A big data approach to predicting crop yield

Patrick Filippi*, Edward Jones, Thomas Bishop, Niranjan Acharige, Sanjeewani Dewage, Liana Johnson, Sabastine Ugbaje, Thomas Jephcott, Stacey Paterson, Brett Whelan
Sydney Institute of Agriculture, The University of Sydney

Abstract

Many broadacre farmers have a time series of crop yield monitor data for their paddocks which are often augmented with additional spatial data such as gamma radiometrics surveys or ECa (apparent soil electrical conductivity) from an electro-magnetic induction survey (EMI). In addition there are now readily available national and global datasets which can be used to represent the crop-growing environment. Rather than analysing one paddock at a time, there is an opportunity to explore the value of combining data over multiple paddocks and years into one dataset. Using these datasets in conjunction with machine learning approaches allows predictive models of crop yield to be built. In this study we explored this approach with a particular emphasis on the forecasting ability of the models based on pre- and mid-season information from predictor variables. Several large farms in Western Australia were used as a case study. Yield from wheat, barley and canola crops from 3 different seasons that covered ~15,000 hectares in each year were used. The yield data was processed to a 10 m grid, and for each observation we built an associated space-time cube of predictors. This consisted of grower collected data such as EM and gamma radiometrics surveys, and nationally available data such as MODIS NDVI, and rainfall. Random Forest models were used to predict crop yield of wheat, barley and canola using the space-time cube. Three models were created based on pre-sowing, mid-season and late-season conditions to explore the changes in the predictive ability of the model as more within-season information became available. These time points also coincide with points in the season when a management decision is made, such as the application of fertiliser. The models were evaluated using cross-validation based on paddocks and years and this was assessed at the spatial resolution of the paddock. The models performed better as the season progressed, largely because more information about within-season data became available (e.g. rainfall). Cross-validated results showed the models predicted yield very accurately, with an RMSE of 0.36 to 0.42 t/ha, and an LCCC of 0.89 to 0.92 at the paddock resolution. The more years of yield data that were available for a paddock, the better the predictions were. The generic nature of this method makes it possible to apply to other agricultural systems where yield monitor data is available. A data-driven approach to predicting crop yield as an alternative to using mechanistic models has several advantages. Future work should explore integration of more data sources into the models and focus on using the models to inform management decisions such as fertiliser applications.

Background

Agricultural and environmental data is becoming increasingly available at finer spatial and temporal resolutions and at declining costs. While this data is abundant and potentially very useful, it is often in different formats and located in a variety of repositories, which makes it difficult to utilise. Every crop can essentially be considered as an experiment, where the yield is a function of the interaction between a suite of variables that vary in space and time. Many growers have a time series of crop yield monitor data for their paddocks which are generally augmented with additional data such as EM and gamma radiometrics surveys. In addition there are now freely and readily available national and global datasets which can be used to represent the crop-growing environment. Rather than analysing one paddock at a time, there is an opportunity to explore the value of combining this data over multiple paddocks and years into one dataset. Machine learning techniques are well-equipped to deal with large datasets with many variables, and they provide the opportunity to create predictive models of crop yield using this multitude of data. In this study we used grower-collected and nationally available data in combination with Random Forest models (Breiman 2001) to create predictive yield models of wheat, barley and canola at 3 vital time points in the growing season. This study particularly focused on the forecasting ability of the models based on pre- and mid-season information from predictors.

Methods

Datasets

An assortment of spatial and temporal grower-collected and nationally-available environmental and agricultural data was collated into a space-time cube (STC). The data consisted of yield monitor data, soil information, geo-physical data, and remotely sensed information, of varying spatial and temporal resolutions (Table 1). This study focused on a case study of several large farms in Western Australia. Yield from wheat, barley and canola crops from 3 different seasons that covered ~15,000 hectares in each year were used. In addition to this, the study area had been surveyed with EMI and gamma radiometric instruments to 10 m resolution. A STC can essentially be described as a large dataset where each row in the dataset possesses; spatial coordinates (easting + northing), time (year), yield, and a large string of associated covariates that relate to yield (predictor variables). Each row represents a spatial entity for a particular time point. Some spatial locations in the dataset possessed several years of yield data, while others only had one. Despite the varying spatial resolutions of the variables, they were all snapped to a common 10 m grid without changing their native spatial resolution (Table 1).

Table 1. Data sources used in the space-time cube to create predictive yield models

Data type	Model input	Resolution	Source of data
Yield	Yield monitor data	10 m	Grower
	Silo weighed- field averaged yields	Field/paddock	Grower
Soil	Soil maps - clay	10 m	Grower & Project
	Soil maps - sand	10 m	Grower & Project
Geo-physical data	EM surveys	10 m	Grower
	Gamma surveys	10 m	Grower
Remote sensed data	MODIS NDVI	250 m (8 days)	National
Climate data	Received rainfall	5000 m (daily)	National
	Forecasted rainfall	5000 m (daily)	National

Yield monitor data at 10 m resolution was corrected and standardised using measured paddock average silo weight, as it is known that different yield monitors can vary in their calibrations. Maps of soil sand and clay content at 10 m resolution were created using soil test results collected by growers. These sand and clay maps were creating using Random Forest models with spatial coordinates, EM and radiometric surveys used as covariates. Within-season MODIS - NDVI measurements were included in the model and collected in the middle of July and September. Total rainfall received for each year for Jan 1st–Mar 31st, Apr1st–June 30th, Jul 1st–Aug 31st was also used as predictor variables. In addition, the forecast rainfall was used, which is the probability of exceeding the median rainfall for the ensuing three months. The dates for NDVI and aggregation of the rainfall were chosen to coincide with different important points in the winter crop season, e.g. sowing (April), mid-season N-fertiliser top-dressing allocation (July), and anthesis (September).

Modelling

Random Forests (a machine learning approach) were used in conjunction with this STC to create predictive models of crop yield. Rather than creating individual models for wheat, barley and canola, one model was created and crop type was included as a predictor variable. Three models were created based on pre-sowing, mid-season and late-season conditions to explore the changes in the predictive ability of the model as more within-season information became available. These time points also coincide with points in the season when a management decision is made, such as the application of fertiliser. The models were built at a 10 m resolution, and then predicted at 100 m. This was then aggregated up to the paddock-scale, and the prediction quality was then assessed at the paddock-scale spatial resolution. The model was evaluated using cross-validation based on paddocks and years. The paddock-year-out cross validation involved creating a model without one seasons worth of yield data for a particular paddock and then using that model to predict the yield for that paddock for the missing year. The paddock-out cross validation was similar, but involved leaving all prior yield out for that particular paddock to create the model, and then predicting on that paddock for a missing year.

Results

The models performed better as the season progressed, with the September model possessing the lowest root mean squared error (RMSE), and the highest Lin's concordance correlation coefficient (LCCC). At the paddock-resolution the models had a RMSE of ~0.36-0.42 t/ha for the paddock-year-out cross validation (Table 2; Fig. 1). The paddock-year-out cross validation always provided much better results compared to the paddock-out approach, which suggests the importance of prior yield information in model predictions (Table 2). This is supported by Fig. 2, which shows that as more seasons of prior data were available for an individual paddock, the predictions improved dramatically.

Table 2. Cross-validated results of crop yield predictions at the paddock resolution

Time point	**April (sowing)**		**July (top-dressing)**		**September (anthesis)**	
	RMSE (t/ha)	LCCC	RMSE (t/ha)	LCCC	RMSE (t/ha)	LCCC
Paddock-out CV	0.64	0.19	0.63	0.20	0.62	0.27
Paddock-year-out CV	0.42	0.89	0.39	0.91	0.36	0.92

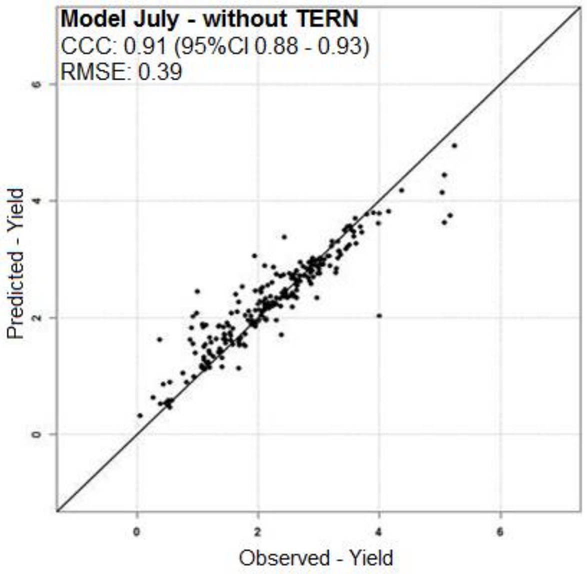

Figure 1. Plot of observed and predicted yield for July model for all paddocks using paddock-year-out cross validation approach

a) zero years b) one year 3) two years

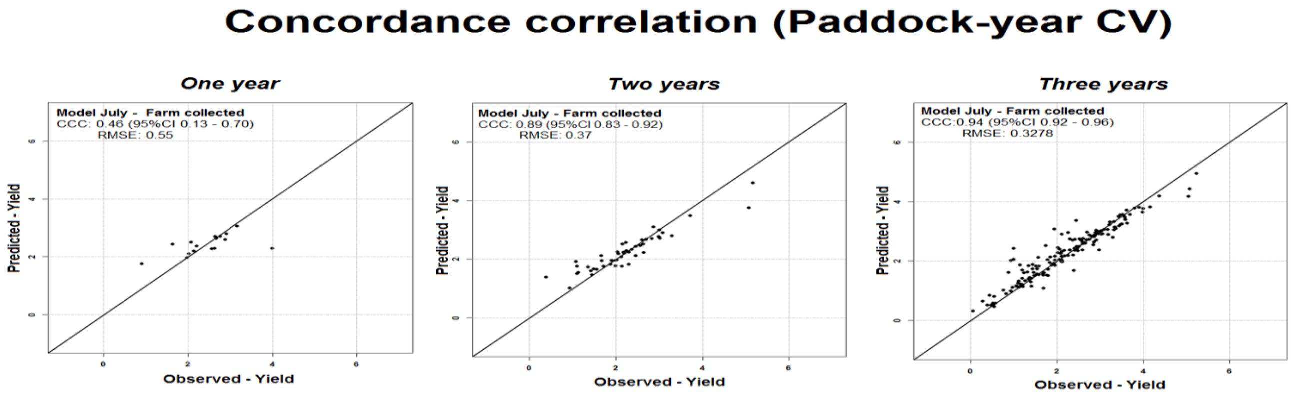

Figure 2. Cross validated results for paddocks that contained a) zero, b) one, c) two years of prior yield data

Within-season predictor variables proved to be very important covariates in the models. As an example, the most important predictors in the July model were; received rainfall, forecast rainfall, and within-season NDVI (Fig. 3). The soil and geo-physical data were less important predictors, however, many of these variables were highly correlated with each other, which may mask their combined significance. It could be assumed the importance of these types of predictors would increase if some were removed.

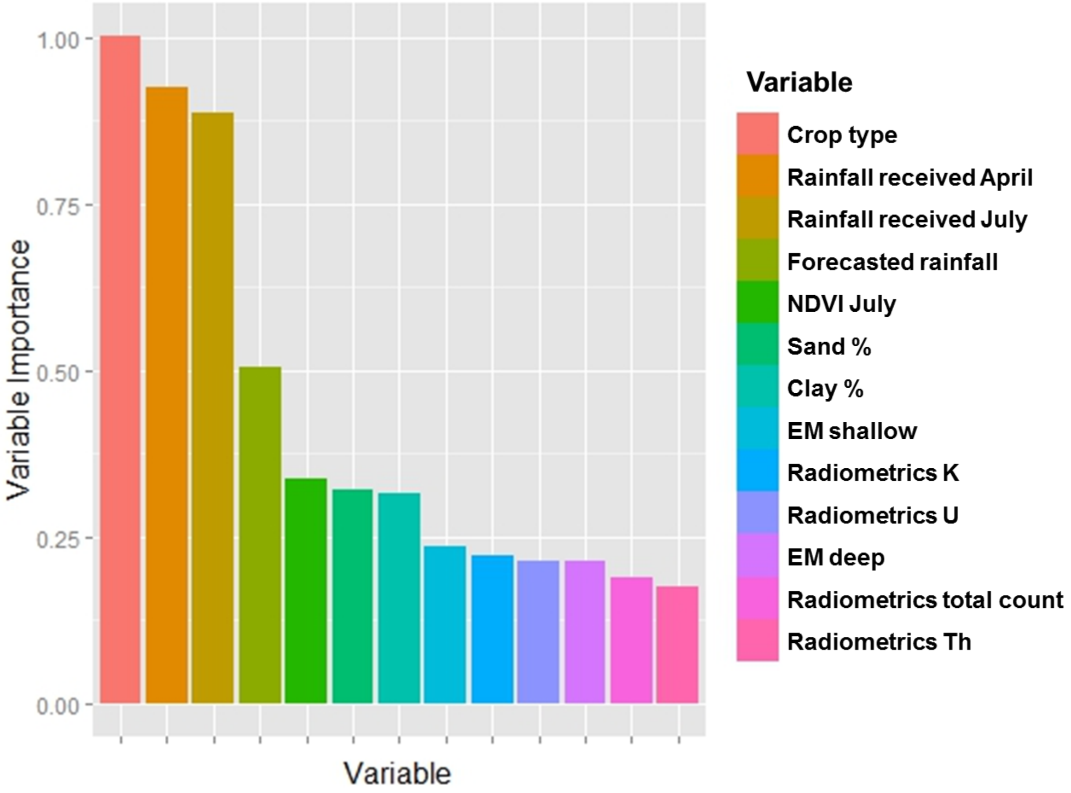

Figure 3. Predictor variable importance graph from July model

Discussion

Overall, the predictive crop yield models presented in this study performed well during cross-validation and had high predictive power (RMSE = 0.36 to 0.42 t/ha, and LCCC = 0.89 to 0.92). It was shown that prior yield data for predicting in a paddock resulted in much better models (Fig. 2). This is logical, as the model would have a better understanding of expected yield, and this suggests that a larger time-series of yield data for paddocks would greatly improve the prediction accuracy. Yield monitoring at harvest in the grains industry has been present in Australia agriculture for about two decades, and there is a great opportunity to maximally benefit from this available information to predict crop yields by using big-data approaches. The size of the dataset and how this affects predictions needs to be further explored. This needs to be done in terms of an expansion of temporal data (more seasons of yield), and an expansion of spatial data (more paddocks/farms). In this study we have successfully predicted yield for a collection of large farms, but this does not provide insight into whether or not this would work at a smaller spatial extent – e.g. for a single farm. The possibility of data sharing among growers is an option; however, this presents some challenges and limitations. It may be ideal to have one model for a region, or it may be better for individual farms to have a specific model, and this ideal area that the model covers should be further evaluated. Regardless, the results in this study show that a greater time-series of paddock yield would be extremely beneficial in improving yield predictions.

Within-season predictor variables proved to be very valuable in the models. The models improved in quality as the season progressed, and this is likely due to an increased amount of within-season predictor variables being used in these models. Integrating more of these within-season data sources into the model should be considered. This may include remotely sensed data from UAVs (Unmanned Aerial Vehicles), soil moisture products, degree days and Landsat data at 30 m resolution. Further work should consider the quality of the models under data-poor (only national datasets available) and data-rich scenarios (grower collected data available). This could identify the value proposition for growers when deciding on the type of data collect, as well as the best spatial and temporal resolution.

Currently, most approaches to predicting crop yield are through the use of mechanistic/simulation models, such as APSIM (Agricultural Production Systems sIMulator) (Keating et al. 2003). The disadvantages of mechanistic models are that they generally require numerous inputs and there are many assumptions made. The advantage of our empirical approach is that real, on-farm data is used to drive predictions, allowing fewer assumptions to be made. Predictive models of the upcoming season's crop yield are extremely useful, particularly when the predictions are at fine spatial resolutions and a high accuracy. There are opportunities to use this information to identify yield gaps, decide on futures contracts and market speculation, and to inform decisions on precision agricultural management practices. In particular, the incorporation of management inputs with these models is a promising avenue for future research, for example fertiliser application, gypsum/lime application or seeding rates.

Conclusion

In this study, we have presented a data-driven approach to predicting wheat, barley, and canola crop yield as an alternative to approaches that use mechanistic models. The approach presented was a success and its generic nature makes it possible to apply it to many other agricultural systems where yield monitor data is available. Future work should explore integration of more data sources, particularly within-season measurements (UAV, soil moisture products etc.) into the model. In addition, focus should be placed on using the predictive model to inform management decisions such as fertiliser applications.

Acknowledgments

We would like to thank Lawson Grains, Precision Agronomics, Precision Cropping Technologies and CSIRO for providing access to the data.

References

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

Keating BA, Carberry PS, Hammer GL, Probert ME, Robertson MJ, Holzworth D, Huth NI, Hargreaves JNG, Meinke H, Hochman Z, McLean G, Verburg K, Snow V, Dimes JP, Silburn M, Wang E, Brown S, Bristow KL, Asseng S, Chapman S, McCown RL, Freebairn DM, Smith CJ 2003. An overview of APSIM, a model designed for farming systems simulation. European Journal of Agronomy 18: 267–288.