

**MACHINE LEARNING-BASED CLASSIFICATION OF SOYBEAN GRAIN
COMPOSITION PROFILES USING METEOROLOGICAL DATA****CLASSIFICAÇÃO POR APRENDIZADO DE MÁQUINA DE PERFIS
COMPOSICIONAIS DA SOJA USANDO DADOS METEOROLÓGICOS****CLASIFICACIÓN DE PERFILES DE COMPOSICIÓN DE GRANOS DE SOJA
MEDIANTE APRENDIZAJE AUTOMÁTICO UTILIZANDO DATOS
METEOROLÓGICOS**

DOI: 10.54899/dcs.v23i95.6903

Submitted: 9/3/2026

Approved: 9/9/2026

Published on: 9/18/2026

Scheila Cristiane Angnes Willers Klein

Master in Mathematical Modeling and Computation

Institution: Universidade Regional do Noroeste do Estado do Rio Grande do Sul
(UNIJUÍ)**Manuel Osorio Binelo**

Ph.D. in Teleinformatics Engineering

Institution: Universidade Regional do Noroeste do Estado do Rio Grande do Sul
(UNIJUÍ)**Ivan Ricardo Carvalho**

Ph.D. in Agronomy

Institution: Universidade Federal de Pelotas (UFPeI)

José Antonio Gonzalez da Silva

Ph.D. in Agronomy

Institution: Universidade Federal de Pelotas (UFPeI)

Claudete Luciane Teixeira

Master's Degree in Environment and Sustainable Technologies

Institution: Universidade Federal da Fronteira Sul (UFFS)

ABSTRACT

Meteorological conditions play an important role in determining soybean grain composition, although their relationships with chemical constituents are often complex and nonlinear. This study investigated the association between weather conditions and the proximate composition of soybean grains collected from commercial fields in Rio Grande do Sul during the 2024/2025 growing season. Protein, ash, lipids, fiber, and carbohydrates were quantified by near-infrared spectroscopy, while meteorological data covering the growing season were obtained for each sampling location. Linear regression analyses revealed weak linear associations, with coefficients of determination no greater than 0.202 for all evaluated components. To better represent these complex interactions, an artificial neural network

classification framework was developed. Chemical composition data were partitioned into high and low groups using K-means clustering, and the resulting labels were used to train independent artificial neural network classifiers based on mean air temperature and weekly precipitation. After 300 independent executions, the proposed models achieved mean training accuracies above 99.95% and mean testing accuracies above 88.13% for all evaluated constituents. The best-performing models reached 100% accuracy on both the training and testing datasets, while analysis of variance confirmed statistically significant differences between the predicted groups for protein, lipids, fiber, and carbohydrates. The proposed methodology demonstrates that machine learning techniques can identify meteorological patterns associated with distinct soybean grain composition profiles, providing a complementary approach for evaluating environmental influences on grain quality.

Keywords: artificial neural network; K-means clustering; nonlinear relationships; seed quality.

RESUMO

As condições meteorológicas desempenham papel importante na composição dos grãos de soja, embora suas relações com os constituintes químicos sejam complexas e não lineares. Este estudo investigou a associação entre as condições meteorológicas e a composição centesimal de grãos de soja coletados no Rio Grande do Sul durante a safra 2024/2025. Proteína, cinzas, lipídios, fibra e carboidratos foram quantificados por espectroscopia no infravermelho próximo, enquanto os dados meteorológicos do ciclo foram obtidos para cada local. As regressões lineares revelaram associações fracas, com coeficientes de determinação não superiores a 0.202 para todos os componentes. Para representar essas interações, foi desenvolvida uma estrutura de classificação por redes neurais artificiais. Os dados de composição química foram divididos em grupos de alto e baixo conteúdo por agrupamento K-means, e os rótulos resultantes foram usados para treinar classificadores independentes de redes neurais artificiais com base na temperatura média do ar e na precipitação semanal. Após 300 execuções, os modelos alcançaram acurácias médias de treinamento acima de 99.95% e de teste acima de 88.13% para todos os constituintes. Os melhores modelos atingiram 100% de acurácia nos conjuntos de treinamento e teste, enquanto a análise de variância confirmou diferenças significativas entre os grupos preditos para proteína, lipídios, fibra e carboidratos. A metodologia proposta demonstra que técnicas de aprendizado de máquina podem identificar padrões meteorológicos associados a distintos perfis de composição dos grãos de soja, fornecendo uma abordagem complementar para avaliar influências ambientais sobre a qualidade dos grãos.

Palavras-chave: rede neural artificial; agrupamento K-means; relações não lineares; qualidade das sementes.

RESUMEN

Las condiciones meteorológicas desempeñan un papel importante en la composición de los granos de soja, aunque sus relaciones con los componentes químicos son complejas y no lineales. Este estudio investigó la asociación entre las condiciones meteorológicas y la composición centesimal de granos de soja recolectados en Rio Grande do Sul durante la cosecha 2024/2025. Las proteínas, cenizas, lípidos, fibra y carbohidratos fueron cuantificados mediante espectroscopia de infrarrojo cercano, mientras que los datos meteorológicos del ciclo fueron obtenidos para cada localidad. Las regresiones lineales revelaron asociaciones débiles, con coeficientes de determinación no superiores a 0,202 para todos los componentes. Para representar estas interacciones, se desarrolló una estructura de clasificación mediante redes neuronales artificiales. Los datos de composición química se dividieron en grupos de alto y bajo contenido mediante el agrupamiento K-means, y las etiquetas resultantes se utilizaron para entrenar clasificadores independientes de redes neuronales artificiales basados en la temperatura media del aire y la precipitación semanal. Después de 300 ejecuciones, los modelos alcanzaron precisiones medias de entrenamiento superiores al 99,95 % y de prueba superiores al 88,13 % para todos los constituyentes. Los mejores modelos alcanzaron un 100 % de precisión tanto en los conjuntos de entrenamiento como de prueba, mientras que el análisis de varianza confirmó diferencias significativas entre los grupos predichos para proteínas, lípidos, fibra y carbohidratos. La metodología propuesta demuestra que las técnicas de aprendizaje automático pueden identificar patrones meteorológicos asociados con distintos perfiles de composición de los granos de soja, proporcionando un enfoque complementario para evaluar las influencias ambientales sobre la calidad de los granos.

Palabras clave: red neuronal artificial; agrupamiento K-means; relaciones no lineales; calidad de las semillas.

1 INTRODUCTION

Soybean (*Glycine max* L.) is one of the world's most important agricultural crops, representing a major source of vegetable protein and oil for human consumption, animal feed, and numerous industrial applications. Besides grain yield, seed quality is a critical factor determining the economic value of soybean production, as it directly affects its nutritional and technological properties.

The chemical composition of soybean grains is primarily determined by the relative concentrations of protein, lipids, carbohydrates, fiber, and ash. These constituents are influenced by genetic characteristics, crop management practices, and environmental conditions during plant development. Consequently, variations in grain composition are commonly observed among cultivars, production regions, and growing seasons, reflecting the interaction between genotype and environment (Staniak *et al.*, 2023; Wu *et al.*, 2024; Bruinsma *et al.*, 2025).

Hu *et al.* (2025) analyzed long-term field experiments conducted between 1993 and 2016 using 14 short-season soybean cultivars grown in eastern Canada to evaluate the effects of weather conditions and rising atmospheric carbon dioxide concentrations on seed yield and composition. Seed yield was positively associated with precipitation but negatively affected by maximum temperature and vapor pressure deficit. Regarding seed composition, greater precipitation increased protein concentration while reducing oil content, whereas higher temperatures and vapor pressure deficit favored oil accumulation and reduced both yield and protein quality. These results indicate that vapor pressure deficit may be an even stronger predictor of soybean productivity and seed composition than temperature alone.

Although the influence of meteorological conditions on soybean grain composition has been widely recognized, accurately characterizing these relationships remains challenging due to the simultaneous interaction of multiple environmental factors throughout the crop cycle. Consequently, approaches based solely on individual variables or simple statistical relationships may not fully represent the complexity of these interactions.

To address this challenge, this study proposes a data-driven framework that combines unsupervised learning and artificial neural networks to identify

meteorological patterns associated with different soybean grain composition profiles (André *et al.*, 2022; Lima *et al.*, 2024). Initially, K-means clustering was applied to define composition groups based on the proximate composition of soybean grains. These groups were subsequently used to train an artificial neural network capable of classifying meteorological conditions according to their associated composition profile. The proposed methodology demonstrates the potential of machine learning techniques to support the identification of weather conditions associated with variations in soybean grain quality, providing complementary information for crop monitoring and agricultural decision-making.

2 METHODOLOGY

2.1 OBSERVATIONAL STUDY

This observational study was conducted using soybean grain samples collected from commercial production fields selected to represent the municipalities included in the study. The selection criteria comprised soybean fields larger than 10 ha, rainfed production systems, growers adopting good agricultural practices, grain moisture standardized at 13%, and the absence of defective, heat-damaged, or mechanically damaged grains.

Soybean grain samples were collected and prepared using procedures that included sample identification, homogenization, storage, and handling to preserve their physical and chemical characteristics prior to analysis. The chemical composition of the grains was subsequently determined using near-infrared spectroscopy, which was employed to quantify protein, ash, crude fiber, lipids, and carbohydrates through calibration models previously established for soybean grains.

Table 1 summarizes the SoilGrids-derived soil attributes at 0-30 cm depth for the sampled municipalities, including geographic coordinates, texture fractions, soil organic carbon, cation exchange capacity, and soil pH.

Table 1 Soil attributes at 0-30 cm depth for the sampled municipalities in Rio Grande do Sul state

n	Municipality	Latitude (°)	Longitude (°)	Clay (%)	Silt (%)	Sand (%)	SOC (g kg ⁻¹)	CEC (cmolc kg ⁻¹)	pH
1	AJURICABA	-28.2397	-53.7693	48.8	28.2	23.0	24.1	21.82	5.12

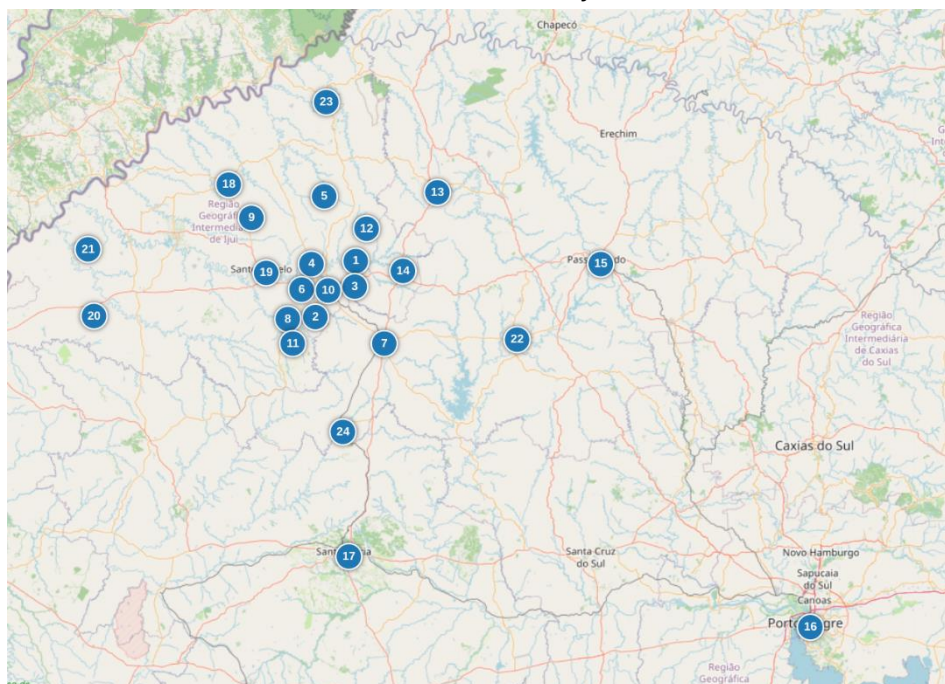
n	Municipality	Latitude (°)	Longitude (°)	Clay (%)	Silt (%)	Sand (%)	SOC (g kg ⁻¹)	CEC (cmolc kg ⁻¹)	pH
2	A. PESTANA	- 28.5148	-53.9906	43.5	29.7	26.8	23.3	20.57	5.04
3	BOZANO	- 28.3670	-53.7705	50.0	25.7	24.3	26.4	23.85	5.12
4	CATUIPE	- 28.2554	-54.0132	47.6	30.5	21.9	23.0	21.21	5.05
5	CHIAPETA	- 27.9232	-53.9419	49.4	30.1	20.5	27.3	19.01	5.11
6	C. BARROS	- 28.3826	-54.0704	50.7	29.3	20.0	23.8	21.28	5.06
7	CRUZ ALTA	- 28.6451	-53.6058	38.3	26.9	34.7	24.0	20.98	5.07
8	E. DE CASTRO	- 28.5286	-54.1475	45.1	31.5	23.4	23.2	19.70	5.10
9	GIRUA	- 28.0299	-54.3507	42.9	33.0	24.2	28.1	18.53	5.15
10	IJUI	- 28.3879	-53.9202	44.9	31.1	23.9	24.7	18.94	5.01
11	JOIA	- 28.6465	-54.1199	41.6	31.1	27.2	22.8	19.65	5.11
12	N. RAMADA	- 28.0821	-53.7064	45.4	30.4	24.2	26.7	19.52	5.06
13	PALM. DAS M.	- 27.9011	-53.3136	42.0	29.3	28.7	31.6	18.63	5.11
14	PANAMBI	- 28.2899	-53.5018	43.3	29.9	26.8	27.2	18.21	5.06
15	PASSO FUNDO	- 28.2551	-52.3967	49.5	28.0	22.5	27.7	16.66	5.04
16	PORTO ALEGRE	- 30.0325	-51.2304	28.5	28.6	42.9	68.2	25.66	5.20
17	SANTA MARIA	- 29.6861	-53.8069	21.8	26.7	51.4	28.9	18.22	5.24
18	SANTA ROSA	- 27.8644	-54.4779	44.5	31.9	23.6	25.1	18.15	5.13
19	SANTO ANGELO	- 28.3000	-54.2670	48.5	31.7	19.8	28.6	19.01	5.11
20	S. A. MISSES	- 28.5118	-55.2290	42.6	35.2	22.2	24.2	21.29	5.22
21	SAO NICOLAU	- 28.1857	-55.2615	41.1	37.6	21.3	21.8	20.13	5.21
22	TAPERA	- 28.6263	-52.8681	51.0	28.6	20.4	25.0	21.57	5.09
23	TRES PASSOS	- 27.4556	-53.9294	51.3	29.5	19.3	29.4	16.78	5.11
24	TUPANCIRETA	- 29.0799	-53.8387	37.9	34.3	27.8	24.2	18.53	5.05

Source: authors

The values were obtained by depth-weighted averaging of the 0-5, 5-15, and 15-30 cm layers, allowing a consistent representation of the upper soil profile across all sampling locations. Figure 1 shows the spatial distribution of the sampled

municipalities within the study region, with numbered markers corresponding to the municipality identifiers reported in Table 1. Soil properties were obtained from the SoilGrids 2.0 global digital soil mapping system (Poggio *et al.*, 2021).

Figure 1. Location of the municipalities in the state of Rio Grande do Sul considered in the observational study.



Source: authors

Table 2 summarizes the meteorological conditions observed at each location between October 15, 2024, and April 30, 2025. The variable Tmean corresponds to the overall mean air temperature, Tmax to the mean daily maximum temperature, Tmin to the mean daily minimum temperature, and Precip. to the cumulative precipitation during the study period. Among the evaluated locations, Passo Fundo exhibited the lowest mean temperature (21.98 °C), while São Nicolau recorded the highest mean temperature (25.07 °C).

Table 2 - Meteorological variables by municipality, from October 15, 2024, to April 30, 2025

Location	Tmean (°C)	Tmax (°C)	Tmin (°C)	Precip. (mm)
AJURICABA	23.24	29.69	17.34	601.68
A. PESTANA	23.03	30.09	17.00	615.73
BOZANO	23.03	30.09	17.00	615.73
CATUIPE	23.24	29.69	17.34	601.68
CHIAPETA	23.24	29.69	17.34	601.68
C. BARROS	23.03	30.09	17.00	615.73
CRUZ ALTA	23.03	30.09	17.00	615.73
E. DE CASTRO	24.02	31.76	17.40	620.26

Location	Tmean (°C)	Tmax (°C)	Tmin (°C)	Precip. (mm)
GIRUA	24.34	31.22	17.90	608.57
IJUI	23.03	30.09	17.00	615.73
JOIA	24.02	31.76	17.40	620.26
N. RAMADA	23.24	29.69	17.34	601.68
PALM. DAS M.	22.91	29.14	17.07	556.57
PANAMBI	23.03	30.09	17.00	615.73
PASSO FUNDO	21.98	28.24	16.41	520.71
PORTO ALEGRE	23.17	28.61	18.49	503.81
SANTA MARIA	22.53	29.35	16.45	565.54
SANTA ROSA	24.34	31.22	17.90	608.57
SANTO ANGELO	24.02	31.76	17.40	620.26
S. A. MISSES	24.93	33.15	18.00	562.27
SAO NICOLAU	25.07	32.97	18.35	592.50
TAPERA	22.64	28.73	16.94	554.02
TRES PASSOS	24.00	30.63	17.94	565.54
TUPANCIRETA	22.32	29.85	16.15	591.25

Tmean - overall mean air temperature; Tmax - mean daily maximum temperature; Tmin - mean daily minimum temperature; and Precip. - precipitation during the study period.

Source: authors

2.2 CLASSIFICATION MODEL

The dataset used in this study comprises the proximate composition of soybean grains and meteorological data collected from 24 municipalities in the state of Rio Grande do Sul, Brazil. Meteorological records were selected for the period from October 15, 2024, to April 30, 2025, corresponding to the typical soybean growing season in Brazil, from planting to harvest.

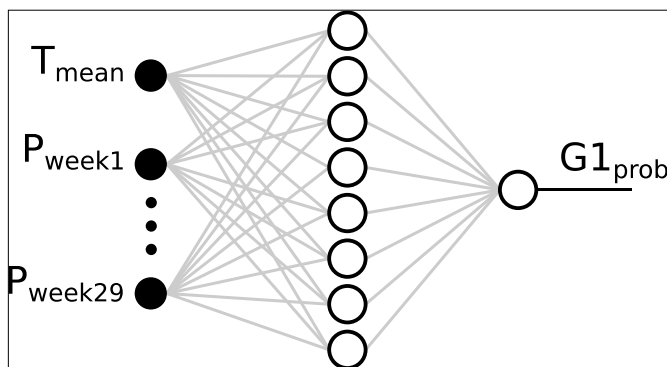
As will be demonstrated in the Results section, there is no direct linear relationship between meteorological factors and the proximate composition of soybean grains. In light of this, the present study proposes an artificial neural network (ANN) model capable of classifying different meteorological conditions according to their potential to promote either high or low levels of each chemical constituent. It is important to note that the categories “high” and “low” are defined relative to the distribution of values observed within the study region.

Artificial neural networks were selected due to their ability to capture complex nonlinear relationships between input and output variables. Furthermore, grain composition is influenced by several factors beyond the meteorological variables considered in this study, including soil properties, cultivation practices, pest pressure, and other environmental conditions. These factors introduce additional uncertainty into

the modeling process, making a classification-based approach more appropriate than deterministic continuous prediction models. Consequently, the proposed ANN estimates the probability of belonging to predefined groups (high or low levels of each chemical constituent), providing a more robust framework for decision-making under uncertainty.

The classification model proposed in this study is the artificial neural network architecture illustrated in Figure 2. The selected input variables were the average temperature over the entire growing season (T_{mean}) and the weekly precipitation values ($P_{\text{week1}} \dots 29$) throughout the study period.

Figure 2. Classification artificial neural network scheme.



Source: authors

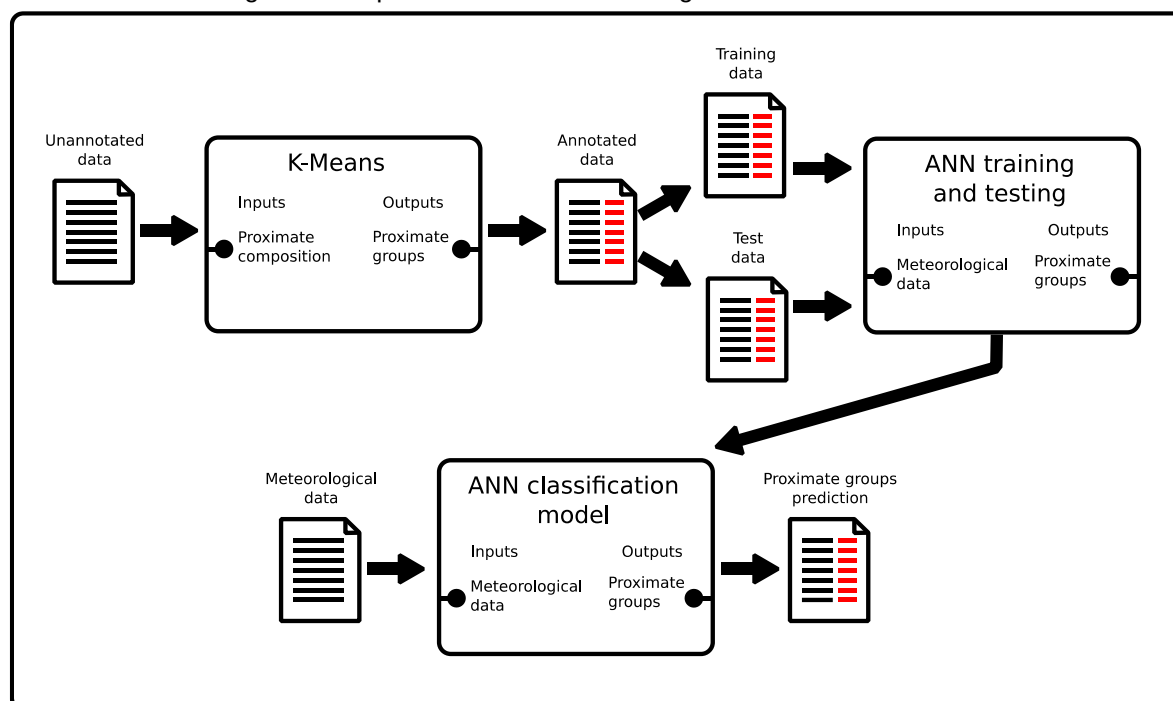
As will also be demonstrated in the Results section (Figure 4), the temporal temperature profiles exhibit a similar dynamic pattern across all municipalities, differing mainly in their average levels. For this reason, only the average temperature was used as an input variable. In contrast, precipitation exhibits distinct temporal patterns among municipalities (Figure 6), which motivated the inclusion of weekly precipitation values as separate inputs to the network.

The artificial neural network output layer consists of a single neuron whose value represents the probability ($G1_{\text{prob}}$) that a given meteorological profile belongs to Group 1 (High). Accordingly, a meteorological profile is classified as High when $G1_{\text{prob}} \geq 0.5$ and as Low when $G1_{\text{prob}} < 0.5$ for each proximate composition parameter.

The hidden layer was composed of eight neurons. This number was determined through successive experiments in which the number of hidden neurons was gradually increased until no further improvement in predictive performance was observed. A separate neural network was developed for each proximate composition parameter.

To avoid potential bias in the annotation of artificial neural network training data, the dataset was processed according to the procedure described in Figure 3. Initially, the unannotated data, that is, data without a predefined assignment to a proximate composition group, were subjected to a K-means clustering procedure (Hartigan & Wong, 1979). Based on the clustering results, each sample, represented by its meteorological variables, was automatically annotated as belonging to one of two groups (Low or High) for each proximate composition parameter.

Figure 3. Proposed workflow for training the classification model



Source: authors

The annotated data were subsequently divided into training (80%) and testing (20%) datasets, and an artificial neural network was created, trained, and tested. The resulting classifications were then subjected to analysis of variance to verify whether the groups identified by the model were statistically distinct.

Since both the K-means clustering procedure and artificial neural network training involve stochastic processes, the entire workflow was repeated 300 times. For each proximate composition parameter, the average training and testing accuracies were calculated, together with the corresponding accuracies of the best-performing model. The best model was defined as the one achieving 100% accuracy on both the training and testing datasets while also presenting the lowest analysis-of-variance p-value.

The final outcome of this procedure is a classification model capable of receiving the climatic characteristics of a given location as input and predicting whether the expected level of a specific proximate composition parameter is High or Low.

All computational procedures were implemented in Python using the Scikit-learn machine learning framework. The K-means clustering algorithm was implemented through the `sklearn.cluster` module, whereas the artificial neural network models were developed using the `sklearn.neural_network` module.

3 RESULTS AND DISCUSSIONS

The chemical composition of soybean grains presented in Table 3 reveals substantial variability among the evaluated municipalities.

Table 3 - Chemical composition of soybean grains from the municipalities in the state of Rio Grande do Sul considered in the study. (mean $\pm$ standard deviation)

Location	Protein (%)	Ash (%)	Lipids (%)	Fiber (%)	Carbohydrates (%)	Samples
AJURICABA	37.55 $\pm$ 1.12	5.22 $\pm$ 0.23	16.27 $\pm$ 3.13	2.94 $\pm$ 2.69	29.30 $\pm$ 0.25	2
A. PESTANA	35.51 $\pm$ 3.80	5.36 $\pm$ 0.71	18.35 $\pm$ 3.17	3.07 $\pm$ 1.63	27.11 $\pm$ 3.84	704
BOZANO	35.70 $\pm$ 3.52	5.17 $\pm$ 0.32	17.45 $\pm$ 1.43	3.84 $\pm$ 1.78	28.07 $\pm$ 4.21	5
CATUIPE	32.32	5.14	17.68	2.35	33.03	1
CHIAPETA	33.70 $\pm$ 0.79	5.10 $\pm$ 0.25	17.89 $\pm$ 4.75	1.68 $\pm$ 0.47	28.75 $\pm$ 1.02	4
C. BARROS	32.32	7.79	14.17	3.60	28.70	1
CRUZ ALTA	41.05	4.92	24.38	3.86	31.61	1
E. DE CASTRO	38.99 $\pm$ 4.50	5.03 $\pm$ 0.27	18.81 $\pm$ 3.36	2.44 $\pm$ 0.82	28.31 $\pm$ 3.37	4
GIRUA	33.05 $\pm$ 5.28	5.37 $\pm$ 0.17	22.09 $\pm$ 3.07	2.54 $\pm$ 1.59	29.00 $\pm$ 2.70	4
IJUI	35.78 $\pm$ 3.59	5.32 $\pm$ 0.66	17.98 $\pm$ 3.30	3.11 $\pm$ 1.54	26.91 $\pm$ 4.11	102
JOIA	34.39 $\pm$ 3.32	5.12 $\pm$ 0.25	15.91 $\pm$ 1.89	2.04 $\pm$ 0.94	28.44 $\pm$ 2.99	7
N. RAMADA	35.23 $\pm$ 0.79	6.00 $\pm$ 1.55	21.71 $\pm$ 3.64	3.02 $\pm$ 2.36	22.90 $\pm$ 1.87	3
PALM. DAS M.	35.60 $\pm$ 2.90	5.61 $\pm$ 1.09	18.72 $\pm$ 3.11	3.07 $\pm$ 1.62	28.62 $\pm$ 2.47	6
PANAMBI	39.50 $\pm$ 5.92	5.16 $\pm$ 0.30	16.78 $\pm$ 1.96	1.66 $\pm$ 0.41	27.72 $\pm$ 3.96	3
PASSO FUNDO	36.62 $\pm$ 3.76	5.25 $\pm$ 0.30	18.60 $\pm$ 2.90	2.49 $\pm$ 2.11	26.04 $\pm$ 2.52	3
PORTO ALEGRE	40.51 $\pm$ 6.40	5.17 $\pm$ 0.19	20.71 $\pm$ 2.89	3.32 $\pm$ 1.47	29.17 $\pm$ 1.81	4
SANTA MARIA	44.99	4.97	17.91	4.78	29.13	1
SANTA ROSA	32.53 $\pm$ 0.92	5.27 $\pm$ 0.13	19.00 $\pm$ 3.21	1.95 $\pm$ 0.47	25.43 $\pm$ 3.99	5

Location	Protein (%)	Ash (%)	Lipids (%)	Fiber (%)	Carbohydrates (%)	Samples
SANTO ANGELO	35.76 ± 4.17	5.41 ± 0.74	18.05 ± 2.96	2.88 ± 1.62	27.13 ± 4.00	124
S. A. MISSOES	33.35 ± 4.05	5.97 ± 1.60	18.36 ± 3.27	3.86 ± 2.49	29.10 ± 2.53	3
SAO NICOLAU	34.85 ± 3.67	6.01 ± 1.23	17.54 ± 2.32	3.49 ± 1.44	26.96 ± 2.56	7
TAPERA	42.60	5.46	24.38	2.35	20.77	1
TRES PASSOS	30.30 ± 1.86	6.45 ± 1.90	23.20 ± 0.00	1.48 ± 0.32	29.79 ± 4.58	2
TUPANCIRETA	35.11 ± 2.14	6.04 ± 1.53	14.38 ± 0.55	2.44 ± 2.03	27.39 ± 1.71	3

Source: authors

Protein content ranged from 30.30% to 44.99%, whereas lipid content varied from 14.17% to 24.38%. Overall, the observed values fall within the ranges commonly reported for soybean crops, whose typical composition comprises approximately 35-45% protein and 18-25% oil on a dry matter basis (Bellaloui *et al.*, 2020). This wide variation confirms that both environmental and genetic factors significantly influence the final chemical composition of soybean grains. In a meta-analysis encompassing experiments conducted across different production regions, Rotundo & Westgate (2009) demonstrated that factors such as water availability, temperature, and nitrogen supply significantly affect the accumulation of protein and oil in soybean seeds.

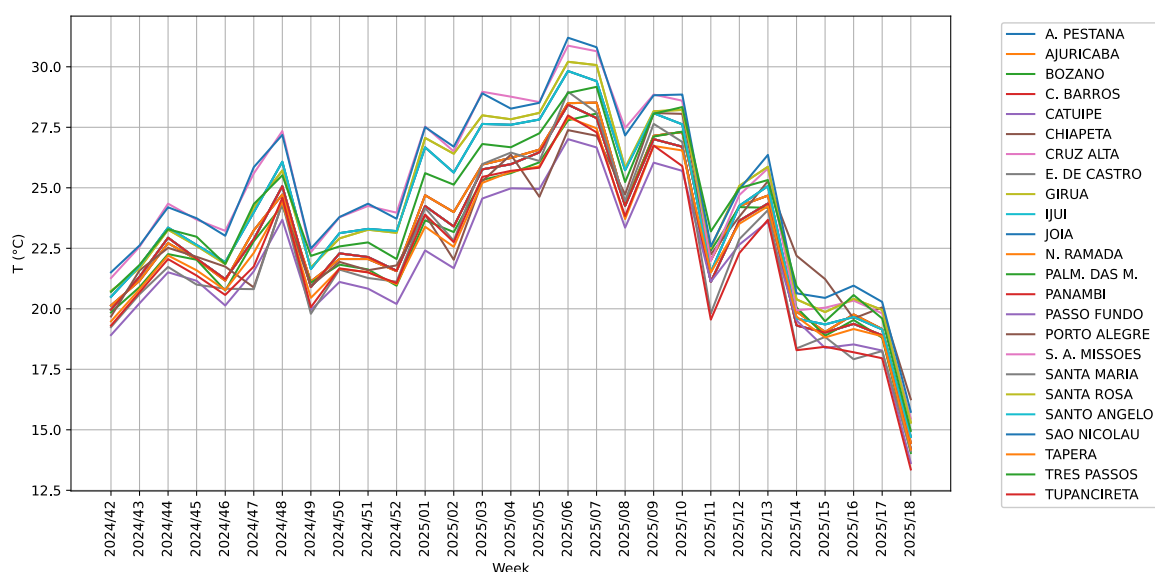
The results presented in Table 3 also support the well-established inverse relationship between protein and oil concentrations reported in the literature (Hwang *et al.*, 2014). Municipalities such as Santa Maria, Cruz Alta, and Tapera exhibited relatively high protein contents, whereas others were characterized by higher lipid concentrations. Consequently, the development of cultivars capable of simultaneously achieving high protein and oil contents remains one of the major challenges faced by modern soybean breeding programs.

The findings of the present study are also consistent with those reported by Kumar *et al.* (2006), Bellaloui *et al.* (2020), and Wu *et al.* (2024), who observed that geographic location can significantly influence the concentrations of protein, oil, and fatty acids in soybean grains. These authors highlighted the importance of genotype-by-environment interactions in determining seed composition, reinforcing the hypothesis that the differences observed among the municipalities evaluated in this study are associated not only with genetic variability but also with distinct environmental conditions during crop development.

A limitation of this observational study is that the number of soybean grain samples was not uniform across the evaluated municipalities, with some locations being represented by only a single sample due to sample availability. However, this limitation did not appear to adversely affect the proposed methodology. As will be shown later, the high predictive performance of the artificial neural network models and the statistically significant separation between most composition groups indicate that the meteorological patterns identified were sufficiently robust despite the unequal sampling distribution among locations.

An important aspect in the development of the classification model is determining whether average meteorological conditions are sufficient to characterize each location or whether temporal variations throughout the growing season must also be considered. Figure 4 illustrates the temporal dynamics of air temperature during the study period. Although slight differences in temperature levels can be observed among municipalities, the overall temporal patterns are remarkably similar across locations.

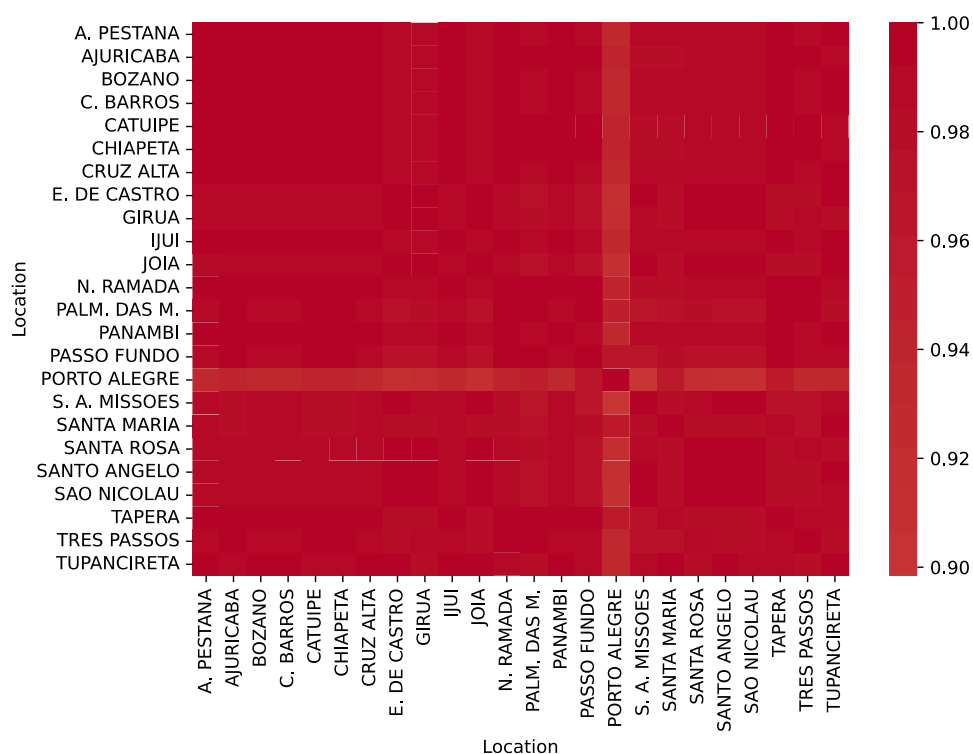
Figure 4. Temporal dynamics of mean air temperature throughout the study period for municipalities in the state of Rio Grande do Sul considered in the study.



Source: authors

This observation is further supported by Figure 5, which presents the correlation matrix of temperature dynamics among municipalities. The results reveal an average correlation coefficient of 0.985, indicating an almost identical temporal behavior of temperature across the evaluated locations. Therefore, while the municipalities differ in their average temperature levels, they exhibit very similar temporal temperature profiles throughout the study period.

Figure 5. Correlation matrix of temperature dynamics among locations, showing an average correlation coefficient of 0.985 for municipalities in the state of Rio Grande do Sul considered in the study.



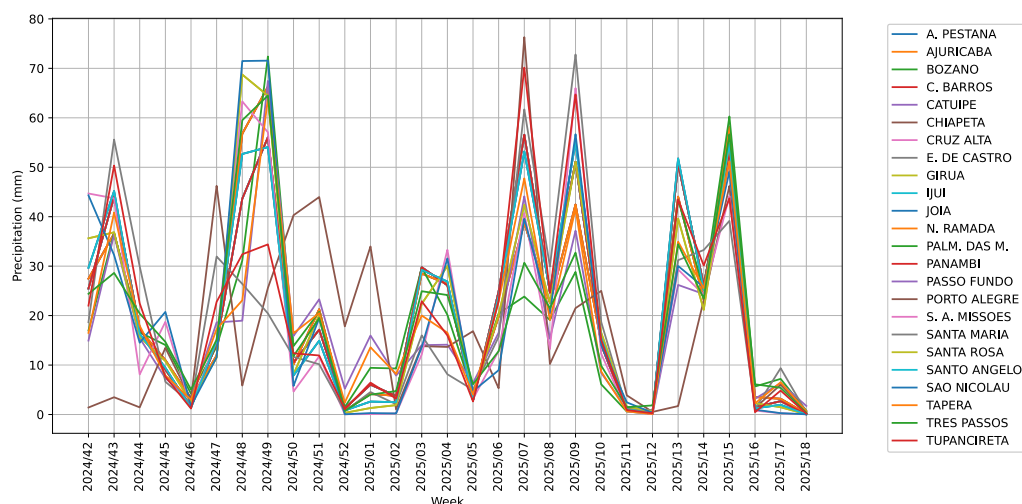
Source: authors

Given the extremely high correlation observed among temperature time series, it is reasonable to assume that the mean temperature alone is sufficient to capture the climatic differences among the municipalities considered in this study. Consequently, incorporating the complete temperature dynamics into the classification model is unlikely to provide substantial additional information beyond that already represented by the average temperature values.

Figure 6 illustrates the weekly precipitation dynamics throughout the study period. In contrast to the temperature patterns, greater variability can be observed among the municipalities, indicating more pronounced differences in rainfall distribution over time. This observation is further supported by Figure 7, which presents the correlation matrix of weekly precipitation patterns across municipalities. The

average correlation coefficient was 0.862, suggesting a relatively strong overall relationship, but noticeably lower than that observed for temperature.

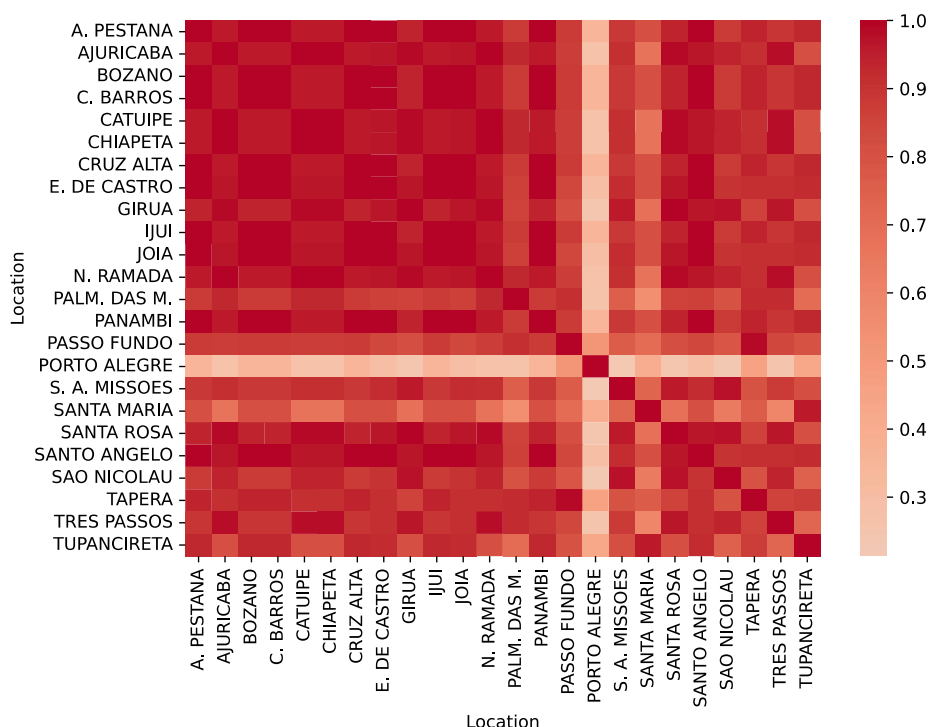
Figure 6. Temporal dynamics of precipitation during the study period for municipalities in the state of Rio Grande do Sul considered in the study.



Source: authors

Although the precipitation time series remain highly correlated, some municipalities exhibit distinct rainfall patterns throughout the growing season. These differences indicate that average precipitation alone may not adequately capture the variability among locations.

Figure 7. Correlation matrix of weekly precipitation patterns among locations, showing an average correlation coefficient of 0.862



Source: authors

Therefore, the use of the complete weekly precipitation profile, rather than only the mean precipitation value, is justified as an input to the classification model, as illustrated in Figure 2. By incorporating the temporal distribution of rainfall, the model can better represent location-specific environmental conditions that may influence soybean grain composition.

Table 4 summarizes the linear regression results for each chemical component as a function of mean temperature and precipitation.

Table 4. Linear regression parameters between meteorological variables and chemical composition

X variable	Y variable	R ²	Angular coefficient	Intercept
Mean temperature	Protein	0.201631	-2.040309	83.866148
	Ash	0.013433	0.092642	3.345961
	Lipids	0.005476	0.258089	12.726511
	Fiber	0.006143	-0.081457	4.750189
	Carbohydrates	0.016288	0.404836	18.419981
Precipitation	Protein	0.071797	-5.793064	53.437789
	Ash	0.000016	-0.015128	5.558286
	Lipids	0.081992	-4.751731	32.954365
	Fiber	0.012114	-0.544283	4.470104
	Carbohydrates	0.013535	1.755922	22.646402

Source: authors

In all cases, the observed correlation coefficients were low, suggesting weak linear associations between the meteorological variables and grain composition. These findings indicate that the underlying relationships are likely more complex and cannot be accurately represented by simple linear models. Consequently, more sophisticated approaches capable of capturing nonlinear interactions among environmental factors may be required to adequately model the phenomenon.

The results obtained in this study are consistent with previous reports indicating that meteorological conditions influence soybean grain composition, while also highlighting the complexity of these relationships. Several studies have reported associations between temperature, precipitation, and the accumulation of protein, oil, and other chemical constituents in soybean seeds (Bennett & Krishnan, 2005; Song *et al.*, 2016; Novikova *et al.*, 2018; Song *et al.*, 2018; Hu *et al.*, 2025). The present work found weak linear relationships between average temperature or cumulative precipitation and the evaluated proximate composition parameters, with all coefficients of determination remaining below 0.21. These findings suggest that, under the environmental conditions evaluated in southern Brazil, soybean grain composition cannot be adequately explained by isolated meteorological variables.

The low linear correlations observed in this study do not necessarily contradict the literature. Previous investigations have shown that the effects of weather conditions on soybean composition depend not only on the magnitude of temperature and precipitation but also on their temporal distribution throughout the crop cycle, as well as on interactions with genotype, soil characteristics, and other environmental factors (Song *et al.*, 2018; Zhu *et al.*, 2024; Hu *et al.*, 2025). In agreement with these observations, the present results indicate that the influence of meteorological conditions is predominantly nonlinear, supporting the use of machine learning techniques capable of representing complex multivariate relationships.

The classification model developed to predict the potential of a given location to produce either high or low levels of a specific chemical component based on meteorological conditions demonstrated a substantially greater ability to capture the underlying behavior of the phenomenon. As shown in Table 5, after 300 independent executions of the procedure illustrated in Figure 3, the mean training classification accuracy exceeded 99.95% in all cases, while the mean testing accuracy remained above 88.13%.

Table 5. Classification accuracy metrics for chemical composition variables

Constituent	Mean train accuracy (%)	Mean test accuracy (%)	Best train accuracy (%)	Best test accuracy (%)	p-value
Protein	99.96	89.40	100	100	0.00204
Ash	99.96	89.00	100	100	0.37020
Lipids	99.98	89.13	100	100	0.01019
Fiber	99.95	90.13	100	100	0.00426
Carbohydrates	99.95	88.13	100	100	0.02485

Source: authors

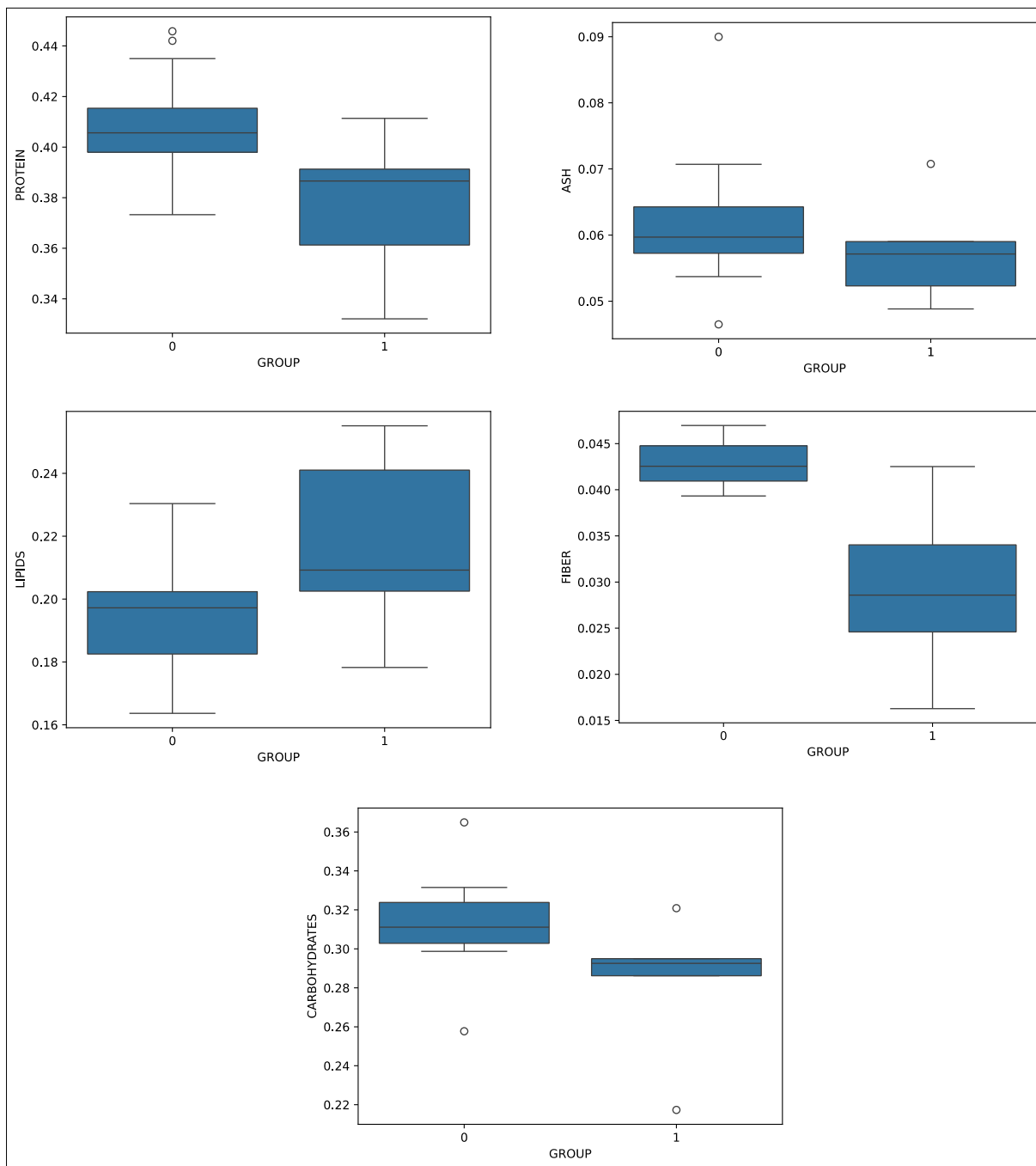
For each chemical component, the best-performing artificial neural network achieved 100% accuracy on both the training and testing datasets. These results indicate that the proposed approach was highly effective in identifying the complex relationships between environmental conditions and soybean grain composition.

To further evaluate the consistency of the predicted classes, an analysis of variance was performed. With the exception of ash content, all chemical components produced p-values below 0.05, indicating statistically significant differences between the groups identified by the classifier. These findings provide additional evidence that the proposed classification framework successfully captured meaningful patterns in the data and was able to distinguish locations associated with distinct chemical composition profiles.

Figure 9 presents the analysis-of-variance box plots for each chemical component. The figure also provides a visual assessment of the ability of the artificial neural network to separate the samples into groups with high and low production potential for each chemical property. For most components, the predicted groups exhibit clear differences in their distributions, corroborating the statistical significance

identified by the analysis-of-variance results and demonstrating the effectiveness of the proposed classification approach.

Figure 9. Box plots of the ANOVA results for the high- and low-production groups predicted by the best-performing ANN for each chemical property



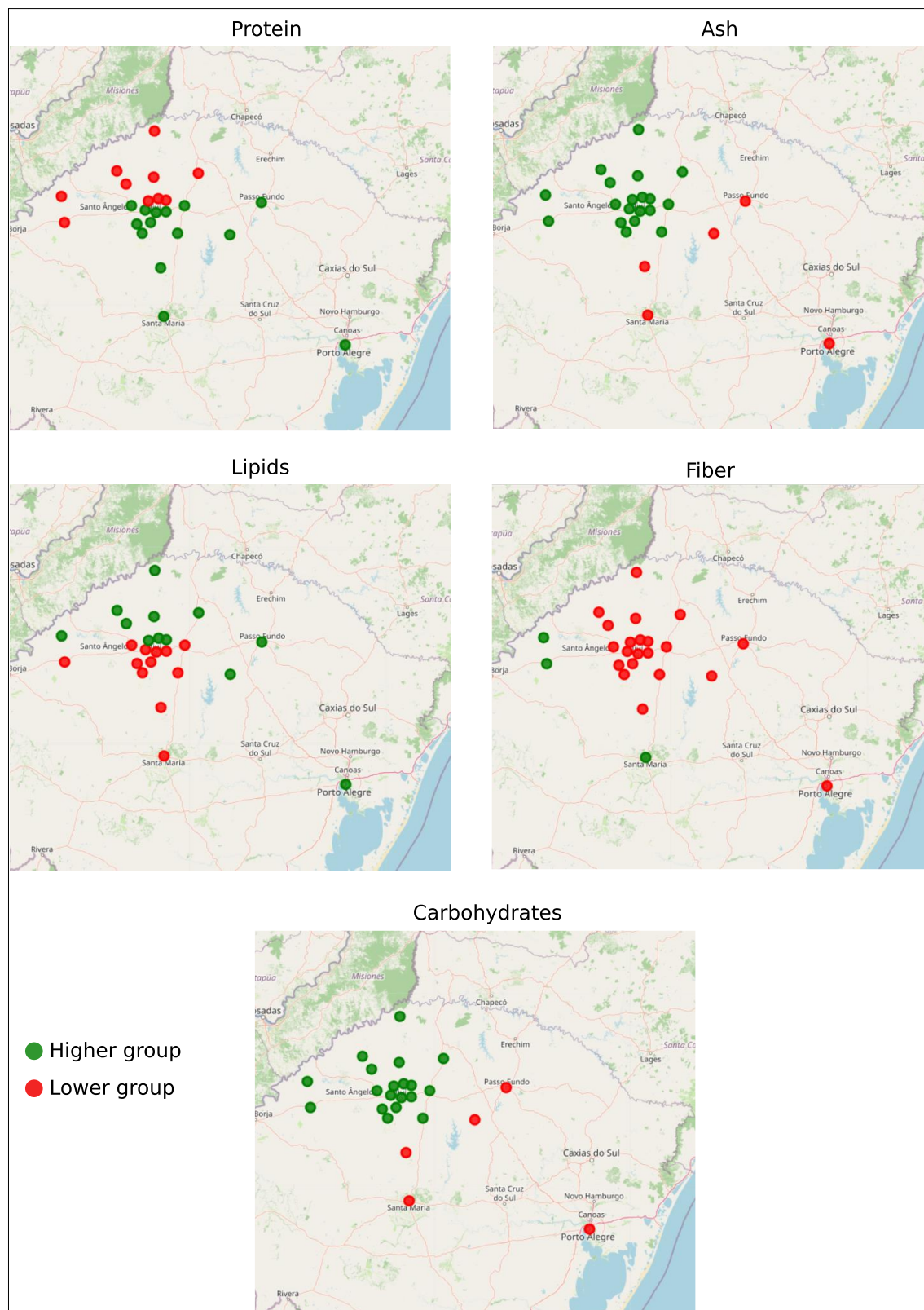
Source: authors

The proposed artificial-neural-network-based classification framework substantially improved the representation of these relationships compared with simple linear regression. While linear models explained only a small fraction of the observed variability, the classification models consistently achieved testing accuracies above

88%, suggesting that meaningful meteorological patterns associated with soybean grain composition were successfully identified. Rather than predicting the exact concentration of each chemical constituent, the proposed methodology identifies meteorological profiles associated with higher or lower levels of protein, ash, lipids, fiber, and carbohydrates, providing a robust alternative for situations in which deterministic prediction is difficult due to the complexity of environmental interactions.

Figure 10 illustrates the spatial distribution of the municipalities classified as having either high or low potential for the production of each chemical property according to the proposed artificial neural network model. An interesting observation is that, although geographic location was not explicitly included as an input variable, the resulting classifications exhibit a clear spatial coherence. In several cases, the predicted groups tend to be organized according to their geographic position, with a noticeable separation between municipalities located in the northern and southern portions of the study area.

Figure 10. Spatial distribution of the high- and low-content groups predicted by the best-performing ANN for each soybean chemical property for municipalities in the state of Rio Grande do Sul considered in the study.



Source: authors

This spatial consistency provides additional evidence of the robustness of the proposed classification framework. The emergence of geographically coherent patterns from a model based solely on meteorological variables suggests that the artificial neural network successfully captured meaningful environmental gradients associated with soybean grain composition. Consequently, the observed spatial structure reinforces the plausibility of the predictions and supports the effectiveness of the developed classification approach.

4 CONCLUSION

Simple linear regression models did not adequately represent the relationship between meteorological variables and soybean grain composition.

The K-means and artificial neural network classification framework achieved mean training accuracies above 99.95% and mean testing accuracies above 88.13% for all evaluated chemical properties.

The best-performing models achieved 100% accuracy in the training and testing datasets.

The predicted groups differed significantly for protein, lipids, fiber, and carbohydrates, but not for ash.

The predicted groups showed spatial coherence even though geographic coordinates were not included as model inputs.

REFERENCES

- ANDRÉ, G. da S.; CORADI, P. C.; TEODORO, L. P. R.; TEODORO, P. E. Predicting the quality of soybean seeds stored in different environments and packaging using machine learning. **Scientific Reports**, v. 12, 8793, 2022. DOI: <https://doi.org/10.1038/s41598-022-12863-5>.
- BELLALOUI, N.; MCCLURE, A. M.; MENGISTU, A.; ABBAS, H. K. The influence of agricultural practices, the environment, and cultivar differences on soybean seed protein, oil, sugars, and amino acids. **Plants**, v. 9, 378, 2020. DOI: <https://doi.org/10.3390/plants9030378>.
- BENNETT, J. O.; KRISHNAN, H. B. Long-term study of weather effects on soybean seed composition. **Korean Journal of Crop Science**, v. 50, p. 32-38, 2005.
- BRUINSMA, G. M. W.; CARVALHO, I. R.; COLET, C. de F.; BABESKI, C. M.; SANGIOVO, J. P.; ZUSE, G. H. Fifteen years of soybean productivity in Rio Grande do Sul and the decomposition of phenotypic variation along with relations with the production environment. **Crop Science**, v. 65, e70141, 2025. DOI: <https://doi.org/10.1002/csc2.70141>.
- HARTIGAN, J. A.; WONG, M. A. Algorithm AS 136: A K-means clustering algorithm. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, v. 28, p. 100-108, 1979. DOI: <https://doi.org/10.2307/2346830>.
- HU, Y.; SOOLANAYAKANAHALLY, R. Y.; GUY, R. D.; LU, X.; SILIM, S. N.; COBER, E. R.; MORRISON, M. J. Long-term field trials assess temporal trends and growing season weather on soybean seed yield, nutrient composition and nitrogen dynamics. **Plant and Soil**, v. 514, p. 1957-1972, 2025. DOI: <https://doi.org/10.1007/s11104-025-07499-z>.
- HWANG, E. Y.; SONG, Q.; JIA, G.; SPECHT, J. E.; HYTEN, D. L.; COSTA, J.; CREGAN, P. B. A genome-wide association study of seed protein and oil content in soybean. **BMC Genomics**, v. 15, 1, 2014. DOI: <https://doi.org/10.1186/1471-2164-15-1>.
- KUMAR, V.; RANI, A.; SOLANKI, S.; HUSSAIN, S. M. Influence of growing environment on the biochemical composition and physical characteristics of soybean seed. **Journal of Food Composition and Analysis**, v. 19, p. 188-195, 2006. DOI: <https://doi.org/10.1016/j.jfca.2005.06.005>.
- LIMA, R. E.; CORADI, P. C.; RODRIGUES, D. M.; TEODORO, P. E.; TEODORO, L. P. R.; OLIVEIRA, D. P. de. Monitoring and predicting the quality of soybeans for different drying and storage technologies on a real scale using sensors and machine learning models. **Journal of Stored Products Research**, v. 108, 102386, 2024. DOI: <https://doi.org/10.1016/j.jspr.2024.102386>.
- NOVIKOVA, L. Y.; SEFEROVA, I. V.; NEKRASOV, A. Y.; PERCHUK, I. N.; SHELENGA, T. V.; SAMSONOVA, M. G.; VISHNYAKOVA, M. A. Impact of weather and climate on seed protein and oil content of soybean in the North Caucasus.

Vavilov Journal of Genetics and Breeding, v. 22, p. 708-715, 2018. DOI: <https://doi.org/10.18699/VJ18.414>.

POGGIO, L.; DE SOUSA, L. M.; BATJES, N. H.; HEUVELINK, G. B. M.; KEMPEN, B.; RIBEIRO, E.; ROSSITER, D. SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. **Soil**, v. 7, p. 217-240, 2021. DOI: <https://doi.org/10.5194/soil-7-217-2021>.

ROTUNDO, J. L.; WESTGATE, M. E. Meta-analysis of environmental effects on soybean seed composition. **Field Crops Research**, v. 110, p. 147-156, 2009. DOI: <https://doi.org/10.1016/j.fcr.2008.07.012>.

SONG, W.; YANG, R.; WU, T.; WU, C.; SUN, S.; ZHANG, S.; JIANG, B.; TIAN, S.; LIU, X.; HAN, T. Analyzing the effects of climate factors on soybean protein, oil contents, and composition by extensive and high-density sampling in China. **Journal of Agricultural and Food Chemistry**, v. 64, p. 4121-4130, 2016. DOI: <https://doi.org/10.1021/acs.jafc.6b00008>.

SONG, W.; YANG, R.; YANG, X.; SUN, S.; MENTREDDY, S. R.; JIANG, B.; WU, T.; TIAN, S.; SAPEY, E.; WU, C.; HOU, W.; REN, G.; HAN, T. Spatial differences in soybean bioactive components across China and their influence by weather factors. **The Crop Journal**, v. 6, p. 659-668, 2018. DOI: <https://doi.org/10.1016/j.cj.2018.05.001>.

STANIAK, M.; SZPUNAR-KROK, E.; KOCIRA, A. Responses of soybean to selected abiotic stresses - photoperiod, temperature and water. **Agriculture**, v. 13, 146, 2023. DOI: <https://doi.org/10.3390/agriculture13010146>.

WU, C.; ACUÑA, A.; FLOREZ-PALACIOS, L.; HARRISON, D.; ROGERS, D.; MOZZONI, L.; MIAN, R.; VIEIRA, C. C. Across-environment seed protein stability and genetic architecture of seed components in soybean. **Scientific Reports**, v. 14, 16452, 2024. DOI: <https://doi.org/10.1038/s41598-024-67035-4>.

ZHU, W.; LI, J.; XIE, T. Impact of climate change on soybean production: Research progress and response strategies. **Advances in Resources Research**, v. 4, p. 474-496, 2024. DOI: https://doi.org/10.50908/arr.4.3_474.