

Practical opportunities to leverage existing on- & off-farm data layers to support decision making

Brett Whelan and Mario Fajardo.

University of Sydney.

GRDC project code: 9176493

Keywords

- yield data, publicly available data, digital agriculture, machine learning.

Take home messages

- The pool of publicly available off-farm data that may be relevant to combine with on-farm data is increasing and can now be swiftly gathered for any farm or field. Collecting and using this data to make more informed decisions is an opportunity for growers.
- Machine learning and hybrid models derived from large data sets and field validation should be tested against crop simulation models currently in use for estimating yield potential and input requirements/crop response.
- Using these techniques provides the opportunity to use large data sets that cover a local area for analysis of the drivers of variability in crop performance and profit rather than just using individual field data in the analysis as is the current Precision Agriculture technique. Building a freely available weather database at a much finer scale than is available now to improve predictions should be an industry and government imperative.
- Using power output or fuel use data recorded while working with ground-engaging implements may be a low-cost, novel way to map changes in soil strength/type.

Background

Australian broadacre crop production currently provides approximately 35% of agricultural's gross domestic product (GDP) and the nations farmers export approximately 75% of production into competitive international markets. This is achieved with a very low level of external financial support compared to almost all competitors (Organisation for Economic Co-operation and Development (OECD) Producer Support Estimate 2017, Australia = <2%; OECD = 18%, USA and Canada = 10%, Kazakhstan = 4%). Optimum business performance in a competitive environment requires the application of relevant information to critical decisions relating to improving efficiencies and production quantity/quality. In cropping businesses which operate in a variable environment, information on variability in resources, environmental conditions and output

is an important component of the relevant information required.

GRDC future farm program

The future farm program (a joint investment by GRDC, CSIRO, USYD, USQ, QUT and Agriculture Victoria) aims to tackle the issue of variability in resources, environmental conditions and output by utilising off-farm data and historical on-farm data to re-examine and improve the way in which current in-season field monitored data (soil, crop, climatic) is used to inform decisions about input management. The outcome should be an automated process from data acquisition, through analysis, to the formulation and implementation of decision options with manager input. The initial focus is on improving the efficiency and profitability of applied nitrogen (N).



The initial operational parameters are:

- N fertiliser decision making should be supported by measures of plant N status (which in turn requires estimation of biomass), soil N status and soil water status/availability (i.e. a multi-sensor approach is required).
- In-season sensor data will be a key input and employ machine learning methods of data integration for development of location-specific decision options.
- Both remote and proximal sensing of the crop canopy will make an important contribution to N fertiliser decision making but should be supported by some form of on-farm experimentation, with a zero N treatment (plot or strip) functioning as a critical enabler for interpretation.
- The process should be deployable in a way that will be complementary to the inclusion of other inputs/assessments that managers may also bring to bear in decision making.
- The data that is available for use is both public and privately owned.

Privately owned data

Yield monitor and terrain data

In terms of high value, low cost information for broadacre cropping, yield monitor data should be high on a cropping managers' list. With the overwhelming majority of farms already accessing high accuracy global navigation satellite systems (GNSS) and yield monitors becoming standard equipment on most harvesters, the yield mass (t/ha), grain moisture (%) and elevation (m) data available from these systems automatically during harvest operations comes at a low financial cost (Figure 1). The elevation data can be used to produce a range of useful information relating to changes in the landscape and its impact on soil development, water movement and solar radiation aspect. Calibrated crop yield and moisture data directly records variability in production across fields and years, and the high spatial resolution yield data is a simple, yet crucial method for monitoring or modelling the effect of management changes on production and is a layer of data vital to ground-truth data gathered from off-farm.

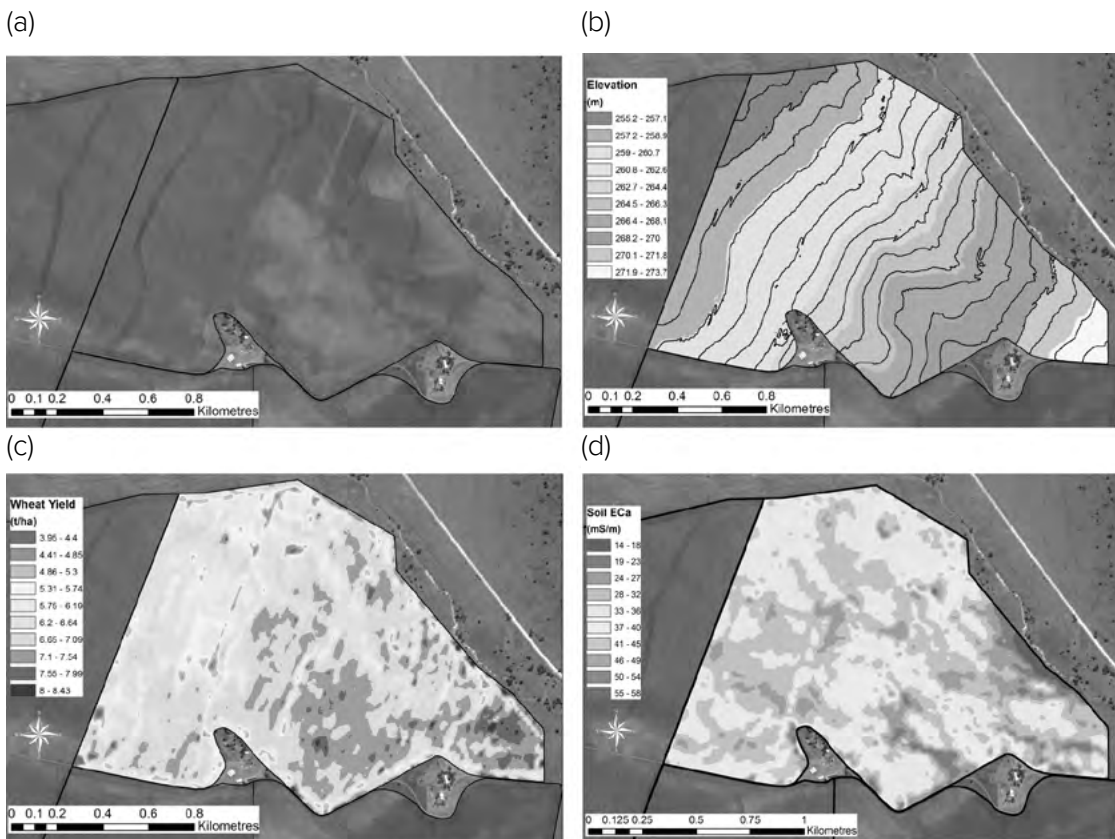


Figure 1. Paddock imagery on Google Earth (a), elevation data (b), crop yield data (c) and SOIL ECa (d).



Soil data

Identifying links between variability in crop production with local variation in soil properties should enable more considered crop management decisions. Data from point sample analysis is obviously valuable, but the value is increased if combined with higher spatial resolution measurements such as soil apparent electrical conductivity (ECa) (Figure 1). Together this information can be used to map the variability in the soil resource at the same resolution as crop yield monitors and remotely sensed crop imagery.

Vehicle performance data

Performance data is routinely recorded by newer tractors and self-propelled implements. Data on variation in fuel use and other relevant operational parameters can have economic and efficiency dividends. Novel ways to use this free data include using power output or fuel use while working with ground-engaging implements to map changes in soil strength/type (Figure 2). There is also the potential to use the fuel use data in a carbon and nitrogen auditing process.

Publicly available data

The progress towards increasing use of digital data in agriculture is being led by a combination of improvements in sensor development, computing power, data storage/delivery, data analytical techniques and reduced costs. The synchronisation of these occurrences has in turn fuelled a greater interest in the data and its potential, thereby

stimulating more development in all areas. A complementary benefit has been a rising number of data sources being made publicly, and more easily, available. Table 1 records a number of the most relevant data sources as of early 2020.

Publicly available information can be downloaded from a range of individual providers (for example; Geoscience Australia, CSIRO, ESA and NASA). This can be achieved on a number of platforms (for example; Python, R and Android) using Application Program Interface (APIs).

In 2010, Google released the 'Google Earth Engine'; a platform dedicated to providing access to a multitude of different data layers (including most listed in Table 1) at the cost of registration time only. A substantial advantage of this platform is that all the information is stored in a database standardised to different resolutions, in the same geographic reference system. Data can be downloaded in user-configurable locations and resolutions for use as required. For example, yield data (or the location of yield data points) can be uploaded to Google Earth Engine as private data and then the public data layers can be extracted to match the geographic locations and extent of the input layers. Alternatively, boundaries of farms or fields can be used to clip and extract data layers. The platform also enables real-time analysis using Google's computing infrastructure that runs processes across thousands of computers in parallel, enabling large analytical tasks to be performed or task time to be drastically reduced.

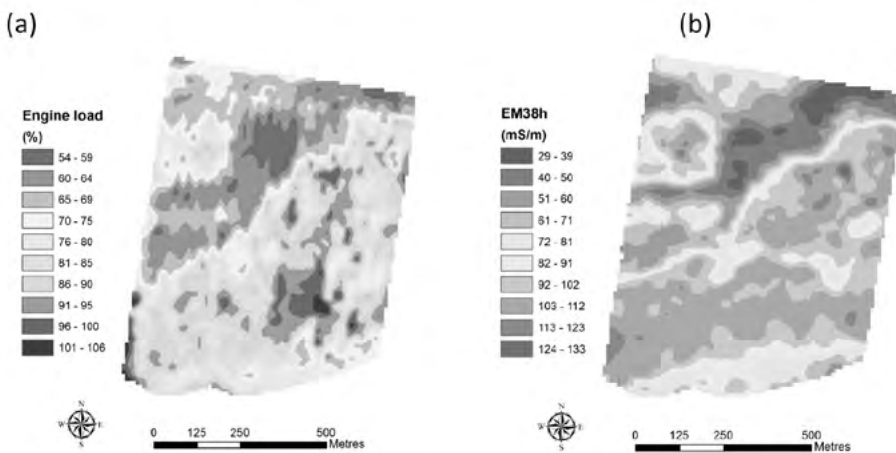


Figure 2. Tractor engine load recorded during the sowing operation (a) and soil apparent electrical conductivity (ECa) (b). Correlation coefficient = 0.85. (Source: data supplied by Rupert McLaren, 'Glenmore' Barmedman, NSW).



Table 1. Public sources of data for potential use in describing variability in resources and production.

Provider	Resource	Data	Spatial Resolution	Temporal Resolution
CSIRO	SLGA	Bulk density, organic carbon, clay, sand, silt, pH, available water capacity, total N, total P, effective cation exchange capacity, depth of regolith, soil depth, coarse fragments.	90 x 90 x 2 m	Static
BOM	Gridded Daily Data	Rainfall, temp, vapour pressure, solar exposure, NDVI, atmospheric circulation	5 x 5 km	Daily from 1889
NASA	SRTM	Digital Elevation Model (DEM)	90 x 90 m	Static
Geoscience Australia	ELVIS	Digital Elevation Models (DEM) and (DEM-S)	5 x 5 m	Static
Geoscience Australia	ELVIS	Hydrological DEM (Hydrological features are enhanced)	30 x 30 m	Static
ESA	Sentinel 2	13 bands from ~ 430 to 2190 nm	10-20-60	Weekly
Geoscience Australia	AWAGS	Radiometric map of Australia	100 x 100 m	Static
NASA	ASTER	14 Bands from visible to thermal IR	15-30-90 m	Weekly from 2000
NASA	MODIS	36 spectral bands	250-500-1000 m	Weekly until 2010
NASA	LANDSAT 7	8 bands from ~450 to 2350 + ~10.400 to 12.500	15-30-60 m	Second week
NASA	LANDSAT 8	11 bands from ~435 to 2294 + ~10.600 to 12.500	15-30-60 m	Second week
NASA	SMAP	Soil moisture and Carbon Net Ecosystem Exchange	9 x 9 km	Weekly and Second week
QANDL	Financial	Futures and commodity exchange data	-	various

Machine learning and data fusion approaches

The increase in availability of digital data and processing capabilities is leading to the application of data fusion techniques and machine learning to search for new insights from the data. There have been significant developments in machine learning analytical methods, which differ from mechanistic or process-based models that are commonly used in cropping because they use data-driven approaches to discover relationships between variables.

A major advantage is that they can make use of both quantitative and qualitative data from a wide range of data sources. On-farm data from sensors currently used in precision agriculture, along with what will be an increasing variety of sources, volumes, scales and structures of off-farm data (from other local/regional farms and the non-farm domains shown above) can be inputted into analysis and decision-making back on-farm.

An example testing of the use of data combinations

Data from a total of eleven paddocks (approximately 300ha each) were used to build six different data sets for this study, covering the period between 2016 to 2018. Data from seven of the paddocks were used in all six data sets and data for the four extra paddocks were added to the last two data sets to test for any benefit of increasing the spatial extent of the information.

The first two data sets were built using data from publicly available sources extracted at a 5m resolution through Google Earth Engine for each of the seven main paddocks. The data included surface soil clay content (%), available water content (%), bulk density (mg/Kg), soil depth (m), regolith depth (m), effective cation exchange capacity (meq/100g), total nitrogen (%) and total phosphorus (%) modelled by the Terrestrial Ecosystem Research network (TERN) project, plus a digital elevation model (DEM). The data also included satellite-monitored surface reflectance information from the clearest cloud free day in both the autumn and winter growing periods. The winter date meant the latest reflectance information was obtained approximately three months prior to harvest. The data sets differed in the source of this reflectance information, with data from the Landsat 8 platform (bands 1 to 7) making up the ‘Public Landsat’ dataset and information from the Sentinel 2 platform (bands 2 to 8A, 11 and 12) making up the ‘Public Sentinel’ dataset.

The second group of data sets were built using the first two datasets and adding on-farm proximally sensed data which comprised apparent soil electrical conductivity (ECa) for 0-50cm and 0-150cm depths and soil gamma-radiometric data (Thorium, Potassium, Uranium and Total Count) at the same spatial resolution. These data sets were labelled ‘Private Landsat’ and ‘Private Sentinel’.

A third group of data sets used the first two data sets as a base and were augmented with the same

publicly available data from the four extra fields. These data sets were denoted 'Extra Public Landsat' and 'Extra Public Sentinel', respectively.

All paddocks had harvester-collected wheat yield information at a 5m resolution available within the years 2016 to 2018.

Modelling

Two modelling methods were tested. The first followed a Bootstrapped Regression Tree (BRR) technique and the second a Convolved Neural Network (CNN) technique, and both modelling techniques were validated against an independent subset of the training data sets, following a "Leave One Field Out" validation process. Samples from all the fields except one, were used in creating a model that then was used for predicting the one field that had been excluded. This makes it a test for the ability to predict production within a field without using actual historic yield data from the target field, just surrounding fields.

Whole field results

As the resolution (Landsat 8 < Sentinel 2) increased, the predictive performance increased according to the Lin's Concordance Correlation Coefficient (LCCC) assessment for all approaches (Figure 3). An interesting fact is the consistent low performance of CNN when used with Landsat 8 imagery (30m resolution) versus BRR. Possible

explanations are the known effects of the 'fine tuning' of the complex CNN modelling, as the architecture of the network greatly affects the final model results (Nevavuori et al., 2019).

With respect to the different data sets (Public, Private and Extra-Public), for the BRR approaches, the use of proximal sensing layers improved the results as the LCCC values for Private were consistently higher than the Public approach. This agrees with a whole body of research studying the relation between soil properties and proximal soil sensing surveys.

The use of an augmented dataset (Extra-Public), excluding the use of CNN with Landsat 8 imagery, showed better LCCC results compared with the other the strategies. The best forecasting approaches was CNN Extra-Public and BRR Private. This modelling approach used Sentinel 2 imagery plus freely available soil and landform attributes. The main characteristics of this type of modelling were the higher resolution of the input imagery (10m from Sentinel 2 compared with 30m for Landsat 8) and a bigger dataset (extra training samples from the four extra contiguous paddocks). The hypothesis behind the improvement using the Extra-Public strategy relies on the inclusion of a higher variance in the training dataset, increasing the generalisation power of the model, in other words it can cover a wider range of predictions.

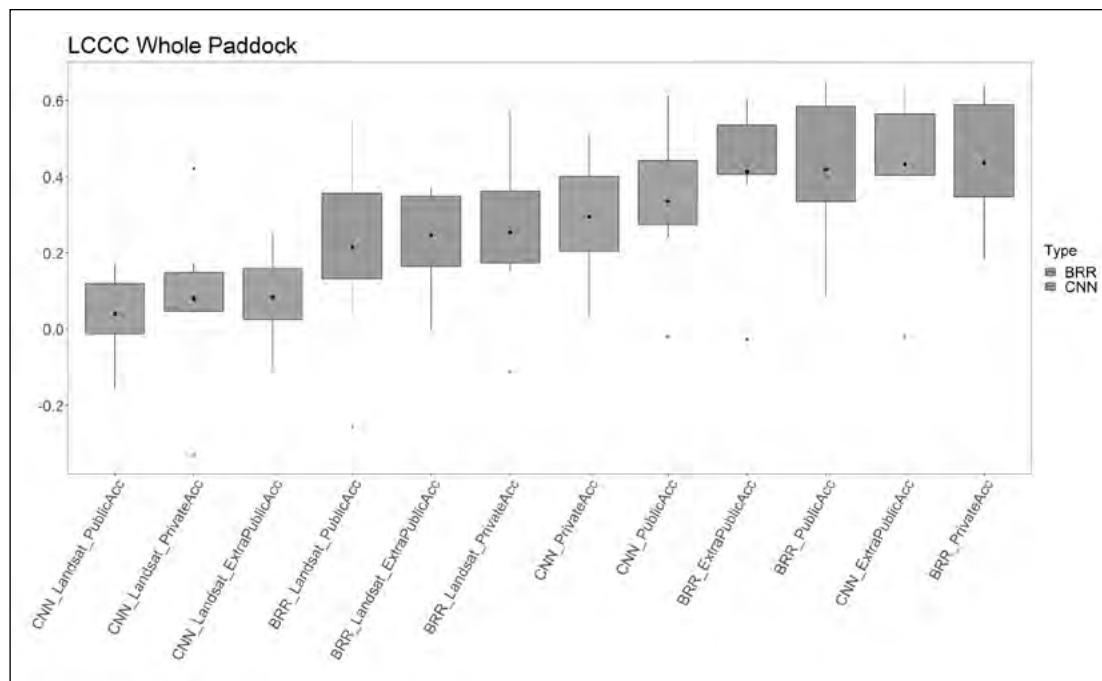


Figure 3. Lin's Concordance Correlation Coefficient (LCCC) for the different forecasting approaches. Tested using both modelling methods; a Bootstrapped Regression Tree (BRR) and a Convolved Neural Network (CNN) technique.



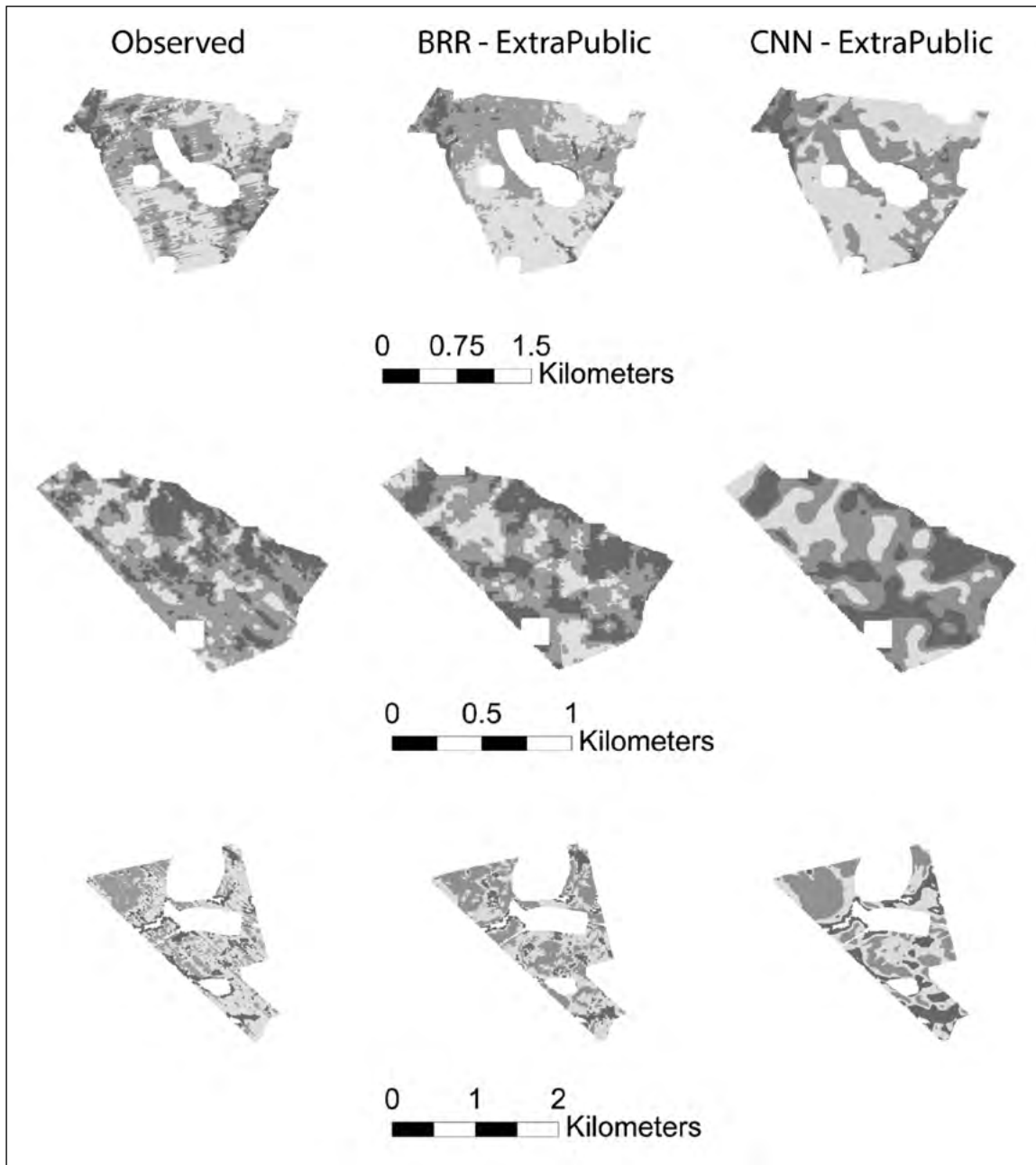


Figure 4. Three class K-means clustering of observed and predicted yield patterns for three example paddocks.

Within-field pattern results

Figure 4 shows a K-means clustering (3 classes) for three selected paddocks, with actual yield and predictions from the best BRR and CNN models.

A completeness analyses (C) was calculated to compare the spatial patterns of the different maps. Since the predictions are in a continuous scale, they were first clustered using a K-means algorithm for three classes. This number of classes was used as an example of reflecting a basic low-mid-high classification often used by growers in order to reduce the complexity of variable-rate applications.

Figure 4 shows the K-means clustering (3 classes) for three selected paddocks, with actual yield and predictions from the BRR and CNN models.

The results of C (Figure 5) showed a similar trend to the whole-field statistical model performance indicator (LCCC). Again, Landsat 8 approaches had the lowest performance compared with finer resolution datasets. The best results were observed in CNN Extra-Public and BRR-Private followed by BRR Extra-Public. These results confirm the hypothesis that an augmented dataset will improve predictions and confirming the added value of soil



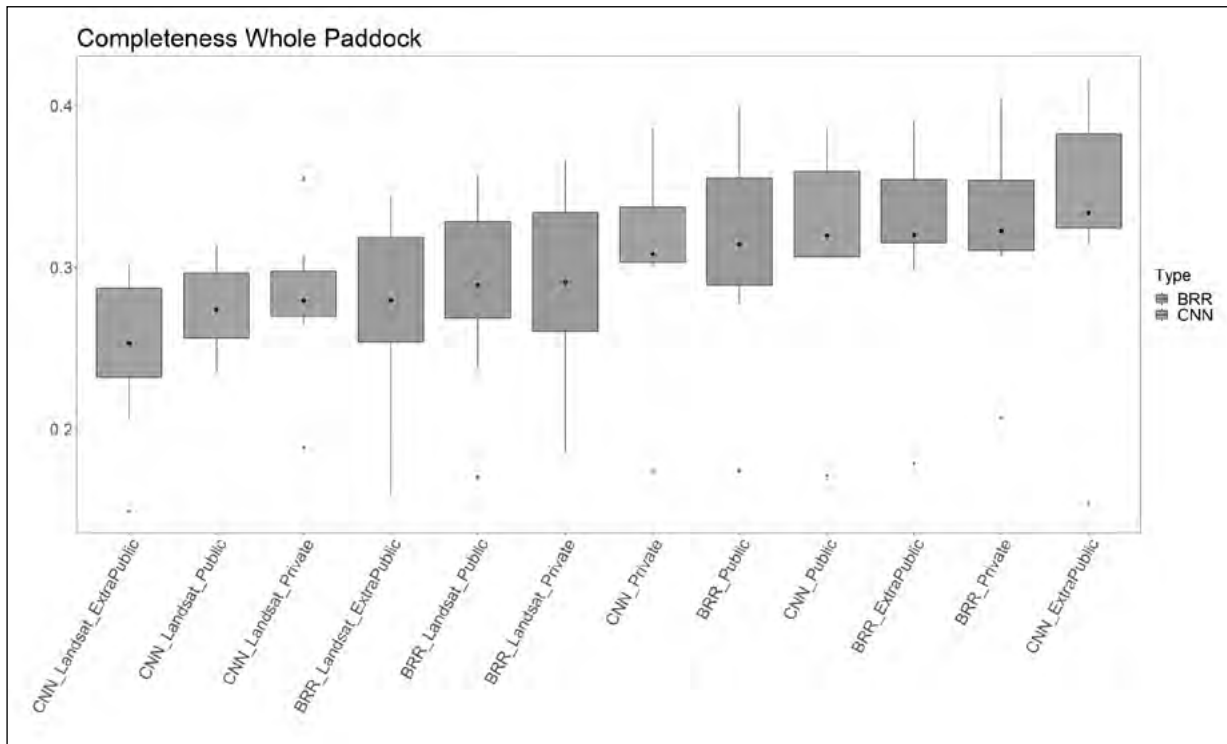


Figure 5. Completeness assessment for the different prediction approaches.

proximal sensing surveys as useful tool for crop yield pattern prediction, especially if using the lower resolution off-farm products.

Acknowledgements

The research undertaken as part of this project is made possible by the significant contributions of growers through both trial cooperation and the support of the GRDC, the authors would like to thank them for their continued support.

Contact details

Brett Whelan
 Precision Agriculture Laboratory, Sydney Institute of
 Agriculture, The University of Sydney
 Biomedical Building, Australian Technology Park,
 Eveleigh NSW 2015
 02 8627 1132
 brett.whelan@sydney.edu.au

 **Return to contents**

