Fundamentos de Geo-estatística

Rachid Muleia & Mauro Langa

Universidade Eduardo Mondlane
Faculdade de Ciências
Departamento de Matemática e Informática

April 23, 2018

Conteúdo

Breve introdução ao Software R

Antes de darmos início ao estudo da Geo-estatística, iremos fazer uma breve introdução a programação em **R**. O **R** é uma linguagem de programação e também um ambiente de desenvolvimento integrado para cálculos estatísticos e gráficos.

Esta linguagem foi originalmente criada por Ross Ihaka e por Robert Gentleman no departamento de Estatística da universidade de Auckland, Nova Zelândia, e foi desenvolvido em um esforço colaborativo de pessoas em vários locais do mundo.

A linguagem R é largamente usada entre estatísticos e analistas de dados para desenvolver software de estatística e análise de dados. Pesquisas e levantamentos com profissionais da área mostram que a popularidade do R aumentou substancialmente nos últimos anos. *Importa referir que o R é um software grátis, disponível na internet para qualquer um.*

Porque usar o R

- Custo zero, isto é o software é grátis
- Contém implementações de métodos avançados, que não são facilmente encontrados em outros programas estatísticos
- Capacidade de produção de gráficos de alta qualidade.
- Possibilidade de criar e compartilhar pacotes
- É amplamente utilizado não apenas na academia, mas em empresas como Google, New York Times, Pfizer, Bank of America, Merck, InterContinental Hotels Group, Shell, etc.

Instalando o R

Para instalar o R basta aceder a página web https://cran.r-project.org/bin/windows/base/R-3.3.2-win.exe. Esta versão é apenas para windows.

Ou vá a página https://www.r-project.org/ e baixe a versão que for compatível ao seu computador.

Após instalar o R podemos inicializar programa fazem um duplo "click" no icon visível no desktop.

Caso este não esteja visível no desktop, podemos procurar no startmenu.

Foco do R para esta disciplina

Não vamos ver a programação ao fundo, mas apenas vamos aprender alguns tópicos básicos para nos auxiliarem nas nossas análises ao longo desta disciplina.

A aprendizagem sobre o software vai se focalizar na importação de dados que esteja armazenados em diferentes formatos, tais como, `.csv`, `.txt`.... O R também pode aceder a bases de dados relacionais(onde a informação está armazenada em tabelas, e tais tabelas estão relacionadas através de chaves primárias, tal como é o caso de bases de dados desenhadas em SQL).

Além de importar dados, o R pode servir como uma calculadora científica, isto é, pode-se fazer operações aritméticas básicas. À parte a operações aritiméticas, o R é uma linguagem de programação virada para análises estatística, o que significa que pode-se muito bem calcular as medidas de posição, de central, de dispersão e muito mais.

Introdução à Geo-estatística

A geoestatística tem por objecto a caracterização da dispersão espacial e espácio-temporal das grandezas que definem a quantidade e a qualidade de recursos naturais, tais como florestas, recursos geológicos, hidrológicos, etc.

A geoestatística nasceu da necessidade da modelização de recursos geológicos-caracterização da dispersão espacial da concentração de metais em jazigos por volta da década da 50.

Os primeiros passos foram dados por na África do Sul por D.G. Krige e H.S. Sichel, engenheiros de minas. Krige observou que podia melhorar as estimativas da concentração de minérios em blocos, se levasse em conta a concentração em blocos vizinhos (indicativo de que a geo-estatística lida com observações correlacionadas no espaço)

Introdução à Geo-estatística- Definição

Pode-se definir a *geoestatística* como sendo um ramo da estatística dedicada à caracterização de fenómenos espaciais e temporais. Esta é muito útil quando se tenta interpolar dados dispersos de observações de campo STEPPE(2016).

A *geoestatística* é um subconjunto de estatísticas especializadas em análise e interpretação de dados geograficamente referenciados. Em outras palavras, a *geoestatística* compreende técnicas estatísticas que são ajustadas aos dados espaciais (Hengl, 2007)

Porquê a geoestatística

A geoestatística é importante para modelizar/perceber a variabilidade espacial, bem como para quantifica-la. *São poucos os fenómenos/ ou grandezas que não variam ao longo do espaço (ex: gravidade, abundância de oxigénio)*. Maior parte das grandezas variam ao longo do espaço, por exemplo: a precipitação, a vegetação, densidade populacional, etc.

A estatística clássica assume que as observações são independentes, enquanto que maior parte de dados geo-referenciados apresentam dependência espacial, com a premissa de que observações mais próximas apresentam maior dependência e observações mais afastadas apresentam menor dependência espacial (Yost, 1982). Ignorar este comportamento pode ser prejudicial, pois pode nos levar a conclusões falsas ou inconsequentes. A geo-estatística apresenta ferramentas e técnicas para modelizar tal dependência espacial.

Tipo de dados em análise espacial

Dados geo-estatísticos (superfícies contínuas)

Seja $\boldsymbol{s} \in \mathbb{R}^d$ uma localização genéria num espaço euclidiano de dimensão d e seja $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in \mathbb{R}^d\}$ função aleatória espacial, Z denota o atributo de interesse.

1. $Z(\boldsymbol{s})$ pode ser observado em qualquer ponto do domínio (contínuo);
2. Todos os pontos no domínio D são não-estocástico(fixos, o domínio é o mesmo para todas a realização da função aleatória espacial)

Ex: Concentração carbono nas minas de carvão de Moatize; Valores de precipitação em uma região. (dados geológicos, climatológicos)

Em suma, podemos dizer que neste tipo de dados, espera-se que estes estejam distribuídos de forma contínua no espaço.

Normalmente, o processo espacial é observado em alguns pontos de uma região(domíni
e com base nos valores observados, a análise geo-estatística reproduz o com-
portamento do processo espacial em toda região de interesse/domínio.

Neste contexto, a geoestatística tem como objectivo :

- Prever alguns pontos não amostrados na região de interesse;
- Estimar um valor médio sobre a área de interesse ou para uma parte desta.

O interesse primordial é a quantificação da correlação espacial entre as observações com base na estimação do *semivariograma* e a posterior usar essa informação para atingir os objectivos acima.

Dados de Área

1. A região D de interesse é discreta, isto é $Z(\boldsymbol{s})$ pode ser observado em locais fixos que podem ser enumerados/indexados. Estes podem ser pontos, regiões, províncias etc.
2. Os locais/pontos em D são não-estocásticos.

Ex: Taxa de desemprego por província; Rendimento agrícola por parcela(canteiro);

Dados de ponto padrão

Diferentemente dos dados geoestatísticos e regionais, nos dados de ponto padrão a região de interesse não é fixa, mas sim aleatória. Este tipo de dado surge quando o interesse reside em estudar/analisar os locais onde os enventos de interesse ocorrem.

Ex: Localização dos incêndios na Cidade de Maputo; Localização espécies vegetais

Modelos Geoestatísticos

Nos modelos geoestatísticos os dados amostrais são interpretados como provenientes de um processo aleatório . O facto destes modelos incorporarem incerteza na sua conceptualização não significa que o fenómeno em si – a floresta, aquífero, o jazigo mineral – tenha resultado de um processo aleatório , mas serve somente de base metodológica a inferência espacial ou estimação de grandezas em áreas não amostradas e à quantificação da incerteza associada ao estimador (Soares, 2014)

Modelos Geoestatísticos-Variável Aleatória

Um valor localizado espacialmente em x_1(denominação genérica de um conjunto de coordenadas geográficas) é interpretado como uma realização da variável aleatória (v.a) $Z(x_1)$. Na região $\boldsymbol{A}$, onde se dispersa o conjunto de amostras, temos as realizações de N v.a $Z(x_1)$, $Z(x_2)$, $\cdots$, $Z(x_N)$ correlacionadas entre si.

Para uma v.a pode-se definir dois primeiros momentos:

$$E(Z(x_i)) = \mu(x_i) = \int_{-\infty}^{+\infty} z dF_{x_i}(z) = \int_{-\infty}^{+\infty} z f_{x_i}(z)\, dz \tag{1}$$

$$\text{var}\{Z(x_i)\} = \int_{-\infty}^{+\infty} [z - \mu(x_i)]^2 dF_{x_i}(z) \tag{2}$$

$f(z)$ e $F(z)$ representam as funções densidade de probabilidade e de distribuição de probabilidade

Para além da média e a variância, pode-se definir a covariância entre duas v.a $Z(x_1)$ e $(Z(x_2)$

$$C(Z(x_1), Z(x_2)) = E\{Z(x_1)Z(x_2)\} - \mu(x_1)\mu(x_2)$$
$$E\{Z(x_1)Z(x_2)\} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf_{x,y}(x,y)dxdy$$

O coeficiente correlação e o variograma são respectivamente dado por

$$\rho(Z(x_1), Z(x_2)) = \frac{C(Z(x_1), Z(x_2))}{\sqrt{\text{var}\{Z(x_1)\}\text{var}\{Z(x_2)\}}} \quad (3)$$

$$\gamma(Z(x_1), Z(x_2)) = E\Big\{[Z(x_1) - Z(x_2)]^2\Big\} \quad (4)$$

Em outras palavras pode-se definir o variograma como sendo a variância das primeiras diferenças das v.a: $\text{var}\Big\{Z(x_1) - Z(x_2)\Big\}$

Modelos Geoestatísticos-Função Aleatória Estacionária

O conjunto de variáveis aleatórias $Z(x_i)$, $i = 1, \cdots, N$, constituem uma função aleatória da qual só se conhece uma realização $z(x_i)$

Com uma só realização de cada v.a é teoricamente impossível determinar qualquer parâmetro estatístico.

A solução consiste em assumir diversos graus de estacionariedade,de tal forma que a inferência de algumas estatística seja possível. Assim, pode-se assumir que todas as variáveis aleatoórias tem a mesma média:

$$E\{Z(x_1)\} = E\{Z(x_2)\} = \cdots = E\{Z(x_i)\} = E\{Z(x)\} = \mu \qquad (5)$$

Desta forma este parâmetro passa a ser independente da localização.

$$\mu = \frac{\sum_{i=1}^{N} Z(x_i)}{N}$$

- A hipótese de estacionariedade da média é parte integrante e fundamental do modelo probabilista geoestatístico validade ou refutada na prática uma vez que, na realidade, só existe uma realização da função aleatória
- Contudo, ela deve ser julgada apropriada, ou não, dependendo da homogeneidade da amostra na área A em que a variável se distribui.
- A hipótese de estacionariedade da média implica que esta pode ser estimada pela média aritmética dos valores amostrais.

Estacionariedade

Visto que o conceito de *estacionariedade* é fundamental para fazer inferências é importante que este seja definido com "rigor"

De uma forma vaga o conceito de *estacionariedade* em geoestatística é sinónimo de de homogeneidade da v.a (também conhecida por variável regionalizada).

A hipótese de estacionariedade implica que as propriedades probabilísticas de um conjunto de observações não depende da localização especifica onde as observações foram amostradas, mas apenas depende da sua separação.

Portanto,em termos matemáticos e probabilísticos , a hipótese de estacionariedade refere-se ao comportamento regular dos momentos de uma função aletória sobre uma região ou intervalo de tempo (constância dos momentos, isto é não variação da média ou da variância)

Estacionariedade estrita

- Em estacionariedade estrita a distribuição conjunta de probabilidades das variáveies aleatórias é invariante para qualquer translação.
- Um processo geoestatístico $\{Z(\boldsymbol{x}) : \boldsymbol{x} \in D\}$ é considerado estritamente estacionario se :

  $$F\{Z(x_1), \cdots, Z(x_n)\} = F\{Z(x_1 + h), \cdots, Z(x_n + h)\}$$
- Isto significa que o processo ou fenómeno é homogéneo ao longo da região. Assim como, se a média bem como a variacia existirem, estes momentos não mudam quando há uma translação.

Estacionariedade de segunda ordem

Um processo ou fenómeno é considerado estacionário no sentido lato ou estacionário de segunda ordem se:

- Os dois primeiros momentos da distribuição da função aleatória estiverem definidos, nomeadamente a média e a covariância
- A média é constante, isto é não depende da localização

$$E\{Z(\boldsymbol{x})\} = \mu(\boldsymbol{x}) = \mu$$

- A covariância para todos pares de v.a $Z(\boldsymbol{x})$ e $Z(\boldsymbol{x}+\boldsymbol{h})$ existe e depende apenas de um vector de separação $\boldsymbol{h}$ (distância e direcção), e não da localização especiífica

$$C\{Z(\boldsymbol{x}), Z(\boldsymbol{x}+\boldsymbol{h})\} = C(\boldsymbol{h})$$

- Se a covariancia existe e é finita, então tem-se que $C(\mathbf{0}) = \text{Var}\{Z(\boldsymbol{x})\} = \sigma^2$

A hipótese de estacionariedade de segunda ordem pode ser interpretada como se a variável regionalizada toma-se valores que flutuam em torno de um valor constante(média)e que por sua vez a variação dessas flutuações é a mesma em todo domínio(variância constante)

Sob a hipótese de estacionariedade de segunda ordem, o variograma e o co-variagrma são equivalentes:

$$
\begin{aligned}
\gamma(\boldsymbol{h}) &= \frac{1}{2}\mathsf{Var}\{Z(\boldsymbol{x}+\boldsymbol{h}) - Z(\boldsymbol{x})\} \\
&= \frac{1}{2}\Big\{\mathsf{Var}[Z(\boldsymbol{x}+\boldsymbol{h})] + \mathsf{Var}[Z(\boldsymbol{x})] - 2C[Z(\boldsymbol{x}+\boldsymbol{h}), Z(\boldsymbol{x})]\Big\} \\
&= \frac{1}{2}C(\mathbf{0}) + \frac{1}{2}C(\mathbf{0}) - \frac{2}{2}C(\boldsymbol{h}) \\
&= C(\mathbf{0}) - C(\boldsymbol{h})
\end{aligned}
$$

- Observe que se um processo geoestatístico $\{Z(\boldsymbol{x}) : \boldsymbol{x} \in D\}$ for estritamente estacionário e com média e covariância definida, é também estacionário de segunda ordem. O recíproco, geralmente, não é válido.
- Para processos gaussianos, estacionariedade de segunda ordem é equivalente a estacionariedade estrita
- Um processo geoestatístico pode ser considerado quasi-estacionário se as respectivas hipóteses de estacionariedade forem válidas para uma certa distância $|\boldsymbol{h}| < d$

Estacionariedade intrínseca

A hipótese de estacionariedade de segunda ordem por vezes, pode ser considerada como sendo uma suposição estrita, pois ela requer que a variância esteja definida.

Um fenómeno pode ter uma varinĉia infinita e ser difícil modeliza-lo usando a hipótese de estacionariedade de segunda ordem.

Porém há casos, em que os incrementos ou diferenças $Z(\boldsymbol{x}+\boldsymbol{h})-Z(\boldsymbol{x})$ tem variância finita, e por consequência são estacionário de segunda ordem.

Quando as primeiras diferenças de um processo geoestatistico são estacionárias de segunda ordem, este é considerado um processo *intrinsecamente estacionário*

Para que uma função aleatória seja intrinsecamente estacionária é necessario que o valor esperado da variável não dependa da posição $\boldsymbol{x}$:

$$E\{Z(\boldsymbol{x})\} = \mu \quad \text{para qualquer } \boldsymbol{x} \tag{6}$$

Para além da disso, a variância dos incrementos/diferenças deve ser finita. Desse modo, o variogram $\gamma(\boldsymbol{h})$, que é definido como sendo a metade do valor esperado do quadrado das diferenças entre pares de v.a's existe e depende apenas de $\boldsymbol{h}$.

$$\gamma(\boldsymbol{h}) = \frac{1}{2} E\Big\{ \big[Z(\boldsymbol{x}+\boldsymbol{h}) - Z(\boldsymbol{h}) \big]^2 \Big\} \tag{7}$$

Isto significa que o valor da semivariance é o mesmo para todas as observações separadas um certo desfasamento/lag independentemente da localização dos pares de observações (Loyd & Atkinson, 2004).

Estacionariedade de segunda ordem implica estacionariedade intrínseca, mas o inverso não é válido.

O variograma

O variograma é uma ferramenta chave em análises geoestatísticas. O variograma serve para caracterizar a dependência espacial da variável de interesse. Resumindo, o variograma descreve como é que duas observações se tornam diferente a medida que a separação aumenta.

- a medida que os valores de $\boldsymbol{h}$ aumentam, a dispersão das amostras vai aumentado, e por sua vez a correlação vai diminuindo.
- visto que o variograma é uma função teorética, urge a necessidade de introduzir o variograma experimental que é calculado com base nos dados observados.

O variograma experimental/empírico pode ser estimado a partir de $N(\boldsymbol{h})$ pares de observações $z(\boldsymbol{x}_i)$, $z(\boldsymbol{x}_i + \boldsymbol{h})$, $i = 1, 2, 3 \cdots$

$$\hat{\gamma}(\boldsymbol{h}) = \frac{1}{2N(\boldsymbol{h})} \sum_{i=1}^{N(\boldsymbol{h})} \Big[z(\boldsymbol{x}_1) - z(\boldsymbol{x}_i + \boldsymbol{h}) \Big]^2 \tag{8}$$

O variograma é estimado com base na metade da média do quadrado das diferenças entre observações que são separados por um vector $\boldsymbol{h}$

O variograma é usado quando se assume a hipótese de estacionariedade intrínseca

Em caso de se assumir estacionariedade de segunda ordem, a modelização do fenómeno é feita com base no covarigrama, visto que a covariância só existe quando o fenómeno é estacionário de segunda ordem (Loyd & Atkinson, 2004)

$$\hat{C}(\boldsymbol{h}) = \frac{1}{N(\boldsymbol{h})} \sum_{i=1}^{N(\boldsymbol{h})} \Big[z(\boldsymbol{x}_i) - \bar{z} \Big] \Big[z(\boldsymbol{x}_i + \boldsymbol{h}) - \bar{z} \Big] \tag{9}$$

Observe que $\hat{\gamma}(\boldsymbol{h}) \neq \hat{C}(\boldsymbol{0}) - \hat{C}(\boldsymbol{h})$, porém para $N(\boldsymbol{h})/n \approx 1$, a diferença será minúscula (Cressie, 2015).

Geralmente, o variograma é mais preferido em relação ao covariograma, visto que este não necessita do conhecimento da média para a sua estimação(a média na prática não é conhecida, a sua estimação introduz um viés na estimação do covariograma)(Montero & Mateu, 2015).

Estimação robusta do variograma

O variograma experimental apresenta algumas desvantagens:

1. é sensível a outliers
2. o quadrados das diferença($Z(\boldsymbol{x}+\boldsymbol{h}-Z(\boldsymbol{x}))$ podem se apresentar distorcido(dados não seguem distribuição normal)
3. os quadrados das diferenças também serão v.a's dependentes.
4. a mais subtil objeção surge devido a assimetria da distribuição. Se o fenómeno geoestatístico for um processo de Gauss, então $\Big[Z(x+\boldsymbol{h}) - Z(x)\Big]^2 \sim 2\gamma(\boldsymbol{h})\chi^2_{(1)}$. A distribuicao qui-quadrado é assimétrica. Porém, se $X \sim \chi^2_{(1)}$, então $X^{1/4}$ terá uma distribuição razoavelmente simétrica. Desse modo, espera-se que $\Big[Z(x+\boldsymbol{h}) - Z(x)\Big]^2$ tenha também uma distribuição razoavelmente simétrica.

Por consequência deste resultado, Crissie & Hawkins(1980) propuseram um estimador robusto não enviesado para variograma:

$$2\bar{\gamma}(\boldsymbol{h}) = \frac{1}{0.457 + 0.494/N(\boldsymbol{h})} \left\{ \frac{1}{N(\boldsymbol{h})} \sum_{i=1}^{N(\boldsymbol{h})} \Big[|Z(x_i + \boldsymbol{h}) - Z(x_i)| \Big]^{1/2} \right\}^4$$

&

$$2\tilde{\gamma}(\boldsymbol{h}) = \frac{\left\{ \text{med} \Big[|Z(x_i + \boldsymbol{h}) - Z(x_i)| \Big]^2 \right\}^4}{0.457}$$

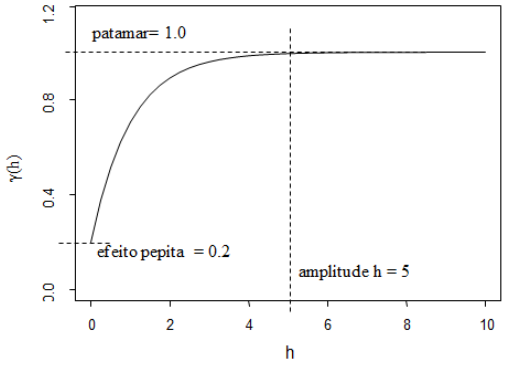
Mais detalhes sobre estimação robusta do variograma podem ser encontrados em Cressie(2015).

Características dum variograma

Efeito pepita- representa micro variações devido a erros de medição. Pode ser estimado a partir do variograma experimental para $h = 0$

Amplitude- A distância em que o variograma atinge o patamar, isto é, a distância a partir da qual os dados não estão mais correlacionados.

Patamar- Corresponde a variância da v.a $V\left[Z(x)\right]$. Este valor corresponde a valor do variograma para uma h = amplitude



Exemplo- Cálculo do variograma

Para o cálculo do variograma e respectiva representação podemos usar o software R e fazer o uso do pacote **geoR**, e a posterior usamos a função "`variog()`"

```
> install.packages("geoR") instalação do "geoR"
> library(geoR)
> data(s100) # aceder os dados simulados do geoR

############### calculando o variograma experimental##############

> bin1<- variog(s100, uvec=seq(0,1,l=11)) #estimador classico
> bin2  <- variog(s100, uvec=seq(0,1,l=11),
+ estimator.type= "modulus") #estimador robusto Crissie & Hawkins(1980)

> plot(bin1, main = "classical estimator") # visualizacao grafica
> plot(bin2, main = "modulus estimator") # visualizacao grafica
```

Exemplo- Cálculo do variograma

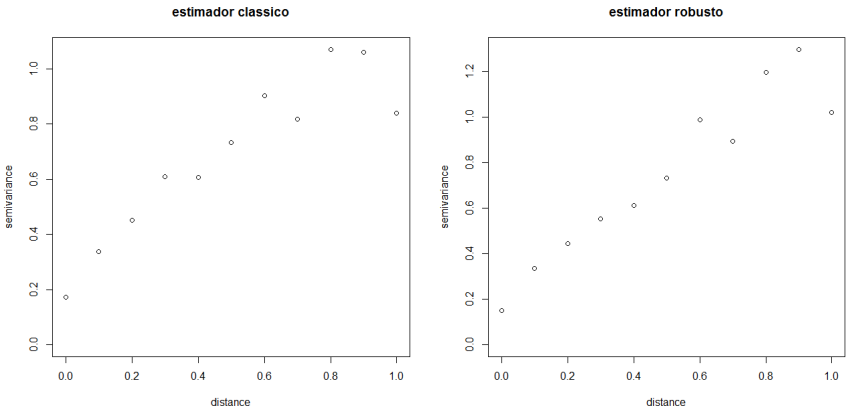


Figure 1: a figura a esquerda corresponde ao variograma clássico e a figura a direita corresponde ao variograma robusto sugerido por Crissie & Hawkins(1980).O variograma robusto é menos sensível a outliers e a desvios da normalidade.

Modelos teóricos de semivariograma

Os modelos teóricos de variogramas são funções usadas para representar (que são ajustadas ao) o variograma experimental, visto que este não cumpre com algumas propriedades que são fundamentais para a krigagem(método de interpolação espacial). Existem variás funções de variogram para modelizar fenómenos geoestatísticos, mas apenas iremos apresentar alguns modelos comumente usados para a modelização espacial. Os variogramas que serão apresentados são de natureza **isotrópica**, isto é, não depende da direcção.

Os modelos de semivariograma ou covariograma não podem ser funções arbitrárias. Estes devem satisfazer algumas algumas propriedades teóricas

Modelos teóricos de semivariograma

Propriedades do variograma

1. O variograma na origem é igual a zero $\gamma(\mathbf{0}) = 0$, embora na prática apresenta uma descontinuidade. A descontinuidade na origem é designada por *efeito pepita*.
2. O variograma é uma função par $\gamma(-\boldsymbol{h}) = \gamma(\boldsymbol{h})$.
3. O variograma sempre assume valores não negativos $\gamma(\boldsymbol{h}) \geqslant 0$, enquanto que o covariograma pode assumir valores negativos .
4. A funcao de variograma é negativa definida $\sum_{i=1}^{n} \sum_{j=i}^{n} \lambda_i \lambda_j \gamma(\boldsymbol{x_i} - \boldsymbol{x_j}) \leqslant 0$.

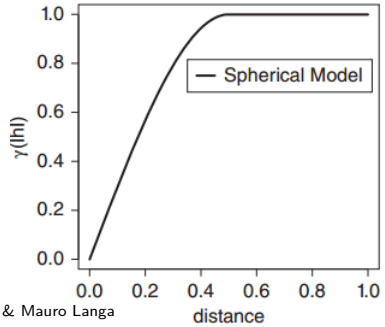
TPC: Demonstre as propriedades 2 e 4.

Modelos teóricos de semivariograma

Modelo esférico

$$\gamma(\boldsymbol{h}) = \begin{cases} C(\mathbf{0})\left(1.5\dfrac{|\boldsymbol{h}|}{a} - 0.5\left(\dfrac{|\boldsymbol{h}|}{a}\right)^3\right) & \text{se} \quad |\boldsymbol{h}| \leqslant a \\ C(\mathbf{0}) & \text{se} \quad |\boldsymbol{h}| > a \end{cases}$$

- $C(\mathbf{0})$ representa o **patamar**, limite superior para qual tendem os valores com o aumento dos valores de $|\boldsymbol{h}|$
- $|\boldsymbol{h}| = a$ representa a amplitude, a distâancia a partir da qual os valores de $\gamma(h)$ param de crescer e são iguais a um patamar que é normalmente coincidente com a variância de $Z(x)$
- A *amplitude* mede a distância a partir da qual os valores de $Z(x)$ deixam de estar correlacionados

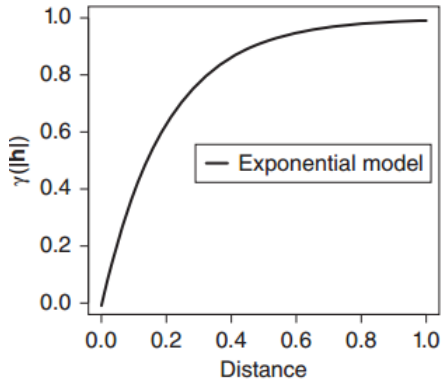


Modelo exponencial

$$\gamma(\boldsymbol{h}) = C(\mathbf{0})\left(1 - \exp\left(-\frac{\boldsymbol{h}}{a}\right)\right)$$

- O variograma exponencial atinge o patamar assimptoticamente a medida que $\boldsymbol{h} \to \infty$

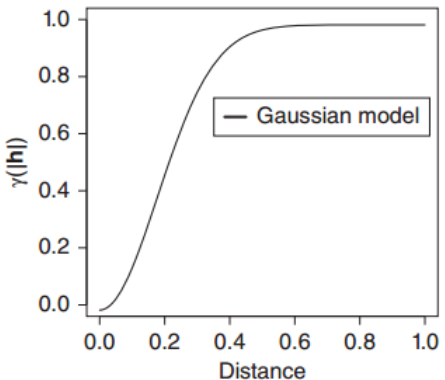
- Para o variograma exponencial tem-se a *amplitude efectiva* a', que corresponde ao valor de $\boldsymbol{h}$ a covariância é aproximadamente igual a 5% do seu valor na origem($\boldsymbol{h} = 0$), em outras palavras pode-se dizer que o valor da amplitude efectiva num modelo exponencial é a distância em que o modelo atinge 95% do patamar: $\gamma(a') = 0.95\,C(\mathbf{0})$



Modelo Gaussiano

$$\gamma(\boldsymbol{h}) = C(\mathbf{0})\left(1 - \exp\left(-\frac{\boldsymbol{h}^2}{a^2}\right)\right)$$

- Ao contranto do modelo esférico e exponencial, o modelo Gaussiano apresenta um comportamento parabólico na origem.
- Tal como no modelo exponencial, a amplitude efectiva é a distância para a qual o modelo atinge 95% do patamar: $\gamma(a') = 0.95C(\mathbf{0})$.



Modelo de potência

Os variogramas descritos atrás têm todos um patamar como limite para o qual tendem os valores de $\gamma(\boldsymbol{h})$ quando $h \to \infty$. Estes tipos de variogramas são para modelizar os chamados fenómenos de transição, os quais são caracterizados por uma distância (amplitude) a partir da qual deixa de haver correlação entre as amostras. Nos fenómenos de transição existe sempre uma relação entre a covariância e o variograma (Soares, 2014).

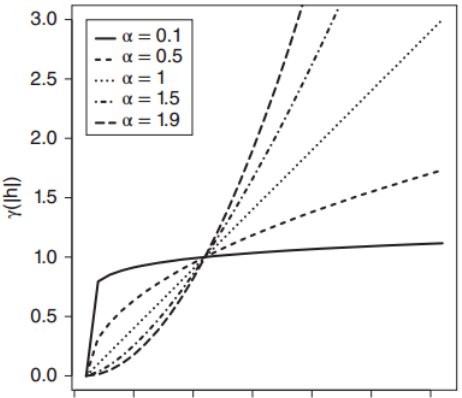
Contudo, existem outros fenómenos em que o crescimento de $\gamma(h)$ é contínuo e não tende para um patamar. Estes são os fenómenos não estacionários, os quais não apresentam variância finita ou noção de covariância, por serem grandezas que crescem com a dimensão do campo de dispersão de $Z(x)$.

O variograma de potência é dado por:

$$\gamma(h) = C(\mathbf{0})\boldsymbol{h}^{\alpha}$$

com α compreedido de 0 a 2.

Dependendo do expoente α, o variograma pode ter um andamento linear($\alpha = 1$), logarítimico ($0 < \alpha < 1$) ou parabólico ($1 < \alpha < 2$). Quando o $\alpha = 0$ temos um variorama com *efeito pepita puro*. O modelo com efeito pepita puro reflecte a ausência da dependência espacial.



Ajustamento do variograma

Os variogramas experimentais não podem ser usados directamente para fazer interpolações espaciais, visto que este não satisfazem as propriedades de um modelo teórico de semivariograma (Cressie, 2015). Ademais, o uso dos modelos téoreticos garante que a variância do erro de interpolação resultante da krigagem seja não negativa (Montero & Mateu, 2015).

O ajustamento do variograma pode ser feito consoante o ajustamento manual que consiste numa inspeção visual, ou usando procedimentos estatísticos. Porém Webster & Oliver (2011) recomendam o uso dos dois métodos. Onde primeiro pode se usar a inspeção visual para seleccionar os variogramas que melhor captam as principais características do variograma. Depois, pode-se usar os procedimentos estatísticos para selecionar o modelo teórico dentre os modelos pre-seleccionados.

Ajustamento do variograma

Contudo, Wackernagel (2003) afirma que não é deveras relevante se o ajustamento é feito de forma visual ou usando procedimentos estatísticos. O que realmente importa é o tipo de continuidade e a hipótese de estacionariedade assumida. Pois estes pressupostos irão conduzir a selecção do modelo apropriado. Armstrong & Wackernagel (1988) por sua vez reiteram, que defacto, a expressão analítica do variograma/semivariograma não é tão relevante desde que as característica do fenómeno em estudo sejam respeitadas.

Ajustamento do variograma

Procedimento estatístico

Existem vários procedimentos estatísticos que permitem estimar o conjunto de parâmentros (efeito pepita, patamar e a amplitude) dum semivariograma, dentre os quais, pode se destacar:

1. Método dos minímos quadrados (mínimos quadrados ordinários, mínimos quadrados generalizados e mínimos quadrados ponderados.
2. Métdos baseados na maximização da função de verossimilhança (Métod de máxima verossimilhança e método de máxima verossimilhança restrita)
3. Verossimilhança composta, assim como, o método de máxima verossimilhança usa pseudo-dados no lugar de dados observados.

Método dos mínimos quadrados ordinários

O método dos mínimos quadrados ordinários consiste em estimar o vector $\boldsymbol{\theta}$ minimizando :

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{k} \left\{ \hat{\gamma}(\boldsymbol{h_i}) - \gamma(\boldsymbol{h_i}, \boldsymbol{\theta}) \right\}^2 \tag{10}$$

onde $\hat{\gamma}(\boldsymbol{h_i})$ representa o valor do semi-varigrama experimental para uma distância $\boldsymbol{h_i}$ e . $\gamma(\boldsymbol{h_i}, \boldsymbol{\theta})$ é uma função válida de semi-variograma

Contudo, este método tem algumas limitações pois assume que as observações são independentes, enquanto que, os valores do semivariograma $\gamma(\boldsymbol{h}_i)$ pondem estar correlacionados.

Mínimos Quadrados Generalizados (MQG)

Visto que o método dos mínimos quadrados ordinários não leva em conta o facto de os valores do semivariograma experimental, bem como, a distribuição desses valores . A estimação pode ser melhorada usando o método dos mínimos quadrados generalizados.

Seja $\mathbf{V}$ a matriz de variância e covâriancia de

$$\hat{\gamma} = \Big(\hat{\gamma}(\boldsymbol{h}_1), \cdots, \hat{\gamma}(\boldsymbol{h}_k)\Big)'$$

Desse modo, o problema de minimização consiste em achar $\boldsymbol{\theta}$ de tal forma que minimize a expressão

$$\Big(\hat{\gamma} - \gamma(\boldsymbol{\theta})\Big)\mathbf{V}^{-1}\Big(\hat{\gamma} - \gamma(\boldsymbol{\theta})\Big)'$$

Mínimos Quadrados Ponderados (MQP)

O processo de estimação com base nos MQG é bastante iterativo e complicado. Cressie (1985) apresenta uma discussão detalhada sobre ajustamento do semivariograma usando MQP, e recomenda que a minimização de

$$\sum_{i=1}^{k} N(h_i)\left[\frac{\hat{\gamma}(h_i)}{\gamma(h_i;\theta)} - 1\right]^2 \tag{11}$$

como sendo uma boa aproximação dos MQP. Observe que quanto maior for o número de observações para uma certa distância h_i maior será o peso para atribuído as resídos para essa distância.

Ademais, Cressie(2015) aconselha a usar o variograma robusto no lugar do variograma experimental.

Exemplo

Ajustamento do semivariogram usando método dos MQO e MQP ao semivariograma experimental usando os dados s100 do pacote "geoR" para mais detalhes pode consultar https://rdrr.io/cran/geoR/man/variofit.html. Corra as linhas de código no programa R.

```
> vario100 <- variog(s100, max.dist=1)
> ini.vals <- expand.grid(seq(0,1,l=5), seq(0,1,l=5))
> mqo <- variofit(vario100, ini=ini.vals, fix.nug=TRUE,
 + wei="equal") # mínimos quadrados ordinários
> summary(mqo)
> mqp <- variofit(vario100, ini=ini.vals, fix.nug=TRUE)
> summary(mqp)
> plot(vario100)
> lines(mqp)
> lines(mqo, lty=2)
```

```
summary(ols)
$pmethod
[1] "OLS (ordinary least squares)"

$cov.model
[1] "matern"

$spatial.component
  sigmasq       phi
1.1070408 0.4006288

$spatial.component.extra
kappa
  0.5

$nugget.component
tausq
    0

$fix.nugget
[1] TRUE
```



```
$fix.kappa
```

Interpolação espacial

A interpolação é o processo de estimar valores em locais não amostrados com base nas observações vizinhas . A interpolação espacial assume que o atributo é continuo em toda a superfície. Ademais, a interpolação espacial assume que o atributo é espacialmente dependente, indicando maior similaridade para dados mais próximos e maior dessemelhança para valores mais afastados.

A interpolação espacial pode ser classificada em *determinística e estatística*. Os métodos de interpolação são vários. Dentre os métodos deterministicos pode se destacar a Interpolação Inversa da Distância ponderada bem como as suas variações (veja, Franke, 1982; Nader & Wein, 1998). Dentre os métodos geoestatísticos pode-se destacar Krigagem Simples, Krigagem Ordinária, Krigagem Universal Cokrigagem, etc.

Interpolação Inversa da Distância Ponderada (IDW)

A IDW é uma das técnicas de interpolação mais simples e popular. Este método é definido como sendo a média pondera de uma amostra de valores distribuídos no espaço numa certa região de pesquisa. A interpolação de um espaço/ponto não amostrado pode ser calculado com base em:

$$Z(\boldsymbol{x}_0) = \sum_{i=1}^{n} \lambda_i Z(\boldsymbol{x}_i) \tag{12}$$

onde, $\boldsymbol{x}_0$ é o ponto a ser estimado, λ_i representa o peso a ser atribuído a cada ponto amostrado. Os pesos são dados por:

$$\lambda_i = \frac{d_i^{-p}}{\sum_{i=1}^{n} d_i^{-p}} \tag{13}$$

d_i representa a distância euclidiana entre dois pontos.

Interpolação Inversa da Distância Ponderada (IDW)

Observe que a soma dos pesos λ_i é igual a unidade, isto é, $\sum_{i=1}^{n} \lambda_i = 1$.

O valor comummente usando para p é 2, daí o estimador passa a ser chamado *inverso do quadrado da distância*. Porém, p pode assumir qualquer valor. O método IDW é basicamente determinístico, o que significa que toda a estimação ou interpolação com base neste método não está associado a uma incerteza, isto é não se pode quantificar com que precisão se estima um determinado valor.

Babak & Deutsch (2009) apresentam um formalismo estatístico para este método permitindo que a variância do estimador na localização não amostrada x_0 seja dada por:

$$\sigma^2_{est} = E\big[Z(x) - Z(x_0)\big]^2 = \sigma^2 - 2\sum_{i=1}^{n} \lambda_i Cov\big(Z(x_i), Z(x_0)\big) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j Cov\big(Z(x_i), Z(x_j)\big) \tag{14}$$

Observações:

A medida que se aumenta o número de observações, mantendo fixo o expoente (p), as estimativas com base IDW tornam-se suaves.

Quanto maior for o valor do expoente (p) menos suave se apresentara a superfície estimada, e a medida que o expoente aumenta o número de observações deixa de ter um impacto relevante sobre a estimação de pontos não amostrados (Babak & Deutsch, 2009).

Além disso, quanto maior for o expoente p, maior será a variância das estimativas.

Importa referir que, o cálculo da variância só faz sentido se considerarmos que o processo em observância é estacionário, visto que, tradicionalmente este método é determinístico.

Krigagem

A *Krigagem* é uma técnica geoestatística para fazer interpolações espaciais. A *Krigagem* tem como objectivo interpolar valores referentes a uma função aleatória em locais não amostrados.

Esta técnica é bastante famosa, pois fornece a melhor previsão linear não-enviesada (MPLNE). Porém, a mesma apresenta uma limitação, pois exige que o processo seja estacionário da segunda ordem.

Tal como na *Interpolação Inversa da Distância Ponderada*, a *Krigagem* usa valores circundantes para interpolar locais não amostrados (ou estimar a média sobre um determinado bloco).

$$Z^*(x_0) = \sum_{=i}^{n} \lambda_i Z(x_i) \tag{15}$$

Embora a *Krigagem* seja MPLNE, importa salientar que a qualidade da interpolação depende de vários factores, nomeadamente: a qualidade dos dados; a localização/ disposição das observações (se os dados estiverem distribuídos de forma uniforme, haverá maior cobertura e muita informação sobre o processo, o que não acontece quando os dados apresentam-se em clusters);a distância entre os pontos amostrados e não amostrados.

É de salientar que a *Krigagem* apresentam alguma vantagem sobre outros métodos de interpolação, visto que, este considera a estrutura espacial do processo. Ademais, a *Krigagem* permite conhecer *a exactidão da interpolação* com base na variância do erro de previsão. A krigagem é um interpolador exacto, o que significa que para lugares amostrados, as interpolações coincidem com os valores observados, e dessa forma a variância do erro de previsão é zero.

Noção de Vizinhança

A *vizinhança* é um conceito de extrema importância em geoestatística. Como foi referido anteriormente, a *Krigagem* usa valores circunvizinhos para interpolar locais não amostrados. Contudo, no processo de interpolação tem-se duas opções

1. Todos os pontos amostrados são incluídos no processo de interpolação;
2. Apenas os pontos mais próximos do local a ser interpolado é que são considerados.

No primeiro caso, o impacto das observações distantes em relação ao ponto a ser interpolado é menor (até pode-se pensar que observações que se encontram além da amplitude do semivariograma terão um impacto nulo, porém isto não é verdade, pois estão espacialmente correlacionadas com o ponto a ser interpolado)(Montero & Mateu, 2015).

O uso de todas observações no processo de interpolação, que é um conceito de "vizinhança global" exige que o processo seja estacionário de segunda ordem or intrinsecamente estacionário sobre toda a vizinhança, o que na maioria das vezes não acontece. Além disso, esta abordagem tem consigo algumas desvantagens, visto que, quando o volume de dados é maior, o processo de estimação pode se tornar tedioso.

No segundo caso, apenas se exige a hipótese de quase estacionaridade de segunda ordem ou quase estacionaridade intrínseca. Porém, a dimensão bem como a configuração da vizinhança continuam a ser crucial. Visto que, se a vizinhança for muito menor, a precisão da interpolação será afectada e irá depender das observações inclusas na interpolação. Assim como se a vizinhança for muito maior, a hipótese de estacionaridade sobre toda a vizinhaca poderá ser questionável.

Não existe uma regra clara para definir a dimensão da vizinhança, contudo Webster & Oliver (2001) sugerem algumas directrizes:

1. Se os dados forem densos e o semivariograma tiver efeito pepita menor, então o raio da vizinhança pode ser igual a amplitude ou amplitude prática;
2. Se os o efeito pepita for maior, observações distantes do ponto a ser previsto terão um impacto significativo na interpolação e por essa razão devem ser incluídos na vizinhança.
3. Por outra, pode-se definir a vizinhança em termos de número mínimo e máximo de observações próximos do ponto a ser interpolado. Geralmente recomenda-se um mínimo de $n \approx 7$ e um máximo $n \approx 20$.

Krigagem Ordinária

Considere que a variável regionalizada em estudo tenha valores $z_i = z(x_i)$. Considere também que a variável regionalizada é estacinária de segunda ordem, com média:

$$\mathsf{E}\Big[Z(x)\Big] = \mu$$

que geralmente deverá ser estimada, e com uma covariância centrada

$$\mathsf{E}\Big[Z(x)Z(x+h)\Big] - \mu^2$$

e um variograma

$$\mathsf{E}\Big\{[Z(x+h) - Z(x)]^2\Big\} = 2\gamma(h)$$

O *interpolador de krigagem* será uma combinação linear dos valor circundantes do ponto a ser interpolado:

$$Z^*(x_0) = \sum_{=i}^{n} \lambda_i Z(x_i)$$

Para poder prever/interpolar o ponto não amostrados é preciso calcular os ponderados λ_i, de tal forma que :

1. O estimador seja não-enviesado.
2. A variância do erro de estimação (erro quadrático médio) seja mínima.

A primeira condição é satisfeita assegurando que a soma dos ponderadores seja igual a unidade, isto é, $\sum \lambda_i = 1$

$$\mathsf{E}[Z^*(x_0)] = \mu \sum \lambda_i = \mathsf{E}[Z(x)] \tag{16}$$

A segunda condição diz que a variância do erro de estimação deve ser miníma. Sob a condição de estacionaridade de segunada ordem, a variânica do erro de estimação é dada por:

$$\mathrm{V}\Big[Z^*(x_0) - Z(x_0)\Big] = \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C(x_i - x_j) - 2\sum_{i=1}^{n}\lambda_i C(x_i - x_0) + C(0) \quad (17)$$

O principal objectivo é achar os ponderadores λ_i que minimizam a equação (17) sob a condição de $\sum \lambda_i = 1$. Este é um problema que pode ser resolvido com base em multiplicadores de Lagrange, onde a função de lagrange é dada por

$$\phi(\lambda_i, \alpha) = \mathrm{V}\Big[Z^*(x_0) - Z(x_0)\Big] - \alpha\Big(\sum_{i=1}^{n}\lambda_i - 1\Big) \quad (18)$$

Para achar os ponderadores λ_i, acham-se as derivadas parciais em relação a λ_i e α, e iguala-se a zero.

As derivadas parciais conduzem-nos ao seguinte sistema de equações:

$$\begin{cases} \sum_{j=1}^{n} \lambda_j C(\boldsymbol{x}_i - \boldsymbol{x}_j) - \alpha = C(\boldsymbol{x}_i - \boldsymbol{x}_0), \quad i = 1, 2, ..., n \\ \sum_{i=1}^{n} \lambda_i = 1 \end{cases} \tag{19}$$

Substituindo $\sum_{j=i}^{n} \lambda_j C(x_i - x_j)$ por $C(x_i - x_0) + \alpha$ na expressão (17), a variância da estimativa resultante da krigagem ordinária é dada por:

$$\sigma^2_{OK}(\boldsymbol{x}_0) = C(\mathbf{0}) - \sum_{i=1}^{n} \lambda_i C(\boldsymbol{x}_i - \boldsymbol{x}_0) + \alpha \tag{20}$$

Com base na equação (20) pode se notar que a variância da estimativa resultante da krigagem sempre será menor que a variância do processo, visto que $C(0) = V(Z(x)) = \sigma^2$.

O sistema de equações (21) também pode ser expresso em função do semivariograma:

$$\begin{cases} \sum_{j=1}^{n} \lambda_j \gamma(\boldsymbol{x}_i - \boldsymbol{x}_j) + \alpha = \gamma(\boldsymbol{x}_i - \boldsymbol{x}_0), \quad i = 1, 2, ..., n \\ \sum_{i=1}^{n} \lambda_i = 1 \end{cases} \tag{21}$$

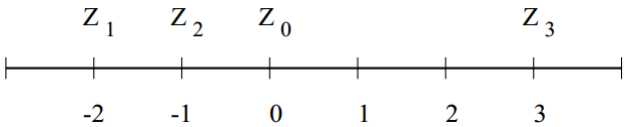
e a variância da estimativa é dada por :

$$\sigma^2_{KO}(x_0) = \sum_{i=1}^{n} \lambda_i \gamma(x_i - x_0) + \alpha \tag{22}$$

No caso em que o processo não é estacionário de segunda ordem, mas sim *intrinsecamente estacionário, o sistema de equações só pode ser expresso em função do semivariograma*.

Exemplo

Considere o seguinte caso:



onde se pretende estimar o valor de Z_0 com base $\{Z_1, Z_2, Z_3\}$. Considere também que a estrutura espacial é descrita por um semivariograma esférico:

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C_1\left(\frac{3}{2}\frac{h}{a} - \frac{1}{2}\frac{h^3}{a^3}\right) & 0 < h < a \\ C_0 + C_1 & a \leqslant h \end{cases} \tag{23}$$

Suponha que $a = 6,\ C_0 = 0,\ C_1 = 1$

h	0	1	2	3	4	5	6
$\gamma(h)$	0	0.2477	0.4815	6875	0.8519	0.9606	1

Para facilitar os cálculos basta representar o sistema de equações na forma matricial $AX = B$

$$\begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & 1 \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & 1 \\ \gamma_{31} & \gamma_{32} & \gamma_{33} & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \alpha \end{pmatrix} = \begin{pmatrix} \gamma_{10} \\ \gamma_{20} \\ \gamma_{30} \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 0.0000 & 0.2477 & 0.9606 & 1 \\ 0.2477 & 0.0000 & 0.8519 & 1 \\ 0.9606 & 0.8519 & 0.0000 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \alpha \end{pmatrix} = \begin{pmatrix} 0.4815 \\ 0.2477 \\ 0.6875 \\ 1 \end{pmatrix}$$

Para achar os pesos basta achar a inversa da matriz A, e desse modo teremos $X = A^{-1}B$

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \alpha \end{pmatrix} = \begin{pmatrix} -2.0658594 & 1.8973217 & 0.1685377 & 0.3681361 \\ 1.8973217 & -2.3294571 & 0.4321354 & 0.1618973 \\ 0.1685377 & 0.4321354 & -0.6006730 & 0.4699666 \\ 0.3681361 & 0.1618973 & 0.4699666 & -0.4915519 \end{pmatrix} \begin{pmatrix} 0.4815 \\ 0.2477 \\ 0.6875 \\ 1 \end{pmatrix}$$

$\lambda_1 = -0.04073892$, $\lambda_2 = 0.79554423$, $\lambda_3 = 0.24519469$ e $\alpha = 0.04890968$

Abaixo segue um pequeno código em R para o exemplo na página 62-64. Tente exercitar no programa R !!!

```
dist=matrix(data=c(0,1,5,2,1,0,4,1,5,4,0,3,2,1,3,0),nrow=4,ncol=4)
G=matrix(data=NA, nrow=4, ncol=4)  # matriz das semivariancias
G[4,]=c(rep(1,3),0)
G[,4]=c(rep(1,3),0)
c0=0 # efeito pepita
c1=1  # soleira parcial
a=6 # amplitude
for(i in 1:3){
  for(j in 1:3){
    if(i==j){G[i,j]=0}
            G[i,j]=c0+c1*(1.5*dist[i,j]/a-0.5*(dist[i,j]/a)^3)
  }
}

B=c(rep(NA,3),1)
for(i in 1:3){
  B[i]=c0+c1*(1.5*dist[i,4]/a-0.5*(dist[i,4]/a)^3)
}
w=solve(G)%*%B # ponderadores
```

Krigagem Simples

Por vezes a média do processo é conhecida (ou pode-se assumir) a partir de estudos ou experimentos passados. Desse modo, é sensato aproveitar esse conhecimento para melhorar as previsões. A *krigagem simples* não se difere tanto da krigagem ordinária, visto que, as interpolações também são uma combinação linear dos dados em disposição, porém o interpolador de krigagem incorpora a média (o conhecimento da média) no processo de interpolação/previsão.

Para a *krigagem simples* a equação de estimação é dada por :

$$Z^*(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) + \Big\{1 - \sum_{i=1}^{n} \lambda_i\Big\}\mu \tag{24}$$

Visto que não precisamos mais da restrição $\sum \lambda_i = 1$ para que (24) seja um estimador não-enviesado, o cálculo dos pesos λ será apenas com base na função de covariância. Além disso, com a ausência da restrição, não usamos mais o multiplicador de lagrange para calcular os pesos, e o sistema de equações para krigagem simples é dado por:

$$\sum_{j=1}^{n} \lambda_i C(x_i - x_j) = C(x_i - x_0) \qquad i = 1, 2, ..., n \tag{25}$$

A variância da estimativa para a kigagem simples é dada por :

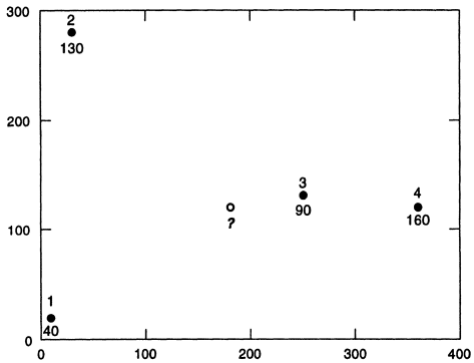
$$\sigma^2_{KS}(x_0) = C(0) - \sum_{i=1}^{n} \lambda_i C(x_i - x_0) \tag{26}$$

Caso assuma-se que o fenómeno em estudo é Gaussiano, pode-se muito bem associar a estimativa a um intervalo de confiança.

$$\Big[Z^*(x_0) - 1.96\sigma(x_0); Z^*(x_0) + 1.96\sigma(x_0)\Big] \tag{27}$$

Exemplo-Krigagem Simples

Considere a figura (lado esquerdo) e a tabela abaixo (lado direito)



índice	X	Y	Medições
1	10	20	40
2	30	280	130
3	250	130	90
4	360	120	160

Assuma que o atributo em estudo tem uma média de 110 e uma função de covariância $C(h) = 2000 \exp\Big(-\frac{h}{250}\Big)$. Se o ponto a ser estimado é $x_0 = (180, 120)$, responda as seguintes questões:

1. Calcule os pesos para a estimação do ponto não amostrado.
2. Que comentários pode fazer em relação aos pesos (ponderadores)?
3. Calcule o valor da estimativa no ponto não amostrado.
4. Calcule a variância da estimativa .

```
x=c(10,30,250,360,180)
y=c(20,280,130,120,120)
XY=as.matrix(cbind(x,y))
dist=matrix(data=NA, nrow=5,ncol=5)

for(i in 1:5){
  for(j in 1:5){
    dist[i,j]=sqrt((XY[i,1]-XY[j,1])^2+(XY[i,2]-XY[j,2])^2)
  }
}

Cov=matrix(data=NA, nrow=5,ncol=5)
for(i in 1:5){
  for(j in 1:5){
    Cov[i,j]=2000*exp(-dist[i,j]/250)
  }
}

Cov1=Cov[1:4,1:4]
Cov0=Cov[1:4,5]
Cov0=t(Cov0)
w=solve(Cov1)%*%t(Cov0) #ponderadores
```

1. Os pesos são:

```
        [,1] [,2] [,3] [,4]
[1,] 0.185 0.128 0.646 -0.001
```

2. (a) Em geral ,pode-se observar que as observações distantes em relação ao ponto não amostrado apresenta um peso menor em relação as observações mais próximas do ponto não amostrados. (b) Também pode-se observar que o menor peso é negativo. Importa salientar que na krigagem simples os valores dos pesos não estão limitados a um certo intervalo, isto é, os pesos podem assumir qualquer valor real. Isto faz com que a combinação linear do estimador da krigagem simples seja uma combinação não convexa. (c) A possibilidade de ter valores negativos faz com que as estimativas da krigagem simples não esteja confinadas ao intervalo dos dados.

Krigagem Universal

A Krigagem oridnária, assim como, a Krigagem simples assumem que a média do processo no campo aleatório é costante. Porém, na prática, campos ambientais e geológicos muitas vezes apresentam valores médios não constantes(a média do processo não é constante em todo espaço aleatório). Para lidar com este fenómeno, Matheron(1969) desenvolveu o método de *Krigagem Universal*.

A *Krigagem Universal* assume que a função aleatória é uma combinação linear de duas componentes. Sendo a primeira determinística, e uma função que depende da localização. A segunda componente é probabilística e estacionária de segunda ordem.

$$Z(\boldsymbol{x}) = \mu(\boldsymbol{x}) + \epsilon(\boldsymbol{x}) \tag{28}$$

onde, $\mu(\boldsymbol{x})$ é uma função que depende da localização $(\boldsymbol{x})$, e $\epsilon(\boldsymbol{x})$ é um processo estacionário de segunda ordem (com média zero).

A componente $\mu(\boldsymbol{x})$ caracteriza a tendência do processo (e designa-se por *drift*). Suponha que $\mu(\boldsymbol{x})$ pode ser representado como uma combinação linear de funções conhecidas $\Big\{f_l(x), l = 1, \cdots, k\Big\}$, com coeficientes desconhecidos $\{a_l\}$

$$\mu(\boldsymbol{x}) = \sum_{l=1}^{k} a_l f_l(x) \tag{29}$$

A média do processo bem como a covariância podem ser expressas da seguinte forma:

$$E\Big[Z(x)\Big] = \sum_{l=1}^{k} a_l f_l(x)$$

$$E\Big\{[Z(x_1) - \mu(x_1)][(Z(x_2) - \mu(x_2)]\Big\} = E\Big[\epsilon(x_1)\epsilon(x_2)\Big] = C(x_1 - x_2)$$

Tal como a Krigagem Ordinária, a *Krigagem Universal* também é um interpolador que resulta da combinação linear das observações circundantes ao local a ser estimado,

$$Z^*(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{30}$$

onde λ_i é escolhido de tal maneira que o estimador seja não enviesado e erro de estimação seja mínimo. O estiamdor será não enviesado, se e somente se, $E[Z^*(x_0)] = E[Z(x_0)]$, ou

$$\sum_{i=1}^{n} \lambda_i \mu(x_i) = \mu(x_0) \Rightarrow \sum_{l=0}^{k} a_l \sum_{i=1}^{n} \lambda_i f_l(x_i) = \sum_{l=0}^{k} a_l f_l(x_0)$$

O interpolador só será não-enviesado, se e somente se, $\sum_{i=1}^{n} \lambda_i f_l(x_i) = f_l(x_0)$.

A variância para este interpolador pode ser dada por:

$$\sigma^2_{KU}(x_0) = E\Big\{[Z^*(x_0) - Z(x_0)]^2\Big\} = \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C(x_i, x_j) - 2\sum_{i=1}^{n}\lambda_i C(x_i, x_0) + C(0) \quad (31)$$

Esta expressão deve ser minimizada sob a condição de não-enviesamento $\sum_{i=1}^{n}\lambda_i f_l(x_i)$ $f_l(x_0)$. isto pode ser feito usando multiplicador de lagrange, que irá resultar no seguinte sistema de equações:

$$\begin{cases} \sum_{j=1}^{n}\lambda_j C(\boldsymbol{x}_i - \boldsymbol{x}_j) - \sum_{l=0}^{k}\alpha_l f_l(x_i) = C(\boldsymbol{x}_i - \boldsymbol{x}_0), & i = 1, 2, ..., n \\ \sum_{i=1}^{n}\lambda_i f_l(x_i) = f_l(x_0), & l = 0, 1, 2, \cdots, k \end{cases} \quad (32)$$

Substituindo $\sum_{j=i}^{n} \lambda_j C(x_i - x_j)$ por $C(x_i - x_0) + \sum_{l=0}^{k} \alpha_l f_l(x_i)$ na expressão (31), a variância da estimativa resultante da *krigagem universal* é dada por:

$$\sigma^2_{KU}(\boldsymbol{x}_0) = C(\mathbf{0}) - \sum_{i=1}^{n} \lambda_i C(\boldsymbol{x}_i - \boldsymbol{x}_0) + \sum_{l=0}^{k} \alpha_l f_l(x_0) \tag{33}$$

A média do processo $\mu(x)$, geralmente, é modelizada por numa função polinomial na forma: $\mu(x) = a_0 + a_1 x + a_2 y + a_3 x^2 + a_4 xy + a_5 y^2$, onde x e y representam coordenadas.

Krigagem em Bloco

O interesse é estimar/predizer a média de um processo sobre um bloco/área V com base nas observações $Z(x_1)$, $Z(x_2)$,$\cdots$, $Z(x_n)$.

Uma das possíveis soluções é discretizar a área onde se pretende fazer a estimativa/previsão em vários pontos e depois calcular a média das estimativas pontuais individuais para obter a média sobre a área de interesse.

Este método é computacionalmente oneroso, pois teríamos de resolver vários sistemas de equações(isto porque, para cada ponto teríamos de aplicar a krigagem ordinária).

Krigagem em Bloco permite a estimação da média de uma variável aleatória sobre uma área/bloco resolvendo apenas um sistema de krigagem. Além disso, permite obter a variância para a krigagem.

Sob a suposição de estacionaridade de segunda ordem, a krigagem ordinária por bloco pode ser feita médiante a substituição do covariograma (semivariograma) para a distância entre x_i e x_0 pela média do covariograma $\bar{C}(x_i, V)$(semivariograma $\bar{\gamma}(x_i, V)$) na equação (21).

$$\begin{cases} \sum_{j=1}^{n} \lambda_j \gamma(\boldsymbol{x}_i - \boldsymbol{x}_j) + \alpha = \bar{\gamma}(\boldsymbol{x}_i, V), \quad i = 1, 2, ..., n \\ \sum_{i=1}^{n} \lambda_i = 1 \end{cases} \tag{34}$$

Com isto, temos que a variância para krigagem ordinária em bloco é dada por:

$$\sigma^2_{KOB} = \sum_{i=1}^{n} \lambda_i \bar{\gamma}(x_i, V) + \alpha - \bar{\gamma}(V, V) \tag{35}$$

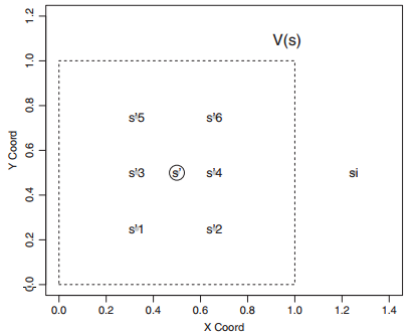
onde

$$\bar{\gamma}(x_i, V) \simeq \frac{1}{m} \sum_{j=1}^{n} \gamma(x_i - x'_j) \tag{36}$$

$$\bar{\gamma}(V, V) \simeq \frac{1}{m^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma(x'_i - x'_j) \tag{37}$$

Exemplo

Considere o bloco $V(s)$ que foi discretizado em seis pontos $s'_1, ..., s'_6$, considere também um ponto observado s_i. Assuma que o processo é caracterizado por um semivariograma esférico com soleira/patamar igual a 1 e amplitude igual 1.5.



Como pode se aproximar o valor de $\bar{\gamma}(V, V)$ e $\bar{\gamma}(x_i, V)$?

Distância entre o ponto si e os pontos sj'

(s_i, s'_j)	h	$\gamma(\mathbf{h})$
(s_i, s'_1)	0.950146	0.8230689
(s_i, s'_2)	0.634647	0.5967772
(s_i, s'_3)	0.916666	0.8025545
(s_i, s'_4)	0.583333	0.5539263
(s_i, s'_5)	0.950146	0.8230689
(s_i, s'_6)	0.634647	0.5967772

Distância entre os pontos no bloco

	s'_1	s'_1	s'_3	s'_4	s'_5	s'_6
s'_1	0					
s'_2	0.3333	0				
s'_3	0.2500	0.4167	0			
s'_4	0.4167	0.2500	0.3333	0		
s'_5	0.5000	0.3611	0.2500	0.4167	0	
s'_6	0.3611	0.5000	0.4167	0.2500	0.3333	0

Semivariograma pára a região discretizada

h	$\gamma(\mathbf{h})$	Number of pairs
0	0	6
0.2500	0.2476852	8
0.3333	0.3278147	6
0.4167	0.4059807	8
0.5000	0.4814815	4
0.6009	0.5687558	4

Validação Cruzada/ Cross-Validation

O processo de krigagem baseia-se, principalmente, na escolha de um modelo de semivariograma teórico e também no tipo de estacionaridade do processo.

Porém, é preciso avaliar a performance do modelo. Uma das formas, de avaliar a performance do modelo é dividir o conjunto de dados em duas partes, e usar uma das partes para ajustar o modelo de semivariograma teórico e a outra para para avaliar a performance do modelo de semivariograma escolhido para processo de krigagem , comparando as estimativas da krigagem e os valores observados.

Leave-one-out Cross Validation

O processo de Leave-one-out Cross Validation consiste em:

1. Obter as estimativas da krigagem $\hat{Z}(x_i)$ nos pontos amostrados $\boldsymbol{x}_i$, $i = 1, \cdots, n$ com base nas $n-1$ observações. E depois calcula-se as variância da estimativa da krigagem para cada ponto($\hat{\sigma}^2(\boldsymbol{x}_i)$)
2. Calcular as seguintes estatísticas de diagnóstico a partir dos resultados obtidos em (1):

- Erro médio de previsão(estimativa)

$$ME = \frac{1}{n}\sum_{i=1}^{n}(Z(\boldsymbol{x}_i) - \hat{Z}(\boldsymbol{x}_i)) \tag{38}$$

- Erro quadrático médio da estimativa

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Z(\boldsymbol{x}_i) - \hat{Z}(\boldsymbol{x}_i)\right)^2 \tag{39}$$

- Erro quadrático médio padronizado

$$MSDE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z(\boldsymbol{x}_i) - \hat{Z}(\boldsymbol{x}_i)}{\hat{\sigma}^2(\boldsymbol{x}_i)}\right)^2 \tag{40}$$

Se a escolha do semivariograma for bem sucedida:

- Espera-se que ME esteja próximo zero
- É desejável que MSE seja menor.
- MSDE deve ser aproximadamente igual a 1.
- Além disso, o coeficiente de correlação entre os valores previsto e observado deve ser próximo de 1.

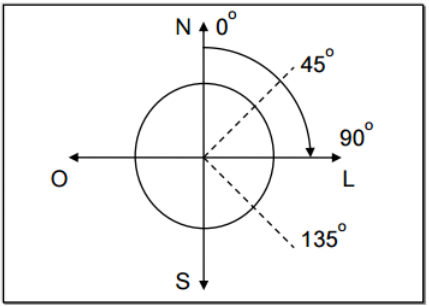
Cressie(1993) afirma que se o modelo de semivariograma tiver sido validado com sucesso, pode-se ter certeza de que a previsão baseada nesse semivariograma é aproximadamente centrada(não enviesada) e que o erro quadrático médio da estiativa é aproximadamente certo. CV não pode provar que o modelo ajustado está correto, apenas que não é grosseiramente incorreto (não há razão para rejeitá-lo). Assim, deve se deixar claro que o sucesso do CV não garante que o semivariograma escolhido (ajustado) seja correto; O que ele garante é que não é incorreto.

Anisotropia

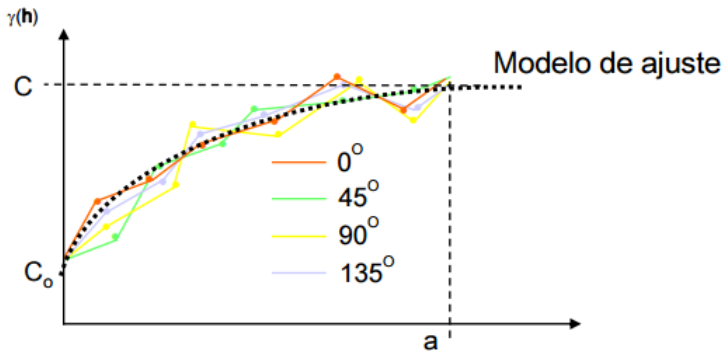
Um modelo de semivariograma teórico $\gamma(\boldsymbol{h})$ é definido como anisotrópico quando depende da direcção do vector $\boldsymbol{h}$. Caso contrário é considerado como sendo isotrópico. A isotropia implica que os semivariogramas direccionais coincidem e, portanto, o semivariograma "global" é omnidirecional .

A anisotropia pode ser identificada com base no semivariograma experimental, calculado para várias direcções. A análise da anisotropia objetiva detectar as direções de maior e menor continuidade espacial do fenómeno investigado

Convenções direcionais usadas na geoestatística

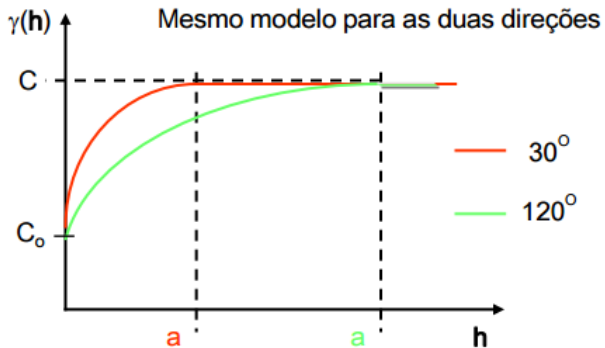


Considere os semivariogramas ilustrados na figura abaixo. Desta figura pode se verificar que a distribuição espacial do fenomeno é isotrópica. Neste cenário, um único modelo é suficiente para descrever a variabilidade espacial do fenómeno em estudo



Anisotropia geométrica

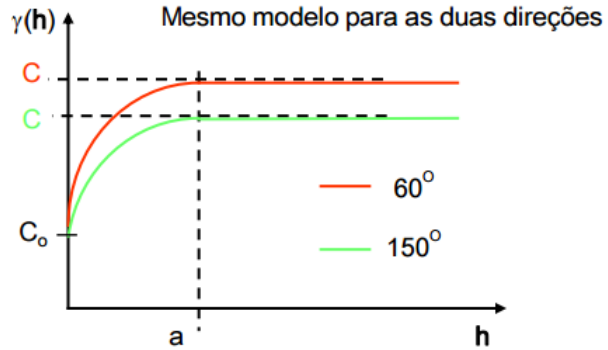
Neste caso, os semivariogramas apresentam o mesmo patamar (C) com diferentes amplitudes (a) para o mesmo modelo



Anisotropia zonal

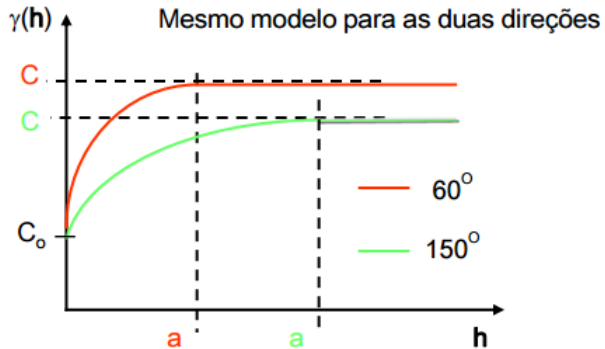
Neste caso, os semivariogramas apresentam diferentes patamares (C) com mesma amplitude (a) para o mesmo modelo.

Assim como a isotropia, a **anisotropia zonal** é um caso menos frequente nos fenómenos naturais.

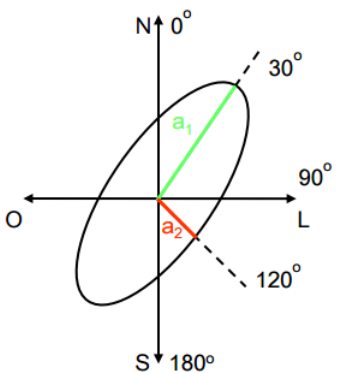


Anisotropia combinada (geométrica + zonal)

Neste caso, os semivariogramas apresentam diferentes patamares (C) e diferentes amplitudes (a) para o mesmo modelo. Pode apresentar também diferentes efeitos pepita.



Uma forma prática de visualizar e calcular os parâmetros (factor e ângulo) da anisotropia é através do esboço gráfico de uma elipse (ou diagrama de rosa)



Ângulo de anisotropia é o ângulo que o eixo das ordenadas forma com o eixo e maior amplitude. Para a figura ao lado o ângulo é 30°

O factor de anisotropia será o quociente entre a maior e menor amplitude a_1/a_2. Desta forma, o factor de anisotropia sempre será maior ou igual a unidade. Para o caso onde o processo é isotrópico, o factor de anisotropia é igual a 1.

A anisotropia (geométrica) pode-se corrigir subistituíndo as coordenadas elípticas pelas coordenadas correspondentes no circulo, isto é, temos de transformar a elipse em um circulo, de tal forma que amplitude do semivariograma nas várias direcções seja igual a 1. Para tal teremos:

$$\gamma_{a=1}(h) = \gamma_x(h_x) \quad \text{com} \quad h = h_x/a_x$$

$$\gamma_{a=1}(h) = \gamma_y(h_y) \quad \text{com} \quad h = h_y/a_y$$

$$\gamma_{a=1}(h) = \gamma_z(h_z) \quad \text{com} \quad h = h_z/a_z$$

Isto corresponde, no espaco cartesiano, a as distânicas serem "isotropizadas" do seguinte modo.

$$h = \sqrt{\left(\frac{h_x}{a_x}\right)^2 + \left(\frac{h_y}{a_y}\right)^2 + \left(\frac{h_z}{a_z}\right)^2} \tag{41}$$

Se optarmos por transformar a anisotropia num semivariograma de referência (por exemplo, o de maior amplitude em vez do semivariograma com amplitude igual a 1), então a distância h "isotropizada" fica igual a :

$$h = \sqrt{h_x \left(\frac{a_x}{a_x}\right)^2 + h_y \left(\frac{a_x}{a_y}\right)^2 + h_z \left(\frac{a_x}{a_z}\right)^2} \tag{42}$$

onde a_x é a amplitude se semivariograma de referência e $r_x = a_x/a_x$, $r_y = a_x/a_y$, $r_z = a_x/a_z$ são os factores de anisotropia nos 3 eixos principais.

Este método de transformação geométrica de coordenadas é aplicável igualmente a modelos não estacionários que não atingem o patamar, por exemplo o modelo linear.

Se a direcção de maior amplitude não coincidir com os eixos do sistema de coordenadas, então é necessário, em primeiro lugar, uma operação de rotação para que o eixo xx' coincida com a_x, yy' com a_y e zz' com a'_z, antes da transformação geométrica das distâncias. Para o caso de duas dimensões a rotação é dada por:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{43}$$

Depois de fazer a rotação é só multiplicar as coordenadas y' por factor de anisotropia $r = a_2/a_1$ caso a amplitude de referência seja a maior amplitude.

Ex: Suponha que o semivariograma na direcção Norte-Sul e Oeste-Este($\theta = \pi/2$) seja dado por

$$\gamma_{\alpha_{O-E}}(h_{\alpha_{O-E}}) = 5\left(1.5\frac{h_{\alpha_{O-E}}}{20} - 0.5\frac{h^3_{\alpha_{O-E}}}{20^3}\right) \tag{44}$$

A rotação dos eixos não é preciso e o quociente de anisotropia é $\lambda = \frac{a_{O-E}}{a_{N-S}} = 2$, $\gamma_{\alpha_{N-S}}$ pode ser expresso em termos de $\gamma_{\alpha_{O-E}}$ multiplicando $h_{\alpha_{N-S}}$ pelo quociente de anisotropia.

$$\gamma_{\alpha_{N-S}}(h_{\alpha_{N-S}}) = 5\left(1.5\frac{\frac{20}{10}h_{\alpha_{N-S}}}{20} - 0.5\frac{\frac{20^3}{10^3}h^3_{\alpha_{N-S}}}{20^3}\right) \tag{45}$$

- Portanto, fazendo a mudança das coordenadas, tem-se $\begin{cases} x^* = x \\ y^* = \lambda y \end{cases}$, isto é $|\boldsymbol{h}^*| = \sqrt{{h_1}^{*2} + {h_2}^{*2}}$
- Entao, o variograma esférico $\gamma_(|\boldsymbol{h}^*|) = 5\left(1.5\frac{|\boldsymbol{h}^*|}{20} - 0.5\frac{|\boldsymbol{h}^*|^3}{20^3}\right)$ pode ser usado para estudar a variabilidade nas duas direcções.

: Observação:

- A anisotropia zonal não pode ser corrigida por meio de uma trasnformação linear das coordenadas.
- O metódo mais comum de lidar com a anisotropia zonal é : Se estivermos diante um espaço bidimensional, e se por exemplo, a soleira alongo do eixo y for maior do que na direcção de x, primeiro ajustamos um modelo isotrópico ao semivariograma empírico considerando a direcção de x e depois acrescentamos o semivariograma com anisotropia geométrica.

Referências

1. Cressie, N. (2015). Statistics for spatial data. John Wiley & Sons.
2. Franke R (1982) Scattered data interpolation: tests of some methods. Math Comput 38: 181-200
3. Hengl, T. (2009). *A practical guide to geostatistical mapping of Environmental Variables* (Vol. 52). Hengl.
4. Lloyd, C. D., & Atkinson, P. M. (2004). Archaeology and geostatistics. Journal of Archaeological Science, 31(2), 151-165.
5. Montero, J. M., & Mateu, J. (2015). Spatial and spatio-temporal geostatistical modeling and kriging (Vol. 998). John Wiley & Sons.
6. Nader IA, Wein RW (1998) Spatial interpolation of climatic normals: test of a new method in the canadian boreal
7. STEPPE (2016, Dezembro 22)https://steppe.org/why-geostatistics/.

Referências

8. Soares, A. (2104) *Geoestatística para as Ciências da Terra e do Ambiente* (3a ed), Lisboa: IST Press
9. Yost, R. S., Uehara, G., & Fox, R. L. (1982). *Geostatistical analysis of soil chemical properties of large land areas. I. Semi-variograms*. Soil Science Society of America Journal, 46(5), 1028-1032.
10. Wackernagel, H. (1995) *Multivariate Geostatistics: An Introduction with Applications.* Springer-Verlag, Berlin.
11. Webster, R. & Oliver, M.A. (2001) *Geostatistics for Environmental Scientists*. John Wiley & Sons, Ltd., Chichester.