Prediction on Soybean Variety Yield and apply to Soybean Variety Selection

ID:489716 Gorden Li

Abstract: Food supply is a very important aspect of sustainability of the human race. Nowadays, still there are a lot of countries and regions suffering from food shortage. It becomes more and more important for farmers to grow efficiently using the resources they have to ensure that the shortage is relieved to some extent. In the study, we collected data set of Soybean Variety from 2003-2009, covering 34212 samples and 27 dimensions including temperature, radiation, precipitation, soil type, weather etc. First we will cluster locations into different group based on latitude and longitude since sample within groups will have similar weather condition. Then by applying Machine Leaning based algorithm like Linear Regression, LASSO, Tree Models, Neural Network etc. to different varieties, we are able to predict yield of Soybean according to the conditions given. Based on the prediction, suggestions on sowing are provided to farmers to get an optimal portfolio and a higher yield to relieve shortage of food supply

Keywords: Food Shortage, Predict yield of Soybean, Machine Learning, Linear Regression, LASSO, Tree Models, Optimal Portfolio

1. Introduction

1.1 Motivation of the study

Woven into the fabric of history, hunger and food shortage experienced by individuals, communities, or large-scale societies are important historical markers, often used to chronicle significant points in time. Hunger and food shortage was and is also the result of intentionally orchestrated conditions including any combination of warfare, politics, poverty, and power[1]. From the Population Institute's point of view, around 230 thousand more babies are born every day. Also, the World Food Programme has evidence showing that about 795 million people are lacking food supply to ensure no physical disease raised by food shortage happen. On the other hand, land used for farming has been decreasing which increase the burden of food shortage. Simply attempting to increase the land available for farming is not enough to sustain the needed food supply[2]. Thus, teaching the farmers how to increase the efficiency of limited land resources is a key solution to food shortage. It is the core problem this study is going to solve, optimizing varieties portfolio and increasing yield in a certain size of land. By this, food shortage will be relieved.

1.2 Goal of the study

The goal of the study is that through quantitative analysis, we can give detail suggestions to Soybean growers under different conditions on how to grow Soybeans more efficiently. The study will serve as an indicator and help them make plan before hand year by year according to different sowing conditions like temperature, precipitation, radiation, soil type, weather etc. As a result of machine learning algorithm in the research, we will give out quantitative indicators on what to sow, how much to sow, where to sow and how much they will gain. The result will not only serve for farmers, but also might affect the entire Soybean sowing industry from seed retailers to farmers. It is because once set a more efficient sowing plan, farmers might change their portfolio at the beginning-seeds purchasing. Retailers can benefit from the study by knowing when to sell which seeds and how much seeds.

1.3 Literature Review

So far, there are many researches analyzing what factors will affect Soybean yield and productivity. Ana POSPIŠIL(2009) found out that Cropping system intensity is positively related with soybean seed yield. Plant density had no significant influence on yield of Soybean. However when Soybean meets with stress like high temperature and inadequate precipitation, weight of Soybean during its growth can be impacted. It will finally lead to a low yield [3]. Heidi Liere's (2015) study indicates that one very essential is landscape component. It helps to understand spatial variability in biocontrol and yield, within which , however, effect of environmental variability and compensatory growth might reverse beneficial effects on crop and lower the yield of Soybean[4]. M. Scott Wells (2014) did researches to evaluate the effects of soybean planting timing and row spacing on soil moisture, weed density, soybean lodging, and yield in a cover crop-based no-till organic soybean production system. The conclusion is that soil volumetric water content (VWC) was higher in the cereal rye mulch treatments compared to the no rye checks. Moreover, lowered soil water evaporation might be aroused by delay in Soybean sowing. As the increased soil VWC in the rolled-rye treatment, translation from soil VWC into increased soybean yield is not significant, indicating there might be some obstacle for Soybean to gain yield from increasing soil VWC. The rolled-rye treatment pointed out that there is a significant increases in soil VWC when compared to the no-rye treatment[5]. According to the review, current study focus more on how Soybean yield is lost, biological environment and effect of seed size itself on the yield Soybean. We are now offering an estimation based on a more micro environment factors like temperature, precipitation, radiation, weather, soil type, genetic group etc.

2. Methodology and Analysis

2.1 Description Analysis

The data set of Soybean varieties yield has 34212 objectives. The dimensions of Soybean yield data cover 2003-2009 7 years data, 118 locations, 18 genetics

group, and 182 varieties. In Figure 1, almost all of the Soybean samples locate in Midwest of the US.

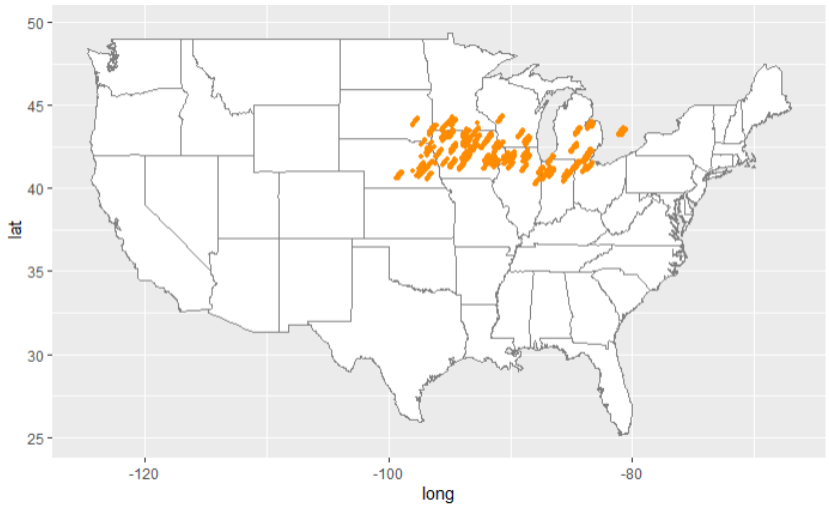


Figure 1 Distribution of Locations

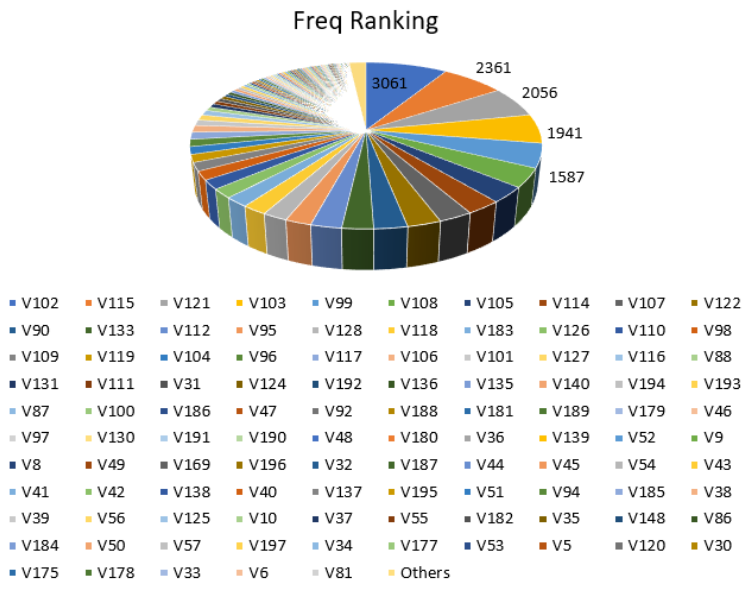


Figure 2 Frequency of Varieties

In Figure 2, we can see 10 varieties almost occupy 50% of total samples. A threshold of 30 is set in this research. Those varieties have samples bigger than 30 will have its own model. All those varieties under 30 samples will be combined into others since sample size smaller than 30 will not be able to support the algorithm containing almost 30 predictors. Accuracy will be low if we set models for each of those variety. Finally, 106 models will be set.

Location	Variety	Freq	Location	Wea1	Freq	Location	Wea2	Freq
4210	V102	451	4210	322	2467	4210	322	2467
4210	V99	350	4310	322	2157	4310	322	2157
4310	V102	325	4490	321	1613	4490	321	1613
4310	V99	293	3210	221	1268	3210	221	1268
4310	V90	268	3270	222	985	3270	222	985
4490	V102	266	3260	322	974	3260	322	974
4210	V103	263	3250	322	939	3250	322	939
3210	V121	242	3230	322	935	3230	322	935
4210	V90	215	3440	321	903	3440	321	903
4210	V95	202	2240	222	869	2240	222	869

Table 1 Relation between Location & Variety/Weather ranking by Frequency

In Table 1, we can see that the most common variety/weather in a certain location. The most popular location for Soybean sowing is 4210, with varieties V102 and V99 leading. Its weather type is 322. It is followed by 4310, with varieties V102, V99, V90 leading and weather type is 322 as well.

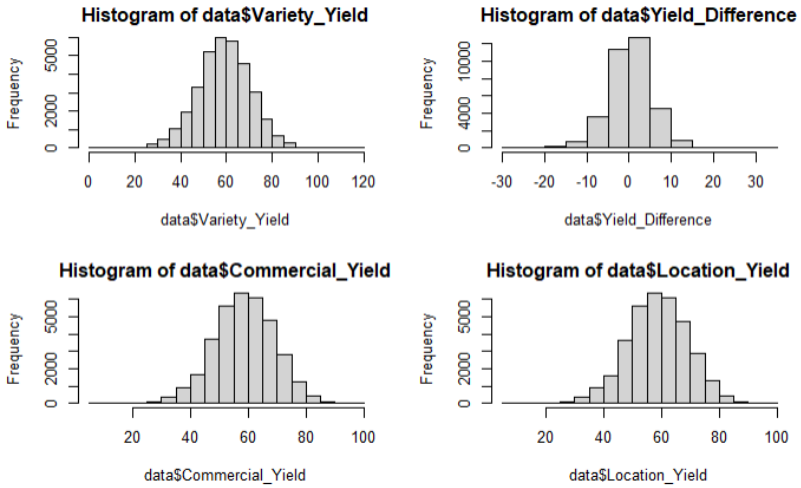


Figure 3 Distribution of Yield Variables

In Figure 3, we can see that all of the Yield-related variables agree with normal distribution. Thus, it is unnecessary to transfer Yield variables and Variety_Yield is chosen here as dependent variable. To reduce dimension and cluster similar weather for growing Soybean, k-means clustering is applied here to increase the accuracy of model by clustering potential pattern of sowing environment hidden beneath the dimension. In this case, scaled latitude and longitude is taken into consideration when calculation the distance between clusters. In Figure 4, we can see the effect of different numbers of clusters. Finally, 30 clusters is chosen to label all samples, its within group sum of square is 47.3. Based on the cluster, we set a new variable, Location_Class to the data set.

Samples within same Location_Class will have similar weather conditions.

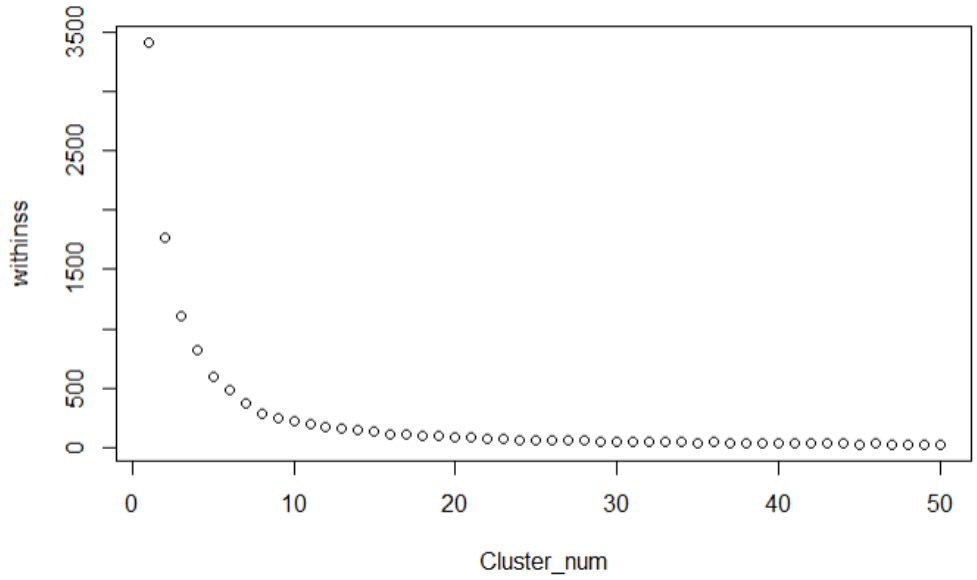


Figure 4 K-means clustering

2.2 Predictive Analysis

2.2.1 Predictors Description

The processed data has 27 predictors, and there will be 105 models for varieties with more than 30 samples each and 1 model for rest of varieties less than 30 samples. Totally, there are 106 models for all the varieties. Predictors in the models are shown in Table 2.

Variables	Description
Variety_Yield (Y)	Bushells per acre adjusted by moisture
Latitude (X_1)	Latitude
Longitude (X_2)	Longitude
Rela_Maturity25 (X_3)	Probability of growing soybeans in the nearby area of the site
Wea1, Wea2 (X_3, X_4)	Climate class, Season Class
Prob (X_5)	Probability of growing soybeans in the nearby area of the site
Prob_IRR (X_6)	Probability of field irrgation nearby the area of the site
Genetics (X_7)	breeding group
Temp (X_8)	Sow Year Daily degree Celsius sum from April 1st to October 31st
Median_Temp (X_9)	Daily degree Celsius sum between 1994 and 2007
Prec (X_{10})	Sow Year Daily Precipitation sum from April 1st to October 31st

Median_Prec (X_{11})	Precipitation sum between 1994 and 2007
Rad (X_{12})	Sow Year Daily Watts per sq. meter solar radiation sum from April 1st to October 31st
Median_Rad (X_{13})	Daily Watts per sq. meter solar radiation sum between 1994 and 2007
Location_Class (X_{14})	Cluster No. based on K-means
Others (X_{15} ~X_{27})	Year, Soil_Type, Density, AWC1, PH1, Clay1, Silt1, Sand1, PH2, Clay2, Silt2, Sand2, CEC

Table 2 Predictors Description

2.2.2 Models for Predictions

Type	Linear Models		Non-Linear Models				
Model	Linear	LASSO	Tree	Bagging	Forest	Boost	Neural
RMSE	9.05	9.09	8.84	7.71	7.61	7.53	7.80

Table 3 Performance of algorithms

In the study, Linear Regression, LASSO, Regression Tree, Bagged Tree, Random Forest, Boosted Tree and Neural Networks. In Table 3, compared to Linear Models, Non-linear Models have better performance with RMSE much lower than that of Linear Models. Neural could have good performance. However, neural network is a data-hunger algorithm. Most of the varieties do not have such big sample size. Thus, the performance of Neural Network is not very well. Among all the algorithms, Boosted Tree has the best performance. After hyperparameter adjustment, boost trees with 40 trees, 10 in depth, 0.2 in shrinkage rate achieve the best performance among test set.

2.2.3 Predictions of evaluation sample.

So far the best model has already been chosen. In order to predict the yield of evaluation sample(sample of 2011), what we need now is to estimate the environment condition of the evaluation year 2011. As is shown in Table 3, among all the predictors, Temp(X_8), Prec(X_9) and Rad(X_{10}) are the most changable predictors related to year of sowing. To reasonably estimate the possible temperature, precipitation and radiation of the evaluation year 2011, distribution of these environment conditions from 2003~2009 in the same

cluster(location_class) as evaluation sample are taken as reference. There is an imbalance in the history data since only few sample in 2003~2009 belongs to evaluation sample cluster. On the other hand, according to history data, temperature, precipitation and radiation is random. Resampling with repetition within each year is needed to build a new environment sample. After resampling, we have 700(100 for each year from 2003~2009) samples of evaluation data with different temperature, precipitation and radiation. Distribution of these resampled environment condition is shown in Figure 5.

Temp	Prec	Rad
Min. :3197	Min. : 345.6	Min. :1026152
1st Qu.:3245	1st Qu.: 540.8	1st Qu.:1065438
Median :3321	Median : 696.6	Median :1086145
Mean :3385	Mean : 675.5	Mean :1089777
3rd Qu.:3548	3rd Qu.: 822.9	3rd Qu.:1092172
Max. :3736	Max. :1053.4	Max. :1186828

Figure 5 Distribution of environment condition estimation for evaluation sample

Predictions is based on the simulated data, distribution of yield of those varieties is shown in Figure 6.

V128	V130	V131	V133	V135	V136
Min. :68.44	Min. :53.93	Min. :40.22	Min. :49.40	Min. :41.59	Min. :46.62
1st Qu.:69.96	1st Qu.:54.80	1st Qu.:45.78	1st Qu.:49.58	1st Qu.:41.59	1st Qu.:47.19
Median :70.45	Median :55.21	Median :46.80	Median :51.19	Median :44.93	Median :48.69
Mean :70.70	Mean :55.77	Mean :47.06	Mean :52.10	Mean :44.00	Mean :48.18
3rd Qu.:71.51	3rd Qu.:56.92	3rd Qu.:49.15	3rd Qu.:53.23	3rd Qu.:45.78	3rd Qu.:48.69
Max. :72.90	Max. :56.92	Max. :52.65	Max. :57.90	Max. :46.75	Max. :48.69
V139	V140	V169	V179	V180	V181
Min. :43.45	Min. :36.16	Min. :44.39	Min. :53.54	Min. :69.37	Min. :65.87
1st Qu.:43.45	1st Qu.:45.59	1st Qu.:52.99	1st Qu.:53.54	1st Qu.:70.17	1st Qu.:68.35
Median :43.45	Median :52.20	Median :52.99	Median :55.82	Median :70.62	Median :69.17
Mean :43.79	Mean :48.83	Mean :51.76	Mean :55.49	Mean :71.50	Mean :71.32
3rd Qu.:44.25	3rd Qu.:53.34	3rd Qu.:52.99	3rd Qu.:56.64	3rd Qu.:73.24	3rd Qu.:75.39
Max. :44.25	Max. :53.89	Max. :52.99	Max. :57.93	Max. :73.70	Max. :76.21
V183	V186	V187	V188	V189	V190
Min. :63.81	Min. :52.93	Min. :59.66	Min. :51.49	Min. :58.05	Min. :59.57
1st Qu.:65.08	1st Qu.:57.06	1st Qu.:59.66	1st Qu.:54.38	1st Qu.:58.05	1st Qu.:60.22
Median :66.55	Median :58.79	Median :59.66	Median :54.66	Median :58.85	Median :60.59
Mean :66.64	Mean :58.39	Mean :60.68	Mean :54.72	Mean :58.98	Mean :60.67
3rd Qu.:68.16	3rd Qu.:61.88	3rd Qu.:62.04	3rd Qu.:56.79	3rd Qu.:59.39	3rd Qu.:60.81
Max. :70.36	Max. :61.88	Max. :62.04	Max. :56.83	Max. :61.35	Max. :61.83

Figure 6 Distribution of yield of varieties

2.3 Prescriptive Analysis

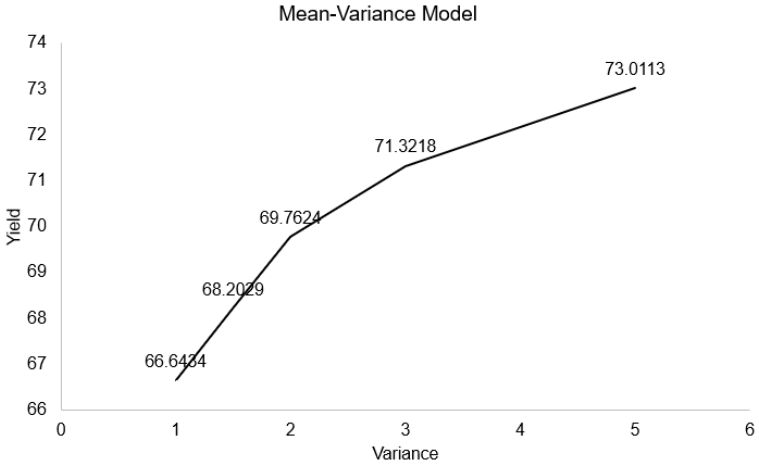
2.3.1 Optimal Portfolio

When building an Optimal Portfolio of varieties to build, we take consider of the top 5 yielding variety. They are shown in Figure 7.

V180	V181	V128	V183	V44
Min. :69.37	Min. :65.87	Min. :68.44	Min. :63.81	Min. :71.15
1st Qu.:70.17	1st Qu.:68.35	1st Qu.:69.96	1st Qu.:65.08	1st Qu.:71.15
Median :70.62	Median :69.17	Median :70.45	Median :66.55	Median :71.15
Mean :71.50	Mean :71.32	Mean :70.70	Mean :66.64	Mean :73.01
3rd Qu.:73.24	3rd Qu.:75.39	3rd Qu.:71.51	3rd Qu.:68.16	3rd Qu.:75.49
Max. :73.70	Max. :76.21	Max. :72.90	Max. :70.36	Max. :75.49

Figure 7 Top yielding varieties

According to the mean, variance and covariance of those top 5 varieties, portfolio is made. According to Markowitz Mean-Variance Model[6], frontier of the portfolio is built to help farmers to identify under a certain variance (risk), the best yield (return) they can have since different farmers and land conditions may have different situation, leading to a different risk torrelance.



	V180	V181	V128	V183	V44
66..6434	0.0	0.0	0.0	1.0	0.0
68.2029	0.0	0.0	0.006	0.994	0.0
69.7624	0.1097	0.0	0.4406	0.4496	0.0
71.3218	0.1444	0.0	0.5820	0.0065	0.2671
73.0113	0.0	0.0	0.0	0.0	1.0

Figure 8 Optimal Portfolio

As we can see in Figure 8, best portfolio under different standard variance (risk) is shown. Actually it's the frontier of Markowitz Mean-Variance Model. The highest return of sowing is to sow V44 only in the evaluation location, with return yield at 73.0113 Bushells per acre. As variance(risk) decreases, return decreases as well.

3. Conclusion

In the study, Boosted Trees is used to predict the yield of different varieties in evaluation location. After training 34212 samples, finally we have a Boosted Trees with Test RMSE equals to 7.53. In Figure 8, recommendation of weight of varieties is shown. If a farmer and his land is variance resisting, he might want to take the plan to grow more V183, which is comparatively low return and

low risk. If the farmer and the land can stand more risk to get a higher return, V44 will be the varieties they want to sow. Base on this model and data, local farmers will know what to sow next year and prepare for it before-hand. Moreover, the model enable them to adjust the yield and risk in different period by applying different weight. It helps farmers become more flexible and productive based on the resources they currently have and finally relieve the food shortage of the area.

Reference:

[1]. Hunger and Food Shortages, Kelly Kean Sharp

[2]. 500S Guidline

[3]. INFLUENCE OF CROPPING SYSTEM INTENSITY ON YIELD AND YIELD COMPONENTS OF NEW SOYBEAN GENOTYPES[J], Ana POSPIŠIL, Milan POSPIŠIL, Svjetlana MATOTAN, Dario JAREŠ, Bogdan KORIĆ, Cereal Research Communications, Vol. 37, Supplement: Proceedings of the VIII. Alps-Adria Scientific Workshop, 27 April–2 May 2009, Neum, Bosnia-Herzegovina (March 2009), pp. 41-44

[4]. Trophic cascades in agricultural landscapes: indirect effects of landscape composition on crop yield[J]. Heidi Liere, Tania N. Kim, Benjamin P. Werling, Timothy D. Meehan, Douglas A. Landis, Claudio Gratton Ecological Applications, Vol. 25, No. 3 (April 2015), pp. 652-661

[5]. Cultural Strategies for Managing Weeds and Soil Moisture in Cover Crop Based No-Till Soybean Production[J].M. Scott Wells, S. Chris Reberg-Horton, Steven B. Mirsky, Weed Science, Vol. 62, No. 3 (July-September 2014), pp. 501-511

[6]. On the Markowitz mean–variance analysis of self-financing portfolios[R]; Bai Zhidong, Liu, Huixia, Wong, Wing-Keung; KLAS, Math & Stat, Northeast Normal University and DSAP & RMI, National University of Singapore, Singapore | Department of Statistics, National University of Singapore, Singapore | Department of Economics, Hong Kong Baptist University, Hong Kong