Lab 4-Binary Classifier
Stat 215A, Fall 2014

Andrew Do, Hye Soo Choi, Jonathan Fischer, Xiang (Lisha) Li

November 11, 2014

1 Introduction

1.1 Data

Based on the recordings from the NASA's Multi-angle Imaging SpectroRadiometer (MISR) imagery, the data used to train the classification models are a collection of measurements from the three different images provided. Data units for these images were provided from MISR measurements over three consecutive orbits. These data were used to construct general classification models. The true classification of pixels is taken to be provided by the expert labels, though some pixels remained unclassified.

MISR collected electromagnetic radiation measurements using nine cameras at nine different angles, each of which views the Earth in four spectral bands (blue, green, red, and near-infrared). Each MISR pixel encompasses a 275m×275m region, yielding tremendous amounts of data. Due to the size of the MISR readings and transmission channel constraints, only the red radiances were transmitted at full resolution from all cameras. It was also stated that all four bands have similar reflectance signatures over ice, snow, and clouds. Therefore, only the red radiances which have high spatial resolution were used in constructing three features SD, CORR, and NDAI. In our data, in addition to the three features, five of the nine angles' radiances were given.

1.2 Features

The nine cameras had different views at the following angles: 70.5°(Df), 60.0°(Cf), 45.6°(Bf), and 26.1°(Af) in the forward direction; 0.0°(An) in the nadir direction and 26.1°(Aa), 45.6°(Ba), 60.0°(Ca), and 70.5°(Da) in the aft direction, where 'f' indicated the forward direction, and 'a' indicates the aft direction. Note that we are only given radiances for angles Df, Cf, Bf, An, Af. The nadir (vertically downward) viewing camera (An) has a strong advantage over cameras from other viewpoints since it is least influenced by atmospheric scattering and least distorted by surface topographic effects. Data derived from the Af and Aa cameras were aimed at determining topographic heights and cloud heights by using parallax from these two view angles. The Bf and Ba cameras are positioned to utilize aerosol sensitivity. Based on theoretical findings that the directionally-oriented reflectance variation over different types of cloud is minimized at a 60.0 degree angle, the Cf and Ca cameras are positioned at a 60.0 degree angle, which is also considered as crucial due to its significant role in estimating the amount of reflection at each ground point. The high angle view of Df and Da cameras is designed to present maximal sensitivity to off-nadir effects.

The feature CORR is an average linear correlation of radiation measurements at different view angles while SD is the standard deviation within groups of MISR angle An camera red radiation measurements. NDAI is the ratio between the difference and sum of the mean radiation measurements from the first and fifth angle associated with a particular pixel region. Extensive exploratory data analysis combined with specific domain knowledge, such as the fact that ice and snow surfaces scatter radiation more isotropically than clouds, motivated the development of NDAI, SD, and CORR. High values of CORR correlate with a lack of clouds or low altitude clouds. SD is designed to aid identifying smooth surface where the correlation between

different MISR viewpoints are muted by measurement noise. Larger NDAI values imply the presence of clouds. For further details on these features, see the paper by Shi, Yu, Clothiaux, and Braverman [].

2 EDA

2.1 Plots of Raw and Expertly Labelled Images

Figure 1 displays the unprocessed image files for comparison with Figure 2, the expertly-labelled files. With the human eye it is not so difficult to parse cloud from ground based on these images, but we see that cloudy and clear pixels alike run through wide spans of AN so other features must be used.

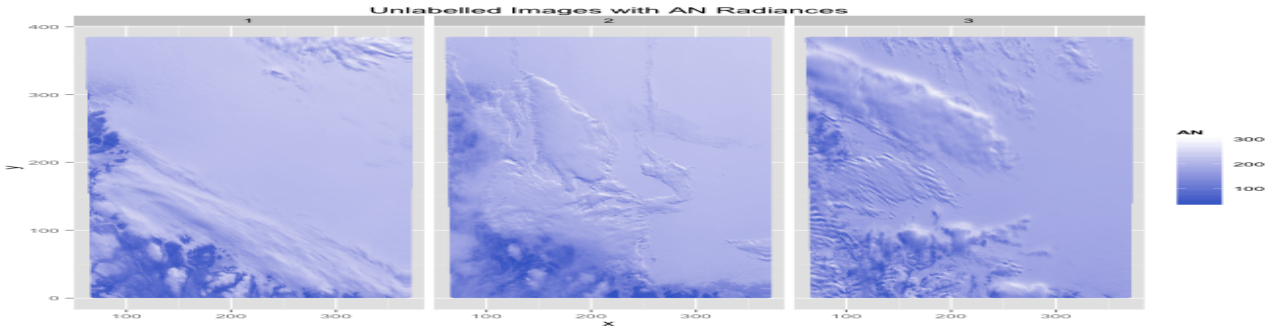


Figure 1: Raw images with AN radiances.

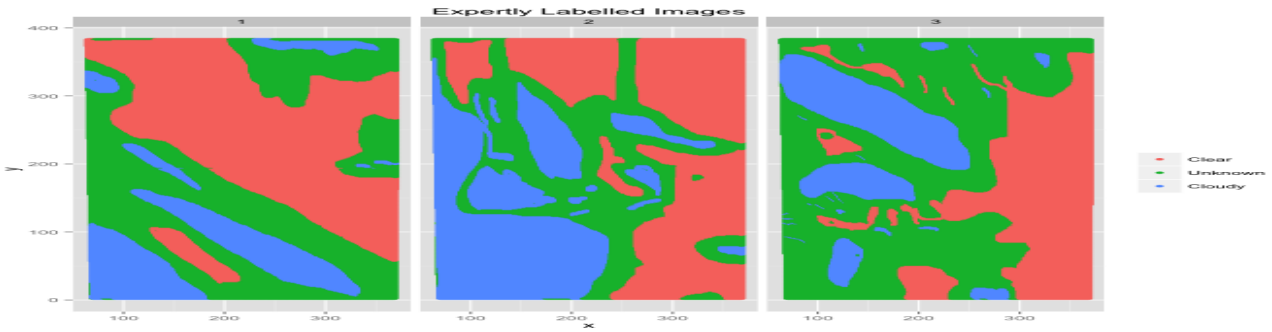


Figure 2: Images with expert classifications. Proportions: 39.8% unknown, 23.4% cloudy, 36.8% clear.

2.2 Densities of NDAI, SD and CORR

The following three plots give us a sense of what can be learned from NDAI, SD and CORR. The densities are grouped by their expert labels, red for 'no cloud', green for 'unknown' and blue for 'cloud'. We see that NDAI has reasonably good separation between cloud and no cloud in all three pictures, which is confirmed in our later modeling sections by Gini importance measures found in the random forest model and its importance as a variable in LDA/QDA and logit models. In comparison, SD has worse separation within the smaller SD values, however it is still clear that pixels labelled as clouds are the only pixels with higher SD values. We thus expect the two features in combination can help determine whether a pixel with high NDAI should be labelled as a cloud by using the SD feature. Finally CORR values appear to be a good separator for image 2, but less so for image 1. This uneven distribution of CORR values across images prompted us to cross validate our models across images in addition to the folds created by dividing each image into 4 quadrants.

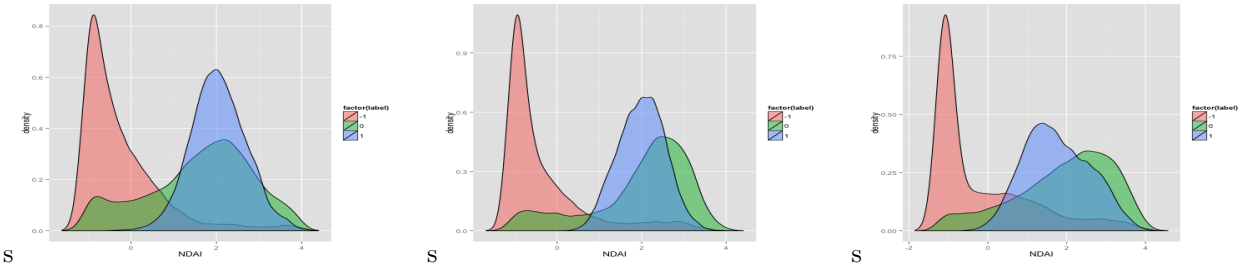


Figure 3: NDAI density plot for Image 1, 2, 3 (respectively).

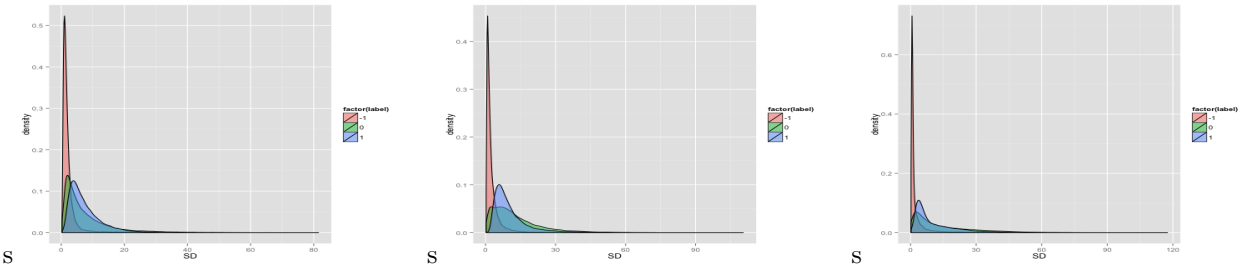


Figure 4: SD density plot for Image 1, 2, 3 (respectively).

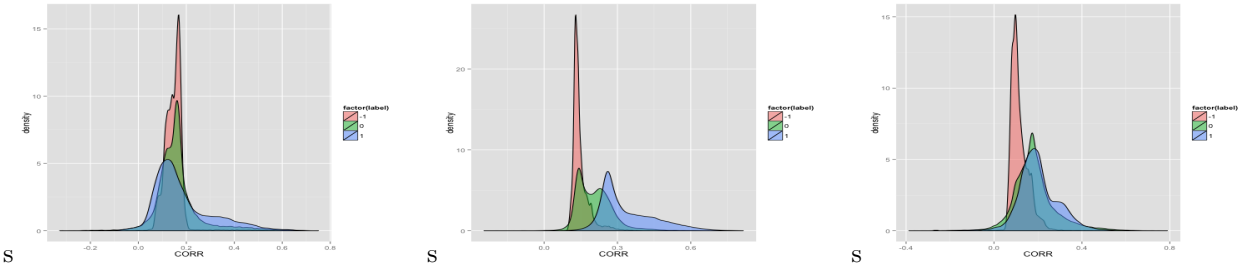
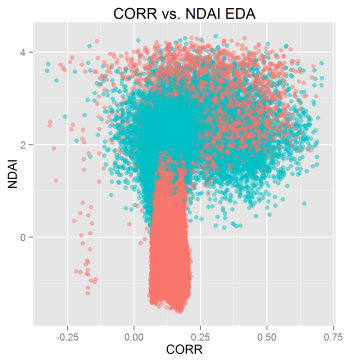


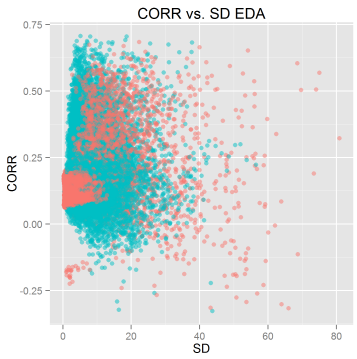
Figure 5: CORR density plot for Image 1, 2, 3 (respectively).

2.3 Mapped Features

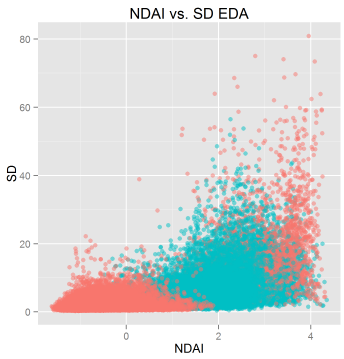
The ensuing figures plot the engineered feature values (NDAI, SD, CORR) spatially. In agreement with the NDAI density plots, higher NDAI values indicate increased likelihood of the presence of clouds, and the NDAI plots look quite similar to the binary classification plots. The SD maps show areas of high variance in the radiances thereby providing a decent outline of cloud boundaries. Furthermore, SD performs well at detecting smooth regions in the image while NDAI is noisy throughout. Unfortunately, this can lead to the highlighting of uneven ground regions. Finally, the CORR images resemble weakened versions of their NDAI counterparts though with some additional strange behavior in the bottom left corners of Images 1 and 2.



(a) CORR vs. NDAI Plot of Image 1



(b) CORR vs. SD Plot of Image 1



(c) NDAI vs. SD Plot of Image1

Figure 6: Scatter plots of the pairs of features used in the analyses. Red points denote pixels that were labeled as not clouds and blue points represent cloudy pixels. This shows hope that our selected features may be able to distinguish between cloudy and icy pixels. The rounded boundaries suggest that non-linear surfaces may be necessary to separate the feature space.

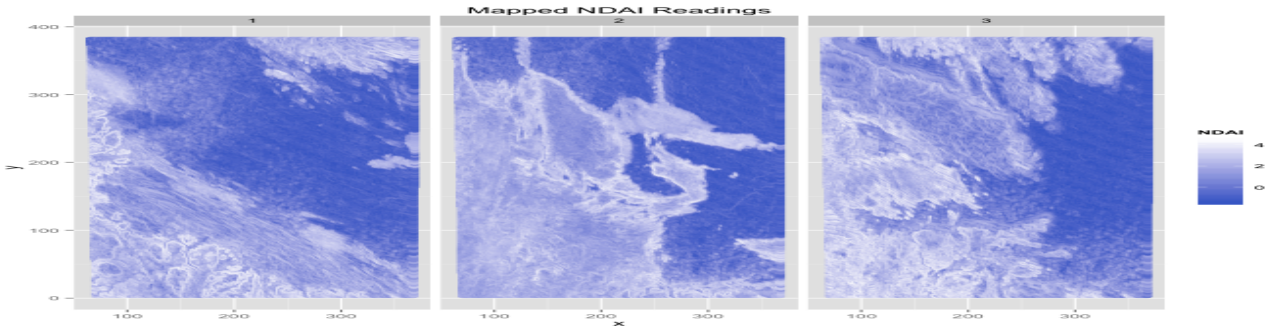


Figure 7: Mapped NDAI readings. We see good correspondence between larger values and presence of clouds.

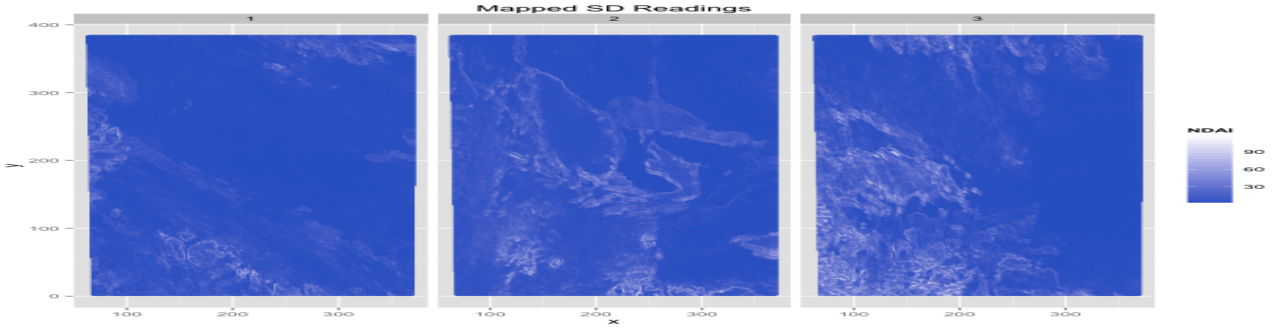


Figure 8: Mapped SD readings. Higher values show cloud boundaries, though also show uneven terrain.

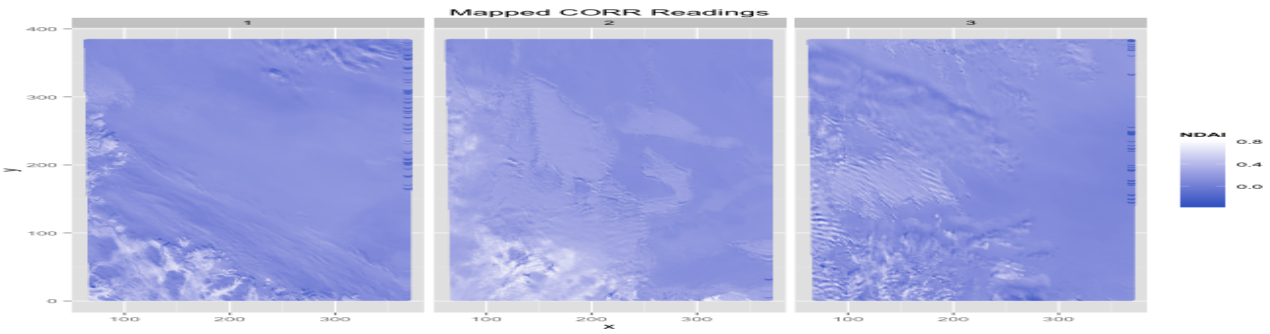
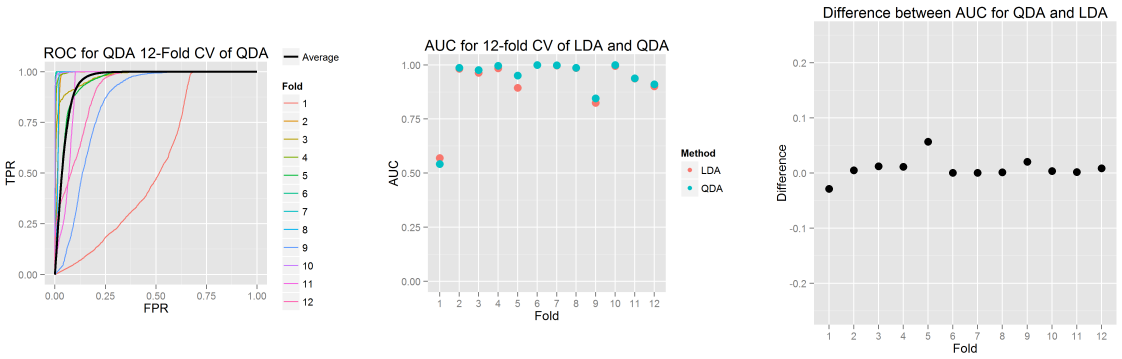


Figure 9: Mapped CORR readings. Cloudy regions tend to be lighter, but not as strongly as in NDAI.

3 Modeling

As shown in the EDA section, it is possible to separate cloudy and clear pixels based on the 3 engineered features. Density plots of the radiances do not display similar separations. This along with the Gini importance from our random forests analysis cause us to focus on NDAI, SD, and CORR.

3.1 QDA/LDA



(a) ROC for 12-fold CV of QDA (b) AUC for LDA QDA Classifiers (c) AUC Differences

Figure 10: (10a) shows the ROC curves for each of the 12 folds of cross validation and the ROC curve for the model with averaged parameters (AUC = .93). (10b) and (10c) show evidence via AUC estimates that QDA performs slightly better than LDA. Note that in both analysis, fold 1 has an abnormally low AUC.

Linear discriminant analysis (LDA) for binary classification aims to separate the feature space using a hyperplane. It assumes that the each class is normally distributed with means μ_0 and μ_1 and that the two classes are homoscedastic with covariance matrix $\Sigma_0 = \Sigma_1 = \Sigma$. Quadratic discriminant analysis works in the same manner except that it makes no assumptions about the covariance matrix. Thus given the multivariate Guassian distribution:

$$f_c(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_c|^{1/2}} e^{-\frac{1}{2}(x-\mu_c)^T \Sigma_c^{-1}(x-\mu_c)} \tag{1}$$

where $c = 1, 2$ is the class label. Let p_c be the prior probability of class c. We can calculate the probability for a case x belonging to class c by:

$$P(x \in \text{ Class } c | X = x) = \frac{f_c(x)p_c}{f_1(x)p_1 + f_2(x)p_2} \tag{2}$$

We can then classify each point by selecting a probability threshold and hard-assigning based on whether or not the posterior probability was higher or lower than that threshold. Both LDA and QDA were carried out during our analyis of the data, but as we suspected from our exploratory data anaylsis, QDA performed slightly better than LDA when we compared the results side-by-side (see figure 10). During cross-validation, each image was divided into four horizontal strips, resulting in 12 folds to be used in our validation procedure. Quadrants were considered but rejected because one quadrant has only clouds and thus QDA breaks down. In each iteration, we withheld one of the strips and trained on the remaining 11. The resulting images after piecing the horizontal strips back together can be seen in figure 12. Cross-validation for QDA revealed that while the method works extremely well in some cases, producing AUC scores of almost 1, it sometimes fails to perform better than even the theoretical random classifier. In particular, the classifier does not seem to be good at discerning snow from clouds in regions where there are many dark pixels corresponding to geographical features. The most indicative example of this is seen in image 1 where the algorithm misclassifies an entire ridge-filled region as clouds. We also noticed that there were some misclassifications within large cloudy regions. Figure 11 shows this nicely.

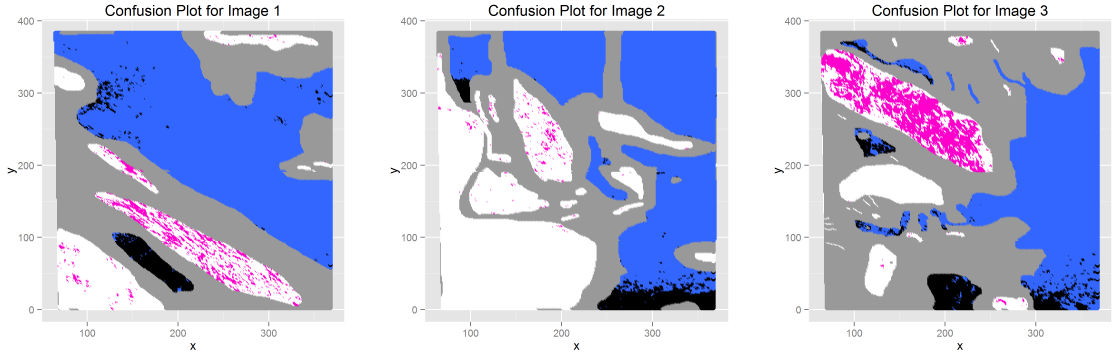


Figure 11: A comparison between the output of QDA and expert labels.

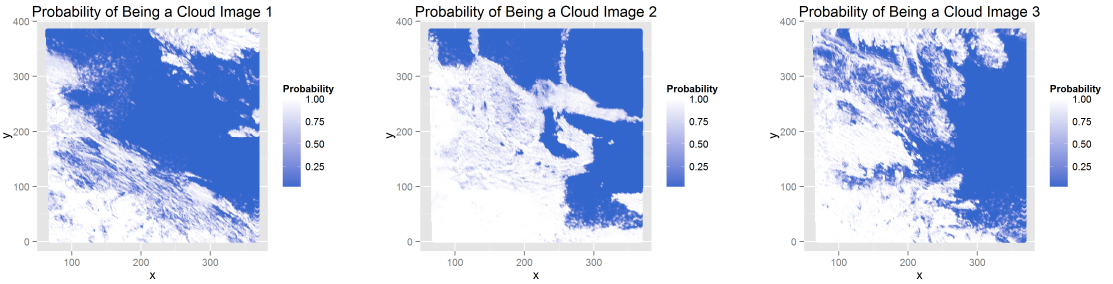


Figure 12: Posterior probabilities of being a cloud as given by QDA

3.2 Logit

Logistic regression is a widely-used discriminative classification method which models the conditional distribution of the outcome given the covariates as a Bernoulli random variable with success probability given by the logistic function evaluated at $X_i\beta$. Here X_i is the covariate vector and β the coefficient vector. It further supposes that the response variables are independent conditioned on the observed features. Classification is

performed in the same manner as in LDA according to a chosen threshold, above which a pixel is labelled cloudy. Preliminary results showed similar performance between the probit and logit models, so we focused on logit due its fewer assumptions (probit requires homoscedasticity and normality of errors) and greater ubiquity in practice. Use of logistic regression also assumes that the model has been properly specified with all significant and no extraneous predictors included.

The logit model obtained via averaging after 12-fold cross-validation has $y_i = -3.356 + 1.900 * NDAI_i - 0.074 * SD_i + 9.002 * CORR_i$ where $P_i(\text{Cloud}) = \frac{1}{1+e^{-y_i}}$. To choose a cutoff value, we varied the cutoffs from .01 to 1 and calculated the misclassification error of our model at that threshold on expertly-labelled pixels. The figure below shows that the optimal threshold for this model is at .38, with an error of just under 10.1%.

Cross-validation was done by dividing each image into quandrants and holding one quadrant out for testing in each iteration of training. There were thus 12-folds in the CV procedure. We chose blocks to comprise each fold rather than a random subsample of points in order to compensate for the correlation of errors among neighboring points. The average CV error was 11.48%. The ROC curve of the logit model is given in Figure 17 and has an AUC of .95 as compared to the AUC of .93 for QDA.

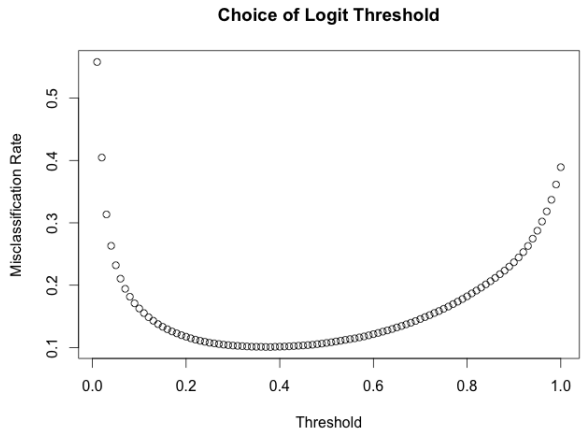


Figure 13: We see a minimum in misclassification rate at .38.

Figure 15 gives the binary classifications overlaid on the images. Visual comparision with the expertly-labelled images suggests that we are slightly biased towards the presence of clouds as 74.2% of unlabelled points have been classified as cloudy, though this may be due to a lack of power in the expert labelling scheme.

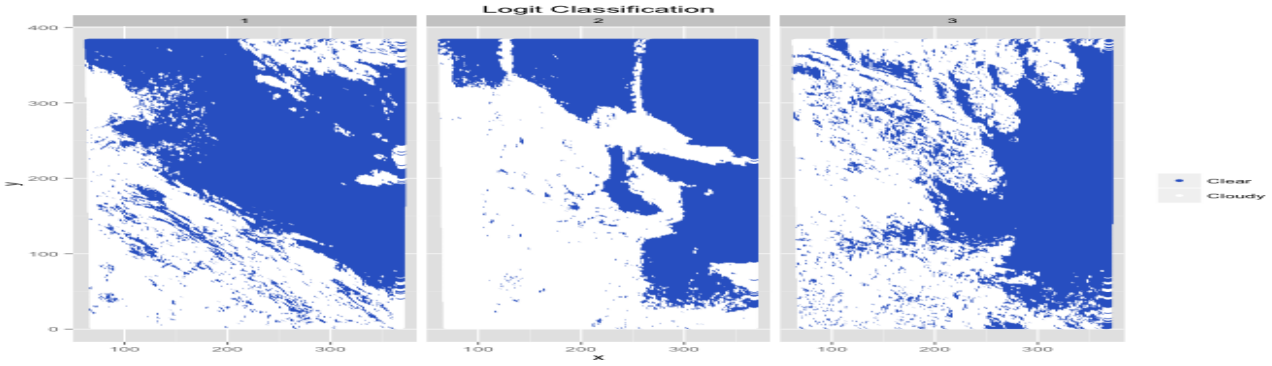


Figure 14: Binary classifications with threshold = .38 for logit trained via 12-fold CV.

Here we see the logit probabilites plotted spatially. This should approximate how the image would appear if the ground were not covered in snow and ice, so comparison with Figure 1 is especially appropriate.

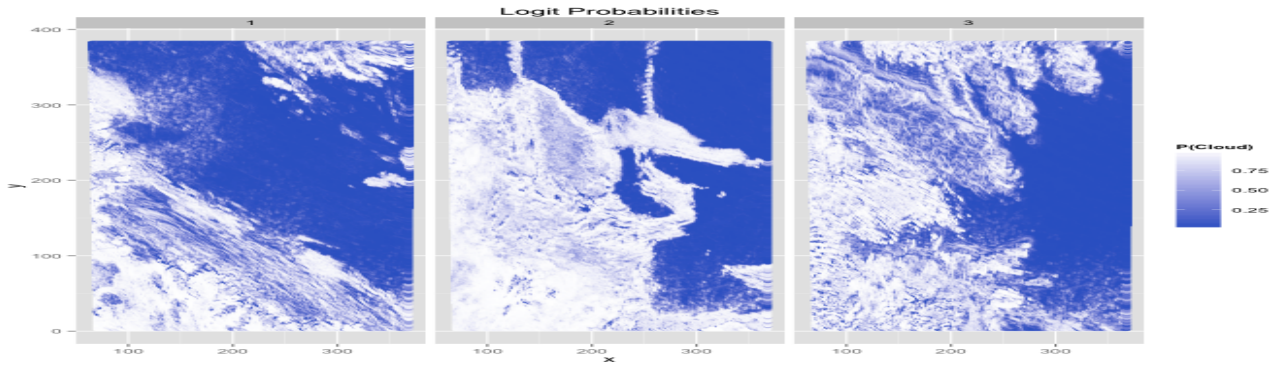


Figure 15: Probabilities for logit trained via 12-fold CV.

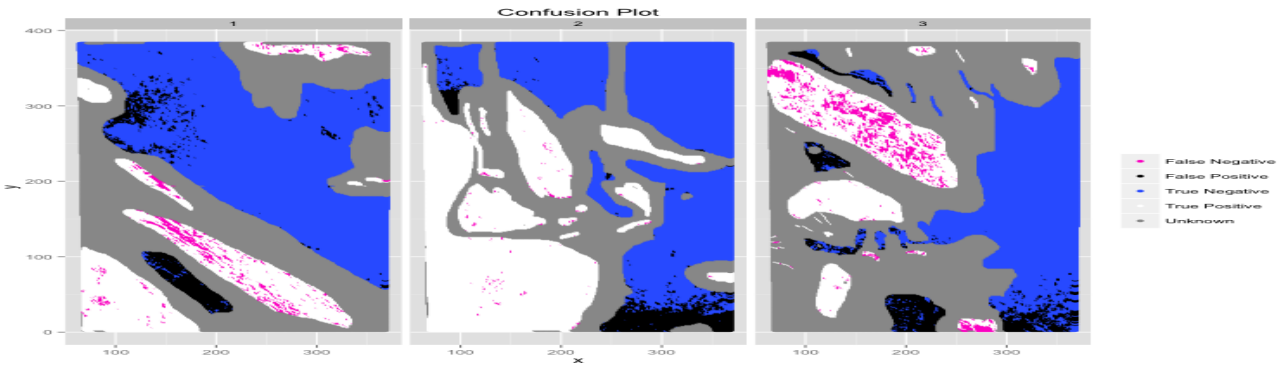


Figure 16: Classification results from logit with respect to the expert labels.

We now discuss the trends in misclassification. As depicted in the above figure, we have many more false positives than false negatives with large regions sometimes completely misclassified. These regions are often situated on boundaries of labelled and unlabelled points. Calculation shows the false negative rate to be 7.5% and the false positive rate to be 11.7%. Models with interaction terms were considered and yielded very slight improvement as measured by drop in data-wide misclassification rate from 10.4% to 9.1%. However, comparison of the misclassification plots showed that the increased complexity did not resolve the large clusters of false positives observed in each image. Especially concerning areas are the large tracts of false positives on the 'island' in the southwest corner of image 1 and the southeast sections of images 2 and 3. Looking at the NDAI images only suggests these areas to be cloudy, but the AN radiance images have dark or uneven patches in these areas. Our model may thus be responding to irregularities in the ground's terrain.

The right half of Figure 17 displays the estimated coefficient values output by logistic regression at each stage of cross-validation. The points are a running average of the previously estimated coefficients, so we expect them to converge towards the higher iteration values. After the first few iterations for all four coefficients, we see convergent behavior, particularly for the intercept and NDAI terms.

As found in our other models, one image outperformed the others when used as a trainer. In the case of logit, training on image 3 gave substantially better predictions than training on either of the other images.The AUC's were .91, .876, and .96 when training was done on image 1, 2, and 3, respectively, with associated misclassification rates 15.5%, 20.4%, and 9.6%. These results inform our hypothesis that images similar to image 3 will be difficult to classify unless similar data were included in the training set. Examination illustrates that image 3 has much sharper variation in its visual features than the other two images, a

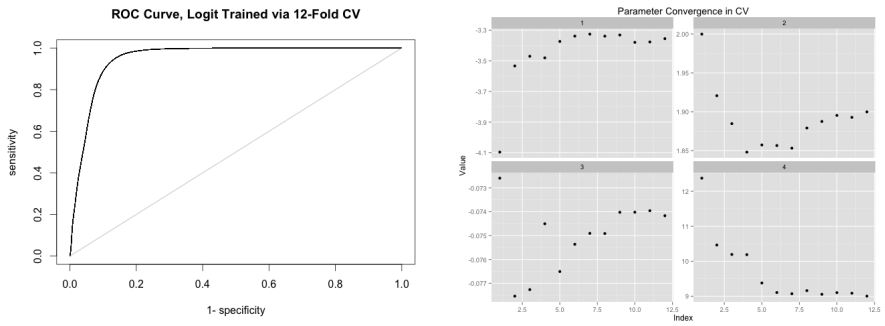


Figure 17: Left: ROC Curve with AUC=.95. Right: Parameter convergence through iterations of CV. 1 - Intercept, 2 - NDAI, 3 - SD, 4 - CORR

phenomenon that is only partially captured by our chosen covariates. Hence we expect our predictions to be fairly accurate on images with smooth changes within cloud and ground surfaces but possibly off in regions of intense heterogeneity.

3.3 Random Forest

Random forest is a decision tree based model that improves generalizability of the model by bootstrapping the training data each time we train a new tree in the forest, and than bagging (bootstrap-aggregate) the forest by having the trees vote on the result of the predictions. This model does not have underlying assumptions on the dataset, expect perhaps that bootstrapping our dataset will produce a representative sample (this will not be true for distributions with heavy tails, which via density plots confirms our data is not distributed such).

In theory we do not need to cross validate as each bootstrap only trains on a randomly selected 2/3 of the data, and also randomly selects a subset of the features to grow the decision tree on. So there is an inherit out of bag error that can be computed by training on the entire dataset. Nonetheless, to compare with the other classifiers, we divided the three images into equal sized quadrants (2X2) in order to do 12 fold validation on the dataset. Keeping the test segments as disjoint chunks from the test set ensured that our models were picking up on 'higher' level structure of the dataset, and not the continuous variation of neighbouring pixels. As such, we also expect the model to general better than if we had randomly subsampled the validation folds. Indeed, after doing random subsampling, we noticed that the AUC was at a high 0.999 for all cross validations, whereas by cross validating on the isolated quadrants, our AUC curves had much higher variance, as seen below:

	AUC
1st fold	0.996600222048787
2st fold	0.960408561886229
3st fold	0.629930590380754
4st fold	0.995687025393392
5st fold	0.999193970074872
6st fold	0.973449221681138
8st fold	0.945314282554696
9st fold	0.996766914151693
10st fold	0.594564820737272
11st fold	0.788939754299337
12st fold	0.946931831284243

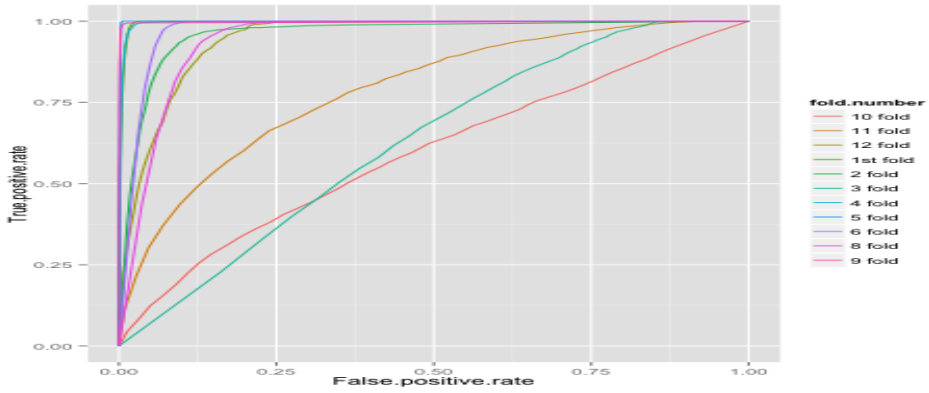


Figure 18: ROC across all 12 folds of quadrants

We also trained on each image and tested on the remaining two, AUC values and ROC curves are plotted below.

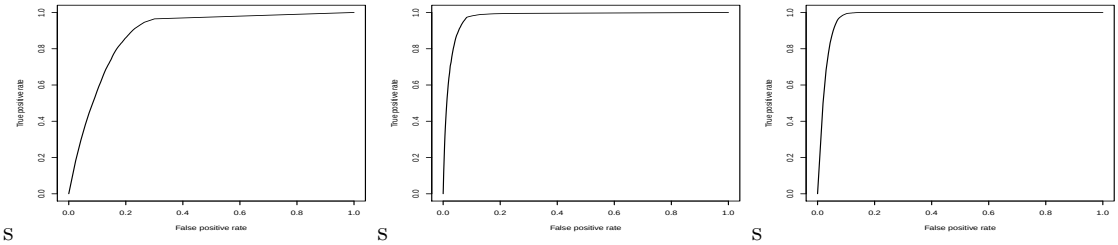


Figure 19: train 1 and 2, test 3: AUC: 0.8857

Figure 20: train 2 and 3, test 1: AUC: 0.9757

Figure 21: train 1 and 3, test 2: AUC: 0.9462

Where was the model performing poorly? The images below show classification errors for models trained on 2 images and tested on the other. Unsurprisingly this model did not perform as well as the 11-fold trained one but plotting their classification error helps us see where the model failed. Below, we plot the classfication error and juxtapose it with the feature with the highest Gini importance (explained below) NDAI. We plot the best and worst of the 3.

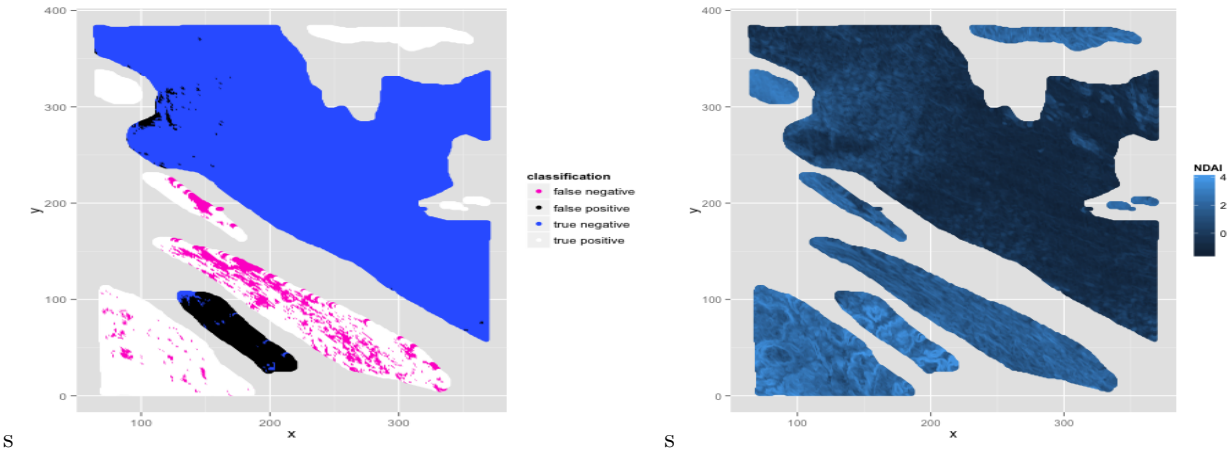


Figure 22: prediction error for image 1-trained on image 2 and 3 juxtaposed with NDAI

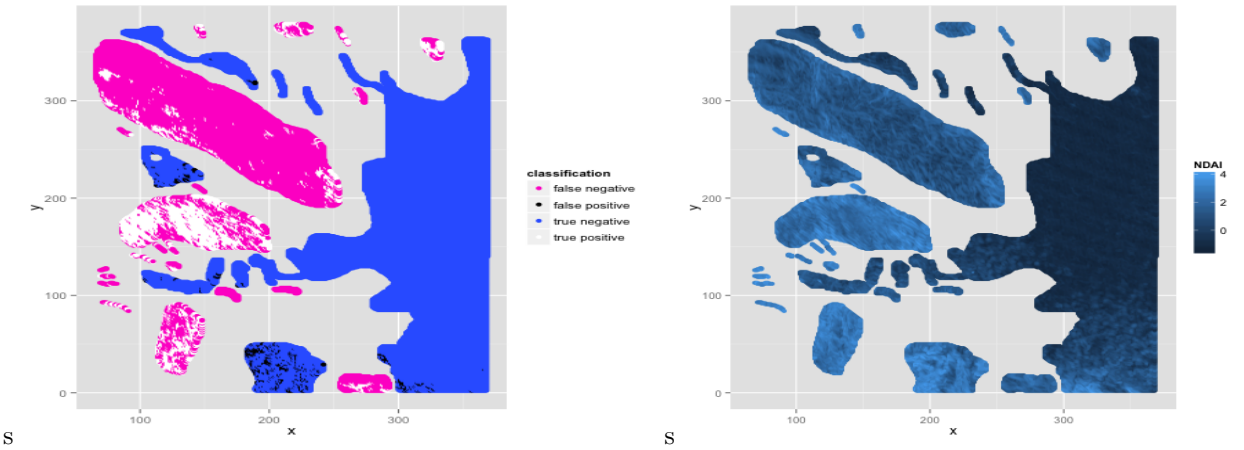


Figure 23: prediction error for image 3-trained on image 1 and 2 juxtaposed with NDAI

Convergence To test convergence, we trained on an increasing number of quadrants and looked at whether the AUC values increased as we increased the training set. We also trained random forest with a range of forest sizes, from 2 trees to 50. By 25 trees, we were seeing very similar predictive behaviour. Finally, since we had less than 9 features to train on, we stuck with the finest level of detail possible and branched on one feature per node. It was interesting to note that the continuity of increasing the training set by neighbouring quadrants displayed much less convergence than by increasing the training quadrant set in a randomly sub-sampled way. The following three plots make this clear. This shows how important the continuity between neighbouring pixels are in prediction, which is something we did not want to pick up on in our models, since that does not generalize well to new data.

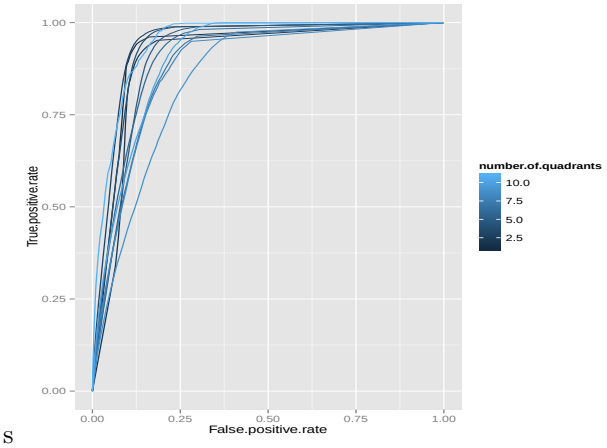


Figure 24: Convergence of ROC training on increasing number of quadrants

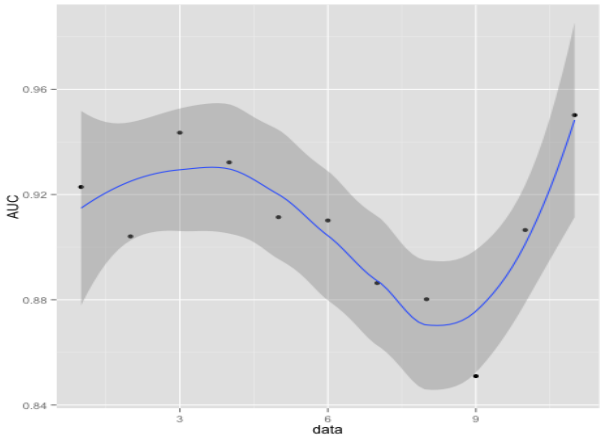


Figure 25: Smoothed convergence of AUC for growing training set with 50 trees and 3 features

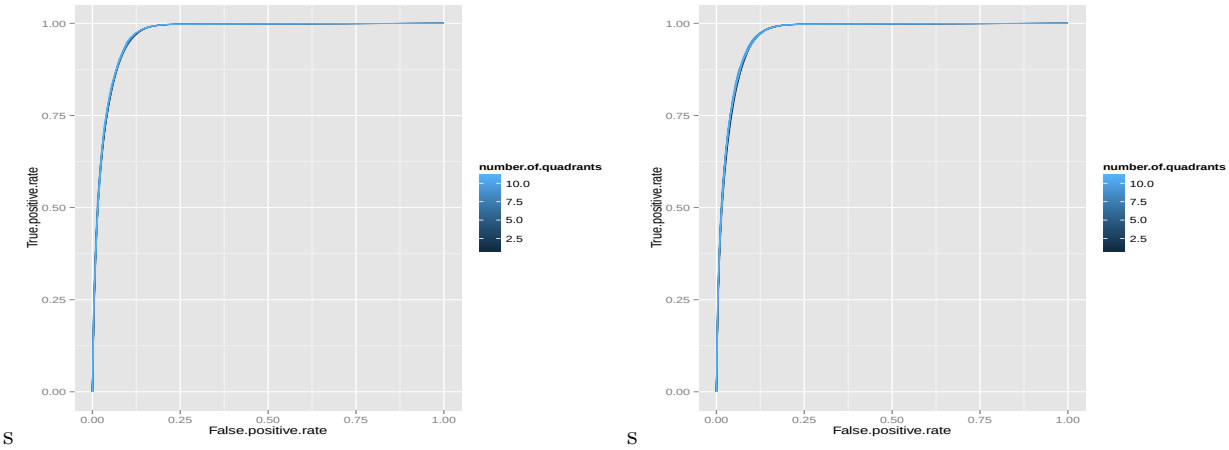


Figure 26: Convergence of ROC training on increasing number of quadrants in shuffled order. Shuffled randomly twice to generate two plots of AUC convergence

Variable Importance To gather a quantitative measurement of the importance of each feature, we first ran random forest using all the features and looked at the Gini importance measure. This is calculated by recording the difference between the Gini measure of a random forest's predictions on a fold and the Gini measure with a particular feature's values randomly shuffled. The intuition is that if a feature were crucial to the forest's trees, than randomly shuffling that feature will drastically decrease its Gini measure. Sure enough, NDAI, SD and CORR consistently ranked as the top three in all cross validations.

	mean	variance	sd
NDAI	31851.137865015	5541519.63141267	2354.04325181435
SD	15181.3036164789	7250037.34195744	2692.58933778574
CORR	16696.8895555848	15065266.3880018	3881.40005513498
DF	5153.57742916125	715182.664503553	845.684731151954
CF	3725.98512846366	238611.088513114	488.478339860749
BF	4741.67090014784	908691.31228918	953.253015882551
AF	6307.45653651382	1489619.01113017	1220.49949247436
AN	7656.24468437671	5730133.08077185	2393.76963820077

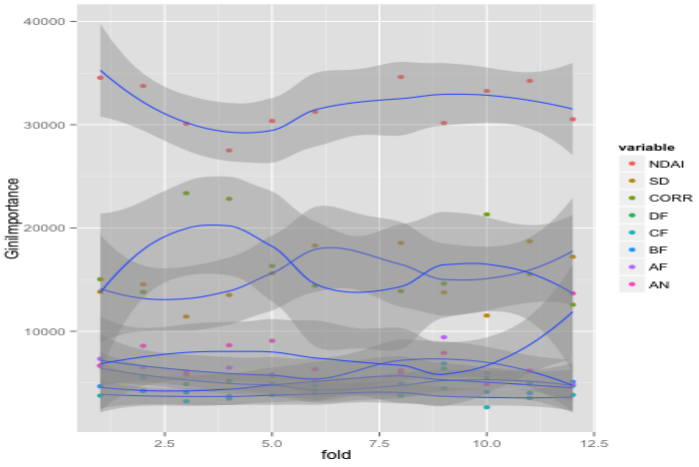


Figure 27: Gini importance measures across folds

4 Reproducibility

How we organized out code and github repo

5 Conclusion

Logit, QDA and random forest all created reasonable predictive models with high AUC values. On average, random forest did best based on the mean and variance the AUC values across the folds. Prediction errors

for all three models shared similar shortcomings as seen in the classification error plots. Nonetheless, it is not clear whether the models are failing for similar reasons. Logit and QDA/LDA have similar assumptions on the distribution of the classifications, but random forest only assumes no heavy tails. In the absence of significant domain knowledge, coupled with the better AUC/ROC performance, we are inclined to conclude random forest generalizes better. It also helped confirm the significance of the derived features NDAI, CORR and SD.

References

[1]