








Temporal variability in soybean sowing and harvesting according to K-means and silhouette scores

Bruno Rodrigues de **Oliveira**^{1*} , Francisco Charles dos Santos **Silva**² , Ricardo **Mezzomo**² , Leandra Matos **Barrozo**² , Tatiane Scilewski da Costa **Zanatta**² , Joel Cabral dos **Santos**² , Carlos Henrique Conceição **Sousa**², Yago Pinto **Coelho**², Aurilucia do Nascimento Silva **Caldas**², Alan Mario **Zuffo**^{2†} 

¹ Editora Pantanal, Nova Xavantina, MT, Brazil.

² State University of Maranhão, Balsas, MA, Brazil.

*Correspondence: bruno@editorapantanal.com.br; alan_zuffo@hotmail.com

Received: 2024-07-08

Accepted: 2024-07-31

Published: 2024-08-01

Main Editors

Jorge Gonzales Aguilera



Copyright: © 2023. Creative Commons Attribution license: [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/).

For citation: Oliveira, B. R.; Silva, F. C.; Mezzomo, R.; Barrozo, L. M.; Zanatta, T. S.; Santos, J. C.; Sousa, C. H.; Coelho, Y. P.; Caldas, A. d.; Zuffo, A. M. (2024). Temporal variability in soybean sowing and harvesting according to K-means and silhouette scores. Trends in Agricultural and Environmental Sciences, (e240010), DOI: 10.46420/TAES.e240010



Abstract: Understanding the impact of seasonal variations on soybean productivity is crucial for optimizing agricultural practices. Given the influence of climatic factors such as temperature and rainfall on crop phenology, this study aims to analyze the effects of sowing and harvesting soybean cultivars in different seasons. This research investigates whether sowing soybeans in November and December leads to varying outcomes in terms of productivity and morphological characteristics, focusing on identifying the most stable cultivars across different climatic conditions. The study employed a comprehensive methodology, including data standardization, statistical tests (Levene, Shapiro-Wilk, ANOVA, Wilcoxon), and K-means clustering, to analyze 40 soybean cultivars across two seasons. Statistical preprocessing ensured data accuracy, while clustering helped identify cultivars with consistent responses to climatic changes. All computational analyses were performed using Python in the Google Colab environment. The findings revealed no significant difference in productivity between the two seasons, despite variations in temperature and rainfall. However, the moisture content of the grains (MTG) showed significant differences, influenced by higher rainfall in March and April and increased temperatures in December. K-means clustering highlighted SYN2282IPRO as the most stable cultivar and 77HO111I2X-GUAPORÉ as the most sensitive to climatic changes. The results emphasize the need for careful cultivar selection based on specific adaptability to seasonal variations. This study underscores the potential of computational tools like K-means clustering in agricultural optimization, offering a data-driven approach to selecting stable soybean cultivars. The adaptable methodology can be tailored to different geographical regions, soil types, and climate conditions, enhancing its relevance and applicability. These insights contribute to a better understanding of the complex interactions between climatic variables and soybean phenology, providing a foundation for improving agricultural practices in the face of climate change.

Keywords: season; machine learning; cultivar selection; *Glycine max* (L.) Merrill

1. Introduction

Soybeans (*Glycine max* (L.) Merrill), the primary global supplier of vegetable protein, are extensively grown in diverse regions, such as the Canadian prairies, the northern Great Plains of the United States, and the tropical areas of the Brazilian Cerrados and the Argentine Pampas (Grassini et al., 2021). This diverse geographic distribution highlights the remarkable adaptability of soybeans to different climatic and soil-climatic conditions, emphasizing their economic and

agronomic relevance in various regions of the world. Moreover, this broad adaptability, coupled with the nutritional importance of soy, highlights its prominent position in global food security and sustainable supply of plant-based proteins (Nóia Júnior & Sentelhas, 2019). Thus, improved soybean cultivation is multifaceted and encompasses aspects of food security, nutrition, economics, sustainability, and environmental impact.

A significant increase in global demand for crops is expected, estimated to occur between 60% and 110% by 2050, with climate change already impacting yields in several countries (Mourtzinis et al., 2019; Nóia Júnior & Sentelhas, 2019). Climatic variables, notably temperature and precipitation, have a determining influence on soybean productivity (de Oliveira et al., 2022). The influence of climate variables on crop yield is complex, with precipitation being crucial for determining the water balance, while temperature affects the development rate, size, and number of grains (L Hoffman et al., 2020). Therefore, identifying limiting climatic conditions and developing agricultural adaptation strategies are essential for mitigating food security concerns. Moreover, the development of agricultural agronomy aimed at adapting to climate change is highly important, especially when focusing on legumes, which are intrinsically susceptible to fluctuations in yield stability (Staniak et al., 2023). To this end, an in-depth understanding of these meteorological factors and their correlation with the physiological responses of plants is crucial for optimizing the productive efficiency of these important crops, thus contributing to sustainability and food security despite ongoing climate change (Gawęda et al., 2020). Overall, understanding these interactions is vital for developing integrated approaches in agricultural research, ensuring the resilience of agricultural production systems in the face of imminent climate change (L Hoffman et al., 2020; Mourtzinis et al., 2019).

In crops, sowing outside the optimal period can cause major losses. In the corn belt region in the United States alone, losses of approximately US\$340 million per year are estimated due to planting outside the recommended period, and climate change exacerbates this situation due to greater climate unpredictability, which affects decision making (Luiz Piaty et al., 2023). The sowing date plays a crucial role in the interaction with the response of soybean seed yield to the seeding rate, considering the findings of previous studies carried out from October 5th to December 15th in Brazil (Corassa et al., 2018). Three weeks of delay in planting soybeans is enough to change yields, according to research conducted in Japan, due to temperature changes (Kumagai & Takahashi, 2020). In another study in the United States, the authors found that the date of soybean sowing affected the oil concentration and seed yield regardless of latitude (Assefa et al., 2019). In a study carried out in Argentina, the authors concluded that among the management variables, sowing date and soybean genotype selection are the most important and help to explain approximately 40% of the total productivity variability (Vitantonio-Mazzini et al., 2021). In a study that analyzed the biometric and phenological characteristics of four contrasting soybean cultivars, the authors found that there was a reduction in growth in all cultivars with delayed sowing (Clovis et al., 2015).

Acknowledging the significance of sowing and harvesting seasons, coupled with climatic variations attributed to climate change and phenomena such as El Niño and La Niña, which impact the rainy season, this research aimed to examine the impact of sowing in distinct subsequent months (November and December) on 40 soybean cultivars. These months were chosen due to the observed variations in temperature and rainfall. Morphological and productivity data were analyzed using an unsupervised machine learning approach. Thus, the main objective of this study was to evaluate the temporal variability in soybean sowing and harvesting using K-means clustering and silhouette scores. The secondary objective was to verify which of the forty cultivars analyzed were more stable (whose variables responded in a more similar way) when sown/harvested in different months.

The K-Means machine learning method, which is based on cluster analysis, represents an extremely important tool for identifying and characterizing patterns in data and plays a crucial role in the context of agriculture (de Oliveira et al., 2021; Li & Niu, 2020; Rahamathunnisa et

al., 2020; Shedthi et al., 2017). The intrinsic ability of K-means to partition datasets into distinct clusters, based on similarities between observations, enables the discernment of latent patterns and the efficient categorization of agronomic information. By applying the K-means algorithm to agricultural datasets, it is possible to identify homogeneous groups of variables, such as soil characteristics, climate conditions and agricultural practices, providing an in-depth understanding of the interactive relationships between these factors (Bekkanti et al., 2020; Sharma et al., 2023; Yadav et al., 2020). This analytical approach offers valuable input for agricultural decision-making, promoting more effective and sustainable management, optimizing the use of resources and maximizing crop yields (de Oliveira et al., 2021; Li & Niu, 2020; Rahamathunnisa et al., 2020). Therefore, the K-Means method proves to be an essential tool in the analysis of agricultural data, contributing significantly to the promotion of efficiency and productivity in the agricultural sector.

2. Material and Methods

2.1 Experimental area, design and treatments

The experiment was conducted in the field at the “Pequizeiro” farm, located at the Experimental Station of “Accert Pesquisa e Consultoria Agronomica” near Balsas, MA, Brazil, during the 2022/2023 harvest to investigate soybean cultivar performance under the region's hot and humid tropical (Aw) climate according to Köppen’s classification. The study area, characterized by a latitude of 07°31’57” S, longitude of 46°02’08” W, and an altitude of approximately 283 m, experiences rainy summers and dry winters, with an average annual rainfall of 1175 mm. The soil, classified as a Yellow Oxisol with a sandy texture, was sampled and analyzed before the experiment, revealing its chemical and physical properties. Figure 1 shows precipitation and temperature data between November and April for the years in which the experiment was carried out (2022 to 2023) and historical data between 1991 and 2020. Table 1 shows the physicochemical characteristics of the soil in the experimental area where the 40 cultivars were sown.

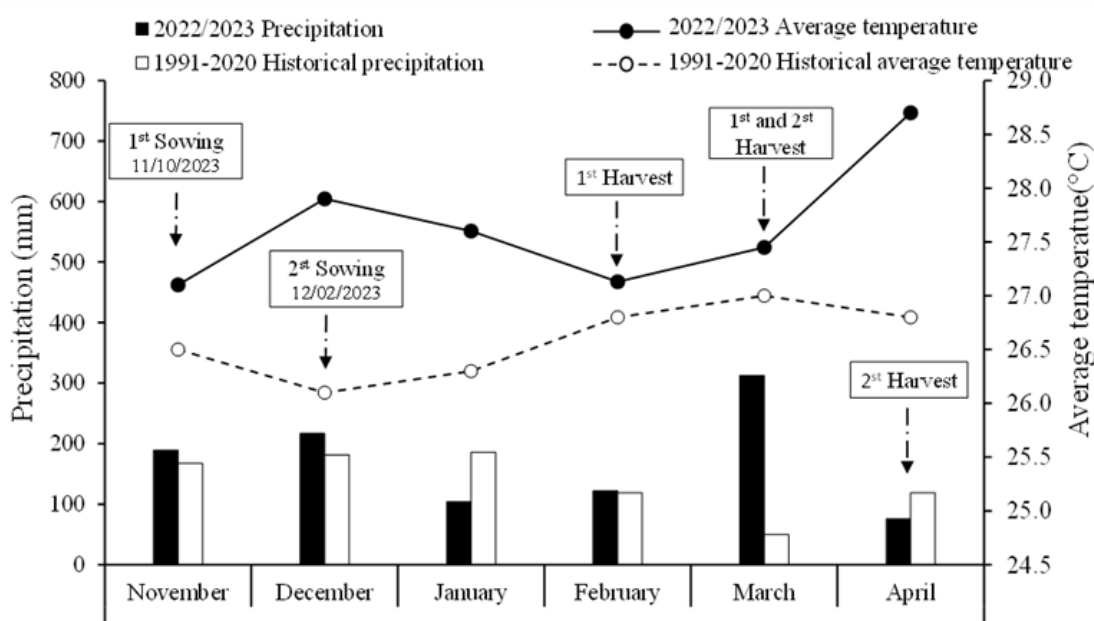


Figure 1. Average precipitation and temperature data for the years 2022 and 2023 and the historical data for 1991 to 2020 for the city of Balsas-MA. Source: Accert (2023) and the National Institute of Meteorology (2023).

Table 1. Main chemical properties of the soils used in the experiment.

Depth cm	pH	OM	P _{Mehlich-1}	H+Al	Al ³⁺	Ca ²⁺	Mg ²⁺	K ⁺	CEC	B
	H ₂ O	dag kg ⁻¹	mg dm ⁻³			cmol _c dm ⁻³				%
0-20	6.00	1.29	54.95	1.20	0.01	2.15	0.71	136.00	4.41	72.78
20-40	4.65	0.23	20.72	1.80	0.54	0.95	0.30	70.00	3.23	44.26
	B	Cu	Fe	Mn	Zn	S	TOC	Clay	Silt	Sand
	mg dm ⁻³						dag kg ⁻¹	%		
0-20	0.22	0.44	113.21	14.28	0.73	6.30	0.75	24.24	9.26	66.49
20-40	0.23	0.40	81.98	4.25	0.37	12.60	0.13			

OM: organic matter. CEC: cation exchange capacity at pH 7.0. B: base saturation. TOC: total organic carbon.

The experimental design followed a randomized block arrangement in a split-plot scheme with four replications. The plot treatments involved two sowing times (season 1: 10/11/2023, season 2: 04/12/2023), while the subplots included 40 soybean cultivars: FTR 3190 IPRO, FTR 4288 IPRO, NK 8770 IPRO, M 8606I2X, M 8644 IPRO, ADAPTA LTT 8402 IPRO, 98R30 CE, FORTALEZA IPRO, MONSOY 8330I2X, SUZY IPRO, TMG 22X83I2X, EXPANDE LTT 8301 IPRO, FORTALECE L090183 RR, 83IX84RSF I2X, 82HO111 IPRO - HO COXIM IPRO, 82I78RSF IPRO, SYN2282IPRO, ATAQUE I2X, NK 8100 IPRO, FTR 4280 IPRO, LYNDIA IPRO, BRASMAX OLÍMPO IPRO, LAT 1330BT.11, FTR 3179 IPRO, 97Y97 IPRO, BRASMAX BÓNUS IPRO, PAULA IPRO, NEO 790 IPRO, LTT 7901 IPRO, GNS7900IPRO – AMPLA, 79I81RSF IPRO, ELISA IPRO, NK 7777 IPRO, 77HO111I2X – GUAPORÉ, GNS7700IPRO, FTR 3868 IPRO, MANU IPRO, NEO 760 CE, 74K75RSF CE, 96R29 IPRO. Each experimental unit consisted of eight rows spaced 0.50 m apart and 10 m long and covered an area of 40 m², with the central 16 m² considered the useful area. Desiccation was carried out using glyphosate + Haloxifoprop-C-methyl, and subsequent soybean sowing was performed via a no-tillage system. The fertilization agents included monoammonium phosphate (MAP) and potassium chloride. The soybean seeds were treated, and throughout plant development, various products were used for weed, pest, and disease management.

At the R8 harvest stage, variables such as plant height (PH), insertion of the first pod (IFP), number of stems (NS), number of legumes per plant (NLP), number of grains per plant (NG), number of grains per pod (NGP), mass of a thousand grains (MTG), and grain yield (GY) were measured from 10 plants per plot. The data, organized into a tabular format with columns representing different variables and conditions, comprised a total of 320 samples, considering 4 replications for each of the 40 cultivars across two harvest seasons. Details about the experiment and the data collected were previously described (de Oliveira et al., 2023).

2.2 K-Means and Silhouette Scores

Machine learning algorithms for performing clustering employ unsupervised learning. In this type of learning, the classes (groups) to which the samples belong are not known. These algorithms group data points that are similar to each other in the same cluster in such a way that the remaining samples grouped in other clusters are less similar (James et al., 2013). A frequently used dissimilarity metric is the Euclidean distance (Theodoridis & Koutroumbas, 2006). The smaller the distance between two vectors (data sample) is, the more similar they are (de Oliveira et al., 2022).

The K-Means algorithm clusters data by attempting to separate samples in K groups of equal variances, minimizing a criterion known as within-cluster sum-of-squares (WCSS). This algorithm requires the number of clusters to be specified (Ahmed et al., 2020). It scales well to large numbers of samples and has been used across a large range of application areas in many different fields (Ikotun et al., 2023). This algorithm divides a set of N samples X into K disjoint clusters, each described by the mean U_j of the samples in the cluster (Oti et al., 2021). The

means are commonly called the cluster “centroids”; note that they are not, in general, points from X , although they live in the same space. K-Means has three main steps. In the first step, the initial centroids are chosen randomly or via an adequate scheme (Buitinck et al., 2013). After initialization, each sample is associated with the nearest centroid using a distance metric such as the Euclidean distance. Finally, the algorithm creates new centroids by taking the mean value of all of the samples assigned to each previous centroid. The difference between the old and the new centroids is computed, and the algorithm repeats these last two steps until this value is less than a certain threshold. In other words, the process repeats until the centroids do not move significantly (Ikotun et al., 2023).

However, while K-Means has been applied to a diverse array of real-world problems (Ahmed et al., 2020), it has advantages and disadvantages (Buitinck et al., 2013; Oti et al., 2021). On the positive side, it is relatively straightforward to implement, scale effectively to large datasets, ensure convergence, and readily adapt to new examples. Conversely, its drawbacks include the impact of the choice of K on clustering results, challenges in handling data of varying sizes and density, displacement of centroids by outliers hindering correct grouping, and susceptibility to issues in high-dimensional data. To address some of these limitations, normalization and standardization have proven useful as preprocessing techniques before applying K-means. Normalization involves scaling data to a specific range, while standardization transforms data to have a mean of zero and a standard deviation of one (Abdulhafedh, 2021; James et al., 2013).

Choosing the value of K is a critical stage in the application of K-means. Several approaches can be used in this decision (Umargono et al., 2020); however, one of the most effective approaches is the silhouette score (Naghizadeh & Metaxas, 2020). This approach serves as a crucial evaluation metric, offering a quantitative measure to assess the quality and appropriateness of clustering results. It evaluates the well-defined nature and distinctiveness of clusters by quantifying how effectively data points fit into their assigned clusters and how distinct they are from other clusters. The score ranges from -1 to +1, with negative values indicating potential misassignments, values close to 0 suggesting ambiguous clustering, and positive values reflecting well-clustered and distinct data points (Tambunan et al., 2020).

2.3 Proposed methodology

First, standardization preprocessing was applied to the data. This technique centralizes the samples for each variable by subtracting the mean and dividing by the standard deviation. Afterwards, statistical tests (Levene and Shapiro–Wilk) and analyses (ANOVA and Wilcoxon ranks) were carried out to verify whether the samples came from different or equal distributions with the same average. This process is performed for each variable individually. For variables that do not pass statistical tests, the Yeo–Johnson transformation is applied so that the distribution is approximated to a normal distribution. From the p values provided by Wilcoxon ranks, the variables that will be used in K-means are chosen. The choice of the number of clusters is carried out empirically by testing the values that return the highest silhouette score. After setting this quantity, K-means was ultimately applied to the data for each season individually to determine how the cultivars were grouped in the same group in different seasons. The entire computational implementation is carried out in Python in the Google Colab environment and is accessible to the public.

3. Results

Summary information calculated from the original data (before any transformation) is recorded in Table 2. Measurements were calculated for each season separately. Although the original values are not used in the proposed machine learning methodology, visualization of some

measurements of these data is necessary. The transformations employed eliminate the dimensions of the variables, complicating the discussion concerning these transformed values.

Table 2. Summary measures for each variable calculated from the raw data.

Measures	PH (cm)	IFP (cm)	NLP (unit)	NG (unit)	NGP (unit)	NS (unit)	MTG (g)	GY (kg/ha ⁻¹)
Season 1								
Average	65.9850	16.4850	57.9650	135.2350	2.2995	4.8166	162.2884	3428.4524
Std.	8.3128	2.3366	19.3414	48.7502	0.3704	1.3962	18.4309	639.0364
Minimum	50.4000	10.8000	20.2000	47.8000	0.9388	2.2000	127.0572	1538.2298
Maximum	91.0000	26.4000	116.4000	272.4000	4.7533	9.0000	215.9964	4930.0000
Season 2								
Average	70.7883	14.4450	60.2116	134.9366	2.2817	3.3266	174.3557	3408.6551
Std.	8.9604	3.2856	20.7694	70.4670	1.1304	1.1379	18.9684	314.7245
Minimum	47.6000	7.2000	24.8000	48.0000	1.1494	0.4000	127.7627	2625.9137
Maximum	94.8000	24.2000	123.0000	683.4000	14.8565	9.0000	213.2448	4164.0574

Std.: standard deviation; cm: centimeter; g: gram; kg/ha: kilogram per hectare.

The statistics and p values of the Levene and Shapiro–Wilk tests were calculated for each variable (Table 3). In the “Variance?” and “Normal?” columns, the final results are presented, considering a confidence level of 5%. The results shown in Table 4 were obtained from analysis of variance (ANOVA) and included three sources of variation, i.e., season (SE), cultivar (CL), and season versus cultivar (SE × CL), as well as the coefficient of variation (CV) for each variable.

Table 3. Results of the Levene and Shapiro–Wilk tests for each variable after undergoing the Yeo–Johnson transformation.

Variable	Levene’s Test			Shapiro–Wilk test		
	Statistic	p value	Variance?	Statistic	p value	Normal?
PH	1.8990	0.1691	Equal	0.9904	0.0347	Yes
IFP	22.2157	0.0000	Not Equal	0.9830	0.0008	No
NLP	0.1851	0.6672	Equal	0.9904	0.0360	No
NG	0.2232	0.6369	Equal	0.8005	0.0806	Yes
NGP	9.2179	0.0025	Not Equal	0.8971	0.0000	No
NS	0.6210	0.4312	Equal	0.9951	0.4199	Yes
MTG	0.0317	0.8587	Equal	0.9879	0.0091	No
GY	58.1039	0.0000	Not Equal	0.9868	0.0052	No

Table 4. Results of the analysis of variance showing the p values for each variable in each source of variation after undergoing the Yeo–Johnson transformation.

Source of variation	Probability > F							
	PH (cm)	IFP (cm)	NLP (unit)	NG (unit)	NGP (unit)	NS (unit)	MTG (g)	GY (kg/ha ⁻¹)
SE	0.0014	0.0005	0.0737	0.8088	0.0517	0.0000	0.0002	0.4970
CL	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
SE x CL	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CV (%)	13.0787	19.5247	33.9099	44.7123	36.610	36.1557	11.6414	14.6909

SE: season; CL: cultivar; CV: coefficient of variation.

To investigate which cultivars underwent fewer changes at different sowing and harvest seasons, we initially checked which of the variables collected were associated with different statistical

distributions. Since the variables did not pass the assumptions of the Levene and Shapiro–Wilk tests, even after undergoing the Yeo–Johnson transformation, we opted for the Wilcoxon test for independent samples. This nonparametric method was chosen due to its independence from fixed characteristics of the data distributions. Our results are presented as the statistics of these tests as well as the p values (Table 5). These results are used to select the variables that will be used in K-means.

Table 5. Wilcoxon rank results showing the p values for each variable.

	PH (cm)	IFP (cm)	NLP (unit)	NG (unit)	NGP (unit)	NS (unit)	MTG (g)	GY (kg/ha ⁻¹)
Statistic	-4.6457	5.8819	-0.7516	1.0193	3.0995	9.1380	-5.6227	-0.0622
p value	0.0000	0.0000	0.4522	0.3080	0.0019	0.0000	0.0000	0.9503

Boxplots of each standardized variable and of the outliers for both the sowing and harvesting seasons were generated (Figure 2). It should be noted that the graphs were not produced with the original values of the variables due to the different scales (units of measurement) used for each of the variables. Therefore, the values themselves do not matter but only the distribution of the data in the different seasons.

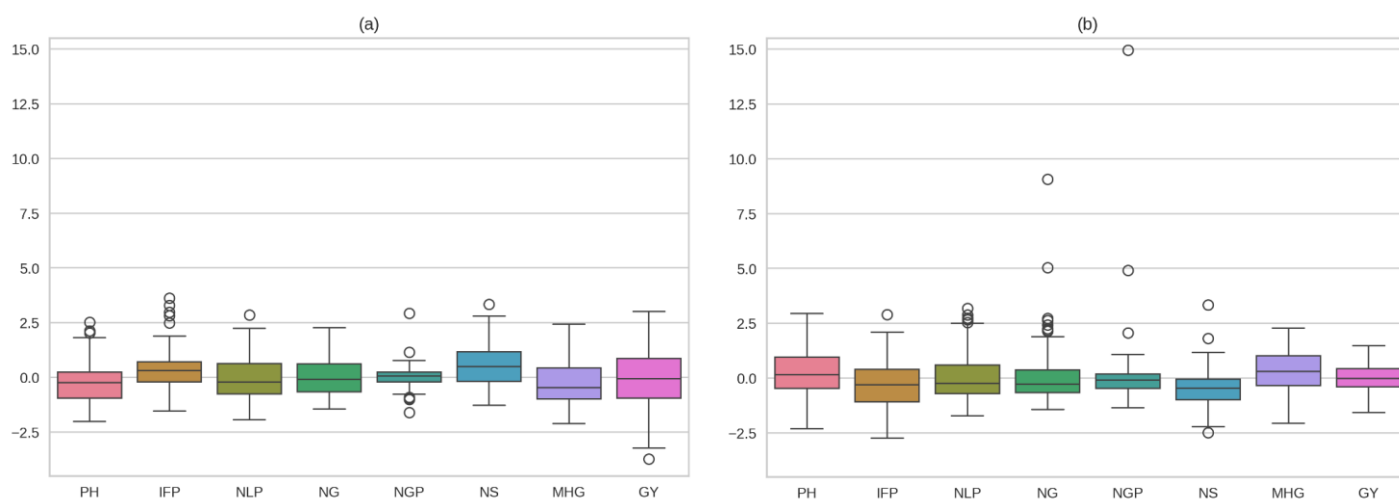


Figure 2. Boxplot of each standardized variable for the different seasons: (a) season 1 and (b) season 2. PH: plant height, IFP: insertion of the first pod, NS: number of stems, NLP: number of legumes per plant, NG: number of grains per plant, NGP: number of grains per pod, MTG: mass of a thousand grains, and GY: grain yield.

The results of choosing the values for the number of centroids in K-means using the elbow method and the silhouette score are shown in Figures 3 and 4. These analyses were conducted independently for each individual season. This step involves tuning the hyperparameter of the machine learning model and must be performed empirically on the entire dataset. Figure 3 shows the silhouette scores for different choices of k, highlighting (dashed black line) the value of k for the highest score. Figure 4 shows how the samples are grouped into different groups as well as their silhouette score values according to fixed k values (Figure 3). In addition, the average silhouette score (dashed red line) is also displayed. Samples with negative scores are those with a greater possibility of having been grouped incorrectly.

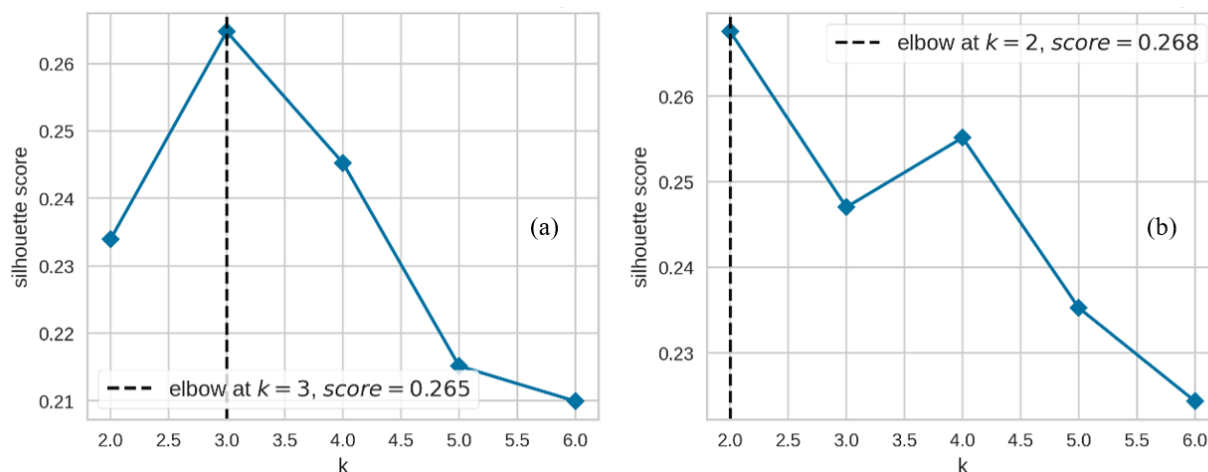


Figure 3. Silhouette score for six chosen from the number of centroids: (a) for season 1 (b) for season 2.

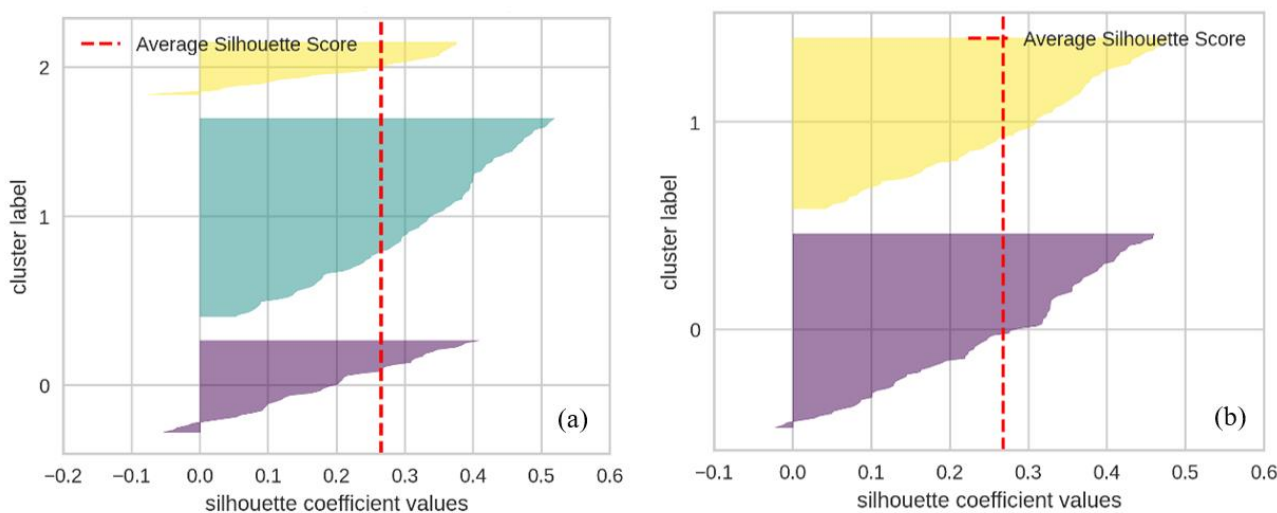


Figure 4. Silhouette plot for 160 samples in the (a) tree cluster and (b) two other clusters. The colors designate the groupings.

Since the number of clusters selected by the elbow method differs for each season, numbers of clusters equal to 2 and 3 were considered. For each choice of cluster number, the number of samples grouped into different groups was computed, considering the groups formed in the first and second seasons separately. Figure 5 shows the results where the colored numbers indicate the number of replications of each cultivar that were grouped into different groups. For example, 3 repetitions of the cultivar 74K75RSF CE were grouped into distinct groups using two clusters in K-Means when comparing the clusters carried out in the first and second seasons. In other words, in the first season (November), these three repetitions were grouped into group 0, and in the second season (December), they were grouped into group 1. However, when 3 clusters were used, only one repetition was grouped together.

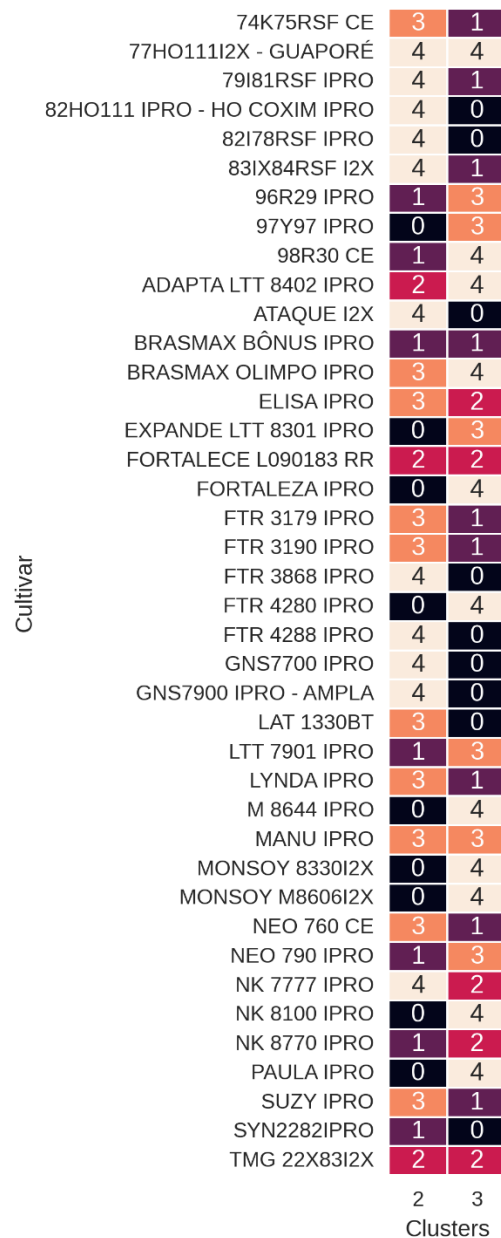


Figure 5. The number of repetitions (samples) grouped into different groups for each cultivar when the number of clusters was 2 or 3 is shown. The colors are used only to differentiate the quantities.

4. Discussion

The main objective of this research was to verify whether sowing and harvesting soybeans in consecutive seasons (November and December) yield varying outcomes. This investigation is valid because the chosen months present different rainfall and temperature characteristics in the growing region. Furthermore, the seeding rate should be adjusted at each sowing date to increase soybean productivity (Umburanas et al., 2019). The phenology of crops is influenced by both climatic factors and agronomic management practices, including sowing date and cultivar characteristics (Gawęda et al., 2020; He et al., 2020). December showed an increase in temperature relative to the average temperature of the historical series recorded since 1991 (Figure 1). Furthermore, this temperature is also higher than that recorded in the previous month. Additionally, the rainfall in March, when part of the 1st and 2nd harvests were carried out, was significantly greater than that in the historical series.

The secondary objective of this research was to determine which of the 40 cultivars analyzed exhibited the greatest variation in the measured variables. Understanding how cultivars react and respond to factors is of great cognitive and practical importance, as there are numerous studies and programs on the selection of genotypes with greater tolerance to abiotic stresses (Soares et al., 2015; Staniak et al., 2023; Zuffo et al., 2018).

Our statistical test results showed that not all the collected variables passed the normality and homogeneity of variance tests, even after applying the standardization and Yeo–Johnson transformations (Table 3). Because the analyzed data sample includes more than 30 examples, the normal distribution requirement can be relaxed for the ANOVA test because of the central limit theorem. However, the variance is not homogeneous for three of the variables, namely, IFP, NGP, and GY. Therefore, ANOVA may lead to a Type I error, where a true null hypothesis is mistakenly rejected. In other words, the observed significant differences between means could be ascribed to variations in variances rather than to differences in population means (Roberts & Russo, 2014). Keeping this in mind, the p values indicate that for the variables PH, IFP, NS, and MTG, there was a significant difference between the seasons (Table 4). In relation to cultivar (CL) and the interaction (SE × CL), there were significant differences in all the variables. On the other hand, Wilcoxon rank analysis (Table 5), which does not require specific assumptions about the data distribution, revealed that there was a significant difference in the variables PH, IFP, NGP, NS and MTG. This second analysis included only one variable in relation to the ANOVA, i.e., NGP. The variables that were significantly different between the seasons were not related to productivity, except for the NGP and MTG variables. The reasons for these exceptions are explained below.

The main variable related to productivity, e.g., GY, had a very similar average in both seasons (Table 1). However, the larger standard deviation in season 1 indicates greater variability in this period. This was also observed when comparing the minimum and maximum values in the two seasons. On the other hand, the results from other research suggest that late sowing reduces productivity because of reductions in aerial biomass per area, leaf area index, final plant height, pod height, pods per area, seeds per area, and seed mass (Umburanas et al., 2019). However, this research was conducted under different soil and climatic conditions, and only one cultivar was analyzed. Observations of the MTG variable revealed that, in the second season, its average was greater than that in the second season, contrary to the results of the abovementioned research (Umburanas et al., 2019). In addition, the variables IFP, NG, NGP, NS, and GY exhibited average decreases in the late season, which partially corroborates the previous results. Therefore, the results obtained here can be partially explained by the overlap in the harvest in March 2023 (Figure 1), when there was a partial harvest from both the November and December sowing seasons. During this period, there was a smaller change in temperature and rainfall in July than in April.

Although the NGP variable was selected using Wilcoxon rank analysis, it had outliers because the standardized values were much greater than those of the other variables for the second season (Figure 2). Therefore, concluding that the average is significantly different between seasons may be misleading. Therefore, additional investigations are necessary to verify the source of these outliers. Because K-Means is sensitive to outliers, only the variables PH, IFP, NS and MTG were used to learn the centroids that separate the groups in each season.

Based on the statistical results, there is no evidence to conclude that one season is better than another in relation to productivity. In contrast, the tests indicated that there was no difference in productivity, which can be explained in part by the overlap in the harvest period (the means and medians are shown in Table 1 and Figure 2). The MTG variable, selected by both statistical analyses, is related to productivity. However, the high moisture content of the grains during the second sowing at the time of harvest (Figure 1, coinciding with increased precipitation in April) led to a variation in the mean moisture content between the seasons. The other variables associated with productivity are shown as medians in Figure 2. This explains why only this

variable related to productivity was selected. Because the mean MTG differed between seasons 1 and 2 but with an approximate standard deviation, this variable was maintained for the application of K-means. Therefore, for the morphological variables and the MTG variable, there was a significant difference between the seasons (Table 5). In other research, the authors observed that the mass of a thousand grains increases linearly as the moisture content increases (Tavakoli et al., 2009). Additionally, as the analysis of variance showed, there was a statistically significant difference between the soybean cultivars. Therefore, it is necessary to investigate which cultivars have different results because of variations in temperature and precipitation in different seasons.

The elbow method is applied to fix the number of centroids returning different values for the sowing season, with k equal to 2 for November (Figure 3 (a)) and k equal to 3 for December (Figure 3 (b)). The use of k equal to 2 (Figure 4 (a)) results in fewer samples with a negative silhouette score, according to the graphs in Figure 4. There are more samples with a score higher than the average silhouette score for this value of k . However, none of the k values resulted in a very high score, and in both cases, it was less than 0.3. This indicates that the patterns observed in the variables, that is, how cultivars respond to variations in precipitation and temperature in different seasons, are not as distinct. Therefore, the data samples have values closer to the centroids of the group to which they were associated, but they are also not very far from the centroids of the other groups.

Although the cultivars' responses are generally similar in relation to the distance from the centroids, some are more stable in terms of grouping. In other words, they are grouped into the same groups regardless of the sowing season. When 2 clusters were used, nine cultivars had four repetitions grouped in the same groups (Figure 5). These are indicated by 0 values in the column. For the three clusters, nine other cultivars also showed the same result. The best cultivar for both cluster choices was SYN2282IPRO. For this cultivar, only one of the repetitions was grouped into a different group when comparing the groupings of the two seasons. On the other hand, the cultivar with the greatest sensitivity to changes in temperature and precipitation in different seasons was 77HO111I2X-GUAPORÉ. These results are expected, as soybean genotypes respond differently to different environmental variations and to sowing date (Clovis et al., 2015).

The use of K-means for some cultivars was more sensitive to the number of clusters. Depending on the choice of k , the repetitions were grouped into completely different groups (number 4 in the column of Figure 5) or into the same groups (number 0 in the column of Figure 5). The cultivars that fit these results are 82HO111 IPRO-HOCOXIM IPRO, 82I78RSF IPRO, FTR 3868 IPRO, FTR 4280 IPRO, FTR 4288 IPRO, GNS7700 IPRO, GNS7900 IPRO-AMPLA, M8644 IPRO, MONSOY 8330I2X, MONSOY M8606I2X, NK 8100 IPRO, and PAULA IPRO.

The results presented here show that the proposed methodology provides a scheme for farmers to select the most stable cultivars depending on the season. For planting in November, 3 clusters were chosen (Figure 4a), for which the values in Figure 5, column 2, were equal to 0. These cultivars generated similar results. On the other hand, for sowing in December, the choice of 2 clusters was more appropriate (Figure 4b). Therefore, cultivars with values equal to 0 in the first column of Figure 5 must be selected. The selection process is different because the cultivars responded differently to variations in temperature and precipitation during these seasons. In such scenarios, the use of computational tools is necessary. Such communication and information technology tools are extremely important for agriculture because they contribute to long-term sustainability (Lindblom et al., 2017).

It is crucial to emphasize that the presented methodology must be customized according to the specific cultivar, soil, and climate characteristics. Consequently, for crops in different geographical regions, the selection of cultivars may deviate from that outlined in this study. In

essence, the machine learning model derived herein should not be universally applied to all scenarios. This particular characteristic of the model is not a constraint; rather, it underscores the adaptability of the methodology. This adaptability extends beyond the variables employed in this study, allowing the model to analyze additional factors. Furthermore, the methodology is not restricted to assessing the stability of soybean crops exclusively; it can also be extended to evaluate the stability of other crops. This versatility is conceivable because the methods employed in the methodology do not consider intrinsic information pertaining to crops, soil, or climate. Instead, they rely solely on the collected data, irrespective of the measured variables.

5. Conclusion

A comparative analysis of soybean cultivar performance across consecutive planting seasons was conducted utilizing a machine learning framework. Forty soybean cultivars were sown in November and December, with subsequent data collection and analysis. The primary objective was to quantify the influence of seasonal temperature and precipitation variability on soybean morphology and yield.

Statistical analysis revealed no significant differences in grain yield (GY) between the two planting periods. Conversely, grain moisture content (MTG) exhibited substantial inter-seasonal variation. Increased precipitation during March and April, coupled with elevated December temperatures, was correlated with these MTG discrepancies.

Employing K-means clustering, cultivars were categorized based on their phenotypic stability across seasons. SYN2282IPRO demonstrated consistent performance, with minimal clustering variation. In contrast, 77HO111I2X-GUAPORÉ exhibited pronounced sensitivity to climatic fluctuations. These findings underscore the importance of cultivar selection aligned with specific environmental conditions.

While morphological traits displayed seasonal variation, yield metrics remained relatively constant. However, the observed MTG discrepancies necessitate refined post-harvest management strategies. The study highlights the utility of machine learning, specifically K-means clustering, in cultivar selection for optimal agronomic performance. This methodology is adaptable to diverse agro-ecological contexts.

This research provides foundational insights into the soybean-climate interaction. Nevertheless, the model's generalizability is contingent upon local environmental factors. Subsequent investigations should focus on model refinement for various crops and regions. The findings contribute to the broader understanding of soybean phenology under changing climatic conditions and inform data-driven agricultural decision-making.

6. References

- Abdulhafedh, A. (2021). Incorporating k-means, hierarchical clustering and pca in customer segmentation. *Journal of City and Development*, 3(1), 12–30.
- Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics*, 9(8). <https://doi.org/10.3390/electronics9081295>
- Assefa, Y., Purcell, L. C., Salmeron, M., Naeve, S., Casteel, S. N., Kovács, P., Archontoulis, S., Licht, M., Below, F., Kandel, H., Lindsey, L. E., Gaska, J., Conley, S., Shapiro, C., Orłowski, J. M., Golden, B. R., Kaur, G., Singh, M., Thelen, K., ... Ciampitti, I. A. (2019). Assessing Variation in US Soybean Seed Composition (Protein and Oil). *Frontiers in Plant Science*, 10. <https://doi.org/10.3389/fpls.2019.00298>
- Bekkanti, A., Gunde, V. P., Itnal, S., Parasa, G., & Basha, C. Z. (2020). Computer based classification of diseased fruit using K-means and support vector machine. *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*, 1227–1232. <https://doi.org/10.1109/ICSSIT48917.2020.9214177>

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Grobler, A. G. and J., Layton, R., VanderPlas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108–122.

Clovis, P. J., Jackson, K., Marcelo, B., Murilo, V. D. C., & Leandro, M. (2015). Phenological and quantitative plant development changes in soybean cultivars caused by sowing date and their relation to yield. *African Journal of Agricultural Research*, 10(6), 515–523. <https://doi.org/10.5897/AJAR2014.9325>

Corassa, G. M., Amado, T. J. C., Strieder, M. L., Schwalbert, R., Pires, J. L. F., Carter, P. R., & Ciampitti, I. A. (2018). Optimum Soybean Seeding Rates by Yield Environment in Southern Brazil. *Agronomy Journal*, 110(6), 2430–2438. <https://doi.org/10.2134/agronj2018.04.0239>

de Oliveira, B. R., da Silva, A. A. P., Teodoro, L. P. R., de Azevedo, G. B., Azevedo, G. T. de O. S., Baio, F. H. R., Sobrinho, R. L., da Silva Junior, C. A., & Teodoro, P. E. (2021). Eucalyptus growth recognition using machine learning methods and spectral variables. *Forest Ecology and Management*, 497, 119496. <https://doi.org/10.1016/j.foreco.2021.119496>

de Oliveira, B. R., Zuffo, A. M., Aguilera, J. G., Steiner, F., Ancca, S. M., Flores, L. A. P., & Gonzales, H. H. S. (2022). Selection of Soybean Genotypes under Drought and Saline Stress Conditions Using Manhattan Distance and TOPSIS. *Plants* 2022, Vol. 11, Page 2827, 11(21), 2827. <https://doi.org/10.3390/PLANTS11212827>

de Oliveira, B. R., Zuffo, A. M., dos Santos Silva, F. C., Mezzomo, R., Barrozo, L. M., da Costa Zanatta, T. S., dos Santos, J. C., Sousa, C. H. C., & Coelho, Y. P. (2023). Dataset: Forty soybean cultivars from subsequent harvests. *Trends in Agricultural and Environmental Sciences*, e230005–e230005.

Gawęda, D., Nowak, A., Haliniarz, M., & Woźniak, A. (2020). Yield and Economic Effectiveness of Soybean Grown Under Different Cropping Systems. *International Journal of Plant Production*, 14(3), 475–485. <https://doi.org/10.1007/s42106-020-00098-1>

Grassini, P., Cafaro La Menza, N., Rattalino Edreira, J. I., Monzón, J. P., Tenorio, F. A., & Specht, J. E. (2021). Soybean. In *Crop Physiology Case Histories for Major Crops* (pp. 282–319). Elsevier. <https://doi.org/10.1016/B978-0-12-819194-1.00008-6>

He, L., Jin, N., & Yu, Q. (2020). Impacts of climate change and crop management practices on soybean phenology changes in China. *Science of The Total Environment*, 707, 135638. <https://doi.org/10.1016/j.scitotenv.2019.135638>

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622, 178–210. <https://doi.org/10.1016/j.ins.2022.11.139>

James, G., Witten, D., Hastie, T., Tibshirani, R., & others. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Kumagai, E., & Takahashi, T. (2020). Soybean (*Glycine max* (L.) Merr.) Yield Reduction due to Late Sowing as a Function of Radiation Interception and Use in a Cool Region of Northern Japan. *Agronomy*, 10(1), 66. <https://doi.org/10.3390/agronomy10010066>

L Hoffman, A., R Kemanian, A., & E Forest, C. (2020). The response of maize, sorghum, and soybean yield to growing-phase climate revealed with machine learning. *Environmental Research Letters*, 15(9), 094013. <https://doi.org/10.1088/1748-9326/ab7b22>

Li, C., & Niu, B. (2020). Design of smart agriculture based on big data and Internet of things. *International Journal of Distributed Sensor Networks*, 16(5), 155014772091706. <https://doi.org/10.1177/1550147720917065>

Lindblom, J., Lundström, C., Ljung, M., & Jonsson, A. (2017). Promoting sustainable intensification in precision agriculture: review of decision support systems development and strategies. *Precision Agriculture*, 18(3), 309–331. <https://doi.org/10.1007/s11119-016-9491-4>

Luiz Piati, G., Ferreira de Lima, S., Lustosa Sobrinho, R., dos Santos, O. F., Vendruscolo, E. P., Jacinto de Oliveira, J., do Nascimento de Araújo, T. A., Mubarak Alwutayd, K., Finatto, T., & AbdElgawad, H. (2023). Biostimulants in Corn Cultivation as a Means to Alleviate the Impacts of Irregular Water Regimes Induced by Climate Change. *Plants*, 12(13), 2569. <https://doi.org/10.3390/plants12132569>

- Mourtzinis, S., Specht, J. E., & Conley, S. P. (2019). Defining Optimal Soybean Sowing Dates across the US. *Scientific Reports*, 9(1), 2800. <https://doi.org/10.1038/s41598-019-38971-3>
- Naghizadeh, A., & Metaxas, D. N. (2020). Condensed Silhouette: An Optimized Filtering Process for Cluster Selection in K-Means. *Procedia Computer Science*, 176, 205–214. <https://doi.org/10.1016/J.PROCS.2020.08.022>
- Nóia Júnior, R. de S., & Sentelhas, P. C. (2019). Soybean-maize succession in Brazil: Impacts of sowing dates on climate variability, yields and economic profitability. *European Journal of Agronomy*, 103, 140–151. <https://doi.org/10.1016/j.eja.2018.12.008>
- Oti, E. U., Olusola, M. O., Eze, F. C., & Enogwe, S. U. (2021). Comprehensive review of K-Means clustering algorithms. *Criterion*, 12, 22–23.
- Rahamathunnisa, U., Nallakaruppan, M. K., Anith, A., & Kumar K.S., S. (2020). Vegetable Disease Detection Using K-Means Clustering And Svm. 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 1308–1311. <https://doi.org/10.1109/ICACCS48705.2020.9074434>
- Roberts, M., & Russo, R. (2014). *A Student's Guide to Analysis of Variance*. Routledge. <https://doi.org/10.4324/9781315787954>
- Sharma, P. K. ; Ferrarezi, S., Kadyampakeni, D. M., Awal, R., Shukla, M. K., & Sharma, P. (2023). Fuzzy K-Means and Principal Component Analysis for Classifying Soil Properties for Efficient Farm Management and Maintaining Soil Health. *Sustainability* 2023, Vol. 15, 3144, 15(17), 13144. <https://doi.org/10.3390/SU151713144>
- Shedthi, B. S., Shetty, S., & Siddappa, M. (2017). Implementation and comparison of K-means and fuzzy C-means algorithms for agricultural data. 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), 105–108. <https://doi.org/10.1109/ICICCT.2017.7975168>
- Soares, I. O., Rezende, P. M., Bruzi, A. T., Zambiazzi, E. V., Zuffo, A. M., Silva, K. B., & Gwinner, R. (2015). Adaptability of soybean cultivars in different crop years. *Genetics and Molecular Research*, 14(3), 8995–9003. <https://doi.org/10.4238/2015.August.7.8>
- Staniak, M., Szpunar-Krok, E., & Kocira, A. (2023). Responses of Soybean to Selected Abiotic Stresses—Photoperiod, Temperature and Water. *Agriculture*, 13(1), 146. <https://doi.org/10.3390/agriculture13010146>
- Tambunan, H. B., Barus, D. H., Hartono, J., Alam, A. S., Nugraha, D. A., & Usman, H. H. H. (2020). Electrical peak load clustering analysis using K-means algorithm and silhouette coefficient. *Proceeding - 2nd International Conference on Technology and Policy in Electric Power and Energy, ICT-PEP 2020*, 258–262. <https://doi.org/10.1109/ICT-PEP50916.2020.9249773>
- Tavakoli, H., Rajabipour, A., & Mohtasebi, S. S. (2009). Moisture-Dependent Some Engineering Properties of Soybean Grains. *Agricultural Engineering International: The CIGR Ejournal*, XI(1110).
- Theodoridis, S., & Koutroumbas, K. (2006). *Pattern recognition*. Elsevier.
- Umargono, E., Suseno, J. E., & Gunawan, S. K. V. (2020). K-means clustering optimization using the elbow method and early centroid determination based on mean and median formula. *The 2nd International Seminar on Science and Technology (ISSTEC 2019)*, 121–129.
- Umburanas, R. C., Yokoyama, A. H., Balena, L., Dourado-Neto, D., Teixeira, W. F., Zito, R. K., Reichardt, K., & Kawakami, J. (2019). Soybean Yield in Different Sowing Dates and Seeding Rates in a Subtropical Environment. *International Journal of Plant Production*, 13(2), 117–128. <https://doi.org/10.1007/s42106-019-00040-0>
- Vitantonio-Mazzini, L. N., Gómez, D., Gambin, B. L., Di Mauro, G., Iglesias, R., Costanzi, J., Jobbágy, E. G., & Borrás, L. (2021). Sowing date, genotype choice, and water environment control soybean yields in central Argentina. *Crop Science*, 61(1), 715–728. <https://doi.org/10.1002/csc2.20315>
- Yadav, S. A., Sahoo, B. M., Sharma, S., & Das, L. (2020). An Analysis of Data Mining Techniques to Analyze the Effect of Weather on Agriculture. *Proceedings of International Conference on Intelligent Engineering and Management, ICIEM 2020*, 29–32. <https://doi.org/10.1109/ICIEM48762.2020.9160110>
- Zuffo, A. M., Steiner, F., Busch, A., & Zoz, T. (2018). Response of early soybean cultivars to nitrogen fertilization associated with *Bradyrhizobium japonicum* inoculation. *Pesquisa Agropecuária Tropical*, 48(4), 436–446. <https://doi.org/10.1590/1983-40632018v4852637>

7. Additional Information

7.1 Acknowledgments

The authors would like to thank “Accert Pesquisa e Consultoria Agronomia” for financial support and for the availability of the experimental area.

7.2 Funding

There was no funding sponsoring the development of this research.

7.3 Conflicts of Interest

The authors declare that there are no conflicts of interest.

7.4 Data availability statement

The data and Python scripts used are available at <https://github.com/brunobro/temporal-variability-in-soybean-sowing-and-harvesting-deciphered-by-k-means-and-silhouette-scores/tree/main>