BCG - Boletim de Ciências Geodésicas - On-Line version, ISSN 1982-2170
http://dx.doi.org/10.1590/S1982-21702014000400043

PROPOSTA DE UM TESTE DE HIPÓTESE PARA A EXISTÊNCIA DE DEPENDÊNCIA ESPACIAL EM DADOS GEOESTATÍSTICOS

Proposal of a hypothesis test for the existence of spatial dependence in geostatistical data

ENIO JÚNIOR SEIDEL[1]
MARCELO SILVA DE OLIVEIRA[2]

[1]Universidade Federal de Santa Maria (UFSM)
Santa Maria – RS – Brasil
enioseidel@gmail.com
[2]Universidade Federal de Lavras (UFLA)
Lavras – MG - Brasil
marcelo.oliveira@dex.ufla.br

RESUMO

A avaliação da significância da dependência espacial é importante para que se possa realizar uma inferência formal sobre a hipótese nula de não existência de dependência espacial em Geoestatística. Com o presente estudo, objetivou-se construir um teste de significância para a hipótese nula de ausência de dependência espacial, para uma melhor decisão sobre a existência ou não de dependência em dados geoestatísticos. Para a construção do teste de dependência espacial levou-se em consideração as características dos modelos ajustados ao semivariograma e, para a construção do teste de hipótese, inicialmente, foram definidos a hipótese nula e uma estatística de teste gerada a partir do conceito de área de dependência espacial obtida no semivariograma. O teste foi construído com base em simulações de fenômenos geoestatísticos apresentando característica de efeito pepita puro, ou seja, fenômenos em que a hipótese nula foi verdadeira. Por fim, foi estudado o poder do teste para diferentes graus de dependência espacial simulados. O teste apresentou bom poder, sendo que este tendeu a 100% quando aumentou o grau de dependência espacial dos fenômenos geoestatísticos simulados.

Palavras-chave: Estatística Espacial; Variografia; Inferência; Simulação.

ABSTRACT

The assessment of the significance of spatial dependence is important for the formal inference about the null hypothesis of non-existence of spatial dependence in geostatistics. Through this study, we aimed at developing a significance test for the null hypothesis of lack of spatial dependence, for the best decision about the existence whether or not the dependence on geostatistical data. For the development of the spatial dependence test, we considered the characteristics of models adjusted to the semivariogram; and for the development of the hypothesis test, firstly the null hypothesis and a statistics of the test generated from the concept of spatial dependence area obtained in the semivariogram were defined. The test was developed based on simulation of geostatistical phenomena presenting features of pure nugget effect, i.e., phenomena in which the null hypothesis was true. Finally, it was studied the power of the test for different degrees of simulated spatial dependencies. The test showed good power, and this tended to 100% when the degree of spatial dependence of the simulated geostatistical phenomena increased.
Keywords: Spatial Statistics; Variography; Inference; Simulation.

1. INTRODUÇÃO

A variabilidade de qualquer variável espacial avaliada em um fenômeno pode ser dividida em dois tipos básicos, a saber, aquela de cunho não-estocástico, frequentemente chamada de tendência, e outra de cunho estocástico, frequentemente denominada de dependência espacial. A Geoestatística considera os dois tipos de variabilidade para explicar a continuidade observada em dados espaciais, reservando a média para modelar a estrutura da tendência, e a covariância, correlação e semivariância, para modelar a estrutura de dependência espacial.

A continuidade geográfica da variável espacial se manifesta pela semelhança de valores da variável em dois pontos vizinhos e valores muito diferentes em pontos distantes, dando o aspecto de ligação entre os pontos (RESENDE, 2007).

Esta continuidade geográfica ocorre pelo fato de existir certa medida de tendência e/ou certo grau de dependência espacial, pois observações próximas são associadas, e essa associação é maior para distâncias menores.

A avaliação da significância da dependência espacial, quando feita, é realizada através de interpretação de algum indicador de dependência espacial, como, por exemplo, a medida do grau de dependência espacial apresentada por Biondi, *et al*., 1994, ou a razão de dependência dada por Cambardella, *et al*., 1994.

Também, em alguns casos realiza-se tal avaliação com base na construção de envelopes simulados no semivariograma, como, por exemplo, quando se utiliza o pacote *geoR* (RIBEIRO JÚNIOR; DIGGLE, 2001) do *software* R (R DEVELOPMENT CORE TEAM, 2012). Os envelopes simulados são gerados a partir de permutações dos dados observados nas coordenadas do *grid* amostral. Isto é, mantêm-se inalteradas as coordenadas geográficas, e permutam-se os valores da variável (atributo) observada. O princípio da realização das permutações é tentar

quebrar a estrutura de dependência espacial dos dados, gerando uma espécie de dados independentes.

Contudo, apesar dos indicadores de dependência espacial e os envelopes simulados gerarem uma indicação sobre a dependência espacial, não geram, diretamente, um valor p para interpretação mais objetiva.

Assim, o emprego dos envelopes simulados, ou a utilização de indicadores de dependência espacial, talvez possa não ser suficiente para realizar uma inferência formal sobre a hipótese nula de não existência de dependência espacial em Geoestatística.

Propor um teste de hipótese, para tal avaliação, possibilita a geração de um valor p (probabilidade de significância) que permite uma avaliação objetiva da significância da dependência espacial, e, portanto, uma melhor interpretação e, consequentemente, uma melhor decisão, sobre a existência ou não, de dependência espacial em dados geoestatísticos. Assim, tem-se como objetivo construir um teste de significância para a hipótese nula de ausência de dependência espacial.

2. METODOLOGIA

O desenvolvimento do trabalho é feito utilizando-se da teoria geoestatística e do recurso de simulações estocásticas. Todo o desenvolvimento é realizado no *software* R (R DEVELOPMENT CORE TEAM, 2012).

Além disso, o trabalho considera semivariogramas de processos estocásticos que atendam a hipótese de estacionariedade de 2ª ordem, sem tendência (média constante) e isotrópicos (semivariogramas que não dependem das direções, somente das distâncias).

O semivariograma é a principal ferramenta utilizada para estudar a estrutura de dependência espacial em Geoestatística, sendo um gráfico que relaciona semivariâncias (γ) com distâncias (h) (SEIDEL; OLIVEIRA, 2013). Mais detalhes sobre o semivariograma e sua construção podem ser vistos em Journel e Huijbregts (2003), Olea (2006) e Seidel (2013).

O estimador clássico de semivariograma é dado por (Matheron, 1963):

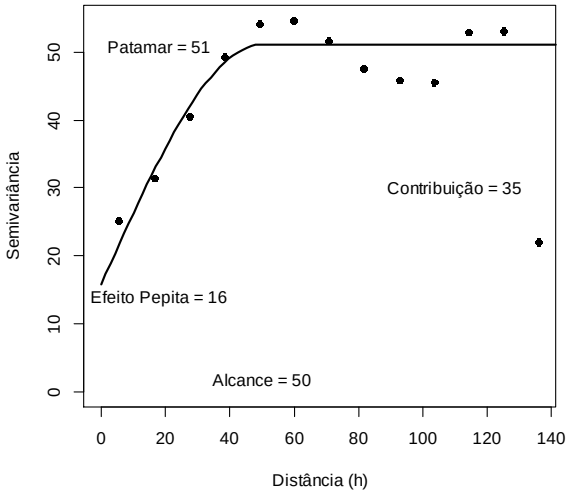
$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \{Z(s_i + h) - Z(s_i)\}^2 , \qquad (1)$$

em que $Z(s_i)$ e $Z(s_i + h)$ são valores da variável Z observada nos pontos s_i e $s_i + h$, respectivamente e $N(h)$ é o número de pontos observados para cada distância h. A Figura 1 apresenta um semivariograma amostral qualquer com seus parâmetros.

Com base na Figura 1 é possível perceber que as semivariâncias médias crescem conforme se aumentam as distâncias h. Assim, pode-se inferir que quanto

maiores forem as distâncias h entre as observações, menores se tornam as relações entre $Z(s+h)$ e $Z(s)$, ou seja, menor se torna a dependência espacial.

Figura 1 – Exemplo de um semivariograma amostral qualquer e seus parâmetros (Efeito pepita, Contribuição, Patamar e Alcance).



Após a construção do semivariograma é necessário ajustar um modelo teórico que explique o comportamento do semivariograma em termos de modelar a variabilidade observada. Segundo Alves *et al*. (2011), a ideia de ajustar um modelo teórico aos dados experimentais é sumarizar as relações espaciais nos dados. Para Carvalho, Silveira e Vieira (2002), os modelos matemáticos ajustados permitem visualizar a natureza da variação espacial das variáveis estudadas.

Os modelos mais utilizados são o modelo esférico, o modelo exponencial e o modelo gaussiano. O modelo esférico é dado na forma (SEIDEL, 2013):

$$\gamma(h)_{esf} = \begin{cases} 0 & , h = 0 \\ C_0 + C_1\left[1{,}5\left(\frac{h}{a}\right) - 0{,}5\left(\frac{h}{a}\right)^3\right] & , 0 < h \leq a \cdot \\ C_0 + C_1 & , h > a \end{cases} \quad (2)$$

Já o modelo exponencial é dado por (SEIDEL, 2013):

$$\gamma(h)_{\exp} = \begin{cases} 0 & , h = 0 \\ C_0 + C_1\left[1 - e^{\left(-\frac{h}{a/3}\right)}\right] & , h \neq 0 \end{cases} . \quad (3)$$

O modelo gaussiano é escrito como (SEIDEL, 2013):

$$\gamma(h)_{gau} = \begin{cases} 0 & , h = 0 \\ C_0 + C_1\left\{1 - e^{\left[-\left(\frac{h}{a/\sqrt{3}}\right)^2\right]}\right\} & , h \neq 0 \end{cases} , \quad (4)$$

em que C_0 é o parâmetro efeito pepita, C_1 é o parâmetro contribuição, a é o parâmetro alcance, $C_0 + C_1$ é o parâmetro patamar e h é a distância entre pontos. Nos casos de ausência de dependência espacial, ajusta-se o modelo de efeito pepita puro.

A construção do teste de dependência espacial, na forma que é proposto neste trabalho, leva em consideração características dos modelos ajustados ao semivariograma. Assim, o primeiro passo antes da aplicação do teste, é ajustar o modelo adequado ao semivariograma em estudo.

Para a construção do teste de hipótese, inicialmente, são definidos:

i) A hipótese nula de ausência de dependência espacial;

ii) Uma estatística de teste que é gerada a partir do conceito de área de dependência espacial obtida no semivariograma. O teste é construído com base em simulações de fenômenos apresentando característica de efeito pepita puro, ou seja, fenômenos em que a hipótese nula de ausência de dependência espacial é verdadeira.

Por fim, é realizada a avaliação do nível nominal e o estudo do poder do teste para diferentes graus de dependência espacial.

3. RESULTADOS E DISCUSSÃO

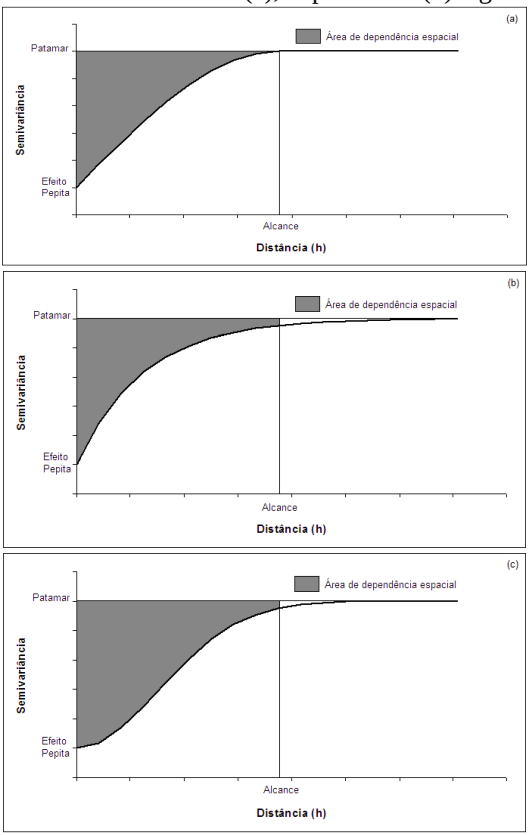
A estatística de teste a ser utilizada é a medida denominada de *ADE* (SEIDEL, 2013). Essa estatística assume seus valores de forma aleatória, caracterizando-se assim, como uma estatística de teste adequada do ponto de vista da inferência estatística.

3.1 Definição da Estatística *ADE*

É possível definir uma área, que caracteriza a dependência espacial, que é calculada com base na geometria do semivariograma, sendo entendida como a área de dependência espacial.

Então, no semivariograma, essa área está entre o patamar e o modelo teórico e entre a origem e o alcance prático. Ela permite interpretar a variabilidade espacial em termos de área, facilitando, dessa forma, por exemplo, a comparação de semivariogramas. As representações da área de dependência espacial, para semivariogramas onde é possível ajustar os modelos esférico, exponencial e gaussiano, por exemplo, estão representadas nas Figuras 2(a), 2(b) e 2(c), respectivamente.

Figura 2 – Representações da área de dependência espacial em semivariogramas com ajuste de modelos esférico (a), exponencial (b) e gaussiano (c).



Com base nas áreas de dependência espacial (Figura 2) é possível construir, por integração, uma estatística que descreve a variabilidade espacial, pois ela é a própria área de dependência espacial. Como essa estatística é a representação geométrica da dependência espacial, optou-se por defini-la como "*ADE*", que é a abreviatura de "Área de Dependência Espacial". Os cálculos completos da estatística *ADE*, para semivariogramas com ajuste dos modelos esférico, exponencial e gaussiano, podem ser encontrados em Seidel (2013).

A estatística *ADE* para um semivariograma com ajuste de modelo esférico é dada por:

$$ADE_{esf} = \text{área de dependência} = 0{,}375.\left(\hat{C}_1.\hat{a}\right), \tag{5}$$

em que $\hat{C}_1$ é a contribuição estimada e $\hat{a}$ é o alcance estimado.

Para semivariogramas com ajuste de modelo exponencial, a estatística *ADE* é dada por:

$$ADE_{\exp} = 0{,}317.\left(\hat{C}_1.\hat{a}\right). \tag{6}$$

E, para um semivariograma com ajuste de modelo gaussiano, a estatística *ADE* é dada por:

$$ADE_{gaus} = 0{,}504.\left(\hat{C}_1.\hat{a}\right). \tag{7}$$

É possível observar nas expressões (5) a (7) que cada modelo apresenta uma constante em sua respectiva estatística *ADE*. Essa constante, inerente a cada modelo, pode ser entendida como um fator de modelo que reflete a força da dependência espacial. A estatística *ADE*, em sua forma geral, é dada por:

$$ADE_{Modelo} = FM.\left(\hat{C}_1.\hat{a}\right), \tag{8}$$

em que, *FM* é o fator do modelo, $\hat{C}_1$ é a contribuição estimada e $\hat{a}$ é o alcance estimado.

A estatística *ADE* é uma variável aleatória, pois é composta por duas componentes aleatórias: a contribuição estimada ($\hat{C}_1$) e o alcance estimado ($\hat{a}$). Portanto, a estatística *ADE* assume valores de forma aleatória (cada amostra gera um valor para a estatística *ADE*).

O fator de modelo (*FM*) pode ser entendido como um valor que expressa a força da dependência espacial que o modelo pode atingir. Mais detalhes sobre o fator de modelo (*FM*) podem ser obtidos em Seidel (2013).

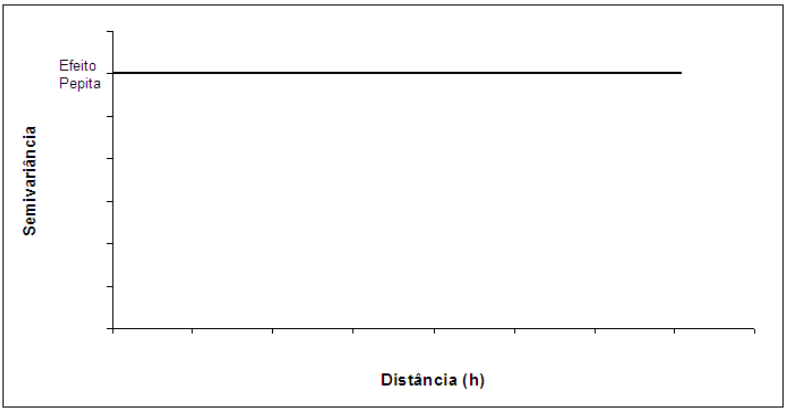
3.2 Procedimento do Teste de Dependência Espacial (*TDE*)

O procedimento do teste de dependência espacial (*TDE*) é computacional, baseado em simulações. As hipóteses a serem testadas são as seguintes:

$$\begin{cases} H_0 : \text{Não existe dependência espacial } (C_1 = 0, \forall\, h > 0) \\ H_1 : \text{Existe dependência espacial } (C_1 \neq 0, \forall\, h > 0) \end{cases}$$

As simulações utilizadas para o teste são baseadas na hipótese nula, de não existência de dependência espacial. Ou seja, são simulados semivariogramas de efeito pepita puro, caracterizando independência espacial, com base na representação da Figura 3.

Figura 3 – Representação de um semivariograma de efeito pepita puro, caracterizando a não existência de dependência espacial.



Para proceder ao teste de dependência espacial, é realizada a seguinte sequência:

a) Calcular a estatística $ADE_{calculada}$ para o teste de dependência espacial, com base no modelo ajustado ao semivariograma observado nos dados em estudo;

b) Simular, sob hipótese nula de independência espacial, *n* semivariogramas com modelo efeito pepita puro;

Neste artigo utiliza-se $n = 99$. *As simulações dos semivariogramas com modelo efeito pepita puro são*

realizadas com a função grf do geoR (RIBEIRO JÚNIOR; DIGGLE, 2001).

c) Ajustar o modelo de interesse (esférico, exponencial ou gaussiano) aos *n* semivariogramas simulados sob hipótese nula;

Para realizar os ajustes, utilizam-se, como valores iniciais, os valores de contribuição, efeito pepita e alcance, gerados pelo ajuste do semivariograma construído a partir dos dados originais em estudo.

Quando o ajuste do modelo ao semivariograma gera valor de alcance maior que o máximo comprimento do vetor de distâncias do semivariograma, atribui-se valor nulo ao alcance. Isto é feito como correção ao ajuste "superestimado", quando ocorre, aumentando a chance de aceitar a hipótese nula, ou seja, priorizando minimizar o Erro Tipo I.

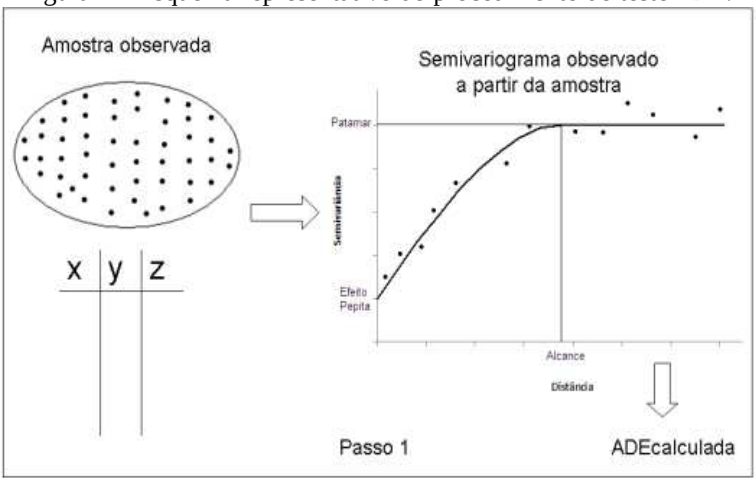
d) Calcular a estatística *ADE* em cada uma das *n* simulações;

e) Construir a distribuição da estatística *ADE* (valor de $ADE_{calculada}$ + os *n* valores de *ADE* das simulações);

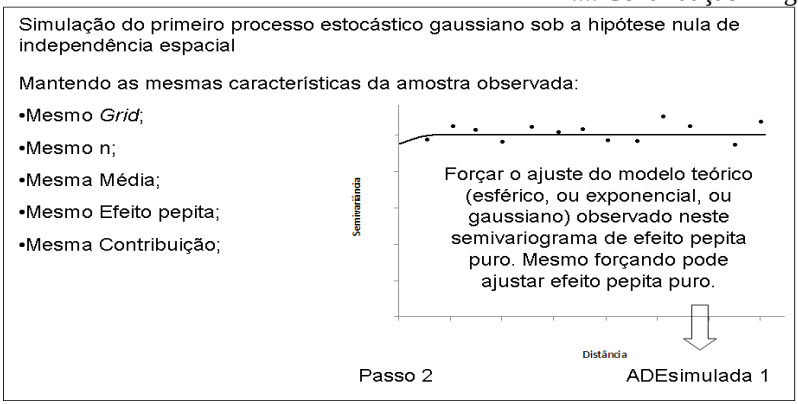
f) Calcular um valor p com base na distribuição da estatística *ADE* . O valor p é dado pela proporção de valores de *ADE* que são maiores ou iguais a $ADE_{calculada}$.

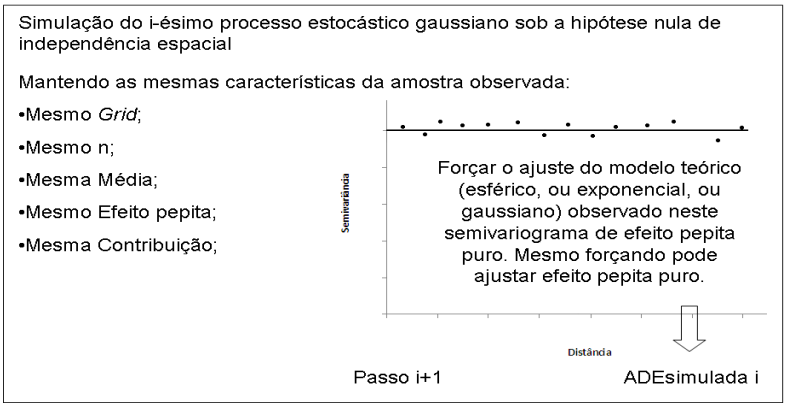
Um esquema com os passos do procedimento para a realização do teste *TDE* pode ser observado na Figura 4.

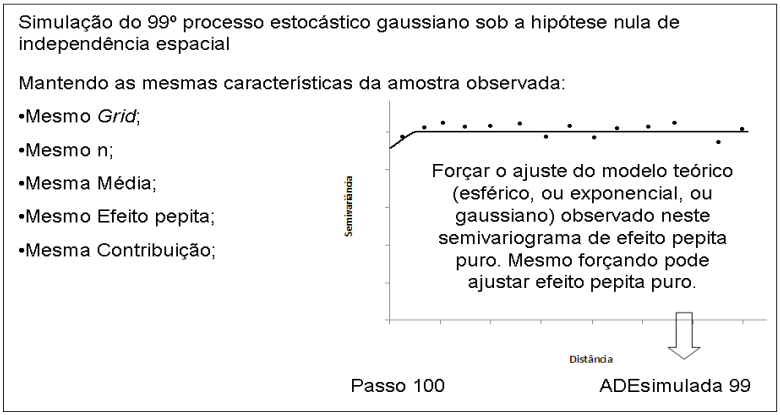
Figura 4 – Esquema representativo do procedimento do teste *TDE*.



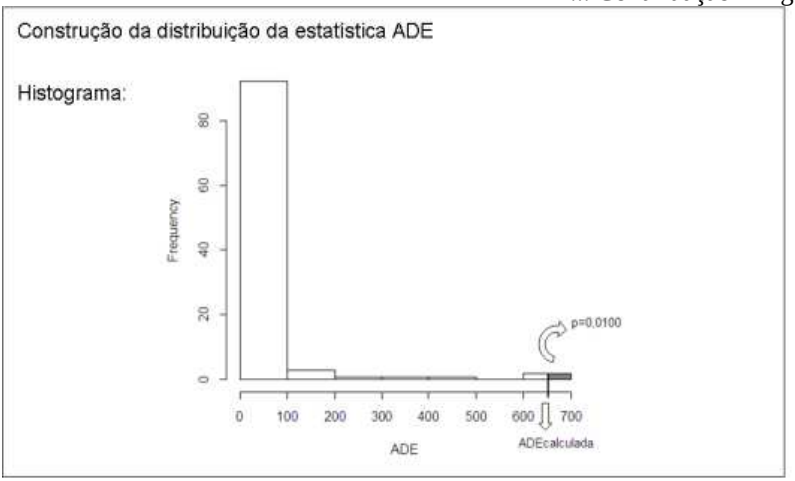
.... Continuação - Figura 4







... Continuação – Figura 4



Seguindo o esquema dado na Figura 4, realiza-se o teste e se pode concluir sobre aceitar ou não a hipótese nula (H_0), ou seja, concluir sobre a existência ou não de dependência espacial.

3.3 Avaliação do Nível Nominal do Teste *TDE*

Para realizar o estudo do nível nominal do teste, são simulados dois cenários de independência espacial para cada um dos modelos teóricos esférico, exponencial e gaussiano e mais um cenário sob modelo de efeito pepita puro. Estes cenários são considerados, pois são construídos sob a hipótese nula (de ausência de dependência espacial), isto é, são cenários em que a hipótese nula (H_0) é verdadeira. A tabela 1 mostra os cenários a serem simulados.

Tabela 1 – Cenários simulados para avaliar o nível nominal do teste.

Cenário	Parâmetros da dependência espacial			*DE (%)*
	C_0	C_1	a	
1*	50	0	0	0
2	50	0	1	0
3	45	5	10	10

* Cenário sob modelo de efeito pepita puro.

Os cenários simulados, com o pacote *Random Fields* (SCHLATHER, 2001), apresentam as seguintes características:

1º) média=0;

2º) patamar=50;

3º) n=169;
4º) *grid* regular de 100x100 m;
5º) processo estocástico gaussiano
6º) estimador clássico de semivariograma;

Em cada um dos cenários simulados, apresentados na tabela 1, realiza-se o teste 100 vezes (o teste é aplicado repetidamente 100 vezes) (3 cenários x 100 replicações = 300 simulações). Isto é feito para avaliar o nível nominal do teste. Em cada uma das 100 vezes, o valor p é calculado, e o percentual de vezes em que a hipótese nula, de independência espacial, é rejeitada define o nível nominal do teste. O nível de significância adotado é de 5% (nível nominal), ou seja, só é considerada rejeitada a hipótese nula, se o valor p é menor que 0,05. A tabela 2 mostra o resultado do estudo do nível nominal do teste.

Tabela 2 – Nível nominal do teste *TDE*, considerando ajuste dos modelos esférico, exponencial e gaussiano.

Modelo	Cenário		
	1	2	3
Esf	0,00	0,00	0,00
Exp		0,00	0,05
Gaus		0,00	0,00

A partir da tabela 2 observa-se que a hipótese nula foi aceita em todas as aplicações, com exceção ao cenário 3 para o modelo exponencial. Assim, obteve-se um nível real de 0% frente a um nível nominal de 5% para os modelos esférico e gaussiano. Já, para o modelo exponencial, obteve-se nível nominal menor ou igual a 5%. Esse fato de obter nível real menor que o nominal pode ser devido ao baixo número de repetições, ou ser o verdadeiro comportamento do teste.

3.4 Estudo do Poder do Teste *TDE*

Para realizar o estudo do poder do teste, são simulados 12 cenários de dependência espacial para cada um dos modelos teóricos esférico, exponencial e gaussiano. Nestes cenários, diferentes valores de contribuição, efeito pepita e alcance são atribuídos. Estes cenários são considerados, pois são construídos sob a hipótese alternativa (de existência de dependência espacial), isto é, são cenários em que a hipótese alternativa (H_1) é verdadeira. A tabela 3 mostra os cenários a serem simulados.

Os cenários simulados, com o pacote *Random Fields* (SCHLATHER, 2001), apresentam as seguintes características:

1º) média=0;
2º) patamar=50;
3º) n=169;
4º) *grid* regular de 100x100 m;

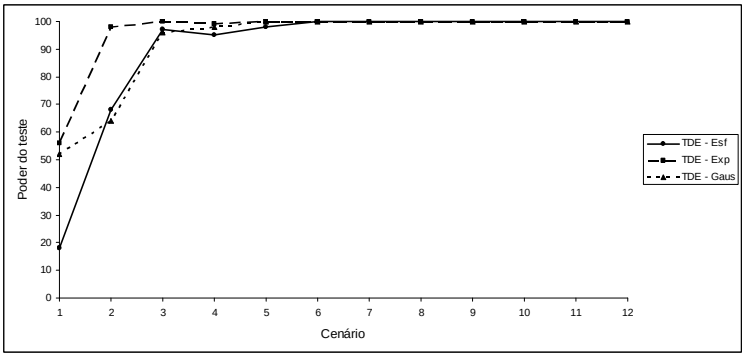
5º) processo estocástico gaussiano
6º) estimador clássico de semivariograma;

Tabela 3 – Cenários simulados para avaliar o poder do teste.

Cenário	Parâmetros da dependência espacial			*DE (%)*
	C_0	C_1	a	
1	37,5	12,5	25	25
2	37,5	12,5	50	25
3	37,5	12,5	75	25
4	25	25	25	50
5	25	25	50	50
6	25	25	75	50
7	12,5	37,5	25	75
8	12,5	37,5	50	75
9	12,5	37,5	75	75
10	5	45	25	90
11	5	45	50	90
12	5	45	75	90

Em cada um dos cenários simulados, apresentados na tabela 3, realiza-se o teste 100 vezes (o teste é aplicado repetidamente 100 vezes) (12 cenários x 100 replicações = 1200 simulações). Isto é feito para avaliar o poder do teste. Em cada uma das 100 vezes, o valor p é calculado, e o percentual de vezes em que a hipótese nula, de independência espacial, é corretamente rejeitada define o poder do teste. O nível de significância adotado é de 5% (nível nominal), ou seja, só é considerada rejeitada a hipótese nula, se o valor p é menor que 0,05. A Figura 5 apresenta o resultado do estudo do poder do teste em relação aos cenários simulados.

Figura 5 – Poder do teste *TDE*, considerando ajuste dos modelos esférico, exponencial e gaussiano.



Com base na Figura 5 pode-se observar que o poder cresce rapidamente para menores valores de *DE(%)* e se aproxima de 100% para maiores valores de *DE(%)*, conforme os cenários simulados. Ou seja, para os cenários com baixa dependência espacial o poder do teste é menor, e para aqueles cenários com maior grau de dependência, o poder do teste é maior.

O teste proposto aqui vem a ser um complemento à avaliação da dependência espacial, sendo que pode ser utilizado conjuntamente com outros métodos já existentes, como, por exemplo, envelopes simulados e indicadores de dependência espacial.

4. CONCLUSÃO

A partir do conceito de áreas de dependência espacial foi possível construir uma nova estatística de dependência espacial, denominada de *ADE* , utilizada para o teste de hipótese de dependência espacial.

O teste construído, denominado de *TDE* , utiliza o princípio de simulação, sob hipótese nula de independência espacial, gerado a partir de dados sob comportamento de efeito pepita puro (modelo de efeito pepita puro).

O teste teve bom poder, sendo que este tende a 100% quando aumenta a dependência espacial do fenômeno em estudo.

Para futuros estudos necessita-se ampliar o número de cenários, o número de simulações e ampliar os estudos do nível nominal e do poder para se atingir uma melhor avaliação do teste.

AGRADECIMENTOS

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro em forma de bolsa.

REFERÊNCIAS BIBLIOGRÁFICAS

ALVES, M. C. et al. Geostatisticalanálisis of the spatial variation of the Berry borer and leaf miner in a coffee agroecosystem. *Precision Agriculture*, v. 12, n. 1, p. 18-31, 2011.

BIONDI, F.; MYERS, D. E.; AVERY, C. C. Geostatistically modeling stem size and increment in an old-growth forest. *Canadian Journal of Forest Research*, v. 24, n. 7, p. 1354-1368, 1994.

CAMBARDELLA, C. A. et al. Field-scale variability of soil properties in Central Iowa soils. *Soil Science Society AmericaJournal*, v. 58, n. 5, p. 1501-1511, 1994.

CARVALHO, J. R. P.; SILVEIRA, P. M.; VIEIRA, S. R. Geoestatística na determinação da variabilidade espacial de características químicas do solo sob diferentes preparos. *Pesquisa Agropecuária Brasileira*, v. 37, n. 8, p. 1151-1159, 2002.

JOURNEL, A. G.; HUIJBREGTS, C. J. *Mining geostatistics*. Caldwell: Blackburn, 2003. 600 p.

MATHERON, G. Principles of geostatistics. *Economic Geology*, v. 58, p. 1246-1266, 1963.

OLEA, R. A. A six-step practical approach to semivariogram modeling. *Stochastic Environmental Research and Risk Assessment*, v. 20, n. 5, p. 307-318, 2006.

R DEVELOPMENT CORE TEAM. *R*: a language and environment for statistical computing. Vienna: R Foundation for StatisticalComputing, 2012. Software.

RESENDE, M. D. V. *Matemática e estatística na análise de experimentos e no melhoramento genético*. Colombo: EMBRAPA Florestas, 2007. 362 p.

RIBEIRO JÚNIOR, P. J.; DIGGLE, P. J. GeoR: a package for geostatisticalanalysis. *R News*, v. 1, n. 2, p. 15-18, 2001.

SCHLATHER, M. Simulation and analysis of random fields.*R News*, v. 1, n. 2, p. 18-20, 2001.

SEIDEL, E. J. *Novas contribuições para avaliação e descrição da estrutura de dependência espacial em geoestatística*. 2013. 146 p. Tese (Doutorado em Estatística e Experimentação Agropecuária), Universidade Federal de Lavras, Lavras, 2013.

SEIDEL, E. J.; OLIVEIRA, M. S. Proposta de uma generalização para os modelos de semivariogramas exponencial e gaussiano. *Semina: Ciências Exatas e Tecnológicas*, v. 34, n. 1, p. 125-132, 2013.

(Recebido em maio de 2014. Aceito em junho de 2014).