Estatística Espacial
Dados Agregados por Área

Raquel Menezes

Departamento de Matemática
Universidade do Minho

Novembro de 2023

Estrutura em Reticulado?

Se a paisagem ou região estiver dividida em subáreas (Cressie, 1993):

- Reticulado[1] regular, por exemplo, imagens de satélite obtidas remotamente (dados sobre solo, vegetação ou temperatura da superfície; note-se que a informação não é especifica de um ponto, mas é definida para uma área)
- Reticulado irregular, por exemplo, se a região for dividida em células com base em bacias fluviais, fronteiras nacionais, distritos ou códigos postais

Objetivos:

- Detecção de padrões espaciais sobre o reticulado (por exemplo, a precipitação em unidades conectadas pode ser semelhante)
- Explicação desses padrões em termos de covariáveis medidas nas mesmas células

[1] Na terminologia inglesa, dito *lattice*.

Algumas definições

Notação:

- Reticulado consistindo em n unidades ou polígonos, $A_1, \ldots, A_n$, tal que $A_1 \cup A_2 \cup \cdots \cup A_n = S \subset \mathbb{R}^2$ e $A_i \cap A_j = \emptyset$ para $i \neq j$
- Variável aleatória na localização A_i definida como $Y_i = Y(A_i)$ com valor observado y_i
- $\mu_i = E[Y_i]$ e $Cov[Y_i, Y_j]$ identificam momentos de 1ª e 2ª ordem

Como as relações espaciais podem ser quantificadas?

Através de uma **matriz de pesos espaciais W** na qual os elementos representam a força da estrutura espacial entre as unidades (auto-correlação espacial).

Matriz de vizinhanças ou pesos

Há várias formas de representar a matriz dos pesos

Exemplos de **W**:

- $w_{ij} = \begin{cases} 1 & \text{se } A_i \text{ e } A_j \text{ têm uma fronteira comum} \\ 0 & \text{caso contrário} \end{cases}$
- $w_{ij} = d_{ij}^{-\gamma}$, $\gamma \geq 0$, d_{ij} =distância entre os centróides A_i e A_j
- $w_{ij} = (l_{ij}/l_i)/d_{ij}^{-\gamma}$, l_{ij} =comprimento da fronteira comum entre A_i e A_j, l_i =perímetro de A_i Fronteiras entre 2 areas, maiores têm maior peso

Então, quanto mais próximas no espaço duas unidades A_i e A_j estiverem (com uma fronteira comum significativa), maior será o fator de ponderação w_{ij}.

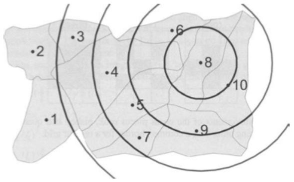
Outros exemplos de **W**

A_i e A_j são considerados contíguos, se seus centróides estiverem a menos de uma **"distância de corte"** especificada, ou seja, se $d_{ij} \leq d_k$. Alternativamente, se

$$w_{ij}^{(k)} = w_{ij}(d_k, d_{k+1}) = \begin{cases} 1 & \text{se } d_k < d_{ij} \leq d_{k+1} \\ 0 & \text{caso contrário} \end{cases} \qquad (1)$$

Para distâncias de corte $\{0, 3, 6, 9, 12\}$, quais são as unidades contíguas à unidade 8 ?

distância de corte d_k	3km	6km	9km	12km
unidades contíguas	$\{10\}$	$\{6,9\}$	$\{4, 5, 7\}$	$\{3\}$
$w_{8,j}$ não nulo	$w_{8,10} = 1$	$w_{8,6} = w_{8,9} = 1$	$w_{8,4} = w_{8,5} = w_{8,7} = 1$	$w_{8,3} = 1$



O coeficiente de Moran I

Este coeficiente quantifica o **grau de correlação espacial** entre unidades vizinhas, como

Estatística de teste:

$$I = \frac{n}{w_{++}} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{n}{\mathbf{1}'\mathbf{W}\mathbf{1}} \frac{\mathbf{u}'\mathbf{W}\mathbf{u}}{\mathbf{u}'\mathbf{u}} \qquad (2)$$

onde w_{++} é a soma dos pesos w_{ij} (em notação matricial, $\mathbf{u}$ contém elementos y_i 'centrados'').

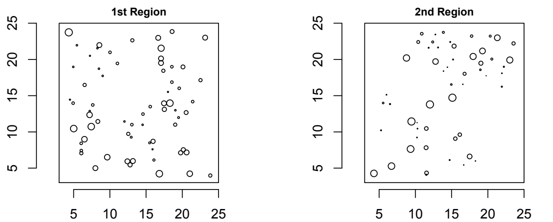
Notas:

- Se dois valores adjacentes são ambos acima da média ou ambos abaixo da média, então há uma **correlação espacial positiva**. Caso contrário, há uma **correlação negativa**.
- Sob **"H_0: não correlação espacial"**, então $E[I|H_0] = -\frac{1}{n-1}$ (próximo de 0 para n grande).
- Um **valor-p pode ser obtido** assumindo normalidade assintótica ou usando um teste de permutação.

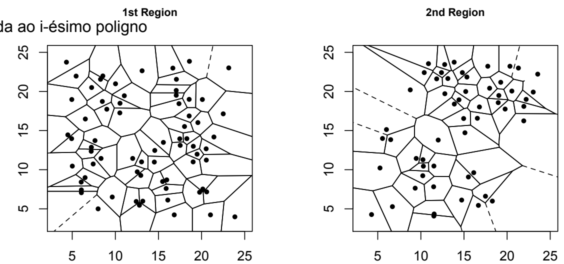
Exemplo: dados de árvores em 2 regiões diferentes numa reserva russa

Dados sobre **distribuição espacial**, **altura** e **diâmetro** da copa para 4 espécies de árvores. As localizações das árvores foram convertidas para um reticulado usando a tesselação de Voronoi (Moller, 1994).

Yi= altura da i-ésima localização - Processos Pontuais Marcados



Y(Ai)= altura assocuada ao i-ésimo poligno



Aqui os pontos não são os centróides, é onde está a árvore

Exemplo: dados de árvores

Ver código R a partir da linha 90 do ficheiro RCode_DadosAgregados

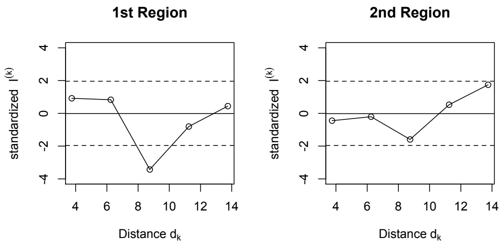
O coeficiente de Moran I é usado para investigar se há alguma auto-correlação espacial entre as alturas das árvores.

H_0: nenhuma correlação espacial na altura das árvores

	n	$E[I\|H_0]$	Estatística I de Moran estatística	Valor p	**Correlação espacial ?**
1ª região	70	-0.015	0.246	< 0.001	presente
2ª região	54	-0.019	0.021	0.62	ausente

Correlograma espacial de Moran – exemplo

Se $w_{ij} = w_{ij}^{(k)}$ em (1), então **o índice de Moran I em (2) dependerá dos "limites de corte"** d_k **e** d_{k+1}. Um índice $I^{(k)}$ pode ser calculado para diferentes k. O gráfico, que relaciona d_k e $I^{(k)}$, mostra como a força da dependência muda com a distância entre as unidades/áreas.



Notas:

Seja d_k={2.5,5,7.5,10,12.5,15}, então $I^{(k)}$ é calculado primeiro usando apenas pontos separados por 2.5 m e depois com pontos entre 2.5 e 5 m, etc. **Para a 1ª Região**, há evidência estatística de que as **alturas das árvores separadas por 7.5 a 10 m são dependentes** ($p<0.001$). Para todas as outras classes, não há evidência para rejeitar H_0 de nenhuma correlação espacial [2].

[2] Linhas tracejadas fornecem bandas de confiança de 95% para "nenhuma correlação espacial".

Modelo de regressão linear (M1) – exemplo

Considere o modelo de regressão linear (RL)

$$\textbf{Altura da árvore}_i = \alpha + \beta\ \textbf{Diâmetro da árvore}_i + \epsilon_i \qquad \textbf{(M1)}$$

então

model_LM <- lm(height ~ diameter, data = plot_1_6)

	estimativa	s.e.	valor p
α	5.820	0.441	<0.001
β	0.311	0.021	<0.001

com erro padrão residual = 2.622 e AIC = 337.547.

nao vamos poder usar o lm

O modelo de RL **(M1)** assume que os resíduos são independentemente distribuídos, mas para

$$\mathbf{H_0}: \textbf{os resíduos são não (espacialmente) correlacionados}$$

I=0.115 (valor p=0.026), então **devemos assumir resíduos dependentes** para α=5%.

model_LM.moran <- lm.morantest(model_LM, W, alternative = "greater",
resfun = weighted.residuals)

Modelos de regressão com auto-correlação espacial

Caso o pressuposto de "independência dos resíduos" for violado, como proceder?

se os residuos tiverem correlação

1. Incluir mais variáveis explicativas em (M1) para reduzir a auto-correlação espacial
2. Modelo auto-regressivo condicional, **modelo CAR**
   A formulação geral do CAR para uma variável de resposta Y numa unidade espacial i é:
   $$Y_i = \rho \sum_{j \in N(i)} w_{ij} Y_j + \epsilon_i$$
   onde
   - ρ é o parâmetro de auto-correlação espacial
   - $N(i)$ é o conjunto de unidades espaciais vizinhas à unidade
   - w_{ij} são os elementos da matriz de ponderação espacial que indicam a força da conexão entre as unidades i e j
   - ϵ_i é o termo de erro

   NOTA: O CAR assume que Y_i é uma média ponderada das respostas nas unidades espaciais vizinhas, com um termo de erro aditivo, capturando assim a auto-correlação espacial nos dados.

Modelos de regressão com auto-correlação espacial

Modelos de regressão espacial: SAR e SMA

3. Modelo auto-regressivo simultâneo, **modelo SAR**
   $$Y_i = \mu_i + \rho \sum_j w_{ij}(Y_j - \mu_j) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \text{ i.i.d.} \qquad \textbf{(M2)}$$
   - ρ é o parâmetro de auto-regressão (similar ao modelo AR em séries temporais)
   - F(variáveis explicativas)=μ_i=E[Y_i]=$\alpha + X_i\beta$, por exemplo, X_i latitude e longitude
4. Modelo de média móvel espacial, **modelo SMA**
   $$Y_i = \alpha + \beta_1 X_{1i} + \cdots + \beta_m X_{mi} + u_i$$
   $$u_i = \lambda \sum_j w_{ij} u_j + \epsilon_i \qquad \textbf{(M3)}$$
   correlação à custa dos vizinhos
   - u_i é um ruído correlacionado espacialmente
   - λ é o coeficiente de auto-regressão do erro (similar ao modelo MA em séries temporais)

Exemplo: dados das árvores – **Modelo SAR (M2)**

Considere o modelo **SAR**

$$\textbf{Altura}_i = \alpha + \beta\ \textbf{Diametro}_i + \rho \sum_j \textbf{w}_{ij}(\textbf{Altura}_j - \alpha - \beta \textbf{Diametro}_j) + \epsilon_i \qquad \textbf{(M2)}$$

então

	estimativa	s.e.	valor p
α	5.144	1.107	<0.001
β	0.304	0.023	<0.001
ρ	0.074	0.113	0.484

com erro padrão residual = 2.574 e AIC = 339.06.

- Embora o $\hat{\rho}$ não seja significativamente diferente de 0, **o novo termo** $\rho \sum_j (\dots)$ (modelando a correlação espacial) **permite resíduos independentes**.
- Além disso, o "erro padrão residual" em (M2) é ligeiramente menor do que em (M1).

moran.test(model_SMA$residuals, W, alternative = "two.sided")
#p-value = 0.9131

Exemplo: dados das árvores – **Modelo SMA (M3)**

Considere o modelo **SMA**

$$\textbf{Altura}_i = \alpha + \beta \textbf{Dimetro}_i + \textbf{u}_i$$
$$\textbf{u}_i = \lambda \sum_j \textbf{w}_{ij}\textbf{u}_j + \epsilon_i$$

ruido branco

então

	estimativa	s.e.	valor p
α	5.806	0.520	<0.001
β	0.310	0.022	<0.001
λ	0.266	0.175	0.137

com erro padrão residual = 2.526 e AIC = 337.55.

- Os resíduos são independentes (como desejado), mas a significância de $\hat{\lambda}$ é baixa (p-valor=0.137), então agora tentaremos integrar uma nova variável exploratória, **espécie**, no modelo de RL (M1).

Embora os parametros ró e lambda não sejam estatisticamente significativos, quando fazemos posteriormente o teste de morgan, referente aos resíduos vemos que não há correlação, ou seja, os residuos já são independentes.

Exemplo: dados das árvores – **Modelo RL (M1')**

Considere a covariável categórica **espécie**[3] no modelo de RL (M1)

$$\textbf{Altura}_i = \alpha + \beta\ \textbf{Diametro}_i + \textbf{fator}(\textbf{Especie}_i) + \epsilon_i \qquad \textbf{(M1')}$$

então

	estimativa	s.e.	valor p
α (Intercepto)	3.058	1.673	0.072
β (Diâmetro)	0.320	0.030	<0.001
Espécie2	2.262	1.605	0.163
Espécie3	3.910	1.311	0.004
Espécie4	0.790	1.792	0.660

A espécie 1 é a mais baixa (pq nenhuma das outras é negativa) a seguir é a especie 4

com erro s.e. residual = 2.354 e AIC = 325.317.

- Os resíduos ϵ_i tornaram-se **espacialmente não-correlacionados** (I de Moran =0.058, valor p=0.12).
- O modelo (M1') oferece o **menor AIC** e o menor "erro padrão residual", **então pode ser considerado o modelo mais ótimo**.

[3]Existem 4 espécies diferentes de árvores.