AgriEngineering

Article

Improving Coffee Yield Interpolation in the Presence of Outliers Using Multivariate Geostatistics and Satellite Data

César de Oliveira Ferreira Silva [1,*], Celia Regina Grego [2], Rodrigo Lilla Manzione [3] and Stanley Robson de Medeiros Oliveira [1,2]

[1] School of Agricultural Engineering, Campinas State University (UNICAMP), Campinas 13083-875, Brazil; stanley.oliveira@embrapa.br
[2] Embrapa Digital Agriculture, Campinas 13083-886, Brazil; celia.grego@embrapa.br
[3] School of Science, Technology and Education, São Paulo State University (UNESP), Ourinhos 19903-302, Brazil; lilla.manzione@unesp.br
* Correspondence: cesaroliveira.f.silva@gmail.com

Abstract: Precision agriculture for coffee production requires spatial knowledge of crop yield. However, difficulties in implementation lie in low-sampled areas. In addition, the asynchronicity of this crop adds complexity to the modeling. It results in a diversity of phenological stages within a field and also continuous production of coffee over time. Big Data retrieved from remote sensing can be tested to improve spatial modeling. This research proposes to apply the Sentinel-2 vegetation index (NDVI) and the Sentinel-1 dual-polarization C-band Synthetic Aperture Radar (SAR) dataset as auxiliary variables in the multivariate geostatistical modeling of coffee yield characterized by the presence of outliers and assess improvement. A total of 66 coffee yield points were sampled from a 4 ha area in a quasi-regular grid located in southeastern Brazil. Ordinary kriging (OK) and block cokriging (BCOK) were applied. Overall, coupling coffee yield with the NDVI and/or SAR in BCOK interpolation improved the accuracy of spatial interpolation of coffee yield even in the presence of outliers. Incorporating Big Data for improving the modeling for low-sampled fields requires taking into account the difference in supports between different datasets since this difference can increase uncontrolled uncertainty. In this manner, we will consider, for future research, new tests with other covariates. This research has the potential to support precision agriculture applications as site-specific plant nutrient management.

Keywords: *Coffea arabica* L.; precision agriculture; cokriging; variogram

Citation: Silva, C.d.O.F.; Grego, C.R.; Manzione, R.L.; Oliveira, S.R.d.M. Improving Coffee Yield Interpolation in the Presence of Outliers Using Multivariate Geostatistics and Satellite Data. *AgriEngineering* **2024**, *6*, 81–94. https://doi.org/10.3390/agriengineering6010006

Academic Editor: Bugao Xu

Received: 20 November 2023
Revised: 13 December 2023
Accepted: 20 December 2023
Published: 10 January 2024

1. Introduction

Coffee yield can be affected by climate [1], the occurrence of pests [2], plant physiology [3], water use, plant density and population [4], slope [5], and other factors [6]. Also, coffee plots present higher and lower yield levels at different locations of the plot in alternate years, a characteristic named "biennial yield" [7]. Many different combinations of tools can be used to deal with this highly heterogeneous scenario [8].

In 2019, the International Society of Precision Agriculture [9] defined precision agriculture (PA) as a management strategy based on the combination of temporal (different seasons, years, etc.), spatial (varying across the farmland), and individual (agronomical knowledge) data to support decision making. This site-specific treatment seeks to optimize the use of resources in terms of agronomic efficiency and financial profits. This means that PA is based on spatial variability to make decisions. Soil and its properties are highly complex; therefore, any sampling at a finite number of sites inevitably gives an incomplete description of natural variation. The recognition of spatial variability takes into account the spatial heterogeneity of soil attributes, ranging from global to micro-scales. Taking into account this range of potential factors, uniform management based on the assumption of spatial homogeneity across the field disregards this wide range of influences. In this

manner, spatial modeling approaches are needed to guide site-specific management [10]. This includes geostatistical methods, such as the different types of kriging [11–16].

Geostatistics refers to the statistical analysis of phenomena that change in a continuous spatial manner. It can be defined as the tools that study and predict the spatial structure of georeferenced variables. Geostatistics has been widely applied in agricultural science to solve the problem of estimating soil and plant properties in unsampled locations from sample data [17,18]. Measurement techniques hardly work on the same scale as the process of interest. Therefore, some small-scale variability may be lost because sampling at a lower scale is necessary and can rarely be achieved. Yost [19] suggests 30 pairs of points as the minimum number of pairs of points to structure a variogram, while Webster [13] suggests 100–140 points as the minimum range.

However, when a field is low-sampled, outliers may occur, making the performance of univariate kriging techniques problematic [20]. In some situations, the information is multivariate: samples are collected from several locations, and several measurements are taken for each one. The tools used in multivariate geostatistical analysis are analogous to those in univariate analysis and include intrinsic hypothesis, covariance, and cokriging [14]. In addition, multivariate geostatistical methods are suited for Big Data applications, since this allows for the use of auxiliary datasets for improving interpolation over unsampled areas, especially when dealing with the presence of outliers and irregular grids [10]. More sophisticated geostatistical models, like cokriging [13], can include auxiliary data.

Big Data are massive volumes of unstructured and structured datasets considered difficult to process, analyze, and manage using traditional data-processing techniques [21,22]. With the increasing number of remotely sensed data provided by sensors coupled on orbital satellites at various spatial and temporal resolutions, the number of data generated has grown exponentially, making multispectral imagery for calculating vegetation indices [21,23] and SAR satellite imagery [24–26] significant sources of Big Data [22].

The present research problem is how to model the spatial dependence (or autocorrelation) of specialty coffee yield in an accurate way since it is an asynchronous crop that tends to have outliers when it is low-sampled. Therefore, the research hypothesis is that the use of Big Data as covariates, integrated with multivariate geostatistics, provides models capable of successfully interpolating the yield of specialty coffee crops, as well as helping to increase the accuracy of this modeling compared with a univariate approach.

In this context, the objective of this study was to evaluate the use of remotely sensed data as auxiliary variables in the block cokriging (BCOK) modeling of coffee yield characterized by the presence of outliers.

2. Materials and Methods

2.1. Description of the Study Site and Agronomic Practices

The experimental site was located in Paraguaçu (southern Minas Gerais, Brazil; 21° 39′ 13″ S and 45° 48′ 07″ W), as shown in Figure 1 together with the soil sampling distribution in May 2021. This is a 4 ha area for coffee cultivation (*Coffea arabica* L.) with the cultivar Catucai Amarelo, transplanted in 2012. There are 2000 plants ha^{-1} separated by distances of 2.5 m between rows and 1 m between plants. This coffee plot has an altitude of 894.3 m. The sampling plan was defined across the coffee plot using a GPS (mean error of 2–5 m), totaling 40 georeferenced sampling points. Samples of coffee yield were collected in May 2021 by obtaining subsamples from each group of 2 coffee trees composing the sampling point. In this area, the soil is classified as Argisols. This is characterized as soil with higher natural fertility (eutrophic), good physical conditions, and more gentle terrain that has greater potential for agricultural use. Its limitations are more related to its low fertility, acidity, high aluminum content, and susceptibility to erosion processes, especially on rougher terrain [27]. Uniform fertilization over the entire coffee plot was conducted directly on the soil in 2021 by applying around 42, 10, and 42 kg·ha^{-1} of N, P, K.

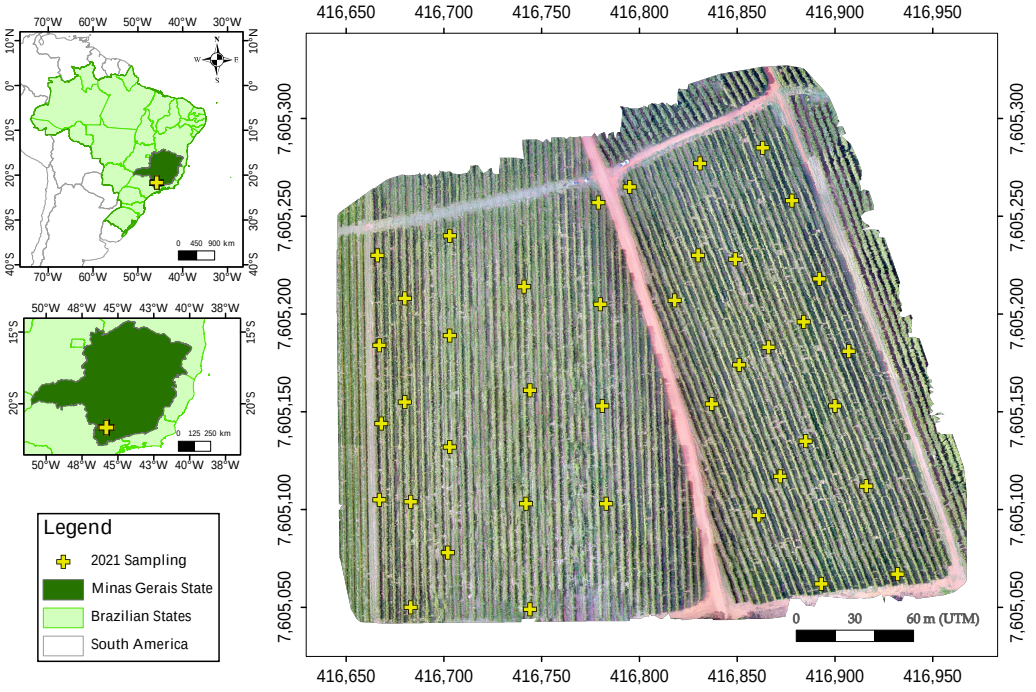

Figure 1. UAV image of the study area with coffee yield sampling points in May 2021.

The climatic conditions of the municipality of Paraguaçu are shown in Figure 2 in terms of monthly accumulated rainfall and average monthly air temperature. According to data from the National Meteorological Institute (INMET) [28], since 1961, the absolute minimum temperature recorded was on 9 June 1985, with a minimum of −1.8 °C, followed by −0.8 °C on 21 July 1981 and −0.6 °C on 18 July 2000. The historical maximum is 37.1 °C on 3 October 2020, with the previous record being October 2014, on the 14th and 15th, when the maximum reached 37 °C. The record for accumulated rainfall in 24 h was 140 mm on 8 November 1970. According to the Köppen–Geiger climatic classification, this area is classified as subtropical humid climate (Cwa) [29].

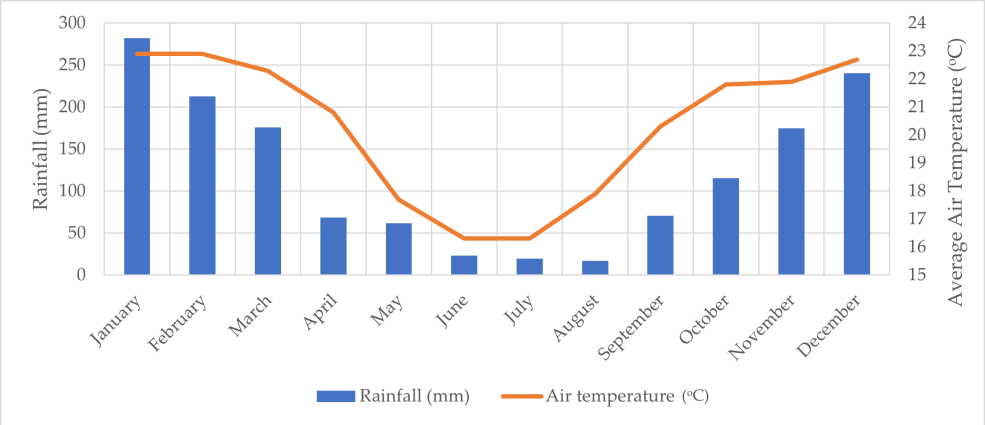

Figure 2. Climatic conditions (rainfall and average air temperature) of the municipality of Paraguaçu (state of Minas Gerais, southeastern Brazil).

2.2. Multivariate Geostatistics

To perform multivariate geostatistical analysis, a previous support check is needed [30], followed by support regularization if the variables have different supports, in other words, if they are of widely differing spatial resolutions, sizes, or depths [31]. Here, we performed descriptive and exploratory statistics on all variables. We wanted to regularize the variables

in a manner that presented mean zero and unit standard deviation. To achieve this distribution, we performed Gaussian anamorphosis transformation [32] and then fitted the linear model of co-regionalization (LMC) and used it for block kriging (BCOK) interpolation. A step-by-step flowchart (Figure 3) synthesizes the methodology.

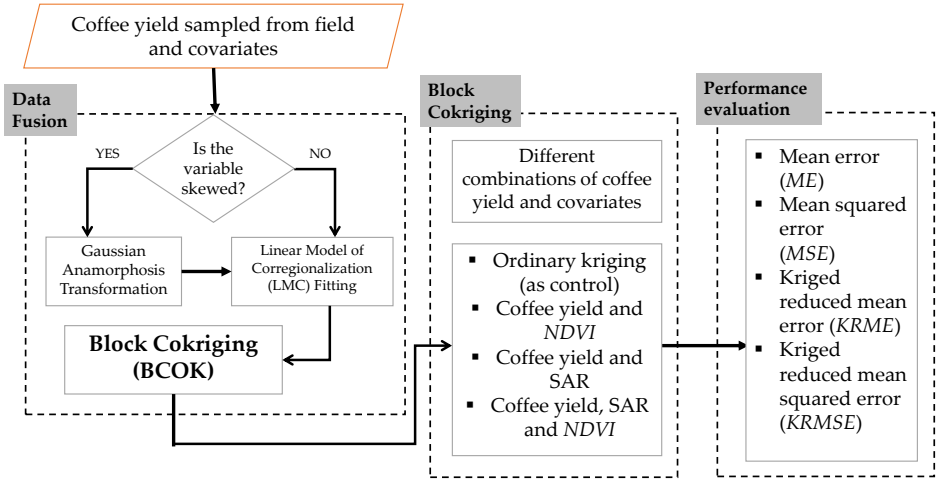

Figure 3. Overview of the methodology.

2.2.1. Preprocessing with Gaussian Anamorphosis Transformation

Gaussian anamorphosis transformation is used to convert skewed and non-Gaussian statistically distributed variables into new ones with mean zero and unit standard deviation [30]. BCOK variance and standard deviation using transformed variables instead of the original ones are closer to the linear and optimal situation [17]. This transformation is based on the fitting of a polynomial expansion, as shown in Equation (1), named Hermite polynomials.

$$\Phi = \sum \Psi_i H_i(Y) \tag{1}$$

where $H_i(Y)$ are Hermite polynomials, Ψ_i are Hermite coefficients. This function is reversible and able to convert a non-Gaussian variable into a new variable with mean zero and unit standard deviation, as shown in Equation (2).

$$Y = \Phi^{-1}(Z) \tag{2}$$

Then, we performed a geostatistical analysis on the new standardized variables. After that, we back-transformed the predictions into the raw distribution by using the same reversible anamorphosis function.

2.2.2. Fitting of Linear Model of Co-Regionalization (LMC)

The linear model of co-regionalization (LMC) is a unified model which considers the experimental direct and cross-variograms of all the n variables and then performs weighted least squares (WLS) on the pairs of samples at each lag [33]. To perform WLS well, the variables need to be highly correlated. The $n\ (n+1)/2$ experimental direct and cross-variograms of all the n variables are fitted with a linear combination of the N_S standardized variograms of unit sill, $g^u(\boldsymbol{u})$. In matrix notation, the LCM follows Equation (3):

$$\boldsymbol{\Gamma}(\boldsymbol{h}) = \sum_{u=1}^{N_S} \boldsymbol{B}^u g^u(\boldsymbol{h}) \tag{3}$$

where $\boldsymbol{\Gamma}(\boldsymbol{h}) = [\gamma_{ij}(\boldsymbol{h})]$ is a symmetric matrix (order of $n \times n$) where diagonal elements contain direct variograms and out-of-diagonal elements contain cross-variograms; $\boldsymbol{B}^u = [b_{ij}^u]$

(the co-regionalization matrix) is a symmetrical semi-definite matrix (order of $n \times n$) containing the sampled values b_{ij}^{u} for spatial support u [34,35].

2.2.3. Block Cokriging (BCOK)

The basis for geostatistical modeling is the variogram [13,14,17,18]. This is the mathematical description of the spatial autocorrelation (or spatial dependence) between a sampled value and its neighboring sampled values. Equation (4) shows the empirical variogram, $\gamma(h)$, which is a discrete variation based on the difference between sampled values separated by a distance h.

$$Z(B) = \sum_{i=1}^{N} \lambda_i Z_i \tag{4}$$

where Z_i is the observed value at location i, $Z(B)$ is the predicted value at a block, N is the number of pairs of observations, B is the block, and λ is the weight.

A variogram shows the spatial structure and variability of a variable over an area. High spatial dependence means that spatial similarity can be found by analyzing the sampled values using Equation (5):

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(\boldsymbol{x_i}) - Z(\boldsymbol{x_i} + h)]^2 \tag{5}$$

where $\gamma I(h)$ is the estimated variogram; $Z(\boldsymbol{x_i})$ and $Z(\boldsymbol{x_i} + h)$ are the observed values at locations $\boldsymbol{x_i}$ and $\boldsymbol{x_i} + h$; $N(h)$ is the number of observation pairs separated by distance h.

Variograms have three main parameters to consider when evaluating the spatial structure of samples: nugget (C_0), sill ($C + C_0$), and range (R) (Figure 4). The variance increases with distance and stabilizes at a constant value ($C + C_0$) at a given separation distance, the so-called range of spatial dependence (or range, R). The sill approximates the variance of the samples for stationary data. Samples that have a distance between them greater than the range are not spatially autocorrelated, because the variance is equal to a random variation with no spatial correlation. If the variogram reaches a plateau (sill) at a distance, the variable is stationary. If the variance increases continuously, without reaching a plateau, it indicates the presence of trend effects and non-stationarity. Under ideal conditions, the experimental variogram (from Equation (5)) should start at the origin (0,0); therefore, its variation should be equal to zero. However, usually, soil attributes in the real world have non-zero variance only when h tends to zero. This discontinuity at the origin is called the nugget effect and is represented by unexplained spatial variation (microvariability at a shorter distance than the shortest sampling distance) or purely random variance (such as measurement or sampling error).

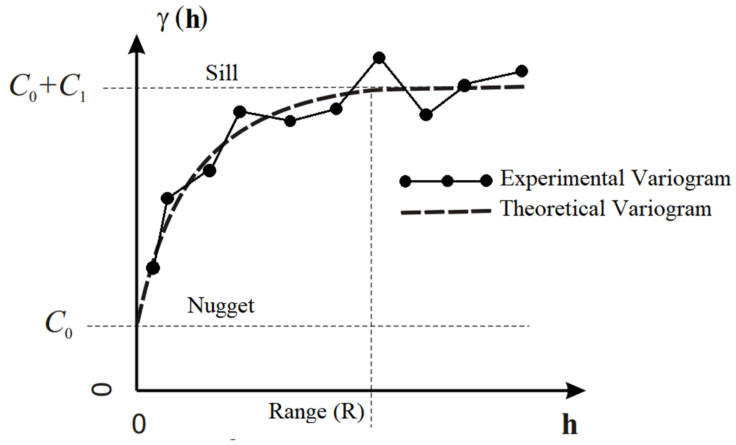

Figure 4. Example of a variogram model.

The experimental variogram must be calculated at different angles to check the existence of anisotropy. If there is no sign of anisotropy (different behaviors in different directions), an "omnidirectional" empirical variogram is calculated (usually at angle 0°) [17]. Then, a theoretical continuous model of the variogram is fitted over the discrete empirical variogram. The most common models are the spherical, Gaussian, exponential, power-law, and linear functions [36].

When using BCOK, the main difference consists in the calculation of the point-to-block covariance [35], according to Equation (1):

$$\overline{C}(B, \boldsymbol{x_i}) = cov(Z(B), Z(\boldsymbol{x_i})) = \int_B \frac{C(v, \boldsymbol{x_i})}{|B|} du' \quad (6)$$

where $\overline{C}$ is point-to-block covariance, $|B|$ is the volume of the block, and C is point-to-point covariance [35].

The punctual LMC for different combinations of variables requires regularization over the same block support (here, it is 10 m by 10 m by 0.2 m) by applying block cokriging (BCOK) over the selected block grid. The BCOK method can be understood as the summation of points in the block grid, and for this reason, the coarsest pixel size is the final spatial resolution. We considered the depth (0.2 m) of soil samples when designing the block grid. In other words, applying BCOK is a solution to the problem of change in support [10,33,37,38].

2.3. NDVI and SAR Mosaic Derivation from Sentinel-1 and -2

Remotely sensed data are quite often used as covariates in precision agriculture applications, since they can be correlated as proxies of key soil-forming factors [39–42]. Quite often used covariates are vegetation indices such as the normalized difference vegetation index (NDVI) [43–45], and recently, Synthetic Aperture Radar (SAR) was also correlated as a covariate for geostatistical modeling in water–soil sciences [46–48].

The *NDVI* from Sentinel-2 satellite imagery was used as the auxiliary variable for BCOK and was calculated using Equation (7):

$$NDVI = \frac{NIR - red}{NIR + red} \quad (7)$$

where NIR is the percent near-infrared reflectance (0.83 to 0.88 µm) and *red* is the percent red reflectance (0.64 to 0.67 µm). Both are bands from Sentinel-2 satellite imagery. The possible values of the *NDVI* range from −1 to 1, where the pixels with values higher than 0.6 and closer to 1 indicate dense vegetation and greater vegetative vigor [49].

SAR C-band regular imaging from the Sentinel-1 mission over the coffee plot during the year before harvesting was retrieved. Different from Sentinel-2 optical imagery, SAR images can penetrate clouds and, for this reason, are widely used for emergency detection during and after flooding and storms.

NDVI and SAR time series were retrieved using Google Earth Engine (GEE). GEE is an intrinsically parallel, high-performance computing service platform for large-scale spatial analysis using Google's computational capabilities to perform the processing of spatialized socio-environmental data, such as satellite imagery, deforestation, drought, disasters, diseases, and environmental protection [50–52].

2.4. Performance Evaluation

Different variogram models were evaluated, and for each, the values of mean error (*ME*) (Equation (8)), mean squared error (*MSE*) (Equation (9)), kriged reduced mean error (*KRME*) (Equation (10)), and kriged reduced mean squared error (*KRMSE*) (Equation (11)) were calculated

$$ME = \frac{1}{N} \sum_{i=1}^{N} (Z(x_i) - Z(B, \boldsymbol{x_i})) \quad (8)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Z(x_i) - Z(B, \boldsymbol{x_i}))^2 \quad (9)$$

$$KRME = \frac{1}{N}\sum_{i=1}^{N}\frac{Z(x_i) - Z(B, \boldsymbol{x_i})}{s} \quad (10)$$

$$KRMSE = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{Z(x_i) - Z(B, \boldsymbol{x_i}))}{s}\right]^2 \quad (11)$$

where $Z(x_i)$ is the value sampled at location i, $Z(x_i)$ is the predicted value for location i, N is the number of pairs of sampled and predicted values, and s is the standard deviation of the sampled values.

ME and *KRME* values close to zero indicate good model performance. *MSE* indicates good model performance when its value is lower than the variance of the sample values. *KRMSE* should be inside the range $1 \pm (2\sqrt{2})/\mathrm{N}$ [53].

Finally, the spatial dependence ratio (*DD*) [54] was calculated with Equation (12). According to Cambardella [54], this index can classify the dataset as indicating (a) strong spatial dependence, <25%; (b) moderate spatial dependence, 25 to 75%; and (c) weak spatial dependence, >75%.

$$DD = \frac{C_0}{C_0 + C} \times 100 \quad (12)$$

where C_0 is the nugget effect and $C_0 + C$ is the sill.

Finally, BCOK interpolation can have its local accuracy measured using a kriging standard deviation metric named interpolation variance ($S^2(x_0)$) [55]. $S^2(x_0)$ is the average of the squared differences between data values and the retained estimates. Yamamoto [55] presents this metric as Equation (13):

$$S^2(x_0) = \sum_{i=1}^{N}\lambda_i[Z(x_i) - Z(B, x_0))]^2 \quad (13)$$

where λ_i are the BCOK weights. This interpretation is valid only if all weights are positive or constrained to be such [55] .

All geostatistics analyses were preformed using the software Geovariances Isatis.neo 2023.08.01 (www.geovariances.com/en/software/isatis-neo-geostatistics-software/, accessed on 21 October 2023).

3. Results and Discussion

A total of 34 images from the Sentinel-2 dataset with TOA (Top-of-Atmosphere) atmospheric correction were imported. These images were captured between 1 June 2020 and 10 May 2021 and presented a cloud percentage lower than 20%. The average value of each pixel was used. A total of 11 images from the SAR datasets from the Sentinel-1 satellite were imported. These images were captured between 22 June 2020 and 29 April 2021.

Descriptive statistics of coffee yield, *NDVI*, and SAR datasets are shown in Table 1. Notably, the number of *NDVI* and SAR values ($N = 1757$ pixels) was 44 times larger than the coffee yield values ($N = 40$ points), indicating that covariates were exhaustively more abundant than the primary variable and also indicating that covariates were characterized as Big Data compared with the primary variable.

Coffee yield showed outliers in its upper values (when the distance between the average value and the maximum value is higher than double their standard deviation). High skewness indicates a non-normal distribution [16,35]. Therefore, all variables were transformed using Gaussian anamorphosis. The kurtosis value of coffee yield showed a leptokurtosis distribution, while the covariates showed a platykurtic distribution. Medium-to-high correlation existed between coffee yield measurements and either the *NDVI* ($\rho = -0.56$) or

SAR (ρ = 0.61). These high correlations indicate the reliability of using these remote sensing measurements for applying BCOK.

Table 1. Descriptive statistics for coffee yield and *NDVI*.

Attribute	N	Min	Mean ± SD	Max	Kurtosis	Skewness
Coffee yield (kg trees^{-1})	40	0.27	3.05 ± 2.60	16.34	−0.40	0.16
NDVI	1757	0.33	0.56 ± 0.058	0.64	8.0	−2.04
SAR	1757	−10.79	−8.27 ± 0.60	−6.79	3.95	−0.75

Figure 5 shows the values transformed with Gaussian anamorphosis (x-axis) against the original values (y-axis). The larger the dataset (Figure 5B,C), the lower the sinuosity of the Hermite polynomials on the extreme sides. Skewed distributions could be more controlled after this data transformation. One can see how different distributions could be squeezed into a Gaussian distribution centered on zero. Of course, this transformation was performed before calculating the experimental variograms and fitting their theoretical models, and it was back-transformed after finishing geostatistical modeling. Considering that several authors showed how data transformation improves variogram modeling [16,30,31,34,56–60], we did not perform tests without data transformation.

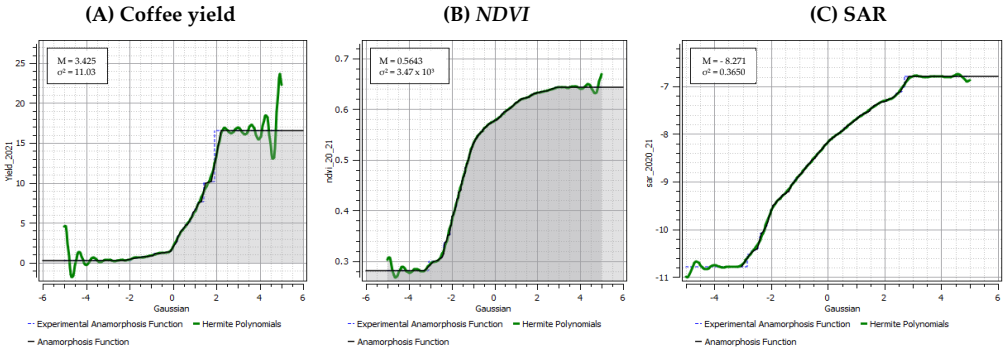

Figure 5. Gaussian anamorphosis of coffee yield (**A**), *NDVI* (**B**), and SAR (**C**).

Anisotropic variograms were calculated for directions separated by angular differences of 45° and 25° (not shown) and ± angular tolerance of 22.5°. No relevant anisotropy sign was found in the variograms. A relevant anisotropy sign can be found when the sill, range, and nugget are different in different directions. Isotropic variogram parameters are presented in Table 2. In the variogram fits (Table 2, Figure 6), the range increased when we used BCOK instead of OK. This indicates that the BCOK method estimates values for a larger range of yield. Because of this, it works for a larger number of neighbors than kriging, in spatial dependence, with greater spatial correlation. According to Gontijo et al. [61], variograms that present greater ranges indicate more spatially homogeneous fields. Also, DD increased from medium to very high values after adding covariates. The nugget value was generally small, even under low-sampling conditions. Here, we only show the interpolated maps and uncertainty from the BCOK application using the *NDVI* and SAR, since it showed the best evaluation metrics.

The variograms for BCOK application using the *NDVI* and SAR are shown in Figure 6. The cross-variogram with SAR shows negative cross-covariance, meaning that the primary and auxiliary variables were negatively spatially correlated, as expected because of the negative Pearson correlation coefficient.

Table 2. Parameters of isotropic variogram models for the coffee plot in May 2021.

Scenario	C_0	$C_0 + C$	R	DD (%)
OK	0.11	0.24	24.76	31.43
BCOK using *NDVI*	0.06	0.42	56.20	12.51
BCOK using SAR	0.07	0.61	59.78	10.30
BCOK using *NDVI* and SAR	0.01	0.70	74.98	2.78

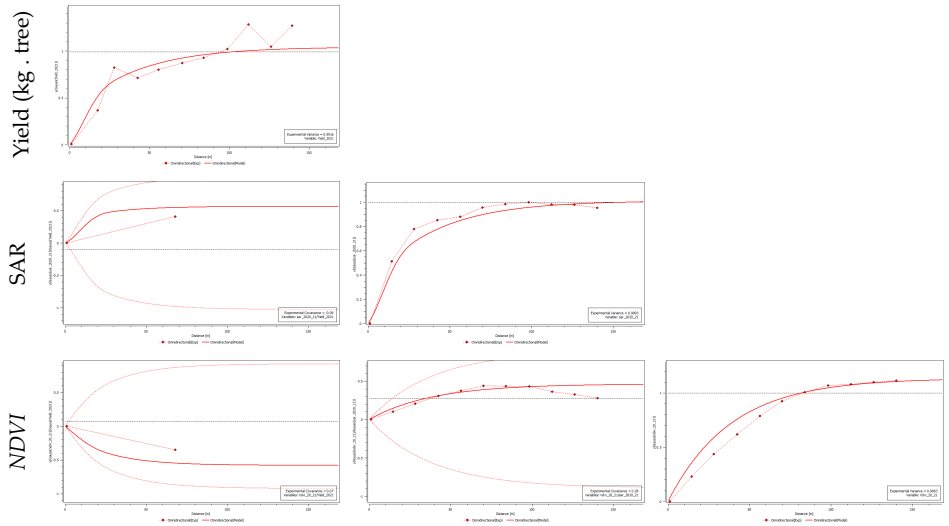

Figure 6. Linear model of co-regionalization (LMC) with experimental variograms and cross-variograms of transformed variables (dashed red with red points) and the theoretical continuous model of the variogram (red lines).

Direct relationships between coffee yield and remotely sensed measurements may not be captured by spatial autocorrelation because of the interactions among soil factors [13,61–65]. Only basing the analysis on the comparisons between OK maps and evaluating the Pearson correlation values may be not sufficient to provide insights about coffee yield spatial variability. In this manner, multivariate geostatistical analysis can be more useful in disclosing the spatial-scale-dependent correlations between remotely sensed data and yield and consequently provide better coffee yield maps if a relationship between the primary variable and covariates is found. Figure 7 shows the coffee yield map interpolated using BCOK with the *NDVI* and SAR.

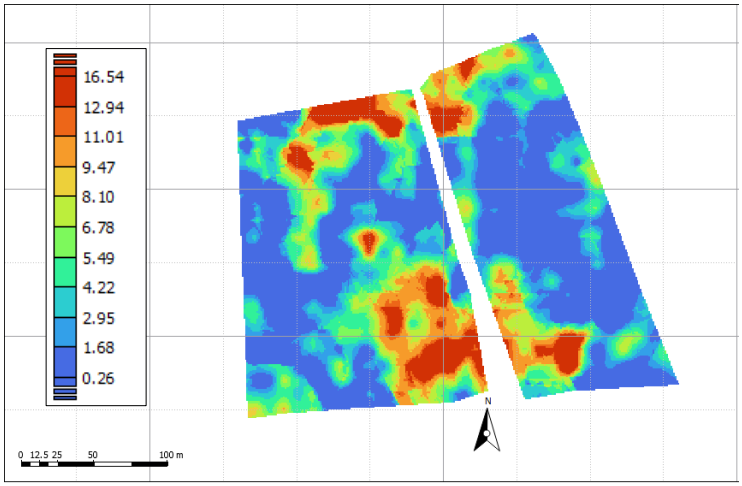

Figure 7. BCOK-interpolated map of coffee yield sampled in May 2021.

The assessment metrics are presented in Table 3. OK presented bad results, while BCOK showed better results, with *ME*, *MSE*, and *KRME* closer to zero and *KRMSE* close to one. This indicates an improvement when using BCOK. The final interpolated maps are not linearly dependent on the bivariate Pearson correlation coefficients but are highly dependent on the structural coefficients (sill, range, and nugget effect), as in any geostatistical analysis. In this sense, different metrics should be jointly analyzed to evaluate the results, because any analysis may be subjective if a single metric is used. In addition, the interpolation variance (Figure 8) showed uncertainty across the coffee plot, with an area in the southern part of the plot having a higher variance.

Table 3. Cross-validation evaluation with the transformed values.

Scenario	*ME*	*MSE*	*KRME*	*KRMSE*
OK	2.75	3.79	3.93	0.713
BCOK using *NDVI*	2.71	2.24	2.88	0.777
BCOK using SAR	2.61	1.91	1.86	0.884
BCOK using *NDVI* and SAR	1.11	1.01	1.12	0.984

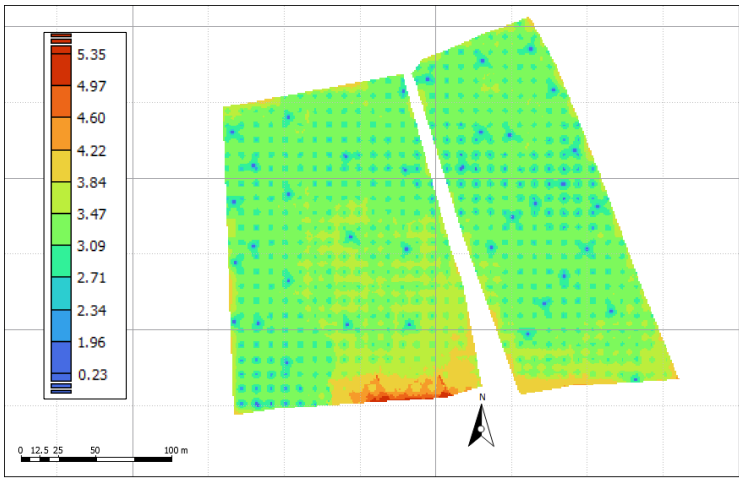

Figure 8. BCOK interpolation variance ($S^2(x_0)$) map of coffee yield sampled in May 2021.

The results shown here demonstrate how the application of a multivariate cokriging technique can be used to improve the estimation of coffee yield, even when impacted by outliers and biennial behavior. For this reason, when using a covariate in a model for improving the spatial interpolation of a primary variable, quite different combinations can be used if there is a high correlation between them is found.

Data fusion methods for interpolation are increasingly being used to enhance the estimation of agronomical attributes for PA applications, as the availability of accurate and large observation datasets is limited. This study investigated and assessed the performance of applying BCOK with a primary variable containing outliers. The case study was based on the spatial prediction of coffee yield in a plot in southeastern Brazil and was more accurate when the block cokriging (BCOK) method was used, with coffee yield as the main variable and the satellite-derived *NDVI* and SAR as auxiliary variables, compared with both ordinary kriging (OK) and BCOK but with *NDVI* or SAR only. BCOK application using the two covariates available achieved a smaller nugget with a larger range in the cross-variogram model compared with the direct variogram.

Improvement in coffee yield predictability using the combination of sampled data and remotely sensed data can be particularly advantageous when the observed dataset is small (less than 75 sampling points). However, careful preparation needs to be conducted for using multi-sensor datasets together, especially when different sources of data are based on different support sizes (spatial resolutions), shapes, and configurations. This

difference among supports is a classic problem in the area of spatial modeling [63–66]. This problem of change in support needs to be dealt with, because using data with different supports without regularization results in wrong results [10]. Here, we dealt with the support difference by applying the BCOK method for regularizing the variables to the same block support.

4. Conclusions and Recommendation

The study demonstrated that exhaustively remotely sensed Big Data variables can be incorporated into low-sampled experiments to improve their accuracy and effectiveness for practical use. Also, taking the support differences into account can help control uncertainty when incorporating Big Data for improving interpolation over low-sampled fields.

It can be concluded that assessing spatial variability cokriging can be used for precision agriculture applications as site-specific management. However, the use of multivariate geostatistics is limited by the correlation between covariates and the primary variable. In this sense, taking advantage of the current high availability of remotely sensed datasets over the Earth's surface and the new facility for retrieving large datasets is crucial to boosting computer applications in agriculture, thus consuming fewer resources and improving the environmental and financial management of farming.

Based on our results, we recommend considering using the BCOK approach for improving spatial interpolation for precision agriculture applications when sampled data present outliers and covariates with high correlation with the sampled data are available.

Author Contributions: Conceptualization, C.d.O.F.S.; methodology, C.d.O.F.S.; software, C.d.O.F.S. and R.L.M.; validation, C.d.O.F.S.; formal analysis, C.d.O.F.S.; investigation, C.d.O.F.S.; resources, C.R.G.; data curation, C.R.G.; writing—original draft preparation, C.d.O.F.S.; writing—review and editing, R.L.M.; visualization, C.d.O.F.S.; supervision, S.R.d.M.O.; project administration, C.R.G. and S.R.d.M.O.; funding acquisition, C.R.G. All authors have read and agreed to the published version of the manuscript.

Funding: This study was financed by the Research Coffee Consortium (in Portuguese, Consórcio Pesquisa Café) within the project coordinated by Embrapa (Seg number 10 18 20 01200000). This study was financed in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001.

Data Availability Statement: The data code and material are available from the corresponding author upon reasonable request.

Acknowledgments: Special thanks go to Josiane Moraes, the farm manager, and her team, who provided the experimental area and all the necessary support and infrastructure for the execution of field activities; to Luciano Vieira Koenigkan for providing a UAV image of the study area; and to Gustavo Costa Rodrigues and Cristina Aparecida Gonçalves Rodrigues for the discussions part of this project. We also thank the reviewers of this article for the excellent observations that enriched this work.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PA	Precision agriculture
R	Range
C_0	Nugget effect
DD	Spatial dependence degree
$C_0 + C$	Sill or structural component
OK	Ordinary kriging
BCOK	Block cokriging
LMC	Linear model of co-regionalization
$NDVI$	Coefficient of variation

SAR	Synthetic Aperture Radar
SD	Standard deviation
KRMSE	Kriged reduced mean squared error
KRME	Kriged reduced mean error
MSE	Mean squared error
ME	Mean error
Min	Minimum value
Max	Maximum value
N	Number of observations

References

1. Carvalho, L.G.d.; Sediyama, G.C.; Cecon, P.R.; Alves, H.M. A regression model to predict coffee productivity in Southern Minas Gerais, Brazil. *Rev. Bras. Eng. Agrícola Ambient.* **2004**, *8*, 204–211. [CrossRef]
2. Johnson, M.A.; Ruiz-Diaz, C.P.; Manoukis, N.C.; Verle Rodrigues, J.C. Coffee berry borer (*Hypothenemus hampei*), a global pest of coffee: Perspectives from historical and recent invasions, and future priorities. *Insects* **2020**, *11*, 882. [CrossRef] [PubMed]
3. Mayoli, R.N.; Gitau, K. The effects of shade trees on physiology of arabica coffee. *Afr. J. Hort. Sci.* **2012**, *6*, 35–42.
4. Sakai, E.; Barbosa, E.A.A.; de Carvalho Silveira, J.M.; de Matos Pires, R.C. Coffee productivity and root systems in cultivation schemes with different population arrangements and with and without drip irrigation. *Agric. Water Manag.* **2015**, *148*, 16–23. [CrossRef]
5. Souza, Z.M.d.; Marques Júnior, J.; Pereira, G.T.; Moreira, L.F. Variabilidade espacial do pH, Ca, Mg e V% do solo em diferentes formas do relevo sob cultivo de cana-de-açúcar. *Ciência Rural* **2004**, *34*, 1763–1771. [CrossRef]
6. Vieira, H.D. *Café Rural*; Interciência/FAPERJ: Rio de Janeiro, Brazil, 2017.
7. Camargo, A.P.; Camargo, M.B.P. Definition and outline for the phenological phases of arabic coffee under Brazilian tropical conditions. *Bragantia* **2001**, *60*, 65–68. [CrossRef]
8. Bernardes, T.; Moreira, M.A.; Adami, M.; Giarolla, A.; Rudorff, B.F.T. Monitoring biennial bearing effect on coffee yield using MODIS remote sensing imagery. *Remote Sens.* **2012**, *4*, 2492–2509. [CrossRef]
9. ISPAG. Precision Ag Definition. Available online: https://www.ispag.org/about/definition (accessed on 18 September 2023).
10. Silva, C.d.O.F.; Manzione, R.L.; Oliveira, S.R.d.M. Exploring 20-year applications of geostatistics in precision agriculture in Brazil: What's next? *Precis. Agric.* **2023**, *24*, 2293–2326. [CrossRef]
11. Juang, K.W.; Lee, D.Y. Comparison of three nonparametric kriging methods for delineating heavy-metal contaminated soils. *J. Environ. Qual.* **2000**, *21*, 197–205. [CrossRef]
12. Lloyd, C.; Atkinson, P.M. Assessing uncertainty in estimates with ordinary and indicator kriging. *Comput. Geosci.* **2001**, *27*, 929–937. [CrossRef]
13. Webster, R.; Oliver, M.A. *Geostatistics for Environmental Scientists*; John Wiley & Sons: Hoboken, NJ, USA, 2007.
14. Oliver, M.A.; Webster, R. *Basic Steps in Geostatistics: The Variogram and Kriging*; Springer: New Jersey, NY, USA, 2015.
15. Emadi, M.; Shahriari, A.R.; Sadegh-Zadeh, F.; Jalili Seh-Bardan, B.; Dindarlou, A. Geostatistics-based spatial distribution of soil moisture and temperature regime classes in Mazandaran province, northern Iran. *Arch. Agron.* **2016**, *62*, 502–522. [CrossRef]
16. Castrignanò, A.; Buttafuoco, G.; Quarto, R.; Parisi, D.; Rossel, R.V.; Terribile, F.; Langella, G.; Venezia, A. A geostatistical sensor data fusion approach for delineating homogeneous management zones in Precision Agriculture. *Catena* **2018**, *167*, 293–304. [CrossRef]
17. Goovaerts, P. *Geostatistics for Natural Resources Evaluation*; Oxford University Press: New York, NY, USA, 1997.
18. Goovaerts, P. Geostatistics in soil science: State-of-the-art and perspectives. *Geoderma* **1999**, *89*, 1–45. [CrossRef]
19. Yost, R.; Uehara, G.; Fox, R. Geostatistical analysis of soil chemical properties of large land areas. II. Kriging. *Soil Sci. Soc. Am. J.* **1982**, *46*, 1033–1037. [CrossRef]
20. Cressie, N. Fitting variogram models by weighted least squares. *J. Int. Assoc. Math. Geol.* **1985**, *17*, 563–586. [CrossRef]
21. Rhif, M.; Ben Abbes, A.; Martinez, B.; Farah, I.R. A deep learning approach for forecasting non-stationary big remote sensing time series. *Arab. J. Geosci.* **2020**, *13*, 1174. [CrossRef]
22. Hu, J.; Zhang, L.; Lee, C.; Gui, R. Advanced big SAR data analytics and applications. *Front. Environ. Sci.* **2022**, *10*, 2097. [CrossRef]
23. Rhif, M.; Abbes, A.B.; Farah, I.R. A Non-stationary NDVI Time Series with Big Data: A Deep Learning Approach. In Proceedings of the Conference of the Arabian Journal of Geosciences, Sousse, Tunisia, 25–28 November 2019; Springer: Cham, Switzerland, 2019; pp. 357–359.
24. Zhang, L.; Dong, J.; Zhang, L.; Wang, Y.; Tang, W.; Liao, M. Adaptive Fusion of Multi-Source Tropospheric Delay Estimates for InSAR Deformation Measurements. *Front. Environ. Sci.* **2022**, *10*, 213. [CrossRef]
25. Liu, H.; Yue, J.; Huang, Q.; Li, G.; Liu, M. A Novel Branch and Bound Pure Integer Programming Phase Unwrapping Algorithm for Dual-Baseline InSAR. *Front. Environ. Sci.* **2022**, *10*, 890343. [CrossRef]
26. Roznik, M.; Boyd, M.; Porth, L. Improving crop yield estimation by applying higher resolution satellite NDVI imagery and high-resolution cropland masks. *Remote Sens. Appl. Soc. Environ.* **2022**, *25*, 100693. [CrossRef]
27. Santos, H.G.; Jacomine, P.K.T.; Anjos, L.H.C.; Oliveira, V.A.; Coelho, M.R.; Lumbrelas, J.R. *Sistema Brasileiro de Classificação de Solos*; Centro Nacional de Pesquisa de Solos: Rio de Janeiro, Brazil, 2006.

28. Machado, R.D.; Bravo, G.; Starke, A.; Lemos, L.; Colle, S. Generation of 441 typical meteorological year from INMET stations-Brazil. In Proceedings of the IEA SHC International Conference on Solar Heating and Cooling for Buildings and Industry, Santiago, Chile, 4–7 November 2019.
29. Reboita, M.S.; Rodrigues, M.; Silva, L.F.; Alves, M.A. Aspectos climáticos do estado de Minas Gerais. *Rev. Bras. Climatol.* **2015**, *17*, 206–226. [CrossRef]
30. Castrignanò, A.; Buttafuoco, G. Data processing. In *Agricultural Internet of Things and Decision Support for Precision Smart Farming*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 139–182. [CrossRef]
31. Castrignanò, A.; Quarto, R.; Venezia, A.; Buttafuoco, G. A comparison between mixed support kriging and block cokriging for modelling and combining spatial data with different support. *Precis. Agric.* **2019**, *20*, 193–213. [CrossRef]
32. Castrignano, A.; Buttafuoco, G. Geostatistical stochastic simulation of soil water content in a forested area of south Italy. *Biosyst. Eng.* **2004**, *87*, 257–266. [CrossRef]
33. Journel, A.G.; Huijbregts, C.J. *Mining Geostatistics*; The Blackburn Press: Caldwell, NJ, USA, 1976.
34. Castrignanò, A.; Giugliarini, L.; Risaliti, R.; Martinelli, N. Study of spatial relationships among some soil physico-chemical properties of a field in central Italy using multivariate geostatistics. *Geoderma* **2000**, *97*, 39–60. [CrossRef]
35. Castrignanò, A.; Buttafuoco, G.; Quarto, R.; Vitti, C.; Langella, G.; Terribile, F.; Venezia, A. A combined approach of sensor data fusion and multivariate geostatistics for delineation of homogeneous zones in an agricultural field. *Sensors* **2017**, *17*, 2794. [CrossRef]
36. Bernardi, A.C.C.; Grego, C.R.; Andrade, R.G.; Rabello, L.M.; Inamasu, R.Y. Variabilidade espacial de índices de vegetação e propriedades do solo em sistema de integração lavoura-pecuária. *Rev. Bras. Eng. Agric. Ambient.* **2017**, *21*, 513–518. [CrossRef]
37. Chiles, J.P.; Delfiner, P. *Geostatistics: Modeling Spatial Uncertainty*; John Wiley & Sons: Hoboken, NJ, USA, 2012; Volume 713.
38. Armstrong, M. *Basic Linear Geostatistics*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1998.
39. Nussbaum, M.; Spiess, K.; Baltensweiler, A.; Grob, U.; Keller, A.; Greiner, L.; Schaepman, M.E.; Papritz, A. Evaluation of digital soil mapping approaches with large sets of environmental covariates. *Soil* **2018**, *4*, 1–22. [CrossRef]
40. Samuel-Rosa, A.; Heuvelink, G.; Vasques, G.; Anjos, L. Do more detailed environmental covariates deliver more accurate soil maps? *Geoderma* **2015**, *243*, 214–227. [CrossRef]
41. Siqueira, D.; Marques, J., Jr.; Pereira, G. The use of landforms to predict the variability of soil and orange attributes. *Geoderma* **2010**, *155*, 55–66. [CrossRef]
42. Wu, Z.; Wang, B.; Huang, J.; An, Z.; Jiang, P.; Chen, Y.; Liu, Y. Estimating soil organic carbon density in plains using landscape metric-based regression Kriging model. *Soil Tillage Res.* **2019**, *195*, 104381. [CrossRef]
43. Boudibi, S.; Sakaa, B.; Benguega, Z.; Fadlaoui, H.; Othman, T.; Bouzidi, N. Spatial prediction and modeling of soil salinity using simple cokriging, artificial neural networks, and support vector machines in El Outaya plain, Biskra, southeastern Algeria. *Acta Geochim.* **2021**, *40*, 390–408. [CrossRef]
44. Du, M.; Noguchi, N.; Ito, A.; Shibuya, Y. Correlation analysis of vegetation indices based on multi-temporal satellite images and unmanned aerial vehicle images with wheat protein contents. *Eng. Agric. Environ. Food* **2021**, *14*, 86–94. [CrossRef]
45. Pusch, M.; Samuel-Rosa, A.; Oliveira, A.L.G.; Magalhães, P.S.G.; do Amaral, L.R. Improving soil property maps for precision agriculture in the presence of outliers using covariates. *Precis. Agric.* **2022**, *23*, 1575–1603. [CrossRef]
46. Zeng, L.; Qingyun, S.; Guo, K.; Shuyun, X.; Herrin, J.S. A three-variables cokriging method to estimate bare-surface soil moisture using multi-temporal, VV-polarization synthetic-aperture radar data. *Hydrogeol. J.* **2020**, *28*, 2129–2139. [CrossRef]
47. Gururaj, P.; Umesh, P.; Sara, P.K.; Shetty, A. Top Surface Soil Moisture Retrieval Using C-Band Synthetic Aperture Radar Over Kudremukh Grasslands. In *Hydrological Modeling: Hydraulics, Water Resources and Coastal Engineering*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2022; pp. 31–42. [CrossRef]
48. Munda, M.K.; Parida, B.R. Soil moisture modeling over agricultural fields using C-band synthetic aperture radar and modified Dubois model. *Appl. Geomat.* **2023**, *15*, 97–108. [CrossRef]
49. Rouse, J., Jr.; Haas, R.; Schell, J.; Deering, D.; Harlan, J. Monitoring the vernal advancement and retrogradation (Green wave effect) of natural vegetation. In *NASA/GSFC, Type III Final Report*; NASA: Washington, DC, USA, 1974.
50. Duong, P.C.; Trung, T.H.; Nasahara, K.N.; Tadono, T. JAXA high-resolution land use/land cover map for Central Vietnam in 2007 and 2017. *Remote Sens.* **2018**, *10*, 1406. [CrossRef]
51. Kumar, L.; Mutanga, O. Google Earth Engine Applications. *Remote. Sens.* **2019**, 11, 591.
52. Tsai, Y.H.; Stow, D.; Chen, H.L.; Lewison, R.; An, L.; Shi, L. Mapping vegetation and land use types in Fanjingshan National Nature Reserve using google earth engine. *Remote Sens.* **2018**, *10*, 927. [CrossRef]
53. Adhikary, P.P.; Dash, C.; Bej, R.; Chandrasekharan, H. Indicator and probability kriging methods for delineating Cu, Fe, and Mn contamination in groundwater of Najafgarh Block, Delhi, India. *Environ. Monit. Assess.* **2011**, *176*, 663–676. [CrossRef]
54. Cambardella, C.; Elliott, E. Carbon and nitrogen dynamics of soil organic matter fractions from cultivated grassland soils. *Soil Sci. Soc. Am. J.* **1994**, *58*, 123–130. [CrossRef]
55. Yamamoto, J.K. An alternative measure of the reliability of ordinary kriging estimates. *Math. Geol.* **2000**, *32*, 489–509. [CrossRef]
56. Manzione, R.L.; Castrignanò, A. A geostatistical approach for multi-source data fusion to predict water table depth. *Sci. Total Environ.* **2019**, *696*, 133763. [CrossRef] [PubMed]
57. Manzione, R.L.; Silva, C.O.F.; Castrignanò, A. A combined Geostatistical approach of data fusion and stochastic simulation for probabilistic assessment of shallow water table depth risk. *Sci. Total Environ.* **2020**, *765*, 142743. . [CrossRef]

58. Buttafuoco, G.; Castrignanò, A.; Colecchia, A.S.; Ricca, N. Delineation of management zones using soil properties and a multivariate geostatistical approach. *Ital. J. Agron.* **2010**, *5*, 323–332. [CrossRef]
59. Shaddad, S.M.; Buttafuoco, G.; Castrignanò, A. Assessment and mapping of soil salinization risk in an Egyptian field using a probabilistic approach. *Agronomy* **2020**, *10*, 85. [CrossRef]
60. Buttafuoco, G.; Quarto, R.; Quarto, F.; Conforti, M.; Venezia, A.; Vitti, C.; Castrignanò, A. Taking into account change of support when merging heterogeneous spatial data for field partition. *Precis. Agric.* **2021**, *22*, 586–607. [CrossRef]
61. Gontijo, I.; Nicole, L.R.; Partelli, F.L.; Bonomo, R.; Santos, E.O.d.J. Variabilidade e correlação espacial de micronutrientes e matéria orgânica do solo com a produtividade da pimenta-do-reino. *Rev. Bras. Ciência Solo* **2012**, *36*, 1093–1102. [CrossRef]
62. Silva, S.d.A.; Lima, J.S.d.S.; Souza, G.S.d.; Oliveira, R.B.d.; Silva, A.F.d. Variabilidade espacial do fósforo e das frações granulométricas de um Latossolo Vermelho Amarelo. *Rev. Ciência Agronômica* **2010**, *41*, 1–8. [CrossRef]
63. Webster, R. Local disjunctive kriging of soil properties with change of support. *J. Soil Sci.* **1991**, *42*, 301–318. [CrossRef]
64. Rivoirard, J. *Introduction to Disjunctive Kriging and Non-Linear Geostatistics*; Clarendon Press: Oxford, UK, 1994.
65. Cressie, N.A. Change of support and the modifiable areal unit problem. *Geogr. Syst.* **1996**, *3*, 159–180.
66. Gelfand, A.E.; Zhu, L.; Carlin, B.P. On the change of support problem for spatio-temporal data. *Biostatistics* **2001**, *2*, 31–45. [CrossRef] [PubMed]