Using machine learning to sharpen agronomic insights to improve decision making in Australian cotton systems

Kavina Dayal[1], Tim Weaver[2], Michael Bange[2], CSD Ltd. Extension & Development Team[3]

[1] CSIRO Agriculture & Food, 15 College Road, Sandy Bay, TAS, 7005, www.csiro.au, kavina.dayal@csiro.au
[2] CSIRO Agriculture & Food, 21888 Kamilaroi Highway, Narrabri, NSW, 2390
[3] Cotton Seed Distributers Ltd., 2952 Culgoora Road, Wee Waa, NSW 2388

Abstract
The ability to understand the impact of genetics x environment x management (GxExM) influences at a farm and paddock scale offers significant opportunities for informing management interventions to raise crop productivity. New means of agronomic data collection and collation, along with machine learning statistical approaches can help realise these opportunities. Cotton Seed Distributors Ltd. agronomy and extension team collect a large number of crop physiological and agronomic characteristics every year in their key varieties across the whole industry. Over the past four seasons this has resulted in the collection of a significant dataset for in-depth modelling. A machine learning algorithm (i.e., Random Forest) has been applied to understand which measured variables affect yield which can be used to identify management interventions. To evaluate the approach, the Random Forest method was applied to the dataset using key variables only during first flower. Variables were then used to predict yield at this stage ($r^2 = 0.74$). The machine learning algorithm is intended to form the back bone of a decision tool so that crop managers can access the insights being generated from the dataset in real time and project current crop performance, giving them the ability to investigate the consequences of management interventions.

Key Words
Machine learning, Random Forest, cotton yield prediction, Cotton Seed Distributors

Introduction

Cotton Seed Distributors Ltd. (CSD) and CSIRO co-jointly invests in the research and breeding of future cotton cultivars for Australian cotton growers. For the past four seasons, CSD has been capturing crop physiological and agronomic parameters from their 'Ambassador Cotton Variety Trials' in both irrigated and rainfed cotton crops across all the major cotton growing regions. This data can be used to understand the key features that impact cotton yield during the course of the season. This is important, as many of the management interventions for improving yield have been a result of in-season management. Industry yield improvement in cotton has increased by 24% due purely to a cultivar by management interaction (Liu et. al., 2013). To continue this improvement is to explicitly exploit the GxExM interactions, and the application of machine learning, using real-time crop specific agronomic information to provide informed decisions is seen as one strategy to do this.

Machine Learning techniques are a major research quest for assisting crop yield prediction.

The two key aims of this endeavour and presented here are: (1) Identify key variables that influence cotton yield during first flower, and (2) Predict cotton yield based on the key variables in their order of importance.

Data and Methods

Crop Physiological and Agronomic Data

This study utilises the CSD Ambassador Variety data from four summer seasons (2014/15, 2015/16, 2016/17 and 2017/18). The data captured various crop phenological and agronomic parameters at key growth stages in cotton, from planting date to final yield (bales/ha) for all cotton growing regions in NSW, QLD and VIC. To demonstrate the assessment approach by machine learning, the analysis presented in this abstract only utilised variables at first flower. To ensure appropriate analysis without bias, all variables with missing data were discarded from the analysis. This resulted in 13 variables used at first flower (FF; early) stage being included in the analysis: general information (GI) variety, GI 7 day germination (days), GI cool germination (%), establishment measurements (EM) final estimate per metre, mid squaring (MS), nodes to first fruiting branch, first flower (FF) days after planting, FF day degrees (DD), FF squaring nodes, FF total nodes, FF

nodes above white flower (NAWF), FF first position retention of fruit (%), FF plant height (cm), and the final yield in bales/ha (1 bale = 227 kg lint). Table 1 provides the definition of these variables.

Table 1: Definition of First Flower variable inputs used in machine learning models. (The data collection was categorised into four stages: General Information, Establishment, First Flower and Mid-squaring)

Data collection Stages	Acronym	Variable	Definition
General information	GI	Variety	Ambassador Varieties: Sicot 714 B3F, Sicot 746 B3F and Sicot 748 B3F
		7 Day Germination	Number of seeds germinated after 7 days from seed imbibition.
		Cool Germination %	Germination at 18^{o}C after 7 days on 200 seeds
Establishment	EM	Final Estimate per metre	Final count of emerged seedlings after 10 days
First Flower	FF	Days after planting	Days after seed imbibition to the first white flower
		Day Degree Accumulation	Day Degrees = (maximum temperature – 12) + (minimum temperature – 12) / 2
		Squaring Nodes	Number of nodes on a plant with branches producing squares on the main stem (the development of cotton fruit).
		Total Nodes	Total nodes above the cotyledon nodes
		Nodes above last white flower (NAWF)	Nodes above the last white flower
		First Position Retention of Fruit	The first fruiting node on the main stem from the cotyledon nodes that retained a viable fruit
		Plant Height	Total height from ground to growth tip.
Mid-squaring	MS	Nodes to First Fruiting Branch	The nodes to the first fruiting branch from the cotyledon nodes.

Random Forest Modelling

A Random Forest model was used to identify the variables in term of their order of importance in affecting final cotton yield. The Random Forest model used a regression type with 500 trees and 13 independent variables. The data was then separated into training (85%) and testing (15%). The Random Forest model was trained on the training dataset in order to learn patterns and create relationships between predictor variables and final cotton yield, and then used to predict the yield. The predicted yield was compared with the observed yield (observed vs predicted for validation).

Results

Key Determinants of Cotton Yield

Figure 1 shows the order of importance of variables from the Random Forest analysis at first flower. The first flower (FF) in days after planting and day degrees to first flower (FF DD (day degrees)) variables were identified as the two key determinants of cotton yield with 372 and 351 weights (the higher the weighting indicates the increased importance in predicting yield), respectively. In terms of cotton growth and physiology this highlights that the actual timing of flowering is important most likely reflecting the regional differences in the dataset. The rate of development to first flower represented by the FF DD most likely

reflects the impacts of the environment leading up to flowering that effects the size of the crop (leaf area and number of nodes). Leaf area was not measured, however, the number of nodes to first branch was measured and was a significant variable affecting yield. Interestingly the cotton variety, however, was shown to be the least weighted variable alone to influence yield. While the variety would have influence on some of the variables, it does highlight that the GxExM effect is a significant opportunity to influence yield.

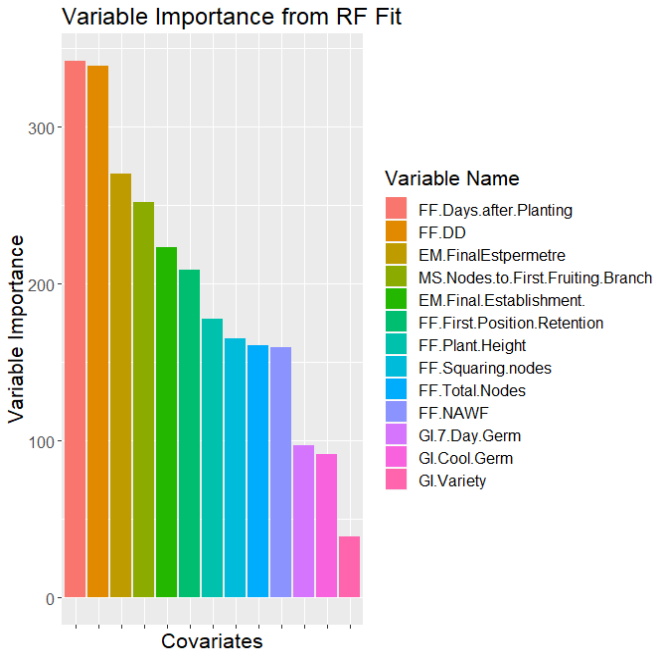

Figure 1: Random Forest generated variable order of importance (higher values of importance = higher importance of the variable to predict yield). (FF = First Flower; DD = Day Degrees; EM = Establishment Measurements; NAWF = Nodes after White Flower; MS= Mid-squaring; GI = General Information).

Cotton Yield Prediction

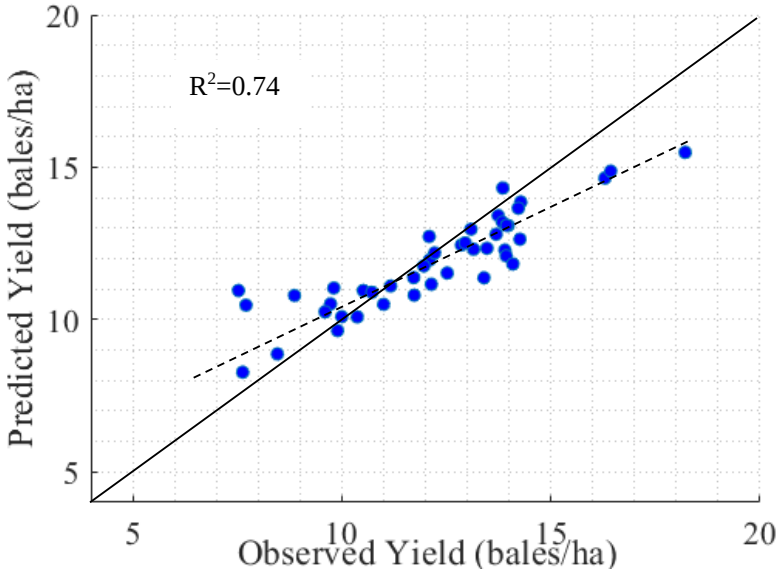

Figure 2: Scatterplot of observed versus predicted cotton yield using the key crop physiological variables captured to first flower. (The red line is the best fit.)

Using the variable weights according to their order of importance, the Random Forest can be used to make a yield prediction. Figure 2 shows the scatterplot of the observed versus predicted yield that yielded an R-squared value of 0.74 and RMSE of 1.32 bales/ha.

It is important to note that machine learning models are data driven and therefore require large quantities of data in order to train the model. This has been a limitation in our study where total data points were only 621. In order to obtain higher prediction accuracy, more training data is necessary. Although it is possible that there are key variables missing or the required knowledge is not available to allow better predictions. One key element being investigated is the addition of climate indices that represent the forecast or seasonal outlook to improve yield predictions.

Conclusion

This study utilised a Random Forest machine learning algorithm to determine the variables from early phenological stage of cotton growth to predict key determinants affecting final cotton yield. The number of days to first flower after planting, and day degrees to first flower were shown to have the highest effect, and were the top two key variables for predicting cotton yield. Accuracy can be increased with the addition of more training data collected from the mid and late cotton physiological stages. The power of machine learning adds the ability to learn patterns and better predict outcomes from the models developed. If more data can be added to train the model, the prediction would be more accurate and reflective of the changing environments as opposed to linear model that are fixed in their predicting or estimating ability.

Future studies will consider mid and late season crop data to identify key determinants of the final yield. It will also group the cotton growing regions based on climatic zones.

References

Liu SM, Constable GA, Reid PE, Stiller WN and Cullis BR (2013). The interaction between breeding and crop management in improved cotton yield. Field Crops Research 148, 49-60.

Acknowledgments

The collection and collation of the CSD Ltd. Ambassador data was undertaken by the CSD Ltd. Extension and Development Team: James Quinn (Lead), Alice Curkpatrick (Gwydir), Chris Teague (Border Rivers and Balonne), Bob Ford (Namoi Valley and Walgett), Craig McDonald (Central NSW), Jorian Millyard (Southern NSW), Chris Barry (Darling Downs and Central Queensland), Sam Lee (Queensland), Lucy Burrows (Riverina), Angus Marshall (Namoi) and Larissa Holland (Darling Downs).