Statistical Methods of Classification on Arctic Cloud Data

Omri Newman SID: 3032273024
Phoebe Abramowitz SID: 26386343
Friday, May 3, 2019

1. Data Collection and Exploration

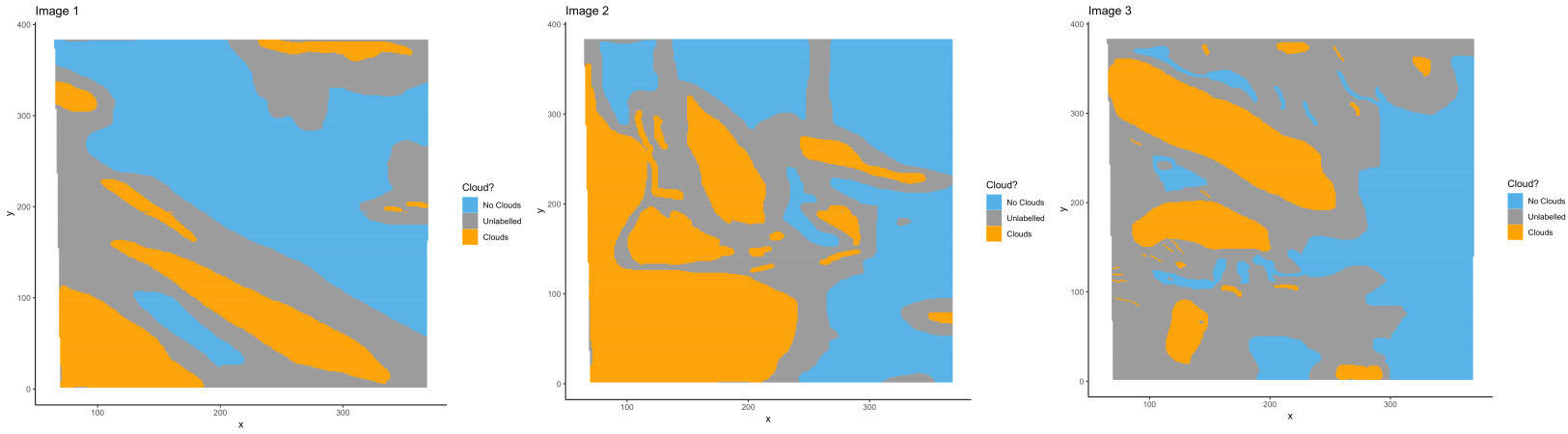
Summary of Paper

The purpose of this study is to build an accurate cloud detection algorithm that can process Multi-angle Imaging SpectroRadiometer (MISR) data without human intervention. In the last 20 years, there has been a global increase in green and sustainable climate initiatives, discussions which often begin with an analysis of cloud formation. Clouds in the polar regions have the most effect on global climate, and it is difficult to discern between white snow and white clouds from satellite images. Thus, it is imperative to have reliable algorithms that accurately distinguish between snow and clouds in order to have accurate global climate models.

The data in this study were compiled in 2002 by NASA's Terra satellite, which is equipped with a Multi-angle Imaging SpectroRadiometer and collected from ten orbits over one path. The paper analyzed 57 data units containing 7,114,248 1.1-km resolution pixels, with 36 radiation measurements for each pixel. The analysts focused on the 275-m red radiation measurements to build some of the features, so the entire data set is even larger than what is expounded upon in the paper. The study includes repeated visits over time to improve the expert labeling process, which created accurate classifications to build models from.

This study demonstrates the power of statistical models and algorithms, and their potential impact on human understanding of complex scientific problems. This research aims to help scientists better understand the flow of visible and infrared radiation in the atmosphere, subsequently studying its response on clouds and changes in the arctic climate. Accurate arctic cloud models will yield a more complete and sound global climate model. Eventually these studies will further illuminate how changing cloud properties relate to broader changes in the arctic as concentrations of carbon increase in the atmosphere.

Plots of Satellite Images



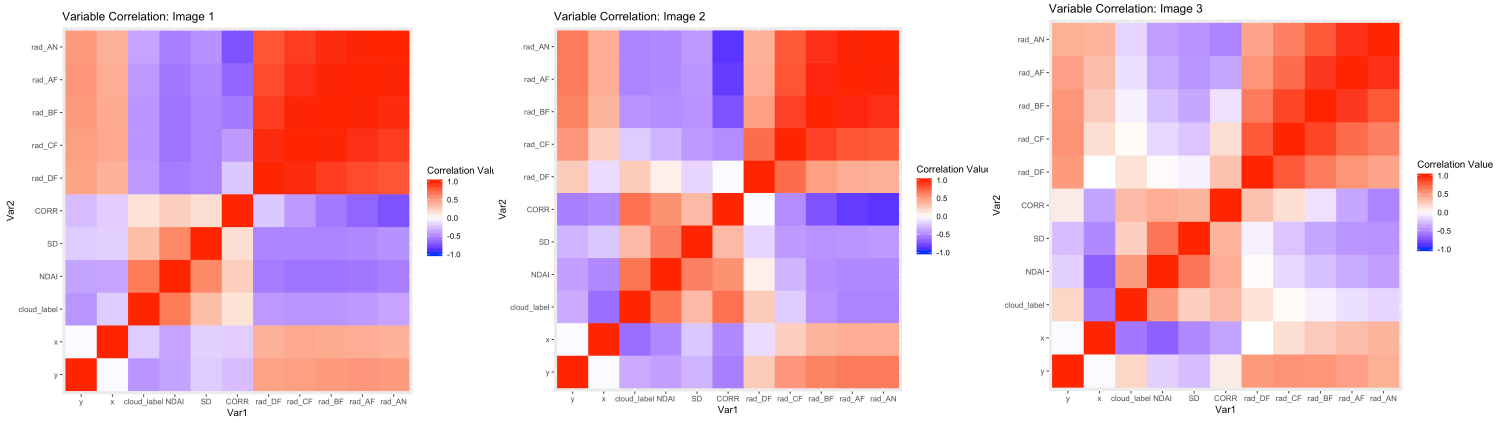
Percent of Pixels by Class

	Image 1	Image 2	Image 3
Cloud	18%	34%	18%
No Cloud	44%	37%	29%
Unlabelled	38%	29%	52%

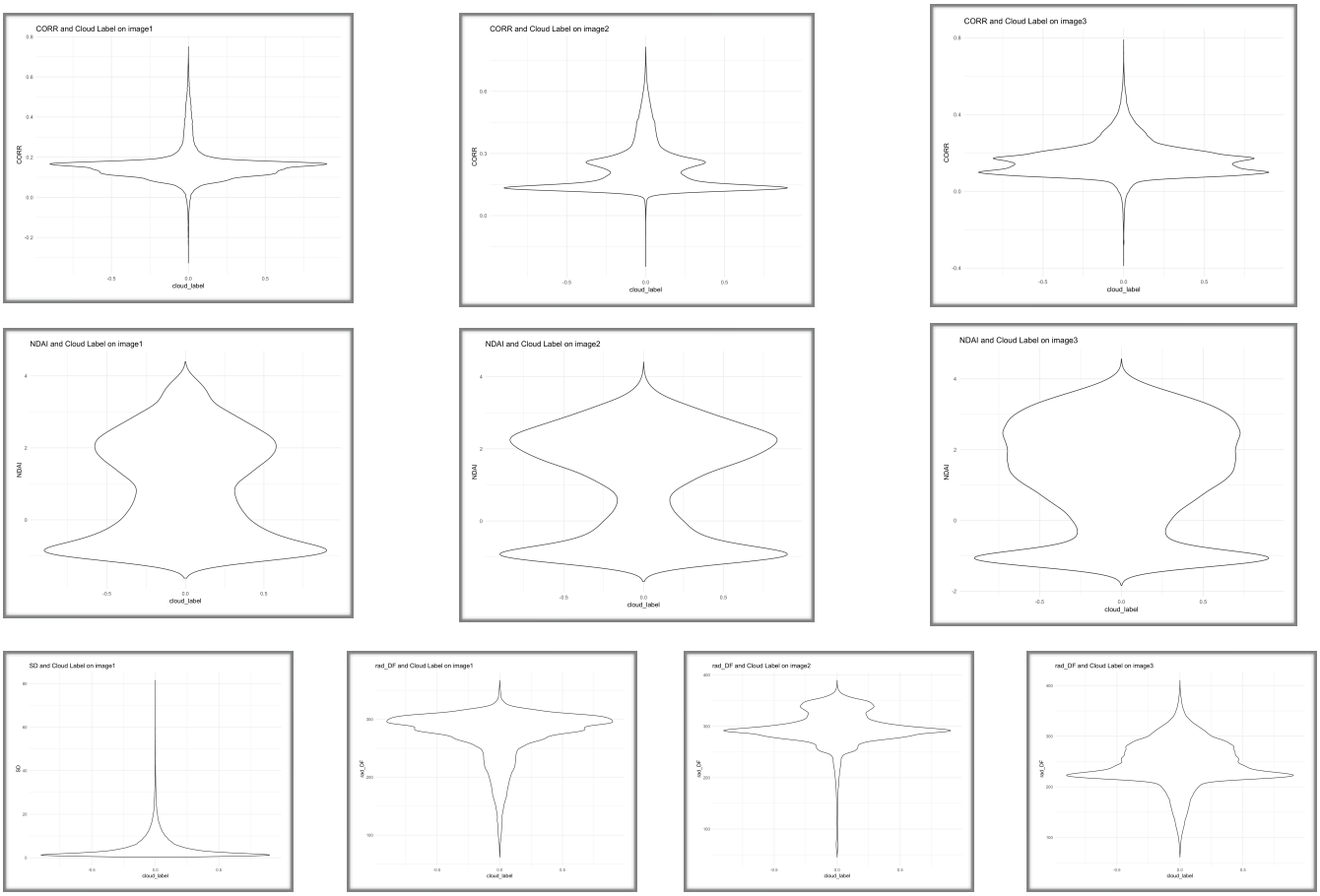
There is a large quantity of unlabelled data, generally on the border between cloud and no-cloud areas in the three satellite images. Clouds exist in regions much larger than one pixel, and thus are grouped spatially.

An i.i.d. assumption for the samples is not justified, because they are not independent. The label of one pixel is dependent on that of the pixels next to it, as demonstrated by the shape of the cloud and no-cloud regions above.

The following heat maps depict the pairwise relationships between all features in each image. We notice high correlation between the radiance angles, which follows intuitively since they measure radiance in the same location from different angles. These heat maps also illustrate the correlation between the features and cloud labels. NDAI and CORR appear to have relatively high correlation with cloud label in comparison to the other features.



Violin Plots of CORR, NDAI, SD, rad_DF with Cloud Labels



Since the radiance angles are highly correlated with one another, their relationships to the expert labels -- and consequently their respective violin plots — are largely similar. Thus, we only look at the violin plot of rad_DF and the expert labels, and notice some differences in the plots across images. These differences make sense because of substantially differing terrain and cloud formations between the three images, which can be seen in the plots of satellite images on page 2. .

We see differences between the violin plots for NDAI and CORR across images, but the plots for SD look the same - only one is shown above. The SD feature is used in the paper to identify smooth surfaces in the satellite images. Given that the violin plot of SD and the expert labels hugs the x-axis with an SD value close to zero, and that we are using a very small subset of images in our project, it is likely that our images have largely smooth surfaces. CORR is used as an average linear correlation of radiation measurements at different view angles, and has noticeable differences across the three images when considered with cloud labels. NDAI is defined as the normalized difference angular index, which has the widest and largest spread when compared to cloud labels. This wide spread for NDAI is consistent across the three images.

The correlation heat map shows darker boxes for cloud_label and NDAI as well as cloud_label and CORR. The box for cloud_label and SD is less shaded. This follows in the respective violin plots -- the higher the correlation, the greater the area inside the violin plot.

2. Preparation

Given the non i.i.d nature of these data, we use two different split methods and asses our models on both to compare. The first method splits all the image data points randomly into 80% training set, 10% validation set, and 10% test set without any prior transformations. The second method divides the data into 3 pixel X 3 pixel blocks and creates 'super pixels' to transform the data into a new set. This transformation decreases the number of data points, and works to reconcile the dependence relationship between the points by blurring the overall data. Although the classification methods used have less data to train on as a result of this transformation, more assumptions can be made about the data at hand if they are i.i.d. and thus will yield a higher model accuracy.

In order to establish a baseline for this classification problem, we report the accuracy of a trivial classifier which sets all labels to cloud-free on the validation and test sets, for both split methods. This kind of classifier will have high average accuracy if the satellite images depict mostly cloudless regions. Since this serves as a baseline accuracy for our models, we can expect our classification methods to perform much higher than 60%.

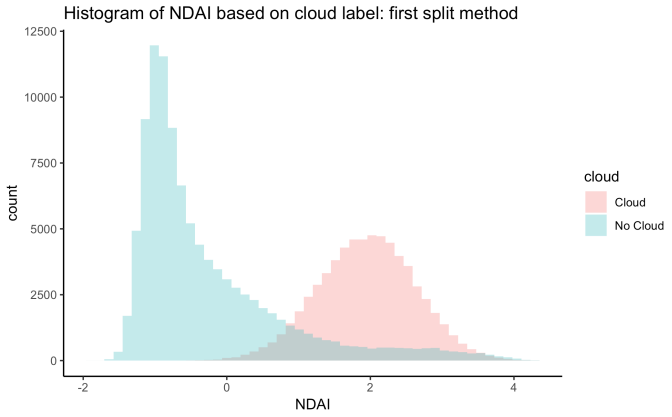
Data Set	Trivial Classifier Accuracy
1st method Val. Set	60.9%
1st method Test Set	61.5%
2nd method Val. Set	59.1%
2nd method Test Set	61.1%

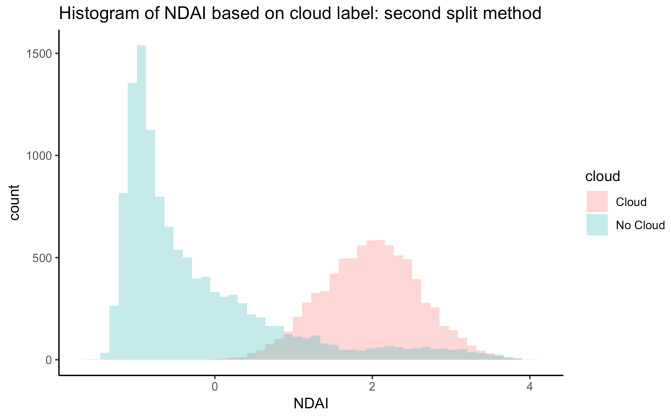
Feature Importance

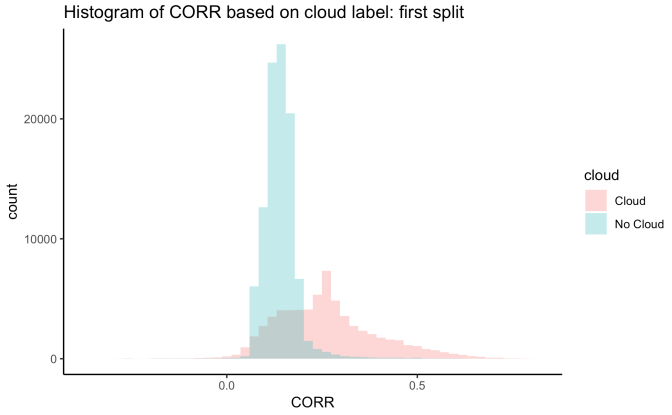
We used overlapped histograms to determine the "best" three features in the dataset. The farther the spread and the smaller the overlap between the two distributions--cloud and no cloud-- the better. We expect the distribution of cloud vs. no cloud to be different. Moreover, we want to pick features that can minimize ambiguity when distinguishing between the two labels. Based on our findings, NDAI seems to be the best feature, since it has minimal error (shown here as overlap between the histograms).

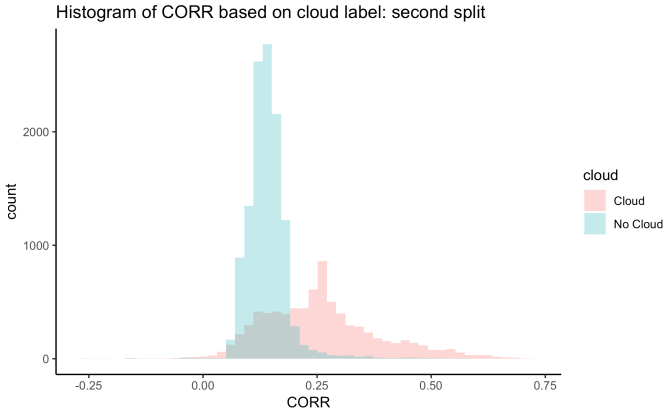
We also compare CORR and SD, and while they have greater error than to NDAI, they have clearer distinctions than the individual radiance angles. In an attempt to aggregate the collective information provided by the five radiance angles, we performed Principal Component Analysis on these angles, and plotted the overlapped histogram to determine whether or not the first principal component is particularly useful as a classification feature.

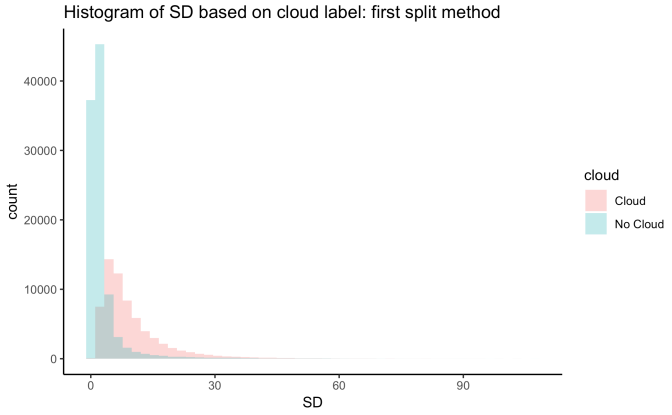
From the scree plot, PC1 captures 85% of the variance of all five radiance angles using the first split method. It does not do a good job here of predicting cloud labels however, and so we will not initially choose any of the radiance angles, or the respective principal component, as features to train our classifier.

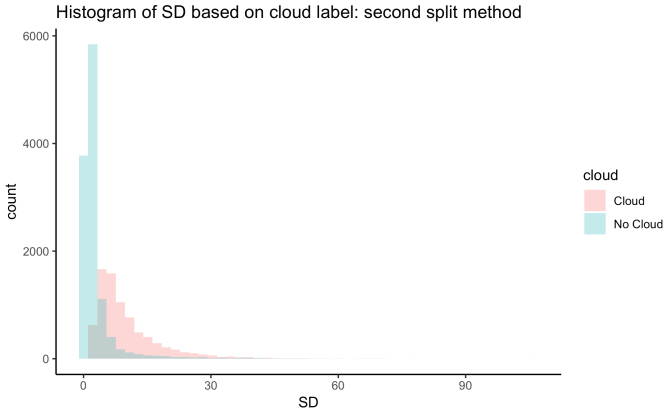

Histogram of NDAI based on cloud label: first split method
count
NDAI
cloud
Cloud
No Cloud

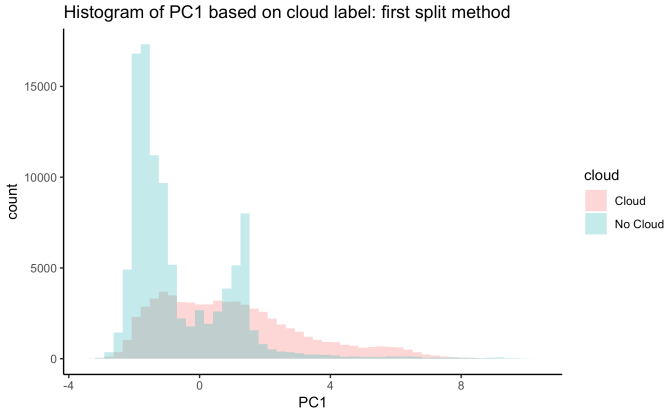

Histogram of NDAI based on cloud label: second split method
count
NDAI
cloud
Cloud
No Cloud

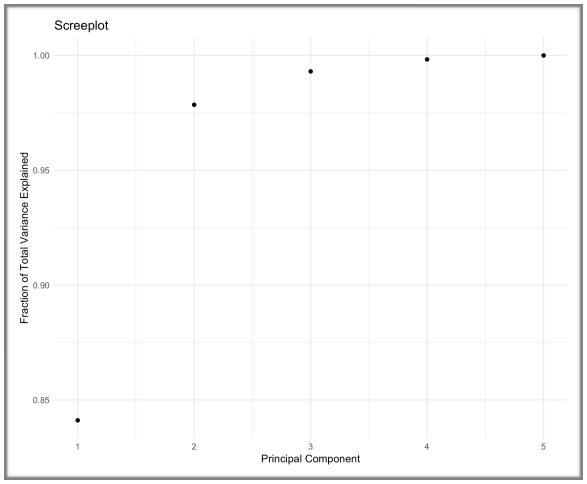

Histogram of CORR based on cloud label: first split
count
CORR
cloud
Cloud
No Cloud


Histogram of CORR based on cloud label: second split
count
CORR
cloud
Cloud
No Cloud


Histogram of SD based on cloud label: first split method
count
SD
cloud
Cloud
No Cloud


Histogram of SD based on cloud label: second split method
count
SD
cloud
Cloud
No Cloud


Histogram of PC1 based on cloud label: first split method
count
PC1
cloud
Cloud
No Cloud


Screeplot
Fraction of Total Variance Explained
Principal Component

3. Modeling

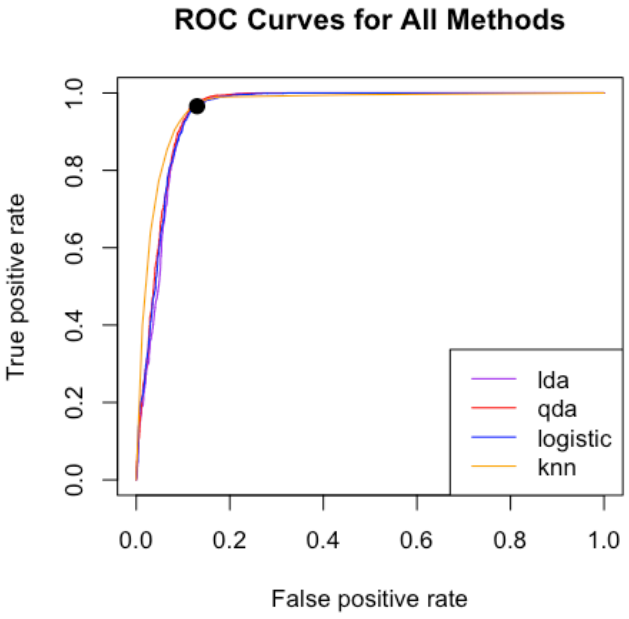
Five-fold Cross-Validation Accuracy on Both Split Methods:

KNN	1st method	2nd method
Fold 1	91.5%	87.6%
Fold 2	91.2%	86.6%
Fold 3	91.2%	86.5%
Fold 4	91.3%	86.9%
Fold 5	91.1%	85.7%

Logistic	1st method	2nd method
Fold 1	89.1%	90.8%
Fold 2	89.3%	89.3%
Fold 3	89.5%	89.9%
Fold 4	89.3%	89.5%
Fold 5	89.2%	88.9%

LDA	1st method	2nd method
Fold 1	90%	91%
Fold 2	89.6%	90.6%
Fold 3	89.7%	90.7%
Fold 4	89.8%	90%
Fold 5	89.6%	89.9%

QDA	1st method	2nd method
Fold 1	89.6%	91.2%
Fold 2	89.7%	89.5%
Fold 3	89.5%	90.1%
Fold 4	89.7%	90.9%
Fold 5	89.7%	89.9%



Model Accuracy on Test Set:

Model	Accuracy
Logistic	88.96%
QDA	89.83%
LDA	89.89%
KNN	91.41%

Model Assumptions and Commentary

LDA and Logistic regression produce linear decision boundaries; the main difference between them exists in the way each method computes the respective coefficients. Logistic regression coefficients are computed with an estimated mean and variance from the normal distribution, while LDA coefficients are computed with maximum likelihood estimates. Furthermore, LDA assumes the observations are drawn from a Gaussian distribution with a common covariance matrix, and outperforms Logistic regression when these assumptions are met. If the data are not drawn from a Gaussian distribution then Logistic regression will have higher accuracy.

KNN makes no assumptions about the decision boundaries, and hence will outperform LDA and Logistic regression when the boundary is non-linear, as it is here.

QDA assumes a quadratic decision boundary, and thus can, in general, accurately model a wider range of problems when compared to linear boundaries. QDA can perform better in the presence of less training observations, because of the assumptions it makes about the decision boundary.

Our LDA model outperforms Logistic regression in test accuracy by 1% on the transformed data. This could imply that the Gaussian assumptions are met, and confirm our theory that transforming the data reconciles the dependence relation among the data. LDA and QDA perform similarly, while KNN outperforms all models. Thus, it is evident the decision boundaries for our classification problem are highly non-linear, and that there are a sufficient number of data points to train on in the transformed data, despite having 1/9th the original amount. This conclusion about non-linear decision boundaries also follows from the satellite images on page 2.

We see from these ROC curves that KNN outperforms the other three classification methods. We chose a cutoff value as close to the point (0,1) as possible, because we want low false positive rates and high true positive rates. Given the classification problem at hand, it is not catastrophic to misclassify a few values as it would be in a cancer diagnosis. For this reason we don't have to worry about picking a slightly lower false positive rate, and can focus instead on choosing a value which will yield as high an accuracy as possible.

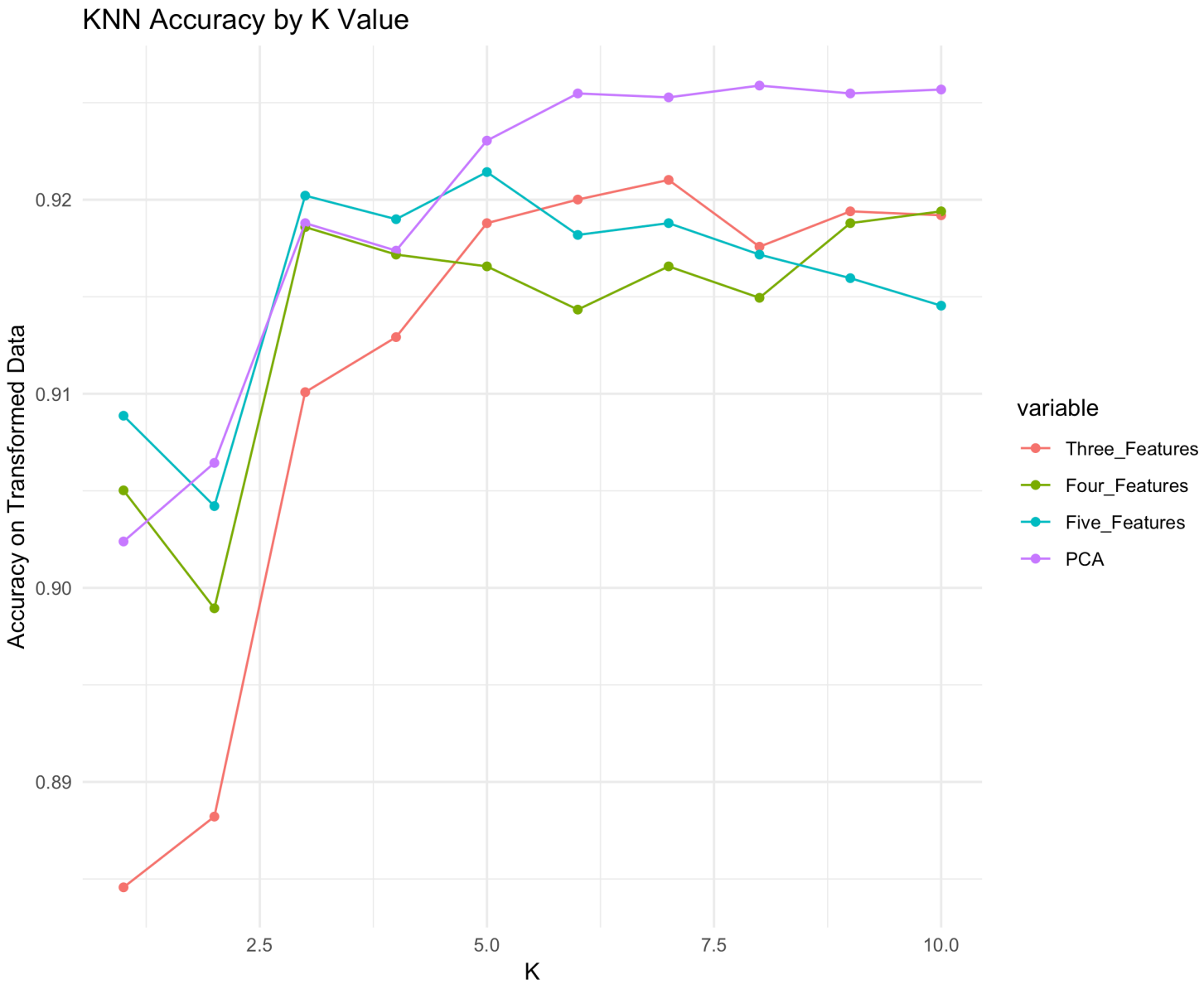
Model accuracies on the test set were performed using the split method that reported highest overall cross-validation accuracy.

4. Diagnostics

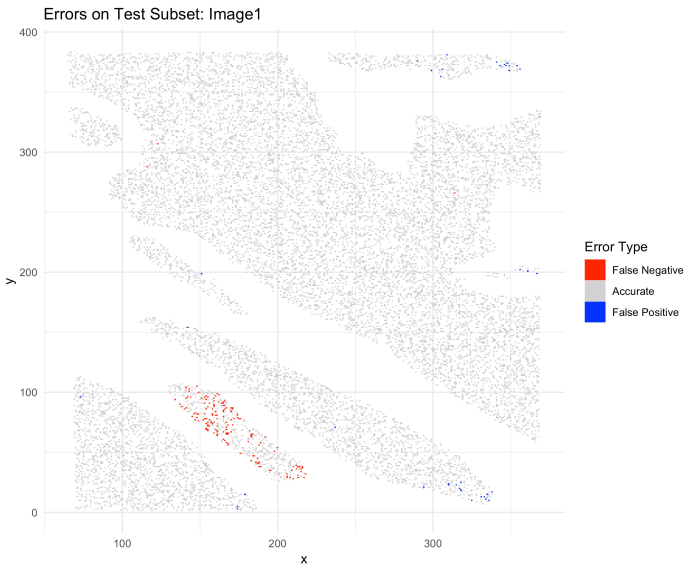
KNN outperforms the other classifiers and has a higher accuracy using the transformed data set. In part three, we only used NDAI, CORR, and SD to train our models, in this section we explore the KNN accuracy as K increases while adding more training features.

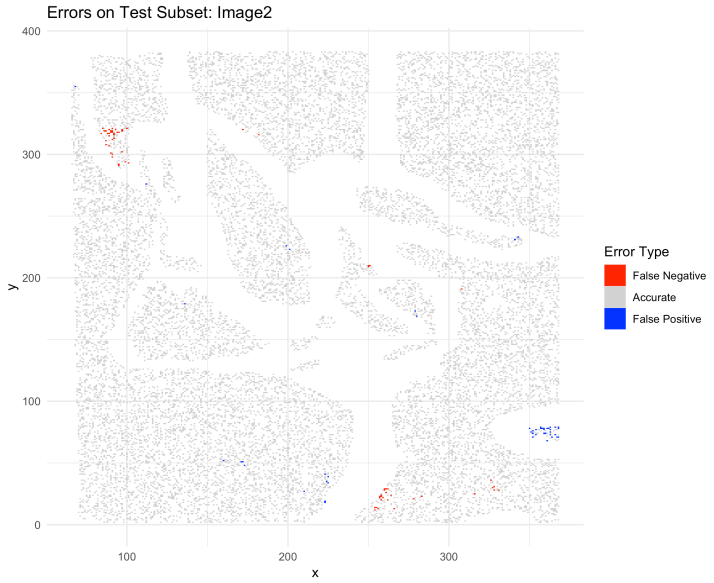
We see an increase in model accuracy when adding radiance angles as features, and a substantial increase in model accuracy when including the first principal component after running PCA on the radiance angles. In part two we deduced that PC1 of the radiance angles was not a good feature since the histograms of cloud label were largely overlapping. We see from this analysis however that it does improve the model accuracy for K-nearest neighbors significantly.

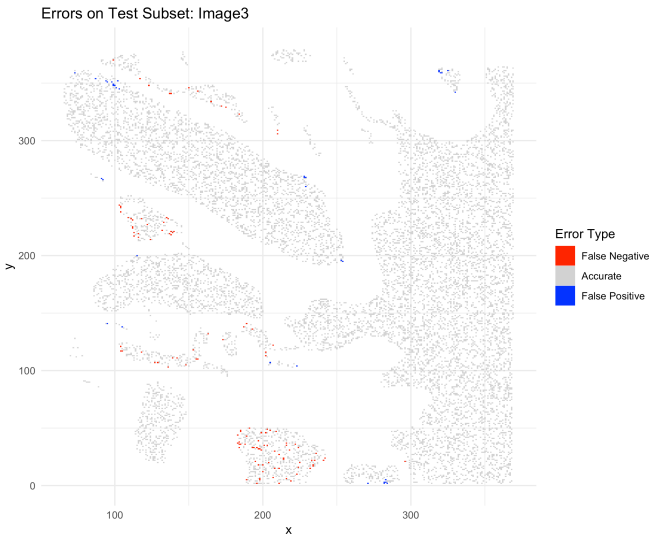
It is clear that highest model accuracy for KNN occurs for k=10, which is what we used throughout the project.



KNN Misclassification Errors:

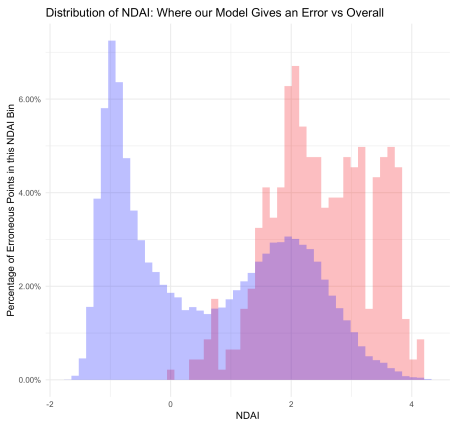


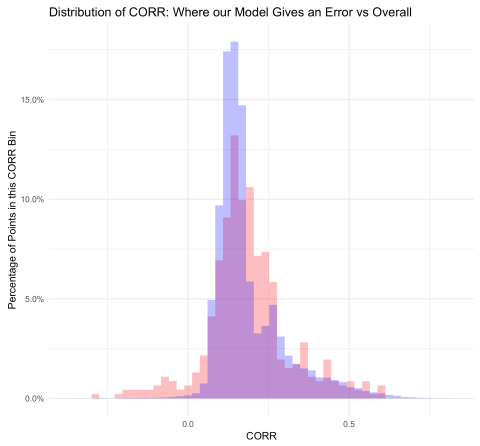


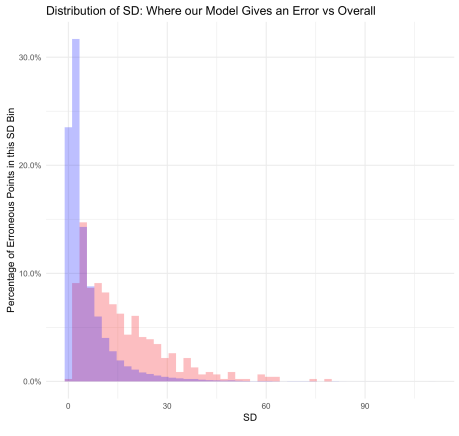


The vast majority of the area in our model for all three images is error free. It's clear from these images that errors are clustered in certain clouds or cloudless regions. Furthermore, those spatially clustered errors are of the same type(i.e. there are clouds that have relatively high rates of being mislabelled as no cloud, a few small cloudless regions with high rates of being mislabelled as a cloud).

It's worth considering what could be making our model perform worse with these particular regions of the data. Thus, we examine the relationship between errors and certain features.







The distribution of NDAI values is substantially different among values where our model misclassifies than in the general test set. It appears that our model fails at much higher rates with higher NDAI values. It would be valuable to consult with experts about the possible influence of NDAI on certain regions where our model has a relatively high failure rate. The distributions for SD and CORR are similar.

Boosting:

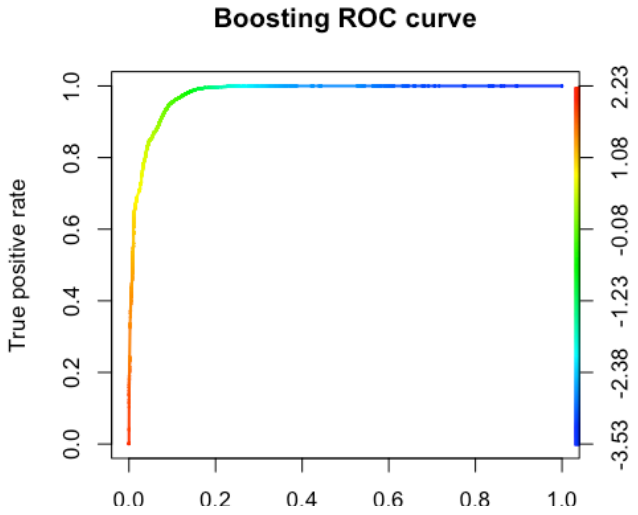
Boosted regression trees is one classifier we used to improve model accuracy. Boosting fits the decision trees on a modified version of the original data set, growing them sequentially, and almost recursively. This boosted approach learns slowly, and uses errors from previous trees when training the next ones. In general this approach should perform well, but we see in our data that is has a slightly lower test accuracy than KNN.

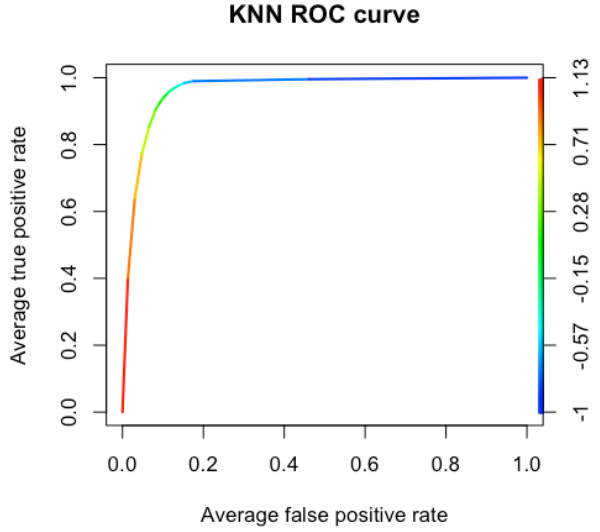
Boosting	1st method	2nd method
Fold 1	90.7%	89.5%
Fold 2	90.9%	89.7%
Fold 3	90.7%	89.7%
Fold 4	90.7%	89.9%
Fold 5	90.8%	90.4%
Test Set Accuracy	90.9%	

While this classifier was initially used to improve on the test accuracy of our models, we notice that the KNN test accuracy of 91.4% is higher than the Boosting test accuracy of 90.9%.

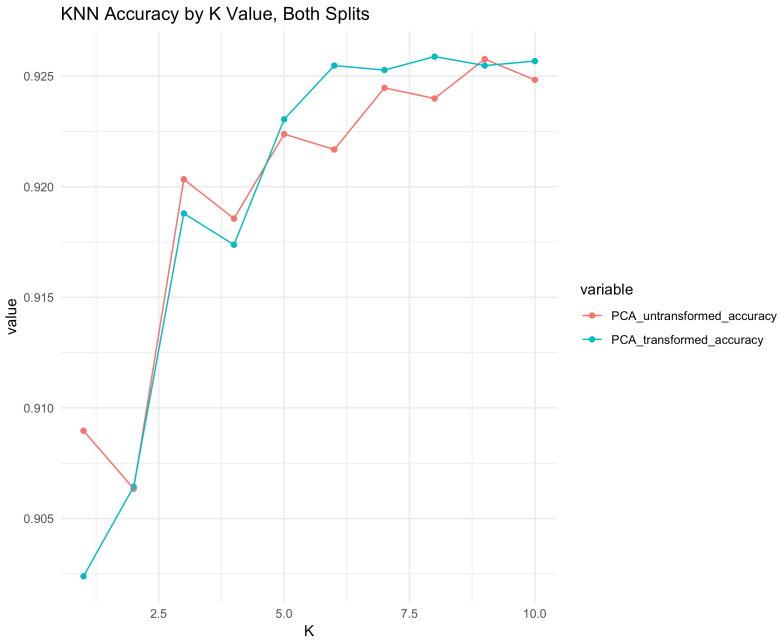
Both of these ROC curves illustrate the effectiveness and similarity of their predictions. Out of the five models used to classify cloud labels in this project, the only ones which reported higher accuracy on the untransformed data within 5-fold CV are KNN and Boosting.

It's possible that our model will work poorly on future data without expert labels, if there exist factors in future data that cause the cloudiness of a region to have a different relationship with our features, than it has in the three images we trained our model on. Ideally, we would train our model on more images as they do in the paper to account for possible discrepancies between images and their respective polar regions.





After discovering the improved accuracy from adding PC1 of radiance angles as a training feature for KNN, we check to see whether there is a noticeable difference between the transformed and untransformed data on model accuracy. We see the difference is slight, but the transformed data does outperform the untransformed data. Since KNN makes no assumptions about decision boundaries or the underlying properties of the data at hand, we expect the effect of reconciling the inherent dependence relation amongst the data to not make a big difference in model accuracy, as is seen in this plot.



Conclusion

In this report we train five separate classification models to predict cloud labels from satellite images in the arctic region. We present the power of classic statistical prediction, with deliberately chosen training features to fit our model. An analysis of the model accuracies was completed to compare the performance of each method. Lastly, we run several diagnostic procedures to asses how to improve our models, as well as explore misclassified results.

GitHub Repository link: https://github.com/phoebeabramowitz/154project2

5. Acknowledgments

Most of this project was completed while working together, often times at Yali's Cafe in Stanley Hall. Lots of questions were answered in office hours, as well as by reading through existing Piazza queries.

Contributions of each member:
1. a) Omri b) Phoebe c) Omri
2. a) Phoebe b) Omri c) Omri d) Phoebe
3. a) Phoebe Omri b) Omri Phoebe
4. a) Omri b) Phoebe c) Omri d) Omri e) Phoebe
5. Omri Phoebe

Resources used for this project include Raaz's office hours, Yuansi's office hours, professor Yu's office hours, ESL textbook, ISL textbook, and the following sites:

References

Chauduri, Kamalika. *Rates of Convergence for Nearest Neighbor Classification*. cseweb.ucsd.edu/~dasgupta/papers/nn-rates.pdf.

"Discriminant Analysis Tutorial ." *RPubs*, rpubs.com/ifn1411/LDA.

kahlokahlo 1, et al. "Multiple ROC Curves in One Plot ROCR." *Stack Overflow*, stackoverflow.com/questions/14085281/multiple-roc-curves-in-one-plot-rocr.

"Linear & Quadratic Discriminant Analysis." *Linear & Quadratic Discriminant Analysis · UC Business Analytics R Programming Guide*, uc-r.github.io/discriminant_analysis.

Linear Discriminant Analysis - Using Lda(). maths-people.anu.edu.au/~johnm/courses/mathdm/2008/pdf/r-exercisesVI.pdf.

"Linear Discriminant Analysis in R: An Introduction." *Displayr*, 14 Apr. 2019, www.displayr.com/linear-discriminant-analysis-in-r-an-introduction/.

lrocalroca 3141618, and Paul HiemstraPaul Hiemstra 48.9k10107134. "Linear Discriminant Analysis Plot Using ggplot2." *Stack Overflow*, stackoverflow.com/questions/20197106/linear-discriminant-analysis-plot-using-ggplot2.

spektraspektra 142249, et al. "How to Plot a ROC Curve for a Knn Model." *Stack Overflow*, stackoverflow.com/questions/11741599/how-to-plot-a-roc-curve-for-a-knn-model.

Thiagogm. "Computing and Visualizing LDA in R." *Thiago G. Martins*, 14 Jan. 2014, tgmstat.wordpress.com/2014/01/15/computing-and-visualizing-lda-in-r/.