VARIETY TRIAL VALIDATION: A FRAMEWORK TO INCORPORATE ON-FARM DATA

Raul Sebastian
Robert G. Hardin
Texas A&M University
College Station, TX
Edward M. Barnes
Cotton Incorporated
Cary, NC
Jason K. Ward
NC State University
Raleigh, NC
Wesley M Porter
University of Georgia
Tifton, GA
Michael T. Plumblee
Clemson University
Blackville, SC
John D. Wanjura
USDA-ARS Cotton Production and Processing Research Unit
Lubbock, TX

Abstract

Variety trial validation is a framework to incorporate on-farm data. This study presents a predictive model for official cultivar trial yields, an innovative solution to the absence of cultivar selection tools based on on-farm data. Historical data indicates that cultivar selection is widely accepted to be the most important decision a farmer makes during the year. This decision establishes a maximum possible yield and quality, based on the genetic potential of the cultivar. Studies based upon the National Variety Trial Data (NCVT) from 2013-2018 indicate that the environment has a high contribution to the variability in yield. Therefore, for this study, several environmental factors were considered given their influence on yields, such as soil texture, pH, and weather data. Public databases NRCS Soil Surveys and NOAA National Climatic Data Center (RNOAA) served as the main sources of environmental data. To identify the association between environmental data and trail data a k-means cluster analysis (unsupervised clustering technique) was used to group trial data per state. The generated results were then used to train and test a predictive model, this model was then evaluated using a confusion matrix. After evaluating the prediction results of the model (for the state of Texas given 2014 information only), the model had an accuracy of 95.4% when predicting the yield at a given trial site in Texas. The 2014 cultivar trial-based model was then used to predict 2015 (Texas) outcomes. The model had an accuracy of 34.4%, this percentage represents the ratio of successes generated by the model. The system compared the model's results against the unsupervised clustering results to generate a confusion matrix. This model can be further improved by incorporating weather data into the model. Currently, cultivar trial data does not include planting dates nor harvesting dates, therefore a method needs to be developed to predict planting dates and harvesting dates indirectly (Ex. Degree days). Then, a new predictive model can be developed to take into account weather data and obtain results based on sufficient environment data.

Introduction

Cultivar selection is widely accepted to be the most important decision a farmer makes during the year. This decision establishes a maximum possible yield and quality, based on the genetic potential of the cultivar. This maximum yield and quality are not achieved due to environmental conditions, which will be less than optimum at some point during the growing season. Cultivar tests are routinely conducted and published, often by university extension personnel. For a farmer to interpret and apply to their farm, several significant obstacles exist:

- Likely some significant differences in environmental factors from even the most representative test site, for example, small-scale variations in climate patterns and soil types
- Commercial cultivars typically have a limited lifespan in the marketplace, less than five years, limiting the amount of cultivar test data available in different environments
- Cultivar trials often evaluate widely adapted cultivars, while cultivars better adapted for a specific environmental niche may have far less data

A key aspect of analyzing and interpreting cultivar trial data is partitioning variation in yield and quality factors into genotype (G), environment (E), and genotype by environment interaction (GxE) components. Cultivar selection involves consideration of G and GxE components of variation to identify the cultivar that will perform best in a specific environment. While the effect of G on yield and quality can be determined from large-scale cultivar testing, the GxE effect for a specific combination of cultivar and a farmer's unique environment is not easily determined. Furthermore, with cotton, the GxE interaction typically accounts for a greater amount of yield variation than G. Meredith et al. (2012) found that GxE accounted for 8.4% of the total variance in lint yield in the National Cotton Variety Test (NCVT) Regional High-Quality tests from 2001-2007, while G accounted for 7.4% of the variance. Previous studies were summarized, with similar findings: an average of 5% of the variance was due to G, while 9% was due to GxE. Therefore, understanding and predicting this interaction is critical to selecting profitable cultivars.

Statisticians and geneticists have developed analyses to optimize multi-environment cultivar trials to identify the most desirable cultivars (correctly identifying genetic effects) while minimizing the number of environments tested to reduce costs. This research addresses the goals of the plant breeder and maximizes genetic progress of the entire industry but does not provide the farmer predictive information of GxE interactions.

Materials and Methods

For this study R Studio was used as the main integrated development environment to run the algorithms developed for this project. The RStudio version 1.1.456 was installed in a Windows 10 environment with the following specifications.

System specifications:

System: Intel(R) Core(TM) i7-7700HQ CPU @2.80GHz, 16.0 GB of RAM
R version 4.0.3 (2020-10-10)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows >= 8 x64 (build 9200)

R-packages (name_version):

factoextra_1.0.7	ggplot2_3.3.2	rnoaa_1.2.0
cluster_2.1.0	tidyverse_1.3.0	caTools_1.18.0
forcats_0.5.0	plyr_1.8.6	rattle_5.4.0
stringr_1.4.0	rgeos_0.5-5	bitops_1.0-6
dplyr_1.0.2	rgdal_1.5-18	tibble_3.0.4
purrr_0.3.4	sp_1.4-4	rpart.plot_3.0.9
readr_1.4.0	aqp_1.25	rpart_4.1-15
tidyr_1.1.2	soilDB_2.5.8	

The predictive model developed for this study was generated using data extracted from the Seedmatrix, web-based application with a database of trial data, from 2005-2015. Data from multiple states was extracted, it was first examined to remove unnecessary data and the information related to location, yield and product were maintained. From a study performed in parallel to this research it was concluded that analyzing the data per state and per year was much efficient and could better results than analyzing the data as a whole. Due to time constraints one state was selected for proof of concept. Texas was selected as the candidate state given the amount of data available, 2171 variety trial locations (2005-2015). To complement the trial data, information was retrieved from the publicly available SoilDB database using the latitude and longitude recorded in the variety trial data. By traversing through the SQL tables using the mukey and cokey obtained based on location of the production sites the soil chemical and physical properties were obtained. A new file was created to hold the newly generated data following the format shown in Table 1.

Table 1: Data format of variety trial data after appending soil chemical and physical properties.

TrialID	Latitude	Longitude	Year	Mean Sand %	Mean Silt %	Mean Clay %	Mean pH	Mean EC

An unsupervised clustering analysis was performed to identify cotton production regions based on yield, soil chemical and physical properties (yield, mean sand %, mean silt %, mean clay %, mean pH and mean EC). The built-in function *fviz_nbclust()* available in "factoextra_1.0.7" package provided an automatic solution to determine the optimal number of clusters per dataset using the average silhouette method. With a defined optimal amount of clusters, the partitioning function *kmeans()* available in "cluster_2.1.0" was used to assign a group to each record in the variety trial dataset. With this method the unlabeled variety trial records were labeled, based on the level of similarity with other variety trial records of the same year and state.

A predictive model was then trained and tested with the labeled datasets generated from the previous steps. The information was first shuffled (built in function *shuffled[]*) then separated into two sets, the training set and the testing set. A decision tree is then built (with *rpart()* available in "rpart_4.1-15" package) and trained with the training set saved, the objective of the decision tree is to identify properly the group number in which the variety trial records belong to. After building the decision tree, this one can be tested using the previously created testing set, the readily available *predict()*. The results of the testing portion of the experiment can then be evaluated using a confusion matrix to measure the overall accuracy of the predictive model. With the numerous techniques applied, the method used for this study can be best described as a case of semi-supervised learning.

Results and Discussion

In the early stages of this study environmental factors were analyzed to define the influence and the impact they had on cotton's yield. One of these factors was irrigation availability, according to the Seedmatrix trial data from 2005-2015, in the state of Texas irrigation availability played a significant role throughout the years in the state's production sites yield. The importance of irrigation availability shows consistency despite the location (latitude and longitude) of the cotton production sites as shown in Figure 1 and Figure 3.

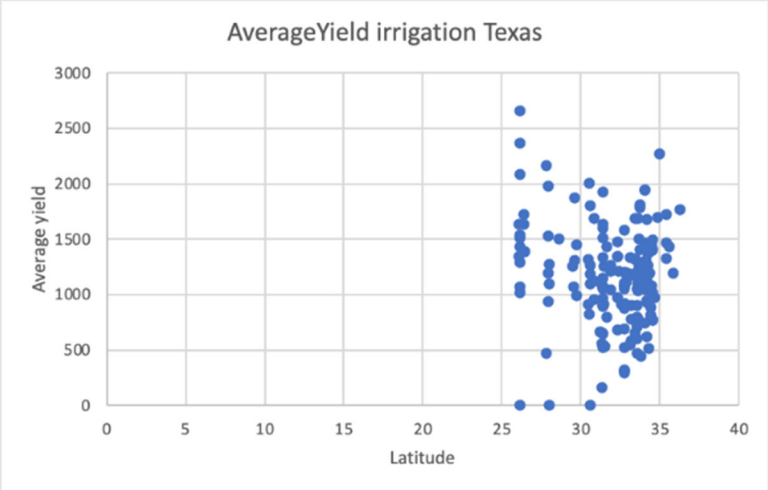

Figure 1: Latitude of irrigated cotton production locations in the state of Texas.

Despite the location cotton production sites display an increase in yield if there is irrigation available, as shown by Figure 2.

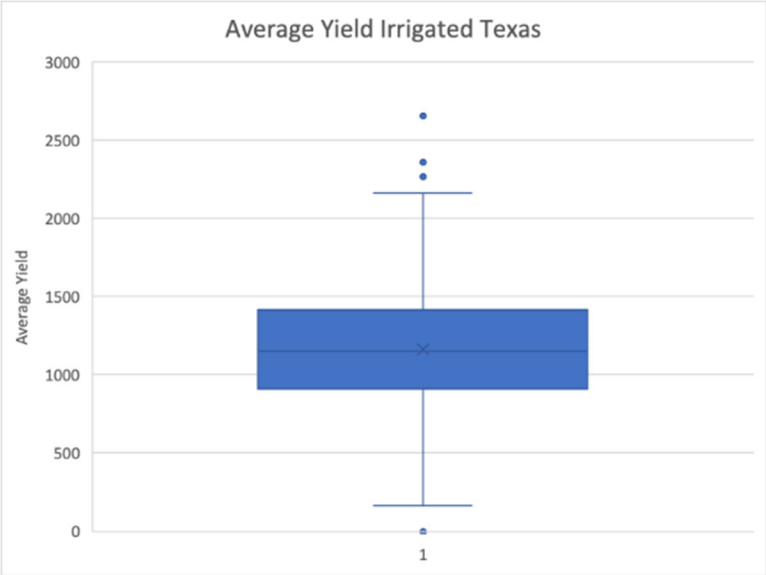

Figure 2: Average cotton yield of irrigated cotton production locations in the state of Texas.

On the other hand, cotton production sites display a decrease in yield if there is no irrigation available, as shown by Figure 4.

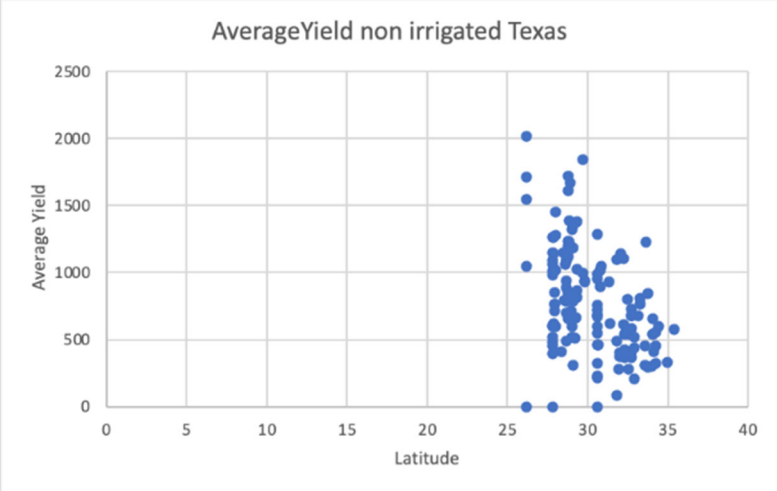

Figure 3: Latitude of non-irrigated cotton production locations in the state of Texas.

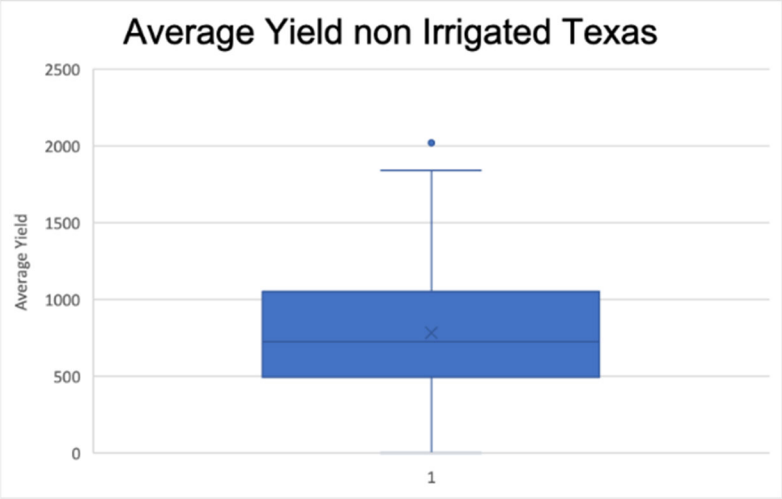

Figure 4: Average cotton yield of non-irrigated cotton production locations in the state of Texas.

An analysis was conducted to determine the relationship between irrigation availability and yield. The results of this analysis are shown in Table 2. According to the results irrigation does affect the yield of cotton in the state of Texas

Table 2: Impact of irrigation systems on cotton yield at cotton production locations in Texas.

t-Test: Two-Sample Assuming Unequal Variances		
	Irrigated Texas	*Non irrigated Texas*
Mean	1159.891268	783.2855928
Variance	197286.7522	154455.2663
Observations	173	137
Hypothesized Mean Difference	0	
df	304	
t Stat	7.908330575	
P(T<=t) one-tail	2.43488E-14	

t Critical one-tail	1.649881428	
P(T<=t) two-tail	4.86976E-14	
t Critical two-tail	1.967798141	

Ho = Irrigation does not change the yield of cotton.
Ha = Irrigation does change the yield of cotton.

Because, since the p-value of 4.86976E-14 is less than the stated significance level (α) of 0.05. Therefore, we reject the null hypothesis. The data support the claim that irrigation affects the yield of cotton in the state of Texas.

In a previous study, the irrigation availability factor was identified as a crucial and decisive factor to predict the productivity of a site. Therefore, irrigation availability was included as a factor to identify the optimum grouping projection. After identifying the impact of irrigation in cotton yield the study focused on exploring alternate relationships between cultivar trials and environment factors. Sand, silt, clay percentages, pH and electrical conductivity were selected due to their contribution to cotton production. Soil texture (a summation of proportions of sand, silt and clay content) is a very stable characteristic that influences soil biophysical properties, such as nutrient retention and drainage capabilities, and is largely unalterable. Soil pH regulates plant nutrient availability by controlling the chemical forms of the different nutrients and also influences their chemical reactions.
A clustering analysis was conducted with variety trial data from the state of Texas from the year 2014. The results of the unsupervised clustering analysis of this data set worked as the training and testing set of the prediction model. This model was tested and predicted the 2015 results.

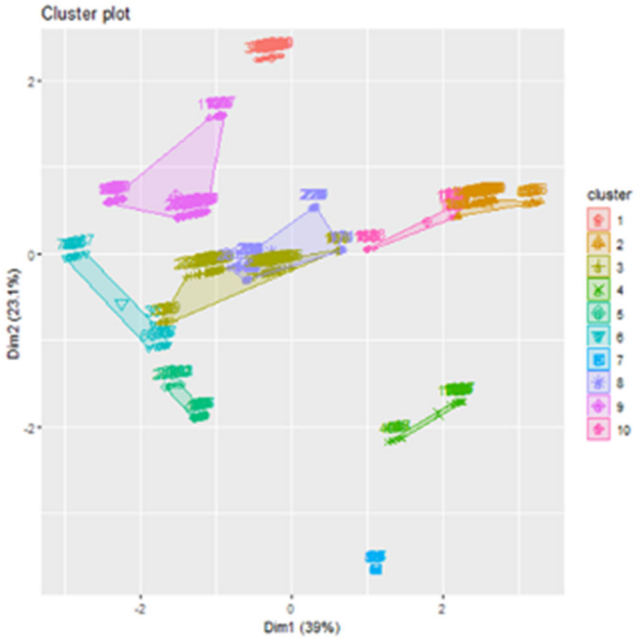

Figure 5: Unsupervised clustering analysis was performed to identify cotton production regions based on soil chemical and physical properties.

```
> yearcurrentPredResults
   groupPredictionForYearcurrent
     1  2  3  4  5  6  7  8  9 10
1   16  0  0  0  0  0  0  0  0  0
2    5 28  0  0  0  0  0  0  0  0
3    0  0 28  0  0  0  0  0  0  0
4    0  0  0 40  0  0  0  0  0  0
5    0  0  1  0 27  0  0  0  0  0
6    0  0  0  0  0 26  0  7  0  0
7    0  0  0  0  0  0 33  0  0  0
8    0  0  0  0  0  0  0 19  0  0
9    0  0  0  0  0  0  0  0 18  0
10   0  0  0  0  0  0  0  0  0 35
```

Figure 6: Training and testing 2014 cluster dataset (10 groups).

```
> yearcurrentPredResults
   groupPredictionForYearcurrent
     1  2  3  4  5  6  7  8  9 10
1    1  0  0 11  0  0  3  0  0 13
2   19 23  0  0  0  0  0  0  0  0
3    0  0  0  0  5  4 14  0  0  0
4    0  0  1  6 22  0  0  0  0  7
5    0  0  0  0  0  0  0  0 22  0
6    0  0  0  0  0  4  0  0  0  0
7    0  0  0  0  6  0 28  0  0  0
8    0  0  0  0  0 27  0 18  2  0
9    0  0  0  0  0 12  6  0  0  0
10   2  0 28  0  4  0  0  0  0 29
```

Figure 7: Confusion matrix results from testing the 2014 model to predict 2015 groups.

```
[1] "Yield summary for group  7"
> summary(allGroupRecords$yield)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  262.0   690.5   877.0   894.0  1130.0  1446.0
```

Figure 8: Example of yield summary for group number 7 from testing 2014 model to predict 2015 groups.

```
Cotton varities
"[Product,Yield]: [ PHY 333 WRF , 1355 ]"
"[Product,Yield]: [ PHY 333 WRF , 1355 ]"
"[Product,Yield]: [ ST 4946 GLB2 , 1382 ]"
"[Product,Yield]: [ ST 4946 GLB2 , 1382 ]"
"[Product,Yield]: [ PHY 499 WRF , 1446 ]"
"[Product,Yield]: [ PHY 499 WRF , 1446 ]"
```

Figure 9: Example of product suggestion generated based on results from testing 2014 model to predict 2015 groups.

After evaluating the prediction results of the model (for the state of Texas given 2014 information only), the model had an accuracy of 95.4% when predicting the yield at a given trial site in Texas. The 2014 cultivar trial-based model was then used to predict 2016 (Texas) outcomes. The model had an accuracy of 34.4%, this percentage represents the ratio of successes generated by the model. The system compared the model's results against the unsupervised clustering results to generate a confusion matrix. This model can be further improved by incorporating weather data into the model.

Summary

Variety Trial Validation: A Framework to Incorporate on-Farm Data predictive model using K means analysis and confusion matrix to develop a predictive model has an accuracy of 95.4% when predicting the yield at a given trial site in Texas. The 2014 cultivar trial-based model was then used to predict 2016 (Texas) outcomes the model had an

accuracy of 34.4% predicting future years. The model is being further developed to take into consideration weather data from planting to harvesting date. The addition of databases like soilDB and RNOAA to complement the current variety trial data is expected to make a more accurate predictive model based on location, soil chemical and physical properties, and weather data. The implementation of the predictive model is expected to determine the best suited cotton cultivar, based on farmers criteria, for a current producing field to a newly developed agricultural land.

Acknowledgements

This work was supported by Cotton Incorporated under Project No. 18-496

References

Breiman L., Friedman J. H., Olshen R. A., and Stone, C. J. (1984) *Classification and Regression Trees.* Wadsworth.

Cothren, J. T. (1999). Physiology of the cotton plant.'Cotton origin, history, technology, and production'.(Eds CW Smith, JT Cothren) pp. 207–268.

Hunt, L., & Jorgensen, M. (2011). Clustering mixed data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(4), 352-361.

Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning* (Vol. 1). Sthda.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344). John Wiley & Sons.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*(Vol. 26). New York: Springer.

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

Meredith Jr, W., Boykin, D. L., Bourland, F. M., Caldwell, W. D., Campbell, B. T., Gannaway, J., . . . Smith, C. (2012). Genotype× environment interactions over seven years for yield, yield components, fiber quality, and gossypol traits in the regional high-quality tests. *Journal of Cotton Science, 16*(3), 160-169.

Reddy, V. R., Baker, D. N., & Hodges, H. F. (1991). Temperature effects on cotton canopy growth, photosynthesis, and respiration. *Agronomy Journal*, *83*(4), 699-704.