Australian Journal of Crop Science AJCS

Aust J Crop Sci. 19(11):1078-1089 (2025) | ISSN:1835-2707
https://doi.org/10.21475/ajcs.25.19.11.p12

Geostatistical models for asymmetric agricultural data

Amilton Luciano Garcia da Silva[1]*, Miguel Angel Uribe-Opazo[1], Jerry Adriani Johann[1], Gustavo Henrique Dalposso[2]

[1]PGEAGRI, Western Paraná State University – (UNIOESTE), Cascavel, Paraná, Brazil
[2]PPGBio, Federal University of Technology Paraná – (UTFPR), Toledo, Paraná, Brazil

***Corresponding author: amiltonlucianogarcia@gmail.com**

Submitted:* *10/04/2025

Revised:* *16/06/2025

Accepted:* *03/08/2025

Abstract: Soybean production (*Glycine max* (L.) Merrill) is key to the global economy and environmental sustainability, but it faces the challenge of increasing productivity without harming the environment. In this context, geostatistics appears as an essential tool for Precision Agriculture (AP), allowing the mapping of spatial variability of factors such as soybean productivity and soil physicochemical attributes, which helps in making more efficient decisions, for optimizing input application, improving crop management, reducing environmental impact, and maximizing yield. This study was carried out in a commercial area of 173.04 ha during the 2022/2023 harvest. We analyzed soybean yield data and soil attributes, such as nutrient content and mechanical resistance to penetration, which required data transformations due to asymmetric distributions. Diagnostic techniques of local influence were used to identify influential observations, whose impacts were evaluated in parameter estimates, in the generation of thematic maps and in the definition of management zones. The exclusion of these observations changed spatial patterns and productivity estimates, highlighting the importance of careful analysis. Although, in some cases, the isolation forest method has identified outliers that coincided with influential observations, it is important to emphasize that this detection is not directly related to the concept of influential observations, since the methods have different approaches. The proposed procedure contributes to a more sustainable agriculture, reducing the environmental impact and optimizing the use of resources, aligning greater profitability with environmental responsibility.

Keywords: precision agriculture, geostatistics, local influence, spatial analysis of asymmetric data.
Abbreviations: AIC_Akaike Information Criterion; BIC_Bayesian Information Criterion; K_soil potassium content; K#42_potassium without the influential observation #42; ML_maximum likelihood; P_soil phosphorus content; P#19_phosphorus without the influential observation #19; PA_precision agriculture; pH_soil pH; pH#45_soil pH without the influential observation #45; Prod_soybean productivity; Prod#17_soybean productivity without the influential observation #17; $RSP_{0.0-0.10m}$_soil penetration resistance at a depth of 0.0 to 0.10 meters depth layer; $RSP_{0.0-0.10m}$#99_soil penetration resistance in the 0.0 to 0.10 meters depth layer without the influential observation #99; $RSP_{0.31-0.40m}$_soil penetration resistance at a depth of 0.31 to 0.40 meters depth layer; $RSP_{0.31-0.40m}$94_soil penetration resistance in the 0.31 to 0.40 meters depth layer without the influential observation #94; SDI_Spatial Dependency Index.

Introduction

Modern agriculture faces a crucial paradox: how to meet the growing global demand for food, intensified by climate change, while seeking to preserve environmental sustainability and strengthen the resilience of the planet (IPCC, 2019; Castaldi et al., 2024). In this challenging scenario, the soybean production chain emerges as one of the most strategic in the world. Besides being one of the main sources of protein for animal nutrition (Monteiro et al., 2021) and human nutrition (Chi et al., 2021), soybean plays a vital role in the global energy matrix, especially in the biodiesel production (Zhu et al., 2021) as a sustainable alternative to fossil fuels.

With the shortage of new agricultural areas available, the future of global soybean production is intrinsically linked to productivity gains at the level of rural properties (Masino et al., 2018). In this context, precision agriculture (PA) stands out as an indispensable ally, enabling management practices adapted to the spatial variability of the factors that influence production. This approach brings benefits such as increased productivity, greater economic return and a reduction in environmental impact, by promoting the rational and efficient use of agricultural inputs (Zain et al., 2024).

Geostatistics appears in this scenario, as a fundamental scientific pillar for the implementation of AP. Its ability to model and describe the spatial variability of natural phenomena is instrumental in estimating values in unsampled areas and to

adapt traditional statistical methods to the study of spatial dependence on data (Uribe-Opazo et al., 2021; 2023). Through techniques such as kriging, geostatistics allows the construction of thematic maps, fundamental for decision-making.
In addition, the identification of influential observations are critical steps in geostatistical analysis. Such points may distort environmental and geological patterns, changing the estimates of parameters and the results interpretation (Uribe-Opazo et al., 2012). To evaluate the impact of disturbances on the data or model, Cook (1986) proposed the local influence technique, widely explored in recent studies. De Bastiani et al. (2015) advanced in this field by developing widespread Zhu disturbance, while Uribe-Opazo et al. (2023) used diagnostic techniques to identify influential points, analyzing their impact on the response variable and on the construction of thematic maps with kriging.
One of the problems in spatial data analysis is the presence of *outliers.* In the literature it is known that an *outlier is* an atypical value that escapes the patterns and can cause anomalies in the results obtained if it is not controlled. Understanding *outliers* is fundamental in an analysis, because *outliers can* negatively experience all the results of a spatial analysis or the behavior of *outliers* can be precisely what is being sought (creation of management zones). The central question is: what to do with them? In the literature, the use of data transformation is recommended, such as Box-Cox (Box and Cox, 1964), which aims to normalize data distribution and reduce the impact of extreme values.
Given this scenario, this study investigates the influence of outliers and atypical observations on asymmetric data, besides using diagnostic techniques of local influence in geostatistical models to explore the spatial dependence of soybean productivity and soil attributes. The results show that the identification and exclusion of these influential observations alter not only the estimates of the parameters and the forecasts of the models, but also the reliability of the thematic maps generated by kriging. By combining geostatistical methods with analysis of local influence, this work highlights the relevance of a thorough analysis for assertive decision making in precision agriculture, promoting a balance among productive efficiency, economic viability and environmental preservation.

Results and discussion

Exploratory analysis

Data were analyzed on soybean yield data (Prod)[$t\ ha^{-1}$], of the chemical contents in the soil of: Potassium (K)[$cmolc\ dm^{-3}$], phosphorus (P)[$mg\ dm^{-3}$], pH (pH), soil resistance to penetration in layers 0.0 to 0.10m ($RSP_{0.0-0.10m}$) [MPa] and 0.31 to 0.40m ($RSP_{0.31-0.40m}$) [MPa] depth were selected due to its relevance for soybean crop development and its asymmetric data behavior. The factors such as nutrient availability, soil acidity and soil resistance to penetration – directly influence crop growth and productivity (Vanderhasselt et al., 2023). The selection of these variables aims to provide a comprehensive analysis of the main elements that impact soybean performance.
The descriptive analysis of the variables considered in the study is presented in Table 1. The average soybean yield in the monitored area was 1.534 $t\ ha^{-1}$, with a coefficient of variation of 39.59%, indicating a moderate variability in the data. The third quartile, with a value of 1.989 $t\ ha^{-1}$, indicates that 75% of the data are below this limit, while the maximum value observed, 2.909 $t\ ha^{-1}$. These results suggest specific challenges for the study area, because it presents productivity values ranging from 0.331 $t\ ha^{-1}$ to 2.909 $t\ ha^{-1}$ that may be related to edaphoclimatic or agricultural management factors.
The average potassium (0.76 $cmolc\ dm^{-3}$) and phosphorus (19.14 $mg\ dm^{-3}$) levels were classified as very high, according to the criteria of Santos e Silva (2001). In contrast, the average soil pH value (5.82) is considered adequate, indicating favorable chemical conditions for soybean crop, according to the same authors.
Regarding soil resistance to penetration, the layer of 0.0 to 0.10 m deep ($RSP_{0.0-0.10m}$) presented an average of 1.554 MPa, indicating low compaction level and little limitation to root development. The average in the layer from 0.31 to 0.40m ($RSP_{0.31-0.40m}$) was 0.774 MPa, considered very low, without restrictions to root growth, according to the Canarache criteria (1990).
Data distribution was analyzed using boxplots (Figure 2), histogram and density (Figure 3), normality test and asymmetry coefficients (Table 1) and the Isolation Forest method (Figure 2) (Liu, Ting, Zhou, 2008), which identified the presence of outliers in all analyzed variables (Table 4). The variables normality was evaluated by the Shapiro-Wilk test, whose p-value was less than 0.05 for all variables, indicating the rejection of the normality hypothesis. To correct the effects of asymmetry and approximate the data of a normal distribution, Box-Cox transformation (1964) was applied, with a specific transformation parameter for each variable, as described in Table 1.

Geostatistical analysis

For spatial dependence analysis, 11 lags were considered up to a distance of 880 meters (50% of the maximum distance) (Clark, 1979). The semi variogram was analyzed in the directions 0º, 45º, 90º and 135º to verify the existence of isotropy, as recommended by Guedes et al. (2013), and the results indicated that the transformed data are anisotropic, suggesting that the spatial variability of the variables under study does not have a privileged direction.
The parameters of spatial dependence structures were estimated using the maximum likelihood method. The adjusted models belong to the Matérn family, considering different values of the k smoothing parameter: 0.5 (exponential model), 0.7, 1.0, 2.0 $k \rightarrow \infty$ and (Gaussian model), in addition to the Wave model (Matérn, 1986; Silva et al, 2025a, 2025b).
The models validation was performed through cross-validation (Faraco et al., 2008) and the information criteria of Akaike (AIC) (Akaike, 1973) and Bayesian of Schwarz (BIC) (Schwarz, 1978). The Gaussian model($k \rightarrow \infty$) presented the best performance to represent the spatial variability of the variables soybean yield (Prod[$t\ ha^{-1}$]), potassium (K[$cmolc\ dm^{-3}$]) and phosphorus levels (P[$mg\ dm^{-3}$]) on the soil. The Wave model was the most suitable for soil pH (pH[$CaCl_2\ dm^{-3}$]) and soil resistance to penetration in layer 0.31 to 0.40 meters deep ($RSP_{0.31-0.40m}$ [MPa]). On the other hand, $RSP_{0.0-0.10m}$,

Table 1. Descriptive statistics of the variables under study.

Statistics	Prod	K	P	pH	$RSP_{0.0-0.10m}$	$RSP_{0.31-0.40m}$
Minimum	0.331	0.23	7.13	4.60	0.821	0.004
1^{st} Quartile	1.045	0.53	11.39	5.50	1.242	0.407
Median	1.442	0.75	16.31	5.90	1.516	0.707
Mean	1.534	0.76	19.14	5.82	1.554	0.774
3^{rd} Quartile	1.989	0.90	21.97	6.20	1.880	1.016
Maximum	2.909	1.70	62.79	6.70	2.248	3.949
SD	0.60	0.28	11.06	0.43	0.39	0.59
CV (%)	39.59	37.44	57.79	7.40	25.23	77.38
$\tilde{u}_3$	0.38	0.63	1.89	-0.49	0.24	1.85
Kur	-0.69	0.23	3.72	0.04	-0.91	6.46
p-value	0.01*	0.01*	0.00*	0.02*	0.02*	0.00 *
λ	0.46	0.35	-0.65	2,00	0.35	0.46

SD: Standard deviation; CV: coefficient of variation; $\tilde{u}_3$: coef. asymmetry; Ku: coef. Kurtosis; p-value: Descriptive level of Shapiro-Wilk normality test; * rejects normality at 5% significance; λ lambda parameter used in Box-Cox transformation; Prod: Soybean yield in harvest year 2022/2023 [$t\ ha^{-1}$]; K: potassium content [$cmolc\ dm^{-3}$]; P: phosphorus content [$mg\ dm^{-3}$]; pH: soil pH [$CaCl_2\ dm^{-3}$];$RSP_{0.0-0.10m}$: soil resistance to penetration in layer 0.0 at 0.10 meters deep [MPa]; $RSP_{0.31-0.40m}$: Soil resistance to penetration in layer 0.31 at 0.40 meters deep [MPa].

the [MPa] in layer 0.0 to 0.10 meters, the Matérn model with a smoothing parameter k = 0.7 was the most appropriate (Table 2).

The spatial dependence indices (SDI) presented in Table 2 show the variability in the degree of spatial association between the observations (Neto et al., 2020; Uribe-Opazo et al, 2023). Soybean yield (Prod) showed a strong spatial dependence (SDI > 24%, Gaussian model classification), indicating that spatial proximity plays a relevant role in the variability of this variable. The radius of spatial dependence estimated for Prod by the Gaussian model was 1.387 meters, which means that for distances less than or equal to this value, soybean yield samples are spatially correlated.

For $RSP_{0.0-0.10m}$ the SDI showed a moderate spatial dependence (6% < SDI ≤ 14%, Matérn model classification with k=0.7), with a spatial dependence radius of 344 meters. In contrast, the potassium (K) and phosphorus (P) levels showed weak spatial dependence, with SDI of 4.93% and 8.75%, respectively, both classified by the Gaussian model. The radius of spatial dependence for K was 204 meters, while for P it was 295 meters, indicating that the spatial correlation is less expressive for these variables. Soil pH, in turn, showed moderate spatial dependence (SDI = 14.36 %, Gaussian model classification), with a spatial dependence radius of 606 meters. Whereas $RSP_{0.31-0.40m}$ showed weak spatial dependence (SDI ≤ 11 608%, classification of the Wave model), with a radius of spatial dependence of 608 meters.

These results highlight the importance of spatial dependence analysis to understand the patterns of variability in the cultivation environment. The strong spatial association observed for soybean productivity demonstrates that factors related to management and edaphoclimatic conditions are spatially structured. The moderate or weak variability observed for the other variables suggests that these attributes may be influenced by local processes or intrinsic heterogeneity of the soil. This information is fundamental for the development of more accurate management strategies, aiming at the optimization of productivity and the sustainability of soybean crop.

Diagnosis of local influence

This section aimed to verify if some observations were influencing the distance from likelihood, using diagnostic techniques of local influence. The charts C_i or$L_{max}|$ versus the order of observations (Cook, 1986) were used in order to identify the influential observations. The study was carried out by applying the generalized Zhu disturbance scheme, as proposed by De Bastiani et al. (2015).

The results of the local influence analysis, presented in Figure 4, highlight the following influential observations: #17 for soybean yield, #42 for soil potassium content, #19 for soil phosphorus content, #45 for soil pH, #99 for soil penetration resistance at the 0.00 to 0.10-meter depth layer, and #94 for soil penetration resistance at the 0.31 to 0.40-meter depth layer. It is important to emphasize that, in this study, some influential observations coincided with the outliers previously identified. However, there is a conceptual distinction between the two methods: while outlier analysis seeks to identify atypical points in relation to the data distribution, influence analysis evaluates the impact of these observations on the statistical model's results. As highlighted by Uribe-Opazo et al. (2012), an outlier may not be influential, just as an influential observation is not necessarily characterized as an outlier.

To evaluate the effect of the influential observations on the spatial dependence structure and on the elaboration of thematic maps, we performed the exclusion of them from the database for each variable, followed by a new analysis of spatial variability. This methodological approach provided a more detailed understanding of the spatial distribution of variables, taking into account the influence of individual observations.

Considering this new context, without the presence of influential observations, the results are detailed in Table 2. Based on cross-validation criteria (Faraco et al., 2008), Akaike information criterion - AIC (Akaike, 1973) and Schwarz Bayesian information criterion - BIC (Schwarz, 1978), it was observed that for soybean productivity, the exclusion of the influential

Table 2. Estimated parameters of the linear spatial model, by the maximum likelihood method, asymptotic standard deviation of the parameters (in parentheses) and spatial dependence index – *SDI*.

Variables	Model	$\hat{\mu}$	$\hat{\varphi}_1$	$\hat{\varphi}_2$	$\hat{\varphi}_3$	$\hat{a}(m)$	SDI (Class)
Prod	Gaus	0.487 (0.291)	0.090 (0.115)	0.255 (0.171)	0.801 (0.198)	1,387	37.25 (strong)
Prod#17	Wave	0.563 (0.279)	0.155 (0.091)	0.236 (0.204)	0.525 (0.172)	1,571	35.49 (strong)
K	Gaus	-0.294 (0.043)	0.067 (0.017)	0.049 (0.033)	0.117 (0.043)	204	4.93 (weak)
K#42	Wave	-0.287 (0.041)	0.094 (0.018)	0.022 (0.014)	0.097 (0.014)	1,446	11.46 (moderate)
P	Gaus	1.284 (0.012)	0.002 (0.0008)	0.002 (0.001)	0.170 (0.043)	295	8.75 (weak)
P#19	Wave	1.317 (0.008)	0.002 (0.001)	0.003 (0.001)	0.062 (0.002)	555	21.19 (moderate)
pH	Wave	16.595 (0.283)	4.010 (0.936)	2.207 (0.950)	0.067 (0.004)	606	14.36 (moderate)
pH#45	Wave	16.674 (0.354)	5.000 (0.929)	1.015 (0.694)	0.145 (0.030)	1,305	9.92 (weak)
$RSP_{0.0-0.10m}$	M0.7	0.466 (0.052)	0.004 (0.014)	0.082 (0.027)	0.099 (0.033)	344	12.92 (moderate)
$RSP_{0.0-0.10m}$#99	Exp	0.457 (0.059)	0 (0.014)	0.077 (0.024)	0.127 (0.052)	382	13.73 (strong)
$RSP_{0.31-0.40m}$	Wave	-0.322 (0.121)	0.391 (0.091)	0.085 (0.062)	0.203 (0.052)	608	7.30 (weak)
$RSP_{0.31-0.40m}$#94	M0.7	0.637 (0.084)	0.018 (0.010)	0.029 (0.014)	0.474 (0.284)	1,636	21.30 (strong)

$\hat{\mu}$: mean; $\hat{\varphi}_1$: peptic effect; $\hat{\varphi}_2$: contribution; $\hat{\varphi}_3$: range function; $\hat{a}$: range; SDI: spatial dependence index; $Class$: spatial dependence classification; # x_i: indicates the removal of the influential observation from the database; M0.7: Matérn model with a smoothing parameter k = 0.7; Prod: soybean productivity[$t\ ha^{-1}$]; K: potassium content [$cmolc\ dm^{-3}$]; P: phosphorus content [$mg\ dm^{-3}$]; pH: soil pH [$CaCl_2\ dm^{-3}$]; $RSP_{0.0-0.10m}$: soil resistance to penetration in layer 0.0 to 0.10 meters deep [MPa]; $RSP_{0.31-0.40m}$: soil resistance to penetration in layer 0.31 to 0.40 meters deep [MPa].

observation # 85 [2.861 $t\ ha^{-1}$] (Prod#85) resulted in a change from the Gaussian model to the Wave model, with an increase of 184 m in the radius of spatial dependence, maintaining a strong spatial dependence according to the SDI index (Table 2).

For the potassium content in the soil, the exclusion of the influential observation #42 [0.50 $cmolc\ dm^{-3}$] (K#42) changed the selected model, from Gaussian to Wave, leading to an increase in the radius of spatial dependence, from 204 m to 1.446 m to (increase of 1.242 m), with a transition in SDI from weak to moderate (Table 2).

For phosphorus content in the soil disregarding influential observation #19 [13.76 $mg\ dm^{-3}$] (P#19), there was a change in the model selected to describe spatial variability, moving from Gaussian to Wave. Spatial dependency radius increased from 295 m to 555 m, and SDI went from weak to moderate.

In the case of soil pH, even after excluding the influential observation #45 [5.90 $CaCl_2\ dm^{-3}$] (pH#45), the Wave model remained the most appropriate. However, the spatial dependence radius increased from 606 m to 1.305 m (699 m increase), resulting in a transition in SDI from moderate to weak (Table 2).

For soil resistance to penetration in the layer from 0.0 to 0.10 m, the exclusion of the influential observation #99 [0.82 MPa] ($RSP_{0.0-0.10m}$ # 99) changed the selected model, from Matérn with smoothing k=0.7 to Matérn with smoothing parameter k=0.5 (exponential). The spatial dependence radius increased from 344 m to 382 m, and the SDI went from moderate to strong (Table 2).

Finally, for the soil resistance to penetration in the layer from 0.31 to 0.40 m, the exclusion of the influential observation #94 [2.20 MPa] ($RSP_{0.31-0.40m}$#94) resulted in the replacement of the Wave model (range of 608 m) the Matérn model with a smoothing parameter k= 0.7 , and the spatial dependence radius increased to 1.636 (1.028 m increase), with the SDI from weak to strong (Table 2).

These results reinforce the importance of evaluating the influence of individual observations on spatial modeling, highlighting how exclusion of influential points can significantly alter the parameters of the models and the characteristics of spatial dependence.

Geostatistical map and map comparison

Based on the interpolation by ordinary kriging and the models selected to describe the spatial variability of the variables, thematic maps were generated for conditions with all observations and excluding observations considered influential. The

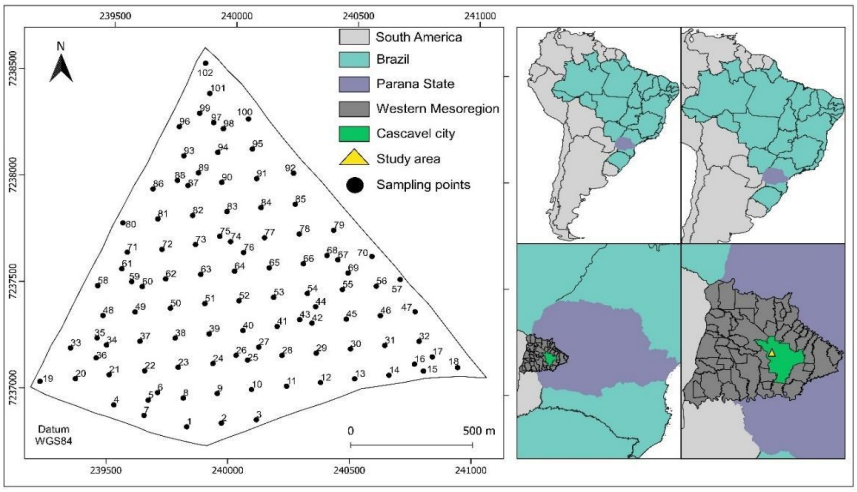

Fig 1. Location of the monitored area and the positioning of the 102 sampling points.

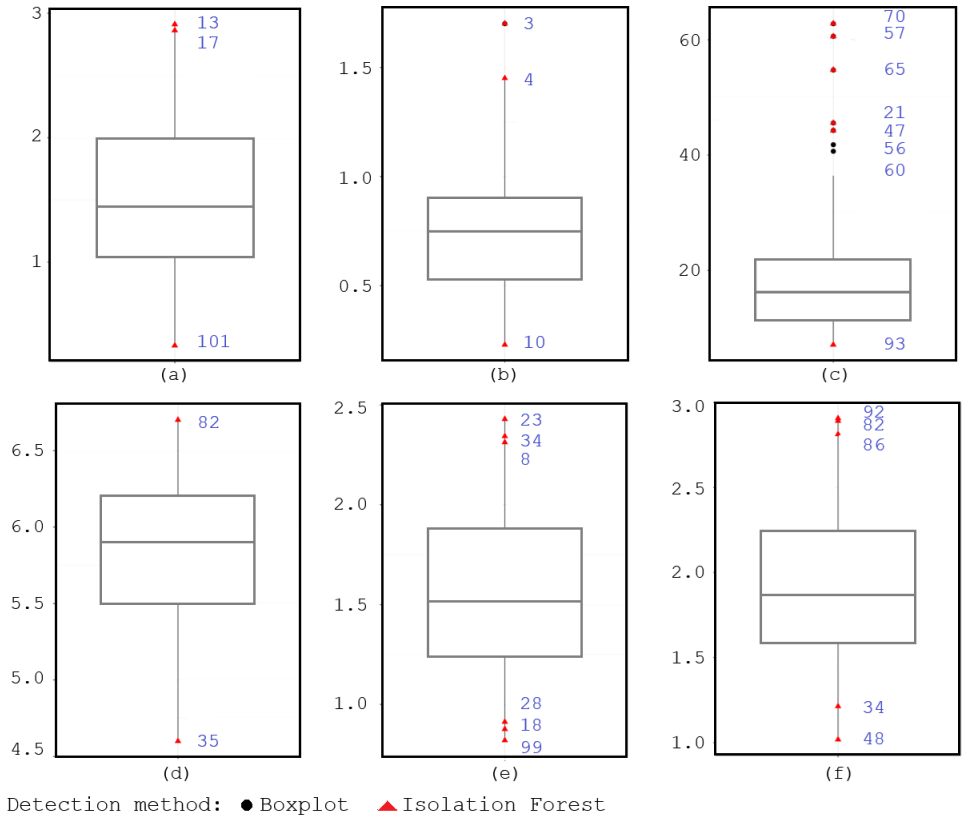

Fig 2. Detection method of outlier Boxplot and Isolation Forest for the identification of outliers in data of: (a) soybean yield; (b) potassium content; (c) phosphorus content; (d) soil pH; (e) soil resistance to penetration in layer 0.0 to 0.10 meters deep; (f) soil resistance to penetration in layer 0.31 to 0.40 meters deep.

results are presented in Figure 5, and the comparison reveals important information about the influence of these observations on the spatial distribution of the analyzed variables.

It is important to highlight that, so far, we have worked with data transformed through Box-Cox transformation in the choice of the model and estimation of parameters. Therefore, the thematic maps were generated using the Box-Cox inverse transformation, considering for each variable the λ corresponding transformation parameter. This approach ensures that the maps reflect the real values of the variables, preserving the interpretability of the observed spatial patterns.

When analyzing the soybean yield maps for the year 2022/2023, it is observed that the maps generated with all observations (Figure 5(a)) and without the influential #observation 17 (Figure (b))) show moderate similarity according to the Kappa accuracy index ($0.4 < \widehat{K} < 0.75$, Table 5). There was a considerable change in the frequency distribution of classes (Table 5), especially in the area of higher productivity, which decreased from 4.55% to 0% of the total area after exclusion of influential observation. This reduction directly impacts the estimated profitability of the area, highlighting the relevance of identifying and treating influential observations in geostatistical studies.

Table 3. Spatial Dependency Index Classification – SDI.

Model	MF	Weak	Moderate	Strong
Wave	0.58900	SDI ≤ 11%	11% < SDI ≤ 24%	SDI > 24%
Matérn k→∞ (Gaussiano)	0.50400	SDI ≤ 9%	9% < SDI ≤ 20%	SDI > 20%
Matérn k= 0.5 (Exponencial)	0.31673	SDI ≤ 6%	6% < SDI ≤ 13%	SDI > 13%
Matérn k= 0.7	0.34833	SDI ≤ 6%	6% < SDI ≤ 14%	SDI > 14%

$SDI = MF\left(\frac{\varphi_2}{\varphi_1+\varphi_2}\right) min\left\{1; \left(\frac{a}{0.5MD}\right)\right\}100$, MF: Specific factor for each model, φ_1: Pepite effect, φ_2: Contribution, a: Range, MD: Maximum distance between two sampled points.
Source: (Seidel and Oliveira, 2014; Neto et al., 2018; Uribe-Opazo et al., 2023).

Table 4. Outlier detected by the boxplot graph and isolation forest method and influential observation by local influence.

Variable	Boxplot outlier points	Isolation forest outlier points	Local influence influential points
Prod	-	13, 17*, 101	17*
K	3	3, 4, 10	42
P	3, 21*, 47*, 56, 57*, 60, 65*, 70*	21*, 47*, 57*, 65*, 70*, 93	19
pH	-	11, 35, 82	45
$RSP_{0.0-0.10m}$	-	8, 18, 23, 28, 34, 99*	99*
$RSP_{0.31-0.40m}$	-	34, 48, 82, 86, 92	94

Prod: Soybean yield; K: Potassium content; P: Phosphorus content; pH: Soil pH; $RSP_{0.0-0.10m}$: soil resistance to penetration in layer 0.0 to 0.10 meters deep; $RSP_{0.31-0.40m}$: soil resistance to penetration in layer 0.31 to 0.40 meters deep; *: identification coincident.

Table 5. Kappa index ($\widehat{K}$) for the comparison between the maps obtained with all observations and the maps obtained excluding the influential observations.

Comparison	$\widehat{K}$
Prod × Prod#17	0.74
K × K#42	0.45
P × P#19	0.67
pH × pH#45	0.24
$RSP_{0.0-0.10m} \times RSP_{0.0-0.10m}$#99	0.89
$RSP_{0.31-0.40m} \times RSP_{0.31-0.40m}$#94	0.0

Rating: $\widehat{K}$ ≥0.75 indicates high similarity between maps;0.4< $\widehat{K}$<0.75 indicates moderate similarity; $\widehat{K}$ ≤ 0.4 indicates low similarity; #x_i: indicates removal of influential observation from the database; Prod ≥: soybean yield; K: potassium; P: phosphorus; pH: soil pH;$RSP_{0.0-0.10m}$:soil resistance to penetration in layer 0.0 to 0.10 meters deep; $RSP_{0.31-0.40m}$: soil resistance to penetration in layer 0.31 to 0.40 meters deep.

Similarly, thematic maps of potassium content in the soil (K) also show considerable differences. The comparison between the map and all observations (Figure 5(c)) and the one without the influential observation #42 (Figure 5(d)) shows low similarity ($0.4 \leq \widehat{K}$, Table 5). The main change is the extinction of the class of very high levels of potassium (1.00 to 1.13 $cmolc\ dm^{-3}$), according to the classification of Santos e Silva (2001). Given the crucial role of potassium in water regulation and nutrient transport (Moreira et al., 2024), it is essential to ensure a homogeneous distribution for satisfactory performance.

The evaluation of soil pH, represented in the maps with all the observations (Figure 5(g)) and without the influential observation #45 (Figure 5 (h)), also presents low similarity ($0.4 \leq \widehat{K}$, Table 5). There was a significant reduction in the area corresponding to the pH class between 5.31 and 5.50, classified as mean, and the extinction of the area in the ideal range from 6.06 to 6.25, according to Santos e Silva (2001). Fagundes et al. (2018) emphasize that the ideal pH for soybean cultivation varies from 5.7 to 7.0, being influenced by factors such as fertilization, organic matter and soil type. Thus, adjustments in pH are fundamental for efficient agricultural production.

As for soil resistance to $RSP_{0.31-0.40m}$, the maps generated by the Wave model with all observations (Figure 5(k)) and the Matérn model with = k 0.7, excluding the influential observation #94 ($RSP_{0.31-0.40m}$ # 94) (Figure5(l)), they show substantial differences, being classified as low similarity($0.4 \leq \widehat{K}$, Table 5). In the first case, the area consists of classes from 0.45 to 1.31 MPa, while in the second case these classes are replaced by values between 1.31 and 2.61 MPa, classified as average resistance to root development (Canarache, 1990). Soil compaction is more pronounced in the regions with higher machine traffic (Keller et al., 2019; Vanderhasselt et al., 2023).

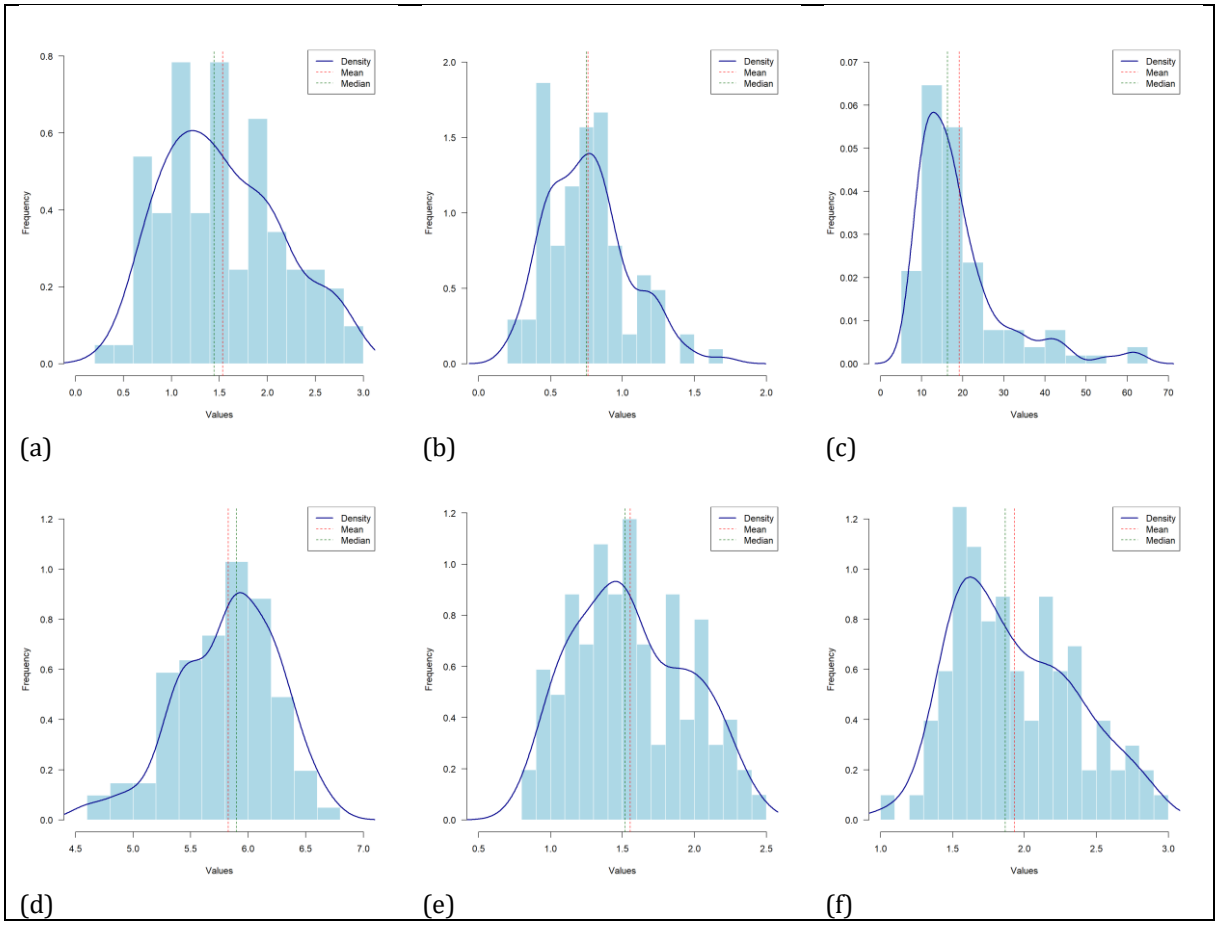

Fig 3. Frequency distribution of the sample values of the data of: (a) soybean yield; (b) potassium content; (c) phosphorus content; (d) soil pH; (e) soil resistance to penetration in layer 0.0 to 0.10 meters deep; (f) soil resistance to penetration in layer 0.31 to 0.40 meters deep.

This study identified that the northern region of the property is the most affected by soil compaction due to slightly inclined relief. This area is also the one with the highest machine traffic, especially in return maneuvers, intensifying the pressure on the ground.

Finally, the phosphorus maps (P) with all observations (Figure 5(e)) and without the influential observation #19 (P#19) (Figure 5(f))) were classified as moderate similarity ($0.4 < \widehat{K} < 0.75$, Table 5), while the maps of soil resistance to penetration in layer 0.0 to 0.10 m (Figure 5(i) and Figure 5(j)) showed high similarity($\widehat{K} \geq 0.75$, Table 5). These analyzes highlight the importance of considering influential observations and the use of appropriate geostatistical models to capture spatial variability and guide agricultural management practices with greater efficiency and sustainability.

Materials and methods

Study of the area and data

Soybean yield data (Prod) [$t\ ha^{-1}$], potassium content (K)[$cmolc\ dm^{-3}$], phosphorus content (P)[$mg\ dm^{-3}$], soil pH (pH) and soil resistance to penetration in layer 0.0 to 0.10 m($RSP_{(0.0\ a\ 0.10m)}$) [MPa] and layer 0.31 to 0.40 m($RSP_{(0.31\ a\ 0.40m)}$) [MPa] depth, were chosen for their relevance to the development of soybean crop. These variables are determinant for the establishment, growth and productivity of the crop, since the availability of nutrients, soil acidity and resistance to penetration directly influence root development, water and nutrients absorption, and consequently plant performance (Keller et al., 2019; Vanderhasselt et al., 2023; Moreira et al., 2024).

Data were collected during the 2022/2023 soybean harvest year in a commercial area of 172.04 ha, located in the municipality of Cascavel, western Paraná, Brazil. This area, is cultivated in a no-tillage system with rotation of corn and soybean crops, has geographic coordinates of approximately 24º57’18 29”S latitude, 53º34’750”W longitude, at an average altitude of 1 m (Figure). The regional climate is mesothermic and super humid temperate, climatic type Cfa (Köppen) and its soil is classified as a *typical dystroferric Red Latosol* of clayey context (Santos et al., 2018).

The 102 sampling points were defined by means of a *lattice plus close* pairs sampling (Diggle and Ribeiro Jr., 2007; Chipeta et al., 2017). The soil chemical attributes were collected at a layer from 0.0 to 0.20 meters, for each sampling point, three subsamples were randomly collected in a radius of 4 meters, allowing a representative and homogeneous final sample. Soil resistance to penetration (RSP) was measured with the penetrometer penetroLOG – PLG 2040 Falker brand up to 0.40 meters deep and soybean yield data (Prod) were collected manually. All samples were georeferenced using GPS in an UTM spatial coordinate system.

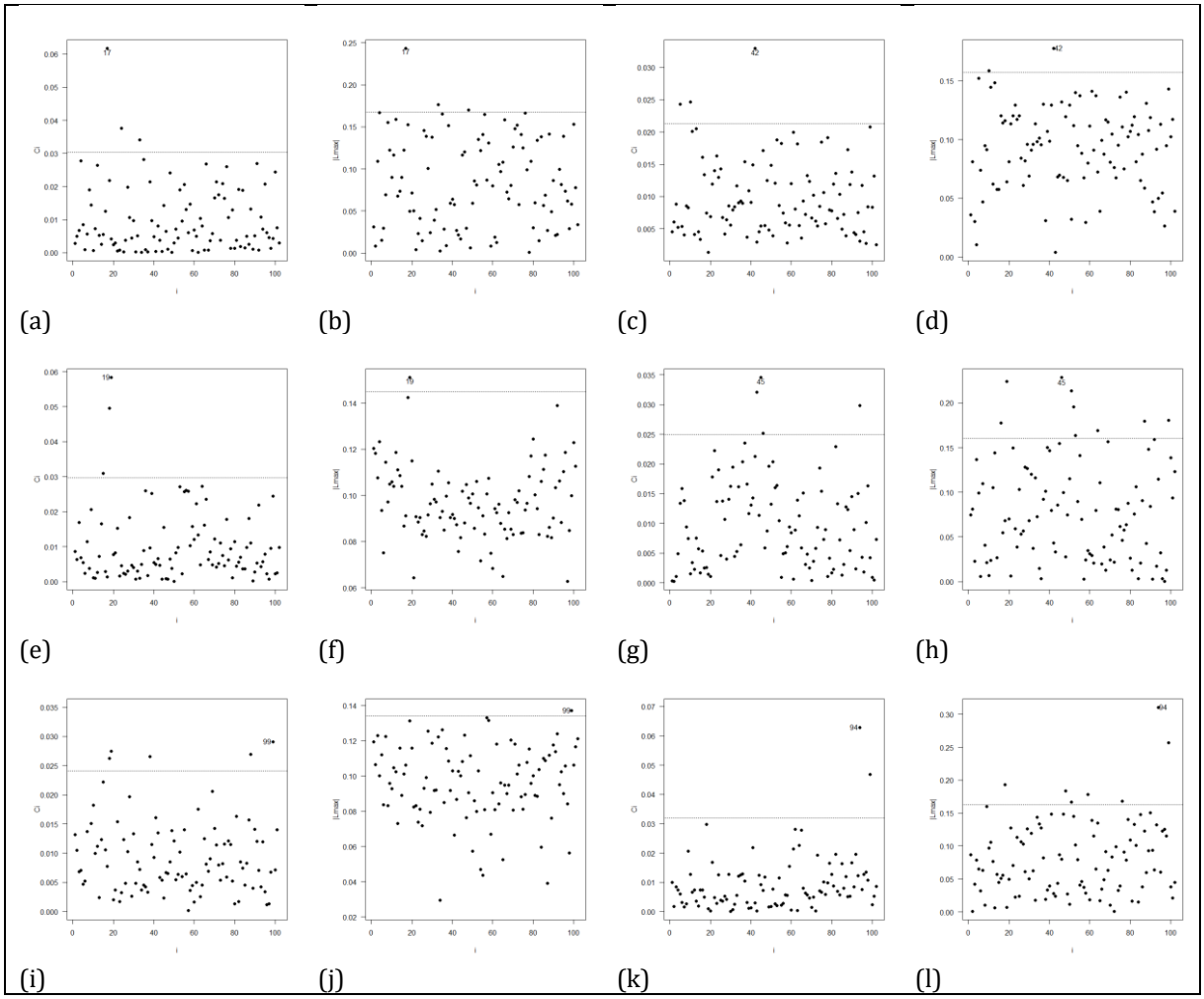

Fig 4. Charts of local influence C_i and $|L_{max}|$ according to the order of observations collected: (a) C_i and (b) $|L_{max}|$ for soybean productivity; (c) C_i and (d) $|L_{max}|$ for potassium content; (e) C_i and (f) $|L_{max}|$ for phosphorus content; (g) C_i and (h) $|L_{max}|$ for pH; (i) C_i and (j) $|L_{max}|$ for soil resistance to penetration in layer 0.0 to 0.10 meters; (k) C_i and (l) $|L_{max}|$ for soil resistance to penetration in layer 0.31 to 0.40 meters.

Exploratory analysis

Descriptive analyzes were performed, which included the calculation of position, dispersion and form measurements. The data normality was evaluated by Shapiro-Wilk test, serving as a decisive tool to determine the need for the data adjustment, ensuring its adequacy to the adjustment of geostatistical models. In the absence of normality, the data were submitted to a transformation using the Box and Cox method (1964).

This transformation aims to correct asymmetries and adjust the data to the premises of geostatistical models. To complement exploratory analysis, boxplot charts were used, which allowed to identify patterns and behaviors of sampling points. In addition, the Isolation Forest algorithm (Liu, Ting, & Zhou, 2008) was applied for outlier detection. This method, based on unsupervised learning principles, isolates anomalous observations through recursive random splits in feature subspaces. The fewer splits required to isolate an observation, the higher its probability of being classified as an outlier, given its deviation from the predominant data distribution.

Geostatistical analysis

To model the spatial dependence structure of a regionalized variable, a Gaussian stochastic $Z = \{Z(s), s \in S\}$ process was considered where $s = (x, y)^{\top}$ represents a specific location in the study area $S \subset R^2$, where R^2 is the two-dimensional euclidean space. It is assumed that the data $Z = \big(Z(s_1), \dots, Z(s_n)\big)^{\top}$ constitute a Gaussian stochastic process is stationary of second order and isotropic, collected in known locations $(s_1, \dots, s_n) \in S \subset R^2$. This process is modeled by the set $Z = \mu(s) + \varepsilon(s)$where the deterministic term $\mu(s) = \mu 1$ is a vector $n \times 1$ of the process averages $Z(s)$, and μ is an unknown parameter to estimate and **1** a unit vector, $\varepsilon = \big(\varepsilon(s_1), \dots, \varepsilon(s_n)\big)^{\top}$ represents the random error vector $n \times 1$, , with normal n-varied distribution, where,$E[\varepsilon(s)] = 0$ and covariance matrix Σ, of dimension $n \times n$, defined as $\Sigma = \Sigma\big[(\sigma_{ij})\big] = C\big(s_i, s_j\big), i, j = 1, \dots, n$. The covariance matrix Σ is symmetrical and defined positive, with elements $C\big(s_i, s_j\big)$ that depend on the Euclidean distance $d_{ij} = \|s_i - s_j\|$ between points s_i and s_j, being sometimes denoted by $C\big(d_{ij}\big)$ or $C(d)$. The structure of the matrix Σ is influenced by the parameters $\varphi = (\varphi_1, \dots, \varphi_s)^{\top}$ as established by Equation (2) (Uribe-Opazo *et* al., 2012):

$$\Sigma = \varphi_1 I_n + \varphi_2 R(\varphi_3), \qquad (2)$$

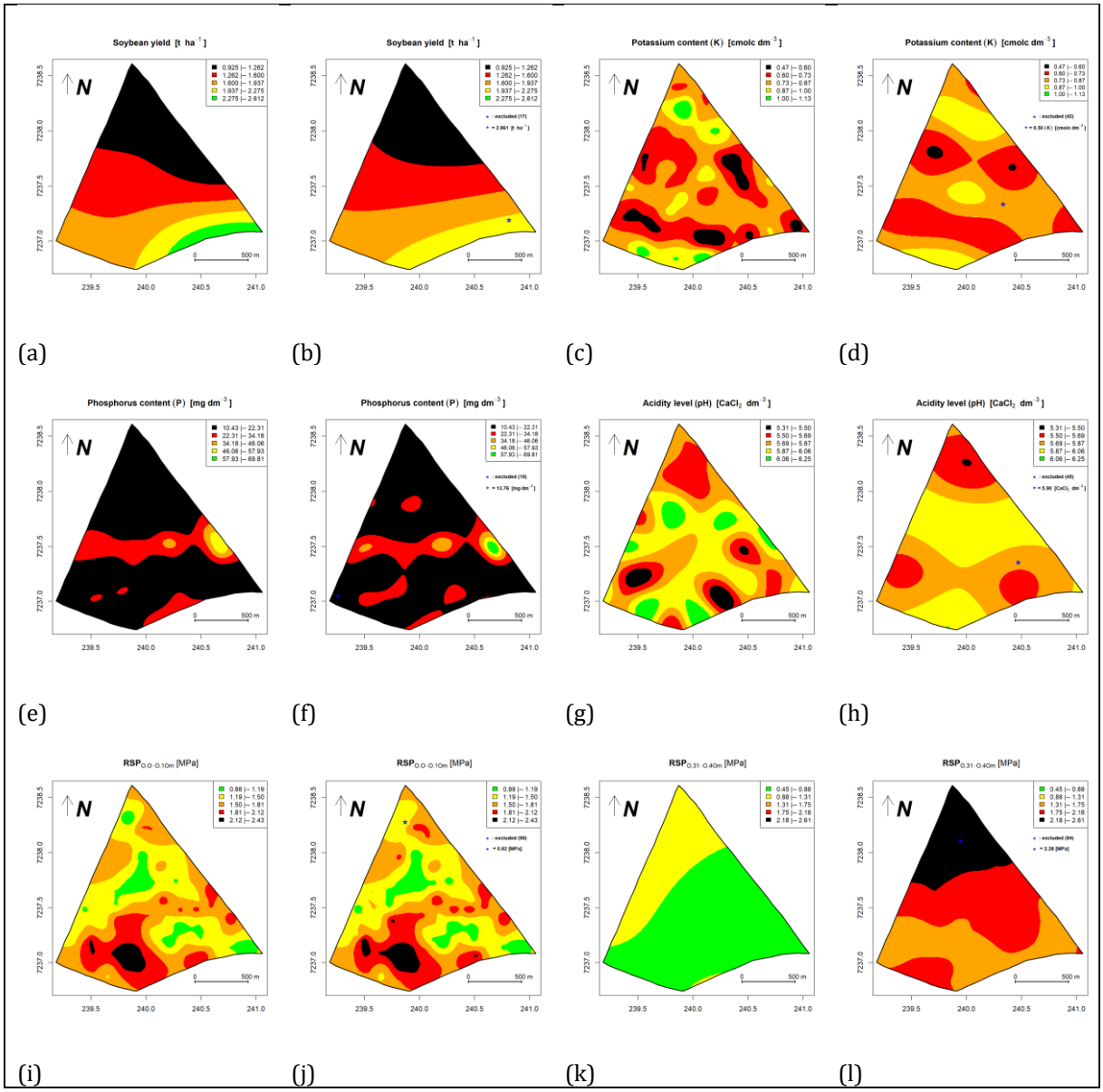

Fig 5. Thematic map interpolated by kriging of: (a) soybean productivity; (b) soybean productivity without the #17 influential observation; (c) potassium content; (d) potassium content without the #42 influential observation; (e) phosphorus content; (f) phosphorus content without the # 19 influential observation; (g) soil pH; (h) soil pH without #45 influential observation; (i) soil resistance to penetration in layer 0.00 to 0.10 meters; (j) soil resistance to penetration in layer 0.0 to 0.10 m without #99 influential observation; (k) soil resistance to penetration in layer 0.31 to 0.40 meters; (l) soil resistance to penetration in layer 0.31 to 0.40 meters without #94 influential observation.

where, $\varphi_1 \geq 0$ it is known as peptic effect ;$\varphi_2 \geq 0$ as contribution;$R(\varphi_3) = \left[(r_{ij})\right]$ is a symmetric matrix $n \times n$, depending on $\varphi_3 > 0$, with elements diagonally $r_{ii} = 1\ j = 1, \ldots, n$, ;$r_{ij} = \varphi_2^{-1} C(s_i, s_j)\ \varphi_2 \neq 0$ to $r_{ij} = 0$ and $\varphi_2 = 0$to , $i \neq j = 1, \ldots, n$, where r_{ij} is dependent on d_{ij};φ_3 it is determined by the range model $(a = g(\varphi_3))$.

To investigate the spatial dependence structure, semi variograms were constructed using the Matheron semi variance function estimators (equation 3) (Cressie, 2015).

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \left[\left(Z(s_i) - Z(s_i + h) \right)^2 \right], \qquad (3)$$

where, $\hat{\gamma}(h)$ it is the estimator of the Matheron semivariance function; $N(h)$is the number of pairs of values sampled in locations separated by distance h ;$Z(s_i + h)$ and $Z(s_i)$ are the values of the variable Z in points $s_i + h$, and s_i , respectively. For a detailed analysis of spatial dependence, 11 gaps were defined, covering up to 880 meters ($cutoff = 0{,}5 \times MD$), that is, half the maximum distance (MD) of 1,760 meters between two sampled points, as recommended by Clark (1979). The semivariogram was examined in the directions 0º, 45º, 90º and 135º, following the guidelines of Guedes *et* al. (2013), allowing to verify potential anisotropy in the data.

For the analysis of the spatial dependence structure, the models of the Matérn family (Matérn, 1986) were evaluated, using different values for the k smoothing parameter: 0.5 (exponential), 0.7, 1.0, 2.0 and $k \rightarrow \infty$ (Gaussian) and Wave model (Olea, 2006).
The estimates of the parameters were performed using the maximum likelihood method (ML) (Mardia & Marshal, 1984), more details in Silva et al. (2025a). The selection of the ideal model was made based on cross-validation and the information criteria of Akaike (AIC) and Shwarz Bayesian (BIC) (Faraco et al., 2008).
The local influence study (Cook, 1986) used the generalized Zhu disturbance in the response variable, (De Bastiani et al., 2015). The influential point charts for the variable response to the Wave model can be consulted in Silva et al. (2025a).
The evaluation of the spatial dependence degree of the adjusted model was performed using the Spatial Dependency Index (SDI), as proposed by Seidel and Oliveira (2014). The classification is according to Table 3 (Neto et al., 2018; Uribe-Opazo et al., 2023).
The comparison of maps with all points and without the influential points was performed using the error matrix and the similarity between interpolated maps was evaluated by the estimates of Kappa accuracy indices ($\hat{K}$) (Cohen, 1960), according to Table 5.

Computational resources

All analyzes were performed in the software R (R Core Team, 2025). The Geor package was used to calculate the semi variances, adjust the models and generate the thematic maps (Ribeiro Jr & Diggle, 2001). The lambda parameter used in the Box-Cox transformation was calculated with the MASS package (Venables & Ripley, 2022). The Kappa index was calculated with the vcd package (Meyer et al, 2021) and the identification of outlier by isolation forest was performed with the isotree package (Cortes, 2025).

Conclusion

The identification of outliers, performed by the Boxplot and Isolation Forest methods, revealed the presence of asymmetry in the data, confirmed the data non-normality by the Shapiro-Wilk test. This characteristic is common in agronomic variables. To correct this asymmetry, Box-Cox transformation was applied, normalizing the data distribution and ensuring the adequacy of geostatistical models and reliability of estimates. After the analysis, the inverse transformation was used to return to the original values, ensuring that the thematic maps represent the actual conditions of the field.
The methodology used was effective in detecting and analyzing influential points, showing that its removal can significantly alter the estimates of the parameters that define the spatial dependence structure. These changes directly impact the construction of thematic maps, as demonstrated by the exclusion of the influential observation #17, which affected both the areas of high productivity of soybean and the general spatial patterns. The area of higher soybean productivity, for example, was reduced from 4.55% to 0%, reinforcing the importance of a careful analysis of influential observations.
In addition, the results highlighted considerable differences between the maps generated with and without the inclusion of influential points, ranging from low to high similarity. An emblematic case was the resistance of soil to penetration in the layer from 0.31 to 0.40 meters deep, where the $\hat{K}$ accuracy index indicated low similarity between the maps. This finding reinforces the need for thematic maps to faithfully reflect the real conditions observed in the field, since they are decisive tools for resource allocation and strategic planning by producers.
Therefore, the analysis of influence of observations is an indispensable component in geostatistical studies, ensuring that thematic maps accurately represent spatial variability and serve as a reliable basis for informed decisions in precision agriculture. The rigorous application of this approach contributes to the advancement of more efficient, sustainable and economically advantageous agricultural practices.

Acknowledgment

Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) Finance code 001; Council for Scientific and Technological Development (CNPq); Araucaria Fundation of Paraná; Post-Graduate Program in Agricultural Engineering of Western State University of Paraná (PGEAGRI - UNIOESTE); Federal Technological University of Paraná (UTFPR) and Spatial Statistics Laboratory (LEE - UNIOESTE).

References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. J R Stat Soc. 51: 469-483.
Box GEP, Cox DR (1964) An analysis of transformations. J R Stat Soc, 26(2): 211-252.
Canarache A (1990) A generelized semi-empirical model estimating soil resistence to penetration. Soil Tillage Res. 16(01): 51-70
Castaldi F, Buttafuoco G, Bertinaria F, Toscano P (2024) A geospatial approach for evaluating impact and potentiality of conservation farming for soil health improvement at regional and farm scale. Soil Tillage Res. 244: 106212.
Chi J, Song S, Zhang H, Liu Y, Zhao H, Dong L (2021) Research on the mechanism of soybean resistance to phytophthora infection using machine learning methods. Front Genet 1(2): 634635.
Chipeta MG, Terlouw DJ, Phiri KS, Diggle PJ (2017) Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure. Environmetrics. 28(1):

e2425.
Clark I (1979) Practical Geostatistics. Applied Science Publishers LTD.
Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas. 20(1): 37-46.
Cook RD (1986) Assessment of local influence. J R Stat Soc. 48(2): 133–169.
Cortes D (2025) Package "isotree": Isolation-Based Outlier Detection. R package version 0.6.1-4. https://doi.org/10.32614/CRAN.package.isotree
Cressie N (2015) Statistics for spatial data. New York: John Wiley and Sons. 928 p.
Dalposso GH, Uribe-Opazo MA, Johann JA, Galea M, De Bastiani F (2018) Gaussian spatial linear of model soybean yield using bootstrap methods. Eng Agric. 38(1): 110-116.
Dalposso GH, Uribe-Opazo MA, De Bastiani F (2021) Spatial-temporal Analysis of Soybean Productivity Using Geostatistical Methods. J Agric Stud. 9(2): 283-303.
De Bastiani F, Cysneiros AHMD, Uribe-Opazo MA, Galea M (2015) Influence diagnostics in elliptical spatial linear models. Test. 24(2): 322–340.
Diggle PJ, Ribeiro Junior PJ (2007) Model based Geoestatístics. New York: Springer.
Fagundes RS, Uribe-Opazo MA, Guedes LPC, Galea M (2018) Slash spatial linear modeling: soybean yield variability as a function of soil chemical properties. Rev Bras Cienc Solo. 42: e0170030.
Faraco MA, Uribe-Opazo MA, Silva EAA, Johann JA, Borssoi JA (2008) Seleção de modelos de variabilidade espacial para a elaboração de mapas temáticos de atributos físicos do solo e produtividade da soja. Rev Bras Cienc Solo. 32(02): 463-476.
Guedes LPC, Uribe-Opazo MA, Ribeiro Jr PJ (2013) Influence of incorporating geometric anisotropy on the construction of thematic maps of simulated data and chemical attributes of soil. Chil J Agric Res. 73(4): 414–423.
IPCC, 2019. Intergovernmental Panel on Climate Change (IPCC). Cambridge University Press, Cambridge, UK and New York, NY, USA. https://doi.org/10.1017/9781009157988.
Keller T, Sandin M, Colombi T, Horn R, Or D (2019) Historical increase in agricultural Machinery weights enhanced soil stress levels and adversely affected soil functioning. Soil Tillage Res. 194: 104293.
Mardia KV, Marshal RJ (1984) Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression. Biometrika. 71(01): 135-146.
Masino A, Rugeroni P, Borrás L, Rotundo JL (2018) Spatial and temporal plant-to-plant variability effects on soybean yield. Eur J Agron. 98: 14-24.
Matérn B (1986) Spatial Variation, 2nd edition. Lecture Notes in Statistics. nº 36. Springer.
Monteiro A, Miranda C, Trindade H (2021) Mediterranean lupines as an alternative protein source to soybean. Biol Life Sci Forum. 3(01): 38.
Meyer D, Zeileis A, Hornik K, Gerber F, Friendly M (2021) Package "vcd": Visualizing Categorical Data. R package version, 1:4-8. https://doi.org/10.32614/CRAN.package.vcd
Moreira LA, Migliavacca RA, Albrecht AJP, Luchese AV, Marino IB, Silva LC, Souza MS, Beck DV, Lambreht TK (2024) Deficiências nutricionais na cultura da soja: guia prático para identificação dos sintomas. Centro de Energia Nuclear na Agricultura / USP.
Neto EA, Barbosa IC, Seidel EJ, Oliveira MS de (2018) Spatial dependence index for cubic, pentaspherical and wave semivariogram models. Bul Geod Sci. 24(01): 142-151.
Olea RA (2006) A six-step practical approach to semivariogram modeling. Stoch Environ Res Risk Assess. 20(5): 307-318.
Pimentel-Gomes F (2009) Curso de estatística experimental. FEALQ.
Ribeiro Jr PJ, Diggle PJ (2001) geoR: Analysis of Geostatistical Data. R package 1(2): 15–18.
R Core Team (2025) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org>
Santos DR, Silva, LS (2001) Fertilidade do solo e nutrição de plantas. Santa Maria: Universidade Federal de Santa Catarina, 20p.
Santos HG, Jacomini PKT, Anjos LHC, Oliveira VA, Lumbreras JF, Coelho MR, Almeida JA, Araújo Filho JC, Oliveira JB, Cunha TJF (2018) Sistema Brasileiro de classificação de solos. Brasília, DF: Embrapa.
Schwarz G (1978) Estimating the dimensiono f a model. The Annals of Stat. 6(2): 461-464.
Seidel EJ, Oliveira MS (2014) Novo índice geoestatístico para a mensuração da dependência espacial. Rev Bras Cienc Solo. 38: 699-705.
Silva ALG, Uribe-Opazo MA, Dalposso GH, Guedes LPC (2025a) Spatial variability of soybean productivity and soil attributes: "hole effect" and local diagnosis with the Wave model. Aust J Crop Sci. 19(1): 44-51.
Silva ALG, Uribe-Opazo MA, Dalposso GH, Guedes LPC, Maltauro TC (2025b) Analysis of Spatial Dependence Using the Wave Covariance Structure in Soybean Productivity Associated with Soil Attributes. Revista de Gestão Social e Ambiental. 19(1): e010971.
Uribe-Opazo MA, Borssoi J, Galea M (2012) Influence diagnostics in Gaussian spatial linear models. J Appl Stat. 39(3): 615–630.
Uribe-Opazo MA, De Bastiani F, Galea M, Schemmer R, Assumpção RAB (2021) Influence diagnostics on a reparametrized *t*-Student spatial linear model. Spat Stat. 41: 100481.
Uribe-Opazo MA, Dalposso GH, Galea M, Johann JA, De Bastiani F, Moyano ENC, Grzegozewski DM (2023) Spatial variability of wheat yield using the gaussian spatial linear model. Aust J Crop Sci. 17(2): 179-189.

Vanderhasselt A, Cool S, D'House T, Cornelis W (2023) How tine characteristics of subsoilers affect fuel consumption, penetration resistance and potato yield of a sandy loam soil. Soil Tillage Res. 228: 105631.
Venable WN, Ripley BD (2022) Modern Applied Statistics with S. Springer.
Zain M, Ma H, Rahman SU, Nuruzzaman M, Chaudhary S, Azeem I, Mehmood F, Duan A, Sun C (2024) Nanotechnology in precision agriculture: Advancing towards sustainable crop production. Plant Physiol Bioch. 206: 108244.
Zhu Z, Liu Y, Cong W, Zhao X, Janaun J, Wei T (2021) Soybean biodiesel production using synergistic CaO/Ag nano catalyst: Process optimization, kinetic study, and economic evaluation. Ind Crops Prod. 166: 113479.