Algoritmo de Semivariância para *Big Data*

André Felipe Berdusco Menezes (PIC/UEM), Diogo Rossoni (Orientador),
e-mail: andrefelipemaringa@gmail.com.
Universidade Estadual de Maringá / Centro de Ciências Exatas

Palavras-chave: Geoestatística, Estimação, Simulação.

Resumo:

Em virtude do avanço das tecnologias de informação e captação de dados, os grandes conjuntos de dados (Big Data), apresentam-se cada vez mais frequentes em análises estatísticas. No que tange dados espacialmente correlacionados, as ferramentas da metodologia Geoestatística encontram dificuldades técnicas e computacionais ao lidar com Big Data. Entre elas a semivariância, uma medida de dissimilaridade, indispensável na interpolação de dados não amostrados (krigagem). Desta forma o presente trabalho propõe um método de estimação da semivariância conjuntamente com a otimização computacional. O algoritmo de semivariância para Big Data consiste em: retirar *k* amostras (subamostras) de tamanho *b* do conjunto de dados; para cada subamostra calcula-se a semivariância para determinadas distâncias; por fim a nova estimativa de semivariância é obtida pela média aritmética das *k* semivariâncias (em cada distância). Realizaram-se estudos de simulação e análise em banco de dados reais, no qual se comparou o algoritmo proposto e o método clássico de semivariância, evidenciando melhor desempenho do estimador proposto quanto ao custo computacional.

Introdução

Na geoestatísca, consideramos as observações sendo realizações de um processo estocástico, isto é, para cada localização xi amostrada têm-se uma variável aleatória Z distinta. Entre seus propósitos a geoestatística, estudo a compreensão da variabilidade espacial, um fator imprescindível para estudos posteriores, como por exemplo, realizar predições. Dessa forma, procede-se uma modelagem sobre o fenômeno para determinar e quantificar a variabilidade espacial. No entanto, para se ajustar uma

distribuição teórica é inevitável fazer uso de medidas de associação, tais como covariância e principalmente semivariância.

A semivariância é uma medida de dissimilaridade, ou seja, seu valor é maior à medida que as variáveis estão menos associadas. Na prática conhecemos algumas realizações do processo espacial, assim devemos estimar a semivariância com base nessas realizações. Neste projeto tivemos foco sobre o estimador clássico de semivariância proposto por Matheron (1962), uma vez que é o estimador mais comum na literatura.

Com o avanço das tecnologias de informação e captação de dados, os grandes conjuntos de dados (*Big Data*), tornaram-se cada vez mais presentes em análises estatísticas. No que tange a geoestatística, em específico a semivariância, o método clássico para sua estimação é considerado computacionalmente maçante. Assim sendo, buscou-se neste projeto a proposição e implementação de um algoritmo para redução do custo computacional na estimação da semivariância.

Objetivou-se neste projeto a proposição e implementação de um algoritmo para redução do custo computacional na estimação da semivariância. Dessa forma, o algoritmo de semivariância para Big Data consiste em: retirar "*k*" amostras (subamostras) de tamanho "*b"* do conjunto de dados; para cada subamostra calcula-se a semivariância para determinadas distâncias; por fim a nova estimativa da semivariância é obtida pela média aritmética das k semivariâncias.

Materiais e métodos

Desenvolvido por Matheron (1962) a partir do método dos momentos, o estimador clássico de semivariância é definido pela seguinte expressão:

$$\hat{\gamma}(h) = \frac{1}{2\,N(h)} \sum_{i=1}^{N(h)} [z(x_i) - z(x_i + h)]^2$$

Sendo:

- $z(x_i)$ a realização da função aleatória *Z* no ponto x_i ;
- $z(x_i + h)$ a realização da função aleatória *Z* no ponto x_i mais uma distância *h;*
- *h* a distância entre as observações;
- $N(h)$ o número de pares de valores medidos, separados por uma distância *h.*

Dada uma amostra de tamanho *n,* proveniente de um processo estocástico espacial, a estimação da semivariância por meio do algoritmo proposto consiste em:

i. Selecionar aleatoriamente e sem reposição uma subamostra de tamanho *b*, tal que $b < n$;
ii. A partir da subamostra de tamanho b procede-se com o cálculo da semivariância pelo estimador de Matheron;
iii. Repete-se a etapa (i) e (ii) *k* vezes, gerando um vetor de semivariâncias para cada nova subamostra de tamanho *b*;
iv. Ao final a semivariância será definida como:

$$\hat{\gamma}(h) = k^{-1} \sum_{i=1}^{k} \gamma(h)_{bi}$$

Em que:

- *b:* tamanho da subamostra;
- *k:* número de subamostras ou número de iterações;
- *h:* vetor distância;
- $\gamma(h)_{bi}$ semivariância da distância *h* da subamostra de tamanho *b,* na i-ésima iteração;

A implementação computacional, foi realizada no ambiente estatístico R, com auxilio da biblioteca ***geoR***.

Com intuito de comparar a performance, isto é, o tempo computacional percorrido pelo método clássico e o algoritmo proposto para estimação da semivariância foi conduzido um estudo de simulação, variando o tamanho amostral n = 5000, 10000 e 15000. Para cada n foi gerado M = 1000 realizações de um processo espacial com modelo gaussiano e os seguintes parâmetros: C = 60 (patamar), a = 30 (alcance) e C_0 = 0 (efeito pepita).

O estimador clássico, foi calculado através da função ***variog*** da biblioteca ***geoR***. Já na execução do algoritmo consideramos 100 subamostras com tamanho 100, isto é, k = b = 100.

Resultados e Discussão

Na Figura 1, observamos a vantagem computacional ao utilizar o algoritmo proposto para estimar a semivariância. Primeiramente, verifica-se que conforme o tamanho da amostra aumenta maior o tempo computacional na estimação. Analisando para amostras com 15000 observações, o

estimador de Matheron demorou em média aproximadamente 15 segundos, em contrapartida, o algoritmo para *Big Data* demorou em média 0,98 segundos.

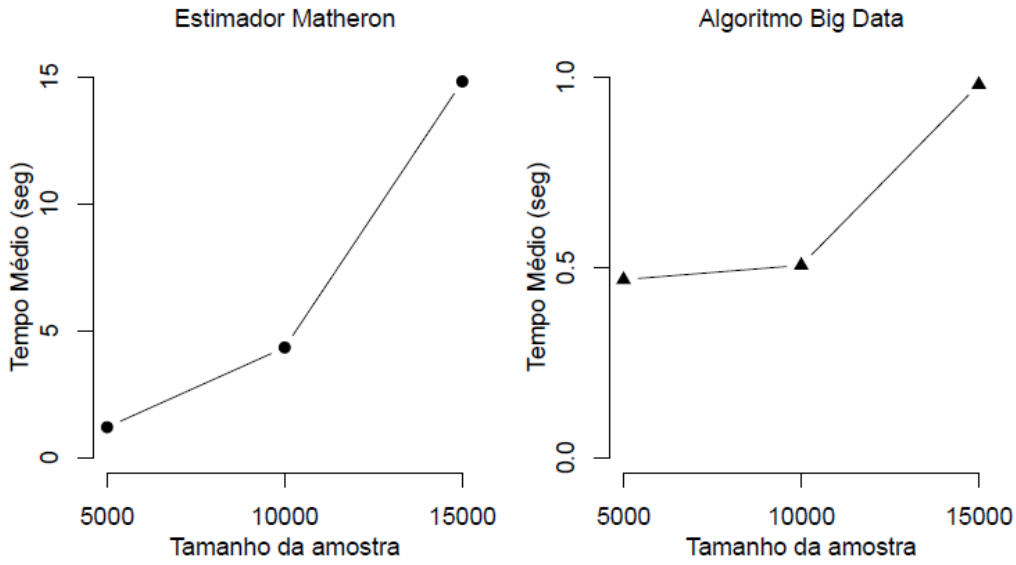

Figura 1 – Tempo médio de estimação.

Conclusões

A fim de reduzir o custo computacional na estimação da semivariância para grandes conjuntos de dados (Big Data), foi proposto uma método baseado na reamostragem. De acordo com os resultados obtidos pelo estudo de simulação verificou-se que o algoritmo proposto oferece resultados melhores comparado ao método clássico.

Referências

ISAAKS, E. H. et al. **Applied geostatistics.** [S.l.]: Oxford University Press, 1989

RIBEIRO JÚNIOR, Paulo. J.; DIGGLE, Peter J. **geoR: Analysis of Geostatistical** Data, 2015. R package version 1.7-5.1.

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