Facing Crop Yield Uncertainty with Bayesian Neural Network

I. Chinesta,[a,*] A. Ammar,[a] F. Chinesta[b]

[a] LAMPA Lab, Arts et Métiers Institute of Technology, 2 Boulevard du Ronceray, BP 93525 cedex 01, F-49035 Angers, France

[b] PIMM Lab, Arts et Métiers Institute of Technology, 151 Boulevard de l'Hôpital, 75013 Paris, France.

Abstract. Focus on sustainability in the agricultural industry projects, this work analyses how to reduce uncertainty in the agriculture production. We build a panel fixed effects four-degree polynomial regression, a Random Forest (RF) and a Bayesian Neural Network (BNN) to compare their prediction power for soft winter wheat crop yield at metropolitan France regional level. The polynomial regression results are, a correlation coefficient(R^2) of 0.8978 and a root mean squared (RMSE) of 0.2323. With RF model we obtain a 0.9716 value in the R^2 and a RMSE of 0.2378, with a relevant reduction of the variable dimension with respect to the polynomial regression, from 117 to 20. Finally, after a data mining process, we build our BNN model, that shows the best prediction accuracy with six independent variables, providing a R^2 of 0.7122 and a RMSE of 0.2291.

Keywords: Machine-Learning, Crop yield prediction, Industrial uncertainty, Bayesian inference, Precision Agriculture.

***First Author E-mail: ignacio.chinesta@ensam.eu

1 Introduction

Agriculture is the cornerstone of the human society and nowadays remains as one of the most relevant industries worldwide. The last Statistical Yearbook of the Food and Agricultural Organization of the United Nations (FAO Statistical Yearbook 2024) shows interesting figures about the industry for the 2000-2022 period. Despite the global agricultural land area has decreased in a 2% and the number of people working in agriculture, including forestry and fishing, went down in a 13%, we also can see that share of agriculture in the global gross domestic product (GDP) has been stable for this period, around 4%, and the added value increased in 89%, to 3.8 USD trillion. Agriculture is also in the core of the 2030 sustainable development agenda, as food security and nutrition and sustainable agriculture conform the "zero hunger" goal of the agenda. The goal of transform agriculture in a more sustainable industry and climate change, have encouraged academics to focus their researches into study crop yields. By typing the term "Crop yield prediction" into the Scopus search engine, we can see how publications in this area have grown steadily in recent years, from 260 in the year 2016 to 1669 in 2024.

Our literature review starts with an analysis of the statistic models. These models-focus in provide empirical evidences of how climate variables affect to crop yield. Panel fixed effects models are commonly developed for this kind of researches. One great example is the recent work of D. Wang et al. (2023). With the aim of identify adaptive measures for climate change in the future, authors leverage quasi-experimental variations in irrigation induced by a natural experiment for irrigation expansion started in 1996 and quantify the contribution of irrigation access to the overall adaptation effect. Their results indicate that improving irrigation access may help for mitigating yield loss from a warming climate.

Polynomial regressions are largely used and provide good results showing the non-linear relationship between independent climate variables and crop yield. We highlight the work of Matthew Gammans et al, (2017). In this work, authors developed a flexible statistical yield model

using a long panel data from France to study the impacts of temperature and precipitation changes on wheat and barley yields. Their model predicts a 21% and 17.3% decline by the end of the century in winter wheat and winter barley yields under the Representative Concentration Pathway 8.5(RCP8.5) scenario. RCP models simulate scenarios following different Greenhouse Gas (GHG) emission concentration paths. RCP8.5 represents a warming path scenario, for the period 2011-2100, of 0.63°C/10. The authors work also predicts that continuing technology trends will counterbalance most of the effects of the climate change.

For us this was a very relevant work as they studied the same country we are interested in and really important products within the French agricultural industry. France is the main cereal-producing country of the EU. The country produced de 27% and 22% of total EU's wheat and barley production, respectively. This means more than 120 tonnes of wheat and more than 40 of barley in 2022, as reflects the Statistical Book 2023 published by the Department of Statistics and Foresight Analysis (SSP) of the Ministry of Agriculture and Food Sovereignty.

Our research is focused into study the agricultural industry from the smart territory (ST) framework. I. Gorelova et al. (2024) following a systematic literature review (SLR) approach to study the ST concept built a detailed table with the research domains observed in their studied literature and the codes retrieved for each domain. In this table we can find the domain "Territorial Management" link to de codes: Smart management, management platform, decision support system, management of the infrastructure, smart growth and smart governance. Another domain we can find is "Rural Development" linked to the codes: Sustainable agriculture, land scape planning, land consumption and rural spatial planning. With this information, we understand that, for this first work focused in crop yield estimation, we have to go further than with statistical models with the aim of get the most accurate estimations.

Despite the classical statistical models work quite well, specially fixed effects polynomial regressions, in terms of show causal relations between variables, machine learning (ML) tools generally provide more accurate predictions. In our study of the related literature, we observe that authors commonly compare the predictive power of traditional regressions with different ML tools. R.A. Schwalbert et al. (2020) compare multivariate Ordinary Least Squares (OLS) linear regression with a random forest (RF) and Long-Short Term Memory (LSTM), Neural Network for soybean crop yield prediction. Our work goes in the same direction in the sense that we start from a regression model to analyse the empirical relationships between dependent and independent variables, but we built a panel fixed effects polynomial regression instead of an OLS multivariable linear regression. We build also a RF model with the aim of get better estimations. Due to the length of our data, was not possible to build a LSTM model. In a similar work, M. Kuradusenge et al. (2023) test the power of polynomial regressor, RF and Support Vector Regressor (SVR) for maize and potatoes production prediction, where RF provides the best performance, and the polynomial regressor works better that the SVR.

Another technique that is common seen in the literature is to combine different tools. Join convolutional neural networks (CNN) and LSTM is a recurrent idea that provides that make possible to take profit of the image datasets. Sun et al. (2019), working on maize, wheat and potatoes forecasting, evidence in their research that combining these two tools provides better results than use each method separately.

From the smart territory framework and its sustainable development target, crop yield prediction is a key point for crop planning and allocation tasks. At the same time, these tasks are translated in medium(long) term investment decisions, insurance planning and territorial policies that resulting in a more complex scenario, with a higher level of uncertainty. When working in so many complemented tasks, we need to manage a range of possible scenarios and quantify the probability of each scenario occurring. First, we have a set of possible scenarios that can occur when we face a problem, and data that allow us to build prior believes about each scenario, in other words, with the data we can estimate the likelihood of each scenario occurring. Then if we calculate the proportion of each scenario's likelihood over the entire set of scenario likelihoods, we can estimate the probability of each scenario occurring.

What has been described in the previous paragraph is basically how to estimate a probability following the Bayesian method. Another important feature of the Bayesian logic is that we can update our previous believes every time we get new data. Bayesian methods are common to quantify risk and face uncertainty in different fields of knowledge. Specifically, for this work we see interesting to work with Bayesian Neural Networks (BNN).

A BNN is a probabilistic implementation of a standard neural network with the key difference being that the weights and biases are represented via the posterior probability distributions rather than single point values (Chandra and Simmons, 2024). This structure allows us to make predictions within a 95% confidence level band of values, what may help to reduce the uncertainty with a better mathematical explanation of how risky is a given project.

Despite nowadays BNN are not commonly used for this works yet, we found some evidences of its good performance in crop yield forecasting task. A good example is given by Ma et al. (2021), in their work authors showed the outperformance of BNN over RF, LSTM, SVR, Ridge regression and classic Multi-Layer Perceptron(MLP) in corn crop yield forecasting, and also was demonstrated the capacity of BNN for reduce the uncertainty: "We also assessed the predictive uncertainty, and more than 84% of the observed yield records were successfully enveloped in the 95% confidence interval of the predictive yield distribution".

As said above, this research consists in to analyse the agricultural industry from the point of view of the smart territory framework, provide a series of tools for producers, financial agents and policy makers to let the industry reach a maximum degree of sustainable development and reduce the level of uncertainty related with each operation within the industry. Provide tools to face crop yield uncertainty means to build algorithms that work in a scenario in which we cannot control all relevant variables, and this also implies to carry out a data analysis work focus in to detect the appropriate set of variables. In this work first, we reduced set of independent variables to a six, after a data features analysis and secondly, we take profit of the BNN's properties to make consistent predictions with the given independent variables set. Our results show that is possible to make soft winter wheat yield predictions with a reduced set of easy to access variables.

In the next section we will describe our data, sources and a data analysis done to detect the relevant features and statistics of our dataset. In section 3 we expose the techniques considered and in sections 4 and 5 we will show the results obtained and share our final conclusions and thoughts.

2 Data

2.1 Data presentation

The geographical area of this study is the metropolitan France, divided in the 21 regions that formed the metropolitan France, with exception of Corse, before the regional changes of the year 2016. Our agricultural dataset contains the crop production and land surface used for the cultivation of a high range of products in the period 2010-2023. The data is available in the web site of AGRESTE, the French ministry of agriculture. The data is annual and we use production and surface to obtain for each year, the product crop yield in tonnes per hectare, which is our dependent variable. This study will focus on soft winter wheat yield.

The selection of the independent variables is based on what we have observed in similar researches and studies focus in the literature review as the work of T. van Klompenburg et al. (2020), where the authors performed a systemic literature review to analyse which variables and tools are more used in this kind of works. Obviously, temperatures and rainfall level are very relevant variables, but literature also highlight the relevance of soil characteristics and vegetation, that could explain the good performance of satellite image-based CNN models.

For this research, the climate independent variables were provided by the Climate Change Knowledge Portal of the Word Bank Group. With respect to the soil characteristics, we used the Soil Wetness Index (SWI) provided by the open data catalogue of Météo France. We observed a notable improve of the results adding this variable. The SWI is defined by Meteo-France as *"a soil moisture index documented in the scientific literature. It represents, over a depth of approximately two meters, the state of the soil water reserve in relation to the useful reserve (water available for feeding plants). It is therefore the water status of the surface soil and not the filling of the water tables. If the SWI is equal to zero, the soil is very dry and plants can no longer draw water from it, while if the SWI is equal to one, the soil is saturated with water and has reached its useful reserve".*

All the data is available monthly and quarterly. We selected the quarterly format as aggregation period because many of agricultural products grow during one or two seasons of the year. In the case of soft winter wheat, during almost a complete year. This means that each variable has four values for each year. We set the subindex "1Q" for the aggregate or mean value, depending of the variable's nature, for the quarterly that starts in the previous year's December, and finishes with the end of February of the current year. In other words, for the winter season. In the same way we have the subindex "2Q" for the spring season (from March to the end of May), subindex "3Q" for the summer season (from June to the end of August), and "4Q" for the autumn season (from September to the end of November). As, depending of the region and the climatic conditions, soft winter wheat is planted between October and the first week of September, we will also refer as "4Q_t-1" as the autumn season of the previous year.

Table 1 shows all the independent variables used in this research with the definition given by the previous cited data sources.

Table 1 Data format

Variable	Definition	Units	Code
Growing Season Length Start	Annual series with the day of the year (1st Jan to June 30 in Northern Hemisphere, NH, and 1st July to 31st Dec in Southern Hemisphere, SH) that reflects the first span of at least 6 consecutive days with daily mean temperature T >5C.	Days	gslstart
Growing Season Length End	Annual series with the day of the year (1st Jan to June 30 in Southern Hemisphere, SH, and 1st July to 31st Dec in Northern Hemisphere, NH) that reflects the first span of at least 6 consecutive days with daily mean temperature T <5C.	Days	gslend
Cold Spell Duration Index	The number of days each year in a sequence of at least six consecutive days during which the value of the daily minimum temperature is less than the 10th percentile of daily minimum temperature calculated for a five-day window centered on each calendar day, using all data for the given calendar day-pentad from the data period for a reference climate (e.g., present-day climate).	Days	csdi
Warm Spell Duration Index	The number of days in a sequence of at least six consecutive days during which the value of the daily maximum temperature is greater than the 90th percentile of daily maximum temperature calculated for a five-day window centered on each calendar day, using all data	Days	wsdi

	for the given calendar day-pentad from the data period for a reference climate.		
Average Mean Surface Air Temperature	Average mean temperature over the aggregation period	°C	tas
Average Maximum Surface Air Temperature	Average daily maximum temperature over the aggregation period	°C	tasmax
Average Minimum Surface Air Temperature	Average daily minimum temperature over the aggregation period	°C	tasmin
Average Largest 5-Day Cumulative Precipitation	The average highest precipitation amount over a consecutive 5-day period during each month in the data period.	mm	rx5day
Precipitation	Aggregated accumulated precipitation.	mm	pr
Relative Humidity	Based on daily mean relative humidity at 2m as reported by climate models, or derived from specific humidity reported by climate models.	%	hurs
Number of Frost Days (Tmin<0C)	Number of Frost Days (Tmin<0C).	Days	td
Soil Wetness Index	Represents, over a depth of approximately two meters, the state of the soil's water reserve in relation to the useful reserve (water available for plant nutrition).	%	SWI

2.2 Data Mining

As there is a huge range of variables that can affect to the agricultural crop yield, face uncertainty on this context means to build models taking into account that is not possible to control all relevant variables. With the aim of select as reduced as possible set of variables, we made a data mining process. The research developed by P. Kamath, P. Patil, E.S et al (2021) analyse how data mining can facilitate crop yield prediction, in this work authors studied the Random Forest approach. In our work, first we start from the basic statistic features of our dataset: standard deviation, the mean, maximum, and minimum values for each individual, and take a preliminary view of our dataset. After this first look, we identify some interesting patrons and study the correlations between variables. Then, to make a deeper analysis of the data from different approaches, we run advanced

feature selection tools such as the principal component analysis (PCA), minimum redundance maximum relevance (mRMR) algorithm, the Naive Bayes Classifier (NBC) and Boruta algorithm, a little bit more sophisticated version of the RF variable importance estimation approach.

In our preliminary view, we observed that for 16 of the 21 regions, the difference between the minimum and the average yield value was greater than the difference between the maximum and average values. This could mean that extreme low values of crop yield are more linked to unforeseen events than extreme high values.

To check if the previous hypothesis has fundaments, we checked the worst crop yield for each individual and we discovered that the worst regional yields were concentrated in four of the 13 years that contains our dataset: 2011, 2016, 2020 and 2022. One of the regions had its worst year in 2011. Each one of the years 2020 and 2022 was the worst year for four regions, and 2016 was the worst year for 12 regions, more than half total region number of our dataset.

After the commented above, we study deeply the independent variables of our dataset looking for patrons or extraordinary events that could explain us more, and we found that 2016 was the year in which the variable "pr_2Q", accumulated precipitation in the aggregated period of the spring season, takes a maximum value for all the regions that had his worst yield this year. Also in this year, the variable SWI_3Q, soil wetness index in the aggregated period of summer season, reach its highest value for 10 of these regions. On the other hand, in the years 2011, 2020, 2022, we observed a considerable increase of temperatures with a stable level of precipitations. 2022 is the hottest year in the history of the country, with a 14.5°C temperature average, 0.4°C degrees more that the second average hottest year, 2020. In fact, the four regions that had their worst yield in 2022, reached its maximum of "tas_3Q", average mean air surface temperature in the summer season, this year. For the regions that had 2020 as its worst year we found that two regions had its maximum value of "tas_1Q" this year, one region had its maximum value of "tas_2Q" and one region had the maximum value of both variables. For "Poitou-Charente", the region that has its worst yield in 2011, we found a record value of "tas_2Q" and a minimum record value of "pr_2Q" and "pr_4Q".

As final step, we calculate the correlation coefficient between the highlighted variables in the previous analysis. Equation (1) shows the correlation coefficient, where y_i and $\bar{y}$ are the value of crop yield in the i-th observation, and its average value respectively. x_{ki} represents the i-th value of the k-th independent variable, and $\overline{x_k}$ represents the mean value of the k-th independent variable. Tables 2, 3, 4, show the correlation coefficients obtained after dropping from our datasets the years 2016, 2020 and 2022 respectively, table 5 show the correlation coefficients with all the years of our dataset.

$$Correlation\ Coefficient = 1 - \frac{\sum_i^N (y_i - \bar{y})(x_{ki} - \overline{x_k})}{\sqrt{\sum_i^N (y_i - \bar{y})^2 (x_{ki} - \overline{x_k})^2}} \quad (1$$

Table 2 Correlation between first look highlighted variables and yield, after dropping year 2016

Variable	Soft Winter Wheat Yield(t/h) correlation.
pr_4Q_t-1	**-0.41**
pr_1Q	**-0.10**
pr_2Q	**-0.40**
pr_3Q	**-0.14**
tas_4Q_t-1	**-0.10**
tas_1Q	**0.11**
tas_2Q	**-0.18**
tas_3Q	**-0.38**
SWI_4Q_t-1	**0.18**
SWI_1Q	**0.05**
SWI_2Q	**-0.30**
SWI_3Q	**-0.25**

Table 3 Correlation between first look highlighted variables and yield, after dropping year 2020

Variable	Soft Winter Wheat Yield(t/h) correlation.
pr_4Q_t-1	**-0.36**
pr_1Q	**-0.12**
pr_2Q	**-0.43**
pr_3Q	**-0.08**
tas_4Q_t-1	**-0.01**
tas_1Q	**0.05**
tas_2Q	**-0.04**
tas_3Q	**-0.35**
SWI_4Q_t-1	**0.14**
SWI_1Q	**0.15**
SWI_2Q	**-0.36**
SWI_3Q	**-0.33**

Table 4 Correlation between first look highlighted variables and yield, after dropping year 2022

Variable	Soft Winter Wheat Yield(t/h) correlation.
pr_4Q_t-1	**-0.36**
pr_1Q	**-0.10**
pr_2Q	**-0.46**
pr_3Q	**-0.04**
tas_4Q_t-1	**-0.06**
tas_1Q	**0.01**
tas_2Q	**-0.07**
tas_3Q	**-0.35**
SWI_4Q_t-1	**0.15**
SWI_1Q	**0.10**
SWI_2Q	**-0.38**
SWI_3Q	**-0.37**

Table 5 Correlation between first look highlighted variables and yield

Variable	Soft Winter Wheat Yield(t/h) correlation.
pr-4Q_t-1	**-0.37**
pr_1Q	**-0.09**
pr_2Q	**-0.45**
pr_3Q	**-0.07**
tas_4Q_t-1	**-0.06**
tas_1Q	**0.02**
tas_2Q	**-0.08**
tas_3Q	**-0.35**
SWI_4Q_t-1	**0.15**
SWI_1Q	**0.11**
SWI_2Q	**-0.35**
SWI_3Q	**-0.33**

We see in these tables that the aggregated period 1Q, winter season for us, is the period with less correlation with soft winter wheat crop yield. According with "S.C.A VIVESCIA", a French cereal cooperative group, this is an early period on the winter wheat lifetime. Other features we can extract from the tables are absence of correlation of the precipitations during period 3Q, summer, with yield, and the moderated correlation of the soil wetness index with yield.

Finally, we calculate the average values of these variables and run a correlogram to analyse how a percentual change over the average values of this variables are correlated between them. Figure 1 shows the correlation between variables “pr” and “SWI” percentual variation over its mean value, and the Figure 2 shows the correlation between variables “tas” and “SWI” percentual variation over its mean value. We observe how the precipitations in a given season are strongly correlated to the SWI of the next season. This is not appreciated with temperatures and SWI, where we only can see something similar with temperatures and SWI in the 2Q and 3Q aggregation periods, respectively. Air surface temperatures and SWI are more correlated in the same season.

	var_pr_4Q_t-1(%)	var_pr_1Q(%)	var_pr_2Q(%)	var_pr_3Q(%)	var_swi_4Q_t-1(%)	var_swi_1Q(%)	var_swi_2Q(%)	var_swi_3Q(%)	var_yield(%)
var_pr_4Q_t-1(%)	1.000000	0.160000	-0.250000	0.100000	0.450000	0.790000	0.120000	-0.150000	-0.170000
var_pr_1Q(%)	0.160000	1.000000	0.320000	0.050000	0.200000	0.460000	0.670000	0.470000	-0.220000
var_pr_2Q(%)	-0.250000	0.320000	1.000000	-0.280000	-0.160000	-0.160000	0.670000	0.780000	-0.140000
var_pr_3Q(%)	0.100000	0.050000	-0.280000	1.000000	0.030000	0.120000	-0.220000	0.130000	-0.090000
var_swi_4Q_t-1(%)	0.450000	0.200000	-0.160000	0.030000	1.000000	0.590000	0.180000	0.040000	-0.170000
var_swi_1Q(%)	0.790000	0.460000	-0.160000	0.120000	0.590000	1.000000	0.300000	0.030000	-0.180000
var_swi_2Q(%)	0.120000	0.670000	0.670000	-0.220000	0.180000	0.300000	1.000000	0.670000	-0.170000
var_swi_3Q(%)	-0.150000	0.470000	0.780000	0.130000	0.040000	0.030000	0.670000	1.000000	-0.120000
var_yield(%)	-0.170000	-0.220000	-0.140000	-0.090000	-0.170000	-0.180000	-0.170000	-0.120000	1.000000

Figure 1 Correlation analysis

	var_tas_4Q_t-1(%)	var_tas_1Q(%)	var_tas_2Q(%)	var_tas_3Q(%)	var_swi_4Q_t-1(%)	var_swi_1Q(%)	var_swi_2Q(%)	var_swi_3Q(%)	var_yield(%)
var_tas_4Q_t-1(%)	1.000000	0.110000	-0.250000	0.070000	-0.130000	-0.060000	-0.050000	0.090000	0.400000
var_tas_1Q(%)	0.110000	1.000000	-0.080000	0.080000	0.010000	0.120000	0.170000	0.080000	-0.180000
var_tas_2Q(%)	-0.250000	-0.080000	1.000000	0.160000	-0.000000	0.030000	-0.450000	-0.680000	-0.150000
var_tas_3Q(%)	0.070000	0.080000	0.160000	1.000000	-0.180000	-0.320000	-0.000000	-0.330000	0.110000
var_swi_4Q_t-1(%)	-0.130000	0.010000	-0.000000	-0.180000	1.000000	0.590000	0.180000	0.040000	-0.170000
var_swi_1Q(%)	-0.060000	0.120000	0.030000	-0.320000	0.590000	1.000000	0.300000	0.030000	-0.180000
var_swi_2Q(%)	-0.050000	0.170000	-0.450000	-0.000000	0.180000	0.300000	1.000000	0.670000	-0.170000
var_swi_3Q(%)	0.090000	0.080000	-0.680000	-0.330000	0.040000	0.030000	0.670000	1.000000	-0.120000
var_yield(%)	0.400000	-0.180000	-0.150000	0.110000	-0.170000	-0.180000	-0.170000	-0.120000	1.000000

Figure 2 Correlation analysis

After analyse the correlations, we continue our data mining process with more advanced feature selection tools. First, we will explain how works each algorithm and then we will show the results provided by each of them.

2.2.1 Minimum Redundancy Maximum Relevance (mRMR)

To explain how this algorithm, introduced by Pen et al. (2005), works, first we must define mutual information (MI):

$$I(x,y) = -\frac{1}{2} ln(1 - p(x_i,y)^2) \quad (3)$$

Where $\boldsymbol{p(x_i,y)}$ denotes the correlation between the i-th feature and the target variable. Then, we have a target variable (y) and a set of features $X = \{X_0, X_1, .., X_n\}$. This set is rank based in MI, and the feature with the highest MI is which initialize the set (S) of selected features. After that, the next step is to add a new variable with the highest relevance with the target variable and the lowest redundancy with the previous selected feature/s, maximizing the score show in the equation (4):

$$q_i = I(x_i,y) - \frac{1}{|S|} \sum_{k \in S} I(x_i,x_k) \quad (4)$$

We repeat this step till reach the desired length of (S).

2.2.2 Principal Component Analysis (PCA)

PCA consists in to build a covariance matrix such as shown in equation (5). Where the diagonal are the self-covariance values and the rest of the matrix is formed by the covariance between variables.

$$M(X,Y) = \begin{matrix} Cov(X,X) & Cov(Y,X) \\ Cov(X,Y) & Cov(Y,Y) \end{matrix} \quad (5)$$

Given this matrix we can obtain as eigenvalues as variables we have. Then, the eigenvector V_i of λ_i indicates the i-th principal component, and, therefore, vector V_i will capture more variability than vector V_j for i < j.

2.2.3 Boruta algorithm

Random Forest feature importance is based in the contribution of the features to the reduction of the tree's impurity, estimated by the variance of the predictions within each node of a tree. When a split is created in a tree, and this new split produces an important reduction in the average variance between the nodes of the tree, we can interpret that the feature used in this split was

relevant or important. As a random forest is a set of trees, we calculate the average variation reduction that implies each feature in the entire forest.

The Boruta algorithm is a sort of variant of the RF feature importance that generates more robust results. The algorithm generates a copy, $\check{x}_i$, of each independent variable, x_i, column of our dataset, changing randomly the order of its values, and then, test the Random Forest feature importance. If the importance of the original variable, x_i, is lower or close to its copy $\check{x}_i$, the algorithm interprets that the variable is not relevant.

2.2.3 Naïve Bayesian Classifier (NBC)

To use NBC, we first transform the continuous dependent variable of our dataset, the crop yield(y), into five discrete categories. NBC is a conditional probability model that calculates the probabilities of a given vector, x, to belong to a determinate class, k. As show in the equation (6):

$$p(C_k|X) = \frac{p(X|C_k)\, p(C_k)}{p(X)} \quad (6)$$

2.3 Feature selection results

Table 6 show the results obtained after running the previous introduced feature selection tools. The length of variables is larger in the Boruta algorithm than the other tools due to the characteristics of the algorithm, that removes the irrelevant features. For mRMR, PCA and NBC we fix the desired number of relevant variables to 10.

Table 6 Relevant Variables by each tool criteria

Feature Selection Tool	**Selected set of relevant variables.**
mRMR	**[tasmax_4Q_t-1, gslstart, SWI_3Q, tas_1Q, tas_4Q_t-1, rx5day_3Q, tasmin_1Q, pr-2Q, hurs_2Q, wsdi]**
PCA	**[gslstart, wsdi, SWI_4Q_t-1, td_2Q, rx5day_3Q, rx5day_1Q, csdi, pr_1Q, pr_3Q, tas_3Q]**
NBC	**[rx5day_1Q, pr_1Q, tas_1Q, tasmin_1Q tasmax_1Q, rx5day_2Q, hurs_3Q, tas_4Q_t-1, tasmin_4Q_t-1, tasmax_4Q_t-1,]**
Boruta	**[rx5day_1Q, pr_1Q, tasmin_1Q, pr_2Q, tas_2Q, tasmin_2Q, pr_3Q, hurs_3Q, tas_4Q_t-1, tasmax_4Q_t-1, gslstart, gslend, csdi, SWI_3Q]**

mRMR, NBC and Boruta algorithms, provide a list with of the relevant variables as an output, the PCA provides a matrix composed by vectors that represent each principal component. To estimate the importance of each feature we calculate the Euclidean distance of each variable from the origin in the 10-dimensional principal component space.

3 Models and methodology

3.1 Panel Regression

As commented above, we start our research analysing the power of the statistical tools. We build a fixed effects panel four-degree polynomial regression. As a fixed effects regression means to assume the existence of a constant in time unobserved variable, different for each individual, in this case, each region, we apply the so call "within method" to remove this unobserved effect from our data. This method consists in centre the data before estimate the coefficients of the regression. This also implies to remove all the constant variables of our dataset as for example, geographical coordinates of each region.

A good form of introduce the within method is starting from one independent variable regression as in the work of J.M. Wooldridge (2012). Equation (7) shows a regression where "y_{it}" is the predicted value of the dependent variable for the i-th individual at period t. The expression "$\beta_1 x_{it}$" represents the coefficient of the independent variable and the value of the variable for the i-th individual at period t. The expression "α_i" represents the unobserved fixed effect of the i-th individual, and the expression "u_{it}" represents the error term for the i-th individual prediction at period t. In this case, our individuals are the French regions and or time length starts in 2010 and finish in 2022.

Equation (8) shows the process of data centring, where: $\bar{y}_i = \frac{\sum_{t=1}^{T}(y_{it})}{T}$, $\bar{x}_i = \frac{\sum_{t=1}^{T}(x_{it})}{T}$, $\bar{u}_i = \frac{\sum_{t=1}^{T}(u_{it})}{T}$, $\bar{\alpha}_i = \frac{\sum_{t=1}^{T}(u_{it})}{T} = \alpha_i$.

Equation (9) shows the final result. The dotted hat represents the difference between the mean value of the independent variable, the error term and the dependent variable, and its values for each period t. After equation (9) is calculated, we obtain the coefficients for each independent variable and the intersection term by OLS and build our four-degree polynomial regression.

$$y_{it} = \beta_1 x_{it} + \alpha_i + u_{it} \; ; t = 1, 2, \dots, T \quad (7)$$

$$y_{it} - \bar{y}_i = \beta_1 (x_{it} - \bar{x}_i) + (\alpha_i - \bar{\alpha}_i) + (u_{it} - \bar{u}_i) \; ; t = 1, 2, \dots, T \quad (8)$$

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it} + \ddot{u}_{it} \; ; t = 1, 2, \dots, T \quad (9)$$

3.2 Random Forest.

We use the Scikit-Learn library for Python to build the Random Forest model. We needed 200 decision trees to get the best result. The "random_state" and "max_depth" parameters are 42 and 15 respectively. As an output we obtain an estimation and a feature importance list based in the relevance of each variable in the decision trees. The feature importance list is similar to the output generated by the Boruta algorithm, but due to the previous commented characteristics of the Boruta

algorithm we only had into account the output provided by this algorithm. We apply the within method to the data before train and test the model.

3.2 Bayesian Neural Network.

A Bayesian neural network is a probabilistic implementation of a standard neural network with the key difference being that the weights and biases are represented via the posterior probability distributions rather than single point values (Chandra and Simmons, 2024). In our case we combine the standard neural network layers with a final Bayesian layer with the aim of get a distribution as an output instead of a fixed value. Figure 3 represents a visualization of our BNN models.

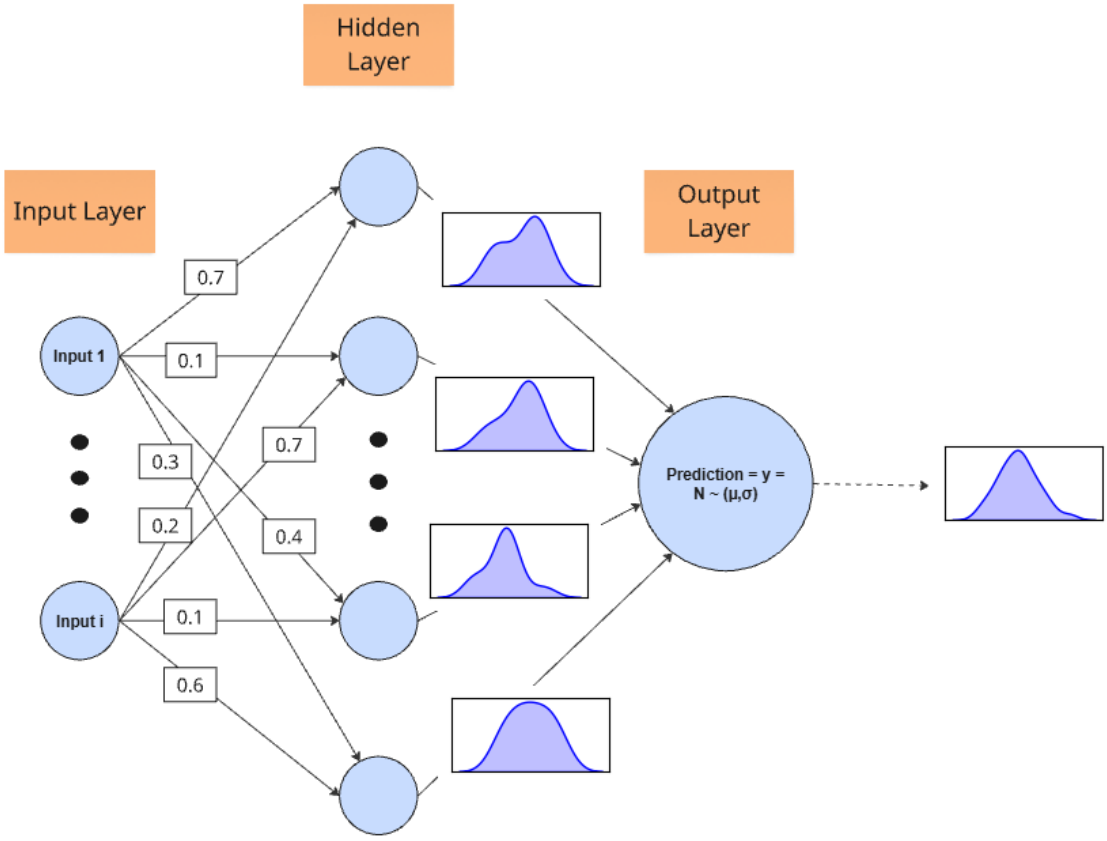

Figure 3 Hybrid Bayesian Neural Network

Markov Chain Monte Carlo (MCMC) is the most common method for bayesian inference, and for sampling multi-modal probability distributions (Gunapati et al. 2021). This method provides a straightforward approach to numerically estimate uncertainties in the parameters of a model using a sequence of random samples (Speagle, 2020). In some situations, where is not possible to determine the analytical solution for a BNN or use the MCMC methods, you need to use techniques to approximate the Bayesian model (Dürr and Sick, 2020). Variational inference (VI), an alternative faster method to MCMC introduced by the Google DeepMind scientists Blundell et al. (2015) is the method used to update the weights' distribution of our BNN models.

Our dataset contains 273 observations. We use 246 observations to train the model, 13 for the validation process and our test set is formed by 14 observations. Train, validation and test sets were selected by randomly combinations till find the best results.

4 Results

We use as metrics the correlation coefficient(R^2), equation (10), to evaluate the goodness-of-fit of the predictions versus the real values, and the Root Mean Squared Error (RMSE), equation (11), to analyse the distance between the predicted and real values. RMSE is largely used in the literature and makes easy to compare our results with other works.

$$R^2 = 1 - \frac{\sum_i^N (y_i - \hat{y}_i)}{\sum_i^N (y_i - \bar{y})} \quad (10)$$

$$RMSE = \frac{\sqrt{\sum_i^N (y_i - \hat{y}_i)^2}}{N} \quad (11)$$

Thanks to the data analysis task, and the feature selection tools and the properties of BNN, we finally built a model that provide us the best results with a reduced set of independent variables. Table 7 shows the main results of this research. The number of variables of the Polynomial Regression includes the fourth polynomial version of each variable with exception of "Growing season length start", "Growing season length end", "Warm spell duration index" and "Cold spell duration index". The RF model contains these previous commented variables, and all variables for "spring" and "summer" seasons, in addition to the soil wetness index for the "autumn" season. Finally, the BNN model contains six variables: precipitation accumulation, average air surface temperature and the SWI, all of them for the seasons of "spring" and "autumn".

Table 7 Models analysis

Model	**R^2**	**RMSE**	**Number of independent variables**
Polynomial Regression	0.8978	0.2323	117
Random Forest	**0.9715**	0.2378	20
Bayesian Neural Network	0.7122	**0.2291**	**6**

In addition to a prediction, BNN also provides a 95% confidence band that is another important feature of this tool. In Figure 4 we can see the output of our BNN, where the 78.57% of the test dataset values are within this confidence band, with observations 10, 12 and 13 as the only outbounds observation.

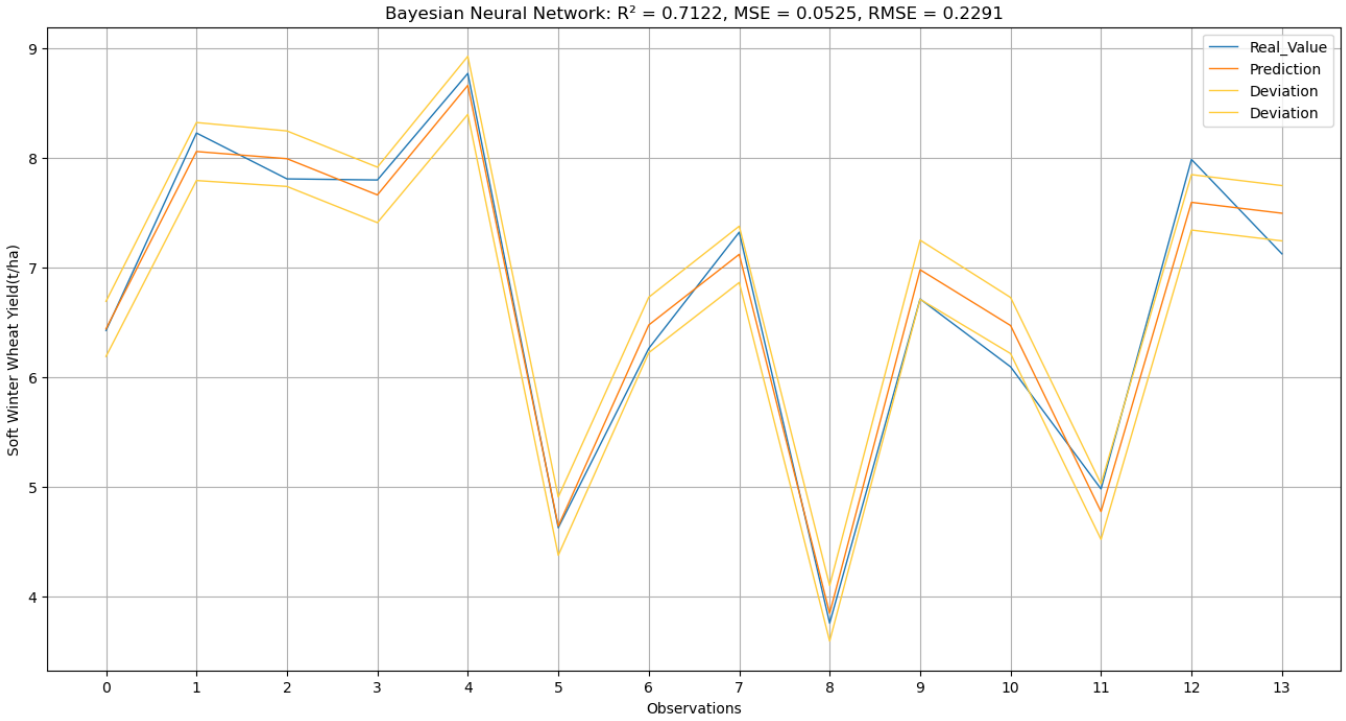

Figure 4 Bayesian Neural Network output

5 Conclusions and future work

The results follow the line of the literature, with similar random forest and polynomial regression producing similar results and the advanced machine learning regression providing the best results. We have shown how a previous deep data analysis with a data mining process, to reduce the length of relevant variables, and the properties of Bayesian Neural Networks, can reduced the uncertainties linked in the agricultural production. It is striking that the final selected set of variables to predict the "Soft Winter Wheat" through our BNN model is formed for "spring" and "summer" aggregated period variables. This could be explained by, first, the observed in the data analysis. We see how the variable SWI captures information from itself and other variables in previous periods. Secondly, as seen in the information provided by "S.C.A VIVESCIA", in our aggregated period "4Q_t-1" is the planting season, and in the period "1Q" the tip of the ear reaches one cm above the shoot apex, inside the stem. Therefore, the periods "2Q" and "3Q", when the main phases of the crop growth and the harvest occur, are more relevant.

Despite the main literature focused in crop yield analysis and forecasting are studies working in a lower dimension, the results reinforce the idea that regional dimension is adequate to analyse the crop yield and build new tools for policy makers and producers. The fact of find the worsts crop yields in 13 years for 21 regions, concentrated in four years, also could reinforce this idea. This evidences that climatological events can affect in the same way to more large territory than a department, in the case of France.

For future works, we are evaluating the possibility of develop an insurance system based in the BNN properties and an allocation planner algorithm. The output generated by the BNN opens the door to analyse new ways of study the risk uncertainty and develop modern insurance contracts where insurer and insured have the same information and tools to evaluate their investment. On the other hand, the changing climate patrons and the search for sustainable development invite us to develop models to face allocation problems in this context.

6 References

Agreste Statistical Book 2023.
https://www.agreste.agriculture.gouv.fr/agreste-web/download/publication/publie/MemSta2023-en/Handbook2023.pdf

Chandra, R., Simmons, J. 2024. "Bayesian Neural Networks via MCMC: A Python-Based Tutorial," in *IEEE Access*, vol. 12, pp. 70519-70549. 10.1109/ACCESS.2024.3401234

Blundell,C., Cornebise, J., Kavukcuoglu, K., Wiestra, D. 2015. Proceedings of the 32nd International Conference on Machine Learning, Lille, France. JMLR: W&CP volume 37.

Dürr, O., Sick, Beate. 2020. Probabilistic Deep Learning: With Python, Keras, and Tensorflow Probability. Manning Publications.

FAO. 2024. World Food and Agriculture – Statistical Yearbook 2024. Rome. https://doi.org/10.4060/cd2971en

Gammans, M., Mérel, P., Ortiz-Bobea, A. 2017. Negative impacts of climate change on cereal yields: statistical evidence from France. Environmental Research Letters. 12. 054007. https://iopscience.iop.org/article/10.1088/1748-9326/aa6b0c

Golerova, I., Bellini, F., D'Ascenzo, F. 2024. Understanding smart territories : A conceptual framework .Cities. 152. 105146. https://doi.org/10.1016/j.cities.2024.105146

Gunapati, G., Jain, A., Srijith, P. K., & Desai, S. (2022). Variational inference as an alternative to MCMC for parameter estimation and model selection. *Publications of the Astronomical Society of Australia*, *39*, e001. doi:10.1017/pasa.2021.64

Li M. Cheng, Zhixun Su, Bo Jiang. Mathematical Problems in Data Science: Theoretical and Mathematical Methods. 2015. Springer.

Ma, Y., Zhang, Z., Kang, Y., & Ozdogan, M. (2021). Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach. *Remote Sensing of Environment, 259*, 112408. https://doi.org/10.1016/j.rse.2021.112408

Maya Gopal P S & Bhargavi R. (2019). Selection of Important Features for Optimizing Crop Yield Prediction. *International Journal of Agricultural and Environmental Information Systems (IJAEIS)*, *10*(3), 54-71. https://doi.org/10.4018/IJAEIS.2019070104

Nicolas De Jay, Simon Papillon-Cavanagh, Catharina Olsen, Nehme El-Hachem, Gianluca Bontempi, Benjamin Haibe-Kains, mRMRe: an R package for parallelized mRMR ensemble feature selection, *Bioinformatics*, Volume 29, Issue 18, September 2013, Pages 2365–2368. https://doi.org/10.1093/bioinformatics/btt383

Schwalbert, R.A., Amado, T., Corassa, Geomar., Pott, L.P. 2020. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. Agricultural and Forest Meteorology. 284. 107886. https://doi.org/10.1016/j.agrformet.2019.107886

Speagle, J. S., "A Conceptual Introduction to Markov Chain Monte Carlo Methods", arXiv e-prints, Art. no. arXiv:1909.12313, 2019. doi:10.48550/arXiv.1909.12313.

Sun, J., Di, L., Sun, Z., Shen, Y., & Lai, Z. (2019). County-Level Soybean Yield Prediction Using Deep CNN-LSTM Model. *Sensors*, *19*(20), 4363. https://doi.org/10.3390/s19204363

Klompenburg, T. v., Kassahun, A., Catal, C,. 2020.Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture. 177, 105709. https://doi.org/10.1016/j.compag.2020.105709

Kuradusenge, M., Hitimana, E., Hanyurwimfura, D., Rukundo, P., Mtonga, K., Mukasine, A., Uwitonze, C., Ngabonziza, J., & Uwamahoro, A. 2023. Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize. *Agriculture*, *13*(1), 225. https://doi.org/10.3390/agriculture13010225

Visalakshi, S., Radha, V. 2014. "A literature review of feature selection techniques and applications: Review of feature selection in data mining,". IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, India, pp. 1-6. 10.1109/ICCIC.2014.7238499

Vivescia. "Quel est le cycle du Blé ?". Last seen: April 2025. https://www.vivescia.com/grand-angle/tous/cereale-quel-est-le-cycle-du-ble

Wang, D., Zhang, P., Shuai, C., Zhang, N. 2024. Adaptation to temperature extremes in Chinese agriculture, 1981-2010. Journal of Development Economics. 166. 103196. https://doi.org/10.1016/j.jdeveco.2023.103196

Wooldridge, Jeffrey M., 1960-. (2012). Introductory econometrics: a modern approach. Mason, Ohio: South-Western Cengage Learning.