LUCAS ROBERTO DE CASTRO

spANOVA:
BIBLIOTECA PARA ANÁLISE DE VARIÂNCIA DE EXPERIMENTOS COM DEPENDÊNCIA ESPACIAL EM AMBIENTE R

LAVRAS - MG

2019

LUCAS ROBERTO DE CASTRO

spANOVA:
BIBLIOTECA PARA ANÁLISE DE VARIÂNCIA DE EXPERIMENTOS COM DEPENDÊNCIA ESPACIAL EM AMBIENTE R

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

Prof. DSc. Renato Ribeiro de Lima
Orientador

LAVRAS - MG

2019

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Castro, Lucas Roberto de.
spANOVA : Biblioteca para Análise de Variância de Experimentos com Dependência Espacial em Ambiente R / Lucas Roberto de Castro. - 2019.
114 p. : il.

Orientador(a): Renato Ribeiro de Lima.

Dissertação (mestrado acadêmico) - Universidade Federal de Lavras, 2019.
Bibliografia.

1. Análise de Variância. 2. Geoestatística. 3. Modelo Espacial Autorregressivo. I. Lima, Renato Ribeiro de. II. Título.

LUCAS ROBERTO DE CASTRO

spANOVA: BIBLIOTECA PARA ANÁLISE DE VARIÂNCIA DE EXPERIMENTOS COM DEPENDÊNCIA ESPACIAL EM AMBIENTE R
spANOVA: LIBRARY FOR ANALYSIS OF VARIANCE OF EXPERIMENTS WITH SPATIAL DEPENDENCE IN R ENVIRONMENT

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

APROVADA em 26 de Abril de 2019.

Profª. DSc. Alessandra Querino da Silva	UFGD
Prof. DSc. Diogo Francisco Rossoni	UEM
Prof. DSc. José Márcio de Melo	UFLA
Prof. DSc. João Domingos Scalon	UFLA

Prof. DSc. Renato Ribeiro de Lima
Orientador

LAVRAS - MG
2019

AGRADECIMENTOS

A Deus pela vida, cuidado e sabedoria concedida.

À Universidade Federal de Lavras e ao Programa de Pós-Graduação em Estatística e Experimentação Agropecuária por me concederem a oportunidade de realização deste curso.

Ao corpo técnico e docente do Departamento de Estatística (DES) pela paciência e conhecimento compartilhado.

Aos professores Renato e Diogo Rossoni pela orientação recebida, a qual foi fundamental para o desenvolvimento deste trabalho.

Aos integrantes do Grupo de Pesquisa em Estatística Espacial (GPS) pelo apoio científico.

Aos membros da banca examinadora por terem aceitado o convite e terem dado importantes contribuições para este trabalho.

À família, que mesmo de longe, sempre prestou o apoio e o encorajamento necessários para a conclusão do curso.

Aos amigos que fiz na universidade, agradeço pela amizade, companheirismo e momentos que compartilhamos.

O presente trabalho foi realizado com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

RESUMO

Os procedimentos de inferência empregados na análise de variância (ANOVA) necessitam de algumas suposições acerca do erro experimental para que seus resultados sejam válidos, nomeadamente: normalidade, homoscedasticidade e independência. Um dos problemas mais frequentes ocorre quando a independência dos erros não é satisfeita, nem mesmo quando a casualização é realizada. Um dos motivos que conduzem à ocorrência desse fenômeno deve-se ao fato de que em muitos experimentos, sobretudo da área agrícola, há uma forte dependência espacial gerada pela localização da unidade experimental, que não é reduzida com a casualização ou com o controle local. Nesse contexto, foram desenvolvidos trabalhos no com o objetivo de modelar essa dependência e incluí-la na análise para obtenção de resultados mais fidedignos. Algumas abordagens compreendem a modelagem da dependência espacial diretamente a partir da matriz de covariância dos erros, como é o caso da abordagem geoestatística, enquanto outras utilizam transformações na variável resposta com o objetivo de neutralizar o efeito da correlação gerada pelo espaço, como é o caso da análise de variância via modelos autorregressivos espaciais (SAR). Embora essas teorias já estejam desenvolvidas e prontas para serem utilizadas, a ausência de *softwares* ou bibliotecas que realizem tais procedimentos torna-se um empecilho em sua utilização prática, fazendo com que muitos pesquisadores possam tomar uma interpretação equivocada de seus resultados ao optarem pela utilização de modelos que não consideram a informação espacial. Com isso, o objetivo deste trabalho consistiu em desenvolver uma biblioteca para ser utilizada em ambiente de programação, bem como um ambiente gráfico interativo, a fim de possibilitar a inclusão da dependência espacial na análise de variância de forma simples e intuitiva, utilizando tanto a abordagem geoestatística quanto a abordagem via modelos autorregressivos espaciais, viabilizando assim a obtenção de resultados mais precisos. Para isso, foram empregados *softwares* de código aberto. Especificamente, o R foi utilizado na construção de uma biblioteca baseada nessa linguagem de programação, uma vez que grande parte da comunidade científica que faz uso de métodos estatísticos está familiarizada com sua sintaxe. A construção do ambiente gráfico interativo também foi realizada no R por meio da biblioteca *shiny*. Para ilustrar o funcionamento do produto final obtido com a realização deste trabalho, foi analisado um experimento com plantio de candeia (*Eremanthus erythropappus*) realizado na região de Baependi – MG, cujo interesse foi verificar o efeito de 13 tipos de tratamentos de adubação na altura das árvores. Os resultados apontaram diferenças significativas entre os 13 tratamentos, os quais posteriormente foram submetidos a procedimentos de comparações múltiplas revelando que o tratamento adubo formulado NPK 8-28-16 forneceu os melhores resultados.

Palavras-chave: Análise de Variância. Geoestatística. Modelo Espacial Autorregressivo. R. Shiny.

ABSTRACT

The inference procedures used in the analysis of variance (ANOVA) need some assumptions about the experimental error so that its results are valid, namely: normality, homoscedasticity, and independence. One of the most frequent problems occurs when the independence of the errors is not satisfied, not even when the casualization is performed. One of the reasons that lead to the occurrence of this phenomenon is that in many experiments, especially in the agricultural area, there is a strong spatial dependence generated by the location of the experimental unit, which is not reduced by randomization or local control. In this context, work has been developed to model this dependency and include it in the analysis to obtain more reliable results. Some approaches include the modeling of spatial dependence directly from the error covariance matrix, as is the case with the geostatistical approach, while others use transformations in the response variable in order to neutralize the spatial correlation effect, as is the case of analysis of variance via spatial autoregressive models (SAR). Although these theories are already developed and ready to be used, the absence of software or libraries that perform such procedures becomes a drawback in their practical use, causing many researchers to misinterpret their results by opting to use models that do not consider spatial information. Thus, the objective of this work was to develop a library to be used in a programming environment, as well as an interactive graphical interface, in order to allow the inclusion of space dependence in the analysis of variance in a simple and intuitive way, using both the approach geostatistics approach using spatial autoregressive models, thus making it possible to obtain more precise results. For this, open source software was used. Specifically, R was used in the construction of a library based on this programming language, since much of the scientific community that makes use of statistical methods is familiar with its syntax. The construction of the interactive graphical user interface was also performed in the R through the textit shiny library. In order to illustrate the functioning of the final product obtained by this work, an experiment was carried out with candeia (*Eremanthus erythropappus*) carried out in the Baependi - MG region, whose interest was to verify the effect of 13 types of fertilization treatments at the height of the trees. The results showed significant differences among the 13 treatments, which were later submitted to multiple comparison procedures revealing that the formulated NPK 8-28-16 fertilizer treatment provided the best results.

Keywords: Analysis of Variance. Geostatistics. Autoregressive Spatial Model. R. Shiny.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Os experimentos formam uma das bases para a construção do conhecimento científico, por isso devem ser cuidadosamente analisados para se alcançar resultados confiáveis. A análise de dados provenientes de experimentos é o campo de estudo da Estatística Experimental, na qual geralmente se emprega o método denominado análise de variância (ANOVA).

A ANOVA, apresentada por Fisher (1925), consiste em decompor a variabilidade total do experimento em parte atribuída às fontes de variação controladas e parte às fontes de variação não controladas. Esta última, conhecida como erro experimental, o qual desempenha um papel fundamental na análise, uma vez que a partir dele é possível determinar se um tratamento apresenta efeito significativo.

Para tanto, é necessária a observância de três suposições sobre o erro para se garantir a validade do método: normalidade, homoscedasticidade (variância constante) e independência. Dentre elas, a mais difícil de se alcançar em experimentos de campo é a de independência dos erros, mesmo quando a casualização é realizada. Isso porque nesse tipo de experimento há uma forte dependência gerada pela localização da unidade experimental, uma vez que parcelas mais próximas tendem a compartilhar mais características entre si do que as mais afastadas, e essa relação não é neutralizada com a casualização (ZIMMERMAN; HARVILLE, 1991).

De acordo com Stringer et al. (2012), esse fenômeno, conhecido como autocorrelação espacial, pode introduzir um sério viés na análise ao inflacionar a soma de quadrados do erro se não for levado em conta no modelo, reduzindo assim a precisão dos resultados obtidos. Desse modo, tendo em vista que a análise de variância clássica não incorpora a informação espacial, ou seja, considera a independência das observações, alguns trabalhos foram desenvolvidos com o objetivo de modelar e incluir tal dependência na análise de variância.

Uma das primeiras tentativas de sanar as limitações das metodologias usuais e de alguma forma contornar o problema da dependência espacial na análise de variância, foi apresentada por Papadakis (1937), que considerava o uso de vizinhos mais próximos, incluindo no modelo os desvios das parcelas vizinhas com relação à média como uma covariável.

Bartlett (1978) revisou as propriedades do modelo proposto por Papadakis e propôs uma metodologia que incorporava uma estrutura autorregressiva no modelo. Além disso, sugeriu um processo iterativo na aplicação do método de Papadakis, o que resultou em um ganho de eficiência. Porém, este apresentava um viés no valor da estatística F da análise de variância.

Ainda trabalhando com a ideia da modelagem autorregressiva, Long (1996) apresentou um método que utilizava o modelo espacial autorregressivo (SAR) para estimar a estrutura de variabilidade espacial presente no experimento. Em seguida, empregava uma transformação na variável resposta a fim de neutralizar o efeito da correlação espacial, para então proceder à análise de variância. Sua principal vantagem encontra-se na facilidade de implementação, uma vez que, após a transformação da variável resposta a análise de variância é feita exatamente como no modelo usual.

Na literatura é possível encontrar diversos trabalhos (BESAG; KEMPTON, 1986; GLEESON; CULLIS, 1987; CULLIS; GLEESON, 1991; dentre outros) que buscavam a modelagem da dependência espacial por meio da tomada de diferenças sucessivas (diferenciação) baseadas em métodos de vizinho mais próximo. De modo geral, esses métodos baseavam-se em um modelo de "tendência + erro" e usavam a diferenciação dos dados para a remoção da tendência.

Uma abordagem alternativa foi inicialmente discutida nos trabalhos de Gotway e Cressie (1990) e Zimmerman e Harville (1991) em que passaram a considerar as observações como realizações de um campo aleatório e modelar a variação espacial por meio de modelos de correlação, de forma análoga às análises empregadas na geoestatística. Cressie e Hartfield (1996), defendem o uso dessa modelagem, pois, de acordo com esses autores, essa abordagem fornece maior eficiência no processo de inferência.

Isso pôde ser verificado no trabalho de Duarte (2000) que analisou um experimento em blocos aumentados com linhagens de soja e constatou que a correlação espacial nos ensaios genéticos pode ser bastante relevante, o que compromete a validade e eficiência dos métodos estatísticos usuais (que assumem independência). Além disso, verificou que a qualidade do ajuste de um modelo espacial foi bem superior ao modelo usual.

Da mesma forma, Pontes e Oliveira (2004), considerando a utilização de modelos geoestatísticos na análise de ensaios experimentais, ao analisarem um experimento com efeitos aleatórios, propuseram um método iterativo para aprimorar a eficiência das

estimativas dos parâmetros do semivariograma, ferramenta fundamental da abordagem geoestatística. Os resultados obtidos com a aplicação dessa metodologia mostraram uma redução na média dos erros padrões de predição dos efeitos aleatórios de 24,04%, ressaltando a importância de se obter melhores estimativas dos parâmetros do semivariograma para o resultado final da análise.

Caetano (2013), Nogueira (2013) e Rossoni (2011) também fizeram uso de alternativas à caracterização da estrutura de variabilidade espacial presente entre os erros por meio de modelos geoestatísticos e autoregressivos respectivamente, mostrando por meio de simulações que o uso dessa informação melhora de maneira significativa a precisão das análises, podendo ser uma escolha interessante para pesquisadores que manejam experimentos nos quais os erros possuem uma estrutura de correlação espacial definida.

Embora esses e outros trabalhos discutam em detalhes como lidar com a variabilidade espacial na análise de variância, não há bibliotecas ou *softwares* que disponibilizem essas metodologias de forma simplificada exigindo, assim, além de conhecimento estatístico, habilidade com linguagens de programação, o que acaba tornando essas técnicas pouco acessíveis a grande parte dos pesquisadores que poderiam se beneficiar delas.

Diante disso, neste trabalho teve-se por objetivo realizar a implementação computacional da análise de variância de experimentos com dependência espacial por meio da modelagem geoestatística, bem como por meio do modelo espacial autorregressivo, tanto em ambiente de programação, desenvolvendo uma biblioteca em R (R Core Team, 2018), como também em ambiente gráfico por meio do *shiny* (CHANG et al., 2017), pretendendo, dessa forma, alcançar um maior número de usuários e, contribuir para a melhoria de suas análises.

Por fim, o produto final desenvolvido com a realização deste trabalho foi utilizado na análise de um experimento com candeias realizado na região Sul de Minas Gerais, cujo objetivo era comparar o efeito de 13 tratamentos (tipos de adubação) no desenvolvimento de árvores de candeia utilizando como medida a altura da planta. Para isso, foi utilizado o modelo inteiramente ao acaso com a incorporação do efeito de tendência espacial para exemplificar o uso da abordagem geoestatística e ainda, o modelo em blocos casualizados para demonstrar o uso da abordagem por meio de modelos espaciais autorregressivos.

2 REFERENCIAL TEÓRICO

Nesta seção, serão apresentado os principais conceitos e definições que formam a base para os resultados alcançados com este trabalho.

2.1 Estatística Experimental

Em pesquisas científicas, um procedimento comum consiste em formular hipóteses sobre determinado fenômeno e testá-las para avaliar sua veracidade. Essas hipóteses são testadas por meio de métodos estatísticos que estão intimamente relacionados com a maneira como esses dados foram coletados.

A principal razão para utilização da Estatística em análise de experimentos é a presença de variabilidade aleatória nas medidas observadas. Essa variabilidade pode ser causada pela diversidade de fatores de influência (tratamentos[1]) e se manifesta como a variação estocástica de uma unidade experimental[2] (ou parcela) para outra.

A Estatística Experimental é o ramo da Estatística que se propõe a estudar o planejamento, execução, análise e interpretação de dados provenientes de experimentos (BANZATTO; KRONKA, 2006). Seus avanços devem-se principalmente às importantes contribuições do cientista inglês Ronald A. Fisher.

2.1.1 Análise de variância

Um dos métodos mais comuns para análise de experimentos, originalmente desenvolvido por Fisher, durante seus anos de trabalho na estação experimental de Rothamsted, é a análise de variância (ANOVA), formalmente apresentada em seu livro "*Statistical methods for research workers*" (1925).

O método é utilizado a fim de mensurar a importância de possíveis fontes de variação presentes no experimento. Isso é feito com base na decomposição da variabilidade total do experimento em partes ortogonais. Parte dessa variação é atribuída a fontes controladas pelo experimentador e parte atribuída a fatores do acaso (erro experimental) (HINKELMANN; KEMPTHORNE, 2008).

[1] Denominação genérica para designar qualquer método elemento ou material cujo efeito deseja-se medir e comparar.

[2] É a unidade no qual o tratamento é aplicado. É na parcela que se obtém os dados que deverão refletir o efeito de cada tratamento nela aplicado.

É preciso ressaltar que a construção da análise de variância é baseada em três importantes suposições sobre o erro experimental, para que o método forneça resultados válidos (PIMENTEL, 1990):

1. **Independência:** Os erros devem ser independentes, isto é, não devem apresentar uma estrutura de correlação;
2. **Homocedasticidade:** Os erros devem ter a mesma variância;
3. **Normalidade:** Os erros devem ser normalmente distribuídos.

Assim, na tentativa de que as suposições acerca da análise de variância fossem de alguma forma satisfeitas, Fisher (1926) estabeleceu os princípios básicos para a condução de experimentos: repetição, casualização e controle local. De acordo com Montgomery (2012), esses princípios podem ser sumarizados como:

i) **Repetição**: São replicações de cada combinação dos fatores em estudo. Sua finalidade é possibilitar o cálculo da variabilidade do erro experimental a qual será comparada com os efeitos dos tratamentos.

ii) **Casualização**: O ordenamento aleatório dos tratamentos nas parcelas (ou unidades experimentais) é chamado casualização. Montgomery (2012) afirma que casualização é um dos pilares para o uso de métodos estatísticos na análise de experimentos, pois ela tenta garantir que os erros sejam variáveis aleatórias distribuídas de forma independente.

iii) **Controle Local**: O controle local é utilizado para aumentar a precisão das análises na presença de fonte de variação sistemática. Isso é feito dividindo-se o conjunto total (heterogêneo) em subconjuntos homogêneos (blocos).

A garantia de que as suposições sejam válidas é necessária para aplicação do teste F empregado na comparação de variâncias. No contexto da análise de variância, esse teste permite a realização de inferências sobre a igualdade de médias de tratamentos.

A estatística desse teste é construída por meio da razão de quadrados médios (QM), que são obtidos pela soma de quadrados (SQ) de cada fator associado ao modelo divididos por seus respectivos graus de liberdade.

As somas de quadrado refletem a estimativa da variabilidade de cada fonte de variação que são especificadas por um modelo linear. Esse modelo linear descreve a estrutura dos valores observados no experimento e é determinado de acordo com o tipo de delineamento experimental [3] utilizado. Dentre os diversos tipos de delineamentos experimentais, os mais simples e utilizados na literatura são o Delineamento Inteiramente Casualizado (DIC) e o Delineamento em Blocos Casualizados (DBC).

2.1.2 Delineamento Inteiramente Casualizado - DIC

No delineamento inteiramente ao acaso, assume-se homogeneidade entre as unidades experimentais. Esse tipo de delineamento geralmente é utilizado quando se é possível controlar a variabilidade entre as unidades experimentais, como em laboratórios, casas de vegetação, entre outros. Esse delineamento é o mais simples possível de ser conduzido, possuindo apenas duas fontes de variação, uma devida aos tratamentos e outra devido ao erro experimental. O modelo estatístico deste delineamento, considerando I tratamentos e J repetições, é definido de acordo com Montgomery (2012) por

$$y_{ij} = \mu + \psi_i + \epsilon_{ij}, \quad \begin{cases} i = 1, \ldots, I; \\ j = 1, \ldots, J. \end{cases}$$

em que:

y_{ij} representa o valor observado do i-ésimo tratamento na j-ésima repetição;

μ constante inerente a todas as observações;

ψ_i representa o efeito do i-ésimo tratamento;

ϵ_{ij} erro aleatório normalmente distribuído com média 0 e variância σ^2, independente e identicamente distribuído (i.i.d), associado a cada observação, doravante $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$.

Neste delineamento, as hipóteses de interesse são

$$\begin{cases} H_0: \quad \psi_1 = \psi_2 = \cdots = \psi_I = 0 \\ H_1: \quad \psi_i \neq 0 \quad \text{para pelo menos um } i. \end{cases}$$

[3] Plano utilizado para realizar o experimento. Esse plano implica na maneira como os diferentes tratamentos deverão ser distribuídos nas parcelas experimentais, e como serão analisados os dados a serem obtidos.

Assim a análise de variância é realizada conforme a Tabela 2.1

Tabela 2.1 – Tabela de análise de variância - DIC

Fonte de Variação	Grau de Liberdade	Soma de Quadrado	Quadrado Médio	F_0
Tratamento	$I-1$	$SQtrat$	$QMtrat = \dfrac{SQtrat}{I-1}$	$\dfrac{QMtrat}{QMres}$
Resíduo	$I(J-1)$	$SQres$	$QMres = \dfrac{SQres}{I(J-1)}$	
Total	$IJ-1$	$SQtot$		

Fonte: Montgomery (2012)

em que a Soma de Quadrados de Tratamento (SQtrat), Soma de Quadrado de Resíduos (SQres) e Soma de Quadrado Total (SQtot) são definidas por

$$SQtrat = \frac{1}{J}\sum_{i=1}^{I} y_{i+}^2 - \frac{y_{++}^2}{IJ}, \qquad SQtot = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}^2 - \frac{y_{++}^2}{IJ} \qquad \text{e}$$

$$SQres = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}^2 - \frac{1}{J}\sum_{i=1}^{I} y_{i+}^2 = SQtot - SQtrat.$$

Os respectivos Quadrados Médios (QM) são definidos conforme as expressões mostradas na quarta coluna da Tabela 2.1.

As quantidades y_{i+} e y_{++} são definidas por

$$y_{i+} = \sum_{j=1}^{J} y_{ij} \quad \text{e} \quad y_{++} = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}.$$

Tanto no caso do DIC quanto do DBC, a estatística F_0 possui distribuição F de Snedcor, cuja função densidade de probabilidade é dada por

$$f(x) = \frac{\Gamma\left[\frac{u+v}{2}\right]\left(\frac{u}{v}\right)^{\frac{u}{2}} x^{\frac{v}{2}-1}}{\Gamma\left[\frac{u}{2}\right]\Gamma\left[\frac{v}{2}\right]\left[\left(\frac{u}{v}\right)x+1\right]^{\frac{u+v}{2}}} \qquad x \in [0,\infty), \tag{2.1}$$

em que u e v são os graus de liberdade associados aos elementos do numerador e do denominador, respectivamente, da estatística F_0. O símbolo Γ grafado na equação 2.1 denota a função Gama, a qual é definida de acordo com Mood, Graybill e Boes (1974) por

$$\Gamma(t) = \int_0^{\infty} x^{t-1}e^{-x} \quad dx, \quad \text{para} \quad t > 0.$$

À vista disso, a hipótese de igualdade dos tratamentos é avaliada comparando-se o valor da estatística (F_0) obtido com o quantil teórico de sua distribuição ($F_{(\alpha,u,v)}$) e, rejeita-se a hipótese nula sempre que $F_0 > F_{(\alpha,u,v)}$, em que α representa o nível de significância adotado e, nesse caso, u e v representam os graus de liberdade associados à fonte de variação tratamento e resíduo, nessa ordem.

2.1.3 Delineamento em blocos casualisados - DBC

Esse delineamento é provavelmente o mais utilizado na área agrícola, é empregado em situações em que existe heterogeneidade entre as unidades experimentais, e essa heterogeneidade ocorre de forma sistemática o que torna necessário o uso do controle local. Segundo Pimentel (1990), para garantir a eficiência do delineamento é necessário que haja homogeneidade, embora possam existir grandes diferenças entre os blocos. O modelo estatístico para este delineamento, considerando a presença de I repetições e J blocos é dado por

$$y_{ij} = \mu + \psi_i + \eta_j + \epsilon_{ij}, \quad \begin{cases} i = 1, \ldots, I; \\ j = 1, \ldots, J. \end{cases}$$

em que

y_{ij} representa o valor observado do i-ésimo tratamento no j-ésimo bloco;

μ constante inerente a todas as observações;

ψ_i representa o efeito do i-ésimo tratamento;

η_j representa o efeito do j-ésimo bloco;

ϵ_{ij} erro aleatório associado a cada observação, tal que $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$.

Nesse delineamento, a hipótese de interesse é a mesma mencionada no DIC, e além disso, pode-se testar, ainda, o efeito do bloco

$$\begin{cases} H_0: \quad \eta_1 = \eta_2 = \cdots = \eta_J = 0 \\ H_1: \quad \eta_j \neq 0 \quad \text{para pelo menos um } j. \end{cases}$$

E a análise de variância é feita conforme mostrado na Tabela 2.2

Tabela 2.2 – Tabela de Análise de Variância - DBC

Fonte de Variação	Grau de Liberdade	Soma de Quadrado	Quadrado Médio	F_0
Tratamento	$I-1$	$SQtrat$	$QMtrat = \frac{SQtrat}{I-1}$	$\frac{QMtrat}{QMres}$
Bloco	$J-1$	$SQblo$	$QMblo = \frac{SQblo}{J-1}$	$\frac{QMblo}{QMres}$
Resíduo	$(I-1)(J-1)$	$SQres$	$QMres = \frac{SQres}{I(J-1)}$	
Total	$IJ-1$	$SQtot$		

Fonte: Montgomery (2012)

em que $SQblo$ e $QMblo$ referem-se à soma de quadrados da fonte de variação bloco e seu respectivo quadrado médio, definidos como

$$SQtrat = \frac{1}{J}\sum_{i=1}^{I} y_{i+}^2 - \frac{y_{++}^2}{IJ}, \qquad SQblo = \frac{1}{I}\sum_{j=1}^{J} y_{+j}^2 - \frac{y_{++}^2}{IJ},$$

$$SQtot = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}^2 - \frac{y_{++}^2}{IJ} \quad \text{e} \quad SQres = SQtot - SQtrat - SQblo,$$

em que $y_{+j} = \sum_{i=1}^{I} y_{ij}$.

A avaliação das hipóteses de interesse é realizada de modo análogo ao procedimento empregado no DIC, ou seja, compara-se o valor teórico do quantil da distribuição F com o valor obtido na estatística F_0.

2.1.4 Representação matricial

De modo geral modelos estatísticos para análise de experimentos podem ser expressos na forma matricial (Modelo de Gauss-Markov) por

$$\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.2}$$

em que

$\mathbf{Y}$ vetor de respostas observadas, de dimensão $n \times 1$, em que n é o número total de observações;

$\mathbf{X}$ matriz de incidência, definida de acordo com o delineamento experimental utilizado, de modo que sua primeira coluna é sempre composta por um vetor de $\mathbf{1}$ e sua dimensão é $n \times p$, sendo p o número de parâmetros do modelo;

$\boldsymbol{\beta}$ vetor de parâmetros, de dimensão $p \times 1$;

$\boldsymbol{\varepsilon}$ vetor de erros, com dimensão $n \times 1$, tal que $\boldsymbol{\varepsilon} \overset{iid}{\sim} N_n(\mu, \boldsymbol{\Sigma})$.

E a análise de variância é feita conforme a Tabela 2.3 em que $\boldsymbol{P}, \boldsymbol{P_1}$ são projetores dados

Tabela 2.3 – Tabela de análise de variância a partir de matrizes

Fonte de Variação	Grau de Liberdade	Soma de Quadrados	Quadrado Médio	F_0
Tratamento	$GLtrat = posto(\boldsymbol{P} - \boldsymbol{P_1})$	$SQtrat = \boldsymbol{Y'}(\boldsymbol{P} - \boldsymbol{P_1})\boldsymbol{Y}$	$QMtrat = \frac{SQtrat}{GLtrat}$	$\frac{QMtrat}{QMres}$
Resíduo	$GLres = posto(\boldsymbol{I} - \boldsymbol{P})$	$SQres = \boldsymbol{Y'}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$	$QMres = \frac{SQres}{GLres}$	
Total	$GLtot = posto(\boldsymbol{I} - \boldsymbol{P_1})$	$SQtot = \boldsymbol{Y'}(\boldsymbol{I} - \boldsymbol{P_1})\boldsymbol{Y}$		

Fonte: Rencher e Schaalje (2008)

por

$$\boldsymbol{P} = (\boldsymbol{X'X})^{-}\boldsymbol{X'Y}, \quad \text{e} \quad \boldsymbol{P_1} = \boldsymbol{X_1}\boldsymbol{X_1^{-}}$$

em que $\boldsymbol{X_1}$ corresponde à primeira coluna da matriz $\boldsymbol{X}$ e $\boldsymbol{(X'X)^{-}}$ representa a inversa generalizada de $\boldsymbol{(X'X)}$ (ver Rao (1973), seção 4a).

É importante ressaltar que nesse modelo, a estrutura da matriz de covariâncias do erros é dada da seguinte forma

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \sigma^2 = \boldsymbol{I}\sigma^2 = \boldsymbol{V}\sigma^2,$$

isso ocorre pois assume-se que o modelo atende às suposições tradicionais, isto é, com erros não correlacionados, variância constante e normalidade. Sob essas condições, $\boldsymbol{V} = \boldsymbol{I}$ é uma matriz identidade de ordem n, em que n representa o número de elementos da amostra.

Note que a variância, representada pela diagonal principal, é a mesma para todas as observações, além disso, os elementos fora da diagonal principal que representam as covariâncias, são nulos. Essa estrutura é usada para descrever o modelo tradicional, cujos erros são independentes e não correlacionados. Assim, a solução do vetor de parâmetros

$\boldsymbol{\beta}$ pode ser obtida por meio do método de mínimos quadrados, que nesse caso, segundo Rencher e Schaalje (2008) é dada por

$$\boldsymbol{\beta}^o = (\boldsymbol{X'X})^- \boldsymbol{X'Y}. \tag{2.3}$$

Quando $\boldsymbol{V} \neq \boldsymbol{I}$, tem-se o método de mínimos quadrados generalizados e a solução dos parâmetros apresentada em 2.3 passa a incluir $\boldsymbol{V}$,

$$\boldsymbol{\beta}^o = (\boldsymbol{X'V^{-1}X})^- \boldsymbol{X'V^{-1}Y}.$$

Uma outra representação possível para o modelo 2.2 pode ser adotada escrevendo-o de forma particionada (SEARLE, 1987). No caso do DIC, a representação é dada por

$$\mathbf{Y} = \mathbf{X_1}\mu + \mathbf{X_2}\boldsymbol{\psi} + \boldsymbol{\varepsilon}, \tag{2.4}$$

em que

$\mathbf{Y}$ vetor de valores observados da variável resposta, de dimensão $n \times 1$;

μ constante inerente a cada observação;

$\boldsymbol{\psi}$ vetor $k \times 1$ que contém os efeitos de tratamento, em que k é o número de tratamentos;

$\mathbf{X_1}$ vetor de 1's de dimensão $n \times 1$;

$\mathbf{X_2}$ matriz de incidência dos efeitos de tratamento, de ordem $n \times k$;

$\boldsymbol{\varepsilon}$ vetor de erros aleatórios, de dimensão $n \times 1$.

2.1.5 Modelos com covariáveis

No contexto da ANOVA, podem existir variáveis quantitativas que exercem influência nos valores observados. Essas variáveis denominadas de covariáveis possuem relação linear com a resposta do experimento e podem ser utilizadas para aumentar a precisão das análises.

De acordo com Montgomery (2012), se tais covariáveis não forem levadas em conta no modelo podem inflacionar o erro quadrático médio e dificultar a detecção de possíveis diferenças causadas pelos tratamentos na resposta.

Assim, o modelo pode ser adaptado para conter não apenas o efeito dos tratamentos de interesse, mas também o efeito das covariáveis. O modelo resultante é conhecido na literatura como análise de covariância (ANCOVA), sua representação no caso do DIC é dada por

$$y_{ij} = \mu + \psi_i + \theta d_{ij} + \epsilon_{ij},$$

em que

y_{ij} valor observado da variável Y, sob o i-ésimo tratamento na j-ésima repetição, com $i = 1, \ldots, I$ e $j = 1, \ldots, J$;

μ constante associada a cada observação;

ψ_i corresponde ao efeito do i-ésimo tratamento;

θ coeficiente de regressão que relaciona o valor da variável resposta y_{ij} com a covariável d_{ij};

ϵ_{ij} erro aleatório associado a cada observação, tal que $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$.

Nesse modelo, o parâmetro θ pode ser entendido como o coeficiente angular de um modelo de regressão, assim, $\mu + \psi_i$ funcionam como um intercepto. Com isso, uma forma de avaliar a igualdade entre as médias é feita por meio de um teste para a comparação entre os interceptos dos ajustes de regressão para cada tratamento.

Uma forma mais comum de representar esse modelo é

$$y_{ij} = \mu + \psi_i + \theta(d_{ij} - \bar{d}) + \epsilon_{ij}, \tag{2.5}$$

no qual $\bar{d}$ representa a média da covariável e μ representa a média geral. Seus estimadores de mínimos quadrados são dados por

$$\hat{\mu} = \bar{y} \quad e \quad \hat{\psi}_i = \bar{y_i} - \hat{\mu} - \hat{\theta}(\bar{d_i} - \bar{d}).$$

Do modelo representado em 2.5 tem-se que

$$\begin{aligned} E[\bar{y}_i] &= \mu + \psi_i + \theta(\bar{d}_i - \bar{d}) \\ E[\bar{y}_i - \bar{y}_j] &= \psi_i - \psi_j + \theta(\bar{d}_i - \bar{d}_j), \end{aligned}$$

essa característica impede que esses estimadores sejam utilizados para comparar médias entre efeitos de tratamento.

Por isso é utilizada uma média ajustada para o efeito das covariáveis dada por

$$\bar{y}_{i_{adj}} = \bar{y}_i - \hat{\theta}(\bar{d}_i - \bar{d}),$$

cujo valor esperado é

$$\begin{aligned} E[\bar{y}_i] &= \mu + \psi_i \\ E[\bar{y}_i - \bar{y}_j] &= \psi_i - \psi_j. \end{aligned}$$

Com base em uma complementação ao modelo apresentado em 2.4, pode-se obter uma representação matricial da análise de covariância, dada por

$$\boldsymbol{Y} = \boldsymbol{X}_1\mu + \boldsymbol{X}_2\boldsymbol{\psi} + \boldsymbol{D}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{2.6}$$

em que:

$\boldsymbol{D}$ matriz que contém os valores de s covariáveis, cuja dimensão é $n \times s$;

$\boldsymbol{\theta}$ vetor de dimensão $s \times 1$ que contém os coeficientes de regressão que relacionam $\boldsymbol{Y}$ com $\boldsymbol{D}$.

2.1.6 Análise de Resíduos

Os modelos apresentados anteriormente supõem que a parte aleatória do modelo, nesse caso, os resíduos (ϵ) são independentes, identicamente distribuídos de acordo com uma distribuição normal com média 0 e variância constante σ^2 (QUINN; KEOUGH et al., 2002). Embora a veracidade de tais suposições não seja necessária para a estimação por meio do método de mínimos quadrados, elas são fundamentais para garantir a validade dos testes de hipóteses e construção de intervalos de confiança baseados na distribuição t ou F, o que justifica a necessidade de avaliá-las.

A averiguação das suposições do modelo são feitas por meio da checagem dos resíduos, os quais são obtidos por

$$\epsilon_{ij} = y_{ij} - \hat{y}_{ij}$$

em que $\hat{y}_{ij}$ é o estimador de mínimos quadrados do valor esperado de y_{ij} ($E[y_{ij}]$).

Geralmente se opta por trabalhar com os resíduos padronizados, uma vez que eles permitem a identificação de valores extremos (*outliers*) com maior facilidade. A padronização dos resíduos é feita dividindo-os pela raiz quadrada de seu quadrado médio correspondente (citar livro experimental da angela dean), isto é,

$$z_{ij} = \frac{\epsilon_{ij}}{\sqrt{QMres}}.$$

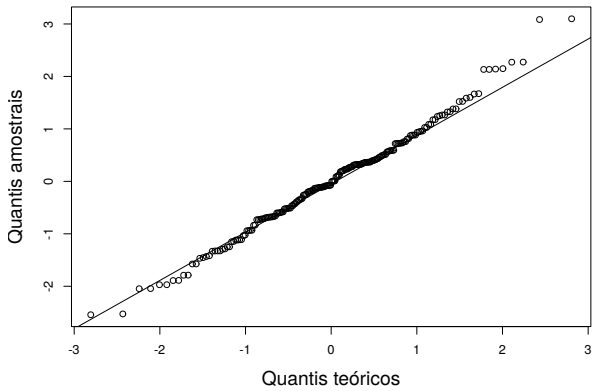
2.1.7 Gráficos de Resíduos

Uma das formas de realizar a verificação dos resíduos é feita por meio da análise gráfica, que permite identificar de forma rápida as possíveis violações de pressupostos no modelo ajustado.

Gráfico Quantil-Quantil (QQ-plot)

O gráfico Quantil-Quantil é um gráfico de amostras ordenadas versus o quantil esperado da distribuição normal padrão (disrtibuição normal com média 0 e variância 1) e pode ser utilizado para verificar se os resíduos atendem a suposição de normalidade. Sob a hipótese de normalidade, espera-se que os pontos apareçam em uma tendência linear (a menos de algumas flutuações aleatórias). Qualquer desvio sistemático de linearidade, indica que os dados sob análise não são normais (QUINN; KEOUGH et al., 2002).

Figura 2.1 – Comportamento esperado do QQ-plot sob normalidade

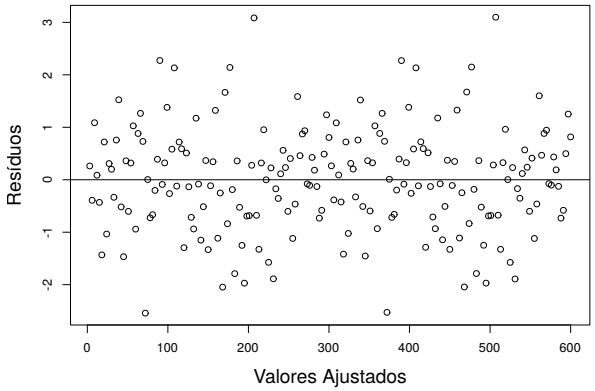


Fonte: Do autor (2019)

Resíduos versus Valores Ajustados

Esse é um gráfico de dispersão dos valores ajustados versus os resíduos do modelo. Ele fornece uma forma visual simplificada de verificar se a homoscedasticidade, isto é, a igualdade das variâncias é uma hipótese plausível. Para tanto, avalia-se o padrão de variação dos pontos, e caso estes não apresentem uma variabilidade que aumenta de acordo com os valores ajustados, conclui-se que a suposição de homoscedasticidade é aceitável (QUINN; KEOUGH et al., 2002).

Figura 2.2 – Comportamento esperado do gráfico de resíduos x valores ajustados

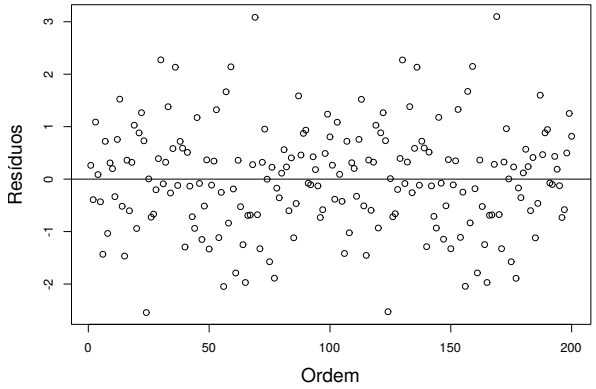


Fonte: Do autor (2019)

Resíduos versus Ordem de Coleta

Em muitos casos, especialmente quando os dados são observados ao longo do tempo ou do espaço, a suposição de independência é violada. Um gráfico de dispersão dos resíduos versus a ordem de coleta (ou tempo) indicará um padrão de altos (ou baixos) valores seguidos por baixos (ou altos) valores, caso a hipótese de dependência seja inverossímil (QUINN; KEOUGH et al., 2002).

Figura 2.3 – Comportamento esperado do gráfico de resíduos x ordem de coleta sob a suposição de independência



Fonte: Do autor (2019)

2.1.7.1 Testes de Hipóteses

Embora a análise gráfica forneça indícios sobre a suposição de normalidade e dos demais pressupostos, ela é subjetiva, e possibilita que indivíduos tomem interpretações diferentes. Por isso torna-se necessária a aplicação de testes de hipóteses para avaliar, por meio da inferência estatística, se essas suposições são verdadeiras com um nível de confiança estabelecido pelo pesquisador.

Teste para normalidade - Shapiro-Wilk

O teste proposto por Shapiro e Wilk (1965), é realizado por meio do cálculo de uma estatística (W_1) a qual é utilizada para testar se uma amostra é proveniente de uma distribuição normal. Valores pequenos de W_1 são prova de ausência de normalidade. O valor da estatística W_1 pode ser calculado por

$$W_1 = \frac{b_1^2}{\sum_{i=1}^{n}(\epsilon_{(i)} - \bar{\epsilon})^2},$$

em que $\epsilon_{(i)}$ são os valores da amostra (resíduos) ordenados e, b_1 é uma constante calculada por

$$b_1 = \begin{cases} \sum_{i=1}^{n/2} a_{n-i+1} \times (\epsilon_{n-i+1} - \epsilon_{(i)}) & \textit{se } n \textit{ é par} \\ \sum_{i=1}^{(n+1)/2} a_{n-i+1} \times (\epsilon_{n-i+1} - \epsilon_{(i)}) & \textit{se } n \textit{ é ímpar} \end{cases}$$

sendo a uma constante obtida por meio das médias, variâncias e covariâncias das estatísticas de ordem de uma amostra de tamanho n de uma distribuição normal, cujos valores

são tabelados e podem ser encontrados no artigo original. As hipóteses avaliadas nesse teste são:

$$\begin{cases} H_0: & \textit{Os dados são provenientes de uma distribuição } N(\mu, \sigma^2) \\ H_1: & \textit{Os dados não são provenientes de uma distribuição normal.} \end{cases}$$

A hipótese nula é rejeitada quando $W_1 < W_\alpha$, em que W_α representa o valor crítico da estatística W_1 considerando um nível de significância α. Os valores de W_α podem ser encontrados em Shapiro e Wilk (1965).

Testes para os demais pressupostos

Os testes de hipóteses também podem ser empregados para a verificação dos pressupostos de independência e homoscedasticidade. Este texto abordará na seção 2.7.3 o teste de permutação para o índice I de Moran que permite verificar a independência (espacial) dos resíduos. Porém, geralmente, na análise de variância usual, adotam-se testes que verificam a suposição de independência por meio da ausência de correlação serial, como por exemplo, o teste de Durbin-Watson. Esse teste não será apresentado aqui, da mesma forma que não será apresentado um teste para a verificação da homoscedasticidade. Ao leitor interessado, recomenda-se a leitura de Kutner et al. (2005).

2.1.8 Procedimentos de comparações múltiplas

Após a detecção de diferença estatisticamente significativa entre os tratamentos por meio da análise de variância, é necessário saber em quais níveis do tratamento essa diferença ocorre, para isso, utiliza-se um procedimento de comparações múltiplas de médias.

Os procedimentos de comparações múltiplas auxiliam o pesquisador na identificação dos grupos que apresentam os melhores ou piores resultados. Dentre os procedimentos os mais conhecidos, pode-se citar o teste t, teste de Tukey e o agrupamento de médias de Scott-Knott.

Diante disso, nesta subseção serão abordados o teste t-multivariado, Tukey e o método de Scott-Knott.

2.1.8.1 Teste t-multivariado

Este teste pode ser utilizado quando o interesse consiste em avaliar hipóteses acerca de um contraste de médias (T). Um contraste é a combinação linear de médias (μ_i) de tal forma que a soma de seus coeficientes seja zero, isto é, considerando-se a comparação de I tratamentos, tem-se que

$$T = c_1\mu_1 + c_2\mu_2 + \ldots + c_I\mu_I \quad \text{em que} \quad \sum_{i=1}^{I} c_i = 0.$$

sendo c_i os coeficientes, determinados de modo conveniente para que as comparações desejadas sejam obtidas.

Para avaliar as hipóteses

$$H_0 : T = A \quad \text{verus} \quad H_1 : T \neq A,$$

pode-se usar a estatística

$$t_c = \frac{\hat{T} - A}{\sqrt{Var(\hat{T})}}.$$

Em que A representa o valor com o qual o contraste será comparado e $\hat{T}$ a estimativa do contraste T. Assim, sob H_0, t_c tem distribuição t de student com ν graus de liberdade. Porém na prática, ocorre frequentemente o interesse em testar mais de dois contrastes entre médias, o que segundo Hothorn, Bretz e Westfall (2008), torna este método ineficiente pois o teste simultâneo de p hipóteses eleva a probabilidade de se cometer erro tipo I.

Por isso, uma solução para esse problema consiste no uso de testes baseados na distribuição t multivariada (BRETZ; WESTFALL; HOTHORN, 2016), cuja função de densidade de probabilidade, de acordo com Ferreira (2008), é dada por

$$f_{\boldsymbol{X}}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\frac{\nu+p}{2})}{(\pi\nu)^{p/2}\Gamma(\frac{\nu}{2})|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \left[1 + \frac{1}{\nu}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right]^{-\frac{\nu+p}{2}}$$

em que $\boldsymbol{X} = [X_1, \ldots, X_p]' \in \mathbb{R}^p$, $\boldsymbol{\mu}$ representa o vetor de médias de $\boldsymbol{X}$ e $\nu\boldsymbol{\Sigma}/(\nu - 2)$ a matriz de covariâncias, no caso de $\nu > 2$.

Com isso, para procedimento de avaliar I médias de tratamentos, expressas por $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_k]'$ podem ser propostos p vetores de contrastes definidos por $\boldsymbol{c_1, \ldots, c_p}$,

assim as hipóteses de interesse são

$$H_{0i} : \boldsymbol{c'}_i\boldsymbol{\mu} = 0, \quad \text{pra} \quad i = 1, 2, \ldots, p. \tag{2.7}$$

Definindo $\boldsymbol{C}$ como a matriz de contrastes, de forma que cada linha do contraste avaliado, tem-se $\boldsymbol{C} = [\boldsymbol{c'_1}, \ldots, \boldsymbol{c'_p}]$, então o teste usado para avaliar simultaneamente as hipóteses em 2.7 tem a seguinte hipótese nula

$$H_0 = \boldsymbol{C'\mu} = \boldsymbol{0}.$$

Denotando por $\boldsymbol{M}$ a matriz de covariâncias do estimador $\boldsymbol{C'\mu}$, ou seja $\boldsymbol{M} = Var(\boldsymbol{C'\mu})$, tem-se que os elementos da diagonal dessa matriz (m_{ii}) representam as variâncias dos p contrastes avaliados. Com isso, a partir da estimação de $\boldsymbol{M}$ obtém-se a estatística do teste, dada por

$$\boldsymbol{t_c} = \boldsymbol{S^{1/2}C\hat{\mu}},$$

em que $\boldsymbol{S^{1/2}}$ representa uma matriz diagonal cujos elementos são dados por $(1/\sqrt{m_{ii}})$. Daí segue que sob H_0, $\boldsymbol{t_c}$ tem distribuição t p-variada com ν graus de liberdade e matriz de correlação dada por

$$\boldsymbol{R = S^{1/2}MS^{1/2}}.$$

Por fim, sob H_0, a hipótese nula é rejeitada quando $\boldsymbol{t_c} > \boldsymbol{t_p}(\nu)$.

2.1.8.2 Teste de Tukey

O teste de Tukey é aplicado para testar se existe diferença entre médias tomadas duas a duas, isto é,

$$\begin{cases} H_0 : & \mu_i = \mu_j \\ H_1 : & \mu_i \neq \mu_j, \end{cases}$$

são feitas todas as combinações possíveis entre i e j, assim ao avaliar I médias, são feitas $I(I-1)/2$ comparações.

Cada diferença entre média, $|\bar{y}_i$ - $\bar{y}_j|$ é avaliada com base na diferença mínima significativa (DMS) e a hipótese nula é rejeitada se $|\bar{y}_i$ - $\bar{y}_j| > DMS$, que no caso de experimentos balanceados é dada por

$$DMS = q_{(\alpha,I,\nu)}\sqrt{\frac{QMres}{J}},$$

em que J representa o número de repetições e I o número de tratamentos. A quantidade $q_{(\alpha,I,\nu)}$ representa o quantil da distribuição de amplitude estudentizada (PIMENTEL, 1990).

2.1.8.3 Método de Agrupamento Scott-Knott

O agrupamento de Scott-Knott, trata-se de um método de agrupamento que consiste em dividir as médias de I tratamentos em grupos distintos e não sobrepostos por meio da maximização da soma de quadrados entre os grupos (RAMALHO; FERREIRA; OLIVEIRA, 2005).

Um dos problemas encontrados na aplicação desse método é que quando o número de tratamentos é grande, o número de grupos cresce consideravelmente, o que dificulta a aplicação do teste. Por exemplo, ao considerar I médias de tratamentos, existem $2^{I-1}-1$ partições possíveis das I médias em grupos distintos (RAMALHO; FERREIRA; OLIVEIRA, 2005).

Para contornar esse problema, uma alternativa é ordenar as médias dos tratamentos, dessa maneira, passam a existir $I-1$ partições possíveis. Assim, considerando as partições realizadas nas médias ordenadas, as hipóteses avaliadas passam a ser

$$\begin{cases} H_0: \ \mu_{(1)} = \mu_{(2)} = \cdots = \mu_{(I)} = \mu \\ H_1: \ \mu_{(1)} = \mu_{(2)} = \cdots = \mu_{(k_1)} = \mu_{g_1} \quad \text{e} \quad \mu_{(k_1+1)} = \mu_{(k_1+2)} = \cdots = \mu_I = \mu_{g_2}, \end{cases}$$

em que $\mu_{(i)}$ representa a média ordenada na posição i, μ_{g_1} e μ_{g_2} representam, respectivamente as médias dos grupos 1 e 2, tal que $\mu_{g_1} \neq \mu_{g_2}$ e ainda $k_1, k_2 \geq 1$ sendo $k_2 = I - k_1$.

Um algortimo para a aplicação do método é dado da seguinte forma:

(i) ordenar as I médias e dividir os tratamentos em dois grupos, para todas as $I-1$ partições possíveis dos valores médios ordenados;

(i) determinar a soma de quadrados máxima entre os dois grupos. Tal valor será definido por B_0 e estimado da seguinte forma:

$$B_0 = \frac{T_1^2}{k_1} + \frac{T_2^2}{k_2} - \frac{(T_1+T_2)^2}{k_1+k_2} \tag{2.8}$$

em que T_1 e T_2 representam a soma das médias dos grupos 1 e 2 enquanto k_1 e k_2 são as quantidades de médias em cada grupo. Assim,

$$T_1 = \sum_{i=1}^{k_1} \bar{Y}_{(i)} \quad \text{e} \quad T_2 = \sum_{i=k_1+1}^{I} \bar{Y}_{(i)};$$

(iii) determinar o valor da estatística λ

$$\lambda = \frac{\pi}{2(\pi - 1)} \frac{B_0}{\hat{\sigma}_0^2}$$

- em que π é o número irracional de valor aproximado 3,141592.
- B_0 é calculado de acordo com a equação 2.8;
- $\hat{\sigma}_0^2$ é o estimador de máxima verossimilhança de $\sigma_{\bar{Y}}^2$, dados por

$$\frac{1}{k+v}\left[\sum_{i=1}^{I}(\bar{Y}_{(i)} - \bar{Y}) + v s_{\bar{Y}}^2\right]$$

em que:

v representa os graus de liberdade do resíduo;

$\bar{Y_{(i)}}$ é a média ordenada do tratamento i;

$\bar{Y}$ é a média geral do experimento;

$s_{\bar{Y}}^2$ é a variância da média dada pela razão entre o quadrado médio do resíduo e o número de repetições de cada tratamento.

(iv) sob H_0, a estatística λ segue aproximadamente uma distribuição $\chi_{\nu_0}^2$. Assim, a hipótese nula é rejeitada se $\lambda > \chi_{\nu_0,\alpha}^2$, em que α é o nível de significância adotado para o teste e ν_0 é o grau de liberdade dado por $\nu_0 = I/(\pi - 2)$.

(v) Caso H_0 seja rejeitada, os dois grupos formados serão independentemente submetidos aos passos (i) e (ii). O processo de cada subgrupo se encerra ao não rejeitar H_0 no passo (v) ou se cada grupo contiver apenas uma média.

2.2 Geoestatística

Nesta seção será apresentada a geoestatística, que é uma das abordagens utilizadas para a modelagem da dependência espacial dos erros na análise de variância.

A geoestatística teve sua origem na década de 50 quando Daniel G. Krige, trabalhando com mineração de ouro na África do Sul, percebeu que a variabilidade dos dados possuía uma estrutura que dependia da distância de amostragem. Isso despertou o interesse do engenheiro francês Georges Matheron, que anos mais tarde formalizou a teoria de variável regionalizada, a qual ficou mais conhecida como geoestatística devido as suas aplicações na área de geologia e mineração (RIVOIRARD, 2005).

Essa teoria baseia-se no princípio de que os fenômenos sob estudo como, concentração de ouro, propriedades de solo, chuvas em determinadas regiões, dentre outros são realizações de um processo estocástico. Dessa forma, o valor observado de um desses fenômenos em qualquer localização x_i, $i = 1, 2, \ldots, n$, em que x_i denota a coordenada geográfica, é apenas uma das infinitas realizações possíveis.

Dessa maneira, o valor observado em x é tratado como uma variável aleatória, a qual será denotada por Y. O conjunto de todas as variáveis aleatórias em todas as localizações x pertencentes a R (um subconjunto do espaço vetorial $\mathbb{R}^s$) constituem um processo estocástico. Essas variáveis aleatórias estão contidas no espaço real s-dimensional $\mathbb{R}^s$ e recebem o nome de variáveis regionalizadas. Esse espaço vetorial $\mathbb{R}^s$, embora possa ser definido em qualquer dimensão, geralmente, em análises espaciais é definido no espaço bidimensional $\mathbb{R}^2$ (VIEIRA et al., 1983).

Matematicamente, esse processo é representado como

$$\{Y(x) : x \in R \subset \mathbb{R}^s\}. \tag{2.9}$$

Diante disso, pode-se apontar uma diferença entre a geoestatística e os métodos estatísticos tradicionais. Enquanto neste último considera-se que as observações são não correlacionadas, na geoestatística a correlação no espaço constitui a base para a modelagem dos fenômenos, uma vez que se parte do princípio de que quanto mais próxima uma observação está da outra mais relacionadas elas estão e, à medida que a distância aumenta essa correlação fica mais fraca.

Além disso, é importante ressaltar que diferentemente dos métodos estatísticos usuais, na geoestatística, trabalha-se com n variáveis aleatórias das quais somente uma realização é observada. Isso acontece, pois na maioria dos fenômenos, dificilmente se observa mais de uma realização da variável ao mesmo tempo e exatamente na mesma localização. Dessa maneira, Soares (2006) apud Yamamoto e Landim (2015) afirma que

é impossível determinar estatísticas como média e variância em um ponto específico, assim, uma solução consiste em assumir algum grau de estacionariedade para a variável regionalizada, os quais serão abordados a seguir.

2.2.1 Estacionariedade

Conforme Ross (2014), um processo é dito estacionário se $Y(x)$ e $Y(x+h)$ têm a mesma distribuição conjunta de probabilidade. Em que a quantidade h — assumindo que o processo seja isotrópico, ou seja, considerando que seu comportamento é o mesmo em todas as direções do fenômeno — pode ser entendida como a distância de uma amostra a outra $h = ||x_i - x_j||$, conhecida na literatura como *lag*.

Essa condição, porém, raramente é encontrada na prática, por isso se adota a estacionariedade de segunda ordem. Um processo estacionário de segunda ordem ou fracamente estacionário, de acordo com Cressie (1993), é aquele que satisfaz as seguintes condições

(i) $E[Y(x)] = \mu$,

(ii) $C(h) = E[\{Y(x)\}\{Y(x+h)\}] - \mu^2$.

Ou seja, a média do processo é assumida constante sobre toda a região de estudo e ainda existe uma função de covariância $C(\cdot)$ que depende somente de h.

Ainda assim, essa condição é difícil de ser atendida. Para contornar esse problema, é adotada uma suposição menos restritiva denominada de hipótese intrínseca, na qual se assume que o valor esperado da diferença entre variável regionalizada na localização x e $x+h$ é nula

$$E[Y(x) - Y(x+h)] = 0,$$

e ainda que a variância dessa diferença depende apenas de h

$$Var[Y(x) - Y(x+h)] = E[\{Y(x) - Y(x+h)\}^2] = 2\gamma(h)$$

que também pode ser representada por

$$\frac{1}{2}Var[Y(x) - Y(x+h)] = \frac{1}{2}E[\{Y(x) - Y(x+h)\}^2] = \gamma(h).$$

A função $\gamma(h)$ é denominada semivariância, essa medida descreve a dependência espacial das variáveis e é a base para a construção do semivariograma, uma ferramenta usada para a modelagem espacial do fenômeno de interesse.

A semivariância, sob a estacionariedade de segunda ordem (e consequentemente sob a hipótese intrínseca), também pode ser obtida por meio da relação dada por

$$C(h) = C(0) - \gamma(h), \tag{2.10}$$

em que $C(0)$ corresponde à $Var[Y(x)]$.

Nesse caso é possível definir, ainda, o coeficiente de correlação ($\rho(h)$), que se trata de uma medida adimensional limitada ao intervalo $[-1, 1]$ também empregada para descrever a relação espacial das variáveis. Tal coeficiente pode ser obtido por

$$\rho(h) = \frac{C(h)}{C(0)} = \frac{C(0) - \gamma(h)}{C(0)} = 1 - \frac{\gamma(h)}{C(0)}. \tag{2.11}$$

2.2.2 Semivariograma

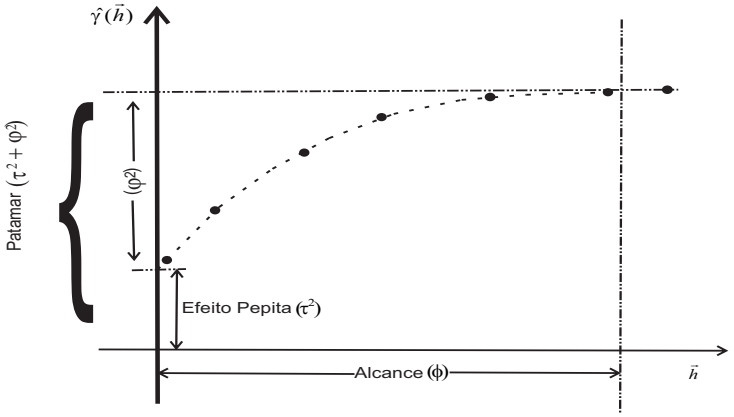
Como já mencionado, uma ferramenta útil para modelar a relação entre as distâncias e a associação espacial das variáveis é o semivariograma, que é baseado na função de semivariância (γ). Sob a hipótese intrínseca, um método usual para estimar a semivariância é conhecido como estimador de momentos de Matheron, que de acordo com Schabenberger e Gotway (2017) é definido como

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

em que $\hat{\gamma}(h)$ é o valor estimado para a semivariância, $y(x_i)$ e $y(x_j)$ são os valores observados de Y nas localizações x_i e x_j separados por uma distância $h = |x_i - x_j|$, e $N(h)$ é o número de pares comparados em h.

Mudando o valor de h, obtém-se um conjunto ordenado de semivariâncias, o qual é utilizado na construção do semivariograma experimental, que se trata de um gráfico de dispersão constituído pelos valores de $\hat{\gamma}(h)$ no eixo das ordenadas e suas respectivas distâncias no eixo das abscissas, e que permite realizar o ajuste de um modelo capaz de descrever a relação espacial dos dados, conforme exemplo mostrado na Figura 2.4, sendo

Figura 2.4 – Exemplo de semivariograma e a representação de seus parâmetros



Fonte: Do autor (2019)

Alcance (ϕ): A distância dentro da qual as amostras são espacialmente correlacionadas. Desse ponto em diante, considera-se que não existe mais dependência espacial entre as amostra;

Efeito Pepita (τ^2): é o valor da semivariância para a distância zero e representa a componente da variabilidade espacial que não pode ser relacionado com uma causa específica;

Contribuição (φ^2): É a diferença entre o patamar e o efeito pepita, este parâmetro está relacionado com a variabilidade que pode ser descrita por um modelo.

Patamar ($\tau^2 + \varphi^2$): O valor da semivariância correspondente ao alcance (ϕ). Teoricamente, esse parâmetro corresponde à variância dos dados.

Para dar continuidade as análises geoestatísticas, é necessário ajustar um modelo teórico ao semivariograma. Isso porque é preciso que a semivariância atenda a condição de não negatividade, o que é garantido por esses modelos (ARMSTRONG, 1998). Dentre os diversos modelos que garantem a não negatividade, conhecidos como modelos autorizados, os mais comuns na literatura são: esférico, exponencial, gaussiano e wave.

O modelo esférico possui uma expressão polinomial simples e sua forma geralmente condiz bem com o que é frequentemente observado: um crescimento quase linear até uma certa distância, em seguida, uma estabilização (ARMSTRONG, 1998). Sua fórmula é

mostrada na equação 2.12.

$$\gamma(h) = \begin{cases} 0 & \text{, se} \quad h = 0; \\ \tau^2 + \varphi^2 \left[\frac{3}{2} \left(\frac{h}{\phi} \right) - \frac{1}{2} \left(\frac{h}{\phi} \right)^3 \right] & \text{, se} \quad 0 < h < \phi; \\ \tau^2 + \varphi^2 & \text{, se} \quad h \geq \phi. \end{cases} \tag{2.12}$$

O modelo exponencial, apresentado na equação 2.13, cresce rapidamente no início, porém apenas tende ao seu patamar em vez de alcançá-lo, por isso seu alcance prático é definido como o valor no qual a distância atinge 95% do patamar, isto é, 3ϕ (CRESSIE, 1993).

$$\gamma(h) = \begin{cases} 0 & \text{, se} \quad h = 0; \\ \tau^2 + \varphi^2 \left[1 - exp \left(\frac{-h}{\phi} \right) \right] & \text{, se} \quad h \neq \phi. \end{cases} \tag{2.13}$$

O modelo gaussiano é utilizado para descrever fenômenos extremamente contínuos, assim como no modelo exponencial, seu alcance é obtido de forma assintótica e é definido como $1,73\phi$ (ARMSTRONG, 1998). Sua representação é dada por:

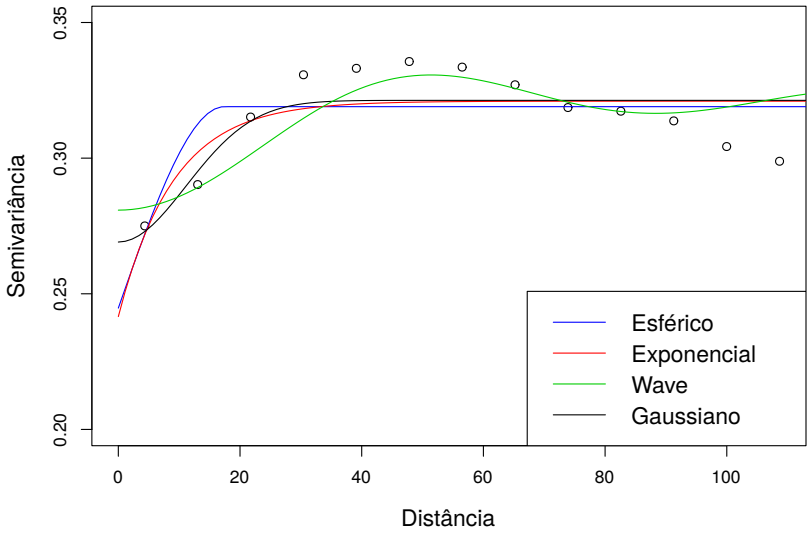
$$\gamma(h) = \begin{cases} 0 & \text{, se} \quad h = 0; \\ \tau^2 + \varphi^2 \left[1 - exp \left[- \left(\frac{h}{\phi} \right)^2 \right] \right] & \text{, se} \quad h \neq \phi. \end{cases} \tag{2.14}$$

O modelo wave, também chamado de modelo cardial seno, pertence à família dos modelos conhecidos como “hole effect”, e são úteis para descrever fenômenos com autocorrelação negativa ou com variação cíclica (SCHABENBERGER; GOTWAY, 2017). Esse modelo alcança o patamar e continua a oscilar após um período b_w. É representado matematicamente por

$$\gamma(h) = \begin{cases} 0 & \text{, se} \quad h = 0; \\ \tau^2 + \varphi^2 \left[1 - \frac{b_w}{h} \sin \left(\frac{h}{b_w} \right) \right] & \text{, se} \quad h \neq \phi. \end{cases} \tag{2.15}$$

Para ilustrar, uma representação gráfica desses modelos é apresentada na Figura 2.5.

Figura 2.5 – Modelos teóricos de semivariograma



Fonte: Do autor (2019)

2.2.3 Estimação dos parâmetros

De acordo com Mello et al. (2005) os métodos de estimação podem ser classificados em dois grupos: os baseados em ajuste de modelos ao semivariograma experimental e os que ajustam um modelo direto aos dados.

Neste trabalho serão utilizados os métodos baseados em ajuste de modelos ao semivariograma, especificamente o de mínimos quadrados (ordinários). É importante ressaltar que conforme h aumenta, o número de pares usados na estimação da semivariância diminui. Com isso sua estimativa fica menos precisa, por isso é comum delimitar uma distância máxima dentro da qual as estimativas serão realizadas, conhecida como *cutoff*. Alguns autores, como Clark (1979), sugerem um *cutoff* de metade da máxima distância entre os pontos observados.

Método de mínimos quadrados ordinários

A estimativa de mínimos quadrados ordinários dos parâmetros do semivariograma podem ser obtidos minimizando-se a seguinte função

$$Q(\boldsymbol{\vartheta}, \gamma_i) = \sum_{i=1}^{k} \{\gamma(h_i) - \gamma(h_i, \boldsymbol{\vartheta})\}^2 ,$$

em que $\boldsymbol{\vartheta}$ representa o vetor de parâmetros que definem o semivariograma, γ_i representa cada estimativa da semivariância nos *lag* h_i, $i = 1, \ldots, k$ de modo que k refere-se ao número de lags dentro da distância máxima definida, e $\gamma(h_i, \boldsymbol{\vartheta})$ é a semivariância esperada do modelo considerado.

2.3 Seleção de Modelos

Após o ajuste de alguns modelos, pode-se determinar aquele que melhor descreve o fenômeno. Na literatura, há alguns critérios que auxiliam nessa escolha, um dos mais conhecidos é o critério de informação de Akaike (AIC) (AKAIKE, 1983), definido por

$$AIC = 2logL + 2K,$$

sendo L a verossimilhança maximizada do modelo avaliado e K o número de parâmetros que o compõe. A seleção do modelo é feita escolhendo-se aquele que possui o menor valor de AIC.

Além do AIC, na geoestatística é comum o uso de validação cruzada para a seleção de modelos. Uma das maneiras de realizar a validação cruzada é por meio da técnica *leave-one-out*, que consiste em excluir uma observação do conjunto de dados e estimá-la através do interpolador de krigagem e calcular o erro de predição, o processo é repetido até que todas as observações tenham sido utilizadas.

Com base nisso, é possível calcular algumas medidas como o Erro Médio Reduzido ($\bar{ER}$) e o Desvio Padrão do Erro Reduzido (S_{ER}) definidos nas equações 2.16 e 2.17, respectivamente, como:

$$\overline{ER} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y(x_i) - \hat{y}(x_i)}{\sigma_{x_i}}\right) \tag{2.16}$$

$$S_{ER} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y(x_i) - \hat{y}(x_i)}{\sigma_{x_i}}\right)^2} \tag{2.17}$$

em que

$y(x_i)$ valor observado na posição x_i;

$\hat{y}(x_i)$ valor predito na posição x_i;

$\hat{y}(x_i)$ desvio padrão da krigagem na posiçãox_i;

n número de amostras.

Cressie (1993) recomenda que o modelo a ser escolhido é aquele que possuir o erro médio reduzido mais próximo de 0 e desvio padrão do erro reduzido mais próximo de 1.

2.4 Tendência espacial

Embora na modelagem de alguns fenômenos seja possível assumir que a média é constante, ou pelo menos não apresenta grandes variações dentro da região de estudo, na prática isso é pouco comum, uma vez que em muitos fenômenos a média pode variar de acordo com a localização espacial, infringindo a primeira condição da hipótese intrínseca.

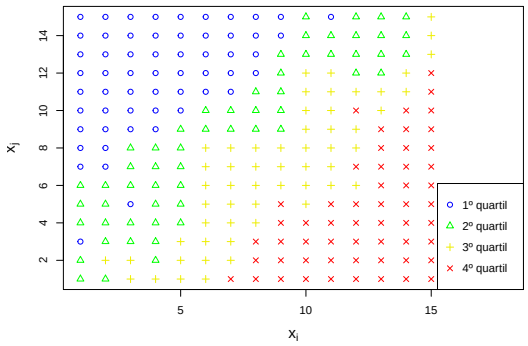
Quando isso ocorre, diz-se que existe efeito de tendência espacial nos dados a qual geralmente pode ser modelada por alguma função polinomial das coordenadas espaciais é dada por (DIGGLE; RIBEIRO JÚNIOR, 2007)

$$\mu(\boldsymbol{x}) = \theta_0 + \sum_{j=1}^{s} \theta_j x_{ij},$$

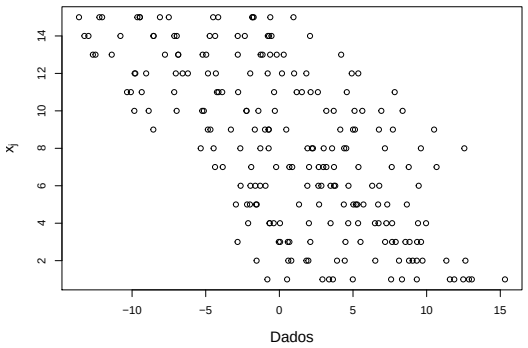
em que x_{ij} representa a j-ésima coordenada espacial da i-ésima realização da variável regionalizada Y, tal que $j = 1, \ldots, s$ e θ_0 e θ_j com $j = 1, \ldots, s$ são os parâmetros que descrevem a relação entre a coordenada espacial e a realização da variável Y.

De acordo com Diggle e Ribeiro Júnior (2007), é possível detectar indícios de tendência espacial já na análise exploratória dos dados. Isso é feito mediante a construção de gráficos de quartis e também de gráficos da variável resposta em relação às suas respectivas coordenadas.

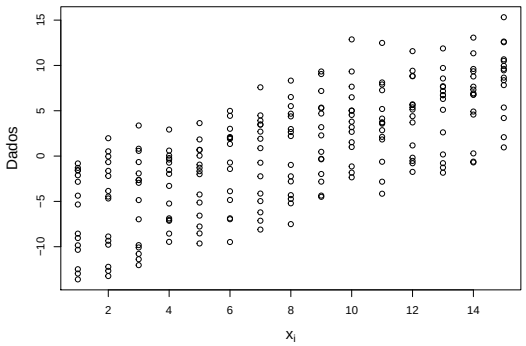
Figura 2.6 – Exemplos de fenômenos com tendência espacial em função das coordenadas



(a) Separação dos dados por quartil



(b) Dados em função da coordenada x_i



(c) Dados em função da coordenada x_j

Fonte: Do autor (2019)

A Figura 2.6 ilustra como essa tendência é detectada por meio da análise exploratória. Nela nota-se que a os valores mais altos (identificados pelos quartis) tendem a se concentrar em valores maiores da coordenada x_i e valores menores da coordenada x_j. Por meio do gráfico de dispersão, essa tendência também é evidente, visto que os pontos não estão espalhados aleatoriamente, mas em vez disso seguem uma tendência linear clara.

Como a avaliação gráfica pode mostrar-se muitas vezes subjetiva, os mesmos autores também sugerem um teste de hipóteses para a verificação de tendência, que é feito avaliando-se a significância do parâmetro $\theta_j = 0$ versus $\theta_j \neq 0$.

2.5 Distribuição de Variáveis Espaciais

O modelo geoestatístico clássico assume que o valor da variável Y na localização x pode ser expressa pela soma de três componentes: Uma componente determinística

associada a um valor médio, uma componente aleatória espacialmente correlacionada e um erro residual (BURROUGH, 1986), isto é,

$$Y(x) = \mu(x) + S(x) + \xi(x), \tag{2.18}$$

em que

$\mu(x)$ componente determinística assumida contínua, podendo ser interpretada como uma tendência associada a um valor médio;

$S(x)$ é uma componente espacialmente dependente com $E[S(x)] = 0$ e função de covariância $C[S(x_i), S(x_j)]$ parametrizada por ϕ, o parâmetro de alcance do semivariograma, se $i = j$ então $C[S(x_i), S(x_j)] = \varphi^2$, caso contrário $C[\cdot]$ assume qualquer outro valor que não necessariamente é zero;

$\xi(x)$ erro aleatório independente com média 0 e variância dada pelo efeito pepita do semivariograma (τ^2).

Dessa definição, decorre que,

$$C[Y(x_i), Y(x_j)] = \begin{cases} Var[Y(x_i), Y(x_j)] = \sigma^2 = \tau^2 + \varphi^2, & se \quad i = j, \\ C[S(x_i), S(x_j)], & se \quad i \neq j. \end{cases}$$

Além disso, se for assumida distribuição gaussiana para ξ, tem-se que $\xi \sim N(\mathbf{0}, \mathbf{I}\tau^2)$, analogamente $S \sim N(\mathbf{0}, \mathbf{F}\varphi^2)$, em que $\mathbf{F}$ é uma matriz de correlação quadrada de ordem n cujos elementos são dados por $f(h_{ij})$ em que $h_{ij} = ||x_i - x_j||$ e f é uma função que corresponde ao modelo adotado para descrever a variabilidade espacial da variável $S(x)$.

Considera-se ainda que as variáveis $S(x)$ e $\xi(x)$ são independentes, isso implica que $C[S(x), \xi(x)] = 0$. Dessa forma, a variável aleatória Y também seguirá uma distribuição gaussiana, tal que $Y \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, em que $\boldsymbol{\Sigma} = \mathbf{F}\boldsymbol{\varphi^2} + \mathbf{I}\boldsymbol{\tau^2}$ a qual pode ser representada como

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & C(h_{12}) & \dots & C(h_{1n}) \\ C(h_{12}) & \sigma^2 & \dots & C(h_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ C(h_{1n}) & C(h_{2n}) & \dots & \sigma^2 \end{bmatrix}$$

Sob suposição de estacionariedade de 2ª ordem e usando a relação entre covariância e a semivariância apresentada em 2.11, é possível mostrar que a matriz $\boldsymbol{\Sigma}$ pode ser reescrita como

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho(h_{12}) & \dots & \rho(h_{1n}) \\ \rho(h_{12}) & 1 & \dots & \rho(h_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(h_{1n}) & \rho(h_{2n}) & \dots & 1 \end{bmatrix} \sigma^2 = \boldsymbol{V}\sigma^2.$$

2.6 Modelos de Resposta Autoregressivos

Outra abordagem utilizada para modelar dados de experimentos com dependência espacial é do modelo espacial autorregressivo (SAR). Esse modelo relaciona os dados de um determinado local com uma combinação linear de valores vizinhos, que representam a estrutura autorregressiva.

2.6.1 Modelo SAR

Segundo Ywata e Albuquerque (2011), um dos modelos mais utilizados para a modelagem de autocorrelação espacial é o modelo espacial autorregressivo (Spatial Autoregressive ou Spatial Lag Model), ou simplesmente designado de modelo SAR, definido por

$$\mathbf{Y} = \rho\mathbf{W}\mathbf{Y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.19}$$

sendo

$\mathbf{Y}$ vetor $n \times 1$ de valores observados da variável resposta;

ρ parâmetro espacial autoregressivo;

$\mathbf{W}$ matriz de peso da vizinhança espacial de dimensão $n \times n$, em que n corresponde ao número de observações;

$\mathbf{X}$ matriz de incidência das variáveis explicativas de dimensão $n \times p$;

$\boldsymbol{\beta}$ vetor $p \times 1$ de parâmetros;

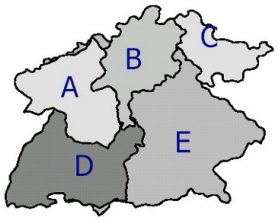
$\boldsymbol{\varepsilon}$ vetor $n \times 1$ de erros inerentes a cada observação.

A matriz $\mathbf{W}$ é composta por w_{ij} elementos que quantificam a dependência espacial entre as parcelas i e j. Os elementos dessa matriz são denominados pesos espaciais, os quais em sua forma mais simples são definidos por

$$w_{ij} = \begin{cases} 1 & \text{se as regiões } i \text{ e } j \text{ compartilham fronteiras;} \\ 0 & \text{caso contrário.} \end{cases}$$

Ao construir $\boldsymbol{W}$ dessa forma ela é intitulada matriz vizinhança espacial binária e uma exemplificação de sua elaboração é apresentada na Figura 2.7.

Figura 2.7 – Exemplo de construção da matriz de vizinhança binária

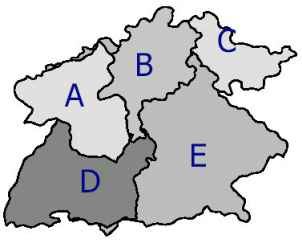


	A	B	C	D	E
A	0	1	0	1	0
B	1	0	1	1	1
C	0	1	0	0	1
D	1	1	0	0	1
E	0	1	1	1	0

Fonte: Câmara e Carvalho (2004)

Algumas vezes é útil fazer a normalização de $\boldsymbol{W}$ devido à necessidade de seu emprego na realização de cálculos. Essa normalização, é feita mediante a divisão de cada elemento w_{ij} pela soma da linha a qual pertence, conforme ilustrado pela Figura 2.8.

Figura 2.8 – Exemplo de construção da matriz de vizinhança normalizada



	A	B	C	D	E
A	0	0,5	0	0,5	0
B	0,25	0	0,25	0,25	0,25
C	0	0,5	0	0	0,5
D	0,33	0,33	0	0	0,33
E	0	0,33	0,33	0,33	0

Fonte: Câmara e Carvalho (2004)

Uma alternativa para a definição da estrutura de vizinhança, discutida em Gumpertz, Graham e Ristaino (1997), é pela adoção de uma parcela de referência, para a qual é definido um raio (r) e então é considerada como vizinha todas as parcelas abrangidas

por este raio. Dessa forma, a matriz $\boldsymbol{W}$ poderia ser construída empregando-se a seguinte regra

$$w_{ij} = \begin{cases} 1 & \text{se as parcelas estiverem na circunferência de raio } r; \\ 0 & \text{caso contrário.} \end{cases}$$

2.6.1.1 Estimação do parâmetro ρ

Após a construção de $\boldsymbol{W}$, o parâmetro ρ precisa ser estimado, e isso pode ser feito pelo método da máxima verossimilhança. Para isso, uma solução proposta por Ord (1975) baseia-se na decomposição do jacobiano $|\boldsymbol{I} - \rho\boldsymbol{W}|$ em termos dos autovalores ω_i, dada por

$$|\boldsymbol{I} - \rho\boldsymbol{W}| = \prod_{i=1}^{n}(1 - \rho\omega_i), \tag{2.20}$$

sendo $\boldsymbol{I}$ a matriz identidade de ordem $n \times n$.

A estimativa é obtida por meio da maximização da equação 2.20 a qual não possui solução única e deve ser feita por métodos iterativos. No software R, esse procedimento é feito a partir da função `lagsarlm` da biblioteca spdep (BIVAND; PEBESMA; GOMEZ-RUBIO, 2013).

2.7 Análise de Variância com Erros Espacialmente Dependentes

Nesta seção será apresentado o procedimento para a obtenção da análise de variância usando a abordagem geoestatística e a abordagem via modelos autorregressivos espaciais.

2.7.1 Abordagem via geoestatística

Na abordagem geoestatística, o objetivo é a obtenção de uma estrutura para a matriz de covariâncias alternativa àquela apresentada para o modelo 2.2 com a finalidade de descrever a dependência espacial das observações.

Para isso, o primeiro passo consiste em extrair o erro do modelo convencional mostrado na equação 2.2, ou seja, $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{\circ}$ em que $\boldsymbol{\beta}^{\circ} = (\boldsymbol{X'X})^{-}\boldsymbol{X'Y}$, e em seguida é construído o semivariograma.

Após esse procedimento, é necessário ajustar um modelo teórico ao semivariograma, a partir do qual será possível obter a matriz de covariância dos erros $\boldsymbol{\Sigma}$ de acordo com o que foi apresentado na seção 2.5, por

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho(h_{12}) & \dots & \rho(h_{1n}) \\ \rho(h_{12}) & 1 & \dots & \rho(h_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(h_{1n}) & \rho(h_{2n}) & \dots & 1 \end{bmatrix} \sigma^2 = \boldsymbol{V}\sigma^2,$$

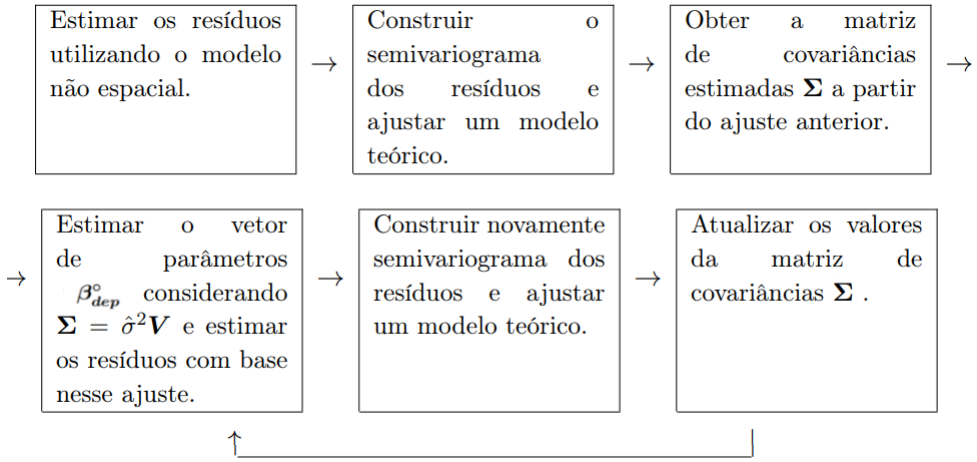
em que $\boldsymbol{V}$ é a chamada matriz de correlação, cujos valores são obtidos por meio dos modelos teóricos de usado para modelar a dependência espacial como por exemplo, modelo esférico (2.12), exponencial (2.13), gaussiano (2.15) dentre outros. E ainda $\sigma^2 = \tau^2 + \varphi^2$, sendo τ^2 e φ^2 os parâmetros efeito pepita e contribuição, respectivamente, estimados a partir do modelo teórico ajustado.

De posse da matriz $\boldsymbol{V}$, é possível obter as estimativas para o vetor de soluções considerando a dependência espacial, por meio do método de mínimos quadrados generalizados

$$\boldsymbol{\beta^{\circ}_{dep}} = (\boldsymbol{X'V^{-1}X})^{-}\boldsymbol{X'V^{-1}Y}.$$

Com $\boldsymbol{\beta^{\circ}_{dep}}$ obtém-se uma estimativa para o erro considerando a dependência espacial. Pontes e Oliveira (2004) sugerem um método iterativo para aprimorar essa estimativa. Tal processo consiste em repetir várias vezes as etapas descritas na Figura 2.9 até que se alcance uma convergência. Essa convergência pode ser avaliada comparando-se a diferença entre as estimativas obtidas na etapa atual e na etapa anterior, se essa diferença for menor que um erro de tolerância, neste trabalho definido com 10^{-3}, considera-se que a convergência foi alcançada.

Figura 2.9 – Processo iterativo para a estimação dos parâmetros do semivariograma



Fonte: Nogueira (2017)

2.7.1.1 Modelo sem tendência espacial

Após a realização do procedimento descrito inicialmente, é possível obter a análise de variância com erros espacialmente dependentes, que é apresentada na Tabela 2.4 a seguir. Para maiores detalhes veja Nogueira (2013).

Tabela 2.4 – Tabela de análise de variância de um DIC considerando dependência espacial

Fonte de Variação	Grau de Liberdade	Soma de Quadrados	Quadrado Médio	F_0
Tratamento	$GLtrat = posto(\boldsymbol{P_1} - \boldsymbol{P})$	$SQtrat = \mathbf{Y}'(\mathbf{P_1} - \mathbf{P})\mathbf{Y}$	$QMtrat = \frac{SQtrat}{GLtrat}$	$\frac{QMtrat}{QMres}$
Resíduo	$GLres = posto(\boldsymbol{P})$	$SQres = \mathbf{Y}'\mathbf{P}\mathbf{Y}$	$QMres = \frac{SQres}{GLres}$	
Total	$GLtot = posto(\boldsymbol{P_1})$	$SQtot = \mathbf{Y}'\mathbf{P_1}\mathbf{Y}$		

Fonte: Nogueira (2013)

Neste caso, a estatística F_0 tem distribuição F de Snedecor não-central com parâmetro de não centralidade δ, isto é, $\frac{QMtrat}{QMres} \sim F(\nu_1, \nu_2, \delta)$, em que $\nu_1 = posto(\boldsymbol{P_1} - \boldsymbol{P})$, $\nu_2 = posto(\boldsymbol{P})$ e $\delta = \frac{1}{2\sigma^2}\boldsymbol{\psi'X_2P_1X_2\psi}$, em que $\boldsymbol{X_2}$ representa a matriz de incidência excluindo-se sua primeira coluna e $\boldsymbol{\psi}$ representa o vetor de efeitos dos tratamentos. As matrizes $\boldsymbol{P}$ e $\boldsymbol{P_1}$ são definidas como

$$\begin{aligned} \boldsymbol{P} &= \mathbf{V^{-1}} - \mathbf{V^{-1}X}(\mathbf{X'V^{-1}X})^{-}\mathbf{X'V^{-1}} \\ \boldsymbol{P_1} &= \mathbf{V^{-1}} - \mathbf{V^{-1}X_1}(\mathbf{X_1'V^{-1}X_1})^{-}\mathbf{X_1'V^{-1}}, \end{aligned}$$

em que $(\mathbf{X'V^{-1}X})^{-}$ representa a inversa generalizada de $(\mathbf{X'V^{-1}X})$ e $\boldsymbol{X_1}$ é definida conforme apresentado em 2.1.4.

Assim como apresentado no modelo convencional, o interesse é testar se existe diferença significativa entre os tratamentos. Assim, sob H_0, isto é, considerando a hipótese em que todos os efeitos são nulos, tem-se que $\boldsymbol{\psi} = \mathbf{0}$, dessa forma o parâmetro de não centralidade δ da distribuição da estatística F_0 se anula conduzindo a uma distribuição F central com graus de liberdade ν_1 e ν_2, então a hipótese nula é rejeitada quando $F_0 > F(\alpha, \nu_1, \nu_2)$, sendo $F(\alpha, \nu_1, \nu_2)$ o quantil da distribuição F.

No caso do delineamento em blocos casualizados, a análise de variância é feita de acordo com a Tabela 2.5

Tabela 2.5 – Tabela de análise de variância de um DBC considerando dependência espacial

Fonte de Variação	Grau de Liberdade	Soma de Quadrado	Quadrado Médio	F_0
Tratamento	$GLtrat = posto(\boldsymbol{X_2}) - posto(\boldsymbol{X_1})$	$SQtrat = \mathbf{Y'(P - P_1)Y}$	$QMtrat = \frac{SQtrat}{GLtrat}$	$\frac{QMtrat}{QMres}$
Bloco	$GLblo = posto(\boldsymbol{X}) - posto(\boldsymbol{X_2})$	$SQblo = \mathbf{Y'(P - P_1 - P_2)Y}$	$QMblo = \frac{SQblo}{GLblo}$	$\frac{QMblo}{QMres}$
Resíduo	$GLres = posto(\boldsymbol{V^{-1}}) - posto(\boldsymbol{X})$	$SQres = \mathbf{Y'(V^{-1} - P)Y}$	$QMres = \frac{SQres}{GLres}$	
Total	$GLtot = posto(\boldsymbol{V^{-1}}) - posto(\boldsymbol{X_1})$	$SQtot = \mathbf{Y'(V^{-1} - P_1)Y}$		

Fonte: Nogueira (2013)

em que $\mathbf{P_2} = \mathbf{R_v X_2}(\mathbf{X_2' R_v X_2})^{-}\mathbf{X_2' R_v}$ e $\mathbf{R_v} = \mathbf{V^{-1}} - \mathbf{V^{-1}X_1}(\mathbf{X_1' V^{-1} X_1})^{-1}\mathbf{X_1' V^{-1}}$. Tem-se ainda que

$$\frac{QMtrat}{QMres} \sim F(\nu_1, \nu_2)$$
$$\frac{QMblo}{QMres} \sim F(\nu_3, \nu_2)$$

sendo

$$\nu_1 = posto(\boldsymbol{X_2}) - 1$$

$$\nu_2 = n - posto(\boldsymbol{X})$$

$$\nu_3 = posto(\boldsymbol{X}) - posto(\boldsymbol{X_2}).$$

A decisão é feita de modo análogo ao caso do DIC.

2.7.1.2 Modelo com tendência espacial

Para o caso em que a tendência espacial é considerada, o modelo empregado é o de análise de covariância mostrado na equação 2.6. Nesse caso, as coordenadas espaciais são utilizadas como covariáveis que ajudam a descrever o comportamento da média.

Retomando o modelo dado em 2.6, tem-se que

$$\boldsymbol{Y} = \boldsymbol{X}_1\mu + \boldsymbol{X}_2\boldsymbol{\psi} + \boldsymbol{D}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

em seu contexto de utilização para fins de incorporação do efeito de tendência e, considerando o caso particular em que as coordenadas espaciais estão indexadas no $\mathbb{R}^2$, a matriz $\boldsymbol{D}$ é especificada da seguinte forma

$$\boldsymbol{D} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix},$$

em que $\boldsymbol{D}$ contém as coordenadas 1 e 2 das n posições espaciais amostradas no plano. E o vetor $\boldsymbol{\theta} = [\theta_1, \theta_2]'$ representa o vetor de parâmetros dos efeitos das coordenadas, enquanto que os demais componentes são definidos da mesma forma que em 2.6.

Com esse modelo, é possível aplicar testes para avaliar a significância do efeito de tendência espacial como também dos tratamentos. Isso é feito através da análise de variância, conforme mostrado em Nogueira (2017).

Vale lembrar que na análise de covariância não há ortogonalidade entre os fatores, por isso é necessário obter os testes de significância em ajustes separados, isto é, o teste para o efeito das covariáveis deve ser corrigido pelo efeito de tratamento e vice-versa.

Para lidar com essa situação, pode-se recorrer à soma de quadrado tipo III. Nesse método, é analisado o efeito da inclusão de cada parâmetro corrigido para o efeito dos já existentes no modelo.

Para isso, parte-se de um modelo completo, ou seja, um modelo que contém todos os parâmetros possíveis, e a partir dele é possível obter as somas de quadrados de forma corrigida para os parâmetros, a partir das seguintes reduções (NOGUEIRA, 2017)

$$(1) \quad \boldsymbol{Y} = \boldsymbol{X_1}\mu + \boldsymbol{X}_2\boldsymbol{\psi} + \boldsymbol{D\theta} + \boldsymbol{\varepsilon}; \quad \text{(modelo completo)}$$

$$(2) \quad \boldsymbol{Y} = \boldsymbol{X_1}\mu + \boldsymbol{X}_2\boldsymbol{\psi} + \boldsymbol{\varepsilon};$$

$$(3) \quad \boldsymbol{Y} = \boldsymbol{X_1}\mu + \boldsymbol{D\theta} + \boldsymbol{\varepsilon};$$

$$(4) \quad \boldsymbol{Y} = \boldsymbol{X_1}\mu + \boldsymbol{\varepsilon}.$$

Dessa forma, as somas de quadrados (SQ) são obtidas de acordo com a Tabela 2.6.

Tabela 2.6 – Soma de quadrados tipo III

Fontes de Variação	SQ tipo III
Coordenadas	$\text{SQ}(\boldsymbol{\theta}\|\mu, \boldsymbol{\psi})$
Tratamentos	$\text{SQ}(\boldsymbol{\psi}\|\mu, \boldsymbol{\theta})$
Resíduos	$\boldsymbol{Y'Y} - SQ(\mu, \boldsymbol{\psi}, \boldsymbol{\theta})$

Fonte: Nogueira (2017)

As somas de quadrados dos parâmetros são obtidas por meio das diferenças dessas quantidades em cada modelo, dessa forma definindo SQparâmetro_i como a i-ésima soma de quadrados de parâmetro associada a cada um dos quatro modelos, tem-se que

$$SQ(\mu, \boldsymbol{\theta}) = \text{SQparâmetro}_3;$$

$$SQ(\mu, \boldsymbol{\psi}) = \text{SQparâmetro}_2;$$

$$SQ(\mu, \boldsymbol{\psi}, \boldsymbol{\theta}) = \text{SQparâmetro}_1;$$

$$SQ(\boldsymbol{\theta}|\mu, \boldsymbol{\psi}) = \text{SQparâmetro}_1 - \text{SQparâmetro}_2;$$

$$SQ(\boldsymbol{\psi}|\mu, \boldsymbol{\theta}) = \text{SQparâmetro}_1 - \text{SQparâmetro}_3.$$

Dessa maneira, a análise de variância pode ser obtida em função de projetores, conforme a Tabela 2.7

Tabela 2.7 – Tabela de análise de variância a partir de soma de quadrados tipo III para erros espacialmente dependentes considerando tendência espacial

Fonte de Variação	Grau de Liberdade	Soma de Quadrado	Quadrado Médio	F_0
Coordenadas	$GLcoord = posto(\boldsymbol{P}_c - \boldsymbol{P}_2)$	$SQcoord = \boldsymbol{Y'}(\boldsymbol{P}_c - \boldsymbol{P}_2)\boldsymbol{Y}$	$QMcoord = \frac{SQcoord}{GLcoord}$	$\frac{QMcoord}{QMres}$
Tratamentos	$GLtrat = posto(\boldsymbol{P}_c - \boldsymbol{P}_3)$	$SQtrat = \boldsymbol{Y'}(\boldsymbol{P}_c - \boldsymbol{P}_3)\boldsymbol{Y}$	$QMtrat = \frac{SQtrat}{GLrat}$	$\frac{QMtrat}{QMres}$
Resíduo	$GLres = posto(\hat{\boldsymbol{V}}^{-1} - \boldsymbol{P}_c)$	$SQres = \boldsymbol{Y'}(\hat{\boldsymbol{V}}^{-1} - \boldsymbol{P}_c)\boldsymbol{Y}$	$QMres = \frac{SQres}{GLres}$	

Fonte: Nogueira (2017)

em que,

$$\boldsymbol{P}_c = \boldsymbol{V^{-1}M}_c(\boldsymbol{M}_c'\boldsymbol{V^{-1}M}_c)^{-}\boldsymbol{M}_c\boldsymbol{V^{-1}}$$
$$\boldsymbol{P}_2 = \boldsymbol{V^{-1}M}_2(\boldsymbol{M}_2'\boldsymbol{V^{-1}M}_2)^{-}\boldsymbol{M}_2\boldsymbol{V^{-1}}$$
$$\boldsymbol{P}_3 = \boldsymbol{V^{-1}M}_3(\boldsymbol{M}_3'\boldsymbol{V^{-1}M}_3)^{-}\boldsymbol{M}_3\boldsymbol{V^{-1}}$$

Neste caso, $\boldsymbol{V}$ é definida do mesmo modo como foi descrito para o modelo que não considera a tendência espacial e ainda $\boldsymbol{M}_c = [\boldsymbol{X}_1|\boldsymbol{X}_2|\boldsymbol{D}]$, $\boldsymbol{M}_2 = [\boldsymbol{X}_1|\boldsymbol{X}_2]$, $\boldsymbol{M}_3 = [\boldsymbol{X}_1|\boldsymbol{D}]$, em que "|" é a simbologia adotada para representar uma matriz aumentada.

Por meio da análise de variância apresentada na Tabela 2.7 é possível testar as seguintes hipóteses

$$(i) = \begin{cases} H_0: \quad \theta_1 = \theta_2 = 0 \\ H_1: \quad \theta_1 \neq \theta_2 \end{cases} \quad e \quad (ii) = \begin{cases} H_0: \quad \psi_1 = \psi_2 = \cdots = \psi_k = 0 \\ H_1: \quad \psi_k \neq 0 \quad \text{para pelo menos um } k. \end{cases}$$

A hipótese (i) refere-se ao efeito das coordenadas, ela é utilizada para verificar se tais coordenadas exercem efeito significativo na variável resposta, ou seja, verifica a presença de tendência espacial em função das coordenadas. H_0 é rejeitada sempre que sua respectiva estatística $F_0 > F_{(\alpha,GLcoord,GLres)}$, sendo $F_{(\alpha,GLcoord,GLres)}$ o quantil teórico da distribuição F com nível de significância α e graus de liberdade $GLcoord$ e $GLres$.

A hipótese (ii) é utilizada para avaliar o efeito dos tratamentos, ela é avaliada de forma análoga ao que foi descrito para a hipótese (i), isto é, comparando-se o valor de sua estatística F_0 com o quantil da F e a rejeição de H_0 ocorre quando $F_0 > F_{(\alpha,GLtrat,GLres)}$.

2.7.2 Procedimentos de comparações múltiplas sob a abordagem geoestatística

Ao detectar-se um efeito significativo no teste F da análise de variância, é necessário o uso de um procedimento de comparações múltiplas para apontar quais dos tratamentos sob estudo diferenciam-se entre si. Para isso, são empregados procedimentos de comparações baseados na distribuição t multivariada, distribuição da amplitude estudentizada e o método de agrupamento de médias de Scott-Knott. Os procedimentos apresentados nesta seção levam em conta a presença da autocorrelação espacial e foram abordados no trabalho de Nogueira (2017).

Como esses procedimentos são realizados por meio de comparações na média (nesse caso a média espacial), serão apresentados inicialmente os métodos utilizados para estimação desse parâmetro. Além disso serão abordados ainda os métodos para estimação da variabilidade da média e da variabilidade da diferença entre médias, os quais são necessários na composição dos procedimentos apresentados.

2.7.2.1 Médias espaciais dos tratamentos

Considere um modelo de experimento balanceado com I tratamentos e J repetições cujos efeitos de tratamento são nulos, dado por

$$\boldsymbol{Y} = \boldsymbol{X}_1\mu + \boldsymbol{\varepsilon}, \tag{2.21}$$

o que equivale a

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{1J} \\ \vdots \\ y_{I1} \\ \vdots \\ y_{IJ} \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix} \mu + \begin{bmatrix} e_{11} \\ \vdots \\ e_{1J} \\ \vdots \\ e_{I1} \\ \vdots \\ e_{IJ} \end{bmatrix}$$

em que

$\boldsymbol{Y}$ vetor de respostas de ordem $n \times 1$ de forma que $n = IJ$;

$\boldsymbol{X_1}$ vetor de 1's;

μ representa a média geral;

$\boldsymbol{\varepsilon}$ vetor de erros aleatórios de dimensão $n\times$ 1.

Considerando este modelo sob a suposição de que os erros são independentes e identicamente distribuídos segundo uma normal com média zero e matriz de covariâncias $\boldsymbol{I}\sigma^2 = \boldsymbol{V}\sigma^2$, tem - se que um estimador para a média geral é dado por

$$\hat{\mu} = (\boldsymbol{X_1'X_1})^{-1}\boldsymbol{X_1Y}$$

E no caso em que $\boldsymbol{V} \neq \boldsymbol{I}$, como no caso espacial em que $\boldsymbol{V}$ representa a matriz de correlação espacial, a estimativa passa a ser obtida pelo método de mínimos quadrados generalizados

$$\hat{\mu}_V = (\boldsymbol{X_1'V^{-1}X_1})^{-1}\boldsymbol{X_1V^{-1}Y}.$$

Porém, ao considerar o caso em que os efeitos de tratamento são não nulos, o modelo mais adequado é o representado na equação 2.4, isto é,

$$\mathbf{Y} = \mathbf{X_1}\mu + \mathbf{X_2}\boldsymbol{\psi} + \boldsymbol{\varepsilon},$$

adotando uma parametrização alternativa, sendo $\mu_i = \mu + \psi_i$, o modelo pode ser apresentado da seguinte forma

$$\boldsymbol{Y} = \boldsymbol{X_2\mu} + \boldsymbol{\varepsilon}, \tag{2.22}$$

em que $\boldsymbol{\mu}' = [\mu_1, \mu_2, \ldots, \mu_k]$ representa o vetor de médias espaciais dos tratamentos e $\boldsymbol{X_2}$ representa a matriz de incidência dos efeitos. Com essa parametrização, o vetor $\boldsymbol{\mu}$ pode ser estimado de modo semelhante ao que foi apresentado para o modelo 2.21

$$\hat{\boldsymbol{\mu}} = (\boldsymbol{X_2'X_2})^{-1}\boldsymbol{X_2Y}$$

quando $\boldsymbol{V = I}$, e

$$\hat{\boldsymbol{\mu}}_V = (\boldsymbol{X_2'V^{-1}X_2})^{-1}\boldsymbol{X_2V^{-1}Y},$$

quando $\boldsymbol{V \neq I}$, que é denominado vetor de médias espaciais dos tratamentos.

Na presença de tendência, o modelo 2.22 passa a incluir as coordenadas como covariáveis

$$\boldsymbol{Y} = \boldsymbol{X_2\mu} + \boldsymbol{D\theta} + \boldsymbol{\varepsilon}.$$

Nesse caso, o estimador da média passa a ser

$$\hat{\boldsymbol{\mu}}_V = (\boldsymbol{X_2'V^{-1}X_2})^{-1}\boldsymbol{X_2'V^{-1}Y} - (\boldsymbol{X_2V^{-1}X_2})^{-1}\boldsymbol{X_2V^{-1}D}\hat{\boldsymbol{\theta}}$$
$$\hat{\boldsymbol{\theta}} = (\boldsymbol{D'PD})^{-1}\boldsymbol{D'PY},$$

em que $\boldsymbol{P} = \boldsymbol{V^{-1}} - \boldsymbol{V^{-1}X_2(X_2'V^{-1}X_2)^{-1}X_2'V^{-1}}$.

Antes de apresentar os procedimentos de comparações múltiplas, é necessário definir a matriz de contrastes $\boldsymbol{C}$. Essa matriz tem dimensão $p \times k$ em que $p = \frac{I!}{2(I-2)!}$ representa o número de combinações de médias tomadas duas a duas possíveis, sendo I o número de tratamentos. Dessa forma, cada linha de $\boldsymbol{C}$ representa os coeficientes de um contraste de médias avaliado.

Por exemplo,ao considerar $I = 4$ tratamentos, a matriz $\boldsymbol{C}$ é dada por

$$\boldsymbol{C} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

Segundo Nogueira (2017), a matriz de covariâncias de todas as comparações duas a duas no caso em que não é considerada a tendência espacial é definida por

$$\boldsymbol{M} = Var(\boldsymbol{C}\hat{\boldsymbol{\mu}}_{\boldsymbol{V}}) = \boldsymbol{C}(\boldsymbol{X_2'V^{-1}X_2'})^{-1}\boldsymbol{C}\sigma^2, \tag{2.23}$$

Já no caso do modelo que considera a tendência espacial é dada por

$$\boldsymbol{M} = Var(\boldsymbol{C}\hat{\boldsymbol{\mu}}_{\boldsymbol{V}}) =$$
$$\boldsymbol{C}(\boldsymbol{X_2'V^{-1}X_2'})^{-1}\boldsymbol{C'}\sigma^2 + \boldsymbol{C}(\boldsymbol{X_2'V^{-1}X_2})^{-1}\boldsymbol{X_2'V^{-1}AV^{-1}X_2}(\boldsymbol{X_2'V^{-1}X_2})^{-1}\boldsymbol{C'}, \tag{2.24}$$

em que $\boldsymbol{A} = \boldsymbol{D}(\boldsymbol{D'PD})^{-1}\boldsymbol{D'}\sigma^2$.

2.7.2.2 Teste baseado na distribuição t multivariada

Primeiramente, se define um vetor com I contrastes de interesse $\boldsymbol{c'} = [c_1, \ldots, c_I]$ sendo $\sum_{i=1}^{I} c_i = 0$, de modo que $\boldsymbol{c'\mu}_V$ representa uma combinação linear entre as médias espaciais.

Como o interesse é em fazer p comparações múltiplas, as quais avaliam a diferença entre médias espaciais tomadas duas a duas, deve-se construir p vetores de contrastes $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_p$, e consequentemente devem ser formuladas p hipóteses definidas por

$$H_{0i} = \boldsymbol{c'\mu}_V = 0, \quad \text{para} \quad i = 1, 2, \ldots, p$$

que podem ser avaliadas por meio da seguinte estatística de teste

$$t_{ci} = \frac{\boldsymbol{c}_i'\hat{\boldsymbol{\mu}}_V}{\sqrt{Var(\boldsymbol{c}_i'\hat{\boldsymbol{\mu}}_V)}}.$$

Pode-se também fazer uso da matriz a matriz $\boldsymbol{M}$, definida por 2.23 no caso do modelo sem tendência e por 2.19 para o modelo com tendência, desse modo tem-se que a estatística t_{ci} pode ser reescrita como

$$t_{ci} = \frac{\boldsymbol{c}_i'\hat{\boldsymbol{\mu}}_V}{\sqrt{\hat{m}_{ii}}}.$$

Sob H_0, cada estatística t_{ci} segue uma distribuição t de Student com ν graus de liberdade dado pelo grau de liberdade do resíduo do modelo utilizado, nesse caso $\nu = n - I - s$, em que s representa a dimensão das coordenadas espaciais utilizadas, neste trabalho $s = 2$.

No entanto como as p comparações não são independentes, sua distribuição conjunta segue uma t de Student multivariada com ν graus de liberdade e matriz de correlação dada por

$$\boldsymbol{R} = \boldsymbol{S}^{-\frac{1}{2}}\boldsymbol{M}\boldsymbol{S}^{-\frac{1}{2}},$$

em que $\boldsymbol{S}^{-\frac{1}{2}}$ representa uma matriz diagonal $p \times p$, cujos elementos da diagonal principal são dados por $(\sqrt{1/w_{ii}})$, em que w_{ii} representa a variância do contraste i.

2.7.2.3 Teste baseado na distribuição amplitude estudentizada

Esse teste é feito de modo semelhante ao teste proposto por Tukey, ou seja, é realizada a estimação dos contrastes e em seguida estes são comparados com uma diferença média significativa (DMS) baseada na distribuição da amplitude estudentizada,

$$\text{se} \quad |\boldsymbol{c}_i'\hat{\boldsymbol{\mu}}_V| \geq DMS_i, \quad \text{há diferença significativa entre as médias}$$

em que

$$DMS_i = q(\alpha, I, \nu)\sqrt{0,5w_{ii}}, \quad \text{para} \quad i = 1, \ldots, p;$$

sendo $q(\alpha, I, \nu)$ o quantil da distribuição da amplitude estudentizada considerando um nível de significância α, e ν o grau de liberdade do resíduo modelo.

2.7.2.4 Agrupamento de médias Scott-Knott

Esse procedimento trata-se de uma técnica de agrupamento de médias, baseado na razão de verossimilhanças. Assim, em um grupo de I médias, avaliam-se as seguintes hipóteses

$$\begin{cases} H_0: \ \mu_1 = \mu_2 = \cdots = \mu_I = \mu \\ H_1: \ \mu_1 = \mu_2 = \cdots = \mu_{k_1} = \mu_{g_1} \quad \text{e} \quad \mu_{k_1+1} = \mu_{k_1+2} = \cdots = \mu_I = \mu_{g_2}. \end{cases}$$

em que $k_1, k_2 \geq 1$, com $k_2 = I - k_1$, são dados pelo número de médias contidas nos grupos 1 e 2. Já μ_{g_1} e μ_{g_2}, representam, respectivamente a média dos grupos 1 e 2.

Sob H_1, é possível dividir um grupo de I médias em dois subgrupos, cujas médias são denominadas por μ_{g_1} e μ_{g_2}. Assim, o seguinte modelo pode ser usado para descrever os dados:

$$\boldsymbol{Y} = \boldsymbol{K}\boldsymbol{g} + \boldsymbol{\varepsilon},$$

em que

$$\mathbf{K} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} \Rightarrow \; k_1 \\ \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} \Rightarrow \; k_2 \end{matrix} \quad ;$$

e além disso, $\boldsymbol{g} = [\mu_{g_1}, \mu_{g_2}]'$ é o vetor que contém as médias dos dois grupos formados e ε o erro aleatório normalmente distribuído com vetor de média zero e matriz de covariâncias $\boldsymbol{\Sigma} = \boldsymbol{V}\sigma^2$.

A estatística utilizada para avaliar as hipóteses desse teste é dada por

$$\lambda = \frac{\pi}{2(\pi - 2)} nB,$$

em que π representa o número irracional com valor aproximado de $3,141592$, n é o número de observações e B é dado por

$$B = \frac{\boldsymbol{Y'V^{-1}K(K'V^{-1}K)^{-1}K'V^{-1}Y - Y'V^{-1}X_1(X_1'V^{-1}X_1)^{-1}X_1'V^{-1}Y}}{\boldsymbol{Y'V^{-1}Y - Y'V^{-1}X_1(X_1V^{-1}X_1)^{-1}X_1'V^{-1}Y}}.$$

Ao comparar I médias, particionadas em dois grupos distintos, a hipótese nula de igualdade entre as médias de dois grupos deve ser rejeitada se, e somente se

$$\lambda = \frac{\pi}{2(\pi - 2)} nB > \chi^2(\nu_0, \alpha), \tag{2.25}$$

sendo $\nu_0 = I/(\pi - 2)$ o grau de liberdade da distribuição qui-quadrado, e α o nível de significância.

Dessa forma, a aplicação do teste pode ser realizada a partir das seguintes etapas

(i) Ordenar as médias espaciais dos tratamentos e, a partir daí, dividir os tratamentos em dois grupos, para as $I - 1$ partições possíveis;

(ii) Ordenar o vetor $\boldsymbol{Y}$, segundo a ordem dos tratamentos, sendo que $\boldsymbol{Y}$ deve ser composto pelas observações provenientes dos tratamentos presentes em cada partição;

(iii) determinar a partição da matriz $\boldsymbol{\Sigma}$, a qual se refere às covariâncias entre as observações contidas em $\boldsymbol{Y}$, conforme especificado no item anterior;

(iv) Calcular o valor de B, para todas as partições possíveis;

(v) calcular a estatística de teste conforme equação 2.25 e comparar com o quantil da distribuição $\chi^2(\nu_0, \alpha)$. Se essa estatística for maior que o quantil determinado, rejeita-se a hipótese de que os dois grupos apresentam médias iguais;

(vii) no caso de rejeitar essa hipótese, os dois subgrupos formados serão independentemente submetidos aos passos (ii) a (vi). O processo em cada subgrupo se encerra ao não se rejeitar H_0 no passo (vi) ou quando cada subgrupo contiver apenas uma média.

2.7.3 Abordagem via Modelo Espacial Autorregressivo

Esta metodologia foi proposta por Long (1996) e, diferentemente da abordagem geoestatística, este método baseia-se em uma transformação na variável resposta ($\boldsymbol{Y_{adj}}$) a fim de neutralizar os efeitos da autocorrelação espacial. Para isso, essa autocorrelação é estimada utilizando-se o modelo apresentado em 2.6, em seguida, subtrai-se de cada observação o valor estimado da autocorrelação espacial . No caso do modelo SAR (equação 2.19), isso pode ser feito por meio da seguinte relação

$$\boldsymbol{Y_{adj}} = \boldsymbol{Y} - (\hat{\rho}\boldsymbol{W}\boldsymbol{Y} - \hat{\rho}\beta_0) \tag{2.26}$$

em que $\boldsymbol{Y_{adj}}$ é o vetor contendo os valores transformados da variável resposta, $\boldsymbol{Y}$ é o vetor de valores observados para a resposta do experimento, β_0 é a média geral, $\boldsymbol{W}$ é a matriz de vizinhança e $\hat{\rho}$ é a estimativa do parâmetro autorregressivo. Com $\boldsymbol{Y_{adj}}$ pode-se empregar os métodos usuais da análise de variância. Assim, os procedimentos de comparações múltiplas são realizados diretamente em $\boldsymbol{Y_{adj}}$ da forma usual.

2.7.4 Análise de Resíduos dos Modelos com a Inclusão da Informação Espacial

Os testes de suposição dos modelos em que a informação espacial é incorporada são realizados de modo análogo ao que foi apresentado na seção 2.1.6, porém para o caso

da geoestatística utiliza-se o resíduo espacialmente independente fornecido pelo modelo 2.18 e, para o caso do modelo espacial autorregressivo, utiliza-se o resíduo proveniente do modelo resultante no qual foi utilizada a variável ajustada $\boldsymbol{Y_{adj}}$ de acordo com a equação 2.26.

Na abordagem geoestatística, há a suposição de que o resíduo do modelo final é espacialmente independente e isso pode ser verificado por meio do índice I de Moran, conforme apresentado a seguir.

2.7.4.1 Índice I de Moran

O índice I de Moran é uma medida global da autocorrelação espacial, pois indica o grau de associação espacial presente no conjunto de dados. O índice é utilizado em um teste que assume como hipótese nula a ausência de autocorrelação espacial, ou seja,

$$\begin{cases} H_0: & \text{Não há dependência espacial} \\ H_1: & \text{Há dependência espacial.} \end{cases}$$

O índice I de Moran é dado por

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(r_i - \bar{r})(r_j - \bar{r})}{\sum_{i=1}^{n}(r_i - \bar{r})^2}$$

nesse caso r_i é o valor do atributo na área i, $\bar{r}$ é o valor médio do atributo na região do estudo e w_{ij} são os elementos da matriz de vizinhança $\boldsymbol{W}$.

Para avaliar a significância estatística do teste sem pressupostos com relação à distribuição, pode-se usar um teste de pseudo-significância. Para isso, é gerada uma distribuição empírica do índice I por meio de permutação e compara-se com o valor obtido para I inicialmente, se o valor for considerado um extremo quando comparado a distribuição simulada, rejeita-se a hipótese de que há independência espacial (CÂMARA; CARVALHO, 2004).

3 MATERIAL E MÉTODOS

Esta seção é dedicada à descrição dos materiais e metodologias utilizados ao longo do desenvolvimento do trabalho.

3.1 Dados de Experimento com Candeias

A ilustração dos resultados apresentados foi feita utilizando os dados de um experimento com candeia (*Eremanthus erythropappus*) instalado no ano de 2004 em Baependi - MG, descrito em Nogueira et al. (2015).

A candeia é uma espécie comumente encontrada no Brasil que se desenvolve em locais com solo pouco férteis, rasos e, predominantemente em áreas de campo de altitude variando de 900 a 1700 m. Vários produtos são originados a partir da exploração comercial da candeia, dentre os quais se destacam a produção de moirão de cercas e a extração do óleo essencial cujo componente principal é o alfabisabolol, que possui alta aplicação na indústria farmacêutica, uma vez que possui propriedades antiflogísticas, antibacterianas, antimicóticas, dermatológicas e espasmódicas (LONGHI et al., 2009).

O objetivo foi avaliar o efeito de 13 diferentes tipos de adubação no crescimento das árvores. Para isso, o experimento foi instalado em uma área de 1,5 ha de acordo com um delineamento em blocos casualizados (DBC) constituído de 4 blocos, em que cada parcela era formada por 50 plantas úteis e 4 utilizadas como bordadura, espaçadas em 2,0 m x 5,0 m.

Os seguintes tratamentos foram avaliados

1. Testemunha absoluta;
2. Completo: Ca + NPK + micros;
3. N: Ca + NPK + micros;
4. P205: Ca + NK + micros;
5. K20: Ca + NP + micros;
6. Micros: PK + micros;
7. Calcário: PK + micros;

8. Completo + adubação orgânica;
9. Testemunha absoluta + adubação orgânica;
10. Adubo formulado NPK (8-28-16);
11. Completo 2: Ca + gesso + NPK + micros;
12. Completo 3: Ca + gesso + NP (100g de superfosfato simples + 200g de fosfato reativo) K + micros;
13. Completo 4: Ca + adubo formulado NPK (6-30-6) + NK + micros.

Os dados foram coletados no ano de 2012, na ocasião foram medidos os diâmetros a 1,30 m do solo (em centímetros) e as alturas H (em metros) de todas as árvores que integraram o experimento. Uma vez que houve perdas de algumas plantas no decorrer do experimento, a distribuição do número de árvores por tratamento foi conforme mostrado na Tabela 3.1.

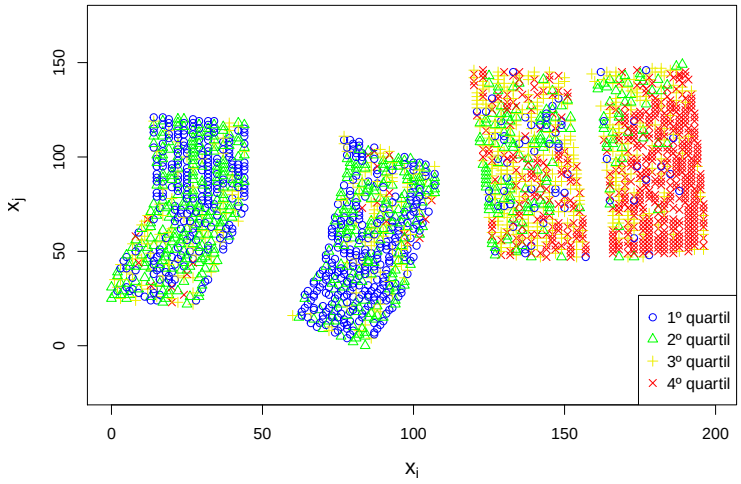
Tabela 3.1 – Número de árvores por tratamento

Tratamento	1	2	3	4	5	6	7	8	9	10	11	12	13
Repetições	191	155	161	181	161	156	152	160	188	183	165	149	177

Fonte: Do autor (2019)

Dessa maneira, a distribuição dos valores de altura das árvores (em metros) dentro da área experimental é sumarizada por meio de quartis na Figura 3.1, em que as linhas verticais em cada um dos quatro blocos representam os 13 tratamentos.

Figura 3.1 – Valores das alturas das árvores de candeia separados por quartis



Fonte: Do autor (2019)

A análise da Figura 3.1 revela indícios da presença de autocorrelação espacial, pois nota-se um padrão de comportamento na altura das árvores de acordo com sua localização. Observa-se que existe uma relação entre o valor das coordenadas e o valor da altura das árvores, isto é, nota-se que as árvores mais altas estão localizadas nos lugares em que e o valor da coordenada X é maior e a coordenada Y é menor, caracterizando além de uma dependência, uma tendência em função do espaço.

3.2 Softwares utilizados

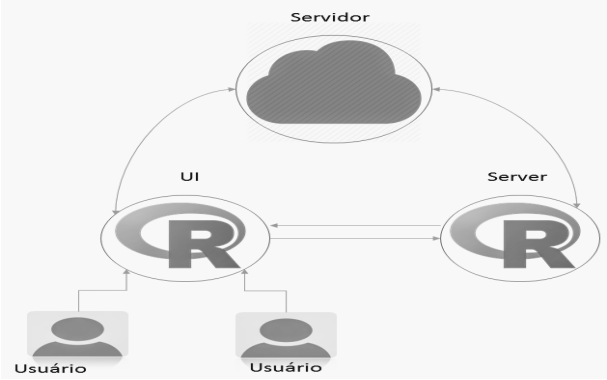
As funções computacionais utilizadas foram desenvolvidas por Rossoni (2011), Nogueira (2013, 2017) em linguagem R e aprimoradas pelo autor do presente trabalho para comporem o produto final desenvolvido neste trabalho.

Um dos pacotes empregados no decorrer deste trabalho foi o *shiny* (CHANG et al., 2017), o qual se destaca por permitir o desenvolvimento de interfaces gráficas por meio de aplicações *web* de forma prática e intuitiva. Em termos gerais, esse pacote utiliza a linguagem R para simplificar os comandos que normalmente seriam utilizados em linguagens de programação voltadas ao desenvolvimento *web*.

As aplicações desenvolvidas em *shiny* são baseadas em dois *scripts*: *ui* e *server*. O *script ui* é o responsável pela interface gráfica da aplicação, nele é possível especificar dentre outras coisas os ícones e botões que ficarão visíveis e farão a interação com o usuário,

enquanto o *server* é o núcleo da aplicação, onde ficam armazenados os códigos em R que responderão às interações do usuário retornando os resultados de cálculos estatísticos e gráficos que serão exibidos durante determinada sessão. A Figura 3.2 ilustra de forma resumida o esquema de funcionamento de uma aplicação *shiny*.

Figura 3.2 – Esquema de funcionamento da aplicação em *shiny*



Fonte: Do autor (2019)

No caso específico deste trabalho, a aplicação desenvolvida por meio do *shiny* foi disponibilizada dentro do pacote criado, dessa maneira é utilizado o próprio computador do usuário como um servidor local. Além disso, a aplicação foi disponibilizada na internet, de forma que qualquer usuário possa acessá-la sem a necessidade de instalação de programas, evitando assim problemas de compatibilidade em relação aos diversos sistemas operacionais existentes. Para isso, ela foi hospedada no servidor do *rstudio* que fornece o serviço de hospedagem e toda a estrutura necessária para seu funcionamento.

Além do *shiny*, outros pacotes, os quais são apresentadas na Tabela 3.2, constituem uma parte importante do produto final aqui apresentado, pois é por meio deles que são realizados grande parte dos cálculos que compõem as análises.

Tabela 3.2 – Outros pacotes utilizados no desenvolvimento do trabalho.

Pacote	Uso
geoR [1]	Funções para modelagem geoestatística
spdep [2]	Funções bases que compõem o modelo autorregressivo espacial
multcomp [3]	Funções empregadas no teste T multivariado
multcompView [4]	Funções para visualização dos procedimentos de comparação múltipla
Matrix [5]	Funções para cálculos matriciais
MASS [6]	Funções para cálculo de inversas generalizadas

Fonte: Do autor (2019)

1 - Ribeiro Júnior e Diggle (2016)
2 - Bivand, Pebesma e Gomez-Rubio (2013)
3 - Hothorn et al. (2016)
4 - Graves, Piepho e Dorai-Raj (2015)
5 - Bates e Maechler (2018)
6 - Venables e Ripley (2013)

4 RESULTADOS E DISCUSSÃO

Nesta seção serão apresentados os resultados obtidos neste trabalho. Inicialmente, será apresentada a biblioteca desenvolvida e suas funções, em seguida a ferramenta interativa que compõe a biblioteca e, por fim, suas funcionalidades serão ilustradas por meio da análise de um experimento real.

4.1 Apresentação do pacote spANOVA e suas funções

A biblioteca desenvolvida recebeu o nome spANOVA, uma vez que na linguagem R o prefixo sp é comumente adotado em funções e pacotes que lidam com alguma abordagem de estatística espacial. Sua instalação é feita utilizando-se o comando `install.packages("spANOVA")`.

Nas subseções seguintes, são apresentados os passos para a realização das análises considerando as duas abordagens e em seguida as principais funções que foram implementadas no pacote spANOVA para possibilitar a execução no R. A documentação com maiores detalhes sobre as funções e seus respectivos argumentos está disponível no Apêndice A.

4.1.1 Abordagem geoestatística

A abordagem geostatística pode ser implementada por meio dos seguinte passos:

1. Construir o modelo usual de análise de variância e obter seus resíduos;
2. Usar os resíduos do passo anterior e construir o semivariograma;
3. Ajustar um modelo teórico ao semivariograma para descrever a variabilidade espacial;
4. Utilizar o algoritmo proposto por Pontes e Oliveira (2004) para tornar as estimativas dos parâmetros do semivariograma mais precisas;
5. Construir a matriz de covariâncias do modelo de acordo com o que foi mostrado na subseção 2.7.1;
6. Utilizar os resultados apontados na subseção 2.7.1 para obter as somas de quadrados e consequentemente a análise de variância.

Esses passos são executados pelo pacote spANOVA por meio das funções descritas a seguir.

- `spVariog`

Esta função é utilizada para a execução dos passos 1 e 2. Ela recebe como argumentos principais um vetor contendo os valores da variável resposta do experimento, o vetor identificando cada um dos tratamentos associados às respostas, um argumento lógico (verdadeiro ou falso) indicando a presença de tendência espacial, a máxima distância a ser considerada no cálculo do semivariograma, o tipo de delineamento podendo ser o Delineamento Inteiramente ao Acaso (na sigla em inglês, "crd") ou o Delineamento em Blocos Casualizados (em inglês, "rcbd"). Além disso, como há muitos cálculos envolvendo as coordenadas, existe um argumento lógico o qual indica se as coordenadas devem ou não serem deslocadas para a origem, evitando assim problemas computacionais. Esta também pode receber outros argumentos adicionais que podem ser passados para a função `variog` do pacote `geoR`. Sua saída é um objeto da classe[1] `spVariog` que possui o método[2] `plot` disponível para a visualização da nuvem de pontos do semivariograma.

- `spVariofit`

Essa função implementa o passo 3 do algoritmo da abordagem geoestatística. Com ela se estima os parâmetros do modelo através do método de mínimos quadrados. Basicamente, ela recebe três argumentos, sendo o primeiro um objeto da classe `spVariog`, em seguida recebe um argumento do tipo *string* informando qual modelo será adotado para descrever a estrutura da matriz de correlação, um argumento indicando o tipo de ponderação que será utilizado na estimação ("equal": mínimos quadrados ordinários, "cressie": mínimos quadrados ponderados), e a distância máxima a ser considerada nos cálculos. Sua saída é um objeto da classe `spVariofit` a qual possui implementado o método `lines`, que combinado com o método `plot` da classe `spVariog`, fornece uma visualização do modelo ajustado à nuvem de pontos do semivariograma.

- `aovGeo`

[1] Em suma, uma classe pode ser entendida como um objeto que é compatível com funções genéricas específicas para o qual ela foi programada tais como `plot`, `summary` etc.

[2] Um método em R é uma função genérica programada para uma determinada classe.

Os passos 4 a 6 são implementados nesta função que recebe dois argumentos, o primeiro um objeto da classe `spVariofit` e o segundo um número no intervalo (0,1], necessário para definir a proporção da máxima distância a ser utilizada na construção do semivariograma do algoritmo apresentado por Pontes e Oliveira (2004). Essa função retorna um objeto da classe `GEOanova` a qual conta com o método `anova` para a visualização dos resultados em forma de tabela.

4.1.2 Abordagem via modelos autorregressivos

A execução de uma análise segundo a abordagem por meio do modelo espacial autorregressivo, é realizada de acordo com os passos seguintes:

1. Ajustar o modelo espacial autorregressivo (SAR) escolhendo o valor do raio que minimiza o critério de informação de Akaike (AIC);
2. Obter a estimativa do parâmetro autorregressivo;
3. Fazer a transformação na variável resposta de acordo com o que foi mostrado na subseção 2.7.3;
4. Proceder a análise de variância usual utilizando a variável resposta transformada.

As funções descritas a seguir permitem a implementação da abordagem autorregressiva por meio do pacote spANOVA.

- `aovSar.crd`

Esta função implementa os passos de 1 a 4 do algoritmo, quando o delineamento utilizado é o inteiramente casualizado. Ela recebe quatro argumentos: o primeiro é um vetor contendo os valores da variável resposta do experimento, o segundo argumento é também um vetor especificando os tratamentos associados a cada um dos valores contidos no vetor da variável resposta, o terceiro argumento é uma matriz de dimensão $n \times 2$ contendo as coordenadas espaciais das unidades experimentais e, finalmente o último argumento é um vetor de sequências de raios a serem testados. Este último não precisa necessariamente ser especificado e, por padrão, é criada uma sequência de dez raios, de tal forma que o valor irá variar entre 0 e metade da máxima distância entre as unidades experimentais. A saída da função é um objeto da classe `SARanova` que também possui o método `anova` implementado para a visualização dos resultados.

- `aovSar.rcbd`

Esta função é similar à função `aovSar.crd`, porém é utilizada quando o delineamento for em blocos casualizados. Os argumentos são os mesmos, é necessário apenas informar um argumento a mais, denominado `block`, que receberá um vetor, indicando o bloco a que cada observação pertence.

4.1.3 Procedimentos de comparações múltiplas

Além das funções básicas disponíveis para a realização da análise de variância incorporando a dependência espacial, o pacote spANOVA também conta com funções para a execução de procedimentos de comparações múltiplas. Na Tabela 4.1, são apresentados os procedimentos implementados, bem como os argumentos necessários para sua utilização.

Tabela 4.1 – Funções utilizadas para os procedimentos de comparação múltipla

Procedimento	Função	Argumentos
Tukey	`spTukey`	Objeto da classe `GEOanova` ou `SARanova`, nível de significância
T multivariado	`spMVT`	Objeto da classe `GEOanova` ou `SARanova`, nível de significância
Scott-Knott	`spScottKnott`	Objeto da classe `GEOanova` ou `SARanova`, nível de significância

Fonte: Do autor (2019)

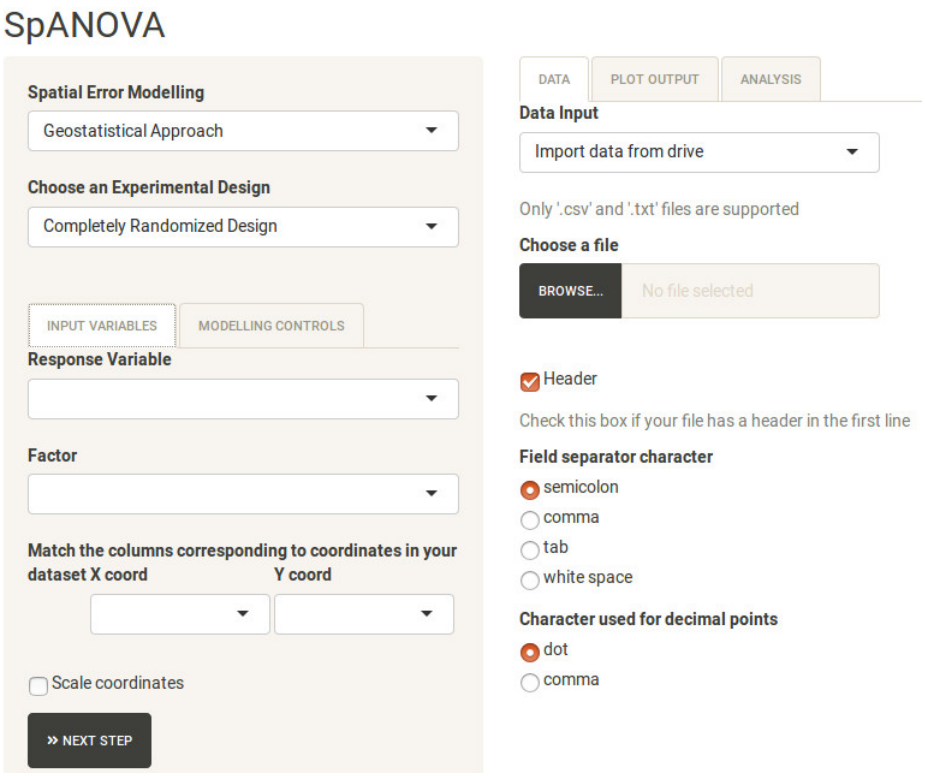
A saída da função é um *data frame* com as médias apresentadas em ordem decrescente seguidas por letras, em que as médias significativamente diferentes ao nível de significância especificado são identificadas por letras distintas.

4.2 Apresentação do ambiente interativo

Como alternativa à realização das análises descritas anteriormente, via linha de comando, está também disponível no pacote criado um ambiente gráfico interativo, que fornece ao usuário uma interface mais amigável para a realização de análises mais simples, servindo também como uma visão geral das funcionalidades dos pacotes para potenciais usuários.

Essa interface é acessada executando-se a função `spANOVAapp()` ou acessando o endereço <https://spanova.shinyapps.io/spanova/>. Ao acessar a aplicação, o usuário encontrará algumas opções que são utilizadas para iniciar suas análises conforme a Figura 4.1.

Figura 4.1 – Página inicial da aplicação



Fonte: Do autor (2019)

No painel lateral esquerdo apresentado na Figura 4.1, há a opção para a escolha do delineamento experimental a ser utilizado e a abordagem para modelar o erro espacialmente dependente, abaixo é possível definir a variável resposta e a variável fator (tratamentos) do experimento quais estarão disponíveis após o carregamento do conjunto de dados.

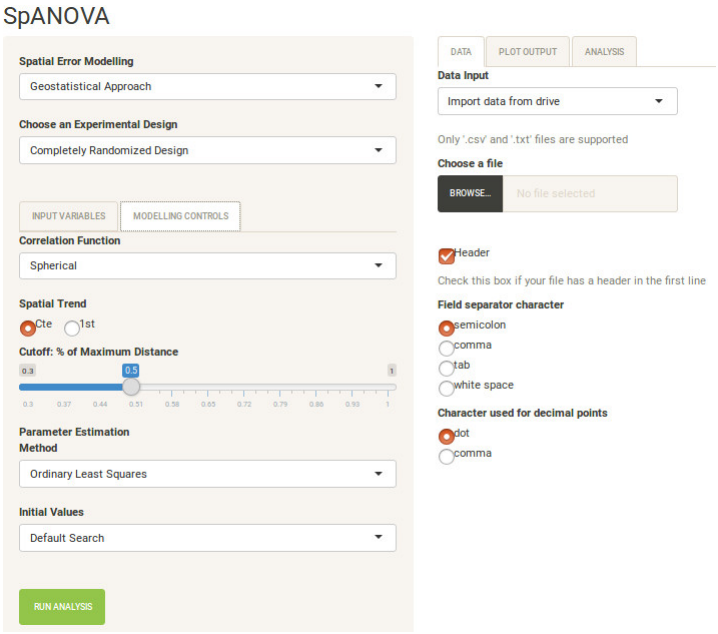
No painel lateral direito, observam-se três abas disponíveis. A primeira aba denominada *data*, permite ao usuário carregar seu conjunto de dados em formato `.txt` ou `.csv` e conta com algumas opções para que isso seja feito de forma correta e, além disso, é possível utilizar os dados incluídos no pacote como exemplos.

Para carregar um conjunto de dados, é necessário acionar o botão *browse*. Com isso é aberta uma janela que permite ao usuário navegar pelas pastas de seu dispositivo a fim de selecionar o arquivo de dados. Após essa seleção os dados são exibidos em uma tabela para que se tenha uma visualização das variáveis e a confirmação de que os dados foram carregados de forma correta.

Em seguida, é necessário configurar algumas características inerentes à modelagem. Na abordagem geoestatística, por exemplo, é preciso que se faça a escolha do mo-

delo teórico de semivariograma e a definição da distância máxima considerada conforme mencionado na subseção 2.2.3. Isso é realizado utilizando a opção *modelling controls* no painel lateral esquerdo, como mostrado na Figura 4.2. Já na abordagem autorregressiva, não é necessário especificar nenhum parâmetro, pois na definição do raio é adotado o valor padrão da função `aovSar.crd` (ou `aovSar.rcbd`), ou seja, uma sequência com dez valores variando de 0 até metade da máxima distância observada entre as amostras.

Figura 4.2 – Parâmetros de modelagem da abordagem geoestatística



Fonte: Do autor (2019)

Nas duas abas restantes, são exibidos alguns gráficos descritivos e, por fim, a tabela de análise de variância bem como o procedimento de comparações múltiplas escolhido. Ao final da análise é possível gerar um relatório no formato docx ou PDF (disponível apenas na versão *online*) com as informações organizadas em tabelas.

4.2.1 Análise através da abordagem geoestatística

Considerando a abordagem geoestatística, foram utilizadas as informações da altura (h) das parcelas do experimento. Dessa forma, foi adotado o modelo em blocos. Assim, o modelo linear é representado como

$$\boldsymbol{Y} = \boldsymbol{X}_1\mu + \boldsymbol{X}_2\boldsymbol{\psi} + \boldsymbol{X}_3\boldsymbol{\eta} + \boldsymbol{\varepsilon}$$

em que

$\boldsymbol{Y}$ representa o vetor contendo a altura das árvores, de dimensão 2179×1;

$\boldsymbol{X}_1$ vetor de 1's, de dimensão 2179×1;

μ constante inerente a todas as observações;

$\boldsymbol{X}_2$ matriz de incidência dos tratamentos, de dimensão 2179×13;

$\boldsymbol{\psi}$ vetor de dimensão 13×1, que representa os efeitos dos 13 tipos de adubação analisados;

$\boldsymbol{X}_3$ matriz de incidência dos blocos, de dimensão 2179×4;

$\boldsymbol{\eta}$ representa o vetor de parâmetros dos efeitos dos blocos, de dimensão 4×1;

$\boldsymbol{\varepsilon}$ vetor de dimensão 2179×1, representando o erro aleatório tal que $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, em que $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{V}$, sendo sua estrutura definida de acordo com o que foi apresentado na subseção 2.7.1.

Neste caso, as hipóteses de interesse são as seguintes

$$\begin{cases} H_0: & \psi_1 = \psi_2 = \cdots = \psi_j = 0 \\ H_1: & \psi_j \neq 0 \quad \text{para pelo menos um } j, \text{ com } j = 13, \end{cases}$$

ou de forma equivalente,

$$\begin{cases} H_0: & \text{Os tratamentos não apresentam diferença significativa entre si} \\ H_1: & \text{Pelo menos um dos tratamentos difere dos demais.} \end{cases}$$

Adicionalmente, pode-se testar as hipóteses

$$\begin{cases} H_0: & \eta_1 = \eta_2 = \cdots = \eta_j = 0 \\ H_1: & \eta_j \neq 0 \quad \text{para pelo menos um } j, \text{ com } j = 4, \end{cases}$$

que podem ser interpretadas como

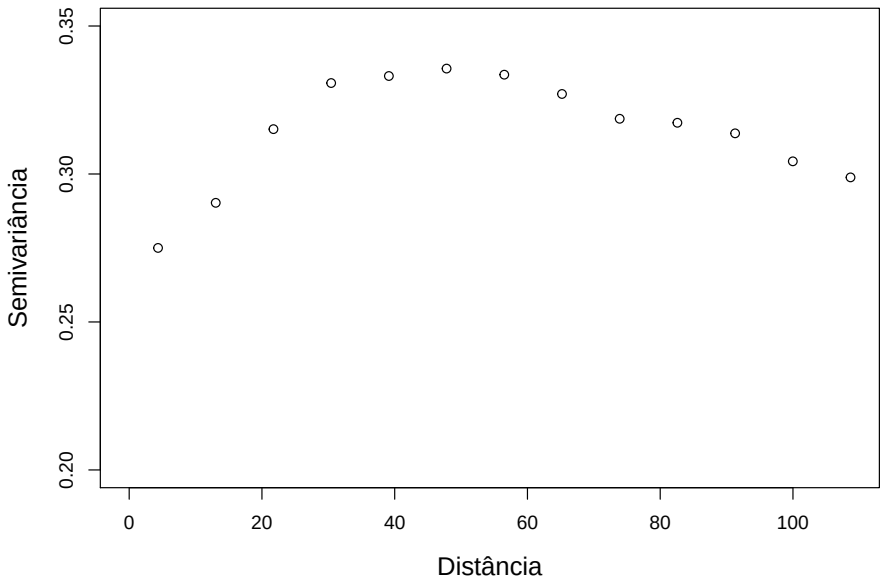
$$\begin{cases} H_0: & \text{Os blocos não apresentam diferença significativa entre si} \\ H_1: & \text{Pelo menos um dos blocos difere dos demais.} \end{cases}$$

Para dar prosseguimento à análise, e partindo do pressuposto que o fenômeno é isotrópico, é necessário inicialmente estimar o semivariograma. Para isso, após a leitura

dos dados no R, utilizaram-se os comandos abaixo para produzir o gráfico mostrado na Figura 4.3.

```
library(spANOVA)
geodados <- as.geodata(candeia, coords.col = 5:6, data.col = 8,
covar.col = c(3,2))
dist <- summary(geodados)[[3]][[2]]*0.50
variog <- spVariog(geodata = geodados, scale = TRUE,
max.dist = dist, design = "rcbd")
plot(variograma, ylab = "Semivariância", xlab = "Distância",
ylim=c(0.2,0.35))
```

Figura 4.3 – Semivariograma experimental



Fonte: Do autor (2019)

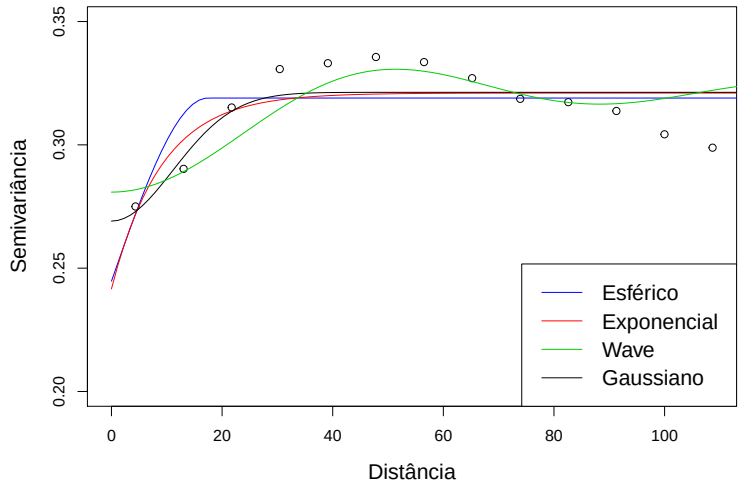
O semivariograma apresentado na Figura 4.3 foi estimado com base nos resíduos do modelo usual de análise de variância, considerando 50% da distância máxima observada entre as amostras. O passo seguinte, consistiu em encontrar um modelo teórico para descrever a variabilidade espacial observada no semivariograma utilizando a estimação por mínimos quadrados ordinais. Aqui foram ajustados os quatro modelos mais utilizados na literatura: Exponencial, Esférico, Gaussiano e Wave. Os comandos utilizados para o

ajuste dos modelos são mostrados a seguir, e suas saídas podem ser visualizados na Figura 4.4.

```
ols1 <- spVariofit(variograma, cov.model = "spherical",
weights = "equal", max.dist = dist)
ols2 <- spVariofit(variograma, cov.model = "exponential",
weights = "equal", max.dist = dist, ini.cov.pars = c(0.4,18))
ols3 <- spVariofit(variograma, cov.model = "gaussian",
weights = "equal", max.dist = dist, ini.cov.pars = c(0.4,18))
ols4 <- spVariofit(variograma, cov.model = "wave",
weights = "equal", max.dist = dist, ini.cov.pars = c(0.4,18))
```

```
lines(ols1, col = 4)
lines(ols2, col = 2)
lines(ols3, col = 3)
lines(ols4, col = 1)
legend("bottomright", legend = c("Esférico", "Exponencial",
"Gaussiano", "Wave"), lty = c(1,1,1),
col = c(4, 2, 3, 1), bty = "n")
```

Figura 4.4 – Ajuste dos modelos ao semivariograma



Fonte: Do autor (2019)

Nota-se que os modelos esférico, exponencial e gaussiano apresentam comportamento semelhante. No entanto, visualmente o modelo wave é o que fornece uma representação mais razoável dos pontos observados no semivariograma. Porém, como a análise gráfica é um critério subjetivo, procedeu-se a seleção do modelo por meio da validação cruzada, na qual se avaliou o erro médio reduzido ($\bar{ER}$) e o seu desvio padrão (S_{er}). Os resultados foram obtidos pela função `spCrossValid`, conforme mostrado a seguir, e os valores obtidos encontram-se na Tabela 4.2.

```
# Validação cruzada
ols1.cv <- spCrossvalid(ols1)
ols2.cv <- spCrossvalid(ols2)
ols3.cv <- spCrossvalid(ols3)
ols4.cv <- spCrossvalid(ols4)
n = summary(geodados)[[1]]
```

```
# Esférico
ERspherical <- sum(ols1.cv$std.erro)/n
Sshperical  <- sqrt((sum(ols1.cv$std.erro^2))/n)
# Exponencial
ERexp <- sum(ols2.cv$std.erro)/n
Sexp  <- sqrt((sum(ols2.cv$std.erro^2))/n)
# Gaussiano
ERgaus <- sum(ols3.cv$std.erro)/n
Sgaus  <- sqrt((sum(ols3.cv$std.erro^2))/n)
# Wave
ERwave <- sum(ols4.cv$std.erro)/n
Swave  <- sqrt((sum(ols4.cv$std.erro^2))/n)
```

Tabela 4.2 – Medidas de ajuste dos modelos comparados

Modelo	$\overline{ER}$	S_{er}
Exponencial	$-1,92 \times 10^{-4}$	1,01
Esférico	$-2,42 \times 10^{-4}$	1,02
Gaussiano	$-1,22 \times 10^{-4}$	1,00
Wave	$-3,24 \times 10^{-5}$	1,00

Fonte: Do autor (2019)

Os dados apresentados na Tabela 4.2 mostram que todos modelos obtiveram bom ajuste e poderiam ser utilizados sem grandes prejuízos para a descrição da variabilidade espacial do fenômeno. Porém, como é necessário selecionar apenas um modelo, a escolha foi feita com base naquele que apresentou o erro médio reduzido mais próximo de 0 e o desvio padrão do erro reduzido mais próximo de 1. Nesse caso, foi selecionado o modelo Wave, cujos parâmetros constam na Tabela 4.3.

Tabela 4.3 – Estimativas iniciais dos parâmetros do modelo

Parâmetro	Estimativa
φ^2	0,04
τ^2	0,28
ϕ^2	11,43

Fonte: Do autor (2019)

Após selecionado o modelo, o passo seguinte consistiu em aplicar o algoritmo sugerido por Pontes e Oliveira (2004) para melhorar as estimativas dos parâmetros e assim construir a matriz de covariância considerando a informação espacial, para então proceder a análise conforme apresentado na seção 2.7.1.2. No pacote spANOVA, isso é realizado a partir da função `aovGeo`, na qual é necessário especificar além do modelo ajustado, o argumento `cutoff` que determinará a porcentagem da máxima distância considerada para o cálculo do semivariograma nas iterações do algoritmo.

```
mod <- aovGeo(ols4, cutoff = 0.50)
```

Considerando o erro de tolerância de 0,001 (valor padrão da função), foram necessárias quatro iterações para que os parâmetros convergissem para as estimativas apresentadas na Tabela 4.4.

Tabela 4.4 – Estimativas atualizadas dos parâmetros do modelo

Parâmetro	Estimativa
φ^2	0,18
τ^2	0,24
ϕ^2	11,73

Fonte: Do autor (2019)

O comando `anova` fornece a tabela de análise de variância, cuja saída é apresentada na Tabela 4.7.

```
anova(mod)
```

Tabela 4.5 – Tabela de análise de variância do experimento considerando a dependência espacial

Fonte de Variação	Grau de Liberdade	Soma de Quadrado	Quadrado Médio	F_0	Pr(>F)	
Tratamentos	12	28,61	2,38	4,97	< 2,22e-16	***
Blocos	3	13,65	4,55	9,50	< 2.22e-16	***
Resíduos	2163	1036,14	0,48			
Total	2178	1078,40				

Fonte: Do autor (2019)

De acordo com a Tabela 4.7, e adotando o nível de significância de 0,05, conclui-se que há evidências para rejeição da hipótese nula (H_0), ou seja, há pelo menos um tratamento que difere dos demais, do mesmo modo, conclui-se que há diferença significativa entre os blocos. Para detectar quais tratamentos diferem entre si, é necessária a utilização de um procedimento de comparações múltiplas. No pacote spANOVA, há três métodos disponíveis, mas, para essa aplicação, optou-se pela utilização do método de agrupamento de Scott-Knott, já que este fornece uma distinção mais clara entre os grupos. Os comandos utilizados para sua aplicação são apresentados abaixo.

```
spScottKnott(mod, sig.level = 0.05)
```

Tabela 4.6 – Método de Agrupamento de Scott-Knott considerando a dependência espacial

Tratamento	Média Original	Média Corrigida	Grupos
10	3,75	3,90	a
2	3,49	3,82	a
13	3,64	3,81	a
9	3,95	3,80	b
8	3,69	3,77	b
7	3,72	3,76	b
6	3,71	3,74	b
11	3,78	3,73	b
4	3,51	3,72	c
5	3,75	3,66	c
3	3,77	3,63	c
12	3,83	3,62	c
1	3,48	3,56	c

Fonte: Do autor (2019)

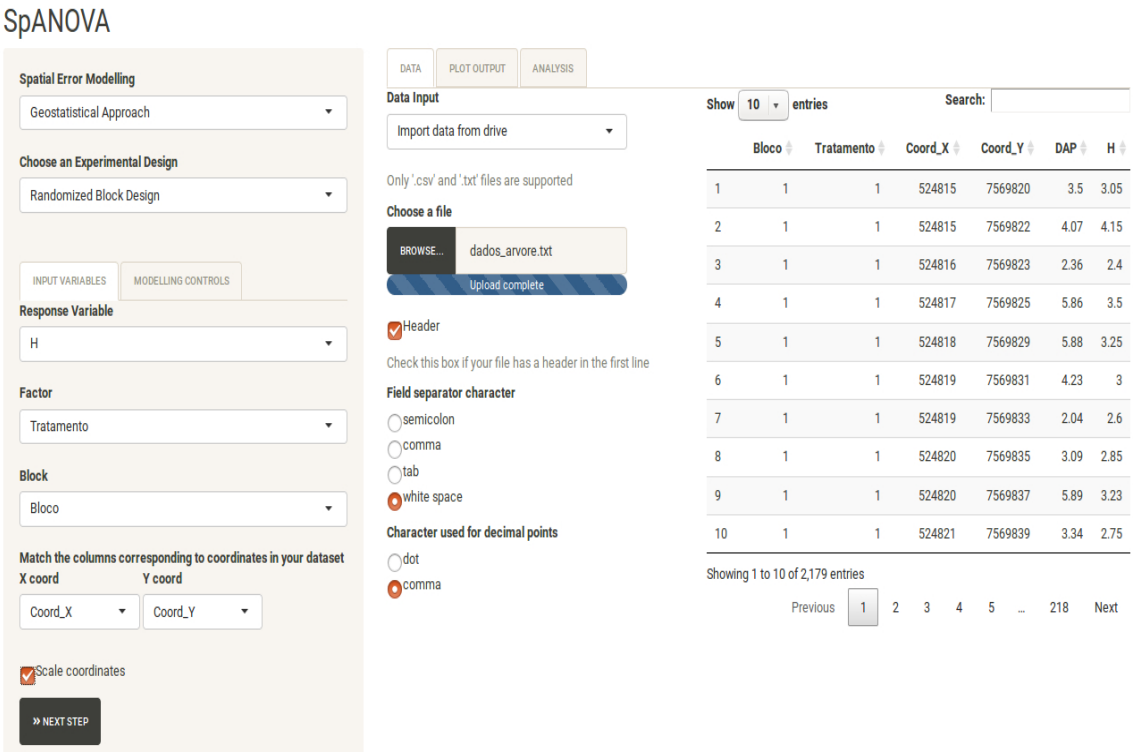
Os dados constantes na Tabela 4.6 correspondem à saída da função `spScottKnott`, em que o ordenamento dos tratamentos é realizado de acordo com as médias corrigidas para o efeito da correlação espacial, conforme apresentado em 2.7.2.1. Levando-se em consideração o nível de significância de 0,05, é possível concluir que os tratamentos 10 (Adubo formulado NPK (8-28-16)), 2 (Ca + NPK + micros), 13 (Ca + adubo formulado NPK (6-30-6) + NK + micros) foram os que apresentaram melhor efeito sobre o desenvolvimento da altura das árvores de candeia, sendo seus efeitos iguais entre si e diferentes dos demais.

Alternativamente, essa análise também foi realizada utilizando a interface gráfica disponível no pacote, para isso bastou acessá-la por meio do comando abaixo.

```
spANOVAapp()
```

Após a execução do comando, o usuário tem acesso à tela inicial da aplicação, na qual é possível carregar os dados, conforme a Figura 4.5.

Figura 4.5 – Carregamento de dados na aplicação



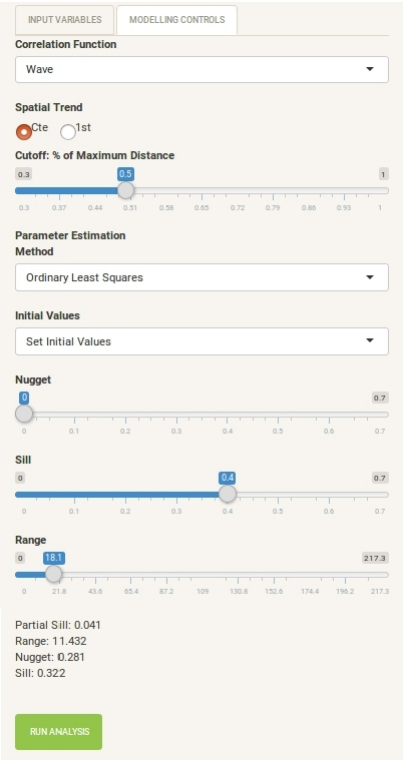
Fonte: Do autor (2019)

Inicialmente foi definido o delineamento experimental a ser utilizado e, depois de feito o *upload* dos dados, certificou-se de que eles foram corretamente reconhecidos por meio da verificação da tabela mostrada no painel direito da aplicação. Em seguida, foram definidos os fatores que compõem o modelo, como mostrado no painel lateral esquerdo da Figura 4.5.

Como as coordenadas espaciais do conjunto de dados em análise foram fornecidas conforme o sistema *Universal Transversa de Mercator* (UTM), que geralmente apresenta valores muito elevados, optou-se por escaloná-las, subtraindo-se de cada uma seu valor mínimo e isso foi feito ao marcar a opção *Scale coordinates*.

Dando prosseguimento à análise, foi necessário acionar o botão *Next* que mostrou uma nova aba chamada *Modelling controls*, na qual são definidos os argumentos mínimos necessários para uma análise, considerando a abordagem geoestatística, como mostrado na Figura 4.6.

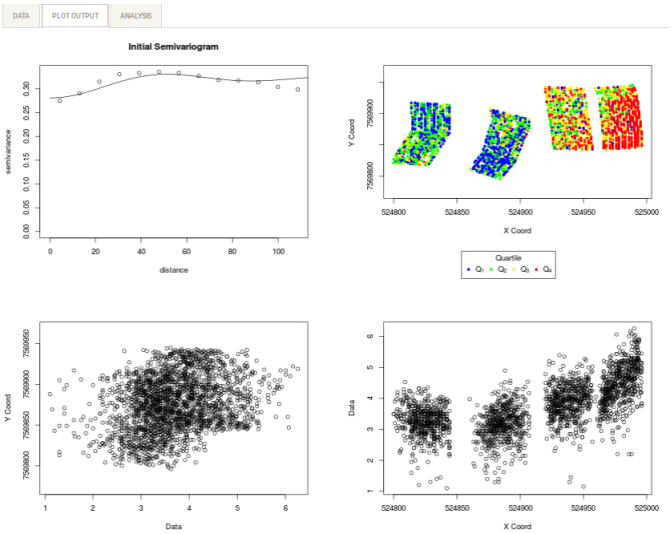
Figura 4.6 – Controles da modelagem geoestatística



Fonte: Do autor (2019)

Após configurados os argumentos necessários, foi fornecido o semivariograma experimental bem como seus respectivos parâmetros estimados e a linha representando o modelo ajustado, além de alguns gráficos descritivos, os quais estão apresentados na Figura 4.7.

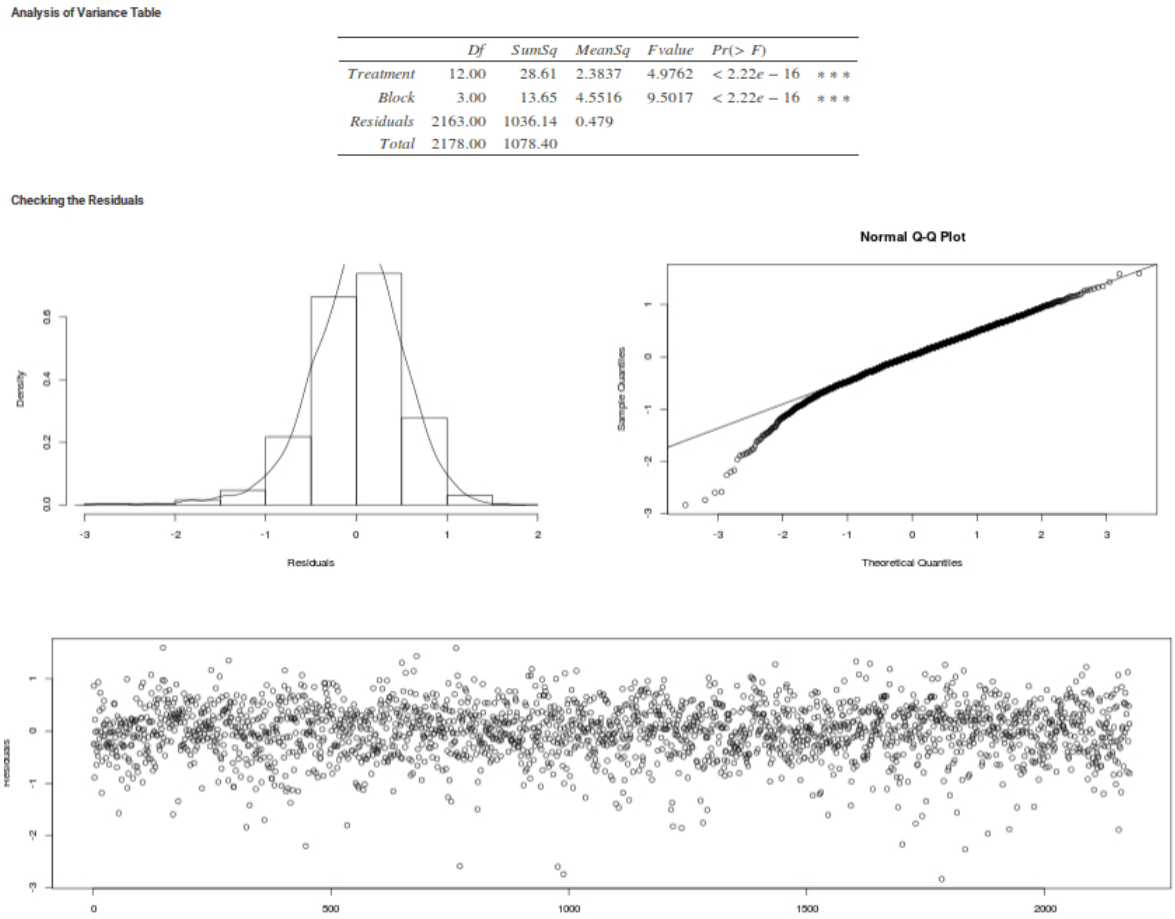
Figura 4.7 – Semivariograma e gráficos descritivos



Fonte: Do autor (2019)

A partir daí obteve-se a tabela de análise de variância ao clicar no botão *Run analysis*, cujos resultados estão apresentados na Figura 4.8.

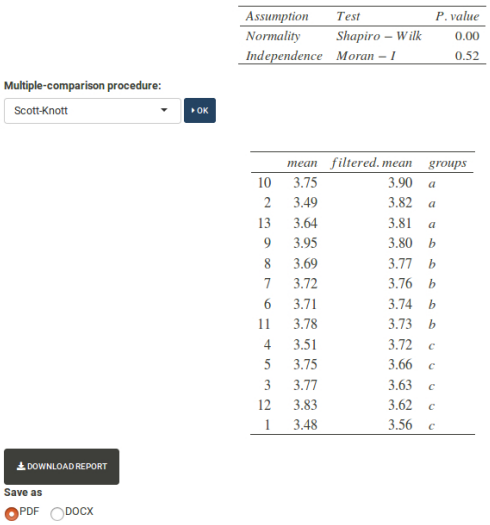
Figura 4.8 – Tabela de análise de variância e gráficos dos resíduos



Fonte: Do autor (2019)

Além da análise de variância, são exibidos alguns gráficos para a análise dos resíduos e os testes de Shapiro-Wilk e Moran ((Figura 4.9) que verificam a condição de normalidade e ausência de dependência espacial, respectivamente. Adotando um nível de significância de 5% e, considerando que o teste de Shapiro-Wilk apresentou valor p abaixo desse nível, entende-se que houve algum desvio de normalidade na distribuição dos resíduos. Porém, Cressie e Whitford (1986) mostraram que os procedimentos empregados na análise de variância são relativamente robustos a desvios de normalidade ocasionado por caudas pesadas, como ocorre nesse caso, o que pode ser observado no gráfico de densidade estimada. Já o valor p do teste de significância para o índice de Moran foi maior que 0,05, não fornecendo evidências para suspeitas de correlação espacial nos resíduos.

Figura 4.9 – Testes de suposições do modelo com a inclusão da informação espacial e procedimento de comparações múltiplas



Fonte: Do autor (2019)

Na Figura 4.9, é possível visualizar a opção para a realização dos procedimentos de comparações múltiplas. Para exemplificar, foi escolhido o método de agrupamento de Scott-Knott. Adicionalmente, é possível gerar um relatório em formato docx ou PDF (disponível apenas na versão *online*) clicando no botão *Download report*, o relatório correspondente a essa análise encontra-se disponível no Apêndice B.

A título de ilustração, foi realizada a modelagem utilizando o modelo usual, o qual não considera a dependência espacial. O resultado da análise de variância é mostrado na Tabela a seguir.

Tabela 4.7 – Tabela de análise de variância do experimento considerando o modelo usual

Fonte de Variação	Grau de Liberdade	Soma de Quadrado	Quadrado Médio	F_0	Pr(>F)	
Tratamentos	12	37,82	3,15	9,79	< 2,22e-16	***
Blocos	3	620,95	206,98	642,90	< 2.22e-16	***
Resíduos	2163	696,39	0,32			
Total	2178	1355,16				

Fonte: Do autor (2019)

Note que a conclusão é a mesma daquela em que se considera o modelo com a informação espacial. No entanto, o teste de significância para o índice de Moran (Tabela 4.8) indica que os resíduos possuem dependência espacial, portanto a utilização desse modelo pode comprometer a validade dos resultados obtidos.

Tabela 4.8 – Teste de pressupostos para o modelo usual

Pressuposto	Teste	Valor p
Normalidade	Shapiro-Wilk	0.00
Independência	Moran-I	0.00

Fonte: Do autor (2019)

4.2.2 Análise através da abordagem espacial autorregressiva

Para esta análise, foi considerado o Delineamento em Blocos Casualizados, cujo modelo pode ser expresso como

$$\boldsymbol{Y}_{adj} = \boldsymbol{X}_1\mu + \boldsymbol{X}_2\boldsymbol{\psi} + \boldsymbol{X}_3\boldsymbol{\eta} + \boldsymbol{\varepsilon}$$

em que

$\boldsymbol{Y}_{adj}$ representa o vetor contendo a altura das árvores ajustado para a autocorrelação espacial, de dimensão 2179×1;

$\boldsymbol{X}_1$ vetor de 1s, de dimensão 2179×1;

μ constante inerente a todas as observações;

$\boldsymbol{X}_2$ matriz de incidência dos tratamentos, de dimensão 2179×13;

$\boldsymbol{\psi}$ vetor de dimensão 13×1, que representa os efeitos dos 13 tipos de adubação analisados;

$\boldsymbol{X}_3$ matriz de incidência dos blocos, de dimensão 2179×4;

$\boldsymbol{\eta}$ representa o vetor de parâmetros dos efeitos dos blocos, de dimensão 4×1;

$\boldsymbol{\varepsilon}$ vetor de dimensão 2179×1, representando o erro aleatório tal que $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$, em que $\boldsymbol{\Sigma} = \boldsymbol{I}\sigma^2$.

Na abordagem utilizando o modelo espacial autorregressivo, é necessário inicialmente construir a matriz de vizinhança. O pacote emprega o método da parcela de referência, conforme mostrado na subseção 2.6.1. Esses procedimentos são todos realizados por meio dos comandos `aovSar.crd` ou `aovSar.rcbd`. Nesse caso específico, como foi utilizado o DBC, os comandos usados estão mostrados abaixo.

```
mod2 <- aovSar.rcbd(candeia$H, candeia$Tratamento, candeia$Bloco,
cbind(candeia$Coord_X,candeia$Coord_Y))
```

Para realizar o ajuste do modelo espacial, é preciso especificar um raio dentro do qual as amostras serão consideradas como vizinhas, isso foi realizado também pelo comando descrito acima e a seleção do raio foi feita de forma a minimizar o valor do AIC. Esse resultado foi obtido aplicando-se o método `summary` no objeto criado, este está apresentado na Tabela 4.9.

```
summary(mod2)
```

Tabela 4.9 – Raios testados e parâmetros espaciais estimados com seus respectivos valores de AIC

Raio	ρ	AIC
11,30	0,78	3461,97
22,6	0,92	3502,51
33,91	0,93	3575,23
45,21	0,94	3595,31
56,51	0,91	3644,81
67,81	0,86	3690,62
79,12	0,72	3715,81
90,41	0,71	3716,13
101,72	0,67	3715,48
113,02	0,55	3722,30

Fonte: Do autor (2019)

Na Tabela 4.9 têm-se os raios testados e seus respectivos valores de AIC, que de acordo com o método de seleção empregado, o valor do raio escolhido foi 11,30. Com isso, a própria função encarregou-se de utilizar a transformação apresentada na equação 2.26, assim, obteve-se a tabela de análise de variância (Tabela 4.10) por meio do método `anova`, conforme o comando a seguir.

```
anova(mod2)
```

Tabela 4.10 – Tabela de análise de variância com efeito de dependência espacial corrigido pelo modelo SAR

Fonte de Variação	Grau de Liberdade	Soma de Quadrado	Quadrado Médio	F_0	Pr(>F)	
Tratmento	12	18,88	1,57	5,65	1,2343e-09	***
Bloco	3	28,51	28,51	34,09	< 2,22e-16	***
Resíduos	2163	602,85	0,28			
Total	2178	704,92				

Fonte: Do autor (2019)

De acordo com a análise de variância, conclui-se que há pelo menos um tratamento que se diferencia dos demais, devido ao baixo valor p apresentado pela fonte de variação "tratamento". Diante disso, foi aplicado o teste de Tukey para indicar em quais níveis do fator essa diferença ocorre. Para tanto, o comando abaixo foi empregado.

```
spTukey(mod2)
```

Tabela 4.11 – Resultado da aplicação do teste de Tukey considerando a abordagem autorregressiva

Tratamento	Média Original	Média Corrigida	Grupo
10	3,95	3,87	a
6	3,78	3,76	ab
2	3,83	3,76	ab
8	3,72	3,76	ab
9	3,69	3,75	ab
13	3,75	3,74	ab
4	3,75	3,71	ab
7	3,71	3,70	abc
11	3,64	3,66	bc
3	3,77	3,66	bc
5	3,51	3,59	bc
12	3,49	3,57	bc
1	3,48	3,52	c

Fonte: Do autor (2019)

Os resultados do teste de Tukey apresentados na Tabela 4.11 apontaram que o melhor tratamento foi o tratamento 10 (Adubo formulado NPK (8-28-16)) e, além disso, ele apresenta efeito comparável aos tratamentos 6, 2, 8, 9, 13, 4 e 7.

Na Tabela 4.12 são apresentados os valores p dos testes de Shapiro-Wilk e a significância do índice de Moran para o modelo. Nota-se que, de modo semelhante ao resultado obtido na abordagem geoestatística, o valor p do teste de Shapiro-Wilk também foi abaixo

do nível de significância de 5%, nesse caso também pode-se utilizar da justificativa anteriormente citada que ressalta a robustez da análise a desvios de normalidade. Já o valor p do teste de significância do índice I de Moran foi superior a 5%, indicando que os resíduos não são espacialmente correlacionados.

Tabela 4.12 – Teste de pressupostos para o modelo autorregressivo

Pressuposto	Teste	Valor p
Normalidade	Shapiro-Wilk	0.00
Independência	Moran-I	0.09

Fonte: Do autor (2019)

A análise utilizando o ambiente gráfico foi omitida, porém é realizada de modo análogo ao exemplo apresentado para a abordagem geoestatística.

5 CONCLUSÕES

A falta de *softwares* que possibilitassem a realização da análise de variância utilizando a abordagem espacial de forma prática levou ao desenvolvimento deste trabalho, que permitiu gerar uma biblioteca em ambiente R que incorpora a informação espacial na análise de variância sob a perspectiva da geoestatística e do modelo espacial autorregressivo (SAR).

A biblioteca desenvolvida mostrou-se funcional uma vez que, reuniu de forma simplificada todos os passos necessários para a inclusão da informação espacial na análise de variância, bem como forneceu uma interface em *shiny* que pode ser utilizada mesmo por usuários que não possuem conhecimentos de linguagens de programação.

Por meio dessa biblioteca, fazendo uso de uma base de dados de plantio de candeia, foi possível comparar uma análise incorporando a dependência espacial e sem incorporar essa componente (modelo de análise de variância usual). Constatou-se nesse exemplo que, apesar de ambas as análises apontarem diferenças significativas entre os tratamentos, a abordagem espacial conseguiu contornar o efeito da correlação gerada pelo espaço, uma vez que os resíduos do modelo final não apresentaram dependência espacial, por outro lado, a verificação dos resíduos provenientes da análise usual apontou que estes eram espacialmente correlacionados, infringindo, dessa maneira, o pressuposto de independência, que é indispensável para a validade dos resultados obtidos.

Com isso, conclui-se que o produto final alcançado por meio da realização deste trabalho pode ser de grande utilidade, sobretudo para pesquisadores de áreas mais aplicadas que, a partir de agora, poderão contar com uma ferramenta construída sob licença de código aberto e totalmente livre de custos para auxiliar na obtenção de resultados mais confiáveis na análise de experimentos com dependência espacial.

REFERÊNCIAS

AKAIKE, H. Information measures and model selection. **Bulletin of the International Statistical Institute**, Rome, v. 44, p. 277–291, 1983.

ARMSTRONG, M. **Basic linear geostatistics**. New York: Springer Science & Business Media, 1998. 154 p.

BANZATTO, D.; KRONKA, S. **Experimentação agrícola**. Jaboticabal: FUNEP, 2006. 154 p.

BARTLETT, M. Nearest neighbour models in the analysis of field experiments. **Journal of the Royal Statistical Society: Series B (Methodological)**, London, v. 40, n. 2, p. 147–158, 1978.

BATES, D.; MAECHLER, M. **Matrix: sparse and dense matrix classes and methods**. [S.l.], 2018. Disponível em: <https://CRAN.R-project.org/package=Matrix>. Acesso em: 15 mar. 2018.

BESAG, J.; KEMPTON, R. Statistical analysis of field experiments using neighbouring plots. **Biometrics**, JSTOR, Alexandria, v. 42, p. 231–251, 1986.

BIVAND, R. S.; PEBESMA, E.; GOMEZ-RUBIO, V. **Applied spatial data analysis with R**. 2. ed. New York: Springer, 2013. 405 p.

BRETZ, F.; WESTFALL, P.; HOTHORN, T. **Multiple comparisons using R**. Boca Raton: Chapman and Hall/CRC, 2016. 202 p.

BURROUGH, P. **Principles of geographical information systems for land resources assessment**. New York: Clarendon Press, 1986. 216 p.

CAETANO, E. R. R. **Análise de variância utilizando modelos autoregressivos em experimentos com dependência espacial**. 116 f. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) — Universidade Federal de Lavras, Lavras, 2013.

CÂMARA, G.; CARVALHO, M. S. Análise espacial de eventos. In: **Análise espacial de dados geográficos**. Brasília: Embrapa, 2004. p. 55–75.

CHANG, W. et al. **Shiny: web application framework for R**. [S.l.], 2017. Disponível em: <https://CRAN.R-project.org/package=shiny>. Acesso em: 15 mar. 2018.

CLARK, I. **Practical geostatistics**. London: Applied Science Pubication, 1979. 129 p.

CRESSIE, N. **Statistics for spatial data**. New York: John Wiley & Sons, 1993. 931 p.

CRESSIE, N.; HARTFIELD, M. N. Conditionally specified gaussian models for spatial statistical analysis of field trials. **Journal of Agricultural, Biological, and Environmental Statistics**, JSTOR, Alexandria, p. 60–77, 1996.

CRESSIE, N.; WHITFORD, H. How to use the two sample t-test. **Biometrical Journal**, v. 28, n. 2, p. 131–148, 1986.

CULLIS, B.; GLEESON, A. Spatial analysis of field experiments-an extension to two dimensions. **Biometrics**, Alexandria, p. 1449–1460, 1991.

DIGGLE, P. J.; RIBEIRO JÚNIOR, P. J. **Model-based geostatistics**. New York: Springer, 2007. 232 p.

DUARTE, J. B. **Sobre o emprego da análise estatística do delineamento em blocos aumentados no melhoramento genético**. 293 f. Tese (Doutorado Agronomia) — Escola Superior de Agricultura "Luiz de Queiroz", Piracicaba, 2000.

FERREIRA, D. F. **Estatística multivariada**. Lavras: Editora UFLA, 2008. 675 p.

FISHER, R. A. **Statistical methods for research workers**. London: Genesis Publishing, 1925. 374 p.

FISHER, R. A. The arrangement of field experiments. **Journal of the Ministry of Agriculture of Great Britain**, [S.l.], v. 33, p. 503–513, 1926.

GLEESON, A. C.; CULLIS, B. R. Residual maximum likelihood (reml) estimation of a neighbour model for field experiments. **Biometrics**, Alexandria, p. 277–287, 1987.

GOTWAY, C. A.; CRESSIE, N. A. A spatial analysis of variance applied to soil-water infiltration. **Water Resources Research**, Hoboken, v. 26, n. 11, p. 2695–2703, 1990.

GRAVES, S.; PIEPHO, H.-P.; DORAI-RAJ, L. S. with help from S. **multcompView: visualizations of paired comparisons**. [S.l.], 2015. Disponível em: <https://CRAN.R-project.org/package=multcompView>. Acesso em: 15 mar. 2018.

GUMPERTZ, M. L.; GRAHAM, J. M.; RISTAINO, J. B. Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: effects of soil variables on disease presence. **Journal of Agricultural, Biological, and Environmental Statistics**, Alexandria, p. 131–156, 1997.

HINKELMANN, K.; KEMPTHORNE, O. **Design and analysis of experiments**. 2nd. ed. New Jersey: John Wiley & Sons, 2008. v. 1. 811 p.

HOTHORN, T.; BRETZ, F.; WESTFALL, P. Simultaneous inference in general parametric models. **Biometrical journal**, New York, v. 50, n. 3, p. 346–363, 2008.

HOTHORN, T. et al. Package 'multcomp'. 2016. Disponível em: <https://CRAN.R-project.org/package=multcomp>. Acesso em: 15 mar. 2018.

KUTNER, M. H. et al. **Applied linear statistical models**. New York: McGraw-hill/Irwin, 2005. v. 5. 1396 p.

LONG, D. S. Spatial statistics for analysis of variance of agronomic field trials. In: ARLINGHAUS, S. e. a. (Ed.). **Practical Handbook of Spatial Statistics**. Boca Raton: CRC Press, 1996. cap. 10, p. 251–278.

LONGHI, P. R. et al. Estudo de caso do processo de extração do óleo essencial da madeira de candeia no Sul de Minas Gerais. **Floresta**, Curitiba, v. 39, n. 3, 2009.

MELLO, J. d. et al. Ajuste e seleção de modelos espaciais de semivariograma visando à estimativa volumétrica de eucalyptus grandis. **Scientia Forestalis**, Piracicaba, v. 69, n. 1, p. 25–37, 2005.

MONTGOMERY, D. C. **Design and analysis of experiments**. 8. ed. New York: John Wiley & Sons, 2012. 757 p.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the theory of statistics.** New York: McGraw-Hill, 1974. 480 p.

NOGUEIRA, C. H. **Análise de variância com dependência Espacial sob uma abordagem geoestatística**. 124 f. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) — Universidade Federal de Lavras, Lavras, 2013.

NOGUEIRA, C. H. **Testes para comparações múltiplas de médias em experimentos com tendência e dependência espacial**. 142 f. Tese (Doutorado em Estatística e Experimentação Agropecuária) — Universidade Federal de Lavras, Lavras, 2017.

NOGUEIRA, C. H. et al. Modelagem espacial na análise de um plantio experimental de candeia. **Revista Brasileira de Biometria**, São Paulo, v. 33, n. 1, p. 14–29, 2015.

ORD, K. Estimation methods for models of spatial interaction. **Journal of the American Statistical Association**, v. 70, n. 349, p. 120–126, 1975.

PAPADAKIS, J. Méthode statistique pour des expériences sur champ. **Institut d'Amélioration des Plantes à Salonique**, p. 30, 1937. (Bullettin, 23).

PIMENTEL, G. F. **Curso de estatística experimental**. 1. ed. Piracicaba: Fealq, 1990. 468 p.

PONTES, J. M.; OLIVEIRA, M. S. de. Uma proposta alternativa para a análise de experimentos de campo utilizando a geoestatística an alternative proposal to the analysis of field experiments using geostatistics. **Ciência e Agrotecnologia**, Lavras, v. 28, n. 1, p. 135–141, 2004.

QUINN, G. P.; KEOUGH, M. J. et al. **Experimental design and data analysis for biologists**. Cambridge: Cambridge University Press, 2002. 553 p.

R Core Team. **R: A language and environment for statistical computing**. Vienna, 2018. Disponível em: <https://www.R-project.org/>. Acesso em: 23 jan. 2019.

RAMALHO, M.; FERREIRA, D.; OLIVEIRA, A. de. **A experimentação em genética e melhoramento de plantas**. Lavras: Ed. UFLA, 2005. 300 p.

RAO, C. R. **Linear statistical inference and its applications**. New York: John Wiley & Sons, 1973. v. 2. 643 p.

RENCHER, A. C.; SCHAALJE, G. B. **Linear models in statistics**. New York: John Wiley & Sons, 2008. 688 p.

RIBEIRO JÚNIOR, P. J.; DIGGLE, P. J. **geoR: Analysis of Geostatistical Data**. [S.l.], 2016. R package version 1.7-5.2. Disponível em: <https://CRAN.R-project.org/package=geoR>. Acesso em: 23 jan. 2018.

RIVOIRARD, J. Concepts and methods of geostatistics. In: BILODEAU, M.; MEYER, F.; SCHMITT, M. (Ed.). **Space, Structure and Randomness**. Paris: Springer, 2005. p. 17–37.

ROSS, S. M. **Introduction to probability models**. Los Angeles: Academic Press, 2014. 800 p.

ROSSONI, D. F. **Análise de variância para experimentos com depedência espacial**. 109 f. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) — Universidade Federal de Lavras, Lavras, 2011.

SCHABENBERGER, O.; GOTWAY, C. A. **Statistical methods for spatial data analysis**. Boca Raton: Chapman and Hall/CRC, 2017. 512 p.

SEARLE, S. **Linear models for unbalanced data**. New York: John Wiley & Sons, 1987. 560 p.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, Oxford, v. 52, n. 3/4, p. 591–611, 1965.

SOARES, A. **Geoestatística para as ciências da terra e do ambiente**. Lisboa: IST Press, 2006. 232 p.

STRINGER, J. K. et al. Spatial analysis of agricultural field experiments. In: HINKELMANN, K. (Ed.). **Design and analysis of experiments: special designs and applications**. New York: Wiley-Blackwell, 2012. v. 3, cap. 3, p. 109–136.

VENABLES, W. N.; RIPLEY, B. D. **Modern applied Statistics with S**. 4. ed. New York: Springer, 2013. 498 p.

VIEIRA, S. et al. Geostatistical theory and application to variability of some agronomical properties. **Hilgardia**, Berkeley, v. 51, n. 3, p. 1–75, 1983.

YAMAMOTO, J. K.; LANDIM, P. M. B. **Geoestatística: conceitos e aplicações**. São Paulo: Oficina de textos, 2015. 213 p.

YWATA, C.; ALBUQUERQUE, P. Métodos e modelos em econometria espacial: uma revisão. **Revista Brasileira de Biometria**, v. 29, n. 2, p. 273–306, 2011.

ZIMMERMAN, D. L.; HARVILLE, D. A. A random field approach to the analysis of field-plot experiments and other spatial experiments. **Biometrics**, Alexandria, v. 47, n. 1, p. 223–239, 1991.

APÊNDICE - A: Documentação do Pacote spANOVA

Package 'spANOVA'

July 2, 2019

Type Package

Title Spatial Analysis of Field Trials Experiments using Geostatistics and Spatial Autoregressive Model

Version 0.99.0

Author Lucas Roberto de Castro, Renato Ribeiro de Lima, Diogo Francisco Rossoni, Cristina Henriques Nogueira

Maintainer Lucas Roberto de Castro <lrcastro@estudante.ufla.br>

Description Perform analysis of variance when the experimental units are spatially correlated. There are two methods to deal with spatial dependence: Spatial autoregressive models (see Scolforo, H.F et al. (2016) <doi:10.1007/s11676-015-0185-y>) and geostatistics (see Pontes, J. M., & Oliveira, M. S. D. (2004) <doi:10.1590/S1413-70542004000100018>). For both methods, there are three multicomparison procedure available: Tukey, multivariate T, and Scott-Knott.

License GPL-3

Encoding UTF-8

LazyData true

Depends R (>= 2.10), stats, utils, graphics, geoR, shiny

Imports MASS, Matrix, ScottKnott, car, gtools, multcomp, multcompView, mvtnorm, devtools, DT, shinyBS, xtable, shinythemes, rmarkdown, knitr, spdep, ape, spatialreg

RoxygenNote 6.1.1

NeedsCompilation no

Repository CRAN

Date/Publication 2019-07-02 14:50:03 UTC

R topics documented:

`aovGeo` *Analysis of variance using a geostatistical model to handle spatial dependence*

Description

Fit an analysis of variance model using geostatistics for modeling the spatial dependence among the observations.

Usage

```
aovGeo(model, cutoff = 0.5, tol = 1e-3)

## S3 method for class 'spVariofitRCBD'
aovGeo(model, cutoff = 0.5, tol = 0.001)
```

Arguments

`model`	an object of class spVariofit.
`cutoff`	a value in (0,1] interval representing the percentage of the maximum distance adopted to estimate the variogram in the algorithm suggested by Pontes & Oliveira (2004). See 'Details'.
`tol`	the desired accuracy.

Details

Three assumptions are made about the error in the analysis of variance (ANOVA):

1. The errors come from a normal distribution.

2. The errors have the same variance.

3. The errors are uncorrelated.

However, in many experiments, especially field trials, there is a type of correlation generated by the sample locations known as spatial autocorrelation, and this condition violates the independence assumption.

In that way, we need to regard this spatial autocorrelation and include it in the final model. It could be done adopting a geostatistical model to characterize the spatial variability among the observations directly in the covariance matrix.

The geostatistical modeling is based on the residuals of the standard model (where the errors are assumed to be independent, uncorrelated and having a normal distribution with mean 0 and constant variance). The basic idea is using them to estimate the residuals of the spatially autocorrelated model in order to fit a theoretical geostatistic model to build the covariance matrix. As Pointed by Pontes & Oliveira (2004), this task can be done using the following algorithm

1 - Extract the residuals from the standard model

2 - Fit a variogram based on residuals obtained in step 1.

3 - Fit a theoretical model to describe the spatial dependence observed in the variogram.

4 - On basis in the theoretical model fitted in step 3 and its parameter estimates, create the covariance matrix.

5 - Estimate the residuals using the covariance matrix obtained in step 4 and use them to create a variogram.

6 - Fit a theoretical model to the residual variogram obtained in step 5 and use its parameters estimates to build a new covariance matrix.

7 - Repeat 5 to 6 until convergence.

Step 1 is implemented in spVariog. Steps 2 and 3 are implemented in spVariofit. aovGeo implements steps 4 to 7 and needs a cutoff argument to define the maximum distance in the computation of the residual variogram described in step 6

In presence of spatial trend, the model is modified in order to include the effect of the spatial coordinates.

Value

aovGeo returns an object of class "GEOanova". The functions summary and anova are used to obtain and print a summary and analysis of variance table of the results. An object of class "GEOanova" is a list containing the following components:

DF	degrees of freedom of coordinates (in presence of spatial trend), treatments, block (if RCBD), residual and total.
SS	sum of squares of coordinates (in presence of spatial trend), treatments, block (if RCBD), residual and total.
MS	mean of squares of coordinates (in presence of spatial trend), treatments, block (if RCBD), residual and total.
Fc	F statistic calculated for coordinates (in presence of spatial trend), treatment and block (if RCBD).
p.value	p-value associated to F statistic for coordinates (in presence of spatial trend), treatment and block (if RCBD).
residuals	residuals extracted from the model
params	variogram parameter estimates from Pontes & Oliveira (2004) algorithm.
type	type of trend in the data. It can be "trend" or "cte".

model	geostatistical model used to describe the spatial dependence.
data	data set used in the analysis.
des.mat	design matrix.
beta	parameter estimates of the linear model taking into account the spatial dependence.
n	number of observations.
vcov	covariance matrix built taking into account the spatial dependence.
design	character string indicating the name of the experimental design.

References

NOGUEIRA, C. H., de LIMA, R. R., & de OLIVEIRA, M. S. (2013). Aprimoramento da Análise de Variância: A Influência da Proximidade Espacial. Rev. Bras. Biom, 31(3), 408-422.

NOGUEIRA, C.H., et al. (2015). Modelagem espacial na análise de um plantio experimental de candeia. Rev. Bras. Biom., São Paulo, v.33, n.1, p.14-29.

Pontes, J. M., & de Oliveira, M. S. (2004). An alternative proposal to the analysis of field experiments using geostatistics. Ciência e Agrotecnologia, 28(1), 135-141.

Gotway, C. A., & Cressie, N. A. (1990). A spatial analysis of variance applied to soil water infiltration. Water Resources Research, 26(11), 2695-703.

Examples

```
data("crd_simulated")

#Geodata object
geodados <- as.geodata(crd_simulated, coords.col = 1:2, data.col = 3,
                           covar.col = 4)
h_max <- summary(geodados)[[3]][[2]]
dist <- 0.6*h_max

# Computing the variogram
variograma <- spVariog(geodata = geodados,
                     trend = "cte", max.dist = dist, design = "crd",
                     scale = FALSE)

plot(variograma, ylab = "Semivariance", xlab = "Distance")

# Gaussian Model
ols <- spVariofit(variograma, cov.model = "gaussian", weights = "equal",
                  max.dist = dist)

lines(ols, col = 1)

# Compute the model and get the analysis of variance table
mod <- aovGeo(ols, cutoff = 0.6)
anova(mod)
```

aovSar.crd *Using a SAR model to handle spatial dependence in a Completely Randomized Design*

Description

Fit a completely randomized design when the experimental units have some degree of spatial dependence using a Spatial Lag Model (SAR).

Usage

```
aovSar.crd(resp, treat, coord, seq.radius)
```

Arguments

resp	Numeric or complex vector containing the values of the response variable.
treat	Numeric or complex vector containing the treatment applied to each experimental unit.
coord	Matrix of point coordinates.
seq.radius	Complex vector containing a radii sequence used to set the neighborhood pattern. The default sequence has ten numbers from 0 to half of the maximum distance between the samples, if no sample is found in this interval the sequence upper limit is increased by 10% and so on.

Details

Three assumptions are made about the error in the analysis of variance (ANOVA):

1. the errors are normally distributed and, on average, zero;

2. the errors all have the same variance (they are homoscedastic), and

3. the errors are unrelated to each other (they are independent across observations).

When these assumptions are not satisfied, data transformations in the response variable are often used to circumvent this problem. For example, in absence of normality, the Box-Cox transformation can be used.

However, in many experiments, especially field trials, there is a type of correlation generated by the sample locations known as spatial correlation, and this condition violates the independence assumption. errors are spatially correlated, by using a data transformation discussed in Long (1996)

$$Y_{adj} = Y - (\hat{\rho}WY - \hat{\rho}\beta_0),$$

where $\hat{\rho}$ denotes the autoregressive spatial parameter of the SAR model estimated by lagsarlm, β_0 is the overall mean and W is a spatial neighborhood matrix which neighbors are defined as the samples located within a radius, this radius is specified as a sequence in `seq.radius`. For each radius in `seq.radius` the model is computed as well its AIC, then the radius chosen is the one that minimizes AIC.

The aim of this transformation is converting autocorrelated observations into non-correlated observations in order to apply the analysis of variance and obtain suitable inferences.

Value

aovSar.crd returns an object of class "SARanova". The functions summary and anova are used to obtain and print a summary and analysis of variance table of the results. An object of class "SARanova" is a list containing the following components:

DF	degrees of freedom of rho, treatments, residual and total.
SS	sum of squares of rho, treatments and residual.
MS	mean square of rho, treatments and residual.
Fc	F statistic calculated for treatment.
residuals	residuals of the adjusted model.
p.value	p-value associated to F statistic for treatment.
rho	the autoregressive parameter.
Par	data.frame with the radius tested and its AIC.
y_orig	vector of response.
y_ajus	vector of adjusted response.
treat	vector of treatment applied to each experimental unit.
modelAdj	model of class aov using the adjusted response.
modelstd	data frame containing the ANOVA table using non-adjusted response.
namey	response variable name.
namex	treatment variable name.

References

Long, D.S., 1996. Spatial statistics for analysis of variance of agronomic field trials. In: Arlinghaus, S.L. (Ed.), Practical Handbook of Spatial Statistics. CRC Press, Boca Raton, FL, pp. 251–278.

ROSSONI, D. F.; LIMA, R. R. . Autoregressive analysis of variance for experiments with spatial dependence between plots: a simulation study. REVISTA BRASILEIRA DE BIOMETRIA, 2019.

Long, D. S. (1998). Spatial autoregression modeling of site-specific wheat yield. Geoderma, 85(2-3), 181-197.

Examples

```
data("crd_simulated")
resp <- crd_simulated$y
treat <- crd_simulated$trat
coord <- cbind(crd_simulated$coordX, crd_simulated$coordY)
cv <- aovSar.crd(resp, treat, coord)

#Summary for class SARanova
summary(cv)

#Anova for class SARanova
anova(cv)
```

aovSar.gen *Using a SAR model to handle spatial dependence in an aov model.*

Description

Fit a completely randomized design when the experimental units have some degree of spatial dependence using a Spatial Lag Model (SAR).

Usage

```
aovSar.gen(formula, coord, seq.radius, data = NULL)
```

Arguments

formula	A formula specifying the model.
coord	A matrix or data.frame of point coordinates.
seq.radius	A complex vector containing a radii sequence used to set the neighborhood pattern. The default sequence has ten numbers from 0 to half of the maximum distance between the samples.
data	A data frame in which the variables specified in the formula will be found.

Details

Three assumptions are made about the error in the analysis of variance (ANOVA):

1. the errors are normally distributed and, on average, zero;

2. the errors all have the same variance (they are homoscedastic), and

3. the errors are unrelated to each other (they are independent across observations).

When these assumptions are not satisfied, data transformations in the response variable are often used to circumvent this problem. For example, in absence of normality, the Box-Cox transformation can be used.

However, in many experiments, especially field trials, there is a type of correlation generated by the sample locations known as spatial correlation, and this condition violates the independence assumption. In this setting, this function provides an alternative for using ANOVA when the errors are spatially correlated, by using a data transformation discussed in Long (1996)

$$Y_{adj} = Y - (\hat{\rho}WY - \hat{\rho}\beta_0),$$

where $\hat{\rho}$ denotes the autoregressive spatial parameter of the SAR model estimated by lagsarlm, β_0 is the overall mean and W is a spatial neighborhood matrix which neighbors are defined as the samples located within a radius, this radius is specified as a sequence in `seq.radius`. For each radius in `seq.radius` the model is computed as well its AIC, then the radius chosen is the one that minimizes AIC.

The aim of this transformation is converting autocorrelated observations into non-correlated observations in order to apply the analysis of variance and obtain suitable inferences.

Value

aovSar.gen returns an object of class "SARaov". The functions summary and anova are used to obtain and print a summary and analysis of variance table of the results. An object of class "SARaov" is a list containing the following components:

DF	degrees of freedom of rho, treatments, residual and total.
SS	sum of squares of residuals and total.
residuals	residuals of the adjusted model.
MS	mean square of residuals and total.
rho	the autoregressive parameter.
Par	data.frame with the radius tested and its AIC.
modelAdj	model of class aov using the adjusted response.
modelstd	data frame containing the ANOVA table using non-adjusted response.
namey	response variable name.

References

Long, D. S. "Spatial statistics for analysis of variance of agronomic field trials." Practical handbook of spatial statistics. CRC Press, Boca Raton, FL (1996): 251-278.

ROSSONI, D. F.; LIMA, R. R. . Autoregressive analysis of variance for experiments with spatial dependence between plots: a simulation study. REVISTA BRASILEIRA DE BIOMETRIA, 2019

Scolforo, Henrique Ferraço, et al. "Autoregressive spatial analysis and individual tree modeling as strategies for the management of Eremanthus erythropappus." Journal of forestry research 27.3 (2016): 595-603.

Examples

```
data("crd_simulated")
coord <- cbind(crd_simulated$coordX, crd_simulated$coordY)
cv <- aovSar.gen(y ~ trat, coord, data = crd_simulated)
cv

#Summary for class SARanova
summary(cv)

#Anova for class SARanova
anova(cv)
```

aovSar.rcbd *Using a SAR model to handle spatial dependence in a Randomized Complete Block Design*

Description

Fit a randomized complete block design when the experimental units have some degree of spatial dependence using a Spatial Lag Model (SAR).

Usage

```
aovSar.rcbd(resp, treat, block, coord, seq.radius)
```

Arguments

resp	Numeric or complex vector containing the values of response variable.
treat	Numeric or complex vector containing the treatment applied to each experimental unit.
block	Numeric or complex vector specifying the blocks.
coord	Matrix of point coordinates or a SpatialPoints Object.
seq.radius	Complex vector containing a radii sequence used to set the neighborhood pattern. The default sequence has ten numbers from 0 to half of the maximum distance between the samples.

Details

Three assumptions are made about the error in the analysis of variance (ANOVA):

1. the errors are normally distributed and, on average, zero;

2. the errors all have the same variance (they are homoscedastic), and

3. the errors are unrelated to each other (they are independent across observations).

When these assumptions are not satisfied, data transformations in the response variable are often used to circumvent this problem. For example, in absence of normality, the Box-Cox transformation can be used.

However, in many experiments, especially field trials, there is a type of correlation generated by the sample locations known as spatial correlation, and this condition violates the independence assumption. In this setting, this function provides an alternative for using ANOVA when the errors are spatially correlated, by using a data transformation discussed in Long (1996)

$$Y_{adj} = Y - (\hat{\rho}WY - \hat{\rho}\beta_0),$$

where $\hat{\rho}$ denotes the autoregressive spatial parameter of the SAR model estimated by lagsarlm, β_0 is the overall mean and W is a spatial neighborhood matrix which neighbors are defined as the samples located within a radius, this radius is specified as a sequence in `seq.radius`. For each

radius in seq.radius the model is computed as well its AIC, then the radius chosen is the one that minimizes AIC.

The aim of this transformation is converting autocorrelated observations into non-correlated observations in order to apply the analysis of variance and obtain suitable inferences.

Value

aovSar.rcbd returns an object of class "SARanova". The functions summary and anova are used to obtain and print a summary and analysis of variance table of the results. An object of class "SARanova" is a list containing the following components:

DF	degrees of freedom of rho, treatments, residual and total.
SS	sum of squares of rho, treatments and residual.
Fc	F statistic calculated for treatment.
residuals	residuals of the adjusted model.
p.value	p-value associated to F statistic for treatment.
rho	the autoregressive parameter.
Par	data.frame with the radii tested and its AIC.
y_orig	vector of response.
y_ajus	vector of adjusted response.
treat	vector of treatment applied to each experimental unit.
modelAdj	model of class aov using the adjusted response.
modelstd	data frame containing the ANOVA table using non-adjusted response.
namey	response variable name.
namex	treatment variable name.

References

Long, D.S., 1996. Spatial statistics for analysis of variance of agronomic field trials. In: Arlinghaus, S.L. (Ed.), Practical Handbook of Spatial Statistics. CRC Press, Boca Raton, FL, pp. 251–278.

ROSSONI, D. F.; LIMA, R. R. . Autoregressive analysis of variance for experiments with spatial dependence between plots: a simulation study. REVISTA BRASILEIRA DE BIOMETRIA, 2019

Scolforo, Henrique Ferraço, et al. "Autoregressive spatial analysis and individual tree modeling as strategies for the management of Eremanthus erythropappus." Journal of forestry research 27.3 (2016): 595-603.

Examples

```
data("rcbd_simulated")

# Fitting the model
model <- aovSar.rcbd(rcbd_simulated$y, rcbd_simulated$trat, rcbd_simulated$block,
                     cbind(rcbd_simulated$coordX, rcbd_simulated$coordY))

# Summary for class SARanova
```

```
summary(model)

# Anova for class SARanova
anova(model)
```

`contr.tuk` *Tukey's contrast matrix*

Description

Compute Tukey's contrast matrix

Usage

```
contr.tuk(x)
```

Arguments

`x`	a vector of means.

Details

Computes the matrix of contrasts for comparisons of mean levels.

Value

The matrix of contrasts with treatments in row names is returned

Examples

```
x <- c(10, 5, 8, 4, 12, 18)
contr.tuk(x)
```

`crd_simulated` *Simulated data set for CRD*

Description

Simulated data set for CRD

Details

This dataset was simulated under a gaussian random field containing 15 treatments and 8 replication with mean 10.

Author(s)

Lucas Castro <lrcastro@estudante.ufla.br>

rcbd_simulated *Simulated data set for RCBD*

Description

Simulated data set for RCBD

Details

This dataset was simulated under a gaussian random field containing 15 treatments and 3 blocks with 3 replication.

Author(s)

Lucas Castro <lrcastro@estudante.ufla.br>

spANOVAapp *Shiny app for spANOVA*

Description

Shiny app for analysis of variance with spatially correlated errors

Usage

```
spANOVAapp(external = TRUE)
```

Arguments

external	logical. If true, the system's default web browser will be launched automatically after the app is started.

Examples

```
spANOVAapp(external = TRUE)
```

`spCrossvalid` *Cross-validation by kriging*

Description

Compute cross-validation for an object of class `spVariofit`.

Usage

```
spCrossvalid(x, ...)
```

Arguments

`x`	an object of class `spVariofit`.
`...`	further arguments to be passed to `xvalid` function.

Details

This function is a wrapper to `xvalid` function of the package geoR. Please check its documentation for additional information.

Examples

```
data("crd_simulated")
dados <- crd_simulated

#Geodata object
geodados <- as.geodata(dados, coords.col = 1:2, data.col = 3,
                            covar.col = 4)
h_max <- summary(geodados)[[3]][[2]]
dist <- 0.6*h_max

# Computing the variogram
variograma <- spVariog(geodata = geodados,
                      trend = "cte", max.dist = dist, design = "crd",
                      scale = FALSE)

plot(variograma, ylab = "Semivariance", xlab = "Distance")

# Spherical Model
ols1 <- spVariofit(variograma, cov.model = "spherical", weights = "equal",
                   max.dist = dist)

#Using crossvalidation to assess the error
ols1.cv <- spCrossvalid(ols1)
```

`spMVT` *Multiple comparison test based on multivariate t student distribution*

Description

Use a multivariate t student distribution to assess the equality of means.

Usage

```
spMVT(x, sig.level = 0.05)

## S3 method for class 'SARanova'
spMVT(x, sig.level = 0.05)

## S3 method for class 'GEOanova'
spMVT(x, sig.level = 0.05)
```

Arguments

`x`	a fitted model object of class SARcrd, SARrcbd or GEOanova.
`sig.level`	a numeric value between zero and one giving the significance level to use.

Details

For objects of class SARcrd or SARrcbd this function performs the general linear hypothesis method provided by the function `glht` on the adjusted response.

For objects of class GEOanova, the test is modified to accommodate the spatial dependence among the observations as pointed out by Nogueira (2017)

Value

a data frame containing the original mean, the spatially filtered mean and its group. For the class GEOanova, the spatial dependence is filtered out using geostatistics, while for the class SARanova the adjusted response based on SAR model is employed.

References

NOGUEIRA, C. H. Testes para comparações múltiplas de médias em experimentos com tendência e dependência espacial. 142 f. Tese (Doutorado em Estatística e Experimentação Agropecuária) | Universidade Federal de Lavras, Lavras, 2017

Examples

```
data("crd_simulated")
```

```
#Geodata object
geodados <- as.geodata(crd_simulated, coords.col = 1:2, data.col = 3,
                       covar.col = 4)
h_max <- summary(geodados)[[3]][[2]]
dist <- 0.6*h_max

# Computing the variogram
variograma <- spVariog(geodata = geodados,
                       trend = "cte", max.dist = dist, design = "crd",
                       scale = FALSE)

plot(variograma, ylab = "Semivariance", xlab = "Distance")

# Gaussian Model
ols <- spVariofit(variograma, cov.model = "gaussian", weights = "equal",
                  max.dist = dist)

lines(ols, col = 1)

# Compute the model and get the analysis of variance table
mod <- aovGeo(ols, cutoff = 0.6)

# Multivariate T test
spMVT(mod)
```

---

spScottKnott *The Scott-Knott Clustering Algorithm*

Description

This function implements the Scott-Knott Clustering Algorithm for objects of class SARcrd, SARrcbd, and GEOanova.

Usage

```
spScottKnott(x, sig.level = 0.05)

## S3 method for class 'SARanova'
spScottKnott(x, sig.level = 0.05)

## S3 method for class 'GEOanova'
spScottKnott(x, sig.level = 0.05)
```

Arguments

x	a fitted model object of class SARcrd, SARrcbd or GEOanova.
sig.level	a numeric value between zero and one giving the significance level to use.

Details

For objects of class SARcrd or SARrcbd this function performs the standard Scott-Knott Clustering Algorithm provided by the function SK on the adjusted response.

For objects of class GEOanova, the method is modified to take into account the spatial dependence among the observations. The method is described in Nogueira (2017).

Value

a data frame containing the means and its group

References

NOGUEIRA, C. H. Testes para comparações múltiplas de médias em experimentos com tendência e dependência espacial. 142 f. Tese (Doutorado em Estatística e Experimentação Agropecuária) | Universidade Federal de Lavras, Lavras, 2017

Examples

```
data("crd_simulated")

#Geodata object
geodados <- as.geodata(crd_simulated, coords.col = 1:2, data.col = 3,
                       covar.col = 4)
h_max <- summary(geodados)[[3]][[2]]
dist <- 0.6*h_max

# Computing the variogram
variograma <- spVariog(geodata = geodados,
                       trend = "cte", max.dist = dist, design = "crd",
                       scale = FALSE)

plot(variograma, ylab = "Semivariance", xlab = "Distance")

# Gaussian Model
ols <- spVariofit(variograma, cov.model = "gaussian", weights = "equal",
                  max.dist = dist)


# Compute the model and get the analysis of variance table
mod <- aovGeo(ols, cutoff = 0.6)

# Scott-Knott clustering algorithm
spScottKnott(mod)
```

`spTukey` *Compute Tukey Honest Significant Differences for a Spatially Correlated Model*

Description

Perform multiple comparisons of means treatments based on the Studentized range statistic when the errors are spatially correlated.

Usage

```
spTukey(x, sig.level = 0.05)

## S3 method for class 'SARcrd'
spTukey(x, sig.level = 0.05)

## S3 method for class 'SARrcbd'
spTukey(x, sig.level = 0.05)

## S3 method for class 'GEOanova'
spTukey(x, sig.level = 0.05)
```

Arguments

`x`	A fitted model object of class SARcrd, SARrcbd or GEOanova.
`sig.level`	A numeric value between zero and one giving the significance level to use.

Details

For objects of class SARcrd or SARrcbd this function performs the standard Tukey's 'Honest Significant Difference' method provided by the function `TukeyHSD` on the adjusted response.

For objects of class GEOanova, the method is modified to take into account the spatial dependence among the observations. First, we estimate a contrast matrix (C) using cont.tuk function and then after estimate the spatial mean of each treatment (μ_i) we can assess the significance of the contrast by

$$|c_i \mu_i| > HSD_i$$

where $HSD_i = q(\alpha, k, \nu) * sqrt(0.5 * w_i i)$ and k is the number of treatments, α is the level of significance, ν is the degree of freedom of the model, $w_i i$ is the variance of the i-th contrast.

Value

a data frame containing the original mean, the spatially filtered mean and its group. For the class GEOanova, the spatial dependence is filtered out using geostatistics, while for the class SARanova the adjusted response based on SAR model is employed.

References

NOGUEIRA, C. H. Testes para comparações múltiplas de médias em experimentos com tendência e dependência espacial. 142 f. Tese (Doutorado em Estatística e Experimentação Agropecuária) | Universidade Federal de Lavras, Lavras, 2017

Examples

```
data("crd_simulated")

#Geodata object
geodados <- as.geodata(crd_simulated, coords.col = 1:2, data.col = 3,
                       covar.col = 4)
h_max <- summary(geodados)[[3]][[2]]
dist <- 0.6*h_max

# Computing the variogram
variograma <- spVariog(geodata = geodados,
                       trend = "cte", max.dist = dist, design = "crd",
                       scale = FALSE)

plot(variograma, ylab = "Semivariance", xlab = "Distance")

# Gaussian Model
ols <- spVariofit(variograma, cov.model = "gaussian", weights = "equal",
                  max.dist = dist)

lines(ols, col = 1)

# Compute the model and get the analysis of variance table
mod <- aovGeo(ols, cutoff = 0.6)

# Tukey's HSD
spTukey(mod)
```

`spVariofit` *Fit a variogram model*

Description

Fit a parametric model to a variogram created by the function spVariog.

Usage

```
spVariofit(x, ...)
```

Arguments

x	an object of class spVariog.
...	further arguments to be passed to variofit function.

Details

This function is a wrapper to variofit and can be used to fit a parametric model to a variogram using either ordinary least squares or weighted least squares. It takes as the main argument a spVariog object and others arguments should be passed to ... such as "cov.model" and so on.

Value

an object of class SpVariofit which is a list containing the following components:

mod	an object of class variofit
data.geo	an object of class geodata
des.mat	the design matrix
trend	a character specifying the type of spatial trend

See Also

variofit

Examples

```
data("crd_simulated")
dados <- crd_simulated

#Geodata object
geodados <- as.geodata(dados, coords.col = 1:2, data.col = 3,
                       covar.col = 4)
h_max <- summary(geodados)[[3]][[2]]
dist <- 0.6*h_max

# Computing the variogram
variograma <- spVariog(geodata = geodados,
                       trend = "cte", max.dist = dist, design = "crd",
                       scale = FALSE)

plot(variograma, ylab = "Semivariance", xlab = "Distance")

# Spherical Model
ols1 <- spVariofit(variograma, cov.model = "spherical", weights = "equal",
                   max.dist = dist)
lines(ols1)
```

spVariog *Compute empirical residual variogram for CRD or RCBD.*

Description

Compute empirical residual variogram for a Completely Randomized Design (CRD) or a Randomized Complete Block Design (RCBD) by a call to variog function of the package geoR.

Usage

```
spVariog(geodata, resp = NULL, treat = NULL, block = NULL, coords = NULL,
data = NULL, trend = c("cte", "1st"), scale = FALSE, max.dist,
design = c("crd", "rcbd"), ...)
```

Arguments

geodata	an object of class geodata in which the response variable should be given in 'data.col', the coordinates in 'coords.col', the treatment vector should be given as the first column of 'covar.col' and block as the second one.
resp	either a vector of response variables or a character giving the column name where it can be found in 'data'. Optional argument, just required if geodata is not provided.
treat	either a vector of treatment factors or a character giving the column name where it can be found in 'data'. Optional argument, just required if geodata is not provided.
block	either a vector of block factors or a character giving the column name where it can be found in 'data'. Optional argument, just required if geodata is not provided.
coords	either a 2-column matrix containing the spatial coordinates or a character vector giving the columns name where the coordinates can be found in 'data'. Optional argument, just required if geodata is not provided.
data	a data frame in which the variables specified as characters will be found. Optional argument, just required if geodata is not provided.
trend	type of spatial trend considered.
scale	logical argument. Should the coordinates be scaled? We recommend this argument to be set as TRUE if your spatial coordinates have high values as in UTM coordinate system otherwise, you could get errors in the calculations. See 'Details'.
max.dist	numerical value defining the maximum distance for the variogram. See variog documentation for additional information.
design	type of experimental design. "crd" corresponds to Completely Randomized Design and "rcbd" corresponds to Randomized Complete Block Design.
...	further arguments to be passed to variog function.

Details

This function provides a wrapper to variog to compute residual variogram for experimental designs. The residuals are obtained by

$$\varepsilon = Y - X\beta,$$

where Y is the vector of response, X is the design matrix built according to the experimental design chosen, and β is the vector of coefficients estimated by the linear model.

If scale = TRUE the spatial coordinates will be scaled for numerical reasons. The scale is made by subtracting the minimum spatial coordinate value from all others.

Value

An object of class spVariog which is a list with the following components:

`vario.res`	an object of class variogram
`data.geo`	an object of class geodata
`des.mat`	the design matrix
`trend`	a character specifying the type of spatial trend

See Also

variog

Examples

```
data("crd_simulated")
dados <- crd_simulated

#Geodata object
geodados <- as.geodata(dados, coords.col = 1:2, data.col = 3,
                               covar.col = 4)
h_max <- summary(geodados)[[3]][[2]]
dist <- 0.6*h_max

# Computing the variogram
variograma <- spVariog(geodata = geodados,
                       trend = "cte", max.dist = dist, design = "crd",
                       scale = FALSE)

plot(variograma, ylab = "Semivariance", xlab = "Distance")
```

Index

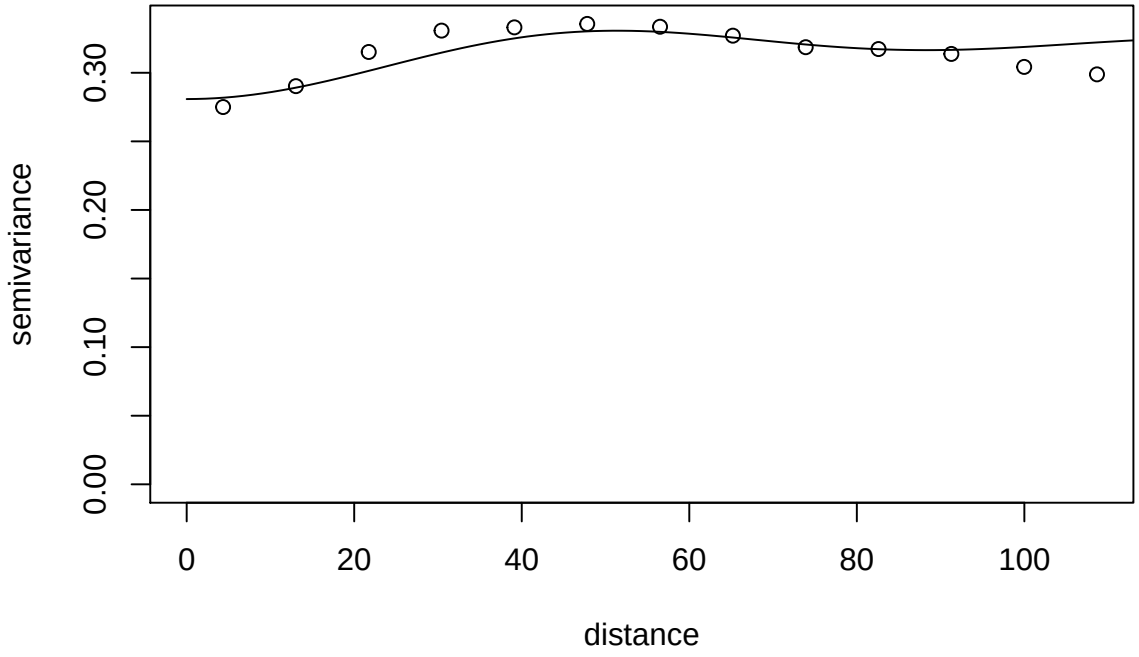
APÊNDICE - B: Relatório Gerado pela Aplicação

Report of SpAnova

Geostatistical Approach

Semivariogram

Correlation function: wave.



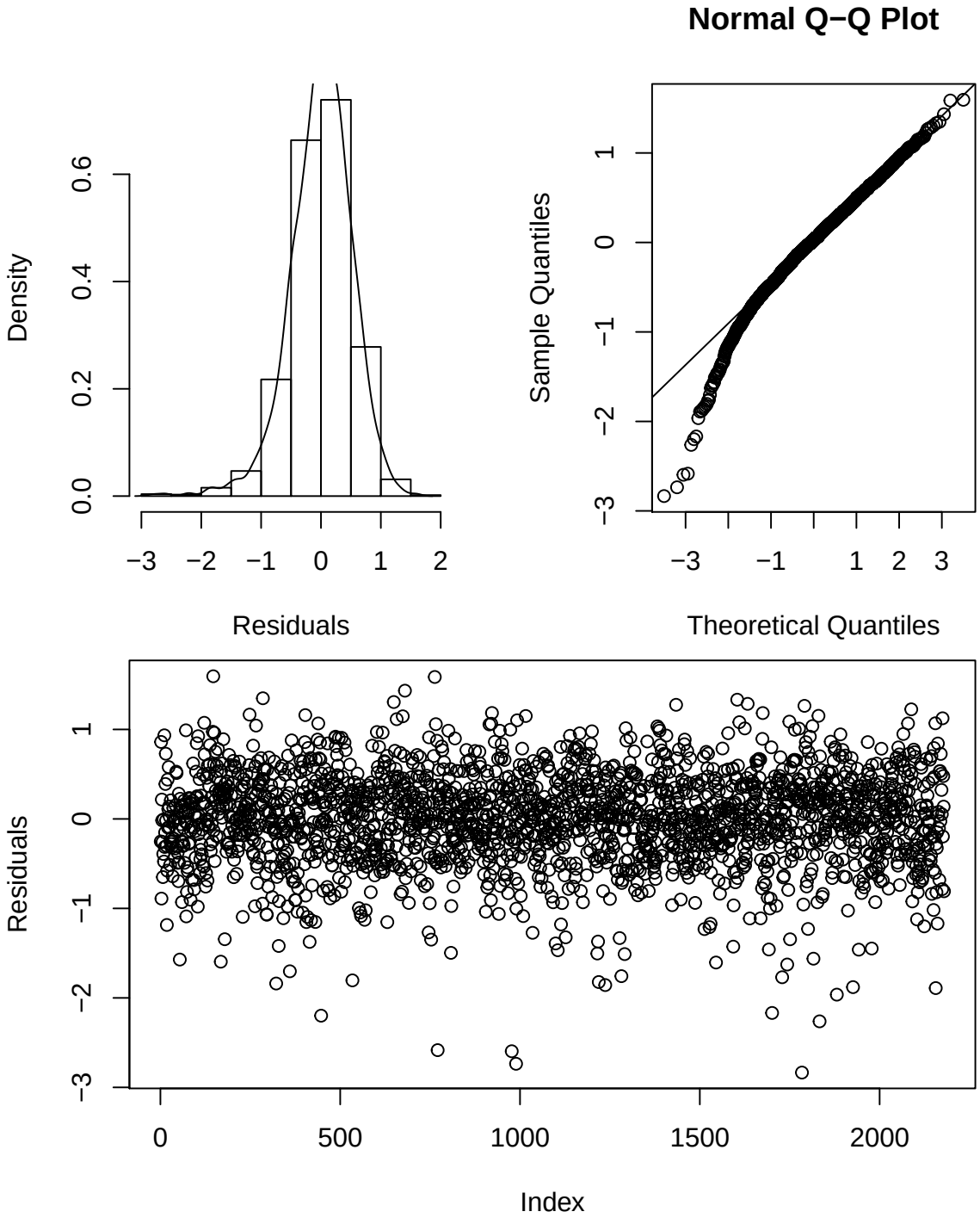
Semivariogram Parameter Estimates

Partial_Sill	Range	Nugget	Sill
0.041	11.432	0.281	0.322

Analysis of Variance

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Treatment	12	28.6050	2.3837	4.9762	< 2.22e-16	***
Block	3	13.6548	4.5516	9.5017	< 2.22e-16	***
Residuals	2163	1036.1441	0.479			
Total	2178	1078.4039				

Checking the Residuals



Assumption	Test	P.value
Normality	Shapiro-Wilk	0.0000000
Independence	Moran-I	0.5205494

According to Shapiro-Wilk normality test at 5% of significance, residuals cannot be considered normal.

According to Moran I test at 5% of significance, there is no spatial correlation among the residuals.

Multiple Comparison Procedure

Procedure: Scott-Knott

	mean	filtered.mean	groups
10	3.747	3.898	a
2	3.490	3.820	a
13	3.640	3.814	a
9	3.946	3.802	b
8	3.685	3.775	b
7	3.716	3.758	b
6	3.712	3.740	b
11	3.783	3.726	b
4	3.514	3.723	c
5	3.750	3.663	c
3	3.765	3.626	c
12	3.826	3.622	c
1	3.485	3.565	c

Treatments with the same letter are not significantly different at 5% of significance.