November 15, 2021

VI Brazilian Study Group
With Industry

CeMEAI
CEPID - Centro de Ciências Matemáticas Aplicadas à Indústria

impa
Instituto de Matemática Pura e Aplicada

ESALQ

Problem 2:
Spatial data interpolation

**José Lucas Safanelli; Afonso Paiva; Francisco Louzada;
Dinalva A. Sales; Vera Tomazella; Gustavo Buscaglia; Oilson Gonzatto;
Glauco Gonçalves; Saulo M. Mastelini; Antonio Castelo;
Samuel R. Silva; Thales Vieira; Guilherme Tubone; Enéas Mendes;Paulo V. A. B. Lisboa;
Jean M. A. Espinosa; Deivid C. L. Alves; Eder Brito; Karla Cruz
Ronaldo Cândido; Alexandre Delbem; Éder S. Brito; Marcos J. Henriques; Renato Silva**

Summary:

Massive data collection has been carried out in both information gathering and decision making in real-time. Due to the nature of the data type (spatial or spatio-temporal), treatment and interpolation become essential steps in this process. In this context, computer systems capable of generating thematic maps are used, through the interpolation of the sample data, to measure information for nonsampled locations. An important aspect is the choice of the interpolation method to be used since most of the interpolators do not preserve the original data, thus affecting the thematic map generated. This work aims to propose different alternatives to solve the problem proposed by Jhon Deere.

Contents

1 INTRODUCTION AND CHARACTERIZATION OF THE PROBLEM

In 1950 the world population was 2.5 billion. This number currently exceeds 8 billion of individuals. There is a 70% chance that in 2050 that contingent will exceed the 9.6 billion population and could reach 12.3 billion in 2100. Everything indicates that Nigeria will become the third most populous country, India will overtake China in the number of inhabitants, and Africa will become the most inhabited continent on the planet [12]. Among the primary needs for the survival of the human race is food. With this significant increase in population to come associated with socio-environmental, behavioural, and economic issues imposed by modern society, additional challenges have been created for agriculture and livestock in general [10].

To minimally meet the expected demand, world agricultural production must be increased by at least 70%. Developing countries will have to double their production to compensate for countries that will not increase them enough. Considering that all humans have a decent diet, to meet the 40% increase in Earth's population, the natural increase will have to be an additional billion tonnes in cereal production and 200 million tonnes in the supply of meat annually [5]. The FAO (Food and Agriculture Organization of the United Nations) still estimates that 70% of this food supplement should come from the use of technologies that improve efficiency, since few countries have areas to be cleared/deforested for implantation. of crops, in addition to the implications that this would generate in the contexts, current and future, of having a sustainable planet [24].

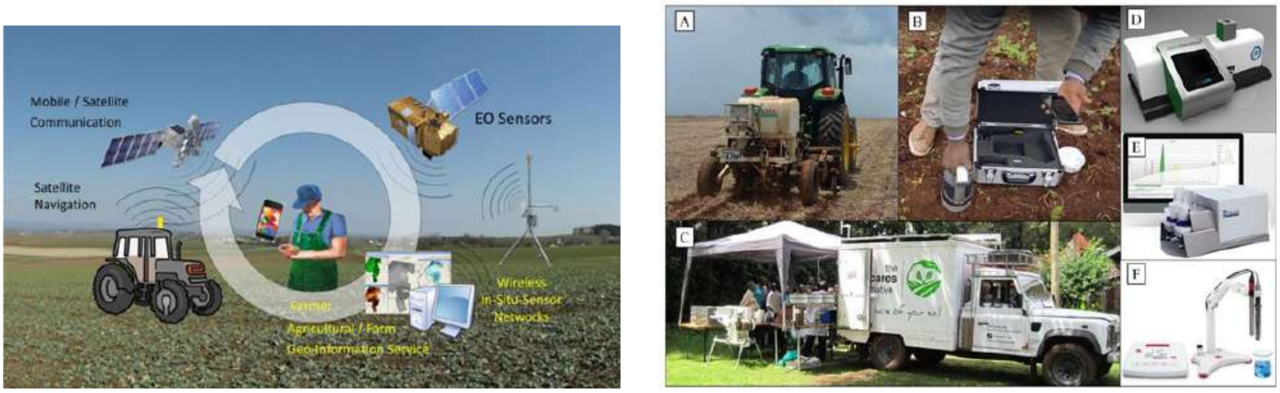
Thanks to scientific research, in the last decades, agricultural production has leveraged substantially in quantity and quality [30]. Concerning crops, we can mention as an example the culture of corn. From 1220 kg/ha in the 1977/1978 crop, the national average yield jumped to 3360 kg/ha in 2002/2003. There is a 2140 kg/ha increase in productivity in just 25 years. It is estimated that 50% of this increase was due to genetic improvement and the other half due to improvements in growing conditions, according to EMBRAPA researcher [27]; these results, derived from several factors, are provided mainly as a result of **precision agriculture** initiated in Brazil in the decade of 90. In the 2014/2015 crop, Brazilian corn production was 204.5 million tonnes, where 5.6% more was harvested compared to the previous crop, where the increase in areas cultivated with corn was only 1.1%. The total Brazilian Harvest of Grains 2020/21 grew by 6% compared to the 2019/20 crop, reaching 272.3 million tons [2]. In a recent interview, the current minister of agriculture, Tereza Cristina, points out that in the past 30 years, Brazil increased agricultural production by 425%, while in the same period, the cultivated areas increased by 43% [8]. What, undoubtedly, was greatly favoured by **precision agriculture**, given that this technique started to be used in Brazil, in this same period, with a gradual increase over the years. **Precision agriculture is a powerful tool for the farmer who, through his results, manages, for example, to correct the soil on his property to the point of significantly**

and substantially increasing his productivity. This technique has been increasingly refined by areas such as statistics, mathematics, computing, among others, through methodologies such as Kriging, Random Forest, Interpolations, Radial Basis Function, Moving Least-Squares among other methodologies.

In recent years, countries such as China, the United States, Brazil, India, Russia, France, Mexico, Japan, Germany and Turkey have emerged among the ten countries that most produce food on the planet. Among those that export most are the United States, Canada, France, Brazil, Italy, the Netherlands, China, Belgium, Germany and the United Kingdom. Currently, only the United States and the European Union export more food than Brazilian farmers and ranchers. Reports indicate that Brazil will be the largest exporter of food in the world in the next decade [1] [3], and for this reason, nowadays, it is already called the future breadbasket of the world.

In this way, all these advances that contribute to the increase and improvement of food production globally are mainly due to researchers and professors from the most diverse areas of knowledge. Moreover, indispensably, approaches and analyzes involving statistics, mathematics, and computation were essential for critical decision-making in this gigantic context. They go from diagnostic surveys to the elaboration of planned experiments and their analyses, based on theoretical statistics, creating interdisciplinary tools, using applied statistics, fundamental directions were taken to reach the current success of the various agricultural sectors that Brazil has. Furthermore, undoubtedly, the emergence and evolution of **precision agriculture** dictated direction and improvements in productivity, both quantitatively and qualitatively.

Agriculture has been impacted by different technologies and has generated the best possible expectations for the future of agribusiness. Faced with a scenario of constant population growth and increased demand for safe food, finding solutions to produce more in the same area is an excellent challenge for rural producers. Through equipment, software, machinery, and intelligent sensors, it has been possible to reduce environmental impacts while increasing the quality of crops (Figure 1).


Mobile / Satellite Communication
EO Sensors
Satellite Navigation
Farmer
Agricultural / Farm Geo-Information Service
Wireless In-Situ-Sensor Networks
A
B
C
D
E
F

(a) FONT: BACH & MAUSER, 2018

(b) FONT: MOLIN & TAVARES, 2019

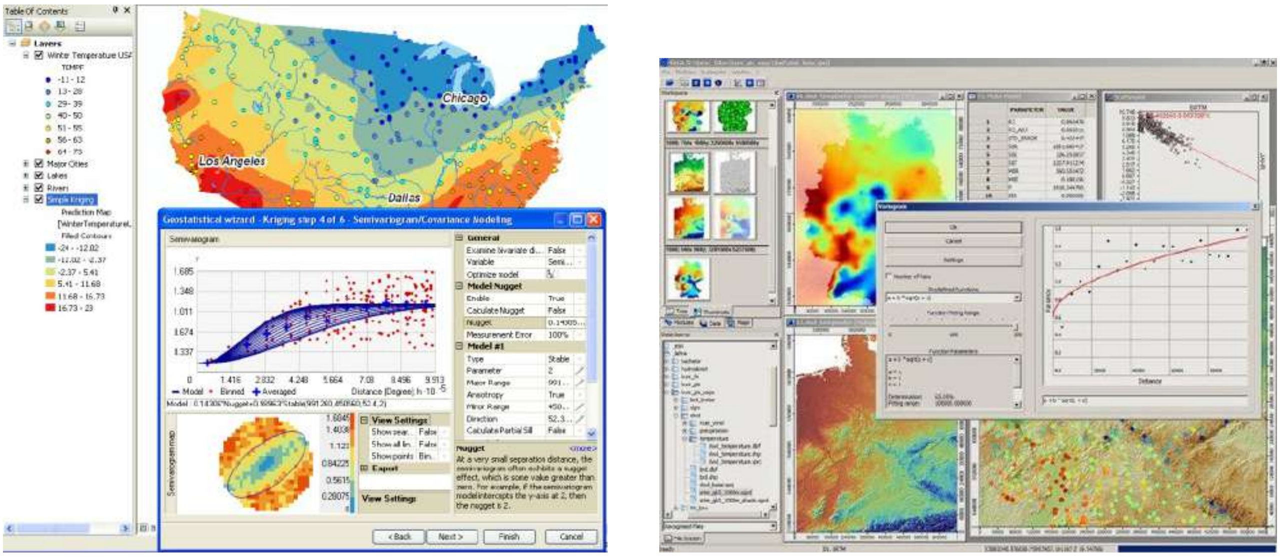
FIGURE 1: TECHNOLOGIES APPLIED TO AGRICULTURAL DEVELOPMENT

Agricultural productivity is influenced, among other factors, by the chemical attributes of the soil, that is, by its fertility. The soil is a mixture of minerals and organic compounds formed by physical, chemical, and biological agents initially on the primary rock. For plants, soils are a source of nutrients necessary for their development in addition to the fixation medium. Samples of soil and plant attributes are collected and analyzed to obtain information that serves as a subsidy for decision-making to correct the problems detected in the analyzed areas.

Farmers need accurate information on the spatial distribution of soil properties to support sustainable agricultural production systems. Conventional soil sampling methods and analysis are generally used to derive soil information at sampling sites, but these methods are expensive and require lengthy field campaigns. In addition, they are often unable to explain the wide spatial variation of soil properties. For soil mapping, the sampling density required is usually significant. Enormous costs arise from dense sampling schemes, which account for a substantial part of the total mapping costs. Consequently, professionals tend to decrease the number of samples collected, undermining the accuracy of soil maps.

In practice, however, soil data is often distorted [15] and may also contain outliers. Robust variogram estimation can prevent outliers from harming the experimental or sample variogram and, therefore, on the model parameters (see[21]).

Geostatistics tools allow the analysis of spatial dependence and data interpolation to be carried out in nonsampled locations, taking into account the spatial variability (Figure 2).



Geostatistical wizard - Kriging step 4 of 6 - Semivariogram/Covariance Modeling

Model : 0.14306*Nugget+0.98963*Stable(991.260,450.660,52.4,2)

General
Examine bivariate d... False
Variable Semi...
Optimize model
Model Nugget
Enable True
Calculate Nugget False
Nugget 0.14305...
Measurement Error 100%
Model #1
Type Stable
Parameter 2
Major Range 991...
Anisotropy True
Minor Range 450...
Direction 52.3...
Calculate Partial Sill False

View Settings
Show sear... False
Show all lin... False
Show points Bin...

Nugget
At a very small separation distance, the semivariogram often exhibits a nugget effect, which is some value greater than zero. For example, if the semivariogram model intercepts the y-axis at 2, then the nugget is 2.

(a) FONT: GEOSTATISTICAL WIZARD, ESRI/ARCGIS

(b) FONT: SAGA GIS

FIGURE 2: GEOSTATISTICS TOOLS FOR MANIPULATING SPATIAL DATA.

1.1 INFORMATION ABOUT THE DATA

The data from this experiment were previously used by [22]. Samples were collected at 458 sites on a regular grid of 100m × 100m at two depths (l = A: 0,0-0,2 m with 458 samples; B: 0,8-1,0 m with 452 sample), representing a total of 910 georeferenced samples (Figure 3)

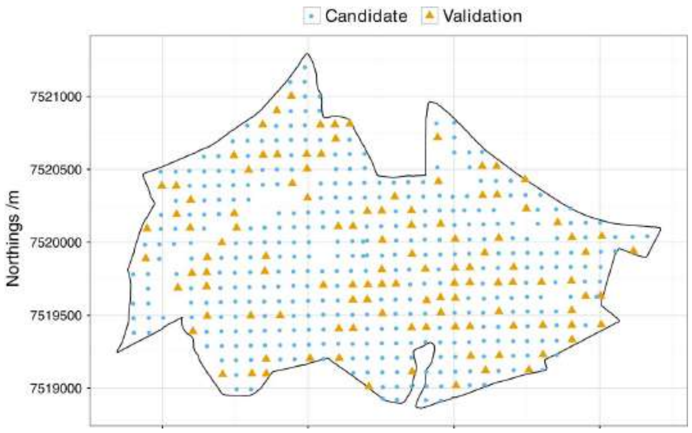


FIGURE 3: STUDY AREA SHOWING SAMPLE LOCATIONS USED FOR VALIDATION AND CANDIDATE SAMPLES FOR CALIBRATION.

The sample design was not initially planned for geostatistical mapping; therefore, this grid may not consider spatial dependence over short distances. The collection area, close to the municipality of Barra Bonita in the state of São Paulo (Brazil), covers 473 ha at altitudes ranging from 550 to 710 m above sea level. A small watershed runs from northeast to southwest where sandstones dominate, but basaltic flows also occur to a lesser extent. The predominant soil types are classified as Typical Quartzipsamment (TQ), Typical Hapludox (TH), Typical Hapludalf (THa), Typical Hapludult (Thu) and Typical Eutrudept (TE).

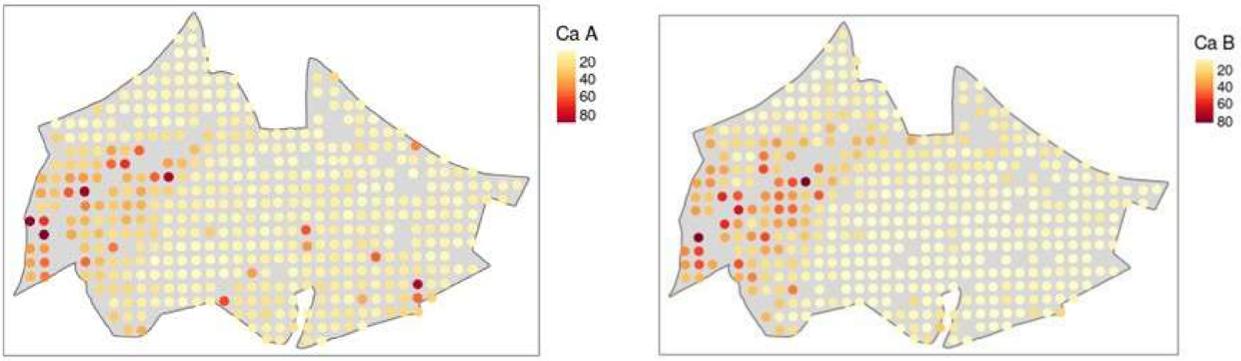
For each georeferenced point, the calcium content of the samples is reported (Ca++ in $mmol_{c}kg^{-}1$) and fractions of sand, silt, and clay granules (percentages). The studies carried out in this report include the evaluation of different regression and interpolation strategies derived from statistics, mathematics, and computation for data spatialization.

For the formulation of this report, we focused on data related to the calcium concentration of the samples. Figure 4 presents the concentration information for the two depths A and B.

1.2 PROBLEMS AND PROPOSALS OF THIS WORK

The problems that arise are:

- How to make the best decision in the process of constructing the semi-variogram and kriging interpolation in an automated and efficient/scalable way?
- Are there alternative (non-parametric) interpolators that provide similar results? (Mean value and uncertainty (variance))



(a) A: 0,0-0,2 M

(b) B: 0,8-1,0 M

FIGURE 4: CALCIUM CONCENTRATIONS SAMPLED AT DEPTHS A AND B.

Proposal for solving problems:

a) Kriging-Based Formulation

- Model problem using the library GeoStats.jl;
- Search for an attempt to optimize kriging techniques with QGIS e ArcGIS;
- Other alternatives to kriging techniques (Stochastic Partial Differential Equations).

b) Hybrid Formulation

- Using meshless interpolation schemes (Radial Basis Functions, Moving Least-Squares);
- Pros: complexity $O(n^3)$, non-parametric model;
- Reintroduce uncertainty to the problem (Quantile Regression);
- Solution via Random Forest or Maximum Entropy approach.

c) Presentation of the data

- Responsive GIS system for the web with colour palettes and layers of compounds and their respective errors.

The accuracy with which spatial distribution maps are produced in the interpolation process is an essential factor in their use.

2 METHODOLOGY

2.1 KRIGING

In general, the geostatistical methodology seeks to extract, from the apparent randomness of the collected data, the probabilistic structural characteristics of the regionalized phenomenon, a correlation function between the values located in a particular neighbourhood and direction in the sampled space. The primary estimation method used is that of kriging.

Kriging is a process of estimating by means mobile, variable values are distributed in space from adjacent values, while considered interdependent by a function called variogram. As in calculating this function, the sum of differences squared is divided by two multiplied by the number of pairs of values; the correct term would be semi-variogram. However, it is usual to use the term more synthetic variogram.

Kriging uses information from the variogram to find the optimal weights associated with samples with known values that will estimate unknown points. In this situation, the method provides the error associated with such estimation in addition to the estimated values, which distinguishes it from other interpolation algorithms. It is understood as a series of regression analysis techniques that seek to minimize the estimated variance from a previous model, which considers the stochastic dependence between data distributed in space. Among the estimation methods commonly used, the geostatistical method of kriging can be considered the best linear estimator without bias, whose objective is to minimize the estimate's variance.

The most common forms are simple kriging and ordinary kriging, and among nonlinear methods, indicative kriging stands out. Simple kriging is used when the mean is assumed to be statistically constant for the entire area. Ordinary kriging, in turn, considers the floating or moving average over the entire area. Indicative kriging is a predictor that uses the technique of Ordinary Kriging or Simple Kriging of the transformed data employing a binary nonlinear function composed of 0 and 1.

One of the significant advantages of Indicative Kriging lies in the fact that it is a nonparametric estimator that allows transforming qualitative variables (presence or absence) or quantitative variables (according to a cut-off point of interest) and estimating the probability of occurrence of the variable.

The geostatistical model, as presented, among several others, in Diggle & Ribeiro Jr. (2007) [9] can be defined as a random-effects model. Let us consider specifying the model for Gaussian responses, although the same structure is valid for other distributions for the observable variable $Y(.)$.

$$[Y|b, D] \sim N(\mu, I\tau^2)$$

$$\nu = D\beta + Zb$$

$$b \sim NMV(0, \Sigma_b),$$

where D is a matrix of known covariables with the vector of linear parameters $beta$ associated with them, as in a usual linear regression model. We associate a random effect $S(x)$ to each position, denoted by defining an identity matrix in $Z = diag(1)$ and a vector b, both of size equal to n, the number of observations. Therefore, the geostatistical model can be interpreted as a random intercept model with a value at each location.

This model that deserves more attention is the Σ_b matrix that describes the structure of the spatial dependence between the observations. Typically, the elements of this matrix are given by a correlation function whose argument is the distance between each pair of observations.

The estimate is made to determine an average value in an unsampled location in the basic kriging process. However, it is also possible to make estimates based on below or above a certain cut-off level.

2.1.1 KRIGING WITH QGIS AND ARCGIS

ArcGIS is a commercial geographic information system for working with maps and geographic information maintained by the Environmental Systems Research Institute with several features presented for online use and mathematical/statistical methods for data science analysis. QGIS is a free and open geographic information system that provides functions for creating, editing, analyzing, and geospatial publications.

Both codes have tools for kriging geospatial data implemented in specific packages. QGIS was used with the *Kriging* technique with the *ordinary* method and *spherical* semivariogram whose results are a Gaussian kriging with a distance of 4 meters and using 12 neighboring points of variable radius for geostatistics (see figure5).

A resolution of 2 meters was also tested. However, the processing time increased considerably, and the errors maintained values similar to the kriging of 4 meters. In this sense, the choice of kriging with 4 meters satisfied the challenge's objective that brought together the requested spatial resolution and the low data processing time.

2.1.2 SOIL GRANULOMETRY

Soils contain particles of different sizes, where certain intervals are given specific names like clay, silt, and sand. This size is defined by the diameter of the particles, considering them as if they were spheres.

According to EMBRAPA's manual of soil analysis methods (2017), [26] we have to

> *Particle size analysis aims to quantify the distribution by size of individual particles of soil minerals. Individual particles are understood as individualized mineral grains, unaltered or partially altered rock fragments (which may contain more than one mineral), cemented concretions, nodules, and similar materials, as defined by the Soil Science Vocabulary (materials that cannot be disaggregated except by high energy applications such as hammer blow).*

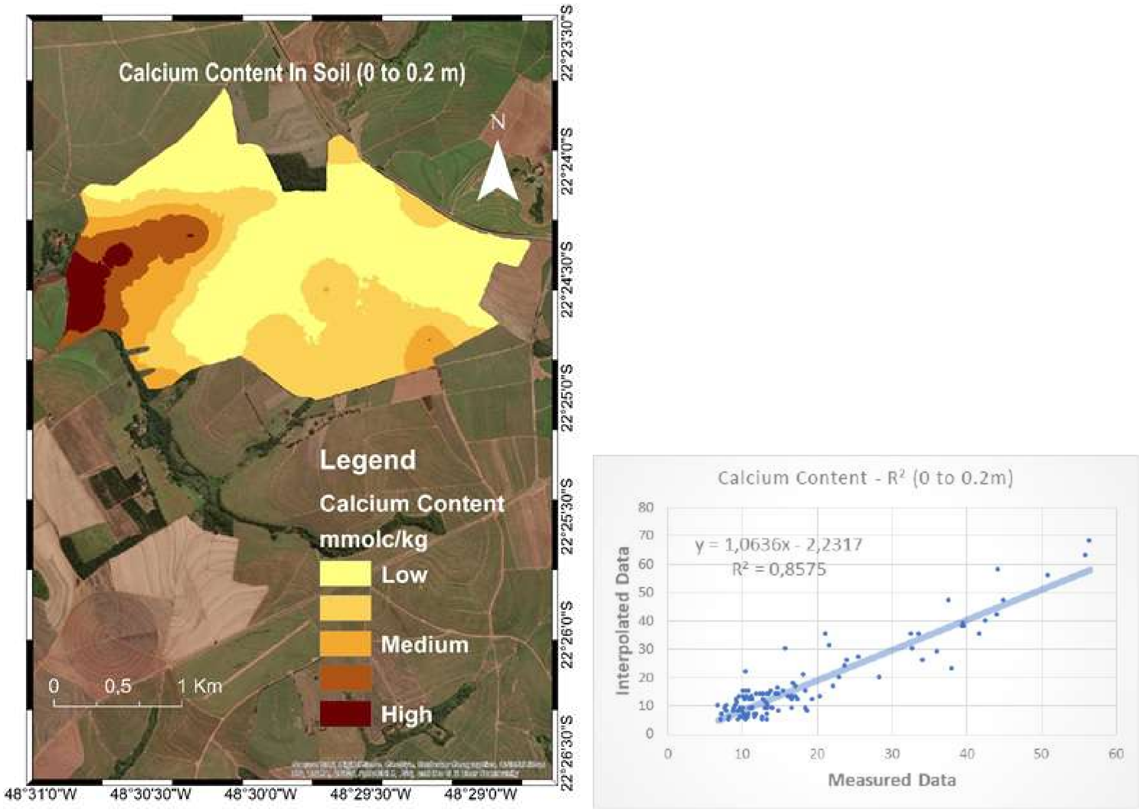


FIGURE 5: KRIGING USING ARCGIS WITH 4 METER RESOLUTION (LEFT). KRIGING VALIDATION USING THE SAME VALIDATION POINTS AS THE ARTICLE (RIGHT).

There are various forms of particle size scales, and the scales were proposed in 1922 by geographer Chester K. Wentworth being smaller sizes like clay with 0.004mm to boulder, which would be larger than 256mm. In this work, we adopted the sizes described in the EMBRAPA granulometry manual (NBR 7181 of 09/2016), they are:

- Clay (smaller than 0.0015 mm);
- Silt (greater than 0.004mm);
- Sand (greater than 0.4mm).

The soil map follows the following equation to estimate the average value of the grain in each pixel it stands:

$$\textit{Grain Size} = \frac{(sand \cdot 0.4) + (silt \cdot 0.004) + (clay \cdot 0.0015)}{100}. \tag{1}$$

Thus, it is possible, via granulometry, to show the predominance of the type of soil on the average value of the grain.

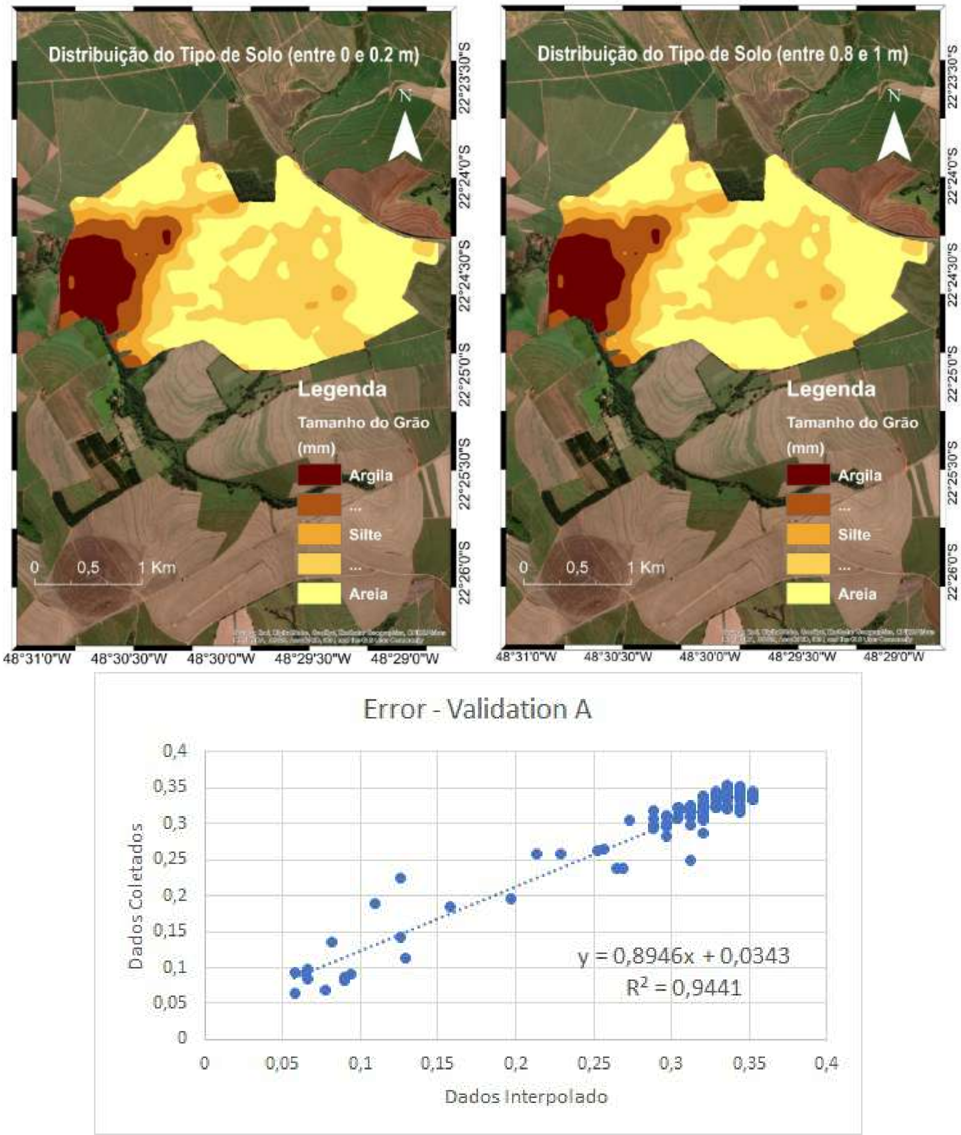


FIGURE 6: STATIC IMAGE OF SOIL GRANULOMETRY FOR POINTS WITH AN ALTITUDE OF 0.2 M (LEFT) AND 1 M (RIGHT).

2.2 GEOREFERENCED INFORMATION SYSTEMS (GIS)

One of the objectives of this work is to present the results generated by the analyzes practically and straightforwardly for interpretation for the end-user. In this sense, interactive thematic map sharing tools will be used, known as Georeferenced Information Systems (SIG) in WEB or WebSIGs.

> The SIG is a system that can locate objects in space or on a map, in which different information is stored in a database and gathered based on similar characteristics, managing them as georeferenced thematic informative layers. Such a system was mainly designed to study geographic analysis integrated into a world computer hub (Longley et al., 2013).

One of the first definitions of GIS in the digital era was introduced in 1986 by Peter A.

Burrough, professor at Oxford University (UK): “GIS consists in a series of software instruments made to acquire, store, extract, elaborate and visualize spatial data in the real world”. The SIG is an integrated system that puts geographic information concerning other information in a database system georeferenced thematic of informative layers. We took advantage of the webSIG user-friendly interface to show worth information for our final user in this work. Therefore, the webSIG is IT-based system chosen to translate our sophisticated data modelling results to the final user.

Our final results are available on the following website https://inoventerfurg.wixsite.com/inoventer at “desafios científicos” bottom.

2.3 STOCHASTIC PARTIAL DIFFERENTIAL EQUATION (SPDE)

Data collected in space or time is often obtained to elicit an underlying pattern. In time, space and many other domains imply the same idea: quantities that occur closer to each other are more similar than those that are more distant. A popular statistical model that represents this idea is the SPDE [16].

The SPDE approach introduced by [16] and implemented in the R-INLA software package [23] is a flexible and efficient method for analyzing this type of data. SPDE can be interpreted as a smoothing penalty.

Consider the following modelling. Let $z(x)$ a random variable at the location x or time x depending on the domain. A statistical model for Z is built on three components:

$$z(x) = \eta(x) + f(x) + \epsilon(x),$$

where

- $\eta(x)$: it is the fixed effect, usually a linear combination of covariates observed with unknown parameters;

- $f(x)$: it is a stochastic process, representing the structured dependence between observations, observations obtained closer to each other in time or space are more likely to be similar than those more distant;

- $\epsilon(x)$: represents the measurement error or unstructured error, often $\epsilon(x) \sim$ N$(0, \sigma^2)$, for the unknown parameter σ and location x.

A mathematically convenient and flexible model for this component is a Gaussian Process (GP) with mean equals 0 and covariance function $c(x_i, x_j) =$ Cov$(f(x_i), f(x_j))$. The covariance function quantifies how related two f values are in two locations.

In general, an SPDE states that the differential of a f function is equal to some known stochastic process, most commonly the white noise process ϵ. SPDE states that $Df = \epsilon$ for some differential operator D. A stochastic process f is called a solution for SPDE if it satisfies this equation. The SPDE $Df = \epsilon$ solution has a covariance structure that is induced by choice of D. This means that one could describe a system using an SPDE and then deduce the associated covariance function.

Particularly, the solution of the stochastic differential equation

$$(\kappa^2 - \Delta)^{(\nu+1)/2}(\tau x(s)) = \mathcal{W}(s), \qquad s \in \Omega, \tag{2}$$

is a stochastic GP, whose correlation structure is equal to Matérn's correlation function. This is a family of functions that is appropriate for the possibility of many behaviors [25, 20] and whose form is given by

$$c(x_i, x_j) = \frac{2^{1-\nu}}{(4\pi)^{d/2}\kappa^{2\nu}\tau^2\Gamma(\nu + d/2)}(\kappa||x_i - x_j||)^\nu K_\nu(\kappa||x_i - x_j||),$$

where

- ν, τ and κ are parameters that refer, respectively, to the smoothness of the field, the variance and the length of the correlation;
- K_ν is the modified Bessel function of the second kind, and;
- d is the dimension of the domain.

Once the Equation (2) is discretized, we transform the infinite GP into a finite-dimensional GP. We arrived at the linear system,

$$Ax(s) = B\xi \qquad \Longleftrightarrow \qquad x(s) = A^{-1}B\xi,$$

whose precision matrix is $Q = A^\top B^{-\top} B^{-1} A$. Additionally, when using finite elements to solve, we obtain A sparse and B diagonal, resulting in objectivity and efficiency in the computational process.

Given these considerations, when establishing an appropriate method to capture the space-time dependency relationship between points. We continue with a usual statistical modelling process, which involves a series of other considerations, such as:

- **Nature of our response variable** strictly positive, counts, proportions, autocorrelated;
- **Quantification of the association** with other observed effects;
- **Validation of model assumptions**;
- **Plausibility of the interpretation**, and subsequent exposure of results.

In particular, the adjustment of the models was carried out by the Bayesian approach with the Integrated Nested Laplace Approximation (INLA) program [7, 16].

> *"The power of the SPDE approach is realised by doing the opposite: find a D that induces the covariance function that you want. The power of finding the SPDE corresponding to the desired covariance function is that the precision matrix can be efficiently computed using the SPDE."* [19, p. 5–6].

2.3.1 INLA

INLA is a deterministic algorithm for Bayesian Inference, proposed by H. Rue, Martino and Chopin (2009) [23], which was developed primarily to be applied to a class of statistical models, known as latent Gaussian models.

Latent Gaussian models are such that their latent field is Gaussian, controlled by some hyperparameters and can also assume non-normality in the response variable. Such models encompass an extensive range of regression models with an additive structure. In most situations, the marginal posteriori distributions do not have closed-form and computational methods to obtain these distributions.

Rue, Martino and Chopin (2009) [23] point out that INLA presents a primary advantage compared to MCMC methods, especially within the context of latent Gaussian models, such methods are not an appropriate tool for routine analysis as they may have a very high computational time or even not achieve convergence. At this point, INLA is an alternative to overcome these limitations. The core of the INLA approach occurs via the Laplace approach.

In the INLA methodology, the interest is to obtain a posteriori approximations of the marginals, both latent variables and the hyperparameters of the latent Gaussian model.

To describe the models that the INLA can adjust, consider a vector $\mathbf{y} = (y_1, \ldots, y_n)$ of n observations. In addition, these observations will have an associated probability (not necessarily from the exponential family).

The mean μ_i of y_i observations can be linked to the linear predictor η_i using a convenient function. Observations will be independent, given their linear predictor, that is,

$$\eta_i = \alpha + \sum_{j=1}^{n_\beta} \beta_j z_{ji} + \sum_{k=1}^{n_f} f^{(k)}(u_{ki}) + \varepsilon_i \qquad \text{with} \qquad i = 1, \ldots, n,$$

where α is the intercept and β_i ($j = 1, \ldots, n_\beta$) are the coefficients of covariates $\mathbf{z}_i = \{z_{ji}\}_{j=1}^{n_\beta}$, the functions $f^{(k)}$ ($k = 1 \ldots, n_f$) define the n_f random effects in the vector of covariates $\mathbf{u} = \{u_{ki}\}_{k=1}^{n_f}$. Finally, ε_i is the random error.

Within the scope of INLA, the latent effects vector $\mathbf{x}$ will be represented as follows:

$$\mathbf{x} = (\eta_1, \ldots, \eta_n, \alpha, \beta_1, \ldots).$$

2.4 RANDOM FOREST

The Random Forest (RF) is a supervised machine learning algorithm used for classification and regression problems. An RF uses a set of independent decision or regression trees, which are grown from a random sample of the original observations. The number of trees (n_{tree}) and the number of variables to be considered at each new node (m_{try}) are the essential free parameters in the training process of a Random Forest [14].

The training algorithm of a Random Forest (both for regression and classification) is given by [6]:

1. Extract n_{tree} independent random bootstrap samples from the original dataset.

2. For each sample, we grow a tree. A new leaf is added to the tree choosing one between a random set of m_{try} predictors. It is chosen the predictor that offers the smaller prediction error.

3. The process of growing new leaves to an existing leaf ends when a pre-defined quantity of observations is associated with this leaf. In order to control overfitting.

A prediction of an RF is obtained by submitting a new sample record to each tree of the trained RF. In a tree, the values of the observed variables of this record are used to go down the nodes until a leaf is selected. The final response of the RF is given by aggregation of the responses of the n_{tree} trees. In a classification problem, the prediction is made by vote, whereas in a regression problem, as the one studied in this report, it is used as the average of the n_{tree} responses.

Formally, following the formulation of Breiman Breiman [4] and Meinshausen [18], let B be the hyperspace of the predictor variables X and $T(\theta)$ a tree grown over a sample X_i, and θ is a vector of parameters determining as the tree was grow (i.e., the variables selected at each partition). Therefore, each leaf $l = 1, \ldots, L$ of a tree is associated with a rectangular subspace $R_l \subseteq B$. For each $x \in B$, there is only one leaf $l(x, \theta)$ in the tree $T(\theta)$ such that $x \in R_l$ (this is the selected leaf when we submit a record x to the tree).

The prediction of one tree $T(\theta)$ for a new record x is obtained averaging the observed values in the leaf $l(x, \theta)$. Let the vector of weights $w_i(x, \theta)$ a positive constant if X_i (the independent set of samples used to grow the tree $T(\theta)$) is part of a leaf $l(x, \theta)$ and 0 if not. These weights sum up one, thus

$$w_i(x, \theta) = \frac{1_{\{X_i \in R_{l_{(x,\theta)}}\}}}{\#\{j : X_j \in R_{l_{(x,\theta)}}\}}.$$

The prediction of one tree, given a record x, is the weighted average from the original observations Y_i,

$$\hat{\mu}(x) = \sum_{i=1}^{n} w_i(x, \theta) Y_i.$$

For a Random Forest, the conditional expectation $E(Y|X = x)$ is approximated by the average prediction of n_{tree} trees, each one grown with an i.i.d. vector $\theta_i = 1, \ldots, n_{tree}$. Let $w_i(x)$ be the average of $w_i(x, \theta)$ over a collection of trees,

$$w_i(x) = \frac{\sum_{t=1}^{n_{tree}} w_i(x, \theta_t)}{n_{tree}}.$$

Finally, the Random Forest approximation $\hat{\mu}_{RF}(x)$ is given by

$$\hat{\mu}_{RF}(x) = \sum_{i=1}^{n} w_i(x) Y_i$$

2.5 RADIAL BASIS FUNCTION INTERPOLATION (RBF)

Given a set of points $\{x_i\}_{i=1}^{N} \subset \mathbb{R}^3\}$ and a set of values $\{f(x_i)\}_{i=1}^{N} \subset \mathbb{R}\}$, the goal is to find an interpolating function $s(x_i) = f_i, i = 1, ..., N$. The interpolant used will be as follows:

$$s(x) = p(x) + \sum_{i=1}^{N} \alpha_i \phi(\|x - x_i\|) \, , \tag{3}$$

where x_i are the called **centers** of the RBF, $p(x)$ is a low-order polynomial, α_i are real coefficients, $\| \cdot \|$ is the Euclidean norm on $\mathbb{R}^3$ and $\phi(x)$ is a *Radial Basis Function* (RBF).

Another required condition is the orthogonality of polynomials

$$\sum_{i=1}^{N} \alpha_i p(x_i) = 0. \tag{4}$$

The equations above can be written in the matrix form as

$$\begin{bmatrix} A & P \\ P^T & O \end{bmatrix} \begin{bmatrix} \alpha \\ c \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}, \tag{5}$$

where

$$A = \begin{bmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1N} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{N1} & \phi_{N2} & \cdots & \phi_{NN} \end{bmatrix}, \tag{6}$$

with $\phi_{ij} = \phi(\|x_i - x_j\|), i, j = 1, 2, ..., N$ and

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p1k \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{Nk} \end{bmatrix}, \tag{7}$$

with $p_{ij} = p_j(x_i), i = 1, 2, ..., N, j = 1, 2, ..., k$, whereas O is a null array of dimensions $(k \times k)$.

As possible choices for the RBF kernel, we initially tested the linear kernel $\phi(r) = r$. Then we solve the positive-definite linear system (5) to determine the coefficients α and c that can be replaced in the interpolating function (3).

In general, direct methods for solving the linear system (5), such as the Cholesky method, take a long time to run and therefore limit the size of the initial set to a few thousand points. In the next section, a quick method to work around this problem will be shown.

2.5.1 ADAPTIVE RBF PARTITION OF UNITY

A major problem of the solution described above is its global characteristic. That is, after having obtained the interpolating function, to evaluate a point in this function, it is necessary to perform many operations since we will have a coefficient for each center of the RBF.

Partition of Unit is a method that makes partition of the function domain into several subdomains and calculates an interpolant for each generated subdomain. Thus, when it is necessary to evaluate a point in the function, the result is a mixture of interpolants from a neighbourhood close to that point. To perform the domain partitioning, the quadtree data structure is used.

This method was used by Ohtake et al. [29] using LS as a local approximation and also with RBF [28].

2.5.2 PARTITION OF UNITY

Let us consider a bounded domain Ω on a Euclidean space and a set of functions ϕ such that

$$\sum_i \Phi_i = 1,$$

where each Φ_i refers to a subdomain of Ω.

Assuming that we have s_i functions that locally approximate the points belonging to each subdomain, then our global function that interpolates the points is given by

$$f(x) \approx \sum_i \Phi_i s_i(x).$$

Given a set of nonnegative functions w_i with compact support such that

$$\Omega \subset \bigcup_i supp(w_i),$$

the partition of the unit can be generated by

$$\Phi_i(x) = \frac{w_i(x)}{\sum_{j=1}^n w_j(x)}.$$

In addition, we use a quadratic B-spline $b(t)$ to generate the weight function

$$w_i(x) = b\left(\frac{3|x-c_i|}{2R_i}\right),$$

where c_i is the center of the RBF and R_i is the radius of the support.

2.5.3 RBF- INVERSE MULTIQUADRIC (IMQ)

Another type of RBF explored in this challenge is the one that uses the Inverse Multiquadric (IMQ) kernel function, given by the expression

$$\phi(r) = \frac{1}{\sqrt{r^2 + \epsilon^2}},$$

where the ϵ parameter is usually called *shape parameter*. Intuitively, the *shape parameter* tells us how far a single training sample influences the RBF interpolation, this can be seen according to Figure 7.

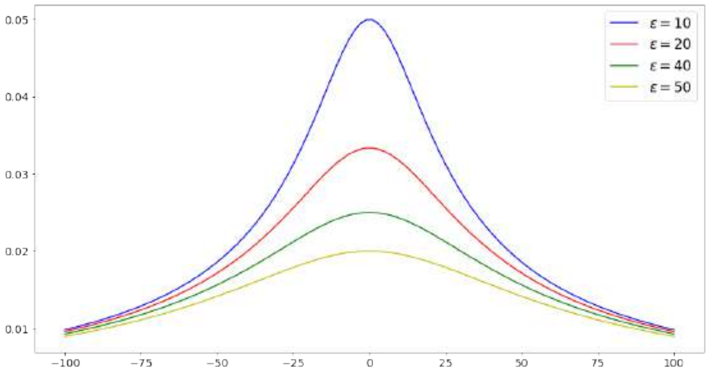


FIGURE 7: GRAPH OF THE INVERSE MULTIQUADRIC FUNCTION FOR DIFFERENT VALUES OF *SHAPE PARAMETER*.

Since this is a crucial parameter for RBF interpolation, it is a good practice to adjust it based on cross-validation strategies. One of these strategies is the K-fold algorithm, which we describe below.

2.5.4 SHAPE PARAMETER ESTIMATION IN RBF-IMQ

In this section, we present a strategy for estimating the best shape parameter in RBF-IMQ, which is characterized by the one that minimizes the RMSE value. For this purpose, we use a popular strategy in Data Sciences, known as *K-fold* [13].

K-fold is a cross-validation method that divides the entire dataset into K mutually disjoint and equally sized subsets, where K models are trained. Each model 1 subset is used for validation, and the remaining $K - 1$ is used for training. This process is performed K times by circularly changing the validation subset.

At the end of the K iterations, the average error is calculated based on the K errors obtained. This provides a more reliable measure regarding the mean error of the model. Figure 8 illustrates this process.

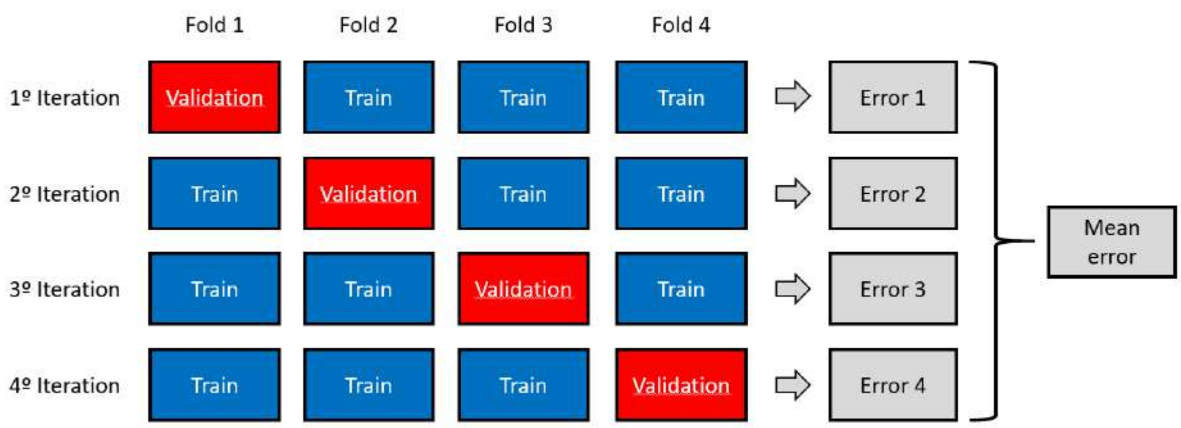


FIGURE 8: ILLUSTRATION OF THE K-FOLD CROSS-VALIDATION ALGORITHM WITH $K = 4.$

This process is performed for different values of the shape parameter, usually sampled regularly from an interval. Finally, the optimal parameter selected is the one that resulted in the lowest average RMSE value during the validation process.

2.5.5 APPROXIMATION BY MOVING LEAST SQUARES (MLS)

Before introducing the Moving Least Squares method [11], it is necessary to understand a traditional version of this method, called the Least Squares method (LS). Given a function $f \in C([a,b])$, where $C([a,b])$ is the space of the real and continuous functions in the interval $[a,b]$, the objective of LS is to obtain an approximation of f given by a linear combination of known functions. An approximation of this type is necessary to simplify complex functions or when the function is discrete.

Let V be a finite-dimensional subspace of $C([a,b])$, whose base $\{\varphi_0,\dots,\varphi_n\}$ is formed by known functions. LS searches the subspace V for a function $f^* = \alpha_0\varphi_0 + \dots + \alpha_n\varphi_n$, that provides the best approximation f to minimize the following functional:

$$Q(\alpha) = \|f - f^*\|_2^2 = \int_a^b [f(x) - f^*(x)]^2 \, dx\,. \tag{8}$$

Therefore, the LS solution is obtained by minimizing the squared norm L^2, hence the name of the method: least squares. The solution $\boldsymbol{\alpha} = (\alpha_0,\dots,\alpha_n)^\top$ of the Equation (8) can be obtained by solving the *system of normal equations*:

$$\begin{bmatrix} \langle\varphi_0,\varphi_0\rangle & \cdots & \langle\varphi_0,\varphi_n\rangle \\ \vdots & \ddots & \vdots \\ \langle\varphi_n,\varphi_0\rangle & \cdots & \langle\varphi_n,\varphi_n\rangle \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} \langle f,\varphi_0\rangle \\ \vdots \\ \langle f,\varphi_n\rangle \end{bmatrix}, \tag{9}$$

where $\langle\cdot,\cdot\rangle$ denotes the standard inner product on $C([a,b])$.

The most common case in LS is when one wants to approximate a function $f \in C([a,b])$ by a polynomial, in other words, when V is the space P_n of polynomials of degree $\leq n$. Thus, the approximation f^* of f is written as $f^*(x) = \mathbf{p}^\top(x)\boldsymbol{\alpha}$, with $\mathbf{p}(x) = (1, x, x^2, \ldots, x^n)^\top$. Another widespread situation is when the function f is discrete, i.e., represented by a set of points $\{(x_i, y_i) \in \mathbb{R}^2 \mid y_i = f(x_i)\,, i = 0, \ldots, m\}$. The approximation of a discrete function defined at points (x_i, y_i) by a polynomial $f^* \in P_n$, with $n < m$, is obtained by substituting the norm L^2, in Equation (8), by the Euclidean norm in $\mathbb{R}^{m+1}$. Considering $\mathbf{y} = (y_0, y_1, \ldots, y_m)^\top$ e $\mathbf{f}^* = (f^*(x_0), f^*(x_1), \ldots, f^*(x_m))^\top$, the approximation of f in the sense of LS in its discrete form is obtained by minimizing $Q(\boldsymbol{\alpha}) = \|\mathbf{y} - \mathbf{f}^*\|_2^2 = \sum_{j=0}^{m} \left[y_j - \mathbf{p}_j^\top \boldsymbol{\alpha}\right]^2$ with $\mathbf{p}_j = \mathbf{p}(\mathbf{x}_j)$. Consequently, the system (9) in the discrete form is generated by replacing the standard inner product in $C([a, b])$ with the dot product in $\mathbb{R}^{m+1}$.

Moreover, note that the system (9) can be rewritten as

$$\mathbf{X}^\top \mathbf{X} \boldsymbol{\alpha} = \mathbf{X}^\top \mathbf{y}\,, \tag{10}$$

where $\mathbf{X}$ is the Vandermonde matrix of order $(m + 1) \times (n + 1)$ constructed from the values x_k. Numerically, the linear system (10) can be solved efficiently by factoring $\mathbf{X} = \mathbf{QR}$ via QR decomposition. Therefore, solving the system (10) is equivalent to solving the upper triangular system $\mathbf{R}\alpha = \mathbf{Q}^\top \mathbf{y}$. Our problems are discrete because they involve approximations of points, so that we will concentrate only on the discrete case of LS and its variants from now on.

Consider a function sampled in the set $\mathcal{S} = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R} \mid y_i = f(\mathbf{x}_i)\,, i = 0, \ldots, m\}$. Given a base of polynomials in several variables $\mathbf{p}(\mathbf{x})$ of degree n and a point $\mathbf{x} \in \mathbb{R}^d$, the approximation Moving Least Squares (MLS) allows local approximations around $\mathbf{x}$ to be performed using a weight function, this function makes points close to $\mathbf{x}$ more relevant in the approximation calculation. Similar to the LS polynomial approximation, the MLS approximation at a point $\mathbf{x}$ is obtained by minimizing the function $Q(\boldsymbol{\alpha}) = \|\mathbf{y} - \mathbf{f}^*\|_W^2 = \sum_{k=0}^{m} W^h(\mathbf{x} - \mathbf{x}_j) \left[y_j - \mathbf{p}_j^\top \boldsymbol{\alpha}\right]^2$. The weight function W^h is a positive function with compact support whose radius of influence depends on h. The norm $\|\cdot\|_W$ is induced by the inner product $\langle u, v\rangle_W = \sum_{j=0}^{m} W^h(\mathbf{x} - \mathbf{x}_j)\, u(\mathbf{x}_j) v(\mathbf{x}_j)$. Replacing this dot product in the system (10), the system (9) can be rewritten as follows:

$$\mathbf{M}_\mathbf{x} \boldsymbol{\alpha} = \mathbf{b}_\mathbf{x} \tag{11}$$

with $\mathbf{M}_\mathbf{x} = \mathbf{X}^\top \mathbf{W}_\mathbf{x} \mathbf{X} = \sum_j W^h(\mathbf{x} - \mathbf{x}_j) \mathbf{p}_j^\top \mathbf{p}_j$ and $\mathbf{b}_\mathbf{x} = \mathbf{X}^\top \mathbf{W}_\mathbf{x} \mathbf{y} = \sum_j W^h(\mathbf{x} - \mathbf{x}_j) \mathbf{p}_j f_j$, where $\mathbf{W}_\mathbf{x} = \operatorname{diag}\left(W^h(\mathbf{x} - \mathbf{x}_0), \ldots, W^h(\mathbf{x} - \mathbf{x}_m)\right)$. The matrix $\mathbf{M}_\mathbf{x}$ is known as *moment matrix*. The solution $\boldsymbol{\alpha} = \boldsymbol{\alpha}(\mathbf{x})$ of the system (11) depends on the $\mathbf{x}$ point for calculating weights , intuitively it is as if the $\boldsymbol{\alpha}$ solution “moves” according to the variation of $\mathbf{x}$, which explains the name “Moving Least Squares”. Therefore, the MLS approach can be written as:

$$f^*(\mathbf{x}) = \sum_{j \in \mathcal{N}(\mathbf{x})} f_j\, \Phi_j(\mathbf{x}) \quad \text{with} \quad \Phi_j(\mathbf{x}) = W^h(\|\mathbf{x} - \mathbf{x}_j\|)\, \mathbf{p}_\mathbf{x}^\top \mathbf{M}_\mathbf{x}^{-1} \mathbf{p}_j\,, \tag{12}$$

where $\mathcal{N}(\mathbf{x})$ denotes the neighborhood around the point $\mathbf{x}$. The functions Φ_j are called *shape functions.*

In addition to providing a good approximation, MLS can be seen as a local regression method, providing an estimate for a variance function and its spatial dependence. An excellent introduction to a statistical view of MLS can be found in Loader [17].

3 RESULTS

In this experiment, the following methods for spatial interpolation were evaluated: Random Forest (RF), Radial Base Functions (RBF), Stochastic Partial Differential Equation (SPDE), and Moving Least Squares (MLS). Furthermore, different modifications of the base methods were evaluated. In this case, we have Random Forest with attribute engineering, linear RBF with Unit Partition (PU), and RBF with inverse multi-quadratic function (IMQ) using K-fold validation. In addition, it has been presented interpolation using *Kriging* technique with the *ordinary* method and *spherical* semivariogram whose results are a Gaussian kriging with a distance of 4 meters (see figure5).

The georeferenced data were divided into datasets A and B, each corresponding to a different depth, as explained in Section 1.1. Furthermore, each of these databases has a data set for training (candidate observations) and another test set (validation observations). The spatial distribution of the samples from each set can be seen in Figure 3. The amount of data in each of the sets is shown in Table 1.

TABLE 1: DISTRIBUTION OF THE 910 OBSERVATIONS IN THE DATA SETS USED IN THE STUDY.

Set	dataset A	dataset B
Train	345	338
Test	113	114

To evaluate the models generated according to each of the strategies, metrics such as RMSE (Root Mean Square Error), MAE (Mean Absolute Error), Bias (Bias), COR (Pearson Correlation), and MAPE (Mean Absolute Percentage Error) were used, which were calculated on the data from the test sets in each database. The results are presented in Section 3.1.

In addition to the quantitative error evaluation, data were spatialized for the region of interest according to each method using a 200×200 grid over the region. These results are presented in Section 3.2, allowing the comparison of the visual quality of the spatialization generated by the models.

3.1 QUANTITATIVE ANALYSIS

Table 2 below displays the Root Mean Square Error (RMSE) for calcium (Ca) in both samples A and B. The laboratory-based and vis-NIR rows in the table correspond to the error results from the reference article (see [22]), both obtained by interpolation using kriging. As can be seen, the methods under study have comparable and even lower error results than those presented in the article.

Table 3 presents the other error metrics calculated for each of the studied strategies, only for the set of tests from dataset A. Table 4 shows the error measures for dataset B.

TABLE 2: COMPARISION USING RMSE.

Method	dataset A	dataset B
laboratory-based	9.42	6.06
vis-NIR	9.67	6.82
RF	9.73	5.86
RF + features	9.26	6.28
RBF + linear poly	9.16	5.98
RBF PU	9.23	5.99
RBF IMQ	**9.09**	**5.64**
MLS quadratic	9.61	6.42
SPDE	9.21	6.59
Kriging	**9.09**	**5.63**

TABLE 3: ERROR MEASURES FOR DATASET A – VARIABLE CA.

Method	MAE	RMSE	MBE	COR	MAPE
RF	5.66	9.73	0.23	0.679	0.38
RF + features	5.25	9.26	0.28	0.711	0.36
RBF + linear poly	5.43	9.16	0.73	0.726	0.36
RBF-PU	5.40	9.23	0.83	0.720	0.36
RBF-IMQ + K-fold	5.21	9.09	0.94	0.727	0.35
MLS quadratic	5.82	9.61	0.78	0.700	0.39
SPDE	5.17	9.21	0.56	0.72	0.34
Kriging	5.15	9.09	0.20	0.723	0.362

TABLE 4: ERROR MEASURES FOR DATASET B – VARIABLE CA.

Método	MAE	RMSE	BIAS	COR	MAPE
RF	4.11	5.86	-1.23	0.867	1.07
RF + features	4.14	6.28	-1.63	0.869	0.94
RBF + linear poly	3.89	5.98	-1.01	0.87	0.84
RBF-PU	3.96	5.99	-0.97	0.870	0.86
RBF-IMQ + K-fold	3.84	5.64	-0.72	0.871	0.91
MLS quadratic	4.23	6.42	-1.23	0.857	0.95
SPDE	4.16	6.59	-1.97	0.870	0.94
Kriging	3.65	5.63	-1.34	0.880	0.93

3.2 QUALITATIVE ANALYSIS

The spatialization of the samples' calcium content (in $mmol_ckg^-1$) for each strategy. Figures 9 to 17 refer to the spatialization of the Ca content at the level A (0.0-0.2 m), while the Figures 18 to 25 are referring to the depth level B (0.8-1.0 m).

Among these results, the Figure 11 shows the lower (5th percentile) and upper (95th percentile) bounds for spatial interpolation using RF with Feature Engineering; and Figure 15 shows a pointwise estimative of the MLS, the estimate of variance for the entire area provided by the method. These results are for depth A. The same results are presented for depth B in Figure 24.

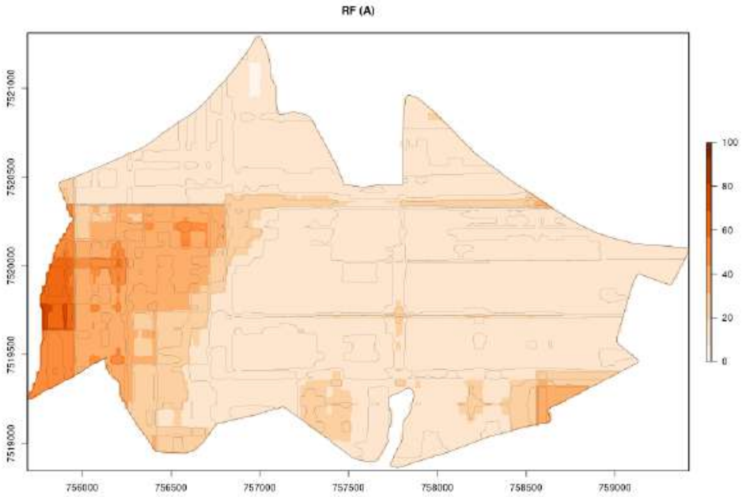


FIGURE 9: SPATIALIZATION OF CALCIUM DATA AT DEPTH **A** FOR THE STUDY REGION USING RF METHOD.

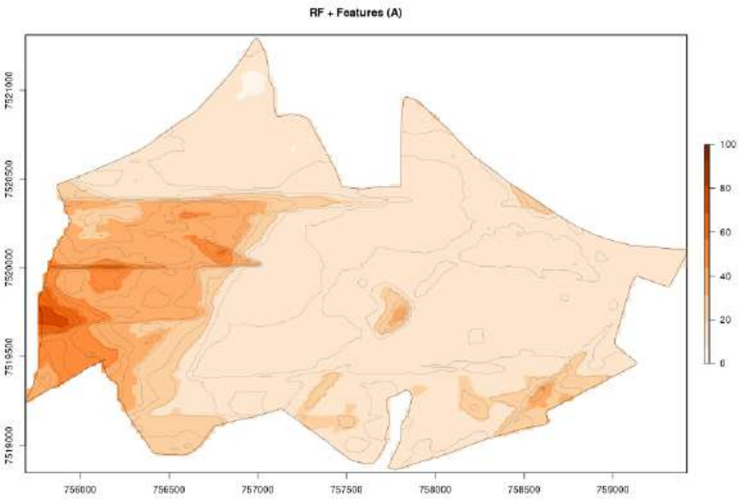


FIGURE 10: SPATIALIZATION OF CALCIUM DATA AT DEPTH **A** FOR THE STUDY REGION USING RF WITH FEATURES.

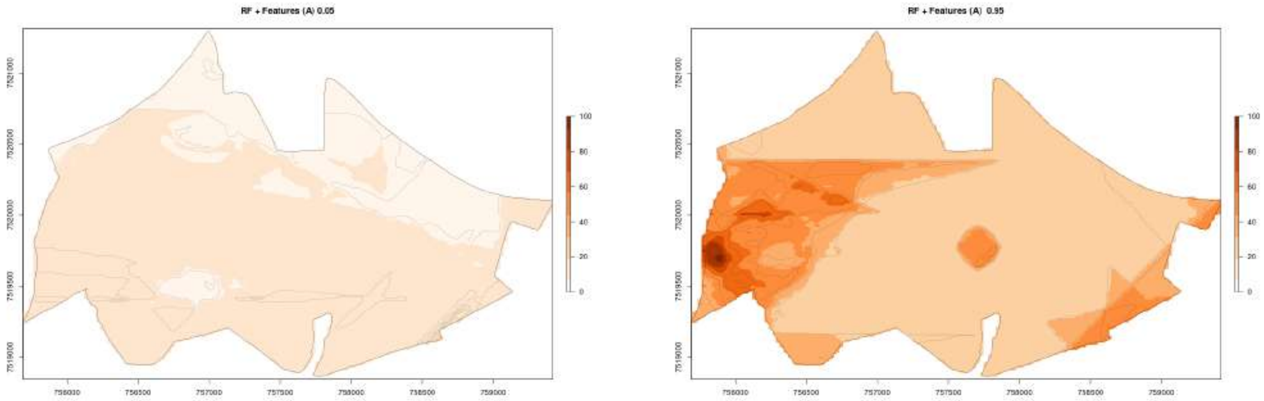


FIGURE 11: SPATIALIZATION OF CALCIUM DATA AT DEPTH **A** FOR THE STUDY REGION USING RF WITH FEATURES: LOWER (LEFT) AND UPPER (RIGHT) BOUNDS.

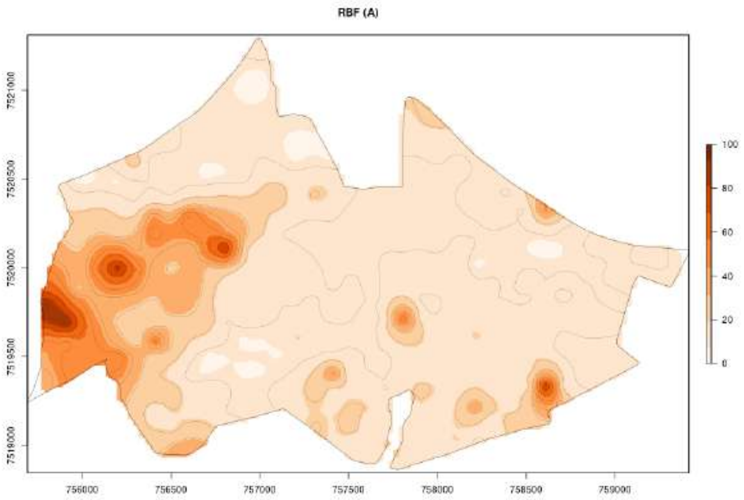


FIGURE 12: SPATIALIZATION OF CALCIUM DATA AT DEPTH **A** FOR THE STUDY REGION USING RBF.

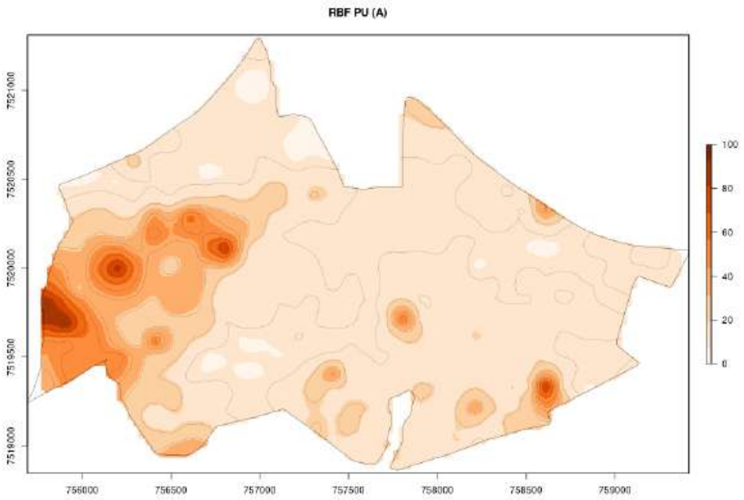


FIGURE 13: SPATIALIZATION OF CALCIUM DATA AT DEPTH **A** FOR THE STUDY REGION USING RBF-PU.

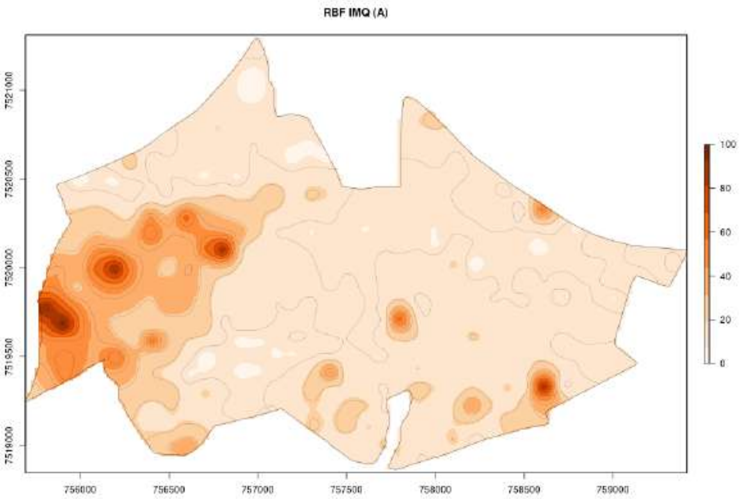


FIGURE 14: SPATIALIZATION OF CALCIUM DATA AT DEPTH **A** FOR THE STUDY REGION USING RBF-IMQ.

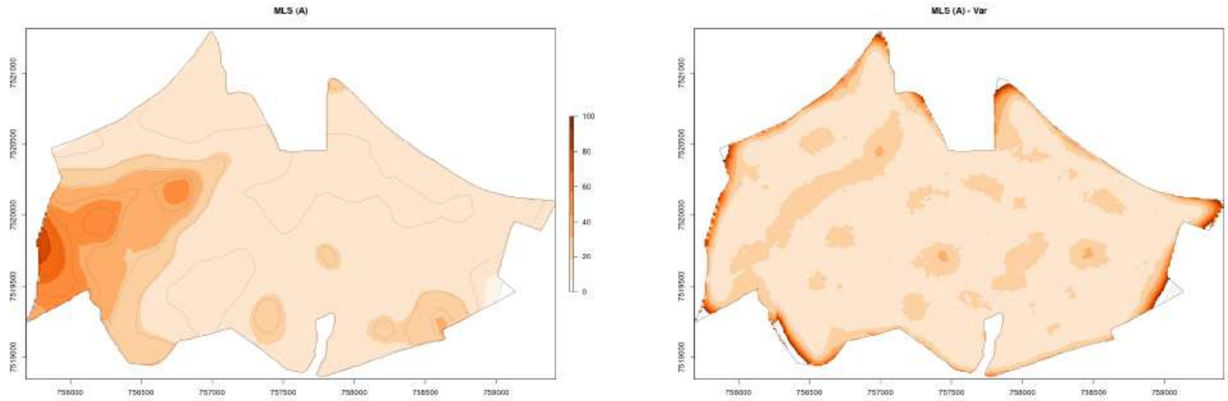


FIGURE 15: SPATIALIZATION OF CALCIUM DATA AT DEPTH **A** FOR THE STUDY REGION USING MLS (LEFT) AND ITS VARIANCE (RIGHT).

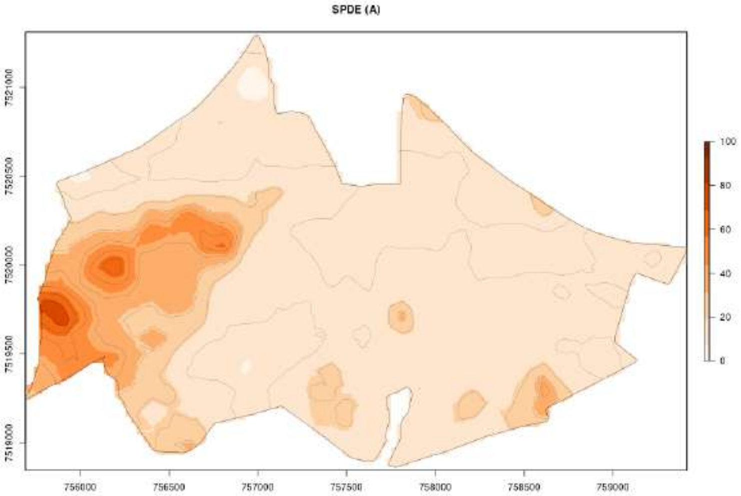


FIGURE 16: SPATIALIZATION OF CALCIUM DATA AT DEPTH **A** FOR THE STUDY REGION USING SPDE.

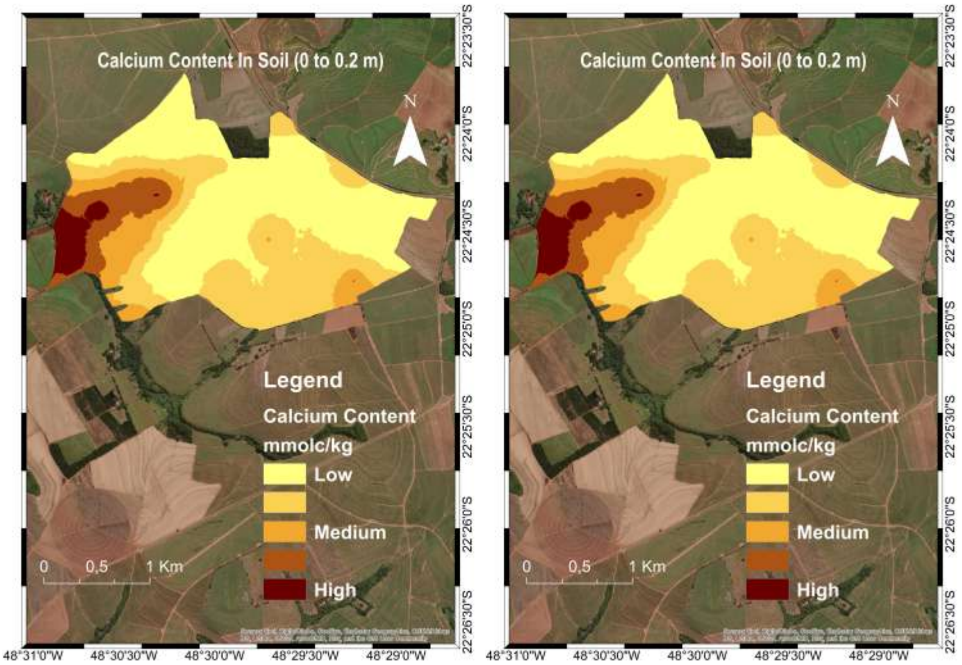


FIGURE 17: SPATIALIZATION OF CALCIUM DATA AT DEPTH **A** (LEFT) AND **B** (RIGHT) FOR THE STUDY REGION USING KRIGING.

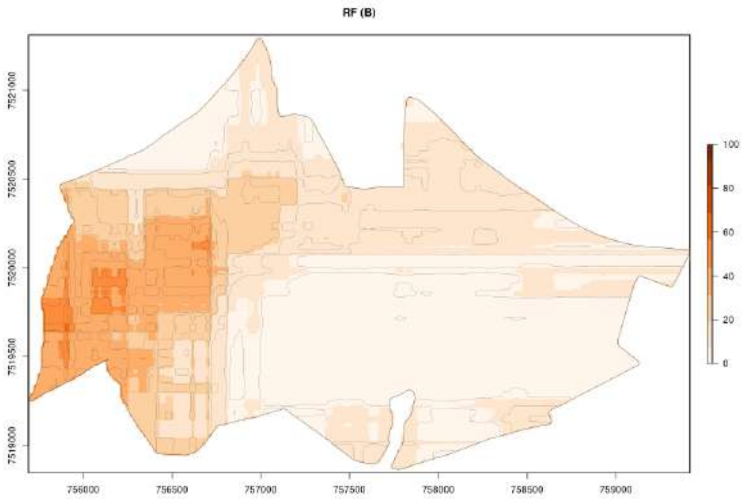


FIGURE 18: SPATIALIZATION OF CALCIUM DATA AT DEPTH **B** FOR THE STUDY REGION USING RF.

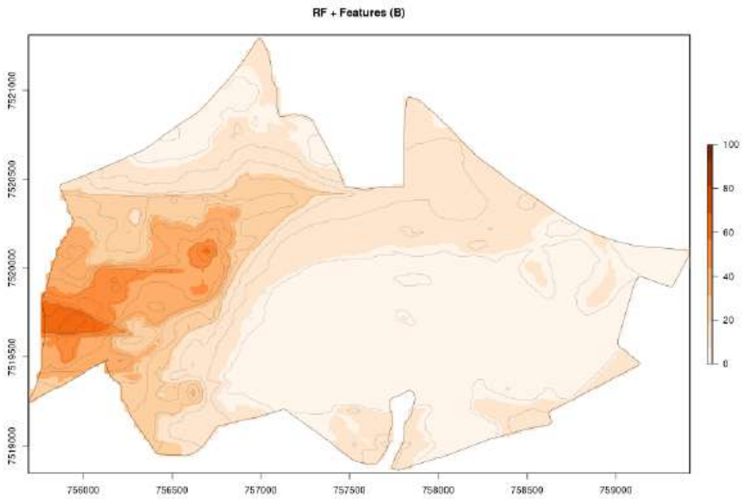


FIGURE 19: SPATIALIZATION OF CALCIUM DATA AT DEPTH **B** FOR THE STUDY REGION USING RF WITH FEATURES.

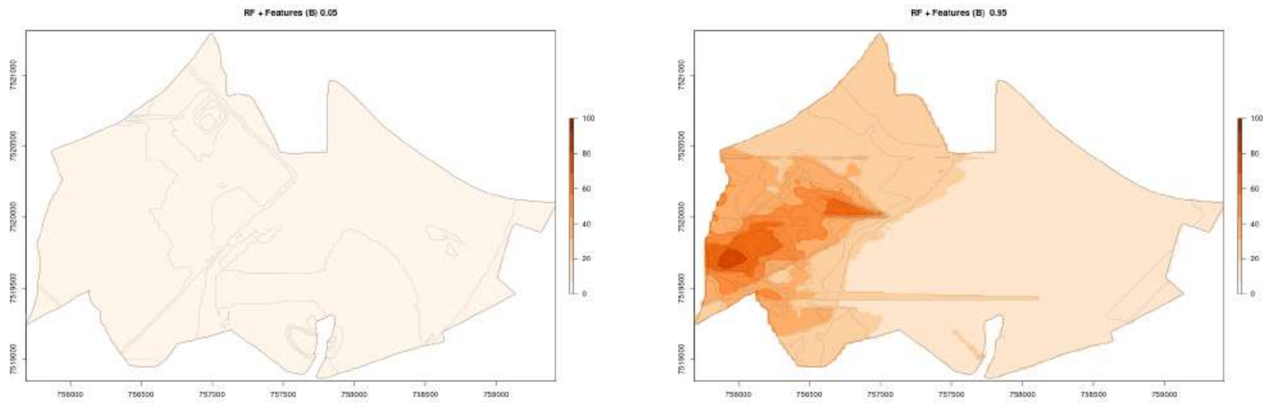


FIGURE 20: SPATIALIZATION OF CALCIUM DATA AT DEPTH **B** FOR THE STUDY REGION USING RF WITH FEATURES: LOWER (LEFT) AND UPPER (RIGHT) BOUNDS.

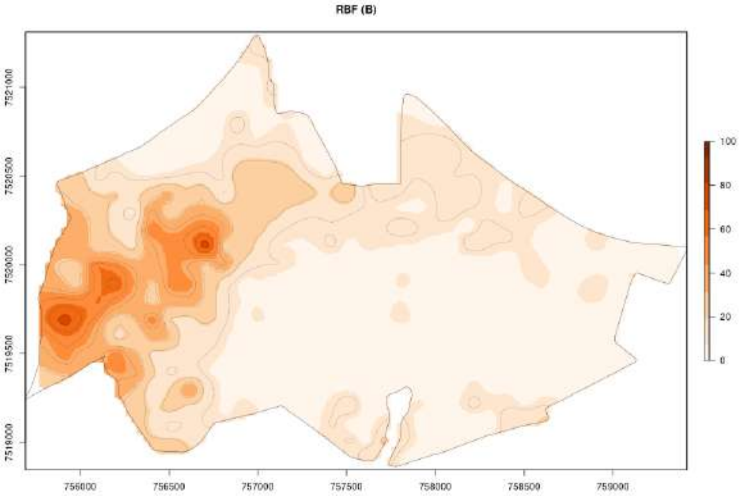


FIGURE 21: SPATIALIZATION OF CALCIUM DATA AT DEPTH **B** FOR THE STUDY REGION USING RBF.

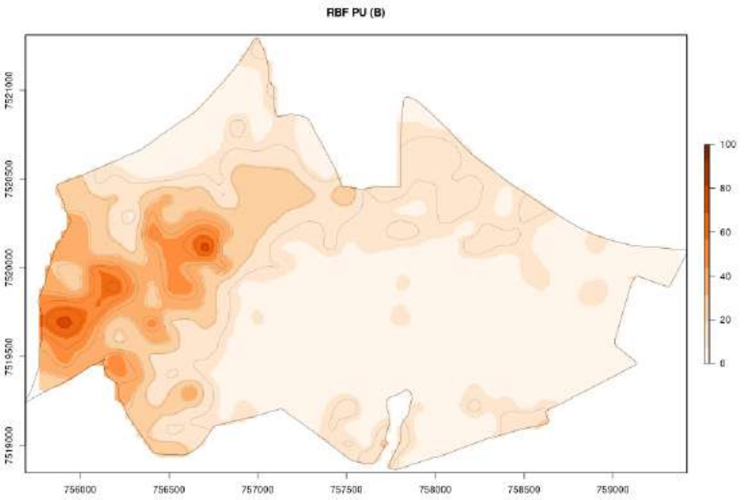


FIGURE 22: SPATIALIZATION OF CALCIUM DATA AT DEPTH **B** FOR THE STUDY REGION USING RBF-PU.

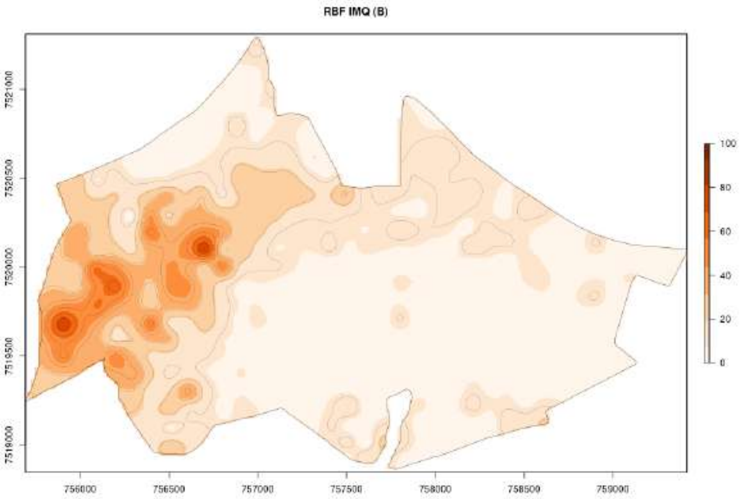


FIGURE 23: SPATIALIZATION OF CALCIUM DATA AT DEPTH **B** FOR THE STUDY REGION USING RBF-IMQ.

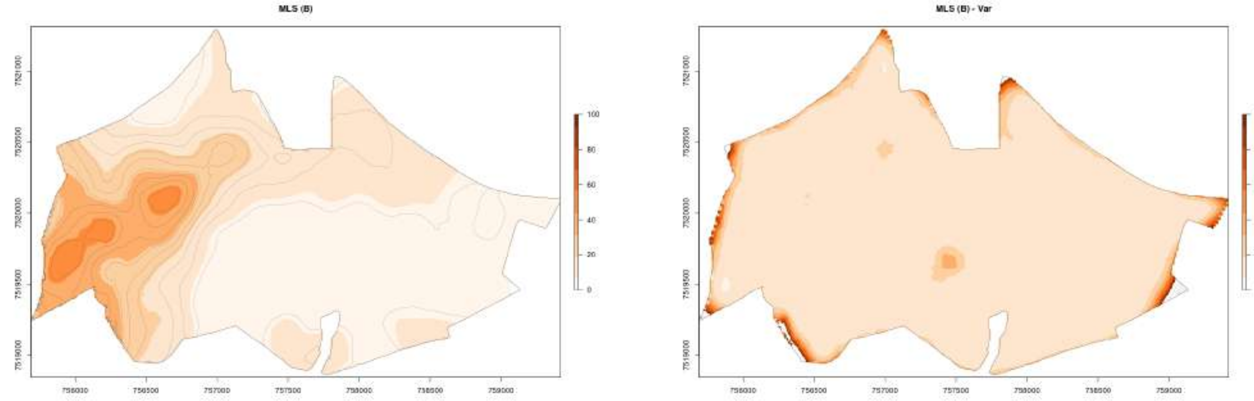


FIGURE 24: SPATIALIZATION OF CALCIUM DATA AT DEPTH **B** FOR THE STUDY REGION USING MLS (LEFT) AND ITS VARIANCE (RIGHT).

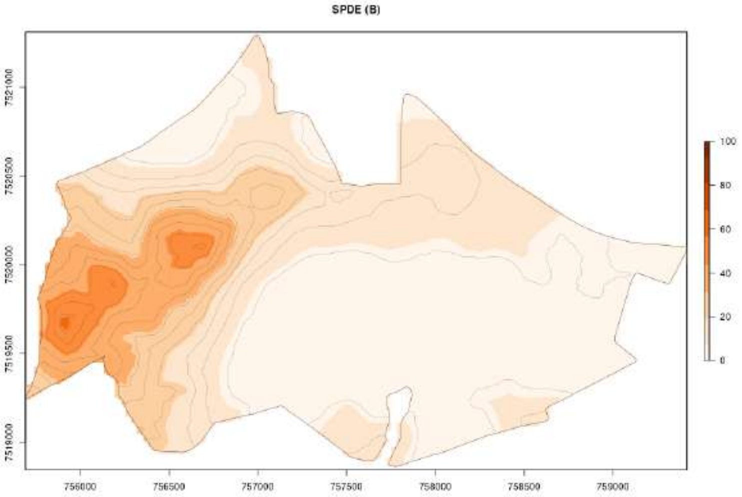


FIGURE 25: SPATIALIZATION OF CALCIUM DATA AT DEPTH **B** FOR THE STUDY REGION USING SPDE.

4 CONCLUSION

The methodologies studied during the VI Brazilian Study Group With Industry showed that the problems brought by the company could be solved in countless ways. In the scenarios analyzed in the period of occurrence of the event and considering the work with a base analysis, we raised some methods that proved to be quite competitive in terms of computational processing time, reduction of subjectivity in capturing spatial dependence, potential possibility of quantification and comparison of effects, control of uncertainty mechanisms and randomness patterns, amongst others. In addition, we also study ways of interactively presenting the results obtained on web platforms free of charge.

In a posterior context, the analysis process can be made more general and flexible, considering other sources of information and even more specific characteristics of the problems in this theme, which allows even more improvements in the forecasts made. There is also the possibility of automating the analysis process and the availability of results to create a computerized analysis system.

REFERENCES

[1] OCDE - Organization for Economic Cooperation and Development. jul 2015.

[2] CONAB - Companhia Nacional de Abastecimento. produção de soja e aumento da área de milho impulsionam supersafra de grãos. *Monitoramento Agrícola*, (2), mar. 2021.

[3] Transformando nosso mundo: A agenda 2030 para o desenvolvimento sustentável. *Nações Unidas no Brasil*, mar. 2021.

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] Jelle Bruinsma. The resource outlook to 2050: By how much do land, water and crop yields need to increase by 2050? Technical report, Expert Meeting on How to Feed the World in 2050. Food and Agriculture Organization of the United Nations. Economic and Social Development Department, Rome, jun 2009.

[6] Angelo J Canty. Resampling methods in r: the boot package. *The Newsletter of the R Project Volume*, 2:3, 2002.

[7] Håvard Rue; Sara Martino; Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 2009.

[8] Tereza Cristina [Ministra da Agricultura do Brasil]. Mulheres positivas. Entrevista concedida à Rádio Jovem Pan (00:22:17 minutos), 2021.

[9] Peter J. Diggle and Paulo J. Ribeiro. *Model-based Geostatistics*. Springer Series in Statistics. Springer, mar. 2007.

[10] K. Euclides Filho. Projeção da demanda futura de carne bovina: Desafios permanentes para o melhoramento animal. *Anais do IX Simpósio Brasileiro de Melhoramento Animal da Sociedade Brasileira de Melhoramento Animal – SBMA*, pages 1–9, 2012.

[11] G. F. Fasshauer. *Meshfree Approximation Methods with MATLAB*. World Scientific, 2007.

[12] P. Gerland, A. E. Raftery, H. Sevcikova, N. Li, D. Gu, T. Spoorenberg, L. Alkema, B. K. Fosdick, J. Chunn, N. Lalic, G. Bay, T. Buettner, G. K. Heilig, and J. Wilmoth. World population stabilization unlikely this century. *Science*, 346(234):234–237, 2014.

[13] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[14] Rasmus Houborg and Matthew F McCabe. A hybrid training approach for leaf area index estimation via cubist and random forests machine-learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 135:173–188, 2018.

[15] Richard M Lark and Dan J Lapworth. Quality measures for soil surveys by lognormal kriging. *Geoderma*, 173:231–240, 2012.

[16] Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

[17] Clive Loader. *Local regression and likelihood*. New York: Springer-Verlag, 1999.

[18] Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. *Journal of Machine Learning Research*, 7(6), 2006.

[19] Richard; Seaton Andrew E. Miller, David L.; Glennie. Understanding the stochastic partial differential equation approach to smoothing. *Journal of Agricultural, Biological, and Environmental Statistics*, 9 2019.

[20] Budiman Minasny and Alex. B. McBratney. The matérn function as a general model for soil variograms. *Geoderma*, 128(3):192–207, 2005. Pedometrics 2003.

[21] Madlene Nussbaum, A Papritz, Andri Baltensweiler, and Lorenz Walthert. Estimating soil organic carbon stocks of swiss forest soils by robust external-drift kriging. *Geoscientific Model Development*, 7(3):1197–1210, 2014.

[22] Leonardo Ramirez-Lopez, AMJ-C Wadoux, Marston HD Franceschini, FS Terra, KPP Marques, VM Sayão, and JAM Demattê. Robust soil mapping at the farm scale with vis–nir spectroscopy. *European Journal of Soil Science*, 70(2):378–393, 2019.

[23] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

[24] J. Simmons. Por que o agronegócio precisa de tecnologia para atender à demanda crescente por alimentos saudáveis, nutritivos e com custos razoáveis. *Nutrition for Tomorrow*, 1(2):5–8, 2010.

[25] M.L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer New York, 1999.

[26] Paulo César Teixeira, Guilherme Kangussu Donagemma, Ademir Fontana, Wenceslau Geraldes Teixeira, et al. *Manual de métodos de análise de solo*. Brasília, 3 edition, 2017. rev. ampl. EMBRAPA Informação Tecnológica.

[27] A. A. Vilarinho. A importância do melhoramento genético na cultura do milho. *Página Rural*, set. 2003. Seção Ponto de Vista.

[28] Hans-Peter Seidel Yutaka Ohtake, Alexander Belyaev. 3d scattered data approximation with adaptivecompactly supported radial basis functions. *Proceedings Shape Modeling Applications*, 2004.

[29] Marc Alexa Greg Turk Hans-Peter Seidel Yutaka Ohtake, Alexander Belyaev. Multi-level partition of unity implicits. *ACM Transactions on Graphics*, 2003.

[30] L Zambolim, L. C. B. Nasser, J. R. Andrigueto, J. M. A. Teixeira, and A. R. Kososki. *Produção Integrada no Brasil: agropecuária sustentável, alimentos seguros*. Biblioteca Nacional de Agricultura – BINAGRI, fachinello, j. c. edition, 2009.