Proposta de Estimador de Semivariância para *Big Data*

André Felipe B. Menezes
andrefelipemaringa@gmail.com
Acadêmico DES – UEM

Diogo Francisco Rossoni
diogo.rossoni@gmail.com
Docente DES – UEM

INTRODUÇÃO

Com o avanço das tecnologias de informação e captação de dados, os grandes conjuntos de dados (*Big Data*), tornaram-se mais frequentes. Dados espacialmente correlacionados são frequentemente coletados em âmbito mundial, como por exemplo, índices pluviométricos, de temperatura e dados de mineração.
Entretanto, as ferramentas clássicas da metodologia Geoestatística (um ramo da estatística espacial) encontram dificuldades computacionais ao lidar com *Big Data*. Entre elas a semivariância, uma medida de dissimilaridade indispensável na interpolação de dados não amostrados. Desta forma o presente trabalho propõe um método de estimação da semivariância conjuntamente com a otimização computacional.

METODOLOGIA

Um dos principais objetivos da Geoestatística é compreender a variabilidade de fenômenos espacias, para isso presume-se que o valor de um ponto no espaço está relacionado, de alguma forma, com valores de pontos situados a certa distância, sendo provável supor que a influência é tanto maior quanto menor for a distância entre eles, isto chamamos de dependência espacial.

A função variograma ou semivariância é responsável por fornecer o grau de dependência espacial entre duas amostras separadas por uma distância h. A magnitude da semivariância entre dois pontos depende da distância entre eles. Usualmente valores próximos apresentam variações pequenas, ao passo que, valores distantes apresentam variações bruscas.

O estimador clássico da semivariância, desenvolvido por Matheron (1962), é definido como:

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

Sendo:

- $Z(x_i)$ a realização da *v.a.* Z no ponto x_i;
- $Z(x_i + h)$ a realização da *v.a.* Z no ponto x_i mais o incremento h;
- h a distância entre os pontos;
- $N(h)$ o número de pares de valores medidos, $Z(x_i)$ e $Z(x_i + h)$, separados pela distância h.

Ao se deparar com grandes conjuntos de dados, o custo computacional para o cálculo da semivariância é demasiado, logo é proposto o seguinte estimador:

1. Dada uma amostra inicial de tamanho n, seleciona-se aleatoriamente e sem reposição uma subamostra de tamanho b, tal que $b < n$;
2. A partir da subamostra de tamanho b procede-se com o cálculo da semivariância de acordo com a Equação (1);
3. Repete-se a etapa 1 e 2 no mínimo 100 vezes, gerando um vetor de semivariâncias para cada nova subamostra de tamanho b;
4. Ao final a semivariância será definida como:

$$\hat{\gamma}(h)_b = k^{-1} \sum_{i=1}^{k} \gamma(h)_{bi} \quad (2)$$

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.

REFERÊNCIAS

LIANG, Faming. et al. **A resampling-based stochastic approximation method for analysis of large geostatistical data.** Journal of the American Statistical Association, 108(501):325–339, 2013.

YAMAMOTO, Jorge Kazuo; LANDIM, Paulo M. Barbosa. **Geoestatística: conceitos e aplicações.** Oficina de Textos, 2015.

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

R Development Core Team (2015). **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

RESULTADOS

Simulação: Foram simuladas 10000 observações distribuídas uniformemente em uma região $[0, 100] \times [0, 100]$, com modelo gaussiano, $\sigma^2 = 100$ e $\phi = 40$.

Definiu-se $k = 200$ e $b = 100$, isto é, 200 subamostras de tamanho 100.

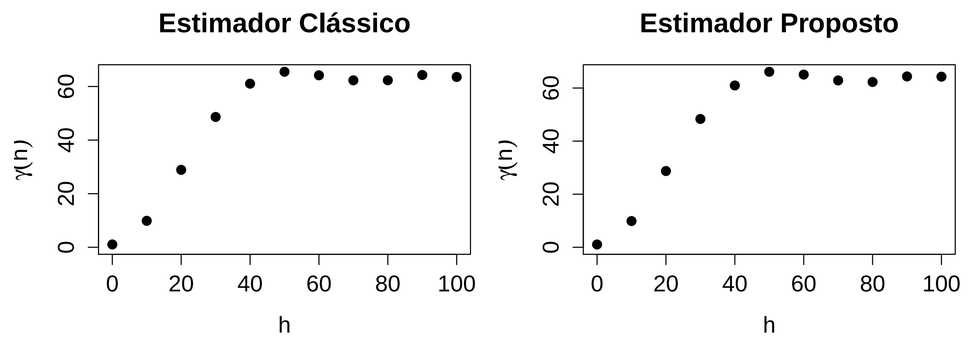
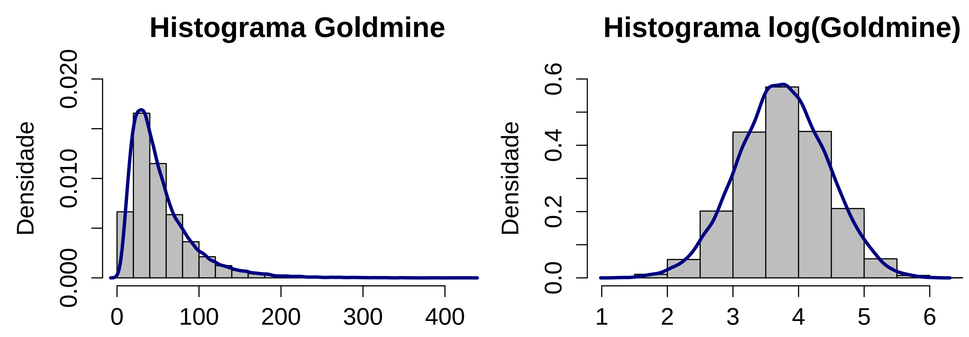


Tabela 1: Comparação entre os estimadores

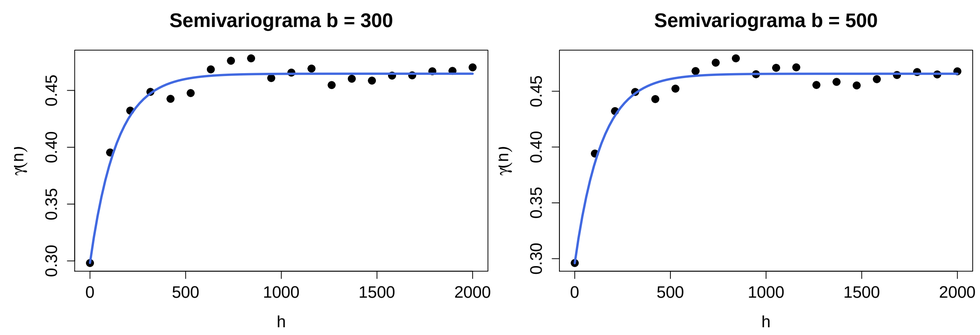
h	0,00	20,00	40,00	60,00	80,00	100,00
$\gamma(h)$	1,06	28,89	61,05	64,18	62,33	63,55
$\gamma(h)_b$	1,05	28,73	60,97	65,07	62,28	64,27
$\lvert\gamma(h) - \gamma(h)_b\rvert$	0,01	0,16	0,08	0,89	0,05	0,72

Verifica-se pequenas diferenças entre as semivariâncias calculadas pelo estimador clássico e proposto. O tempo decorrido em segundos do estimador de Matheron variou entre 5,20 e 5,35, enquanto que, do estimador proposto esteve entre 1,18 e 1,50.

Dados reais (Goldmine Samples): Este conjunto de dados consiste de 22577 observações de posições tomadas a partir de uma verdadeira mina de ouro, em que medem o grau de ouro em gramas. Uma transformação logarítmica foi feita de tal modo que os dados se aproximassem da distribuição normal.



Os semivariogramas abaixo foram calculados e modelados a partir do estimador proposto, com diferentes tamanho de subamostra, porém $k = 200$.



O tempo decorrido quando $b = 300$ ficou entre 2,90 à 3,22 segundos, por outro lado, quando $b = 500$ o custo computacional aumentou ficando entre 5,85 à 6,17 segundos. Embora o tamanho das subamostras aumente, não há diferenças significativas nas estimativas das semivariâncias.

Realizou-se a modelagem da dependência espacial, utilizando o método dos mínimos quadrados, em que o modelo ajustado foi o exponencial. Para $b = 300$ obteve-se: $\sigma^2 = 0,1664$ e $\phi = 139,6425$, já para $b = 500$ os parâmetros foram: $\sigma^2 = 0,1698$ e $\phi = 139,4035$.

CONCLUSÕES

Diante dos estudos realizados, constata-se que o estimador proposto oferece bons resultados quando aplicado em *Big Data*, uma vez que, ele proporciona estimativas com pequenos erros e redução no custo computacional. Para a implementação do estimador deve-se levar em consideração o tamanho e a quantidade de subamostras. Sendo observado que a estimativa da semivariância se aproxima da real quando o tamanho ou número de subamostras aumentam, contudo há o crescimento no tempo computacional.