Spatial predictive modelling for quantifying soybean seed quality using remote sensing and machine learning.

Hernandez C[1], Correndo A[2], Kyveryga P[3], Prestholt A[4], Ciampitti I[1]

[1] Department of Agronomy, Kansas State University, USA.
[2] Sustainable Cropping Systems, University of Guelph, Canadá.
[3] Data Analytics, John Deere.
[4] Soil and Water Outcomes Fund.

Full paper: https://doi.org/10.1016/j.compag.2023.108096

Introduction

Soybean (Glycine max L.) is a key field crop with one of the largest globally planted areas. While yield and productivity have traditionally driven crop expansion and remain a primary focus of many studies, there has been a growing emphasis on seed composition—specifically protein and oil content—in recent years. Soybean meal serves as a crucial protein source for both animal feed and human consumption. From an oil perspective, soybean crops are significant for biofuels, particularly biodiesel and polymer production. Thus, protein and oil are the most vital components of soybean seed quality. Unfortunately, the quality of soybean seeds has been compromised by rising yield trends, with protein concentration notably decreasing over time, which reduces the nutritional value of soybean meal. Consequently, improving the management and prediction of soybean protein and oil concentrations at the field scale is essential to capture the added value in soybean production. Additionally, a recent study found that farmers would focus on quality if offered a small economic incentive, approximately US$ 18 per Mg, to change their approach. A successful framework for shifting the current mindset from quantity-focused production to seed composition should not only provide information but also encourage this change, primarily through the adoption of technology for quality segregation during harvest. Available technologies such as off-combine sensors and remote sensing can help monitor crop seed quality and predict within-field variability before harvest. On the other hand, at-harvest crop seed quality sensing is more accurate (on-combine sensors), but this technology has yet to be widely adopted in crops beyond cereals. Therefore, developing digital solutions to aid soybean farmers in decision-making is critical to advancing the complex process of seed quality segregation in the field. The evolving data processing systems, combined with remote sensor information, are powerful tools for creating agricultural data products. Although there is substantial research on the spatial prediction of soybean yields and the workflows to achieve this, a significant research gap exists in mapping soybean seed quality components before harvest, which could enable on-farm seed segmentation. Machine learning algorithms are becoming increasingly popular across various fields due to their ability to detect subtle patterns in complex datasets. Several supervised learning techniques have shown promise in predicting multiple crop traits. Recent advancements suggest that combining machine learning algorithms through ensemble methods could yield more robust predictions. Regardless of the model, identifying a minimum set of candidate predictors early in the cropping season is desirable.

Materials and methods

The study area encompassed the central regions of Kansas and Iowa in the United States (US). A total of 804 soybean seed samples were collected and georeferenced to evaluate protein and oil concentrations across 47 soybean fields in Iowa and Kansas during 2019, 2020, and 2021. Between 9 and 15 samples per field were collected to capture the spatial variability within the fields. Yield data was not collected along with the seed samples. Both protein and oil concentrations were measured as percentages (%).

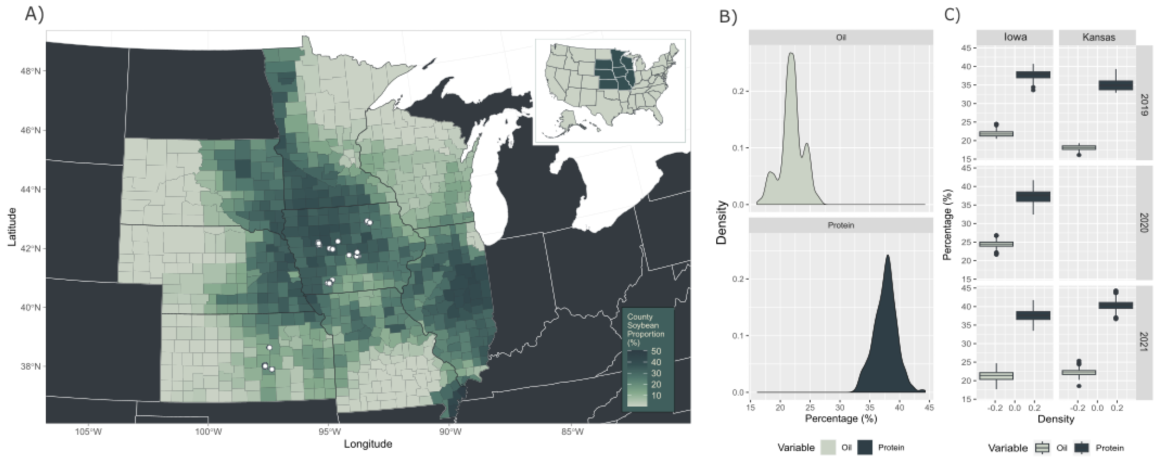

Figure 1 A) Soybean seed samples (white points) located in Kansas in Iowa US states, showing the spatial distribution of proportion of soybean planted in the northwest of US Midwest. B) Soybean protein and oil concentration distribution of the soybean samples analyzed. C) Variability across States and Years of the soybean seed quality parameters.

Remote sensing data was sourced from the Sentinel-2 MSI satellite platform, which offers 13 spectral bands with spatial resolutions of 10, 20, and 60 meters and a temporal revisit time of 5 days.

For each field, spectral bands were obtained from May 1st to October 31st, covering the crop growing season. A total of 80 spectral indices were created using Sentinel-2 MSI spectral bands and map algebra. These indices represented the most used metrics for mapping chlorophyll content, water content, leaf area index (LAI), biomass, and vegetation coverage.

To distinguish full-season from double-cropped soybeans, a time series segmentation was performed using the Time-Series K-means algorithm. The resulting crop type classification (double or full season crop) was added to the dataset as a new categorical variable. For all locations, the Green Chlorophyll Vegetation Index (GCVI) was used during the cropping season due to its strong relationship with the dynamics of LAI in soybean crops. The periods when soybean crops reach maximum LAI values are around the R3-R5 growth stages. The GCVI peak was identified when the first derivative of the double logistic-sigmoid function was equal to zero. After pinpointing the peak, data from all fields were standardized using the peak moment as day 0 or the reference time.

A time frame selection process was carried out to determine the best period for identifying optimal predictors. An iterative incremental algorithm was applied with a Partial Least Square Regression (PLSR) model using all bands, indices, and the crop type as inputs. The PLSR was conducted with window sizes ranging from 5 to 30 days, starting from -40 to +40 days relative to the GCVI peak. The Root Mean Square Error (RMSE) prediction was calculated for each period and used to select the best time frame. All combinations of time and window size were evaluated to identify the period and duration that minimized prediction error, with those combinations having the lowest RMSE being selected. This framework was applied separately for protein and oil concentration models. The response variables were oil and protein concentration, while predictors included integrated bands, spectral indices from the selected best time frame, and a binary variable for crop type (double or single crop).

In the best period identified by the PLSR for applying machine learning models, six models were tested: i) ElasticNet, ii) Random Forest, iii) XGBoost, iv) LightGBM, v) CatBoost, and vi) an ensemble model averaging (i) to (v). To enhance the robustness of this framework, nested cross-validation was implemented in two steps: an inner-loop for parameter optimization and an outer-loop for testing generalization performance. The outer loop consisted of a 70:30 split, with 70% of the fields used for training (n = 30) and 30% for testing (n = 17). In the inner-loop, hyper-parameter optimization was performed using repeated (x3) 10-fold cross-validation. A grid search was conducted for each model to find the best hyper-parameter combinations, which were selected based on the average performance on the inner validation set using RMSE minimization. After optimizing hyper-parameters, predictive performance was assessed using the entire outer-training datasets to predict observations in the outer-testing sets.

To evaluate predictive performance, eight error metrics and indices were used: the coefficient of determination (R²), the concordance correlation coefficient (CCC), the root mean square error (RMSE), the normalized or relative RMSE (RRMSE), the mean bias error (MBE), the segregation of the mean square error into the Percentage Lack of Accuracy (PLA) and the Percentage Lack of Precision (PLP), and the Kling-Gupta efficiency (KGE). For the best machine learning model, the relative importance (%) of each predictor variable was assessed using a permutation-based test, which estimates the relative increase in prediction error (mean squared error) due to the exclusion of a specific variable from the model.

Results

Protein concentration ranged from a minimum of 32.4% to a maximum of 44.3%, with an average of 37.6%. Meanwhile, oil concentration varied from a minimum of 16.1% to a maximum of 26.8%, with an average of 22.0%. The variation in these seed quality traits by crop type, specifically full-season versus double-cropped soybeans, was analyzed to understand the influence of this variable in our dataset. For full-season soybeans, seed protein concentration had an interquartile range (IQR) of 2.27% and a median of 37.78%, whereas for double-cropped soybeans, the IQR was broader at 5.32% with a median of 36.10%. Regarding seed oil concentration, full-season crops exhibited an IQR of 2.06% with a median of 22.16%, while double-cropped soybeans had an IQR of 4.19% with a median of 18.70%. More detailed information can be found in Figure 1 B) and C).

The lowest RMSE values for protein prediction were observed between 2 to 7 days after the GCVI peak. Conversely, the optimal period for predicting oil concentration was from 1 to 7 days after the GCVI peak. Furthermore, the minimum RMSE values were 1.75% for protein concentration and 1.29% for oil concentration. The crop growth curve described by the GCVI pattern can be found in Figure 2.

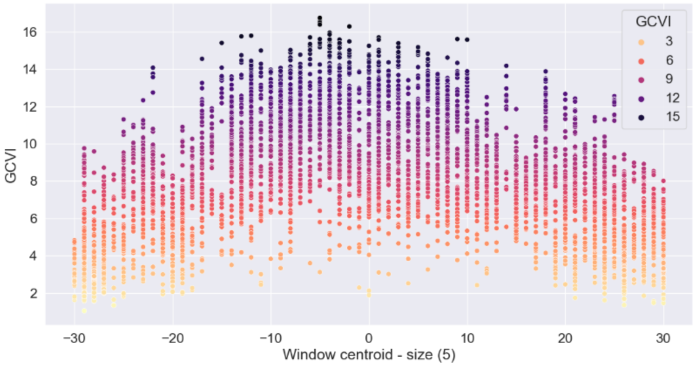

Figure 2: Normalization of the values of the GCVI values centered according to peak date.

For predicting protein and oil concentrations, XGBoost was identified as the best model, demonstrating the highest CCC, lowest RMSE and RRMSE, lowest PLA, and highest KGE. Contrary to expectations, the ensemble of models did not enhance the performance indicators for either protein or oil. Overall, ElasticNet exhibited the poorest predictive performance for both variables. For both protein and oil concentrations, all ML models displayed limitations, commonly over-predicting below-average values and under-predicting above-average values (approximately 38% for protein and 33% for oil). An observed vs predicted scatter plot for the ensemble model and a deployment with application of the trained model can be found in Figure 3.

The most influential variable for the XGBoost model in predicting seed protein concentration was the crop type, distinguishing between full-season and double-cropped soybeans (with +30% feature importance). For oil concentration, XGBoost identified several key variables, with the top three being crop type, TSAVI, and B11, collectively contributing to +15% feature importance.

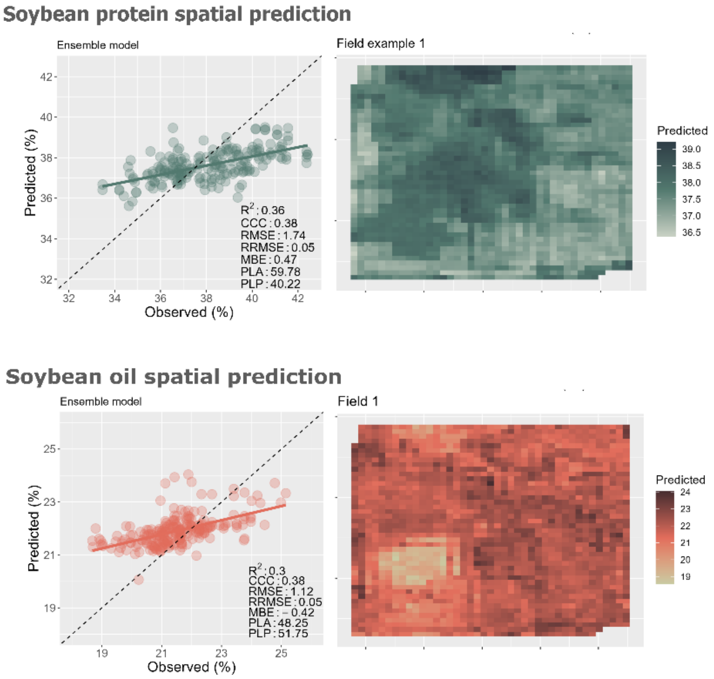

Figure 3: Predicted vs Observed values for soybean seed protein and oil concentration (Left) for the ensemble model. Spatial prediction in a field as an example for both soybean protein and oil concentration.

Conclusion

Based on this, employing predictive models via data analysis and remote sensing might offer a promising solution, aiding in the creation of valuable and scalable digital agricultural tools aimed at quantifying crop quality. Future research should focus on i) exploring new data fusion methods, ii) incorporating seasonal metrics of crops, soil, and climate to create more sophisticated predictions, and iii) investigating the primary factors driving spatial variation in seed quality traits to inform management strategies.

This is a shortened version of the final document published in Hernandez, C. M., Correndo, A., Kyveryga, P., Prestholt, A., & Ciampitti, I. A. (2023). On-farm soybean seed protein and oil prediction using satellite data. Computers and Electronics in Agriculture, 212, 108096.