**Deep Learning Approach For Predicting Crop (Corn) Yield**

**David Joy1\*, Eeshpaul Singh1\*, Madhumita Jadhav1, Rajdeep Arora2****\***

**1** Graduate student at University of Texas at Dallas, TX, USA **2** Teaching Assistant at University of Texas at Dallas,

Abstract

Quantifying the crop yield is essential to make most effective use of policies to ensure food security.  This paper aims at providing quantitative framework for predicting corn hybrid performance in new, untested locations which can help agriculture industry breed better seeds and faster based on its genetics and environmental factors. We tried to develop models that identify robust patterns in corn hybrid genetics in nction with environmental data such as soil and weather that may help scientists more accurately choose seeds that increase the productivity of the crops we plant in the untested locations – and help address the growing global food demand. So, we extracted the features from the three separate datasets such as weather, soil and genetics and yield. We clustered environmental factors into 6 groups based on its properties. Next, we analyzed 2267 corn hybrids in 2122 locations between the year 2008 to 2016 and built the model for predicting corn performance for each cluster.

According to the rules of Syngenta Crop Challenge, for which this research was conducted, we predicted yield using artificial neural network for untested locations across clusters of 6 sub regions and 5 corn hybrid varieties clusters which we derived using K-means Clustering for environment and hybrid variables

*Key words*: K-means Clustering, Linear Regression, Artificial Neural Network

1. Introduction

There are more than 815 million people go hungry despite the fact that there is enough food produced in the world to feed everyone [1]. As mentioned in development agenda by Sustainable Development Goal 2, one of the greatest challenges the world face is how to ensure that a growing global population – as according to the result of the 2017 Revision, the world’s population numbered nearly 7.6 billion in 2017 and will shoot up to 9.7 billion by 2050 [2] - has enough food to meet their nutritional needs. So, to feed another two billion people in 2050, food production will need to increase by 50 percent globally. Today, food security is a complex condition which requires to optimize the food we gain from plans by breeding plants with strongest, highest-yielding genetics. This can be achieved by breeding the stronger plants and selecting the best progenies over time which can be used by farmers to increase crop performance.

As a being bellwether crop, improvement in the performance of corn is one of the hot topic for researchers and scientists. According to the study of Agriculture the major factors which does affects the agriculture performance are genes, environment and the interaction between genes and environment.

Genes are the building blocks to all living things and so does for corn as well. The genes present in a plant affect the productivity of that plant. A gene may influence how tall or short a corn plant is, the length and number of kernel rows, kernel size, its shape and color, or may define the ability to grow in different climates and types of soil, or it may provide resistance to pests. By collecting such a genetic data for the plants that scientists develop, they can make predictions about a plant’s productivity. There are nearly 19,500 unique genetic markers provided in Syngenta challenge and the average number of markers assembled per hybrid is approximately 12,000. As part of the challenge was to determine which genetic marker are useful.

In addition to genes, another factor affecting performance of corn is environmental condition such as weather and soil. As per the study published in the Nature Communication indicates that climate variability accounts for 32–39% approximately of the observed yield variability globally [3]. The relationship between climate variability and crop yield variability, highlighting where variations in temperature, precipitation or their interaction explain yield variability. Plants needs balanced amount of water and sunlight. As, too much rain can cause disease or flooding. Or too much heat, especially in the absence of rainfall, may leads to decrease in productivity. The type of soil also has an effect on a plant. For example, if a plant is grown in soil that is able to hold more water than average, it will be able to better withstand an extended period of low rainfall. By characterizing the environments that plants are grown, we can better understand how plants react to the different environments. Scientists do this by precisely measuring the weather and soil in all growing locations. In case of corn, it is adapted to grow in a warm season, the emergence and germination are optimal when soil temperature between 85 to 90 F [4]. Moist and warmer conditions during first 1 to 2 days after planting can mitigate much of the cold stress during imbibition [4].

The last but most important factor is interaction of environmental factors and genes. As per the Darwin’s theory of evolution and as a property of genetic by environment (GxE) interaction [12], environment activates certain genes that allow the plant to thrive in the environment. There are multiple factors that affect the plant which adapts to grow best in a particular region such as length of growing season, rainfall, temperature, precipitation, soil type, solar radiation and many more. Some plants may tolerate drought better than others. Some plants may prefer a soil that is sandy while others prefer clay. Each genotype is responding to environmental variation in a different way. These interactions are often quite complex, involving multiple genes and multiple facets of the environment.

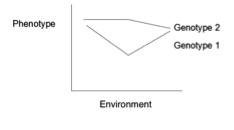


Figure 1: This Norm shows lines that are not parallel indicating a gene by environment interaction.

Syngenta Crop Challenge

The goal of Syngenta Crop Challenge [11] is to create a data driven model that predicts the general performance of a hybrid of a corn under uncertain conditions in untested environment so that breeder can make better decision on which hybrid should be provided to grower for increasing productivity to meet the world’s growing demand.

To find the best performers in the hybrids, the experimental hybrids are planted in diverse environmental locations and kept under meticulous observation on the yield of a corn. The testing place across years from 2008 to 2016 with the goal of advancing the best hybrids each year to test again and this is achieved by Genetics Environmental Interaction (GxE) [12]. However, the number of locations where breeders can test are limited and can cause uncertainty while choosing best hybrids for growers to plant due to lack of observations. So, goal was, how changes in location with varying environmental conditions affects the potential yield of corn. based on the historic observations and genetic compositions.

Data Set:

The Data is divided into three types one training dataset, testing dataset and validation dataset. The training data includes all the current knowledge of 2267 experimental hybrids in 2,122 of locations between 2008 and 2016. The training data itself compiles with three separate datasets,

1. Performance
2. Environment (Weather & Soil)
3. Genetics Dataset

**The performance dataset** having year, hybrid name and yield difference from check where each row indicates the performance of one hybrid at a given location and a year. The environment dataset having features like latitude, longitude, weather information, soil information and the genetic dataset containing near about 19,600 genetic markers.

**Environment Dataset** contains the historical weather and soil conditions for our selected growing region from year 2008 to 2016. Location ID is a unique latitude and longitude combination. Across the growing region, differences in weather conditions and soil types will cause variation in a hybrid’s observed performance.

**Validation dataset:** The validation dataset contains performance information including latitude and longitude, environmental information including weather for previous years and soil conditions for test locations, and the genetic marker information for the hybrids. It has been used to check the performance of the training models using experimental hybrid of 2016

**Test dataset** includes the experimental hybrids tested in 2017. This set will include almost all information as like in training dataset except the yield variable which is the dependent variable or the variable which we have to predict.

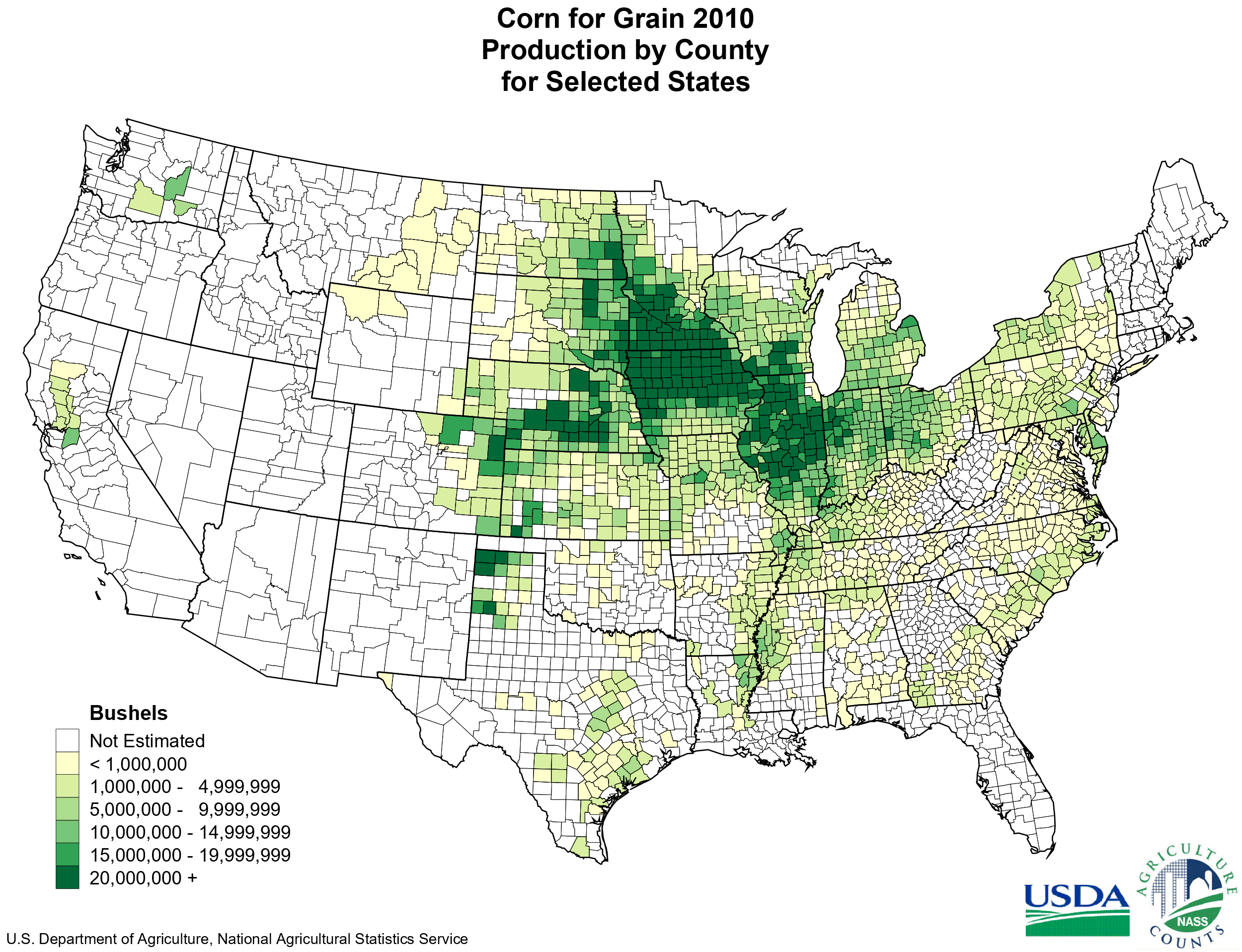


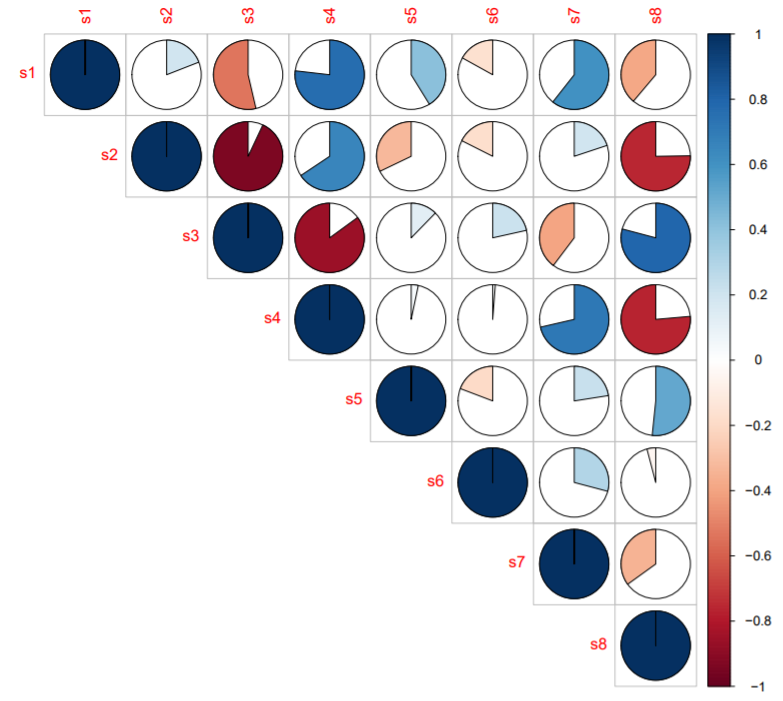
Figure 2: Corn Belt - Location of Experiment and Its performance for the year 2010 [13].

In Figure 2, Corn Belt connotes the United State’s Midwest area dominated by farming. And this is the location of experimenting corn hybrids for many researches which is a part of Syngenta Challenge Dataset.

1. Analysis of data

Genetic information for the experimental hybrids can be useful for predicting how a hybrid will respond in various environmental conditions. Genetic Dataset represents a pair of nucleotides A, C, G, or T. The row defines the hybrid and the column defines a specific position within a chromosome or the marker position. The reason it is a pair, and not a single nucleotide, is that each hybrid has two of each chromosome, one from each parent plant that produced it.

For each marker position, there are typically 2 possible values that can appear for each chromosome, which results in 3 possible values in total. For example, one marker position could contain the nucleotides A or C. Thus, the pair of chromosomes could be AA, CC, or AC. Here AA and CC are considered as homogeneous pair, whereas AC is considered as heterogeneous pair. In every data point values AA indicated as 1, AC as 0, and CC as -1 and the rest of the hybrid that was not tested are indicated as NA. After this conversion, mathematical operations can be applied to extract value from the data, such as defining similarity between pairs of hybrid.

  
Figure 3: Correlation Plot of Soil Variables

The figure 3 indicates correlation between 8 types of soil features provided by Syngenta as a part of the challenge. Here the red pie chart for s2 and s3 variable indicates, feature s2 and s3 are negatively correlated where as s1 and s4 indicates positive correlation between the soil variable.

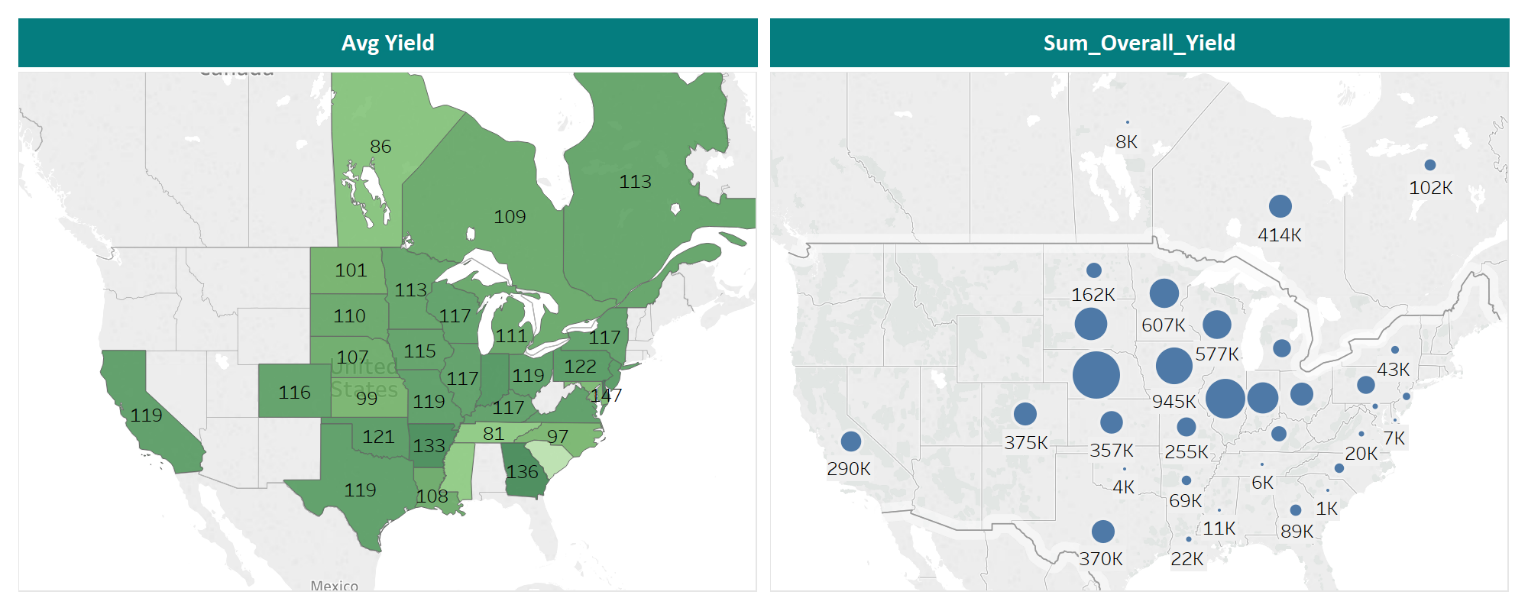


Figure 4: Overall average Yield and Total Yield across United States

To get the more insights we created clusters of regions based on its environmental factors such as soil and weather using k-means Clustering algorithm. We use Tableau as a platform to display the results and highlighted the numbers to show the relation between them.

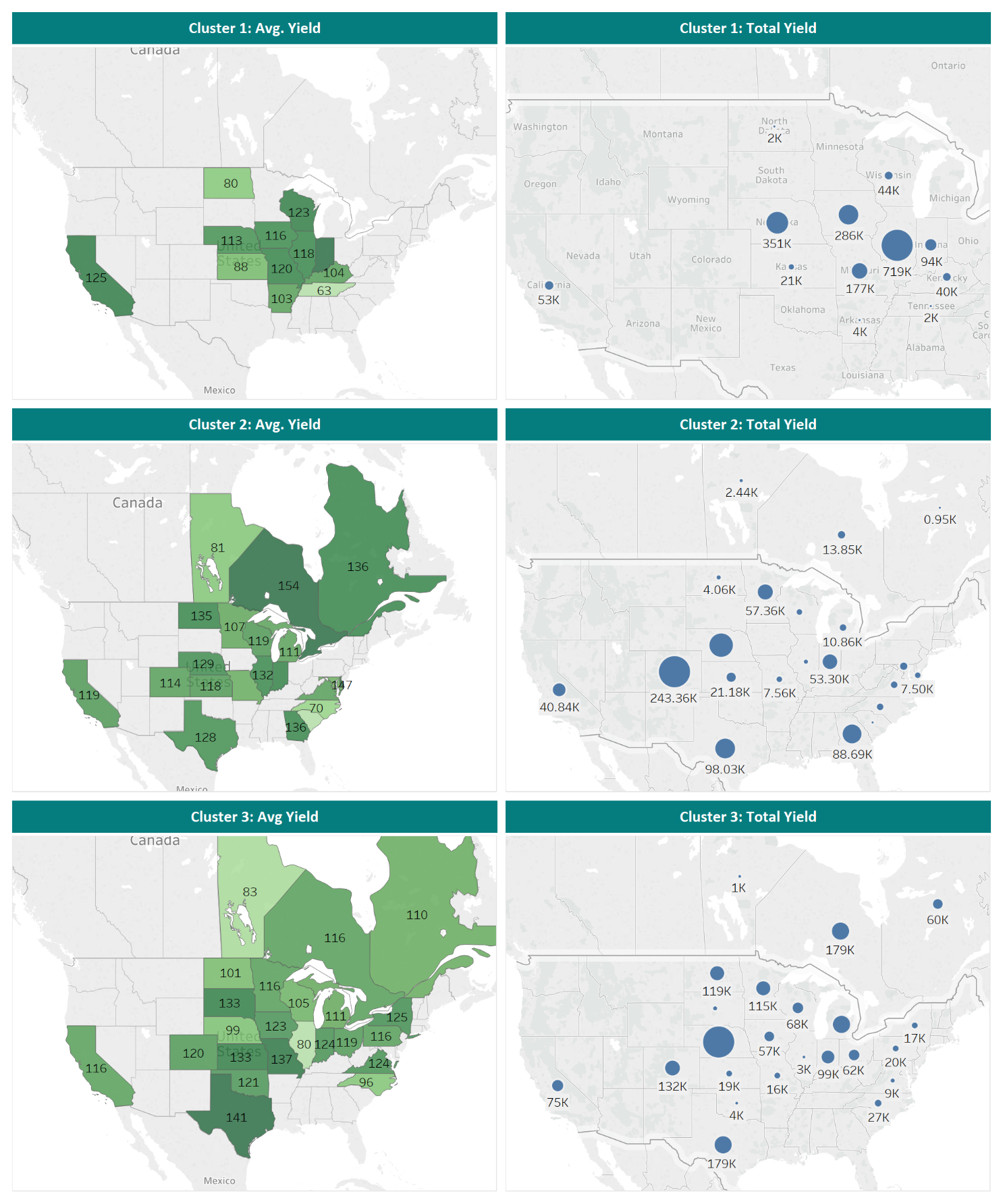


Figure 5: Average Yield and Total Yield across cluster1, cluster 2 and cluster 3

The above results show the difference between clusters in terms of average yield and total yield. Larger is the circle, more is the yield of that particular location.

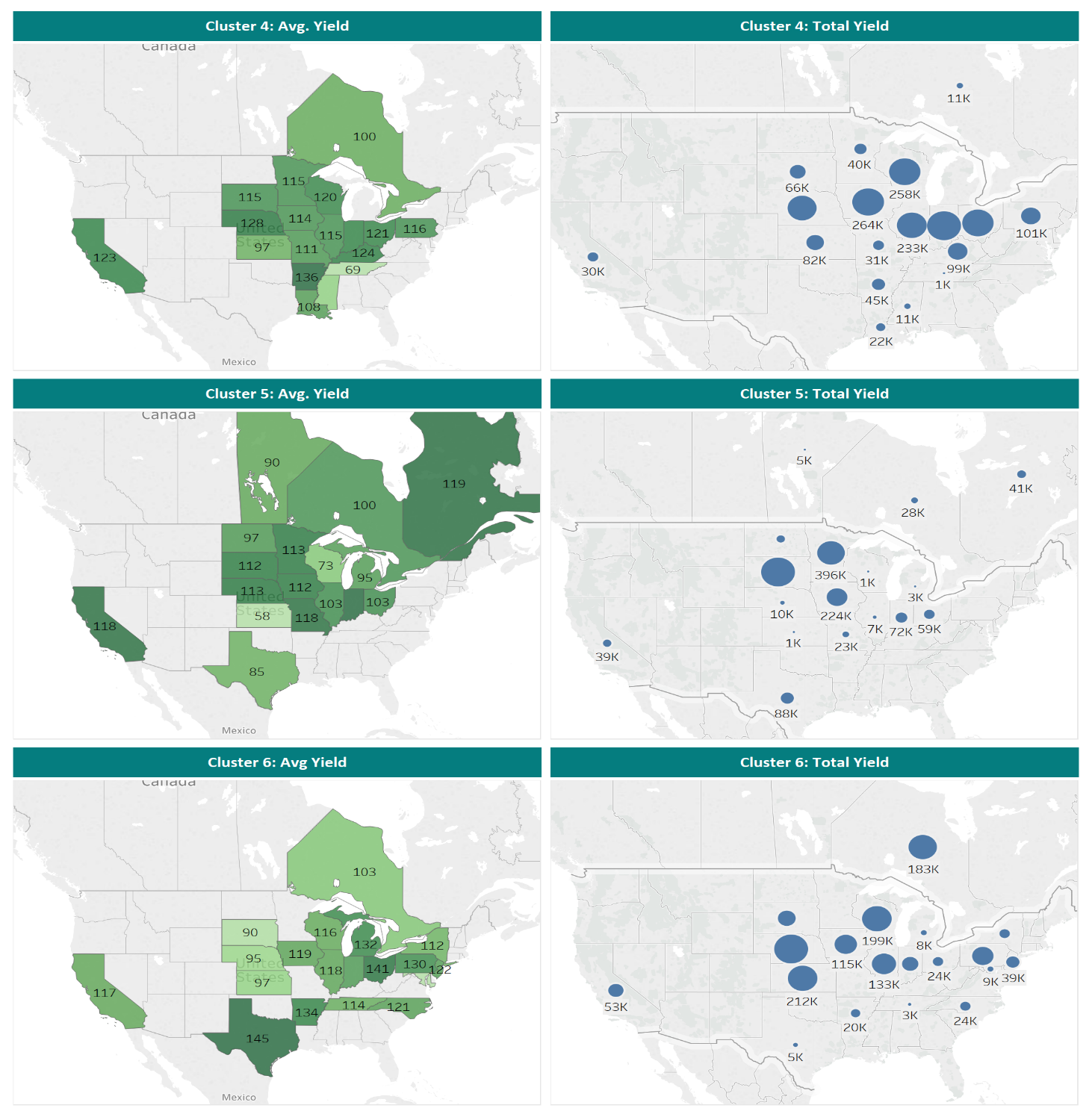


Figure 6: Average Yield and Total Yield across cluster1, cluster 2 and cluster 3



Figure 7: (a) Figure 7: (b)

Figure 7 (a): Corn Hybrid Performance across United States’ Testing locations & Figure 7 (b): Top 15 Planted Hybrids in United States’ testing locations

Above figure shows the performance of the hybrids across different testing location. Darker is the color higher is the yield difference.

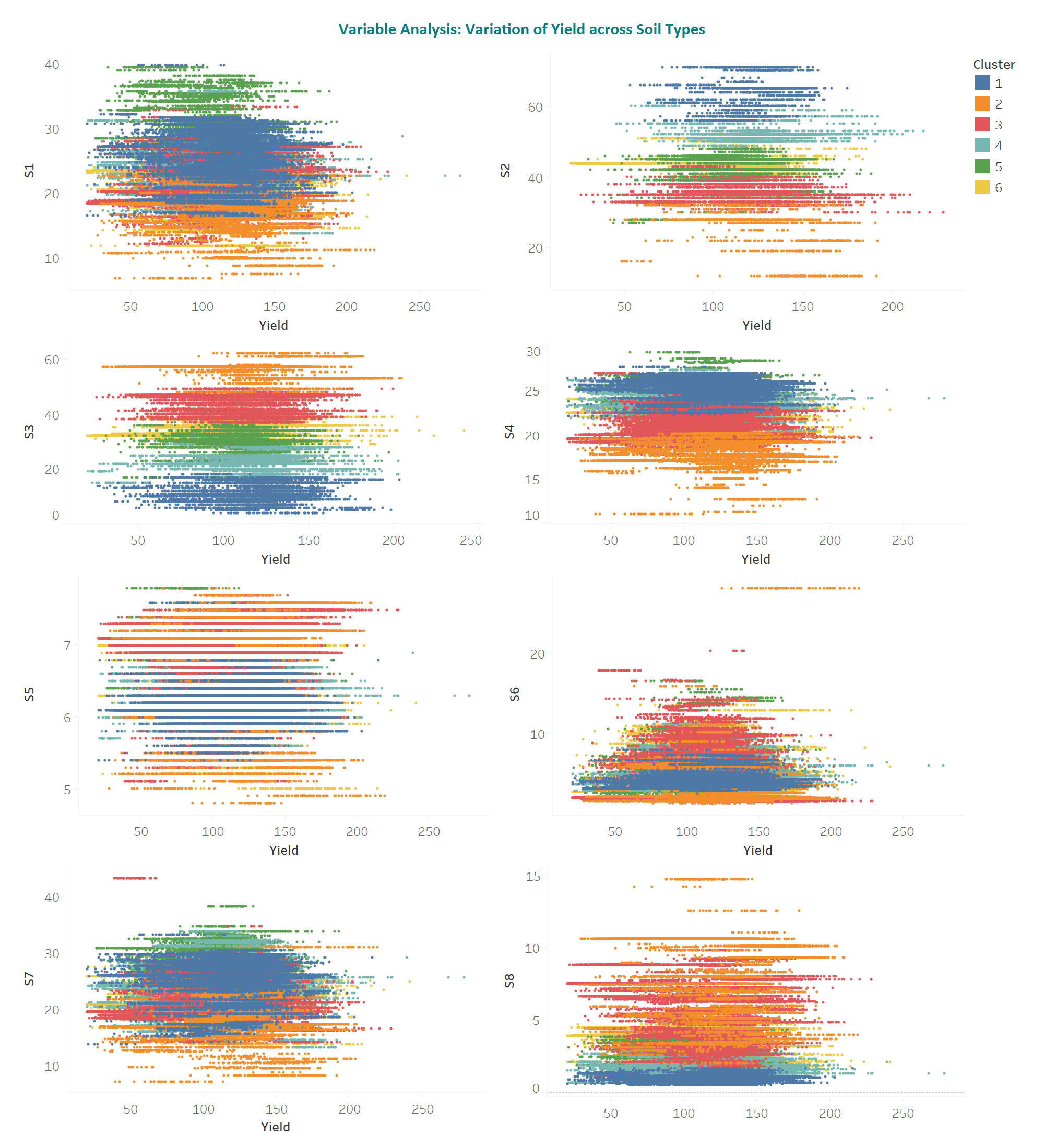


Figure 8: Variation of Yield across Soil Types.

1. M-ethodology and theory

Successfully modeling the effects of corn hybrid in environment will require careful consideration of many topics, including the best way to aggregate the weather and soil data, the possibility of regularizing the genetic data or clustering plants by their genetics, and which machine learning algorithms to use.

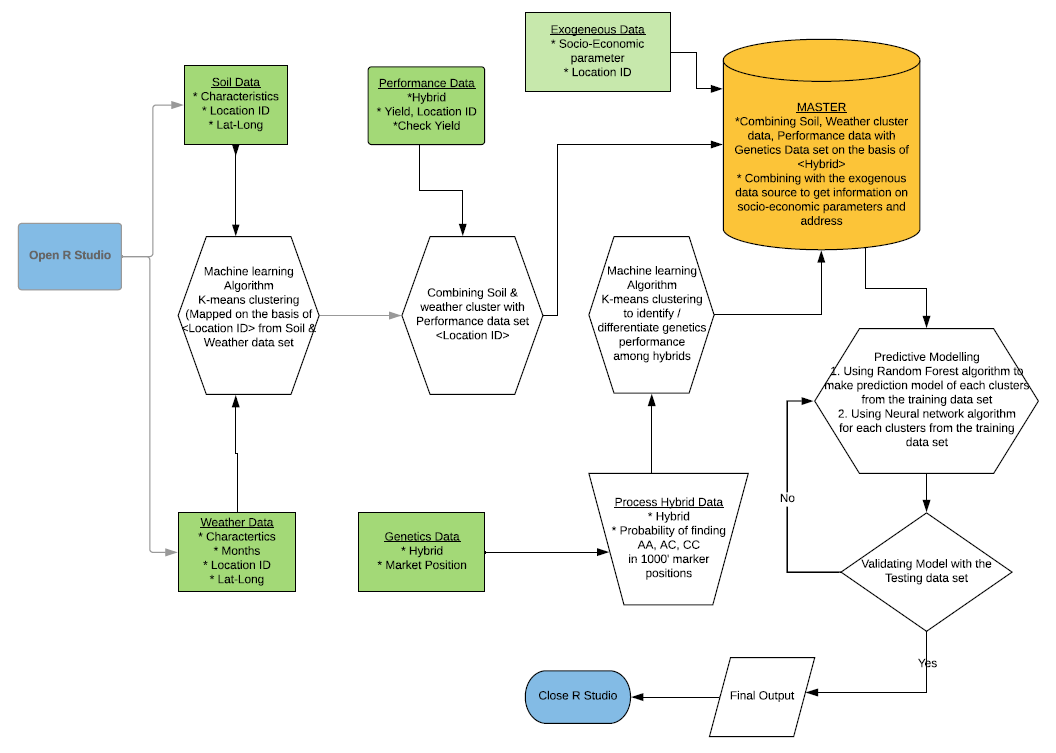


Figure 9: Overall Work flow of the Analysis

Henceforward, we tried to group the environmental factors are into 6 clusters indicating the different regions and having particular soil type and weather conditions. This data is combined with Performance data on the basis of Location by which we are able to get the cluster wise performance for corn. Next, we tried to group the genetics data. There are nearly 19,500 unique genetic markers in the genetics data of corn provided in Syngenta Challenge and the average number of markers assembled per hybrid is approximately 12,000. So, as a part of dimensional reduction we tried to divide the genetics features based on its specific position within a chromosome i.e. its marker position into 20 bins(groups), each contains near about 1000 marker positions. We took the probability of each nucleotides AA, AC and CC for every bin for finding the distribution of marker positions in each bin and combined the final data having 60 features which we got from 19500 features for 2268 unique pairs.

The following table represents the distribution of the percentage of AA, AC and CC nucleotides bases for all marker position bin across 4 different buckets of yield. The horizontal rows indicate, if the distribution of a nucleotide base increases from one yield bucket to the next yield bucket, that cell will be colored in green and if decreases then cell will be colored in orange. Change in the distribution will indicate changes is marker position across bins.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AA | | | | |  | AC | | | | |
| Marker bin | Yield <50 | Yield 50-100 | Yield 100-150 | Yield >150 |  | Marker bin | Yield <50 | Yield 50-100 | Yield 100-150 | Yield >150 |
| 1 | 43.3% | 43.7% | 43.5% | 42.9% |  | 1 | 50.1% | 49.9% | 50.5% | 50.3% |
| 2 | 41.0% | 40.9% | 40.9% | 40.1% |  | 2 | 52.0% | 52.1% | 52.4% | 53.1% |
| 3 | 41.9% | 42.0% | 42.2% | 40.8% |  | 3 | 51.5% | 50.8% | 51.3% | 52.4% |
| 4 | 40.4% | 41.6% | 41.4% | 39.1% |  | 4 | 50.6% | 49.7% | 50.1% | 51.6% |
| 5 | 43.5% | 44.1% | 44.0% | 42.8% |  | 5 | 48.8% | 48.1% | 48.4% | 49.5% |
| 6 | 56.1% | 55.9% | 56.1% | 56.1% |  | 6 | 39.3% | 39.3% | 39.3% | 39.3% |
| 7 | 45.1% | 45.6% | 45.4% | 44.5% |  | 7 | 46.0% | 45.7% | 46.0% | 46.6% |
| 8 | 43.6% | 44.2% | 44.1% | 42.5% |  | 8 | 48.6% | 47.7% | 48.2% | 49.6% |
| 9 | 44.3% | 44.3% | 44.3% | 43.6% |  | 9 | 48.3% | 48.2% | 48.6% | 48.9% |
| 10 | 42.5% | 43.2% | 43.0% | 42.0% |  | 10 | 50.2% | 49.4% | 50.0% | 50.6% |
| 11 | 42.8% | 42.6% | 42.4% | 41.7% |  | 11 | 49.3% | 49.0% | 49.6% | 50.1% |
| 12 | 38.2% | 38.5% | 38.4% | 38.3% |  | 12 | 54.6% | 53.2% | 53.5% | 54.1% |
| 13 | 39.8% | 40.8% | 40.5% | 37.7% |  | 13 | 54.9% | 53.0% | 54.1% | 57.2% |
| 14 | 59.4% | 59.2% | 59.5% | 59.2% |  | 14 | 36.7% | 36.6% | 36.6% | 36.9% |
| 15 | 57.7% | 57.6% | 58.0% | 57.9% |  | 15 | 37.8% | 37.9% | 37.7% | 37.7% |
| 16 | 57.2% | 57.6% | 57.7% | 56.9% |  | 16 | 38.7% | 38.1% | 38.3% | 39.0% |
| 17 | 54.9% | 55.1% | 55.4% | 54.9% |  | 17 | 40.2% | 39.9% | 39.9% | 40.2% |
| 18 | 43.6% | 43.4% | 43.7% | 42.6% |  | 18 | 50.2% | 50.0% | 50.4% | 51.0% |
| 19 | 41.9% | 42.3% | 42.4% | 41.5% |  | 19 | 49.6% | 48.9% | 49.0% | 49.9% |
| 20 | 57.4% | 56.9% | 57.6% | 57.6% |  | 20 | 38.0% | 38.2% | 38.0% | 38.0% |

Table2:

(a) Yield for AA across 20 Marker bins (b) Yield for AC across 20 Marker bins

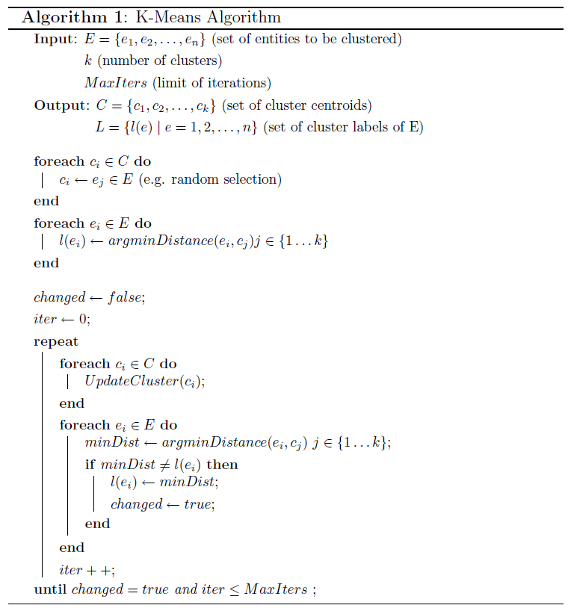
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CC | | | | |
| Marker bin | Yield <50 | Yield 50-100 | Yield 100-150 | Yield >150 |
| 1 | 6.6% | 6.4% | 6.1% | 6.8% |
| 2 | 7.0% | 7.1% | 6.7% | 6.8% |
| 3 | 6.7% | 7.2% | 6.5% | 6.8% |
| 4 | 9.1% | 8.7% | 8.5% | 9.3% |
| 5 | 7.7% | 7.8% | 7.6% | 7.7% |
| 6 | 4.7% | 4.8% | 4.5% | 4.6% |
| 7 | 46.0% | 45.7% | 46.0% | 46.6% |
| 8 | 7.9% | 8.1% | 7.7% | 7.9% |
| 9 | 7.4% | 7.5% | 7.1% | 7.5% |
| 10 | 7.3% | 7.4% | 7.0% | 7.4% |
| 11 | 7.9% | 8.4% | 8.0% | 8.2% |
| 12 | 7.2% | 8.3% | 8.0% | 7.6% |
| 13 | 5.3% | 6.2% | 5.5% | 5.1% |
| 14 | 3.9% | 4.2% | 4.0% | 3.9% |
| 15 | 4.5% | 4.5% | 4.3% | 4.5% |
| 16 | 4.1% | 4.3% | 4.0% | 4.1% |
| 17 | 4.8% | 5.0% | 4.7% | 4.8% |
| 18 | 6.2% | 6.6% | 6.0% | 6.3% |
| 19 | 8.5% | 8.8% | 8.6% | 8.7% |
| 20 | 4.6% | 4.8% | 4.4% | 4.4% |

(c) Yield for CC across 20 Marker bins

Furthermore, in the analysis, we used exogeneous data of socioeconomic factor (Name of the Factor) [] and combined this parameter with the cluster of environmental factors and cluster of genes and created completed structured dataset which we used as a training data for the predicting the performance of corn in untested land using Artificial Neural Network [14].

The benefits of k-means clustering are it divides data rationally into the subsets which can be used for further data processing and the output would be different for different clusters.   
Ideally, it helps in pin-pointing the performance of each clusters and describe different characteristics.

Below is shown the pseudo code for K-means clustering used for our analysis.

[](https://www.google.co.in/url?sa=i&rct=j&q=&esrc=s&source=imgres&cd=&cad=rja&uact=8&ved=0ahUKEwiz0eC3oN7YAhUEYKwKHRlaC10QjRwIBw&url=https%3A%2F%2Fen.wikibooks.org%2Fwiki%2FData_Mining_Algorithms_In_R%2FClustering%2FK-Means&psig=AOvVaw1iiZFtfjoqprXLrpuESE6F&ust=1516252921394012)

Pseudo code – K-means clustering

Below figure shows the Nucleotide gene progression across markers for different yields buckets

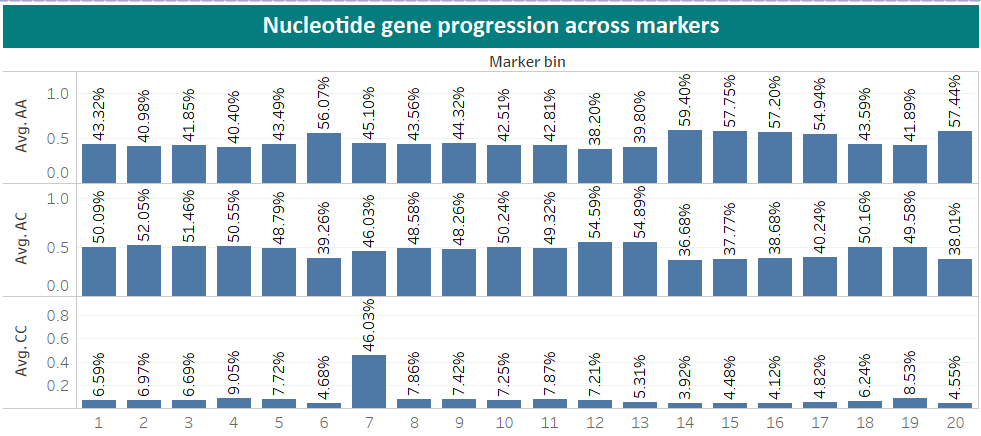


Figure 10: Nucleotide gene progression across markers for yield between 0 - 50

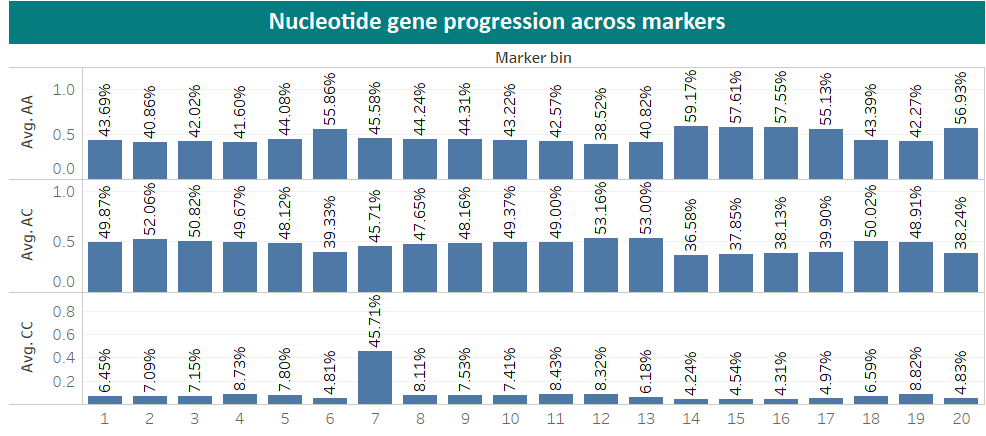
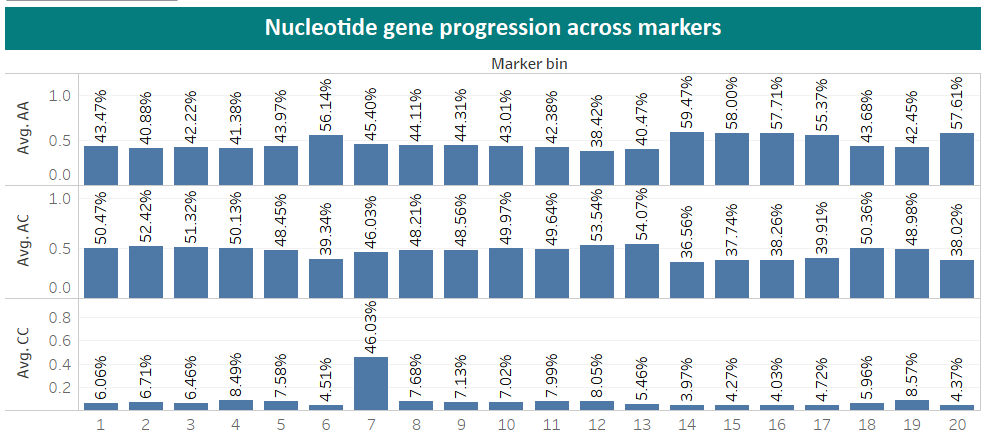
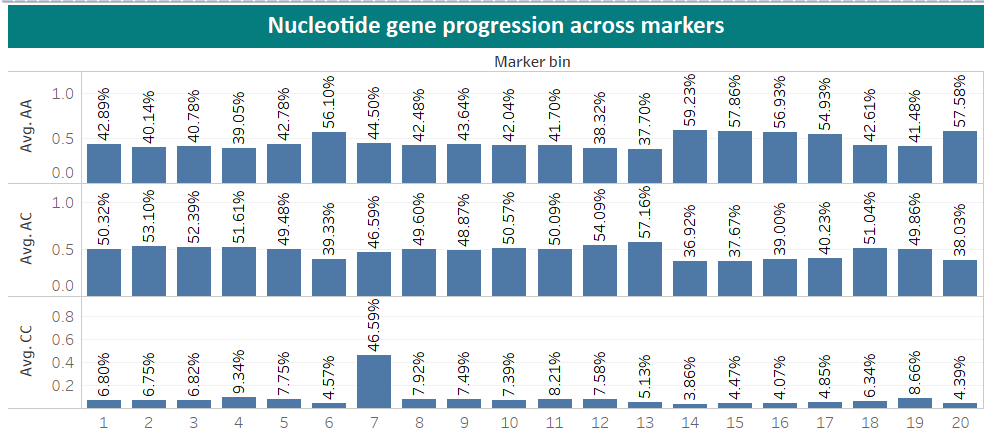


Figure 11: Nucleotide gene progression across markers for yield between 50 & 100  
  
Figure 12: Nucleotide gene progression across markers for yield between 100 & 150

Figure 13: Nucleotide gene progression across markers for yield greater than 150

1. Quantitative results

One of the most important pieces of this challenge was to find relevant subset of markers that interact with the environment in positive or negative ways. There are two different statistical contextual models for the scale of measurement that determines the existence of gene-environment interaction exists. One is additive, and another is multiplicative which uses rations to measure the effects based on interaction between genes and environment. Additive model has been suggested for predicting disease risk in a crop while multiplicative is used for disease etiology. So, in our case we used multiplicative model by taking the ratio as mentioned in the section 3.

The following table 4 indicates how the environmental cluster and gene cluster interacts with each other. Each row represents the performance of each clusters with different hybrid clusters. The best yield is shown with the maximum value along the column and represented with the dark blue colors and lowest with the red color.

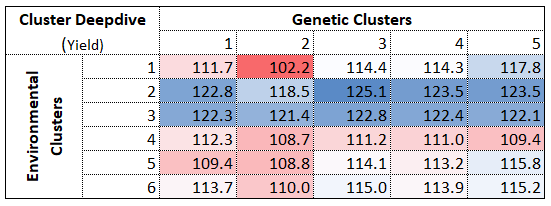


Table3: Gene-environment Interaction

Below graph shows how many hybrids are divided among different clusters for the genetic data set

Figure 14: Hybrids Marker Location division in 5 clusters

Below table shows which genetic cluster has shown best and worst result for each environment clusters.

|  |  |  |  |
| --- | --- | --- | --- |
| Clusters | | Gene Marker | |
| Best | Worst |
| Environment | *1* | 5 | 2 |
| *2* | 3 | 2 |
| *3* | 3 | 2 |
| *4* | 4 | 2 |
| *5* | 5 | 2 |
| *6* | 5 | 2 |

Table4: Identifying best genetic marker locations for each environment clusters

Output:

We ran artificial neural network model for each clusters and calculated the RMSE values. Also in order to have most optimum solution, we ran Artificial Neural Network Algorithm combining all the clusters and calculated the RMSE value.

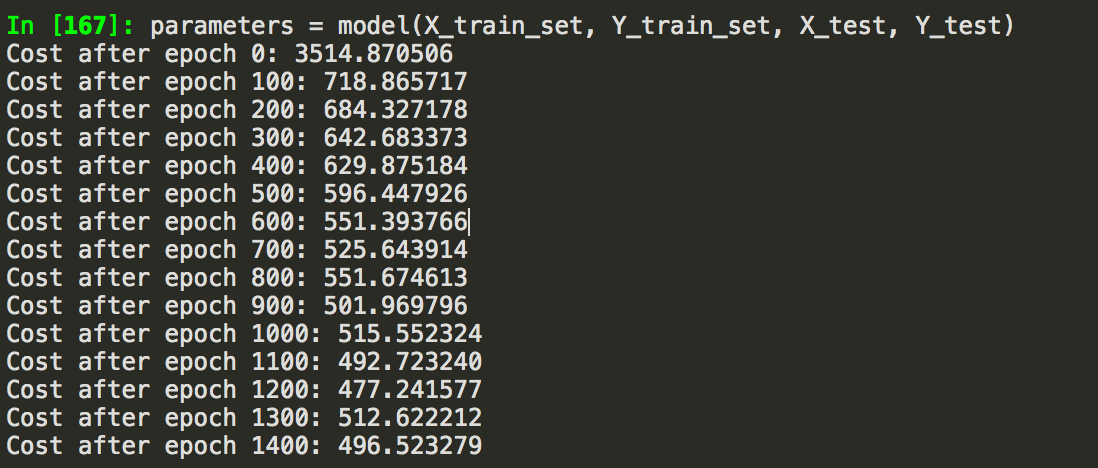
The below results are from the ANN with 10 layers with 8 Hidden Layers.

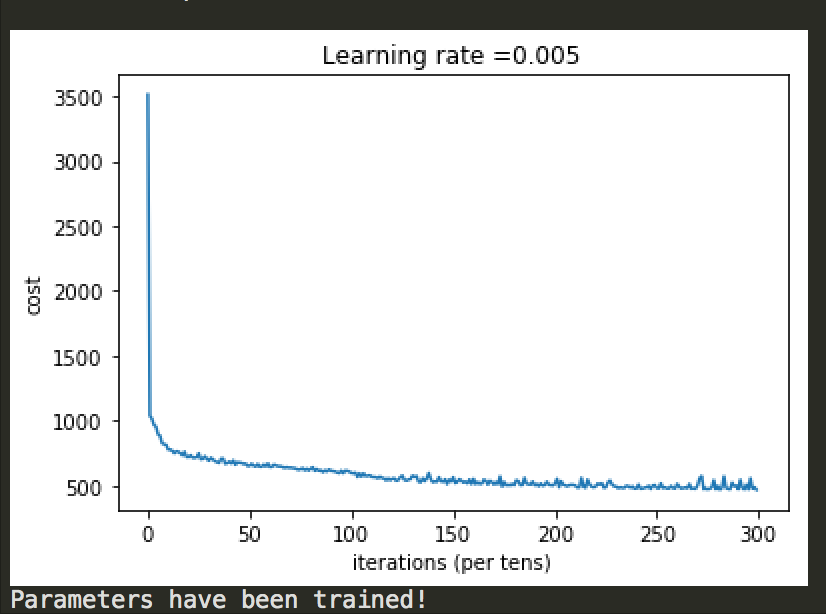
Method of Operations Defined while running ANN

INPUT LAYER - > 8 RELU LAYER - > LINEAR LAYER

The implementation of the network is on the 6 clusters developed and each result shows the cost function and RMSE value for each layer. The network learning rate is .005 and with 1500 EPOCHS and Optimization using ADAM. With the 0ANN method, we can find the global minimum values which are then used for prediction. With the Maximum function in RELU we can learn more from the data and further use it to calculate the linear function result.

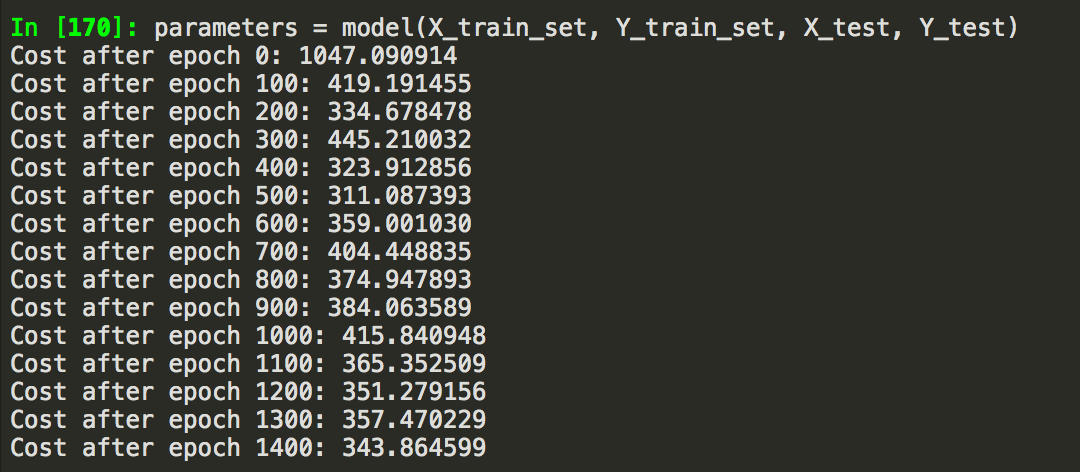
CLUSTER 1:

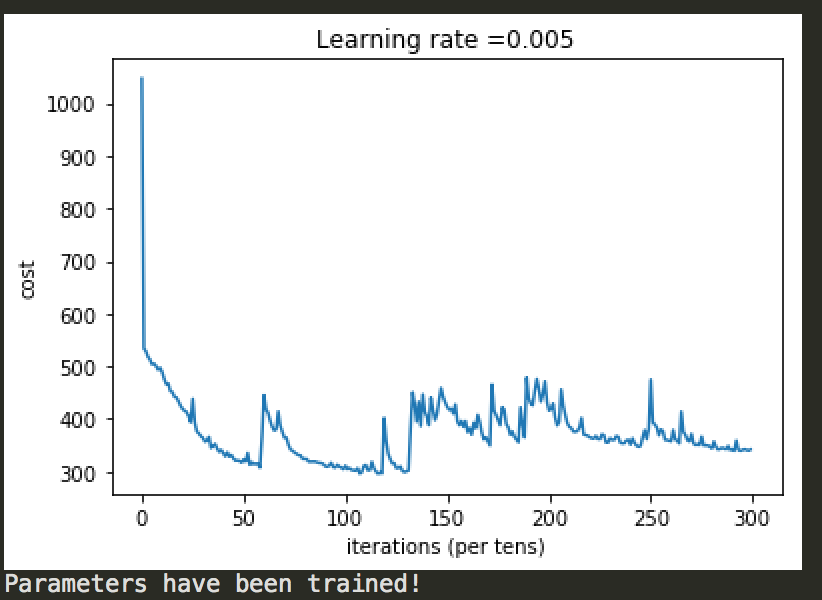




RMSE : 21.86  
  
Figure 15: Learning rate curve for cluster 1

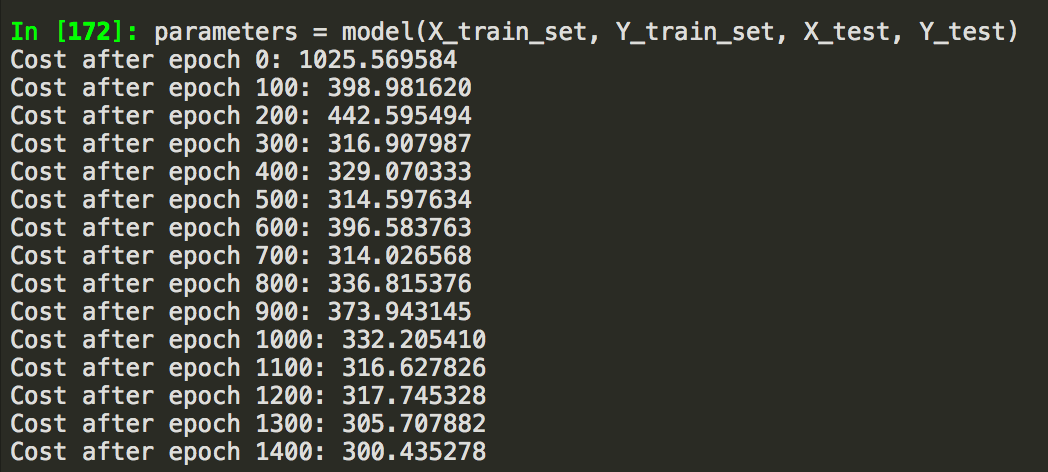
CLUSTER 2:

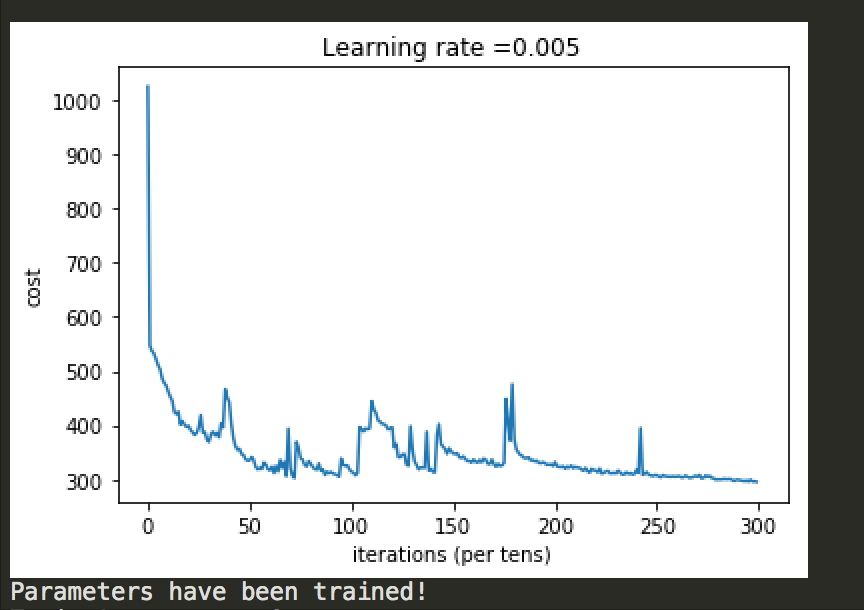




RMSE : 18.5203  
  
  
Figure 16: Learning rate curve for cluster 2

CLUSTER 3:

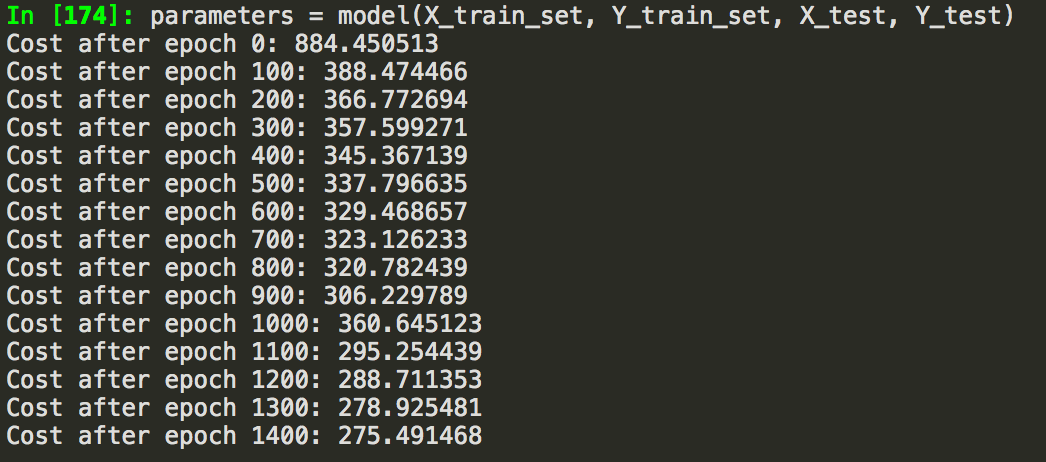


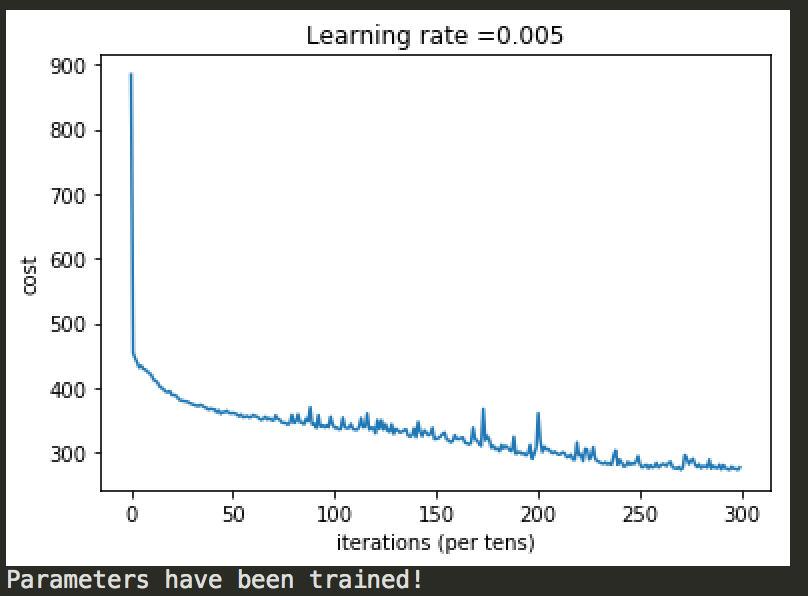


RMSE: 17.3205

Figure 17: Learning rate curve for cluster 3

CLUSTER 4:

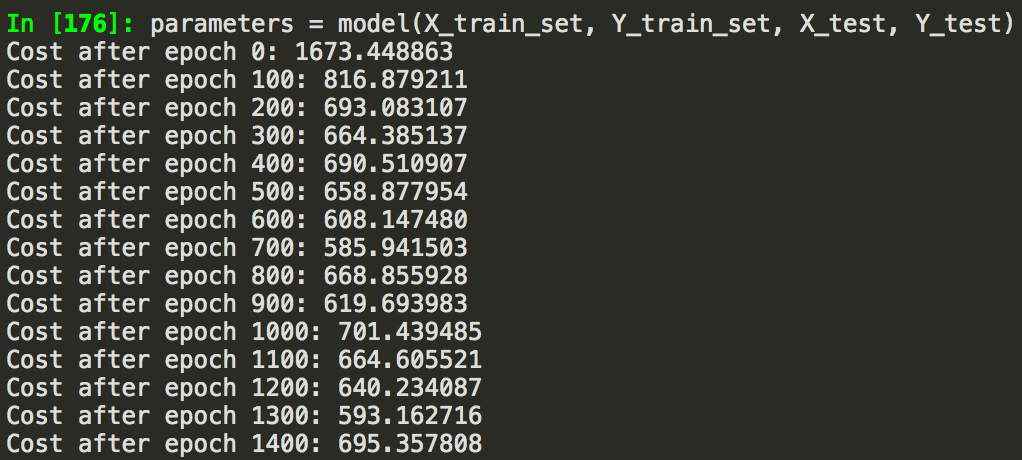


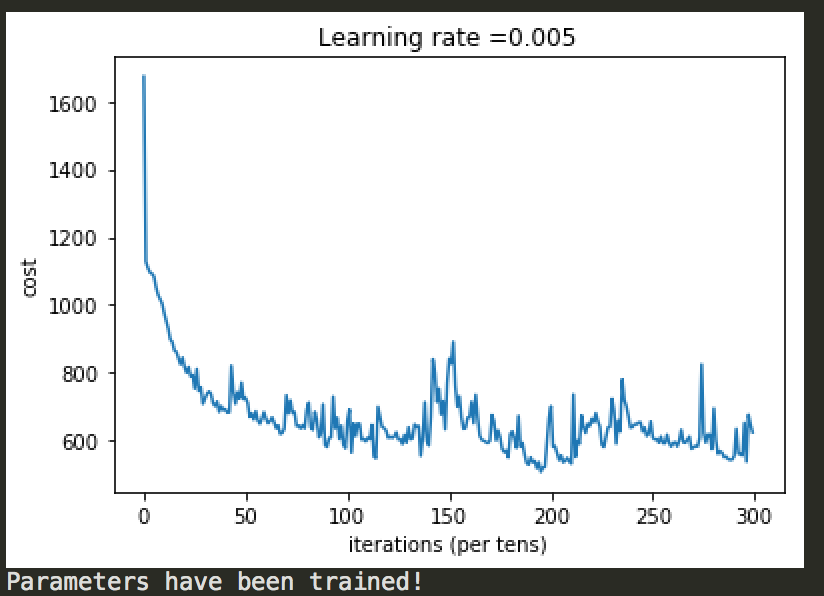


RMSE : 16.5831

Figure 18: Learning rate curve for cluster 4

CLUSTER 5:

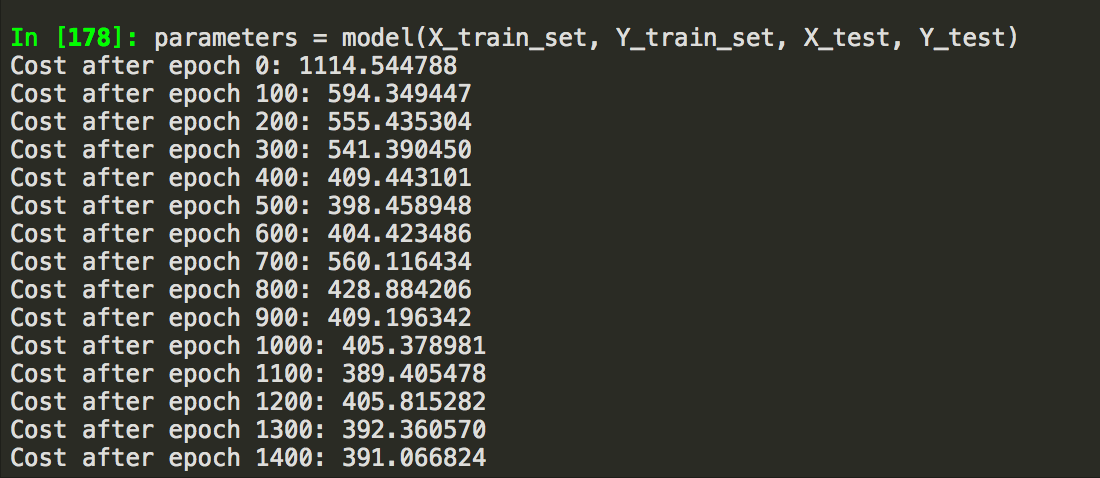


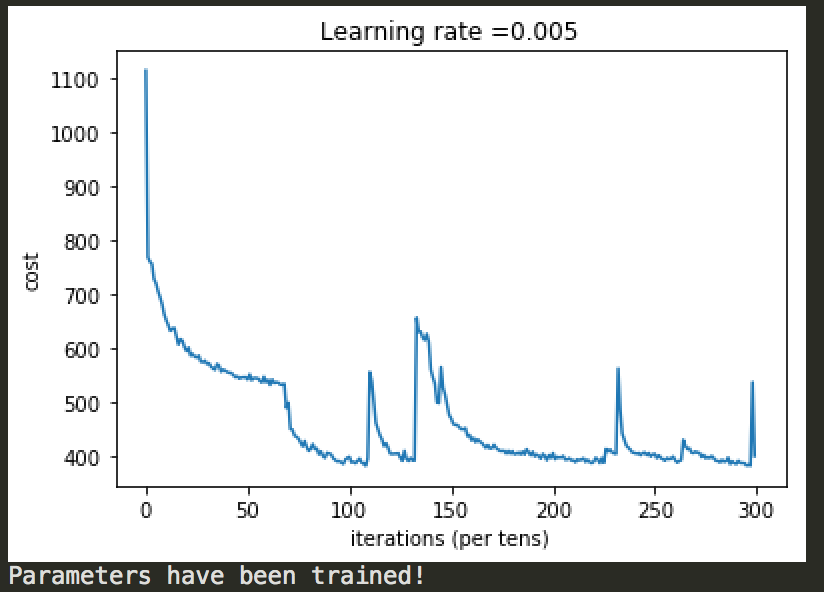


RMSE : 26.3628

Figure 19: Learning rate curve for cluster 5

CLUSTER 6:





RMSE: 19.7737

Figure 20: Learning rate curve for cluster 6

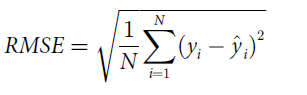
|  |  |  |
| --- | --- | --- |
| Data Clusters | Method | RMSE |
| Cluster1 | ANN | 21.86 |
| Cluster2 | ANN | 18.52 |
| Cluster3 | ANN | 17.32 |
| Cluster4 | ANN | 16.58 |
| Cluster5 | ANN | 26.36 |
| Cluster6 | ANN | 19.77 |

Table 5: Output matrix for the RMSE values from ANN algorithm performance.

Accuracy:

This the measurement of evaluation used for the model selection.

Root Mean Square Error



5. Team members

David Joy

Eeshpaul ­­­­­­­

Madhumita Jadhav

Raj Deep Arora

• Team Member One, University of Texas at Dallas, mitujadhav@gmail.com

• Team Member Two, University of Texas at Dallas, [rajdeep.arora@utdallas.com](mailto:rajdeep.arora@utdallas.com)

• Team Member three, University of Texas at Dallas, [eeshpaul.saggu@utdallas.com](mailto:eeshpaul.saggu@utdallas.com)

• Team Member four, University of Texas at Dallas, [david.joy@utdallas.com](mailto:david.joy@utdallas.com)

6. Exogenous data sets

PCI (Per Capita Income) is a socioeconomic factor that affects the performance of crop. This parameter indicates ability of people to use best practices in agriculture. So, we gathered this data for each geographical location and based on these values we tuned our model. We collected the publicly available data from Wikipedia [10] and from the data provided by the California State University, Dominguez Hills, University of Wisconsin, Parkside and Soka University Japan- Transcend Art and Place

• Exogenous Data Set: <http://www5.csudh.edu/dearhabermas/populcntrl.htm>

7. Supplementary materials

Syngenta\_Crop\_Challenge.R file contains complete analysis and end to end process of the prediction.

Syngenta\_Crop\_Challenge.twdx has complete exploratory data analysis of all types of data provided by Syngenta as a part of the challenge

GitHub Repo Link:

https://github.com/mitujadhav/Syngenta/

Acknowledgments

Necessary acknowledgements to funding agencies or others should be made in this section. This section is optional.

References

[1] http://www.fao.org/state-of-food-security-nutrition/en/

[2] <https://www.compassion.com/multimedia/world-population-prospects.pdf>

[3] <https://www.nature.com/articles/ncomms6989>

[4] https://www.pioneer.com/home/site/us/agronomy/soil-temp-corn-emergence/

[10] <https://en.wikipedia.org/wiki/Population_control>

[11] IdeaConnection. Syngenta Crop Challenge 2018; Available from: <https://www.ideaconnection.com/syngenta-crop-challenge/challenge.php>

[12] Wikipedia. Gene-environment Interaction; Available from: <https://en.wikipedia.org/wiki/Gene%E2%80%93environment_interaction#Gene_%C3%97_Environment_%C3%97_Environment_Interactions>

[13] Wikipedia. Corn Belt; Available From: <https://en.wikipedia.org/wiki/Corn_Belt>

[14] Exogeneous source link. : <http://www5.csudh.edu/dearhabermas/populcntrl.htm>