GUSTAVO HENRIQUE DA SILVA

DETERMINAÇÃO DE BLOCOS EXPERIMENTAIS UTILIZANDO GEOESTATÍSTICA, COMPONENTES PRINCIPAIS E AGRUPAMENTO

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: Antônio Policarpo de Souza Carneiro

Coorientadores: Luiz Alexandre Peternelli
Gerson Rodrigues dos Santos

VIÇOSA – MINAS GERAIS
2022

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

T
S586d
2022

Silva, Gustavo Henrique, 1996-
Determinação de blocos experimentais utilizando geoestatística, componentes principais e agrupamento / Gustavo Henrique Silva. – Viçosa, MG, 2022.
1 dissertação eletrônica (68 f.): il. (algumas color.).

Inclui apêndices.
Orientador: Antonio Policarpo Souza Carneiro.
Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Estatística, 2022.
Referências bibliográficas: f. 58-60.
DOI: https://doi.org/10.47328/ufvbbt.2022.249
Modo de acesso: World Wide Web.

1. Geologia – Métodos estatísticos. 2. Krigagem. 3. Análise espacial (Estatística). 4. Planejamento experimental.
5. Algorítmos. I. Carneiro, Antonio Policarpo Souza, 1973-.
II. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 551.028

Bibliotecário(a) responsável: Bruna Silva CRB6/2552

GUSTAVO HENRIQUE DA SILVA

DETERMINAÇÃO DE BLOCOS EXPERIMENTAIS UTILIZANDO GEOESTATÍSTICA, COMPONENTES PRINCIPAIS E AGRUPAMENTO

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: 25 de fevereiro de 2022.

Assentimento:

Gustavo Henrique da Silva
Autor

Antônio Policarpo de Souza Carneiro
Orientador

A Alane, minha companheira;

A minha mãe, Célia Aparecida;

A meu pai, Cláudio Luciano.

AGRADECIMENTOS

A Universidade Federal de Viçosa e ao programa de pós-graduação em Estatística Aplicada e Biometria (PPESTBIO), pela oportunidade de realizar o curso de pós-graduação.

A Fundação de Amparo à Pesquisa do Estado de Minas Gerais – FAPEMIG – pela concessão da bolsa de estudos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

A Alane, minha companheira, pelo apoio.

Ao meu orientador, professor Antônio Policarpo Souza Carneiro, pela ótima orientação e prontidão em ajudar.

Ao professor Luiz Alexandre Peternelli, pela ajuda na coorientação, ideias e participação na banca.

Ao professor Gérson Rodrigues dos Santos, pela colaboração e coorientação neste trabalho.

A professora Lidiane Maria Ferraz Rosa por integrar a banca, fornecendo importantes contribuições ao trabalho.

Ao Programa de Melhoramento Genético da Cana-de-Açúcar (PMGCA) da UFV, pela disponibilização dos dados.

Aos técnicos do CECA (Centro de Pesquisa e Melhoramento de Cana-de-Açúcar - CECA, Oratórios-MG), um dos campos experimentais do PMGCA-UFV, e aos diversos estudantes que auxiliaram na coleta dos dados experimentais usados nessa pesquisa.

“Chamar o especialista em estatística depois que o experimento foi feito pode ser o mesmo que pedir a ele para fazer um exame post-mortem. Talvez ele consiga dizer de que foi que o experimento morreu.”

(R. A. Fisher)

RESUMO

SILVA, Gustavo Henrique, M.Sc., Universidade Federal de Viçosa, fevereiro de 2022. **Determinação de blocos experimentais utilizando geoestatística, componentes principais e técnicas de agrupamento.** Orientador: Antônio Policarpo Souza Carneiro. Coorientadores: Gérson Rodrigues dos Santos e Luiz Alexandre Peternelli.

Em experimentos que envolvem o princípio do controle local, a determinação mais assertiva dos blocos experimentais é um ponto que se destaca no planejamento experimental. Uma forma interessante de realizar tal procedimento seria a utilização de técnicas que analisam e agrupam regiões do solo que sejam mais semelhantes entre si, definindo assim blocos mais homogêneos. Assim, objetivou-se definir blocos experimentais para experimentos agrícolas utilizando técnicas de geoestatística, análise de componentes principais e técnicas de agrupamento. Foram usados dados de variáveis químicas de solo de um experimento com cana-de-açúcar. As técnicas de geoestatística foram aplicadas de modo a identificar e caracterizar a variabilidade espacial das variáveis químicas, bem como realizar a krigagem em locais não amostrados. Foi aplicado a técnica de análise de componentes principais, visando reduzir a quantidade de variáveis. Por último, a técnica de agrupamento *k-means* foi usada para formar os blocos usando os componentes principais que explicaram boa parte da variabilidade contida nas variáveis originais. Das 12 variáveis químicas do solo 10 apresentaram dependência espacial sendo os locais não amostrados interpolados via krigagem ordinária. Os dois outros atributos foram interpolados via técnica do inverso da distância. A análise de componentes principais permitiu reduzir a dimensionalidade dos dados de 12 variáveis para duas, explicando 82,27% da variância inicial. A área experimental foi dividida em 2, 3, 4 e 5 blocos através do algoritmo *k-means* sendo esses de formato e áreas diversas. A metodologia proposta se mostrou eficaz na delimitação dos blocos, uma vez que esses se apresentam bem uniformes para as variáveis químicas de solo.

Palavras-chave: variabilidade espacial. Controle local. Precisão experimental. Krigagem. Delineamento experimental. K-means.

ABSTRACT

SILVA, Gustavo Henrique, M.Sc., Universidade Federal de Viçosa, February, 2022. **Determination of experimental blocks with the aid of geostatistics, principal components and clustering techniques.** Adviser: Antônio Policarpo Souza Carneiro. Co-advisers: Gérson Rodrigues dos Santos and Luiz Alexandre Peternelli.

In experiments involving the principle of local control, the most assertive determination of the experimental blocks is a point that stands out in the experimental design. An interesting way to carry out such a procedure would be the use of techniques that analyze and group soil regions that are more similar to each other, thus defining more homogeneous blocks. Thus, the objective was to define experimental blocks for agricultural experiments using geostatistical techniques, principal component analysis and clustering techniques. Data from soil chemical variables from an experiment with sugarcane were used. Geostatistical techniques were applied in order to identify and characterize the spatial variability of chemical variables, as well as to perform kriging in unsampled locations. The principal component analysis technique was applied, aiming to reduce the number of variables. Finally, the k-means clustering technique was used to form the blocks using the principal components that explained a good part of the variability contained in the original variables. Of the 12 soil chemical variables, 10 showed spatial dependence and the unsampled sites were interpolated via ordinary kriging. The two other attributes were interpolated via the inverse distance technique. Principal component analysis allowed reducing the dimensionality of the data from 12 variables to two, explaining 82.27% of the initial variance. The experimental area was divided into 2, 3, 4 and 5 blocks using the k-means algorithm, with different formats and areas. The proposed methodology proved to be effective in the delimitation of the blocks, since they are very uniform for the soil chemical variables.

Keywords: Spatial variability. Local control. Experimental precision. Kriging. Experimental design. K-means.

SUMÁRIO

1. INTRODUÇÃO

Para Formaggio e Sanches (2017), a agricultura desempenha um papel insubstituível em qualquer país à medida em que é diretamente responsável pela produção de alimentos, fibras e matérias-primas para a produção de biocombustíveis. O Brasil, por sua vez, sempre manteve uma relação forte com a agricultura. Desde o início do período Colonial, pós 1500, o território brasileiro vem desempenhando um papel importante na produção agrícola mundial. Saltando muitos anos de história e tomando a produção de grãos como exemplo, o país produziu 45 milhões de toneladas em 1975, progredindo para 187 milhões em 2013 (BUAINAIN *et al.*, 2014).

O Brasil se destaca mundialmente, estando entre os quatro maiores exportadores de cana-de-açúcar, soja, café, algodão e milho. Entre os anos 1990 e 2011, o saldo positivo da balança comercial agrícola brasileira cresceu de forma muito superior em relação aos demais países do globo: saltando de U$7 bilhões para cerca de US$70 bilhões (BUAINAIN *et al.*, 2014).

Tomando como exemplo a cultura da cana-de-açúcar, o Brasil possui uma das maiores áreas de cultivo do mundo. Na safra 2014/2015 uma produção de 639 milhões de toneladas das quais resultaram 20% da produção mundial de açúcar. Além disso, a cana-de-açúcar é empregada na produção de etanol – fonte de energia renovável e com menores impactos ambientas – sendo o Brasil o segundo maior produtor (NOCELLI *et al.,* 2017).

Apesar do aumento histórico nos níveis de produção, Formaggio e Sanches (2017) destacam que, até o ano de 2050, será necessário duplicar os níveis de produção agrícola no planeta, visando atender a segurança alimentar. Ao mesmo tempo, os mesmos autores expõem a necessidade de reduzir drasticamente os impactos ambientais relativos à agricultura.

A solução para o problema citado envolve duas possibilidades: aumento da área plantada ou aumento na produtividade. No entanto, cabe ressaltar que o aumento da área plantada está na contramão da redução dos impactos ambientais, pois muitas vezes esse processo está associado a destruição do habitat de várias espécies, emissão de gases de efeito estufa e desgastes de solos (FORMAGGIO; SANCHES, 2017).

Neste contexto, a experimentação agrícola surge como ferramenta indispensável para aumentar os níveis de produtividade das culturas. É por intermédio

dela que novos avanços são possíveis, levando a consolidação de novos conhecimentos e aumentos de produtividade. No entanto, Neto, Scarminio e Bruns (2001) destacam que a experimentação, em muitas das vezes, consome vários meses de trabalho levando a elevados custos com remuneração de profissionais qualificados e despesas com insumos.

Para evitar gastos desnecessários de tempo e de recursos financeiros, o planejamento experimental é fundamental (BANZATTO; KRONKA, 2006). Os autores Neto, Scarminio e Bruns (2001, p. 9) destacam que

> Um pesquisador que desconheça a metodologia do planejamento experimental corre o risco de chegar a conclusões duvidosas. Pior ainda, pode acabar realizando experimentos que não levem a conclusão alguma, duvidosa ou não, e cujo único resultado prático seja o desperdício de tempo e dinheiro.

Dentre os três princípios básicos da experimentação (repetição, casualização e controle local), o controle local, apesar de facultativo, é utilizado com frequência. De acordo com Banzatto e Kronka (2006), em experimentação agrícola, o princípio do controle local consiste em dividir o ambiente sabidamente heterogêneo em subambientes mais homogêneos denominados blocos. Este procedimento tem como objetivo reduzir o erro experimental de modo a tornar o experimento mais preciso, através do controle de uma fonte de variação sistemática.

Gomes (1990) considera que o experimento delineado em blocos casualizados seja o tipo mais importante. Nesse sentido, esse mesmo autor condiciona a eficiência do experimento a necessidade de se estabelecer blocos tão uniformes quanto possível. Isso é justificado pelo fato de que uma eventual variação indesejada dentro dos blocos poderá maximizar os confundimentos em relação aos tratamentos que estão em teste.

Diante desta realidade é fundamental para aumentar a precisão dos experimentos de campo, definir, de forma mais eficiente possível, a posição e o tamanho dos blocos experimentais. Assim, o objetivo do presente trabalho foi propor uma metodologia utilizando ferramentas da geoestatística, da análise de componentes principais e de técnicas de agrupamento para a formação de blocos mais homogêneos, para características químicas do solo, em uma área experimental de cana-de-açúcar.

Na segunda seção deste trabalho (Revisão de bibliográfica) será apresentado um resumo teórico a respeito das técnicas de geoestatística, análise de componentes principais, técnica de interpolação pelo inverso do quadrado da distância e sobre o algoritmo de agrupamento *k-means*. Já na seção 3 (Materiais e métodos) será discutida a metodologia empregada no trabalho. Os resultados obtidos e a discussão serão apresentadas na seção 4. Por fim, a seção 5 será destinada às conclusões mais relevantes do trabalho.

2. REVISÃO BIBLIOGRÁFICA

2.1. A Geoestatística

A Geoestatística é uma subárea da Estatística sendo considerada a melhor metodologia disponível para analisar a distribuição e variação espacial de variáveis regionalizadas. Suas técnicas permitem compreender fenômenos espaciais que inicialmente se escondem atrás dos valores numéricos, *a priori,* considerados aleatórios, das variáveis em estudo (BERNARDI *et al.*, 2014).

Yamamoto e Landim (2015, p. 19) afirmam que "A Geoestatística tem por objetivo a caracterização espacial de uma variável de interesse por meio do estudo de sua distribuição e variabilidade espaciais [...]". Ressalta-se que a metodologia se destaca pelo fato de calcular as incertezas associadas ao processo de estimação. (JOURNEL, HUIJBREGTS, 1976)

Por intermédio das técnicas de interpolação pontual da Geoestatística (krigagem) podemos estimar valores numéricos da variável de interesse em locais não amostrados. Os estimadores empregados na estimação fornecem estimativas não tendenciosas e com variância mínima. De posse dos valores estimados, pode-se elaborar mapas de alta precisão que evidenciam a estrutura e variação espacial do fenômeno em análise (BERNARDI *et al.*, 2014).

2.1.1. Fenômeno espacial e amostragem sistemática

Um fenômeno espacial pode ser entendido como o conjunto de todos os valores de uma certa variável que se distribui no espaço bidimensional ou tridimensional. Portanto, esse conjunto de valores pode ser interpretado como a população estatística (YAMAMOTO; LANDIM, 2015).

Como é de costume na pesquisa, aferir a variável de interesse em toda população, realizando o censo, é praticamente inviável. A falta de tempo, recursos financeiros e inacessibilidade de toda população são apenas alguns dos motivos que indicam a necessidade de se recorrer a algum processo de amostragem.

Neto (2002) cita que "Uma amostra é, pois, um subconjunto de uma população, necessariamente finito, pois todos os seus elementos serão examinados para efeito da realização do estudo estatístico desejado". Nesse sentido, Yamamoto e Landim

(2015, p. 21) expõem que, em análises geoestatística, a amostragem sistemática fornece os melhores resultados.

Geralmente, em geoestatística, define-se um plano de amostragem sistemática através de uma malha regular quadrada ou retangular, onde cada nó dessa malha corresponderia a um local a ser amostrado. A aleatoriedade desse processo viria do fato de se realizar um sorteio para escolher o primeiro local onde a amostra seria colhida. Entretanto, na maioria das vezes, o pesquisador opta por decidir o primeiro local de amostragem, buscando assim potencializar o processo como um todo (YAMAMOTO; LANDIM, 2015).

2.1.2. Variograma experimental e modelos teóricos

Após realizar o processo de amostragem, um passo importante na análise geoestatística é a obtenção do variograma experimental. É esse objeto geoestatístico que indicará se há, ou não, dependência espacial para a variável de interesse.

O variograma, muitas vezes também chamado de semivariograma, é um gráfico (conjuntos de pontos), que descreve o comportamento da variância em função do vetor de distâncias **h**, que separa dois pontos amostrados quaisquer. Deve-se observar que, como o variograma depende de uma grandeza vetorial, esse poderá ser diferente em função do módulo ou direção do vetor **h**. No caso em que o variograma é invariável sobre a mudança de direção considerada dizemos que o fenômeno espacial em estudo é isotrópico. A situação contrária, onde há mudanças no variograma em função da direção, caracteriza um fenômeno espacial anisotrópico (VIEIRA, 2000).

Define-se a semivariância entre dois pontos amostrados como

$$\gamma(h) = \frac{1}{2} E\{[Z(x+h) - Z(x)]^2\}, \quad (1)$$

em que $Z(x)$ é o valor observado do ponto amostrado em x e $Z(x+h)$ o valor observado na posição $x+h$ (ISAAKS; SRIVASTAVA, 1989).

De acordo com Vieira (2000), podemos estimar os valores das semivariâncias plotadas no variograma a partir da Equação 2

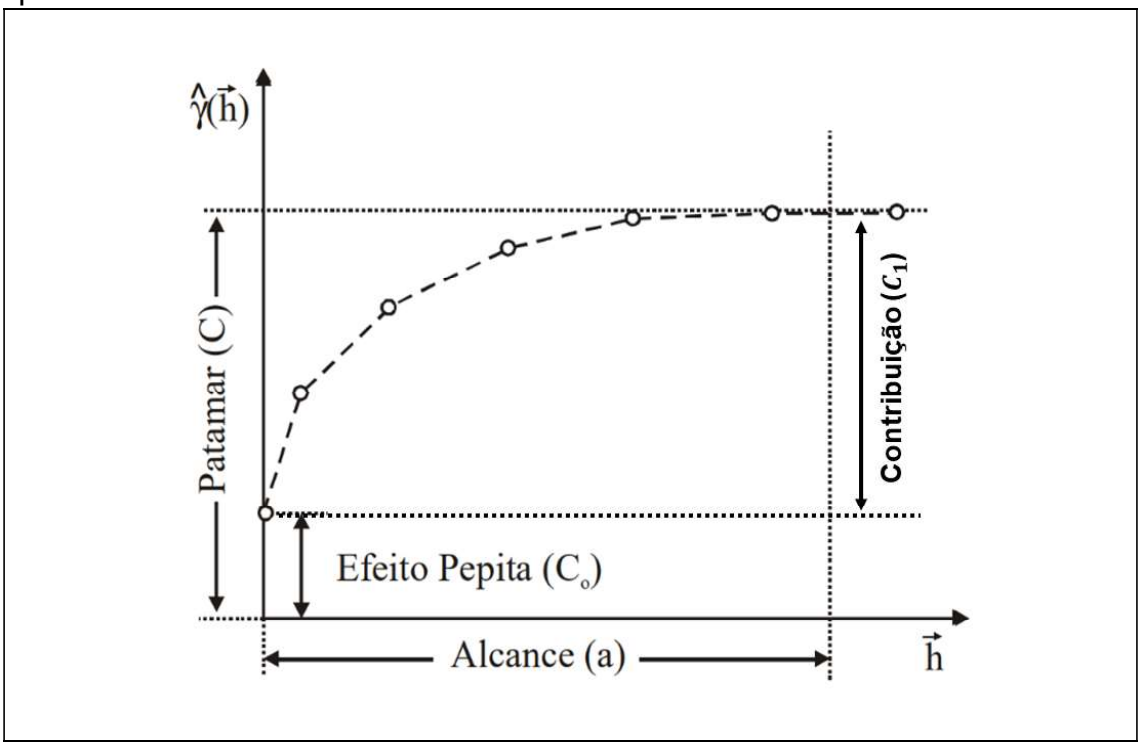
$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \{[Z(x_i + h) - Z(x_i)]^2\}, \quad (2)$$

em que a soma é feita até $N(h)$, número de pontos amostrados separados por uma distância h.

Os variogramas se dividem entre aqueles com ou sem patamar. Para que exista dependência espacial, é preciso que o variograma apresente-se de forma crescente em função do módulo do vetor **h**. A partir de uma determinada distância, conhecida como alcance, os valores das semivariâncias podem se estabilizar, caracterizando o patamar do variograma. Nos casos em que as semivariâncias estimadas crescem indefinidamente, estamos diante de um fenômeno com capacidade infinita de dispersão (YAMAMOTO; LANDIM, 2015).

A Figura 1 ilustra as principais características que devem ser observadas em um variograma experimental com patamar.

Figura 1 – Aspectos fundamentais a serem observados em um variograma com patamar.



Fonte: MONTEIRO et. al., 2004, p. 80.

O alcance, denotado por a, é a distância a partir da qual o variograma alcança seu patamar. Em Geoestatística, esse parâmetro é muito importante, porque indica

até que distância há correlação espacial entre os valores amostrados, o que restringiria as técnicas de estatística inferencial baseadas em independência.

O parâmetro efeito pepita, denotado por C_0, é o valor que o variograma atinge quando a distância tende a zero. Em princípio, à medida em que as distâncias ficam cada vez menores, espera-se que o variograma também tenda a zero. No entanto, na prática, isso pode não acontecer, caracterizando assim a existência do efeito pepita. Yamamoto e Landim (2015) afirmam que a manifestação do efeito pepita está relacionada com variações aleatórias.

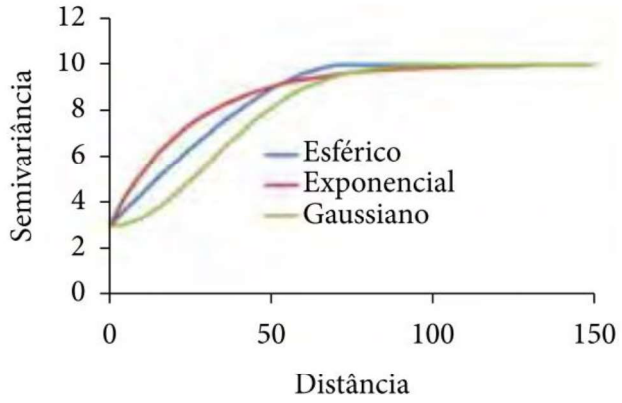
O patamar, simbolizado por C, geralmente é representando como a soma de dois outros fatores: C_0 e C_1. O fator C_1 caracterizaria a contribuição atribuída ao fenômeno espacial em estudo que, somada a contribuição aleatória expressa pelo efeito pepita, seria responsável por levar o variograma ao seu valor assintótico (o patamar).

De posse do variograma experimental é necessário realizar um ajuste de algum modelo teórico de curva conhecida. Bernardi *et al.* (2014, p. 79) destacam que

> Ajuste de modelo ao semivariograma é um dos aspectos mais importantes das aplicações da geoestatística porque os cálculos de geoestatística dependem do valor do modelo do semivariograma para cada distância especificada.

A despeito da existência de vários modelos teóricos para ajuste aos dados do variogramas, apenas três são usados com maior frequência nos processos de modelagem: esférico, exponencial e gaussiano (YAMAMOTO; LANDIM, 2015). Nota-se que as diferenças entre os modelos teóricos se devem a questões de concavidade, inclinação e comportamento assintótico. A Figura 2 apresenta esses três modelos enquanto as Equações 3, 4 e 5 apresentam as equações dos modelos esférico, exponencial e gaussiano, respectivamente.

Figura 2 – Modelos teóricos mais usados para ajuste de variogramas experimentais: esférico, exponencial e gaussiano.



Fonte: Bernadi *et al.,* 2014, p. 79.

$$\gamma(h) = \begin{cases} C_0 + C_1\left[\frac{3}{2}\left(\frac{h}{a}\right) - \frac{1}{2}\left(\frac{h}{a}\right)^3\right], & se\ \ 0 < h < a \\ C_0 + C_1\,, & se\quad h \geq a \end{cases} \tag{3}$$

$$\gamma(h) = C_0 + C_1\left[1 - \exp\left(-\left(\frac{h}{a}\right)\right)\right], \tag{4}$$

$$\gamma(h) = C_0 + C_1\left[1 - \exp\left(-\left(\frac{h}{a}\right)^2\right)\right]. \tag{5}$$

2.1.3. Krigragem ordinária

Observada a dependência espacial, deve-se proceder a krigagem, metodologia própria da Geoestatística que leva em conta a dependência espacial anteriormente detectada.

Yamamoto e Landim (2015, p. 55) citam que:

> Krigagem é um processo geoestatístico de estimativa de valores e variáveis distribuídas no espaço e/ou tempo, com base em valores adjacentes quando considerados interdependentes pela análise variográfica. Pode ser comparado com os métodos tradicionais de estimativa por médias ponderadas ou por médias

> móveis, mas a diferença fundamental é que somente a krigagem apresenta estimativas não tendenciosas e a mínima variância associada ao valor estimado.

Dispondo dos Z_1, Z_2,..., Z_n valores amostrados nas posições 1, 2, ..., n, respectivamente, podemos estimar o valor para uma posição x_0, não amostrada, realizando a soma expressa pela Equação 6

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i \, Z(x_i), \tag{6}$$

em que $\hat{Z}(x_0)$ é o valor estimado em x_0, λ_i os pesos associados a cada valor previamente amostrado e $Z(x_i)$ os valores amostrados.

O que difere a krigagem de outros algoritmos de estimação, é a forma como os pesos λ_i são calculados e aplicados, bem como o número de vizinhos que serão levados em consideração.

Dentre todos os tipos distintos de krigagem, a krigagem ordinária (KO) é o método mais aplicado devido ao fato de apresentar grande simplicidade, bem como trazer resultados bem interessantes (YAMAMOTO; LANDIM, 2015). De acordo com Isaaks e Srivastava (1989) a krigagem ordinária é um estimador pontual linear, sendo a estimativa combinação linear do dados disponíveis.

Considerando o modelo teórico de variograma ajustado e partindo dos pressupostos da krigagem ser um estimador não viesado e com mínima variância, Isaaks e Srivastava (1989) desenvolvem a Equação 7 para se obter os pesos λ_i presentes na Equação 6 e realizar a estimativa no local não amostrado:

$$\begin{cases} \sum_{j=1}^{n} \lambda_j \, \gamma(x_i - x_j) + \mu = \gamma(x_0 - x_i) \;\; para \; i = 1,2,\dots,n \\ \sum_{j=1}^{n} \lambda_j = 1 \end{cases} \tag{7}$$

em que λ_i são os pesos a serem determinados, $\gamma(x_i - x_j)$ a semivariância calculada para a distância que separa $x_i \; e \; x_j$, e μ é o multiplicador de Lagrange.

Além de realizar a estimativa no local não amostrado, a Geoestatística apresenta ferramentas para calcular a variabilidade na qual as estimativas são feitas. (YAMAMOTO; LANDIM, 2015). A Equação 8 é usada para calcular a variância da krigagem ordinária:

$$\sigma_{KO}^2 = \sum_{j=1}^{n} \lambda_j \gamma(x_0 - x_j) + \mu \quad (8)$$

em que σ_{KO}^2 é a variância de krigagem, λ_j os pesos obtidos, $\gamma(x_0 - x_j)$ o valor da semivariância para distância que separa x_0 *e* x_j, μ o multiplicador de Lagrange.

2.2. Análise de componentes principais

A análise de componentes principais (PCA) é uma técnica de análise multivariada que objetiva construir variáveis latentes, também chamadas de componentes principais, a partir de combinações lineares das *p* variáveis originais em estudo. Pode-se, a princípio, determinar *p* variáveis latentes, mas como será explanado abaixo, o sucesso da técnica consiste em determinar *k* componentes principais em que *k* < *p,* de modo que essas abarcam quantidade significativa da variabilidade contida nas *p* variáveis originais.

A técnica PCA foi desenvolvida por Pearson em 1901 e Hoteling em 1933 de forma independente. Seu objetivo central é reduzir a dimensionalidade dos dados, para estudar um grande número de variáveis originais que geralmente apresentam estrutura de correlação complexa, através de poucos componentes principais que não são correlacionadas entre si.

2.2.1. Determinação dos componentes a partir da matriz de correlações

Sejam *p* variáveis de um estudo denotadas por X_1, X_2, ..., X_p em que cada uma delas apresentam *n* observações. Para os casos nos quais as variáveis apresentam escalas de medidas diferentes Ferreira (2018) sugere aplicar a técnica de PCA utilizando a matriz de correlações lineares e dados padronizados, para que variáveis com grande variância não apresentem demasiada influência.

A correlação linear amostral entre duas variáveis pode ser estimada através da Equação 9.

$$r_{ij} = \frac{\widehat{COV}(X_i, X_j)}{\sqrt{\widehat{VAR}(X_i) \cdot \widehat{VAR}(X_j)}}. \quad (9)$$

em que $\widehat{COV}(X_i, X_j)$ é a covariância estimada entre X_i e X_j, e $\widehat{VAR}(X_i)$ e $\widehat{VAR}(X_j)$ é a variância estimada para X_i e X_j, respectivamente.

Após estimar as correlações entre as variáveis originais, os valores são distribuídos em linhas e colunas na matriz de correlações, denotada por R (Equação 10, exemplo para 3 variáveis). Pode-se observar que a matriz de correlações é quadrada e simétrica

$$\boldsymbol{R} = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{pmatrix}. \quad (10)$$

Deseja-se, a partir das *p* variáveis padronizadas Z_1, Z_2, ..., Z_p obter um novo conjunto de *p* combinações lineares (p componentes principais) da forma $Y_j = \sum_{i=1}^{p} a_{ij} Z_i$, para j de 1 a *p*. Essas variáveis latentes são determinadas a partir das condições expressas pelas Equações 11, 12, 13 e 14.

$$\sum_{i=1}^{p} a_{ij}^2 = 1, \quad (11)$$

$$Var(Y_1) \geq Var(Y_2) \geq \cdots \geq Var(Y_p), \quad (12)$$

$$Cov(Y_i, Y_j) = 0 \quad para\ i \neq j, \quad (13)$$

$$\sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(X_i). \quad (14)$$

Para satisfazer todas as condições, Ferreira (2018) indica o uso da técnica matemática dos multiplicadores de Lagrange, concluindo que os vetores contendo os pesos $\boldsymbol{a_i}$ que definem cada componente principal podem ser obtidos a partir da solução do sistema homogêneo indicado na Equação 15.

$$\left(\boldsymbol{R} - \lambda_j \boldsymbol{I}\right)\boldsymbol{a_j} = \boldsymbol{0}, \tag{15}$$

em que $\boldsymbol{R}$ é a matriz de correlação, λ_j uma constante que representa a variância da j-ésima variável original, e $\boldsymbol{I}$ matriz identidade.

Um sistema linear homogêneo como expresso na Equação 15 admitirá sempre a solução trivial. Decorre que não é interessante obter apenas essa solução, pois todos os coeficientes seriam iguais a zero. Para que o sistema tenha solução não-trivial, é necessário impor a condição de que o determinante da matriz associada aos coeficientes do sistema de equações resulte nulo, ou seja

$$det(\boldsymbol{R} - \lambda_i \boldsymbol{I}) = 0. \tag{16}$$

Assim, utilizando de conhecimentos de álgebra linear, nota-se que a Equação 16 é a equação característica que permite determinar os autovalores λ_i da matriz **R**. Posteriormente, os autovetores $\boldsymbol{a_j}$ da matriz de correlação $\boldsymbol{R}$ podem ser calculados.

Os autovetores da matriz $\boldsymbol{R}$ definem as combinações lineares e, portanto, os componentes principais. Para atender ao expresso pela Equação 12 deve-se definir a ordem dos componentes principais observando as magnitudes dos autovalores encontrados. Assim, o maior autovalor estará associado ao autovetor que define o primeiro componente. O autovalor obtido é a própria variância do respectivo componente principal, como expressa a Equação 17 (FERREIRA, 2018)

$$VAR(Y_i) = \lambda_i. \tag{17}$$

2.2.2. Proporção da variância total explicada por cada componente

Após determinar todos os *p* componentes principais, é preciso analisar a porcentagem da variância total explicada por cada um deles. Esses valores auxiliam na decisão de quantos componentes principais manter.

A variância total presente nas variáveis padronizadas, denotada por *t*, é dada pela soma da variância de cada variável individualmente, como expresso pela

Equação 18. No caso particular onde se optou pela padronização das variáveis, a variância total *t* coincidirá com o número de variáveis *p*

$$\sum_{i=1}^{p} VAR(Z_i) = t = p. \tag{18}$$

Com foi indicado na Equação 17, a variância de cada componente principal coincide com o autovalor correspondente. Dessa forma, pode-se calcular o percentual da variância total *t* que pode ser atribuída ao componente individualmente, denotado por E_i, através da Equação 19

$$E_i = \frac{\lambda_i}{t} \cdot 100. \tag{19}$$

Pode-se também determinar o percentual da variância total acumulada até certo componente *k*, denotado por E_k^A, utilizando a Equação 20

$$E_k^A = \frac{\sum_{i=1}^{k} \lambda_i}{t} \cdot 100. \tag{20}$$

2.2.3. Determinação do número de componentes a manter

Diversas metodologias são aplicadas objetivando determinar um valor de *k* de componentes principais que seja inferior a *p*, número de variáveis originais, mas que detenha parte significativa da informação original. Ferreira (2018) cita que quase todos os critérios adotados são empíricos e que apresentam algum componente subjetivo no processo decisório.

Uma primeira metodologia de determinação do valor *k* seria adotar um valor mínimo para o percentual da variância total a ser explicada pelos componentes principais. A subjetividade desse primeiro método advém da escolha do pesquisador acerca do percentual da variância total a ser usado. Em geral, observa-se na literatura, valores de 70% ou 80% (FERREIRA, 2018).

Uma segunda abordagem envolve a média aritmética ou geométrica dos autovalores, expressas pelas Equações 21 e 22. De acordo com esse critério, deve-

se eliminar todos os componentes principais cujos respectivos autovalores sejam menores do que a média aritmética ou menores do que a média geométrica.

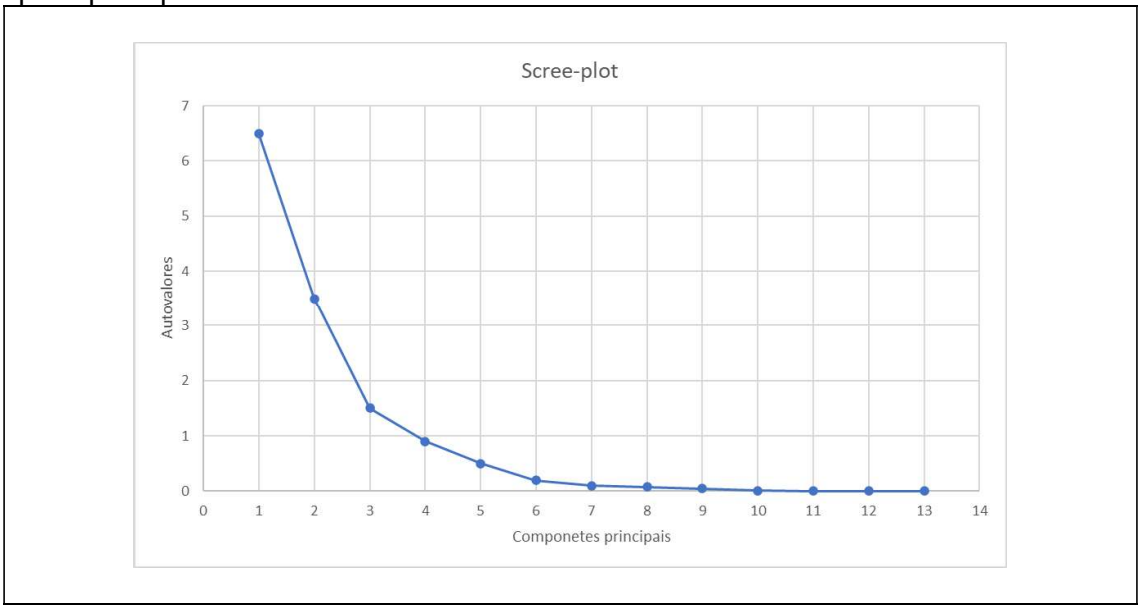
$$\bar{\lambda} = \frac{\sum_{i=1}^{p} \lambda_i}{p}, \tag{21}$$

$$\overline{\lambda_g} = \sqrt[p]{\prod_{i=1}^{p} \widehat{\lambda_i}}. \tag{22}$$

O uso do critério através da média geométrica é mais robusto, sendo essa menos sensível a valores discrepantes (FERREIRA, D., 2018).

Pode-se também aplicar um método gráfico conhecido como *scree-plot*. Trata-se de construir um par de eixos cartesianos adicionando ao eixo x valores inteiros 1, 2, ..., p, relacionados aos componentes principais, e no eixo y os autovalores correspondentes (Figura 3).

Figura 3 – Gráfico *scree-plot* para determinação do número de componentes principais que devem ser retidas.



Fonte: Baseado em FERREIRA, 2018, p. 341.

Com base no *scree-plot*, deve-se escolher aquele número de componentes para o qual o comportamento do gráfico muda repentinamente. Para o exemplo apresentando na Figura 3, deveria ser retido entre 3 e 5 componentes principais.

2.3. Interpolação pelo inverso da distância

Yamamoto e Landim (2015) e Pereira (2021) afirmam que, na ausência da dependência espacial, outros métodos de interpolação não estocásticos podem ser aplicados. Dentre eles, um dos mais usados é a estimação ponderada pelo inverso da distância (IDW) (MONTEIRO *et al.*, 2004).

No IDW, o valor em um ponto não amostrado é estimado, de forma determinística, através de uma média ponderada como expressa na Equação 23

$$\widehat{z_i} = \frac{\sum_{j=1}^{n} w_{ij} z_j}{\sum_{j=1}^{n} w_{ij}}. \tag{23}$$

em que $\widehat{z_i}$ é a estimativa para um local não amostrado, w_{ij} os pesos de estimação e z_j o j-ésimo valor amostrado.

Os pesos, denotados por w_{ij}, são determinados no IDW como o inverso da distância euclidiana entre i-ésimo ponto a ser estimado e aquele j-ésimo amostrado, elevada a uma potência *k*. A Equação 24 apresenta a forma do peso w_{ij}, enquanto a Equação 25 apresenta a expressão convencional para cálculo da distância euclidiana

$$w_{ij} = \frac{1}{d_{ij}^{k}}, \tag{24}$$

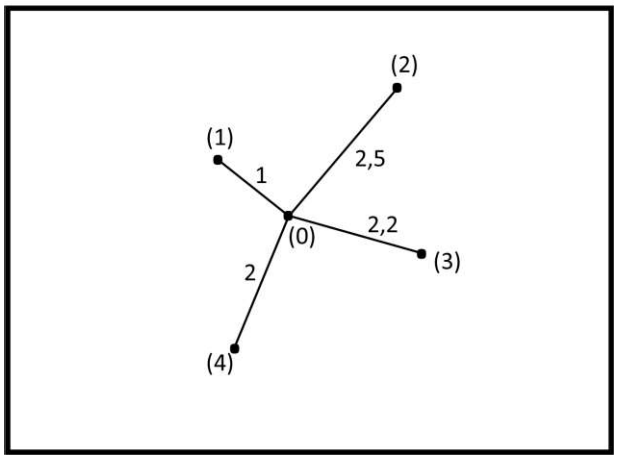
$$d_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}. \tag{25}$$

Considere como exemplo a situação apresentada na Figura 4. Existem 4 pontos amostrados e deseja-se determinar o valor no ponto central, identificado por 0, não amostrado, através da técnica IDW. As distâncias entre o ponto central e os amostrados é indicada na Figura 4. Para este exemplo, considere o expoente *k=2* (estimação pelo inverso do quadrado da distância). A Tabela 1 apresenta os valores amostrados, bem como valores para d_{ij} e w_{ij}.

Tabela 1 – Valores ilustrativos para estimação através do método IDW.

Ponto amostrado	Valor amostrado	Distância euclidiana (d_{ij})	Peso para k=2 (w_{ij})
1	3,2	1	1
2	2,8	2,5	0,16
3	2,2	2,2	0,21
4	3,1	2	0,25

Figura 4 – Exemplo de aplicação da técnica de estimação por IDW. O ponto central será estimado com base nos quatro pontos amostrados.



Fonte: elaborado pelo próprio autor.

É possível observar, através da Tabela 1, que o peso w_{ij} é inversamente proporcional a distância euclidiana. Assim, o ponto amostrado 1, separado por 1 unidade de comprimento do ponto 0, terá efeito maior, ao passo que o ponto 2, mais distante, terá efeito menor.

Aplicando a Equação 23 pode-se então estimar o valor da variável de interesse no ponto 0 como

$$\widehat{z_0} = \frac{3{,}2 \cdot 1 + 2{,}8 \cdot 0{,}16 + 2{,}2 \cdot 0{,}21 + 3{,}1 \cdot 0{,}25}{1 + 0{,}16 + 0{,}21 + 0{,}25} = 3{,}02.$$

Para efeito de comparação, se a estimação fosse baseada na média aritmética, desconsiderando a espacialização dos dados, o valor estimado seria de 2,83.

2.4. Algoritmo *k-means* para agrupamentos

As técnicas de agrupamento buscam, a partir de um conjunto de dados, determinar subconjuntos menores de tal modo que os elementos pertencentes aos subconjuntos sejam mais semelhantes entre si. Um dos algoritmos mais conhecidos e usados na prática é conhecido como *k-means* (KM) (James *et al.,* 2013).

O método KM é classificado como não-supervisionado, cabendo ao pesquisador escolher o número *k* de grupos (*clusters*) que deseja obter ao final. Neste método não há sobreposição dos grupos, ou seja, cada ponto pertencerá a apenas um dos *k* grupos. Além disso, não há resíduo, de tal modo que todos os pontos serão agrupados em um dos grupos resultantes. De forma matemática, essas características são expressas através das Equações 26 e 27, em que C_k denota o k-ésimo agrupamento, e (1, 2, ..., n) o conjunto total de pontos.

$$C_1 \cup C_2 \cup \ldots \cup C_k = \{1, 2, \ldots, n\} \tag{26}$$

$$C_k \cap C_{k'} = \emptyset \tag{27}$$

Segundo James *et al.* (2013), o princípio básico que orienta a determinação do melhor agrupamento possível é a minimização da variação dentro de cada grupo, denotada por $W(C_k)$. Assim, se cada grupo possuir o valor mínimo para variação interna, a soma da variação de todos os grupos também deverá ser mínima.

A variação dentro de cada grupo é calculada com base na soma dos quadrados das distâncias euclidianas entre os pares de pontos pertencentes ao grupo, dividida pelo número de pontos do grupo (Equação 28):

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2 \tag{28}$$

em que i e i' referem-se a pontos no interior do k-ésimo agrupamento e p ao número de variáveis consideradas.

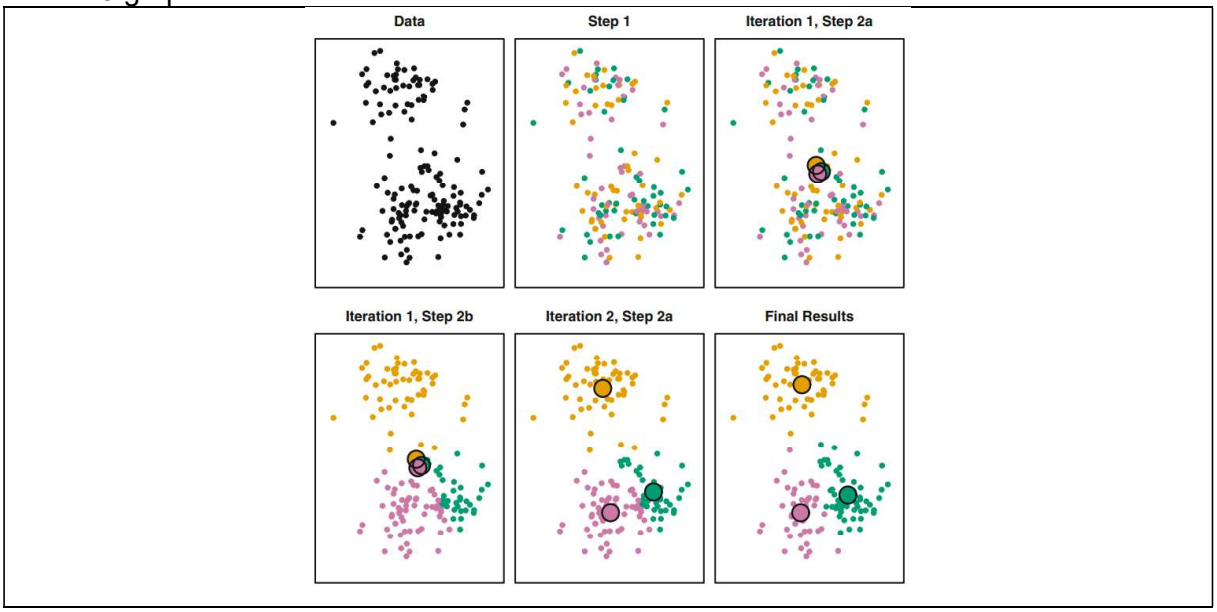
Assim, uma vez definida a variação dentro de cada grupo, deseja-se minimizar a Expressão 29, atribuindo cada uma das *n* observações dentro de um dos *k* grupos

$$mínimo\{\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}\left(x_{ij}-x_{i'j}\right)^2\}. \quad (29)$$

Uma forma de minimizar a Expressão 29, realizando a classificação das *n* observações em *k* grupos, é utilizando o algoritmo KM. O algoritmo consiste em 2 passos básicos.

Inicialmente, o conjunto das *n* observações é classificado aleatoriamente em *k* grupos. A partir deste ponto, são calculados *k* centroides, dados pela média dos pontos pertencentes a cada um dos grupos. Nesse momento é calculado a distância euclidiana entre os centroides e os pontos, reclassificando-os com base na proximidade com cada um dos centroides previamente obtidos. Repete-se então o cálculo dos centroides e a reclassificação até o momento em que a variação dentro dos grupos para de mudar (a Expressão 29 atinge uma mínimo local). A Figura 5 ilustra o funcionamento do algoritmo KM para k=3 grupos.

Figura 5 – Passos do algoritmo de classificação *k-means* para situação de agrupamento em k=3 grupos.



Data: os dados, para duas variáveis, são apresentados como antes da classificação; Step 1: os dados são classificados em 3 grupos de forma aleatória; Interation 1, Step 2a: os centroides (médias das variáveis) são calculados com base na classificação aleatória realizada no passo Step 1; Interation 1, Step 2b: os dados são reclassificados nos três grupos com base na distância até o centroide; Interation 2, Step 2a: os centroides são recalculados com base na classificação anterior; Final Results: os dados são apresentados na classificação final após repetidos processos de interação.

Fonte: James *et al., 2013, p. 389.*

Como o método KM apresenta o resultado de acordo com um valor mínimo local para a Expressão 29, James *et al.* (2013) recomendam que sejam feitas várias repetições do método, observando os valores de variação total obtidos em cada uma das repetições. Deve-se, ao final, escolher aquele agrupamento que apresenta menor valor para a variação total, ao final das repetições.

3. MATERIAIS E MÉTODOS

3.1. Coleta das amostras de solo

Foram utilizados dados de fertilidade de solo coletados em experimento conduzido no Centro de Pesquisa e Melhoramento da Cana-de-açúcar (CECA) da Universidade Federal de Viçosa. Os dados foram disponibilizados por FERREIRA (2020).

A área experimental fica localizada em Oratórios, município de Minas Gerais, e tem dimensões de 42 x 80 metros, resultando em uma área de 3.360 m^2.

De acordo com Yamamoto e Landim (2015, p. 21), "[...] a amostragem sistemática é, sem dúvida, a que fornece o melhor resultado". Assim, a amostragem na área experimental deu-se através da amostragem sistemática, em grade regular 4 x 9, totalizando 36 pontos amostrais. A densidade de pontos foi de aproximadamente 0,01 m^{-2}. Na literatura se encontram diversos valores para densidade de pontos amostrais. Adão *et al.* (2022) utilizaram uma densidade máxima de 0,03 m^{-2} enquanto Pasini *et al.* (2021) usou 0,003 m^{-2}.

A Figura 6 apresenta uma imagem da área experimental, bem como os locais onde as amostras de solo foram coletadas.

Figura 6 – Área experimental em Oratórios (MG). Os pontos amarelos indicam os locais onde as características químicas do solo foram analisadas (grade regular de 4 x 9).

Fonte: Adaptado de FERREIRA, 2020, p. 35.

A coleta das amostras de solo ocorreu no mês de outubro de 2019. Em cada local indicado na Figura 6, uma amostra de solo foi obtida na camada de 0-20 cm de profundidade. Essas amostras foram devidamente armazenadas e encaminhadas para um laboratório de análises de solos no município de Viçosa – MG, onde se procedeu a análise química.

A Tabela 2 indica as variáveis químicas de solo que foram consideradas no estudo, bem como suas respectivas unidades de medida.

Tabela 2 – Variáveis químicas de solo analisadas, suas siglas e unidades/escala de medida.

Variável	**Sigla**	**Unidade/escala de medida**
Potencial hidrogeniônico	pH	0-14
Fósforo	P	mg/dm^3
Potássio	K	mg/dm^3
Magnésio	Mg	$cmolc/dm^3$
Cálcio	Ca	$cmolc/dm^3$
Alumínio	Al	$cmolc/dm^3$
Acidez potencial	H+Al	$cmolc/dm^3$
Soma de bases trocáveis	SB	$cmolc/dm^3$
Capacidade de troca catiônica efetiva	CTC(t)	$cmolc/dm^3$
Capacidade de troca catiônica a pH 7	CTC(T)	$cmolc/dm^3$
Índice de saturação de alumínio	m	%
Índice de saturação de bases	V	%

3.2. Estatística descritiva dos dados

Com objetivo de obter aspectos básicos das variáveis analisadas foram calculadas diversas medidas de estatística descritiva. De acordo com Neto (2002), as medidas de posição e dispersão são importantes no entendimento da localização da distribuição das variáveis em estudo. Assim, foi calculada a média aritmética, mediana, desvio padrão, valor máximo e mínimo, coeficiente de variação, assimetria e curtose.

Além das medidas descritivas, foram confeccionados o histograma e *boxplot* para cada uma das variáveis analisadas, objetivando aumentar o entendimento dos

dados. Ademais, o *boxplot* foi utilizado para detecção e remoção de *outliers* conforme proposto em Devore (2006), pois segundo Isaaks e Strivastava (1989 *apud* GUIMARÃES, 2013, p. 18), *outliers* influenciam significativamente o variograma e a interpolação geoestatística.

3.3. Matriz de correlação

Para avaliar a correlação entre as variáveis em estudo, bem como o uso da metodologia de componentes principais (a ser descrito posteriormente), a matriz de correlações lineares foi obtida.

A matriz de correlação é um objeto matemático que indica se há relacionamento linear entre pares de variáveis. Cada célula da matriz é um valor numérico que indica a correlação entre as variáveis expressas na linha e na coluna. A correlação linear, calculada com base em valores amostrais, é obtida por meio da Equação 30

$$s_{XY} = \frac{COV(X,Y)}{\sqrt{V(X)V(Y)}}, \tag{30}$$

em que COV(X,Y) é a covariância entre as variáveis X e Y, e V(X) e V(Y) as variâncias.

Para obtenção da matriz de correlação neste trabalho foi utilizado o pacote de contribuição *GGally* (SCHLOERKE, et al., 2021) disponível para o *software* R, bem como o pacote *base*.

Para cálculo dos componentes principais foi utilizada a função *cor( ),* do pacote *base,* para obtenção da matriz de correlações. Já para gerar uma imagem da matriz de correlações, mais conveniente para exposição dos resultados, foi utilizado a função *ggcorr( )* do pacote *GGally*.

3.4. Análise geoestatística dos dados originais

Após o cálculo das medidas descritivas, confecção dos gráficos e obtenção da matriz de correlação, foram aplicadas técnicas da Geoestatística a fim de identificar e modelar a dependência espacial das variáveis do estudo.

Foi utilizado o pacote *base* do *software* R (R Development Core Team, 2022) juntamente com o pacote de contribuição chamado *geoR* versão 1.8.1, de autoria de Júnior, R. *et al.* (2020).

As variáveis foram analisadas de forma individual, onde buscou-se inicialmente observar a existência de correlação espacial através da construção do variograma experimental omnidirecional utilizando a função *variog( )* do *geoR*. Neste ponto do estudo foi possível observar quais variáveis não apresentaram dependência espacial, devido ao variograma revelar estrutura de efeito pepita puro.

No entanto, antes de obter o variograma experimental, as variáveis que apresentaram assimetria positiva foram submetidas à transformação logarítmica, como recomendado por Yamamoto e Landim (2015).

Na situação em que o variograma indicou dependência espacial, modelos teóricos conhecidos e descritos na literatura foram ajustados ao variograma. Para este estudo, os modelos ajustados foram o gaussiano, o exponencial e o esférico, utilizando da função *variofit( )* presente no pacote *geoR*, utilizando a metodologia dos mínimos quadrados ordinários (OLS) ou mínimos quadrados ponderados (WLS) (CRESSIE, 1985).

Para avaliar a qualidade do ajuste foi utilizada a técnica de validação-cruzada, ou "jack-knife". Foi usada a função *xvalid( )* do pacote *geoR* para executar a validação. Segundo Isaaks e Strivastava (1989 *apud* SANTOS, 2016), deve-se selecionar o modelo ajustado que apresenta os resultados mais satisfatórios para a validação. Assim, os seguintes aspectos da validação-cruzada foram observados e utilizados para avaliar a qualidade do ajuste (GUIMARÃES, 2013; HERNÁNDEZ, 2021):

- Coeficiente angular da regressão entre valores estimados e observados deve ser igual ou próximo de 1.
- Média do erro da estimação deve ser próxima de zero.
- Média do erro padronizado deve ser próxima de zero.
- Variância do erro de estimação padronizado deve ser próxima de 1.

Após o ajuste do modelo, foi calculado o Índice de Dependência Espacial (IDE) proposto por Zimback (2001 apud. SOUZA et al., 2008) a fim de determinar o grau da intensidade da dependência espacial. O IDE é calculado como expresso na Equação

$$IDE = \frac{C_1}{C_1 + C_0}, \tag{31}$$

em que C_1 é a contribuição e C_0 o efeito pepita, parâmetros obtidos no ajuste do modelo ao variograma experimental.

Classifica-se a correlação espacial através do IDE como expresso na Tabela 3.

Tabela 3 – Classificação IDE (Índice de dependência espacial).

Classificação do IDE	**Valor do IDE**
Fraca	IDE < 0,25
Moderada	0,25 ≤ IDE < 0,75
Forte	IDE ≥ 0,75

3.5. Krigagem

De posse dos dados para características químicas do solo (36 valores amostrados para cada atributo) e do modelo de variograma ajustado, deu-se início a etapa de interpolação via krigagem. Para isso foi definido uma malha regular de 50 por 50 pontos com auxílio da função *expand.grid( )* presente no pacote base do R. Cada ponto pertencente a esta malha formou um par ordenado de longitude e latitude que serviria de base para a futura interpolação via krigagem. O uso da grade regular está de acordo com o exposto por Yamamoto e Landim (2015, p. 21).

Para executar a interpolação via krigagem foi utilizado a função *krig.conv( )* presente no *geoR*. Essa função recebe como parâmetros a variável original georreferenciada (foi analisado uma variável por vez), a malha regular indicando o local onde a interpolação deve ser executada e, por último, o modelo teórico do variograma previamente ajustado para a variável em questão.

O tipo de krigagem escolhida foi a ordinária, método mais usual segundo Yamamoto e Landim (2015). Durante a interpolação, a função *krig.conv( )* usou todos os pontos vizinhos nos cálculos das estimativas, não sendo utilizada uma borda ou então um número máximo de vizinhos para delimitar a krigagem.

3.6. Produção dos mapas

De posse dos dados interpolados, foi usado a função *contour( )* do pacote base do R para gerar as imagens da variabilidade espacial da variável em questão.

Ademais, empregou-se a paleta de cores conhecida como *terrain.colors* que apresentou um visual mais interessante ajudando a visualizar os resultados.

Adicionalmente foi adicionado um retângulo em cada imagem gerada que distingue e indica a região de estudo. Isso se fez necessário porque a malha regular usada na interpolação dos locais não amostrados extrapolava a área de estudo.

3.7. Obtenção dos componentes principais

Para aplicar a técnica de PCA, utilizando os dados estimados para cada variável química de solo (e não apenas os 36 valores amostrados), foi necessário inicialmente manipular a planilha dos dados obtidos via krigagem de modo a selecionar apenas os valores que pertenciam a área de estudo. Para isso foi desenvolvido um código no R que foi capaz de realizar interseção entre a malha regular da interpolação e a área em estudo, resultando em 729 pontos de coordenadas pertencentes a região amostrada.

Esses 729 pontos estimados foram analisados no software R, onde foi obtido a matriz de correlações utilizando a função *cor( )* do pacote *base*. Optou-se por usar a matriz de correlações ao invés da matriz de covariância para se evitar problemas de escala, como foi sugerido em Ferreira (2018).

Utilizando a função *eigen( )* do pacote *base* do R, foram obtidos os autovalores e autovetores da matriz de correlação. Os autovalores, como já foi exposto anteriormente, estão associados à proporção da variância atribuída ao respectivo componente principal e os autovetores aos pesos que definem a combinação linear entre as variáveis.

Para o cálculo dos *scores* em cada componente foi utilizado o *software* R. Em uma planilha foram adicionadas as coordenadas geográficas, bem como os respectivos valores numéricos dos *scores*, calculados a partir dos autovetores obtidos.

Coube ainda definir quantos componentes principais seriam usados já que o emprego da análise de componentes principais objetivou reduzir a dimensionalidade dos dados. De acordo com Jolliffe (2002 *apud* FERREIRA, 2018, p. 426), geralmente é escolhido o número de componentes que está associado a, pelo menos, 70% da variância total. Para este trabalho, foram escolhidos os *k* primeiros componentes referentes a 80% da variância total acumulada.

3.8. Agrupamento dos *scores*

De posse dos *scores* dos primeiros componentes principais, o agrupamento dos dados foi realizado no software R usando a função *kmeans( )* do pacote *base.*

A função *kmeans( )* recebeu três parâmetros para realizar os agrupamentos: *data, centers* e *nstart*.

O parâmetro *data* corresponde ao conjunto de dados no qual será realizada a classificação. Foi passado a função o conjunto de *scores* calculados a partir da análise de PCA que explicaram parte significativa da variância.

O parâmetro *centers* define o número de agrupamentos a serem determinados pelo algoritmo, equivalente ao número de blocos experimentais a serem obtidos. Para este estudo, foram usados os valores 2, 3, 4 e 5.

Finalmente, o parâmetro *nstart* está relacionado ao número de vezes que o algoritmo sorteia, na fase inicial, agrupamentos aleatórios. James *et al.* (2013) recomenda que seja escolhido para esse parâmetro valores de 20 a 50, de modo que o algoritmo selecione ao final aquele com melhores resultados.

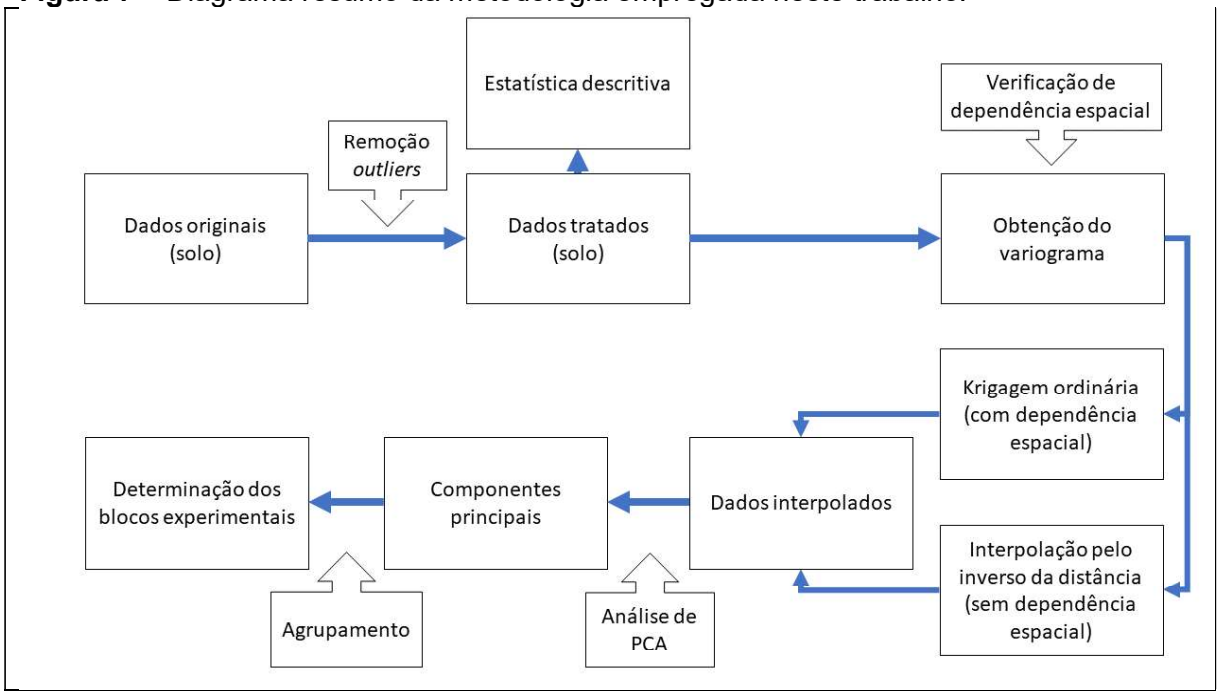
O resultado da função *kmeans( )* é um vetor com os números inteiros que indicam os grupos aos quais as observações pertencem. Esse vetor foi unido ao *data.frame* que continha a localização (longitude e latitude) de cada observação para que fosse possível construir o mapa da área experimental com os agrupamentos indicados.

Para realizar a construção do mapa, foi utilizado a função *grid.image( )* do pacote *phylin*.

3.9. Resumo da metodologia

A metodologia utilizada neste trabalho está resumida na Figura 7.

Figura 7 – Diagrama resumo da metodologia empregada neste trabalho.



Fonte: elaborado pelo próprio autor.

4. RESULTADOS E DISCUSSÃO

4.1. Análise descritiva das variáveis em estudo

4.1.1. Medidas de posição e dispersão

Para compreender aspectos básicos das variáveis químicas do solo foi construída a Tabela 4, que apresenta as estimativas para média aritmética, mediana, desvio-padrão, valor máximo, valor mínimo, coeficiente de variação, assimetria, curtose e p-valor para o teste de Shapiro-Wilk (5%) para todas as variáveis. Todas as medidas estatísticas apresentadas na tabela foram estimadas usando todos os 36 valores observados para as variáveis de solo em estudo.

Observou-se a partir da média aritmética e dos valores máximo e mínimo, que os valores observados se encontram em diferentes intervalos do eixo real. Há presença de valores de baixa e alta magnitude (0,07 a 79,50). Essa comparação, ignorando as diferentes unidades de medida e a natureza das variáveis é importante, pois à medida que o efeito da escala das variáveis aumenta, pode haver predominância indevida de certas variáveis no cálculo dos componentes principais (FERREIRA, 2018).

O coeficiente de variação (CV), apresentado na Tabela 4, é indicado para comparar a variação entre dois conjuntos de dados que apresentam médias e unidades de medidas distintas (NETO, 2002). De acordo com Gomes (1990), classifica-se o CV como baixo quando inferior a 10%, médio quando o valor se encontra entre 10 e 20%, alto para valores entre 20 e 30% e muito altos para valores de CV maiores do que 30%.

Observou-se que apenas as variáveis pH e CTC(T) apresentaram CV baixo. As variáveis H+Al e CTC(t) apresentam CV classificado como médio. As demais variáveis apresentaram CV considerados muito alto, o que revela uma grande variação dos valores observados no interior da área experimental. As cinco variáveis que apresentaram maior CV foram, em ordem decrescente, cálcio (Ca), magnésio (Mg), índice de saturação de bases (V), soma de bases trocáveis (SB) e alumínio (Al). Os valores de CV calculados para essas variáveis (variando de 37,14% até 95,78%) indicaram grande desuniformidade da área experimental o que, por sua vez, pode contribuir para elevar o erro experimental e diminuir a precisão de um experimento.

Tabela 4 – Análise descritiva para as variáveis químicas de solo obtidas a partir dos 36 pontos amostrais na área de estudos.

Atributo	Média aritmética	Mediana	Desvio-padrão	Valor máximo	Valor mínimo	Coeficiente de variação (%)	Assimetria	Curtose	p-valor
pH	4,48	4,45	0,28	5,60	4,10	6,19	2,00	6,50	0,087
P (mg/dm^3)	11,15	10,33	3,40	19,97	6,54	30,47	0,79	0,00	0,030*
K (mg/dm^3)	43,61	40,00	15,99	92,00	23,00	36,66	1,13	1,21	0,049*
Ca (cmolc/dm^3)	0,85	0,72	0,82	5,00	0,22	95,78	4,15	20,05	0,400
Mg (cmolc/dm^3)	0,27	0,27	0,13	0,81	0,07	49,55	1,93	6,74	0,590
Al (cmolc/dm^3)	1,04	1,00	0,38	1,90	0,20	37,14	-0,05	0,07	0,600
H+Al (cmolc/dm^3)	6,01	5,94	0,81	8,25	3,63	13,54	-0,22	1,84	0,012*
SB (cmolc/dm^3)	1,11	1,05	0,54	3,40	0,36	48,94	2,26	8,44	0,479
CTC(t) (cmolc/dm^3)	2,14	2,08	0,34	3,60	1,76	15,90	2,71	9,53	0,668
CTC(T) (cmolc/dm^3)	7,12	7,25	0,65	8,97	5,44	9,16	0,04	1,11	0,184
V (%)	15,63	14,70	7,67	48,40	5,40	49,07	2,30	8,64	0,710
m (%)	49,43	49,15	18,03	79,50	5,60	36,47	-0,46	-0,10	0,567

pH: potencial hidrogeniônico, P: fósforo, K: Potássio, Ca: cálcio, Mg: manganês, Al: alumínio, H+Al: acidez potencial, SB: soma de bases, CTC(t): capacidade de troca catiônica efetiva, CTC(T): capacidade de troca catiônica a pH 7, V: índice de saturação de bases, m: índice de saturação de alumínio.

*Significativo para teste de normalidade Shapiro-Wilk ao nível de significância de 5%.

Em relação a normalidade das variáveis, foi observado que apenas três não apresentaram distribuição normal: H+Al, P e K. Além disso, o fato de P e K apresentarem assimetria positiva indicou a necessidade de realizar uma transformação de variáveis para posterior análise variográfica.

4.1.2. Histogramas e *boxplot*

Com intuito de se observar melhor a distribuição das variáveis foi construído um histograma e um *boxplot* para cada atributo de solo avaliado. Além de indicar aspectos relevantes das distribuições de frequências, o *boxplot* foi utilizado para avaliar a presença de outliers. As Figuras A-1, A-2, A-3 e A-4 com os resultados foram colocados no Apêndice A.

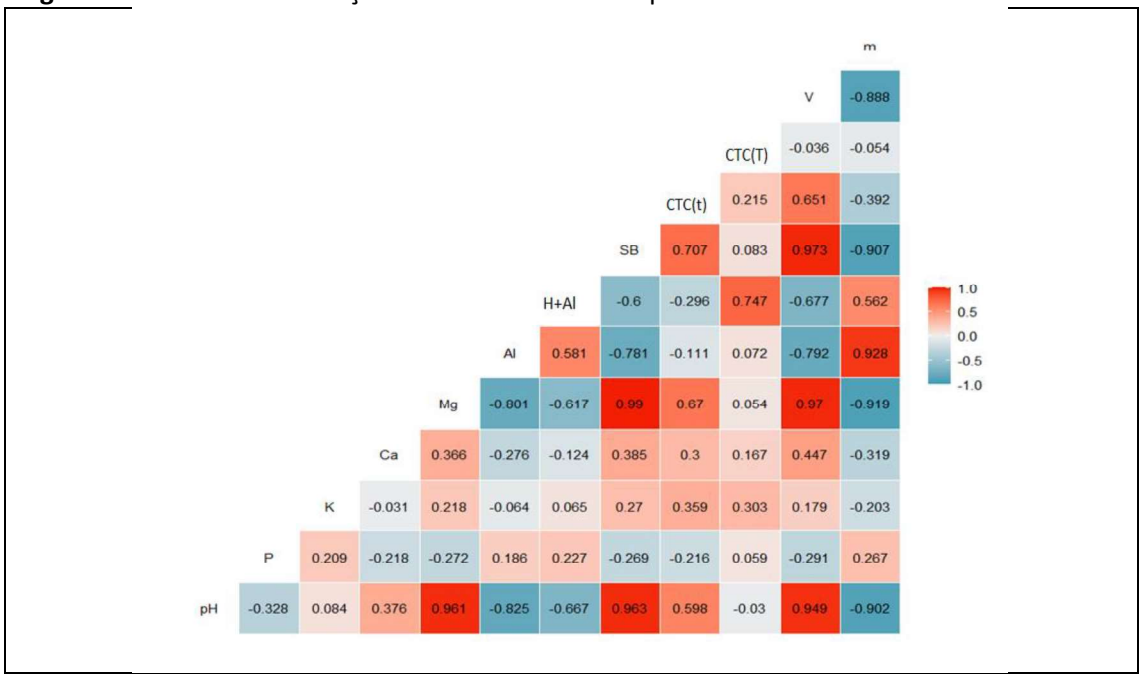
Foi observado no gráfico *boxplot* a presença de possíveis outliers em 8 das 12 variáveis em estudo: pH, K, Ca, Mg, H+Al, SB, CTC(t) e V. Esses valores foram retirados do conjunto de dados antes das análises para não comprometerem a obtenção do variograma.

4.1.3. Matriz de correlação

Buscando entender a relação entre as variáveis em estudo foi construída a matriz de correlação linear de Pearson para todos os 12 atributos químicos do solo (Figura 8).

De acordo com Cohen (1977 *apud.* HERNÁNDEZ, 2021) o coeficiente de correlação pode ser considerado baixo quando seu valor em módulo está entre 0,10 e 0,29; médio entre 0,30 e 0,49; e alto para valores maiores do que 0,50.

Figura 8 – Matriz de correlação amostral de Pearson para as variáveis em estudo.



4.2. Interpolação espacial das variáveis

4.2.1. Variogramas e modelos ajustados

Do conjunto inicial de 12 atributos do solo avaliados apenas 2 não apresentaram dependência espacial: H+Al e CTC(T). Essa conclusão se deu pela observação do variograma experimental, que revelou comportamento de efeito pepita puro. As variáveis pH, P, K, Ca, Mg, Al, SB, CTC(t), V e m apresentaram dependência espacial, uma vez que o variograma se apresentou crescente com a distância, chegando a um patamar a partir do valor do alcance. As Figuras B-1 e B-2 do Apêndice B mostram os variogramas experimentais obtidos, bem como a curva que representa o modelo teórico ajustado.

O modelo esférico apresentou melhor ajuste para as variáveis pH, P, Al, K, Mg, CTC(t) e m. Enquanto o modelo exponencial foi escolhido como melhor para três variáveis: Ca, SB e V. A Tabela 5 mostra os parâmetros da validação-cruzada que levaram a escolha dos respectivos modelos teóricos.

Tabela 5 – Parâmetros obtidos para os melhores valores da validação cruzada.

Variável	$\widehat{\mu}$	$\widehat{\mu_Z}$	$\widehat{S_Z^2}$	$\widehat{\beta_1}$
pH	<0,001	0,001	0,950	1,03
P	0,042	0,006	1,000	0,98
K	-0,002	<0,001	0,990	1,09
Ca	0,001	0,001	0,942	1,00
Mg	<0,001	-0,001	0,913	0,94
Al	-0,001	-0,002	1,020	0,94
SB	<0,001	<0,001	0,965	0,95
SB_t	0,002	0,007	1,060	1,02
V	-0,005	-0,001	0,942	0,96
m	-0,070	-0,002	0,950	0,98

pH: potencial hidrogeniônico, P: fósforo, K: Potássio, Ca: cálcio, Mg: manganês, Al: alumínio, SB: soma de bases, CTC(t): capacidade de troca catiônica efetiva, V: índice de saturação de bases, m: índice de saturação de alumínio, $\widehat{\mu}$: média do erro de estimação, $\widehat{\mu_Z}$: média do erro de estimação padronizado, $\widehat{S_Z^2}$: variância do erro de estimação padronizado, $\widehat{\beta_1}$: inclinação da reta de regressão linear.

Observa-se pela Tabela 5 que os valores dos parâmetros da validação-cruzada foram adequados. A média do erro de estimação é muito próxima de zero para todos os casos. Isso indica que não ocorreram valores estimados muito acima ou muito abaixo dos valores amostrados. Além disso, a média do erro padronizado e a variância do erro padronizado ficaram próximos de 0 e 1, respectivamente.

Por fim, o fato de a inclinação da reta de regressão linear entre os valores amostrados e estimados ser próxima de 1 indicou que os valores preditos são próximos aos observados, o que dá maior credibilidade ao modelo teórico ajustado.

Para sintetizar os parâmetros dos ajustes dos modelos teóricos foi construída a Tabela 6.

Tabela 6 - Parâmetros geoestatísticos estimados e respectivos métodos de estimação.

Variável	C_0	C_1	Patamar	Alcance	Modelo	Método
pH	0,029	0,012	0,045	56,00	Esférico	WLS
P	5,270	6,823	12,094	38,85	Esférico	WLS
K	91,340	100,690	192,032	21,36	Esférico	WLS
Ca	0,037	0,081	0,121	60,01	Exponencial	WLS
Mg	0,005	0,006	0,011	61,02	Esférico	OLS
Al	0,120	0,042	0,160	93,00	Esférico	WLS
SB	0,079	0,150	0,231	67,93	Exponencial	OLS
CTC(t)	0,002	0,029	0,031	50,52	Esférico	WLS
V	<0,001	29,062	29,062	9,70	Exponencial	OLS
m	186,831	255,980	442,816	114,87	Esférico	OLS

pH: potencial hidrogeniônico, P: fósforo, K: Potássio, Ca: cálcio, Mg: manganês, Al: alumínio, SB: soma de bases, CTC(t): capacidade de troca catiônica efetiva, V: índice de saturação de bases, m: índice de saturação de alumínio, C_0: efeito pepita, C_1: contribuição.

De acordo com o exposto na Tabela 6, quase a totalidade das variáveis apresentaram efeito pepita, com exceção da variável V, que apresentou valor nulo até a terceira casa decimal.

O alcance estimado para a dependência espacial foi, na maioria das vezes, abaixo de 100 metros, com média de 57,28 metros. Este valor equivale a 71,6% da maior dimensão (80 m) da área experimental.

Analisando o índice de dependência espacial (IDE), nota-se que 80% dos atributos do solo apresentam dependência espacial moderada. O restante das variáveis apresenta dependência forte: CTC(t) e V. Essa constatação está relacionada ao fato de que o efeito pepita obtido para CTC(t) e V são pequenos quando comparados com o valor da contribuição C_1 para cada variável.

Tabela 7 – Índice de dependência espacial (IDE) obtidos para os atributos de solo.

Variável	IDE(%)	Classificação
pH	34,83	Moderada
P	56,41	Moderada
K	52,44	Moderada
Ca	68,64	Moderada
Mg	57,41	Moderada
Al	25,00	Moderada
SB	65,50	Moderada
CTC(t)	95,08	Forte
V	100,00	Forte
m	57,81	Moderada

pH: potencial hidrogeniônico, P: fósforo, K: Potássio, Ca: cálcio, Mg: manganês, Al: alumínio, SB: soma de bases, CTC(t): capacidade de troca catiônica efetiva, V: índice de saturação de bases, m: índice de saturação de alumínio,

4.2.2. Mapas obtidos via krigagem ordinária

Após a caracterização da dependência espacial através do ajuste dos modelos de variograma, a interpolação usando krigagem ordinária foi procedida no *software* R. As Imagens C-1 e C-2 do Apêndice C apresentam os resultados obtidos.

Nota-se uma semelhança muito grande entre os mapas de variabilidade de potencial hidrogeniônico (pH), magnésio (Mg), cálcio (Ca) e soma de bases trocáveis (SB), apresentados na Figura C-1. Todos eles apresentaram valores mínimos no segundo quadrante dos mapas.

Observa-se também semelhança entre os mapas de alumínio (Al) e índice de saturação de alumínio (m), ambos atributos ajustados com modelo esférico (Figura C-1). Nota-se que os valores mais extremos (positivos) se encontram nos segundos quadrantes dos mapas e os valores menores na parte inferior central.

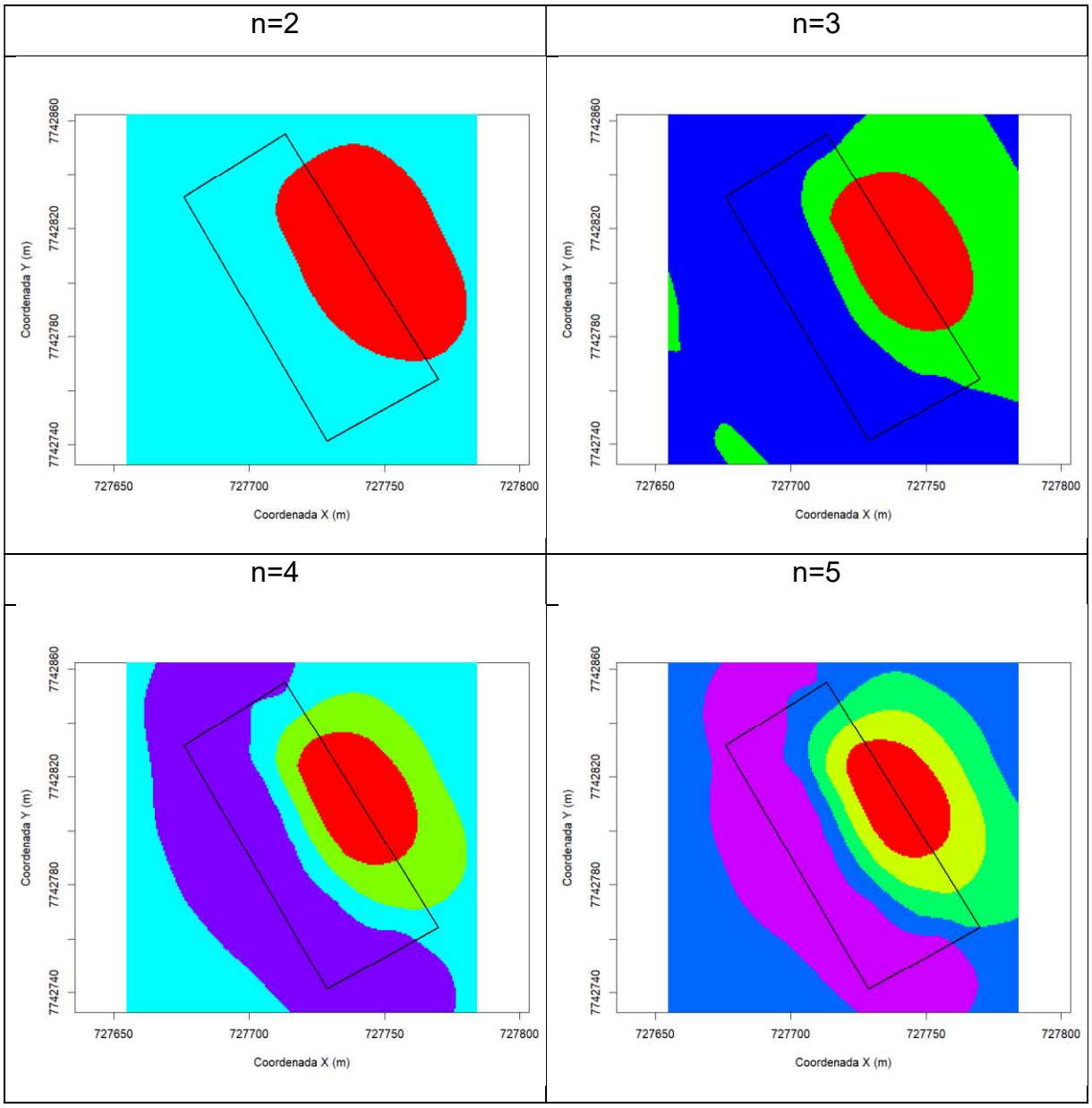
Os mapas de capacidade de troca catiônica efetiva e índice de saturação de bases se apresentam semelhantes entre si (Figura C-2) e com algumas semelhanças aos mapas de pH, Mg, Ca e SB.

O único mapa que foi bem diferente dos demais foi do atributo potássio, apresentado na Figura C-2. Essa variável se distribuiu de forma irregular pela área de

estudo, com a maioria das regiões variando de valores mínimos a valores intermediários.

A partir destes mapas de variabilidade, já pode-se pensar na determinação dos blocos experimentais caso o interesse do pesquisador fosse controlar apenas um único atributo do solo. Por exemplo, para a variável pH, o resultado está apresentado na Figura 9.

Figura 9 – Blocos experimentais sugeridos para controle local do atributo pH para 2, 3, 4 e 5 blocos.



Observa-se na Figura 8 que os blocos experimentais para o atributo pH apresentaram, no geral, formato não regular, com áreas diferentes, o que pode limitar

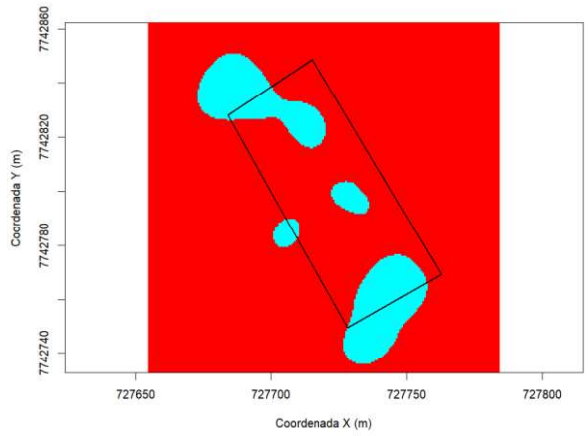
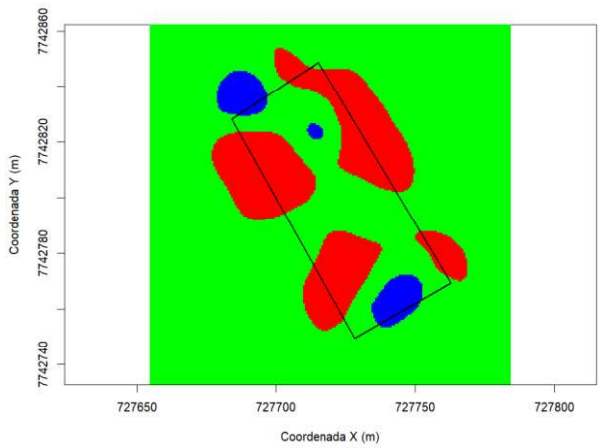
o número de tratamentos a ser avaliado, bem como o tamanho das unidades experimentais.

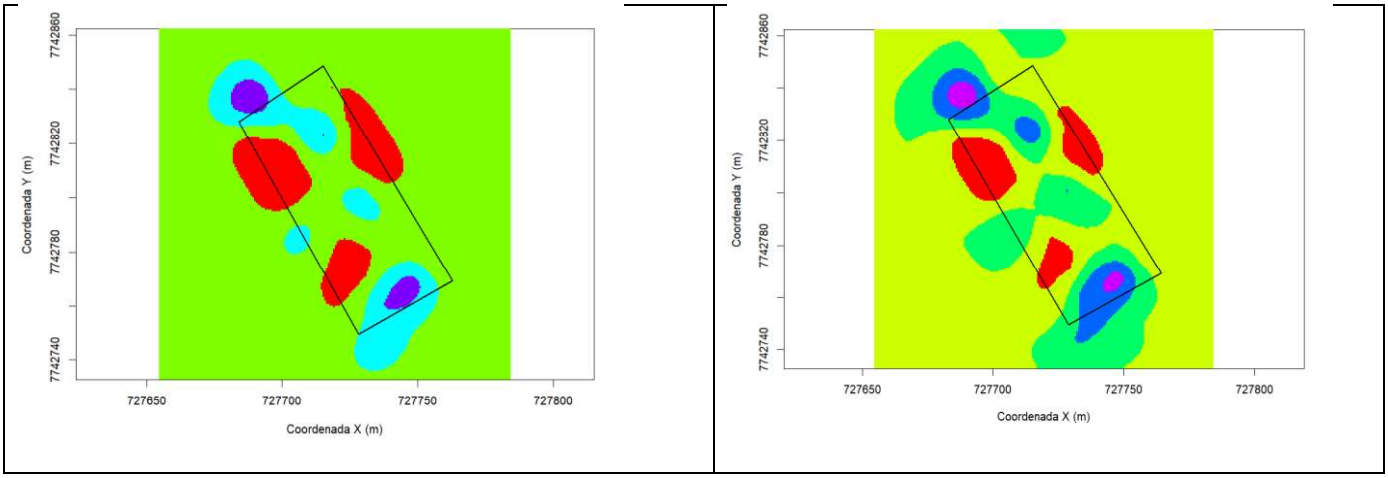
Situações mais complexas podem ocorrer, como para o atributo potássio. Caso o pesquisador tivesse interesse de delimitar os blocos em função desse único atributo, os resultados seriam os indicados na Figura 10.

Os resultados indicados na Figura 10 mostram que os blocos para potássio apresentam pequenas áreas e estão espalhados pela área experimental (estão desconectados). As unidades experimentais que formam cada bloco não estão em uma área contínua, o que pode gerar dificuldades na instalação do experimento.

Vale destacar que estes mapas de variabilidade também podem ser utilizados para orientar a correção de fertilidade para cada atributo, considerando a necessidade específica de cada área. Efetuando tal procedimento aumenta-se a uniformidade da área experimental.

Figura 10 – Blocos experimentais sugeridos para controle local do atributo K para 2, 3, 4 e 5 blocos.

n=2	n=3
	
n=4	n=5



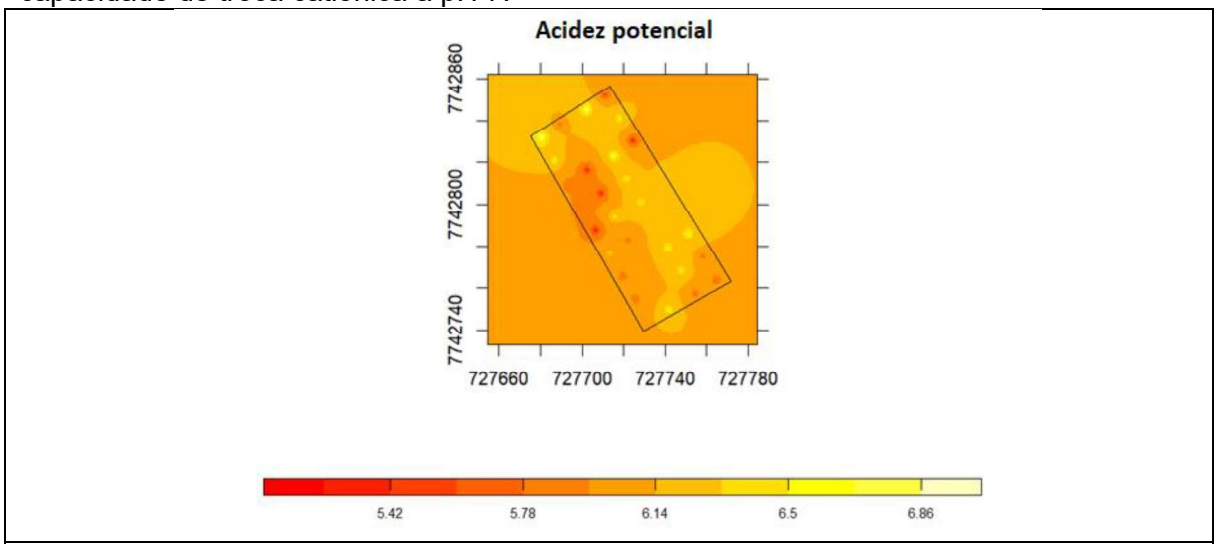
4.2.3. Mapas obtidos via IDW

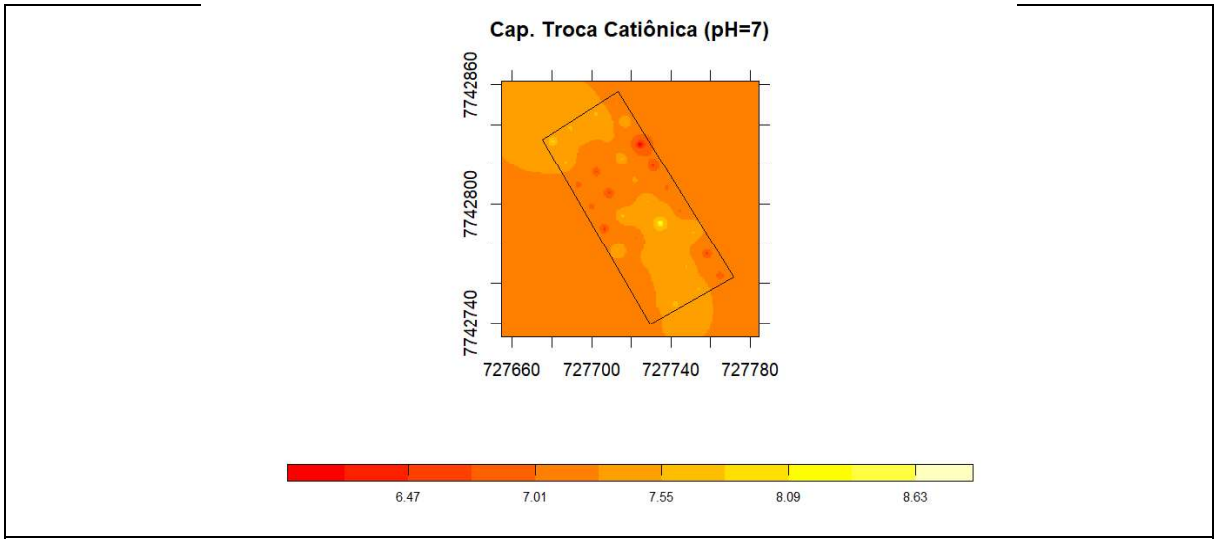
Para as variáveis em que não foi possível identificar dependência espacial foi utilizado o interpolador ponderado pelo inverso da distância (IDW). Os resultados obtidos estão apresentados na Figura 11.

Nota-se que o mapa para acidez potencial (H+Al) apresentou certa semelhança aos mapas obtidos via krigagem ordinário que evidenciaram valores máximos no segundo quadrante. Por outro lado, o mapa para capacidade de troca catiônica a pH 7 (CTC(T)) não apresentou semelhanças com os mapas obtidos via krigagem.

Durante o processo de estimação, a raiz do erro quadrático médio foi calculada como 0,665 para CTC(T) e 0,619 para H+Al.

Figura 11 – Mapas estimados via IDW para os atributos de solo acidez potência e capacidade de troca catiônica a pH 7.





4.3. Análise de componentes principais (PCA)

Utilizando o *software* R foram obtidos os 12 autovalores e autovetores da matriz de correlação linear para as variáveis em estudo. Calculou-se assim o percentual da variância explicada por cada variável latente, bem como o percentual de variância explicada acumulada. Os resultados obtidos estão expressos na Tabela 8.

Tabela 8 – Resumo dos resultados obtidos por meio da técnica de análise de componentes principais.

Componente	**Autovalor**	**% da variância**	**% Acumulado**
1	7,704	64,21	64,21
2	2,167	18,06	82,27
3	0,696	5,80	88,07
4	0,584	4,87	92,93
5	0,322	2,68	95,62
6	0,236	1,97	97,58
7	0,19	1,58	99,17
8	0,071	0,59	99,76
9	0,0183	0,15	99,91
10	0,0056	0,05	99,96
11	0,0038	0,03	99,99
12	0,0013	0,01	100,00
Total	**11,999**		
$\bar{\lambda}$	**0,999**		

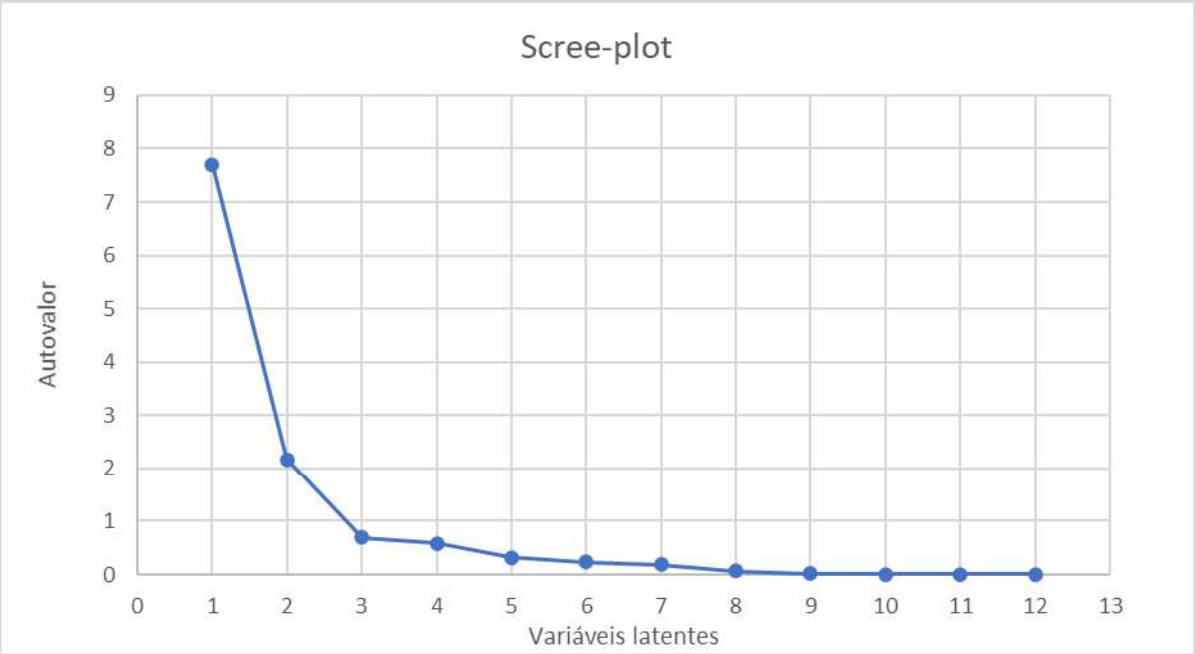
$\bar{\lambda}$: média aritmética dos autovalores.

Foi possível observar que as primeiras 4 variáveis latentes condensam mais de 90% da variância total dos dados. Como indicado na revisão de literatura, a maioria

dos trabalhos acadêmicos retém como componentes principais aquele conjunto de variáveis latentes que explicam de 70% a 90% da variância total. Isso levaria a escolher 2, 3 ou os 4 primeiros componentes.

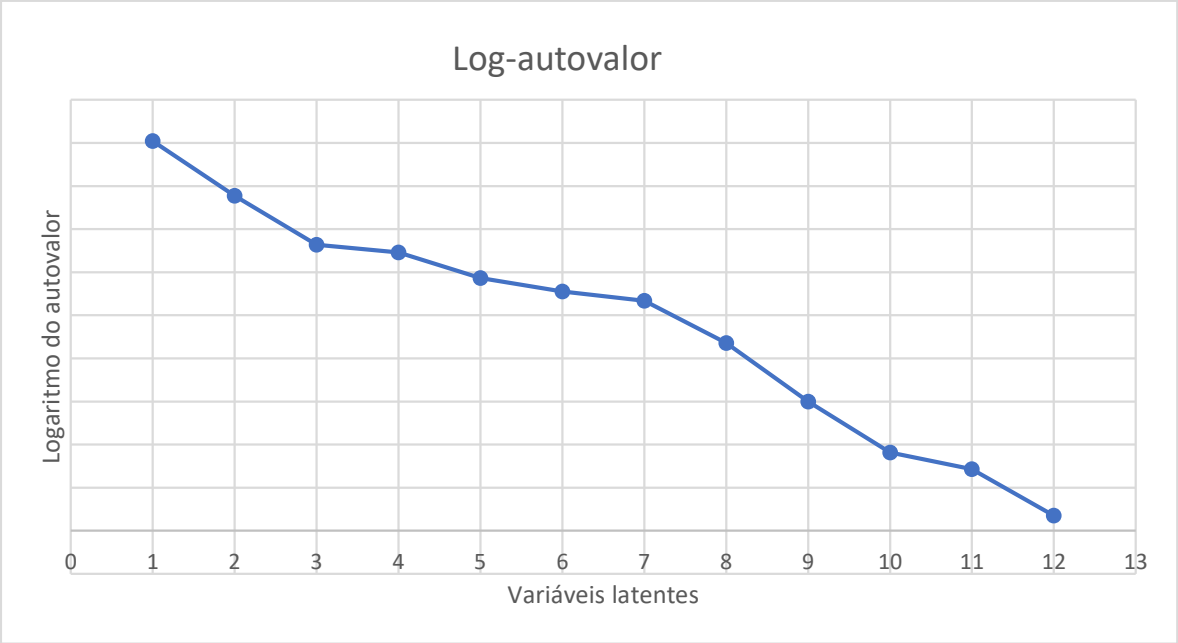
A Figura 12 apresenta o gráfico de *scree-plot* para os autovalores obtidos. Nota-se que, a partir da variável latente número 3, os autovalores e, portanto, a variância associada, apresentam variações bem pequenas. Assim, analisando o gráfico, poderia optar por manter até três componentes principais.

Figura 12 – *Scree-plot* para os autovalores obtidos.



Ferreira (2018) sugere, de forma adicional, construir um gráfico do logaritmo natural do autovalor em função da variável latente (Figura 13). Dessa forma, pode-se observar o comportamento do gráfico, selecionando o número de variáveis com base no local onde o comportamento deixa de ser linear.

Figura 13 – Logaritmo do autovalor em função da variável latente associada.



O gráfico do logaritmo do autovalor apresentou comportamento linear até a terceira variável latente. Isso sugeriu escolher as três primeiras variáveis como componentes principais.

O critério final de escolha adotado foi fixar, de maneira subjetiva, um valor mínimo de explicação de 80%. Da Tabela 8 pôde-se concluir que as variáveis latentes 1 e 2 explicaram parte substancial da variância total (82,27%), o que levou a selecionar as variáveis 1 e 2 como componentes principais.

Como foi apresentado na revisão de literatura, o componente principal é combinação linear de todas as variáveis originais. A Tabela 9 apresenta o peso que foi atribuído a cada uma das variáveis para os componentes selecionados.

Tabela 9 – Pesos que definiram as combinações lineares dos componentes principais 1 e 2.

Pesos	**PC 1**	**PC 2**
m	0,35343659	-0,010573182
V	-0,32441648	0,122592180
CTC(t)	-0,27322821	-0,256562339
SB	-0,35739228	-0,013984200
Al	0,33040853	-0,015219899
Mg	-0,35734480	0,017984528
Ca	-0,34686266	-0,004908099
K	-0,09265879	-0,480409285
P	0,23163067	-0,035920368
pH	-0,35302838	0,054797798
H+Al	0,13892653	-0,559218428
CTC(T)	-0,06408422	-0,608645210

pH: potencial hidrogeniônico, P: fósforo, K: Potássio, Ca: cálcio, Mg: manganês, Al: alumínio, H+Al: acidez potencial, SB: soma de bases, CTC(t): capacidade de troca catiônica efetiva, CTC(T): capacidade de troca catiônica a pH 7, V: índice de saturação de bases, m: índice de saturação de alumínio, PCA 1: componente principal 1, PCA 2: componente principal 2.

A Tabela 10 apresenta a correlação linear entre cada componente principal e as variáveis originais. Observa-se que as maiores correlações positivas com PC 1 se estabeleceram com m e Al. Já as correlações negativas com PC 1 se destacam as variáveis SB, Mg, pH e Ca. Considerando o PC 2, não foi observado correlação positiva forte. Considerando as correlações negativas, as maiores se estabeleceram com CTC(T) e H+Al.

Tabela 10 – Correlações entre os componentes principais 1 e 2 selecionados e as variáveis de solo em análise.

Atributos de solo	PCA 1	PCA 2
m	0,98	-0,02
V	-0,90	0,18
CTC(t)	-0,76	-0,38
SB	-0,99	-0,02
Al	0,92	-0,02
Mg	-0,99	0,03
Ca	-0,96	-0,01
K	-0,26	-0,71
P	0,64	-0,05
pH	-0,98	0,08
H+Al	0,39	-0,82
CTC(T)	-0,18	-0,90

pH: potencial hidrogeniônico, P: fósforo, K: Potássio, Ca: cálcio, Mg: manganês, Al: alumínio, H+Al: acidez potencial, SB: soma de bases, CTC(t): capacidade de troca catiônica efetiva, CTC(T): capacidade de troca catiônica a pH 7, V: índice de saturação de bases, m: índice de saturação de alumínio, PCA 1: componente principal 1, PCA 2: componente principal 2.

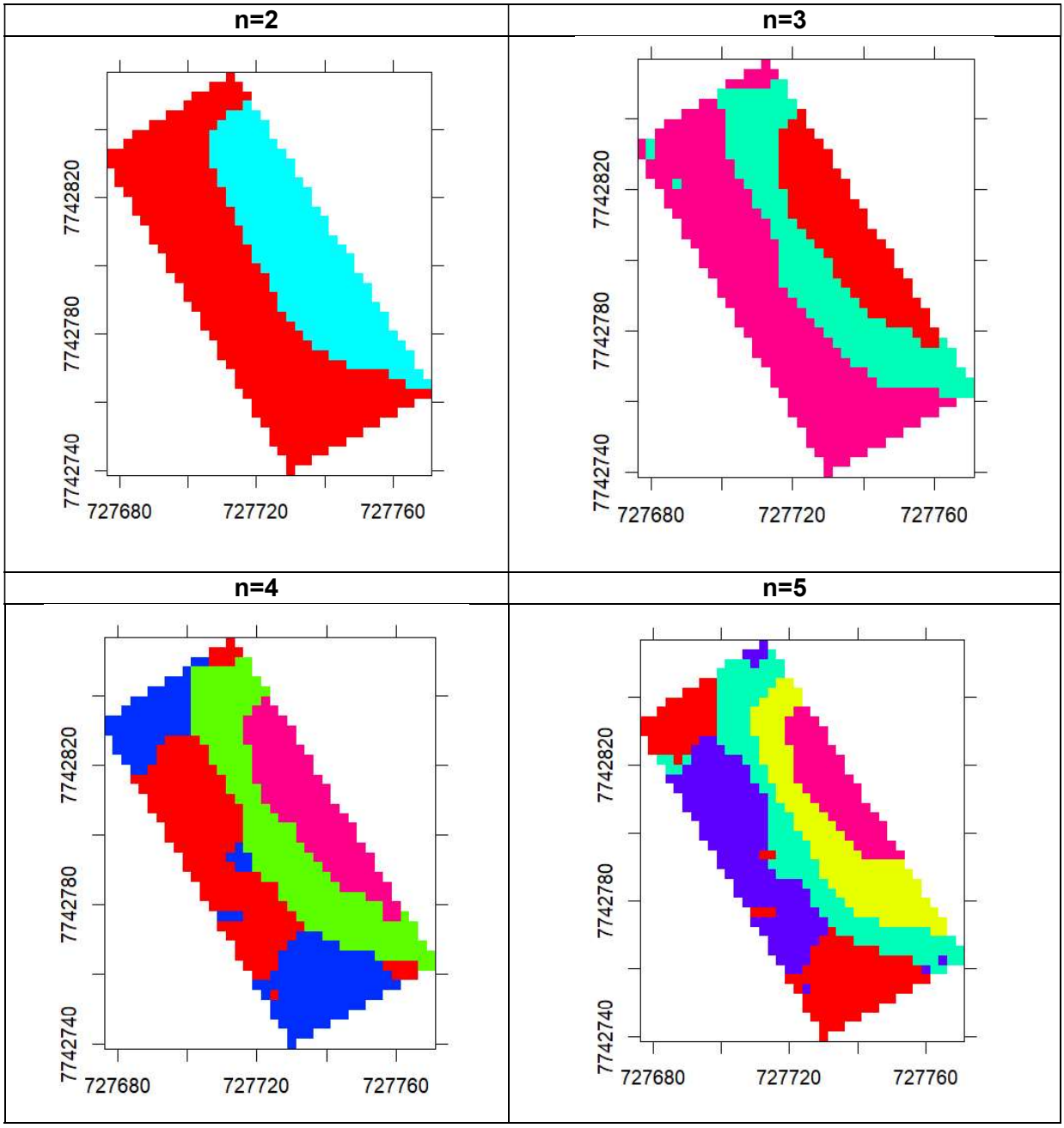
4.4. Análise de agrupamento

Para determinar os blocos experimentais, levando em conta as informações de todas as variáveis químicas, os componentes principais 1 e 2 foram utilizados no algoritmo de agrupamento *k-means*. A Figura 14 apresenta os grupos formados, ou seja, os blocos com maior uniformidade considerando os dois PC, evidenciado o número de blocos em função da cor bem como suas formas.

Salienta-se que a divisão da área experimental em um número maior de blocos irá aumentar a uniformidade dentro dos blocos. Porém, a área de cada bloco será reduzida, o que limita o número de tratamentos a serem testados, bem como o tamanho das unidades experimentais. Assim, a escolha do número de blocos deverá levar em conta o número de tratamentos e a cultura estudada.

Considerando apenas 2 blocos na área experimental, fica evidente a continuidade espacial dos blocos vermelho e azul. Nota-se na Figura 14, que os blocos apresentam áreas distintas, sendo o bloco vermelho maior do que o azul. Além disso, esses blocos não se apresentaram em forma de polígonos, com lados retos, evidenciando curvas que acompanham a variabilidade espacial do terreno.

Figura 14 – Blocos agrupados via algoritmos *k-means* em função do número n de grupos indicado.



Na partição da área em 3 blocos novamente foram observados blocos com formas não regulares e áreas distintas. Apesar do algoritmo *k-means* ter classificado alguns pontos como pertencentes ao bloco verde no interior do bloco rosa, a situação prática em campo talvez possa ignorar esses poucos pontos e tratar os 3 blocos obtidos como contínuos.

Dividindo a área em 4 blocos, novamente observa-se formas e áreas distintas para os blocos. O que chamou a atenção foi o bloco azul devido a evidente descontinuidade. Nota-se que, neste caso, as duas partes do bloco azul são

relativamente grandes, e no planejamento experimental, devem receber uma repetição de cada tratamento.

Considerando a divisão da área em 5 blocos, as observações quanto ao tamanho e forma dos blocos foram semelhantes às apresentadas para 4 blocos.

Nota-se que, à medida em que o número de blocos a serem determinados aumentou, houve crescimento da descontinuidade dos blocos em relação a áreas pequenas. Para fins práticos, talvez seja mais interessante ao pesquisador demarcar e inutilizar essas pequenas áreas, caso não haja área suficiente para implementar uma parcela experimental.

A Tabela 11 apresenta os valores obtidos na classificação dos 729 pontos referentes aos componentes principais 1 e 2 selecionados, que estão relacionados aos resultados da Figura 14.

Tabela 11 – Desuniformidade entre e dentro dos blocos em função das Somas de Quadrados associados aos PC 1 e 2 para divisão da área experimental em 2, 3, 4 e 5 blocos

Parâmetros	**Número de blocos**			
	n=2	**n=3**	**n=4**	**n=5**
SQ_T	7.185,97			
SQ_{TE}	4.254,07	5.030,79	5.699,24	6.013,00
$SQ_{Amarelo}$	N	N	N	242,77
SQ_{Azul}	1.215,84	N	329,64	263,37
SQ_{Rosa}	N	1.201,08	387,07	206,66
SQ_{Verde}	N	496,27	445,87	207,97
$SQ_{Vermelho}$	1.716,06	457,82	324,15	252,20
$N_{Amarelo}$	N	N	N	149
N_{Azul}	276	N	178	181
N_{Rosa}	N	364	134	86
N_{Verde}	N	216	213	160
$N_{Vermelho}$	453	149	204	153

SQ_T: soma de quadrados total, SQ_{TD}: total da soma de quadrado dentro dos blocos determinados, SQ_{TE}: total da soma de quadro entre os blocos, $SQ_{Amarelo}$, SQ_{Azul}, SQ_{Rosa} , SQ_{Verde}, $SQ_{Vermelho}$: soma de quadrados dos blocos amarelo, azul, rosa, verde e vermelho, respectivamente; $N_{Amarelo}$, N_{Azul}, N_{Rosa}, $N_{Verde,}$ $N_{Vermelho}$: número de pontos classificados no bloco amarelo, azul, rosa, verde e vermelho, respectivamente; N: não se aplica.

A soma de quadrados total (SQ_T) obtida foi a mesma para todos os blocos. Esse resultado já era esperado uma vez que esse valor é calculado com base na variância dos componentes principais 1 e 2 (escores).

Como foi observado em todas as situações discutidas anteriormente, os blocos experimentais formados via técnica de agrupamento apresentaram áreas diferentes. Utilizando o *software* R, com auxílio da biblioteca *pracma* foram obtidas as áreas estimadas para cada bloco (Tabela 12).

Tabela 12 – Áreas dos blocos obtidos via *software* R com auxílio da biblioteca *pracma*.

	Número de blocos							
	n=2		**n=3**		**n=4**		**n=5**	
Blocos	**A (m^2)**	**P (%)**	**A (m^2)**	**P (%)**	**A (m^2)**	**P (%)**	**A (m^2)**	**P (%)**
Amarelo	N	N	N	N	N	N	637	18,9
Azul	1.260	37,5	N	N	758	22,6	743	22,1
Rosa	N	N	1723	51,3	609	18,1	408	12,1
Verde	N	N	985	29,3	976	29,0	708	21,0
Vermelho	2100	62,5	652	19,4	1017	30,3	864	25,7

A: área de cada bloco, P: proporção relativa à área ocupada pelo bloco em relação a área total, N: não se aplica.

Blocos de tamanhos diferentes, como apresentado na Tabela 12, implica em algumas restrições que devem ser consideradas durante o planejamento experimental.

Ferreira (2020) utilizou parcelas de 14 m^2 para experimento de seleção com cana-de-açúcar, o que está em acordo com o trabalho de Leite *et al.* (2009), bem como apresentado por Igue *et al.* (1991). Assim, considerando os blocos de menor área, a divisão da área experimental em 2, 3, 4 ou 5 blocos, permitiria testar, respectivamente, 90, 46, 43 e 29 tratamentos em experimentos com cana-de-açúcar.

Assis e Silva (1999) concluíram que para experimentos com milho a parcela experimental ideal deve variar de 0,75 a 6,77 m^2. Considerando uma parcela de 5 m^2, seria possível estabelecer, dentro dos menores blocos para a divisão de 2, 3, 4 e 5 blocos, respectivamente, 252, 130, 121 e 81 unidades experimentais.

Em situações que envolvam muitos tratamentos e/ou unidades experimentais muito extensas, o pesquisador pode optar pelo delineamento em blocos incompletos.

Esse delineamento é mais difícil de ser analisado quando comparado ao delineamento de blocos completos, mas isso é compensado pelo ganho na precisão experimental (GOMES, 1990).

Foi calculado o coeficiente de variação dentro de cada bloco (Tabela 12) para os 5 atributos químicos de solo que apresentaram maiores valores de CV na área total (Tabela 4), a saber, cálcio (Ca), magnésio (Mg), alumínio (Al), soma de bases trocáveis (SB) e índice de saturação de bases (V). Como havia sido apresentado anteriormente, todas essas variáveis apresentaram CV classificados como muito altos de acordo com Gomes (1990), o que reflete a grande heterogeneidade da área experimental.

Para situação com 2 blocos, o bloco vermelho se mostrou homogêneo para quase todas as 5 variáveis avaliadas. Quatro delas apresentaram CV classificado como baixo (inferior a 10%) e apenas o índice de saturação de bases V apresentou com valor classificado como médio. Tal homogeneidade foi menos observada no bloco azul que apresentou quase todos os valores de CV classificados como médios.

À medida em que o número de blocos experimentais aumentou, os valores de CV para as variáveis em análise diminuíram. Para n = 5 blocos o maior valor de CV observado foi para variável V com valor de 13,60% (classificado como médio). Considerando as demais variáveis e os blocos determinados, a maioria dos valores de coeficiente de variação obtidos foram menores do que 10%.

Os resultados apresentados na Tabela 12 permitem verificar que a área experimental inicial apresentava grande variação para os atributos químicos do solo, com variáveis chegando a apresentar 95,78% para o coeficiente de variação. No entanto, quando se aplicou a metodologia exposta neste trabalho, foi possível determinar blocos experimentais bem mais uniformes com valores de CV, no caso mais extremo, sendo classificado como médio.

Tabela 12 – Coeficientes de variação dentro de cada bloco para variáveis cálcio, magnésio, alumínio, soma de bases e índice de saturação de bases.

Blocos	**CV_{Ca} (%)**	**CV_{Mg} (%)**	**CV_{Al} (%)**	**CV_{SB} (%)**	**CV_{V} (%)**
Área total					
	95,78*	49,55*	37,14*	48,94*	49,07*
2 Blocos					
Vermelho	7,69	7,83	7,02	6,48	12,63
Azul	19,00	18,15	4,88	14,93	18,56
3 Blocos					
Vermelho	13,78	13,87	3,11	10,88	16,54
Verde	13,48	9,78	4,95	8,63	14,45
Rosa	5,56	5,59	6,23	4,84	10,97
4 Blocos					
Rosa	13,07	13,41	2,99	10,43	16,24
Verde	13,92	10,02	4,69	8,82	14,83
Azul	6,55	5,46	7,08	4,68	13,37
Vermelho	4,43	5,78	5,11	3,97	9,42
5 Blocos					
Rosa	11,64	11,21	1,66	8,67	13,60
Amarelo	11,45	9,35	3,54	7,97	13,01
Verde	10,77	6,22	4,69	5,65	11,44
Azul	3,57	5,07	4,72	3,46	9,04
Vermelho	11,45	9,35	3,54	7,97	13,01

Abreviações: CV_{Ca}: coeficiente de variação para cálcio, CV_{Mg}: coeficiente de variação para magnésio, CV_{Al}: coeficiente de variação para alumínio; CV_{SB}: coeficiente de variação para soma de bases trocáveis, CV_{V}: coeficiente de variação para índice de saturação de bases.

* Valores calculados a partir dos 36 pontos amostrais iniciais.

5. CONCLUSÕES

A metodologia proposta, utilizando ferramentas da geoestatística, da análise de componentes principais e de técnicas de agrupamento, foi adequada para divisão da área experimental de cana-de-açúcar em blocos bem uniformes para as variáveis químicas do solo. Assim, esta metodologia tem potencial para ser empregada em outras áreas experimentais, bem como em pesquisas com outras culturas.

Das 12 variáveis químicas do solo, 10 apresentaram dependência espacial de moderada a forte, com alcance variando de 9,7 até 114,87 metros. Utilizando técnicas da geoestatística, foi possível interpolar valores para os atributos químicos em locais não amostrados. Para os atributos que não apresentaram dependência espacial, as estimativas foram realizadas com base na técnica do inverso da distância.

Os dados interpolados, submetidos a análise de componentes principais, permitiu reduzir as 12 variáveis originais para dois componentes que explicaram 82,27% da variância total dos dados.

Os *escores* dos componentes principais classificados pelo algoritmo *k-means*, permitiu a divisão da área experimental em 2, 3, 4 e 5 blocos com alta uniformidade. Assim, considerando parcelas de 14 m^2 nesta área experimental com cana-de-açúcar, pode-se avaliar pelo menos 90, 46, 43 e 29 tratamentos, respectivamente com 2, 3, 4 ou 5 repetições.

Os blocos obtidos não apresentaram formas poligonais regulares, sendo obtidos blocos com formatos e tamanho diversos, além de alguns blocos com áreas descontinuas. Indicando que a formação usual de blocos, com formato regular e sem considerar a variabilidade espacial de atributos relacionados com a fertilidade do solo, não é adequada.

REFERÊNCIAS

ADÃO, A. S. et al. **Análise da correlação dos atributos físicos do solo com os componentes de rendimento de grãos de milho em diferentes sistemas de cultivo**. Research, Society and Development, v. 11, n. 2, 2022.

ASSIS, J. P. de; SILVA, P. S. L. **Tamanho e forma ideais da unidade experimental em ensaio com milho**. Agropecuária Técnica, v. 20, p. 42-50, 1999.

BANZATTO, D. A; KRONKA, S. N. **Experimentação agrícola.** 4 ed. Jaboticabal: Funep, 2006. 237 p.

BERNARDI, AC de C. et al. **Agricultura de precisão: resultados de um novo olhar**. Embrapa Instrumentação-Livro técnico (INFOTECA-E), 2014.

BUAINAIN, A. M. et al. **O mundo rural no Brasil do século 21: a formação de um novo padrão agrário e agrícola**. Brasília, DF: Embrapa, 2014., 2014.

CRESSIE, Noel. **Fitting variogram models by weighted least squares**. Journal of the international Association for mathematical Geology, v. 17, n. 5, p. 563-586, 1985.

DEVORE, J. L. **Probabilidade e estatística: para engenharia e ciências.** 6. ed. São Paulo: Cengage Learning, 2006. 708 p.

FERREIRA, D. F. **Análise multivariada**. 3. Ed. Lavras: UFLA, 2018. p. 394.

FERREIRA, M. P. **Geoestatística e aerofotogrametria aplicada à seleção de família de cana-de-açúcar**. 2020. 72 p. Tese (Doutorado em Estatística Aplicada) – Universidade Federal de Viçosa, Viçosa, 2020.

FORMAGGIO, A. R; SANCHES, I. D. **Sensoriamento Remoto em agricultura.** 1. ed. São Paulo: Oficina de Textos, 2017. 284 p.

GOMES, F. P. **Curso de estatística experimental.** 13. ed. Piracicaba: Livraria Nobel, 1990. 461 p.

GUIMARÃES, W. D. **Geoestatística para o mapeamento da variabilidade espacial de atributos física do solo.** 2013. 60 p. Tese (Doutorado em Engenharia Civil) – Universidade Federal de Viçosa, Viçosa, 2013.

HERNÁNDEZ, M. M. **Análise geoestatística de multivariada para definição de zonas de manejo em cana-de-açúcar (*Saccharum officinarum) na Guatemala.*** 2021. 60 p. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa, 2021.

IGUE, T. et al. **Tamanho e forma de parcela experimental para cana-de-açúcar**. Bragantia, v. 50, p. 163-180, 1991.

ISAAKS, E. H.; SRIVASTAVA, R. M. **An Introduction to Applied Geostatistics**. New York: Oxford Univ. 1989.

JAMES, G. et al. **An introduction to statistical learning**. New York: springer, 2013.

JOURNEL, Andre G.; HUIJBREGTS, Charles J. **Mining geostatistics**. Blackburn Press, 1976.

JÚNIOR, R. et al. **geoR: Analysis of Geostatistical Data**. R package version 1.8-1. p. 155, 2020. DOI: 10.1007/978-0-387-48536-2>.

LEITE, M. S. O. et al. **Sample size for full-sib family evaluation in sugarcane**. Pesquisa Agropecuária Brasileira, v. 44, n. 12, p. 1562-1574, 2009.

MONTEIRO, A. M. V. et al. **Análise espacial de dados geográficos**. Brasília: Embrapa, 2004.

NETO, B. B.; SCARMINIO, I. S.; BRUNS, R. E. **Como fazer experimentos: pesquisa e desenvolvimento na ciência e na indústria.** 2. ed. Campinas: Editora Unicamp, 2001. 401 p.

NETO, P. L. O. C. **Estatística.** 2. ed. São Paulo: Edgar Blücher Ltda, 2002. 265 p.

NOCELLI, R. C. F. et al. **Histórico da cana-de-açúcar no Brasil: contribuições e importância econômica. Cana-de-açúcar e seus impactos: uma visão acadêmica**, p. 13, 2017.

PASINI, M. P. B. et al. **Selection of Interpolators to Predict Populations of Tibraca limbativentris in Irrigated Rice**. Brazilian Archives of Biology and Technology, v. 64, 2022.

PEREIRA, R. G. **Características morfológicas e teores de cafeína e teobromina entre ervais nativos sombreados e a pleno sol.** 2021. 62 p. Dissertação (Mestrado em Ciências Florestais) – Universidade Estadual do Centro-Oeste, Irati, 2021.

R CORE TEAM. R: **A language and environment for statistical computing**. Vienna: R Fundation for Statistical Computing, 2020.

SCHLOERKE, B.; CROWLEY, J.; COOK, D. **Package 'GGally': Extension to 'ggplot2.'** v. 713, 2018.

SOUZA, G. S. et al. **Variabilidade espacial de atributos químicos em um Argissolo sob pastagem**. Acta Scientiarum. Agronomy, v. 30, p. 589-596, 2008.

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

YAMAMOTO, J. K.; LANDIM, P. M. B. **Geoestatística: conceitos e aplicações**. Oficina de textos, 2015.

APÊNDICE A – Histograma e *boxplot*

Figura A-1 – Histograma e *boxplot* para os atributos potencial hidrogeniônico, fósforo e potássio.

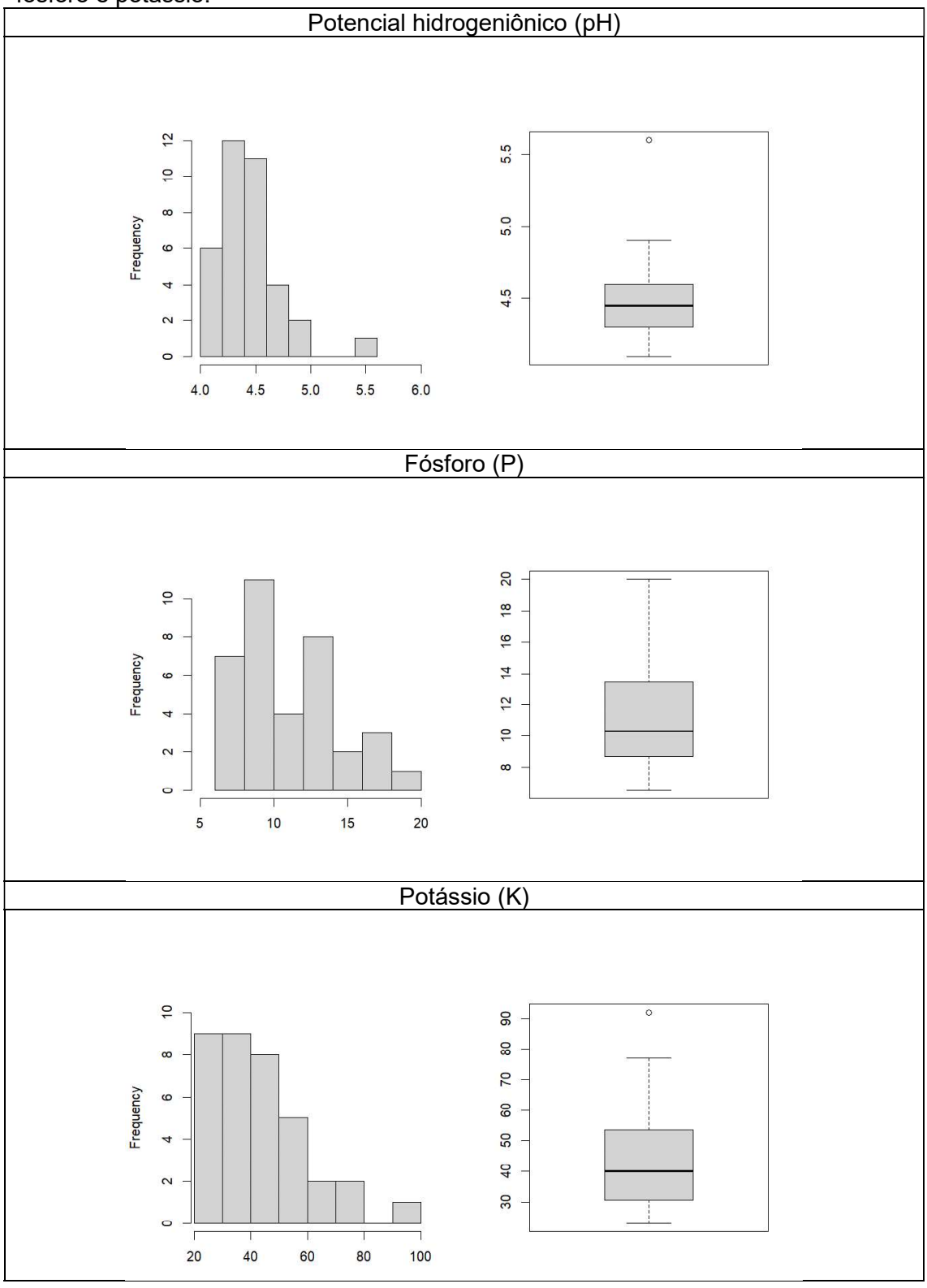


Figura A-2 – Histograma e boxplot para os atributos cálcio, magnésio e alumínio.

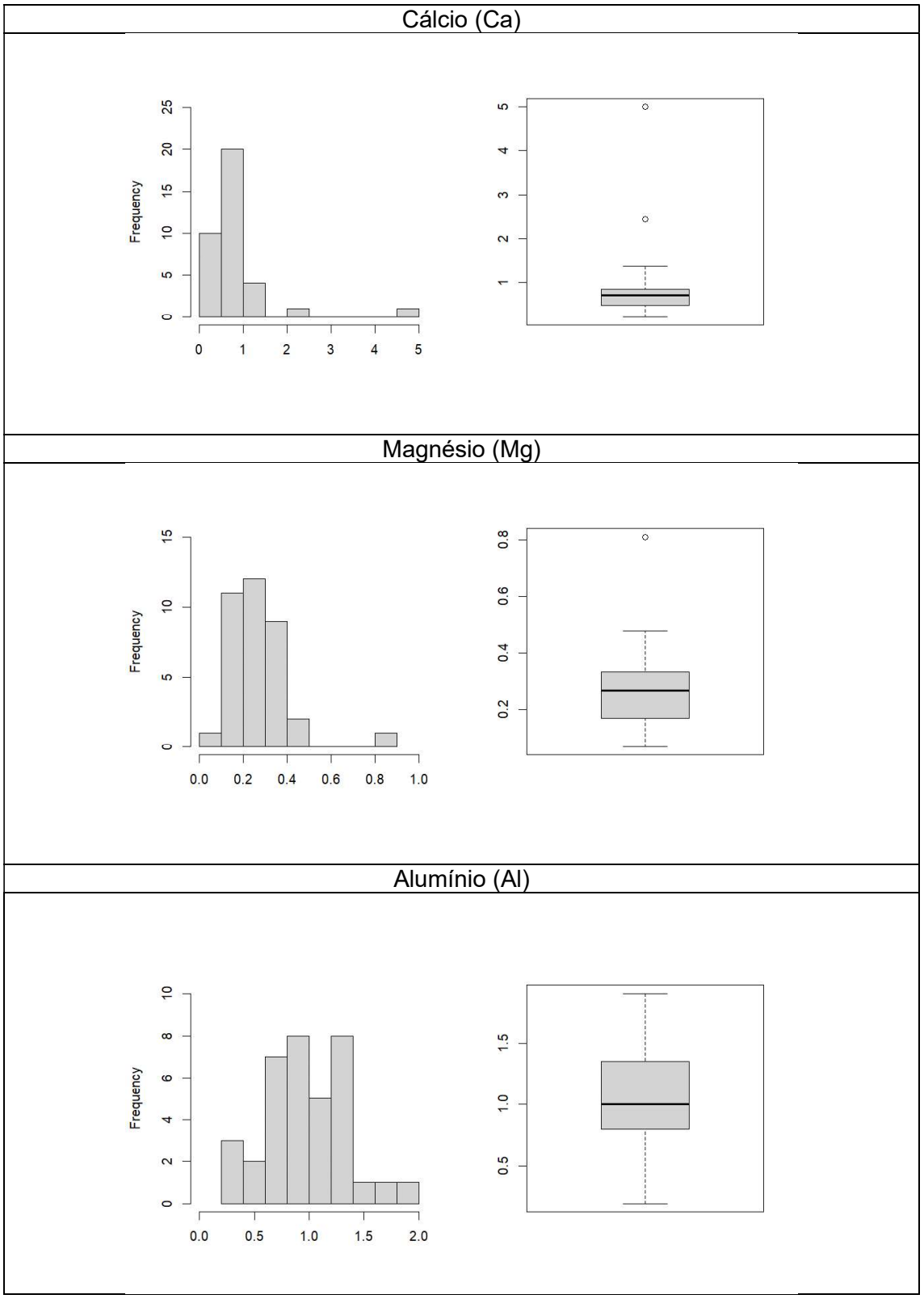


Figura A-3 – Histograma e boxplot para os atributos acidez potencial, soma de bases trocáveis e capacidade de troca catiônica efetiva.

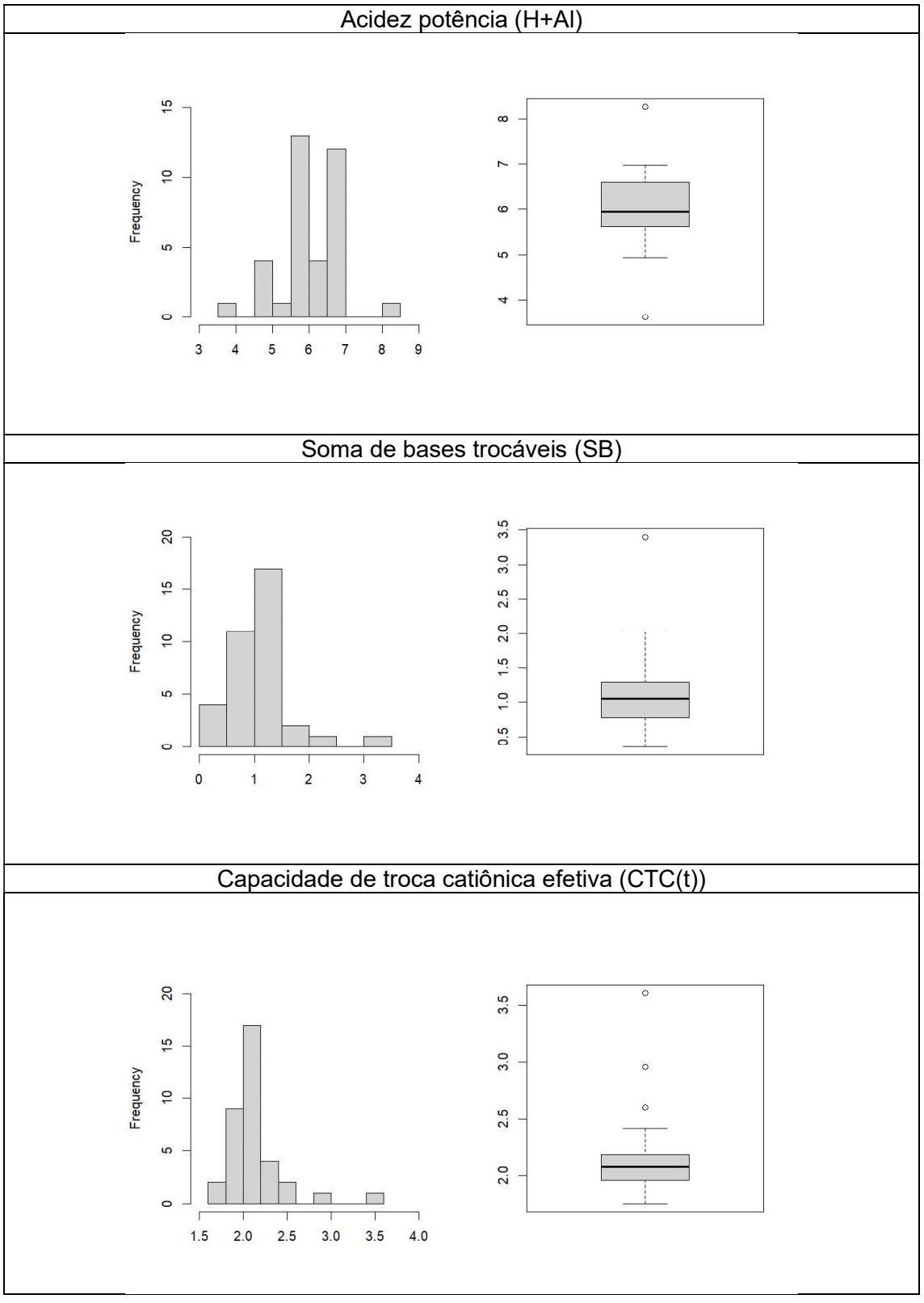
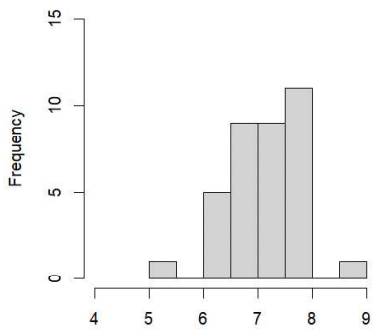
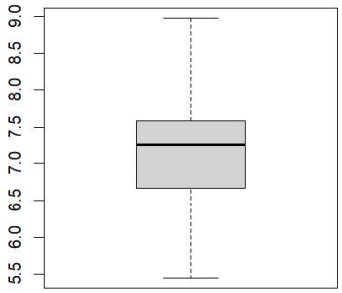


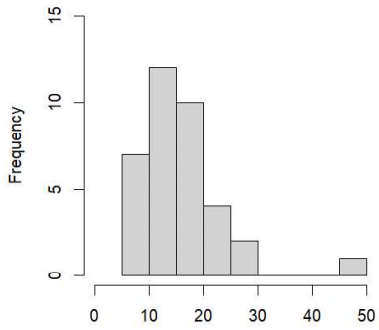
Figura A-4 – Histograma e boxplot para os atributos capacidade de troca catiônica a pH 7, Índice de saturação de bases e índice de saturação de alumínio.

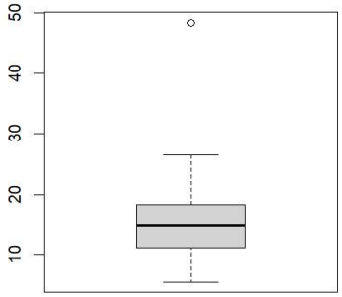
Capacidade de troca catiônica a pH 7 (CTC(T))



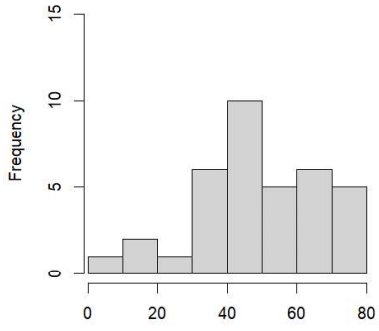


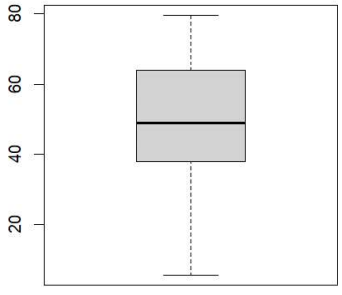
Índice de saturação de bases (V)





Índice de saturação de alumínio (m)





APÊNDICE B – Semivariogramas

Figura B-1 - Variogramas experimental omnidirecional e curva de modelo teórico ajustada para os atributos potencial hidrogeniônico, fósforo, alumínio, potássio, cálcio e soma de bases trocáveis.

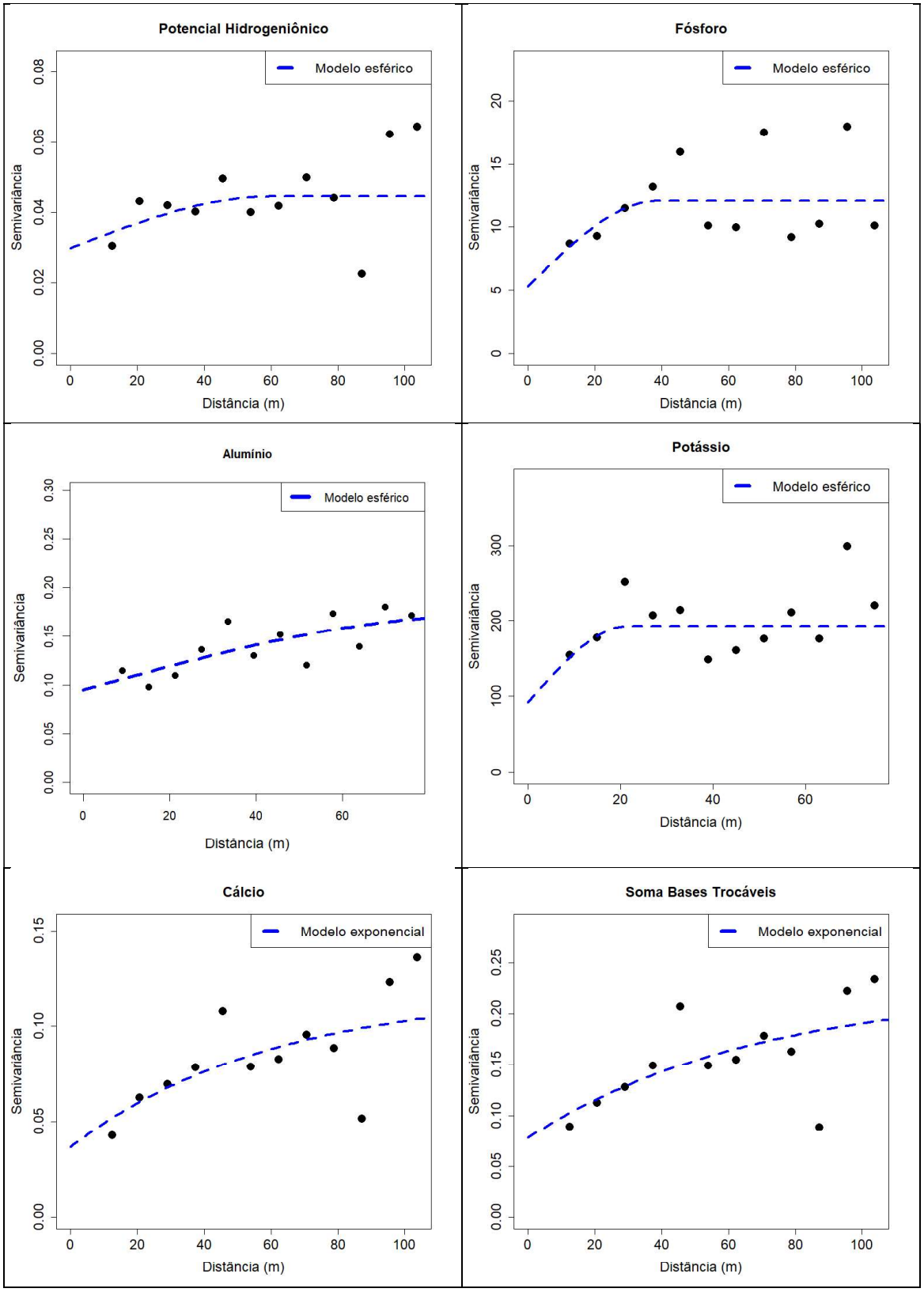
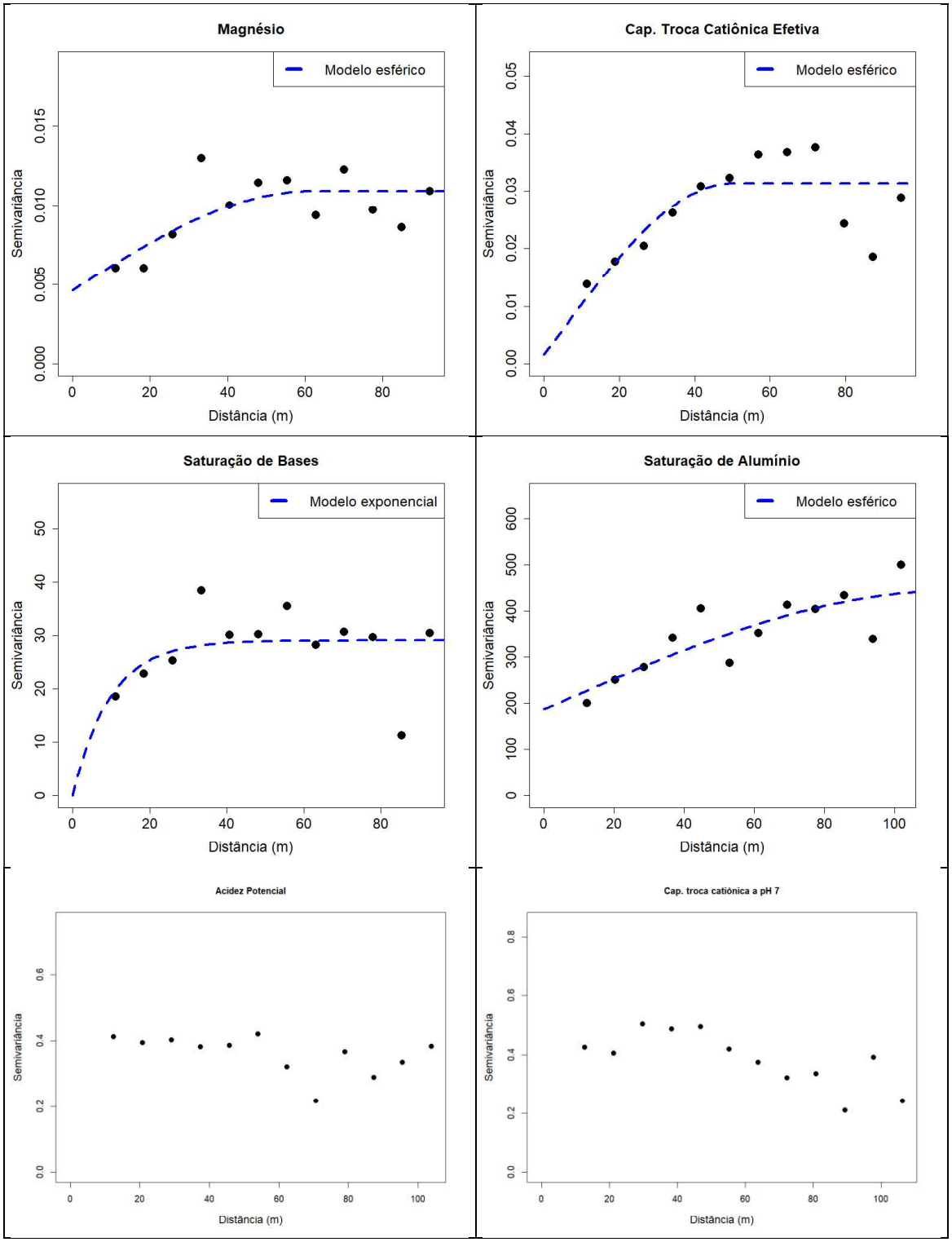


Figura B-2 - Variogramas experimental omnidirecional e curva de modelo teórico ajustada para os atributos magnésio, capacidade de troca catiônica efetiva, índide de saturação de bases, índice de saturacao de alumínio.



APÊNDICE C – Mapas obtidos via krigagem

Figura C-1 – Mapas estimados via krigagem ordinária para as variáveis potencial hidrogeniônico, magnésio, cálcio, soma de bases trocáveis, alumínio e índice de saturação de alumínio.

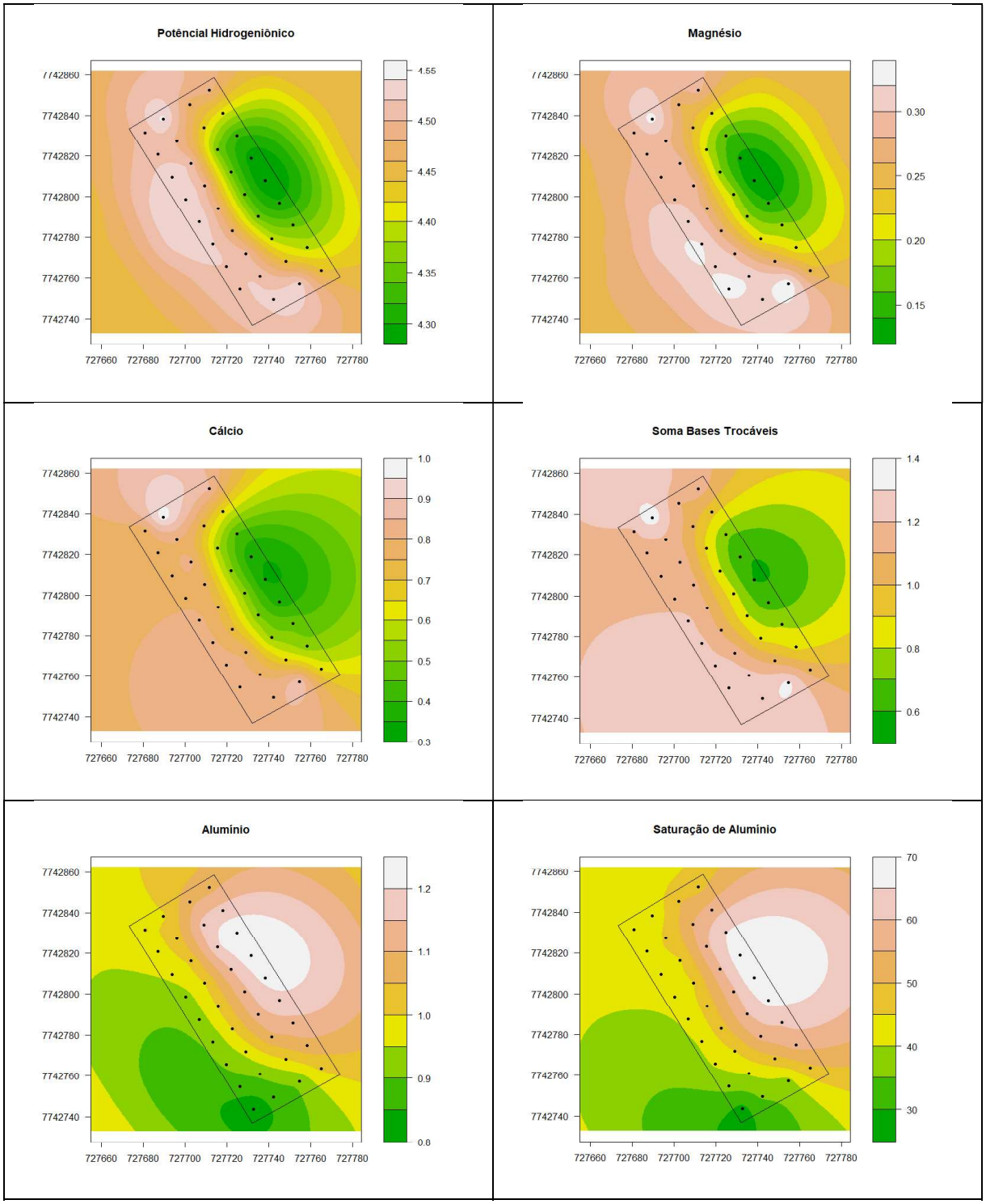


Figura C-2 – Mapas estimados via krigagem ordinária para as variáveis capacidade de troca catiônica efetiva, índice de saturaão de bases e potássio.