MARIANA DE OLIVEIRA SILVA

ANÁLISE DA DEPENDÊNCIA ESPACIAL EM EXPERIMENTOS COM CANA-DE-AÇÚCAR DA RIDESA

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 Souza Carneiro

Coorientadores: Luiz Alexandre Peternelli
Matheus de Paula Ferreira

VIÇOSA - MINAS GERAIS
2024

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

T
S586a
2024

Silva, Mariana de Oliveira, 1999-
Análise da dependência espacial em experimentos com cana-de-açúcar da RIDESA / Mariana de Oliveira Silva. – Viçosa, MG, 2024.
1 dissertação eletrônica (57 f.): il. (algumas color.).

Orientador: Antônio Policarpo Souza Carneiro.
Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Estatística, 2024.
Referências bibliográficas: f. 54-57.
DOI: https://doi.org/10.47328/ufvbbt.2024.653
Modo de acesso: World Wide Web.

1. Análise espacial (Estatística). 2. Geologia - Métodos estatísticos. 3. Cana-de-açúcar - Melhoramento genético. 4. Agricultura de precisão. 5. Correlação (Estatística). I. Carneiro, Antônio 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. 519.5

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

MARIANA DE OLIVEIRA SILVA

ANÁLISE DA DEPENDÊNCIA ESPACIAL EM EXPERIMENTOS COM CANA-DE-AÇÚCAR DA RIDESA

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: 28 de setembro de 2024

Assentimento:

Documento assinado digitalmente
gov.br
MARIANA DE OLIVEIRA SILVA
Data: 08/10/2024 09:56:37-0300
Verifique em https://validar.iti.gov.br

Mariana de Oliveira Silva
Autor

Documento assinado digitalmente
gov.br
ANTONIO POLICARPO SOUZA CARNEIRO
Data: 08/10/2024 15:00:37-0300
Verifique em https://validar.iti.gov.br

Antônio Policarpo Souza Carneiro
Orientador

Dedico aos meus pais e minha irmã.

AGRADECIMENTOS

Agradeço a Deus pelas bençãos em minha caminhada e a São Judas Tadeu por sua proteção e intercessão nos momentos de dificuldade.

Aos meus pais, Alberto e Maria Marta, agradeço pelo amor incondicional, pelo constante incentivo ao longo de todos esses anos de estudo e pelos valores que me foram transmitidos. Vocês são a base de todas as minhas conquistas.

À minha irmã, Manuela, agradeço pela amizade, apoio e por sempre acreditar no meu sucesso.

Aos meus amigos “MatMigos”, com quem construí minha família de Viçosa, e que têm estado ao meu lado desde a época da graduação. E aos demais de Viçosa, gratidão pelo apoio e pela companhia no dia a dia.

Aos meus amigos de Ponte Nova, cuja amizade trouxe conforto e apoio em momentos distantes, mas que permanecem próximos em meu coração.

Aos amigos da “Salinha” do PPESTBIO, pelo companheirismo nos estudos e pela amizade que trouxe momentos de diversão e descontração.

Às minhas amigas, Brenda e Sara, pela amizade, companheirismo e por estarem sempre ao meu lado. Sem vocês, essa caminhada teria sido muito mais difícil.

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, pela oportunidade de realizar a pós-graduação.

Ao Grupo de Estudos em Estatística Aplicada e Biometria - GESTBIO, do qual faço parte, expresso meus agradecimentos pelas contribuições e orientações nos trabalhos realizados ao longo do meu mestrado.

Ao meu orientador, Prof. Antônio Policarpo Souza Carneiro, por toda a paciência, disponibilidade e empenho com que me orientou neste trabalho e nos demais que realizei durante o mestrado. Muito obrigada pelos ensinamentos e pela motivação durante este ciclo.

Aos meus coorientadores, Prof. Luiz Alexandre Peternelli e Matheus de Paula Ferreira, pelas contribuições no trabalho e pelas sugestões nas rotinas.

Aos professores e funcionários do Departamento de Estatística - DET/UFV, por serem sempre solícitos e por contribuírem para minha formação.

Aos membros da banca, por aceitarem o convite e contribuírem para a melhoria deste trabalho.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Código de Financiamento 001, pela concessão da bolsa de estudos.

À Rede Interuniversitária para o Desenvolvimento do Setor Sucroenergético (RIDESA) e ao professor Márcio Henrique Pereira Barbosa, atual coordenador do Programa de Melhoramento Genético da Cana-de-Açúcar da UFV, pela disponibilidade dos dados para a realização dessa pesquisa.

Por fim, a todos que contribuíram de alguma forma para a concretização desta dissertação, meu sincero agradecimento.

“Jamais perca o seu equilíbrio por mais forte que seja o vento da tempestade”.

(Ponto de Equilíbrio)

RESUMO

SILVA, M.O., M.Sc., Universidade Federal de Viçosa, outubro de 2024. **Análise da dependência espacial em experimentos com cana-de-açúcar da RIDESA**. Orientador: Antônio Policarpo Souza Carneiro. Coorientadores: Luiz Alexandre Peternelli e Matheus de Paula Ferreira.

A cana-de-açúcar é uma das culturas agrícolas mais importantes para a economia brasileira, sendo a principal matéria-prima para a produção de açúcar e etanol. Dada sua relevância, o melhoramento genético da cana-de-açúcar é essencial para aumentar a produtividade e a sustentabilidade do setor sucroenergético. A análise de variância tradicional, que assume a independência dos erros, frequentemente confia ao princípio da casualização a tarefa de neutralizar a correlação entre os erros. No entanto, quando a casualização não é realizada corretamente ou a dependência espacial entre parcelas é ignorada, os resultados podem ser comprometidos, reduzindo a eficácia da análise para a seleção de genótipos realmente superiores. Neste estudo, realizou-se uma análise estatística espacial em dois experimentos conduzidos pela Rede Interuniversitária para o Desenvolvimento do Setor Sucroenergético (RIDESA) com o objetivo de avaliar a dependência espacial dos erros aleatórios e verificar se a análise espacial melhora a precisão experimental. A análise inicial dos resíduos empregou o índice de Moran e semivariogramas para identificar a autocorrelação espacial e modelar a estrutura de dependência espacial. Essa estrutura foi posteriormente incorporada aos modelos por meio da matriz de variância e covariância residual (R), possibilitando a comparação entre modelos com erros independentes, modelos com erros dependentes que consideram o controle local do experimento, e modelos com erros dependentes que desconsideram o controle local do delineamento em blocos casualizados. O modelo que desconsiderou o controle local e incorporou a dependência espacial dos erros mostrou o melhor ajuste em um dos experimentos, evidenciando que, em certas condições, a consideração da estrutura espacial pode ser mais eficaz do que o controle local no delineamento.

Palavras-chave: autocorrelação espacial; melhoramento genético; erros correlacionados; geoestatística; precisão experimental.

ABSTRACT

SILVA, M.O., M.Sc., Universidade Federal de Viçosa, October, 2024. **Análise da dependência espacial em experimentos com cana-de-açúcar da RIDESA**. Orientador: Antônio Policarpo Souza Carneiro. Coorientadores: Luiz Alexandre Peternelli e Matheus de Paula Ferreira.

Sugarcane is one of the most important agricultural crops for the Brazilian economy, being the main raw material for the production of sugar and ethanol. Given its relevance, the genetic improvement of sugarcane is essential for increasing productivity and sustainability in the sugar-energy sector. Traditional analysis of variance, which assumes the independence of errors, often relies on the principle of randomization to neutralize error correlation. However, when randomization is not properly carried out or spatial dependence between plots is ignored, results can be compromised, reducing the effectiveness of the analysis for selecting truly superior genotypes. In this study, a spatial statistical analysis was conducted on two experiments carried out by Rede Interuniversitária para o Desenvolvimento do Setor Sucroenergético (RIDESA) to assess the spatial dependence of random errors and determine whether spatial analysis enhances experimental precision. The initial analysis of residuals employed Moran's index and semivariograms to identify spatial autocorrelation and model the spatial dependence structure. This structure was then incorporated into the models through the residual variance-covariance matrix (R), enabling a comparison between models with independent errors, models with dependent errors that consider local control of the experiment, and models with dependent errors that disregard local control in randomized block design. The model that disregarded local control and incorporated spatial error dependence showed the best fit in one of the experiments, demonstrating that, under certain conditions, considering spatial structure can be more effective than local control in design.

Keywords: spatial autocorrelation; genetic improvement; correlated erros; geostatistics; experimental precision.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

A cana-de-açúcar tornou-se importante no setor agropecuário devido sua produção em larga escala, o que concede ao país geração de empregos diretos e indiretos em diversas regiões, além de movimentar toda uma cadeia que envolve desde o cultivo até a produção dos derivados como o açúcar, o etanol e energia renovável pelo bagaço. Atualmente, o Brasil se destaca como o maior produtor mundial de cana-de-açúcar, desempenhando um papel fundamental no mercado global com sua vasta área cultivada (EMBRAPA, 2023). Além disso, o país detém a liderança isolada na produção de álcool e açúcar, sendo também o maior exportador mundial de açúcar. A última estimativa, da safra 2023/24, confirmou o recorde de produção de cana-de-açúcar na série histórica da Companhia Nacional de Abastecimento (Conab), com um total de 713,2 milhões de toneladas, assim registrando um aumento de 16,8% em comparação com a safra anterior (CONAB, 2024).

Diante da importância no mercado financeiro, tem-se a crescente necessidade de maximizar a produtividade. Tal tarefa é atribuída aos programas de melhoramento genético do país que buscam o desenvolvimento de cultivares geneticamente superiores, que combinem o máximo de caracteres desejáveis para contribuir de forma positiva para a produção. Atualmente, o desenvolvimento, avaliação e recomendação de cultivares de cana no país são responsabilidades dos programas de melhoramento genético da cana-de-açúcar de três principais instituições de pesquisa, sendo elas: Rede Interuniversitária para o Desenvolvimento do Setor Sucroenergético (RIDESA – variedades RB); Instituto Agronômico de Campinas (variedades IAC); Centro de Tecnologia Canavieira (variedades CTC), que incorporou o programa das cultivares SP da Copersucar (MORAIS et al., 2015). Havia ainda, até 2015, o programa da CanaVialis - Monsanto (variedades CV).

O melhoramento genético é um processo demorado, no caso da cana-de-açúcar, pode levar de 12 a 15 anos desde a seleção das sementes que vão gerar os *seedlings* até a liberação da variedade. De acordo com Barbosa e Silveira (2010), foi convencionado cinco fases para o desenvolvimento de novas variedades, que são: primeira (T1), segunda (T2) e terceira (T3) fase de teste (ou de seleção), fase de

multiplicação clonal (FM) e fase experimental (FE) . Brasileiro (2013, p.1), doutor em genética e melhoramento, destaca que

> Uma das etapas mais importantes no melhoramento da cana-de-açúcar é a fase inicial (T1), onde são realizadas as primeiras seleções de plantas ou de famílias. Após a fase T1 novos materiais não são mais introduzidos, o que torna a seleção executada nessa fase crucial para o sucesso do programa.

Ao planejar um experimento, o pesquisador deve utilizar alguns princípios básicos para que os dados a serem obtidos permitam uma análise correta e levem a conclusões válidas em relação ao problema em estudo. Esses princípios, propostos por Ronald A. Fisher entre 1919 e 1925, são: princípio da repetição, da casualização e do controle local (FISHER, 1966). Enfatizando o princípio da casualização, a análise de variância tradicional confia a esse princípio a responsabilidade de neutralizar os efeitos da correlação entre os erros, conforme destacado por Duarte (2000 apud Feres 2009, p. 7). Entretanto, com frequência a casualização não é feita de forma correta, assim, caso ocorra a dependência espacial entre parcelas e sua existência seja desconsiderada, os resultados da análise de variância ficam comprometidos e a análise deixa de ser eficaz para a seleção de genótipos realmente superiores. Sendo assim, torna-se necessário técnicas de planejamento e análises que obtenham melhor precisão experimental, como por exemplo, a avaliação da dependência espacial entre parcelas.

Segundo Reis e Miranda Filho (2003), a distribuição não aleatória do tratamento pode originar uma dependência espacial entre as parcelas, manifestando-se como autocorrelação espacial no campo. Essa dependência impactou a análise conduzida em seus estudos e revelou que o modelo espacial ajustado foi mais eficaz em comparação ao modelo tradicional. Os autores ressaltam que a incorporação de análises espaciais deveria ser mais amplamente adotada por melhoristas de plantas, promovendo uma aprimorada eficiência nos programas de melhoramento genético.

Os métodos de análise espacial são métodos estatísticos que incorporam a localização geográfica das observações na análise. Esses métodos reconhecem que as observações próximas umas das outras tendem a ser mais semelhantes do que as distantes, o que reflete uma dependência espacial nos dados. Em resumo, considera-se a proximidade geográfica ao avaliar a relação entre as variáveis, reconhecendo

padrões que podem não ser detectados por métodos tradicionais de análise estatística.

A geoestatística é a área da estatística que possui metodologias para a identificação da existência ou não dessa dependência espacial. Pontes (2002) ressalta que “a correlação espacial entre as observações não é considerada um incômodo a ser evitado, mas sim uma fonte de informações que melhora a análise dos dados”. A desconsideração da dependência, caso ela exista, aumenta consideravelmente o erro experimental, sendo necessário uma análise mais acurada que é obtida por métodos da geoestatística.

Além de Reis e Miranda Filho (2003), que analisaram a autocorrelação espacial existente para resistência à lagarta do cartucho (Spodoptera frugiperda) nos compostos de milho, tem-se diversos outros trabalhos que comparam modelos estatísticos tradicionais e espaciais com a incorporação da autocorrelação dos erros, como Maia *et al.* (2013) que aplicou a análise espacial na avaliação de experimentos de seleção de clones de laranjeira Pêra, Salvador *et al.* (2022) na avaliação de dados do melhoramento genético do feijoeiro e Ferreira *et al.* (2024) para análise dos atributos do solo de uma área experimental de cana-de-açúcar.

O presente trabalho tem como objetivo avaliar a dependência espacial dos erros aleatórios em dois experimentos com cana-de-açúcar realizados pela Rede Interuniversitária para o Desenvolvimento do Setor Sucroenergético (RIDESA), bem como analisar se a incorporação da covariância entre os erros nas análises, por meio da matriz de variância e covariância residual, melhora a precisão experimental.

2. REVISÃO DE LITERATURA

2.1. Programa de Melhoramento Genético da cana-de-açúcar

Os programas de melhoramento da cana-de-açúcar desempenham um papel fundamental na evolução e aprimoramento dessa cultura agrícola tão importante para a economia brasileira. O objetivo principal desses programas é desenvolver variedades de cana-de-açúcar que sejam mais produtivas, resistentes a doenças e pragas, adaptadas a diferentes condições climáticas e com maior qualidade de sacarose para a produção de açúcar e etanol. Além disso, o melhoramento genético da cana-de-açúcar contribui para a sustentabilidade do setor sucroenergético e para a economia do país.

De acordo com Morais *et al*. (2015), no Brasil, o melhoramento genético é realizado por instituições públicas e privadas, com grande interação com o setor produtivo.

A fase inicial de seleção, conhecida como T1, é composta por seedlings provenientes de cruzamentos previamente definidos pelos programas de melhoramento genético. Em seguida, ocorre a fase T2, que inclui clones selecionados na cana-soca da fase anterior. A cana-soca refere-se à cana-de-açúcar que brota e cresce novamente após a colheita da cana-planta (a primeira colheita), diferindo desta por se desenvolver a partir do sistema radicular e parte do colmo remanescente no campo. Um aspecto distintivo da fase T2 é o uso do delineamento em blocos aumentados (DBA), que permite a avaliação de um grande número de clones sem a necessidade de repetições. A fase final de seleção, T3, envolve a avaliação de centenas de clones previamente selecionados. Por fim, ocorre a fase de multiplicação, cujo objetivo é multiplicar os clones selecionados para a produção de mudas a serem utilizadas na fase experimental, onde se avaliam os clones promissores. (PEDROZO, 2006).

2.1.1. RIDESA

A Rede Interuniversitária para o Desenvolvimento do Setor Sucroenergético (RIDESA) é uma organização que reúne esforços de várias universidades federais para impulsionar o setor sucroenergético, promovendo a pesquisa e o

desenvolvimento de novas tecnologias e cultivares. Nesse contexto, as atividades de pesquisa da RIDESA são conduzidas e compartilhadas entre todas as universidades envolvidas, promovendo a troca de informações, conhecimento e resultados. Isso amplia significativamente a capacidade e o alcance nacional das pesquisas e inovações realizadas. Como resultado, a Rede opera em escala nacional e, atualmente, é o principal centro de pesquisa relacionado à cana-de-açúcar sob a égide do Governo Federal.

A definição da nomenclatura das variedades de cana-de-açúcar da rede é utilizada na sigla RB “República do Brasil”, que é registrada no Germplasm Committee of International Society of Sugar Cane Technologists (ISSCT). Segundo Oliveira, Barbosa e Daros (2021), para que estas novas variedades sejam criadas, a RIDESA conta atualmente com 101 bases de pesquisa englobando laboratórios das universidades, estações de cruzamento, estações experimentais e subestações e bases de seleção, sendo essas últimas conduzidas em parceria com as empresas do setor canavieiro.

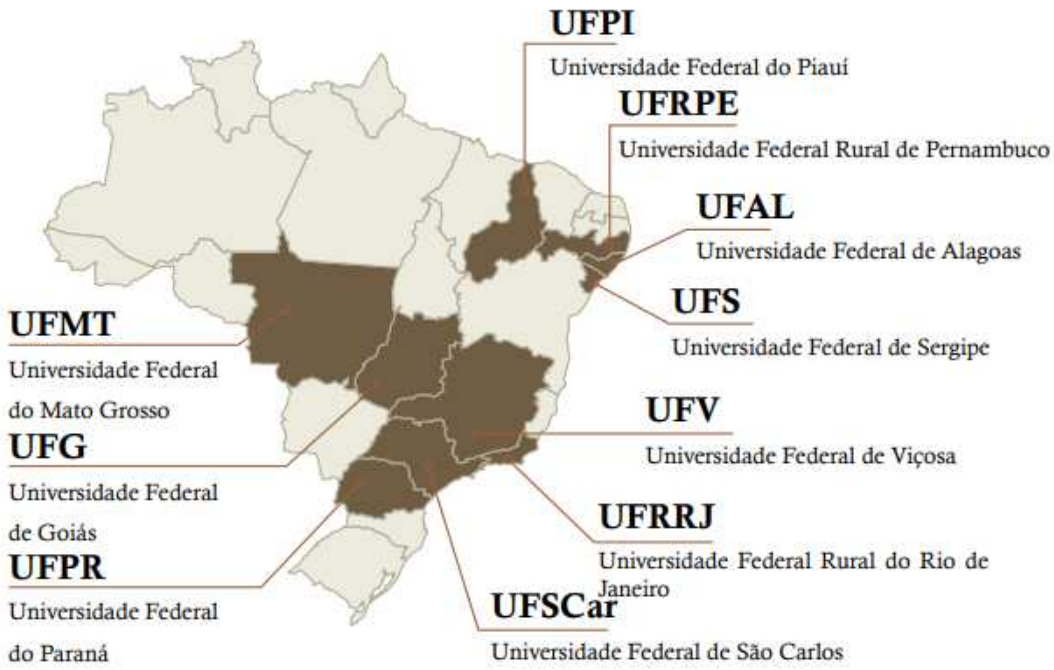
Com notável sucesso no avanço do desenvolvimento de variedades de cana-de-açúcar RB para o território brasileiro, a Universidade Federal de Viçosa (UFV) anunciou, em 2023, o lançamento de duas novas cultivares: RB127825 e RB097217. A primeira apresenta características de robustez, destacando-se pela adaptabilidade a diversos ambientes de produção, proporcionada por uma ampla margem de manejo. Tal versatilidade confere a esta variedade uma probabilidade significativa de aceitação no mercado. Por sua vez, a variedade RB097217 se destaca pela elevada concentração de açúcar, demandando, por conseguinte, condições edafoclimáticas propícias para maximizar a produtividade de sua biomassa. Este fato implica que sua produção seja preferencialmente conduzida em áreas de maior fertilidade. Essas novas variedades não apenas representam avanços significativos no desenvolvimento agronômico, mas também oferecem opções estratégicas aos produtores, permitindo uma adaptação mais eficaz às diferentes demandas e condições de cultivo encontradas no cenário brasileiro.

Além dessas conquistas, a rede demonstra excelência na capacitação de profissionais. A infraestrutura presente nas universidades tem servido de suporte para treinamento de estudantes tanto em nível de graduação quanto de pós-graduação no

estudo desta cultura. Centenas de profissionais qualificados foram formados e estão atualmente desempenhando papéis significativos tanto no setor privado quanto em instituições públicas.

A RIDESA conduz o Programa de Melhoramento Genético da Cana-de-Açúcar (PMGCA), envolvendo as instituições (Figura 1): Universidade Federal de São Carlos, Universidade Federal de Viçosa, Universidade Rural de Pernambuco, Universidade Federal de Alagoas, Universidade Federal do Paraná, Universidade Federal de Goiás, Universidade Federal do Piauí, Universidade Federal de Sergipe, Universidade Federal Rural do Rio de Janeiro e Universidade Federal de Mato Grosso (OLIVEIRA; BARBOSA; DAROS, 2021).

Figura 1**:** Ifes (Instituições Federais de Ensino Superior) participantes da Ridesa.



Fonte: OLIVEIRA; BARBOSA; DAROS (2021, p. 36).

A Universidade Federal de Viçosa (UFV) é uma das instituições que desempenham um papel significativo na Rede. A semeadura e produção dos seeldlings são realizadas nas duas Estações Experimentais da UFV, conhecidas como CECA (Figura 2) e CEPET (Figura 3). Após a seleção, alguns dos clones

escolhidos são plantados em experimentos no CECA, enquanto outros são enviados às usinas parceiras para avaliação inicial em vários municípios de Minas Gerais, abrangendo diferentes condições de solo e clima. Somente na fase 3, também chamada de fase T3, os melhores clones são compartilhados com outras universidades da rede, permitindo que esses clones sejam avaliados em diferentes regiões do Brasil.

Figura 2: Centro de Pesquisa e Melhoramento de Cana-de-Açúcar - CECA, Oratórios-MG.

Fonte: RIDESA/UFV. Disponível em: https://www.ridesaufv.com.br/processo-de-selecao. Acesso em: abr 2024.

Figura 3: Central de Experimentação, Pesquisa e Extensão do Triângulo Mineiro - CEPET, Capinópolis-MG.

Fonte: RIDESA/UFV. Disponível em: https://www.ridesaufv.com.br/processo-de-selecao. Acesso em: abr 2024.

2.2. Geoestatística

A geoestatística é uma subárea da estatística aplicada que desempenha um papel fundamental na análise e modelagem de dados espaciais. Ela se concentra em compreender a variabilidade espacial de fenômenos e na tomada de decisões informadas com base nessa variabilidade. Através da geoestatística, é possível explorar como os dados variam em diferentes locais em uma determinada área geográfica, identificando padrões, tendências e correlações espaciais. Segundo Yamamoto e Landim (2015), a geoestatística tem por objetivo a caracterização espacial de uma variável de interesse por meio do estudo de sua distribuição e variabilidades espaciais.

Conforme destacado por Cressie (2015), a geoestatística se fundamenta em dois conceitos: o semivariograma e a krigagem. O semivariograma é um gráfico que relaciona as semivariâncias entre pares de pontos em função das distâncias que os separam. Esse gráfico desempenha um papel crucial ao descrever a estrutura da variabilidade espacial, revelando como acontece a variância dos dados conforme a distância. Já a krigagem é um interpolador que prediz, não-tendenciosamente e com variância mínima, os valores não mensurados.

2.3. Índice de Moran

Um aspecto fundamental da análise exploratória espacial é a caracterização da dependência espacial, que mostra como os valores de determinadas variáveis estão correlacionados no espaço. Uma maneira eficaz de calcular essa autocorrelação espacial é utilizando o Índice de Moran.

Esse índice quantifica o grau em que uma variável em uma determinada localização espacial é semelhante a valores da mesma variável em localizações vizinhas, assim, avalia se os valores de uma variável tendem a ser similares ou diferentes em locais próximos, auxiliando na identificação de padrões de agrupamento ou dispersão dos dados. A fórmula do Índice de Moran é dada por:

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}\, z_i\, z_j}{S_0 \sum_{i=1}^{n} {z_i}^2} \quad (1)$$

Sendo,

n o total de pontos amostrados associado às localizações;

z_i é o desvio em relação à média, definido como $z_i = x_i - \bar{x}$;

x_i é a variável de interesse

w_{ij} é o peso obtido pelo inverso da distância entre os pontos i e j;

$S_0 = \sum_i^n \sum_j^n w_{ij}$.

Segundo Marconato, Larocca e Quintanilha (2012), o valor do Índice de Moran varia de -1 a +1, onde valores positivos indicam autocorrelação espacial positiva, ou seja, existe a tendência dos valores em posições vizinhas serem semelhantes. Os valores negativos indicam autocorrelação espacial negativa, isto é, a situação inversa da anterior. Valores próximos de 0 sugerem uma distribuição aleatória. Entretanto, o índice, por si só, não determina a existência de autocorrelação. Para avaliar a significância da autocorrelação, é necessário realizar um teste de hipóteses baseado na estatística z, como apresentado por Vendramini *et al.* (2010) e Ferreira (2015).

$$Z_I = \frac{I - E[I]}{\sqrt{V[I]}}, \text{ com } E[I] = \frac{-1}{n-1} \text{ e } V[I] = E[I^2] - E[I]$$

A hipótese nula do teste de Moran supõe que não há autocorrelação espacial, ou seja, a distribuição espacial dos valores é aleatória, enquanto a hipótese alternativa indica a presença de autocorrelação espacial.

2.4. Dependência Espacial e Semivariograma

Na análise espacial, a dependência espacial, ou conhecida também como autocorrelação espacial, é associada à geoestatística, que surgiu na África do Sul quando Krige (1951, citado por Grego *et al.* 2014), trabalhava com dados de concentração de ouro. Em seus estudos concluiu que a variância de duas amostras dependia da distância entre elas. Essa autocorrelação espacial refere-se à correlação espacial de uma mesma variável medida em locais distintos do espaço.

Desta forma, a dependência espacial refere-se à ideia de que a proximidade geográfica ou espacial entre as observações pode influenciar os resultados estatísticos. Isso significa que uma observação em um determinado local pode estar

correlacionada com observações em locais próximos, resultando em uma estrutura de correlação espacial. Portanto, considera-se tanto o valor observado quanto a posição espacial do dado. Com base nessa ideia, a geoestatística descreve e modela a relação entre distância e dependência.

O variograma é a representação gráfica da dependência espacial obtido pela variância versus a distância (De Oliveira, 2015). Na literatura, os variogramas são frequentemente referidos como semivariogramas empíricos ou semivariogramas experimentais. Consiste em um gráfico bidimensional, onde o eixo horizontal (eixo x) representa a distância de separação entre os pontos de amostragem e eixo vertical (eixo y) expressa a estimativa da semivariância. Em outras palavras, ele descreve como a variabilidade entre os pares de pontos varia conforme a distância que os separa.

A semivariância entre dois pontos amostrados é determinada de acordo com a equação 2, apresentada a seguir (Druck *et al.*, 2004; Resende *et al.*, 2014):

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

em que $Z(x)$ é o valor observado do ponto amostrado em x e $Z(x+h)$ o valor observado na posição $x+h$. Conforme destacado por Vieira (2000), bem como por Grego *et al.* (2014) e Resende *et al.* (2014) é possível realizar estimativas dos valores das semivariâncias representadas no semivariograma utilizando a equação 3:

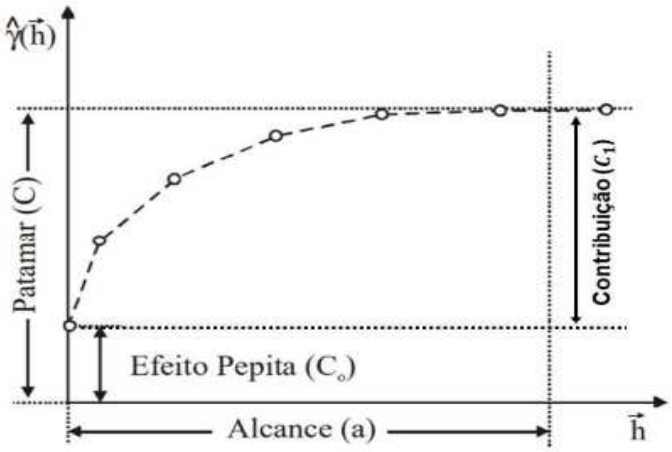
$$\hat{\gamma}(h) = \frac{1}{2\,N(h)} \sum_{i=1}^{N(h)} \{[Z(xi+h) - Z(xi)]^2\} \quad (3)$$

sendo $N(h)$ o número de pontos amostrados separados por uma distância h.

Essa função gera os pontos que serão plotados no semivariograma. A visualização desses pontos permite identificar padrões na variabilidade espacial. Por exemplo, se a semivariância aumenta com a distância, isso sugere uma dependência espacial, indicando que valores semelhantes estão mais próximos uns dos outros. Sendo assim, ele permite verificar e modelar a dependência existente.

A Figura 4 representa as características fundamentais a serem consideradas em um semivariograma experimental com efeito pepita. O patamar (C) é o valor para o qual a semivariância converge à medida que a distância (h) se aproxima do alcance (a), onde a dependência espacial é significativa, a partir desse ponto a semivariância atinge uma estabilização, indicando independência espacial. Desse modo, o alcance (a) representa a distância dentro da qual as amostras estão espacialmente correlacionadas, enquanto o patamar (C) indica a máxima variabilidade observada (Camargo *et al.*, 2004).

Figura 4: Exemplo de semivariograma com os parâmetros estimados em evidência, incluindo o efeito pepita.



Fonte: Camargo *et al.*, 2004. p.96

À medida que a distância tende a zero, espera-se que o valor da semivariância também tenda a zero, entretanto, ocasionalmente isso não ocorre, indicando assim o surgimento de um novo parâmetro chamado de efeito pepita (C_0). Esse parâmetro revela a descontinuidade do semivariograma para distâncias menores que a menor distância entre as amostras. O efeito pepita está associado a variações aleatórias, resultantes de fatores não controláveis e erros experimentais, conforme destacado por Yamamoto e Landim (2015).

Com o efeito pepita, o patamar passa a ser a soma de dois fatores: C_0 e C_1, sendo o segundo conhecido como contribuição. O parâmetro C_1 representa a contribuição associada ao fenômeno espacial em análise, a qual, somada à

contribuição aleatória refletida pelo efeito pepita, é responsável por conduzir o variograma ao seu patamar.

Ao dispor do variograma experimental, torna-se necessário proceder com o ajuste de um modelo teórico representado por uma curva conhecida para se estimar o semivariograma teórico. Os modelos teóricos mais comumente empregados são os modelos esférico, o exponencial e o gaussiano (Vieira, 2000), cujas equações são fornecidas na mesma ordem a seguir, e a Figura 5 exibe a representação desses modelos teóricos:

$$\gamma(h) = \begin{cases} C_0 + C_1\left[\frac{3}{2}\left(\frac{h}{a}\right) - \frac{1}{2}\left(\frac{h}{a}\right)^3\right], & se\ 0 < h < a \\ C_0 + C_1\ , & se\ h \geq a \end{cases} \quad (4)$$

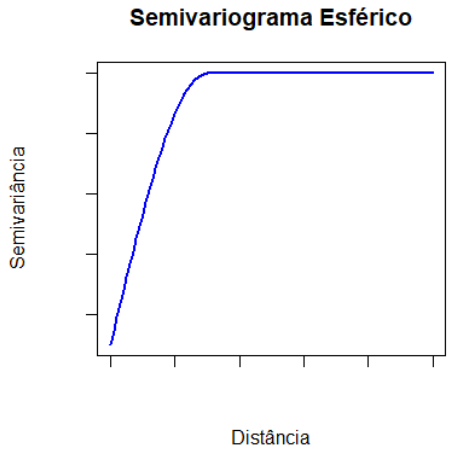
$$\gamma(h) = C_0 + C_1\left[1 - exp\left(-\frac{3h}{a}\right)\right] \quad (5)$$

$$\gamma(h) = C_0 + C_1\left[1 - exp\left(-\frac{3h^2}{a^2}\right)\right] \quad (6)$$

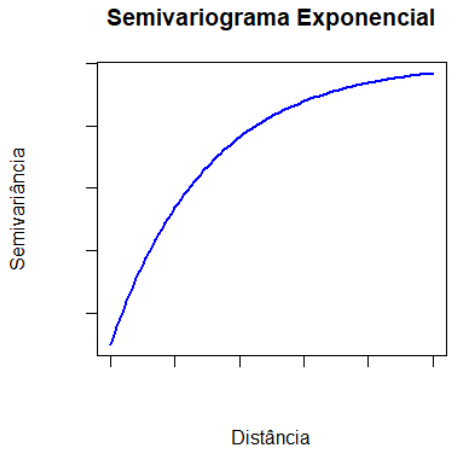
O modelo sem componentes espaciais é equivalente ao modelo de efeito pepita puro, utilizado para descrever a variabilidade espacial de um fenômeno quando não há evidências de uma estrutura de dependência espacial significativa. De acordo com Rodrigues *et al.* (2023) o efeito pepita puro ocorre quando a distribuição da variável na área é aleatória, sem padrões de correlação espacial identificáveis, ou quando distância mínima entre os pontos amostrais é superior à distância da dependência espacial. O semivariograma correspondente a esse modelo é caracterizado por uma constante não nula no eixo das ordenadas e um alcance imediato no eixo das abcissas, sem apresentar o aumento gradual esperado em casos de dependência espacial.

Figura 5: Comportamento dos principais modelos de semivariogramas teórico: (a) esférico, (b) exponencial e (c) gaussiano.

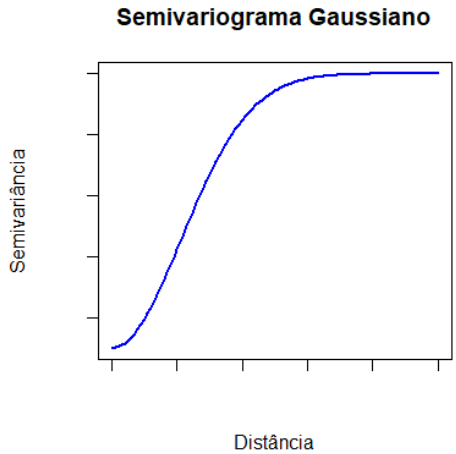
(a)



(b)



(c)



Fonte: Elaborado pela autora

Existem diferentes métodos para ajustar semivariogramas, sendo os mais comuns: métodos de quadrados mínimos ordinários *(Ordinary Least Squares* - OLS*)* e ponderados *(Weight Least Squares* - WLS*)*; métodos da máxima verossimilhança *(Maximum Likelihood* - ML*)* e máxima verossimilhança restrita *(Restricted Maximum Likelihood* - REML*)*. Os métodos OLS e WLS envolvem a ajuste dos valores dos parâmetros de um modelo que minimizam a soma dos quadrados das diferenças entre os valores observados e os estimados. Em contrapartida, os métodos ML e REML buscam determinar o estimador mais verossímil dos parâmetros de um modelo probabilístico a partir dos dados, ou seja, que tornem os dados observados mais prováveis. (Ferreira, 2020).

2.5. Validação dos Modelos

A validação visa avaliar a performance do modelo em prever dados já observados, garantindo sua capacidade de generalização. Conforme destacado por Vieira (2000) e citado por Hernández (2021) é necessária a realização da validação cruzada para avaliar a qualidade do ajuste do modelo teórico.

Uma metodologia amplamente utilizada para validação de modelos preditivos é a *Leave-One-Out Cross-Validation* (LOOCV). Nesse método, cada observação do conjunto de dados é usada uma vez como conjunto de teste, enquanto o restante das observações é utilizado para treinar o modelo, estimando o valor da observação retirada. Ou seja, o modelo é treinado em todos os dados, exceto uma única observação, que será prevista. Esse procedimento é repetido para cada observação no conjunto de dados, permitindo obter tanto o valor real quanto a estimativa correspondente para cada uma delas, o que possibilita o cálculo do erro de estimação. Como descrito por Syed (2011), esse processo de validação envolve a repetição contínua em que cada observação, uma de cada vez, serve como conjunto de teste, garantindo que o modelo seja avaliado em todas as amostras disponíveis.

Na validação deve-se ter média dos erros e dos erros padronizados iguais a zero, variância dos erros finita e variância dos erros padronizados igual a um (Vieira, 2000).

2.6. Critérios para escolha do Modelo

A escolha do melhor modelo em análises estatísticas é fundamental para garantir a precisão e a eficácia das previsões. Para esse propósito, diversos critérios podem ser utilizados, entre os quais se destacam o Teste de Razão de Verossimilhanças - *Likelihood Ratio Test - (LRT*), o Critério de Informação de Akaike - *Akaike Information Criterion - (AIC)* e a Raíz do Erro Quadrático Médio - *Root Mean Square Error - (RMSE)*.

2.6.1. Teste de Razão de Verossimilhanças

O LRT é utilizado para comparar a qualidade de ajuste entre dois modelos aninhados, ou seja, modelos onde um é uma versão restrita do outro. Permite verificar se a inclusão de parâmetros adicionais proporciona um ajuste significativamente melhor. Segundo Duarte (2000), o LRT permite testar estatisticamente a diferença de adequação dos dois modelos. A estatística do teste é calculada da seguinte forma:

$$LRT = -2[log\ L(\theta_0) - log\ L(\theta_1)] \quad (7)$$

Sendo $L(\theta_0)$ a função de verossimilhança maximizada do modelo reduzido (modelo nulo) e $L(\theta_1)$ a função de verossimilhança maximizada do modelo completo.

Quando se considera o teste de razão de verossimilhanças para modelos espaciais, as hipóteses envolvem os parâmetros específicos desses modelos. Podemos, por exemplo, testar a significância do alcance (a) e da contribuição (C_1). Posto isto, na hipótese de nulidade (H_0) temos que o modelo reduzido, que geralmente restringe alguns parâmetros espaciais, é suficiente para explicar a estrutura dos dados espaciais, isso implica que os parâmetros espaciais não são significativos ou que a estrutura espacial pode ser explicada por um modelo mais simples. E na hipótese alternativa (H_a) o modelo completo, que inclui os parâmetros espaciais sem restrições, é necessário para explicar a estrutura dos dados, isso implica que a dependência espacial é significativa e deve ser modelada. Além disso a estatística do LRT segue uma distribuição qui-quadrado com graus de liberdade iguais à diferença no número de parâmetros entre os dois modelos (FERES, 2009).

Utilizar esse teste em modelos espaciais permite avaliar a importância da estrutura espacial dos dados e selecionar o modelo mais apropriado para capturar essa dependência, melhorando assim a precisão e a validade das inferências.

2.6.2. Critério de Informação de Akaike

O AIC (Akaike, 1974), baseado no estudo da função de verossimilhança e no número de parâmetros de cada modelo, é utilizado para avaliar a qualidade relativa de modelos estatísticos aplicados a um determinado conjunto de dados. Este critério penaliza a complexidade do modelo a fim de evitar o sobreajuste, sendo essa complexidade associada ao número de parâmetros do modelo. Desse modo, esse critério favorece o modelo que equilibre adequadamente a complexidade e a capacidade de generalização, isto é, equilíbrio entre ajuste e simplicidade. A fórmula do Critério de Informação de Akaike é dada por:

$$AIC = -2log\ L(\hat{\theta}) + 2k \tag{8}$$

sendo k o número de parâmetros no modelo e $L(\hat{\theta})$ a função de verossimilhança maximizada (Moura, 2021).

Conforme citado por Moura (2021), o modelo selecionado dentre um conjunto de modelos concorrentes é aquele que minimiza o valor calculado do critério.

2.6.3. Raíz do Erro Quadrático Médio

O RMSE é uma medida da magnitude média dos erros de previsão amplamente utilizado para avaliar a precisão de modelos preditivos. Esse critério fornece uma medida da média dos desvios ao quadrado entre os valores observados (y_i) e preditos $(\hat{y}_i)$, sendo sensível a grandes erros. Sua fórmula é dada por (Hodson, 2022):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{9}$$

Na escolha do modelo, um RMSE mais baixo indica uma melhor capacidade preditiva, pois sugere que as previsões do modelo estão mais próximas dos valores reais.

A utilização combinada desses critérios proporciona uma visão abrangente na escolha do melhor modelo. O LRT permite verificar a significância estatística da inclusão de novos parâmetros, o AIC ajuda a balancear ajuste e complexidade, e o RMSE fornece uma medida direta da precisão preditiva. Essa abordagem integrada é essencial para a construção de modelos robustos e eficazes em diversas aplicações analíticas.

2.7. Índice de dependência espacial

O Índice de Dependência Espacial (IDE) constitui uma medida empregada na quantificação da dependência espacial após a estimação dos parâmetros de um modelo teórico espacial. No cálculo do IDE, determina-se a proporção da variância estrutural em relação à variância total, isto é, realiza-se o quociente entre contribuição (C_1) e patamar ($C_0 + C_1$) como é apresentado na equação a seguir:

$$IDE = \left(\frac{C_1}{C_0 + C_1}\right) 100 \tag{10}$$

Conforme classificação proposta por Zimback (2001), tem-se:

- $IDE \leq 25\%$ implica em dependência espacial fraca;
- $25\% < IDE < 75\%$ implica em dependência espacial moderada;
- $IDE \geq 75\%$ implica em dependência espacial forte.

2.8. Modelos para erros dependentes e independentes

A análise tradicional de experimentos de campo pressupõe que todas as observações são não correlacionadas, de modo a atender às pressuposições da análise de variância. Entretanto, em muitos estudos que envolvem dados espaciais, é fundamental levar em conta a correlação entre as observações para garantir

estimativas precisas dos parâmetros do modelo. Através da análise espacial é possível incorporar diferentes estruturas de covariância aos resíduos, assim como feito por Duarte (2000) para competição de linhagens de soja, Maia *et al.* (2013) para a seleção de clones de laranjeira Pêra, Silva (2020) para avaliação da produtividade de famílias de feijão comum e Salvador *et al.* (2022) na avaliação de dados do melhoramento genético do feijoeiro.

Para a análise dos dados considerando os erros independentes adotou-se os seguintes modelos:

Modelo com controle local do delineamento em blocos casualizados

$$y_{ij} = \mu + b_j + t_i + e_{ij} \quad (11)$$

Modelo sem o controle local

$$y_{ij} = \mu + t_i + e_{ij} \quad (12)$$

em que:

y_{ij} é a média de TCH para o tratamento i (i=1,2, ..., f) no bloco j (j=1, 2, ..., b);

μ é a constante associada a todas observações;

b_j é o efeito do bloco j;

t_i é o efeito da família i;

e_{ij} são os erros aleatórios associados às observações, assumindo independência entre os erros, ou seja, $e \sim N(0, R)$ sendo $R = I\sigma^2$.

A matriz R para erros independentes é formada pela multiplicação da variância dos erros σ^2 pela matriz identidade I. A matriz identidade consiste em uma matriz cuja diagonal principal é composta por elementos iguais a 1, enquanto as demais entradas são zero. Isso representa que a covariância entre diferentes erros é nula.

Para a análise considerando os erros autocorrelacionados espacialmente, tem-se a mesma configuração dos modelos citados anteriormente, mas com a incorporação da estrutura de covariância associada aos erros. Segundo Campos *et al.* (2016), na análise espacial utiliza-se uma matriz não diagonal de variâncias e covariâncias residuais, definida a partir de modelos geoestatísticos que representam funções de covariâncias, descrevendo a dependência espacial dos erros em função

da distância entre as parcelas. Assim, tem-se $e \sim N(0, R)$, sendo R a matriz de variância-covariância residual que assume diferentes estruturas de covariância como apresentado por Silva *et al.* (2004), Reis e Miranda Filho (2003), Maia *et al.* (2013) e Campos *et al.* (2016).

A estrutura da matriz de variâncias-covariâncias para erros espacialmente correlacionados é dada por (Reis e Miranda Filho, 2003; Silva *et al.*, 2004):

$$R = Cov(e_i, e_{i'}) = \sigma^2[f(h)] \quad , \qquad se\ h > 0 \tag{13}$$

Sendo $f(h)$ a função que descreve a forma da covariância em função da distância h entre dois erros de parcelas, ela está associada ao variograma modelado para capturar a dependência espacial entre os resíduos. Assim, cada elemento da matriz R é definido de acordo com a dependência espacial estimada pela função de autocovariância $C(h)$.

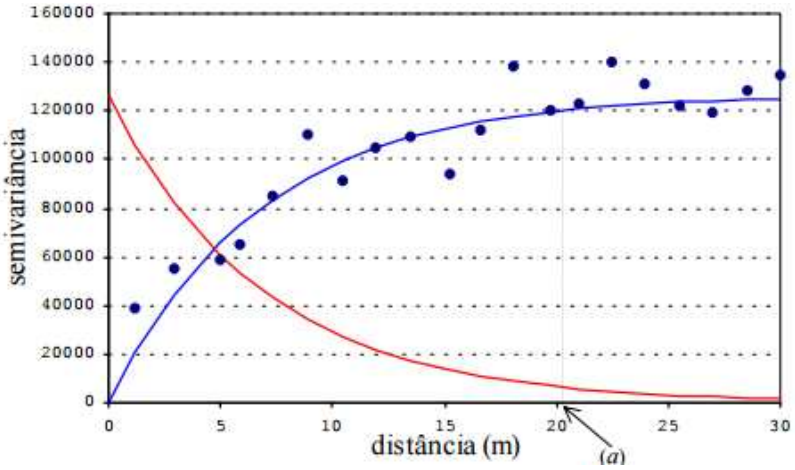
A função de autocovariância, é uma medida estatística que descreve a covariância entre valores de uma variável aleatória em diferentes pontos no espaço em função da distância. Em termos simplificados, ela quantifica como a covariância entre duas observações de uma mesma variável diminui conforme a distância entre essas observações aumenta. Dessa forma, quando $h = 0$, estamos considerando dois erros na mesma localização, ou seja, o mesmo erro. Nesse caso, a função de autocovariância se reduz à variância desse ponto, uma vez que a covariância de um ponto consigo mesmo é, por definição, igual à sua variância.

De acordo com Duarte (2000) uma das vantagens de avaliar a dependência espacial por meio do variograma é que, sob a suposição de estacionaridade, existe uma relação direta entre o variograma e a função de autocovariância $C(h)$ (Figura 6):

$$\gamma(h) = \sigma^2 - C(h) \tag{14}$$

sendo $\sigma^2 = C(h = 0)$.

Figura 6: Representação ilustrativa da função de semivariância (linha azul) e covariância (linha vermelha).



Fonte: Duarte (2000).

Observa-se que, a partir de uma determinada distância (a), a semivariância atinge um patamar, indicando que a variabilidade entre parcelas se estabiliza. Esse patamar reflete a variabilidade própria dos resíduos entre parcelas que são independentes ou que estão separadas por uma distância maior ou igual ao alcance. No gráfico, nota-se que a covariância decresce com o aumento da distância, evidenciando que as observações mais próximas tendem a ser mais semelhantes, enquanto a variância aumenta à medida que a distância entre os pontos aumenta.

A seguir, são apresentadas as estruturas de covariância espacial utilizando as funções de autocovariância mais comumente empregadas (Resende *et al.*, 2014):

Modelo esférico

$$C(h) = \begin{cases} \sigma^2 \left[1 - \frac{3}{2}\left(\frac{h}{a}\right) + \frac{1}{2}\left(\frac{h}{a}\right)^3\right], & se\ h < a \\ 0 \qquad\qquad\qquad , & se\ h \geq a \end{cases} \tag{15}$$

Modelo exponencial

$$C(h) = \sigma^2 \left[exp\left(-\frac{3h}{a}\right)\right] \tag{16}$$

Modelo gaussiano

$$C(h) = \sigma^2 \left[exp\left(-\frac{3h^2}{a^2}\right)\right] \tag{17}$$

3. MATERIAIS E MÉTODOS

3.1. Dados experimentais

Foram utilizados dados de produção de cana-de-açúcar provenientes de dois experimentos conduzidos pelo Programa de Melhoramento Genético da Cana-de-Açúcar (PMGCA) da UFV e pela Rede Interuniversitária para o Desenvolvimento do Setor Sucroenergético (RIDESA) no Centro de Pesquisa e Melhoramento de Cana-de-Açúcar (CECA), localizado no município de Oratórios, MG (latitude 20°25' S, longitude 42°48' W e 494 m de altitude). Ambos os experimentos fazem parte da fase inicial do programa de melhoramento genético.

No experimento D_1, instalado segundo o delineamento em blocos casualizados (DBC), foram avaliadas 98 famílias de cana-de-açúcar com quatro repetições em um campo experimental de 98 metros de comprimento por 39,2 metros de largura. Entretanto foram perdidas informações de quatro parcelas, assim a análise foi feita com 388 observações. As parcelas do campo experimental apresentavam 3,5 metros de comprimento por 2,8 metros de largura e foram constituídas por dois sulcos de 1,5 metros de comprimento, espaçados a 1,4 metros sendo cada sulco composto por doze plantas. De forma análoga, no experimento D_2 foram avaliadas 60 famílias de cana-de-açúcar em quatro blocos em um campo experimental de 80 metros de comprimento por 42 metros de largura, deste modo, totalizando 240 parcelas. As parcelas apresentavam dimensões de 5 metros de comprimento por 2,8 metros de largura, compostas por dois sulcos de 4,5 metros de comprimento, espaçados em 1,4 metros, com dez plantas em cada sulco.

Foram calculados o peso médio dos colmos das parcelas, sendo posteriormente convertidos em toneladas de colmo por hectare (TCH), ou toneladas de cana-de-açúcar por hectare, unidade utilizada para avaliar a produtividade da cana-de-açúcar. Essa conversão foi realizada de acordo com a equação 14, conforme citado Ferreira et al. (2022):

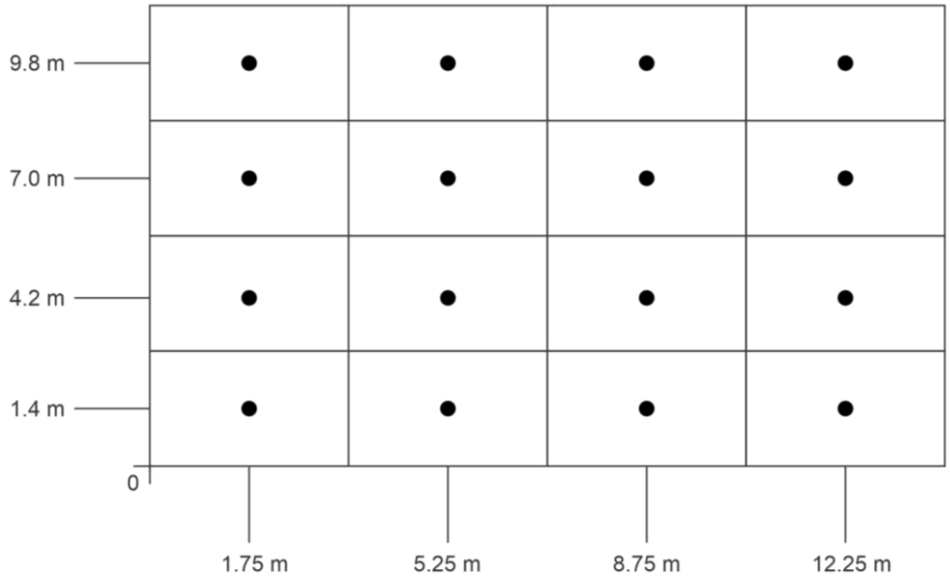
$$TCH = \frac{10(MMC)(NCM)}{TP} \tag{18}$$

em que MMC é a massa média dos colmos, NCM o número de colmos por metro, 10 uma constante de transformação de unidade e TP o tamanho da parcela em m^2.

Foram utilizadas as informações do tamanho das parcelas, comprimento e largura, para determinar as coordenadas cartesianas das observações. Como é de interesse apenas a distância entre as observações foi considerada a posição como sendo o centro da parcela. Essas localizações foram calculadas manualmente fixando um eixo *xy* no croqui do experimento, considerando o vértice inferior esquerdo do campo como a origem do eixo. Assim, para as parcelas do campo experimental no experimento D_1, a primeira parcela, situada no vértice inferior esquerdo, foi atribuída às coordenadas 1,75 para x e 1,4 para y. No experimento D_2, a primeira parcela recebeu as coordenadas 2,5 e 1,4 para os eixos x e y, respectivamente. A Figura 7 ilustra a disposição das parcelas e como essas coordenadas foram atribuídas.

Todas as análises realizadas neste trabalho foram executadas no software R (R Core Team, 2024).

Figura 7: Desenho esquemático ilustrando a disposição parcial das parcelas no campo experimental do experimento D_1, de acordo com as coordenadas atribuídas.



Fonte: Elaborado pela autora

3.2. Análises estatísticas

Segundo Druck *et al*. (2004), a investigação dos resíduos da regressão em busca de sinais de estrutura espacial é o primeiro passo em uma regressão espacial. Para esta análise, os resíduos, considerados como variável dependente, foram estimados de duas maneiras distintas. Inicialmente, considerou-se o controle local, correspondente ao delineamento em blocos casualizados (DBC), utilizado na condução dos experimentos. Em seguida, foram estimados os resíduos sem o controle local, com base em um modelo ajustado pelo delineamento inteiramente casualizado (DIC).

A análise via DIC teve por objetivo capturar a variância no campo experimental pelos modelos espaciais. Em ambos os modelos a variável dependente utilizada para estimar os resíduos para a análise foi a quantidade de toneladas de cana-de-açúcar por hectare (TCH).

Foram investigadas as suposições dos resíduos estimados. Para avaliar a normalidade dos resíduos, utilizou-se o teste de Shapiro-Wilk e a homocedasticidade foi analisada por meio do teste de Bartlett. Além disso, o Índice de Moran foi empregado para verificar a presença de autocorrelação entre os resíduos.

Para o cálculo do Índice de Moran, adotou-se uma metodologia baseada nos K-vizinhos mais próximos e no inverso da distância para a determinação dos pesos. Conforme Almeida (2012), a principal vantagem dessa abordagem reside no fato de que cada observação é atribuída ao mesmo número de vizinhos, garantindo que não existam regiões isoladas sem vizinhos. Inicialmente, identificaram-se os quatro vizinhos mais próximos para cada ponto, utilizando a distância euclidiana como critério. Posteriormente, foram atribuídos pesos a essas relações de vizinhança, com base no inverso da distância entre os pontos, de modo que os pontos mais próximos tivessem maior influência. Essa abordagem permite capturar de forma eficaz a autocorrelação espacial ao considerar tanto a proximidade quanto a intensidade da relação entre os pontos.

3.2.1. Avaliação da dependência espacial dos resíduos

Inicialmente, realizou-se uma análise espacial exploratória para identificar padrões e características dos resíduos. Um mapa foi criado para plotar os dados conforme sua magnitude, utilizando diferentes tamanhos e cores de pontos. Dessa forma, foi possível visualizar se os dados dos pontos vizinhos apresentavam similaridades. Além disso, um *box-plot* foi gerado para detectar possíveis outliers.

Posteriormente foram aplicadas técnicas de geoestatística para modelar a dependência espacial presente entre os resíduos. Construiu-se o variograma experimental omnidirecional utilizando a função *variog( )* da biblioteca *geoR* (RIBEIRO JR et al., 2024) e ajustou-se os parâmetros do semivariograma teórico pelo método de estimação por máxima verossimilhança por meio da função *likfit( )*, também da biblioteca *geoR*. No presente trabalho, foram analisados três tipos de modelos teóricos – Esférico, Exponencial e Gaussiano – para determinar qual se ajustaria melhor aos dados. Segundo Diggle e Ribeiro Jr (2007) o variograma experimental não é estritamente necessário para o método da máxima verossimilhança, mas, no entanto, fornece uma maneira útil de especificar valores iniciais para o algoritmo. Mas para identificar a dependência espacial e incorporar a covariância existente entre os erros no modelo, a análise do variograma é essencial, pois permite identificar o comportamento da covariância espacial em função da distância. Essa análise indica qual modelo de covariância espacial é mais adequado para os dados, orientando a escolha da estrutura de dependência a ser utilizada no modelo de regressão.

A qualidade do ajuste do modelo foi avaliada utilizando a técnica de validação cruzada, implementada através da função *xvalid( )* do pacote *geoR* (RIBEIRO JR *et al.*, 2024). Foram obtidas as seguintes métricas: média dos erros e erros padronizados próxima de zero, variância dos erros finita, e variância dos erros padronizados igual a um. A escolha do melhor modelo espacial foi feita com base nos critérios de menor AIC ou menor RMSE, visando selecionar o modelo mais adequado em termos de ajuste e precisão preditiva.

Após o ajuste do modelo, foi realizado o teste da razão de verossimilhança (LRT) para testar se, estatisticamente, os parâmetros espaciais diferiam de zero. Através do summary gerado pela função likfit(), foram obtidos o valor da verossimilhança maximizada e o número de parâmetros estimados para o modelo

espacial, bem como para o modelo sem componente espacial. Essa abordagem permitiu verificar a significância dos parâmetros espaciais contribuição (C_1) e alcance (a) no contexto do modelo ajustado.

Ademais, para aqueles modelos em que os componentes espaciais foram significativos foi calculado o Índice de Dependência Espacial (IDE) proposto por Zimback (2001) a fim de determinar o grau da intensidade da dependência espacial presente entre os resíduos dos experimentos.

3.2.2. Avaliação das famílias de cana-de-açúcar

Após o ajuste dos modelos espaciais e a identificação da melhor estrutura de covariância para os resíduos, procedeu-se à análise estatística das famílias de cana-de-açúcar utilizando três modelos distintos:

(i) Modelo estatístico com controle local, baseado no delineamento em blocos casualizados, assumindo erros independentes;

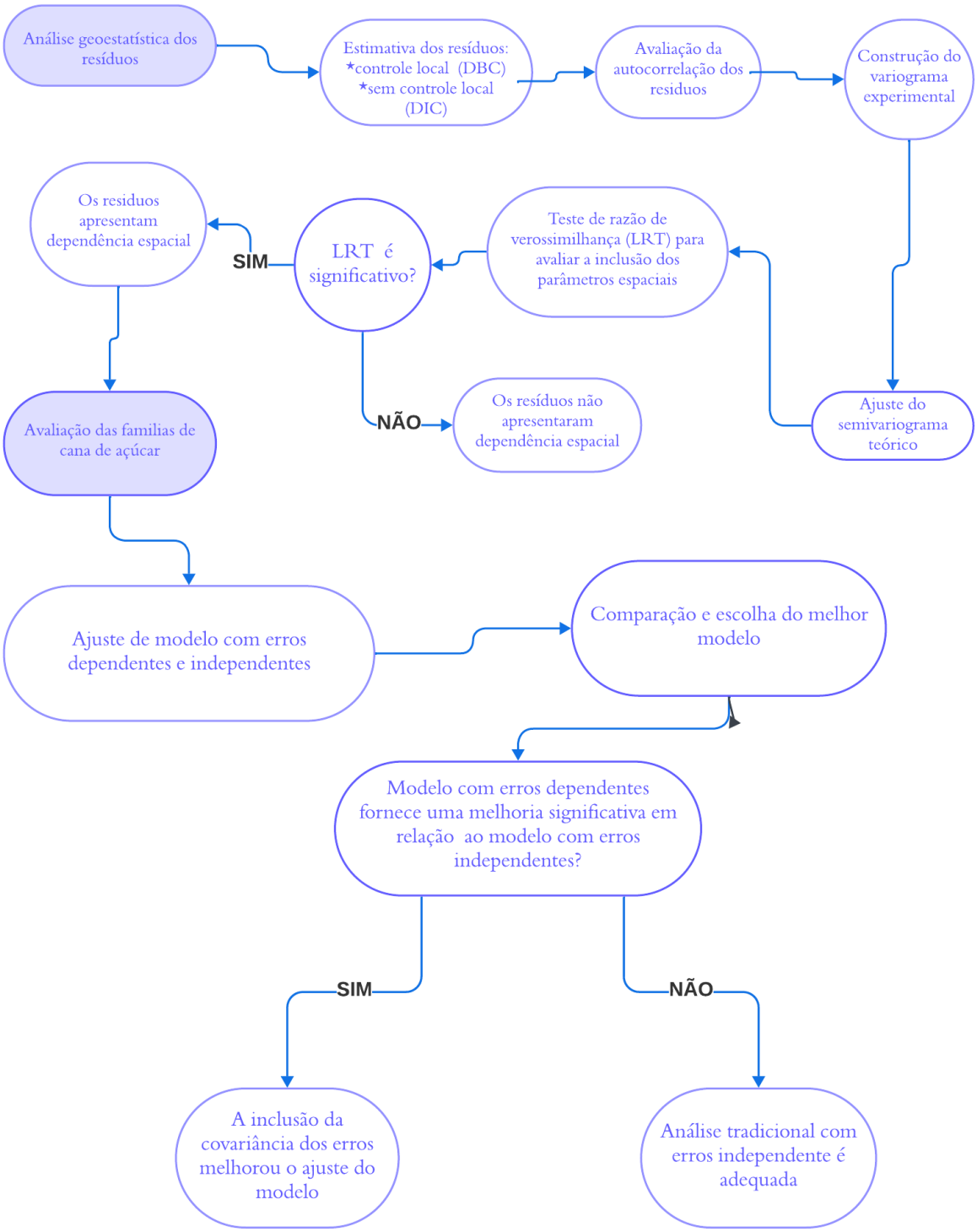
(ii) Modelo estatístico com controle local, também no delineamento em blocos casualizados, mas considerando a autocorrelação espacial dos erros;

(iii) Modelo estatístico sem controle local, utilizando o delineamento inteiramente casualizado, no qual os erros são autocorrelacionados espacialmente.

No modelo (iii) foi desconsiderada a presença dos blocos para tentar explicar o controle da variação sistemática apenas pela análise espacial. Todos os modelos foram ajustados utilizando a função *gls( )* referente a biblioteca *nlme* (PINHEIRO e BATES, 2024) que possibilita a incorporação da estrutura de covariância dos erros no modelo.

A escolha da estrutura de covariância espacial foi baseada no modelo de variograma ajustado. Inicialmente, definiu-se a estrutura de covariância espacial utilizando as funções de autocovariância disponíveis na biblioteca *nlme*. As funções utilizadas foram *corExp*, *corGaus* ou *corSpher*, dependendo do variograma previamente ajustado. Em seguida, ajusta-se o modelo de regressão incorporando essa estrutura pelo argumento '*correlation'* da função *gls( )*.

Figura 8: Fluxograma das etapas envolvidos na análise.



Fonte: Elaborado pela autora

4. RESULTADOS E DISCUSSÃO

A seguir, as figuras e tabelas apresentam os resultados da avaliação da estrutura de dependência espacial entre os resíduos dos dois experimentos estimados nas análises, considerando os blocos do experimento e a ausência deles. Além disso, são apresentados os modelos com a incorporação da covariância espacial dos resíduos, quando tal estrutura foi observada na análise dos resíduos.

Na análise dos resíduos, constatou-se por meio dos testes de Shapiro-Wilk e Bartllet, ao nível de significância de 1%, que eles não seguiram uma distribuição normal e não apresentaram homogeneidade de variância em ambos os experimentos realizados utilizando os delineamentos em blocos casualizados (DBC) e inteiramente casualizado (DIC). Esse resultado foi atribuído à presença de valores atípicos, os *outliers*, os quais, ao serem removidos da análise, resultaram em resíduos que atenderam aos pressupostos de normalidade e homogeneidade de variância, conforme apresentado na Tabela 1. Entretanto, como o objetivo do trabalho é analisar a precisão experimental e não a realização de testes de comparação de médias, optou-se pela não exclusão das parcelas que apresentaram esses valores, assim como feito por Ferreira (2020), que destaca que a presença de outliers para o ranqueamento das famílias pode ser um indicativo importante, uma vez que esses valores podem estar associados a anomalias nas parcelas experimentais, como a mortalidade de indivíduos.

Tabela 1: Testes estatísticos de normalidade (Teste Shapiro-Wilk - SW) e homogeneidade de variância (Teste de Bartlett) para os resíduos estimados através do modelo do DIC e do DBC.

		Sem exclusão das parcelas		Com exclusão das parcelas	
Experimento	Delineamento	SW	Bartlett	SW	Bartlett
D_1	DIC	< 0,0001	< 0,0001	0,2763	0,0491
	DBC	0,0001	0,0068	0,159	0,3754
D_2	DIC	< 0,0001	0,0003	0,0160	0,0326
	DBC	< 0,0001	0,0026	0,0108	0,0907

4.1. Análise geoestatística dos resíduos do experimento D_1

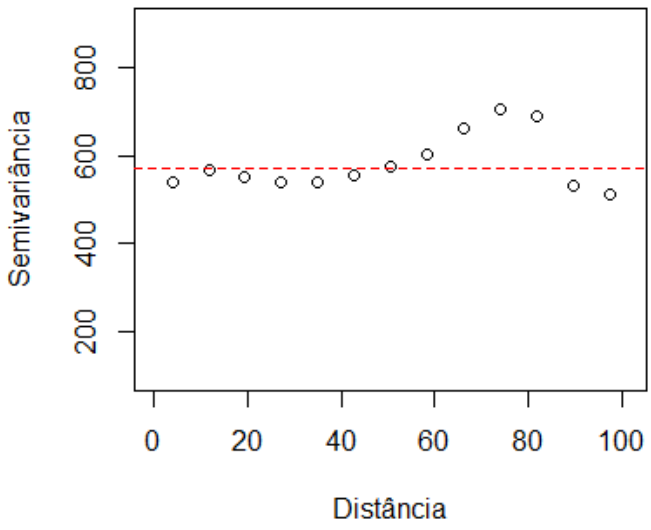
A análise da autocorrelação espacial dos resíduos, realizada por meio do índice de Moran, confirmou as observações iniciais feitas durante a análise exploratória, nas quais os mapas com os dados de resíduos plotados revelaram padrões espaciais. Para os resíduos estimados considerando DBC, a estatística do índice de Moran foi de 0,0613 com p-valor de 0,0363. Logo, ao nível de significância de 1% o índice não foi significativo, sugerindo que os resíduos deste experimento não estão correlacionados. Em relação aos resíduos estimados pelo DIC, a estatística do índice de Moran foi de 0,1822 com p-valor < 0,0001, indicando a existência de estrutura espacial dos resíduos.

O semivariograma empírico dos resíduos foi calculado para ambos os delineamentos, DBC e DIC, a fim de se observar a estrutura de semivariância deles ao longo das distâncias (Figura 9). Os resultados indicam que, para o DIC, observa-se que a semivariância dos resíduos varia à medida que a distância entre as parcelas aumenta, o que sugere uma dependência espacial significativa. Essa dependência é verificada pelos cálculos realizados com o índice de Moran. No caso do DBC, o semivariograma não revelou uma contribuição tão evidente, também em conformidade com o resultado obtido anteriormente.

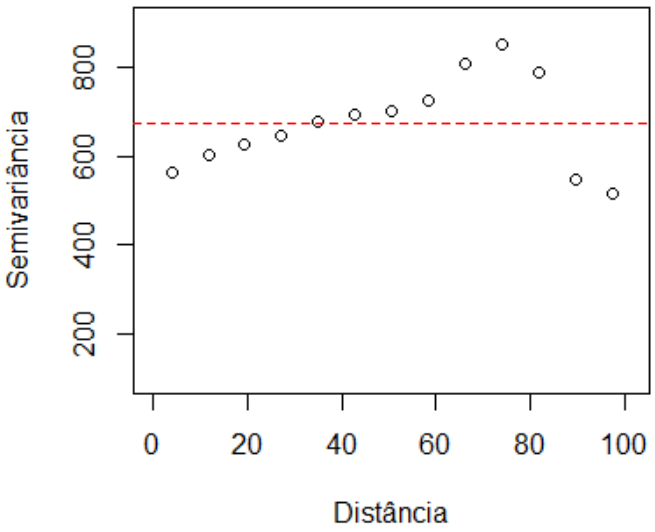
Embora o índice de Moran e o semivariograma empírico (Figura 9) indicassem a ausência de dependência espacial para os resíduos estimados pelo DBC, foi realizado o ajuste de um modelo geoestatístico para ambos semivariogramas.

Figura 9: Semivariogramas empíricos dos resíduos estimados: análise em DBC (a) e DIC (b) para o experimento D_1, com variância dos resíduos representada em linha tracejada vermelha.

(a)



(b)



Para os resíduos estimados considerando o modelo de blocos casualizados, o modelo teórico esférico obteve o melhor ajuste. Os demais modelos geoestatísticos ajustaram apenas o parâmetro efeito pepita (C_0), o que indica que não houve correlação espacial significativa entre as observações. Isso sugere que a variabilidade dos dados ocorre de forma aleatória, sem um padrão de continuidade espacial. O modelo esférico foi ajustado e os seguintes parâmetros foram estimados pelo método da verossimilhança: 57,24; 13,39 e 514,37, respectivamente, contribuição (C_1), alcance (a) e efeito pepita (C_0). Entretanto, ao realizar o teste de razão de

verossimilhança (LRT) para avaliar a necessidade de inclusão de parâmetros espaciais no modelo o resultado não foi significativo, com p-valor igual 0,1029. Isso indica que a inclusão desses parâmetros não melhorou o ajuste do modelo de forma significativa.

Para os resíduos estimados com a desconsideração dos blocos, os modelos teóricos exponencial e esférico foram ajustados, já o modelo gaussiano ajustou apenas o parâmetro efeito pepita (C_0). As estimativas dos parâmetros dos modelos ajustados bem como os critérios de ajuste utilizados para a seleção dos modelos geoestatísticos e o teste da razão da verossimilhança (LRT) que compara o modelo espacial em relação ao modelo sem componentes espaciais estão detalhadas na Tabela 2. No modelo sem componentes espaciais, a variabilidade dos dados ocorre de maneira aleatória, sem apresentar padrões de dependência espacial, o que caracteriza o modelo de efeito pepita puro. Os pressupostos para um bom ajuste do modelo, conforme destacado por Vieira (2000), foram atendidos em todos os modelos ajustados. Esses pressupostos incluem a média dos erros e dos erros padronizados iguais a zero, variância dos erros finita e variância dos erros padronizados igual a um.

Tabela 2: Estimativas dos parâmetros contribuição (C_1), alcance (a) e efeito pepita (C_0), Critério de Informação de Akaike (AIC), Raíz do Erro Quadrático Médio (RMSE), estatística do teste da razão da verossimilhança (LRT) para o modelo espacial em relação ao modelo sem componentes espaciais do experimento D_1 na análise com a desconsideração dos blocos.

Modelo	C_1	a	C_0	AIC	RMSE	LRT
Efeito Pepita Puro	-	-	672,73	3631,5	26,00	-
Exponencial	192,73	68,16	528,99	3594,5	24,18	40,92 ** (<0,001)
Esférico	171,38	51,16	542,57	3592,8	24,16	42,61 ** (<0,001)

** Significativo a 0,1% pelo teste da razão de verossimilhança com 2 graus de liberdade.

Pelo LRT, para ambos os modelos se rejeita a hipótese nula, assim, o modelo completo que inclui os parâmetros espaciais sem restrições, é necessário para

explicar a estrutura dos resíduos. De acordo com os resultados obtidos, o modelo esférico, ao desconsiderar o controle local pelos blocos, mostrou-se o mais adequado, apresentando os menores valores para os critérios de ajuste. Os valores estimados para os parâmetros contribuição (C_1), alcance (a) e efeito pepita (C_0) foram, respectivamente, 171,38; 51,16 e 542,57. Dessa forma, os erros das parcelas localizadas a uma distância máxima de 51,2 metros estão correlacionados entre si. Essa informação possibilitou a identificação das parcelas que se encontram dentro do limite de dependência espacial.

Os parâmetros estimados foram usados para o cálculo do Índice de Dependência Espacial indicando dependência espacial fraca com IDE igual a 24%.

4.2. Análise geoestatística dos resíduos do experimento D_2

Analogamente, realizou-se a análise dos resíduos para o experimento D_2. As estatísticas do índice de Moran foram 0,1739 e 0,1843, respectivamente, para os resíduos estimados pelo DBC e pelo DIC. Para ambos, o índice apresentou significância estatística com p-valor < 0,001, indicando evidência contra a hipótese nula de que não há autocorrelação espacial. Portanto, os resíduos estimados tanto na análise em blocos quanto na análise desconsiderando os blocos revelam a presença da autocorrelação espacial.

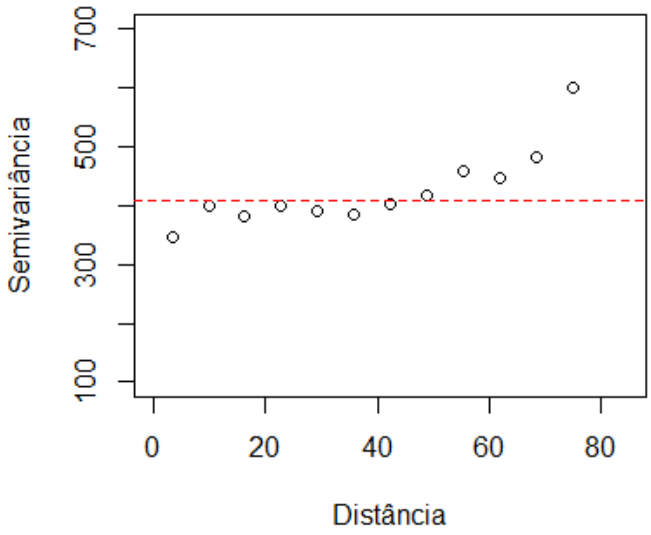
O semivariograma empírico foi calculado para as duas estimativas dos resíduos, tanto para o Delineamento em Blocos Casualizados (DBC) quanto para o Delineamento Inteiramente Casualizado (DIC) (Figura 8). Pode-se notar que a variabilidade dos resíduos muda de forma similar em relação a distância entre os pontos, indicando uma consistência na estrutura de dependência espacial entre os resíduos dos dois delineamentos analisados. Essa semelhança sugere que, independentemente do delineamento analisado, a estrutura espacial dos resíduos permaneceu praticamente inalterada, confirmando, assim, a presença de correlação entre os resíduos pelas duas estimativas.

A configuração semelhante entre os dois semivariogramas pode ser explicada pelo fato de que o efeito do bloco não foi significativo, com p-valor igual a 0,486, na análise de variância, a qual foi realizada inicialmente para identificar as fontes de variação. Esse resultado indica que a contribuição do bloco para a variação total foi

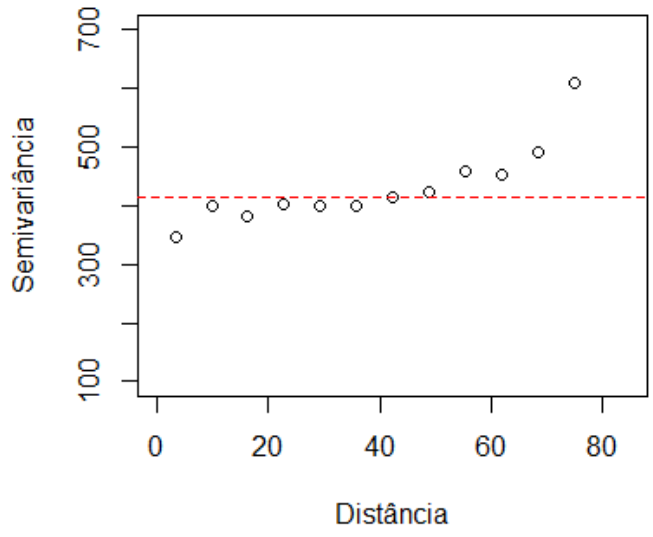
mínima, resultando em uma estrutura espacial dos resíduos que permaneceu praticamente inalterada entre os diferentes delineamentos analisados. Esse comportamento terá reflexo no ajuste do modelo teórico subsequente.

Figura 9: Semivariogramas empíricos dos resíduos estimados: análise em DBC (a) e DIC (b) para o experimento D_2, com variância dos resíduos representada em linha tracejada vermelha.

(a)



(b)



Os três modelos geoestatísticos (exponencial, esférico e gaussiano) ajustaram-se adequadamente ao semivariograma empírico para ambos os delineamentos em que os resíduos foram estimados. Além disso, todos os modelos atenderam aos pressupostos da validação cruzada, apresentaram média dos erros e dos erros

padronizados próximas zero, variância dos erros finita e variância dos erros padronizados próximas a um, conforme apresentado por Duarte (2000).

As tabelas 3 e 4 apresentam as estimativas dos parâmetros ajustados para os modelos teóricos, os valores calculados dos critérios de ajuste utilizados para a seleção do melhor modelo e o valor do teste de razão de verossimilhança (LRT) que compara o modelo espacial em relação ao modelo sem os componentes espaciais (efeito pepita puro).

Tabela 3: Estimativas dos parâmetros contribuição (C_1), alcance (a) e efeito pepita (C0), Critério de Informação de Akaike (AIC), Raíz do Erro Quadrático Médio (RMSE), estatística do teste da razão da verossimilhança (LRT) para o modelo espacial em relação ao modelo sem componentes espaciais do experimento D_2 na análise com blocos.

Modelo	C_1	a	C_0	AIC	RMSE	LRT
Efeito Pepita Puro	-	-	405,80	2126,5	20,23	-
Exponencial	180,57	10,64	227,95	2118,9	19,34	11,636* (0,002)
Esférico	86,88	58,12	359,12	2123,5	19,61	6,979 (0,03)
Gaussiano	111,45	9,68	298,15	2117,7	19,24	12,807* (0,001)

* Significativo a 1% pelo teste da razão de verossimilhança com 2 graus de liberdade.

Todos os modelos apresentaram valores próximos para os critérios de ajuste adotados. No entanto, pelo LRT, apenas os parâmetros espaciais dos modelos exponencial e gaussiano foram significativos. Dentre os modelos, o gaussiano apresentou o melhor ajuste, com os menores valores de AIC e RMSE, e parâmetros estimados para C, a e C_0 de 111,45; 9,68 e 298,15, respectivamente. Assim, as parcelas localizadas dentro de um raio de 9,7 metros no campo experimental de D_2, que possui dimensões de 80 metros por 42 metros, estão correlacionadas

Os parâmetros estimados, levando em consideração o delineamento em blocos casualizados, foram utilizados para o cálculo do Índice de Dependência Espacial,

resultando em um IDE igual a 27%, o qual implica em dependência espacial moderada conforme indicado por Zimback (2001).

Tabela 4: Estimativas dos parâmetros contribuição (C_1), alcance (a) e efeito pepita (C0), Critério de Informação de Akaike (AIC), Raíz do Erro Quadrático Médio (RMSE), estatística do teste da razão da verossimilhança (LRT) para o modelo espacial em relação ao modelo sem componentes espaciais do experimento D_2 na análise com a desconsideração dos blocos.

Modelo	C_1	a	C_0	AIC	RMSE	LRT
Efeito Pepita Puro	-	-	411,66	2129,9	20,37	-
Exponencial	178,27	11,94	236,96	2120,4	19,36	13,493* (0,001)
Esférico	78,20	29,09	346,66	2123,0	19,58	10,935* (0,004)
Gaussiano	118,15	9,91	297,79	2119,4	19,27	14,511* (<0,01)

* Significativo a 1% pelo teste da razão de verossimilhança com 2 graus de liberdade.

A estatística do teste de razão de verossimilhança (LRT) apresentou significância com p<0,01 para os três modelos ajustados. Esses resultados indicaram que o modelo completo, que incorpora os parâmetros espaciais, é necessário para a adequada explicação da estrutura dos resíduos. De acordo com os critérios de ajuste, o modelo gaussiano apresentou os menores valores em relação aos demais modelos, o que evidencia que ele foi o modelo que melhor se ajustou a estrutura dos resíduos. Os parâmetros estimados para contribuição (C_1), alcance (a) e efeito pepita (C_0) foram respectivamente, 118,15; 9,91 e 297,79. As conclusões sobre o número de parcelas que se encontram com os resíduos correlacionados é semelhante à conclusão anterior feita quando o alcance estimado foi 9,68. Esse resultado era esperado, pois o efeito do bloco não foi significativo pela análise de variância. Assim, a análise com o erro estimado pelo DBC se assemelha àquela realizada sem o controle local, como no DIC, resultando no ajuste do mesmo modelo teórico com valores estimados próximos.

O Índice de Dependência Espacial (IDE) foi calculado utilizando os parâmetros estimados, resultando em um valor de 28%, o que indica uma dependência espacial moderada dos resíduos.

Assim, ambos experimentos apresentaram dependência espacial entre os resíduos quando foi desconsiderado os blocos e apenas o D_2 apresentou dependência na análise em blocos, de acordo com o Índice de Dependência Espacial (IDE) estimado (Tabela 5).

Tabela 5: Índices de Dependência Espacial (IDE) e valores do índice de Moran para os resíduos estimados na análise em DBC e DIC.

Experimento	Delineamento	Índice de Moran	IDE	Dependência Espacial
D_1	DIC	0,1822 *	0,24	Fraca
	DBC	0,0613	-	Nula
D_2	DIC	0,1843 *	0,28	Moderada
	DBC	0,1739 *	0,27	Moderada

* Significativo a 1%

Esses resultados forneceram uma fundamentação para a modelagem dos erros dependentes e aprimoraram a compreensão das dependências espaciais dos resíduos em ambos experimentos.

4.3. Análise espacial das famílias de cana-de-açúcar

Com base nos resultados obtidos, foi realizada a análise espacial das famílias de cana-de-açúcar utilizando os três modelos descritos no item 3.2.3. para cada experimento.

No experimento D_1, observou-se uma fraca dependência espacial nos resíduos estimados, quando desconsiderado os blocos experimentais. O modelo esférico foi o ajustado para descrever a estrutura de covariância dos resíduos. Na análise em blocos, nenhum modelo geoestatístico foi ajustado, assim como a autocorrelação dos resíduos não foi significativa, conforme indicado pelo Índice de Moran. A Tabela 6 resume os critérios de ajuste para a seleção do melhor modelo entre o modelo

tradicional (que considera o controle local e erros independentes - DBC) e o modelo que desconsidera o controle local e considera os erros dependentes (DIC + ED).

Tabela 6: Valores dos critérios de ajuste adotados para escolha do melhor modelo para o experimento D_1. Critério de Informação de Akaike (AIC), coeficiente de variação experimental (CV), estatística do teste da razão da verossimilhança (LRT) para o modelo completo em relação ao modelo reduzido.

Modelo	Número de parâmetros estimados	AIC	CV	LRT
DBC	101	3767,41	26,27%	-
DIC + ED	100	3780,40	28,80%	14,99 ^{ns}

DBC = modelo considerando o controle local e erros independentes; DIC + ED = modelo desconsiderando o controle local incorporando a covariância nos resíduos; ^{ns} = não significativo.

Os critérios de ajuste revelaram que o modelo com controle local do experimento apresentou melhor precisão experimental em comparação ao modelo que desconsiderou o controle local e incorporou erros dependentes. Este resultado é respaldado pelo menor valor do Critério de Informação de Akaike (AIC) e um coeficiente de variação (CV) inferior. Além disso, o teste da razão de verossimilhança (LRT), com valor de 14,99 e não significativo, sugere que a inclusão de parâmetros espaciais não é necessária para uma explicação adequada da estrutura dos dados. Portanto, o modelo tradicional com erros independentes demonstrou ser mais eficiente e preciso na captura da estrutura dos dados e na previsão dos resultados. Diferente de Reis e Miranda Filho (2003) que utilizaram a mesma metodologia de incorporação da covariância dos erros através da matriz R de variância e covariância para avaliar a covariância espacial entre parcelas vizinhas para resistência à lagarta do cartucho em cultura de milho e tiveram como resultado que o modelo espacial, isto é, considerando os erros dependentes, se mostrou superior ao modelo tradicional com erros independentes. Silva *et al.* (2004) também incorporaram a covariância dos erros por meio da matriz de variâncias-covariâncias, utilizando a função da distância entre os

erros. O modelo com erros dependentes apresentou um melhor ajuste aos dados de progênies de milho em relação à resistência à ferrugem comum.

No experimento D_2, observou-se uma dependência espacial moderada em ambos os modelos ajustados, tanto no DBC quanto no DIC. Em ambos os casos, o modelo gaussiano apresentou o melhor ajuste, independentemente da consideração dos blocos do experimento. Ressalta-se que o controle local, avaliado pela análise de variância, não foi significativo.

A Tabela 7 apresenta os critérios de ajuste utilizados na seleção do modelo mais adequado. O teste da razão de verossimilhança (LRT) foi empregado para comparar o modelo tradicional, que utiliza controle local e erros independentes, com dois outros modelos: (i) o modelo que também inclui controle local, mas com erros dependentes (DBC + ED), e (ii) o modelo sem controle local, porém com erros dependentes (DIC + ED).

Tabela 7: Valores dos critérios de ajuste adotados para escolha do melhor modelo para o experimento D2. Critério de Informação de Akaike (AIC), coeficiente de variação experimental (CV), estatística do teste da razão da verossimilhança (LRT) para o modelo completo em relação ao modelo reduzido.

Modelo	Número de parâmetros estimados	AIC	CV	LRT
DBC	64	2250,50	16,95%	-
DBC + ED	66	2237,57	17,24%	16,93 * (<0,01)
DIC + ED	63	2230,71	17,38%	17,79 * (<0,01)

DBC = modelo considerando o controle local e erros independentes; DBC + ED = modelo considerando o controle local e erros dependentes; DIC + ED = modelo desconsiderando o controle local incorporando a covariância nos resíduos; * significativo a 1%.

De acordo com o AIC, o modelo DIC + ED, que desconsidera o controle local e assume erros espacialmente correlacionados, obteve o melhor ajuste. Este modelo obteve o menor valor de AIC, indicando uma melhor adequação aos dados em comparação com os demais modelos avaliados. No entanto, apesar do resultado obtido para o AIC, o modelo apresentou um aumento no coeficiente de variação que

pode ser explicado pelo aumento da variância residual, como argumentado por Maia *et al.* (2013). Esse aumento foi observado nas análises, onde a variância residual foi de 407,50 TCH² sem correlação dos erros, aumentando para 421,43 TCH² e 428,39 TCH² ao considerar a correlação nos modelos DBC + ED e DIC + ED, respectivamente.

Adicionalmente, o teste LRT, com estatística igual a 17,79 e significativo, confirma a relevância estatística do modelo com parâmetros espaciais, indicando que a inclusão da covariância espacial melhorou o ajuste do modelo.

As médias de toneladas de cana por hectare (TCH) das diferentes famílias de cana-de-açúcar foram estimadas utilizando os três modelos. A Tabela 8 apresenta o ranqueamento de 25% das melhores famílias de cana-de-açúcar no experimento D_2, com base nas médias estimadas de TCH para cada modelo analisado.

Tabela 8: Ranqueamento das quinze melhores famílias de cana de açúcar do experimento D_2 com base nas médias de TCH estimadas para cada modelo analisado.

DBC		**DBC + ED**		**DIC + ED**	
Família	TCH	Família	TCH	Família	TCH
499	167,32	499	170,49	499	170,44
289	166,84	289	163,56	289	164,33
285	158,27	473	153,71	473	153,30
473	151,12	497	149,03	497	150,07
219	149,64	285	148,58	470	148,68
497	149,41	219	148,11	424	147,12
470	146,43	470	147,54	285	147,03
348	144,97	424	145,58	219	146,87
424	143,33	466	144,06	439	143,71
384	139,18	439	142,87	466	143,20
466	138,88	348	139,93	348	140,70
439	138,23	384	138,14	384	138,80
539	137,75	539	136,22	539	137,90
514	137,17	514	134,35	514	134,31
343	136,89	468	133,87	468	133,12

O ranqueamento demonstra que, embora algumas famílias, como a 499 e a 289, mantenham posições semelhantes entre os três modelos, outras apresentam variações. A família 473, por exemplo, que ocupa a quarta posição no modelo DBC, é classificada na terceira posição nos modelos DBC + ED e DIC + ED, enquanto a família 285, inicialmente terceira no DBC, foi reclassificada para a quinta e sétima posição nos outros dois modelos. Essas diferenças destacam a influência da estrutura de erros dependentes no processo de seleção. Com base no índice de coincidência, 93% das famílias classificadas no DBC com erros independentes também aparecem entre as melhores posições no DIC com erros dependentes.

Os resultados obtidos neste experimento mostraram semelhanças com os observados por Maia *et al.* (2013) em estudos de seleção de clones de laranjeira Pêra. Nesses estudos, a análise que desconsiderou o controle local pelos blocos, mas incorporou a covariância espacial dos erros, foi capaz de capturar as variações na maioria das colheitas realizadas, apresentando qualidade de ajuste superior ou equivalente ao modelo que não considerou a dependência espacial existente entre os erros.

Duarte (2000), em um ensaio de competição de soja, comparou a eficiência de seleção entre o método de análise com erros dependentes espacialmente e blocos aumentados. Considerando uma seleção de 25% das linhagens mais produtivas, ele observou uma coincidência de apenas 46% entre as duas seleções. Esses resultados evidenciaram uma grande diferença no ordenamento das linhagens entre a análise clássica e o método de análise espacial. No entanto, resultados semelhantes não foram observados no presente estudo, onde o índice de coincidência foi de 93%, indicando maior proximidade entre as seleções nas metodologias testadas.

A tabela 9 a seguir apresenta 25% das piores médias estimadas de TCH para as famílias de cana-de-açúcar. Essas famílias, com desempenhos mais baixos, são candidatas a serem descartadas no processo de melhoramento, permitindo focar nas famílias que têm maior potencial produtivo.

Tabela 9: Ranqueamento das quinze piores famílias de cana de açúcar do experimento D_2 com base nas médias de TCH estimadas para cada modelo analisado.

DBC		DBC + ED		DIC + ED	
Família	TCH	Família	TCH	Família	TCH
426	102,70	535	106,34	535	106,09
472	102,33	457	104,41	457	104,27
323	101,79	426	99,33	426	98,07
324	98,24	323	96,46	324	96,07
368	95,67	368	95,57	323	95,77
368	92,23	324	95,10	368	95,14
244	91,70	342	92,78	342	91,24
342	91,48	279	90,15	279	90,38
564	88,27	564	87,12	564	87,13
279	88,03	244	85,42	244	86,75
515	87,33	515	84,91	515	85,87
317	85,73	317	83,97	317	85,19
224	83,58	224	80,09	224	80,37
555	81,68	555	76,66	555	76,59
250	54,27	250	48,79	250	48,438

A análise revela que a família 250 consistentemente apresenta as menores médias de TCH em todos os modelos, destacando-se como a principal candidata ao descarte devido ao seu desempenho baixo. Outras famílias que se destacam para possível descarte incluem as famílias 555 e 224, que também apresentam baixos índices de TCH, especialmente quando a dependência espacial é incorporada nos modelos. De acordo com o índice de coincidência, o ranqueamento das 25% das famílias com menor desempenho, com base na média estimada de TCH para o modelo considerando DBC com erros independentes e o modelo de DIC com erros dependentes, apresentou um valor de 0,9333, assim como no índice de coincidência para as melhores famílias. Como destacado por Ferreira (2020), nos programas de melhoramento genético, a seleção de famílias na fase T1 é crucial para o prosseguimento adequado do programa, pois uma seleção inadequada pode resultar

no desperdício de material genético, gerando perdas valiosas. Nesse contexto, o fato de as famílias identificadas como as piores nas análises terem sido consistentemente associadas a médias baixas de TCH em ambas as avaliações é um resultado positivo, uma vez que reforça a eficácia dos modelos em identificar corretamente o desempenho inferior dessas famílias, minimizando o risco de perda de material genético promissor.

De maneira geral, a incorporação da estrutura de covariância residual nos modelos, levando em conta a dependência espacial dos erros, resultou em diferenças sutis no ranqueamento das famílias de cana-de-açúcar, conforme indicado pelo índice de coincidência, quando comparado à análise que considerou erros independentes nos experimentos avaliados. A análise dos critérios de ajuste entre os modelos evidencia pequenas variações, embora a modelagem da dependência espacial tenha se mostrado eficaz na captura da correlação presente nos dados. Esses resultados ressaltam a relevância de se avaliar a dependência espacial dos resíduos na classificação das famílias, pois tal abordagem contribui também para a compreensão da estrutura dos dados.

5. CONCLUSÕES

Para o experimento D_1, observou-se que a dependência espacial dos erros estimados pelo modelo DIC foi fraca. No entanto, ao considerar o modelo DBC, o controle local proporcionado pelos blocos mostrou-se eficaz, resultando em erros independentes. Para o experimento D_2, tanto a análise pelo DIC quanto pelo DBC revelou uma dependência espacial moderada dos erros.

A incorporação da covariância espacial dos erros não melhorou a precisão experimental no experimento D_1 quando comparado ao modelo tradicional com erros independentes. Esse resultado indica que, para esse experimento, o controle local empregado foi suficiente para capturar a variabilidade dos dados, sem a necessidade de modelagem espacial adicional. Já no experimento D_2, o modelo que considerou os erros dependentes e desconsiderou o controle local apresentou uma precisão experimental superior, em comparação ao modelo tradicional.

Os ranqueamentos das famílias de cana-de-açúcar do experimento D_2 demonstraram uma coincidência de 93% na classificação das 25% melhores e das 25% piores famílias, entre as análises realizadas com o modelo de DBC com erros independentes e o modelo de DIC com erros dependentes. Esse resultado é positivo para o programa de melhoramento genético, pois evidencia que as famílias com baixo desempenho foram consistentemente identificadas em ambos os modelos. Isso assegura que não há desperdício de material genético de alto potencial, visto que as famílias com maiores médias foram corretamente mantidas nas melhores classificações, enquanto as de menor valor genético foram descartadas. Consequentemente, evita-se a perda de material geneticamente promissor, contribuindo para a eficiência e precisão na seleção.

REFERÊNCIAS

AKAIKE, H. **A new look at the statistical model identification**. IEEE Transaction on Automatic Control, v. 19, n. 6, p. 716-723, dez. 1974. DOI: https://doi.org/10.1109/TAC.1974.1100705

BARBOSA, M.H.P.; SILVEIRA, L.C.I. **Melhoramento Genético e Recomendação de Cultivares**. In: Santos, F.; Borém, A. e Caldas, C. Editores. Cana-de-açúcar: Bioenergia, Açúcar e Álcool - Tecnologias e Perspectivas. Viçosa, MG – Suprema, 578 p. 2010.

CAMARGO, E. C. G.; FUCKS, S. D.; CÂMARA, G. **Análise espacial de superfícies**. Análise espacial de dados geográficos. Planaltina: Embrapa Cerrados, p. 79-122, 2004.

CAMPOS, J. F.; CARNEIRO, A. P. S.; PETERNELLI. L. A.; CARNEIRO, J. E. de S.; SILVA, M. J.; CECON, P. R. **Classificação de famílias do feijoeiro sob diferentes cenários de dependência espacial e precisão experimental**. Pesquisa Agropecuária Brasileira, v. 51, n. 2, p. 105–111, 2016.

CONAB - COMPANHIA NACIONAL DE ABASTECIMENTO. **Acompanhamento da safra brasileira de cana-de-açúcar**, Brasília, DF, v. 11, n. 4, abril 2024.

CRESSIE, Noel. **Statistics for spatial data**. John Wiley & Sons, 2015.

DE OLIVEIRA, R. P.; GREGO, C. R.; BRANDÃO, Z. N. **Geoestatística aplicada na agricultura de precisão utilizando o Vesper**. Embrapa: Brasilia, Brazil, 2015.

DIGGLE, P. J.; RIBEIRO JR, P. J. **Model-Based Geostatistics**. Nova Iorque, NY, USA: Springer, 2007.

DUARTE, J. B. **Sobre o emprego e a análise estatística do delineamento em blocos aumentados no melhoramento genético vegetal**. 2000. 293p. Tese (Doutorado em Agronomia/Genética e Melhoramento de Plantas)-Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, SP, 2000.

DRUCK, S.; CARVALHO, M.S.; CÂMARA, G.; MONTEIRO, A.V.M. (eds). **Análise Espacial de Dados Geográficos**. Brasília, EMBRAPA, 2004 (ISBN: 85-7383-260-6).

EMBRAPA. CANA. **Agência de informação tecnológica**. Cultivos. 2023. Disponível em:<https://www.embrapa.br/agencia-de-informacao-tecnologica/cultivos/cana> Acesso em: maio 2024.

FERES, A.L.G. **Análise estatística na avaliação de produtividade no melhoramento genético do feijoeiro**. 2009.79p. Tese (Mestrado em Estatística Aplicada e Biometria) - Universidade Federal de Viçosa, Viçosa, MG, 2009.

FERREIRA, M. P. **Redução do adensamento amostral no ajuste de modelos de semivariogramas**. 2015. 52 p. Dissertação (Mestrado em Estatística Aplicada e Biometria) Universidade Federal de Viçosa, Viçosa, MG, 2015.

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

FERREIRA, M. P.; SOUZA, M. L. C. M.; PETERNELLI, L. A.; BARBOSA, M. H. P.; BARBOSA, D. P.; CARNEIRO, A. P.S. **New insights into adjustment for spatial dependence os soil attributes and use aerial images in the initial selection stage of sugarcane families**. Scientia Agaria. 2024.

FERREIRA, P. H. S. GONÇALVES, M. T. V.; TEIXEIRA, G.; FERREIRA, M. D. P.; DE OLIVEIRA, R. L.; BARBOSA, M. H. P.; PETERNELLI, L. A. **Comparison of family selection methodologies used in the initial phase of sugarcane breeding**. Crop science, v. 62, n. 2, p. 679–689, 2022.

FISHER, Ronald A. **The design of experiments**. 8. ed. Edinburgh: Oliver and Boyd, 1966.

GREGO, C. R.; OLIVEIRA, R. P. de; VIEIRA, S. R. Geoestatística aplicada a agricultura de precisão. In: BERNARDI, A. C. de C.; NAIME, J. de M.; RESENDE, A. V. de; BASSOI, L. H.; INAMASU, R. Y. (Ed.). **Agricultura de precisão: resultados de um novo olhar**. Brasília, DF: Embrapa, 2014. cap. 5, p. 74-83.

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.

HODSON, T. O. **Root mean square error (RMSE) or mean absolute error (MAE)**: When to use them or not. Geoscientific Model Development Discussions, v. 2022, p. 1-10, 2022.

MAIA,E.; SIQUEIRA, D. L. de; CARVALHO, S. A. de; PETERNELLI, L. A.; LATADO R. R. **Aplicação da análise espacial na avaliação de experimentos de seleção de clones de laranjeira Pêra**. Revista Ciência Rural, Santa Maria, v.43, n.1, p.8-14, jan, 2013.

MARCONATO, R.; LAROCCA, A. P. C.; QUINTANILHA, J. A. **Análise de tecnologias em estabelecimentos agropecuários por meio dos índices de Moran global e local**. Revista de Política Agrícola. Ano XXI, n.1, jan/fev/mar, 2012.

MORAIS, L. K. de; CURSI, D. E.; SANTOS, J. M. dos; SAMPAIO, M.; CAMARA, T. M. M.; SILVA, P. de A.; BARBOSA, G. V.; HOFFMANN, H. P; CHAPOLA, R. G.; FERNANDES JUNIOR, A. R.; GAZAFFI, R. **Melhoramento Genético da Cana-de Açúcar**. Aracaju: Embrapa Tabuleiros Costeiros, 2015. 40 p. (Embrapa Tabuleiros Costeiros. Documentos, 200). Disponível em: <https://www.bdpa.cnptia.embrapa.br >. Acesso em: out 2023.

MOURA, A. R. **Critérios de seleção de modelos**: um estudo comparativo. 82 p. Dissertação (Mestrado em Modelagem Matemática e Computacional) – Universidade Federal da Paraíba, João Pessoa, 2021.

OLIVEIRA, R. A.; BARBOSA, G. V. de S.; DAROS, E. (org). **50 anos de variedades RB de cana-de-açúcar**: 30 anos de RIDESA. UFPR, RIDESA: Curitiba, Brasil, 2021.

PEDROZO, C. A. **Eficiência da seleção em fases iniciais no melhoramento da cana-de-açúcar**. 2006. 109 p. Dissertação (Mestrado em Genética e Melhoramento) – Universidade Federal de Viçosa, Viçosa, 2006.

PINHEIRO, J.; BATES, D. R Core Team. **nlme**: Linear and Nonlinear Mixed Effects Models_. R package version 3.1-165, 2024 Disponível em: <https://CRAN.R-project.org/package=nlme>.

PONTES, J. M. **Geoestatística**: aplicações em experimentos de campo. 2002. 82 p. Dissertação (Mestrado em Agronomia - Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras, 2002.

R Core Team. R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria, 2024 Disponível em: <https://www.R-project.org/>.

REIS, A. J. dos S.; MIRANDA FILHO, J. B. de. **Autocorrelação Espacial na avaliação de compostos de milho para resistência à lagarta do cartucho (Spodoptera frugiperda)**. Pesquisa Agropecuária Tropical, 33 (2): 65-72, 2003.

RESENDE, M. D. V.; SILVA, F. F.; AZEVEDO, C. F. **Estatística matemática, biométrica e computacional**. Suprema, Visconde do Rio Branco, 881p, 2014.

RIBEIRO JR, P. J.; DIGGLE P. J., CHRISTENSEN O., SCHLATHER M., BIVAND R., RIPLEY B. **geoR**: Analysis of Geostatistical Data. R package version 1.9-4, 2024 Disponível em: <https://CRAN.R-project.org/package=geoR>.

RIDESA. **Rede Interuniversitária para o Desenvolvimento do Setor Sucroenergético**. Disponível em: <https://www.ridesa.com.br/>. Acesso em: 2024.

RODRIGUES, L. C.; ROQUE, C. G.; DA CUNHA, F. F.; DE ASSUNÇÃO, P. C. G.; SOUZA, G. V.; BAIO, F. H. R.; DE OLIVEIRA, J. T. **Variabilidade espacial dos componentes produtivos da cultura da soja**. Agrarian, 2023.

SALVADOR, F. V.; PEREIRA, G. dos S.; SOUZA, M. H.; SILVA, L. M. B.; SANTANA, A. S.; PAULA, I. G.; STECKLING, S. de M.; FERNANDES, R. S.; MARÇAL, T. de S.; CARNEIRO, A. P. S.; CARNEIRO, P. C. S.; CARNEIRO, J. E. de S. **Correcting experimental data for spatial trends in a common bean breeding program**. Crop Science, [S.l.], v. 62, n. 2, p. 825-838, Mar./Apr. 2022. DOI: 10.1002/csc2.20703.

SILVA, H. D.; GUIMARÃES, E. C.; PEDROSA, M. G. **Incorporação da dependência espacial na análise de um experimento de avaliação de progênies de milho quanto à resistência à ferrugem comum**. Ciência e Agrotecnologia, 2004, v. 28, n. 5, p. 1144–1150.

SILVA, M. J. da; CARNEIRO, A. P. S.; FERES, A. L. G.; CARNEIRO, J. E. S.; CECON, P.R. **Experimental precision of spatial analysis methods to evaluate the productivity of common bean families**. Revista de Ciências Agrárias, Belém, v. 63, 2020.

SYED, A. R. **A Review of Cross Validation and Adaptive Model Selection**. Thesis, Georgia State University, 2011.doi: https://doi.org/10.57709/1997958

VENDRAMINI, S. H. F.; Santos, N. S. G. M. D.; Santos, M. D. L. S. G.; Chiaravalloti-Neto, F.; Ponce, M. A. Z.; Gazetta, C. E.; Ruffino Netto, A. **Análise espacial da co-infecção tuberculose/HIV**: relação com níveis socioeconômicos em município do sudeste do Brasil. Revista da Sociedade Brasileira de Medicina Tropical, v. 43, p. 536-541, 2010.

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

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

ZIMBACK, C. R. L. **Análise espacial de atributos químicos de solo para fins de mapeamento da fertilidade do solo**. Botucatu, 2001. Tese (Livre-docência em Ciências Agronômicas) - Faculdade de Ciências Agronômicas, Universidade Estadual Paulista, Botucatu, 2001.