Estimador de Semivariância para *Big Data*

André Felipe Menezes

Universidade Estadual de Maringá

Departamento de Estatística

7 de Outubro de 2016

Organização

1. Aspectos gerais da Geoestatística
   - Fenômeno espacial
   - Semivariância
2. Estimador de Semivariância para Big Data
   - Estudo de simulação
   - Aplicação em dados reais
3. Conclusões

Aspectos gerais da Geoestatística

Aspectos gerais

A Geoestatística é um ramo da estatística espacial, cujo os principais propósitos são:

- Compreensão de fenômenos espaciais;
- Identificar e quantificar a variabilidade espacial;
- Predizer observações não amostradas.

Fenômeno espacial

O que é?

Defini-se $Z(x_i)$ um fenômeno espacial como:

$$Z(x_i) = \mu(x_i) + \varepsilon'(x_i) + \varepsilon. \qquad (x_i \in \mathbb{R}^n) \tag{1}$$

Sendo:

- $\mu(x_i)$ uma função determinística que representa a componente estrutural;
- $\varepsilon'(x_i)$ um termo estocástico que varia localmente e é espacialmente correlacionado;
- ε um ruido aleatório, não correlacionado.

Semivariância

Definição

A semivariância é definida como uma medida de dissimilaridade, na qual fornece o grau de dependência espacial entre duas amostras separas por uma distância h.

Semivariância teórica

Matheron (1962) e Cressie (1993, p.58) definiram a semivariância teórica $\gamma(h)$ sendo:

$$\gamma(h) = \frac{1}{2}\,\mathrm{Var}(Z(x+h) - Z(x)) \tag{2}$$

Semivariância

Semivariância teórica

Sob hipótese de estacionariedade de 2° ordem ou hipótese intrínseca, temos que $E[Z(x)] = \mu, \; \forall \, x$. Logo têm-se:

$$
\begin{aligned}
\gamma(h) &= \frac{1}{2}\,\mathrm{Var}(Z(x+h) - Z(x)) \\
&= \frac{1}{2}\big(E[(\,Z(x+h) - Z(x)\,)^2] - (E[Z(x+h) - Z(x)])^2\big) \\
&= \frac{1}{2}\big(E[(\,Z(x+h) - Z(x)\,)^2] - (E[Z(x+h)] - E[Z(x)])^2\big) \\
&= \frac{1}{2}\big(E[(\,Z(x+h) - Z(x)\,)^2] - (\mu - \mu)^2\big) \\
\therefore \; \gamma(h) &= \frac{1}{2}\,E[(\,Z(x+h) - Z(x)\,)^2]
\end{aligned}
$$

Semivariância

Semivariância de nuvem

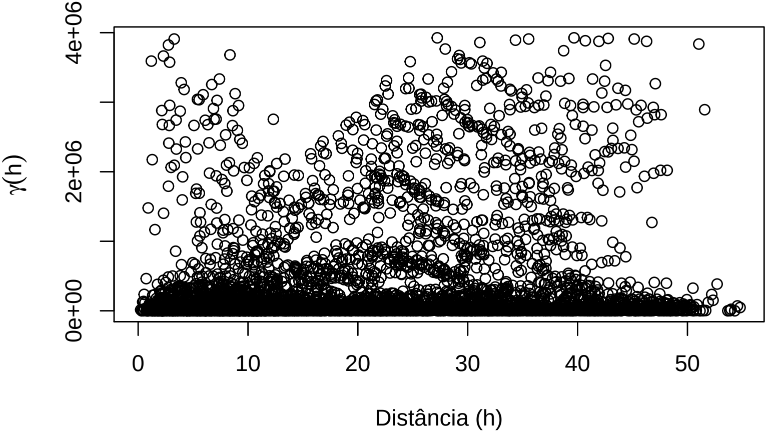
Utilizado na análise exploratória, a semivariância de nuvem é definida como:

$$\hat{\gamma}(h) = \frac{\left(Z(x_i) - Z(x_i + h)\right)^2}{2} \tag{3}$$

em que:

- $Z(x_i)$ é a realização da variável aleatória Z no ponto x_i;
- $Z(x_i + h)$ é a realização da variável aleatória Z no ponto x_i mais uma distância h;
- h é a distância entre as observações;

O variograma de nuvem, produz um gráfico de dispersão entre os valores da variograma e a os $\frac{n(n-1)}{2}$ pares de distâncias.



Estimador de Matheron

Desenvolvido por Matheron em 1962, a partir do método dos momentos.

$$\hat{\gamma}(h) = \frac{1}{2\,N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2 \tag{4}$$

Sendo:

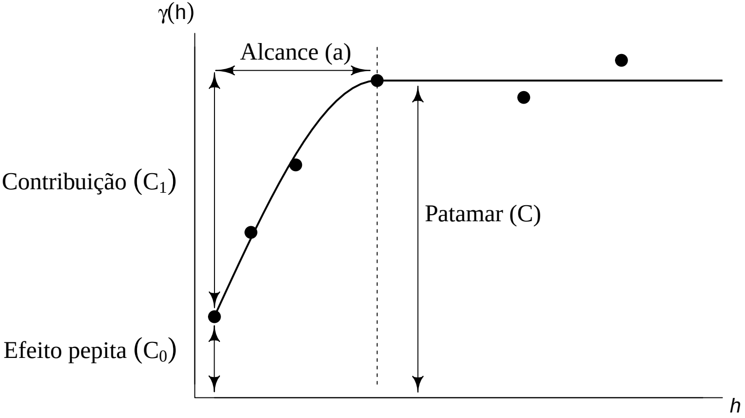
- $Z(x_i)$ a realização da variável aleatória Z no ponto x_i;
- $Z(x_i + h)$ a realização da variável aleatória Z no ponto x_i mais uma distância h;
- h a distância entre as observações;
- $N(h)$ o número de pares de valores medidos, separados por uma distância h;

Semivariância

Semivariograma empírico

O semivariograma empírico é o gráfico da semivariância em função de uma classe de distâncias definidas. A componente estrutural ε é encontrada no ajuste do semivariograma.

Semivariograma



Estimador de Semivariância para *Big Data*

Estimador de Semivariância para Big Data

Algoritmo

Passo 1: Selecionar aleatoriamente e sem reposição uma subamostra de tamanho b;

Passo 2: A partir da subamostra de tamanho b, procede-se com o cálculo da semivariância, utilizando o estimador de Matheron;

Passo 3: Repete-se os passos anteriores k vezes, gerando um vetor de semivariâncias para cada nova subamostra;

Passo 4: Por fim a estimativa da semivariância é obtida pela média aritmética das k semivariâncias em cada distância;

Algoritmo

Portanto definimos a semivariância como:

$$\hat{\gamma}(h)_{bk} = k^{-1} \sum_{i=1}^{k} \hat{\gamma}(h)_{bi} \tag{5}$$

em que:

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

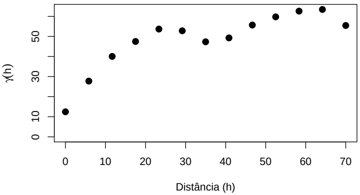
Estudo de simulação

Estudo de simulação

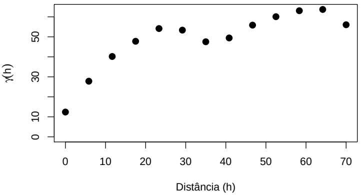
Realizou-se simulações com 15000 observações em diferente configurações. Para o cálculo do estimador proposto padronizou-se $b = 200$ e $k = 400$, isto é 400 subamostras de tamanho 200.

Tabela: Configurações da simulação

Modelo	*Patamar*	*Alcance*	*Efeito Pepita*
Exponencial	60	10	3
Esférico	30	19	5
Gaussiano	80	10	10

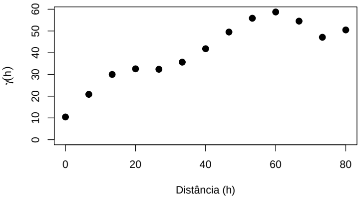


(a) Estimador de Matheron

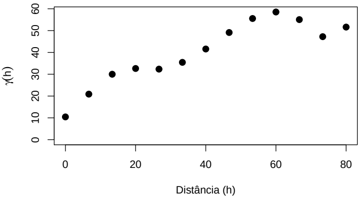


(b) Estimador para *Big Data*

Modelo Esférico

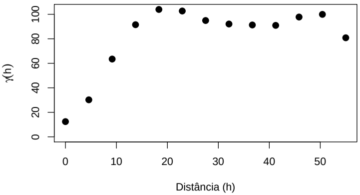


(a) Estimador de Matheron

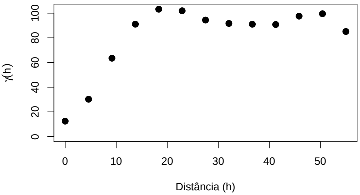


(b) Estimador para *Big Data*

Modelo Gaussiano



(a) Estimador de Matheron



(b) Estimador para *Big Data*

Estudo de simulação

Erro quadrático médio entre estimadores

Além da verificação do tempo computacional, calculamos o erro quadrático médio entre os estimadores, definido por:

$$EQM = \frac{1}{Cl(h)} \sum_{i=1}^{Cl(h)} \left(\hat{\gamma}(h) - \hat{\gamma}_{bk}(h)\right)^2 \tag{6}$$

Sendo:

- $Cl(h)$ classe de distância h;
- $\hat{\gamma}(h)$ o estimador de Matheron;
- $\hat{\gamma_{bk}}(h)$ o estimador para *Big Data*

Síntese

Modelo	EQM	Tempo Est. Matheron	Tempo *Big Data*
Exponencial	0.11	7.91	3.84
Esférico	0.15	7.58	3.48
Gaussiano	1.58	7.58	3.39

Aplicação em dados reais

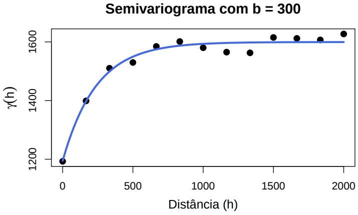
Aplicação em dados reais

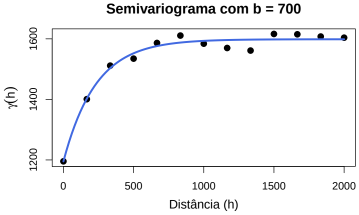
Goldmine Samples

O banco de dados "*goldmine samples*", contém a localização e o grau de ouro em grama, de 21577 observações tomadas a partir de uma mina de ouro.

Aplicação em dados reais

Definiu-se $k = 300$ e variou-se a quantidade de subamostras (b).





Aplicação em dados reais

Em resumo os parâmetros estimados e o tempo gasto em segundos, para cada semivariograma:

b	300	700
Modelo	Exponencial	Exponencial
Patamar	405.23	403.29
Alcance	715.64	692.43
Efeito Pepita	1194.31	1195.34
Tempo (seg.)	3.92	12.16

Além disso o erro quadrático entre as estimativas foi de 55,34. Portanto há diferenças significativas quando tamanho da subamostra aumenta.

Conclusões

- Deve-se padronizar uma classe de distâncias igualmente espaçadas;
- Estimador fornece resultados viáveis;
- Tempo computacional aumenta a medida que b ou k aumentam;
- Estudo de simulação mais detalhado.