Case Study

Garima Garg

April 2016

1 Plant Harvest Point Association Algorithm

To find the association between the harvest points and planting points, it is reasonable to find the nearest neighbour from among the planting points, for each of the harvest point, and then based on the nearest neighbour, merge the information about planting points with the information about harvest points. Some of the algorithms are-

- Distance Based Algorithm - The naive algorithm for nearest neighbour search can be distance based, but this would mean, computing the distance for each of the harvest point to all the planting points, and then select the minimum distance planting point. So this algorithm would require almost 6314 comparisons(number of planting points) for each harvest point.
- Recursive Partitioning based on latitude - The second method is based on recursive partitioning the planting point space until you are left with few neighbours for that harvest point, and then used a distance metric(euclidean distance) to find the minimum distance between harvest point and its neighbours. For this algorithm, we can recursively sample a point(latitude) from planting points, and then compare it with harvest point(latitude) in order to reduce the planting point space while making sure that harvest point comes under reduced planting space. This algorithm runs iteratively until the distance between sample point and harvest point is very very less, around .0001. Only few neighbours are left in the reduced planting space, where we can then compute the euclidean distance to find the minimum distance planting point. This definitely reduces the number of comparisons when compared with distance based algorithm, but it requires sampling a point that adds complexity. A better way would be to already create a decision boundary, so that we don't have to sample a point and save it in a list, and then keep on traversing the list until you reach the end. It is still inefficient as after reaching the reduced planting space, one need to subset the data from the entire data that adds more complexity and inefficiency.
- Distance based Algorithm on reduced planting space - We already know the points are very close to each other, and the nearest neighbour would

be in this range - (latitude of harvest point -.0001, latitude of harvest point +.0001). In order to find the closest neighbour for each of the harvest point, we can just look at this reduced planting space rather than comparing with entire planting points and find the minimum distance planting point. But we would need to filter the planting data based on reduced range that adds to some complexity.

- Most efficient algorithm based on KD tree nearest neighbour search - Kd tree for nearest neighbour search is the best algorithm among all the algorithms mentioned above. First, it stores the data in a tree data structure(to avoid filtering the data), and then query the nearest point after traversing the tree. The search for the nearest neighbour takes only O(log N) time, where N is the number of planting points.

So, I used nearest neighbour search using kd tree and then combined the harvest and planting data to get a final complete data. Fig 1 shows the nearest neighbour for a sample of harvest points.

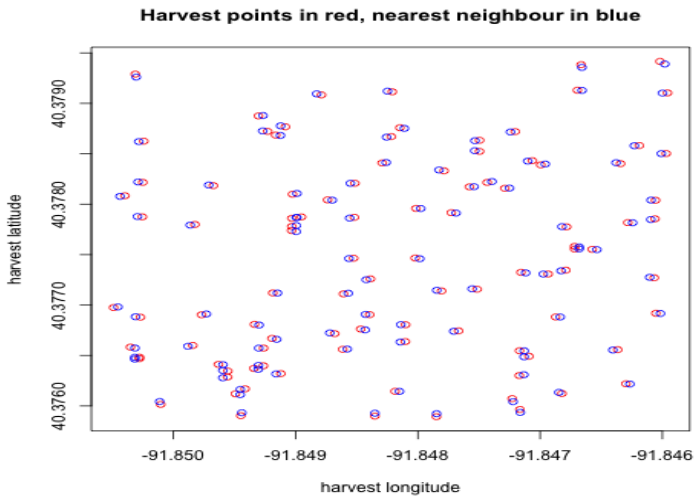

Figure 1: Nearest neighbour for a sample of harvest points

2 Part2: Understanding relationship between variables

2.1 Exploratory Data Analysis

Fig 2 shows the relationship between different variables in the complete dataset. Seeding rate and seed spacing do not have a linear relationship with yield. There

is hardly any significant correlation between these variables and yield. Seeding rate and seed spacing have a non-linear type of relationship. The mean value of yield for both the seed varieties is almost equal.

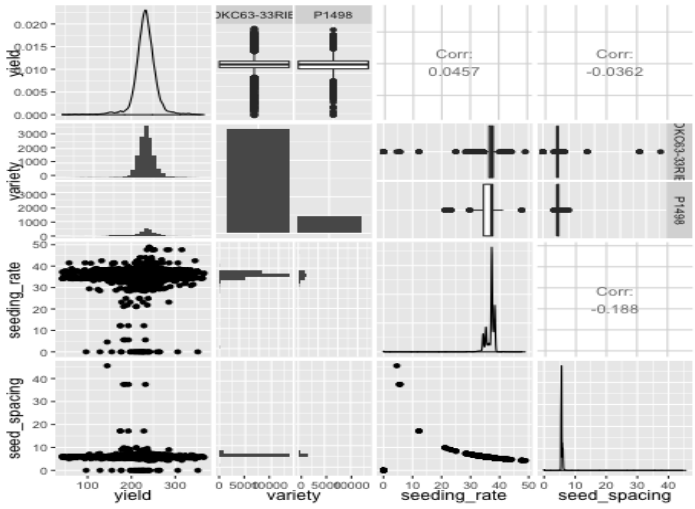

Figure 2: Plot between different variables

Fig 3 shows the plot between yield and seeding rate grouped by the two seed varities; the seeding rate ranges from (0, 48.8162) for seed variety 'DKC63-33RIB' whereas seeding rate varies from (21.1561 47.6412) for seed variety 'P1498', the mean value of seeding rate for both of the seed varieties are almost same. There is also slight difference in yield between the two seed varieties. Table 1 shows the mean of different variables grouped by seed variety.

variety	yield	seeding rate	seed spacing
DKC63-33RIB	231.8932	36.84635	5.666967
P1498	226.5109	36.29613	5.782165

Table 1: Table showing mean of different variables grouped by seed variety

Fig 4 shows the relationship between yield and seed spacing grouped by the two seed variety.

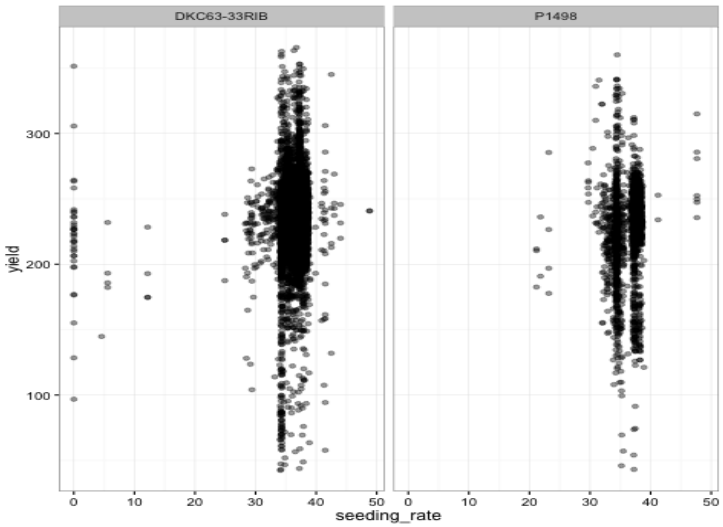

Figure 3: Plot between yield and seeding rate grouped by seed variety

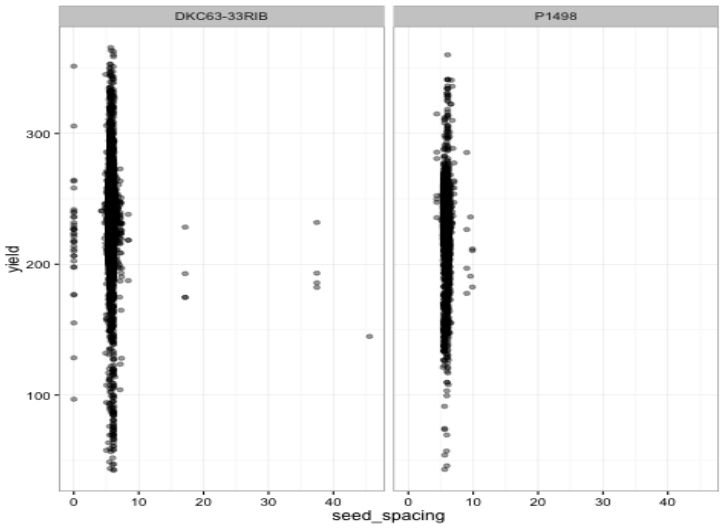

Figure 4: Plot between yield and seed spacing grouped by seed variety

To understand the relationship between these variables, I used decision tree; first I scaled the variables - yield, seeding rate, seed spacing by dividing it with their max value so that they are in the range of (0,1). Fig 5 shows the decision tree for yield. The predictor space is split into many partitions corresponding to different values of yield. Pruning was done to make the tree more compact, and avoid problems of over-fitting. I selected the parameter by cross-validation.

The different values of yield are dependent on seeding rate, seed spacing and variety.

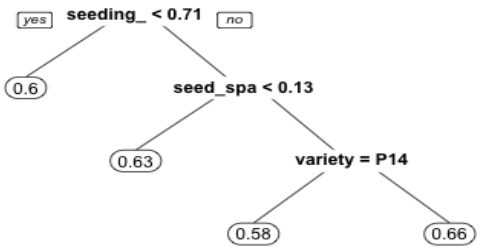

Figure 5: Decision Tree for yield

Fig 6 shows the decision tree for seed variety. Using random Forest, I was able to find the variable importance plot for yield. Seed variety is the most important predictor based on MSE. Fig 7 shows the variable importance plot.

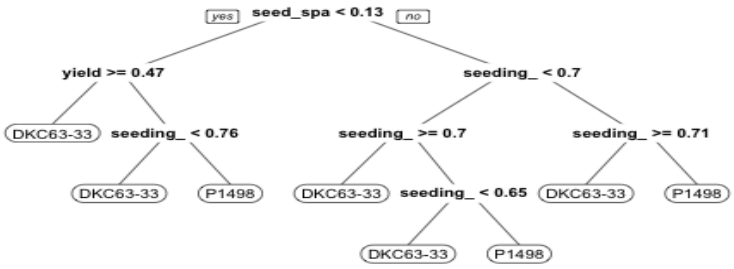

Figure 6: Decision Tree for seed variety

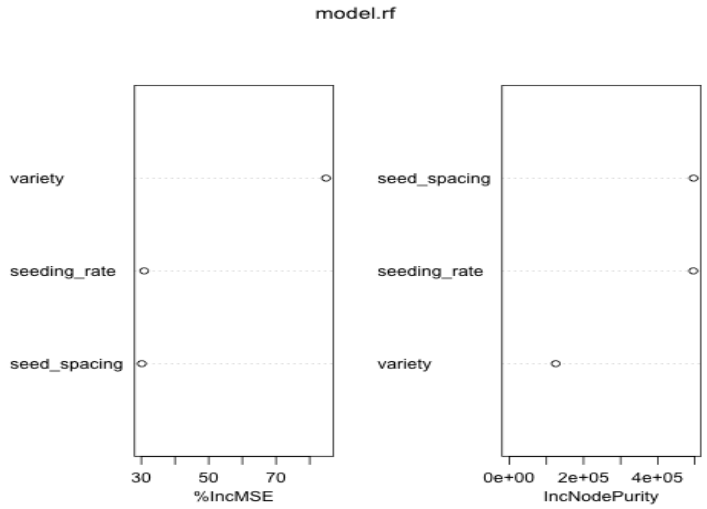

Figure 7: Variable Importance Plot for yield

2.2 Predicting Yield using random forest

Random forests are a way of averaging multiple decision trees. It was used to train the model on a subset of complete data(70% of the data), and then finally test it on remaining subset of the complete data. When predictors are more, pruning should be done in order to select the split candidates that minimizes out of bag(OOB) error. The Root Mean Square Error is 24.79 on test data, which is very high, and thus prediction accuracy is not good. This is because of very few predictors(and that too, not directly related with yield). To be more accurate in reporting RMSE, I used 10-fold cross-validation, trained my model on 9-folds, tested it on remaining fold, and repeated the process 10 times. From cross-validation approach too, the RMSE came around to be 24.59.

Fig 8 shows the plot between original yield and predicted yield for training and test data, one can see the predictions doesn't generalizes to entire data range. I tried predicting yield with gradient boosted algorithm, but I got almost same RMSE. Getting more predictors that are related with yield will solve the problem.

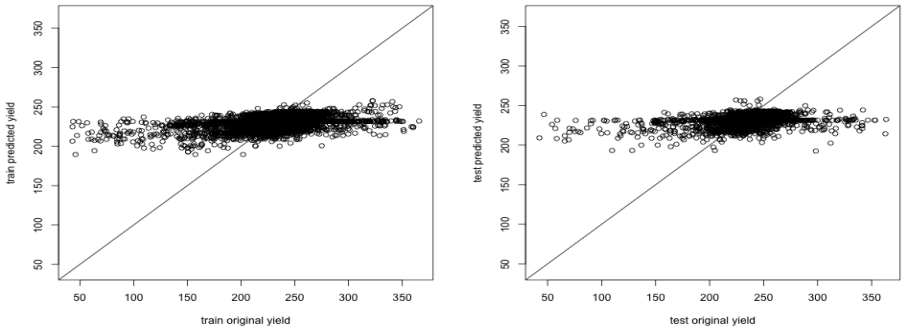

Figure 8: Predicting yield on training and test data