RANDOM FORESTS FOR ESTIMATING CULTIVAR TRIAL PLANTING DATES

Raul Sebastian Martinez
Robert G. Hardin IV
Department of Biological and Agricultural Engineering
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

This study presents a predictive model to generate practical timeframes for cotton planting (planting week) applying a machine-learning method, Random Forests (RF). The model is proposed as an innovative solution to the absence of planting dates in datasets of past agricultural experiments. Advances in computing and data science allow new knowledge to be developed from analysis of combined data from past experiments; however, imputing missing data is a significant challenge. Planting date is a major management variable that needs to be tailored to anticipated weather conditions or weather phenomena. Farmers rely on their past experiences to select a particular planting date; however, this would only work for a particular area or location, due to the variability in environmental factors. For this study, the planting dates for a cotton cultivar trial dataset from 2005-2015 were predicted and then the developed model was used to predict the planting dates for subsequent years (2016-2018). The predicted plating dates for years 2016-2018 were then compared with the original recorded available planting dates of the respective years. The model used a dataset compiled solely for this project which included cotton cultivar trial data and data from the following public databases: NRCS soil surveys and National Oceanic and Atmospheric Administration (NOAA) daily weather data. Among the environmental factors (covariates) collected were soil texture, pH, and weather data which were then analyzed to denote each variable's importance to the dependent variable (planting week). To score the RF model we used the root-mean-square error (RMSE) metric. The RF model trained with our 2005-2015 cultivar trial complete dataset had an RMSE of 1.32. The out-of-bag validation score of the previously mentioned RF model was 1.41. Testing the RF model with 2016-2018 resulted in an RMSE of 1.40. Historical data indicates that planting date is a highly important decision given that it affects the maximum expected yield and cotton lint quality since only a certain quantity of growing degree days can be accumulated throughout the growing season.

Introduction

Planting date is a major management variable that needs to be tailored to anticipated weather conditions or weather phenomena. Farmers do not plant cotton at one planting date, even in a particular area or location, due to variability in the production environment. Therefore, for this study environmental factors (covariates) were collected (soil texture, pH, and weather data) and analyzed to define their influence along with their importance on the prediction accuracy of a planting date. Cotton planting date is a highly important decision a farmer makes during the year. This decision establishes a maximum possible yield and cotton lint quality since only a certain quantity of growing degree days (GDD) can be accumulated throughout the growing season. Mauget, Ulloa, and Dever (2019) conducted research regarding the effects of cotton planting date in the U.S. Southern High Plains (SHP), concluding that the planting date highly influences cotton yield due to its relevance with heat unit accumulation. To maximize GDD farmers must analyze weather factors such as precipitation, daily temperature, along with weather patterns of the current year to finally select the date that provides them with appropriate planting conditions. To approximate the planting and harvesting dates of the multiple locations being analyzed in this study, the GDD concept was used along with daily weather data retrieved from NOAA certified weather stations. The concept of growing degree days is based on a developmental threshold above which the crop grows. Cotton growth increases linearly as temperature increases between 60 and 88 °F according to studies performed at the University of Georgia (Ritchie et al., 2007). Cotton has a threshold temperature of 60˚F; averaging the maximum and minimum temperatures for each day and subtracting the threshold temperature results in heat units that accumulate consecutively and serve to indicate a particular growth stage (Landivar and Benedict, 1996). Accumulated heat units can be used to explain and predict the occurrence of events or duration of stages in crop development. Research studies from Georgia (Ritchie et al., 2007), Florida (Wright and Sprenkel, 2005) and Texas (Sansone et al., 2002) indicate that the total amount of heat units from planting to harvest is 2600 heat units.

With these environmental parameters and the phenology of cotton in mind, a prediction model was developed using the machine-learning method, random forests (RF). Breiman (2001) presented the RF algorithm, and it has shown to be a very successful general-purpose classification and regression tool. The method, which mixes numerous randomized decision trees and averages their predictions, has proved to perform well in situations where the number of variables exceeds the number of observations. This last characteristic is advantage compared with other machine learning methods given that cultivar trial datasets like the one used for this study can be as extensive as 112 variables. The application of RF on multi-variable problems such as the one presented will allows us to further expand our understanding of the genotype by environment interaction (GxE) components and their influence on the variation in yield and quality of cotton.

Materials and Methods

For this study, RStudio was used as the main integrated development environment to run the algorithms developed for this project. RStudio version 1.1.456 was installed in a Windows 10 environment with the following 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	tidyverse_1.3.0	rattle_5.4.0
cluster_2.1.0	plyr_1.8.6	bitops_1.0-6
forcats_0.5.0	rgeos_0.5-5	tibble_3.0.4
stringr_1.4.0	rgdal_1.5-18	rpart.plot_3.0.9
dplyr_1.0.2	sp_1.4-4	rpart_4.1-15
purrr_0.3.4	aqp_1.25	randomForest_4.6-14
readr_1.4.0	soilDB_2.5.8	caret_6.0-90
tidyr_1.1.2	rnoaa_1.2.0	lattice_0.20-45
ggplot2_3.3.2	caTools_1.18.0	lubridate_1.8.0

Planting Date Predictive Model Based on Soil Biophysical Properties and Daily Weather Data

The predictive model developed for this study was generated using data originally extracted from Seedmatrix, a web-based application with a database of trial data from 2005-2015. Data from multiple states was extracted and first examined to remove unnecessary data. Information related to location, yield and product were maintained. 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 cultivar trial locations x year combinations from 2005-2015. To complement the trial data, information was retrieved from the publicly available State Soil Geographic (STATSGO) dataset.

STATSGO is one of three soil geographic databases established by the United States Department of Agriculture's (USDA) Natural Resources Conservation Service (NRCS). The thorough soil survey geographic (SSURGO) information is used to create STATSGO soil maps. Soil survey maps are elaborate spatial environmental datasets and for many surface and subsurface process models, they constitute the principal source of soil physical and chemical attributes. The SSURGO data uses map units, identified with a unique MUkey for each soil survey, which is used in the SQL code to retrieve the soil properties for every cultivar trial location in the dataset. Incorporating the SQL code in a R script with the R package, SoilDB, allowed us to retrieve the complementary data based on the latitude and longitude recorded in the variety trial data.

Planting and Harvest Date Estimation

The first aspect to discuss is maturity of the boll. Maturation starts right after the plant has achieved cut-out. Cotton physiologists define cutout to be when the terminal is five nodes above the uppermost first position white flower (Byrd, 2018). Before then, approximately 100 flowers will produce 1 pound of seed cotton. After cotton reaches cutout, the number of flowers needed to produce 1 pound of seed cotton increases dramatically. The boll maturation phase begins as the boll reaches its full size and maximum weight. During this phase, fiber and seed maturation take place and boll dehiscence occurs. The capsule walls of the boll dry, causing the cells adjacent to the dorsal suture to shrink unevenly. This shrinking causes the suture between the carpel walls to split, and the boll opens (Sansone et al., 2002). After the first boll opens the next action to consider is to either apply or not apply harvest aids. Deciding when to apply harvest aids are a compromise between slowing or halting further development of green bolls and minimizing the weathering of open bolls. If the aids are applied too early, the plant may not defoliate (stops further development) properly; if applied too late, yield and lint quality may be compromised. applied between 50 and 70 percent open bolls. Seed development also is a good indicator of boll (Sansone et al., 2002). The level of maturity and the defoliation application are aspects that each study must have considered to conclude that 2600 heat units is the total amount from planting to harvest ready.

Multiple researchers (Ritchie et al., 2007; Sansone et al., 2002; Wright & Sprenkel, 2005) agree on the total amount of heat units, 2600, required in cotton from planting until harvest ready. However, there are scenarios where harvesting after reaching 2600 heat units is not possible, then this activity needs to be postponed. Studies show that delaying harvest for 4 to 6 weeks does not reduce yield or quality significantly in years when there is less than 1 inch of rain after bolls open (Dodds, Reynolds, Barber, & Raper, 2017).

Based on the research discussed on the previous paragraphs an approximation algorithm was developed. This algorithm takes known soil biophysical properties and daily weather data from an area (trial site) as inputs and generates an approximation. The developed method to approximate the planting date works as follows. The first stage consists of approximating the week and month when planting could potentially take place. To accomplish this, the algorithm first analyzes the year-round temperature distribution in the area (trial site) and determines the time frame where the minimum monthly average temperature is suitable for cotton development. To further narrow down the timeframe where planting could potentially take place, the U.S. Climate Normals are considered (Skaggs & Irmak, 2012), specifically the historical estimated date of the year with 30% probability of the last 36°F occurrence. The official normals are calculated for a uniform 30-year period, and consist of annual/seasonal, monthly, daily, and hourly averages and statistics of temperature, precipitation, and other climatological variables from almost 15,000 U.S. weather stations. The indicated date by the normals then becomes the lower bound of the timeframe that will then be reviewed to define the harvest date, total accumulated degree days in season and length of season. According to multiple studies (Wright & Sprenkel, 2005; Sansone et al., 2002) the estimated amount of heat units required for planting to seedling emergence are 50-60. Applying the 5-day degree-day forecast for cotton planting (Munier, Goodell, & Strand, 2004), allowed us to estimate the planting date for the given year. The harvest date estimation process then takes place. this process consists of computing the accumulated heat units throughout a timeframe and determining if the total amount of heat units is greater than or equal to 2600 and if the minimum day requirement is equal or greater than 130.

The following data elements total accumulated degree days, total days in season, planting day of the year and planting week of the year are by-products originated from the process described before. These elements constitute all the weather data-based elements in the final dataset for this study.

Oceanic Niño Index (ONI) Influence on Yield and Planting Date

The insecurity that comes with weather fluctuation is a never-ending problem for agriculture. Crop management that is ideal one year may result in crop losses, overproduction, or pollution the next year when the weather is distinct. Evidence is growing that El Niño–Southern Oscillation (ENSO) influences crop production through its influence on weather patterns in some regions (Paz et al., 2012). The official ENSO indicator is the Oceanic Niño Index (ONI), which is based on sea surface temperature (SST) in the east-central tropical Pacific Ocean. Hansen, Hodges and Jones

(1998) concluded that cotton along with other crops (peanut, tomato, tobacco, corn, and soybean) in Alabama, Florida, Georgia, and South Carolina are vulnerable to the influence of ENSO. The ENSO phase significantly influenced the yields and regional patterns of yield in soybean and cotton. Due to the identified significant responses of cotton yields to SST, the six-month ONI tendency before planting was included in the final dataset. The mentioned ONI tendency was inferred by taking the ONI value mean of the previous six months before planting date, this is best represented by Eq. (1) shown below.

$$\text{Six Month ONI Tendency} = \frac{1}{6}\sum_{k=0}^{6} ONI_k \tag{1}$$

Random Forest Model

To build the model the function randomForest() from the package randomForest_4.6-14 was used. The 2005-2015 cultivar trial complete dataset was split into training (70%) and validation (30%) subsets. Figure 1 shows a county map of the state of Texas identifying the locations where the cultivar trials took place, as specified by the coordinates in our 2005-2015 dataset. In addition to its predictive capability, RF can also provide useful information about variable importance and dependence. The variable importance rank and the partial impact of the variable on the response can be evaluated for the purpose of systems analysis. We used a commonly used variable importance measure (Khalilia, Chakraborty, & Popescu, 2011), mean increase in accuracy (%IncMSE in r randomForest output), to identify the most influential variables in our model.

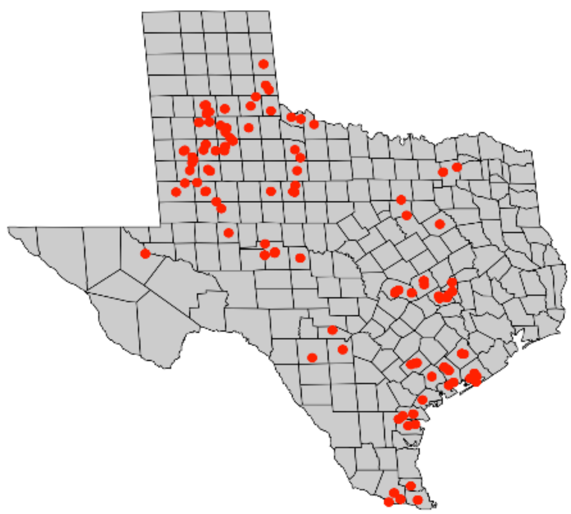

Figure 1. Texas cultivar trial locations (2005-2015) in training dataset.

Cultivar trial data from 2016-2018 was used as the testing set due to our interest on the performance of the model on new, unseen data points. This dataset included actual planting and harvest dates.

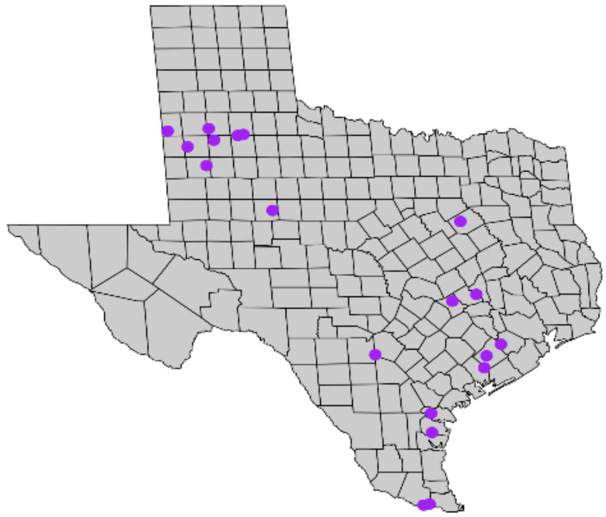

Figure 2. Texas cultivar trial locations (2016-2018) in test dataset.

Results and Discussion

We evaluated a machine-learning method, Random Forests (RF), for its ability to predict the planting week in response to climate and biophysical variables at a state scale (Jeong et al., 2016). The variable importance measurements are shown in Table 1. Not surprisingly, latitude was the most important variable. The next most influential variable was the ONI tendency before planting, which had a much larger mean increase in prediction accuracy than the year, indicating that ENSO accounts for much of the year-to-year variability in the planting date for the cotton cultivar trials. This result has important implications for cotton farmers in areas with more marginal growing conditions. Predicting in advance if planting is likely to proceed earlier or be significantly delayed due to ENSO could be useful for optimizing cropping decisions.

Table 1. Final dataset variable importance measurements.

Variable	%IncMSE
Latitude	76.7
Longitude	9.1
Year	14.8
Average Sand Content	-0.2
Average Silt Content	4.7
Average Clay Content	3.9
Average EC	2.5
Six-Month ONI tendency Before Planting Date	40.1
Average pH	7.2

RF was found highly capable of predicting the potential planting week for the cotton cultivar trial. To score the RF model we used the root-mean-square deviation or root-mean-square error (RMSE), it is a frequently used measure of the differences between values predicted by a model or an estimator and the values observed (Breiman, 2002). After training the RF model with our 2005-2015 cultivar trial complete dataset, the RMSE was 1.32 (Figure 3).

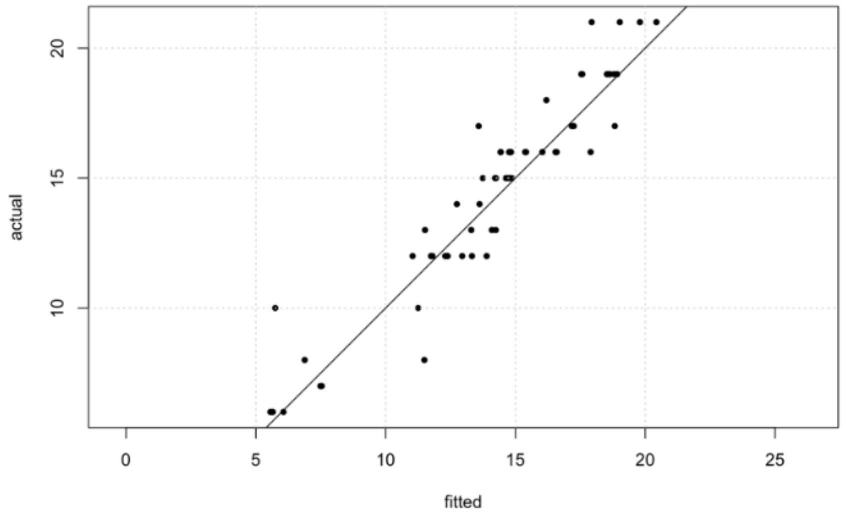

Figure 3. Actual and RF predicted planting week, training data (2005-2015 cultivar trial dataset).

To obtain a more realistic view of the predictive power of the model, the previous model was tested applying the hold-one-out validation method. RMSE of the validation test was 1.41.

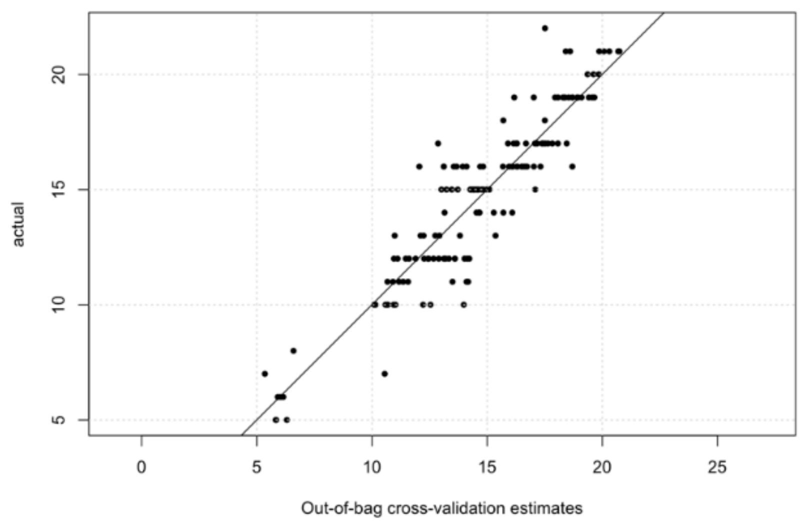

Figure 4. Actual planting week and validation estimates of RF model based on 2005-2015 cultivar trial dataset.

Finally, when the predictive model based on 2005-2015 dataset was tested on 2016-2018 data, the RMSE for predicting planting week was 1.40 (Figure 5).

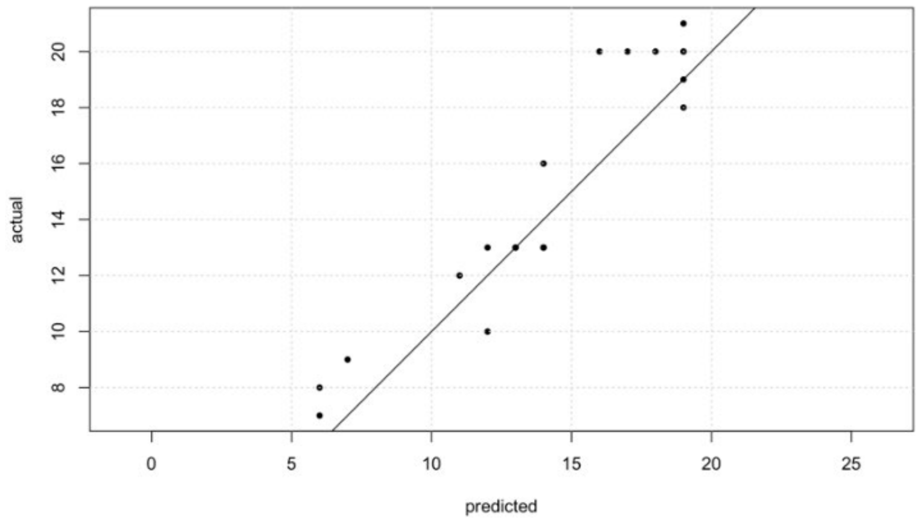

Figure 5. Actual and RF predicted planting week, 2016-2018 test data.

Our results show that RF is an effective and versatile machine-learning method for cultivar trial planting week estimation at state scale because of its high accuracy and precision, ease of use, and utility in data analysis.

Summary

Planting date is a highly important decision given that it affects the maximum expected yield and cotton lint quality since only a certain quantity of degree days can be accumulated throughout the growing season. The cotton cultivar trials data obtained from the Seedmatrix site lacked this valuable data element for many observations. Therefore, we could not determine the weather effects fully. Therefore, a method was developed to indirectly estimate planting dates based on location, soil chemical and physical properties, and weather data. This problem of missing data in field experiments is not unique, so a solution to this problem may be applicable in other studies. An RF predictive model was trained and validated with data from 2005-2015 and tested for its ability to predict 2016-2018 planting dates. The RF model trained with our 2005-2015 cultivar trial complete dataset had an RMSE of 1.32. The out-of-bag validation score of the previously mentioned RF model was 1.41. Testing the RF model with 2016-2018 data resulted in an RMSE of 1.40. The model developed in this study will be included in the toolset of future models that are expected to determine the best suited cotton cultivars for any given location, given a particular farmer's management strategy.

Acknowledgements

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

References

Breiman, L. 2001. Random forests. Machine Learning 45:5-32.

Byrd, S. 2018. Defining Cutout in Cotton. Oklahoma Cooperative Extension Service. Publication No. PSS-2181.

Dodds, D. M., Reynolds, D. B., Barber, L. T., and Raper, T. B. 2017. 2017 Mid-South cotton defoliation guide.

Hansen, J. W., Hodges, A. W., and Jones, J. W. 1998. ENSO influences on agriculture in the southeastern United States. Journal of Climate, 11(3), 404-411.

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., . . . Reddy, V. R. 2016. Random forests for global and regional crop yield predictions. PLoS One, 11(6), e0156571.

Khalilia, M., Chakraborty, S., and Popescu, M. 2011. Predicting disease risks from highly imbalanced data using random forest. BMC medical informatics and decision making, 11(1), 1-13.
Mauget, S., Ulloa, M., and Dever, J. 2019. Planting date effects on cotton lint yield and fiber quality in the US Southern High Plains. Agriculture, 9(4), 82.

Munier, D., Goodell, P., and Strand, J. 2004. Accuracy of cotton-planting forecasts assessed in the San Joaquin Valley. California agriculture, 58(3), 164-168.

Paz, J. O., Woli, P., Garcia, A. G., and Hoogenboom, G. 2012. Cotton yields as influenced by ENSO at different planting dates and spatial aggregation levels. Agricultural Systems, 111, 45-52.

Ritchie, G. L., Bednarz, C. W., Jost, P. H., and Brown, S. M. 2007. Cotton growth and development. University of Georgia Bulletin 1252.

Sansone, C., Isakeit, T., Lemon, R., and Warrick, B. 2002. Texas cotton production: Emphasizing integrated pest management. Texas Cooperative Extension: College Station, TX.

Skaggs, K. E. and Irmak, S. 2012. Long-term trends in air temperature distribution and extremes, growing degree-days, and spring and fall frosts for climate impact assessments on agricultural practices in Nebraska. Journal of Applied Meteorology and Climatology, 51(11), 2060-2073.

Wright, D. L., & Sprenkel, R. (2005). Cotton growth and development. University of Florida/Institute of Food and Agricultural Sciences Extension Service. Publication No. SS-AGR-238.