Universidade do Minho

Departamento de Matemática

Estatística Espacial

Geoestatística

2023-2024

Docente: Raquel Menezes
E-mail: rmenezes@math.uminho.pt

Tópicos principais

- Análise exploratória de dados espaciais;
- Estacionariedade e isotropia;
- Estimação e predição espacial;
- Modelos geoestatísticos Gaussianos;
- Exemplos prácticos.

Bibliografia

- Diggle P. and Ribeiro P. *Model-based Geostatistics*. Springer Series in Statistics, 2007.
- Cressie, N.A.C. *Statistics for Spatial Data*. Wiley, New York, 1993.
- Soares, A. *Geoestatística para Ciências da Terra e do Ambiente*. Ensino da Ciência e Tecnologia 9, IST Press, 2000.

Nota: Esta sebenta é essencialmente baseada no livro *Model-based Geostatistics* de Diggle e Ribeiro (2007).

Conteúdo

Capítulo 1

Introdução

Na actualidade, a estatística espacial ocupa um lugar importante, graças ao desenvolvimento tecnológico que permite a fácil obtenção de dados espaciais. Para além da existência das fontes tradicionais de dados espaciais, tal como mapas, material recolhido nos censos, fotos aéreas, surgem agora novas fontes com excelente fiabilidade, nomeadamente com os dados obtidos por satélite. Estas novas tecnologias encontram-se associadas à disponibilidade de uma maior capacidade computacional e software específico, tal como software de processamento de imagem e sistemas de informação geográfica.

Os **modelos espaciais** trabalham com dados recolhidos em distintas localizações espaciais. Estes modelos **medem a relação entre observações obtidas em diversos locais**, devendo representar a noção intuitiva da existência de uma correlação entre dados situados na proximidade espacial. Tipicamente, esta correlação diminui com o aumento da distância. Estes modelos devem igualmente reflectir a **presença de erros espacialmente correlacionados**. A análise de correlação espacial permite então observar a forma como variáveis, tal como algum indicador de poluição, obtidos em pontos distintos do espaço se relacionam. O conhecimento destas relações é particularmente relevante de um ponto de vista prático, uma vez que permite estimar valores para as localizações nas quais não foram efectuadas medições.

Uma amostra de dados pode ser formada com base em observações obtidas em uma, duas ou três dimensões. Medições da qualidade da água efectuadas ao longo de um percurso fluvial são unidimensionais. Por outro lado, medições de pluviosidade e outras variáveis meteorológicas são obtidas em pontos particulares, mas colectivamente formam um campo aleatório bidimensional. Por último, medições de concentrações de minerais no solo são por vezes tratadas como problemas tridimensionais dado que podem ocorrer em diversas profundidades.

Um **processo espacial é um processo estocástico**. Pode ser representado por um conjunto de variáveis aleatórias (ou vectores) $Y(\mathbf{x})$, indexado sobre $\mathbf{x}$ definido num conjunto $A \subset I\!R^d$, um espaço euclidiano d-dimensional com $d = 1, 2, 3$. A sua notação usual é

$\{Y(\mathbf{x}) : \mathbf{x} \in A\}$. Considerem-se as localizações espaciais $\mathbf{x_1}, ..., \mathbf{x_n}$, então $Y(\mathbf{x_1}), ..., Y(\mathbf{x_n})$ identificam dados observados nestas localizações. Estas observações podem ser obtidas a partir de uma, ou mais, variáveis discretas ou contínuas.

De acordo com Cressie (1993), a natureza de A permite identificar três processos espaciais principais, nomeadamente processos reticulados, processos pontuais e processos contínuos. Estes últimos são geralmente referidos como geoestatística (Matheron, 1963). Nos processos contínuos, ao contrário dos processos reticulados, entre quaisquer dois pontos espaciais associados a observações existentes, existe sempre um outro ponto onde a variável aleatória poderia ser observada.

A geoestatística refere-se, então, ao ramo da estatística espacial para o qual os dados consistem numa amostra finita de valores relacionados com um fenómeno espacialmente contínuo. Exemplos incluem: altitudes acima do nível do mar num levantamento topográfico; medições de poluição à custa de uma rede finita de estações de monitorização; caracterização de propriedades do solo à custa uma amostra representativa; e contagens de insectos num conjunto de localizações pré-seleccionadas.

Capítulo 2

Modelos Geoestatísticos Gaussianos

Na Geoestatística estudam-se fenómenos geo-referenciados definidos de um modo contínuo numa determinada região de interesse. Por conseguinte, os modelos geoestatísticos são considerados modelos espacialmente contínuos que se pretendem ver ajustados a dados espacialmente discretos (Chilès and Delfiner, 1999). Os dados seguem as seguintes características:

- Os valores $Y_i : i = 1 \ldots, n$ são observados num conjunto discreto de localizações amostrais $\mathbf{x_i}$ dentro de alguma região espacial A;
- Cada valor observado Y_i é uma medição directa, ou está estocasticamente relacionada, com um valor de um fenómeno subjacente espacialmente contínuo, $S(\mathbf{x})$, na correspondente localização amostral $\mathbf{x_i}$;
- Os diversos valores Y_i **são identicamente distribuídos mas não independentes**, uma vez que poderão estar espacialmente correlacionados; não são independentes, são id e não iid
- Por último, no caso particular dos modelos Gaussianos, os valores observados Y_i são realizações de uma distribuição normal (ou aproximadamente normal).

Neste capítulo iremos apresentar alguns modelos para dados Gaussianos. Na próxima secção, apresentam-se alguns exemplos de motivação.

2.1 Exemplos motivadores

De seguida, descrevem-se alguns exemplos de dados que poderão recorrer a modelos geoestatísticos Gaussianos. As respectivas bases de dados encontram-se disponíveis nas packages `geoR` e `geoRglm` do R, desenvolvidas por Ribeiro Jr. e Diggle (2001). Estes exemplos enquadram-se dentro da geoestatística, pois, em primeiro lugar, os valores $\{Y_1, \ldots, Y_n\}$ são observados num conjunto discreto de locais $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ dentro de uma região espacial A,

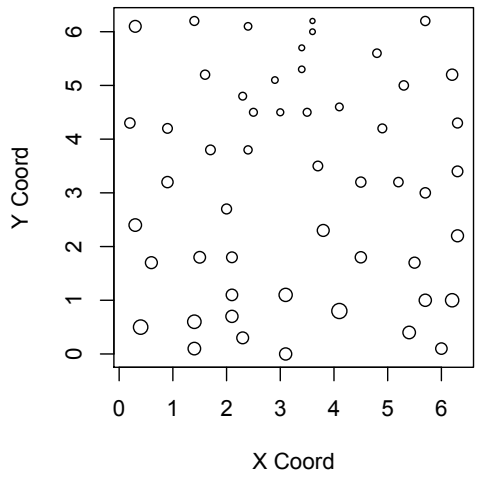

Figura 2.1: Exemplo: elevações de uma superfície. Para as coordenadas a distância unitária é especificada como 50 pés. As elevações observadas variam entre 690 a 960 unidades, onde 1 unidade representa 10 pés de elevação. Os círculos são desenhados com centros nas localizações amostradas e raio proporcional ao valor da elevação amostrada.

e em segundo lugar, cada valor observado de Y_i ou é uma medida directa de, ou estocasticamente relacionada com, um fenómeno contínuo $S(\mathbf{x})$ espacialmente referenciado no local amostrado correspondente.

Exemplo 1: Elevações de uma superfície

Os dados para este exemplo dão-nos as elevações de superfície y_i medidas em 52 locais $\mathbf{x_i}$ (ver detalhes na Figura 2.1). A região de observação A é um quadrado de aproximadamente 300 por 300 pés. A variável que está a ser estudada é a variável Y que identifica a elevação. A análise deste tipo de dados tem por objectivo construir um mapa de elevação contínuo, para toda a região do quadrado A. Uma vez que a elevação territorial pode ser medida com uma margem de erro negligível, neste exemplo, cada y_i é aproximadamente igual a $S(\mathbf{x_i})$.

Na Figura 2.2, apresentam-se alguns gráficos com os principais resultados de uma análise preliminar para os dados de elevação. Pela observação destes gráficos, mais especificamente o painel superior direita, pode-se afirmar que existe tendência espacial, uma vez que à medida que um observador se desloca de norte para sul, as elevações vão aumentando. O gráfico apresentado no painel inferior direito parece apontar para a eventual gaussianidade dos dados.

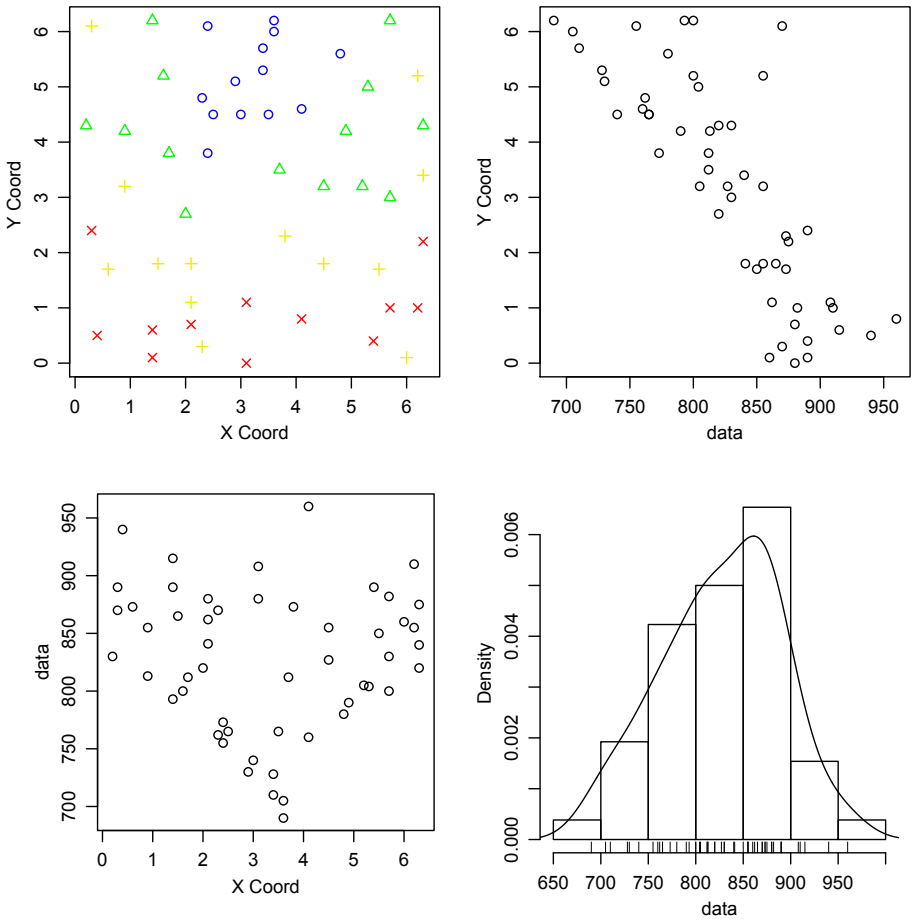

Figura 2.2: Uma breve análise exploratória dos dados "elevações de uma superfície". O painel superior esquerdo apresenta os 4 quartis associados às elevações observadas (Q1-azul; Q2-verde; Q3- amarelo; e Q4-vermelho). O painel inferior direita apresenta um histograma para os valores de elevação amostrados. Os restantes painéis representam a relação entre as elevações observadas e uma das coordenadas.

Exemplo 2: Contaminação resultante de testes de armas nucleares

Os dados para este exemplo foram recolhidos na ilha Rongelap no pacífico sul em 157 locais (ver painel superior esquerdo da Figura 2.3). Os dados têm o formato $(\mathbf{x_i}, y_i, t_i) : i = 1, \ldots, 157$, onde $\mathbf{x_i}$ identifica a localização espacial, y_i a contagem de partículas de radioactividade na localização $\mathbf{x_i}$, e t_i o período de tempo em segundos ao longo do qual y_i foi acumulado (note-se que este período de tempo pode ser diferente nas diferentes localizações). Os dados recolhidos fazem parte de uma investigação mais abrangente e multidisciplinar sobre contaminação residual devido a um programa americano sobre testes de armas nucleares, que originou muita precipitação radioactiva ao longo de toda a ilha nos anos 50. Um dos objectivos científicos da análise dos dados de Rongelap é obter um mapa da contaminação $S(\mathbf{x})$ estimada. Contudo, ao contrário do exemplo anterior, este mapa deverá interpolar os rácios y_i/t_i observados.

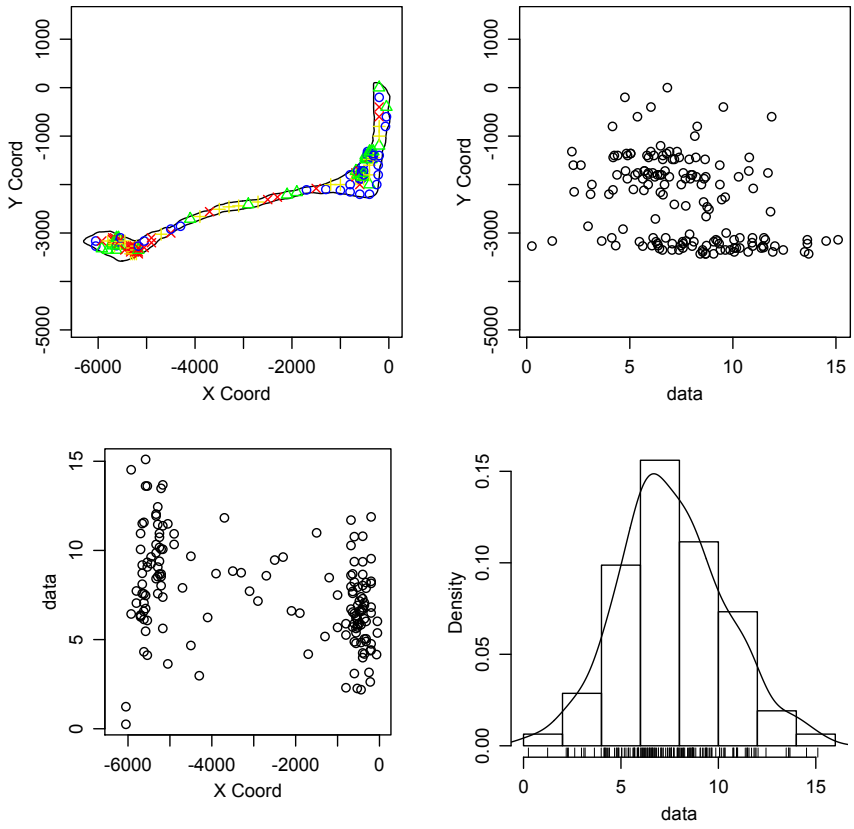

Figura 2.3: Uma breve análise exploratória dos dados "contaminação resultante de testes de armas nucleares". O painel superior esquerdo apresenta os 4 quartis associados aos rácios observadas y_i/t_i. O painel inferior direita apresenta um histograma para estes valores amostrados. Os restantes painéis representam a relação entre os valores de contaminação observados e uma das coordenadas.

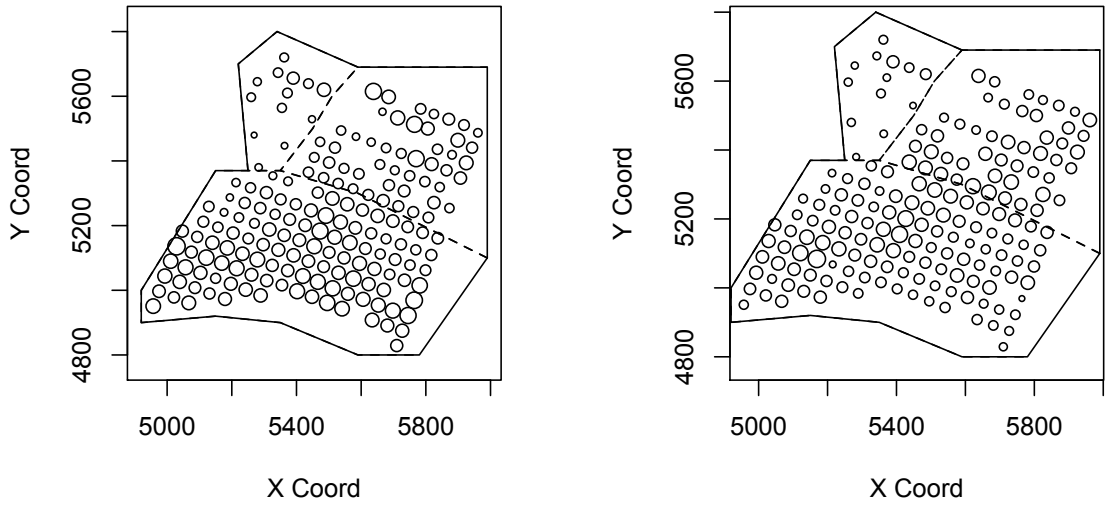

Figura 2.4: Exemplo: dados de solo. Os círculos representam os valores de cálcio (painel da esquerda) e valores de magnésio (painel da direita) com as linhas a tracejado delimitando sub-regiões com diferentes prácticas de manutenção do solo.

Exemplo 3: Dados do solo

Estes dados têm o formato $(\mathbf{x_i}, y_{i1}, y_{i2}, d_{i1}, d_{i2})$ onde $i = 1, \ldots, 178$, onde $\mathbf{x_i}$ identifica a localização da amostra de solo recolhida; y_{i1} e y_{i2} são duas variáveis resposta, identificando o conteúdo de cálcio e magnésio em cada localização; e d_{i1} e d_{i2} são duas covariáveis, identificando a elevação em $\mathbf{x_i}$ e o respectivo código da sub-região.

As amostras de solo foram recolhidas numa profundidade de 0 a 20cm em cada uma das 178 localizações (ver Figura 2.4). A região em estudo foi dividida em 3 sub-regiões, de acordo com o tipo de solo. A primeira, no canto superior esquerdo, está geralmente inundada nas estações das chuvas. A segunda corresponde à parte inferior da região em estudo e, a terceira ao canto superior direito, ambas receberam fertilizantes no passado. A segunda está geralmente ocupada com campos de arroz e a terceira é geralmente utilizada como área experimental. Também a segunda sub-região foi a última das três a receber cálcio para neutralizar os efeitos do alumínio no solo, o que em parte explica os elevados níveis de cálcio medidos nessa sub-região. Por último, faz-se notar que de acordo com os valores obtidos para o conteúdo de cálcio e magnésio, poderemos considerar estas duas variáveis como Gaussianas.

Um dos objectivos pretendidos com a análise destes dados é construir mapas sobre as variações espaciais do conteúdo de cálcio ou de magnésio. Outro objectivo é o de investigar a

relação entre os conteúdos de cálcio e de magnésio e as duas covariáveis d_{i1} e d_{i2}.

2.2 Terminologia e objectivos

Os dados geoestatísticos univariados podem ser representados por $\{(\mathbf{x_i}, y_i) : i = 1, \ldots, n\}$ onde:

- $\mathbf{x_i}$ identifica a localização espacial (geralmente um espaço bi-dimensional, embora possam também ocorrer exemplos a um dimensão ou a três dimensões);
- y_i é uma variável de medição ou resposta associada à localização $\mathbf{x_i}$.

Note-se que, tipicamente, se assume que as localizações $\mathbf{x_i}$ são determinísticas (por exemplo, formando uma grelha sobre a região de estudo) ou estocasticamente independentes do processo que gera as medidas y_i. Cada y_i é uma realização de uma variável aleatória Y_i, cuja distribuição depende do valor na localização $\mathbf{x_i}$ de um processo estocástico espacial contínuo $S(\mathbf{x})$, que não é directamente observável. Em casos particulares, tal como apresentado no Exemplo 1, poderemos assumir $Y_i = S(\mathbf{x_i})$, mas de um modo geral é importante preservar a diferença entre a quantidade observável Y_i e a não observável $S(\mathbf{x})$ (trata-se de um processo latente). No Exemplo 2, essa distinção é feita: Y_i- contagem de partículas de radioactividade; e $S(\mathbf{x})-$ mapa de índice de radioactividade.

Respostas multivariadas e variáveis explicativas

A diferença entre respostas multivariadas e variáveis explicativas nem sempre é clara. O modelo da resposta multivariada requer a especificação do processo estocástico sobre o formato de um vector sobre a região em estudo A, enquanto que as variáveis explicativas são tratadas como quantidades determinísticas sem nenhum modelo estocástico associado. Uma variável explicativa deverá estar disponível em qualquer local dentro de A, pois pode ser utilizada para estimar respostas em locais x não amostrados. Esta situação ocorre no Exemplo 3, onde as covariáveis são d_{i1} e d_{i2}, e a resposta poderá ser considerada bivariada, caso se modele conjuntamente y_{i1} e y_{i2}.

Os modelos geoestatísticos multivariados são relevantes quando duas ou mais variáveis resposta são medidas em localizações espaciais dentro de uma região espacial contínua. Tal situação pode surgir quando as variáveis têm igual interesse científico e se pretende descrever as suas distribuições espaciais e respectiva ligação, ou quando se pretende descrever a distribuição condicional da variável resposta de maior interesse.

Desenho amostral

Os locais $\mathbf{x_i}$ onde são feitas as medições, são colectivamente chamados de "desenho amostral" para os dados. Um desenho é **não-uniforme** se a intensidade da amostragem variar sistematicamente ao longo da região de estudo. Tal ocorrerá se em algumas regiões forem deliberadamente recolhidas amostras com mais intensidade do que outras. Um desenho é uniforme se todos os pontos na região de estudo tiverem igual probabilidade de serem amostrados.

Um desenho é **não-preferencial** se for determinístico (exemplo de uma grelha regular) ou se for estocasticamente independente de $S(.)$. Métodos geoestatísticos convencionais assumem, eventualmente apenas de forma implícita, que o desenho amostral é não-preferencial, permitindo que a análise de dados prossiga condicional ao desenho amostral. Se o processo de amostragem é não-preferencial, então a escolha do desenho não deve afectar o modelo assumido para os dados, mas pode afectar a precisão das inferências feitas à custa dos dados.

Objectivos científicos

Na maioria das aplicações, os principais objectivos científicos da análise geoestatística são estimação e predição.

Estimação refere-se a inferências sobre parâmetros do modelo estocástico obtidas à custa dos dados. Tais parâmetros podem incluir parâmetros de interesse científico directo, como um coefeciente de regressão que relaciona uma variável resposta com uma explicativa; ou parâmetros de interesse indirecto, como aqueles que definem a estrutura de covariância do modelo para $S(.)$.

Predição refere-se a inferências sobre a realização do processo latente não observado $S(.)$ ou de $Y(.)$ em localizações não observadas. Nas aplicações, os objectivos específicos das predições podem incluir uma predição do valor realizado de $S(\mathbf{x})$ numa localização arbitrária x dentro de uma região de interesse A, tipicamente apresentado como mapa de valores previstos para $S(\mathbf{x})$. Alternativamente, pode-se prever alguma propriedade da realização completa de $S(.)$, o que é de particular relevância para alguns problemas. Por exemplo, identificar sub-regiões em A, associadas aos valores máximos de $S(\mathbf{x})$.

Um terceiro tipo de problema, denominado por **testes de hipóteses** pode também surgir nos problemas geoestatísticos; por exemplo, para decidir se queremos ou não incluir uma dada variável explicativa (covariável).

2.3 Análise exploratória de dados

A análise exploratória de dados é uma práctica comum na estatística moderna, e a geoestatística não é excepção. Em geoestatística, esta análise é naturalmente orientada para uma investigação inicial dos aspectos espaciais dos dados, que sejam relevantes para validar as assunções exigidas pelo modelo candidato a ser adoptado. No entanto, aspectos não-espaciais podem e devem também ser investigados.

Análise exploratória não espacial

A análise não espacial deve incluir **técnicas clássicas da variável resposta** Y_i, nomeadamente: cálculo de medidas de localização e de dispersão; análise da normalidade; análise da presença de outliers; e construção de gráficos como histogramas, *boxplots* e diagrama de caule-e-folhas. Vejamos alguns exemplos.

Para o Exemplo 1, nos 52 locais registam-se elevações que variam entre 6900 a 9600 pés, com uma média de 8271, mediana de 8300, e desvio padrão de 620. O histograma para as 52 elevações indica uma leve assimetria e não sugere outliers (Figura.2.2, painel inferior direito). Tal sugere que se pode usar um modelo Gaussiano para aproximar esses dados.

Uma parte importante da análise exploratória de dados é analisar a relação entre a resposta e as possíveis covariáveis. Para o Exemplo 3, a estimação do coefeciente de regressão entre o cálcio e a elevação aponta para uma relação positiva estatisticamente significativa entre as duas variáveis.

Análise exploratória espacial

Na primeira fase da análise exploratória de dados, faz-se o gráfico dos valores observados da variável resposta na região de observações; por exemplo, gráficos análogos aos vários apresentados nas Figuras 2.1 – 2.4. Estes gráficos podem revelar presença ou ausência de uma **tendência (média do processo) não constante e dependente das localizações**; ou outliers (respostas discordantes com as respostas do vizinho espacial).

Se os dados estão localizados numa grelha regular, sugere-se que se calcule a média e a mediana dos dados, por linhas e por colunas. De seguida, obtem-se o gráfico que resume esta informação, permitindo identificar presença ou ausência de tendência. Este tipo de análise indica-nos se, e de que forma, a localização espacial tem influência nos valores das variáveis. Ver exemplo da Figura 2.4 de Cressie (1993).

Uma abordagem, também interessante, é a de desenhar um gráfico bivariado e $Y(\mathbf{x})$ e $Y(\mathbf{x}+\mathbf{u})$, para uma direcção fixa u, enquanto que x varia sobre os dados das localizações. Os

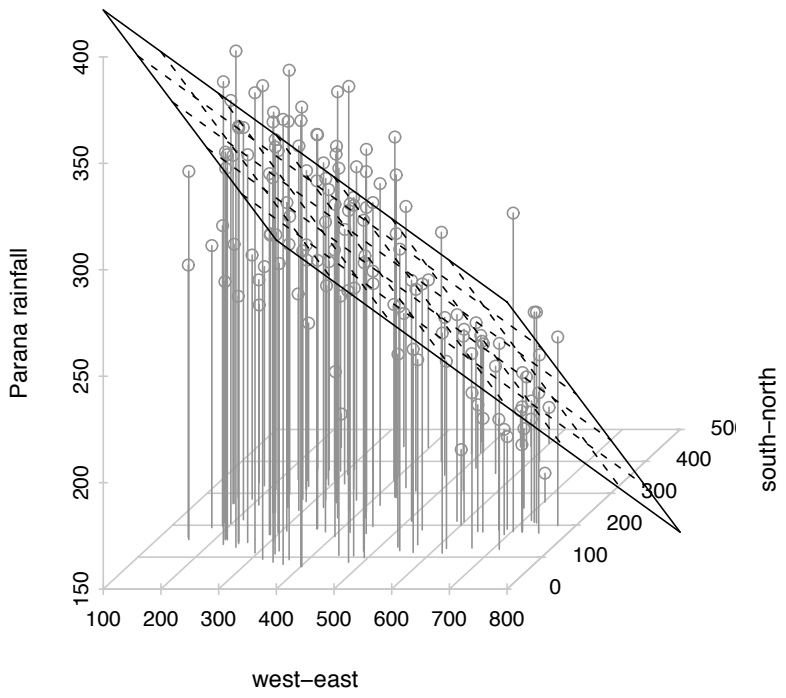

Figura 2.5: Exemplo: pluviosidade no Estado do Paraná no Brasil. Cada círculo representa um valor observado de pluviosidade nas respectivas coordenadas.

pontos isolados e afastados da bissectriz são outliers. Ver exemplo da Figura 2.5 e 2.6 de Cressie (1993).

O gráfico da variável resposta em ordem a cada uma das coordenadas espaciais pode revelar uma maior clareza na tendência (ver Figuras 2.2 e 2.3 painel superior direito e painel inferior esquerdo). Adicionalmente, a análise de uma tendência não constante pode ser feita através de um gráfico que relaciona a variável reposta simultaneamente com as duas coordenada (se $A \subset I\!R^2$). Por exemplo, a análise dos dados da pluviosidade no Estado do Paraná no Brasil (ver Figura 2.5), disponíveis no `geoR`, aponta para uma tendência clara de este para oeste e de sul para norte que, conforme iremos ver posteriormente, poderá ser modelada por um polinómio de 1.grau.

Os gráficos anteriores são úteis para analisar se a média do processo associado aos dados é ou não constante. No entanto, não é fácil através destes gráficos **analisar a existência de correlação espacial**. Para a análise da correlação espacial, devemos recorrer a gráficos específicos que relacionam a associação (ou disassociação) entre duas variáveis Y_i e Y_j e

a distância entre os respectivos pontos $\mathbf{x_i}$ e $\mathbf{x_j}$. Por exemplo, o gráfico do **variograma empírico** poderá ser bastante útil para este objectivo. Este assunto irá ser cuidadosamente discutido na Secção 2.5.

2.4 Apresentação de um modelo geoestatístico básico

O objectivo desta secção é apresentar um modelo geoestatísitico simples, que possa ser adoptado caso a variável resposta possa ser assumida como gaussiana. O modelo deverá considerar um processo estocástico espacialmente contínuo, $S(\mathbf{x})$, associado a localização x. Dependendo da natureza dos dados, pode-se pretender que $S(\mathbf{x})$ seja contínuo e diferenciável de ordem um ou superior. O modelo mais simples que vai ao encontro destes requisitos é o modelo Gaussiano estacionário, que a seguir se define.

1. $\{S(\mathbf{x}) : x \in I\!R^2\}$ é um processo Gaussiano de média μ, variância $\sigma^2 = Var[S(\mathbf{x})]$ e a função de correlação $\rho(u) = Corr[S(\mathbf{x}), S(\mathbf{x}')]$ onde $u = ||x - x'||$ e $||.||$ denota a distância euclidiana;

2. quando condicionados a $\{S(\mathbf{x}) : x \in I\!R^2\}$, os y_i são realizações mutuamente independentes das variáveis de medição Y_i, normalmente distribuídas, com média condicional $E[Y_i|S(.)] = S(\mathbf{x_i})$ e variância condicional τ^2.

Equivalentemente, pode se definir o modelo como:

$$Y_i = S(\mathbf{x_i}) + Z_i : i = 1, \ldots, n \tag{2.1}$$

onde $\{S(\mathbf{x}) : x \in I\!R^2\}$ é definido pela assunção 1. acima, os Z_i são variáveis mutuamente independentes $N(0, \tau^2)$, e n o tamanho da amostra.

De forma a definir um bom modelo, a função correlação $\rho(u)$ tem de ser definida-positiva. Esta condição permite garantir que, para qualquer inteiro m, conjunto de localizações $\mathbf{x_i}$ e constantes reais a_i, a combinação linear $\sum_{i=1}^{m} a_i S(\mathbf{x_i})$ terá uma variância não negativa. Na prática, tal é tipicamente assegurado escolhendo-se um modelo paramétrico para $\rho(u)$ entre um conjunto de modelos teóricos conhecidos (ver Secção 2.5).

Algumas extensões ao modelo proposto

A variação estocástica de uma quantidade física nem sempre é bem descrita pela distribuição Gaussiana. Uma das fórmulas simples para extender o modelo Gaussiano é assumir que o modelo passa a ser adequado após transformação nos dados originais. Para variáveis resposta

de valores positivos, uma classe útil de transformações é a família Box-Cox (Box-Cox, 1964):

$$Y^* = \begin{cases} (Y^\lambda - 1)/\lambda, & \text{se } \lambda \neq 0 \\ \log Y, & \text{se } \lambda = 0 \end{cases}$$

Outra extensão simples do modelo básico é permitir uma média não-constante que dependa da localização x. Por exemplo, pode-se substituir a constante μ por um modelo de regressão linear para a esperança condicional de Y_i dado $S(\mathbf{x_i})$, obtendo-se uma média variável $\mu(\mathbf{x})$.

Uma terceira possibilidade é permitir que $S(\mathbf{x})$ tenha uma estrutura de covariância não estacionária. A maioria dos fenómenos espaciais exibe alguma forma de não estacionariedade, e o modelo Gaussiano estacionário deve ser visto como uma aproximação inicial conveniente cuja a utilidade deve ser cuidadosamente avaliada.

2.5 Função Covariograma e Variograma

O processo espacial Gaussiano $\{S(\mathbf{x}) : \mathbf{x} \in I\!R^2\}$ é um processo estocástico tal que, para qualquer conjunto de localizações $\mathbf{x_1}, \ldots, \mathbf{x_n}$, a distribuição conjunta de $S = \{S(\mathbf{x_1}), \ldots, S(\mathbf{x_n})\}$ é uma Gaussiana multivariada. Qualquer processo deste tipo é completamente definido pela função média $\mu(\mathbf{x}) = E[S(\mathbf{x})]$ e pela função covariância, também denominada **covariograma**, $c(\mathbf{x}, \mathbf{x}') = Cov[S(\mathbf{x}), S(\mathbf{x}')]$.

Por simplificação de notação considere-se μ_S um vector com n elementos $(\mu(\mathbf{x_1}), \ldots, \mu(\mathbf{x_n}))$ e Σ_S uma matriz $n \times n$ em que cada elemento é dados por $c(\mathbf{x_i}, \mathbf{x_j})$. Pode-se, então, escrever

$$S \sim MVN(\mu_S, \Sigma_S)$$

Note-se que o requisito da função covariância $c(\mathbf{x}, \mathbf{x}')$ ser definida-positiva é equivalente a Σ_S ser uma matriz definida positiva.

Um **processo espacial Gaussiano é estacionário** se:

- $\mu(\mathbf{x}) = \mu, \ \forall \mathbf{x} \in A$
- $c(\mathbf{x}, \mathbf{x}') = c(\mathbf{u})$, onde $\mathbf{u} = \mathbf{x} - \mathbf{x}'$, i.e. a covariância depende apenas do vector diferença entre $\mathbf{x}$ e $\mathbf{x}'$

Adicionalmente, um **processo estacionário** é **isotrópico** se $c(\mathbf{u}) = c(||u||)$, ou seja a covariância entre valores de $S(\mathbf{x})$ depende apenas da distância entre as respectivas localizações. Note-se que a variância do processo estocástico é uma constante representada por $\sigma^2 = c(0)$. Por conseguinte, podemos definir a função correlação como

$$\rho(u) = c(u)/\sigma^2,$$

que é simétrica em u dado que $\rho(u) = \rho(-u)$ para qualquer u.

Nestes apontamentos, tipicamente iremos nos referir a estacionário como forma abreviada para estacionário e isotrópico. Note-se que o processo para o qual $S(\mathbf{x}) - \mu(\mathbf{x})$ é estacionário, é denominado um processo **estacionário na covariância** (veja-se exemplo apresentado na Figura 2.5).

Iremos agora apresentar o variograma como uma caracterização alternativa da dependência de segunda ordem para um processo estocástico espacial.

O **variograma** de um processo estocástico espacial $S(\mathbf{x})$ é dado pela função

$$\gamma(\mathbf{x}, \mathbf{x}') = \frac{1}{2} Var[S(\mathbf{x}) - S(\mathbf{x}')]. \tag{2.2}$$

Note-se que $\gamma(\mathbf{x}, \mathbf{x}') = \frac{1}{2}\{Var[S(\mathbf{x})] + Var[S(\mathbf{x}')] - 2Cov[S(\mathbf{x}), S(\mathbf{x}')]\}$. No caso estacionário, tal implica que

$$\gamma(u) = \sigma^2(1 - \rho(u)) \tag{2.3}$$

o que explica a razão pela qual o factor $\frac{1}{2}$ é incluido na expressão do variograma.

Note-se que as expressões (2.2) e (2.3) dizem respeito ao processo $S(\mathbf{x})$, a versão "sem ruído" do processo espacial. Se tivermos em conta o modelo apresentado em (2.1), a nova expressão para o variograma deverá incluir a variância do ruído, tornando-se

$$\gamma(u) = \tau^2 + \sigma^2(1 - \rho(u)).$$

Tipicamente, o variograma tende para um valor constante à medida que a distância u aumenta; este valor é conhecido como **sill** na terminologia inglesa, correspondendo à variância total do processo $Y(.)$ dada por $\sigma^2 + \tau^2$. Quando o variograma tem *sill* (sendo possível especificar um valor para a variância total), tal significa que existe uma distância a partir da qual a correlação entre variáveis é zero; esta distância é chamada **raio de influência** (*range* na terminologia inglesa), e corresponde a um parâmetro tipicamente identificado por ϕ. Aproveita-se para referir que à variância do ruído, τ^2, também se chama **efeito pepita** (*nugget effect*) por razões históricas da geoestatística.

O variograma empírico

usamos o variograma empirico para calcular o teorico pelo MMQ

Iremos agora apresentar o **variograma empiríco**, uma ferramenta muito útil para a análise exploratória de dados, que poderá ser adoptado para encontrar o modelo de correlação espacial adequado para os nossos dados.

Debaixo da estacionariedade, a expressão do variograma (2.2) pode ser re-escrita como

$$\gamma(u) = \frac{1}{2}E[(Y(\mathbf{x}) - Y(\mathbf{x}+u))^2].$$

Consequentemente, pode-se estimar o variograma à custa dos dados amostrados $\{(\mathbf{x_i}, Y(\mathbf{x_i})) : i = 1, \ldots, n\}$, substituindo a esperança teórica anterior pela correspondente média amostral. Para um dado *lag* u, podemos calcular a média das diferenças ao quadrado entre os pares de observações $Y(\mathbf{x_i})$ e $Y(\mathbf{x_j})$ cujas respectivas localizações $\mathbf{x_i}$ and $\mathbf{x_j}$ têm $\|\mathbf{x_i} - \mathbf{x_j}\| = u$.[1]

Tal sugere um modo de estimar o variograma para um processo estacionário à custa dos dados observados. Note-se que é possível considerar pesos no cálculo da média de forma a se obter uma estimativa mais suave do variograma.

Estimadores de variograma frequentemente adoptados

A primeira proposta, na presença de processos estacionários, para um estimador do variograma deve-se a Matheron (1962). Este estimador é basedo no método dos momentos, sendo tipicamente referido como o **estimador clássico**:

$$\widehat{\gamma}(u) = \frac{1}{2|N(u)|} \sum_{N(u)} (Y(\mathbf{x_i}) - Y(\mathbf{x_j}))^2 \qquad (2.4)$$

nº de pares

onde $N(u) = \{(\mathbf{x_i}, \mathbf{x_j}) : \|\mathbf{x_i} - \mathbf{x_j}\| = u,\ u \in I\!R\}$ e $|N(u)|$ é o número total de pares em $N(u)$. O estimador proposto por Matheron é centrado[2], contudo apresenta algumas desvantagens tais como ser afectado por valores atípicos devido ao termo ao quadrado que aparece no somatório de (2.4). De um modo geral, as suas propriedades estatísticas são difíceis de estudar. Se $Y(.)$ é um processo Gaussiano, então $\widehat{\gamma}(u)$ é uma combinação linear de variáveis aleatórias χ^2 com um grau de liberdade.

De acordo com Journal (1978), o número mínimo de 30 pares é recomendado em $|N(u)|$. Quando os dados não são regularmente espaçados, este estimador deve ser obtido considerando uma região de tolerância em torno de u.

Note que o termo ao quadrado é também usado para propor um outro variograma, denominado **variograma nuvem**. Se $\{(Y(\mathbf{x_i}), \mathbf{x_i}) : i = 1, ..., n\}$ representa o conjunto de dados amostrados, então o gráfico de pontos $\{(u_{ij}, v_{ij}) : j > i,\ u_{ij} = \|\mathbf{x_i} - \mathbf{x_j}\|,\ v_{ij} = \frac{1}{2}(Y(\mathbf{x_i}) - Y(\mathbf{x_j}))^2\}$ identifica o variograma nuvem correspondente. A Figura 2.6 apresenta o exemplo de um variograma nuvem e de um empírico para o mesmo conjunto de dados. Tal

[1] Note-se que tipicamente é considerada uma região de tolerância para permitir incluir pares adicionais de valores $Y(\mathbf{x_i})$ e $Y(\mathbf{x_j})$ tais que $\|\mathbf{x_i} - \mathbf{x_j}\| \approx u$.

[2] É centrado para $\gamma(.)$ quando $Y(.)$ é intrinsicamente estacionário (Cressie 1993, page 71).

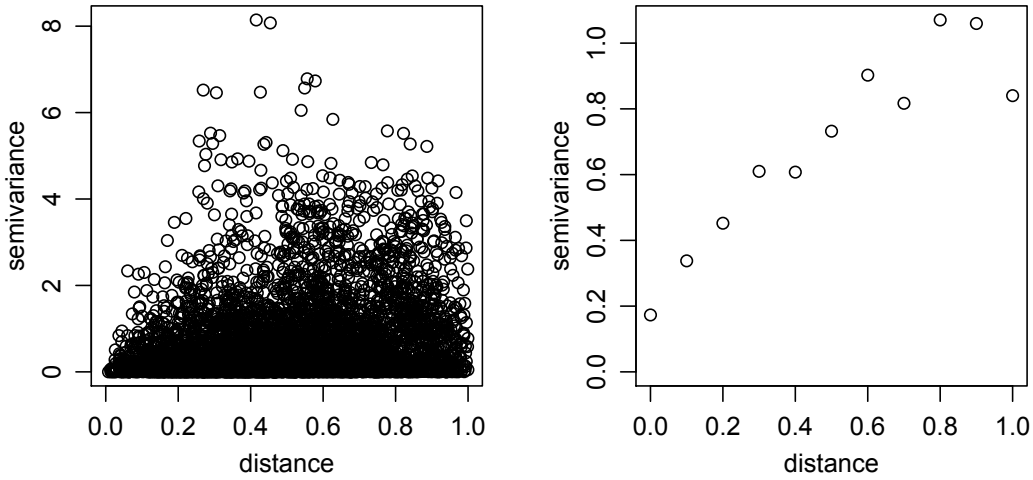

Figura 2.6: Variograma nuvem (painel esquerda) e variograma empírico proposto por Matheron (painel direita).

como esperado, o estimador nuvem também é sensível a outliers.

Debaixo da gaussianidade, Cressie (1980) propõe um estimador mais robusto a outliers, substituindo o termo ao quadrado pela raiz quadrada das diferenças absolutas

$$\widehat{\gamma}(u) = \frac{\left\{\frac{1}{2|N(u)|}\ \sum_{N(u)} |Y(\mathbf{x_i}) - Y(\mathbf{x_j})|^{\frac{1}{2}}\right\}^4}{0.457 + \frac{0.494}{|N(u)|}} \tag{2.5}$$

onde o termo $0.457 + \frac{0.494}{|N(u)|}$ é usado para tornar o estimador centrado.

Infelizmente, os estimadores dados em (2.4) e (2.5) não devem ser usados num contexto de inferência e predição. Tal acontece, porque eles podem falhar a propriedade de condicionalmente negativo-definido, podendo originar valores negativos absurdos para a média dos erros de predição quadráticos, conforme provado em Cressie(1993).

Propriedades importantes

Conforme referido no final da secção anterior, uma propriedade critica do variograma é tratar-se de uma função condicionalmente definida-negativa[3], ou seja

[3]Tal como para o covariograma é exigido a propriedade de condicionalmente definido-positivo.

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma(\mathbf{x_i} - \mathbf{x_j}) \leq 0$$

para qualquer conjunto finito de localizações espaciais, $\{\mathbf{x_1}, ..., \mathbf{x_n}\}$, e qualquer conjunto de números reais $\{a_1, ..., a_n\}$, tais que $\sum_{i=1}^{n} a_i = 0$.

A ideia é selecionar, entre famílias de variogramas definidos-positivos, um variograma que melhor aproxime a dependência espacial subjacente aos dados amostrados disponíveis.

Outras propriedades importantes dos variogramas são agora descritas. Considere-se $A_1 = \{\mathbf{x} = \mathbf{x_1} - \mathbf{x_2} : \mathbf{x_1}, \mathbf{x_2} \in A\}$. Então, para todo $\mathbf{x} \in A_1$, tem-se:

- $2\gamma(0) = 0$, i.e. é nulo origem;
- $2\gamma(\mathbf{x}) \geq 0$;
- $\lim_{\|\mathbf{x}\|\to 0} \gamma(\mathbf{x}) = \theta_0$, onde $\theta_0 \geq 0$ é o efeito pepita;
- $2\gamma(\mathbf{x}) = 2\gamma(-\mathbf{x})$, i.e. é uma função simétrica;
- $\lim_{\|\mathbf{x}\|\to\infty} \gamma(\mathbf{x}) / \parallel \mathbf{x} \parallel^2 = 0$, i.e. a taxa de crescimento de $\gamma(.)$ deve ser menor que $\parallel \mathbf{x} \parallel^2$.

Se a última propriedade falha, não existirá estacionariedade de segunda-ordem para os incrementos, e será esperada a presença de tendência não constante $\mu(\mathbf{x})$.

A primeira propriedade confirma que o variograma não é necessariamente uma função contínua. Na teoria, para distâncias muito pequenas a di-associação entre valores da variável tende para zero. Na prática, contudo, o variograma pode ser significativamente diferente de zero (o tal efeito pepita), reflectindo algum efeito local ou **erro de medição** conforme anteriormente discutido. O comportamento do variograma perto da origem ajuda a definir propriedades de continuidade para o processo aleatório $Y(.)$. O tipo mais comum pode ser categorizado da seguinte forma:

i. Se $\theta_0 = 0$, então $Y(.)$ é L_2-contínuo.

ii. Se $\theta_0 \neq 0$, então $Y(.)$ não é L_2-contínuo e é bastante irregular.

iii. Se $\gamma(.)$ é uma constante positiva (excepto na origem onde é zero), então $Y(\mathbf{x_i})$ e $Y(\mathbf{x_j})$ são não correlacionados para qualquer $\mathbf{x_i} \neq \mathbf{x_j}$, independentemente de quão próximos eles estão; $Y(.)$ é então tipicamente chamado ruído branco.

Alguns modelos teóricos isotrópicos

Alguns exemplos de variogramas válidos que possuem as propriedades anteriores são apresentados na Figura 2.7. Estas curvas suaves são membros de alguma família paramétrica válida do tipo

$$P = \{\gamma : \ \gamma(.) = \gamma(.;\theta), \ \theta \in \Theta\}$$

onde $\gamma(.;\theta)$ é uma função condicionalmente definida-negativa dependendo de valores dados no vetor de parâmetros θ.

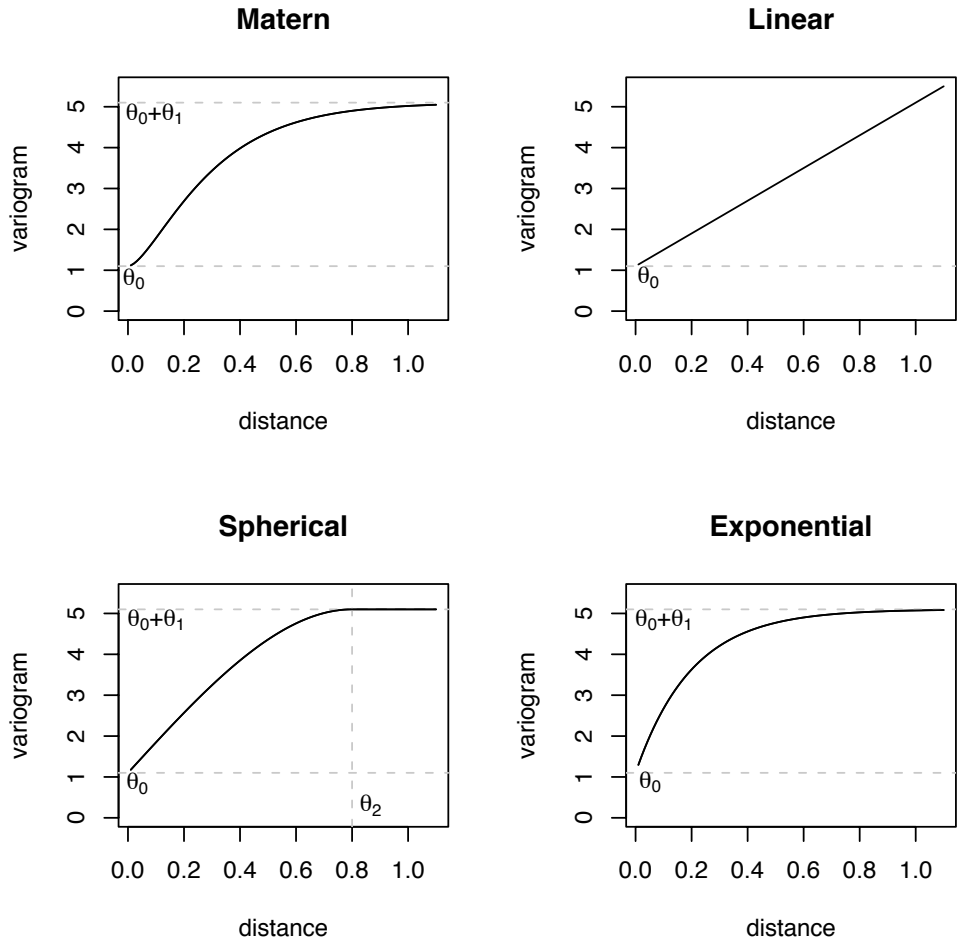

Figura 2.7: Exemplos de modelos de variogramas isotrópicos: Matérn (com $\kappa = 1.0$), linear, esférico e exponencial.

Consideram-se quatro modelos isotrópicos básicos: Matérn, linear, esférico e exponencial.

- Modelo Matérn:
$$\gamma(u;\theta) = \theta_0 + \theta_1(1 - \rho(u;\phi = \theta_2, \kappa = \theta_3)) \tag{2.6}$$
onde $\theta = (\theta_0, \theta_1, \theta_2, \theta_3)^t$ e
$$\rho(u;\phi,\kappa) = (2^{\kappa-1}\Gamma(\kappa))^{-1}\ (u/\phi)^{\kappa}K_{\kappa}(u/\phi)$$
com $K_\kappa(.)$ denotando a função Bessel modificada de 2ªordem em κ.

- Modelo linear:
$$\gamma(u;\theta) = \theta_0 + \theta_1 u$$
onde $\theta = (\theta_0, \theta_1)^t$. Esta função não tem *sill*, $\lim_{u\to\infty}\gamma(u) = \infty$ e, por conseguinte, não corresponde a um processo estacionário.

- Modelo esférico:
$$\gamma(u;\theta) = \begin{cases} \theta_0 + \theta_1\left(\frac{3u}{2\theta_2} - \frac{1}{2}\left(\frac{u}{\theta_2}\right)^3\right) & , 0 < u \leq \theta_2 \\ \theta_0 + \theta_1 & , u > \theta_2 \end{cases} \tag{2.7}$$
onde $\theta = (\theta_0, \theta_1, \theta_2)^t$. Neste modelo, $\theta_0 + \theta_1$ corresponde ao *sill* e θ_2 ao "raio de influência". Tem um comportamento linear perto da origem e, na prática, é um dos modelos mais adoptados uma vez que pode facilmente ser ajustado aos dados.

- Modelo exponencial: Qdº k=0.5 na função de matérn dá-nos a exponencial
$$\gamma(u;\theta) = \theta_0 + \theta_1\left(1 - exp\left(-\frac{u}{\theta_2}\right)\right),\ u \neq 0 \tag{2.8}$$
onde $\theta = (\theta_0, \theta_1, \theta_2)^t$. Neste modelo, $\theta_0 + \theta_1$ identifica o *sill* apenas no sentido assimptótico, sendo $\sqrt{3}\theta_2$ o "raio de influência" correspondente. Tem um comportamento parabólico perto da origem.

Note-se que, com $\kappa = 0.5$ in (2.6), tem-se que os modelos de Matérn e exponencial coincidem.

Capítulo 3

Inferências no modelo Geoestatístico

Os principais objectivos científicos da análise geoestatística são estimação e predição. A estimação refere-se a inferências sobre parâmetros do modelo estocástico obtidas à custa dos dados. Por sua vez, a predição refere-se a inferências sobre a realização do processo não observado $S(.)$.

3.1 Estimação de parâmetros

No modelo estacionário Gaussiano os parâmetros a serem estimados são dados pela média μ, e por qualquer parâmetro adicional que defina a estrutura de covariância dos dados, nomeadamente a variância de $S(.)$ representada por σ^2, a variância do erro de medição representada por τ^2, e o "raio de influência" ϕ; no caso do modelo de Matérn, com o parâmetro adicional κ. Na secção 2.5, os parâmetros $\theta_0, \theta_1, \theta_2$ e θ_3 correspondem a τ^2, σ^2, ϕ e κ, respectivamente.

Para se estimar os parâmetros, pode se utilizar o método de mínimos quadrados ou o método de máxima verosimilhança.

Máxima Verosimilhança likfit(dados,...)

A estimação por máxima verosimilhança é um método estatístico universalmente aceite, com boas propriedades para amostras de grande dimensão. Debaixo de algumas condições de regularidade, o estimador de máxima verosimilhança é assimptoticamente: normalmente distribuído, centrado e eficiente. Dentro do contexto da Geoestatística, a implementação da máxima verosimilhança é apenas directa quando os dados são gerados por um modelo Gaussiano. Contudo, este método é útil para muitas aplicações geoestatísticas para as quais Y é uma quantidade com valor contínuo. Adicionalmente, note-se que as dificuldades associadas à implementação da estimação por máxima verosimilhança em modelos não-Gaussianos são apenas computacionais. A maioria das boas propriedades deste método de estimação

mantêm-se fora do contexto Gaussiano.

Suponha-se então

$$Y = (Y(X_1), \ldots, Y(X_n)) \sim MVN(X\beta,\ \Sigma(\theta))$$

onde X é a matriz $n \times p$ com as observações de p covariáveis ($p < n$), β o vetor de coeficientes de regressão e $\Sigma(\theta)$ a matriz de covariância que pode ser escrita como

$$\Sigma(\theta) = \tau^2 I + \sigma^2 R(\phi)$$

onde $R(\phi)$ corresponde à matriz de correlação sendo $r_{ij} = \rho(||\mathbf{x_i} - \mathbf{x_j}||; \phi)$, $i, j = 1, \ldots, n$ e I a matriz identidade de dimensão $n \times n$. Consequentemente, todos os parâmetros do modelo podem ser estimados maximizando a função de log–verosimilhança que é dada por

função likfit do R:

$$L(\beta, \tau^2, \sigma^2, \phi) = -\frac{1}{2}\left\{n\log(2\pi) + \log(|\tau^2 I + \sigma^2 R(\phi)|) + (Y - X\beta)^t(\tau^2 I + \sigma^2 R(\phi))^{-1}(Y - X\beta)\right\}$$

L(parametros/dados)

determinante

Alguns dos softwares geoestatísticos, como o `geoR`, resolvem o problema de optimização de forma iterativa sendo possível especificar valores iniciais para os parâmetros.

Método dos mínimos quadrados

variofit(variograma.empirico, ..., weights=pairs)

Os critérios basedos em métodos de mínimos quadrados (MQ) são conhecidos como sendo mais abrangentes, uma vez que requerem menos pressupostos na distribuição de $Y(x)$. Em notação matricial, um problema MQ pode ser escrito

$$min_\theta \left\{(\widehat{\gamma} - \gamma_\theta)^T W^{-1} (\widehat{\gamma} - \gamma_\theta)\right\}$$

onde $\hat{\gamma}$ identifica um estimador empírico e γ_θ um modelo paramétrico válido, cuja forma exacta é conhecida excepto para os parâmetros desconhecidos definidos em θ. A matriz W é a matriz de pesos tal que:

1. Se W é a matriz identidade estamos perante o critério de MQ ordinários (MQO). Neste caso, o problema MQO pode ser escrito como

   $$min_\theta \left\{\Sigma_{i=1}^{L} (\widehat{\gamma}(u_i) - \gamma_\theta(u_i))^2\right\},$$

   onde L representa o número total de *lags* (i.e. distâncias) usadas na estimação.

   A principal vantagem do método MQO é a simplicidade de aplicação. Contudo, não envolve informação sobre a variância e covariância da distribuição do estimador empírico, nem do número de pares de pontos envolvidos na estimação do *lag* u.

2. Se $W = V$ onde V é a matriz covariância cujos elementos são do tipo $V_{ij} = Cov[\widehat{\gamma}(u_i), \widehat{\gamma}(u_j)]$, então estamos perante o critério de mínimos quadrados generalizados (MQG).

   A principal desvantagem do método MQG é ser demasiado complexo e depender da função desconhecida $\widehat{\gamma}$.

3. Se V é a matriz diagonal com $V_{ii} = Var[\widehat{\gamma}(u_i)]$, então estamos perante o critério de mínimos quadrados pesados (MQP). Este caso é um compromisso pragmático entre a eficiência do critério de mínimos quadrados generalizados e a simplicidade do critério de mínimos quadrados ordinários.

Em Cressie (1985), o autor prova que $Var[\widehat{\gamma}(u_i)] \approx \frac{2\gamma(u_i)^2}{|N(u_i)|}$, e sugere os seguintes pesos

$$w_i = \frac{|N(u_i)|}{\gamma(u_i)^2},$$

onde a função desconhecida γ pode ser aproximada por γ_θ através de um procedimento iterativo (pesos podem ser iniciados, por exemplo, com valor igual 1). Algumas notas sobre estes pesos:

- Se $N(u_i)$ é maior, então o peso associado w_i também é maior; argumento default weights=pairs
- Se o valor de γ_θ é menor, então o peso associado w_i é maior (permite uma melhor caracterização perto da origem);
- Se a variância é maior, então o peso associado w_i é menor.

Validação cruzada do variograma ajustado

Suponha que um modelo de variograma $\gamma_{\widehat{\theta}}$ foi ajustado aos dados $\{Y(\mathbf{x_i}) : i = 1, ..., n\}$. Adicionalmente, suponha que conhecemos algum método de predição baseado em $\gamma_{\widehat{\theta}}$. Uma forma de diagnosticar a eventual existência de algum problema com o ajuste passa por fazer a validação cruzada do modelo escolhido.

A ideia fundamental desta técnica é estimar a medição $Y(\mathbf{x})$ em cada ponto amostrado $\mathbf{x_i}$ à custa dos dados vizinhos $Y_j = Y(\mathbf{x_j})$, $j \neq i$, considerando $Y_i = Y(\mathbf{x_i})$ como sendo desconhecido. Desta forma, em cada ponto amostrado $\mathbf{x_i}$, obtem-se uma predição $\widehat{Y}_{-i} = \widehat{Y}(\mathbf{x_i})$ e a respectiva variância σ^2_{-i}. Note-se que os valores de predição são influenciados por $\gamma_{\widehat{\theta}}$.

Como na realidade se conhece o valor verdadeiro de Y_i, pode-se calcular um erro predição $PE_i = Y_i - \widehat{Y}_{-i}$. Se o variograma teórico $\gamma(u)$ for conhecido, PE_i é uma variável aleatória com média zero e variância σ^2_{-i}. Adicionalmente, o erro padronizado $e_i = PE_i/\sigma_{-i}$ é uma

variável aleatória com média zero e variância unitária.

A proximidade dos valores estimados $\widehat{Y}_{-i}$ e verdadeiros Y_i pode ser caracterizada considerando o erro quadrático médio (EQM) e o erro quadrático médio standardizado (EQMS)

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_{-i})^2, \tag{3.1}$$

$$EQMS = \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \widehat{Y}_{-i})^2}{\widehat{\sigma^2_{-i}}}. \tag{3.2}$$

Para avaliar a qualidade de ajuste do variograma $\gamma_{\widehat{\theta}}$, a média em (3.1) deve ser aproximadamente 0 e a raiz da expressão dada em (3.2) deve ser aproximadamente 1. Alternativamente, pode-se analisar o histograma dos erros standardizados e confirmar que seguem (aproximadamente) uma normal standard.

Como comentário final, note que a comparação dos resultados de duas validações-cruzadas pode ser útil na escolha entre dois modelos razoáveis.

3.2 Predição espacial

Na predição espacial pretende-se saber, dado um conjunto de n observações do processo $Y(.)$ nos pontos $\{\mathbf{x_i} : i = 1, ..., n\}$, qual o valor assumido pela variável num ponto $\mathbf{x_0}$, onde os dados não estão disponíveis. A abordagem para resolução do problema difere da regressão pelo facto das características locais poderem afectar a solução. Por princípio, todas as medições devem ser consideradas. Tendo em conta que algumas medições na vizinhança do ponto $\mathbf{x_0}$ investigado, ou por vezes noutros locais, estão mais relacionadas com o valor verdadeiro nesse ponto, o procedimento mais adequado passa por adoptar uma *média pesada*.

$$\widehat{Y}(\mathbf{x_0}) = \textstyle\sum_{i=1}^{n} \lambda_i Y(\mathbf{x_i}).$$

Esta combinação linear pode ser considerada uma estimativa óptima se os coeficientes λ_i, também chamados pesos de predição, são tais que somam um, o estimador é centrado e tem variância mínima[1]. Apenas os dados dentro do raio de influência devem ser considerados.

Tipicamente, a estimação de um variograma ou covariograma válido joga aqui um papel decisivo, uma vez que é habitualmente usado para encontrar a solução óptima para os valores

[1]Mais precisamente, este estimador é classificado como BLUE, i.e. *Best Linear Unbiased Estimator*.

dos pesos λ_i. O método é chamado **kriging** e pode ser aplicado a um processo estacionário de segunda ordem ou intrínseco. O nome kriging foi escolhido por Matheron (1963) em honra ao engenheiro de minas D.G.Krige.

O método mais simples de predição denomina-se de **kriging simples**. Este assume que as médias locais são relativamente constantes de valor muito semelhante à média da população que é conhecida, ou seja, **assume a função $\mu(x)$ como conhecida**. A média da população é utilizada para cada estimação local, em conjunto com os pontos vizinhos estabelecidos como necessários para a estimação.

O que é pouco usual acontecer na vida real

Kriging simples tem por objectivo minimizar o erro quadrático médio, considerando as estimativas dos parâmetros do modelo como sendo verdadeiras. A predição $\hat{S}(\mathbf{x_0})$ é obtida minimizando o erro quadrático médio, ou seja, minimizando a quantidade $E[(\hat{S}(\mathbf{x_0}) - S(\mathbf{x_0}))^2]$ considerando a nova localização $\mathbf{x_0}$ e $\mathbf{y} = (y_1, \ldots, y_n)$.

Na secção seguinte iremos analisar com detalhe o caso de um processo Gaussiano estacionário, para o qual a função linear de y_i é dada por

$$\hat{S}(\mathbf{x_0}) = \mu + \sum_{i=1}^{n} w_i(x)(y_i - \mu)$$

onde w_i são pesos tipicamente definidos como funções dos parâmetros da estrutura de covariância, nomeadamente σ^2, τ^2 e ϕ.

Outro método de predição denomina-se **kriging ordinário**, onde se **assume que $\mu(x)$ é igual a uma constante μ desconhecida**. As médias locais não são necessariamente próximas da média da população usando-se apenas os pontos vizinhos para a estimação. Para o cálculo dos pesos é utilizada a média local dos pontos amostrados, por conseguinte deve-se normalizar a média dos pesos e, consequentemente, tem-se um resultado mais preciso do que de kriging simples.

Para considerar o caso em que a média do processo é variável torna-se necessário adaptar a fórmula acima para $\hat{S}(\mathbf{x_0})$, substituindo a constante μ por uma **tendência espacial $\mu(x)$ variável**. Este procedimento é tipicamente associado ao método de **kriging universal**.

Dados Gaussianos

Debaixo do modelo Gaussiano semelhante ao proposto em (2.1), tem-se

$$Y(\mathbf{x}) = S(\mathbf{x}) + \mathrm{N}(0, \tau^2),$$

onde $Y(.)$ é o processo medição e $S(.)$ é um processo estocástico não observado, o nosso objectivo de predição. Por conseguinte, o foco de interesse é o processo $S(.)$, condicionado

aos dados observados $\mathbf{y}$.

Iremos considerar

$$\mathbf{Y} = (Y_1, ..., Y_n)^t \sim \text{MVN}(\mu_Y \mathbf{1}, \mathbf{\Sigma_Y}),$$

onde $\mathbf{1}$ denota o vetor de n-uns. Tendo em conta que

- $Y_i = S(\mathbf{x_i}) + Z_i,\ i = 1, ..., n,$ onde $S(.)$ tem média μ, variância σ^2 e função de correlação $\rho(u; \phi)$,
- e $Z_1, ..., Z_n$ são i.i.d. com $Z_i \sim \text{N}(0, \tau^2),\ i = 1, ..., n,$

então tem-se

$$\mu_Y = \mu \qquad \text{e} \qquad \mathbf{\Sigma_Y} = \sigma^2 \mathbf{R_Y}(\phi) + \tau^2 \mathbf{I_n}$$

onde $\mathbf{I_n}$ é a matriz $n \times n$ identidade e $\mathbf{R_Y}(\phi)$ é a matriz $n \times n$ com o elemento (i, j) igual a $\rho(u_{ij})$ sendo $u_{ij} = \|\mathbf{x_i} - \mathbf{x_j}\|$.

Distribuição conjunta $[S(\mathbf{x_0}), Y]$

Teorema: *Seja $X = (X_1, X_2)$ uma v.a. Gaussiana multivariada conjunta, com vetor média $\mu = (\mu_1, \mu_2)$ e matriz de covariância*

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

i.e. $X \sim MVN(\mu, \Sigma)$. Então a distribuição condicional de X_1 dado X_2 também é uma Gaussiana multivariada, $X_1 | X_2 \sim MVN(\mu_{1|2}, \Sigma_{1|2})$, onde

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2) \qquad e \qquad \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Para aplicar o teorema anterior ao problema de predição, note que a distribuição conjunta $[S(\mathbf{x_0}), Y]$ é Gaussiana multivariada com vetor média $(\mu, \mu\mathbf{1})^t$ e matriz covariância

$$\begin{pmatrix} \sigma^2 & \sigma^2 \mathbf{r}^t \\ \sigma^2 \mathbf{r} & \tau^2 \mathbf{I_n} + \sigma^2 \mathbf{R} \end{pmatrix}$$

onde $\mathbf{r}$ é um vetor com elementos $r_i = \rho(\|\mathbf{x_0} - \mathbf{x_i}\|) : i = 1, ..., n,$ $\mathbf{R}$ é uma matriz $n \times n$ com o elemento (i, j) igual a $\rho(\|\mathbf{x_i} - \mathbf{x_j}\|)$; e a distribuição condicional de interesse é a $[S(\mathbf{x_0}) \mid Y = \mathbf{y}]$.

Estimativa de Kriging pela formula

Por conseguinte, o preditor obtido minimizando o erro quadrático médio é dado por

$$\widehat{S}(\mathbf{x_0}) = \mathrm{E}[S(\mathbf{x_0})|\mathbf{y}] = \sigma^2\mathbf{r}^t(\tau^2\mathbf{I} + \sigma^2\mathbf{R})^{-1}(\mathbf{y} - \mu\mathbf{1}) + \mu \tag{3.3}$$

e respectiva variância de predição

$$\mathrm{Var}[S(\mathbf{x_0})|\mathbf{y}] = \sigma^2 - \sigma^2\mathbf{r}^t(\tau^2\mathbf{I} + \sigma^2\mathbf{R})^{-1}\sigma^2\mathbf{r} \tag{3.4}$$

É importante realçar que a variância de predição (3.4) depende do modelo de correlação, depende da configuração espacial dos dados e depende da localização de predição $\mathbf{x_0}$, mas não depende directamente das medições diponíveis $\mathbf{y}$.

Média não-constante $\mu(\mathbf{x})$

Uma extensão do modelo Gaussiano pode ser considerada ao permitir que a média do processo espacial seja não constante e dependa da localização x. Tipicamente, pode ser útil considerar

$$\mu(x) = \Sigma_{j=1}^{p}\beta_j f_j(x),$$

onde $f_1(x), ..., f_p(x)$ são funções observadas da localização x, ou funções de covariáveis observadas, levando ao **kriging universal** ou **kriging com um modelo de tendência**. Um estimador para o vetor $\beta = (\beta_1, ..., \beta_p)^t$ desconhecido pode ser obtido por máxima verosimilhança, levando a

$$\widehat{\beta} = (\mathbf{F}^t\mathbf{V}^{-1}\mathbf{F})^{-1}\mathbf{F}^t\mathbf{V}^{-1}\mathbf{y}, \tag{3.5}$$

onde $\mathbf{F}$ é uma matriz $n \times p$ com o elemento (i, j) igual a $f_j(\mathbf{x_i})$ e $\mathbf{V} = \mathbf{R} + \tau^2/\sigma^2\mathbf{I}$.

Neste caso, a expressão (3.3) para o preditor altera ligeiramente, obtendo-se

$$\widehat{S}(\mathbf{x_0}) = \sigma^2\mathbf{r}^t(\tau^2\mathbf{I} + \sigma^2\mathbf{R})^{-1}(\mathbf{y} - \mathbf{F}\widehat{\beta}) + \mathbf{F_0}\widehat{\beta}$$

onde $\mathbf{F_0}$ é uma matriz $1 \times p$ com o elemento $(1, j)$ igual a $f_j(\mathbf{x_0})$.

Por último, note-se que o caso de **kriging ordinário** ($\mu(\mathbf{x}) = \mu$, desconhecido) ocorre quando $p = 1$ e $\mathbf{F}$ é igual a um vetor de uns. A expressão (3.5) irá então retornar um valor real $\widehat{\beta}$ como um estimador para $\widehat{\mu}$.