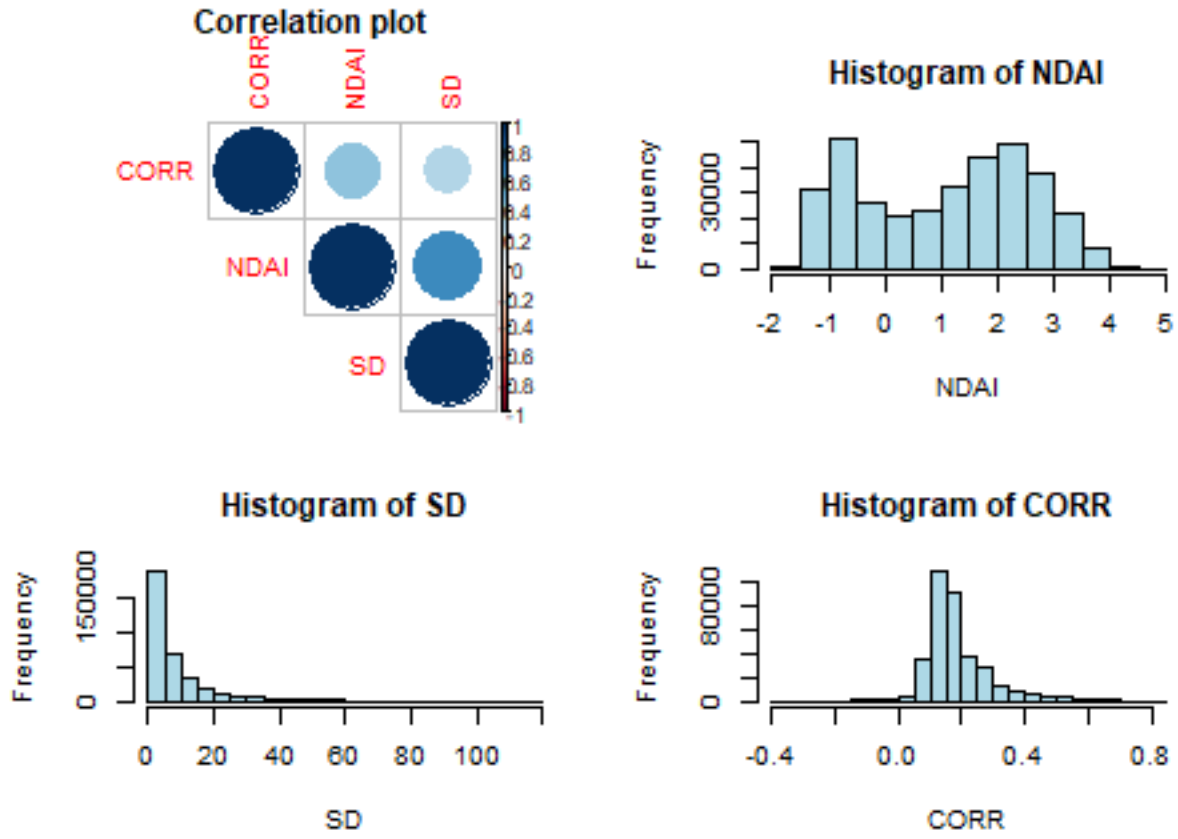
Code for the generic cross validation function is posted on Github.

Problem 3: Modeling

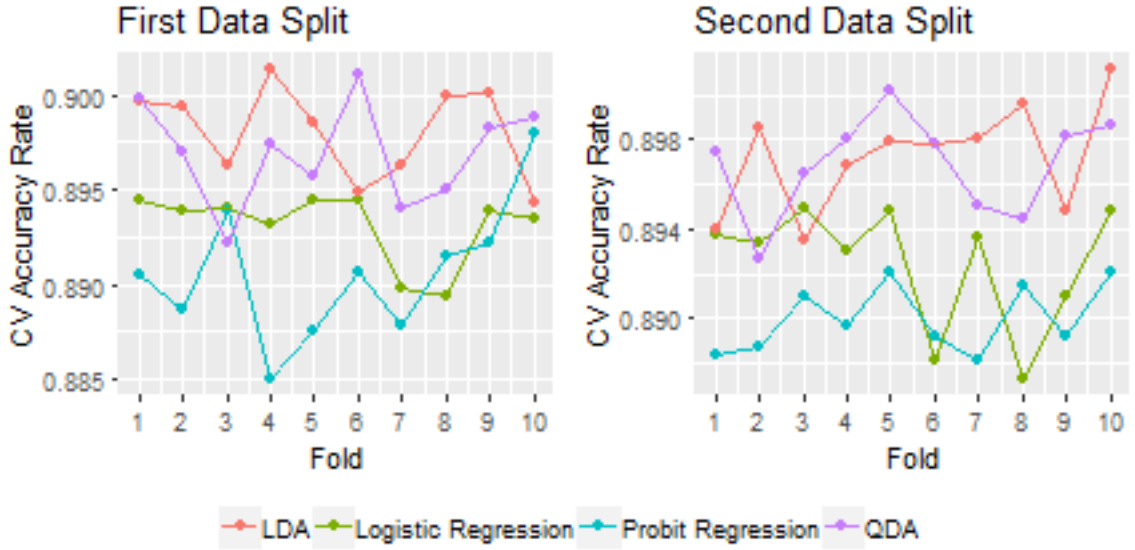
Part (a)

Logistic regression and probit regression assumes that the observations are independent of each other, and that the independent variables show little or no multicollinearity. The assumptions for logistic regression are not satisfied well in this study since the independent variables are highly-correlated, and the observations are not independently distributed due to the clustering of cloudy and cloud-free pixels.

Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) assume that each variable follows a normal distribution. The LDA model assumes that both classes of the data have equal variances. As demonstrated by the histograms below, the NDAI, SD and CORR variables do not follow a normal distribution.

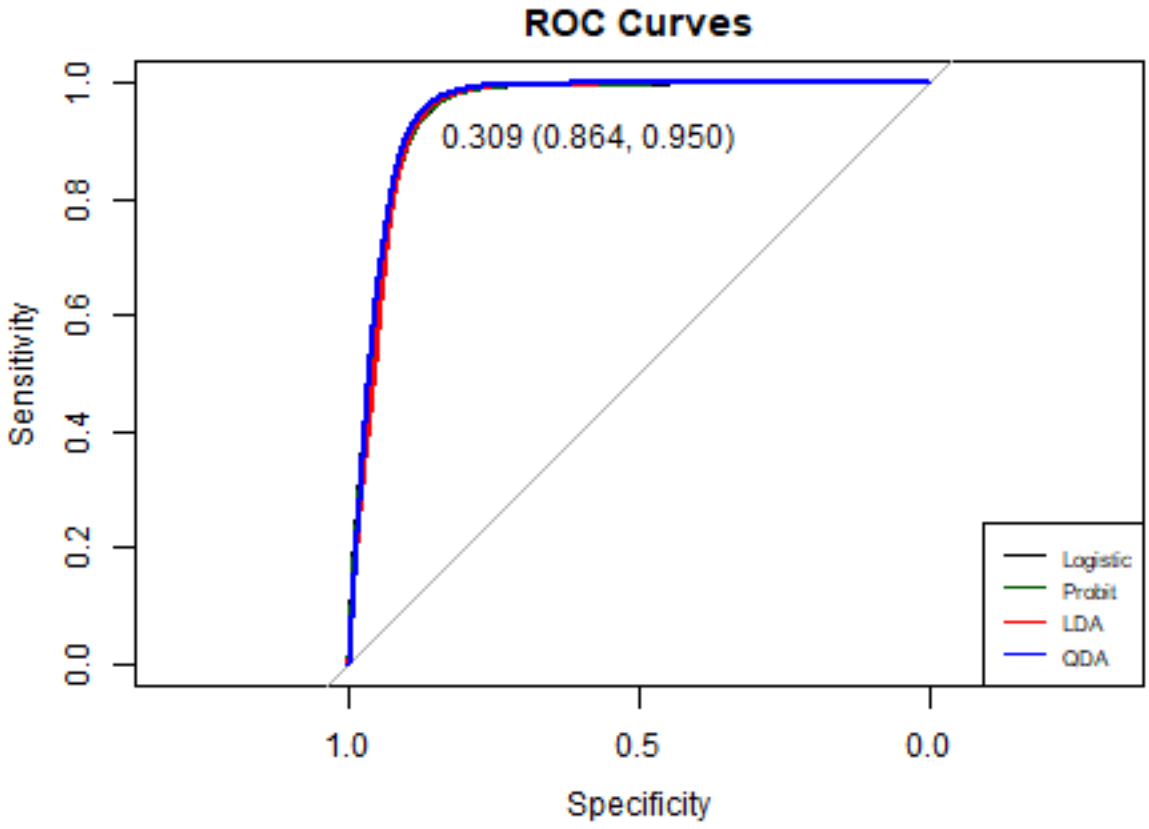


The cross-validation accuracy rate for the combined training/validation set is generally lowest for probit regression. Logistic regression has a slightly higher accuracy rate, while LDA and QDA have the highest accuracy rates. The cross validation accuracy rates are shown by fold in the charts below. On the test set, the best performing method for the first data split is QDA (89.6% accuracy), followed in order by LDA, logistic regrssion, and probit regression; and the best performing method for the second plit was LDA (89.9% accuracy), followed in order by QDA, logistic regression and probit regression. Averaging the accuracy rates from the two methods, LDA has the highest test accuracy rate.



Part (b)

The ROC curve plots the Specificity on the x-axis (defined as 1 - FPR, where FPR is the False Positive Rate) and the Sensitivity (or the True Positive Rate) on the y-axis for all possible thresholds. The plot below shows the ROC curves for all four models; the ROC curves are very similar, which makes the lines in the chart difficult to distinguish. The chart includes the point of the best threshold for the logistic regression curve (0.309). The chart demonstrates that, while a threshold of 0.5 has been used for all models up to this point, there may be accuracy gains by varying the threshold by model.



Part (c)

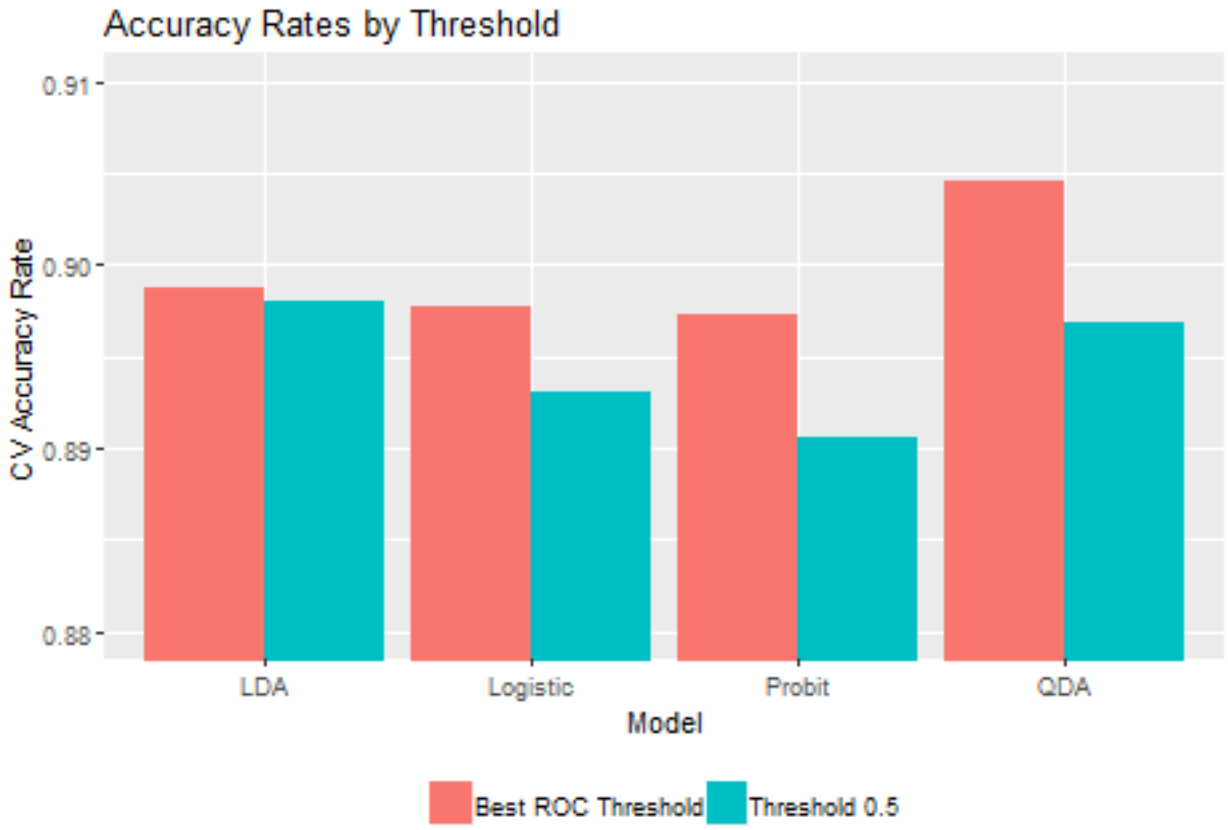
Precision is the ratio of correctly predicted positive observations to the total predicted positive observations (True Positive / [True Postive + False Positive]). High precision corresponds to a lower false positive rate. Recall is the ratio of correctly predicted positive observations to the all observations in actual class positive (True Positive / [True Positive + False Negative]). The F1 Score is the weighted average of Precision and Recall. Using the optimal thresholds for each model as determined in the previous section, the QDA model has the highest precision and the highest F1 Score, while the LDA model has the highest recall.

Problem 4: Diagnostics

Part (a)

Using the updated threshold values from problem 3(b) above for each model, QDA has consistently higher accuracy rates than the other three models. Therefore, we will focus on the results from the QDA model for this section. The chart below demonstrates the increase in the cross-validation accuracy rates for the four

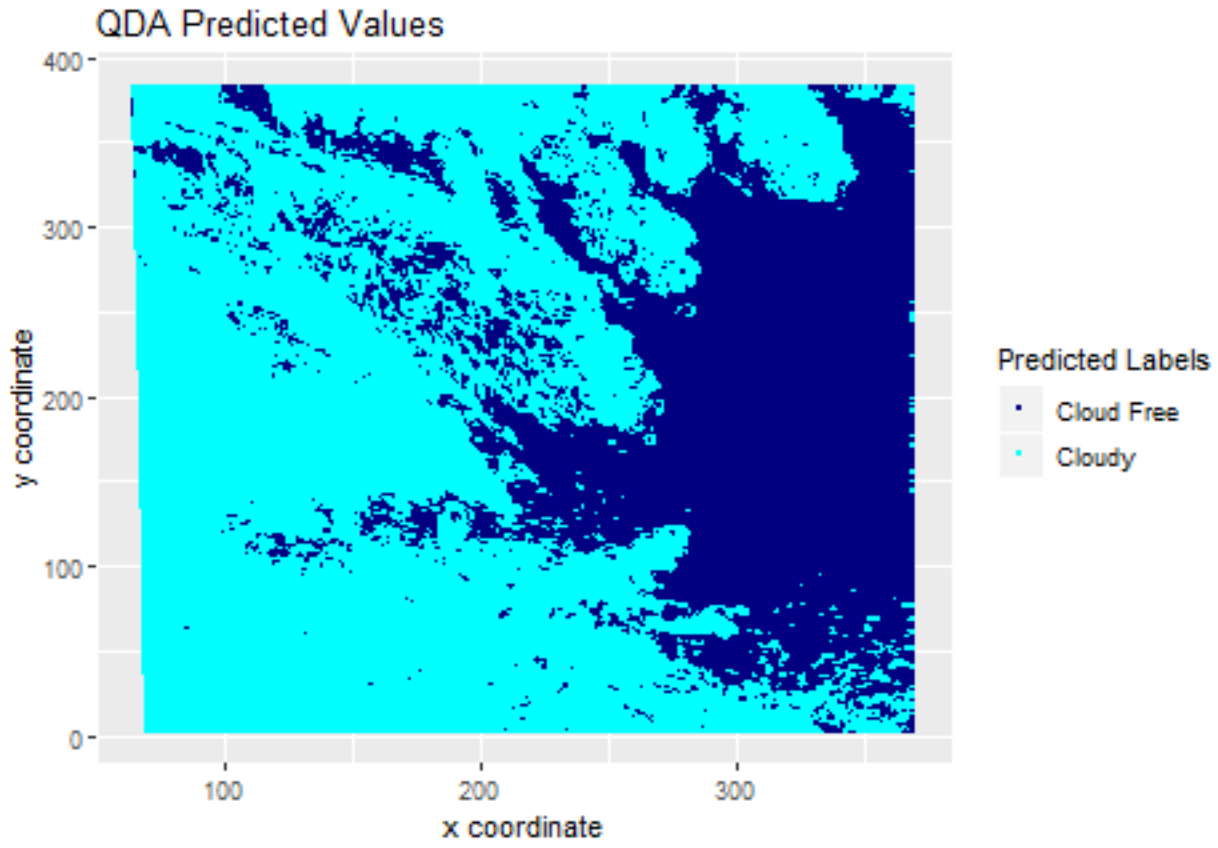
models by using the updated threshold values from the ROC curves instead of using 0.5 as the threshold for all models. The chart also demonstrates that QDA is the most accurate model.



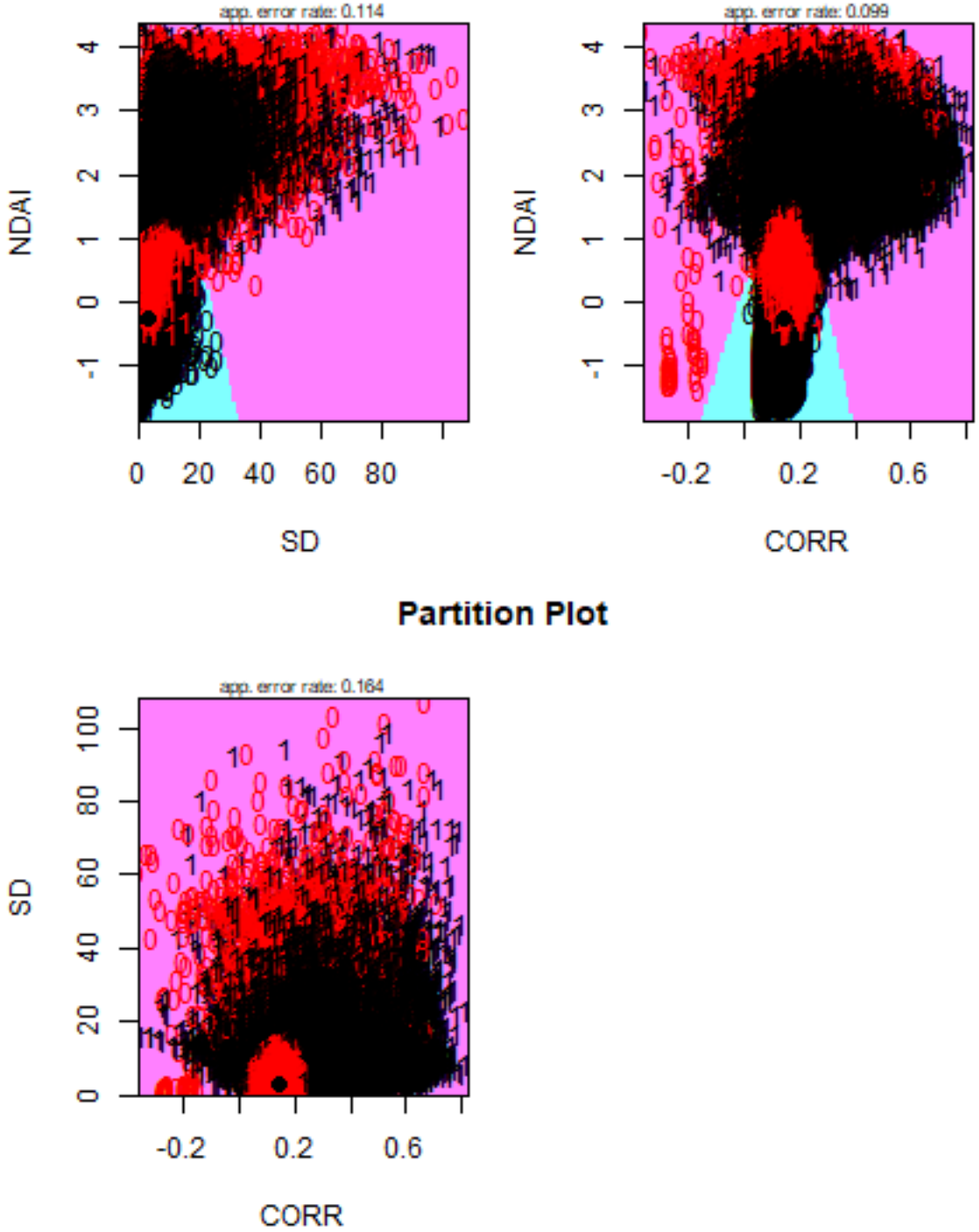
The table below shows the confusion matrix for the expert labels against the QDA predicted values for the test set, based on the first data split. The test error rate for the QDA model is 9.72%.

```
##     qda.predictions.test
##          -1      1
##   -1 0.5322 0.0786
##   1  0.0186 0.3706
```

Below is the map shown in section 1(b) of this paper, with the colors based on the labels predicted by the QDA model rather than the expert labels.



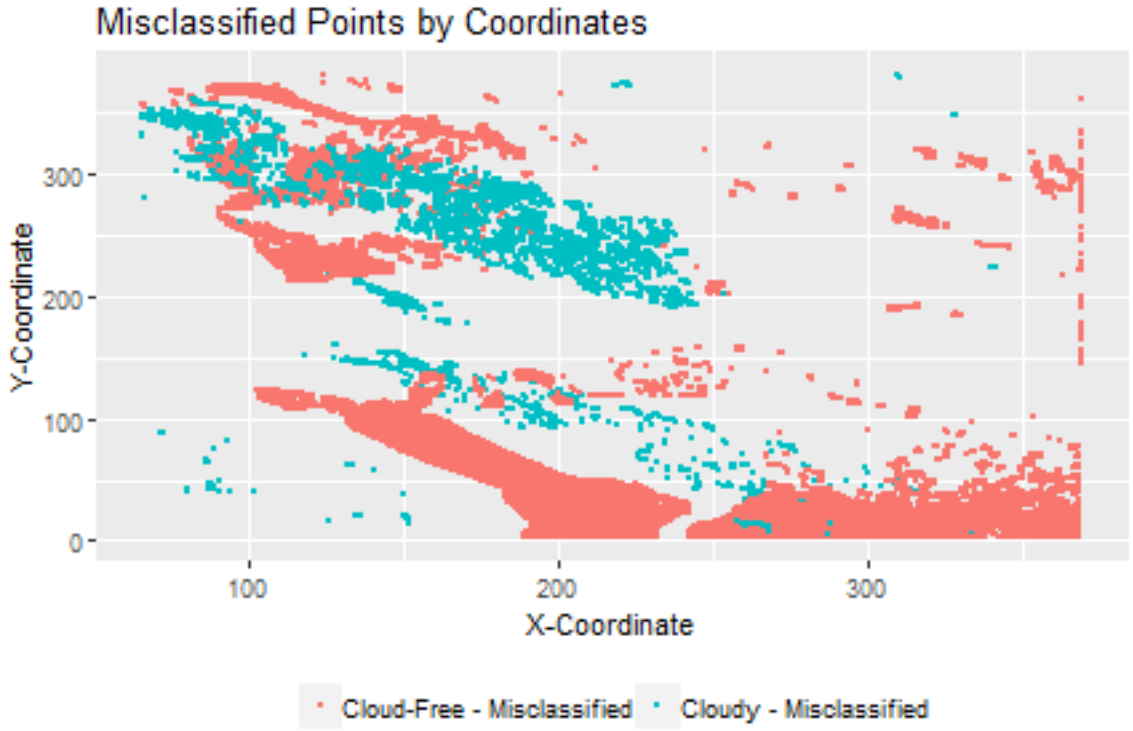
The charts below show scatterplots for each coupling of the three variables (NDAI, SD and CORR), demonstrating the actual labels along with the classification lines drawn by the QDA model.



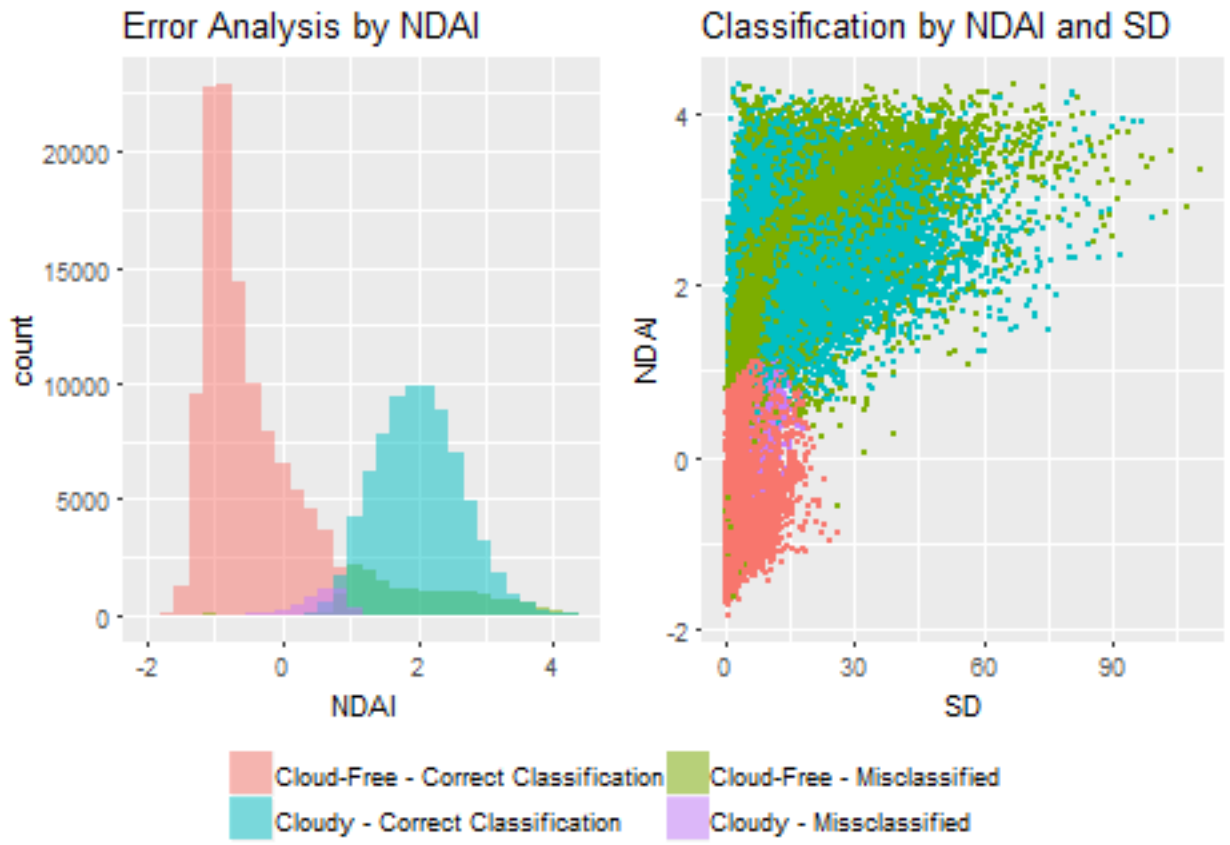
Part (b)

The chart below shows the misclassified points only, plotting by the x-coordinate and the y-coordinate. The misclassified points which were expert labeled cloud-free are found largely at the bottom of the chart (below

y-coordinate 150), and alongside the right edge of the chart; however, there is a large section at the top left coordinate of the chart where a large number of misclassified points from both categories are found.



Below, the chart at left is a histogram of NDAI values classified by the expert labels and whether or not the QDA model misclassifies the pixels; and and right, a scatter plot of NDAI and SD, again based on the different classifications. The left chart demonstrates that cloudy and cloud-free pixels have separate but overlapping distributions by NDAI, and the misclassified points tend to fall in the overlapping portion of the distribution (i.e. cloud-free pixels at the high end of the cloud-free distribution have NDAI values more commonly associated with cloudy pixels, and are therefore misclassified as cloudy). The right chart demonstrates that points with high values of SD (larger than 30) and high values of NDAI (larger than 3) tend to be misclassified at a higher rate than the rest of the data.



Part (c)

As shown above, the QDA model performs poorly when values of NDAI and SD are both high. As a solution, we have added an interaction term to the model, which multiplies the values of NDAI and SD. The classification table resulting from the updated model is below. By including the interaction term, we improve the test error rate from 9.72% in the previous model to 9.45% in the updated model. We expect that the updated model will work well on future data without expert labels, assuming that the data provided for this model is representative of future data to be modeled.

```
##     qda.predictions.update
##          -1      1
##   -1 0.5311 0.0797
##   1  0.0148 0.3745
```

Part (d)

Our results are not materially different between the two ways of splitting the data.

Part (e)

Conclusion: Our paper used three variables from the MISR data (NDAI, CORR and SD) to fit four different models in order to predict whether a given pixel is cloudy or cloud-free. The models we chose to analyze were logistic regression, probit regression, LDA and QDA. Prior to training the model, we split the data in

two different ways that adjusted for the fact that the data is not independently distributed. After adjusting our threshold values for each model based on the output from the ROC curves, our team determined that the QDA model had the lowest error rate on test data (9.72%). In analyzing the data that was misclassified by the QDA model, we determined that data points with both high NDAI and high SD values were frequently misclassified. Due to this realization, we added an interaction term to the QDA model which was the product of the NDAI and SD values. With the inclusion of the interaction term, the error rate was further reduced from 9.72% to 9.45%.

GitHub Repo

Our GitHub repo can be found at the following link: https://github.com/vincent-myers/stat154-proj2

Acknowledgements

Anh was responsible for the initial data analysis for the project, and is responsible for Problem 2 (including the CV function). Vincent wrote the summary in section 1(a). The remainder of the sections were developed by both Anh and Vincent together. Our process started with Anh working through the data and providing an initial draft of the report, and Vincent subsequently working to fill in the remainder of the report. We would like to acknowledge the following link (https://rstudio-pubs-static.s3.amazonaws.com/35817_2552e05f1d4e4db8ba87b334101a43da.html) as the source for the idea and code for the partition plots in section 4(a).