Universidade Estadual de Maringá
Programa de Iniciação Científica – PIC
Departamento de Estatística
Orientador: Prof. Dr. Diogo Francisco Rossoni
Acadêmico: André Felipe Berdusco Menezes

Estimador de Semivariância para *Big Data*

Maringá 31 de Julho de 2016

Universidade Estadual de Maringá
Programa de Iniciação Científica – PIC
Departamento de Estatística
Orientador: Prof. Dr. Diogo Francisco Rossoni
Acadêmico: André Felipe Berdusco Menezes

Estimador de Semivariância para *Big Data*

Relatório contendo os resultados finais do projeto de iniciação científica vinculado ao Programa PIC-UEM.

Maringá 31 de Julho de 2016

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 estimador 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 estimador proposto e o estimador clássico de semivariância, evidenciando melhor desempenho do estimador proposto quanto ao custo computacional.

Palavras-chave: Geoestatística, Semivariância, *Big Data*.

Conteúdo

1 Introdução

Abordagens clássicas da estatística pressupõem que as variáveis aleatórias sejam independentes entre si. Por outro lado, em metodologias da estatística espacial as variáveis aleatórias estão relacionadas, de alguma forma, com a posição espacial que ocupam. Em específico na geoestatísca, consideramos as observações sendo realizações de um processo estocástico, isto é, para cada localização x_i amostrada têm-se uma variável aleatória Z distinta.

Entre os propósitos da geoestatística, a compreensão da variabilidade espacial é um fator imprescindivel 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 (1963), 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.

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.

2 Objetivos

2.1 Objetivo Geral

Apresentar um método alternativo para o cálculo da semivariância, denominado de método de subamostras, a fim de minimizar o custo computacional na estimação do semivariograma teórico, ferramenta indispensável na geoestatística.

2.2 Objetivos Específicos

- Propor um método alternativo para o cálculo da semivariância, realizado a partir da reamostragem das observações;
- Comparar o tempo computacional do método clássico e o proposto, via estudo de simulação;
- Apresentar aplicações em conjuntos de dados grandes, com o intuito de verificar o tempo computacional e o ajuste.

3 Desenvolvimento

3.1 Aspectos Gerais da Geoestatística

A geoestatística é um ramo da estatística espacial que estuda fenômenos com alguma referência espacial, por exemplo índices pluviométricos, temperatura do ar, ph da água de um lago e densidade do solo. Tendo em vista, como propósitos, a elaboração de modelos probabilístico para identificar e compreender a estrutura de dependência espacial do fenômeno em estudo. A elaboração desses modelos ocorre a priori por medidas de correlação espacial, isto é, covariância ou semivariância. Embora a modelagem seja fundamental para descrição e entendimento do fenômeno, na atividade científica o interesse não se limita em obter apenas o modelo de dependência espacial, desejando-se também predizer valores em pontos não amostrados.

3.2 Função Aleatória

Os conceitos teóricos da geoestatística são fundamentados em processos estocásticos ou também conhecido como função aleatória ou ainda processo espacial. De modo geral, considera-se que os dados são provenientes de um processo estocástico, isto é, para cada ponto x_i amostrado tem-se uma variável aleatória Z distinta. Formalmente temos:

$$\{Z(x_i) : x \in D \subset \mathbb{R}^p\} \tag{1}$$

Sendo:

- Z a variável aleátoria que varia continuamente em D;
- x a posição da variável, considerada fixa;
- D a região em estudo;
- $\mathbb{R}^p$ o espaço p-dimensional (p = 1,2,3 ou 4)

Devido a limitações de ordem prática, na maioria das análises existe somente uma realização do processo estocástico, o que torna impossível realizar inferência sobre este processo. Assim, é necessário que algum tipo de estacionariedade, adequada com o problema em estudo, seja imposta de maneira a possibilitar estimação de ao menos os dois primeiros momentos da função aleatória, isto é média, covariância e/ou semivariância.

Estacionariedade Intrínseca

A função aleatória $Z(x)$ é intrinsecamente estacionária se:

i. A esperança matemática existe e não depende do ponto x, isto é:

$$\mathbb{E}\left[Z(x)\right] = \mu \quad \forall x$$

ii. Para qualquer vetor distância h a variância da diferença $[Z(x+h) - Z(x)]$ existe e não depende de x, isto é:

$$Var[Z(x+h) - Z(x)] = \mathbb{E}\left[\left(Z(x+h) - Z(x)\right)^2\right] = 2\gamma \quad \forall x$$

Estacionariedade de Segunda Ordem

A função aleatória $Z(x)$ é estacionária de segunda ordem se:

i. A esperança matemática existe e não depende do ponto x, isto é:

$$\mathbb{E}\left[Z(x)\right] = \mu \quad \forall x$$

ii. Para cada par de variáveis aleatórias $\{Z(x), Z(x+h)\}$ a covariância existe e depende apenas do vetor distância h, isto é:

$$Cov(h) = Cov[Z(x), Z(x+h)] = \mathbb{E}\left[(Z(x)\, Z(x+h)\right] - \mu^2$$

Quando o processo espacial assume estacionariedade de segunda ordem temos:

$$\gamma(h) = C(0) - C(h) \tag{2}$$

Portanto, a covariância e a semivariância são ferramentas equivalentes para caracterizar a correlação espacial entre duas variáveis aleatórias $Z(x+h)$ e $Z(x)$. A estacionariedade de 2° ordem exige a existência de uma variância finita da função aleatória. Todavia, alguns fenômenos apresentam infinita dispersão, por isso, a hipótese intrínseca é menos restritiva, consequentemente a semivariânica é a ferramenta mais difundida na geoestatística.

3.3 Semivariância

A ideia da semivariância ou função variograma baseia-se no pressuposto que a relação espacial entre dois pontos amostrais não dependem de suas próprias localizações espaciais, mas apenas da localização relativa (distância). Assim sendo, o interesse está em encontrar uma medida de dependência espacial para n distintos pontos amostrais x_i pertencentes a um domínio espacial D, em que os valores observados $z(x_i)$ são considerados realizações de variáveis aleatórias $Z(x_i)$ com função aleatória $Z(x)$, $x \in D$. Essa medida é denominada de semivariância (WACKERNAGEL, 2013).

Matheron (1963) e Cressie (1993) definiram a semivariância teórica $\gamma(h)$ sendo:

$$\gamma(h) = \frac{1}{2} Var[Z(x+h) - Z(x)] = \frac{1}{2} \mathbb{E}\left[(Z(x+h) - Z(x))^2\right] \tag{3}$$

Podemos observar pela sua expressão 3 que a semivariância é uma medida de dissimilaridade, ou seja, seu valor é maior à medida que as variáveis estão menos associadas. Usualmente valores próximos apresentam variações pequenas, enquanto que valores distantes apresentam variações bruscas. Nota-se também que seu cálculo depende somente da distância entre as variáveis aleatórias.

Suponha que $Z(x)$ é uma função aleatória intrinsecamente estacionária. Conforme visto em Wackernagel (2013), a semivariância possui as seguintes propriedades:

(i) $\gamma(0) = 0$

(ii) $\gamma(h) \geq 0$

(iii) $\gamma(-h) = \gamma(h)$

Demonstrações

(i) Dado que a variância de uma constante é zero temos:

$$\gamma(0) = \frac{1}{2} Var[Z(x+0) - Z(x)] = \frac{1}{2} Var[Z(x) - Z(x)] = \frac{1}{2} Var[0] = 0 \quad (4)$$

(ii) A semivariância será sempre positiva, uma vez que não existe variância negativa.

(iii) A semivariância é invariante a translações, isto é:

$$\gamma(-h) = \frac{1}{2} Var[Z(x-h) - Z(x)] = \frac{1}{2} Var[Z(x) - Z(x+h)] = \gamma(h) \quad (5)$$

Como vimos a semivariância depende apenas das distâncias entre as variáveis aleatórias de um processo espacial. Todavia, na prática conhecemos algumas realizações do processo espacial, logo devemos estimar a semivariância com base nessas realizações. Existem diversos estimadores de semivariância, em seu artigo Félix et al. (2016) fazem uma revisão detalhada de oito estimadores de semivariância. Neste trabalho o enfoque foi sobre o estimador clássico de semivariância proposto por Matheron (1963), uma vez que é o estimador mais comum na literatura. Contudo, ressalta-se que o estudo realizado pode ser aplicado em qualquer estimador de semivariância.

3.4 Estimador Clássico de Matheron

Desenvolvido por Matheron (1963) a partir do método dos momentos, o estimador clássico de semivariância é definido pela seguinte expressão:

$$\hat{\gamma}(h) = \frac{1}{2\,N(h)} \sum_{i=1}^{N(h)} [z(x_i) - z(x_i + h)]^2 \quad (6)$$

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;

Cressie (1993) apontou que este estimador é não viesado, isto é, a média da distribuição amostral do estimador é igual ao verdadeiro valor do parâmetro. No entanto, observou-se que o mesmo é sensível a presença de outliers, dado que o seu cálculo é a diferença entre as observações ao quadrado.

Tendo em vista um estimador de semivariância é possível identificar e quantificar a variabilidade espacial do fenômeno em estudo. Na literatura, tal variabilidade pode ser representada graficamente por duas formas distintas, a *nuvem variográfica* ou o *semivariograma experimental*.

3.4.1 Nuvem Variográfica (*Variogram Cloud*)

Raramente utiliza-se este procedimento, pois o *variogram cloud* considera somente as distâncias entre as observações da amostra, todavia é baseado nele que podemos observar a presença de outliers. Sua expressão matemática é dada por:

$$\hat{\gamma}(h) = \frac{(z(x_i) - z(x_i + h))^2}{2} \tag{7}$$

em que:

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

A representação gráfica do *variogram cloud* resulta em um gráfico de dispersão da semivariância pela distância, como exemplo podemos observar a Figura 1.

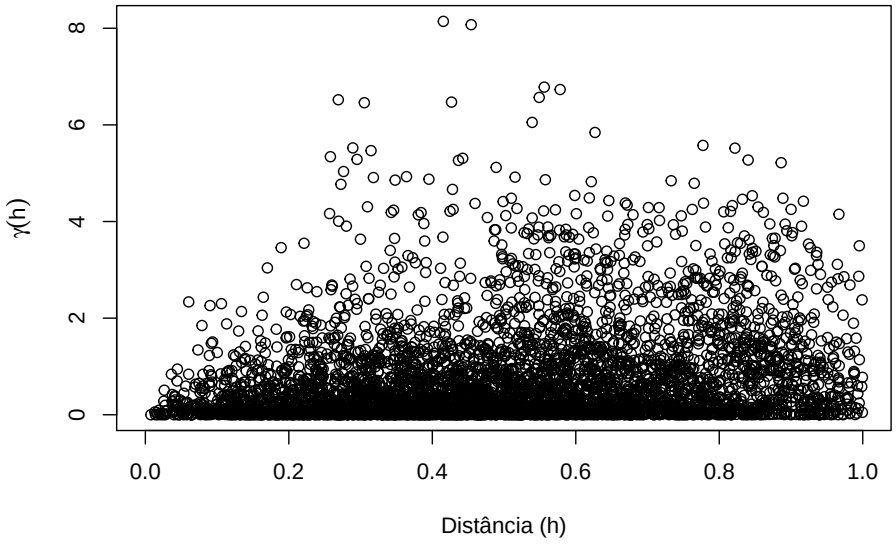


Figura 1: Exemplo *variogram cloud*

3.4.2 Semivariograma Experimental

Diferentemente do *variogram cloud*, um semivariograma experimental fornece as estimativas da semivariância para todas as distâncias entre as variáveis aleatórias de um processo espacial. Assim, dizemos que seu cálculo é baseado em uma classe de distâncias. A Figura 2 apresenta um semivariograma com característica desejáveis.

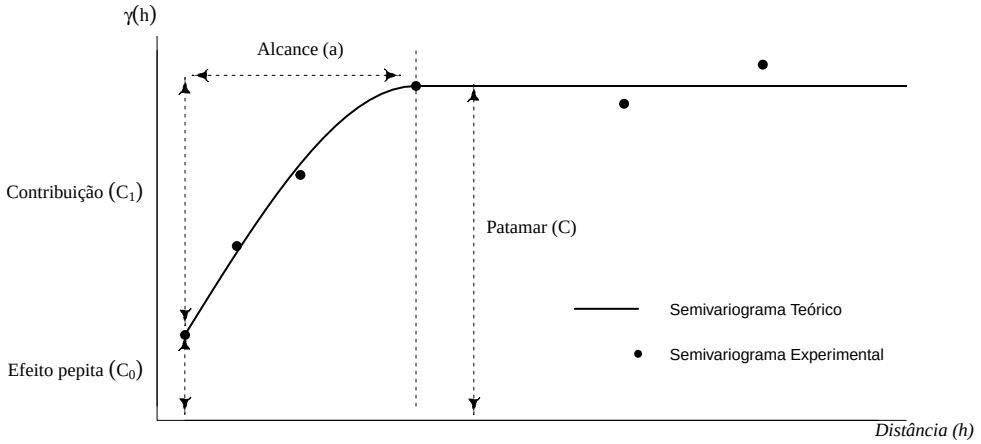


Figura 2: Exemplo Semivariograma

Isaacks e Srivastava (1989) definem os parâmetros de um semivariograma sendo:

- **Alcance (a):** a distância máxima na qual as realizações das variáveis aleatórias possuem uma dependência espacial;
- **Efeito pepita (C_0):** uma descontinuidade do semivariograma para distâncias menores que a menor distância observada. Conforme propriedade (i) (ver Equação 4) a semivariância para $h = 0$ é nula. Entretanto quando próximo a origem o semivariograma sofre uma descontinuidade. Isso ocorre em virtude de erros de medição ou variabilidade de pequena escala não captada pela amostragem.
- **Patamar (C):** o valor da semivariância correspondente ao alcance (a), isto é, $\gamma(a) = C$. Dizemos que $C \approx \text{Var}[Z(x)]$;
- **Contribuição (C_1):** a diferença entre o patamar (C) e o efeito pepita (C_0).

Para ajustar um modelo teórico (não linear) ao semivariograma empírico existem uma série de métodos, como por exemplo mínimos quadrados ordinários, ponderados e generalizados ou ainda máxima verossimilhança, tais métodos são abordados em Cressie (1985) e McBratney e Webster (1986).

3.4.3 Modelos Teóricos de Semivariograma

Segundo Rossoni (2014), a quantificação da variabilidade espacial entre as realizações das variáveis aleatórias está ligada ao ajuste de um modelo teórico ao

semivariograma experimental. Aquele modelo que melhor se ajusta aos pontos representa a magnitude, o alcance e a intensidade da variabilidade espacial do fenômeno.

Os principais modelos encontrados na literatura contemplam semivariogramas com e sem patamar. Os modelos sem patamar são próprios para estimação de parâmetros em fenômenos com dispersão infinita, isto é, aqueles que não atendem a hipótese intrínseca. Usualmente neste tipo de fenômeno é designado o modelo potência, definido a seguir:

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ (C_0 + C_1)\,|h|^b, & h \neq 0 \end{cases} \tag{8}$$

em que b é o parâmetro de intensidade da dispersão, $0 < b < 2$.

A Figura 3 apresenta o modelo potência com diferentes valores para seu parâmetro de intensidade (b). Observa-se que quando $b = 1$, temos uma reta crescente. Por outro lado, se $b > 0$ o modelo é dado por uma curva crescente, isto é, a semivariância aumenta indefinidamente conforme a distância aumenta. Além disso, se $b < 0$, então a curva tende assintóticamente a uma reta constante.

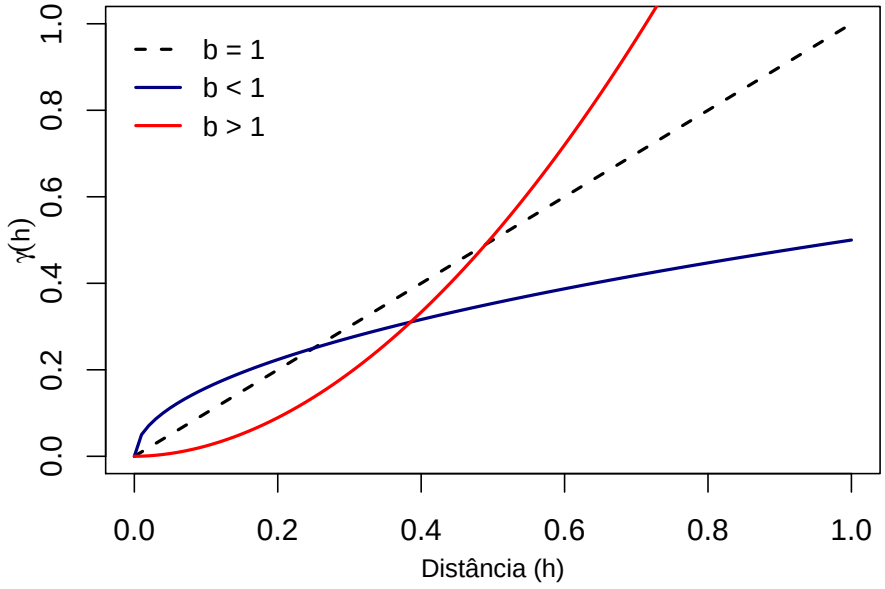


Figura 3: Modelos teóricos de semivariograma sem patamar

Já os modelos com patamar mais utilizados são:

- **Modelo Esférico**: possui um comportamento linear na origem, diferenciando-se dos demais pois o patamar alcançado é real. Destacamos na Figura 4 a maneira não assintótica que o modelo esférico atinge o patamar,

$$\gamma(h) = \begin{cases} 0 & ,h = 0 \\ C_0 + C_1 \left[1.5 \left(\frac{h}{a}\right) - 0.5 \left(\frac{h}{a}\right)^3\right] & ,0 < h < a \\ C_0 + C_1 & ,h > a. \end{cases} \tag{9}$$

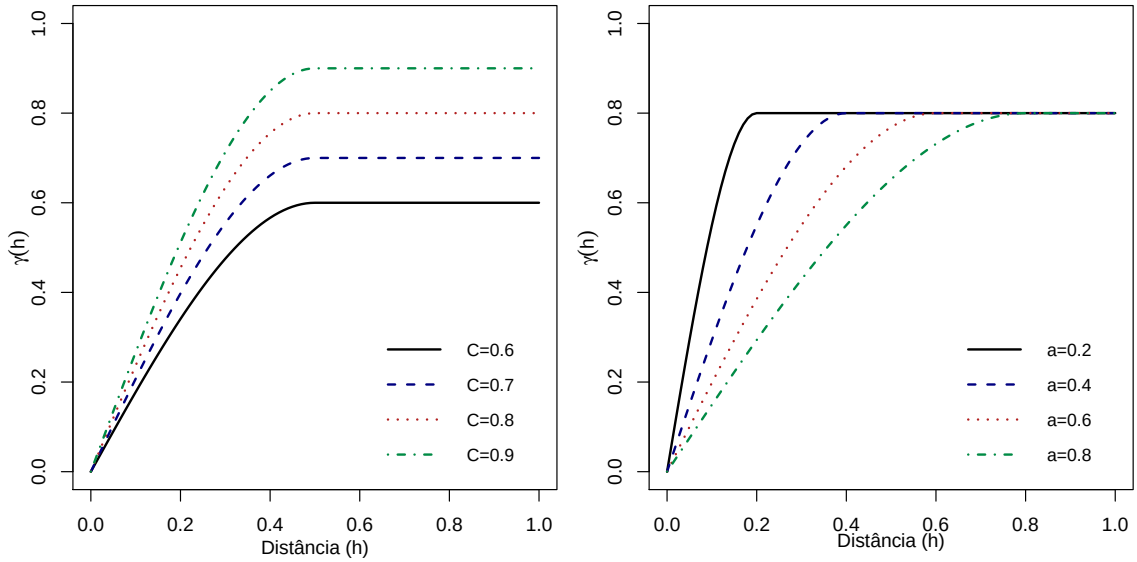


Figura 4: Comportamento do modelo Esférico para diferentes valores de C e a

- **Modelo Exponencial**: percebe-se na Figura 5, que também apresenta comportamento linear na origem, no entanto o patamar é alcançado assintoticamente.

$$\gamma(h) = \begin{cases} 0 & ,h = 0 \\ C_0 + C_1 \left[1 - e^{-\frac{h}{a}}\right] & ,h \neq 0 \end{cases} \tag{10}$$

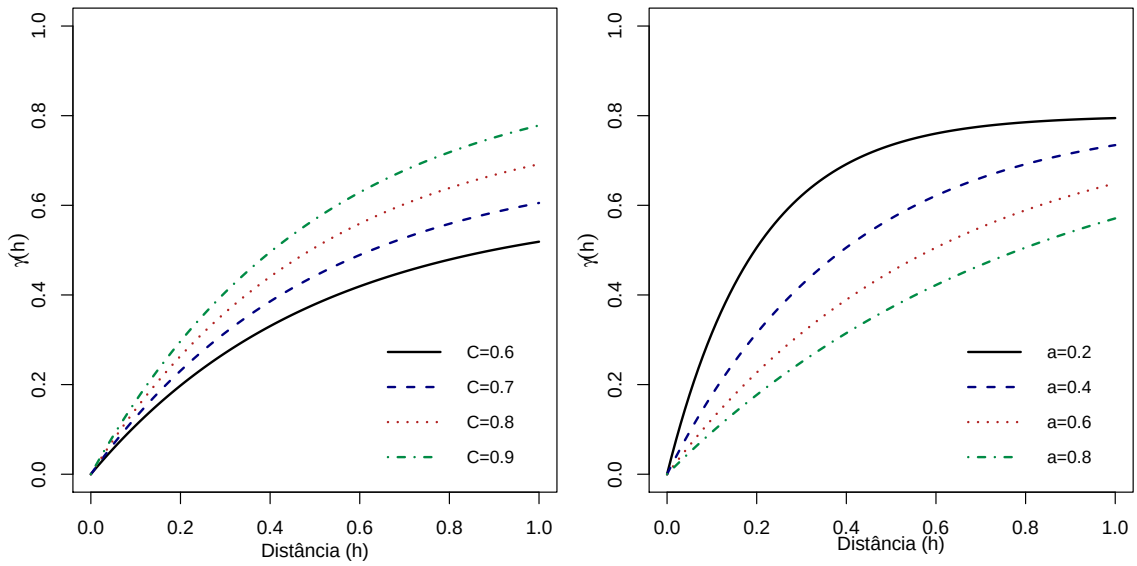


Figura 5: Comportamento do modelo Exponencial para diferentes valores de C e a

- **Modelo Gaussiano**: muito semelhante ao modelo exponencial, também atinge o patamar assintoticamente. Sua peculiaridade está na forma parabólica próxima a origem e por possuir ponto de inflexão, conforme Figura 6.

$$\gamma(h) = \begin{cases} 0 & , h = 0 \\ C_0 + C_1 \left[1 - e^{-\left(\frac{h}{a}\right)^2}\right] & , h \neq 0 \end{cases} \tag{11}$$

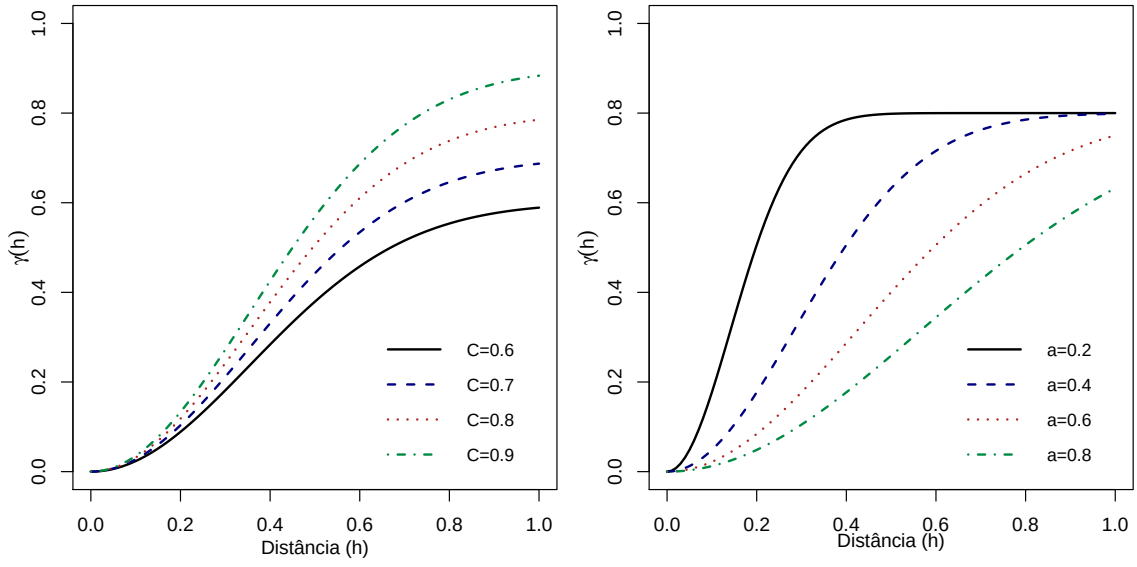


Figura 6: Comportamento do modelo Gaussiano para diferentes valores de C e a

3.5 Materiais e Métodos

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.

As etapas do algoritmo de estimação são:

(i) Dada uma amostra inicial de tamanho n, seleciona-se 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 conforme Equação 6;

(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:

$$\hat{\gamma}(h) = k^{-1} \sum_{i=1}^{k} \gamma(h)_{bi} \tag{12}$$

em que:

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

A implementação computacional, foi realizada no ambiente estatística R, com alguns auxilio da biblioteca `geoR`. Ressalta-se que a escolha do estimador de semivariância, neste caso o estimador de Matheron, não altera todo o procedimento, somente a etapa (ii).

Com intuito de averiguar 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: $\sigma^2 = 60$ (*patamar*), $\phi = 30$ (*alcance*) e $\tau = 0$ (*efeito pepita*).

O estimador clássico (Equação 6), foi calculado através da função `variog` da biblioteca `geoR`. Já na execução do algoritmo consideramos 100 subamostras com tamanho 100, isto é, $k = b = 100$.

Além disso, foi realizado um estudo de aplicação em que estudamos o conjunto de dados apresentado em Clark (1979). Os dados consistem de 22577 observações a respeito do grau de ouro em gramas e a respectiva coordenada, as medições foram extraídas de uma mina de ouro localizada na Africa do Sul.

3.6 Resultados e Discussão

Nesta seção serão apresentados resultados de simulação e análise em um conjunto de dados reais, para que possamos comparar a performance do estimador clássico de semivariância e o algoritmo proposto.

3.6.1 Estudo de Simulação

Na Figura 7, 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.

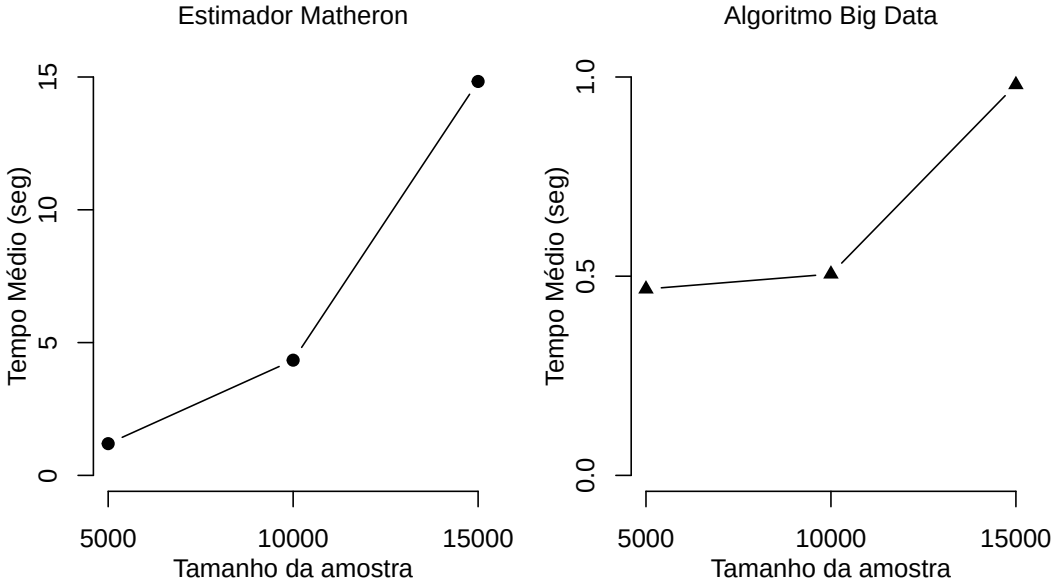


Figura 7: Tempo Médio de Estimação

3.6.2 Aplicação em Dados Reais

Realizou-se uma análise exploratória para verificar o comportamento da variável em estudo. Na Figura 8 é possível observar a distribuição espacial e o histograma do grau de ouro em gramas. No que se refere a distribuição da variável verifica-se uma assimetria à direita. No mapa com a distribuição espacial dos dados, classificamos as observações conforme as medidas descritivas mínimo, quartis e máximo.

Na análise geoestatística tivemos o interesse de identificar e quantificar o modelo de dependência espacial do grau de ouro. Devemos então, encontrar o semivariograma empírico e ajustá-lo a um modelo. Devido ao grande número de observações utilizamos o algoritmo proposto, no qual definimos $k = 100,400$ e $b = 100,400$, isto é, dois semivariograma com diferentes iterações e tamanhos de subamostras.

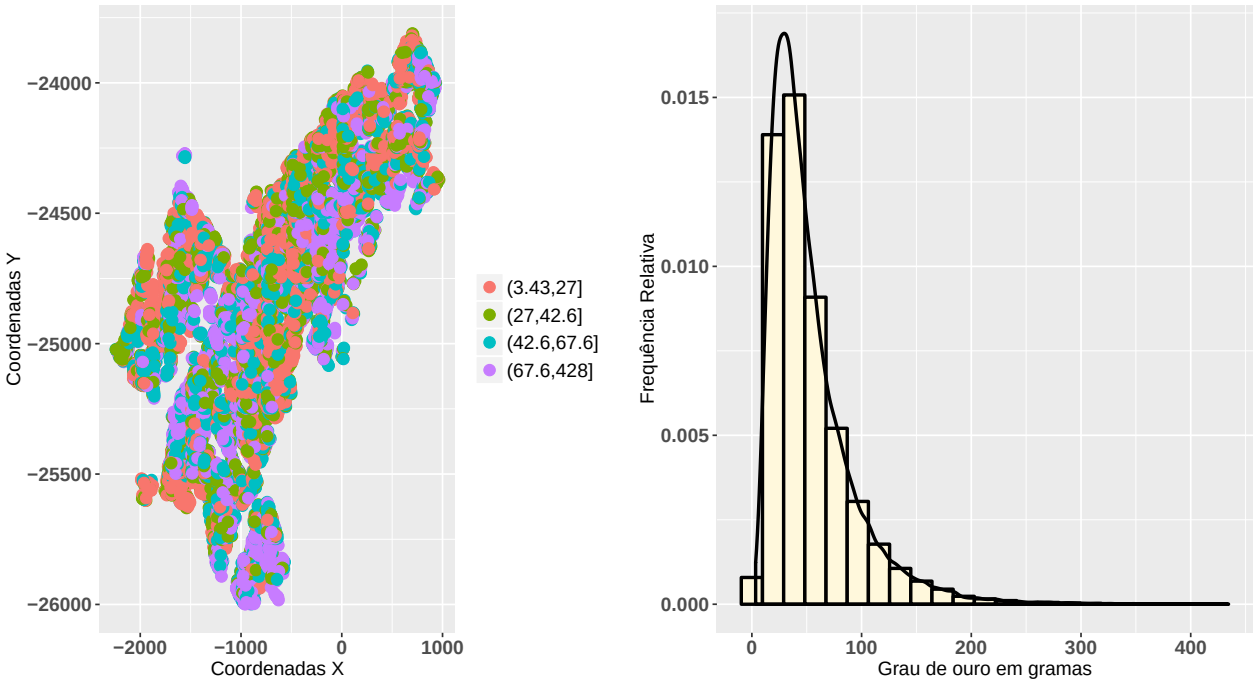


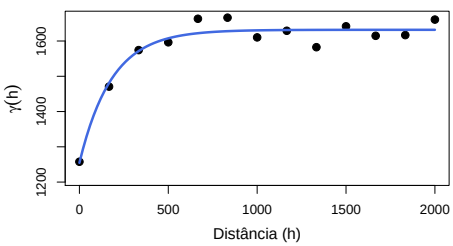
Figura 8: Comportamento do grau de ouro

A Tabela 1 fornece as estimativas dos parâmetros e o modelo ajustado quando o número de iterações e tamanho das subamostras variaram. Embora o tempo computacional gasto quando $k = b = 100$ foi pequeno as estimativas dos parâmetros diferenciaram quando $k = b = 400$, em destaque o alcance.

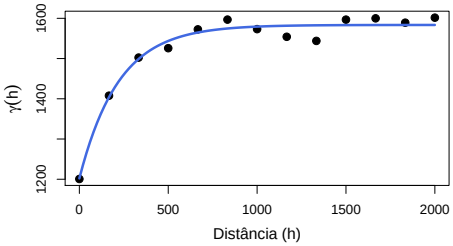
Tabela 1: Resumo dos modelos ajustados para diferentes valores de k e b

Modelo	Patamar	Alcance	Pepita	k	b	Tempo (seg)
Exponencial	377,13	544,58	1254,61	100	100	0,56
Exponencial	387,79	746,41	1207,87	400	400	7,74

Os semivariogramas ajustados ao modelo exponencial podem ser visto na Figura 9, como dito anteriormente ocorreu uma ampla diferença quando k e b aumentaram. Constatamos que o aumento de k e b implicam em boas estimativas do semivariograma. Além disso, nesse caso ocorreu um aumento no tempo computacional, mas sem demandar muito esforço.



(a) Semivariograma quando $k = b = 100$



(b) Semivariograma quando $k = b = 400$

Figura 9: Semivariogramas ajustados para diferentes iterações e tamanhos de subamostra

No que se refere a interpretação da estrutura espacial da variável estudada, verificamos que a distância na qual o semivariograma atinge seu patamar é 746 km, isto é, em distância inferiores a essa as amostradas apresentam-se correlacionadas espacialmente, em contrapartida dessa distância em diante dizemos que não existe mais dependência espacial.

4 Conclusão

De acordo com os resultados obtidos pelo estudo de simulação verificou-se que, o algoritmo proposto para a estimação da semivariância em grandes conjuntos de dados, é considerados mais eficiente que o método clássico, uma vez que possui menor esforço computacional e apresenta resultados semelhantes na estimativa da semivariância.

Ao ser aplicado em dados reais com diferentes tamanhos de subamostras e iterações, o algoritmo proposto descreve bem a variabilidade do fenômeno espacial. Sendo assim, uma boa alternativa para estimar o semivariograma de grandes conjuntos de dados espaciais.

Além disso, verificamos pelo estudo de simulação que conforme o tamanho e o número de subamostras aumentam maior o tempo computacional e mais precisa são as estimativas.

Referências

CLARK, I. **Practical geostatistics**. [S.l.]: Applied Science Publishers London, 1979. v. 3.

CRESSIE, N. Fitting variogram models by weighted least squares. **Journal of the International Association for Mathematical Geology**, Springer, v. 17, n. 5, p. 563–586, 1985.

CRESSIE, N. Statistics for spatial data: Wiley series in probability and statistics. **Wiley-Interscience, New York**, v. 15, p. 105–209, 1993.

DIGGLE, P. J.; TAWN, J.; MOYEED, R. Model-based geostatistics. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 47, n. 3, p. 299–350, 1998.

FÉLIX, V. B. et al. Estimadores de semivariância: Uma revisão. **Ciência e Natura**, v. 38, n. 3, p. 1157–1167, 2016.

FURRER, R.; GENTON, M. G.; NYCHKA, D. Covariance tapering for interpolation of large spatial datasets. **Journal of Computational and Graphical Statistics**, Taylor & Francis, 2012.

GRINGARTEN, E.; DEUTSCH, C. V. Teacher's aide variogram interpretation and modeling. **Mathematical Geology**, Springer, v. 33, n. 4, p. 507–534, 2001.

ISAACS, E. H.; SRIVASTAVA., R. M. **Applied geostatistics**. [S.l.]: Oxford University Press, 1989.

JR, P. J. R.; DIGGLE, P. J. geor: a package for geostatistical analysis. **R news**, London, v. 1, n. 2, p. 14–18, 2001.

LIANG, F. et al. A resampling-based stochastic approximation method for analysis of large geostatistical data. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 108, n. 501, p. 325–339, 2013.

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

MCBRATNEY, A.; WEBSTER, R. Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates. **Journal of Soil Science**, Wiley Online Library, v. 37, n. 4, p. 617–639, 1986.

ROSSONI, D. F. Testes bootstrap para detecção de anisotropia espacial em fenômenos contínuos. UNIVERSIDADE FEDERAL DE LAVRAS, 2014.

TEAM, R. C. et al. R: A language and environment for statistical computing. Vienna, Austria, 2013.

WACKERNAGEL, H. **Multivariate geostatistics: an introduction with applications**. [S.l.]: Springer Science & Business Media, 2013.

YAMAMOTO, J. K.; LANDIM, P. M. B. **Geoestatística: conceitos e aplicações**. [S.l.]: Oficina de Textos, 2015.

Anexo - Implementação Estimador de Semivariância para *Big Data.*

```
var_bigdata <- function(dados, k, b, cl.dist, max.dist,
   est.type, modelar = FALSE, mod)
{

        require(geoR)

        ini <- proc.time()

        semivar <- lapply(1:k, function(...)
           variog(sample.geodata(dados, size = b,
           replace = F),uvec = cl.dist, estimator.type =
           est.type, max.dist = max.dist))

        dist <- rowMeans(sapply(1:k, function(j)
           semivar[[j]]$u))
        var  <- rowMeans(sapply(1:k, function(j)
           semivar[[j]]$v))
        npar <- rowMeans(sapply(1:k, function(j)
           semivar[[j]]$n))

        fim <- proc.time() - ini

        md = NULL

        if(modelar == TRUE){
                semivar[[k]]$v <- var
                md <- variofit(semivar[[k]], cov.model =
                   mod, weights = "equal")
        }

        lista <- list(vg = var, dist = dist, npar = npar,
        var.amostra = semivar[[k]], tempo = fim, modelo
           = md)

        return(lista)
}
```