Agron. Sustain. Dev. (2016) 36: 27
DOI 10.1007/s13593-016-0364-z

RESEARCH ARTICLE

Accurate prediction of sugarcane yield using a random forest algorithm

Yvette Everingham[1,2] · **Justin Sexton**[1,2] · **Danielle Skocaj**[2,3] · **Geoff Inman-Bamber**[2,4]

Accepted: 22 March 2016 / Published online: 19 April 2016

Abstract Foreknowledge about sugarcane crop size can help industry members make more informed decisions. There exists many different combinations of climate variables, seasonal climate prediction indices, and crop model outputs that could prove useful in explaining sugarcane crop size. A data mining method like random forests can cope with generating a prediction model when the search space of predictor variables is large. Research that has investigated the accuracy of random forests to explain annual variation in sugarcane productivity and the suitability of predictor variables generated from crop models coupled with observed climate and seasonal climate prediction indices is limited. Simulated biomass from the APSIM (Agricultural Production Systems sIMulator) sugarcane crop model, seasonal climate prediction indices and observed rainfall, maximum and minimum temperature, and radiation were supplied as inputs to a random forest classifier and a random forest regression model to explain annual variation in regional sugarcane yields at Tully, in northeastern Australia. Prediction models were generated on 1 September in the year before harvest, and then on 1 January and 1 March in the year of harvest, which typically runs from June to November. Our results indicated that in 86.36 % of years, it was possible to determine as early as September in the year before harvest if production would be above the median. This accuracy improved to 95.45 % by January in the year of harvest. The R-squared of the random forest regression model gradually improved from 66.76 to 79.21 % from September in the year before harvest through to March in the same year of harvest. All three sets of variables—(i) simulated biomass indices, (ii) observed climate, and (iii) seasonal climate prediction indices—were typically featured in the models at various stages. Better crop predictions allows farmers to improve their nitrogen management to meet the demands of the new crop, mill managers could better plan the mill's labor requirements and maintenance scheduling activities, and marketers can more confidently manage the forward sale and storage of the crop. Hence, accurate yield forecasts can improve industry sustainability by delivering better environmental and economic outcomes.

Keywords APSIM · Agriculture · Nitrogen · Fertilizer · Value chain · Random forest

1 Introduction

The Green Revolution saw agricultural industries worldwide increase productivity through advancements in research, development, and technology transfer during the first half of the twentieth century. Unfortunately, around the 1980s, agricultural yields of major crops grown all around the world reached their ceiling and flat-lined. This plateau presents an enormous challenge for society. In 2009, the Food and Agriculture Organization of the United Nations (FAO) predicted that agricultural production would need to increase by 70 % to sustain a population that is expected to exceed 9 billion by 2050 (FAO 2009). This challenge is made even more difficult by

✉ Yvette Everingham
yvette.everingham@jcu.edu.au

[1] Centre for Tropical Environmental & Sustainability Science, James Cook University, Townsville, Australia

[2] College of Science, Technology and Engineering, James Cook University, James Cook Drive, Townsville 4811, QLD, Australia

[3] Sugar Research Australia, Tully 4068, QLD, Australia

[4] Crop Science Consulting, Townsville 4811, QLD, Australia

constraints imposed by climate change, climate extremes, and new laws and regulations that govern industry practices.

While the Green Revolution ended several decades ago, the Big Data Revolution has only just begun. Every day, the world collects more than 2.5×10^{18} bytes or 2.5 exabytes of data. That is equivalent to 1.7 trillion 3.5-in. floppy disks of data per day or 100 million 25 GB smartphones. The term "Big Data" refers to (i) the volume and variety of data collected; (ii) the velocity these data can be captured; and (iii) our ability to filter, analyze, and discover patterns in large data sets. Many companies have exploited this explosion of Big Data to dramatically increase profit margins. For example, Paul (2012) describes how a data mining approach was used to develop a Facebook campaign to promote a candy bar to 17-year-old males. The Facebook campaign resulted in production consumption growth of 24 % over a 6-month period. Wal-Mart has integrated smart technologies and a data mining approach with mobile devices to develop and transmit a shopping list to its in-store customers via an app (Cao and Manrai 2014). Users of the Wal-Mart app were found to spend 77 % more than non-app users every month.

Governments have also turned to Big Data to make "Smart Cities" (Caragliu et al. 2011; Perera et al. 2014). A Smart City is a city that is made more efficient through effective integration of Big Data technologies in ways that benefit the city's inhabitants. In an overview of a special issue of the *Journal of Urban Technology* on Smart Cities, Allwinkle and Cruickshank (2011) identified cities such as San Diego, Amsterdam, and Brisbane as forerunners of this revolution. The world eagerly waits to learn if agricultural industries can become "Smart" agricultural enterprises. The "Green Data Revolution" is a term born from the optimism that Big Data can and will deliver benefits to agricultural industries and global society, in a similar way the Green Revolution sparked an increase in agricultural productivity.

Recognizing the potential of Big Data to revolutionize the Australian sugar industry, the major sugarcane funding body in Australia (Sugar Research Australia) has invested in this research area. Key priorities like precision agriculture, plant breeding, and spatial data hubs for research and extension incorporate Big Data technologies. Yield prediction is another area that relies on modern data mining methods. Forecasting the size of the crop can improve industry sustainability. For example, farmers could target their applied nitrogen rates to the size of the forthcoming crop, marketers could decide how much crop to sell on the futures market, and millers could ensure mill maintenance schedules are completed in time for the start of the sugarcane crushing season.

Everingham et al. (2015b) reported the benefits that a modern data mining method offers over contemporary, time-honored methods like stepwise linear regression modelling. These authors used a random forest modelling technique (Breiman 2001) to investigate how climate attributes relate to sugarcane productivity in the Victoria, Bundaberg, and Condong sugar mill regions in Australia. The key advantage of the random forest technique is it can investigate nonlinear and hierarchical relationships between the predictors and the response using an ensemble learning approach. Ensemble methods involve making multiple attempts from different data or models to predict a response variable like sugarcane yields. Using multiple efforts to predict a response can increase the robustness and accuracy of predictions compared to using any single data set or model (Breiman 2001; Everingham et al. 2009). We stress that random forest models should not be confused with the single decision tree approach like that adopted in classification and regression trees (De'ath and Fabricius 2000). Although there exist situations when random forests have been outperformed by traditional linear approaches (García-Gutiérrez et al. 2015), there are many cases where random forests have outperformed traditional linear regression (Craig and Huettmann 2009; Guo et al. 2015) and linear discriminant analysis approaches (Everingham et al. 2007b; Gromski et al. 2014). Consequently, random forests have been applied in a number of agricultural related applications.

Random forests have been used to predict yields directly for mangoes (Fukuda et al. 2013) and have been incorporated into a complex seasonal yield forecasting model for crops in Canada (Newlands et al. 2014). Tulbure et al. (2012) used random forest regression to identify important variables for switchgrass yields across the USA. They identified nitrogen fertilizer, cultivar, rainfall, stand age, and soil silt levels as the most influential of 22 predictor variables. The variables identified by random forests were then used to build better models of switchgrass yield.

Random forest analysis of Big Data sets has also been used to investigate other important issues in agriculture such as nitrous oxide (N_2O) emissions (Philibert et al. 2013), leaf nitrogen levels (Abdel-Rahman et al. 2013), and drought forecasting (Chen et al. 2012). Abdel-Rahman et al. (2013) used random forest regression to build predictive models of sugarcane leaf nitrogen levels from hyperspectral satellite images, while Philibert et al. (2013) were able to identify nitrogen fertilization, crop type, and experiment duration as the most important predictor variables of N_2O emissions. Saussure et al. (2015) incorporated random forests as part of a data processing framework to develop preventive solutions for the sustainable control of wireworms and Everingham et al. (2007b) successfully used random forests to classify sugarcane variety and the number of times the sugarcane has been harvested and allowed to regrow.

Uniquely situated between world heritage rainforest areas and the Great Barrier Reef, the Tully sugarcane-growing region in Australia (Fig. 1) is under increasing pressure to integrate new knowledge and technologies that promote

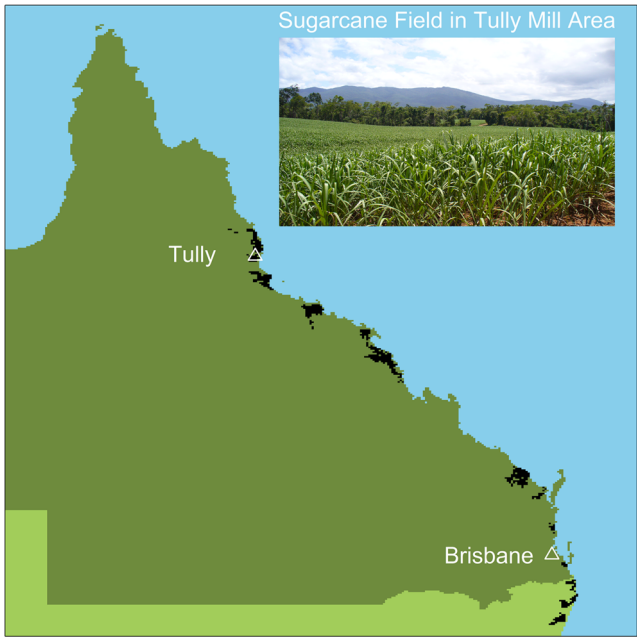

Fig. 1 Sugarcane grown along the eastern coast of Australia. *Black areas* indicate regions where sugarcane is grown. The location of the Tully sugarcane mill is noted as the focus area of this study. *Inset* image of a sugarcane field in the Tully sugar mill area

sustainable agricultural practices. The Tully sugarcane-growing region averages 4000 mm rain each year, making the town of Tully the wettest in Australia and few places around the world can compete with its natural swings in year-to-year climate variability (Nicholls et al. 1997). Owing to the high rainfall, this sugarcane-growing region tends to produce the highest yields following an El Niño event which favors below average rainfall during the major growing periods of austral spring and summer (Everingham et al. 2003; Skocaj et al. 2013). If industry practices can be shaped to suit the size of the forthcoming crop, a number of benefits to both industry and the environment can be realized. Given the advantages offered from early and accurate predictions of sugarcane yields and the ability of data mining techniques to extract patterns in large and complex data sets, the objective of this paper is to determine how accurately the random forest data mining method can estimate sugarcane productivity in the Tully sugarcane-growing region in the pursuit of advancing industry sustainability.

2 Materials and methods

2.1 Research approach

Classification and regression random forest models (Breiman 2001) were built using climate and productivity data from 1992 to 2013. The response variable for classification models was above or below the observed median yield (t ha^{-1}) for the Tully, Australia, sugarcane-growing region (Fig. 1). The response variable for regression models was regional yield in the same region. Yield estimates were produced on the 1 September of the year prior to harvest and on the 1 January and the 1 March of the year of harvest.

Daily climate indices, seasonal climate forecasting indices, yields of previous years, and APSIM (Agricultural Production Systems sIMulator) (Keating et al. 1999) simulated biomass indices were used as predictor variables in the classification and regression models. Daily climate, seasonal climate, and the APSIM biomass predictor variables were calculated until the end of each month. Only variables computed before the yield estimation dates were supplied to the random forest modelling techniques. For example, if the random forest classification model attempted to explain yields on the 1 January in the year of harvest, predictor variables computed no later than the 31 December in the year before harvest could form part of the model's inputs.

2.2 Productivity data

Regional yields from 1992 to 2013 were obtained for the Tully sugarcane-growing region in Australia (Fig. 1). The start date of 1992 was chosen because it followed the rapid expansion of sugarcane into the nearby Murray Valley. The finish date of 2013 was selected because this represents a time that data were available when the research commenced. Trends in yields for this period were tested using a t-procedure for the slope of a straight line equation fitted by the method of least squares (Zar 1999). Yields for this period contained no significant trends over time ($p = 0.078$). As requested by industry, classification models were used to forecast the direction of the crop as either above or below the observed median of 86.65 t ha^{-1} while regression models were used to forecast cane yields (t ha^{-1}).

2.3 Model predictor variables

A range of predictor variables that could be related to crop size was entered in the random forest classification and regression models (Table 1). These included variables based on indices for simulated biomass; previous yields; local climate data consisting of rainfall, radiation, and maximum and minimum temperature. Long-range climate indices that included the Southern Oscillation Index which is derived from sea level air pressures between Tahiti and Darwin and Niño 3.4 region SST anomalies which are derived from sea surface temperatures in the central equatorial Pacific Ocean were also independent variables. All local climate data were obtained from the SILO patched point data repository for the Bureau of Meteorology weather station located at the Tully Sugar Mill station number 32042. These climate data are available from the Long Paddock

Table 1 Predictor variables supplied to the random forest classification and regression analysis

Name	Units	Details
AI	g m^{-2}	The Agricultural Production Systems Simulator (Keating et al. 1999) was used to simulate sugarcane accumulated biomass throughout the growing season. The APSIM biomass index (AI) was generated using an ensemble modelling approach (Everingham et al. 2015a; Everingham et al. 2009). The Ensemble approach considered a range of APSIM parameterizations represented in the Tully region. The AI was generated for the end of each month of the growing season from June, the year before harvest (AI_Jun), to October, the year of harvest (AI_Oct2)
Yield_1SA	t ha^{-1}	The regional cane yield as tonnes of cane per hectare for the previous season. This was collected from the Tully mill
Yield_2SA	t ha^{-1}	The regional cane yield as tonnes of cane per hectare from two seasons ago. This was collected from the Tully mill
rain	mm	The cumulative rainfall from the planting date (15 May). Index values were calculated at the end of each month from May, the year before harvest (rain_May), to December, the year of harvest (rain_Dec2)
rain_lag	mm	The cumulative rainfall from 1 February, the year before harvest. This was used to capture possible climate effects before planting. Index values were calculated at the end of each month from February, the year before harvest (rain_lag_Feb), to December, the year of harvest (rain_lag_Dec2)
radn	MJ m^{-2}	The cumulative radiation from the plant date (15 May). Index values were calculated at the end of each month from May, the year before harvest (radn_May), to December, the year of harvest (radn_Dec2)
radn_lag	MJ m^{-2}	The cumulative radiation from 1 February, the year before harvest. This was used to capture possible climate effects before planting. Index values were calculated at the end of each month from February, the year before harvest (radn_lag_Feb), to December, the year of harvest (radn_lag_Dec2)
trange	°C	An index based on daily temperature range (t_{min} to t_{max}) calculated from the planting date (15 May). Index values were calculated at the end of each month from May, the year before harvest (trange_May), to December, the year of harvest (trange_Dec2)
trange_lag	°C	An index based on daily temperature range (t_{min} to t_{max}) from 1 February, the year before harvest. This was used to capture possible climate effects before planting. Index values were calculated at the end of each month from February, the year before harvest (trange_lag_Feb), to December, the year of harvest (trange_lag_Dec2)
SOI	-	Troups' monthly Southern Oscillation Index. Troups' SOI is derived from normalized Tahiti minus Darwin mean sea level pressure anomalies using the base period 1887–1989. SOI were obtained for each month from February, the year before harvest (SOI_Feb), to December, the year of harvest (SOI_Dec2). Data were obtained from the Long Paddock website (https://www.longpaddock.qld.gov.au/seasonalclimateoutlook/southernoscillationindex/soidatafiles/MonthlySOIPhase1887-1989Base.txt)
Nino	°C	The 3-month running average sea surface temperature anomalies in the Niño 3.4 region available from the NOAA website (http://www.cpc.ncep.noaa.gov/data/indices/3mth.nino34.81-10.ascii.txt). Data were available for February, the year before harvest (NINO_DJF), to December, the year of harvest (NINO_OND2)

website at https://www.longpaddock.qld.gov.au/silo/. Details about the calculation of these indices can be found in Table 1.

Predictor variables used in the models were calculated on a monthly basis. Predictor variables were calculated to cover each month from February, the year before harvest, to December, the year of harvest, to capture data within the growing season as well as possible effects of climatic indices before planting (Fig. 2).

Variable names followed the convention of being named for the last month of observed data which were used in its calculation. For example, the AI_Aug used observed data up to and including 30th August in the year before harvest. Variable names ending in "2" represented observed data in the harvest year. For example, AI_Jan2 used observed data up to and including the 31 January of the harvest year (Fig. 2).

2.4 Random forests

Random forests (Breiman 2001) are an ensemble learning algorithm that can be used for classification, that is predicting a categorical response variable and they can also be used for regression which involves predicting a continuous response variable. Random forest regression and classification models fit an ensemble of decision tree models to a set of data. For each tree, the data are recursively split into more homogenous units, which are commonly referred to as nodes, in order to improve the predictability of the response variable. Split points are based on values of predictor variables. Thus, variables used to split the data are considered important explanatory variables. Random forests fit separate decision trees to a predefined number of bootstrapped data sets. The predicted value of a categorical response is the mode of the classes from all the individual fitted decision trees, and the predicted value of a continuous response is the mean fitted response from all the individual trees that resulted from each bootstrapped sample.

Random forest classification and regression models were built using the "randomForest" package (Liaw and Wiener 2002) in the R free statistical software (R Core Team 2014). Random forest regression models were built using 500 trees

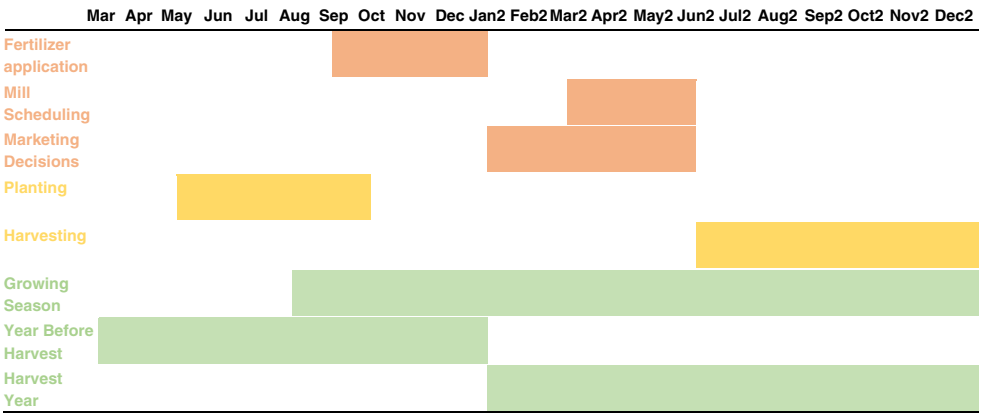

Fig. 2 Gantt chart of key phases of the sugarcane season in Tully. *Columns* represent months from March, the year before harvest, to December, the year of harvest. Months in the year of harvest are denoted with *2*. Key decision-making windows for fertilizer application, mill scheduling, and marketing are shown in relation to planting and harvesting windows, the growing season, harvest year, and year before harvest

derived from 500 bootstrapped data sets. Split points were chosen from a random subset of all available predictor variables (Breiman 2001). By default, the random subset size of the randomForest package is the square root of the number of predictors for classification models and one third of all available predictor variables for regression models (Liaw and Wiener 2002). Also by default, each node is restricted to a minimum size of one for classification or five for regression. Larger node sizes result in smaller trees to be grown reducing computational time (Liaw and Wiener 2002). In this study, the algorithm default values were used as preliminary testing showed little to no improvement if default values were modified. This had the added advantage of keeping inputs constant for all three forecast dates.

The random forest algorithm can rank the relative importance of each predictor variable. Variable importance is based on the regression prediction error of the out-of-bag, also called the OOB, portion of the data (Breiman 2001; Liaw and Wiener 2002). Approximately 30 % of data is OOB and is not used in building the tree (Abdel-Rahman et al. 2013). For classification models, the prediction error is calculated as the classification error rate, while for regression, the mean squared error is calculated. In the randomForest package, predictor variable importance is reported as mean percent decrease in classification rate for the classification model or mean increase in mean square error for the regression model if that variable was removed from the analysis.

2.5 Modelling process

For each forecast date, the randomForest algorithm was used to identify variable importance for all available predictor variables. Variable importance based on the OOB prediction error (Breiman 2001; Liaw and Wiener 2002) has been used in numerous studies such as Abdel-Rahman et al. (2013) and Everingham et al. (2015b). Although other measures of variable importance exist, and are becoming popular, like the conditional variable importance of Strobl et al. (2008), we opted to use the traditional and widely applied approach of Breiman (2001) and Liaw and Wiener (2002). Following previous studies such as Abdel-Rahman et al. (2013), a forward selection process was used to optimize the random forest models. That is, the models were rebuilt starting with the single most important predictor variable and additional variables were sequentially added that optimized the OOB classification error rate for classification models and the OOB R-squared for regression models. This offers the model the opportunity to improve performance while keeping the number of predictor variables low to minimize the risk of overfitting. The selected predictor variables used in the final models were recorded.

The OOB percentage of correctly classified years was used as an indicator to assess how well the random forest model could determine if yields were more likely to be above the median or below the median at the end of harvest. The correct classification rate can range from 0 % for the situation when every year is misclassified to 100 % when every year correctly classified. A higher value is preferred. The square root of the OOB mean square error (RMSE) and R-squared were determined to assess the performance of the random forest regression method. The RMSE gives a measure of the average error between model outputs and observations in appropriate units. A lower RMSE is preferred. The R-squared explains how much variation in the response is explained by the model. The R-squared value can range from 0, if no variation in the response is explained by the model, to 1, if 100 % of the variation in the response is explained by the model. A value closer to 1 is preferred. Correct classification rates, RMSE, and R-squared values for the optimized final model and the model using all available predictor variables were recorded.

3 Results and discussion

3.1 Classification random forest

Random forest models were quite successful at determining if the crop was likely to be above or below the median with an

OOB correct classification rate of at least 86.36 % (Table 2). Classification models did not include any APSIM biomass indices. While temperature in June was the most important predictor variable for the 1 September classification model, the 1 March model relied more heavily on spring and summer rainfall variables. The shift towards spring rainfall variables at later forecast dates aligns with previous research that uncovered the links between high austral spring rainfall and small yields in Tully the following year (Skocaj and Everingham 2014).

Nineteen out of the 22 years were correctly classified as above or below the median at the 1 September and 21 years out of 22 could be correctly classified by the 1 January. For the 1 September forecast, 1998, 2006, and 2013 were incorrectly classified. The 1998 harvest was only slightly above the median suggesting that source of error is most likely due to random chance. The random forest model predicted the 2006 crop to be above median, but a severe tropical cyclone adversely impacted the crop that year. In 2013, the model predicted a below median crop when in fact the crop was above the median. One reason for this misclassification could be due to industry harvesting a slightly larger percentage of younger and higher yielding sugarcane across the region.

For the 1 January and 1 March forecasts, only 1 year was misclassified—2007 was misclassified as above median. Ordinarily, low levels of precipitation would favor a bigger crop in Tully (Everingham et al. 2007b), but in November 2006, the crop was stressed by extremely low precipitation levels. In this year, Tully experienced the driest November and recorded only 5 mm compared to an average of 200 mm. This was followed by 1420 mm in February, almost double the long-term average of 743 mm which created more

Table 2 The predictor variables selected into each model listed in order of importance and model performance statistics of the random forest method used for classification and regression

Date	Classification		Regression		
	Correct classification rate (%) [all variables]	Selected predictor variables	R-squared [all variables]	RMSE (t ha^{-1}) [all variables]	Selected predictor variables.
1 September	86.36 [72.72]	trange_Jun Niño_MJJ	0.67 [0.41]	8.00 [10.66]	AI_Jul Niño_JJA SOI_Aug
1 January	95.45 [72.72]	SOI_Dec rain_Nov	0.72 [0.62]	7.32 [8.65]	SOI_Oct AI_Dec AI_Oct AI_Sep radn_lag_Dec SOI_Nov radn_Nov AI_Jul SOI_Dec rain_Nov trange_lag_Dec radn_Dec radn_lag_Nov rain_Jul AI_Nov rain_Dec rain_lag_Dec
1 March	95.45 [68.18]	SOI_Dec rain_Feb2 SOI_Oct Niño_JAS rain_Nov	0.79 [0.64]	6.33 [8.34]	SOI_Oct rain_Feb2 rain_lag_Feb2 AI_Jan2 AI_Sep AI_Dec trange_Jun

Date refers to the forecast date from 1 September, the year before harvest, to 1 March, the year of harvest. Values in square brackets represent performance statistics for the model using all available predictor variables

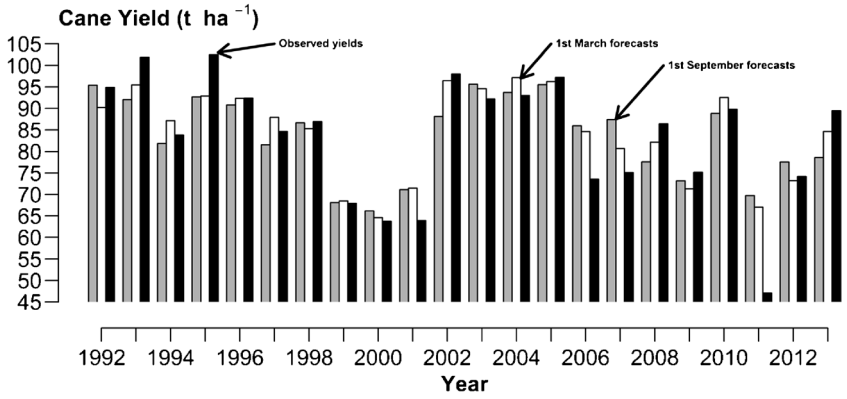

Fig. 3 Comparison of observed cane yields (*black*) and yields forecasted using random forest regression models. *Gray bars* represent yields forecasted on 1 September while *white bars* represent yields forecasted on 1 March. The later 1 March forecast yields were closer to observed yields than the earlier 1 September forecast

unfavorable growing conditions such as heavy cloud cover and conditions conducive to prolific early flowering which is difficult to represent in the APSIM crop model.

3.2 Regression random forest

The model produced on the 1 of September could explain 67 % of the total variability in yield responses. This increased to 72 % by the 1 January and 79 % by the 1 March. The forward selection process greatly reduced the number of predictor variables in each model and improved the forecast performance of the random forest prediction models (Table 2). The OOB RMSE reduced from 8.00 to 6.33 t ha^{-1} as the forecast date became later in the season.

The 1 September regression model relied most heavily on the APSIM biomass index in July and ENSO indices around August in the year before harvest. The Southern Oscillation Index (SOI) in October was important for the 1 January and 1 March forecasts. Other predictor variables that were selected included cumulative radiation, rainfall, and temperature. Everingham et al. (2007a) reported how positive SOI values favored above average rainfall and below average cane yields in the northern, wetter regions of Australia. The impact of rainfall and radiation on yields has been investigated by Muchow et al. (1997) who reported a general trend for an asymptotic increase in sugarcane yield potential with non-limiting water as the latitude declined from 30° S to 15° S, on account of increasing radiation and temperatures, as one approaches the equator. Yield potential in high rainfall regions such as Tully at 19° S was equivalent to yields of regions in the sub-tropics at approximately 25° S because of limited radiation. Radiation and rainfall are negatively correlated with course and would tend to be selected in the forecasting models for the same reason. However, high rainfall would tend to have additional negative effects associated with waterlogging, disruption to the harvest schedule, and additional compaction by harvesting machinery. Disruption of the planting and harvesting schedule can have long-term effects over several seasons as growers attempt to return to schedules that have proven successful in the past. Temperature range is also known to be a good predictor of sucrose content (Kingston 2002) because low night temperatures discourage stalk (cane) and leaf growth, in favor of sucrose accumulation (Inman-Bamber et al. 2010). We chose sugarcane yield as the target for our predictive model because of its importance in determining nitrogen requirements for the crop. Low night temperatures would affect cane yield adversely (Inman-Bamber et al. 2010).

Years that were poorly forecasted as of 1 March included the lowest yielding years such as 2011 and the highest yielding years such as 1993 and 1995 (Fig. 3). The 2011 growing season which runs approximately from August 2010 to December 2011 was greatly affected by a Tropical Cyclone which made landfall on February 3, 2011 and passed over the Tully region soon after. The high February rainfall as an important predictor variable (Table 2) would have further contributed to a low forecast yield; however, the model was not able to take account of the amount of damage to sugarcane due to wet weather harvesting in the previous La Niña year of 2010. The 1993 and 1995 seasons were characterized by low rainfall, and the simulated biomass indices were relatively high. Despite this, the model was challenged to predict the higher than normal yields observed. Clearly, more research is required to better understand how extreme yield events occur and how they can be more accurately predicted.

The results from this study agree with previous studies that show the random forest model can accurately estimate crop yields (Fukuda et al. 2013; Everingham et al. 2015b). The random forest models used in this study were able to identify important variables such as ENSO indices which could be explained from a biophysical perspective and agreed with earlier studies in the region (Skocaj and Everingham 2014; Everingham et al. 2015b). In this study, changes in variable importance were identified between different forecast dates while previous studies have shown that random forests can identify differences in important variables between regions (Everingham et al. 2015b). Future research should investigate if random forests can be used to provide yield estimates at a finer resolution rather than one estimate for the entire region akin to what has been accomplished for wheat (Newlands et al. 2014) and switchgrass (Tulbure et al. 2012). This would

then provide different productivity zones which could have management strategies to suit.

4 Conclusion

The purpose of this paper was to determine if a data mining approach could offer new insights that can explain sugarcane productivity in the Wet Tropics, Australia. Predicting the size of the crop can influence on farm decisions such as how much nitrogen fertilizer to apply and help millers carefully plan maintenance and labor schedules to be ready for the start of the milling season. Marketers can apply the same knowledge to maximize industry profits through more effective and targeted forward selling strategies and logistical planning. Collectively, these improvements in economic and environmental outcomes are important for delivering sustainable solutions to industry.

The random forest models were quite successful at predicting sugarcane yields very early in the season. In September in the year before harvest, when many farmers plan their nutrient management, the random forest classification model could determine if the crop was likely to be above 86.7 t ha^{-1} or less than 86.7 t ha^{-1} with a reasonable level of accuracy. Farmers expecting a smaller crop could consider applying less nitrogen fertilizer than in years when the model predicts a larger crop. This is because cane yields are lower in years experiencing high rainfall during the austral spring and summer. Reducing nitrogen fertilizer rates in these years will improve nitrogen fertilizer use efficiency and deliver positive economic and environmental outcomes. Using this strategy, there would have been 1 year when a farmer would have applied too much fertilizer and 2 years when the farmer applied not enough fertilizer. The regression and classification cane yield estimates can be supplemented with other expert industry knowledge to strengthen forward selling plans and improve industry confidence in selling sugar on the futures market. These estimates could also be used to better plan harvesting and milling operations. When a large crop is forecast, the local sugar industry could consider starting harvesting and milling operations earlier than normal to reduce the risk of cane being left unharvested until the following year, whereas when a small crop is forecast, it might be better to delay the start of the crushing season to maximize sugar production and profitability. Knowledge about crop size produced from the random forest model predictions could also be useful in managing mill maintenance operations to ensure the mill will be ready for the crushing season.

It must be stressed that the data mining approach put forth in this paper is not perfect, and the cost of an inaccurate estimate must be compared to the benefits of an accurate forecast. Other constraints also exist in that climate change does not disrupt the model. Thus, this research should be regularly reviewed to ensure that the approach is still relevant under a changing climate. Notwithstanding these investigations, the key findings of this paper support the use of data mining and Big Data technologies to increase industry guidance on key industry decisions that affect sustainable agricultural systems and contribute a partial solution to food shortages. Most importantly, the approach outlined in this paper can easily be extended to other sugarcane-growing regions and agricultural industries throughout the world to better inform agricultural practices.

Acknowledgments The authors would like to thank Tully Sugar Limited for providing access to the productivity data. This research was funded by Sugar Research Australia.

References

Abdel-Rahman EM, Ahmed FB, Ismail R (2013) Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. Int J Remote Sens. doi:10.1080/01431161.2012.713142

Allwinkle S, Cruickshank P (2011) Creating smarter cities: an overview. J Urban Technol. doi:10.1080/10630732.2011.601103

Breiman L (2001) Random forests. Mach Learn. doi:10.1023/A:1010933404324

Cao S, Manrai AK (2014) Big data in marketing & retailing. J Int Interdiscip Bus Res 1:23–42, http://www-bcf.usc.edu/~jkarayan/base%20rates.JIIBR-Volume-1.pdf#page=27

Caragliu A, Del Bo C, Nijkamp P (2011) Smart cities in Europe. J Urban Technol. doi:10.1080/10630732.2011.601117

Chen J, Li M, Wang W (2012) Statistical uncertainty estimation using Random Forests and its application to drought forecast. Math Probl Eng. doi:10.1155/2012/915053

Core Team R (2014) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, http://www.R-project.org/

Craig E, Huettmann F (2009) Using "blackbox" algorithms such as TreeNET and Random Forests for data-mining and for finding meaningful patterns, relationships and outliers in complex ecological data: An overview and example using G. In: Wang H (ed) Intelligent data analysis: developing new methodologies through pattern discovery and recovery. Information Science Reference, Hershey, pp 65–84. doi:10.4018/978-1-59904-982-3.ch004

De'ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology. doi:10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2

Everingham YL, Muchow RC, Stone RC, Coomans DH (2003) Using southern oscillation index phases to forecast sugarcane yields: a case study for northeastern Australia. Int J Climatol. doi:10.1002/joc.920

Everingham YL, Inman-Bamber NG, Thorburn PJ, McNeill TJ (2007a) A Bayesian modelling approach for long lead sugarcane yield forecasts for the Australian sugar industry. Aust J Agric Res. doi:10.1071/AR05443

Everingham YL, Lowe KH, Donald DA, Coomans DH, Markley J (2007b) Advanced satellite imagery to classify sugarcane crop characteristics. Agron Sustain Dev. doi:10.1051/agro:2006034

Everingham YL, Smyth CW, Inman-Bamber NG (2009) Ensemble data mining approaches to forecast regional sugarcane crop production. Agric For Meteorol. doi:10.1016/j.agrformet.2008.10.018

Everingham YL, Inman-Bamber NG, Sexton J, Stokes C (2015a) A dual ensemble agroclimate modelling procedure to assess climate change impacts on sugarcane production in Australia. Agric Sci. doi:10.4236/as.2015.68084

Everingham YL, Sexton J, Robson A (2015b) A statistical approach for identifying important climatic influences on sugarcane yields. In: Proc Aust Soc Sugar Cane Technol. Bundaberg, Australia, pp 8–15

FAO (2009) How to feed the world in 2050., http://www.fao.org/fileadmin/templates/wsfs/docs/expert_paper/How_to_Feed_the_World_in_2050.pdf

Fukuda S, Spreer W, Yasunaga E, Yuge K, Sardsud V, Müller J (2013) Random Forests modelling for the estimation of mango (Mangifera indica L. cv. Chok Anan) fruit yields under different irrigation regimes. Agric Water Manag. doi:10.1016/j.agwat.2012.07.003

García-Gutiérrez J, Martínez-Álvarez F, Troncoso A, Riquelme JC (2015) A comparison of machine learning regression techniques for LiDAR-derived estimation of forest variables. Neurocomputing. doi:10.1016/j.neucom.2014.09.091

Gromski PS, Xu Y, Correa E, Ellis DI, Turner ML, Goodacre R (2014) A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data. Anal Chim Acta. doi:10.1016/j.aca.2014.03.039

Guo P-T, Li M-F, Luo W, Tang Q-F, Liu Z-W, Lin Z-M (2015) Digital mapping of soil organic matter for rubber plantation at regional scale: an application of random forest plus residuals kriging approach. Geoderma. doi:10.1016/j.geoderma.2014.08.009

Inman-Bamber NG, Bonnet GD, Spillman MF, Hewitt ML, Glassop D (2010) Sucrose accumulation in sugarcane is influenced by temperature and genotype through the carbon source-sink balance. Crop Pasture Sci. doi:10.1071/CP09262

Keating BA, Robertson MJ, Muchow RC, Huth NI (1999) Modelling sugarcane production systems I. Development and performance of the sugarcane module. Field Crop Res. doi:10.1016/S0378-4290(98)00167-1

Kingston GR (2002) Recognising the impact of climate on CCD of sugarcane across tropical and sub-tropical regions of the Australian sugar industry. In Proc Aust Soc Sugar Cane Technol, pp. 145–152

Liaw A, Wiener M (2002) Classification and regression by Random Forest., RNews 2/3:18–22. ftp://131.252.97.79/Transfer/Treg/WFRE_Articles/Liaw_02_Classification%20and%20regression%20by%20randomForest.pdf

Muchow RC, Robertson MJ, Keating BA (1997) Limits to the Australian sugar industry: climatic and biological factors. In: Keating B, Wilson J (eds) Intensive Sugarcane Production: Meeting the Challenges Beyond 2000. CAB International, Wallingford, pp 37–54

Newlands NK, Zamar DS, Kouadio LA, Zhang Y, Chipanshi A, Potgieter A, Toure S, Hill HSJ (2014) An integrated, probabilistic model for improved seasonal forecasting of agricultural crop yield under environmental uncertainty. Front Environ Sci. doi:10.3389/fenvs.2014.00017

Nicholls N, Drosdowsky W, Lavery B (1997) Australian rainfall variability and change. Weather. doi:10.1002/j.1477-8696.1997.tb06274.x

Paul J (2012) Big data takes centre ice. Marketing, vol 117., http://www.marketingmag.ca/brands/big-data-takes-centre-ice-66917/2

Perera C, Zaslavsky A, Christen P, Georgakopoulos D (2014) Sensing as a service model for smart cities supported by Internet of Things. Trans Emerg Telecommun Technol. doi:10.1002/ett.2704

Philibert A, Loyce C, Makowski D (2013) Prediction of N2O emission from local information with Random Forest. Environ Pollut. doi:10.1016/j.envpol.2013.02.019

Saussure S, Plantegenest M, Thibord J-B, Larroudé P, Poggi S (2015) Management of wireworm damage in maize fields using new, landscape-scale strategies. Agron Sustain Dev. doi:10.1007/s13593-014-0279-5

Skocaj DM, Everingham YL (2014) Identifying climate variables having the greatest influence on sugarcane yields in the Tully mill area. In: Proc Aust Soc Sugar Cane Technol. Goldcoast, Australia, pp 53–61

Skocaj D, Everingham Y, Schroeder B (2013) Nitrogen management guidelines for sugarcane production in Australia: can these be modified for Wet Tropical conditions using seasonal climate forecasting? Springer Sci Rev. doi:10.1007/s40362-013-0004-9

Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for random forests. BMC Bioinform. doi:10.1186/1471-2105-9-307

Tulbure MG, Wimberly MC, Boe A, Owens VN (2012) Climatic and genetic controls of yields of switchgrass, a model bioenergy species. Agric Ecosyst Environ. doi:10.1016/j.agee.2011.10.017

Zar JH (1999) Biostatistical analysis. Prentice Hall, Upper Saddle River