Australian Journal of Crop Science AJCS

AJCS 17(2):179-189 (2023) ISSN:1835-2707
doi: 10.21475/ajcs.23.17.02.p3742

Spatial variability of wheat yield using the gaussian spatial linear model

Miguel Angel Uribe-Opazo[1*], Gustavo Henrique Dalposso[2], Manuel Galea[3], Jerry Adriani Johann[1], Fernanda De Bastiani[4], Emma Norma Cambillo Moyano[5], Denise Maria Grzegozewski[1]

[1]CCET/PGEAGRI, Western Paraná State University – (UNIOESTE), Cascavel, Paraná, Brazil
[2]PPGBio, Federal University of Technology – Paraná (UTFPR), Toledo, Paraná, Brazil
[3]Department of Statistics, Pontifical Catholic University of Chile – (UC), Santiago, Chile
[4]Department of Statistics, Federal University of Pernambuco – (UFPE), Recife, Pernambuco, Brazil
[5]Department of Statistics, National University of San Marcos – (UNMSM), Lima, Perú

***Corresponding author: miguel.opazo@unioeste.br**

Abstract

Wheat production has grown over the years and is one of the most important grain food sources for humans. This work analyzed the yield of two varieties of wheat planted in a regular sampling grid in an experimental area in the south of Brazil, using some explanatory variables. For the study of the spatial variability of wheat yield of the COODETEC 101 (CD101) and COODETEC 103 (CD103) varieties, which were cultivated by the farmer in an area of 22.62 ha, 84 samples were defined considering a regular grid of 50 x 50 m. In the sampled sites, the following explanatory variables were collected: average plant height in 60 days - avheight 60 (cm), the average number of tillers in 60 days - avtillers60 (cm), spike length in 120 days - splength (cm) and the wheat variety considered as a dummy variable (CD101 = 0 and CD103= 1). The wheat yield was analyzed using gaussian spatial linear models with different geostatistical models for the parametric form of the variance-covariance matrix. The significance of the parameters to select the explanatory variables were determined by the likelihood ratio test, and also a hypothesis test was presented to confirm that a model that deal with the spatial dependence was required by the data. To assess the global and local influence of some observations, diagnostics techniques based on Cook's approach were considered.
The disregard of potentially influential observations caused changes in the parameters estimates that define the spatial dependence structure, and consequently then in the profitability in sectors of the wheat yield maps. The study of statistical inference and diagnostics on spatial data should be part of all geostatistical analysis.

Keywords: diagnostics; geostatistics; maximum likelihood; yield map.
Abbreviations: CD101_wheat variety COODETEC 101; CD103_wheat variety COODETEC 103; CVA_cross-validation; GSLM_gaussian spatial linear models; LMV_log-likelihood maximum value; LR_likelihood ratio statistic; ML_maximum likelihood; SDI_spatial dependence measure; T_r_trace of the asymptotic covariance matrix of an estimated mean.

Introduction

Wheat is a popular source of animal feed, particularly in years where harvests are adversely affected by rain and significant quantities of the grain are made unsuitable for food use. According to Shewry (2007) it is considered a good source of protein, minerals, B-group vitamins, and dietary fiber. Wheat accounts for about 20 % of the world's total cereal energy and protein, highlighting the importance of wheat production to safeguard global food supply and mitigate powerful greenhouse gases emissions (Ma et al., 2022).
In 2017, China, India, Russia and United States produced 40.5% (366 million tons) of the wheat in the world, i.e., around 98 million of ha (Faostat, 2019). Brazil cultivated around 1.90 million of ha and produced 4.32 million of wheat during this same year, where Paraná is responsible for 53.5% of the production (2.31 million tons) and 49.1% (0.93 million ha) of the area in Brazil, followed by Rio Grande do Sul with 27.56% of the production and 36.4% of the area,

both belonging to the South of Brazil (Ibge, 2019).

Between 1974 and 1987, Brazil had an increasing annual rate of 7.4% of production, totaling the production of 6.03 million tons in 1987. Then, that production decreased to 1.53 million tons until 1995, but recovered the production until 2003, achieving 6.15 million tons. Oscillation in the wheat production has been recorded, and in 2016, Brazil reached the maximum wheat production in history (6.83 million tons in 2.17 million ha) (Ibge, 2019). Brazil's wheat production does not meet the domestic consumption, and National Supply Company - CONAB projections estimated that 2019 should consume around 11 million tons, that is, in comparison to what was produced in 2017, a deficit of 6.68 million tons.

Therefore, it is important the insertion of new technologies of crop management, as advocated by precision agriculture, in which the farming areas, need to be managed according to their soil characteristics and fertility. Since this management is usually implemented through sampling methods, it is necessary to use appropriate statistical methods, such as the ones developed within the geostatistical framework, so that the spatial variability of the collected data can be adequately addressed.

Geostatistics allows the construction of maps that show the spatial variability. It is crucial in precision agriculture, because it has the principle that the sample elements of a regionalized variable are correlated up to a distance and, it has some influence on the closest points unsampled and to be predicted (Cressie, 2015).

Geostatistics offers a way of describing the spatial continuity of natural phenomena and provides adaptations of classical regression techniques to take advantage of this continuity. It studies a response variable (and potentially explanatory variables) that are measured at points in space. Important work by Krige and Matheron laid the foundation for the field of geostatistics where some of the first methods for modelling spatial dependence were proposed, see Cressie (2015), for more details. To estimate the values at unsampled sites, a technique called kriging can be used. There are dozens of kriging methods, that are distinguished by the assumption regarding the spatial trend model, by the data transformation and by the use of auxiliary variables in the prediction. Ameer et al. (2022) modeled wheat yield in a 10 ha area in Pakistan using ordinary kriging, Dalposso et al. (2021) modeled soybean yield using kriging with external drift and Jurado-Expósito et al. (2021) used indicator kriging to generate probability maps for risk assessment when implementing weed control in wheat fields.

In geostatistics, an atypical observation can cause changes in environmental and geological patterns. Influential points may change the parameters estimates and/or the statistical inference. Cook's distance (Cook, 1977) is the more traditional measure to detect influential observations. However, the greater the number of sampling points, the greater the computational costs required. To assess the effect of small perturbations in the model (or data) on the parameter estimates, Cook (1986) proposed an interesting method, named local influence. This analysis does not involve recomputing the parameter estimates for each case deletion, so it is often computationally simpler.

Works using geoestatistics are present in the literature. Sidorova et al. (2012) performed a geostatistical analysis of the spatial variability of the soil properties, the sowing parameters, and the wheat yield in an experimental field under precision agriculture conditions, where the spherical and exponential geostatistical model were selected. Yuan et al. (2022) investigated the spatial variability of soil chemical attributes, apparent soil electrical conductivity and wheat productivity for optimization of management zone delineation for precision crop management in an intensive farming system. In the study of diagnostics for georeferenced variables, Militino et al. (2006) studied methods of global influence based on the elimination of one or more observations to vary the effect in the parameters estimates. Uribe-Opazo et al. (2012) discussed global and local methods of influence to verify the sensibility of the maximum likelihood (ML) and restricted maximum likelihood methods in Gaussian spatial linear models (GSLM). De Bastiani et al. (2015) developed local influence techniques for elliptical spatial linear models considering the appropriate scheme of perturbation. Uribe-Opazo et al. (2021) carried out studies with the reparametrized t-student distribution and showed more robust results, whose model estimates are less sensitive to outliers.

Regarding the spatial dependence measure (*SDI*), it is highlighted in this study its use in the work by Guedes et al. (2020) who investigated the nugget effect influence on spatial variability of agricultural data and in the work of Dalposso et al. (2021) who performed a spatial-temporal analysis of soybean productivity.

This paper aimed to analyze the wheat yield data(response variable) in function four explanatory variables (average plant height in 60 days, average number of tillers in 60 days, spike length in 120 days and wheat variety) in an agricultural area in southern Paraná, in Brazil. To the study of the spatial variability of wheat yield the gaussian spatial linear model (GSLM) was used. Diagnostic techniques were used to detect influential points in the response variable, highlighting observations that might influence in the parameter estimations, the predicted values by the model, and in the construction of the map of the wheat yield by kriging with external drift.

Results and Discussions

The descriptive analysis of the response variable, wheat yield and the explanatory variables are shown in Table 3. The wheat yield mean is 3.37 t ha^{-1}. The mean wheat yield of CD101 was greater than that of CD103 by 0.208 t ha^{-1}. The average plant height in 60

Table 1. Special cases of the Matérn Family covariance function.

smooth parameter ϕ_4	covariance function	model
$\phi_4 = 0.5$	$C(d_{uv}) = \phi_2\, exp(-d_{uv}/\phi_3)$	Exponential
$\phi_4 = 1$	$C(d_{uv}) = = \phi_2 (d_{uv}/\phi_3)\, K_{\phi_4}(d_{uv}/\phi_3)$	Whittle
$\phi_4 \to \infty$	$C(d_{uv}) = \phi_2\, exp(-(d_{uv}/\phi_3)^2)$	Gaussian

K_{ϕ_4} is the modified Bessel function of third type of order ϕ_4, witch $\phi_4 > 0$.

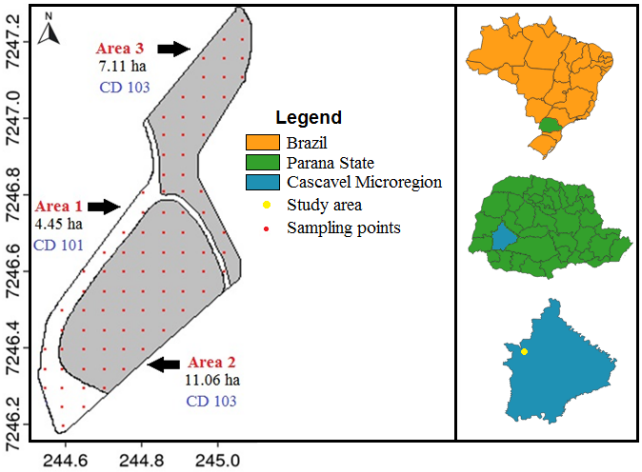

Fig 1. Sampling scheme in a total area of 22.62 hectares, with a regular grid of 50 x 50 m, sited in Cascavel, Paraná, Brazil.

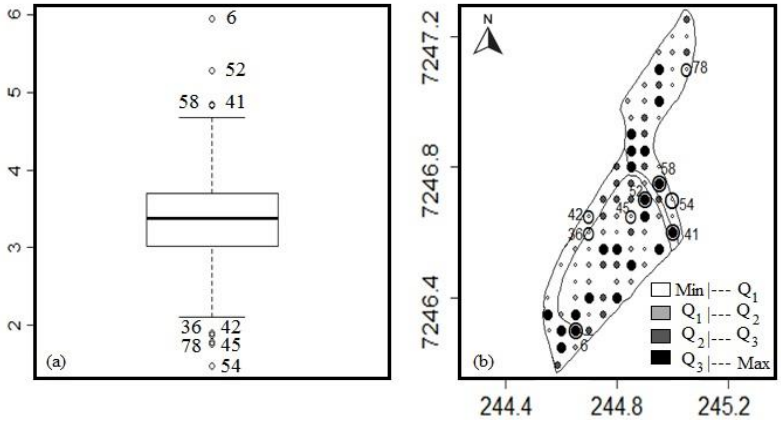

Fig 2. (a) Boxplot plot for the identification of outliers in wheat yield data (b) Postplot plot indicating the location of sample points classified by quartiles and outliers.

days (avheight60) after sowing varies from 13.4 cm to 36.6 cm, values lower than those found by Patel et al. (2019) under organically managed soils (39.20 cm) and under inorganically managed soils (44.51 cm). The average number of tillers in 60 days (avtillers60) presents the greatest value for the variance coefficient; however it still can be considered homogeneous. Singh et al. (2015) also observed a variation in the number of tillers in their experiment and, as demonstrated by Gill et al. (2022), this number can be increased by the application of both NPK and organic fertilizer. The mean and median of the spike length in 120 days (splength) are very similar, and, in general, the length measurements were lower than those obtained by Upadhyay and Kaur (2019), who in the same 120 days obtained measurements in the range (7.63 to 9.35).

Fig. 2(a) presents the boxplot for wheat yield where the observations 06, 36, 41, 42, 45, 52, 54, 58 and 78 are outliers with wheat yield values of 5.95, 1.90, 4.85, 1.88, 1.76, 5.28, 1.48, 4.83 and 1.78 t ha^{-1}, respectively. The sites of these observations are highlighted in Fig. 2(b).

The spatial linear model for the wheat yield - wheat at site $\boldsymbol{s}_i$, considering the explanatory variables average plant height (avheight60) and average number of tillers (avtillers60) in 60 days, spike length (splength) in 120 days and the wheat variety treated as a *dummy* variable (0 or 1), 0 if the variety is CD101 and 1 if it is

Table 2. The spatial dependence measure *SDI* classification for Matérn Family model with different smoothing ϕ_4.

Smooth parameter ϕ_4	*MF*	Weak	Moderate	Strong
0.5	0.316738	SDI ≤ 6 %	6 %< SDI ≤13 %	SDI >13 %
0.7	0.348333	SDI ≤ 6 %	6 %< SDI ≤14 %	SDI >14 %
1.0	0.379003	SDI ≤ 7 %	6%< SDI ≤15%	SDI >15 %
1.5	0.408758	SDI ≤ 7%	7%< SDI ≤16%	SDI >16 %
2.0	0.432194	SDI ≤ 8%	7%< SDI ≤17%	SDI >17 %
2.5	0.439467	SDI ≤ 8%	8%< SDI ≤18%	SDI >18 %
3.0	0,448393	SDI ≤ 8%	8%< SDI ≤18%	SDI >18 %
3.5	0.462040	SDI ≤ 8%	8%< SDI ≤18%	SDI >18 %
$\phi_4 \to \infty$	0.504000	SDI ≤ 9%	9%< SDI ≤ 20%	SDI >20%

MF: model fator.

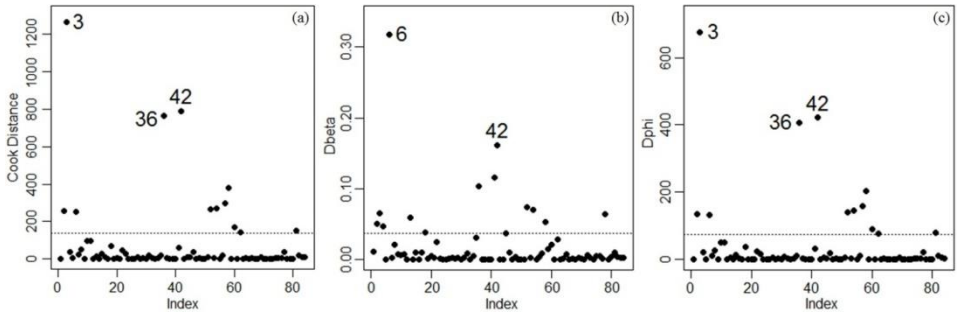

Fig 3. Identification of influential points through global influence plots (a) $D^1_{i\theta}$ versus index, (b) $D^1_{i\beta}$ versus index, and (c) $D^1_{i\phi}$ versus index.

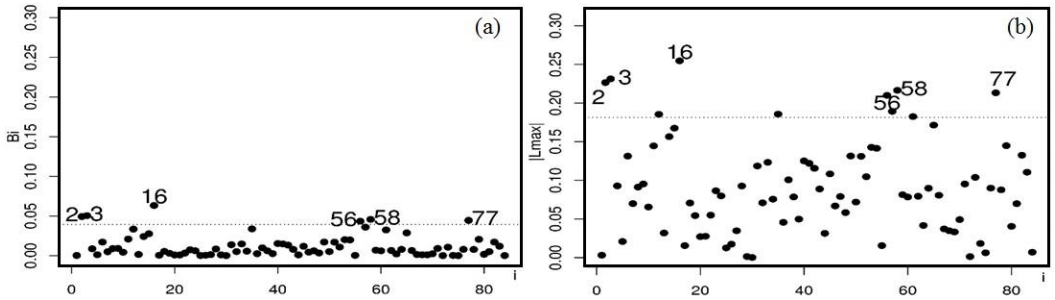

Fig 4. Identification of influential points through local influence plots (a) B_i versus index and (b) $|L_{max}|$ versus index considering the appropriate perturbation scheme.

CD103, is given by $\mu(\boldsymbol{s_i}) = \beta_0 + \beta_1 \text{dummy}(\boldsymbol{s_i}) + \beta_2 \text{avheight60}\ (\boldsymbol{s_i}) + \beta_3 \text{avtillers60}(\boldsymbol{s_i}) + \beta_4 \text{splength}(\boldsymbol{s_i})$.

For the spatial dependence analysis, 14 lags were considered until the distance 0.580 m (cutoff of 50%) (Uribe-Opazo et al., 2012). The semivariogram was checked at directions 0º, 45º, 90º and 135º and confirmed that the data are isotropic (Guedes et al., 2013).

The estimate the covariance matrix structure of Matérn family class, varying the softening parameter with values of ϕ_4 between 0.3 to ∞. It was observer that in all the fitted models the values the nugget effect of $\hat{\phi}_1$ were equal to zero and the values of the parameter $\hat{\phi}_2$ were very similar among themselves, however the $\hat{\phi}_3$ estimates (range function) vary and consequently the so did the spatial dependence distance (a = g($\hat{\phi}_3$)).

According to the criteria cross-validation (CVA), trace (T_r) and the log-likelihood maximum value (LMV) shown in Table 4, the chosen covariance matrix function is the value of $\phi_4 \to \infty$, which corresponds to the Gaussian covariance function. In a comparative study between geostatistical models, Pu et al. (2019) concluded that the Gaussian model was the most suitable for characterizing agricultural lands. However, although it is often selected as the best in automatic selection criteria, in some cases the fit may correspond to a process that is often unrealistically smooth (Abdallah, 2018). Ultimately, the final choice of model must reflect both the results of the model adjustment procedure and a coherent scientific interpretation of the variable under study.

The estimated parameters of the chosen Gaussian model ($\phi_4 \to \infty$) and their respective asymptotic standard errors (in parenthesis) are shown in Table 5. The spatial dependence radius found indicates that, for distances lower than or equal to 60.44 m, the wheat yield samples are spatially correlated. This value is lower than that found by Dalposso et al. (2012), modeled the wheat productivity in an area located in the same city as this experiment using

Table 3. Descriptive analysis of response and explanatory variables.

Statistics	Wheat total (t ha^{-1})	Variety wheat		avheight60 (cm)	avtillers60 (cm)	splength (cm)
		CD101 (t ha^{-1})	CD103 (t ha^{-1})			
Samples (n)	84	17	67	84	84	84
Minimum	1.48	1.88	1.48	13.40	0.40	5.00
1st Quartile	3.02	3.11	3.01	20.65	1.20	6.10
Median	3.37	3.49	3.31	22.50	1.70	6.45
Mean	3.37	3.53	3.33	23.17	1.66	6.47
3rd Quartile	3.70	4.05	3.66	24.62	2.10	6.80
Maximum	5.95	5.95	5.28	36.60	3.40	7.90
CV (%)	23.36	28.92	21.58	17.00	38.00	8.85

avheight60: average plant height in 60 days; avtillers60: average number of tillers in 60 days; splength: spike length in 120 days; CV: coefficient of variation.

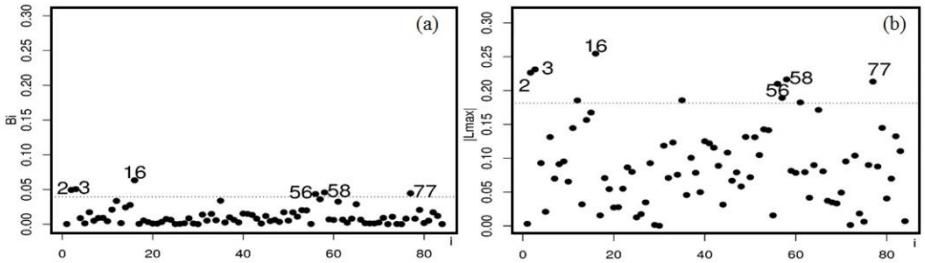

Fig 4. Identification of influential points through local influence plots (a) B_i versus index and (b) $|L_{max}|$ versus index considering the appropriate perturbation scheme.

a Gaussian model, obtained a range of 125.6 m. According to the Spatial Dependence Index - *SDI*, it is possible to conclude that there is a weak spatial dependence among the observations (*SDI* ≤ 9 %). Considering the likelihood ratio test (*LR*) presented in Equation (3), the null hypothesis $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ is rejected at a 5% level of significance, then all the explanatory variables will be considered in the final model. The number of tillers is an important contributor towards final yield (Ahmad et al., 2020), the spike length is one of the important components of grain yield formation in wheat (Mladenov et al., 2019) and the plant height is an important trait that influences the yield and sustainability of wheat productions (Jiang et al., 2020). Thus, a model that incorporates these variables can certainly provide more realistic predictions about wheat productivity.

The likelihood ratio test (*LR**) is presented in Equation (4), $H_0: \phi_2 = 0$ versus $H_1: \phi_2 > 0$. The critical value at significance level of 5% is 2.705. Since, *LR**= 3.7798, the null hypothesis was rejected at significance level of 5%, i.e., one must take account the spatial dependence structure. This is an important result of the analysis because the specification of spatial dependence structure is often used to improve the estimates precision (Sun et al., 2022). Fig. 3 presents the global influence plots. Observations #3, #6, #36, and #42 are detected as potential influential. Note that observations #6, #36 and #42 were also identified by the boxplot (Fig. 2(a)), and it is located in the Southern region of the area (Fig. 2(b)). Fig. 4 presents B_i versus index and $|L_{max}|$ versus index plots for local influence analysis. The observation #16 is detected as the most potential influence in the response variable. The observations pointed out in Fig. 4 are different from the observations highlighted in the boxplot given in Fig. 2(a). It is important to note that in spatial statistics, an influential point is not necessarily an outlier and vice versa (Leiva et al., 2020).

Considering three new scenarios, firstly removing observation #3 (C1), secondly removing only observation #6 (C2), and lastly deleting observation #16 (C3). The results are presented in Table 6. According to criteria LMV, CVA, and T_r, the chosen model for the covariance function remains the Gaussian one ($\phi_4 \to \infty$). The asymptotic standard errors estimate of the $\beta's$ estimators are very similar; however, the asymptotic standard errors estimate of the $\phi's$ estimators vary depending on the scenario. According to the *SDI*, it is possible to conclude that there is weak spatial dependence among the observations of the three scenarios (*SDI* ≤ 9%).

Fig. 5 shows the maps with all the observations and the scenarios mentioned above. The wheat yield maps constructed by kriging with external drift (Hengl et al., 2003) present well defined zones. Note that there is a slight difference among the maps in the Northern area. A difference between the varieties CD101 and CD103 was also noted.

With the information available, the model with all the observations is chosen as the final model. Table 7 shows the average profitability in dollars. It is possible to see how the deletion of points #3, #6, and #16 modify the frequency distribution of the yield areas.

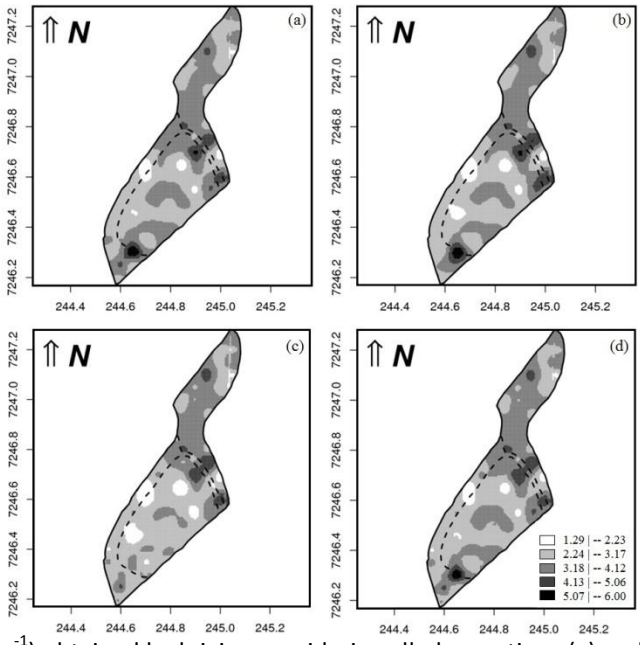

Fig 5. Maps of wheat yield (tha^{-1}) obtained by kriging considering all observations (a) and without influent observation #3 (scenario C1) (b), without influent observation #6 (scenario C2) (c) and without the influent observation #16 (scenario C3) (d).

Table 5. Parameters estimates of GSLM model by maximum likelihood considering the gaussian covariance function ($\phi_4 \to \infty$), and asymptotic standard errors in parenthesis.

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$	*a* (km)	*SDI*	Class
0.1221	0.3541	0.0691	0.0784	0.1885	0.0000	0.5481	0.0349	0.06044	5.28	weak
(1.257)	(0.266)	(0.024)	(0.142)	(0.146)	(0.4058)	(0.4281)	(0.0001)	(0.00017)		

$\beta_0, \beta_1, \beta_2, \beta_3,$ and β_4 denote the regression parameter; ϕ_1: nugget effect; ϕ_2: sill; ϕ_3 is a function range; *a* = range; *SDI*: spatial dependent measure; Class: Classification.

This caused a change in the average profitability in dollars. The greatest change was recorded by the deletion of point #6 ($20,455.35), where we have an effect on the profitability distribution, decreasing the frequency in the areas classified with high wheat yield (last three classes) relative to the analysis considering all the points, with average profitability $ 21,638.30.

Materials and methods

This section presents the description and localization of the wheat data set. To model the mean of the wheat yield, a GSLM model was used with parameters estimated by maximum likelihood method. The likelihood ratio tests were used are presented to study the significance of the parameters estimated of the explanatory variables. A hypothesis test was also presented for the parameter sill that determines the spatial dependent. Global and local diagnostics techniques are used to assess the influence of some observations and an appropriate perturbation scheme in the response variable, and finally, measurements of the spatial dependence degree to the Matérn family models were obtained using the Spatial Dependence Measure - *SDI*.

The data set

The data were collected in Cascavel, Paraná-Brazil, in southern Brazil, in an area of 22.62 hectares, whose geographic location is approximately latitude 24°52'31" S, longitude 53°31'33" W (Fig. 1). According to K¨oppen, the climate is Cfa (Embrapa, 2013), temperate mesothermal and super humid and annual precipitation mean of 1925 *mm*.

According to Embrapa (2013), the soil is of type Red Latosol, with a clayey texture. 84 element samples were collected in a regular grid of 50 x 50 m. The area was divided in three subareas: Area 1, Area 2 and Area 3 with 4.45 ha, 11.06 ha and 7.11 ha, respectively. Two wheat varieties were planted: COODETEC 101 (CD101) in Area 1, and COODETEC 103 (CD103) in Areas 2 and 3, according to the farmer interest. The explanatory variables are average plant height in 60 days - avheight60 (cm) and average number of tillers in 60 days - avtillers60 (cm), spike length in 120 days - splength (cm) and the wheat variety treated as a dummy variable (CD101= 0 and CD103= 1).

Gaussian spatial linear models – GSLM

Let $\boldsymbol{Y} = \boldsymbol{Y}(\boldsymbol{s}) = \big(Y(\boldsymbol{s}_1), \dots, Y(\boldsymbol{s}_n)\big)^T$, be an $n \times 1$ random vector of an isotropic and stationary stochastic process, that belong to the family of

Table 6. Parameters estimates by maximum likelihood considering the gaussian covariance function and asymptotic standard errors in parenthesis, considering three scenarios.

Sc	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$	*a* (km)	*SDI*	Class
C1	0.1407 (1.2157)	0.4134 (0.2632)	0.0783 (0.0230)	0.1377 (0.1478)	0.1753 (0.1380)	0.0000 (0.2093)	0.5632 (0.2587)	0.0412 (0.0001)	0.0713 (0.00017)	6.19	weak
C2	0.1221 (1.2275)	0.3541 (0.2649)	0.0691 (0.0240)	0.0784 (0.1452)	0.1885 (0.1481)	0.0000 (0.3698)	0.5481 (0.3868)	0.0319 (0.0002)	0.0552 (0.00034)	4.80	weak
C3	0.1095 (1.2685)	0.3465 (0.2674)	0.0688 (0.0243)	0.0785 (0.1425)	0.1922 (0.1482)	0.0000 (0.4725)	0.5528 (0.4922)	0.0345 (0.0002)	0.0598 (0.00034)	5.19	weak

Sc: Scenarios; C1: removing observation #3; C2: removing observation #6; C3: removing observation #16; $\beta_0, \beta_1, \beta_2, \beta_3$, and β_4 denote the regression parameter; ϕ_1: nugget effect; ϕ_2: sill; ϕ_3 is a function range; *a* = range; *SDI*: spatial dependent measure; Class: Classification.

multivariate Gaussian distributions and depend on the sites $\boldsymbol{s}_j \in \boldsymbol{S} \subset \mathbb{R}^2$, for $j = 1, \dots, n$, $\boldsymbol{s} = (\boldsymbol{s}_1, \dots, \boldsymbol{s}_n)^T$. For the wheat productivity (wheat) $\boldsymbol{Y}$ represents a vector 84 × 1. This stochastic process can be written as $\boldsymbol{Y}(\boldsymbol{s}) = \boldsymbol{\mu}(\boldsymbol{s}) + \boldsymbol{\varepsilon}(\boldsymbol{s})$, where, the deterministic term $\boldsymbol{\mu}(\boldsymbol{s})$ is an $n \times 1$ vector, the means of the process $\boldsymbol{Y}(\boldsymbol{s})$, $\boldsymbol{\varepsilon}(\boldsymbol{s})$ is an $n \times 1$ vector of a stationary process with zero mean vector, $E[\boldsymbol{\varepsilon}(\boldsymbol{s})] = \boldsymbol{0}$, and $n \times n$ covariance matrix $\boldsymbol{\Sigma} = [C(\boldsymbol{s}_u, \boldsymbol{s}_v)]$. The mean vector $\boldsymbol{\mu}(\boldsymbol{s})$ can be written as a spatial linear model by $\mu(\boldsymbol{s}) = \boldsymbol{X\beta}$, where, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is a $p \times 1$ vector of unknown parameters, $\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{s}) = [x_{j1}(\boldsymbol{s}) \dots x_{jp}(\boldsymbol{s})]$ is an $n \times p$ matrix of p explanatory variables, for $j = 1, \dots, n$. The matrix $\boldsymbol{\Sigma}$ is symmetric and positive defined, where the elements $C(\boldsymbol{s}_u, \boldsymbol{s}_v)$ depend on the Euclidean distance $d_{uv} = \| \boldsymbol{s}_u - \boldsymbol{s}_v \|$ between points $\boldsymbol{s}_u$ and $\boldsymbol{s}_v$, sometimes $C(\boldsymbol{s}_u, \boldsymbol{s}_v)$ is also denoted by $C(d_{uv})$ or $C(d)$. The covariance matrix structure which depends on parameters $\phi = (\phi_1, \dots, \phi_s)^T$ as given in Equation (1) (Uribe-Opazo et al., 2012):

$$\boldsymbol{\Sigma} = \phi_1 \boldsymbol{I}_n + \phi_2 \boldsymbol{R}, \qquad (1)$$

where, $\phi_1 \geq 0$ is the parameter known as nugget effect; $\phi_2 \geq 0$ is known as sill; $\boldsymbol{R} = \boldsymbol{R}(\phi_3, \phi_4) = [(r_{uv})]$ or $\boldsymbol{R} = \boldsymbol{R}(\phi_3) = [(r_{uv})]$ is an $n \times n$ symmetric matrix, which is a function of $\phi_3 > 0$, and sometimes also function of $\phi_4 > 0$, with diagonal elements $r_{uu} = 1$, $(u = 1, \dots, n)$; $r_{uv} = \phi_2^{-1} C(\boldsymbol{s}_u, \boldsymbol{s}_v)$ for $\phi_2 \neq 0$, and $r_{uv} = 0$ for $\phi_2 = 0$, $u \neq v = 1, \dots, n$, where r_{uv} depends on d_{uv}; ϕ_3 is a function of the model range (*a* = g(ϕ_3)), ϕ_4 when it exists it is known as the smoothness parameter, and $\boldsymbol{I_n}$ is an $n \times n$ identity matrix. The Matérn family (Jin and Kelly, 2017) is an example of covariance functions and Table 1 presents special cases of this particularly attractive class of models.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \phi^T)^T$ be the vector of unknown parameters. The log-likelihood for the GSLM is given in Equation (2):

$$\mathcal{L}(\boldsymbol{\theta}) = -\left(\frac{n}{2}\right) log(2\pi) - \frac{1}{2} log|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X\beta})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \boldsymbol{X\beta}), \qquad (2)$$

and the scores functions by $\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} = \boldsymbol{Y} - \boldsymbol{X\beta}$, and

$$\boldsymbol{U}(\phi) = \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \phi} = -\frac{1}{2} \frac{\partial vec^T(\boldsymbol{\Sigma})}{\partial \phi}^T vec(\boldsymbol{\Sigma}^{-1}) + \frac{1}{2} \frac{\partial vec^T(\boldsymbol{\Sigma})}{\partial \phi}^T vec(\boldsymbol{\Sigma}^{-1} \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T \boldsymbol{\Sigma}^{-1}).$$

From the solution of the score function of $\boldsymbol{\beta}$, $\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \boldsymbol{0}$, the maximum likelihood estimator $\boldsymbol{\beta}$ is given by $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{Y}$. The derivatives of first and second order of the scale matrix $\boldsymbol{\Sigma}$, with respect to ϕ_1, ϕ_2 and ϕ_3, for some covariance functions are presented in Uribe-Opazo et al. (2012), however the score equation for $\boldsymbol{\phi}$ does not lead to a closed-form solution for $\widehat{\boldsymbol{\phi}}$.

The parameter ϕ_4 is considered as fixed. The criteria considered to choose the geostatistical model for the covariance matrix were the cross-validation (CVA), trace (T_r) of the asymptotic covariance matrix of an estimated mean and the log-likelihood maximum value (LMV) (De Bastiani et al. 2015).

Asymptotic standard errors can be calculated by inverting either observed information matrix, $\boldsymbol{I}(\boldsymbol{\theta})$ or the expected information matrix, $\boldsymbol{F}(\boldsymbol{\theta})$, where $\boldsymbol{I}(\boldsymbol{\theta})$ is $\boldsymbol{I}(\boldsymbol{\theta}) = -\boldsymbol{L}(\boldsymbol{\theta})$, evaluated in $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, with $\boldsymbol{L}(\boldsymbol{\theta}) = \partial^2 \mathcal{L}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T$ and $\boldsymbol{F}(\boldsymbol{\theta})$ is given by

$$\boldsymbol{F}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{F}_{\boldsymbol{\beta\beta}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{F}_{\boldsymbol{\phi\phi}} \end{pmatrix},$$

where, $\boldsymbol{F}_{\beta\beta} = \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X}$, and $\boldsymbol{F}_{\phi\phi} = \frac{1}{2} \frac{\partial vec^T(\boldsymbol{\Sigma})}{\partial \phi} (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}) \frac{\partial vec(\boldsymbol{\Sigma})}{\partial \phi^T}$.

Likelihood ratio statistic- Hypothesis test for $\boldsymbol{\beta}$ vector

Consider the partitioned vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^T$, where $\boldsymbol{\beta}_1 = (\beta_1, \dots, \beta_q)^T$ and $\boldsymbol{\beta}_2 = (\beta_{q+1}, \dots, \beta_p)^T$ of dimension q and $(p - q)$, respectively, and $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)^T$ in such way that $\boldsymbol{X\beta} = \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{X}_2 \boldsymbol{\beta}_2$. Let $\boldsymbol{\beta}_1$ be the parameter of interest.

Table 7. Average profitability (a.p.) in dollars ($).

	All points		Without #3 (C1)		Without # 6 (C2)		Without #16 (C3)	
Classes (t ha^{-1})	Área (ha)	a.p. ($)	Área (ha)	a.p. ($)	Área (ha)	a.p. ($)	Área (ha)	a.p. ($)
1.29 - 2.23	0.77	402.97	1.24	648.93	1.76	921.07	0.79	413.43
2.24 - 3.17	10.40	8,380.51	10.44	8,412.74	12.20	9,830.98	10.42	8,396.63
3.18 - 4.12	10.14	11,005.22	9.21	9,995.87	7.63	8,281.05	10.03	10,885.83
4.13 - 5.06	1.15	1,569.56	1.54	2,101.84	1.03	1,405.78	1.26	1,719.69
5.07 - 6.00	0.17	280.04	0.19	312.99	0.01	16.47	0.12	197.68
Total	22.62	21,638.30	22.62	21,472.37	22.62	20,455.35	22.62	21,613.26

Let $H_0: \boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^0$, be the hypothesis of interest versus the alternative hypothesis $H_1: \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_1^0$, where $\boldsymbol{\beta}_1^0$ is a fixed vector of dimension q $(q \leq p)$, where, $\widehat{\boldsymbol{\theta}} = \left(\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T, \widehat{\boldsymbol{\phi}}^T\right)^T$ is the unrestricted ML estimator for $\boldsymbol{\theta}$, and denote with a tilde the restricted ML estimator. So, $\widetilde{\boldsymbol{\theta}} = (\boldsymbol{\beta}_1^{(0)T}, \widetilde{\boldsymbol{\beta}}_2^T, \widetilde{\boldsymbol{\phi}}^T)^T$ is the restricted ML estimator of $\boldsymbol{\theta}$, where $\widetilde{\boldsymbol{\beta}}_2$ and $\widetilde{\boldsymbol{\phi}}$ are the restricted ML estimators of $\boldsymbol{\beta}_2$ and $\boldsymbol{\phi}$ under H_0.

The likelihood ratio statistic (*LR*) to test $H_0: \boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^0$ versus $H_1: \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_1^0$, is defined by

$LR = 2\left(\mathcal{L}(\widehat{\boldsymbol{\theta}}) - \mathcal{L}(\widetilde{\boldsymbol{\theta}})\right)$,

where, $\mathcal{L}(\widehat{\boldsymbol{\theta}}) = -\frac{n}{2}log(2\pi) - \frac{1}{2}log|\widehat{\boldsymbol{\Sigma}}| - \frac{1}{2}\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$, and

$\mathcal{L}(\widetilde{\boldsymbol{\theta}}) = -\frac{n}{2}log(2\pi) - \frac{1}{2}log|\widetilde{\boldsymbol{\Sigma}}| - \frac{1}{2}\left(\boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}\right)^T \widetilde{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}\right)$.

Thus, the likelihood ratio statistic has the form given in Equation (3),

$$LR = log\left(\frac{|\widetilde{\boldsymbol{\Sigma}}|}{|\widehat{\boldsymbol{\Sigma}}|}\right) + \frac{1}{2}\left(\tilde{\delta} - \hat{\delta}\right), \qquad (3)$$

Where, $\tilde{\delta} = (\boldsymbol{Y} - \boldsymbol{X}_1\boldsymbol{\beta}_1^{(0)} - \boldsymbol{X}_2\widetilde{\boldsymbol{\beta}}_2)^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y} - \boldsymbol{X}_1\boldsymbol{\beta}_1^{(0)} - \boldsymbol{X}_2\widetilde{\boldsymbol{\beta}}_2)$, is evaluated in $\widetilde{\boldsymbol{\theta}}$, and

$\hat{\delta} = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$, is evaluated in $\widehat{\boldsymbol{\theta}}$.

Asymptotically and under the null hypothesis, the *LR* statistic is distributed as a chi-squared random variable, with degrees of freedom equal to q.

Likelihood ratio statistic - Hypothesis test for covariance structure

The main goal is to test whether the model should take into account the spatial structure, or not, which can be achieved by testing $H_0: \phi_2 = 0$ versus $H_1: \phi_2 > 0$. When $\phi_2 = 0$, $\boldsymbol{\Sigma} = \phi_1\boldsymbol{I_n}$, and when $\phi_2 > 0$ then $\boldsymbol{\Sigma}$ is given in Equation (1).

The corresponding likelihood ratio test (*LR**) is given by

$$LR^* = 2\left[\mathcal{L}(\widehat{\boldsymbol{\theta}}) - \mathcal{L}(\widetilde{\boldsymbol{\theta}})\right], \qquad (4)$$

where $\hat{\theta} = \left(\widehat{\boldsymbol{\beta}}^T, \hat{\phi}^T\right)^T$ is the unrestricted ML estimator for $\boldsymbol{\theta}$, and $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\beta}}^T, \tilde{\phi}^T)^T$ is the restricted ML estimator of $\boldsymbol{\theta}$, where $\widetilde{\boldsymbol{\beta}}$ and $\tilde{\phi}$ are the restricted ML estimators of $\boldsymbol{\beta}$ and ϕ under H_0, i.e., $\tilde{\phi} = (\tilde{\phi}_1, 0, \tilde{\phi}_3)$.

It is important to mention that with this type of null hypothesis, we are at the boundary of the admissible parameter space. According to Self and Liang (1987), assuming that the errors are normal distributed and under the null hypothesis, the asymptotic distribution of the likelihood ratio corresponds to a 50:50 mixture of chi-squares with zero and one degree of freedom.

Diagnostics - Global influence

Detecting influential observations are an important step in the analysis of a data set. There are some papers in the literature on diagnostic for spatial linear models. Warnes (1986) observed the sensitivity of predictions to perturbations in the covariance function. Christensen et al. (1992) discussed case deletion diagnostics for detecting observations that are influential for prediction based on universal kriging. Militino et al. (2006) showed that case deletion diagnostics do suffer from masking and suggest robust procedures based on subsets of data free from outliers. More recently, De Bastiani et al. (2017, 2018) developed local and global influence for Gaussian spatial linear models with repetitions, respectively.

Case-deletion is a diagnostic technique that evaluates the impact on the parameter estimates given by the model, by eliminating one or more observations from the data set. The typical measure is the Cook's distance (Cook, 1977). For the GSLM, the Cook's distance is given by $D_{i\theta} = (\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{[i]})^T\boldsymbol{F}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{[i]})$.

Because the expected information matrix is block diagonal, $D_{i\theta}$ can be written as

$D_{i\theta} = D_{i\beta} + D_{i\phi}$,

where $D_{i\beta} = (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{[i]})^T\boldsymbol{F}(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{[i]})$ and $D_{i\phi} = (\hat{\phi} - \hat{\phi}_{[i]})^T\boldsymbol{F}(\hat{\phi})(\hat{\phi} - \hat{\phi}_{[i]})$.

Denote $\mathcal{L}_{[i]}(\boldsymbol{\theta})$ the log-likelihood with the i-th observation deleted, and $\widehat{\boldsymbol{\theta}}_{[i]}$ the maximum likelihood estimates under $\mathcal{L}_{[i]}(\boldsymbol{\theta})$. To calculate $\widehat{\boldsymbol{\theta}}_{[i]}$, we used one-step approximation to Cook's distance (Pan et al., 2014), $\widehat{\boldsymbol{\theta}}_{[i]} = \widehat{\boldsymbol{\theta}} + [\ddot{\mathcal{L}}(\widehat{\boldsymbol{\theta}})]^{-1}\dot{\mathcal{L}}_{[i]}(\widehat{\boldsymbol{\theta}})$, where $\dot{\mathcal{L}}_{[i]}(\boldsymbol{\theta}) = \partial\,\mathcal{L}_{[i]}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and $\ddot{\mathcal{L}}(\boldsymbol{\theta}) = \partial^2\mathcal{L}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$, evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$.

The one-step Cook's distance becomes $D_{i\theta}^1 = \dot{\mathcal{L}}_{[i]}(\widehat{\boldsymbol{\theta}})[\boldsymbol{F}(\widehat{\boldsymbol{\theta}})]^{-1}\dot{\mathcal{L}}_{[i]}(\widehat{\boldsymbol{\theta}}) = D_{i\beta}^1 + D_{i\phi}^1$ with corresponding $D_{i\beta}^1$ and $D_{i\phi}^1$.

Diagnostics - Local influence

One of the purposes of diagnostic techniques is to evaluate the stability of the fitted model in a data set and should be part of all statistical analysis since influential observations may distort the values of the statistic of interest and lead us to misleading results.

In the local influence method, introduced by Cook (1986), a perturbation scheme is introduced into the postulated model through a perturbation vector $\boldsymbol{\omega} = (\omega_1, \dots, \omega_k)^T (\boldsymbol{\omega} \in \Omega \subset \mathbb{R}^k)$, generating the perturbed model, where $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{\omega})$ is the corresponding log-likelihood function. The influence measure is constructed using the basic geometric idea of curvature of the likelihood displacement given by

$LD(\boldsymbol{\omega}) = 2[\mathcal{L}(\widehat{\boldsymbol{\theta}}) - \mathcal{L}(\widehat{\boldsymbol{\theta}}_\omega)]$,

where $\widehat{\boldsymbol{\theta}}$ is the ML estimator of $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \phi^T)^T$ in the postulated model, with $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T, \phi = (\phi_1, \dots, \phi_s)^T$ and $\widehat{\boldsymbol{\theta}}_\omega$ is the ML estimator of $\boldsymbol{\theta}$ in the perturbed model.

Cook (1986) proposed the use of the normal curvature C_l of $LD(\boldsymbol{\omega})$ at ω_0 in the direction of some unit vector l,

$C_l = C_l(\boldsymbol{\theta}) = 2|\boldsymbol{l}^T \boldsymbol{\Delta}^T \boldsymbol{L}^{-1} \boldsymbol{\Delta} \boldsymbol{l}|$,

with $||\boldsymbol{l}|| = 1$, and $-\boldsymbol{L} = -\boldsymbol{L}(\boldsymbol{\theta})$ is the observed information matrix, evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_\beta^T, \boldsymbol{\Delta}_\phi^T)^T$, where $\boldsymbol{\Delta}_\beta = \partial^2 \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{\omega}) / \partial \boldsymbol{\beta} \partial \boldsymbol{\omega}^T$ and $\boldsymbol{\Delta}_\phi = \partial^2 \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{\omega}) / \partial \phi \partial \boldsymbol{\omega}^T$, evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ and at $\boldsymbol{\omega} = \boldsymbol{\omega}_0$.

The plot of the elements $|l_{max}|$ versus index (order of data) can reveal what type of perturbation has more influence on $LD(\boldsymbol{\omega})$, in the neighborhood of $\boldsymbol{\omega}_0$, (Cook, 1986). Poon and Poon (1999) proposed the conformal normal curvature $B_l = C_l / tr(2\boldsymbol{J})$, where $\boldsymbol{J} = \boldsymbol{\Delta}^T \boldsymbol{L}^{-1} \boldsymbol{\Delta}$. The conformal curvature in the unit direction with $j - th$ entry 1 and all the other entries 0 is given by $B_i = 2|j_{ii}| / tr(2\boldsymbol{J})$. The plot of B_i versus index can reveal potential influential observations.

To verify if a perturbation scheme is appropriate, Zhu et al. (2007) proposed the use of the Fisher information matrix of $\boldsymbol{\omega}$ in the perturbed model considering the vector $\boldsymbol{\theta}$ as fixed.

Following De Bastiani et al. (2015), let us consider as perturbation scheme the model shift in mean, i.e $\boldsymbol{Y} = \boldsymbol{\mu}(\boldsymbol{\omega}) + \boldsymbol{\varepsilon}$ with $\boldsymbol{\mu}(\boldsymbol{\omega}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\omega}$ where $\boldsymbol{A}$, *n x n,* is a matrix that does not depend on $\boldsymbol{\beta}$ or on $\boldsymbol{\omega}$. In this case $\boldsymbol{\omega_0} = \boldsymbol{0}$.

Equivalently, we can write $\boldsymbol{Y}_{\boldsymbol{\omega}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{Y}_{\boldsymbol{\omega}} = \boldsymbol{Y} + (-1)\boldsymbol{A}\boldsymbol{\omega}$ that corresponds to a perturbation scheme of the response vector.

The perturbed log-likelihood is given by

$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{\omega}) = -\left(\frac{n}{2}\right) log(2\pi) - \frac{1}{2} log|\boldsymbol{\Sigma}| - \frac{1}{2}\big(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\omega})\big)^T \boldsymbol{\Sigma}^{-1} \big(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\omega})\big)$.

To select an adequate matrix $\boldsymbol{A}$ we can use the methodology proposed by Zhu et al (2007). In effect, the score function for $\boldsymbol{\omega}$ in the perturbed log-likelihood function is given by:

$$\mathbf{U}(\boldsymbol{\omega}) = \frac{\partial \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial \boldsymbol{\omega}} = \mathbf{A}^T \boldsymbol{\Sigma}^{-1} \big(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\omega})\big).$$

Let

$\mathbf{G}(\boldsymbol{\omega}) = \mathrm{E}_\omega[\mathbf{U}(\boldsymbol{\omega})\mathbf{U}^T(\boldsymbol{\omega})] = \mathrm{diag}[\mathrm{g}_{11}(\boldsymbol{\omega}_1), \dots, \mathrm{g}_{nn}(\boldsymbol{\omega}_n)]$ be the Fisher information matrix with respect to the perturbation vector $\boldsymbol{\omega}$. A perturbation $\boldsymbol{\omega}$ is appropriate if it satisfies $g_{jj}(\boldsymbol{\omega_0}) = c\boldsymbol{I_n}$,

where $c > 0$. In our case, we have $g_{jj}(\boldsymbol{\omega_0}) = c\boldsymbol{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{A}$, with $c = 1$. Note that usually $\boldsymbol{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{A} \neq \boldsymbol{I_n}$. However if $\boldsymbol{A} = \boldsymbol{\Sigma}^{1/2}$ then $g_{jj}(\boldsymbol{\omega_0}) = c\boldsymbol{I_n}$ and so $\boldsymbol{\mu}(\boldsymbol{\omega}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\Sigma}^{1/2}\omega$ is a perturbation scheme appropriate, as shown in De Bastiani et al (2015).

Considering the appropriate perturbation scheme for the response variable, where $\boldsymbol{\Delta}_\beta$ is an $p \times n$ matrix and $\boldsymbol{\Delta}_\phi$ is an $3 \times n$ matrix given by

$\boldsymbol{\Delta}_\beta = \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\omega}^T} = -\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1/2}$, and $\boldsymbol{\Delta}_\phi = \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial \phi \partial \boldsymbol{\omega}^T} = \frac{\partial vec^T(\boldsymbol{\Sigma})}{\partial \phi} vec(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1/2})\, vec(\,\boldsymbol{\varepsilon} \otimes \boldsymbol{1}^T)$,

evaluated in $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ and $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\varepsilon}} = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$ and $\boldsymbol{1}$ is an $n \times 1$ vector of ones.

Measure of the spatial dependence

The measurement of the spatial dependence degree of the adjusted models was obtained using the Spatial Dependence Measure - *SDI* models developed by Neto et al. (2020), shown in Equation (5),

$$SDI = MF\left(\frac{\varphi_2}{\varphi_1 + \varphi_2}\right) min\left\{1; \left(\frac{a}{0.5MD}\right)\right\} 100\,, \qquad 5)$$

in which, *a* is the range, *MF* is the model factor (specific to each semivariogram model) and *MD* is the maximum distance between two sampling points.

Using Equation 5, *SDI* spatial dependence measures were obtained for the Matérn family with different smoothing parameters ϕ_4. The categorization of the *SDI* index was obtained using the criterion of Seidel and Oliveira (2016), and the results are presented in Table 2.

All computations are performed with software R (R Development Core Team, 2021) using the package geoR (Ribeiro JR. and Diggle, 2001).

Conclusion

The spatial linear models enabled us to verify the spatial dependence between the wheat yield data in the study area, according to the two varieties and plant attributes. The likelihood tests presented confirmed the importance of the explanatory variable to explain the response variable, wheat yield and confirmed the need to consider explicative variables.

The maps constructed allowed us to predict the wheat yield in the studying area. This can be used to create management zones with low or high yields with the purpose of unifying similar areas, apply localized inputs, and then maximize the profit reducing the environmental impact. The disregard of potentially

influential observations caused changes in the parameters estimates that define the spatial dependence structure, and consequently then in the profitability in sectors of the wheat yield maps.

The study of statistical inference and diagnostics on spatial data should be part of all the geostatistical analysis.

The developed methodology in this paper can be applied to study other crops yield in different areas, or from different years.

Acknowledgments

The authors are grateful for the financial support from Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Finance Code 001, National Council for Scientific and Technological Development (CNPq), FONDECYT Chile (Project No. 1150325), and Spatial Statistics Laboratory – LEE-UNIOESTE.

References

Abdallah MA (2018) A special analytical methodology for variogram modeling and interpolation of terrain elevation data by kriging method. Damascus University Journal For The Engineering Sciences. 34(1): 7-26.

Ahmad S, Raza MAS, Saleem MF, Iqbal R, Zaheer MS, Haider I, Aslam MU, Ali M, Khan IH (2020) Significance of partial root zone drying and mulches for water saving and weed suppression in wheat. The Journal of Animal & Plant Sciences. 30(1): 154-162.

Ameer S, Cheema MJM, Khan MA, Amjad M, Noor M, Wei L (2022) Delineation of nutrient management zones for precise fertilizer management in wheat crop using geo-statistical techniques. Soil Use Manage. 38:1430–1445.

Christensen R, Johnson W, Pearson L (1992) Prediction diagnostics for spatial linear models. Biometrika. 79(3): 583–591.

Cook RD (1977) Detection of influential observations in linear regression. Technometrics. 19(1): 15–18.

Cook RD (1986) Assessment of local influence. Journal of the Royal Statistical Society. 48(2): 133–169.

Cressie N (2015) Statistics for spatial data. New York: John Wiley & Sons. 936 p.

Dalposso GH, Uribe-Opazo, MA, Mercante E, Johann JA, Borssoi JA (2012) Comparison measures of maps generated by geostatistical methods. Engenharia Agrícola. 32(1): 174-183.

Dalposso GH, Uribe-Opazo MA, De Bastiani F (2021) Spatial-temporal Analysis of Soybean Productivity Using Geostatistical Methods. 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.

De Bastiani F, Galea M, Cysneiros A, Uribe-Opazo MA (2017) Gaussian spatial linear models with repetitions: an application to soybean productivity. Spatial Statistics. 21(Part A): 319–335.

De Bastiani F, Uribe-Opazo MA, Galea M, Cysneiros AHMA (2018) Case-deletion diagnostics for spatial linear mixed models. Spatial Statistics. 28: 284-303.

Embrapa (2013) Sistema Brasileiro de Classificação de Solos. 3rd edn. Brasília: Embrapa Solos. 353p.

Faostat (2019) Food and Agriculture Statistics. <https://www.fao.org/food-agriculture-statistics/en/>

Gill PK, Kaur K, Singh A, Kaur M (2022) Soil characteristic features and availability of nutrients under various fertilization strategies in wheat. International Journal of Modern Developments in Engineering and Science. 1(7): 1-6.

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. Chilean Journal of Agricultural Research. 73(4): 414–423.

Guedes LPC, Bach RT, Uribe-Opazo MA (2020) Nugget effect influence on spatial variability of agricultural data. Engenharia Agrícola. 40(1): 96-104.

Hengl T, Heuvelink GBM, Stein A (2003) Comparison of kriging with external drift and regression-kriging. Techinical note, International Institute for Geo-information Science and Earth Observation (ITC). <http://www.itc.nl/library/Academic-output>

Ibge (2019) Produção Agrícola Municipal: Área plantada, área colhida, quantidade produzida, rendimento médio e valor da produção das lavouras temporária. <https://sidra.ibge.gov.br/tabela/1612>

Jiang T, Lui J, Gao Y, Sun Z, Chen S, Yao N, Ma H, Feng H, Yu Q, He J (2020) Simulation of plant height of winter wheat under soil Water stress using modified growth functions. Agricultural Water Management. 232: 106066.

Jin R, Kelly G (2017) A comparison of sampling grids, cut-off distance and type of residuals in parametric variogram estimation. Communications in Statistics - Simulation and Computation. 46(3): 1781-1795.

Jurado-Expósito M, López-Granados F, Jiménez-Brenes FM, Torres-Sánchez (2021) Monitoring the Spatial Variability of Knapweed (Centaurea diluta Aiton) in Wheat Crops Using Geostatistics and UAV Imagery: Probability Maps for Risk Assessment in Site-Specific Control. Agronomy. 11(5): 880.

Leiva V, Sánchez L, Galea M, Saulo H (2020) Global and local diagnostic analytics for a geostatistical model based on a new approach to quantile regression. Stochastic Environmental Research and Risk Assessment. 34:1457–1471

Ma C, Liu Y, Wang J, Xue L, Hou P, Xue L, Yang L (2022) Warming increase the N2O emissions from wheat fields but reduce the wheat yield in a rice-wheat rotation system. Agriculture, Ecosystems and Environment. 337: 108064.

Militino A, Palacius M, Ugarte M (2006) Outliers detection in multivariate spatial linear models. Journal of Statistical Planning and Inference. 136(1): 125–146.

Mladenov V, Dimitrijević S, Boćanski J, Banjac B, Kondić-Špika A, Trkulja D (2019) Genetic analysis of spike length in wheat. Genetika. 51(1): 167-178.

Neto EA, Seidel EJ, Oliveira MS (2020) Geostatistical-based index for spatial variability in soil properties. Revista Brasileira de Ciência do Solo. 44: e0200086.

Pan J, Fei Y, Foster P (2014) Case-deletion diagnostics for linear mixed models. Technometrics. 56(3): 269–281.

Patel VS, Patel KC, Patel PK, Patel KM (2019) Study on Long Term Effect of Organic and Inorganic Farming on Growth, Yield and Nutrient Content of Wheat Crop. International Journal of Agriculture Sciences. 11(5): 8010-8013.

Poon W, Poon YS (1999) Conformal normal curvature and assessment of local influence. Journal of the Royal Statistical Society. Series B. 61(1): 51–61.

Pu Y, Zhao X, Chi G, Zhao S, Wang J, Jin Z, Yin J (2019) Design and implementation of a parallel geographically weighted k-nearest neighbor classifier. Computers & Geosciences. 127: 111-122.

R Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org>

Ribeiro JR PJ, Diggle PJ (2001) geoR: A package for geostatistical analysis. R-NEWS 1:15-18.

Seidel EJ, Oliveira MS (2016) A Classification for a Geostatistical Index of Spatial Dependence.

Revista Brasileira de Ciência do Solo. 40:e0160007

Self SG, Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of American Statistical Association. 82(398): 605–610.

Shewry PR (2007) Improving the protein content and composition of cereal grain. Journal of Cereal Science. 46(3): 239–250.

Sidorova VA, Zhukovskiib EE, Lekomtsevb PV, Yakushev VV (2012) Geostatistical analysis of the soil and crop parameters in a field experiment on precision agriculture. Eurasian Soil Science. 45(8): 783–792.

Singh R, Tiwari R, Sharma D, Tiwari V, Sharma I (2015) Variability in yield traits of Tilling population of bread wheath (Triticum aestivum L.). Journal of Applied and Natural Science. 7(1): 443-446.

Sun A, Parker PA, Holan SH (2022) Analysis of household pulse survey public-use microdata via unit-level models for informative sampling. Stats. 5:139-153.

Upadhyay H, Kaur K (2019) Response of combined application of Inorganic and organic fertilizers on the growth and yield attributes of wheat. Think India Journal. 22(30): 400-408.

Uribe-Opazo MA, Borssoi J, Galea M (2012) Influence diagnostics in Gaussian spatial linear models. Journal of Applied Statistics. 39(3): 615–630.

Uribe-Opazo MA, De Bastiani F, Galea M, Schemmer RC, Assumpção RA (2021) Appropriate perturbation scheme for the covariance matrix of a t-student spatial linear model. Spatial Statistics. 41: 100481.

Warnes J (1986) Sensitivity analysis for universal kriging. Mathematical Geology. 18: 653–676.

Yuan Y, Shi B, Yost R, Liu X, Tian Y, Zhu Y, Cao W, Cao Q (2022) Optimization of Management Zone Delineation for Precision Crop Management in an Intensive Farming System. Plants. 11(19): 2611.

Zhu H, Ibrahim J. Lee S, Zhang H (2007) Perturbation selection and influence measures in local influence analysis. Annals of Statistics 35(6): 2565–2588.