Detecting causes of spatial variation in crop yield with interpretive machine learning

Patrick Filippi, Brett M. Whelan, Thomas F.A. Bishop

Precision Agriculture Laboratory, Sydney Institute of Agriculture, School of Life and Environmental Sciences, Faculty of Science, The University of Sydney, Sydney, NSW 2006, Email: patrick.filippi@sydney.edu.au

Abstract
Crop yield is driven by a suite of spatial and temporal variables and their interaction. Farmers and agronomists often have an idea of the causes of crop yield variability, but this is generally qualitative and poorly recorded. They need a more quantitative and systematic approach to identify and understand the causes of variation in crop yield in order to implement appropriate management responses. This study created empirical gradient boosted decision tree models of cotton yield to understand the spatial drivers of yield. Cotton yield monitor data from 38 different fields from neighbouring irrigated cotton farms for the 2013 season in the Gwydir Valley, NSW, was used. Predictor variables used for modelling included 30 m EVI satellite imagery, and 30 m digital soil maps of constraints (EC, ESP, pH). The model could describe the spatial variation in yield well, with a Lin's Concordance Correlation Coefficient (LCCC) of 0.75. Interpretive machine learning (SHAP values) was then used to identify variables that were contributing most to increasing or decreasing yield at 30 m resolution. This was mapped across the study area, as well as for a case study field. The preliminary results from this study provide promise in assisting growers to understand the relative causes of low yielding areas. This could lead to rectifying soil constraints or altering management inputs according to a constrained yield potential.

Keywords
Yield modelling, soil constraints, yield variability.

Introduction
Crop yields are affected by many different spatial and temporal variables, and it is often not easy to detect the exact cause of yield variability. This makes it difficult to implement management strategies to address constraints to production. There is a need for a more quantitative and systematic approach to detect the causes of yield variability and low yields. Many studies have focused on creating empirical models of crop yield using a variety of spatial and temporal datasets, such as satellite imagery, weather data, and soil maps. Such studies have been built for various purposes, such as forecasting (Filippi *et al.* 2020), and hindcasting (Donohue *et al.* 2018), but there is also an opportunity to create empirical yield models for the purpose of understanding the primary drivers of yield variation.

Many machine learning models have in-built methods for assessing the relative importance of included predictor variables, but these are quite general and have many limitations. These methods simply assess variable importance at a global scale, providing no insight as to what is driving yield in particular locations, and in certain seasons. Interpretive machine learning (IML) techniques are an opportunity to overcome this limitation, as they can identify which variables contribute to increasing, or decreasing yield at discrete locations, and their relative contribution. This study uses a large yield mapping dataset from 38 different fields to create a machine learning model with satellite imagery and soil constraint maps as predictor variables. IML was then used to identify the local primary spatial drivers of cotton yield across the study area, and the variables causing the largest increase and decrease in yield are then mapped at 30 m resolution for the whole study area, as well as a case study field.

Methods
Study area and datasets
The study region is a collection of 38 different irrigated cotton fields from neighbouring farms in the Gwydir River catchment, NSW, Australia. The annual average rainfall is 585 mm, and the soils are

primarily heavy-textured cracking clays (Grey and Brown Vertosols). Cotton yield monitor data for the 2013 season for all 38 fields was cleaned and processed to a 30 m grid (Filippi *et al.* 2020). To create a model of crop yield, satellite imagery and soil maps were used as predictor variables. The satellite imagery used was the maximum EVI (Enhanced Vegetation Index) value between October and January. The soil maps were created by Tilse *et al.* (2021) with a combination of soil survey data and diverse spatial covariates. The soil properties used in this work were EC (electrical conductivity), ESP (exchangeable sodium percentage), and pH for the 30-60 cm. The 30-60 cm depth was chosen as subsoil constraints are known to be important drivers of yield in the study area.

Modelling and interpretive machine learning

An XGBoost model (an implementation of gradient boosted decision trees) was used to create a predictive yield model at 30 m resolution with the 'xgboost' package (Chen et al. 2020) within the software R (R Core Team 2020). Interpretive machine learning was then used to identify the factors most contributing to yield. More specifically, SHAP (SHapley Additive exPlanations) values were calculated using the SHAPforxgboost package (Liu and Just 2020). SHAP values provide local model explanation and the contribution of each predictor variable for each data point. It is an advanced method of interpreting results from tree-based models and presents variable importance based on the marginal contribution to the model outcome. SHAP values are calculated for each cell in the training dataset, and the sum of each variable's SHAP values (and the bias) is the predicted model output.

Mapping variables contributing most to increasing/decreasing yield

The variable that contributed most to increasing yield (highest SHAP value) at each location was determined and then mapped across the study area, and a case study field. Likewise, the variable with the smallest SHAP value was determined at each location to show the variable most negatively impacting yield.

Results

The model described the variation in yield well, with an 0.75 LCCC and RMSE of 1.15 b ha^{-1} (Fig. 1). The SHAP variable importance plot ranks the global feature importance from top to bottom. Each point on the SHAP plot represents a data point in the model. Overall, EVI was the most important predictor for the model, with a high EVI resulting in a positive impact on yield. Soil maps of EC, pH, and then ESP then followed.

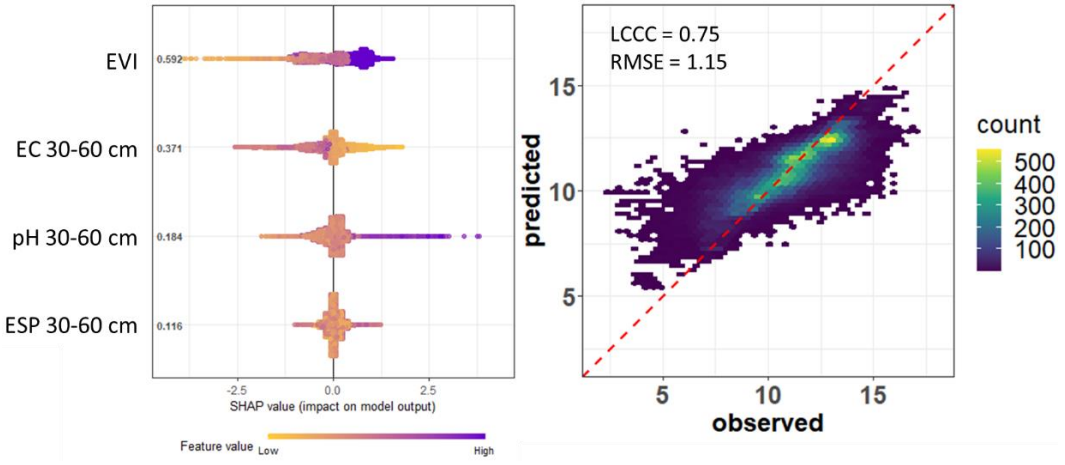

Figure 1. SHAP variable importance plot from XGBoost model across study area (left) and density plot of observed and predicted values (right)

The variable contributing to the largest increase in yield across the whole study area was EVI at 48%, of locations, followed by pH at 19%, ESP at 18%, and then EC at 16% of locations (Fig. 2). The

variables that contributed to the largest decreases in yield was EVI at 32% of locations, which was then followed by EC at 29%, pH at 21% and ESP at 17%.

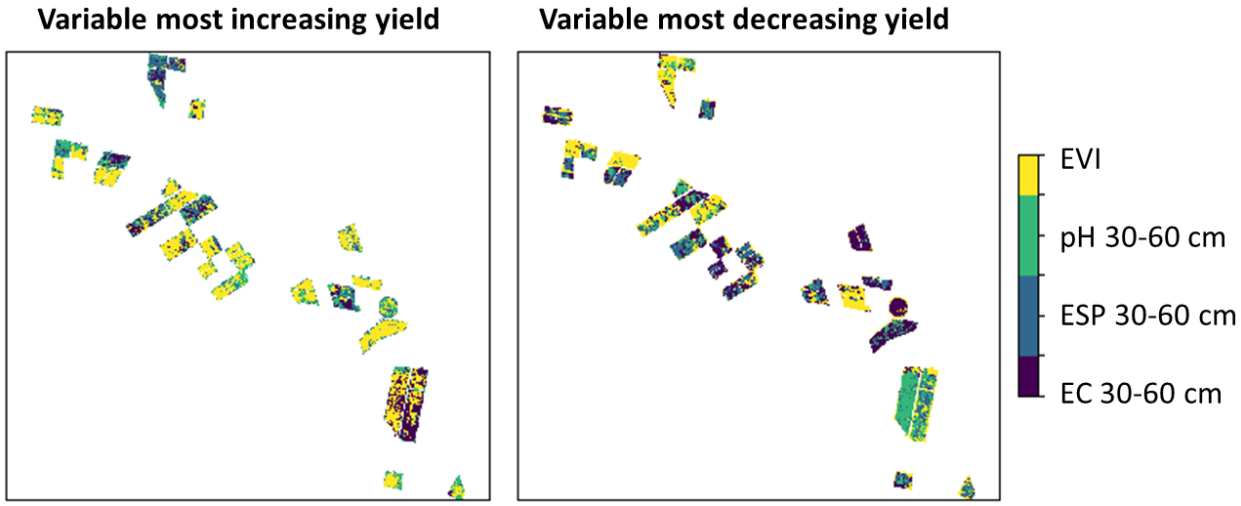

Figure 2. Variable contributing to largest increase in yield (left) and variable contributing to largest decrease in yield (right) across the whole study area

As Figure 2 is visually difficult to interpret due to the large number of fields, a case study field was selected to better demonstrate the approach. It is clear for this field that primarily EVI is contributing to positive impacts on yield (Fig. 3). However, the variables with the largest negative impact on yield were soil constraints ESP, EC, and pH. It must be acknowledged that there is a moderate level of auto-correlation between the three soil parameters used in this study, which somewhat limits the value of the analysis. These maps demonstrate that drivers of yield can be highly spatially variable within fields and that there is a changing importance of these influences on yield.

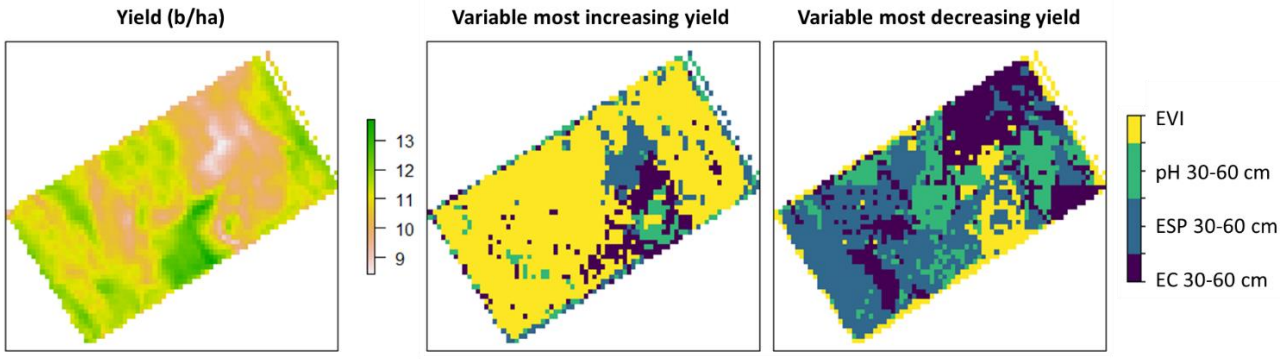

Figure 3. Cotton yield for 2013 season (left), variable contributing to largest increase in yield (middle) and variable contributing to largest decrease in yield (right) for a case-study field

Figure 4 shows the maps for the case study field of the predictor variables used in the modelling process, and the subsequent SHAP value maps showing how each variable impacts on the yield prediction for that field in bales per hectare. These maps allow the causes of spatial yield variation to be better understood, and help to identify the magnitude that each variable positively or negatively influences yield at different locations within fields in interpretable units (b ha^{-1}).

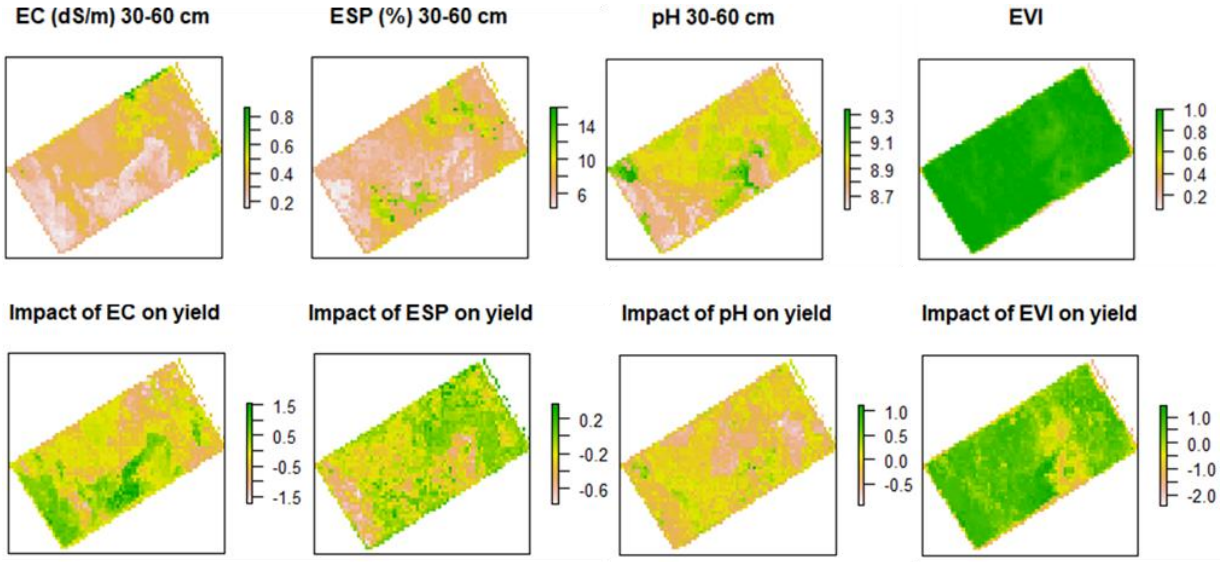

Figure 4. Map of predictor variables for case study field (top row) and corresponding map of impact of that variable on yield in b ha^{-1} (SHAP value) for case study field (bottom row)

Conclusion

This study used an XGBoost model to model cotton yield across 38 neighbouring fields for the 2013 season using yield monitor datasets, satellite imagery, and digital soil maps of important constraints. The model could describe yield variability well, with an LCCC of 0.75. The use of interpretive machine learning and SHAP values proved insightful for identifying the primary spatial drivers of yield between and within fields. Visualising these SHAP values into maps showing the variables most positively and negatively impacting yield eased the interpretation of the results. Overall, the approach implemented in this study shows potential for a more quantitative way for growers and agronomists to implement management strategies to manage and overcome the primary limitations of yield. This work will be extended in the future to include both spatial and temporal variables, and will also be implemented across multiple seasons.

References

Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, Li M, Xie J, Lin M, Geng Y, Li Y (2020) xgboost: Extreme Gradient Boosting. R package version 1.1.1.1. https://CRAN.R-project.org/package=xgboost

Donohue RJ, Lawes RA, Mata G, Gobbett D, Ouzman J (2018) Towards a national, remote-sensing-based model for predicting field-scale crop yield. *Field Crops Research* **227**, 79-90.

Filippi P, Whelan BM, Vervoort RW, Bishop TFA (2020) Mid-season empirical cotton yield forecasts at fine resolutions using large yield mapping datasets and diverse spatial covariates. *Agricultural Systems* **184**, 102894.

Liu Y, Just A (2020) SHAPforxgboost: SHAP Plots for 'XGBoost'. R package version 0.0.4. https://CRAN.R-project.org/package=SHAPforxgboost

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Tilse M, Bishop TFA, Triantafilis J, Filippi P (2021) Quantifying the impact of subsoil constraints on soil available water capacity and potential crop yield across multiple fields/farms. *Crop and Pasture Science*, Under review.