A comparison between machine learning and simple mechanistic-type models for yield prediction in site-specific crop yield predictions

Dhahi Al-Shammari[1]**,** Thomas F.A. Bishop[1] , Chen Wang[2], Brett M. Whelan[1] and Robert G.V Bramley[3]

[1] Sydney Institute of Agriculture, School of Life & Environmental Science, The University of Sydney, Central Ave, Eveleigh, Sydney, NSW, 2015 Email: dhahi.al-shammari@sydney.edu.au

[2] CSIRO Data61, Sydney,

[3] CSIRO, Waite Campus, Locked Bag 2, Glen Osmond, SA 5064,

Abstract

Crop yield prediction is a challenging topic that has been studied worldwide. Many approaches have been proposed to address this challenge. Generally, these approaches follow two paths: physical models (PMs) calibrated with statistical analysis and data-driven machine learning models (MLs). There has been little study on the difference between the two approaches in crop yield prediction. A comparison between a simple empirical model with a physical basis and ML is presented in this paper. Data for wheat was used in this study. The PM has been developed using satellite-based vegetation indices and climate data. The PM has been compared to eXtreme Gradient Boosting (XGBOOST) to examine the potential of ML in learning from the data without the need for domain knowledge of the process. The results suggest that ML models can be comparable to PM models in their prediction performance. XGBOOST does not require domain knowledge to generate robust predictions, yet it requires data representativeness to develop reliable outcomes.

Keywords

Mechanistic models, XGBOOST, crop yield prediction, three-red edge vegetation index.

Introduction

The relationship between crop yield and climate is crucial to understand crop production. Many studies explored the linkage between climatic variables and crop yield using either mechanistic (PM) or machine learning (ML) models. The difference between PM and ML is that PM are built using empirical equations based on the knowledge of the process, whereas ML models are data-driven and built with a minimum knowledge of the process (Roberts et al., 2017). One of the leading PM is The Agricultural Production Systems sIMulator (APSIM) (Keating et al., 2003). This model has been used in several studies to understand the impact of climate on crop yield (Luo et al., 2005; Deihimfard et al., 2018; Ahmed et al., 2016). Other studies have focused on ML models (Fajardo and Whelan, 2021; Filippi et al., 2019). Simpler models can be more effective than the complex model if the causal relationships between the most influential factors are captured well; for example, French and Schultz (1984; F&S) developed a simple model to predict wheat yield by determining a relationship between water use and wheat yield. The F&S model is very simple and widely used for estimating potential crop yield. The differences in the types and complexities of models raise a question of which model to use to achieve reliable and superior results for yield prediction? To answer this question, a comparison between a simple PM and ML has been performed to explore the predictive power of these two types using climatic and vegetation indices data.

Methods

Study area

The study area includes 49 New South Wales (NSW) paddocks near Albury with crop yield data available between 2016 and 2019; 2016 was a year of record high rainfall, whereas 2018 had the lowest rainfall on record. The in-season rainfall reached a mean across the paddocks of 398 mm (2016), 205 mm (2017), 121 mm (2018) and 201 mm (2019). This fluctuation in the amount of rainfall is important and can be used to quantify the wheat yield.

Data

Pixel-based interpolated wheat yield maps were obtained from an AgTech company. The wheat maps were provided at a 10-metre spatial resolution. Daily rainfall and solar radiation rasters were sourced from the ‘SILO’ Australian climate database at ~5 kilometre resolution (Jeffrey et al., 2001). Daily evapotranspiration (ET) rasters were sourced from The Moderate Resolution Imaging Spectroradiometer (MODIS 16) at ~500 metre resolution (Mu et al., 2011). Accumulated rainfall, solar radiation and ET were calculated from the 1st of May to 15th of September for each season. Cloud-free Sentinel-2 data were collected and used to calculate a three-red-edge vegetation index (TREI) (Eq.1) at 20 metre resolution, the resolution of the red-edge bands in the Sentinel-2. A time window was set from the 15th to the 30th of September for each season to collect the high-quality images (cloud-free images). This time window was selected as one during which having a yield prediction would help farmers for decisions related to marketing and for managing supply chain logistics. All data were resampled to the resolution of the yield maps (10 metre) and used for further analysis.

Models

A simple empirical model with a physical basis (PM)

For this study, a simple empirical model with a physical-basis was used which uses rainfall and vegetation index for wheat yield prediction. The PM follows similar concept to the F&S, however, the PM accounts for the within-field variability by exploiting the information from the vegetation index so that pixels with high values are predicted differently than the pixels with low values. Furthermore, by capturing the changes of within-season rainfall, solar radiation, ET and using the vegetation indices, yield can be calculated. The model was built using a mathematical equation that accounts for the energy conversion (EC) (Eq.2) to biomass and the correlation between the result of the EC and the TREI to calculate the potential yield in t ha^{-1} (Eq.3). A parameter (*A*) was used to scale the relationship between EC, TREI and yield depending on seasonal rainfall. In this work this means that we had a different parameter value for rainfall values. The equations are:

$$\textbf{TREI} = (\textbf{RE2} - \textbf{RE1}) * (\textbf{RE3}/\textbf{RE1}); \qquad (1)$$

$$\textbf{Energy conversion (EC)} = \textbf{Solar irradiance} + \textbf{Rainfall} - \textbf{Evapotranspiration}; \qquad (2)$$

$$\textbf{Yield t ha}^{-1} = \textbf{TREI} * \textbf{EC} * \boldsymbol{A}; \qquad (3)$$

where RE1, RE2 and RE3 in Eq.1 are the red-edge bands 1 (705nm), 2 (740nm) and 3 (783nm). The TREI is a three-red-edge bands vegetation index which captures the changes in the red-edge region. The red-edge region based indices have been successfully used for biomass and grain yield prediction (Kanke et al., 2016).

For each season, Eq.1, Eq.2 and Eq.3 were used to calculate the potential yield on a 10m pixel basis. Then the mean of potential and actual yield per-field was calculated for the assessment of prediction quality. The parameter *A* changes with the change of the amount of rainfall and two rainfall intervals (<250 mm, and >251 mm) were determined for the parameter tuning. Therefore, model calibration included tuning the parameter until lowest RMSE was achieved. Years 2017, 2018 and 2019 fell into one interval and these years were used to validate against each other. For 2016, the calibration was performed, but the validation from different years was not available. Therefore, the results from this year are calibrated results only.

Data-driven model (XGBOOST)

We used the eXtreme Gradient Boost (XGBOOST) to develop a data-driven model in this study. XGBOOST is a decision-tree-based ensemble method that applies boosting on the weak learners (regression trees) where the weak learners of the XGBOOST learn sequentially from the residual of the previous weak learner (Chen and Guestrin, 2016). XGBOOST takes into account the trade-off between bias and variance (Nielsen, 2016). XGBOOST was used for the comparison with the simple PM. A leave-one-year-out-cross-validation was used for training and validation where three years were used for training and one year for validation as for the PM. As with the PM, the mean of the predicted and actual yield for each paddock was calculated.

Results and discussion

The results from the PM and ML are shown in Fig.1 and 2. It is clear that the PM model predicted yield better than XGBOOST in 2016 (2016 was calibration results in the PM) and 2017, but

XGBOOST was better than the PM in 2018 and 2019 (Fig.2). In 2016, the predicted yield for XGBOOST was higher than the observed and this was because of the amount of rainfall which the model used as predictor was the highest (mean rainfall 398 mm) compared to the other years. In 2017, XGBOOST did not predict yield well where the concordance correlation coefficient (CCC) = 0.14, and this was because the observed yield (Fig.1) was very high compared to the amount of rainfall (mean = 205 mm) in that year. Therefore, we conclude that XGBOOST requires more data from different years and inclusion of more factors that might help XGBOOST to explain this contradiction (yield vs rainfall). In 2018 and 2019, XGBOOST had higher predictive quality (CCC = 0.83 for 2018 and 0.81 for 2019) and this was for two reasons. First, in 2018, the observed yield was the lowest (< 4 t ha^{-1}) compared to the other years and this could be explained by the rainfall amount (mean = 121 mm). In 2019, the observed yield was also explained by the rainfall amount (mean = 201 mm). This indicates that XGBOOST could identify the relationship between the yield and rainfall. On the other hand, PM did not show any fluctuation in the results. However, it was based on a simple regression equation which was defined based on the knowledge of the impact of the rainfall on crop yield. The PM required the parameter to be tuned based on the rainfall intervals until an optimal result was achieved; the values used were 0.0035 and 0.0040 for the rainfall intervals <250 mm and ≥250mm.

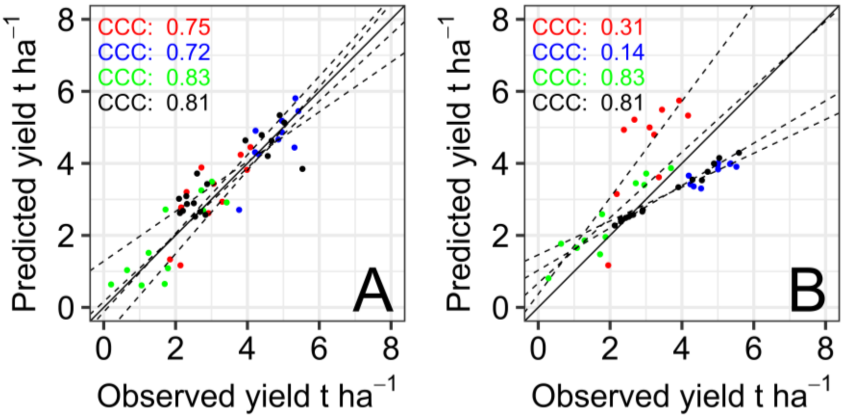

Figure 1. Actual versus predicted wheat yield obtained from (A) the PM and (B) the XGBOOST. Red, blue, green, and black points represent 2016, 2017, 2018 and 2019 respectively. CCC is the Lin's Concordance Correlation Coefficient.

It also can be seen from Fig.2 that XGBOOST yielded higher MAE (>1.5 t ha^{-1}), MAPE (~0.5 t ha^{-1}), MSE (~3 t ha^{-1}) and RMSE (>1.5 t ha^{-1}) in 2016, MAE (>1 t ha^{-1}), MAPE (~0.2 t ha^{-1}), and RMSE (>1.25 t ha^{-1}) and MSE (>1.5 t ha^{-1}) in 2017. This also can be explained by the fluctuation of the rainfall vs yield relationship in those years. In contrast, XGBOOST gave almost similar MAE, MAPE, MSE and MSE to the PM in 2018, and lower in 2019. This indicates the ability of the ML in finding patterns similar or even better than the PM when it was supplied with representative data.

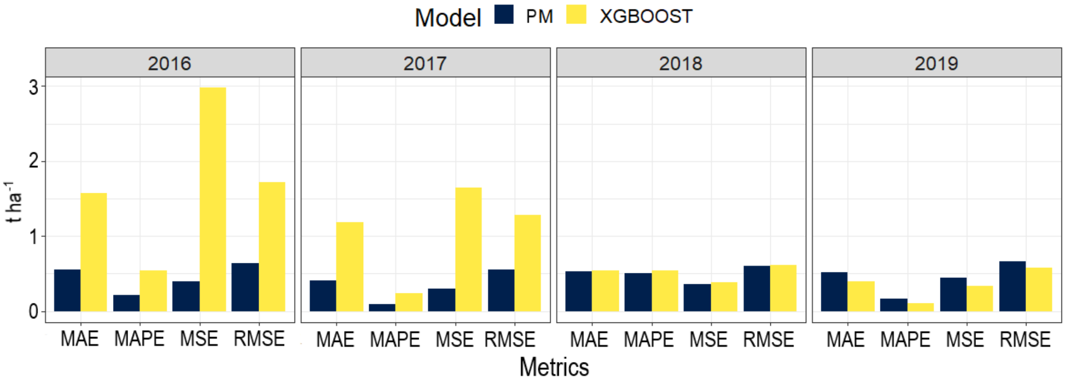

Figure 2. Bar-plot of the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE) and Root Mean Square Error (RMSE).

Finally, the strength of the PM is that it is less complex and easier to interpret compared to the ML model. However, it is known that PM may not adequately account for all processes in crop production, and in particular, the impact of management practices. The PM also requires careful tuning of the parameter to achieve lowest MAE, MAPE, MSE and RMSE. Therefore, the PM requires expert knowledge to achieve reasonable results. Unlike the PM, the XGBOOST requires minimum knowledge about these various processes when fitting the model. It relies on the observations to find casual relationships between yield and the supplied factors.

Conclusion

In this study, a PM and ML method (XGBOOST) were compared for wheat yield prediction. The PM provided more stable predictions than XGBOOST. The advantage of the PM is that it is simple and easy to interpret. However, this model required expert knowledge for parameter fitting to achieve optimal results. In contrast, XGBOOST only requires representative data to achieve good results, and it is very flexible in respect of the addition of more data. Future studies are required to include more observations from different years to cover the interaction between yield and rainfall to achieve better predictions. Future studies can also determine specific constants for the PM based on specific rainfall intervals.

References

Ahmed M, Akram MN, Asim M, et al. (2016) Calibration and validation of APSIM-Wheat and CERES-
Wheat for spring wheat under rainfed conditions: Models evaluation and application. Computers and electronics in agriculture 123: 384-401.

Chen T and Guestrin C. (2016) Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 785-794. Deihimfard R, Eyni-Nargeseh H and Mokhtassi-Bidgoli A. (2018) Effect of future climate change on wheat yield and water use efficiency under semi-arid conditions as predicted by APSIM-wheat model. International Journal of Plant Production 12: 115-125.

Fajardo M and Whelan B. (2021) Within-farm wheat yield forecasting incorporating off-farm information. Precision Agriculture: 1-17.

Filippi P, Jones EJ, Wimalathunge NS, et al. (2019) An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. Precision Agriculture 20: 1015-1029.

Jeffrey SJ, Carter JO, Moodie KB, et al. (2001) Using spatial interpolation to construct a comprehensive archive of Australian climate data. Environmental Modelling & Software 16: 309-330.

Kanke Y, Tubana B, Dalen M, et al. (2016) Evaluation of red and red-edge reflectance-based vegetation indices for rice biomass and grain yield prediction models in paddy fields. Precision Agriculture 17: 507-530.

Keating BA, Carberry PS, Hammer GL, et al. (2003) An overview of APSIM, a model designed for farming systems simulation. European journal of agronomy 18: 267-288.

Luo Q, Bellotti W, Williams M, et al. (2005) Potential impact of climate change on wheat yield in South Australia. Agricultural and Forest Meteorology 132: 273-285.

Mu Q, Zhao M and Running SW. (2011) Improvements to a MODIS global terrestrial evapotranspiration algorithm. Remote sensing of environment 115: 1781-1800. Nielsen D. (2016) Tree boosting with xgboost-why does xgboost win" every" machine learning