Lab 4 - Detecting Clouds Using Satellite Data

11/09/2018

1 Introduction

As the Earth's climate changes, interest in atmospheric carbon dioxide continues to grow. Research has shown that the strongest dependency of climate change on atmospheric carbon dioxide will take place in the Arctic. In order to study this dependence more in-depth, accurate measurements of cloud coverage in the Arctic are needed, as clouds help modulate sensitivity to increasing surface air temperatures in the Arctic. However, because of the prominence of ice- and snow-covered surfaces, it is difficult for existing technologies to detect clouds. This poses a challenge, as identifying clouds over the Arctic is necessary for assessing their impact, whether positive or negative, on electromagnetic radiation flow through the atmosphere.

2 Satellite Data

2.1 Data Collection and Cleaning

In order to address this challenge, Shi et al. [1] collected and analyzed atmospheric data from the Multiangle Imaging SpectroRadiometer (MISR) on the NASA Terra satellite, which made available electromagnetic radiation measurements at nine view angles in four spectral bands. The nine angles include 70.5 (DF), 60.0 (CF), 45.6 (BF), 26.1 (AF), and 0.0 (AN) degrees. In addition to the radiation measurements at these five angles, data used in this study consisted of three features chosen after extensive exploratory data analysis by the authors. These three features are the correlation (CORR) of MISR images of the same scene from different viewing angles, the standard deviation (SD) of the nadir camera pixel values across a scene, and a normalized difference angular index (NDAI) that characterizes the changes in a scene with changes in the view direction.

Of the 233 geographically distinct MISR paths, this study utilized data collected from 10 orbits over the Arctic, northern Greenland, and Baffin Bay. The 10 orbits span about 144 days during daylight season in the Arctic in 2002. Each orbit had 6 data units, though 3 of the 60 were dropped, leaving 57 total for this study. The resulting data consists of over 7 million pixels, where each pixel covers a 1.1km $\times$ 1.1km region on the ground and has 36 radiation measurements. In order to evaluate performance of the classification methods used in the study, one of the authors hand-labeled clear versus cloudy pixels. This was only done for the pixels they were highly confident of labeling, as this was the best known method for obtaining validation data. Only about 71.5% of the total number of valid pixels were labeled, and these labels will be used here to validate our classification models built for images of three data units.

2.2 Exploratory Data Analysis

Exploratory data analysis was conducted through data visualization and numerical summaries. First we look at the true cloud labels for each of the three images, and then we visualize the features available for use in building classification models.

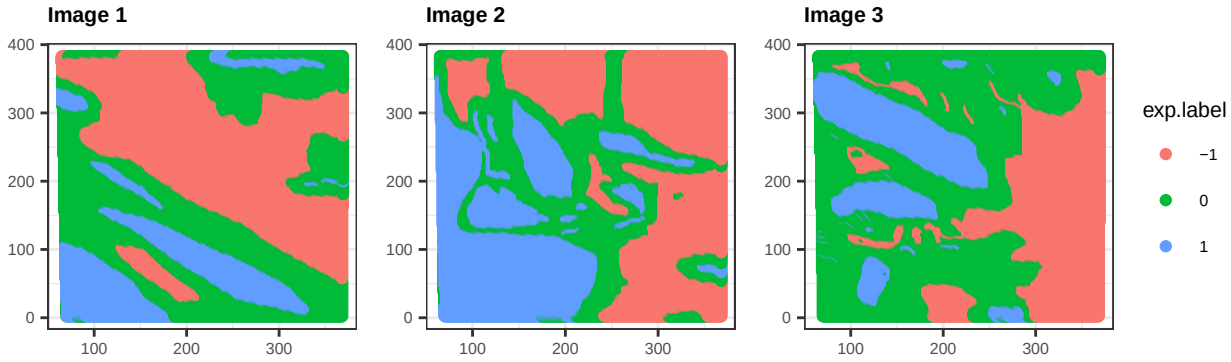

Figure 1: Plot of expert labels for each image from the raw satellite data.

In Figure 1, expert labels are shown for each image. Red indicates a negative label, where there are no clouds. Blue indicates a positive label, where clouds are present. Green indicates an area where experts were unable to determine the presence of clouds, leaving these points unlabeled.

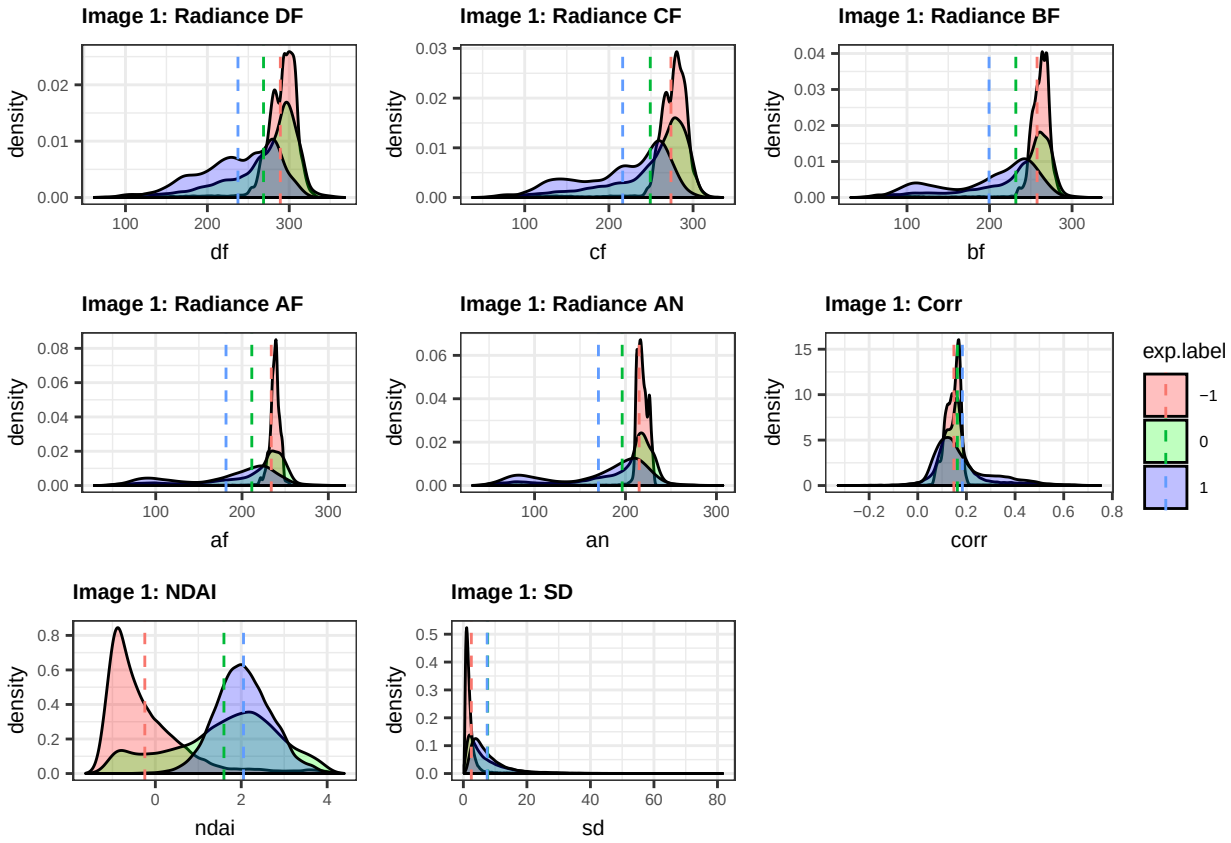

Figure 2: KDE plots of each feature in image 1, with mean as a dotted line. Other images followed similar trends, so we only display image 1.

In Figure 2, kernel density estimation (KDE) plots are shown for each of the eight features available in the data set. There is clear separation in much of the data between positive and negative labels for clouds, while the unlabeled and more challenging data points appear to have a distribution that lies in between positive and negative labels. There appears to be less variance in the negative labels when compared with the positive labels, which may be useful in predicting unlabeled points.

3 Feature Selection

3.1 Visual Analysis

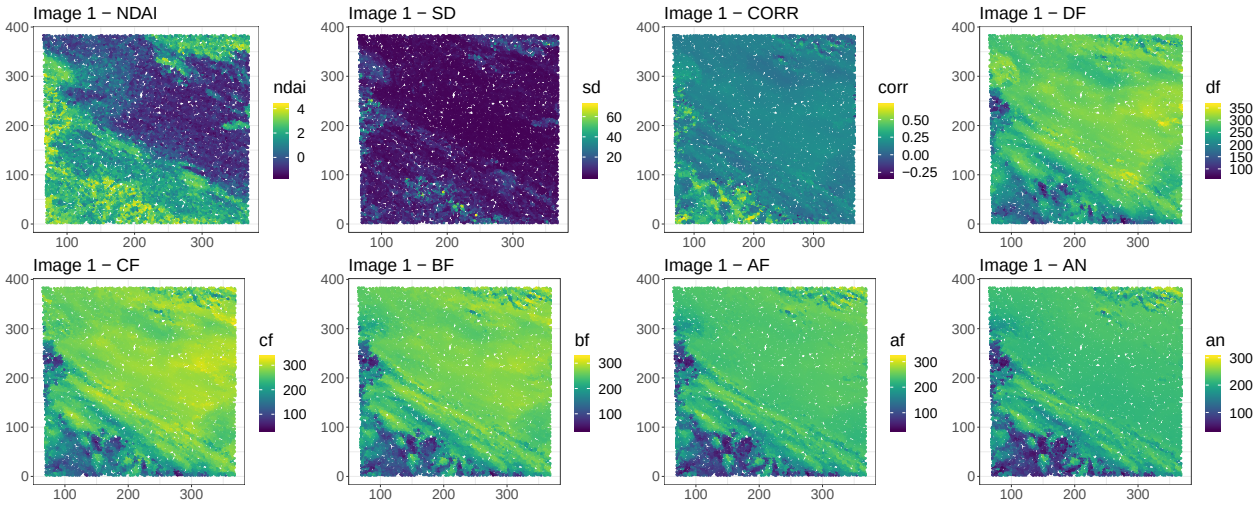

Figure 3: Visualization of each features of image 1, with subsampled data of proportion m = 0.2 for ease of visualization.

In order to get a better sense of what variables to choose, we first map the features and radiance measures in Figure 3. We only plot image 1 here, but other images had similar visualizations. We also subsampled 20% of the data to show instead of using all of the data, since using all data points cluttered our figures unnecessarily. The plots suggest that the mappings of NDAI, CORR, and the radiances show similar patterns as the expert label mapping. However, we also note that the five radiance measurements do not seem too distinct from each other, which means that we need to be careful of multicollinearity in the columns. We will further analyze this condition in the next section where we check model assumptions. With the visualization in mind, it seems that NDAI, CORR, and one of the radiance measurements will provide us the best summarizing features to perform regression on.

3.2 Quantitative Analysis

For quantitative analysis, we perform PCA on the three images. We do not include the label 0 data in our analysis, since it made virtually no difference in the principal components we obtained. We also visualize the data that has been standardized, as standardization helped in our case to differentiate the features better.

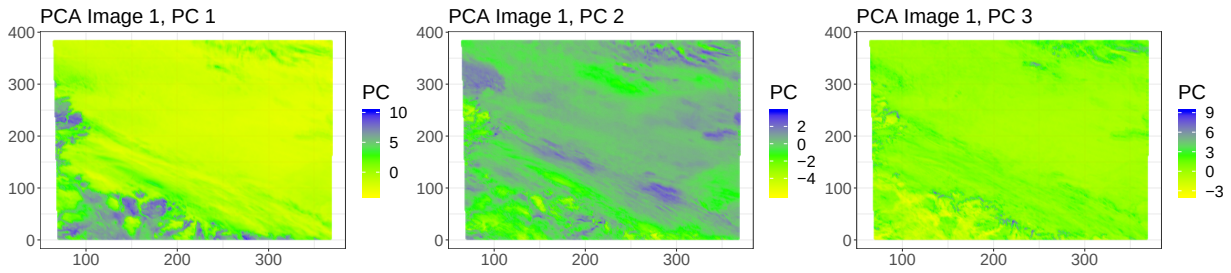

Figure 4: Projection of image 1 to the three largest principal components.

In Figure 4, we plot the projections of image 1 onto the three most prominent principal components. The three principal components each pick out different features of the image and we see clear separation between the different parts of the images. We chose not to plot the other images, since they also had similar levels of desirable separation.

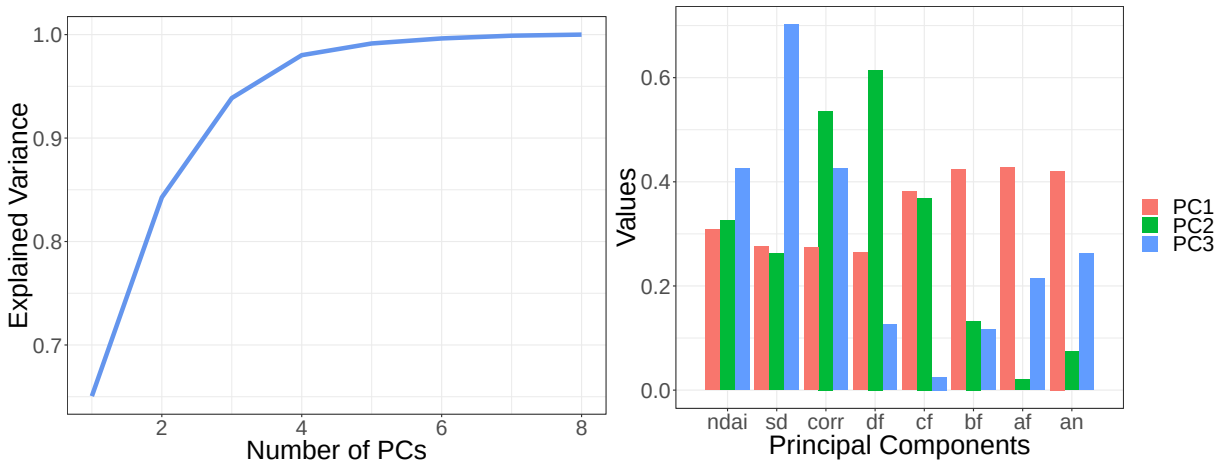

Figure 5: Amount of variance explained by the PCs (left), bar plot of the amplitudes of principal component elements for PCs 1 through 3 (right).

In the left subfigure of Figure 5, we plot the amount of variance explained by the principal components (PCs). We observe that using the top three PCs already captures around 95% of the variance in the data, so we will focus on using these three PCs for feature selection.

In the right subfigure of Figure 5, we plot the amplitudes of the principal components we obtain with ± 1 labeled data that has been standardized. The first PC picks up most of the radiances, and the second and third PCs picks up the other three features. The radiance measurement DF seemed to be the most prominent across the three PCs compared to all the other radiance measurements, and we have seen in the visual analysis that all the radiance measurements look very similar. Thus, we will only use radiance DF, which captures most of the dynamics out of all 5 radiance features. Amongst the three original features, it looked like NDAI and CORR seemed to be more prominent in both second and third PCs. Thus, we choose DF, NDAI, and CORR as the three best predictors to fit our prediction models.

4 Comparison of Classifiers

Before we fit the three classifiers of our choice for this project - random forest, support vector machine (SVM), and logistic regression - we must thoroughly check the model assumptions before we can confidently use them for our classification schema.

4.1 Logistic Regression

The more obvious assumptions of logistic regression include binary categorical output variable, and linear relationship between the regressors and the outcome. Our data satisifes the binary output assumption, and we think linear modeling of the regressors to predict the output is relatively sensible given the visual and quantitative analyses of the features. Below, we check the multicollinearity and no influential values assumptions, which are not as immediate when scrutinizing the data.

4.1.1 Multicollinearity

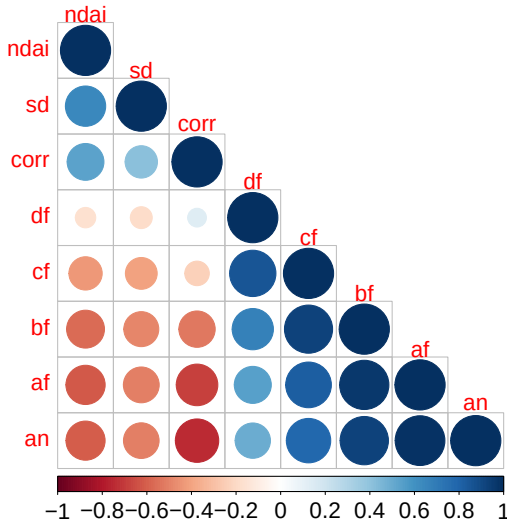

Figure 6: Plot of pairwise correlation between variables, where size and color indicate the amount of correlation.

First, we check the existence of multicollinearity in the observed data in Figure 6. While the three features NDAI, SD, and CORR do not seem to be as correlated with each other, we see a huge correlation trend within the radiance measurements. Amongst all of them, it seems like DF was the least correlated to all the other measurements and features, which justifies our choice of using DF as one of our features, on top of NDAI and CORR. If we were to include the other radiance measures in logistic regression, we might see it perform very poorly or have extremely low stability due to multicollinearity. This verifies our usage of NDAI, CORR, and DF, as these three features are not as correlated with each other.

4.1.2 No Influential Values

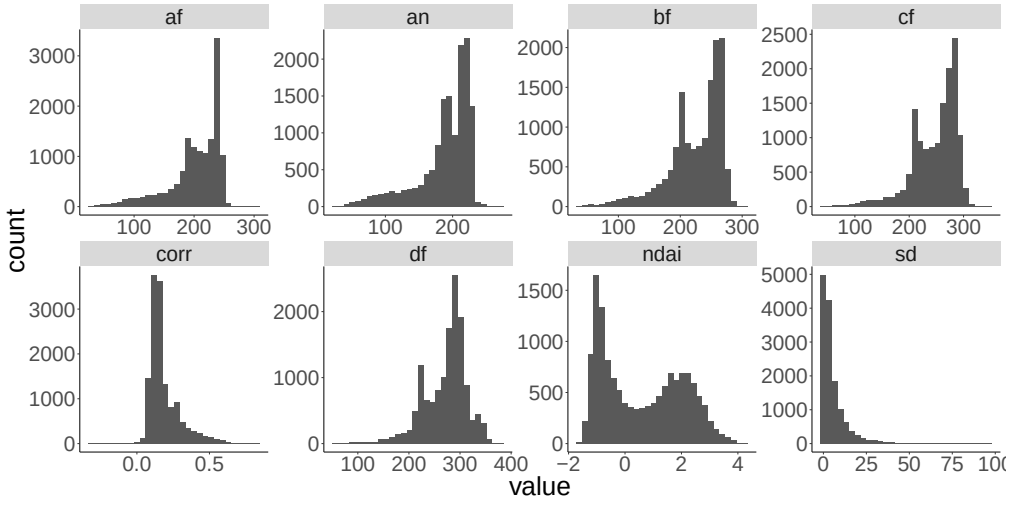

Figure 7: Histogram of all features for the image data.

In Figure 7, we check the assumption of having no influential values, which often refer to outliers that bias the features. As we have seen through other analyses, the radiance measures share a similar pattern, although no visible influential values can be seen. Likewise, NDAI and CORR seem to contain minimal influential values, which once again justifies our usage of the three chosen features. However, we note that SD seems to have a nontrivial amount of influential values, which supports our decision to not include it in our further analyses.

4.2 Support Vector Machine (SVM)

The main implicit assumption of SVM is that there exists a separating hyperplane between the two classes of binary outputs, give or take some slack variables to allow for partial cross-overs. In fact, the usual way check this is to actually run SVM on the dataset and see how much separation we get, which is a rather circular argument in our case since we are using SVM as one of our classifiers. However, in our analyses, the SVM performed just as well as the other classifiers, which suggests that there is a viable separation of the two classes.

4.3 Random Forest

The three assumptions of random forest are given as following. First, observations are not autocorrelated over time or space. We might slightly violate this assumption of having no autocorrelation over space since each image data are taken from the same satellite at once, but it is relatively negligible since we only have three image files collected through the same protocol over multiple days, and we have around 300,000 data points total. Other assumptions include observations not being drawn from designed experiment and no matched case and control samples, which our data satisfies.

4.4 Classifier Evaluations

To evaluate and compare the performances of these classifiers, we look at the prediction results and accuracies on each methods. In this analysis, we will fit random forest, SVM, and logistic regression with our three best features, and with all 8 given features.

First, we compare the cross validation (CV) prediction accuracies of each methods. We initially split the data up into 70% training and test set for CV and fitting, and 30% final validation set that we do not touch until the final fitting. All the images have been aggregated into one dataset, and the subsets were chosen at random. We chose k-fold to be $k = 5$. This value of k is empirically found to be one of the most reliable values in creating balanced folds, which minimize the bias-variance tradeoff. Furthermore, the fold was chosen uniformly across all three images to create unbiased separation of data. Indeed, we found $k = 5$ to reliably create folds with unbiased subsamples that could be used for training and validation.

In the table below, we enumerate the CV accuracies of the fits. We see that in general, using all the features performs slightly better than using only three features, but not by a big margin. In fact, the gain in predictive accuracy is not very substantial, only accounting for around 1-2% boost. This suggests that using the three best predictors will actually result in simpler and more interpretable model fits since there are less variables to keep track of. Among the three model fits, random forest performed the best.

Table 1: Mean and SD of CV accuracies

	Random Forest	(RF SD)	SVM	(SVM SD)	Logistic Regression	(LR SD)
Best Features	0.9309814	0.0019961	0.9275346	0.0016358	0.8877522	0.0024316
All Features	0.9589198	0.0012726	0.9461213	0.0009158	0.8924418	0.0016170

In Figure 8, we show the resulting predictions from the 6 different fitting methods on the subset of 30%

final validation set that belong to image 2. We choose to show image 2, since image 2 contained the most interesting and visible differences within the image, but other images also followed similar trends. Expectedly, the predictions do not significantly differ from using top three features versus using all features. In fact, there isn't a noticeable difference between all the fitting results, but random forest classification seemed to have the least misclassification errors compared to other prediction methods.

Table 2: Final validation accuracy

	Random Forest	SVM	Logistic Regression
Best Features	0.9328570	0.9294606	0.8892485
All Features	0.9610055	0.9473398	0.8924847

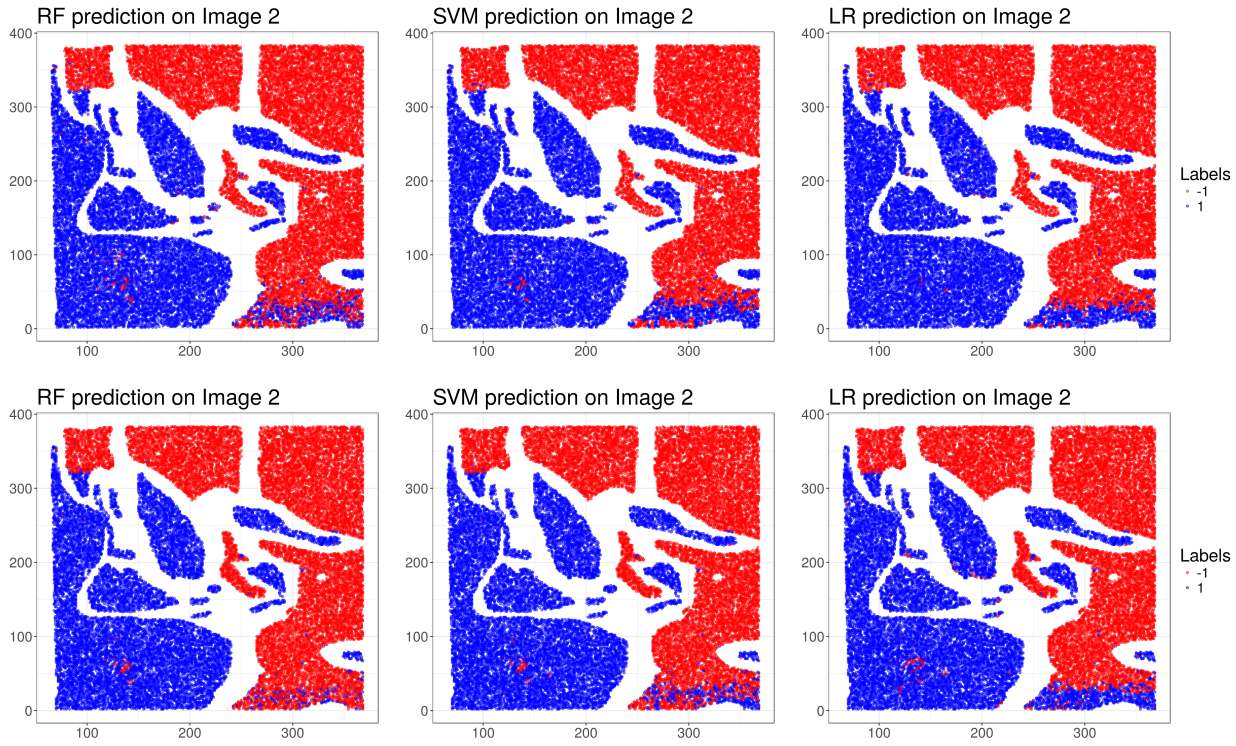

Figure 8: Fitting using our chosen features (top subfigures), and all 8 given features (bottom subfigures).

Finally, we take a look at the ROC curve for each of the methods. In Figure 9, we plot the ROC curves of each methods and provide AUC for each of these curves in the table below. Random forest once again dominates the other methods. Also, using all features do not give us a huge performance boost when looking at the ROC either, which means our choice of three best features is well justified for the data set at hand, and we will be using NDAI, CORR, and DF throughout our final model analysis.

Table 3: AUC of all the ROC curves

	Random Forest	SVM	Logistic Regression
Best Features	0.9830990	0.9779957	0.9521757
All Features	0.9931959	0.9869229	0.9554163

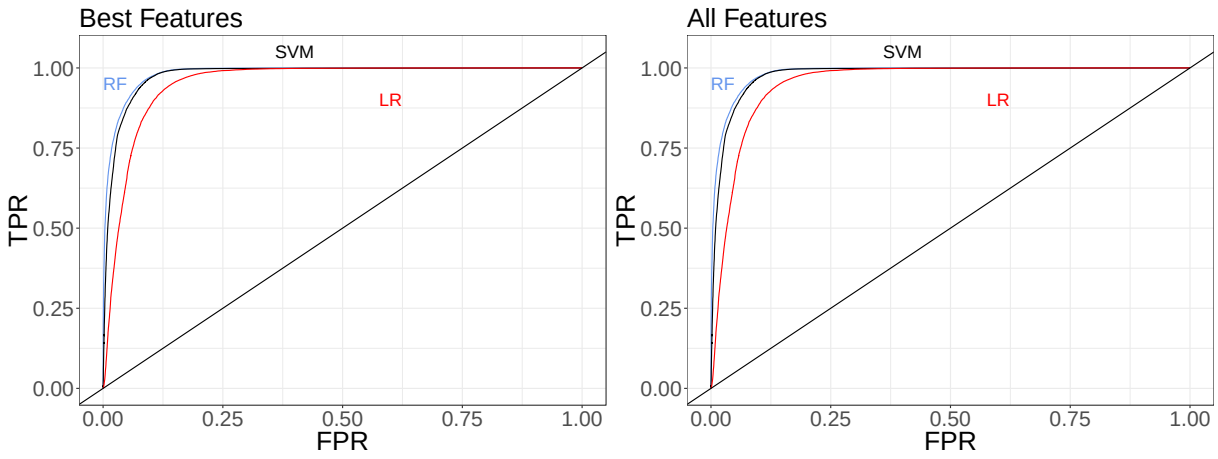

Figure 9: ROC curve comparison for our choice of features (left) and all features (right).

5 Model Diagnostics

5.1 Final Model Choice

With all the considerations of model assumption checks, feature visualizations, prediction accuracy and ROC checks, we decided to use random forest fitted with three features NDAI, CORR, and radiance DF as our final model. With this final model, we evaluate its performance on our dataset. First, we plot all the predictions that have been generated by using our model of choice in Figure 10. We see that the prediction results look very similar to the original expert labels, although some regions do suffer from misclassification. These regions tend to be on the boundary of the unknown labels and known labels. Among all the images, image 3 suffered the most from image classification results, which may suggest that image 3 might have been recorded under slightly different settings compared to the other two images. However, the prediction accuracy seems high enough across all three images.

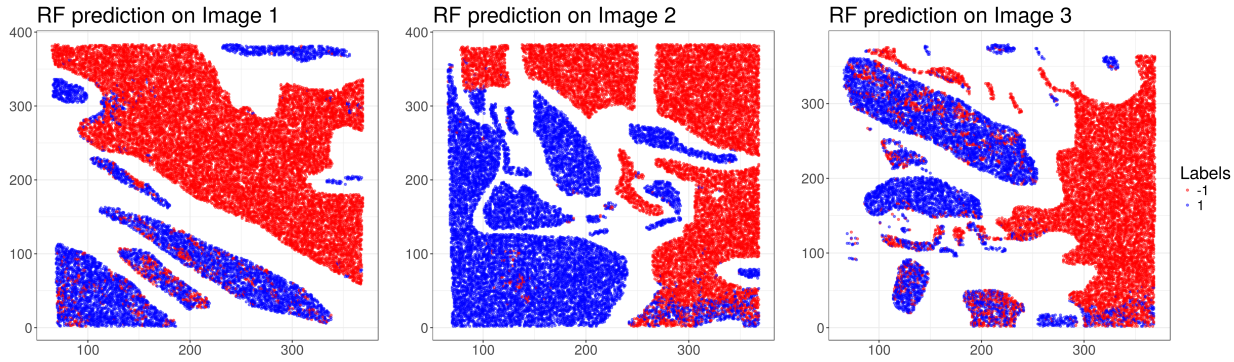

Figure 10: Prediction results of our random forest model with three chosen features on our final validation set.

We now look at the stability of parameters. While random forest does not have parameters to fit, it has variable importance that we can check. In the table below, we show the stability of the importance measurements from subsampling analysis we performed. We ran subsampling with portion $m = 0.2$ that we train the random forest model on, and recorded the importance of each variables for $N = 100$ iterations. We see that most of

the feature importance values are relatively stable, with NDAI being the most predominant feature compared to the two other features.

Table 4: Variable importance summary for $m = 0.2$ subsampling with $N = 100$ iterations. MDA refers to MeanDecreaseAccuracy and MDG MeanDecreaseGini.

Feature	-1	(-1 SD)	+1	(+1 SD)	MDA	(MDA SD)	MDG	(MDG SD)
corr	56.39895	3.235873	121.28256	8.028778	78.01042	4.262293	6415.160	159.53385
df	75.01272	5.321299	110.55192	6.635718	93.44625	6.282199	3169.219	56.48843
ndai	744.93688	32.733628	82.33686	8.751305	218.95306	16.953652	10197.999	172.80912

5.2 Misclassification Patterns

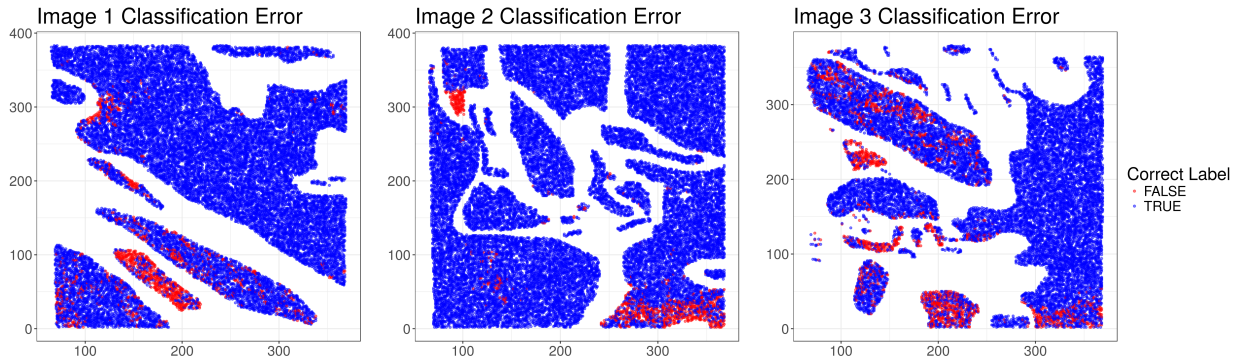

Figure 11: Misclassification patterns of our random forest model with three chosen features on our validation set.

Misclassification errors appear in distinct regions of the images. In Figure 11, we look at the misclassification patterns of our final model on each image. False positives appear around the edges of the non-cloud and unlabeled regions. Meanwhile, false negatives appear in the smaller cloud regions that border closely with unlabeled regions on all sides. We see that there is a small chunk of points with false positive predictions in the top left corner. This region is surrounded on all sides by unlabeled points, and it is therefore likely similar in features to the those difficult-to-label points. In addition, the bottom right corner of image 2 has a region of many misclassified points, which again are surrounding a region of unlabeled points. Overall, however, the random forest predictions had a low classification error, and the figure above shows that the majority of the labels distributed throughout the three images were correct, indicating that this model is effective for identifying clouds within our data set.

5.2.1 Confusion Matrix

Here we look at the confusion matrix to compare misclassification error in the two different labels. We also consider specificity and sensitivity as measures of efficacy of our classification model.

Table 5: Confusion matrix of final fit

	-1	1
-1	35614	1615
1	2576	22614

```
##   Sensitivity
## 1   0.9333443
```

```
##   Specificity
## 1   0.9325478
```

The random forest model used here had an overall accuracy of 93%. In the confusion matrix above, we can see that the misclassification error is similar for false positives and false negatives. There does not appear to be an imbalance in mislabeling points. Sensitivity (true positive rate) of 0.933 and specificity (true negative rate) of 0.932 are both indicative of a fairly strong model. The random forest applied here could predict new cloud patterns with a low error rate, but we need more validation that this is actually the case.

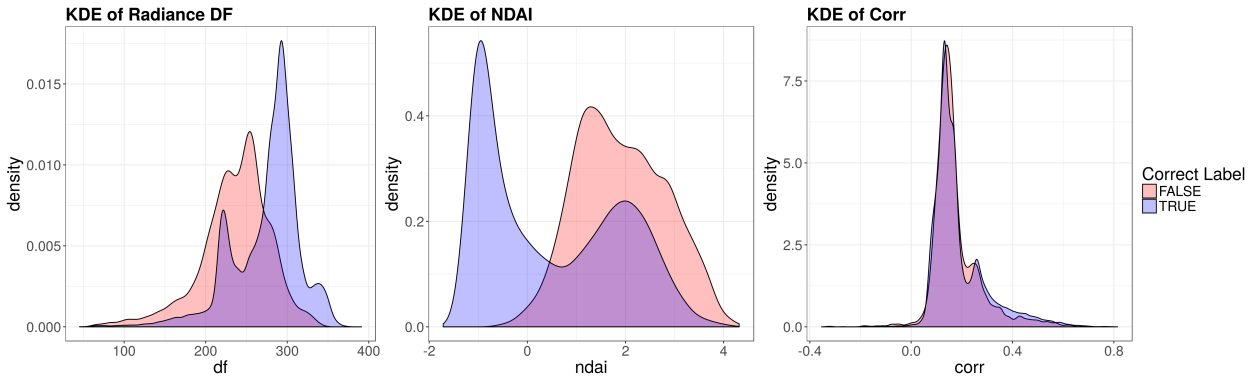

Figure 12: KDE plots of the classified points with our random forest model for our three chosen features, tested on our validation set.

Of the three features used in the model, we can see that the distribution of correctly classified pixels is significantly different from the distribution of misclassified points for two of those features. For radiance DF and NDAI, there appears to be a bimodal distribution amongst the correctly labeled points. Meanwhile, the misclassified points have more of a unimodal distribution, centered around one of the two modes of the correctly classified points. Lower values of the radiance DF and higher values of NDAI are apparently more likely to be misclassified. The CORR feature does not appear to have any particular pattern of misclassification error.

5.3 Future Performance Prediction

Finally, we look at the future performance of our model when given a completely new satellite image data that's been collected using the same protocol, except without expert labels. To get an idea of how our final model will perform, we withold an image to use for test, and train on the other two images, for all three possible permutations. In Figure 13, we show the prediction performance of our three feature random forest model under the selective witholding of each image. Somewhat expectedly, we see that our classifier usually does well when we treat images 1 and 2 as previously unknown, but it performs terribly for image 3. It seems like the structure of the data is different for image 3 compared to images 1 and 2, as we have noticed earlier in the final model choice where image 3 suffered the most from classification results. Thus, we can infer that if we obtain other images that are collected in different settings compared to images 1, 2, and 3, we could expect a drastic decrease in predictive performance due to new unforeseen trends of the features.

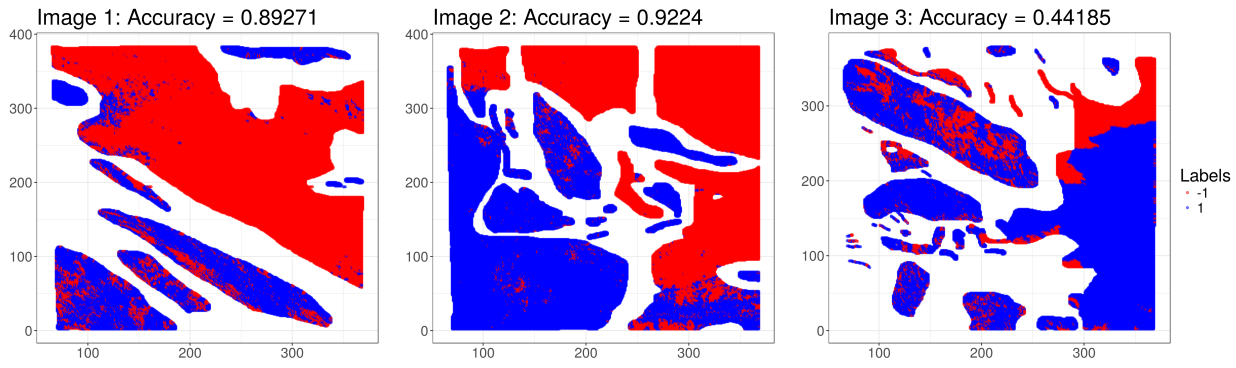

Figure 13: Prediction performance of our model on a simulated unlabeled new data, where we test on a witheld image, and train on the rest.

6 Conclusion

Through our study of satellite image data collected by Shi et al. [1], we have determined the three best representative features to use. We furthermore tested several classifier models, making sure the model assumptions are approximately met, which could be candidates for predicting the expert labels on cloud data. While our final model with our chosen features performed relatively well within our data set, analyzing unlabeled data prediction performance seems to suggest that we must be aware of the superpopulation of the data we collect. Specifically, the satellite images may not be taken under similar conditions, which may drastically affect the predictive performance of our models, or even any other models. Therefore, while being careful about making further generalizations to unseen data, we can still optimize both the predictive performance and the simplicity and interpretability of our models to predict the climate trends that have huge impact on our ecosystem.

7 Bibliography

[1] Shi, T. Yu, B., Clothiaux, E. and Braverman, A. (2008). "Daytime Arctic Cloud Detection Based on Multi-Angle Satellite Data with Case Studies." Journal of the American Statistical Association 103(482): 584-593.