SAMANTHA GOUVÊA OLIVEIRA

Otimização do mapeamento de micronutrientes do solo com base em macronutrientes e técnicas de aprendizado estatístico

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

Coorientadores: Luiz Alexandre Peternelli e Marcelo Marques Costa

VIÇOSA - MINAS GERAIS

2024

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

T
O48o
2024

Oliveira, Samantha Gouvea, 1997-
Otimização do mapeamento de micronutrientes do solo com base em macronutrientes e técnicas de aprendizado estatístico / Samantha Gouvea Oliveira. – Viçosa, MG, 2024.
1 dissertação eletrônica (55 f.): il. (algumas color.).

Orientador: Nerilson Terra Santos.
Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Estatística, 2024.
Referências bibliográficas: f. 50-55.
DOI: https://doi.org/10.47328/ufvbbt.2024.370
Modo de acesso: World Wide Web.

1. Correlação (Estatística). 2. Aprendizado do computador. 3. Plantas - Nutrição - Estatística. 4. Micronutrientes - Amostragem. 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. 519.537

Bibliotecário(a) responsável: Euzébio Luiz Pinto CRB-6/3317

SAMANTHA GOUVÊA OLIVEIRA

Otimização do mapeamento de micronutrientes do solo com base em macronutrientes e técnicas de aprendizado estatístico

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: 22 de fevereiro de 2024.

Assentimento:

gov.br Documento assinado digitalmente
SAMANTHA GOUVEA OLIVEIRA
Data: 17/07/2024 11:38:55-0300
Verifique em https://validar.iti.gov.br

Samantha Gouvêa Oliveira

Autor

gov.br Documento assinado digitalmente
NERILSON TERRA SANTOS
Data: 17/07/2024 16:06:55-0300
Verifique em https://validar.iti.gov.br

Nerilson Terra Santos

Orientador

Dedico este trabalho aos meus pais, que sob muito sol, fizeram-me chegar até aqui, na sombra.

AGRADECIMENTOS

Depois de oito anos de uma longa jornada acadêmica pela mais linda do Brasil, muitos cafés, situações de procedência duvidosa (mas quase sempre divertidas), 6 casas e 2 departamentos, venho aqui deixar meus agradecimentos e tentar não me esquecer de ninguém (mesmo sabendo que será impossível).

Gostaria de começar agradecendo os meus pais, **Maria Aparecida** e **Edilmício** (o nome é esse mesmo, não se preocupe), por todo carinho, força, amor e principalmente pelo apoio financeiro durante toda a vida, o qual nunca serei capaz de pagar. Amo vocês demais e as ligações diárias não matam a saudade.

Ainda agradecendo aos pais, gostaria de deixar um enorme abraço à mãe que Viçosa me deu, **Lígia Pereira**. Te conhecer e ser adotada como filha do seu coração enorme foi um presente e eu serei eternamente grata!

Não poderia me esquecer do meu irmão, **Max**. Você vibrou por todas as minhas conquistas como se fossem suas, me deu apoio e mobiliou a minha primeira casa (graças a sua fase rebelde de cismar em morar sozinho, só para depois descobrir que morar com nossos pais era mais barato).

Não posso deixar de mencionar e agradecer ao **Guilherme Henrique**, por todo amor, cafés, conversas mirabolantes, joguinhos tarde da noite, por topar todas as minhas ideias erradas (mesmo quando ele sabia que não daria certo) e por escolher continuar ao meu lado, mesmo nos momentos difíceis.

Agradeço também ao meu irmão, **Hugo**, pela parceria em quase todas as presepadas nas quais me enfiei nesses anos de UFV, por ser meu companheiro de todos os planos de viagem (que um dia iremos cumprir), por todos os joguinhos, conversas até tarde da noite e por ser a minha fonte inesgotável de batata frita em todos os lanches compartilhados.

Não tem como não pensar nos amigos que ganhei e não me lembrar da **Jamy** (a melhor engenheira de produção que Jundiaí já viu), que me acompanhou em toda a graduação (mesmo trocando de curso) e que se fez presente em todos os momentos (quando eu chorei pela primeira nota vermelha e quando vibrei pela conquista do diploma). Aos amigos do mestrado, agradeço a **Luciano**, **Alice**, **Rodrigo**, **Gingim**, **Thaynara**, **Jhennyfer**, **Marco Luís**, **Matheus de Paula**, **Vinícius**, **Mayara** e **Pablo**. Quando ingressei, não imaginava que encontraria pessoas tão incríveis e que tivessem o poder de aliviar a carga do trabalho de todo dia, que bom que estava errada.

Aos professores que até aqui me orientaram, **Nerilson** e **Peternelli**, deixo o meu muito obrigada pelo apoio, disponibilidade, incentivo, parceria, conselhos e ensinamentos.

Também deixo meu muito obrigado ao **Laboratório de Pesquisas em Estatística Aplicada - LAPEA**, por ter me acolhido durante a graduação, me presenteado com amigos incríveis, ensinamentos acadêmicos e de vida. Agradeço ao **Grupo de Estudos em Estatística Aplicada e Biometria - GESTBIO** que me apresentou uma boa parte dos amigos do mestrado e me deu a oportunidade de organizar eventos para a pós-graduação.

Sou grata aos demais professores e funcionários da Universidade Federal de Viçosa pela formação acadêmica durante minha graduação e mestrado. Em especial deixo o meu muito obrigada para o **Edson** e a **Cleuza**, funcionários da Livraria-UFV e fornecedores da minha dose de café diária.

Também deixo o meu muito obrigada a família e aos amigos que deixei em Ubá quando me mudei (Viçosa tem sido uma aventura e tanto, mas é sempre muito bom voltar para casa).

Por fim, agradeço às agências de fomento Fundação de Amparo à Pesquisa do Estado de Minas Gerais (**FAPEMIG**), Conselho Nacional de Desenvolvimento Científico e Tecnológico (**CNPq**) e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (**CAPES**) - Código de Financiamento 001 pelo financiamento direto ou indireto de todos os trabalhos desenvolvidos nesta dissertação.

Quando comecei essa aventura em 2022, fiz diversos planos e imaginei situações que nunca aconteceram. Não amei cada segundo do processo, mas sou grata ao conglomerado de decisões e acasos que me trouxeram até aqui.

E agora?

(Procurando Nemo, 2003).

RESUMO

OLIVEIRA, Samantha G., M.Sc., Universidade Federal de Viçosa, fevereiro de 2024. **Otimização do mapeamento de micronutrientes do solo com base em macronutrientes e técnicas de aprendizado estatístico**. Orientador: Nerilson Terra Santos. Coorientadores: Luiz Alexandre Peternelli e Marcelo Marques Costa.

O constante crescimento da população mundial acarreta diretamente no setor agronômico, resultando em um aumento na demanda por produção de alimentos, além de gerar preocupações relacionadas a limitações de áreas de cultivo e escassez de mão de obra. Surgem então a agricultura de precisão e a agricultura digital, que são responsáveis por processar um grande volume de informações com o objetivo de promover retorno econômico, vantagem competitiva para o produtor e minimizar os efeitos ao meio ambiente. Nota-se, portanto, a necessidade intrínseca de lidar de forma mais eficiente com os recursos e a variabilidade dos atributos do solo. Um dos ferramentais utilizados para a descrição da variabilidade espacial e mapeamento de atributos é conhecido como geoestatística. Contudo, um dos grandes desafios do método está relacionado com um número mínimo de amostras para realizar as interpolações, o que pode aumentar consideravelmente os gastos e necessidade de mão de obra para um projeto, pois a amostragem envolve a coleta e análise de atributos de todos os pontos previamente estipulados. Com o intuito de contornar a problemática relacionada a amostragem de dados em campo, este trabalho tem como objetivo reduzir o número de amostras analisadas quimicamente para micronutrientes ao predizer suas concentrações com base nos macronutrientes, utilizando uma combinação de krigagem e métodos de *machine learning* (KNN). A área experimental é referente a uma parcela da fazenda "Sozinha" localizada em Goianápolis. As 150 amostras foram recolhidas nas profundidades de 0 a $0{,}2\ m$, sendo cada uma composta por dez subamostras coletadas a uma distância de até $5\ m$ do ponto. Posteriormente foram realizadas análises físicas e químicas para quantificar os atributos presentes. Em seguida foram selecionadas grades modificadas (através dos métodos de amostragem aleatória simples (AAS) e *Conditioned Latin Hypercube Sampling* ($cLHS$)) com redução de 15, 25 e 35% dos pontos originais, os quais resultaram em conjuntos de treinamento para o KNN. Posteriormente, o algoritmo KNN foi utilizado para predizer esses 23, 38 e 53 pontos amostrados e esses valores preditos foram então substituídos no conjunto de dados original. A seguir os mapas interpolados por malha e tipo de amostragem de cada um dos métodos empregados (krigagem ordinária (OK) e da diferença entre a OK e a krigagem ordinária combinada com KNN) foram obtidos. Todo o processo, desde a amostragem até as interpolações por krigagem, foi repetido por 50 vezes. Para comparar as interpolações da krigagem ordinária no banco de dados original e nas grades modificadas foi analisada a razão entre a média da raiz quadrada do erro quadrático médio ($RMSE$) e do erro absoluto médio (MAE) de ambas amostragens e o $RMSE$ e MAE da krigagem dos dados originais. A amostragem $cLHS$ se mostrou melhor em manter as características espaciais do solo (com perda da variabilidade espacial) para os atributos estudados frente a todas as reduções de dimensionalidade quando comparada a AAS. Sugere-se para trabalhos futuros, que sejam estudadas novas metodologias de *machine learning* combinadas à krigagem ordinária, além de tipos de amostragem diferentes como forma a avaliar seu comportamento frente a redução do adensamento amostral.

Palavras-chave: Redução do adensamento amostral; Krigagem; KNN, Random Forest.

ABSTRACT

OLIVEIRA, Samantha G., M.Sc., Universidade Federal de Viçosa, February, 2024. **Optimizing Soil Micronutrient Mapping Based on Macronutrients and Statistical Learning Techniques**. Advisor: Nerilson Terra Santos. Co-advisors: Luiz Alexandre Peternelli and Marcelo Marques Costa.

A continuous growth in the world population directly impacts the agronomic sector, resulting in an increased demand for food production and raising concerns related to limitations in cultivation areas and a shortage of labor. Precision agriculture and digital farming emerge as solutions responsible for processing a large volume of information aimed at promoting economic returns, providing a competitive advantage for producers, and minimizing environmental effects. Therefore, an intrinsic need arises to handle resources and soil attribute variability more efficiently. One of the tools used for describing spatial variability and mapping attributes is known as geostatistics. However, a significant challenge in this method is associated with a minimum number of samples required for interpolations, which can considerably increase expenses and the need for labor in a project. This is because the sampling involves collecting and analyzing attributes from all predetermined points. To address the issues related to field data sampling, this study aims to reduce the number of chemically analyzed samples for micronutrients by predicting their concentrations based on macronutrients. This is achieved using a combination of kriging and machine learning methods (KNN). The experimental area pertains to a section of the "Sozinha" farm located in Goianápolis. One hundred and fifty samples were collected at depths of 0 to 0.2 meters, with each composed of ten subsamples collected within a distance of up to 5 meters from the point. Subsequently, physical and chemical analyses were conducted to quantify the present attributes. Modified grids were then selected (using the methods of random sampling (AAS) and Conditioned Latin Hypercube Sampling ($cLHS$)) with a reduction of 15, 25, and 35% of the original points, resulting in training sets for KNN. The KNN algorithm was used to predict these 23, 38, and 53 sampled points, and these predicted values were then replaced in the original dataset. Maps interpolated by mesh and sampling type for each of the employed methods (ordinary kriging (OK) and the difference between OK and ordinary kriging combined with KNN) were obtained. The entire process, from sampling to kriging interpolations, was repeated 50 times. To compare the interpolations of ordinary kriging in the original database and modified grids, the ratio between the mean square root of the mean error ($RMSE$) and the mean absolute error (MAE) of both samplings and the $RMSE$ and MAE of kriging of the original data was analyzed. The $cLHS$ sampling proved to be more effective in preserving the spatial characteristics of the soil (with loss of spatial variability) for the studied attributes compared to all dimensionality reductions when compared to AAS. It is suggested for future work to explore new machine learning methodologies combined with ordinary kriging, as well as different sampling techniques, to assess their behavior in the face of sample density reduction.

Keywords: Sample density reduction; Kriging; KNN; Random Forest.

SUMÁRIO

1. Introdução

O crescimento da população mundial, estimado para 9,7 bilhões de indivíduos em 2050 (United Nations, 2022), acarreta um aumento na demanda do sistema agropecuário relacionado à produção de alimentos, o que intensifica preocupações do setor ligadas a escassez de mão de obra, mudanças climáticas, sustentabilidade e limitações das áreas de cultivo (Mcfadden; Njuki; Griffin, 2023).

Para atender essas necessidades, é fundamental obter informações sobre o desperdício de recursos, a variabilidade de nutrientes do solo a ser utilizado, predizer a quantidade de fertilizantes para aplicação, entre outros (Kwaghtyo; Eke, 2023).

Para tal, surge o conceito de agricultura de precisão (AP) que, de acordo com o Ministério de Agricultura e Pecuária (2016), é definida como "um sistema de gerenciamento agrícola baseado na variação espacial e temporal da unidade produtiva e visa ao aumento de retorno econômico ao agricultor, à sustentabilidade e à minimização do efeito ao ambiente".

Como forma de complementar os conceitos apresentados pela AP surge a agricultura digital (AD), que aprimorou ainda mais a tomada de decisão baseada em dados, pois dessa vez os dados previamente armazenados são levados em consideração, além da logística agrícola e alimentar, as operações da fazenda e também pessoas de interesse. Todas essas informações sobre a fazenda, o solo, o clima, entre outras, podem ser coletadas em tempo real com a intenção de encontrar soluções cada vez mais ágeis e assertivas (Wolfert; Goense; Sorensen, 2014).

Portanto, a demanda por informações precisas e georreferenciadas fornecidas pela agricultura de precisão e digital tem seu crescimento alavancado. Essa evolução do setor é responsável pelo processamento de um grande volume de dados, advindos da inserção da tecnologia à agricultura convencional, como forma de promover vantagem competitiva dos produtos, além de possuir inúmeros benefícios ambientais (Massruhá, 2020).

Neste contexto, a geoestatística, ao utilizar tais informações georreferenciadas, tem um papel importante ao descrever a variabilidade espacial de diversos fenômenos, além de mapear atributos e predizer seu valor desconhecido em pontos não amostrados através da interpolação dos pontos observados.

Contudo, um dos grandes entraves da utilização dos métodos geoestatísticos é a necessidade de um número mínimo de amostras. Segundo Landim (2006), este tipo de análise exige um mínimo de 30 a 40 pontos amostrados, a depender da área

do experimento. Entretanto, esta amostragem é um trabalho árduo e custoso, pois envolve a coleta do solo em todos os pontos previamente estipulados e a posterior análise laboratorial de cada uma destas amostras de solo.

Em geral, o valor cobrado para a mensuração do teor de elementos químicos em amostras de solo é dividido em dois grupos. Para o primeiro grupo que contém os macronutrientes (fósforo, potássio, cálcio, etc.) cobra-se uma taxa fixa. Já para a análise do segundo grupo que compreende alguns micronutrientes (zinco, ferro, manganês, cobre, etc.) pode-se cobrar uma taxa adicional por elemento. A mensuração destes atributos pode, portanto, aumentar consideravelmente o custo da análise química das amostras coletadas.

Portanto, o estudo e utilização de metodologias que permitam reduzir o adensamento amostral, mantendo a qualidade dos resultados obtidos, é essencial. Dentre os possíveis métodos que podem ser aplicados para contornar tal problema pode-se citar os de *machine learning* (ML) que, de acordo com De Iaco et. al. (2022), fornecem estruturas robustas para o processamento de dados espaciais.

Das metodologias de ML aplicadas na literatura pode-se citar o KNN para regressão, que prediz o valor de determinado atributo com base na média de seus K-vizinhos mais próximos e o *Random Forest* (RF), que combina as predições de múltiplas árvores de decisão para obter a predição de um atributo. Já para métodos geoestatísticos tem-se a krigagem ordinária, que considera que o valor predito para o atributo é dado pela combinação linear ponderada dos valores observados nos pontos amostrados (Isaaks; Srivastava, 1989).

2. Objetivos

Este trabalho tem como objetivo geral investigar o efeito da redução do número de amostras analisadas quimicamente para atributos específicos (micronutrientes) através da predição de suas concentrações com base nos atributos padrão (macronutrientes), utilizando, para esta finalidade, uma combinação de métodos de interpolação (MI) com os de ML.

Dentre os objetivos específicos deste trabalho destacam-se avaliar o efeito da redução da densidade amostral:

- nas predições por KNN;
- nos mapas resultantes da modelagem da dependência espacial;
- nas predições das combinações da krigagem com KNN.

3. Revisão Bibliográfica

3.1 Métodos de interpolação

Os métodos de interpolação têm se mostrado um ferramental importante no estudo das ciências agrárias, sendo aplicados para mapeamento da precipitação pluvial (Viola et al., 2010), avaliação de indicadores de fertilidade (Motomiya; Corá; Pereira, 2006), mapeamento de propriedades do solo (Barrena-González; Lavado Contador; Pulido Fernández, 2022), entre outros.

Segundo Hartkamp, Stein e White (1999) os interpoladores podem ser particionados em três categorias:

- Locais ou globais: os interpoladores globais geram um modelo de predição que leva em consideração todos os pontos da área estudada. Já os locais consideram apenas um certo número de vizinhos para a predição de um ponto não amostrado na área experimental.
- Determinísticos ou estocásticos (geoestatísticos): métodos determinísticos realizam as interpolações baseados em critérios de geometria e distribuição espacial das amostras coletadas. Já os métodos estocásticos utilizam a teoria da probabilidade na determinação do peso de cada uma das amostras para interpolação.
- Exatos e inexatos: a interpolação exata retorna valores preditos iguais aos observados nos pontos amostrados enquanto a inexata ou aproximada produz uma função que não passa necessariamente por todos os pontos estudados, buscando um melhor ajuste entre os valores medidos (Jakob; Young, 2006).

A escolha de qual método de interpolação utilizar depende da área da pesquisa, da natureza dos dados, da existência de dependência espacial, da complexidade do modelo a ser utilizado, entre outros.

O estudo das ciências agrárias necessita de métodos que permitam descrever a variabilidade da área experimental. A geoestatística cumpre esse objetivo, pois leva em consideração a correlação espacial entre os valores observados.

3.2 Geoestatística

A estatística clássica é amplamente utilizada para descrever inúmeros eventos, mas por vezes possui certas limitações, como, por exemplo, as conhecidas

pressuposições de independência (CARVALHO; SILVEIRA; VIEIRA, 2002). Em contrapartida, a estatística espacial ou geoestatística, pressupõem a existência de correlação entre os dados coletados. Portanto, estas fazem uso de informações georreferenciadas ou espaciais. Entretanto, esse fato não torna as duas áreas de estudo completamente distintas, visto que a geoestatística foi idealizada com base em fundamentos da estatística clássica, em especial o conceito de funções aleatórias (Ribeiro Junior, 1995).

Os primeiros indícios da aplicação da geoestatística datam de 1951 por Daniel Gerhardus Krige (Krige, 1951) em seu trabalho realizado para estimar o teor de ouro em minas da África do Sul, no qual o mesmo concluiu que a variabilidade dos dados não pode ser explicada apenas pela média das amostras coletadas, já que sua variação também era influenciada pela distância da amostragem dos pontos.

Segundo Matheron (1971), a geoestatística nada mais é que a aplicação da teoria das variáveis regionalizadas. Do seu ponto de vista matemático, uma variável regionalizada é uma função de um ponto x qualquer, $f(x)$, que possui dois enfoques complementares, sendo um aleatório e o outro estruturado. O primeiro seria referente a aleatoriedade ou imprevisibilidade das variáveis em diferentes pontos e o segundo diz respeito às características de continuidade inerentes ao fenômeno.

Tal continuidade pode ser observada em ocorrências naturais visto que se espera que valores de atributos coletados em pontos geograficamente próximos entre si sejam mais semelhantes do que os coletados em pontos mais distantes.

Entre os propósitos da geoestatística está a possibilidade de mapeamento de determinado atributo e, consequentemente, a predição de seu valor desconhecido em pontos não amostrados através da interpolação dos valores observados de uma variável e/ou de outra variável correlacionada.

Dentre os métodos de predição abrangidos pela geoestatística pode-se citar a krigagem (Isaaks; Srivastava, 1989). Porém, nesse caso, se faz necessário analisar a dependência espacial dos atributos estudados utilizando os semivariogramas (Isaaks; Srivastava, 1989).

3.2.1 Semivariograma

O semivariograma experimental é uma ferramenta utilizada para representar graficamente a variabilidade espacial de um determinado fenômeno em função da distância e da direção entre os pontos coletados. Também pode ser definido como a

representação das estimativas de semivariância, $\hat{\gamma}(h)$, para uma dada distância h de separação dentre todas as possíveis combinações de pares de pontos amostrados (Ferreira, 2005; Sousa, 2020).

É importante levar em consideração que os fenômenos podem apresentar ou não anisotropia, que ocorre quando a função de semivariância (Equação 1) é modificada a depender da direção considerada (Isaaks, Srivastava, 1989; Yamamoto, Landim, 2013). Se $\gamma(h)$ independe da direção, pode-se dizer que o fenômeno é isotrópico.

A construção de um semivariograma pode ser dividida em três etapas:

- Primeiro define-se a direção e as distâncias de separação entre os pontos (*lag*) para os quais a semivariância será calculada. A Figura 1 ilustra exemplos de uma grade amostral, *lags* e direções que podem ser consideradas;
- Para cada *lag* escolhida anteriormente calcula-se a semivariância através da Equação 1. Esse processo é repetido para todas as distâncias de separação e, ao final, obtém-se um gráfico que representa os valores de semivariância (eixo y) para todas as distâncias de separação consideradas (eixo x) para a direção previamente definida;

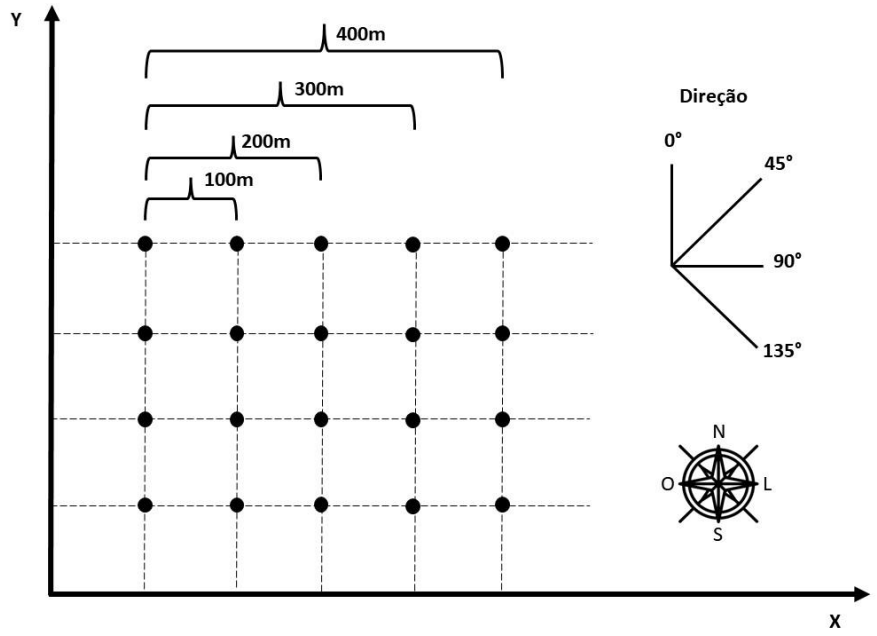
  A semivariância empírica, $\gamma(h)$, para a distância h é obtida por:

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

Em que:

- $N(h)$ é o número de pares de pontos amostrados separados por uma distância h;
- $Z(x_i)$ *e* $Z(x_i + h)$ são os valores observados para o atributo nos pontos localizados nas posições x_i e $x_i + h$, respectivamente.

Figura 1 - Representação gráfica de um exemplo de grade de amostragem e direção.



Fonte: Adaptado de (Spiazzi, 2011).

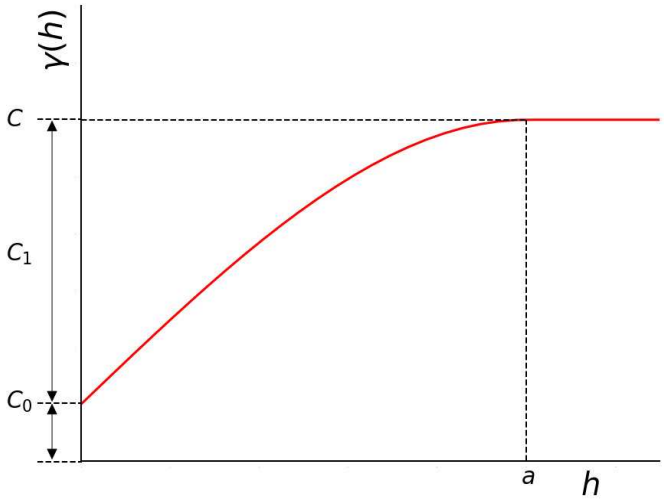
Com o semivariograma experimental em mãos, necessita-se de um modelo teórico que melhor se adeque aos dados provenientes da área experimental para descrever sua estrutura de dependência espacial.

Deve-se levar em consideração que o ajuste dos modelos só deixou de ser completamente realizado de forma visual (sem o apoio de procedimentos matemáticos) em meados da década de 80 devido ao avanço dos recursos computacionais, o que tornou o processo de estimação dos parâmetros do semivariograma menos subjetivo. Entre os métodos que passaram a ser aplicados pode-se citar o método dos mínimos quadrados ordinários ponderados e também o método da máxima verossimilhança (Mello, 2004).

Os modelos teóricos de semivariograma podem ser divididos em duas grandes categorias que se diferenciam pela presença ou não de um patamar. Os semivariogramas com patamar atingem um valor constante de semivariância a partir de uma determinada distância h de separação.

Isaaks e Srivastava (1989) relatam algumas propriedades importantes de um semivariograma teórico com patamar, as quais são apresentadas na Figura 2**.**

Figura 2 - Exemplo de um semivariograma com patamar ilustrando os parâmetros efeito pepita (C_0), contribuição (C_1), patamar (C) e alcance (a).

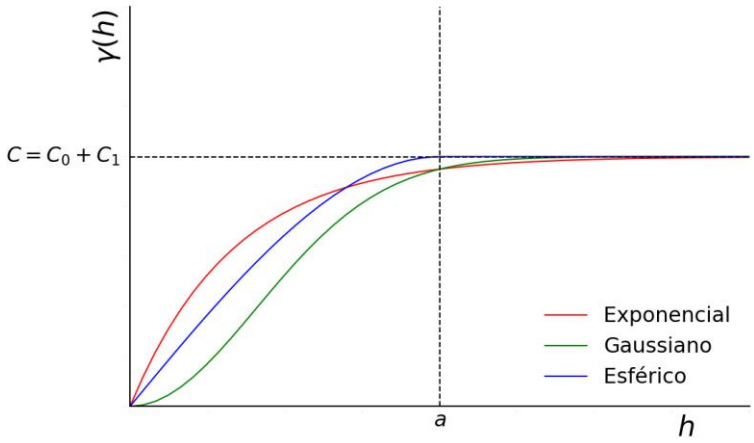


Os parâmetros de um modelo de um semivariograma teórico são:

- Efeito Pepita (C_0): representa a descontinuidade na origem do variograma. Em teoria, para uma distância $h = 0$, o valor da semivariância deveria ser zero, mas existem situações como o erro durante a amostragem e questões relacionadas à variabilidade do fenômeno que fazem com que a descontinuidade ocorra.
- Alcance (a): é a distância na qual a semivariância atinge o seu valor máximo ou se aproxima de um máximo assintótico. Nota-se na Figura 2 que à medida que a distância h entre os pontos aumenta, o valor da semivariância também aumenta. Contudo, a partir do alcance a semivariância se estabiliza e a dependência espacial deixa de existir.
- Patamar (C): é o valor máximo que a semivariância assume. Corresponde ao valor de semivariância quando o alcance é atingido, ou seja, quando a distância de separação entre as amostras é grande o suficiente para torná-las independentes;
- Contribuição (C_1): é a diferença entre o patamar (C) e o Efeito Pepita (C_0).

Dentre os modelos teóricos com patamar, os mais comuns incluem o modelo esférico, modelo exponencial e gaussiano (Rios, 2018), conforme ilustrado na Figura 3.

Figura 3 - Modelos teóricos de semivariograma mais utilizados. Todos com os mesmos parâmetros de alcance, patamar e efeito pepita.



3.2.2 Modelo Esférico

Segundo Isaaks e Srivastava (1989), este é o modelo mais comumente utilizado. Nota-se através da Figura 3 que o mesmo possui um crescimento rápido e linear na origem.

A representação do modelo é dada por:

$$\gamma(h) = \begin{cases} C\left[\frac{3}{2}\frac{h}{a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right], & se\ h \leq a \\ C, & c.c. \end{cases} \quad (2)$$

3.2.3 Modelo Exponencial

Tem um crescimento rápido na origem, linear a curtas distâncias e atinge o patamar de forma assintótica, sendo que o alcance prático é estabelecido como valor da distância tal que o valor da semivariância é 95% do valor do patamar assintótico.

A representação do modelo é dada por:

$$\gamma(h) = C\left[1 - \exp\left(-\frac{3h}{a}\right)\right] \quad (3)$$

3.2.4 Modelo Gaussiano

É utilizado para a modelagem de fenômenos mais contínuos e assim como o modelo exponencial, atinge o patamar de forma assintótica. Dos três modelos apresentados, esse é o único que apresenta um ponto de inflexão.

A representação do modelo é dada por:

$$\gamma(h) = \begin{cases} 0, & se\ \ h = 0 \\ C\left[1 - exp\left(-\frac{3h^2}{a^2}\right)\right], & se\ h \neq 0 \end{cases} \quad (4)$$

Como critério de escolha do melhor modelo ajustado pode-se utilizar a métrica do erro quadrático médio, do inglês *mean square error* (MSE), que leva em consideração a diferença quadrática entre a semivariância empírica e a predita através da curva do modelo teórico estudado.

$$MSE = \frac{\sum_{j=1}^{J}\left(\widehat{\gamma_j}(h) - \gamma_j(h)\right)^2}{J} \quad (5)$$

Em que:

- $\widehat{\gamma_j}(h)$ é o valor da semivariância predita através do modelo teórico;
- $\gamma_j(h)$ é o valor da semivariância empírica;
- $j = 1,2,\cdots,J$ são as diferentes *lags* utilizadas para obter cada uma das semivariâncias empíricas $\gamma(h)$.

Após a construção de um modelo de dependência espacial para os dados estudados pode-se prosseguir para a fase de predições e mapeamento da área experimental utilizando a técnica da krigagem.

3.3 Krigagem

O termo krigagem foi cunhado no meio científico em 1967, (Agterberg, 2004), através de uma homenagem a Daniel G. Krige, um engenheiro de minas da África do Sul, que realizou contribuições essenciais para a teoria das variáveis regionalizadas (Yamamoto; Landim, 2013) a qual foi posteriormente desenvolvida por Matheron (1965).

O método consiste em um procedimento de interpolação do ramo da geoestatística aplicado na predição de valores de características com base em seus pontos vizinhos (Yamamoto; Landim, 2013). Posteriormente, o processo ficou conhecido como sinônimo de “predição ótima” ou “ideal” (Cressie, 1990).

Um dos maiores diferenciais entre a krigagem e outros métodos de interpolação é que a mesma não possui viés e tem a finalidade de minimizar a variância dos erros de predição. A obtenção das predições segue uma função aleatória $Z(u)$ que se

desdobra em uma parte residual, $R\ (u)$, e outra com tendência, $m(u)$, que é referente à média dos valores obtidos para os atributos estudados (Carvalho; Vieira, 2001).

$$Z(u) = R(u) + m(u) \quad (6)$$

A literatura apresenta múltiplas formas de interpolação por krigagem, dentre elas a krigagem universal, ordinária, simples, em bloco e fatorial, que diferem entre si através da forma com que cada uma considera o seu componente de tendência (Carvalho; Vieira, 2001).

A título de ilustração, a técnica da krigagem universal considera seu componente de tendência (média) como um valor local e desconhecido, ou seja, deve ser recalculado caso a vizinhança estudada mude (Omran, 2012). Por outro lado, na krigagem ordinária, a média dos atributos é constante em toda a área de estudo, mesmo que seu valor exato seja desconhecido (Saito et al., 2005; Meul, Van Meirvenne, 2003).

3.3.1 Krigagem ordinária

É considerada como o método mais difundido entre os pesquisadores e leva em consideração que o valor predito para o atributo de interesse é uma combinação linear ponderada dos dados coletados (Isaaks; Srivastava, 1989):

$$\hat{Z}_k = \sum_{i=1}^{n} w_i\, Z_i \quad (7)$$

Em que:

- $i = 1, 2, \cdots,\ n$ representam cada um dos pontos amostrados;
- $\hat{Z}_k$ é o valor predito do atributo em um ponto não amostrado k;
- Z_i é o valor medido do atributo em um ponto amostrado i;
- w_i é o peso atribuído a cada ponto amostrado i.

Também é conhecida como BLUP, do inglês, *Best Linear Unbiased Predictor*, ou seja, o melhor preditor linear não viesado. Possui essa nomenclatura, pois suas predições são calculadas através da combinação linear das amostras obtidas e é não viesada porque a média dos erros de predição é nula. É o melhor porque dentre todos os preditores lineares não viesados é o que minimiza a variância dos erros de predição (Isaaks; Srivastava, 1989).

Para assegurar que o preditor seja não viesado a soma de todos os pesos atribuídos aos valores observados nos pontos amostrados deve ser obrigatoriamente igual a 1, ou seja, $\sum_{i=1}^{n} w_i = 1$ (Matheron, 1963).

Estes pesos são obtidos através da solução de um sistema de equações lineares representado no formato matricial por $\Gamma w = g$.

$$\begin{bmatrix} \gamma_{11} & \gamma_{21} & \cdots & \gamma_{1n} & 1 \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{21} & 1 \\ \vdots & \vdots & \ddots & \vdots & 1 \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{nn} & 1 \\ 1 & 1 & & 1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma_{k1} \\ \gamma_{k2} \\ \vdots \\ \gamma_{kn} \\ 1 \end{bmatrix}$$

Em que:

- γ_{ij} é estimativa da semivariância entre os valores dos pontos amostrados i e j;
- μ é o multiplicador de Lagrange;
- γ_{ki} é a estimativa da semivariância entre os valores do ponto amostrado i e o ponto k para o qual se deseja obter a predição.

Para encontrar a matriz de pesos basta multiplicar a inversa da matriz Γ pelo vetor g, resultando em $w = \Gamma^{-1} g$. A partir dos pesos é possível predizer o valor de um determinado atributo estudado em um ponto não amostrado através da krigagem, (Equação 7). Para análise da qualidade das predições obtidas por krigagem ordinária pode-se empregar as estatísticas de validação cruzada definidas como a raiz quadrada do erro quadrático médio $(RMSE)$ e o erro absoluto médio (MAE).

Nos últimos anos se intensificou o interesse por confrontar ou até mesmo combinar a abordagem tradicional da krigagem com as mais recentes técnicas de predição, como os algoritmos de *machine learning*. Para ilustrar essa tendência crescente de comparação pode-se mencionar: Farooq et al. (2022), que compararam *Random Forest* e krigagem para mapear carbono orgânico do solo, Derdouri e Murayama (2020) que mapearam a precificação de terras no Japão através do método de krigagem de regressão e algoritmos de aprendizado de máquina, além de Erdogan Erten, Yavuz e Deutsch (2022), que combinaram aprendizado de máquina e krigagem para a predição de atributos geológicos.

3.4 *Machine Learning*

O aprendizado de máquina, do inglês *machine learning* (ML), é considerado um ramo da inteligência artificial (AI) responsável por elaborar sistemas automatizados

que tem capacidade para realizar tomadas de decisão com base em experiências acumuladas de processos anteriores (Monard; Baranauskas, 2003). Além disso, também é uma área de estudo conhecida por processar conjuntos de dados numerosos, resultando em economia de tempo e de recursos computacionais (Xu et al., 2021).

Pode-se dividir o ML em três grandes categorias conhecidas como supervisionado, não supervisionado e aprendizagem por reforço. O método supervisionado compreende os modelos que aprendem através de resultados passados de entrada e saída, os chamados dados de treinamento, para poder prever uma saída para uma determinada entrada dos dados de teste (Liu; Wu, 2012). Ou seja, a experiência adquirida através da etapa de treinamento é aplicada na predição de novas classes ou valores numéricos para o teste.

Já o aprendizado não supervisionado não trabalha com a rotulação prévia dos dados, ou seja, é uma abordagem que não possui informações *a priori* que a auxiliem na tomada de decisão. O algoritmo agrupa ou segmenta os dados com base nas similaridades, ou discrepâncias encontradas nos mesmos. A intenção nesse caso é descobrir padrões nos dados analisados (Liakos et al., 2018; Dridi, 2021).

Na aprendizagem por reforço, assim como no aprendizado não supervisionado, a máquina inicia o processo sem nenhuma informação prévia da atividade ou sobre qual decisão tomar. O processo funciona com base em tentativa e erro de forma que a interação entre algoritmo e ambiente seja positiva quando a ação certa é escolhida (“recompensa”) e negativa caso contrário (“castigo”). O método visa a melhora constante e possivelmente a maximização de ações que gerem recompensas (Dey, 2016; François-Lavet et al., 2018).

Dentre os métodos de *machine learning* supervisionados existentes, pode-se mencionar o dos K-vizinhos mais próximos, já utilizado no ramo da agricultura para detecção de doenças e pragas (Turkoglu; Hanabay, 2019), avaliações de secagem de solo para planejamento agrícola (Coopersmith et al., 2014), além da predição de produtividade de algodão através de dados agrometeorológicos (Meneses, 2021).

Como anteriormente mencionado, certas metodologias de ML necessitam da partição dos dados em treinamento e teste. A segmentação do conjunto de dados pode ser realizada através de diversos métodos de amostragem. Dentre eles, pode-se citar:

- Amostragem aleatória simples (AAS):

É um tipo de amostragem probabilística no qual se realiza um sorteio aleatório de n números entre os N pertencentes à população estudada de forma que todos os pontos têm a mesma probabilidade de serem incluídos na amostra. A amostragem é realizada com reposição quando os N elementos da população permanecem em todas as retiradas, ou seja, um ponto selecionado é devolvido à população e pode ser novamente escolhido. Já na amostragem sem reposição cada elemento da população original só pode aparecer uma vez na amostra (Bolfarine; Bussab, 2004).

- *Conditioned Latin Hypercube Sampling* ($cLHS$):

A abordagem em questão é baseada no método LHS proposto por McKay, Beckman e Conover (1979) e fornece uma maneira de coletar amostras de variáveis a partir de suas distribuições multivariadas, como proposto por Minasny e McBratney (2006).

O LHS segue a ideia de um quadrado latino em que apenas uma amostra é alocada em cada linha e coluna e generaliza esse conceito para um certo número de dimensões. O método indica que, para um conjunto contendo K variáveis representadas por $(X_1, \ldots, X_k)$, o intervalo de variação de cada variável é subdividido em n intervalos equiprováveis, denominados estratos.

O intuito do método é garantir que cada variável seja bem representada por meio de seus estratos, garantindo uma amostragem representativa da distribuição dos dados estudados.

Entretanto, um dos problemas encontrados ao utilizar o LHS é que as amostras geradas pelo mesmo podem não existir no campo experimental.

Como forma de contornar esse problema, Minasny e McBratney (2006), propuseram um método condicionado capaz de fornecer uma estratégia de amostragem para situações nas quais se trata de uma área com informações prévias.

Minasny e McBratney (2006) ainda concluem que sua metodologia pode ser aplicada para amostras contínuas e categóricas, e deve reproduzir de maneira satisfatória o solo e sua variabilidade, garantindo boas chances de representações de relacionamentos adequadas, caso existam.

3.4.1 Método dos K-vizinhos mais próximos para regressão (KNN)

Idealizado inicialmente por Fix e Hodges (1951) e posteriormente aprimorado por Cover e Hart (1967) o algoritmo dos K-vizinhos mais próximos, do inglês, *K-nearest neighbours,* é um método supervisionado e não paramétrico amplamente conhecido na área de aprendizado de máquina, principalmente quando se fala de classificação, porém, o mesmo também pode ser usado para regressão (Zhang et al., 2018; James et al., 2013)

De forma simplificada o método pode ser resumido em alguns passos:

1) O primeiro passo é a definição do número K de vizinhos utilizados para a predição de um ponto x específico.
2) Em seguida, calculam-se as distâncias (conforme o método mais apropriado) entre o ponto de interesse (x) e todos os pontos do conjunto de treinamento;
3) Selecionam-se as observações que estão mais próximas do ponto x, representadas por N;
4) A predição para o valor do atributo neste ponto é então calculada com base em todos os dados de treinamento, Z_i, que se concentram dentro da vizinhança escolhida, N, como pode ser observado na Equação 8 (James et al., 2013);

$$\hat{Z} = \frac{1}{K} \sum_{xi \,\epsilon\, N} Z_i \tag{8}$$

5) Por fim, analisa-se a qualidade do modelo através das métricas de erro previamente determinadas, como $RMSE$ e MAE.

Uma das grandes dificuldades encontradas ao implementar o método é a escolha do número de vizinhos. Segundo Rahim e Ahmar (2022), existem algumas abordagens para encontrar um valor ótimo para K, sendo elas:

- Validação cruzada: onde o conjunto de dados é dividido em várias camadas de forma que cada camada é escolhida como teste enquanto o restante dos dados é direcionado para o treinamento. O melhor K será aquele que resultar em um algoritmo mais eficiente conforme as medidas de erro utilizadas. Esse método permite que o modelo seja testado com diversas partições dos dados.
- Conhecimento prévio sobre o histórico dos dados: os autores afirmam que para o caso estudado pelos mesmos, onde são analisados registros médicos, um

valor razoável para K é a raiz quadrada do número de dados considerados no treinamento.

- Tentativa e erro: em que diferentes valores de K são testados e o que resultar em predições de maior qualidade segundo as métricas de erro utilizadas é empregado no processo.

Apesar de não existir uma regra pré-estabelecida para a escolha de K, é importante ressaltar que valores muito baixos tendem a resultar em um ajuste com alta variância, já que a estimativa virá de uma região com poucas observações, o que pode não representar bem a vizinhança estudada. Em contrapartida, valores muito altos de K resultam em uma variação menor, mas pode causar certa perda na variabilidade dos dados, já que a predição é realizada com base na média de seus vizinhos, ou seja, quanto maior o valor de K mais parecidas tendem a ser as predições (James et al., 2013).

Determinado o valor de K, o próximo passo para o processo é a escolha da distância a ser utilizada, sendo a mais comumente empregada a distância euclidiana. Entretanto, para o caso de dados correlacionados, a distância de Mahalanobis é mais indicada (Xiang; Nie; Zhang, 2008).

Entre as vantagens do método pode-se mencionar sua simplicidade e robustez para com dados de treinamento com ruídos e também sua eficácia, desde que o conjunto de treinamento seja grande. Já sobre as desvantagens tem-se a sensibilidade do método para com variáveis redundantes já que todas contribuem para o cálculo da predição (Imandoust; Bolandraftar, 2013).

Como forma de reduzir o esforço computacional e melhorar o processo de predição, necessita-se entender quais variáveis preditoras são as mais importantes para a predição de cada uma das variáveis dependentes estudadas. Para tal, o algoritmo *Random Forest* pode ser empregado.

3.4.2 *Random Forest* para regressão

O algoritmo conhecido como Floresta Aleatória, ou *Random Forest* (RF) em inglês, foi idealizado por Breiman (2001). Essa abordagem é baseada nas árvores de decisão, sejam para classificação ou regressão. No entanto, seu diferencial está no fato de não utilizar uma única árvore, mas sim um conjunto variado delas.

Porém, antes de compreender o método RF é interessante entender sobre as árvores. De forma simplificada podemos definir as *Decision trees* (DT), como algoritmos recursivos baseados na divisão contínua de um problema de classificação ou regressão (nó raiz) em vários subproblemas menores (nós de decisão). Tal divisão ocorre até que todos os subproblemas criados sejam solucionados, ou seja, até que todos os atributos pertençam a alguma classe (nó de término) (Garcia, 2003).

Uma vez compreendido o processo de construção de uma árvore de decisão, é possível aplicar esse conhecimento para entender o funcionamento das florestas aleatórias.

De acordo com Liaw e Wiener (2002), o processo de predição ou classificação do algoritmo de floresta aleatória pode ser resumido da seguinte forma:

1. Obtém-se um número denominado "$ntree$" de amostras *bootstrap* (amostragem aleatória e com reposição) dos dados originais;
2. Para cada amostra será obtida uma árvore, seja para classificação ou regressão, sem poda. Uma particularidade do método é que para cada nó será selecionado aleatoriamente um número "$mtry$" entre os atributos preditores, e a partir dessa seleção é determinado o melhor critério de quebra, ou seja, a localização de um nó.
3. As predições serão calculadas levando em consideração o resultado de todas as n árvores de forma que para classificação serão utilizados os votos mais frequentes e para regressão será utilizada a média das predições.

As amostras utilizadas para construção de cada uma das árvores da floresta são retiradas de um conjunto de treinamento, de forma que parte dos dados originais são destinados para validação. Essa parcela de teste também é conhecida como amostra *out-of-bag* (OOB).

Após gerar as predições, é possível identificar as variáveis independentes que exercem maior influência para o resultado encontrado como forma de simplificar o modelo de predição. No contexto do algoritmo RF as duas formas mais encontradas de se quantificar essa importância são conhecidas como *mean decrease in accuracy* (MDA), utilizada para dados quantitativos, e *decrease of gini impurity* (MDG) para dados categóricos, (Hoare, 2018; Hastie, Tibshirani, Friedman, 2009).

Para a métrica MDA ($\%\ de\ aumento\ do\ MSE$) calcula-se a precisão da predição na amostra OOB. Logo em seguida, os valores da variável de interesse nessa amostra

OOB são embaralhados aleatoriamente. Posteriormente, mensura-se a diminuição da precisão nos dados embaralhados. Esse processo é repetido para todas as árvores de forma que o seu resultado informa o quanto retirar ou acrescentar uma variável preditora aumenta, ou diminui a acurácia da predição (Hoare, 2018; Hastie, Tibshirani, Friedman, 2009).

Além das métricas individuais de importância de variáveis, é importante enfatizar a necessidade de se avaliar o desempenho de qualquer método de interpolação (OK, KNN e RF), o que pode ser realizado através da validação cruzada *leave-one-out* (LOOCV).

3.5 Validação cruzada (método *Leave-one-out)*

É um método amplamente conhecido para avaliar o desempenho de algoritmos.

O procedimento de avaliação envolve dividir um conjunto de dados em duas partes. Em uma delas, $(n-1)$ pontos são utilizados para treinamento, a fim de fazer a predição do ponto restante. Isso significa que, para cada atributo em todos os pontos da amostra, tem-se um valor verdadeiro e um valor predito. Com essa abordagem, é possível calcular os erros do processo. Essa técnica é repetida n vezes, e em cada rodada, uma observação diferente é excluída (Ferreira, 2005).

A partir dos valores preditos e observados das n rodadas pode-se avaliar o desempenho dos métodos através do cálculo de estatísticas de erro, as quais são apresentadas a seguir:

- Erro médio absoluto, do inglês, *Mean Absolute Error* (MAE): representa o valor médio dos erros de predição, portanto, espera-se que seja perto de zero.

$$MAE = \frac{\sum_{i=1}^{n} \left| \widehat{Z}_i - Z_i \right|}{n} \tag{9}$$

- Raiz Quadrada do Erro Quadrático Médio, do inglês, *Root Mean Square Error* $(RMSE)$: responsável por avaliar a variabilidade do erro de predição, portanto é desejável que o erro seja o menor possível.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(\widehat{Z}_i - Z_i \right)^2}{n}} \tag{10}$$

Em que, para ambos os casos:

- $\hat{Z}_i$ é o valor predito para o atributo através do método de interpolação;
- Z_i é o valor observado;
- n número total de pontos amostrados.

4. Material e métodos

4.1 Caracterização da área de estudo

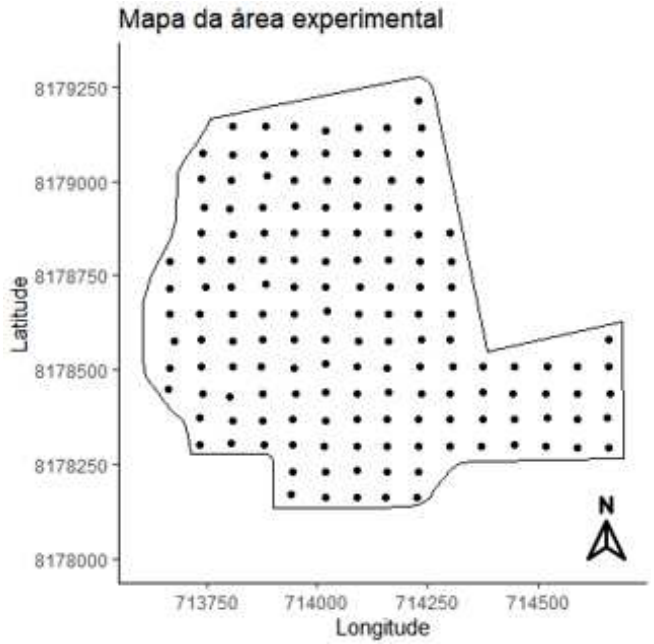
A área em estudo onde a base de dados deste trabalho foi obtida abrange uma parcela da fazenda "Sozinha" (16º28'20'' Sul e 49º00'32'' Oeste) situada em Goianápolis, região metropolitana de Goiânia, conforme documentado por Costa (2011). O município está localizado a 964 m de altitude (Cidade Brasil, 2021), tem como bioma o cerrado (IBGE, 2023) e o tipo de solo característico é o latossolo vermelho amarelo (Costa, 2011).

4.2 Coleta de dados

A base de dados utilizada neste trabalho foi gentilmente cedida por Marcelo Marques Costa (Costa, 2011) e as informações retratadas se referem a um projeto de pesquisa desenvolvido por ele na Universidade Federal de Viçosa para obtenção do seu título de Mestre em Engenharia Agrícola.

O conjunto de dados é constituído por uma grade amostral com 150 pontos (Figura 4) com uma densidade de dois pontos por hectare (ha) em uma área correspondente a $75\ ha$ (Costa, 2011). Em cada um dos pontos foi coletada uma amostra de solo nas profundidades de 0 a $0{,}2\ m$, sendo cada amostra composta por dez subamostras retiradas aleatoriamente a uma distância de até $5\ m$ do ponto.

Figura 4 - Malha regular das amostras de solo coletadas.



As amostras de solo foram analisadas para determinação dos teores de Magnésio (Mg), Cálcio (Ca), Fósforo (P), Zinco (Zn), Ferro (Fe), Manganês (Mn), Cobre (Cu), fósforo remanescente (P-rem), pH_H_2O e acidez potencial (H+Al) (Costa, 2011).

Neste estudo os atributos foram classificados em duas categorias: padrão e específicos. Os atributos padrão abrangem normalmente um conjunto básico de medições realizadas nos laboratórios de análise de solo. Nesse contexto, os atributos padrão incluem pH_H_2O, P, K, Ca, Mg, Al, H+Al e P-rem. Por outro lado, os atributos específicos são aqueles que necessitam de uma solicitação exclusiva para sua análise. Nesse caso, os atributos específicos são Fe, Zn, Mn e Cu.

4.3 Análises estatísticas

Inicialmente foram obtidas as estatísticas descritivas com o objetivo de conhecer o conjunto de dados a ser estudado e as particularidades dos valores de cada um dos atributos mensurados em todos os pontos amostrados.

4.3.1 Krigagem ordinária dos atributos específicos

Após as análises estatísticas dos atributos considerados neste estudo, foi realizada a krigagem ordinária para cada um dos atributos específicos utilizando a grade original, ou seja, os 150 pontos amostrados.

Na etapa de modelagem da dependência espacial foram ajustados os modelos esférico, exponencial e gaussiano. O método dos mínimos quadrados ordinários, do inglês, *ordinary least squares* (OLS), foi utilizado para o ajuste de cada um destes modelos teóricos. O modelo selecionado foi aquele que apresentou o menor erro quadrático médio (MSE), valor este obtido entre os pontos do semivariograma experimental e os respectivos do modelo teórico.

Após a seleção do modelo teórico foi realizada a interpolação por krigagem ordinária utilizando um mínimo de 8 e um máximo de 10 pontos vizinhos. O mapa interpolado foi composto por um *grid* com 10.000 pontos. A qualidade da interpolação foi mensurada por meio das estatísticas do erro médio absoluto (MAE) e raiz quadrada do erro quadrático médio ($RMSE$) obtidas após um estudo da validação cruzada.

4.3.2 *Random Forest* para regressão

Em seguida, o algoritmo *Random Forest* (RF) para regressão foi utilizado a fim de identificar os atributos padrão que exercem maior influência na predição de cada um dos atributos específicos. Como treinamento para o RF foram utilizados 85% dos dados originais e como métrica para a qualidade do modelo construído foi utilizado o $RMSE$ dos dados de teste.

Para cada atributo específico o parâmetro $ntree$ foi estabelecido com base na magnitude dos erros das amostras OOB em função do aumento do número de árvores. Já a escolha do parâmetro $mtry$ foi realizada através de um processo de tentativa e erro. Os valores de $mtry$ que resultaram em menores $RMSE$ nos dados de teste foram escolhidos para dar continuidade ao estudo.

Os resultados da RF permitiram ranquear os atributos independentes com base em sua contribuição para a qualidade do modelo através da percentagem de aumento do MSE (PA_{MSE}). Essa contribuição é medida por meio da comparação entre o MSE do modelo original contendo o atributo específico e o MSE do modelo RF sem o atributo.

Dessa forma, quanto maior a PA_{MSE}, mais importante é o atributo para o modelo gerado, pois esse aumento indica que a retirada desse atributo influencia diretamente na qualidade do ajuste.

A seleção dos atributos foi conduzida mediante a observação de um gráfico que representou os atributos padrão em função do aumento percentual do erro quadrático médio (MSE). Optou-se por incluir na seleção final os atributos que estavam localizados após uma mudança na curvatura do gráfico, de forma que todos os atributos localizados após essa mudança na concavidade fossem incluídos na seleção final como mais importantes.

Após a seleção dos atributos padrão pelo RF, o número de atributos do conjunto de dados original foi reduzido para conter apenas as variáveis essenciais para análise, sendo elas: a latitude, longitude, os atributos mais importantes selecionados e o atributo específico de interesse, logo, foram criados 4 novas grades simplificadas, ou seja, um conjunto simplificado para cada atributo específico.

4.3.3 Redução da dimensionalidade, krigagem ordinária e KNN

A partir de cada uma destas grades simplificadas, foram selecionadas grades reduzidas correspondentes às reduções de $15, 25$ e 35% dos 150 pontos, as quais foram utilizadas como conjuntos de treinamento para o algoritmo de *machine learning* (KNN). Portanto, os respectivos conjuntos de treinamento foram formados por $127, 112$ e 98 pontos.

As grades reduzidas foram obtidas por meio de dois processos: a amostragem aleatória simples (identificadas por AAS_i tal que $i = 127, 112$ e 98) e através do algoritmo *Conditioned Latin Hypercube Sampling* (identificada por $cLHS_i$ tal que $i = 127, 112$ e 98). Para aumentar a acurácia das predições foram realizadas 50 repetições na obtenção de cada AAS_i e $cLHS_i$, ou seja, as grades reduzidas selecionadas foram AAS_{ij} e $cLHS_{ij}$ tal que $i = 127, 112, 98$ e $j = 1, \cdots, 50$.

Os atributos padrão selecionados pelo RF foram utilizados como variáveis explicativas no KNN para predizer o valor de cada atributo específico em cada ponto do conjunto de teste das grades AAS_{ij} e $cLHS_{ij}$.

A escolha do número K $(K = 1, \cdots, 50)$ de vizinhos empregados no algoritmo KNN foi realizada por meio do menor $RMSE$ proveniente da validação cruzada *leave-*

one-out ($LOOCV$). Deu-se preferência para K ímpar para evitar empates durante o processo de escolha da vizinhança.

A localização dos K-vizinhos mais próximos empregados para as predições foi determinada através da distância de Mahalanobis entre o ponto faltante e o restante dos pontos do conjunto.

Ao término do processo KNN cada uma das percentagens ($15, 25$ e 35%) resultou em $23, 38$ e 53 pontos preditos, respectivamente. As coordenadas Leste-Oeste destas predições foram utilizadas para substituir os respectivos valores observados na grade original.

Cada um dos conjuntos de dados modificados foi submetido à interpolação por krigagem ordinária, seguindo o mesmo procedimento de seleção de vizinhos, *grid*, modelagem da dependência espacial e critérios de qualidade do modelo aplicado à grade original.

Em seguida, foram elaborados os mapas de diferença com o intuito de visualizar o quanto os resultados se distanciaram da krigagem dos dados reais. As diferenças foram calculadas ponto a ponto entre as interpolações originais e as interpolações resultantes das amostragens e reduções de dimensionalidade para cada atributo específico.

Foram também calculados os valores médios das estatísticas de validação cruzada $RMSE$ e MAE com base nas 50 repetições de redução amostral, aplicadas aos métodos AAS e $cLHS$, a fim de analisar e comparar o desempenho desses métodos de amostragem.

$$\overline{RMSE}(KO)_{AAS_i} = \frac{\sum_{j=1}^{50} RMSE(Krigagem)_{AAS_{ij}}}{50} \tag{11}$$

$$\overline{RMSE}(KO)_{cLHS_i} = \frac{\sum_{j=1}^{50} RMSE(Krigagem)_{cLHS_{ij}}}{50} \tag{12}$$

$$\overline{MAE}(KO)_{AAS_i} = \frac{\sum_{j=1}^{50} MAE(Krigagem)_{AAS_{ij}}}{50} \tag{13}$$

$$\overline{MAE}(KO)_{cLHS_i} = \frac{\sum_{j=1}^{50} MAE(Krigagem)_{cLHS_{ij}}}{50} \tag{14}$$

Por fim, para comparar as interpolações da krigagem ordinária no banco de dados original e nas grades modificadas foi analisada a razão entre $RMSE$ e MAE

médio das 50 amostragens por $cLHS_{ij}$ e AAS_{ij} e o $RMSE$ e MAE da krigagem dos dados originais.

4.3.4 Ferramental computacional

O Software R versão 4.3.1 (R Development Core Team, 2023) foi utilizado para realizar todas as análises estatísticas e obtenção das grades reduzidas.
Um panorama de todas as etapas do processo é apresentado na Figura 5.

Figura 5 - Síntese da metodologia aplicada no presente trabalho.

5. Resultados

5.1 Análises Estatísticas

A estatística descritiva dos atributos mensurados nos 150 pontos amostrados e as correlações entre eles são apresentadas na Tabela 1 e Figura 6, respectivamente.

Tabela 1 – Estatística descritiva dos atributos estudados.

Atributo	Unidade	Média	Min.	Max.	DP	CV(%)	Curtose	Assimetria
pH_H_2O	-	6,75	5,80	7,60	0,30	4,43	0,49	-0,74
P	$mg.dm^{-3}$	6,84	1,70	21,60	3,96	57,88	1,60	1,27
K	$mg.dm^{-3}$	52,63	24,00	108,00	14,20	26,98	1,86	1,02
Ca^{2+}	$cmolc.dm^{-3}$	3,27	1,90	4,20	0,46	14,04	-0,10	-0,39
Mg^{2+}	$cmolc.dm^{-3}$	0,84	0,60	1,40	0,14	16,53	1,42	0,72
H_Al	$cmolc.dm^{-3}$	1,72	0,00	5,60	0,89	51,92	4,17	1,51
P_rem	$mg.L^{-1}$	17,35	9,50	27,40	3,38	19,48	0,24	0,43
Zn	$mg.dm^{-3}$	3,94	1,50	27,10	2,93	74,52	29,51	4,66
Cu	$mg.dm^{-3}$	1,33	0,80	6,70	0,59	44,35	43,28	5,38
Fe	$mg.dm^{-3}$	21,90	11,00	41,10	5,50	25,12	0,56	0,88
Mn	$mg.dm^{-3}$	27,23	13,70	66,70	9,28	34,07	4,09	1,80

Min = mínimo; Max = máximo, DP= desvio padrão; CV = coeficiente de variação; pH_H_2O = potencial hidrogeniônico; P = fósforo; K = potássio; Ca = cálcio; Mg = magnésio; H_Al = acidez potencial; P_rem = fósforo remanescente; Zn = zinco, Cu = cobre; Fe = ferro e Mn = Manganês.

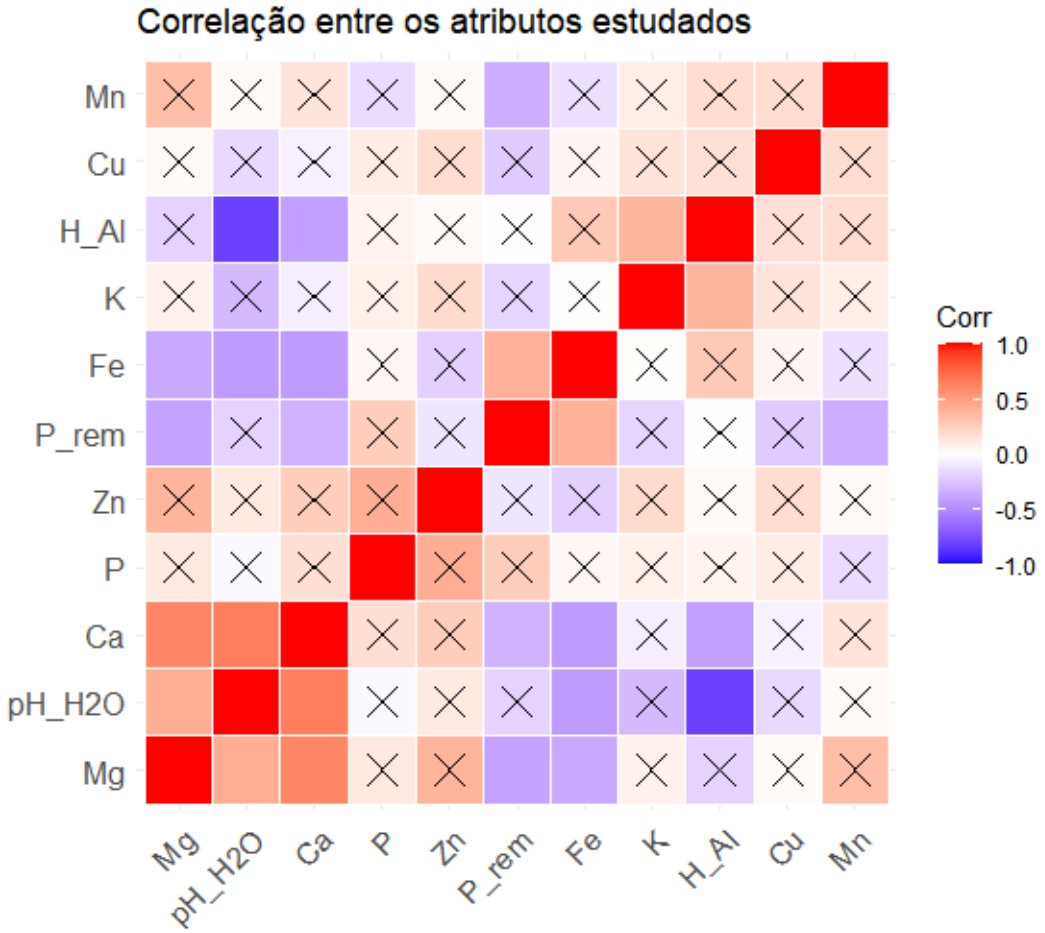
Em relação à descrição dos atributos, Tabela 1, o Zn se destaca como o que apresenta maior dispersão (representada pelo coeficiente de variação), seguido por P e H_Al. No que diz respeito à assimetria das distribuições, o pH, e o Ca demonstram características de distribuições assimétricas a esquerda, onde a média é menor que a mediana e a moda. Em contraste, os demais atributos apresentam assimetria positiva (Portella et al., 2015).

Quanto ao achatamento das distribuições, é notável que Ca exibe uma distribuição platicúrtica, sugerindo uma maior dispersão dos valores e uma menor concentração em torno da média. Os demais atributos apresentam distribuições

leptocúrticas, indicando maior concentração de valores em determinados intervalos (Portella et al., 2015)

Com relação à correlação linear de *Pearson* entre os atributos ressalta-se uma correlação negativa significativa entre os atributos H_Al e pH_H2O e entre Ca e H_Al. Além desta, notam-se as correlações positivas entre Ca e Mg e entre Ca e pH, Ca e pH, Ca e H_Al entre outros, conforme evidenciado na Figura 6.

Figura 6 - Correlação linear de Pearson entre os atributos do solo.



X: não apresenta significância estatística para o teste t a $\alpha = 5\%$.

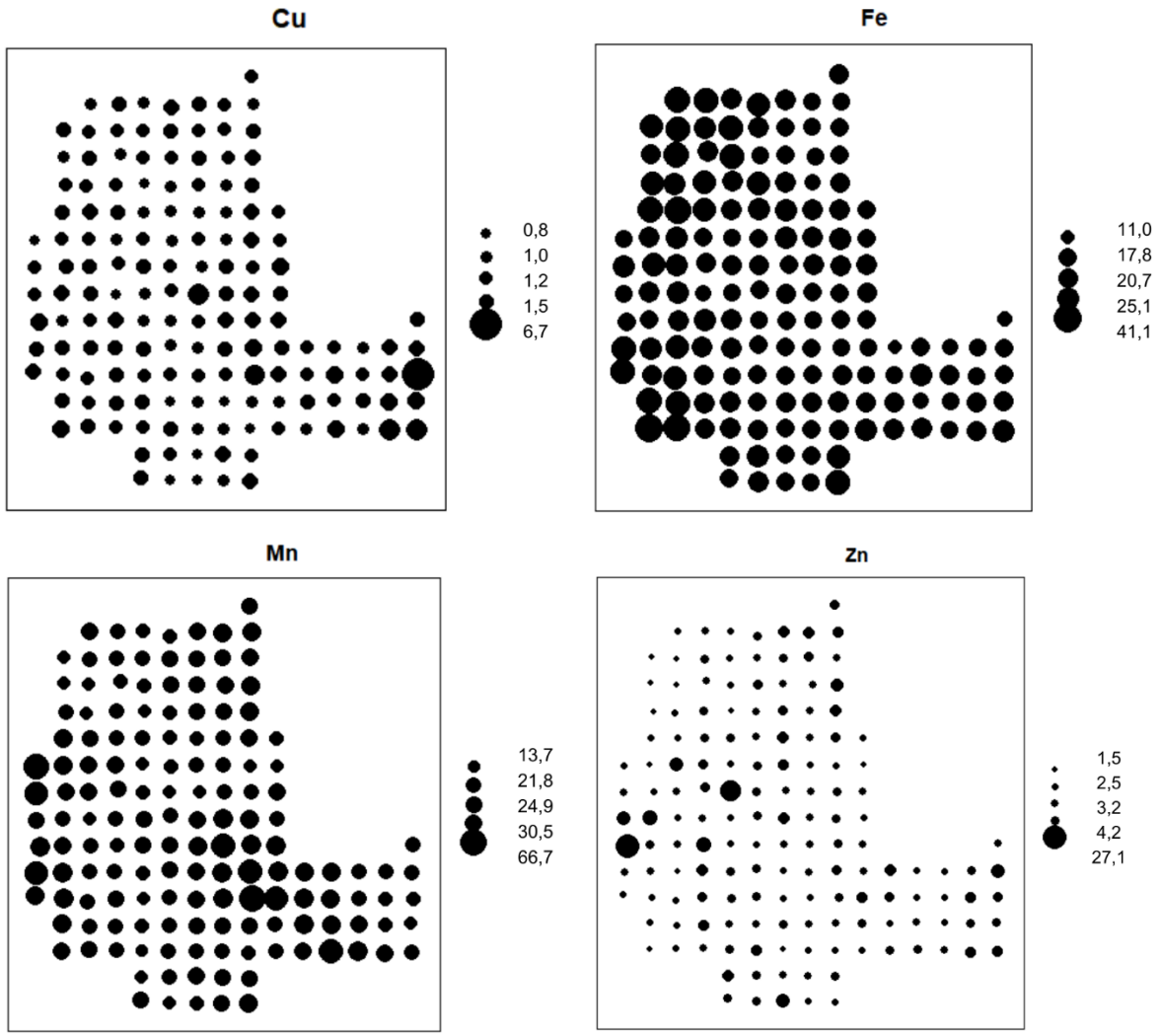
Mg = magnésio, pH = potencial hidrogeniônico; Ca = Cálcio; P = fósforo; Zn = zinco; P_rem = fósforo remanescente, Fe = ferro; K = potássio; H_Al = acidez potencial; Cu = cobre; Mn = manganês.

Os *bubble maps,* Figura 7, descrevem espacialmente a magnitude da concentração de cada um dos atributos nos pontos amostrados. Em geral, nota-se que parece existir uma continuidade espacial para os quatro atributos, ou seja, pontos próximos tendem a apresentar valores semelhantes para o mesmo atributo. Diferentemente dos atributos específicos Fe e Mn, pode-se observar que o Zn e o Cu

parecem possuir menor variabilidade espacial, apresentando apenas alguns pontos com valores discrepantes.

É fundamental ressaltar a importância da continuidade espacial na geoestatística, pois ela desempenha um papel essencial no aprimoramento das técnicas de interpolação, permitindo modelagens e predições mais assertivas.

Figura 7 - *Bubble map* dos atributos específicos estudados.



Cu = cobre, Fe = ferro, Mn = manganês e Zn = zinco.

5.1.1 Krigagem Ordinária dos atributos específicos

Os mapas de interpolação para cada um dos atributos específicos utilizando todos os 150 pontos da grade original e suas estatísticas da validação cruzada são apresentados na Figura 8 e na Tabela 2, respectivamente.

Tabela 2 – Estatísticas de erro obtidas através da validação cruzada para a krigagem ordinária dos 150 pontos para cada um dos atributos específicos.

Atributo específico	Modelo	Efeito Pepita	Alcance	Patamar	$RMSE$	MAE
Cu	-	-	-	-	-	-
Fe	Esférico	5,15	249,47	30,03	3,97	3,14
Mn	Esférico	5,18	314,82	94,74	5,91	4,06
Zn	Exponencial	4,62	224,33	5,44	2,97	1,70

(-) não apresentou dependência espacial.

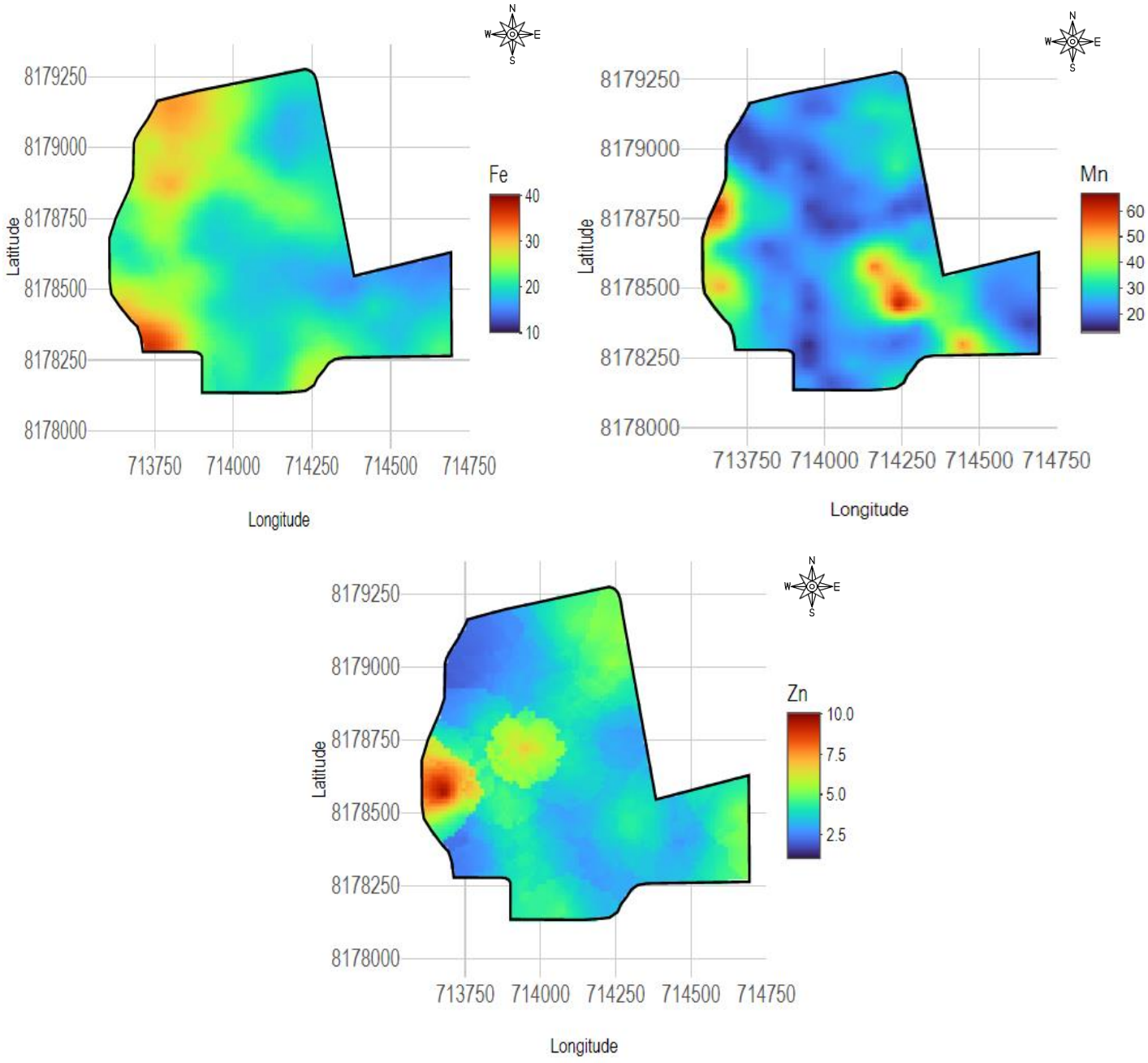
Cu = cobre; Fe = ferro; Mn = manganês; Zn = zinco.

O efeito pepita (Tabela 2) não nulo pode indicar uma menor dependência espacial, o que, por sua vez, pode comprometer a precisão das predições obtidas por meio da krigagem (Guedes; Bach; Uribe-Opazo, 2020). Além disso, este valor não nulo pode ocorrer devido ao processo de amostragem, perfuração e erros intrínsecos do experimento (Camana; Deutsch, 2019). Também pode-se notar que como os modelos teóricos ajustados não foram os mesmos, o padrão da dependência espacial não é idêntico para os atributos específicos.

Entretanto, todas as krigagens ordinárias dos atributos representados na Tabela 2 possuem valores de $RMSE$ próximos do desejável (zero), o que indica que este método de interpolação apresentou um bom desempenho. Na teoria, quanto menor o valor do MAE e do $RMSE$, melhor é a interpolação. No entanto, é importante destacar que a qualidade do modelo pode variar dependendo da quantidade de pontos coletados e da representatividade da amostragem.

Nota-se, através da observação dos mapas interpolados (Figura 8), que o solo é majoritariamente constituído de concentrações mais baixas de Mn e Zn. Já para o Fe, pode-se observar maiores concentrações nas regiões localizadas a noroeste e sudoeste.

Figura 8 - Mapa interpolado referente à krigagem ordinária dos 150 pontos originais para os atributos ferro (Fe), manganês (Mn) e zinco (Zn).



Na Figura 8 também é possível observar que os mapas interpolados pela OK produzem transições graduais entre regiões de concentrações diferentes de um determinado atributo. Além disso, os mapas possuem poucas áreas de mudanças bruscas de concentração (candidatas a *outliers*), os chamados, *Bull's-eyes* (Malvic et al., 2019).

É importante ressaltar que o maior valor de concentração predito para o atributo zinco ($10\ mg.dm^{-3}$) (Figura 8) é menor do que o maior valor real medido deste atributo ($27{,}10\ mg.dm^{-3}$), apresentado na Tabela 1. Tal resultado evidencia um efeito retratado por Isaaks e Srivastava (1989), de que a interpolação por krigagem ordinária produz um efeito suavizante nos resultados, ou seja, as predições tendem a apresentar menor variabilidade que a obtida com base valores observados.

Com o intuito de identificar quais atributos padrão exercem maior influência na predição de cada um dos atributos específicos, foi utilizado o algoritmo *Random Forest.*

5.1.2 *Random Forest* para regressão

Para a utilização do algoritmo RF foi necessário definir o número de árvores ($ntree$), assim como a quantidade de variáveis utilizadas para a divisão de cada nó ($mtry$) das árvores da floresta. Tais parâmetros são apresentados na Tabela 3, juntamente com o $RMSE$ dos dados de treinamento do algoritmo RF.

Tabela 3 - Variáveis de entrada e estatística de erro do algoritmo *Random Forest* para regressão.

Atributo específico	$ntree$	$mtry$	$RMSE$
Fe	≈ 150	2	22,92
Mn	≈ 150	3	63,73
Zn	≈ 150	4	8,99

Fe = ferro; Mn = manganês; Zn = zinco

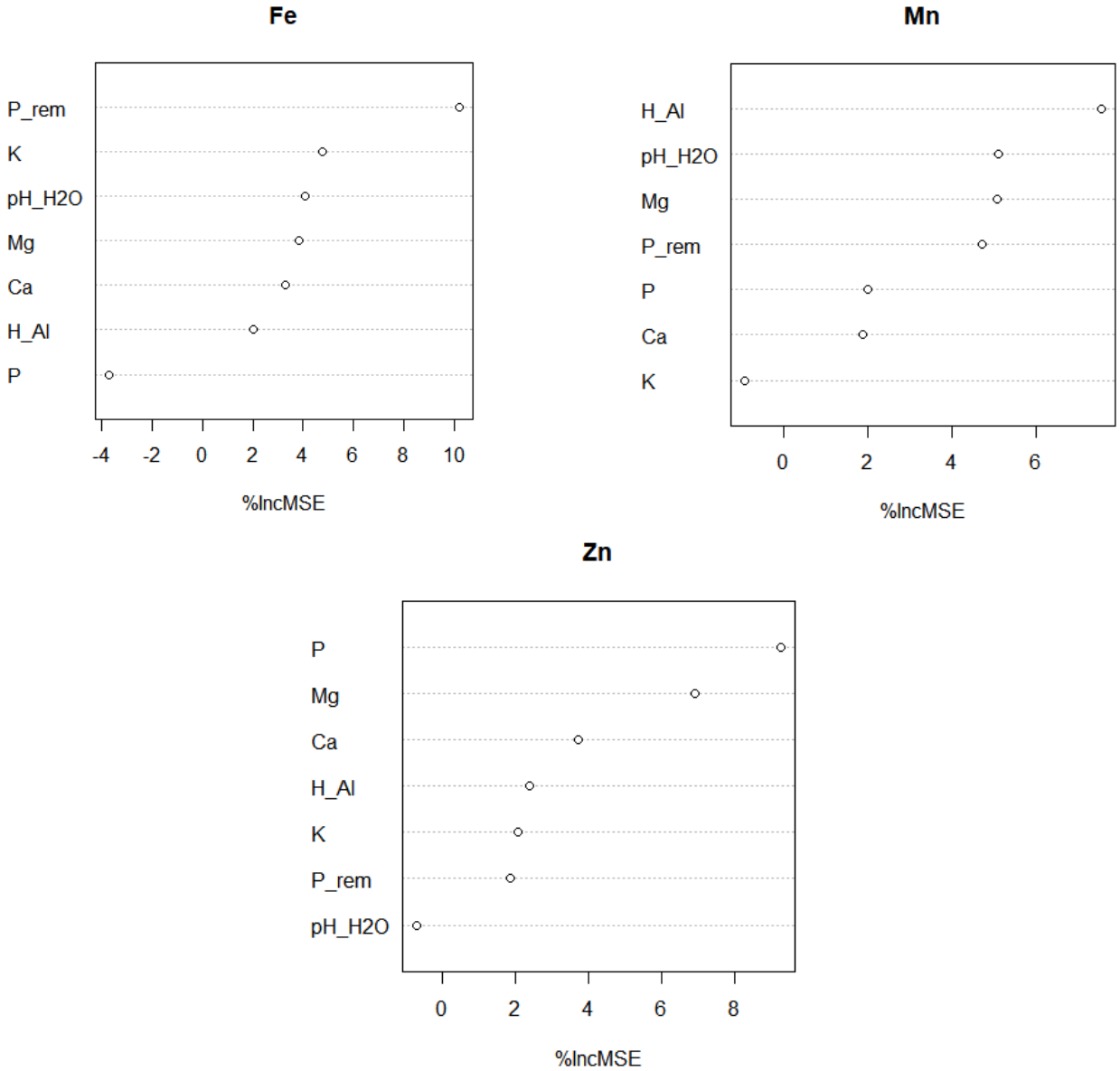
Para todos os atributos específicos, o número de árvores que minimizou o erro das amostras OOB foi de aproximadamente 150. Já a escolha do parâmetro $mtry$ foi realizada através de um processo de tentativa e erro, no qual os valores de $mtry$ variaram entre 2 e 7. Os valores de $mtry$ que resultaram em menores $RMSE$ nos dados de teste foram utilizados para gerar as árvores da floresta (Tabela 3).

Por fim, com o resultado da análise através do RF foi possível ranquear os atributos independentes com base em sua contribuição para a qualidade do modelo, $\%\ de\ aumento\ MSE$ (%IncMSE). Para os processos seguintes de predição através do algoritmo KNN optou-se por reduzir o número de variáveis da base de dados original

para conter apenas os atributos mais importantes para a predição de cada atributo específico.

Essa seleção de atributos foi realizada com base na observação de uma mudança na curvatura do gráfico dos atributos padrão em função do aumento percentual do MSE, onde os atributos localizados após a mudança na concavidade da curva foram incluídos na seleção final como mais importantes, como pode ser observado na Figura 9.

Figura 9 – Importância dos atributos padrão na predição de cada um dos atributos específicos através do aumento percentual de MSE (%IncMSE).



pH_H_2O = potencial hidrogeniônico; P = fósforo; K = potássio; Ca = cálcio; Mg = magnésio; H_Al = acidez potencial; P_rem = fósforo remanescente.

Sendo assim, para o ferro foram utilizados o P_rem, K, pH_H2O e Mg para o manganês foram selecionados o H_Al, pH_H2O, Mg e P_rem. Por fim, para o zinco tem-se o P, Mg, Ca e H_Al.

5.1.3 Redução da dimensionalidade, krigagem ordinária e KNN

As estatísticas de erro médias obtidas por validação cruzada no processo do KNN em todas as porcentagens de redução estudadas (15, 25 e 35%) e diferentes técnicas de amostragem (AAS_i *e* $cLHS_i$) são apresentadas na Tabela 4.

Tabela 4 - Estatísticas de erro médias obtidas por validação cruzada por porcentagem de redução da malha amostral para cada atributo específico e técnica de amostragem.

Atributo específico	Amostragem	$\overline{RMSE}_{KNN}$	Amostragem	$\overline{RMSE}_{KNN}$
Fe	$cLHS_{127}$	6,57	AAS_{127}	7,28
	$cLHS_{112}$	7,15	AAS_{112}	7,39
	$cLHS_{98}$	7,28	AAS_{98}	7,49
Mn	$cLHS_{127}$	10,07	AAS_{127}	12,27
	$cLHS_{112}$	11,30	AAS_{112}	12,65
	$cLHS_{98}$	11,65	AAS_{98}	12,74
Zn	$cLHS_{127}$	1,29	AAS_{127}	2,66
	$cLHS_{112}$	1,36	AAS_{112}	2,97
	$cLHS_{98}$	1,80	AAS_{98}	2,98

Fe = ferro; Mn = manganês; Zn = zinco

Através da observação dos $RMSE$ médios encontrados na Tabela 4 pode-se notar que à medida que o conjunto de treinamento do algoritmo KNN diminui ($i = 127 \rightarrow 98$) o $RMSE$ médio aumenta, ou seja, o modelo passa a errar mais em suas predições independentemente do tipo de amostragem considerado. Tal resultado é condizente com o esperado, visto que o treinamento do modelo foi realizado com cada vez menos informação, o que tende a acarretar predições mais imprecisas devido a uma menor representatividade do conjunto de dados real nos dados de treinamento (Ramezan et al., 2021).

Após a predição realizada por meio do algoritmo KNN, os métodos de amostragem (AAS e $cLHS$) e suas respectivas reduções de dimensionalidade ($i = 127 \rightarrow 98$) foram substituídos no conjunto de dados original e posteriormente submetidos à técnica de krigagem ordinária, seguindo o mesmo procedimento utilizado para o conjunto de dados original. Os resultados das estatísticas de erro médio comparadas com os erros relacionados à krigagem da grade original estão representados na Tabela 5.

É interessante ressaltar que não foi possível ajustar um modelo de dependência espacial para o atributo zinco para a amostragem aleatória simples nas reduções $i = 112$ e 98 (Tabela 5). Já para o método $cLHS$ esse problema não foi encontrado, pois, segundo Minasny e Mcbratney (2006), esse método visa amostrar eficientemente para manter a variabilidade dos atributos encontrados no solo estudado.

Tabela 5 – Razão entre as médias das estatísticas de erro RMSE e MAE das grades reduzidas e original para cada atributo específico e tipo de amostragem.

Atributo específico	Amostragem	$\frac{\overline{RMSE}_{Krig}}{RMSE}$	$\frac{\overline{MAE}_{Krig}}{MAE}$	Amostragem	$\frac{\overline{RMSE}_{Krig}}{RMSE}$	$\frac{\overline{MAE}_{Krig}}{MAE}$
Fe	$cLHS_{127}$	1,73	1,70	AAS_{127}	1,69	1,64
	$cLHS_{112}$	1,70	1,69	AAS_{112}	1,67	1,60
	$cLHS_{98}$	1,65	1,62	AAS_{98}	1,63	1,52
Mn	$cLHS_{127}$	1,98	2,14	AAS_{127}	1,85	1,97
	$cLHS_{112}$	1,91	2,04	AAS_{112}	1,74	1,82
	$cLHS_{98}$	1,86	1,98	AAS_{98}	1,64	1,71
Zn	$cLHS_{127}$	1,05	1,06	AAS_{127}	0,99	1,01
	$cLHS_{112}$	1,05	1,06	AAS_{112}	-	-
	$cLHS_{98}$	1,04	1,02	AAS_{98}	-	-

(-) não apresentou dependência espacial.

Fe = ferro; Mn = manganês; Zn = zinco

Contudo, diferentemente do processo KNN, a krigagem ordinária resultou em uma diminuição do $RMSE$ médio à medida que a porcentagem de redução aumenta independentemente do processo de amostragem considerado (Tabela 5). Sabe-se

que o resultado da krigagem é extremamente dependente da etapa anterior de predição por KNN.

Como relatado por Ramezan et al. (2021), se o KNN é treinado com poucos dados, ele tende a gerar predições mais semelhantes entre si e discrepantes com relação à variabilidade real do conjunto de dados inicial. Como resultado, obtém-se uma interpolação por krigagem com cada vez menos variabilidade espacial e menor erro (como representado pelo aumento do $RMSE$ e MAE na Tabela 5), já que como as predições por KNN se tornam cada vez mais parecidas, as da krigagem não poderiam ser diferentes.

Além da observação das estatísticas de erro, também é possível observar as diferenças entre o mapa da krigagem ordinária dos dados originais com cada um dos mapas resultantes das amostragens e reduções de dimensionalidade.

Para a confecção de cada um dos mapas representados nas Figuras 10, 11 e 12 foi calculada a diferença ponto a ponto entre as interpolações retratadas na Figura 8 e as interpolações resultantes das amostragens e reduções de dimensionalidade para cada atributo específico.

Pode-se observar que, para o Fe na amostragem AAS e Mn em ambos os métodos, há um padrão de aumento nas áreas representadas por maiores diferenças (denotadas pelos tons de amarelo e vermelho nos mapas das Figuras 10 e 11) à medida que se aumenta a redução de dimensionalidade de 15 para 35%.

No caso da amostragem $cLHS$ para o Fe, embora as mudanças de coloração não sejam tão perceptíveis, nota-se que o intervalo de variação das diferenças segue uma tendência crescente, indicada pelo aparecimento da coloração azul nos mapas de redução de 25% e 35% (Figura 10). Ao comparar os mapas gerados por ambas as amostragens em cada uma das reduções de dimensionalidade, observa-se que o método $cLHS$ performa melhor que a AAS. Isso se deve à menor concentração de resultados que destoam das diferenças nulas, indicadas pelos tons de amarelo e verde na Figura 10.

Em relação ao atributo Zn, Figura 12, ao utilizar as reduções de dimensionalidade de 25% e 35% com a amostragem AAS, não foi possível obter o mapa das diferenças, pois não se conseguiu ajustar um modelo teórico de semivariância. No entanto, ao comparar os mapas das diferenças do AAS de 15% com os mapas das diferenças das reduções de $cLHS$, observa-se que as diferenças nulas

se distribuíram espacialmente de forma semelhante, sendo que os maiores valores absolutos das diferenças foram encontrados na mesma região.

Nota-se também que independente do atributo específico estudado e do método de amostragem escolhido, quanto maior a porcentagem de redução, maior é a discrepância entre os mapas interpolados da grade original e das grades modificadas (Figuras 10, 11 e 12). O que reafirma que quanto menor a porção de treinamento dos dados (maior a redução de dimensionalidade), piores as predições através do KNN e menor a variabilidade espacial, o que reflete diretamente na krigagem.

A importância de uma amostragem com a quantidade de pontos adequada é destacada pelos resultados obtidos por Sahu, Ghosh e Seema (2021) em sua pesquisa sobre a predição do teor de carbono orgânico do solo na Índia. Os autores concluíram que a coleta de amostras em intervalos menores, ou seja, com uma maior densidade de pontos, conduziu a predições mais precisas por meio da krigagem. Isso ocorre porque uma das principais limitações desse método é a sua tendência a gerar predições menos confiáveis quando a densidade amostral é baixa (Pouladi et al., 2019).

Tal afirmação também foi comprovada por Qu et al. (2023) e Li et al. (2007) em seus estudos sobre a predição de areia e salinidade do solo utilizando o método de krigagem ordinária. Ambos os estudos indicaram que a diminuição do tamanho amostral resultou em um aumento nos valores de $RMSE$ e MAE associados à krigagem. Esses resultados mostram que a qualidade das predições encontradas pela krigagem está diretamente ligada à quantidade e distribuição das amostras coletadas.

Assim, o estudo de métodos que busquem reduzir a densidade de pontos amostrados no solo, mantendo uma representação satisfatória da área de estudo, pode representar uma vantagem relacionada ao mapeamento de atributos. A implementação de tais métodos pode ocasionar uma economia de recursos e maior velocidade na obtenção de informações.

Figura 10 - Mapas interpolados referentes as diferenças entre as amostragens $cLHS$ e AAS e as predições por krigagem ordinária para o atributo Fe nas reduções de dimensionalidade de 15, 25 e 35%, respectivamente.

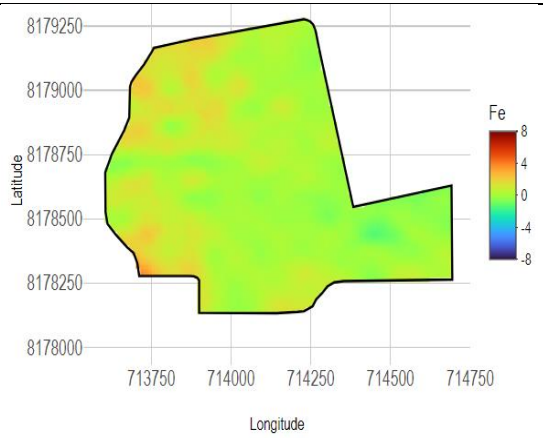
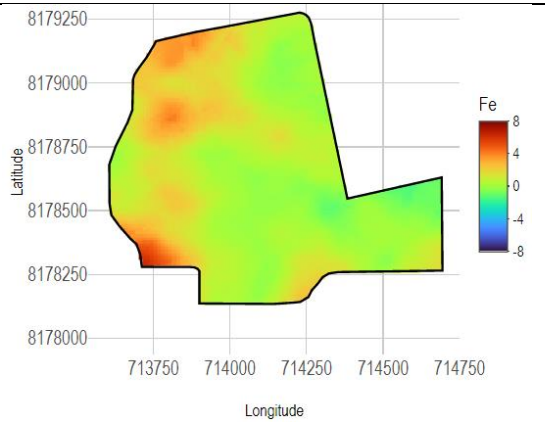
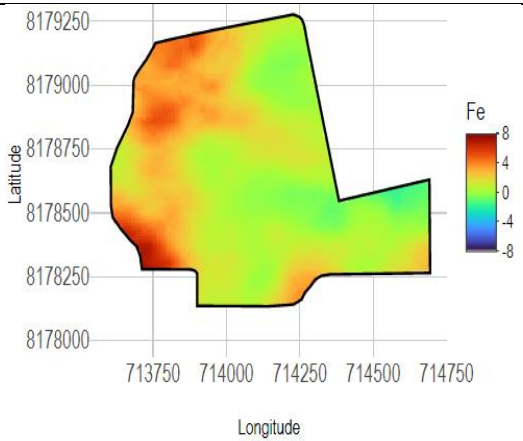
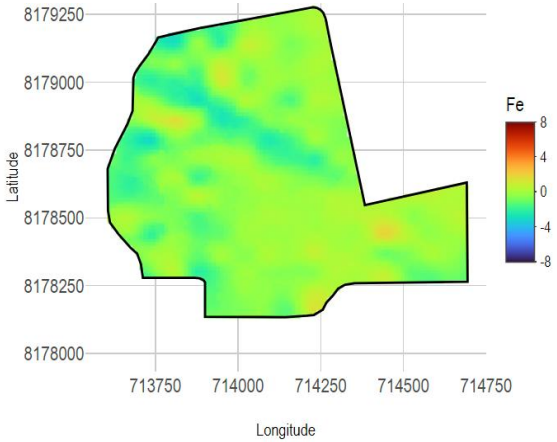
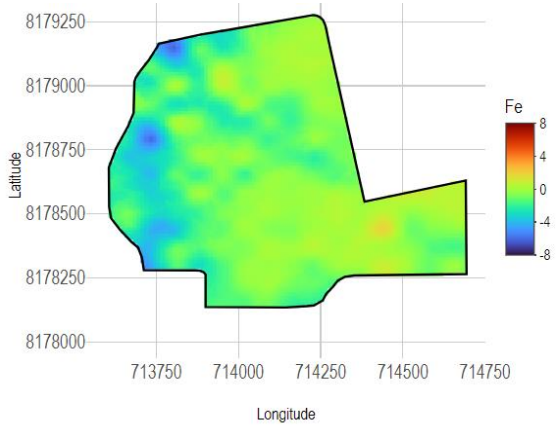
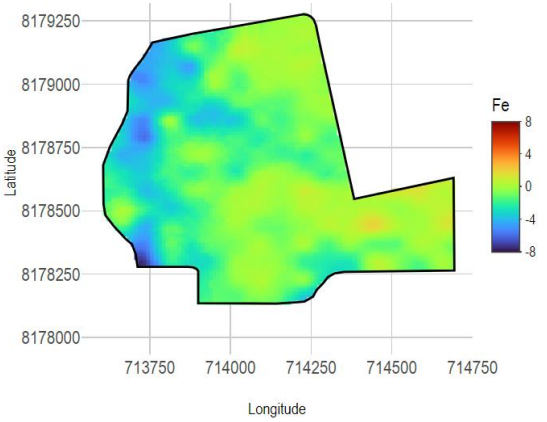
	15%	**25%**	**35%**
AAS			
	15%	**25%**	**35%**
cLHS			

Figura 11 - Mapas interpolados referentes as diferenças entre as amostragens $cLHS$ e AAS e as predições por krigagem ordinária para o atributo Mn nas reduções de dimensionalidade de 15, 25 e 35%, respectivamente.

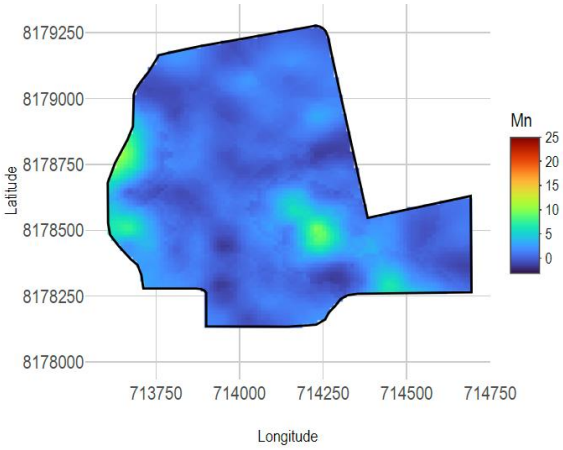
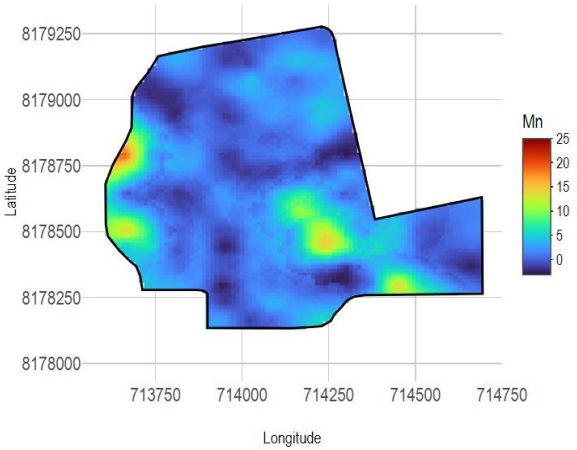
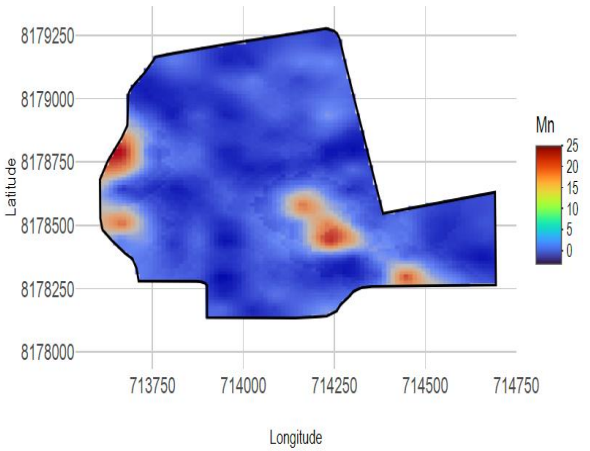
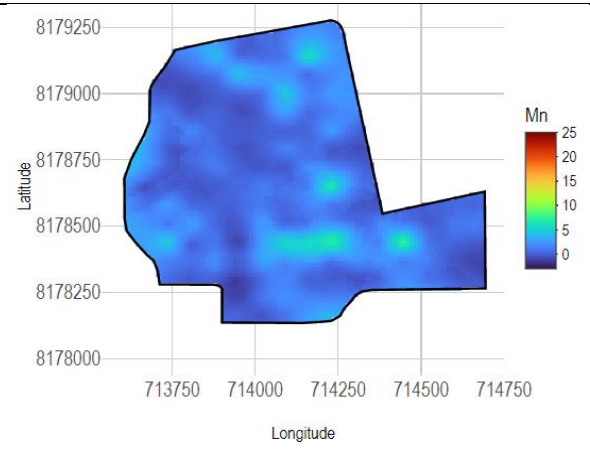
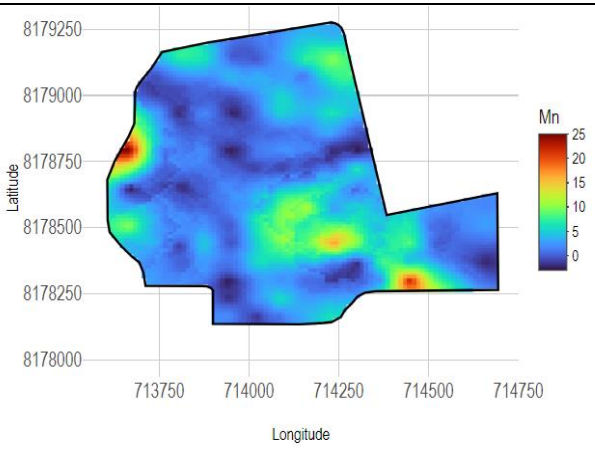
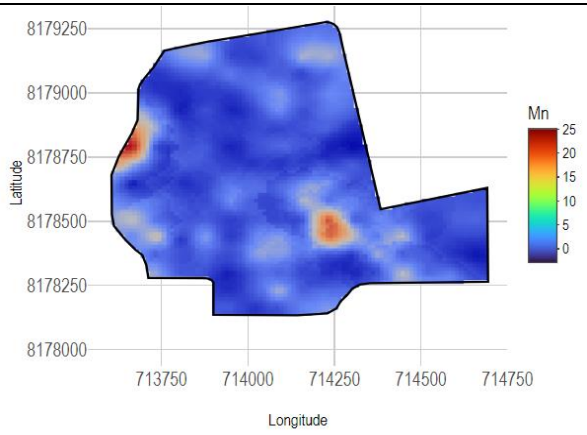
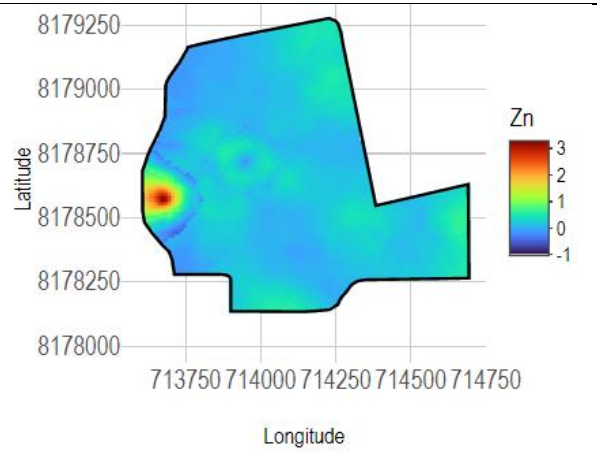
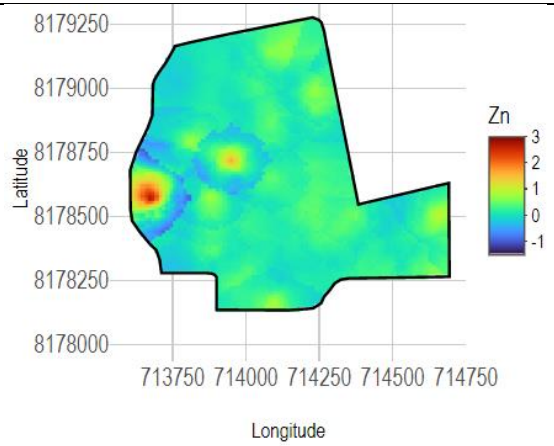
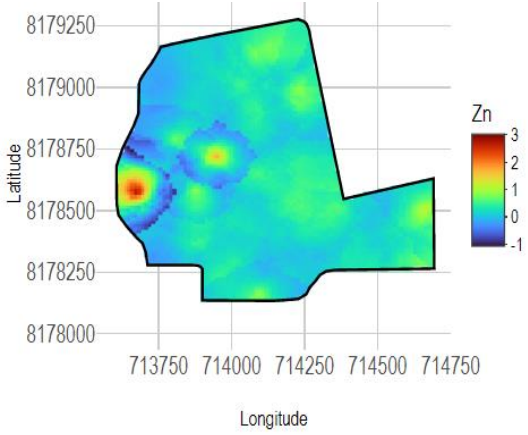
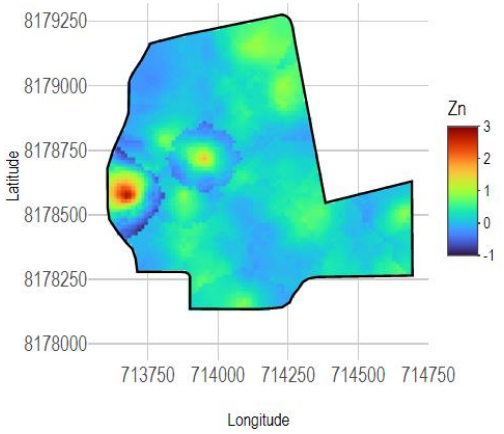
	15%	**25%**	**35%**
AAS			
cLHS			

Figura 12 - Mapas interpolados referentes as diferenças entre as amostragens $cLHS$ e AAS e as predições por krigagem ordinária para o atributo Zn nas reduções de dimensionalidade de 15, 25 e 35%, respectivamente.

	15%	**25%**	**35%**
AAS		-	-
cLHS	**15%**	**25%**	**35%**
			

(-) não apresentou dependência espacial.

6. Conclusão

A mensuração de atributos do solo, apesar de extremamente importante para seu manejo, pode elevar consideravelmente o custo do processo a depender do número de análises das amostras coletadas.

É possível reduzir parte do trabalho relacionado a coleta e análise de atributos do solo a partir da predição de uma parte de suas características com perda da variabilidade espacial através de algoritmos de *machine learning*, em especial o KNN.

Entretanto, deve-se ter em mente que quanto maior for o tamanho da amostra predita (maior a redução de dimensionalidade), menos acurada a representação da área, independentemente do método de amostragem utilizado. Isto ocorre porque o KNN é treinado com poucos dados, e, portanto, tende a gerar predições mais semelhantes entre si e discrepantes com relação à variabilidade real do conjunto de dados inicial. Como resultado obtém-se uma interpolação por krigagem com cada vez menos variabilidade espacial.

Dentre os dois métodos de amostragem estudados, o $cLHS$ se mostrou melhor em manter as características espaciais do solo para os três atributos específicos estudados (ferro, manganês e zinco) frente às reduções de dimensionalidade de 15, 25 e 35% quando comparado a amostragem aleatória simples, que foi incapaz de representar os dados nas reduções de 25 e 35% para o atributo zinco.

Sugere-se, ainda, para trabalhos futuros, que sejam estudadas novas metodologias de ML (além do KNN) combinadas à krigagem ordinária, além de tipos de amostragem distintos como forma a avaliar seu comportamento frente a redução de amostras analisadas quimicamente.

7. Referências

AGTERBERG, F. Georges Matheron: Founder of Spatial Statistics. **Earth Sciences History**, v. 23, n. 2, p. 205–334, 1 jan. 2004.

BARRENA-GONZÁLEZ, J.; LAVADO CONTADOR, J. F.; PULIDO FERNÁNDEZ, M. Mapping Soil Properties at a Regional Scale: Assessing Deterministic vs. Geostatistical Interpolation Methods at Different Soil Depths. **Sustainability**, v. 14, n. 16, p. 10049, 13 ago. 2022.

BOLFARINE, H.; BUSSAB, W. O. **Elementos de Amostragem**. São Paulo: Edgard Blucher, 2004.

BREIMAN, L. Machine Learning. **Random Forests**, p. 28, 2001.

CAMANA, F. A.; DEUTSCH, C. V. **The Nugget Effect. In J.L. Deutsch (Ed.), Geostatistics Lessons.** , 2019. Disponível em: <http://geostatisticslessons.com/lessons/nuggeteffect>

CARVALHO, J. R. P. DE; VIEIRA, S. R. Avaliação e Comparação de Estimadores de Krigagem para Variáveis Agronômicas – Uma Proposta. n. 1ª, p. 24, 2001.

CARVALHO, J. R. P. D.; SILVEIRA, P. M. D.; VIEIRA, S. R. Geoestatística na determinação da variabilidade espacial de características químicas do solo sob diferentes preparos. **Pesquisa Agropecuária Brasileira**, v. 37, n. 8, p. 1151–1159, ago. 2002.

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

MALVIĆ, T. et al. Kriging with a Small Number of Data Points Supported by Jack-Knifing, a Case Study in the Sava Depression (Northern Croatia). **Geosciences**, v. 9, n. 1, p. 36, 11 jan. 2019.

PORTELLA, A. C. F. et al. **Estatística Básicapara os cursos de Ciências Exatas e Tecnológicas**. [s.l.] EDUFT, 2015.

RAMEZAN, C. A. et al. 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, 21 jan. 2021.

CIDADE BRASIL. **Município de Goianápolis**. Disponível em: <https://www.cidade-brasil.com.br/municipio-goianapolis.html>. Acesso em: 15 ago. 2023.

COOPERSMITH, E. J. et al. Machine learning assessments of soil drying for agricultural planning. **Computers and Electronics in Agriculture**, v. 104, p. 93–104, jun. 2014.

COSTA, M. M. **CONDUTIVIDADE ELÉTRICA APARENTE DO SOLO COMO FERRAMENTA PARA AGRICULTURA DE PRECISÃO EM UMA ÁREA SOB CERRADO**. Dissertação - Mestado em Engenharia Agrícola—Viçosa MG: Universidade Federal de Viçosa, 20 jul. 2011.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE Transactions on Information Theory**, v. 13, n. 1, p. 21–27, jan. 1967.

CRESSIE, N. The origins of kriging. **Mathematical Geology**, v. 22, n. 3, p. 239–252, abr. 1990.

DE IACO, S.; HRISTOPULOS, D. T.; LIN, G. Special Issue: Geostatistics and Machine Learning. **Mathematical Geosciences**, v. 54, n. 3, p. 459–465, abr. 2022.

DERDOURI, A.; MURAYAMA, Y. A comparative study of land price estimation and mapping using regression kriging and machine learning algorithms across Fukushima prefecture, Japan. **Journal of Geographical Sciences**, v. 30, n. 5, p. 794–822, maio 2020.

DEY, A. Machine Learning Algorithms: A Review. v. 7, 2016.

DRIDI, S. **Unsupervised Learning - A Systematic Literature Review**. [s.l.] Open Science Framework, 28 dez. 2021. Disponível em: <https://osf.io/kpqr6>. Acesso em: 16 ago. 2023.

ERDOGAN ERTEN, G.; YAVUZ, M.; DEUTSCH, C. V. Combination of Machine Learning and Kriging for Spatial Estimation of Geological Attributes. **Natural Resources Research**, v. 31, n. 1, p. 191–213, fev. 2022.

FAROOQ, I. et al. Comparison of Random Forest and Kriging Models for Soil Organic Carbon Mapping in the Himalayan Region of Kashmir. **Land**, v. 11, n. 12, p. 2180, 1 dez. 2022.

FIX, E.; HODGES, J. L. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. **I USAF School of Aviation Medicine,** Randolph Field, Tex., projeto 21-49-004, Rept. 4, contrato AF41(128)-31, fev. 1951.

FRANÇOIS-LAVET, V. et al. An Introduction to Deep Reinforcement Learning. **Foundations and Trends® in Machine Learning**, v. 11, n. 3–4, p. 219–354, 2018.

GARCIA, S. C. **O Uso de Árvores de Decisão na Descoberta de Conhecimento na Área da Saúde**. Dissertação - Mestrado em Ciência da Computação—Porto Alegre: UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL, 2003.

GUEDES, L. P. C.; BACH, R. T.; URIBE-OPAZO, M. A. NUGGET EFFECT INFLUENCE ON SPATIAL VARIABILITY OF AGRICULTURAL DATA. **Engenharia Agrícola**, v. 40, n. 1, p. 96–104, fev. 2020.

HARTKAMP, A. D.; STEIN, A.; WHITE, J. W. Interpolation Techniques for Climate Variables. 1999.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. Random Forests. Em: **The Elements of Statistical Learning**. Springer Series in Statistics. New York, NY: Springer, 2009. p. 745.

HOARE, J. **How is Variable Importance Calculated for a Random Forest? Display R**, 30 jul. 2018. Disponível em: <https://www.displayr.com/how-is-variable-importance-calculated-for-a-random-forest/>. Acesso em: 1 ago. 2023

IBGE, I. B. DE G. E E. **IBGE Goianápolis**. , 2023. Disponível em: <https://cidades.ibge.gov.br/brasil/go/goianapolis/panorama>. Acesso em: 15 ago. 2023

IMANDOUST, S. B.; BOLANDRAFTAR, M. Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background. v. 3, n. 5, 2013.

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

JAKOB, A. A. E.; YOUNG, A. F. O uso de métodos de interpolação espacial de dados nas análises sociodemográficas. 2006.

JAMES, G. et al. **An introduction to statistical learning with applications in R**. [s.l: s.n.]. v. 6

KRIGE, D. G. Journal of the Southern African Institute of Mining and Metallurgy. **A statistical approach to some basic mine valuation problems on the Witwatersrand**, n. 52, p. 119–139, 1951.

KWAGHTYO, D. K.; EKE, C. I. Smart farming prediction models for precision agriculture: a comprehensive survey. **Artificial Intelligence Review**, v. 56, n. 6, p. 5729–5772, jun. 2023.

LANDIM, P. M. B. Sobre Geoestatística e mapas. **Terrae Didatica**, v. 2, n. 1, p. 19, 2006.

LI, Y. et al. Improved Prediction and Reduction of Sampling Density for Soil Salinity by Different Geostatistical Methods. **Agricultural Sciences in China**, v. 6, n. 7, p. 832–841, jul. 2007.

LIAKOS, K. et al. Machine Learning in Agriculture: A Review. **Sensors**, v. 18, n. 8, p. 2674, 14 ago. 2018.

LIAW, A.; WIENER, M. Classification and Regression by randomForest. v. 2, 2002.

LIU, Q.; WU, Y. Supervised Learning. Em: SEEL, N. M. (Ed.). **Encyclopedia of the Sciences of Learning**. Boston, MA: Springer US, 2012. p. 3243–3245.

MASSRUHÁ, S. M. F. S. M. Agricultura digital: pesquisa, desenvolvimento e inovação nas cadeias produtivas. 2020.

MATHERON, G. Principles of geostatistics. **Economic Geology**, v. 58, n. 8, p. 1246–1266, 1 dez. 1963.

MATHERON, G. **Les variables régionalisées et leur estimation une application de la théorie des fonctions aléatoires aux**. Paris: Maisson, 1965.

MATHERON, G. Em: **The Theory of Regionalized Variables and Its Applications**. [s.l.] École Nationale Supérieure des Minas de Paris, 1971. p. 211.

MATHEUS DE PAULA FERREIRA. **REDUÇÃO DO ADENSAMENTO AMOSTRAL NO AJUSTE DE MODELOS DE SEMIVARIOGRAMAS**. Dissertação—Viçosa MG: Universidade Federal de Viçosa, 30 jul. 2005.

MCFADDEN, J.; NJUKI, E.; GRIFFIN, T. Precision Agriculture in the Digital Era: Recent Adoption on U.S. Farms. p. 53, fev. 2023.

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. 21, n. 2, p. 239–245, maio 1979.

MELLO, J. M. D. **Geoestatística aplicada ao inventário florestal.** Doutorado em Recursos Florestais—Piracicaba: Universidade de São Paulo, 7 out. 2004.

MENESES, K. C. DE. **RUMO À AGRICULTURA INTELIGENTE: PREVISÃO DE PRODUTIVIDADE AGRÍCOLA COM DADOS AGROMETEOROLÓGICOS USANDO MACHINE LEARNING**. Tese - Doutora em Agronomia (Ciência do Solo)—Jaboticabal: Universidade Estadual Paulista, 2021.

MEUL, M.; VAN MEIRVENNE, M. Kriging soil texture under different types of nonstationarity. **Geoderma**, v. 112, n. 3–4, p. 217–233, mar. 2003.

MINASNY, B.; MCBRATNEY, A. B. A conditioned Latin hypercube method for sampling in the presence of ancillary information. **Computers & Geosciences**, v. 32, n. 9, p. 1378–1388, nov. 2006.

MINISTÉRIO DA AGRICULTURA E PECUÁRIA. **Agricultura de Precisão**. **gov.br**, 30 nov. 2016. Disponível em: <https://www.gov.br/agricultura/pt-br>. Acesso em: 23 nov. 2023

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre Aprendizado de M´aquina. Em: **Sistemas Inteligentes Fundamentos e Aplicações**. 1ª ed ed. Barueri - SP: Manole Ltda, 2003. p. 89–114.

MOTOMIYA, A. V. D. A.; CORÁ, J. E.; PEREIRA, G. T. Uso da krigagem indicatriz na avaliação de indicadores de fertilidade do solo. **Revista Brasileira de Ciência do Solo**, v. 30, n. 3, p. 485–496, jun. 2006.

OMRAN, E.-S. E. Improving the Prediction Accuracy of Soil Mapping through Geostatistics. **International Journal of Geosciences**, v. 03, n. 03, p. 574–590, 2012.

POULADI, N. et al. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. **Geoderma**, v. 342, p. 85–92, maio 2019.

QU, L. et al. 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, out. 2023.

RAHIM, R.; AHMAR, A. S. Cross-Validation and Validation Set Methods for Choosing K in KNN Algorithm for Healthcare Case Study. **JINAV: Journal of Information and Visualization**, v. 3, n. 1, p. 57–61, 31 jul. 2022.

RIBEIRO JUNIOR, P. J. **Métodos geoestatísticos no estudo da variabilidade espacial de parâmetros do solo**. Mestrado em Estatística e Experimentação Agronômica—Piracicaba: Universidade de São Paulo, 3 mar. 1995.

RIOS, É. D. S. **O EFEITO DE BORDA NA GEOESTATÍSTICA**. Dissertação—Viçosa MG: Universidade Federal de Viçosa, 2018.

SAHU, B.; GHOSH, A. K.; SEEMA. Deterministic and geostatistical models for predicting soil organic carbon in a 60 ha farm on Inceptisol in Varanasi, India. **Geoderma Regional**, v. 26, p. e00413, set. 2021.

SAITO, H. et al. Geostatistical interpolation of object counts collected from multiple strip transects: Ordinary kriging versus finite domain kriging. **Stochastic Environmental Research and Risk Assessment**, v. 19, n. 1, p. 71–85, fev. 2005.

SOUSA, Í. V. D. **CONSTRUÇÃO ANALÍTICA DE SEMIVARIOGRAMAS MÉDIOS PARA KRIGAGEM DE BLOCOS**. Dissertação - Mestado em Estatística e Experimentação Agropecuária—Lavras - MG: Universidade Federal de Lavras, 3 mar. 2020.

SPIAZZI, F. R. UNIVERSIDADE DO ESTADO DE SANTA CATARINA – UDESC CENTRO DE CIÊNCIAS AGROVETERINÁRIAS – CAV PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS AGRÁRIAS MESTRADO EM MANEJO DO SOLO. 2011.

TÜRKOĞLU, M.; HANBAY, D. Plant disease and pest detection using deep learning-based features. **TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES**, v. 27, n. 3, p. 1636–1651, 15 maio 2019.

UNITED NATIONS DEPARTMENT FOR ECONOMIC AND SOCIAL AFFAIRS. **WORLD POPULATION PROSPECTS 2022: summary of results**. S.l.: UNITED NATIONS, 2023.

VIOLA, M. R. et al. Métodos de interpolação espacial para o mapeamento da precipitação pluvial. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v. 14, n. 9, p. 970–978, set. 2010.

WOLFERT, S.; GOENSE, D.; SORENSEN, C. A. G. **A Future Internet Collaboration Platform for Safe and Healthy Food from Farm to Fork**. 2014 Annual SRII Global Conference. **Anais**... Em: 2014 ANNUAL SRII GLOBAL CONFERENCE (SRII). San Jose, CA, USA: IEEE, abr. 2014. Disponível em: <http://ieeexplore.ieee.org/document/6879694/>. Acesso em: 26 nov. 2023

WUEST, T. et al. Machine learning in manufacturing: advantages, challenges, and applications. **Production & Manufacturing Research**, v. 4, n. 1, p. 23–45, jan. 2016.

XIANG, S.; NIE, F.; ZHANG, C. Learning a Mahalanobis distance metric for data clustering and classification. **Pattern Recognition**, v. 41, n. 12, p. 3600–3612, dez. 2008.

XU, Y. et al. Machine learning in construction: From shallow to deep learning. **Developments in the Built Environment**, v. 6, p. 100045, maio 2021.

YAMAMOTO, J. K.; LANDIM, P. M. B. **Geoestatística conceitos e aplicações**. [s.l.] Oficina de Textos, 2013.

ZHANG, S. et al. A novel k NN algorithm with data-driven k parameter computation. **Pattern Recognition Letters**, v. 109, p. 44–54, jul. 2018.