Crop yield prediction with ensemble algorithms and Artificial Neural Networks (ANN)

Dissertation to apply for the degree:
Master's in Management & Analytics
BUSINESS SCHOOL

Tobias Ruiz Moreno
28/06/2019

Master's Committee: Dr. Merener, Martín

ABSTRACT

ABSTRACT

World cereal production is set to grow by around 1% per year for the next decade, and while crop areas are not expanding, the major driver for the growth production is expected to come from yield improvements. Crop yields have been commonly modelled in two ways: process-based modelling (also known as crop simulation) and statistical modelling. Recently, machine learning started to deliver interesting results, mainly because it has the advantage of dealing with non-linear relationships between factors. Weather plays an important role in defining crop yields. Being able to simulate accurate weather conditions and predict crop yield has been an important topic in the industry. The objective of this work is to model crop yields using Random Forest regressor and Long Short-Term Memory (LSTM) Neural Networks (NN) in 9 annual crops in Argentina: wheat, barley, maize, soybean, sunflower, sorghum, rice, cotton and peanut. Soil and weather data was collected and transformed for 80 counties in Argentina. Hyperparameters for the 2 models were optimized and accuracy metrics were compared. Weather information was simulated estimating the distribution of the historical information using KDE (Kernel Density Estimator) and Monte Carlo to generate random sampling. Feature importance analysis allowed to reduce the number of factors up to 7 without compromising model accuracy. From the 9 crops studied, soybean, maize, sunflower, sorghum, wheat and barley models returned reasonable accuracy metrics. Except for the last two (wheat and barley) which are winter crops, the remaining 4 summer crops (soybean, maize, sorghum and sunflower) were forecasted simulating rainfall in different stages of the growing season and returned estimations with an error below 20% (MAPE) before harvest. Random forest outperformed classic MLR statistical model by more than 30% on average over all the crops, but overfitting was significantly high. LSTM did not perform as well as Random forest: although LSTM did not overfit, performance was slightly better than baseline with large variations between crops. This work demonstrates that machine learning algorithms are a competitive alternative to statistical modelling for crop yield prediction, and weather simulations can return reasonably accurate predictions before harvest. This allows the agricultural community to anticipate strategic decisions based on crop production forecasts.

INDEX

INDEX

TABLES

FIGURES

INTRODUCTION

INTRODUCTION

AGRICULTURE

During the past decade agricultural products demand increased remarkably. Two main events are responsible for driving the trend: China's economic growth caused the consumption of feed to grow by almost 6% per year, and on the other hand the use of feedstock inputs to produce biofuel grew by almost 8% per year. The replenishment of cereal stocks by 230 Mt also augmented demand (OECD/FAO 2017).

Cereals world production reached a historical high in 2016, especially for wheat and maize. Soybean production increased strongly due to record in crops in the United States and Brazil. World production for other oilseeds (rapeseed, sunflower and groundnuts) increased in 2016 for the first time in three years (OECD/FAO 2017).

Argentina and Brazil experienced the strongest expansion in crop areas over the past ten years, adding respectively 10Mha and 8Mha to global crop land (OECD/FAO 2017). However, global agricultural land decreased by 63 Mha between 1960 and 1993, a trend which is expected to continue.

Under this scenario, yield growth will continue to drive global crop production. According to (OECD/FAO 2017), cereal production is set to grow by around 1% per year for the next decade, being yield improvements the bulk of production increase.

Maize is expected to grow 14% in 10 years. Latin America will contribute with 28% of the total increase, Asia and Pacific with 24% and North America with 22%, leaving the rest to European Union and Africa (OECD/FAO 2017). Wheat is expected to grow 11%, with an increase in area of only 1.8%. The increase in wheat production is therefore expected to occur through higher yields (OECD/FAO 2017), mostly from Asia and Pacific (46%), European Union (13%) and Russia (9%). Rice yields are expected to grow 12%, and major gains are projected for Asian countries.

Agriculture plays an important role in countries GDP. According to (WorldBank 2017), world agriculture contribution to GDP was 3.5% in 2016, but it varies significantly between countries. While India's agriculture contribution to GDP was 15.6%, United States' was 1%, Brazil's agricultural value added was 4.6% and Argentina's 5.6% (WorldBank 2017).

Although yield growth is expected to satisfy most of the increasing demand for cereals in the upcoming decade, there may be variations depending on weather and climate conditions, such as the "El Niño" phenomenon (OECD/FAO 2017). This opens the discussion about how yield predictions can be achieved using crop modelling techniques with weather information to account to crop production variability.

This work will analyse 9 of the world's principal crops in terms of area planted and total production and will contemplate last 40-50 years of crop yields obtained in different regions of Argentina.

CROP MODELLING

The agricultural system is very complex and deals with a large number of scenarios, which come from a number factors (Aditya Shastry 2017a). Yield variability due to extreme weather events such as drought and high temperatures remains a major concern in the agriculture industry.

Process-based modelling (also known as crop simulation) and statistical modelling are two common approaches for predicting crop yield responses to climate variability (Kim et al. 2016). Process-based crop models are powerful tools for crop yield predictions, particularly at a field level scale, because they simulate physiological processes of crop growth and development in response to environmental conditions and management practices. However, this approach is not scalable for regional and global estimations due to intensive data and calibration requirements. On the other hand, statistical models estimates direct relationships between predictor variables (e.g. climate and soil factors) and crop yield without considering the underlying processes in crop physiology and ecology (Kim et al. 2016). They return simple but reasonable predictions, provided that sufficient and reliable data is used for training the model and predictions are made within the boundaries of training data. Simple and Multi Linear Regressions (MLR) are commonly used statistical models.

Several statistical studies contemplate weather and soil data to model crop yield in a regional or global scale. (Lobell and Burke 2010) analysed maize crop yields and compared the ability of statistical models to predict yield responses to changes in mean temperature and rainfall against process-based model predictions. They concluded that statistical modelling is more suitable for broader spatial scales and global climate projections, and accurate enough to develop robust statistical inferences with weather simulated information. The relationship between wheat yield and climate trends in Mexico were analysed by (Lobell et al. 2005), understanding how much of crop yield increase could be attributed to climate change. Similarly, (Lobell et al. 2011) investigated the non-linear relationship between perennial crop yield and climate change in California. They found that climate change would likely make yields go down for almonds, walnuts, avocados and table grapes by 2050.

Most of the investigation around crop modelling takes similar environmental factors into account, mostly water availability and temperature as the critical weather factors responsible for crop yield. Regarding statistical modelling, both linear and non-linear techniques have been tested for estimating crop yield and results vary depending on the crop and geography.

This work will model annual crop yields in a regional scale using soil characteristics and regional weather historical information. Although statistical modelling was used in the past to accomplish similar objectives in Argentinian agriculture, in this thesis machine learning algorithms will be tested and optimized to predict accurate crop yields based on data from 1970 until 2018.

MACHINE LEARNING ALGORITHMS FOR CROP MODELLING

While still not as popular as crop simulation and statistical modelling, machine learning is a third choice for predicting crop yields in Agriculture. Machine learning presents methods that define rules and patterns in large datasets and can also adapt the predictive model by itself.

As (Paswan and Begum 2013) indicate, crop production is influenced by a great variety of interrelated factors and it is difficult to describe their relationships by conventional methods. As statistical models need to fulfil regression assumptions and multiple co-linearity between independent and dependent variables, it causes classic statistical methods to be inefficient. Statistical modelling methods based on machine learning algorithms are able to overcome of the limitations of traditional regression approaches (Kim et al. 2016).

(Kim et al. 2016) evaluated prediction performance of Random Forest (RF) regressor using MLR as a baseline in 4 crops: wheat (global), maize, silage maize and potato (county level in US). They concluded that RF outperformed MLR in all crops and scales tested. RMSE (Root Mean Square Error) for wheat trained with RF was 320 kg/ha (MLR scored 1,320 kg/ha), 1,130 kg/ha for maize (MLR scored 1,940 kg/ha), 2,770 kg/ha for potato (MLR scored 5,620 kg/ha) and silage maize RF scored 1,900 kg/ha (MLR 4,540 kg/ha).

On the other hand, (Cunha, Silva, and Netto 2018) studied the implementation of Long Short Term Memory (LSTM) Neural Networks as an alternative to cope with different time scale information and deal with monthly variable weather data combined with timeless soils properties in order to predict annual crop yield for maize (US) and soybean (Brazil and US). They reported that performances achieved with LSTM neural networks are comparable to those obtained with Normalized Difference Vegetation Index (NDVI) and remote sensing data, with the advantage of being able to forecast yield before crop is planted and resulting in a more scalable model: no need of large amount of remote sensing data and scalable to regional level as well.

Barley biomass and grain yield was predicted using ANN models by (Ayoubi and Sahrawat 2011). They compared ANN predictability performance with earlier tested statistical models based on multivariate regression. ANN yield models resulted in higher coefficient of determination and lower RMSE compared to multivariate regression. A comparison of most important factors between the two models revealed that soil organic matter was misled by MLR because of non-linear relationship with other soil properties. This factor is well known to be a key variable in terms of soil fertility and is proven to be related to crop yield. Similarly, (Kaul, Hill, and Walthall 2005) also compared ANN yield prediction performance with MLR, in US corn and soybean and concluded that ANN outperformed MLR.

Recently, a combination of statistical modelling and neural networks was investigated by (Crane-Droesch 2018) to predict impact of climate change in crop yield. They proposed an approach that uses semiparametric variant of deep neural network, which can simultaneously account for complex non-linear relationships in high dimensional dataset, as well as known parametric structure to predict maize yield in US agriculture. They claim that the hybrid model implemented outperforms both classical statistical methods and fully non-parametric neural networks in predicting climate change impact on maize yields.

This study will compare and analyse 3 different machine learning algorithms ability to predict crop yields:

1. Lasso (*Least Absolute Shrinkage and Selection Operator*), a MLR algorithm that will be used as benchmark.
2. Random Forest Regressor: A non-linear regression model.
3. LSTM (*Long Short-Term Memory*) ANN (Artificial Neural Network).

WEATHER SIMULATIONS

One of the main goals of crop simulation models is to estimate agricultural production as a function of weather and soil conditions (Murthy and Radha 2003). According to (Murthy and Radha 2003), simulation is defined as "reproducing the essence of a system without reproducing the system itself". Essential characteristics of the system are reproduced and model, which is then studied in an abbreviated time scale.

Different curve fitting techniques, interpolation, extrapolation functions are being followed to use weather data in the model operation. (Van Wart, Grassini, and Cassman 2013) analysed the reliability of gridded weather databases (GWD) to simulate crop yield potential and water-limited yield potential, which are considered as benchmarks to quantify climate change scenarios on crop productivity and land use change. Among their discoveries, they concluded that studies based on GWD to simulate agricultural productivity in current and future climates are highly uncertain. They compared models based on GWD with control weather data (CWD), resulting the first one with poor prediction ability, strong bias and large root mean square errors (RMSEs). They conclude that climate scenario on location-specific observed daily weather databases combined with appropriate upselling method could lead to much better results.

Stochastic weather models can be used as random number generators whose input resembles the weather data to which they have been fit (Murthy and Radha 2003). While the earliest crop simulation models were based on photosynthesis and carbon balance, next crop modelling generation involved combination of multiple inputs that resulted in superior analytical, statistical and empirical models (Murthy and Radha 2003). However, when many inputs are added the models become more complex.

In planning and analysing agricultural systems it is essential to consider variability and think in terms of the components of the system (Murthy and Radha 2003). As (Murthy and Radha 2003) indicate, the principal effect of weather on crop growth and development are understood and predictable. In consequence, simulation models can predict responses to large variations in weather.

Many applications are derived from successful weather simulation models. From an agronomic management perspective, simulation models can help to set the optimum planting date, choose the appropriate hybrids or varieties and evaluate weather risk (Murthy and Radha 2003).

Crop simulations models can also help policy management by estimating global production for famine early warning, which can allow policy makers save time to prepare and work to ameliorate the effect of global food shortages (Murthy and Radha 2003).

Among the first crop simulation models, Decision Support System for Agro Technology Transfer (DSSAT) was widely used in research and as a teaching tool. The model combines ecosystem data (soil, weather genetics), management inputs, crop simulation models with weather generation programs to predict crop growth and yields (Murthy and Radha 2003).

(Wu et al. 2010) worked on adapting soil and weather data from VEMAP (Vegetation Ecosystem Modelling and Analysis Project) into a format readable by DSSAT modelling software. They managed to convert the dataset and compared the resulting weather simulations with ground stations observations. Contrary to (Van Wart, Grassini, and Cassman 2013) findings, which proved that control weather data (CWD) was more accurate to predict crop yields compared to simulated grid data (GWD), T-test conducted by (Wu et al. 2010) between the daily weather data and the simulated VEMAP dataset showed that no significant differences were found on yield prediction and concluded that VEMAP can be used for crop model applications.

In this study, regional weather dataset will be interpolated and used as a basis for stochastic simulated scenarios that will provide crop yield response to weather variability. Furthermore, a combination between actual and simulated information will emulate real situations with partial level of certainty, and forecasts will be modelled to analyse the ability of models to predict crop yields with different levels of real information.

JUSTIFICATION

There are no records of studies made using machine learning modelling to predict crop yield in Argentinian agriculture. Local data processing and consolidation is a challenge and could partially explain why the adoption of machine learning algorithms is taking longer in Argentina than in the rest of important agricultural countries, such as United States, India, Brazil and Asian and European countries.

The outcome of this study will complement other tools used to estimate regional crop production and will support policy management to understand the impact of soil and weather variability across the country. Weather simulations will help to measure the risk into which each crop and region is exposed and will contribute to give more accurate information to the agricultural industry. Crop insurance industry, commodity traders and manufacturers among others could benefit from better understanding yield variability across regions.

Furthermore, the ability to forecast yield before the crop season is finished gives the whole industry valuable information for decision making and perform strategic changes based on this information. Farmers could fix prices upon accurate future production, traders could anticipate deficit and surplus and foreseen price volatility, and government can adjust trading policies to meet domestic market demands.

OBJECTIVE

This work propose to model yield crops using 2 different machine learning techniques (ensemble modelling and Artificial Neural Networks) taking MLR as a benchmark for regional crop yield predictions for 9 annual crops: wheat (*Triticum aestivum*), barley (*Hordeum vulgare*), maize (*Zea mays*), soybean (*Glycine max*), sunflower (*Helianthus annuus*), sorghum (*Sorghum bicolor*), rice (*Oryza sativa*), cotton (*Gossypium herbaceum*) and peanut (*Arachis hypogaea*).

Specifically, Random Forest and Long Short-Term Memory (LSTM) Neural Networks (NN) are trained with Argentinian soil and weather data to predict yields for the 9 crops mentioned, spread across 6 provinces and 85 counties within seasons 1970 to 2018.

Additionally, this study will analyse and understand how the different soil and weather factors contribute to explain crop yield and will discuss the differences between the importance levels of the mentioned factors across crops evaluated.

Although all factors will be considered in the first modelling phase, it is a side objective to generate a simplified model that could be easier to extrapolate to the agricultural industry without depending too much on detailed information to be used, but at the same time could keep a reasonably good performance compared to an initial model. Ideally, a few soil and weather variables would be able to describe accurately yield crops variability.

Finally, it is of the interest of this study to conduct weather simulations to forecast crop yield before the end of the season. This will be implemented by emulating monthly scenarios based on actual weather data and simulating the remaining period to predict crop production.

METHODS

METHODS

DATASET

The datasets used in this study are public. Soil data was accessed from (GeoINTA, n.d.), weather data was collected from the weather station network program from (INTA, n.d.), a national research institute. Crop yields were collected from (Agroindustria, n.d.), and Total Factor Productivity (TFP) was gathered from (Lema 2015).

Soil dataset contains geospatial information for the whole Argentina, at a scale 1:500,000 (1 cm in the map equals 5 km in the terrain). The data is presented in polygons, which can be related to specific locations (counties and provinces) within the country.

These are the soil variables took under consideration, excluding the first 3 variables, the rest are categorical:

- Productivity index (numerical)
- Soil deepness (numerical)
- Slope (numerical)
- Main soil limitation
- Secondary limitation
- Third limitation
- Group classification
- Texture of superficial soil
- Texture of sub superficial soil
- Drain capacity
- Alkalinity degree
- Susceptibility to hydric erosion
- Susceptibility to wind erosion
- Susceptibility to floods

Weather dataset consists of 240 weather stations reporting daily data, each of them with different time series window (some stations records go back to 1930, whereas some others only have 10 years of data or even less). Each weather station is geolocated and can be linked to a county within Argentina. These are the main variables gathered in the dataset:

- Heliophany
- Sum of cold hours
- Humidity
- Snow
- Rainfall
- Average pressure
- Global and net radiation
- Mean temperature (shelter)

- Maximum temperature (shelter)
- Minimum temperature (shelter)
- Mean soil temperature
- Average wind speed (1000 cm)
- Average wind speed (200 cm)

TFP is an annual index that establishes the level of efficiency of agriculture in Argentina. According to ((IFPRI) 2018), it is an indicator of how efficiently agricultural land, labour, capital, and materials (agricultural inputs) are used to produce a country's crops and livestock (agricultural output). It is calculated as the ratio of total agricultural output to total production inputs. When more output is produced from a constant amount of resources, meaning that resources are being used more efficiently, TFP increases. (Lema 2015) calculated different TFP for agriculture and for livestock, in this study the agricultural TFP will be used.

Crop yield dataset holds annual data at a county level for different crops across Argentina. This study focuses in 9 main crops of Argentinian agriculture: soybean, maize, wheat, sunflower, sorghum, barley, rice, peanuts and cotton. Except for winter wheat and barley, the remaining 7 crops are summer crops. The season for summer crops in Argentina spreads from September until May, whereas for winter crops the season starts in May and finished in January. For this study and according to Argentinian agriculture season layout, a standard season is defined from July until June of the next year.

A generic model trains with known 'X' features (set of independent variables) to predict unknown 'y' values (dependent variable). In this study, X represents annual observations of soil and weather data. Because soil characteristics do not change over the period analysed in this study, soil factors remain the same year after year for each region. Weather variables are expressed annually as a combination of average, minimum, maximum and cumulative observations. Each weather record represents an annual observation composed by a set of variables in different times of the year (e.g. minimum temperature in June, average mean annual temperature, cumulative rainfall from July to September). The dependent "y" variable is the annual crop yield for a given crop, expressed as kg/ha (kilograms per hectare). This variable depends on soil characteristics (productivity of the land) and seasonal weather factors. The resulting configuration defines the dimension of the coordinate X, y for each crop as following:

- Observations (rows)
    - y: Number of years of crop yields for each region.
    - X: Number of years with soil and weather data for each region. Same number of observations for X and y.
- Features (columns)
    - y: 1 feature (crop yield)
    - X: Up to 600 features between soil and weather factors depending on the complexity of the model and the degree of completeness of the dataset for a given region.

It is important to mention that crop yield in a given season will be affected by weather conditions starting from July until June of the following year. Therefore, weather features will

be transformed to the mentioned calendar, for example accumulated rainfall will start in July, and accumulate monthly until June next year.

MODELS

With the objective of bringing together different techniques to predict crop yields, three algorithms with different properties are tested and compared in this study:

I. LASSO

Lasso (*Least Absolute Shrinkage and Selection Operator*) is and advance linear regression model suitable for high-dimensional regression problems. Is was proposed by Tibshirani in 1996 as a new method for estimation in linear models, and became very popular in genomics in the 2000's. As the name indicates, it serves two purposes that ordinary least squares (OLS) do not have: a penalty parameter controlling the amount of shrinkage and variable selection. According to Tibshirani, OLS estimates often have low bias, but suffer from high variance (Bergersen 2013). Lasso estimator β' is defined by:

$$\beta' = argmin \{ \frac{||y - X\,\beta||^2}{2} + \lambda \sum |\beta j| \}$$

From the equation above, lambda is a regularization factor that can take values from 0 to 1. When lambda = 0, the equation reduces to an OLS algorithm. This type of regularization can lead to zero coefficients (i.e. some of the features being completely ignored for the evaluation of the output), therefore Lasso not only helps reducing overfitting but also helps in feature selection (analysing which variables are more important to the prediction).

In this study, Lasso algorithm will be used as a baseline to benchmark the other 2 models' ability to predict crop yields.

II. RANDOM FOREST REGRESSOR

Random forest is a supervised learning algorithm that builds multiple decision trees and merges them together (this is also known as *ensembling*). As the name indicates, is a collection of multiple regression trees, where each tree is generated based on the values of an independent set of random vectors (Tan, Steinbach, and Kumar 2006).

The training algorithm applies a technique called bootstrap aggregating, also known as *bagging*. Given a set $X = x_1, \ldots x_n$ with responses $Y = y_1, \ldots y_n$, bagging repeatedly select a random sample with replacement of the training set and fits trees to these samples (Biau 2010). Each time a split in a tree is considered, a random sample of *m* predictors is chosen as a split candidate from the full of *p* predictors. The split is allowed to use only one of those *m* predictors (James et al. 2009).

Each randomize tree predicts the output variable and results from all trees are averaged to get an aggregated prediction (James et al. 2009). Bagging leads to better model performance because it decreases the variance of the model, without increasing the bias. This means that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated (Amnon Shashua 2017).

In this work, Random Forest algorithm will be trained with soil and weather information to predict crop yields and results will be compared against baseline algorithm (Lasso).

III. LONG-SHORT TERM MEMORY NEURAL NETWORK

Neural Networks (NN) can learn complex non-linear patterns thanks to their strong self-learning and self-adaptive capabilities (Azzouni and Pujolle 2017). NNs can estimate almost any linear or non-linear function when the underlying data relationships are unknown. This model is an adaptive approach which relies on the observed data rather than on an analytical model.

A neural network consists of interconnected nodes, called neurons. The interconnections are weighted (weights) and neurons are organized in layers (Azzouni and Pujolle 2017): An input layer, one or more hidden layers and an output layer.

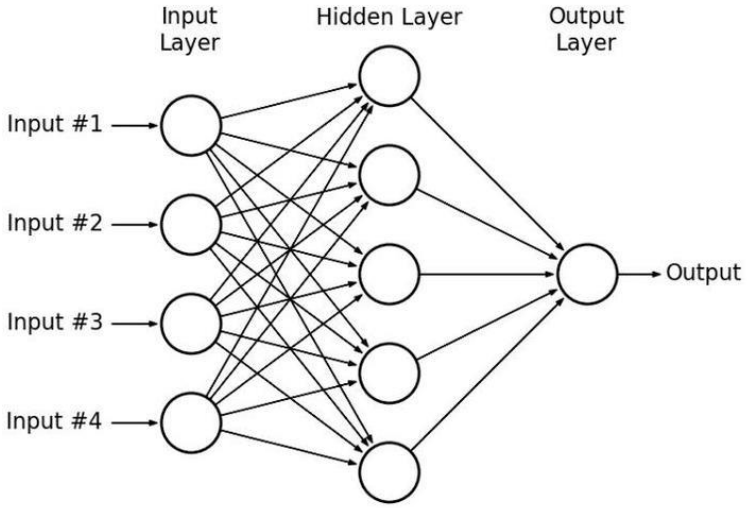


Fig. 1: Illustration of Multi-Layer Perceptron (MLP) for simple non-linear regression (Jaokar 2019).

Long-short term memory (LSTM) is a specific recurrent neural network (RNN) architecture that is well suited to learn from experience to classify, process and predict time series with time lags of unknown size. LSTM neural networks are composed by units called memory blocks, and each memory block contains memory cells with self-connections storing the temporal state of the network, in addition to multiplicative units (gates) to control the flow of information (Azzouni and Pujolle 2017). LSTM are useful when dealing with data with a

temporal relationship and can learn to recognize temporally extended patterns in noisy sequences (Cunha, Silva, and Netto 2018).

As proposed by (Cunha, Silva, and Netto 2018), to train a LSTM the dataset is divided into a static set and a dynamic set. The static set is integrated by the soil data, which for the time scales consider in this work soil properties do not change over time. The dynamic set is integrated by the monthly weather set, containing meteorological data with seasonal variability.

MODEL PIPELINE

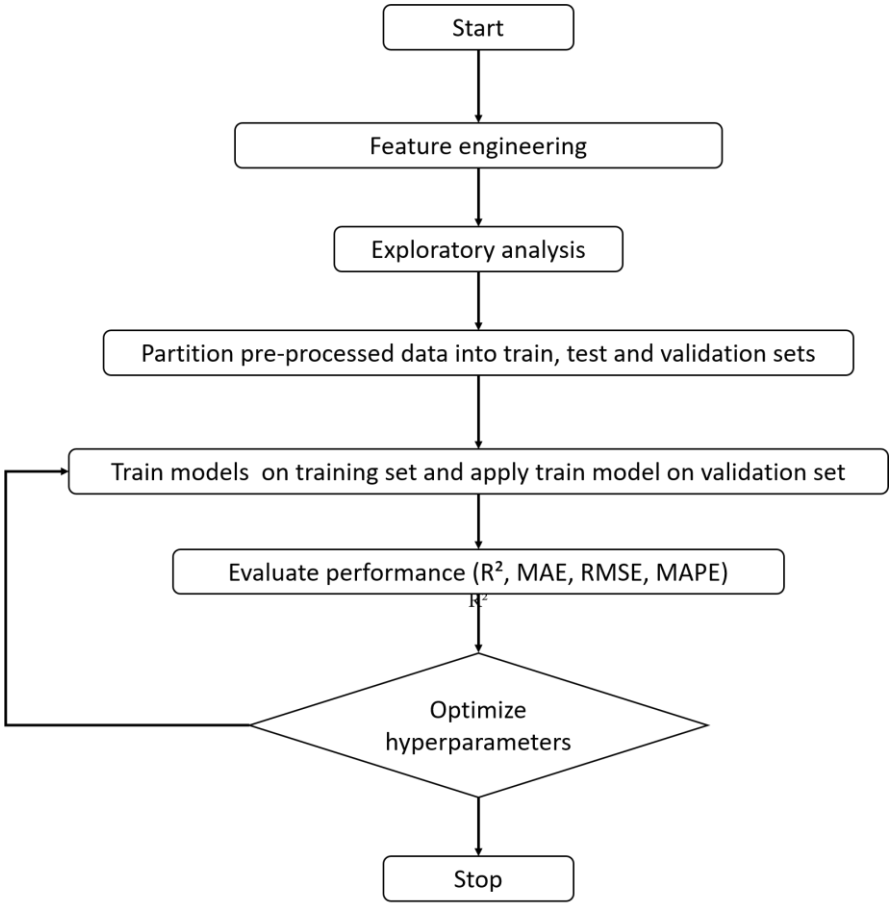


Fig. 2: Pipeline for modelling crop yields with 3 machine learning algorithms (Lasso, Random Forest and LSTM Neural Networks)

I. FEATURE ENGINEERING

Feature engineering is the process of transforming the given data into a form which is easier for the model to interpret (Grabiński 2018). As (Zheng and Casari 2018) indicates, it is a crucial step in the machine learning pipeline, because the right features can ease the difficulty of modelling, and therefore enable the pipeline to output results of higher quality.

Each of the different datasets has different temporal and spatial levels of granularity. To get meaningful data, it is necessary to standardize the information into a same time and space. This study analyses annual data at county level.

Soil data is timeless, and spatially represented by around 7,000 polygons across 500 counties in Argentina. Transformations consist in calculating the area represented by each polygon for each county. Then, weighted averages are obtained for each variable using the area of each polygon as the weight. The resulting transformed dataset consolidates timeless data at a county level, with the variables mentioned before.

Weather data has more challenges than the rest of the datasets. To begin with, the data is not consistent across weather stations. Each station has different amount of missing data and different time frame. Spatially, each station is related to a coordinate, a specific point in the map, and the target for the analysis requires aggregated county level information.

A completeness analysis was conducted to evaluate each weather station degree of consistency, combined with a variable analysis: analysing each individual variable to determine if a variable has enough information to be included in the modelling stage. Threshold levels were introduced to classify wanted/unwanted variables and regions. A time frame was also fixed from 1970 onwards, which is the starting point for 2 of the other datasets (crop yields and TFP index).

Secondly, an *Inverse Distance Weighting* (IDW) interpolation was performed across the weather dataset. For each daily missing observation, the missing value is replaced by a weighted interpolation from the neighbour weather stations, and the weight is inversely proportional to the distance between the target and the interpolated station. Although IDW has well known short comings when performing spatial interpolations, proves to be very effective for the weather parameters analysed (Yang 2010).

Additionally, new features were created derived from the original variables, mostly related to temperature and water availability. According to (Prieto 2010), summer crops suffer with temperatures higher than 28° C, and can also be affected by lower amplitude temperatures (difference between maximum and minimum temperature). Mean temperatures below 24° C can also affect yield. Taking all this into consideration, new variables were generated:

- Evapotranspiration calculated with Hargreaves equation (Strzepe 1994).

$$\mathrm{Erc} = 0.0022 * \mathrm{RA} * (\mathrm{Tmax} - \mathrm{Tmin}) * 0.5 * (\mathrm{Tmean} + 17.8)$$

  - RA: Radiation
  - T: temperature

- Growing degree days

$$GDD = \left(\frac{T\,\mathrm{max} - T\,min}{2}\right) - T\,base$$

  - T base: 8° C

- Number of days with maximum temperatures higher than 28° C.
- Number of days with mean temperatures lower than 24° C.
- Number of days with minimum temperatures lower than 0° C

Finally, the resulting daily dataset was aggregated annually and by county. For spatial aggregation, variables were classified into mean, maximum and minimum. For annual aggregation, apart from mean, maximum and minimum calculations, sums where computed for

variables that needed addition, such as rainfall, growing degree days and number of days with temperature below 0° C.

Total Factor Productivity (TFP) index dataset is an annual time series index for the whole country. The index was applied equally across all counties for each year.

Random regression models cannot deal with missing information or incomplete datasets. Null data was removed from the training dataset in two dimensions:

1. Remove records where datapoint for the predictable variable (yield) are missing.
2. Remove fields with missing information. Thanks to the interpolation in the most relevant weather dataset pursued during the data processing stage and thanks to the completeness of the soil map, most of the variables were able to pass the completeness filter and were considered in the modelling stage. Having said that, the remaining variables varied for each crop modelled.

Normally, when adapting a dataset to fit a model, variables are split into numerical and categorical. Another necessary transformation to fit regression models is to convert categorical variables to numerical. While there are a few ways of doing this, in this study two mechanisms were tested and selected the one that performed better according to the performance evaluation of each model.

1. *Label binarizer*: Also known as ordinal coding, converts categorical variables into numerical. This methodology keeps the same number of fields in the dataset, by only replacing each category field by a number that is assigned to each category value. A potential drawback of ordinal coding is that implies an order to the variable that may not actually exist.
2. *One hot encoding*: Compares each level of the categorical variable to a fixed reference level. One hot encoding transforms a single variable with n observations and d distinct values into a d binary variables with n observations each (Potdar, S., and D. 2017).

One hot encoding proved to perform better compared to label binarizer and was implemented across all models tested.

Numerical data can be scaled using different techniques. A *MinMaxScaler* was implemented into the dataset but the results showed slight better performance with the original scale, so the transformation was dismissed. What this technique does is transforms features by scaling each value to a given range (e.g. between zero and one).

Feature selection techniques prune away non-useful features in order to reduce the complexity of the resulting model (Zheng and Casari 2018). The goal is to end up with a simplified model, quicker to compute and with little or no degradation in predictive accuracy. While there are different techniques to perform feature selection, such as filtering and wrapper methods, this study leverages the use of random forest model to extract feature importance from the trained model. This technique is called embedded (Zheng and Casari 2018), and occurs as part of the model training process.

Initially, each crop model was trained with more than 600 variables, combining both soil and weather features. After analysing feature importance and identifying the critical

features, a feature selection was performed. This work tries to reduce the number of features up to 10, trying to prioritize the features that are easier to collect, without compromising too much the prediction accuracy. Another advantage of keeping a reduced number of variables is the ability to perform simulations when the number of variable factors is limited. Soil properties are considered not to change within a season, leaving only the weather variables subject to seasonal variability. In summary, Random Forest feature importance property will be used in this study to minimize the number of weather variables without giving up too much performance of the models.

II. EXPLORATORY ANALYSIS

Weather information was evaluated by distribution analysis. Monthly data for each region was plotted to detect patterns in the most important climate variables, such as rainfall, moisture, extreme and mean temperatures.

Despite season variations, historical yields follow an ascending trend due to technological improvements. This is what TFP index is about and the agricultural factors involved during the period analysed are seed genetics, machinery, pesticides efficiency and precision in agriculture.

To isolate the importance of time scale, crop yield was adjusted by TFP index prior to the training process. This helped to better differentiate the remaining important variables when performing feature selection.

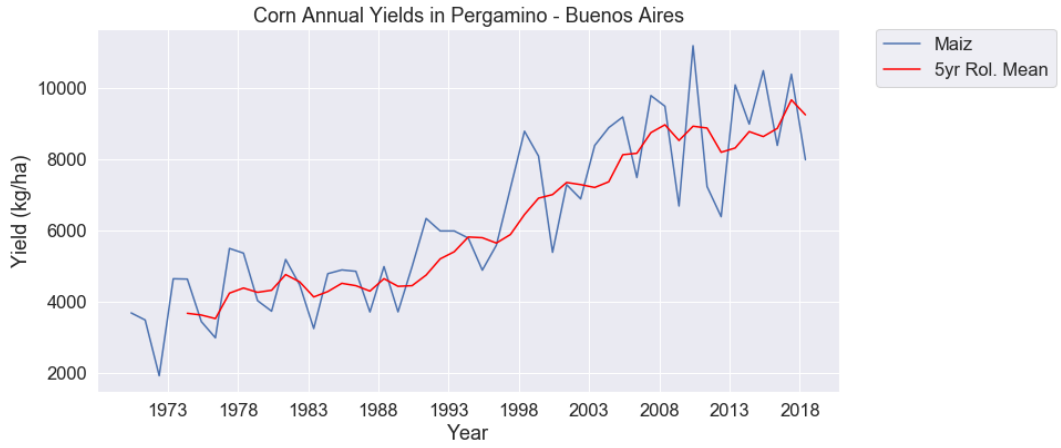


Fig. 3: Maize historical yields for Pergamino, Buenos Aires [blue] and maize 5 year rolling average annual yield for the same region [red].

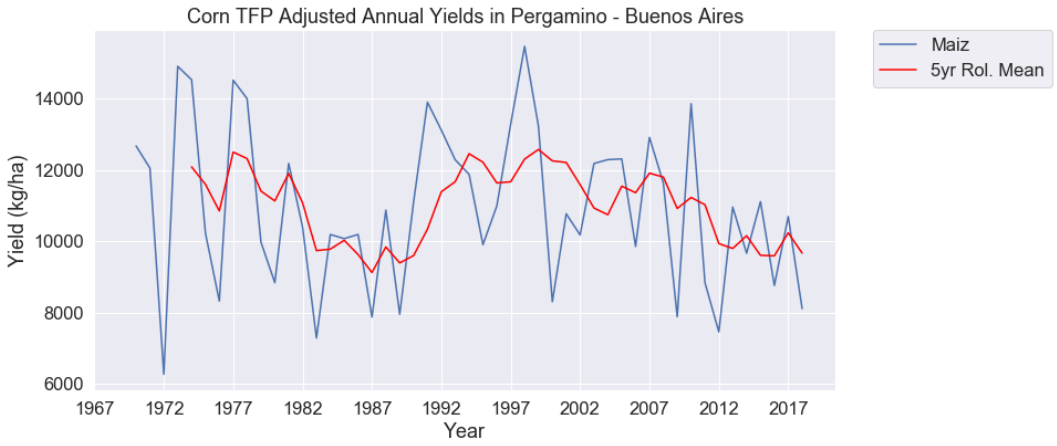


Fig. 4: Maize historical yield for Pergamino, Buenos Aires adjusted by Total Factor Productivity (TFP) [blue] and 5 year rolling average of the adjusted annual yield [red].

A coverage analysis was performed to understand the scope of the models to be trained. For each crop, the number of counties was registered together with the proportion of the national production for the given crop represented by the training dataset. This is a heuristic approach to evaluate the importance of the training set for each crop and gives an indication of level of confidence to the model trained.

III. SEPARATE INTO TRAINING, VALIDATION AND TEST DATASET

Once the data was ready to input into the different models, 10% of the whole dataset was put apart as a test set, which will be use at the end of the research to evaluate each model's performance.

The remaining 90% of the dataset was introduced into the model pipeline. 70% of the data was used as training set and 30% as validation set. During training process, a seed was planted to neutralize the random state effect when splitting the data. This helps to isolate the performance of the model from random effect of sampling.

The same training, validation and test set was used to train all three algorithms to ensure a fair comparison.

Table 1: Dataset size. Number of observations per crop (total) spitted into train, validation and test.

crop	total	train	validation	test
Cotton	142	89	39	14
Rice	234	147	64	23
Barley	585	368	159	58
Sunflower	968	609	262	97
Corn	1175	739	318	118
Peanut	165	104	45	16
Soybean	1007	634	272	101
Sorghum	998	628	270	100
Wheat	1084	683	293	108

IV. HYPERPARAMETER OPTIMIZATION

Machine learning use the term hyperparameter to distinguish from standard model parameters. While standard parameters are the ones learned from the data, hyperparameters are the properties that govern the entire training process (Prabhu 2018). A regularization hyperparameter controls the capacity of the model, how flexible the model is, how many degrees of freedom it has in fitting the data. Proper capacity can prevent overfitting, which happens when the model is too flexible, and the training process adapts too much to the training data (Zheng 2015).

Hyperparameter tuning works by running multiple trials in a single training job. Each trial of a hyperparameter setting involves training a model. The outcome of hyperparameter tuning is the best hyperparameter setting, and the outcome of model training is the best model parameter setting (Zheng 2015).

For all three algorithms considered in this work (Lasso, Random Forest and LSTM Neural Networks), hyperparameter tuning was implemented with Random Search Cross Validation technique instead of Grid Search, which tries every combination of a pre-set list of values of the hyperparameters and choose the best combination based on the cross-validation score. Reaching every possible combination can take a lot of time. Instead, Random Search tries random combinations of a range of values (number of iterations needs to be defined). Normally, Random Search reaches a very good combination of hyperparameters faster than Grid Search does. The problem is that it does not guarantee to give the best parameters combination. For Lasso and Random forest, the number of iterations was set to 20 and the cross-

validation size was set to 10. Because training LSTM took significantly more time, the number of iterations was set to 10 for this model.

Lasso hyperparameter and values tested:

- `lambda`: Regularization factor that controls the amount of shrinkage. Lambda can take values from 0 to 1 (when lambda = 0, Lasso algorithm is equivalent to OLS algorithm). In this study, the range tested was [0 - 1].

Random forest hyperparameters and grid are described below:

- `n_estimators`: Represents the number of trees in the forest. Usually, the higher the number of trees the better to learn the data. However, adding a lot of trees can slow down the training process considerably (Mohtadi 2017). In this study, the range tested for n_estimators is [100, 300, 500].
- `max_depth`: Indicates how deep a tree can be. The deeper the tree, the more splits it has, and it captures more information about the data. But deeper trees can lead to overfitting, which means over learning from the dataset, taking the noise in consideration. Overfitting tends to have high variance and low bias. The range tested is [1 - 6].
- `min_samples_split`: The minimum number of samples required to split an internal node. This can vary from considering at least one sample at each node to considering all of the samples at each node (Mohtadi 2017). By increasing this parameter, the forest is more constrained, and the deepness is limited. A range from 5 to 100 was tested.
- `min_samples_leaf`: The minimum number of samples required to be at a leaf node (Mohtadi 2017). Like min_samples_splits. Increasing the value of this parameter, same as the previous one, can cause underfitting, which can be defined as the failure to capture the underlying pattern of the data. An underfitted model, in contrary to overfitting, has low variance and high bias. The options tested were: [5, 10, 15].
- `max_features`: Represents the number of features to consider when looking for the best split. This parameter remained unrestricted in this study. Options tested: [1,2,3].
- `min_impurity_decrease`: A node will be split if the split indicated a decrease of the impurity greater of equal to this value. Random samples from 0 to 1 were tested.
- `min_weight_fraction_leaf`: Defined as the minimum weighted fraction of the sum of weights required to be a leaf node. Tested values in the range of [0 – 0.5].
- `oob_score`: Out of bag error score, a way of measuring prediction error.

Optimization of Neural Network performance was accomplished by computing predictions on training and validation set while passing through the dataset. A full pass of the dataset is known as an *epoch*, and prediction and accuracy were computed after each epoch. This allows to observe the behaviour of training and validation performance across the epochs ran by the model.

Because Neural Network offer large flexibility, there are many hyperparameters to tweak. Not only how neurons are interconnected, but even in a simple Multi-Layer Perceptron (MLP) the number of layers can be changed, the number of neurons per layer, the type of activation function to use in each layer, the weight initialization logic, among others. Taking

the work done by (Cunha, Silva, and Netto 2018) as a reference, the number of layers and the activation function was implemented based on their experience working on yield predictions in maize and soybean in United States and Brazil. The optimizer chosen is '*Adam'*. and the number of hidden layers was set to 4. The hyperparameters optimized in this work follow the order set below:

- `epochs`: An epoch is defined as a full pass of the dataset. Too few epochs don't give the network enough time to learn the good parameters; too many might overfit the training data. Epochs tested: [10 - 150]
- `batch_size`: Refers to the number of examples used at a time when computing gradients and parameter updates. Values tested: [5 -150].
- `neurons`: The number of neurons per hidden layer is determined by the type of input and output the task requires. A recommendable practice is to start with all hidden layers with the same number of neurons, and gradually increase the number of neurons until the networks starts overfitting. This study tested the following values: [1 – 50].
- `learning_rate`: Is one of the most important hyperparameters. Defines how fast is the model going to learn from the train set. For this work, the values tested were [0,1; 0.001, 0,00001]

In each optimization instance, the process was repeated from 5 to 10 times (depending on the time consumed to execute the calibration) and averages where computed from these samples.

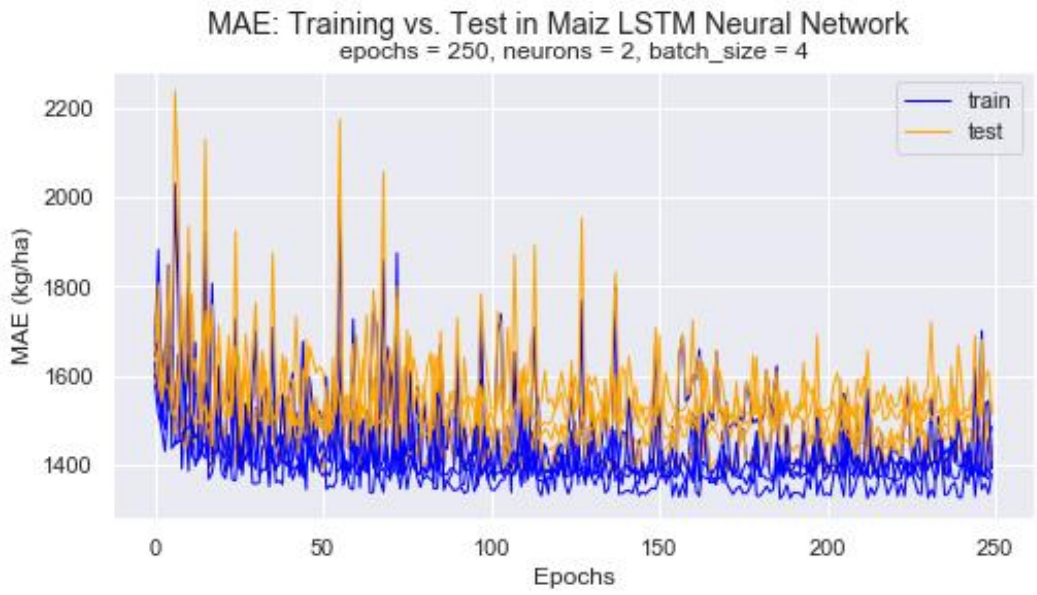


Fig. 5: Loss for training and validation set across epochs for Maize LSTM Neural Network model. Configuration: 250 epochs, 2 neurons, batch size = 4 and learning rate = 0.01.

V. EVALUATE PERFORMANCE

In a regression task, the model learns to predict numeric scores. The most commonly used metric for regression models is RMSE (root-mean-square-error), defined as the square root of the average squared distance between the actual score and the predicted score (Zheng 2015):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(yi - yi')^2}{n}}$$

Here, 'y_i' denotes the true score for the *ith* data point, and y_i' denotes the predicted value. While RMSE is very popular, it has some problems. Because is an average, it is sensitive to large outliers. If the regressor performs really badly on a single data point, the average error could be very big, which the model is not robust (Zheng 2015). Taking this into consideration, is useful to look at the mean absolute percentage error. MAPE gives a relative measure of the typical error:

$$\text{MAPE} = \frac{1}{n} * 100 \sum_{i=1}^{n} \left(\left|\frac{Yi - Yi'}{Yi}\right|\right)$$

MAPE has important shortcomings that need to be considered. It cannot be used if there are zero values and puts heavier penalty in negative errors than on positive ones. On the other hand, because it is a relative error to the mean, it is useful when comparing targets with different scale units and is commonly used as a quality score for forecasting models.

An alternative, Mean Absolute Error (MAE) is calculated as an average of absolute differences between the target values and the predictions. MAE is a linear score with means that all the individual differences are weighted equally (Drakos 2018). MAE penalizes huge errors and is less sensitive to outliers compared to RMSE.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |Yi - Yi'|$$

Coefficient of determination R^2 is closely related to RMSE, with the advantage of being scale free (it doesn't matter if the output values are very large or very small, the R^2 is between $-\infty$ and 1. When R^2 is negative it means that the model is worse than predicting the mean. R^2 can be also defined as the ratio between how good the model is and how good a naïve mean model is (Drakos 2018):

$$R^2 = 1 - \frac{\text{MSE (model)}}{\text{MSE (baseline)}}$$

This study will compute RMSE, MAPE, MAE and R^2 and will compare model performance using Lasso estimator as a baseline for Random Forest regressor and LSTM neural networks.

SIMULATIONS

An important section of this work is the application of machine learning models to predict outcome from simulated scenarios. It has already been mentioned that the variable factor in this crop model is the weather component. By simulating *n* times how weather is going to behave in a season, yield predictions can be estimated for each *n* simulation, and then statistical analysis can be performed to determine mean and variation of the outcomes from those simulations.

Monte Carlo simulation is a powerful statistical analysis tool widely used across different industries: engineering, finance and physics are clear examples. Its ability to deal with a large number of random variables and various distribution types and nonlinear models makes Monte Carlo suitable for complex problems (Askew 2012).

Monte Carlo simulation performs random sampling and statistical characteristics of the experiments (model outputs) are observed, and conclusions on the model output are drawn based on the statistical experiments (Askew 2012). In each experiment, the possible values of the input random variables X = (x1, x2, … xn) are sampled (generated) according to their distributions. Then the values of the output variable Y are calculated through the performance function. Repeating the same process for a number of experiments leads to a set of samples of output variable Y from which statistical analysis can be conducted.

The purpose of sampling on the input random variables is to generate samples that represents distribution of the input variable from their CDF's (*Cumulative Distribution Function*). The samples of the random variables will then be used as inputs to the simulation experiments (Askew 2012).

The two steps to carry out a Monte Carlo simulation are:

1. Generate random variables that are uniformly distributed between zero and one. This can be achieved using a random-generator.
2. Transforming [0, 1] uniform variables into random variables that follow the given distributions. While several methods exist to perform such transformations, *inverse transformation method* is used in this work.

In order to accomplish Monte Carlo experiments, its necessary to now beforehand the distribution of the variables to be simulated, which in this case are unknown. However, based on historical data, distribution can be estimated using different statistical tools.

Density estimation methods are statistical tools used to reconstruct an unknown PDF (*Probability Density Function*) from a set of samples drawn from the PDF. These samples are used to reconstruct the PDF using either nonparametric or parametric estimators. A non-

parametric estimator doesn't assume the underlying distribution in order to obtain estimates about that distribution, whereas a parametric estimator assumes the functional form of the underlying PDF (Burke 2016). In other words, non-parametric estimators make no assumption about the shape of the distribution, but parametric estimators do (e.g. assume Gaussian distribution beforehand).

This study will conduct non-parametric estimations by calculating Kernel Density Estimators (KDE) for each variable, for each region, in each of the 12 months of the season. KDEs obtain estimates of the underlying density at discrete points and was first referenced by Pazen in 1962 (Parzen 1962). The performance of the KDEs is heavily dependent upon the bandwidth. An optimal bandwidth is the one that minimizes the Mean Integrated Square Error (MISE), sum of integrated square bias and the integrated variance (Burke 2016).

On the other hand, the choice of the kernel has a minimal impact on the estimate of the underlying density compared to the choice of the bandwidth.

Once KDE was obtained for a given variable for each month on a given county, 1,000 Monte Carlo simulations (random number from 0 to 1) were computed and by the inverse transformation method, the CDF was obtained, from which the value from the variable that fall under that CDF was taken. This process was automated for main monthly weather variables across all the regions and a new dataset was generated from the simulations, which would then be used to make yield forecast combining actual weather data with simulated information.

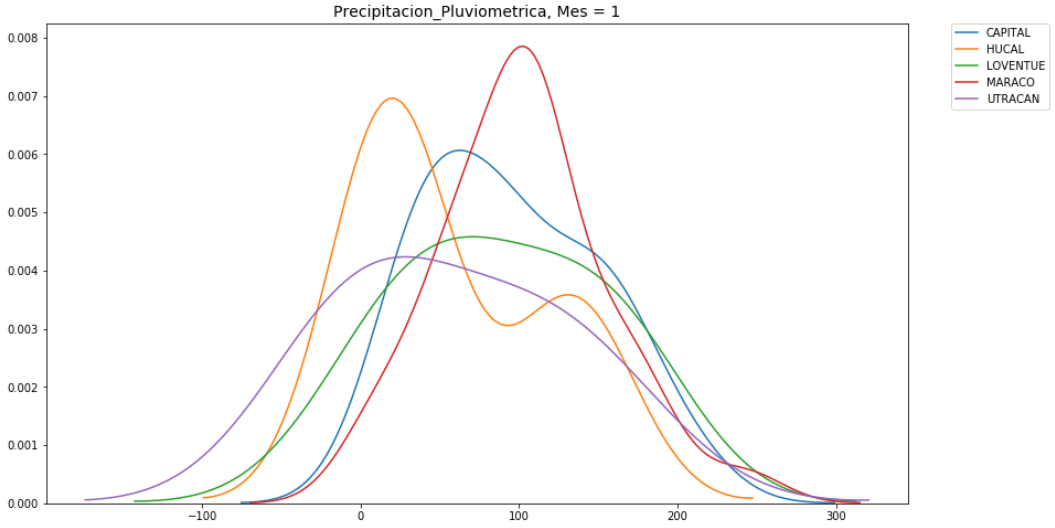


Fig. 6: Kernel Density Estimator (KDE) for historical January rainfall in 5 counties in La Pampa province. X axis indicated monthly rainfall (mm) and Y axis the density.

For the model to be a good representation of the reality and at the same time be manageable, a trade-off must be done towards reducing the number of dimensions or features to be used to train the model. The model needs to be accurate, but easy to use in practice. Therefore, modelling took part in 2 stages:

I. Stage 1: A complex model with more than 600 features involved (after applying one hot encoding). This allowed to identify the most important variables and quickly dismiss the ones that didn't' add value to the prediction.

II. Stage 2: A simplified model, with less than 10 variables and all being the same for all 9 crops. To simplify things even more, feature selection towards weather variables pushed to select only one weather feature, knowing that each variable would need to be simulated for each month, and dealing with more than one variable at the same time is computationally expensive, not to mention the difficulty to put the model in practice.

After training a simplified model with a reduced number of variables with a Random Forest Regressor for each crop, monthly weather simulations were conducted to forecast yield on 12 scenarios each one with different level of uncertainty:

1. For example, for a given crop staring the season in July, 100 simulations where conducted for each of the 12 months ahead.
    i. All 100 simulated weather data were then randomly grouped to conform 100 seasons.
    ii. Prediction was made on all 100 seasons and accuracy metrics were calculated.
2. Once predictions are made simulating all 12 months of the season, step 1 is repeated, but n-1 months each time, using the current month as a certain fact. For example, next step is to take July data as actual data and simulate the remaining 11 months. After that, July and August actual data plus October until June simulated data, and so on.

This results on 12 different stages of prediction, each one with different level of uncertainty. Starting from 100% simulations in July and finishing with 100% real information in June of the following year.

Moreover, to evaluate performance of the forecasted model, MAPE was calculated at each stage and then thresholds were implemented to evaluate at which stage of the season the model could make reliable recommendations based on the real data plus the simulations.

Two thresholds were implemented as acceptance level of accuracy for forecasting modelling:

1. Threshold 20: MAPE less or equal to 20 means the model is *qualified*.
2. Threshold 15: MAPE less or equal to 15 means the model is *good*.

SOFTWARE

This study was coded in is python and several python modules were implemented:

- Pandas and NumPy: Feature engineering and data analysis.
- Geopandas, Fiona and Shapely: Geospatial feature engineering and analysis.
- Matplot lib and Seaborn: Plotting packages.
- Scikit Learn:

        - Kernel Density Estimator (KDE)
        - Lasso Estimator
        - Random Forest Regressor
        - Model performance metrics (RMSE, MAE and R^2)
- Keras and Tensorflow High level API
        - LSTM Neural Networks

Code is available in Github:

https://github.com/truizmoreno/Crop-Modelling-Thesis

RESULTS

RESULTS

DATA REPRESENTATIVENESS

For each of the 9 crops county level and national production information was collected to calculate the amount of crop production that the counties involved in the training set represented from the national production. For example, if Argentina produces 60 million tons of soybeans on average every year, the training regions accounted for 60% of the total production (36 million tons). This qualitative analysis allows to make inferences about how trusted results can be for each crop.

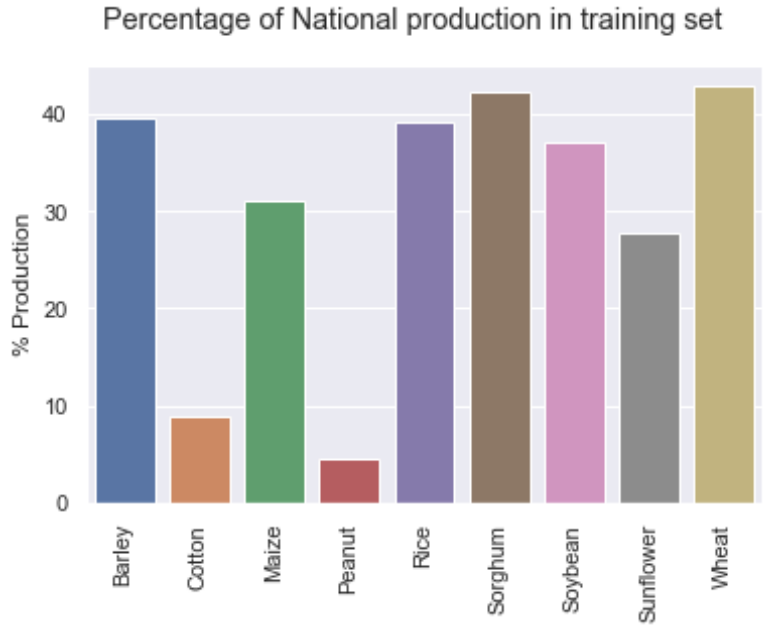


Fig. 7: Indicator of level of representativeness of training set over national production, calculated as percentage of production associated to the counties in training set over total national production of each crop.

Apart from cotton and peanuts, the counties used in the training set accounted for 30 to 40% of the national production, which is a reasonably good representation considering the extension of the country and the fact that some of these crops are widely spread across the country, specially soybean, maize and wheat.

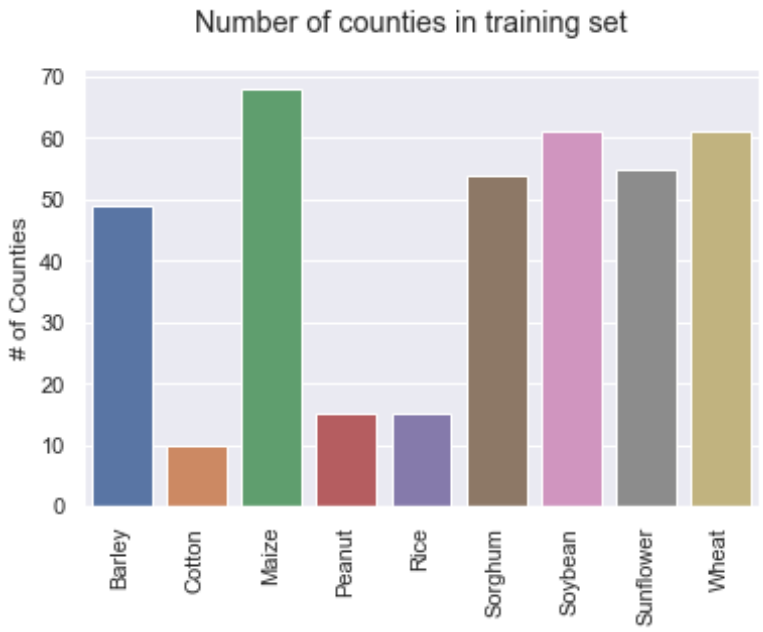


Fig. 8: Number of counties used to train the model for each crop.

Looking at the number of counties used in the training set, the numbers are consistent with the percentage of national production described above. Except for rice, which is confined to specific areas of the north east of the country, and the few counties in the training set represent a 40% of the national production.

MODEL PIPELINE

I. FEATURE IMPORTANCE

It is desirable to determine which variables and group of variables contribute most to the predictive skill of a model. *Important measures* were developed in the context of Random Forest training process. These statistics measure the decline in the accuracy when a variable in the out-of-bag sample is randomly permuted (Crane-Droesch 2018). Random permutation destroys their correlation with the outcome and with the variables with which they interact, remaining uninformative.

Although remarkable differences were found across the 9 crops, some common features can be extracted as the ones that appear to be important cross to all crops. Soil productivity index, (*IND_PROD* in the dataset), is a productivity score from 0 to 1 that reflects the ability of the soil to produce. Is a very powerful variable that summarizes many of the other soil features, such as type and degree of limitation, fertility, slope, risk of erosion and degradation, among others. This index is usually applied in the industry to classify farms between arable (highest score), mixed crop and livestock, dairy, livestock, and sheep farms. Three other soil

variables remained present across many of the crops, soil deepness of sub superficial horizon (*PROFUND_S1*) and two features related to soil fertility: *Natracuoles tipico* accounts for acid soils (ph. < 7) , with high levels of organic matter, while *Natracualfes tipico* are classified as alkaline soils with high concentration levels of exchangeable sodium from surface (Borrajo, Braco, and Ezcurdia 2011).

The most popular weather variables were the ones associated with water availability: accumulated rainfall in a certain month (*Precipitacion Pluviometrica acum|[month]*), hydric stress represented as the difference between precipitation and evapotranspiration (*pp_ETP*) and estimated Hargreaves evapotranspiration (*ET_h*).

Starting with the winter crops, wheat most important variables were related to soil composition. Important weather variables were maximum temperatures (*Bool_tmax_acum*) accumulated in September and November, Growing Degree Days or GDD (*GradosDia_acum|[month]*) accumulated to December and until the end of crop season (January), average annual soil moisture (*Humedad_media_14_8_20*) and annual rainfall. The fact that averages and annual variable were more important than accumulated ones could be related to the generalization of the crop season (July-June), when the real season occurs from May until January. Barley, the other important winter crop, shares a very similar pattern of feature importance.

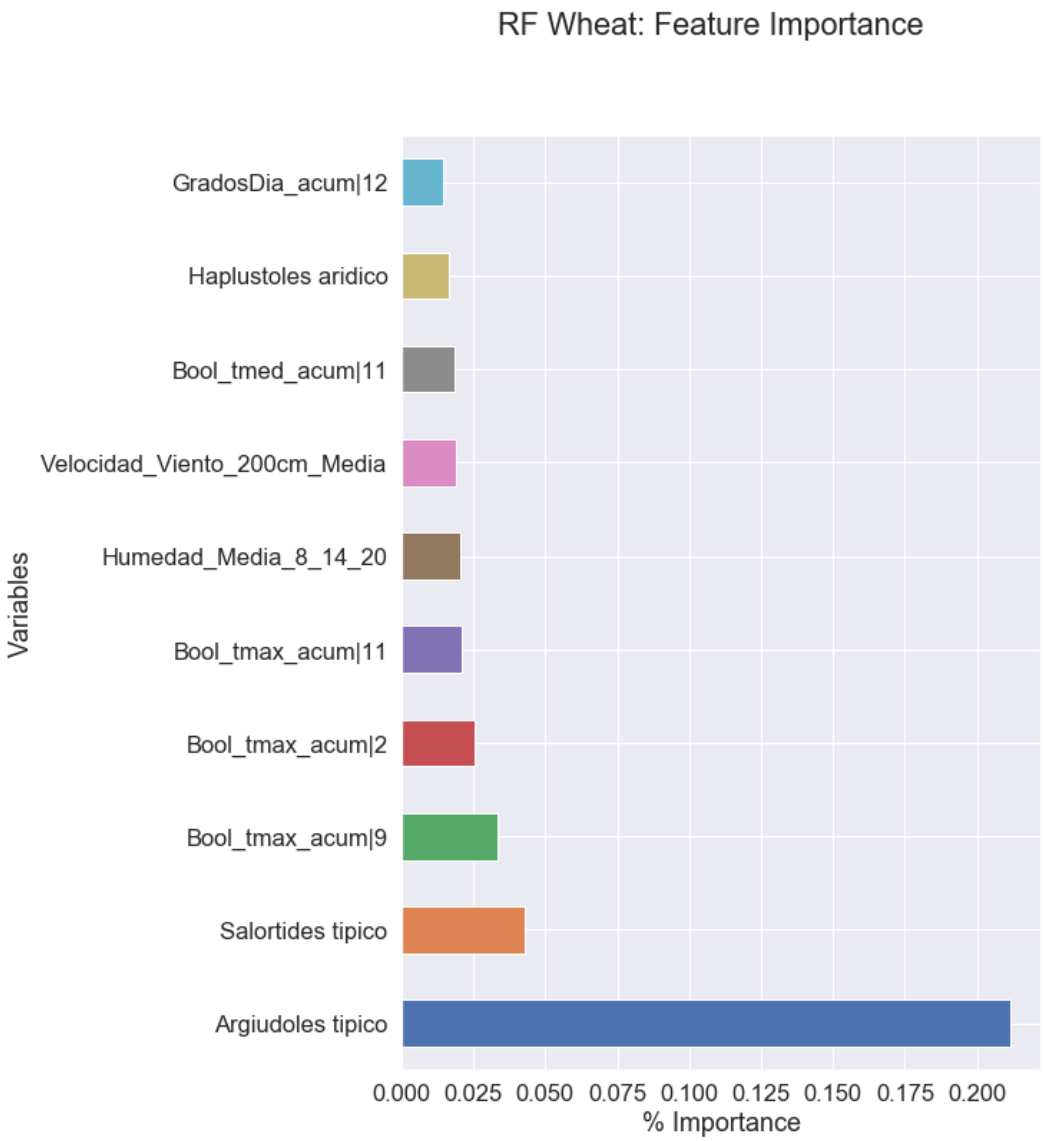


Fig. 9: Random Forest Regressor top 10 most important features for wheat according to Feature Importance.

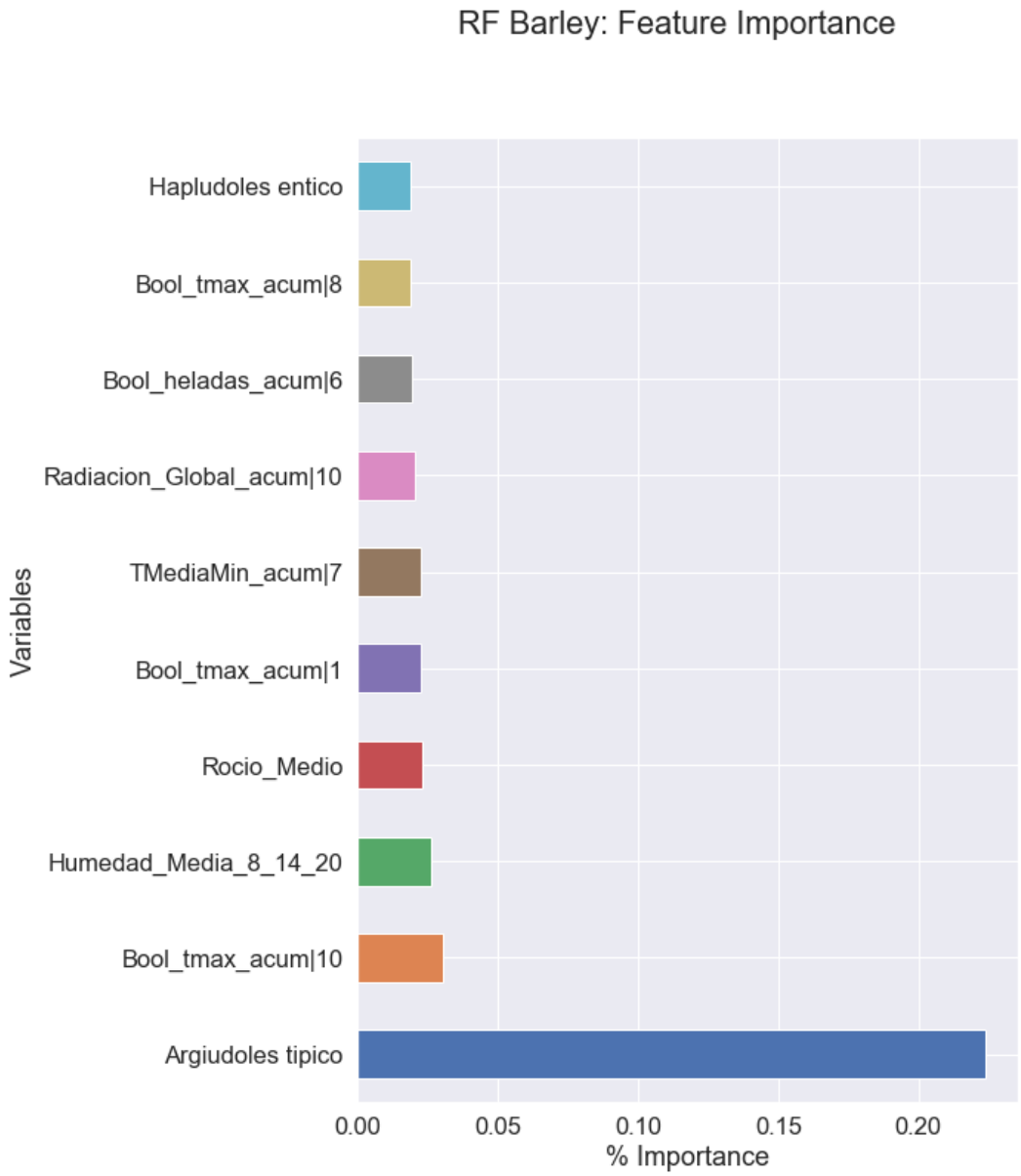


Fig. 10: Random Forest Regressor top 10 most important features for barley according to Feature Importance.

Moving on to summer crops, soybean most important variables were soil productivity index, followed by accumulated rainfall until February and April, and water stress accumulated in March. This makes total sense, since soybean is a crop very well adapted to a wide range of temperatures in Argentina and according to research mostly limited by water availability.

Sunflower most important variable is also soil productivity, followed by temperature amplitude accumulated in September (*TMedia_min_acum|9*) which matched with planting season (September). Wind speed appears to be important for this crop (*Velocidad_Viento_200cm_Media* and *Velocidad_Viento_200cm_Media*), explained by the fact that a substantial area where sunflower is planted suffers from high speed winds coming from the south of the country. Also, the heavy weight of the flower when the crop approaches

to maturity makes it very vulnerable to falls, and storms can make considerable losses in sunflower fields.

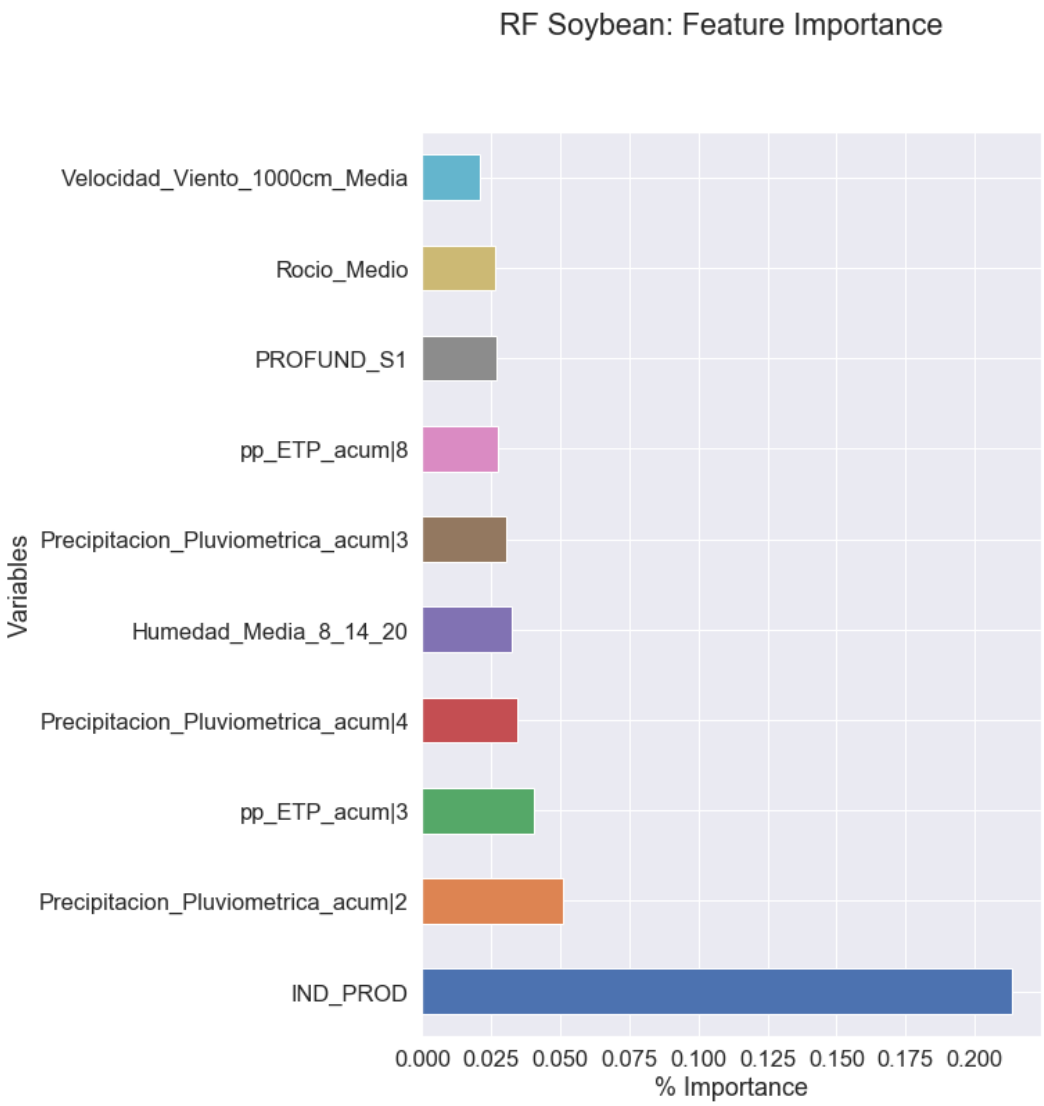


Fig. 11: Random Forest Regressor top 10 most important features for soybean according to Feature Importance.

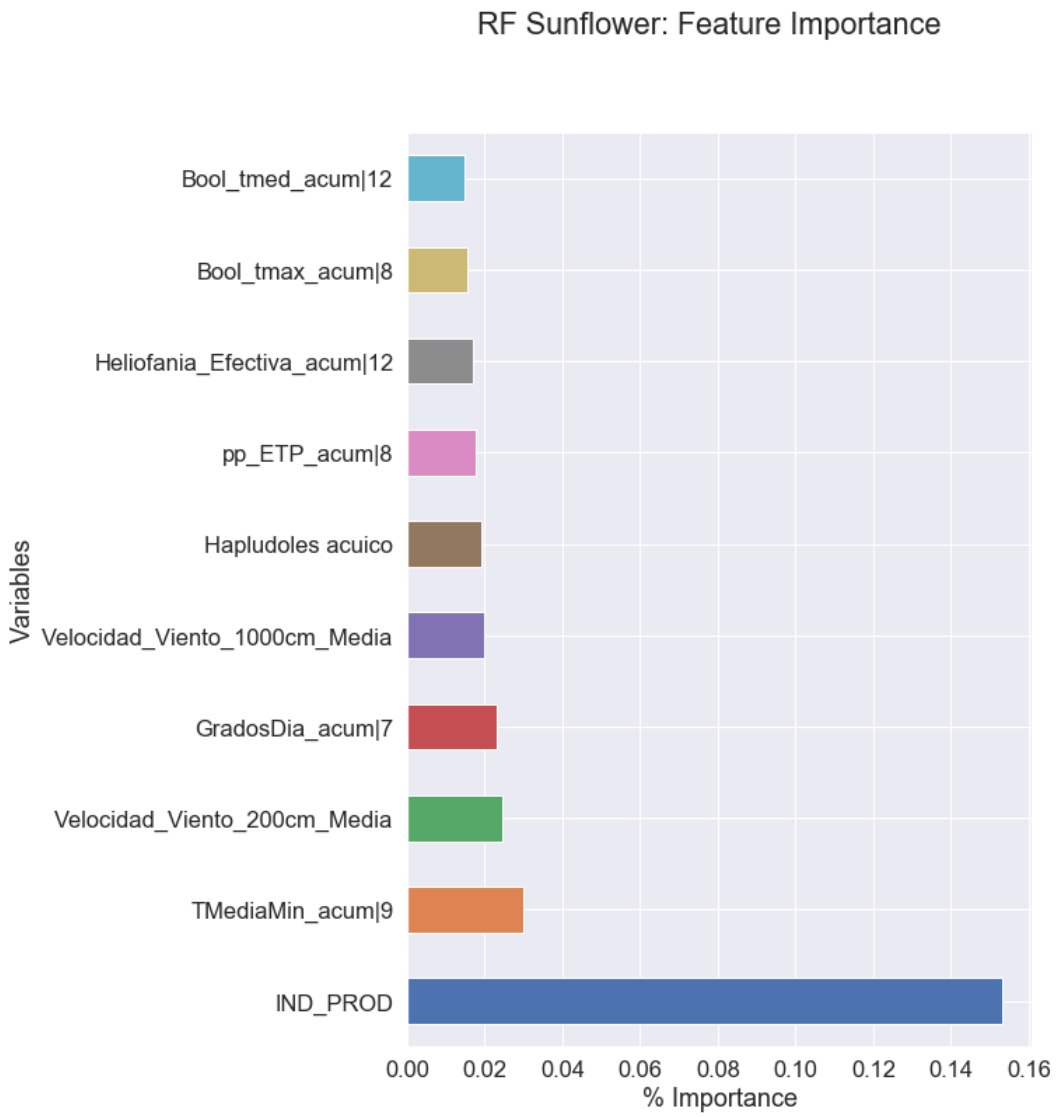


Fig. 12: Random Forest Regressor top 10 most important features for sunflower according to Feature Importance.

Maize most important feature by far is also soil productivity, followed by rainfall accumulated until December and January. This can be explained by the fact that maize is very demanding in nutrients, higher yields correspond to fertile soils; and very demanding of water in specific moments of the growing period (flowering is a critical stage in maize) (Prieto 2010). Sorghum most important features are very similar to maize 's, with the caveat that sorghum is less affected by rainfall (can tolerate drought better than maize) and is more restricted by maximum temperatures.

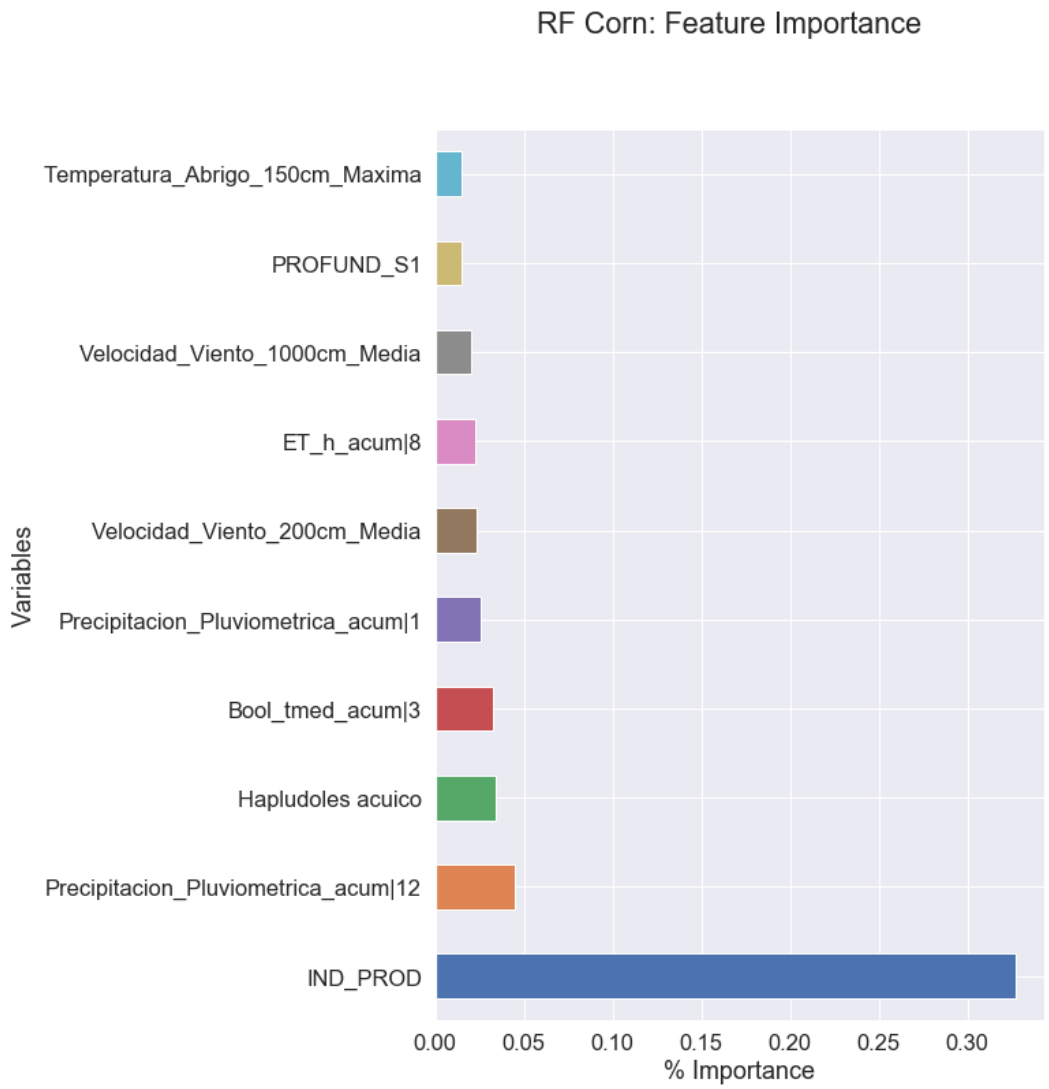


Fig. 13: Random Forest Regressor top 10 most important features for maize according to Feature Importance.

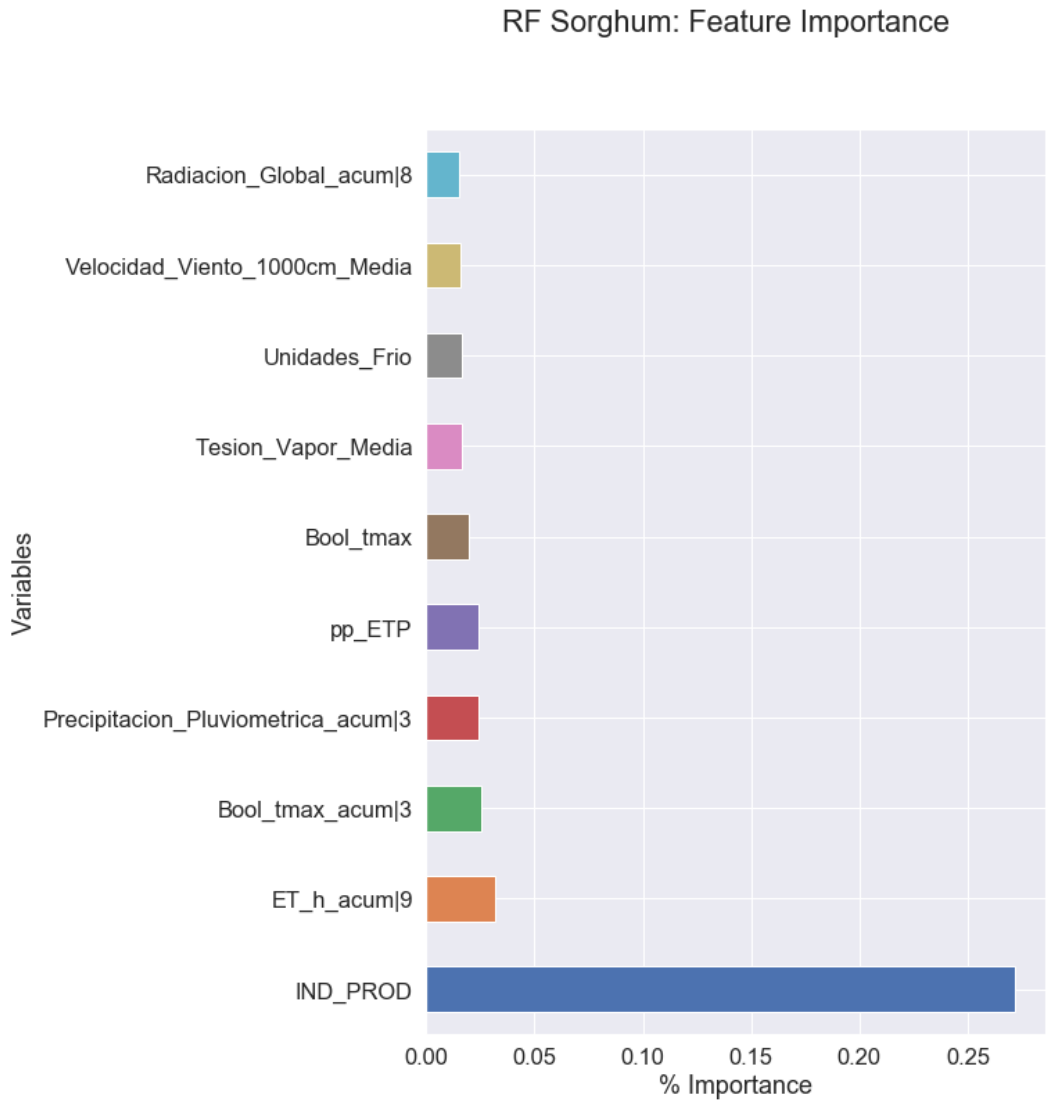


Fig. 14: Random Forest Regressor top 10 most important features for sorghum according to Feature Importance.

Although rice in Argentina is mostly irrigated, rainfall is still classified as an important variable for this crop. Heliophany (*Heliofania_Relativa_acum|[month]*) happens to be an important factor, together with mean temperature below 24 degrees (*Bool_tmed_acum|[month]*).

Cotton most important variables are associated to temperature. GDD, mean temperatures below 24 C, and accumulated days with maximum temperatures higher than 28 C (*Bool_tmax*). Wind speed has also considerable importance. As rice, is not as influenced by soil productivity as the other crops.

Peanut doesn't have a dominant feature (neither does rice and cotton), but all top 10 important variables are related to weather and not even one to soil (mean and maximum temperature, wind speed, heliophany and accumulated rainfall are some of the most important ones). Peanut production is located in the centre of Argentina (mainly in Cordoba province),

requires quite specific soil conditions and therefore is limited to homogeneous soil properties. Other important weather variables are accumulated days with temperatures below 0 C (*Bool_heladas_acum|[month]*) and sum of cold hours (*Horas_frio_acum|[month]*), which are similar variables.

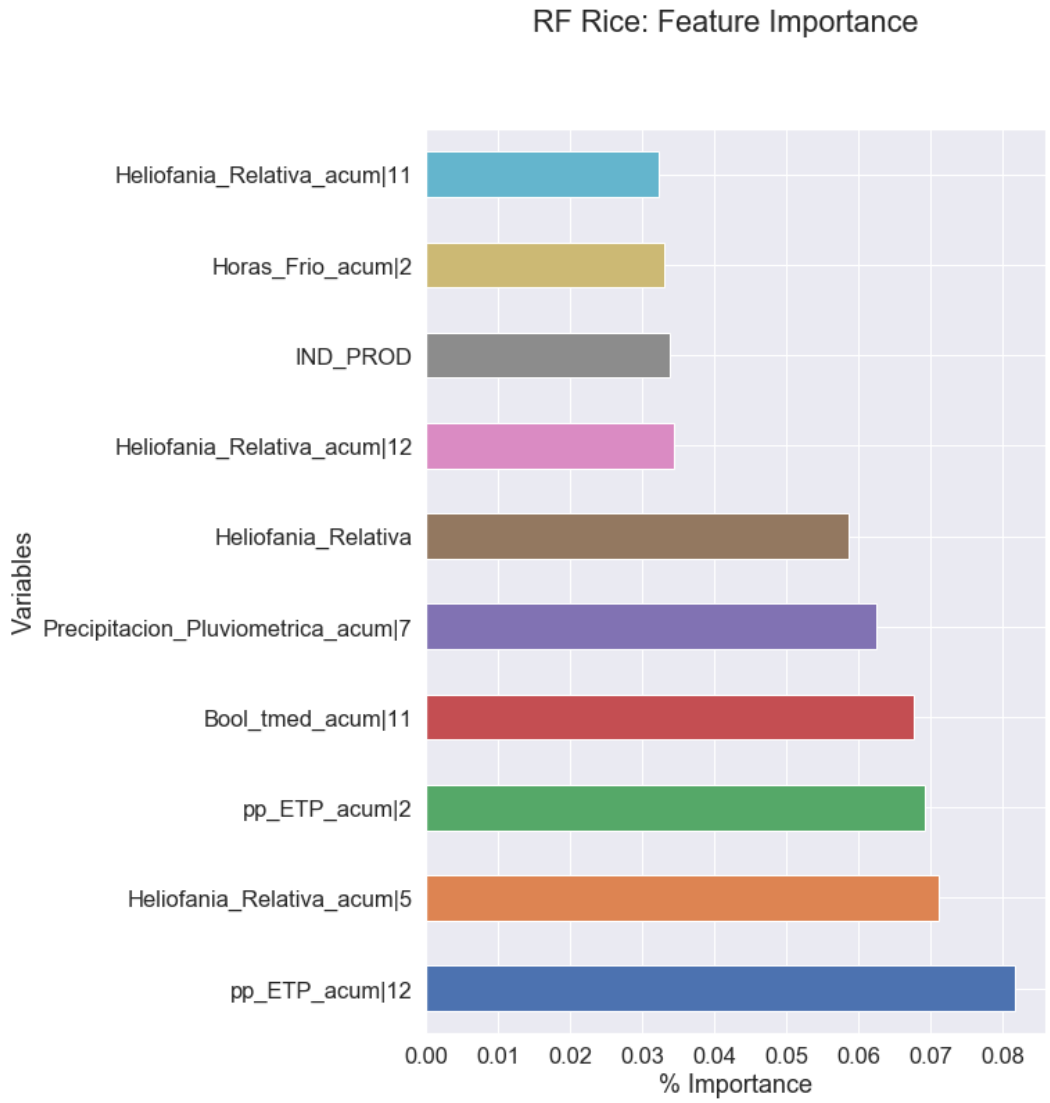


Fig. 15: Random Forest Regressor top 10 most important features for rice according to Feature Importance.

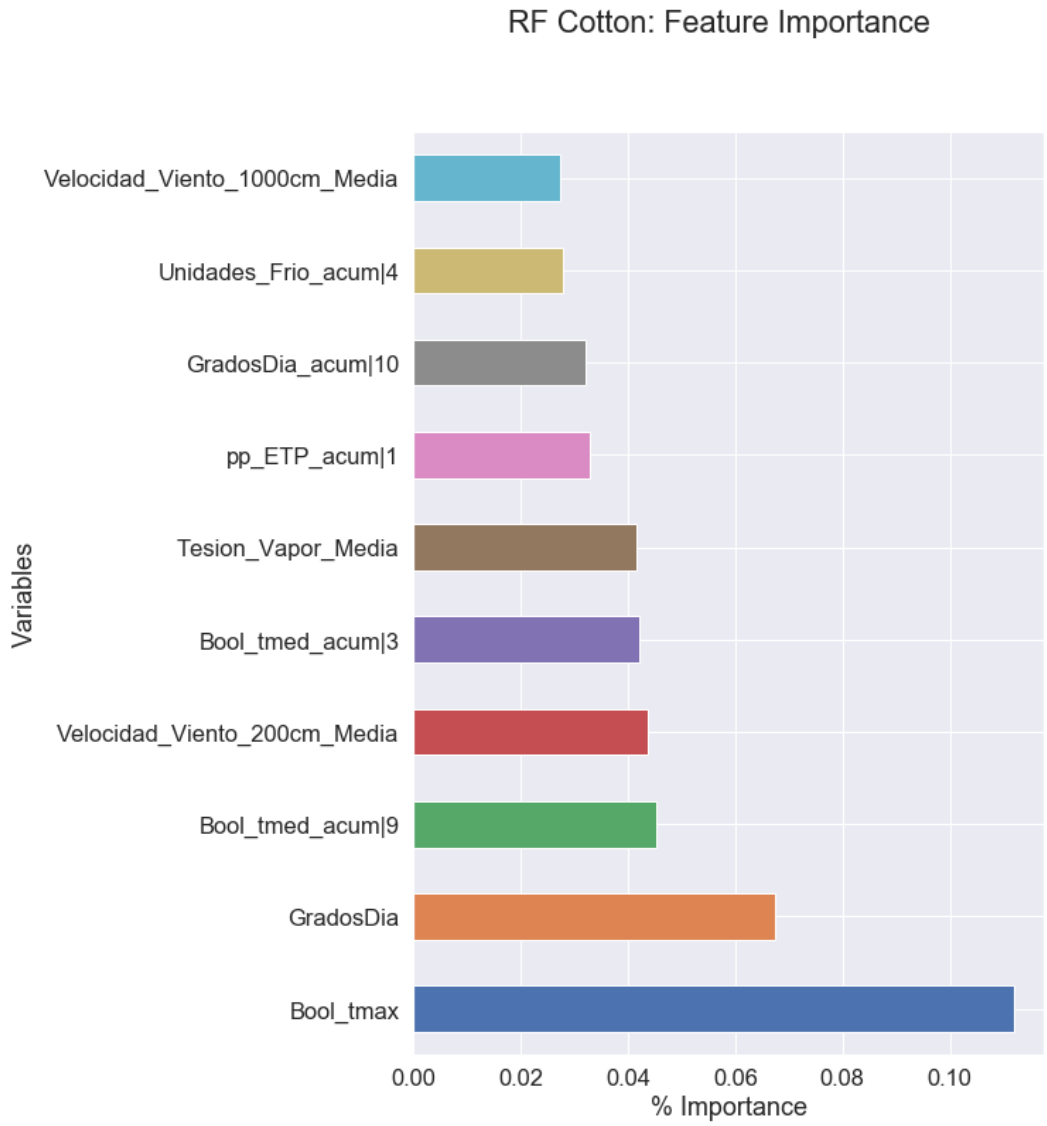


Fig. 16: Random Forest Regressor top 10 most important features for cotton according to Feature Importance.

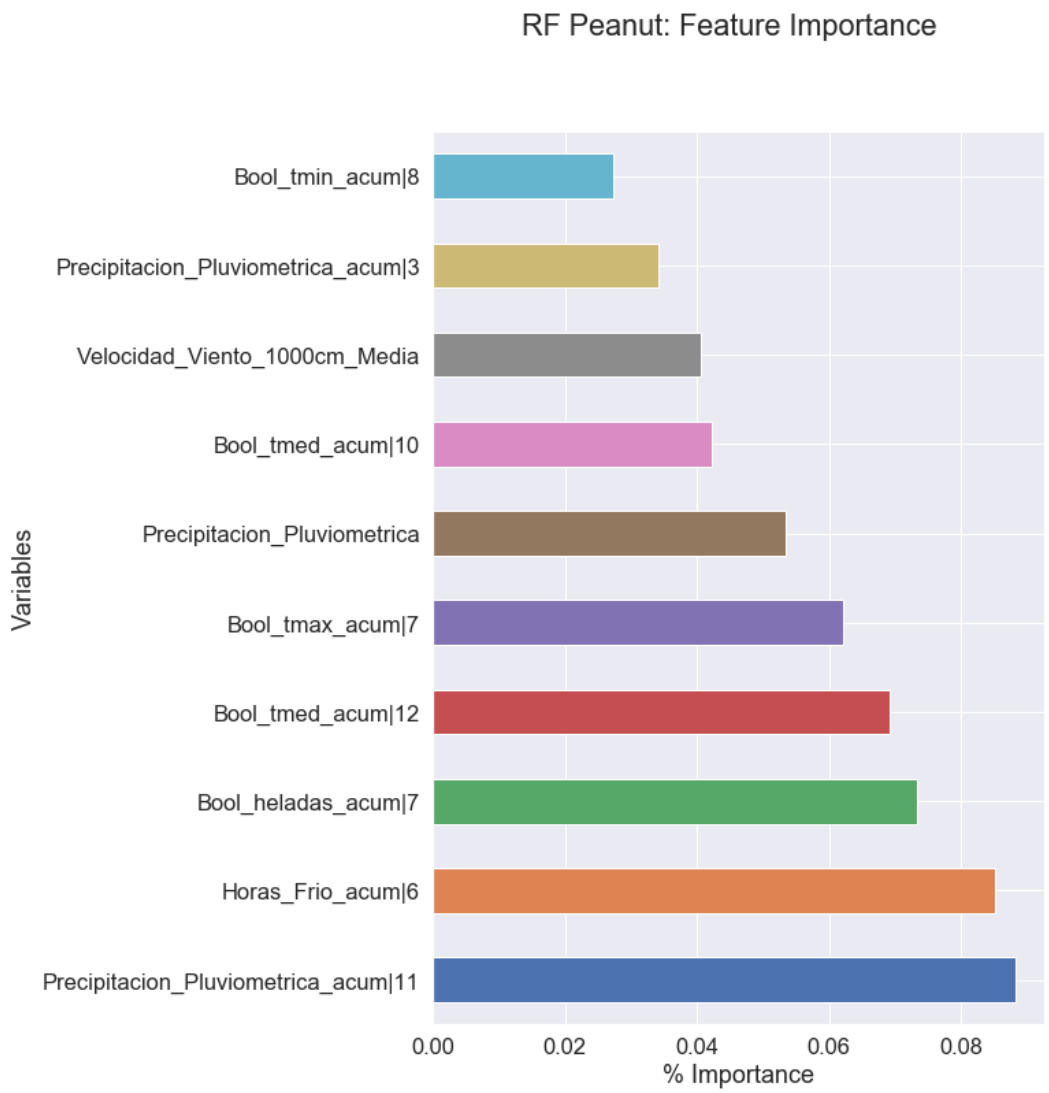


Fig. 17: Random Forest Regressor top 10 most important features for peanut according to Feature Importance.

After running a complex model with a large number of variables and analysing feature importance individually for each crop, 7 variables (4 soil features and 3 weather features) were selected to train a simpler model that will later be used to perform simulation scenario analysis. These are the variables selected to integrate a simplified model:

1. Soil productivity index (*IND_PROD*)
2. Soil deepness of sub superficial horizon (*PROFUND_S1*)
3. Proportion of Natracuoles tipico in soil (*Natracuoles_tipico*)
4. Proportion of Natracualfes tipico in soil (*Natracualfes_tipico*)
5. Rainfall accumulated from July until September (*Precipitacion Pluviometrica acum|9*)
6. Rainfall accumulated from July until January (*Precipitacion Pluviometrica acum|1*)

7. Rainfall accumulated from July until March (*Precipitacion Pluviometrica acum|3*)

II. HYPERPARAMETERS

Each crop was optimized for hyperparameters for Random Forrest regressor and LSTM using the 7 variables described before. The combination of hyperparameters that Random Search selected to reduce mean squared error (MSE) for each crop is listed below:

1. Cotton

Random Forest

```
{'oob_score': True,
 'n_estimators': 500,
 'min_weight_fraction_leaf':
         0,
  'min_samples_split': 5,
  'min_impurity_decrease':
        0.6,
   'max_leaf_nodes': 87,
    'max_features': 1,
     'max_depth': 2}
```

LSTM

```
{'epochs': 60,
 'neurons': 24,
 'batch_size': 105,
 'learning_rate':
    0.1,
 'loss': 'mae',
'optimizer': 'Adam'}
```

2. Rice

Random Forest

```
{'oob_score': False,
 'n_estimators': 300,
 'min_weight_fraction_leaf':
         0,
  'min_samples_split': 65,
  'min_impurity_decrease':
        0.9,
   'max_leaf_nodes': 52,
    'max_features': 1,
     'max_depth': 1}
```

LSTM

```
{'epochs': 20,
 'neurons': 39,
 'batch_size': 45,
 'learning_rate':
    0.1,
 'loss': 'mae',
'optimizer': 'Adam'}
```

3. Barley

Random Forest

```
{'oob_score': False,
 'n_estimators': 100,
 'min_weight_fraction_leaf':
         0,
  'min_samples_split': 20,
  'min_impurity_decrease':
        0.9,
   'max_leaf_nodes': 97,
    'max_features': 1,
     'max_depth': 2}
```

LSTM

```
{'epochs': 140,
 'neurons': 19,
 'batch_size': 5,
 'learning_rate':
    0.001,
 'loss': 'mae',
'optimizer': 'Adam'}
```

4. Sunflower

Random Forest

```
{'oob_score': True,
 'n_estimators': 300,
 'min_weight_fraction_leaf':
        0,
  'min_samples_split': 20,
  'min_impurity_decrease':
        0.25,
   'max_leaf_nodes': 47,
    'max_features': 2,
     'max_depth': 4}
```

LSTM

```
{'epochs': 100,
 'neurons': 43,
 'batch_size': 15,
'learning_rate': 0.1,
   'loss': 'mae',
'optimizer': 'Adam'}
```

5. Maize

Random Forest

```
{'oob_score': False,
 'n_estimators': 500,
 'min_weight_fraction_leaf':
        0,
  'min_samples_split': 85,
  'min_impurity_decrease':
        0.9,
   'max_leaf_nodes': 47,
    'max_features': 2,
     'max_depth': 5}
```

LSTM

```
{'epochs': 110,
 'neurons': 11,
 'batch_size': 75,
 'learning_rate':
    0.001,
  'loss': 'mae',
'optimizer': 'Adam'}
```

6. Peanut

Random Forest

```
{'oob_score': False,
 'n_estimators': 500,
 'min_weight_fraction_leaf':
        0,
  'min_samples_split': 20,
  'min_impurity_decrease':
        0.4,
   'max_leaf_nodes': 7,
    'max_features': 1,
     'max_depth': 1}
```

LSTM

```
{'epochs': 110,
 'neurons': 40,
 'batch_size': 35,
'learning_rate': 0.1,
   'loss': 'mae',
'optimizer': 'Adam'}
```

7. Soybean

Random Forest

```
{'oob_score': True,
 'n_estimators': 300,
 'min_weight_fraction_leaf':
0.2, 'min_samples_split': 5,
  'min_impurity_decrease':
        0.25,
   'max_leaf_nodes': 57,
    'max_features': 2,
```

LSTM

```
{'epochs': 90,
 'neurons': 4,
 'batch_size': 35,
 'learning_rate':
    0.001,
  'loss': 'mae',
'optimizer': 'Adam'}
```

```
'max_depth': 1}
```

8. Sorghum

Random Forest

```
{'oob_score': False,
 'n_estimators': 300,
 'min_weight_fraction_leaf':
        0,
 'min_samples_split': 20,
 'min_impurity_decrease':
        0.7,
 'max_leaf_nodes': 47,
 'max_features': 2,
 'max_depth': 5}
```

LSTM

```
{'epochs': 60,
 'neurons': 13,
 'batch_size': 25,
 'learning_rate':
        0.1,
 'loss': 'mae',
 'optimizer': 'Adam'}
```

9. Wheat

Random Forest

```
{'oob_score': True,
 'n_estimators': 500,
 'min_weight_fraction_leaf':
        0,
 'min_samples_split': 25,
 'min_impurity_decrease':
        0.6,
 'max_leaf_nodes': 22,
 'max_features': 1,
 'max_depth': 4}
```

LSTM

```
{'epochs': 60,
 'neurons': 24,
 'batch_size': 105,
 'learning_rate':
        0.1,
 'loss': 'mae',
 'optimizer':
 'Adam'}
```

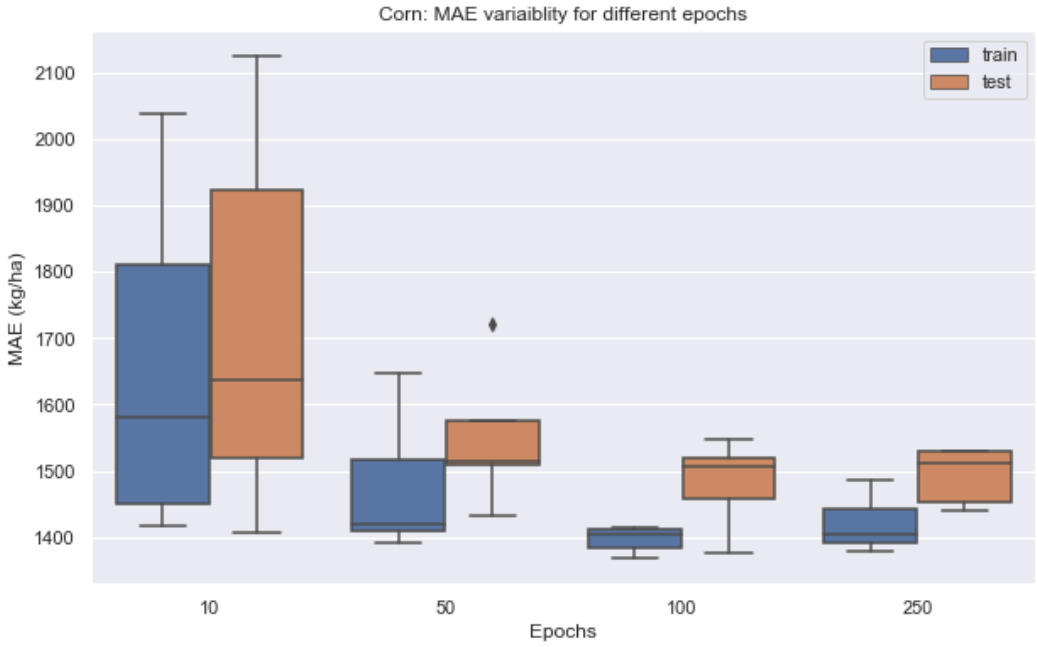


Fig. 18: Comparison of loss variability [MAE] in training and validation for LSTM Neural Network model in maize testing different number of epochs

III. MODEL PERFORMANCE: RANDOM FOREST AND LSTM NEURAL NETWORKS

The evaluation of model performance was conducted by calculating accuracy metrics for all 3 models in two stages: one for a complex model with almost all the soil and weather features, and another fit with the 7 selected variables.

Random Forest average improvement across all crops compared to Lasso is 32% in MAE and 35% in RMSE. Reduction in MAE score was significant across the 9 crops analysed (minimum improvement was observed for soybean, were RF outperformed Lasso by 20%). On average, LSTM performed 9% better than Lasso when measuring MAE and 18% better in RMSE. However, substantial variations were observed between crops: wheat LSTM model performed 20% worse than the baseline, while rice and cotton LSTM model outperformed baseline over 40%.

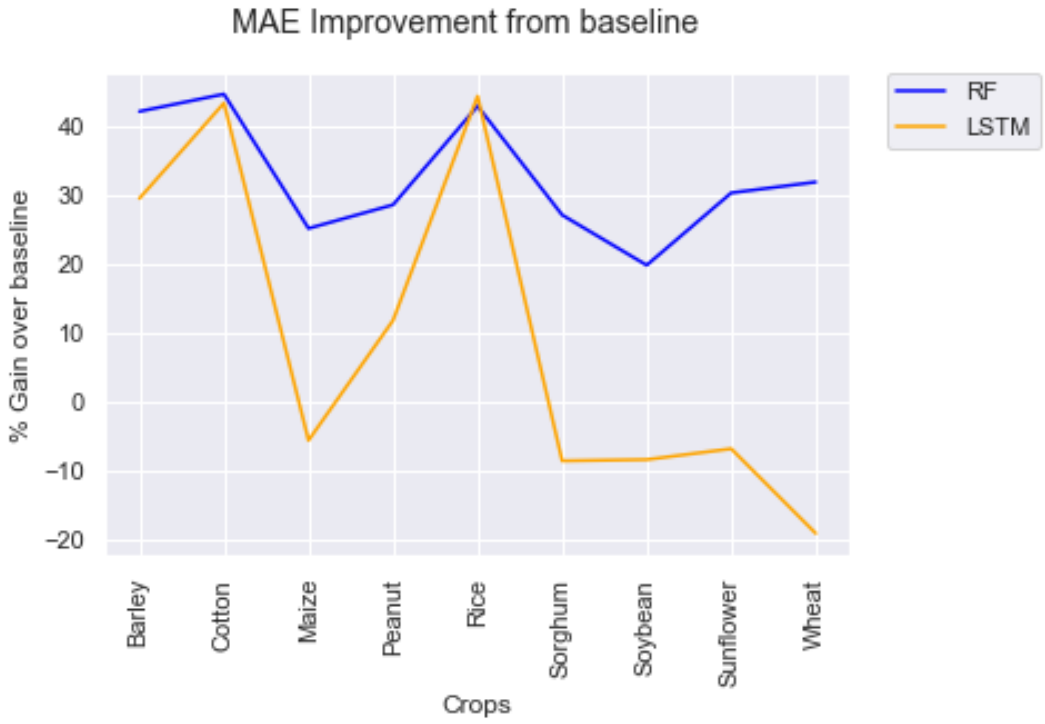


Fig. 19: Reduction in MAE in Random Forest and LSTM compared to Lasso's over the 9 crops analysed trained with all variables.

For a complex model with many features Random Forest outperformed Lasso and LSTM. Only in 2 crops and for specific scores LSTM got best scores by minimum difference.

Table 2: Model scores for 3 algorithms (LSTM, Random Forest and Lasso) trained with the complete set of variables. *Lowest* column indicate the algorithm that obtained the best score. *LSTM (%)* and *RF (%)* indicate each algorithms error reduction compared to Lasso Regressor (baseline).

MODEL		LSTM	RF	lasso	Lowest	LSTM (%)	RF (%)
CROP	Metric						
Barley	MAE	616.4	506.1	874.0	RF	29.47	42.09
	MAPE	37.2	14.2	25.4	RF	-46.46	44.09
	RMSE	830.3	727.0	1404.1	RF	40.87	48.22
Cotton	MAE	398.3	389.1	702.5	RF	43.30	44.61
	MAPE	25.1	21.1	36.2	RF	30.66	41.71
	RMSE	516.8	559.1	1423.2	LSTM	63.69	60.72
Maize	MAE	1400.4	992.7	1325.7	RF	-5.63	25.12
	MAPE	60.6	19.2	25.5	RF	-137.65	24.71
	RMSE	1833.3	1396.2	1854.4	RF	1.14	24.71
Peanut	MAE	549.5	445.1	622.9	RF	11.78	28.54
	MAPE	41.6	36.9	39.2	RF	-6.12	5.87
	RMSE	761.7	715.2	1125.3	RF	32.31	36.44
Rice	MAE	1018.8	1045.2	1829.5	LSTM	44.31	42.87
	MAPE	14.8	10.7	20.8	RF	28.85	48.56
	RMSE	1911.5	1564.2	3241.6	RF	41.03	51.75
Sorghum	MAE	1254.4	842.3	1154.9	RF	-8.62	27.07
	MAPE	41.6	14.9	20.5	RF	-102.93	27.32
	RMSE	1581.2	1257.8	1502.1	RF	-5.27	16.26
Soybean	MAE	658.4	487.1	607.3	RF	-8.41	19.79
	MAPE	43.0	20.1	23.1	RF	-86.15	12.99
	RMSE	879.6	708.1	865.4	RF	-1.64	18.18
Sunflower	MAE	525.5	342.9	491.8	RF	-6.85	30.28
	MAPE	32.2	15.1	22.2	RF	-45.05	31.98
	RMSE	693.5	509.9	687.7	RF	-0.84	25.85
Wheat	MAE	721.6	412.8	605.7	RF	-19.13	31.85
	MAPE	43.0	13.4	19.8	RF	-117.17	32.32
	RMSE	939.3	603.2	871.4	RF	-7.79	30.78

For the simple model with 7 variables, Random Forest got the best performance for all crops and scores analysed except for cotton, were LSTM performed slightly better. RF improved MAE on average 35% compared to Lasso while LSTM performance compared to Lasso averaged 7% for the same metric. RF had lower standard deviation between crops performance than LSTM.

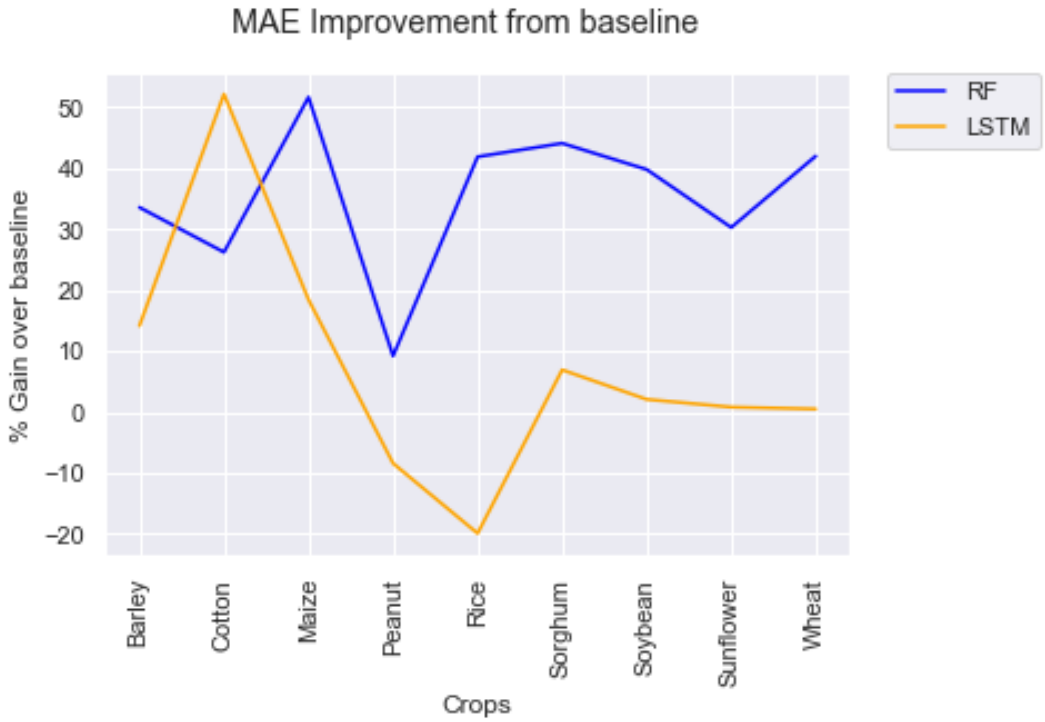


Fig. 20: Reduction in MAE in Random Forest and LSTM compared to Lasso's over the 9 crops analysed trained with simplified model (7 variables).

Table 3: Model scores for 3 algorithms (LSTM, Random Forest and Lasso) trained with reduced number of variables. *Lowest* column indicate the algorithm that obtained the best score. *LSTM (%)* and *RF (%)* indicate each algorithms error reduction compared to Lasso Regressor (baseline).

	MODEL	LSTM	RF	lasso	Lowest	LSTM (%)	RF (%)
CROP	Metric						
Barley	MAE	757.5	586.0	882.2	RF	14.14	33.58
	MAPE	33.3	20.5	31.9	RF	-4.39	35.74
	RMSE	970.2	812.3	1072.4	RF	9.53	24.25
Cotton	MAE	329.0	507.2	687.4	LSTM	52.14	26.21
	MAPE	20.2	23.4	31.8	LSTM	36.48	26.42
	RMSE	426.1	702.6	818.7	LSTM	47.95	14.18
Maize	MAE	1559.4	924.4	1911.6	RF	18.42	51.64
	MAPE	54.8	15.7	37.1	RF	-47.71	57.68
	RMSE	2009.5	1406.0	2407.5	RF	16.53	41.60
Peanut	MAE	753.6	631.2	695.1	RF	-8.42	9.19
	MAPE	44.2	23.8	25.6	RF	-72.66	7.03
	RMSE	972.9	958.3	874.6	lasso	-11.24	-9.57
Rice	MAE	1643.3	796.8	1370.4	RF	-19.91	41.86
	MAPE	17.2	9.0	16.0	RF	-7.50	43.75
	RMSE	2614.2	1124.6	1680.5	RF	-55.56	33.08
Sorghum	MAE	1509.5	906.8	1621.5	RF	6.91	44.08
	MAPE	42.4	17.5	31.2	RF	-35.90	43.91
	RMSE	1857.4	1432.7	2050.4	RF	9.41	30.13
Soybean	MAE	738.7	454.2	754.3	RF	2.07	39.79
	MAPE	41.0	17.2	29.3	RF	-39.93	41.30
	RMSE	992.3	685.4	925.6	RF	-7.21	25.95
Sunflower	MAE	528.4	371.4	532.6	RF	0.79	30.27
	MAPE	28.5	14.4	21.4	RF	-33.18	32.71
	RMSE	702.4	541.1	705.4	RF	0.43	23.29
Wheat	MAE	851.3	496.3	855.5	RF	0.49	41.99
	MAPE	35.1	15.7	29.6	RF	-18.58	46.96
	RMSE	1098.2	743.8	1060.5	RF	-3.55	29.86

Comparisons of performance scores between training and test dataset allowed to detect when a model is overfitting. Although Random forest outperformed LSTM and Lasso, it overfitted the training dataset in each crop model.

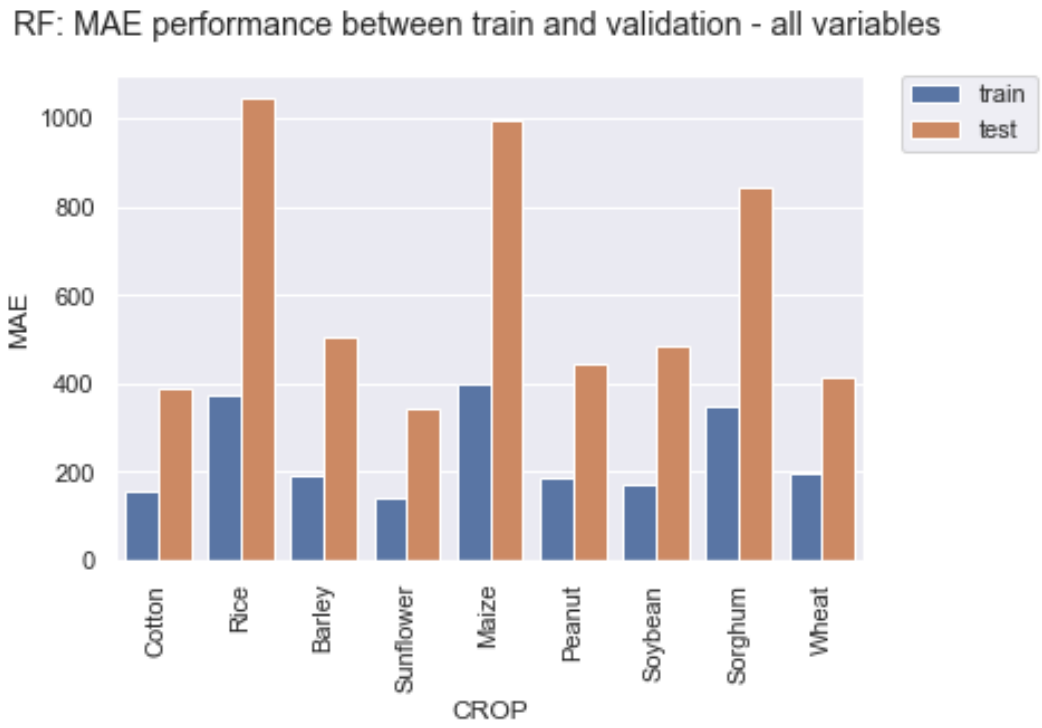


Fig. 21: Random forest overfitting for MAE score. Complex model (all variables).

LSTM, on the other hand, with less attractive accuracy metrics, had a very similar scores for training and test set.

Fig. 22: LSTM comparison of training and test dataset performance (MAE). Complex model (all variables). Results show a very close relationship between test and training set accuracies, with no signs of overfitting.

IV. MODEL PERFORMANCE: COMPLEX AND SIMPLE MODEL

Reducing complexity of the model had different impact depending on the crop. For the summer crops were RF originally performed well (maize, soybean, sorghum and sunflower) the reduction of features had softer effects. These could be related to the fact that the variables chosen for the simplified model are quite representative of crop yield variability. Wheat and barley, the winter crops, deteriorated MAE by 20% for both LSTM and RF. As was mentioned when discussing feature importance, these two crops depended very much on specific soil features that were removed in the simplified model, plus the fact that water availability is less critical compared to summer crops. Peanut model was severely affected by feature reduction on both models. None of the top 10 most relevant variables (from RF feature importance) for peanut were kept in the simplified model. Finally, rice and cotton performance changed opposite from each other after feature reduction. While RF decreased performance in cotton, LSTM increased, and while RF improved performance in rice, LSTM performance deteriorated.

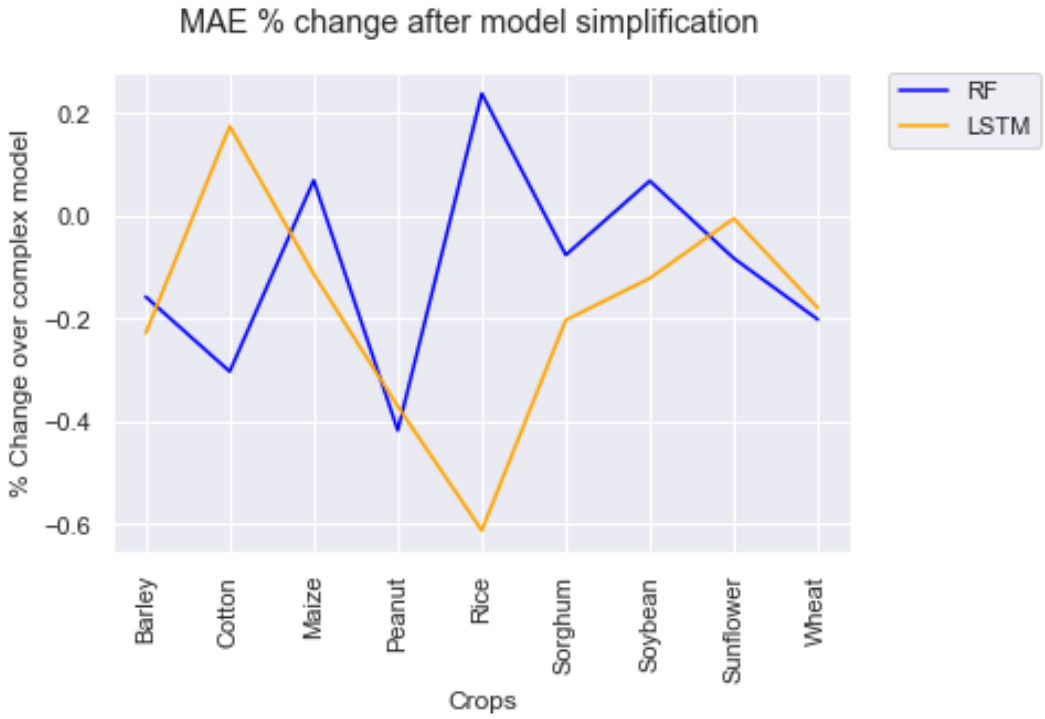


Fig. 23: Change in MAE score after reducing the number of variables from a complex model to a simple model with 7 variables.

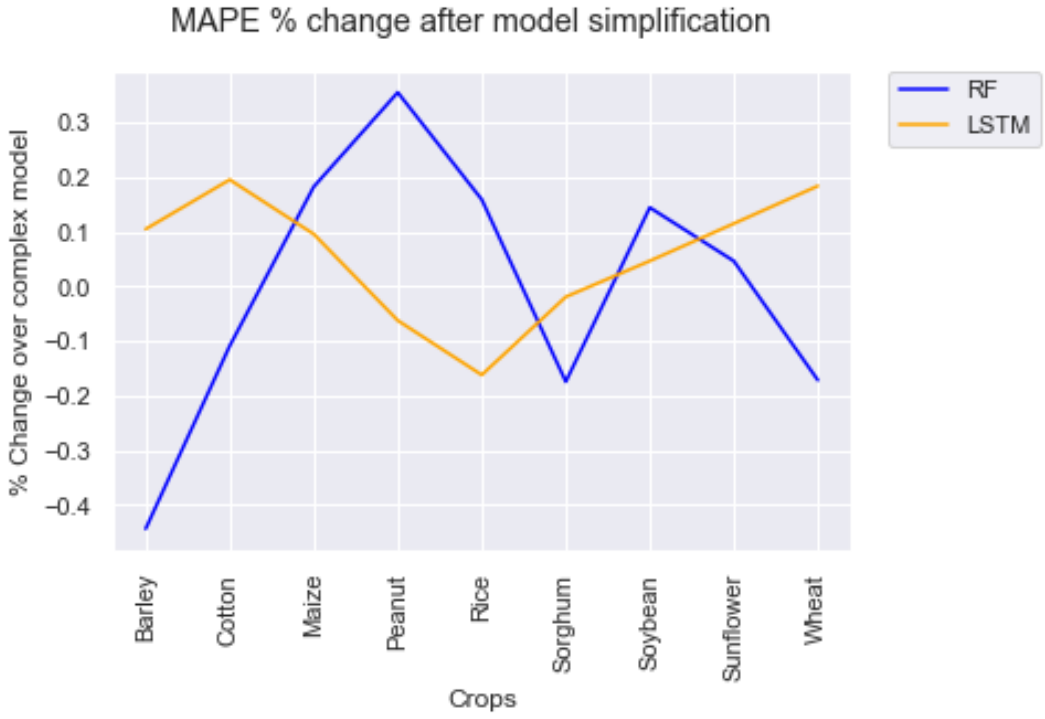


Fig. 24: Change in MAPE score after reducing the number of variables from a complex model to a simple model with 7 variables.

Coefficient of determination R^2 experimented very small reductions after reducing the number of dimensions in the model. Except for cotton and peanut, which coefficients of determination resulted very damaged after model simplification.

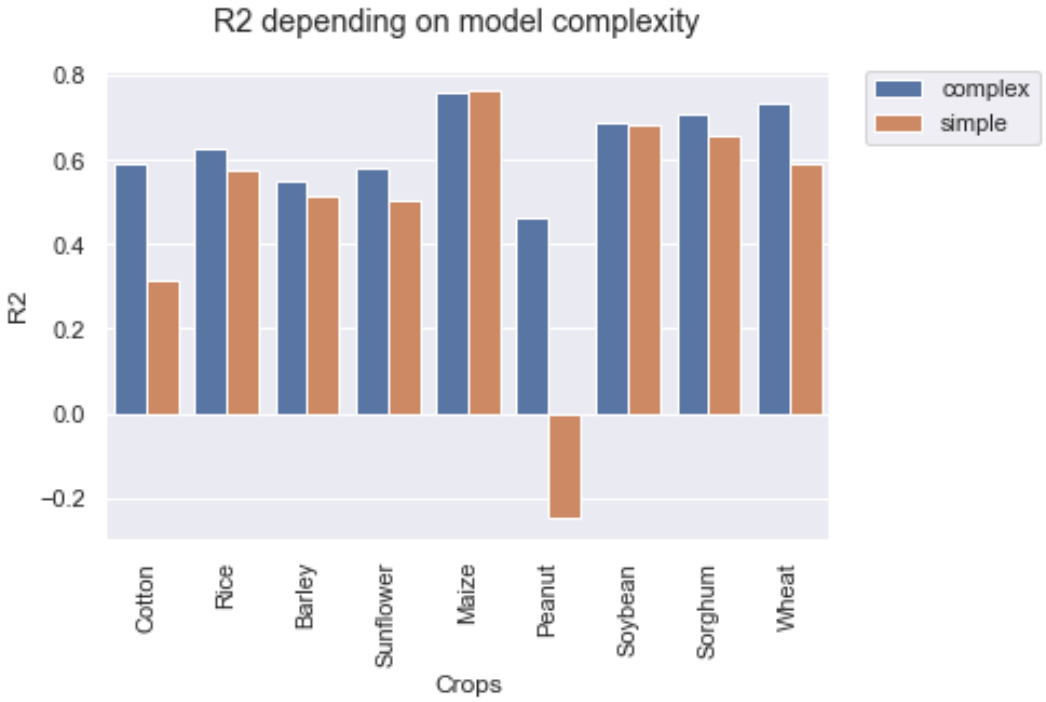


Fig. 25: Comparison of Coefficient of Determination R^2 for Random Forest Regressor between a complex model and simple models with 7 variables.

V. MODEL PERFORMANCE: CROPS

Because crops have different scale of yields, and the models learned with the original scale of values (no normalization was effectively implemented), it is not convenient to use MAE score to compare the accuracy of the models between crops. Maize for example, can yield more than 12 tons/ha (12,000 kg/ha), whereas sunflower average yields could be around 3,5 tons/ha in some regions.

MAPE (Mean Average Percentage Error) is used to compare model accuracy between crops. It gives a relative measure of the typical error (Zheng 2015) and solves the scaling problem. Coefficient of determination R^2 is also analysed as an indicator of how good the trained model against a naïve mean model is.

Table 4: Evaluation of Random Forest performance scores for 9 crops between training and test dataset. Left: Complex model, with all soil and weather variables. Right: Simplified model

CROP	SET	R2	MAPE	MAE
Cotton	train	0.89	7.47	156.49
	test	0.59	21.08	389.09
Rice	train	0.88	4.16	375.54
	test	0.62	10.70	1045.16
Barley	train	0.93	6.19	194.26
	test	0.55	14.17	506.09
Sunflower	train	0.92	6.07	141.24
	test	0.58	15.10	342.89
Maize	train	0.96	6.86	397.13
	test	0.76	19.15	992.67
Peanut	train	0.92	9.90	184.92
	test	0.46	36.94	445.09
Soybean	train	0.95	6.27	172.81
	test	0.69	20.10	487.13
Sorghum	train	0.94	6.61	350.13
	test	0.71	14.87	842.28
Wheat	train	0.94	5.97	198.62
	test	0.73	13.38	412.80

CROP	SET	R2	MAPE	MAE
Cotton	train	0.88	7.39	164.58
	test	0.31	23.43	507.24
Rice	train	0.92	4.22	391.55
	test	0.58	8.98	796.75
Barley	train	0.88	7.63	237.89
	test	0.51	20.48	586.05
Sunflower	train	0.90	6.86	151.09
	test	0.51	14.39	371.39
Maize	train	0.94	7.52	439.31
	test	0.76	15.67	924.45
Peanut	train	0.94	10.06	168.10
	test	-0.25	23.85	631.16
Soybean	train	0.93	7.68	192.59
	test	0.68	17.25	454.23
Sorghum	train	0.94	6.74	366.48
	test	0.66	17.47	906.81
Wheat	train	0.92	6.94	220.36
	test	0.59	15.72	496.32

Looking at Random Forest R^2 and MAPE, maize, soybean and wheat models performed very well on both metrics. Barley and sunflower, with lower R^2 performed reasonably well on MAPE.

As was mentioned before, significant difference between training and test performance was observed for Random forest which indicates that the model overfitted the training data.

Predictions where made on the test dataset and regression analysis was conducted between predicted and actual yields for each crop. The difference between prediction and actual values was normalized and *z scores* were calculated as following:

$$z = \frac{(error - \mu)}{\sigma}$$

Looking at the performance scores obtained for the test dataset, results look very promising for 6 of the 9 crops analysed: barley, wheat, maize, sorghum, soybean and sunflower.

Correlation between actual and prediction yields was statistically significant for all 6 crops mentioned above, with very acceptable Pearson coefficient ($r \geq 0.75$ except for sunflower). This are the main crops in terms of national production and geographical extension compared to the remaining 3 crops which didn't show good results in the predicted vs. actual analysis.

The amount of data available for training and the representativeness of the training set was higher for the 6 crops with best scores compared to the 3 scores that performed poorly.

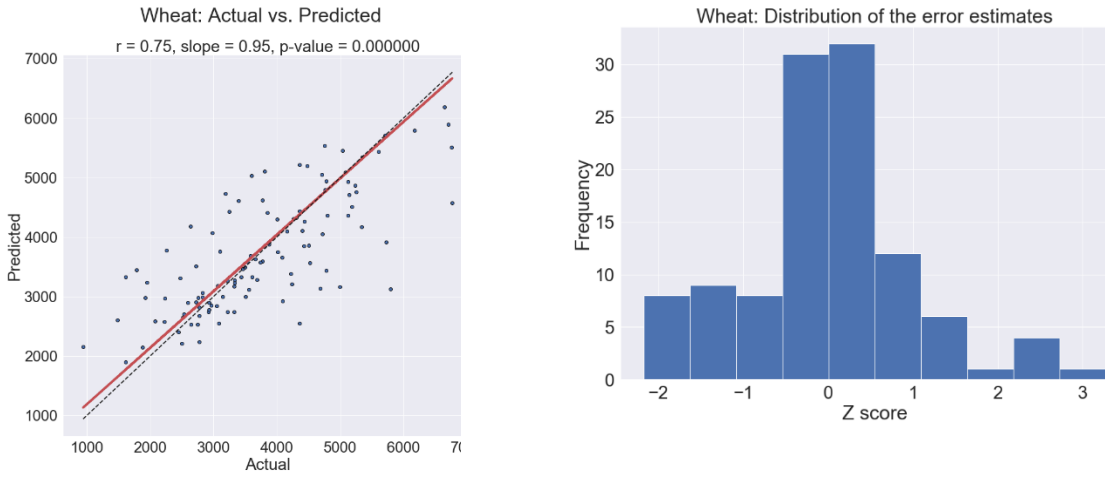


Fig. 26: Wheat prediction performance analysis with Random Forest simplified model. Left: Scatter plot of actual vs. predicted yields. Right: Distribution of the normalized error, calculated as the difference between actual and predicted yield, normalized to the mean.

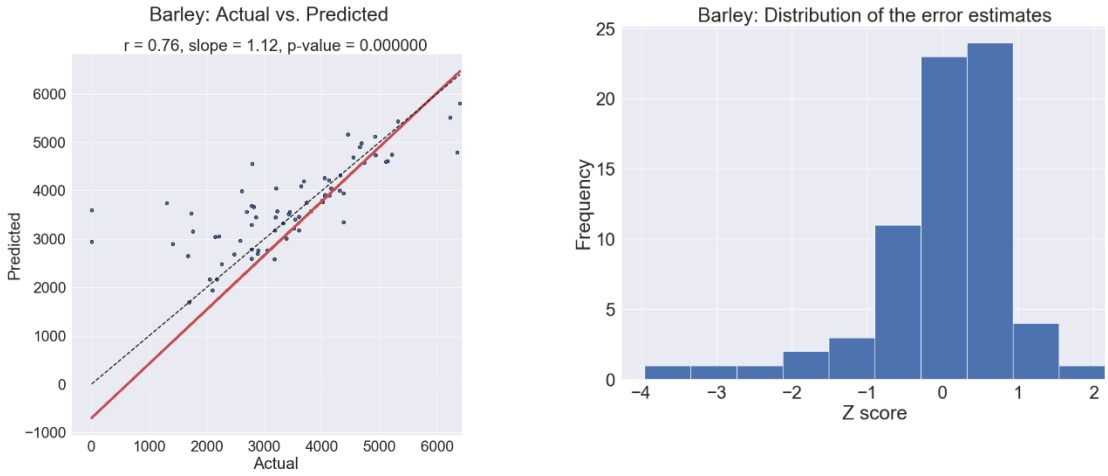


Fig. 27 Barley prediction performance analysis with Random Forest simplified model.

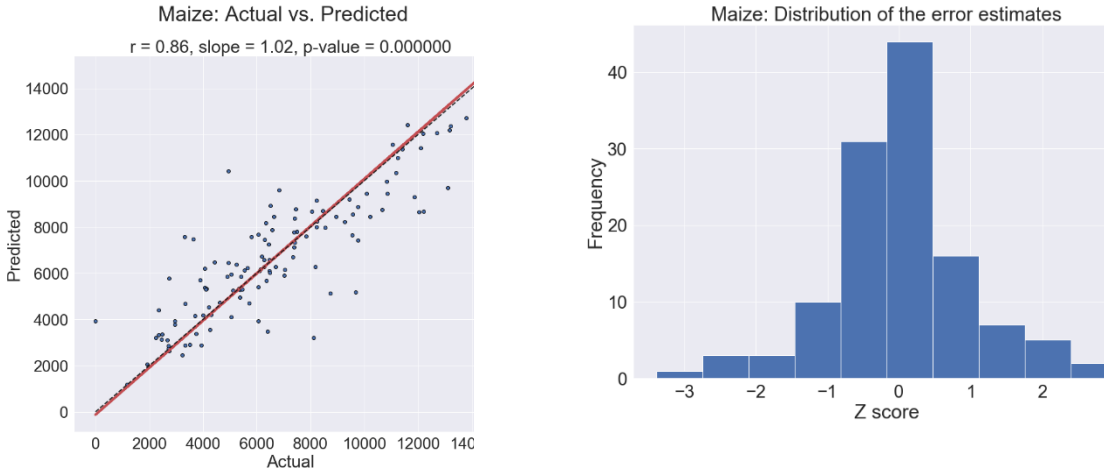


Fig. 28: Maize prediction performance analysis with Random Forest simplified model.

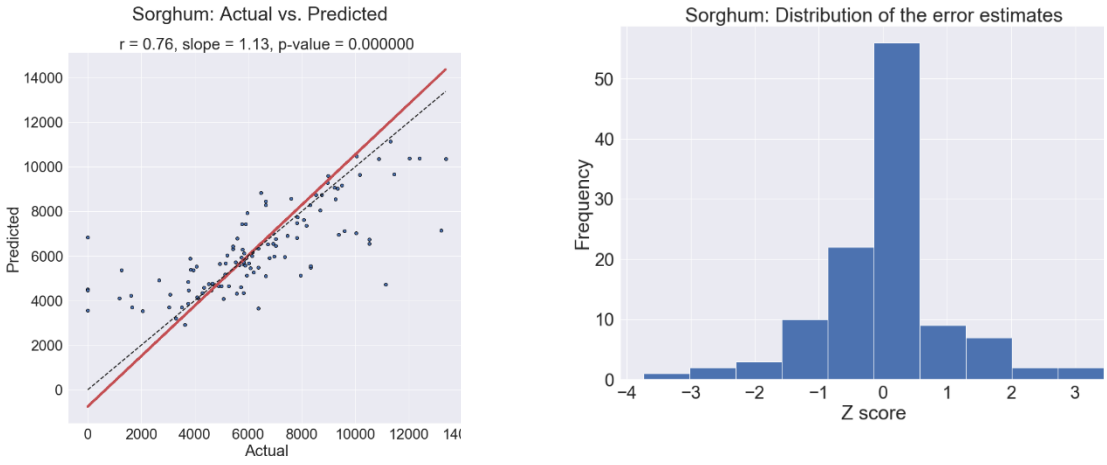


Fig. 29: Sorghum prediction performance analysis with Random Forest simplified model.

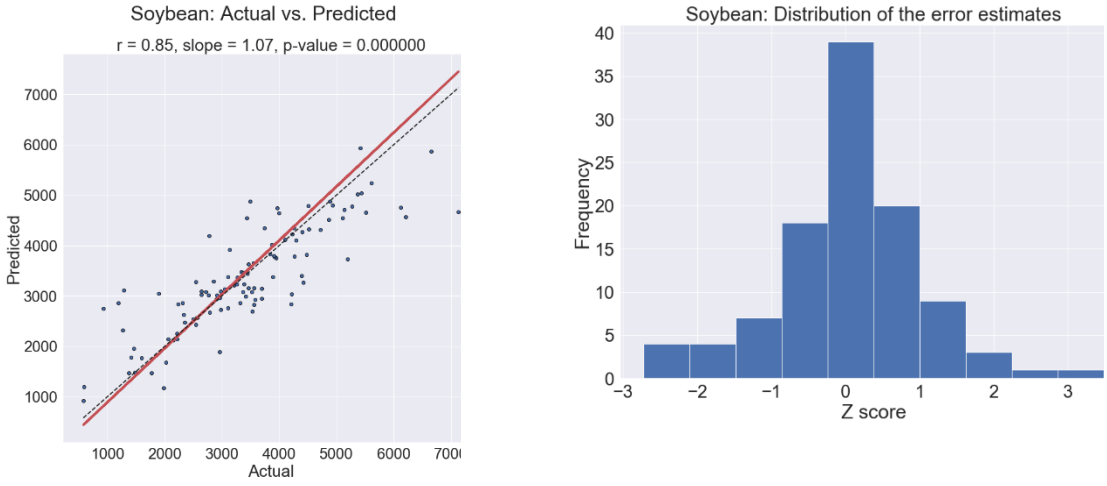


Fig. 30: Soybean prediction performance analysis with Random Forest simplified model.

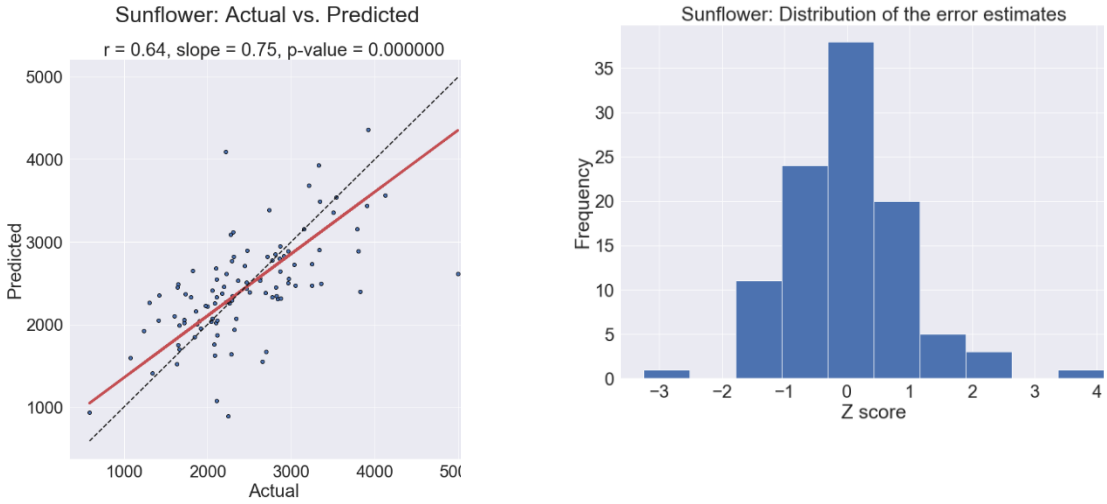


Fig. 31: Sunflower prediction performance analysis with Random Forest simplified model.

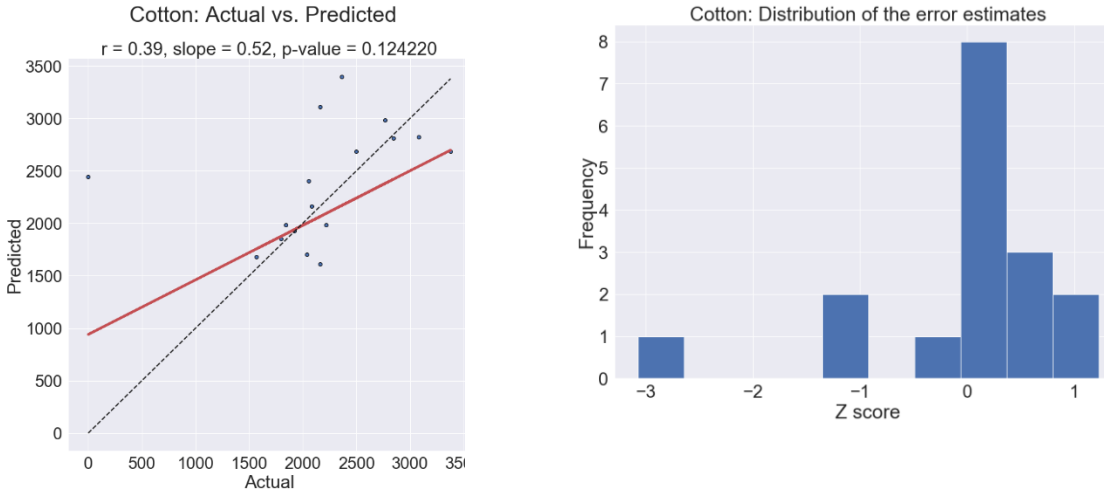


Fig. 32: Cotton prediction performance analysis with Random Forest simplified model.

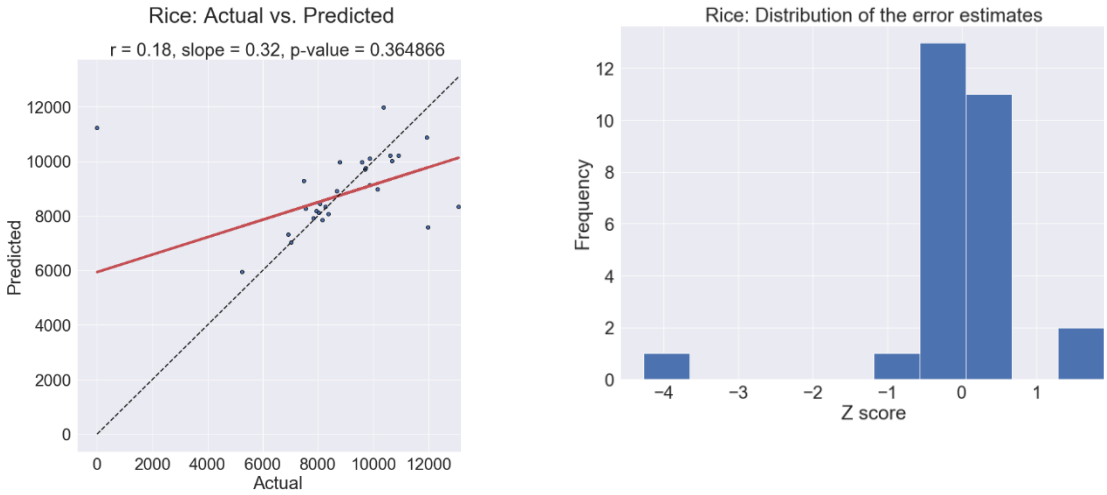


Fig. 33: Rice prediction performance analysis with Random Forest simplified model.

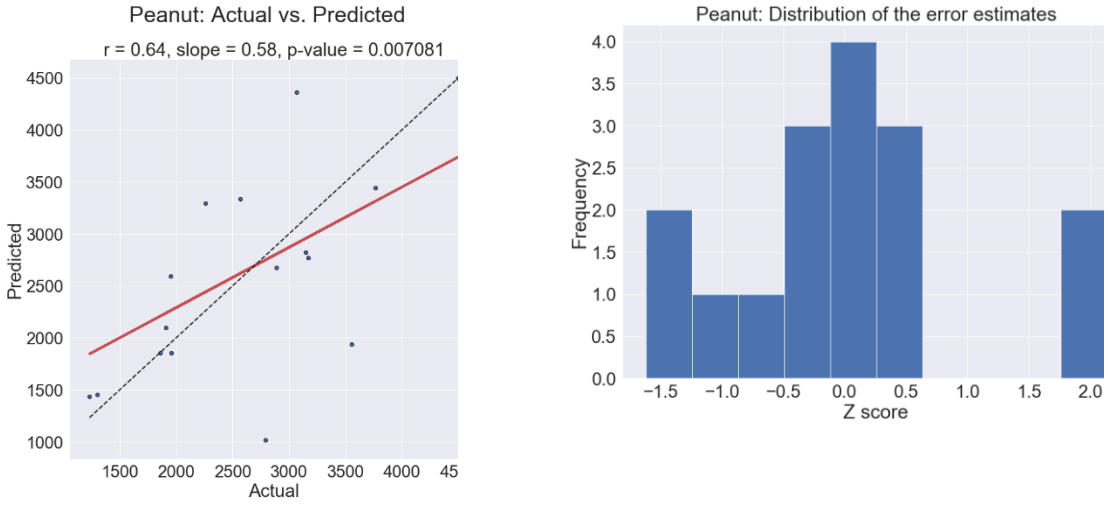


Fig. 34: Peanut prediction performance analysis with Random Forest simplified model.

SIMULATIONS

Step by step monthly rainfall simulations for each crop in each region resulted in seasonal MAE curves that express level of confidence of simulations compared to the prediction using real data. Although crop season changes regionally, this model is a generalization for the national crop season.

Maize season in Argentina usually goes from September-October (planting) until May-June of the following year (harvest) (INTA 2016). Accuracy of the model was good with 100% real data (no simulations), MAPE was just below 15%. Simulations from January, were there is still 4-5 months of the season to go, resulted in a qualified model (MAPE ≤ 20).

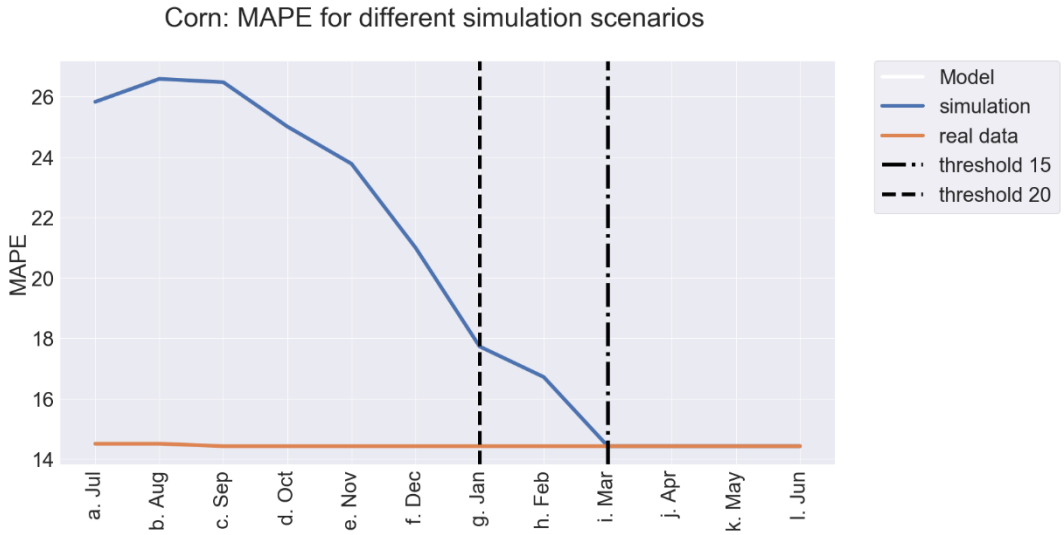


Fig. 35: Reduction in MAPE score for maize RF model across different simulation scenarios, starting with 100% simulated data (July) up to 100% real data

(June). Vertical dashes indicate the moment when a threshold is achieved, T15: MAPE ≤ 15%, T20: MAPE ≤ 20%.

Soybean (as first crop and not as double crop after a winter crop) season is very similar to maize's, although could have later sowing (October-November) and earlier harvest (April-May) (INTA 2016). Soybean model accuracy was even better than maize'. In December 3 to 4 months after planting, a forecast can be made with a qualified accuracy, meanwhile a good accuracy is achieved in February.

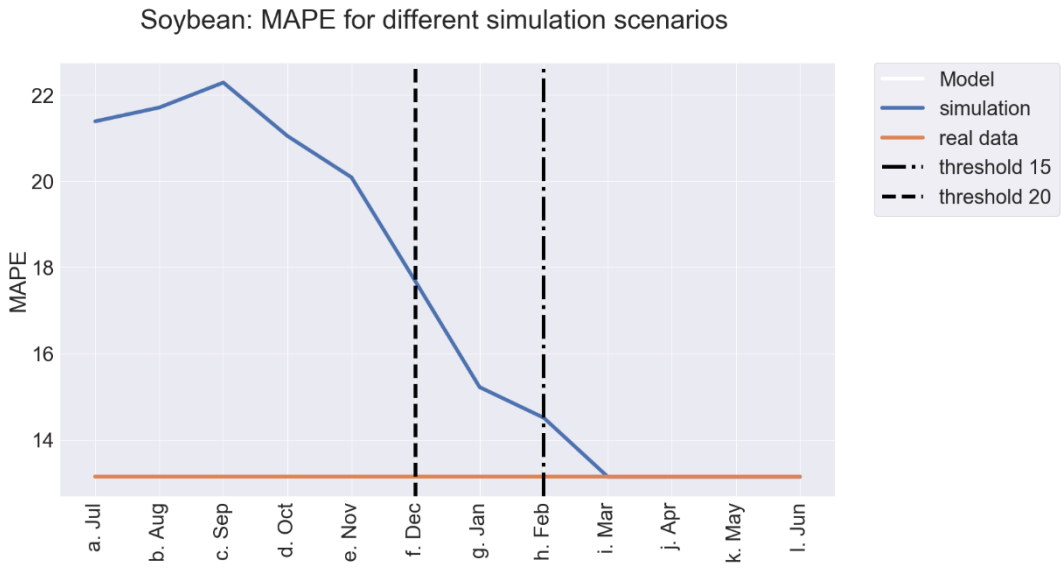


Fig. 36: Reduction in MAPE score for soybean RF model across different simulation scenarios, starting with 100% simulated data (July) up to 100% real data (June). Vertical dashes indicate the moment when a threshold is achieved, T15: MAPE ≤ 15%, T20: MAPE ≤ 20%.

Sunflower can be planted between August until November and harvested in February-March (INTA 2016). The accuracy obtained with simulations allows to forecast sunflower yields as from October with a MAPE ≤ 20.

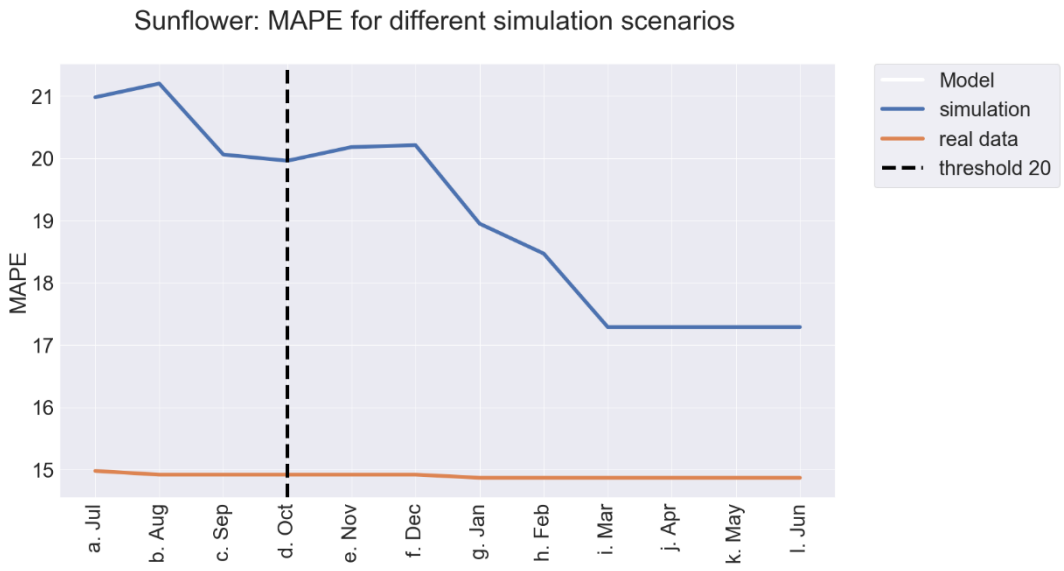


Fig. 37: Reduction in MAPE score for sunflower RF model across different simulation scenarios, starting with 100% simulated data (July) up to 100% real data (June). Vertical dashes indicate the moment when a threshold is achieved, T15: MAPE ≤ 15%, T20: MAPE ≤ 20%.

Sorghum growing season is very similar to soybean's, sowing happens around October-November and harvest in April-May (INTA 2016). Results show that with 100% data in July a Threshold 20 is achieved, and in January, simulating 42% of the season, an accuracy of MAPE ≤ 15 is obtained.

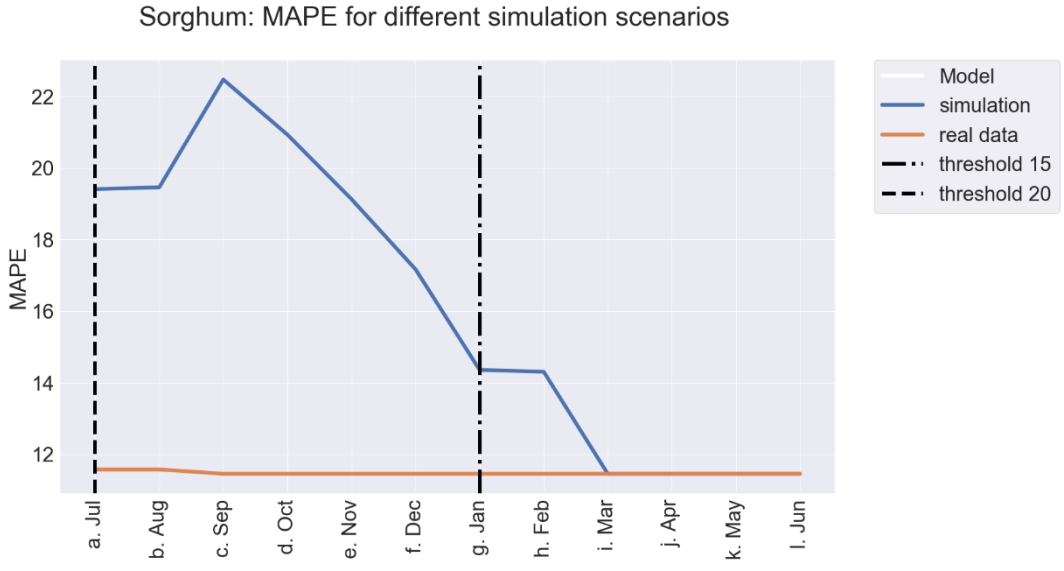


Fig. 38: Reduction in MAPE score for sorghum RF model across different simulation scenarios, starting with 100% simulated data (July) up to 100% real data (June). Vertical dashes indicate the moment when a threshold is achieved, T15: MAPE ≤ 15%, T20: MAPE ≤ 20%.

Rice is majorly planted in October and harvested around February-March (INTA 2016). In November, only a month after planting, a MAE ≤ 15 is achieved by simulating the rest of the season.

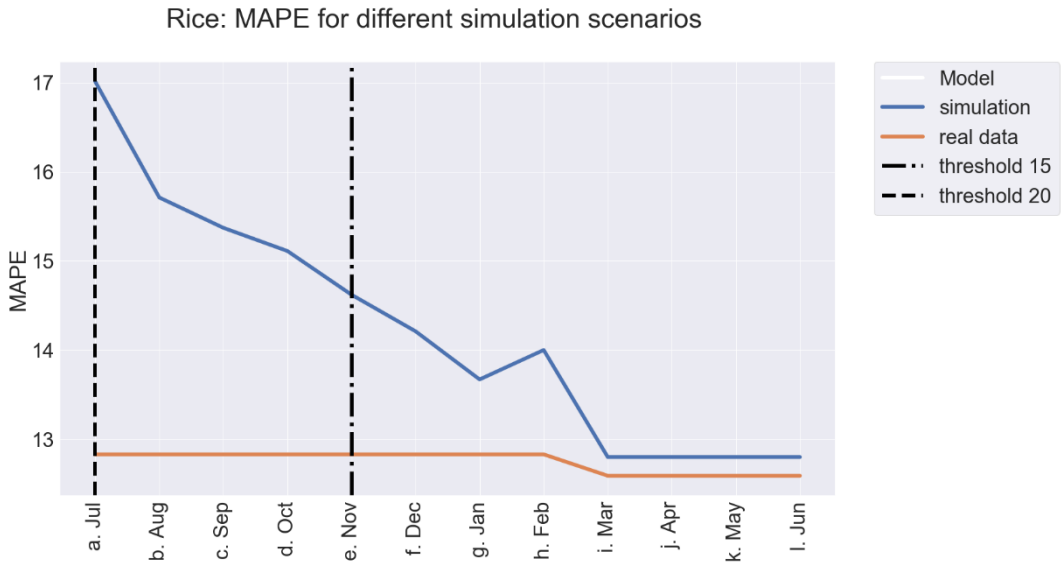


Fig. 39: Reduction in MAPE score for rice RF model across different simulation scenarios, starting with 100% simulated data (July) up to 100% real data (June). Vertical dashes indicate the moment when a threshold is achieved, T15: MAPE ≤ 15%, T20: MAPE ≤ 20%.

Cotton is planted between October and November and harvested between March and May (INTA 2016). The accepted accuracy level for forecasting cotton yield arrive to late in the season to be beneficial. Only in March the model can predict with an error less than 20%, and by that time the crop is about to be harvested.

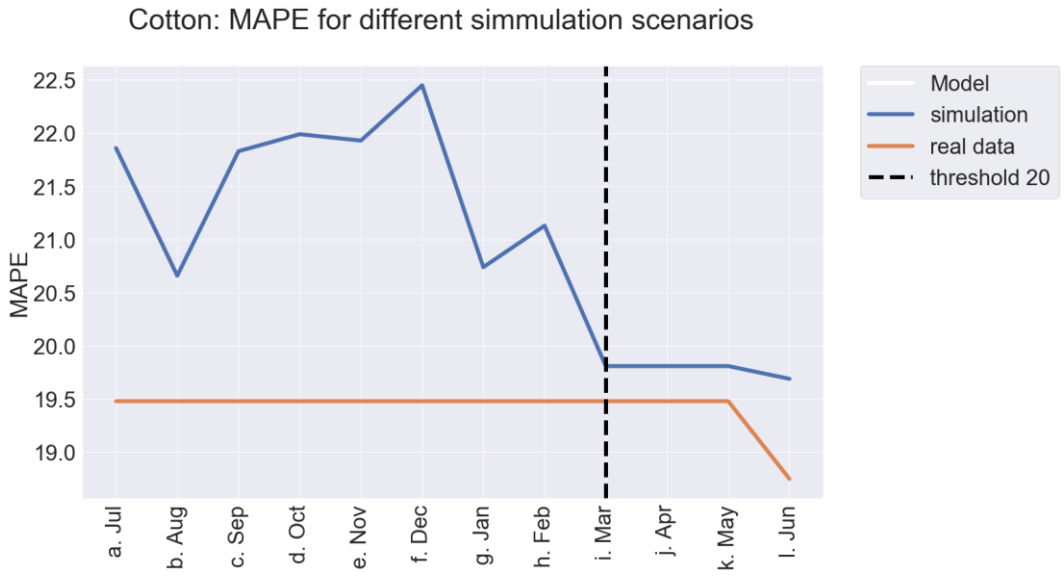


Fig. 40: Reduction in MAPE score for cotton RF model across different simulation scenarios, starting with 100% simulated data (July) up to 100% real data

(June). Vertical dashes indicate the moment when a threshold is achieved, T15: MAPE ≤ 15%, T20: MAPE ≤ 20%.

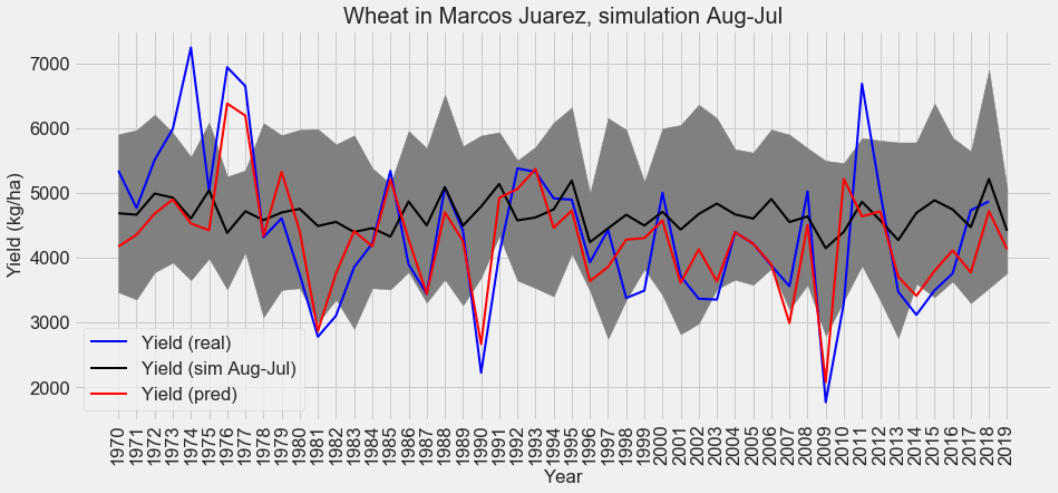
Because peanut MAPE values using real data were already higher than the thresholds stablished, monthly scenario analysis was not analysed for this crop. On the other hand, winter crops (wheat and barley) simulations can't be model in the same way that summer crops due to the differences in growing season and how this study defined a common denominator for crop seasonality. However, a time-series analysis was performed combining real data with predicted data using different levels of simulations.

Picking up data for 1 region, 3 instances where modelled:

1. Real data from July, simulations from August.
2. Real data from July until September, simulations from October.
3. Real data from July until November, simulations from December

For each stage, 100 simulations were generated calculating the average and quartiles of the distribution of the predictions and the results were compared against the actual data and against the predictions made by 100% real data.

As the number of months simulated decreases, so does the wideness between lower and upper bands (±2 standard deviations from the mean). In other words, as the levels of uncertainty decreases, the variability in the calculated yields decreases too.



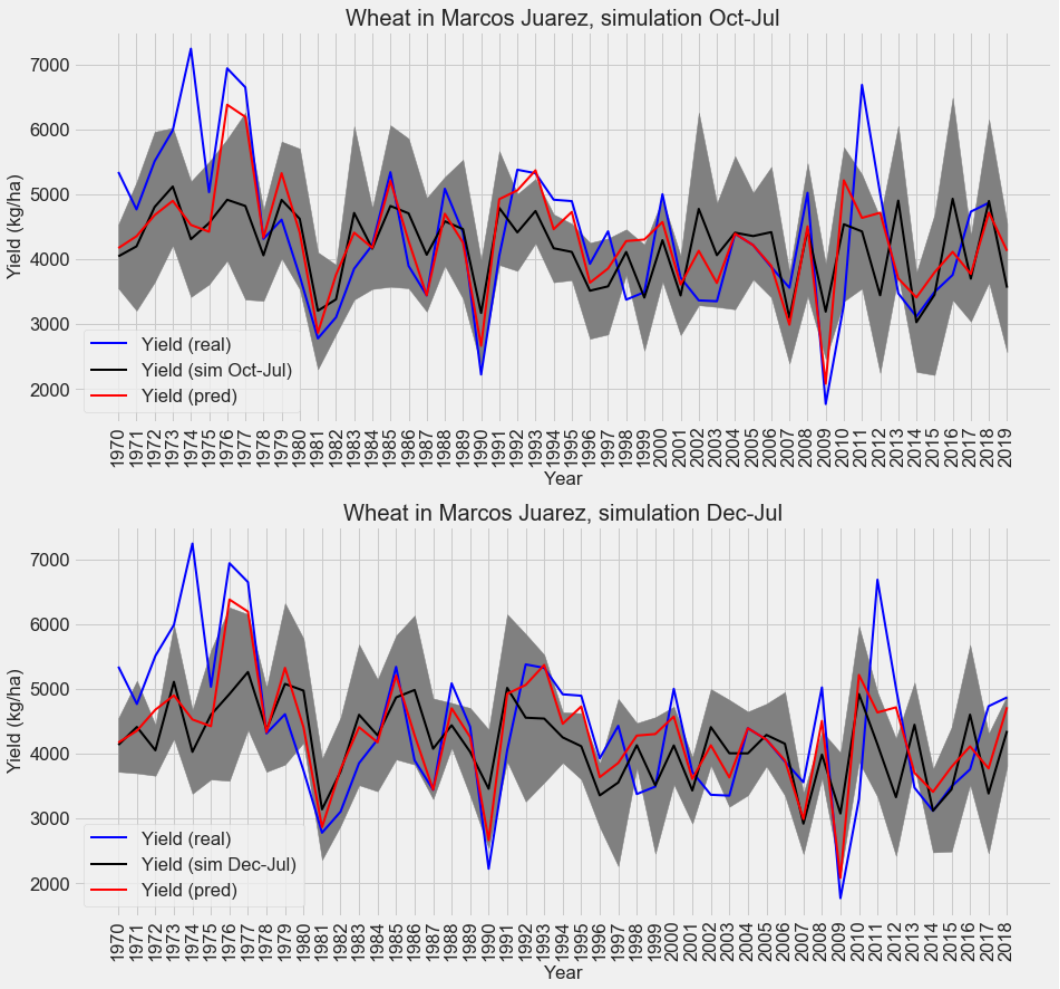


Fig. 41: Time series 1970-2018 simulations for wheat in Marcos Juarez (Cordoba) with 3 levels of simulations. On top: 11 months simulations (August to July), centre: 9 months simulations (October to July) and bottom: 7 months simulations (December to July). Grey bands delimit +- 2 standard deviations from the mean prediction of 100 simulations, which is represented by the black line. Red line is the prediction with real data (no simulation) and blue line is the actual yield time series. Yields are adjusted by TFP.

Looking at the barley example, the mean of the simulations improves its accuracy and copies the pattern of the real yields better when the number of months simulated decreases.

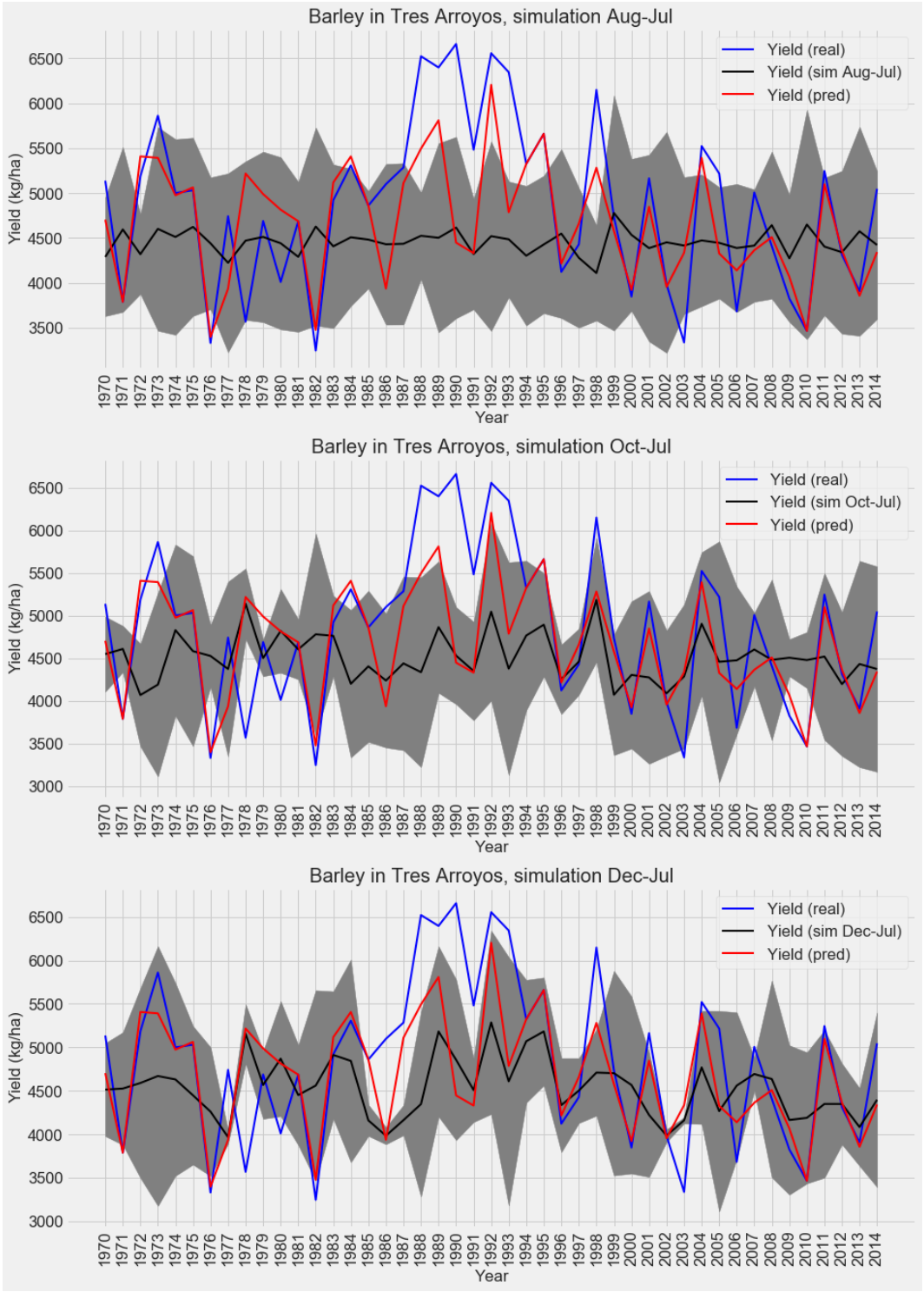


Fig. 42: Time series 1970-2014 simulations for barley in Tres Arroyos (Buenos Aires) with 3 levels of simulations. On top: 11 months simulations (August to July), centre: 9 months simulations (October to July) and bottom: 7 months simulations (December to July). Grey bands delimit +- 2 standard deviations from the mean prediction of 100 simulations, which is represented by the black line. Red line is the prediction with real data (no simulation) and blue line is the actual yield time series. Yields are adjusted by TFP.

DISCUSSION

DISCUSSION

This study illustrates that Random Forest Regressor is highly effective for regional scale crop yield predictions. Random Forest models outperformed baseline by more than 30% on average across 9 crops analysed and outperform LSTM Neural Networks in almost every comparison. The only situations were LSTM came closer to RF performance was for crops with fewer amount of training information. On the other hand, RF strongly overfitted the training dataset while LSTM results were consistent between training and test dataset.

Results correspond with findings made by (Kim et al. 2016) whom compared predictability performance between Random Forest Regressor and Multi Linear Regression (MLR) algorithm. Their study concluded that RF outperformed MLR in all the 4 crops analysed (maize, silage maize and potato in United States and global wheat).

In addition to RF predictive capability, RF also provides useful information about variable importance and dependence. The variable importance rank and the partial impact of the variable on the response can be evaluated for the purpose of systems analysis (Díaz-Uriarte and Alvarez de Andrés 2006). This study used Mean Squared Error (MSE) to identify the most influential variables determining crop yield in the 9 crops and more than 80 counties analysed.

RF intrinsically separates a random subset of data for performance testing from calibration data by using the remaining set of data for model training Therefore, splitting data for training and validation is likely a redundant procedure when applying RF for crop modelling and its performance may increase as more data is included for training (Kim et al. 2016).

INTA's soil productivity index proved to be a very powerful feature to estimate yield for any crop. Other soil characteristics such as deepness, alkaline levels were also useful across most of the crops analysed.

Weather variables associated to water availability and water stress were substantially important too. All summer crops seem to be highly dependent on water availability, even rice which is majorly irrigated. Sunflower suffers from high wind speed (so does cotton) and cotton is affected by mean and maximum temperatures. Winter crops (barley and wheat) share common feature importance's, with special importance of maximum temperatures and GGD (Growing Degree Days).

Although feature importance was variable across crops, a simple model using a combination of soil most important variables and accumulated rainfall in 3 critical stages of the season proved to be as accurate as the complex multi-feature model initially trained.

The ability to standardise the same features across all crops was a major success and made simulation become more powerful. By analysing the monthly rainfall historical distribution for each region and modelling Kernel Density Estimators (KDEs), simulations were performed for all regions from which scenario analysis was conducted to evaluate forecasting with different levels of uncertainty.

MAIZE

Maize's training dataset represents 31% of the total national production and expands to 68 counties and 6 provinces.

Simplified RF Regressor for maize was one of the best accuracies obtained in this study. RF 7 featured model explained 76% yield variance for the test set, with a statistical significance agreement (p-value < 0.05) between the observed values and predicted ones from the test data. The MAE was 924 kg/ha, which was 15.7% of the observed yield and RMSE 1,405 kg/ha. The ability to predict maize yield allows simulations to successfully forecast yield with MAPE of 18% as from January. This means that 4-5 before harvest the crop yield can be predicted with reasonable accuracy.

(Cunha, Silva, and Netto 2018) predicted maize (US) and soybean (US and Brazil) yields using LSTM Neural Networks. In their findings they reported MAE for US maize of 1,031 kg/ha (based on 2,204 counties observations), MAPE 11.3% and RMSE 1,390 kg/ha.

(Kim et al. 2016) estimators for maize yield using Random Forest Regressor are very similar to the ones obtained in this work. They reported a Coefficient of Determination of 0.73 and RMSE = 1,130 kg/ha.

(Aditya Shastry 2017a) analysed maize, wheat and cotton yields modelling comparing linear and non-linear statistical regressor algorithms. They reported coefficient of determination R^2 between 0.75 (linear model) and 0.9 (quadratic model) and RMSE between 600 and 400 kg/ha respectively for maize yields.

(Kaul, Hill, and Walthall 2005) evaluated Artificial Neural Networks (ANN) model performance and compared the effectiveness of multiple linear regression models to ANN models. ANN models consistently produced more accurate yield predictions than regression models. ANN maize yield models resulted in R^2 and RMSEs of 0.77 and 1,036 kg/ha versus 0.42 and 1,356 kg/ha for linear regression, respectively.

SOYBEAN

Soybean's training dataset accounts for 37% of total national production spread over 61 counties. RF estimator was able to predict 68% of yield variance with a relative error of 17.25% over observed yields. MAE was 454 kg/ha, RMSE of 685 kg/ha and association between predicted and actual values was statistically significant (p-value < 0.05). Simulations were able to predict soybean yield from December with a MAPE < 18%. Again, being able to anticipate harvest actual yields 5-6 months beforehand proves to be very powerful.

(Kaul, Hill, and Walthall 2005) reported ANN soybean yield models for Maryland resulted in R^2 and RMSEs of 0.81 and 214 kg/ha versus 0.46 and 312 kg/ha for linear regression, respectively.

(Cunha, Silva, and Netto 2018) trained dataset with Brazilian soybean yields with LSTM Neural Networks and obtained R^2 0.55, MAE 288 kg/ha, RMSE 386 kg/ha and MAPE 11%.

WHEAT

Wheat's training dataset is integrated by 61 counties that account for 43% of total national production. Coefficient of determination R^2 for RF wheat estimator was 0.59. This resulted in a MAE of 496 kg/ha representing 15.7% of observed yields. Correlation between actual and prediction was statistically significant (p-value < 0.05).

(Aditya Shastry 2017a) reported R^2 score around 0.9 and RMSE values between 200 and 300 kg/ha for winter wheat. In this study, RMSE for wheat was 750 kg/ha and R^2 equal 0.59.

(Kim et al. 2016) also studied wheat yields, using Random Forest Regressor for global yield records derived from wheat atlas mega environment classification (CIMMYT, n.d.). They reported coefficient of determination of 0.96, RMSE of 320 kg/ha and MAPE 11.9%.

Although forecasting wasn't evaluated monthly as it was for summer crops time series analysis showed an acceptable fit of yield predictions using weather simulated data against actual yields.

SUNFLOWER

Sunflower's training set represents 28% of sunflower's national production and is conformed by 55 counties. Correlation between observed and predicted values was also significant (p-value < 0.05). RF regressor model was able to explain 51% of yield variance with a MAE of 371 kg/ha, which represents 14.3% of observed yield. Forecast simulations were able to predict sunflower yields with a relative error of 20% as from October for main sunflower regions. This is the earliest stage of monthly forecast prediction obtained in this work.

SORGHUM

Sorghum's training dataset represents 42% of total national production all together and data was gathered from 54 counties. RF model explained 66% of yield variance with MAE of 906 kg/ha, that represents 17.5% of observed yield. Correlation between actual and prediction test data was also significant. In relation to forecasting, with less than 15% relative error, May yields can be predicted in January using weather simulations.

BARLEY

Barley's training set accounts for 39.5% of national production integrated by 49 counties. Correlation between actual and predicted data was statistically significant. Coefficient of determination R^2 was 0.51 and MAE equal 586 kg/ha, MAPE equal 20.5%.

(Ayoubi and Sahrawat 2011) designed ANN models to predict biomass and grain yield of barley from soil properties; and they compared the performance of ANN models with multivariate regression models. In their study, they reported R^2 of 0.93 and 0.89 for ANN and regression models respectively.

RICE

Rice training dataset is conformed only by 15 counties, but because rice production is so concentrated in a specific region of the country, those counties account for 39% of national production. Rice RF performed poorly. The model account for 58% of yield variance and correlations between actual and predicted data are not statistically significant. Nevertheless, MAPE values obtained for rice are pretty good (9%), which in fact were the lowest MAPE values obtained for the crops analysed. This resulted in quite attractive forecasts, leading to an error less than 15% forecasting yields from November for a crop that will be harvested in February.

(Ji et al. 2007) investigated the performance of ANN model to predict Fujian rice yield for typical climatic conditions of the mountainous region. They also compared the effectiveness of multiple linear regression models with ANN models. Their study revealed that ANN models consistently produced more accurate yield predictions than regression models. ANN rice grain yield models for Fujian resulted in R^2 and RMSE of 0.67 and 891 kg/ha vs. 0.52 and 1977 kg/ha for linear regression, respectively.

PEANUT

Peanuts training information comes from 10 counties that represent only 4.5% of total national production. For future studies, weather information should be gathered for the main regions were peanuts grow, in Cordoba province. The low representativeness of the dataset was reflected in peanut' model performance. In terms of model accuracy, peanuts obtained the lowest performance across all 9 crops, with negative coefficient of determination (-0.25) and highest MAE reported in the study (23.8%). No simulations were conducted for this crop due to the poor MAE score.

COTTON

Cotton training dataset represents only 9% of national production and in integrated by 10 counties. RF model accounted for 31% of yield variance in cotton and reported a reasonably good RMSE score (702 kg/ha) and MAE score (507 kg/ha)/ Simulations acceptance levels of

accuracy do not occur until the end of cotton season (harvest occurs in March-April) and therefore results are not very interesting in terms of the ability to anticipate yields predictions. Correlation between observed and actual yields is very poor and not significant.

(Aditya Shastry 2017a) trained US cotton yield dataset with different statistical lineal and non-linear algorithms and reported accuracy of 0.6 to 0.8 coefficient of determination R^2 and 350 to 300 kg/ha RMSE for linear and Supported Linear Regressor (SLR) respectively.

(Mwasiagi, Huang, and Wang 2008) designed an ANN model by selecting cotton-growing cost factors to predict cotton yield in Kenya. They reported to find an optimum ANN with 12 neurons resulting in a R^2 of 0.88 and RMSE of 204 kg/ha.

CONCLUSIONS

CONCLUSIONS

This work demonstrates that machine learning algorithms are a competitive alternative to statistical modelling for crop yield prediction. Machine learning algorithms performed better than classical linear regression models, mainly because of the non -linear relationship between soil and weather variables and crop yields, already mentioned by (Kaul, Hill, and Walthall 2005) and by (Lobell et al. 2011).

Random forest outperformed baseline algorithm by 32% when training with all the variables of the dataset and 35% better when training with a simplified model. Similar results were found by (Kim et al. 2016), who analysed RF performance against MLR in wheat, maize, potato and silage maize, and concluded that RF outperformed linear models in all crops analysed. However, further efforts should intend to reduce Random forest overfitting, which occurred systematically in all crops modelled.

LSTM neural networks perform slightly better than MLR, but with large variability between crops. With practically no difference for 4 crops analysed (sorghum, soybean, sunflower and wheat) in a full model (with all soil and weather variables), outperformed baseline by more than 50% for cotton and performed worst for rice (20% lower than baseline). On average, LSTM outperformed benchmark by 8% in MAE in the full model and by 7.4% in the simplified model.

Wheat and barley are more influenced by soil types. Summer crops are affected by water availability, particularly soybean and maize. Together with sorghum and sunflower, soybean and maize are very depending on soil productivity. Sorghum behaves similarly to maize, with capacity to tolerate water stress, but affected by maximum temperatures, explained by the fact that sorghum is planted under more extreme temperature conditions than maize. Sunflower suffers from high wind speed more than any other weather factor. Cotton most important variables are related to temperature and peanut is similar, but also affected by rainfall. Rice is affected by temperature, water stress and heliophany.

Random Forest models trained returned satisfactory results for 6 of the 9 crops tested: wheat, barley, soybean, maize, sorghum and sunflower. The models for these crops had statistically significant correlation between actual and predicted yield for the test dataset, and performance of the 4-accuracy metrics analysed (R^2, MAE, RMSE, MAPE) were very acceptable too. Accuracy metrics obtained in this work are comparable to the ones obtained by (Cunha, Silva, and Netto 2018) and (Kim et al. 2016) for maize, (Cunha, Silva, and Netto 2018) and (Kaul, Hill, and Walthall 2005) for soybean and (Aditya Shastry 2017b) for wheat.

The three remaining crops are less important in terms of area and production in Argentina, this also explains why less amount of training information was available for the latter crops. In consequence, models trained for cotton, peanuts and rice performed poorly, with no significant correlation between actual and predicted yields and poor accuracy for the metrics reported.

From over 600 weather and soil variables, 4 soil attributes and accumulated rainfall in 3 stages (July to September, July to January, July to March) were able to explain crop yields

almost as accurate as the full model trained initially. Soil productivity, soil deepness, proportion of acid soil and alkaline soils with high concentration of sodium are the 4 soil factors.

A 7-variable simplified model had very similar accuracy values for MAE, MAPE and R^2 compared to a complex model with all variables contemplated. This illustrated the importance of feature selection and how identifying the correct factors to explain the predictable variable can make a model simpler and robust. The ability to simplify a model training with only 7 variables, 4 from which remain constant through the season for a given region, makes weather simulation and yield scenarios feasible and easy to explain.

Monthly rainfall simulations conducted for all regions returned reasonable accurate yield prediction scenarios, with promising results for maize, soybean, sorghum, rice and sunflower. These crops' models were able to predict crop yield with MAPE $\leq$ 20% before the crop harvest, making the model acceptable for forecasting. The advantages for forecasting crop yields are various and concern many stakeholders of the agricultural chain. Global and regional estimations are useful for government to plan possible shortages in production to meet demand or exports needs. Farmers forecast production to fix contracts beforehand, compromising production they yet have not harvest, with the benefit of fixing the price of the commodity.

REFERENCES

REFERENCES

(IFPRI), International Food Policy Research Institute. 2018. “Agricultural Total Factor Productivity (TFP), 1991-2014: 2018 Global Food Policy Report Annex Table 5.” Edited by International Food Policy Research Institue (IFPRI). Harvard Dataverse. https://doi.org/doi/10.7910/DVN/IDOCML.

Aditya Shastry, H.A. Sanjay and E. Bhanusree. 2017a. “Prediction of Crop Yield Using Regression Techniques.” *International Journal of Soft Computing*.

———. 2017b. “Prediction of Crop Yield Using Regression Techniques.” *International Journal of Soft Computing* 12 (2): 96–102.

Agroindustria, Ministerio de. n.d. “Datos Agroindustriales.” https://datos.agroindustria.gob.ar/.

Amnon Shashua. 2017. “An Introduction to Machine Learning.” *An Introduction to Machine Learning*, 1–2. https://doi.org/10.1007/978-3-319-63913-0.

Askew, Mike. 2012. “Transforming Primary Mathematics.” *Transforming Primary Mathematics*, 1–152. https://doi.org/10.4324/9780203806746.

Ayoubi, Shamsollah, and Kanwar Lal Sahrawat. 2011. “Comparing Multivariate Regression and Artificial Neural Network to Predict Barley Production from Soil Characteristics in Northern Iran.” *Archives of Agronomy and Soil Science* 57 (5): 549–65. https://doi.org/10.1080/03650341003631400.

Azzouni, Abdelhadi, and Guy Pujolle. 2017. “A Long Short-Term Memory Recurrent Neural Network Framework for Network Traffic Matrix Prediction.” http://arxiv.org/abs/1705.05690.

Bergersen, Linn Cecilie. 2013. “Guiding the Lasso : Regression in High Dimensions Linn Cecilie Bergersen Dissertation Presented for the Degree of Philosophiae Doctor (PhD).”

Biau, Gérard. 2010. “Analysis of a Random Forests Model” 13: 1063–95. http://arxiv.org/abs/1005.0208.

Borrajo, Celina I, Marcelo Braco, and Pedro Ezcurdia. 2011. “ROTACIONES AGRICOLO/GANADERAS EN SUELOS BAJOS DE LA CUENCA DEL SALADO,” 1–20.

Burke, Timothy P. 2016. “Kernel Density Estimation Techniques for Monte Carlo Reactor Analysis.”

CIMMYT. n.d. “Wheat Atlas.” http://wheatatlas.org/megaenvironments.

Crane-Droesch, Andrew. 2018. “Machine Learning Methods for Crop Yield Prediction and Climate Change Impact Assessment in Agriculture.” *Environmental Research Letters* 13

(11). https://doi.org/10.1088/1748-9326/aae159.

Cunha, Renato L.F., Bruno Silva, and Marco A.S. Netto. 2018. "A Scalable Machine Learning System for Pre-Season Agriculture Yield Forecast." *Proceedings - IEEE 14th International Conference on EScience, e-Science 2018*, 423–30. https://doi.org/10.1109/eScience.2018.00131.

Díaz-Uriarte, Ramón, and Sara Alvarez de Andrés. 2006. "Gene Selection and Classification of Microarray Data Using Random Forest." *BMC Bioinformatics* 7: 3. https://doi.org/10.1186/1471-2105-7-3.

Drakos, Georgios. 2018. "How to Select the Right Evaluation Metric for Machine Learning Models: Part 1 Regression Metrics." *Medium*, 7. https://towardsdatascience.com/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regrression-metrics-3606e25beae0.

GeoINTA. n.d. "Suelos de La República Argentina." 2017. http://catalogo.geointa.inta.gob.ar/geonetwork/srv/spa/catalog.search#/metadata/7762b256-adfa-4bd7-a68e-e9cebcf51a4a.

Grabiński, Pawel. 2018. "Feature Engineering, Explained." *KD Nuggets*. https://www.kdnuggets.com/2018/12/feature-engineering-explained.html.

INTA. n.d. "SISTEMA DE INFORMACIÓN Y GESTIÓN AGROMETEOROLÓGICA." http://siga.inta.gov.ar/#/.

———. 2016. "Calendario de Siembra y Cosecha Por Provincia En Argentina." 2016. https://public.tableau.com/views/Cultivosxprovincia/Cultivosporprovincias?:embed=y&:display_count=yes&:origin=viz_share_link.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2009. *An Introduction to Statistical Learning*. *A Modern Approach to Regression with R*. Vol. 102. https://doi.org/10.1016/j.peva.2007.06.006.

Jaokar, Ajit. 2019. "The Mathematics of Data Science: Understanding the Foundations of Deep Learning through Linear Regression." Data Science Central. 2019. https://www.datasciencecentral.com/profiles/blogs/the-mathematics-of-data-science-understanding-the-foundations-of.

Ji, B., Y. Sun, S. Yang, and J. Wan. 2007. "Artificial Neural Networks for Rice Yield Prediction in Mountainous Regions." *Journal of Agricultural Science* 145 (3): 249–61. https://doi.org/10.1017/S0021859606006691.

Kaul, Monisha, Robert L. Hill, and Charles Walthall. 2005. "Artificial Neural Networks for Corn and Soybean Yield Prediction." *Agricultural Systems* 85 (1): 1–18. https://doi.org/10.1016/j.agsy.2004.07.009.

Kim, Soo-Hyung, Vangimalla R. Reddy, Nathaniel D. Mueller, James S. Gerber, Kyo-

Moon Shim, Jig Han Jeong, Ethan E. Butler, et al. 2016. “Random Forests for Global and Regional Crop Yield Predictions.” *Plos One* 11 (6): e0156571. https://doi.org/10.1371/journal.pone.0156571.

Lema, Daniel. 2015. “Crecimiento y Productividad Total de Factores En La Agricultura Argentina y Países Del Cono Sur 1961-2013.” *Serie de Informes Técnicos Del Banco Mundial En Argentina, Paraguay y Uruguay N°1*, 1–63. http://documents.worldbank.org/curated/en/970151468197997810/Crecimiento-y-productividad-total-de-factores-en-la-agricultura-Argentina-y-países-del-cono-sur-1961-2013.

Lobell, David B., J. Ivan Ortiz-Monasterio, Gregory P. Asner, Pamela A. Matson, Rosamond L. Naylor, and Walter P. Falcon. 2005. “Analysis of Wheat Yield and Climatic Trends in Mexico.” *Field Crops Research* 94 (2–3): 250–56. https://doi.org/10.1016/j.fcr.2005.01.007.

Lobell, David B, Marianne Bänziger, Cosmos Magorokosho, and Bindiganavile Vivek. 2011. “Nonlinear Heat Effects on African Maize as Evidenced by Historical Yield Trials.” *Nature Climate Change* 1 (March): 42. https://doi.org/10.1038/nclimate1043.

Lobell, David B, and Marshall B Burke. 2010. “On the Use of Statistical Models to Predict Crop Yield Responses to Climate Change.” *Agricultural and Forest Meteorology* 150 (11): 1443–52. https://doi.org/https://doi.org/10.1016/j.agrformet.2010.07.008.

Mohtadi, B.F. 2017. “In Depth: Parameter Tuning for Random Forest.” *Medium*, 2017. https://medium.com/all-things-ai/in-depth-parameter-tuning-for-random-forest-d67bb7e920d.

Murthy, V, and Krishna Radha. 2003. “Crop Growth Modeling and Its Applications in Agricultural Meteorology.” *Satellite Remote Sensing and GIS Applications in Agricultural Meteorology*, 235–61. http://www.wamis.org/agm/pubs/agm8/Paper-12.pdf.

Mwasiagi, Josphat Igadwa, Xiu Bao Huang, and Xin Hou Wang. 2008. “Prediction of Cotton Yield in Kenya.” *South African Journal of Science* 104 (7–8): 249–50.

OECD/FAO. 2017. *OECD-FAO Agricultural Outlook 2017-2026. OECD Publishing, Paris,*. https://doi.org/http://dx.doi.org/10.1787/agr_outlook-2017-en.

Parzen, E. 1962. “On Estimation of a Probability Density Function and Mode.” *The Annals of Mathematical Statistics* 33: 1065–76.

Paswan, Raju Prasad, and Shahin Ara Begum. 2013. “Regression and Neural Networks Models for Prediction of Crop Production.” *International Journal of Scientific & Engineering Research* 4 (9): 98–108.

Potdar, Kedar, Taher S., and Chinmay D. 2017. “A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers.” *International Journal of Computer Applications* 175 (4): 7–9. https://doi.org/10.5120/ijca2017915495.

Prabhu. 2018. “Understanding Hyperparameters and Its Optimisation Techniques.”

Medium, 5. https://towardsdatascience.com/understanding-hyperparameters-and-its-optimisation-techniques-f0debba07568.

Prieto, Gabriel. 2010. “Pautas Para El Manejo Del Cultivo de Maíz,” 1–4.

Strzepe, K. 1994. “Yates K. Strzepe.”

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. 2006. *Introduction to Data Mining*.

Wart, Justin Van, Patricio Grassini, and Kenneth G. Cassman. 2013. “Impact of Derived Global Weather Data on Simulated Crop Yields.” *Global Change Biology* 19 (12): 3822–34. https://doi.org/10.1111/gcb.12302.

WorldBank. 2017. “Agriculture, Forestry, and Fishing, Value Added (% of GDP).” 2017. https://data.worldbank.org/indicator/NV.AGR.TOTL.ZS.

Wu, Wei, Ji long Chen, Hong bin Liu, Axel Garcia y Garcia, and Gerrit Hoogenboom. 2010. “Parameterizing Soil and Weather Inputs for Crop Simulation Models Using the VEMAP Database.” *Agriculture, Ecosystems and Environment* 135 (1–2): 111–18. https://doi.org/10.1016/j.agee.2009.08.016.

Yang, Xiaojun. 2010. “Spatial Interpolation.” *Handbook of Research on Geoinformatics*, 129–36. https://doi.org/10.4018/978-1-59140-995-3.ch017.

Zheng, Alice. 2015. *Evaluating Machine Learning Algorithms*. *Springer*.

Zheng, Alice, and Amanda Casari. 2018. *Feature Engineering for Machine Learning and Data Analytics - Principles and Techniques for Data Scientists*. http://bit.ly/featureEngineering_for_ML.

APPENDICES

APPENDICES

FIGURES

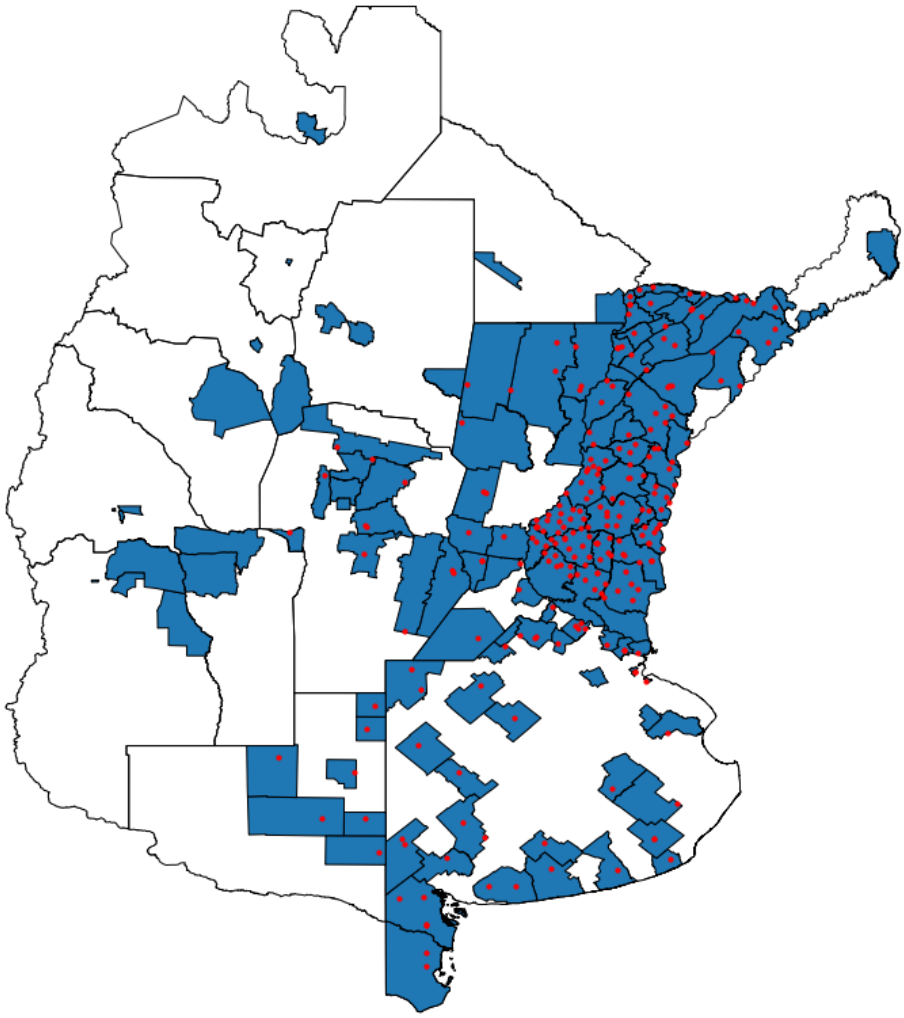


Fig. A 1: Argentinian county level map of the weather stations location used that integrated the dataset. Red dots indicate the location of the weather stations, and blue polygons delimit the counties assigned to weather stations.

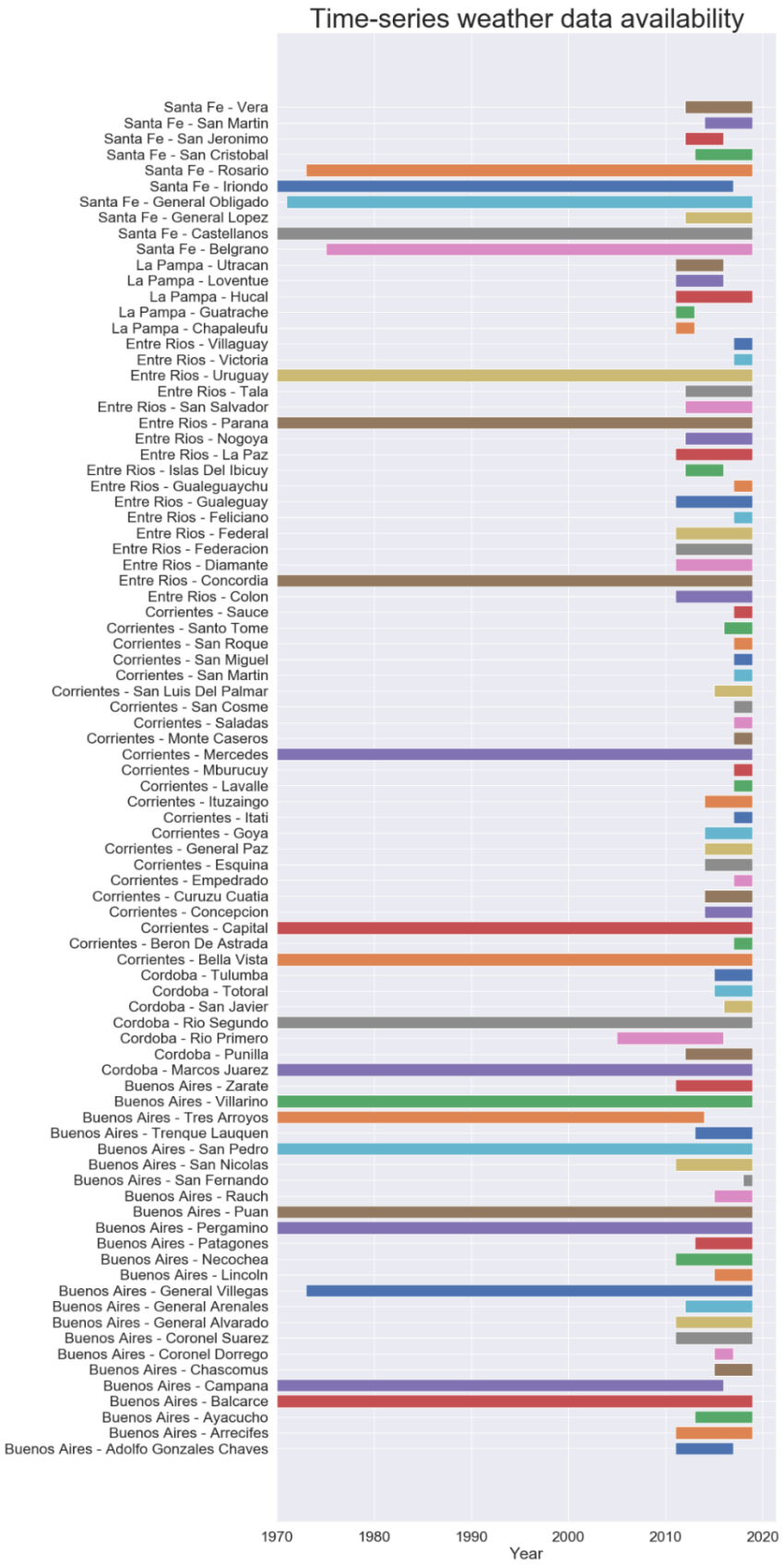


Fig. A 2: Time series data availability of weather stations data by county.

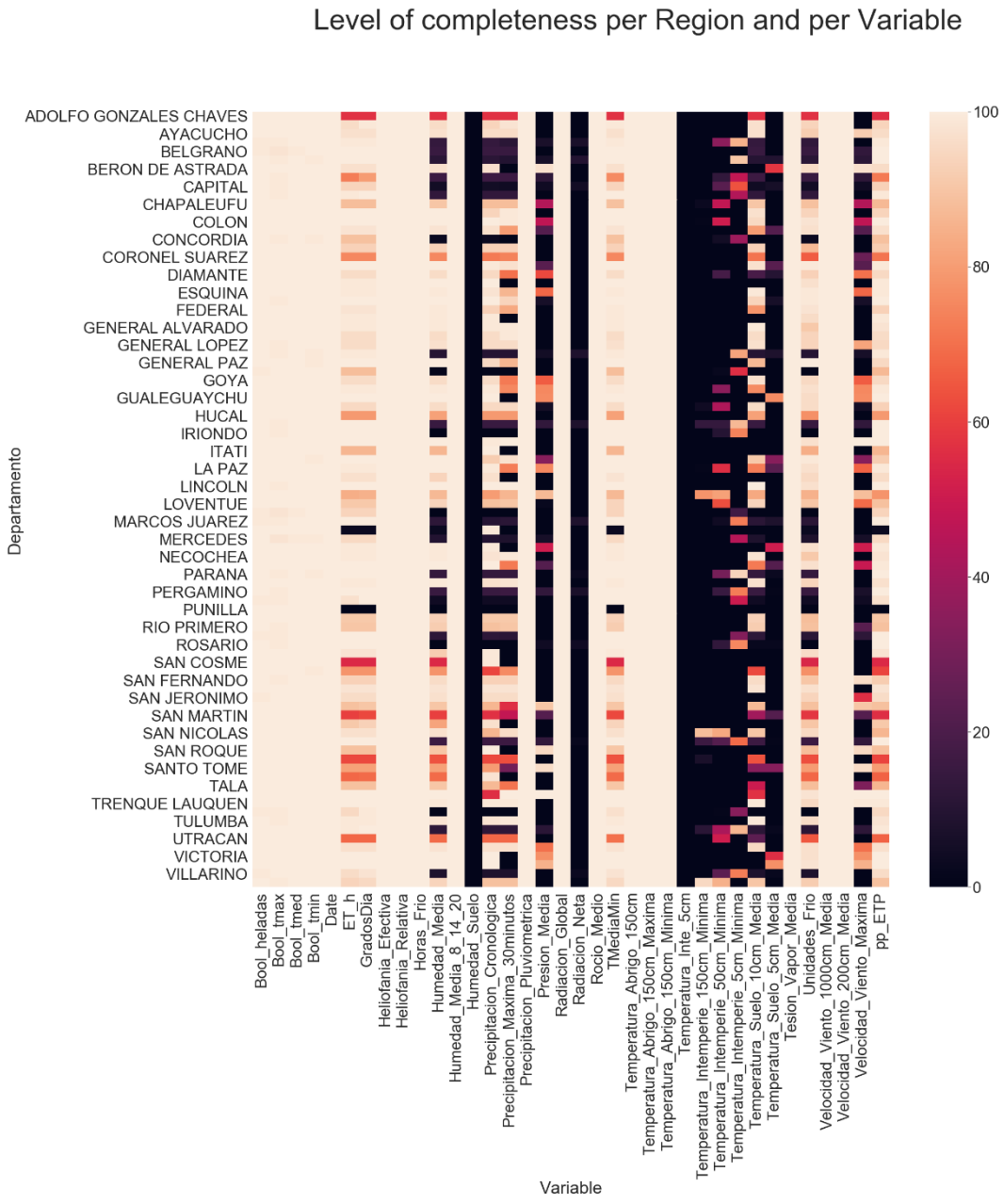


Fig. A 3: Completeness analysis by regions and by weather variable. Values go from 0 (no data) to 100% (fully complete), represented by dark and bright scale respectively.

Fig. A 4: Winter crops sowing and harvest calendar for across provinces in Argentina (INTA 2016).

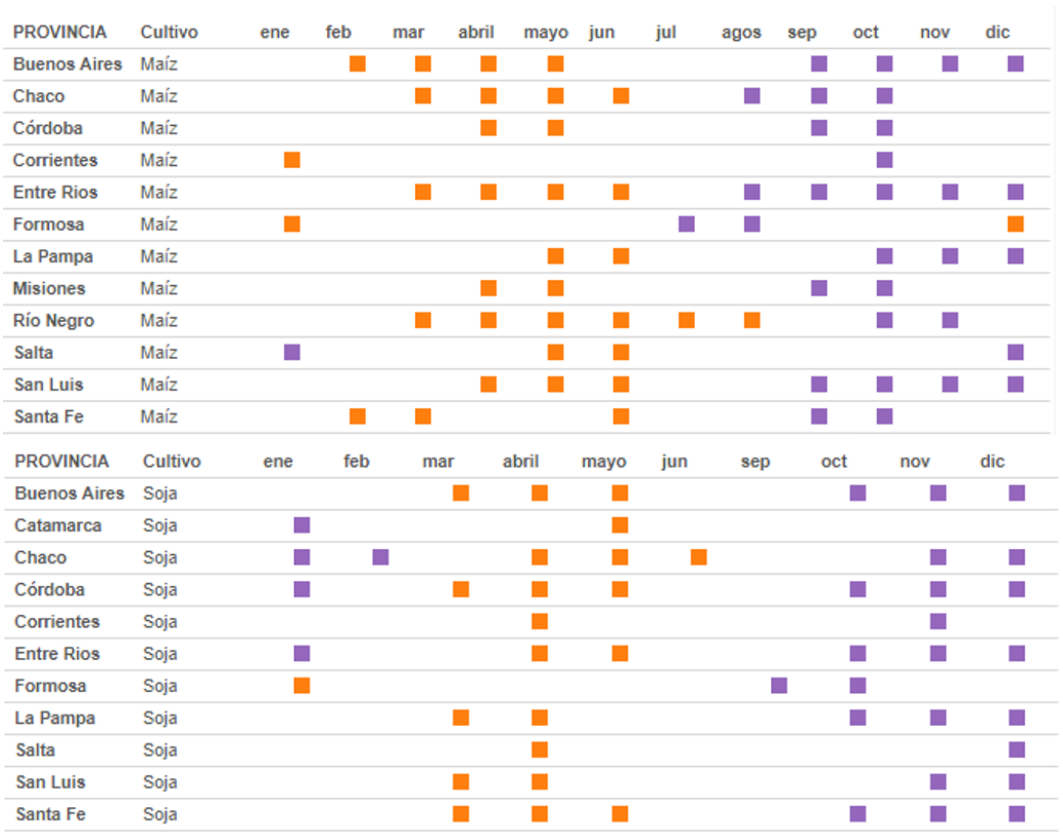


Fig. A 5: Summer crops sowing and harvest calendar for across provinces in Argentina. (INTA 2016).

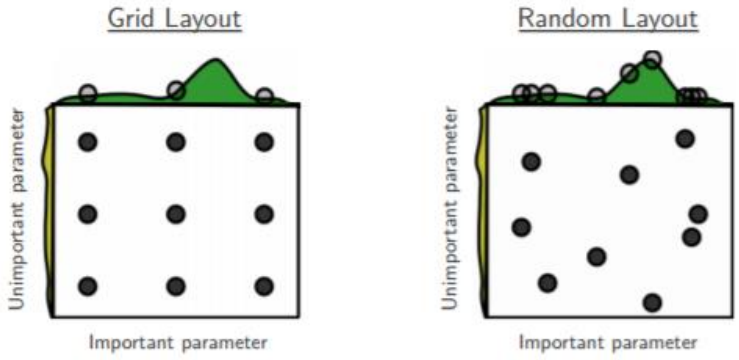


Fig. A 6: Hyperparameter optimization; illustration of Grid search and Random search

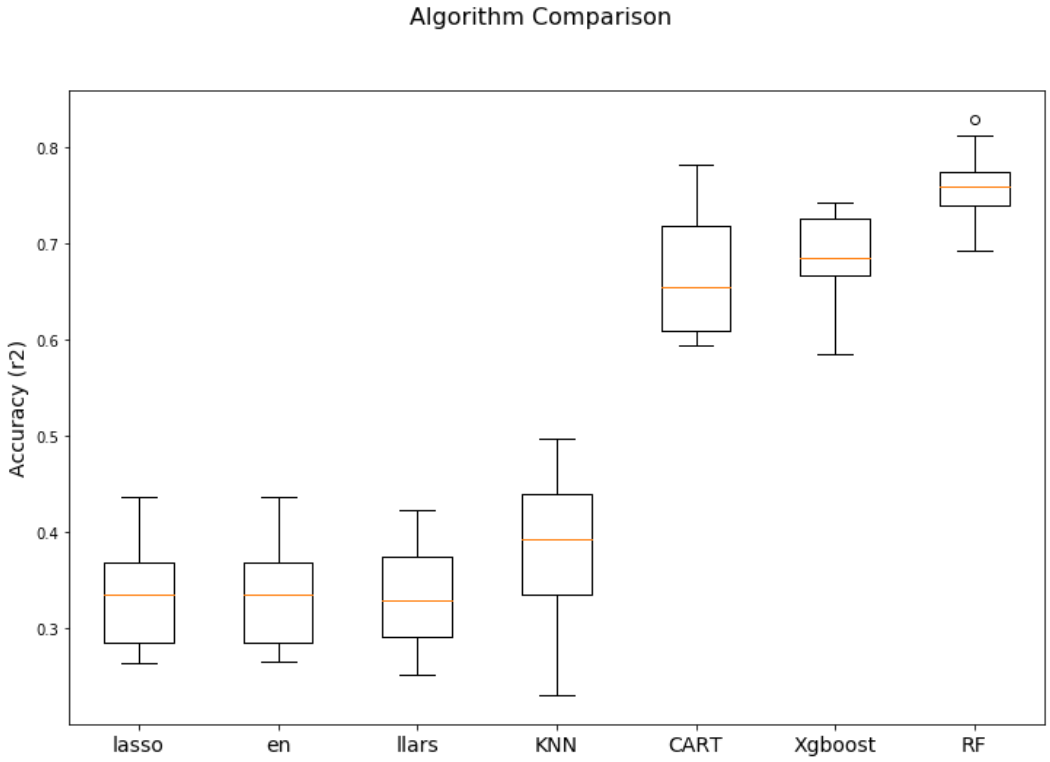


Fig. A 7: Accuracy levels, measured as Coefficient of Determination of different linear and non-linear algorithms tested on maize for the reduced dataset (7 variables); en (Elastic Net), llars (Least Angle Regression), KNN (K-Nearest Neighbours), CART (Decision Tree Regressor), RF (Random Forest Regressor).

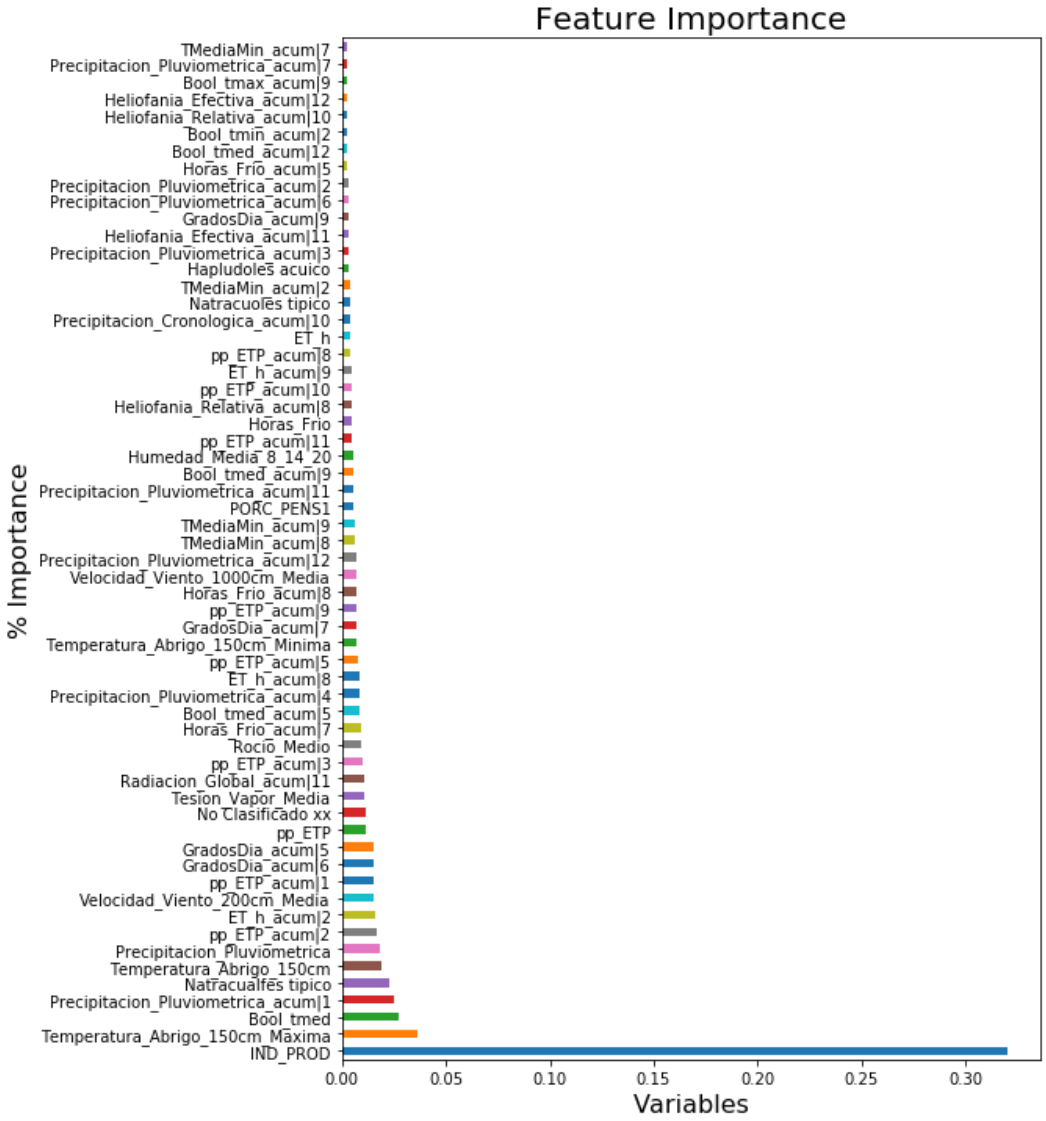


Fig. A 8: Random Forest Regressor top 60 features according to Feature Importance for maize using all variables available in dataset.

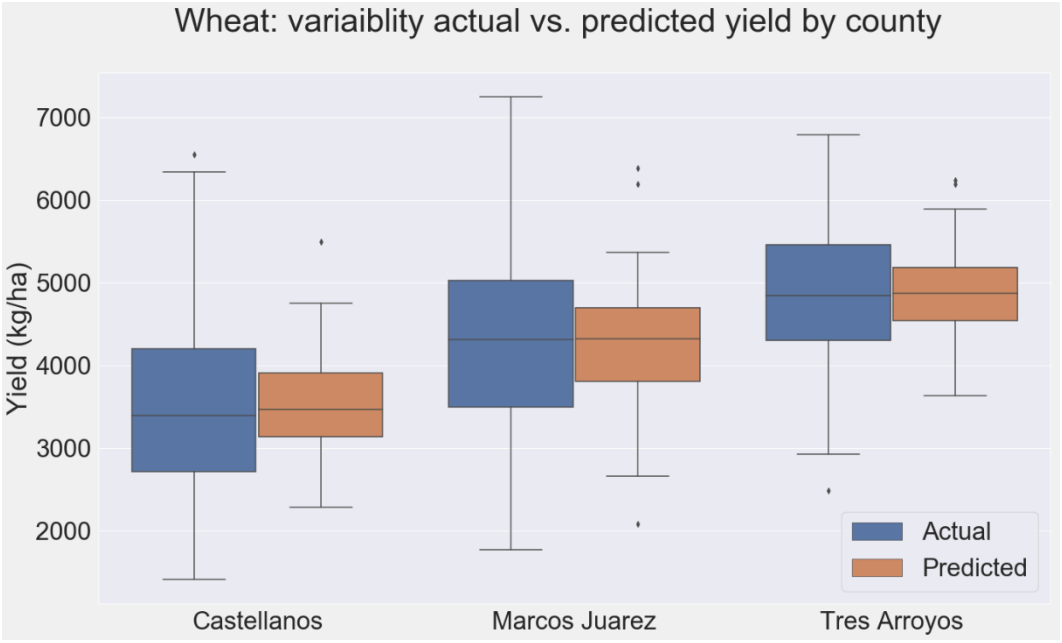


Fig. A 9:Actual vs. Predicted analysis for 3 main wheat production counties using Random Forest simplified model.

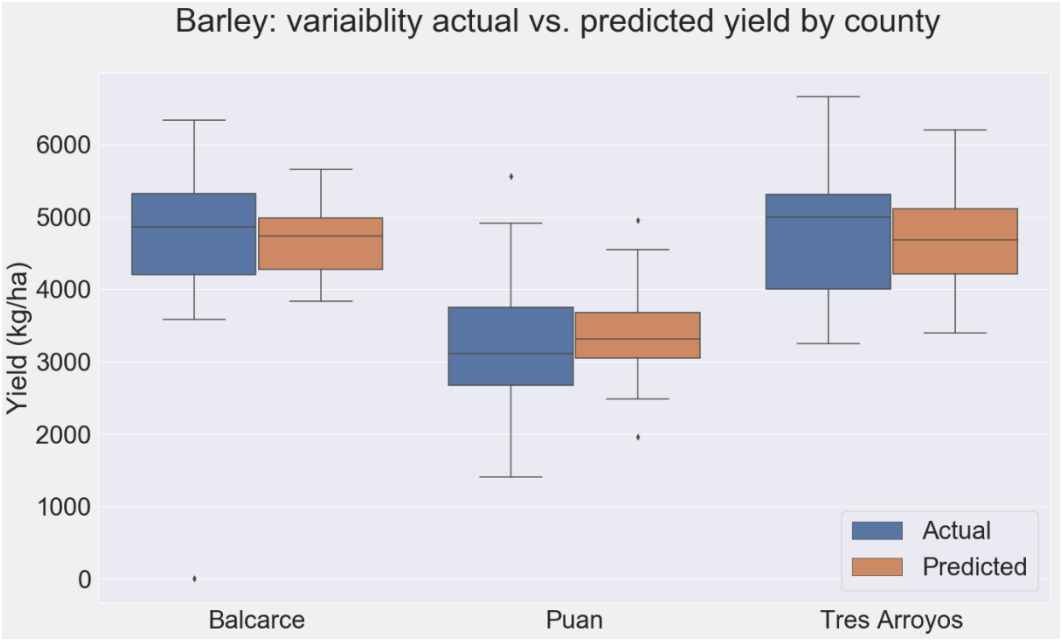


Fig. A 10: Actual vs. Predicted analysis for 3 main barley production counties using Random Forest simplified model.

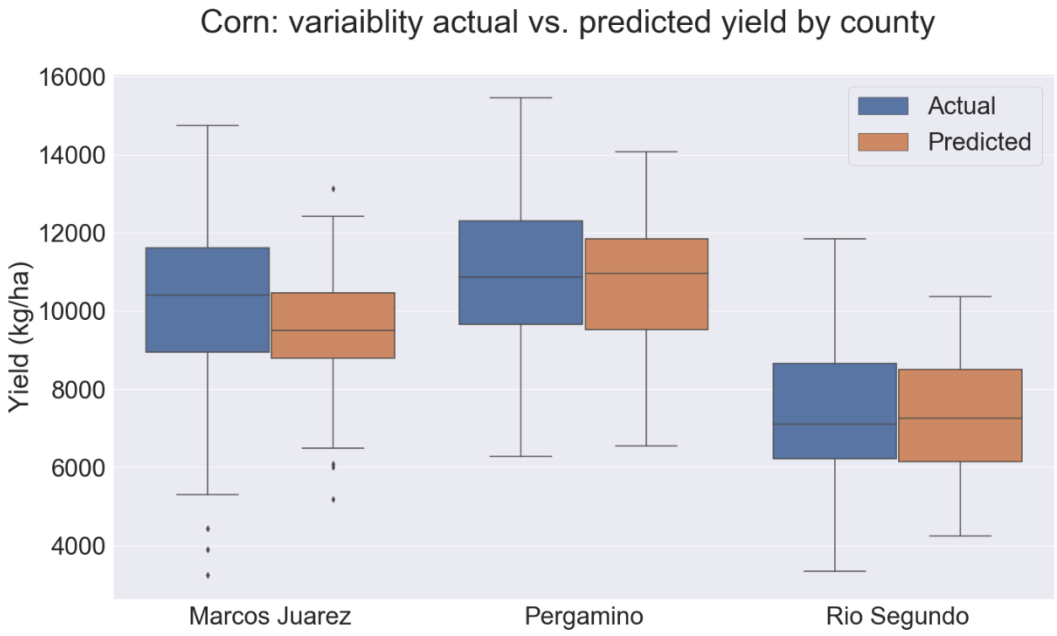


Fig. A 11: Actual vs. Predicted analysis for 3 main maize production counties using Random Forest simplified model.

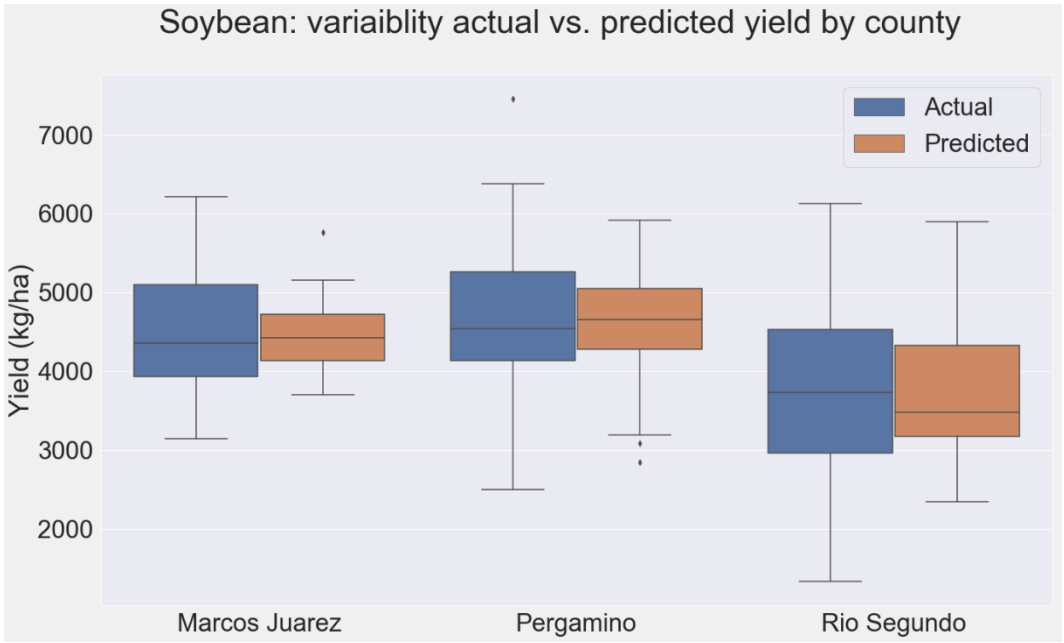


Fig. A 12: Actual vs. Predicted analysis for 3 main soybean production counties using Random Forest simplified model.

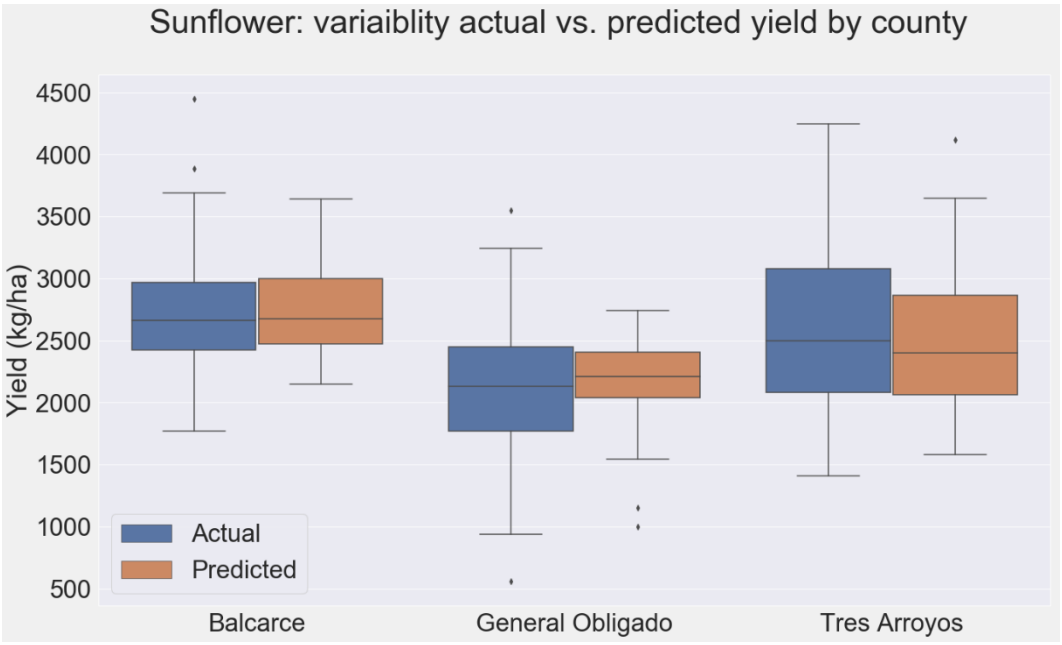


Fig. A 13: Actual vs. Predicted analysis for 3 main sunflower production counties using Random Forest simplified model.

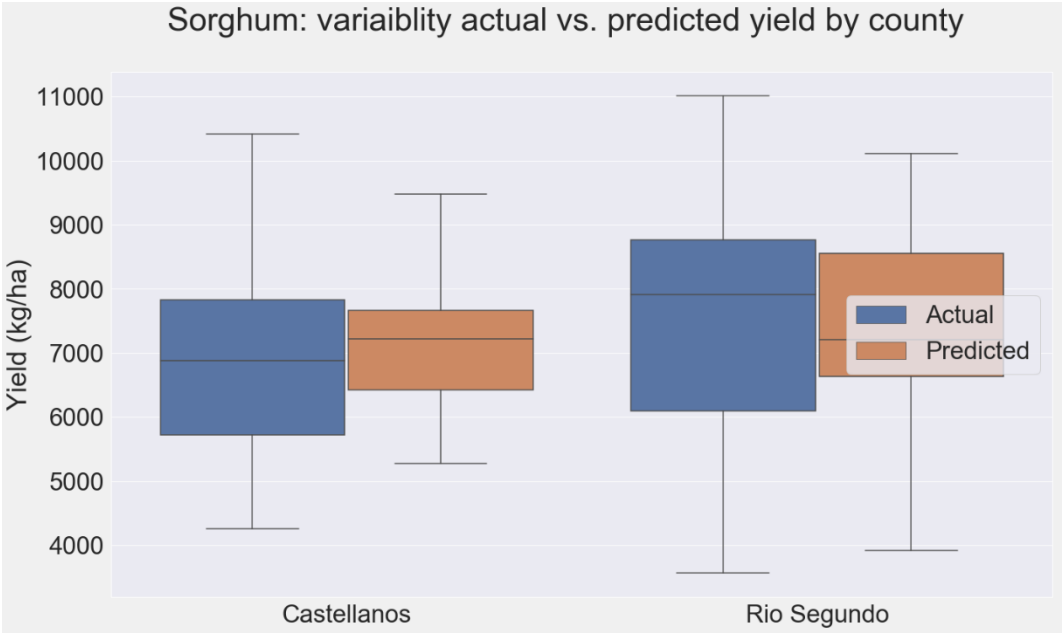


Fig. A 14: Actual vs. Predicted analysis for 3 main sorghum production counties using Random Forest simplified model.

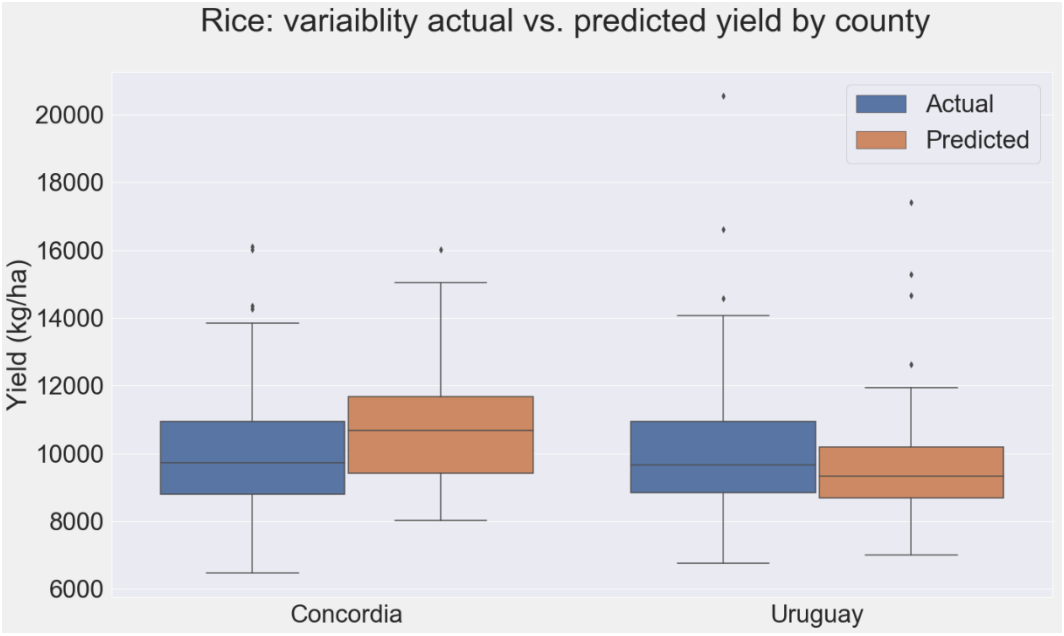


Fig. A 15: Actual vs. Predicted analysis for 2 main rice production counties using Random Forest simplified model.

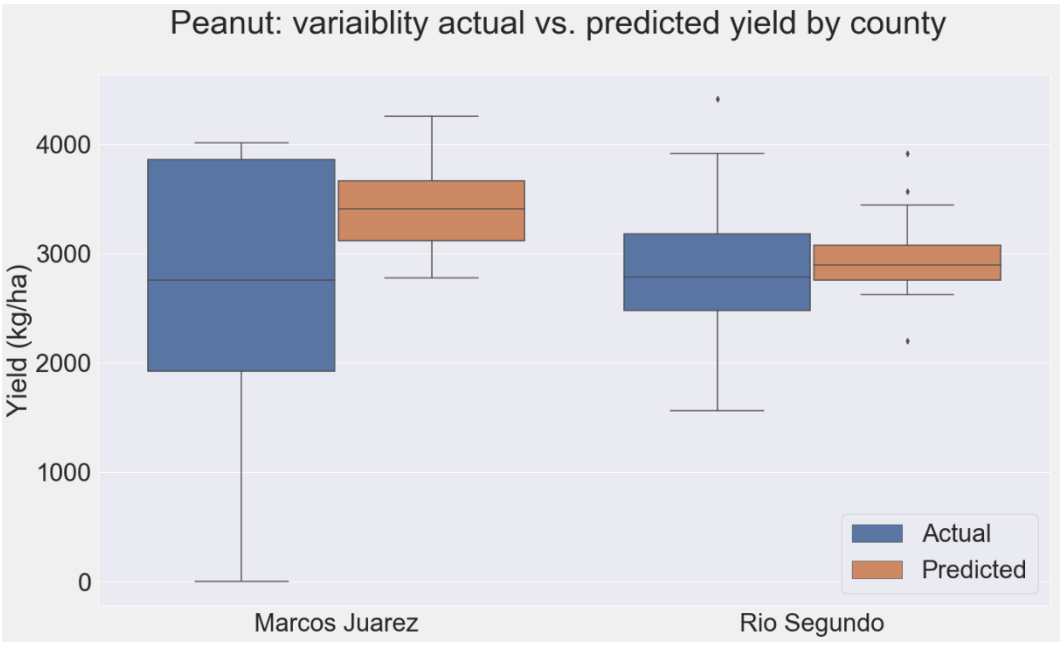


Fig. A 16: Actual vs. Predicted analysis for 2 peanut production counties using Random Forest simplified model.