Lab 4: Cloud Data
Statistics 215A, Fall 2014

Lindsey Hearn

12 November 2014

https://github.com/lindsey7/STAT-215A-Lab-4

1 EDA

Figure 1 plots the presence and absence of clouds, `expert==1` and `expert==-1`, respectively, by MISR image, according to X-Y coordinates. Figure 1 uses a subset of the data for which expert labels exist, omitting `expert==0`. The dark shading indicates the presence of a cloud, `expert==1`, while the light shading indicates the absence of a cloud, `expert==-1`.

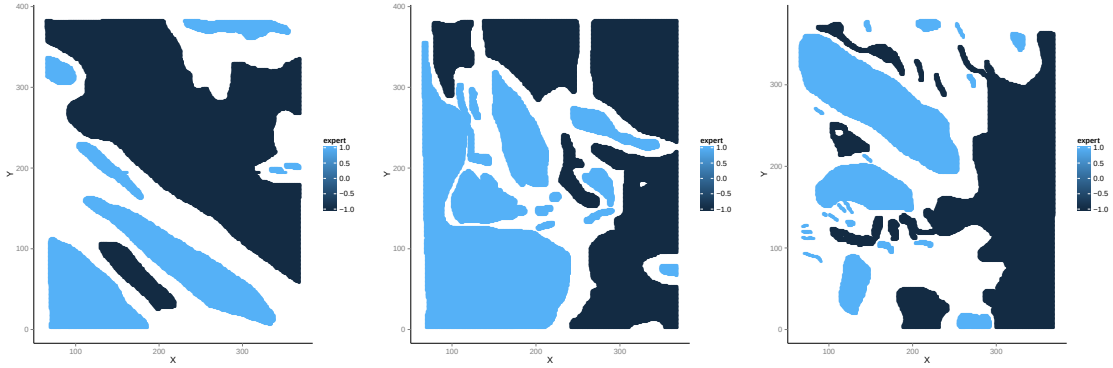
Figure 2 plots the spatial distribution of radiance for each angle in the image files. Figures 2 (a), (b), (c) present markedly similar spatial distributions, while figures 2 (d) and (e) present similar arrangements. Nevertheless, radiance distribution exhibits a high degree of similarly across all angles. Figure 3 presents the conditional density of radiance by angle. Figure 3 restricts attention to classified observations. The blue region corresponds to `expert==1`, indicating the presence of a cloud, while the gray region corresponds to `expert==-1`, indicating the absence of a cloud. The conditional densities are similar across angles, with slight variation in the distribution of radiance for `expert==-1`.

Figures 3, 4, and 5 plot the conditional densities for CORR, NDAI, and SD, respectively. Within each figure, the densities are presented separately, by image. Collectively, CORR, NDAI, and SD exhibit more correspondence to the presence and absence of clouds, than the angles in the previous plots. Furthermore, the CORR, NDAI, and SD features are less correlated than the angles. This, in conjunction with the greater association with the `expert` label, make the features compelling candidates for predictors in the modeling exercise. The table below provides the pairwise correlation matrix across angles, demonstrating the high level of correlation between angles' radiances. Whereas the subsequent

	DF	CF	BF	AF	AN
DF	1.0000000	0.8503037	0.6703445	0.5377937	0.4892642
CF	0.8503037	1.0000000	0.9189584	0.8259473	0.7795202
BF	0.6703445	0.9189584	1.0000000	0.9624793	0.9255600
AF	0.5377937	0.8259473	0.9624793	1.0000000	0.9819174
AN	0.4892642	0.7795202	0.9255600	0.9819174	1.0000000

table below documents a lower average level of correlation between features CORR, NDAI, and SD.

	ndai	corr	sd
ndai	1.0000000	0.5350207	0.6474474
corr	0.5350207	1.0000000	0.4073057
sd	0.6474474	0.4073057	1.0000000



(a) Image 1 (b) Image 2 (c) Image 3

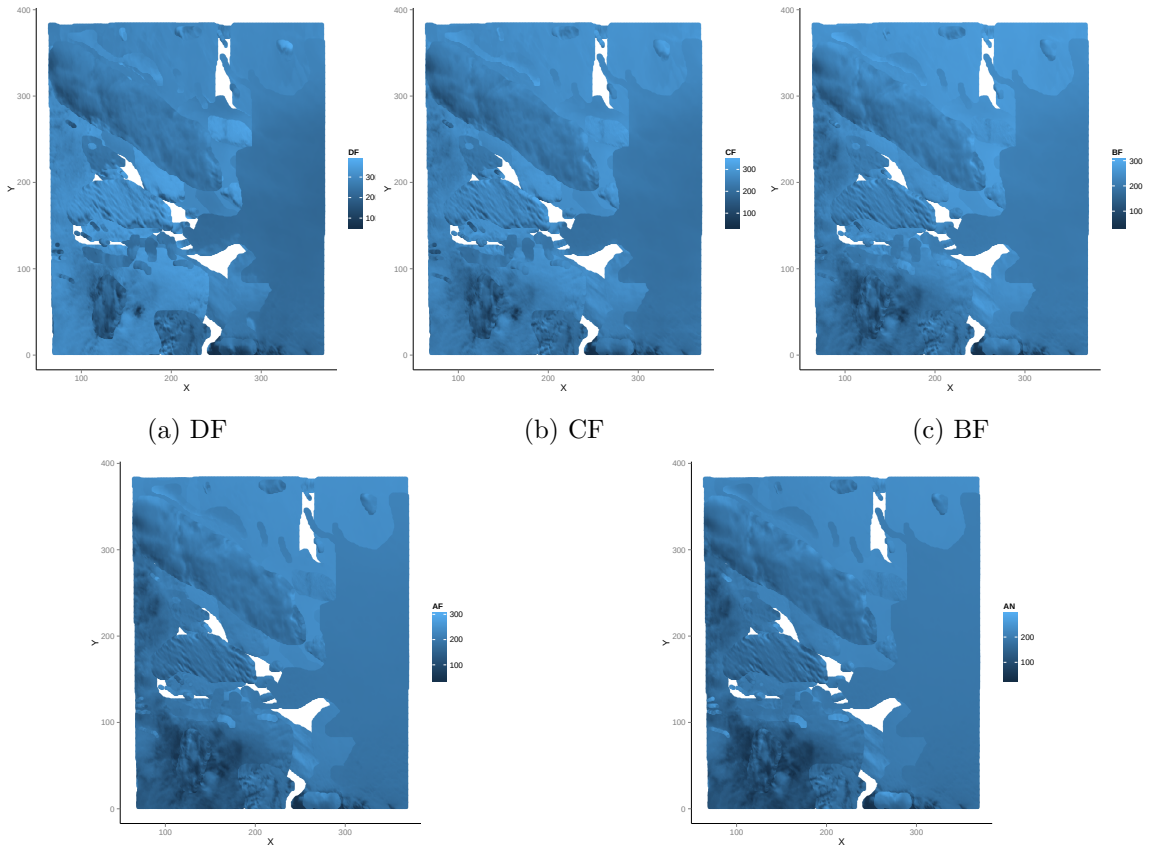
Figure 1: Expert labels according to X-Y coodinates

2 Modeling

2.1 Selection of Predictors

Assuming the expert labels represent the true classification of the presence of clouds, CORR, NDAI, and SD predict the presence of clouds better than the radiances of angles. The conditional densities in figures 4, 5, and 6 generally exhibit less overlap in the two distributions than the conditional densities represented in figure 3.

Calculating the false positive rate (FPR) and true positive rate (TPR) using predicted responses from logistic regressions using angles as predictors in one specification and the alternate predictors in another specification: the alternate predictors (CORR, NDAI, and SD) produce a much higher TPR value, 0.8718 compared to 0.6925. The FPR and TPR for the two models are assessed using a threshold of 0.5. Though the alternate predictors yield a higher false positive rate (0.094 compared to 0.0527) the gain in correct positive classifications is greater than the increase in false positive classifications.



(a) DF (b) CF (c) BF

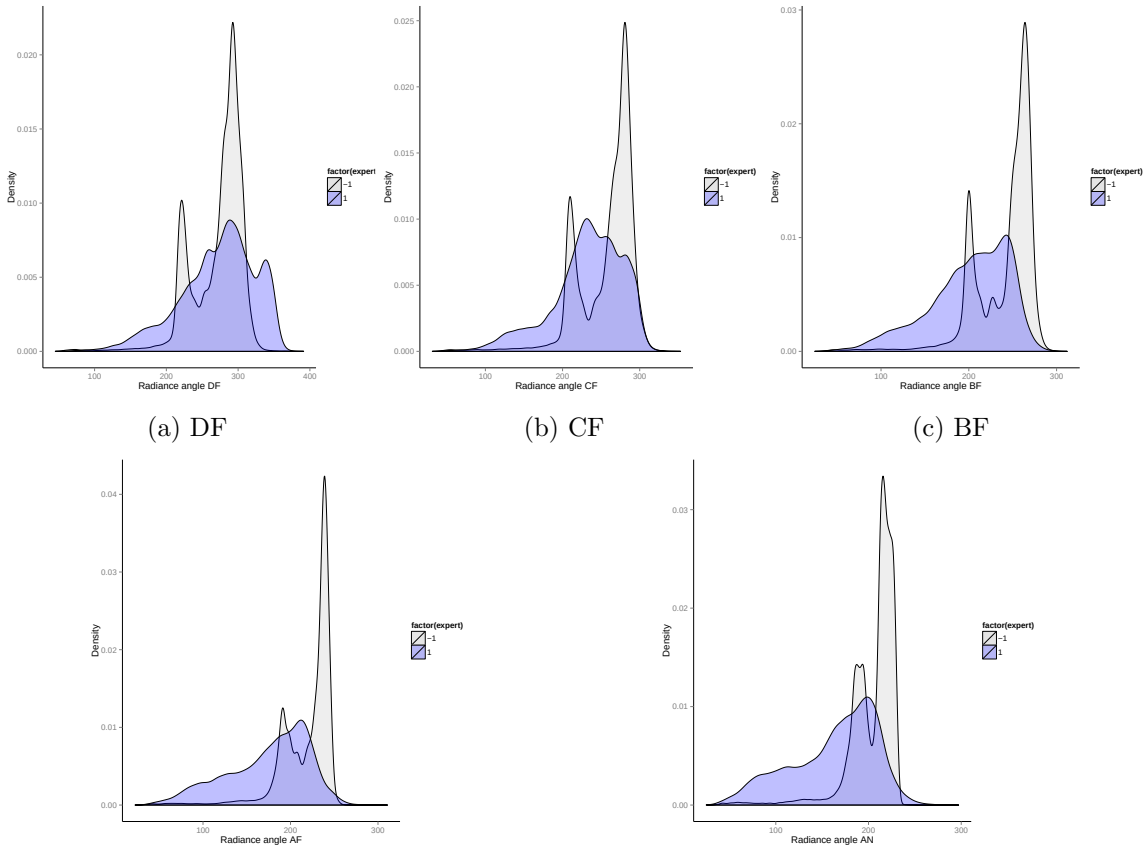
(d) AF (e) AN

Figure 2: Spatial distribution of radiance by angle

2.2 Binary Classification Methods

Logistic Regression Model (GLM)

Logistic regression predicts a binary outcome from a set of continuous or categorical predictors. The binary outcome of interest, in this case, `cloud` (the factor transformation of the `expert` term from the previous section), is distributed according to the Bernoulli distribution. The model predicts the probability that `cloud==1` according to the logistic distribution. Logistic regression may be preferred to discriminant function analysis as the necessary assumptions are less prohibitive.



(a) DF (b) CF (c) BF

(d) AF (e) AN

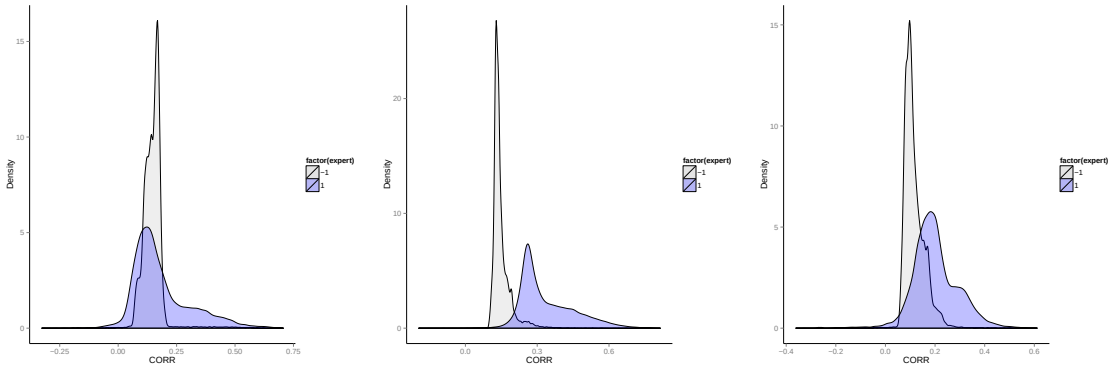
Figure 3: Density plots by presence and absence of clouds

Linear Discriminant Analysis (LDA)

The linear discriminant analysis (LDA) predicts classes of events or outcomes. The critical motivating assumptions are: (a) normally distributed predictors; (b) full rank covariance matrices; and (c) homoskedastic random variables.

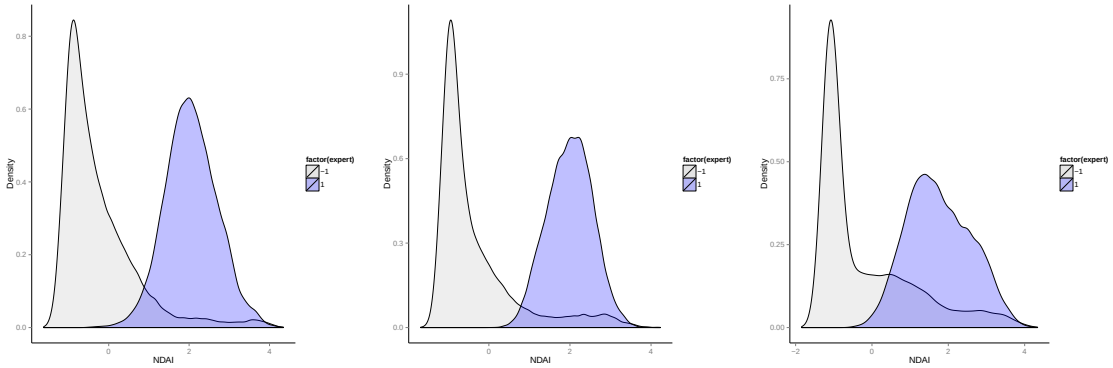
Random Forest Model

The random forest model applies the technique of bootstrap aggregating (bagging) to trees, averaging results by tree. Random forests correct for the potential overfitting that may occur in tree learning systems, by averaging deep trees.



(a) Image 1 (b) Image 2 (c) Image 3

Figure 4: Density plots, CORR

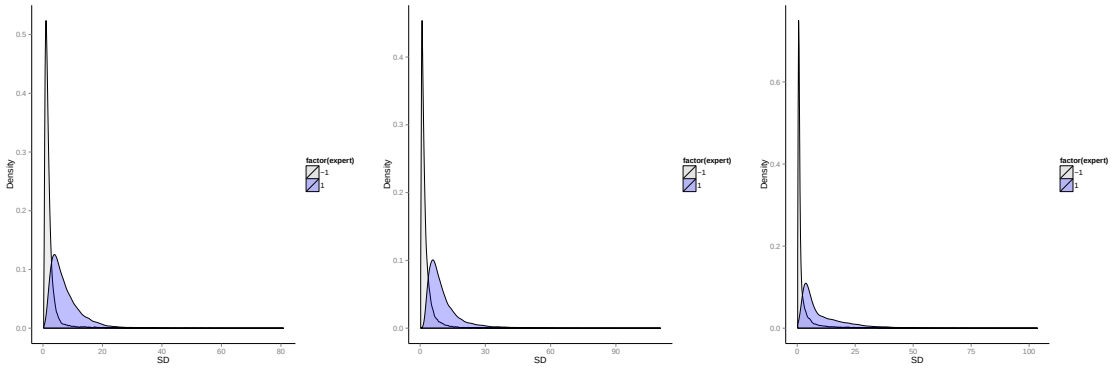


(a) Image 1 (b) Image 2 (c) Image 3

Figure 5: Density plots, NDAI

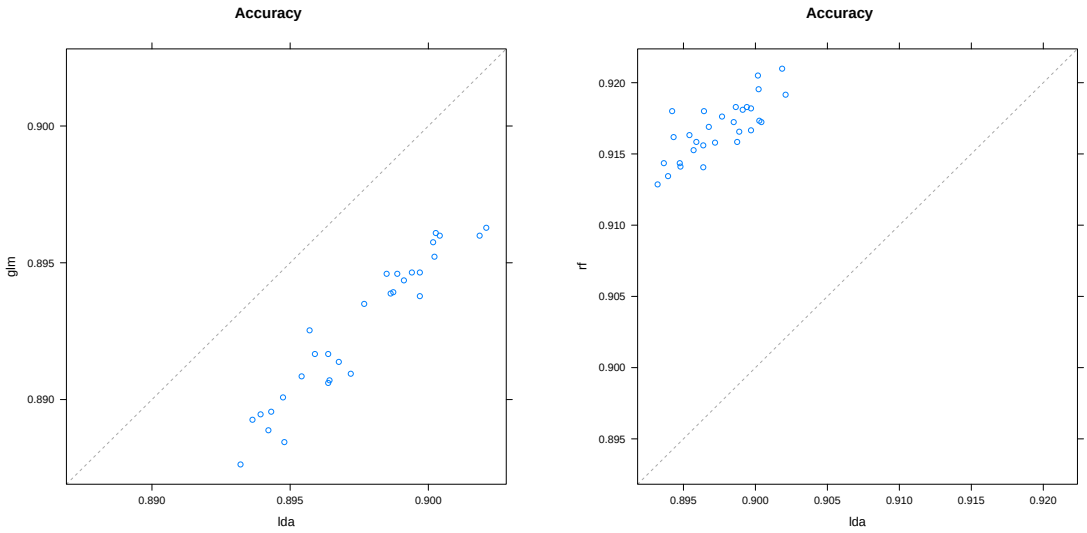
2.3 Cross-validation

Figure 7 presents comparisons of the accuracy of the logistic model, linear discriminant analysis, and random forest model. Accuracy represents the average agreement rate. In this case, the agreement fraction is averaged over 10 cross-validation iterations. The default number of iterations using the `caret` training package is set at $K = 10$ (see R code for more detail). The specification is reasonable as the 10-fold cross-validation allows the model to correct for overfitting in a particular subset, while producing shorter run times than $K > 10$. Using the `preProc` function in the `caret` package, the data is centered and scaled, prior to training. Random forests appear to preform better than LDA, as the average agreement fraction is greater with the random forest method (figure 7 (b)), while LDA appears to yield higher average agreement than the logistic model (figure 7 (a)).



(a) Image 1 (b) Image 2 (c) Image 3

Figure 6: Density plots, SD



(a) LDA-GLM (b) LDA-RF

Figure 7: Accuracy comparison GLM-LDA-RF

The random forest model has the highest average accuracy of the 3 methods examined in detail within this report (see R code for additional specifications): an average accuracy of approximately 0.92 compared with 0.90 for LDA. Figure 8 plots the receiver operating characteristic (ROC), using repeated cross-validation, against the number of randomly selected parameters. The ROC is highest for 2 predictors. Figure 9 presents the ROC curve for the tuned random forest model, using `mtry==2`, the mtry value returned from the initial random forest model.

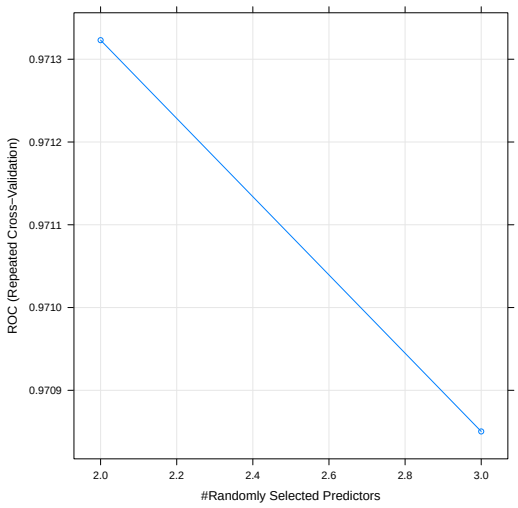


Figure 8: ROC

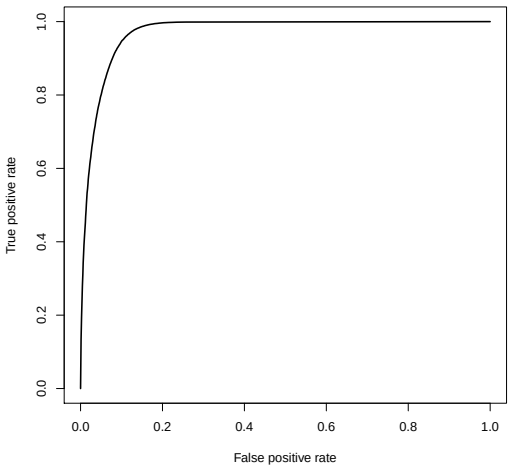
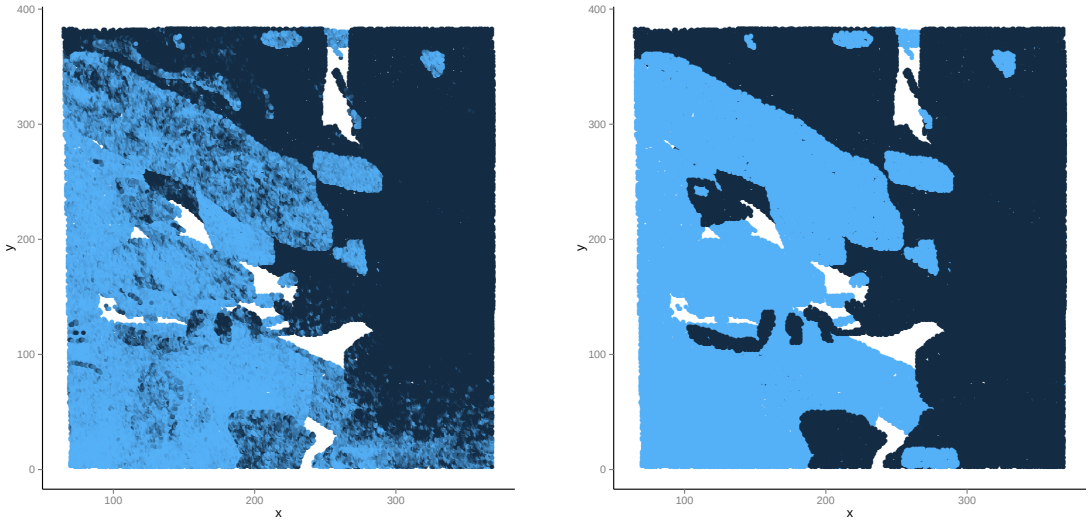


Figure 9: ROC curve for random forest model

2.4 Predicted Values

Figure 10 (a) plots the spatial distribution of the predicted presence and absence of clouds in the `test.data` dataset, while figure 10 (b) plots the actual distribution the random forest model attempts to predict. The dark shading indicates the absence of clouds, while the light shading indicates the presence of clouds. The spatial distributions are relatively similar with the predicted distribution following generally the same pattern according to X-Y coordinates. Nevertheless, the predictions produce false positives in the lighter region, while false negatives are sparse.



(a) Predictions

(b) Observed `cloud` indicator

Figure 10: Spatial distribution of predictions and true indicators using `test.data`

While the predictions produce reliable indicators for presence of clouds within the aggregated image files used for this lab, the predictions could correspond to specific features of the particular image files used in conducting the analysis. Though the random forest model produces generally accurate predictions, the model could be overly responsive to particular features of the `image.txt` files. However, the model appears to predict the presence of clouds well, with a low false positive rate. Therefore, perhaps this particular random forest model would produce more reliable estimates of the presence of clouds, rather than the absence of clouds, in future data without expert labels.

3 Reproducibility

The repository for this project is provided preceding the first section of this report. The repository contains the requisite files: a README file describing how to reproduce this report; an R file with code to reproduce the statistical and graphical analysis; and the raw .tex file.