Aplicação de Métodos Geoestatísticos Univariados e Multivariados em Problemas de Mapeamento de Variáveis de Solo e Plantas

Edson Antonio A. Silva
Prof. PhD. Paulo J. Ribeiro Jr.

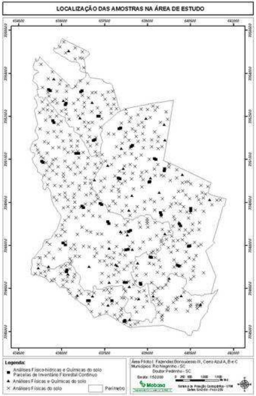
Universidade Federal do Paraná
Programa de Pós-Graduação em Métodos Numéricos em Engenharia

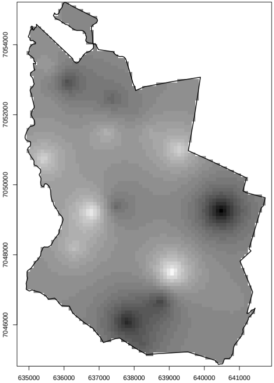
12 de março de 2007

- A definição exata da palavra não consta dos dicionários da língua portuguesa;
- Muitos autores dão definições próprias. A nossa é: Conjunto de métodos relacionados a modelagem de fenômenos espacialmente contínuos.

Área de reflorestamento com estudo em parcelas de inventários florestais contínuos, com plantio de Pinus da espécie P.Taeda L., localizada no município de Rio Negrinho-SC, com área de 2.252 hectares, onde foram levantados dados pedológicos, amostras para análises físicas e químicas do solo e dados relacionados ao rendimento produtivo.

Problema Geoestatístico





63 pontos de análises Físico-hídricas e Químicas, 18 pontos de análises Químicas, 555 pontos de análise Física.

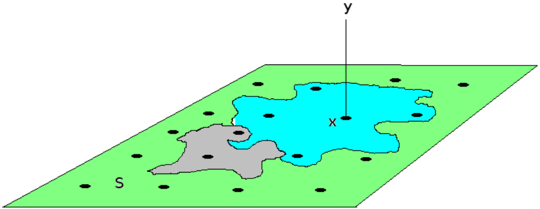
- Propriedades Físicas e Químicas do solo;
- Propriedades ambientais e regionais;
- Propriedades das plantas;
- Rendimento agrícola.

Principais Aplicações Agrícolas

- Caracterização das correlações espaciais entre as variáveis;
- Identificação de diferentes áreas de manejo;
- Elaboração de mapas de informações dos recursos naturais disponíveis, priorizando um conteúdo temático;
- Elaboração de mapas de previsão de rendimento agrícola.

$$\{(x_i; y_i) : x \in \mathbb{R}^2, y \in \mathbb{R}, i = 1, 2, \ldots, n\}$$

- x_i: Localização espacial de uma coordenada (geo)referenciada
- y_i: Medida escalar ou vetorial de uma variável, em uma coordenada.



Premissas

- Definimos o processo estocástico S como um conjunto infinito da variáveis aleatórias que descrevem um fenômeno em uma região do espaço. Esse processo é desconhecido.
- Y é um conjunto finito de observações nessa região do espaço e corresponde a uma realização parcial do processo S.
- Para os propósitos desse trabalho Y é uma v. a. contínua.
- As observações da variável Y são autocorrelacionadas. As observações mais próximas são mais similares entre si do que as observações mais distantes, geograficamente;
- O conjunto de observações y_i nas localizações x_i corresponde a uma única observação do processo estocástico Y.
- Como o processo estocástico Y envolve um conjunto de variáveis da mesma natureza, dizemos que o processo é univariado.

$$Y(x_i) = \mu(x_i) + S(x_i) + \delta_i \qquad i = 1, 2, \ldots, n.$$

- $Y(x_i)$ é uma v. a. com distribuição normal.
  - $E[Y(x_i)|S(x_i)] = \mu(x_i) + S(x_i)$
  - $Var[Y(x_i)|S(x_i)] = \tau^2$
- $\mu(x_i) = \alpha + \beta_1 d_1(x_i) + \beta_2 d_2(x_i) + \ldots + \beta_p d_p(x_i)$ é uma tendência espacial associada às variáveis externas $d_k(x_i)$;
- $\left\{S(x_i) : x_i \in \mathbb{R}^2\right\}$ é um processo gaussiano multivariado com média zero, variância σ^2 e função de correlação $\rho(\cdot)$;
- δ_i são erros aleatórios i.i.d. tal que $\delta_i \sim N(0; \tau^2)$.

Representação Matricial do Modelo Geoestatístico Básico

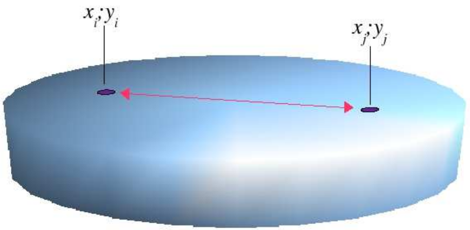
Para um modelo geoestatístico básico temos:

$$\mathbf{Y} \sim MVN(\mathbf{D}\boldsymbol{\beta}, \sigma^2\mathbf{R}(\phi) + \tau^2\mathbf{I})$$

- $\mathbf{D}\boldsymbol{\beta}$ é a representação matricial de $\mu(x_i)$;
- σ^2 é a variância (constante no processo);
- $\mathbf{R}_{n\ x\ n}$ é uma matriz de correlação onde os elementos $[r_{ij}]_{n\ x\ n} = \rho(\|x_i - x_j\|) = \rho(u_{ij})$;
- τ^2 é a variância do erro aleatório δ_i.

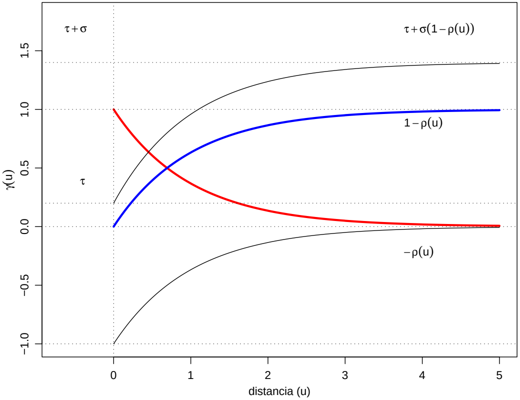
Se $\mathbf{D}\boldsymbol{\beta} = \mu\mathbf{1}$ dizemos que o processo é estacionário no sentido amplo

- A **VARIÂNCIA** mede, de certa maneira, o quão diferentes são essas duas medidas.
- $(Y(x_i) - Y(x_j))$ dá a diferença de duas observações separadas por uma distância $u_{ij} = \|x_i - x_j\|$
- A variância da diferença das medidas irá descrever as similaridades e dissimilaridades entre variáveis tomadas em localizações diferentes separadas por certa distância.



- $Var(Y(x_i) - Y(x_j)) =$
  $Var(Y(x_i)) + Var(Y(x_j)) - 2\ Cov(Y(x_i); Y(x_j))$
  - $Var(Y(x_i)) = Var(\mu_i + S(x_i) + \delta_i) = \sigma^2 + \tau^2$
  - $Cov(Y(x_i); Y(x_j)) = \sigma^2 \rho(u_{ij})$

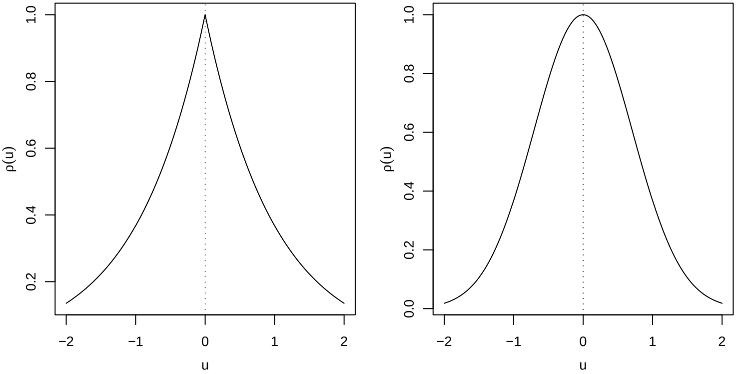
  ..... substituindo .....
- $Var(Y(x_i) - Y(x_j)) = 2(\tau^2 + \sigma^2(1 - \rho(u_{ij})))$
- $\frac{1}{2} Var(Y(x_i) - Y(x_j)) = \tau^2 + \sigma^2(1 - \rho(u_{ij}))$
- $\gamma(u_{ij}) = \tau^2 + \sigma^2(1 - \rho(u_{ij}))$ **SEMIVARIÂNCIA**



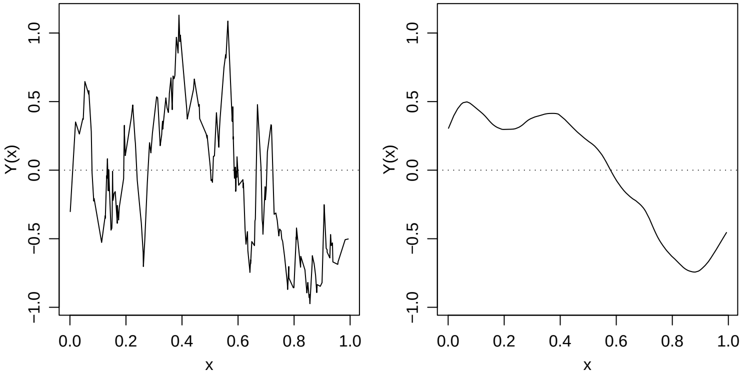
$$\gamma(u_{ij}) = \tau^2 + \sigma^2(1 - \rho(u_{ij}))$$

No modelo $Y(x_i) = \mu_i + S(x_i) + \delta_i$, a função de correlação é a **LEI** que estabelece a associabilidade entre os pares observações geográficas do processo em estudo. Os principais aspectos dessa função são:

- CONTINUIDADE
- DIFERENCIABILIDADE
- $\rho(0) \leq 1$ e $\lim_{u \to \infty} \rho(u) = 0$, tipicamente



A forma da função de correlação determina o quão suave pode ser a variação da informação ao se dela afastar.

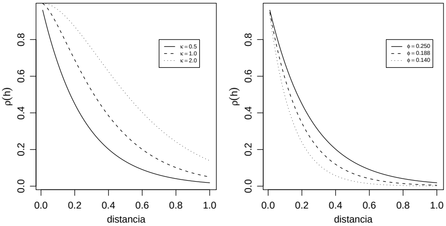


Exemplo unidimensional simulado

FAMÍLIAS PARAMÉTRICAS DE FUNÇÕES DE CORRELAÇÃO

$$\rho(u,\phi,\kappa) = \frac{1}{2^{\kappa-1}\Gamma(\kappa)}\left(\frac{u}{\phi}\right)^{\kappa} K_{\kappa}\left(\frac{u}{\phi}\right)$$

onde K_κ é a função modificada de Bessel (ver Abramowitz, 1965)



- κ é um parâmetro de diferenciabilidade da função;
- ϕ é um parâmetro de alcance prático para a função.

Métodos de Estimação dos Parâmetros do Modelo

- **EXPLORATÓRIO**:
  Feito sobre os pontos de um semivariograma experimental;
- **MÍNIMOS QUADRADOS**:
  Feito sobre os pontos de um semivariograma experimental;
- **MÁXIMA VEROSSIMILHANÇA**:
  Feito sobre o conjunto de observações;
- **BAYESIANO**:
  Método computacionalmente intensivo;

O método consiste em maximizar a função de log-verossimilhança para a função distribuição conjunta $f_Y(y_1, y_2, \ldots, \; y_n)$

$$
\begin{aligned}
l(\beta, \tau^2, \sigma^2, \phi) \quad \propto \quad & -\tfrac{n}{2}\log(\sigma^2\mathbf{R}(\phi) + \tau^2\mathbf{I}) - \\
& -\tfrac{1}{2}(\mathbf{y} - \mathbf{D}\beta)^T(\sigma^2\mathbf{R}(\phi) + \tau^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{D}\beta)
\end{aligned}
$$

- $\mathbf{Y} \sim MVN(\mathbf{D}\beta, \sigma^2\mathbf{R}(\phi) + \tau^2\mathbf{I})$
  Processo gaussiano multivariado.

- Suposição ou conjectura sobre um resultado geo-localizado Y, desconhecido, que poderá ou não ocorrer conforme o valor predito;
- Prever um resultado em uma localização, com base em um número discreto (normalmente pequeno) de observações obtidas dispersamente na área (Krige, 1951);
- Estimar, por regressão linear, baseado naquele modelo proposto inicialmente (Goovaerts, 1997).

$$\hat{Y}(x_0) = \sum_{i=1}^{n} \omega_i Y(x_i) \qquad \text{onde } \omega_{n\text{x}1} = C^{-1}_{n\ x\ n} D_{n\ x\ 1}$$

- C: matriz de autocorrelação baseada nos pontos conhecidos do processo Y;
- D: vetor de correlações entre os pontos conhecidos e um ponto x_0 onde se deseja ser estimado $\hat{Y}(x_0)$;
- ω: peso aplicado a cada ponto $Y(x_0)$ tal que $\sum_{i=1}^{n} \omega_i = 1$.

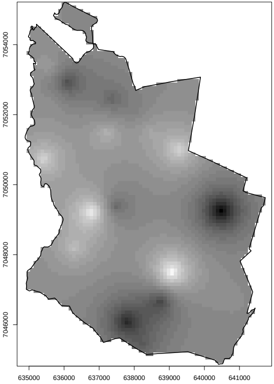
Journel & Huijbregts (1978) e Isaaks & Srivastava (1989)

$$E(\hat{Y}(x_0)|Y) = \mu + \mathbf{r}'(\tau^2\mathbf{I} + \sigma^2\mathbf{R}(\phi))^{-1}(\mathbf{y} - \mu\mathbf{1})$$

- μ: média dos valores observados;
- $\mathbf{r}$: vetor de correlações entre os pontos conhecidos e um ponto x_0 onde se deseja que seja estimado $\hat{y}(x_0)$;
- $\mathbf{y}$: valores observados do processo **Y**

Stein (1999); Schabenberger & Gotway (2005) e Diggle & Ribeiro Jr. (2006)

Mapa de Predição (MOBASA)



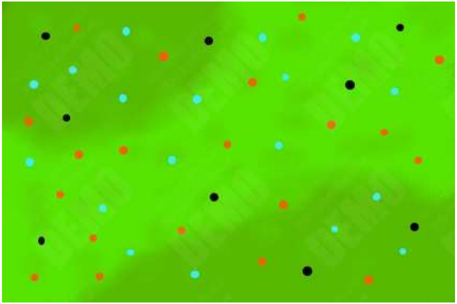
Mapa de predição por krigagem ordinária em 7176 coordenadas a partir de 18 locais observados.

Principais aplicações ...

- Associar a variação espacial de uma variável primária com um conjunto de variáveis secundárias, não necessariamente posicionadas nas mesmas coordenadas;
- Útil para ampliar a quantidade de informações de uma variável primária de difícil medição com informações de variáveis "baratas" que sejam correlacionadas com ela.

Premissa

Dados geoestatísticos multivariados não precisam estar localizados nas mesmas coordenadas para todas as variáveis.



Pontos azuis representam locais onde foi medida a variável Y_1;

Pontos vermelhos representam locais onde foi medida a variável Y_2;

Pontos pretos representam locais onde foram medidas as variáveis Y_1 e Y_2;

Para Journel e Huijbregts (1978) um fenômeno regionalizado por ser representado através de algumas variáveis intercorrelacionadas, estudando-as simultaneamente.

Sob a hipótese de estacionariedade, define a covariância cruzada e o variograma cruzado como:

- $C_{kk'}(h) = E\left\{Z_{k'}(x+h)\ Z_k(x)\right\} - \mu_{k'}\mu_k$
- $2\gamma_{kk'} = E\left\{[Z_{k'}(x+h) - Z_{k'}(x)][Z_k(x+h) - Z_k(x)]\right\}$

Para Isaaks e Srivastava (1989) e Goovaerts (1997) a covariância cruzada e a função de correlação cruzada descrevem a relação espacial entre duas variáveis. Seus estimadores são:

- $C_{uv}(h) = \frac{1}{N(h)} \sum_{(ij)|h_{ij}=h} u_i v_j - m_{u_{(-h)}} m_{v_{(+h)}}$

  $m_{u_{(-h)}} = \frac{1}{N(h)} \sum_{i|h_{ij}=h} u_i$ e $m_{v_{(+h)}} = \frac{1}{N(h)} \sum_{j|h_{ij}=h} v_j$
- $\gamma_{uv}(h) = \frac{1}{2N(h)} \sum_{(ij)|h_{ij}=h} (u_i - u_j)(v_i - v_j)$

Para Diggle e Ribeiro Jr (2007)

Y_1: Processo gaussiano estacionário primário.

Y_2: Processo gaussiano estacionário secundário.

Modelo Bivariado plausível

$$\begin{cases} Y_{1i} = \mu_1 + \sigma_{01} R_0(\phi_a) + \sigma_1 R_1(\phi_b) + \tau_1 & i = 1, 2, \ldots, m \\ Y_{2i} = \mu_2 + \sigma_{02} R_0(\phi_a) + \sigma_2 R_2(\phi_c) + \tau_2 & j = 1, 2, \ldots, n \end{cases}$$

Matriz de Covariância

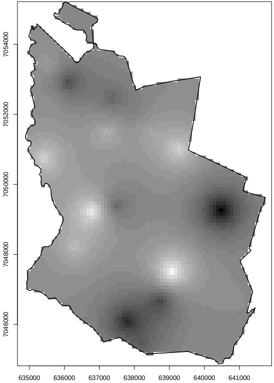
$$\boldsymbol{\Sigma} = \left(\begin{array}{c|c} Cov(Y_1; Y_1) & Cov(Y_1; Y_2) \\ \hline Cov(Y_2; Y_1) & Cov(Y_2; Y_2) \end{array}\right)$$

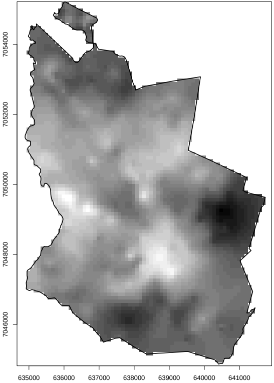
Predição Espacial Bivariada

- Predição de uma variável Y_1 com o "apoio" de uma segunda variável Y_2 (co-variável) tomada nas mesmas coordenadas na primeira (Krigagem com co-variável);
- Predição de uma variável de interesse primário Y_1 em coordenadas de uma segunda variável Y_2, sabidamente correlacionada com a primeira (Co-Krigagem).

- Estimar a função de correlação e os parâmetros de um modelo bivariado (likBGCC) – *Bivariate Gaussian Common Componente Model* (no geoR);
- Predizer valores da variável primária nas coordenadas da variável secundária (ampliação);
- Predizer valores da variável primária em coordenadas compatíveis com a produção de um mapa.

Mapa de Predição (MOBASA)





Mapa de predição por krigagem ordinária em 7176 coordenadas a partir de 18 locais observados (esquerda) e por co-krigagem, a partir de 555 locais da segunda variável.

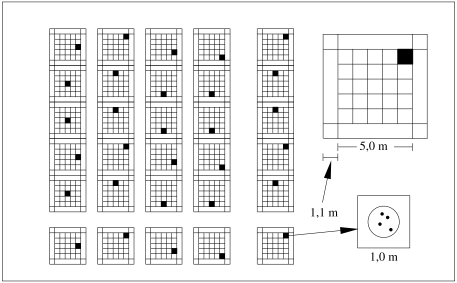
- Ampliar a revisão bibliográfica;
- Focar propostas que adotem um conjunto de processos estocásticos gaussianos multivariados $\{S_1(x_i), S_2(x_j), \ldots S_p(x_k) : i \neq j \neq \ldots \neq k; p \in \mathbb{Z}^+\}$ nem todos independentes;
- Avaliar métodos multivariados (análise de componentes principais e análise fatorial) para a redução no número de processos.
- Implementar computacionalmente a função MGCCM - *Multivariate Gaussian Common Component Model*;

Objetivos

- Aplicar e comparar o método multivariado em estudo de caso;
- Avaliar criticamente as estratégias de aplicação de modelos multivariados tomando como base dados simulados.

Estudo de Caso #2

COODETEC: Dados de pesquisa da Unioeste financiada pelo CNPq em área de agricultura de precisão com 1,33 ha, localizada no Centro de Pesquisa Eloy Gomes da Cooperativa Central Agropecuária de Desenvolvimento Tecnológico e Econômico Ltda, no município de Cascavel-PR, cultivada com soja na safra 1998, onde foram coletadas 256 amostras de solo em grid semi-regular, para análise de atributos químicos e dados de produtividade.


5,0 m
1,1 m
1,0 m

Toolbox

- Linguagem e ambiente operacional **R**;
- Pacote geoestatístico **geoR**;
- Sistema operacional GNU/Linux.

Recursos computacionais sob licença GPL (*General Public Licence*)

- ABRAMOWITZ, M. & STEGUN I., *Handbook of Mathematical Functions*. 9.ed., New York: Dover, 1965.
- DIGGLE, P. J. & RIBEIRO Jr P. J., *Model-based Geostatistics*. USA: Springer Series in Statistics, 2006.
- GOOVAERTS, P., *Geostatistics for Natural Resources Evaluation*. Oxford:, Oxford University Press, 1997.
- ISAAKS, E. H. & R. SRIVASTAVE, M., *Applied GEostatistics*. New York: Oxford University, 1989.
- JOURNEL A. G. & HUIJBREGTS Ch. J., *Mining Geostatistics*. London: Academic Press, 1978.
- KRIGE D. G., *A Statistical Approach to Some Mine Valuations and Allied Problems at Witwatersrand*. University of Witwatersrand, 1951. Master's thesis.

REFERÊNCIAS

- *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna:, 2006, http://www.R-project.org.
- RIBEIRO Jr, P. J. & DIGGLE P. J., geoR: A package for geostatistical analysis, *R-NEWS*, v01, n2, 2001, http://cran.r-project.org/doc/Rnews.
- SCHABENBERGER O. & GOTWAY, A., *Statistical Methods for Spatial Data Analysis*, New York: Chapman-Hall, 2005.
- STEIN, M. L. *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer series in statistics. 1999.
- WAKERNAGEL, H., *Multivariate geostatistics: an introduction with applications*, 3.ed., Germany: Springer series in statistics, 2003.