Predicting, mapping, and understanding the drivers of grain protein content variability – utilizing harvested mounted grain protein sensors.

Mikaela J. Tilse*, Thomas F. A. Bishop, Patrick Filippi

Level 1, Biomedical Building (C81), 1 Central Avenue, Australian Technology Park, Eveleigh, New South Wales, Australia, 2015.

* mikaela.tilse@sydney.edu.au

A paper from the Proceedings of the
16th International Conference on Precision Agriculture
21-24 July 2024
Manhattan, Kansas, United States

Abstract.

Grain protein content (GPC) is a key determinant of the prices that grain growers receive, and the rising cost of production is shifting management focus towards optimizing this to maximise return on investment. Harvester-mounted grain protein sensors have been used to map grain protein on-the-go for more than 20 years (e.g. CropScan, John Deere). While there is growing interest in measuring and mapping within-field GPC to better understand variability and support management decisions, the uptake of these sensors has been slow and GPC maps are not available for every field, farm, or season, resulting in considerable gaps. There is the potential to utilize this grain protein sensor data to understand the nature and drivers of variability in GPC for improved management. Further, utilising dense yield and GPC data layers together provides an opportunity to explore local relationships between each to guide management decisions which optimize both yield and quality. In this work, we present the use of GPC sensor data for winter wheat from across 46 fields over 4 seasons (2020 – 2023; i.e. 63 FieldYears worth of data in total) across Western Australia and northern New South Wales, Australia, for mapping and modelling GPC. This study aims to 1) create a model to predict GPC in fields without a grain protein sensor, by a) using readily available yield, agronomic, and publicly-available data, b) validating model performance using two different validation approaches (leave-one-FieldYear-out cross validation, LOFYOCV, and two-fold cross-validation, 2FCV); and c) mapping predictions at different spatial resolutions (fine 30 m resolution, management zones); and 2) assess the relationship between GPC and wheat grain yield, both spatially and temporally, within fields. This research utilises a data-driven, machine learning (Random Forest) approach to process multiple layers of yield, agronomic (e.g. sowing and harvest dates, cropping history) and publicly-accessible data (e.g. digital elevation model, terrain attributes, satellite remote sensing imagery) into a useful format

The authors are solely responsible for the content of this paper, which is not a refereed publication. Citation of this work should state that it is from the Proceedings of the 16th International Conference on Precision Agriculture. Tilse, M. J., Bishop, T. F. A., Filippi, P. (2024). Predicting, mapping, and understanding the drivers of grain protein content variability – utilizing harvested mounted grain protein sensors. In Proceedings of the 16th International Conference on Precision Agriculture (unpaginated, online). Monticello, IL: International Society of Precision Agriculture.

for growers by predicting GPC within fields. The LOFYOCV and 2FCV approaches, which were used to simulate instances where grain protein sensor data is not available for parts-of (2FCV) or an entire field (LOFYOCV), demonstrated the potential to fill gaps and predict GPC across fields with little-or-no protein sensor data using readily available data layers. Results showed that a combination of yield, agronomic and publicly-available data layers produced the best quality predictions of GPC. Local correlations between GPC and yield were not always inverse, which is contrary to what is commonly expected, and varied spatially and season-to-season in both strength and distribution within a field. Overall, this research highlights the potential of utilizing readily available data layers to predict and map GPC, and demonstrates the value in utilizing GPC maps and sensing technologies to better understand variability and inform management decisions.

Keywords.
Precision agriculture, grain protein content, machine learning, grain protein sensor, yield

Introduction

Grain protein content (GPC) is one of the key determinants of the prices that grain growers receive. Like grain yield, within and between field variation of GPC is widespread (Figure 1). Grain protein content is determined by a range of genetic, environmental and management factors, including grain crop type, crop variety, nitrogen (N) in the soil and applied as fertiliser, and soil moisture availability throughout the growing season (Whelan and Taylor 2013). Accurately measuring and mapping GPC within a field, across a farm, and over multiple seasons, can be useful to better understand the nature and drivers of variability in GPC, manage the quality of marketed grain, and better understand, evaluate, and improve N nutrition decisions (Whelan 2019).

Harvester-mounted grain protein sensors have been used to map grain protein on-the-go for more than 20 years (e.g. CropScan), and in 2023, John Deere commercially released the HarvestLab 3000™ Grain Sensing system in Australia for real-time, on-the-go measurement of protein, starch, and oil values for wheat, barley, and canola. The sensor is mounted onboard the harvester and uses near infrared (NIR) spectroscopy to take measurements of continuous grain flow. The sensor emits radiation which passes through a glass window onto the grain sample. A portion of this radiation is absorbed by the grain, while some is reflected back to the sensor. This NIR reflectance is then measured, and the wavelengths are analyzed and used to determine properties such as grain protein, oil, or moisture content.

While there is growing interest in measuring and mapping within-field GPC to better understand variability and support management decisions, the uptake of these sensors has been slow and GPC maps are not available for every field, farm, or season. This is resulting in considerable knowledge-gaps. There is the potential to utilize this grain protein sensor data to understand how and why GPC varies and to improve management. By building a predictive model to predict GPC in fields without a grain protein sensor, growers and advisors can be equipped with the necessary information and tools to make better management decisions for more profitable and environmentally sustainable production systems. Together with grain yield maps, GPC maps can provide an opportunity to make future N management decisions and optimize both yield and quality, for example. Maps of GPC can be used in conjunction with wheat gain yield maps, input costs, and the final grain price to map gross margins and better understand the costs of variable GPC. Likewise, GPC maps can be useful to understand N dynamics and agronomy, including variation in N availability and the implications of fertilizer decisions prior to or during the growing season. Improving this understanding can have positive outcomes for on-farm economics, production efficiencies, and environmental sustainability.

Today, we have vast amounts of public data that is free to access, including remote sensing imagery. These data layers can represent variability and the factors driving GPC, including soil constraints or nutrient deficiencies, both within a season and over longer timescales. These publicly-available data layers can be used on their own, or in conjunction with on-farm data such as yield maps or cropping history information, to model and map GPC.

We present a data-driven, machine learning approach which utilizes a combination of yield, agronomic, and/or publicly-available data layers to model and predict GPC within fields and fill knowledge gaps across farms. The aims of this research were to:

1) Create a model to predict GPC in fields without a grain protein sensor, by:
    a. Using readily available yield, agronomic, and publicly-available data;
    b. Validating model performance using two different validation approaches; and
    c. Mapping predictions at different spatial resolutions.
2) Assess the relationship between GPC and wheat grain yield, both spatially and temporally, within fields.

This research aims to demonstrate the value of collecting grain protein data, and the use of this information alongside the growing amount of on-farm and publicly-available data layers to better understand and manage GPC.

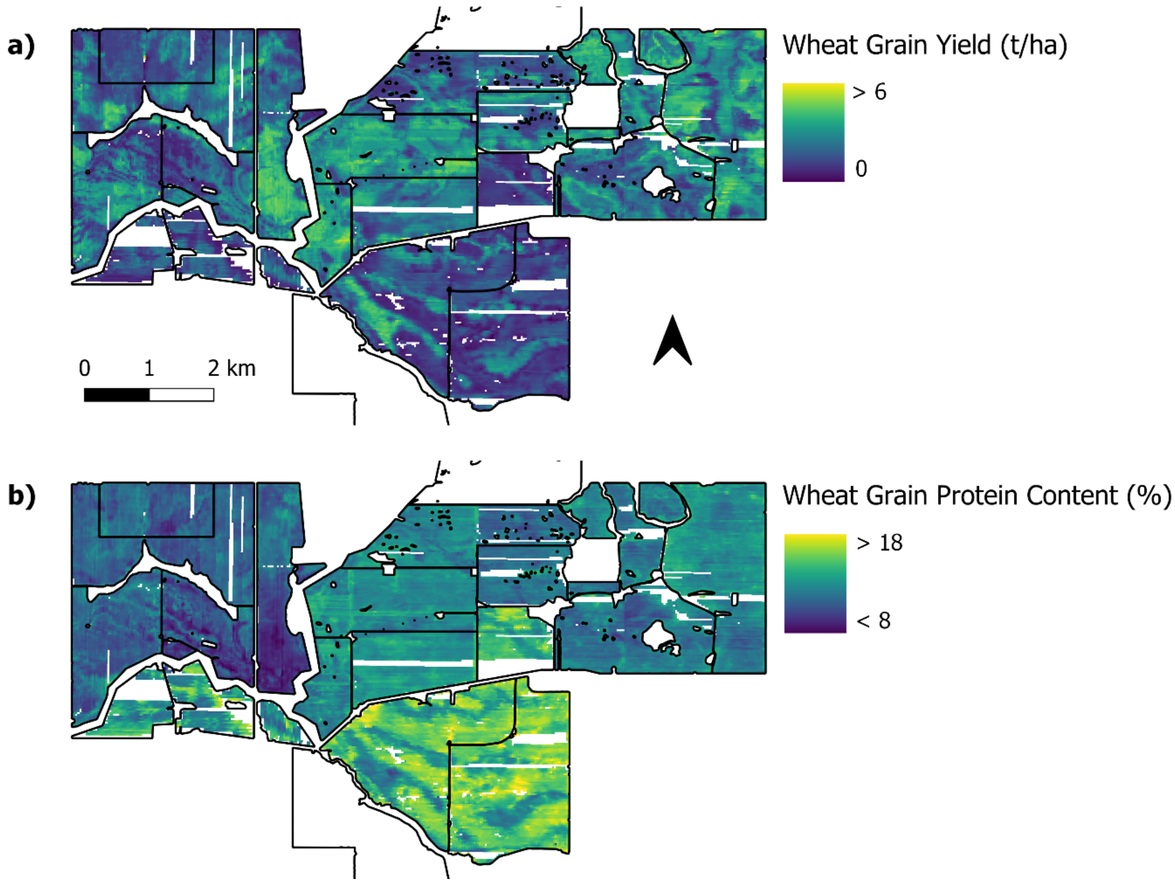

Figure 1. Spatial variation of wheat grain yield (t/ha, a) and wheat grain protein content (%, b) across a farm.

Methods

Study area

We present the use of grain protein sensor data for mapping and modelling the GPC of winter wheat between 2020 and 2023 across two large broadacre, dryland farms in Western Australia (WA) and northern New South Wales (NSW), Australia (Figure 2). Wheat grain yield and protein sensor data was collected at harvest for each season onboard harvesters equipped with the John Deere HarvestLab 3000™ NIR spectroscopy sensor. All yield and protein data points were collected at the same locations on each pass, and each yield/protein map was assigned a unique FieldYear identifier which represented the respective field name and season. There was 22 FieldYears worth of data for the NSW aggregation, and 41 for the WA aggregation. Fields ranged from 44 – 1248 hectares (ha) in size.

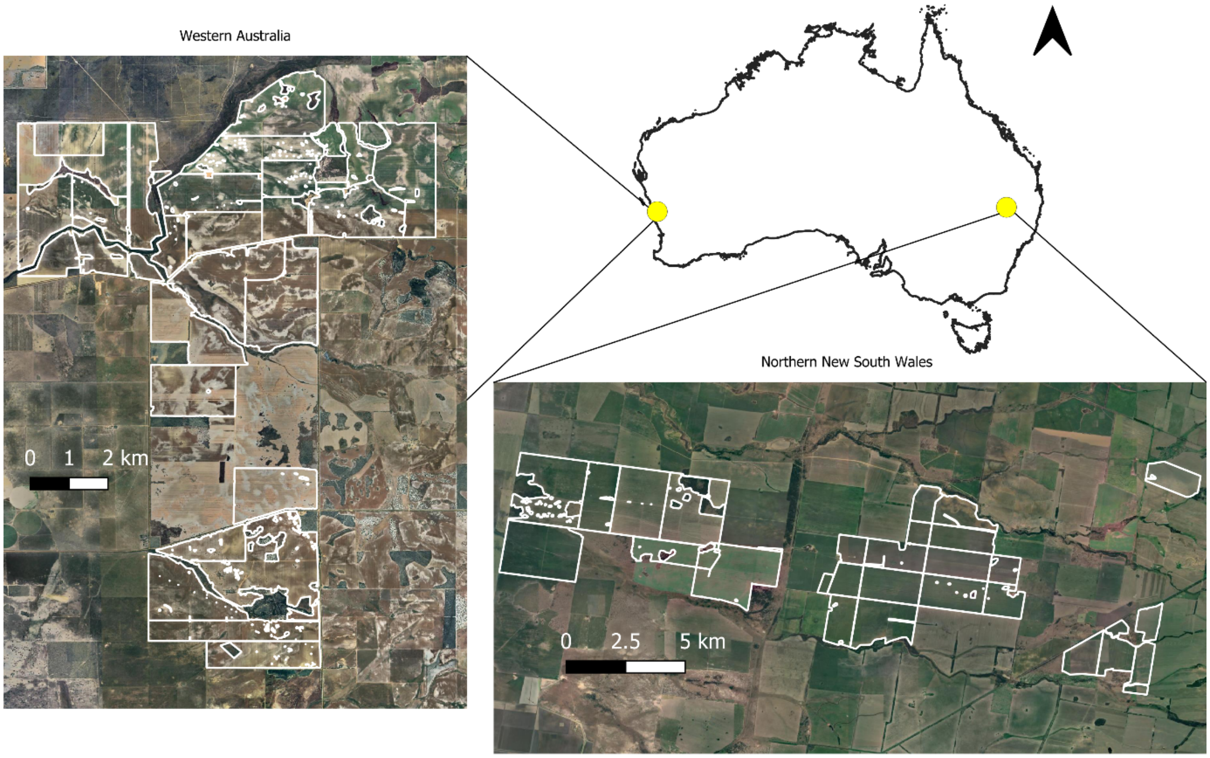

Figure 2. Map of farm aggregations in Western Australia and northern New South Wales, Australia.

Experiments

We now have access to vast amounts of on-farm and publicly-available data that can be used to represent variability and the factors that drive GPC. Different combinations of yield, agronomic and/or publicly-available data layers were used with machine learning (Random Forest) models to predict and map grain protein content (Table 1). All yield and agronomic data was accessed via Precision Cropping Technology (PCT) AgCloud. All publicly-available data layers were accessed via the R package *'dataharvester'* (Harianto *et al.* 2023), and are available for every field and farm across Australia. Modelling was performed for five Experiments using different combinations of yield, agronomic, and publicly-available data (Table 1):

1) Experiment 1: Yield + Agronomic + Publicly-available
2) Experiment 2: Publicly-available
3) Experiment 3: Agronomic + Publicly-available
4) Experiment 4: Yield + Publicly-available
5) Experiment 5: Yield

Table 1. On-farm and publicly available data layers for modelling grain protein content using machine learning (Random Forest) models.

Data	Source	Data category		Data layers
Yield	PCT AgCloud	Yield		
Agronomic	PCT AgCloud	Field data		Sowing Date Harvest Date Variety
		Cropping History		1 season prior 2 seasons prior 3 seasons prior
Publicly-available	*dataharvester*	Remote sensing Sentinel 2A, 10 m spatial resolution	Current season maximum	Normalised Difference Vegetation Index (NDVI) Normalised Difference Red Edge (NDRE) Enhanced Vegetation Index (EVI)
			Long-term averages	EVI: 1, 5, 10 year averages NDVI and Red band: 5th, 50th, 95th percentiles
			Bare earth Imagery	
		Terrain attributes	Radiometrics	Dose Rate, Thorium, Uranium, Potassium
			Digital Elevation Model	

Validation approaches

Grain protein sensors may not be available across all fields, farms, or for every season. In some cases, entire fields may not have maps of GPC. In other cases across large fields where multiple headers are operating (e.g. in Australia or the United States of America), only one header may be equipped with a grain protein sensor. This leaves information gaps across parts-of for an entire field. These two scenarios were tested using two validation approaches (Figure 3):

1) A leave one FieldYear-out cross validation (LOFYOCV, Figure 3 a) method, which was used to simulate cases where grain protein sensor data was not available for an entire field; or
2) A two-fold cross validation (2FCV, Figure 3 b) method was used to simulate cases where only one header is equipped with a grain protein sensor, and GPC data is only available for part (half) of a field.

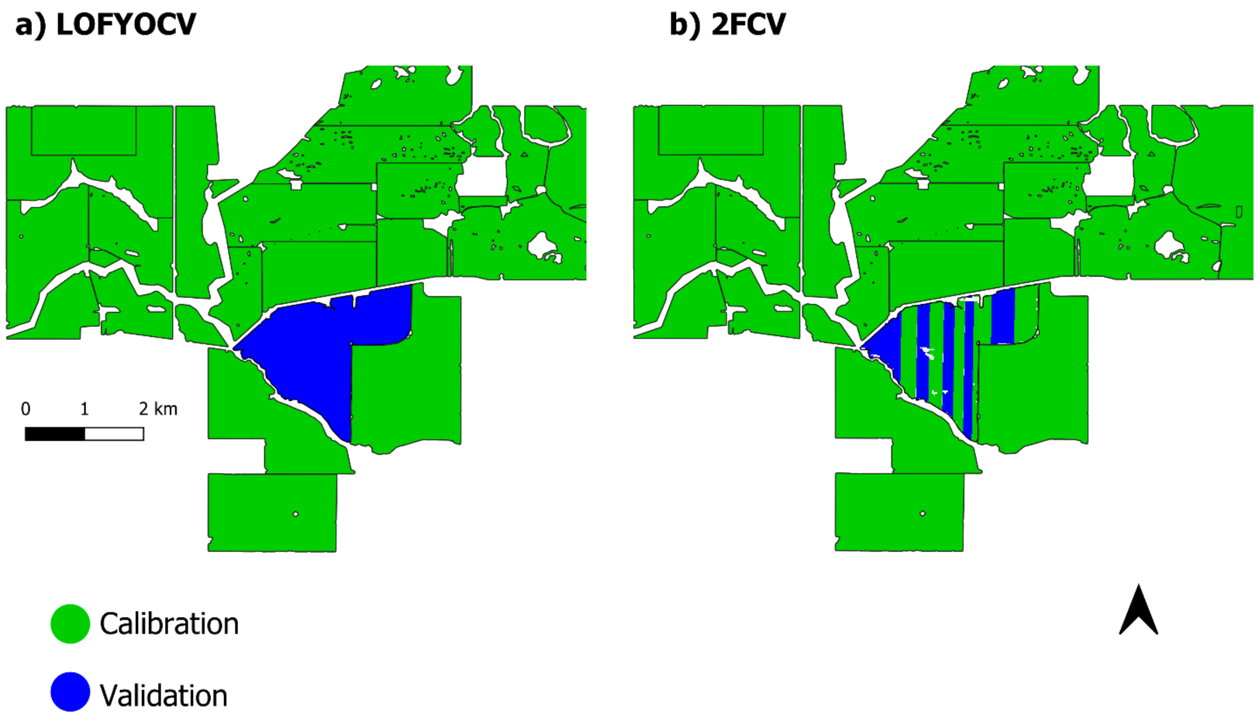

Figure 3. Leave-one-FieldYear-out cross-validation (LOFYOCV, a) and two-fold cross-validation (2FCV, b) approaches used in this study.

The LOFYOCV method involved removing all data for one FieldYear combination, using this as the validation dataset, and leaving the remaining FieldYears as the calibration dataset to build the Random Forest model (Figure 3 a). This LOFYOCV method was repeated for all FieldYear combinations (i.e. a total of 22 for the NSW aggregation, and 41 times for the WA aggregation). The 2FCV method was used to simulate instances when GPC data is only available for part of a field (e.g. there are two headers and only one header is equipped with a GPC sensor). In the 2FCV approach, each field was split into 10 equal areas based on Longitude and allocated alternating as either “even” or “odd”. The “even” allocated data for each FieldYear was removed and used as the validation dataset. The “odd” portion of the current FieldYear and all other FieldYears across the aggregation were used as the calibration dataset to build the RF model (Figure 3 b). This 2FCV method was repeated for both the “even” and “odd” halves of all FieldYear combinations (i.e. a total of 44 times for NSW, and 82 times for WA).

Validation scale

Predictive models for GPC were then built and grain protein maps were produced for each Experiment and Validation approach. Predictions were made at a fine (30 m) resolution, and were also aggregated to management zones within each field to reduce noise and provide maps of GPC that are more informative for management decisions such as for N removal and prescription maps. Yield data for the current season was used to aggregate each field into five management zones using k-means clustering.

Moving window correlations

Local correlations between wheat grain yield and protein content in a 100 m moving window were mapped within each field to investigate their relationship. A 100 m moving window was used to balance sample size with computational efficiency, and a 100 m buffer was used around the field boundary to limit potential edge effects.

Results and discussion

Model quality and performance

The model quality was assessed by calculating the root-mean-square error (RMSE) and the Lin's Concordance Correlation Coefficient (LCCC). The RMSE represents the accuracy of the predictions (how close the predictions are to the true values) and provides a measure of prediction accuracy in the variables units. The LCCC is a measure of both the precision (how close the predictions are to each other) and the accuracy of the predictions. The LCCC value explains the fit of the observed and predicted values to a 1:1 line, where values of 0 are a poor fit (poor agreement between observed and predicted values) and 1 for a perfect fit (perfect agreement between observed and predicted values). The LCCC is unitless and is useful for comparing the precision and accuracy of predictions between variables of different magnitudes (Lin 1989).

Model quality statistics (RMSE and LCCC) for each Experiment, Validation approach, and Validation scale in both WA and NSW are presented in Figure 4. Overall, Experiment 1 (yield + agronomic + publicly-available) performed best for both WA and NSW, indicating that a combination of both on-farm and publicly-available data layers are needed to produce the best quality predictions of GPC. As the best performer, only results for Experiment 1 are presented in Figures 5 and 6. By far, Experiment 5 (yield) performed the worst overall, as expected.

Model quality improved for all Experiments in both NSW and WA when predictions were aggregated to management zones compared to validation at a fine-resolution. Unsurprisingly, the 2FCV approach performed better than the LOFYOCV approach overall. Model quality was generally better for WA than for NSW, which may be attributed to the higher variability in GPC for the NSW aggregation (coefficient of variation, CV = 0.18) compared to WA (CV = 0.14).

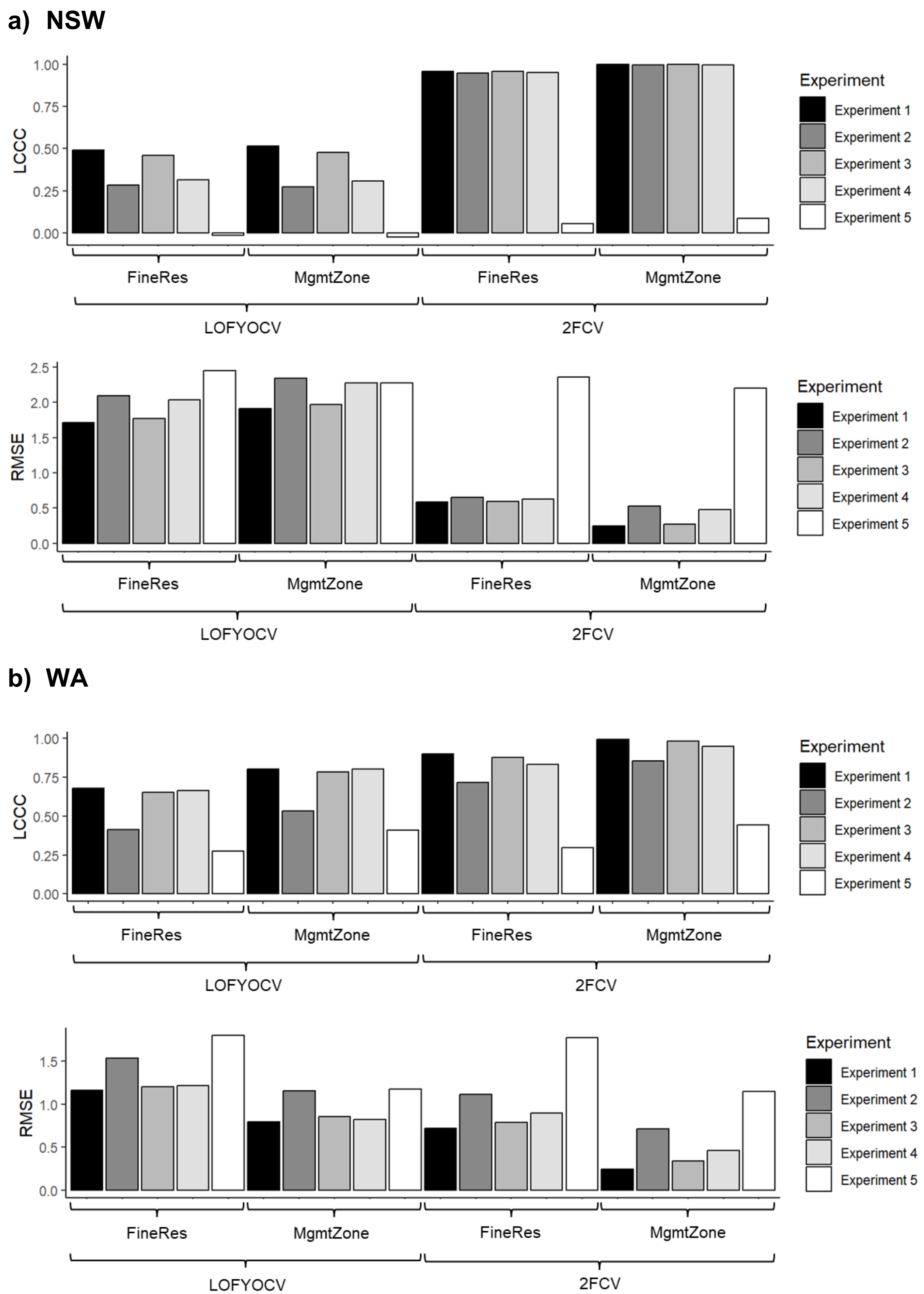

Figure 4. Lins Concordance Correlation Coefficient (LCCC) and Root Mean Square Error (RMSE) values for a) northern New South Wales (NSW) and b) Western Australia (WA) aggregations for five Experiments, validated at a Fine-resolution (FineRes) and aggregated to management zones (MgmtZone) using a leave-one -FieldYear-out cross-validation (LOFYOCV) and two-fold cross-validation (2FCV) approach.

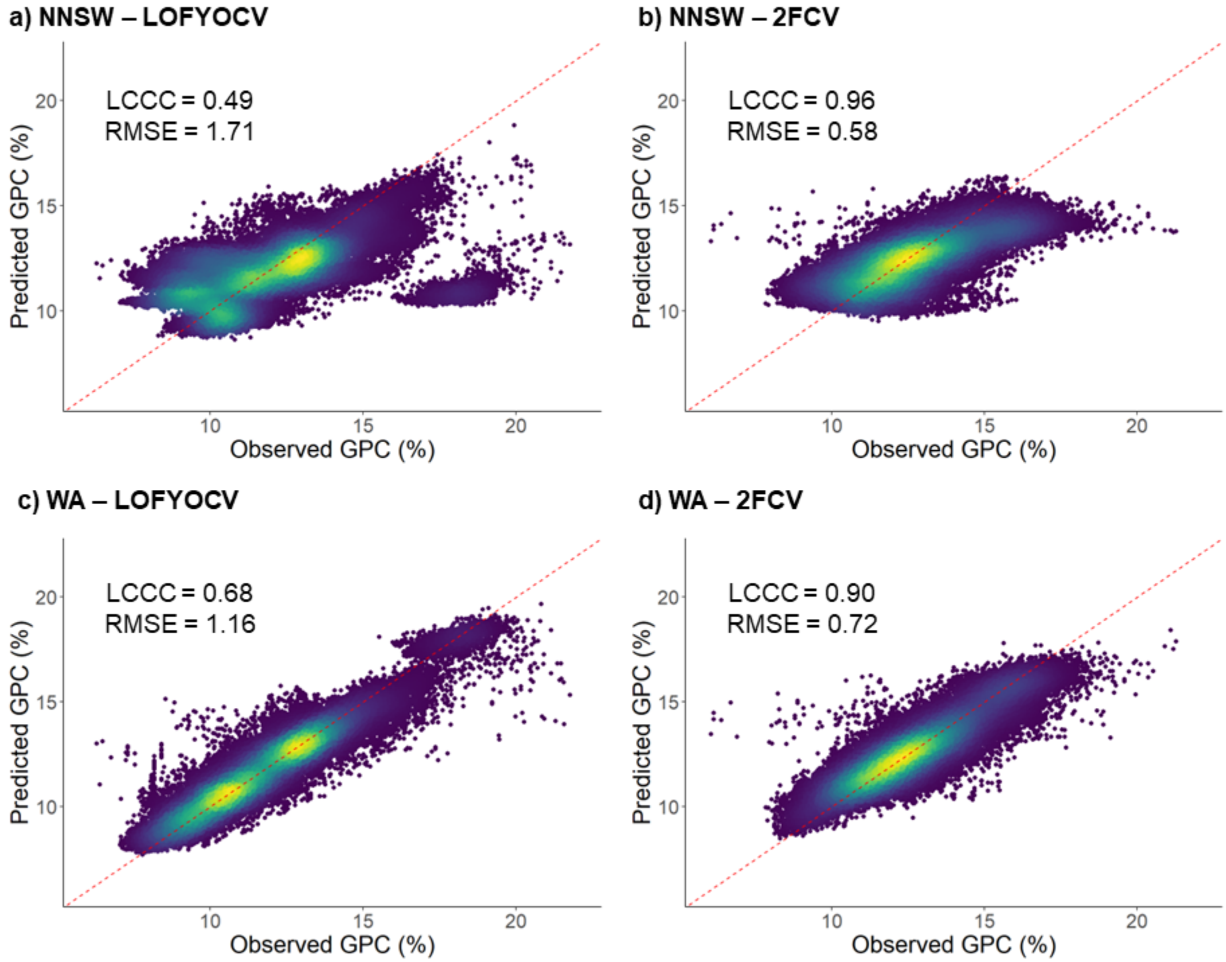

Figure 5. Observed and predicted values of grain protein content (GPC) from Random Forest models for Experiment 1 (Yield + Agronomic + Publicly-available) for northern New South Wales (NSW, a and b) and Western Australia (WA, c and d) aggregations. Models were validated using leave-one-FieldYear-out cross validation (LOFYOCV, a and c) and two-fold cross-validation (2FCV, b and d). Lins Concordance Correlation Coefficient (LCCC) and Root Mean Square Error (RMSE) values are presented for each aggregation and validation approach.

While fine-resolution maps of grain protein provide a high degree of detail describing the spatial variability of GPC, these may be difficult to use to make operational decisions. When implementing precision agriculture (PA) practices, it is common to divide a field into management zones. Aggregating GPC predictions into management zones can smooth small-scale noise and may be useful for informing management decisions such as N prescription maps.

Figure 6 shows a comparison of observed and predicted GPC values for two fields in WA at a fine-resolution and aggregated to five management zones (Figure 6 g) based on yield data for the current season. Predicted values are presented for the LOFYOCV (Figure 6 b and e) and 2FCV (Figure 6 c and f) methods.

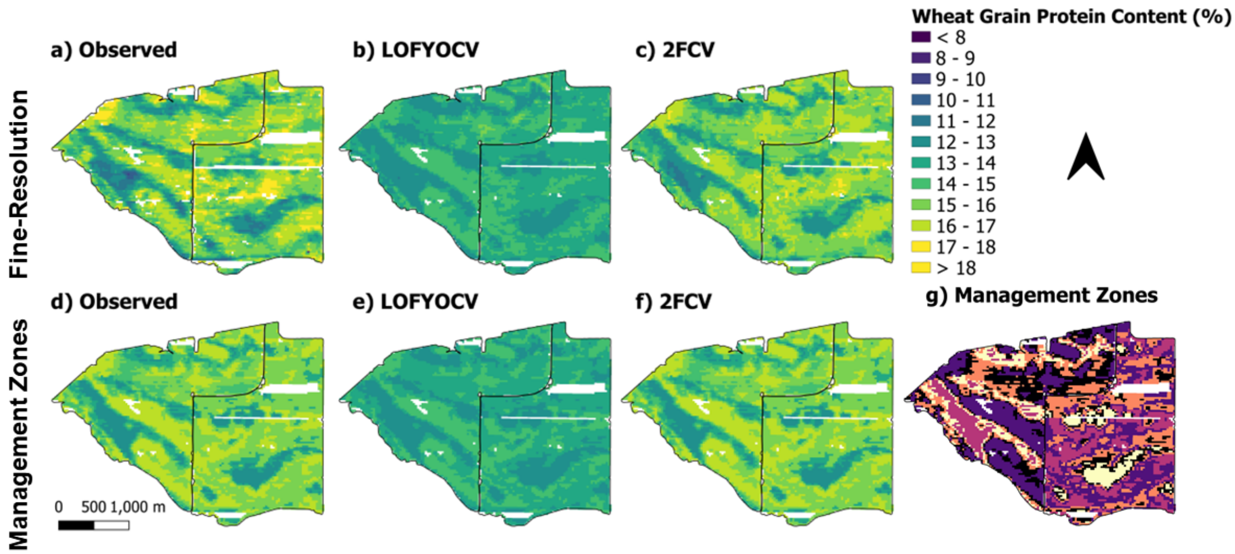

Figure 6. Observed and predicted values of wheat grain protein content (%) at a fine-resolution and aggregated to five management zones (g) for two fields in Western Australia. Predicted values are presented for the leave-one-FieldYear-out cross-validation (LOFYOCV, b and e) and two-fold cross-validation (2FCV, c and f).

Validation approaches

Compared to random K-fold cross validation strategies, which involve randomly splitting observations into K-equal random training and test subsets without assuming any spatial structure in the data, spatial cross-validation methods such as LOFYOCV can provide a more reliable measure of model prediction accuracy, and provide more realistic measures for predicting unknown locations (Christy 2008; Stevens *et al.* 2012; Filippi *et al.* 2020; Habibi *et al.* 2024). As expected, ignoring the spatial dependence of the data can underestimate model prediction errors (Ruß and Brenning 2010; Ferraciolli *et al.* 2019) and result in poor predictions in unknown locations. Despite the poorer validation statistics of the LOFYOCV approach compared to the 2FCV in this analysis, it is evident that the spatially-aware (Habibi *et al.* 2024) LOFYOCV method is a robust validation strategy that can provide a more realistic and reliable assessment of model performance for predicting GPC on unseen fields.

Model quality was better when the 2FCV method was used compared to the LOFYOCV method, as expected. The 2FCV approach involved splitting each FieldYear into two alternating halves based on Longitude. While the 2FCV method is not entirely spatially independent, as FieldYear-level agronomic information such as variety, sowing, and harvest dates for the current season are common across the “odd” and “even” halves of each FieldYear, this is not undesirable in the context of utilizing existing grain protein data to fill-in knowledge gaps within fields with incomplete grain protein maps, and the spatial structure in the dataset is still considered by splitting the training and test datasets via spatial blocks based on Longitude. Incorporating valuable field-specific information describing seasonal-interactions between grain protein and environmental (e.g. rainfall or temperature), soil conditions (e.g. constraints or moisture), or management implications (e.g. variety choice, fertiliser application) is important for capturing and describing variability in GPC and for improving model predictive performance to fill-in knowledge gaps within fields. This demonstrates that predictions of grain protein may be improved if at least some harvest data within a field is collected for the current season.

As highlighted by Smith *et al.* (2023) when implementing a similar leave-one-group-out cross-validation approach for developing transferrable remote-sensing pasture estimates using a group of experimental plots, the 2FCV method implemented here may be more an interpolative evaluation rather than extrapolative as the training and test halves of each FieldYear are located closely to each other and the spatial domain of the model is similar to where predictions are being made. What is missing from this analysis, however, is interpolation solely within a field. The 2FCV approach can also be used to fill in gaps within a field by using data from the “odd” half of a field to predict the “even” half, thus using data from just that one field only. This is operationally more useful for a single season when harvest data is incomplete. Future work should consider if we can interpolate for missing data within a single field, which may include the incorporation of kriging.

Experiments and available data

While the uptake of grain protein sensors is increasing, it is unlikely that we will see a map of GPC for every field, farm, or season in the near future. Here, we highlight the potential to use existing yield, agronomic and publicly-available data layers to model and map GPC to fill-in previously unmapped areas of a farm. Publicly-available data layers were chosen to represent the factors that drive variability in GPC, meaning that bespoke soil samples or Electromagnetic (EM) surveys, for example, are not required for individual fields and growers should not be burdened with additional data collection. Further, this approach did not aim to produce a bespoke model for every field, and instead one model was built for each aggregation. For the LOFYOCV approach, the addition or more fields and seasons worth of data within an aggregation should improve model performance by capturing a greater range of growing conditions. If several seasons of yield and protein data can be collected which represent a range of environmental conditions and management scenarios, it is likely that we will be able to map previous seasons worth of GPC data to better understand long-term trends, or make forecasts for the current season. For the 2FCV approach, within field interpolation should be explored in future when mapping grain protein within a single field for a single season when harvest data is incomplete.

Yield-Protein relationship

Overall, model quality for both the WA and NSW aggregations was moderate-to-good, but it is still unclear what is driving this variability. The factors driving grain protein content predictions within models will be examined in future research, but seasonal fluctuations in environmental conditions and management decisions may influence predictions between fields and seasons. High-yielding, high-protein grain may be desirable in some markets, but grain yield and protein are understood to be negatively correlated (Terman *et al.* 1969). This inverse relationship is generally the result of grain protein dilution by total carbohydrates, which is predominately driven by soil moisture and N availability. In non-limiting soil moisture situations, increasing the soil N supply will typically increase grain yield, whereas increasing the N supply where soil moisture is severely limited will typically increase grain protein (Whelan *et al.* 2009). Generally, high yield/low protein at harvest may be the result of sub-optimal N management, whereas low yield/high protein may be the result of a lack of soil moisture supply and a dry finish (Scott 2022). Other factors such as variety, environmental conditions, and soil constraints may also influence the grain yield/protein relationship.

Local correlations between wheat grain yield and protein content in a 100 m moving window were mapped within each field to investigate their relationship (Figure 7). While an inverse yield-protein relationship was expected and was observed across the entire dataset overall, moving window yield-protein correlation maps showed considerable variation within fields. The strength and direction of the relationship between yield and protein was highly variable both within and between fields, farms, and seasons, and Spearman-Rank correlations within the moving windows ranged between – 0.94 and + 0.94. Future work is needed to better understand this relationship, including better understanding how and why this dynamic varies within and between fields, farms, and seasons, and subsequent management options to optimize this dynamic.

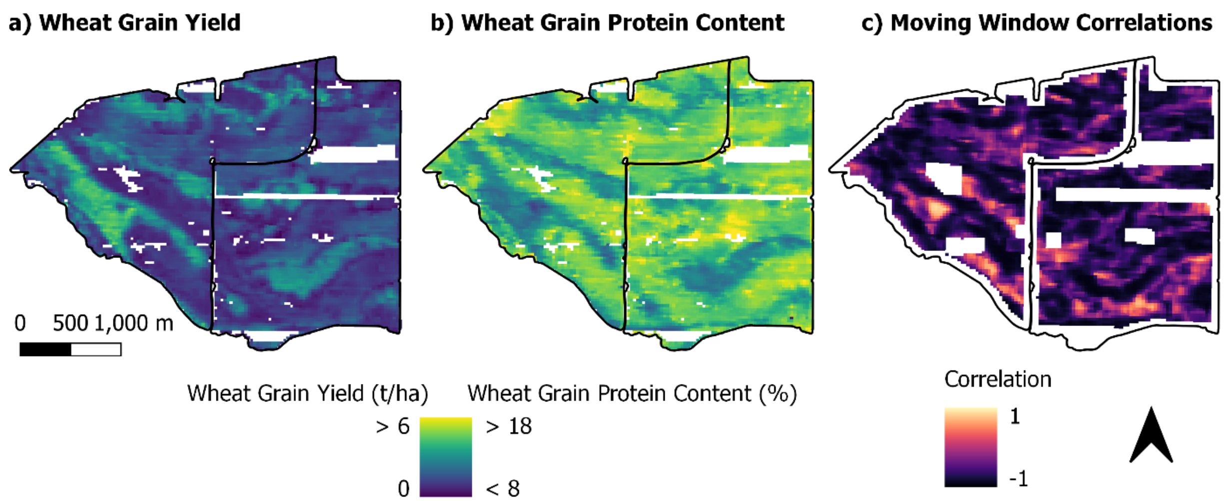

Figure 7. Observed Wheat grain yield (t/ha; a) and Protein content (%; b) values for two fields in WA, and yield-protein correlations in a 150 m moving window across each field (c). Values closer to -1 (black) indicate a negative relationship between yield and protein, whereas values closer to 1 (white) indicate a positive relationship between yield and protein.

Future directions

Future work aims to investigate yield-protein relationships further through the use of interpretive machine learning models and additional data layers. Typically, machine learning models like Random Forest models are considered a "black box", where it can be difficult to understand what factors are driving predictions within the model. Interpretive machine learning can be used to overcome this limitation. Interpretive machine learning refers to a collection of techniques developed to identify the importance of individual predictors in a model and determine what was used to make a prediction (Jones *et al.* 2022). Interpretive machine learning has been used to identify the causes of crop yield variability in cotton (Jones *et al.* 2022), where digital soil maps and terrain information was used to map cotton lint yield and interpretive machine learning was then used to identify the contribution of each predictor variable to the modelled yield prediction. Interpretive machine learning can be used to understand what may be driving variations in GPC and what may explain these changing relationships between yield and protein within and between fields, farms, and seasons. By identifying the contribution of each variable to modelled grain protein predictions, we can then map the major drivers of grain protein content within a field and across farms. Applying this over multiple seasons may also help to identify any seasonal fluctuations or changes in the drivers of grain protein over time.

Conclusion

In the absence of grain protein sensor data, a combination of yield, agronomic and publicly-available data layers can be used to build a predictive model to map GPC. In this study, for each of the five Experiments tested, Random Forest models for GPC were validated using either a LOFYOCV or 2FCV approach at a fine (30 m) resolution or aggregated to management zones. Model performance was moderate-to-good overall and a combination of yield, agronomic and publicly-available data layers produced the best quality predictions of GPC. As expected, model quality improved when predictions were validated using 2FCV compared to LOFYOCV, and performance also generally improved when predictions were aggregated from a fine-resolution to management zones. Spatial correlations between GPC and yield varied spatially and season-to-season, and these correlations fluctuated in both strength and distribution across a field. The LOFYOCV and 2FCV methods utilised here provide a robust and more realistic assessment of model performance for predicting GPC on unseen fields. These spatial validation strategies also highlight the potential to fill gaps and predict GPC across fields with little-or-no protein sensor data using readily available data layers. Future work will investigate the use of interpretive machine learning to better understand the drivers of variability in wheat GPC, and the yield-protein relationship between fields, farms, and seasons.

Acknowledgments

We would like to thank the growers for providing access to the data and for being partners in this research, and Precision Cropping Technologies (PCT) for their collaboration in this research and facilitating access to the yield and protein datasets. We would also like to thank John Deere for their ongoing collaborations. We also acknowledge the support of the Sydney Informatics Hub, a Core Research Facility of The University of Sydney, and the Agricultural Research Federation (AgReFed) for their development of the data harvester.

References

Christy CD (2008) Real-time measurement of soil attributes using on-the-go near infrared reflectance spectroscopy. *Computers and electronics in agriculture* **61**, 10–19.

Ferraciolli MA, Bocca FF, Rodrigues LHA (2019) Neglecting spatial autocorrelation causes underestimation of the error of sugarcane yield models. *Computers and Electronics in Agriculture* **161**, 233–240. doi:10.1016/j.compag.2018.09.003.

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. doi:10.1016/j.agsy.2020.102894.

Habibi LN, Matsui T, Tanaka TST (2024) Critical evaluation of the effects of a cross-validation strategy and machine learning optimization on the prediction accuracy and transferability of a soybean yield prediction model using UAV-based remote sensing. *Journal of Agriculture and Food Research* **16**, 101096. doi:10.1016/j.jafr.2024.101096.

Harianto J, Haan S, Butterworth N (2023) dataharvester: Download and Process Geospatial Data. https://sydney-informatics-hub.github.io/dataharvester/.

Jones EJ, Bishop TFA, Malone BP, Hulme PJ, Whelan BM, Filippi P (2022) Identifying causes of crop yield variability with interpretive machine learning. *Computers and Electronics in Agriculture* **192**, 106632. doi:10.1016/j.compag.2021.106632.

Lin L.-K (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255–268. doi:10.2307/2532051.

Ruß G, Brenning A (2010) Data mining in precision agriculture: Management of spatial information. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **6178 LNAI**, 350–359. doi:10.1007/978-3-642-14049-5_36.

Scott E (2022) Protein mapping - getting more bang for your fertiliser buck. In 'GRDC 2022 Grains Res. Updat. – South',

Smith HD, Dubeux JCB, Zare A, Wilson CH (2023) Assessing Transferability of Remote Sensing Pasture Estimates Using Multiple Machine Learning Algorithms and Evaluation Structures. *Remote Sensing* **15**,. doi:10.3390/rs15112940.

Stevens A, Miralles I, van Wesemael B (2012) Soil Organic Carbon Predictions by Airborne Imaging Spectroscopy: Comparing Cross-Validation and Validation. *Soil Science Society of America Journal* **76**, 2174–2183. doi:10.2136/sssaj2012.0054.

Terman GL, Ramig RE, Dreier AF, Olson RA (1969) Yield-Protein Relationships in Wheat Grain, as Affected by Nitrogen and Water 1 . *Agronomy Journal* **61**, 755–759. doi:10.2134/agronj1969.00021962006100050031x.

Whelan B (2019) On-the-go protein sensors. In 'GRDC 2019 Grains Res. Updat. – Pallamallawa.',

Whelan BM, Taylor J (2013) 'Precision agriculture for grain production systems.' doi:10.1080/17538947.2013.817183.

Whelan BM, Taylor JA, Hassall JA (2009) Site-specific variation in wheat grain protein concentration and wheat grain yield measured on an Australian farm using harvester-mounted on-the-go sensors. *Crop and Pasture Science* **60**, 808–817. doi:10.1071/CP08343.