Unveiling advanced data-informed predictive approaches for characterizing spatial variability in agriculture.

by

Carlos Manuel Hernandez

B.S., Universidad Nacional de Rio Cuarto, 2019

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Agronomy
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2025

Approved by:

Major Professor
Dr. Ignacio Ciampitti

Copyright

Abstract

Emerging technologies such as spatial statistics, machine learning, and remote sensing have been shown to play a crucial role in digital agriculture, offering promising tools that enhance data-driven decision-making and contribute to the sustainability of farming systems. However, despite the growing volume of data collected each season, information on yield, environmental conditions (e.g., soil data), and crop variables are often incomplete or becomes too costly to obtain. In response to these challenges, this dissertation explores the development of predictive models that can support agricultural applications.

The study is structured into three chapters. The first chapter introduces the expanding capabilities of digital technologies for data collection and predictive modeling, highlighting their potential for developing digital agricultural products at an on-farm level. The second chapter presents the design and implementation of a robust framework for generating spatial predictions of soybean seed quality parameters at the farm level. Recognizing that on-farm data often requires preprocessing before it can be effectively used in predictive modeling, the third chapter proposes a hierarchical Bayesian approach to address spatial misalignment issues. This method is designed to address the issue of missing values that are a result of the merging of datasets from disparate geographic locations. This approach facilitates the process of both data imputation and the integration of continuous spatial information with prior knowledge.

Overall, this dissertation employs a combination of statistical techniques, machine learning algorithms, and spatial data management strategies to advance the design and implementation of agricultural predictive models. These models facilitate the spatial characterization of a broad range of agriculturally relevant variables, thereby contributing to more precise and informed decision-making in the field.

Table of Contents

List of Figures

List of Tables

Acknowledgements

The design and development of this document, as well as the projects presented herein, are the result of the interaction, support, and contributions of numerous individuals who, in various ways, made this work possible.

I would like to express my sincere appreciation to Dr. Ignacio, who served not only as my supervisor but also as a mentor. His trust, encouragement, and unwavering support were instrumental in the development of this work. His guidance opened doors and constantly challenged me to achieve excellence.

I am also deeply grateful to Dr. Vara Prasad and Dr. Trevor Hefley, members of my evaluation committee, for their thoughtful advice and guidance throughout the course of my studies.

I would like to sincerely thank the Flickner family, not only for generously sharing us knowledge and insights about their farming systems, but also for their remarkable kindness and commitment to supporting the Kansas community.

My sincere thanks to Adrian, Victor, Luiz, Andres, Leo, Josefina, Luciana, Ana, Francisco and Federico, students and postdoctoral researchers who contributed significantly to the success of our projects. Their commitment, insights, and collaboration were essential to reaching our research objectives and ensuring the quality of our results.

I would also like to extend my heartfelt gratitude to my closest friends in Manhattan—Nicolas, Hilario, Sofia, Valentina, Diego, Ignacio, Emmanuela, Lucas and Juan Francisco—whose support and companionship were vital throughout this journey. They became a surrogate family during my time away from home. In the same spirit, I thank Agustin, German, Renzo, and Agustin, my dearest friends in Argentina. Though not family by blood, they have been a source

of strength and solidarity over the past years, always standing by my side without expecting anything in return.

Lastly, I am thankful to all the interns and visiting students who were part of our team between 2021 and 2025. Their contributions, enthusiasm, and dedication played an important role in making this work possible.

To all of them, I express my heartfelt appreciation for their support, dedication, and invaluable contributions throughout this journey.

Dedication

To my parents, Daniel and Nancy, and my brothers Luciano, Franco and Gonzalo.

Chapter 1 - Introduction

Agricultural systems are poised to face significant challenges in the coming years (Fao, 2019). These challenges will not only impact farmers, but also all stakeholders in the agricultural supply chain and its related sectors (Sharma et al., 2021). In this context, sustainable agricultural systems are expected to face several major challenges (El Jarroudi et al., 2024). These challenges are not only linked to the growing demand for greater efficiency in agriculture (Muhie, 2022), but also to the increasing risk of production disruptions and changes in consumer behavior (Raymond et al., 2022; White et al., 2019).

Addressing these critical bottlenecks in future agricultural production will require a profound transformation of cropping systems, considering not only agronomic aspects but also all other factors that could constrain the agricultural supply chain (Rockström et al., 2017). Emerging digital tools could serve as a cornerstone in tackling these issues (Parra-López et al., 2024). Data generated from various sources, including satellites, sensors, and agricultural machinery, can be leveraged to overcome these challenges by supporting decision-making, enhancing knowledge exchange, and improving process monitoring (Fuentes-Peñailillo et al., 2024). A deeper understanding of spatial variability has the potential to not only boost field productivity and resource use efficiency of inputs but also facilitate access to better markets and mitigate production risks (Hernandez et al., 2023; Zhang et al., 2002).

In this respect, the digital data collection process, integrated with on-farm experiments, plays an important role by expanding the scale and capacity to address agricultural challenges (Bramley et al., 2022; Lacoste et al., 2021). While the concept of digital agriculture has been traditionally defined in terms of data-driven farming (O’Donoghue et al., 2024), it can also be redefined as a paradigm in which agronomical science incorporates innovative analytical methods and higher-quality data to enhance decision-making (Basso & Antle, 2020). Emerging technologies such as spatial statistics, machine learning, and advancements in remote sensing have demonstrated their essential role in digital agriculture, providing promising tools that improve agriculture data-driven decision-making (Araújo et al., 2023; Mathenge et al., 2022).

On one hand, the development of predictive models presents an opportunity to estimate complex or costly-to-measure variables using easily obtainable and low-cost data (Ben Ayed & Hanana, 2021; Elbasi et al., 2023), facilitating large-scale implementation. On the other hand, despite the increasing volume of data collected each season—including yield collection, environmental conditions, and crop quality parameters—data gaps remain a significant challenge due to high acquisition costs or incomplete records. Developing models capable of filling these data gaps while simultaneously quantifying estimation errors is a crucial step for ensuring accurate model evaluation (Barbedo, 2022).

In response to these challenges, this thesis explores various approaches for developing predictive models to support agricultural applications. This study is structured into three chapters. The first chapter provides a general introduction to the expanding capabilities of digital technologies for data collection and predictive modeling, emphasizing their potential in developing at on-farm level digital agricultural products. The second chapter presents as the main goal to design and implement a robust framework for generating spatial remotely sensed derived predictions of soybean seed quality parameters at the field level (before harvest time), acknowledging that farm-level data often require preprocessing before they can be effectively utilized in predictive modeling. The third chapter introduces as the main goal the utilization of a Hierarchical Bayesian approach to address spatial misalignment issues, mainly focused on managing missing values within a spatial context when integrating different data sources facilitating both data imputation (quantification of uncertainty) and the incorporation of continuous spatial information with prior knowledge.

Overall, this dissertation employs a combination of statistical techniques, machine learning algorithms, and spatial data management strategies to enhance the design and implementation of predictive agricultural models. These models enable the spatial characterization of a wide range of agriculturally relevant variables, contributing to more precise and informed decision-making for farmers, agronomists, and consultants working at field-scale.

Chapter 2 - On-farm soybean seed protein and oil prediction using satellite data

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.

Abstract

Soybean (*Glycine max* L.) seed composition is receiving increased attention among farmers, agronomists, and commodity traders. Increasing the ability to predict seed quality traits such as protein and oil at the field level before harvest will provide a competitive ability to segregate quality and create an economic advantage to position the production at both domestic and global markets. Therefore, this study aims to use remote sensing satellite data to spatially predict soybean seed protein and oil concentrations at the field level before harvest time. The dataset consisted of 47 fields located in Kansas and Iowa, United States, from the 2019 to 2021 seasons. Six machine-learning approaches (ElasticNet, Random Forest, XGBoost, LightGBM, CatBoost, and an ensemble) were tested evaluating different vegetation indices and spectral bands to predict before harvest seed protein and oil concentration from satellite imagery. The optimal timing for training prediction models was identified within a week after the peak of the green chlorophyll vegetation index, with different spectral indices and bands of importance for each seed quality component. The XGBoost outperformed the rest of the algorithms for both quality traits. Overall, models reported an absolute error of 1.80 % for protein and 1.04 % for oil concentration. Our research describes a pipeline that combines on-farm data, open access satellite imagery, an intensive use of spectral bands, and machine learning to forecast seed

quality before harvest. Future research guiding crop management interventions should be directed to i) integrating and assessing major drivers of spatial variation of seed quality traits such as soil and weather data, and ii) exploring satellite data-fusion approaches and with larger moving towards deep learning methods.

Introduction

During the last two decades, the soybean crop harvested area registered a consistent rising trend of 1.78 M Ha. Yr-1 (Food and Agriculture Organization of the United Nations, 2023). While production is still a major driver for crop expansion and many more studies are focused on the estimation of yield and productivity, an increasing focus on seed composition, herein referring to protein and oil, has been documented during the last years (Borja Reis et al., 2022). Soybean meal is the main supply of animal and human protein. Managers and nutritionists use soybean meals in feed formulations to add proteins of high quality to the animal diet (Ibáñez et al., 2020). On the other hand, a wide variety of fermented and non-fermented high-protein products can be produced from soybeans for human consumption (Wijewardana et al., 2019). From an oil standpoint, soybean crops can play a key role in biofuel production (Sheehan et al., 1998). In summary, both protein and oil constitute the most important components of soybean seed quality. Improving overall management and prediction at field-scale for soybean protein and oil concentrations is critical to potentially capture the added value in soybean production. In addition, a recent study (Borja Reis et al., 2022) reported that farmers will manage for quality if a small economic incentive, roughly US$ 18 per Mg, is provided to produce a change in mindset. A successful framework for changing the current mindset of production only based on quantities and shifting the focus to seed composition should not only provide information but also facilitate this change (Moss, 2019), mainly via technology adoption for quality segregation during harvest time.

Available technologies to monitor crop seed quality such as off-combine sensors and remote sensing (Song et al., 2017) may facilitate segregating seed quality at the farm by predicting within-field variability before harvest (Bastos et al., 2021). On the other end, at harvest crops, seed quality sensing is more precise (on-combine sensors) (Taylor & Whelan, 2007), but the technology has yet to be adopted and implemented in other crops beyond cereals. Therefore, the development of digital solutions to assist soybean farmers in making decisions (Correndo, McArtor, et al., 2022) is critical to move forward with this complex seed quality segregation process at farmer fields.

The rising evolution of data processing systems combined with information obtained from remote sensors is a powerful tool in the creation of data products for agriculture. The extraction of crop features from the earth's surface through satellite data has been widely explored. The diversity of applications ranges from application in scouting and crop monitoring (Bajwa et al., 2017), crop segmentation (Zhong et al., 2016), and management areas delineation (Georgi et al., 2018), Leaf Area Index (LAI) (Nandan et al., 2022), crop phenology (Diao, 2020), biomass (Reisi Gahrouei et al., 2020), crop height (Alonso-Gonzalez et al., 2015). Additionally, a considerable number of publications focus on the spatial prediction of soybean crop yields and the workflows to achieve it (Schwalbert et al., 2020). Nonetheless, there is a considerable research gap in terms of mapping soybean seed quality components before harvest, which could open opportunities for the on-farm segmentation of seed. Machine learning algorithms are gaining popularity across multiple disciplines due to their capacity of reading hard-to-perceive patterns from complex datasets (Abraham et al., 2020). Various supervised learning techniques have demonstrated promising performance for predicting multiple crop traits (Ramanantenasoa et al., 2019). Recent advances suggest the combination of machine learning algorithms through ensemble techniques may allow the achievement of more robust predictions (Shahhosseini et al., 2021). Regardless of the model, a minimum set of candidate predictors as early as possible during the cropping season is desirable (Correndo et al., 2021).

Following this rationale, this study aims to deploy a pipeline to predict the on-farm spatial variability of soybean seed protein and oil concentrations before harvest using temporal satellite data. The specific goals of this study are to identify: (i) the optimal timing (temporal) during the crop growing season, and (ii) the optimal spectral data to develop a prediction of soybean seed protein and oil concentration on farmer fields (training dataset) before harvest time.

Materials and Methods

Database

The study area comprised the central region in the states of Kansas and Iowa, United States (US), respectively (Figure 1). Precipitation and temperature levels (May 1st – September 30th) across site years are described in Table S1.

Soybean seed samples were collected right before harvest from a total of 707 geo-referenced points from 47 commercial fields (Figure 1) (15 in Kansas, 32 in Iowa) in 2019 (n=19), 2020 (n=4), and 2021 (n=24). A total of 9-to-15 distributed samples per field were collected attempting to capture field spatial variability. No yield data was collected along with the seed samples. Seed samples were determined via near-infrared spectroscopy (NIR) using the Perten DA7200 Feed Analyzer (Perten Instruments, Stockholm, Sweden). Both protein and oil concentrations were expressed as mass percentages (%). A general summary of the descriptive statistics of soybean seed samples by field is presented in Table S2.

Remote sensing data were obtained from the Sentinel-2 MSI satellite platform. Sentinel-2 MSI provides 13 spectral bands at 10, 20, and 60 m of spatial resolution with a temporal revisit time of 5 days (Gatti & Bertolini, 2013). For each field, the spectral bands were retrieved from May 1 to October 31. Spatial coordinates were used to extract the values from every spectral band using a buffer area of 20 by 20 m for each georeferenced point. To filter those images with the presence of clouds that could influence the values of bands and indices, two types of filters at the selection level were applied. While the first type of filter contemplates the percentage of cloud occupancy in the entire tile, the second type of filter uses the QA mask, filtering out those dates where each pixel in the image was classified as cirrus or clouds. All satellite data were selected and retrieved using Google Earth Engine (GEE) through geemap (Q. Wu, 2020) and GEE Python packages (Gorelick et al., 2017).

Feature Engineering

Spectral Indices

A total of 80 spectral indices were created using Sentinel-2 MSI spectral bands and map algebra (Table S3). Spectral indices created represented the most used indices to map chlorophyll content, water content, leaf area index (LAI), biomass, and vegetation coverage.

Soybean crop type clustering

A time series segmentation (L. Zhong et al., 2016) was applied to segregate different patterns in the Green Chlorophyll Vegetation Index (GCVI) signature, which could indicate differences in crop management. Part of the crop management oriented to crop rotation, such as crops planting in full season or soybean planted as a double-cropped (e.g., after winter cereal crop), can produce differences in the duration of the season (Santos Hansel et al., 2019). Changes in the length of the soybean season could affect the main seed quality components under investigation (Ray et al., 2008). Regarding crop type segmentation, the signature values obtained among samples collected within a field were averaged for the same date to obtain a single average GCVI time series per field. To improve the temporal resolution of GCVI, a resampling was applied to a scale of three days and a linear interpolation between two consecutive dates. In addition, a time series smoothing was carried out through a Savitzky-Golay filter with a window size of 7 days. Subsequently, a time series k-means algorithm was applied to select 2 clusters, one for the curves representing the full-season crops, and the other one referring to the double crops. The results of the last process were added to the dataset as a new synthetic categorical variable defining the crop type as a double- or full-season crop (Fig 3A). The time-Series K-means algorithm was implemented through the tslearn package (Tavenard et al., 2020) for Python language.

Temporal normalization

To establish a common time reference for spectral indices, a peak detection analysis was performed (Fig. 3B). For all locations, we used the GCVI signature during the cropping season due to its close relationship with the dynamics of the LAI in soybean crops (Lobell et al., 2015). Based on the linear relationship that exists between LAI vs. GCVI (Cai et al., 2018), the GCVI peak values are expected to coincide with the maximum LAI values. In addition, based on previous publications (Sakamoto et al., 2010), the periods in which soybean crops reach the maximum LAI values would be around the R3-R5 growth stages. Furthermore, the period from Soybean flowering (R1; Fehr et al., 1971) to full maturity (R8; Fehr et al., 1971) is critical for seed yield and quality determination. Lastly, the GCVI evolution during the season was modeled for each site using a double logistic sigmoid function according to Eq. (1) using the *lmfit* package in Python (Newville et al., 2019).

$$GCVI(t) = m_1 + (m_2 - m_7 t)\left(\frac{1}{1+e^{\frac{m_3-t}{m_4}}} - \frac{1}{1+e^{\frac{m_5-t}{m_6}}}\right) \tag{1}$$

Where GCVI(t) is GCVI evolution fitted, it is time in day of year, m1, m2, m3, m4, m5, m6, and m7 are the parameters of the function.

The GCVI peak was obtained when the first derivate of the double logistic sigmoid function was zero (Fig. 4). After identifying the peak, data from all fields were standardized using the peak's moment as day 0 or reference time.

Time-frame selection

A time frame selection was carried out to select the best period to make predictions. 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 implemented considering a window size between a minimum of 5 days to a maximum of 30 days, starting from -40 to +40 days with respect to the GCVI peak. In this step, a nested cross-validation was performed, where an outer-loop consisted of a random train-test split (70:30 for train and test, respectively); and an inner-loop consisted of a 10-fold structure. The inner loop served to fine-tune the number of components in the PLSR as the main hyper-parameter; while the outer-loop served to test the performance on unseen observations. The criterion to select the best model was based on the minimization of the root mean square error (RMSE). The RMSE of prediction was obtained in each period listed and used as an error measurement to select the best time frame. All the combinations between time and window size were obtained, to know at what moment of the growth curve and during how many days we can make the predictions by minimizing the error and selecting those combinations with the lowest RMSE. This framework was applied separately for protein and oil concentration models.

Predictive Modeling

The response variables were oil and protein concentration. Predictors consisted of integrated bands, spectral indices (Table S3) from the best time frame selected in Section 2.2.4., plus a binary variable representing the crop type (double or single crop) obtained in the feature engineering process. Six machine-learning models were tested. The candidate models were: i) ElasticNet (Zou & Hastie, 2005), ii) Random Forest (Breiman, 2001), iii) XGBoost (Chen & Guestrin, 2016), iv) LightGBM (Ke et al., 2017), v) CatBoost (Prokhorenkova et al., 2018), and vi) an ensemble using the mean of models (i) to (v).

Nested cross-validation was implemented in two steps: (i) an inner loop for the optimization of hyper-parameters, and (ii) an outer loop for testing the generalization performance. The latter consisted of a 70:30 split, where 70% and 30% of entire fields were used for training (n = 30) and testing (n = 17) the models, respectively. This split also considered maintaining a balanced sample of fields across years and states. At the inner loop, the hyper-parameters optimization was conducted using a repeated (x3) 10-fold cross-validation. We performed a grid search for each model to fine-tune the hyper-parameters (Table S4). The best hyperparameter combinations for each algorithm were selected based on average performance on the inner validation set using the root mean square error (RMSE) minimization criterion. More details about the hyperparameters tuning are shown in Table S4.

Field Predictions Performance

Once hyper-parameters were optimized, predictive performance was assessed using the entire training datasets to predict the observations on the outer-testing sets (hold-out) For a comprehensive predictive performance evaluation, a total of eight complementary error metrics and indices were used: i) the classical coefficient of determination (R^2) that exclusively represents a measure of precision (not accuracy); ii) the concordance correlation coefficient (CCC, higher is better) as a normalized metric that weighs the correlation coefficient (precision) by an index of accuracy; iii) the root mean square error (RMSE, % lower is better) as an average squared errors-based statistic that Kling et al., (2012). Specific formulae and descriptions of the metrics can be found in Table S5.

Feature importance

For the best ML model, the relative importance (%) of each predictor variable on the overall performance of the best model was assessed using a permutation-based test, which estimates the relative increase of prediction error (mean squared error) due to the exclusion of a certain variable from the model (Breiman, 2001).

Software

The ML libraries used for training, testing, and applying the models were *Sklearn* (Pedregosa et al., 2011), *XGBoost* (Chen & Guestrin, 2016), *LightGBM* (Ke et al., 2017), and *CatBoost* (Prokhorenkova et al., 2018) in Python language. Predictive performance metrics were estimated in R software using the *metrica* package (Correndo, Rosso, et al., 2022).

Results

Soybean seed quality components description

Protein concentration ranged between a minimum of 32.4% and a maximum of 44.3 %, with a mean of 37.6 %. On the other hand, oil concentration ranged between a minimum of 16.1% and a maximum of 26.8 %, with a mean of 22.0 %. The general relationship between soybean seed protein and oil concentration is shown in Fig. S1. The variation of these seed quality traits by crop type, full versus double-cropped soybeans, was described to understand the influence of this variable in our database. For full-season soybeans, seed protein concentration presented an IQR of 2.27 % and a median of 37.78 %, while for double-cropped the IQR was wider at 5.32 % with a median of 36.10 %. Lastly, for seed oil concentration, the full-season crops showed an IQR of 2.06 % with a median of 22.16 % and the double-cropped presented an IQR of 4.19 % with a median of 18.70 %.

Time-frame selection

The lowest RMSE values for protein prediction were identified between 2 to 7 days after reaching the GCVI –peak (Fig. 5B). On the other hand, the best time to predict oil concentration was from 1 to 7 days after achieving the GCVI –peak (Fig. 5C). Additionally, the lowest RMSE values were 1.75 % and 1.29 % for protein and oil concentration, respectively.

Predictive performance

For both protein and oil concentration, XGBoost was identified as the best predictive algorithm with the least PLA from the RMSE and high KGE (Table 1). In this sense, it is preferable that the PLP be of a greater value than the PLA, indicating that the error is due to lack of precision and not accuracy, defining this as a model selection criterion. Despite the expectations, the ensemble of models did not improve the performance indicators for either protein or oil. In terms of predictive performance, Random Forest and ElasticNet resulted in the poorest predictive performance for protein and oil concentration, respectively. For both protein and oil concentration, all ML models showed limitations, with a common tendency of over-prediction below- and under-prediction above-average values (~38% of protein, and ~33% of oil) (Fig. 6). The full set of metrics obtained from the performance process using testing dataset are in Table 1. Predictive performance metrics using the training dataset are shown in Table S6. The distribution of both protein and oil concentrations for observed and predicted data are shown in Fig S2.

Feature importance

The most relevant variable for the XGBoost model for seed protein concentration was the crop type (Fig. 7A), separating full relative to double-cropped soybeans (+30% feature importance). For oil concentration, the XGBoost identified several critical variables, with the first three as crop type, TSAVI, and B11 (summing up +15 feature importance) (Fig. 7B). A detailed description of the entire ranking of variable importance can be found in Fig. S3.

Discussion

Our research describes a pipeline that combines on-farm data, open access satellite imagery, an intensive use of spectral bands, and machine learning to forecast seed quality before harvest. The approach developed is based on an iterative incremental algorithm integrating bands and spectral indices together with a binary variable segmenting the type of soybean crop in the rotation. Furthermore, seven new vegetation indices were proposed integrating red-edge and SWIR bands. The analysis here permitted us to identify the most suitable moment and the interval of time to obtain the best predictions for soybean seed quality before harvest time. By identifying the best time during the growing season (crop phenology time) to retrieve the images and the combination of spectral bands (i.e., indices), it was possible to capture the spatial variability for protein and oil concentrations of seed within the field. A recent review synthesis (Bastos et al., 2021) portraying the state-of-the-art and current research progress for protein sensing on field crops, highlights still low progress (2 out of 84 entries) for soybean crops. The latter pinpoints the scarce research progress achieved on remote sensing for characterizing seed quality traits, with limited focus on translating applications from the science to the practical side, farmer fields.

The present research effort clearly aligns with the need to add value to farmers' production. The importance of soybean seed quality and composition is receiving increased attention among farmers, agronomists, and commodity traders. The higher nutritional content of US soybeans provides a competitive edge that can be widely exploited to produce increased economic value through the targeted marketing of each quality level. Within-field knowledge of soybean quality will enable and encourage value chain disruption both up and downstream via more leveraged negotiations. A recent survey of farmers in the Midwestern US regarding their perception of soybean quality (n = 271 responses) revealed that most of them are willing to invest in technology for the characterization of spatial variability of soybean protein at the field level if they could get an additional $0.5/bushel as a premium (Borja Reis et al., 2022). In addition, for the same study, almost all respondents highlighted that with a relatively small economic incentive. Creating awareness and generating relevant data about soybean quality variation in farmer fields will support the goal of increasing the competitiveness of US farmers in both global and regional niche markets.

Our work remarks on the plausibility of using remote sensing to forecast soybean seed quality before harvest. This study could foster the development of decision-making guidelines related to soybean seed quality at different levels of the soybean production chain value. From the industry perspective, this type of model would assist in more efficiently delineating regional soybean markets based on expected quality. From an agronomist standpoint, these models could assist in decision-making not only for improved seed quality but also for more effective fertilization strategies across their fields. Lastly, farmers could use these models to more efficiently segregate soybean seed quality directly on their fields, which implies a significant aid to maximizing the revenue of their production.

Although prediction performance did not result flawlessly, the use of only satellite imagery was demonstrated to be useful enough for identifying within-field spatial patterns of protein and oil concentration. This study also highlights the importance of the distinction between full-season from double-cropped soybeans, the latter being more likely to express a greater level of oil and larger variation for both seed quality traits. The effect of changes in planting date (either early or late timing) are reported in the current literature, but with a non-consistent effect on seed protein concentration (Assefa et al., 2019; Mourtzinis et al., 2018; Salmerón et al., 2017), with a variation on this trait highly dependent on the specific genotype x management x environment conditions. Besides the potential positive impact on soybean export value, further implications may include improved sustainability from the standpoint of nutrient budgets (Bastos et al., 2021) with better planning for the next crop nitrogen fertilization management. In summary, satellite-based models for developing the prediction of seed quality traits as an initial (less complex, without considering other environmental factors) step will assist to provide farmers with a map of the spatial variation on these traits before harvest time.

The proposed models could help farmers for a better understanding of the spatial variability of soybean protein could help to explore and achieve markets with differential payment for quality levels. The models could be used to more efficiently segregate quality soybeans directly into their fields, which would significantly help maximize their production revenue. Future research should expand the domains for the development of these models, which may assist in addressing challenges such as disentangling erratic patterns in the seed composition (Anthony et al., 2012).

From a modeling standpoint, this study offers a relatively simple pipeline to identify suitable timing to retrieve satellite imagery and select the best algorithm for seed quality prediction. Previous research efforts on remote sensing applied to agriculture have been focused on the development of workflows for crop type mapping (Pierre Pott et al., 2022), yield forecasting (Schwalbert et al., 2020), phenology estimation (Gao & Zhang, 2021), and input prescription (Mulla, 2013). However, comparatively less development has been pursued in estimating the spatial variation of soybean seed quality at the farmer field scale (Kravchenko & Bullock, 2002a). Efforts on characterizing seed quality traits via satellite data for other field crops were focused on wheat (*Triticum aestivum* L.) (Xu et al., 2020).

Machine learning models based on satellite imagery represent the most suitable technology in terms of scalability (Basnet et al., 2003). Alternative options for spatially characterizing seed quality variation at the farmer field scale are on-combine (at harvest time), proximal and remote including options ranging from handheld (Oøvergaard et al., 2010), or mounted on unmanned- and manned-aerial vehicles (Hama et al., 2020). The integration of aerial vehicles into crop management and breeding programs can facilitate mapping multiple plant traits including seed quality using hyperspectral imagery (Dilmurat et al., 2022). Yet, compared to satellite imagery, the main limitation of handheld, ground-based, and aerial vehicle sensors is linked to their comparatively low scalability. Moreover, the growing use of daily or semi-daily images with high spectral resolution could represent different alternatives to real-time monitoring of the quality status of a soybean crop, reducing the use of UAV flights (Kimm et al., 2020).

Several aspects are worth highlighting in terms of the development of the models. First, regarding feature engineering, by including a variable that segregates between full season and double-cropped soybean, a considerable improvement in capturing traits variability was achieved (data not shown). This is clearly depicted by the high importance of the ranking of this binary feature of our models. This agrees with the documented seed yield and quality differences based on crop rotation and planting date (Assefa et al., 2019; Houx III et al., 2014). According to the predictive modeling, XGBoost showed the best result compared to ElasticNet, Random Forest, LightGBM, and CatBoost. This could be because XGBoost, like LightGBM and Catboost, are models that use boosting techniques as an optimization algorithm (unlike Random Forest, which uses bagging), which combine different weak learners in sequence; to improve the prediction of the general model. Furthermore, the use of sequential trees and the use of loss functions using derivatives give XGBoost a competitive advantage over the rest. In addition, and like ElasticNet, XGBoost uses a technique called "regularization" that allows different weights to be given to variables according to their importance. Additionally, it is possible that an ensemble through the average of all model outputs is not the best alternative. In this sense, future research should further explore other techniques such as stacking or blending, which allow segmenting the predictions with the different machine learning models. Second, concerning the predictive modeling, we observed that, based on multiple metrics, algorithms implementing boosting (XGBoost, LightGBM, and Catboost) showed better performance compared to other alternatives (ElasticNet, and Random Forest). Bagging (Random Forest) and boosting are both powerful regression trees ensemble-techniques and more effective that non-ensemble algorithms (e.g., ElasticNet), our results reinforce the evidence that combining weak learners in a sequential manner (boosting) outperforms bagging (bootstrapped trees) for both regression and classification problems (Ribeiro & Dos Santos Coelho, 2020; Shahhosseini et al., 2019; Yaman et al., 2021). In our case, XGBoost appeared to be the most suitable among boosting alternatives, possibly because of its tuning options that include regularization, learning rate, among others (Chen & Guestrin, 2016). Finally, it is possible that there are better ensemble alternatives compared to the averaged ensemble we used such as an optimized weighted ensemble (Shahhosseini et al., 2021) or stacking techniques like stacked generalization or blending (Zhou, 2012), which may allow segmenting the predictions with the different machine learning models.

In relation to soybean seed quality, the idea of carrying out differentiated harvest prescriptions at the farm level was stated by previous (Gaso et al., 2023; Kravchenko & Bullock, 2002b; Martin et al., 2007). The first steps to delineate zones within fields should characterize contrasting seed protein and oil concentration levels. For example, Kravchenko & Bullock, (2002a) spatially characterized the variability of seed protein and oil concentrations relative to the topography, but they have not used in-season satellite imagery or machine learning algorithms. Previous research has also explored the relationship between seed quality and soil fertility but have found inconsistent spatial structure for both protein and oil concentration, with patterns differing across years (Anthony et al., 2012).

More recently, the use of remote sensing and predictive models started to dominate the scene. Among others, this approach has been used to characterize genotypes with various levels of quality (Santana et al., 2023), to evaluate the impact of water stress on both soybean seed yield and quality (Tavares et al., 2022). Other publications have focused on defining ideal phenotypes that maximize both seed yield and quality (Roth et al., 2022), selecting genotypes with greater biological nitrogen fixation and seed protein content (Vollmann et al., 2022), or studying potential hyperspectral bands that serve in early selection for breeding programs (Chiozza et al., 2021). Yet, the development of easily scalable models that use remote sensing and machine learning to spatially characterize soybean quality parameters before harvest have not been explored in depth. Previous similar research has made use of small-plots trials and drone imagery, obtaining comparable predictive performance (e.g., RMSE) (Dilmurat et al., 2022; Vollmann et al., 2022). However, due to the nature of our data (i.e., observational on-farm data), the spectral resolution of sentinel compared to images from drones, it could be expected to obtain a reduced decreased predictive performance.

Certainly, the proposed framework is susceptible to improvement. The presence of proportional bias (under- and over-predictions) in the models (Fig. 6A, 6C) could be corrected by addressing the following main limitations: i) a geographical coverage focused on only two of the major soybean producing states (Kansas and Iowa), ii) a single sensor type used (satellite), with the potential of other sources to improve temporal, spatial, and spectral resolution data, iii) the lack of inclusion of environmental characteristics as predictors (soil, topography, and meteorological factors), iii) potentially limited number of observations (9-15) per field to represent actual variability, iv) lack of crop phenology metadata as a reference to refine the use of spectral curves, v) lack of yield data to accompany the seed quality assessment with potential dilution or concentration trade-offs with crop productivity and vi) state of the art deep learning were not tested for this research due the constraints related to the size and non-balanced dataset. Future research investments should direct more attention to i) improving the integration of sensors with new data processing (data fusion approaches), ii) assessing the integration of seasonal crop metrics, soil, and weather variables to improve the prediction of seed quality traits, and iii) assessing the major drivers of field spatial variation of seed quality traits to develop guidelines for field managing interventions related to these plant traits. Concomitantly, future extension efforts should focus on translating predictive models for soybean quality into user-friendly decision-making tools that assist in an enhanced adoption of the technology.

Conclusion

The development of predictive models using satellite imagery (different spectral bands, B-10, 11, TSAVI) and machine learning algorithms resulted in an RMSE of 1.80 and 1.04 in protein and oil concentrations, with the best crop in-season timing centered around a week after the peak of GCVI. This work sets the foundation for developing before harvest or near harvest spatially explicit soybean seed quality maps with implications not only for the domestic market, with potential to increase profitability of farmers via segregation, but also for the international scenario for enhancing the competitiveness of the US soybean value chain. Future steps should focus on i) the integration of satellite with crop and environmental metadata, collecting more observations and exploring deep learning approaches, and ii) the development of a user-friendly web application that assists farmers in the adoption of the proposed approach.

Figures

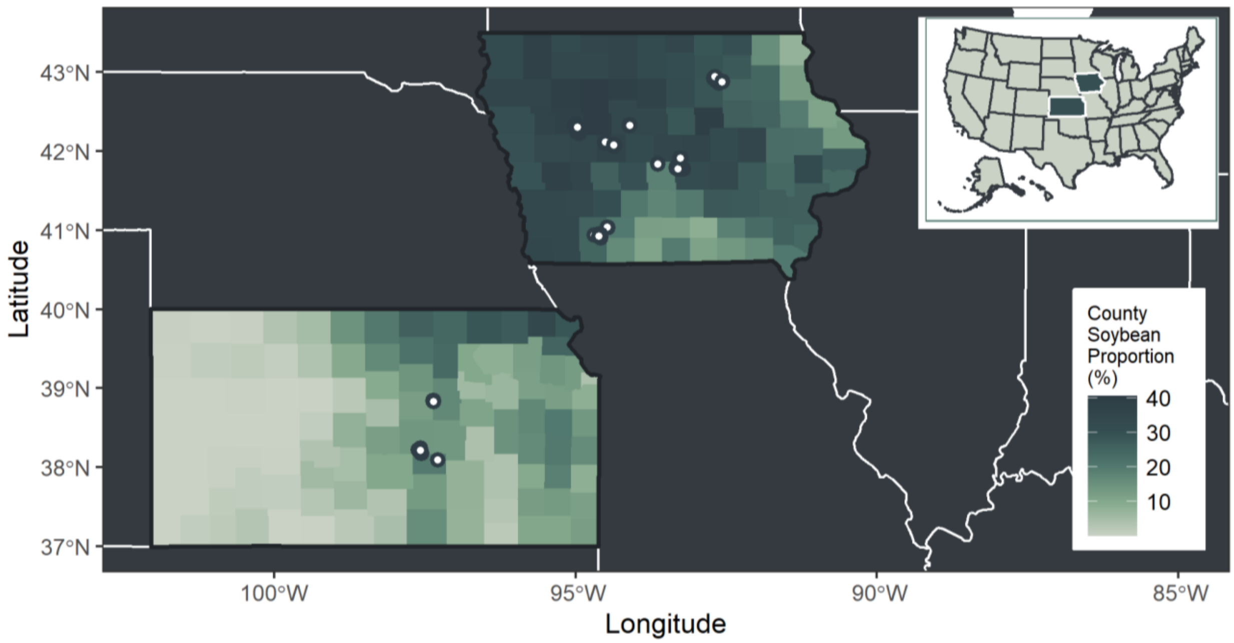

Figure 1. Geographical distribution of soybean seed samples collected from farmer fields (white circles) across the states of Kansas and Iowa, United States.

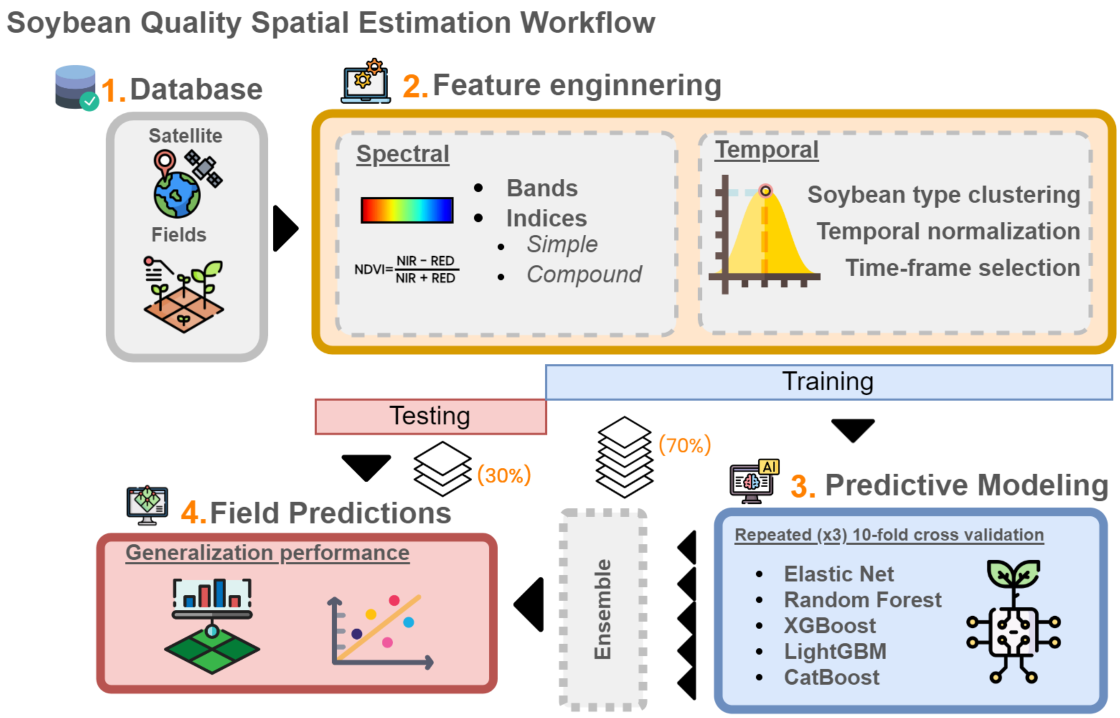

Figure 2. Conceptual workflow of the overall data collection, processing, analysis, and validation for soybean seed quality prediction at field scale, integrating satellite and ground field seed data into machine learning models.

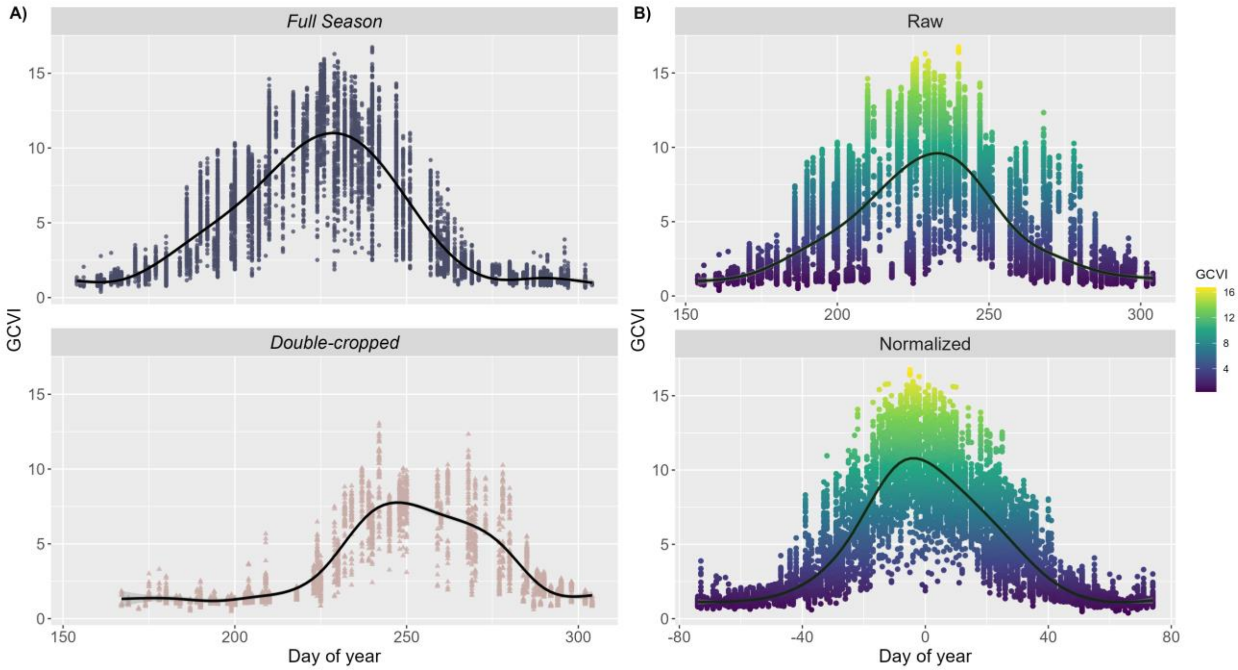

Figure 3. Fig. 3. A) Soybean crop clustering analysis delineating the growing season pattern of two crop types (full season or double-cropped). B) Temporal normalization to establish a common time reference using peak's moment as day 0.

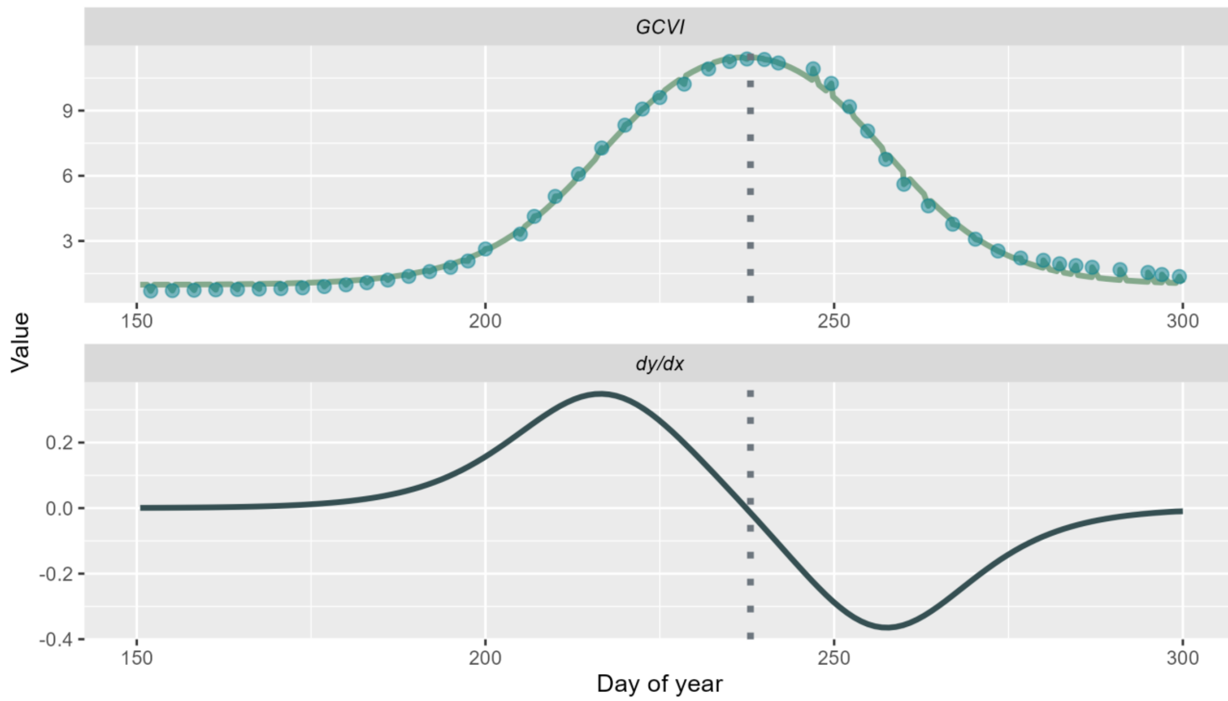

Figure 4. Green Chlorophyll Vegetation Index (GCVI) time series showing the time in where the curve peak is located. The maximum value of GCVI is pointed out when the value of the first derivate of the GCVI times series is equal to zero.

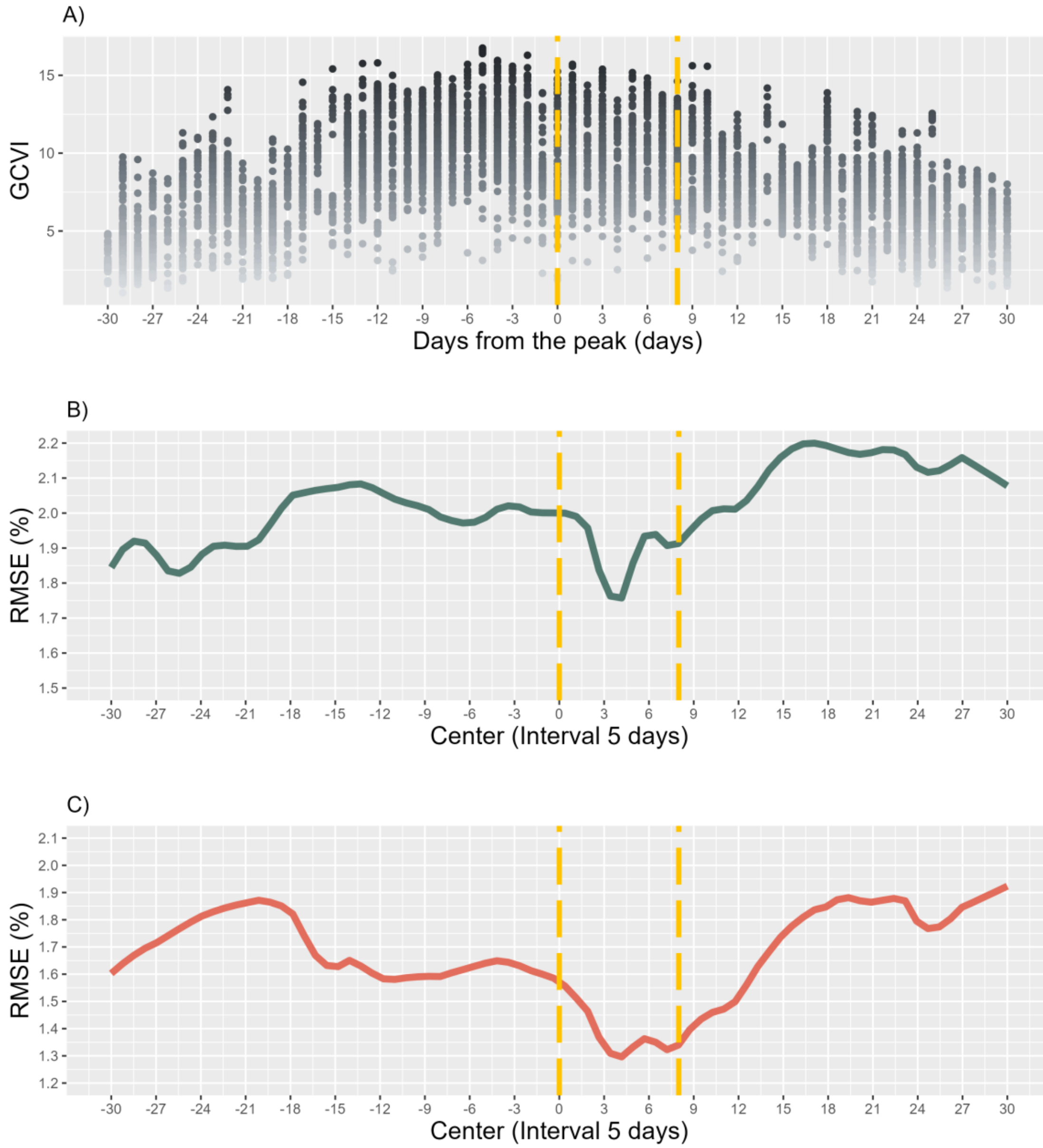

Figure 5. A) Overall Green Chlorophyll Vegetation Index (GCVI) time series normalized using the curve peak as a reference. B) and C) Time frame selection for protein and oil concentration prediction respectively using as a criterion the lowest root mean square

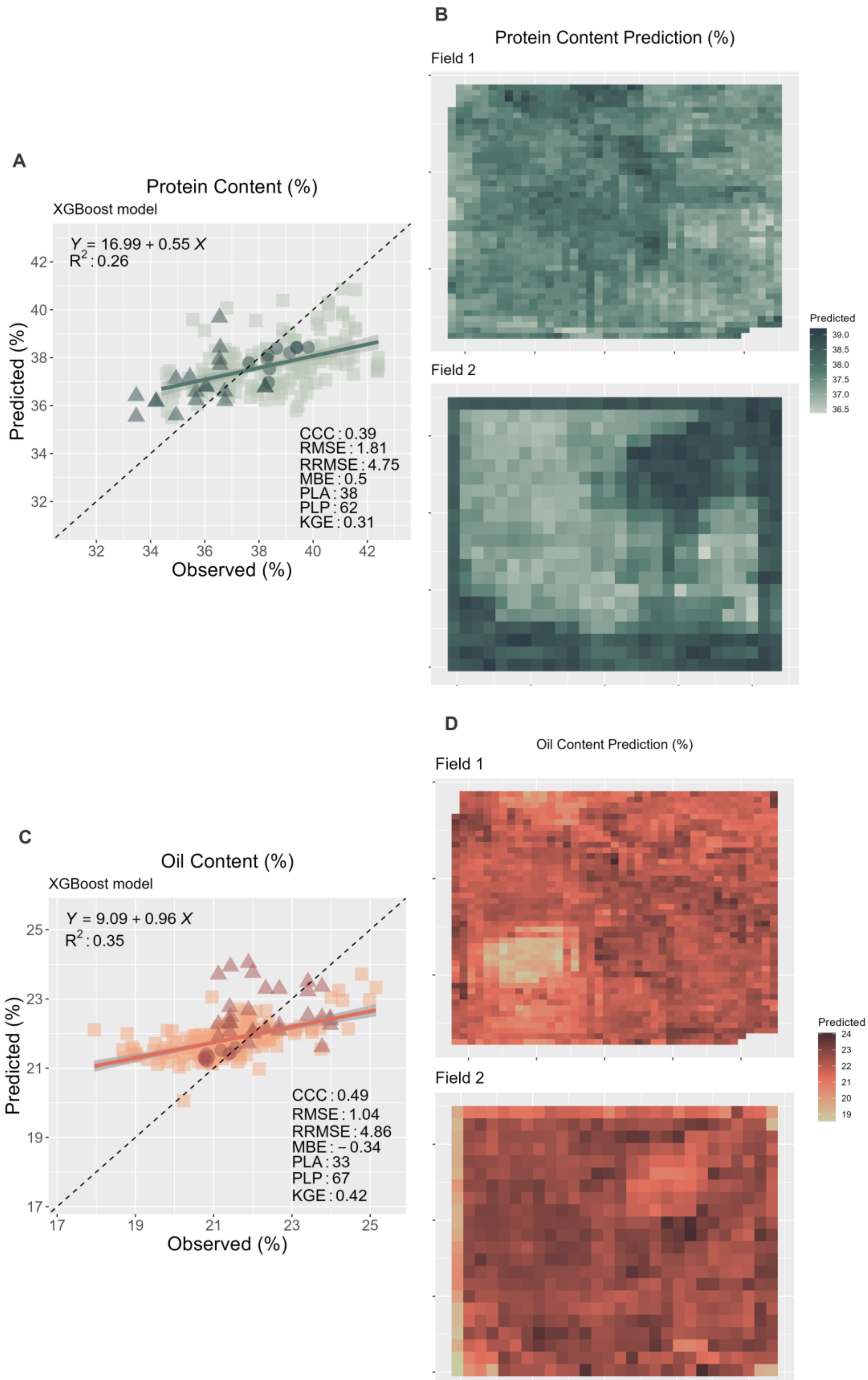
A
Protein Content (%)
XGBoost model
Y = 16.99 + 0.55 X
R2 : 0.26
CCC : 0.39
RMSE : 1.81
RRMSE : 4.75
MBE : 0.5
PLA : 38
PLP : 62
KGE : 0.31
Predicted (%)
Observed (%)
B
Protein Content Prediction (%)
Field 1
Field 2
Predicted
C
Oil Content (%)
XGBoost model
Y = 9.09 + 0.96 X
R2 : 0.35
CCC : 0.49
RMSE : 1.04
RRMSE : 4.86
MBE : − 0.34
PLA : 33
PLP : 67
KGE : 0.42
Predicted (%)
Observed (%)
D
Oil Content Prediction (%)
Field 1
Field 2
Predicted

Figure 6: Prediction performance for protein (A) and oil (C) concentration in soybean seeds using XGBoost machine learning models based on remote sensing. The shapes in the scatterplot represent the two selected fields for protein and oil concentration (B and D)

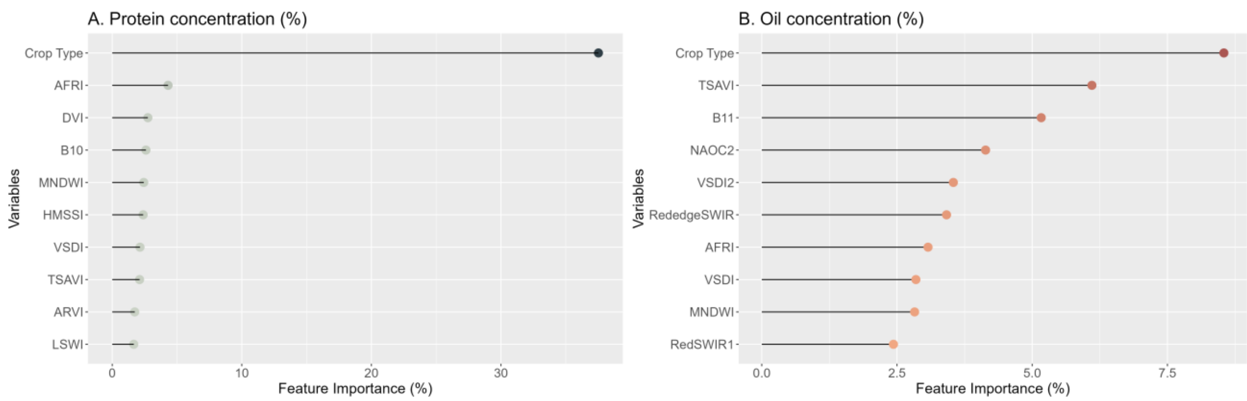

Figure 7: Feature importance assessment showing the 10 most important variables in the soybean seed protein (A) and oil concentration models (B). Double or Single crop (Crop Type), Aerosol Free Vegetation Index (AFRI), Difference Vegetation Index (DVI), Band 10

Tables

Table 1: Predictive performance metrics of the different candidate models for both protein and oil concentration. Comparison of changes in RMSE between training and testing are detailed in Table S6

		Predicting performance metrics							
	Models	R^2	CCC	RMSE	RRMSE (%)	MBE	PLA (%)	PLP (%)	KGE
Protein concentration (%)	ElasticNet	0.25	0.36	1.79	4.72	0.39	41	59	0.26
	XGBoost	0.26	0.39	1.80	4.75	0.5	38	62	0.31
	LightGBM	0.31	0.39	1.75	4.60	0.46	49	51	0.29
	Random Forest	0.25	0.31	1.83	4.82	0.51	54	46	0.21
	CatBoost	0.33	0.37	1.75	4.61	0.46	57	43	0.25
	Ensemble	0.36	0.38	1.74	4.58	0.47	60	40	0.26
Oil concentration (%)	ElasticNet	0.15	0.28	1.19	5.57	-0.37	38	62	0.19
	XGBoost	0.35	0.49	1.04	4.86	-0.34	33	67	0.42
	LightGBM	0.23	0.37	1.17	5.47	-0.47	36	64	0.32
	Random Forest	0.23	0.32	1.17	5.47	-0.47	48	52	0.24
	CatBoost	0.29	0.29	1.11	5.17	-0.42	42	58	0.33
	Ensemble	0.30	0.38	1.12	5.18	-0.41	48	52	0.29

Where R^2: Coefficient of determination; CCC: Concordance correlation coefficient; RMSE: Root Mean Squared Error; RRMSE: Relative Root Mean Squared Error; MBE: Mean Bias Error; PLA: Percentage Lack of Accuracy; PLP: Percentage Lack of Precision; KGE: Kling-Gupta Model Efficiency.

Chapter 3 - Hierarchical Bayesian kriging for spatial misalignment data. A case study applied in agriculture.

Abstract

Digital agriculture tools, such as site-specific management and predictive modeling, require increasing amounts of field-level data, often collected at different spatio-temporal scales. At spatial scale, when measurement locations of different variables do not coincide results in spatial misalignment. Bayesian kriging models have emerged as a promising method for quantifying uncertainty in interpolation and regression models. This study aimed to develop a hierarchical Bayesian kriging framework to impute missing values and quantify uncertainty in spatially misaligned datasets. The dataset included a soil sample grid with data on Cation Exchange Capacity, Nitrate Nitrogen, Organic Matter, pH, and Sulfur, along with georeferenced soybean seed protein concentration and a crop yield map. Using auxiliary variables such as a digital elevation model, compound topographic index, slope, and vegetation index named GCVI, a Bayesian hierarchical kriging scheme was implemented. While mean expected values were similar across estimated variables, variance patterns differed, with pH showing the greatest variation in standard deviation and soybean protein concentration the least spatial variability. The spatial distribution of expected variance was linked to data location, with lower variance near sampled sites. Future research should expand validation, improve computational efficiency, and develop user-friendly interfaces for agronomists, plant breeders, and other professionals. Furthermore, there is a necessity to enhance uncertainty quantification to assess error propagation in future models.

Introduction

The use of innovative model-based technologies and data-driven approaches increasingly play a key role in addressing the challenges of sustainable agriculture in the forthcoming years (El Jarroudi et al., 2024; Rozenstein et al., 2024). Consequently, digital technologies provide a plethora of data science instruments that can be instrumental in deriving insights for decision-making in agriculture (Parra-López et al., 2024). Furthermore, the potential of predictive models to achieve some of the proposed goals for sustainable development is well known (Holloway & Mengersen, 2018).

Digital agriculture could represent a feasible alternative that not only provides a framework for improving efficiencies in farming systems but also drives digital transformation (Aldoseri et al., 2024). Presently, digital agriculture tools, including precision agriculture and predictive modelling, necessitate an escalating volume of data collected at the field level (Lacoste et al., 2021). The quality of the data can be a potential barrier that needs to be overcome in order to apply some of those predictive models. In turn, countless diverse data types are collected by different sensors for subsequent integration into databases (Fuentes-Peñailillo et al., 2024).

Data integration represents one of the most challenging stages in the development process of data-driven predictive models (Benos et al., 2021). In this sense, different datasets may have different spatio-temporal attributes that make their integration difficult. These attributes include their spatial and temporal characteristics.

For several years now, unifying incompatible spatial data sets has been a complex task (Gotway & Young, 2002). In the context of data integration, when the measurement locations of a given variable differ in the locations from another variable measurement emerge a specific change of support called spatial misalignment (Banerjee & Gelfand, 2002; Finley et al., 2014; Lopiano et al., 2011). This phenomenon arises due to the integration of data at varying spatial scales, resulting in the presence of gaps or missing values during the process of data fusion (R. Zhong & Moraga, 2024). Within data analysis, there is a wide variety of data techniques for addressing spatially misaligned data problems, which sometimes overlap with multiple imputation or data padding techniques (Gelfand, 2012; Madsen et al., 2008; Pouliot, 2023). The use of probabilistic Bayesian approaches allows for flexible management not only of the data and the models to be developed, but also of their parameters, allowing for highly customizable models of all model components (Gelfand & Banerjee, 2017). In agriculture, unification is a crucial task when it comes to data integration (Córdoba et al., 2016a). However, many of the methodologies currently used do not consider or do not adequately consider the uncertainty associated with the predictions of the different data layers (Schenatto et al., 2017; Y.-H. Wu et al., 2013). The Bayesian kriging model are emerging as one of the most promising methods for quantifying uncertainty in interpolation and regression models (Pilz & Spöck, 2008).

Following this rationale, the overall objective of this project was to build a hierarchical Bayesian kriging framework to impute missing values and quantify uncertainty in spatially misaligned datasets. The specific objectives of this work are to: (I) integrate the different data sources into a complete dataset to apply the developed functions; (II) design and apply the hierarchical Bayesian kriging framework; and (III) analyze the spatial variability of predicted continuous spatial layers to datasets with spatial misalignment.

Materials and Methods

Data collection

The study area focused on an agricultural farmer field located in McPherson, Kansas. To integrate the available spatial data, five distinct types of data were gathered to characterize the spatial variability. In addition, the data collected were divided into two types of data according to if they do not match spatially (spatially misaligned) or additional variables used as explanatory variables that do not have spatial misalignment or are spatially complete without missing values. The spatially misaligned variables consisted of soil samples, soybean seed samples, and a soybean yield map. On the other hand, a set of variables named explanatory variables encompassed a digital elevation model (DEM), and a vegetation index obtained from satellite sources.

Soil samples were taken before planting using a grid design sampling at a depth of 0-15 cm with samples evenly distributed in the field covering the entire surface. The soil features extracted from the samples were Sulphur (S), Organic (OM), pH (1:1), Cation Exchange Capacity (CEC) and Nitrogen from nitrates (NO3_N).

The soybean seed samples were retrieved from a total of ten sites right before harvest at the crop maturity stage in distinct locations within the field attempting to capture the field spatial variability. The soybean seed quality parameters such as protein concentrations (%) were determined via near-infrared spectroscopy (NIR) using the Perten DA7200 Feed Analyzer (Perten Instruments, Stockholm, Sweden). The soybean seed samples were georeferenced with a GPS at the time of sampling.

Crop yield data were obtained through a yield monitor immediately after the crop harvest process. These data spatially describe the spatial variability of soybean crop yield at the end of the crop season. Given the potential for extreme yield values in these data due to various factors, a series of steps were implemented to depurate the yield map using the protocol developed by Córdoba et al., (2016). Spatial distribution of variables with spatial misalignment are displayed in Figure 1.

Regarding the explanatory layers, a digital elevation model (DEM) layer for the field study was recovered from Lidar Explorar 3DEP web platform (USGS, 2024). These DEM layers are provided through the processing of LIDAR sensor surveys, which provide products with very high spatial resolution (~1 m).

While field topography can explain differences in soil texture and in chemical properties , other topographic attributes such as the Compound Topographic Index or also known as Topographic Wetness Index and Slope information layer can be particularly useful in explaining patterns of erosion processes and drainage pathways within the field (Sheshukov et al., 2018). These phenomena can attribute variability to water distribution and nutrient distribution within the field (Liu et al., 2013).

To capture the spatial variability in the water distribution of the surface of soil, a Compound Topographic Index (CTI) layer was calculated using as input the DEM downloaded in the last step. The layers required to create the CTI layer are the Specific Catchment Area (SCA) and a Slope (SLP) layer. Both SCA and SLP are derived from DEM processing using specific algorithms described in the next steps. The DEM was first processed with a fill depression step. Additionally, a gaussian filter with a sigma equal to 12 was implemented to smooth the DEM surface making to remove small defects. The output of the last process was used as an input to create the SLP and SCA layer. The SCA layer was created using the D-Infinity flow accumulation algorithm with the Whitebox-tools R package (Lindsay, 2016) in R software (R Core Team, 2025). The CTI layer was created according to Beven & Kirkby, (1979) presented in equation 1.

$$CTI = \log\left(\frac{SCA}{\tan(SLP)}\right) \tag{1}$$

Where CTI means Compound Topographic Index, SCA is the Specific Catchment Area layer and SLP is the slope layer.

Spatial topographic attributes such as DEM or CTI allow us to understand how water flows from the highest part of the field to the lowest areas, delineating the drainage ways within field (Moore et al., 1993). In this sense, it is imperative to highlight that many characteristics of the soil are related and correlated with them, these are considered withing the most relevant soils forming factors together with the parental material, climate, biological organism, and time (Minasny et al., 2013).

To include the crop status at the critical period, a layer of Green Chlorophyll Vegetation Index (GCVI) (Gitelson et al., 2003) was included. The calculation of GCVI was done using the Green and NIR channel (Band 3 and Band 8 of Sentinel-2) according to Equation 2.

$$GCVI = \frac{NIR}{Green} - 1 \tag{2}$$

The time to retrieve the GCVI layer was selected according to the date on which each field reached its maximum value of GCVI achieved during the growing season. Several researchers mentioned that time around such as R3-R5 (from pod formation to beginning of seed filling) based on the soybean phenological stages (Fehr et al., 1971). The explanatory layers are shown in Figure 2.

Several number of past studies demonstrated advantages from the use of GCVI over other vegetation indexes (Gitelson, 2005; Lobell et al., 2015; Nguy-Robertson et al., 2012). One advantage is related to the overall capability of this index to overcome the index saturation at high levels of biomass or leaf area index (Gitelson et al., 2003).

To summarize, we sought to fill in the missing values of S, OM, pH, CEC N_NO3, Soybean seed protein concentration, and soybean crop yield values using spatial continuous layers like DEM, GCVI, SLP, and CTI as predictor variables named as explanatory layers.

Bayesian Hierarchical modeling framework

Model description

To impute the gaps of data with the missing values a Bayesian kriging model was implemented. This model has the advantage of being able to leverage a Gaussian process as an underlying process. In addition, variables described as explanatory variables were used as input data to add continuous spatial information without spatial misalignment that helps to improve the predictions.

The Bayesian kriging model is shown in the next equation according to the data, process and parameter model notation.

Data model

$$y_{j,i} \mid \mu_{j,i}, \sigma^2_{j,i} \sim N(\mu_{j,i}, \sigma^2_{j,i}\Sigma_j) \tag{3}$$

Process model

$$\mu_{j,i} = \beta_{0,j} + \sum_{X=1}^{4} \beta_{X,j} X_{i,j} \tag{4}$$

$$\Sigma_j = \exp\left(-\frac{d_{i,j}}{\lambda_j}\right) \tag{5}$$

$$d_{i,j} = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2} \tag{6}$$

Parameter model

$$\beta_{0,j} \sim Gamma(1,6) \tag{7}$$

$$\beta_{X,j} \sim N(0,2) \tag{8}$$

$$\sigma^2_{j,i} \sim Gamma(2,2) \tag{9}$$

$$\lambda_j \sim Uniform\left(0, \overline{d_{\iota,J}}\right) \tag{10}$$

Where $y_{j,i}$ Observed value of the variable for the observation i for misaligned variable j, $\mu_{j,i}$, represent the mean of the misaligned variable for the observation i, $\sigma^2_{j,i}$ is the variance for the observation i , Σ_j represents the covariance matrix for misaligned variable j where uses a distance matrix $d_{i,j}$ with the distances between the latitude and longitude x and y for two observations a and b. In the parameter model, $\beta_{0,j}$ represent the intercept of the misaligned variable j and $\beta_{X,j}$ are the coefficient given by the explanatory variables X (GCVI, DEM, SLP and CTI). Lastly, λ_j is a parameter named range that controls the spatial structure of the process.

In order to describe the Bayesian framework, the model was specified through a data model that predicts each of the spatially misaligned variables characterized through a normal distribution with mean μ and a variance defined as a $\sigma^2\Sigma$, that is composed by a variance term σ^2 and a covariance matrix Σ describing the spatial random effect (Equation 3). As for the process model, μ was modeled through a linear regression using the explanatory variables specifying five β coefficients (β_0, β_1, β_2, β_3 and β_4) according to equation 4. The covariance matrix was defined through the specification of σ^2 and a distance matrix using an exponential function (Equation 5) which uses a distance matrix $d_{i,j}$, and a spatial range parameter called λ. The distance matrix was constructed according to equation 6 which is created by calculating the Euclidean distances between each of the sites where the spatially misaligned variables were sampled.

The model parameters and their priors are specified in equations 7, 8, 9 and 10.

Posterior Prediction

With the purpose of having continuous predictions using the predicted values and conditioning the gaussian process it is possible obtain predictions spatially continuous in the study area using the expression in equation 11 for the conditional mean and equation 12 for the conditional variance. Thus, usually in geostatistical science this expression is called simple kriging predictor and is the best linear predictor under squared-error loss (Schabenberger & Gotway, 2017).

$$E[y^*|\, y\,] = \mu + k_*^T\, \sigma^2\Sigma^{-1}(y-\mu) \tag{11}$$

$$Var\,[y^*|\, y\,] = \sigma^2 - k_*^T\, \sigma^2\Sigma^{-1}k_*^T \tag{12}$$

Where the $E[y^*|\, y\,]$ represent the expected value of y^*conditional to y, μ is the mean of the expected value k_* is a covariance vector between the site to predict and the site of the observed value, $\sigma^2\Sigma$ is the covariance matrix that define the spatial structure of the observed values and σ^2 is the variance of the expected value.

The model was fitted using R (R Core Team, 2025) through the NIMBLE package (De Valpine et al., 2017). One of the advantages of NIMBLE over other statistical packages is that it allows missing values to be managed through latent states. On the other hand, NIMBLE is flexible in terms of which sampling algorithm to use, having the option to select Metropolis-Hastings algorithm or Gibbs sampling depending on how the model was configured. However, when an analytical form of the full conditional distribution is available, it opts for Gibbs sampling. Furthermore, in computational terms, NIMBLE has demonstrated greater efficiency than probabilistic languages such as JAGS and STAN (Baraha, 2021). In this case, the sampling method used was Gibbs sampling. In summary, the configuration to generate the sequence of samples that approximate the parameters of the posterior probability distribution functions used were 10,000 iterations with a burn-in period of 2000, a thin of one, and setting a seed for reproducibility.

Results

Data Collection

The spatially misaligned variables demonstrated varying degrees of missing data. For the seed protein concentration variable, ninety-eight values were missing. For pH, CEC, OM, NO3_N and S, seventy-seven values were missing. Finally, the yield variable had twenty-five missing data. A detailed description of the descriptive statistics for each variable is described in Table 1.

Imputation analysis

Figure 3 presents a comparison between the distributions of observed and predicted data, revealing strong similarities in their mean values for each variable. Specifically, the mean protein concentration in soybean seeds was 40.27% for observed data and 40.37% for predicted data, while pH values were 6.07 and 6.06, respectively. For CEC, the values were 13.28 and 13.39 meq/100g, OM recorded 2.31% and 2.34%, NO3-N measured 4.34 and 4.40 ppm, S values were 15.23 and 15.20 ppm, and soybean yield was 3,534 and 3,528 kg/ha for observed and predicted data, respectively. Although most variables exhibited similar minimum and maximum values, the predicted data had a narrower range of variation, with lower minimum values and higher maximum values compared to the observed data.

In terms of percentage differences between observed and predicted mean values, using the observed data as a reference, variations ranged from 0.16% to 1.38%, with pH showing the smallest difference and NO3-N the largest. However, greater discrepancies were observed in standard deviation values, which varied between 7.81% and 45.83%. The soil pH variable exhibited the highest variation in standard deviation, whereas soybean protein concentration displayed the least spatial variability. This suggests that, while the mean pH values were similar, the predicted values captured less spatial variability than the observed data. A detailed summary of these values is provided in Table 2.

Predictive performance

Generating posterior predictive distributions and predicting each layer of the model revealed distinct spatial variability patterns across the misaligned data layers (Figure 4.I). These differences were evident not only in the predicted posterior means but also in the posterior variances. The latter varied in direct relation to both the quantity and spatial distribution of observed data, with areas of higher or lower uncertainty reflected in the variance around the predicted mean values. Specifically, variance increased with distance from observed data points, leading to greater uncertainty in regions with a lower density of observations (Figure 4.II).

Discussion

The data framework introduced on this study presents a hierarchical Bayesian geostatistical approach designed to fill missing values in several variables caused by spatial misalignment data. Additionally, this workflow allows for the estimation of continuous layers across the entire study area. The use of spatially continuous variables, such as topographic layers from DEM, CTI, and SLP, alongside vegetation indices like GCVI, provided an advantage by incorporating auxiliary data layers to estimate the variables of interest (Pusch et al., 2022). Thus, the use of Bayesian schemes through the inclusion of covariates to estimate the expected value or variance is as technique widely used in spatial data modeling both to model the value or the expected variance and the spatial data structure (Yarali & Rivaz, 2020). Furthermore, the developed approach has the advantage of integrating prior knowledge via the selection of distributions and their parameters, using probabilistic models that enable the quantification of uncertainty. In this context, modern analytical methods that provide predictions along with their associated uncertainty facilitate a better understanding of spatial variability and the errors around the estimation (Samsonova et al., 2017).

The use of new digital tools enables the appropriate integration of various layers of information (Mgendi, 2024). Moreover, the integration of computational tools and on-farm data offers the possibility to provide a better framework to understand the factors that which define spatial and temporal variability (Basso & Antle, 2020; Lacoste et al., 2021). In this regard, Anthony et al. (2012) conducted an extensive analysis of the influence of soil properties on soybean crop yield and quality parameters. In addition, Cox et al., (2003) showed an in-depth analysis of the spatial variability of topography and its relationship with crop yield, demonstrating a strong influence of the soil properties and topographical position of the land has an influence on the soybean crop yield. On the other hand, Kravchenko & Bullock, (2002) showed how the effect of spatial variability in turn alters seed quality patterns in soybeans. More recently, Hernandez et al., (2023) showed the promising capabilities of using machine learning models to spatially map soybean quality parameters by including the use of remotely sensed vegetation indices. This reinforces the rationale for including data layers such as remote sensing data and topographic attributes as predictors in spatial prediction models. Similarly, incorporating satellite data during the growing season provides up-to-date insights into crop growth status and its impact on spatial yield variability. Despite these advancements, the application of spatial misalignment models in agriculture remains limited, particularly in addressing missing spatial data and its use for inference. In forestry, (Finley et al., 2014) developed methodologies similar to those employed in this research to perform spatial predictions of dependent and independent variables, underscoring the potential for broader application in agricultural contexts.

Beyond agriculture, spatial misalignment of data is a widespread problem in various scientific fields. In the health sciences, such frameworks have been used to analyze life expectancy in relation to deprivation using data at different geographic scales (Johnson et al., 2020). Additionally, Gramatica et al., (2021) studied the prevalence of diabetes and mortality in patients with this disease, employing multiple membership approaches to address areal misalignment. In economics, Pouliot, (2023) explored this issue by presenting methodologies such as two-step bootstrap and Krig-and-regress to reanalyze data published by (Maccini & Yang, 2009) under spatially misaligned regression models. Similarly, in the environmental sciences, Madsen et al. (2008) used data from the Environmental Protection Agency (EPA) to compare the aforementioned Krig-and-regress approach with maximum likelihood models using deterministic approaches. More recently, Zhong & Moraga, (2024) tackled similar challenges by predicting fine particulate matter (PM2.5) concentrations using monitoring station data, satellite observations, and environmental information, leveraging other Bayesian approaches such as Integrated Nested Laplace Approximation (INLA) and Stochastic Partial Differential Equation (SPDE) methods. In our case, this framework allowed us to generate predictions for soil variables, crop yield, and protein concentration in soybean seeds at locations where no data had been collected. This data could serve as a starting point for constructing models that not only address spatial misalignment but also enable an appropriate quantification of uncertainty associated with measurement and prediction errors.

Hierarchical Bayesian models are widely regarded as one of the most suitable techniques for modeling data in situations of spatial misalignment (Zhu & Carlin, 2000). Due to this, numerous researchers have explored a wide variety of these predictive approaches in recent years (Cameletti et al., 2019; Finley et al., 2014; Johnson et al., 2020; Liang & Kumar, 2013). Part of the increased adoption of these models is attributed to advancements in model development and processing techniques (Berger, 2000). However, a sizable portion of this growth is also due to improvements in large-scale computing power and reduced computational costs (Levy & McNeish, 2023). Among the most widely used methods for generating samples, MCMC and INLA stand out, providing great flexibility in model design and structuring (De Smedt et al., 2015). Nevertheless, programming and applying these methodologies in their purest form can be complex and require extensive knowledge of Bayesian statistics, with still a high requirement of computing level high level (Bon et al., 2023). The increasing adoption of these methodologies has also driven the creation of new high-level computational tools that offer more intuitive interfaces (Abbasi et al., 2019). In this regard, embedded languages in R or Python, such as NIMBLE (De Valpine et al., 2017), JAGS (Plummer, 2004), and STAN (Stan Development Team, 2015), allow for substantial flexibility in constructing complex hierarchical Bayesian models (Beraha et al., 2021), though they still require a high level of understanding of the underlying Bayesian statistical theory. Other function packages, such as SpBayes (Finley et al., 2007) provide less flexibility for handling complex models but offer more direct results with lower coding complexity. In our case, NIMBLE allowed highly flexible design and execution.

From the perspective of potential users, this framework is not restricted solely to agronomy professionals. Given the nature of these challenges, it is applicable to anyone aiming to perform spatial imputation or prediction under conditions of spatial misalignment or missing data. For instance, an agronomist with data analysis expertise can generate information to characterize spatial variability, which could serve as input for clustering or regression models.

The delineation of management zones in crop fields demands spatially continuous layers to achieve optimal results. Similarly, a crop breeder could use this approach to fill missing values in an experiment affected by unforeseen issues. Some authors, such as (Yang et al., 2023) have employed similar techniques under these circumstances.

Despite the promising results obtained, the proposed framework is still far from ideal. Several aspects could be addressed to enhance its performance and capabilities, including: I) The dataset used in this study was from a single agricultural field; additional independent datasets may be required for proper evaluation, II) Model performance evaluation could benefit from cross-validation to highlight how well the model predicts spatially misaligned variables using an out-of-sample dataset, III) The implementation and comparison of alternative covariance functions, as only the exponential function was used in this study, would help determine the appropriate conditions for applying each function, IV) The use of MCMC for multiple variables and large datasets would help to improve the performance of the proposed framework. Alternative hierarchical Bayesian frameworks, such as INLA, should be explored to evaluate their performance in comparison (Rue et al., 2009).

Future research should focus on improving the understanding of spatial statistical models, enhancing the workflow between spatial product generation and accurate uncertainty quantification through a more accessible interface for end users.

Conclusion

This study demonstrates the effectiveness of the hierarchical Bayesian geostatistical framework developed to address spatial misalignment problems and generate continuous predictions of agronomic layers. The integration of ancillary data layers, including topographic attributes and vegetation indices, proved beneficial in improving prediction accuracy and capturing spatial variability. The workflow successfully estimated missing values for soil properties, soybean yield and seed protein concentration, while quantifying uncertainty and providing valuable insight into spatial patterns.

Future research should focus on expanding validation, improving computational efficiency, and developing intuitive interfaces to facilitate adoption by agronomists, plant breeders, and other professionals working with spatially misaligned data. Thes study highlights the potential of hierarchical Bayesian approaches to optimize decision-making in precision agriculture and other areas requiring integration of spatially misaligned data, while also quantifying the uncertainty associated with predictions to assess potential error propagation in future models.

Figures

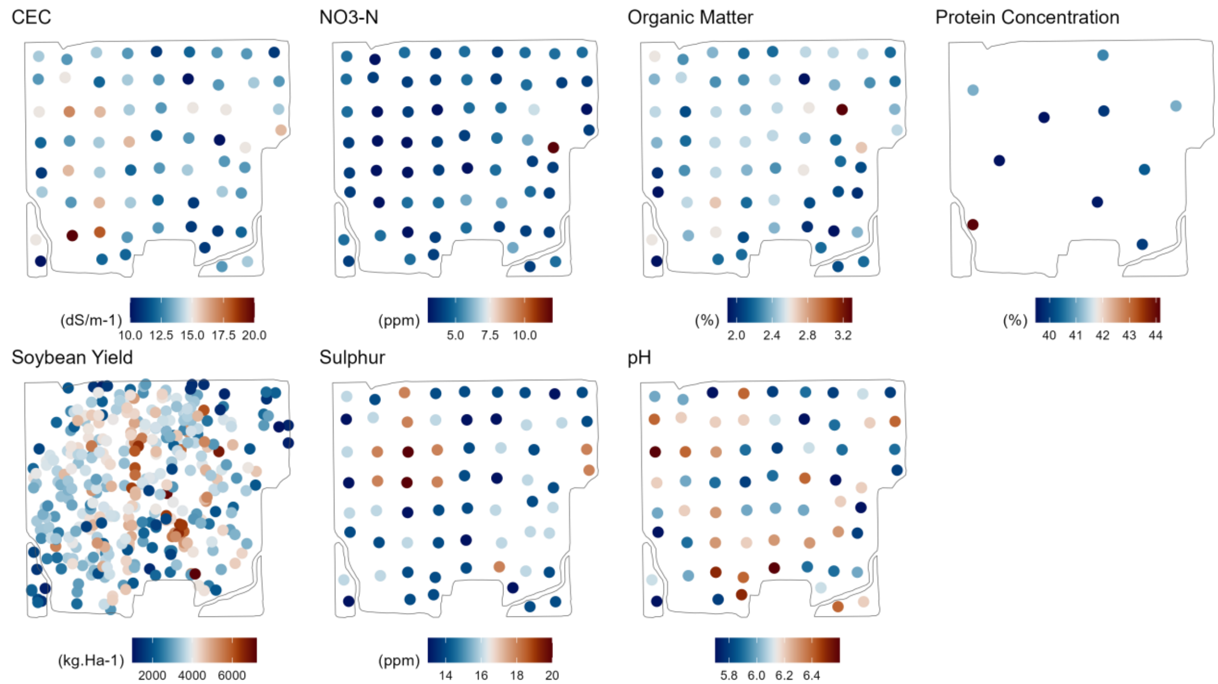

Figure 8: Spatial distribution of the different data sources with spatial misalignment. Merging these data from a single data source produces missing values for the other variables.

The distribution pattern changes according to the nature of the data collection. For data collected through our soil surveys, such as CEC, NO3-N, organic matter, sulfur, and pH, the pattern corresponds to grid sampling. Soybean yield values correspond to a sample from a yield map. Due to the standard volume of observations in these data, they correspond to a random subsample using 0.5% of the total data. In the case of the georeferenced soybean seed protein concentration samples, the samples were taken at the most representative sites in the study area.

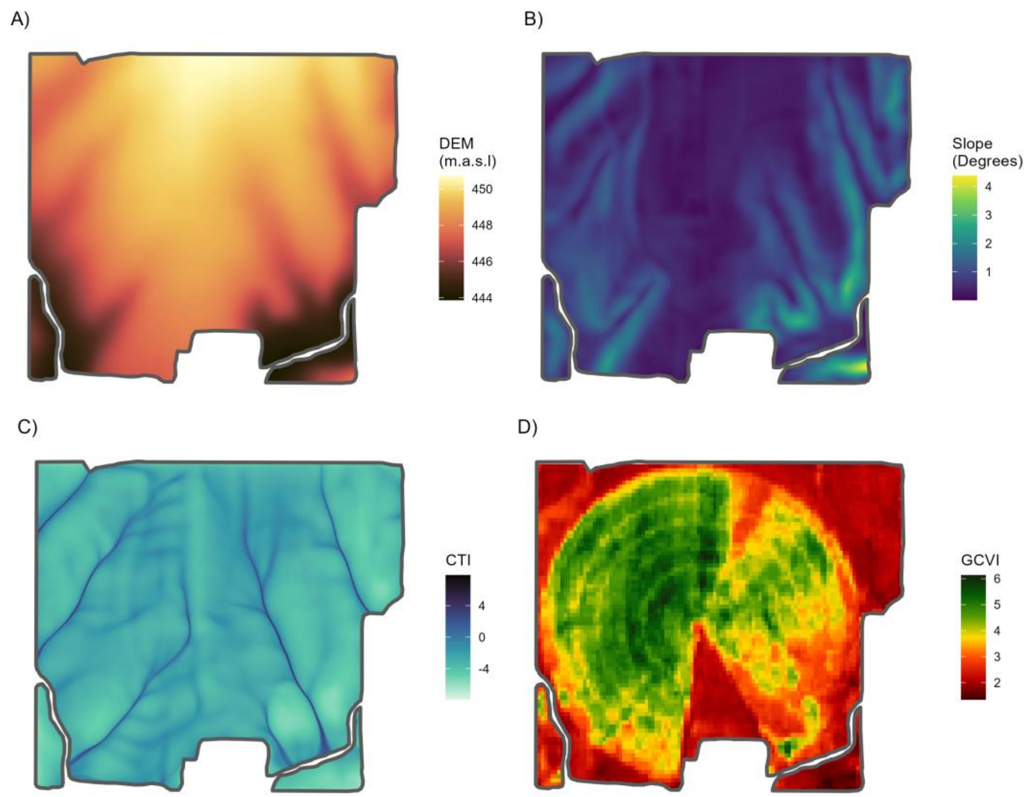

Figure 9: Special variability of the variables used as explanatory variables A) Digital Elevation Model (DEM). Sectors with lower values are depressed areas within the field that have greater water input from runoff during rainfall. B) Visualization of a slope map showing the degree of inclination measured in degrees with respect to the horizontal axis. Sectors with greater slopes are more prone to water erosion and degradation due to nutrient loss. C) Map of compound topographic index (CTI) map. Higher CTI values indicate areas with greater runoff and surface water movement, highlighting possible drainage pathways. D) Green Chlorophyll Vegetation Index. This vegetation index is used to monitor crop status due to its relationship with accumulated biomass and the leaf area index.

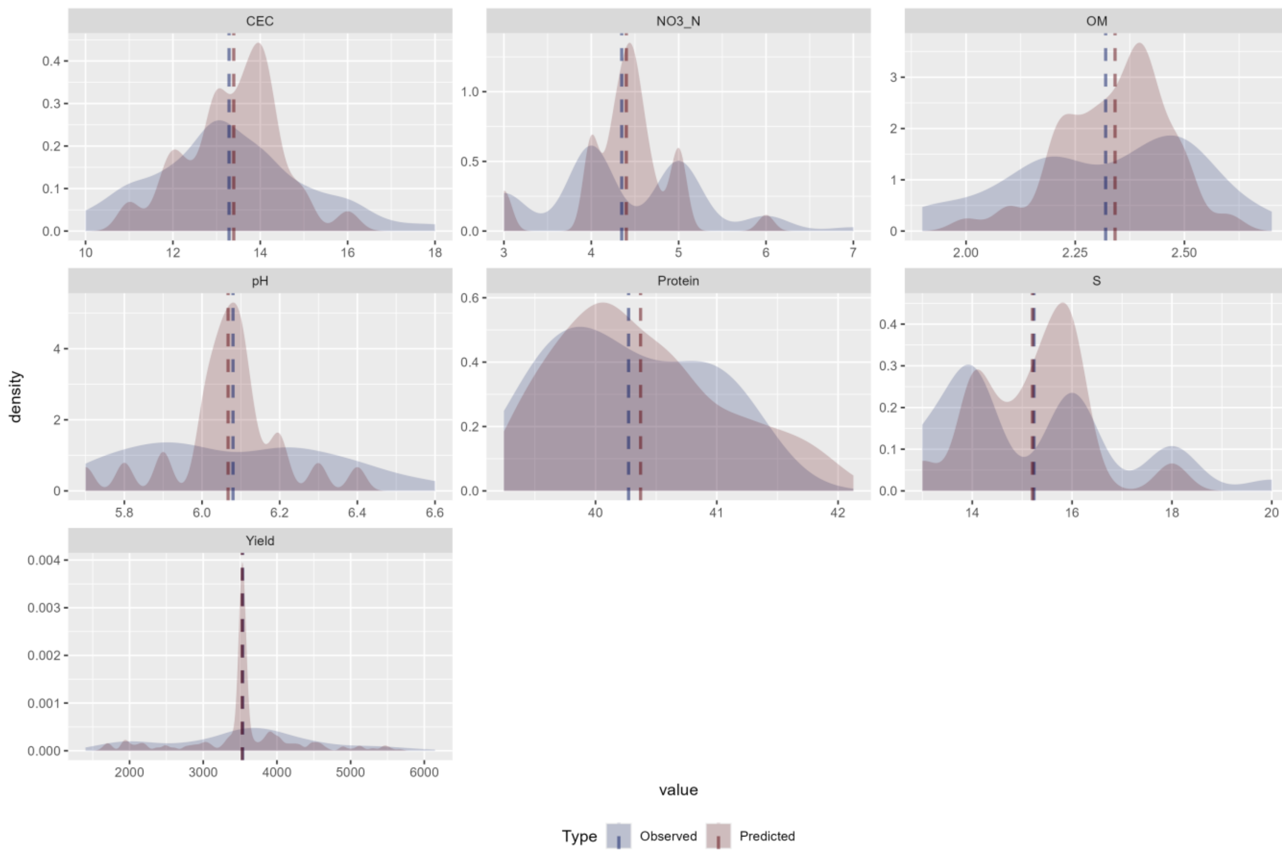

Figure 10: Distributions of observed and estimated data for variables with spatial misalignment. The dashed lines represent the mean of the distribution for each data set.

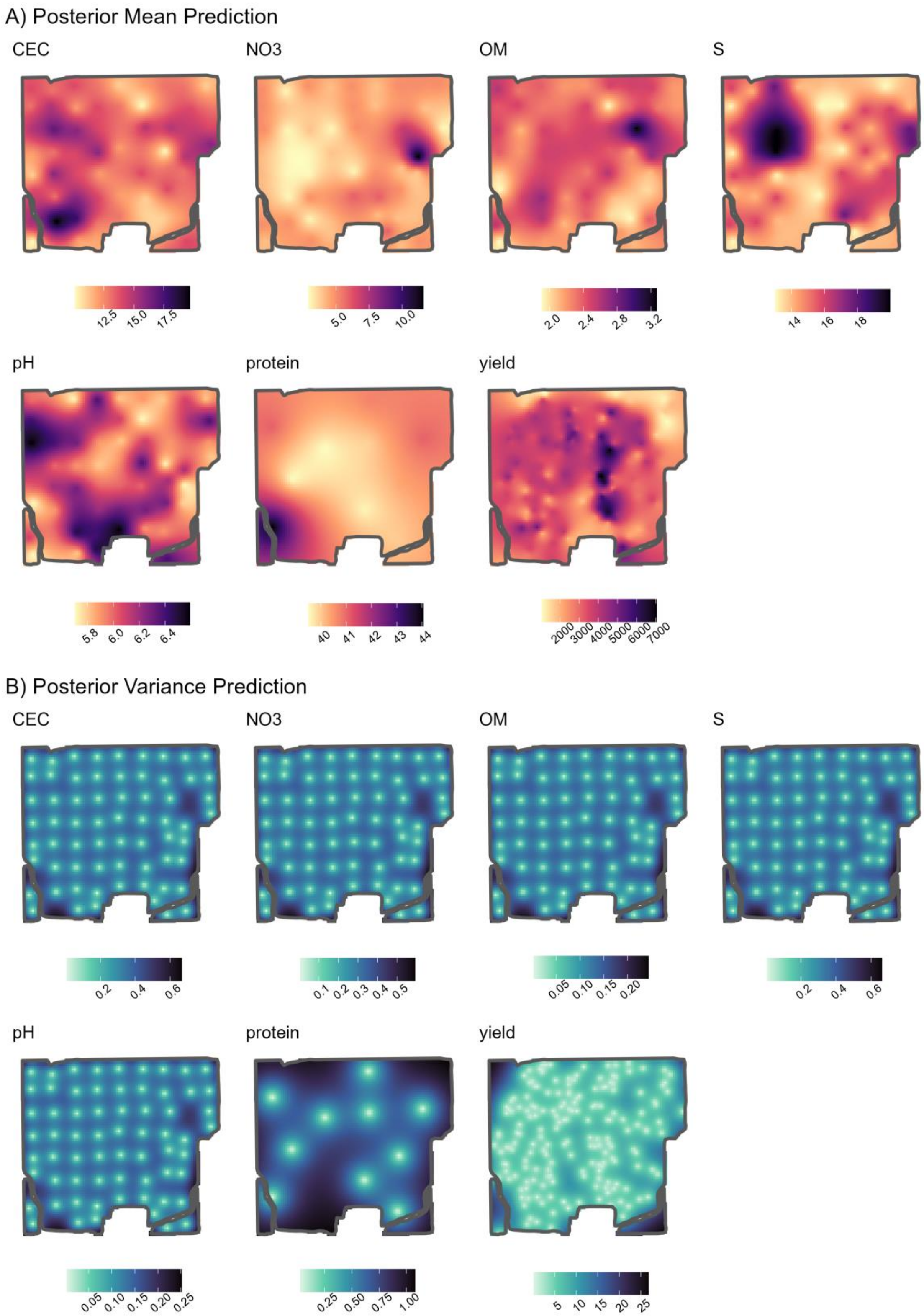
A) Posterior Mean Prediction
CEC
12.5 15.0 17.5
NO3
5.0 7.5 10.0
OM
2.0 2.4 2.8 3.2
S
14 16 18
pH
5.8 6.0 6.2 6.4
protein
40 41 42 43 44
yield
2000 3000 4000 5000 6000 7000
B) Posterior Variance Prediction
CEC
0.2 0.4 0.6
NO3
0.1 0.2 0.3 0.4 0.5
OM
0.05 0.10 0.15 0.20
S
0.2 0.4 0.6
pH
0.05 0.10 0.15 0.20 0.25
protein
0.25 0.50 0.75 1.00
yield
5 10 15 20 25

Figure 11 I) Spatial visualization of the posterior mean and II) variance prediction. These maps show how the spatial predictions of variables with spatial misalignment are distributed. While some of these variables showed greater spatial variability (CEC, OM, S, pH, and Yield), others, such as NO3-N, exhibited a more consistent pattern across the field area. On the one hand, these maps show that the variance values decrease near the data sample collection sites. In contrast, as can be seen in the soybean protein concentration map, due to a smaller proportion of samples covering the area, various parts of the field exhibited higher uncertainty values in the expected values.

Tables

Table 2: Description of the characteristics of variables with spatial misalignment. While the "n" column represents the number of observed values, the "Missing Values" column represents the missing values due to data merging. The Mean, SD, Min, and Max columns represent the mean, standard deviation, and the minimum and maximum values within the observed data.

Misaligned variables	n	Missing values	Mean	SD	Min	Max
Soybean Protein Concentration (%)	10	298	40.58	1.39	39.48	44.15
pH (1:1)	114	194	6.07	0.24	5.7	6.6
Cation Exchange Capacity (CEC) (meq/100g)	114	194	13.40	1.93	10	20
Organic Matter (%)	114	194	2.33	0.24	1.9	3.3
NO3-N (ppm)	114	194	4.48	1.35	3	12
S (ppm)	114	194	15.23	1.84	13	20
Soybean Yield (kg. Ha^{-1})	184	124	3561	1213	1023	7360

Table 3: Description of the means and standard deviations of the observed and predicted values, and their differences at the point estimate level using a 95% interval of the data. The difference between the observed and predicted values is based on the observed values and is expressed as absolute percentages.

	Observed		**Predicted**		**Difference (%)**	
Misaligned variables	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Soybean Protein Concentration (%)	40.27	0.64	40.37	0.69	0.25	7.81
pH (1:1)	6.07	0.24	6.06	0.13	0.16	45.83
Cation Exchange Capacity (CEC) (meq/100g)	13.28	1.73	13.39	1.07	0.83	38.15
Organic Matter (OM) (%)	2.31	0.20	2.34	0.12	1.30	40.00
NO3-N (ppm)	4.34	0.91	4.40	0.53	1.38	41.76
S (ppm)	15.23	1.84	15.20	1.11	0.20	39.67
Soybean Yield (kg. Ha^{-1})	3534.50	1027.50	3528.03	724.51	0.18	29.49

Chapter 4 - Final remarks

This study introduces various methodological frameworks aimed at characterizing the spatial variability of key agronomic variables, encompassing both environmental and crop-related factors.

In Chapter 2, predictive modeling leveraging satellite imagery—specifically vegetation indices—combined with machine learning algorithms achieved an RMSE of 1.80 and 1.04 for seed protein and oil concentrations, respectively. The most effective period for performing these estimations was identified as approximately one week after the GCVI peak. This research establishes a foundation for generating spatially explicit soybean seed quality maps prior to or near harvest, offering significant implications not only for the domestic market for potential for field segregation, but also for the global market by strengthening the competitiveness of the U.S. soybean value chain.

In Chapter 3, a hierarchical Bayesian geostatistical framework is applied to address the challenge of spatial misalignment by estimating continuous spatial variables from agronomic data layers using auxiliary predictors. Integrating supplementary datasets, such as topographic attributes and vegetation indices, proved instrumental in refining prediction accuracy and capturing spatial heterogeneity. This approach effectively estimated missing values for soil properties, soybean yield, and seed protein concentration, while simultaneously quantifying uncertainty and offering valuable insights into spatial distribution patterns.

Future research should prioritize the integration of the strengths of these methodologies while maintaining their core objectives. The development of predictive models capable of incorporating multi-scale spatial data and providing comprehensive assessments of associated uncertainties will be crucial. In addition, these models should have the flexibility to capture complex non-linear relationships and interactions between explanatory variables, while effectively accounting for the spatial structure inherent in both input and response data.

References

Abbasi, A. A., Abbasi, A., Shamshirband, S., Chronopoulos, A. T., Persico, V., & Pescapè, A. (2019). Software-defined cloud computing: A systematic review on latest trends and developments. *Ieee Access*, *7*, 93294-93314.

Abraham, G. K., Jayanthi, V., & Bhaskaran, P. (2020). Convolutional neural network for biomedical applications. En *Computational Intelligence and Its Applications in Healthcare* (pp. 145-156). Elsevier.

Aldoseri, A., Al-Khalifa, K. N., & Hamouda, A. M. (2024). AI-Powered Innovation in Digital Transformation: Key Pillars and Industry Impact. *Sustainability*, *16*(5), 1790. https://doi.org/10.3390/su16051790

Alonso-Gonzalez, A., Jagdhuber, T., & Hajnsek, I. (2015). *Agricultural monitoring with polarimetric SAR time series*. 2015 8th International Workshop on the Analysis of Multitemporal Remote Sensing Images, Multi-Temp 2015. https://doi.org/10.1109/Multi-Temp.2015.7245798

Anthony, P., Malzer, G., Sparrow, S., & Zhang, M. (2012). Soybean Yield and Quality in Relation to Soil Properties. *Agronomy Journal*, *104*(5), 1443-1458. https://doi.org/10.2134/agronj2012.0095

Araújo, S. O., Peres, R. S., Ramalho, J. C., Lidon, F., & Barata, J. (2023). Machine learning applications in agriculture: Current trends, challenges, and future perspectives. *Agronomy*, *13*(12), 2976.

Assefa, Y., Purcell, L. C., Salmeron, M., Naeve, S., Casteel, S. N., Kovács, P., Archontoulis, S., Licht, M., Below, F., Kandel, H., Lindsey, L. E., Gaska, J., Conley, S., Shapiro, C., Orlowski, J. M., Golden, B. R., Kaur, G., Singh, M., Thelen, K., … Ciampitti, I. A. (2019). Assessing Variation in US Soybean Seed Composition (Protein and Oil). *Frontiers in Plant Science*, *10*, 298. https://doi.org/10.3389/fpls.2019.00298

Bajwa, S., Rupe, J., & Mason, J. (2017). Soybean Disease Monitoring with Leaf Reflectance. *Remote Sensing*, *9*(2), Article 2. https://doi.org/10.3390/rs9020127

Banerjee, S., & Gelfand, A. (2002). Prediction, interpolation and regression for spatially misaligned data. *Sankhyā: The Indian Journal of Statistics, Series A*, 227-245.

Barbedo, J. G. A. (2022). Data Fusion in Agriculture: Resolving Ambiguities and Closing Data Gaps. *Sensors*, *22*(6), 2285. https://doi.org/10.3390/s22062285

Basnet, B. B., Apan, A., Kelly, R., Jensen, T., Strong, W., & Butler, D. (2003). *Relating satellite imagery with grain protein content*. 1-11.

Basso, B., & Antle, J. (2020). Digital agriculture to design sustainable agricultural systems. *Nature Sustainability*, *3*(4), 254-256. https://doi.org/10.1038/s41893-020-0510-0

Bastos, L. M., Froes de Borja Reis, A., Sharda, A., Wright, Y., & Ciampitti, I. A. (2021). Current Status and Future Opportunities for Grain Protein Prediction Using On- and Off-Combine Sensors: A Synthesis-Analysis of the Literature. *Remote Sensing*, *13*(24), Article 24. https://doi.org/10.3390/rs13245027

Ben Ayed, R., & Hanana, M. (2021). Artificial intelligence to improve the food and agriculture sector. *Journal of Food Quality*, *2021*(1), 5584754.

Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). Machine Learning in Agriculture: A Comprehensive Updated Review. *Sensors*, *21*(11), 3758. https://doi.org/10.3390/s21113758

Beraha, M., Falco, D., & Guglielmi, A. (2021). *JAGS, NIMBLE, Stan: A detailed comparison among Bayesian MCMC software* (Versión 1). arXiv. https://doi.org/10.48550/ARXIV.2107.09357

Berger, J. O. (2000). Bayesian Analysis: A Look at Today and Thoughts of Tomorrow. *Journal of the American Statistical Association*, *95*(452), 1269-1276. https://doi.org/10.1080/01621459.2000.10474328

Beven, K. J., & Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology / Un modèle à base physique de zone d’appel variable de l’hydrologie du bassin versant. *Hydrological Sciences Bulletin*, *24*(1), 43-69. https://doi.org/10.1080/02626667909491834

Bon, J. J., Bretherton, A., Buchhorn, K., Cramb, S., Drovandi, C., Hassan, C., Jenner, A. L., Mayfield, H. J., McGree, J. M., Mengersen, K., Price, A., Salomone, R., Santos-Fernandez, E., Vercelloni, J., & Wang, X. (2023). Being Bayesian in the 2020s: Opportunities and challenges in the practice of modern applied Bayesian statistics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *381*(2247), 20220156. https://doi.org/10.1098/rsta.2022.0156

Borja Reis, A. F., Rosso, L., Davidson, D., Kovács, P., Purcell, L. C., Below, F. E., Casteel, S., Kandel, H. J., Naeve, S., Archontoulis, S. V., & Ciampitti, I. A. (2022). Soybean management for seed composition: The perspective of U.S. farmers. *Agronomy Journal*, *114*(4), Article 4. https://doi.org/10.1002/agj2.21082

Bramley, R. G., Song, X., Colaço, A. F., Evans, K. J., & Cook, S. E. (2022). Did someone say “farmer-centric”? Digital tools for spatially distributed on-farm experimentation. *Agronomy for Sustainable Development*, *42*(6), 105.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), Article 1.

Cai, Y., Guan, K., Peng, J., Wang, S., Seifert, C., Wardlow, B., & Li, Z. (2018). A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. *Remote Sensing of Environment*, *210*, 35-47. https://doi.org/10.1016/j.rse.2018.02.045

Cameletti, M., Gómez-Rubio, V., & Blangiardo, M. (2019). Bayesian modelling for spatially misaligned health and air pollution data through the INLA-SPDE approach. *Spatial Statistics*, *31*, 100353. https://doi.org/10.1016/j.spasta.2019.04.001

Chen, T., & Guestrin, C. (2016). *Xgboost: A scalable tree boosting system*. 785-794.

Chiozza, M. V., Parmley, K. A., Higgins, R. H., Singh, A. K., & Miguez, F. E. (2021). Comparative prediction accuracy of hyperspectral bands for different soybean crop variables: From leaf area to seed composition. *Field Crops Research*, *271*, 108260. https://doi.org/10.1016/j.fcr.2021.108260

Córdoba, M. A., Bruno, C. I., Costa, J. L., Peralta, N. R., & Balzarini, M. G. (2016a). Protocol for multivariate homogeneous zone delineation in precision agriculture. *Biosystems Engineering*, *143*, 95-107. https://doi.org/10.1016/j.biosystemseng.2015.12.008

Córdoba, M. A., Bruno, C. I., Costa, J. L., Peralta, N. R., & Balzarini, M. G. (2016b). Protocol for multivariate homogeneous zone delineation in precision agriculture. *Biosystems Engineering*, *143*, 95-107. https://doi.org/10.1016/j.biosystemseng.2015.12.008

Correndo, A., McArtor, B., Prestholt, A., Hernandez, C., Kyveryga, P. M., & Ciampitti, I. A. (2022). Interactive soybean variable-rate seeding simulator for farmers. *Agronomy Journal*, *114*(6), Article 6.

Correndo, A., Rosso, L. M., Hernandez, C., Bastos, L. M., Nieto, L., Holzworth, D., & Ciampitti, I. A. (2022). metrica: An R package to evaluate prediction performance of regression and classification point-forecast models. *Journal of Open Source Software*, *7*(79), Article 79.

Cox, M. S., Gerard, P. D., Wardlaw, M. C., & Abshire, M. J. (2003). Variability of Selected Soil Properties and Their Relationships with Soybean Yield. *Soil Science Society of America Journal*, *67*(4), 1296-1302. https://doi.org/10.2136/sssaj2003.1296

De Smedt, T., Simons, K., Van Nieuwenhuyse, A., & Molenberghs, G. (2015). Comparing MCMC and INLA for disease mapping with Bayesian hierarchical models. *Archives of Public Health*, *73*(S1), O2. https://doi.org/10.1186/2049-3258-73-S1-O2

De Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., & Bodik, R. (2017). Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE. *Journal of Computational and Graphical Statistics*, *26*(2), 403-413. https://doi.org/10.1080/10618600.2016.1172487

Diao, C. (2020). Remote sensing phenological monitoring framework to characterize corn and soybean physiological growing stages. *Remote Sensing of Environment*, *248*, 111960. https://doi.org/10.1016/j.rse.2020.111960

Dilmurat, K., Sagan, V., Maimaitijiang, M., Moose, S., & Fritschi, F. B. (2022). Estimating Crop Seed Composition Using Machine Learning from Multisensory UAV Data. *Remote Sensing*, *14*(19), Article 19.

El Jarroudi, M., Kouadio, L., Delfosse, P., Bock, C. H., Mahlein, A.-K., Fettweis, X., Mercatoris, B., Adams, F., Lenné, J. M., & Hamdioui, S. (2024). Leveraging edge artificial intelligence for sustainable agriculture. *Nature Sustainability*, *7*(7), 846-854. https://doi.org/10.1038/s41893-024-01352-4

Elbasi, E., Zaki, C., Topcu, A. E., Abdelbaki, W., Zreikat, A. I., Cina, E., Shdefat, A., & Saker, L. (2023). Crop prediction model using machine learning algorithms. *Applied Sciences*, *13*(16), 9288.

Fao. (2019). *The state of food and agriculture 2019: Moving forward on food loss and waste reduction*. UN.

Fehr, W. R., Caviness, C. E., Burmood, D. T., & Pennington, J. S. (1971). Stage of Development Descriptions for Soybeans, *Glycine Max* (L.) Merrill [1]. *Crop Science*, *11*(6), Article 6. https://doi.org/10.2135/cropsci1971.0011183X001100060051x

Finley, A. O., Banerjee, S., & Carlin, B. P. (2007). spBayes: An R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of statistical software*, *19*(4), 1.

Finley, A. O., Banerjee, S., & Cook, B. D. (2014). Bayesian hierarchical models for spatially misaligned data in R. *Methods in Ecology and Evolution*, *5*(6), 514-523. https://doi.org/10.1111/2041-210X.12189

Food and Agriculture Organization of the United Nations. (2023). Statistical databases. *Food and Agriculture Organization of the United Nations*. https://search.library.wisc.edu/catalog/999890171702121

Fuentes-Peñailillo, F., Gutter, K., Vega, R., & Silva, G. C. (2024). Transformative Technologies in Digital Agriculture: Leveraging Internet of Things, Remote Sensing, and Artificial Intelligence for Smart Crop Management. *Journal of Sensor and Actuator Networks*, *13*(4). https://doi.org/10.3390/jsan13040039

Gao, F., & Zhang, X. (2021). Mapping crop phenology in near real-time using satellite remote sensing: Challenges and opportunities. *Journal of Remote Sensing*, *2021*.

Gaso, D. V., De Wit, A., De Bruin, S., Puntel, L. A., Berger, A. G., & Kooistra, L. (2023). Efficiency of assimilating leaf area index into a soybean model to assess within-field yield variability. *European Journal of Agronomy*, *143*, 126718. https://doi.org/10.1016/j.eja.2022.126718

Gatti, A., & Bertolini, A. (2013). Sentinel-2 products specification document. *Available online (accessed February 23, 2015) https://earth. esa. int/documents/247904/685211/Sentinel-2+ Products+ Specification+ Document*.

Gelfand, A. E. (2012). Hierarchical modeling for spatial data problems. *Spatial Statistics*, *1*, 30-39. https://doi.org/10.1016/j.spasta.2012.02.005

Gelfand, A. E., & Banerjee, S. (2017). Bayesian modeling and analysis of geostatistical data. *Annual review of statistics and its application*, *4*(1), 245-266.

Georgi, C., Spengler, D., Itzerott, S., & Kleinschmit, B. (2018). Automatic delineation algorithm for site-specific management zones based on satellite remote sensing data. *Precision Agriculture*, *19*(4), 684-707. https://doi.org/10.1007/s11119-017-9549-y

Gitelson, A. A. (2005). Remote estimation of canopy chlorophyll content in crops. *Geophysical Research Letters*, *32*(8), Article 8. https://doi.org/10.1029/2005GL022688

Gitelson, A. A., Viña, A., Arkebauer, T. J., Rundquist, D. C., Keydan, G., & Leavitt, B. (2003). Remote estimation of leaf area index and green leaf biomass in maize canopies: REMOTE ESTIMATION OF LEAF AREA INDEX. *Geophysical Research Letters*, *30*(5), Article 5. https://doi.org/10.1029/2002GL016450

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, *202*, 18-27.

Gotway, C. A., & Young, L. J. (2002). Combining Incompatible Spatial Data. *Journal of the American Statistical Association*, *97*(458), 632-648. https://doi.org/10.1198/016214502760047140

Gramatica, M., Congdon, P., & Liverani, S. (2021). Bayesian Modelling for Spatially Misaligned Health Areal Data: A Multiple Membership Approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *70*(3), 645-666. https://doi.org/10.1111/rssc.12480

Hama, A., Tanaka, K., Mochizuki, A., Tsuruoka, Y., & Kondoh, A. (2020). Estimating the protein concentration in rice grain using UAV imagery together with agroclimatic data. *Agronomy*, *10*(3), Article 3.

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.

Holloway, J., & Mengersen, K. (2018). Statistical Machine Learning Methods and Remote Sensing for Sustainable Development Goals: A Review. *Remote Sensing*, *10*(9), 1365. https://doi.org/10.3390/rs10091365

Houx III, J. H., Wiebold, W. J., & Fritschi, F. B. (2014). Rotation and tillage affect soybean grain composition, yield, and nutrient removal. *Field Crops Research*, *164*, 12-21. https://doi.org/10.1016/j.fcr.2014.04.010

Ibáñez, M. A., De Blas, C., Cámara, L., & Mateos, G. G. (2020). Chemical composition, protein quality and nutritive value of commercial soybean meals produced from beans from different countries: A meta-analytical study. *Animal Feed Science and Technology*, *267*, 114531. https://doi.org/10.1016/j.anifeedsci.2020.114531

Johnson, O., Diggle, P., & Giorgi, E. (2020). Dealing with spatial misalignment to model the relationship between deprivation and life expectancy: A model-based geostatistical approach. *International Journal of Health Geographics*, *19*(1), 6. https://doi.org/10.1186/s12942-020-00200-w

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Kimm, H., Guan, K., Jiang, C., Peng, B., Gentry, L. F., Wilkin, S. C., Wang, S., Cai, Y., Bernacchi, C. J., & Peng, J. (2020). Deriving high-spatiotemporal-resolution leaf area index for agroecosystems in the US Corn Belt using Planet Labs CubeSat and STAIR fusion data. *Remote Sensing of Environment*, *239*, 111615.

Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of hydrology*, *424*, 264-277.

Kravchenko, A. N., & Bullock, D. G. (2002a). Spatial Variability of Soybean Quality Data as a Function of Field Topography. *CROP SCIENCE*, *42*, 12.

Kravchenko, A. N., & Bullock, D. G. (2002b). Spatial Variability of Soybean Quality Data as a Function of Field Topography: II. A Proposed Technique for Calculating the Size of the Area for Differential Soybean Harvest. *CROP SCIENCE*, *42*, 6.

Lacoste, M., Cook, S., McNee, M., Gale, D., Ingram, J., Bellon-Maurel, V., MacMillan, T., Sylvester-Bradley, R., Kindred, D., Bramley, R., Tremblay, N., Longchamps, L., Thompson, L., Ruiz, J., García, F. O., Maxwell, B., Griffin, T., Oberthür, T., Huyghe, C., … Hall, A. (2021). On-Farm Experimentation to transform global agriculture. *Nature Food*, *3*(1), 11-18. https://doi.org/10.1038/s43016-021-00424-4

Levy, R., & McNeish, D. (2023). Perspectives on Bayesian inference and their implications for data analysis. *Psychological Methods*, *28*(3), 719-739. https://doi.org/10.1037/met0000443

Liang, D., & Kumar, N. (2013). Time-space Kriging to address the spatiotemporal misalignment in the large datasets. *Atmospheric Environment*, *72*, 60-69.

Lindsay, J. B. (2016). Whitebox GAT: A case study in geomorphometric analysis. *Computers & Geosciences*, *95*, 75-84. https://doi.org/10.1016/j.cageo.2016.07.003

Liu, Y., Lv, J., Zhang, B., & Bi, J. (2013). Spatial multi-scale variability of soil nutrients in relation to environmental factors in a typical agricultural region, Eastern China. *Science of The Total Environment*, *450-451*, 108-119. https://doi.org/10.1016/j.scitotenv.2013.01.083

Lobell, D. B., Thau, D., Seifert, C., Engle, E., & Little, B. (2015). A scalable satellite-based crop yield mapper. *Remote Sensing of Environment*, *164*, 324-333.

Lopiano, K. K., Young, L. J., & Gotway, C. A. (2011). A comparison of errors in variables methods for use in regression models with spatially misaligned data. *Statistical Methods in Medical Research*, *20*(1), 29-47. https://doi.org/10.1177/0962280210370266

Maccini, S., & Yang, D. (2009). Under the Weather: Health, Schooling, and Economic Consequences of Early-Life Rainfall. *American Economic Review*, *99*(3), 1006-1026. https://doi.org/10.1257/aer.99.3.1006

Madsen, L., Ruppert, D., & Altman, N. S. (2008). Regression with spatially misaligned data. *Environmetrics*, *19*(5), 453-467. https://doi.org/10.1002/env.888

Martin, N. F., Bollero, A. G., & Bullock, D. G. (2007). Relationship between secondary variables and soybean oil and protein concentration. *Transactions of the ASABE*, *50*(4), 1271-1278.

Mathenge, M., Sonneveld, B. G., & Broerse, J. E. (2022). Application of GIS in agriculture in promoting evidence-informed decision making for improving agriculture sustainability: A systematic review. *Sustainability*, *14*(16), 9974.

Mgendi, G. (2024). Unlocking the potential of precision agriculture for sustainable farming. *Discover Agriculture*, *2*(1), 87. https://doi.org/10.1007/s44279-024-00078-3

Minasny, B., McBratney, A. B., Malone, B. P., & Wheeler, I. (2013). Digital mapping of soil carbon. *Advances in agronomy*, *118*, 1-47.

Moore, I. D., Gessler, P. E., Nielsen, G. A., & Peterson, G. A. (1993). Soil Attribute Prediction Using Terrain Analysis. *Soil Science Society of America Journal*, *57*(2), 443-452. https://doi.org/10.2136/sssaj1993.03615995005700020026x

Moss, S. (2019). Integrated weed management (IWM): Why are farmers reluctant to adopt non-chemical alternatives to herbicides? *Pest management science*, *75*(5), Article 5.

Mourtzinis, S., Borg, B. S., Naeve, S. L., Osthus, J., & Conley, S. P. (2018). Characterizing Soybean Meal Value Variation across the United States: A Swine Case Study. *Agronomy Journal*, *110*(6), Article 6. https://doi.org/10.2134/agronj2017.11.0624

Muhie, S. H. (2022). Novel approaches and practices to sustainable agriculture. *Journal of Agriculture and Food Research*, *10*, 100446.

Mulla, D. J. (2013). Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems engineering*, *114*(4), Article 4.

Nandan, R., Bandaru, V., He, J., Daughtry, C., Gowda, P., & Suyker, A. E. (2022). Evaluating Optical Remote Sensing Methods for Estimating Leaf Area Index for Corn and Soybean. *Remote Sensing*, *14*(21), 5301. https://doi.org/10.3390/rs14215301

Newville, M., Otten, R., Nelson, A., Ingargiola, A., Stensitzki, T., Allan, D., Fox, A., Carter, F., Pustakhod, D., & Ram, Y. (2019). lmfit/lmfit-py 0.9. 14. *Zenodo*.

Nguy-Robertson, A., Gitelson, A., Peng, Y., Viña, A., Arkebauer, T., & Rundquist, D. (2012). Green Leaf Area Index Estimation in Maize and Soybean: Combining Vegetation Indices to Achieve Maximal Sensitivity. *Agronomy Journal*, *104*(5), Article 5. https://doi.org/10.2134/agronj2012.0065

O'Donoghue, T., Minasny, B., & McBratney, A. (2024). Digital Regenerative Agriculture. *Npj Sustainable Agriculture*, *2*(1), 5. https://doi.org/10.1038/s44264-024-00012-6

Oøvergaard, S. I., Isaksson, T., Kvaalc, K., & Korsaeth, A. (2010). Comparisons of two hand-held, multispectral field radiometers and a hyperspectral airborne imager in terms of predicting spring wheat grain yield and quality by means of powered partial least squares regression. *Journal of Near Infrared Spectroscopy*, *18*(4), 247-261. https://doi.org/10.1255/jnirs.892

Parra-López, C., Ben Abdallah, S., Garcia-Garcia, G., Hassoun, A., Sánchez-Zamora, P., Trollman, H., Jagtap, S., & Carmona-Torres, C. (2024). Integrating digital technologies in agriculture for climate change adaptation and mitigation: State of the art and future perspectives. *Computers and Electronics in Agriculture*, *226*, 109412. https://doi.org/10.1016/j.compag.2024.109412

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

Pierre Pott, L., Jorge Carneiro Amado, T., Augusto Schwalbert, R., Mateus Corassa, G., & Antonio Ciampitti, I. (2022). Crop type classification in Southern Brazil: Integrating remote sensing, crop modeling and machine learning. *Computers and Electronics in Agriculture*, *201*, 107320. https://doi.org/10.1016/j.compag.2022.107320

Pilz, J., & Spöck, G. (2008). Why do we need and how should we implement Bayesian kriging methods. *Stochastic Environmental Research and Risk Assessment*, *22*(5), 621-632. https://doi.org/10.1007/s00477-007-0165-7

Plummer, M. (2004). *JAGS: Just another Gibbs sampler*.

Pouliot, G. A. (2023). Spatial econometrics for misaligned data. *Journal of Econometrics*, *232*(1), 168-190. https://doi.org/10.1016/j.jeconom.2021.04.011

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.

Pusch, M., Samuel-Rosa, A., Oliveira, A. L. G., Magalhães, P. S. G., & Do Amaral, L. R. (2022). Improving soil property maps for precision agriculture in the presence of outliers using covariates. *Precision Agriculture*, *23*(5), 1575-1603. https://doi.org/10.1007/s11119-022-09898-z

R Core Team. (2025). R: A language and environment for statistical computing. *(No Title)*.

Ramanantenasoa, M. M. J., Génermont, S., Gilliot, J.-M., Bedos, C., & Makowski, D. (2019). Meta-modeling methods for estimating ammonia volatilization from nitrogen fertilizer and manure applications. *Journal of environmental management*, *236*, 195-205.

Ray, C. L., Shipe, E. R., & Bridges, W. C. (2008). Planting Date Influence on Soybean Agronomic Traits and Seed Composition in Modified Fatty Acid Breeding Lines. *Crop Science*, *48*(1), 181-188. https://doi.org/10.2135/cropsci2007.05.0290

Raymond, C., Suarez-Gutierrez, L., Kornhuber, K., Pascolini-Campbell, M., Sillmann, J., & Waliser, D. E. (2022). Increasing spatiotemporal proximity of heat and precipitation extremes in a warming world quantified by a large model ensemble. *Environmental Research Letters*, *17*(3), 035005.

Reisi Gahrouei, O., McNairn, H., Hosseini, M., & Homayouni, S. (2020). Estimation of Crop Biomass and Leaf Area Index from Multitemporal and Multispectral Imagery Using Machine Learning Approaches. *Canadian Journal of Remote Sensing*, *46*(1), 84-99. https://doi.org/10.1080/07038992.2020.1740584

Ribeiro, M. H. D. M., & Dos Santos Coelho, L. (2020). Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. *Applied Soft Computing*, *86*, 105837. https://doi.org/10.1016/j.asoc.2019.105837

Rockström, J., Williams, J., Daily, G., Noble, A., Matthews, N., Gordon, L., Wetterstrand, H., DeClerck, F., Shah, M., & Steduto, P. (2017). Sustainable intensification of agriculture for human prosperity and global sustainability. *Ambio*, *46*, 4-17.

Roth, L., Barendregt, C., Bétrix, C.-A., Hund, A., & Walter, A. (2022). High-throughput field phenotyping of soybean: Spotting an ideotype. *Remote Sensing of Environment*, *269*, 112797. https://doi.org/10.1016/j.rse.2021.112797

Rozenstein, O., Cohen, Y., Alchanatis, V., Behrendt, K., Bonfil, D. J., Eshel, G., Harari, A., Harris, W. E., Klapp, I., Laor, Y., Linker, R., Paz-Kagan, T., Peets, S., Rutter, S. M., Salzer, Y., & Lowenberg-DeBoer, J. (2024). Data-driven agriculture and sustainable farming: Friends or foes? *Precision Agriculture*, *25*(1), 520-531. https://doi.org/10.1007/s11119-023-10061-5

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *71*(2), 319-392.

Sakamoto, T., Wardlow, B. D., Gitelson, A. A., Verma, S. B., Suyker, A. E., & Arkebauer, T. J. (2010). A Two-Step Filtering approach for detecting maize and soybean phenology with time-series MODIS data. *Remote Sensing of Environment*, *114*(10), Article 10. https://doi.org/10.1016/j.rse.2010.04.019

Salmerón, M., Purcell, L. C., Vories, E. D., & Shannon, G. (2017). Simulation of genotype-by-environment interactions on irrigated soybean yields in the U.S. Midsouth. *Agricultural Systems*, *150*, 120-129. https://doi.org/10.1016/j.agsy.2016.10.008

Samsonova, V. P., Blagoveshchenskii, Yu. N., & Meshalkina, Yu. L. (2017). Use of empirical Bayesian kriging for revealing heterogeneities in the distribution of organic carbon on agricultural lands. *Eurasian Soil Science*, *50*(3), 305-311. https://doi.org/10.1134/S1064229317030103

Santana, D. C., Teodoro, L. P. R., Baio, F. H. R., Santos, R. G. D., Coradi, P. C., Biduski, B., Silva Junior, C. A. D., Teodoro, P. E., & Shiratsuchi, L. S. (2023). Classification of soybean genotypes for industrial traits using UAV multispectral imagery and machine learning. *Remote Sensing Applications: Society and Environment*, *29*, 100919. https://doi.org/10.1016/j.rsase.2023.100919

Santos Hansel, D. S., Schwalbert, R. A., Shoup, D. E., Holshouser, D. L., Parvej, R., Prasad, P. V. V., & Ciampitti, I. A. (2019). A Review of Soybean Yield when Double-Cropped after Wheat. *Agronomy Journal*, *111*(2), Article 2. https://doi.org/10.2134/agronj2018.06.0371

Schabenberger, O., & Gotway, C. A. (2017). *Statistical Methods for Spatial Data Analysis: Texts in Statistical Science* (1.ª ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315275086

Schenatto, K., De Souza, E. G., Bazzi, C. L., Gavioli, A., Betzek, N. M., & Beneduzzi, H. M. (2017). Normalization of data for delineating management zones. *Computers and Electronics in Agriculture*, *143*, 238-248. https://doi.org/10.1016/j.compag.2017.10.017

Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V., & Ciampitti, I. A. (2020). Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agricultural and Forest Meteorology*, *284*, 107886.

Shahhosseini, M., Hu, G., Huber, I., & Archontoulis, S. V. (2021). Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-020-80820-1

Shahhosseini, M., Martinez-Feria, R. A., Hu, G., & Archontoulis, S. V. (2019). Maize yield and nitrate loss prediction with machine learning algorithms. *Environmental Research Letters*, *14*(12), Article 12.

Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2021). Machine Learning Applications for Precision Agriculture: A Comprehensive Review. *IEEE Access*, *9*, 4843-4873. https://doi.org/10.1109/ACCESS.2020.3048415

Sheehan, J., Camobreco, V., Duffield, J., Graboski, M., & Shapouri, H. (1998). *Life cycle inventory of biodiesel and petroleum diesel for use in an urban bus*. National Renewable Energy Lab.(NREL), Golden, CO (United States).

Sheshukov, A. Y., Sekaluvu, L., & Hutchinson, S. L. (2018). Accuracy of topographic index models at identifying ephemeral gully trajectories on agricultural fields. *Geomorphology*, *306*, 224-234. https://doi.org/10.1016/j.geomorph.2018.01.026

Song, X., Yang, G., Yang, C., Wang, J., & Cui, B. (2017). Spatial variability analysis of within-field winter wheat nitrogen and grain quality using canopy fluorescence sensor measurements. *Remote Sensing*, *9*(3), Article 3.

Stan Development Team. (2015). *Stan: A C++ library for probability and sampling, version 2.8. 0*.

Tavares, C., Ribeiro Junior, W., Ramos, M., Pereira, L., Casari, R., Pereira, A., de Sousa, C., da Silva, A., Neto, S., & Mertz-Henning, L. (2022). Water Stress Alters Morphophysiological, Grain Quality and Vegetation Indices of Soybean Cultivars. *Plants*, *11*(4), Article 4. https://doi.org/10.3390/plants11040559

Taylor, J., & Whelan, B. (2007). *On-the-go protein monitoring: A review*. Conference: 4th International Symposium on Precision Agriculture (SIAP07). Federal University of Viçosa, Viçosa.

USGS. (2024). *3DEP LidarExplorer*. 3DEP LidarExplorer. https://apps.nationalmap.gov/lidar-explorer/#/

Vollmann, J., Rischbeck, P., Pachner, M., Đorđević, V., & Manschadi, A. M. (2022). High-throughput screening of soybean di-nitrogen fixation and seed nitrogen content using spectral sensing. *Computers and Electronics in Agriculture*, *199*, 107169. https://doi.org/10.1016/j.compag.2022.107169

Walker, C. K., Assadzadeh, S., Wallace, A. J., Delahunty, A. J., Clancy, A. B., McDonald, L. S., Fitzgerald, G. J., Nuttall, J. G., & Panozzo, J. F. (2023). Technologies and data analytics to manage grain quality on-farm—A review. *Agronomy*, *13*(4), 1129.

White, K., Habib, R., & Hardisty, D. J. (2019). How to SHIFT consumer behaviors to be more sustainable: A literature review and guiding framework. *Journal of marketing*, *83*(3), 22-49.

Wijewardana, C., Reddy, K. R., & Bellaloui, N. (2019). Soybean seed physiology, quality, and chemical composition under soil moisture stress. *Food Chemistry*, *278*, 92-100. https://doi.org/10.1016/j.foodchem.2018.11.035

Wu, Q. (2020). geemap: A Python package for interactive mapping with Google Earth Engine. *Journal of Open Source Software*, *5*(51), Article 51.

Wu, Y.-H., Hung, M.-C., & Patton, J. (2013). Assessment and visualization of spatial interpolation of soil pH values in farmland. *Precision Agriculture*, *14*(6), 565-585. https://doi.org/10.1007/s11119-013-9316-7

Xu, X., Teng, C., Zhao, Y., Du, Y., Zhao, C., Yang, G., Jin, X., Song, X., Gu, X., & Casa, R. (2020). Prediction of wheat grain protein by coupling multisource remote sensing imagery and ECMWF data. *Remote Sensing*, *12*(8), Article 8.

Yaman, M. A., Rattay, F., & Subasi, A. (2021). Comparison of Bagging and Boosting Ensemble Machine Learning Methods for Face Recognition. *Procedia Computer Science*, *194*, 202-209. https://doi.org/10.1016/j.procs.2021.10.074

Yang, F., Zhang, D., Zhang, Y., Zhang, Y., Han, Y., Zhang, Q., Zhang, Q., Zhang, C., Liu, Z., & Wang, K. (2023). Prediction of corn variety yield with attribute-missing data via graph neural network. *Computers and Electronics in Agriculture*, *211*, 108046. https://doi.org/10.1016/j.compag.2023.108046

Yarali, E., & Rivaz, F. (2020). Incorporating covariate information in the covariance structure of misaligned spatial data. *Environmetrics*, *31*(6), e2623. https://doi.org/10.1002/env.2623

Zhang, N., Wang, M., & Wang, N. (2002). Precision agriculture—A worldwide overview. *Computers and electronics in agriculture*, *36*(2-3), 113-132.

Zhong, L., Hu, L., Yu, L., Gong, P., & Biging, G. S. (2016). Automated mapping of soybean and corn using phenology. *ISPRS Journal of Photogrammetry and Remote Sensing*, *119*, 151-164. https://doi.org/10.1016/j.isprsjprs.2016.05.014

Zhong, R., & Moraga, P. (2024). Bayesian Hierarchical Models for the Combination of Spatially Misaligned Data: A Comparison of Melding and Downscaler Approaches Using INLA and SPDE. *Journal of Agricultural, Biological and Environmental Statistics*, *29*(1), 110-129. https://doi.org/10.1007/s13253-023-00559-w

Zhu, L., & Carlin, B. P. (2000). Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion. *Statistics in Medicine*, *19*(17-18), 2265-2278.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), Article 2.