UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CURSO DE LICENCIATURA EM MATEMÁTICA

VIVIANE VANESSA DÖHL

VALIDAÇÃO CRUZADA EM GEOESTÁTISTICA

TRABALHO DE CONCLUSÃO DE CURSO

TOLEDO - PR
2015

VIVIANE VANESSA DÖHL

VALIDAÇÃO CRUZADA EM GEOESTATÍSTICA

Trabalho de conclusão de Curso apresentado a Coordenação do Curso Superior de Licenciatura em Matemática (COMAT) da Universidade Tecnológica Federal do Paraná (UTFPR) Campus Toledo, como requisito parcial para obtenção do título de Licenciado em Matemática.

Orientadora Prof. Dra. Rosangela Aparecida Botinha Assumpção

TOLEDO

2015

AGRADECIMENTOS

A minha orientadora, Prof. Drª. Rosangela Aparecida Botinha Assumpção pela oportunidade de pesquisarmos juntas me orientando no caminho do conhecimento científico.

A todos os professores e colegas do curso de Licenciatura em Matemática, pelas contribuições, ensinamentos e companheirismo durante este período e em especial a Profª. Larissa Hagedorn Vieira pelo apoio com o LaTeX, ao Profº. Marcio Paulo de Oliveira nas contribuições e ensinamentos sobre o *software* R, a Profª. Daniela Trentin Nava pelas contribuições na interpretação dos gráficos e pela participação em minha banca e a professora Araceli Ciotti de Marins pela participação em minha banca.

Agradeço imensamente aos meus pais Elemar e Cleide pelo incentivo nos estudos desde meus primeiros dias de vida.

Ao meu noivo Jean pela companhia e suporte nas horas difíceis, como também por suas palavras motivadoras e acolhedoras em todos os momentos de dificuldade.

A todos aqueles que não foram citados mas que de alguma forma contribuiram para que este trabalho fosse realizado.

Meus agradecimentos!

RESUMO

Este trabalho buscou apresentar um estudo geoestatístico, utilizando modelos espaciais lineares a fim de testar o método de validação cruzada. Para testar esse método foram simulados dados com dependência espacial conhecida e identificado o melhor modelo segundo o método. A intenção principal é mostrar que o melhor modelo identificado pelo método é de fato o modelo utilizado na simulação dos dados, garantindo assim, que a validação cruzada pode ser considerada um método robusto. Os dados experimentais nos quais foram aplicadas as técnicas geoestatísticas e em especial a validação cruzada se referem a produtividade de soja em um Latossolo Vermelho Distroférrico Típico (EMBRAPA, 2009) com 57 ha, localizada no município de Cascavel - PR. Os resultados foram positivos e a técnica de validação denominada validação cruzada foi vista como eficiente neste trabalho. Os dados experimentais foram então analisados segundo suas características descritivas, homogeneidade, dispersão, indícios de normalidade e foram feitas as conclusões cabíveis com a precisão e segurança desejada em relação ao método de validação. Foi constatado que a validação cruzada indicou o modelo exponencial para os dados simulados exponencialmente e indicou o modelo gaussiano para os dados simulados com esse mesmo comportamento.

Palavras-chave: Simulação de Dados, Método Robusto, Krigagem.

ABSTRACT

This study aimed to present a geostatistical study using linear spatial models in order to test the cross-validation method. To test this method were simulated data with known spatial dependence and identified the best model using the method. The main intention was to show that the best model identified by the method is the model used in the simulation data, thus ensuring that cross-validation can be considered a robust method. The experimental data on which were applied geostatistics and especially cross-validation refer to soybean yield in an Oxisol (EMBRAPA, 2009) with 57 ha, located in the city of Cascavel - PR. It has been found that cross-validation indicated the exponential model for exponentially simulated data and indicated the Gaussian model for the simulated data to such behavior. The analyzes were done as if we did not know the behavior of simulated data looking only for the result of cross-validation. The results were positive, then the validation technique called cross-validation was seen as efficient in this study. The experimental data were then analyzed using descriptive characteristics, homogeneity, dispersion, normal evidence and the conclusions were made applicable to the desired accuracy and safety in relation to the validation method.

Keywords: Data Simulation, Robust Method, Kriging .

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Os processos de análise do solo e produtividade de culturas agrícolas que visam aplicações da Agricultura de Precisão vêm crescendo a cada dia. Atualmente os estudos que a envolvem consistem na compreensão da distribuição espacial dos dados que são coletados em uma referida área agrícola. Esses dados são analisados de modo que possamos inferir sobre toda a área com determinada precisão, de um jeito que o agricultor possa tratar cada parte de sua área de acordo com as necessidades específicas do local.

Nesse sentido, a geoestatística vem com o propósito de ser instrumento para esse tipo de análise, desde que os dados coletados apresentem algum tipo de dependência espacial.

A geoestatística consiste basicamente em um estudo sobre as variáveis regionalizadas, que consiste em uma função espacial numérica que varia de um local para outro com uma certa continuidade, ou seja, existe uma função que é capaz de expressar essa dependência. Esta função é denominada Modelo Espacial Linear. Os modelos mais comuns são o Exponecial, Gaussiano e Esférico.

O método da validação cruzada identifica o melhor modelo para representar os dados reais, permitindo assim que se façam boas inferências. Esse método consiste em retirar um ponto da amostra e estimá-lo verificando a diferença entre o ponto amostrado e o estimado, faz-se isso com todos os pontos separadamente.

A partir dos critérios de comparação para o Erro Médio, Erro Médio Reduzido, Desvio Padrão dos Erros Médios, Desvio Padrão dos Erros Médios Reduzidos, Erro Absoluto e Máxima Verossimilhança, determina-se qual o modelo que melhor representa a realidade daquela área.

Pensando na eficiência que deve ter o modelo, apresenta-se neste trabalho um estudo a respeito da técnica de validação cruzada, que é o método que utilizaremos para a escolha do melhor modelo dado seus parâmetros, e consequentemente confirmada a precisão desta técnica estaremos seguros de inferências realizadas posteriormente.

2 REFERENCIAL BIBLIOGRÁFICO

2.1 GEOESTATÍSTICA

Por Geoestatística entende-se o estudo feito sobre determinada amostra, quando é identificada continuidade espacial, ou seja, quando é verificado um grau de organização ou continuidade, e então parâmetros como a média e o desvio padrão da Estatística Clássica não são suficientes para representar o fenômeno em questão, pois tratam os dados como independentes. Para determinarmos qual estatísitica usaremos: clássica ou espacial, construímos o semivariograma que expressa essa dependência espacial entre os dados da amostra (VIEIRA, 2000).

A Geoestatística foi desenvolvida por Daniel Kriege, quando através de suas atividades percebeu que para que as variâncias obtidas pela Estatística Clássica fizessem sentido deveria considerar a distância entre os locais onde foram retiradas as amostras.

Desde seu surgimento na década de 50 a Geoestatísitica vem ganhando seu espaço e junto com ela surgiram algumas teorias, uma que podemos citar é a de Matheron que foi desenvolvida na década de 60 e supõe a existência de dependência espacial entre os pontos amostrais de uma determinada variável, sendo assim, existe uma função entre o valor das variáveis e a distância entre os pontos amostrados.

Segundo (ANDRIOTTI, 1989), a Geoestatística se baseia em conceitos probabilísticos, que utilizam os dados coletados para estimar a correlação espacial e para fazer as estimativas. Na Geoestatística o ideal é que os dados coletados na área desejada sejam consistentes em toda a área.

A Geoestatística é utilizada para estimar incertezas associadas a locais onde não foram retiradas amostras e ainda predizer valores nesses locais utilizando uma técnica muito importante conhecida como Krigagem.

Para que possamos estimar valores desconhecidos, ou seja, para predizermos valores de certo ponto no espaço onde não foi coletada amostra, a partir dos dados amostrais, devemos verificar a variabilidade espacial dos dados e postular um

modelo que descreva essa variabilidade, estimar os parâmetros do modelo se existirem e por fim, caso o modelo seja aceito, ainda nos resta testá-lo e predizer com embasamento estatístico as informações obtidas por esse processo.

2.2 VARIÁVEIS REGIONALIZADAS

Uma variável distribuída no espaço que apresenta aspectos aleatórios, com variações inesperadas e também aspectos estruturados é conhecida como variável regionalizada (VR).

Uma VR corresponde basicamente a uma variável distribuída no espaço, cujo valor esteja relacionado de alguma maneira com a sua posição espacial. Dessa forma não podemos nos referir a geoestatística sem estarmos nos referindo automaticamente as variáveis regionalizadas que compõe o estudo geoestatístico.

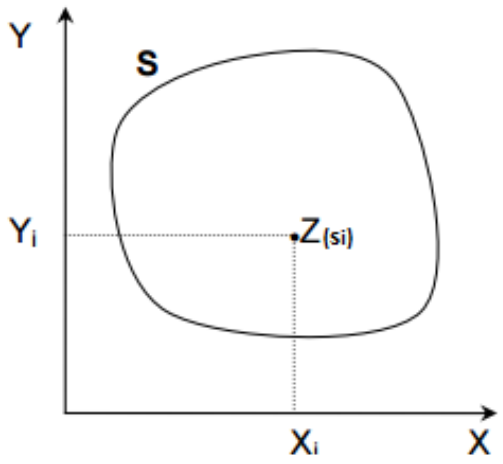

Figura 1: Variável aleatória regionalizada $Z_{(s_i)}$.

, onde $Z(si)$ é o atributo de interesse e $si = (xi, yi)$.

Segundo (ANDRIOTTI, 1989), as VR podem ser **estacionárias**, quando o mesmo fenômeno ocorre em toda a amostragem, **estacionárias de segunda ordem**, quando a média e a covariância são invariantes por translação e **intrínsecas**, quando a média e a covariância dos crescimentos são invariantes por translação.

A região de estacionariedade é onde se deseja fazer estimativas, sendo assim, quanto mais dados em uma área tivermos, menor será a quantidade de pontos para estimar naquela área, ou seja, menor a região de estacionariedade.

A isotropia é uma caracteristica marcante a ser verificada ao tratarmos das

VR, uma vez que refere-se a existência ou não de uma direção privilegiada ao longo da amostra. Dizemos que o conjunto é isotrópico, ou que há isotropia, quando não é possível verificar uma direção privilegiada, caso contrário podemos considerar o conjunto de dados anisotrópico, ou seja, existe uma direção privilegiada. Essa direção, segundo a literatura pode corresponder a 0°, 45°, 90° ou 135°.

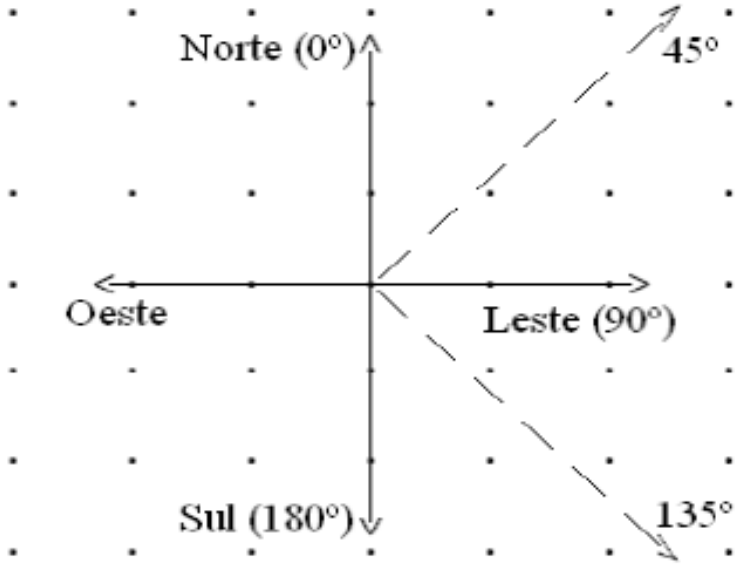

Figura 2: Direções utilizadas no estudo da anisotropia da variável regionalizada.

2.3 SEMIVARIOGRAMA

Comumente vemos a utilização das palavras variograma e semivariograma se referindo aos mesmos valores, entretanto, o variograma é o dobro do semivariograma.

O semivariograma possui parâmetros que permitem ajustar um modelo para obtermos a autocorrelação em função da distância. São eles: Efeito Pepita, Patamar e Alcance.

O semivariograma é definido por (MATHERON, 1963) como um estimador da função semivariância para variáveis regionalizadas com distribuição normal de probabilidade apresentado pela equação 1:

$$\gamma^*(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} [Z(s_i) - Z(s_i + h)]^2 \tag{1}$$

O alcance $(a = g(\varphi_3))$ é a distância a partir do qual os valores passam a ser independentes, ele separa os campos estruturados dos campos aleatórios. Em outras palavras ele reflete o grau de homogeneização entre as amostras (ANDRIOTTI, 1989). O alcance marca exatamente a distância em que o ponto em estudo não sofre influências do ponto vizinho, ou seja, ele não é considerado dependente.

O patamar $(C = \varphi_1 + \varphi_2)$ é o ponto de estabilidade do semivariograma, igual a variância dos valores da variável correspondente ao ponto em que o mesmo estabiliza. Deste ponto em diante, considera-se a não existência espacial entre as amostras.

Por fim, o efeito pepita (φ_1) é atribuído a erros de mensuração aliado ao fato dos dados não terem sido coletados em intervalos pequenos para representar o comportamento espacial do fenômeno a distância praticamente nula (SANTANA, 2011). É o valor da função semivariograma na origem $(h = 0)$. Teoricamente esse valor deveria ser zero, pois duas amostras tomadas nas mesmas coordenadas não deveriam possuir diferença entre os valores da variável Z. A interpretação que damos ao efeito pepita é de erros de medição ou de variabilidade em pequena escala (CRESSIE, 1989).

Este efeito exerce influência sobre os ponderadores e principalmente sobre a krigagem, onde sua correta avaliação é uma forma de intepretar o comportamento do semivariograma à origem.

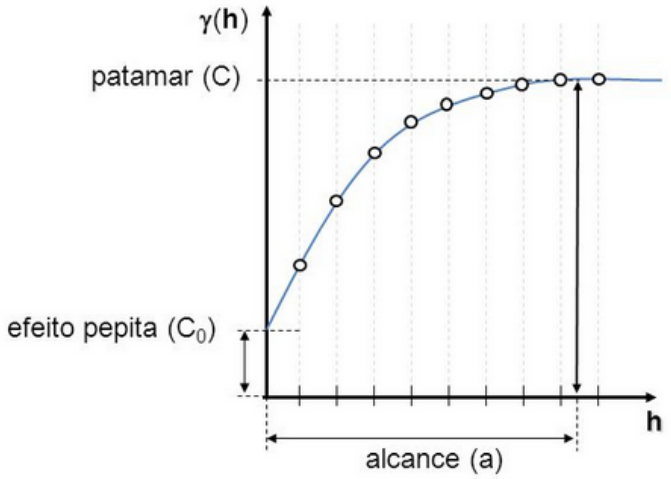

Figura 3: Semivariograma experimental.

Dessa forma, o patamar $(C = \varphi_1 + \varphi_2)$, o alcance $(a = g(\varphi_3))$ e o efeito pepita (φ_1), se tornam os parâmetros que devemos buscar ao quantificar a dependência das variáveis regionalizadas em estudo.

2.4 VARIÂNCIA DE ESTIMAÇÃO

A Variância de Estimação quantifica o valor do erro que se comete ao se avaliar Z_i por meio de $\hat{Z}_i$. Sendo Z_i o valor real e $\hat{Z}_i$ o valor estimado (ANDRIOTTI, 1989).

2.5 MODELOS ESPACIAIS LINEARES

Segundo (VIEIRA, 2000), o ajuste de um modelo teórico ao semivariograma experimental é importante na aplicação da teoria das variáveis regionalizadas, pois todos os cálculos geoestatísticos que seguem dependem do modelo do semivariograma. Nesse sentido é importante que se observe com atenção os parâmetros, para que a partir do modelo obtido possamos criar mapas precisos, que é um dos objetivos do estudo em geoestatística.

O semivariograma mostra as características discretas dos pares: valor e localização, porém esse gráfico é contínuo e não pontual como parece, para tanto é necessário que ajustemos uma função, ou modelo, como também podemos nos referir, de modo que se aproxime da melhor maneira possível dos pontos do semivariograma.

Uma das caracterísiticas que esse modelo, função, precisa garantir é que as variâncias calculadas sejam todas positivas, e então temos três modelos que se adaptam a maioria das situações encontradas, são eles: esférico, exponencial e gaussiano.

Para que possamos analisar a melhor adaptação do modelo a ser utilizado, precisamos verificar os quatro parâmetros essenciais para que definamos a estrutura de dependência espacial, são eles:

i) **Efeito pepita** (φ_1)

ii) **Contribuição** (φ_2)

iii) **Alcance** $(a = g(\varphi_3))$

iv) **Patamar** $(C = \varphi_1 + \varphi_2)$

Considerando os parâmetros acima temos os modelos:

a) Modelo Esférico

Este modelo apresenta crescimento rápido na origem e atinge o patamar a $2/3$ do alcance. Segundo (CRESSIE, 1989), este modelo é válido em $\mathbb{R}$, $\mathbb{R}^2$ e $\mathbb{R}^3$ e tem

como expressão a Equação (2):

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ \varphi_1 + \varphi_2 \left[\frac{3}{2} \left(\frac{h}{\varphi_3} \right) - \frac{1}{2} \left(\frac{h}{\varphi_3} \right)^3 \right], & 0 < h \leq \varphi_3 \\ \varphi_1 + \varphi_2, & h > \varphi_3 \end{cases} \tag{2}$$

A função de covariância é expressa por:

$$C(h) = \begin{cases} \varphi_1 + \varphi_2, & h = 0 \\ \varphi_2 \left[1 - \frac{3}{2} \left(\frac{h}{\varphi_3} \right) + \frac{1}{2} \left(\frac{h}{\varphi_3} \right)^3 \right], & 0 < h \leq \varphi_3 \\ 0, & h > \varphi_3 \end{cases} \tag{3}$$

A função de correlação espacial é definida como:

$$\rho(h) = \begin{cases} 1, & h = 0 \\ 1 - \frac{3}{2} \left(\frac{h}{\varphi_3} \right) + \frac{1}{2} \left(\frac{h}{\varphi_3} \right)^3, & 0 < h \leq \varphi_3 \end{cases} \tag{4}$$

É importante destacar que o modelo esférico não tem segunda derivada e por isso não serve para estudos de inferência.

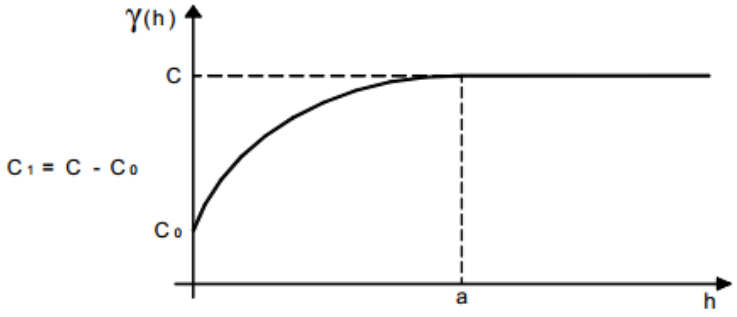

Figura 4: Representação gráfica do modelo esférico.

b) Modelo Exponencial

Este modelo apresenta comportamento aproximadamente linear na origem e atinge o patamar assintoticamente com alcance prático definido como a distância na qual o valor do modelo é 95% de φ_2, sendo o alcance prático dado por a = 3φ_3. Este modelo é válido em $\mathbb{R}$, $\mathbb{R}^2$ e $\mathbb{R}^3$ (CRESSIE, 1989), e tem como expressão a Equação (5):

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ \varphi_1 + \varphi_2 \left[1 - exp\left(-\frac{h}{\varphi_3}\right)\right], & 0 < h \leq \varphi_3 \\ \varphi_1 + \varphi_2, & h > \varphi_3 \end{cases} \tag{5}$$

A função de covariância é expressa por:

$$C(h) = \begin{cases} \varphi_1 + \varphi_2, & h = 0 \\ \varphi_2 \left[exp\left(-\frac{h}{\varphi_3}\right)\right], & 0 < h \leq \varphi_3 \\ 0, & h > \varphi_3 \end{cases} \tag{6}$$

A função de correlação espacial é definida como:

$$\rho(h) = \begin{cases} 1, & h = 0 \\ exp\left(-\frac{h}{\varphi_3}\right), & 0 < h \leq \varphi_3 \end{cases} \tag{7}$$

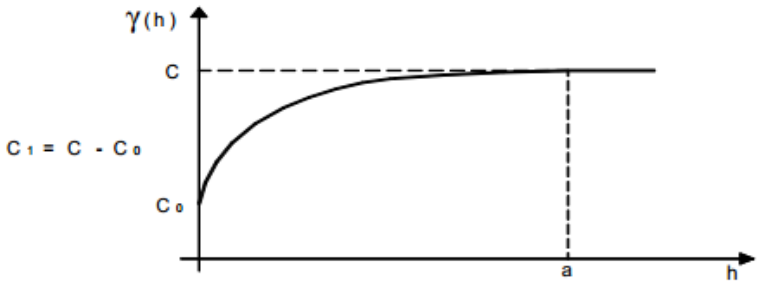

Figura 5: Representação gráfica do modelo exponencial.

c) Modelo Gaussiano

O Modelo gaussiano apresenta comportamento parabólico na origem e é utilizado para modelar um fenômeno extremamente contínuo. Também atinge o patamar apenas assintoticamente e o alcance prático é dado por $a = \sqrt{3}\varphi_3$. Este modelo é válido em $\mathbb{R}$, $\mathbb{R}^2$ e $\mathbb{R}^3$ (CRESSIE, 1989) e tem como expressão a Equação (8):

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ \varphi_1 + \varphi_2 \left\{ 1 - exp\left[-\left(\frac{h}{\varphi_3}\right)^2 \right] \right\}, & 0 < h \leq \varphi_3 \\ \varphi_1 + \varphi_2, & h > \varphi_3 \end{cases} \tag{8}$$

A função de covariância é expressa por:

$$C(h) = \begin{cases} \varphi_1 + \varphi_2, & h = 0 \\ \varphi_2 \left\{ exp\left[-\left(\frac{h}{\varphi_3}\right)^2 \right] \right\}, & 0 < h \leq \varphi_3 \\ 0, & h > \varphi_3 \end{cases} \tag{9}$$

A função de correlação espacial é definida como:

$$\rho(h) = \begin{cases} 1, & h = 0 \\ exp\left[-\left(\frac{h}{\varphi_3}\right)^2 \right], & 0 < h \leq \varphi_3 \end{cases} \tag{10}$$

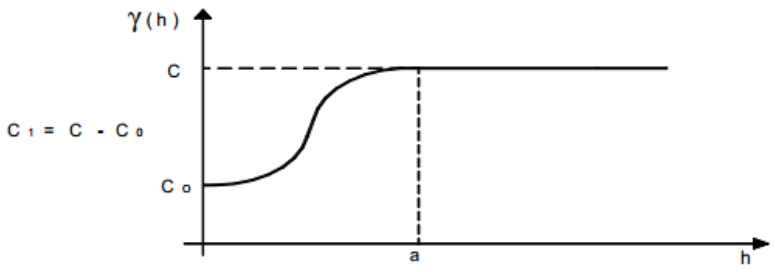

Figura 6: Representação gráfica do modelo gaussiano.

2.6 VALIDAÇÃO CRUZADA

Ao longo da pesquisa salientamos a necessidade de escolher o modelo que melhor se ajusta aos pontos do semivariograma de modo a tornar precisa a krigagem posterior, dessa forma neste tópico trataremos de um dos métodos de escolha do modelo denominado validação cruzada.

Segundo (ISAAKS; SRIVASTAVA, 1989), a validação cruzada é um método que permite comparar valores amostrados e estimados para que o melhor modelo de estimação seja escolhido.

Esse método seleciona o modelo que melhor descreve a dependência espacial das variáveis em função das distâncias, e consiste em supor que um dado não tenha sido observado e então retira-se ele da amostra e obtém-se uma nova estimativa com base nos dados restantes. Isso é feito com todos os pontos da amostra e assim, para todos eles existirá o valor real e o valor estimado, e portanto será possível determinar o erro de estimação e optar pelo melhor modelo semivarigráfico (SANTANA, 2011).

O valor da amostra, em certa localização (s_i), é temporariamente descartado do conjunto e então é feita uma estimação por krigagem no mesmo local $\hat{Z}(s_i)$ usando as amostras restantes. Dessa forma conseguimos obter o erro médio (EM), conforme a equação:

$$EM = \frac{1}{n}\sum(Z(s_i) - \hat{Z}(s_i)) \tag{11}$$

em que n é o número de dados; $Z(s_i)$, valor observado no ponto (s_i); e $\hat{Z}(s_i)$, valor predito por krigagem ordinária no ponto (s_i). Este procedimento de "deixar um ponto de fora" é repetido para todas as amostras disponíveis.

Além do EM, (CRESSIE, 1989) apresenta o erro médio reduzido (ER), desvio padrão dos erros médios (S_{EM}), desvio padrão dos erros médios reduzidos (S_{ER}) e o erro absoluto (EA), como instrumento para avaliar os modelos no método da validação cruzada.

O erro médio reduzido (ER) é definido como:

$$ER = \frac{1}{n}\sum_{i=1}^{n}\frac{(Z(s_i) - \hat{Z}(s_i))}{\sigma(\hat{Z}(s_i))} \tag{12}$$

em que, $\sigma(\hat{Z}(s_i))$ é o desvio padrão da krigagem no ponto (s_i), sem considerar a observação no ponto $Z(s_i)$.

O desvio padrão dos erros reduzidos (S_{ER}) é definido como:

$$S_{ER} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(Z(s_i) - \hat{Z}(s_i))}{\sigma(\hat{Z}(s_i))}} \tag{13}$$

Segundo (CRESSIE, 1989) o valor populacional para o EM e o ER deverá ser o

mais próximo de 0, enquanto o valor do S_{ER} deve ser o mais próximo possível de 1 e os valores de S_{EM} e EA também o menor possível. Seguindo esses critérios é feita a escolha do modelo que melhor se ajusta a nuvem de pontos.

Se a validação cruzada apresentar os seus resultados sobre uma reta de regressão próxima da reta bissetriz e com pequena dispersão, podemos estar confiantes de uma boa estimativa (CRESSIE, 1989).

2.7 MÁXIMA VEROSSIMILHANÇA - ML

A Máxima Verossimilhança - ML é um método de estimação considerado não viciado e eficiente o qual consiste em maximixar a função de densidade de probabilidade (MARDIA; MARSHALL, 1984). O melhor modelo para um processo será aquele que apresentar maior valor do logaritmo da função verossimilhança. Neste trabalho servirá como “critério de desempate” nos casos que a validação cruzada indicar iguais parâmetros para diferentes modelos.

2.8 KRIGAGEM

A krigagem é um interpolador geoestatítico, que permite estimar os valores das variáveis distribuídas no espaço, utilizando-se o semivariograma. Esse método se destaca entre os demais por levar em consideração a dependência espacial. A krigagem consite em um conjunto de técnicas de estimação baseado na minimização da variância do erro (SANTANA, 2011).

A geoestatítica, através da krigagem, preocupa-se em dizer o quão distante um valor estimado está do valor real. Ela é utilizada para criar mapas de krigagem, mapas de erros padrão, mapas de probabilidade e mapas de percentis.

A krigagem é conhecida como um processo de cálculo que minimiza a variância de estimação de amostras determinadas para então determinar o melhor conjunto de ponderadores. Ela tem como verdade que vizinhanças diferentes conduzem a resultados diferentes. Podemos apontar duas krigagens distintas, uma conhecida como **ordinária** onde sua média é considerada desconhecida e outra conhecida como **simples** onde a média das variáveis regionalizadas é considerada conhecida.

A krigagem ordinária é mais utilizada por não exigir conhecimento nem estacionariedade da média sobre a área estudada.

É desejável sempre fazer uma melhor estimativa possível em um local não amostrado, minimizando a variância dos dados coletados. A simulação então, permite fazer infinitas

aproximações do variograma real de dados originais. Ao minimizar a variância de estimação, obtem-se a estimação mais precisa possível a partir das informações que dispomos (ANDRIOTTI, 1989).

2.8.1 KRIGAGEM ORDINÁRIA

A krigagem ordinária (KO) traz a ideia de regressão linear ao estimar um valor desconhecido $\hat{Z}(s_i)$ por meio da combinação linear de valores conhecidos $Z(s_i)$. O estimador é definido como:

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) \tag{14}$$

em que $\hat{Z}(s_0)$ é o valor predito do local e n é a quantidade de valores $Z(s_i)$ medidos nos pontos amostrados e λ_i é o peso associado a cada um dos valores $Z(s_i)$ medidos.

A krigagem ordinária é considerada um ótimo estimador pelo fato de produzir estimativas com variância mínima e não viciadas. Para garantirmos que o estimador não seja tendencioso tem-se que garantir que:

$$\sum_{i=1}^{n} \lambda_i = 1 \tag{15}$$

.

3 MATERIAIS E MÉTODOS

Analisou-se uma amostra de dados relacionada a produtividade da soja no ano de agrícola $2004/2005$ na cidade de Cascavel - PR, como também utilizou-se a decomposição de Cholesky para a simulação de dados com a dependência espacial conhecida.

Foram simulados dados com um ajuste exponencial e verificado a eficiência da técnica de validação cruzada em fornecer parâmetros que indicam que o melhor modelo é de fato o exponencial, e de modo análogo foi simulado um conjunto de dados com ajuste gaussiano para de igual forma verificar a eficiência do método de validação cruzada.

Uma vez confirmado e testado o método de validação cruzada foi aplicada a técnica aos dados experimentais confiantes de um bom ajuste.

3.1 DADOS SIMULADOS

Segundo (CRESSIE, 1989), a simulação de processos espaciais estacionários de segunda ordem pode ser feita pelo método de decomposição de Cholesky. Esta é uma forma de garantir que a geração de sequência aleatória respeite uma matriz de correlações.

A decomposição de Cholesky é uma operação matricial que, aplicada ao vetor de números aleatórios, produz outro vetor de números aleatórios que tem a característica de obedecer a uma dada matriz de correlação entre eles (ASSUMPÇÃO, 2010).

Seja $\boldsymbol{Y} = (Y(s_1), \ldots, Y(s_n))^T$ o vetor $n \times 1$ dos dados simulados, os quais representam a realização de um processo estocástico ou função de variáveis aleatórias $Y(s_i),\ s \in S$, em que $s \subset \mathbb{R}^2$ e $\mathbb{R}^2$ é um espaço euclidiano bi-dimensional em diferentes localizações $s_1, \cdots, s_n$.

Considere agora o vetor de médias, $\boldsymbol{\mu} = (\mathrm{E}(Y(s_1)), \ldots, \mathrm{E}(Y(s_n)))^T$ do processo e a matriz de covariância para a distribuição t-student n-variada. Para processos estocásticos que satisfazem a hipótese de estacionaridade de segunda ordem e isotropia, tem-se:

$$\mathrm{E}[Y(s_i)] = \mu, \tag{1}$$

para $i = 1, 2, \cdots, n$ e

$$C(Y(s_i), Y(s_u)) = C(h_{iu}),\ \text{em que } h_{iu} = \|s_i - s_u\|. \tag{2}$$

Nesse caso, cada elemento do vetor é igual a um valor constante μ e cada (i,u)-ésimo elemento da matriz $\mathbf{\Sigma}$, $n \times n$, é igual a $C(h_{iu})$.

Assim, escolhendo o valor de μ e a função covariância $C(h)$, o vetor $\boldsymbol{Y}$, pode ser simulado pela relação:

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{\epsilon} \tag{3}$$

em que $\boldsymbol{L}$ é uma matriz triângular inferior $n \times n$, obtida mediante a decomposição de $\mathbf{\Sigma}$ no produto $\boldsymbol{L}\boldsymbol{L}^T$, chamada de decomposição de cholesky, e $\boldsymbol{\epsilon} = (\epsilon(s_1), \cdots, \epsilon(s_n))^T$ é um vetor de variáveis aleatórias não correlacionadas.

Em resumo esse método a partir de uma matriz dada a decompõe de modo que a matriz se iguale ao produto entre uma outra matriz com a sua transposta, ou seja, dada a matriz M a decomposição de cholesky obtém matrizes N e N^t que multiplicadas resultam em M.

$$M = N.N^t$$

3.2 DADOS EXPERIMENTAIS

A amostra de produtividade da soja foi obtida no ano agrícola $2004/2005$, em uma área de Latossolo Vermelho Distroférrico Típico (EMBRAPA, 2009) com 57 ha, localizada no município de Cascavel - PR, com coordenadas geográficas de 24,95º sul de latitude e 53,57º oeste de longitude, com altitude média de 650 m. O clima da região apresenta-se como temperado mesotérmico e úmido, tipo climático Cfa (Köppen) onde a temperatura anual possui média de 21 ºC. O levantamento topográfico e a verificação do posicionamento dos locais de amostragem foram realizados por meio de receptores GPS, pelo método estático com correção diferencial pós-processada. A grade amostral regular espacialmente georreferenciada utilizada nessa pesquisa foi denotada pelo valor em metros da distância entre os pontos amostrais, isto é, 75×75 m, onde obteve-se o número de amostras de 66 para a referida grade amostral.

A produtividade da soja foi obtida em pontos georreferenciados. Essas informações foram utilizadas para realização de toda análise estatística e posterior confrontação dos resultados. A variedade da soja COODETEC 216 (CD 216) foi semeada na área.

3.3 ANÁLISE DE DADOS

Os dados obtidos pelo método da decomposição de Cholesky e a amostra de produtividade de soja do ano agrícola $2004/2005$, foram analisados utilizando o *software* livre R e seu módulo GeoR (JR; DIGGLE, 2001), sendo feita uma análise exploratória dos dados apre-

sentando estatísticas descritivas, como a média, mediana, quartis, desvio padrão e coeficiente de variação, além da construção de gráficos como o histograma e o box-plot que nos fornecem informações a respeito da existência ou não de pontos discrepantes, simetria e homogeneidade dos dados e também indícios de normalidade.

Tomadas as características de cada conjunto de dados ajustou-se um modelo teórico que melhor representasse o conjunto de pontos, o qual foi selecionado de acordo com o método da validação cruzada, que forneceu os parâmetros através das estimativas do erro médio EM; erro médio reduzido ER; o desvio padrão do erro médio reduzido S_{ER}; desvio padrão do erro médio S_{EM}, o erro absoluto EA e a máxima verossimilhança ML.

O processo de análise utilizou o *software* R, com seus pacotes GeoR (JR; DIGGLE, 2001), Splancs (BIVAND; GEBHARDT, 2000) e Mass (HUBER; GENTLEMAN, 2004).

4 RESULTADOS E DISCUSSÕES

A análise dos dados forneceu os resultados necessários para as conclusões nas duas diferentes amostras de dados: Dados Simulados e Dados Experimentais.

A análise consistiu em testar o método da validação cruzada, com dados simulados, quando já se sabia a priori o modelo que a técnica de validação deveria indicar. O objetivo era testar a técnica antes de aplicá-la aos dados experimentais.

4.1 DADOS SIMULADOS

Utilizando a decomposição de Cholesky e o script do *software* R, foi gerado duas amostras de dados A e B, a qual foi chamada de dados simulados e onde foram aplicadas as técnicas geoestatísticas.

A amostra A foi simulada de modo que os dados tivessem um comportamento exponencial, enquanto a amostra B foi simulada para que os dados tivessem um ajuste gaussiano. Verificou-se então as características nas duas amostras: estatísticas descritivas, homegeneidade, dependência e indícios de normalidade.

A análise exploratória é apresentada na tabela a baixo de acordo com as estatísticas descritivas.

Tabela 1: Estatísticas descritivas dados simulados

Estatísticas	Amostra Simulada A	Amostra Simulada B
n	64	64
Média	3,002	2,909
Mediana	2,951	2,916
Q1	2,772	2,638
Q3	3,231	3,129
Mínimo	2,392	2,373
Máximo	3,992	3,881
Desvio Padrão	0,398	0,326
CV%	13,259	11,209

n: número de elementos amostrais; Q1: primeiro quartil; Q3: terceiro quartil; CV: coeficiente de variação.

Conforme simulado, as amostras apresentam média em torno de $3u.m$ com baixa

variabilidade e ambas foram consideradas homogêneas por apresentarem um coeficiente de variação menor que 30%. Foi possível verificar que as amostras possuem propositalmente a mesma quantidade de elementos.

Para compreender a distribuição espacial dos dados, homogeneidade, dispersão, indícios de normalidade e outras características do conjunto A apresenta-se o gráfico abaixo:

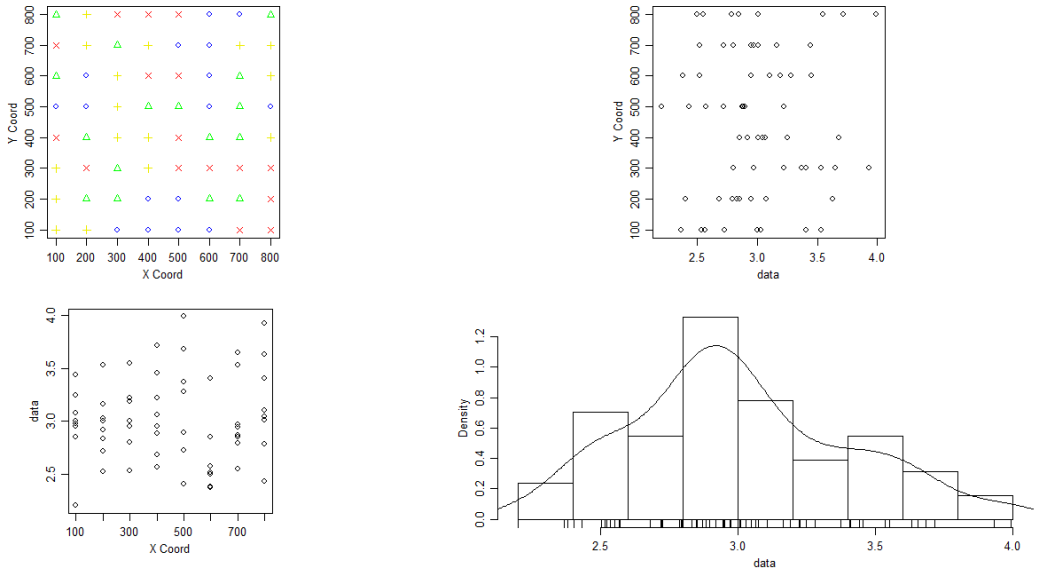

Figura 7: Mapa de estudo espacial para dados simulados A

A primeira imagem mostrou a distribuição dos dados espacialmente de acordo com a proximidade, foi possível ver que eles estão distribuídos de forma aleatória e não foi possível verificar nenhum sinal de anisotropia, ou seja, a amostra aparentou ser isotrópica.

Foi possível verificar que por se tratar de dados simulados o gride de pontos foi determinado de forma regular.

A segunda e terceira imagem foram capazes de analisar a existência de tendências nos dados quando plotados nas direções dos eixos cartesianos, ou seja, foi possível verificar que a hipótese de isotropia observada na primeira imagem foi fortalecida.

Por fim, a quarta imagem tratou-se de um histograma que permitiu visualizar um pequeno indício leptocúrtico com poucos dados acima da curva de probabilidade.

O gráfico box-plot dos dados simulados auxiliou na verificação da simetria do conjunto de dados, como também forneceu informações a repeito da homogeneidade da amostra, além de permitir a visualização de pontos discrepantes ou *outliers*. Apresenta-se abaixo o box-plot da amostra.

Analisando o gráfico box-plot observou-se tratar de dados homogêneos aparentando estarem bem distribuídos com concentração de 50% deles em torno da média, além da

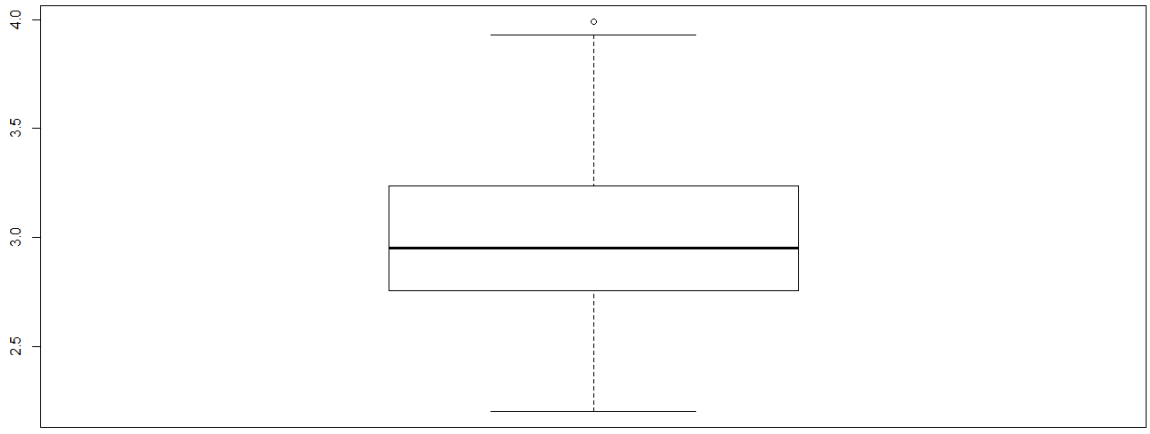

Figura 8: Box-plot dados simulados A.

distância entre o mínimo e o primeiro quartil ser próxima da distância do máximo até o segundo quartil, o que o tornou ideal. Observou-se a existência de apenas um ponto discrepante que se trata exatamente do número de máximo.

O gráfico abaixo apresenta o semivariograma da amostra no qual ajustou-se o modelo sugerido pela validação cruzada posteriormente.

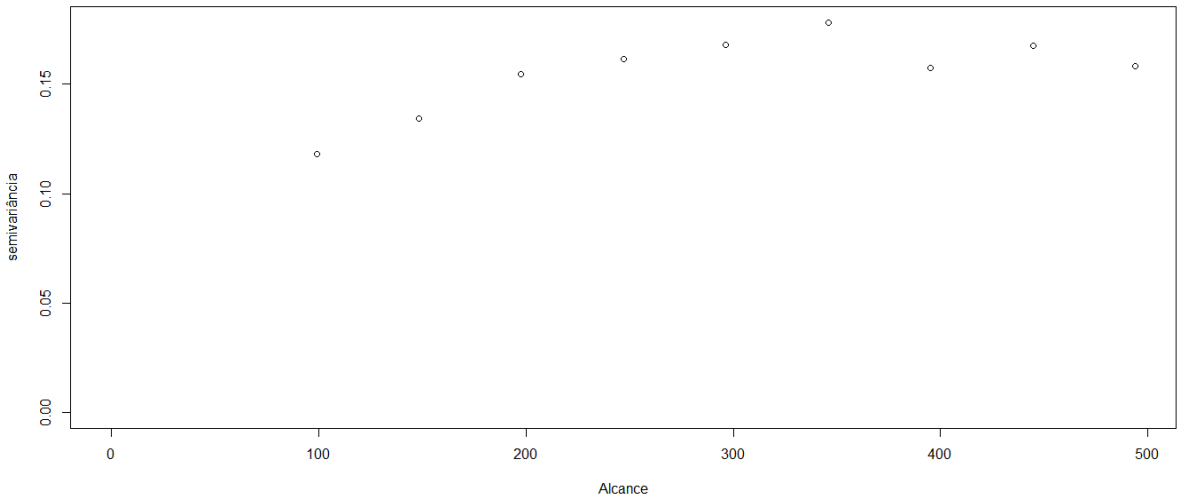

Figura 9: Semivariograma de dados simulados A.

4.1.1 VALIDAÇÃO CRUZADA DADOS SIMULADOS A

A próxima etapa da análise consistiu na verificação da validação cruzada, que conforme anunciado no item 2.6 tem como objetivo indicar qual é o modelo que melhor se ajusta a nuvem de pontos.

Tabela 2: Validação cruzada dados simulados A

Modelo	φ_1	φ_2	$(g(\varphi_3))$	EM	ER	S_{EM}	S_{ER}	EA	ML
Exponencial	0.0	0.158	203,476	**0,0005**	**0,00080**	0,3709	**1,0061**	**19,4644**	-28,92
Esférico	0,0742	0,0857	259,7117	0,0007	0,0009	**0,3698**6	1,0122	19,4982	**-28,72**
Gaussiano	0.0832	0.0754	206,4833	0,0006	0,00085	0,3704	1,0123	19,4777	**-28,72**

φ_1: efeito pepita; φ_2: contribuição; $(g(\varphi_3))$: alcance; EM: erro médio; ER: erro reduzido; S_{EM}: desvio padrão do erro médio; S_{ER}: desvio padrão do erro reduzido, EA: erro absoluto e ML: Máxima Verossimilhança.

Analisados os valores dos parâmetros foi possível perceber que o modelo exponencial foi o que melhor representou a amostra A, apresentando melhores valores de EM : $0,005$, ER : $0,00080$ S_{ER} : $1,0061$ e EA : $19,4644$ comparados com os valores dos demais parâmetros dos outros modelos. Pode ser confirmado visualmente no gráfico abaixo o modelo exponencial descrevendo a dependência espacial de uma boa maneira em relação aos demais modelos.

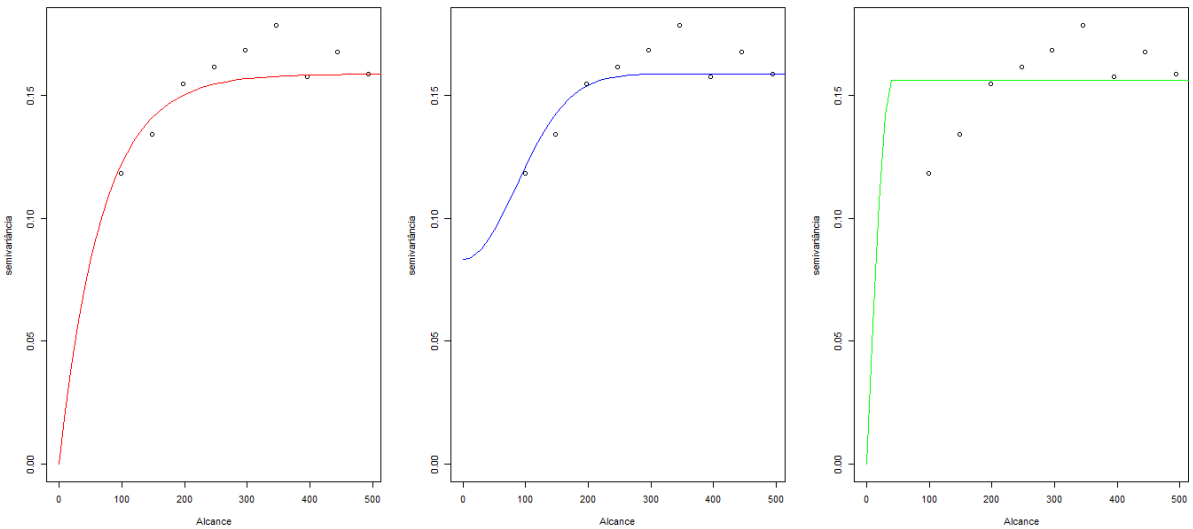

Figura 10: Modelos exponencial, gaussiano e esférico ajustados respectivamente ao semivariograma da amostra A.

Analisou-se de igual forma a amostra B, e para compreender a distribuição espacial dos dados, homogeneidade, dispersão, indícios de normalidade e outras características do conjunto B apresenta-se a figura abaixo:

A primeira imagem mostrou que a amostra B também está distribuída de forma aleatória e não foi possível verificar nenhum sinal de anisotropia, ou seja, a amostra aparentou

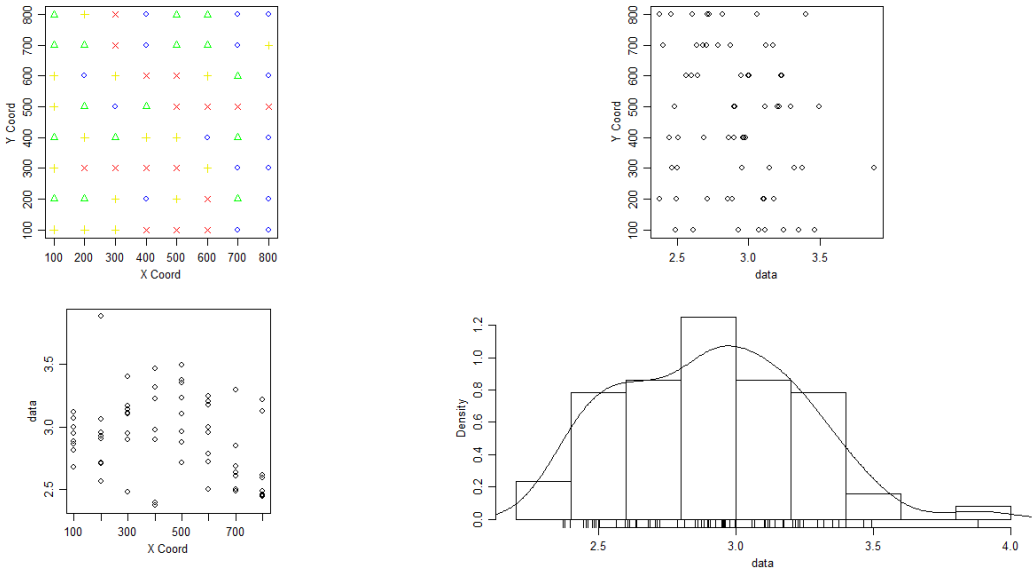

Figura 11: Mapa de estudo espacial para dados simulados B.

ser isotrópica.

A segunda e terceira imagem trazem os gráficos plotados nas direções dos eixos cartesianos, fortalecendo a hipótese de isotropia sugerida na primeira imagem.

A quarta imagem trata-se do histograma que indicou um pequeno indício de ser leptocúrtico com poucos dados acima da curva de probabilidade.

Apresenta-se abaixo o box-plot da amostra.

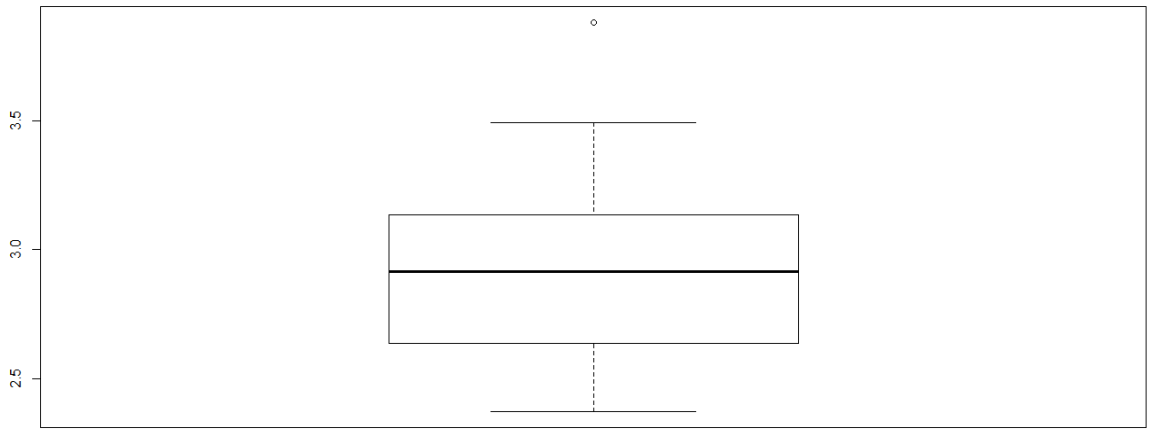

Figura 12: Box-plot dados simulados B.

Analisando o gráfico box-plot observou-se tratar também de dados homegêneos aparentando uma boa distribuição. Foi possível observar que também nesta amostra o ponto

de máximo é um ponto discrepante.

O gráfico abaixo apresenta o semivariograma da amostra B no qual ajustou-se o modelo sugerido pela validação cruzada posteriormente.

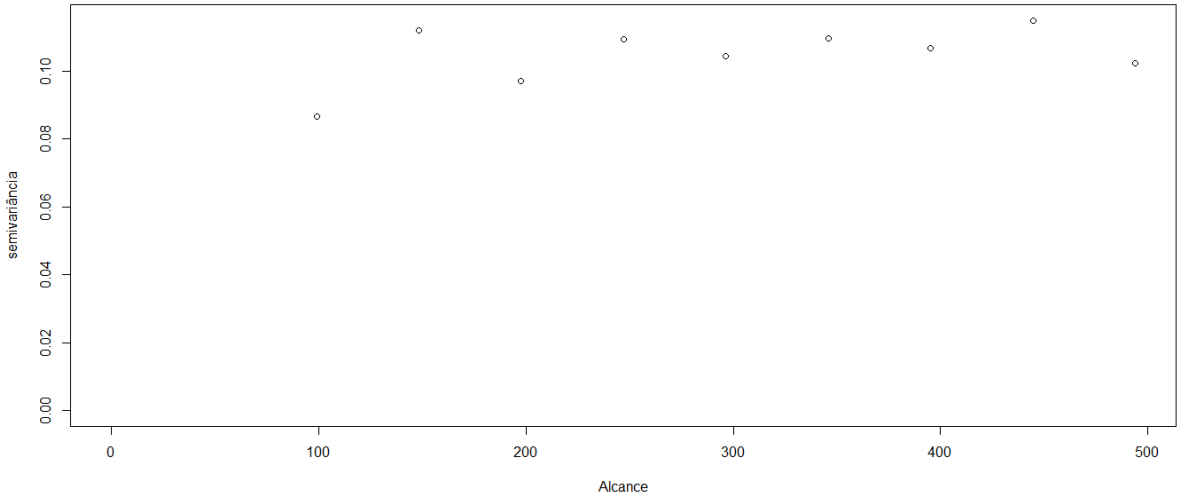

Figura 13: Semivariograma de dados simulados B.

4.1.2 VALIDAÇÃO CRUZADA DADOS SIMULADOS B

Tabela 3: Validação cruzada dados simulados B

Modelo	φ_1	φ_2	$(g(\varphi_3))$	EM	ER	S_{EM}	S_{ER}	EA	ML
Exponencial	0.0	0.1047	162,1174	-0,0003	-0,0006	0,3136	**1,0066**	15,793	-17,18
Esférico	0,0	0,1047	40	**-2,082e-17**	**-7,493e-17**	0,3313	1,0158	17,207	-18,6
Gaussiano	0.0	0.1041	129,709	-0,0005	-0,0009	**0,3098**	1,0085	**15,539**	**-16,78**

φ_1: efeito pepita; φ_2: contribuição; $(g(\varphi_3))$: alcance; EM: erro médio; ER: erro reduzido; S_{EM}: desvio padrão do erro médio; S_{ER}: desvio padrão do erro reduzido, EA: erro absoluto e ML: Máxima Verossimilhança.

Analisando os valores dos parâmetros foi possível perceber que o modelo gaussiano melhor representou o conjunto de dados simulados B, apresentando bons valores para $S_{EM} : 0,3098$, $EA : 15,539$ e $ML : -16,78$ o que pode ser confirmado visualmente no gráfico abaixo, onde o modelo gaussiano melhor descreveu a dependência espacial quando comparado aos demais modelos.

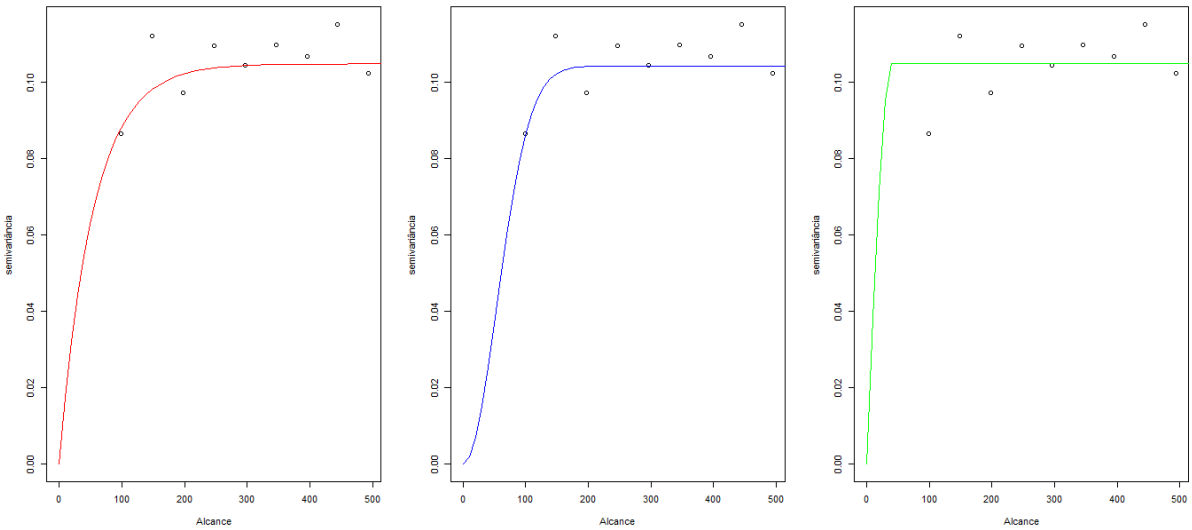

Figura 14: Modelos exponencial, gaussiano e esférico ajustados respectivamente ao semivariograma da amostra B.

Foi possível observar que a validação cruzada indicou os modelos exponencial e gaussiano para as amostras simuladas exponencialmente e gaussianamente em estudos respectivos, o que indicou que o método é válido e podemos utilizá-lo com os dados experimentais, cientes de que a validação indicará de fato o melhor modelo.

4.2 DADOS EXPERIMENTAIS

A amostra caracterizada no item 3.1 se refere a produção da soja em toneladas por hectare que permite avaliar especificamente as regiões que estão produzindo mais ou menos soja para que se possa aplicar as técnicas agrícolas preditas para aquele local.

A análise exploratória de dados apresenta-se na tabela a baixo de acordo com as estatísticas descritivas.

Tabela 4: Estatísticas descritivas amostra experimental

Estatísticas	Amostra Experimental
n	66
Média	3,217
Mediana	3,190
Q1	2,963
Q3	3,478
Mínimo	2,090
Máximo	4,090
Desvio Padrão	0,376
CV%	11,708

n: número de elementos amostrais; Q1: primeiro quartil; Q3: terceiro quartil; CV: coeficiente de variação.

Percebe-se que a média apresenta-se em torno de $3,2t.ha^{-1}$ com uma variabilidade considerável ao analisar a diferença entre o mínimo $2,090$ e o máximo $4,090$, além de ter apresentado homogeneidade verificada pelo coeficiente de variação menor que 30%.

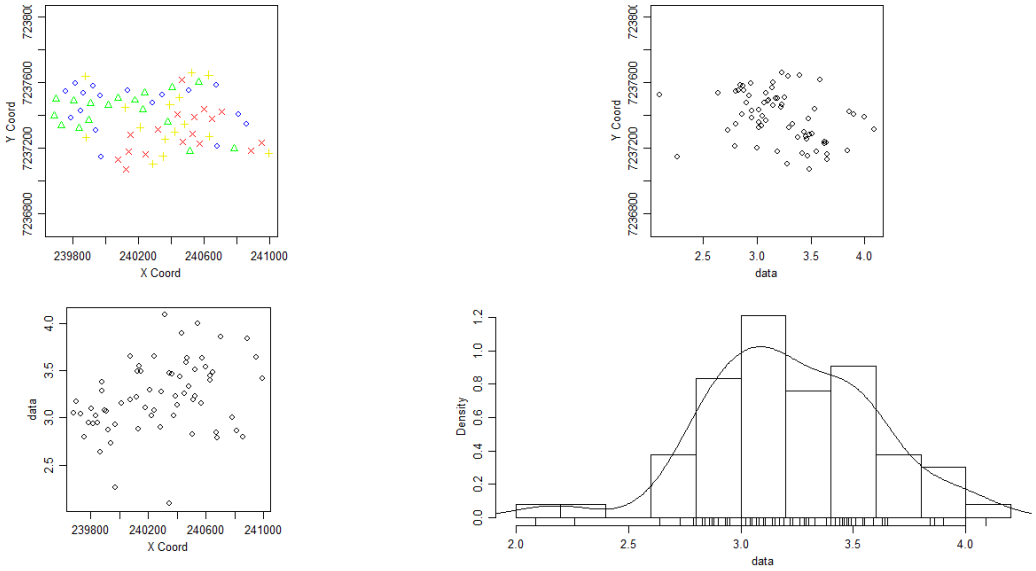

Figura 15: Mapa de estudo espacial para dados experimentais.

O gráfico acima mostrou a distribuição dos dados espacialmente de acordo com a proximidade, vemos que eles estão distribuídos de forma aleatória e não foi possível verificar nenhum sinal de anisotropia.

A segunda e terceira imagens são capazes de mostrar se há alguma tendência nos dados nas direções dos eixos cartesianos e verificamos que a hipótese de isotropia foi fortalecida.

A quarta imagem tratou-se de um histograma aparentemente leptocúrtico com alguns dados acima da curva de probabilidade.

Apresenta-se abaixo o box-plot da amostra experimental.

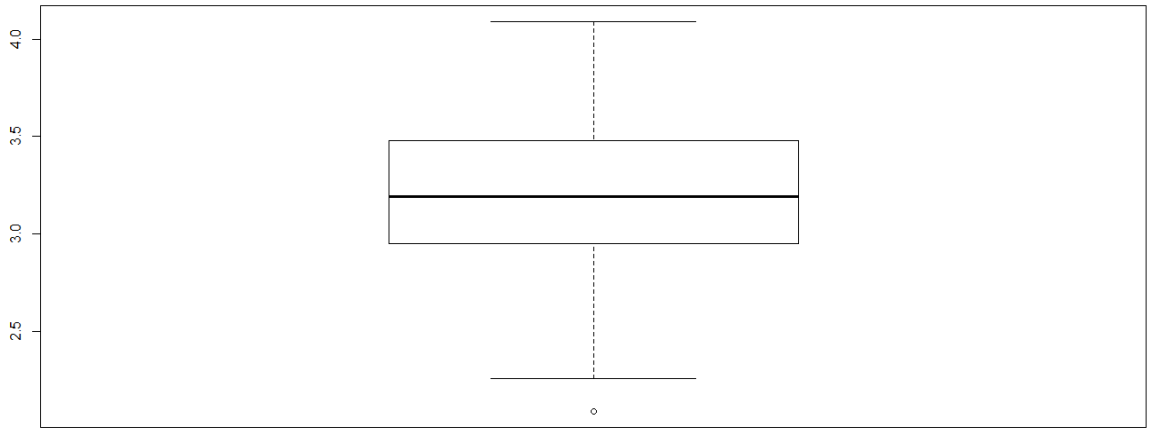

Figura 16: Box-plot dados experimentais.

Analisando o gráfico acima observou-se tratar de dados homogêneos aparentando estarem bem distribuídos com concentração de 50% deles em torno da média além da distância entre o mínimo não discrepante e o primeiro quartil ser próxima da distância do máximo até o segundo quartil, o que o torna ideal, a não ser pelo ponto discrepante que se apresentou também como o mínimo deste conjunto.

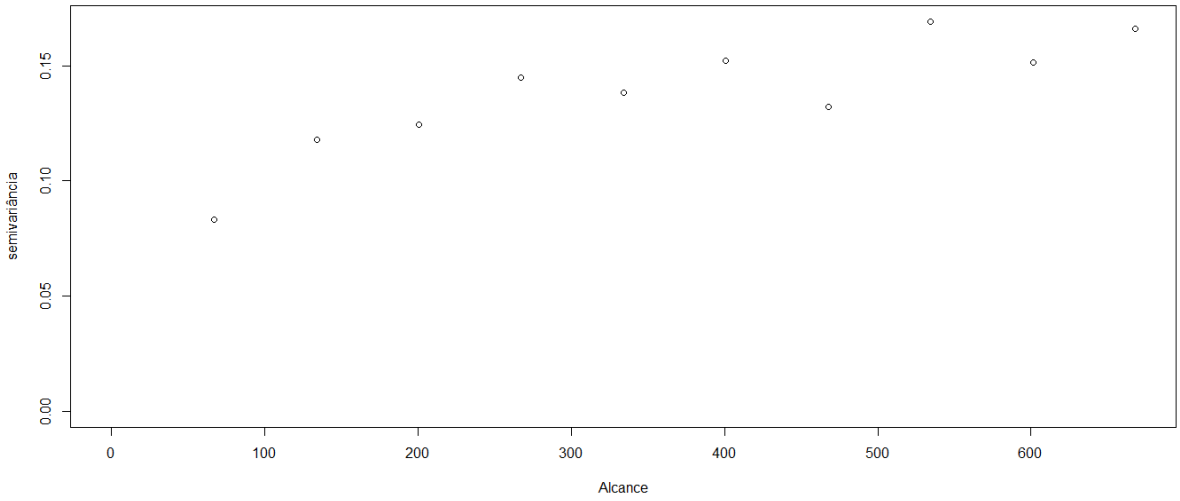

Figura 17: Semivariograma dos dados experimentais.

4.2.1 VALIDAÇÃO CRUZADA DADOS EXPERIMENTAIS

Tabela 5: Validação cruzada dados experimentais

Modelo	φ_1	φ_2	$(g(\varphi_3))$	EM	ER	S_{EM}	S_{ER}	EA	ML
Exponencial	0.0424	0.0937	261.7293	-0.0017	-0.0024	0.3447	1.0173	16.5131	-24,04
Esférico	0.0	0.1398	40	**3,366e-17**	**8,709e-17**	0,3824	1,0171	19,7944	-28,71
Gaussiano	0.0871	0.0488	266.1338	-0,0018	0.0026	**0.34379**	**1,0153**	**16.0445**	**-23,90**

φ_1: efeito pepita; φ_2: contribuição; $(g(\varphi_3))$: alcance; EM: erro médio; ER: erro reduzido; S_{EM}: desvio padrão do erro médio; S_{ER}: desvio padrão do erro reduzido e EA: erro absoluto.

Analisando os valores dos parâmetros foi possível perceber que o modelo gaussiano foi o que melhor representou o conjunto de dados experimentais, fornecendo melhores valores de $S_{EM} : 0,34379$, $S_{ER} : 1,0153$, $EA : 16,0445$ e $ML : -23,90$ comparados com os parâmetros dos demais modelos. Pode-se ver o ajuste do modelo gaussiano ao semivariograma com relação aos outros modelos no gráfico abaixo:

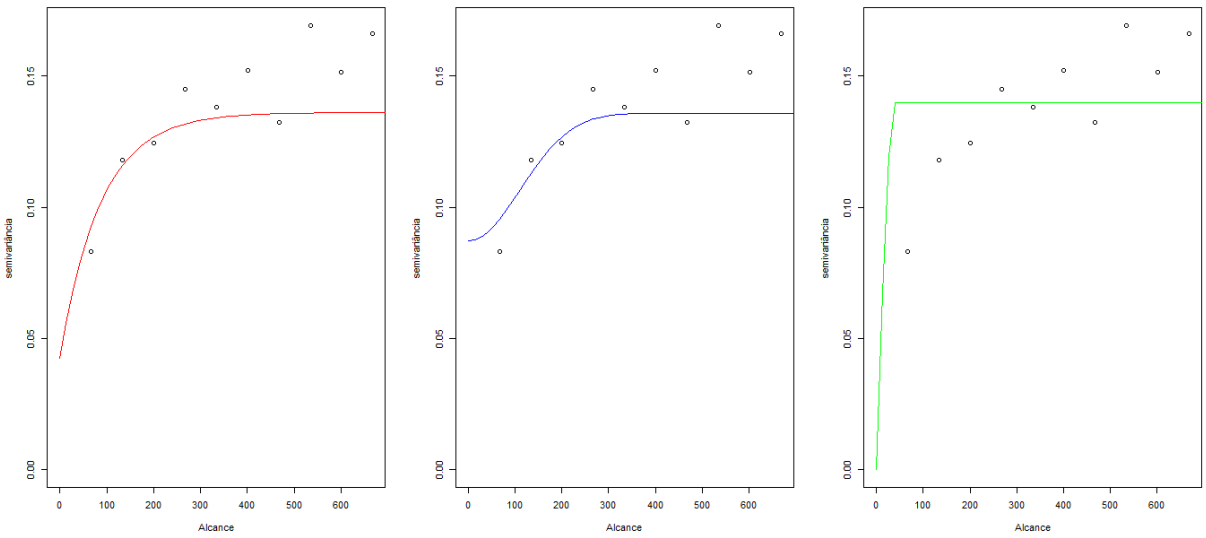

Figura 18: Modelos exponencial, gaussiano e esférico ajustados respectivamente ao semivariograma da amostra experimental.

5 CONCLUSÕES

Tendo em vista os dois resultados positivos, concluímos que o método de validação cruzada, neste trabalho, trata-se de um bom estimador de modelos e pode ser utilizado em um conjunto de dados experimentais, onde podemos confiar no resultado que o método indicar.

Aplicando a técnica na amostra experimental, obtivemos que o modelo gaussiano melhor ajustou os pontos amostrais e podemos então utilizar o modelo gaussino, com a certeza de que é de fato o modelo que melhor representa a amostra, para interpolar utilizando-se da krigagem.

REFERÊNCIAS

ANDRIOTTI, J. L. S. Introdução à geoestatìstica. **Acta Geologica Leopoldensia**, v. 11, 1989.

ASSUMPÇÃO, R. A. Influência local em um modelo espacial linear da produtividade da soja utilizando distribuição t-student naimara v. do prado, miguel a. uribe-opazo 2, manuel galea 3. SciELO Brasil, 2010.

BIVAND, R.; GEBHARDT, A. Implementing functions for spatial statistical analysis using the language. **Journal of Geographical Systems**, Springer, v. 2, n. 3, p. 307–317, 2000.

CRESSIE, N. Geostatistics. **The American Statistician**, Taylor & Francis Group, v. 43, n. 4, p. 197–202, 1989.

HUBER, W.; GENTLEMAN, R. matchprobes: a bioconductor package for the sequence-matching of microarray probe elements. **Bioinformatics**, Oxford Univ Press, v. 20, n. 10, p. 1651–1652, 2004.

ISAAKS, E. H.; SRIVASTAVA, R. M. **An introduction to applied geostatistics**. Ames, USA: Oxford university press, 1989. 561 p.

JR, P. J. R.; DIGGLE, P. J. geor: A package for geostatistical analysis. **R news**, London, v. 1, n. 2, p. 14–18, 2001.

MARDIA, K. V.; MARSHALL, R. Maximum likelihood estimation of models for residual covariance in spatial regression. **Biometrika**, Biometrika Trust, v. 71, n. 1, p. 135–146, 1984.

MATHERON, G. Principles of geostatistics. **Economic geology**, Society of Economic Geologists, v. 58, n. 8, p. 1246–1266, 1963.

SANTANA, R. A. Avaliação de técnicas geoestatísticas no inventário de povoamentos de tectona grandis lf. Universidade Federal de Viçosa, 2011.

VIEIRA, S. R. Geoestatística aplicada à agricultura de precisão. **GIS Brasil**, v. 98, 2000.