Revista da Estatística UFOP, Vol III(3), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS
Trabalhos apresentados na versão poster.

Utilização de diferentes estimadores de semivariância com modelo teórico exponencial

Marcilia Bruna dos Reis Teixeira [1] [2]
João Domingos Scalon [3]

1 Introdução

A Geoestatística aborda a detecção da estrutura da variabilidade espacial em superfícies contínuas. Para tanto, comumente utiliza-se o semivariograma, que é um gráfico das semivariâncias (γ) em função das distâncias (h) entre os pares de pontos.

Um dos estimadores de semivariância mais utilizados é o modelo clássico proposto por Matheron. Segundo Genton (1998) e Mingoti e Rosa (2008), este estimador é muito afetado pela presença de *outliers* nos dados, não apresentando propriedades de robustez.

Existem outros estimadores na literatura, tendo muitos sido elaborados com o intuito de serem mais robustos, como os de Cressie e Hawkins (1980), que foi construído para ser robusto contra *outliers*; das Medianas (CRESSIE, 1993); o estimador altamente robusto de Genton (1998), que busca ser mais robusto diante de *outliers*; o estimador das diferenças proposto por Haslett (1997), utilizado num contexto de séries temporais, especialmente para dados não estacionários; e os estimadores New-1 e New-2 (LI; LAKE, 1994) que, segundo os autores, são consideravelmente robustos.

É importante utilizar estimadores que levem em conta o máximo possível da informação disponível nos dados, obtendo modelos que representem melhor a estrutura de dependência espacial estudada. Assim, torna-se interessante e útil a comparação destes estimadores diante de diferentes características do banco de dados. Em Teixeira e Scalon (2013) foram realizadas comparações entre estimadores utilizando o modelo teórico esférico. Este trabalho apresenta uma complementação do trabalho anterior, sendo utilizado o modelo exponencial.

2 Material e Métodos

As comparações entre os estimadores foram realizadas por meio de simulações. Sendo geradas malhas regulares quadradas, com dois diferentes tamanhos de *gride*: 5x5 e 10x10. Nas

[1]DEX - UFLA. e-mail: *marciliabruna@yahoo.com.br*
[2]Agradecimento à CAPES e à FAPEMIG pelo apoio financeiro.
[3]DEX - UFLA. e-mail: *scalon@dex.ufla.br*

simulações foi utilizado o modelo teórico exponencial:

$$\gamma(h) = \begin{cases} 0, & se \;\; h = 0 \\ \gamma(h) = C_0 + C_1 \left[1 - e^{[-3(h/a)]}\right], & se \;\; h \neq 0 \end{cases} \tag{1}$$

em que $\gamma(h)$ é a semivariância para um determinado h; C_0 é o efeito pepita; C_1 é a contribuição; a é o alcance e h são os valores das distâncias.

Também foi estudada a influência de *outliers* nas estimativas obtidas por cada estimador. Assim, foram realizadas contaminações no dados com valores discrepantes, sendo adotadas quatro porcentagens de contaminação: 0%, 1%, 5% e 10% de *outliers*. É importante observar que no banco de dados com tamanho 25, as porcentagens de dados discrepantes não correspondem a valores inteiros. Assim, foi considerado o menor inteiro superior ao valor calculado.

Todas as simulações foram feitas no *software* R (R DEVELOPMENT CORE TEAM, 2013), utilizando-se o pacote *Randomfields* (SCHLATHER, 2014). Foram comparados os seguintes estimadores de semivariância:

- **Estimador clássico de Matheron**

$$2\hat{\gamma}_M(h) = \frac{1}{N(h)} \sum_{i=1}^{N(h)} (Z(x_i + h) - Z(x_i))^2 \tag{2}$$

- **Estimador robusto de Cressie e Hawkins**

$$2\hat{\gamma}_{CH}(h) = \left[\frac{1}{N(h)} \sum_{i=1}^{N(h)} |(Z(x_i + h) - Z(x_i))|^{\frac{1}{2}})\right]^4 / \left(0,457 + \frac{0,494}{N(h)}\right) \tag{3}$$

- **Estimador das medianas de Cressie**

$$2\hat{\gamma}_{Md}(h) = \frac{med\left[|(Z(x_i + h) - Z(x_i)|^{\frac{1}{2}}\right]^4}{0,457} \tag{4}$$

em que $med\{\bullet\}$ denota a mediana da sequência $\{\bullet\}$

- **Estimador das diferenças de Haslett**

$$2\hat{\gamma}_H(h) = \frac{1}{N(h) - 1} \sum_{i=1}^{N(h)} (d_{hi} - \overline{d_h})^2 \tag{5}$$

em que $d_{hi} = (Z(x_i + h) - Z(x_i))$.

- **Estimador altamente robusto de Genton**

$$2\hat{\gamma}_G(h) = (Q_{N(h)})^2 \tag{6}$$

em que $Q_{N(h)} = 2,2191\{(|V_i(h) - V_j(h)|; i < j\}_{(k)}$; $V(h) = Z(x+h) - Z(x)$; 2,2191 é a consistência da distribuição gaussiana e $k = \left(\frac{[\frac{N(h)}{2}]+1}{2}\right)$;

$[\frac{N(h)}{2}]$ denota a parte inteira de $\frac{N(h)}{2}$.

- **Estimadores New-1**

$$2\hat{\gamma}_{N1}(h) = \frac{2}{n}\sum_{i=1}^{n}\left\{\frac{1}{2m}\sum_{j\in D_{i,h}}(Z(x_i) - Z(x_j))^2\right\} \tag{7}$$

em que n é o número total de dados; $D_{i,h}$ é o índice de um conjunto de valores de dados em uma janela móvel $\Delta i, h$ (de tamanho h centrada no ponto bloco i), excluindo o ponto x_i e m é o número de dados em $D_{i,h}$.

- **Estimadores New-2**

$$2\hat{\gamma}_{N2}(h) = 2(\hat{\gamma}_{N1}(h) + \frac{h}{d}\hat{\gamma}'_{N1}(h)); \tag{8}$$

em que $\hat{\gamma}'_{N1}(h)$ é a derivada de $\hat{\gamma}_{N1}(h)$ em relação a h, calculada pelo Método da Diferença Central; h vetor de distâncias; e d é a dimensão no espaço euclidiano.

As comparações foram realizadas por meio do Erro Médio Quadrático (EMQ). Foram consideradas duas situações: na primeira utilizou-se todas as estimativas, na segunda, foi considerado um *cutoff* de 50%.

2.1 Resultados e discussões

Para uma melhor interpretação dos resultados obtidos nos dois *grides*, foram construídos gráficos das quantidades de contaminação *versus* o EMQ. Cada linha representa um estimador: Matheron (M), Cressie e Hawkins (CH), Medianas (MD), Hastlett (H), Genton (G), New1 (N1) e New2 (N2).

Na Figura 1 são apresentados os resultados obtidos nos bancos de dados de tamanho 25. Nela, é possível perceber a influência da utilização do *cutoff*, o qual diminui, consideravelmente, o erro da maioria dos estimadores. Esta diferença é muito realçada, principalmente, nos estimadores altamente robusto de Genton e das medianas de Cressie.

Isto acontece porque as estimativas das maiores distâncias são pouco confiáveis, pois utilizam poucos pares de pontos. Este problema foi bem controlado pelos estimadores New1 e New2. Estes estimadores utilizam todas as observações em cada estimativa.

Também, pode-se perceber que o estimador New-1, na maior parte do tempo, obteve os menores erros. Os estimadores de Genton, Cressie e Hawkins e das Medianas apresentaram um mal comportamento para os casos sem contaminação. Sendo que seus resultados apresentaram melhoras nos casos de contaminação.

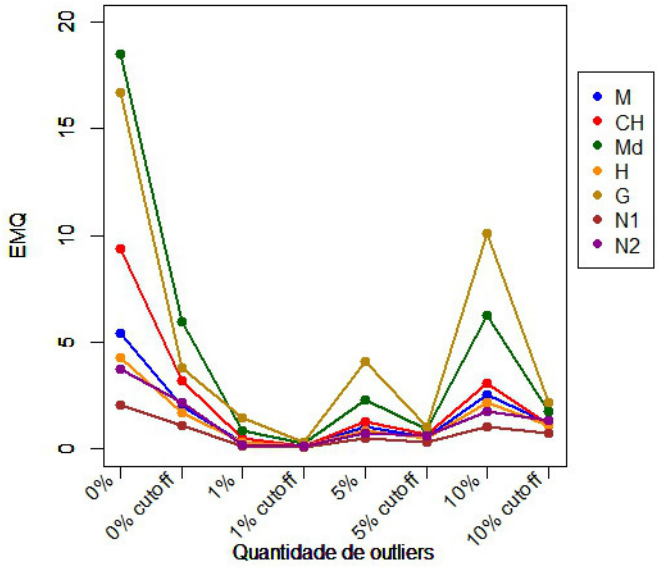

Figura 1: EMQ *versus* quantidade de *outliers* dos dados com n=25

Os resultados, referentes às análises dos bancos de dados de tamanho 100 estão dispostas na Figura 2. Pode-se perceber que as contaminações de 10% de *outliers* causaram forte influência nos resultados, ocasionando um aumento nos erros.

O estimador New-1 apresentou o melhor desempenho em quase todas as situações, sendo seguido de perto pelo estimador New-2. A utilização de *cutoff* apresentou uma forte influência nos resultados, reduzindo os erros, principalmente no estimador de Genton e no das Medianas.

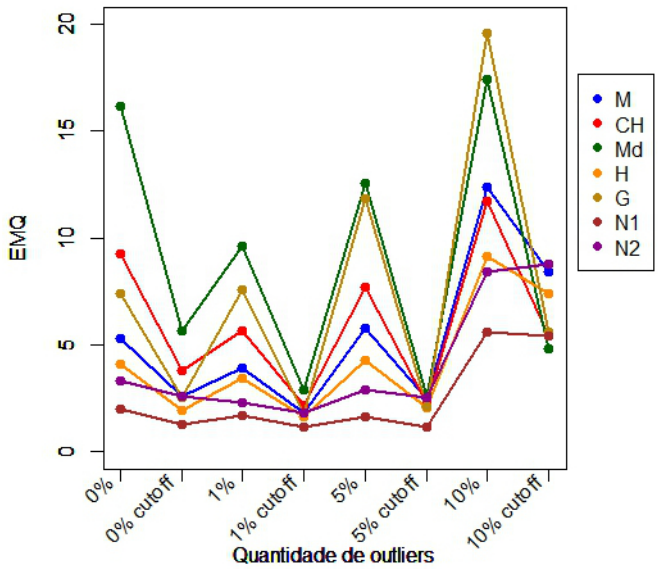

Figura 2: EMQ *versus* quantidade de *outliers* dos dados com n=100

2.2 Conclusões

Por meio das simulações realizadas, foi possível verificar a grande influência da utilização de *cutoff*, o qual reduz consideravelmente os erros. Contudo, a utilização do *cutoff* não foi rele-

vante para a comparação entre os estimadores. O estimador New-1, de forma geral, apresentou os melhores resultados.

A literatura, muitas vezes, apresenta o estimador de Matheron como o único estimador existente. Contudo, com o presente trabalho, mostrou-se que, dependendo das condições dos dados, a escolha do estimador de semivariância tem grande influência no resultado final. Assim, para redução dos erros e melhor construção do semivariograma, recomenda-se cautela e atenção na escolha do estimador de semivariância, sempre considerando as caracteristucas dos dados estudados.

Referências

[1] CRESSIE, N. A. C.. **Statistics for spatial data**. New York: J. Wiley, 1993. 900 p.

[2] CRESSIE, N. A. C.; HAWKINS, D. M. Robust estimation of the variogram. **Mathematical Geology**, New York, v. 12, n. 2, p. 115-125, 1980.

[3] GENTON, M. G. Highly robust variogram estimation. **Mathematical Geology**, New York, v. 30, n. 2, p. 213-221, 1998.

[4] HASLETT, J. On the sample variogram and the sample autocovariance for non-stationary time series. **The Statistician**, Washington, v. 46, n. 4, p. 475-485, 1997.

[5] LI, D.; LAKE, L. W. A moving window semivariance estimator. **Water Resources Research**, Washington, v. 30, n. 5, p. 1479-1490, May 1994.

[6] MINGOTI, S. A.; ROSA, G. A note on robust and non-robust variogram estimators. **Revista Escola de Minas**, Ouro Preto, v. 61, n. 1, p. 87-95, 2008.

[7] R DEVELOPMENT CORE TEAM. **R**: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2013. Disponível em: http://www.R-project.org.

[8] Martin Schlather, Alexander Malinowski, Marco Oesting, Daphne Boecker, Kirstin Strokorb, Sebastian Engelke, Johannes Martini, Peter Menck, Sebastian Gross, Katharina Burmeister, Juliane Manitz, Richard Singleton, Ben Pfaff and R Core Team (2013). **RandomFields: Simulation and Analysis of Random Fields**. R package version 3.0.5. http://CRAN.R-project.org/package=RandomFields

[9] TEIXEIRA, M. B. R.; SCALON, J. D. Comparação entre estimadores de semivariância, **Rev. Bras. Biom.** v.31, n.2, abr.-jun. 2013, 23p.