LUCIANO GONÇALVES BATISTA

DECISION TREE E GEOESTATÍSTICA NA REDUÇÃO DO NÚMERO DE ANÁLISES DE MICRONUTRIENTES DO SOLO

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

Orientador: Nerilson Terra Santos

Coorientador: Luiz Alexandre Peternelli

VIÇOSA - MINAS GERAIS
2024

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

B333d
2024

Batista, Luciano Gonçalves, 1997-

Decision tree e geoestatística na redução do número de análise de micronutrientes do solo / Luciano Gonçalves Batista. – Viçosa, MG, 2024.

1 dissertação eletrônica (48 f.): il. (algumas color.).

Orientador: Nerilson Terra Santos.

Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Estatística, 2024.

Inclui bibliografia.

DOI: https://doi.org/10.47328/ufvbbt.2024.475

Modo de acesso: World Wide Web.

1. Árvores de decisão. 2. Amostragem (Estatística). 3. Aprendizado do computador. 4. Krigagem . I. Santos, Nerilson Terra, 1966-. II. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 511.52

Bibliotecário(a) responsável: Alice Regina Pinto Pires CRB-6/2523

LUCIANO GONÇALVES BATISTA

DECISION TREE E GEOESTATÍSTICA NA REDUÇÃO DO NÚMERO DE ANÁLISES DE MICRONUTRIENTES DO SOLO

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 23 de fevereiro de 2024.

Assentimento:

Documento assinado digitalmente
gov.br
LUCIANO GONCALVES BATISTA
Data: 19/08/2024 12:11:57-0300
Verifique em https://validar.iti.gov.br

Luciano Gonçalves Batista
Autor

Documento assinado digitalmente
gov.br
NERILSON TERRA SANTOS
Data: 19/08/2024 14:18:03-0300
Verifique em https://validar.iti.gov.br

Nerilson Terra Santos
Orientador

Ao meu eu.

AGRADECIMENTOS

Agradecer é um ato de reconhecer a importância daquilo que foi essencial na nossa trajetória. Agradeço primeiramente a Deus que é a matriz de todo o nosso conhecimento, pois sem ele, nada disso seria possível. Agradeço a minha família, que sempre esteve me apoiando ao longo dos anos. Sou muito grato à Universidade Federal de Viçosa, em especial ao Departamento de Matemática, ao Programa de Pós-graduação em Estatística Aplicada e Biometria, ao Laboratório de Análises e Pesquisas em Estatística Aplicada (LAPEA) e ao Grupo de Estudos em Estatística Aplicada e Biometria (GESTBIO) que me acolheram e me tornou a pessoa que sou hoje.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) - Código de Financiamento 001. Agradeço à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), pela concessão da bolsa de estudos.

Muito obrigado a banca pelas valiosas sugestões.

Não poderia deixar de agradecer a todos os meus amigos que fizeram, fazem e farão parte de todo meu processo de aprendizado. Agradeço em especial a Samantha, que sempre esteve meu lado e foi essencial na minha trajetória durante o mestrado.

A felicidade é a única forma de se sentir imortal

(Padecestes_Thálles Dias)

RESUMO

BATISTA, Luciano Gonçalves, M.Sc., Universidade Federal de Viçosa, fevereiro de 2024. ***Decision tree*** **e geoestatística na redução do número de análises de micronutrientes do solo**. Orientador: Nerilson Terra Santos. Coorientador: Luiz Alexandre Peternelli.

Para realizar a interpolação por krigagem, é importante que cada ponto num semivariograma seja obtido com base no mínimo da combinação de 30 pares de pontos. Além disso, alguns autores alegam que é necessário ter pelo menos 100 amostras para fazer a interpolação. Sendo assim, o processo de amostragem se torna caro para o produtor rural. Como alternativa de contornar este problema de amostragem, foi utilizado metodologias de *machine learning*. O objetivo principal deste trabalho é avaliar o uso da metodologia de *decision tree* na redução do adensamento amostral para atributos do solo visando a realização da krigagem ordinária com tamanho amostral reduzido. Para isso, foi realizado 50 amostragem pelo algoritmo *Latin Hypercube Sampling* (LHS), com malhas contendo 82, 112 e 127 pontos amostrados e os valores faltantes foram preditos com *decision tree*, até completar 150 pontos e logo em seguida foi realizado a krigagem ordinária para as malhas MR_{127}, MR_{112} e MR_{82} , que foi gerado pela combinação das 50 predições por *decision tree* e avaliados os valores da Raiz Quadrada do Erro Médio (RMSE) e Média do Erro Absoluto (MAE), denominados RMSE_Krig e MAE_Krig. Foi percebido que há uma redução nestas estatísticas ao passo que aumentamos a redução amostral. A redução das estatísticas de validação indica que à medida que aumentamos a quantidade de amostras preditas com *decision tree*, há uma melhoria no modelo de krigagem ordinária. Ao fazer o mapa de atributos para as malhas reduzidas, é percebido que o padrão de concentração de nutrientes dos solos nas malhas reduzidas segue semelhante ao padrão original, ou seja, regiões com maiores concentrações ainda mantêm níveis elevados, enquanto aquelas com menores concentrações continuam a apresentar índices reduzidos. Ao fazer o mapa de atributos das malhas reduzidas é percebido que o padrão de concentração de micronutrientes dos solos nas malhas reduzidas segue semelhante ao padrão original, ou seja, zonas com maiores concentrações ainda continuam com concentrações elevadas e regiões com menores concentrações continuam com concentrações menores. Com isso, a *decision tree*, se mostrou eficiente em preservar o padrão de distribuição dos micronutrientes.

Palavras-chave: Adensamento amostral; Aprendizado estatístico; Krigagem ordinária.

ABSTRACT

BATISTA, Luciano Gonçalves, M.Sc., Federal University of Viçosa, February 2024. ***Decision tree* and geostatistics in reducing the number of soil micronutrient analyzes**. Adviser: Nerilson Terra Santos. Co-adviser: Luiz Alexandre Peternelli.

To perform kriging interpolation, it is important that each point in a semivariogram is obtained based on a minimum combination of 30 pairs of points. Additionally, some authors argue that at least 100 samples are necessary to conduct the interpolation. Therefore, the sampling process becomes expensive for the farmer. As an alternative to address this sampling issue, machine learning methodologies were employed. The main objective of this work is to evaluate the use of the decision tree methodology in reducing sample density for soil attributes, aiming to perform ordinary kriging with a reduced sample size. For this purpose, 50 samplings were performed using the Latin Hypercube Sampling (LHS) algorithm, with grids containing 82, 112, and 127 sampled points, and the missing values were predicted with a decision tree until 150 points were completed. Ordinary kriging was then conducted for the grids MR_{127}, MR_{112} and MR_{82}, which were generated by combining the 50 predictions made by the decision tree, and the values of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), referred to as RMSE_Krig and MAE_Krig, were evaluated. It was observed that these statistics decrease as the sample reduction increases. The reduction in validation statistics indicates that as the number of samples predicted by the decision tree increases, there is an improvement in the ordinary kriging model. When creating the attribute map for the reduced grids, it is observed that the pattern of nutrient concentration in the soils of the reduced grids remains similar to the original pattern, meaning that regions with higher concentrations still maintain elevated levels, while those with lower concentrations continue to show reduced levels. Therefore, the decision tree proved to be effective in preserving the distribution pattern of micronutrients.

Keywords: Sample density; Statistical learning; Ordinary kriging.

SUMÁRIO

1. INTRODUÇÃO

Estudar o padrão de variabilidade espacial de atributos químicos e físicos de propriedades do solo é tema importante para a agricultura de precisão, que consiste em aumentar a eficiência da propriedade rural, através do mapeamento e monitoramento da atividade agrícola (MIAO; MULLA; ROBERT, 2006; TSCHIEDEL; FERREIRA, 2002).

A agricultura de precisão é peça fundamental no manejo sustentável da produção agrícola, minimizando os danos causados ao meio ambiente e prejudicando o mínimo possível as reservas naturais (TSCHIEDEL; FERREIRA, 2002), uma vez que utilizamos a quantidade correta de insumos nos locais apropriados, sem excedentes (FATORGIS, 1999; MANZATTO; BHERING; SIMÕES, 1999).

Como ferramenta para mensurar a variabilidade espacial das propriedades do solo, utilizamos a geoestatística (OLIVER, 1999), que é um ramo da estatística que estuda as variáveis regionalizadas (YAMAMOTO; LANDIM, 2013) e considera a dependência espacial entre as variáveis.

Usando metodologias da geoestatística, como o semivariograma, que é uma função matemática para modelar a continuidade espacial (VIEIRA, 2000), conseguimos fazer previsões sobre o rendimento da lavoura e identificar quais áreas são mais pobres em nutrientes. Assim, podemos corrigir com fertilizantes os nutrientes faltantes na quantidade apropriada para cada área da fazenda, fazendo com que a lavoura atinja seu potencial máximo de crescimento (DAHIKAR; RODE, 2014).

A previsão de rendimento das culturas nas lavouras não é tarefa fácil, pois a junção entre os fatores ecológicos e suas explicações são complexas e não lineares (D'AMARIO *et al*., 2019; GONZALEZ-SANCHEZ; FRUSTO-SOLIS; OJEDA-BUSTAMANTE, 2014).

Além disso, nos deparamos com algumas limitações quanto ao uso da geoestatística para predizer a produção da lavoura. Muitos agricultores reclamam da grande quantidade de amostras que a geoestatística exige (MENDES *et al*., 2020), visto que, de acordo com Cherubin *et al*. (2015) e Nanni *et al*. (2011), são necessários um ponto por hectare e um ponto por 7,2 hectares, respectivamente, para conseguir construir o semivariograma e modelar a variabilidade espacial. Além do mais, Cherubin *et al*. (2015) alertam que o uso de malhas que utilizam poucos pontos de amostragem deve ser evitado para não afetar a acurácia do poder preditivo do modelo, o que torna o processo de amostragem e análise do solo mais caro para o produtor (MENDES *et al*., 2020).

Como alternativa para contornar este problema da geoestatística, podem ser utilizadas técnicas de machine learning para mapeamento do solo em atividades agrícolas (MENDES *et al*., 2020). As técnicas de machine learning têm sido muito utilizadas, pois suas metodologias são capazes de lidar com problemas complexos e não lineares (PANTAZI *et al*., 2016; TANTALAKI; SOURAVLAS; ROUMELIOTIS, 2019), que é uma das complicações de previsões de rendimento das culturas, que estão interligadas com a distribuição dos nutrientes dos solos (PANTAZI *et al*., 2016).

Alguns trabalhos recentes, como Pereira *et al*. (2022) e Martins *et al*. (2023), têm utilizado técnicas de machine learning como método de interpolação e comparado com a krigagem ordinária, que é um método de interpolação na geoestatística. Em suas pesquisas, a krigagem ordinária ainda continuou sobressaindo em comparação às técnicas usadas para interpolação com machine learning.

As pesquisas realizadas por Martins *et al*. (2023) e Pereira *et al*. (2022) têm mostrado que a krigagem ordinária ainda é um bom método de interpolação e que o machine learning é um grande aliado para redução do adensamento amostral. A justificativa para o desenvolvimento deste trabalho baseia-se em combinar os dois métodos, pois até então há poucos trabalhos seguindo esta vertente.

Há uma variedade de metodologias de machine learning; porém, para este trabalho, foi empregada a decision tree. A ideia não é parar de usar a krigagem para interpolar, mas sim usar ferramentas de machine learning, em especial decision tree, para auxiliar a interpolação por krigagem ordinária com uma quantidade reduzida de amostras.

O objetivo principal deste trabalho é avaliar o uso da metodologia de decision tree na redução do adensamento amostral para atributos do solo visando a realização da krigagem ordinária com tamanho amostral reduzido.

2. REVISÃO DE LITERATURA

2.1 Agricultura de precisão

As recomendações tradicionais para correções do solo utilizam a média dos atributos como parâmetro para aplicação dos insumos agrícolas, como fertilizantes, defensivos agrícolas etc. (TSCHIEDEL; FERREIRA, 2002). No entanto, esta prática considera a homogeneidade da área, e o tratamento é feito de forma igualitária para todo o terreno. Utilizar a média como parâmetro de correção atende apenas às necessidades da média e não às especificidades de cada parte do campo (TSCHIEDEL; FERREIRA, 2002).

De acordo com Blackmore (2000), usar a média como critério para aplicar insumos agrícolas pode resultar em uma distribuição desigual: algumas áreas podem receber insumos em excesso, enquanto outras recebem menos do que necessitam. Isso acaba gerando uma produção desuniforme na lavoura.

Para evitar esse problema, é recomendável adotar técnicas mais precisas de manejo agrícola, como a agricultura de precisão. De acordo com Zhang *et al*. (2002), a agricultura de precisão utiliza dados detalhados sobre o solo para aplicar a quantidade de insumos necessária em cada área específica, evitando assim o desperdício.

A agricultura de precisão utiliza mapas de produtividade para identificar fatores que afetam a produtividade das culturas no campo (SANA *et al*., 2014). Tais mapas são fundamentais para gerir a produção agrícola do campo e ilustram a variabilidade espacial do solo (GUO; MAAS; BRONSON, 2012), e têm como objetivo orientar práticas de manejo (MOLIN; RABELO, 2011; MILANI; SOUZA; URIBE-OPAZO, 2006; BLACKMORE; GODWIN; FOUNTAS, 2003).

Com o avanço das tecnologias de informação e comunicação (TIC), a agricultura tem passado por grandes transformações em suas práticas de manejo. As primeiras formas de manejo agrícolas, conhecidas como agricultura 1.0, utilizavam a força física dos animais para arar a terra. Com a modernização, os produtores passaram a utilizar máquinas para gradear a terra e usar variedades melhoradas e outros insumos em suas lavouras, conhecidas como agricultura 2.0 e agricultura 3.0, respectivamente (BORÉM, 2021).

Agora, estamos começando uma nova fase na agricultura, conhecida como agricultura 4.0 (KLERKX; ROSE, 2020), no qual os produtores rurais vêm incrementando altas tecnologias (ROSE; CHILVERS, 2018), como uso de drones, robôs e imagens de satélites para mapear a produtividade da lavoura. Muitos pesquisadores estão chamando a agricultura 4.0 como uma Nova Revolução Verde (BORÉM, 2021).

A geoestatística é extremamente importante para a agricultura 4.0, pois estas novas metodologias de análise do solo utilizam dados georreferenciados (GREGO; OLIVEIRA; VIEIRA, 2014).

A agricultura de precisão está intimamente ligada à agricultura 4.0, pois, para que possamos colocar em prática os equipamentos tecnológicos no campo, precisamos entender o padrão de variabilidade espacial, que é um dos fundamentos da agricultura de precisão.

2.2. Variabilidade espacial e geoestatística

O delineamento inteiramente casualizado (DIC), delineamento em blocos casualizado (DBC) e delineamento em quadrado latino (DQL) são delineamentos experimentais muito utilizados, mas eles consideram que as unidades experimentais são homogêneas ou homogêneas em cada bloco.

Para estudar, entender e modelar a variabilidade espacial, utilizamos a geoestatística, que tem suas origens nas minas de ouro da África do Sul, nas décadas de 1950 e 1960, com trabalhos de Daniel Krige e Georges Matheron. Trabalhando com dados de concentração de ouro, Krige percebeu que as variâncias amostrais só faziam sentido se olhasse a posição em que as amostras foram coletadas (VIEIRA, 2000). Já em 1963, Matheron formalizou esta teoria de Krige, denominando como Teoria das Variáveis Regionalizadas (VIEIRA, 2000), que é o pilar da geoestatística.

Segundo Matheron (1963), variável regionalizada é caracterizada como uma função matemática, que considera a posição geográfica em que as amostras foram coletadas. Além disso, essa função considera a continuidade espacial, que não pode ser modelada por uma simples função matemática. Para modelar a continuidade espacial, fazemos uso do semivariograma (VIEIRA, 2000).

2.3. Semivariograma e suas propriedades

O método estatístico mais utilizado dentro das ciências agrárias é a análise de variância (ANOVA). Esta metodologia considera a hipótese de independência, homoscedasticidade e a aleatoriedade entre as variáveis. Mas sabemos que, do ponto de vista prático, algumas destas hipóteses, como independência e aleatoriedade, não são satisfeitas, pois as propriedades químicas e físicas do solo são dependentes de acordo com a sua posição geográfica e são autocorrelacionadas (VIEIRA, 2000).

A análise de variância só pode ser aplicada se for provada a independência espacial entre as amostras. Caso haja dependência espacial, utilizamos a autocorrelação para estimar a dependência espacial. A covariância mede a relação entre duas variáveis distintas, X e Y. A autocorrelação mede a relação entre os valores de uma mesma variável separados por uma distância h. A autocorrelação tem sido muito utilizada para estudos em ciências do solo, mas suas origens se deram na análise de séries temporais (WEBSTER, 1973; WEBSTER; CUANALO, 1975; VIEIRA *et al*., 1981).

Quando coletamos amostras de solo, consideramos as suas coordenadas geográficas no plano cartesiano (x,y), ou seja, estamos fazendo coletas em duas dimensões. Quando isso acontece, segundo Silva (1988), o método mais indicado para estimar a dependência espacial é o semivariograma. Quando precisamos fazer a interpolação, o semivariograma também é muito útil (VIEIRA *et al*., 1983).

Para estudar a dependência espacial, precisamos obter o semivariograma para todas as distâncias possíveis. Logo abaixo, é apresentada a demonstração do cálculo da semivariância.

2.4. Cálculo da semivariância

A semivariância é definida como na equação (1):

$$\gamma(h) = \frac{1}{2} E\{Z(u_i) - Z(u_i + h)\}^2 \quad (1)$$

Em que, $z(u_i)$ representa o valor do atributo u, na posição i e $(u_i + h)$ é o valor do atributo u, na posição i, separado por um vetor de distância h.

De acordo com Braga (1990), no semivariograma não há a necessidade do conhecimento do valor esperado da função $Z(u)$.

A semivariância total separado por uma distância h, é estimada por: (equação (2))

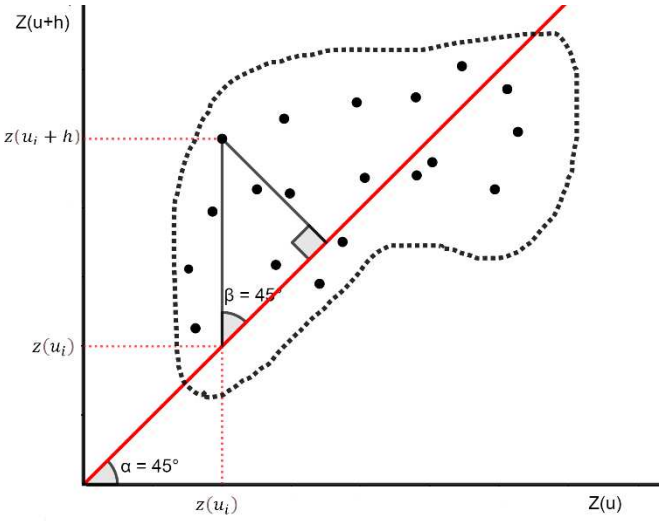
$$\gamma(h) = \frac{1}{2\, n(h)} \sum_{i=1}^{n(h)} [Z(u_i) - Z(u_i + h)]^2 \quad (2)$$

No qual,

- $\gamma(h)$ é a semivariância;
- h é o vetor de separação entre as amostras;
- $n(h)$ é o número de pares de observações amostrais medidos por uma separação de distância h;
- $z(u_i)$ é o valor do atributo u, na posição i;
- $z(u_i + h)$ é o valor do atributo u, na posição i, separado por um vetor de distância h.

Para entendimento do cálculo da semivariância, vamos fazer um gráfico de dispersão das variáveis separadas por uma distância h. Como apresentado na Figura 1.

Figura 1 - Gráfico de dispersão



Fonte: Imagem adaptada do livro. An Introduction to Applied Geostatistics, ISAAKS, E. H; SRIVASTAVA, R. M, 1989.

Trace a reta de correlação, ou reta identidade no gráfico de dispersão, como destacado em vermelho na imagem acima. Note que o ângulo entre a reta identidade e os eixos das abcissas, mede 45^{o}. Para estudar a correlação, vamos calcular a distância (d) do valor observado no ponto de coordenadas $(z(u_i),\ z(u_i\ +\ h))$ até a reta de correlação. Sendo assim, por semelhança de triângulos, pelo teorema de Pitágoras e pela relação trigonométrica no triângulo retângulo, chegamos na equação (3).

$$d = |\, z(u_i)\ -\ z(u_i\ +\ h)|sen\ 45^{\circ} \tag{3}$$

Como, $sen\ 45^{\circ}\ =\ \frac{\sqrt{2}}{2}$, temos da equação (3).

$$d = |\, z(u_i)\ -\ z(u_i\ +\ h)|\ \frac{\sqrt{2}}{2} \tag{4}$$

Elevando ambos os lados ao quadrado, temos.

$$d^2 = \frac{1}{2}[z(u_i) - z(u_i\ +\ h)]^2 \tag{5}$$

A predição da semivariância é a soma de todas as distâncias até a reta de correlação. Sendo assim, chegamos na equação (6).

$$\gamma\,(h) = \frac{1}{n(h)} \sum_{i=1}^{n(h)} {d_i}^2 \tag{6}$$

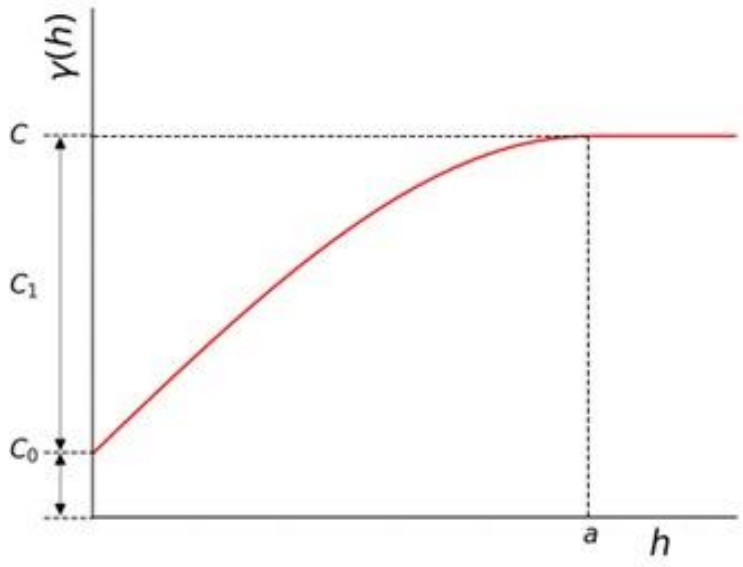
Substituindo a equação (5) em (6) chegamos na equação da semivariância.

$$\gamma(h) = \frac{1}{2\,n(h)} \sum_{\mathrm{i}=1}^{n(h)} [z(u_i) - z(u_i\ +\ h)]^2 \tag{7}$$

2.5. Gráfico da semivariância

O gráfico da semivariância é chamado de semivariograma, que é uma função de $\gamma(h)$ por h (Figura 2).

Figura 2 - semivariograma e suas propriedades: alcance (a), efeito pepita (C_0), patamar (C_0 + C),contribuição (C_1)



Fonte: O autor (2023)

Note que, ao passo que vamos aumentamos a distância de separação, vamos aumentando a semivariância. O que é natural, pois sabemos que amostras mais distantes são menos semelhantes entre si, tendo assim muita variabilidade.

O gráfico da semivariância tem os seguintes parâmetros.

- **Alcance (a):** indica a distância na qual já não há dependência espacial;

- **Efeito pepita ($\mathbf{C_0}$):** um valor positivo que representa a descontinuidade do semivariograma para distâncias inferiores à menor distância de separação;
- **Patamar ($\mathbf{C_0 + C_1}$):** valor de estabilização da semivariância. A partir da distância correspondente a este valor, é inferido que não há qualquer dependência espacial entre as observações.
- **Contribuição ($\mathbf{C_1}$):** é a diferença entre o patamar e o efeito pepita (C_0).

A partir do momento em que modelamos a dependência espacial e escolhemos qual é o melhor semivariograma, partimos para a interpolação por krigagem. Antes de falar da krigagem, vamos estudar alguns modelos de semivariograma mais utilizados na geoestatística.

2.6. Modelos de semivariograma

De acordo com Viera (2000), semivariograma é uma função de correlação espacial sem viés e variância mínima. O modelo de semivariograma a ser ajustado depende do comportamento dos dados. Existem duas categorias de semivariograma: modelos sem patamar e modelos com patamar.

2.6.1. Modelos sem patamar

2.6.1.1. Modelo Linear

O modelo linear é um exemplo de semivariograma sem patamar, com a seguinte equação.

$$\gamma(h) = C_0 + Ah^B, \ 0 < B < 2 \tag{8}$$

Os parâmetros A e B, são constantes do modelo. Para o parâmetro B, temos a restrição $0 < B < 2$, pelo fato de ser necessário uma função que faça sentido prático. Na hora de montar a matriz de correlação, para fazer a interpolação por krigagem, é necessário que esta matriz admita inversa (MENDES, 2020). Assim como veremos mais adiante.

2.6.2. Modelos com patamar

Os modelos com patamares são aqueles nos quais a semivariância aumenta conforme a distância aumenta, até um ponto em que se estabiliza, indicando que não há mais correlação espacial.

Aqui neste tópico, vamos estudar apenas alguns semivariogramas que são mais utilizados na geoestatística, tais como: efeito pepita puro, esférico, exponencial e gaussiano.

2.6.2.1. Efeito pepita puro

A equação do modelo de semivariograma do efeito pepita puro para qualquer distância h, é dado por.

$$\gamma(h) = \mathrm{C_0} \tag{9}$$

O semivariograma do efeito pepita puro indica que não foi possível detectar variabilidade espacial nos dados, apontando assim que o fenômeno é aleatório (MENDES, 2020). O fato de não se conseguir capturar a variabilidade espacial está atrelado a malha utilizada, que foi superior ao alcance e/ou por variações relacionada as medições e coleta das amostras.

2.6.2.2. Semivariograma esférico

O modelo de semivariograma esférico, possui comportamento linear próximo a origem para pequenos valores de h. Este é representado pela seguinte equação.

$$\gamma(\mathrm{h}) = \begin{cases} 0 & , \quad h = 0 \\ \mathrm{C_o} + \mathrm{C_1}\left[\frac{3}{2}\left(\frac{\mathrm{h}}{\mathrm{a}}\right) - \frac{1}{2}\left(\frac{\mathrm{h}}{\mathrm{a}}\right)^3\right] & , 0 < \mathrm{h} < \mathrm{a} \\ \mathrm{C_o} + \mathrm{C_1} & , \quad h > a \end{cases} \tag{10}$$

2.6.2.3. Semivariograma exponencial

De acordo com Isaaks e Srivastava (1989), o modelo de semivariograma exponencial apresenta pouca continuidade para intervalos pequenos. A equação deste modelo é representada pela seguinte equação.

$$\gamma(h) = \begin{cases} C_o + C_1\left(1 - e^{\frac{-3h}{a}}\right) & , \quad 0 < h < a \\ \mathrm{C_o} + \mathrm{C_1} & , \quad h > a \end{cases} \tag{11}$$

2.6.2.4. Semivariograma gaussiano

O modelo gaussiano possui um comportamento parabólico na origem e é o único semivariograma que possui um ponto de inflexão. De acordo com Isaaks e Srivastava (1989), o modelo gaussiano é utilizado para fenômenos contínuos. A equação deste modelo é representada por.

$$\gamma(h) = \begin{cases} 0 & , \quad h = 0 \\ C_o + C_1\left(1 - e^{-3\left(\frac{h}{a}\right)^2}\right) & , \quad h \neq 0 \end{cases} \tag{12}$$

2.7. Interpolação e Krigagem

2.7.1. Interpolação

Interpolação é o ato de predizer o valor de um atributo em áreas que não foram amostradas utilizando informações das amostras vizinhas (MENDES, 2020).

Para que possamos predizer o valor em um ponto não amostrado, precisamos atribuir pesos às amostras vizinhas de acordo com a sua distância. Os pesos variam entre 0 e 1 e o somatório de todos os pesos deve ser igual a 1. Sabemos que amostras mais próximas tendem a ser mais parecidas entre si (SOUZA, *et al.* 2003). Sendo assim, vamos dar mais importância a ela, pois essa amostra nos ajuda a entender melhor o fenômeno do que amostras mais distantes. Atribuímos mais pesos para as amostras mais próximas.

Assim como salienta Miranda (2005), o processo de interpolação é gerado pela relação entre a vizinhança e pela definição de qual método usar para interpolar.

A equação de interpolação é representada da seguinte maneira.

$$\hat{Z}(u) = \sum_{i=1}^{n} \lambda_i Z(u_i) \tag{13}$$

No qual,

- u refere-se a um ponto qualquer que não foi amostrado;
- Z (u) é o valor predito na localização u, onde existem n dados $Z(u_i)$, $i = 1, \ldots, n$ na circunvizinhança de u;
- λ_i refere-se aos pesos atribuídos aos pontos vizinhos.

O meio como atribuímos os pesos depende do método de interpolação que estamos utilizando. Existem métodos que consideram a distância cartesiana e outros que consideram a distância estatística, que é a correlação. Amostras mais próximas, tendem estar mais correlacionadas (MENDES, 2020).

Alguns métodos de interpolação como inverso da distância, polígonos de influência e splines, dentre outros, são muito utilizados para fazer interpolação e são conhecidos como métodos determinísticos (CAMARGO, 1997). Estes, por sua vez, são simples, intuitivos e facilmente programáveis no computador. No entanto, apresentam a desvantagem de serem métodos que não consideram o padrão de variabilidade espacial, e para os estudos de ciências dos solos, a variabilidade espacial é importante.

Assim como apresentado por Costa e Souza (2014), alguns fatores devem ser considerados na hora de fazermos a interpolação, tais como: proximidade das amostras e redundância amostral, ou seja, amostragem preferencial, anisotropia e magnitude espacial.

Um método que atende essas exigências é a Krigagem, que leva em consideração as características espaciais de autocorrelação das variáveis regionalizadas.

2.8. Krigagem Ordinária

Segundo Ribeiro Junior (1995) e Camargo (1997) o processo de interpolação por krigagem se diferencia das demais técnicas, pela forma como atribui os pesos as amostras. O peso é atribuído utilizando o semivariograma, que foi utilizado para modelar a continuidade espacial.

A krigagem ordinária é considerada como o melhor método de predição, pois dentre os métodos de interpolação, este é o que apresenta menor variância do erro de predição (ISAAKS; SRIVASTAVA, 1989). Krigagem também é considerado um preditor não viesado. Sendo assim, krigagem é um preditor linear não viesado de variância mínima (BLUP).

Os pesos da krigagem ordinária são calculados a partir da resolução do seguinte sistema matricial.

$$\begin{bmatrix} C(Z(u_1),Z(u_1)) & C(Z(u_1),Z(u_2)) & \cdots & C(Z(u_1),Z(u_n)) & 1 \\ C(Zu_2),Z(u_1)) & C(Z(u_2),Z(u_2)) & \cdots & C(Z(u_2),Z(u_n)) & 1 \\ \vdots & \vdots & \ddots & \vdots & 1 \\ C(Z(u_n),Z(u_1)) & C(Z(u_n),Z(u_2)) & \cdots & C(Z(u_n),Z(u_n)) & 1 \\ 1 & 1 & & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ -\mu \end{bmatrix} = \begin{bmatrix} C(Z(u_0),Z(u_1)) \\ C(Z(u_0),Z(u_2)) \\ \vdots \\ C(Z(u_0),Z(u_n)) \\ 1 \end{bmatrix}$$

No qual,

$C = \begin{bmatrix} C(Z(u_1), Z(u_1)) & C(Z(u_1), Z(u_2)) & \cdots & C(Z(u_1), Z(u_n)) & 1 \\ C(Zu_2), Z(u_1)) & C(Z(u_2), Z(u_2)) & \cdots & C(Z(u_2), Z(u_n)) & 1 \\ \vdots & \vdots & \ddots & \vdots & 1 \\ C(Z(u_n), Z(u_1)) & C(Z(u_n), Z(u_2)) & \cdots & C(Z(u_n), Z(u_n)) & 1 \\ 1 & 1 & & 1 & 0 \end{bmatrix}$, é a matriz de covariância entre os valores observados $Z(u_i)$ nos pontos amostrados u_i, $i = 1 \ldots n$.

$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ -\mu \end{bmatrix}$, é o vetor de pesos.

$D = \begin{bmatrix} C(Z(u_0), Z(u_1)) \\ C(Z(u_0), Z(u_2)) \\ \vdots \\ C(Z(u_0), Z(u_n)) \\ 1 \end{bmatrix}$, é o vetor de covariância entre os pontos amostrados, $u_i, i = 1, \ldots, n$, e o ponto no qual se quer predizer u_0.

Logo, o vetor de pesos da krigagem ordinária é obtido pela resolução do seguinte produto matricial.

$$\boldsymbol{\lambda = C^{-1}D} \tag{14}$$

Note que para calcularmos o vetor de pesos λ, vamos precisar multiplicar o inverso na matriz $\boldsymbol{C}$, pela matriz $\boldsymbol{D}$. A matriz $\boldsymbol{C}$ é obtida pelo semivariograma e para que ela admita inversa, precisamos escolher os modelos de semivariogramas que permita que isto aconteça.

2.9. Relação entre covariograma e semivariograma

Note que para calcular a matriz $\boldsymbol{C}$, no sistema matricial da krigagem ordinária, precisamos das covariâncias entre as amostras, mas sabemos que o semivariograma modela a semivariância. A seguinte equação representa a relação entre covariância e semivariância (COSTA; SOUZA, 2014).

$$C(h) = C(0) - \gamma(h) \tag{15}$$

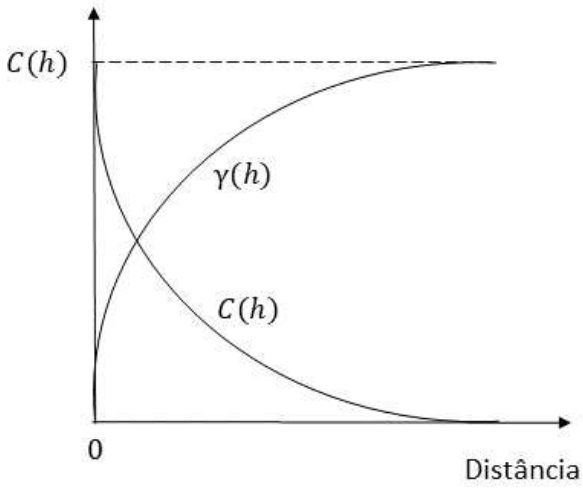
Em que,

- $C(h)$ é a função de covariância para uma distância h;
- $C(0)$ é a função de covariância para a distância 0;

- $\gamma(h)$ é a função de semivariância para a distância h.

A título de ilustração a Figura 3 representa a relação entre covariograma e semivariograma.

Figura 3 - Gráfico de covariância e semivariância



Fonte: o autor (2023)

Ao passo que aumentamos a distância, a semivariância também aumenta e por consequência a covariância diminui, tendendo a zero. Note que o gráfico de covariograma é o inverso do semivariograma.

2.10. Validação Cruzada

Para avaliar a qualidade das predições do método de interpolação da geoestatística fazemos o uso da validação cruzada (FERREIRA, 2015). A validação cruzada parte do princípio de que um elemento da amostra não foi previamente mensurado. Para prever este atributo, usamos as técnicas de krigagem, no qual utilizamos os valores das amostras vizinhas para predizer este valor ausente. Esse procedimento é aplicado a todos os pontos amostrados. Sendo assim, para cada um desses pontos, teremos o valor verdadeiro e o valor predito pela krigagem, o que nos permite calcular as estatísticas de erro de predição. Existem várias técnicas de validação cruzada, porém serão citadas apenas as que foram utilizadas neste trabalho.

2.10.1. Raiz quadrada do erro quadrático médio (RMSE)

A raiz quadrada do erro quadrático médio é muito empregada com o intuito de avaliar a qualidade da predição de um método estatístico. Para escolher o melhor método de interpolação por krigagem, por exemplo, é escolhido o semivariograma que apresentar o menor RMSE (FERREIRA, 2015). Sua equação é dada por:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{\left(\hat{Z}(u_i) - Z(u_i)\right)^2}{n}} \quad (16)$$

No qual, $\hat{Z}(u_i)$ é o valor predito pela krigagem e $Z(u_i)$ é o valor verdadeiro para o ponto amostrado i e n é o número de pontos amostrados.

2.10.2. Erro absoluto médio (MAE)

É o cálculo da média da diferença absoluta entre os valores preditos e o valor real, suja equação é dada por:

$$MAE = \frac{\sum_{i=1}^{n} \left|\hat{Z}(u_i) - Z(u_i)\right|}{n} \quad 17$$

2.11. *Machine Learning* e geoestatística

Temos um grande problema quando o quesito é a amostragem na geoestatística. Para fazermos a interpolação por krigagem, é importante que cada ponto num semivariograma seja obtido com base na combinação de no mínimo 30 pares de pontos (GUERRA, 1988). Além do mais, de acordo com Webster e Oliver (2007), é importante ter 100 pontos amostrais para conseguirmos fazer a krigagem. Sendo assim, é imprescindível planejar a quantidade de amostras adequadas, pois sabemos que o processo de amostragem e análise de solo é caro e isso pode dificultar a execução do processo (BOLFE; GOMES, 2005; MENDES *et al*., 2020).

Sabemos que é importante a proximidade das amostras para que a interpolação consiga capturar as manchas de variabilidade (GREGO; OLIVEIRA; VIEIRA, 2014). Para direcionar as amostras, podemos utilizar dados históricos da área (VIEIRA; XAVIER; GREGO, 2008) e podemos utilizar imagens aéreas, nas quais identificamos as diferenças nos índices de vegetação

e, com isso, é possível diminuir o adensamento amostral em áreas homogêneas e priorizar as áreas que têm maior variabilidade (GREGO; OLIVEIRA; VIEIRA, 2005).

Dentro desta nova era digital em que a agricultura está vivenciando, faz-se uso de várias técnicas para análise dos dados como big data, deep learning, realidade aumentada, robotização, inteligência artificial, machine learning, dentre outras (BORÉM, 2021).

De acordo com Iaco *et al*. (2022), machine learning tem sido muito utilizado para resolver problemas em ciências do solo, como manejo da terra (RODRIGO-COMINO *et al*., 2018), definição de zonas de manejo (CASTRO-FRANCO *et al*., 2018), mapeamento dos diferentes tipos de solo (DEMATTÊ; RIZZO; BOTTEON, 2015), previsão de rendimento das lavouras (ADAMCHUK *et al*., 2017), dentre outros.

Uma rápida pesquisa realizada por Iaco *et al*. (2022) mostrou que a palavra "machine learning" tem aparecido 105 e 319 vezes nas revistas Mathematical Geosciences e Computers and Geosciences do IAMG, respectivamente, desde o ano de 2016, o que indica o crescente número de artigos publicados sobre machine learning para análise de dados em ciências do solo (IACO *et al*., 2022).

2.12. *Machine Learning*

Machine learning tem sido muito utilizado no meio agrícola por fornecer soluções mais assertivas (BURDETT; WELLEN, 2022) e pela sua capacidade de lidar com problemas complexos e não lineares (PANTAZI *et al*., 2016; TANTALAKI; SOURAVLAS; ROUMELIOTIS, 2019).

Os modelos de *machine learning* são divididos em aprendizado supervisionado e aprendizado não supervisionado. Para os métodos de aprendizagem supervisionada, consideramos que, para cada observação, há uma medida associada, e o objetivo é entender a relação entre as variáveis preditoras e a resposta (JAMES *et al*., 2013, p. 26).

Por outro lado, o aprendizado não supervisionado descreve um desafio maior, pois temos um vetor de observações, mas nenhuma resposta associada. Assim, não é possível ajustar um modelo de regressão, uma vez que não há nenhuma variável resposta para prever. Neste cenário, estamos trabalhando às cegas e deixamos o algoritmo aprender com os dados de acordo com seus padrões (JAMES *et al*., 2013, p. 26).

2.13. Aprendizado Supervisionado

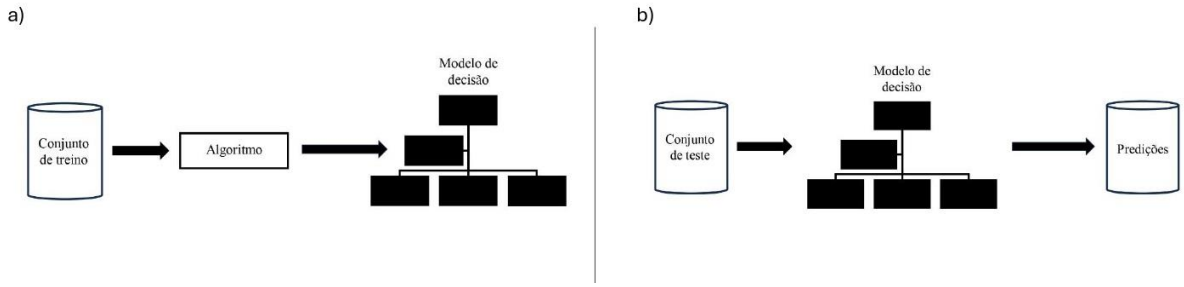
No aprendizado supervisionado, o conjunto de dados é dividido em conjunto de treinamento e conjunto de teste. Um modelo é treinado usando um conjunto de exemplos em que as entradas estão associadas às saídas correspondentes. Para validar este modelo, usamos o conjunto de teste.

Do ponto de vista matemático, queremos encontrar uma função $g(x)$, na qual denominamos de função classificadora, que se aproxime da função real, $f(x_i)$ (RUSSELL; NORVIG, 2002).

Em outras palavras, o conjunto de treino é da forma $Y = \{(x_1, z_1), (x_2, z_2), .., (x_m, z_m)\}$, em que $x \in X$ e $z \in Z$. X é a variável de entrada e Z é a variável de saída. O algoritmo irá aprender a partir do conjunto Y e depois, quando novas entradas forem apresentadas à função de classificação, o algoritmo irá tomar decisões com base no seu treinamento.

Para ilustrar o aprendizado supervisionado, observe a Figura 4.

Figura 4 - Fase de Treinamento a) e fase de teste b)



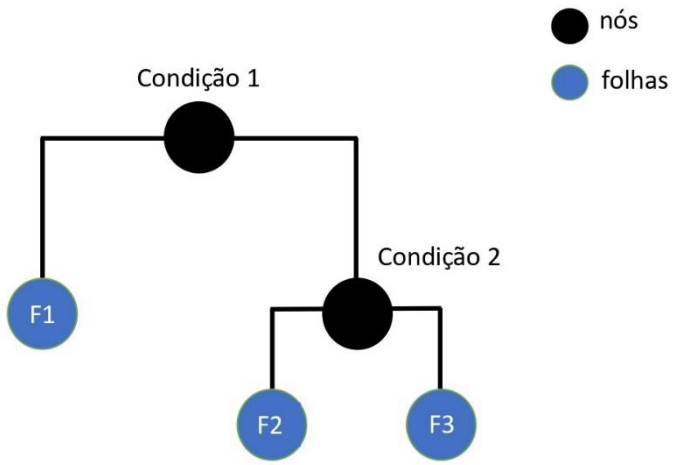
Fonte: o autor (2023)

2.14. *Decision tree*

Uma metodologia muito útil e de fácil interpretação em *machine learning*, para aprendizado supervisionado, é a *decision tree* ou árvore de decisão (JAMES *et al.*, 2013), que é uma técnica não paramétrica (IZBICKI; SANTOS. 2020). Assim como mencionado por James *et al.* (2013), *decision tree* podem ser utilizadas para resolver problemas de classificação e regressão. Para desenvolvimento deste trabalho vamos nos atentar apenas a resolver problemas de regressão, no qual, sempre que mencionarmos *decision tree*, retrataremos a árvores de regressão.

Uma *decision tree* é muito parecido com uma árvore tradicional, porém invertida, no qual possui folhas, ramos e raízes. As regiões de separação são denominadas de nós, os ramos são os valores do atributo de um nó e as folhas são os resultados das predições. A título de exemplificação, consideremos a Figura 5.

Figura 5 - Estrutura de uma decision tree



Fonte: o autor (2023)

Para prever um resultado com uma *decision tree* incialmente verifica a condição do primeiro nó. Se a condição for satisfeita, segue para a esquerda. Caso não seja, segue para a direita. Este processo é realizado até chegar em uma folha. Para o caso descrito na Figura 5, temos como predição F1, se a condição 1 for satisfeita. Caso não seja, segue para a direita e olha para a condição 2. Temos como predição F2 se a condição 2 for satisfeita. Caso não seja, temos F3 como predição.

Matematicamente, o processo de construção de uma árvore se dá pelo particionamento em conjuntos disjuntos do espaço das covariáveis (IZBICKI; SANTOS, 2020), ou seja, vamos dividir o nosso espaço amostral em regiões $R_1, R_2, .., R_n$, em que.

- $R_i \cap R_j = \emptyset$, para quaisquer $i \neq j$;
- $R_1 \cup R_2 \cup, .., \cup R_n = U$, em que U, é o conjunto universo das covariáveis.

A predição para a resposta associada, digamos Z, de covariáveis X é definida como sendo a média das observações de treinamento (JAMES *et al.*, 2013), entre todas as amostras que caíram naquela folha, como representado na equação (18).

$$g(x) = \frac{1}{|\{i : x_i \in R_n\}|} \sum_{i:\, x_i \in R_n} Z_i \tag{18}$$

Em que, x_i é a amostra e, Z_ié o valor mensurado da amostra x_i.

De acordo com Izbicki e Santos (2020), a estrutura e o processo de construção de uma *decision tree* é feita em duas etapas.

I. Criação de uma árvore completa;

II. Poda da árvore.

Para atender as exigências da etapa I, é necessário criar uma árvore com partições "puras", ou seja, os valores Z das covariáveis X, são particionadas de forma recursiva até que as folhas sejam homogêneas (IZBICKI; SANTOS, 2020). Consideremos T, como sendo uma *decision tree*. O objetivo é encontrar T, que minimize o erro quadrático médio $RSS\ (T)$ (JAMES *et al.*, 2013), expressado na equação abaixo.

$$RSS(T) = \sum_{R} \sum_{i:\, x_i \in R} \frac{\left(Z_i - \hat{Z}_R\right)^2}{n} \tag{19}$$

Em que $\hat{Z}_R$ é o valor predito.

Assim como justifica Izbicki e Santos (2020), descobrir a *decision tree* T, que minimize o RSS (T) é inviável. Sendo assim, para encontrar uma decison tree com baixo valor de RSS, será necessário fazer divisões binárias recursivas em cada nó e avaliar a combinação entre a covariável e o parâmetro de corte que leva ao menor RSS (T) (JAMES *et al.*, 2013). Para selecionar essa partição, procura-se a combinação ideal entre todas as variáveis x_i e os pontos de corte t_1, de modo a obter uma divisão (R_1, R_2) que minimize o erro quadrático das predições (Equação(20)).

$$\sum_{i:\, x_i \in R_1}^{n} (Z_i - \hat{Z}_{R_1})^2 + \sum_{i:\, x_i \in R_2}^{n} (Z_i - \hat{Z}_{R_2})^2 \tag{20}$$

No qual, $\hat{Z}_{R_k}$ é o valor predito para a região R_k ($k = 1,2..,n$). Note que neste caso, estamos apenas calculando o erro quadrático médio apenas para dois ramos com a seguinte divisão.

$$R_1 = \{x_i < t_1\} e\, R_2 = \{x_i \geq t_1\} \quad (21)$$

No qual, x_i é a covariável e t_1é o parâmetro de corte. Este processo de divisão binária é realizado consecutivamente até o ponto que tenhamos poucas observações em uma folha (IZBICKI; SANTOS, 2020).

Agora que temos a *decision tree*, será necessário fazer a poda para diminuir a variância do estimador (IZBICKI; SANTOS, 2020). A poda é realizada retirando um nó de cada vez e avaliando como o erro de predição se comporta no conjunto de validação (IZBICKI; SANTOS, 2020).

3. MATERIAIS E MÉTODOS

3.1. Software utilizado e principais pacotes

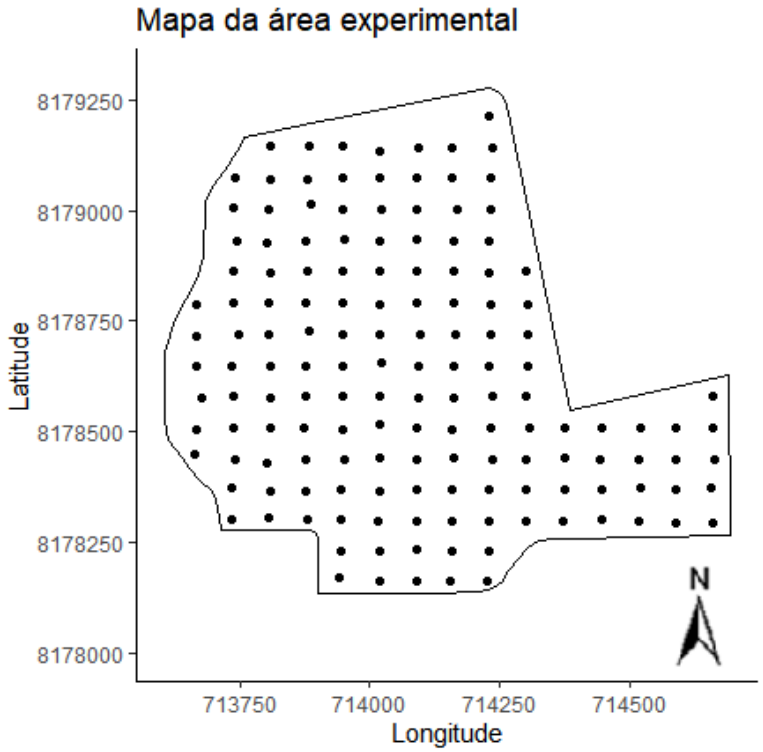
Para a análise dos dados, foi utilizado o software R (R DEVELOPMENT CORE TEAM, 2023) versão 4.3.1 e os principais pacotes utilizados foram o Rpart para construção da *decision tree*, o geoR para fazer a interpolação por krigagem e o ggplot para construção dos mapas de atributo do solo.

3.2. Banco de dados e área de estudos

Os dados utilizados neste estudo foram gentilmente fornecidos pelo professor Marcelo Marques Costa e incluem informações sobre os atributos químicos e físicos do solo. Esses dados foram coletados em um talhão da fazenda "Andréia ou Sozinha", localizada em Goianápolis, no Centro-Oeste do Brasil. A região possui solo classificado como Latossolo Vermelho Amarelo.

Foram coletadas 150 amostras em uma área de 75 hectares, como mostrado na Figura 6.

Figura 6 - Mapa da área experimental



Fonte: O autor (2023)

3.3. Análises químicas e físicas realizadas

Para mensurar os atributos químicos e físicos do solo, as amostras foram coletadas usando um trado de rosca com a profundidade de 0 a 0,20 m. E, a amostra foi composta por 10 subamostras do solo, que foram coletadas de forma aleatória, variando em um raio de 5 m do ponto de coleta.

Para análise química do solo, foram mensurados os seguintes componentes: pH, acidez potencial (H + Al), teores de cálcio (Ca), magnésio (Mg), potássio (K), alumínio (Al), fósforo (P), fósforo remanescente (Prem), matéria orgânica, soma de bases (SB), capacidade de troca catiônica (CTC) efetiva (t) e CTC total, saturação por bases (V) e os teores de zinco (Zn), ferro (Fe), manganês (Mn) e cobre (Cu). Já, para análises físicas, foram mensurados os teores de argila, silte e areia total.

3.4. Atributos utilizados

Quando realiza a análise de solos, os laboratórios entregam para o produtor a mensuração de um conjunto de atributos que são considerados de rotina. O produtor paga por

um valor fixo para a mensuração destes atributos. Caso o cliente queira analisar mais algum componente, é cobrado separadamente por cada atributo.

Os atributos de rotina são: pH-H2O, P, K, Ca, Mg, Al, H+Al, P-rem, SB, t, T e V. Para desenvolvimento deste trabalho, utilizaremos os atributos de rotina para predizer uma porcentagem dos micronutrientes (Fe, Zn, Mn e Cu), que são atributos muito importante para correção do solo e que são cobradas um valor a mais para mensurar cada um destes micronutrientes.

Como SB, t, T e V são combinações de outros atributos, foi decidido não as utilizar, para não afetar o poder de predição. Sendo assim, foram utilizados apenas atributos que não são função de outros. Em resumo, utilizamos os atributos pH-H2O, P, K, Ca, Mg, H+Al e P-rem para predizer Fe, Zn, Mn e Cu.

3.5. Redução do adensamento amostral

A base de dados original utilizada contém uma malha de 150 pontos amostrados e georreferenciados. Neste estudo foram avaliadas as reduções da densidade da malha em 15%, 25% e 45% para os micronutrientes o que correspondem a malhas reduzidas contendo 127, 112 e 82 pontos amostrados. Daqui em diante identificadas como MR_{127}, MR_{112} e MR_{82}.

Estas malhas reduzidas foram obtidas a partir da seleção de pontos da malha original ao utilizar o algoritmo *Latin Hypercube Sampling (LHS*) proposto por Mickey *et al.* (1979). Em resumo, o algoritmo recebe a malha original contendo todos os pontos amostrados como entrada e, produz como resultado grades reduzidas contendo i pontos uniformemente distribuídos em toda a área de estudo (PEREIRA *et al.*, 2022), tal que $i = 82, 112$ e 127.

Para diminuir o efeito da variabilidade das predições, para cada uma das reduções, foram retiradas 50 amostras originando assim as malhas reduzidas MR_{ij} tal que $i = 127, 112$ e 82 e $j = 1, \cdots, 50$.

3.6. Construção da *decision tree*

Após a redução dos pontos, o método *decision tree* foi utilizado em cada uma das grades reduzidas para preencher a informação faltante do atributo, retirado por LHS, e substituído na grade reduzida contendo os demais pontos.

O critério de parada das *decision trees*, diz o ponto em que a divisão de um nó deve ser parada. Para este estudo em questão foi utilizado como critério de parada até ao passo em

que o nó tenha apenas uma observação. Este critério adotado, certamente está criando árvores que tenha um ajuste excessivo (*overfitting*) no conjunto de treinamento, mas este fato não é tão preocupante, pois é realizado a poda de todas as árvores e, como é sabido, a poda diminui a variabilidade da árvore, evitando assim o *overfitting* (IZBICKI*;* SANTOS, 2020; JAMES *et al.*, 2013).

O parâmetro de corte utilizado foi o menor valor de CP (*complexity parameter)*, que é baseado no *xerror* (erro cruzado). O erro cruzado é determinado pela comparação das previsões do modelo com os valores reais no conjunto de teste. A métrica de erro utilizada varia de acordo com o tipo de problema. Em problemas de regressão, por exemplo, é comum utilizar o erro quadrático médio ($RSS(T)$, como descrito na Equação (19).

Como está sendo utilizado aprendizado supervisionado, precisamos de um conjunto de treinamento e outro conjunto de teste. Para o treinamento do modelo da *decision tree* foram utilizadas as amostras selecionadas pelo algoritmo LHS e para o conjunto de teste foi utilizado o complementar do conjunto de treinamento. Este processo de criação da *decision tree*, foi realizado para todas as repetições, MR_{ij} tal que $i = 127, 112$ e 82 e $j = 1, \cdots, 50$, dos micronutrientes.

Para avaliar a qualidade das predições dadas pelas *decision trees*, será utilizado o RMSE médio das 50 *decision trees*, definidas como $RMSE\_DT$, cuja equação é definida como sendo:

$$RMSE\_DT = \frac{\sqrt{\sum_{i=1}^{n} \frac{\left(\hat{Y}(x_i) - Y(x_i)\right)^2}{n}}}{50} \tag{22}$$

No qual $\hat{Y}(x_i)$, representa o valor predito pela *decision tree* e, $Y(x_i)$ o valor real, n é o número de dados no conjunto de teste ($n = 23, 38\ e\ 68$) e $i = 1, .., n$.

3.7. Interpolação por krigagem

Os modelos esférico, exponencial e gaussiano foram utilizados como modelagem da continuidade espacial para cada micronutriente. Os semivariogramas para a malha original e para as malhas MR_{127}, MR_{112} e MR_{82}, foram ajustados utilizando o método dos mínimos quadrados ordinários, *ordinary least squares* (OLS). Quando o método OLS indicava que o

melhor modelo era o efeito pepita puro, não foi realizado a krigagem ordinária e o algoritmo parava de rodar.

Para escolha do melhor modelo de semivariograma teórico, foi utilizado o erro quadrático médio (MSE). O modelo escolhido foi aquele que resultou em menor valor de MSE.

As grades reduzidas foram preenchidas com os valores preditos por *decision tree*. Este processo foi realizado 50 vezes e logo em seguida foi realizado a interpolação por krigagem ordinária para a média destas 50 amostragens e foram obtidos os gráficos das malhas MR_{127}, MR_{112} e MR_{82}.

A interpolação por krigagem ordinária também foi realizada para cada um dos atributos Fe, Mn, Zn e Cu na malha de dados original. Para a malha original foi realizado a krigagem ordinária estabelecendo uma vizinhança de busca com um mínimo de 8 e um máximo de 10 pontos ao entorno do ponto que se quer predizer. Foi utilizado um grid composto por 10000 pontos dentro da área estudada. Os mesmos parâmetros estabelecidos para realizar a krigagem ordinária na malha original foram mantidos para as malhas MR_{127}, MR_{112} e MR_{82}.

A média da validação cruzada das interpolações da krigagem combinada com *decision tree*, foi utilizada para comparar com as estatísticas de validação da krigagem ordinária com a grade original. Essas comparações permitiram conhecer o quanto é possível reduzir o adensamento amostral para se obter mapas de interpolação associando a krigagem ordinária com *decision tree.*

4. RESULTADOS E DISCUSSÕES

4.1. Análise descritiva

Com o intuito de entender e sumarizar o comportamento dos dados, foi realizado uma análise descritiva dos atributos, como observado na Tabela 1.

Observando a Tabela 1, notamos que o atributo Potássio (K), apresenta a maior variância dentre todos os atributos. Para as variáveis de rotina (pH-H2O, P, K, Ca, Mg, H+Al e P-rem), o Magnésio (Mg), apresenta a menor variância.

Quando olhamos para os micronutrientes, podemos observar que o cobre tem menor variância ($\sigma^2 = 0{,}40$) e o atributo que possui maior variância é Mn ($\sigma^2 = 0{,}8610$).

Tabela 1 - Estatística descritiva dos dados originais

Atributo	Média	Mínimo	Máximo	S^2	CV(%)	Assimetria	Curtose
pH_H_2O	6,75	5,80	7,60	0,10	4,43	-0,74	0,49
P	6,84	1,70	21,60	15,70	57,88	1,27	1,60
K	52,63	24,00	108,00	201,70	26,98	1,02	1,86
Ca	3,27	1,90	4,20	0,20	14,04	-0,39	-0,10
Mg	0,84	0,60	1,40	0,0196	16,53	0,72	1,42
H+Al	1,72	0,00	5,60	0,80	51,92	1,51	4,17
P-rem	17,35	9,50	27,40	11,40	19,48	0,43	0,24
Zn	3,94	1,50	27,10	8,60	74,52	4,66	29,51
Cu	1,33	0,80	6,70	0,40	44,35	5,38	43,28
Fe	21,90	11,00	41,10	30,30	25,12	0,88	0,56
Mn	27,23	13,70	66,70	86,10	34,07	1,80	4,09

potencial hidrogeniônico (pH_H_2O), Fósforo (P), Potássio (K), Cálcio (Ca), Magnésio (Mg), acidez potencial (H+Al), P-rem Fósforo remanescente, Zinco (Zn), Cobre (Cu), Ferro (Fe), Manganês (Mn). variância amostral (S^2), Coeficiente de variação CV (%).

Com o intuito de entender a variabilidade relativa entre os dados, vamos olhar para o $CV(\%)$, que é obtido pela razão entre o desvio padrão e a média de cada atributo. Existem alguns critérios para comparar a variabilidade dos dados. Destaco aqui, os limites impostos por Vanni (1998), de que: $CV(\%) < 35\%$, indica que os dados são homogêneos e a média é bem representativa, para, $35\% < CV(\%) < 65\%$, indica que os dados são heterogêneos e a média é pouco representativa, já para, $CV(\%) > 65\%$, indica heterogeneidade do dados e média nada representativa.

Sendo assim, podemos dizer, de acordo com a Tabela 1, que os atributos homogêneos são: pH-H2O, K, Ca, Mg, P-rem, Fe e Mn. Sendo que pH-H2O é o atributo com menor valor de $CV(\%)$, indicando assim, que ele é o mais homogêneo entre todos os atributos.

De acordo com a Tabela 1, os atributos aos quais possuem assimetria a esquerda são: pH_H_2O e Ca. Já os demais, possuem assimetria a direita. Quando olhamos para o coeficiente de curtose, os atributos: pH_H_2O, P, K, Ca, Mg, P-rem e Fe, são classificados como distribuição platicúrtica e os atributos: H+Al, Zn, Cu e Mn são leptocúrtica. Observe que não há atributos com valores de curtose aproximadamente igual a 3, o que indica que não há distribuições que se assemelham a distribuição normal.

Algo de interessante a se observar é que o atributo cobre, possui o maior valor de assimetria e o maior valor de curtose. Sendo assim, este atributo possui caudas mais leves e picos mais achatados e assimetria a direita.

4.2. Predições por *decision tree* das malhas reduzidas (MR_{127}, MR_{112} e MR_{82}) para os micronutrientes

Para a malha MR_{127}, foram utilizadas 127 amostras para treinamento e 23 para teste, para a malha MR_{112}, 112 amostras para treinamento e 38 para teste e para a malha MR_{82}, 82 amostras para treinamento e 68 para teste.

Para a criação das *decision trees*, podemos ressaltar algumas diferenciações quanto a quantidade de folhas quando se cria as árvores e ao podá-las, para cada micronutriente. Assim como pode ser observado na Tabela 2.

Tabela 2 - Quantidade de folhas das 50 *decision trees* sem e após a poda para os micronutrientes Fe, Mn e Zn

Atributo	Quantidade de folhas sem a poda	Quantidade de folhas após a poda
Fe	MR_{127} = 20 a 30	MR_{127} = 4, 3 e 2
	MR_{112} = 20 a 30	MR_{112} = 4, 3 e 2
	MR_{82} = 15 a 20	MR_{82} = 1, 2, 3, 4, 5, 6, 8 e 9
Mn	MR_{127} = 20 a 25	MR_{127} = 1 e 4
	MR_{112} = 15 a 20	MR_{112} = 1 e 4
	MR_{82} = 15 a 20	MR_{82} = 1, 2, 3 e 5
Zn	MR_{127} = 12 a 15	MR_{127} = 1
	MR_{112} = 12 a 15	MR_{112} = 1, 2, 6, 7, 8 e 9
	MR_{82} = 8 a 14	MR_{82} = 1, 2, 3, 4, 6 e 7

Para entendimento da Tabela 2, vamos considerar como exemplo o atributo Fe. Para a malha MR_{127}, a quantidade de folhas das 50 *decision trees*, obtidas pela amostragem LHS, variaram entre 20 e 30. Quando ocorreu a poda, a quantidade de folhas diminuiu para 4, 3 e 2 folhas.

Após a poda, para a malha MR_{127} do atributo Mn, quase que metade das árvores tiveram 1 e 4 folhas. O que também se repetiu para a malha MR_{112}, deste mesmo atributo.

Algo de interessante que aconteceu com a malha MR_{127} do atributo Zn, e que diferencia dos atributos Fe e Mn, é que ao realizar a poda das *decision trees,* todas as 50 árvores

tiveram apenas uma folha. Sendo assim, todas as 23 amostras do conjunto de teste que caíram naquela folha tiveram as mesmas predições, pois é sabido que a predição de cada folha é dada pela média de todas as amostras daquele conjunto (JAMES *et al.*, 2013).

Foi observado que para todos os micronutrientes algumas folhas tiveram apenas uma única observação. Sendo assim, o atributo que caiu naquela folha não sofreu nenhuma alteração.

Como pode ser observado na Tabela 2, ao podar as árvores, nota-se que ao aumentar redução amostral, a quantidade de folhas nas *decision trees* aumenta. Isso pode ser justificado pelo fato das árvores se tornarem mais profundas para acomodar os padrões mais detalhados nos dados de teste (IZBICKI; SANTOS, 2020; JAMES *et al.*, 2013).

O micronutriente Cu, não apresentou dependência espacial para a malha original, sendo assim, não foi realizado as reduções para este atributo, pois a nossa proposta é estudar krigagem ordinária combinadas com *decision tree*.

James *et al.* (2013) e Izbicki*;* Santos (2020), afirmam que as árvores de decisão são de fácil interpretação. Eles alertam também que as árvores de decisão têm algumas limitações, como propensão ao *overfitting* em conjuntos de dados pequenos e sensibilidade a pequenas variações nos dados de entrada. Apesar dessas limitações, sua versatilidade e facilidade de interpretação continuam a torná-las uma escolha popular em muitas aplicações de *machine learning* (JAMES *et al.*, 2013).

Tabela 3 - RMSE médio das 50 *decision trees* (RMSE_DT)

Atributo	**Malhas**	**RMSE_DT**
Fe	MR_{127}	4,92
	MR_{112}	5,20
	MR_{82}	5,45
Mn	MR_{127}	6,35
	MR_{112}	8,48
	MR_{82}	8,93
Zn	MR_{127}	1,53
	MR_{112}	2,58
	MR_{82}	2,70

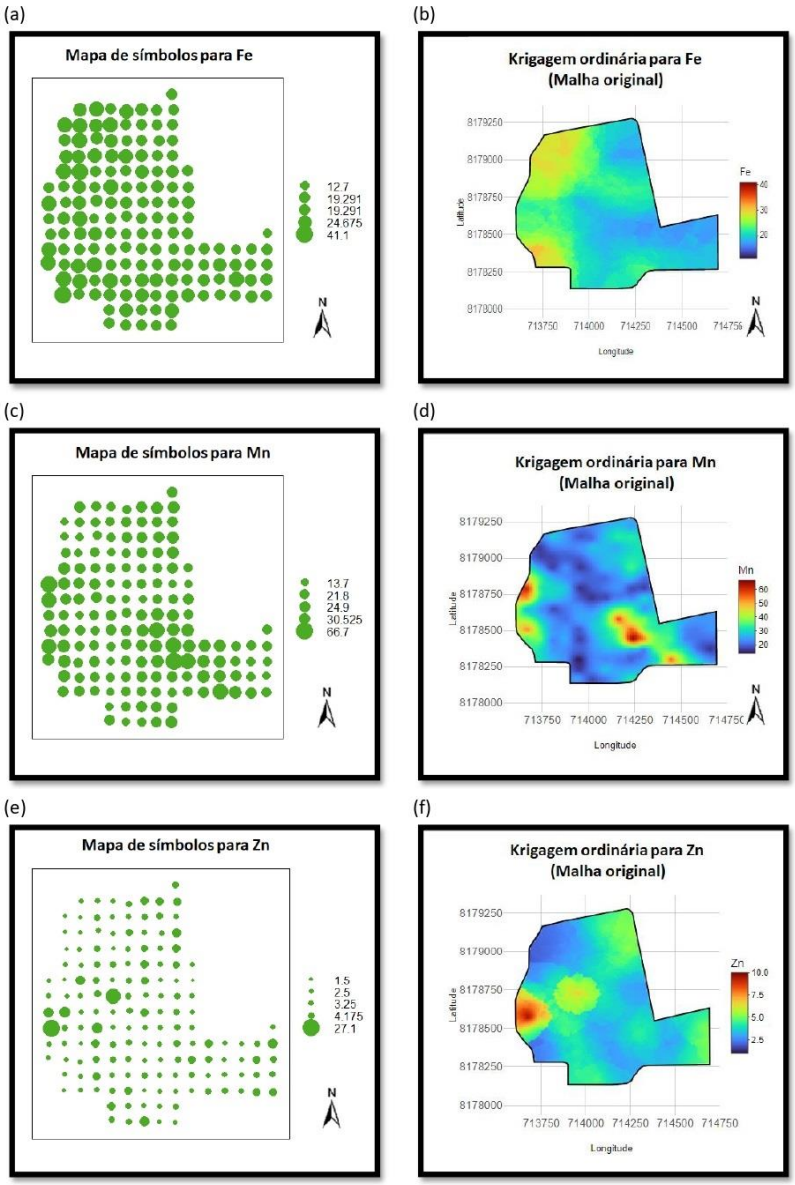
Observando os RMSE médios (RMSE_DT) na Tabela 3, nota-se que, à medida que o conjunto de treinamento do algoritmo *decision tree* diminui, o RMSE médio aumenta, indicando um aumento nos erros de predição do modelo. Esse resultado é esperado, já que o

modelo foi treinado com uma quantidade reduzida de informações, o que resulta em predições menos precisas devido à insuficiência de dados que capturem adequadamente a complexidade e a variabilidade dos dados reais (RAMEZAN *et al*., 2021).

4.3. Krigagem ordinária da malha original para os micronutrientes

Antes de iniciar o processo de redução do adensamento amostral para os micronutrientes, foi realizado a interpolação para a grade original destes atributos, no qual obtivemos os gráficos como apresentado na Figura 7.

Figura 7 - Mapas interpolados por krigagem ordinária para a malha original dos micronutrientes. (a) mapa de símbolos para o Fe, (b) mapa interpolado para o Fe, (c) mapa de símbolos para o Mn, (d) mapa interpolado para o Mn, (e) mapa de símbolos para o Zn, (f) mapa interpolado para o Zn.



Fonte: O autor (2023)

A Figura 7 (a), (c) e (e), representam o mapa de símbolos dos atributos Fe, Mn e Zn, respectivamente, no qual a dimensão das bolhas indicam o seu grau de concentração. Quanto maior o diâmetro da bolha, maior a concentração do atributo no ponto e vice-versa.

Observe que na Figura 7, não é apresentado o gráfico para o atributo Cu, pois este não apresentou dependência espacial. Com isso não foi aplicado krigagem ordinária, pois as técnicas da geoestatística são utilizadas para variáveis que apresente dependência espacial (YAMAMOTO; LANDIM, 2013).

Os parâmetros e os modelos de semivariograma que melhor se adequaram aos dados da malha original são apresentados na Tabela 4.

Tabela 4 - Parâmetros e modelos de semivariogramas para a malha original

Atributo	**Semivariograma**	**Alcance** (a)	**Efeito pepita** (C_0)	**Patamar** $(C_0 + C)$
Fe	Modelo esférico	249,47	5,15	30,03
Mn	Modelo esférico	187,11	19,93	93,94
Zn	Modelo esférico	343,02	4,36	7,82

De acordo com a Tabela 4, o modelo de semivariograma que melhor se ajustou aos atributos Fe, Mn e Zn foi o modelo esférico, o que corrobora com pesquisas de Vieira (2000), Corá *et al.* (2004), Souza *et al.* (2003), Grego*;* Vieira (2005), Cambardella *et al.* (1994), Boyer *et al.* (1996), Albuquerque *et al.* (1996), Tsegaye*;* Hill, (1998) e Paz-González *et al.* (2000), que afirmam que este modelo é o de maior ocorrência para atributos químicos do solo.

Na Figura 7 (b), (d) e (f), são apresentados os mapas interpolados por krigagem ordinária para os atributos Fe, Mn e Zn. Nestes mapas podemos observar zonas com maiores e menores concentrações de Fe, Mn e Zn e que estão em consonância com os mapas de símbolos para estes micronutrientes.

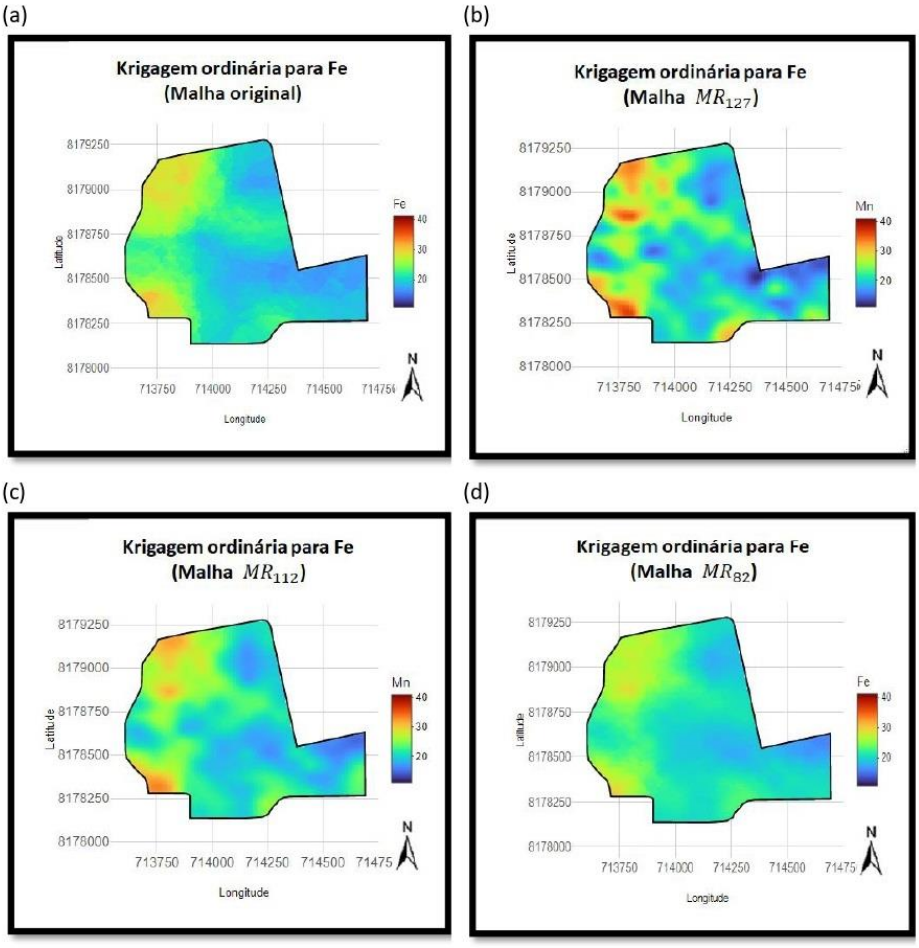
De acordo com a Figura 7 (b), é observado que nas regiões noroeste e sudoeste, estão as maiores concentrações de Fe, destacado nas cores vermelha e amarela. Para o atributo Mn (Figura 7 (d)), é observado que as maiores concentrações de Mn estão nas zonas norte e sudeste, como pode ser observado nas cores vermelha da legenda deste mapa. Além disso, também pode ser observado que nesta área, há pequenas concentrações de Mn em sua totalidade. Já para o Zn, observa-se que na região oeste estão as maiores concentrações e nas regiões nordeste, leste e na parte central, estão os de concentrações medianas.

4.4 Krigagem ordinária para as malhas reduzidas dos micronutrientes

Para a krigagem ordinária foram escolhidas as malhas MR_{127}, MR_{112} e MR_{82} que representam reduções de 15%, 25% e 45%, respectivamente, nos pontos amostrados. Optou-se por valores de redução em até 45%, pois quando a redução amostral era maior, alguns atributos não mostraram dependência espacial, indicado pela ocorrência de efeito pepita puro no ajuste do semivariograma (ISAAKS; SRIVASTAVA, 1989).

Os gráficos interpolados por krigagem ordinária para o atributo Fe com a malha original e com as malhas MR_{127}, MR_{112} e MR_{82}, são apresentadas na Figura 8.

Figura 8 - (a) Mapa interpolado do atributo Fe para a malha original. (b) Mapa interpolado do Fe para a malha MR_{127}, (c) Mapa interpolado do Fe para a malha MR_{112}, (d) Mapa interpolado do Fe para a malha MR_{82}.



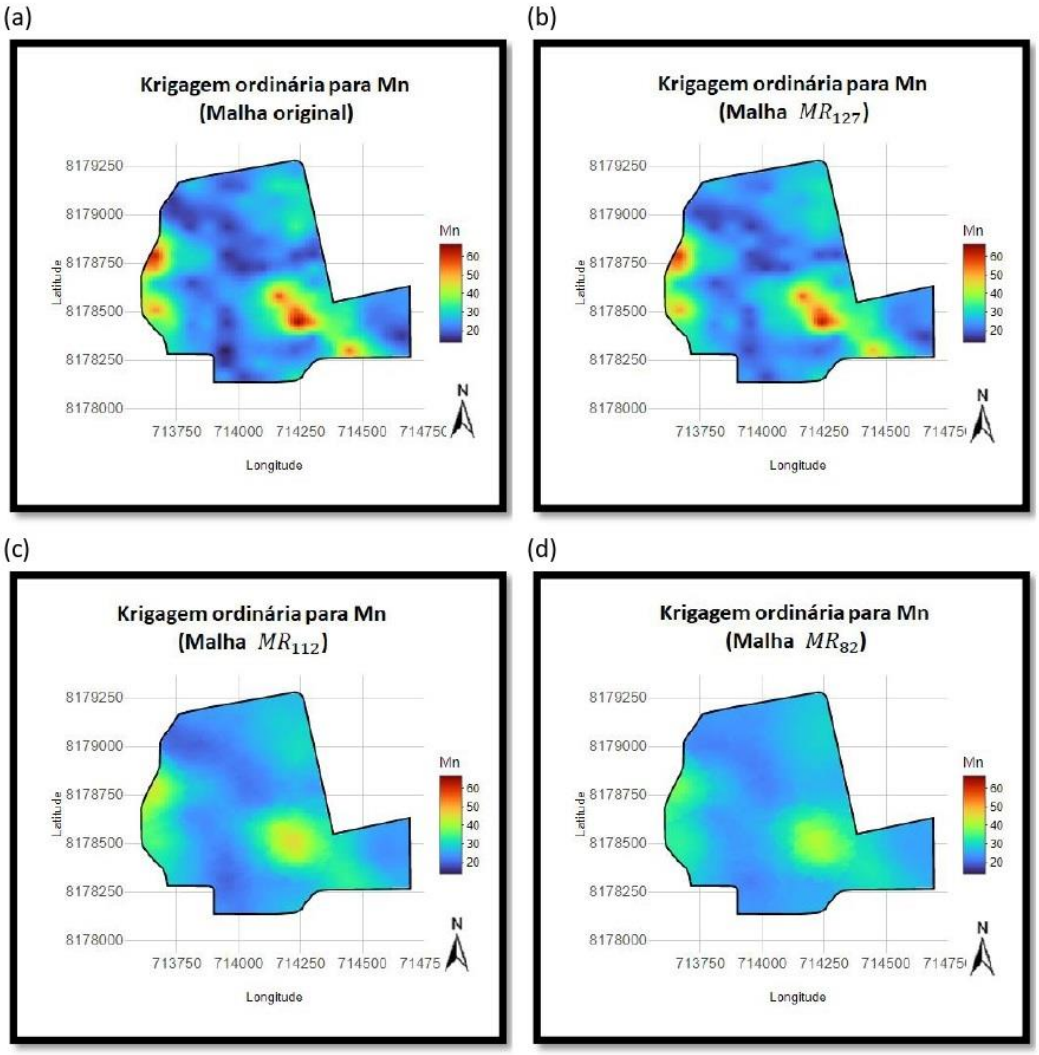
Fonte: O autor (2023)

Comparando o mapa original (Figura 8 (a)) com o mapa da malha MR_{127} (Figura 8 (b)), é possível observar que regiões com maiores concentrações de Fe no mapa original tendem a aumentar suas concentrações e regiões com menores concentrações diminuem.

Para a redução de 25% dos dados (MR_{112}), as regiões com altas concentrações voltam a decair com relação a malha (MR_{127}) e aumentar para zonas com menores atributos. Para a redução de 45% (MR_{82}), os valores dos atributos com maiores concentrações decaem e os valores com menor concentração aumentam em comparativo com a malha MR_{112}.

Os gráficos interpolados por krigagem ordinária para o atributo Mn com a malha original e com as malhas MR_{127}, MR_{112} e MR_{82}, são apresentadas na Figura 9.

Figura 9 - (a) Mapa interpolado do atributo Mn para a malha original. (b) Mapa interpolado do Mn para a malha *MR127* (c) Mapa interpolado do Mn para a malha *MR112*. (d) Mapa interpolado do Mn para a malha *MR82*.



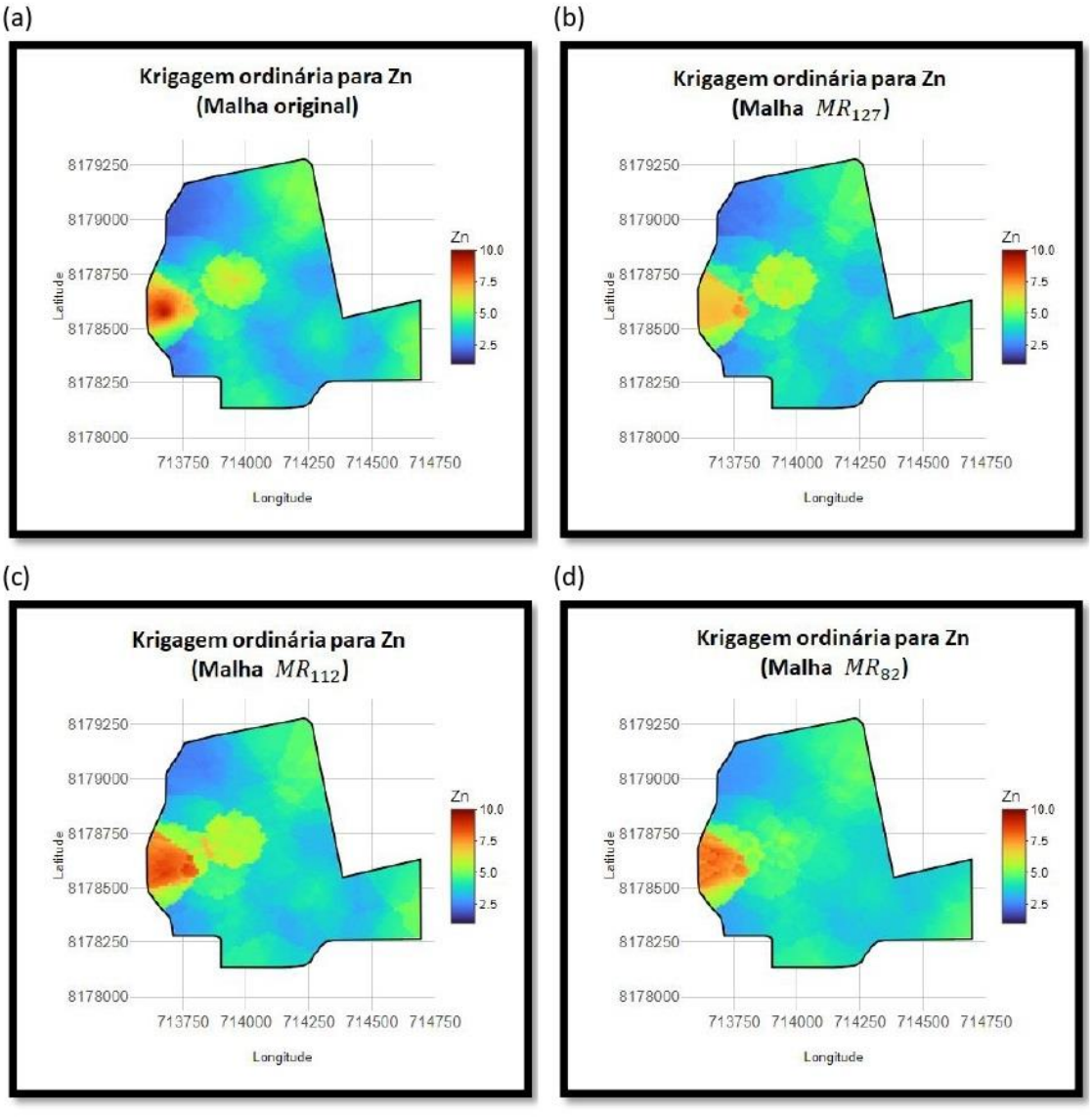
Fonte: O autor (2023)

Comparando o mapa original (Figura 9 (a)), com o mapa sob redução de 15% (Figura 9 (b)), percebe-se uma pequena variação quase imperceptível nos teores de Mn.

Comparando o mapa original do atributo Mn com a malha MR_{112}, percebe-se que há uma grande diferença. Há um aumento da concentração do Mn, nas regiões as quais os teores são menores e uma diminuição da concentração nas regiões com maiores teores de Mn. Este efeito também pode ser observado na malha MR_{82}.

Os gráficos interpolados por krigagem ordinária para o atributo Zn com a malha original e com as malhas MR_{127}, MR_{112} e MR_{82}, são apresentadas na Figura 10.

Figura 10 - (a) Mapa interpolado do atributo Zn para a malha original. (b) Mapa interpolado do Zn para a malha *MR127* (14.2) Mapa interpolado do Zn para a malha *MR112*. (c) Mapa interpolado do Zn para a malha *MR82*



Fonte: O autor (2023)

Comparando o mapa da krigagem original do Zn (Figura 10 (a)) e o mapa com redução de 15%, malha MR_{127}, observa-se que para a região oeste, houve uma redução da concentração dos atributos de maior magnitude. Já nas outras regiões, houve pouca variação.

Para a redução de 25%, malha MR_{112}, Figura 10 (c), pode-se observar que a mancha com maiores concentrações de Zn, localizada na parte central e na zona oeste sofreu uma pequena alteração em comparativo com a malha original. Mesmo com esta alteração, pode-se observar que a área da concentração deste atributo seguiu o mesmo padrão.

Para a redução de 45%, malha MR_{82}, percebe-se que a concentração de Zn na parte central, mudou bruscamente em comparativo com a malha original. Alguns pontos, ainda permaneceram com concentrações altas, mas a área da concentração deste atributo não permaneceu a mesma.

Recentes estudos como os de Pereira *et al.* (2022) e Martins *et al.* (2023), têm explorado o emprego de técnicas de *machine learning* como um método de interpolação utilizando todo o conjunto de dados e comprado com krigagem ordinária. Segundo eles, a krigagem ordinária ainda demonstrou superioridade em comparativo as técnicas de *machine learning.* Este trabalho segue uma vertente diferente, que é utilizar de forma mista as duas metodologias, *machine learning* e krigagem ordinária, para aproveitar um pouco de cada método e não usar as técnicas de *machine learning* como substituto da krigagem ordinária.

Estudos como os de Guo *et al.* (2018), Yang *et al.* (2020), Qu *et al.* (2024) avaliaram o desempenho de algumas metodologias estatísticas para mapeamento do solo, inclusive *machine learning,* sob diferentes condições amostrais, e concluíram que ainda são necessários uma grande quantidade de amostras para se obter predições mais precisas sobre a distribuição de micronutrientes dos solos.

A Tabela 5 exibe os valores da validação cruzada da krigagem ordinária para o conjunto de dados original, como os valores de RMSE e MAE (RMSE_Krig e MAE_Krig) das malhas MR_{127}, MR_{112} e MR_{82} para os atributos Fe, Mn e Zn.

Ao observar a Tabela 5, nota-se também que ao passo que aumenta a quantidade de amostras preditas com *decision tree*, os valores de RMSE_Krig e MAE_Krig tendem a cair. De acordo com El-Sayed Ewis (2012), quando estamos escolhendo o melhor método estatístico, optamos por aqueles que tem o menor RMSE ou o menor MAE. Com isso, a redução das estatísticas de validação, RMSE_Krig e MAE_Krig, indica estar havendo uma melhoria nas predições. Precisamos ter cautela quanto a interpretação destas estatísticas, pois as *decision tree* geram valores iguais para todas as observações que caem na mesma folha e isto pode estar afetando a interpolação na hora da krigagem ordinária.

Tabela 5 - Tabela com os valores das estatísticas de validação da krigagem ordinária para a malha original e as malhas reduzidas dos atributos Fe, Mn e Zn.

Atributo	Malhas	RMSE_Krig	MAE_Krig
Fe	Original	4,24	3,13
	MR_{127}	6,66	5,26
	MR_{112}	6,57	5,15
	MR_{82}	6,07	4,59
Mn	Original	5,90	4,09
	MR_{127}	10,92	7.91
	MR_{112}	9.50	6,64
	MR_{82}	8,37	5,43
Zn	Original	3,00	1,88
	MR_{127}	3,72	1,99
	MR_{112}	3,38	1,94
	MR_{82}	2,98	1,58

RMSE_Krig: Raiz quadrada do erro quadrático médio das malhas MR_{127}, MR_{112} e MR_{82}; MAE_Krig: Média do Erro Absoluto médio das malhas MR_{127}, MR_{112} e MR_{82}.

5. CONCLUSÃO

Neste trabalho foi mostrado que é necessário ter um pouco de cuidado ao usar esta metodologia de redução amostral, pois é percebido que há uma perda da variabilidade espacial ao passo que aumenta a redução amostral. Mesmo assim, foi percebido que o padrão de concentração de micronutrientes dos solos nas malhas reduzidas segue semelhante ao padrão original, ou seja, regiões com maiores concentrações ainda mantêm níveis elevados, enquanto aquelas com menores concentrações continuam a apresentar índices reduzidos.

Com isso, a *decision tree,* se mostrou eficiente em preservar o padrão de distribuição dos micronutrientes.

REFERENCIAL BIBLIOGRÁFICO

ADAMCHUK, V.; LACROIX, R.; SHINDE, S.; TREMBLAY, N.; HUANG, H. **An uncertainty-based comprehensive decision support system for site-specific crop management**. Advances in Animal Biosciences, v. 8, p. 625-629, 2017.

ALBUQUERQUE, J. A.; REINERT, D. J.; FIORIN, J. E. **Variabilidade de solo e planta em Podzólico Vermelho-Amarelo**. Revista Brasileira de Ciência do Solo, v. 20, n. 1, p. 151-157, 1996.

BLACKMORE, S. **The interpretation of trends from multiple yield maps**. Computers and Electronics in Agriculture, v. 26, n. 1, p. 37-51, 2000.

BLACKMORE, B. S.; GODWIN, R. J.; FOUNTAS, S. **The analysis of spatial and temporal trends in yield map data over six years**. Biosystems Engineering, v. 84, p. 455-466, 2003.

BOLFE, E. L.; GOMES, J. V. B. **Geoestatística subsidia agricultura de precisão**. Agroline, 2005.

BORÉM, A. **Nova Revolução Verde**. In: QUEIROZ, Daniel Marçal; VALENTE, Domingos Sárvio Valente; PINTO, Francisco de Assis; BORÉM, Aluízio (org.). Agricultura Digital. 2. ed. São Paulo: Oficina de Textos, 2021. p. 11-19.

BOYER, D. G.; WRIGHT, R. J.; FELDHAKE, C. M.; BLIGH, D. P. **Relações de variabilidade espacial do solo em um ambiente de solo ácido de declive acentuado**. Soil Science, v. 161, n. 5, p. 278-287, 1996.

BRAGA, L. P. V. Geoestatística e aplicações. Minicurso do 9º Simpósio Brasileiro de Probabilidade e Estatística, IME, Universidade de São Paulo, São Paulo, 1990. 36 p.

BURDETT, W.; WELLEN, C. **Statistical and machine learning methods for crop yield prediction in the context of precision agriculture**. Precision Agriculture, v. 23, p. 1553–1574, 2022. Disponível em: https://doi.org/10.1007/s11119-022-09897-0. Acesso em: 25 de out 2023.

CAMARGO, E. C. G. **Desenvolvimento, implementação de testes de procedimentos geoestatísticos (krigagem) no sistema de processamento de informações georeferenciadas**. Spring. 1997. 123 p. Dissertação (Mestrado) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos.

CAMBARDELLA, C. A.; MOORMAN, T. B.; NOVAK, J. M.; PARKIN, T. B.; KARLEN, D. L.; TURCO, R. F.; KONOPKA, A. E. **Field-scale variability of soil properties in central Iowa soils**. Soil Science Society of America Journal, v. 58, n. 5, p. 1501-1511, 1994.

CASTRO-FRANCO, M.; CÓRDOBA, M. A.; BALZARINI, M. G.; COSTA, J. L. **A pedometric technique to delimitate soil-specific zones at field scale**. Geoderma, v. 322, p. 101-111, 2018.

CHERUBIN, M. R.; SANTI, A. L.; EITELWEIN, M. T.; AMADO, T. J. C.; SIMON, D. H.; DAMIAN, J. M. **Dimensão da malha amostral para caracterização da variabilidade espacial de fósforo e potássio em Latossolo Vermelho**. Pesquisa Agropecuária Brasileira, v. 50, p. 182-177, 2015.

CORÁ, J. E.; FERRARESE, C. P.; OLIVEIRA, F. S.; CORÁ, J. A.; BORGES, J. C.; LIMA, J. R.; HERRMANN, R.; CRUZ, L. M**. Variabilidade espacial de atributos do solo para adoção do sistema de agricultura de precisão na cultura de cana-de-açúcar**. Revista Brasileira de Ciência do Solo, Viçosa, v. 28, n. 6, p. 1013-1021, 2004. Disponível em: https://doi.org/10.1590/S0100-08232004000600010. Acesso em: 01 nov. 2023.

COSTA, J. F.; SOUZA, E. L**. Modelos 2D e 3D estimativas por técnicas de krigagem**. UFRGS. SlydePlayer, maio 2014. Online. 14 slides, color. Disponível em: https://slideplayer.com.br/slide/1228013/. Acesso em: 25 ago. 2022.

DAHIKAR, S. S.; RODE, D. S. V. **Agricultural crop yield prediction using artificial neural network approach**. *International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*, v. 2, n. 1, p. 823–826, 2014. Disponível em: https://www.ijireeice.com/upload/2014/ijireeice-2844.pdf. Acesso em: 23 out. 2023.

D'AMARIO, S. C.; REARICK, D. C.; FASCHING, C.; KEMBEL, S. W.; PORTER-GOF, E.; SPOONER, D. E.; WILLIAMS, C. J.; WILSON, H. F.; XENOUPOULOS, M. A. **The prevalence of nonlinearity and detection of ecological breakpoints across a land use gradient in streams**. Scientific Reports, v. 9, n. 1, p. 1–11, 2019. doi: 10.1038/s41598-019-51472-1.

DEMATTÊ, J.; RIZZO, R.; BOTTEON, V. W. **Pedological mapping through integration of digital terrain models, spectral sensing, and photopedology**. Revista Ciência Agronômica, v. 46, p. 669–678, 2015.

EL-SAYED EWIS, O. **Improving the prediction accuracy of soil mapping through geostatistics**. International Journal of Geosciences, v. 2012, 2012.

FATORGIS. **Agricultura de precisão: A tecnologia de GIS/GPS chega às fazendas**. Curitiba, 1998. Disponível em: http://www.fatorgis.com. Acesso em: 21 out. 2023.

GONZALEZ-SANCHEZ, A.; FRUSTO-SOLIS, J.; OJEDA-BUSTAMANTE, W. **Predictive ability of machine learning methods for massive crop yield prediction**. Spanish Journal of Agricultural Research, v. 12, n. 2, p. 313–328, 2014.

GREGO, C.R.; OLIVEIRA, R.P.; VIEIRA, S.R. **Geoestatística aplicada a agricultura de precisão**. In: BERNADI, A.C.C.; NAIME, J.M.; RESENDE, A.V.; BASSOI, L.H.; INAMASU, R.Y. Agricultura de precisão: resultados de um novo olhar. Brasília: EMBRAPA, 2014. p. 74-83. Disponível em: https://www.embrapa.br/busca-de-publicacoes/-/publicacao/1002959/agricultura-de-precisao-resultados-de-um-novo-olhar. Acesso em: 22 fev. 2023.

GREGO, C.R.; VIEIRA, S.R. **Variabilidade espacial de propriedades físicas do solo em uma parcela experimental**. Revista Brasileira de Ciência do Solo, Viçosa, v. 29, n. 2, p. 169-

177, 2005. Disponível em: https://www.scielo.br/j/rbcs/article/view/32799. Acesso em: 01 nov. 2023. doi: 10.1590/S0100-08232005000200002.

GUERRA, P.A.G. **Geoestatística operacional**. Brasília: Ministério das Minas e Energia, 1988.

IACO, S.; HRISTOPULOS, D.T.; LIN, G. **Special Issue: Geostatistics and Machine Learning. Mathematical Geosciences**, v. 54, p. 459–465, 2022. Disponível em: https://doi.org/10.1007/s11004-022-09998-6. Acesso em: 24 fev.2023.

ISAAKS, E. H.; SRIVASTAVA, R. M. **Applied geostatistics**. New York: Oxford University Press, 1989.

IZBICKI, R; SANTOS, T. M. **Aprendizado de máquina: uma abordagem estatística**. 1ª edição. 2020. 272 páginas. ISBN: 978-65-00-02410-4.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R.; TAYLOR, J. **An introduction to statistical learning: With applications in R**. Springer Nature, 2013.

KLERKX, L.; ROSE, D. **Dealing with the game-changing technologies of Agriculture 4.0: How do we manage diversity and responsibility in food system transition pathways?** Global Food Security, v. 24, 2020.

MCKAY, M. D.; BECKMAN, R. J.; CONOVER, W. J. **Comparison of three methods for selecting values of input variables in the analysis of output from a computer code**. *Technometrics*, v. 42, n. 1, p. 239–245, 1979.

MANZATTO, C. V.; BHERING, S. B.; SIMÕES, M. **Agricultura de precisão: propostas e ações da Embrapa Solos**. EMBRAPA Solos, 1999. Disponível em: http://www.cnps.embrapa.br/search/pesqs/proj01/proj01.html. Acesso em: 02 fev. 2023.

MARTINS, R. N.; CARVALHO PINTO, F. D. A.; QUEIROZ, D. M.; VALENTE, D. S. M.; ROSAS, J. T. F.; PORTES, M. F.; CERQUEIRA, E. S. A. **Digital mapping of coffee ripeness using UAV-based multispectral imagery**. Computers and Electronics in Agriculture, v. 204, p. 107499, 2023. https://doi.org/10.1016/j.compag.2022.107499

MATHERON, G. **Principles of geostatistics. Economic Geology**, Lancaster, v. 58, p. 1246-1266, 1963.

MENDES, W. S.; DEMATTÊ, J. A. M.; BARROS, A. S.; SALAZAR, D. F. U.; AMORIM, M. T. A. **Geostatistics or machine learning for mapping soil attributes and agricultural practices.** Revista Ceres, v. 67, n. 4, p. 330–336, 2020. Disponível em: https://doi.org/10.1590/0034-737X202067040010. Acesso em: 29 jul. 2023.

MIAO, Y.; MULLA, D. J.; ROBERT, P. C. **Identificação de fatores importantes que influenciam a variabilidade da produtividade e da qualidade do milho usando redes neurais artificiais**. Agricultura de Precisão, v. 7, n. 2, p. 117–135, 2006.

MILANI, L.; SOUZA, E. G. de; URIBE-OPAZO, M. A.; GABRIEL FILHO, A.; JOHANN, J. A.; PEREIRA, J. O. **Unidades de manejo a partir de dados de produtividade**. Acta Scientiarum. Agronomy, v. 28, p. 591–598, 2006.

MIRANDA, J. I. **Fundamentos de sistemas de informações geográficas**. Brasília: EMBRAPA Informação Tecnológica, 2005. 425 p.

MOLIN, J. P.; RABELLO, L. M. **Estudos sobre a mensuração da condutividade elétrica do solo**. Engenharia Agrícola, v. 31, p. 90-101, 2011.

NANNI, M. R.; POVH, F. P.; DEMATTÊ, J.; OLIVEIRA, R. B.; CHICATI, M. L.; CEZAR, E. **Optimum size in grid soil sampling for variable rate application in site-specific management**. Scientia Agricola, v. 82, p. 386-392, 2011.

OLIVER, M.A. **Exploring soil spatial variation geostatistically**. In: EUROPEAN CONFERENCE ON PRECISION AGRICULTURE, 2, 1999. Odense. Proceedings.. Silsoe: Sheffield, 1999. p. 03-18.

PANTAZI, X. E.; MOSHOU, D.; ALEXANDRIDIS, T.; WHETTON, R. L.; MOUAZEN, A. M. **Wheat yield prediction using machine learning and advanced sensing techniques**. Computers and Electronics in Agriculture, v. 121, p. 57–65, 2016.

PAZ-GONZÁLEZ, A.; SÁNCHEZ, J. S.; FRUTOS, M. A.; MORENO, J. A.; ALBA, L. **The effect of cultivation on the spatial variability of selected properties of an umbric horizon**. Geoderma, v. 97, p. 272–292, 2000.

PEREIRA, G. W.; VALENTE, D. S. M.; DE QUEIROZ, D. M.; SANTOS, N. T.; FERNANDES-FILHO, E. I. **Soil mapping for precision agriculture using support vector machines combined with inverse distance weighting**. Precision Agriculture, v. 23, p. 1189-1204, 2022. Disponível em: https://doi.org/10.1007/s11119-022-09880-9. Acesso em: 01 nov. 2023.

QU, L.; LU, H.; TIAN, Z.; SCHOORL, J. M.; HUANG, B.; LIANG, Y. **Spatial prediction of soil sand content at various sampling density based on geostatistical and machine learning algorithms in plain areas**. Catena, v. 234, p. 107572, 2024.

RAMEZAN, C. A.; WARNER, T. A.; MAXWELL, A. E.; PRICE, B. S. **Effects of training set size on supervised machine-learning land-cover classification of large-area high-resolution remotely sensed data**. Remote Sensing, v. 13, n. 3, p. 368, 2021.

RIBEIRO JUNIOR, P. J. **Métodos geoestatísticos no estudo da variabilidade espacial dos parâmetros do solo.** 1995. 99 p. Dissertação (Mestrado) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 1995.

RODRIGO-COMINO, J.; MARTÍNEZ-HERNÁNDEZ, C.; ISERLOH, T.; CERDÀ, A. **Contrasted impact of land abandonment on soil erosion in Mediterranean agriculture fields.** Pedosphere, v. 28, n. 4, p. 617-631, 2018.

ROSE, D. C.; CHILVERS, J. **Agriculture 4.0: broadening responsible innovation in an era of smart farming**. Frontiers in Sustainable Food Systems, v. 2, n. 87, 2018.

RUSSELL, S.; NORVIG, P. **Artificial intelligence: a modern approach**. 2. ed. Upper Saddle River: Prentice Hall, 2002.

SANA, R. S.; HOLZSCHUH, M. J.; ANGHINONI, I.; BRANDÃO, Z. N. **Spatial variability of physical-chemical attributes of soil and its effects on cotton yield**. Revista Brasileira de Engenharia Agrícola e Ambiental, v. 18, n. 10, p. 994–1002, 2014. Disponível em: https://doi.org/10.1590/1807-1929/agriambi.v18n10p994-1002. Acesso em: 29 jul. 2023.

SILVA, A. P. da. **Variabilidade Especial de Atributos Físicos do Solo**. Piracicaba, 1988. Tese (Doutorado em Solos) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, 1988.

SOUZA, C. K.; MARQUES JÚNIOR, J.; MARTINS FILHO, M. V.; PEREIRA, G. T. **Influência do relevo na variação anisotrópica dos atributos químicos e granulométricos de um Latossolo em Jaboticabal**, SP. Engenharia Agrícola, v. 23, n. 3, p. 486-495, 2003.

TANTALAKI, N.; SOURAVLAS, S.; ROUMELIOTIS, M. **Data-driven decision making in precision agriculture: The rise of big data in agricultural systems.** Journal of Agricultural & Food Information, v. 20, n. 4, p. 344-380, 2019.

TSCHIEDEL, M.; FERREIRA, M. F. **Introdução à Agricultura de Precisão: Conceitos e Vantagens.** Ciência Rural, Santa Maria, v. 32, n. 1, p. 159-163, 2002.

TSEGAYE, T.; HILL, R. L. **Intensive tillage effects on spatial variability of soil test, plant growth, and nutrient uptake measurements**. Soil Science, v. 163, p. 155-165, 1998.

VANNI, S. M. **Modelos de regressão: estatística aplicada**. São Paulo: Legmar Informática, 1998. 177 p.

VIEIRA, S. R**. Geoestatística em estudos de variabilidade espacial do solo**. In: NOVAIS, R. F. de; ALVAREZ V., V. H.; SCHAEFER, C. E. G. R. (Ed.). Tópicos em ciência do solo. Viçosa, MG: Sociedade Brasileira de Ciência do Solo, 2000. v. 1, p. 1-54.

VIEIRA, S. R.; HATFIELD, T. L.; NIELSEN, D. R.; BIGGAR, J. W. **Geostatistical theory and application to variability of some agronomical properties.** Hilgardia, Berkeley, v. 51, n. 3, p. 1-75, 1983.

VIEIRA, S. R.; NIELSEN, D. R.; BIGGAR, J. W. **Spatial variability of field-measured infiltration rate.** Soil Science Society of America Journal**,** Madison, v. 45, p. 1040-1048, 1981.

VIEIRA, S. R.; XAVIER, M. A.; GREGO, C. R. **Aplicações de geoestatística em pesquisas com cana-de-açúcar**. In: DINARDO-MIRANDA, L. L.; VASCONCELOS, A. C. M.; LANDELL, M. G. A. (Ed.). Cana de açúcar. Ribeirão Preto: Instituto Agronômico, 2008. p. 839-852.

WEBSTER, R.; CUANALO, H. E. de la C. **Soil transects correlograms of north Oxfordshire and their interpretation**. The Journal of Soil Science, Oxford, v. 26, p. 176-194, 1975.

WEBSTER, R. **Automatic soil boundary location for transect data**. Mathematical Geology, New York, v. 5, p. 27-37, 1973.

WEBSTER, R.; OLIVER, M. A. **Geostatistics for environmental scientists**. 2. ed. Chichester: John Wiley & Sons, 2007.

YAMAMOTO, J. K.; LANDIM, P. M. B. **Geoestatística: conceitos e aplicações**. São Paulo: Oficina de Textos, 2013. Disponível em: https://www.ofitexto.com.br. Acesso em: 24 fev. 2023.

YANG, L.; LI, X.; SHI, J.; SHEN, F.; QI, F.; GAO, B.; ZHOU, C. **Evaluation of conditioned Latin hypercube sampling for soil mapping based on a machine learning method**. Geoderma**,** v. 369, p. 114337, 2020.

ZHANG, N.; WANG, M.; WANG, N. **Precision agriculture—a worldwide overview**. Computers and Electronics in Agriculture, v. 36, n. 2-3, p. 113-132, 2002.