119. Decoding spatio-temporal drivers of yield variability with interpretive machine learning

S. Poole*, T.F.A Bishop, D. Al-Shammari and P. Filippi
*Sydney Institute of Agriculture, School of Life and Environmental Sciences, The University of Sydney, Eveleigh, NSW 2015, Australia; * sally.poole@sydney.edu.au*

Abstract

Interpretive machine learning (IML) techniques were utilised to determine if the spatio-temporal drivers of yield could be identified, quantified and map. For two case study farms, in different agroecological zones of Australia, digital soil maps and elevation data were used in XGBoost predictive models for multiple seasons of yield data. The Shapley Additive exPlanations (SHAP) values provided a localised explanation within the field of each year's predictive yield model. Temporally, the most commonly occurring yield limiting variable was determined for different crop types and rainfall seasonality. This study shows, the IML provides an interpretable way of determining the spatio-temporal drivers of yield that can vary in different agronomic contexts.

Keywords: crop yield, limiting factors, soil constraints, soil water, spatial variability, within-field variability

Introduction

In Australia, agricultural cropping systems are very diverse across the country due to the climate, rainfall patterns, soil type, and management systems. Yet, Australian broadacre systems have the commonality of being deemed some of the most water limited agricultural production of the world (Anderson *et al*., 2016). At a localised area, either on a farm or within a field, many cropping systems experience high spatial and temporal crop production variability. There are numerous spatial and temporal variables that affect crop yield in a range of complex interactions (Jaynes and Colvin, 1997). Farmers and agronomists often anecdotally know or understand the spatio-temporal drivers of crop variability through knowing the agronomic context of a field, the seasonal conditions, or understanding the soil. Yet, to meet the growing global demand for climate smart and sustainable food, these spatio-temporal drivers of crop variability need to be quantified using a systematic approach from which constraints can be addressed through site-specific crop management (SSCM) to help tailor inputs and reduce the yield gap.

Currently, these complex spatio-temporal interactions are infrequently analysed by studies, with many relying on simple correlation analysis between yield and spatial data layers such as soil apparent electrical conductivity (ECa) maps. Machine learning (ML) models are increasingly being employed to model key agricultural properties, such as crop yield, due to the increasing availability of spatio-temporal datasets and the ability of ML models to handle large, complex, and nonlinear relationships. However, whilst ML models have in-built methods of identifying the key predictor variables, these models are often deemed to be 'black boxes'(Jones *et al*., 2022). Interpretive machine learning (IML) techniques, in particular SHapley additive exPlanation values (SHAP) enables the contribution of each predictive variable within the model to be identified at each data point and the impact quantified in the same units as the response variable (e.g. for crop yield models, kg/ha) (Nehbandani *et al*., 2023). Jones *et al.* (2022) demonstrated that this methodology could identify the most yield limiting constraints spatially in an irrigated cotton field and quantified the impact on the crop at each location across the field. However, it was carried out only on one season of production data, in one distinct production system, which leads to the question of how this methodology

could be used to understand spatio-temporal drivers of yield in differing production systems and how these change from season to season. Thus, this paper utilised IML techniques to analyse and determine if the spatio-temporal drivers of yield could be identified, quantified and mapped for two case study fields in different agroecological zones and if these spatio-temporal drivers vary under different agronomic contexts or management.

Methods

Case study farms

This research focuses on two case study fields from two different dryland broadacre cropping farms located on the opposite sides of Australia. Farm A is 1099-ha broadacre field located on high clay content Vertosols (Isbell, 2016) in northern New South Wales (NSW), Australia. This NSW farm produces a range of broadacre crops, including cotton, wheat, and chickpeas and is a semi-arid climate receiving an average of 580 mm of annual rainfall. Farm B is a 210-ha field located on the high sand content Tenosols (Isbell, 2016) of the northern Western Australian (WA) wheat belt (Grundy *et al*., 2015). This WA farm receives an average of 476 mm of annual rainfall (Bureau of Meterology, 2024) and produces wheat, canola, barley and lupins.

Spatial input variables

Digital soil maps (DSM) of soil chemical and physical soil properties were created based on the approach of Filippi *et al*. (2019) that utilised a Random Forest model and a suite of spatial datasets, including proximally sensed (electromagnetic induction and gamma radiometric surveys), remotely sensed (RS) data (terrain data and multispectral bare earth imagery) and site specific soil data (35 soil cores for Farm A and 25 at Farm B, collected at four depth increments from 0-1 m down the soil profile and analysed for chemical and physical properties). The DSM models were validated using a leave-one-site-out-cross-validation approach and was assessed with the Lin's concordance correlation coefficient (LCCC) and root-mean-square-error (RMSE) before applying the model across the field. A pedotransfer function was used to model the soil water holding properties utilising the clay, sand, and organic carbon DSM (Padarian Campusano, 2014). To capture the flow of water across the field, a layer described as 'landscape change' (LC), was created using a moving window on the elevation data to give a map of the percentage above or below the average elevation within the window. Crop yield monitor data across multiple years was cleaned and interpolated for both farms using the approach of Taylor *et al*. (2007). Farm A had ten seasons of yield data including six seasons of wheat, three of chickpeas, and one of cotton yield. Farm B had six seasons of yield data including three seasons of canola, two of wheat, and one season of barley yield. To determine the influence of rainfall seasonality, that can vary substantially between seasons in Australia, the monthly rainfall data that was either collected on farm (Farm A since 1934) or at the local government weather station (Farm B, sourced from Bureau of Meterology, 2024) was combined to determine the growing season rainfall (GSR) (this included the fallow period prior to planting and in crop rain). This GSR was then classified for each season based on the lower (dry), upper (wet) and middle (average) quartiles that it fell within over the recorded length of the rainfall dataset. All yield and spatial data layers were collated into an individual dataset for each field based on a standardised 10 m grid of the fields.

Modelling and interpretive machine learning

Pearson's correlation analysis was used to examine the correlation between all the years of yield each field to analyse the spatial variability in yield data temporally. An extreme gradient boosting (XGBoost) model, which is an implementation of a gradient-boosted decision trees (Chen and Guestrin, 2016), was utilised to model each individual season of yield data. A suite of soil and elevation-related predictor variables (from the collated datasets) were used in the XGBoost model

of yield. The suite of variables for each farm were based on the local understand of the primary drivers of crop variability and identified subsoil constraints. For Farm A, the soil exchangeable sodium percentage (ESP), drained upper limit (DUL), and LC were utilised as the predictor variables. Whereas, for Farm B, the predictor variables were soil pH, soil sand percentage (%), and cation exchange capacity (CEC). Each predictive yield model was assessed using 10-fold cross-validation, and the quality of each model was assessed using LCCC and RMSE. For each season and each point in the field, the SHAP values were calculated using the 'SHAPforxgboost' package (Liu and Just, 2020). SHAP values provide a local model explanation by quantifying each predictor variables contribution to the prediction. At each location, for each individual season, the variable that had the lowest SHAP value was considered to cause the biggest relative reduction in yield and was used to create maps of the most yield limiting variable. To examine the temporal stability of these yield-limiting variables, the most commonly occurring yield limiting factors were examined by the crop types with multiple seasons of data (wheat, chickpea, or canola) and by rainfall seasonality (wet, dry, or average). Additionally, for both fields a case study season was examined to understand the influence that a variable rate (VR) urea management decision had, in combination with the other predictor variables, on the crop grown that season.

Results

Correlation analysis

The Pearsons correlation analysis of the 10 seasons of available yield data at Farm A (Figure 1a) shows that there is moderate correlation between some seasons such as the 2021 wheat and 2017 wheat. However, the correlation between consecutive year is often very weak with some being negatively correlated, such as 2014 wheat and 2016 chickpea. This reflects the inconsistence seen in the yield maps and suggests that the yield at Farm A has a less predictable spatial pattern of variability. However, the Pearsons correlation analysis for Farm B shows strong high correlations between all season of available yield data, indicating a temporally stable spatial pattern in yield variability.

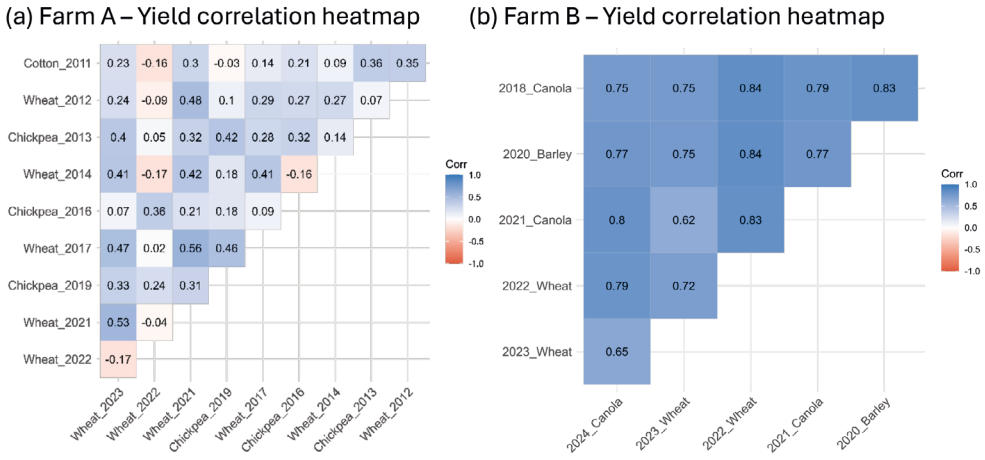

Figure 1. The Pearsons correlation plots of all seasons of yield data for (a) Farm A and (b) Farm B.

Interpretive machine learning

The results of the XGBoost model's predictive quality for each season's individual model were high, with LCCC ranging from 0.84 to 0.92 for Farm A and 0.69 to 0.86 for Farm B. Both models also have low RMSE of 0.14 to 0.33 t/ha for Farm A and 0.24 to 0.47 t/ha for Farm B. This shows the XGBoost can effectively predict the spatial variability in yield across all the seasons of available yield data using the predictor variables. This gives us more confidence to use the SHAP values.

The most yield limiting variables for Farm A are shown in Figures 2 by crop types (wheat and chickpea) and rainfall seasonality (wet or dry). There are some similar spatial trends in these maps, particularly for DUL through the central region of the field for the wheat (Figure 2a) and dry season (Figure 2c) maps which aligns with areas of the field with the highest DUL. The DUL was identified as the most limiting variable for chickpeas in map (Figure 2b) although in a more scattered pattern. The 'wet' season map (Figure 2d) shows no real distinct patterns.

However, the results for Farm B (Figure 3) show a very consistent and stable spatio-temporal patterns in the most yield limiting variable by crop type and rainfall seasonality. This is because a very similar spatial patterns occurred in crop yield each year (Figure 1b). These patterns follow the same spatial patterns as the soil texture which is highly influenced by the high sand and low clay contents.

The variable rate management (VR) case study for Farm A in the 2021 season, with the addition of the VR urea map (Figure 4a) shows an improvement in the interpretability of the most yield limiting variable map for the 2021 season (Figure 4b). The areas identified as being the most limiting by urea (light blue areas in Figure 4b) corresponds with the VR areas where the lowest rate of 60 kg/ha of urea was applied (pink areas in Figure 4a). For the VR case study for Farm B in the 2024 season, the variable rate application map (Figure 5a), which is based on the farm management zones, again follows that same spatio-temporal pattern of yield in the field and the 2024 canola yield. The most yield limiting variable map (Figure 5b) indicates that the area where the VR urea was the most yield limiting variable, largely corresponds with area of the field that received 0-80 kg/ha of urea.

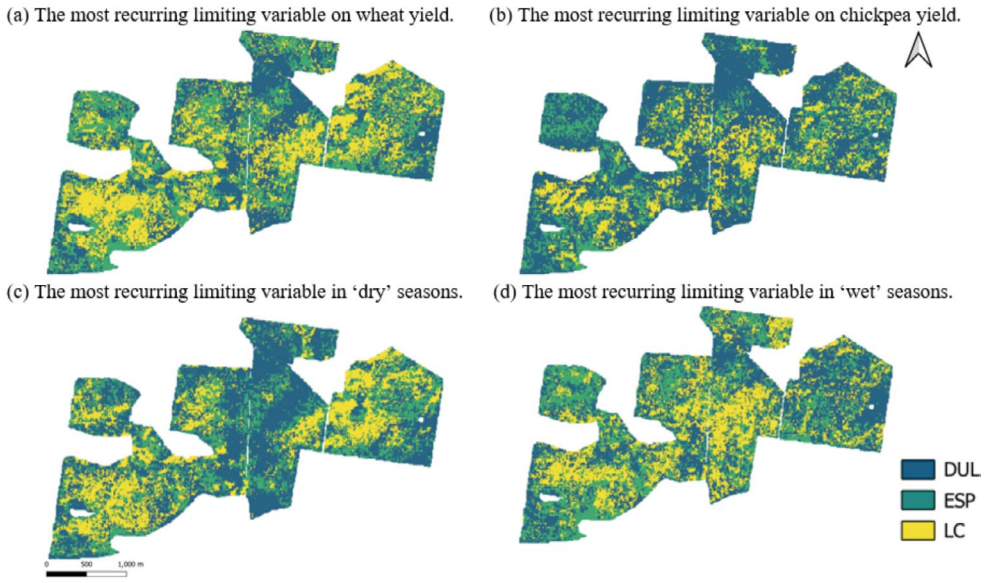

Figure 2. The spatio-temporal most commonly occurring limiting variable on yield at Farm A for (a) all wheat crops, (b) for all chickpea crops, (c) in dry seasons, and (d) in wet seasons.

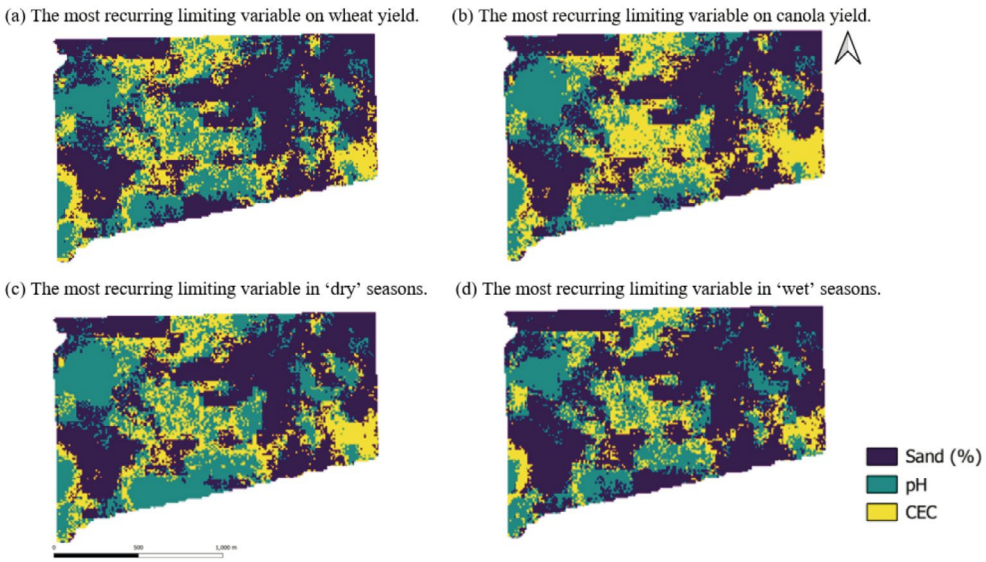

Figure 3. The spatio-temporal most commonly occurring limiting variables on yield at Farm B for (a) all wheat crops, (b) for all canola crops, (c) in dry rainfall season, and (d) in average rainfall seasons.

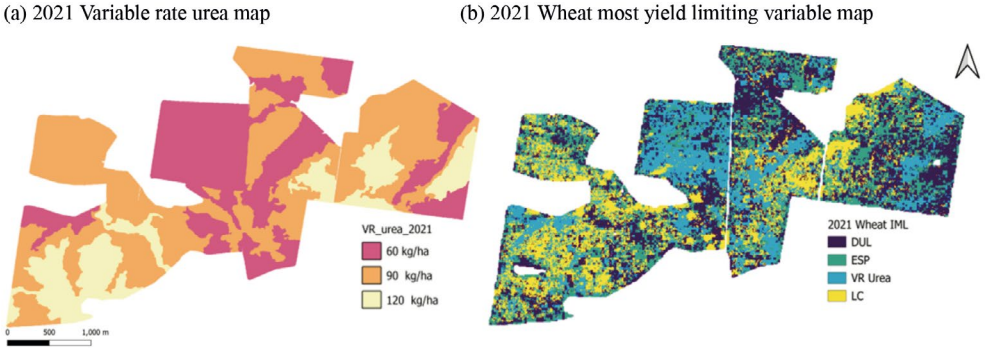

Figure 4. (a) the variable rate urea map applied the wheat crop at Farm A in 2021, and (b) the associate most yield limiting variable map for the wheat yield in 2021.

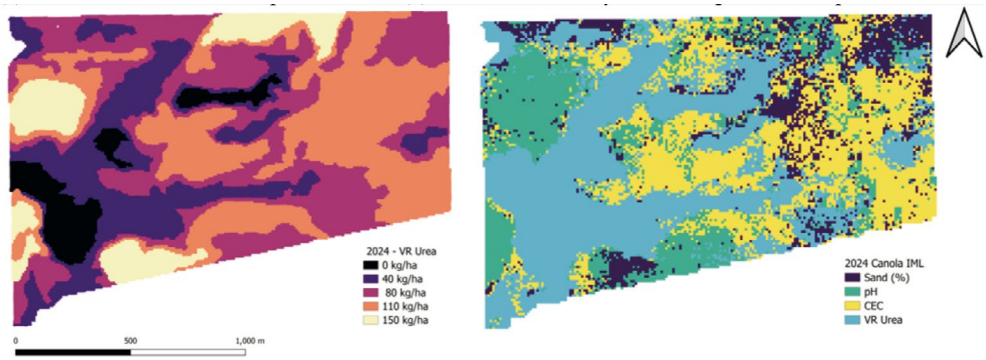

Figure 5. (a) the variable rate urea map applied the canola crop at Farm B in 2024, and (b) the associate most yield limiting variable map for the canola yield in 2024.

Discussion

The use of IML to identify the key spatial limiting variables to crop yield, for an individual season, has been demonstrated by Jones *et al*. (2022). This study shows that IML can be used to identify and easily interpret the key spatio-temporal yield limiting factors within a field and across multiple (6+) seasons. Additionally, it highlighted that spatio-temporal patterns of yield can also vary significantly in different cropping regions of Australia.

The spatio-temporal trends of the most commonly occurring yield limiting variable at Farm B are very consistent regardless of crop type or rainfall seasonality (Figure 3a-d) and follow the same distinct spatial trend of the highly correlated annual yield data. This suggests that spatial yield variability is less affected by temporal variables such as crop type or rainfall seasonality and is more highly affected by spatial changes in soil texture, in particular sand percentage. Making the Farm B easily manageable by soil-based zones from which inputs could be tailored to yield potentials and seasonal requirements.

However, the inconsistent spatio-temporal patterns of yield at Farm A, show the complex nature of farming on the Vertosol soils of northern NSW where it is inherently known to be affected by extreme rainfall seasonality, and soil constraints. The pattern in the most limiting constraints were more consistent, particularly through the central region of the field, for the wheat crops and in drier years. However, there was little interpretable pattern for the chickpea crop or wet season. The inconsistency in yield limiting variable in the 'wet' season map (Figure 2d) is likely due to the varying severity of the wet weather and associated waterlogging or flooding (3 years had major flood events), as different crop stages and types have different levels of sensitivity to the impact of waterlogging. The chickpea crops (Figure 2b) were largely affected by either extremely wet or dry season, with 2019 being the driest GSR on record. Additionally, these chickpea seasons were affected by other external factors such as disease which currently cannot be spatially mapped. However, in the future, the addition of such spatial data could be used to help quantify the impact and improve the interpretability of the IML. For Farm A, this inconsistence in spatial yield variability, most yield limiting constraints, and additional external variables, makes SSCM more challenging to implements based on yield. However, the IML does highlight frequently occurring areas of yield limiting variable, such as sodicity, that could be managed with SSCM based on the IML, to limit the impact on future crops.

The variable rate urea application case study at Farm A (Figure 4) showed that where the urea was identified as the most limiting constraint (Figure 4b) corresponded with the lower urea application rate of 60 kg/ha. This urea rate is only half that of the normal grower standard practice of 120 kg/ha of urea for wheat in the region, thus indicating that limited nitrogen was the most limiting driver of yield in those areas. The VR case study at Farm B indicates that areas where the VR urea was the most yield limiting variable (Figure 5b) also corresponded with the VR urea application rate of 80 kg/ha or less (Figure 5a). Whilst this could indicate the underapplication of urea, due to the static spatial pattern in yield each season for this farm, it may also indicate that the management zones, which the VR map is based on, are representative of yield potential of this field which is suited to SSCM. These case studies show that the IML models can be further improved with the addition of key spatial management layers, such as variable rate inputs.

Conclusion

Interpretive machine learning offers a way of identifying, mapping and easily interpreting the spatial and temporal drivers of yield. This study also showed the diversity, complexities and challenges that occur within different cropping systems. Farm B from WA had a stable and consistent spatio-temporal yield pattern and pattern of yield limiting variables. However, Farm A from NSW had more complex spatio-temporal patterns of yield limiting variables. This study also shows that the

addition of variable rate management decisions and perhaps other spatial data, such as pest, disease and management layers, may further improve the model's ability to predict and map limitations to crop yield. From this understanding of the spatio-temporal drivers of crop variability, new agronomic management decisions can be tailored and assessed to reduce the limitation of these variables on future crop production.

References

Anderson, W.K., Stephens, D., Siddique, K.H.M., Farooq, M., & Siddique, K.H.M. (2016). Dryland agriculture in australia: experiences and innovations. In Innovations in dryland agriculture. Springer International, Cham, pp. 299–319. https://doi.org/10.1007/978-3-319-47928-6_11

Bramley, R.G.V., & Ouzman, J. (2019). Farmer attitudes to the use of sensors and automation in fertilizer decision-making: nitrogen fertilization in the Australian grains sector. Precision Agriculture, 20(1), 157–175. https://doi.org/10.1007/s11119-018-9589-y

Bureau of Meterology. (2024). Monthly rainfall – Twin Hills. Available online at http://www.bom.gov.au/jsp/ncc/cdio/weatherData/av?p_nccObsCode=139&p_display_type=dataFile&p_stn_num=008289 (accessed November 2024).

Chen, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–795. https://doi.org/10.1145/2939672.2939785

Filippi, P., Jones, E.J., Ginns, B.J., Whelan, B. M., Roth, G.W., & Bishop, T.F.A. (2019). Mapping the depth-to-soil ph constraint, and the relationship with cotton and grain yield at the within-field scale. Agronomy, 9(5), 251. https://doi.org/10.3390/agronomy9050251

Grundy, M.J., Rossel, R.A.V., Searle, R.D., Wilson, P.L., Chen, C., & Gregory, L.J. (2015). Soil and landscape grid of Australia. Soil Research, 53(8), 835–844. https://doi.org/10.1071/SR15191

Haan, S., Harianto, J., Butterworth, N., Bishop, T.,. (2023). Geodata-harvester: a python package to jumpstart geospatial data extraction and analysis. Journal of Open Source Software, 8(89), 5205.

Isbell, R. (2016). The Australian Soil Classification. Csiro Publishing, Clayton, VIC.

Jaynes, D.B., & Colvin, T S. (1997). Spatiotemporal variability of corn and soybean yield. Agronomy Journal, 89(1), 30–37. https://doi.org/10.2134/agronj1997.00021962008900010005x

Jones, E.J., Bishop, T.F.A., Malone, B.P., Hulme, P.J., Whelan, B.M., & Filippi, P. (2022). Identifying causes of crop yield variability with interpretive machine learning. Computers and Electronics in Agriculture, 192, 106632. https://doi.org/10.1016/j.compag.2021.106632

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

Nehbandani, A., Filippi, P., Alizadeh-Dehkordi, P., Dadrasi, A., & Soltani, A. (2023). Use of interpretive machine learning and a crop model to investigate the impact of environment and management on soybean yield gap. Crop and Pasture Science, 75(1), CP23032. https://doi.org/10.1071/CP23032

Taylor, J.A., McBratney, A.B., & Whelan, B.M. (2007). Establishing management classes for broadacre agricultural production. Agronomy Journal, 99(5), 1366–1376. https://doi.org/10.2134/agronj2007.0070