Forest Cover Type Prediction

1st Project Report for CSE 581

Dong Han
October 27, 2014
dhan@oakland.edu

Overview

- Introduction of sample data
- Process features to numerical variable or categorical variable
- View of features correlations
- Neural network classifier
    - Evaluation method
    - Neural network
        - The influence from number of hidden nodes
- Preliminary experiment results
    - Experiment setup
    - Experiment result
    - Score on leader board
- Next step

Introduction of data

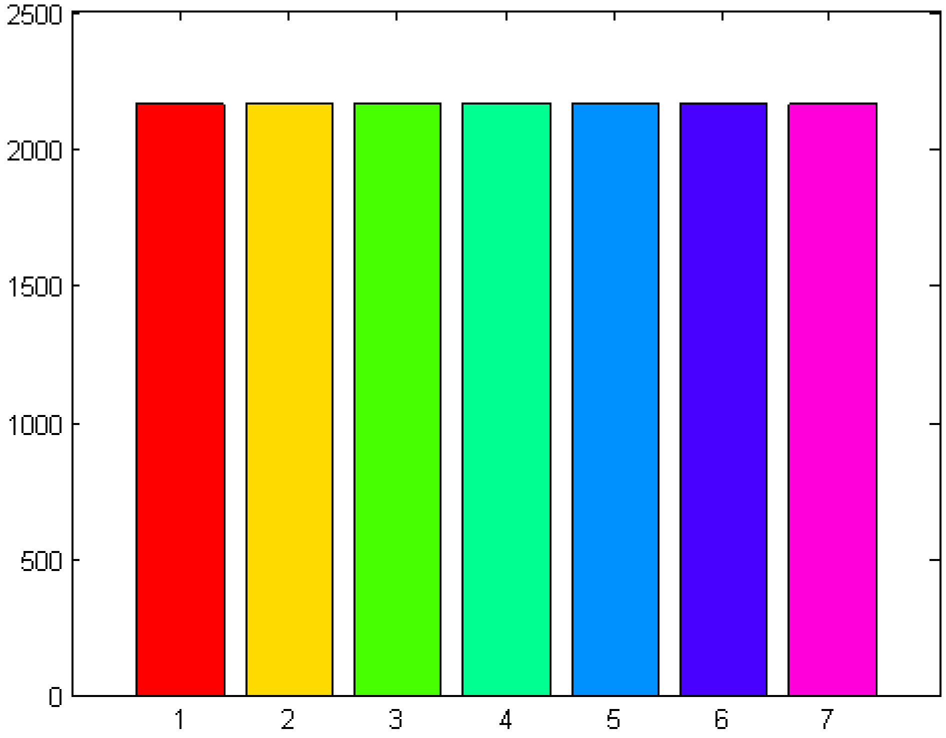
- Sample description
  - 7 Different types of coverage, that we need predict.
    - We convert the result to categorical variable. Use a vector with 7 elements to describe the 7 types coverage.
    - E.g. [0, 1, 0, 0, 0, 0, 0] is the second cover type.
  - The sample size of Training set is 15120
  - The sample size of Testing set is 565892
  - We take features of Continuous data as Numerical variable
    - Elevation, Aspect, Slope
    - Horizontal_Distance_To_Hydrology
    - Vertical_Distance_To_Hydrology
    - Horizontal_Distance_To_Roadways
    - Horizontal_Distance_To_Fire_Points

Sample processing

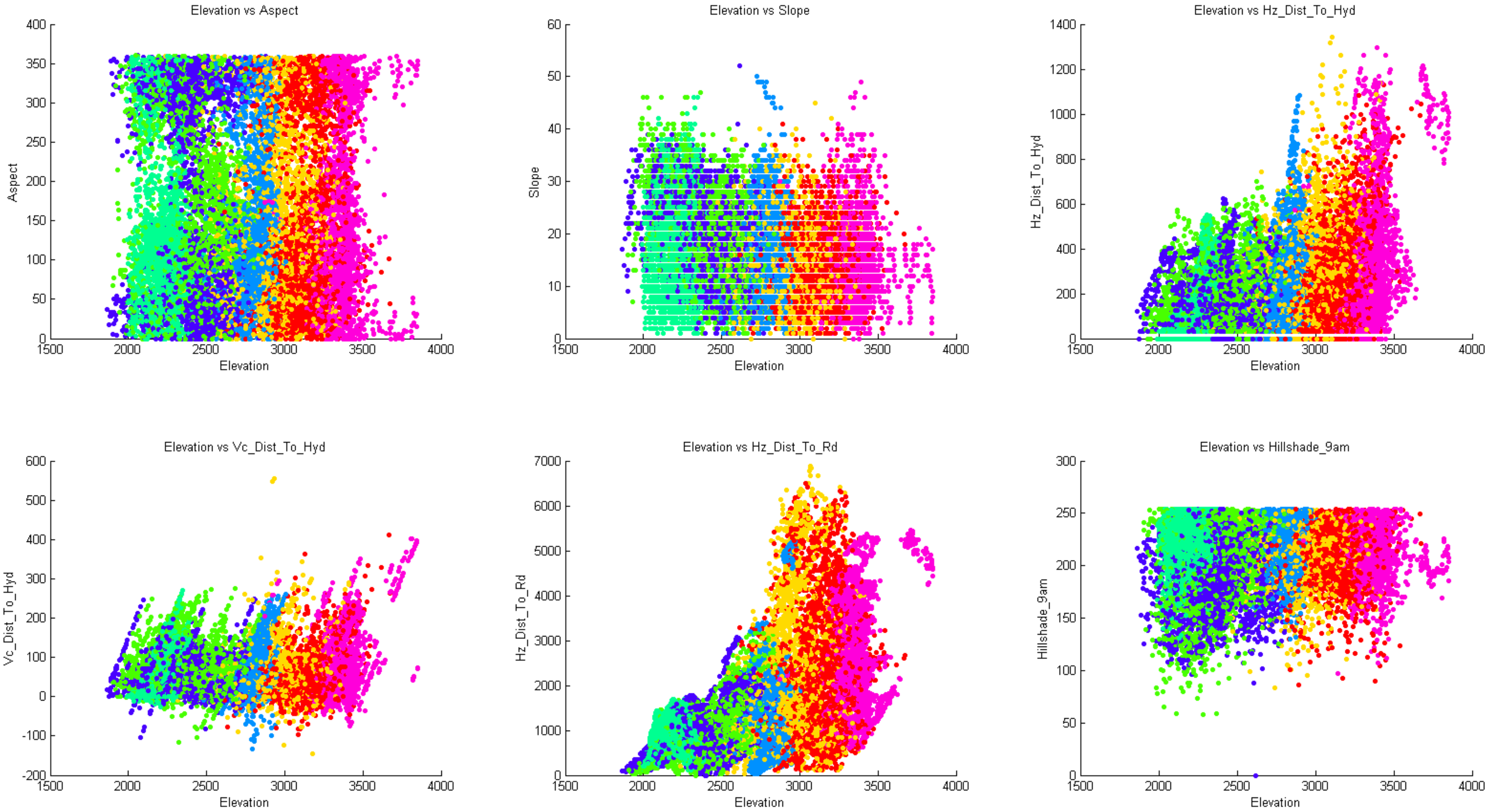
- Categorical data
    - ordinal variables (We assume these are numerical variables)
        - Hillshade_9am
        - Hillshade_Noon
        - Hillshade_3pm
    - nominal variable
        - Wilderness_Area (4 binary vector)
        - Soil_Type (40 binary vector)
        - Cover_Type (7 binary vector)

Training dataset

- Number of samples for each cover type in training dataset.

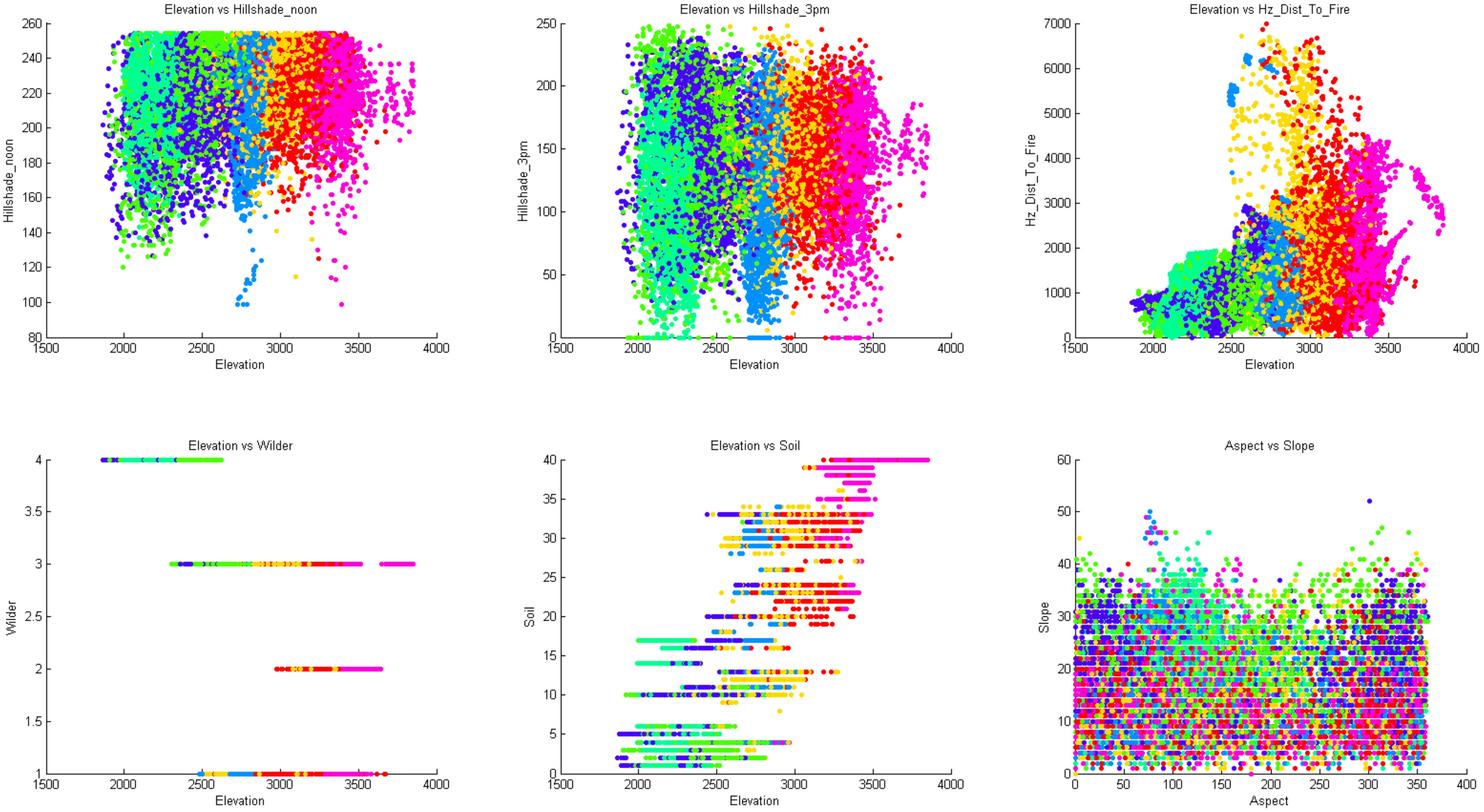


The plot for a pair of features(1/11)



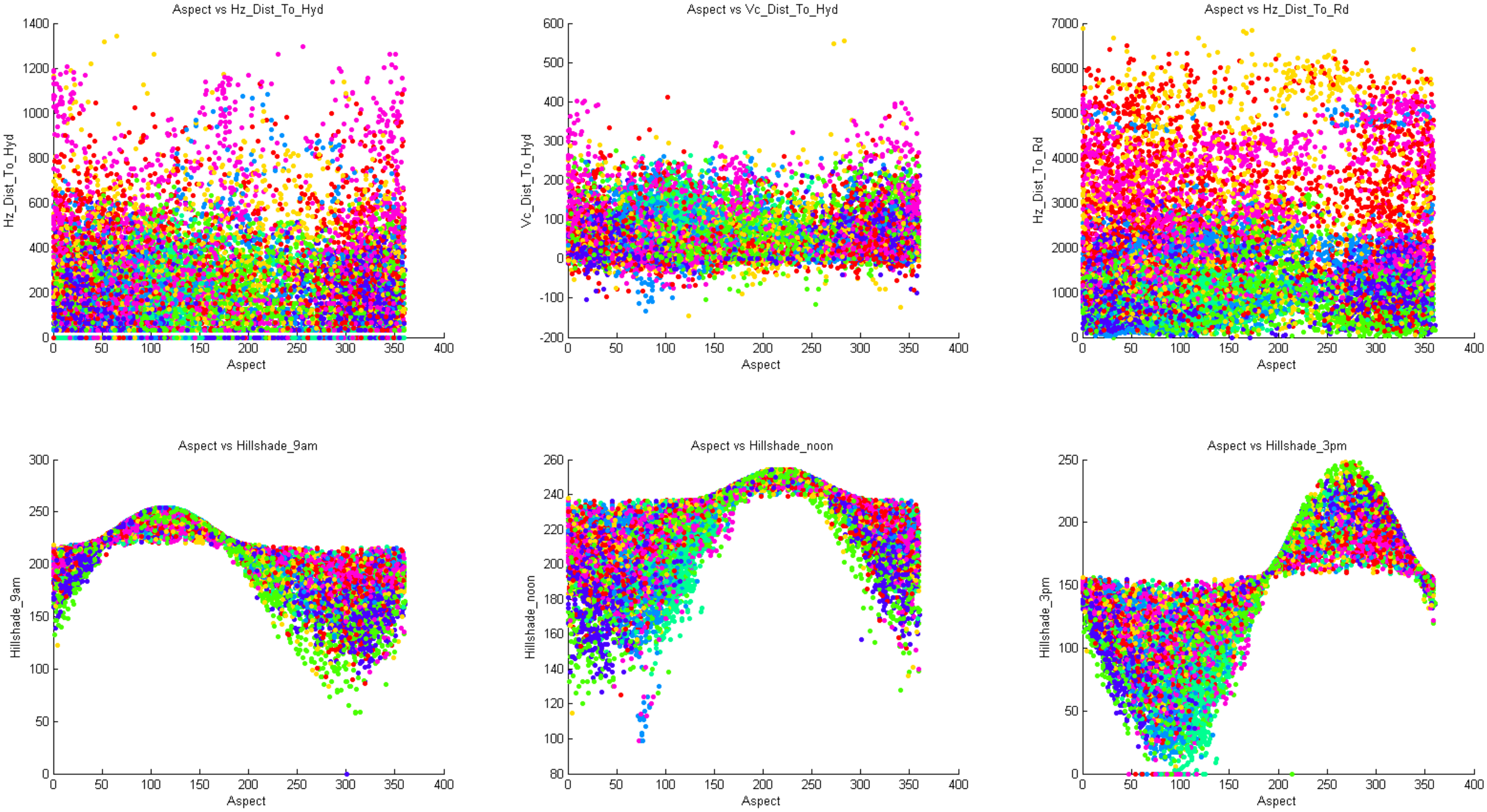
Seven coverages distribute at different scale of Elevation.

The plot for a pair of features(2/11)



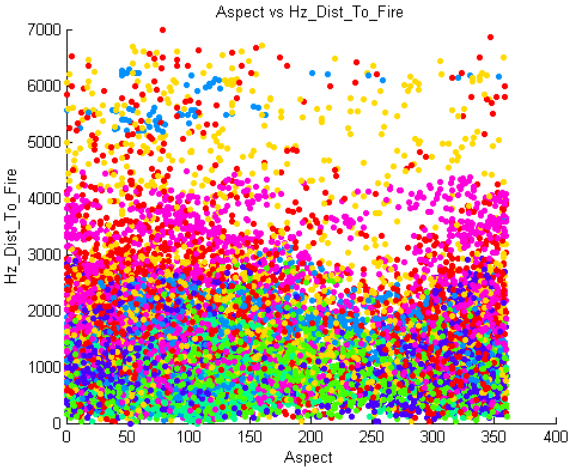
Different Soil types grow different trees.

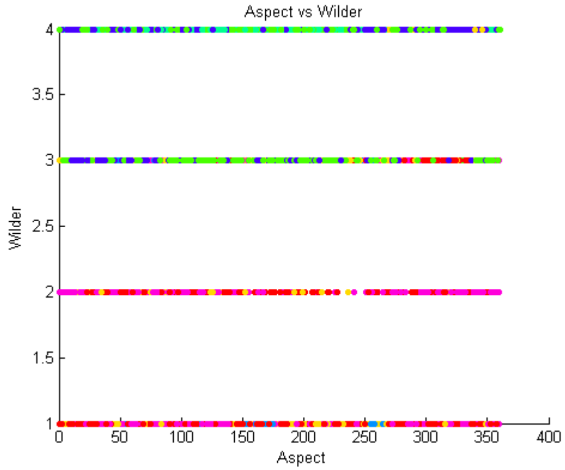
The plot for a pair of features(3/11)

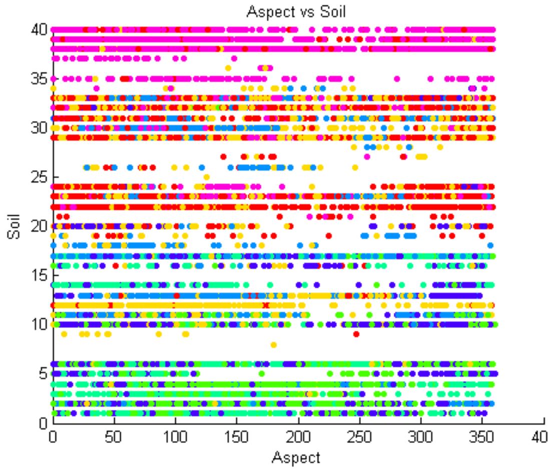


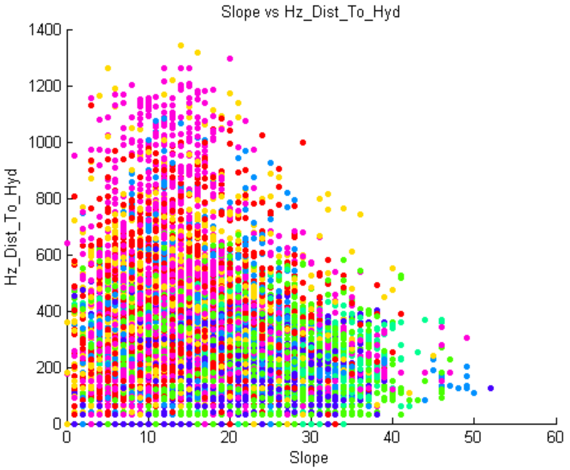
Interesting correlations between Aspect and Hillshades

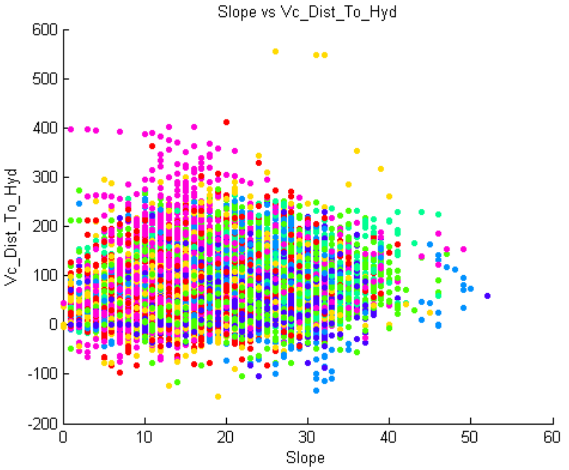
The plot for a pair of features(4/11)

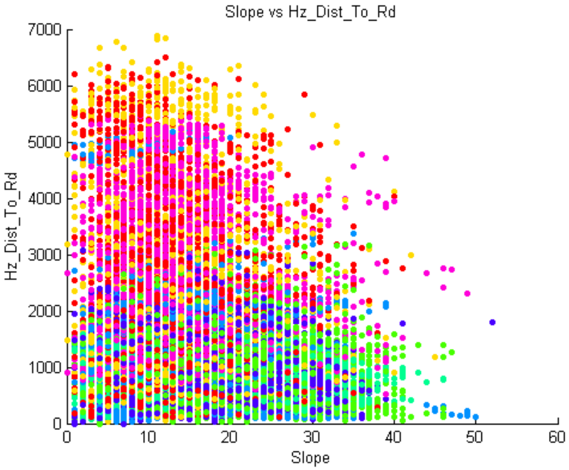

Aspect vs Hz_Dist_To_Fire


Aspect vs Wilder


Aspect vs Soil

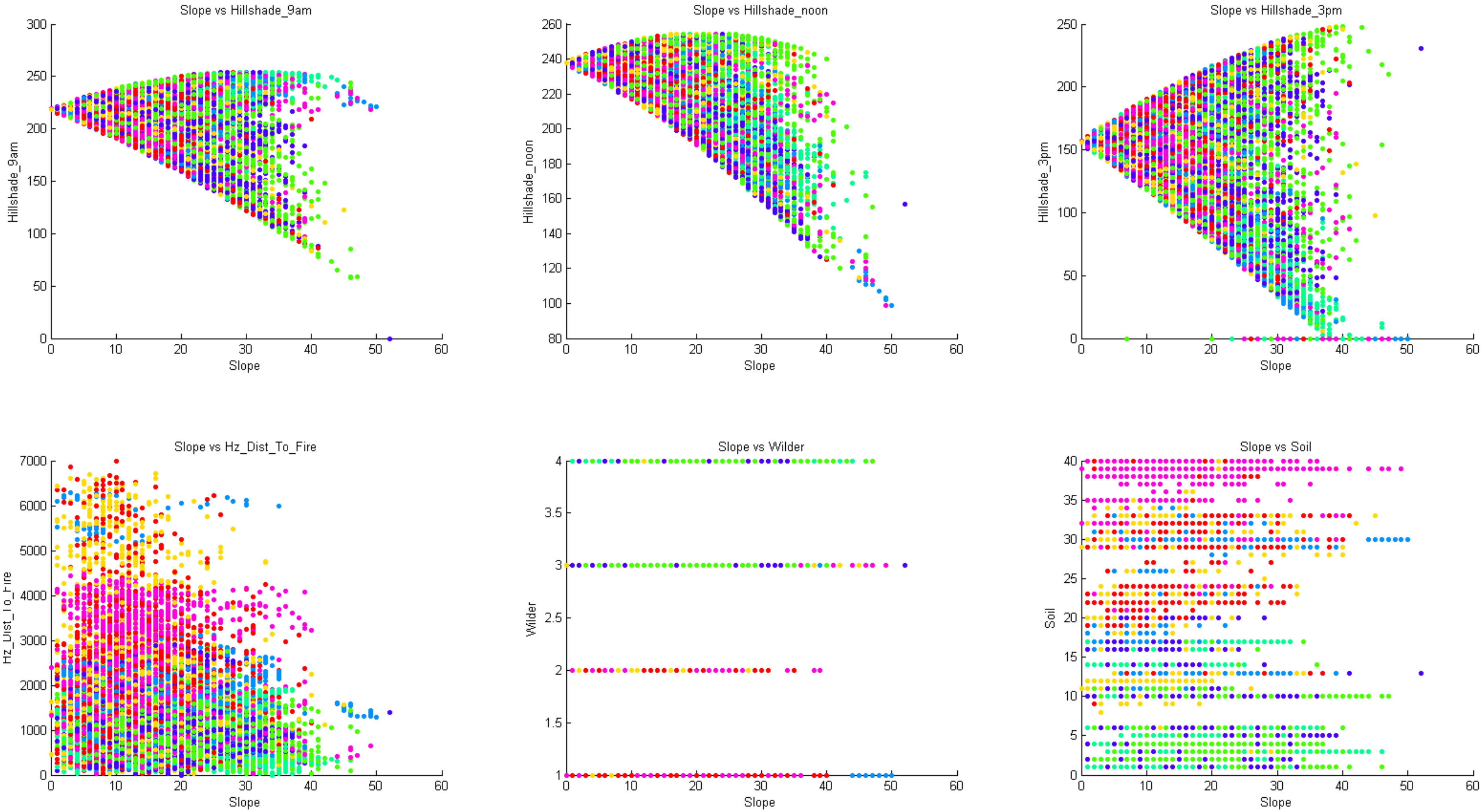

Slope vs Hz_Dist_To_Hyd


Slope vs Vc_Dist_To_Hyd


Slope vs Hz_Dist_To_Rd

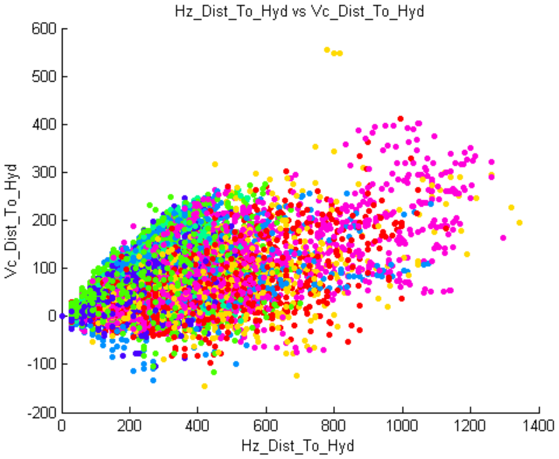
Different types of wilderness and soils grow different trees

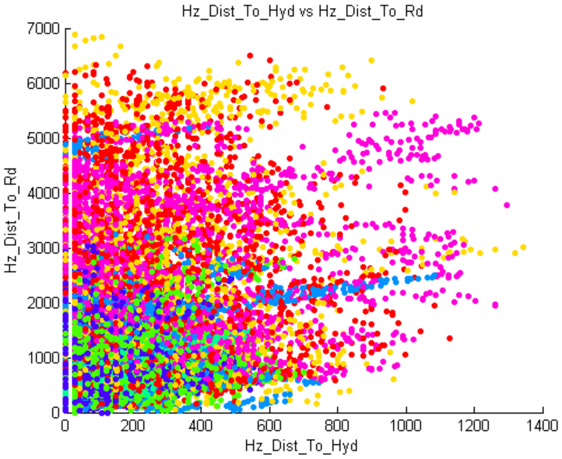
The plot for a pair of features(5/11)

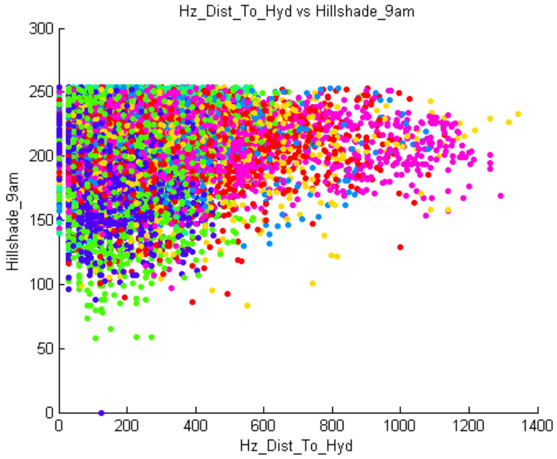


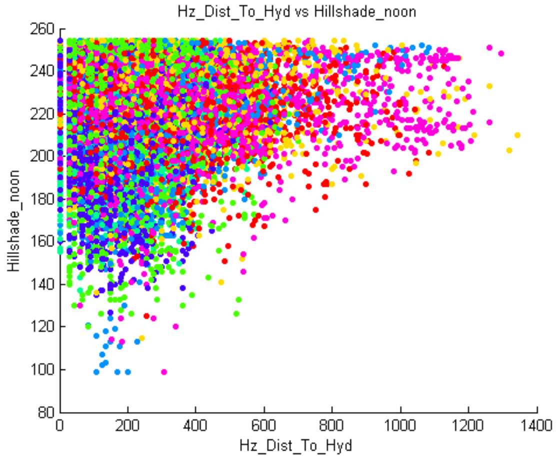
Samples for Hillshade_3pm features have zero values.

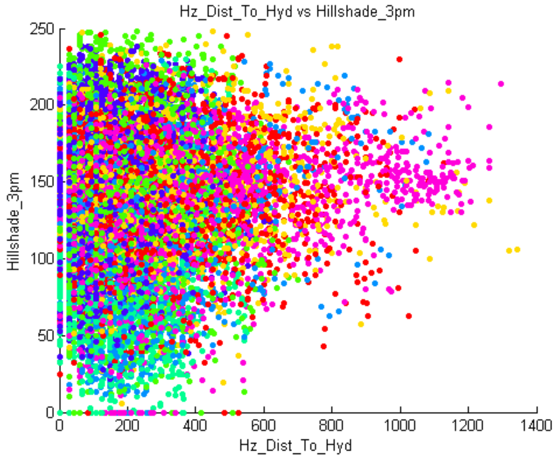
The plot for a pair of features(6/11)

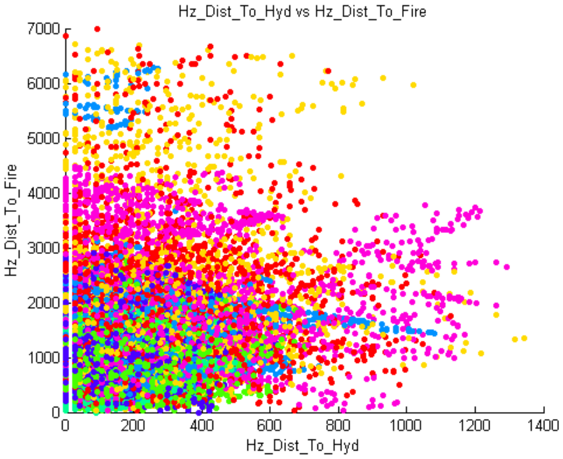

Hz_Dist_To_Hyd vs Vc_Dist_To_Hyd


Hz_Dist_To_Hyd vs Hz_Dist_To_Rd

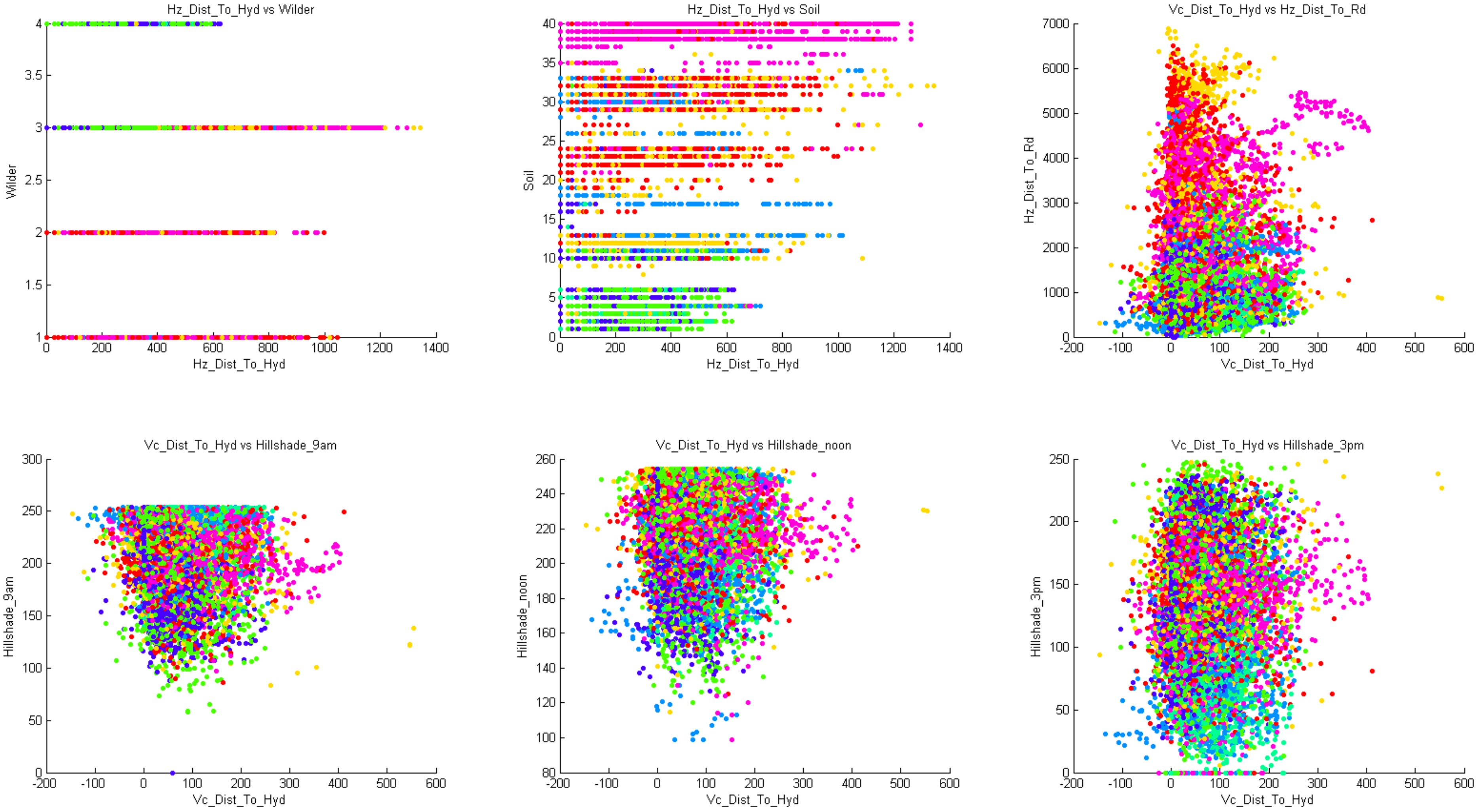

Hz_Dist_To_Hyd vs Hillshade_9am


Hz_Dist_To_Hyd vs Hillshade_noon

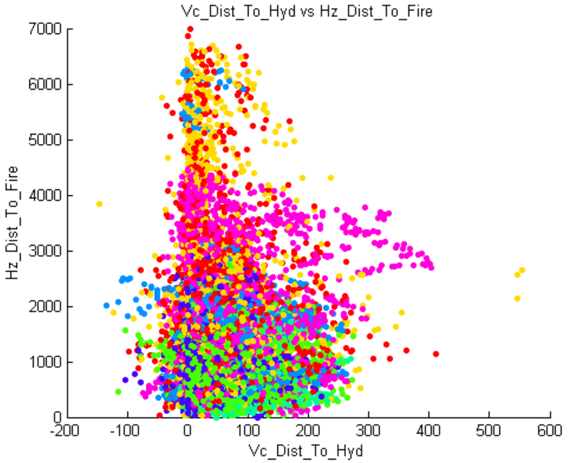

Hz_Dist_To_Hyd vs Hillshade_3pm

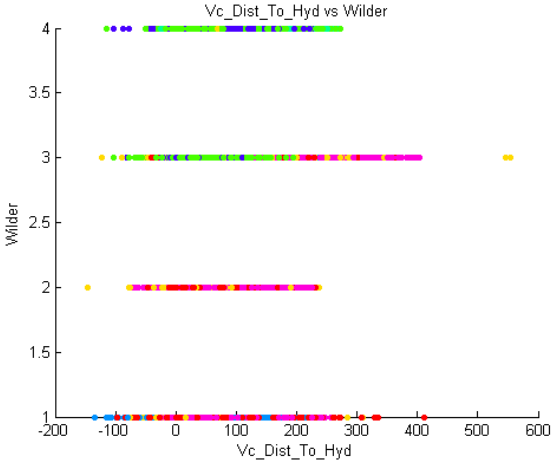

Hz_Dist_To_Hyd vs Hz_Dist_To_Fire

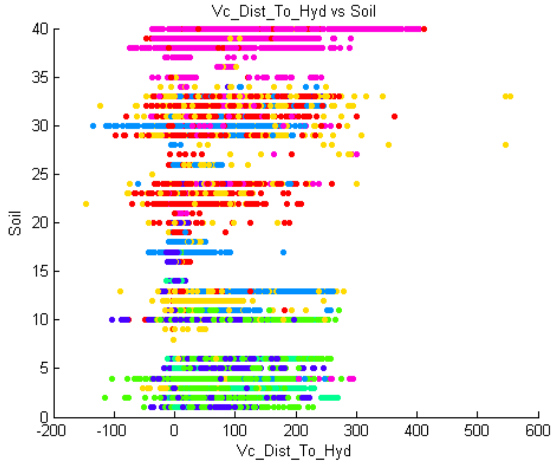
The plot for a pair of features(7/11)

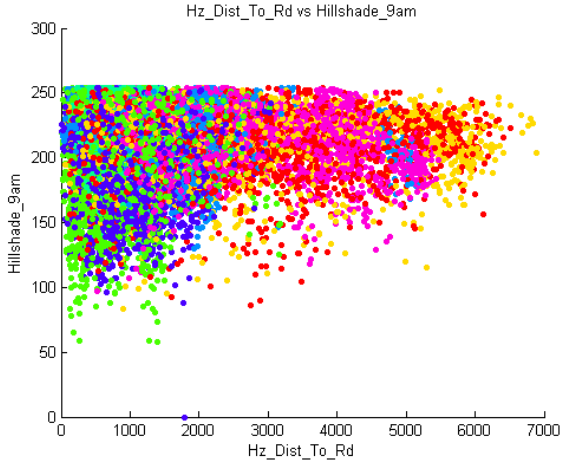


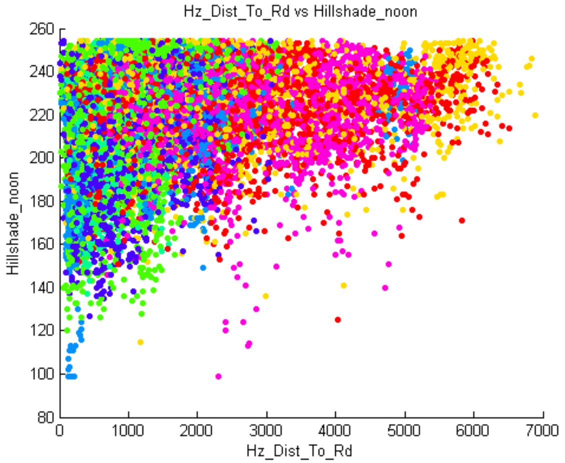
The plot for a pair of features(8/11)

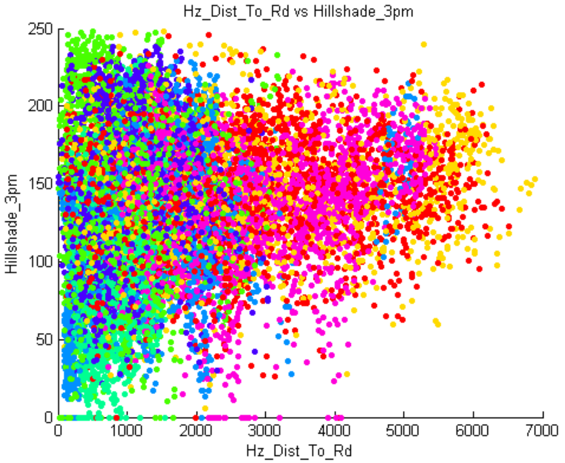




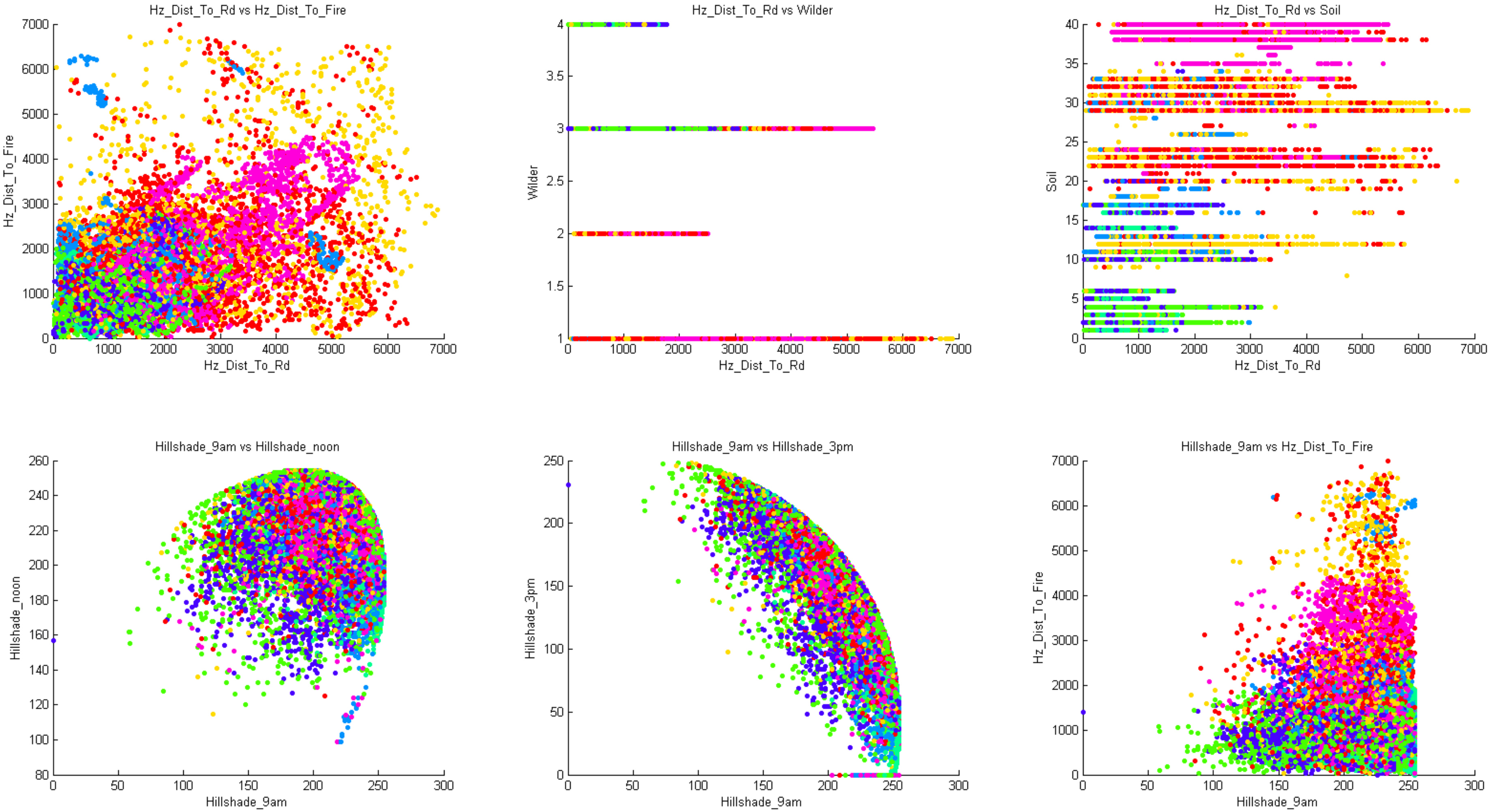




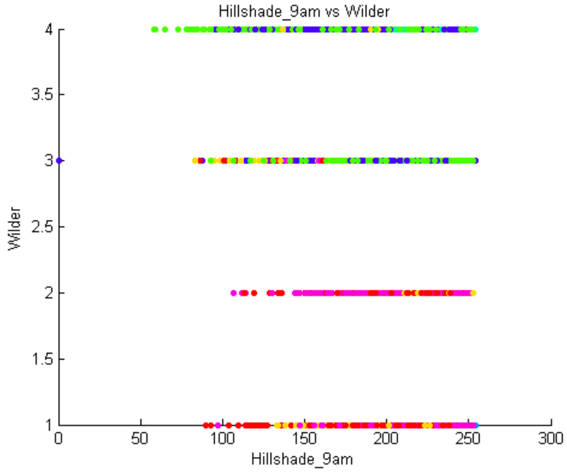


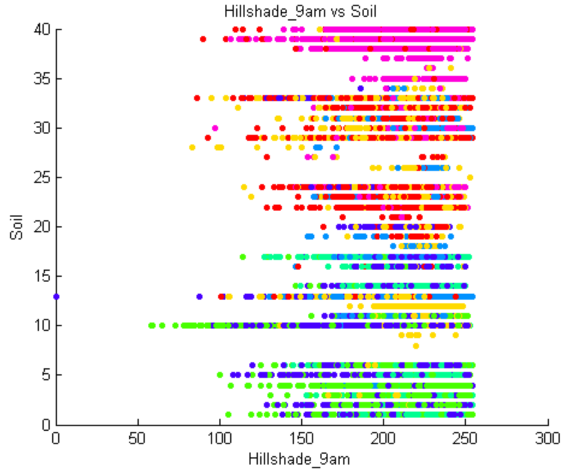


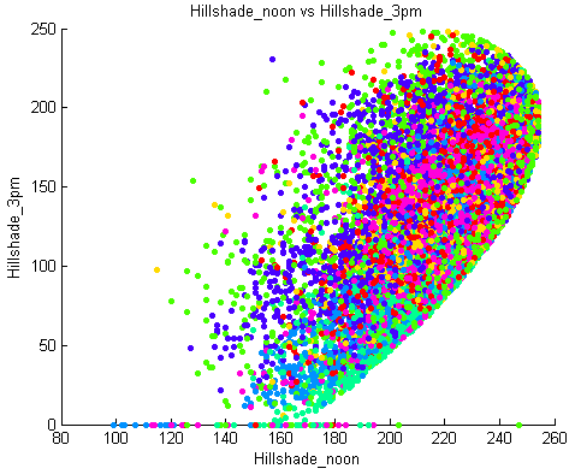
The plot for a pair of features(9/11)

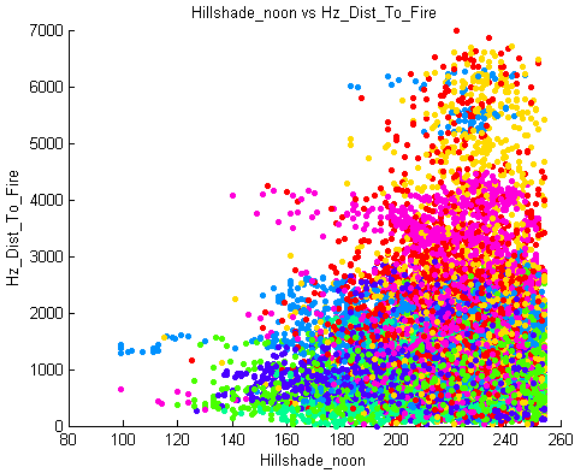


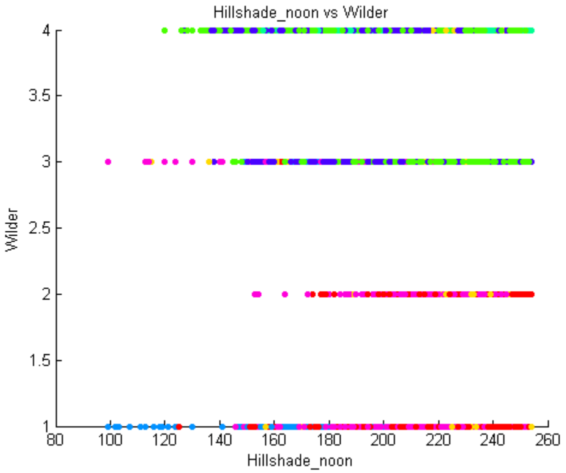
The plot for a pair of features(10/11)

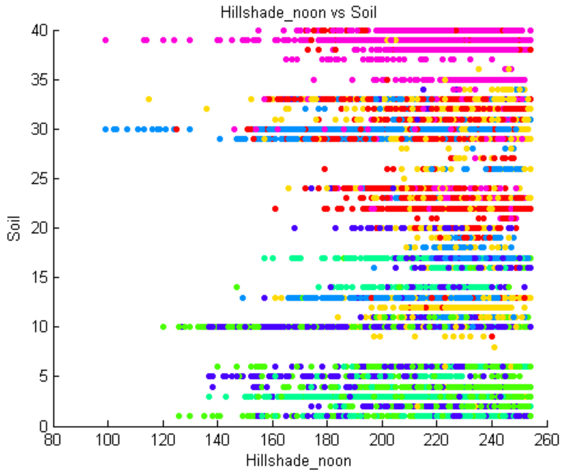




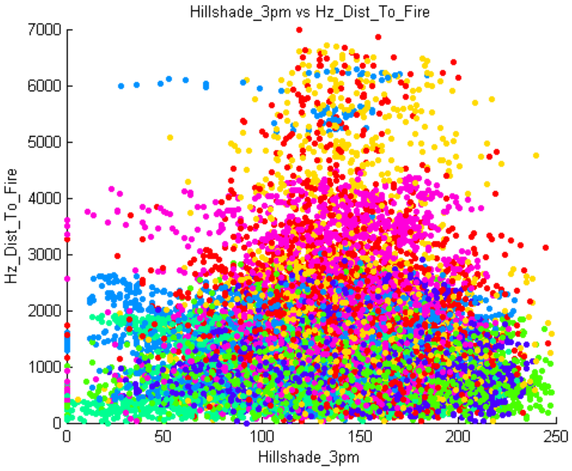


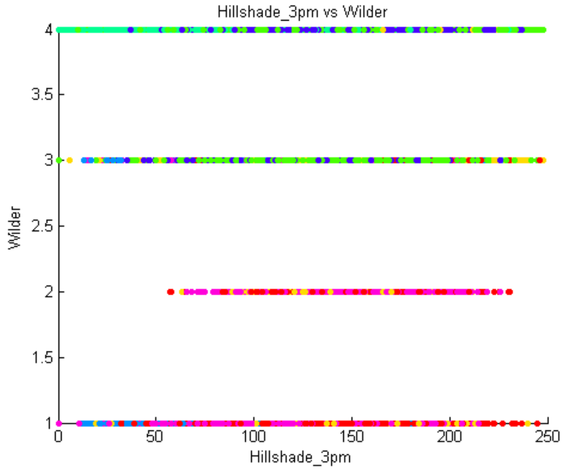


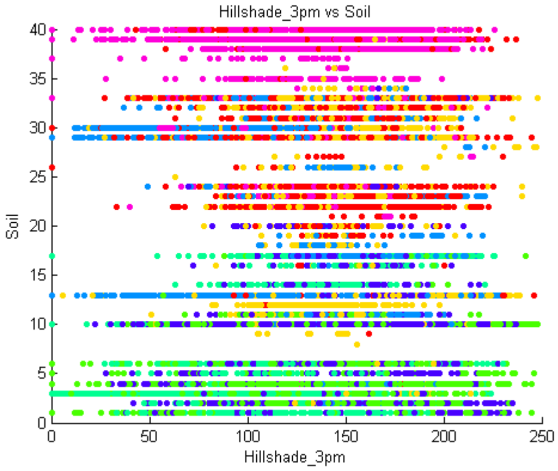


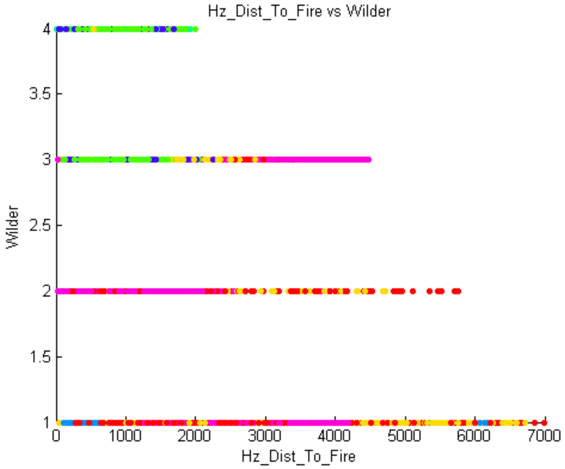


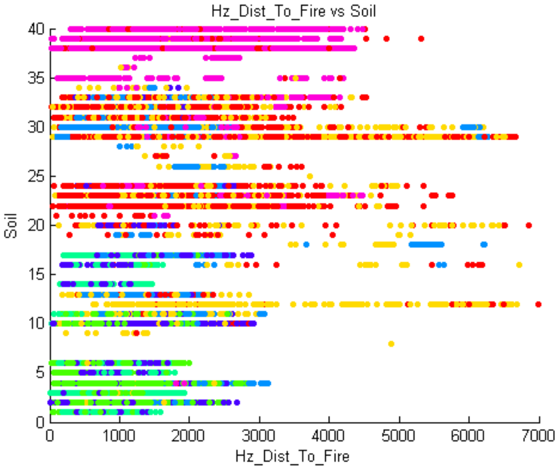
The plot for a pair of features(11/11)

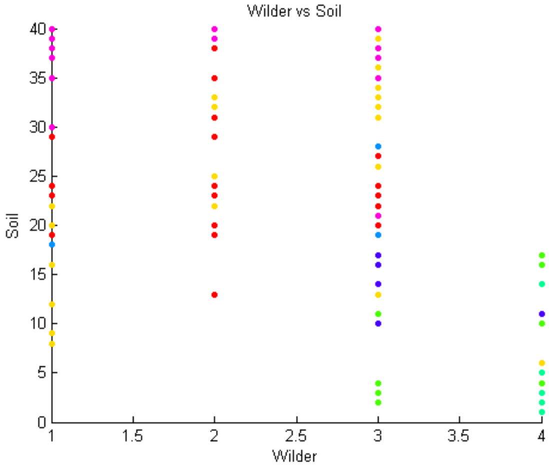






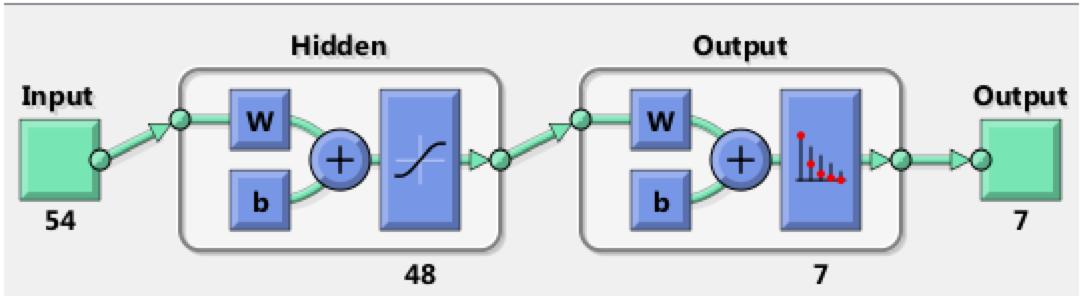






Preliminary experiment

- Experiment Setup
  - We attempt to build a neural network that can classify forest cover type from seven types.
  - The reason we choose neural network is that it a good candidate model for solving problems that has many variables with complicated decision.

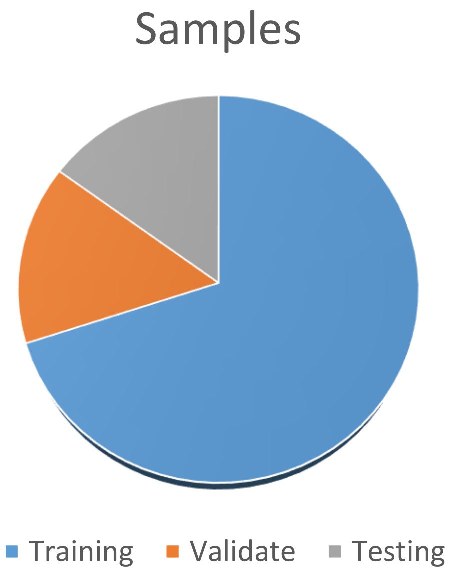


The input layer has 54 features.
The hidden layer has 48 nodes.
The number of nodes in hidden layer is determined by experiments.
The output uses 7 elements to define a cover type.

The result of different number of hidden nodes.

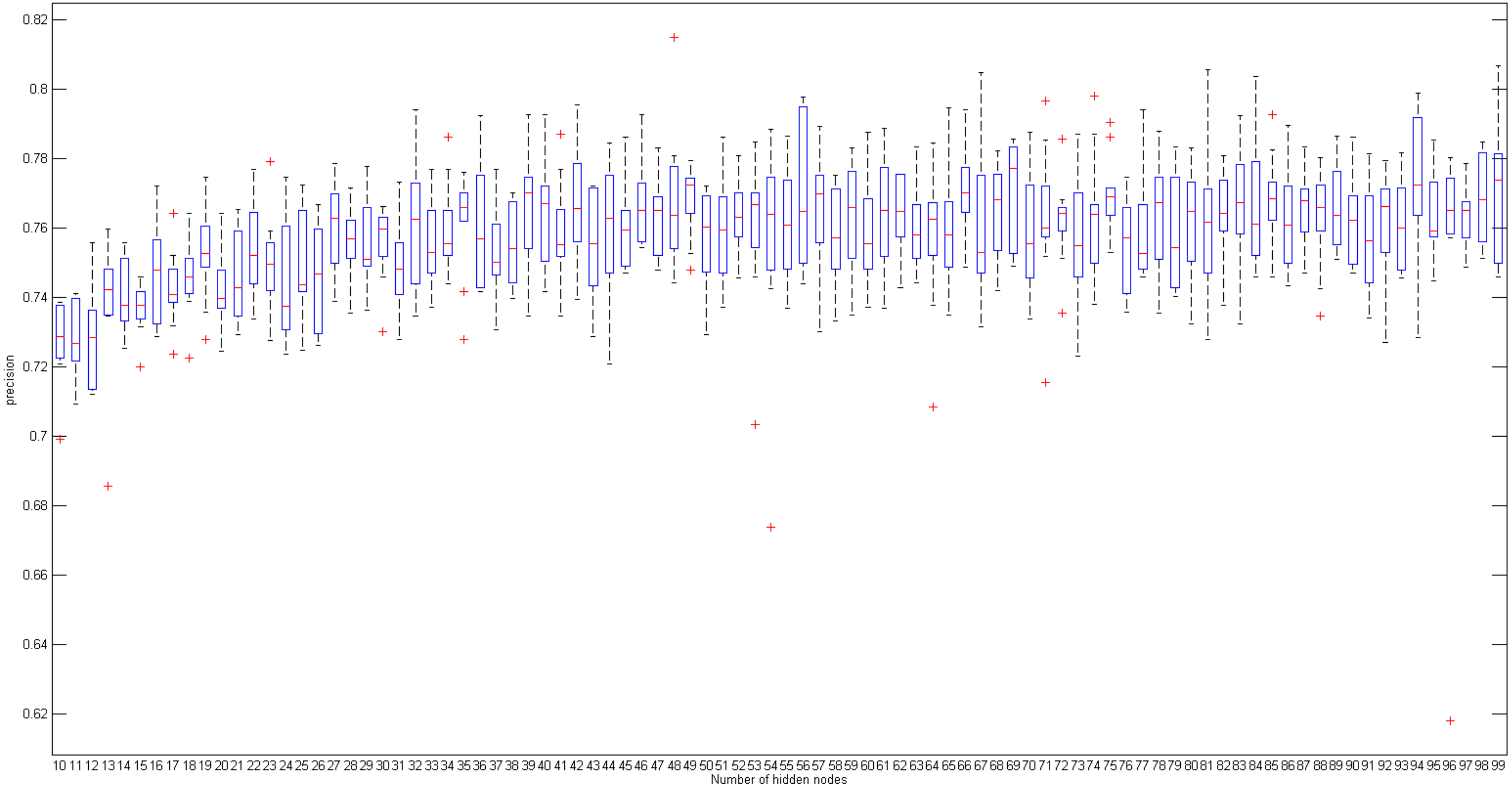
- We do experiment for number of hidden nodes from 10 to 99.
  - For each specific number of hidden nodes, we do 10 times experiment.
  - We divide original training set to 3 different subsets
    - 70% samples for training
    - 15% samples for validation
    - 15% samples for testing
  - Use precision to evaluate result.

$$Precision = \frac{TP}{TP + FP}$$

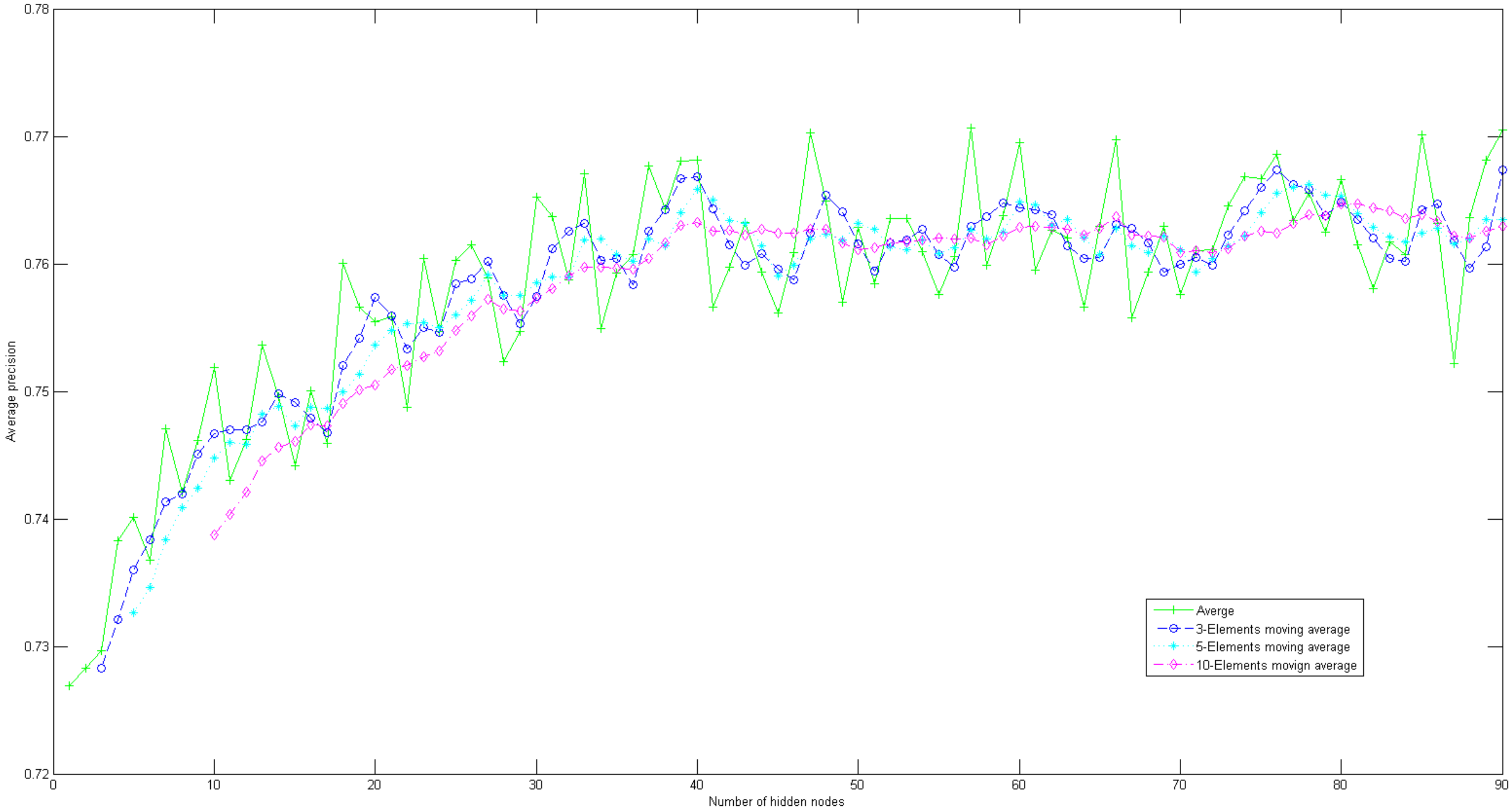


The result of different number of hidden nodes.

- Boxplot of experiment for hidden nodes from 10 to99



Trend of precision

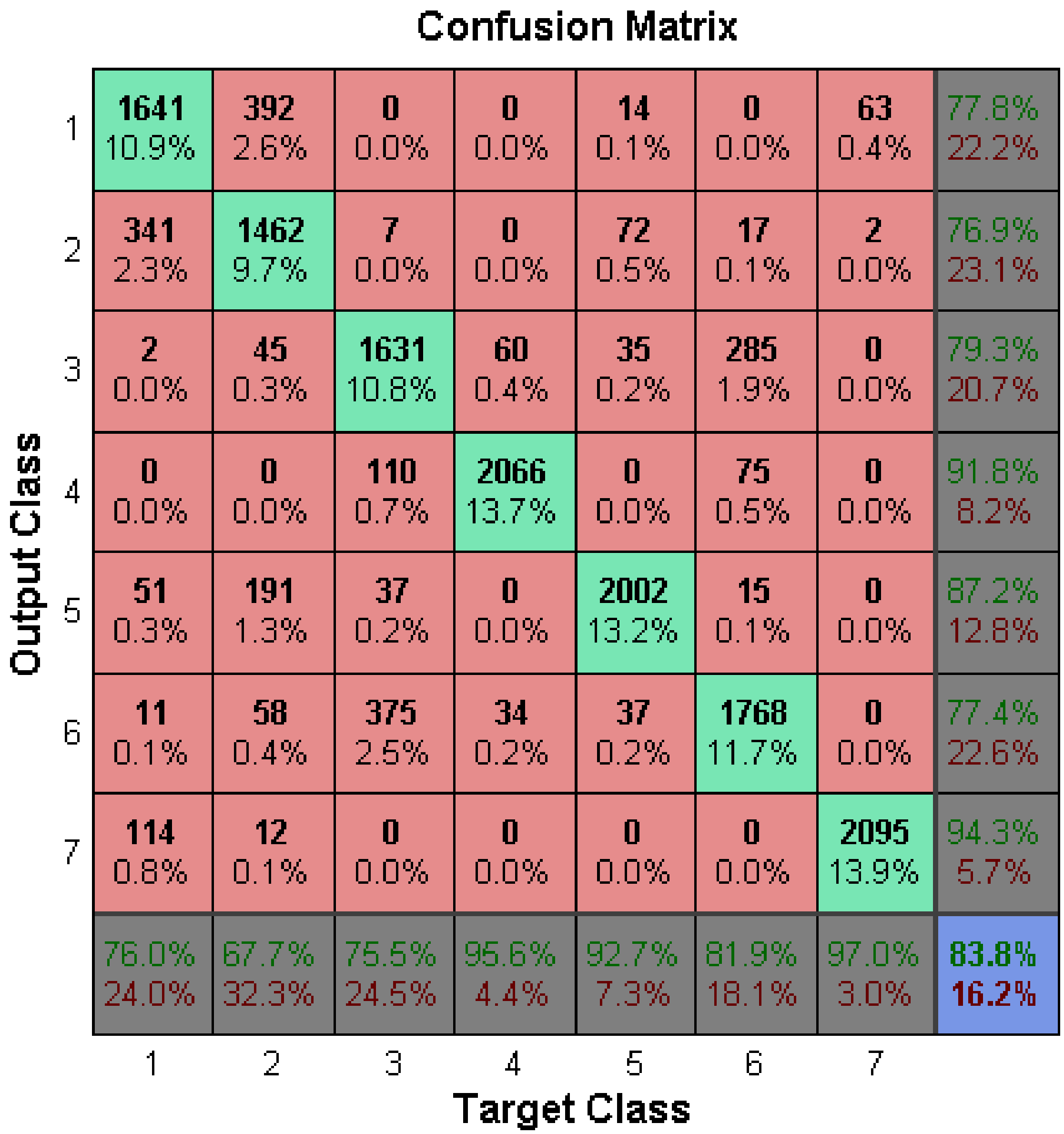


From the plot, when size of hidden nodes is greater than 40, we can get fare good precision.

Top N precisions in neural network. (N=10)

Size of hidden nodes	Precision on training dataset
48	0.8148
99	0.8068
81	0.8055
67	0.8046
84	0.8037
94	0.7989
74	0.7981
56	0.7976
56	0.7971
71	0.7967

Confusion Matrix when hidden layer nodes is 48



Confusion Matrix

Output Class

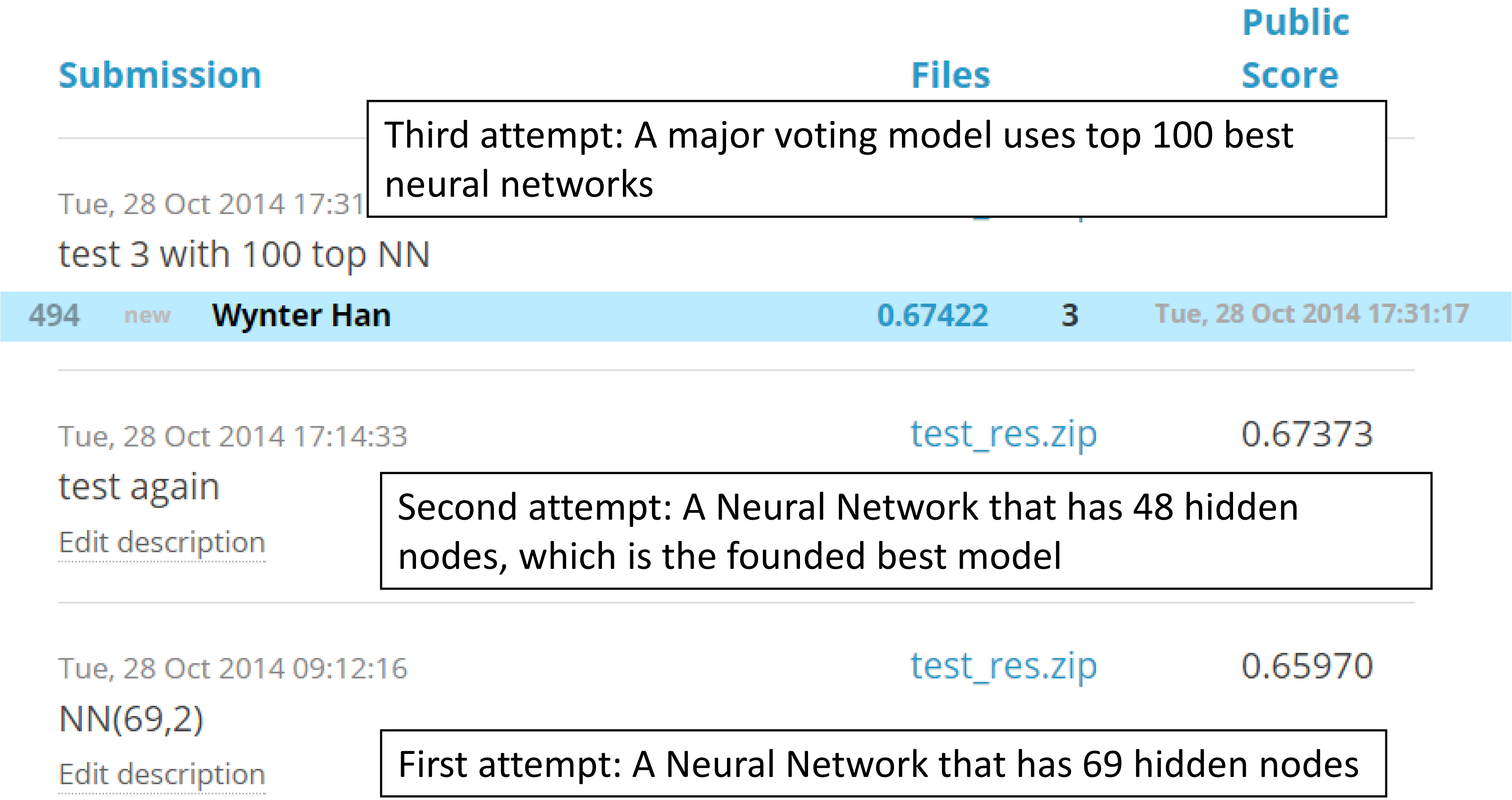
	1	2	3	4	5	6	7	
1	1641 10.9%	392 2.6%	0 0.0%	0 0.0%	14 0.1%	0 0.0%	63 0.4%	77.8% 22.2%
2	341 2.3%	1462 9.7%	7 0.0%	0 0.0%	72 0.5%	17 0.1%	2 0.0%	76.9% 23.1%
3	2 0.0%	45 0.3%	1631 10.8%	60 0.4%	35 0.2%	285 1.9%	0 0.0%	79.3% 20.7%
4	0 0.0%	0 0.0%	110 0.7%	2066 13.7%	0 0.0%	75 0.5%	0 0.0%	91.8% 8.2%
5	51 0.3%	191 1.3%	37 0.2%	0 0.0%	2002 13.2%	15 0.1%	0 0.0%	87.2% 12.8%
6	11 0.1%	58 0.4%	375 2.5%	34 0.2%	37 0.2%	1768 11.7%	0 0.0%	77.4% 22.6%
7	114 0.8%	12 0.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2095 13.9%	94.3% 5.7%
	76.0% 24.0%	67.7% 32.3%	75.5% 24.5%	95.6% 4.4%	92.7% 7.3%	81.9% 18.1%	97.0% 3.0%	83.8% 16.2%

Target Class

Ensemble Methods

- We use top N neural networks to establish an ensemble model. Then we use major vote to get output result.
- Result:
  - Use top 100 neural networks.
  - The precision on the training set is 81.83%
- Discussion:
  - The precision 81.83% is lower than the Top-1 method, which is 83.76%, however, the ensemble methods works better on test set.

Result on Leader board



Submission	Files	Public Score
Tue, 28 Oct 2014 17:31 test 3 with 100 top NN		

494 new **Wynter Han** 0.67422 3 Tue, 28 Oct 2014 17:31:17

Submission	Files	Public Score
Tue, 28 Oct 2014 17:14:33 test again Edit description	test_res.zip	0.67373
Tue, 28 Oct 2014 09:12:16 NN(69,2) Edit description	test_res.zip	0.65970

Third attempt: A major voting model uses top 100 best neural networks

Second attempt: A Neural Network that has 48 hidden nodes, which is the founded best model

First attempt: A Neural Network that has 69 hidden nodes

The best score on leader board is 0.985, my preliminary result gets score of 0.674, still has a large potential to pursue.

Next step

- Use the observed correlation to reconstruct some new features, which may improve prediction model.
- Compare a few different classifiers and choose a better model.
  - K-NN
  - Neural Network
  - Ensemble Model