Combining data-driven models and mechanistic carbon assimilation models to predict sugarcane yield for improved management

Si Yang Han, Filippi, P., Bishop, TFA.

Precision Agriculture Laboratory, Sydney Institute of Agriculture, School of Life and Environmental Sciences, The University of Sydney, Sydney, New South Wales, Email: siyang.han@sydney.edu.au

Abstract
Accessing sugarcane harvest data over large areas is difficult, although more data is collected today than ever before. Sugarcane yield data is a critical variable for research involving the Great Barrier Reef. From an agricultural perspective, knowledge of yield variability facilitates the identification of yield potential and yield gaps across fields, farms, and regions. This study developed a mechanistic carbon-assimilation model for sugarcane (CAMS), and then used it as a covariate in a data-driven yield forecast model at the block (field) resolution. The CAMS model included satellite imagery, elevation, latitude, and daily weather records. A 380 ha sugarcane property in the Isis district, Queensland, was used as a case study. The scope of this yield prediction approach could be extended to the entire region, or throughout the industry, if the location of sugarcane blocks and harvest dates are known.

Keywords
Remote sensing; site-specific; machine learning; CAMS; scalable

Introduction

Sugarcane (Saccharum spp.) is a major crop in Queensland, Australia, with an estimated economic contribution of $4 billion AUD annually (ASMC 2021). In recent years, the sugarcane industry has become heavily regulated for on-farm inputs, due to nutrient-rich farm runoff flowing into the Great Barrier Reef (GBR). Given the increasing privacy concerns for growers and farm data, accessing large-scale sugarcane production and harvest data is challenging, even though more data is collected today than ever before.

Availability of sugarcane yield data is a prerequisite in estimating critical environmental processes such as water pollution and carbon sequestration, as well as optimising on-farm productivity through the concept of yield potential and gaps. A scalable sugarcane yield model calibrated to a small dataset could be expanded across regions, and even the entire industry.

There are two main approaches to predicting crop yield. Mechanistic simulation models such as the Agricultural Production Systems Simulator (APSIM) (Holzworth et al. 2014) use a suite of inputs, such as plant, soil, climate, and management data to simulate the crop growth with biological theory. On the other hand, regression and machine learning models optimise the fit using remotely sensed imagery and relevant covariates to yield data, allowing statistics to uncover the biological relationships.

The aim of this study was to incorporate a mechanistic crop model in a machine learning model to obtain the benefits of each. Rather than using all the inputs of a mechanistic model in a machine learning model, we explored the use of the output of the mechanistic model as a covariate, with the intent for simpler, more interpretable, and accurate models. Donohue et al. (2018) defined a scalable mechanistic crop model for wheat and canola based off Monteith's radiation use efficiency concept (Lobell 2013). The model was scalable, as inputs were either publicly available national datasets, or were otherwise estimated (i.e. sowing and harvest dates). The adapted carbon assimilation model for sugarcane (CAMS) was used as a covariate with other spatial and temporal variables in a random forest regression model to predict yield.

Methods

Study Area

A 380 ha irrigated sugarcane farm in the Isis District (QLD, Australia) was used as a case study. The farm consisted of four separate properties, with harvest data available from 12 growing seasons, 2007 to 2018.

Data Collation

Five main types of predictor variables were collated: harvest/management data, terrain attributes, radiometrics, weather and remote sensing. A set of 40 initial covariates (Table 1) was culled by removing redundant variables, as in Bishop et al. (2015). Pairs of covariates within each group which had Pearson's correlations greater than 0.80 were identified, and the covariate with the weaker correlation with yield was removed.

Table 1. A summary of the predictor variables considered for the random forest models.

Covariate Type	Resolution	Description	Source
Management Data	Block	7 covariates – e.g. harvest date, ratoon number etc.	Isis County Sugar
Terrain Attributes	30 m	12 covariates derived from a digital elevation model or similar, e.g. silica index (Gray et al. 2016).	Mill CSIRO Data Portal
Radiometrics	105 m	8 covariates of K, Th and U radiation, as well as their ratios.	
Weather	5 km	10 covariates of weather attributes cumulated from Spring-Autumn.	Geoscience Australia Google Earth Engine
Remote Sensing	30 m	1 covariate (NDVI) from Landsat 7 Imagery.	SILO, BOM

Carbon Assimilation Covariate

CAMS was developed with an adapted methodology from Donohue et al. (2018) for 'C-CROP', although with Landsat 7 imagery, as opposed to MODIS 16-day composites. Other model inputs included daily minimum and maximum temperature measurements, elevation, and latitude. In addition, the model was refined with block specific growing dates (which for sugarcane, is essentially the previous and current harvest dates) as well as varying carbon allocation for root to shoot ratios depending on the ratoon, i.e. crop age. Crop specific constants were adapted for sugarcane, using values observed in literature or fitted empirically. The maximum CAMS value, which occurred towards the end of each season, was used as covariate in the subsequent yield model.

Yield Modelling

A random forest model from the R package 'ranger' (Wright and Ziegler 2015) was used to predict yield at a block resolution using the remaining covariates. A fine grid (10 m) was used to extract the covariates, which was then aggregated to each block, the resolution of the harvest data. The relative permutation importance of each covariate to the model was ranked by the random forest algorithm.

Two models were compared: a model including the output from the mechanistic model CAMS, where the covariates utilised or related to the development of CAMS were excluded (e.g. temperature), and a non-mechanistic model, which contained all variables other than CAMS. The biomass estimator to substitute for CAMS predictions in the non-mechanistic model was the maximum NDVI and GNDVI for each block, in each season.

The model quality was assessed with leave-one-block-out cross validation. A single field from a single year was left out of the dataset, to be predicted by all other available data. The predicted yield from all 12 seasons was accumulated and assessed against the observed yield data using Lin's concordance correlation coefficient (LCCC) and root mean square error (RMSE).

Results and Discussion

In terms of Pearson's correlation with yield, maximum CAMS had the highest correlation of 0.76, whereas maximum NDVI was notably poorer, at 0.31 (Figure 1). As CAMS is a cumulative mechanistic model, it incorporates some on-farm management such as the ratoon, and the 'start' of each season, or the previous

harvest, which ranges between June to December. On the other hand, the maximum NDVI of each season is a single measurement and cannot reliably identify fallows due to weeds or break crops.

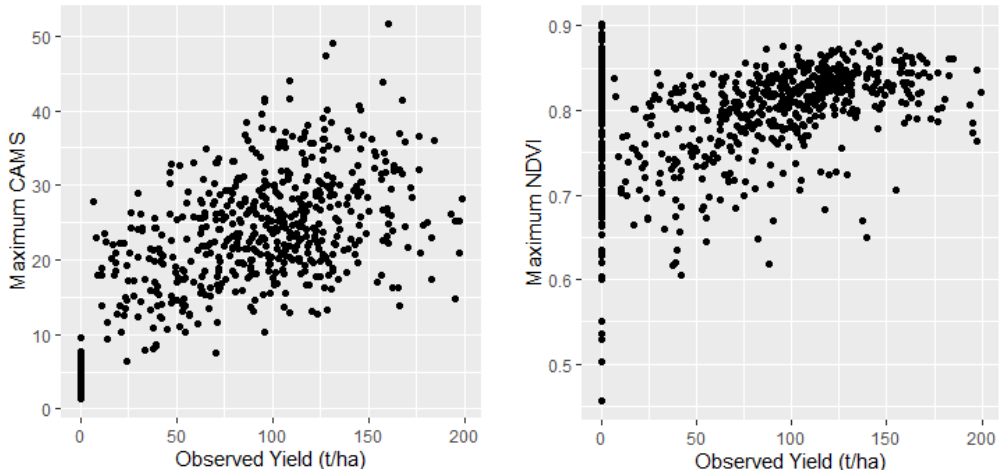

Figure 1. Scatter plot of maximum NDVI for each block in each season against yield (right) and the maximum carbon assimilation value against yield (left).

The yield predictions made by the random forest model with CAMS had a higher LCCC and lower RMSE, though the models were comparable (Figure 2). The random forest model was able to largely compensate for the confusion in maximum NDVI for the non-mechanistic model by incorporating the aspect of fallow through the categorical variables Ratoon and Variety.

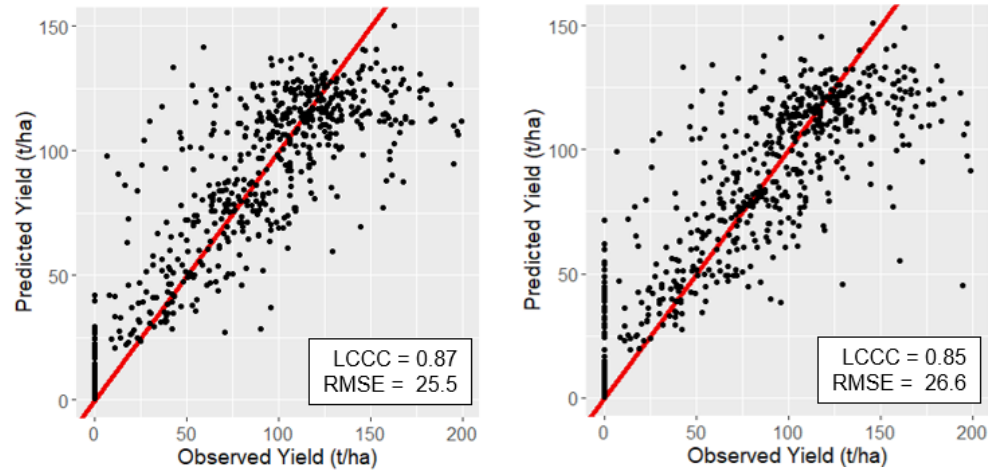

Figure 2. The fit of the observed and predicted sugarcane yield (Lin's concordance correlation coefficient and root mean square error), for all 12 seasons in the CAMS random forest model.

The ranked variable importance for each random forest model demonstrates that the inputs of CAMS (Harvest Month, Ratoon, Latitude, Season Length, Max_NDVI) were key variables driving sugarcane yield predictions (Figure 3). The relatively high importance of spatial covariates (i.e. radiometrics, silica index) suggests the variation in yield across the farm is largely driven by variability in soil.

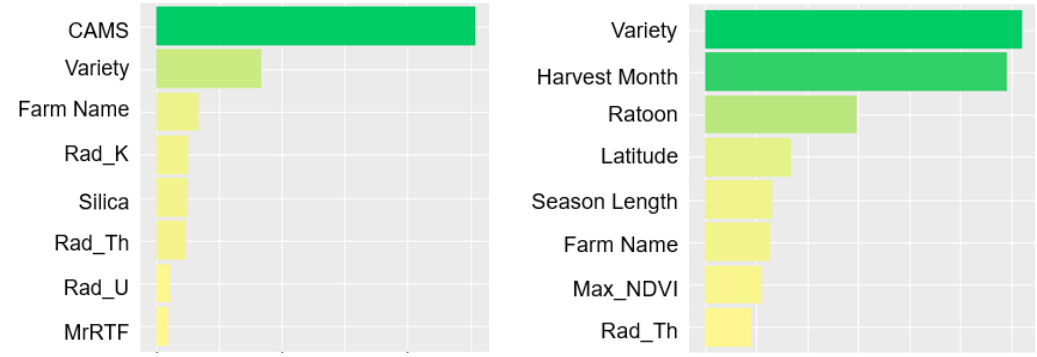

Figure 3. Variable importance plots for the CAMS model (left) and the separate inputs model (right). The 8 most important variables are included.

The current industry standard for predicting yield involves fitting a single image of Green NDVI (GNDVI) taken during the peak growth stage of each season (March-April) against a quadratic of the historical growth pattern for a particular harvest month (Rahman and Robson, 2020). Maximum NDVI was used as there was a higher correlation with yield, compared to maximum GNDVI (r = 0.21), and a time series of NDVI was used in the development of the CAMS covariate. In the context of sugarcane, it is often difficult to obtain a clear satellite image during the wet season (December-April), when peak growth occurs. As CAMS is a cumulative carbon assimilation model, it is more accommodating for prolonged breaks in imagery. While maximum NDVI is a single measurement, CAMS is the output from a mechanistic model, which has room for optimisation and improvement. Even in a small case study, it performs better than the maximum NDVI of each season, implying that there is value in using mechanistic models to combine potential predictor variables for simpler and improve data-driven models.

Conclusion

This study investigated the use of large-scale mechanistic carbon assimilation model as an input in a machine learning model to predict sugarcane yield. CAMS had a strong relationship with yield (r = 0.76) and the validated yield prediction was also strong (LCCC 0.87, RMSE = 25.5 t ha^{-1}). The model with the mechanistic covariate did perform better than the model with separate components, even in a small case study. The scalable nature of CAMS paves the path to a regional or national sugarcane yield map, which would be invaluable to a number of stakeholders as it could be used to identify yield gaps and their causes.

References

Australian Sugar Milling Council (ASMC), 2021, Sugar Industry Summary Statistics, https://asmc.com.au/policy-advocacy/sugar-industry-overview/economic-contribution-sugar/.

Bishop, T.F.A., Horta, A. and Karunaratne, S.B., 2015. Validation of digital soil maps at different spatial supports. Geoderma, 241, pp.238-249.

Donohue, R.J., Lawes, R.A., Mata, G., Gobbett, D. and Ouzman, J., 2018. Towards a national, remote-sensing-based model for predicting field-scale crop yield. Field Crops Research, 227, pp.79-90.

Gray, J.M., Bishop, T.F. and Wilford, J.R., 2016. Lithology and soil relationships for soil modelling and mapping. Catena, 147, pp.429-440.

Holzworth, D.P., Huth, N.I., deVoil, P.G., Zurcher, E.J., Herrmann, N.I., McLean, G., Chenu, K., van Oosterom, E.J., Snow, V., Murphy, C. and Moore, A.D., 2014. APSIM–evolution towards a new generation of agricultural systems simulation. Environmental Modelling & Software, 62, pp.327-350.

Lobell, D.B., 2013. The use of satellite data for crop yield gap analysis. Field Crops Research, 143, pp.56-64.

Rahman, M.M. and Robson, A., 2020. Integrating Landsat-8 and Sentinel-2 time series data for yield prediction of sugarcane crops at the block level. Remote Sensing, 12(8), p.1313.

Wright, M.N. and Ziegler, A., 2015. ranger: A fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint arXiv:1508.04409.