ISSN 2359-6562 (ONLINE) 2359-6562 (CD-ROM)

ESTIMAÇÃO DO TAMANHO AMOSTRAL NA GEOESTATÍSTICA USANDO UM MODELO DE VARIOGRAMA GAUSSIANO NA PRESENÇA DE *OUTLIERS*

ANDRÉ MENDES[1], GERSON RODRIGUES DOS SANTOS[2], PAULO CÉSAR EMILIANO[3], AMY LEIGH KALEITA[4], MATHEUS DE PAULA FERREIRA[5]

[1] *Departamento de Estatística, Universidade Federal de Viçosa-UFV, Av. Peter Henry Rolfs, s/n, Campus Universitário, cep: 36570-900, Viçosa-MG, Brasil, e-mail: amendesmat@yahoo.com.br*
[2] *Departamento de Estatística, Universidade Federal de Viçosa-UFV, Av. Peter Henry Rolfs, s/n, Campus Universitário, cep: 36570-900, Viçosa-MG, Brasil, e-mail: gerson.santos@ufv.br*
[3] *Departamento de Estatística, Universidade Federal de Viçosa-UFV, Av. Peter Henry Rolfs, s/n, Campus Universitário, cep: 36570-900, Viçosa-MG, Brasil, e-mail: paulo.emiliano@ufv.br*
[4] *Department of Agricultural and Biosystems Engineering, Iowa State University, 605 Bissel Road, Ames, IA 50011-3270, Estados Unidos, e-mail: kaleita@iastate.edu*
[5] *Departamento de Estatística, Universidade Federal de Viçosa-UFV, Av. Peter Henry Rolfs, s/n, Campus Universitário, cep: 36570-900, Viçosa-MG, Brasil, e-mail: matheus.paula@ufv.br*

RESUMO: A determinação de um tamanho de amostragem 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*.

ESTIMATION OF THE SAMPLING SIZE OF GEOSTATISTICS CONSIDERING GAUSSIAN VARIOGRAM MODEL IN THE PRESENCE OF *OUTLIERS*

ABSTRACT: The determination of suitable sample size for population reconstruction in the analysis of spatial data is something that has been studied in several scientific papers. Regardless 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 scientific areas knowledge. The purpose of this work is 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 Gaussian. What we intend to achieve is a sampling density necessary for the population maps reconstruction 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 outliers absence, 115 points, was well below the 228 points obtained in the outliers presence.

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.

Recebido em 02/12/2017 e aprovado para publicação em 23/08/2019
DOI: http://dx.doi.org/10.17224/EnergAgric.2019v34n3p429-440

Nas diversas áreas do conhecimento, a geoestatística vem sendo utilizada como principal método de análise de dados de amostras (YAMAMOTO; LANDIM, 2015) e está 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), permitindo 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 de Nyquist para a determinação de um tamanho de amostragem ótimo para pesquisadores que utilizam grade regular quadrática em que o modelo de dependência ajustado é o Exponencial ou Esférico.

Vašát, Heuvelink e Borùvka (2010) mostraram uma alternativa para reduzir o tamanho de amostragem 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; TOLOI, 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, Haibo e Hong (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 de outliers para dados geoespaciais contínuos através da geoestatística e teoremas da estatística clássica, independentemente da causa geradora das inconsistências. Estes autores fazem referência a uma Tese de Doutorado de Santos (2016) onde foi feita uma proposta via geoestatística para eliminação de *outliers* do banco de dados.

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

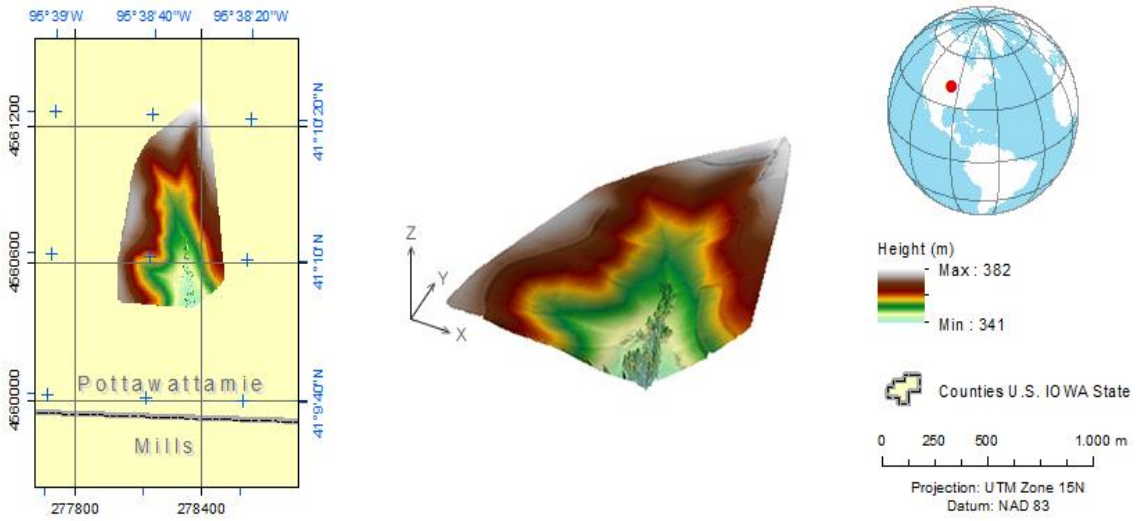
2 MATERIAL E MÉTODOS

Visando alcançar os objetivos desse trabalho, a seguir é apresentada a descrição da região estudada, uma síntese da metodologia de detecção de *outliers* e da taxa de Nyquist bem como a 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. Área de estudo, que compreende uma parcela de aproximadamente 34 hectares da cidade de Treynor, Estado de Iowa, Estados Unidos



Fonte: Santos et al. (2016)

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; 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 coordinate system). Esses dados compreendem pouco mais de 192 mil pontos com altitude conhecidas, 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:

1) 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 nula e variância unitária, ou seja, $\varepsilon_p^{"} \sim Z(0,1)$, em que $\varepsilon_p^{"}$ são os resíduos padronizados, segundo Vieira (2000);

2) 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;

3) Para obtenção dos resíduos, a variável regionalizada estudada, Y, foi decomposta em três componentes, conforme Equação 1.

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

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

4) 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);

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

6) Construiu-se intervalos de confiança (IC) com probabilidade $(1-\alpha)$ para os resíduos adotando-se a distribuição normal padrão $\big(Z(0,1)\big)$ 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.

$$P\left[\bar{x} - Z_{\alpha}\frac{S}{\sqrt{n}} < \mu < \bar{x} + Z_{\alpha}\frac{S}{\sqrt{n}}\right] = 1-\alpha \qquad (2)$$

Silva (2012) mostra que todos os valores que não pertencerem 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 outliers;

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

8) 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; MOSTELLER; TUKEY, 1983).

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

2.3 Taxa de Nyquist

Modis e Papaodysseus (2006) mostraram que é possível obter um tamanho de amostragem 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 às funções do correlograma da geoestatística. Modis e Papaodysseus (2006) apresentaram uma solução somente para os modelos teóricos de variogramas esféricos e exponenciais, apresentando, ao final, um algoritmo prático, a saber:

1) Começa-se com uma densidade de amostragem planejada;

2) Ajusta-se o modelo de covariância correspondente;

3) Determina-se o limite superior prático do espectro do modelo obtido (taxa Nyquist);

4) Usando o passo 3 deste algoritmo, determina-se a densidade de amostragem ótima;

5) Se essa densidade ótima não for financeiramente viável, redimensiona-se o tamanho amostral.

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 variograma 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.

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

Os parâmetros do modelo Gaussiano apresentado na Equação 3, são: C, patamar, h, vetor distância entre os pontos e a o alcance de dependência espacial.

Segundo Ferreira, Santos e Rodrigues (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 variograma. 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:

1) Abramowitz e Stegun (1972) apresentaram, a partir da Equação 3, a Equação 4, utilizando a Transformada de Fourier para a função de Correlação do modelo Gaussiano, que é inversamente relacionada ao variograma.

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

Sendo ω a taxa de amostragem relacionada à frequência dos sinais e t um instante de amostragem;

2) Ainda de acordo com Abramowitz e Stegun (1972), o espectro de potência da função aleatória subjacente, que é o modelo que descreve o comportamento da função de correlação do modelo Gaussiano, é dado pela Equação 5.

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

Devido à estabilização do correlograma, $S(\omega)$ tende a zero. Assim, $exp\left(-\frac{\omega^2 a^2}{12}\right)$ vale zero para ω infinito. Entretanto, Journel e Huijbregts (1978) afirmam que o modelo Gaussiano, assintótico ao eixo das abscissas, deve ter nulidade considerada em 0,05. Logo $\omega = \frac{6}{a}$. Decorre do teorema da taxa de Nyquist que o tamanho amostral T é dado pela Equação 6.

$$T \leq \frac{\pi}{\omega} = \frac{\pi}{\frac{6}{a}} = \frac{\pi}{6} a \quad (6)$$

O que corresponde a aproximadamente metade do alcance teórico, conforme a taxa obtida por Modis e Papaodysseus (2006).

Entretanto, segundo Olea (1999), alguns modelos variográficos e, assim sendo o correlograma, não alcançam a estabilização da curva no alcance teórico a, sendo o Gaussiano um deles. Assim sendo, este autor apresenta a transformação do alcance teórico para o alcance prático, a_p, dada pela Equação 7.

$$a_p = \sqrt{3}a \Rightarrow a = \frac{\sqrt{3}}{3} a_p \quad (7)$$

Portanto, o tamanho amostral T, em função do alcance prático a_p é dado pela Equação 8.

$$T \leq \frac{\pi\sqrt{3}}{18} a_p \quad (8)$$

Esse resultado mostra que a maior distância entre dois pontos de uma grade regular quadrática de amostragem deve ser aproximadamente 30% do alcance prático. Isto significa que, comprovada a existência das condições de regularidade para o modelo de correlograma Gaussiano e, consequentemente, para o variograma, uma primeira amostragem, chamada de amostragem experimental, deve ser feita para que a densidade da amostra possa ser estimada como 30% do alcance prático.

Como numa grade regular quadrática a maior distância entre dois pontos está nas diagonais, para se obter a distância máxima para os lados deve-se estabelecer antes a relação $d = l\sqrt{2}$, em que d significa a diagonal e l o lado do quadro.

2.5 Caracterização dos dados

O conjunto de dados 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, com uma altitude média de cerca de 363 metros e variância média de cerca de 62 metros quadrados. 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 amostragem.

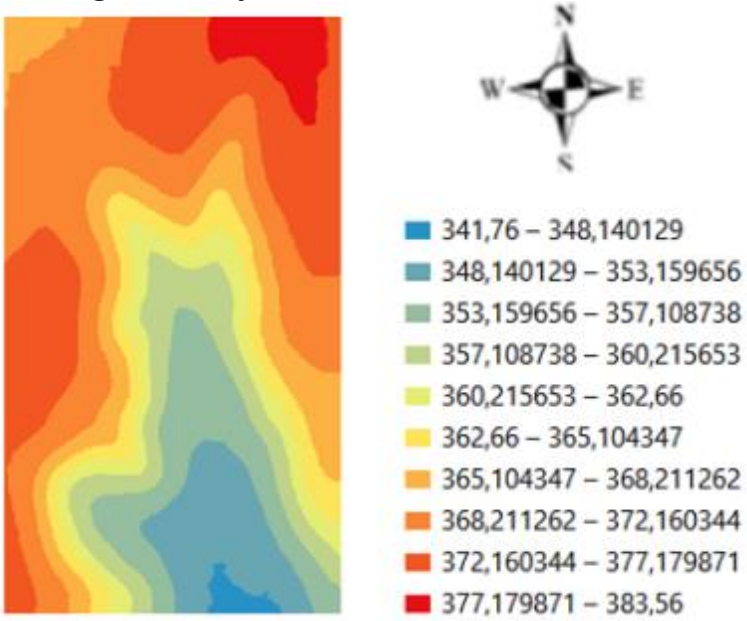
Como critério de eficiência, considerou-se a variância média de krigagem (RMS), os coeficientes da Regressão Linear Simples (RLS)

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; PAPAODYSSEUS, 2006; SANTOS et al., 2011; FERREIRA; SANTOS; RODRIGUES, 2013; YAMAMOTO; LANDIM, 2013).

Para a análise computacional deste trabalho, foi utilizado o software ArcGIS 10.2.2 (ENVIRONMENTAL SYSTEMS RESEARCH INSTITUTE, 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 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 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 Analisador da Proximidade do Toolbox, usando o comando Near (ENVIRONMENTAL SYSTEMS RESEARCH INSTITUTE, 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 da região de Treynor-Iowa, Estados Unidos



Fonte: Santos et al. (2017)

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, redução de amostras, médias, variâncias e espaçamento das grades regulares quadráticas do levantamento altimétrico de parte da região de Treynor-Iowa, Estados Unidos

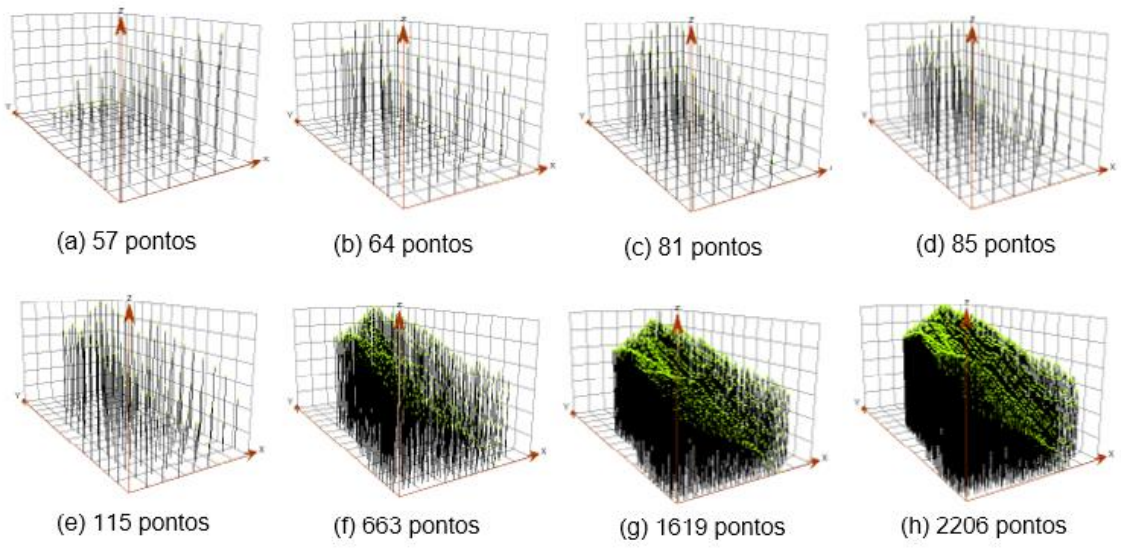
Tamanho amostral	Redução (%)	Média	Variância	Espaçamento (metros)
4067	0	363,62	62,02	"população"
2206	45,75	363,42	60,38	12
1619	60,19	363,43	60,89	14
663	83,69	363,43	61,80	22
558	86,28	363,46	60,34	24
315	92,25	363,43	64,46	32
280	93,12	363,49	61,47	34
183	95,50	363,51	61,61	42
167	95,89	363,36	62,47	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

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.

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.

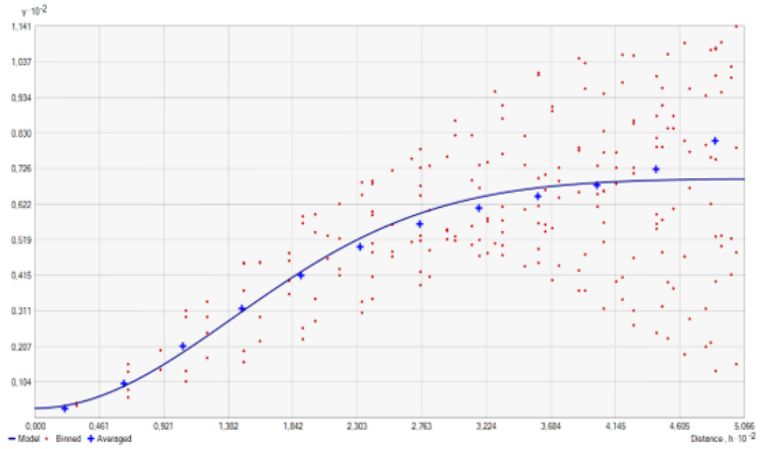
Figura 3. Representação tridimensional amostral de grade regular quadrática do levantamento altimétrico da cidade de Treynor-Iowa, Estados Unidos



Conforme esperado, de acordo com Oliveira et al. (2014), aumentando o tamanho de amostragem 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 variograma para todos os 4067 pontos utilizados no estudo.

Figura 4. Variograma experimental (sinais "+") e modelo Gaussiano ajustado (linha) para a dependência espacial dos dados de altimetria da região da cidade de Treynor-Iowa, Estados Unidos para o tamanho de amostragem de 4067 pontos



Modelo : 2.8223.Nugget+66.865Gaussian(346.6)

Os demais variogramas, embora não tenham sido apresentados, seguiram o mesmo comportamento. Pode-se notar que o modelo ajustado ao variograma experimental foi o modelo Gaussiano, apresentado na Equação 3.

Considerando-se que o alcance prático obtido foi de 252,6 metros, utilizando-se a metodologia proposta por Modis e Papaodysseus (2006) adaptada ao modelo Gaussiano de dependência espacial ajustado aos dados desse trabalho, determinou-se que o tamanho ideal de amostragem é 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. Vale ressaltar, ainda, que as condições de regularidade geoestatística foram verificadas e comprovadas para cada tamanho de amostragem no estudo.

Segundo Oliveira et al. (2014) sempre houve uma preocupação em determinar os principais indicadores da representatividade amostral para a Estatística Clássica. Já Yamamoto e Landim (2013) destacam que, apesar das muitas e variadas tentativas, indicadores da representatividade amostral são igualmente importantes também para a Estatística Espacial. Portanto, na ausência desses "melhores" indicadores se faz necessário avaliar os trabalhos científicos sobre este assunto a fim de obter mecanismos viáveis para tal determinação.

Observa-se que, na prática, é utilizado o erro quadrático médio (RMS) como critério de eficiência, conforme Clark e Harper (2000), Modis e Papaodysseus (2006) e Yamamoto e Landim (2013).

Além da utilização do RMS, Vieira (2000), Santos et al. (2011) e Ferreira, Santos e Rodrigues (2013), utilizaram, como critério de eficiência, a Regressão Linear Simples entre os valores preditos e observados por krigagem, após o processo de validação cruzada.

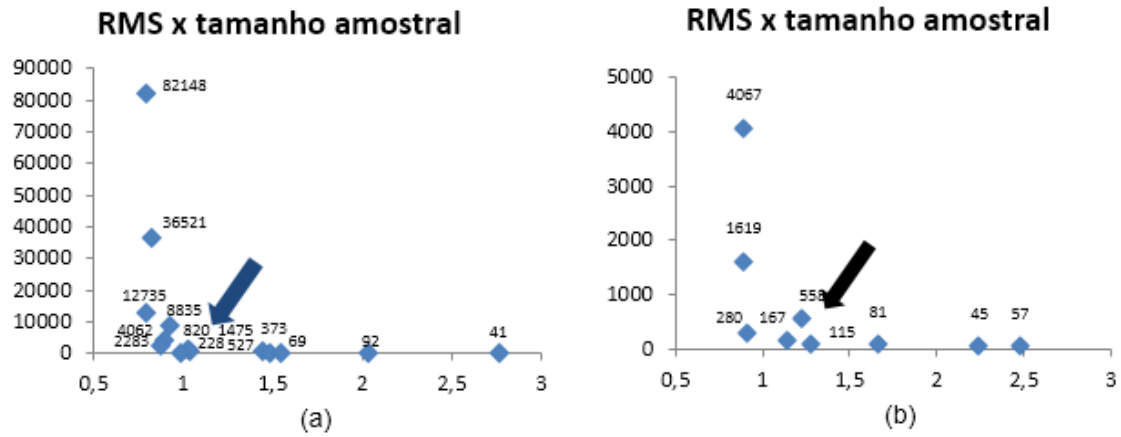
Os erros médios quadráticos, os tamanhos de amostragem e o espaçamento das grades quadráticas para a variável altimetria estudada são apresentados na Tabela 2.

Tabela 2. Apresentação dos erros quadráticos médios (RMS), os tamanhos de amostragem 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

Na Figura 5 é apresentado o gráfico do tamanho de amostragem em função do RMS.

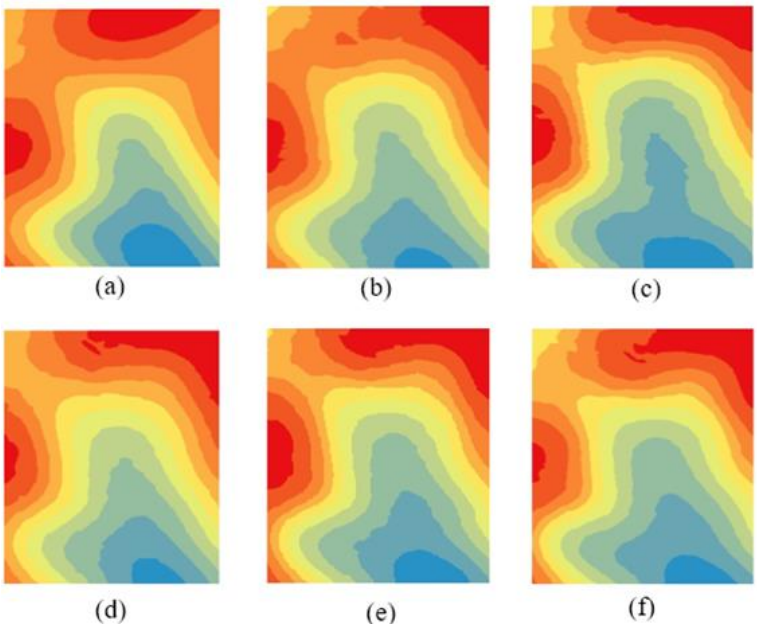
Figura 5. Representação gráfica da relação entre o erro quadrático médio (RMS) e o tamanho de amostragem para os dados de altimetria próxima à cidade de Treynor-Iowa, Estados Unidos. O conjunto de dados em (a) contém outliers e em (b) não contém outliers



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

Segundo Santos et al. (2011), citando a literatura científica, a produção de mapas populacionais através da interpolação de dados é uma etapa muito importante da análise geoestatística uma vez que as pessoas tendem a aceitá-los como verdadeiras. Assim, são apresentados na Figura 6 os mapas obtidos através da interpolação por krigagem simples, conforme recomendado por Santos et al. (2011).

Figura 6. Krigagem simples para os dados de altimetria próxima à cidade de Treynor-Iowa, Estados Unidos. Os tamanhos de amostragem são (a) 48 pontos, (b) 57 pontos, (c) 64 pontos, (d) 81 pontos, (e) 85 pontos e (f) 115 pontos



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

A validação cruzada é outro processo de suma importância em uma análise geoestatística. Neste processo, é possível obter a média e a variância dos resíduos gerados entre os valores observados e os preditos. Embora se espere que a média dos resíduos obtidos por este processo seja nula e a variância seja igual a 1, conforme citado por Vieira (2000), Santos et al. (2011), Ferreira, Santos e Rodrigues (2013), na prática, é analisado a proximidade desses valores. Para este estudo, ambos se mostraram estatisticamente iguais aos valores de referência.

4 CONCLUSÃO

Na presença de *outliers*, o tamanho de amostragem 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 amostragem 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 cara.

5 AGRADECIMENTOS

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

6 REFERÊNCIAS

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

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. 285, 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.

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

FARACO, M. A.; URIBE-OPAZO, M. A.; SILVA, E. A. A.; 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, 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.** New York: Wiley, 1983.

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

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

MODIS, K.; PAPAODYSSEUS, K. Theoretical Estimation of the Critical Sampling Size for Homogeneous Ore Bodies with Small Nugget Effect. **Mathematical Geology**, Berlin, 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.

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

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

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

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

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. Tese (Doutorado em Engenharia Civil) –Universidade Federal de Viçosa, Viçosa, 2016.

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.; MARQUES 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.

VAŠÁT, R.; HEUVELINK, G. B. M.; BORÙVKA, L. Sampling design optimization for multivariate soil mapping. **Geoderma**, Amsterdam, 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: UFV, v. 1, p. 1-54, 2000.

YAMAMOTO, J.; LANDIM, P. **Geoestatística**: conceitos e aplicações. São Paulo: Oficina de Textos, 2013.