ANDRÉ MENDES

DETERMINAÇÃO DO TAMANHO DE AMOSTRA PARA A GEOESTATÍSTICA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae*.

Orientador: Gerson Rodrigues dos Santos

Coorientador: Wellington Donizete Guimarães

VIÇOSA – MINAS GERAIS
2020

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

T
M538d
2020

Mendes, André, 1977-
Determinação do tamanho de amostra para a geoestatística / André Mendes. – Viçosa, MG, 2020.
94 f. : il. (algumas color.) ; 29 cm.

Inclui apêndice.
Orientador: Gerson Rodrigues dos Santos.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Geologia - Métodos estatísticos. 2. Amostragem. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22 ed. 551.028

ANDRÉ MENDES

DETERMINAÇÃO DO TAMANHO DE AMOSTRA PARA A GEOESTATÍSTICA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae*.

APROVADA: 21 de fevereiro de 2020.

Assentimento:

André Mendes
Autor

Gérson Rodrigues dos Santos
Orientador

À mulher da minha vida, Regiane, pelo apoio incondicional em todos os momentos, principalmente nos de incerteza. Sem você, nenhuma conquista valeria a pena!

À minha filha, Maria Eduarda, amor incondicional, que me faz acreditar que os sonhos podem ser realizados.

Aos meus pais, Marlene e Evandro, que dignamente me fizeram entender a importância de uma família e me ensinaram o caminhado da honestidade e da persistência.

AGRADECIMENTOS

A realização desta Tese de Doutorado contou com importantes apoios e incentivos sem os quais não teria se tornado uma realidade e aos quais estarei eternamente grato.

A Deus, acima de tudo, que me concedeu forças, saúde, sabedoria e determinação ao longo do desenvolvimento deste trabalho, bem como durante toda a minha vida.

Aos familiares, especialmente aqueles que me incentivaram durante esta etapa da minha vida.

À minha amada esposa, Regiane, amiga e companheira, que vivenciou cada etapa do meu Doutorado. Sem você eu não teria tido forças para concluir este nosso sonho. Juntos, eternamente.

À minha filha amada, Maria Eduarda, amor incondicional, razão de todo o meu esforço e comprometimento, na superação de todos os obstáculos encontrados no decorrer desta jornada.

Aos meus pais, Marlene e Evandro, pelos exemplos, ensinamentos, palavras de incentivo e principalmente por acreditarem na minha capacidade.

Ao professor Doutor Gerson Rodrigues dos Santos, pela sua orientação, apoio, incentivo, disponibilidade (inclusive nos finais de semana e pelo WhatsApp), pelo saber que transmitiu, pelas opiniões, sugestões e críticas, total colaboração no solucionar de dúvidas e problemas que foram surgindo ao longo da realização deste trabalho, além das incansáveis horas de contribuição. A você "*brother*", a minha eterna gratidão.

Aos professores do Departamento de Estatística da UFV pelos ensinamentos e incentivos.

Aos servidores do Departamento de Estatística da UFV, Anita e Júnior, pela disponibilidade em esclarecer e auxiliar nas atividades burocráticas, bem como solucionar pendências.

Aos amigos (as) que tive o privilégio de conquistar ao longo do Doutorado. Vocês foram imprescindíveis durante esta caminhada, quer seja tirando dúvidas, fazendo análises estatísticas, resolvendo exercícios, oferecendo e pegando caronas ou ainda virando noites preparando seminários e estudando para as provas.

Aos membros da banca que fizeram observações relevantes visando a melhoria deste trabalho.

Ao Instituto Federal de Minas Gerais – IFMG, Campus Avançado Ponte Nova, e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo investimento.

RESUMO

MENDES, André, D.Sc., Universidade Federal de Viçosa, fevereiro de 2020. **Determinação do tamanho de amostra para a Geoestatística.** Orientador: Gerson Rodrigues dos Santos. Coorientador: Wellington Donizete Guimarães.

A estimativa do tamanho da amostra na geoestatística é de grande importância para o planejamento e tomada de decisão, especialmente quando se objetiva a reconstrução total da população estudada. Por este motivo, muitos trabalhos sobre o tamanho da amostra geoestatística surgem com este propósito. Assim, o objetivo geral deste trabalho é utilizar a geoestatística associada ao teorema da taxa Nyquist para determinar um tamanho de amostra ideal quando se utiliza uma grade regular quadrática, na qual o modelo de dependência espacial ajustado é o gaussiano, identificando especificamente mudanças no tamanho ideal da amostra na presença de *outliers*. Dois conjuntos de dados altimétricos (Viçosa-MG, Brasil e Treynor-Iowa, EUA) foram analisados e o tamanho amostral ideal para ambos os conjuntos foi obtido. Posteriormente, os *outliers* foram removidos do conjunto de dados norte-americano e comparados os tamanhos de amostra ideais obtidos anteriormente. Além disso, utilizando os *softwares* R e ArcGIS, as estimativas dos parâmetros do modelo gaussiano, da média e da variância dos resíduos, provenientes da validação cruzada, foram comparadas através da construção de intervalos de confiança. Com o presente estudo concluiu-se que: (i) a distância máxima entre os pontos da grade regular quadrática é de aproximadamente 30% do alcance prático observado no semivariograma da primeira amostragem experimental; (ii) o tamanho amostral ideal obtido na presença de *outliers* é praticamente o dobro do tamanho de amostra ideal na ausência de *outliers*; (iii) o *software* R é o mais adequado na comparação das estimativas da média e da variância dos resíduos pois apresentou uma menor variabilidade (menores amplitudes dos intervalos de confiança construídos).

Palavras-chave: Tamanho de amostra. Taxa Nyquist. Geoestatística. *Outliers*.

ABSTRACT

MENDES, André, D.Sc., Universidade Federal de Viçosa, February, 2020. **Determination of sample size for a Geostatistics.** Adviser: Gerson Rodrigues dos Santos. Co-Adviser: Wellington Donizete Guimarães.

The estimation of sample size in Geostatistics has a great importance for planning and decision-making, specially when aims at the total reconstruction of the studied population. For this reason, many papers on the size of Geostatistics sampling are emerging. Thus, the general objective of this work is to use Geostatistics associated with the Nyquist Rate Theorem to determine an ideal sampling size when using a regular quadratic grid, in which the spatial dependence model is Gaussian, thus specifically identifying changes in ideal sampling size in the presence of outliers. Two sets of altimetry data (Viçosa-MG, Brazil and Treynor-Iowa, USA) were analyzed and the ideal sampling size for both sets was obtained. Subsequently, the outliers were removed from the North American dataset and compared to the ideal sampling sizes previusly obtained. In addition, using the R and ArcGIS softwares, the estimates of the Gaussian model parameters, the mean and the variance of the residues obtained from the cross validation were compared through the construction of confidence intervals. With the present study it was concluded that: (i) the maximum distance between the points of the regular quadratic grid is approximately 30% of the practical range observed in the semivariogram of the first experimental sample; (ii) the ideal sampling size obtained in the presence of outliers is practically double the ideal sampling size in the absence of outliers; (iii) the software R is the most adequate in the comparison of the mean and the variance estimates of the residues since it presented a lower variability (smaller amplitudes of constructed confidence intervals).

Keywords: Sample size. Nyquist rate. Geostatistics. *Outliers*.

LISTA DE ILUSTRAÇÕES

INTRODUÇÃO GERAL

CAPÍTULO 1

CAPÍTULO 2

CAPÍTULO 3

LISTA DE TABELAS

SUMÁRIO

INTRODUÇÃO GERAL

Estatística Espacial é o ramo da Estatística que permite estudar a ocorrência de eventos, ao considerar sua localização espacial (SANTOS e SOUZA, 2007). Por meio dela, além de identificar, localizar e visualizar os eventos que se materializam no espaço, é possível modelar o fenômeno, ao incorporar a estrutura de distribuição espacial ou identificar padrões. Para Landim (1998), ao utilizar a Estatística Clássica para representar propriedades de valores amostrais distribuídos espacialmente, assume-se que essas realizações sejam uma variável aleatória, e, portanto, todos os valores amostrais apresentam a mesma probabilidade de serem escolhidos.

A não comprovação das pressuposições da Estatística Clássica ocorridas frequentemente em diversas áreas do saber científico reforça a necessidade da utilização de metodologias mais informativas, devido a incorporação da dimensão espacial. Na análise dessas metodologias são essenciais, pelo menos, as informações sobre a localização, os atributos e o pressuposto de dependência espacial dos dados. A utilização de metodologias da Estatística Espacial, como por exemplo, padrão de pontos, dados de área e geoestatística são recomendadas na literatura científica, como é o caso de Bailey e Gatrell (1995), Assunção (2001), Santos e Souza (2007), Bivand et al. (2013).

Desde o início do século XX busca-se obter padrões de dependência espacial em amostras não aleatórias. Inicialmente a geoestatística foi utilizada por Daniel Gerhardus Krige (Krige, 1951), professor e engenheiro de minas sul-africano, na avaliação de jazidas de ouro na África do Sul. Krige concluiu que as variâncias dos dados de concentração de ouro só faziam sentido ao se considerar a distância entre os pontos amostrados (VIEIRA, 2000).

Segundo Yamamoto e Landim (2013), a partir do trabalho de Krige, o professor George Matheron apresentou na década de 1960 alguns trabalhos que formalizaram a teoria das variáveis regionalizadas, objetivo de estudo da geoestatística e foi considerado, portanto, o criador dessa ciência.

Vale ressaltar que uma variável regionalizada é uma função espacial numérica, que varia de um local para o outro, com continuidade aparente e cujos valores são relacionados com a posição espacial que ocupam.

Uma variável regionalizada é estacionária de segunda ordem se o primeiro e o segundo momento estatístico da variável aleatória $Z(x + \mathbf{h})$ forem os mesmos para qualquer vetor distância **h** (VIEIRA, 2000). Assim, $Z(x)$ é uma função aleatória estacionária de segunda ordem quando satisfeitas as Equações 1 e 2:

$$E[Z(x)] = \mu^2 \quad (1)$$

$$C(x, x + \mathbf{h}) = E[Z(x)Z(x + \mathbf{h})] - \mu^2 \quad (2)$$

A Equação (1) indica que o valor esperado $E[Z(x)]$ existe e independe da posição do ponto x. Já a Equação (2) indica que para cada par de variáveis aleatórias, $Z(x)$, $Z(x + \mathbf{h})$, a função de covariância $C(x, x + \mathbf{h})$ existe e depende apenas do vetor distância **h**.

O semivariograma, primeiro passo para uma análise geoestatística, é capaz de descrever, tanto qualitativamente quanto quantitativamente a variação espacial, além de ser primordial na determinação do preditor geoestatístico. Segundo Isaaks e Srivastava (1989), o semivariograma experimental ou empírico é um gráfico no sistema cartesiano, no qual o eixo das abscissas representa o vetor distância **h** entre dois pontos amostrados e o eixo das ordenadas a função de semivariância. A função de semivariância é definida como a metade da esperança matemática do quadrado da diferença entre as realizações de duas variáveis aleatórias localizadas no espaço, separadas por um vetor distância **h**, dada pela Equação (3):

$$\gamma(\mathbf{h}) = \frac{1}{2} E\left\{[Z(x) - Z(x+\mathbf{h})]^2\right\} \quad (3)$$

em que $Z(x)$ e $Z(x + \mathbf{h})$ são os valores da variável regionalizada Z em pontos amostrados separados por um vetor distância **h**.

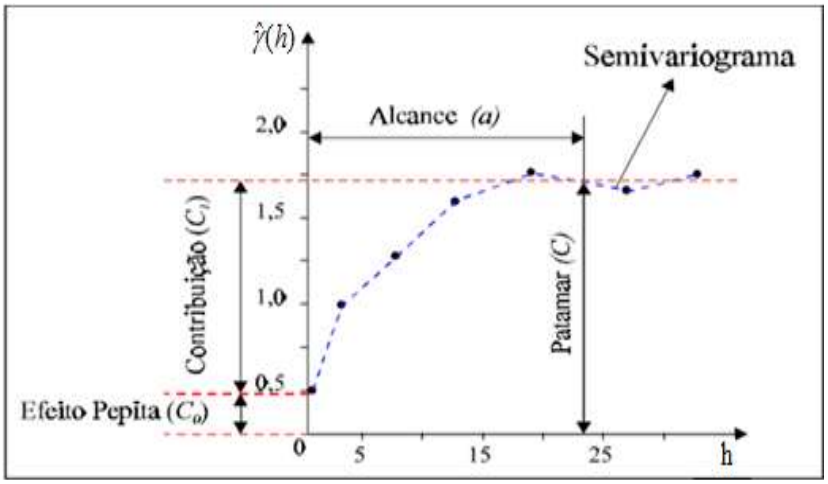
Segundo Vieira (2000), o estimador de semivariância mais utilizado é o baseado no método de momentos, proposto por Matheron em 1963, dado pela Equação (4):

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + \mathbf{h})]^2 \quad (4)$$

em que $N(h)$ é o número de pares de valores observados separados entre si por uma distância **h**.

O semivariograma possui parâmetros que auxiliam na modelagem da estrutura da dependência espacial, conforme Figura 1.

Figura 1 – Exemplo de um semivariograma experimental e seus parâmetros.



Fonte: Câmara e Medeiros (1998)

O conceito de cada um desses parâmetros é apresentado conforme (WACKERNAGEL, 2003; YAMAMOTO e LANDIM, 2013):

- Efeito Pepita (C_0): valor onde a função $\gamma(\mathbf{h})$ intercepta o eixo das ordenadas, o qual representa a descontinuidade na origem e está relacionado com $h = 0$.
- Alcance (a): distância dentro da qual os valores amostrais apresentam-se correlacionados espacialmente e corresponde ao raio de dependência espacial. Para distâncias maiores do que o alcance a semivariância não se altera.
- Patamar (C): é o valor da semivariância no qual o semivariograma atinge o alcance e representa o menor valor para o qual não existe mais dependência espacial.
- Contribuição (C_1): é a diferença entre os valores do patamar (C) e o efeito pepita (C_0). É a parte da variabilidade dos dados que é explicada pelo semivariograma.

Diferentes modelos teóricos com patar podem ser ajustados a um semivariograma experimental, sendo os mais utilizados os modelos esférico, exponencial e gaussiano.

O modelo esférico é o mais utilizado (ISAAKS e SROVASTAVA, 1989). Possui comportamento linear para valores de h pequenos e é representado pela Equação (5):

$$\gamma(\mathrm{h}) = \begin{cases} 0\,; & h = 0 \\ C_0 + C_1\left[\dfrac{3}{2}\left(\dfrac{\mathrm{h}}{a}\right) - \dfrac{1}{2}\left(\dfrac{h}{a}\right)^3\right] & ; 0 < \mathrm{h} < a \\ C_0 + C_1\,; & h \geq a \end{cases} \tag{5}$$

Já o modelo exponencial atinge o patamar de forma assintótica, sendo o alcance definido como a distância em que o modelo apresenta valor igual a 95% do patamar. Além disso, pressupõe continuidade pequena em pequenos intervalos (ISAAKS e SRIVASTAVA, 1989). O modelo é representado pela Equação (6):

$$\gamma(h) = \begin{cases} 0\,; & h = 0 \\ C_0 + C_1\left[1 - \exp\left(-\dfrac{3h}{a}\right)\right]; & 0 < h < a \\ C_0 + C_1\,; & h > a \end{cases} \quad (6)$$

O modelo gaussiano, além do comportamento parabólico na origem tem como particularidade a presença de ponto de inflexão, sendo que o patamar é atingido de forma assintótica. O modelo é representado pela Equação (7):

$$\gamma(\mathrm{h}) = \begin{cases} 0\,; & h = 0 \\ C_0 + C_1\left[1 - \exp\left(-\dfrac{3h^2}{a^2}\right)\right]; & 0 < h < a \\ C_0 + C_1\,; & h > a \end{cases} \quad (7)$$

Diferentes modelos teóricos, a princípio, podem ser ajustados a um semivariograma experimental. Entretanto, espera-se que haja diferença na qualidade do ajuste destes modelos.

A partir da escolha do modelo do semivariograma, pode-se proceder à interpolação geoestatística, conhecida como krigagem. Esse método permite interpolar valores em qualquer posição da área em estudo, sem tendência e com variância mínima, desde que seja conhecido o semivariograma teórico da variável em estudo e que haja dependência espacial entre as amostras (VIEIRA, 2000).

A krigagem, como método de predição, é preferível por apresentar predições não tendenciosas e variância mínima associada ao valor predito (YAMAMOTO e LANDIM, 2013) e pressupõe que a variável em estudo possua estacionariedade de segunda ordem (CRESSIE, 1993).

A krigagem leva em consideração o número de amostras, as posições e a distância entre as mesmas, a área a ser estimada e a continuidade espacial da variável estudada por meio do semivariograma, primeiro passo para uma análise geoestatística.

Diferentemente dos outros métodos de estimação ponderados, a Krigagem leva em consideração a dependência estatística espacial existente entre os valores dos pontos (amostrados e não amostrados), bem como a distância entre tais pontos. Estatisticamente, a Krigagem é considerada o melhor método de estimação, pois entre os métodos de estimação,

é o método que produz estimativas com a menor variância do erro de estimação (ISAAKS e SRIVASTAVA, 1989).

Entre os tipos de krigagem existentes, como a krigagem simples, a krigagem ordinária, a krigagem da média e a krigagem universal (YAMAMOTO e LANDIM, 2013), destaca-se a simples, pois de acordo com Santos et al. (2011) ela é um preditor mais preciso do que as outras krigagens.

Segundo Yamamoto e Landim (2013), a krigagem simples assume que a média da variável regionalizada é conhecida. Assim, o valor predito não viesado da variável aleatória Z no ponto x_0 não amostrado é dado pela Equação (8):

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) + \left\{1 - \sum_{i=1}^{n} \lambda_i\right\} \mu \tag{8}$$

em que λ_i representa o peso de cada ponto amostrado x_i, n é o número de pontos amostrados, $Z(x_i)$ é o valor da variável aleatória Z no ponto x_i e μ é a média dos valores da variável Z.

Um método para quantificar o erro da predição nos locais que foram amostrados é a valiação cruzada. De acordo com Vieira (2000) esse erro quantifica a incerteza sobre as hipóteses assumidas para a realização da krigagem e sobre o ajuste dos parâmetros de cada modelo.

A validação cruzada consiste em eliminar temporariamente uma das *n* observações e utilizar as demais $n - 1$ observações no ajuste de um modelo, sendo este utilizado para predizer predizer aquela observação que havia sido retirada por meio da krigagem (WACKERNAGEL, 2003). A realização desse procedimento deve ser feita para cada um dos n valores observados, obtendo-se, ao final, o erro absoluto de cada um deles, definido como a diferença entre o valor predito e o observado para cada ponto amostrado (VIEIRA, 2000).

Algumas estatísticas referentes ao erro de predição que utilizam a validação cruzada auxiliam na verificação de qual modelo melhor prediz nos locais dos pontos amostrados. As mais comuns são: o erro médio (ME), raiz quadrada do erro médio (RMSE) e erro padrão médio (ASE). Segundo Isaaks e Srivastava (1989), o modelo teórico que apresentar valores mais satisfatórios para essas estatísticas; ME: próximo de zero, RMSE e ASE: quanto menor, melhor, deve ser escolhido para representar a dependência espacial do fenômeno em estudo.

Conhecida como variância do erro por Yamamoto e Landim (2013), variância mínima das predições por Vieira (2000), a variância de krigagem é uma medida que pode ser utilizada

na comparação de krigagens quanto à incerteza na predição, escolhendo aquela que apresentar a menor variância média de krigagem (SANTOS et al., 2011).

Segundo Wackernagel (2003), ao ser realizada uma amostragem irregular a variância de krigagem é ainda mais importante, pois fornece uma análise da variação da precisão das predições proporcionada pela distribuição irregular dos pontos amostrados. De acordo com Webster e Oliver (2007), a variância de krigagem para a krigagem simples é dada pela Equação (9):

$$\sigma^2(x_0) = C(0) - \sum_{i=1}^{n} \lambda_i \, C(x_i, x_0) \tag{9}$$

em que $C(0)$ é a variância da variável regionalizada Z, λ_i o peso de cada ponto amostrado x_i e $C(x_i, x_0)$ a função de covariância entre o i-ésimo ponto e aquele que se deseja predizer.

Note que, para **h** = 0, tem-se da Equação (2) que $C(0) = E[(Z(x))^2] - \mu^2 = Var[Z(x)]$. Assim, a função de covariância $C(\mathbf{h})$ pode ser obtida a partir da função de semivariância $\gamma(\mathbf{h})$ sendo dada por $\sigma(h) = C(0) - \gamma(\mathbf{h})$. Portanto, a covariância $C(\mathbf{h})$ e a semivariância $\gamma(\mathbf{h})$ são funções que podem ser utilizadas para caracterizar a dependência espacial.

Segundo Modis e Papodysseus (2006), os estimadores teóricos, altamente sofisticados em muitos casos, são influenciados pela densidade de amostragem. Para estes autores, o parâmetro principal que afeta a precisão da estimativa é o tamanho da grade de amostragem. Na prática, para determinar um tamanho da grade de amostragem ideal utiliza-se a variância dos erros como critério de eficiência (BROOKER, 1991; DAVID, 1977; DOWD e MILTON, 1987). Assim, em um experimento contendo diferentes tamanhos amostrais, obtém-se a variância da estimativa média dos erros em função da densidade de amostragem. O tamanho ideal da grade de amostragem é o limite para o qual uma diminuição adicional não oferece melhorias na variação da estimativa.

Modis e Papaodysseus (2006) propuseram um estudo envolvendo amostras de mineração e calcularam o tamanho teórico ideal da grade amostral suficiente para a representação precisa da área estudada. A abordagem matemática utilizada por estes autores é baseada na teoria da informação, apresentada a seguir.

Originalmente desenvolvido para sinais elétricos (WHITTAKER, 1915; SHANNON, 1949), a teoria da informação afirma que uma forma de onda aleatória pode ser reconstruída por suas amostras se a densidade de amostragem for maior ou igual a um valor crítico, dependendo das características do sinal (LLOYD, 1959). A especificação dessa taxa, denominada taxa Nyquist, é feita no domínio da frequência. A transformação para o domínio

da frequência é obtida tomando a transformada de Fourier da função aleatória, que é conhecida como o espectro de potência da função de correlação.

Uma condição necessária e suficiente para a existência de uma taxa Nyquist é que a função aleatória seja de faixa limitada, o que significa que a transformada de Fourier da função aleatória possui valores nulos fora da faixa definida por uma frequência mínima e máxima. Este teorema se estende diretamente a duas ou mais dimensões (PETERSEN e MIDDLETON, 1962), geralmente a uma amostragem de grade regular. A utilização desse teorema possibilita obter um tamanho de amostra ideal para amostras georreferenciadas que utilizam grades regulares.

De acordo com Modis e Papaodysseus (2006) a utilização de modelos de campo aleatório estacionário com função semivariograma isotrópico em região de influência finita ou assintótica deve ser considerada para:

(i) examinar se as funções de correlação são de faixa limitada;

(ii) verificar se a taxa Nyquist corresponde a um tamanho adequado de espaçamento entre as amostras.

A restrição aos modelos estacionários é definida, sem perda de generalidade, uma vez que os modelos de estimação usuais filtram a tendência pela aplicação de combinações lineares (MATHERON, 1971).

Modis e Papaodysseus (2006) apresentam todo o desenvolvimento teórico necessário para obtenção da densidade de amostragem ideal, para modelos de semivariograma exponencial e esférico. Como resultado prático, obteve-se um tamanho amostral "T" no máximo igual à metade do alcance prático de influência oriundo da modelagem do semivariograma. Posteriormente, esses resultados teóricos foram validados em uma aplicação prática utilizando amostras de minério de cobre. Para um tratamento detalhado desse procedimento, veja Modis e Papodysseus (2006).

De forma geral, o objetivo deste trabalho é estudar alguns temas da geoestatística que necessitam de estudos mais aprofundados, quer sejam teóricos ou práticos.

Especificamente, pretende-se atingir esse objetivo geral através dos específicos:

a) Utilizar a geoestatística, associada à teoria da informação de sinais eletricos (teorema da taxa Nyquist) para determinar um tamanho de amostra ideal (com e sem *outliers*) a partir de uma grade regular quadrática na qual o modelo de dependência espacial ajustado é o gaussiano.

b) Identificar qual o *software* (ArcGIS ou R) mais adequado ao comparar, para a mesma base de dados, as estimativas dos parâmetros do modelo gaussiano e as estimativas da média e da variância dos resíduos provenientes da validação cruzada, por meio da construção de intervalos de confiança, para a mesma base de dados.

Este trabalho está estruturado em três capítulos, organizados da seguinte maneira: No Capítulo 1 apresenta-se a geoestatística associada à teoria da informação de sinais elétricos, ao utilizar o teorema da taxa Nyquist para a determinação de um tamanho de amostra ideal quando o modelo de dependência espacial ajustado a uma grade regular quadrático for o gaussiano. No Capítulo 2 apresenta-se o teorema da taxa Nyquist para a determinação, na presença de *outliers*, de um tamanho de amostra ideal quando o modelo de dependência espacial ajustado a uma grade regular quadrática for o gaussiano. Para finalizar, no Capítulo 3 apresenta-se um estudo comparativo entre os *softwares* ArcGis e R para identificar qual é o mais adequado ao comparar as estimativas dos parâmetros do modelo variográfico gaussiano e as estimativas da média e variância dos resíduos provenientes da validação cruzada, a partir da construção de intervalos de confiança para dados altimétricos.

REFERÊNCIAS BIBLIOGRÁFICAS

ASSUNÇÃO, R. M. Estatística Espacial com Aplicações em Epidemiologia, Economia e Sociologia. **7ª Escola de Modelos de Regressão**, São Carlos, SP. 2001.

BAILEY, T.; GATTREL, A. **Spatial Data Analysis by Example.** London, Longman, 1995.

BIVAND, R. S.; PEBESMA, E. J.; GOMEZ-RUBIO, V. **Spatial data import and export in applied spatial data analysis with R**. Springer, New York, p. 83-125, 2013.

BROOKER, P. I. **A geostatistical primer**. Singapore: World Scientific, 1991. 95 p.

CÂMARA, G.; MEDEIROS, J. S. **Geoprocessamento para projetos ambientais**. V. 1, Online Book, 1998. São José dos Campos, Brasil. INPE.

CRESSIE, N. A. C. **Statistics for Spatial Data.** Wiley Series in Probability and Mathematical Statistics: applied probability and statistics. New York: John Wiley & Sons, Inc. 1993. 900 p.

DAVID, M. **Geostatistical ore reserve estimation.** Developments in Geomathematics 2. Elsevier, Amsterdam, 1977. 364 p.

DOWD, P. A.; MILTON, D. W. Geoestatistical Estimation of a Section of the Perseverance Nickel Deposit. In: MATHERON, G.; ARMSTRONG, M., eds. **Geoestatistical case studies.** Reidel, Dordrecht, p. 39-67, 1987.

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

KRIGE, D. G. A statistical approach to some mine evaluation problems on the Witwatersrand. **Johanesburg Chemistry Metallurgy Mining society South African**, Johanesburg, v. 52, n. 6, p. 119-139, 1951.

LANDIM, P. M. P. **Análise estatística de dados geológicos.** São Paulo: Editora UNESP, 1998. 226 p.

LLOYD, S. P. A Sampling Theorem for Stationary (Wide Sense) Stochastic Processses. **Transactions of the American Mathematical Society**, v. 92, n. 1, p. 1-12, 1959.

MATHERON, G. **The theory of regionalized variables and its applications**. Les Cahiers du Centre de Morphologie Mathématique, Fascicule 5. Ecole Nationale Superieure des Mines de Paris, Fontainebleau, 1971. 212 p.

MODIS, K.; PAPAODYSSEUS, K. Theoretical Estimation of the Critical Sampling Size for Homogeneous Ore Bodies with Small Nugget Effect. **Mathematic Geology**, v. 38, n. 8, p. 489–501, 2006.

PETERSEN, D. P.; MIDDLETON, D. Sampling and Reconstruction of Wave-Number-Limited Functions in N-dimensional Euclidean Spaces. **Information and Control**, v. 5, p. 279–323, 1962.

SANTOS, G. R., OLIVEIRA, M. S., LOUZADA, J. M., SANTOS, A. M. R. T. Krigagem Simples versus Krigagem Universal: qual o preditor mais preciso? **Revista Energia na Agricultura**, Botucatu, v. 26, n. 2, p. 49-55, 2011.

SANTOS, S. M.; SOUZA, W. V. (Org.). **Introdução à Estatística Espacial para a Saúde Pública.** Brasília: Ministério da saúde, 2007. 122 p.

SHANNON, C. E. **The Mathematical Theory of Communication**. University of Illinois Press, 1949.

VIEIRA, C. A. O. **Accuracy of remotely sensing classification of agricultural: a comparative study.** 2000. 352f. Thesis (Degree of Doctor of Philosophy) – University of Nottingham, Nottingham.

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

WACKERNAGEL, H. **Multivariate geostatistics: an introduction with applications.** Springer, Berlin Heidelberg New York, 2003. 387 p.

WEBSTER, R.; OLIVER, M. A. **Geostatistics for Environmental Scientists.** 2 ed. Chichester: John Wiley & Sons, 2007. 271 p.

WHITTAKER, E. T. On the Functions wich are represented by the Expansions of the Interpolation- Theory. **Proceedings of the Royal Society,** Edinburgh**,** v. 35, p. 181-194, 1915.

YAMAMOTO, J.; LANDIM, P. **Geoestatística: Conceitos e Aplicações.** Oficina de Textos: São Paulo, 2013. 216 p.

CAPÍTULO 1 - ESTIMAÇÃO TEÓRICA DO TAMANHO AMOSTRAL NA GEOESTATÍSTICA USANDO UM MODELO DE SEMIVARIOGRAMA GAUSSIANO

RESUMO

Na Geoestatística Clássica ou Geoestatística baseada em design, existe uma grande necessidade de pesquisas que criem e/ou investiguem métodos de amostragem de dados geoespaciais. Além da complexidade do assunto, alguns trabalhos apresentam soluções que utilizam mecanismos teóricos e práticos de diferentes áreas do conhecimento científico que atendem demandas específicas de pesquisadores da área. O objetivo deste artigo é utilizar a teoria da informação de sinais elétricos, principalmente considerando o teorema da taxa Nyquist, a fim de determinar um tamanho ideal para amostras georreferenciadas que usam grade quadrática regular, no qual o modelo de dependência espacial é o gaussiano. O que se deseja alcançar teoricamente é uma densidade de amostragem necessária para a reconstrução de mapas populacionais de variáveis nas quais as condições de regularidade necessárias em geoestatística foram atendidas, a saber: estacionariedade de 1ª e 2ª ordem e/ou estacionariedade do semivariograma, ausência de *outliers* e tendências, e semivariograma isotrópico. Como resultado, pode-se afirmar que a distância máxima entre os pontos da grade regular quadrática é de aproximadamente 30% do alcance prático observado no semivariograma da primeira amostragem experimental.

Palavras-chave: Geoestatística. Amostragem. Taxa Nyquist.

ABSTRACT

In Classical Geostatistic or Design Based Geostatistic, there is a great need for research that creates and/or discusses methods of geospatial data sampling. In addition to the complexity of the subject, some papers present solutions that use theoretical and practical mechanisms of different areas of scientific knowledge that meet the specific demands of researchers in the field. The purpose of this paper is to use the Electric Signal Information Theory, especially considering the Nyquist Rate Theorem, to determine an optimal size for georeferenced samples using regular quadratic grid, in which the spatial dependence model is the Gaussian one. What is theoretically desired to achieve is a sampling density necessary for the reconstruction of population maps of variables in which the regularity conditions required in geostatistics were verified, namely: 1st and 2nd order stationarity and / or stationarity of the variogram, absence of levels and trends, and isotropic variogram. As a result, we can state that for the data set used, the maximum distance between points of the regular quadratic grid is approximately 30% of the practical range observed in the variogram of the first experimental sample.

Key-words: Geostatistics. Sampling. Nyquist Rate.

1 INTRODUÇÃO

Para qualquer análise de dados científicos, é de extrema importância destacar informações de qualidade, denominada representatividade, especialmente quando se trata de amostras (OLIVEIRA et al., 2014). Esse comportamento também é necessário quando você considera informações em que o tempo e o espaço são incorporados (chamadas de regionalização) no processo (YAMAMOTO e LANDIM, 2013). É, precisamente, neste aspecto de regionalização das variáveis de um processo, que a Estatística Espacial ganha entusiastas em todo o mundo, principalmente porque as questões podem ser esclarecidas localmente, isto é, as variáveis são georreferenciadas.

Cressie e Wikle (2011) afirmam que regionalizar as informações sempre foi essencial para o ser humano. Para questões de sobrevivência e/ou conquistas, as civilizações usaram esse mecanismo para melhorar a representação dos sistemas de orientações.

Yamamoto e Landim (2013) afirmam que a Estatística Clássica é importante para muitos estudos, embora tenham que assumir um conjunto de pressuposições que são difíceis ou mesmo impossíveis de serem verificados, como a exigência de amostras aleatórias independentes e o conhecimento da distribuição da probabilidade da variável estudada.

Já a estimação do tamanho de amostra na Estatística Espacial é um dos problemas citado por Webster e Oliver (1992), Santos et al. (2011), Sartori e Zimback (2011) e Souza et al. (2014).

Segundo Ver Hoef (2002) e Clark (2009) apresentam outro problema da Estatística Clássica no que diz respeito ao tamanho de amostra necessário para obtenção da mesma precisão que a Estatística Espacial: aproximadamente 9 vezes maior.

Oliveira et al. (2014) apresentam as subdivisões de estatísticas espaciais. Entre elas, a geoestatística se destaca, como apresentado por Clark (1979), Armstrong (1998), Brooker (1991), Clark e Harper (2000), Druck et al. (2004), Olea (2009), Webster e Oliver (2007), Diggle et al. (2010), Santos et al. (2011) e Ferreira et al. (2013).

Santos et al. (2011) mostram que é vantajoso usar a Geoestatística na modelagem de fenômenos em que não são atentidos os pressupostos da estatística clássica uma vez que a geoestatística usa a vizinhança amostrada com o objetivo de reforçar a percepção de que a estrutura da dependência espacial de um fenômeno contribuiu para a melhoria das previsões, pois existem sem tendência e com variância mínima.

Além disso, Vieira (2000) mostra que é possível ainda conhecer a incerteza envolvida nas previsões, através da variância de krigagem.

Yamamoto e Landim (2013) apresentam uma lista de pesquisas, em diversas áreas, que utilizaram a geoestatística como principal método de análise de dados.

Dada esta tendência, muitos trabalhos sobre o tamanho de amostra geoestatística estão surgindo. Podem-se citar: Brus e Heuvelink (2007), Modis e Papaodysseus (2006), Peigné et al. (2009), Vásat et al. (2010) e Diggle et al. (2010).

O objetivo principal deste trabalho é utilizar a geoestatística, associada à teoria da informação de sinais elétricos, principalmente o teorema da taxa Nyquist, para determinar um tamanho de amostra ideal para os pesquisadores que utilizam grade regular quadrática em que o modelo de dependência espacial ajustado é o gaussiano. O que se deseja é obter teoricamente uma densidade amostral necessária para a reconstrução de mapas populacionais da variável estudada em que as condições de regularidade exigidas pela geoestatística foram verificadas, a saber: estacionaridade de 1ª e 2ª ordem e/ou estacionaridade do semivariograma, ausência de outiliers e tendências, e semivariograma isotrópico.

As referências utilizadas como apoio ao objetivo principal são: Modis e Papaodysseus (2006) que apresentaram esta metodologia teórica para os modelos esférico e exponencial; Yfantis et al. (1987) que mostraram a eficiência teórica entre diferentes grades amostrais; Vásat et al. (2010) que mostraram uma alternativa para reduzir o tamanho de amostra para um processo multivariado; Ferreira et al. (2013) que apresentaram uma metodologia sistemática e interativa da análise dos dados geoestatísticos e Santos et al. (2011) que apresentaram um estudo comparativo de precisão com os mapas gerados pelos diferentes interpoladores lineares geoestatísticos.

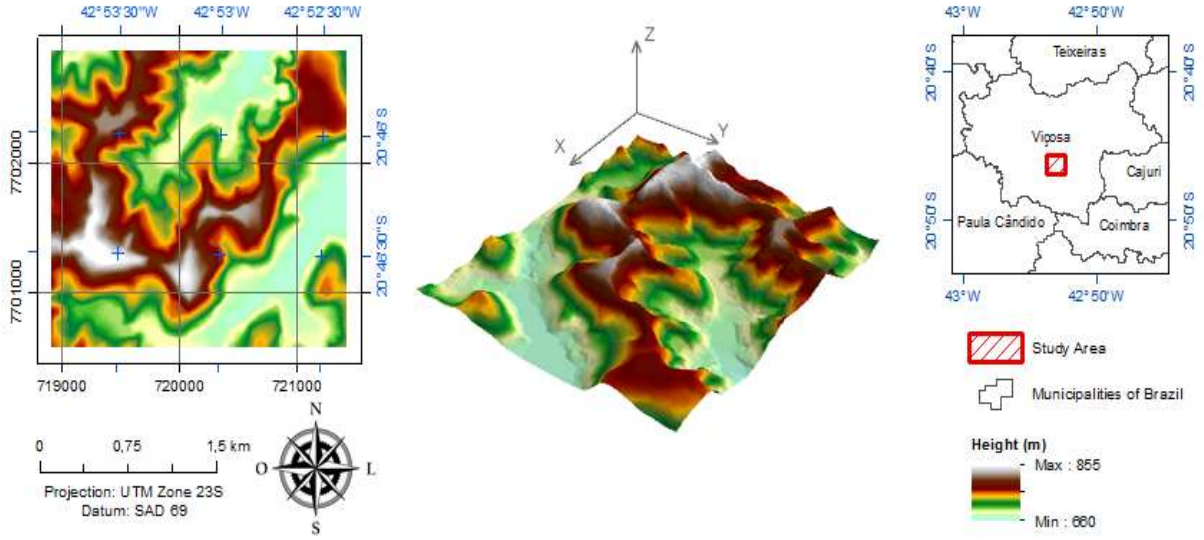
2 MATERIAL E MÉTODOS

Apresenta-se em seguida a metodologia adotada e a descrição da área estudada, descrevendo a proposição do método e apresentação dos dados.

2.1 Descrição da área de estudo

A área de estudo compreende uma parcela de 5,7 km^2 do município de Viçosa, situada na Zona da Mata, no Estado de Minas Gerais, Brasil, delimitada pelas latitudes 20º45’39” S a 20º46’53” S, e as longitudes 42º52’24” W a 42º53’49” W, como mostrado na Figura 1.

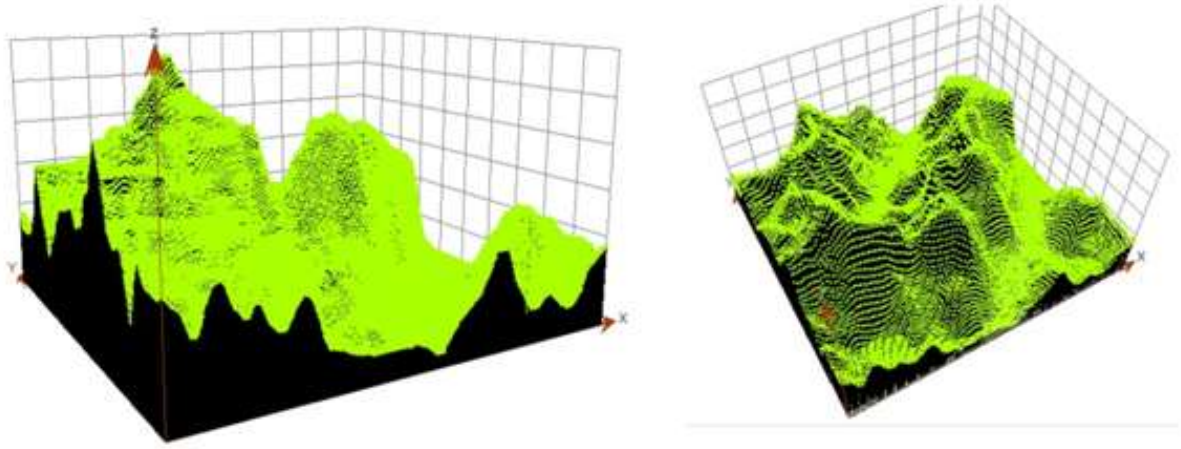
Figura 1 – Representação da área de estudo, município de Viçosa, no Estado de Minas Gerais, Brasil.



Fonte: Rosa (2017)

Os dados de altimetria utilizados neste trabalho são referenciados ao sistema geodésico Sirgas 2000 e representado no sistema de projeção UTM (Universal Transversa de Mercator) zona 23 Sul. Esses dados compreendem, aproximadamente, 230 mil pontos de altitudes conhecidas, com um espaçamento de cerca de 5 metros nas direções X e Y. Os valores das altitudes apresentam valores mínimos de 660 metros e valores máximos de 885 metros, mostrados na Figura 2.

Figura 2 – Representação tridimensional do levantamento altimétrico de parte da Zona da Mata, do município de Viçosa-MG, Brasil, contendo aproximadamente 230 mil pontos amostrais.



Fonte: Rosa (2017)

2.2 A proposição do método

Modis e Papaodysseus (2006), ao utilizar o teorema da teoria da informação de sinais elétricos, denominado taxa Nyquist e a transformada de Fourier às funções de correlação da geoestatística, mostrou que é possível obter um tamanho de amostra adequado para a geoestatística, visando a reconstrução total da população estudada.

No entanto, esses pesquisadores apresentaram apenas uma solução para os modelos teóricos de semivariogramas esféricos e exponenciais.

Segundo Modis e Papaodysseus (2006), como sugestão final de uma pesquisa, um algoritmo prático foi apresentado, a saber:

i) comece com o tamanho de amostra disponível;

ii) ajuste o modelo de covariância correspondente;

iii) determine o limite superior prático do espectro do modelo obtido (taxa Nyquist);

iv) usando iii determine o tamanho de amostra ideal;

v) se esse tamanho de amostra ideal não for atingido, repita o processo de amostragem, se possível;

Frequentemente usado para fenômenos naturais com elevada continuidade, destaca-se o modelo gaussiano, caracterizado pela Equação (1):

$$\gamma(\mathrm{h}) = \begin{cases} 0\,; & h = 0 \\ C_0 + C_1\left[1 - \exp\left(-\dfrac{3h^2}{a^2}\right)\right]\,; & 0 < h < a \\ C_0 + C_1\,; & h > a \end{cases} \qquad (1)$$

No modelo gaussiano da Equação (1), C_0 representa o efeito pepita, interseção do gráfico com o eixo das ordenadas, $C_0 + C_1$ representa o patamar, valor da semivariância relacionado ao alcance a da dependência espacial; C_1 representa a contribuição; h vetor distância entre os pontos.

De acordo com Ferreira et al. (2013), os semivariogramas ajustados pelo modelo gaussiano são caracterizados por uma dependência espacial que apresenta baixas variações entre os vizinhos mais próximos e maiores variações para os vizinhos mais distantes, ainda dentro do alcance do semivariograma.

Geralmente, variáveis como altimetria e batimetria apresentam características como essas, o que justifica o uso da variável altimetria neste trabalho.

De acordo com Modis e Papaodysseus (2006), com base no teorema da taxa Nyquist, a reconstrução dos mapas populacionais é possível se a densidade da amostra, utilizando amostras de grades regulares, é maior ou igual a um valor crítico, que pode ser obtido a partir da função de correlação sob condições de regularidade, a saber: estacionaridade, isotropia, forte dependência espacial e alcance finito.

Conforme citado, esses autores apresentaram a teoria e os resultados apenas para modelos Esférico e Exponencial. Além disso, mostraram que para as variáveis na área de mineração (minério homogêneo), a prática corresponde à utilização dessa teoria na geoestatística e que a densidade da amostra deve ser menor ou igual a metade do alcance prático do experimento, ajustado no semivariograma.

Neste sentido, este trabalho visa, especificamente: desenvolver a teoria da taxa Nyquist para o modelo gaussiano; verificar que, para os dados de altimetria, essa teoria se aplica e estimar a densidade da amostra para o modelo da Equação (1).

A partir do modelo gaussiano caracterizado pela Equação (1) Abramowitz e Stegun (1972) obtiveram a Equação (2), utilizando a transformada de Fourier para a função de correlação do modelo gaussiano, que é inversamente relacionada ao semivariograma.

$$R(\omega) = \exp\left(-\frac{3t^2}{a^2}\right)\cos(\omega t) \tag{2}$$

sendo ω a taxa de amostragem relacionada à frequência dos sinais e t um instante de amostragem.

Ainda de acordo com Abramowitz e Stegun (1972), a função de densidade espectral de potência, que é o modelo que descreve o comportamento da função de correlação do modelo gaussiano, é dada pela Equação (3):

$$S(\omega) = \frac{1}{2} a \sqrt{\frac{\pi}{3}} \exp\left(-\frac{\omega^2 a^2}{12} \right) \quad (3)$$

Como $S(\omega)$ tende a zero, devido à estabilização da função de correlação, $\exp\left(-\frac{w^2 a^2}{12} \right)$ tende a zero quando ω tende ao infinito. No entanto, Journel e Huijbregts (1978) afirmam que o modelo gaussiano, por ser assintótico ao eixo, deve ter nulidade considerada em 5%. Assim,

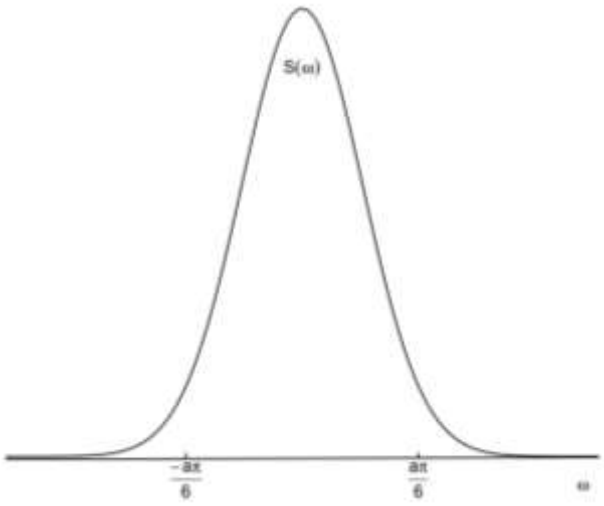
$$\exp\left(-\frac{\omega^2 a^2}{12} \right) = 0{,}05 \Leftrightarrow \ln\left[\exp\left(-\frac{\omega^2 a^2}{12} \right) \right] = \ln(0{,}05) \Leftrightarrow -\omega^2 a^2 \cong -36 \Leftrightarrow \omega \cong \frac{6}{a} \quad (4)$$

Do teorema da amostragem, originalmente desenvolvido para sinais elétricos (também conhecido como taxa Nyquist), o tamanho amostral *T* é dado pela Equação (5).

$$T \leq \frac{\pi}{\omega} \cong \frac{\pi}{\frac{6}{a}} = \frac{\pi}{6} a \cong 0{,}5a \quad (5)$$

Pela igualdade 5 percebe-se que o tamanho amostral *T* é aproximadamente metade do alcance teórico *a*, conforme obtido por Modis e Papaodysseus (2006), e representado na Figura 3.

Figura 3 – Gráfico da função de densidade espectral.



Fonte: Modis e Papaodysseus (2006)

No entanto, de acordo com Olea (1999), alguns modelos de variograma teórico e, consequentemente a função de correlação, não alcançam a estabilização da curva no alcance

teórico *a*, sendo o gaussiano um desses modelos. Dessa forma, este autor apresenta a transformação do alcance teórico para o alcance prático a_p, por meio de:

$$a = \frac{a_p}{\sqrt{3}} \Rightarrow a = \frac{\sqrt{3}}{3} a_p \tag{6}$$

Portanto, o tamanho amostral *T*, em função do alcance prático é dado pela Equação (7):

$$T \leq \frac{\pi}{18} a_p \cong 0{,}30 a_p \tag{7}$$

O resultado da Equação (7) mostra que a distância máxima entre dois pontos de uma grade quadrática regular de amostragem deve ser aproximadamente 30% do alcance prático. Isso significa que uma primeira amostragem, chamada de amostragem experimental, deva ser feita para que a densidade da amostra possa ser estimada como 30% do alcance prático, se comprovada a existência das condições de regularidade para o modelo de correlação gaussiano e, consequentemente, para o semivariograma.

Observando uma grade quadrática regular, a maior distância entre os pontos está nas diagonais. Dessa forma, o tamanho da distância máxima para os lados deve ser convertido antes, usando a relação $d = S\sqrt{2}$, em que d significa diagonal e S significa lado do quadrado.

2.3 Descrição dos dados

O conjunto de dados, referente à região brasileira, situada na Zona da Mata, no município de Viçosa-MG, Brasil, consistiu de 229.414 pontos. Com o objetivo de verificar a teoria aplicada neste estudo, este conjunto de dados foi reduzido (seguindo e verificando as condições de regularidade necessárias na descrição do método) 17 vezes (atingindo o número de 49 pontos para o tamanho de amostra).

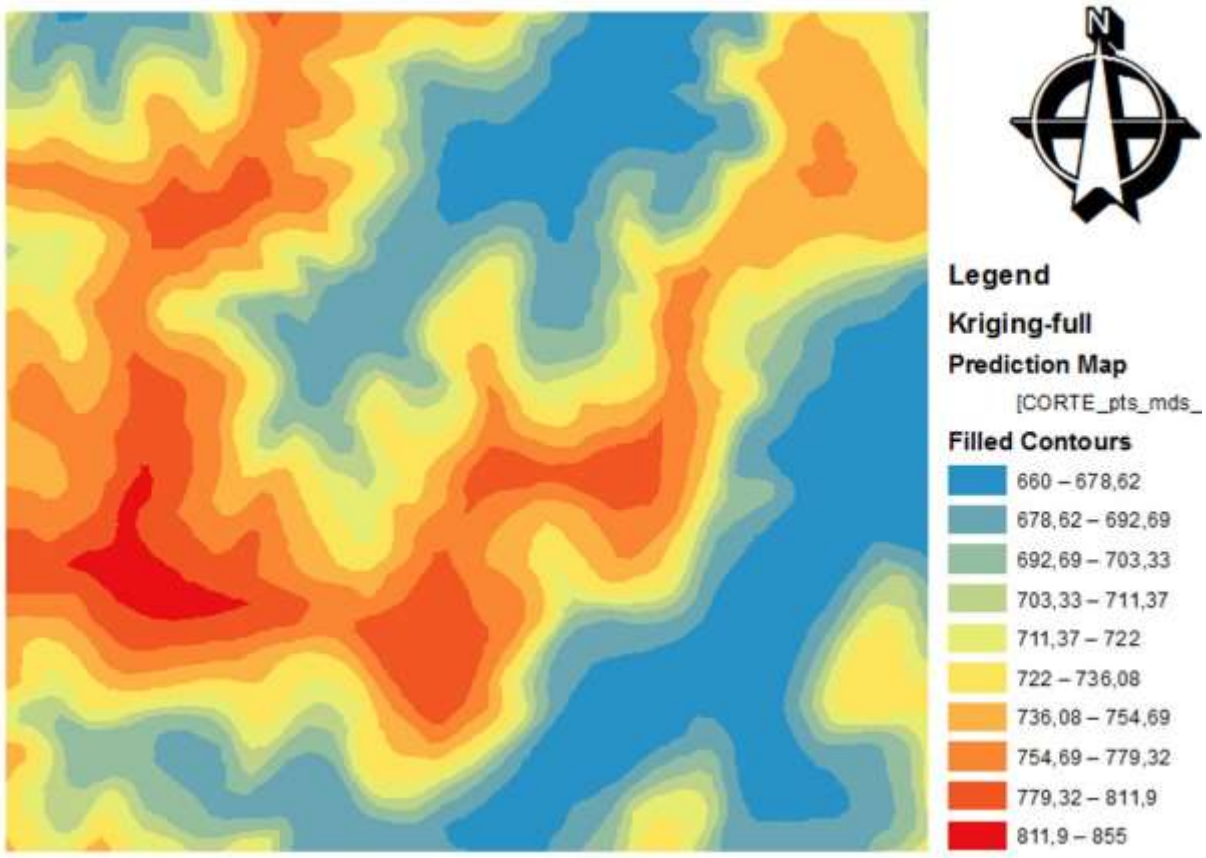
Como critério de eficiência, considerou-se a variância de krigagem, os coeficientes da Regressão Linear Simples (apenas RLS) entre os valores preditos e observados da validação cruzada e a variância dos dados para a estimativa do parâmetro alcance na modelagem dos semivariogramas (VIEIRA, 2000; MODIS e PAPAODYSSEUS, 2006; SANTOS et al., 2011; FERREIRA et al., 2013; YAMAMOTO e LANDIM, 2013).

Devido à limitação de alguns *softwares* em relação ao número de informações no conjunto de dados, foi escolhida a utilização do ArcGIS (ESRI, 2014) para toda a análise computacional do trabalho. Para a redução do tamanho de amostra, adotou a seleção regular

dos dados de altimetria, usando a ferramenta de amostragem regular do *software* ArcGIS 10.2.2. O primeiro passo desse processo foi definir o espaçamento em ambos os sentidos X e Y para executar a seleção regular. Desta forma, foi criada uma grade intermediária de pontos baseando-se no espaçamento definido e, em seguida, obtendo o ponto da base de dados de altimetria mais próximo de cada ponto criado na grade intermediária. Depois disso, a seleção desses pontos mais próximos foi feita, tendo como resultado um conjunto de dados de altimetria com uma amostra "quase regular". Como os conjuntos de dados deste trabalho são grandes, com uma alta densidade de pontos, esse resultado "quase" pode ser considerado, de forma prática, uma seleção regular, que foi previamente comprovada com a ferramenta analisador da proximidade do Toolbox, usando o comando Near (ESRI, 2014).

Portanto, foram adotados como mapas populacionais da área estudada, os resultados da krigagem simples (SANTOS et al., 2011) utilizando todo o conjunto de dados de altimetria, em metros, conforme Figura 4.

Figura 4 – Krigagem simples dos dados de altimetria de parte da Zona da Mata, município de Viçosa-MG, Brasil.



Fonte: Rosa (2017)

3 RESULTADOS E DISCUSSÃO

Pela Tabela 1 é possível notar que as médias e variâncias não sofreram grandes alterações, apenas quando a amostra foi 99,96% e 99,98% menor, atingindo nesta última o limite mínimo do tamanho de amostra para a geoestatística, conforme recomendado por Yamamoto e Landim (2013).

As reduções foram feitas com base no espaçamento regular entre pontos, selecionados em grades quadráticas com lados de 7m, 10m, 15m, 20m, 25m, 30m, 40m, 50m, 70m, 90m, 130m, 160m, 190m, 220m, 260m, 300m e 340m.

De acordo com os dados analíticos apresentados na Tabela 1, pode-se notar que a decisão sobre a representação de uma amostragem não pode ser julgada, simplesmente, por tamanho, média e variância.

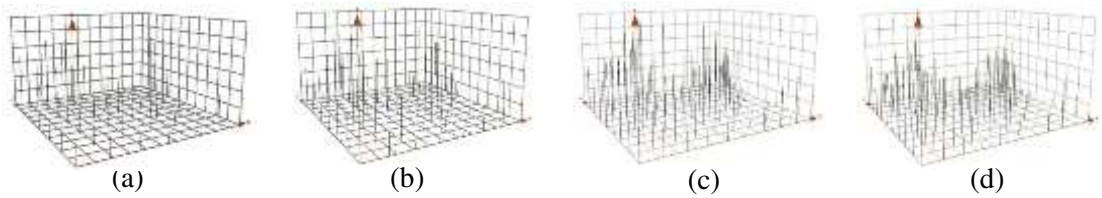
Tabela 1 – Apresentação dos tamanhos de amostragem, redução de amostras, médias, variâncias e espaçamento das grades regulares quadráticas do levantamento altimétrico de parte da Zona da Mata, perto da cidade de Viçosa-MG, Brasil.

Tamanho amostral	Redução (%)	Média	Variância	Espaçamento
229.414	0,00	721,47	1.473,48	"população"
116.708	49,13	721,50	1.474,94	7 metros
57.228	75,05	721,46	1.474,94	10 metros
25.384	88,94	721,57	1.475,87	15 metros
14.364	93,74	721,37	1.473,79	20 metros
9.292	95,95	721,63	1.476,17	25 metros
6.308	97,25	721,65	1.477,33	30 metros
3.591	98,43	721,36	1.474,10	40 metros
2.300	99,00	721,49	1.474,48	50 metros
1.188	99,48	721,23	1.466,05	70 metros
700	99,69	721,40	1.483,56	90 metros
342	99,85	721,26	1.460,84	130 metros
224	99,90	721,16	1.498,39	160 metros
156	99,93	721,96	1.538,68	190 metros
110	99,95	721,99	1.514,61	220 metros
90	99,96	719,88	1.477,86	260 metros
64	99,97	721,78	1.569,59	300 metros
49	99,98	719,12	1.390,62	340 metros

Fonte: Mendes *et al*. (2018)

É apresentado, nas Figuras 5, 6 e 7, as representações em três dimensões das grades regulares de amostra visando mostrar a perda significativa da representatividade da população quando a amostragem não apresenta um tamanho adequado, de acordo com o critério utilizado.

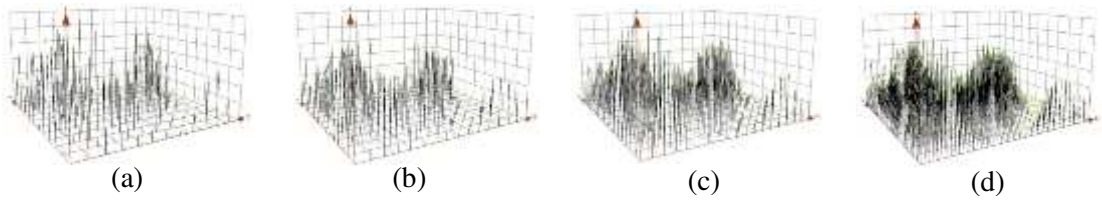
Figura 5 – Representação tridimensional amostral de grade regular quadrática do levantamento altimétrico de parte da Zona da Mata, município de Viçosa-MG, Brasil. As grades apresentam (a) 49 pontos, (b) 64 pontos, (c) 90 pontos e (d) 110 pontos.



(a) (b) (c) (d)

Fonte: Mendes et al. (2018)

Figura 6 – Representação tridimensional amostral de grade regular quadrática do levantamento altimétrico de parte da Zona da Mata, município de Viçosa-MG, Brasil. As grades apresentam (a) 156 pontos, (b) 224 pontos, (c) 342 pontos e (d) 700 pontos.



(a) (b) (c) (d)

Fonte: Mendes et al. (2018)

Figura 7 – Representação tridimensional amostral de grade regular quadrática do levantamento altimétrico de parte da Zona da Mata, município de Viçosa-MG, Brasil. As grades apresentam (a) 9292 pontos, (b) 14364 pontos, (c) 25384 pontos e (d) 57228 pontos.

(a) (b) (c) (d)

Fonte: Mendes et al. (2018)

Como era esperado, de acordo com Oliveira et al. (2014), aumentando o tamanho de amostra, a visualização gráfica tridimensional simples dos dados informa o comportamento real da população, o que pode ser observado nas Figuras 6 e 7.

De acordo com Oliveira et al. (2014) sempre houve uma preocupação em determinar os principais indicadores de uma amostra representativa para as Estatísticas Clássicas, no entanto, essas medidas não alcançam esse objetivo por si mesmas.

Para Yamamoto e Landim (2013), em Estatística Espacial, apesar das muitas e variadas tentativas, essa pesquisa mostra igual importância do assunto. Dessa forma, até que os "melhores" indicadores sejam encontrados, é fundamental para avaliar os trabalhos científicos sobre esse assunto, buscar mecanismos viáveis para tal determinação.

Modis e Papaodysseus (2006), Clark (2009) e Yamamoto e Landim (2013), apresentam trabalhos que indicam tentativas deste 1975. Observa-se a partir desses trabalhos que a prática é utilizar a variância de krigagem como critério de eficiência.

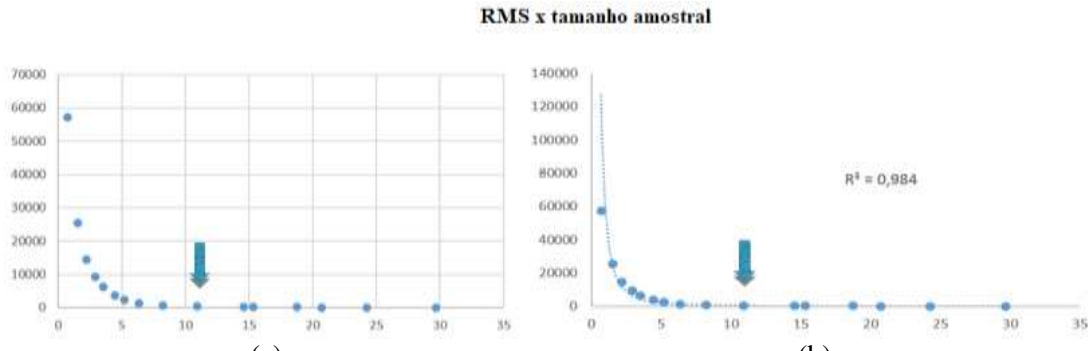
Vieira (2000), Santos et al. (2011) e Ferreira et al. (2013), além de utilizar a variância de krigagem como critério, também utilizaram a Regressão Linear Simples (RLS) entre os valores preditos por krigagem e os valores observados, após o processo de validação cruzada.

De acordo com Vieira (2000) e Ferreira et al. (2013), o coeficiente angular teórico deve ser igual a 1. No entanto, na prática, deve-se observar a proximidade desse valor teórico. A RLS deste coeficiente em função da variância média de krigagem (RMS) apresentou o modelo estimado $\hat{Y} = -1{,}75X + 1{,}72$, com um coeficiente de determinação $R^2 = 0{,}9593$, o que significa que, além da representação de uma relação linear inversa (aumentando o RMS o coeficiente angular é reduzido), cerca de 9% da variação da variância RMS pode ser explicada pela variação do coeficiente angular.

Segundo Modis e Papaodysseus (2006), o indicador espacial usado em trabalhos científicos desta natureza, é uma função entre as variâncias médias de krigagem e a densidade da amostra nas mesmas condições e de diferentes tamanhos, conservando as mesmas condições de amostra da variável estudada (como a amostra apresentada neste trabalho).

Esses autores ainda afirmam que o tamanho ideal de amostragem foi determinado pela estabilização da curva ajustada ao gráfico, cuja diferença entre as variâncias médias de dois vizinhos não melhorou o desempenho da precisão pela krigagem. O gráfico desta função para a área estudada, apresentado na Figura 8, mostra uma estabilização da variância média de krigagem, denominada RMS (Root Mean Square), em torno do tamanho de amostra 156, 224 e 342 pontos.

Figura 8 – Representação gráfica da relação entre a variância média de krigagem e o tamanho de amostra para os dados de altimetria de parte da Zona da Mata, município de Viçosa-MG, Brasil. (a) diagrama de dispersão, (b) diagrama com a linha de tendência e o coeficiente de determinação, R^2.



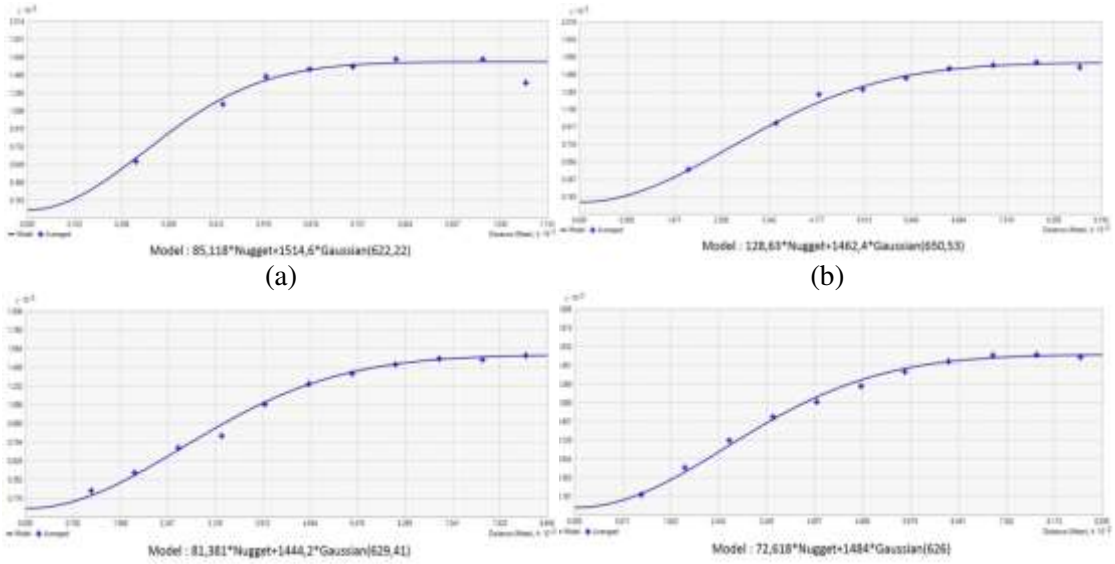
(a) (b)

Fonte: Mendes et al. (2018)

Entre estas, pode ocorrer a dúvida de escolher entre a amostra de 224 e 342 pontos, mas pela teoria apresentada neste artigo, o tamanho de amostra ideal é 342 pontos (destacado na Figura 8 através de uma flecha), pois a distância lateral entre os pontos é cerca de 130 metros (ou 190 metros na diagonal), o que corresponde a 21,6% (ou 31,5% se considerar a diagonal) em relação ao alcance prático de 603 metros.

É apresentado, na Figura 9, o comportamento do semivariograma para alguns diferentes tamanhos de amostragem do estudo. Pode-se perceber que o modelo ajustado aos semivariogramas experimentais foi o modelo gaussiano, apresentado na Equação 1.

Figura 9 – Semivariogramas experimentais (sinais "+") e modelos gaussianos ajustados (linhas) para a dependência espacial dos dados de altimetria de parte da Zona da Mata, município de Viçosa-MG, Brasil. Os tamanhos de amostragem são (a) 156 pontos, (b) 224 pontos, (c) 342 pontos e (d) 700 pontos.



Fonte: Mendes et al. (2018)

Clark (2000) diz que muitos dos pesquisadores geoestatísticos acreditam que o efeito pepita pode ser causado pela baixa densidade da amostra. Apesar de os semivariogramas apresentarem esses valores muito próximos, o efeito pepita aumentou significativamente, variando, desde a grade mais densa até a menos densa, de 42,51 m^2 para 460 m^2 (pode-se observar que esta unidade de medida se refere à variância e não à área).

De fato, pode-se ver que na proposição do método adotado neste trabalho, uma unidade de grande importância é o alcance, porque teoricamente a densidade de amostra estará em função dele. No entanto, aqui é o ponto de grande preocupação deste tipo de estudo, porque o intervalo não se mantém constante com a redução do tamanho de amostra.

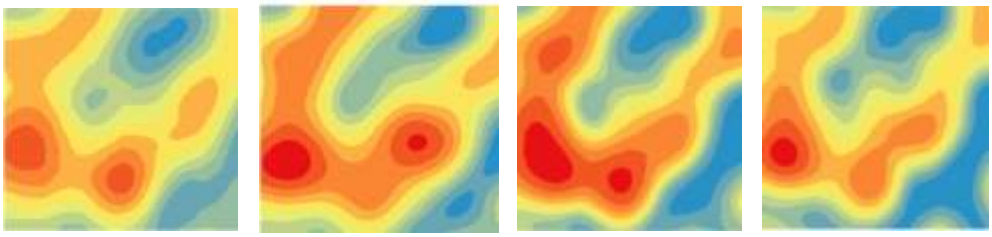
Modis e Papaodysseus (2006), trabalhando com minérios homogêneos, não perceberam a variação na estimativa do alcance variográfico, no entanto, neste estudo, o alcance estimado variou, da grade mais densa para a grade menos densa, de 603 metros para 650 metros.

Mesmo diante desta variabilidade percebida nos alcances, a distância lateral média da grade regular quadrática recomendado neste trabalho é de 130 metros, o que equivale a 342 pontos.

Este valor obtido na análise dos dados está em conformidade com a demonstração apresentada neste trabalho para o modelo gaussiano. Deve ser destacado também que as condições de regularidade geoestatística foram verificadas e comprovadas.

Santos et al. (2011), citando a literatura científica, afirmam que, uma vez que os mapas visam representar a realidade, as pessoas tendem a aceitá-los como verdadeiras. Dessa forma, a produção dos mapas através da interpolação dos dados é uma etapa muito importante da análise geoestatística, porque os mapas produzidos serão, no mínimo, criticados por aqueles que conhecem a região mapeada. Assim, são apresentados, nas Figuras 10 e 11, os mapas obtidos através da interpolação por krigagem simples, conforme recomendado por Santos et al. (2011).

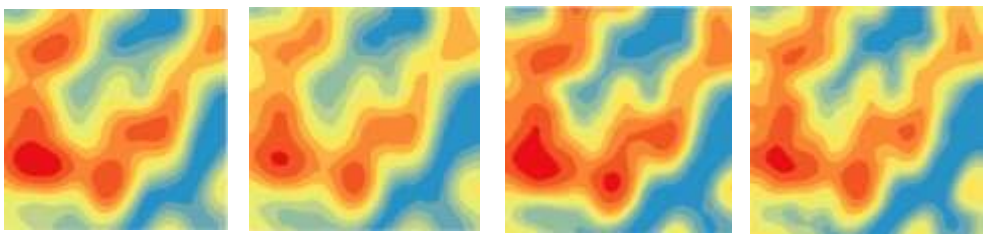
Figura 10 – Krigagem simples do levantamento altimétrico de parte da Zona da Mata, município de Viçosa-MG, Brasil. Os tamanhos de amostragem são (a) 49 pontos, (b) 64 pontos, (c) 90 pontos e (d) 110 pontos.



(a) (b) (c) (d)

Fonte: Mendes et al. (2018)

Figura 11 – Krigagem simples do levantamento altimétrico de parte da Zona da Mata, município de Viçosa-MG, Brasil. Os tamanhos de amostragem são (a) 156 pontos, (b) 224 pontos, (c) 342 pontos e (d) 700 pontos.



(a) (b) (c) (d)

Fonte: Mendes et al. (2018)

Conforme apresentado nas Figuras 10 e 11, o mapa populacional começa a ser visualizado pelo tamanho da amostra de 342 pontos, cuja precisão do estudo atingiu a estabilização, conforme apresentado na Figura 8.

De acordo com Vieira (2000), Modis e Papaodysseus (2006), Santos et al. (2011), Ferreira et al. (2013) e Yamamoto e Landim (2013), na modelagem do semivariograma, o alcance estimado deve estimar a variação dos dados, que por sua vez é estimada pela variação da amostra. Mesmo com a redução da amostragem, o mesmo aconteceu com este conjunto de dados.

Outro importante estágio em uma análise geoestatística é o processo de validação cruzada. Entre os passos importantes desse processo estão a média e a variância dos resíduos gerados entre os valores observados e os preditos. De acordo com Vieira (2000), Santos et al. (2011), Ferreira et al. (2013), espera-se que a média dos resíduos obtidos por este processo tenha valor 0 e a variância tenha valor 1, como mostrado em Mood et al. (1974). Na prática, o que é analisado é a proximidade desses valores.

4 CONCLUSÕES

Neste trabalho, utilizando-se da taxa Nyquist, que sugere uma densidade mínima de amostra para que a população seja reconstruída a partir das amostras, desenvolveu-se a parte teórica importante para o modelo gaussiano de dependência espacial, cuja densidade mínima teórica foi em função de um alcance prático estimado, um dos parâmetros do semivariograma experimental ajustado por este modelo. Em termos práticos, a partir da primeira amostragem experimental, a densidade mínima é cerca de 30% do valor estimado para o parâmetro alcance, em condições de regularidade.

A teoria desenvolvida foi aplicada a um grande conjunto de dados e a densidade teórica foi comprovada pela prática, adotando como critérios de avaliação alguns procedimentos e medidas já consolidadas na área de geoestatística clássica.

Sugere-se uma pequena alteração em um algoritmo prático com o objetivo de alcançar o caminho satisfatório para as questões que precisam ser esclarecidas no planejamento da amostra de pesquisa que irão utilizar as estatísticas espaciais.

Como recomendação para trabalhos futuros, sugere-se o estudo do comportamento do alcance prático quando a dimensão da amostra é reduzida, considerando acima de tudo a modelagem desse comportamento; o estudo das funções críticas, como a função de perda, por exemplo, quando a dimensão da amostra teoricamente desejada não é financeiramente viável; e o desenvolvimento teórico de uma densidade de amostragem mínima para a amostra experimental denominada.

REFERÊNCIAS BIBLIOGRÁFICAS

ABRAMOWITZ, M.; STEGUN, I. A. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, Washington, D.C.: U.S. **Government Printing Office**, 1972. 1046 p.

ARMSTRONG, M. **Basic Linear Geostatistics**. Springer, Berlin, 1998. 153p.

BROOKER, P. I. **A geostatistical primer**. Singapore: World Scientific, 1991. 95 p.

BRUS, D. J.; HEUVELINK, G. B. M. **Optimization of sample patterns for universal kriging of environmental variables**. Geoderma, n. 138, p. 86-95, 2007.

CLARK, I. **Practical geostatistics**. London: Applied Science Publishers, 1979. 130p.

CLARK, I. **SAIMM Conference, Fourth World Conference on Sampling and Blending**, p. 21-23, 2009.

CLARK, I.; HARPER, W. V. **Practical geostatistics 2000**. Geostokos (Ecosse) Limited, 2000. 416 p.

CRESSIE N. A. C.; WIKLE, C. K. **Statistics for spatio-temporal data.** Wiley Series in Probability and Statistics. Hoboken, New Jersey, 2011. 588 p.

DIGGLE, P. J., MENEZES, R.; SU, T. Geostatistical inference under preferential sampling. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, v. 59, n. 2, p. 191-232, 2010.

DRUCK, S.; CARVALHO, M.S.; CÂMARA, G.; MONTEIRO, A. M. V. **Análise Espacial de Dados Geográficos**. Brasília: EMBRAPA, 2004. 209 p.

ESRI. **ArcGIS 10.2.2 for Desktop.** ESRI: Redlands, USA, 2014.

FERREIRA, D. F. **Estatística básica**. 2. ed. Lavras: Editora UFLA, 2009. 664 p.

FERREIRA, I. O.; SANTOS, G. R.; RODRIGUES, D. D. Estudo sobre a utilização adequada da krigagem na representação computacional de superfícies batimétricas. **Revista Brasileira de Cartografia,** n. 65/5, p. 831-842, 2013.

JOURNEL, A. G., HUIJBREGTS, C. J. **Mining Geostatistics**. Academic Press, London, 1978. 600 p.

MENDES, A.; SANTOS, G. R. dos; EMILIANO, P. C. ; ILAMBWETSI, P. S. ; KALEITA, A. L. . Theoretical Estimation of the Sampling Size of Geostatistics considering Gaussian Variogram Model. **SIGMAE**, v. 7, p. 17-30, 2018.

MODIS, K.; PAPAODYSSEUS, K. Theoretical Estimation of the Critical Sampling Size for Homogeneous Ore Bodies with Small Nugget Effect. **Mathematic Geology,** v. 38, n. 8, p. 489-501, 2006.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the Theory of Statistics**. McGraw Hill, 1974. 480 p.

OLEA, R. A. **A practical primer on geostatistics**. U.S. Geological Survey Open-File Report, 2009-1103, 2009. 346 p.

OLEA, R. A. **Geostatistics for engineers and earth scientists**. Kluwer Academic Publishers, London, 1999. 303 p.

OLIVEIRA, M. S. de; BEARZOTI, E.; VILLAS BOAS, F. L.; NOGUEIRA, D. A.; NICOLAU, L. A.; OLIVEIRA, H. S. S. de. **Introdução à Estatística**, 2. ed. Lavras: UFLA, 2014. 462 p.

PEIGNÉ, J.; VIAN, J. F.; CANNAVACCIUOLO, M.; BOTTOLLIER, B.; CHAUSSOD, R. Soil Sampling based on field spatial variability of soil microbials indicators. **European Journal of Soil Biology**, v. 45, p. 488-495, 2009.

PINTO, E. S. O.; SANTOS, G. R. ; OLIVEIRA, F. L. P. Análise Espaço-Temporal Aplicada às Ocorrências de Hipertensão e Diabetes nos Municípios do Estado de Minas Gerais. **Revista Brasileira de Biometria**, v. 32, p. 238-266, 2014.

ROSA, L. M. . **Estudos sobre a influência de afirmações populares na Geoestatística clássica**. 2017. 103 p. Tese (Doutorado em Estatística Aplicada e Biometria). Viçosa: UFV.

SANTOS, G. R., OLIVEIRA, M. S., LOUZADA, J. M., SANTOS, A. M. R. T. Krigagem Simples versus Krigagem Universal: qual o preditor mais preciso? Revista **Energia na Agricultura**, Botucatu, v. 26, n. 2, p. 49-55, 2011.

SARTORI, A. A. C.; ZIMBACK, C. R. L. Recomposição florestal visando à conservação de recursos hídricos na bacia do rio Pardo, São Paulo. **Revista Energia na Agricultura**, v. 26, n. 4, p. 43-53, 2011.

SOUZA, Z. M.; SOUZA, G. S.; JÚNIOR, J. M.; PEREIRA, G. T. Número de amostras na análise geoestatística e na krigagem de mapas de atributos do solo. **Revista Ciência Rural**, Santa Maria, v. 44, n. 2, p. 261-268, 2014.

VÁSAT, R.; HEUVELINK, G. B. M.; BORŮVKA, L. Sampling design optimization for multivariate soil mapping. **Geoderma**, v. 155, n. 3-4, p. 147–153, 2010.

VER HOEF, J. M. Sampling and geostatistics for spatial data. **Ecoscience**, v. 9, p. 152-161, 2002.

VIEIRA, S. R. Geoestatística em estudos de variabilidade espacial do solo. In: NOVAIS, R.F.; ALVAREZ, V.H.; SCHAEFER, G.R. **Tópicos em ciência do solo**. Viçosa, v.1, p.1-54, 2000.

WEBSTER, R.; OLIVER, M. A. **Geostatistics for Environmental Scientists**. 2. ed. Chichester: John Wiley & Sons, 2007. 315p.

WEBSTER, R.; OLIVER, M. A. Sample adequately to estimate variograms of soil properties. **Journal of Soil Science**, v. 43, p. 177-192, 1992.

YAMAMOTO, J. K.; LANDIM, P. M. B. **Geoestatística: Conceitos e Aplicações.** Oficina de Textos: São Paulo, 2013. 216 p.

YFANTIS, E. A.; FLATMAN, G. T.; BEHAR, J. V. Efficiency of kriging estimation for square, triangular and hezagonal grids. **Mathematical Geology**, v. 19, p. 183-205, 1987.

CAPÍTULO 2 – ESTIMAÇÃO DO TAMANHO AMOSTRAL NA GEOESTATÍSTICA USANDO UM MODELO DE SEMIVARIOGRAMA GAUSSIANO NA PRESENÇA DE *OUTLIERS*

RESUMO

A determinação de um tamanho de amostra que seja adequado para a reconstrução da população, na análise de dados espaciais, é algo que tem sido estudado em vários trabalhos. Independentemente da área de estudo, qualquer variável pode conter *outlier*. Conforme sugerido por alguns pesquisadores, no intuito de eliminar tais dados discrepantes, metodologias vêm sendo criadas para atender às demandas das diversas áreas do conhecimento científico. O objetivo deste trabalho é utilizar o teorema da taxa de Nyquist para determinar um tamanho ideal para amostras georreferenciadas contendo *outliers*, oriundas de uma grade quadrática regular, no qual o modelo de dependência espacial é o gaussiano. O que se pretende atingir é uma densidade de amostragem necessária para a reconstrução de mapas populacionais de variáveis nas quais as condições de regularidade necessárias em geoestatística foram verificadas. Como resultado pode-se concluir que o tamanho ideal de amostragem obtido na ausência de *outliers*, 115 pontos, foi bem inferior aos 228 pontos obtidos na presença dos *outliers*.

Palavras-chave: Geoestatística. Amostragem. Taxa Nyquist. *Outliers*.

ABSTRACT

The determination of a sample size that is suitable for population reconstruction in the analysis of spatial data is something that has been studied in several studies. Regardless of the study area, any variable may contain outlier. As suggested by some researchers, in order to eliminate such discrepant data, methodologies have been created to meet the demands of various areas of scientific knowledge. The purpose of this work is to use the Nyquist Rate Theorem to determine an ideal size for georeferenced samples containing outliers from a regular quadratic grid in which the spatial dependence model is the Gaussian. What we intend to achieve is a sampling density necessary for the reconstruction of population maps of variables in which the necessary regularity conditions in Geostatistics were verified. As a result it can be concluded that the ideal sampling size obtained in the absence of outliers, 115 points, was well below the 228 points obtained in the presence of outliers.

Keywords: Geostatistics. Sampling. Nyquist Rate. *Outliers*.

1 INTRODUÇÃO

Em várias abordagens de análise espacial se faz necessário coletar uma quantidade considerável de amostras georreferenciadas a fim de produzir um mapeamento da região de estudo e, dependendo do tamanho e da localização da região estudada, a aquisição de tais informações demanda tempo e investimentos financeiros consideráveis.

Nas diversas áreas do conhecimento, a geoestatística vem sendo utilizada como principal método de análise de dados de amostras (YAMAMOTO e LANDIM, 2015). Fundamentada no estudo de uma função espacial que varia localmente com continuidade, cujos valores são relacionados com a posição espacial que ocupam (FARACO et al., 2008), permite a estimativa de uma determinada variável em locais não amostrados e aplicações em planejamentos de amostragens e modelagens e em mapeamentos (GOMES et al., 2007; GOMES et al., 2008).

Na análise de dados espaciais, a determinação de um tamanho amostral que seja adequado para a reconstrução da população é algo que tem sido estudado em vários trabalhos: Modis e Papaodysseus (2006) apresentaram uma metodologia teórica baseada no teorema da taxa Nyquist para a determinação de um tamanho de amostra ideal para pesquisadores que utilizam grade regular quadrática em que o modelo de dependência ajustado é o exponencial ou esférico; Vásat et al. (2010) mostraram uma alternativa para reduzir o tamanho de amostra para um processo multivariado; Souza et al. (2014) analisaram diferentes intensidades de amostragem do solo com relação à precisão na análise geoestatística e interpolação de mapas, para fins de agricultura de precisão em área de cana-de-açúcar.

Qualquer variável, independentemente da área do conhecimento, pode conter discrepâncias (*outliers*) em diferentes escalas, sendo que suas causas podem estar associadas, dentre outros, a erros instrumentais, erros dos observadores e problemas na mecanização de monitoramento (MORETTIN e TOLÓI, 2002).

Metodologias para a detecção de *outliers* vêm sendo criadas para atender às demandas das diversas áreas do conhecimento científico, como proposto por Barua e Alhajj (2007) para processamento de imagens, Qiao et al. (2013) para dados provenientes de satélites e Appice et al. (2014) para fluxo de dados geofísicos.

Santos et al. (2017) propuseram um método de detecção e eliminação de *outliers* para dados geoespaciais contínuos através da geoestatística e teoremas da estatística clássica, independentemente da causa geradora das inconsistências.

Utilizando a geoestatística associada ao teorema da taxa Nyquist proposto por Modis e Papaodysseus (2006) e a proposta para eliminação de *outliers* feita por Santos et. al (2017), objetivou-se neste trabalho obter, a partir de um banco de dados contendo *outliers*, um tamanho ideal de amostra utilizando uma grade quadrática regular na qual o modelo de dependência espacial é o gaussiano. Mais especificamente, pretendeu-se obter um tamanho de amostra ideal necessária para a reconstrução de mapas populacionais para a variável altimetria, na presença de *outliers*.

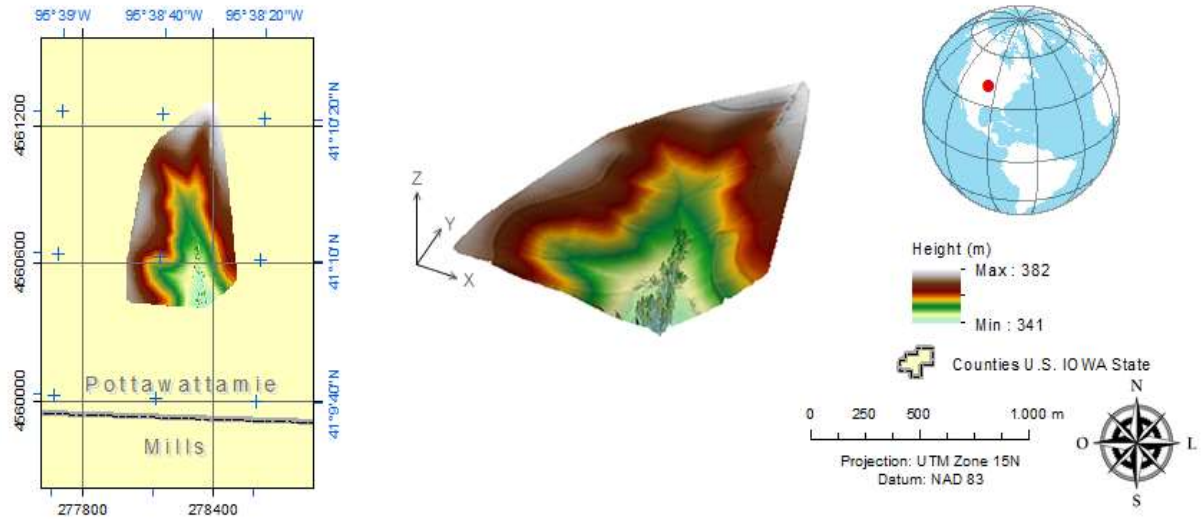
2 MATERIAL E MÉTODOS

Visando alcançar os objetivos desse trabalho, a seguir é apresentada a descrição da região e da proposta do estudo e a caracterização dos dados.

2.1. Descrição da região de estudo

A área estudada compreende uma parcela de 34,3 hectares da cidade de Treynor, situada no município de Pottawattamie, no Estado de Iowa, Estados Unidos. A região estudada é delimitada pelas latitudes 41º10’23” N e 41º09’53” N e longitudes 95º38’24’’ W a 95º38’47” W, como mostrado na Figura 1.

Figura 1 – Representação da área de estudo, próximo da cidade de Treynor-Iowa, Estados Unidos



Fonte: Santos et al. (2017)

Atualmente, para o mapeamento de média e grande escalas, modelos digitais de elevação (MDE), podem ser produzidos utilizando principalmente a tecnologia LiDAR (Light Detection and Ranging). Esse método mostrou-se eficiente e acurado, além de proporcionar alta densidade de pontos planialtimétricos (HÖHLE e HÖHLE, 2009).

Os dados de altimetria utilizados nesse trabalho são de um mapeamento LiDAR, sendo referenciados ao sistema geodésico NAD 83 (Datum norte-americano de 1983) e representados no sistema de projeção UTM (Universal Transverse Mercator). Esses dados compreendem pouco mais de 192 mil pontos de altitude conhecidas, com uma densidade de 0,55 pontos/m^2 e um espaçamento de aproximadamente 1,7 e 1,2 metros nas direções X e Y, respectivamente.

2.2. Detecção de *outliers*

Santos et al. (2017) propuseram um método de detecção de dados inconsistentes, *outliers*, para dados geoespaciais contínuos baseando-se na geoestatística, independentemente dos fatores geradores da inconsistência (erros de medição, execução ou variabilidade inerente aos dados).

Uma breve síntese a respeito deste método faz-se necessário, a saber:

i) O método baseou-se nas pressuposições teóricas dos resíduos de uma modelagem estatística, segundo Rencher e Schaalje (2008). Esses resíduos, caracterizados como ruído branco, seguem, em sua forma padronizada, uma distribuição de probabilidade gaussiana com média 0 e variância 1, ou seja, distribuição normal padrão, $\varepsilon_p^{''} \sim Z(0,1)$, em que $\varepsilon_p^{''}$ são os resíduos padronizados, segundo VIEIRA (2000).

ii) Visando atender às pressuposições teóricas dos resíduos, conforme as recomendações de Yamamoto e Landim (2013), Santos et al. (2011) e Vieira (2000), adotou-se a análise geoestatística para os dados geoespaciais.

iii) Para obtenção dos resíduos, a variável regionalizada estudada, *Y*, foi decomposta em três componentes, conforme Equação (1):

$$Y(x) = \mu(x) + \varepsilon^{'}(x) + \varepsilon^{''} \tag{1}$$

em que $\mu(x)$ é a função determinística que descreve a componente estrutural de *Y* em *x*; $\varepsilon^{'}$ é o termo estocástico correlacionado localmente; $\varepsilon^{''}$ é o ruído branco não correlacionado, com distribuição normal com média zero e variância σ^2.

iv) Utilizou-se a metodologia geoestatística para analisar os dados geoespaciais com dependência espacial comprovada e caracterizada, e, consequentemente, obtiveram-se os resíduos dessa modelagem a partir da autovalidação leave-one-out (cada resíduo foi obtido pela diferença entre um valor observado e seu respectivo valor predito).

v) Testou-se independência e distribuição normal com média nula e variância constante para o ruído branco, obtendo-se resultados satisfatórios.

vi) Construiu-se intervalos de confiança (*IC*) com probabilidade $(1 - \alpha)$ para os resíduos adotando-se a distribuição normal padrão *Z*(0,1) e nível de significância α de 1% (arbitrário). Em outras palavras, desejou-se determinar o quanto estas estimativas dos resíduos são prováveis $(1 - \alpha)$ de confiança, com $\alpha \in (0,1)$, conforme Equação (2):

$$P\left[\overline{X} - z_{\alpha}\frac{S}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha}\frac{S}{\sqrt{n}}\right] \quad (2)$$

Silva (2012) mostrou que todos os valores não pertencentes ao (IC) construído, sem viés, com variância mínima e levando em consideração a estrutura de dependência espacial, são possíveis *outliers*. Estatisticamente, se $x_i \in IC_{(1-\alpha)}$ então x_i é ruído branco; caso contrário, é um provável outlier.

vii) Foi possível, ainda, apontar quantos, quais e onde estavam os resíduos com alta probabilidade de serem *outliers*, usando recursos de georreferenciamento de dados.

viii) O método proposto foi comparado e/ou validado a partir da comparação com um método de detecção de *outliers* dos mais robustos e usual, o Box Plot (HOAGLIN et al., 1983).

Com a utilização dessa metodologia foi possível detectar e mapear dados discrepantes. Adicionalmente, Santos (2016) comparou-o com o método muito utilizado de detecção de *outliers*, *BoxPlot*, verificando sua importância e funcionalidade, já que o BoxPlot não detectou nenhum dado como discrepante.

2.3. Taxa Nyquist

Modis e Papaodysseus (2006) mostraram que é possível obter um tamanho de amostra adequado para a geoestatística, visando a reconstrução total da população estudada, usando o teorema da teoria da informação de sinais elétricos, denominada taxa Nyquist e a transformada de Fourier à função de autocorrelação.

Estes autores apresentaram uma solução somente para os modelos teóricos de semivariogramas esféricos e exponenciais, apresentando, ao final, um algoritmo prático, a saber:

i) comece com o tamanho de amostra disponível;

ii) ajuste o modelo de covariância correspondente;

iii) determine o limite superior prático do espectro do modelo obtido (taxa Nyquist);

iv) usando iii determine o tamanho de amostra ideal;

v) se esse tamanho de amostra ideal não for atingido, repita o processo de amostragem, se possível;

Adicionalmente, mostraram que para as variáveis na área de mineração (minério homogêneo), a prática corresponde à utilização dessa teoria na geoestatística e que a

densidade da amostra depende da metade do alcance prático do experimento, estimado na análise do semivariograma empírico.

2.4 Proposta do estudo

Assim como os modelos Esféricos e Exponenciais são muito importantes em muitas áreas do conhecimento, o modelo gaussiano, caracterizado pela Equação (3), é de grande importância para outras variáveis regionalizadas.

$$\gamma(\mathrm{h}) = \begin{cases} 0\,; & h = 0 \\ C_0 + C_1\left[1 - \exp\left(-\dfrac{3h^2}{a^2}\right)\right]\,; & 0 < h < a \\ C_0 + C_1\,; & h > a \end{cases} \tag{3}$$

Os parâmetros do modelo gaussiano apresentado na Equação 3, são: efeito pepita (C_0); patamar (C_0 + C_1); contribuição (C_1); vetor distância entre os pontos (h); alcance de dependência espacial (a).

Segundo Ferreira et al. (2013), os variogramas que são ajustados pelo modelo gaussiano são caracterizados por uma dependência espacial que apresenta baixas variações entre os vizinhos mais próximos e maiores variações para os vizinhos mais distantes, ainda dentro do alcance do semivariograma. Devido ao fato de que variáveis como altimetria e batimetria apresentam tais características, o uso da variável altimetria neste trabalho é justificado.

Conforme mencionado, Modis e Papaodysseus (2006) apresentaram a teoria e os resultados apenas para modelos Esféricos e Exponenciais. Como a variável utilizada neste trabalho é altimetria, se faz necessário uma pequena adaptação da metodologia proposta por estes autores, a saber:

(i) Abramowitz e Stegun (1972) apresentaram, a partir da Equação (3), utilizando a transformada de Fourier para a função de correlação do modelo gaussiano, a Equação (4), que é inversamente relacionada ao semivariograma.

$$R(\omega) = \exp\left(-\frac{3t^2}{a^2}\right)\cos(\omega t) \tag{4}$$

sendo ω a taxa de amostragem relacionada à frequência dos sinais e t um instante de amostragem.

ii) Ainda de acordo com Abramowitz e Stegun (1972), a função de densidade espectral de potência, que é o modelo que descreve o comportamento da função de correlação do modelo gaussiano, é dada pela Equação (5):

$$S(\omega) = \frac{1}{2} a \sqrt{\frac{\pi}{3}} \exp\left(-\frac{\omega^2 a^2}{12} \right) \tag{5}$$

iii) Como $S(\omega)$ tende a zero, devido à estabilização da função de correlação, $\exp\left(-\frac{w^2 a^2}{12} \right)$ tende a zero quando ω tende ao infinito. Entretanto, Journel e Huijbregts (1978) afirmam que o modelo gaussiano, por ser assintótico ao eixo, deve ter nulidade considerada em 5%. Logo $\omega \cong \frac{6}{a}$ **.**

iv) Decorre do teorema da taxa Nyquist que o tamanho amostral *T* é dado por $T \cong 0{,}5a$, o que corresponde a aproximadamente metade do alcance teórico, conforme obtido por Modis e Papaodysseus (2006).

v) Segundo Olea (1999), alguns modelos de variograma teórico não alcançam estabilização da curva no alcance teórico *a*, sendo o gaussiano um desses modelos. Este autor apresenta uma transformação do alcance teórico para o alcance prático a_p, dado por $a = \frac{\sqrt{3}}{3} a_p$.

vi) O tamanho amostral *T*, em função do alcance prático a_p é dado por $T \leq \frac{\pi\sqrt{3}}{18} a_p$.

Portanto, a maior distância entre dois pontos de uma grade regular quadrática de amostragem deve ser aproximadamente 30% do alcance prático. Isto significa que, na prática, uma primeira amostragem, chamada de amostragem experimental, deve ser feita para que a densidade da amostra possa ser estimada como 30% do alcance prático.

Dado que a maior distância entre dois pontos numa grade regular quadrática encontra-se nas diagonais, para se obter a distância máxima para os lados deve-se estabecer a relação entre a diagonal *d* e o lado *l* do quadro dada por $d = l\sqrt{2}$.

2.5 Caracterização dos dados

O conjunto de dados referente à região norte Americana, situada próxima à cidade de Treynor-Iowa, Estados Unidos, consistiu de 192.017 pontos. Após eliminação dos *outliers* desse banco de dados, seguindo a proposta feita por Santos et. al (2017), obteve-se um novo

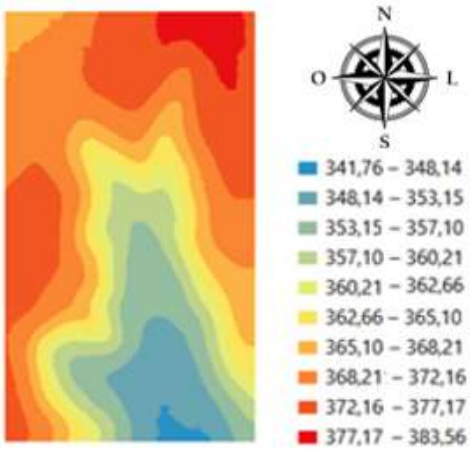
conjunto de dados contendo 4067 pontos. Posteriormente, este novo conjunto de dados foi reduzido 15 vezes (seguindo e verificando as condições de regularidade necessárias, conforme proposto por Modis e Papaodysseus, 2006), atingindo o número de 48 pontos para o tamanho de amostra.

Como critério de eficiência, considerou-se a variância média de krigagem (RMS), os coeficientes da Regressão Linear Simples (RLS) entre os valores preditos e observados da validação cruzada e a variância dos dados para a estimativa do parâmetro alcance na modelagem dos variogramas (VIEIRA, 2000; MODIS e PAPAODYSSEUS, 2006; SANTOS et al., 2011; FERREIRA et al., 2013; YAMAMOTO e LANDIM, 2013).

Para a análise computacional desse trabalho, foi utilizado o *software* ArcGIS 10.2.2 (ESRI, 2014). Para a realização das 8 reduções a partir dos 4067 pontos, adotou a seleção regular dos dados de altimetria, utilizando a ferramenta de amostragem regular do *software* ArcGIS. O primeiro passo desse processo foi definir o espaçamento em ambos os sentidos X e Y para executar a seleção regular. Assim, foi criada um grade intermediária de pontos baseando-se no espaçamento definido e, em seguida, obtendo o ponto da base de dados de altimetria mais próximo de cada ponto criado nesta grade intermediária. Posteriormente, a seleção desses pontos mais próximos foi feita, tendo como resultado um conjunto de dados de altimetria com uma amostra "quase regular". Esse resultado "quase" foi previamente comprovado com a ferramenta Ferramentas de Analisador da Proximidade do Toolbox, usando o comando Near (ESRI, 2014).

Conforme Figura 2, é adotado como mapa populacional da área estudada, a krigagem simples (Santos et al., 2011) dos 4067 pontos.

Figura 2 – Krigagem simples dos dados de altimetria obtidos por LiDAR Cloud de uma pequena bacia hidrográfica próximo da cidade de Treynor-Iowa, Estados Unidos.



Fonte: Santos et al. (2017)

3 RESULTADOS E DISCUSSÃO

Para atingir os objetivos propostos neste trabalho, utilizou-se um conjunto de dados sobre altimetria, relativo à cidade de Treynor-Iowa, nos Estados Unidos, onde o relevo é plano e levemente montanhoso, com uma altitude média de cerca de 363 metros e variância media de cerca de 62 metros quadrados.

A partir da amostra inicial contendo 4067 pontos foram feitas 15 reduções nos tamanhos amostrais, conforme Tabela 1.

Pela Tabela 1 é possível notar que as médias e variâncias não sofreram grandes variações, mesmo com a redução de 98,82 % do tamanho original da amostra, tendo este ultrapassado o limite mínimo recomendado por Yamamoto e Landim (2013). As reduções da amostra original foram feitas com base no espaçamento regular entre os pontos, removidos em grades quadráticas de lados 12m, 14m, 22m, 24m, 32m, 34m, 42m, 44m, 52m, 54m, 62m, 64m, 72m, 74m, 82m e 84m.

Tabela 1 - Apresentação dos tamanhos de amostragem, redução de amostras, médias, variâncias e espaçamento das grades regulares quadráticas do levantamento altimétrico de parte da região de Treynor-Iowa, Estados Unidos.

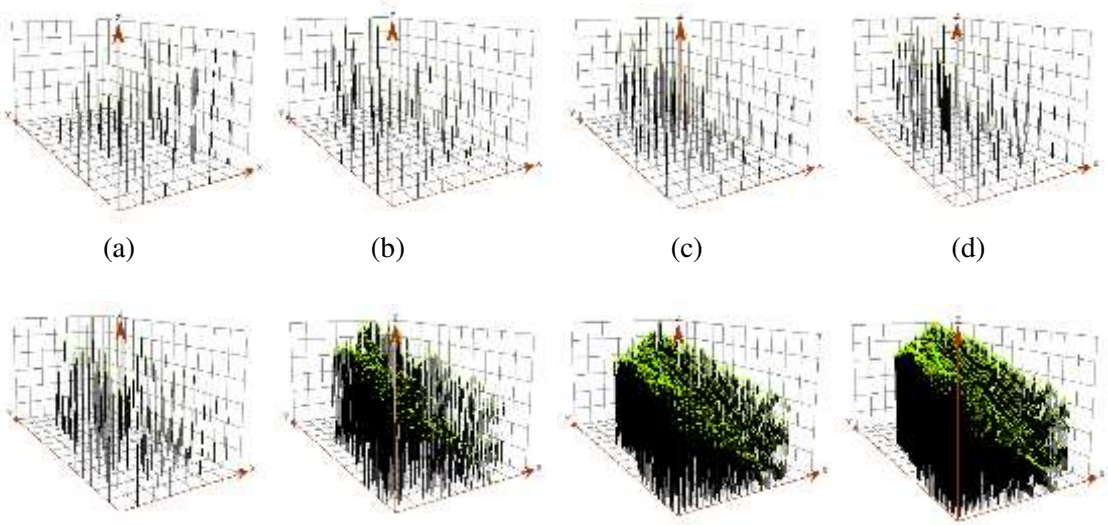
Tamanho amostral	Redução (%)	Média	Variância	Espaçamento (metros)
4067	0	363,62	62,02	"população"
2206	45,75	363,42	60,38	12
1619	60,19	363,43	60,89	14
663	83,69	363,43	61,80	22
558	86,28	363,46	60,34	24
315	92,25	363,43	64,46	32
280	93,12	363,49	61,47	34
183	95,50	363,51	61,61	42
167	95,89	363,36	62,47	44
117	97,12	363,28	60,74	52
115	97,17	363,68	64,29	54
85	97,91	363,59	62,07	62
81	98,01	363,73	66,55	64
64	98,43	363,57	60,83	72
57	98,59	363,15	64,48	74
48	98,82	363,60	59,12	82

Fonte: Mendes et al. (2019)

Pelos dados analíticos apresentados na Tabela 1 pode-se notar que a decisão sobre a representação de uma amostragem não pode ser considerada simplesmente por tamanho, média e variância dos dados.

É apresentado na Figura 3 as representações em três dimensões de algumas das grades regulares de amostra visando mostrar a perda da representatividade da população quando a amostragem não mostra um tamanho adequado, de acordo com o critério utilizado.

Figura 3 – Representação tridimensional de grade regular quadrática do levantamento altimétrico próximo da cidade de Treynor-Iowa, Estados Unidos. Os tamanhos de amostragem são (a) 57 pontos, (b) 64 pontos, (c) 81 pontos, (d) 85 pontos, (e) 115 pontos, (f) 663 pontos, (g) 1619 pontos e (h) 2206 pontos.



(a) (b) (c) (d)

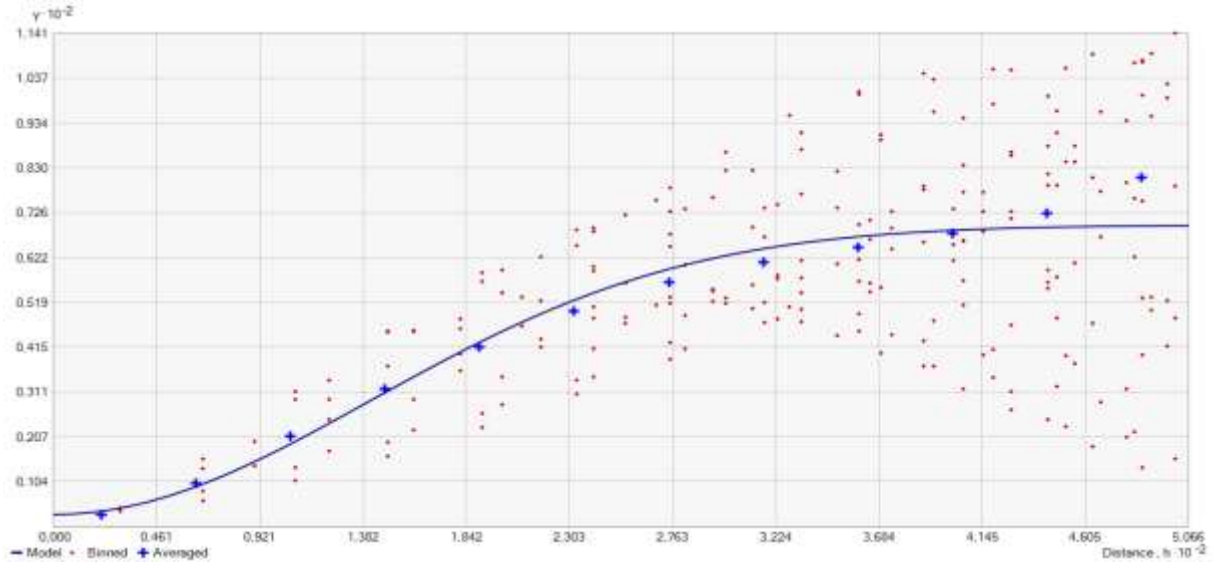
(e) (f) (g) (h)

Fonte: Mendes et al. (2019)

Conforme esperado, de acordo com Oliveira et al. (2014), aumentando o tamanho de amostra a visualização gráfica tridimensional simples dos dados mostra o comportamento real da população, o que pode ser observado na Figura 3 (g).

É apresentado na Figura 4 o comportamento do semivariograma para todos os 4067 pontos utilizados no estudo. Os demais semivariogramas, embora não tenham sido apresentados, seguiram o mesmo comportamento. Pode-se notar que o modelo ajustado ao semivariograma experimental foi o modelo gaussiano, apresentado na Equação 3.

Figura 4 – Semivariograma experimental (sinais "+") e modelo gaussiano ajustado (linha contínua) para a dependência espacial.



Fonte: Mendes et al. (2019)

Utilizando-se a metodologia proposta por Modis e Papaodysseus (2006) adaptada ao modelo gaussiano de dependência espacial ajustado aos dados desse trabalho, determinou-se o tamanho ideal de amostragem: 115 pontos, sendo a distância lateral entre os pontos cerca de 54 metros (ou 76 metros na diagonal), correspondendo a 21,38% (ou 30,09% considerando a diagonal), em relação ao alcance prático de 252,6 metros.

Os erros quadráticos médios, os tamanhos de amostragem e o espaçamento das grades quadráticas para a variável altimetria são apresentados na Tabela 2.

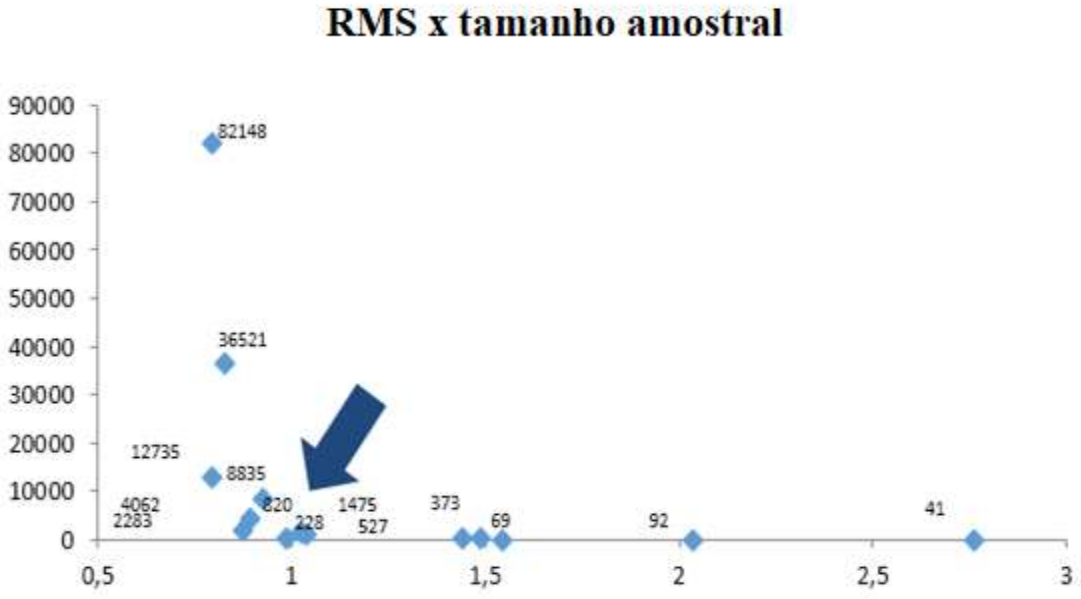
Tabela 2 – Apresentação dos erros quadráticos médios (RMS), os tamanhos de amostra e o espaçamento das grades quadráticas do levantamento altimétrico de parte da região de Treynor-Iowa, Estados Unidos.

RMS	Tamanho amostral	Espaçamento (metros)
0,8943	4067	"população"
0,9792	2206	12
0,8913	1619	14
0,8223	663	22
1,2323	558	24
1,0659	315	32
0,9154	280	34
1,9105	183	42
1,1368	167	44
1,6522	117	52
1,2751	115	54
1,7468	85	62
1,6617	81	64
2,6152	64	72
2,4773	57	74
2,2384	48	82

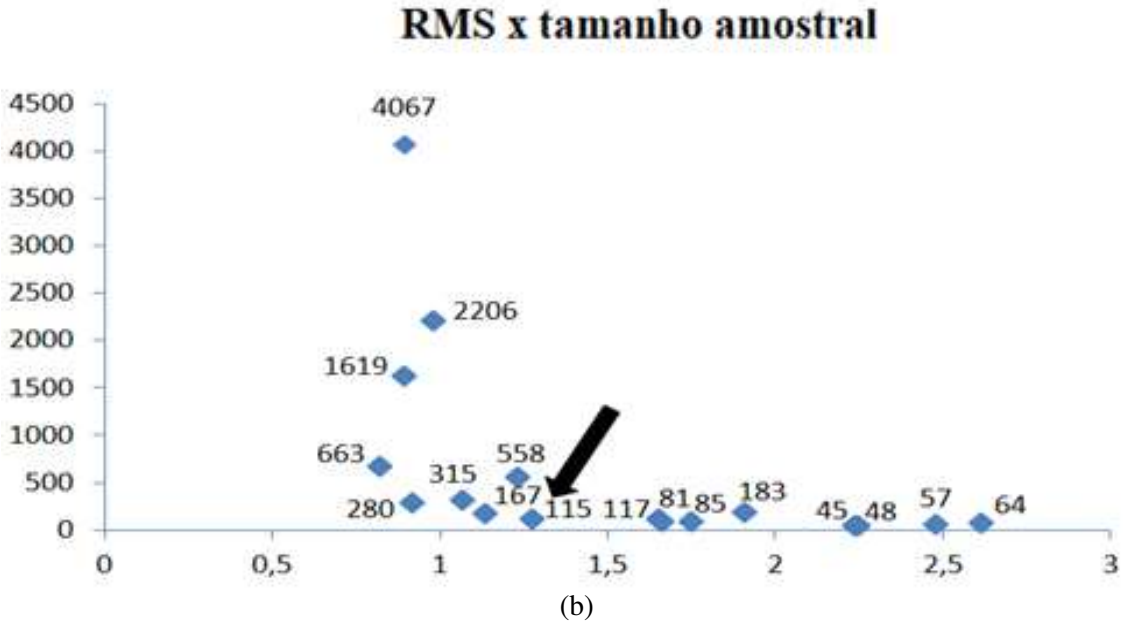
Fonte: Mendes et al. (2019)

O gráfico do tamanho de amostra em função do RMS é apresentado na Figura 5.

Figura 5 – Representação gráfica da relação entre o erro quadrático médio (RMS) e o tamanho de amostra. (a) conjunto de dados com *outliers*, (b) conjunto de dados sem *outliers*.



(a)



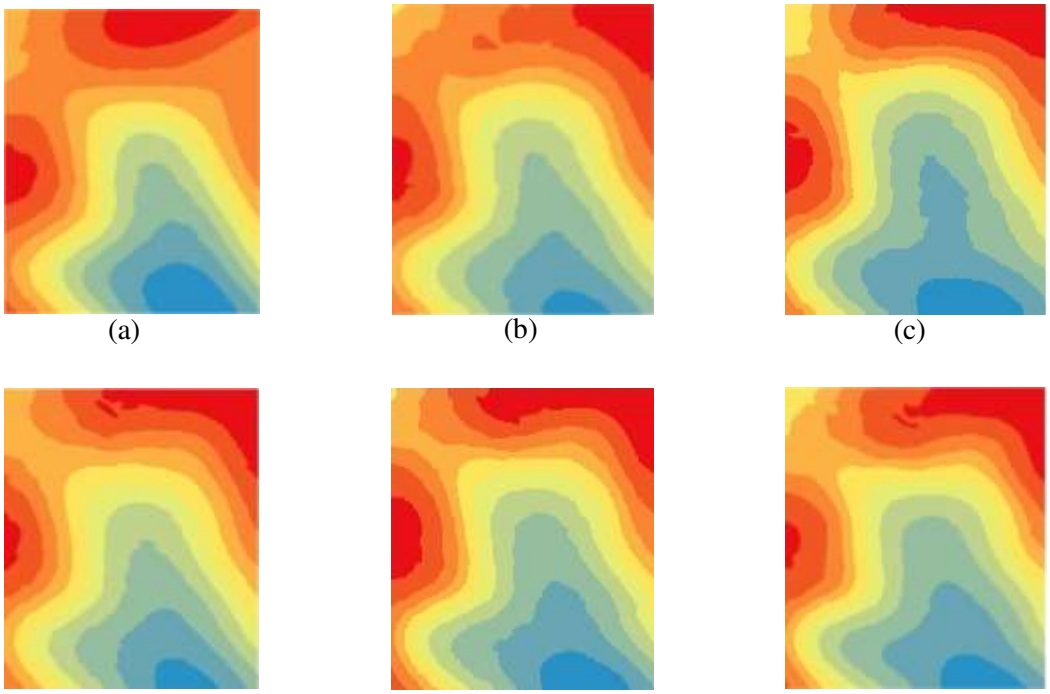
(b)

Fonte: Mendes et al. (2019)

Na presença de *outliers*, o tamanho de amostra ideal é de 228 pontos, indicado pela seta na Figura 5a. Já na ausência de *outliers*, o tamanho de amostra ideal é 115 pontos, indicado pela seta na Figura 5b. Esses resultados indicam que é necessário aproximadamente o dobro de pontos amostrais para se obter um tamanho de amostra ideal quando o conjunto de dados contém *outliers*.

São apresentados na Figura 6 os mapas obtidos através da interpolação por krigagem simples para os dados de altimetria.

Figura 6 – Krigagem simples para os dados de altimetria. Os tamanhos de amostras são (a) 48 pontos, (b) 57 pontos, (c) 64 pontos, (d) 81 pontos, (e) 85 pontos, (f) 115 pontos.



(a) (b) (c)

(d) (e) (f)

Fonte: Mendes et al. (2019)

É possível observar pela Figura 6f que o mapa populacional é representado satisfatoriamente pelo tamanho de amostra de 115 pontos, que, conforme apresentado na Figura 5b, corresponde ao ponto de estabilização da curva.

4 CONCLUSÕES

Na presença de *outliers*, o tamanho de amostra ideal para amostras georreferenciadas que usam uma grade quadrática regular, na qual o modelo de dependência espacial é o gaussiano é superior ao tamanho de amostra ideal, na ausência de *outliers*. Portanto, para a reconstrução de mapas populacionais de variáveis que contenham *outliers* e que satisfazem as condições de regularidade necessárias em geoestatística, é necessário amostrar mais pontos tornando a amostragem mais onerosa.

REFERÊNCIAS BIBLIOGRÁFICAS

ABRAMOWITZ, M.; STEGUN, I. A. **Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables**. Washington, D.C.: U.S. Government Printing Office, 1972, 1046 p.

APPICE, A.; GUCCIONE, P.; MALERBA, D.; CIAMPI, A. Dealing with temporal and spatial correlations to classify outliers in geophysical data streams. **Information Science**, Alberta, v. 28, n. 5, p. 62-80, 2014.

BARUA, S.; ALHAJJ, R. High performance computing for spatial outliers detection using parallel wavelet transform. **Intelligent Data Analysis**, Alberta, v. 11, n. 6, p. 707-730, 2007.

CLARK, I.; HARPER, W. V. **Practical geostatistics 2000**. Columbus: Ecosse North America L1c, 2000. 416 p.

ENVIRONMENTAL SYSTEMS RESEARCH INSTITUTE - ESRI. **ArcGIS 10.2 for Desktop**, 2014.

FARACO, M. A; URIBE-OPAZO, M. A.; SILVA, E. A. A. da; JOHANN, J. A.; BORSSOI, J. A. Seleção de modelos de variabilidade espacial para elaboração de mapas temáticos de atributos físicos do solo e produtividade da soja. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 32, n. 2, p. 463-476, 2008.

FERREIRA, I. O.; SANTOS, G. R.; RODRIGUES, D. D. Estudo sobre a utilização adequada da krigagem na representação computacional de superfícies batimétricas. **Revista Brasileira de Cartografia**, Rio de Janeiro, n. 65/5, p. 831-842, 2013.

GOMES, N. M.; MARCIANO, A. S.; ROGÉRIO, C. M.; ALVES, M. F.; MARA, P. O. Métodos de ajuste e modelos de semivariograma aplicados ao estudo da variabilidade espacial de atributos físico-hídricos do solo. **Revista Brasileira de Ciência do Solo,** Viçosa, v. 31, n. 3, p. 435-443, 2007.

GOMES, J. B. V.; BOLFE, E. L.; CURI, N.; FONTES, H. R.; BARRETO, A. C.; VIANA, R. D. Variabilidade espacial de atributos de solos em unidades de manejo em área piloto de produção integrada de coco. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 32, n. 6, p. 2471-2482, 2008.

HOAGLIN, D.C.; MOSTELLER, F.; TUKEY, J.W. **Understanding robust and exploratory data analysis.** Wiley, New York, 1983.

HÖHLE, J.; HÖHLE, M. Accuracy assessment of digital elevation models by means of robust statistical methods. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 64, n. 4, p. 398**-**406, 2009.

JOURNEL, A. G., HUIJBREGTS, C. J. **Mining Geostatistics**. London: Academic Press, 1978. 600 p.

MENDES, A.; SANTOS, G. R.; EMILIANO, P. C.; KALEITA, A. M.; FERREIRA, M. P. Estimação do tamanho amostral na geoestatística usando um modelo de variograma gaussiano na presença de outliers. **Energia na Agricultura,** Botucatu, v. 34, n. 3, p. 429-440, 2019.

MODIS, K.; PAPAODYSSEUS, K. Theoretical Estimation of the Critical Sampling Size for Homogeneous Ore Bodies with Small Nugget Effect. **Mathematical Geology**, v. 38, n. 4, p. 489**-**501, 2006.

MORETTIN, P. A.; TOLOI, C. M. C. **Análise de séries temporais.** 2. ed. São Paulo: Edgard Blücher, 2006. 538p.

OLEA, R. A. **Geostatistics for engineers and earth scientists**. London: Kluwer Academic Publishers, 1999. 303 p.

OLIVEIRA, M. S. de; BEARZOTI, E.; VILLAS BOAS, F. L.; NOGUEIRA, D. A.; NICOLAU, L. A.; OLIVEIRA, H. S. S. de. **Introdução à Estatística**. 2. ed. Lavras: UFLA, 2014. 462 p.

QIAO, C.; HAIBO, H.; HONG, M. Spatial outlier detection based on iterative selforganizing learning model. **Neurocomputing,** v. 117, p. 161-172, 2013.

RENCHER, A. C., SCHAALJE, G. B**. Linear Models in Statistics.** 2. ed. New Jersey: John Wiley & Sons, 2008. 672 p.

SANTOS, A. M. R. T.; SANTOS, G. R.; EMILIANO, P. C.; MEDEIROS, N. G.; KALEITA, A. L.; PRUSKI, L. O. S. Detection of inconsistencies in geospatial data with Geostatistics. **Boletim de Ciências Geoedésicas**, Curitiba, v. 23, n. 2, p. 296-308, 2017.

SANTOS, A. M**.** R. T. ***Outliers*** **em variáveis geoespaciais: proposições utilizando Geoestatística.** 2016. 82 p. Tese (Doutorado em Engenharia Civil). Viçosa: UFV.

SANTOS, G. R.; OLIVEIRA, M. S.; LOUZADA, J. M.; SANTOS, A. M. R. T. Krigagem Simples versus Krigagem Universal: qual o preditor mais preciso? **Revista Energia na Agricultura**, Botucatu, v. 26, n. 2, p. 49-55, 2011.

SILVA, A. N.; SANTOS, G. R.; SANTOS, N. T.; PRUSKI, F. F.; ILAMBWETSI, P. S. Detecção de *outliers* em séries espaço-temporais: análise de precipitação em Minas Gerais. **Revista da Estatística**, Ouro Preto, v. 6, p. 121-131, 2012.

SOUZA, Z. M. SOUZA, G. S.; JÚNIOR, J. M.; PEREIRA, G. T. Número de amostras na análise geoestatística e na krigagem de mapas de atributos do solo. **Revista Ciência Rural**, Santa Maria, v. 44, n. 2, p. 261-268, 2014.

VÁSAT, R; HEUVELINK, G. B. M.; BORUVKA, L. Sampling design optimization for multivariate soil mapping. **Geoderma**, v. 155, n. 3-4, p. 147–153, 2010.

VIEIRA, S. R. Geoestatística em estudos de variabilidade espacial de propriedades do solo. In: NOVAIS, R. F.; ALVAREZ, V. H.; SCHAEFER, G. R. **Tópicos em ciência do solo**, Viçosa, v. 1, p. 1-54, 2000.

YAMAMOTO, J.; LANDIM, P. **Geoestatística: Conceitos e Aplicações.** Oficina de Textos: São Paulo, 2013. 216 p.

CAPÍTULO 3 - AVALIAÇÃO DA ACURÁCIA DOS RECURSOS COMPUTACIONAIS NA ESTIMAÇÃO INTERVALAR DOS PARÂMETROS DO SEMIVARIOGRAMA GAUSSIANO AJUSTADO A DIFERENTES TAMANHOS DE AMOSTRA

RESUMO

A geoestatística vem sendo amplamente utilizada na modelagem de fenômenos espaciais, devido ao grande número de *softwares* que realizam sua interpolação, sendo a mais usual denominada simplesmente por krigagem. Com o intuito de identificar e mapear padrões espaciais da superfície terrestre, a geoestatística permite determinar se existe autocorrelação espacial entre dados de pontos que são apresentados por meio de mapas obtidos por krigagem. De forma geral, a krigagem consiste em combinar os parâmetros do modelo variográfico com os dados para a produção de superfícies de predição, em sua forma automática (auto-ajuste), o que, por sua vez, pode produzir resíduos que fornecem informações distorcidas a cerca dos dados. Portanto, o presente trabalho objetivou-se identificar qual *software* (ArcGIS ou R) é o mais adequado ao comparar as estimativas dos parâmetros do modelo variográfico gaussiano, as estimativas da média e as da variância dos resíduos provenientes da validação cruzada, por meio da construção de intervalos de confiança, ao utilizar a mesma base de dados (dados de altimetria). Como resultado, pode-se concluir que as estimativas da média e da variância dos resíduos obtidos pelo *software* R foram aproximadamente 4 e 5 vezes menores, respectivamente, comparadas ao *software* ArcGIS.

Palavras-chave: Geoestatística. *Softwares* R e ArcGIS. Intervalo de confiança.

ABSTRACT

Geostatistics has been widely used in the modeling of spatial phenomena, due to the large number of software that perform its interpolation, being the most usual denominated simply by kriging. In order to identify and map spatial patterns of the terrestrial surface, Geostatistics allows to determine if there is spatial autocorrelation between data points that are presented through maps obtained by kriging. In general, kriging consists of combining the parameters of the variographic model with the data for the production of prediction surfaces, in their automatic form (self-tuning), which, in turn, can produce residues that provide distorted information to about the data. Therefore, the present work aims to identify which software (ArcGIS or R) is most appropriate when comparing the estimates of the parameters of the Gaussian variographic model and the estimates of the mean and the variance of the residuals from the cross validation, through the construction of confidence intervals, for the same database (altimetry data). As a result, it can be concluded that the mean and variance estimates of residues obtained from cross-validation obtained by R were about 4 times and 5 times lower, respectively, compared to ArcGIS.

Keywords: Geostatistics. Software R and ArcGIS. Confidence interval.

1 INTRODUÇÃO

Atualmente, há uma variedade de *softwares* estatísticos, proprietários e livres disponíveis para a comunidade científica. Para utilização de *software*s proprietários, como o ArcGIS, se faz necessário a aquisição de sua licença para uso, implicando em custos. Por outro lado, os ambientes de programação livre, como o *software* R e seu console RStudio, facilitam o acesso do público, além de propiciar ao pesquisador o acompanhamento de todos os passos efetuados ao longo de suas análises (SMOLSKI, F. M. S. et al, 2018).

Segundo Batista et al. (2015), devido ao grande número de *software*s que realizam interpolação via geoestatística, a geoestatística tem se mostrado bastante eficiente quanto ao uso de dados georreferenciados e vem sendo amplamente utilizada na modelagem de fenômenos espaciais.

A melhor maneira de se obter a modelagem da dependência espacial é feita através do semivariograma (VIEIRA, 2000). Posteriormente, predições para valores não amostrados são obtidos, sendo a krigagem o interpolador geoestatístico mais utilizado por apresentar características ótimas (estimativas sem viés e com variância mínima) (SANTOS et al., 2011; YAMAMOTO e LANDIM, 2013).

Com o desenvolvimento computacional, alternativas para o ajuste do semivariograma experimental foram surgindo, como o método dos mínimos quadrados ordinários (OLS) e da máxima verossimilhança (REML).

Após a construção do semivariograma experimental, deve-se optar por um modelo teórico que melhor se ajusta aos dados amostrais. A qualidade dos ajustes dos modelos utilizados pode ser avaliada, entre outros, pelo critério de informação de Akaike (AIC) (Akaike, 1974), validação cruzada ou coeficiente de determinação (R^2).

Dentre as opções disponíveis para ajuste de modelos de semivariograma e estimação de parâmetros, destacam-se o *software* R (The R Foundation for Statistical Computing, Viena, Áustria; http://www.r-project.org), e o *software* ArcGIS (ESRI, 2014).

Tais análises geoestatísticas, já consolidadas no meio científico e profissional, devem ser melhor investigadas, já que estes procedimentos são realizados automaticamente, podendo comprometer no resultado final. De acordo com Ferreira et al. (2013) o procedimento de ajuste não deve ser feito de forma direta e/ou automática, mas sim de forma interativa.

Dessa forma, o objetivo deste trabalho foi verificar se existem diferenças nos resultados específicos de uma análise geoestatística entre os *software*s R e ArcGIS, para o mesmo conjunto de dados. De forma mais específica, objetivou-se analisar os intervalos de

confiança construídos para as estimativas dos parâmetros do modelo gaussiano ajustado aos dados de altimetria e também para as estimativas da média e da variância dos resíduos provenientes da validação cruzada, por meio dos *software*s supracitados.

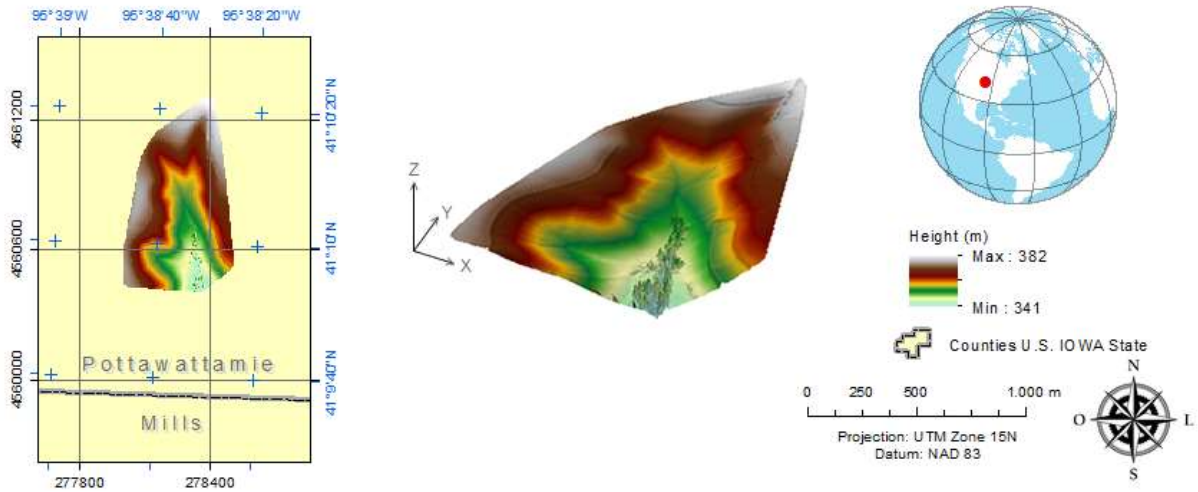
2 MATERIAL E MÉTODOS

A seguir é apresentada a descrição da região e da proposta do estudo e a caracterização dos dados.

2.1 Descrição da região de estudo

A área estudada compreende uma parcela de 34,3 hectares da cidade de Treynor, situada no município de Pottawattamie, no Estado de Iowa, Estados Unidos. A região estudada é delimitada pelas latitudes 41º10’23” N e 41º09’53” N e longitudes 95º38’24’’ W a 95º38’47” W, como mostrado na Figura 1.

Figura 1 – Representação da área de estudo, próximo da cidade de Treynor-Iowa, Estados Unidos



Fonte: Santos et al. (2017)

Os dados altimétricos utilizados nesse trabalho são de um mapeamento LiDAR, referenciados ao sistema geodésico NAD 83 (Datum norte-americano de 1983) e representados no sistema de projeção UTM (Universal Transverse Mercator), espaçados horizontalmente e verticalmente, de 1,7 e 1,2 metros, respectivamente.

2.2 Caracterização dos dados

O conjunto de dados consistiu de 192.017 pontos referentes à região norte Americana, situada próxima à cidade de Treynor-Iowa, Estados Unidos. Utilizando a proposta feita por Santos et al. (2017) eliminou-se os *outliers* desse banco de dados e obteve-se um novo conjunto contendo 4067 pontos. Conforme proposto por Mendes et al. (2019), reduziu-se este

novo conjunto de dados 15 vezes, seguindo e verificando as condições de regularidade necessárias, conforme proposto por Modis e Papaodysseus (2006). Adicionalmente, algumas estatísticas descritivas foram obtidas para cada amostra.

A modelagem do semivariograma experimental, realizada por meio do método dos mínimos quadrados ordinários (OLS), ajustou aos dados o modelo gaussiano, caracterizado pela Equação (1):

$$\gamma(\mathrm{h}) = \begin{cases} 0\,; & h = 0 \\ C_0 + C_1\left[1 - \exp\left(-\dfrac{3h^2}{a^2}\right)\right]\,; & 0 < h < a \\ C_0 + C_1\,; & h > a \end{cases} \qquad (1)$$

No modelo gaussiano da Equação (1), C_0 representa o efeito pepita, $C_0 + C_1$ representa o patamar; C_1 representa a contribuição, h vetor distância entre os pontos e a o alcance da dependência espacial.

A krigagem adotada para interpolação foi a krigagem simples, conforme recomendações de Santos et al. (2011). Considerou-se o Erro Quadrático Médio da validação cruzada (RMS) como critério de comparação (VIEIRA, 2000; MODIS e PAPAODYSSEUS, 2006; SANTOS et al., 2011; FERREIRA et al., 2013; YAMAMOTO e LANDIM, 2013).

Para a análise computacional desse trabalho, foram utilizados os *software*s ArcGIS 10.2 (ESRI, 2014) – ferramenta *Geostatistical Wizard: kriging/Cokriging*, e R (R Development Core Team, 2016) – através do pacote *geoR* (RIBEIRO JUNIOR e DIGGLE, 2001).

3 RESULTADOS E DISCUSSÃO

A partir da amostra inicial contendo 4067 pontos foram feitas 15 reduções nos tamanhos amostrais, conforme Tabela 1.

Tabela 1 – Apresentação dos tamanhos de amostragem, médias, variâncias e espaçamento das grades regulares quadráticas do levantamento altimétrico de parte da região de Treynor-Iowa, Estados Unidos.

Tamanho amostral	Média	Variância	Espaçamento
2206	363,42	60,38	12
1619	363,43	60,89	14
663	363,43	61,80	22
558	363,46	60,34	24
315	363,43	64,46	32
280	363,49	61,47	34
183	363,51	61,61	42
167	363,36	62,47	44
117	363,28	60,74	52
115	363,68	64,29	54
85	363,59	62,07	62
81	363,73	66,55	64
64	363,57	60,83	72
57	363,15	64,48	74
48	363,60	59,12	82

Fonte: Mendes et al. (2019)

Nota-se, pela Tabela 1, que as médias e variâncias não sofreram grandes alterações mesmo com a redução do tamanho amostral tendo ultrapassado o limite mínimo recomendado por Yamamoto e Landim (2013).

Por meio do método dos mínimos quadrados ordinários (OLS), um modelo de dependência espacial gaussiano foi ajustado ao semivariograma experimental, utilizando ambos os *softwares*, para cada amostra, e, assim, definidos os parâmetros do modelo gaussiano: Pepita (C_0), Patamar ($C_0 + C_1$) e Alcance (a), conforme Tabela 2.

Tabela 2 – Estimativa dos parâmetros do semivariograma gaussiano para os dois *software*s utilizados.

Tamanho amostral	*Software* ArcGis			*Software* R		
	Pepita	Patamar	Alcance	Pepita	Patamar	Alcance
2206	2,71938	63,73926	336,0504	2,0	60,09237	320,9166
1619	1,81499	62,2406	316,0028	1,6	62,57854	322,7482
663	2,21401	65,82223	339,1942	1,5	62,96064	324,6505
558	2,5816	60,90041	323,0409	2,5	60,07251	323,2407
315	3,03345	68,8853	354,6073	2,0	64,77803	330,5078
280	2,24651	61,8791	331,6313	2,2	61,12546	332,5892
183	4,27513	61,6716	340,2037	4,5	61,31979	352,5509
167	2,37744	67,16929	338,7016	2,5	68,07638	347,554
117	2,61158	64,03913	317,3036	3,8	60,87864	312,195
115	2,56458	68,65202	341,3317	3,0	66,53306	333,0184
85	3,15614	63,08294	320,3926	0,2	67,70519	320,7539
81	4,33917	70,14134	383,3784	1,5	72,37984	374,7419
64	6,14022	57,51324	323,3136	5,0	58,21629	324,0048
57	5,42983	65,98168	347,0119	4,7	66,74065	349,9399
48	3,91397	60,39424	330,8446	3,9	67,94309	336,895

Fonte: Elaborada pelo autor

Vale resaltar que cada modelo foi validado pela técnica da validação cruzada a fim de avaliar a qualidade das estimativas obtidas pelo modelo gaussiano, conforme recomendado por Vieira (2000).

As medidas apresentadas pela validação cruzada são de grande importância, uma vez que a literatura científica cita tais indicadores como os mais relevantes na tarefa de avaliar a qualidade de ajuste do semivariograma. Segundo Vieira (2000), devemos observar o comportamento das estimativas da média dos resíduos, que deve ser próxima de 0, e das estimativas da variância dos resíduos, considerada variância de Krigagem por muitos autores da geoestatística, que deve ser o menor possível.

Na Tabela 3 são apresentadas as estimativas da média e da variância dos resíduos da validação cruzada para cada amostra.

Tabela 3 – Estimativas da média e da variância dos resíduos da validação cruzada para os dois *software*s utilizados.

Tamanho amostral	*Software* ArcGIS		*Software* R	
	Média	Variância	Média	Variância
2206	0,02043487	0,9792002	-0,000757	0,024548
1619	0,036815	0,8913326	-0,001559	0,0270226
663	0,08526815	0,8223807	-0,004198	0,0377395
558	0,08808083	1,232383	-0,004016	0,0566018
315	0,1480776	1,065937	-0,006568	0,065055
280	0,1103602	0,9154581	-0,005464	0,065419
183	0,1577239	1,910541	-0,007307	0,1481816
167	0,08888711	1,136806	0,009461	0,1035248
117	0,0286094	1,652218	0,003214	0,1629191
115	-0,01680914	1,275172	0,003505	0,1276275
85	-0,07504598	1,746814	0,001057	0,1639137
81	-0,02637911	1,661691	0,005365	0,1712779
64	-0,1458565	2,615209	0,060049	0,3201257
57	-0,1936786	2,477332	0,064424	0,3186674
48	-0,3326648	2,238366	0,107101	0,3037283

Fonte: Elaborada pelo autor

Pela Tabela 3 é possível notar que tanto as estimativas da média dos resíduos quanto as estimativas da variância dos resíduos, obtidas utilizando o *software* R estão mais próximas dos valores de referência, conforme recomendado por Vieira (2000). Esses resultados mostram que em média o *software* R é mais preciso ao ser comparado ao *software* ArcGIS por apresentar menor variabilidade.

As estimativas da média e da variância dos resíduos apresentadas na Tabela 3 foram submetidas ao teste de normalidade de Shapiro-Wilk (1965), sendo ambas significativas a 1%, para os dois *softwares*. Também foram significativas a 1% pelo teste de Shapiro-Wilk (1965) as estimativas dos parâmetros do semivariograma obtidas pelos *softwares* ArcGIS e pelo R.

Adicionalmente, com o intuito de verificar a similaridade entre as estimativas dos parâmetros do semivariograma e da validação cruzada produzidas pelos *software*s ArcGIS e R, foram construídos intervalos de $100(1 - \alpha)\%$ de confiança para cada um dos parâmetros "θ", IC(θ,α), uma vez que tais parâmetros não são conhecidos na população. Nesse sentido, ao invés de estimar o parâmetro "θ" por um único valor, considera-se um intervalo de estimativas prováveis para cada parâmetro "θ". O quanto estas estimativas são prováveis é determinado pelo coeficiente de confiança $(1 - \alpha)$, com $\alpha \in (0,1)$. Dessa forma, pode-se ter $100(1 - \alpha)\%$ de certeza de que o verdadeiro valor do parâmetro está dentro do intervalo com $(1 - \alpha)\%$ de confiança e n graus de liberdade, aplicando-se:

$$IC(\theta,\alpha)=\left[\overline{X}-t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}},\overline{X}+t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right] \quad (2)$$

Sendo "θ" o parâmetro populacional de interesse; $\overline{X}$ e S são, respectivamente, a média e o desvio padrão amostral do parâmetro de interesse; α é o nível de significância; $t_{\frac{\alpha}{2}}$ é o valor tabelado na distribuição t de Student e n é o tamanho da amostra.

Levando em consideração as 15 amostras apresentadas na Tabela 3, adotando $\alpha = 1\%$ e sob a hipótese nula de que as médias das estimativas dos parâmetros obtidos no ArcGIS e no R são estatisticamente iguais, foram construídos intervalos com 99% de confiança, conforme Tabela 4.

Tabela 4 – Intervalos com 99% de confiança para os parâmetros do semivariograma gaussiano e da validação cruzada.

	Software ArcGIS			*Software* R		
	IC (99%) [1]			IC (99%) [1]		
	LI [2]	LS [3]	Amplitude	LI [2]	LS [3]	Amplitude
Efeito Pepita	0,1659	6,5023	6,3364	-0,7318	6,3318	7,0636
Patamar	55,2763	73,2667	17,9904	53,8891	73,9945	20,1054
Alcance	292,9156	381,8061	88,8905	291,3688	374,4036	83,0348
Média dos resíduos	-0,3799	0,3504	0,7303	-0,0716	0,1072	0,1788
Variância dos resíduos	0,0062	3,1021	3,0959	-0,1335	0,4370	0,5705

[1] IC (99%) = Intervalo com 99% de confiança; [2] LI = Limite Inferior; [3] LS = Limite Superior
Fonte: Elaborada pelo autor

Pela Tabela 4 pode-se notar que os *software*s tem comportamento similar na estimativa do efeito Pepita devido a existência de uma intersecção entre esses intervalos. Para os parâmetros Patamar e Alcance, o comportamento também se manteve similar. Tal comportamento se justifica pois ambos os *software*s utilizam a mesma fórmula na estimativa desses parâmetros.

Ao se comparar os intervalos de confiança dos parâmetros da validação cruzada apresentados na Tabela 4, nota-se que o *Software* R merece destaque. Para a média dos resíduos, o erro da autovalidação no R chega a ser, aproximadamente, quatro vezes menor. Já para a variância dos resíduos, o erro da autovalidação no R chega a ser, aproximadamente, cinco vezes menor.

De acordo com Esri (2004a), na validação cruzada, o erro associado a cada ponto amostrado é obtido pela diferença entre o valor verdadeiro e sua respectiva estimativa. Vale resaltar que, por default, a estimação no *software* ArcGIS utiliza no máximo os 12 pontos

amostrados mais próximos, dentro da área de busca (alcance prático), enquanto que o *software* R utiliza todos os pontos dentro dessa área de busca. Portanto, o *software* R produz estimativas mais acuradas para os parâmetros da validação cruzada.

4 CONCLUSÕES

Os *softwares* ArcGIS e R apresentaram comportamento similar na estimação dos parâmetros do modelo de semivariograma gaussiano para a mesma base de dados de altimetria. Porém, ao se comparar as estimativas da média e da variância dos resíduos provenientes da validação cruzada foi possível concluir que o *software* R apresentou uma menor variabilidade, sendo 4 e 5 vezes menor a amplitude do intervalo de confiança construído para a média e para a variância dos resíduos, respectivamente, comparado ao ArcGIS.

REFERÊNCIAS BIBLIOGRÁFICAS

AKAIKE, H. **A new look at the statistical model identification, IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716–723, 1974.

ArcGIS. An overview of the geodatabase, 2014. Disponível em: <http://resources.arcgis.com/en/help/main/10.2/index.html#//003100000032000000>. Acessado em 20 de agosto de 2018.

CLEMENTE, M. K. **Uso de dados SRTM para análise do relevo do Parque Nacional de Aparados da Serra/SC-RS**. 2016. 22 p. Especialização (Especialista em Análise Ambiental) – Programa de Pós-Graduação da Universidade Federal do Paraná, Curitiba.

ENVIRONMENTAL SYSTEMS RESEARCH INSTITUTE - ESRI. **ArcGIS 10.2 for Desktop**, 2014.

ESRI. USING ArcGIS Geostatistical Analyst. **Redlands**, CA – USA: ESRI Press, 2004a. 300p.

FERREIRA, I. O.; SANTOS, G. R. RODRIGUES, D. D. Estudo sobre a utilização adequada da krigagem na representação computacional de superfícies batimétricas. **Revista Brasileira de Cartografia**, Rio de Janeiro, v. 65, n.5, p. 831-842, 2013.

MENDES, A.; SANTOS, G. R.; EMILIANO, P. C.; KALEITA, A. M.; FERREIRA, M. P. Estimação do tamanho amostral na geoestatística usando um modelo de variograma gaussiano na presença de outliers. **Energia na Agricultura,** Botucatu, v. 34, n. 3, p. 429-440, 2019.

MODIS, K.; PAPAODYSSEUS, K. Theoretical Estimation of the Critical Sampling Size for Homogeneous Ore Bodies with Small Nugget Effect. **Mathematical Geology**, v. 38, n. 4, p. 489**-**501, 2006.

R DEVELOPMENT CORE TEAM. **R: A language and environment for statistical computing R Foundation for Statistical Computing**, Vienna. Disponível em:< http://www.r-project.org >. Acessado em 20 agosto de 2018.

RIBEIRO, Jr. P. J; DIGGLE, P. J. GeoR: A Package for Geostatistical Analysis. **R NEWS**, v. 1, n. 2, p. 14-18, 2001.

SANTOS, A. M. R. T.; SANTOS, G. R.; EMILIANO, P. C.; MEDEIROS, N. G.; KALEITA, A. L.; PRUSKI, L. O. S. Detection of inconsistencies in geospatial data with Geostatistics. **Boletim de Ciências Geoedésicas**, Curitiba, v. 23, n. 2, p. 296-308, 2017.

SANTOS, G. R.; OLIVEIRA, M. S.; LOUZADA, J. M.; SANTOS, A. M. R. T. Krigagem simples versus Krigagem Universal: qual o preditor mais preciso? **Revista Energia na Agricultura**, Botucatu, v. 26, n. 2, p. 49-55, 2011.

SHAPIRO, S. S.; WILK, M. B. An Analysis of Variance Test for Normality. **Biometrika Trust**, London, v. 52, p. 591-609.

SILVA, XAVIER. **Geoprocessamento para Análise Ambiental**. Rio de Janeiro: Ed. do Autor, 2001.

SMOLSKI, F. M. S.; BATTISTI, I. E.; CHASSOT, T.; REIS, D. I.; KASZUBOWSKI, E.; RIEGER, D. S. Capacitação em análise estatística de dados com uso do *software* livre R. **Revista Ciência em Extensão**, v. 14, n. 3, p. 123-134, 2018.

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

YAMAMOTO, J. K., LANDIM, P. M. B. **Geoestatística: conceitos e aplicações**. Oficina de Textos: São Paulo, 2013, 216 p.

CONCLUSÕES GERAIS

O presente trabalho apresenta uma metodologia que associa a geoestatística ao teorema da taxa Nyquist para obtenção da estimativa do tamanho ideal de amostragem.

Essa estimativa, de grande importância para o planejamento e tomada de decisão, especialmente quando se objetiva a reconstrução total da população, está relacionada ao alcance, um dos parâmetros obtidos na modelagem da dependência espacial.

Para o modelo gaussiano de dependência espacial ajustado aos dados de altimetria deste estudo, tal metodologia se mostrou satisfatória, tanto na presença quanto na ausência de *outliers*, conforme os resultados apresentados nos três capítulos do trabalho.

No capítulo 1, para um conjunto de dados reais de altimetria, sabidamente sem *outliers*, concluiu-se que a distância máxima entre os pontos da grade regular quadrática é de aproximadamente 30% do alcance prático observado no semivariograma da primeira amostragem experimental.

Já no capítulo 2, para um conjunto de dados reais de altimetria, sabidamente contendo *outliers*, foram obtidos os tamanhos de amostragem ideal, na presença e na ausência destes *outliers*, e conclui-se que o tamanho ideal de amostragem na presença de *outliers* é praticamente o dobro do tamanho ideal de amostragem na ausência dos mesmos: 228 e 115, respectivamente.

No capítulo 3, ainda utilizando o mesmo conjunto de dados reais de altimetria, sem *outliers*, foram comparadas as estimativas dos parâmetros do modelo gaussiano ajustado, além da média e da variância dos resíduos provenientes da validação cruzada, através da construção de intervalos de confiança utilizando os *softwares* R e ArcGIS. Por apresentar menor variabilidade (menores amplitudes dos intervalos de confiança construídos para a média e variância dos resíduos) o *software* R se mostrou mais adequado.

APÊNDICE

APÊNDICE A – Capítulo 1 escrito na língua Inglesa

Estimação teórica do tamanho amostral na geoestatística usando um modelo de semivariograma gaussiano

Theoretical Estimation of the Sampling Size of Geostatistics considering Gaussian Variogram Model

RESUMO

Na geoestatística clássica existe uma grande necessidade de pesquisas que criem e/ou investiguem métodos de amostragem de dados geoespaciais. Além da complexidade do assunto, alguns trabalhos apresentam soluções que utilizam mecanismos teóricos e práticos de diferentes áreas do conhecimento científico que atendem demandas específicas de pesquisadores da área. O objetivo deste artigo é utilizar a teoria da informação de sinais elétricos, principalmente considerando o teorema da taxa Nyquist, para determinar um tamanho ótimo para amostras georreferenciadas que usam grade quadrática regular, no qual o modelo de dependência espacial é o gaussiano. O que se deseja alcançar teoricamente é uma densidade de amostragem necessária para a reconstrução de mapas populacionais de variáveis nas quais as condições de regularidade necessárias em geoestatística foram verificadas, a saber: estacionariedade de 1ª e 2ª ordem e/ou estacionariedade do semivariograma, ausência de *outliers* e tendências, e semivariograma isotrópico. Como resultado, pode-se afirmar que a distância máxima entre os pontos da grade regular quadrática é de aproximadamente 30% do alcance prático observado no semivariograma da primeira amostragem experimental.

Palavras-chave: Geoestatística. Amostragem. Taxa Nyquist.

ABSTRACT

In Classical Geostatistics there is a great need for research that creates and/or explores methods of geospatial data sampling. In addition to this complex subject, some papers present solutions that use theoretical and practical mechanisms from different areas of scientific knowledge addressing specific demands of researches in the field. The purpose of this paper is to apply the Electrical Signal Information Theory, especially considering the Nyquist Rate Theorem, to determine an optimal size for georeferenced samples using a regular quadratic grid based on the spatial dependence Gaussian model. We expect to achieve a necessary sampling density to rebuild population maps of variables in which the following regularity conditions required in geostatistics were certified: 1^{st} and 2^{nd} order stationarity and/or stationarity of the variogram, absence of outliers and trends, and isotropic variogram. As a result, we can state that from the data set used, the greatest distance between points of the regular quadratic grid is nearly 30% of the practical range observed on the variogram of the first experimental sample.

Keywords: Geostatistics. Sampling. Nyquist Rate.

1. INTRODUCTION

For any scientific data analysis, it is extremely important to highlight value information (stated as representativeness), especially when we are dealing with samples (OLIVEIRA et al., 2014). This performance is also required when information regarding time and space is incorporated in the process (stated as regionalization) (YAMAMOTO; LANDIM, 2013). The aspect of regionalization of variables in a process is the reason why spatial statistics conquers enthusiasts all over the world. It happens mainly because questions can be clarified as considers the georeferenced location of the target variable.

Cressie and Wikle (2011) stated that regionalized information has always been essential to human being. When the topic was related to survival and/or achievements, the civilizations applied this mechanism to improve the display of orientation systems.

Yamamoto and Landim (2013) declared that Classical Statistics is important for many studies, although they have to assume a set of assumptions that are difficult or even impossible to verify, such as the requirement for random samples and the knowledge of the probability distribution of the variable studied.

The estimation of sample size in Spatial Statistics is one of the problems cited by Webster and Oliver (1992), Moraes, Santos et al. (2011), Souza et al. (2014) and Sartori and Zimback (2015).

Hoef (2002) and Clark (2009) present another problem of Classical Statistics with regard to the sampling size needed to obtain the same precision as the Spatial Statistic: approximately 9 times higher.

Pinto, Santos, and Oliveira (2014) suggested the subdivision of Spatial Statistics. Among them, the Geostatistics is highlighted as declared by Clark (1979), Armstrong (1998), Brooker (1991), Clark and Harper (2000), Druck et al. (2004), Olea (2009), Webster and Oliver (2007), Diggle, Menezes, and Su (2010), Santos et al. (2011), and Ferreira, Santos, and Rodrigues (2013).

Santos et al. (2011) mentioned the advantage of using Geostatistics modelling phenomena that do not exhibit the assumptions of Classical Statistics. Geostatistics uses the sampled neighbourhood sampled in order to reinforce the perception that the spatial dependence structure in a phenomenon to improve interpolations that are not a trend and have the least variance. Besides that, Vieira (2000) demonstrates that is still possible to recognize the uncertainty around predictions throughout the kriging variance.

Yamamoto and Landim (2015) registered the studies from distinct fields using Geostatistics as the main tool to analyse sampled data. Following this trend, plenty of papers concerning the size of geostatistics sampling are rising such as Brus and Heuvelink (2007), Modis and Papaodysseus (2006), Peigne et al. (2009), Vasat, Heuvelink and Boruvka (2010), Diggle, Menezes and Su (2010), and others.

Modis and Papaodysseus (2006) calculated (measured) the theoretical size to obtain an ideal density sampling based on the Information theory, formerly developed to Electrical Signals (WHITTAKER, 1915; SHANNON, 1949).

The paramount of this research is to apply Geostatistics in association with the Electrical Signal Information Theory, mainly considering the Nyquist Rate Theorem, in order to establish an ideal sampling size to those geostatistics/researchers. Those scientists use regular quadratic grid based on the spatial dependence Gaussian model. Then, our aim is to obtain a theoretical sampling density required to rebuild populational maps of the chosen variable in which all the regularity conditions required by Geostatistics were certified: 1st and 2nd order stationarity and/or stationarity of the variogram, absence of outliers and trends, and isotropic variogram.

The literature employed to our aim are: Modis and Papaodysseus (2006) introducing the theoretical methodology applied to spherical and exponential models; Yfantis et al. (1987) demonstrating the theoretical efficiency between different sampling grids; Vasat et al. (2010) indicating an alternative to reduce the sampling size in a multivariate process; Ferreira, Santos and Rodrigues (2013) explaining a systematic and interactive methodology of the geostatistics data analysis; and Santos et al. (2011) suggesting a comparative precision study with maps produced by distinct linear geostatistical interpolators.

2. MATERIAL AND METHODS

Hereafter, we are displaying the methodology adopted in this study and also the description of the researched area. We are also describing the purpose of the method and introducing the data.

2.1 Description of there search area

We analysed an area comprising 5.7 km^2 in the municipality of Viçosa, belonging to the "Zona da Mata" region in Minas Gerais State, Brazil. The site is surrounded by the latitudes 20º45'39'' S and 20º46'53'' S, and the longitudes 42º52'24'' W and 42º53'49'' W as seen in Figure 1.

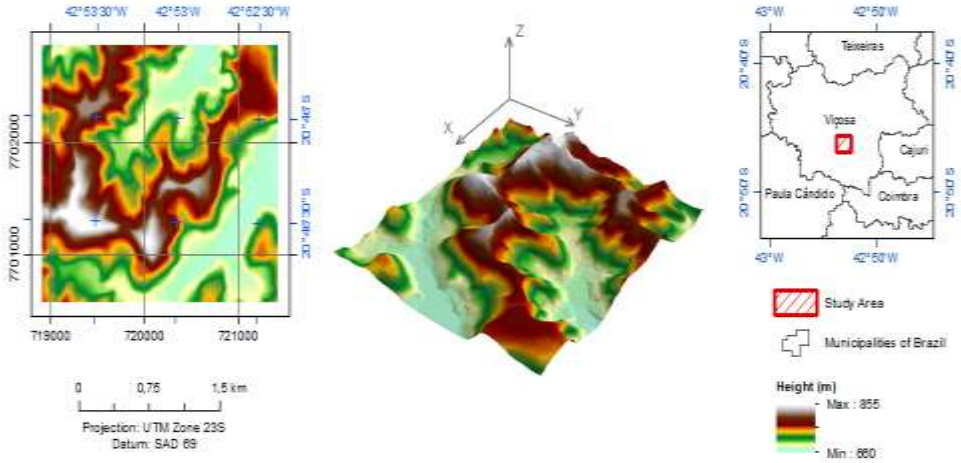


Figure 1 - Study site encompassing an area of 5.7 km^2 in the municipality of Viçosa, in Minas Gerais State, Brazil.
Source: Rosa (2017)

The altimetry data used are in accordance with the geodetic reference system Sirgas 2000 and the UTM (Universal Transverse Mercator coordinate system) projection system zone 23 S. Those data correspond to roughly 230 thousand spot heights known, which are interspersed around 5 meters both in X and Y directions. The minimum height recorded is 660 meters while the maximum is 885 meters (Figure 2).

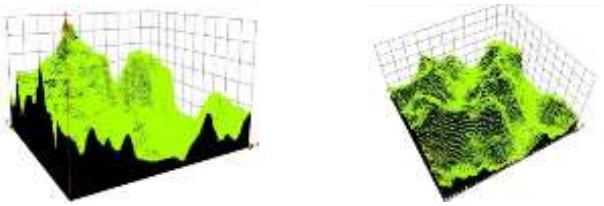


Figure 2 - Three-dimensional representation of the altimetry survey from a section in "Zona da Mata" in the municipality of Viçosa, Minas Gerais State, Brazil with roughly 230 thousand spot heights.
Source: Rosa (2017)

2.2 The methodology purposes

Modis and Papaodysseus (2006), based upon the Theorem of the Electrical Signal Information Theory called Nyquist Rate and also the Fourier transformed to the Geostatistics correlogram functions, proved that is possible to achieve a sampling size suitable to Geostatistics aiming a complete rebuild of the population studied.

However, those authors presented only one solution to the theoretical models of both spherical and exponential variograms. As a final suggestion (MODIS; PAPAODYSSEUS, 2006), a practical algorithm was proposed:

i) Start with an appropriate sampling density;

ii) Regulate to an equivalent covariance model;

iii) Establish the superior practical upper limit of the spectrum from the model (Nyquist Rate);

iv) Based on iii, determine an ideal sampling density;

v) If the ideal sampling density is not achieved, the sampling process should be repeated by preference;

vi) If the ideal density is economically unattainable, the available ways should be recalculated.

Indeed, the spherical and exponential models are frequently applied in many fields. the Gaussian model, defined by Equation 1, has a great value to other regionalized variables.

$$\gamma(h) = \begin{cases} 0 & \text{if } |h| = 0 \\ C\left[1 - \exp\left(-\dfrac{3|h|^2}{a^2}\right)\right] & \text{if } |h| > 0 \end{cases} \quad (1)$$

In the Gaussian model form Equation 1, C represents the level, h is a distance vector between the points and a the range of spatial dependence.

According to Ferreira, Santos, and Rodrigues (2013), those variograms adjusted by the Gaussian model are characterized by a spatial dependence with low deviations among closer neighbours and higher ones to those who are more distant. All of them inside the variogram range. Those features are typical to variables like altimetry and bathymetry supporting the use of the first one in this survey.

In compliance with Modis and Papaodysseus (2006) and based on the Theorem of Nyquist Rate, the rebuilding of populational maps is possible only if the sampling density adopting regular grids is greater or the same as a critical value. This value can be calculated in regard to the correlation function under some conditions such as stationarity, isotropy, strong spatial dependence, and finite range.

As mentioned before those authors introduce the theory and its results only for spherical and exponential models. Moreover, they demonstrated that for the variables in the mining field (homogeneous ore) the practice resembles the Geostatistics theory and the

sampling density must be greater or the same as half of the effective reach of the experiment, all regulated in the variogram.

Following this idea, this study focuses specifically on developing the Nyquist Rate Theory to the Gaussian model, demonstrating this theory could be applied to altimetry, and finding the sampling density for this model.

As part of Equation 1, Abramowitz and Stegun (1972) suggested the Equation 2 using the Fourier Transform in the correlation function of the Gaussian model, which is inversely related to the variogram.

$$R(\omega) = \exp\left(-\frac{3t^2}{a^2}\right)\cos(\omega t) \tag{2}$$

where ω is the sampling rate related to the signal frequency, and t is a sampling instant.

The same authors stated that the power spectrum of the random subjacent function is in Equation 3. This equation is the model which displays the performance of the correlation function in the Gaussian model.

$$S(\omega) = \frac{a}{2}\sqrt{\frac{\pi}{3}}\exp\left(-\frac{\omega^2 a^2}{12}\right) \tag{3}$$

As $S(\omega)$ tends to zero due to stabilization, $\exp\left(-\frac{\omega^2 a^2}{12}\right)$ ends on zero for ω infinite. Nevertheless, Journel and Huijbregts (1978) stated the Gaussian model should have its nullity on 0.05 since it is an asymptotic axis. Therefore, $\omega = \frac{6}{a}$.

The sampling size T of the sampling Theorem (Nyquist Rate) is given by Equation 4.

$$T \leq \frac{\pi}{\omega} = \frac{\pi}{\frac{6}{a}} = \frac{\pi}{6}a \tag{4}$$

which means almost half of the theoretical achievement as the rate claimed by Modis and Papaodysseus (2006) as shown in Figure 3.

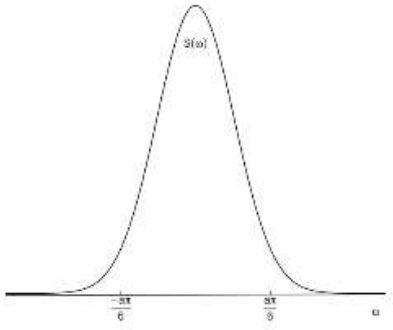


Figure 3 - Depiction of Fourier Transform to the Gaussian correlation model.
Source: Modis and Papaodysseus (2006)

Nonetheless, Olea (1999) reported that some variografic models including the correlogram do not reach stabilization of the curve in the theoretical range *a*. Furthermore, the Gaussian model is one of them. Thereby, this author demonstrates the transformation from the theoretical to the practical achievement a_p throughout:

$$a_p = \sqrt{3}a \;\Rightarrow\; a = \frac{\sqrt{3}}{3} a_p \tag{5}$$

Accordingly, the sampling size T in terms of practical achievement is given by Equation 6.

$$T \leq \frac{\pi\sqrt{3}}{18} a_p \tag{6}$$

The result of this equation indicates that the greatest distance between two points in a sampling regular quadratic grid must be nearly 30% of the practical achievement. This means that a first sample called experimental sampling should aim a 30% sampling density in the practical achievement if the regularity conditions to the Gaussian correlogram and consequently to the variogram are approved.

Examining a regular quadratic grid, we can notice the greatest distance between points is on the diagonals. Seeing that, the size of the maximum distance for sides should be converted before using the ratio $d = l\sqrt{2},$ where *d* meaning diagonal and *l* the square size.

2.3 Data description

The data set of a Brazilian region sited in Zona da Mata, municipality of Viçosa, Minas Gerais State, Brazil comprised 229,414 points. In order to verify the theory used in this study, the group aforementioned was reduced (following and checking the regularity conditions required for the method) 17 times reaching a sampling size of 49 points.

As an efficiency criterion, we considered kriging variance, the coefficients of Simple Linear Regression (SLR) among predicted and observed values of cross-validation, and data variance to the parameter assessment in variogram modelling (VIEIRA, 2000; MODIS and PAPAODYSSEUS, 2006; SANTOS et al., 2011; FERREIRA, SANTOS and RODRIGUES, 2013; YAMAMOTO and LANDIM, 2013).

Due to some *software* limitations regarding data set input, we adopted ArcGIS (ESRI, 2014) to all computational analyses. In order to reduce the sampling size, we applied the regular selection of altimetry data using the regular sampling tool of ArcGIS. Firstly, we

defined spatial arrangement in X and Y directions to employ the regular selection. As a result, the program created an intermediate grid of points based on the established arrangement, and also assembled the closer point of the altimetry database to each point produced in the intermediate grid. Afterward, we selected those closer points gathering an “almost regular” altimetry data set. As long as this study has a big data set with a high density of points, this “almost regular” can be considered as a regular selection. This fact was also proved by the Proximity Analyzer tool using the command Near in the Toolbox (ESRI, 2014).

Moreover, the simple kriging output (SANTOS et al., 2011) of all altimetry data set created the populational maps of sites studied in this survey according to Figure 4.

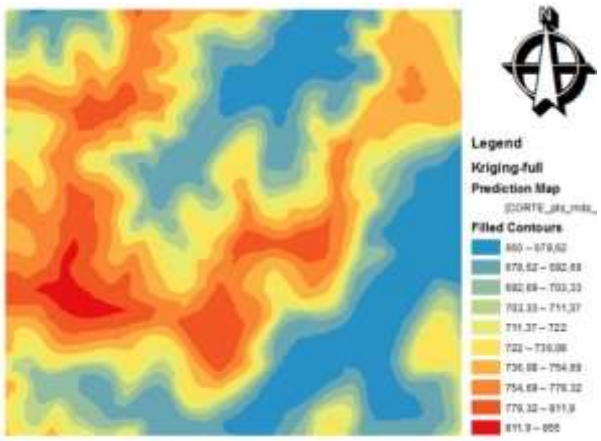


Figure 4 - Simple Kriging of altimetry data from a section of “Zona da Mata” in the municipality of Viçosa, Minas Gerais State, Brazil.
Source: Rosa (2017)

3. RESULTS AND DISCUSSION

Observing Table 1, we can notice that the averages and variances have not varied broadly. Only when the sampling variation was 99.96% and 99.98% smaller, we registered a larger range of deviation on those measurements reaching on this last sample the minimum limit of sampling size for Geostatistics as recommended by Yamamoto and Landim (2013).

The reductions were performed according to the regular spatial arrangement between points chosen in quadratic grids with sizes of 7 m, 10 m, 15 m, 20 m, 25 m, 30 m, 40 m, 50 m, 70 m, 90 cm, 130 m, 160 m, 190 m, 220 m, 260 m, 300 m, and 340 m.

The analytical data displayed in Table 1 indicate that decisions on sampling representation cannot be accomplished exclusively by size, average, and variance.

Table 1 - Sampling sizes, sample reduction, averages, variances, and spatial arrangement of quadratic regular grids of the altimetric survey from a section in “zona da mata” region near to Viçosa, in Minas Gerais state, Brazil.

SamplingSize	Reduction (%)	Average	Variance	Spatial Arrangement
229,414	0.00%	721.47	1,473.48	“population”
116,708	49.13%	721.50	1,474.94	7 meters
57,228	75.05%	721.46	1,474.94	10 meters
25,384	88.94%	721.57	1,475.87	15 meters
14,364	93.74%	721.37	1,473.79	20 meters
9,292	-95.95%	721.63	1,476.17	25 meters
6,308	-97.25%	721.65	1,477.33	30 meters
3,591	-98.43%	721.36	1,474.10	40 meters
2,300	-99.00%	721.49	1,474.48	50 meters
1,188	-99.48%	721.23	1,466.05	70 meters
700	-99.69%	721.40	1,483.56	90 meters
342	-99.85%	721.26	1,460.84	130 meters
224	-99.90%	721.16	1,498.39	160 meters
156	-99.93%	721.96	1,538.68	190 meters
110	-99.95%	721.99	1,514.61	220 meters
90	-99.96%	719.88	1,477.86	260 meters
64	-99.97%	721.78	1,569.59	300 meters
49	-99.98%	719.12	1,390.62	340 meters

Source: Prepared by the author

Figures 5, 6, and 7 exhibit the three-dimensional picture of sampling regular grids exposing the representative loss of population, when the sample does not have a suitable size suggested by the criteria being adopted.

(a) (b) (c) (d)

Figure 5 - Three-dimensional sampling figure of the quadratic regular grid of the altimetric survey in a section of “Zona da Mata” in the municipality of Viçosa, Minas Gerais State, Brazil. The grids displayed (a) 48 points, (b) 64 points, (c) 90 points, and (d) 110 points.

Source: Prepared by the author

(a) (b) (c) (d)

Figure 6 - Three-dimensional sampling figure of the quadratic regular grid of the altimetric survey in a section of “Zona da Mata” in the municipality of Viçosa, Minas Gerais State, Brazil. The grids displayed (a) 156 points, (b) 224 points, (c) 342 points, and (d) 700 points.

Source: Prepared by the author

(a) (b) (c) (d)

Figure 7 - Three-dimensional sampling figure of the quadratic regular grid of the altimetric survey in a section of “Zona da Mata” in the municipality of Viçosa, Minas Gerais State, Brazil. The grids displayed (a) 9,292 points, (b) 14,364 points, (c) 25,384 points, and (d) 57,228 points.
Source: Prepared by the author

According to expectation, on the report of Oliveira et al. (2009) as the sampling size increases the simple three-dimensional view of data displays the real performance of population, what is observed in Figures 6 and 7.

Oliveira et al. (2009) stated that researchers are always concerned about establishing the major indicators of a typical sample to Classical Statistics. Nevertheless, those measurements did not overtake this aim by themselves.

Yamamoto and Landim (2013) suggested that although many and distinct trials, this survey highlights an equal matter of the subject including Spatial Statistics. Therefore, until we find the “best” indicators, it is essential to evaluate scientific surveys on this topic looking forward feasible mechanisms for that determination. Modis and Papaodysseus (2006), Clark (2000), and Yamamoto and Landim (2013) presented articles that have tried it since 1975. Based on those papers, we noticed the common procedure is to apply the kriging as the efficiency criteria.

Authors such Vieira (2000), Santos et al. (2011), and Ferreira, Santos, and Rodrigues (2013) employed the SLR (Simple Linear Regression) as a criterion along with the kriging variance among kriging predicted and observed values after the cross-validation process.

As determined by Vieira (2000), and Ferreira, Santos, and Rodrigues (2013), the theoretical slope must be 1. Notwithstanding, we must notice how near this value is from the theoretical one. The RLS of this slope in function of kriging average variance (RMS) established the estimated model $\hat{Y} = -1{,}75X + 1{,}72$ with a determination coefficient of $R^2 = 0{,}9593$. That means about 96% of RMS variation can be explained by the slope deviation.

Modis and Papaodysseus (2006) stated that the spatial indicators adopted in scientific research of this field are a function among the kriging medium variances and the sampling density under the same conditions and different sizes conserving equivalent sample conditions

of the examined variable (as the sample exhibited in this paper). Those authors further claim that the ideal sample size was determined by the stabilization of the curve fitted to the graph. Furthermore, the difference among average variances of neighbours did not improve precision performance throughout kriging.

The graph of this function to the examined area on Figure 8 displays a stabilization of the kriging average variance named as RMS (Root Mean Square) around the sampling size 156, 224, and 342 points.

Among them, we could choose between sample 224 or 342. However, following the theory presented in this article, the ideal sampling size is 342 points (highlighted with an arrow in Figure 8), with the side distance between the points around 130 meters (or a diagonal of 190 meters) corresponding to 21.6% (or 31.5% if we consider the diagonal) regarding the practical reaching of 603 meters.

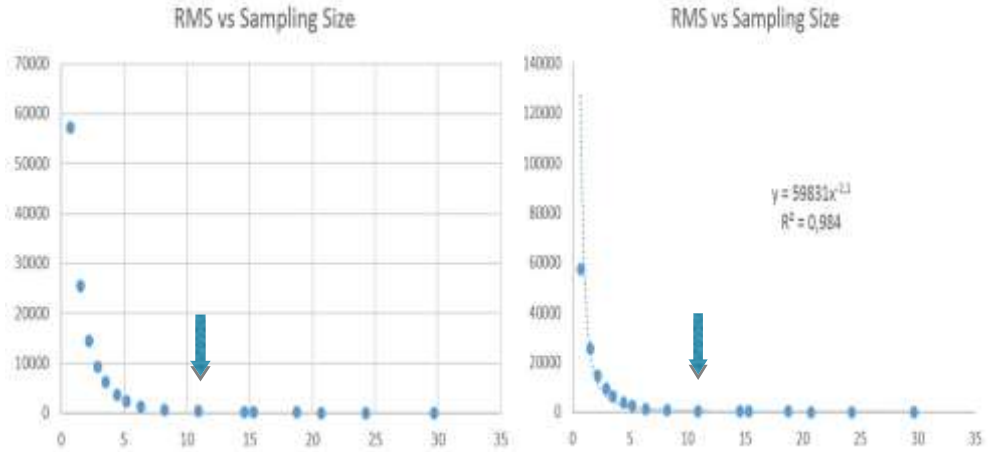
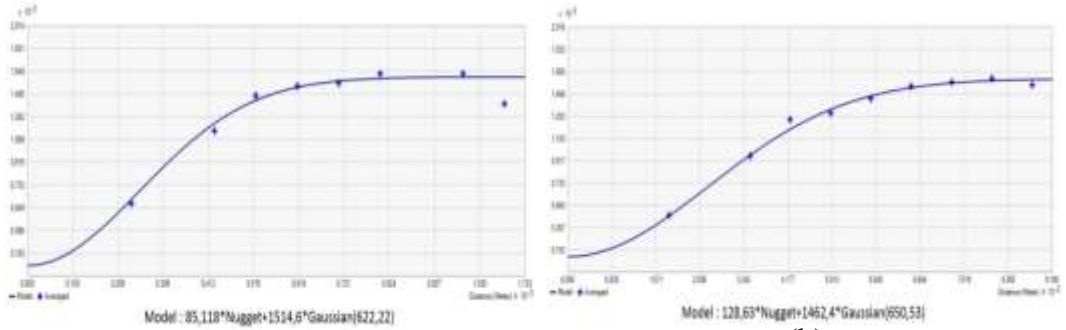
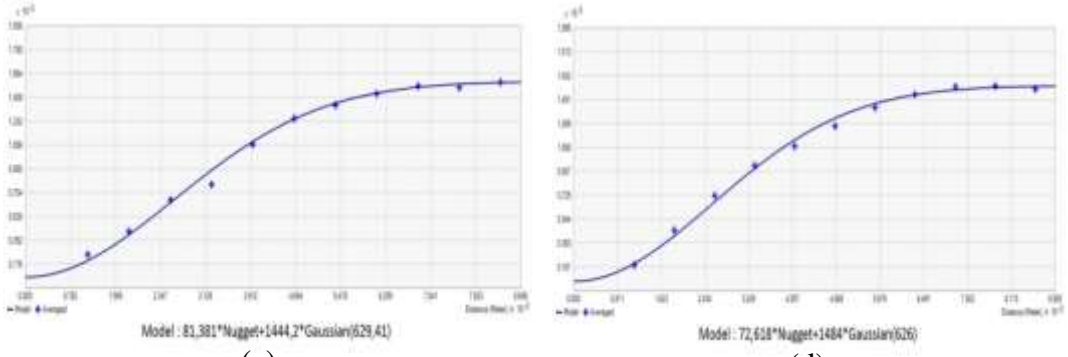


Figure 8 - Graphic demonstrating the relationship between kriging average variance and the sampling size of altimetry data from a section of “Zona da Mata”, Minas Gerais State, Brazil. On the left, a dispersion diagram, and on the right, a diagram with a tendency line, the model and R^2.
Source: Prepared by the author

In Figure 9, we are emphasizing variogram performance applied to distinct sampling sizes. Moreover, in Equation 1 the adapted the Gaussian model to experimental variograms.



(a) (b)



(c) (d)

Figure 9 - Experimental Variograms (“+”) and adapted Gaussian models (lines) to the spatial dependence of altimetry data from a section of “Zona da Mata”, Minas Gerais State, Brazil. The sampling size are (a) 156 points, (b) 224 points, (c) 342 points, and (d) 700 points.
Source: Prepared by the author

Clark (2000) said that many Geostatistic researchers believe the nugget effect can be produced by a low sampling density. Although variograms are displaying very similar values, the nugget effect increased significantly deviating since the grid most dense up to the less dense grid from 42.51 m^2 to 460 m^2 (considering this unit is referring to the variance rather than the area).

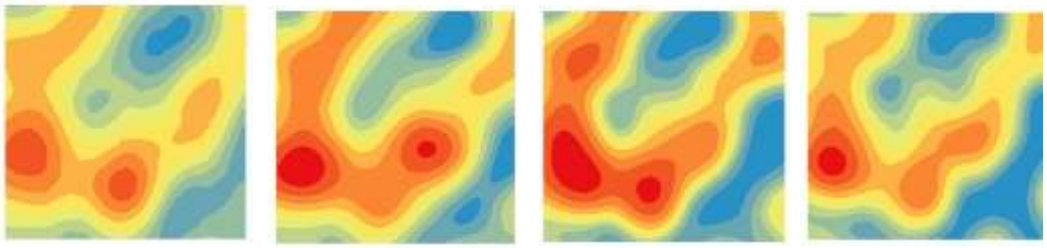
The method proposed in this work highlights the reach as a unit of great importance, since the sampling density will vary theoretically in its function. However, here is the point of great concern of this type of study, because the range is not kept constant with the reduction of sampling size.

Modis and Papaodysseus (2006) dealt with homogenous ores and did not detect variation when estimating the variografic range. On the other hand, the range estimated here fluctuated since the most to the less dense grid from 603 to 650 meters.

Even though we identified oscillations on those ranges, the average lateral distance of the quadratic regular grid we suggest is 130 meters corresponding to 342 points.

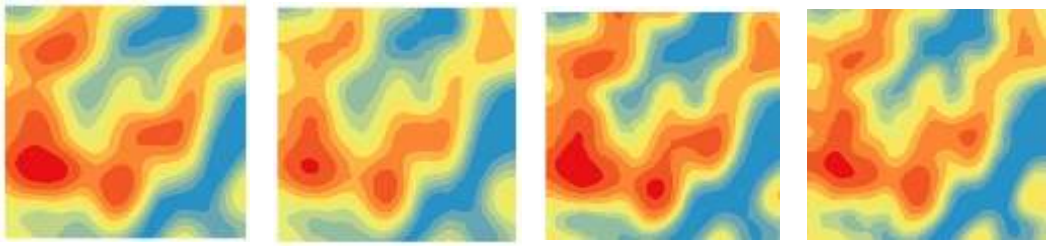
The choice of this value is in accordance with the Gaussian model performed here, and all Geostatistic regularity conditions were also certified and proofed.

Santos et al. (2011) mentioned that as maps reproduce reality, people tend to accept them as truth. Therefore, maps production through data interpolation is a relevant step to Geostatistic analyses since they will be at least criticized for those who know the mapped region. Accordingly, Figures 10 and 11 display maps obtained based on interpolation by simple kriging as recommended by Santos et al. (2011).



(a) (b) (c) (d)

Figure 10 - Simple kriging of the altimetric survey in a section of “Zona da Mata”, Minas Gerais State, Brazil. The sampling size are (a) 49 points, (b) 64 points, (c) 90 points and (d) 110 points.
Source: Prepared by the author



(a) (b) (c) (d)

Figure 11- Simple kriging of the altimetric survey in a section of “Zona da Mata”, Minas Gerais State, Brazil. The sampling size are (a) 156 points, (b) 224 points, (c) 342 points and (d) 700 points.
Source: Prepared by the author

As exhibited in Figures 10 and 11, the populational map starts to be reasonably depicted by a sampling size of 342 points which survey accuracy achieved stabilization according to Figure 8.

Vieira (2000), Modis and Papaodysseus (2006), Santos et al. (2011), Ferreira, Santos, and Rodrigues (2013), and Yamamoto and Landim (2013) specified that in the variogram modelling, the estimated range should appraise data deviation, which in turn is assessed by the sample variation. Even with a simple reduction of sampling, the same happened with this data set.

Another significant stage of a Geostatistic analysis is the process of cross-validation. Among the essential steps in this process are the average and the residuals produced by observed and predicted values. As stated by Vieira (2000), Santos et al. (2011), Ferreira, Santos, and Rodrigues (2013), we expected the average of residuals obtained in this process is null, and the variance is unitary as shown by Mood, Graybill, and Boes (1974). For truth, the closeness of those values is the focus of the analysis.

4. CONCLUSIONS

The present work aimed to use the Nyquist Rate Theorem to determine an ideal size for georeferenced samples using a regular quadratic grid. The important theoretical part for the Gaussian model of spatial dependence fitted to the data was developed and, as a result, a minimum density was presented as a function of an estimated practical range, one of the parameters of the experimental variogram adjusted by this model. In practical terms, this density is about 30% of the estimated value for the parameter, under regularity conditions. The developed theory was applied to a large data set and the theoretical density was proven by practice, adopting as criteria of evaluation some procedures and measures already consolidated in the area of Classical Geostatistics.

REFERENCES

ABRAMOWITZ, M.; STEGUN, I. A. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, Washington, D.C.: U.S. **Government Printing Office**, 1972. 1046 p.

ARMSTRONG, M. **Basic Linear Geostatistics**. Springer, Berlin, 1998. 153p.

BROOKER, P. I. **A geostatistical primer**. Singapore: World Scientific, 1991. 95 p.

BRUS, D. J.; HEUVELINK, G. B. M. **Optimization of sample patterns for universal kriging of environmental variables**. Geoderma, n. 138, p. 86-95, 2007.

CLARK, I. **Practical geostatistics**. London: Applied Science Publishers, 1979. 130p.

CLARK, I. **SAIMM Conference, Fourth World Conference on Sampling and Blending**, p. 21-23, 2009.

CLARK, I.; HARPER, W. V. **Practical geostatistics 2000**. Geostokos (Ecosse) Limited, 2000. 416 p.

CRESSIE N. A. C.; WIKLE, C. K. **Statistics for spatio-temporal data.** Wiley Series in Probability and Statistics. Hoboken, New Jersey, 2011. 588 p.

DIGGLE, P. J., MENEZES, R.; SU, T. Geostatistical inference under preferential sampling. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, v. 59, n. 2, p. 191-232, 2010.

DRUCK, S.; CARVALHO, M.S.; CÂMARA, G.; MONTEIRO, A. M. V. **Análise Espacial de Dados Geográficos**. Brasília: EMBRAPA, 2004. 209 p.

ESRI. **ArcGIS 10.2.2 for Desktop.** ESRI: Redlands, USA, 2014.

FERREIRA, D. F. **Estatística básica**. 2. ed. Lavras: Editora UFLA, 2009. 664 p.

FERREIRA, I. O.; SANTOS, G. R.; RODRIGUES, D. D. Estudo sobre a utilização adequada da krigagem na representação computacional de superfícies batimétricas. **Revista Brasileira de Cartografia,** n. 65/5, p. 831-842, 2013.

JOURNEL, A. G., HUIJBREGTS, C. J. **Mining Geostatistics**. Academic Press, London, 1978. 600 p.

MENDES, A.; SANTOS, G. R. dos; EMILIANO, P. C. ; ILAMBWETSI, P. S. ; KALEITA, A. L. . Theoretical Estimation of the Sampling Size of Geostatistics considering Gaussian Variogram Model. **SIGMAE**, v. 7, p. 17-30, 2018.

MODIS, K, PAPAODYSSEUS, K. Theoretical Estimation of the Critical Sampling Size for Homogeneous Ore Bodies with Small Nugget Effect. **Mathematic Geology,** v. 38, n. 8, p. 489-501, 2006.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the Theory of Statistics**. McGraw Hill, 1974. 480 p.

OLEA, R. A. **A practical primer on geostatistics**. U.S. Geological Survey Open-File Report, 2009-1103, 2009. 346 p.

OLEA, R. A. **Geostatistics for engineers and earth scientists**. Kluwer Academic Publishers, London, 1999. 303 p.

OLIVEIRA, M. S. de; BEARZOTI, E.; VILLAS BOAS, F. L.; NOGUEIRA, D. A.; NICOLAU, L. A.; OLIVEIRA, H. S. S. de. **Introdução à Estatística**, 2. ed. Lavras: UFLA, 2014. 462 p.

PEIGNÉ, J.; VIAN, J. F.; CANNAVACCIUOLO, M.; BOTTOLLIER, B.; CHAUSSOD, R. Soil Sampling based on field spatial variability of soil microbials indicators. **European Journal of Soil Biology**, v. 45, p. 488-495, 2009.

PINTO, E. S. O.; SANTOS, G. R. ; OLIVEIRA, F. L. P. Análise Espaço-Temporal Aplicada às Ocorrências de Hipertensão e Diabetes nos Municípios do Estado de Minas Gerais. **Revista Brasileira de Biometria**, v. 32, p. 238-266, 2014.

ROSA, L. M. . **Estudos sobre a influência de afirmações populares na Geoestatística clássica**. 2017. 103 p. Tese (Doutorado em Estatística Aplicada e Biometria). Viçosa: UFV.

SANTOS, G. R., OLIVEIRA, M. S., LOUZADA, J. M., SANTOS, A. M. R. T. Krigagem Simples versus Krigagem Universal: qual o preditor mais preciso? Revista **Energia na Agricultura**, Botucatu, v. 26, n. 2, p. 49-55, 2011.

SARTORI, A. A. C.; ZIMBACK, C. R. L. Recomposição florestal visando à conservação de recursos hídricos na bacia do rio Pardo, São Paulo. **Revista Energia na Agricultura**, v. 26, n. 4, p. 43-53, 2011.

SOUZA, Z. M.; SOUZA, G. S.; JÚNIOR, J. M.; PEREIRA, G. T. Número de amostras na análise geoestatística e na krigagem de mapas de atributos do solo. **Revista Ciência Rural**, Santa Maria, v. 44, n. 2, p. 261-268, 2014.

VÁSAT, R.; HEUVELINK, G. B. M.; BORŮVKA, L. Sampling design optimization for multivariate soil mapping. **Geoderma**, v. 155, n. 3-4, p. 147–153, 2010.

VER HOEF, J. M. Sampling and geostatistics for spatial data. **Ecoscience**, v. 9, p. 152-161, 2002.

VIEIRA, S. R. Geoestatística em estudos de variabilidade espacial do solo. In: NOVAIS, R.F.; ALVAREZ, V.H.; SCHAEFER, G.R. **Tópicos em ciência do solo**. Viçosa, v.1, p.1-54, 2000.

WEBSTER, R.; OLIVER, M. A. **Geostatistics for Environmental Scientists**. 2. ed. Chichester: John Wiley & Sons, 2007. 315p.

WEBSTER, R.; OLIVER, M. A. Sample adequately to estimate variograms of soil properties. **Journal of Soil Science**, v. 43, p. 177-192, 1992.

YAMAMOTO, J. K.; LANDIM, P. M. B. **Geoestatística: Conceitos e Aplicações.** Oficina de Textos: São Paulo, 2013. 216 p.

YFANTIS, E. A.; FLATMAN, G. T.; BEHAR, J. V. Efficiency of kriging estimation for square, triangular and hezagonal grids. **Mathematical Geology**, v. 19, p. 183-205, 1987.