



HAL
open science

A surrogate model based on feature selection techniques and regression learners to improve soybean yield prediction in southern France

David Camilo Corrales, Céline Schoving, Hélène Raynal, P Debaeke, Etienne-Pascal Journet, Julie Constantin

► To cite this version:

David Camilo Corrales, Céline Schoving, Hélène Raynal, P Debaeke, Etienne-Pascal Journet, et al.. A surrogate model based on feature selection techniques and regression learners to improve soybean yield prediction in southern France. *Computers and Electronics in Agriculture*, 2022, 192, pp.106578. 10.1016/j.compag.2021.106578 . hal-03454504

HAL Id: hal-03454504

<https://hal.science/hal-03454504v1>

Submitted on 29 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A surrogate model based on feature selection techniques and regression learners to improve soybean yield prediction in southern France

David Camilo Corrales ^{a, b *}, Céline Schoving ^c, H  l  ne Raynal ^a, Philippe Debaeke ^a, Etienne-Pascal Journet ^{a, d} and Julie Constantin ^a

^a *Universit   de Toulouse, INRAE, UMR AGIR, F-31326, Castanet-Tolosan, France*

^b *Grupo de Ingenier  a Telem  tica, Universidad del Cauca, Sector Tulc  n, Popay  n, Colombia*

^c *Terres Inovia, Bazi  ge, France*

^d *Universit   de Toulouse, INRAE, CNRS, LIPME, F-31326, Castanet-Tolosan, France*

* Corresponding author. E-mail address: davidcamilo.corralesmunoz@inrae.fr, dcorrales@unicauca.edu.co

Abstract

Empirical and process-based models are currently used to predict crop yield at field and regional levels. A mechanistic model named STICS (Multidisciplinary Simulator for Standard Crops) has been used to simulate soybean grain yield in several environments, including southern France. STICS simulates at a daily step the effects of climate, soil and management practices on plant growth, development and production. In spite of good performances to predict total aboveground biomass, poor results were obtained for final grain yield. In order to improve yield prediction, a surrogate model was developed from STICS dynamic simulations, feature selection techniques and regression learners. STICS was used to simulate functional variables at given growth stages and over selected phenological phases. The most representative variables were selected through feature selection techniques (filter, wrapper and embedded), and a subset of variables were used to train the regression learners Linear regression (LR), Support vector regression (SVR), Back propagation neural network (BPNN), Random forest (RF), Least Absolute Shrinkage and Selection Operator (LASSO) and M5 decision tree. The subset of variables selected by wrapper method combined with regression models SVR ($R^2 = 0.7102$; subset of variables = 6) and LR ($R^2 = 0.6912$; subset of variables = 14) provided the best results. SVR and LR models improved significantly the soybean yield predictions in southern France in comparison to STICS simulations ($R^2 = 0.040$).

Keywords: *STICS; regression learners; filter; wrapper; embedded*

1- Introduction

Soybean (*Glycine max* L.) is grown on 125 million ha worldwide, with a total average production of 340 million tons on the 2016-2020 period (Oil Word, 2020). On an absolute basis, soybean is the fourth most important grain crop after wheat, maize and rice. USA, Brazil, and Argentina are the three most producing countries, accounting collectively for 81 % of the global production (Grassini et al., 2021). EU-27 is a marginal producer (2.6 million tons, <0.8 % of world production), importing *ca.* 95 % of its soybean domestic needs as rich-protein GMO (Genetically Modified Organisms) and non-GMO meals for animal feed. In Europe, France is the second most important producer after Italy (186 500 ha in 2020), the two main producing

43 regions being South-West and Center-East parts with an increasing contribution of organically-
44 grown production. One main objective in France is to achieve self-sufficiency at least in non-
45 GMO soybean meals at 2025 horizon. Therefore, EU-27 members share a common objective:
46 “*reducing markedly the dependency upon soybean imports by developing European*
47 *production*”. Soybean crop requires few pesticides, no N-fertilizer and less irrigation than
48 maize, results in low emissions of greenhouse gases, hence bringing environmental benefits. In
49 addition, it could contribute as a summer crop to the diversification of winter cereal-based
50 systems.

51 Grain yield in France slightly increased since the 80s (Terres Univia, 2021). In 2019, yields
52 were 2.61 t.ha⁻¹ for France, 2.09 t.ha⁻¹ for all Europe, but 3.19, 3.18, and 3.33 t.ha⁻¹ for USA,
53 Brazil and Argentina respectively (FAOSTAT, 2021). Climate change and its impacts on
54 temperature, precipitation, and CO₂ concentration, but also on water resources available for
55 irrigation, will certainly impact the future production (Porter et al., 2014; Guilpart et al., 2020;
56 Kothari et al., 2020). In addition, expanding soybean growing areas northward and introducing
57 new cropping systems (e.g. double cropping with cereals, rainfed or irrigated soybean, etc.) will
58 change the potential and attainable grain yields.

59 Therefore predicting soybean yield in various environments and a range of cropping systems
60 will be necessary to evaluate the ability of France and European countries to achieve their
61 objectives in terms of protein self-sufficiency by growing more soybean in cropland. Modeling
62 can be efficient in yield analysis and investigation of the limiting factors due to easy
63 manufacturing, testing, applying, understanding and interpretation of results (Nehbandani et al.,
64 2020).

65 Yield prediction models are based on historical or future climate data for evaluating production
66 potentials; also, yield prediction models assimilate remote sensing information when applied to
67 in-season prediction. Nowadays, both statistical and mechanistic approaches are used in
68 agricultural modelling, especially for yield prediction. Statistical approaches search and explore
69 the relations between data to explain the variables of interest whereas mechanistic models are
70 based on the description of biophysical processes. Dynamic crop models simulate daily growth
71 and development in relation with environmental resources and agricultural inputs; they allow
72 the testing of functional hypotheses and the identification of potential constraints to crop growth
73 and yield (Purcell & Roedel, 2019). However, mechanistic and statistical approaches can be
74 combined in order to improve the crop modeling predictions (Casadebaig et al., 2011, 2020).

75 Statistical models from traditional Artificial Neural Networks (ANN) and Deep Learning (DL)
76 have been used for soybean yield prediction. ANN models were proposed by (Kaul et al., 2005)
77 in order to predict Maryland soybean yield at state, regional, and local levels. ANN were
78 developed using historical yield data (1978–1998). Field-specific rainfall data and Soil Rating
79 for Plant Growth (SRPG) values were used for each location. The work developed in
80 (Maimaitijiang et al., 2020) estimated the soybean grain yield through multispectral images
81 (information type: canopy spectral, structure, thermal and texture features) and DLN in
82 Columbia, Missouri. A Convolutional Neural Network (CNN) for soybean yield prediction in
83 15 states of CONUS (United States) is proposed in (Sun et al., 2019). The model was trained
84 by crop growth and environment variables, which include weather data, MODIS Land Surface
85 Temperature data, and MODIS Surface Reflectance data. In Cachoeira do Sul, Brazil, a Multi-
86 Layer Perceptron (MLP) was used to adjust a predictive model for estimating the yield of

87 soybean crop based on 9 vegetation indices (Eugenio et al., 2020). A soybean yield model was
88 created by deep learning framework using CNN and recurrent neural networks (Khaki et al.,
89 2020). Model was built based on environmental data and management practices from Corn Belt
90 (including 13 states) in the United States. In southern Brazil different type of indices as
91 Normalized Difference Vegetation Index (NVDI), Enhanced vegetation index (EVI), land
92 surface temperature (LST) and precipitation were used to build a model using Long-Short Term
93 Memory (LSTM), Neural Networks (Schwalbert et al., 2020). ANN was developed to evaluate
94 the relative importance of predictor variables as vegetation indices (NDVI, red edge NDVI and
95 simple ratio-SR) and elevation derived variables (slope, flow accumulation, aspect) for the
96 prediction of soybean in Ontario, Canada (Kross et al., 2020).

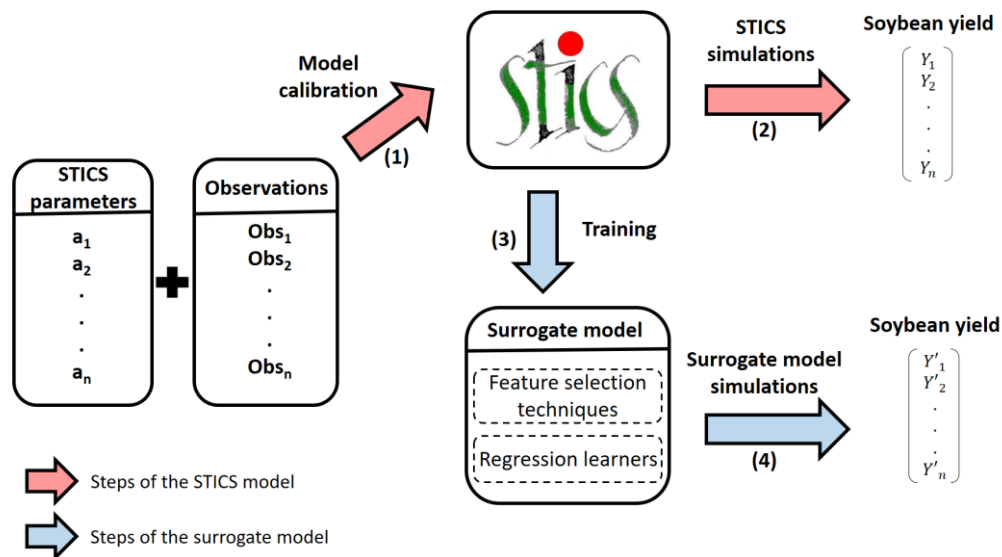
97 Traditional regression models/analysis were used with the same purpose. Authors in (Stepanov
98 et al., 2020) used backward stepwise in order to build a regression model in Khabarovsk,
99 Russia. Several data sources were used as Moderate Resolution Imaging Spectroradiometer
100 (MODIS), arable land mask and meteorological stations. The NDVI was used to improve
101 soybean yield predictions using the flexible Fourier transform model in United States (Xu &
102 Katchova, 2019). The objective was to explore the relationships between soybean yield and
103 number of grains (NG) and thousand grains weight (TGW), generate equations to estimate yield
104 in several countries in the years 2010 and 2019 (Wei & Molin, 2020). A regression analysis
105 was developed by (Ma et al., 2001) in order to study the correlations between plant canopy
106 reflectance and aboveground biomass for early prediction of soybean yield in Canada.

107 In addition, mechanistic models were calibrated in order to predict soybean yield. Authors in
108 (Robertson & Carberry, 1998) used Agricultural Production Systems Simulator (APSIM) with
109 aim to simulate the soybean yield. The model was tested on an independent set of experiments,
110 from northern Australia, with factors such as cultivars, sowing date, irrigation, soil type, plant
111 population density row spacing varying. The research proposed by (Jagtap & Jones, 2002)
112 developed a procedure to simulate soybean yield and production by linking the CROPGRO-
113 soybean model with a regional resolution (about a 50 km grid cell) database of weather, soils,
114 management, and varieties in the state of Georgia over the 1974–1995 time period. The
115 CROPGRO-Soybean model was calibrated to estimate potential yields and yield gaps of
116 soybean for 21 locations regions in India (Bhatia et al., 2008). Authors in (Ovando et al., 2018)
117 simulated soybean yield using a DSSAT model through weather data from Clouds and
118 Earth's Radiant Energy System (CERES) and Tropic Rainfall Measurement Mission (TRMM)
119 for 2006 – 2016 in Oliveros, Argentina. In (Battisti et al., 2017), the model for Nitrogen and
120 Carbon in Agroecosystems (MONICA) was used to simulate soybean grain yield for 14 sites
121 in Southern Brazil. The models CSM-CROPGRO-Soybean and STICS (Multidisciplinary
122 Simulator for Standard Crops) were used to simulate soybean yield responses under near (2041–
123 2070) and distant (2071–2100) future climate scenarios in eastern Canada (Jing et al., 2017).

124 The AgMip initiative started an inter-comparison of 10 soybean crop models at 5 locations in
125 major global production areas with high quality observed data for calibration (Kothari et al.,
126 2020). Among the tested models, STICS (Brisson et al., 2009), a widely used soil-plant crop
127 model applied on a wide range of crops (305 papers in Web of Science), appeared as moderately
128 performing with a prediction gap. Previous attempts to validate STICS on soybean concluded
129 to good performance in aboveground biomass prediction but poor results on grain yield and
130 protein concentration (Schoving, 2020). In eastern Canada, (Jégo et al., 2010) obtained
131 scattered results for biomass, LAI and yield, with Root Mean Square Error (RMSE) from 23 to

132 38 %. Heretofore, this model has been used less in soybean crops, but encouraging results were
 133 recently obtained through proper calibration on pea and faba bean, two other grain legumes
 134 (Falconnier et al., 2019, 2020). In addition, STICS describes with a good accuracy the dynamics
 135 of water and nitrogen in soils and a module was introduced for considering explicitly N₂-
 136 fixation in legumes.

137 Improving the STICS prediction for soybean could imply a thorough calibration of the model
 138 and/or a deep re-examination of the underlying biophysical processes. The experimental data
 139 required for such an improvement could be insufficient. The data-driven modeling approach
 140 has emerged as an alternative to model the biophysical system purely from the data available.
 141 A data-driven model, also known as a meta-model or surrogate model, is a “model of the model”
 142 (Cui et al., 2016). A surrogate model is a statistical model trained from simulations/variables
 143 difficult to measure in field (e.g. leaf area index, aboveground crop biomass, N crop uptake,
 144 crop transpiration, etc.). The surrogate model can be deployed to replace or support the original
 145 biophysical simulation module to accurately approximate the simulation output. Figure 1
 146 explains the interactions between surrogate and STICS model.



147
 148 **Figure 1.** Steps of STICS and surrogate model to predict the soybean yield. The red arrows show the process of soybean yield
 149 simulations generated by STICS. (1) Parameters are calibrated and field observations are used to run STICS. (2) Soybean yield
 150 simulations $Y_1, Y_2 \dots, Y_n$ generated by STICS. The blue arrows depict the steps to create the surrogate model. (3) The STICS
 151 simulations are used to select the relevant variables and train the surrogate model. (4) Soybean yield simulations $Y'_1, Y'_2 \dots, Y'_n$
 152 generated by the surrogate model.

153 In this sense, we proposed a surrogate model based on feature selection techniques and
 154 regression learners to predict soybean yield in southern France. The surrogate model is trained
 155 from the data produced by STICS simulations generated by (Schoving, 2020) (effects of
 156 climate, soil and management practices on dynamic variables of soybean crop functioning) to
 157 improve the prediction of final grain yield. This study progressed through three steps:

- 158 a) Calculate crop variables at different phenological stages with STICS as evaluated by
- 159 (Schoving, 2020) in southern France.
- 160 b) Find the representative variables of soybean yield based on feature selection techniques.
- 161 c) Build a regression model to predict soybean yield based on representative variables found
- 162 by feature selection techniques.

163 **2- Materials and methods**

164 **Multidisciplinary Simulator for Standard Crops (STICS)**

165 The STICS model simulates at a daily step the effects of climate, soil and management practices
166 on plant growth, development and production (quantity and quality) and environmental
167 impacts. The combination of these input variables is termed a USM (Unit of SiMulation). Each
168 USM corresponds to one execution of the STICS model (Brisson et al., 1998). STICS can be
169 tuned to a single crop, two intercropped or several successive crop cycles. STICS has been
170 evaluated over a large data set for 15 different crops and different conditions of soil and climate
171 in France (Coucheney et al., 2015).

172 In order to calibrate STICS, the crop files contain species parameters, ecophysiological options
173 (e.g. effect of photoperiod and/or cold requirements on crop phenology, potential radiation use
174 efficiency) and cultivar specific parameters (e.g. flowering precocity, maximum number of
175 grains per m²). Crop temperature (calculated from weather variables) and photoperiod drive
176 crop phenology. The model dynamically simulates (i) the development of the root system that
177 takes up N and water according to root density over the whole soil profile and (ii) the
178 establishment of the canopy that transpires water and intercepts light to produce the crop
179 biomass (Brisson et al., 2009).

180 **Study area and datasets**

181 The data used in this work were collected by (Schoving, 2020). Seventeen experimental sites
182 were conducted during 2010-2018 from six regions in the south of France: Mauguio (2010),
183 Béziers (2010 - 2012), Mondonville (2010 - 2014), Rivières (2010 - 2014), En Crambade (2013
184 - 2014) and Auzeville (2017 - 2018) as shown in Figure 2.



185
186 **Figure 2.** Locations in Southern France where experimental sites were conducted. Locations are depicted by red markers
187 (Mauguio, Béziers, Mondonville, Rivières, En Crambade and Auzeville). This figure was created by Google Earth.

188 The eleven tested soybean varieties belonged to four maturity groups corresponding to different
 189 crop durations and potential yields: 000 (very early-maturing), 0, I and II (late-maturing). Three
 190 late-maturing varieties were tested in all experiments since 2010 (Ecuador, Santana, Isidor and
 191 Sarema). Detailed information on varieties and maturity groups are presented in Table A1
 192 (Appendix A). Weather data were collected near to the experimental sites. Soil samples contain
 193 texture and physico-chemical analyzes. These data are essential to correctly initialize STICS
 194 with realistic values of soil moisture and mineral nitrogen (nitrate, ammonium). The water
 195 pressure of the soils was monitored in a micro-plot of Santana variety, at 30, 60 and 90 cm
 196 depth.

197 The dataset contains 227 simulation units (USM) created from combination of experimental
 198 sites, years and cropping practices (cultivar, water management and sowing date); We used the
 199 same training (105 USMs) and test (122 USMs) datasets as defined by (Schoving, 2020). The
 200 train-test split was defined based on the number of variables measured in field by experimental
 201 sites (variables such as phenology, biomass, leaf area index, grain yield and seed protein
 202 content). Most complete observations measured in field were selected to train the surrogate
 203 model while data with less observations were retained for the test set. We preprocessed 87
 204 variables based on agronomist knowledge from a selection of 19 STICS state variables
 205 calculated daily by the model during its simulation (Table 1). The preprocessed variables
 206 concern either state variables at different crop phenology stages (Ilev: emergence, Idrp: grain
 207 filling onset, Iflo: flowering, Imat: physiological maturity), descriptive values (Cum:
 208 cumulative, Max: maximum, Avg: average) and thresholds set for variables as MinTemp (days
 209 MinTemp < 18°C), MaxTemp (days MaxTemp > 28°C), Swfac (days Swfac < 0.6) and Inn
 210 (days Inn < 0.6). The soybean grain yield is represented by variable Mafruit (Table 1; dependent
 211 variable).

STICS variables	Units	Description
Nbgrmax	m ²	Maximal grain number
Stlevdrp	°C.days	Heat sum from emergence to grain filling onset
Stfloodrp	°C.days	Heat sum from flowering to grain filling onset
Stdrrpmat	°C.days	Heat sum from grain filling onset to physiological maturity
Masec(n)	t.ha ⁻¹	Aboveground crop biomass
Lai(n)	m ² .m ⁻²	Leaf area index
Qnplante	kg.ha ⁻¹	Cumulative amount of N taken up by the crop
Qfix	kg.ha ⁻¹	Cumulative amount of N fixed by the crop
Zrac	m	Water excess stress index on roots
Jul	Julian day	Julian day
Raint	MJ.m ⁻²	Photosynthetic active radiation intercepted by the canopy
Etp(n)	mm.d ⁻¹	Daily potential evapotranspiration
Precip	mm.d ⁻¹	Precipitation
Ep	mm.d ⁻¹	Daily actual transpiration
AvgTemp	°C	Average air temperature
MinTemp	°C	Minimum air temperature
MaxTemp	°C	Maximum air temperature
Swfac	0-1	Stomatal water stress index
Inn	0-2	Nitrogen nutrition index
Mafruit	t.ha ⁻¹	Biomass of harvested organs (grain yield)

Table 1. Variables used from Multidisciplinary Simulator for Standard Crops (STICS).

212

213 For instance, preprocessed variables computed by phenology stages as *Lai(n)_Iflo* variable
 214 indicates the leaf area index *Lai(n)* at flowering (*Iflo*). Other variables are expressed over
 215 phenophases, for example *Precip_cum_Iflo-Imat* represents the cumulative precipitation
 216 (*Precip*) between two phenological stages: flowering (*Iflo*) and physiological maturity (*Imat*).
 217 In addition, we define variables by phenophases and thresholds as *Days_MinTemp_18_Ilev-*
 218 *Imat* variable which indicates the number of days when minimum temperature was less than
 219 18°C between emergence and physiological maturity stages. Table 2 presents a summary of
 220 preprocessed variables from STICS by stages or phases, descriptive values and thresholds.
 221 Table A2 (Appendix A) lists all of the preprocessed variables.

STICS variables	Stages or phases	Descriptive value	Threshold	Number of variables
Nbgrmax	-	-	-	1
Stlevdrp	-	-	-	1
Stflodrp	-	-	-	1
Stdrpmat	-	-	-	1
Masec(n)	Iflo, Idrp, Imat	-	-	3
Lai(n)	Iflo, Idrp, Imat	-	-	3
Qnplante	Iflo, Idrp, Imat	-	-	3
Qfix	Iflo, Idrp, Imat	-	-	3
Zrac	Iflo, Idrp, Imat	-	-	3
Jul	Ilev-Imat, Iflo-Imat, Idrp-Imat, Iflo-Idrp	Cum	-	4
Raint	Ilev-Imat, Iflo-Imat, Idrp-Imat, Iflo-Idrp	Cum	-	4
Etp(n)	Ilev-Imat, Iflo-Imat, Idrp-Imat, Iflo-Idrp	Cum	-	4
Precip	Ilev-Imat, Iflo-Imat, Idrp-Imat, Iflo-Idrp	Cum	-	4
Ep	Ilev-Imat, Iflo-Imat, Idrp-Imat, Iflo-Idrp	Cum	-	4
AvgTemp	Ilev-Imat, Iflo-Imat, Idrp-Imat, Iflo-Idrp	Avg	-	4
MinTemp	Ilev-Imat, Iflo-Imat, Idrp-Imat, Iflo-Idrp	Avg	# days MinTemp < 18	8
MaxTemp	Ilev-Imat, Iflo-Imat, Idrp-Imat, Iflo-Idrp	Avg, Max	# days MaxTemp > 28	12
Swfac	Ilev-Imat, Iflo-Imat, Idrp-Imat, Iflo-Idrp	Avg, Min	# days Swfac < 0.6	12
Inn	Ilev-Imat, Iflo-Imat, Idrp-Imat, Iflo-Idrp	Avg, Min	# days Inn < 0.6	12

222 **Table 2.** Preprocessed variables from Multidisciplinary Simulator for Standard Crops (STICS). STICS variables were
 223 preprocessed by crop phenology stages (Ilev: emergence, Idrp: grain filling onset, Iflo: flowering, Imat: physiological maturity),
 224 descriptive values (Cum: cumulative, Max: maximum, Avg: average) and thresholds (MinTemp, MaxTemp, Swfac and Inn).
 225 Last column (Var) corresponds to number of variables processed from a STICS, phenology stages, descriptive values and
 226 thresholds. Table A2 (Appendix A) lists all of the preprocessed variables.

227 In order to create an interpretable soybean yield model with a minimum number of variables,
 228 we applied feature selection techniques to previous dataset based on three approaches: filter,
 229 embedded and wrapper.

230 Feature selection techniques

231 Feature selection is the process (automatic or manual) of selecting a subset of relevant variables
 232 which contribute most to learner (Corrales et al., 2018). Feature selection techniques can be
 233 grouped in three categories:

- 234 • **Filter methods** are based only on the intrinsic properties of the data (Solorio-Fernández
 235 et al., 2020). Filter method computes an importance value between one independent
 236 variable and the dependent variable. Variables with highest importance values are
 237 selected based on user criteria. Filter methods are usually computationally less
 238 expensive than embedded and wrapper methods. We used classical feature selection

239 methods based on Pearson (Pearson, 1920) and Spearman (Spearman, 1961)
 240 coefficients and Information Gain (Shannon, 1948). In order to explain the statistical
 241 coefficients, independent variable is named X and dependent variable Y.

- 242
- 243 ○ **Pearson coefficient** measures the linear correlation between two variables. If
 244 both variables are linearly dependent, then their correlation coefficient is close
 245 to ± 1 . If the variables are uncorrelated, the correlation coefficient is 0 (Pearson,
 246 1920). When Pearson coefficient is used as filter method only positive values
 247 are considered following the equation (1):

$$248 \quad r = \frac{\left| \sum_{i=1}^n cov(x_i, y_i) \right|}{\sum_{i=1}^n \sigma x_i \sigma y_i} \quad (1)$$

249
 250 Where x_i and y_i are the i^{th} observations of independent and dependent variable
 251 respectively; *cov* corresponds to covariance and σ indicates the standard
 252 deviation of x and y .

- 253
- 254 ○ **Spearman coefficient** through a monotonic function measures the correlation
 255 between two variables (Spearman, 1961). A monotonic function is defined as
 256 function which is either entirely increasing or decreasing. It is similar to Pearson
 257 coefficient except that it operates on the ranks of the data rather than the raw
 258 data (Gauthier, 2001). The Spearman correlation rank is defined by equation (2):

$$259 \quad \rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (2)$$

260
 261 Where d_i is the difference between ranks for each x_i, y_i data pair and n is the
 262 number of data pairs.

- 263 ○ **Entropy-based information gain** discretizes the independent variable and
 264 subsequently the entropy is computed between x and continuous y variable
 265 (Yang et al., 2010) by equation (3):

$$266 \quad InfoGain = H(y) + H(x) - H(y, x) \quad (3)$$

267
 268 Where $H(x)$ and $H(y)$ correspond to Shannon's Entropy for x and y variables.
 269 $H(y, x)$ is a joint Shannon's Entropy for a variable y with a condition to x .
 270 Detailed explanation of Shannon's Entropy is explained in (Shannon, 1948).

- 271
- 272 • **Embedded methods** integrate the variables selection as part of training process into
 273 learner (Guyon & Elisseeff, 2003). We used as embedded methods the learners Random
 274 Forest (Breiman, 2001), M5 decision tree (Quinlan, 1992) and Least Absolute
 275 Shrinkage and Selection Operator (Tibshirani, 1996).
- 276 • **Wrapper methods** selects a subset of variables according to performance criteria
 277 (regression tasks, measure of errors as mean absolute error, mean square error, root
 278 square mean error; classification tasks measures as accuracy, precision, overall, recall,

279 etc.) of a learner (Guyon & Elisseeff, 2003). In this paper, we use the method Recursive
280 Feature Elimination (Guyon et al., 2002).
281

282 **Regression learners**

283 In order to predict soybean yield in southern France crops, we used regression learners from
284 different families of algorithms. They are explained briefly below.

- 285 • **Linear regression (LR)** explains the relationship between dependent variable and one
286 or more independent variables by fitting a linear equation to observed data (Prion &
287 Haerling, 2020). Coefficients multiply the values of dependent values; the coefficient
288 signs represent the direction of the relationship between a dependent variable and the
289 independent variable.
290
- 291 • **Support vector regression (SVR)** is based on same principles as Support Vector
292 Machine (Vapnik, 1995). SVR determines a regression function in the feature space
293 considering only data points within the decision boundary lines called support vectors.
294 In nonlinear data, a kernel function is used in order to transform the feature space into
295 a linear hyperplane (Brereton & Lloyd, 2010).
296
- 297 • **Back propagation neural network (BPNN)** calculates the gradient of the error
298 function with respect to the weights of the neural network (Rumelhart et al., 1986). The
299 computed error is propagated in a backward manner from one layer to the other until
300 the minimum Mean Squared Error (MSE) is attained and weights can be modified
301 accordingly (Deshwal et al., 2020).

302 In addition, we used each of Random forest, Least Absolute Shrinkage and Selection Operator
303 and M5 decision tree as both an embedded method and a wrapper.

- 304 • **Random forest (RF)** builds several decision trees using a different bootstrap sample of
305 data training set (Breiman, 2001). The decision trees are built using CART learner
306 (Breiman et al., 1984). In regression tasks, RF final prediction is obtained by averaging
307 the results of all the CART trees.
308
- 309 • **Least Absolute Shrinkage and Selection Operator (LASSO)** is a linear regression
310 method which imposes a bound on the L_1 -norm of the regression coefficients, resulting
311 in coefficient shrinkage (Tibshirani, 1996). LASSO adds a L_1 penalty equal to the
312 absolute value of the magnitude of coefficients (Equation 1). Variables are discarded
313 when the coefficients take values equal to zero. Larger penalties are expressed by
314 coefficient values closer to zero. The objective function for finding the minimum is
315 shown by Equation (4):

$$316 \text{ minimize } \beta_0, \beta \left(\frac{1}{2N} \sum_{i=1}^N (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \right) \quad (4)$$

317 Where N is the total number of observations, λ is a nonnegative regularization
318 parameter corresponding to one value of Lambda, y_i is the dependent variable, p is the

319 number of independent variables $x_i = (x_{i1}, \dots, x_{ip})^T$, β_0 is the intercept, and β_j are the
 320 coefficients (Shi et al., 2020).

- 321
- 322 • **M5** is a conventional decision tree composed by different nodes such as root,
 323 intermediates and leaves (Quinlan, 1992). Root and intermediate nodes are chosen over
 324 the dependent variable that maximizes the expected error reduction as a function of the
 325 standard deviation of output parameter (Wang & Witten, 1996). Leaves nodes predict
 326 the dependent variable through linear regression functions; they are fitted from data that
 327 follows one branch between root and leaf node.

328 3- Results

329 This section reports the subset of variables selected by feature selection methods and the
 330 evaluation of soybean yield models built from selected variables.

331 3.1 Feature selection

332 3.1.1 Filter methods

333 Filter methods were computed through R 'mlr' package (Bischl et al., 2016). With aim to create
 334 simple models with great explanatory predictive power with a minimum number of features,
 335 we defined two criteria to select the variables: (i) top-15 of variables with highest importance
 336 values for each filter method; (ii) variables selection by threshold based on importance values
 337 of top-15. Regarding to Pearson and Spearman coefficients, we selected the features with
 338 importance value greater than or equal to 0.6. Concerning to entropy-based information gain,
 339 we selected the features with importance value greater than or equal to 0.4. We defined these
 340 thresholds following the “conventional interpretation of the correlation coefficients” proposed
 341 in (Schober et al., 2018). Values between 0.60 - 0.79 are defined as “moderately correlated”
 342 and coefficient values between 0.70 – 0.89 are interpreted as “strongly correlated”. Table 3
 343 shows Top-15 of soybean variables selected by filter methods (Appendix B contains the entire
 344 ranking).

Pos	Pearson coefficient		Spearman coefficient		Entropy-based information gain	
	Variable	Importance	Variable	Importance	Variable	Importance
1	Masec(n)_Imat	0.7923	Lai(n)_Imat	0.7699	Lai(n)_Imat	0.5929
2	Lai(n)_Imat	0.7726	Masec(n)_Imat	0.7311	Masec(n)_Imat	0.5875
3	Qnplante_Ildr	0.7710	Qnplante_Ildr	0.7137	Qnplante_Imat	0.5515
4	Qnplante_Imat	0.7523	Masec(n)_Ildr	0.7003	Ep_cum_Ildr-Imat	0.4223
5	Masec(n)_Ildr	0.7334	Ep_cum_Iflo-Imat	0.6886	Qfix_Imat	0.4072
6	Ep_cum_Ildr-Imat	0.6819	Ep_cum_Ildr-Imat	0.6734	Inn_min_Iflo-Ildr	0.3799
7	Qfix_Imat	0.6805	Qnplante_Imat	0.6714	Inn_avg_Iflo-Ildr	0.3652
8	Ep_cum_Iflo-Imat	0.6650	Ep_cum_Ilev-Imat	0.6456	Ep_cum_Iflo-Imat	0.3486
9	Raint_cum_Ildr-Imat	0.6326	Lai(n)_Ildr	0.6439	Days_Swfac_0.6_Iflo-Imat	0.3443
10	Lai(n)_Ildr	0.6292	Raint_cum_Ildr-Imat	0.6288	Inn_avg_Ilev-Imat	0.3400
11	AvgTemp_avg_Ildr-Imat	0.6157	Swfac_min_Ildr-Imat	0.6160	Inn_avg_Iflo-Imat	0.3380
12	Swfac_avg_Iflo-Imat	0.6088	Swfac_min_Iflo-Imat	0.6136	Masec(n)_Ildr	0.3300
13	MinTemp_avg_Ildr-Imat	0.5937	AvgTemp_avg_Ildr-Imat	0.6081	Qnplante_Ildr	0.3285
14	Ep_cum_Ilev-Imat	0.5842	Swfac_min_Ilev-Imat	0.6009	Raint_cum_Ildr-Imat	0.3080
15	Days_Swfac_0.6_Iflo-Imat	0.5811	MinTemp_avg_Ildr-Imat	0.5977	Days_Swfac_0.6_Iflo-Ildr	0.3025

345 Table 3. Top-15 of variables selected by filter methods: Pearson, Spearman and entropy-based information gain.

346 In this sense, 12 and 14 variables were selected by Pearson and Spearman respectively. The
347 variables *Lai(n)_Imat*, *Masec(n)_Imat*, *Qnplante_Imat*, *Ep_cum_Idrp-Imat* and *Qfix_Imat* were
348 selected by entropy-based information gain. The subdatasets with the selected variables by filter
349 methods were used to train the regression learners presented in Section 2. Besides, RF, LASSO
350 and M5 decision tree act as regression learners and feature selection techniques due these
351 learners are considered embedded methods. In other words, RF, LASSO and M5 are trained
352 with subset of variables selected by filter methods and subsequently the embedded
353 methods/learners select a new subset of variables in the training process into learner. The results
354 are presented in Table 5.

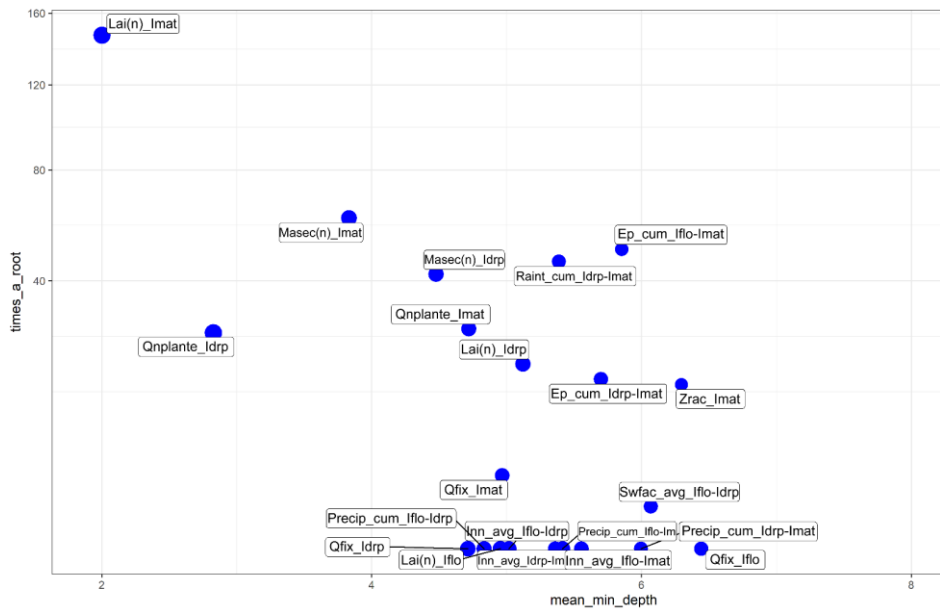
355 **3.1.2 Embedded methods**

356 We used as embedded methods the learners: Random forest, Least Absolute Shrinkage and
357 Selection Operator and M5 decision tree. The variable selection process for each one is
358 explained below.

359 **Random Forest**

360 Random Forest (RF) gathers a set of CART trees in order to obtain the soybean yield prediction
361 by averaging the results of all of trees. We used R packages 'randomForest' to create RF model
362 and 'randomForestExplainer' (Liaw & Wiener, 2002) to design multi-way plot shown in Figure
363 3. Five hundred CART trees were built (ntree parameter) as result, 42 variables were sampled
364 as candidates at each split. The multi-way plot focuses on three importance measures that derive
365 from the structure of trees in the forest: (i) the mean_min_depth variable refers to the depth of
366 first split on the variable to the top of the tree; (ii) the times_a_root variable measures the
367 number of times a variable is set as top of a decision tree. Figure 3 presents multi-way plot for
368 first 15 relevant variables.

369 *Lai(n)_Imat* was the most used variable as top split criterion (148 times) followed by
370 *Masec(n)_Imat* (67 times) and *Ep_cum>Iflo-Imat* (58 times). Variables as *Qfix>Iflo* and
371 *Precip_cum_Idrp-Imat* were never used as top of decision trees and they have the longest
372 distance (mean minimum depth of 6.728 and 6.398 respectively) to the top of decision trees
373 considered less associated with the dependent variable *Mafruit*. Other variables as *Qfix_Idrp*,
374 *Lai(n>Iflo* and *Precip_cum>Iflo-Idrp* can be considered as intermediate nodes of the trees
375 (times_a_root = 0) with mean minimum depth less than *Precip_cum_Idrp-Imat* and *Qfix>Iflo*
376 (mean_min_depth = 4.856, 5.045 and 5.297 respectively).



377

378 **Figure 3.** Multi-way plot between two measures of importance: mean_min_depth (x-axis) and times_a_root (y-axis). First 15
 379 relevant variables are depicted. X-axis correspond to mean depth of first split on the variable, y-axis the number of trees in
 380 which the root is split on the variable.

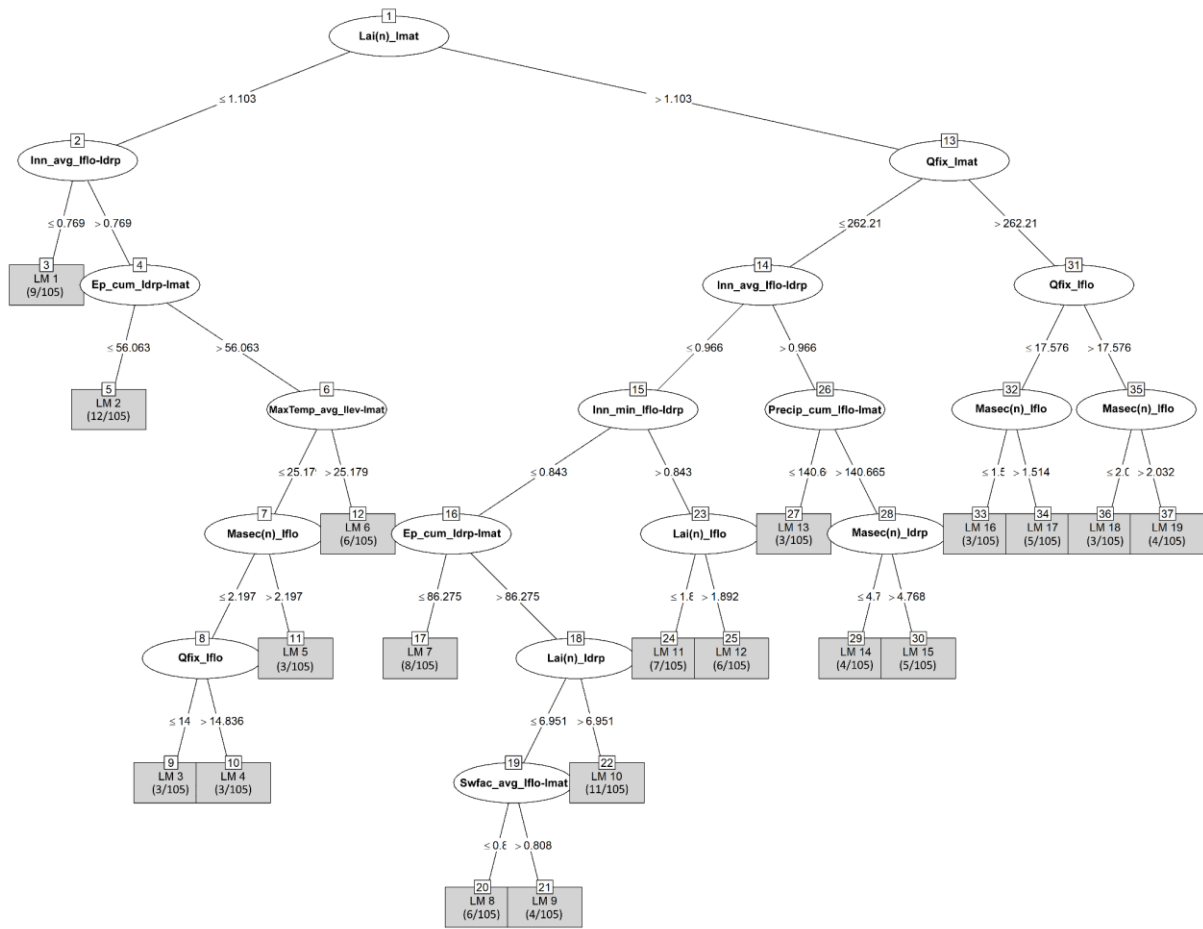
381 Least Absolute Shrinkage and Selection Operator

382 We used the R package 'caret' to build LASSO model (Kuhn, 2008). Root Mean Square Error
 383 (RMSE) was used to select the optimal model using the smallest value. LASSO model was run
 384 with parameters fraction = 0.1 and lambda = 0.01 (Equation 1). The LASSO model set the
 385 regression coefficients of 21 variables to zero by imposing the L1 penalty. Table C1 (Appendix
 386 C) contains the regression coefficients of 66 variables calculated by LASSO.

387

388 M5 decision tree

389 M5 was built by R package 'RWeka' (Hornik et al., 2009). The construction of M5 tree is based
 390 on recursive splitting of the standard deviation of *Mafruit* (dependent variable) that reach a node
 391 as a measure of the error at the node. The variable that maximizes the expected error reduction
 392 is selected for splitting at the node. The expected reduction of the error is obtained as a result
 393 of testing each variable at that node. To remove the problem of over fitting, M5 uses a method
 394 to prune back the over grown tree. Figure 4 shows the structure of the pruned M5 tree to predict
 395 the soybean yield (*Mafruit*) regarding the thirteen selected variables (Table 4). Left branches
 396 show conditions below the node threshold, for example, if the top node *Lai(n)_Imat* ≤ 1.103
 397 and if *Inn_avg_Iflo-Idrp* ≤ 0.769 , the linear model LM 1 must be selected. Thirty-seven nodes
 398 were created including nineteen linear models as decision rules. We defined 3 observations as
 399 minimum number at the leaf node. Appendix C contains the linear models generated by M5.



400

401 **Figure 4.** M5 pruned decision tree. Thirteen variables were selected and 19 decision rules (linear regressions) were created. At
 402 the leaf node, the number in parenthesis corresponds to observations for each linear regression.

403 3.1.3 Wrapper methods

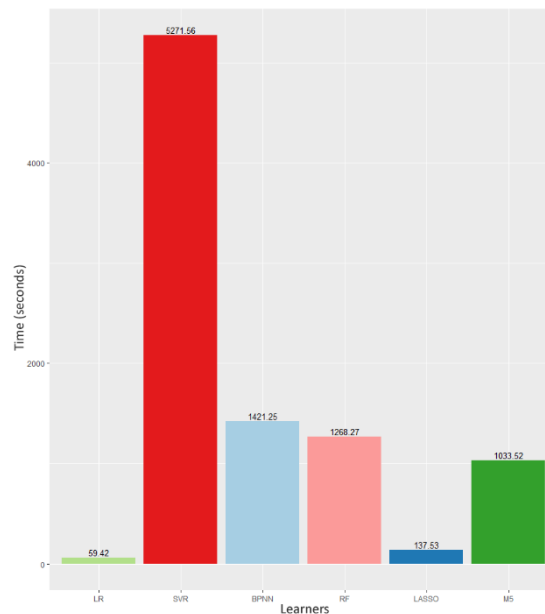
404 A greedy algorithm named Recursive Feature Elimination (RFE) was used as wrapper method
 405 (Kohavi & John, 1997; Guyon et al., 2002). RFE selects a subset of variables to improve a
 406 learner performance by removing the least important features. This process is repeated
 407 recursively (based on backward elimination) until the optimal number of features is obtained.
 408 The recursive step plays a key role due the relative importance of each variable can change
 409 substantially over a different subset of variables during the backward elimination process
 410 (particularly for highly correlated variables) (Granitto et al., 2006). We used the RFE version
 411 coded in R package 'caret' (Kuhn, 2008). Six learners were selected (explained in section 2) to
 412 determine the performance of subsets of variables selected by RFE shown in Table 4. The
 413 variables subset selection were conducted through lowest Mean Absolute Error obtained by
 414 learners. Recursive Feature Elimination selected all of the variables for RF model. The variables
 415 subset for remaining models are less than 17.

Learner	Variables	Selected variables	RMSE
LR	<i>MaxTemp_avg_Ildr-Imat, Qnplante_Ildr, Masec(n)_Imat, Ep_cum_Ildr-Imat, Lai(n)_Imat, Raint_cum_Ildr-Imat</i>	6	0.5865
SVR	<i>Lai(n)_Imat, Masec(n)_Imat, Qnplante_Ildr, Qnplante_Imat, Masec(n)_Idrp, Ep_cum_Iflo-Imat, Ep_cum_Ildr-Imat,</i>	14	0.6750

	<i>Ep_cum_Ilev-Imat, Qfix_Imat, Zrac_Imat, Zrac_Idrp, Lai(n)_Idrp, Raint_cum_Idrp-Imat, MaxTemp_avg_Idrp-Imat</i>		
BPNN	<i>Lai(n)_Imat, Masec(n)_Imat, Qnplante_Idrp, Qnplante_Imat, Masec(n)_Idrp, Ep_cum_Iflo-Imat, Ep_cum_Idrp-Imat, Qfix_Imat, Zrac_Imat, Zrac_Idrp, Ep_cum_Ilev-Imat, Lai(n)_Idrp, Raint_cum_Idrp-Imat, MaxTemp_avg_Idrp-Imat, Swfac_min_Ilev-Imat, Swfac_min_Iflo-Imat</i>	16	1.294
RF	All variables were selected	87	0.4639
LASSO	<i>Lai(n)_Imat, Masec(n)_Imat, Qnplante_Idrp, Qnplante_Imat, Masec(n)_Idrp, Ep_cum_Iflo-Imat, Ep_cum_Idrp-Imat, Qfix_Imat, Ep_cum_Ilev-Imat, Zrac_Idrp, Zrac_Imat, Raint_cum_Idrp-Imat, Lai(n)_Idrp, MinTemp_avg_Idrp-Imat, MaxTemp_avg_Idrp-Imat, Swfac_min_Iflo-Imat</i>	16	6.141e-01
M5	<i>Lai(n)_Imat, Inn_avg_Idrp-Iflo, Ep_cum_Idrp-Imat, MaxTemp_avg_Ilev-Imat, Masec(n)_Iflo, Qfix_Iflo, Qfix_Imat, Inn_min_Iflo-Idrp, Lai(n)_Idrp, Swfac_avg_Iflo-Imat, Lai(n)_Iflo, Precip_cum_Iflo-Imat, Masec(n)_Idrp</i>	13	0.5465

416 **Table 4.** Subset of variables selected by Recursive Feature Elimination (RFE) and base learners: Linear Regression (LR),
417 Support Vector Regression (SVR), M5 decision tree, Random Forest (RF) and Backpropagation Neural Network (BPNN).
418 Learner performance is based on Root Mean Square Error (RMSE).

419 Concerning time complexity, RFE is slower than filter and embedded methods, since RFE
420 needs to evaluate performance criteria for each iteration besides the computational cost of the
421 model training. In this sense, learners based on linear models as LR and LASSO obtained much
422 less computational cost (59.42 and 137.53 seconds) than BPNN, RF and M5 (1421.25, 1268.27
423 and 1033.52 seconds). In contrast to Support Vector Regression which imposed considerable
424 computational cost (5271.56 seconds) due to margin maximization to find the support vectors
425 and nonlinear transformations of the feature space (Yu et al., 2003). Figure 5 presents the time
426 complexity of the variables subset selection by Recursive Feature Elimination and base
427 learners. Wrapper methods were run on Windows 10 comprised of Intel Core i5-774HQ CPU
428 2.80GHz – 16GB RAM based on sequential computing.



429 **Figure 5.** Time complexity to select subset of variables by Recursive Feature Elimination (RFE) and base learners: Linear
430 Regression (LR), Support Vector Regression (SVR), M5 decision tree, Random Forest (RF) and Backpropagation Neural
431 Network (BPNN).
432

3.2 Regression models

In order to examine the performance of subset of variables selected by feature selection methods, we used traditional statistical criteria to estimate the prediction accuracy of regression learners as Coefficient of determination (R^2), Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The metric R^2 was adopted to measure the correlation of the observed and predicted values.

Regression learners were trained with 105 observations and evaluated with 122 examples (explained in section 2). Table 5 presents the results of regression learners trained with subset of variables selected by feature selection methods. The underlined values correspond to statistical criteria obtained by best two learners and feature selection method.

Feature selection approach	Feature selection method	SC	Learners					
			LR	SVR	BPNN	RF	LASSO	M5
Filter	Pearson correlation	R^2	0.6185	0.6389	0.1837	0.3487	0.6274	0.6193
		MSE	0.9085	0.6539	2.1301	1.0325	0.7250	0.9211
		RMSE	0.9531	0.8086	1.4595	1.0161	0.8514	0.9597
		MAE	0.7227	0.6267	1.2375	0.8422	0.6552	0.7258
		No. Selected Variables	12	12	12	12	12	12
	Spearman correlation	R^2	0.5890	0.6394	0.1841	0.3463	0.6101	0.6276
		MSE	1.0418	0.6948	1.4388	1.0818	0.8167	0.6776
		RMSE	1.0207	0.8335	1.1995	1.0401	0.9037	0.8231
		MAE	0.7889	0.6551	1.0158	0.8505	0.6862	0.6397
		No. Selected Variables	14	14	14	14	14	14
	Entropy-based information gain	R^2	0.3974	0.3916	0.2701	0.2622	0.3961	0.2502
		MSE	0.9479	1.0085	4.0698	1.1859	0.9554	1.2713
		RMSE	0.9736	1.0042	2.0173	1.0890	0.9774	1.1275
		MAE	0.7951	0.8215	1.7373	0.8815	0.7975	0.9212
		No. Selected Variables	5	5	5	5	5	5
Embedded	Learners: RF, LASSO and M5	R^2	-	-	-	0.5020	0.1249	0.4010
		MSE	-	-	-	0.8300	6.1841	0.9466
		RMSE	-	-	-	0.9110	2.4867	0.9729
		MAE	-	-	-	0.7258	1.9206	0.7802
		No. Selected Variables	-	-	-	42	66	87
Wrapper	Recursive Feature Elimination (RFE)	R^2	<u>0.6912</u>	<u>0.7102</u>	0.1829	0.5020	0.6718	0.4010
		MSE	<u>0.4807</u>	<u>0.4170</u>	1.4916	0.8300	0.4760	0.9466
		RMSE	<u>0.6933</u>	<u>0.6458</u>	1.2213	0.9110	0.6899	0.9729
		MAE	<u>0.5469</u>	<u>0.5230</u>	0.9746	0.7258	0.5747	0.7802
		No. Selected Variables	6	14	16	87	16	13

Table 5. Results of regression learners trained with subset of variables selected by feature selection methods. Validation dataset (Section 2) was used to evaluate the regression learners. Statistical criteria (SC) used to estimate the performance of regression learners: Coefficient of determination (R^2), Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The underlined values correspond to statistical criteria obtained by best two learners and feature selection method.

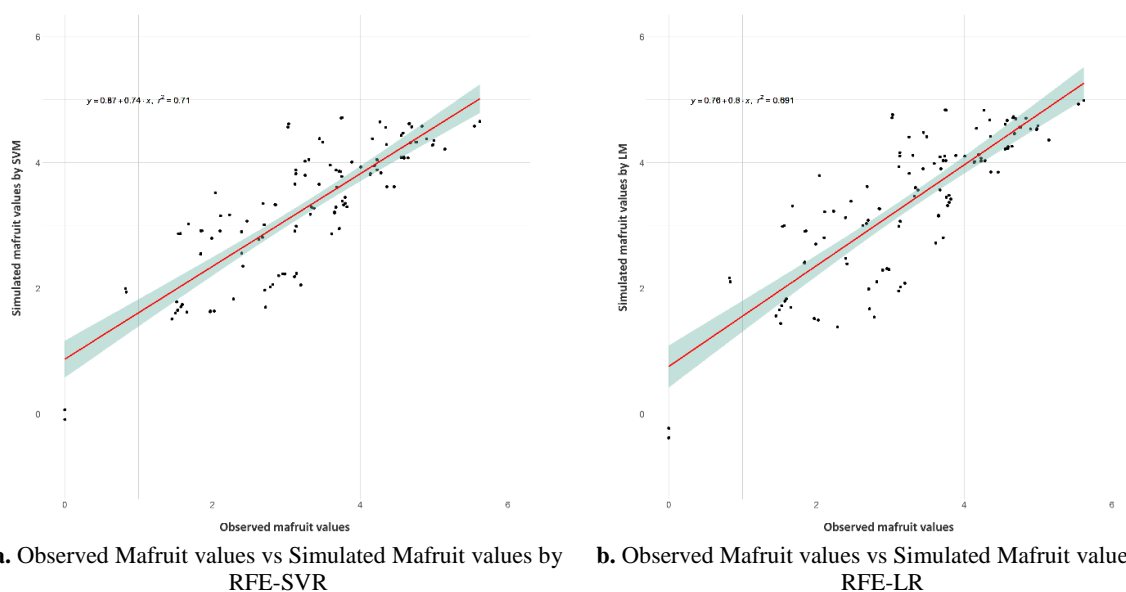
According to filter methods, Pearson, Spearman and Entropy coefficients selected 12, 14 and 5 variables respectively (Section 3). The regression learners LR, SVR and BPNN were trained with subset of variables selected by filter methods; SVR obtained the best results with subset of feature selected by Pearson coefficient ($R^2 = 0.6389$, MSE = 0.6539, RMSE = 0.8086 and MAE = 0.6267). Regression learners RF, LASSO and M5 decision tree act as ensemble

452 approaches for feature selection due these learners are considered embedded methods as well.
 453 In other words, embedded methods were trained with subset of variables selected by filter
 454 methods. Subsequently, embedded methods select a new subset of variables in the training
 455 process into learner. However, the ensemble of filter and embedded methods do not improve
 456 the performance obtained by SVR trained with subset of variables chosen by Pearson
 457 coefficient.

458 The number of variables selected by two embedded methods were higher respecting to others
 459 feature selection approaches. Random Forest and LASSO selected 87 and 66 variables
 460 respectively, whereas M5 decision Tree 13 variables (Section 3). Random Forest reached the
 461 best performance in the ensemble methods ($R^2 = 0.5020$, $MSE = 0.8300$, $RMSE = 0.9110$ and
 462 $MAE = 0.7258$), however RF does not overcome the performance obtained by SVR and Pearson
 463 coefficient in the filter methods.

464 Concerning to wrapper method, Recursive Feature Elimination selected 6, 14 and 16 variables
 465 for LR, SVR and BPNN. SVR achieved the best results compared to all of combination of
 466 feature selection approaches and learners ($R^2 = 0.7102$, $MSE = 0.4170$, $RMSE = 0.6458$ and
 467 $MAE = 0.5230$). Remain of learners that work as ensemble feature selection among wrapper
 468 and embedded methods (RF, LASSO and M5 decision tree), RFE proposes different subset of
 469 variables created in the backward elimination process. The embedded methods are tested with
 470 subset of variables proposed by wrapper method and they select a new subset of variables in
 471 the training process into learner. The best subset of variables are selected regarding to
 472 performance criteria of the embedded method. LASSO trained with 16 variables selected by
 473 RFE (Table 4) reached the best results ($R^2 = 0.6718$, $MSE = 0.4760$, $RMSE = 0.6899$ and
 474 $MAE = 0.5747$) compared to RF and M5 of the wrapper method. Besides, RFE improve the
 475 performance of LASSO compared to LASSO's version of filter and embedded.

476 In summary, the wrapper methods RFE-SVR ($R^2 = 0.7102$, $MSE = 0.4170$, $RMSE = 0.6458$
 477 and $MAE = 0.5230$) and RFE-LR ($R^2 = 0.6912$, $MSE = 0.4807$, $RMSE = 0.6933$ and MAE
 478 $= 0.5469$) achieved the best results from validation dataset. Figure 6 depicts the scatter plots of
 479 observed vs simulated soybean yield values by RFE-SVR and RFE-LR.



480 **Figure 6.** Scatter plot of observed (x-axis) vs simulated (y-axis) Mafruit values by RFE-SVR (14 variables) and RFE-LR (6
 481 variables)

482 For real values of soybean yield equal to zero, RFE-SVR and RFE-LR predict negative soybean
 483 values or close to zero. RFE-SVR simulates Mafruit values equal to 0.066, -0.086 and -0.085
 484 for 3 observations where Mafruit = 0 of the validation dataset (Béziers in 2011 with maturity
 485 group I in varieties Isidor, Santana and maturity group II in Ecuror variety). Similarly, LR-RFE
 486 predicts -0.227, -0.379 and -0.378 for same observations of validation dataset.

487 3.3 Comparative study

488 In order to demonstrate the performance of RFE-SVR and RFE-LR, we compared both
 489 regression models against STICS simulations developed in (Schoving, 2020). The soybean
 490 yield model proposed by Schoving was calibrated with the same training set as presented in
 491 section 2. Table 6 shows the results of RFE-SVR, RFE-LR and STICS models evaluated from
 492 validation dataset (Section 2).

Models	Statistical criteria	
	R ²	RMSE
RFE-SVR	0.710	0.645
RFE-LR	0.691	0.693
(Schoving, 2020)	0.040	1.320

493 **Table 6.** Comparison of Support Vector Regression and Linear Regression (LR) trained with subset of variables selected by
 494 Recursive Feature Selection (RFS) vs soybean yield model proposed by (Schoving, 2020). Validation dataset (Section 2) was
 495 used to evaluate the models. Statistical criteria used to estimate the performance of models: Coefficient of determination (R²)
 496 and Root Mean Square Error (RMSE).

497 The two regression models explained around 70 % of the grain yield variation and they achieved
 498 half of RMSE values obtained for STICS simulations developed in (Schoving, 2020). Although
 499 total aboveground biomass was correctly simulated by (Schoving, 2020) ($r^2 = 0.64$), final grain
 500 yield of semi-indeterminate and indeterminate soybean cultivars was poorly represented (Table
 501 6; $r^2 = 0.04$). This is probably because STICS uses the standard formalism of wheat and maize
 502 crops to simulate the final grain yield in soybean.

503 4- Discussion

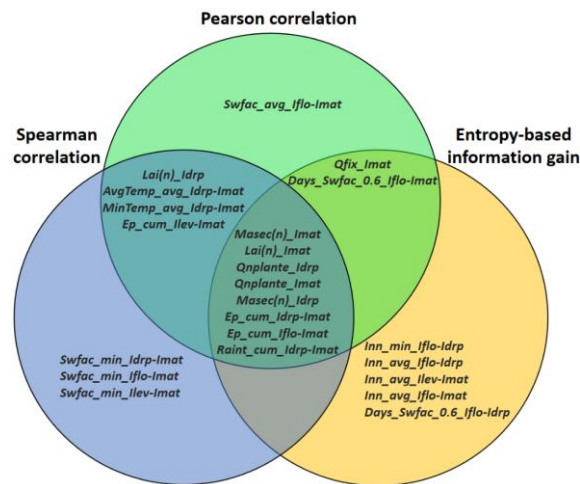
504 Feature selection methods are relevant in order to reduce the computational complexity and
 505 improve the model generalization ability (Maldonado & Weber, 2009). High dimensional
 506 variables impose a high computational cost and a high cost of data acquisition. On the other
 507 hand, a low-dimensional representation reduces the risk of overfitting (Famili et al., 1997; Liu
 508 & Zheng, 2006). The value added by applying feature selection techniques is to determinate a
 509 subset of available variables to build a good model, which is a combinatorial problem in the
 510 number of original variables (Wolsey & Nemhauser, 1999; Guyon & Elisseeff, 2003).

511 With this rationale, we used popular filter, embedded and wrapper methods in order to select
 512 the relevant variables to predict soybean yield.

513 Filter methods are independent of any learners and they are based on performance evaluation
 514 metrics calculated directly from the data. We used two correlation-based filters (Pearson and
 515 Spearman) followed by a measure based on information theory (Information Gain). Figure 7
 516 presents Venus diagram of top-15 variables selected by filter methods. Correlation-based filters
 517 selected twelve same variables. The correlations found by Pearson and Spearman are equivalent
 518 to rank the common twelve variables; whereas Pearson inspects straight connections, Spearman
 519 evaluates monotonic connections (regardless of whether direct or not) (Thirumalai et al., 2017).
 520 However, of the three filter methods share less features in common (eight variables) due to

521 entropy-based information is designed to observe the amount of information gained between
 522 two discrete variables, and dataset used contains only numeric variables.

523 The principle of embedded methods (feature selection as part of the training process) is to
 524 reduce the computation time taken up for testing several subsets of variables which is done in
 525 wrapper methods (Chandrashekar & Sahin, 2014). In this work, the number of variables
 526 selected by RF and LASSO were higher. LASSO selected 66 variables and Random Forest all
 527 of the initial variable set. Although M5 decision Tree selected less variables (13), the results
 528 were worse than RF ($R^2 = 0.4010$, $MSE = 0.9466$, $RMSE = 0.9729$ and $MAE = 0.7802$).
 529 Embedded methods showed an improvement when they were used as ensemble approaches of
 530 filter and wrapper techniques.



531
 532 **Figure 7.** Venus diagram of top-15 variables selected by filter methods: Pearson, Spearman and Entropy-based information

533 Wrapper methods are widely recognized and considered a superior alternative for two reasons:
 534 (i) they evaluate variables iteratively with respect to performance learner. Therefore, variables
 535 selected by wrapper approach are more likely to suit the learner (Kohavi & John, 1997); (ii)
 536 wrapper approaches evaluate variables jointly and are effective in capturing intrinsic
 537 relationships such as interactions among multiple variables (Freitas, 2001). However
 538 computational cost is high, even for learners that exhibit a moderate complexity, the number of
 539 iterations required by the search of subset variables is high, especially as more complex search
 540 strategies are used (Talavera, 2005). In this paper, we used a dataset with few observations (105
 541 instances) but it contains a large number of features (87 variables). Recursive Feature
 542 Elimination through SVR and LR learners reached the best performance. RFE-SVR and RFE-
 543 LR are recommended to be used as surrogate models to predict soybean yield in southern
 544 France. Concerning to time complexity, the lowest time was obtained by RFE-LR (59.42
 545 seconds) against RFE-SVR (5271.56 seconds).

546 On the other hand, Support Vector Regression and Linear Regression belong different type of
 547 learners. SVR is considered a black-box model whereas LR an interpretable model (Loyola-
 548 González, 2019). SVR builds a hyperplane or set of hyperplanes in a high-dimensional space,
 549 which are very hard to explain and to be understood by experts in practical applications (Rudin,
 550 2019). LR generates a linear equation to explain the correlation among variables in a language
 551 close to a human expert.

552 In this sense, RFE-LR is represented by the following linear equation:

$$Mafruit = -0.3095*MaxTemp_avg_Idrp-Imat + 0.0148*Qnplante_Idrp + 0.0057*Masec(n)_Imat + 0.0066*Ep_cum_Idrp-Imat + 0.0458 *Lai(n)_Imat + 0.0008*Raint_cum_Idrp-Imat + 8.5049 \quad (9)$$

553 RFE-LR model explains biologically the higher leaf area duration during grain filling through
 554 the increase of grain yield (*Mafruit*) with aboveground biomass at maturity (*Masec(n)_Imat*),
 555 precipitation amount (*Raint_cum_Idrp-Imat*), cumulative crop transpiration during grain filling
 556 (*Ep_cum_Idrp-Imat*), the mineral nitrogen accumulated by the plants at the onset of grain filling
 557 (*Qnplante_Idrp*) and residual leaf area index at maturity (*Lai(n)_Imat*). All these variables
 558 demonstrate that radiation, water and nitrogen resources are highly representative variables of
 559 soybean grain yield. Further, the high temperatures may affect crop photosynthesis and grain
 560 filling (*MaxTemp_avg_Idrp-Imat*) which is in accordance with our knowledge of soybean
 561 physiology (Grassini et al., 2021).

562

563 **Acknowledgments**

564 This research was supported by Horizon 2020 SusCrop-ERA-NET Cofound on Sustainable
 565 Crop Production: LegumeGap project (2019-2021) “Increasing productivity and sustainability
 566 of European plant protein production by closing the grain legume yield gap”:
 567 <https://www.suscrop.eu/projects-first-call/legumegap>

568 **REFERENCES**

- 569 Battisti, R., Parker, P. S., Sentelhas, P. C., & Nendel, C. (2017). Gauging the sources of uncertainty in
 570 soybean yield simulations using the MONICA model. *Agricultural Systems*, 155, 9–18.
 571 <https://doi.org/10.1016/j.agsy.2017.04.004>
- 572 Bhatia, V. S., Singh, P., Wani, S. P., Chauhan, G. S., Rao, A. V. R. K., Mishra, A. K., & Srinivas, K. (2008).
 573 Analysis of potential yields and yield gaps of rainfed soybean in India using CROPGRO-
 574 Soybean model. *Agricultural and Forest Meteorology*, 148(8), 1252–1265.
 575 <https://doi.org/10.1016/j.agrformet.2008.03.004>
- 576 Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., & Jones, Z. M.
 577 (2016). mlr: Machine Learning in R. *The Journal of Machine Learning Research*, 17(1), 5938–
 578 5942.
- 579 Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- 580 Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*.
 581 CERN Document Server; Wadsworth & Brooks/Cole Advanced Books & Software.
 582 <https://cds.cern.ch/record/2253780>

583 Brereton, R. G., & Lloyd, G. R. (2010). Support Vector Machines for classification and regression.
584 *Analyst*, 135(2), 230–267. <https://doi.org/10.1039/B918972F>

585 Brisson, N., Launay, M., Mary, B., & Beaudoin, N. (2009). *Conceptual basis, formalisations and*
586 *parameterization of the STICS crop model*. Editions Quae.

587 Brisson, N., Mary, B., Ripoche, D., Jeuffroy, M. H., Ruget, F., Nicoullaud, B., Gate, P., Devienne-Barret,
588 F., Antonioletti, R., & Durr, C. (1998). *STICS: a generic model for the simulation of crops and*
589 *their water and nitrogen balances. I. Theory and parameterization applied to wheat and corn*.

590 Casadebaig, P., Debaeke, P., & Wallach, D. (2020). A new approach to crop model calibration:
591 Phenotyping plus post-processing. *Crop Science*, 60(2), 709–720.
592 <https://doi.org/10.1002/csc2.20016>

593 Casadebaig, P., Guilioni, L., Lecoeur, J., Christophe, A., Champolivier, L., & Debaeke, P. (2011).
594 SUNFLO, a model to simulate genotype-specific performance of the sunflower crop in
595 contrasting environments. *Agricultural and Forest Meteorology*, 151(2), 163–178.
596 <https://doi.org/10.1016/j.agrformet.2010.09.012>

597 Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical*
598 *Engineering*, 40(1), 16–28. <https://doi.org/10.1016/j.compeleceng.2013.11.024>

599 Corrales, D. C., Lasso, E., Ledezma, A., & Corrales, J. C. (2018). Feature selection for classification
600 tasks: Expert knowledge or traditional methods? *Journal of Intelligent & Fuzzy Systems*, 34(5),
601 2825–2835.

602 Coucheney, E., Buis, S., Launay, M., Constantin, J., Mary, B., de Cortázar-Atauri, I. G., Ripoche, D.,
603 Beaudoin, N., Ruget, F., & Andrianarisoa, K. S. (2015). Accuracy, robustness and behavior of
604 the STICS soil–crop model for plant, water and nitrogen outputs: Evaluation over a wide
605 range of agro-environmental conditions in France. *Environmental Modelling & Software*, 64,
606 177–190.

607 Cui, C., Hu, M., Weir, J. D., & Wu, T. (2016). A recommendation system for meta-modeling: A meta-
608 learning based approach. *Expert Systems with Applications*, 46, 33–44.
609 <https://doi.org/10.1016/j.eswa.2015.10.021>

610 Deshwal, D., Sangwan, P., & Kumar, D. (2020). A Language Identification System using Hybrid
611 Features and Back-Propagation Neural Network. *Applied Acoustics*, 164, 107289.
612 <https://doi.org/10.1016/j.apacoust.2020.107289>

613 Eugenio, F. C., Grohs, M., Venancio, L. P., Schuh, M., Bottega, E. L., Ruoso, R., Schons, C., Mallmann,
614 C. L., Badin, T. L., & Fernandes, P. (2020). Estimation of soybean yield from machine learning
615 techniques and multispectral RPAS imagery. *Remote Sensing Applications: Society and*
616 *Environment*, 20, 100397. <https://doi.org/10.1016/j.rsase.2020.100397>

617 Falconnier, G. N., Journet, E.-P., Bedoussac, L., Vermue, A., Chlébowski, F., Beaudoin, N., & Justes, E.
618 (2019). Calibration and evaluation of the STICS soil-crop model for faba bean to explain
619 variability in yield and N₂ fixation. *European Journal of Agronomy*, 104, 63–77.
620 <https://doi.org/10.1016/j.eja.2019.01.001>

621 Falconnier, G. N., Vermue, A., Journet, E.-P., Christina, M., Bedoussac, L., & Justes, E. (2020).
622 Contrasted response to climate change of winter and spring grain legumes in southwestern
623 France. *Field Crops Research*, 259, 107967. <https://doi.org/10.1016/j.fcr.2020.107967>

624 Famili, A., Shen, W.-M., Weber, R., & Simoudis, E. (1997). Data preprocessing and intelligent data
625 analysis. *Intelligent Data Analysis*, 1(1), 3–23.

626 FAOSTAT. (2021). *Food and agriculture organization of the united nations—Crops*.
627 <http://www.fao.org/faostat/en/#data/QC>

628 Freitas, A. A. (2001). Understanding the crucial role of attribute interaction in data mining. *Artificial*
629 *Intelligence Review*, 16(3), 177–199.

630 Gauthier, T. D. (2001). Detecting trends using Spearman’s rank correlation coefficient. *Environmental*
631 *Forensics*, 2(4), 359–362.

632 Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with
633 random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent*
634 *Laboratory Systems*, 83(2), 83–90.

635 Grassini, P., Cafaro La Menza, N., Rattalino Edreira, J. I., Monzón, J. P., Tenorio, F. A., & Specht, J. E.
636 (2021). Chapter 8—Soybean. In V. O. Sadras & D. F. Calderini (Eds.), *Crop Physiology Case*
637 *Histories for Major Crops* (pp. 282–319). Academic Press. [https://doi.org/10.1016/B978-0-12-](https://doi.org/10.1016/B978-0-12-819194-1.00008-6)
638 [819194-1.00008-6](https://doi.org/10.1016/B978-0-12-819194-1.00008-6)

639 Guilpart, N., Iizumi, T., & Makowski, D. (2020). Data-driven yield projections suggest large
640 opportunities to improve Europe’s soybean self-sufficiency under climate change. *BioRxiv*,
641 2020.10.08.331496. <https://doi.org/10.1101/2020.10.08.331496>

642 Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine*
643 *Learning Research*, 3(Mar), 1157–1182.

644 Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using
645 Support Vector Machines. *Machine Learning*, 46(1), 389–422.
646 <https://doi.org/10.1023/A:1012487302797>

647 Hornik, K., Buchta, C., & Zeileis, A. (2009). Open-source machine learning: R meets Weka.
648 *Computational Statistics*, 24(2), 225–232. <https://doi.org/10.1007/s00180-008-0119-7>

649 Jagtap, S. S., & Jones, J. W. (2002). Adaptation and evaluation of the CROPGRO-soybean model to
650 predict regional yield and production. *Agriculture, Ecosystems & Environment*, 93(1), 73–85.
651 [https://doi.org/10.1016/S0167-8809\(01\)00358-9](https://doi.org/10.1016/S0167-8809(01)00358-9)

652 Jégo, G., Pattey, E., Bourgeois, G., Morrison, M. J., Drury, C. F., Tremblay, N., & Tremblay, G. (2010).
653 Calibration and performance evaluation of soybean and spring wheat cultivars using the
654 STICS crop model in Eastern Canada. *Field Crops Research*, 117(2), 183–196.
655 <https://doi.org/10.1016/j.fcr.2010.03.008>

656 Jing, Q., Huffman, T., Shang, J., Liu, J., Pattey, E., Morrison, M., Jégo, G., & Qian, B. (2017). Modelling
657 soybean yield responses to seeding date under projected climate change scenarios. *Canadian*
658 *Journal of Plant Science*. <https://doi.org/10.1139/cjps-2017-0065>

659 Kaul, M., Hill, R. L., & Walthall, C. (2005). Artificial neural networks for corn and soybean yield
660 prediction. *Agricultural Systems*, *85*(1), 1–18. <https://doi.org/10.1016/j.agsy.2004.07.009>

661 Khaki, S., Wang, L., & Archontoulis, S. V. (2020). A CNN-RNN Framework for Crop Yield Prediction.
662 *Frontiers in Plant Science*, *10*. <https://doi.org/10.3389/fpls.2019.01750>

663 Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1),
664 273–324. [https://doi.org/10.1016/S0004-3702\(97\)00043-X](https://doi.org/10.1016/S0004-3702(97)00043-X)

665 Kothari, K., Salmeron, M., Battisti, R., Boote, K., Archontoulis, S., Confalone, A., Constantin, J., Cuadra
666 Sanatiago, V., Debaeke, P., Faye, B., Grant, B., Hoogen-Boom, G., Jing, Q., Michael Van Der
667 10, L., Macena, F., Marin, F., Nehbandani, A., Nendel, C., Larry, P., ... Viera Nilson, A. (2020,
668 February). First Soybean Multi-model Sensitivity Analysis to CO₂, Temperature, Water, and
669 Nitrogen. *ICROP2020: Second International Crop Modelling Symposium*, Montpellier.
670 <https://hal.inria.fr/hal-02950318>

671 Kross, A., Znoj, E., Callegari, D., Kaur, G., Sunohara, M., Lapen, D. R., & McNairn, H. (2020). Using
672 Artificial Neural Networks and Remotely Sensed Data to Evaluate the Relative Importance of
673 Variables for Prediction of Within-Field Corn and Soybean Yields. *Remote Sensing*, *12*(14),
674 2230. <https://doi.org/10.3390/rs12142230>

675 Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical*
676 *Software*, *28*(5), 1–26.

677 Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

678 Liu, Y., & Zheng, Y. F. (2006). FS_SFS: A novel feature selection method for support vector machines.
679 *Pattern Recognition*, *39*(7), 1333–1345.

680 Loyola-González, O. (2019). Black-Box vs. White-Box: Understanding Their Advantages and
681 Weaknesses From a Practical Point of View. *IEEE Access*, 7, 154096–154113.
682 <https://doi.org/10.1109/ACCESS.2019.2949286>

683 Ma, B. L., Dwyer, L. M., Costa, C., Cober, E. R., & Morrison, M. J. (2001). Early Prediction of Soybean
684 Yield from Canopy Reflectance Measurements. *Agronomy Journal*, 93(6), 1227–1234.
685 <https://doi.org/10.2134/agronj2001.1227>

686 Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., & Fritschi, F. B. (2020). Soybean yield
687 prediction from UAV using multimodal data fusion and deep learning. *Remote Sensing of*
688 *Environment*, 237, 111599. <https://doi.org/10.1016/j.rse.2019.111599>

689 Maldonado, S., & Weber, R. (2009). A wrapper method for feature selection using Support Vector
690 Machines. *Information Sciences*, 179(13), 2208–2217.
691 <https://doi.org/10.1016/j.ins.2009.02.014>

692 Nehbandani, A., Soltani, A., Nourbakhsh, F., & Dadrasi, A. (2020). Estimating crop model parameters
693 for simulating soybean production in Iran conditions. *OCL*, 27, 58.
694 <https://doi.org/10.1051/ocl/2020057>

695 Oil Word. (2020). *The OIL WORLD ANNUAL encyclopedia*.
696 <https://www.oilworld.biz/t/publications/annual>

697 Ovando, G., Sayago, S., & Bocco, M. (2018). Evaluating accuracy of DSSAT model for soybean yield
698 estimation using satellite weather data. *ISPRS Journal of Photogrammetry and Remote*
699 *Sensing*, 138, 208–217. <https://doi.org/10.1016/j.isprsjprs.2018.02.015>

700 Pearson, K. (1920). Notes on the History of Correlation. *Biometrika*, 13(1), 25–45.
701 <https://doi.org/10.2307/2331722>

702 Porter, J. R., Xie, L., Challinor, A. J., Cochrane, K., Howden, S. M., Iqbal, M. M., Lobell, D. B., &
703 Travasso, M. I. (2014). *Food security and food production systems*.

704 Prion, S. K., & Haerling, K. A. (2020). Making Sense of Methods and Measurements: Simple Linear
705 Regression. *Clinical Simulation in Nursing*, 48, 94–95.
706 <https://doi.org/10.1016/j.ecns.2020.07.004>

707 Purcell, L. C., & Roekel, R. J. V. (2019). Simulating Soybean Yield Potential under Optimum
708 Management. *Agrosystems, Geosciences & Environment*, 2(1), 190029.
709 <https://doi.org/10.2134/age2019.04.0029>

710 Quinlan, J. R. (1992). Learning with continuous classes. *5th Australian Joint Conference on Artificial
711 Intelligence*, 92, 343–348.

712 Robertson, M. J., & Carberry, P. S. (1998). Simulating growth and development of soybean in APSIM.
713 *Proceedings, 10th Australian Soybean Conference*, 130–136.
714 [https://publications.csiro.au/rpr/pub?list=BRO&pid=procite:e5446cf7-51a2-4606-a660-
715 28dadedfbf68d](https://publications.csiro.au/rpr/pub?list=BRO&pid=procite:e5446cf7-51a2-4606-a660-28dadedfbf68d)

716 Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use
717 interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
718 <https://doi.org/10.1038/s42256-019-0048-x>

719 Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating
720 errors. *Nature*, 323(6088), 533–536.

721 Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and
722 Interpretation. *Anesthesia and Analgesia*, 126(5), 1763–1768.
723 <https://doi.org/10.1213/ANE.0000000000002864>

724 Schoving, C. (2020). *Analyse écophysiological et modélisation dynamique des interactions génotype x
725 environnement x conduite de culture chez le soja* [PhD Thesis, Université fédérale Toulouse
726 Midi-Pyrénées]. <http://www.theses.fr/s164533>

727 Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V. V., & Ciampitti, I. A. (2020).
728 Satellite-based soybean yield forecast: Integrating machine learning and weather data for

729 improving crop yield prediction in southern Brazil. *Agricultural and Forest Meteorology*, 284,
730 107886. <https://doi.org/10.1016/j.agrformet.2019.107886>

731 Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*,
732 27(3), 379–423.

733 Shi, X., Wang, K., Cheong, T. S., & Zhang, H. (2020). Prioritizing driving factors of household carbon
734 emissions: An application of the LASSO model with survey data. *Energy Economics*, 92,
735 104942. <https://doi.org/10.1016/j.eneco.2020.104942>

736 Solorio-Fernández, S., Martínez-Trinidad, J. Fco., & Carrasco-Ochoa, J. A. (2020). A Supervised Filter
737 Feature Selection method for mixed data based on Spectral Feature Selection and
738 Information-theory redundancy analysis. *Pattern Recognition Letters*, 138, 321–328.
739 <https://doi.org/10.1016/j.patrec.2020.07.039>

740 Spearman, C. (1961). *The Proof and Measurement of Association Between Two Things* (p. 58).
741 Appleton-Century-Crofts. <https://doi.org/10.1037/11491-005>

742 Stepanov, A., Dubrovin, K., Sorokin, A., & Aseeva, T. (2020). Predicting Soybean Yield at the Regional
743 Scale Using Remote Sensing and Climatic Data. *Remote Sensing*, 12(12), 1936.
744 <https://doi.org/10.3390/rs12121936>

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

747 Talavera, L. (2005). An evaluation of filter and wrapper methods for feature selection in categorical
748 clustering. *International Symposium on Intelligent Data Analysis*, 440–451.

749 Terres Univia. (2021). *Plantes riches en protéines* (Chiffres Clés 2020, p. 24). Chiffres Clés 2020.
750 <http://www.terresunivia.fr/documentation-presse/chiffres-cles/chiffres-cles>

751 Thirumalai, C., Chandhini, S. A., & Vaishnavi, M. (2017). Analysing the concrete compressive strength
752 using Pearson and Spearman. *2017 International Conference of Electronics, Communication
753 and Aerospace Technology (ICECA)*, 2, 215–218.

754 Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical*
755 *Society: Series B (Methodological)*, 58(1), 267–288.

756 Vapnik, V. (1995). *The nature of statistical learning theory*. Springer science & business media.

757 Wang, Y., & Witten, I. H. (1996). *Induction of model trees for predicting continuous classes*.

758 Wei, M. C. F., & Molin, J. P. (2020). Soybean Yield Estimation and Its Components: A Linear
759 Regression Approach. *Agriculture*, 10(8), 348.

760 Wolsey, L. A., & Nemhauser, G. L. (1999). *Integer and combinatorial optimization* (Vol. 55). John
761 Wiley & Sons.

762 Xu, C., & Katchova, A. L. (2019). Predicting Soybean Yield with NDVI Using a Flexible Fourier
763 Transform Model. *Journal of Agricultural and Applied Economics*, 51(3), 402–416.
764 <https://doi.org/10.1017/aae.2019.5>

765 Yang, P., Zhou, B. B., Zhang, Z., & Zomaya, A. Y. (2010). A multi-filter enhanced genetic ensemble
766 system for gene selection and sample classification of microarray data. *BMC Bioinformatics*,
767 11(1), S5. <https://doi.org/10.1186/1471-2105-11-S1-S5>

768 Yu, H., Yang, J., & Han, J. (2003). Classifying large data sets using SVMs with hierarchical clusters.
769 *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and*
770 *Data Mining*, 306–315.

771

APPENDIX A

Experimental site	Year	Maturity group	Variety	Planting date	Water management	Soil	Observed variables
Auzeville	2017	000, I, II	Blancas, Ecudor, Santana, Isidor, Sultana, RGT, Shouna	80, 130	IRR, DRY	Clay, Loam	LAI, BNF, AGPN, AGB, GY, GNC, Oil
	2018	000, I, II	Blancas, Ecudor, Santana, Isidor, ES, Pallador, Sultana, RGT, Shouna	114, 155	IRR, DRY	Silty, Clay, Loam	BNF, AGPN, AGB, GY, GNC, Oil, roots depth
Béziers	2010	I, II	Isidor, Sumatra, Ecudor, Fukui, S109554	55, 76, 112	IRR	Loam	GY
	2011	0, I, II	Sarema, Isidor, Ecudor	67, 96, 132	IRR	Loam	GY
	2012	0, I, II	Sarema, Isidor, Ecudor	76, 103, 131	IRR	Silt, Loam	GY, GNC
En Crambade	2013	I, II	Isidor, Santana, Ecudor	74, 115	IRR, DRY	Clay	LAI, BNF, AGPN, AGB, GY, GNC, Oil
	2014	I, II	Isidor, Santana, Ecudor	73, 120	IRR, DRY	Clay	LAI, BNF, AGPN, AGB, GY, GNC, Oil
Mauguio	2010	I, II	Isidor, Sumatra, Ecudor, Fukui, S109554	74, 98, 145	IRR	Clay, Loam	GY
Mondonville	2010	I, II	Isidor, Sumatra, Ecudor, Fukui, S109554	61, 92, 138	IRR	Silt, Loam	GY
	2011	0, I, II	Sarema, Isidor, Ecudor	80, 102, 124	IRR	Silt, Loam	GY
	2012	I, II	Isidor, Ecudor	76, 97, 124	IRR	Silt, Loam	GY, GNC
	2013	I, II	Isidor, Santana, Ecudor	81, 147	IRR, DRY	Silt, Loam	LAI, BNF, AGPN, AGB, GY, GNC, Oil
	2014	I, II	Isidor, Santana, Ecudor	126	IRR, DRY	Silt, Loam	LAI, BNF, AGPN, AGB, GY, GNC, Oil
Rivières	2010	I, II	Isidor, Sumatra, Ecudor, Fukui, S109554	62, 99	IRR	Clay, Loam	GY
	2011	0, I, II	Sarema, Isidor, Ecudor	70, 102, 131	IRR	Clay, Loam	GY
	2012	I, II	Isidor, Ecudor	76, 108, 138	IRR	Clay, Loam	GY, GNC
	2013	I, II	Isidor, Santana, Ecudor	81, 126	IRR, DRY	Clay, Loam	LAI, BNF, AGPN, AGB, GY, GNC, Oil
	2014	I, II	Isidor, Santana, Ecudor	77, 126	IRR, DRY	Clay, Loam	LAI, BNF, AGPN, AGB, GY, GNC, Oil

Table A1. Dataset description of seventeen experimental sites during 2010-2018 from six regions in the south of France. The dataset contains 227 simulation units (USM) created from combination of experimental sites, years and cropping practices (cultivar, water management and sowing date). LAI: Leaf area index (m².m⁻²), BNF: fixed nitrogen (kg.ha⁻¹), AGPN: Total nitrogen (kg.ha⁻¹), AGB: Biomass of aerial parts (t.ha⁻¹), GY: Grain yield (t.ha⁻¹), GNC: Nitrogen concentration in the grains (%), Oil: percentage of oil (%), roots depth: rooting depth (cm). Source: (Schoving, 2020).

Table A2 lists all of the preprocessed variables from STICS basic variables, crop phenology stages, descriptive values and thresholds.

#	Variable	#	Variable	#	Variable
1	Nbgrmax	30	Etp(n)_cum_Ildr-Imp	59	MaxTemp_max_Iflo-Ildr
2	Stlevdrp	31	Etp(n)_cum_Iflo-Ildr	60	Days_MaxTemp_28_Ilev-Imp
3	Stflodrp	32	Precip_cum_Ilev-Imp	61	Days_MaxTemp_28_Iflo-Imp
4	Stdrpmp	33	Precip_cum_Iflo-Imp	62	Days_MaxTemp_28_Ildr-Imp
5	Masec(n)_Iflo	34	Precip_cum_Ildr-Imp	63	Days_MaxTemp_28_Iflo-Ildr
6	Masec(n)_Ildr	35	Precip_cum_Iflo-Ildr	64	Swfac_avg_Ilev-Imp
7	Masec(n)_Imp	36	Ep_cum_Ilev-Imp	65	Swfac_avg_Iflo-Imp
8	Lai(n)_Iflo	37	Ep_cum_Iflo-Imp	66	Swfac_avg_Ildr-Imp
9	Lai(n)_Ildr	38	Ep_cum_Ildr-Imp	67	Swfac_avg_Iflo-Ildr
10	Lai(n)_Imp	39	Ep_cum_Iflo-Ildr	68	Swfac_min_Ilev-Imp
11	Qnplante_Iflo	40	AvgTemp_avg_Ilev-Imp	69	Swfac_min_Iflo-Imp
12	Qnplante_Ildr	41	AvgTemp_avg_Iflo-Imp	70	Swfac_min_Ildr-Imp
13	Qnplante_Imp	42	AvgTemp_avg_Ildr-Imp	71	Swfac_min_Iflo-Ildr
14	Qfix_Iflo	43	AvgTemp_avg_Iflo-Ildr	72	Days_Swfac_0.6_Ilev-Imp
15	Qfix_Ildr	44	MinTemp_avg_Ilev-Imp	73	Days_Swfac_0.6_Iflo-Imp
16	Qfix_Imp	45	MinTemp_avg_Iflo-Imp	74	Days_Swfac_0.6_Ildr-Imp
17	Zrac_Iflo	46	MinTemp_avg_Ildr-Imp	75	Days_Swfac_0.6_Iflo-Ildr
18	Zrac_Ildr	47	MinTemp_avg_Iflo-Ildr	76	Inn_avg_Ilev-Imp
19	Zrac_Imp	48	Days_MinTemp_18_Ilev-Imp	77	Inn_avg_Iflo-Imp
20	Jul_cum_Ilev-Imp	49	Days_MinTemp_18_Iflo-Imp	78	Inn_avg_Ildr-Imp
21	Jul_cum_Iflo-Imp	50	Days_MinTemp_18_Ildr-Imp	79	Inn_avg_Iflo-Ildr
22	Jul_cum_Ildr-Imp	51	Days_MinTemp_18_Iflo-Ildr	80	Inn_min_Ilev-Imp
23	Jul_cum_Iflo-Ildr	52	MaxTemp_avg_Ilev-Imp	81	Inn_min_Iflo-Imp
24	Raint_cum_Ilev-Imp	53	MaxTemp_avg_Iflo-Imp	82	Inn_min_Ildr-Imp
25	Raint_cum_Iflo-Imp	54	MaxTemp_avg_Ildr-Imp	83	Inn_min_Iflo-Ildr
26	Raint_cum_Ildr-Imp	55	MaxTemp_avg_Iflo-Ildr	84	Days_Inn_0.6_Ilev-Imp
27	Raint_cum_Iflo-Ildr	56	MaxTemp_max_Ilev-Imp	85	Days_Inn_0.6_Iflo-Imp
28	Etp(n)_cum_Ilev-Imp	57	MaxTemp_max_Iflo-Imp	86	Days_Inn_0.6_Ildr-Imp
29	Etp(n)_cum_Iflo-Imp	58	MaxTemp_max_Ildr-Imp	87	Days_Inn_0.6_Iflo-Ildr

Table A2. Name of variables processed from Multidisciplinary Simulator for Standard Crops (STICS). Crop phenology stages: Ilev (emergence), Ildr (grain filling onset), Iflo (flowering), Imp (physiological maturity); descriptive values: cum (cumulative), max (maximum), avg (average); thresholds (MinTemp, MaxTemp, Swfac, Inn variables)

APPENDIX B

Pos	Variable name	Var Importance	Pos	Variable name	Var Importance
1	Masec(n)_Imat	0.7923	45	Inn_avg_Iflo-Idrp	0.3542
2	Lai(n)_Imat	0.7726	46	MinTemp_avg_Ilev-Imat	0.3508
3	Qnplante_Idrp	0.7710	47	Days_MaxTemp_28_Iflo-Imat	0.3478
4	Qnplante_Imat	0.7523	48	Precip_cum_Iflo-Idrp	0.3460
5	Masec(n)_Idrp	0.7334	49	MaxTemp_avg_Iflo-Imat	0.3432
6	Ep_cum_Idrp-Imat	0.6819	50	Days_MinTemp_18_Iflo-Idrp	0.3327
7	Qfix_Imat	0.6805	51	Inn_avg_Iflo-Imat	0.3099
8	Ep_cum_Iflo-Imat	0.6650	52	MinTemp_avg_Iflo-Idrp	0.2944
9	Raint_cum_Idrp-Imat	0.6326	53	Inn_avg_Ilev-Imat	0.2924
10	Lai(n)_Idrp	0.6292	54	Days_Inn_0.6_Iflo-Imat	0.2923
11	AvgTemp_avg_Idrp-Imat	0.6157	55	Days_Inn_0.6_Iflo-Idrp	0.2923
12	Swfac_avg_Iflo-Imat	0.6088	56	Days_Inn_0.6_Ilev-Imat	0.2903
13	MinTemp_avg_Idrp-Imat	0.5937	57	Inn_min_Iflo-Imat	0.2824
14	Ep_cum_Ilev-Imat	0.5842	58	Inn_min_Iflo-Idrp	0.2684
15	Days_Swfac_0.6_Iflo-Imat	0.5811	59	Inn_min_Idrp-Imat	0.2673
16	MinTemp_avg_Iflo-Imat	0.5654	60	Precip_cum_Idrp-Imat	0.2453
17	Swfac_min_Idrp-Imat	0.5638	61	Inn_min_Ilev-Imat	0.2405
18	Days_MinTemp_18_Idrp-Imat	0.5618	62	AvgTemp_avg_Ilev-Imat	0.2320
19	Swfac_min_Iflo-Imat	0.5600	63	Jul_cum_Ilev-Imat	0.1721
20	MaxTemp_avg_Idrp-Imat	0.5578	64	Masec(n)_Iflo	0.1698
21	AvgTemp_avg_Iflo-Imat	0.5551	65	Lai(n)_Iflo	0.1594
22	Swfac_min_Ilev-Imat	0.5521	66	MaxTemp_max_Ilev-Imat	0.1560
23	Days_MinTemp_18_Iflo-Imat	0.5328	67	Inn_avg_Idrp-Imat	0.1557
24	Days_MaxTemp_28_Idrp-Imat	0.5119	68	Days_MaxTemp_28_Ilev-Imat	0.1442
25	Ep_cum_Iflo-Idrp	0.5113	69	AvgTemp_avg_Iflo-Idrp	0.1030
26	Jul_cum_Idrp-Imat	0.5069	70	Zrac_Iflo	0.0934
27	Swfac_avg_Ilev-Imat	0.5054	71	Etp(n)_cum_Idrp-Imat	0.0858
28	Qfix_Idrp	0.4953	72	MaxTemp_avg_Iflo-Idrp	0.0856
29	Swfac_avg_Idrp-Imat	0.4929	73	Qfix_Iflo	0.0821
30	Swfac_avg_Iflo-Idrp	0.4909	74	Nbgrmax	0.0781
31	Days_Swfac_0.6_Ilev-Imat	0.4899	75	MaxTemp_max_Iflo-Imat	0.0773
32	Days_Swfac_0.6_Iflo-Idrp	0.4834	76	Etp(n)_cum_Ilev-Imat	0.0708
33	Raint_cum_Iflo-Imat	0.4712	77	Etp(n)_cum_Iflo-Imat	0.0701
34	Days_Swfac_0.6_Idrp-Imat	0.4627	78	Days_MaxTemp_28_Iflo-Idrp	0.0643
35	Zrac_Idrp	0.4592	79	MaxTemp_max_Iflo-Idrp	0.0618
36	Zrac_Imat	0.4536	80	MaxTemp_avg_Ilev-Imat	0.0613
37	Swfac_min_Iflo-Idrp	0.4464	81	Etp(n)_cum_Iflo-Idrp	0.0565
38	MaxTemp_max_Idrp-Imat	0.4257	82	Raint_cum_Iflo-Idrp	0.0461
39	Jul_cum_Iflo-Imat	0.3997	83	Jul_cum_Iflo-Idrp	0.0456
40	Days_MinTemp_18_Ilev-Imat	0.3926	84	Stflodrp	0.0372
41	Qnplante_Iflo	0.3866	85	Stlevdrp	0.0308
42	Raint_cum_Ilev-Imat	0.3802	86	Stdrpmat	0.0074
43	Precip_cum_Ilev-Imat	0.3714	87	Days_Inn_0.6_Idrp-Imat	0.0000
44	Precip_cum_Iflo-Imat	0.3585			

Table B1. Ranking of variables computed by Pearson correlation.

Pos	Variable name	Var Importance	Pos	Variable name	Var Importance
1	Lai(n)_Imat	0.7699	45	Precip_cum_Iflo-Imat	0.3140
2	Masec(n)_Imat	0.7311	46	Days_MinTemp_18_Iflo-Idrp	0.3018
3	Qnplante_Idrp	0.7137	47	Inn_avg_Iflo-Idrp	0.2983
4	Masec(n)_Idrp	0.7003	48	MinTemp_avg_Ilev-Imat	0.2851
5	Ep_cum_Iflo-Imat	0.6886	49	Inn_min_Idrp-Imat	0.2817
6	Ep_cum_Idrp-Imat	0.6734	50	Precip_cum_Ilev-Imat	0.2602
7	Qnplante_Imat	0.6714	51	Days_Inn_0.6_Ilev-Imat	0.2599
8	Ep_cum_Ilev-Imat	0.6456	52	Days_Inn_0.6_Iflo-Imat	0.2599
9	Lai(n)_Idrp	0.6439	53	Days_Inn_0.6_Iflo-Idrp	0.2599
10	Raint_cum_Idrp-Imat	0.6288	54	Inn_avg_Iflo-Imat	0.2570
11	Swfac_min_Idrp-Imat	0.6160	55	Precip_cum_Iflo-Idrp	0.2407
12	Swfac_min_Iflo-Imat	0.6136	56	Inn_min_Iflo-Idrp	0.2370
13	AvgTemp_avg_Idrp-Imat	0.6081	57	Inn_avg_Ilev-Imat	0.2353
14	Swfac_min_Ilev-Imat	0.6009	58	MinTemp_avg_Iflo-Idrp	0.2174
15	MinTemp_avg_Idrp-Imat	0.5977	59	Masec(n)_Iflo	0.2164
16	Qfix_Imat	0.5964	60	Inn_min_Iflo-Imat	0.2155
17	MaxTemp_avg_Idrp-Imat	0.5937	61	Lai(n)_Iflo	0.2105
18	Days_MinTemp_18_Idrp-Imat	0.5876	62	Inn_min_Ilev-Imat	0.1991
19	MinTemp_avg_Iflo-Imat	0.5758	63	Jul_cum_Ilev-Imat	0.1979
20	Swfac_avg_Iflo-Imat	0.5655	64	Days_MaxTemp_28_Ilev-Imat	0.1840
21	AvgTemp_avg_Iflo-Imat	0.5548	65	AvgTemp_avg_Ilev-Imat	0.1831
22	Days_MinTemp_18_Iflo-Imat	0.5510	66	MaxTemp_max_Ilev-Imat	0.1801
23	Ep_cum_Iflo-Idrp	0.5482	67	Inn_avg_Idrp-Imat	0.1667
24	Jul_cum_Idrp-Imat	0.5380	68	Precip_cum_Idrp-Imat	0.1519
25	Days_Swfac_0.6_Iflo-Imat	0.5361	69	AvgTemp_avg_Iflo-Idrp	0.1400
26	Days_MaxTemp_28_Idrp-Imat	0.5033	70	Qfix_Iflo	0.1263
27	Qfix_Idrp	0.4679	71	Nbgrmax	0.1160
28	Jul_cum_Iflo-Imat	0.4667	72	Etp(n)_cum_Idrp-Imat	0.1095
29	MaxTemp_max_Idrp-Imat	0.4629	73	Raint_cum_Iflo-Idrp	0.1015
30	Raint_cum_Iflo-Imat	0.4590	74	MaxTemp_max_Iflo-Imat	0.0949
31	Days_Swfac_0.6_Ilev-Imat	0.4546	75	Jul_cum_Iflo-Idrp	0.0849
32	Swfac_avg_Ilev-Imat	0.4364	76	MaxTemp_avg_Ilev-Imat	0.0763
33	Swfac_avg_Idrp-Imat	0.4319	77	Etp(n)_cum_Ilev-Imat	0.0763
34	Raint_cum_Ilev-Imat	0.4209	78	Etp(n)_cum_Iflo-Idrp	0.0736
35	Qnplante_Iflo	0.4206	79	Etp(n)_cum_Iflo-Imat	0.0731
36	Days_Swfac_0.6_Idrp-Imat	0.4148	80	Zrac_Iflo	0.0720
37	Swfac_avg_Iflo-Idrp	0.4131	81	Days_MaxTemp_28_Iflo-Idrp	0.0538
38	Swfac_min_Iflo-Idrp	0.4105	82	Stflodrp	0.0465
39	Days_Swfac_0.6_Iflo-Idrp	0.3993	83	MaxTemp_avg_Iflo-Idrp	0.0345
40	Days_MinTemp_18_Ilev-Imat	0.3857	84	MaxTemp_max_Iflo-Idrp	0.0235
41	Days_MaxTemp_28_Iflo-Imat	0.3552	85	Stlevdrp	0.0085
42	Zrac_Idrp	0.3287	86	Stdrpmat	0.0085
43	MaxTemp_avg_Iflo-Imat	0.3273	87	Days_Inn_0.6_Idrp-Imat	0.0000
44	Zrac_Imat	0.3214			

Table B2. Ranking of variables computed by Spearman correlation filter.

Pos	Variable name	Var Importance	Pos	Variable name	Var Importance
1	Lai(n)_Imat	0.5929	45	Inn_min_Iflo-Imat	0.1745
2	Masec(n)_Imat	0.5875	46	Inn_avg_Idrp-Imat	0.1744
3	Qnplante_Imat	0.5515	47	MinTemp_avg_Iflo-Imat	0.1687
4	Ep_cum_Idrp-Imat	0.4223	48	Inn_min_Idrp-Imat	0.1673
5	Qfix_Imat	0.4072	49	Days_Swfac_0.6_Ilev-Imat	0.1628
6	Inn_min_Iflo-Idrp	0.3799	50	Days_MaxTemp_28_Ilev-Imat	0.1595
7	Inn_avg_Iflo-Idrp	0.3652	51	Lai(n)_Iflo	0.1540
8	Ep_cum_Iflo-Imat	0.3486	52	Precip_cum_Ilev-Imat	0.1213
9	Days_Swfac_0.6_Iflo-Imat	0.3443	53	AvgTemp_avg_Ilev-Imat	0.1204
10	Inn_avg_Ilev-Imat	0.3400	54	MaxTemp_avg_Ilev-Imat	0.1204
11	Inn_avg_Iflo-Imat	0.3380	55	Etp(n)_cum_Iflo-Imat	0.0993
12	Masec(n)_Idrp	0.3300	56	Etp(n)_cum_Ilev-Imat	0.0993
13	Qnplante_Idrp	0.3285	57	Etp(n)_cum_Iflo-Idrp	0.0993
14	Raint_cum_Idrp-Imat	0.3080	58	Masec(n)_Iflo	0.0000
15	Days_Swfac_0.6_Iflo-Idrp	0.3025	59	Qfix_Iflo	0.0000
16	Precip_cum_Iflo-Imat	0.2988	60	Zrac_Iflo	0.0000
17	Qnplante_Iflo	0.2895	61	Etp(n)_cum_Idrp-Imat	0.0000
18	Ep_cum_Ilev-Imat	0.2833	62	Precip_cum_Idrp-Imat	0.0000
19	Swfac_avg_Iflo-Imat	0.2758	63	MaxTemp_avg_Iflo-Imat	0.0000
20	Lai(n)_Idrp	0.2711	64	MaxTemp_max_Iflo-Imat	0.0000
21	Swfac_min_Iflo-Imat	0.2476	65	Raint_cum_Ilev-Imat	0.0000
22	Swfac_min_Ilev-Imat	0.2476	66	MinTemp_avg_Ilev-Imat	0.0000
23	Zrac_Idrp	0.2457	67	Swfac_avg_Ilev-Imat	0.0000
24	Zrac_Imat	0.2457	68	MaxTemp_max_Ilev-Imat	0.0000
25	MinTemp_avg_Idrp-Imat	0.2374	69	Inn_min_Ilev-Imat	0.0000
26	Swfac_min_Iflo-Idrp	0.2309	70	Jul_cum_Ilev-Imat	0.0000
27	Swfac_min_Idrp-Imat	0.2289	71	Raint_cum_Iflo-Idrp	0.0000
28	AvgTemp_avg_Iflo-Imat	0.2266	72	AvgTemp_avg_Iflo-Idrp	0.0000
29	MaxTemp_avg_Idrp-Imat	0.2222	73	MinTemp_avg_Iflo-Idrp	0.0000
30	Days_MinTemp_18_Ilev-Imat	0.2214	74	MaxTemp_avg_Iflo-Idrp	0.0000
31	Jul_cum_Iflo-Imat	0.2191	75	MaxTemp_max_Iflo-Idrp	0.0000
32	Days_MinTemp_18_Iflo-Imat	0.2063	76	Jul_cum_Iflo-Idrp	0.0000
33	AvgTemp_avg_Idrp-Imat	0.2049	77	Days_Inn_0.6_Ilev-Imat	0.0000
34	Days_MinTemp_18_Idrp-Imat	0.2035	78	Days_MaxTemp_28_Iflo-Imat	0.0000
35	MaxTemp_max_Idrp-Imat	0.2000	79	Days_Inn_0.6_Iflo-Imat	0.0000
36	Qfix_Idrp	0.1980	80	Days_Inn_0.6_Idrp-Imat	0.0000
37	Ep_cum_Iflo-Idrp	0.1971	81	Days_MinTemp_18_Iflo-Idrp	0.0000
38	Jul_cum_Idrp-Imat	0.1969	82	Days_MaxTemp_28_Iflo-Idrp	0.0000
39	Days_MaxTemp_28_Idrp-Imat	0.1954	83	Days_Inn_0.6_Iflo-Idrp	0.0000
40	Raint_cum_Iflo-Imat	0.1927	84	Nbgrmax	0.0000
41	Swfac_avg_Idrp-Imat	0.1883	85	Stlevdrp	0.0000
42	Days_Swfac_0.6_Idrp-Imat	0.1883	86	Stflodrp	0.0000
43	Precip_cum_Iflo-Idrp	0.1872	87	Stdrpmat	0.0000
44	Swfac_avg_Iflo-Idrp	0.1829			

Table B3. Ranking of variables computed by filter of entropy-based information gain.

APPENDIX C

Coefficients of linear regression created by LASSO

#	Variable name	Coefficient value	#	Variable name	Coefficient value
1	Masec(n)_Iflo	2.8188e+00	34	Inn_min_Iflo-Imat	-1.3651e+01
2	Lai(n)_Iflo	9.3241e-01	35	Raint_cum_Ilev-Imat	-1.1394e-02
3	Qnplante_Iflo	7.4166e-02	36	Precip_cum_Ilev-Imat	5.6827e-03
4	Qfix_Iflo	-1.7155e-01	37	Ep_cum_Ilev-Imat	-7.9358e-02
5	Zrac_Iflo	7.6535e-02	38	MinTemp_avg_Ilev-Imat	1.2562e-01
6	Masec(n)_Idrp	-4.0427e-01	39	MaxTemp_avg_Ilev-Imat	-3.7455e-04
7	Lai(n)_Idrp	7.3189e-01	40	Swfac_avg_Ilev-Imat	6.1859e+00
8	Qnplante_Idrp	-9.3804e-02	41	Inn_avg_Ilev-Imat	1.6985e+01
9	Qfix_Idrp	9.1514e-02	42	MaxTemp_max_Ilev-Imat	7.1837e-01
10	Zrac_Idrp	5.0038e-02	43	Swfac_min_Ilev-Imat	1.4684e-01
11	Masec(n)_Imat	5.1417e-01	44	Inn_min_Ilev-Imat	-3.9823e+00
12	Lai(n)_Imat	-4.0912e-01	45	Jul_cum_Ilev-Imat	-7.5989e-02
13	Qnplante_Imat	3.8100e-02	46	Precip_cum_Iflo-Idrp	6.9931e-03
14	Qfix_Imat	-5.2369e-02	47	Ep_cum_Iflo-Idrp	2.4697e-02
15	Zrac_Imat	-6.7964e-02	48	MinTemp_avg_Iflo-Idrp	3.9772e+00
16	Raint_cum_Idrp-Imat	9.6513e-03	49	MaxTemp_avg_Iflo-Idrp	-2.6303e+00
17	Precip_cum_Idrp-Imat	-2.8790e-03	50	Swfac_avg_Iflo-Idrp	1.6022e+01
18	MinTemp_avg_Idrp-Imat	4.7465e+00	51	Inn_avg_Iflo-Idrp	1.1904e+01
19	MaxTemp_avg_Idrp-Imat	-3.7388e+00	52	MaxTemp_max_Iflo-Idrp	-8.1299e-01
20	Swfac_avg_Idrp-Imat	1.2500e+01	53	Swfac_min_Iflo-Idrp	-4.1521e-01
21	Inn_avg_Idrp-Imat	3.1469e+01	54	Inn_min_Iflo-Idrp	2.0532e+01
22	MaxTemp_max_Idrp-Imat	-4.1193e-01	55	Jul_cum_Iflo-Idrp	4.9970e-01
23	Swfac_min_Idrp-Imat	-1.3343e+00	56	Days_MinTemp_18_Ilev-Imat	1.8624e-02
24	Inn_min_Idrp-Imat	-1.2703e+00	57	Days_MaxTemp_28_Ilev-Imat	4.6168e-02
25	Jul_cum_Idrp-Imat	-2.3218e-01	58	Days_Swfac_0.6_Ilev-Imat	3.6354e-03
26	Raint_cum_Iflo-Imat	5.7320e-03	59	Days_MinTemp_18_Iflo-Imat	-2.3197e-02
27	Ep_cum_Iflo-Imat	7.4176e-02	60	Days_MaxTemp_28_Iflo-Imat	-1.7790e-01
28	MinTemp_avg_Iflo-Imat	-9.4243e+00	61	Days_Swfac_0.6_Iflo-Imat	1.0801e-02
29	MaxTemp_avg_Iflo-Imat	7.7822e+00	62	Days_MinTemp_18_Iflo-Idrp	-1.0127e-01
30	Swfac_avg_Iflo-Imat	-2.8792e+01	63	Days_MaxTemp_28_Iflo-Idrp	-8.3875e-04
31	Inn_avg_Iflo-Imat	-4.6737e+01	64	Days_Swfac_0.6_Iflo-Idrp	5.2675e-03
32	MaxTemp_max_Iflo-Imat	-1.6202e-03	65	Nbgrmax	5.1039e-04
33	Swfac_min_Iflo-Imat	6.6187e-02	66	Stdrpmat	9.8675e-03

Table C1. Coefficients of linear regression created by LASSO. Coefficients equal to zero were assigned to 11 variables.

Linear regressions created by M5 decision tree

Linear regression number 1:

$$Mafruit = -0.2193 * Lai(n)_Iflo + 0.0187 * Qnplante_Iflo - 0.0314 * Qfix_Iflo + 0.0869 * Lai(n)_Idrp + 0.0759 * Masec(n)_Imat + 0.002 * Qfix_Imat + 3.9514 * Inn_avg_Iflo-Idrp - 1.665 * Inn_min_Iflo-Idrp - 1.0665$$

Linear regression number 2:

$$Mafruit = -0.243 * Masec(n)_Iflo - 0.2193 * Lai(n)_Iflo + 0.0305 * Qnplante_Iflo - 0.0179 * Qfix_Iflo + 0.0047 * Zrac_Iflo + 0.0869 * Lai(n)_Idrp + 0.0434 * Masec(n)_Imat + 0.002 * Qfix_Imat + 0.0036 * Ep_cum_Idrp-Imat + 0.0287 * MaxTemp_avg_Ilev-Imat + 3.2551 * Inn_avg_Iflo-Idrp - 1.665 * Inn_min_Iflo-Idrp - 1.7154$$

Linear regression number 3:

$$\begin{aligned} \text{Mafruit} = & -0.4282 * \text{Masec}(n)_I\text{flo} - 0.2193 * \text{Lai}(n)_I\text{flo} + 0.0242 * \text{Qnplante}_I\text{flo} - 0.0398 * \text{Qfix}_I\text{flo} + 0.0869 \\ & * \text{Lai}(n)_I\text{drp} + 0.0434 * \text{Masec}(n)_I\text{mat} + 0.002 * \text{Qfix}_I\text{mat} + 0.0032 * \text{Ep_cum}_I\text{drp-Imat} + 0.0258 * \\ & \text{MaxTemp_avg_Ilev-Imat} + 3.2551 * \text{Inn_avg}_I\text{flo-Idrp} - 1.665 * \text{Inn_min}_I\text{flo-Idrp} - 0.1411 \end{aligned}$$

Linear regression number 4:

$$\begin{aligned} \text{Mafruit} = & -0.4282 * \text{Masec}(n)_I\text{flo} - 0.2193 * \text{Lai}(n)_I\text{flo} + 0.0242 * \text{Qnplante}_I\text{flo} - 0.0398 * \text{Qfix}_I\text{flo} + 0.0869 \\ & * \text{Lai}(n)_I\text{drp} + 0.0434 * \text{Masec}(n)_I\text{mat} + 0.002 * \text{Qfix}_I\text{mat} + 0.0032 * \text{Ep_cum}_I\text{drp-Imat} + 0.0258 * \\ & \text{MaxTemp_avg_Ilev-Imat} + 3.2551 * \text{Inn_avg}_I\text{flo-Idrp} - 1.665 * \text{Inn_min}_I\text{flo-Idrp} - 0.1438 \end{aligned}$$

Linear regression number 5:

$$\begin{aligned} \text{Mafruit} = & -0.4587 * \text{Masec}(n)_I\text{flo} - 0.2193 * \text{Lai}(n)_I\text{flo} + 0.0242 * \text{Qnplante}_I\text{flo} - 0.0401 * \text{Qfix}_I\text{flo} + 0.0869 \\ & * \text{Lai}(n)_I\text{drp} + 0.0434 * \text{Masec}(n)_I\text{mat} + 0.002 * \text{Qfix}_I\text{mat} + 0.0032 * \text{Ep_cum}_I\text{drp-Imat} + 0.0258 * \\ & \text{MaxTemp_avg_Ilev-Imat} + 3.2551 * \text{Inn_avg}_I\text{flo-Idrp} - 1.665 * \text{Inn_min}_I\text{flo-Idrp} - 0.0967 \end{aligned}$$

Linear regression number 6:

$$\begin{aligned} \text{Mafruit} = & -0.3026 * \text{Masec}(n)_I\text{flo} - 0.2193 * \text{Lai}(n)_I\text{flo} + 0.0242 * \text{Qnplante}_I\text{flo} - 0.0312 * \text{Qfix}_I\text{flo} + 0.0869 \\ & * \text{Lai}(n)_I\text{drp} + 0.0434 * \text{Masec}(n)_I\text{mat} + 0.002 * \text{Qfix}_I\text{mat} + 0.0032 * \text{Ep_cum}_I\text{drp-Imat} + 0.0258 * \\ & \text{MaxTemp_avg_Ilev-Imat} + 3.2551 * \text{Inn_avg}_I\text{flo-Idrp} - 1.665 * \text{Inn_min}_I\text{flo-Idrp} - 0.5323 \end{aligned}$$

Linear regression number 7:

$$\begin{aligned} \text{Mafruit} = & -0.2811 * \text{Masec}(n)_I\text{flo} + 0.0803 * \text{Lai}(n)_I\text{flo} + 0.0067 * \text{Qnplante}_I\text{flo} + 0.0383 * \text{Qfix}_I\text{flo} + 0.0881 \\ & * \text{Masec}(n)_I\text{drp} + 0.0103 * \text{Lai}(n)_I\text{drp} + 0.0017 * \text{Qfix}_I\text{mat} - 0.0046 * \text{Ep_cum}_I\text{drp-Imat} + 0.7974 * \\ & \text{Swfac_avg}_I\text{drp-Imat} + 0.577 * \text{Swfac_avg}_I\text{flo-Imat} + 3.8547 * \text{Inn_avg}_I\text{flo-Idrp} - 2.4314 * \text{Inn_min}_I\text{flo-Idrp} \\ & + 0.256 \end{aligned}$$

Linear regression number 8:

$$\begin{aligned} \text{Mafruit} = & -0.1899 * \text{Masec}(n)_I\text{flo} + 0.0274 * \text{Lai}(n)_I\text{flo} + 0.0067 * \text{Qnplante}_I\text{flo} + 0.0223 * \text{Qfix}_I\text{flo} + 0.0881 \\ & * \text{Masec}(n)_I\text{drp} + 0.0256 * \text{Lai}(n)_I\text{drp} + 0.0017 * \text{Qfix}_I\text{mat} - 0.0041 * \text{Ep_cum}_I\text{drp-Imat} + 0.7974 * \\ & \text{Swfac_avg}_I\text{drp-Imat} + 0.6392 * \text{Swfac_avg}_I\text{flo-Imat} + 3.8547 * \text{Inn_avg}_I\text{flo-Idrp} - 2.4314 * \text{Inn_min}_I\text{flo-Idrp} \\ & + 0.1624 \end{aligned}$$

Linear regression number 9:

$$\begin{aligned} \text{Mafruit} = & -0.1928 * \text{Masec}(n)_I\text{flo} + 0.0274 * \text{Lai}(n)_I\text{flo} + 0.0067 * \text{Qnplante}_I\text{flo} + 0.0223 * \text{Qfix}_I\text{flo} + 0.0881 \\ & * \text{Masec}(n)_I\text{drp} + 0.0256 * \text{Lai}(n)_I\text{drp} + 0.0017 * \text{Qfix}_I\text{mat} - 0.0041 * \text{Ep_cum}_I\text{drp-Imat} + 0.7974 * \\ & \text{Swfac_avg}_I\text{drp-Imat} + 0.6669 * \text{Swfac_avg}_I\text{flo-Imat} + 3.8547 * \text{Inn_avg}_I\text{flo-Idrp} - 2.4314 * \text{Inn_min}_I\text{flo-Idrp} \\ & + 0.1587 \end{aligned}$$

Linear regression number 10:

$$\begin{aligned} \text{Mafruit} = & -0.1928 * \text{Masec}(n)_I\text{flo} + 0.0274 * \text{Lai}(n)_I\text{flo} + 0.0067 * \text{Qnplante}_I\text{flo} + 0.0223 * \text{Qfix}_I\text{flo} + 0.0881 \\ & * \text{Masec}(n)_I\text{drp} + 0.0256 * \text{Lai}(n)_I\text{drp} + 0.0017 * \text{Qfix}_I\text{mat} - 0.0041 * \text{Ep_cum}_I\text{drp-Imat} + 0.7974 * \\ & \text{Swfac_avg}_I\text{drp-Imat} + 0.3752 * \text{Swfac_avg}_I\text{flo-Imat} + 3.8547 * \text{Inn_avg}_I\text{flo-Idrp} - 2.4314 * \text{Inn_min}_I\text{flo-Idrp} \\ & + 0.3922 \end{aligned}$$

Linear regression number 11:

$$\begin{aligned} \text{Mafruit} = & -0.163 * \text{Masec}(n)_I\text{flo} - 0.2727 * \text{Lai}(n)_I\text{flo} + 0.0067 * \text{Qnplante}_I\text{flo} + 0.0228 * \text{Qfix}_I\text{flo} + 0.0881 \\ & * \text{Masec}(n)_I\text{drp} + 0.0528 * \text{Lai}(n)_I\text{drp} + 0.0017 * \text{Qfix}_I\text{mat} - 0.0032 * \text{Ep_cum}_I\text{drp-Imat} + 1.002 * \\ & \text{Swfac_avg}_I\text{drp-Imat} + 0.1212 * \text{Swfac_avg}_I\text{flo-Imat} + 3.8547 * \text{Inn_avg}_I\text{flo-Idrp} - 2.0948 * \text{Inn_min}_I\text{flo-Idrp} \\ & + 0.6173 \end{aligned}$$

Linear regression number 12:

$$\begin{aligned} \text{Mafruit} = & -0.163 * \text{Masec}(n)_I\text{flo} - 0.1797 * \text{Lai}(n)_I\text{flo} + 0.0067 * \text{Qnplante}_I\text{flo} + 0.0234 * \text{Qfix}_I\text{flo} + 0.0881 \\ & * \text{Masec}(n)_I\text{drp} + 0.0528 * \text{Lai}(n)_I\text{drp} + 0.0017 * \text{Qfix}_I\text{mat} - 0.0032 * \text{Ep_cum}_I\text{drp-Imat} + 1.0215 * \\ & \text{Swfac_avg}_I\text{drp-Imat} + 0.1212 * \text{Swfac_avg}_I\text{flo-Imat} + 3.8547 * \text{Inn_avg}_I\text{flo-Idrp} - 2.0948 * \text{Inn_min}_I\text{flo-} \\ & \text{Idrp} + 0.3868 \end{aligned}$$

Linear regression number 13:

$$\begin{aligned} \text{Mafruit} = & -0.0608 * \text{Lai}(n)_I\text{flo} + 0.0067 * \text{Qnplante}_I\text{flo} + 0.0881 * \text{Masec}(n)_I\text{drp} + 0.0528 * \text{Lai}(n)_I\text{drp} + \\ & 0.0017 * \text{Qfix}_I\text{mat} - 0.0032 * \text{Ep_cum}_I\text{drp-Imat} + 0.7974 * \text{Swfac_avg}_I\text{drp-Imat} + 0.0005 * \text{Precip_cum}_I\text{flo-} \\ & \text{Imat} + 0.1593 * \text{Swfac_avg}_I\text{flo-Imat} + 5.2879 * \text{Inn_avg}_I\text{flo-Idrp} - 4.1489 * \text{Inn_min}_I\text{flo-Idrp} + 0.3625 \end{aligned}$$

Linear regression number 14:

$$\begin{aligned} \text{Mafruit} = & -0.0608 * \text{Lai}(n)_I\text{flo} + 0.0067 * \text{Qnplante}_I\text{flo} + 0.102 * \text{Masec}(n)_I\text{drp} + 0.0528 * \text{Lai}(n)_I\text{drp} + \\ & 0.0017 * \text{Qfix}_I\text{mat} - 0.0032 * \text{Ep_cum}_I\text{drp-Imat} + 0.7974 * \text{Swfac_avg}_I\text{drp-Imat} + 0.0004 * \text{Precip_cum}_I\text{flo-} \\ & \text{Imat} + 0.1593 * \text{Swfac_avg}_I\text{flo-Imat} + 5.2879 * \text{Inn_avg}_I\text{flo-Idrp} - 4.1489 * \text{Inn_min}_I\text{flo-Idrp} + 0.3484 \end{aligned}$$

Linear regression number 15:

$$\begin{aligned} \text{Mafruit} = & -0.0608 * \text{Lai}(n)_I\text{flo} + 0.0067 * \text{Qnplante}_I\text{flo} + 0.1013 * \text{Masec}(n)_I\text{drp} + 0.0528 * \text{Lai}(n)_I\text{drp} + \\ & 0.0017 * \text{Qfix}_I\text{mat} - 0.0032 * \text{Ep_cum}_I\text{drp-Imat} + 0.7974 * \text{Swfac_avg}_I\text{drp-Imat} + 0.0004 * \text{Precip_cum}_I\text{flo-} \\ & \text{Imat} + 0.1593 * \text{Swfac_avg}_I\text{flo-Imat} + 5.2879 * \text{Inn_avg}_I\text{flo-Idrp} - 4.1489 * \text{Inn_min}_I\text{flo-Idrp} + 0.36 \end{aligned}$$

Linear regression number 16

$$\begin{aligned} \text{Mafruit} = & 0.0227 * \text{Masec}(n)_I\text{flo} - 0.2846 * \text{Lai}(n)_I\text{flo} + 0.0067 * \text{Qnplante}_I\text{flo} + 0.0131 * \text{Qfix}_I\text{flo} + 0.2127 \\ & * \text{Masec}(n)_I\text{drp} + 0.0528 * \text{Lai}(n)_I\text{drp} + 0.0025 * \text{Qfix}_I\text{mat} - 0.0077 * \text{Ep_cum}_I\text{drp-Imat} + 1.9247 * \\ & \text{Swfac_avg}_I\text{drp-Imat} - 1.1523 * \text{Swfac_avg}_I\text{flo-Imat} + 4.2938 * \text{Inn_avg}_I\text{flo-Idrp} - 3.7675 * \text{Inn_min}_I\text{flo-Idrp} \\ & + 1.6917 \end{aligned}$$

Linear regression number 17:

$$\begin{aligned} \text{Mafruit} = & 0.0184 * \text{Masec}(n)_I\text{flo} - 0.2846 * \text{Lai}(n)_I\text{flo} + 0.0067 * \text{Qnplante}_I\text{flo} + 0.0131 * \text{Qfix}_I\text{flo} + 0.2127 \\ & * \text{Masec}(n)_I\text{drp} + 0.0528 * \text{Lai}(n)_I\text{drp} + 0.0025 * \text{Qfix}_I\text{mat} - 0.0077 * \text{Ep_cum}_I\text{drp-Imat} + 1.9247 * \\ & \text{Swfac_avg}_I\text{drp-Imat} - 1.1523 * \text{Swfac_avg}_I\text{flo-Imat} + 4.2938 * \text{Inn_avg}_I\text{flo-Idrp} - 3.7675 * \text{Inn_min}_I\text{flo-Idrp} \\ & + 1.7071 \end{aligned}$$

Linear regression number 18:

$$\begin{aligned} \text{Mafruit} = & -0.2114 * \text{Masec}(n)_I\text{flo} - 0.2846 * \text{Lai}(n)_I\text{flo} + 0.0067 * \text{Qnplante}_I\text{flo} + 0.0143 * \text{Qfix}_I\text{flo} + 0.2127 \\ & * \text{Masec}(n)_I\text{drp} + 0.0528 * \text{Lai}(n)_I\text{drp} + 0.0025 * \text{Qfix}_I\text{mat} - 0.0077 * \text{Ep_cum}_I\text{drp-Imat} + 1.9247 * \\ & \text{Swfac_avg}_I\text{drp-Imat} - 1.1523 * \text{Swfac_avg}_I\text{flo-Imat} + 4.2938 * \text{Inn_avg}_I\text{flo-Idrp} - 3.7675 * \text{Inn_min}_I\text{flo-Idrp} \\ & + 2.2973 \end{aligned}$$

Linear regression number 19:

$$\begin{aligned} \text{Mafruit} = & -0.1891 * \text{Masec}(n)_I\text{flo} - 0.2846 * \text{Lai}(n)_I\text{flo} + 0.0067 * \text{Qnplante}_I\text{flo} + 0.0143 * \text{Qfix}_I\text{flo} + 0.2127 \\ & * \text{Masec}(n)_I\text{drp} + 0.0528 * \text{Lai}(n)_I\text{drp} + 0.0025 * \text{Qfix}_I\text{mat} - 0.0077 * \text{Ep_cum}_I\text{drp-Imat} + 1.9247 * \\ & \text{Swfac_avg}_I\text{drp-Imat} - 1.1523 * \text{Swfac_avg}_I\text{flo-Imat} + 4.2938 * \text{Inn_avg}_I\text{flo-Idrp} - 3.7675 * \text{Inn_min}_I\text{flo-Idrp} \\ & + 2.2385 \end{aligned}$$