Jilin Cao (3033278367)
Sizhuo (Cindy) Liu (3032082925)
Statistics 154
Spring 2019

Data Splitting Matters

Introduction

The purpose of this report is to conduct detailed data exploration and modeling of cloud detection based on data collected through MISR sensor. The primary goal of such analysis is to develop a classification model that can effectively distinguish the clouds and ice / snow surfaces using given features.

The paper consists of four major components: data collection and exploration, data cleaning, modeling, and diagnostics. In data collection the background and source of data was explained, and explorative data analysis was carried out. In data cleaning, the information was split into multiple sets and particular features were chosen for later analysis. In modeling, different classification methods were used and their fit was assessed through cross validation and other techniques. In the diagnostics section, the best classifier from the previous part was assessed for its accuracy in predicting the response variables.

1. Data Collection and Exploration

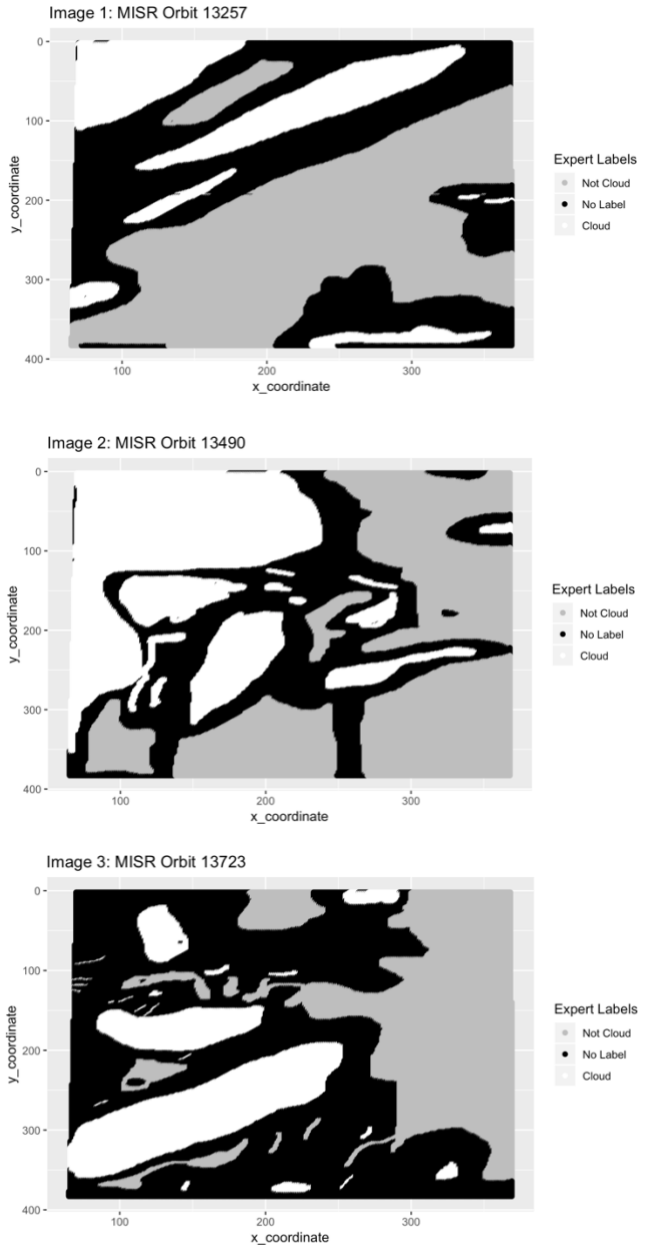
1.1 Summary of Paper

The purpose of this study is to develop a new operational Arctic cloud detection algorithm using Multiangle Imaging Spectroradiometer (MISR) imagery. In the contemporary world, the effect of the amount of atmospheric carbon dioxide is of great scientific interest to researchers. Arctic appears to be an ideal area for investigations in this field as it displays the strongest dependences of surface air temperatures on increasing atmospheric carbon dioxide levels. However, a major issue is that accurate Arctic-wide measurements are often difficult to obtain--it is hard to detect clouds over other surface types, such as liquid and ice-water cloud particles that have similar scattering properties. Therefore, it is crucial to develop a set of algorithms that accurately characterize the properties of clouds over the daylight Arctic.

The data used throughout the study were collected from 10 MISR orbits of path 26 over the Arctic, northern Greenland, and Baffin Bay through a time span of approximately 144 days. Each MISR pixel covers a 275 m x 275 m region on the ground. Six data units from each orbit are included in the observations. In addition to the key parameters such as correlation, NDAI index, standard deviation, and information from the six data units, the dataset also includes a column of expert label values, which represents the experts' hand-labeled data indicating whether a particular patch of image pixels signify cloudy or clear scenes: all of which that are based on highly-confident domain knowledge were marked as clear or cloudy and those with ambiguity were unlabeled.

The data also suggests that the ELCM algorithm is more accurate and provides more comprehensive spatial information than the existing MISR operational algorithm. The conclusions from this research is significant in that it provides a better understanding of the polar cloud properties and its potential effect to the changes in the Arctic region brought about by increasing concentrations of atmospheric carbon dioxide. Another key

aspect of this research is that it shows the power of statistical thinking and the key role that statistics play in contributing to modern scientific problems.



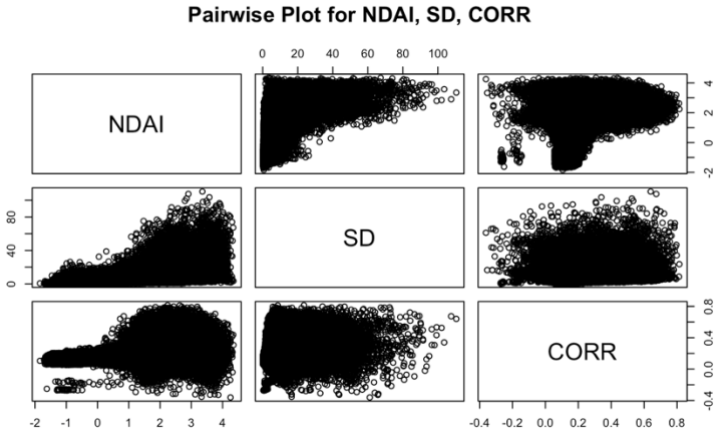
1.2 Data Summary

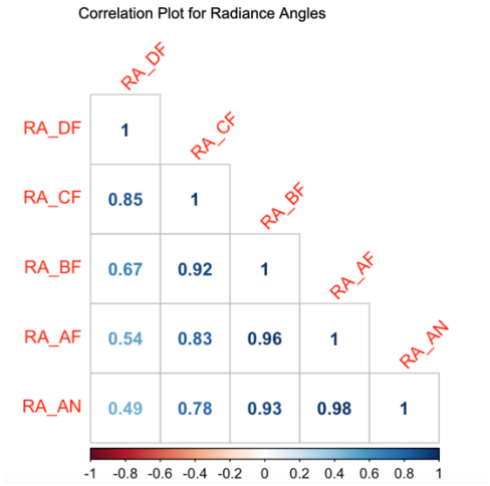
When exploring the relationships between different variables, we purposely combined rows of datasets from all three images so that we could observe the overall relationships across all images. In addition, we stored a data frame in which we omitted all data with an expert label of zero (no information) as they, at this stage, does not contribute to the investigation. It will also make the classification process later easier as most of the classification methods used only support binary response variables. The original data is still stored under a different name just in case it is needed in later analysis.

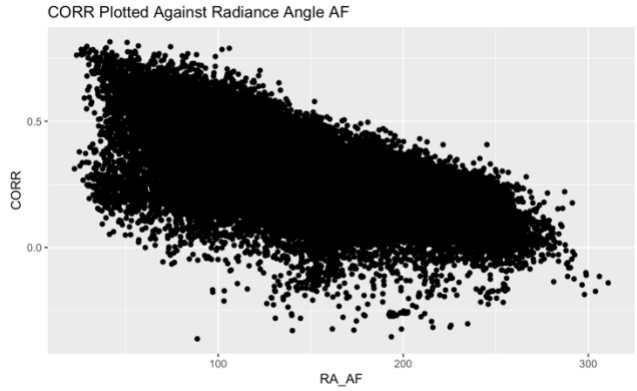
As shown by the plots of the three images on the left, the region of ice/snow surfaces and cloudy areas are clustered without obvious patterns. For this reason, we deduce that there are spatial correlations between pixels. Therefore, the independent and identical distribution assumption is not valid in this dataset.

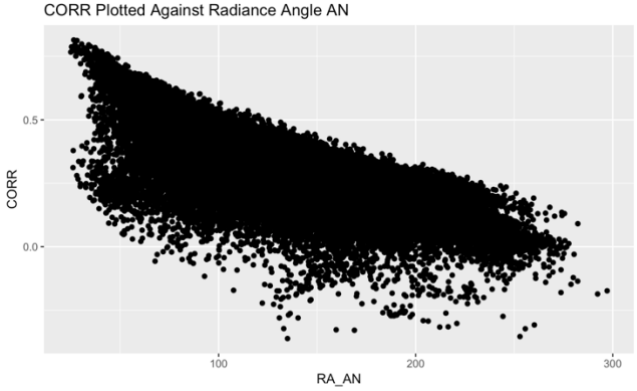
The top left figure on the following page displays the pairwise plots for NDAI, SD, and CORR, which allows us to observe the pairwise relationship between the three variables. However, no significant pattern was found between the features except for NDAI plotted against SD, which seems to suggest a wider range of SD values as the value of NDAI index increases. Nevertheless, such correlation does not appear to be very strong. Such observations are similar across all three images. The top right figure on the following page shows the correlation values between different radiance angles and all values are highly correlated with each other. This may be because the data for the radiance angles are collected at around the same region with the same set of tools, just with slightly different angles. Relationships between NDAI, CORR, and the different radiance angle were observed by plotting the angles against NDAI or CORR. It was discovered that when CORR was plotted against the radiance angles AF, AN, and BF, there exists a negative,

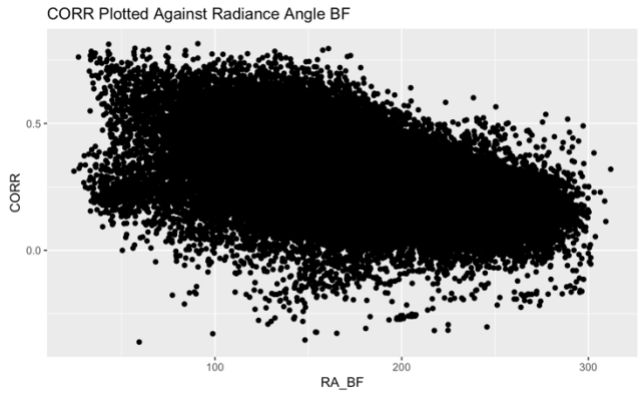
approximately linear relationship, as displayed in the three figures in the following page. This confirms the calculation of CORR used AF, AN, and BF in Bin's paper (2008).











1.3 Data Exploration

Histograms displaying the distributions of SD, CORR, NDAI, and a selection of radiance angles were also plotted, color-coded by the expert label. It is obvious from such histograms that the distributions of SD and CORR for both categories (cloudy and clear)

are approximately the same, while the count for clear is significantly higher than the count for cloudy. The histogram for NDAI (shown on the right above) for all three images suggest that the lower range of NDAI index corresponds to a high number of "clear" labels, while all the "cloudy" labels were made at higher NDAI indices: the distributions for the two labels are relatively distinct from each other. This indicates that NDAI may be a good feature for predicting if a particular part of the images is cloudy or not. The other histograms, on the other hand, show that the distributions of the cloudy and clear labeled data have very similar distributions. The abundance of overlap suggest that it can be difficult to distinguish the two labels using the features.

2. Data Cleaning and Preparation

2.1 Data Splitting

In order to build and evaluate the model, we decided to separate data for each image into training, validation, and test set independently and then combine the sets for each image.

Taking the spatial relationship between adjacent pixels into account, we know that the data points are not independent and identically distributed. A naive splitting approach might cause the training dataset to consist of all cloudy or clear pixels. This will lead to a misrepresentation of the original data and thus negative affect the model performance. For example, if our training dataset has 99 out of 100 pixels as cloudy, then a trivial classifier will get 99% training accuracy. In order to keep the balance in our three datasets as well as preserve local spatial correlation, we proposed two ways of splitting to avoid the issues. In the first method, we partition each image into n x n rectangular boxes, and then, viewing each box as an independent entity, we randomly sample training, validation and testing set among the boxes.

The second method is similar to the first one but slightly different. Instead of choosing an entire box, we randomly sample from the data within each box into training, validation and testing sets. Then, we combine all training, validation and testing from each box into our final training, validation and testing sets. The second method may damage spatial relationship but can still manage to ensure a balance. It can also serve as a comparison against the first method. Our conjecture at this point is that our first method can do better as it also preserves the spatial relationships. However, such conjecture can only be assessed in the later modeling and diagnostics sections.

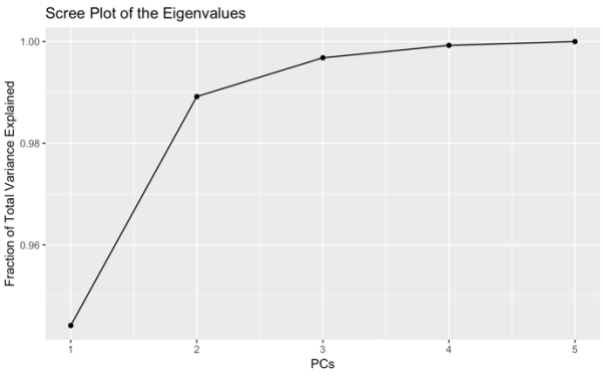
2.2 Baseline Accuracy

A trivial classifier that automatically sets all labels to be -1 was created. The accuracies of such classifier on the validation set and test set were then calculated in comparison to the actual expert label: they are 60.85% and 60.82%, respectively. This sets a minimal requirement for our research by implying that any model created should obtain an accuracy that is at least as high as the baseline accuracy, as such baseline model is the most basic and simple model one can possibly create.

2.3 Variable Selection

To investigate on the predictive power of the raw features, we implemented PCA on all five radiance angle features. As shown by the scree plot below, the first three PCs

capture about 99.7% of the overall variance, which means the first three features are almost capable of explaining the entire dataset. Therefore, we decided to use the first three principal components as our features.



To see if our choice is reasonable, we performed a quick and simple logistic regression and tested it on validation set. We got an accuracy around 84.5%, which is higher than SDCM, ASCM, and Offline SVM methods mentioned in Bin's paper (shown on the left). The coverage is also 100%.

However, this is not as powerful as sophisticated features: NDAI, CORR, and SD in Bin's paper as these three features produce even higher accuracies across all folds. Therefore, the modeling and diagnostics sections will be implemented using these three features instead.

Table 1. Agreement (accuracy) rates relative to expert labels and coverages of the ELCM, MISR operational SDCM and ASCM, and offline SVM algorithms

	ELCM	SDCM	ASCM	Offline SVM
Agreement with expert	91.80%	80.00%	83.23%	80.99%
Coverage	100%	26.64%	70.12%	100%

3. Data Modeling

3.1 Classification Methods and Fit Assessment

Several classification methods were implemented, and their fit was assessed using cross-validation.

The logistic regression model assumes that the predictor variables are all independent from each other, so that the problem of multicollinearity will not arise. From the pairwise relationship analysis from an earlier section, it was observed that the five radiance angles are highly correlated because their values were measured in very similar ways. However, from the variable selection section, it was concluded that NDAI, SD, and CORR do not display any obvious pattern or trend with each other. Since the actual model used only involves these three features, the assumption of non-multicollinearity is satisfied.

The LDA model makes two crucial assumptions: that the data under each label is multivariate normal and has the same variance. We can make a simple check if it is Gaussian by plotting a histogram each individual feature in the model. The histograms indicate that none of the features display a bell curve that is characteristic of a normal distribution. To check the variances across different features, two covariance matrices for the three variables were created, one for a set that contains only "cloudy" data, and another including only "clear" data. The final matrices are clearly different from each other. Therefore, neither of the assumptions of LDA are satisfied in our case.

The QDA model holds the same assumptions as LDA except that it does not require the variables to have equal variance. Since neither of the assumptions for LDA was fulfilled, so is the case for QDA.

SVM is very computationally expensive and so we decided not to implement for our model.

The k-nearest neighbor algorithm, decision tree, and random forest are all nonparametric methods. Therefore, their models make no assumptions over the data.

The accuracies of both methods of creating folds are as the following:

Logistic Model Cross_Validation			
Split Method 1		Split Method 2	
Folds	Accuracy	Folds	Accuracy
1	0.89138049	1	0.89320796
2	0.89069368	2	0.89064849
3	0.88901099	3	0.893173
4	0.89209794	4	0.89274602
5	0.89076236	5	0.89260994

QDA Model Cross_Validation			
Split Method 1		Split Method 2	
Folds	Accuracy	Folds	Accuracy
1	0.89668542	1	0.89781864
2	0.89619437	2	0.8954029
3	0.89839756	3	0.89698326
4	0.89723442	4	0.89532932
5	0.89791717	5	0.89810543

LDA Model Cross_Validation			
Split Method 1		Split Method 2	
Folds	Accuracy	Folds	Accuracy
1	0.89637661	1	0.89667226
2	0.89771129	2	0.89851779
3	0.89822948	3	0.89564761
4	0.89671265	4	0.89861912
5	0.89671265	5	0.89722436

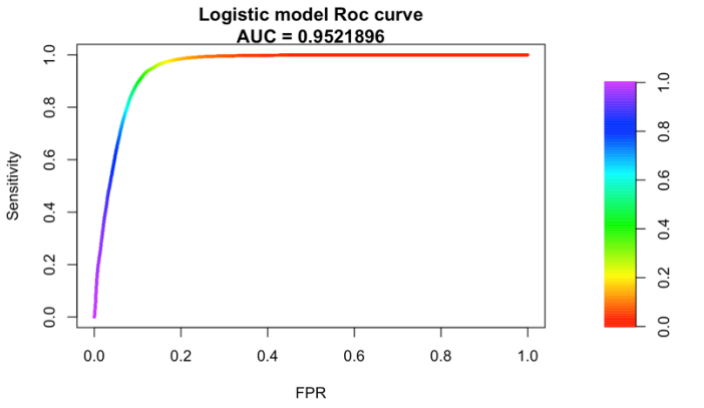
KNN Model Cross_Validation			
Split Method 1		Split Method 2	
Folds	Accuracy	Folds	Accuracy
1	0.9128809	1	0.9146712
2	0.9154857	2	0.913232
3	0.9129465	3	0.9143506
4	0.9135612	4	0.9117698
5	0.9120544	5	0.9121163

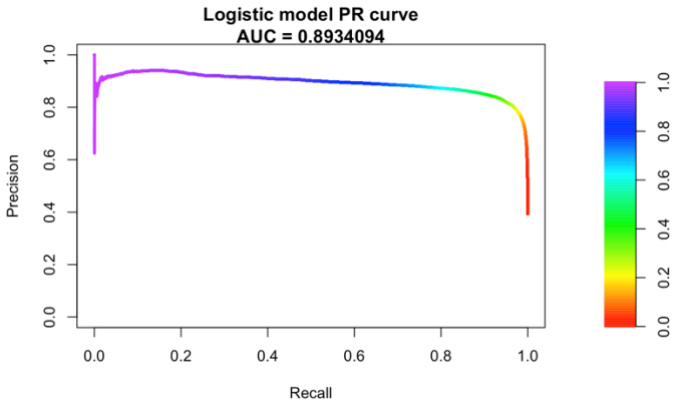
Decision Tree Model Cross_Validation			
Split Method 1		Split Method 2	
Folds	Accuracy	Folds	Accuracy
1	0.88721134	1	0.88823632
2	0.88586919	2	0.88470547
3	0.88532803	3	0.88817422
4	0.88758879	4	0.88960358
5	0.88827506	5	0.88342014

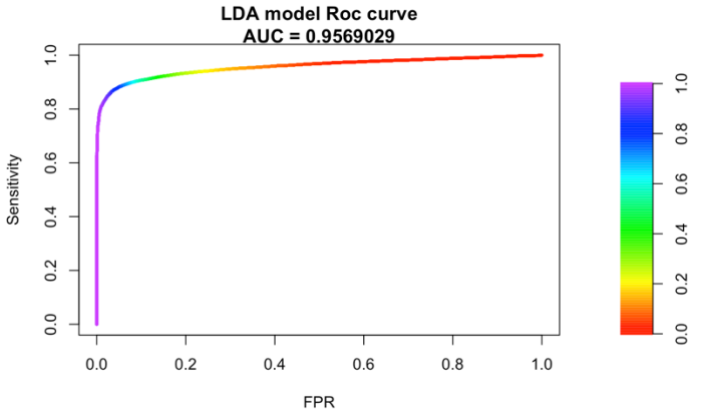
The tables above indicate that for all classification methods implemented, the accuracy is relatively consistent between different folds. Among all methods used, the KNN algorithm's output has slightly higher accuracy than the others. The two different splitting methods did not seem to have a significant effect on the accuracies.

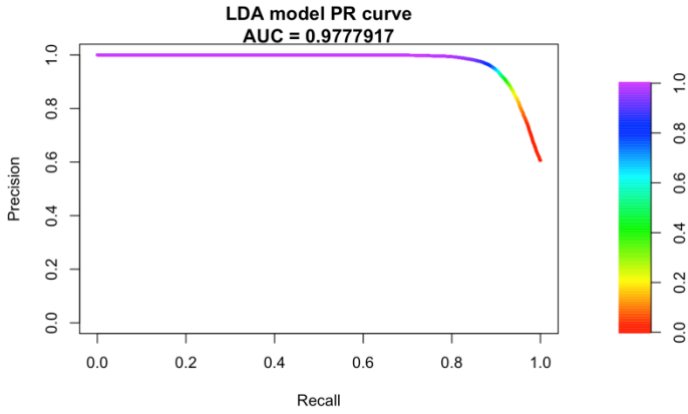
3.2 ROC curves

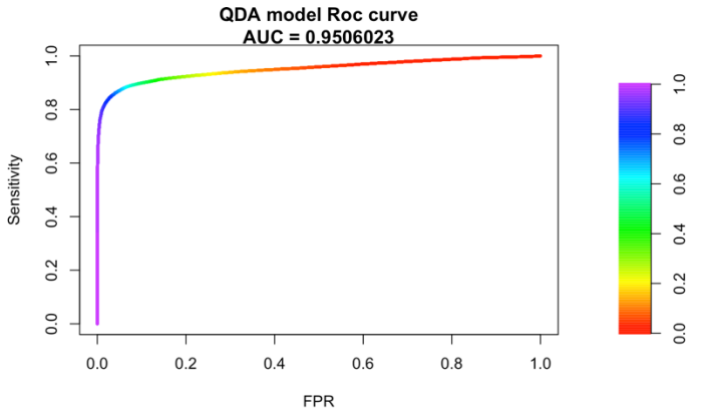
ROC curves for all fitted models are plotted in the following page (left panel). The x-axis of ROC curve is false positive rate, which is the type 1 error. The y-axis is the true positive rate, also known as sensitivity or recall. Sometimes we may want to choose the cutoff value depending on our tasks, but in general we want to pick a cutoff value for which we can get the highest true positive rate and lowest false positive rate. For example, in the plot "Logistic Model ROC curve", we would like to choose a cutoff that is to the top left, which is shown in green at approximately 0.5. Therefore, this value is a good choice. Following the same thought process, 0.9 is chosen as the cutoff for KNN and 0.6 is chosen for decision tree. For LDA and QDA, 0.8 is chosen as the cutoff value. Such value is reasonable because involves MLE calculation, which usually seeks for large probability as a cut off value.

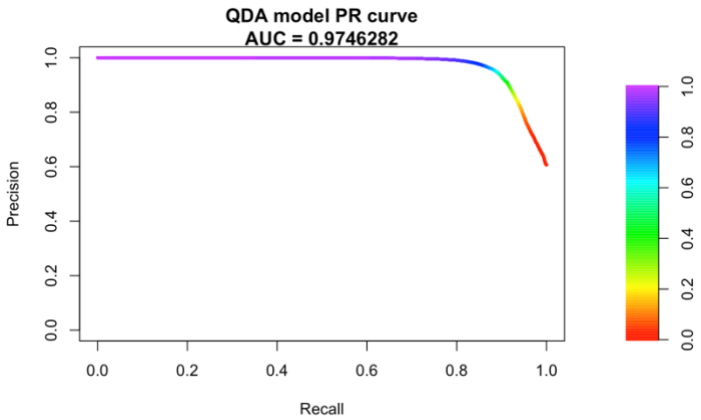

Logistic model Roc curve
AUC = 0.9521896
Sensitivity
FPR

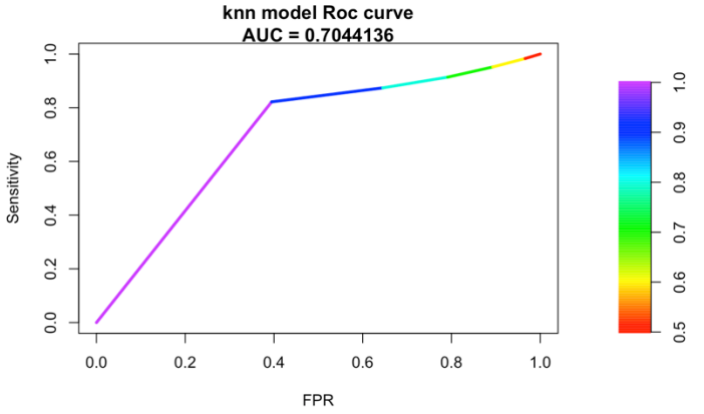

Logistic model PR curve
AUC = 0.8934094
Precision
Recall

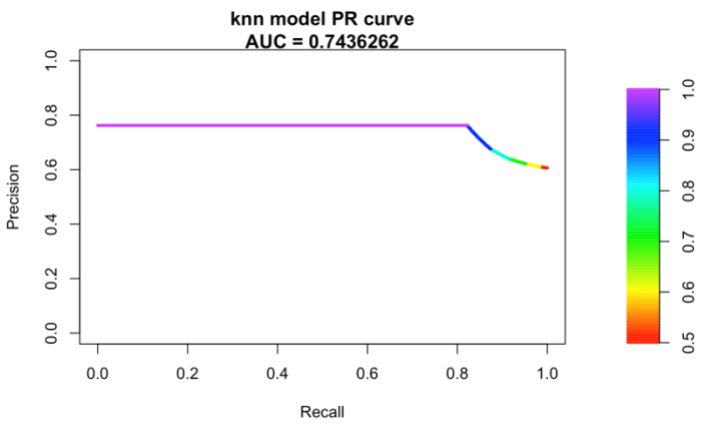

LDA model Roc curve
AUC = 0.9569029
Sensitivity
FPR

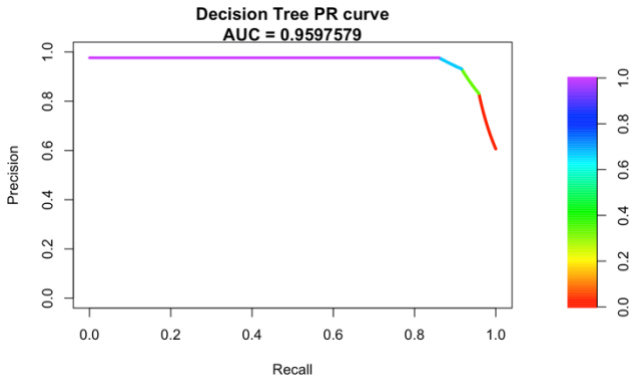

LDA model PR curve
AUC = 0.9777917
Precision
Recall

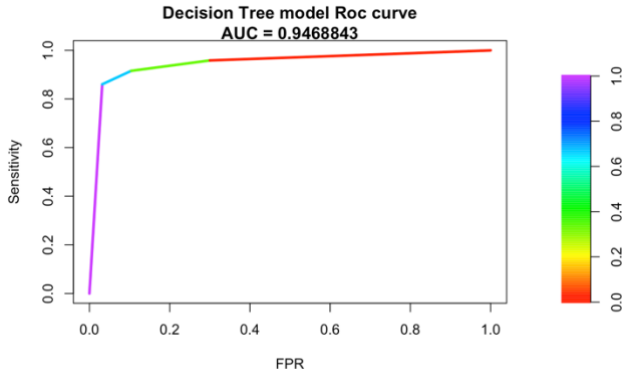

QDA model Roc curve
AUC = 0.9506023
Sensitivity
FPR


QDA model PR curve
AUC = 0.9746282
Precision
Recall


knn model Roc curve
AUC = 0.7044136
Sensitivity
FPR


knn model PR curve
AUC = 0.7436262
Precision
Recall





3.3 Precision Recall curves

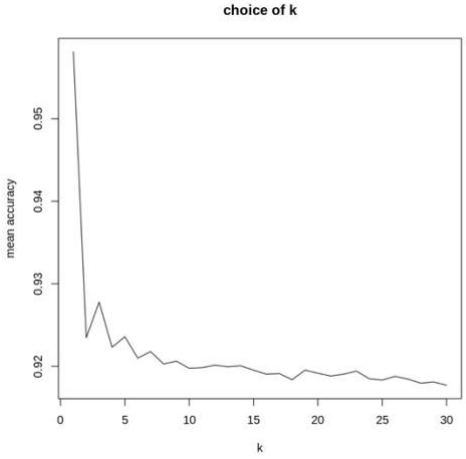
As another way to assess the fit of the models, we also plotted precision recall curves for all models that were implemented. The ideas behind these plots are similar to ROC curves. The x-axis is precision, while the y-axis is recall. Their formulas are displayed as follows:

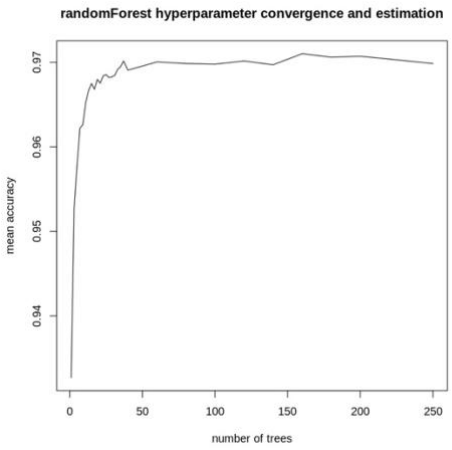
$$\text{Precision} = \frac{TruePositive}{TruePositive + FalsePositive} \qquad \text{Recall} = \frac{TruePositive}{TruePositive + FalseNegative}$$

Suppose that 1 is the label to be predicted, which in this case represents a cloudy region. The precision informs us of the proportion of true values among the pixels that are predicted to be cloudy. Recall, on the other hand, states the proportion of pixels that are predicted to be cloudy among all pixels with an expert label of "cloudy" from the original dataset. For a model to be considered descent, it should have both high accuracy and high coverage. Therefore, we aim to pick a cut off value that produces us the highest precision and recall. Therefore, the cutoff value should be chosen at a region close to the top right corner in the plot, as that is the area where both x and y values are high. Based on the PR curves above (right panel), we observed similar cutoffs as in ROC curves. This is reasonable as ROC and PR involves similar concepts.

4. Diagnostics

4.1 Analysis of Chosen Classification Model





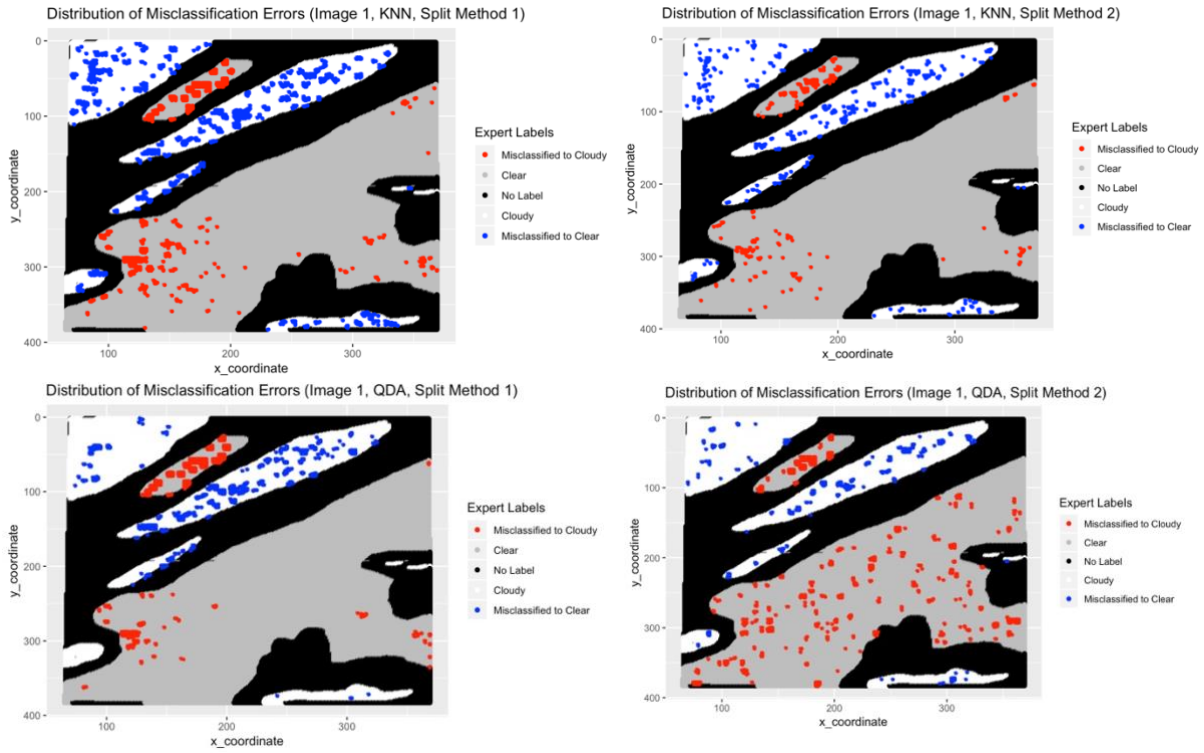
From the accuracy tables from the previous section, it is obvious that the KNN method produces the highest cross-validation accuracies across all folds. Therefore, we count this as the most ideal model for now. In the previous section, the hyperparameter k was picked as 10 as an initial guess. In this section, we calculate for the k value that outputs the highest accuracies. In order to do this, we tried to fit different models with different k values. From k = 1 to k = 30, for each k, we build a model and report the accuracy on validation dataset. As shown in the left graph above, surprisingly, k = 1 produces the best validation accuracy. The plot also indicates that there exists a dramatic fall in mean accuracy rate from k = 1 to k = 2. After this, the rate of fall of accuracy decreases.

4.2 Analysis of Misclassification Errors

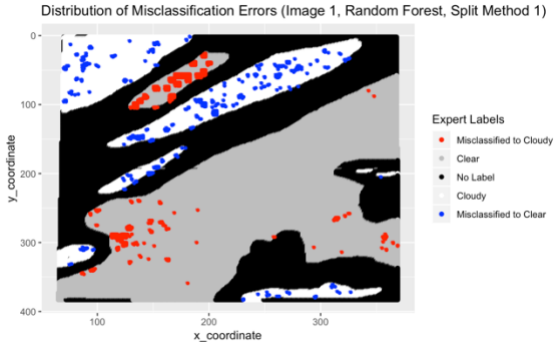
In order to better visualize the physical distribution of the misclassified points in each image and to look for potential problems in the model, we plotted the original well-labeled map shown at the start as the background and plotted the misclassified points on top of the map. The points were also color-coded based on the type of misclassified errors. Such plots were created for the KNN model, the QDA model, and the Random Forest model, as these three are considered the best models that produce the highest accuracies.

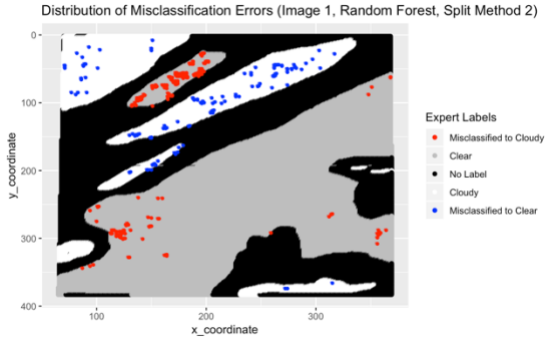
The plots for KNN models indicate that for both splitting methods, it is difficult for the model to distinguish cloudy and clear regions in areas where the clouds and ice/snow surfaces are clustered (one surrounded by the other). This is probably because the KNN algorithm looks the nearest data points. In this case, the pixels that are very close to each have high similarities. Thus, it is observed that KNN did a good job distinguishing the two labels in the large uniform area but predicted poorly on the borderline.

The distribution of misclassification errors from the QDA regression does not appear to have a strong pattern: the misclassified points seem to be uniformly distributed. This may be due to the multivariate normal assumption of QDA.



By the comparison of the two plots for KNN, there does not seem to appear a strong difference in the distribution of misclassified points across the two splitting methods. However, the plots for QDA does display a difference in pattern due to the different splits.





4.3 Potential Improvement in Classification Model

Based on the graphs from the previous section, the model did not perform well in some particular areas of the images: it fails to distinguish the cloudy and clear areas especially close to the borderline, and in areas where cloudy and clear regions are clustered. Therefore, an ensemble learning idea was proposed.

As shown by the modeling accuracy tables, a single decision tree has high variance and is not very stable. Random forests, however, solves such problem by training multiple different trees with randomly sampled subsets of the data (bagging), and subsets of the features to de-correlate the trees. In summary, the overarching idea is that a group of weaker learners are combined with the hope that some weaker learners could specialize in difficult-to-distinguish areas and their knowledge could be utilized in these particular regions. Since Random Forest is doing bagging on data, it is similar to feeding into new data. A new model using random forest was created. The ideal value of the hyperparameter was calculated in the same way that the k value for KNN was found (the plot is shown above). After having tested this new model on the validation and testing datasets, we achieved an accuracy of 97.35%, which is considerably higher than the previous record. The plots for the misclassification errors above further supports the conjecture that Random Forest is an improvement to the previous classifiers as the number of misclassified points has clearly decreased. In addition, the difference in splitting method does not lead to significantly different outcomes.

4.4 Effect of Data Splitting Methods

The random forest model accuracy was compared between the two ways of splitting data and there appears to be significance difference. Split method 1 produces an accuracy rate of 97.45% while the other is only 91.6%, though both were based on their corresponding testing datasets. Their respective confusion matrices are the following:

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 24753   360
         1   715 15870
```

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 23147  1161
         1  2376 15480
```

The first splitting method produces a sensitivity value of 0.9719 and specificity value of 0.9778. The second outputs 0.9069 and 0.9302 for sensitivity and specificity, respectively. Since we aim to maximize these criteria above, splitting method 1 is better than splitting method 2. In addition, method 1 satisfied the assumption that it preserves local spatial correlations.

4.5 Conclusion

Based on the analysis in the previous sections, we concluded that while the choice of the classifier, the features used to build the classification model, and the choice of hyperparameters can all influence the final prediction accuracy of the model itself, the initial assumptions of the original data and the way in which the raw data is separated may also significantly affect the performance of the models.

5. References

Brownlee, Jason. "Linear Discriminant Analysis for Machine Learning." *Machine Learning Mastery*, 22 Sept. 2016, machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning/.

Katz, Mitchell H. "Assumptions of Multiple Linear Regression, Multiple Logistic Regression, and Proportional Hazards Analysis." *Multivariable Analysis*, pp. 38–67., doi:10.1017/cbo9780511811692.006.

Shi, Tao, et al. "Daytime Arctic Cloud Detection Based on Multi-Angle Satellite Data With Case Studies." *Journal of the American Statistical Association*, vol. 103, no. 482, 2008, pp. 584–593., doi:10.1198/016214507000001283.

6. Acknowledgement

Jilin Cao: Read research paper and took notes; brainstormed ideas for EDA; carried out data cleaning; proofread final paper
Cindy Liu: Brainstormed ideas for EDA; Organized formatting of final paper; carried out data cleaning; proofread final paper

Thanks to Raaz for answering questions relating to the project during Office Hours and Lab sessions.