

COMPARAÇÃO DE MÉTODOS PARA TESTE DE BIGAUSSIANIDADE

Jorge Kazuo YAMAMOTO ¹ & Li CHAO ²

(1) Departamento de Geologia Sedimentar e Ambiental, Instituto de Geociências, USP. Rua do Lago, 562 – Cidade Universitária. CEP 05508-080. São Paulo, SP. Endereço eletrônico: jkyamamo@usp.br

(2) Institute of Hydrology and Water Resources, Tsinghua University – Beijing 100084, China. Endereço eletrônico: l-c06@mails.tsinghua.edu.cn

Introdução
Teste de Variogramas de Variáveis Indicadoras
Teste de Variogramas de Ordem ω
Materiais e Métodos
Resultados e Discussão
Conclusões
Agradecimentos
Referências Bibliográficas

RESUMO – Em geoestatística, alguns métodos trabalham sob a hipótese de multiGaussianidade. Simulação Gaussiana seqüencial e krigagem multiGaussiana são exemplos desses métodos que requerem os dados originais transformados para escores da distribuição normal. Contudo, isso não garante que as funções de distribuições acumulativas para dois, três ou mais pontos sigam uma distribuição Gaussiana também. Todos os testes, disponíveis até o momento, usam a estatística de dois pontos que é melhor descrita através da função variograma. Se o teste de dois pontos for positivo, então o modelo multiGaussiano pode ser adotado. Nesse trabalho, dois métodos são testados com três amostras retiradas de um conjunto completo. Os resultados desses métodos são comparados entre si e concluindo um deles pode ser considerado estatisticamente robusto.

Palavras-chave: biGaussianidade, transformação Gaussiana, variogramas de indicadoras, variogramas de ordem ω .

ABSTRACT – *J.K. Yamamoto & Li Chao - Comparison of methods for bigaussianity tests.* In geostatistics some methods work under the strong hypothesis of multiGaussianity. Sequential Gaussian simulation and multiGaussian kriging are examples of these methods that require original data transformed into normal score. However, it does not guarantee that cumulative distributions functions for two, three or more points follow a Gaussian distribution as well. All available tests so far have used two-point statistics that is best described by a semivariogram function. If the test for two points is positive then the multiGaussian model can be adopted. In this paper two methods are tested with three samples drawn from an exhaustive data set. Results of these methods are compared each other and in conclusion one of them can be considered statistically robust.

Keywords: biGaussianity, normal score transform, variograms of indicator variables, variograms of order ω .

INTRODUÇÃO

Certos procedimentos em geoestatística trabalham sob a hipótese de multiGaussianidade dos dados. Estes métodos requerem a transformação Gaussiana dos dados, mas a condição que a função de distribuição acumulada é gaussiana não é suficiente. É necessário testar se a distribuição de dois pontos, três ou mais pontos é também Gaussiana. Entretanto, é muito difícil na prática realizar o teste para distribuições multi-pontos. Assim, se o teste de biGaussianidade for positivo supõe-se que as demais distribuições multi-pontos são também Gaussianas e, conseqüentemente, aceita-se a hipótese de multiGaussianidade dos dados. Os testes de biGaussianidade são feitos através da comparação de variogramas, os quais, por excelência, representam as estatísticas de dois pontos.

Na prática, existem dois métodos que permitem fazer o teste de biGaussianidade dos dados: Goovaerts (1997) e Emery (2005). Os dois métodos requerem a transformada Gaussiana dos dados. O primeiro, proposto por Goovaerts (1997), compara variogramas experimentais da variável indicadora para vários percentis da distribuição acumulada normal com variogramas teóricos deduzidos do modelo Gaussiano. Se o ajuste entre os variogramas experimentais e teóricos for bom para todos os percentis analisados, então se aceita a hipótese de biGaussianidade dos dados. O segundo método (Emery, 2005), compara a relação entre pontos do variograma de ordem ω com o variograma padrão ($\omega=2$) em um diagrama log-log, se os pontos se alinharem à reta de referência, então se

aceita a hipótese de biGaussianidade. Os dois testes de biGaussianidade foram implementados em um programa chamado Bigauss, dentro do Sistema Geovisual (Yamamoto, 2001).

Programas para teste de biGaussianidade dos

dados nem sempre estão disponíveis ao usuário de geoestatística. O objetivo deste artigo é mostrar os métodos existentes e também divulgar e colocar à disposição os programas que vêm sendo introduzidos no Sistema Geovisual.

TESTE DE VARIOGRAMAS DE VARIÁVEIS INDICADORAS

O teste de biGaussianidade proposto por Goovaerts (1997) é baseada na função de distribuição acumulada Gaussiana:

$$G(h; y_p) = \Pr \{Y(u) \leq y_p, Y(u+h) \leq y_p\} \\ = p^2 + \frac{1}{2\pi} \int_0^{\arcsin C_Y(h)} \exp\left[\frac{-y_p^2}{1 + \sin \theta}\right] d\theta \quad (1)$$

Segundo Goovaerts (1997), os dados são transformados para o domínio Gaussiano, em seguida, o

variograma $\gamma_Y(h)$ é calculado e modelado, do qual se deriva a covariância $C_Y(h) = 1 - \gamma_Y(h)$. Para um número de *scores* (y_{p_i}), correspondentes aos percentis p_i , calcula-se a distribuição Gaussiana de dois pontos, conforme a equação (1).

O variograma da variável indicadora $\gamma_I(h; y_p)$ para o *score* y_p da função de distribuição acumulada Gaussiana pode ser calculada como:

$$\gamma_I(h; y_p) = p - G(h; y_p) \quad (2)$$

TESTE DE VARIOGRAMAS DE ORDEM ω

O variograma de ordem ω é, segundo Emery (2005):

$$\gamma_\omega(h) = \frac{1}{2} E \left[|Y(x+h) - Y(x)|^\omega \right] \text{ com } 0 < \omega \leq 2$$

onde $Y(x)$ é a variável Gaussiana e ω é a potência.

Quando $\omega=2$, a expressão proporciona o cálculo do variograma tradicional da variável Gaussiana e quando $\omega=1$, tem-se o madograma.

Segundo Emery (2005), a relação entre o variograma de ordem ω e o variograma tradicional é dada por:

$$\gamma_\omega(h) = \frac{2^{\omega-1}}{\sqrt{\pi}} \Gamma\left(\frac{\omega+1}{2}\right) [\gamma(h)]^{\omega/2} \quad (3)$$

onde $\Gamma(\omega)$ é a função gama, que pode ser calculada numericamente (Press et al., 1996):

$$\Gamma(z) = \left(\frac{\sqrt{2\pi}}{z} \left(p_0 + \sum_{n=1}^6 \frac{p_n}{z+n} \right) \right) (z+5.5)^{-z+0.5} e^{-(z+5.5)} \quad (4)$$

onde $p_0 = 1.000000000190015$; $p_1 = 76.18009172947146$; $p_2 = -86.50532032941677$; $p_3 = 24.01409824083091$; $p_4 = -1.231739572450155$; $p_5 = 1.208650973866179 \times 10^{-3}$; $p_6 = -5.395239384953 \times 10^{-6}$.

A função gama aproximada pela equação (4), no intervalo de interesse ao teste de biGaussianidade, ou seja, $0 < \omega < 2$, tem a seguinte forma (Figura 1).

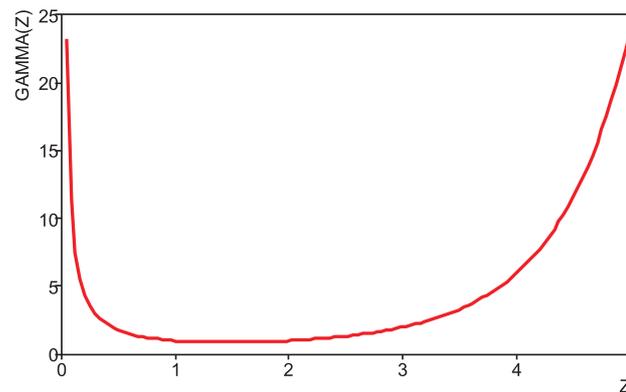


FIGURA 1. Gráfico da função gama no intervalo [0,5], conforme aproximação proposta por (Press et al., 1996).

A relação (3), de acordo com Emery (2005), pode ser usada para o teste de biGaussianidade dos dados, assumindo que a distribuição marginal dos dados é Gaussiana, ou seja, que os dados originais foram transformados para uma distribuição normal $N(0,1)$. Emery (2005) sugere dois testes, quais sejam:

- para uma dada ordem ω , calcular os variogramas de ordem ω e o variograma tradicional, os quais são divididos pelos respectivos variogramas para uma distância de referência h_0 :

$$\frac{\gamma_\omega(h)}{\gamma_\omega(h_0)} = \left[\frac{\gamma(h)}{\gamma(h_0)} \right]^{\omega/2} \quad (5)$$

As duas razões são então lançadas em um gráfico log-log; se a distribuição for biGaussiana, então os pontos deverão se alinhar conforme uma reta com inclinação $\omega/2$. Observe-se que a relação (5) resulta da equação (3) calculada para a distância h_0 (equação 6) e dividida lado a lado e simplificada.

$$\gamma_\omega(h_0) = \frac{2^{\omega-1}}{\sqrt{\pi}} \Gamma\left(\frac{\omega+1}{2}\right) [\gamma(h_0)]^{\omega/2} \quad (6)$$

Segundo Emery (2005), a distância de referência h_0 deve ser tal que ambos os variogramas sejam confiáveis, isto é dentro do campo geométrico e com suficiente número de pares. Conforme o mesmo autor, h_0 corresponde geralmente à amplitude do variograma.

- o segundo teste é feito para um valor h_0 fixo e ω variável. A expressão (3) é calculada para uma distância constante h_0 e rearranjada para que resulte na seguinte relação:

$$\frac{\sqrt{2\pi}\gamma_\omega(h_0)}{2^{\omega-1}[\gamma(h_0)]^{\omega/2}\Gamma\left(\frac{\omega+1}{2}\right)} = 1 \quad (7)$$

A razão expressa pela equação (7) deve resultar em valor próximo de 1, se a distribuição for biGaussiana. Segundo Emery (2005), este teste deve ser feito somente para pequenas distâncias, ou seja, dentro do campo estruturado do variograma.

MATERIAIS E MÉTODOS

Três amostras com 64 pontos foram extraídas de um conjunto completo usando a amostragem aleatória estratificada. O conjunto completo é conhecido sobre uma malha regular de 50 x 50 nós, totalizando 2500 pontos de dados. Portanto, as amostras representam 2,56% da população. As estatísticas descritivas das 3 amostras e os parâmetros populacionais encontram-se na Tabela 1. A Figura 2 mostra as distribuições de frequências acumuladas e os diagramas P-P comparando as amostras com o conjunto completo.

As amostras representam razoavelmente o conjunto completo. As distâncias médias medidas nos diagramas P-P foram iguais a 3,18, 1,96 e 2,07,

respectivamente para as amostras teste1.txt, teste2.txt e teste3.txt. Estas medidas apenas confirmam a inspeção visual dos diagramas P-P em que teste2.txt e teste3.txt são melhores aproximações que o teste1.txt.

Para as amostras assim constituídas foram calculados e modelados os variogramas experimentais da variável original após a transformação Gaussiana. Os variogramas obtidos encontram-se na Figura 3.

As amostras descritas nesta seção representam os materiais deste estudo. Assim, o objetivo é testar os dois testes de biGaussianidade (Goovaerts, 1997 e Emery, 2005) e comparar os resultados, em termos de eficácia e robustez.

TABELA 1. Estatísticas descritivas das amostras para estudo.

Estatística	Completo.txt	Teste1.txt	Teste2.txt	Teste3.txt
No. de dados	2500	64	64	64
Média	1,815	1,752	1,765	1,379
Desvio padrão	2,617	2,605	2,441	1,206
Coef. variação	1,441	1,487	1,383	0,875
Máximo	39,787	12,479	13,101	5,711
Quartil superior	2,096	1,935	1,762	1,693
Mediana	0,999	0,755	1,051	0,986
Quartil inferior	0,476	0,488	0,418	0,496
Mínimo	0,025	0,056	0,091	0,093

RESULTADOS E DISCUSSÃO

Os resultados dos dois testes são gráficos; no caso de Goovaerts (1997), variogramas experimentais e teóricos são comparados e caso o ajuste entre eles

seja bom, o teste é considerado satisfatório para a hipótese de biGaussianidade dos dados; no caso de Emery (2005), os resultados são expressos em dois

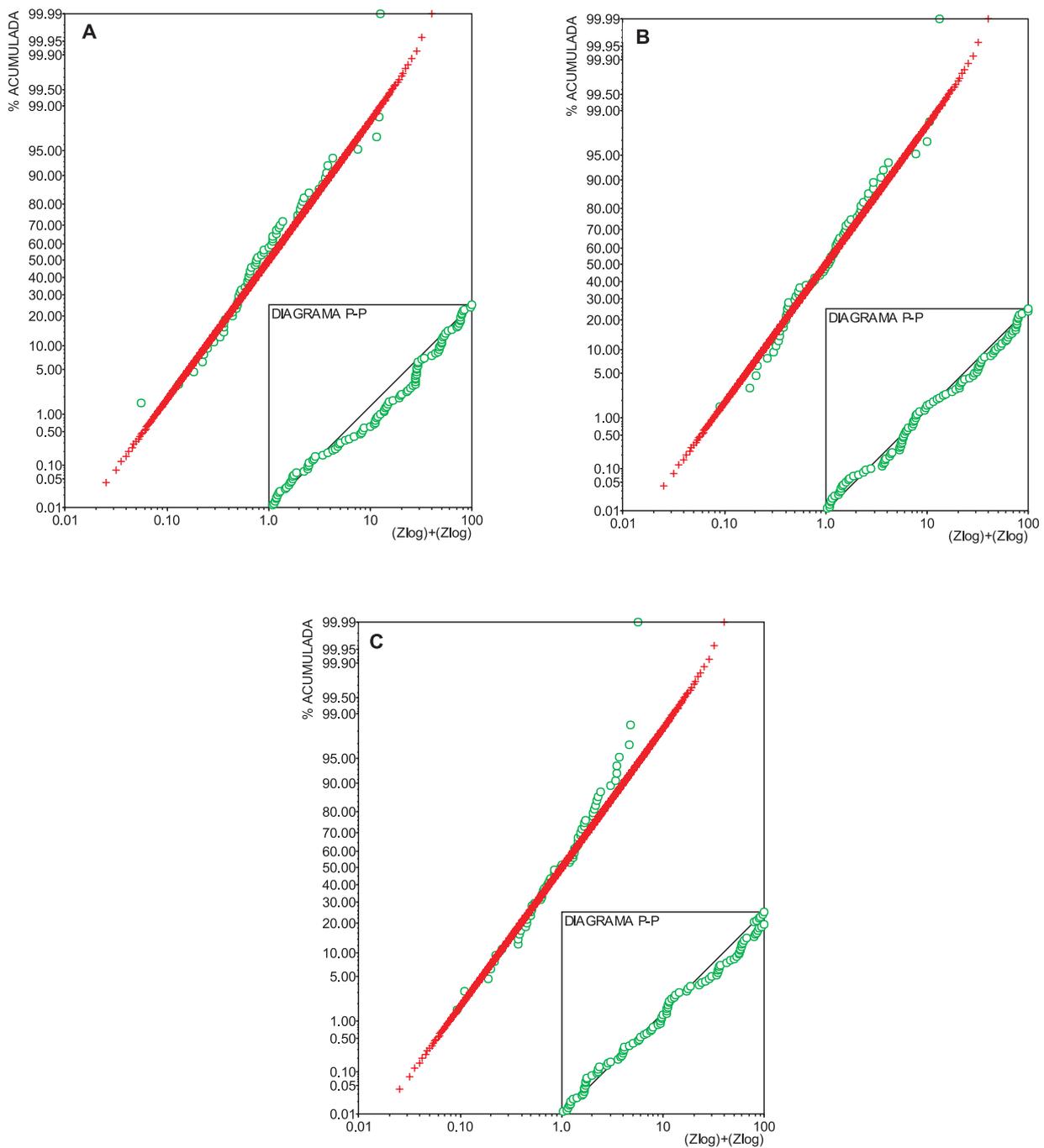


FIGURA 2. Distribuições de freqüências acumuladas e diagramas P-P, para as amostras: teste1.txt (A); teste2.txt (B) e teste3.txt (C).

diagramas, sendo que no primeiro os pontos devem se alinhar às retas padrão e no segundo os pontos devem mostrar valores próximos a um. As Figuras 4, 5 e 6 apresentam os resultados do teste de biGaussianidade, conforme a metodologia de Goovaerts (1997). O programa apresenta a opção de

cálculo dos níveis de corte das indicadores para 3 quartis ou então 9 decis. Os resultados são todos para 9 decis.

Os resultados do teste de Emery (2005) encontram-se nas Figuras 7, 8 e 9, respectivamente para as amostras teste1.txt, teste2.txt e teste3.txt.

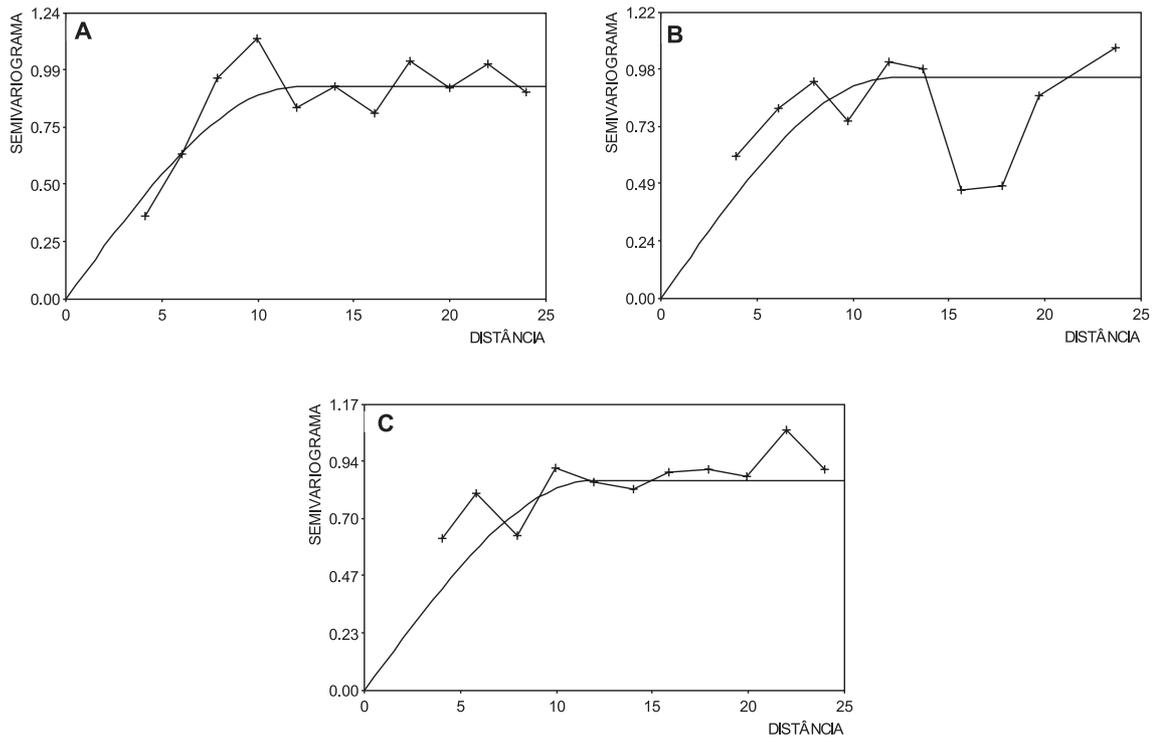


FIGURA 3. Variogramas experimentais e modelos ajustados para as amostras: teste1.txt (A); teste2.txt (B) e teste3.txt (C).

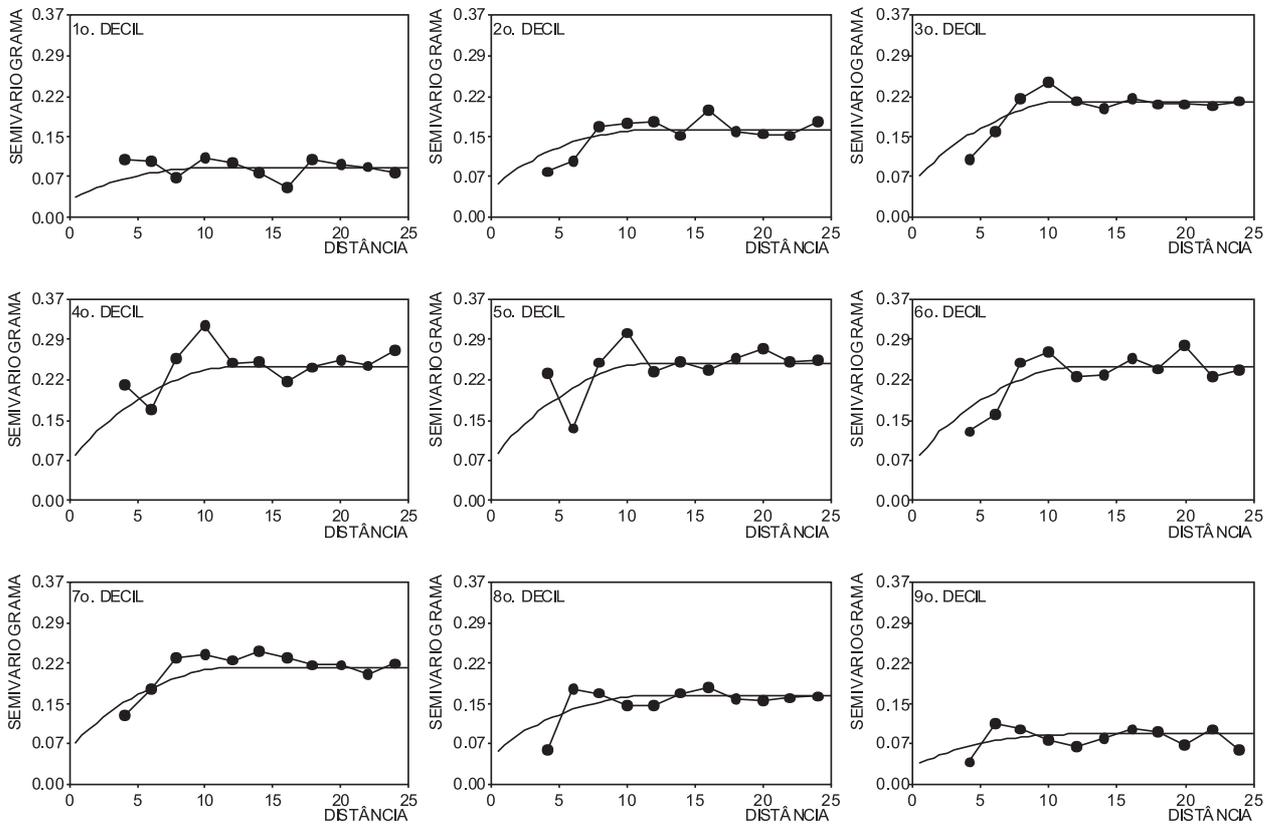


FIGURA 4. Variogramas experimentais (círculos cheios) e teóricos deduzidos do modelo Gaussiano teórico para a amostra teste1.txt.

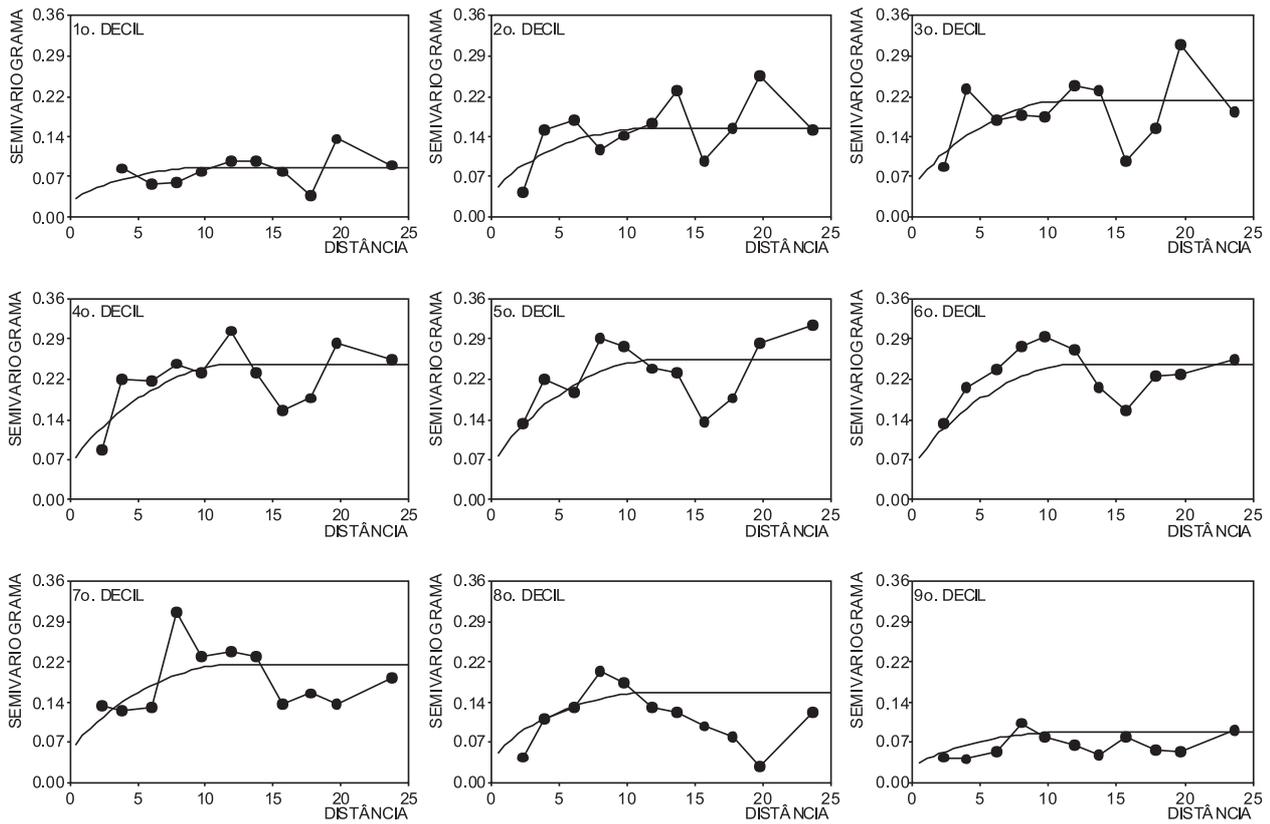


FIGURA 5. Variogramas experimentais (círculos cheios) e teóricos deduzidos do modelo Gaussiano teórico para a amostra teste2.txt.

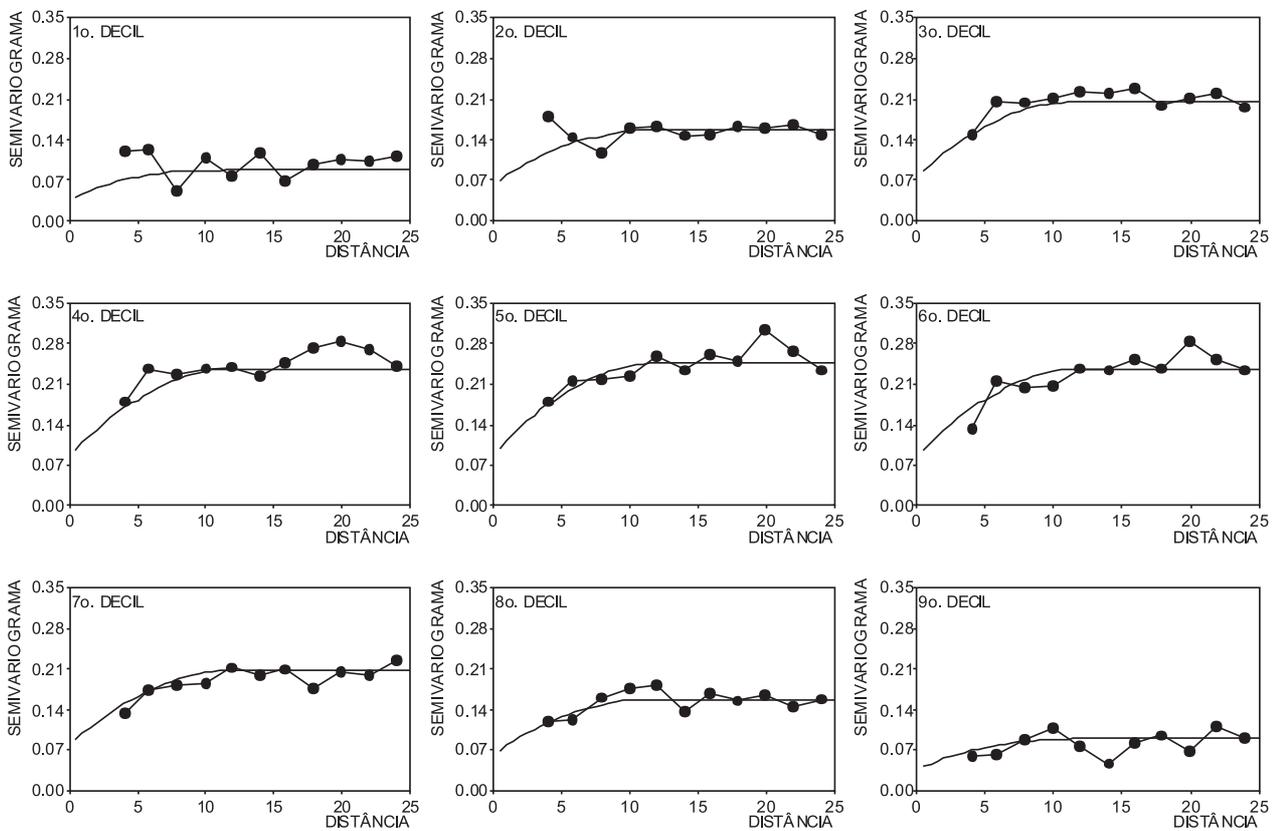


FIGURA 6. Variogramas experimentais (círculos cheios) e teóricos deduzidos do modelo Gaussiano teórico para a amostra teste3.txt.

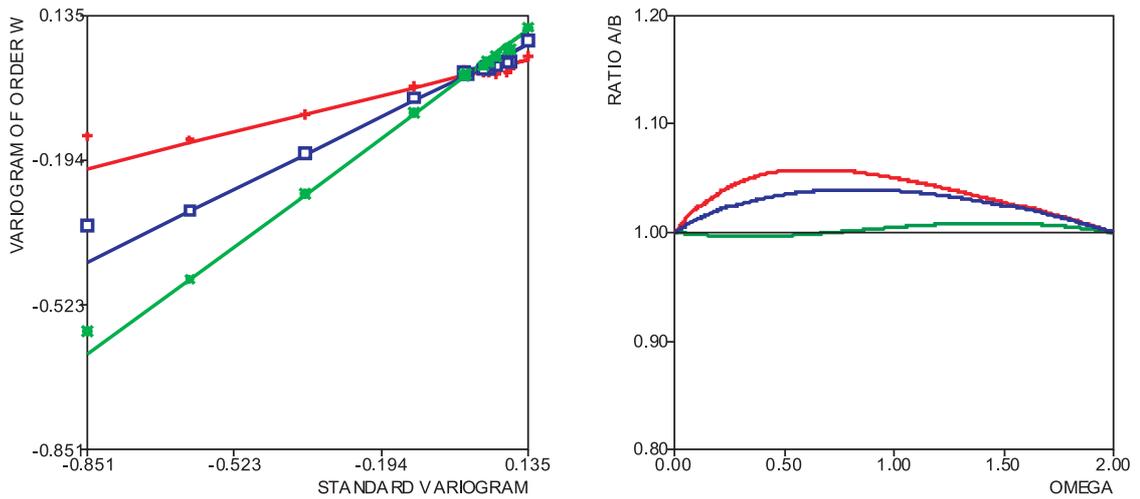


FIGURA 7. Gráfico bi-log (valores em logaritmos) e razões calculadas conforme a equação (7) para a amostra teste1.txt (no gráfico bi-log a cor vermelha representa $\omega=0,5$; azul $\omega=1,0$ e verde $\omega=1,5$).

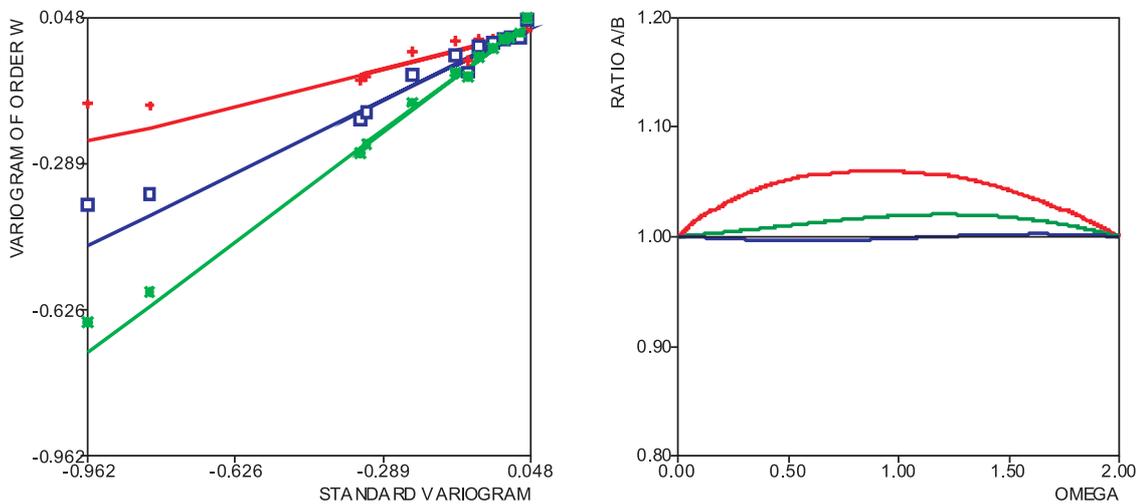


FIGURA 8. Gráfico bi-log (valores em logaritmos) e razões calculadas conforme a equação (7) para a amostra teste2.txt (no gráfico bi-log a cor vermelha representa $\omega=0,5$; azul $\omega=1,0$ e verde $\omega=1,5$).

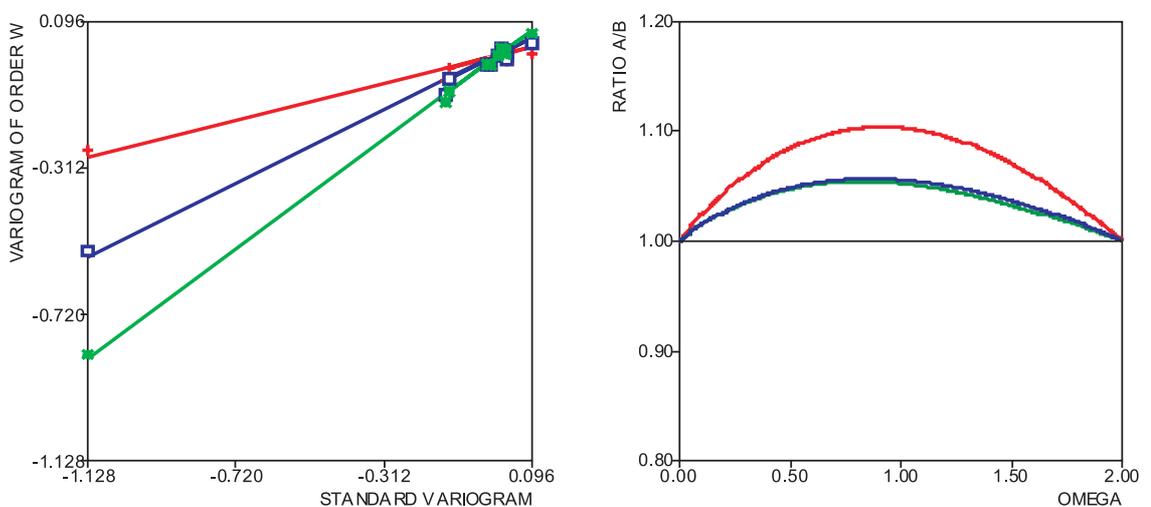


FIGURA 9. Gráfico bi-log (valores em logaritmos) e razões calculadas conforme a equação (7) para a amostra teste3.txt (no gráfico bi-log a cor vermelha representa $\omega=0,5$; azul $\omega=1,0$ e verde $\omega=1,5$).

Analisando-se os resultados obtidos pelo método de Goovaerts (1997), verifica-se que em termos rigorosos, as amostras teste1.txt e teste2.txt seriam rejeitadas para o teste de biGaussianidade. Assim, somente a amostra teste3.txt poderia ser considerada satisfatória para o teste de biGaussianidade e, conseqüentemente, poder-se-ia aceitar a hipótese de multiGaussianidade desse conjunto amostral.

Com relação aos resultados apresentados pelo método de Emery (2005), a única amostra que apresenta todos os pontos alinhados no gráfico bi-log é teste3.txt e, portanto poderia ser considerada satisfatória em termos de biGaussianidade. Por outro lado, é também a amostra que apresenta as maiores razões conforme a equação (7).

Observando-se os gráficos bi-log das Figuras 7 e 8, verifica-se que apenas o primeiro ponto não se alinha com a reta e analisando-se as razões próximas de um, poder-se-ia aceitar essas duas amostras como satisfatórias em termos de biGaussianidade. Em vista

do exposto, o teste de Emery (2005) não é robusto e dá resultados ambíguos como no caso do gráfico de razões da amostra teste3.txt.

Assim, o método de Goovaerts (1997) pode ser considerado mais robusto, pois não deixa dúvidas quanto à biGaussianidade da amostra3.txt.

Os programas que realizam os testes descritos requerem apenas como arquivos de entrada os dados amostrais e o modelo de semivariograma calculado para os dados transformados para o campo Gaussiano. Todos os demais cálculos são feitos pelos programas, sem nenhuma interferência do usuário. Por exemplo, no caso do programa biGauss (método de Goovaerts, 1997), ele calcula todos os variogramas experimentais usando os parâmetros aqueles empregados para o cálculo do variograma experimental dos dados transformados. Todos os programas do Sistema Geovisual seguem essa filosofia, qual seja a de facilitar a utilização de procedimentos muito complexos.

CONCLUSÕES

Este trabalho apresentou os dois métodos existentes para testes de biGaussianidade dos dados, conforme Goovaerts (1997) e Emery (2005). Apesar da proposta de Emery (2005) ser interessante em termos de rápida visualização de pontos alinhados

às retas de referência, as duas verificações não são consistentes com a conclusão final. Assim, o método de Goovaerts (1997) ainda é a melhor opção para o teste de biGaussianidade, haja vista os resultados obtidos.

AGRADECIMENTOS

O autor expressa os seus agradecimentos ao CNPq, através da concessão de bolsa de produtividade em pesquisa (Processo CNPq 303505/2007-9), que permitiu a realização da pesquisa.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Emery, X. Variograms of order ω : a tool to validate a bivariate distribution model. **Math. Geology**, v. 37, p. 163-181, 2005.
2. Goovaerts, P. **Geostatistics for natural resources evaluation**. Oxford University Press, New York, 512 p., 1997.
3. Press, W.H.; Flannery, B.P.; Teukolsky, S.A.; Vetterling, W.T. **Numerical recipes in Pascal: the art of computer programming**. Cambridge University Press, New York, 759 p., 1996.
4. Yamamoto, J.K. **Sistema GeoVisual: guia do usuário**. Instituto de Geociências-USP, 81 p. 2001.

*Manuscrito Recebido em: 1 de agosto de 2009
Revisado e Aceito em: 28 de setembro de 2009*