Journal of
Applied Remote Sensing

RemoteSensing.SPIEDigitalLibrary.org

Using phenology-based enhanced vegetation index and machine learning for soybean yield estimation in Paraná State, Brazil

Jonathan Richetti
Jasmeet Judge
Kenneth Jay Boote
Jerry Adriani Johann
Miguel Angel Uribe-Opazo
Willyan Ronaldo Becker
Alex Paludo
Laíza Cavalcante de Albuquerque Silva

SPIE.

Jonathan Richetti, Jasmeet Judge, Kenneth Jay Boote, Jerry Adriani Johann, Miguel Angel Uribe-Opazo, Willyan Ronaldo Becker, Alex Paludo, Laíza Cavalcante de Albuquerque Silva, "Using phenology-based enhanced vegetation index and machine learning for soybean yield estimation in Paraná State, Brazil," *J. Appl. Remote Sens.* **12**(2), 026029 (2018), doi: 10.1117/1.JRS.12.026029.

Using phenology-based enhanced vegetation index and machine learning for soybean yield estimation in Paraná State, Brazil

**Jonathan Richetti,[a,b,*] Jasmeet Judge,[b] Kenneth Jay Boote,[c]
Jerry Adriani Johann,[a] Miguel Angel Uribe-Opazo,[a]
Willyan Ronaldo Becker,[a] Alex Paludo,[a] and
Laíza Cavalcante de Albuquerque Silva[a]**
[a]State University of West Paraná, Applied Statistics Laboratory,
Department of Agricultural Engineering, Cascavel, Brazil
[b]University of Florida, Center for Remote Sensing, Department of Agricultural and
Biological Engineering, Gainesville, Florida, United States
[c]University of Florida, Department of Agricultural and Biological Engineering,
Gainesville, Florida, United States

Abstract. Accurate and timely regional estimates of agricultural production are key for decision makers. This study aims to understand how different machine learning techniques impact soybean yield estimation in extracting maximum information from remotely sensed MODIS enhanced vegetation index (EVI) that is constrained by phenology. Specifically, a methodology is developed for incorporating phenological information aligned with EVI acquisition for each pixel and selecting the most significant predictors out of 36 predictors using feature selection. These predictors were then used in four machine learning algorithms (MLA) to obtain soybean yield estimates for observed farms in the Paraná State, Brazil. The optimal MLA was then implemented for the whole state to obtain regional soybean yield. The gradient boosting model (GBM) with all 36 predictors performed well with a mean difference of 3.5 $\mathrm{kg\,ha^{-1}}$, an RMSD of 373 $\mathrm{kg\,ha^{-1}}$, and Willmott's d of 0.85, however, the random forest (RF) algorithm using five optimal EVI predictors presented similar results, but with considerably less computational time. Both GBM and RF provided higher regional yields compared to the officially reported yields by 1775×10^3 and 2059×10^3 metric tons, respectively. The RF with five EVI predictors provided the best results for regional soybean estimations, considering the accuracy and computational performances. *© 2018 Society of Photo-Optical Instrumentation Engineers (SPIE)* [DOI: 10.1117/1.JRS.12.026029]

Keywords: data-driven yield estimation; gradient boosting model; generalized linear model; random forests; extreme machine learning; Gaussian process.

Paper 180131 received Feb. 11, 2018; accepted for publication May 17, 2018; published online Jun. 15, 2018.

1 Introduction

Reliable and timely information on agricultural production is essential for ensuring world food security.[1] Generally, census and surveys are the main tools for agricultural statistics. The census of agriculture is one of the key pillars of a national statistical system, and in many developing countries, it is often the only means of producing statistical information on the structure and other relevant aspects of the agriculture sector.[2] However, the time and cost of ground surveys and census are directly proportional to the size of the area being surveyed. Satellite remote sensing data can provide timely, accurate, and objective information on a cultivated area by

*Address all correspondence to: Jonathan Richetti, E-mail: j_richetti@hotmail.com; jonathan.richetti@unioeste.br crop type and, in turn, facilitate accurate estimates of agricultural production.[3] Also, satellite observations, owing to their synoptic and repetitive nature, have the unique advantage of providing spatially contiguous information on crop growth at local, regional, and global scales.[1]

Remote sensing data have been used for estimating agricultural productivity, either as inputs, integration, and assimilation data, or in data driven models, such as traditional statistical models and machine learning. For example, Gusso et al.[4] used the enhanced vegetation index (EVI) from the moderate resolution imaging spectroradiometer (MODIS) as inputs to a coupled model to estimate soybean yield. The soybean yields in southern Brazil from the model obtained differences of less than 15% when compared to official statistics.[5] El Hajj et al.[6] used leaf area index (LAI) from MODIS, weather from ground stations, and soil data from as inputs to the Boreal Ecosystem Productivity Simulator for rice yield estimation within the middle and lower reaches of the Yangtze River with errors lower than 10% compared to official data. Li et al.[7] integrated remote-sensing-derived parameters (LAI, harvest and irrigation dates) with *in situ* data in a crop model that simulates vegetation growth for hay crops in different scenarios, concluding that incorporating remote-sensing-derived estimates of the initial and maximal LAI values may spare costly *in situ* measurements while still ensuring correct model execution with root mean square error of $410\,\mathrm{kg\,ha^{-1}}$ and mean absolute percentage error of 22%. Chakrabarti et al.[8] assimilated LAI information into the wheat CERES-model by employing particle filters and the proper orthogonal decomposition-based ensemble four-dimensional variational strategies in Hengshui city, China, obtaining relative errors lower than 8%. They assimilated downscaled remote sensing soil moisture from the soil moisture and ocean salinity mission in the DSSAT-CROPGRO model using an ensemble Kalman filter-based augmented state-vector technique that estimates states and parameters simultaneously. The framework was implemented in La Plata basin in Brazil for 2 years and the root mean square differences (RMSD) between the assimilated and observed crop yields were 16.8% during the first growing season and 4.37% during the second season.

Data driven models, including statistical and machine learning algorithms, can be used to attain yield estimates based upon remote sensing information. For example, Bolton and Friedl[9] used the MODIS nadir bidirectional reflectance distribution function adjusted surface reflectance (NBAR) data to calculate normalized difference vegetation index (NDVI), EVI, and normalized difference water index as inputs in a linear model to forecast soybean and maize yields in central United States and concluded that including information related to crop phenology, e.g., the emergence date, improved yield estimations. Aligning MODIS EVI acquisition dates closely with phenology at each pixel may further improve yield predictions.[10] Jeong et al.[11] used multiple linear regression, M5-Prime regression trees, perceptron multilayer neural networks, support vector regression, and k-nearest neighbor methods using climate data as inputs for yield prediction. Kim and Lee[12] tested random forest (RF) and multiple linear regression for global and regional estimates of maize (grain and silage), potato, and wheat based in climate and management information. Bose et al.[13] tested support vector machine, RF, extremely randomized trees, and deep learning for corn estimation in Iowa based on remote sensing and climate information as inputs. Fieuzal et al.[14] used spiking neural networks for remote sensing spatiotemporal analysis of image time series of NDVI for winter wheat yield estimation in China obtaining a mean error of 235 kg/ha and an overall accuracy of 95%. They presented two methods to estimate yields using artificial neural networks in southwestern France. A diagnostic approach based on all the satellite data acquired throughout the agricultural season and a real-time approach, where estimates are updated after each image was acquired in the microwave and optical domains (Formosat-2, Spot-4/5, TerraSAR-X, and Radarsat-2) throughout the crop cycle with a relative error of 6%. In spite of recent data driven algorithms used in various studies, the impact of different data driven algorithms on yield estimation is still not well understood.

This study aims to understand how different machine learning techniques impact soybean yield estimation in extracting maximum information from EVI that is constrained by phenology. Specific objectives of this study are to (1) develop a methodology for incorporating phenological information aligned with MODIS EVI acquisition for each pixel, (2) compare the performance of four widely used machine learning algorithms, viz., stochastic gradient boosting model (GBM), generalized linear model (GLM), RF, and Gaussian process (GP), for soybean yields in one of the primary soybean producing regions in Brazil—Paraná State, and (3) implement the optimal algorithm from objective (2) to obtain regional soybean yield for the whole state of Paraná.

2 Material and Methods

2.1 *Study Area*

The study was conducted in Paraná State in southern Brazil, located between parallels 22°29′S and 26°43′S and the meridians 48°2′W and 54°38′W (Fig. 1) with an area of 199,307.945 km^2. The climatic regions in Paraná include tropical wet and dry, monsoon-influenced humid subtropical climate, humid subtropical climate, and temperate oceanic climate.[15] The soils in the state are predominantly Oxisols, clay, and neo soils.[16] Paraná State is responsible for almost 18% of the Brazilian production and the state produces more soybeans than China, the fourth great world producer.[17] Typically, soybean is sown from September to November and harvested from January to April. The 2013/2104 growing season was affected by low rainfall and high temperatures, resulting in lower yields compared to previous years.[18]

2.2 *Dataset*

In this study, observations of yield (kg ha^{-1}), and sowing and harvest dates were collected from 86 commercial rainfed farms during the crop-season in 2013/2014 (Fig. 1). In addition, 37 EVI images from September 1 to April 15 from MODIS Terra and Aqua products, MOD13Q1 and MYD13Q1, were used. These products have a spatial resolution of 250 m and, combined, a temporal resolution of 8-day. The study area encompassed 1253 MODIS pixels and with an average farm size of 102 ha, each farm covered about 16 MODIS pixels. The sowing and harvest dates for the region were estimated based on MODIS EVI time-series following. Johann et al.[21] determined the sowing and the harvest dates based upon the temporal changes in EVI for each MODIS pixel. The peak of the vegetation coincided with maximum EVI value. The time at which EVI was minimum, prior to the peak of vegetation, was considered as the sowing date and the minimum EVI after the peak was considered as the harvest date. The mid development was at the mid point of sowing and harvest dates. For this study, the mean differences (MDs) between the estimated and the observed values were 4 and 24 days for sowing and harvest dates, respectively. These differences were reasonable because farmers usually either report the period for the activity (sowing or harvest) or the activity start date.

2.3 *Methodology Overview*

For this study, phenology-based EVI predictors were created by incorporating phenology dates into the MODIS EVI data. These predictors were normalized using Eq. (1) and used as inputs for

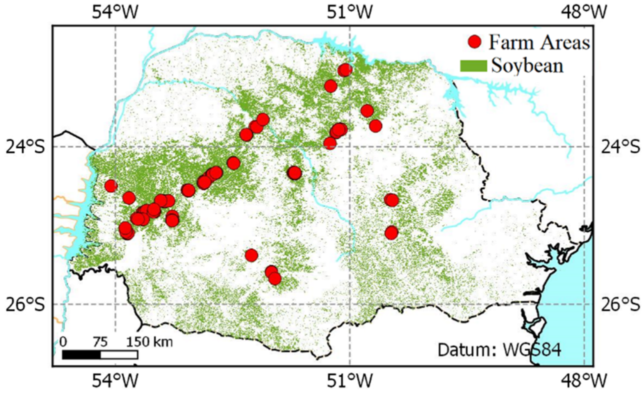


Fig. 1 Study area of the Paraná State in Brazil. The red circles represent the 86 farms that were observed and the green regions are the soybean growing areas in the state during the 2013 to 2014 season.[19,20]

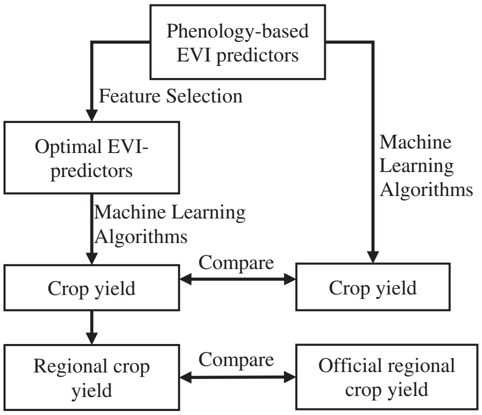

Phenology-based EVI predictors
Feature Selection
Optimal EVI-predictors
Machine Learning Algorithms
Machine Learning Algorithms
Crop yield
Compare
Crop yield
Regional crop yield
Compare
Official regional crop yield

Fig. 2 Yield estimation from the Phenology-based EVI used in this study from four MLAs (GBM, GLM, RF, and GP).

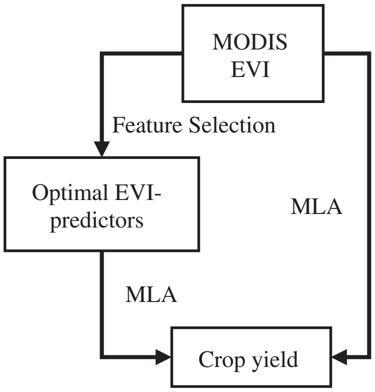


Fig. 3 Yield estimation based upon nonphenology-based MODIS-EVI as inputs to the MLA.

the four machine learning algorithms. In addition, feature selection based on recursive feature selection was used to determine the optimal predictors related to yield:

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}, \tag{1}$$

where z is the normalized value and x is the predictor value. The selected phenology-based EVI predictors were used in the machine learning algorithms to estimate yield. The estimated yields were spatially averaged and compared with observed yields and official governmental data, as shown in the flow chart of the methodology in Fig. 2.

To demonstrate the added value of using phenology-based EVI, the yield estimates were compared those obtained using MODIS EVI (nonphenology-based), as shown in Fig. 3.

2.4 *Phenology-Based EVI*

Rather than using MODIS products directly as inputs, EVI predictors were created from MODIS time series data considering the phenological information, such as sowing, peak vegetative, mid-development, and harvest. As shown in Fig. 4, for each pixel, the EVI information was extracted at a specified phenological date making the information strongly correlated to phenology. For example, if adjacent farms, each consisting of multiple MODIS pixels, have different sowing dates, e.g., October 8, 9, 16, 17, 18, and 24, as shown in Fig. 4, then the "EVI-sowing" for all the pixels in one farm will be obtained from the MODIS product acquired closest to the planting dates, such as image acquired on October 8 for the fields planted on Oct 8 and 9, image on Oct 16 for fields planted on Oct 16, 17, and 18, and image on Oct 24 for the field planted on Oct 24.

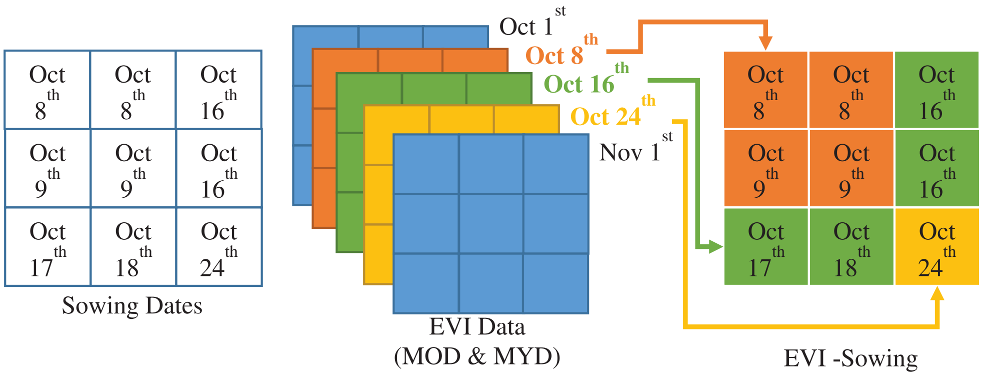


Fig. 4 Illustrational example for incorporating phenology in EVI to create a phenology-based variable: EVI on sowing date.

Therefore, each pixel has a closer relation to phenological date in that area instead of the EVI image for a given calendar date. Using this methodology, 36 phenology-based EVI variables (Table 1), viz., 5 around sowing, 10 around peak vegetative, 12 around mid-development, 5 around harvest, and 4 for the season length, were created from the EVI profile. These 36 variables served as inputs to the feature selection and to the machine learning algorithms, as shown in Fig. 2.

2.5 *Recursive Feature Elimination*

Feature selection facilitates data visualization and data understanding and improves prediction performance.[22] The recursive feature elimination (RFE) implements a backward selection of predictors based on predictor importance ranking. The predictors are ranked and the less important ones are sequentially eliminated prior to modeling. The goal is to find a subset of predictors that can be used to produce an accurate model.[23] In this study, the RFE with five repetitions, each using 10-fold cross-validation were applied on the 36 phenology-based EVI variables to obtain the optimal ranked predictors. The algorithm used RF to estimate yield. The RMSD shown in between estimated and observed yields [Eq. (4)] was used to rank the predictors for the RFE. The normalized RMSD [Eq. (5)] is a dimensionless number between 0 and 1, the normalized RMSD with a threshold of 0.3, or 30% was used as a metric for selecting the predictors.

2.6 *Machine Learning Algorithms*

Four machine learning algorithms, stochastic GBM, GLM, RF, and GP were used to estimate soybean yield.

(1) GBM constructs additive regression models by sequentially fitting a simple parameterized function (base learner) to current "pseudo"-residuals by least squares at each iteration. The pseudo-residuals are the gradient of the loss functional being minimized, with respect to the model values at each training data point evaluated at the current step.[24] The tuning parameters that need adjustment are the number of boosting iterations (N trees), the maximum tree depth (interaction depth), the shrinkage, and the minimum terminal node size (nminobsinnode).

(2) GLM is a technique of iterative weighted linear regression can be used to obtain maximum likelihood estimates of the parameters with observations distributed according to some exponential family and systematic effects that can be made linear by a suitable transformation. A generalization of the analysis of variance is given for these models using log-likelihoods. These GLMs can be used in different distributions.[25] There is no tuning parameter for this method.

(3) RF is a regression tree technique, where a number of randomly constructed bags and trees are generated in parallel. As in bagging, a number of decision trees on bootstrapped training samples are built. When building these trees, each time a split in a tree is

Table 1 Description of the 36 predictors used as input in the machine learning regression models.

	Variable description
1	Cycle length (# of days from sowing to harvest)
2	EVI on sowing
3	EVI one scene before sowing
4	EVI two scene before sowing
5	EVI three scene before sowing
6	Total EVI on sowing
7	EVI from sowing to peak
8	EVI one scene before the middle date
9	EVI two scene before the middle date
10	EVI three scene before the middle date
11	EVI one scene after the middle date
12	EVI two scene after the middle date
13	EVI three scene after the middle date
14	Total EVI on middle
15	Partially total EVI on middle
16	Partially total EVI on middle
17	EVI at peak
18	EVI one scene after peak
19	EVI two scene after peak
20	EVI three scene after peak
21	EVI one scene before peak
22	EVI two scene before peak
23	EVI three scene before peak
24	Total EVI on peak
25	Center peak Total EVI around peak 1
26	Center peak Total EVI around peak 2
27	EVI from peak to harvest
28	EVI on harvest
29	EVI one scene before harvest
30	EVI two scene before harvest
31	EVI three scene before harvest
32	Total EVI on harvest
33	EVI on harvest
34	EVI total (sum of all EVI's)
35	EVI total without the total EVI from middle
36	EVI total without the total EVI from peak

considered, a random sample of predictors is chosen as split candidates from the full set of predictors.[26] Each predicted class is voted and the forest prediction is the class that gets the most votes, for classification or the average for regression.[27] The tuning parameter used was the number of randomly selected predictors (mtry) and the number of trees (ntrees).

(4) GP is a generalization of the Gaussian probability distribution, whereas a probability distribution describes random variables, which are scalars or vectors (for multivariate distributions), a stochastic process governs the properties of functions. Leaving mathematical sophistication aside, one can loosely think of a function as a very long vector, each entry in the vector specifying the function value $f(x)$ at a particular input x.[28] There is no tuning parameter for this method.

A training set, consisting of 70% of the data (878 pixels), was used for parameter adjustment/tuning of each of the four MLA with 10 repetitions, each repetition using a 10-fold cross-validation. The optimal combination of parameters for each model was chosen based on the lowest RMSD and higher R^2 between estimated and observed yield. The models were tested using the remaining 30% of the data (375 pixels) that were not used for training, and the accuracy analysis was performed. The R version 3.3.4[29] with the package CARET[23] was used, and the seed was set to 153 to ensure reproducible results.

2.7 *Accuracy Assessment*

The accuracy metrics of mean absolute difference (MAD), MD, RMSD, the enhanced Willmott concordance index (Willmott's d),[30] as shown in Eqs. (2)–(6), respectively, and the Pearson's correlation were calculated for assessing the differences between the MLA and the observed yields in kg ha^{-1} from the farms:

$$\mathrm{MAD} = \frac{1}{n}\sum_{i=1}^{n}|Y_{\mathrm{est}} - Y_{\mathrm{obs}}|, \tag{2}$$

$$\mathrm{MD} = \frac{1}{n}\sum_{i=1}^{n}(Y_{\mathrm{est}} - Y_{\mathrm{obs}}), \tag{3}$$

$$\mathrm{RMSD} = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(Y_{\mathrm{est}} - Y_{\mathrm{obs}})^2\right)}, \tag{4}$$

$$\text{Normalized RMSD} = \frac{\mathrm{RMSD} - \mathrm{RMSD}_{\mathrm{min}}}{\mathrm{RMSD}_{\mathrm{max}} - \mathrm{RMSD}_{\mathrm{min}}}, \tag{5}$$

$$d = 1 - \frac{\sum_{i=1}^{n}|Y_{\mathrm{est}} - Y_{\mathrm{obs}}|}{2 * \sum_{i=1}^{n}|Y_{\mathrm{obs}} - \overline{Y_{\mathrm{obs}}}|}, \tag{6}$$

where Y_{est} is estimated yield; Y_{obs} is observed yield; and n is the total number of pixels. To determine total regional production (kg) in the state, the estimated yield (kg ha^{-1}) in each MODIS pixel was multiplied by the area of the pixel (6.25 ha) to obtain production for each pixel. These per pixel production values were combined to obtain the total regional production. In addition, the computational time using one core of an Intel i5-2500 processor for each MLA was recorded.

3 Results and Discussions

3.1 *Feature Selection*

Figure 5 shows the normalized RMSD between estimated and observed yields with increasing number of predictors from 1 to 36. From the normalized RMSD and threshold of 30% (Fig. 5), the five optimal ranked predictors provided sufficient information to obtain acceptable estimates and were used in the MLA. In addition, a further increase in the number of predictors to 10 increased the computational time of the MLA. These five optimal predictors were EVI 16 days prior the middle of the cycle, EVI 24 days after the vegetative peak, the cycle length (the cycle length varied from 98 to 229 days), EVI 24 days prior the middle of the cycle, and EVI 16 days after the vegetative peak.

3.2 *Machine Learning Algorithms*

For each MLA, the respective parameters were adjusted (Table 2) based on the 10-fold cross-validation with five repetitions and these parameters were used for the implementation of the models.

Table 3 shows the performance metrics MD, MAD, RMSD, Willmott's *d*, Pearson's *r*, and the processing times for the MLA used in this study. In general, all MLAs presented satisfactory performance with MD < 20 $\mathrm{kg\,ha^{-1}}$. The GBM, on average, presented a small overestimation by

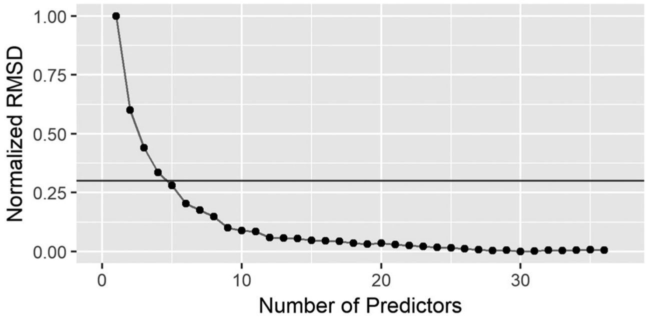


Fig. 5 Normalized RMSD between estimated and observed soybean yields as a function of number of predictors used. The horizontal line shows the threshold values of RMSD at 30%.

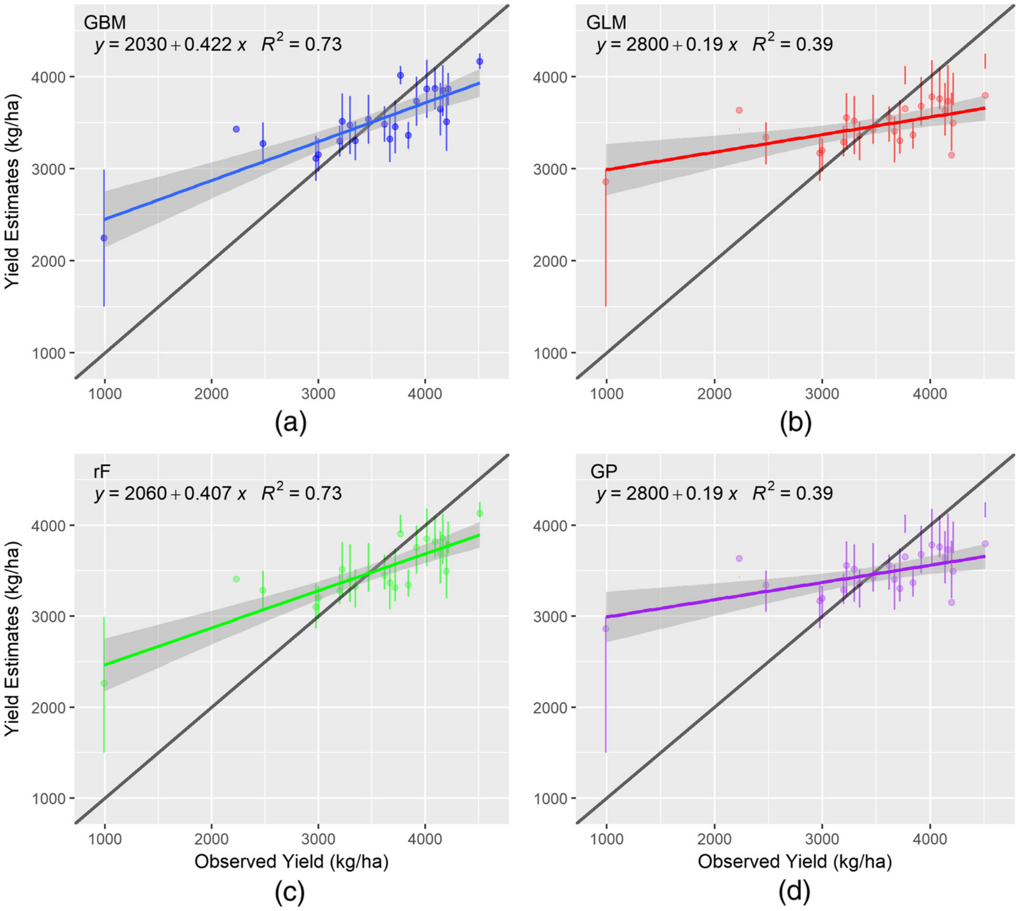
Table 2 Adjusted parameters for each model using the five optimal predictors and all 36 predictors.

MLA	Parameters with five selected predictors		Parameters with all 36 predictors	
GBM	Interaction depth	3	Interaction depth	10
	N trees	100	*N* trees	400
	Shrinkage	0.1	Shrinkage	0.1
	Nminobsinnode	10	Nminobsinnode	10
GLM	—	—	—	—
RF	Mtry	3	Mtry	17
	Ntrees	550	ntrees	550
GP	—	—	—	—

Note: No parameters to be adjusted. *N* trees: number of boosting iterations; interaction depth: the maximum tree depth; nminobsinnode: the minimum terminal node size; mtry: number of randomly selected predictors; ntrees: number of trees.

Table 3 Accuracy analysis of all MLAs with the five optimal predictors and with all 36 predictors. Bold values represent the optimal result.

	GBM	GLM	RF	GP
Statistics		Five predictors		
MD (kg ha^{-1})	10.73	−13.28	**−7.06**	−13.38
MAD (kg ha^{-1})	293.35	336.72	**274.02**	336.64
RMSD (kg ha^{-1})	403.38	461.75	**390.16**	461.73
Willmott's *d*	0.79	0.66	**0.79**	0.66
Pearson's *r*	0.68	0.54	**0.70**	0.54
Computational Time (s)	168.88	**1.08**	486.50	51.39
		All 36 predictors		
MD (kg ha^{-1})	**3.52**	−18.13	−8.16	−17.50
MAD (kg ha^{-1})	259.13	353.68	**252.41**	343.01
RMSD (kg ha^{-1})	**372.56**	475.14	395.38	463.55
Willmott's *d*	**0.85**	0.70	0.80	0.71
Pearson's *r*	**0.74**	0.53	0.69	0.55
Computational time (s)	989.68	**2.05**	4683.89	46.63



(a) (b) (c) (d)

Fig. 6 Scatterplot by farm of observed yields compared with yields estimated with the five selected predictors from Recursive Feature Elimination using (a) GBM, (b) GLM, (c) rF, and (d) GP algorithms.

0.31% of the observed yield (3471 $\mathrm{kg\,ha^{-1}}$, as shown in Table 5), using five selected predictors, and only by 0.10% of the observed yield with all 36 predictors. All other MLAs presented underestimations, with the RF model underestimating yield by 0.20% of the observed yield with five selected EVI-predictors and by 0.24% of the observed yield with all 36 predictors. The fastest MLA (GLM) took little more than 1 s to compute and the slowest (RF) took little more than 8 min when using the five optimal predictors. When using all predictors, the fastest one (GLM) took little more than 2 s and the slowest (RF) almost 80 min. In general, the use of the five optimal predictors greatly reduced the computational time, but with some loss in the accuracy. However, RF showed an improvement in all metrics except MAD and was almost 10 times faster when using the five optimal predictors.

Figures 6 and 7 show the scatterplots by farm of yields estimated by the MLA and those observed, when using five predictors and all 36 predictors, respectively. The RF with 5 optimal predictors (Fig. 6) and the GBM with all 36 predictors (Fig. 7) presented the least dispersion, as shown by low RMSD, and more accurate, as shown by low MD and MAD, when compared to the other MLAs. The GLM and GP MLAs presented more variability for low yield and less variability for high yields, not being able to completely capture the changes from one farm to the other (Willmott's d around 0.5 and $r < 0.75$—Table 3). The RF with 5 selected predictors and GBM with all 36 predictors better capture the changes as seen in higher r and d values when compared to other MLAs. Other similar studies, such as Refs. 31 and 32, found similar Pearson's r of 0.96 and 0.74, respectively.

Table 4 shows the comparison of yield results from the two best performing MLAs, GBM, and RF, using phenology-based EVI and nonphenology-based EVI. The MD was reduced by 6.43 kg/ha and the Willmott's d index was increased by 0.13 in the GBM algorithm. Similar improvements were also obtained for the RF algorithm.

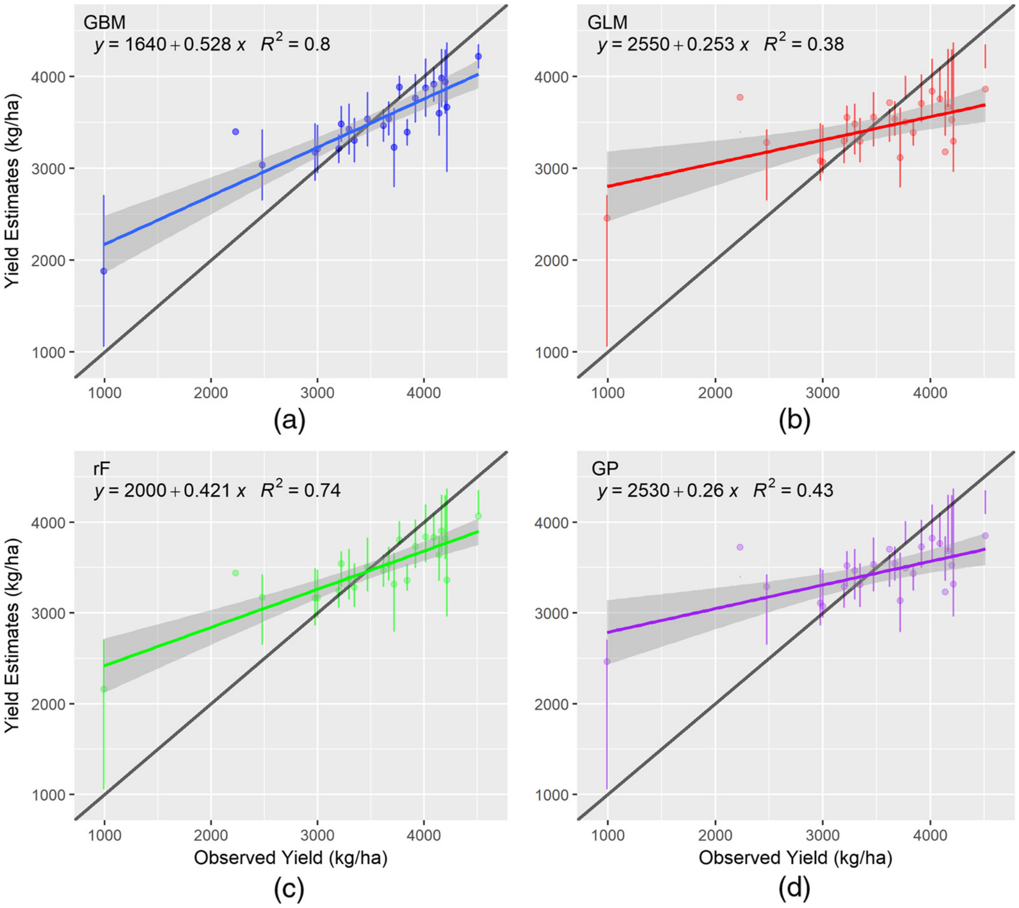


Fig. 7 Scatterplot by farm of observed yields compared with yields estimated with all 36 predictors using (a) GBM, (b) GLM, (c) rF, and (d) GP algorithms.

Table 4 Comparison of yield estimations using the phenology-based EVI and the MODIS EVI (nonphenology-based) with the GBM and RF algorithms.

	GBM		RF	
	Nonphenology-based	Phenology-based	Nonphenology-based	Phenology-based
MD (kg ha^{-1})	9.95	3.52	−11.97	−7.06
MAD (kg ha^{-1})	375.35	259.13	350.28	274.02
RMSD (kg ha^{-1})	553.82	372.56	525.33	390.16
Willmott's *d*	0.72	0.85	0.73	0.79
Pearson's *r*	0.58	0.74	0.64	0.70
Computational time (s)	2127.75	989.68	566.72	486.50

3.3 *Regional Soybean Yield*

Both GBM and RF were applied to the whole Paraná State. The GBM with all 36 predictors presented higher deviation with a coefficient of variation 5.8% higher than the RF with five optimal predictors, however, both MLAs have similar descriptive statistics except for the GBM minimum estimated yield that is almost 10 times lower than the RF minimum estimated yield (Table 5).

For both GMB (Fig. 8) and RF (Fig. 9), the west of the state provided highest yields following the high yield soybean belt from west to northeast, especially for RF. In both figures, there is predominance of yield from 3000 to 3750 for the two MLAs, with yellow, light, and dark green pixels, in the whole state. The GBM presented more areas with yields lower than 3000 kg ha^{-1} concentrated in the extreme northeast area but spread throughout the state while RF had fewer pixels showing less than 3000 kg ha^{-1} and also are concentrated in the very northeast region.

Comparing the average of 2950 kg ha^{-1} reported by CONAB[18] with the GBM estimated higher yield than the officially reported yield in 394 kg ha^{-1}, however, the difference between the average observed yield in 86 farms and the estimated yield by GBM was only 3.52 kg ha^{-1}. With the GBM model, the total estimated production of the state is 16,579.306 thousands of metric tons produced by the state in the 2013/2014 season and 1774.7 thousands of metric tons higher than the official report by CONAB.[18] The RF overestimated the average officially reported yield by 451 kg ha^{-1}, however, the difference between the average actual yield and the estimated yield by RF was only −7.02 kg ha^{-1} from the 86 farms. With the RF model, the total

Table 5 Descriptive statistics of crop yields (kg ha^{-1}) from farms using the GBM and rF algorithms.

Statistics	Actual (kg ha^{-1})	GBM (kg ha^{-1})	RF (kg ha^{-1})
Minimum	991.7	147.5	1102
First quartile	3223	3156	3223
Median	3471	3432	3383
Third quartile	3917	3672	3596
Maximum	4512	4915	4451
Mean	3471	3344	3401
Standard deviation	579.6	538.5	349.5
Coefficient of variation (%)	16.6	16.1	10.3

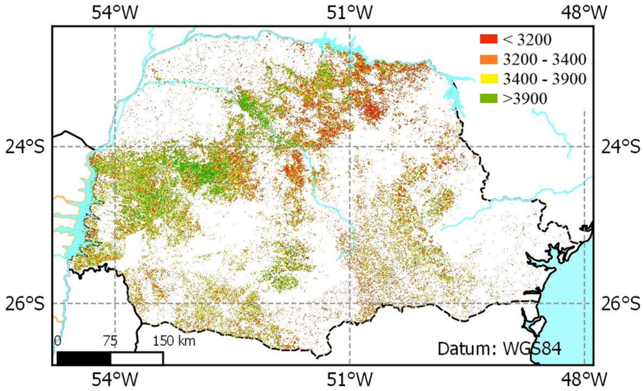


Fig. 8 Soybean yield estimation for the whole state for crop season 2013 to 2014 by the GBM algorithm using all predictors.

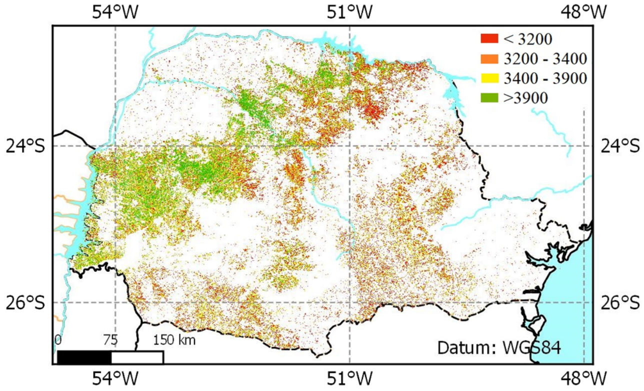


Fig. 9 Soybean yield estimation for the whole state for crop season 2013 to 2014 by the rF algorithm using five optimal predictors.

estimated production of the state is 16,864.383 thousands of metric tons produced in the 2013/2014 season and 2,059.8 thousands of metric tons more than the official report by CONAB.[18] Both the MLAs estimated higher yields than the officially reported yields. The RF presented similar results as the GBM with lower inputs and lower computational time. In addition, even though both algorithms estimated higher production than officially reported data, the estimated yield matches well with the observed yield at the farms. Interestingly, the average yield of 86 farms was 521 $\mathrm{kg\,ha^{-1}}$ higher than the regionally reported yield. The RF algorithm with five EVI predictors is recommended for soybean yield estimation in this region.

4 Conclusions

In this study, a method was developed for incorporating phenological information into the EVI data for soybean yield estimation using machine learning algorithms. Phenology was incorporated in MODIS EVI obtaining 36 predictors. A RFE-based feature selection was performed to obtain five optimal predictors. The five optimal predictors were: EVI 16 days prior the middle of the cycle, EVI 24 days after the vegetative peak, the cycle length, EVI 24 days prior the middle of the cycle, and EVI 16 days after the vegetative peak. Four machine learning methods were implemented using all 36 predictors and the five optimal predictors in one of the primary soybean producing regions in Brazil—Paraná State. The GBM presented optimal performance with all 36 EVI predictors, with low MDs of 3.52 $\mathrm{kg\,ha^{-1}}$, an RMSD of 373 $\mathrm{kg\,ha^{-1}}$, and d of 0.85. However, the RF presented similar results using only the five optimal EVI predictors from the feature selection but with considerably reduced computational time. Incorporating phenology in EVI MODIS data provided improved yield estimations compared to the nonphenology-based remote sensing data to the MLA. Therefore, the phenology-based EVI can be used in MLA obtaining accurate regional estimates of soybean yields. In addition, the use of feature selection considerably reduced the computational time without losing accuracy when using the RF algorithm.

Disclosure

No potential conflict of interest was reported by the authors.

Acknowledgments

This work was supported by the CAPES foundation (Coordination for the Improvement of Higher Education Personnel—Ministry of Education—Brazil) under Grant 88881.131979/2016-01. In addition, the authors are grateful to the farmers for sharing their management and yield data.

References

1. X.-P. Song et al., "National-scale soybean mapping and area estimation in the United States using medium resolution satellite imagery and field survey," *Remote Sens. Environ. J.* **190**, 383–395 (2017).
2. FAO, *World Programme for the Census of Agriculture 2020. Volume 1 Programme, Concepts and Definitions*, FAO, Rome (2015).
3. L. King et al., "A multi-resolution approach to national-scale cultivated area estimation of soybean," *Remote Sens. Environ.* **195**, 13–29 (2017).
4. A. Gusso et al., "Spectral model for soybean yield estimate using MODIS/EVI data," *Int. J. Geosci.* **4**, 1233–1241 (2013).
5. F. Yao et al., "Estimation of rice yield with a process-based model and remote sensing data in the middle and lower reaches of Yangtze River of China," *J. Indian Soc. Remote Sens.* **45**, 477–484 (2017).
6. M. El Hajj et al., "Agricultural water management integration of remote sensing derived parameters in crop models: application to the PILOTE model for hay production," *Agric. Water Manage.* **176**, 67–79 (2016).
7. H. Li et al., "Improving winter wheat yield estimation from the CERES-wheat model to assimilate leaf area index with different assimilation methods and spatio-temporal scales," *Remote Sens.* **9**(3), 190–215 (2017).
8. S. Chakrabarti et al., "Assimilation of SMOS soil moisture for quantifying drought impacts on crop yield in agricultural regions," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **7**(9), 3867–3879 (2014).
9. D. K. Bolton and M. A. Friedl, "Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics," *Agric. For. Meteorol.* **173**, 74–84, Elsevier B.V. (2013).
10. A. Gonzalez-Sanchez, J. Frausto-Solis, and W. Ojeda-Bustamante, "Predictive ability of machine learning methods for massive crop yield prediction," *Spanish J. Agric. Res.* **12**, 313–328 (2014).
11. J. H. Jeong et al., "Random forests for global and regional crop yield predictions," *PLoS One* **11**(6), 1–5 (2016).
12. N. Kim and Y.-W. Lee, "Machine learning approaches to corn yield estimation using satellite images and climate data: a case of Iowa State," *J. Korean Soc. Surv. Geod. Photogramm. Cartogr.* **34**(4), 383–390 (2016).
13. P. Bose et al., "Spiking neural networks for crop yield estimation based on spatiotemporal analysis of image time SERIES," *IEEE Trans. Geosci. Remote Sens.* **54**(11), 6563–6573 (2016).

14. R. Fieuzal, C. Marais Sicre, and F. Baup, "Estimation of corn yield using multi-temporal optical and radar satellite data and artificial neural networks," *Int. J. Appl. Earth Obs. Geoinf.* **57**, 14–23 (2017).
15. L. E. O. de Aparecido et al., "Köppen, Thornthwaite and Camargo climate classifications for climatic zoning in the State of Paraná, Brazil," *Ciência e Agrotecnol.* **40**(4), 405–417 (2016).
16. EMBRAPA, *Sistema Brasileiro de Classificação de Solos*, 2nd ed., EMBRAPA, Rio de Janeiro (2009).
17. FAO, "Production quantities by country," *Statistics Division*, 2015, http://faostat3.fao.org/browse/Q/QC/E (11 August 2015).
18. CONAB, *Acomp. safra bras. grãos, v. 1—Safra 2013/14, n. 10—Décimo Levantamento*, CONAB, Brasília (2014).
19. C. H. W. de Souza et al., "Mapping and discrimination of soya bean and corn crops using spectro- temporal profiles of vegetation indices," *Int. J. Remote Sens.* **36**(7), 1809–1824 (2015).
20. D. M. Grzegozewski et al., "Mapping soya bean and corn crops in the State of Paraná, Brazil, using EVI images from the MODIS sensor," *Int. J. Remote Sens.* **37**(6), 1257–1275 (2016).
21. J. A. Johann et al., "Uso de Imagens do Sensor Orbital MODIS na Estimação de Datas do Ciclo de Desenvolvimento da Cultura da Soja para o Estado do Paraná—Brasil," *J. Braz. Assoc. Agric. Eng.* **36**(1), 126–142 (2016).
22. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).
23. M. Kuhn, "Building predictive models in R using the caret package," *J. Stat. Softw.* **28**(5), 1–26 (2008).
24. J. H. Friedman, "Stochastic gradient boosting," *Comput. Stat. Data Anal.* **38**, 367–378 (2002).
25. J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," *J. R. Stat. Soc. Ser. A* **135**(3), 370–384 (1972).
26. G. James et al., *An Introduction to Statistical Learning: with Applications in R*, 1st ed., p. 429, Springer, London (2013).
27. L. Breiman, "Statistical modeling: the two cultures," *Stat. Sci.* **16**(3), 199–231 (2001).
28. C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, Massachusetts (2006).
29. R Core Team, "R: a language and environment for statistical computing," 3.3.4, Foundation for Statistical Computing, Vienna, Austria (2017).
30. C. J. Willmott, "On the validation of models," *Phys. Geogr.* **2**(2), 219–232 (1981).
31. J. Betbeder, R. Fieuzal, and F. Baup, "Assimilation of LAI and dry biomass data from optical and SAR images into an agro-meteorological model to estimate soybean yield," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **9**(6), 2540–2553 (2016).
32. A. Gusso, D. Arvor, and J. R. Ducati, "Model for soybean production forecast based on prevailing physical conditions," *Pesqui. Agropecuária Bras.* **52**(2), 95–103 (2017).

Jonathan Richetti is a PhD student in agricultural engineering at the State University of West Paraná. His research interests include application of remote sensing data to agricultural production, mapping, and land use. He was a visitor research scholar at the Center of Remote Sensing in the Agricultural and Biology Engineering Department in the University of Florida.

Jasmeet Judge is currently the director of the Center for Remote Sensing and an associate professor in the Agricultural and Biological Engineering Department, University of Florida. Her research interests include microwave remote sensing applications to agriculture under dynamic conditions; modeling of energy and moisture interactions at the land surface and in the vadose zone; spatial and temporal scaling of remotely sensed observations in heterogenous landscapes, and data assimilation.

Kenneth Jay Boote is professor Emeritus of Agronomy, specializing in measuring and modeling crop response to climatic factors. He has developed process-based crop simulation models for soybean, peanut, dry bean, and forages as part of the DSSAT family models. He participates

in crop modeling projects especially the Agricultural Model Improvement and Intercomparison Project, for which he is co-coordinator where he advises scientists on multi-model intercomparison, and improving of models for response to climate change.

Jerry Adriani Johann is an adjunct professor at State University of West Paraná of undergraduate and graduate courses in Agricultural Engineering (PGEAGRI - MSc and PhD). He is a CNPq level 1B research fellow. The research is related to agribusiness, mainly in the following areas: GIS, remote sensing, crop forecasting, precision agriculture, geostatistics, spatial statistical areas, agro-meteorological data, and data mining.

Miguel Angel Uribe-Opazo is an associate professor at the State University of West Paraná. Experiences in the areas of Theory and Statistical Methods, Spatial Statistics and Precision Agriculture. He work as visiting professor in the doctorate program in Statistics of the University of Valparaiso-Chile since 2012, in the area of Space Statistics. He is a CNPq level 1B research fellow.

Willyan Ronaldo Becker is a PhD student in the State University of West Paraná. His research includes remote sensing applied to agriculture, mainly in the subjects: mapping and estimation of area with agricultural crops, and determination of dates of the phenological cycle of agricultural crops.

Alex Paludo is a MS student in the State University of West Paraná. His research includes remote sensing applied to agriculture. Mainly for mapping and area estimation with agricultural crops such as soybean, maize, and wheat. He works with phenological cycle estimates of agricultural crops using remote sensing.

Laíza Cavalcante de Albuquerque Silva received her BA degree in agricultural engineering in 2017 and currently, she is a MS degree student in agricultural engineer at State University of West Paraná. Her research interests includes soybean, maize and wheat mapping using remote sensing, agricultural crops yield estimations, and phenological cycle recognition and identification using remote sensing.