ADVANCED DATA ANALYSIS METHODS TO OPTIMIZE CROP MANAGEMENT DECISIONS

BY

RODRIGO GONCALVES TREVISAN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Crop Sciences
in the Graduate College of the
University of Illinois Urbana-Champaign, 2021

Urbana, Illinois

Doctoral Committee:

Assistant Professor Nicolas Martin, Chair
Professor David Bullock
Professor Adam Davis
Research Assistant Professor Christopher Harbourt

ABSTRACT

The lack of knowledge of limiting factors and optimal management practices at the field level is one of the main reasons for the inefficient use of inputs and low productivity, profitability, and sustainability of agricultural systems. Agricultural research aims to update and improve crop management recommendations to match the spatiotemporal variability and the dynamism of production systems. The advances in remote sensing, precision agriculture, the adoption of information and communication technologies by farmers, and the ability to collect and process large amounts of data create an opportunity to reimagine agricultural research and extension. Advanced data analysis methods are needed to take full advantage of the new data sources and other technological innovations. Therefore, the objectives of this Ph.D. research were i) to develop an image-based high-throughput phenotyping system for evaluating soybean maturity in breeding programs, ii) investigate the spatial variability of optimal input rates in on-farm precision experimentation and the potential economic benefit of site-specific input management, iii) develop a data-driven decision support system for maize in Mexico

The first chapter addresses the need for scalable and accurate methods to develop imagery-based high-throughput phenotyping in breeding programs. Images were acquired with unmanned aerial vehicles twice a week, starting when the earlier lines began maturation until the latest ones were mature. Two complementary convolutional neural networks were developed to predict the maturity date. The first using a single date, and the second using the five best image dates identified by the first model. The proposed neural network architectures were validated using more than 15,000 ground truth observations from five trials, including data from three growing seasons and two countries. The trained model showed good generalization capability with a root mean squared error lower than two days in four out of five trials. Four methods of estimating prediction uncertainty showed potential at identifying different sources of errors in the maturity date predictions. The architecture developed solves limitations of previous research and can be used at scale in commercial breeding programs.

The second chapter demonstrates how on-farm precision experimentation can be a valuable tool for estimating in-field variation of optimal input rates and improving agronomic decisions. Within-field variability of crop yield levels has been extensively investigated, but the spatial variability of crop yield *responses* to agronomic treatments is less understood. Mixed geographically weighted regression models were used to estimate local yield response functions. The methodology was applied to investigate the spatial variability in corn response to nitrogen and seed rates in four cornfields in Illinois, USA. The results showed that spatial heterogeneity of model parameters was significant in all four fields evaluated. On average, the root mean squared error of the fitted yield decreased from 1.2 Mg ha^{-1} in the non-spatial global model to 0.7 Mg ha^{-1} in the geographically weighted regression model, and the r-squared increased from 10% to 68%. The average potential gain of using optimized uniform rates of seed and nitrogen was US$ 65.00 ha^{-1}, while the added potential gain of the site-specific application was US$ 58.00 ha^{-1}. The reported results encourage more research on response-based input management recommendations instead of the still widespread focus on yield-based algorithms.

The third chapter integrates domain knowledge and explainable machine learning methods to optimize management decisions using observational data. The data comes from the Sustainable Modernization of Traditional Agriculture (MasAgro) project in the southern state of Chiapas - Mexico. The dataset was assembled using field observations, including yield, cultivars and management, and environment variables from soil mapping and gridded weather datasets. Random forest models were trained with the dataset and explained up to 75% of the variation. However, the ability of the model to predict crop performance in future weather scenarios was limited. Overall, nitrogen was the management decision that influenced yields the most, with different yield responses depending on the year and variety. This research exemplifies the use of explainable machine learning to offer farmers the opportunity to benchmark their management decisions with peers in similar growing conditions and visualize what would have happened if they made different decisions.

"All models are wrong, but some are useful!"

George Box

ACKNOWLEDGEMENTS

I would like to thank God with whom all things are possible.

I would like to express my sincere gratitude to my advisor Dr. Nicolas Martin for giving me the opportunity to pursue my Ph.D. His expertise and enthusiasm were invaluable not only along with the development of this research but also through my academic formation. I would like to thank my thesis committee: Dr. David Bullock, Dr. Adam Davis, and Dr. Christopher Harbourt, for their insightful feedback and support.

I would like to thank all the Crop Sciences department staff, especially Dianne, Linda, Nathan, and Christina, who helped navigate during the Ph.D. process. I would also like to thank my colleagues German Mandrini, Agustin Alesso, Alexandre Ormiga, Giovani Fontes, Andres Prada, Samuel Fernandes, and Felippe Hoffmann, for their feedback, support, and friendship.

I would like to acknowledge the Don Mario Group (GDM Seeds) for providing the images and part of the ground truth maturity data used in Chapter 1, especially Nathan Schmitz. Also, to the Diers' lab for providing the ground truth maturity data from their field experiments, especially Brian Diers and Osvaldo Perez, who helped immensely in preparing the manuscript.

I would also like to acknowledge the Data-Intensive Farm Management Project under the United States Department of Agriculture - National Institute of Food and Agriculture (USDA-NIFA) Food Security Program that financed the research presented in Chapter 2 under the grant award number 2016-68004-24769. I would also like to express my sincere gratitude to all the farmers who let us run crazy experiments in large parts of their fields and were responsible for providing all the data used in Chapter 2.

I would like to especially thank the ACES Office of International Programs for awarding me the ACES International Global Food Security Fellowship, the CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS) under the project Towards a Digital Climate Smart Agriculture transformation in Latin America, the CGIAR Platform for Big Data in Agriculture under the community of practice Data-Driven Agronomy, the Center for Tropical Agriculture (CIAT) in Colombia, and the International Maize and Wheat Improvement Center

(CIMMYT) in Mexico. The work presented in Chapter 3 would not be possible without this fellowship and the support of these institutions and the people involved in those projects. I would like to acknowledge the exceptional help, support, and welcoming from Hugo Dorado, Juan Camilo, Daniel Jimenez, and Brian King from CIAT, and Andrea Gardeazabal, Samuel Flores, Luis Vargas, and Jorge Garcia from CIMMYT.

I would like to express my sincere gratitude to Saint John's Catholic Newman Center, Bridges International, GLOBE, the Brazilian community, and the Armas, Martinez, and Seufferheld families for their support, welcoming, and greatly enhancing my living experience here.

I would also like to acknowledge “Escola Municipal Rural Produtiva Ranchão”, “Universidade Federal de Mato Grosso”, “Universidade Tecnológica Federal do Paraná”, and “Escola Superior de Agricultura Luiz de Queiroz” (ESALQ – USP) for their contributions in my academic training, especially to Lee Yun Sheng, Onã da Silva Freddi and José Paulo Molin who advised me in this journey.

This thesis is dedicated to my loving family: my mother, Adelaide, my father, Enio, and my sister Roberta. Finally, I could not have completed this dissertation without the love and support of my wife Laura De La Guardia, and my friends here in Illinois and Brazil.

TABLE OF CONTENTS

CHAPTER 1

HIGH-THROUGHPUT PHENOTYPING OF SOYBEAN MATURITY USING TIME SERIES UAV IMAGERY AND CONVOLUTIONAL NEURAL NETWORKS

ABSTRACT

Soybean maturity is a trait of critical importance for the development of new soybean cultivars, nevertheless, its characterization based on visual ratings has many challenges. Unmanned aerial vehicles (UAVs) imagery-based high-throughput phenotyping methodologies have been proposed as an alternative to the traditional visual ratings of pod senescence. However, the lack of scalable and accurate methods to extract the desired information from the images remains a significant bottleneck in breeding programs. The objective of this study was to develop an image-based high-throughput phenotyping system for evaluating soybean maturity in breeding programs. Images were acquired twice a week, starting when the earlier lines began maturation until the latest ones were mature. Two complementary convolutional neural networks (CNN) were developed to predict the maturity date. The first using a single date, and the second using the five best image dates identified by the first model. The proposed CNN architecture was validated using more than 15,000 ground truth observations from five trials, including data from three growing seasons and two countries. The trained model showed good generalization capability with a root mean squared error lower than two days in four out of five trials. Four methods of estimating prediction uncertainty showed potential at identifying different sources of errors in the maturity date predictions. The architecture developed solves limitations of previous research and can be used at scale in commercial breeding programs.

INTRODUCTION

As the most important source of plant protein in the world, soybean (*Glycine max* L.) is widely grown and heavily traded and plays a significant role in global food security (Hartman et al., 2011). In this context, crop breeding aims to increase the grain yield potential and improve the adaptation of new cultivars to environmental changes. Improving traits of interest, such as grain yield, depends on the ability to accurately assess the phenotype of a large number of experimental lines developed annually from breeding populations (Cobb et al., 2013; Liu et al., 2020). However, the labor-intensive and costly nature of classical phenotyping limits its implementation when large populations are used. This may result in breeders not selecting potentially valuable germplasm and reduced genetic gain (Moreira et al., 2019; Morales et al., 2020).

Among the many plant phenotyping tasks, the most critical phenological traits characterized in breeding programs are usually emergence, flowering, and physiological maturity (Reynolds et al., 2020). In soybean, physiological maturity or the R8 stage is defined as the date when 95% of the pods have reached their mature color (Fehr et al., 1977). For soybean, maturity is especially important because, besides defining the crop cycle length, many management decisions are associated with it. The ideal cultivar for a given region is the one that can take full advantage of the growing season to maximize yields, but at the same time avoids delayed harvest, which increases risks and costs. In most cases, if all other characteristics are the same, relatively early-maturity cultivars are preferred. One of the reasons for this preference is for better management of soybean diseases, especially Asian soybean rust. The shorter growing cycle decreases the time for epidemic development, thus preventing yield loss by the disease (Koga et al., 2014). Besides the actual costs, the cycle length is also associated with opportunity costs. The possibility of successful development of a second cash crop or a cover crop is increased when early maturity cultivars are used, which can be an important step towards the sustainable intensification of production (Andrea et al., 2020). The accurate measurement of maturity is also important in breeding trials. Only the performance of experimental lines that have similar maturity dates should be directly compared. This information is also used to take into account the effects that earlier maturing lines may have on the neighboring plots (Reynolds et al., 2020).

Soybean phenology is directly affected by the interactions of photoperiod and temperature, therefore, one observation of cycle length from a single year and location is insufficient to characterize a cultivar. This led to the development of the relative maturity concept, which is a rating system designed to account for all of the factors that affect the number of days from emergence to maturity and allow for comparisons of cultivars that were not directly compared in tests (Alliprandini et al., 2009; Song et al., 2019). Maturity groups are estimated by comparing experimental lines to well-known cultivars grown in the same conditions. The choice of these references is usually guided by published lists of the most stable cultivars and, consequently, of the most suitable check genotypes for each maturity group (Alliprandini et al., 2009; Mourtzinis and Conley, 2017; Zdziarski et al., 2018).

The technological advances in other breeding sciences such as marker-assisted selection and genomic selection, where phenotyping provides critical information for developing and testing statistical models, has increased the demand for phenotypic data resulting in phenotyping becoming the major bottleneck of plant breeding (Araus et al., 2018). In this context, the term high-throughput field phenotyping (HTFP) is used to refer to the field-based phenotyping platforms developed to deliver the necessary throughput for large scale experiments and to provide an accurate depiction of trait performance in real-world environments (Yu et al., 2016). Most HTFP technologies are based on remote sensing, taking advantage of light and other properties that can be measured without direct contact (Araus et al., 2018). Recent advances in proximal remote sensing, in which sensors are usually a few meters from the plants, paired with new sensors and computer science applications, has enabled cost-effective HTFP (Moreira et al., 2019). Among the many options of remote sensing platforms, unmanned aerial vehicles (UAVs) equipped with different sensors have received considerable attention recently. UAVs have become an important approach for fast and non-destructive HTFP due to their growing autonomy, reliability, decreasing cost, flexible and convenient operation, on-demand access to data, and high spatial resolution (Yang et al., 2017; Araus et al., 2018). RGB (red-green-blue) cameras are the most commonly used sensor due to their lower cost and much higher resolution when compared with multispectral cameras (Araus et al., 2018). These factors contribute to the fact that UAVs equipped with RGB cameras are currently the most affordable and widely adopted proximal sensing based HTFP tools (Reynolds et al., 2019; Borra-Serrano et al., 2020).

The costs associated with image capture represent a limited fraction of the overall cost of HTFP. The massive number of images produced and the intense computational requirements to accurately locate images and extract data for corresponding experimental units contribute to a significant increase in the cost of the analysis (Tsaftaris et al., 2016; Reynolds et al., 2019). Routine use of phenotypic data for breeding decisions requires a rapid data turnaround, and image processing remains a significant bottleneck in breeding programs (Morales et al., 2020). Systems for data management, including user-friendly components for data modeling and integration, are fundamental for the adoption of these technologies (Araus et al., 2018). The phenotyping pipeline also has to include metadata and integrate other sources of information following best practices and interoperability guidelines (Schnaufer et al., 2020).

Recently, free and open-source alternatives such as the Open Drone Map integrated into cloud computing platforms have been made available, which helps to reduce the costs of mosaicing the images (Ampatzidis et al., 2020). This makes the construction of the orthomosaic mostly an automated process, which is similar to the needs of many other scientific uses. However, the delineation of experimental units and the extraction of plot-level features poses additional difficulties in processing the information from HTFP platforms (Matias et al., 2020). These challenges have been addressed in recent publications, with optimized methods for semi-automatic detection of the microplots (Khan and Miklavcic, 2019; Tresch et al., 2019) and open-source software packages in python (Chen and Zhang, 2020) and R (Matias et al., 2020). Another contribution that can improve the usefulness of the data collected is the projection of individual microplots generated from the orthomosaic back onto the raw aerial UAV images. This allows the final plot image to retain higher quality and allows the extraction of many replicates from the overlapping images, resulting in several plot images of different perspectives from the same sampling date (Tresch et al., 2019; Moreira et al., 2019). This is also an essential step towards direct georeferencing the geometric position of the microplot in the raw image, avoiding the expenses related to building the orthomosaic and allowing high accuracy with smaller overlaps so that the time and amount of redundant data is minimized (Zhou et al., 2019).

Another strategy to simplify the processing is to move from the image to an aggregated value early in the pipeline. The use of vegetation indices and other averages of reflectance from all pixels in the plot is widespread. From a computer vision perspective, this is the equivalent of using handcrafted features to reduce the dimensionality of the data. Recently, methods that automate feature extraction integrated with the final classification or regression model have been shown to outperform classic feature extraction in many image processing tasks such as image classification/regression, object recognition, and image segmentation (Jiang and Li, 2020). Within machine learning, the term deep neural networks is used to characterize models in which many layers are sequentially stacked together, allowing the model to learn hierarchical features that encode the information in the image in lower dimensions. In this way, the features are learned automatically from input data. Deep convolutional neural networks (CNNs) have become the most common type of deep learning model for image analysis. CNNs are especially well-suited for these tasks because they take advantage of the spatial structure of the pixels. The kernels are shared across all the image positions, which dramatically reduces the number of parameters to be learned, improves computational performance, reduces the risk of overfitting, and requires fewer examples for training. CNNs have been successfully applied in plant phenotyping for plant stress evaluation, plant development, and postharvest quality assessment (Jiang and Li, 2020).

The training of most deep learning models is supervised, thus requiring a great number of training examples with annotated labels. The availability of annotated data is among the main limitations to the use of these advanced supervised algorithms in plant phenotyping problems (Tsaftaris et al., 2016; Araus et al., 2018). For example, the availability of several large, annotated image datasets for plant stress classification accelerated the evaluation of various CNNs for stress phenotyping (Jiang and Li, 2020). Although the number of publicly available datasets and the diversity of phenotyping tasks covered is growing (David et al., 2020; Dobrescu et al., 2020), there are still many tasks that have yet to be addressed. In general, these datasets have been used to compare new CNN architectures and to pretrain CNNs models to be used in transfer learning. However, training a robust model for field applications still requires a great effort to prepare the dataset. For some traits, such as grain yield, ground truth data can only be obtained in the field because the phenotype cannot be directly observed in the image (Maimaitijiang et al., 2020). When the large number of observations needed is not met, strategies such as synthetic data augmentation may be used to improve the robustness of models trained with fewer examples (Jiang and Li, 2020).

In most published research, the features chosen to build maturity prediction models are related to the canopy reflectance. Because pod maturity and canopy senescence are usually well correlated, it is possible to estimate the plant maturity level based on the spectral reflectance (Zhou et al., 2019). However, physiological maturity, defined by the R8 stage, is assigned by the pod maturity and not by the canopy senescence. Delayed leaf senescence, green stems, and the presence of weeds may cause significant errors in the predictions based only on canopy reflectance. This may explain why transformations applied to high-resolution images that extract additional color and texture information may improve the precision and accuracy of the predicted values (Yuan et al., 2019). The robustness of the model may also be affected by variation in reflectance during the acquisition of the images. Factors such as the relative position between the sun and the camera, cloudiness, and the image stitching process that may cause artifacts such as blurred portions of the orthomosaic, are some examples (Zhou et al., 2019).

Increasing the robustness of the model to the factors listed above may require the use of additional features and more observations during the training. The use of synthetic data augmentation could substantially increase the sample size and the variation within the observations. However, the augmented images are still highly correlated, presenting potential problems due to overfitting (Jiang and Li, 2020). Even though the use of specific features and variable selection based on expert knowledge may be preferred when the biological interpretation of the parameters is important (Borra-Serrano et al., 2020), the use of models with automatic feature extraction may increase the accuracy of the model (Jiang and Li, 2020). CNNs have become state of the art in many computer vision tasks, with an increasing number of applications in plant phenotyping tasks such as plant stress detection (Jiang and Li, 2020). Recently, CNNs have also been applied to monitoring the phenology in rice and wheat crops (Wang et al., 2019; Yang et al., 2020). However, this type of advanced model still needs to be validated for predicting physiological maturity in soybean breeding programs using an HTFP approach.

Working with time-series of images poses additional challenges to the phenotyping pipeline, mainly because it is difficult to assure consistency of reflectance values and spatial alignment over time. Some researchers have focused on analyzing individual dates to overcome this challenge, however, these algorithms may lack generalization robustness and lose accuracy drastically when applied in other experiments (Yu et al., 2016). The importance of multi-temporal data to describe crop growth and to predict specific parameters such as maturity is well recognized (Borra-Serrano et al., 2020). The number of available image dates, and the intervals between dates, may also be different from one trial to another. This requires a great deal of flexibility in the model so that it can be tested in other locations. The resolution of the images, which is a function of flight height and sensor characteristics, can also vary and therefore pose additional challenges for the model generalization.

In order to decrease the cost of dating tens of thousands of plots in the field, there is a need to improve the tools to predict the maturity date of soybean progenies in breeding programs. UAV-based imagery is the most promising candidate for this task (Yu et al., 2016; Zhou et al., 2019). However, there are still many challenges and bottlenecks with the tools used to extract the desired information from the images. These tools could be significantly enhanced by incorporating the latest scientific developments in other areas into an integrated, cost-efficient, robust, flexible, and scalable high-throughput phenotyping pipeline. Therefore, the objective of this study was to develop a high-throughput phenotyping system based on aerial images for evaluating soybean maturity in breeding trials.

MATERIALS AND METHODS

Experimental Setup

Five trials were conducted in partnership with public and private breeding programs. Each trial was comprised of various blocks with experimental lines in different generations of the selection cycle (Figure 1.1). A summary of the trials is presented in Table 1.1. The ground truth maturity date (GTM), equivalent to the R8 phenological stage, was recorded by field visits every three or four days, starting at the end of the growing season when the early lines achieved maturity. About 5% of the plots were used as checks, and for these, the maturity group (MG) was known. Only the plots with GTM were used for training and evaluating the models. The total number of plots is included to allow realistic estimates of image acquisition and storage space requirements for different plot sizes and experiment scales.

Table 1.1. Field trials from different breeding programs used for data collection.

Trial	Year	Location	Plot Length (m)	Plot Width (m)	#Plot*	#GTM *
T1	2018	Savoy, IL-USA	2.2	1 × 0.76	9360	9230
T2	2019	Champaign, IL-USA	5.5	2 × 0.76	8608	1421
T3	2019	Arcola, IL-USA	5.5	2 × 0.76	6272	1408
T4	2019	Litchfield, IL-USA	5.5	2 × 0.76	6400	883
T5	2019	Rolândia, PR-Brazil	5.5	2 × 0.50	7170	2680

* #Plot: total number of plots in the trial; #GTM: number of plots with ground truth maturity date observations.

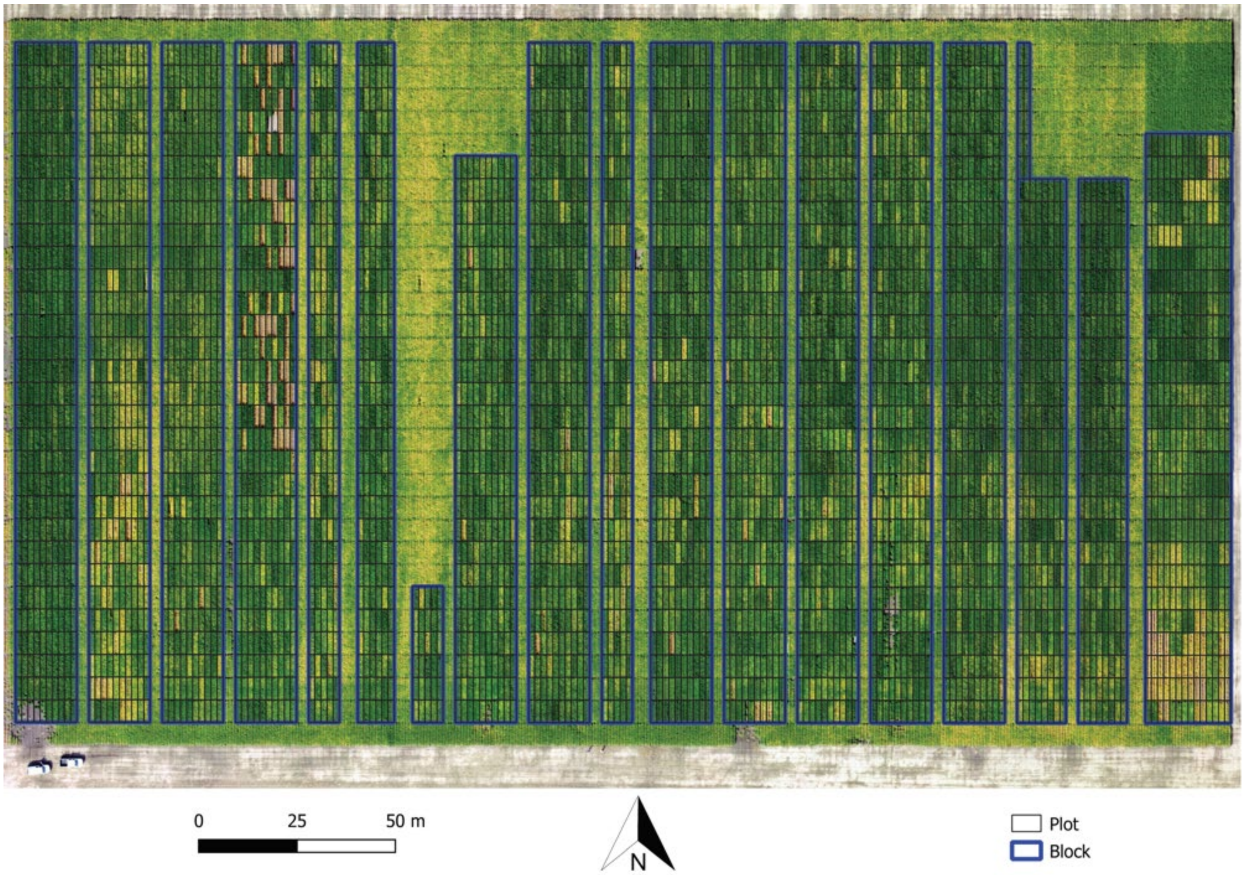

Figure 1.1. Example of soybean breeding field trial (T4) with the layout of plots overlaid on top of the UAV mosaic from images acquired 112 days after seeding.

Image Acquisition

Images were acquired using DJI Phantom 4 Professional UAVs (SZ DJI Technology Co., Ltd., Shenzhen, China), with the built-in 20 MP RGB camera (DJI FC6310) and GPS. The camera has a field of view (FOV) of 84º, and an image resolution of 5472 × 3648 pixels, which were stored as JPEG compressed files with an average size of 8 MB. All images were acquired at a flight height of 80 m, yielding a ground sample distance (GSD) of 25 mm/pixel. The image overlap was set to 80% to the front and 60% to the side. The setting up of the flight plan and the acquisition of the images usually took less than one hour, unless there were clouds shading the trials. In such conditions, the flights were paused and resumed. The acquisition of the images followed a similar schedule of the field visits to record GTM data, with about two images per week recorded from the beginning of leaf senescence in the early lines until the latest lines matured (Figure 1.2). Therefore, the number of flight dates varied according to the range of maturity present in each trial. A summary of the image acquisition step is presented in Table 1.2.

The reduction in data size from the raw images to the image representing each plot for each date is about 20 times. Half of this reduction came from the areas not occupied by plots, such as the paths and borders. However, the most significant reduction of about ten times is from the elimination of overlaps.

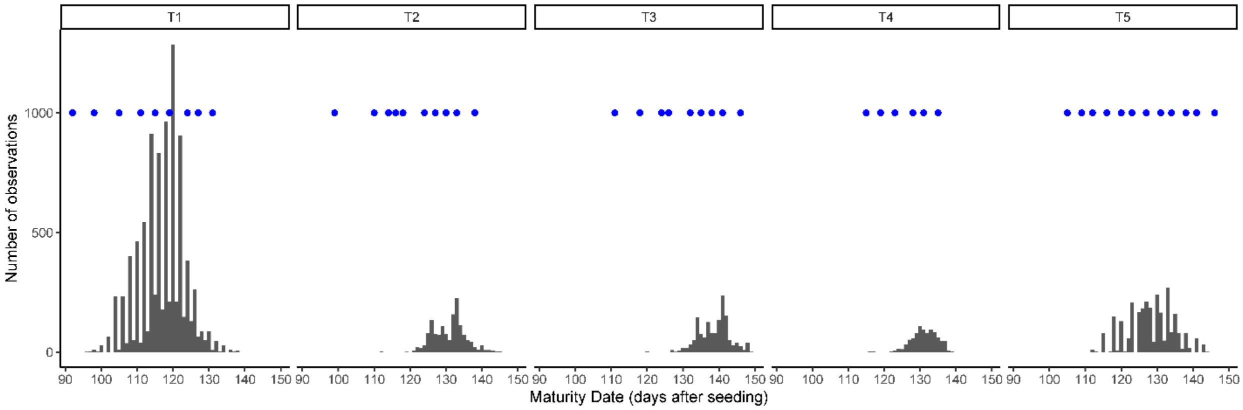

Figure 1.2. Distribution of ground truth maturity dates and image acquisition dates (blue dots) in each trial.

Table 1.2. Image acquisition details and total storage used for each breeding trial.

Trial	Images	Dates	Height (px/plot)	Width (px/plot)	Raw Data (GB)	Processed Data (GB)
T1	100	9	32	96	7.2	0.1
T2	250	10	64	224	20	0.53
T3	150	9	64	224	10.8	0.32
T4	200	6	64	224	9.6	0.22
T5	200	12	40	224	19.2	0.3

Image Processing

After the acquisition, the images were processed using the commercial photogrammetry software (Metashape v1.6, Agisoft LLC, St. Petersburg, Russia). The images were matched with the high accuracy setting, followed by the construction of a dense cloud, the digital elevation map, and the orthomosaic. A total of 12 to 18 ground control points (GCPs) were used in each trial. The targets were placed in the field before the first flight and kept in place until the last flight. The coordinates of the markers were extracted from the first date orthomosaic and used in all subsequent dates. In this way, the points are not necessarily globally accurate, but they ensure the temporal consistency of the images. The first image was also used for manual alignment of the trial layout using QGIS software (QGIS Development Team, 2020). The georeferenced orthomosaic was exported to a three-band (RGB) GeoTIFF file and used to extract the image for each plot using the python packages "geopandas" and "rasterio". Each individual orthophoto was also exported and used to extract replicated observations for each plot.

Resolution

Another important aspect of the images that may affect the model is resolution. Images with downsampled resolution simulating a GSD of 50, 100, and 750 mm/pixel were used to train and compare models. The images were resized accordingly and then compressed to JPEG. For training the model, after decompressing the images, they were scaled back to the original resolution in order to use the same model architecture (Figure 1.3). The visual difference between images with a GSD of 25 and 50 mm/px is very subtle. With 100 mm/px, the difference becomes more evident. The images at 750 mm/px lose all texture information. These were used to help understand the importance of color versus texture and other high-level features.

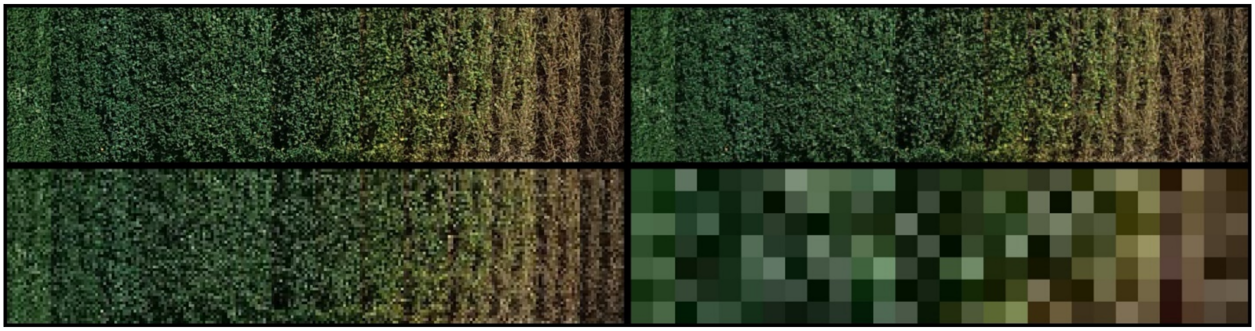

Figure 1.3. Time series of plot images with resolution of 25 mm/px (**top left**), 50 mm/px (**top right**), 100 mm/px (**bottom left**), and 750 mm/px (**bottom right**).

Data Augmentation

One of the disadvantages of using low-cost RGB sensors is their sensitivity to variation in light conditions (Figure 1.4). This motivated the comparison of different data augmentation strategies to improve the model's robustness. The first type of image augmentation consisted of digital transformations of the images by applying variation in contrast and luminosity. On the other hand, the availability of many replicates from each plot may be seen as more natural data augmentation. The availability of many replicates can reproduce geometric errors, distortions, blur, and shadow effects that are hard to reproduce with synthetic data augmentation. Therefore, three different strategies of augmentation were compared: no augmentation, synthetic data augmentation, and using the image replicates. At this time, the image digital numbers stored as

8bit integers were converted to 32-bit floats and scaled from the original range (0–255) to have zero mean and unit variance.

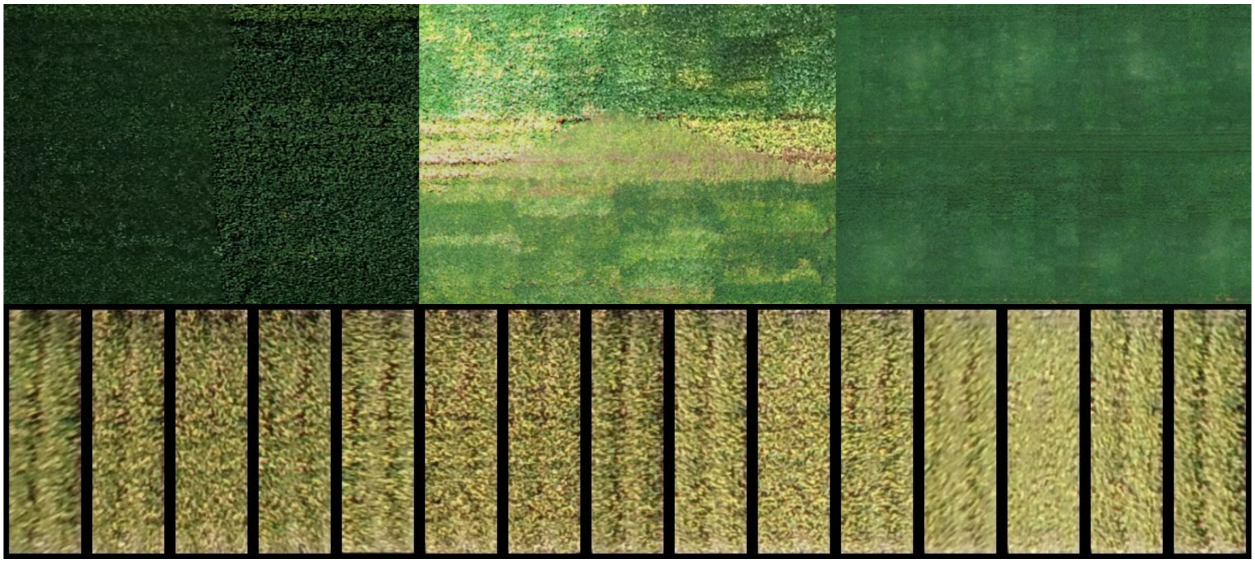

Figure 1.4. Examples of image variations caused by shadows, out of focus images and direct reflection of sunlight (**top**), and differences found among replicated images of the same plot (**bottom**).

Model Development

The model was developed with two steps: The architectures used are referred to as single-date (SD) and multi-date (MD) models. In the first step, the model takes one image and predicts the maturity date. The variable ground truth difference (GTDiff), was calculated to represent the difference between the GTM date and the image acquisition date. A set of SD models were trained using 10-fold cross-validation with GTM data for each trial. The predictions in the test set (PREDDiff) were then used to calculate the average root mean squared error (RMSE) for each trial:

$$RMSE = \sqrt{\left(\frac{1}{n}\right)\sum_{i=1}^{n}\left(DOY_{pred} - DOY_{obs}\right)^2} \qquad [\ 1.1\ ]$$

where: DOY_{pred} and DOY_{obs} are the days of the year in which maturity was predicted and observed, respectively. This allowed the estimation of which GTDiff interval provided the best accuracy in the prediction. The image with the PREDDiff closer to the best GTDiff, and the two images acquired immediately before and after were selected for the next step. The MD model uses the features extracted by the SD at the layer before the predictions, instead of running the model again over the full images, which reduces the number of parameters to be trained. In this way, the

SD model, which has more parameters, can be trained with a greater number of observations and data variation, while the MD model only uses a small number of extracted features and few parameters to refine the prediction.

Single-Date Model Architecture

Based on the layers used and the intention behind their use, the architecture for the SD model can be divided into two groups. In the first group, each block contains a 2D convolution with 3 × 3 kernels, a max-pooling layer that halves the number of pixels in the output, a dropout layer, and a rectified linear activation function (RELU) activation. The convolutions are zero-padded to keep the output sizes the same as the inputs. This block is repeated sequentially five times. Therefore, the output has its spatial dimensions reduced by a factor of 2^5 or 32 times. The dimensions shown in Figure 1.5 are valid for input sizes used in the largest plots. The main purpose of this group of operations is to extract meaningful spatial information and condense it in a lower resolution representation. The next block contains only convolutions with 1 × 1 kernel sizes followed by a dropout layer. Therefore, only the different features of the same pixel are used to calculate the values in the next layer. This block is repeated sequentially four times to obtain the output. The output is then subtracted from the image acquisition date to generate the prediction. This second block does not change the spatial dimension of the output, but it forces the information to be represented by lower-dimensional spaces since the number of channels is being reduced. The result from the layer immediately before the output will be used as features to the temporal model. The reasoning behind the choice to use 1 × 1 convolutions instead of flattening the features was to conserve the variability within the plot to be used in one of the estimates of model uncertainty later. By subtracting the image acquisition date, internally, the model is learning to estimate the difference between the maturity date and the date the image was taken.

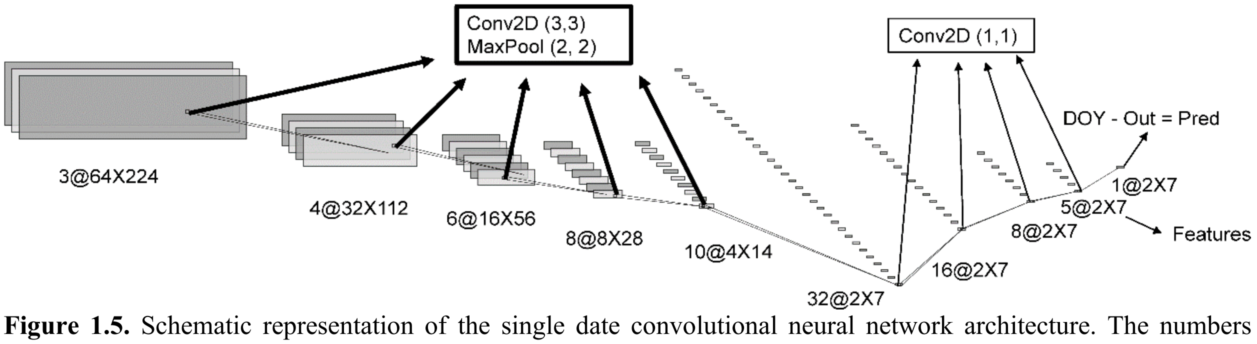

Figure 1.5. Schematic representation of the single date convolutional neural network architecture. The numbers represent the dimensions of the tensors and the names in the boxes are the operations applied.

Multi-Date Model Architecture

The architecture for the MD model was developed to operate over groups of five images, selected from the results of the SD model. The difference between the day of the year (DOY) of each image and the DOY of the central image was concatenated as an additional feature for each image. The difference date from the center image is always zero and can be omitted. However, it is easier to keep it and have all tensors with the same dimensions. Therefore, six features from each of the five images were concatenated into the 30 features that were used as inputs in the MD model. In case the acquisition dates span through two different years, as happens, for instance, in the South Hemisphere where maturity starts in December, the DOY from the previous year can be negative, or on the contrary, it can be extended beyond 365 for the next year. It is also possible to use days after planting or emergence instead of the day of the year. Because the value is subtracted before entering the model and is added back at the end, it is only the intervals that matter. The architecture used in the MD model is straightforward and follows the same layers of the second block in the single date model (Figure 1.6). To keep the number of parameters to be trained to a minimum, the convolutions with 1 × 1 kernel sizes followed by a dropout layer were repeated sequentially three times. The output is then subtracted from the DOY of the central image to generate the final prediction. The order in which the DOY is subtracted and then added back may not be very intuitive. However, this is necessary to keep the same relationship when the difference is greater because the image was taken earlier or when the soybean line presents delayed maturity.

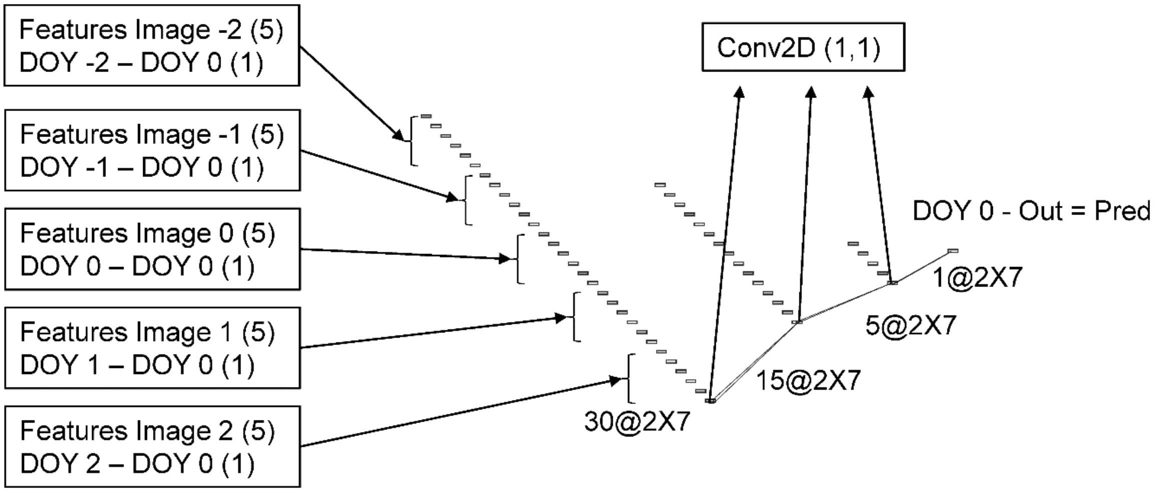

Figure 1.6. Schematic representation of the multi-date convolutional neural network architecture. The numbers represent the dimensions of the tensors, and the names in the boxes are the operations applied. The DOY stands for the day of the year.

Model Parameters

The distribution of the parameters in each step of the model is presented in Table 1.3. The total number of parameters for the full model was 5682, which characterizes a small and light-weight model, with more observations available than parameters to be estimated. This number is the same independent of the size of the input images. The number of parameters in the SD model was 5131, while the number of parameters in the MD model was 551. The last number represents the effective samples to train the MD model, which is about 10% of the available data to train the SD model.

Table 1.3. Details of model architecture and number of parameters in each layer.

Layer	Kernel Dim	Tensor Shape	Param #
Conv2D-S1	[3,3,3]	[-1, 3, h, w]	112
Conv2D-S2	[4,3,3]	[-1, 4, h/2, w/2]	222
Conv2D-S3	[6,3,3]	[-1, 6, h/4, w/4]	440
Conv2D-S4	[8,3,3]	[-1, 8, h/8, w/8]	730
Conv2D-S5	[10,3,3]	[-1, 10, h/16, w/16]	2912
Conv2D-S6	[32,1,1]	[-1, 32, h/32, w/32]	528
Conv2D-S7	[16,1,1]	[-1, 16, h/32, w/32]	136
Conv2D-S8	[8,1,1]	[-1, 8, h/32, w/32]	45
Conv2D-S9	[5,1,1]	[-1, 5, h/32, w/32]	6
Total SD		[-1, 1, h/32, w/32]	5131
Conv2D-M1	[30,1,1]	[-1, 30, h/32, w/32]	465
Conv2D-M2	[15,1,1]	[-1, 15, h/32, w/32]	80
Conv2D-M3	[5,1,1]	[-1, 5, h/32, w/32]	6
Total MD		[-1, 1, h/32, w/32]	551
Total			5682

Model Training

The training and testing were performed in a computer equipped with an Intel i7 processor (Intel Corporation, Santa Clara, CA, USA) and an NVIDIA Quadro P4000 GPU (NVIDIA, Santa Clara, CA, USA) with 8GB memory using the PyTorch deep learning package v. 1.5 (Paszke et al., 2019). The Adam optimizer, with a learning rate of 0.001 was used. The RMSE was used as the loss function (Equation [1.1]). The models were trained using a 10% dropout rate. The models were trained to a maximum of 100 epochs, using early stopping criteria to monitor the validation set and stop training after the loss did not decrease for 10 consecutive epochs. The architectures and hyper-parameters were fine-tuned based on the amount of data available and the overall results in the validation sets.

Model Validation

The dataset was split into three different sets used for training, validation, and testing. The validation set is primarily used for early stopping the model. All metrics presented are calculated over the test set. All comparisons were made using 10-fold cross-validation so that all data were evaluated in all sets. The data split was set to 80% for the training set, 10% for the validation set, and 10% for the test set. The split was fully randomized, which represents the most common method used in the literature. The models trained in one trial were also tested in all other trials. Testing in different trials assures more independence of the testing set and reflects a more desirable model.

Model Uncertainty

In the proposed architecture using only convolutional layers, every 32 × 32 pixels in the input will produce one pixel in the output. The final prediction is taken as the average of the pixels in the prediction. The standard deviation of the predictions is used as an estimate of model uncertainty due to within plot variability. As a consequence of the 10-fold cross-validation, there were ten resulting models for each trial. The standard deviation of these predictions was also evaluated as a metric of uncertainty.

The use of replicated images was also evaluated at test time to estimate the uncertainty caused by variation in light intensity and the overall aspect of the images. This also reflects some of the uncertainty due to the geometric differences in the images, since the distortions are greater for plots close to the borders of the images. Finally, multiple predictions with dropout layers enabled at test time were also used to estimate the uncertainty of the model parameters and architecture. The standard deviation of the predictions with the image replicates, and dropout enabled was computed with 10 random initializations for each plot and method. The four estimates of uncertainty were compared with the average error at a trial level and also correlated to the absolute error of each plot.

RESULTS

Single Date Model

In four of the five trials, the lowest RMSE was observed when the images were acquired about one week before maturity, while for T5 the lowest error was obtained when the images were taken about two weeks prior to maturity (Figure 1.7). Looking at the images of T5, it was noted that in many plots the plants were lodged on the neighboring plots. Furthermore, it was noted that weed growth occurred simultaneously with the crop senescence, and most importantly, leaf retention after pod senescence. These factors contributed to larger errors when the used images are taken closer to senescence. So, even though under optimal experimental conditions, images close to maturity would be preferred, the confounding factors could affect the predicted values and increase the error. When considering all data, the errors remain relatively low for about 12 days before maturity; outside of this range, the errors increase substantially. Based on these observations, the value of the GTDiff was set to −6, meaning that from all the available image dates, the one that predicted maturity would occur about 6 days after the acquisition was used as the center image. The two images acquired immediately before and after this center image usually fell within this 12 day time window. The choice of five images was based mostly on the minimum number of images usually available for the trials. Choosing a fewer number of images may degrade the performance; this is because at prediction time the GTM value is unknown, and the estimates from individual images are used to find the center image. The use of more images confers robustness to the model, in case the choice of images was not optimal.

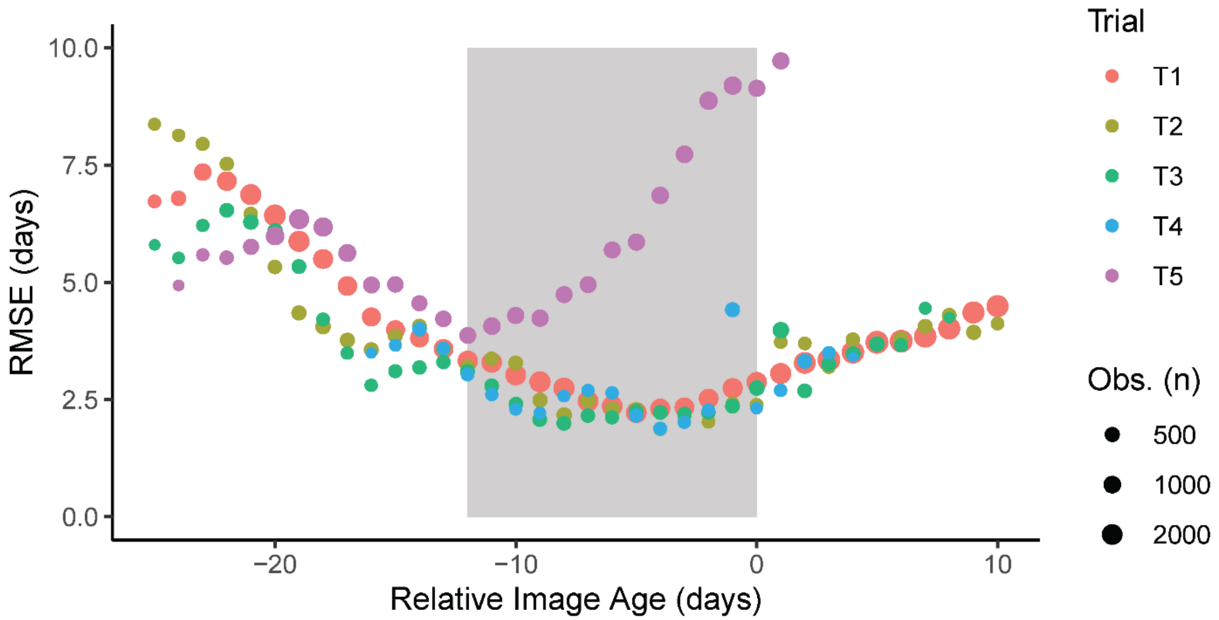

Figure 1.7. Prediction performance measured by the root mean squared error (RMSE) as a function of the difference between the image acquisition date and the ground truth maturity date. The shaded area represents the time window comprising the five images with the least error.

Overall Performance

The overall performance of the models trained and evaluated within the same trial indicated an RMSE inferior to 2 days in all trials except T5, in which the RMSE was about 3 days (Figure 1.8). The lower performance in the last trial is attributed to the lower quality of image acquisition, with more shadows, and to the higher frequency of lines with leaf retention. The performance of models trained in other trials and seasons varied among trials. For most cases with high RMSE there was a bias of a few days in the distributions of predicted and observed values. This could be due to some offset in the relationship of leaf senescence and pod senescence caused by environmental factors and their interaction with the genotype. Part of this bias may also be due to differences in the GTM data acquisition, since the maturity date is an estimate subject to human error. The bias in the raw predictions was corrected using the information from the reference check plots, which greatly reduced the extremely high values of RMSE.

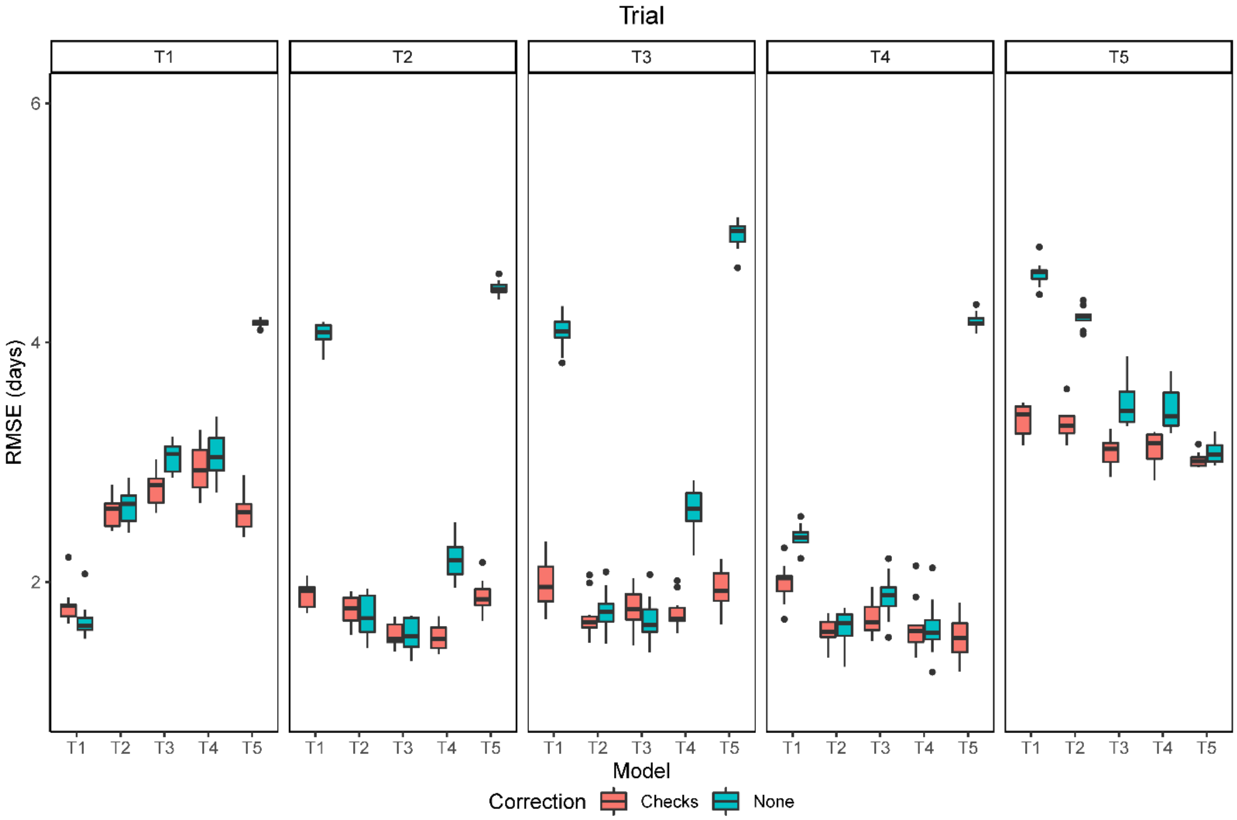

Figure 1.8. Prediction performance measured by the root mean squared error (RMSE) for the raw model outputs and after the correction using the check plots.

When evaluating the RMSE of the adjusted maturity dates, the models showed good generalization (Figure 1.9). It can be noted that the number of model parameters and dropout were effective at avoiding overfitting, since the validation loss did not show any trend to increase within the number of epochs used. The RMSE values were lower when the conditions were similar, but the increased errors when the conditions of the trial changed. For example, all models performed well in trials T2, T3, and T4, which had good quality images and no confounding factors in the trial. However, all the models that were trained in other trials, had higher errors in T1. One reason for that was due to the emergence of a new generation of seedlings after the harvest of the earlier maturing lines. This caused some plots to be green again in the last two acquisition dates. Even though this effect added a low error, considering that the predictions in the single date model were good enough to choose the early images, under other conditions, few large errors could cause an overall increase in the RMSE.

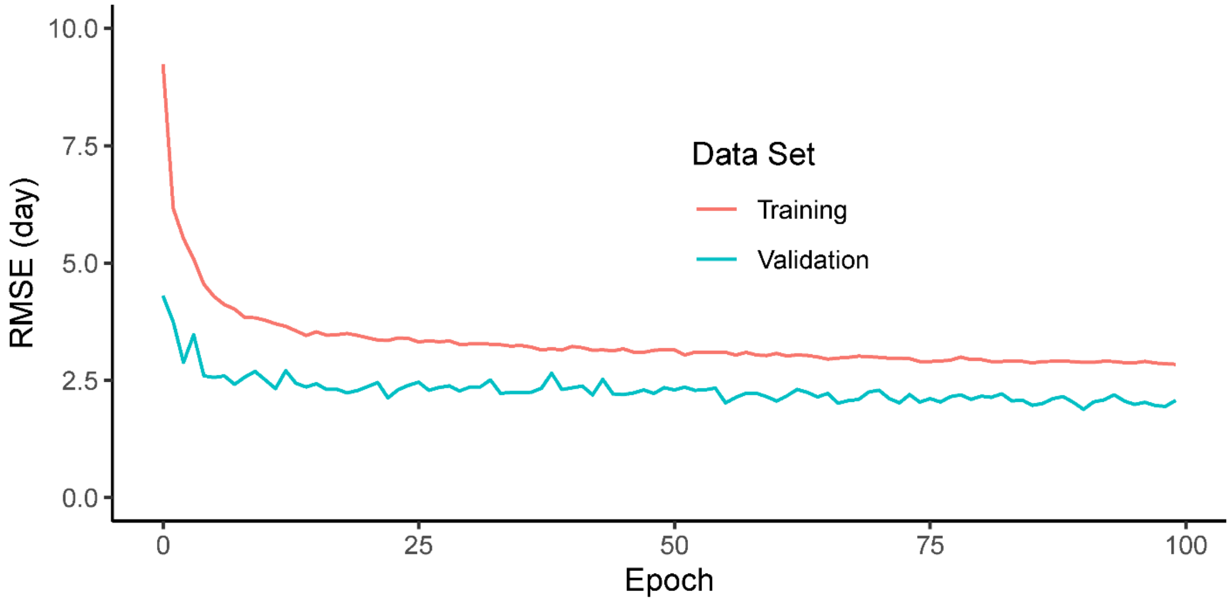

Figure 1.9. Prediction performance measured by the root mean squared error (RMSE) as a function of the number of iterations (epochs) for the training and validation data sets.

Resolution

The effect of resolution was small, resulting in similar model performance in most trials with variations between 25 and 100 mm/px (Figure 1.10). The most significant increase in the RMSE was observed in T1 with the lowest resolution (750 mm/px). This shows that the features learned by the model in T1 depend on the texture of images and not only on the color. For the other trials, the small differences may be related to the number of observations used to train the model, which were five to 10 times fewer than what was available for T1. The trials in which the quality of resolution was less important also had more problems with out of focus images such as the examples shown earlier (Figure 1.4). It is also important to note that even with the best resolution (25 mm/px), the pods cannot be distinguished from the leaves, which would be necessary to improve the models when germplasm expressing leaf retention is present in the trials.

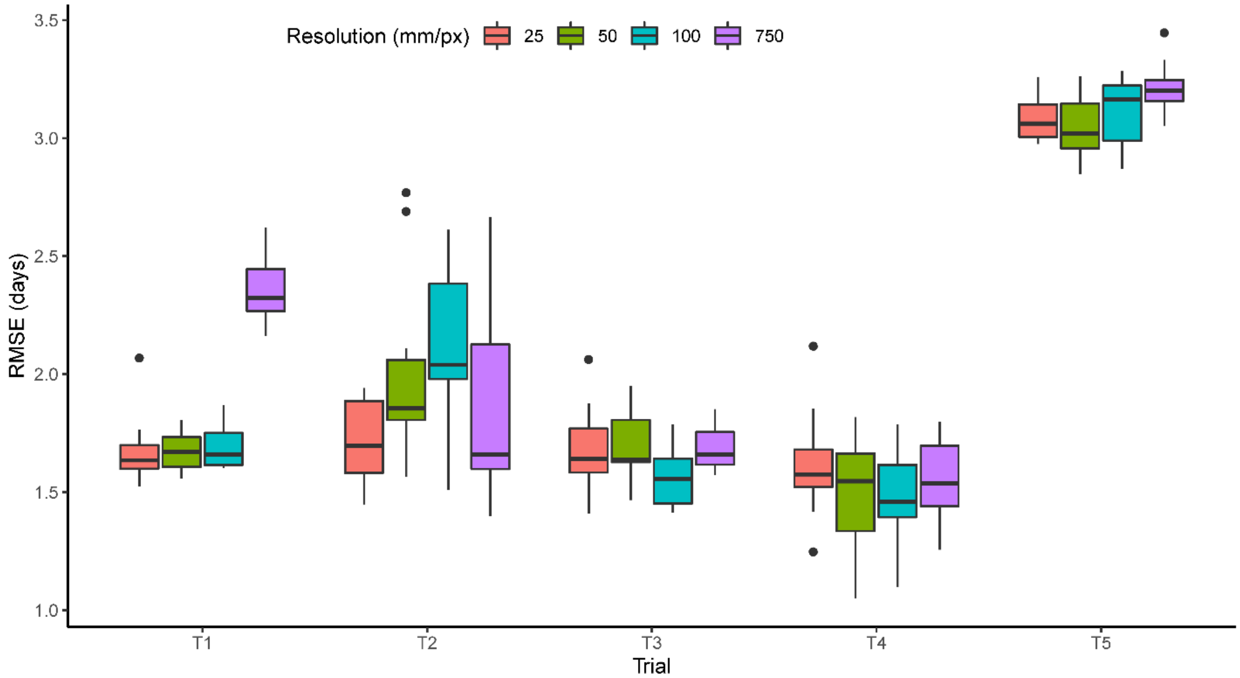

Figure 1.10. Prediction performance measured by the root mean squared error (RMSE) as related to simulated image resolutions.

Data Augmentation

The two strategies of data augmentation used to train the models did not improve the results, compared to no data augmentation (Figure 1.11). Overall, the use of synthetic augmentation decreased model performance when it was evaluated at the same trial, and even more, when it was evaluated in the other trials. The use of the image replicates had mixed results, with increased generalization when the model was tested in other trials in a few cases, but with decreased performance being still more frequent. These results give further evidence to what was observed from the image resolution analysis. Since the model relies mostly on the average color of each image, applying augmentation techniques that change the color of the images (brightness, contrast), leads to a decrease in the model accuracy. In contrast, the augmentation technique has had more success in other computer vision problems, where more complex features related to image texture and shape of objects are more important than color.

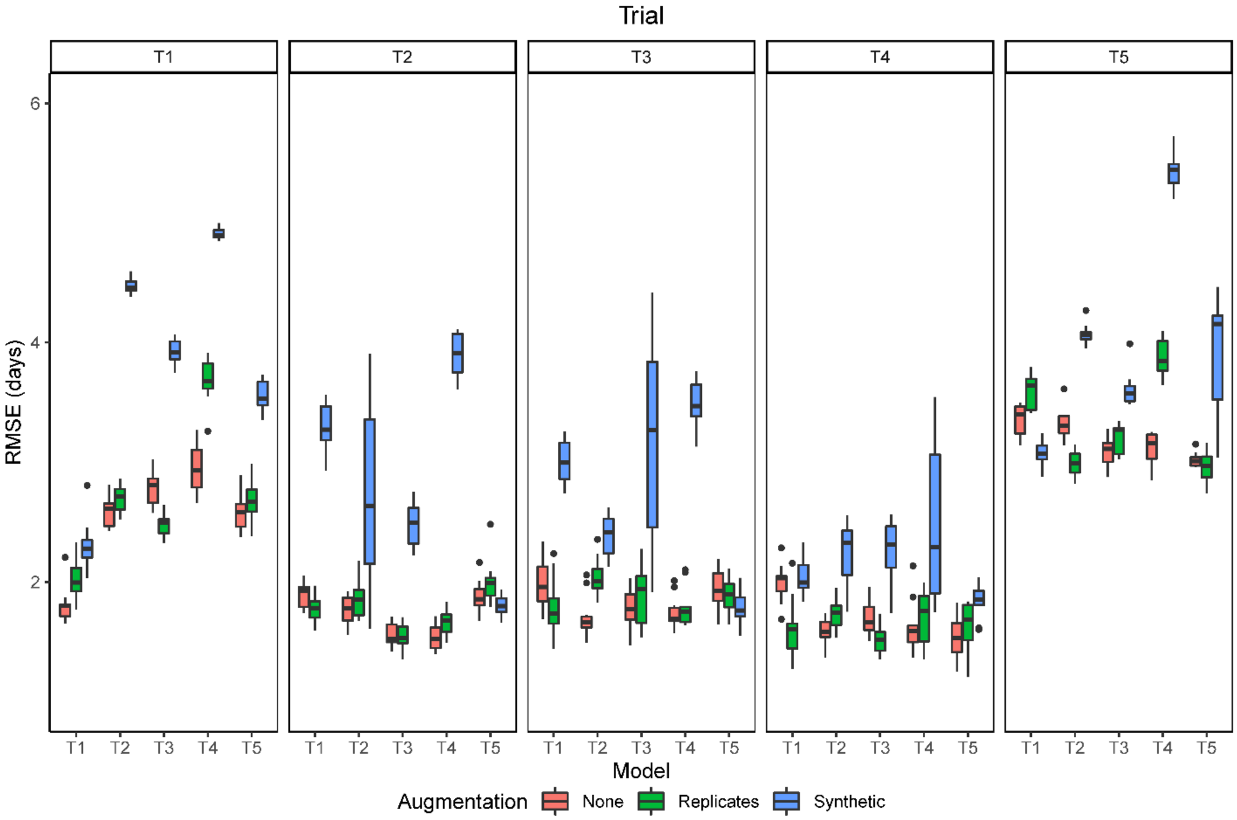

Figure 1.11. Prediction performance measured by the root mean squared error (RMSE) as related to different data augmentation strategies.

Uncertainty

The standard deviation of the within-plot predictions (spatial) was more related to the data used for training than the trial in which the predictions were made (Figure 1.12). The values were lower than 1 day for models trained in trials with bigger plots and higher than 1 day for the models trained in T1 and T5. The standard deviation of using models trained with different subsets of the data (folds) showed a clear difference towards lower RMSE when the model included data from the same trial and when it did not. There was also a distinction between two groups of trials, as models trained in T2, T3, or T4 had lower variation when tested within this group but higher values when tested in T1 or T5, and vice-versa. The standard deviation of using image replicates was lower than 1 day for all trials except in T2, for which its values were higher than the uncertainty estimated by other methods. The standard deviation of predictions with dropout layers enabled was usually higher than the other methods, but similar for all trials and models.

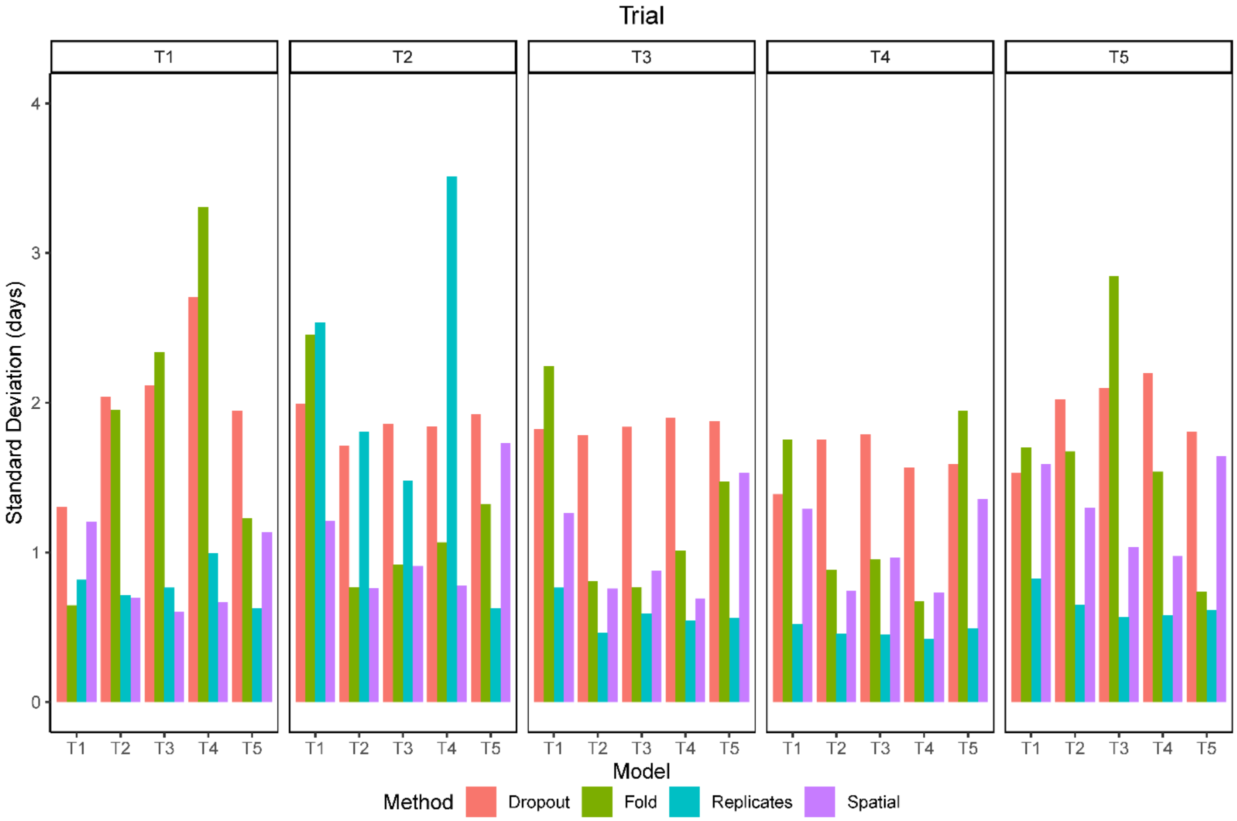

Figure 1.12. Standard deviation of maturity predictions using different methods to estimate uncertainty.

The correlations between the standard deviation of predictions and the absolute error varied from −0.1 to 0.3 depending on the trial and the model (Figure 1.13). In most scenarios the correlation was positive, although some negative values were observed, mostly when using dropout. The overall correlation was higher in T1 and lower in T2, independently of the method. Using dropout presented mixed results, with the best correlation when the model trained in T2 was used in T3. Using the replicates produced the best results in T1 and T5.

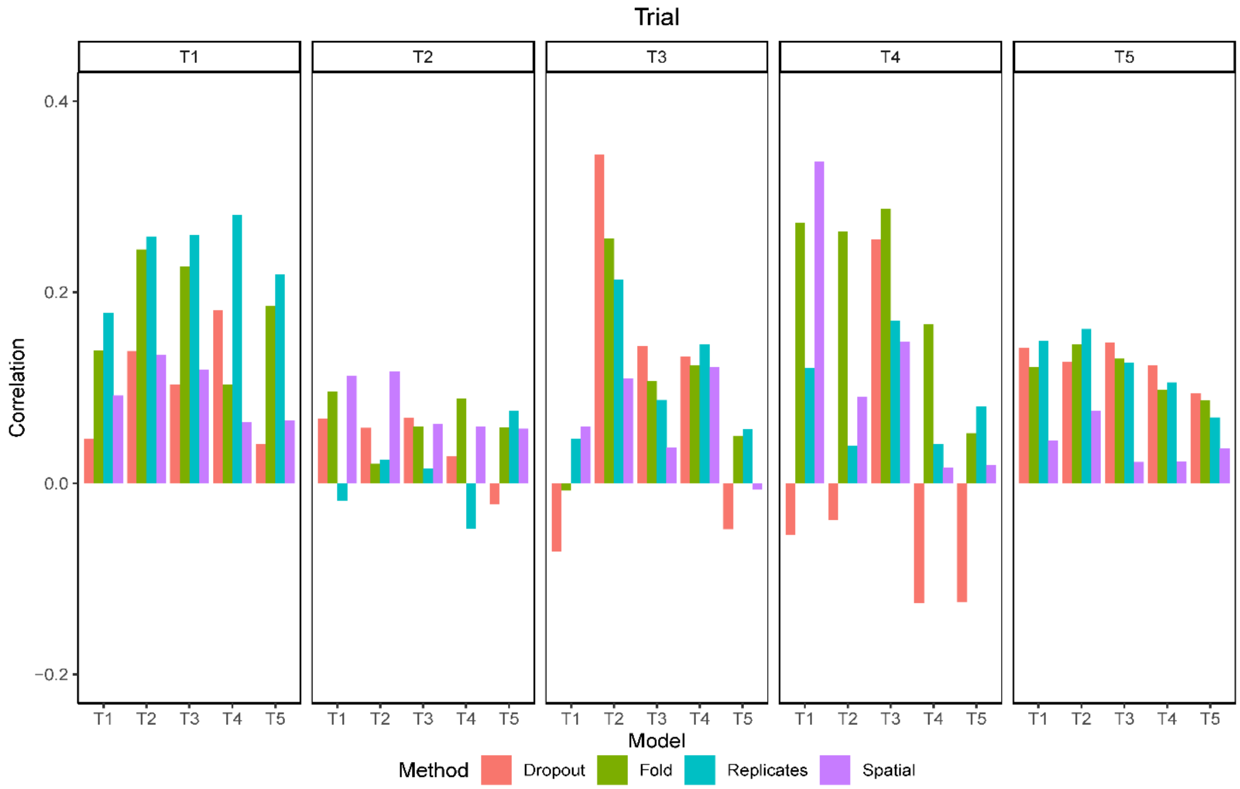

Figure 1.13. Correlation between the standard deviation and the absolute error of maturity predictions using different methods to estimate uncertainty.

DISCUSSION

The maturity prediction was framed as a regression model, aiming to predict the maturity date as a continuous variable, instead of classifying each plant row as mature or immature for a given date (Yu et al., 2016). This eliminates the need for post-processing steps before getting the final result. It also makes it easy to include local information from the check plots in a simple linear regression to account for the environmental factors and assign the maturity group. Reporting the results in terms of the RMSE enables a better evaluation of the model than using classification accuracy, as images taken far from the maturity date are easier to classify but do not contribute to improved model performance.

The overall performance of the model was superior to what has been reported in previous studies. One study, using partial least square regression (PLSR) and three vegetation indices to predict maturity in a diverse set of soybean genotypes, achieved an RMSE of 5.19 days (Christenson et al., 2016). Another recent study, also using PLSR models and 130 handcrafted features from five-band multispectral images, achieved an RMSE of 1.4 days (Zhou et al., 2019).

However, this study used 326 GTM observations with a range of maturity dates of only 10 days, which makes low RMSE easier to achieve. The relatively low importance of image resolution, which is an indicator of the importance of using CNN as feature extractors, shows that this was not the main reason to explain the good performance of the model.

The CNN model can learn how to extract the best combination of features. This flexibility would allow using the model for the extraction of many traits of interest at the same time. For example, the same model could be trained to predict maturity date, senescence rate, lodging and pubescence color. The importance of image resolution and the automated feature extraction with CNNs was demonstrated in a similar study in rice (Yang et al., 2020). In that study, the accuracy of the phenological stages estimation was higher with image resolutions of 20–40 mm/px and decreased sharply when they were reduced to 80–160 mm/px. The maturity in rice is observed in the panicles, which are at the top of the canopy, more visible in the images than the soybean pods, which are in the middle of the canopy. Therefore, it is likely that the best resolution tested in this work (25 mm/px) is still too coarse to allow the model to learn any feature specific to the pods, which is an explanation for why there was little impact of reductions in resolution. A future research direction could be to evaluate the importance of much higher resolutions, which could be obtained with lower flying altitudes or using autonomous ground vehicles (Young et al., 2019).

Contrary to the expected, using replicated observations from different images of the same plot did not increase the model performance. More surprising, applying synthetic data augmentation markedly decreased the model performance in most cases. This result is mostly attributed to the relative importance of color, rather than more complex plant features. Another reason for the low performance when using augmentation may be the simultaneous use of dropout. Some works have shown that for most models there is an equivalence between dropout and data augmentation (Zhao et al., 2019), both introducing some randomness to reduce the risk of overfitting. Since dropout was used in all models, it is possible that the combination of dropout and data augmentation created excessive randomness, reducing the effectiveness of the model training. Considering that dropout is easier to implement and does not require assumptions about what types of augmentation are meaningful, this would be preferred instead of data augmentation. However, a more thorough evaluation of hyper-parameters could be done in future research to confirm these findings.

Developing a model with low prediction error using RGB images makes it more likely to be used due to the low cost. Besides, an RMSE of about 1.5 to 2.0 days is usually considered the acceptable limit in breeding programs (Zhou et al., 2019). Considering that errors above this limit were observed in T5, the use of multispectral images could provide better results when leaf retention is a significant concern. The challenge to correctly predict maturity in plots where plants with mature pods still retain green leaves has been previously reported (Yu et al., 2016). This type of error is more important than a random error because some lines consistently would have higher errors than others, possibly affecting the selection decisions. Future works to predict physiological maturity should consider foliar retention as a trait to be analyzed. Another consideration is what stage maturities should be predicted or visually rated in breeding programs. Breeders will develop and test tens of thousands of experimental lines annually and evaluate them in small plots. It is very labor-intensive to evaluate all of these lines visually for maturity, and it is not critically important to obtain accurate maturity estimates at this stage as the estimates are used to place lines in tests with similar maturities. Predicting maturities with a UAV would most benefit breeding programs at this stage. At later stages of breeding programs, more accurate estimates are needed so that the maturity groups of cultivars can be determined.

One particularity of the proposed architecture is the use of only convolutional layers instead of using fully connected layers for the final prediction. Although this is common in semantic segmentation tasks, it is less used for regression tasks. The goal of applying this strategy in this context was also different. Rather than improving performance, the main purpose was to add flexibility and to estimate prediction uncertainty due to within plot variability. This was demonstrated in Figure 1.13, and was helpful to identify the sources of prediction errors in some plots (Figure 1.14). In a similar way, the use of image replicates also identified an overall higher uncertainty in T2 and was positively correlated with errors of individual plots in T1. Therefore, the different methods of uncertainty estimation can be used for two different purposes. The first is to evaluate the overall quality of the images and procedures used at the trial level, which can identify problems with image stitching or radiometric calibration. The second use is to select individual plots in which the error is likely to be higher, which should be targeted for new data acquisition in order to improve the model.

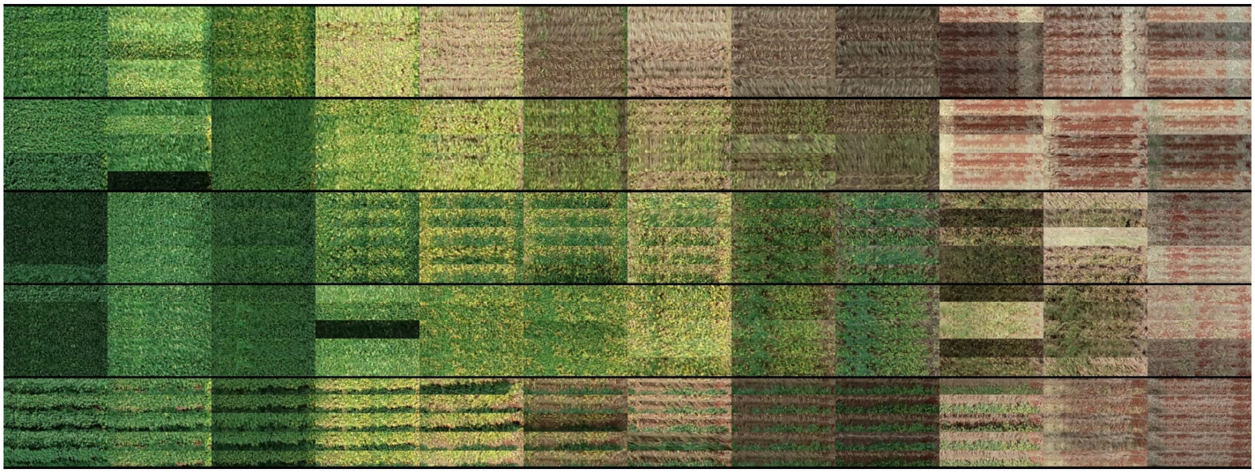

Figure 1.14. Examples of replicated images from trial T5 illustrating leaf retention, weeds, and influence from neighboring plots.

Another source of uncertainty comes from the imprecision of GTM ratings, and is related to the observation frequency, the experience level of the people collecting the data, and the number of people taking notes for the same field. In order to estimate this source of uncertainty it would be necessary to conduct independent maturity assessments by different people in the same plots with a higher frequency of field visits, ideally daily. This is beyond the scope of this work and is left as a suggestion for future research.

Most of the processing time is spent preparing the images for each plot and training the models, but making the predictions is actually very fast. With more than one thousand predictions per second, in the hardware used, using the GPU, this shows the potential scalability of the method once other bottlenecks in image processing are solved. Fast predictions are also important to enable the test time augmentation and evaluate the model uncertainty. The ability to understand when the predictions fail is one of the foundations for model improvement. This also opens the possibility of using model ensembles to improve predictions and to better identify the uncertainty (Lakshminarayanan et al., 2017).

CONCLUSIONS

The strategy of choosing a subset of images that contribute the most to model accuracy proved to be successful in conferring flexibility to the model. Models trained in other trials and years, with different plot sizes and image acquisition intervals, were able to predict soybean maturity date with an RMSE lower than 2.0 days in four out of five trials. Compared to previous studies, additional challenges were addressed, focusing on the scalability of the proposed solutions. This was possible after using more than 15,000 ground truth maturity date observations from five trials, including data from three growing seasons and two countries. Data augmentation did not improve model performance and was harmful in many cases. Changing the resolution of images did not affect model performance. Model performance decreased when tested in trials with conditions unseen during training. Using ground truth information from check plots helped to correct for environmental bias. Four methods of estimating prediction uncertainty showed potential at identifying different sources of errors in the maturity date predictions. The main challenge remaining to improve model accuracy is the low correlation between leaf senescence and pod senescence for some genotypes.

REFERENCES

Alliprandini, L.F., C. Abatti, P.F. Bertagnolli, J.E. Cavassim, H.L. Gabe, et al. 2009. Understanding soybean maturity groups in brazil: Environment, cultivar classification, and stability. Crop Sci. 49(3): 801–808. doi: 10.2135/cropsci2008.07.0390.

Ampatzidis, Y., V. Partel, and L. Costa. 2020. Agroview: Cloud-based application to process, analyze and visualize UAV-collected data for precision agriculture applications utilizing artificial intelligence. Comput. Electron. Agric. 174(February): 105457. doi: 10.1016/j.compag.2020.105457.

Andrea, M.C. da S., R. Dallacort, R.C. Tieppo, and J.D. Barbieri. 2020. Assessment of climate change impact on double-cropping systems. SN Appl. Sci. 2(4): 1–13. doi: 10.1007/s42452-020-2325-z.

Araus, J.L., S.C. Kefauver, M. Zaman-Allah, M.S. Olsen, and J.E. Cairns. 2018. Translating High-Throughput Phenotyping into Genetic Gain. Trends Plant Sci. 23(5): 451–466. doi: 10.1016/j.tplants.2018.02.001.

Borra-Serrano, I., T. De Swaef, P. Quataert, J. Aper, A. Saleem, et al. 2020. Closing the Phenotyping Gap: High Resolution UAV Time Series for Soybean Growth Analysis Provides Objective Data from Field Trials. Remote Sens. 2020, Vol. 12, Page 1644 12(10): 1644. doi: 10.3390/RS12101644.

Chen, C.J., and Z. Zhang. 2020. GRID: A Python Package for Field Plot Phenotyping Using Aerial Images. Remote Sens. 12(11): 1697. doi: 10.3390/rs12111697.

Christenson, B.S., W.T. Schapaugh, N. An, K.P. Price, V. Prasad, et al. 2016. Predicting soybean relative maturity and seed yield using canopy reflectance. Crop Sci. 56(2): 625–643. doi: 10.2135/cropsci2015.04.0237.

Cobb, J.N., G. DeClerck, A. Greenberg, R. Clark, and S. McCouch. 2013. Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype-phenotype relationships and its relevance to crop improvement. Theor. Appl. Genet. 126(4): 867–887. doi: 10.1007/s00122-013-2066-0.

David, E., S. Madec, P. Sadeghi-Tehran, H. Aasen, B. Zheng, et al. 2020. Global Wheat Head Detection (GWHD) dataset: a large and diverse dataset of high resolution RGB labeled images to develop and benchmark wheat head detection methods. arXiv: 1–15.

Dobrescu, A., M.V. Giuffrida, and S.A. Tsaftaris. 2020. Doing More With Less: A Multitask Deep Learning Approach in Plant Phenotyping. Front. Plant Sci. 11(February): 1–11. doi: 10.3389/fpls.2020.00141.

Fehr, W.R., C.E. Caviness, D.T. Burmood, and J.S. Pennington. 1977. Stage of soybean development. Spec. Rep. 80: 929–931.

Hartman, G.L., E.D. West, and T.K. Herman. 2011. Crops that feed the World 2. Soybean—worldwide production, use, and constraints caused by pathogens and pests. Food Secur. 3(1): 5–17. doi: 10.1007/s12571-010-0108-x.

Jiang, Y., and C. Li. 2020. Convolutional Neural Networks for Image-Based High-Throughput Plant Phenotyping: A Review. Plant Phenomics 2020: 1–22. doi: 10.34133/2020/4152816.

Khan, Z., and S.J. Miklavcic. 2019. An Automatic Field Plot Extraction Method From Aerial Orthomosaic Images. Front. Plant Sci. 10(May): 1–13. doi: 10.3389/fpls.2019.00683.

Koga, L.J., M.G. Canteri, E.S. Calvo, D.C. Martins, S.A. Xavier, et al. 2014. Managing soybean rust with fungicides and varieties of the early/semi-early and intermediate maturity groups. Trop. Plant Pathol. 39(2): 129–133. doi: 10.1590/S1982-56762014000200003.

Lakshminarayanan, B., A. Pritzel, and C. Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems. p. 6402–6413

Liu, S., M. Zhang, F. Feng, and Z. Tian. 2020. Toward a "Green Revolution" for Soybean. Mol. Plant 13(5): 688–697. doi: 10.1016/j.molp.2020.03.002.

Maimaitijiang, M., V. Sagan, P. Sidike, S. Hartling, F. Esposito, et al. 2020. Soybean yield prediction from UAV using multimodal data fusion and deep learning. Remote Sens. Environ. 237(December 2019): 111599. doi: 10.1016/j.rse.2019.111599.

Matias, F.I., M. V Caraza-Harter, and J.B. Endelman. 2020. FIELDimageR: An R package to analyze orthomosaic images from agricultural field trials. Plant Phenome J. 3(1): 1–6. doi: 10.1002/ppj2.20005.

Morales, N., N.S. Kaczmar, N. Santantonio, M.A. Gore, L.A. Mueller, et al. 2020. ImageBreed: Open-access plant breeding web–database for image-based phenotyping. Plant Phenome J. 3(1): 1–7. doi: 10.1002/ppj2.20004.

Moreira, F.F., A.A. Hearst, K.A. Cherkauer, and K.M. Rainey. 2019. Improving the efficiency of soybean breeding with high-throughput canopy phenotyping. Plant Methods 15(1): 139. doi: 10.1186/s13007-019-0519-4.

Mourtzinis, S., and S.P. Conley. 2017. Delineating soybean maturity groups across the United States. Agron. J. 109(4): 1397–1403. doi: 10.2134/agronj2016.10.0581.

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems. p. 8026–8037

QGIS Development Team. 2020. QGIS Geographic Information System Software, Version 3.10.

Reynolds, D., F. Baret, C. Welcker, A. Bostrom, J. Ball, et al. 2019. What is cost-efficient phenotyping? Optimizing costs for different scenarios. Plant Sci. 282(June 2018): 14–22. doi: 10.1016/j.plantsci.2018.06.015.

Reynolds, M., S. Chapman, L. Crespo-Herrera, G. Molero, S. Mondal, et al. 2020. Breeder friendly phenotyping. Plant Sci. 295(July 2019): 110396. doi: 10.1016/j.plantsci.2019.110396.

Schnaufer, C., J.L. Pistorius, and D. LeBauer. 2020. An open, scalable, and flexible framework for automated aerial measurement of field experiments. In: Thomasson, J.A. and Torres-Rua, A.F., editors, Autonomous Air and Ground Sensing Systems for Agricultural Optimization and Phenotyping V. SPIE. p. 9

Song, W., S. Sun, S.E. Ibrahim, Z. Xu, H. Wu, et al. 2019. Standard Cultivar Selection and Digital Quantification for Precise Classification of Maturity Groups in Soybean. Crop Sci. 59(5): 1997–2006. doi: 10.2135/cropsci2019.02.0095.

Tresch, L., Y. Mu, A. Itoh, A. Kaga, K. Taguchi, et al. 2019. Easy MPE: Extraction of Quality Microplot Images for UAV-Based High-Throughput Field Phenotyping. Plant Phenomics 2019: 1–9. doi: 10.34133/2019/2591849.

Tsaftaris, S.A., M. Minervini, and H. Scharr. 2016. Machine Learning for Plant Phenotyping Needs Image Processing. Trends Plant Sci. 21(12): 989–991. doi: 10.1016/j.tplants.2016.10.002.

Wang, X., H. Xuan, B. Evers, S. Shrestha, R. Pless, et al. 2019. High-throughput phenotyping with deep learning gives insight into the genetic architecture of flowering time in wheat. Gigascience 8(11): 1–11. doi: 10.1093/gigascience/giz120.

Yang, G., J. Liu, C. Zhao, Z. Li, Y. Huang, et al. 2017. Unmanned aerial vehicle remote sensing for field-based crop phenotyping: Current status and perspectives. Front. Plant Sci. 8(June). doi: 10.3389/fpls.2017.01111.

Yang, Q., L. Shi, J. Han, J. Yu, and K. Huang. 2020. A near real-time deep learning approach for detecting rice phenology based on UAV images. Agric. For. Meteorol. 287(July 2019): 107938. doi: 10.1016/j.agrformet.2020.107938.

Young, S.N., E. Kayacan, and J.M. Peschel. 2019. Design and field evaluation of a ground robot for high-throughput phenotyping of energy sorghum. Precis. Agric. 20(4): 697–722. doi: 10.1007/s11119-018-9601-6.

Yu, N., L. Li, N. Schmitz, L.F. Tian, J.A. Greenberg, et al. 2016. Development of methods to improve soybean yield estimation and predict plant maturity with an unmanned aerial vehicle based platform. Remote Sens. Environ. 187: 91–101. doi: 10.1016/j.rse.2016.10.005.

Yuan, W., N.K. Wijewardane, S. Jenkins, G. Bai, Y. Ge, et al. 2019. Early Prediction of Soybean Traits through Color and Texture Features of Canopy RGB Imagery. Sci. Rep. 9(1): 1–17. doi: 10.1038/s41598-019-50480-x.

Zdziarski, A.D., M.H. Todeschini, A.S. Milioli, L.G. Woyann, A. Madureira, et al. 2018. Key soybean maturity groups to increase grain yield in Brazil. Crop Sci. 58(3): 1155–1165. doi: 10.2135/cropsci2017.09.0581.

Zhao, D., G. Yu, P. Xu, and M. Luo. 2019. Equivalence between dropout and data augmentation: A mathematical check. Neural Networks 115: 82–89. doi: 10.1016/j.neunet.2019.03.013.

Zhou, J.J., D. Yungbluth, C.N. Vong, A. Scaboo, and J.J. Zhou. 2019. Estimation of the Maturity Date of Soybean Breeding Lines Using UAV-Based Multispectral Imagery. Remote Sens. 11(18): 2075. doi: 10.3390/rs11182075.

CHAPTER 2

SPATIAL VARIABILITY OF CROP RESPONSES TO AGRONOMIC INPUTS IN ON-FARM PRECISION EXPERIMENTATION

ABSTRACT

Within-field variability of crop yield levels has been extensively investigated, but the spatial variability of crop yield *responses* to agronomic treatments is less understood. On-farm precision experimentation (OFPE) can be a valuable tool for the estimation of in-field variation of optimal input rates and thus improve agronomic decisions. Therefore, the objectives of this study were to investigate the spatial variability of optimal input rates in OFPE and the potential economic benefit of site-specific input management. Mixed geographically weighted regression (GWR) models were used to estimate local yield response functions. The methodology was applied to investigate the spatial variability in corn response to nitrogen and seed rates in four cornfields in Illinois, USA. The results showed that spatial heterogeneity of model parameters was significant in all four fields evaluated. On average, the RMSE of the fitted yield decreased from 1.2 Mg ha^{-1} in the non-spatial global model to 0.7 Mg ha^{-1} in the GWR model, and the r-squared increased from 10% to 68%. The average potential gain of using optimized uniform rates of seed and nitrogen was US$ 65.00 ha^{-1}, while the added potential gain of the site-specific application was US$ 58.00 ha^{-1}. The combination of OFPE and GWR proved to be an effective tool for testing precision agriculture's central hypothesis of whether optimal input application rates display adequate spatial variability to justify the costs of the variable rate technology itself. The reported results encourage more research on response-based input management recommendations instead of the still widespread focus on yield-based algorithms.

INTRODUCTION

Site-specific technologies, including yield monitoring, remote sensing imaging, and variable rate input application, have become increasingly available to farmers in recent decades. Additionally, many farmers have access to software tools to process this information and use it to guide site-specific management decisions. This decision-making process is based primarily on the knowledge about yield response to crop input management in agronomic trials, which involve changing management practices and subsequent monitoring of the effects on the system output (Pringle *et al*. 2004).

The potential for economic gains from the implementation of site-specific crop management technologies depends on the assumption that yield response to an agronomic input can be described as a function of managed input strategies, field characteristics, and weather. Because field characteristics are spatially dependent, how yields respond to managed inputs changes over space the shape of the traditional "yield curve" plot with input application rate on the horizontal axis and yield on the vertical axis will vary among sites within a field) (Bullock and Bullock, 1994). However, most variable rate recommendations are calculated using the same methods, and calibrations developed for uniform field management. These are based on agronomic trials that are usually planned to be representative of the conditions in the geographical region of interest to the researchers aiming to infer the management insights coming from small-scale plots to larger regions. To make the recommendations more accessible to users, these models are usually very simplified versions of the true yield response function (Morris et al., 2018). In the case of nitrogen fertilizer, this simplification has led to the widespread use of yield levels as a proxy for estimating the optimal input application rates, even though academic research provides only weak evidence of any correlation between yield levels and economically optimal input application rates (Scharf et al., 2006; Bachmaier and Gandorfer, 2009; Rodriguez et al., 2019).

To improve site-specific management requires information about how crop yield responds to varying treatments and how those responses vary over space (Bullock and Bullock, 1994). Variable-rate technology can be used to systematically control input levels in highly mechanized, large-scale production systems (Piepho et al., 2011). In addition, these operations make possible the running of large-scale, on-farm precision experiments (OFPE), that generate large amounts of site-specific response data, which can be used to understand the spatial variation of optimal input application rates (Bullock et al., 2019). Moreover, because OFPE data are gathered in the same fields for which management recommendations are desired, field and site-specific yield response functions can be estimated. Thus, the ultimate purpose of OFPE is to develop site-specific input applications (Piepho et al., 2011). OFPEs also allow testing the fundamental hypothesis of precision agriculture, which is that the rate at which inputs are applied can be profitably varied within fields to match site-specific requirements (Lark and Wheeler, 2003; Bachmaier and Gandorfer, 2009). Of course, estimating site-specific yield response requires spatial data analysis (Hurley et al., 2005; Bullock et al., 2007).

Most of the examples of OFPE consider only the effect of a single factor on crop yield (Kindred *et al*. 2017; Piepho *et al*. 2011; Pringle *et al*. 2004). The statistical analysis of this type of trial often involves using geostatistical interpolation methods to estimate the effects of all tested treatment levels in all points of a regular grid. With that, yield estimates for each treatment level are obtained not only for the treatment tested on that point but for all other treatments as well. Then local response function parameters need to be estimated using the interpolated values for each point in the interpolated grid. The confidence of each estimate will depend on the distance of neighboring experimental units with the same treatment level, and for that, systematic designs are preferable over randomized designs (Pringle *et al*. 2004).

The main limitation of these geostatistical methods is that they rely on interpolating yield maps using only one level of the treatment. If the treatment has five levels, for example, only 20% of the data is used at every interpolation. In a factorial design with four levels of the first treatment and five levels of the second, totaling twenty combinations, only 5% of the observations would be used in each one of the twenty interpolations. Due to the suboptimal use of the neighboring plots in the estimation of the yield response function parameters, the geostatistical interpolation is not well suited for the analysis of continuous variables and factorial designs with many levels (Pringle *et al*. 2004).

Most spatial inference methods, including most generalized least square and spatial econometrics methods, were developed to focus on the implications of spatial dependence (Anselin, 2010). However, spatial heterogeneity has been overlooked, even though it is crucially important to model spatial data appropriately (Geniaux and Martinetti, 2018; Murakami and Griffith, 2019).

To model the spatial heterogeneity as local yield responses, it is necessary to have some local variable as an input, either as part of the design or some covariable that was measured in the field (Bachmaier and Gandorfer, 2009; Thöle et al., 2013). Besides the need for additional data, the main limitation of this approach is that the results will be dependent on the correlation between the covariables and the yield response. The hypotheses that can be tested with such a model are restricted to whether the yield response function has any interaction with the spatial covariables, but it is not possible to test whether there is significant spatial variability in the parameters of the yield response function itself. Since there will always remain variables that can not be observed, drawing conclusions based only on the known variables may lead to wrong generalizations.

In recent years, spatially varying coefficient (SVC) models have attracted considerable attention in various fields of applied sciences. However, their use in agronomic research is still limited (Cai *et al*. 2014; Murakami *et al*. 2018; Trevisan, *et al*. 2019a). These methods have been proposed to investigate the spatial variability or no stationarity of coefficient estimates in regression models. In other words, SVC methods are equivalent to the direct estimation of the site-specific yield production function parameters. The estimates, along with associated inference diagnostics, can be mapped to the original measurement locations or a new set of locations. Among the SVC models, the geographically weighted regression (GWR) has been one of the methods commonly applied. In the GWR, the parameters of the models vary in space and can be mapped and interpreted as a spatial variable (Brunsdon et al., 1996; Fotheringham, 1997).

The GWR method allows the specification of more complex and less restrictive models, in which all parameters are estimated for each location. The neighboring points are included in the estimation even if they are from different treatment levels by applying weighted least squares estimation to neighboring subsamples. Weights are estimated via a distance-decay kernel, similar to the weights given by the semivariogram to neighbors in kriging interpolation.

Therefore, this study proposes to integrate the recent advances in precision agriculture and methods of spatial analysis to develop a new methodology for understanding the spatial variability of crop responses, and thereby test the viability and profitability of precision agriculture. The novelty of the proposed methodology is the use of GWR as a data analysis technique applied to OFPE, without the need for spatial covariates. The hypothesis being tested is that the yield response function with spatially varying coefficients is adequate relative to the alternative of using a yield response function with only global parameters. The specific objectives of this study were to investigate the spatial variability of optimal input application rates, and the potential economic benefit of joint use of OFPE and site-specific management.

MATERIALS AND METHODS

OFPE design

The fields used to generate the datasets for this study come from Data-Intensive Farm-Management (DIFM) project on-farm field trials. DIFM is a multi-university project supported by the United States Department of Agriculture - National Institute of Food and Agriculture (USDA – NIFA) to use precision agriculture technology to generate original, high-quality, full-field, on-farm trial data at low cost (Bullock *et al*. 2019). Four Illinois fields representing typical U.S. Corn Belt maize production systems were used for the trials (Table 2.1), with Fields 1 and 2 hosting trials in 2017, and Fields 3 and 4 in 2018.

Table 2.1. Description of the four cornfields used for the on-farm precision experimentation.

Field*	Year	Location	Area (ha)	Elev. (m)	Exp. Units	Dim. (m)	Obs. Units
Field 1	2017	Effingham – IL	17	180	322	88 X 6	2898
Field 2	2017	Moultrie – IL	32	210	208	85 X 18	3276
Field 3	2018	Effingham – IL	12	175	234	82 X 6	3137
Field 4	2018	Effingham – IL	20	192	128	86 X 18	2816

*Elev: elevation; Exp. Units: number of experimental units in the trial design; Dim: plot dimensions; Obs. Units: number of observations in the final dataset after aggregation of raw observations into regular polygons.

The OFPEs were designed to generate data for localized estimation of crop response to guide site-specific crop management (Piepho et al., 2011). Each experiment had two managed input factors, seeding rate (*S*) and nitrogen fertilizer application rate (*N*), with at least four levels of each factor. The allocation of the rates in the experimental units followed a completely randomized factorial design replicated over the entire field.

All treatment levels were implemented in the field using variable rate planters and fertilizer applicators. Each experimental unit (“plot”) was given a dimension to fit the swath width of the machinery available (Table 2.1). The ranges of variation for the tested rates were chosen according to the producer’s experience and expectations for the field, typically allowing a 20% variation in each direction around the *status quo* rate (Table 2.2). Other farming practices were kept constant throughout the field and were conducted by the farmers by standard protocols for the region.

Table 2.2. Treatment rates and yield of the four Illinois cornfields used in the OFPE trials.

Field*	Seed Rate		Base N	Total N		Yield	
	(kseed ha^{-1})	(levels)	(kg ha^{-1})	(kg ha^{-1})	(levels)	(count)	(Mg ha^{-1})
Field 1	66 – 96	4	30	164 – 232	7	9562	11.2 (1.7)
Field 2	76 – 96	4	62	190 - 257	4	13311	11.9 (0.4)
Field 3	67 – 89	5	52	208 - 275	6	7109	9.8 (1.8)
Field 4	67 – 91	4	179	179 - 246	4	11612	13.5 (1.3)

*Base N: nitrogen rate applied at a uniform rate, at or before planting. Total N: total nitrogen rate consisting of the base N added to the experimental rates applied at side-dressing. Yield: grain yield estimated by the yield monitor system; the count represents the number of raw sensor observations, followed by the average yield and the standard deviation within parenthesis.

Data analysis

Treatment applications were monitored using the feedback sensors in the variable rate applicators, generating the as-applied data. The as-applied data was filtered to remove extreme values, defined as values distant from the average more than three times the interquartile range. Yield data were collected during harvest using combine yield monitoring systems. Yield data were filtered to remove global and spatial outliers, taking into account the boundary of the plots (Trevisan *et al*. 2019b). The data from the headlands and borders of the field were also discarded.

Each plot was subdivided into ten to twenty smaller regular polygons, depending on the plot's original length (Figure 2.1). The area of observational units was constant within each field and ranged from 37 m^2 to 87 m^2 among fields because of differing machinery dimensions. The average value of all points within a polygon was used as the observational unit in the final dataset, which in each field consisted of approximately 3000 observations (Table 2.1).

All steps and models were individually applied to each field. Letting *I* represent the exact number of observations in a field, with observations are indexed by *i* = 1, …, *I*. Let the (longitude, latitude) geographic coordinates of the centroid of an observational unit (a "point") be denoted (l_i, t_i). Based on weighted least squares, the GWR method modifies the estimates of the regression coefficients as a function of the geographic position (l_i, t_i). In matrix notation, the regression coefficients were estimated by Equation [2.1]:

$$b(l_i, t_i) = (X'G(l_i, t_i)X)^{-1} X'G(l_i, t_i)Y, \qquad [2.1]$$

where Y is the ($I \times 1$) vector whose components are the $Yield_i$, and X_i is the ($I \times 1$) vector of the values of the J explanatory variable at the location (l_i, t_i). X is the ($I \times J$) matrix in which X_i is the i^{th} column, for i = 1, …, I. In this notation, the GWR model incorporates local variation in the estimated coefficients b(l_i, t_i) using a geographic weighting matrix G(l_i, t_i). This diagonal matrix incorporates the distance d_{ij} between predictors X_i at (l_i, t_i) and dependent observations Y_j at the location (l_j, t_j). As d_{ij} increases, the explanatory variables are expected to decrease in influence over the response variable (Plant, 2012).

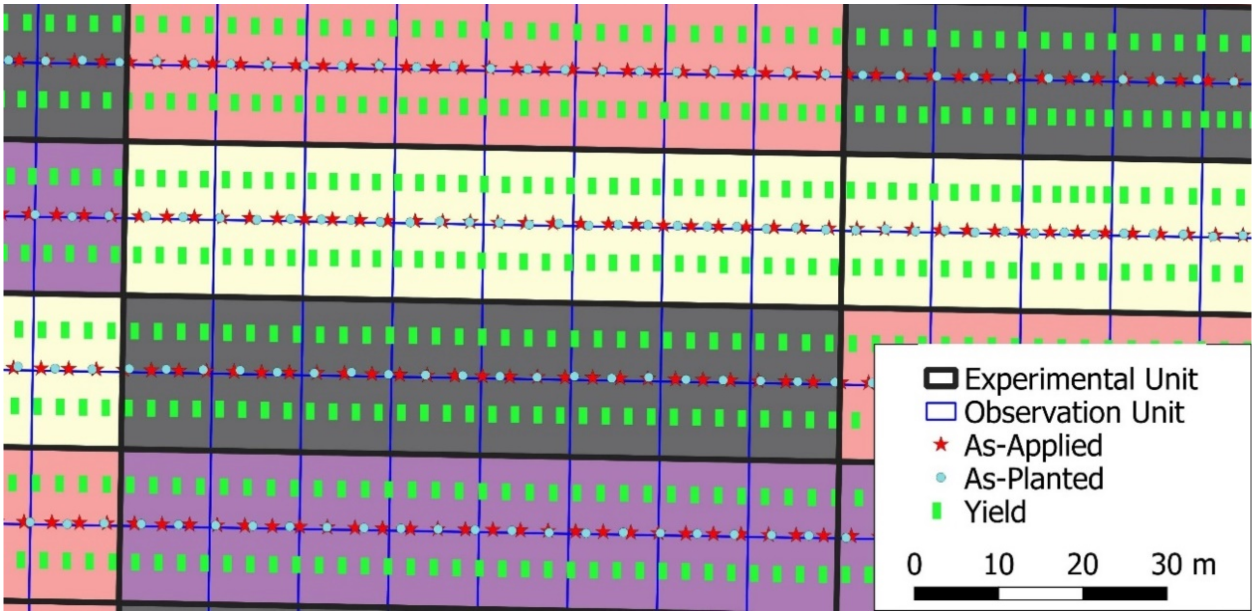

Figure 2.1. Examples of spatial data collected in one of the OFPE trials and the units used to aggregate the information.

One particular case of GWR is the moving window regression, which applies ordinary least squares estimation to neighboring subsamples by specifying a kernel weighting scheme in which all weights are equal to one within the kernel and zero otherwise (Lloyd, 2009). Using distance-decay weighting provides added flexibility to local regression modeling, allowing more data to have local influence, and tends to yield more smoothly varying coefficient surfaces. This decay is captured by elements g_{ij} of a matrix G, for example, as would be defined by the Gaussian spatial kernel function g_{ij} in Equation [2.2]:

$$g_{ij}(l_i, t_i) = exp\left(-0.5\left[\frac{d_{ij}}{h}\right]^2\right) \qquad [2.2]$$

Function g_{ij} calculates the weights given to the regression points at each location, with a bandwidth h adjusted by cross-validation and selected based on the corrected Akaike information criterion (AICc) (Lu et al., 2018). In addition to the Gaussian, the exponential and the bi-square spatial kernels were also compared. The spatial kernel can use a fixed distance for the bandwidth *h*, or it can be adapted at each location to include a minimum number of neighbors. Although the observations in the final dataset were regularly spaced in the field, an adaptive distance bandwidth was chosen to improve the results at the borders of the experiment, where a fixed kernel would include fewer observations. An example of the weights at one location is presented in Figure 2.2.

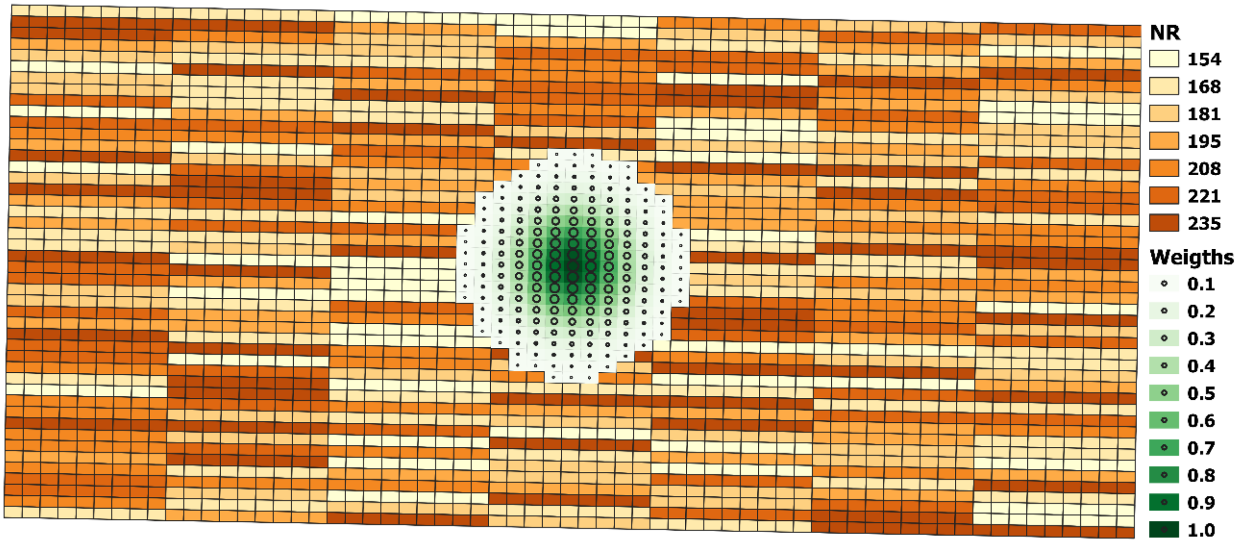

Figure 2.2. Illustration of the weights given to neighbor points for the parameter estimates at the center of the field using a Gaussian kernel. NR stands for nitrogen rate and represents the input variation according to the trial design.

Literature often suggests that non-linear functions such as the quadratic-plateau function, are more suited to describe yield response in corn than are their linear-in-variables counterparts (Scharf et al., 2005; Pahlmann et al., 2016). Although nonlinear-in-variables models may be necessary for describing the response to a wide range of rates, a linear-in-variables function can be an adequate approximation over a sufficiently small subset of the function's domain. Since, as was the case in the experiments reported here, the ranges of rates tested in whole field OFPEs are restricted to reduce potential profit losses caused by experimental rates being far different from economically suboptimal rates, linear-in-variables models may be deemed best after the trade-off between goodness-of-fit and research expenses is considered. The polynomial function in Equation [2.3] presents the full model considered to describe yield response for each field:

$$Y_i = \beta_{0(i)} + \beta_{1(i)}.S_i + \beta_{2(i)}.N_i + \beta_{3(i)}.S_i.N_i + \beta_{4(i)}.S_i^2 + \beta_{5(i)}.N_i^2 + \beta_{6(i)}.S_i^2.N_i + \beta_{7(i)}.S_i.N_i^2 + \beta_{8(i)}.S_i^2.N_i^2 + \varepsilon_i. \qquad [2.3]$$

In Equation [2.3], Y_i is the observed yield, S_i is the seed rate, and N_i is the nitrogen fertilizer application rate at a location $i \in \{1, \ldots, I\}$. The βs in Equation [2.3] can assume values according to one of the three scenarios:

a) $\beta_{k(i)} = 0$ for all i, which is equivalent to dropping term k from the model (where term 1 is S_i, term 2 is N_i, etc.);

b) $\beta_{k(1)} = \ldots = \beta_{k(I)} \neq 0$, which makes term k a global variable, also denoted as $\beta_{k\,g}$;

c) $\beta_{k(m)} \neq \beta_{k(n)}$ for some $m, n \in \{1, \ldots, I\}$, which makes term k a local variable, also denoted as $\beta_{k\,l}$.

The standard GWR considers that all model parameters vary spatially, meaning that case "c" is assumed for every k. But in the research reported here, a mixed GWR model was applied, meaning that case "b" could hold for some k and case "c" for others—that is, that some terms could be local and others global. By limiting the number of coefficients that can vary over space, the mixed GWR approach can sidestep multicollinearity problems, which can lead to artificial spatial patterns of the coefficients (Geniaux and Martinetti, 2018). This could be a problem, for example, when including coefficients of both the linear and the quadratic terms, for example, $\beta_{1(i)}$ and $\beta_{4(i)}$ in Equation [2.3], as spatially varying. It has also been demonstrated that setting some parameters of a yield response function to a common value for all subareas within a field and year can improve both the goodness of fit and interpretability of a model (Bachmaier and Gandorfer, 2012; Pahlmann et al., 2016). Based on these recommendations and to reduce the number of scenarios tested, model selection was performed setting all interaction and higher-order terms (i.e., terms 3, …, 8) as global variables.

The mixed GWR was also used to test whether a local parameter was needed for the intercept, S_i, and N_i terms in Equation [2.3]. Model selection was based on the AICc criterion (Lu et al., 2018). The change in the root mean squared error (RMSE) when a term was added as a local or a global variable was used to represent a term's importance. The significance of the coefficient at each location was also tested using the p-values from pseudo-t-tests, with the alpha adjusted by the Fotheringham-Byrne method (Lu et al., 2014).

The “MGWRSAR” R package (Geniaux and Martinetti, 2018) was used for model calibration and the term inclusion. The “GWmodel” R package (Gollini et al., 2015) was used to provide additional diagnostics and to test parameter significance at every location. Both packages were used because MGWRSAR computed the results much faster, which was especially important for testing multiple values of kernel type, bandwidth, and parameters in the model, while GWmodel provided more detailed output.

Comparison of optimal rate scenarios

Table 2.3 summarizes the equations used to compare optimal rates under the various scenarios. The assumptions described above about coefficients implied location-specific yield response functions $f_i(N, S)$, where units for yield, N, and S were $(Mg\ ha^{-1})$, $(kg\ ha^{-1})$, and $(kseed\ ha^{-1})$, where $kseed$ denotes thousands of seeds. The *status quo* seed rate S_{sq} and nitrogen fertilizer application rate N_{sq} were those that the producers said they would have applied had they not taken part in the on-farm precision experiments.

The following prices were used to calculate the economically optimal rates: corn price of p = US$ 160.00 Mg^{-1} (US$ 4.00 bu^{-1}), corn seed price of w_S = US$ 3.50 kseed^{-1}, and nitrogen price of w_N = US$ 0.88 kg^{-1} (US$ 0.40 lb^{-1}). These values imply approximate price ratios w_S/p = 0.022 Mg kseed^{-1} (0.88 bu kseed^{-1}), and w_N/p = 0.0055 Mg kg^{-1} (0.10 bu lb^{-1}). The quantitative results reported below were calculated by replacing the expected yield response function $f_i(N, S)$ in Table 3 with the predicted expected yield response function in Eq. 3, and solving the maximization problems using standard first-order conditions from multivariate calculus, obtaining the values of the arguments S and N that maximize (argmax) the expected net revenue function (Bullock and Bullock 1994). In these scenarios, it is assumed that all equipment needed is already available at no extra cost. For simplicity, costs associated with information management and data processing are also not considered in the economic analysis. The calculations were conducted with base functions in the R software package (R Core Team, 2020).

Table 2.3. Summary of the equations used in the comparison of optimal rates scenarios.

Equation	Description
$y_i = f_i(N,S)$	the expected yield response function
$y_i^{sq} = f_i(N^{sq}, S^{sq})$	*status quo* expected yield also referred to as reference yield
$r_i^{sq} = py_i^{sq} - w_N N^{sq} - w_S S^{sq}$	*status quo* expected net revenue
$r_i(N,S) = \ (pf_i(N,S) - w_N N - w_S S)$	the expected net revenue function
$\Delta r_i(N,S) = r_i(N,S) - \ r_i^{sq}$	expected net revenue loss (gain if < 0) function
$(N^u, S^u) = \underset{(N,S)}{\text{argmax}} \left\{ \sum_{i=1}^{I} r_i(N,S) \right\}$	optimal nitrogen fertilizer and seeding plan under uniform management
$y_i^u = f_i(N^u, S^u)$	expected yield under optimal uniform management
$r_i^{u*} = py_i^u - w_N N^u - w_S S^u$	expected net revenue under optimal uniform management
$(N_i^*, S_i^*) = \underset{(N,S)}{\text{argmax}} \{[r_i(N,S)]\}$	the optimal site-specific management plan
$y_i^* = f_i(N_i^*, S_i^*)$	expected yield under optimal site-specific management
$r_i^{ss*} = py_i^* - w_N N_i^* - w_S S_i^*$	expected net revenue under optimal site-specific management
$\Delta r_i^{u+} = r_i^{u*} - r_i^{sq}$	expected net revenue increase due to optimal uniform management
$\Delta r_i^{ss+} = r_i^{ss} - r_i^{sq}$	expected net revenue increase due to optimal site-specific management

RESULTS AND DISCUSSION

OFPE implementation

Variable-rate applicators generally apply inputs with some degree of error, rarely exactly matching the target rate. Nevertheless, monitors can record very accurately the "as-applied" and "as-planted" rates, which are the actual rates at which the inputs "went into the ground". Therefore, the as-planted and as-applied maps presented highly accurate implementations of the trial designs (Figure 2.3a,b). Historically, the main reason for using the targeted rates or grouping the as-applied rates into a few levels was to analyze variance using treatments as categorical variables. Since the proposed methodology treats treatments as continuous variables in all steps, such categorization was not necessary.

The yield maps of the four fields reveal that most yield variation could be attributed to factors other than the treatment rates since there are no obvious patterns of treatment-yield relations (Figure 2.3c). Field 1 presents a wide range of yield values, but not a clear spatial structure of the variability. Yields in Field 2 were the most spatially homogeneous among the four, with a yield range of only 2.0 Mg ha^{-1}. Spatial variability is more evident in Fields 3 and 4, with Field 3 on average producing the lowest yields and Field 4 the highest.

The factorial combinations of the experiment's several nitrogen and seed rates resulted in many (*N*, *S*) treatment levels in each field's trial design, making a visual interpretation of the results difficult. The quality of the yield data is currently a major concern in OFPE and also affect how the results look in the map. Errors in yield data can come from the different lengths of time needed for combines' internal threshing mechanisms to process the grains, which can reduce the contrast between the yields at different input rates (Lark et al., 1997; Lark and Wheeler, 2003). To improve yield data quality, the convoluted yield monitoring process was taken into account in trial design, with plot lengths sufficiently long to allow the harvester to pass through any treatment plot for at least 30 seconds. Therefore, the effects of the yield convolution in the treatment transition zones were considered to be small, since most of the points included in each location were outside of the transition zones. To maintain methodological simplicity, the points in the transitions were kept in the data analyzed, and no explicit effect was included in the model. The importance of these effects should be explored in future research.

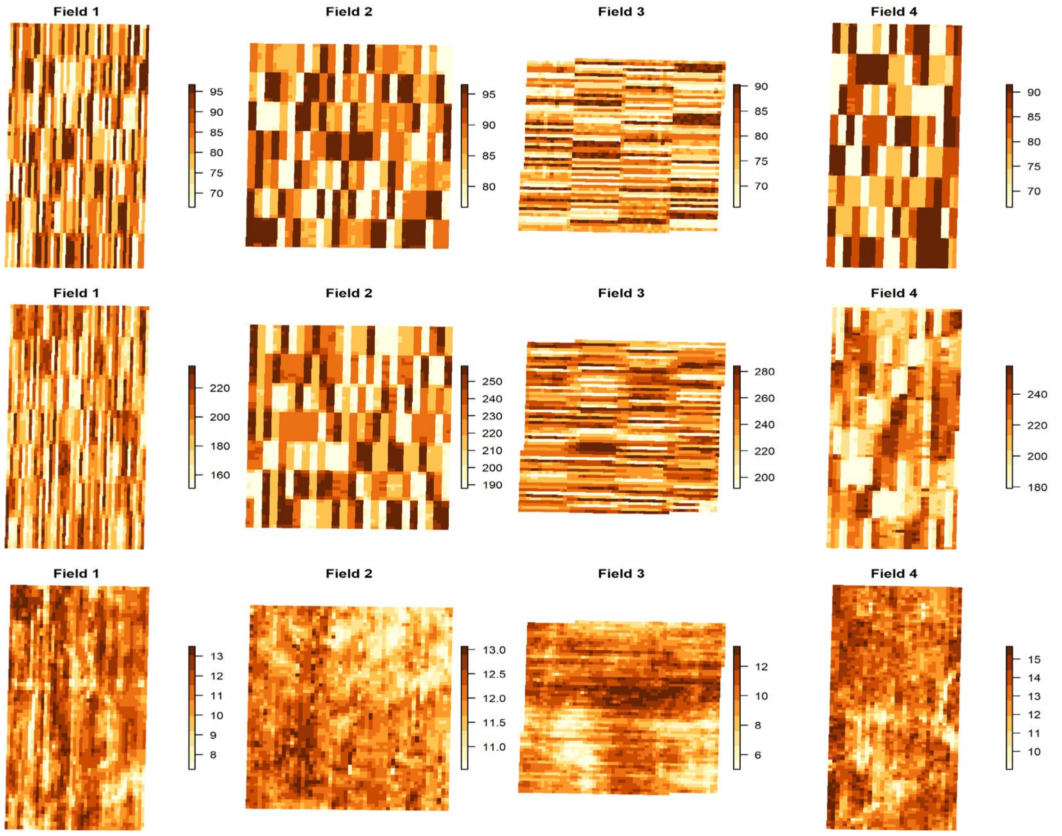

Figure 2.3. Spatial distribution of (a) as-planted seed rates (kseed ha^{-1}), (b) as-applied nitrogen rates (kg ha^{-1}), and (c) corn yield (Mg ha^{-1}), as registered by the yield monitor and aggregated to each observational unit in the four on-farm precision experiments.

The dispersion of yield values for the same input rate is much higher in OFPE than what is commonly observed in small plot research (Figure 2.4). The noise in the data comes from many sources, including the errors associated with the yield monitoring process and the spatial variability of yield that would be observed even if inputs were applied at a uniform rate throughout the field. The only scenario in which an overall response can be observed was for the nitrogen rate in Field 1. The distribution of points also supports the decision of using linear regression to represent the yield responses.

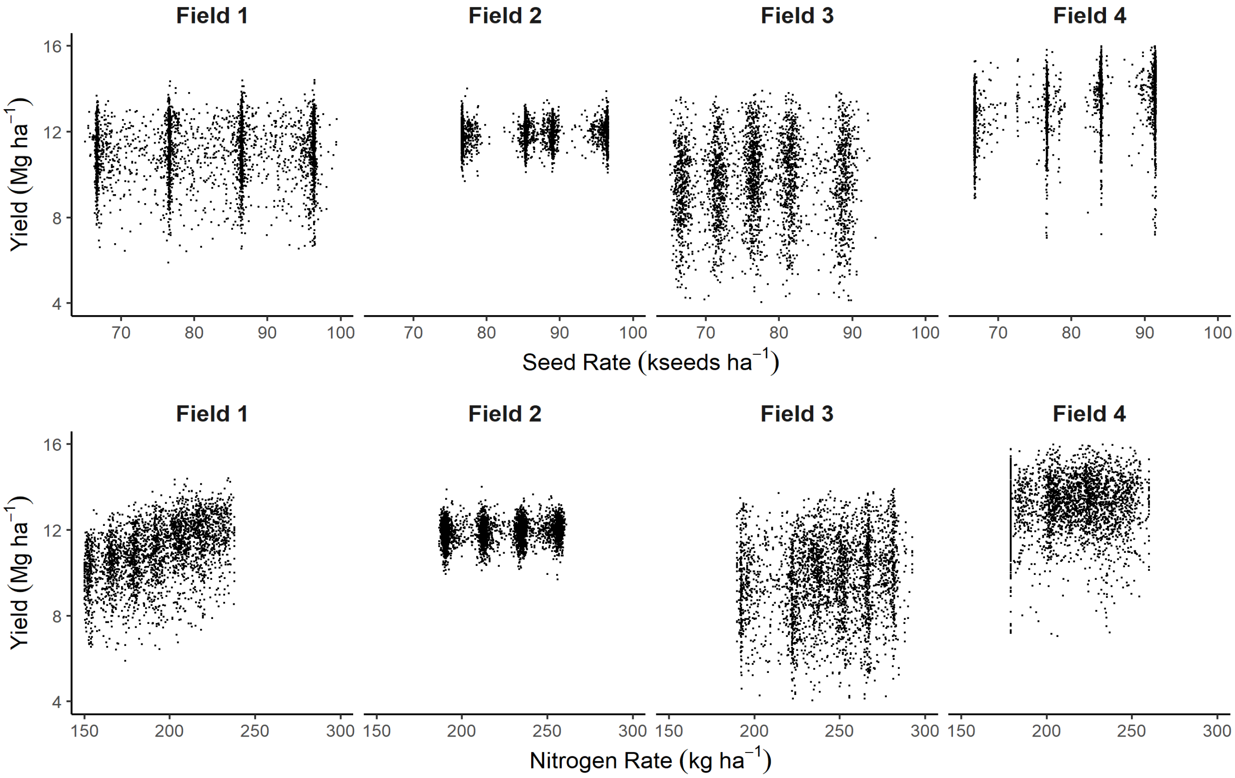

Figure 2.4. Scatterplots of the relationship between (a) yield and as-planted seed rates, and (b) yield and as-applied nitrogen rates, in the four on-farm precision experiments.

GWR results

The optimal bandwidth converged to about the same number of neighbors independently of the kernel type and the field. This optimal bandwidth determined that between 200 and 300 parameters were necessary for the Gaussian and exponential kernels, and the number of parameters in the bi-square kernel was between 1000 and 1500. There are indications in the literature that the cross-validation tends to lead to an overfitted model with a too-small bandwidth, which can lead to instability and overestimation of the spatial variability in the fitted parameters (Murakami et al., 2018). This seemed to be the case for the bi-square kernel. The Gaussian kernel presented lower AICc than the exponential in all fields, and it was therefore selected for subsequent steps.

In each field, the optimal size of the adaptive bandwidth parameter was found to allow the inclusion of about twenty neighbors. With the Gaussian kernel, these twenty neighbors captured about 67% of the total weights, and another eighty neighboring observations were needed to capture 99% of the total weights (Figure 2.2). The convergence to the same bandwidth size even though the support area of the observation units varied from 37 to 87 m^2, further demonstrates the cross-validation used to optimize the bandwidth is taking into account the variance-bias trade-off more than the spatial variability of the inputs (Lu et al., 2018; Murakami et al., 2018). A discussion of alternative methods of parameter estimation is beyond the scope of the present article, but these nuances of the GWR method may significantly impact results and should be further explored.

A summary of the comparisons between the fitted GWR model and the alternative model using ordinary least squares (OLS) is presented in Table 2.4. The AICc and BIC values show that the GWR model fits the data significantly better than does the OLS model. The largest difference was observed in Field 3, where the RMSE decreased from 1.88 Mg ha^{-1} to 0.81 Mg ha^{-1}.

The relative importance of each variable, as measured by the relative decrease in the RMSE, is one of the keys to understanding why there is such a significant difference between the two models (Figure 2.5). The spatially varying intercept (β_0) explained 25% to 50% of observed yield variability; this percentage is related to the spatial structure of the field variability. For example, Field 3 had the most extensive range of the predicted reference yield and was also the field for which β_0 explained most of the variability. Comparing the results in Table 2.4 and Figure 2.5, it becomes evident that the explanatory power of the GWR model is also more related to the spatial structure of the yield variability than to the magnitude of the crop response to the treatments.

The global response caused a reduction in the RMSE greater than the local terms only in explaining the variability of the responses for seed rate ($\beta_{1g} > \beta_{1l}$) in Field 4 and nitrogen rate ($\beta_{2g} > \beta_{2l}$) in Field 1. For all other field and input combinations, the local terms were more prominent in explaining yield than the global terms. The spatial variability in response to *S* and *N* were consistent across all fields, with β_{1l} explaining on average 3% and β_{2l} explaining on average 5% of the total variability. The unexplained variations, represented by the error term, are caused mainly by the short distance variance caused by the noise in the yield monitoring system, and by other sources of variation such as crop damage by machinery and yield losses caused by pests and diseases.

Table 2.4. Summary statistics for the yield response models fitted for each field.

Field*	Model	DF	ENP	AICc	BIC	RMSE	R-squared
Field 1	OLS	2889	9	46556	46574	1.23	0.21
	GWR	2701	197	44969	46081	0.87	0.60
Field 2	OLS	3267	9	47092	47111	0.53	0.04
	GWR	3042	234	43922	45276	0.30	0.69
Field 3	OLS	3128	9	53083	53101	1.88	0.03
	GWR	2876	261	48349	49846	0.81	0.82
Field 4	OLS	2807	9	45346	45364	1.25	0.14
	GWR	2625	191	43357	44428	0.82	0.63

*OLS: ordinary least squares; GWR: mixed geographically weighted regression; DF: degrees of freedom; ENP: effective number of parameters; AICc: corrected Akaike information criteria; BIC: Bayesian information criteria; RMSE: root mean squared error.

The outcome of accounting for most of the yield variability by the variation in factors other than controllable inputs have been reported in other studies (Kindred et al., 2017). The limited range of variation in the treatment rates, in comparison to what is used in small plots, also plays a significant role in the explanatory power of the model. While in small plot trials, the nitrogen rates may vary from zero to twice the usual rates (0 to 200% of the *status quo* rate), in the OFPE reported here, the variation was of only 20% (80% to 120% of the *status quo* rate). The small range of variation in *N* and *S* rates, coupled with the broader importance of spatial effects, the large number of observations, and the spatial correlation between observations make the analysis and interpretation of OFPE results challenging. Initially, input ranges were chosen in an attempt to minimize field implementation costs and encourage farmer's participation. Trial design scenarios with more extensive input rate ranges are planned in future studies.

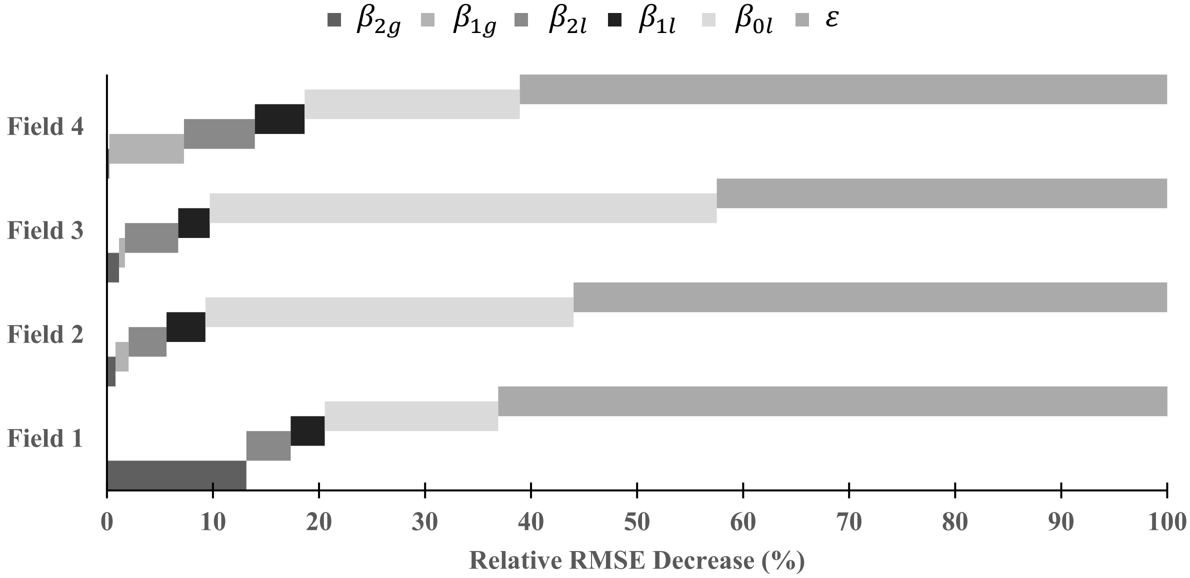

Figure 2.5. Decrease relative RMSE with the inclusion of each variable in the regression model in the four on-farm precision experiments. For interpretation of the terms, refer to Eq. 3.

Although the β_1 and β_2 coefficients explained only about 10% of yield variation, these are the most important model parameters from a crop management perspective and they will be the focus of the discussion (Figure 2.6). The values of $\beta_1$ranged from about -150 to 150 kg kseed^{-1}. If coefficients $\beta_k = 0$ for $k = 3, 4, \ldots, 8$, then β_1 represents the change in the expected yield in kg ha^{-1} when the seed rate is increased by 1.0 kseed ha^{-1} (the *marginal expected product* of kseed), and β_2 represents the *marginal expected product* of N. Since the grain equivalent cost of increasing the seed rate by 1.0 kseed ha^{-1} is about 22 kg ha^{-1} (buying one kseed costs w_S = \$3.50, and selling 22 kg (i.e., 0.022 Mg) of grain provides 160×0.022 ≈ \$3.50 in revenue). Thus, the “break-even” marginal product of kseed is β_1^{be} = 22 kg ha^{-1}. It is also possible to compare the distributions between fields, evidencing, for example, that Field 3 has a wider variation than Field 2. Similar results can be seen for the β_2 values, with an overall range within -40 to 40 kg kg^{-1}. The grain-equivalent cost of increasing the nitrogen rate by 1.0 kg ha^{-1} is $w_N/p \approx$ \$5.50 kg ha^{-1}, which, if $\beta_3 = \ldots = \beta_8 = 0$, is the break-even marginal product of nitrogen, called β_2^{be}.

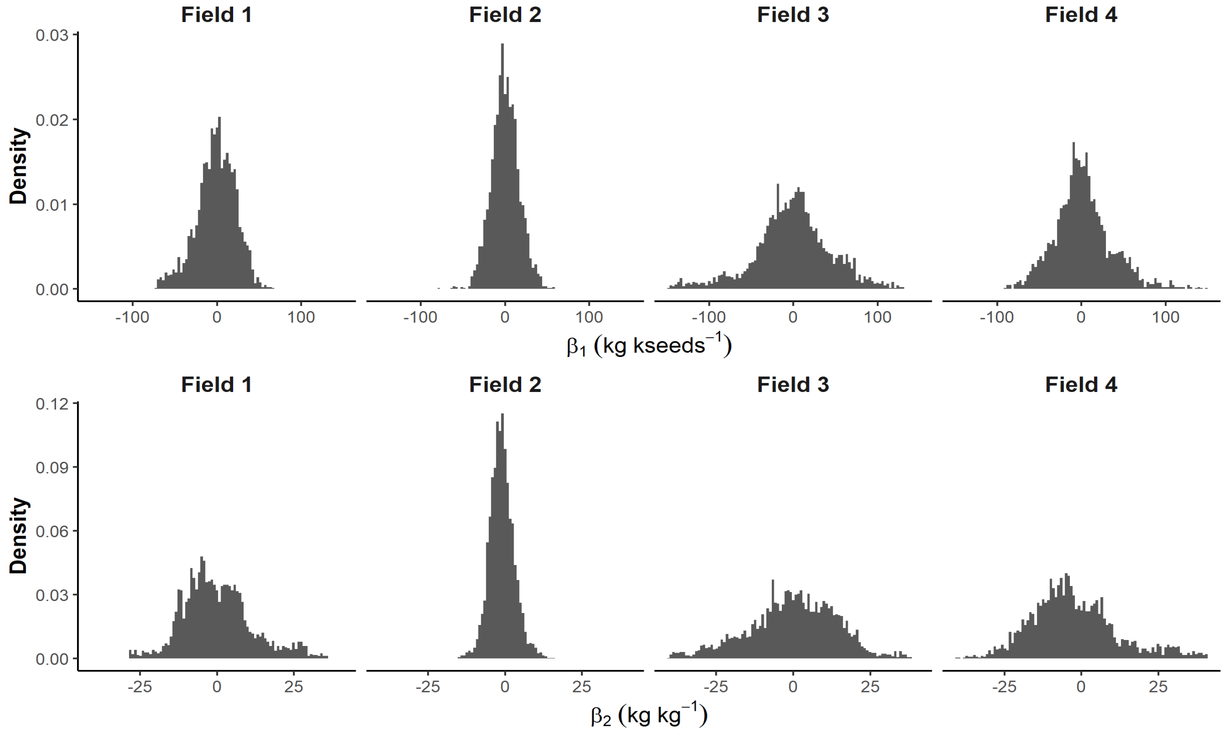

Figure 2.6. Frequency distribution of the fitted parameters of local yield response to (a) as-planted seed rates, and (b) as-applied nitrogen rates, in the four on-farm precision experiments.

The spatial variation of β_1 and β_2 can also be interpreted as proxies or latent variables that account for the combined effects that crop development, soil parameters, weather factors, and their interactions exert on the crop response to each input. The use of proxies of yield response is a common practice in most of the agronomic recommendations, since measuring all the factors influencing yields is either prohibitively expensive or virtually impossible. To be effective, the proxy variable has to be well correlated with the crop response, which may not always be the case for variables such as electrical conductivity or vegetation indices. The advantage of this type of OFPE and the GWR analysis is that the crop itself is used as a "sensor" to estimate the spatial variabilities of the crop response and the optimal rates. The use of soil and weather characteristics may still be necessary for modeling yield response under different growing conditions, such as in other fields and years. The main difference is that the focus is no longer yield prediction; instead, the target variable could be the yield response, optimal rates, or simply the β_1 and β_2.

The distribution of these parameters is the main difference between the yield response functions in each location, and are the main drivers of the differences in optimal input rates.

Therefore, the statistical significance of these variations is strong evidence of the within-field spatial variability of crop responses to agronomic inputs. The maps in Figure 2.7 show the regions where that was the case. The significance is reflected in the raw p-values, and in the p-values adjusted by the Fotheringham-Byrne procedure (Lu et al., 2014). The percentage of locations with significantly different yield responses to (N, S) varies among fields, with Field 3 showing the most substantial variation. With the conservative protection given by the Fotheringham-Byrne procedure, less than 10% of the locations remain as significant in Fields 1 and 2. In general, no more than 50% of the locations are significantly different from the average. If temporal variability can be ignored, these maps could be used as management zones, clearly showing which parts would benefit the most from site-specific management. Considering temporal variability is possible, but beyond the scope of the research reported here.

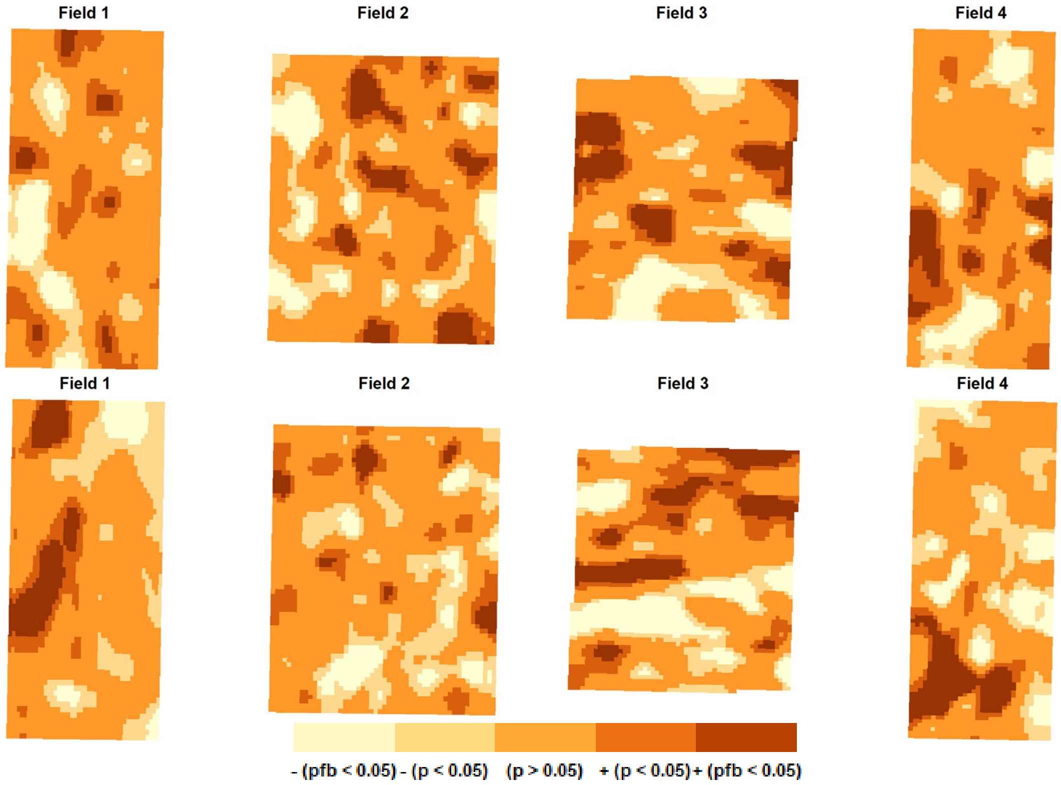

Figure 2.7. Spatial distribution of the statistical significance of the fitted parameters of local yield response to (a) seed rates (β_1), and (b) nitrogen rates (β_2), in the four on-farm precision experiments. Darker colors represent positive values, and lighter colors represent negative values of the parameters. "Pfb" denotes the p-value adjusted by the Fotheringham-Byrne procedure.

Comparison of crop management scenarios

The predicted yields at the *status quo* rates showed smooth variations across the field (Figure 2.8). These are the best estimations of the reference yield, and the variation is mainly due to differences in soil parameters, which affect yield both directly and indirectly through interactions among soil characteristics and weather.

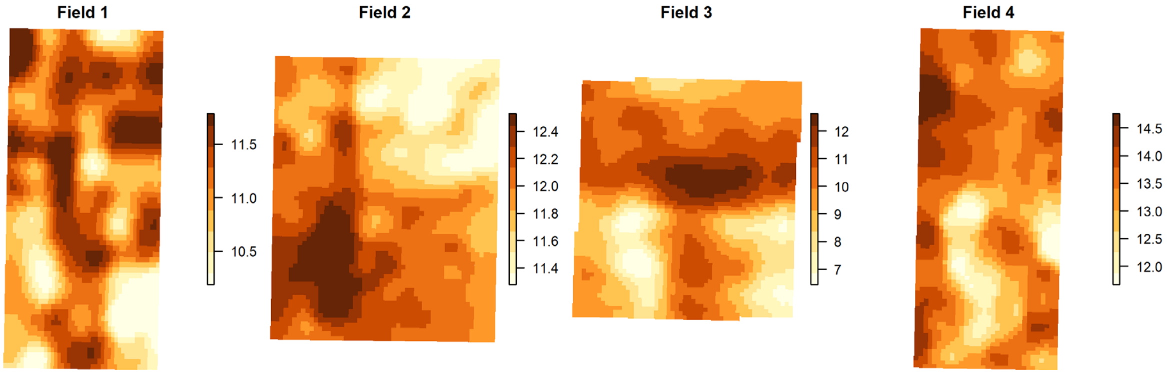

Figure 2.8. Spatial distribution of predicted corn yield (Mg ha^{-1}) if a uniform *status quo* rate of seed and nitrogen had been used on each of the four fields in the on-farm precision experiments.

The estimated optimal seed rates ranged from less than 70 kseed ha^{-1} to the maximum of the experimental seed rates, slightly more than 90 kseed ha^{-1} (Figure 2.9a). The number of fields evaluated was not sufficient to conclude how weather influenced crop response variability in each year. By examining the cumulative frequency distributions of optimal seed rates (Figure 2.9a), it is possible to conclude that the optimal rate was lower than the *status quo* rate in over 85% of Field 2, but in less than 15% of Field 3. The range of variation in optimal nitrogen rates was about 100 kg ha^{-1} for Field 3 and only 40 kg ha^{-1} for Field 1 (Figure 2.9b). However, the optimal nitrogen rate was higher than the *status quo* rate in the totality of Field 1, and the estimates of optimum rates were limited by the maximum rates used in the experiment, thus restricting the perceived variability.

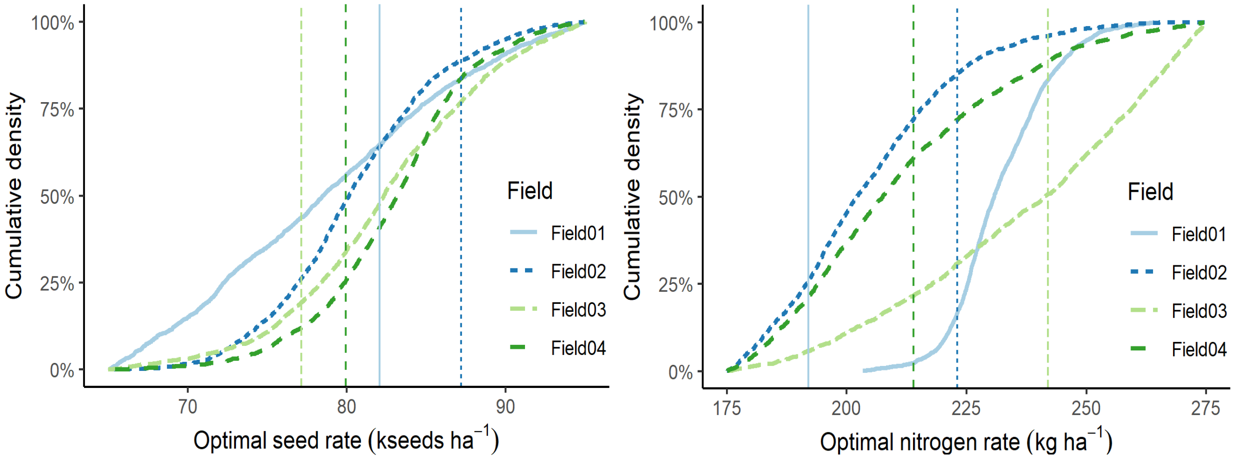

Figure 2.9. Cumulative frequency distribution of optimal (a) seed and (b) nitrogen rates. Dashed vertical lines represent the fields' *status quo* rates.

For the four fields' geographic regions, the Maximum Return to Nitrogen MRTN program's recommended rates, which are the "best practices" promoted by several midwestern universities, ranged between 180 and 220 kg ha^{-1} (Sawyer et al., 2006). For Fields 2 and 4, the higher end of this range contained this study's estimated economically optimal N rates. However, the empirical results predict that Fields 1 and 3 could have benefitted from even higher rates. Note that Field 3 had the lowest yields (Table 2.2) and the highest optimal nitrogen rates among the four fields, providing additional evidence against yield-based N management recommendation algorithms.

The patterns of the spatial variability for the optimal rates of seed and nitrogen (Figure 2.10a,b) are consistent with the significance of the local coefficients presented in Figure 2.7. In some regions, there is an indication of a weak negative correlation between both inputs, with low optimal seed rates being more frequently associated with higher optimal nitrogen rates. There were also many locations where the optimal rates were either high or low for both inputs. It is possible that in the regions with a negative correlation, the main driver for the optimal nitrogen rates was the soil availability, while in the region with a positive correlation, the main driver was the plant demand (Morris et al., 2018). For the reasons outlined above, understanding the reasons behind the crop response is important, and a key limitation of GWR analysis is that it cannot provide such understanding. To draw robust conclusions about the causes of yield response, more fields and weather scenarios should be investigated, and formal statistical methods need to be applied. This is left for future research.

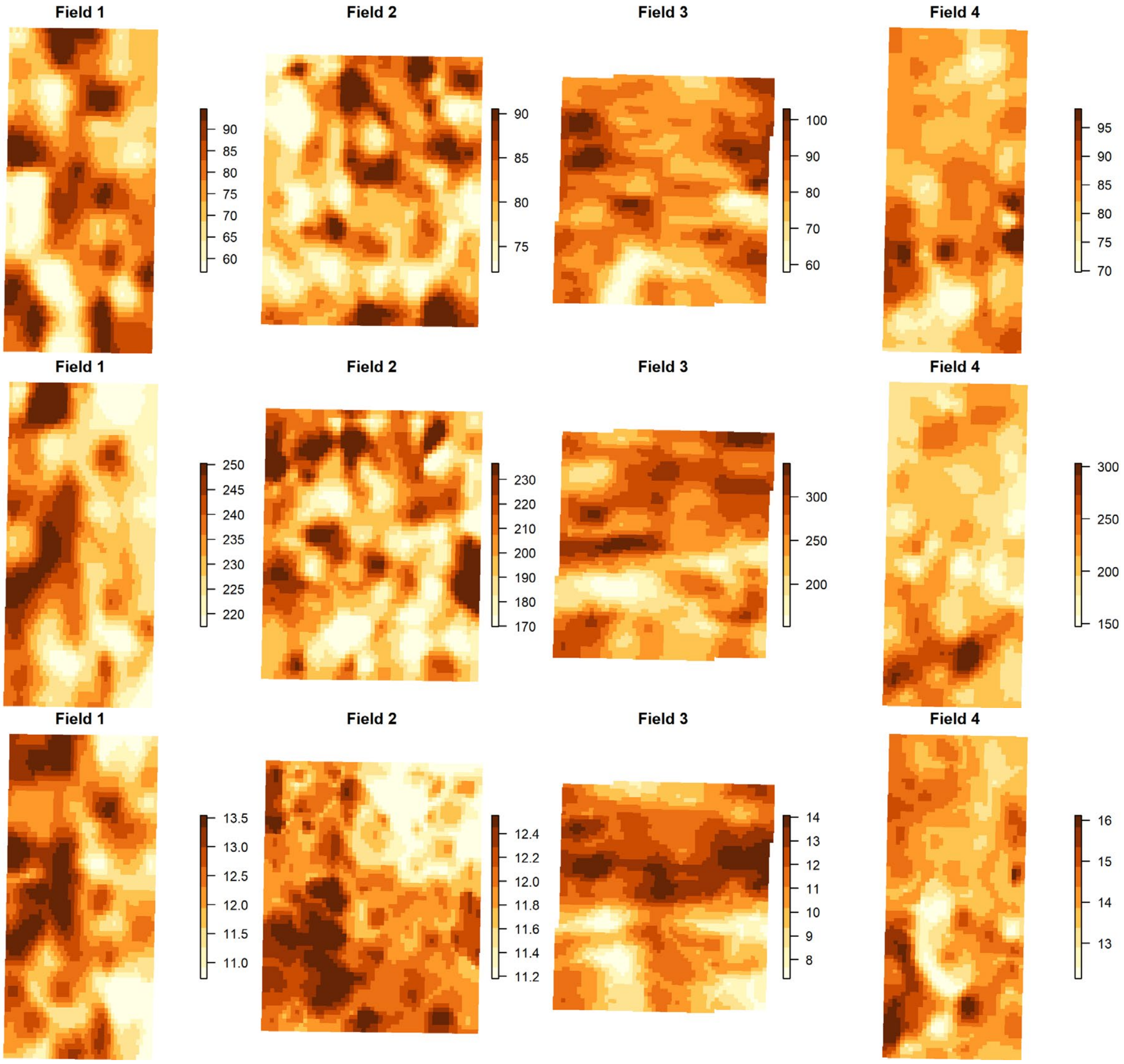

Figure 2.10. Spatial distribution of predicted optimal rates for (a) seed (kseed ha^{-1}), and (b) nitrogen (kg ha^{-1}) and predicted corn yield (Mg ha^{-1}) if the optimized rates of seed and nitrogen had been used in the fields.

The resulting range of the spatial variability of optimal input rates was shorter than the range of the reference yield variability. This again raises questions about the efficacy of using yield-based management zones to establish prescriptions for variable rate input rate management; for the four fields examined here, yield-based algorithms would recommend similar management strategies for areas having similar yields but different yield *responses* to managed inputs (Rodriguez et al., 2019). The results also provide an understanding of why other strategies, such as the nitrogen-rich strip and the ramped calibration methods used to calibrate sensor-based nitrogen applications, often fail to provide reliable nitrogen requirement predictions (Roberts et al., 2011; Colaço and Bramley, 2019). Since they are allocated with no previous knowledge of crop response, and the range of the spatial variability is limited, recommendations based on the nitrogen-rich strip will be highly variable, depending on the field to which it was applied.

Figure 2.11 illustrates the impact of using site-specifically optimized input rates. The difference between the reference yield and the optimized yield reflects the potential for improving yields by applying site-specifically according to the optimal rate of each portion of the field. There are no distinguishable differences in yield for the different scenarios in Field 2, indicating the *status quo* rates were already close to the optimal uniform rates for the whole field. The yield improvements were also observed in the high yielding areas of Field 1. The more compact the distributions are, the smaller is the yield variability. Comparing the cumulative frequency distribution of the reference yield and the optimized yield for Field 1, it is also possible to note that the variability of the optimized yields is higher than the variability in the reference yield. This shows that the objective of variable rate application is not to make the yield more homogenous, but rather to accept the spatial variability and take advantage of it by maximizing profits at every location within the field.

The spatial variability of the optimized yields (Figure 2.10c) closely follows the variations in the reference yield (Figure 2.8). This is related to the observation that most of the yield variability cannot be explained by controllable inputs (Kindred et al., 2017). In some areas, such as in Field 4, it is possible to observe some “hot-spots” with high yields, associated with high optimal rates of the inputs (Figure 2.10). These are the locations where the local estimates of crop response had the higher values and could be due to a combination of factors that make those areas highly responsive, or due to random chance and the noise in the measurements. Statistical methods such as bootstrap may be used to estimate the confidence intervals of these estimations and refine the local predictions by excluding some observations from the analysis (Harris et al., 2017).

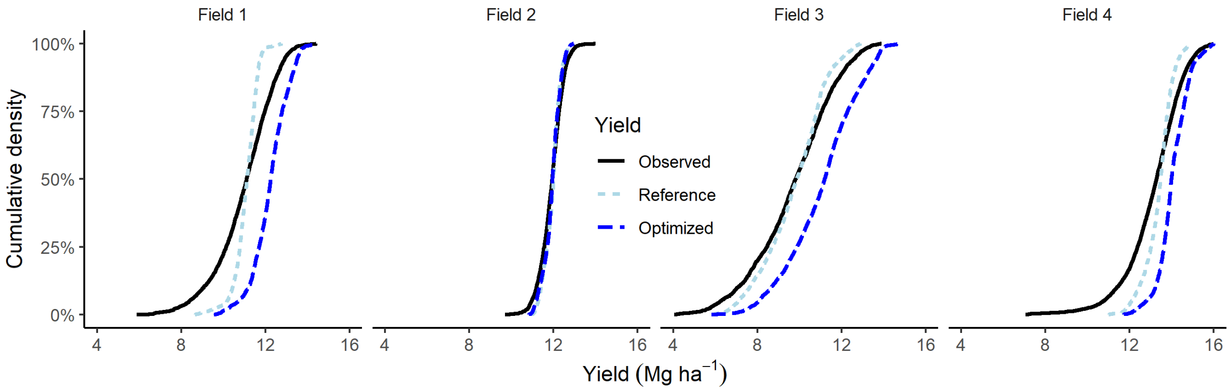

Figure 2.11. Cumulative frequency distribution of yields comparing the observed yield, the expected yield if the *status quo* rates had been applied and, the expected yield if the optimized rates of seed and nitrogen had been applied in each field.

Economic results

The estimation of any opportunity index of the potential economic benefits of spatially adjusting crop inputs requires the spatial resolution used to characterize the variability to be compatible with the resolution at which the recommendations can be applied (Leroux and Tisseyre, 2018). The economic opportunity presented here implicitly considers the machinery sizes as given in the analysis because the spatial resolution of the data was defined by the ability of the available machinery to apply the assigned treatments.

Adjusting the uniform rates of both inputs would have an economic impact substantially greater than the loss of revenue due to the suboptimal trial rates, except for Field 4, (Table 2.5). Considering the eight possible combinations of fields and inputs, in five of them, profits from the *status quo* rate were within US$ 11.00 from the maximized profit. The spatial homogeneity of Fields 1 and 2 resulted in small differences in the local yield response to the agronomic inputs. For Fields 3 and 4, taking into account the spatial variability of optimal rates would have generated a positive economic impact greater than adjusting the optimal uniform rate for both seed and nitrogen rates. The average increase in net revenue of using the site-specific optimal seed and nitrogen rates would be US$ 17.00 ha^{-1} and US$ 36.00 ha^{-1}, respectively. These results are similar to the overall average profit of US$ 30.00 ha^{-1} reported with the use of crop sensors to drive variable rate nitrogen application (Colaço and Bramley, 2018).

Table 2.5. Summary of net revenue losses due to yield loss in suboptimal treatment rates, and potential economic benefits if an optimized uniform or variable rate had been used instead of the farmer's *status quo* rates.

Field*	**Revenue losses**	**Optimal uniform rate**			**Optimal variable rate**			**VR-UR**		
		SR	***NR***	***SRNR***	***SR***	***NR***	***SRNR***	***SR***	***NR***	***SRNR***
Field 1	-17	43	129	137	55	131	169	12	2	32
Field 2	-7	11	11	27	19	18	39	8	7	13
Field 3	-9	1	58	79	34	122	181	32	64	102
Field 4	-41	11	3	16	28	76	103	17	73	87

*All values are in US\$ ha^{-1} at a corn price of US\$ 160.00 Mg^{-1} (US\$ 4.00 bu^{-1}). SR: seed rate; NR: nitrogen rate; SRNR: seed and nitrogen rate; VR-UR: Difference in profit of applying the optimal variable rate compared to a uniform optimal rate.

Although the number of fields used in this work was adequate for demonstrating the method and making initial considerations, the extrapolation of these results will require more data. Nevertheless, the tendency of OFPE resulting in higher economic benefits on fields with higher yield variability, even though yield levels and optimal input rates are not necessarily correlated, is a promising result. This tendency means that fields with higher yield variability should be prioritized for the implementation of whole field OFPE, while in fields with less variability, the experiments may be replicated only in a small part of the field.

The main question that remains is the extent to which the results from one year of data can be used to guide the decisions of the next year, especially considering that some input strategies, such as seed rate, cannot be changed in-season. Answering that question requires knowledge of the temporal stability in the spatial variability of crop responses to each input, which is not extensively explored in the scientific literature. The majority of the related studies focus only on the temporal stability of the spatial variability in yields, which interacted with the weather of each growing season (Maestrini and Basso, 2018). Many studies and reviews have focused on within- and among-field variations in optimal nitrogen rates, and have concluded that most of the variability is related to dynamic variables, such as precipitation and temperature (Puntel et al., 2019); and their interaction with field-specific variables, such as previous crop, tillage practice, soil drainage class, and N form and timing (Morris et al., 2018; Tao et al., 2018).

In terms of the model described in Equation [3], the results available in the literature suggest that at least the global βs will be season-specific, and therefore may need to be adjusted for every season. Depending on the degree of temporal variability in β_1 and β_2, there could be no need to repeat the whole field experiment every year. Instead, a small subset of plots in representative areas of the field could be used to adjust the other parameters of the equation. Even more promising, the season-specific parameters may be adjusted by the use of other tools such as crop modeling systems and pre-plant soil nitrate tests. Optimizing the trial designs to account for the temporal variability should also be explored in future research. The relative importance of temporal and spatial variability in the overall crop response variability is likely to be different for each field, crop and, input considered. The methods presented here have the potential to be applied in many other scenarios to improve management decisions by accounting for these sources of variability.

CONCLUSIONS

The main contribution of this work is to demonstrate an alternative method to test and characterize the spatial variability of crop response to inputs. The combination of OFPE and GWR proved to be an effective methodology to test precision agriculture central hypothesis' of whether there is significant within-field variability in optimal rates. It also allows changing the focus from yield-based to response-based variable-rate prescriptions for crop input application. Future research on trial design and models with spatially varying coefficients for OFPE is advised.

Incorporating spatial heterogeneity of yield responses into model parameters improved model performance in all four fields evaluated. On average, the RMSE of the fitted yield decreased from 1.2 Mg ha^{-1} in the non-spatial model to 0.7 Mg ha^{-1} in the GWR model, and the r-squared increased from 10% to 68%. In 10% to 50% of the observations, the coefficients of the local parameters were found to be significantly different from the average, providing further evidence of the need for increased knowledge about local yield response functions.

In Fields 1 and 2 the greatest benefits of OFPE would come from optimizing the field's uniform rate, while in Fields 3 and 4 the highest revenue increases would account for the spatial variability in crop response, and implementing the site-specifically optimal variable rates would result in the highest increase in revenue. The average potential gain of using optimized uniform rates of seed and nitrogen was US$ 65.00 ha^{-1}, while the added potential gain of using variable rate application was US$ 58.00 ha^{-1}.

REFERENCES

Anselin, L. 2010. Thirty years of spatial econometrics. Pap. Reg. Sci. 89(1): 3–25. doi: 10.1111/j.1435-5957.2010.00279.x.

Bachmaier, M., and M. Gandorfer. 2009. A conceptual framework for judging the precision agriculture hypothesis with regard to site-specific nitrogen application. Precis. Agric. 10(2): 95–110. doi: 10.1007/s11119-008-9069-x.

Bachmaier, M., and M. Gandorfer. 2012. Estimating Uncertainty of Economically Optimum N Fertilizer Rates. Int. J. Agron. 2012: 1–10. doi: 10.1155/2012/580294.

Brunsdon, C., A.S. Fotheringham, and M.E. Charlton. 1996. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. Geogr. Anal. 28(4): 281–298. doi: 10.1111/j.1538-4632.1996.tb00936.x.

Bullock, D.S., M. Boerngen, H. Tao, B. Maxwell, J.D. Luck, et al. 2019. The Data-Intensive Farm Management Project: Changing Agronomic Research Through On-Farm Precision Experimentation. Agron. J. 111(6): 2736. doi: 10.2134/agronj2019.03.0165.

Bullock, D.S., and D.G. Bullock. 1994. Calculation of Optimal Nitrogen Fertilizer Rates. Agron. J. 86(5): 921–923. doi: 10.2134/agronj1994.00021962008600050030x.

Bullock, D.S., J. Lowenberg-DeBoer, David S. Bullock, James Lowenberg-DeBoer, D.S. Bullock, et al. 2007. Using spatial analysis to study the values of variable rate technology and information. J. Agric. Econ. 58(3): 517–535. doi: 10.1111/j.1477-9552.2007.00116.x.

Cai, R., D. Yu, and M. Oppenheimer. 2014. Estimating the spatially varying responses of corn yields to weather variations using geographically weighted panel regression. J. Agric. Resour. Econ. 39(2): 230–252. doi: 10.22004/ag.econ.186586.

Colaço, A.F., and R.G.V. Bramley. 2018. Do crop sensors promote improved nitrogen management in grain crops? F. Crop. Res. 218(January): 126–140. doi: 10.1016/j.fcr.2018.01.007.

Colaço, A.F., and R.G.V. Bramley. 2019. Site-Year Characteristics Have a Critical Impact on Crop Sensor Calibrations for Nitrogen Recommendations. Agron. J. 111(4): 2047–2059. doi: 10.2134/agronj2018.11.0726.

Fotheringham, A.S. 1997. Trends in quantitative methods I: stressing the local. Prog. Hum. Geogr. 21(2): 283–292. doi: 10.1191/030913299667756016.

Geniaux, G., and D. Martinetti. 2018. A new method for dealing simultaneously with spatial autocorrelation and spatial heterogeneity in regression models. Reg. Sci. Urban Econ. 72(April 2017): 74–85. doi: 10.1016/j.regsciurbeco.2017.04.001.

Gollini, I., B. Lu, M. Charlton, C. Brunsdon, and P. Harris. 2015. GWmodel: an R package for exploring spatial heterogeneity using geographically weighted models. J. Stat. Softw. 63(17). doi: 10.1080/10095020.2014.917453.

Harris, P., C. Brunsdon, B. Lu, T. Nakaya, and M. Charlton. 2017. Introducing bootstrap methods to investigate coefficient non-stationarity in spatial regression models. Spat. Stat. 21: 241–261. doi: 10.1016/j.spasta.2017.07.006.

Hurley, T.M., K. Oishi, and G.L. Malzer. 2005. Estimating the potential value of variable rate nitrogen applications: A comparison of spatial econometric and geostatistical models. J. Agric. Resour. Econ. 30(2): 231–249. doi: 10.22004/ag.econ.31210.

Kindred, D.R., R. Sylvester-Bradley, A.E. Milne, B. Marchant, D. Hatley, et al. 2017. Spatial variation in Nitrogen requirements of cereals, and their interpretation. Adv. Anim. Biosci. 8(02): 303–307. doi: 10.1017/S2040470017001327.

Lark, R.M., J.V. Stafford, and H.C. Bolam. 1997. Limitations on the Spatial Resolution of Yield Mapping for Combinable Crops. J. Agric. Eng. Res. 66(3): 183–193. doi: 10.1006/jaer.1996.0132.

Lark, R.M., and H.C. Wheeler. 2003. A Method to Investigate Within-Field Variation of the Response of Combinable Crops to an Input. Agron. J. 95(5): 1093–1104. doi: 10.2134/agronj2003.1093.

Leroux, C., and B. Tisseyre. 2018. How to measure and report within-field variability: a review of common indicators and their sensitivity. Precis. Agric. 20(3): 562–590. doi: 10.1007/s11119-018-9598-x.

Lloyd, C.D. 2009. Nonstationary models for exploring and mapping monthly precipitation in the United Kingdom. Int. J. Climatol. 30(3): n/a-n/a. doi: 10.1002/joc.1892.

Lu, B., P. Harris, M. Charlton, and C. Brunsdon. 2014. The GWmodel R package: Further topics for exploring spatial heterogeneity using geographically weighted models. Geo-Spatial Inf. Sci. 17(2): 85–101. doi: 10.1080/10095020.2014.917453.

Lu, B., W. Yang, Y. Ge, and P. Harris. 2018. Improvements to the calibration of a geographically weighted regression with parameter-specific distance metrics and bandwidths. Comput. Environ. Urban Syst. 71(April): 41–57. doi: 10.1016/j.compenvurbsys.2018.03.012.

Maestrini, B., and B. Basso. 2018. Drivers of within-field spatial and temporal variability of crop yield across the US Midwest. Sci. Rep. 8(1): 1–9. doi: 10.1038/s41598-018-32779-3.

Morris, T.F., T.S. Murrell, D.B. Beegle, J.J. Camberato, R.B. Ferguson, et al. 2018. Strengths and limitations of Nitrogen rate recommendations for corn and opportunities for improvement. Agron. J. 110(1): 1–37. doi: 10.2134/agronj2017.02.0112.

Murakami, D., and D.A. Griffith. 2019. Spatially varying coefficient modeling for large datasets: Eliminating N from spatial regressions. Spat. Stat. 30: 39–64. doi: 10.1016/j.spasta.2019.02.003.

Murakami, D., B. Lu, P. Harris, C. Brunsdon, M. Charlton, et al. 2018. The Importance of Scale in Spatially Varying Coefficient Modeling. Ann. Am. Assoc. Geogr. 109(1): 1–21. doi: 10.1080/24694452.2018.1462691.

Pahlmann, I., U. Böttcher, and H. Kage. 2016. Evaluation of small site-specific N fertilization trials using uniformly shaped response curves. Eur. J. Agron. 76: 87–94. doi: 10.1016/j.eja.2016.01.017.

Piepho, H.-P.P., C. Richter, J. Spilke, K. Hartung, A. Kunick, et al. 2011. Statistical aspects of on-farm experimentation. Crop Pasture Sci. 62(9): 721–735. doi: 10.1071/CP11175.

Plant, R.E. 2012. Spatial data analysis in ecology and agriculture using R. CRC Press.

Pringle, M.J., S.E. Cook, and A.B. McBratney. 2004a. Field-scale experiments for site-specific crop management. Part I: Design considerations. Precis. Agric. 5(6): 617–624. doi: 10.1007/s11119-004-6346-1.

Pringle, M.J., A.B. McBratney, and S.E. Cook. 2004b. Field-scale experiments for site-specific crop management. Part II: A geostatistical analysis. Precis. Agric. 5(6): 625–645. doi: 10.1007/s11119-004-6347-0.

Puntel, L.A., A. Pagani, and S. V. Archontoulis. 2019. Development of a nitrogen recommendation tool for corn considering static and dynamic variables. Eur. J. Agron. 105(March): 189–199. doi: 10.1016/j.eja.2019.01.003.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. https://www.r-project.org/.

Roberts, D.C., B.W. Brorsen, R.K. Taylor, J.B. Solie, and W.R. Raun. 2011. Replicability of nitrogen recommendations from ramped calibration strips in winter wheat. Precis. Agric. 12(5): 653–665. doi: 10.1007/s11119-010-9209-y.

Rodriguez, D.G.P., D.S. Bullock, and M.A. Boerngen. 2019. The Origins, Implications, and Consequences of Yield-Based Nitrogen Fertilizer Management. Agron. J. 111(2): 725–735. doi: 10.2134/agronj2018.07.0479.

Sawyer, J., E. Nafziger, G. Randall, L. Bundy, G. Rehm, et al. 2006. Concepts and rationale for regional nitrogen rate guidelines for corn concepts and rationale for regional nitrogen rate guidelines for corn. Iowa State Univ. Univ. Ext. (April 2006): 1–28.

Scharf, P.C., N.R. Kitchen, K.A. Sudduth, and J.G. Davis. 2006. Spatially variable corn yield is a weak predictor of optimal nitrogen rate. Soil Sci. Soc. Am. J. 70(6): 2154–2160. doi: 10.2136/sssaj2005.0244.

Scharf, P.C., N.R. Kitchen, K.A. Sudduth, J.G. Davis, V.C. Hubbard, et al. 2005. Field-scale variability in optimal nitrogen fertilizer rate for corn. Agron. J. 97(2): 452–461. doi: 10.2134/agronj2005.0452.

Tao, H., T.F. Morris, P. Kyveryga, and J. McGuire. 2018. Factors affecting nitrogen availability and variability in cornfields. Agron. J. 110(5): 1974–1986. doi: 10.2134/agronj2017.11.0631.

Thöle, H., C. Richter, and D. Ehlert. 2013. Strategy of statistical model selection for precision farming on-farm experiments. Precis. Agric. 14(4): 434–449. doi: 10.1007/s11119-013-9306-9.

Trevisan, R.G., D.S. Bullock, and N.F. Martin. 2019a. Site-specific treatment responses in on-farm precision experimentation. Precision agriculture '19. Wageningen Academic Publishers, The Netherlands. p. 925–931

Trevisan, R.G.G., L.S. Shiratsuchi, D.S. Bullock, and N.F. Martin. 2019b. Improving yield mapping accuracy using remote sensing. Precis. Agric. '19 (January): 14–15. doi: 10.3920/978-90-8686-888-9_111.

CHAPTER 3

DEVELOPMENT OF A DATA-DRIVEN DECISION SUPPORT SYSTEM FOR MAIZE IN MEXICO

ABSTRACT

The need to improve management decisions in crop production systems has not been fully achieved by conventional agricultural research and extension methodologies, especially in more complex agricultural production systems of small farms in tropical regions. Machine learning methods offer a promising alternative to model the relationships between environmental and management variables with crop yields, which can be embedded into decision support tools to provide recommendations to crop advisors and farmers. However, trusting the model predictions without understanding what type of relationships they learned is not advisable. Therefore, it becomes essential to explain why individual predictions were made and the factors that contribute the most to those predictions. The objective of this work was to develop a data-driven decision support system for maize in Mexico. The data comes from the Sustainable Modernization of Traditional Agriculture (MasAgro) project in the southern state of Chiapas. The dataset was assembled using field observations, including yield, cultivars and management, and environment variables from soil mapping and gridded weather datasets. Random forest models were trained with the dataset and explained up to 75% of the variation. However, the ability of the model to predict crop performance in future weather scenarios was limited. Domain knowledge and explainable machine learning methods allowed the use of the model as a source of information to create and validate hypotheses. Overall, nitrogen was the management decision that influenced yields the most, with different yield responses depending on the year and variety. This research exemplifies the use of explainable machine learning to offer farmers the opportunity to benchmark their management decisions with peers in similar growing conditions and visualize what would have happened if they made different decisions.

INTRODUCTION

Crop management recommendations should align production systems and environmental characteristics. The lack of knowledge of limiting factors and is one of the main reasons for the inefficient use of inputs and low productivity and profitability of agricultural systems. Traditionally, this knowledge has been generated with small-plots highly controlled agricultural experimentation. The results from these small -scale experiments are then diffused to farmers with linear extension services (Chambers and Jiggins, 1987). Although this agricultural research and extension methodology helped improve many production systems worldwide, it has faced limitations in many contexts, especially in more complex agricultural production systems of small farms in tropical regions. This complexity extends beyond the main production factors (genetics, environment, management), including access to resources, vulnerability to risk, labor supply, market opportunities, household needs, management ability, and cultural beliefs (Wortmann et al., 2020).

The development of farmer participatory research and plant breeding (Humphries et al., 2015; Camacho-Villa et al., 2016; Snapp et al., 2019; Eldon et al., 2020), the use of crowdsourced data from the farmers (Schmidt et al., 2018; van Frank et al., 2019), along with the ability to collect and process large amounts of data (Donnet et al., 2017; Cui et al., 2018), and the adoption of information and communication technologies by farmers (Steinke et al., 2020), creates an opportunity to reimagine agricultural research and extension. Using enough data, proper analytical methods, and efficient, scalable tools, it is possible to deliver information and recommendations specific to each farmer's field (Jiménez et al., 2019). Effectively, these innovations enable the diversity of treatment responses in heterogeneous agricultural production systems to be embraced rather than avoided (Vanlauwe et al., 2019).

The focus of research innovations in highly mechanized extensive production systems has been on better accounting for the within-field spatial variability of optimal management decisions (Bullock et al., 2019; Barbosa et al., 2020; Trevisan et al., 2020). However, in the smallholder production environment, the between-field spatial variability and the temporal variability are substantially more important (Jiménez et al., 2019; Vanlauwe et al., 2019; Eldon et al., 2020). The temporal variability of optimal crop management strategies is mainly a function of the weather and its interactions with the soil and crop genetics. The year-to-year variations in growing conditions increase the challenges of making decisions that result in the best outcome. The effects of climate change and weather variability have caused adverse impacts on agricultural production and food security in recent years (de Sousa et al., 2018; Westermann et al., 2018). Even though farmers cannot control the weather, knowing how it interacts with factors that can be managed is useful to help them decide on the management strategies to be adopted. Changing the planting date or choosing cultivars more adapted to the expected weather are strategies that can usually be implemented at low cost and result in higher returns (Delerce et al., 2016). Many climate-smart adaptation strategies have been suggested. However, their acceptance depends on context-specific recommendations.

In this context, the Sustainable Modernization of Traditional Agriculture (MasAgro) in Mexico is a large-scale program that goes beyond participatory plant breeding and addresses the whole production system, emphasizing crop management and conservation agriculture practices (Camacho-Villa et al., 2016). The MasAgro project started in 2010 aiming to increase maize productivity of rain-fed areas in Mexico and enhance the country's maize self-sufficiency as a joint effort between the International Maize and Wheat Improvement Center (CIMMYT) and the Mexican Government's Secretariat of Agriculture, Livestock, Rural Development, Fisheries and Food (SAGARPA). The maize cropping systems in Mexico are heterogeneous, with approximately 6 million hectares of maize covering seven different rain-fed maize growing regions. Farmers in Mexico have been growing maize for centuries and have been adjusting their practices to optimize their results for multiple purposes under diverse growing conditions. In general, there are two types of maize farmers, the traditional and the commercial, who apply different management practices (Donnet et al., 2017). This unique dataset capturing the diversity of Mexican agriculture represents many opportunities to showcase the use of data-driven decision support systems to address the need for time and location-specific recommendations.

Mobile phone-based services have been considered an ideal tool to complement traditional extension services and enable smallholder farmers to access context-specific information, increasing their decision-making capacities (Inwood and Dale, 2019; Ortiz-Crespo et al., 2020). As part of the MasAgro effort, CIMMYT and the International Institute for Applied Systems Analysis (IIASA) developed the smartphone application AgroTutor (Bayas et al., 2020). It is freely available, allowing farmers to geolocate and register plots by using the phone GPS and collect insitu information like soil management and yield data, which are then used to develop and improve models. In return, farmers have access to highly specific and timely agricultural recommendations, potential yield and financial information, historical and forecasted weather data, and other agricultural information sources (Bayas et al., 2020).

The models relating environmental and management variables to crop yield powering the recommendations offered to crop advisors and farmers in this type of decision support tool can usually be separated into statistical and analytical crop models (Jones et al., 2017). Analytical crop models are dynamic system simulations based on the physiological processes driving crop growth and development. These models' results can be accurate, but only if the model parameters are correctly calibrated, which requires data not measurable by farmers. In general, those models have been used to estimate potential production, which serves as a benchmarking tool to compare the farmer's results with what was possible with perfect management decisions, to compute resource use efficiency indicators, and to conduct yield gap analysis (Bayas et al., 2020; Riccetto et al., 2020).

Historically, statistical models have been more frequent in controlled experiments to understand the optimal management decisions for a single production factor, such as the variety and the amount of fertilizer. The use of these methods requires a pre-defined trial design and assumes identical and independently distributed residuals. Applying these same methods with on-farm trials and field production data is challenging because most of the data have no replication or randomization and can even be entirely observational, in which no deliberate treatment was applied. The differences in the spatial and temporal resolution of data sources and spatiotemporal correlation add other layers of complexity. However, in recent years the increased ability to collect large amounts of observational data in farmers' fields and the advances in statistical learning methods renewed the research community interest in this topic. Many studies combining multiple sources of information and expert-guided machine learning (ML) methods have offered alternatives to understand these observational datasets and uncover relationships that can be used to predict crop yields and optimize crop management decisions (Delerce et al., 2016; Dorado et al., 2018; Jiménez et al., 2019).

These works follow a similar structure, in which a model is fitted to minimize the yield prediction error. Then different optimization techniques are used to derive the best management practices. In most of these, the authors reiterate the importance of expert knowledge to validate the results since they are often not meaningful (Jiménez et al., 2019). One of the reasons the model results may be misleading is the presence of spurious correlations among the variables used, which may be caused by other factors that were not measured or not included in the model. This shows a significant difference between the two groups of tools used to create decision support systems. While analytical crop models rely on causation, the statistical crop models rely almost exclusively on correlation (Jones et al., 2017). In many cases, the relationships learned by the model would be considered by humans as cheating rather than valid problem-solving (Lapuschkin et al., 2019).

Similar limitations in using ML models have been identified in other fields of research, with models and algorithms leading to severe violations of fairness and ethical principles, which raised legal exigences and prohibitions in their use (Goodman and Flaxman, 2017). Explainable ML can be defined as using domain knowledge to explain black-box model decisions with a certain level of detail, making systems more understandable and transparent, and is considered a prerequisite to ensure the results' scientific consistency (Roscher et al., 2020). These concepts have been recently incorporated in agriculture research using ML methods for yield prediction and land use classification (Campos-Taberner et al., 2020; Wolanin et al., 2020). One of the advantages of explainable ML is the possibility of using the model as a source of information to create and validate hypotheses, which can be particularly useful in agricultural research (Lorentzen and Mayer, 2020).

OBJECTIVES

The main objective of this work was to develop a data-driven decision support system tool based on machine learning methods and multiple years of field data from the MasAgro project in Chiapas - Mexico to help farmers increase agriculture profitability and sustainability.

The specific objectives were:

- To develop machine learning-based models to represent the spatial and temporal variability of alternative management decisions based on observational data;
- To develop tools to evaluate model accuracy and uncertainty and to explain and visualize the results;

MATERIAL AND METHODS

Dataset

The data used in this work comes from the MasAgro project (Donnet et al., 2017). The program, which consists of 12 innovation hubs located strategically throughout Mexico, seeks to integrate farmers and local and regional value-chain actors with an innovation system approach. The hubs comprise four levels of agronomic experimentation: research platforms are typical on-station controlled experiments; demonstration modules are on farmers' land and involve side-by-side fields comparisons of different technologies and management practices; extension areas are fields where farmers have implemented management changes after testing them in demonstration modules; impact areas are other places where farmers adopted MasAgro's innovations without being directly connected to the hubs (Molina-Maturano et al., 2020). The research platforms follow a valid statistical design and aim to test a specific hypothesis. The impact areas often do not have information about all the farmers' practices, thus have limited value to model yield response. Therefore, the data used in this work comes from the demonstration modules and extension areas, which lack the replications needed in classical statistical models while still useful for ML models.

The dataset consists of eight years of field records of maize production by Mexican farmers in the state of Chiapas. Crop production in this hub is characterized by rainfed maize farming systems with a mix of small-scale low-input self-consumption farmers and medium-scale medium-input mechanized semi-commercial farmers selling to local markets (Camacho-Villa et al., 2016). These two systems were differentiated in the dataset by the use of landraces in the low-input systems and the use of hybrids in the medium and high-input systems and are used as synonyms in this text. Each observation represents one crop event, from planting to harvesting, with a set of correspondent practices in one parcel of land, which was usually a small field (< 5 ha) or part of the field for a given season (year). The variables consisted mainly of the harvested yield, the variety or hybrid planted, the tillage method, and the fertilizer rates used (Table 3.1 and 3.2). The data was filtered to remove points with coordinates outside of the boundaries of the state, the observation with missing data for the variables used, and some extreme values. In the case of nitrogen, the observations with zero rate were also excluded because they likely represented missing data. Data wrangling was performed using the R software version 4.1 (R Core Team, 2020) and the package "dplyr" (Wickham et al., 2020).

Soil information was obtained from a polygon map of soil units with functional properties constructed using soil samples from Mexico's National Institute of Statistics and Geography (INEGI) open-access datasets (Delerce, 2018). The point coordinates from the field observations were spatially overlaid in the soil map polygons using the R software package "sf" (Pebesma, 2018). The soil attributes clay, cation exchange capacity (CEC), soil organic matter (SOM), and pH were extracted. Elevation and slope were derived from SRTM digital elevation data (Jarvis et al., 2008).

The weather dataset was assembled using DAYMET gridded daily surface weather data with 1 km spatial resolution (Thornton et al., 2016). The point coordinates from the field observations were spatially overlaid in the DAYMET gridded data using the R software package "raster" (Hijmans, 2020). The values correspondent to each pixel for precipitation, solar radiation, maximum temperature, minimum temperature, and vapor pressure were extracted. The daily data were aggregated into ten-day intervals to reduce the number of features used in the model. The weather data were organized according to the planting dates, starting 60 days before planting and running up to 240 days after planting, thus creating 30 new features for each variable.

Model fitting

The models were fitted using the random forest algorithm implemented in the R software package "ranger" (Wright and Ziegler, 2017). A grid search of hyperparameters was performed based on the cross-validation errors. Two different strategies of k-fold cross-validation were tested. The first assigns the data into testing and validation sets randomly, while the second used leave-one-year-out cross-validation, such that data from one year was used only for validation in each fold (Qin et al., 2018). Model training and cross-validation were performed with the help of the R software package "mlr3" (Lang et al., 2019). Different sets of features were also considered, which evaluated model performance with only controllable and static variables that do not change with time (cultivar, management, soil, topography) or with the inclusion of temporally dynamic variables, either represented by the weather features or the year as a categorical variable. Years were treated as a categorical factor with unordered levels. The time series was too short to only use observations from the past in the comparisons, although it is not possible to use future years to predict the past in practice. The yield was used as the dependent variable in all models, which were trained to minimize the mean squared error between the observed and predicted yields. Model performance was always evaluated in the testing set and was aggregated over the seven folds of the cross-validation.

Interpretability and visualization

Explainable machine learning is a relatively new area (Roscher et al., 2020). Methods are continually being updated, and new tools developed. The methods used in this work were selected based on their usefulness to extract meaningful information from the models and the wide availability of software implementations (Lorentzen and Mayer, 2020). Whenever possible, the interpretation methods were compared to those used in the more rigorous modeling framework of mixed models and other standard statistical tools.

For the purpose of developing a decision support system, model interpretation aims to answer questions such as which features (crop event characteristics) contributed the most to the explanatory or predictive power of the model, what would happen if the values of selected features were changed, and why a particular value was predicted for a specific crop event. These questions relate to the decomposition of model predictions variability into different explanatory variables and their interactions, either at the dataset level or to the level of individual observations. The general idea of decomposing the variance is similar to the analysis of variance (ANOVA) performed with statistical models. The lack of statistical design and the inability to test if residuals are independent and identically distributed is replaced to some extent by the use of permutations, conditional probabilities, and domain knowledge. More specifically, variable importance, interaction strength, partial dependence profiles, and variable attributions were calculated using the R software packages "DALEX" (Biecek, 2018) and "flashlight" (Lorentzen and Mayer, 2020). The graphical visualization of results was prepared using the R software packages "ggplot2" (Wickham, 2016).

RESULTS

Descriptive statistics

After removing crop events with missing values, the dataset had 4585 observations encompassing seven years (Table 3.1). The yield values ranged from 0.1 to 10.0 Mg ha^{-1}, with 75% of the values lower than 5.0 Mg ha^{-1} and an average of 3.6 Mg ha^{-1}. The variation in elevation was equally remarkable, ranging from almost sea level to about 3000 m, although observations were concentrated around the median of 700 m. The planting dates had a standard deviation of 20 days, with some observations up to five months apart from others. There were considerable differences between the number of observations recorded in some years, with a smaller number in 2014. It is also possible to observe that the operations shifted to later dates in some years, especially in 2015 (Figure 3.1). The total season duration had an average of 175 days. It is important to note that the corn is kept in the field for long periods after physiological maturity until it is used for self-consumption or sold in local markets. Therefore, the season length is not a good representation of the growing period.

More than 25% of farmers did not use phosphorus fertilizer, while more than 50% did not use potassium fertilizer. The average nitrogen rate was 110 kg ha^{-1}, with some farmers applying up to tripled this average amount. There were also some observations without any nitrogen application. Most zero nitrogen rates were attributed to farmers not reporting the application or applying nitrogen from other sources such as manure. This led to the decision to remove observations with zero nitrogen rates. The weather variables include the overall values of 300 days distributed in 30 features each, therefore extending beyond the growing season and not representing the conditions in which plants were growing.

Most of the observations came from recent years (2016-2018), and about two-thirds were from production systems using hybrids (Table 3.2). Conventional tillage was the most common, followed by no-till. A total of 250 unique cultivars were recorded, with many of them in low frequency. After removing cultivars with a frequency lower than 50, there were ten hybrids and five landraces remaining in the dataset, with a total of 3761 observations (Figure 3.2). The number of observations for each cultivar greatly changes between years. For example, most of the observations for DEKALB 370 were in 2012, and most of the observations for DEKALB 7500 were in 2015. This type of unbalanced distribution is common in observational datasets and needs to be taken into consideration when interpreting the model. The temporal distribution of landraces is more stable over time, apart from 2013 and 2015 that had only hybrids. The yields from the landraces were about half of what was usually obtained with hybrids (Figure 3.3). The lowest yields in the landraces were observed in 2014, while for the hybrids, 2015 was the worse year. There are almost no landraces observations in 2015, so it is not possible to tell whether yields in the landrace system would be even lower in 2015 than they were in 2014.

There is a clear separation in the spatial distribution of system type and maize yield (Figure 3.4). The central west part of the state is characterized by medium-scale medium-input mechanized semi-commercial farmers selling to local markets, using mainly hybrids. The north and east municipalities are characterized by small-scale low-input self-consumption farmers using mainly landraces (Camacho-Villa et al., 2016). The yields follow the same pattern, with most observations in the range of 1.6 to 3.0 Mg ha^{-1} for the landrace system and 4.2 to 5.3 Mg ha^{-1} in the hybrids.

Table 3.1. Descriptive statistics of the numerical variables in the dataset of maize crop events using seven years of field observations from Chiapas – Mexico.

Variable	Mean	SD*	P0	P25	P50	P75	P100	Hist
Yield (Mg ha^{-1})	3.6	1.9	0.1	1.9	3.6	5.0	9.8	▆▆▇▂▁
Elevation (m)	884	443	7	592	712	1079	2849	▃▇▃▁▁
Slope (%)	6.0	6.8	0	1.3	3.0	8.7	61.2	▇▂▁▁▁
Clay (%)	30	10	5	23	28	37	57	▁▆▇▃▂
CEC (cmolc dm^{-3})	22.7	7.4	4.3	16.2	21.1	27.6	50.8	▁▇▅▂▁
SOM (%)	1.6	0.9	0.3	1.0	1.2	2.0	4.0	▇▇▂▂▂
PH	6.8	0.7	4.9	6.6	6.8	7.3	8.3	▁▂▇▃▃
Planting (DOY)	165	22	91	153	167	180	242	▁▃▇▃▁
Nitrogen (kg ha^{-1})	109	64	0	64	110	156	349	▆▇▆▁▁
Phosphorus (kg ha^{-1})	23	26	0	0	23	46	143	▇▃▁▁▁
Potassium (kg ha^{-1})	9	17	0	0	0	12	100	▇▂▁▁▁
Precipitation (mm day^{-1})	4.05	4.86	0	0	2.1	6.8	54.9	▇▁▁▁▁
Solar Rad. (MJ m^{-2} day^{-1})	17.7	2.8	6.8	15.6	17.4	19.5	26.8	▁▂▇▃▁
Max Temp. (C^{o})	29.8	2.5	16.6	28.4	30.2	31.5	39.2	▁▁▇▇▁
Min Temp. (C^{o})	17.4	3.1	-2.1	15.5	18.0	19.7	36.2	▁▁▇▁▁
Vapor Press. (Pa)	1761	628	252	1336	1900	2240	6452	▃▇▁▁▁

*SD: standard deviation; P0 – P100: data distribution percentiles; DOY: day of the year.

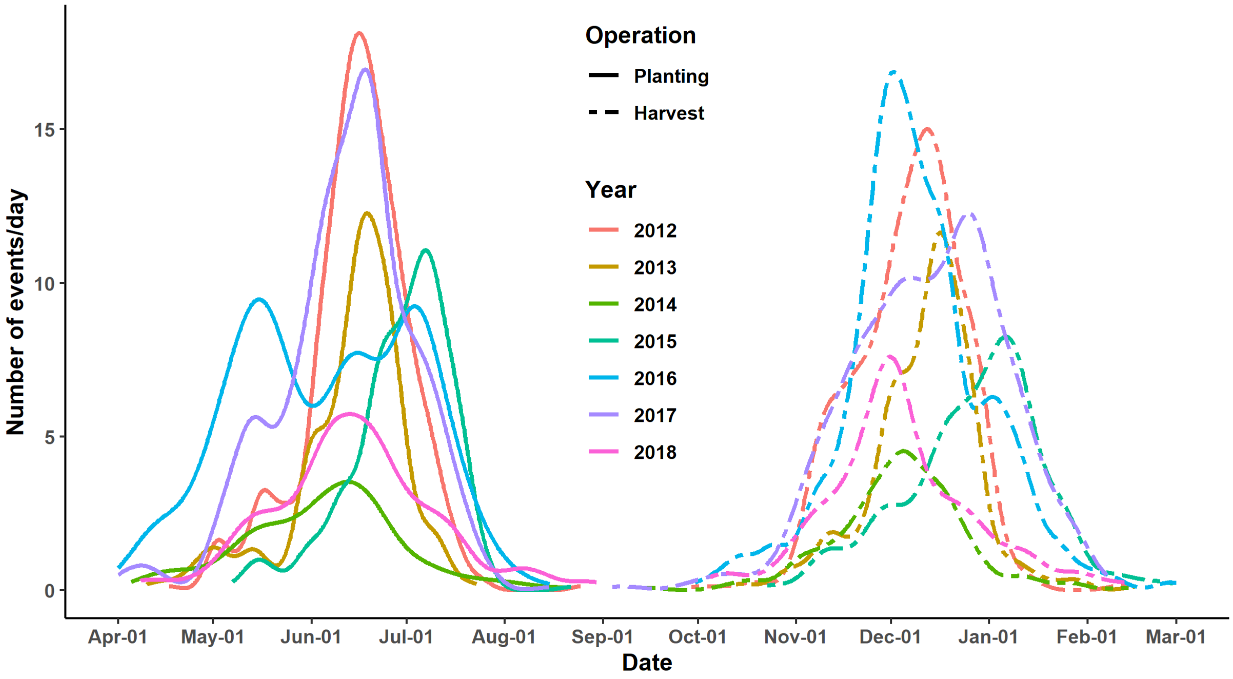

Figure 3.1. Temporal distribution of planting and harvesting maize crop events in seven years of field observations from Chiapas – Mexico.

Table 3.2. Descriptive statistics of the categorical variables in the dataset of maize crop events using seven years of field observations from Chiapas – Mexico.

Variable	Number of levels	Most frequent
Year	7	2017: 960, 2016: 816, 2018: 807, 2012: 723
System	2	Hybrid: 3034, Landrace: 1551
Cultivar	250	CRIOLLO: 629, P4082W: 420
Tillage	3	Convention: 2092, No-till: 1634, Reduced: 859

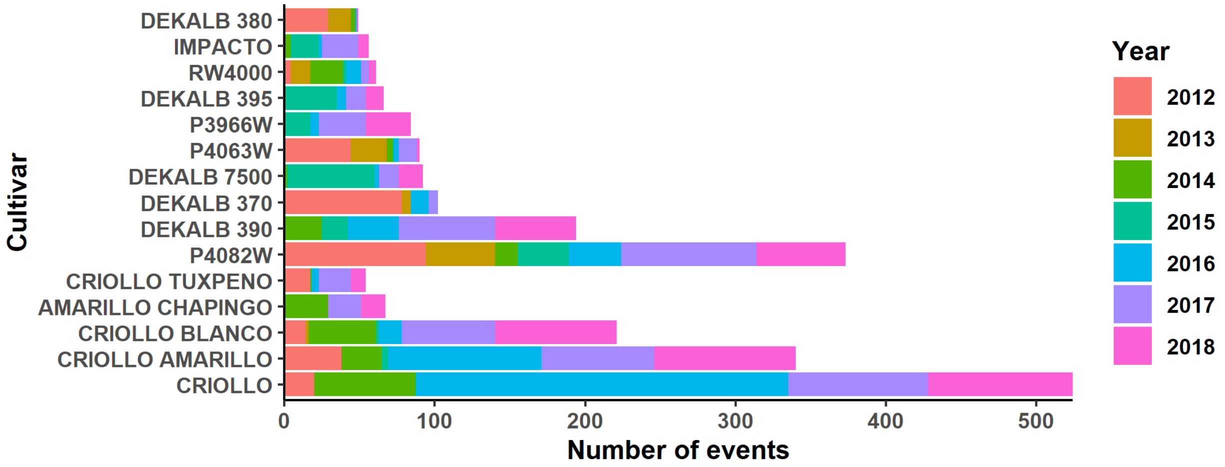

Figure 3.2. Temporal distribution of most commonly observed maize cultivars in seven years of field observations from Chiapas – Mexico.

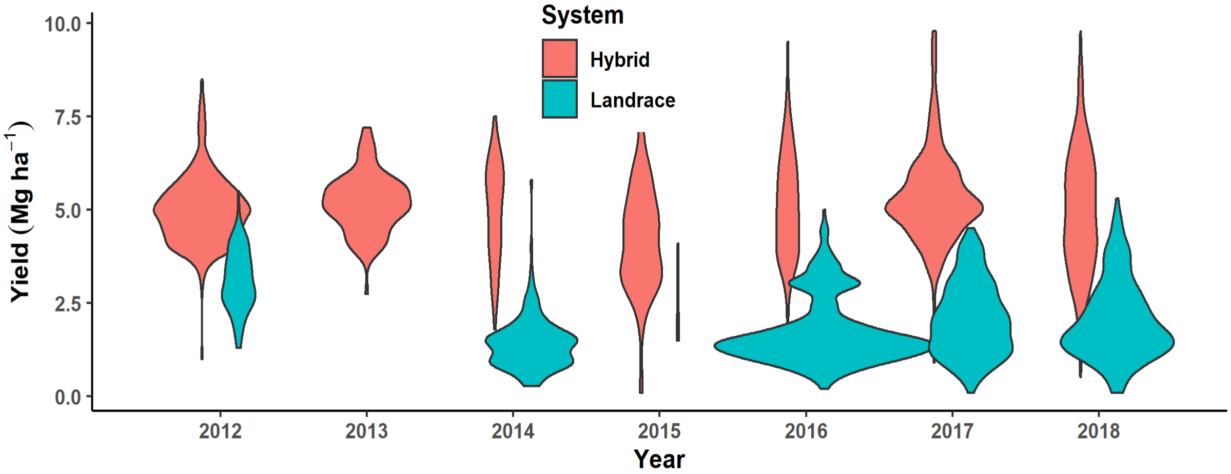

Figure 3.3. Temporal distribution of maize yield in seven years of field observations from Chiapas – Mexico.

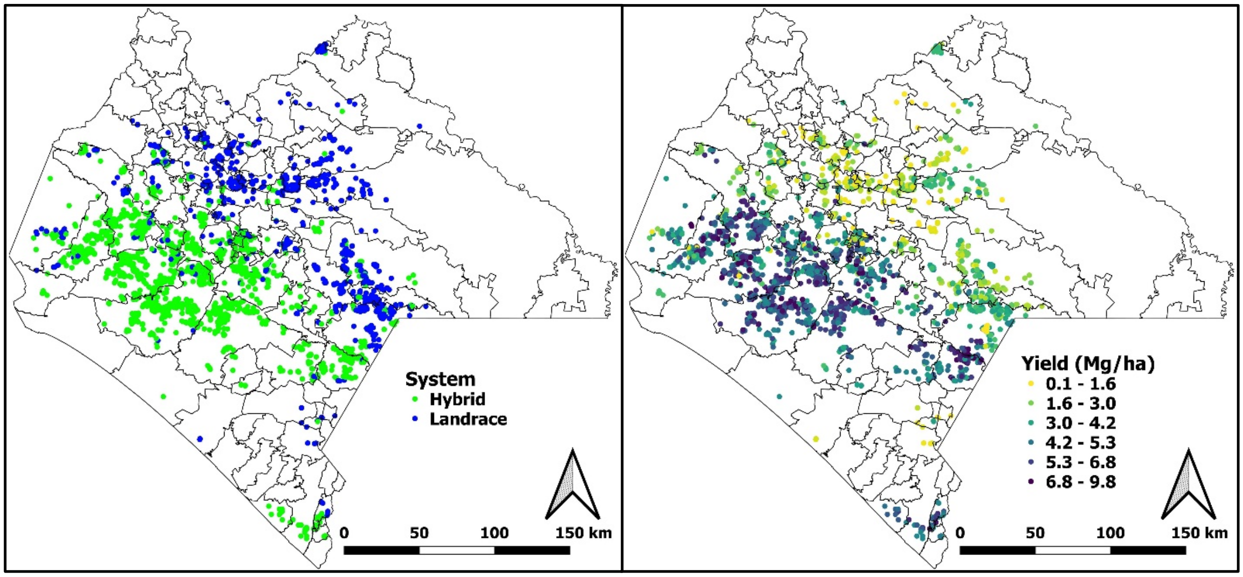

Figure 3.4. Spatial distribution of system type and maze yield using seven years of field observations from Chiapas – Mexico.

Model performance

After optimizing the hyperparameters, the final random forest models were trained with 250 trees and a minimum of 5 observations per node. The cross-validation in the first row of figures (Figure 3.5a-c) was performed with validation sets chosen randomly, while the figures at the bottom (Figure 3.5d-e) used the leave-one-year-out cross-validation. The overall root mean squared error (RMSE) in the random cross-validation was 0.92 Mg ha^{-1}, increasing to 1.16 Mg ha^{-1} using the cross-validation by year. At the same time, the r-squared decreased from 0.75 to 0.60. The performance differences were similar, whether only static features were used or including either year or weather variables as dynamic features. The first row's model performance is useful for describing the importance of variables (descriptive analytics).

In contrast, the second is a better representation of the model performance in future years (predictive analytics). It is important to understand the purpose of the model to decide on the best way to validate it. In this case, the geographic area of interest was defined as roughly the same area where the data was collected. There is also a clear spatial separation of observations from different systems (Figure 3.4), so that enforcing spatial clusters to split the data would not make sense (Valavi et al., 2019).

Although the model's overall predictive performance still explained 60% of the variation, it is also essential to know the performance when isolating the most important non-controllable factors. In the years highlighted, 2016 is an example of a year in which the model had a good performance in both cross-validation scenarios, while 2015 is an example of poor results in the leave-one-year-out cross-validation. The observed yields were usually lower than the model predicted. The differences could be caused by some feature that wasn't part of the model, such as a high insect or disease incidence, or even due to some combination of feature values not previously observed, such as a drought in a critical time for crop development.

Adding weather variables increased model accuracy slightly in the random cross-validation, although including year as a factor increased it even more (Figure 3.6). However, when the cross-validation was performed using years as a grouping variable, the inclusion of weather variables did not improve the models. The predictions were low for years 2012 in Landrace systems and 2015 in Hybrid systems. The negative values of r-square show that model predictions were worse than using just the average of past observations. The relative trends in performance were maintained for the two types of cross-validation, with a significant decrease when the cross-validation was performed grouping by years. This shows that models could be helpful when years are similar to previous years included in the model, while their usefulness in years with extreme weather or other significant differences may be limited. Therefore, most of the remaining results in this paper focus on these models' explanatory power to understand the factors influencing yields.

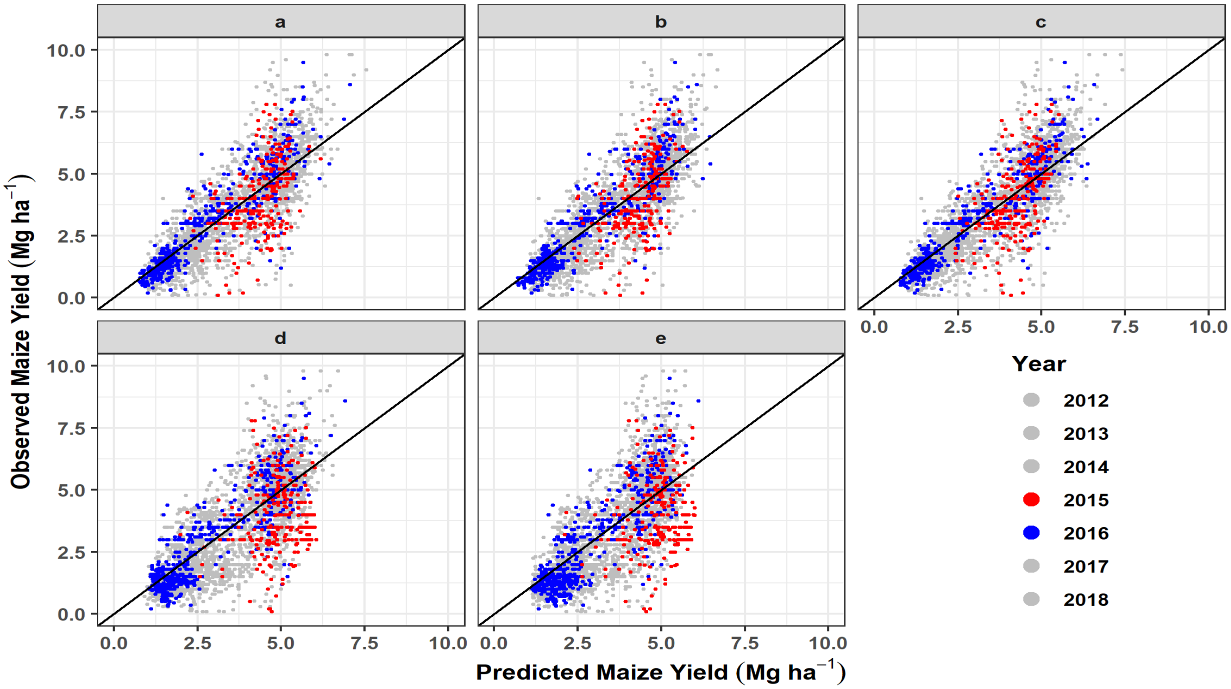

Figure 3.5. Scatterplots of predicted and observed maize yield in five random-forest models with different input features (a,d: baseline; b,e: addition of weather features; c: addition of year and planting date) and cross-validation methods (a,b,c: random split; d,e: grouped by year) using seven years of field observations from Chiapas – Mexico.

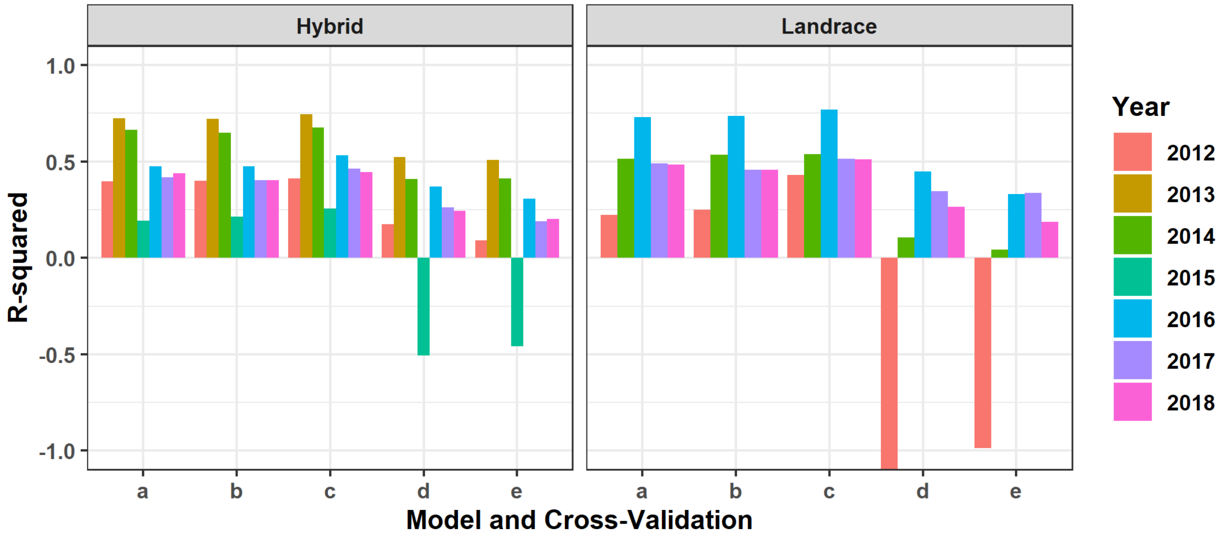

Figure 3.6. Model performance represented by the r-squared values for the maize yield predictions by five random-forest models with different input features (a,d: baseline; b,e: addition of weather features; c: addition of year and planting date) and cross-validation methods (a,b,c: random split; d,e: grouped by year) using seven years of field observations from Chiapas – Mexico.

Feature importance

The interpretation of feature importance is similar to that of classical analysis of variance (ANOVA). The variables' overall effects are decomposed into main effects and interactions, which in turn is used to decide how to decompose the interactions and which posthoc tests are appropriate. In ML terms, the ranking of variable importance and interaction strength helps select how to further perform model exploration. Many of the ANOVA simplifications depend on the assumptions of normally independent and identically distributed residuals, which are most likely broken in the observational data without any trial design being used. In this case, the assumptions are overcome by cross-validation procedures, which works to maintain the model's validity in observations never seen by the model. Therefore, the interpretation of the model relies also upon domain knowledge to filter which hypotheses are meaningful from those that are impaired by the lack of independence.

The feature importance is measured by the decrease in model performance when one of the features' values are randomly permuted. A larger reduction means the feature is more important for model predictions. The most critical feature in the model trained with the full dataset was the system (Figure 3.7). This is consistent with what has been observed in the distributions of yield values (Figure 3.3 and 3.4). Nitrogen was the second most important variable, followed by the temporal variability represented by the year. The interaction strength measures how much of the variability can not be explained by the additive main effect, which indicates to what extent a feature interacts in the model with all the other features (Friedman and Popescu, 2008). The order of variables was similar when looking at the overall interaction strength (Figure 3.8). More than 20% of the variation explained by system and year comes from their interactions with other variables. On the other hand, it also means that almost 80% of the contribution is the main effect. Since the system had the largest contribution and the most important interactions, separate analyses were conducted with each subset.

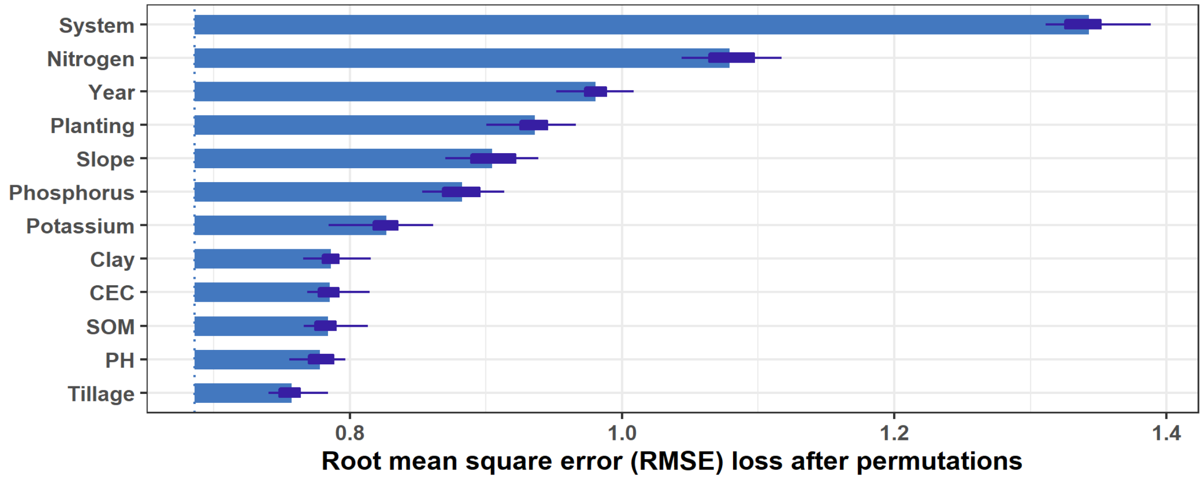

Figure 3.7. Overall feature importance per variable in the random forest model used to predict field observations of maize yield in Chiapas – Mexico, from 2012 to 2018.

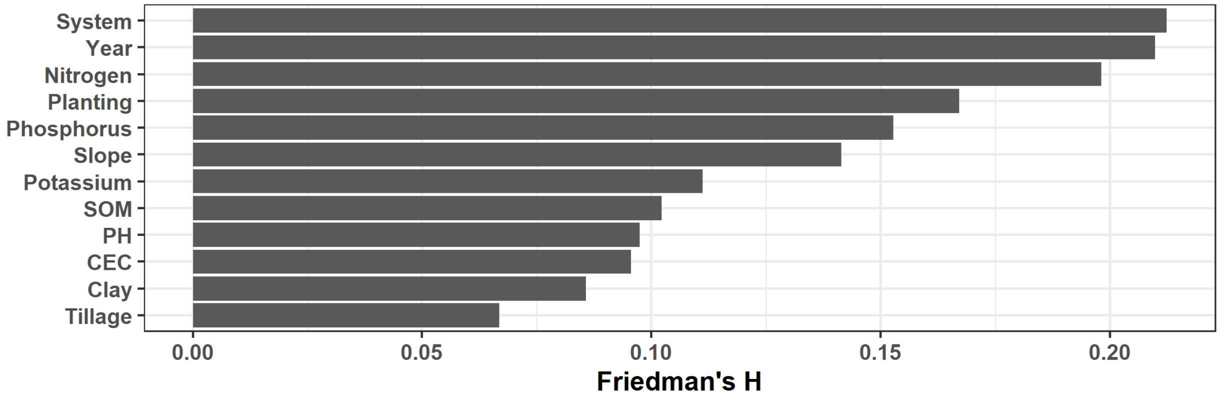

Figure 3.8. Overall interaction strength per variable in the random forest model used to predict field observations of maize yield in Chiapas – Mexico, from 2012 to 2018.

In the high-input (hybrid) system, the temporal variability represented by year was the most important feature, followed by nitrogen and variety (Figure 3.9). The order of variables was similar when looking at the overall interaction strength (Figure 3.10). Almost 40% of the variation explained by year and more than 30% for nitrogen comes from their interactions with other variables. As the number of features increases, higher-level interactions are possible. However, they tend to be lower in magnitude and hard to explain. This was the reason the full dataset was split into two groups, and then the two-way interactions were decomposed. The most important pairwise interaction was phosphorus and year (Figure 3.11). This turned up to be of little practical importance because the data was inflated with zeros, and the interaction was related to the few high rates observed in some years. The following interactions were further investigated: variety with year, variety with nitrogen, and nitrogen with year.

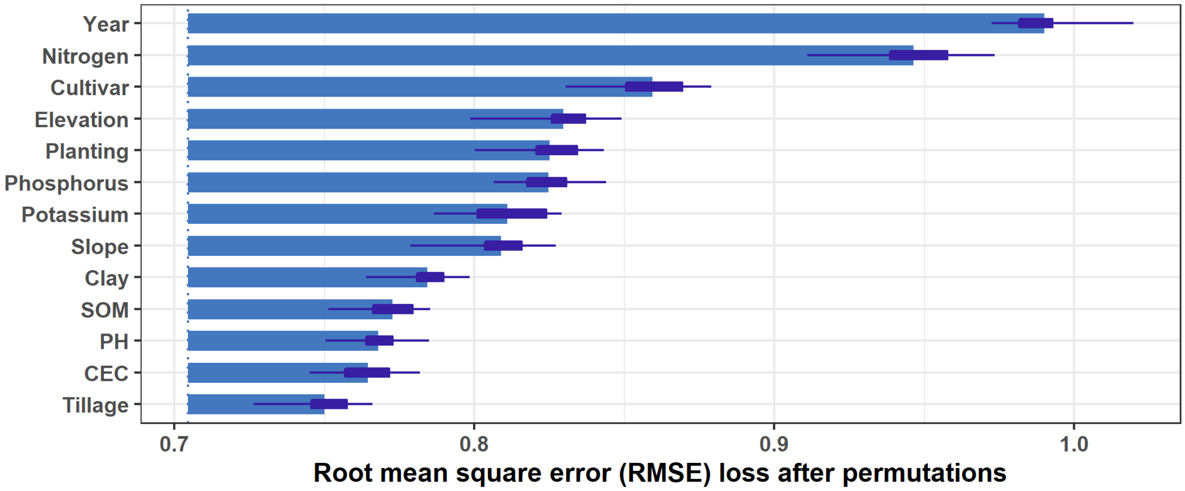

Figure 3.9. Overall feature importance per variable in the random forest model used to predict maize yield in the high-inputs system in Chiapas – Mexico, from 2012 to 2018.

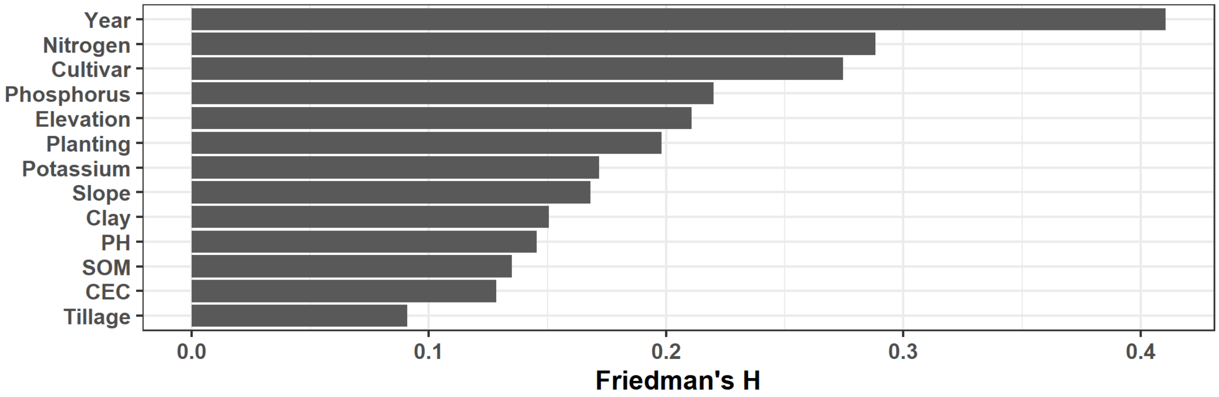

Figure 3.10. Overall interaction strength per variable in the random forest model used to predict maize yield in the high-inputs system in Chiapas – Mexico, from 2012 to 2018.

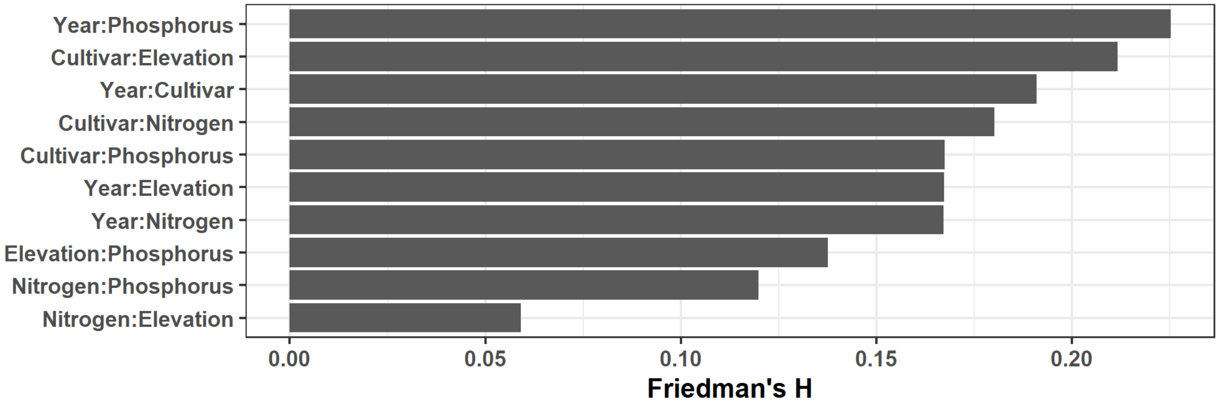

Figure 3.11. Pairwise interaction strength per variable combination in the random forest model used to predict maize yield in the high-inputs system in Chiapas – Mexico, from 2012 to 2018.

Slope and elevation were among the most important variables in the low-input system (Figure 3.12), which account for more than only the topographical characteristics but are instead a proxy for other variables not included in the model. These are related to logistics, access to extension services and inputs, the feasibility of mechanization, and soil degradation. Even though these are environmental variables that cannot be changed, some of the underlying factors that they may be correlated with can change over time. Nitrogen, planting date, and year completed the list of the most important variables with similar values. As in the high-input system, the interactions again followed almost the same order as the main effects (Figure 3.13), with nitrogen and planting date as the most important interactions from the management group of variables. Their interactions were mainly with temporal variability (Figure 3.14). Therefore, the interactions that were further explored were planting date with year and nitrogen with year.

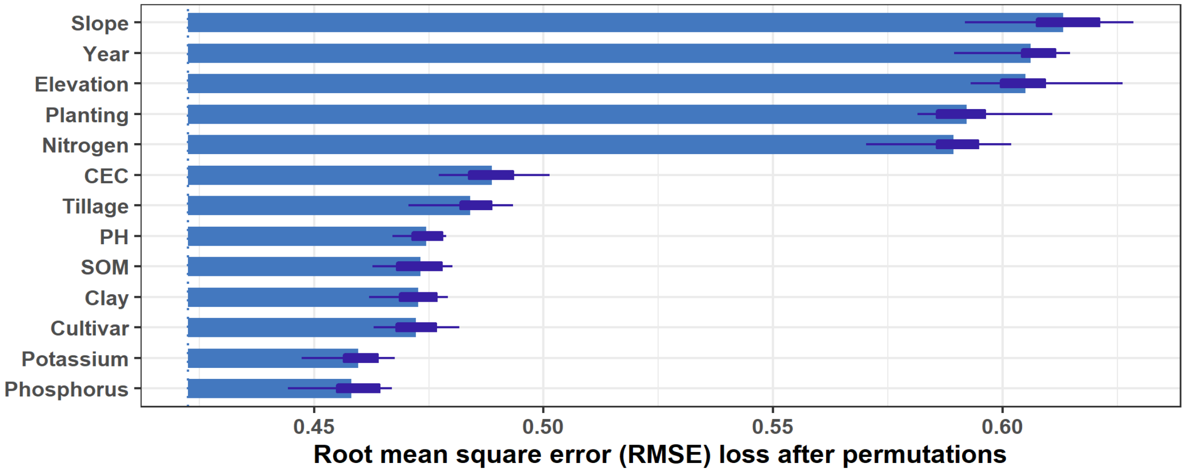

Figure 3.12. Overall feature importance per variable in the random forest model used to predict maize yield in the low-inputs system in Chiapas – Mexico, from 2012 to 2018.

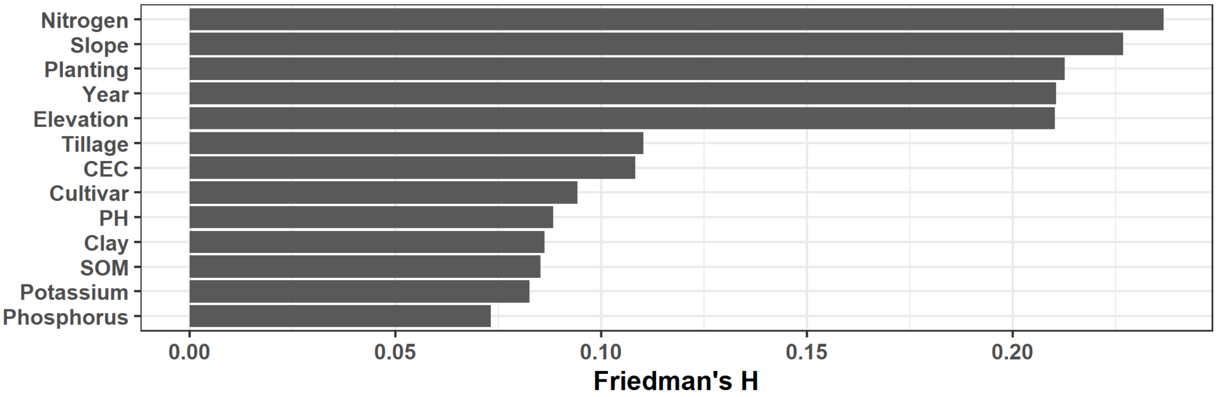

Figure 3.13. Overall interaction strength per variable in the random forest model used to predict maize yield in the low-inputs system in Chiapas – Mexico, from 2012 to 2018.

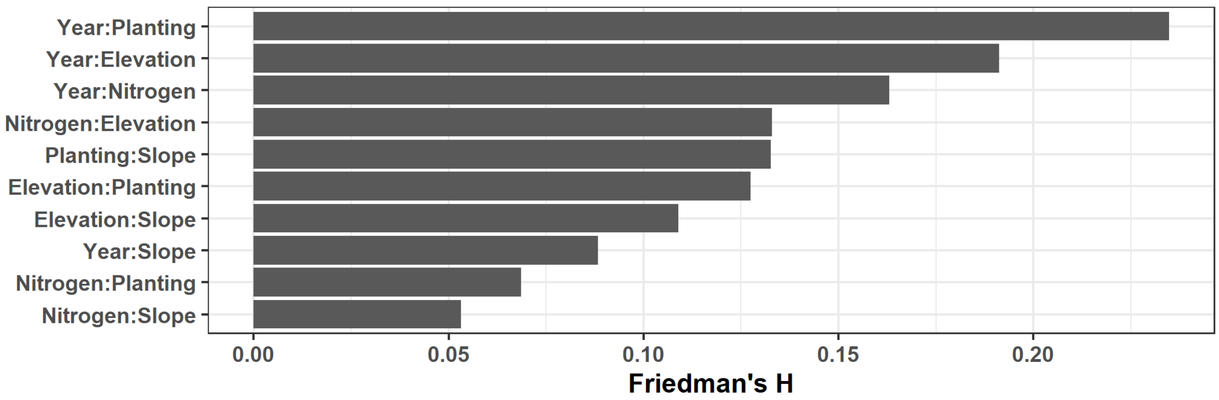

Figure 3.14. Pairwise interaction strength per variable combination in the random forest model used to predict maize yield in the low-inputs system in Chiapas – Mexico, from 2012 to 2018.

Yield response

Partial dependences were used to calculate yield response to changes in each variable. The approach is analogous to the best linear unbiased prediction analysis (BLUP) used in classical statistical models (mixed models). Both methods are designed to calculate expected values when controlling for all other effects included in the model (Montesinos-López et al., 2019). However, only a subset of all factors that may affect yields was effectively being controlled. Grouped partial dependence plots were used to evaluate the most important interactions from a model and management perspective. The yield responses were generated by fixing all other factors and running predictions for each observation's possible nitrogen rates (Figure 3.15). Then the predicted yield response curves were averaged for each variety. This can be used to characterize how each variety responds to increased nitrogen rates, which, after including costs, would allow the estimation of the economically optimum nitrogen rates. The cultivars can be characterized according to their yields with limited or sufficient levels of nitrogen (Mastrodomenico et al., 2018b). For example, the hybrid RW4000 shows the highest yields up to 200 kg ha^{-1} of nitrogen fertilizer, showing small yield gains with increased nitrogen rates. This hybrid has a high nitrogen use efficiency and would probably be the best choice overall because it can deliver high yields with fewer inputs. Another good example is the hybrid DEKALB 370, which also has low responsiveness to nitrogen, and the economically optimum nitrogen rates would be lower. However, the yield potential of this hybrid is limited, which limits its use. Finally, DEKALB 395 and P4082W are the most responsive hybrids, requiring higher nitrogen rates to express their full potential since yields increased by more than 20% when N rates go from 100 to 200 kg ha^{-1}.

There are clear differences in yield responses to N among years (Figure 3.16), with higher responses in 2015 and 2017 and lower responses in 2012 and 2014. Interestingly, 2015 and 2017 were opposite years in terms of yield potential (main effect), but the optimum nitrogen rates were high in both years, confirming that yield potential is not a good predictor of economically optimum nitrogen rates (Rodriguez et al., 2019). It is important to remember that these are overall observations, which are somehow affected by each year's specific observations. Similar to what can be observed for the cultivars, the yield response in the range of 50 to 100 kg ha^{-1} is almost flat. This may be explained by the correlation between practices, where farmers apply low rates of synthetic fertilizer when they expect a low response to it, for example, when a legume was the previous crop was a legume, using other nitrogen sources, or when they have soils with more organic matter. Similarly, it is also possible that those using more than 100 kg ha^{-1} are also more likely to adopt other practices that increase yields, such as applying fungicides and better weed control, and some of the yield increase is due to correlated practices rather than nitrogen alone.

Most of the interaction seen in the hybrids with years was due to the low yields observed in 2015 (Figure 3.17). The hybrid DEKALB 7500 moved from being an average to the lowest-yielding variety, while DEKALB 370 was more stable in this bad year, moving from the lowest yield to about the average. Looking at yield stability and nitrogen response, DEKALB 370 can be considered a typical workhorse hybrid (Mastrodomenico et al., 2018a). Since stability is one of the MasAgro program's objectives, recommending cultivars that do not perform the best but are more stable may be desirable. This form of bet-hedging or risk-averse portfolio optimization could also be essential to guarantee genetic diversity if some disease appears.

In the low-input system, the year's main effect appears much larger than any interactions (Figure 3.18). For example, the yield in 2012 was double that in 2014, no matter how much nitrogen was used. The year 2013 was the only one in which a clear interaction appears, with no response in the range tested. This result is not reliable because there were no landrace observations this year. It would be advisable to remove factor levels when the number of observations is lower than a certain threshold. The decision to keep all levels here was to show that the model interpretation needs to consider the distribution of observations used. This applies to the range tested as well. That is the reason to keep the marks along the x-axis to show for which value observations were available. Combining the marks with each year's colors is also possible but becomes uninformative with a static graph because many symbols would overlap. Using interactive plots provides a richer experience to interpret the models and is advisable whenever possible.

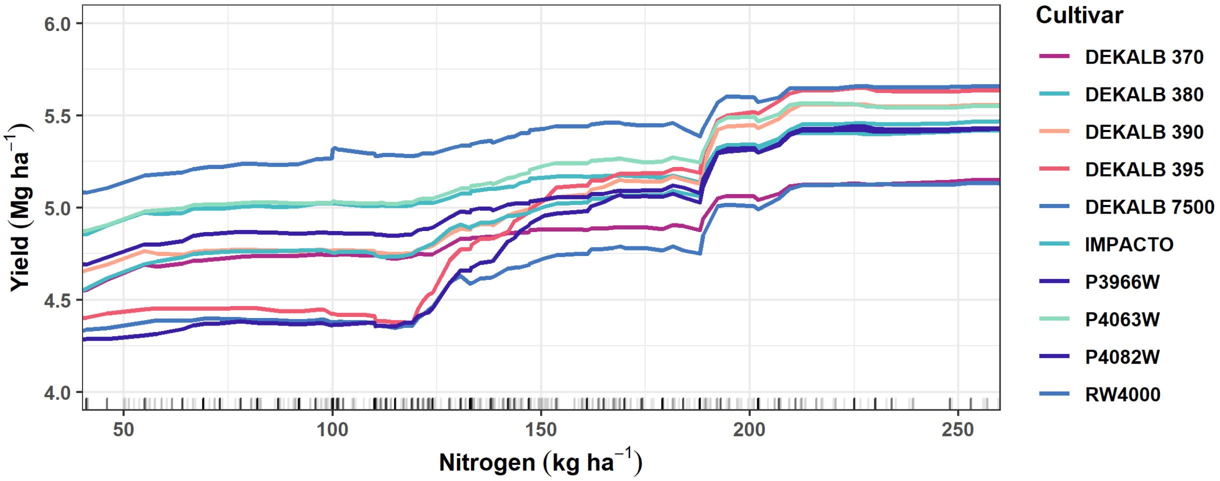

Figure 3.15. Partial dependence plots for the yield response to nitrogen rates in different maize hybrids cultivated in Chiapas – Mexico, from 2012 to 2018. The marks on the x-axis represent the distribution of nitrogen rates tested.

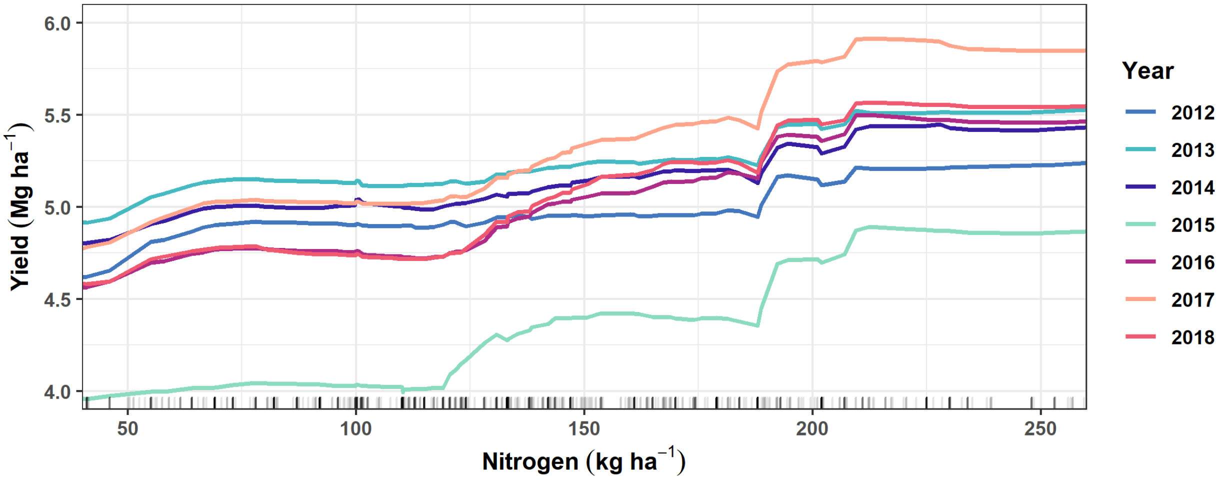

Figure 3.16. Partial dependence plots for maize yield response to nitrogen rates in different years in Chiapas – Mexico. The marks on the x-axis represent the distribution of nitrogen rates tested.

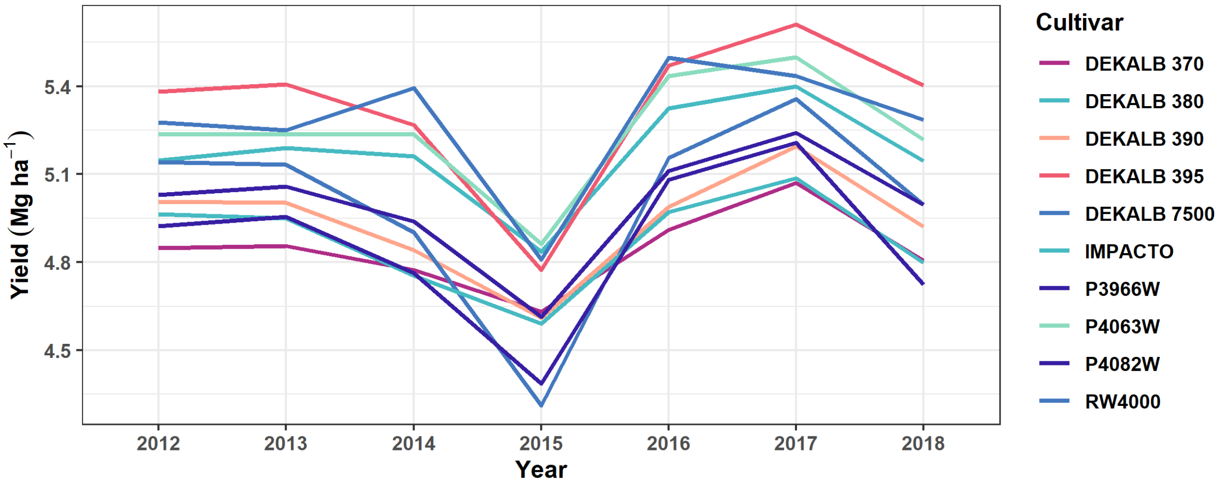

Figure 3.17. Partial dependence plots for maize yield response of hybrids in seven years in Chiapas – Mexico.

The planting date and year interactions were more about how yield responded to planting delays than which planting date maximized yields (Figure 3.19). Although plating after July 1st provided consistently higher yields, delaying planting may have some tradeoffs. Usually, early harvests achieve higher marketing prices. The pressure to have the food ready when people need it may also favor earlier planting dates. On the other hand, later planting requires less storage time, which could reduce storage losses. Since the variations were small, the use of more than one planting date may be advisable for reducing risk by diversification and accounting for needs other than grain yield.

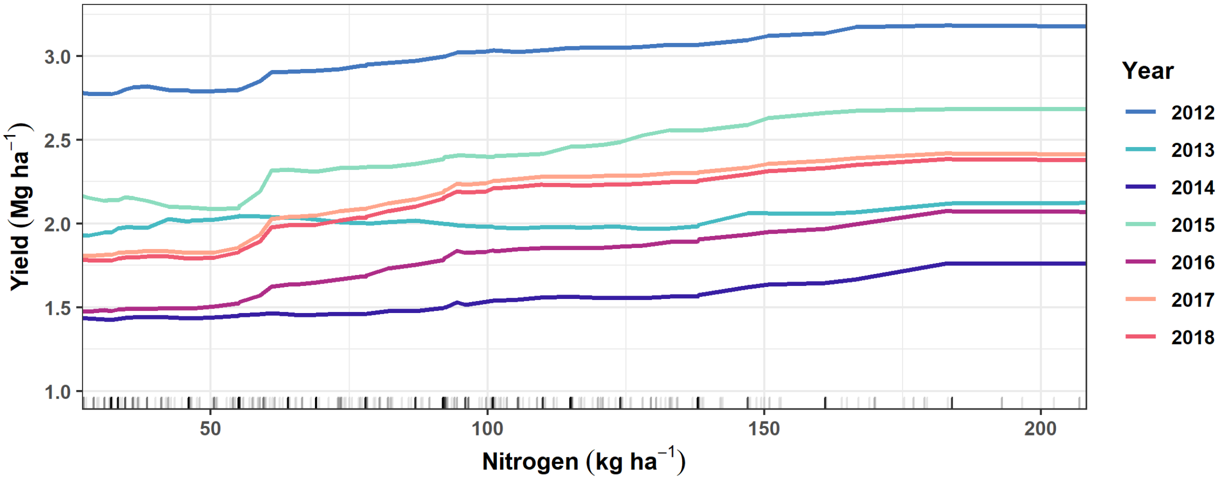

Figure 3.18. Partial dependence plots for maize yield response to nitrogen rates over seven years in the low-inputs system in Chiapas – Mexico. The marks on the x-axis represent the distribution of nitrogen rates tested.

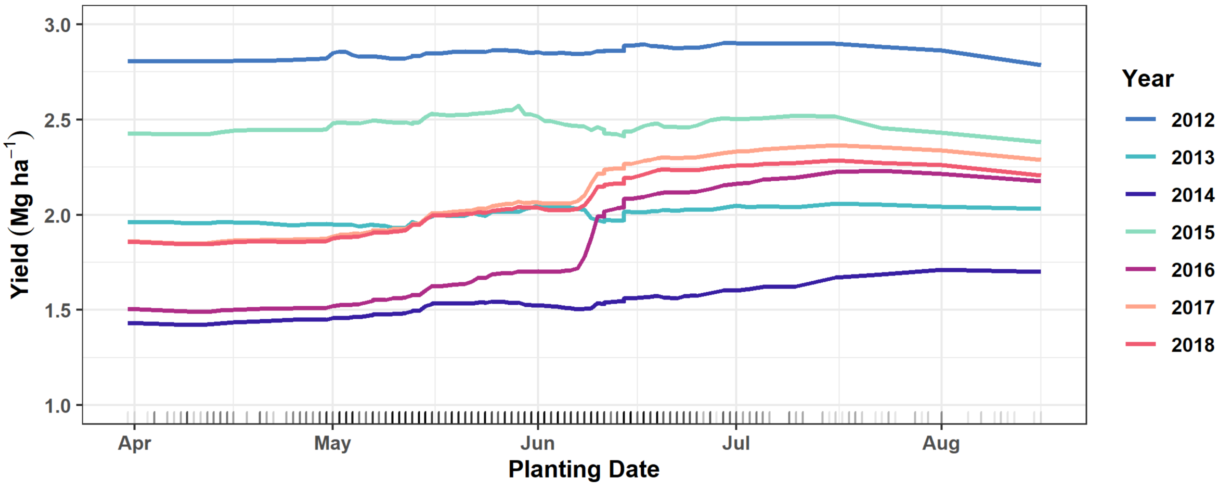

Figure 3.19. Partial dependence plots for maize yield response to planting dates over seven years in the low-inputs system in Chiapas – Mexico. The marks on the x-axis represent the distribution of nitrogen rates tested.

Feature contribution

This section focus on explanations for single crop events. The breakdown of feature contributions explains how each combination of variable and value for a given observation contributed to the model prediction for that same observation. The values represent the change in the prediction when comparing the conditional probability, sequentially fixing the value of one variable at a time, with the original prediction. Since the models are not additive, due to the interactions, the order in which the variables are fixed will affect the result. To solve this, the process is repeated, permuting the order of variables. In both examples (Figure 3.20 and Figure 3.21), in the graph on the left side, the contributions are based on permutations of the order of variables. The variables are then ordered by their absolute importance. The graph on the right side represents one of the possible explanations for that same observation. In this case, the interpretation of the order is different, with each level down representing the contribution of the variable-value pair given that all other variables on top were already fixed.

In the high-input system example (Figure 3.20), the year was the most critical feature. Fixing the nitrogen rate at 188 increased the prediction by about 120 kg ha^{-1}, compared to the average prediction when using the actual rates in each observation. The bar represents the main effect, while the boxplot represents the variation from the interaction with changes in other variables. In the low-input system example (Figure 3.21), the three most important features are environmental variables that cannot be controlled. In the third position, the nitrogen rate was the most important management variable, with a positive contribution. Tillage type, planting date, and potassium rate reduced the yield.

The interpretation has to be done carefully because the contribution depends on the importance of the feature and also on the value itself. The farther from the average, the more critical it will be. This also means that variables with small contributions may still be optimized. This tool is also valuable for interactive visualization, allowing the user to fix environment variables' values and test what would happen by changing management variables. This can also be aided by partial dependence curves for each variable, specific for this observation rather than aggregated. This would also allow the visualization of an optimized version of management and to compare with what the model suggested would be the best decisions.

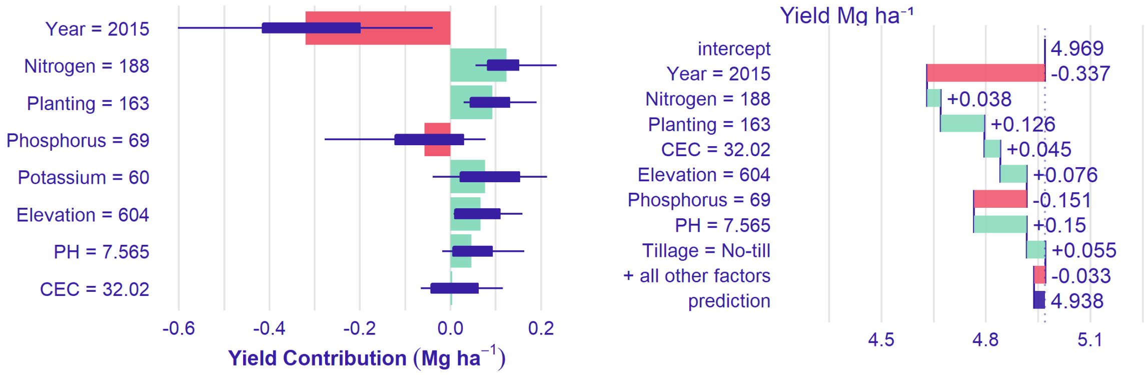

Figure 3.20. Feature contribution in the maize yield prediction of one randomly selected observation using the random forest model for the high-input system in Chiapas – Mexico. The left side reports the uncertainty in the contributions based on permutations in the order of variables, while the right side represents one of the possible explanations for that same observation.

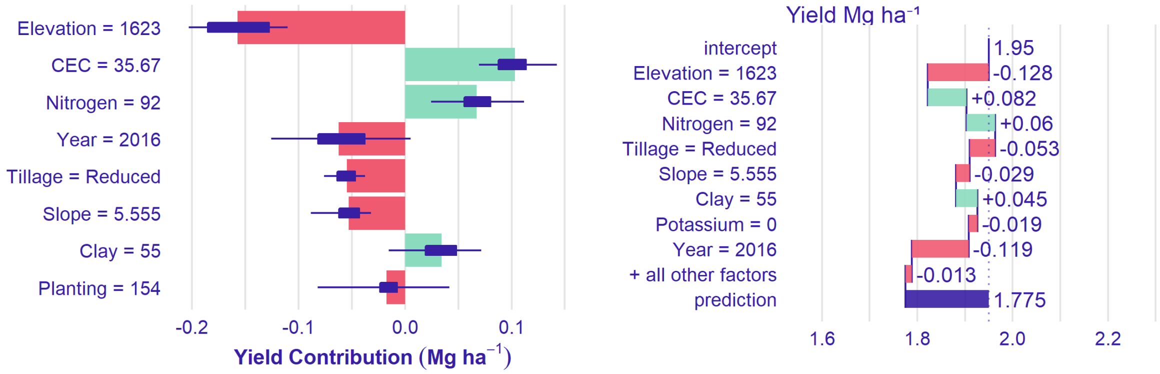

Figure 3.21. Feature contribution in the maize yield prediction of one randomly selected observation using the random forest model for the low-inputs system in Chiapas – Mexico. The left side reports the uncertainty in the contributions based on permutations in the order of variables, while the right side represents one of the possible explanations for that same observation.

DISCUSSION

Model performance

Although there are more than four thousand observations, there are inherent correlations between different groups of data points. This causes many observations to be correlated, and the correct number of genuinely independent observations would be much smaller. Even though different planting dates and regions contribute to more independent observations from a weather perspective, there is considerably more similarity among observations from the same year, thus making predictions into future years challenging. Temporal correlation is probably the most important factor explaining the reduction in model performance in some years (Figure 3.6). From a management perspective, the interaction between weather and management decisions is more important since the weather effect itself is a non-controllable factor (Jiménez et al., 2016). However, from a statistical perspective, the analysis of interactions requires a greater number of observations than what would be necessary to achieve the same statistical power looking only at the main effects (Rothery et al., 2003).

The number of independent observations from on-farm research trials has also been a limiting factor in other studies (Eldon et al., 2020). The authors also pointed out that the number of observations available in their study was sufficient to test existing general management recommendations and even resulted in new ones, but they were not enough to develop site-specific recommendations. Researches have demonstrated that yield prediction errors using ML models decreased by 10%–40% as the training dataset increased from 0.5 to 1.8 million data points, thus showing the importance of having more observations (Shahhosseini et al., 2019).

Another aspect of weather data related to the number of observations is its high dimensionality. Including daily data as individual features should provide the maximum explanatory power in the model. However, the total number of features would be greater than the number of observations in the dataset, thus diluting each feature's effect and leading to unstable model training. On the other side, using only the growing season average would not represent the growing conditions in critical stages of plant development such as flowering. The aggregation of weather variables into ten-day intervals was an attempt to use a reasonable number of features. Other dimensionality reduction or feature engineering techniques may provide better results.

Temporal convolutional networks have achieved state-of-the-art status in many time-series problems (Hewage et al., 2020; Wu et al., 2020). Although temporal convolutional networks could be used to learn these features automatically in an end-to-end fashion, a much larger number of observations would be needed.

Integration of ML methods with mechanistic crop simulation models is probably one of the best alternatives to reduce the dimensionality of the data without losing important information (Langensiepen et al., 2020). Crop simulation models are skillful in using weather variables, but it is difficult to properly calibrate all parameters and to represent management differences within the models (Leng and Hall, 2020). Different ways have been proposed to integrate crop simulation models with ML into hybrid models that are superior to both. One way to integrate both approaches is to use crop simulation models results as additional features that incorporate the weather variability more efficiently in the model. This strategy has been shown to improve wheat yield projections in Australia, using a combination of the process-based APSIM model and a Random Forest ML model (Feng et al., 2020). Because there are still interactions with other environmental variables and management decisions, instead of using only the yield potential, a handful of different scenarios could be simulated and included as additional features.

An alternative to this scenario is to use the crop simulation models to generate the ground truth data and then train ML models based on this. This approach has been used to provide reliable estimates of model performance (Shahhosseini et al., 2019; Yamamoto, 2019; Saikai et al., 2020). One less explored possibility derived from this would be to apply transfer-learning techniques using ML models pre-trained with synthetic data. This could allow most of the model parameters to be trained in synthetic data generated with the crop simulation models and only a small part of the parameters to be trained with field data, thus allowing more complex models to be trained with a small number of observations (Kim et al., 2019). There are clearly many opportunities for the integration of data-driven and science-based methodologies (Messina et al., 2020).

Dataset quality

One of the challenges with this dataset was the number of different cultivars. Random Forest models cannot handle categorical variables with many factor levels efficiently because each factor level is represented internally as an additional feature. Keeping low-frequency factor levels in the training dataset increases computational time. Other ML methods, such as neural networks, can handle this type of data more efficiently using categorical embeddings (Khaki et al., 2020). However, future predictions would still be limited to the same cultivars used in training. Ideally, the conversion from a single categorical variable to a series of numerical features should be performed before training the model in a way that could be applied to new cultivars. A practical method to achieve this would be to replace the variety with features describing its main agronomic characteristics, such as cycle length, fertility requirements, and disease tolerances.

The uncertainties of the information available also accentuate the challenge of modeling the genetic effect. The name attributed to some cultivars may not represent a single genetic material, as some farmers may store the grain to use as seeds and attribute the same name, although it is no longer the original hybrid. This happens even more intensely with the landraces. The same landrace commonly denominated Criollo is likely to encompass many different genetics depending on the region.

Part of the effect of slope and elevation and the lack of clear interactions between landrace cultivars and elevation is attributed to the poor discrimination of different landrace populations in the present dataset. Early researches conducted in the same region found that environmental differences are the primary factor determining the overall pattern of maize genetic variability, but even social origin had a significant effect on maize populations in all environments (Brush and Perales, 2007). A study evaluating 21 populations of maize landrace populations from three altitudinal ranges in the same region concluded that cultivars did not perform well planted in different elevations (Mercer et al., 2008). The performance was especially poor when highland landraces (>2000 m) were cultivated in midland sites (1200–2000 m), which is a concerning signal of poor adaptation to climate change (Mercer and Perales, 2010). Within a limited latitude and time range, the elevation is also strongly correlated with temperature, and other weather variables, which could also explain why including weather variables in the model did not improve model performance since most of the weather effect could be already explained by the topographical variables. Increased temperatures are expected to be the main factor reducing rainfed maize yields under future scenarios (Ureta et al., 2020). Therefore, relying on elevation to make predictions is risky because it does not account for the temporal variability in temperatures, which were probably less important in the seven years spanning the training data than the spatial variability.

Although there were some issues with the variety and fertilizer data from the field observations, the overall quality of these variables was superior to the data obtained from other sources. This becomes clear when comparing the general low importance of soil and weather variables in the models (Figure 3.7) with the usually large effects in other works (Khaki and Wang, 2019; Ureta et al., 2020). The importance of high-resolution characterization of the environment has been discussed mostly in the context of crop breeding (Resende et al., 2020). Researchers recognize the need to increase the resolution to the level of specific experimental plots and individual plants, which requires the development of low-cost, high-throughput envirotyping platforms (Xu, 2016). These needs extend beyond lineage selections, reaching the last step of breeding programs to recommend the most suitable cultivars for each environment. The development and use of these platforms would greatly benefit this type of data-driven approach.

With increased data quality, environmental variables and interactions would be expected to become more important in the model, which would allow more accurate recommendations. The main limitation with gridded weather datasets is usually precipitation, which in turn is one of the most critical factors in rainfed agriculture (Feng et al., 2020; Ureta et al., 2020). Better radar and satellite data may provide improved data in the near future. However, the spatial variability of rain is so intense, occurring in short distances, that low-cost devices to measure precipitation at the field level could still be crucial.

Improved envirotyping data quality includes location-specific data to represent each field and more variables to represent chemical, physical, and biological soil quality. An evaluation of 236 soil samples taken in Vilaflores, a municipality in Chiapas with 25000 ha of maize production, revealed coefficients of variation greater than 100% for most soil attributes. The majority of samples were low in potassium and organic matter, and 40% of the sites were below the critical limit of 15 ppm of phosphorus (López Báez et al., 2019). Recent research in the same area has identified soil compaction as one of the main constraints to maize production. The decrease of soil porosity as a result of the compaction was correlated to yields losses, which in years with drought events in the crop's critical period reduced maize yields up to 58% (López Báez et al., 2018). Evaluating only the amount of fertilizer applied without information about the soil nutrient contents can be misleading because there is an inverse correlation between the amount of fertilizer used and soil fertility. However, the yields may still be lower in poor soil with more fertilizer than in more fertile soils with less fertilizer, effectively creating a negative correlation between applied fertilizer and yield. Phosphorus and potassium's low importance in the models reflects the lack of knowledge about these nutrients' soil levels and their low variability within the dataset. Acquiring the right and accurate information remains a significant challenge to the development of decision support systems and the adoption of prediction methods (Messina et al., 2020).

Delivering results

The most interesting results at the dataset level were the interactions between genotypes and N rates. The differences in nitrogen use efficiency indicate an excellent opportunity to make better choices within the currently available hybrids and develop new cultivars that are more productive and require fewer inputs (Mastrodomenico et al., 2018a). It is important to note that these are aggregated results, and the best hybrid on average may not be the same as the more frequent best hybrid for each location since there are higher-level interactions that are not considered. A comparison of genotype performance from on-farm trials and on-station trials found similar results in precision of a single plot, but with significant interaction effects between genotype and trial system (Schmidt et al., 2018). The interactions are likely caused by uncontrolled differences in management between the research stations and farmer's fields. A similar effect is likely to occur with the results of this research, which advises for careful considerations when translating model results into field recommendations since many essential factors were left out of the model. These factors could be tolerance or resistance to insects and diseases, for example.

These aspects are crucial in the context of on-farm experimentation. Some authors have argued that management history and biological factors are rarely described, although these are important descriptors of the research population and therefore needed to establish the populations for which studies seek to generalize their findings (Kool et al., 2020). Others have argued that it is useful to look at the actual crop yield effects at the whole system rather than control how the crop is managed because those differences represent the real world (Coe et al., 2019). Based on this work's results, it is important to be explicit with how the data was collected, how the model was trained, and the intended use of the results. Communicating the strengths and weaknesses of the methods employed helps set the right level of expectations and build trust over time, which is necessary to allow for new iterations of model improvements. It is essential to accept the limitations in using the model as a prescriptive system and reiterate how the resulting recommendations can be used as suggestions for further testing. One of the goals of a decision support system is to encourage farmers to make personalized decisions based on a set of adaptive options and on factors that are not well captured by agronomic research (Eldon et al., 2020).

The decisions to decompose the system interaction were motivated both by its interaction strength in the model and the known differences between the two systems that are not easily captured with the variables used. Governmental agriculture programs have played an essential role in fostering the adoption of high-yielding hybrids, while cultural preferences have encouraged landrace retention (Bellon and Hellin, 2011; Hoogendoorn et al., 2018). The farmers using landraces are unlikely to switch to hybrids just because they are more productive. Even if there were no cultural barriers, input-intensive practices are unlikely to be cost-effective for many subsistence farmers (Eldon et al., 2020). Rather than suggesting optimal management practices, this research exemplifies the use of explainable ML to offer farmers the opportunity of benchmarking their management decisions with peers in similar growing conditions (Figure 3.20 and 3.21) and visualize expected outcomes if different decisions were made. The next step to ensure that farmers have access to the information would be to deploy interactive versions of these tools in smartphone applications such as AgroTutor (Bayas et al., 2020).

CONCLUSIONS

Using the random forest algorithm, machine learning models explained up to 75% of the variation of maize yield in various environments and cropping systems scenarios in Chiapas. The variance explained for years not seen during training dropped to 60%. The model performance further decreased when evaluated in a specific system and year combinations. The ability to use the model to predict crop performance in future weather scenarios is still limited. The main challenges faced during model development were the high dimensionality of weather variables and the unbalanced spatial and temporal distribution of the cultivars. The total number of independent observations limits the use of more complex models. There are opportunities to integrate machine learning and crop simulation models to solve these challenges.

Domain knowledge and explainable machine learning methods allowed the use of the model as a source of information to create and validate hypotheses. Feature importance and interaction strength were used to identify the most critical variables. Partial dependence plots were used to determine the trends in the main interactions. Nitrogen was the management decision that influenced yields the most, with different yield responses depending on the year and variety. The differences in nitrogen use efficiency indicate an excellent opportunity to make better choices within the currently available hybrids and develop new cultivars that are more productive and require fewer inputs. Breakdown plots were used to reveal contributions and uncertainties at the individual observation level. This can allow each farmer to answer why they obtained a particular result and what would have happened if they made different decisions.

REFERENCES

Barbosa, A., R. Trevisan, N. Hovakimyan, and N.F. Martin. 2020. Modeling yield response to crop management using convolutional neural networks. Comput. Electron. Agric. 170(February): 105197. doi: 10.1016/j.compag.2019.105197.

Bayas, J.C.L., A. Gardeazabal, M. Karner, C. Folberth, L. Vargas, et al. 2020. Agrotutor: A mobile phone application supporting sustainable agricultural intensification. Sustain. 12(22): 1–10. doi: 10.3390/su12229309.

Bellon, M.R., and J. Hellin. 2011. Planting Hybrids, Keeping Landraces: Agricultural Modernization and Tradition Among Small-Scale Maize Farmers in Chiapas, Mexico. World Dev. 39(8): 1434–1443. doi: 10.1016/j.worlddev.2010.12.010.

Biecek, P. 2018. DALEX: Explainers for Complex Predictive Models in R. J. Mach. Learn. Res. 19(84): 1–5. https://jmlr.org/papers/v19/18-416.html.

Brush, S.B., and H.R. Perales. 2007. A maize landscape: Ethnicity and agro-biodiversity in Chiapas Mexico. Agric. Ecosyst. Environ. 121(3): 211–221. doi: 10.1016/j.agee.2006.12.018.

Bullock, D.S., M. Boerngen, H. Tao, B. Maxwell, J.D. Luck, et al. 2019. The Data-Intensive Farm Management Project: Changing Agronomic Research Through On-Farm Precision Experimentation. Agron. J. 111(6): 2736. doi: 10.2134/agronj2019.03.0165.

Camacho-Villa, T.C., C. Almekinders, J. Hellin, T.E. Martinez-Cruz, R. Rendon-Medel, et al. 2016. The evolution of the MasAgro hubs: responsiveness and serendipity as drivers of agricultural innovation in a dynamic and heterogeneous context. J. Agric. Educ. Ext. 22(5): 455–470. doi: 10.1080/1389224X.2016.1227091.

Campos-Taberner, M., F.J. García-Haro, B. Martínez, E. Izquierdo-Verdiguier, C. Atzberger, et al. 2020. Understanding deep learning in land use classification based on Sentinel-2 time series. Sci. Rep. 10(1): 1–12. doi: 10.1038/s41598-020-74215-5.

Chambers, R., and J. Jiggins. 1987. Agricultural research for resource-poor farmers Part I: Transfer-of-technology and farming systems research. Agric. Adm. Ext. 27(1): 35–52.

Coe, R.I.C., J. Njoloma, and F. Sinclair. 2019. TO CONTROL or NOT to CONTROL: HOW DO WE LEARN MORE about HOW AGRONOMIC INNOVATIONS PERFORM on FARMS? Exp. Agric. 55(S1): 303–309. doi: 10.1017/S0014479717000102.

Cui, Z., H. Zhang, X. Chen, C. Zhang, W. Ma, et al. 2018. Pursuing sustainable productivity with millions of smallholder farmers. Nature 555(7696): 363–366. doi: 10.1038/nature25785.

Delerce, S.J. 2018. Polygon map of soil units with functional properties for Mexico (1:250000), Harvard Dataverse, V1. doi: 10.7910/DVN/QNMIZR.

Delerce, S., H. Dorado, A. Grillon, M.C. Rebolledo, S.D. Prager, et al. 2016. Assessing weather-yield relationships in rice at local scale using data mining approaches. PLoS One 11(8). doi: 10.1371/journal.pone.0161620.

Donnet, M.L., I.D.L. Becerril, J.R. Black, and J. Hellin. 2017. Productivity differences and food security: A metafrontier analysis of rain-fed maize farmers in MasAgro in Mexico. AIMS Agric. Food 2(2): 129–148. doi: 10.3934/agrfood.2017.2.129.

Dorado, H., S. Delerce, D. Jimenez, and C. Cobos. 2018. Finding Optimal Farming Practices to Increase Crop Yield Through Global-Best Harmony Search and Predictive Models, a Data-Driven Approach. Mexican International Conference on Artificial Intelligence. p. 15–29

Eldon, J., G. Baird, S. Sidibeh, D. Dobasin, P. Rapaport, et al. 2020. On-farm trials identify adaptive management options for rainfed agriculture in West Africa. Agric. Syst. 182(May): 102819. doi: 10.1016/j.agsy.2020.102819.

Feng, P., B. Wang, D.L. Liu, C. Waters, D. Xiao, et al. 2020. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. Agric. For. Meteorol. 285–286(February): 107922. doi: 10.1016/j.agrformet.2020.107922.

van Frank, G., I. Goldringer, P. Rivière, and O. David. 2019. Influence of experimental design on decentralized, on-farm evaluation of populations: a simulation study. Euphytica 215(7): 126. doi: 10.1007/s10681-019-2447-9.

Friedman, J.H., and B.E. Popescu. 2008. Predictive learning via rule ensembles. Ann. Appl. Stat. 2(3): 916–954. doi: 10.1214/07-AOAS148.

Goodman, B., and S. Flaxman. 2017. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation." AI Mag. 38(3): 50–57. doi: 10.1609/aimag.v38i3.2741.

Hewage, P., A. Behera, M. Trovati, E. Pereira, M. Ghahremani, et al. 2020. Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station. Soft Comput. 24(21): 16453–16482. doi: 10.1007/s00500-020-04954-0.

Hijmans, R.J. 2020. raster: Geographic Data Analysis and Modeling. https://cran.r-project.org/package=raster.

Hoogendoorn, J.C., G. Audet-Bélanger, C. Böber, M.L. Donnet, K.B. Lweya, et al. 2018. Maize seed systems in different agro-ecosystems; what works and what does not work for smallholder farmers. Food Secur. 10(4): 1089–1103. doi: 10.1007/s12571-018-0825-0.

Humphries, S., J.C. Rosas, M. Gómez, J. Jiménez, F. Sierra, et al. 2015. Synergies at the interface of farmer-scientist partnerships: Agricultural innovation through participatory research and plant breeding in Honduras. Agric. Food Secur. 4(1): 1–17. doi: 10.1186/s40066-015-0046-0.

Inwood, S.E.E., and V.H. Dale. 2019. State of apps targeting management for sustainability of agricultural landscapes. A review. Agron. Sustain. Dev. 39(1): 8. doi: 10.1007/s13593-018-0549-8.

Jarvis, A., H.I. Reuter, A. Nelson, E. Guevara, and others. 2008. Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database.

Jiménez, D., S. Delerce, H. Dorado, J. Cock, L.A. Muñoz, et al. 2019. A scalable scheme to implement data-driven agriculture for small-scale farmers. Glob. Food Sec. 23(May): 256–266. doi: 10.1016/j.gfs.2019.08.004.

Jiménez, D., H. Dorado, J. Cock, S.D. Prager, S. Delerce, et al. 2016. From observation to information: Data-driven understanding of on farm yield variation. PLoS One 11(3): 1–20. doi: 10.1371/journal.pone.0150015.

Jones, J.W., J.M. Antle, B. Basso, K.J. Boote, R.T. Conant, et al. 2017. Brief history of agricultural systems modeling. Agric. Syst. 155: 240–254. doi: https://doi.org/10.1016/j.agsy.2016.05.014.

Khaki, S., Z. Khalilzadeh, and L. Wang. 2020. Predicting Yield Performance of Parents in Plant Breeding: A Neural Collaborative Filtering Approach. http://arxiv.org/abs/2001.09902.

Khaki, S., and L. Wang. 2019. Crop yield prediction using deep neural networks. Front. Plant Sci. 10(May): 1–10. doi: 10.3389/fpls.2019.00621.

Kim, N., K.J. Ha, N.W. Park, J. Cho, S. Hong, et al. 2019. A comparison between major artificial intelligence models for crop yield prediction: Case study of the midwestern United States, 2006–2015. ISPRS Int. J. Geo-Information 8(5). doi: 10.3390/ijgi8050240.

Lang, M., M. Binder, J. Richter, P. Schratz, F. Pfisterer, et al. 2019. mlr3: A modern object-oriented machine learning framework in R. J. Open Source Softw. doi: 10.21105/joss.01903.

Langensiepen, M., M.A.K. Jansen, A. Wingler, B. Demmig-Adams, W.W. Adams, et al. 2020. Linking integrative plant physiology with agronomy to sustain future plant production. Environ. Exp. Bot. 178(May). doi: 10.1016/j.envexpbot.2020.104125.

Lapuschkin, S., S. Wäldchen, A. Binder, G. Montavon, W. Samek, et al. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. Nat. Commun. 10(1): 1–8. doi: 10.1038/s41467-019-08987-4.

Leng, G., and J.W. Hall. 2020. Predicting spatial and temporal variability in crop yields: An inter-comparison of machine learning, regression and process-based models. Environ. Res. Lett. 15(4). doi: 10.1088/1748-9326/ab7b24.

López Báez, W., R. Reynoso Santos, J. López Martínez, R. Camas Gómez, and A. Tasistro. 2018. Diagnóstico de la compactación en suelos cultivados con maíz en la Región Fraylesca, Chiapas. Rev. Mex. Ciencias Agrícolas 9(1): 65–79. doi: 10.29312/remexca.v9i1.848.

López Báez, W., R. Reynoso Santos, J. López Martínez, B. Villar Sánchez, R. Camas Gómez, et al. 2019. Caracterización físico-química de suelos cultivados con maíz en Villaflores, Chiapas. Rev. Mex. Ciencias Agrícolas 10(4): 897–910. doi: 10.29312/remexca.v10i4.1764.

Lorentzen, C., and M. Mayer. 2020. Peeking into the Black Box: An Actuarial Case Study for Interpretable Machine Learning. SSRN Electron. J. doi: 10.2139/ssrn.3595944.

Mastrodomenico, A.T., J.W. Haegele, J.R. Seebauer, and F.E. Below. 2018a. Yield Stability Differs in Commercial Maize Hybrids in Response to Changes in Plant Density, Nitrogen Fertility, and Environment. Crop Sci. 58(1): 230–241. doi: 10.2135/cropsci2017.06.0340.

Mastrodomenico, A., C. Hendrix, and F. Below. 2018b. Nitrogen Use Efficiency and the Genetic Variation of Maize Expired Plant Variety Protection Germplasm. Agriculture 8(1): 3. doi: 10.3390/agriculture8010003.

Mercer, K., Á. Martínez-Vásquez, and H.R. Perales. 2008. Asymmetrical local adaptation of maize landraces along an altitudinal gradient. Evol. Appl. 1(3): 489–500. doi: 10.1111/j.1752-4571.2008.00038.x.

Mercer, K.L., and H.R. Perales. 2010. Evolutionary response of landraces to climate change in centers of crop diversity. Evol. Appl. 3(5–6): 480–493. doi: 10.1111/j.1752-4571.2010.00137.x.

Messina, C.D., M. Cooper, M. Reynolds, and G.L. Hammer. 2020. Crop science: A foundation for advancing predictive agriculture. Crop Sci. 60(2): 544–546. doi: 10.1002/csc2.20116.

Molina-Maturano, J., N. Verhulst, J. Tur-cardona, D.T. Güerena, and A. Gardeazábal-. 2020. Understanding smallholder farmers ' intention to adopt agricul- tural apps : the role of mastery-approach and innovation hubs. (December). doi: 10.20944/preprints202012.0396.v1.

Montesinos-López, O.A., A. Montesinos-López, R. Tuberosa, M. Maccaferri, G. Sciara, et al. 2019. Multi-Trait, Multi-Environment Genomic Prediction of Durum Wheat With Genomic Best Linear Unbiased Predictor and Deep Learning Methods. Front. Plant Sci. 10(November): 1–12. doi: 10.3389/fpls.2019.01311.

Ortiz-Crespo, B., J. Steinke, C.F. Quirós, J. van de Gevel, H. Daudi, et al. 2020. User-centred design of a digital advisory service: enhancing public agricultural extension for sustainable intensification in Tanzania. Int. J. Agric. Sustain. 0(0): 1–17. doi: 10.1080/14735903.2020.1720474.

Pebesma, E. 2018. Simple Features for R: Standardized Support for Spatial Vector Data. R J. 10(1): 439–446. doi: 10.32614/RJ-2018-009.

Qin, Z., D.B. Myers, C.J. Ransom, N.R. Kitchen, S.Z. Liang, et al. 2018. Application of machine learning methodologies for predicting corn economic optimal nitrogen rate. Agron. J. 110(6): 2596–2607. doi: 10.2134/agronj2018.03.0222.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. https://www.r-project.org/.

Resende, R.T., H.P. Piepho, G.J.M. Rosa, O.B. Silva-Junior, F.F. e Silva, et al. 2020. Enviromics in breeding: applications and perspectives on envirotypic-assisted selection. Theor. Appl. Genet. (0123456789). doi: 10.1007/s00122-020-03684-z.

Riccetto, S., A.S. Davis, K. Guan, and C.M. Pittelkow. 2020. Integrated assessment of crop production and resource use efficiency indicators for the U.S. Corn Belt. Glob. Food Sec. 24(November 2019): 100339. doi: 10.1016/j.gfs.2019.100339.

Rodriguez, D.G.P., D.S. Bullock, and M.A. Boerngen. 2019. The Origins, Implications, and Consequences of Yield-Based Nitrogen Fertilizer Management. Agron. J. 111(2): 725–735. doi: 10.2134/agronj2018.07.0479.

Roscher, R., B. Bohn, M.F. Duarte, and J. Garcke. 2020. Explainable Machine Learning for Scientific Insights and Discoveries. IEEE Access 8: 42200–42216. doi: 10.1109/ACCESS.2020.2976199.

Rothery, P., S.J. Clark, and J.N. Perry. 2003. Design of the farm-scale evaluations of genetically modified herbicide-tolerant crops. Environmetrics 14(7): 711–717. doi: 10.1002/env.619.

Saikai, Y., V. Patel, and P.D. Mitchell. 2020. Machine learning for optimizing complex site-specific management. Comput. Electron. Agric. 174(March): 105381. doi: 10.1016/j.compag.2020.105381.

Schmidt, P., J. Möhring, R.J. Koch, and H.P. Piepho. 2018. More, larger, simpler: How comparable are on-farm and on-station trials for cultivar evaluation? Crop Sci. 58(4): 1508–1518. doi: 10.2135/cropsci2017.09.0555.

Shahhosseini, M., R.A. Martinez-Feria, G. Hu, and S. V Archontoulis. 2019. Maize yield and nitrate loss prediction with machine learning algorithms. Environ. Res. Lett. 14(12): 124026. doi: 10.1088/1748-9326/ab5268.

Snapp, S.S., J. Dedecker, and A.S. Davis. 2019. Farmer participatory research advances sustainable agriculture: Lessons from Michigan and Malawi. Agron. J. 111(6): 2681–2691. doi: 10.2134/agronj2018.12.0769.

de Sousa, K., F. Casanoves, J. Sellare, A. Ospina, J.G. Suchini, et al. 2018. How climate awareness influences farmers' adaptation decisions in Central America? J. Rural Stud. 64(September): 11–19. doi: 10.1016/j.jrurstud.2018.09.018.

Steinke, J., J. van Etten, A. Müller, B. Ortiz-Crespo, J. van de Gevel, et al. 2020. Tapping the full potential of the digital revolution for agricultural extension: an emerging innovation agenda. Int. J. Agric. Sustain. 0(0): 1–17. doi: 10.1080/14735903.2020.1738754.

Thornton, P.E., M.M. Thornton, B.W. Mayer, Y. Wei, R. Devarakonda, et al. 2016. Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 3. doi: 10.3334/ORNLDAAC/1328.

Trevisan, R.G., D.S. Bullock, and N.F. Martin. 2020. Spatial variability of crop responses to agronomic inputs in on-farm precision experimentation. Precis. Agric. (0123456789). doi: 10.1007/s11119-020-09720-8.

Ureta, C., E.J. González, A. Espinosa, A. Trueba, A. Piñeyro-Nelson, et al. 2020. Maize yield in Mexico under climate change. Agric. Syst. 177(December 2018): 102697. doi: 10.1016/j.agsy.2019.102697.

Valavi, R., J. Elith, J.J. Lahoz-Monfort, and G. Guillera-Arroita. 2019. blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods Ecol. Evol. 10(2): 225–232. doi: 10.1111/2041-210X.13107.

Vanlauwe, B., R.I.C. Coe, and K.E. Giller. 2019. BEYOND AVERAGES: NEW APPROACHES to UNDERSTAND HETEROGENEITY and RISK of TECHNOLOGY SUCCESS or FAILURE in SMALLHOLDER FARMING. Exp. Agric. 55(S1): 84–106. doi: 10.1017/S0014479716000193.

Westermann, O., W. Förch, P. Thornton, J. Körner, L. Cramer, et al. 2018. Scaling up agricultural interventions: Case studies of climate-smart agriculture. Agric. Syst. 165(July): 283–293. doi: 10.1016/j.agsy.2018.07.007.

Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wickham, H., R. François, L. Henry, and K. Müller. 2020. dplyr: A Grammar of Data Manipulation. https://cran.r-project.org/package=dplyr.

Wolanin, A., G. Mateo-Garciá, G. Camps-Valls, L. Gómez-Chova, M. Meroni, et al. 2020. Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt. Environ. Res. Lett. 15(2). doi: 10.1088/1748-9326/ab68ac.

Wright, M.N., and A. Ziegler. 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J. Stat. Softw. 77(1): 1–17. doi: 10.18637/jss.v077.i01.

Wu, P., J. Sun, X. Chang, W. Zhang, R. Arcucci, et al. 2020. Data-driven reduced order model with temporal convolutional neural network. Comput. Methods Appl. Mech. Eng. 360: 112766. doi: 10.1016/j.cma.2019.112766.

Xu, Y. 2016. Envirotyping for deciphering environmental impacts on crop plants. Theor. Appl. Genet. 129(4): 653–673. doi: 10.1007/s00122-016-2691-5.

Yamamoto, K. 2019. Distillation of crop models to learn plant physiology theories using machine learning. PLoS One 14(5).

CHAPTER 4

CONCLUSIONS AND DISCUSSION

This work's main contribution is to offer flexible alternatives to analyze datasets and answer questions that challenge most of the traditional statistical methods in different contexts. This flexibility can be seen in using the same model with varying inputs, relying on the spatial dependence without using covariables, and analyzing observational data that lacks independence. The discussion of the results focused on aspects that are overlooked in most publications. The emphasis on model performance in different years and locations and understanding the main factors contributing to decreased performance (Chapter 1 and 3) illustrates the importance of explainable machine learning.

Different uncertainty estimation methods (Chapter 1) were used to evaluate the overall quality of the data at the trial level and select individual plots in which the error is likely to be higher, which should be targeted for new ground truth data acquisition in order to improve the model. The ability to understand when the predictions fail is one of the foundations for model improvement. Although the methods were tested within a plant breeding context, there are many possibilities to use the same tools to collect higher resolution data and improve on-farm precision experimentation (Chapter 2).

The combination of on-farm precision experimentation and geographically weighted regression proved to be an effective methodology to test precision agriculture central hypothesis' of whether there is significant within-field variability in optimal rates (Chapter 2). The results also allowed decomposing the expected benefits from improved management of temporal variability (optimizing the average rate) and spatial variability (variable-rate application). Yield is the most common dependent variable used to evaluate the effect of agronomic treatments. However, the yield is the result of many interactions between production factors, and most of those cannot be directly controlled by decision-makers. The literature is full of examples of research that implicitly or explicitly assume that yield variability is directly associated with variations in optimum management decisions. The methodology presented in Chapter 2 avoids these assumptions and gives some evidence that they are inappropriate. Yield response, that is, how the yield change when an alternative management decision is tested, should be the focus of the research.

This also raises the concern that evaluating model performance by how well the yield variation was explained is not aligned with the objective of using the results to optimize management decisions. In the models used in Chapter 2, the spatial variability of the intercept is more important to overall model performance than all higher-level parameters combined. However, improving the estimates of the intercept does not help to improve recommendations. The same applies to the observational data used in Chapter 3. Improved model performance from additional non-controllable production factors will only be meaningful to decision-makers if they change yield response to management decisions.

Prescriptive analytics is considered the ultimate level of data analysis, the hardest to achieve, but also the most valuable. All the limitations identified in this research showed that there is still a long way to achieve this level in agriculture, especially with more complex production systems. In the meantime, working with farmers, understanding their needs, and offering them the opportunity of benchmarking their management decisions with peers in similar growing conditions and visualize expected outcomes if different decisions were made is the most promising strategy.