

Estatística Espacial

Exemplo - um mapa representando a quantidade de votos de Fernando Henrique em 1998.

O mapa é definido em seis classes (sextil)

Material

Bailey, T.C., Gatrell, A.C. Interactive spatial Data Analysis.

Assunção, R. M. Estatística Espacial em Epidemiologia, Economia, Sociologia. < [www.est.ufmg.br/leste/publicações.htm#20017](http://www.est.ufmg.br/leste/publicacoes.htm#20017)

Anselin, L. Exploring Spatial Data with GeoDa: A Workbook.
<http://geodacenter.asu.edu/system>

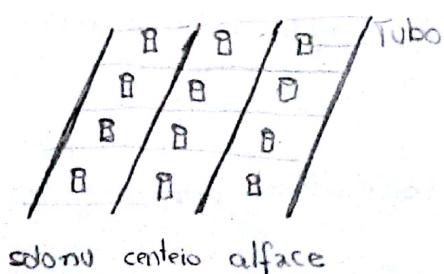
www.dpi.inpe.br/gilberto/analise

Tópico do Curso

1. Análise Espacial
2. Análise de Padrões em Dados de Áreas
3. Análise de Padrões Pontuais
4. Análise de Dados Espacialmente contínuos

Abordagem do Curso

Análise Espacial de Dados



Analizar a umidade do terreno

dist das árvores = 3,5 m

Tipo de Gráfico - Dendrograma

Tipos de Problemas

Epidemiologia - doença. Sua distribuição é igual no estado? Existem regiões com maior ocorrência? Há alguma associação com partes de poluição ou habitat ambiental?

Agricultura

Mapear a região de acordo com produtividade. Será que a produção é igual em toda a propriedade? Há partes da propriedade que tenham necessidade de demais fertilizantes?

Sociologia

Padrão espacial na distribuição dos assaltos. Roubos que ocorrem em determinada área estão relacionados com a questão socio-econômica?

Geologia

Dado um conjunto de amostras, qual a extensão de um depósito mineral?

Comércio

Analizar espacialmente a possibilidade de abrir novas lojas.

Exemplo

A epidemia de Colera em Londres - John Snow (1854)

Análise Espacial

Est. Convencional - Independência

Estat. Espacial - "Todas as coisas se parecem, coisas mais próximas são mais parecidas com aquelas distantes".
- correlação espacial

"Se onde é importante para o seu estudo, então a Análise Espacial é a sua ferramenta"

- estuda métodos científicos para a coleta, visualização e análise de dados que possuem coorden. geográficas".

enfase: mensurar propriedades e relacionamentos, levando em conta a localização espacial

Inferência Estatística para dados espaciais

- as inferências nesses tipos de dados não serão tão eficientes quanto no caso de amostras indepen. do mesmo tamanho
(Variancia maior - i.o maior)

Tipos de Dados

1. Dados de Superfície Aleatória

- v.a. contínua (temperatura)
- n. pontos de coleta de dados em localizações
- gerar uma superfície ($Z(x)$) - descrição de fen. de interesse.

. Krig (1951) - considerar a distância entre observações.

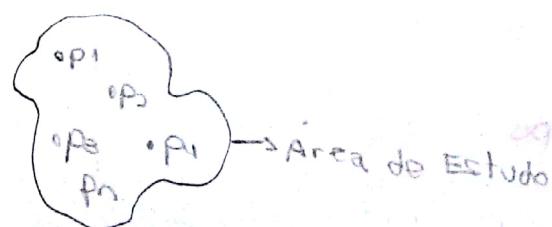
. Matheron (1971)

No inicio, geostatística era chamada agricultura da precisão - começou a ser utilizada por Austrália, EVA e depois pelo Brasil (países grandes) geostatística - consideração da localização geográfica e dependência espacial.

Dados de Processos Pontuais

Dados em que o principal interesse está no conjunto de coordenadas geográficas representando as localizações exatas de eventos

ex: ocorrência de dengue (varia apenas a posição)



Verificar se existe algum agrupamento ("cluster") ou regularidade.

Unidos

- Métodos - Estimador de Intensidade de Kernel.

Como se fosse colocar uma tigela sobre o ponto a ser estudado.



Dados de Área

- Localização está associada a áreas delimitadas por polígonos.
- Informações disponibilizadas por Ministério e Secretaria e geralmente são contíguas.
- Índices ou taxas, proporções, média ...

Exemplo:

Votos para Lula em 2006



Matriz de vizinhança

Máx. valor os elementos representam as distâncias entre as regiões.

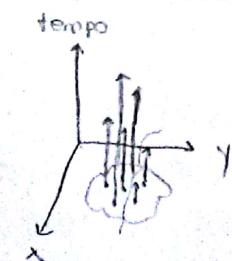
Dados de Intereração Espacial

Rota aérea

A largura da rota depende do fluxo de passageiros.

Exemplo: Migração, fluxo de passageiros

Intereração Espaço-Tempo



- ESTATCART - IBGE
- Diferentes Software - interface - TabWin, TerraView, GeoDa, Sas, R, etc
INFORMAT, crimestat

Peter Diggle.

www.dpi.inpe.br

www.est.ufmg.br

www.rc.unesp.br/igce/aplicada/textodi.html

09/08/2010

Correlação: mede a relação entre 2 v.a. (X e Y)

$$\rho_{xy} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}} = \frac{\text{Cov}(X,Y)}{\text{D.P}(X) \text{D.P}(Y)}$$

Covariância - Variância conjunta entre os v.a X, Y .

Variância - variação (espalhamento, dispersão) de uma variável em torno de sua média.

Obs: Usa-se más el coeficiente de relación pues estamos ahora trabajando en una misma medida y , además tenemos un valor entre -1 y 1.

- 61

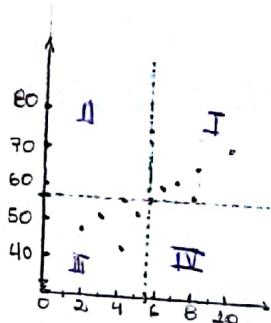
Ex.

$n = 10$ agentes de farmacia

X : años de servicio

Y : no de clientes.

Agente	A	B	C	D	E	F	G	H	I	J
X	2	3	4	5	4	6	7	8	9	10
Y	48	50	56	52	43	60	62	58	64	72



$$\begin{aligned}
 & \text{Covariância } (X, X) \\
 & = \text{Variância } (X) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \\
 & = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})
 \end{aligned}$$

$$\begin{aligned}
 \text{Covariância } (X, Y) & = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\
 \text{I} & + + + + = + \\
 \text{II} & - + + - = - \\
 \text{III} & + - - + = + \\
 \text{IV} & + - - - = -
 \end{aligned}$$

O cálculo de p_{xy} pode ser simplificado:

$$p_{xy} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sqrt{(\sum x_i^2 - n \bar{x}^2)(\sum y_i^2 - n \bar{y}^2)}}$$

pues

$$p_{xy} = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right] \left[\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \right]}} = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1} \sum (x_i - \bar{x}) \sum (y_i - \bar{y})}} =$$

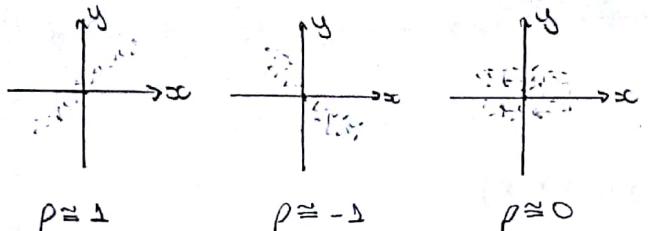
$$= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{1}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \cdot \frac{\sum_{i=1}^n \{ x_i y_i - \bar{x} y_i - \bar{y} x_i + \bar{x} \bar{y} \}}{\sum (x_i - \bar{x}) \sum (y_i - \bar{y})} \cdot \frac{I}{A}$$

$$I = \sum_{i=1}^n (x_i y_i - \bar{x} \bar{y} \sum_{j=1}^n y_j - \bar{y} \sum_{j=1}^n x_j + \bar{x} \bar{y} n) = \sum x_i y_i - \bar{x} \sum y_i - \bar{y} \sum x_i + \bar{x} \bar{y} n =$$

$$= \sum x_i y_i - \frac{n \bar{x} \bar{y} n}{n} - \bar{y} \bar{x} n + \sum \bar{x} \bar{y} = \sum x_i y_i - 2n \bar{x} \bar{y} + n \bar{x} \bar{y} = \sum x_i y_i - n \bar{x} \bar{y}$$

No exemplo $\rho_{xy} = 0,88$

Resumindo



No trabalho

correlação: umidade do solo

$\begin{cases} X: \text{umidade do solo} \\ Y: " " \end{cases}$

Como é uma única variável, temos autocorrelação

autocorrelograma \rightarrow gráfico de autocorrelação

hecho en trabajo.

Hay también el variograma - gráfico de varianzas

Es contrario al correlograma

Y el semivariograma - solo partitionar las varianzas por 2

Análise de Dados de Área

A localização está associada à áreas delimitadas por polígonos

(geralmente contagem)

polígonos (quadras, setores considerados, município, estado, país, etc.)

Forma usual de apresentação - uso de mapas coloridos com padrão espacial do fenômeno.

Após o mapa, ou deparar com algum padrão espacial:

- ele é aleatório?

- tem agrupamento definido?

- pode estar associado a causas mensuráveis?
- os valores observados são suficientes p/ analisar o fenômeno espacial de interesse?
- existem grupos de áreas c/ diferentes padrões na área de estudo?

Modelos de distribuição de dados de áreas

Considere γ_i como a v.a que descreve a contagem, indicador ou taxa associada à área A_i .

Temos um valor observado y_i .

A hipótese mais comum é supor que a v.a. $\gamma \sim \text{Poisson}$.

Outras distrib. podem ser mais adequadas, dependendo da variável de interesse.

As taxas podem ter distrib. normal.

Alguns cuidados.

problema de unidade de área

Ex: Censo Demográfico \rightarrow + de 500 variáveis

B. H \rightarrow análise c/ os setores censitários e unidades de planejamento (UP) SC

O estudo com 1000 correlações de pares de variáveis.

773 correlações são menores p/ SC que UP, e apenas 110 tem comportamento oposto.

X_i : nº de chefes de família com renda entre 0,5 e 1,5 mil

Y : nº de chefes de família c/ 1 a 3 anos de estudo.

$$P_{SC} = 0,79 \quad P_{UP} = 0,96$$

obs: UP contém mais dados que SC.

Verifica-se que a redução de escala (áreas maiores), tende a homogeneizar os dados, reduzir a flutuação aleatória e reforçar correlações que, assim, aparentam ser mais fortes que áreas menores.

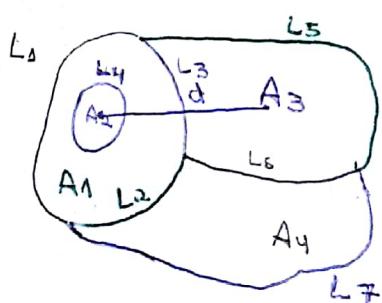
Não se pode afirmar qual escala seja "certa", mas apenas qual dos dois modelos serve melhor ao que se deseja esclarecer.

Proximidade Espacial

Para outros tipos de dados (pontuais e superfícies aleatórias), geralmente a proximidade é medida pela distância euclidiana.

A principal diferença p/ dados de área está na formalização da proximidade espacial.

Exemplos de medida



A) Proporção da fronteira pelo perímetro

$$w_{14} = \frac{L_2}{L_1 + L_2 + L_3 + L_6 + L_7} \quad w_{23} = 0$$

$$w_{12} = \frac{L_4}{L_1 + L_2 + L_3 + L_4}$$

B) distância linear entre centroides dos objetos

$$w_{23} = \begin{cases} 0 & \text{para } d \geq \limiar \\ 1 & \text{... } d \leq \limiar \end{cases}$$

C) Inverso da Distância Linear

$$w_{23} = \frac{1}{d}$$

D) Existência de fronteira comum

$w_{34} = 1$, A_3 faz fronteira com A_4

$w_{24} = 0$, A_2 não tem II com A_4

Constrói-se a matriz de proximidade espacial

$$W = \begin{bmatrix} w_{11} & w_{12} & w_{13} & w_{14} \\ w_{21} & w_{22} & w_{23} & w_{24} \\ w_{31} & w_{32} & w_{33} & w_{34} \\ w_{41} & w_{42} & w_{43} & w_{44} \end{bmatrix}$$

w_{ij} : "distância" do objeto i ao objeto j , às vezes, os valores w_{ij} são "normalizados", a soma dos elementos de uma linha da matriz.

$$W = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \quad \text{- Usando a medida D, temos:}$$

Padronizando

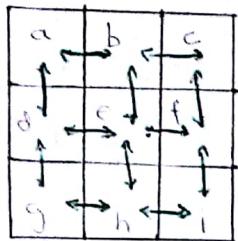
$$W^* = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ \frac{1}{3} & 0 & \frac{1}{2} & 0 \end{bmatrix}$$

Há também 3 tipos de definições de contiguidade para dados espaciais: torre, bispo e rainha.

Localizações

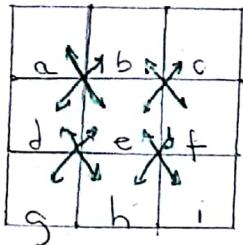
a	b	c
d	e	f
g	h	i

Contiguidade da Torre (Rook's Case)



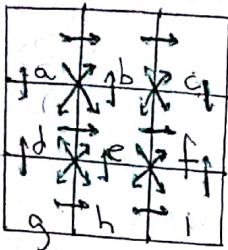
	a	b	c	d	e	f	g	h	i
a	0	1	0	1	0	0	0	0	0
b	1	0	1	0	1	0	0	0	0
c	0	1	0	0	0	1	0	0	0
d	1	0	0	0	1	0	1	0	0
e	0	1	0	1	0	1	0	1	0
f	0	0	1	0	1	0	0	0	1
g	0	0	0	1	0	0	0	1	0
h	0	0	0	0	0	1	0	1	0
i	0	0	0	0	0	1	0	1	0

Contiguidade do Bispo (Bishop's Case)



	a	b	c	d	e	f	g	h	i
a	0	0	0	0	1	0	0	0	0
b	0	0	0	1	0	1	0	0	0
c	0	0	0	0	1	0	0	0	0
d	0	1	0	0	0	0	0	0	1
e	1	0	1	0	0	0	1	0	1
f	0	1	0	0	0	0	0	1	0
g	0	0	0	1	0	0	0	0	0
h	0	0	0	0	1	0	1	0	0
i	0	0	0	0	1	0	0	0	0

Contiguidade da Rainha (Queen's Case)



	a	b	c	d	e	f	g	h	i
a	0	1	0	1	1	0	0	0	0
b	1	0	1	1	1	1	0	0	0
c	0	1	0	0	1	1	0	0	0
d	1	1	0	0	1	0	1	1	0
e	1	1	1	1	0	1	1	1	1
f	0	1	1	0	1	0	0	1	1
g	0	0	0	1	1	0	0	1	0
h	0	0	0	1	1	1	1	0	1
i	0	0	0	0	1	1	0	1	0

Medidas Básicas de Autocorrelação

Testes para dados qualitativos nominais

classificação binária \rightarrow cada região é codificada como P(preta) ou B(branca)

$$Z_i = \begin{cases} 1, & \text{se a região é P} \\ 0, & \text{se a " " é B} \end{cases}$$

P \rightarrow se possui a característica de interesse

B \rightarrow se não " " "

As contiguidades possíveis são:

(BP, BB, PP, PB) \rightarrow consideramos PB = BP

Quando tiver n° grande de vizinhos

PP \rightarrow agrupamento ou cluster (correlação positiva)

Quando tiver n° grande de vizinhos PB \rightarrow padrões alternados, correlação negativa.

A estatística do produto cruzado geral é dada por:

$$r = \sum_i \sum_j w_{ij} y_{ij}$$

$w \Rightarrow \{w_{ij}\}$ é a medida de proximidade espacial dos locais i e j.

$y \Rightarrow \{y_{ij}\}$ é a medida de proximidade de i e j em alguma outra dimensão.

Se os valores observados em locais i e j são x_i e x_j é comum usar

$$y_{ij} = (x_i - x_j)^2$$

Ex: Os valores de x

a	b	c
d	1	1
g	1	1

$$Y = \begin{matrix} a & b & c & d & e & f & g & h & i \\ \left[\begin{array}{ccccccccc} 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \end{array} \right] \end{matrix}$$

Como só há 8 elementos em comum = 1, temos que:

$$r = \sum_i \sum_j w_{ij} y_{ij} = 8$$

Fazer outras permutações

1	0	1
1	0	0
1	1	1

$$C_{9,4} = \frac{9!}{5! 4!} = 126$$

Nosso trabalho

a	b	c
d	e	f
g	h	i

$$Y = a \begin{bmatrix} a & b & c & d & e & f & g & h & i \\ b & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ c & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ d & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ e & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ f & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ g & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ h & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ i & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Assim

$$\sigma = \sum_i \sum_j w_{ij} y_{ij} = 14$$

16/08/2010

No trabalho - faltou realizar o test de hipóteses

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$



$$p_{obs} = \infty$$

$$p(|\rho| > \infty) = p\text{-valor}$$

Upton e Fingleton (1985) $\rightarrow +98$ permutações aleatórias

No caso do trabalho, só se pode obter correlação par

Tabela 2 - Distribuição de r obtida de 99 permutações aleatórias da matriz Y .

	8	10	12	14	16	18	20	22	24	Total
freq	8	11	25	25	7	10	0	2	99	

A tabela mostra que, enquanto o valor observado $r=8$ é um valor extremo, porém ocorreu em. Portanto, não suficiente para ser julgado significativo.

Para valores extremos e muito menos provável é o resultado $r=24$, que surge da permutação



Cliff e Ord (1981) → Aproximação Normal

$$E(r) = \frac{S_0 T_0}{n(n-1)}$$

$$\text{Var}(r) = \frac{S_1 T_1}{2n^{(2)}} + \frac{(S_2 - 2S_1)(T_2 - 2T_1)}{4n^{(3)}} + \frac{(S_0^2 + S_1 + S_2)(T_0^2 + T_1 - T_2)}{n^{(4)}} - [E(r)]^2$$

$$S_0 = \sum_{i+j} w_{ij}$$

$$S_1 = \frac{1}{2} \sum_{i+j} (w_{ij} + w_{ji})^2$$

$$S_2 = \sum_i (w_{i0} + w_{0i})^2$$

$$w_{i0} = \sum_j w_{ij} \quad w_{0i} = \sum_j w_{ji}$$

$$n^{(2)} = n(n-1)$$

$$n^{(3)} = n(n-1)(n-2) \quad n^{(4)} = n(n-1)(n-2)(n-3)$$

As fórmulas p/ T_0 , T_1 e T_2 são idênticas na forma às fórmulas p/ S_0 , S_1 e S_2 , com W sendo substituído por γ .

S_0 = soma de todos os elementos = 24

$$S_1 = \frac{1}{2} ([1+1]^2 + [1+1]^2 + \dots + [1+1]^2) = \frac{1}{2} 24(4) = 48$$

$$W_{10} = \sum_j w_{1j} = w_{11} + w_{12} + \dots + w_{1n} = 2$$

$$w_{01} = \sum_j w_{j1} = w_{11} + w_{21} + w_{31} + \dots + w_{n1} = 2$$

$$\begin{aligned} S_2 &= (2+2)^2 + (3+3)^2 + (2+2)^2 + (3+3)^2 + (4+4)^2 + (3+3)^2 + (2+2)^2 + (3+3)^2 + (2+2)^2 \\ &= 16 + 36 + 16 + 36 + 64 + 36 + 16 + 36 + 16 = 272 \end{aligned}$$

Repetir-se, usando a matriz γ .

$$T_0 = \sum_{i \neq j} \gamma_{ij} = 40$$

$$T_1 = \frac{1}{2} \sum_{i \neq j} \sum (y_{ij} + \gamma_{ji})^2 = \frac{1}{2} ([1+1]^2 + [1+1]^2 + \dots + [1+1]^2) = \frac{1}{2} 40(4) = 80$$

$$T_2 = \sum_i (\text{total da linha} + \text{total da coluna})^2 =$$

$$= (4+4)^2 + (4+4)^2 + (4+4)^2 + (4+4)^2 + (5+5)^2 + (5+5)^2 + (4+4)^2 + (5+5)^2 + (5+5)^2$$

$$= 5(8)^2 + 4(10)^2 = 5.64 + 4.100 = 720$$

Assim

$$E(n) = \frac{S_0 T_0}{n(n-1)} = \frac{24 \cdot 40}{9 \cdot 8} = \frac{40}{3} = 13,333$$

$$\text{Var}(n) = \frac{S_1 T_1}{2n^{(2)}} + \frac{(S_2 - 2S_1)(T_2 - 2T_1)}{4n^{(3)}} + \frac{(S_0^2 + S_1 + S_2)(T_0^2 + T_1 - T_2)}{n^{(4)}} - [E(n)]^2$$

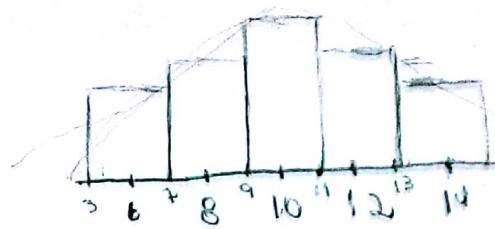
$$= \frac{48 \cdot 80}{2 \cdot 9 \cdot 8} + \frac{(272 - 96)(720 - 160)}{4 \cdot 9 \cdot 8 \cdot 7} + \frac{(24^2 + 48 - 272)(40^2 + 80 - 720)}{9 \cdot 8 \cdot 7 \cdot 6} - [13,33]^2$$

$$26,667 + 48,889 +$$

$$= 9,52$$

O valor observado de $\sigma = 8$

$$z = \frac{|n - E(n)| - 1}{\sqrt{\text{Var}(n)}}$$



n assume somente valores pares.

$$z = \frac{|8 - 13,33| - 1}{\sqrt{9,52}} = 1,40$$

$$\text{p-valor} \rightarrow P(z > 1,40) = \text{p-valor} = 2(0,5 - P(0 \leq z \leq 1,4)) = 2(0,5 - 0,46076) = 0,13848$$

A probabilidade de uma variável normal padrão exceder 1,40 é igual a 16,1%. Isso é muito próximo da probabilidade empírica de 16,5% derivada das simulações sumariadas na Tabela. O resultado é muito próximo da probabilidade empírica de 16,5%.

discreto (localização no canto sudeste) não é um caso raro. A estatística apresentada anteriormente vale para todas as situações (não só categoricas binárias).

No caso de dados binários

Seja

$p = \text{nº de caselas pretas (ou c/s)}$

$b = \text{nº de caselas brancas (ou c/N)}$

$n = \text{nº total de caselas}$

$$S_0 = 2(2l - l - c)$$

$$S_1 = 2S_0$$

$$S_2 = 8(8lc - 7l - 7c + 4)$$

$l = \text{nº de linhas}$

$c = \text{nº de colunas}$

$$T_0 = 2pb$$

$$T_1 = 2T_0$$

$$T_2 = 4npb = nT_1$$

Notas Tabwin

- DATASUS: www2.datasus.gov.br/DATASUS/index.php

└ Sistemas e Aplicativos

 └ Tabulação

 └ Tabwin

 └ Download mapas: mapbr.zip

└ Informações

 └ Estatísticas Vitais

Arquivo

1 → Abrir / Importar mapas

2 → Incluir tabela

Operações (calcular taxa)

 └ Calcular Indicador

① - constrói mapas

 └ A - legenda

24/08/2020

Contagem de Juncções

Caso particular: resposta binária.

VIII	VII
VI	V
IV	III
II	I

continua → 2 categorias

ex: acima da mediana (1) - (Preto)

abaixo da mediana (0) - (Branco)

discreta → Transforma em 2 categorias → a de maior interesse (1) e o restante 0.

Dessa forma, as possíveis junções são: PP, PB e BB.

Quando tiver um nº grande de vizinhos PP \rightarrow agrupado (corr. positiva)
" " " " " " " " PB \rightarrow padrões alternativos (corr. negativa)

O nº de junções PB é $r/2$, de acordo com o $r = \sum_{i,j} w_{ij} y_{ij}$, obtido da matriz W e Y, este último definido como $y_{ij} = (x_i - x_j)^2$.

O nº de junções PP é $r^*/2$, r^* é o valor de r quando $y_{ij} = x_i x_j$.

Se o nº total de junções no sistema for J \Rightarrow o nº de junções BB é $J - PB - PP$.

Podemos usar as três estatísticas como indicadores de autocorrelação, mas os resultados de Cliff e Ord (1981) sugerem que PB é mais informativo.

Desde que o nº de junções PB é dado por $r/2$, a média $E(PB) = E(r/2) = \frac{E(r)}{2}$ e a variância $V_r(PB) = V_r\left(\frac{r}{2}\right) = \frac{Var(r)}{4}$.

E como já tínhamos resultado de r.

$$E(PB) = \frac{1}{2} \frac{S_0 T_0}{n(n-1)}$$

No caso particular binário e se $l=c$ (área quadrada)

$$S_0 = 2(2lc - l - c) = 2(2l^2 - l - l) = 2(2l^2 - 2l) = 2 \cdot 2l(l-1) \Rightarrow$$

$$\Rightarrow S_0 = 4l(l-1)$$

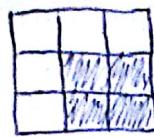
$$\text{É como } T_0 = 2pb$$

$$\Rightarrow E(PB) = \frac{1}{2} \frac{4l(l-1)(2pb)}{l^2(l^2-1)}, \text{ pois } n = l \cdot c = l^2,,$$

$$\Rightarrow E(PB) = \frac{4l(l-1)pb}{l^2(l+1)(l-1)} = \frac{4pb}{l(l+1)} //$$

A distribuição de $(X_1 - X_2)^2$ depende

no ex. da aula



$$E(PB) = \frac{4 \times 4 \times 5}{3 \cdot 4} = \frac{20}{3} = 6,7$$

$$E(\pi) = 13,84$$

Também da mesma forma é possível obter:

$$E(PP) = \frac{2p(p-1)}{l(l+1)}$$

Obs: A aproximação normal é adequada somente quando n , np e $n(1-p)$ não forem muito pequenos, em que p é a proporção de localidades pretas no total de n localidades.

Em outras situações, usar a distribuição de aleatorização (permutações aleatórias) ou distribuição de Poisson se a esperança for quase igual à variância ($E(PB) \approx \text{Var}(PB)$) → maiores detalhes em Upton e Fingleton (1985).

Média Móvel Espacial

Uma forma bastante simples de explorar a variação da tendência espacial na análise de padrões de área é o cálculo da média com os valores dos vizinhos de primeira ordem.

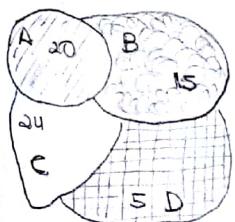
→ uma aproximação da variabilidade espacial, pois a operação produz uma superfície menos descontínua (mais suave) que os dados originais i. podendo ainda apresentar locais de transições entre regimes espaciais.

Considerando a matriz de proximidade espacial W , a estimativa de média móvel espacial pode ser:

$$\hat{\mu}_i = \frac{\sum_{j=1}^n w_{i,j} y_j}{\sum_{j=1}^n w_{i,j}}, \quad i=1, 2, \dots, n$$

em que y_j : valor do atributo considerado em cada área j .
 n : nº de polígonos (áreas).

Exemplo:



$$\text{Amplitude} = \text{Máx} - \text{Min} = 20 - 5 = 15$$

Dividir em 4 classes.

até 10 \rightarrow amarelo (C1)

10 \rightarrow 15 \rightarrow azul (C2)

15 \rightarrow 20 \rightarrow vermelho (C3)

+ de 20 \rightarrow verde (C4)

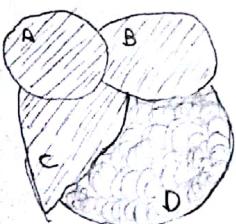
$$\mu_A = \frac{20 + 15 + 24}{3} = \frac{59}{3} = 19,67, \quad (\text{C3})$$

$$\mu_B = \frac{20 + 15 + 24 + 5}{4} = \frac{64}{4} = 16, \quad (\text{C3})$$

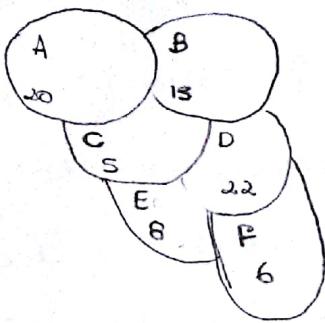
$$\mu_C = \frac{20 + 15 + 24 + 5}{4} = \frac{64}{4} = 16, \quad (\text{C3})$$

$$\mu_D = \frac{15 + 24 + 5}{3} = \frac{44}{3} = 14,67, \quad (\text{C2})$$

Depois



Ou em forma de produto de matrizes



$$W = \begin{bmatrix} A & B & C & D & E & F \\ A & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ B & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ C & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 \\ D & 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ E & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ F & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \quad Y = \begin{bmatrix} 20 \\ 15 \\ 5 \\ 22 \\ 8 \\ 6 \end{bmatrix}$$

Logo

$$\hat{\mu} = WY = \begin{bmatrix} 53,33 \\ 45,20 \\ 34,00 \\ 55,20 \\ 10,25 \\ 32,00 \end{bmatrix}$$

Indicadores Globais de Autocorrelação Espacial

Índice de Moran e de Geary

Vamos explorar a dependência espacial dos desvios dos valores dos atributos em relação ao valor médio, ou seja, os efeitos de 2º ordem.

Autocorrelação Espacial = quando o valor observado de um atributo em uma região é independente dos valores dessa mesma variável nas localizações vizinhas → ela mede o nível de interdependência geográfica entre as variáveis e a natureza e a força desse relacionamento.

As duas medidas mais utilizadas são o índice global I de Moran e índice C de Geary.

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\left[\sum_{i=1}^n (y_i - \bar{y})^2 \right] \left[\sum_{i \neq j} w_{ij} \right]}$$

n = nº de áreas em estudo

y_i = valor do atributo considerado na área i .

\bar{y} = valor médio do atributo na região de estudo.

w_{ij} = pesos atribuídos conforme a conexão entre as áreas i e j .

O índice de Moran testa se as áreas conectadas apresentam maior semelhança quanto ao indicador estudado do que o esperado num padrão aleatório.

A hipótese nula é de completa aleatoriedade espacial, quando o indicador se distribui ao acaso entre as áreas sem relação com a posição.

Uma estatística fortemente relacionada ao índice de Moran é o índice C de Geary.

$$C = \frac{(n-1) \sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - y_j)^2}{2 \left[\sum_{i=1}^n (y_i - \bar{y})^2 \right] \left[\sum_{i \neq j} w_{ij} \right]}$$

sendo as mesmas variáveis e valores definidos em I de Moran.

Nenhuma das 2 estatísticas é obrigada a situar na faixa entre $(-1, 1)$, como é o caso das correlações não espaciais convencionais.

Quando as variáveis são espacialmente independentes, o valor esperado de I é próximo de zero.

Quando há similaridade entre localidades próximas, o índice I tende a ser positivo, e tende a ser negativo quando as

localidades próximas são dissimilares.

Quando não há autocorrelação presente \Rightarrow

$$E(I) = \frac{-1}{n-1} \quad \text{e} \quad E(C) = 1$$

Quando a autocorrelação positiva máxima estiver presente
C estará próximo de 0 e I aproximará de 1.
 \therefore O índice I tem o comportamento mais similar ao tradicional coeficiente de correlação linear que C de Geary.

Sob a hipótese nula (não existe dependência espacial entre as localidades).

$$I \sim N(E(I), \text{Var}(I)), \quad n \geq 20$$

$$\text{Var}(I) = \frac{n \{(n^2 - 3n + 3)s_1 - ns_2 + 3s_0^2\} - k \{n(n-1)s_1 - 2ns_2 + 6s_0^2\}}{(n-1)^{(3)} s_0^2}$$

$$= \frac{1}{(n-1)^2}$$

$k = \frac{m_4}{m_2^2}$: coeficiente de curtose, $m_n = \frac{1}{n} \sum (x_i - \bar{x})^n$

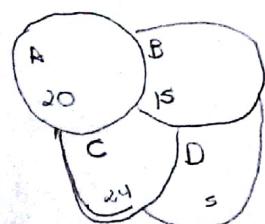
(Upton e Fingleton (1985))

e segundo INPE,

$$\text{Var}(I) = \frac{n^2(n-1)s_1 - n(n-1)s_2 - 2s_0^2}{(n+1)(n-1)^2 s_0^2}$$

$$\frac{641 \cdot 68 \cdot 3 \cdot 3}{6 \cdot 5 \cdot 11 \cdot 10} = 1$$

Exemplo: Consideremos os exemplos anteriores.



Obtenha I de Moran

matriz de proximidade

$$\begin{matrix} & A & B & C & D \\ A & 0 & 1 & 1 & 0 \\ B & 1 & 0 & 1 & 1 \\ C & 1 & 1 & 0 & 1 \\ D & 0 & 1 & 1 & 0 \end{matrix}$$

Precisamos:

$$\bar{y} + \frac{20+15+24+5}{4} = 16 = \mu$$

$$\text{Variância: } \sigma^2 = \frac{\sum_{i=1}^n (y_i - \mu)^2}{n} = \frac{(20-16)^2 + (15-16)^2 + (24-16)^2 + (5-16)^2}{4} =$$
$$= \frac{16+1+64+121}{4} = \frac{202}{4} = 50,5$$

$$\sigma = \sqrt{\text{Var}(Y)} = \sqrt{50,5} = 7,106$$

A equação de \mathbb{J} pode ser simplificada se padronizarmos as observações e alterarmos a matriz W , de forma que a soma dos elementos de cada linha seja igual a 1.

$$W = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/3 & 0 & 1/3 & 1/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 0 & 1/2 & 1/2 & 0 \end{bmatrix}$$

$$\sum_{i \neq j} \sum w_{ij} = 4 = n$$

$$\text{Padronização: } \frac{y_i - \mu}{\sigma} = z_i$$

$$\Rightarrow \mathbb{J} = \frac{\sum_{i=1}^n \sum_{j=1}^n z_i z_j w_{ij}}{\sum_{i=1}^n z_i^2} \quad (\text{A})$$

$$z_A = \frac{20-16}{7,106} = \frac{4}{7,106} = 0,553$$

$$z_B = \frac{15-16}{7,106} = \frac{-1}{7,106} = -0,141$$

$$z_C = \frac{24-16}{7,106} = \frac{8}{7,106} = 1,126$$

$$z_D = \frac{5-16}{7,106} = -1,548$$

$$\begin{aligned}
 (A) &= \frac{(0,563)(-0,141)}{2} + \frac{(0,563)(1,126)}{2} + \frac{(0,563)(-0,141)}{3} + \frac{(-0,141)(1,126)}{3} + \\
 &+ \frac{(-0,141)(-1,548)}{2} + \frac{(0,563)(1,126)}{3} + \frac{(-0,141)(1,126)}{3} + \frac{(1,126)(-1,548)}{3} \\
 &+ \frac{(-0,141)(-1,548)}{2} + \frac{(1,126)(-1,548)}{2} = -0,914
 \end{aligned}$$

B)

$$\sum z_i^2 = 4 = n, \text{ pois } \sum z_i^2 = \sum \left(\frac{z_i - \bar{z}}{\sigma} \right)^2 = n \sum \frac{(x_i - \bar{x})^2}{\sigma^2} = n, \Rightarrow I = -0,228$$

$$I = -0,228.$$

Usando matrizes:

$$W_{ij}(z_i z_j) = W_{ij} \underbrace{Z_i Z_j^T}_{\text{em que}}$$

$$Z = \begin{bmatrix} z_A \\ z_B \\ z_C \\ z_D \end{bmatrix}$$

→ somar-se os elementos p/ obter o numerador.

n é um valor pequeno → permutações aleatórias $4! = 24$ cenários possíveis.

A = 20	A = 20	A = 50	A = 5	A = 24	A = 15	15	24	5
B = 15	B = 15	B = 5	B = 15	B = 15	B = 20	5	20	24
C = 24	C = 5	C = 24	C = 20	C = 20	C = 24	20	15	20
D = 5	D = 24	D = 15	D = 24	D = 5	D = 5	24	5	5



Nossos

- A = 5
- B = 15
- C = 20
- D = 24

30/10/2020

08

Obs:

ZZ' = multiplicação de matrizes = W

$W \cdot M$ = multiplicando W e M (elemento por elemento)

$$W \cdot M = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} 0,3167 & -0,0792 & 0,6335 & -0,8711 \\ -0,0792 & 0,0197 & -0,1583 & 0,2177 \\ 0,6335 & -0,1583 & 1,2672 & -1,7424 \\ -0,8711 & 0,2177 & -1,7424 & 2,3959 \end{bmatrix} =$$

$$= \begin{bmatrix} 0 & -0,0396 & 0,31675 & 0 \\ -0,0264 & 0 & -0,0527 & 0,07256 \\ 0,2112 & -0,0527 & 0 & -0,5868 \\ 0 & 0,10885 & -0,8712 & 0 \end{bmatrix}$$

$$\sum_i \sum_j (wm)_{ij} = A \quad e \quad I = \frac{A}{\sum_j z_j^2}$$

Distribuição das permutações aleatórias

Existem no total $4! = 24$ permutações possíveis (permutando os valores nas regiões)

→ ABCD, ABDC, ADBC, ~~B~~ADBC, DABC, ACBD, CABD, BACD, ...

A(20), B(15), C(5), D(24) → I = -0,557

A(5), B(24), C(20), D(15) → I = -0,488

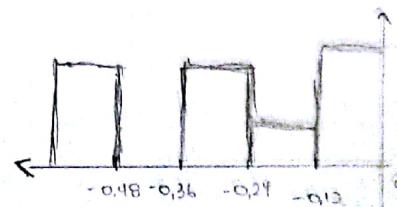
A(15), B(20), C(24), D(5) → I = -0,48865

A(20), B(5), C(24), D(15) → I = -0,3276

A(5), B(15), C(20), D(24) → I = -0,0533

:

Depois fazer um histograma



Outra medida de autocorrelação espacial.

C. de Geary

$$C = \frac{(n-1) \sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - y_j)^2}{2 \left[\sum_{i=1}^n (y_i - \bar{y})^2 \right] \left[\sum_i \sum_j w_{ij} \right]}$$

Obtenha C.

$$C = \frac{3 \left[\frac{1}{2}(25) + \frac{1}{2}(16) + \frac{1}{3}(25) + \frac{1}{3}(81) + \frac{1}{3}(100) + \frac{1}{3}(16) + \frac{1}{3}(81) + \frac{1}{3}(361) + \frac{1}{2}(100) + \frac{1}{2}(361) \right]}{2 [202] [4]}$$

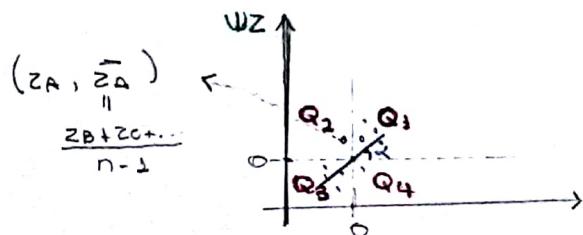
$$C = \frac{3(472,333333)}{1616} = 0,876685644,$$

$$\text{e } I = -0,228$$

$C \rightarrow 0$, autocorrelação espacial

$C \rightarrow 1$, não existe " " " .

Diagrama de Espalhamento de Moran



Relacionamento entre os valores do vetor de desvios Z e os valores das médias locais WZ .

$$I = \frac{\sum WZ}{\sum Z} \rightarrow \text{coeficiente de regressão linear (inclinação da reta).}$$

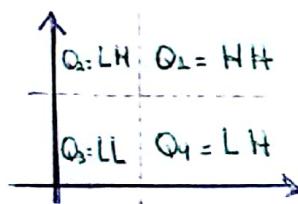
Q_1 (valores +, médias +) e Q_3 (valores -, médias -)

Indicam pontos de associação espacial positiva \Rightarrow uma localização possui vizinhos com valores semelhantes.

Q_4 (valores +, médias -) e Q_2 (valores -, médias +).

Indicam pontos de associação espacial negativa \Rightarrow uma localização possui vizinhos com valores distintos.

O diagrama de espalhamento de Moran pode ser apresentado na forma de um mapa \rightarrow coroplótico (mapa em que a intensidade de uma pré-escolhida cor determina a frequência), bidimensional, no qual cada polígono é apresentado indicando-se seu quadrante no diagrama.



$Q_1 = \text{alto-alto}$
 $Q_2 = \text{baixo-alto}$
 $Q_3 = \text{baixo-baixo}$
 $Q_4 = \text{alto-baixo}$

—
Laboratório

GeoDA

File

L Open project
L Search.shp

! Key Variable \rightarrow valor chave

Hacer el mapa coroplótico \rightarrow Mapa
L Quantile

Podemos duplicar el número de mapas y para esto se hace
Edit - Duplicate Maps o pulsar en ícono correspondiente.

Índices Locais de Associação Espacial

- Índice global fornece um único valor para toda região de estudo (única medida de associação espacial).
- Mas muitas vezes temos interesse em examinar padrões em uma escala maior.
- Índices de associação espacial → cada localização tem a sua medida.

0,000179

Os índices locais:

- permitem avaliar diferentes regimes espaciais existentes na área de estudo;
- medem a associação espacial entre uma observação i e a sua vizinhança.

Requisitos:

- A soma dos índices locais deve ser proporcional ao índice global
- Deve indicar a significância da associação espacial para cada observação.

$$I \propto \sum_{i=1}^n I_i \quad \left(I = \frac{\sum I_i}{n} \right)$$

proporcional

$$I_i = \frac{(y_i - \bar{y}) \sum_{j=1}^n (y_j - \bar{y}) w_{ij}}{\sum_{k=1}^n (y_k - \bar{y})^2 / n}, \quad i = 1, 2, \dots, n$$

Dados padronizados

$$I_i = \frac{\sum_{j=1}^n w_{ij} z_j}{\sum_{j=1}^n z_j / n}$$

Esses índices produzem um valor específico para cada área possibilitando: a identificação de agrupamentos (clusters = objetos com valores de atributos semelhantes), dados que tem maior impacto na medida global, outliers (objetos anormais), etc.

$I_i > 0$ "clusters de valores similares (altos ou baixos)

$I_i < 0$ " " " " distintos (ex: uma área com valor alto rodeada por uma vizinhança com valores baixos).

De forma similar aos indicadores globais a significância do índice local de Moran (I_i) deve ser avaliada utilizando a hipótese de normalidade ou simulação de distribuição por permutação aleatoria nos valores dos atributos. $\Rightarrow I_i$ podem ser classificados como não significantes ou significantes com nível de:

5%	1%	0,1%
95%	99%	99,9%
1,96	2,54	3,29

Uma vez determinada a significância estatística de I_i de Moran é muito útil gerar um mapa indicando as regiões que apresentam correlação espacial significativamente diferente do resto dos dados. Este mapa é denominado LISA Map.

Indicadores Logais G_i e G_i^* (Getis e Ord)

19

$$G_i = \frac{\sum_{j=1}^n w_{ij} y_j}{\sum_{k=1}^n y_k}, j \neq i$$

$$G_i^* = \frac{\sum_{j=1}^n w_{ij} y_j}{\sum_{k=1}^n y_k}, j \neq i$$

$G_i \rightarrow$ Soma dos dados da vizinhança de i relativa à soma de todos os dados excluindo y_i .

$G_i^* \rightarrow$ Idem a G_i , considerando todos os valores (inclusive y_i) no denominador.

São medidas de agrupamento na vizinhança ao redor de i .

Valores altos da estatística G indicam alta concentração espacial (agrupamento).

G_i e G_i^* têm aproximadamente distribuição normal

$$E(G_i) = \frac{w_i}{n-1} \quad \text{e} \quad E(G_i^*) = \frac{w_i^*}{n}$$

sendo

$$w_i = \sum_{j=1}^n w_{ij} \quad \text{e} \quad w_i^* = \sum_{j=1}^n w_{ij}$$

As variâncias são

$$\text{Var}(G_i) = \frac{w_i(n-1-w_i)}{(n-1)^2(n-2)} \left[\frac{s(i)}{\bar{x}(i)} \right]^2,$$

$$\text{Var}(G_i^*) = \frac{w_i^*(n-w_i^*)}{n^2(n-1)} \left[\frac{s}{\bar{x}} \right]^2$$

$$\bar{x}(i) = \frac{\sum_{j=1, j \neq i}^n x_j}{n-1}, \quad s^2(i) = \frac{\sum_{j=1, j \neq i}^n x_j^2}{n-1} - [\bar{x}(i)]^2.$$

Obs: A variável observada está sendo denotada por y_j nas expressões da variância substituir por essa variável.

Também podem ser definidas em função da distância ~~de~~
d entre as localidades (centroíde)

$$G(d) = \frac{\sum_{j=1, j \neq i}^n w_{ij}(d)y_j}{\sum_{k=1, k \neq i}^n y_k}, \quad G^*(d) = \frac{\sum_{j=1, j \neq i}^n w_{ij}(d)y_j}{\sum_{k=1}^n y_k}.$$

Gráfico de d x G(d) ou d x G*(d).

→ Tem vários artigos utilizando esta técnica, especialmente em Sensoriamento Remoto.

A estatística geral G da associação espacial global é dada por:

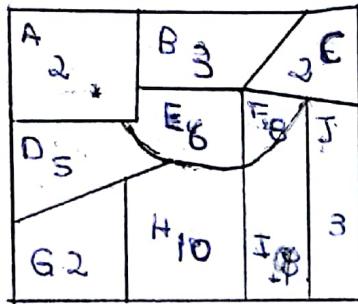
$$G = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}y_i y_j}{\sum_{i=1}^n \sum_{j=1}^n w_{ij}}, \quad \forall i \neq j$$

$G \in N(E(G), V(G))$

$$E(G) = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}}{n(n-1)}, i \neq j$$

e a obtenção de $V(G)$ envolve muitos cálculos.

Ex:



I_i , G_i e G_i^*

$$\bar{y} = \frac{\sum y_i}{n} = \frac{(2+5+2+3+6+10+8+8+2+3)}{10} = \frac{49}{10} = 4,9$$

$$W = \begin{pmatrix} a & b & c & d & e & f & g & h & i & j \\ a & 0 & \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ b & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 \\ c & 0 & \frac{1}{5} & 0 & 0 & \frac{1}{5} & \frac{1}{5} & 0 & 0 & \frac{1}{5} \\ d & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & 0 \\ e & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & 0 & \frac{1}{7} & 0 & \frac{1}{7} & \frac{1}{7} \\ f & 0 & \frac{1}{6} & \frac{1}{6} & 0 & \frac{1}{6} & 0 & 0 & \frac{1}{6} & \frac{1}{6} \\ g & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 \\ h & 0 & 0 & 0 & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & 0 & \frac{1}{8} & 0 \\ i & 0 & 0 & \frac{1}{8} & 0 & \frac{1}{8} & \frac{1}{8} & 0 & \frac{1}{8} & 0 \\ j & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \end{pmatrix}$$

$$z_1 = \underline{2.4, 9}$$

1426

$$0 = \sum_{i=1}^n (y_i - \bar{y})^2 \leq (2-4,9)^2 + (3-4,9)^2 + \dots + (3-4,9)^2$$

$$0 = 2,809$$

$$z_A = \frac{2-4,9}{2,809} = -1,0323$$

$$z_B = \frac{3-4,9}{2,809} = -0,6367$$

cont. exemplo

$$I_A = \frac{n(y_A - \bar{y})}{\sum_{j=1}^n w_{ij}(y_j - \bar{y})} \sim A$$
$$\sum_{j=1}^n (y_j - \bar{y})^2 \sim B$$

A

$$(y_A - \bar{y}) \sum_{j=1}^n w_{ij}(y_j - \bar{y}) =$$

$$(2 - 5) \frac{1}{3} [(3 - 5) + (5 - 5) + (6 - 5)] = +11,6 \approx$$

$$B = \frac{\sum (y_j - \bar{y})^2}{n} = \frac{86}{10} = 8,6$$

Logo

$$I_A = \frac{-1}{8,6} = +0,1163 \approx$$

$$I_G = \frac{n(y_0 - \bar{y}) \sum_{j=1}^n w_{0j}(y_j - \bar{y})}{\sum (y_j - \bar{y})^2}$$

$$= \frac{10(-3) \cdot \frac{1}{2} [(5-5) + (10-5)]}{86} = -0,8721$$

$$I_H = \frac{10(5) \frac{1}{5} [(5-5) + (6-5) + (8-5) + (2-5) + (9-5)]}{86}$$

$$= \frac{10(5)}{86} = \frac{50}{86} = 0,5814$$

Quando nos deparamos com a sigla,

HH

HL

LH

LL

queremos dizer que a região que estudamos possui, no caso HL por exemplo, valor alto e sua vizinhança possui valor baixo.

Temos

$$I_A = 0,1163 \quad I_J = -0,310$$

$$I_G = -0,8721$$

$$I_H = 0,5814$$

$$I_B = 0,3721$$

$$I_C = 0,1163$$

$$I_D = -0,279$$

$$I_E = 0$$

$$I_F = 0,00664$$

G de Getis e Ord

$$G_i = \frac{\sum_{j=1}^n w_{ij} y_j}{\sum_{k=1}^{K \neq i} y_k}$$

$$G_i^* = \frac{\sum_{j=1}^n w_{ij} y_j}{\sum_{k=1}^n y_k}$$

$$G_A = \frac{1/3(3+5+6)}{48} = 0,097$$

$$G_A^* = \frac{1/3(3+5+6)}{50} = 0,093$$

$$G_G \in G_S^*$$

$$G_G = 0,156$$

$$G_S^* = 0,150.$$

$$G_H = 0,150$$

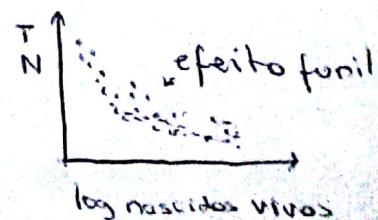
$$G_H^* = 0,120$$

Problemas de Estimação em Áreas Pequenas

- Valores extremos ocorrem nas áreas com pequenas populações
- O que mais chama a atenção em um mapa são os valores extremos (pode ser menos confiável).

Ex: Mortalidade Infantil em MG

TN = taxa de natalidade padronizada



- taxas municipais em 1994 → 756 municípios → variaram de 0 a 600 (608,9).
- 15 municípios c/ nenhuma morte e menos de 30 nascerdos vivos.
- Se uma única morte fosse registrada as taxas passariam de 0 p/ valores entre 116 e 1048.

Como resolver?

- agrregar áreas p/ formar áreas maiores.

Desvantagem: perder a informação localizada.

mapas de probabilidade - estimar melhor o risco de uma área i, com abordagens bayesianas

• empírica → fácil implementação

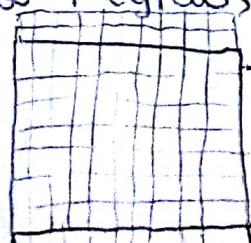
• puramente bayesiana → requer mais esforço computacional.

teste de normalidade - shapiro.test(a)

Abordagem Bayesiana.

principal conceito: probabilidade a priori e a posteriori

Ex: Encontrar depósito de um determinado mineral em uma região, cuja área é de 10000 Km².



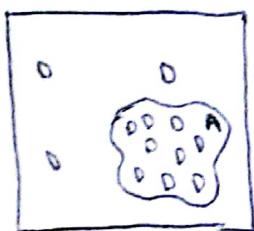
→ a área foi dividida em células de Km² e ocorre somente um depósito em cada célula.

Total = 10000 célula

$$P(D) = \frac{200}{10000} = 0,02$$

Nova Evidência:

No mapa de anomalia magnética da região, observou-se que 180 dos 200 depósitos ocorreram dentro dessa área.



$$P(D/A) \geq 0,02$$

$$P(D/\bar{A}) \leq 0,02$$

$$P(D/A) = ?$$

tamanho da área A = 3600 Km²

$$P(D/A) = \frac{P(D \cap A)}{P(A)} = \left\{ \begin{array}{l} \text{é a proporção da área total onde} \\ \text{ocorre simultaneamente depósito e anomalia} \end{array} \right.$$

$$P(A) = \frac{3600}{10000} = 0,36$$

$$P(D/A) = ?$$

$P(A/D) = ?$ prob. de uma célula estar na região de anomalia A, dado que esta célula contém um depósito.

$$P(A/D) = \frac{180}{200} = 0,9$$

$$P(A/D) = \frac{P(A \cap D)}{P(D)}$$
 como $P(A \cap D) = P(D \cap A)$

$$P(D/A) = \frac{P(D) \cdot P(A/D)}{P(A)} = P(D) \cdot \frac{P(A/D)}{P(A)}$$

Dai a ideia Bayesiana

$$P(D/A) = P(D) \cdot \frac{P(A/D)}{P(A)}$$

↓ ↓

prob. a posteriori prob. priori fator adicional (evidência).

$$P(D/A) = 0,02 \times \frac{0,90}{0,36} = 0,02 \times 2,5 = 0,05.$$

A presença de anomalia magnética faz com que a prob. de depósito seja 2,5 vezes maior que a probab. a priori.

Quando esta evidência, a verificação de novos depósitos do mesmo tipo será muito mais eficiente com uma área pequena reduzida, de 1000 km^2 p/ $\exists 600 \text{ km}^2$.

Exercício: Obtenha $P(D/\bar{A})$

Obs: a função densidade a posteriori = f a priori \times f de verossimilhança

$$\int f(x)dx = 1$$

$$f(x) \geq 0.$$

Abordagem Bayesiana Empírica

A taxa verdadeira desconhecida de cada região θ_i e seja $x_i = \hat{\theta}_i$, a taxa observada.

n_i

Clássica $\rightarrow \hat{\theta}_i = x_i$

Bayesiana empírica \rightarrow vamos supor que temos uma distribuição a priori para cada θ_i , com média $\bar{\theta}_i$ e variância ϕ_i .

$$\Rightarrow \hat{\theta}_i = w_i x_i + (1 - w_i) \bar{\theta}_i$$

sendo $w_i = \frac{\phi_i}{\phi_i + \frac{\bar{\theta}_i}{n_i}}$

04/10/2020

Métodos para Processos Pontuais

→ Estacionariedade e Isotropia estão relacionados a momentos de 1^a e 2^a ordem

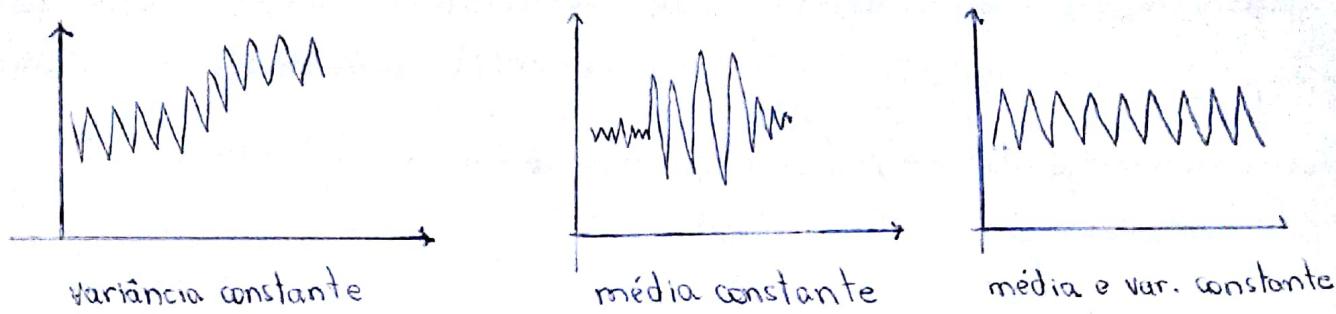
→ Um processo é dito estacionário se os momentos constantes em toda a região de estudo R.

→ Se a média é constante e variância diferente ao longo do processo, tem-se estacionariedade de 1^a ordem

→ Se a variância é constante e média varia ao longo do processo, denominase estacionariedade de 2^a ordem

Exemplo: série temporal, assim em vez de termos uma linha no tempo temos

uma superfície no espaço.



- A isotropia aparece, quando além do processo ser estacionário, a covariância é invariante a direção, ou seja, seu comportamento é igual em todas as direções (norte-sul, leste-oeste) dependendo somente da distância entre os pontos e não da direção entre eles (círculo)
- Caso a covariância, além de variar com a distância varia simultaneamente em função da direção, ela é considerada anisotrópica (elipse).

Análise de Padrão Pontual

Operações

Variável aleatória \Rightarrow localização pontual

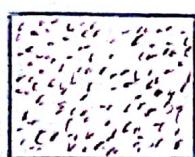
\rightarrow variável de interesse: ocorre o evento (sucesso)

\rightarrow podem ser relacionadas a atributos.

\rightarrow Interesse: é verificar se os eventos observados existem um padrão sistemático ou estão distribuídos aleatoriamente.

Conjunto de dados

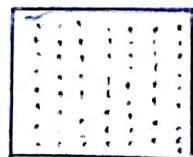
distâncias entre os



Aleatório



Aglomerado



Regular

→ Padrão aleatório: apresenta uma distribuição.
Está ligado ao conceito de estacionariedade, em que as propriedades estatísticas da variável independem de sua localização absoluta, ou seja, a média e a variância são constantes.

Processos de Poisson

→ Padrão Aglomerado

→ aglomeração de pontos, um conjunto de eventos muitas próximas uns dos outros.

O estudo parte para a identificação dos possíveis fatores que pode influenciar na ocorrência dos conglomerados.

→ Padrão Regular

Os pontos estão espalhados uniformemente por toda a região de estudo R (dist. uniforme)

Observação:

Aglomerado: distância média menor

desvio padrão maior

Regular: média grande

desvio padrão da distância pequeno

Aleatório: distância média e desvio padrão razoável.

Descrição Estatística

$N(A)$ = número de pontos na área A

→ Densidade teórica de pontos (intensidade de 1ª ordem) em torno da localização $s = (x, y)$

$$E\{N(A)\} \approx \lambda(x, y) \cdot A$$

sendo A uma pequena área em $s = (x, y)$.

Exploração de padrões pontuais

- método Quadrat
- Dividi-se R em sub-regiões em áreas iguais
- Conta-se o número de eventos que caíram dentro da área Quadrat.

$$\lambda(s) = \frac{\# \text{ de eventos}}{\text{área de quadrat}} \quad (\text{média de intensidade})$$

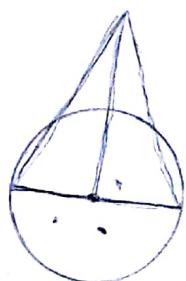
O resultado fará alguma indicação de como a intensidade do processo está variando sobre a região estudada.

Estimação de densidade

Locação $s = (x, y)$ em que a densidade é estimada

$$\hat{\lambda}(x, y) = \frac{N(s_v)}{s_v} \rightarrow \begin{matrix} \text{núm. de casos} \\ \rightarrow \text{área} \end{matrix}$$

Estimação de Densidade por Kernel

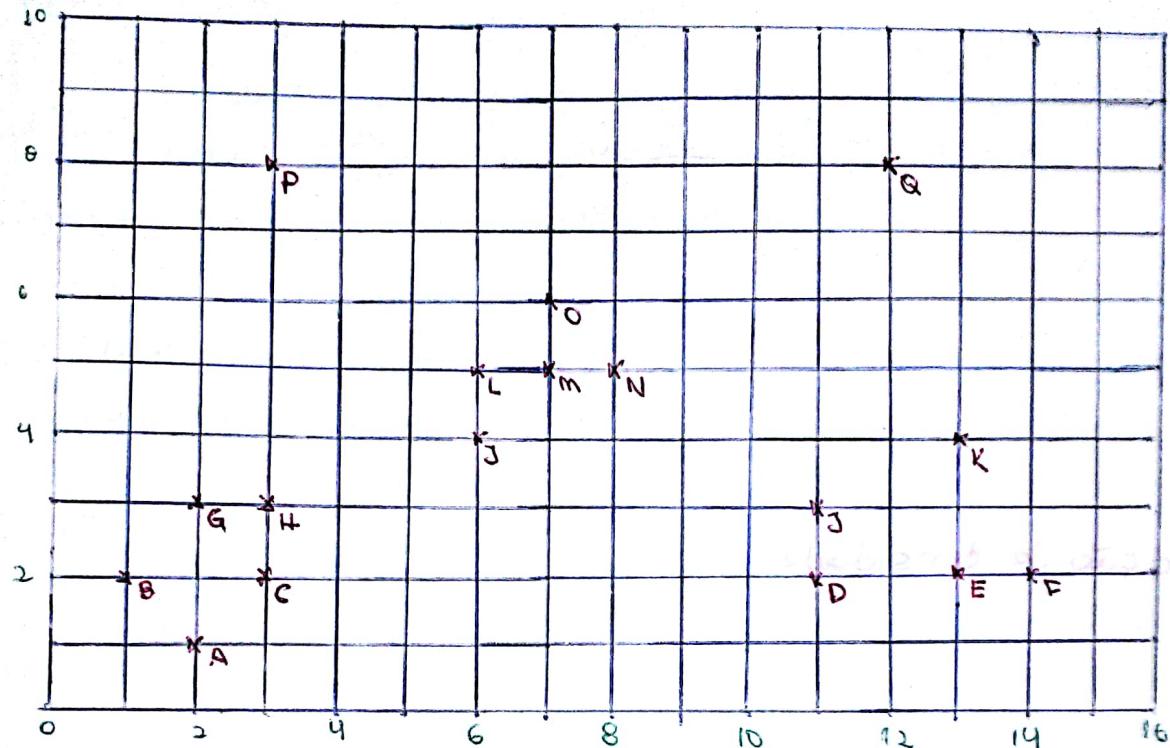


Ao invés de contar somente a quantidade de pontos, cada ponto é ponderado pela distância a s .

Método quadrat -

→ espalhar aleatoriamente quadrantes (2×2) no total de 20 (contar os elementos e aplicar o teste de aleatoriedade).

→ dividir a área em quadrantes (2x2) e contar o número de eventos. (convenção: pontos no limite contar para a área inferior e para esquerda.)



Lista de exercícios

1 - Dados eleição 2010 - Unidades de Federação - 26 e DF

Calcular o índice de Moran local para 2 unidades por variável %. de votos para Marina.

2 - Dados pontuais → considerando capitais variável %. votos para Dilma (acima de 50%)

rec. 14/10/2010

Dados de Padrão Pontual

Estimar a ocorrência de eventos na área de estudo.

Estimador de Kernel

Objetivo: dar uma estimativa suave da intensidade de um padrão pontual de uma amostra de observações

$$h_r(s) = \frac{1}{\pi} e^{-\frac{s^2}{r^2}} = \frac{1}{\pi r^2} e^{-\frac{|s|}{r}}$$

$h(\cdot)$ → função densidade de prob. bivariada (kernel) que é simétrica.

r - raio

Funções Mais Utilizadas

Kernel gaussiano ou normal

$$h_r(x) = \frac{1}{2\pi r^2} e^{-\frac{|x|^2}{2r^2}}, \quad r = \sigma$$

Kernel Quartico

$$h_r(x) = \frac{3}{\pi} (1-x^2)^{\frac{3}{2}}$$

Kernel Triangular

$$h_r(x) = 1 - |x|$$

Kernel Exponencial negativo

$$h_r(x) = \frac{1}{2\pi} e^{-\frac{|x|}{r}}$$

Kernel Uniforme

$$k_u(h) = 1/2.$$

A distribuição normal pondera os pontos dentro do círculo de forma que os pontos mais próximos têm maior peso comparado com os mais afastados.

A distribuição quântica também dá pesos maiores para os pontos mais próximos do que os distantes, mas o decrescimento é gradual.

A distribuição uniforme pondera todos os pontos dentro do círculo igualmente.

A função triangular dá maior peso aos pontos próximos do que os distantes, mas o decrescimento é mais rápido.

A função exponencial negativa pondera os pontos próximos com pesos muito mais intensos do que os pontos distantes.

Vizinho mais próximo

A \rightarrow 2

K \rightarrow 4

Dist² 1 2 4 17 18

B \rightarrow 2

L \rightarrow 1

freq 12 2 1 1 1

C \rightarrow 1

M \rightarrow 1

D \rightarrow 1

N \rightarrow 1

sorteados
Vizinhos

E \rightarrow 1

O \rightarrow 1

evento - evento (W)

F \rightarrow 1

P \rightarrow 18

evento - evento (X)

G \rightarrow 1

Q \rightarrow 17

Estimar a distrib. de prob. acumulada empírica (Função de G(w) de W ou F(x) de X).

H \rightarrow 1

I \rightarrow 1

J \rightarrow 1

$$T \quad 1 \quad \sqrt{2} \quad \sqrt{4} \quad \dots \quad \sqrt{17} \quad \sqrt{18}$$

Método do Vizinho mais próximo

calc. os vizinhos próximos (men. dist.).

$$G(d) = \text{Prob}(W \leq d)$$

$$G_0(d) = 1 - \exp(-\lambda \pi d)$$

$G(d) > G_0(d) \rightarrow$ aglomerado

$G(d) < G_0(d) \rightarrow$ aleatório

Para este caso, temos:

$$\lambda = \frac{17}{16 \times 10} = 0,10625$$

Como $G(d) > G_0(d)$, temos que há aglomerados

$$\hat{G}(1) = \frac{1^2}{17} = 0,059$$

$$G_0(1) = 0,264$$

$$\hat{G}(\sqrt{2}) = \frac{1^2 + 2^2}{17} = 0,08235$$

$$G_0(\sqrt{2}) = 0,487$$

$$\hat{G}(\sqrt{4}) = \frac{1^2 + 2^2 + 4^2}{17} = 0,188$$

$$G_0(\sqrt{4}) = 0,737$$

$$\hat{G}(\sqrt{17}) = \frac{1^2 + 2^2 + 4^2 + 17^2}{17} = 0,74$$

$$G_0(\sqrt{17}) = 0,99$$

$$\hat{G}(\sqrt{18}) = \frac{1^2 + 2^2 + 4^2 + 17^2 + 18^2}{17} = 1$$

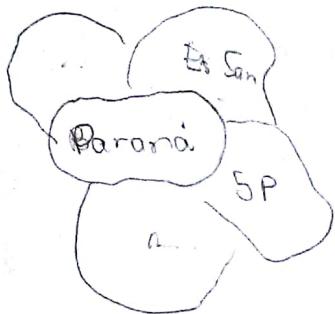
$$G_0(\sqrt{18}) = 0,997$$

Trabalho

UF

Calcular I_i , Índice de Moran Local (Pará, Paraná)

% de votos para Marina



www.rc.unesp.br/igce/aplicada/landim.html

Paulo M. Barbosa. Landim

Curso: Análise Estatística de Variáveis Regionalizadas

Vamos considerar a análise de dados que são "contínuos no espaço".

Vamos procurar entender a distribuição espacial de valores de um atributo sobre toda a região de estudo, dados valores em pontos amostrais ficados.

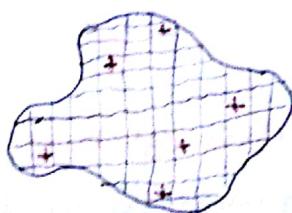
O objetivo é modelar o padrão de variabilidade e estabelecer algum fator que possa estar relacionado com esse padrão. A partir de tais modelos obter boas previsões de valores em pontos nos quais o atributo não foi observado.

Teve grande utilização em geociências - análise de solo, mineração, etc.

→ Grande desenvolvimento nos anos 80/90

Essas técnicas são conhecidas como **Geoestatísticas** → análise de dados de superfície.

De forma geral, esses dados estão disponíveis em forma de amostras pontuais para gerar uma representação na forma de uma grade regular.



em uma mesma localidade (z_1, z_2, \dots, z_k).

O objetivo é reconstruir a superfície da qual as amostras foram retiradas (levantadas)



estimar um modelo de dependência espacial que permite a interpolação da superfície e apresentá-la em um mapa.

Interpolação: procedimento matemático de ajuste de uma função de pontos não amostrados, baseando-se em valores obtidos em pontos amostrados.

Os modelos para gerar superfícies podem ser classificados segundo três grandes abordagens:

1) Modelos determinísticos de efeitos locais: cada ponto da superfície é estimado com base apenas na interpolação dos valores das amostras mais próximas. A suposição implícita é de que predominam os efeitos puramente locais.

São: triangulação, inverso da potência da distância, Krigagem e splines.

2) Modelos determinísticos de efeitos globais. Consideramos todos os pontos da área para interpolar o valor da função em qualquer ponto dentro do domínio dos dados originais. A suposição é que há predominância de variação em larga escala e a variabilidade local não é relevante.

Superfície de Tendência (polinômios)

3) Modelos estatísticos de efeitos globais e locais. Cada ponto da superfície é estimado apenas tendo como fundamento a interpolação dos valores das amostras, utilizando um estimador estatístico. Esses procedimentos requerem que a variabilidade local e global

sejam modeladas por um conjunto de funções básicas, em geral, polinômias.

Krigagem.

Avaliação dos Métodos de Interpolação

- Um método é melhor do que o outro?
- O resultado obtido é fiel dos dados originais?
- A superfície estimada ajusta-se aos dados a um determinado nível de precisão.
- A superfície interpolada é contínua e suave em todos os locais
- Cada valor interpolado depende apenas do subgrupo local de dados (os membros do subgrupo estão próximos do ponto interpolado)
- O método de interpolação pode ser aplicado à todas as configurações e padrões e densidade dos dados.

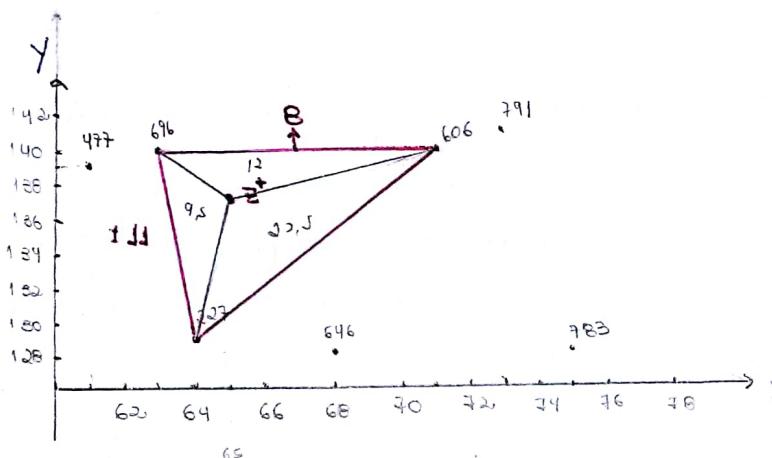
Métodos de Estimação

Triangulação: conecta os pontos amostrados através de triângulos e interpola os valores entre eles. São considerados métodos de estimação diretos, pois os contornos derivam do padrão original dos dados, não permitindo a extração. O estimador limita-se à amostra amostrada.

Uma equação matemática é utilizada para ajustar a superfície. Vários algoritmos podem ser usados.

Os pontos estimados de igual valor (isovaleores) entre os dados medidos e posicionados nos vértices dos triângulos são conectados para intervalos específicos.

Amostras	X	Y	Z	\hat{f}_{Z,Z^*}
1	63	139	477	9,5
2	63	140	696	3,6
3	64	129	227	8,1
4	68	128	646	9,5
5	71	140	606	6,7
6	73	141	791	8,9
7	75	128	783	13,5
	65	137	Z^*	



problema
escala.

O valor z das três amostras próximas

A equação de um plano pode ser escrita como

$$z = ax + by + c$$

Com os valores e as coordenadas, temos o seguinte sistema:

$$\begin{aligned} 696 &= 63a + 140b + c \\ 227 &= 64a + 129b + c \\ 606 &= 71a + 140b + c \end{aligned} \quad \left[\begin{matrix} 696 \\ 227 \\ 606 \end{matrix} \right] = \left[\begin{matrix} 63 & 140 & 1 \\ 64 & 129 & 1 \\ 71 & 140 & 1 \end{matrix} \right] \left[\begin{matrix} a \\ b \\ c \end{matrix} \right]$$

A solução do sistema é:

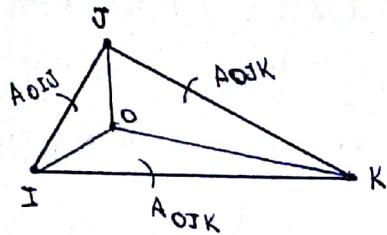
$$a = -11,25, \quad b = 41,619 \quad e \quad c = -4421,159$$

Que dá a seguinte equação como estimador:

$$z^* = -11,25x + 41,61y - 4421,159$$

↓ ↓
65 137

$$\Rightarrow z^* = 548,709,$$



$$\hat{z}_o = \frac{A_{OJK} z_I + A_{OIK} z_J + A_{OIJ} z_K}{A_{IJK}}$$

$$\hat{z}^* = \frac{696 \times 22,5 + 606 \times 9,5 + 227 \times 12}{44} = 548,7$$

Vantagens:

- fácil de ser entendido
- rápido de cálculo e visualização
- fiel aos dados originais
- superfície pode ser interpolada entre os pontos amostrados.

Desvantagens.

- valores acima ou abaixo dos valores reais não podem ser extrapolados
- gera superfícies angulares
- área triangular não é única e isso pode distorcer os resultados.

Reticulação (Reticulado - Gridding)

- estabelece uma grade regular (grid) sobre a área de estudo e calcula os valores nos nós da reticulação com base nos valores dos pontos já amostrados.

São considerados métodos de estimação indiretos, uma vez que os contornos são construídos a partir de dados estimados para os nós da grade e não a partir dos dados originais.

Permite tanto a interpretação quanto à extração dos valores.

Estimativa de Reticulado

Há vários algoritmos para ajustar uma superfície através dos dados estimados para os nós. Os mais utilizados são: inverso ponderado da distância, curvatura mínima, superfície de tendência e Krigagem.

Fornecidos n valores conhecidos z_1, z_2, \dots, z_n (regularmente distribuídos ou não), o valor \bar{z}^* a ser interpolado para qualquer nó da rede será igual a

$$\bar{z}^* = \sum_{i=1}^n p_i z_i$$

A diferença entre os métodos está na maneira como os z_i 's são escolhidos e os respectivos pesos.

(janelas móveis - criar retâng. na reg. e transladar, fazendo a média dos pontos dentro da janela para estimar um ponto).

Inverso Ponderado da Distância

Considerando os valores z_i e as respectivas distâncias d_{ij} do ponto cujo valor a ser estimado, podemos usar a expressão.

$$z_j^* = \frac{\sum_{i=1}^n \frac{z_i}{d_{ij}^p}}{\sum_{i=1}^n \frac{1}{d_{ij}^p}}$$

p = expoente de ponderação (peso)

z_j^* = valor interpolado para o nó do reticulador

O valor p pode ser escolhido pelo pesquisador, segundo seus objetivos:

$0 < p \leq 2 \rightarrow$ destacam anomalias locais

$2 < p \leq 5 \rightarrow$ suavizam anomalias locais

$p=0$ resulta em estimativas \bar{z}^* para os casos:

$p=1; p=0,5; p=2$ e $p=5$.

que aproximam a superfície.

x	y	z	dis	d_{ij}^p	y_{dis}
61	139	477	4,5		
63	140	696	3,6		
64	129	227	8,1		
68	128	646	9,5		
71	140	606	6,7		
73	143	791	8,9		
75	128	783	13,5		

Se $p=1 \Rightarrow z_3 = \frac{\left[\frac{477}{4,5} + \frac{696}{3,6} + \frac{227}{8,1} + \frac{646}{9,5} + \frac{606}{6,7} + \frac{791}{8,9} + \frac{783}{13,5} \right]}{\left[\frac{1}{4,5} + \frac{1}{3,6} + \frac{1}{8,1} + \frac{1}{9,5} + \frac{1}{6,7} + \frac{1}{8,9} + \frac{1}{13,5} \right]} = \frac{630}{1,106} = 595$

Se

$$p=0,5 \quad z_3 = \frac{\left[\frac{477}{\sqrt{4,5}} + \frac{696}{\sqrt{3,6}} + \frac{227}{\sqrt{8,1}} + \frac{646}{\sqrt{9,5}} + \frac{606}{\sqrt{6,7}} + \frac{791}{\sqrt{8,9}} + \frac{783}{\sqrt{13,5}} \right]}{\left[\frac{1}{\sqrt{4,5}} + \frac{1}{\sqrt{3,6}} + \frac{1}{\sqrt{8,1}} + \frac{1}{\sqrt{9,5}} + \frac{1}{\sqrt{6,7}} + \frac{1}{\sqrt{8,9}} + \frac{1}{\sqrt{13,5}} \right]} = \frac{1,593,4}{2,668} = 597,23 //$$

Superfície de Tendência

São interpoladores globais. A superfície é aproximada por um ajuste polinomial dos dados, por um processo de regressão múltipla entre os valores do atributo e as localizações geográficas.

Essa função polinomial é então utilizada para estimar os valores dos pontos em todas as localizações de uma grade regular que aproxima a superfície.

Régressão Linear Simples

Cap 16 - Estatística Básica - Bussab e Morettin

mais detalhes no livro de Montgomery, Pick...

Seja Z a v.a. de interesse, ou seja, queremos informações sobre esta variável, por exemplo, a média $E(Z) = \mu_Z$.

Supondo que a informação populacional é desconhecida.

Tomamos uma amostra aleatória z_1, z_2, \dots, z_n e o melhor estimador de μ_Z é a média amostral $\bar{z} = \frac{\sum z_i}{n}$.

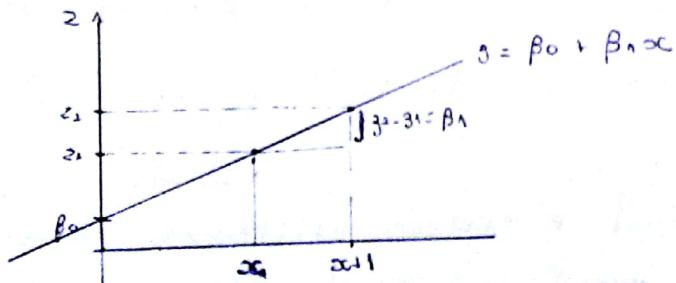
Por exemplo: tempo de reação a um estímulo

$$n=20 \quad \bar{z} = 107,5 \quad s_z = 8,5$$

$$I.C(95\%) \rightarrow \left(\bar{z} \pm 1,96 \frac{s_z}{\sqrt{n}} \right) = \left(107,5 \pm 1,96 \cdot \frac{8,5}{\sqrt{20}} \right)$$

X: Idade \rightarrow 20 anos $\rightarrow \bar{z} = 106,98 \rightarrow s_z = 7,21$
25 anos $\rightarrow \bar{z} = 103,25 \rightarrow s_z = 5,12$
40 anos $\rightarrow \bar{z} = 117,25 \rightarrow s_z = 6,85$

(Z : variável resposta (var. dependente, interesse)
(X : variável explicativa (var. independ., auxiliar))



$$z_i = \beta_0 + \beta_1 x_i + e_i, \quad i=1, 2, \dots, n$$

$$z_1 = \beta_0 + \beta_1 x_1$$

$$z_2 = \beta_0 + \beta_1 (x_2 + 1) = \beta_0 + \beta_1 x_2 + \beta_1$$

$$z_2 - z_1 = \beta_0 + \beta_1 x_2 + \beta_1 - \beta_0 - \beta_1 x_1 = \beta_1$$

β_0 = intercepto = representa o ponto onde a reta corta o eixo das ordenadas.

β_1 = coeficiente angular representa o quanto varia a média de z para um aumento de uma unidade da variável X .

Estimação dos Parâmetros

Algumas suposições são necessárias:

$E(e_i/x) = 0$, $\text{Var}(e_i/x) = \sigma^2_e$ e os erros não são correlacionados.

A partir da amostra, para cada observação temos os pares.

$$(z_i, x_i) \Rightarrow z_i = \beta_0 + \beta_1 x_i + e_i, \quad i=1, \dots, n.$$

Temos n equações e $(n+2)$ incógnitas ($\beta_0, \beta_1, e_1, e_2, \dots, e_n$).

Critério. Encontrar os valores β_0 e β_1 que minimizem a soma dos erros.

$$e_i = z_i - (\beta_0 + \beta_1 x_i), \quad i=1, 2, \dots, n$$

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n \{ z_i - (\beta_0 + \beta_1 x_i) \}^2$$

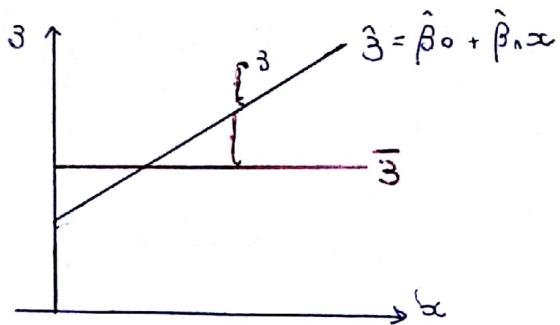
A solução é obtida:

i) derivando (A) em relação a β_0 e igualando a zero.

ii) derivando (A) em relação a β_1 e igualando a zero

iii) resolvendo o sistema obtido em i e ii.

Obtidos $\hat{\beta}_0$ e $\hat{\beta}_1$ temos que avaliar o modelo.



$$(z - \bar{z})^2 = [(z - \hat{z}) + (\hat{z} - \bar{z})]^2$$

$$\sum (z - \bar{z})^2 = \sum [(z - \hat{z}) + (\hat{z} - \bar{z})]^2 \quad (B)$$

A soma de produtos cruzados em B se cancelam \Rightarrow

$$\underbrace{\sum_{i=1}^n (z_i - \bar{z})^2}_{\text{SQ Total}} = \underbrace{\sum_{i=1}^n (z_i - \hat{z}_i)^2}_{\text{SQ Resíduo}} + \underbrace{\sum_{i=1}^n (\hat{z}_i - \bar{z})^2}_{\text{SQ Regressão}}$$

SQ Total

↓
desvio da
observação
em relação
à média

SQ Resíduo

↓
desvio do observado
em relação ao
valor ajustado

SQ Regressão

↓
desvio do valor
ajustado em
relação à média

Estatísticas para verificações de ajuste

$$F = \frac{\text{SQ Reg} / g.l_1}{\text{SQ Res} / g.l_2}, \quad g.l_1 = \text{nº de parâmetros} - 1$$
$$g.l_2 = n - \text{nº de parâmetros}$$

$R^2 = \frac{\text{SQ Reg}}{\text{SQ Total}}$: proporção da variação explicada pelo modelo ou
coeficiente de explicação do modelo

Essas informações podem ser agrupadas em uma única tabela.

ANOVA (Abreviação de Analysis Of Variance).

fonte de variação	graus de liberdade	Soma de Quadrados	Quadrado Médio	F
Regressão	$p-1$	SQ_{Reg}	$\frac{SQ_{Reg}}{p-1}$	$\frac{QM_{Reg}}{\Delta e^2}$
Resíduo	$n-p$	SQ_{Res}	$\frac{SQ_{Res}}{n-p} = \sigma_e^2$	
Total	$n-1$	SQ_{Total}	$\frac{SQ_{Total}}{n-1} = S_p^2$	

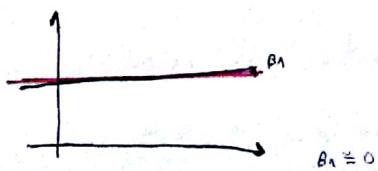
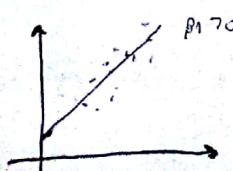
p = no de parâmetros do modelo (no caso, $p=2$)

Por exemplo, $\Delta_3^2 = 72,26$, $\sigma_e^2 = 31,28$ e $R^2 = 0,59$

↳ o modelo consegue explicar 59% da variabilidade

O modelo proposto diminui a variância residual em mais da metade e explica 59% da variabilidade total. Justificamos, então, que é vantajosa a adoção do modelo linear para explicar o tempo médio de reação ao estímulo em função da idade.

$$\rightarrow \begin{cases} y = \beta_0 + \beta_1 x + \epsilon \\ y = \mu + \epsilon \end{cases} \Rightarrow H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0$$



Régressão Múltipla

$$z_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

ou

$$z_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \dots + \beta_k x_{ki} + e_i$$

Em notação matricial:

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Objetivo: Obter valores de $\beta_0, \beta_1, \dots, \beta_k$.

$$Z = X\beta + \varepsilon$$

$$Z \cong X\beta$$

$$X'Z \cong X'X \beta$$

↓

$$\hat{\beta} = (X'X)^{-1} X'Z$$

$$\hat{\beta} = (X'X)^{-1} X'Z$$

Régressão Polinomial

O modelo de regressão linear $Z = X\beta + \varepsilon$ é um modelo geral para ajustar qualquer relação que seja linear nos parâmetros desenhados.

Isso inclui uma classe importante de modelos de regressão polinomial.

• regressão de 2ª ordem com uma variável

$$\hat{z} = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon.$$

- regressão polinomial de 2ª ordem com duas variáveis

$$\hat{z} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_{12} x_1 x_2 + \epsilon$$

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon$$

Esses modelos são usados nas situações em que a resposta é curvilínea.

Retornando a análise de superfície de tendência, queremos correlacionar a distribuição de uma variável resposta (dependente) z em função das coordenadas x e y .

Vamos efetuar a análise a partir de polinômios, tentando preliminarmente uma superfície linear, em seguida quadrática e assim por diante.

O método p/ ajustamento aos dados é o de regressão pelos mínimos quadrados.

O modelo para representação da superfície pelo método dos polinômios é:

$$z_i = \beta_0 + \beta_1 x_i + \beta_2 y_i + \beta_3 x_i^2 + \beta_4 x_i y_i + \beta_5 y_i^2 + \dots + \epsilon_i$$

$$\sum \epsilon_i^2 = \sum (z_i - (\beta_0 + \beta_1 x_i + \dots))^2 \text{ deriva}$$

Ex. A representação de uma superfície linear é dada por:

$$z_i = \beta_0 + \beta_1 x_i + \beta_2 y_i + \epsilon_i, i=1, 2, \dots, n$$

Para o cálculo das coeficientes β_i , dispomos os dados em um sistema de equações normais:

$$\sum z_i = \beta_0 n + \beta_1 \sum x_i + \beta_2 \sum y_i$$

$$\sum z_i x_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2 + \beta_2 \sum x_i y_i$$

$$\sum z_i y_i = \beta_0 \sum y_i + \beta_1 \sum x_i y_i + \beta_2 \sum y_i^2$$

$$\begin{bmatrix} \sum z_i \\ \sum z_i x_i \\ \sum z_i y_i \end{bmatrix} = \begin{bmatrix} n & \sum x_i & \sum y_i \\ \sum x_i & \sum x_i^2 & \sum x_i y_i \\ \sum y_i & \sum x_i y_i & \sum y_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

$ZXY = A$

$XY = C$

$(XY)^{-1}$

$\hat{\beta} = C^{-1} ZXY$

$$\hat{\beta} = (XY)^{-1} ZXY \quad \text{ou}$$

$$\hat{\beta} = C^{-1} A$$

$$\hat{z}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 y_i$$

$$\text{erro } e_i = z_i - \hat{z}_i$$

A superfície quadrática é representada por:

$$z_i = \beta_0 + \beta_1 x_i + \beta_2 y_i + \beta_3 x_i^2 + \beta_4 x_i y_i + \beta_5 y_i^2 + e_i$$

$$\sum e_i^2 = \sum [z_i - (\beta_0 + \beta_1 x_i + \beta_2 y_i + \beta_3 x_i^2 + \beta_4 x_i y_i + \beta_5 y_i^2)]^2$$

$$z_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \beta_4 x_1 y_1 + \beta_5 y_1^2$$

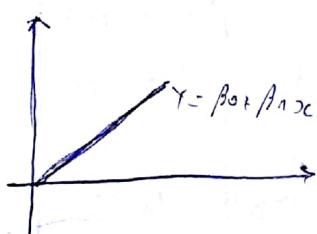
$$z_2 = \beta_0 + \beta_1 x_2 + \beta_2 y_2 + \beta_3 y_2^2 + \beta_4 x_2 y_2 + \beta_5 y_2^3$$

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & y_1 & x_1^2 & x_1 y_1 & y_1^2 \\ 1 & x_{21} & y_2 & x_2^2 & x_2 y_2 & y_2^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & y_n & x_n^2 & x_n y_n & y_n^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

\Rightarrow

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} n & \sum x_i & \sum y_i & \sum x_i^2 & \sum x_i y_i & \sum y_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i y_i & \sum x_i^3 & \sum x_i^2 y_i & \sum x_i y_i^2 \\ \sum y_i & \sum x_i y_i & \sum y_i^2 & \sum x_i^2 y_i & \sum x_i y_i^2 & \sum y_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^2 y_i & \sum x_i^4 & \sum x_i^3 y_i & \sum x_i^2 y_i^2 \\ \sum x_i y_i & \sum x_i^2 y_i & \sum x_i y_i^2 & \sum x_i^2 y_i & \sum x_i y_i^3 & \sum x_i^2 y_i^2 \\ \sum y_i^2 & \sum x_i y_i^2 & \sum y_i^3 & \sum x_i^2 y_i^2 & \sum x_i y_i^3 & \sum x_i y_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum z_i \\ \sum z_i x_i \\ \sum z_i y_i \\ \sum z_i x_i^2 \\ \sum z_i x_i y_i \\ \sum z_i y_i^2 \end{bmatrix}$$

As superfícies de ordem superior seguem o mesmo processo polinomial.



Alguns cuidados:

- a) fazer considerações em relação à curva coberta pelas partes

evitando as extremidades dos mapas,

- b) o nº de pontos deve ser maior que o número de coeficientes do polinômio calculado.
- c) o arranjo dos pontos, ainda que irregular, deve ser aleatório e razoavelmente bem distribuído, evitando agrupamento.
- d) pode haver problemas com resultados obtidos para superfícies de mais alto grau (inversão da matriz).

Ex. sistemas do tipo UTM.

Neste caso, transformar para os valores entre 0 e 1.

$$x^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad \text{e idem para } y^*$$

A verificação do ajuste da superfície aos dados observados é feita da mesma forma pela tabela ANOVA.

A comparação entre modelos também pode ser realizada usando-se a tabela:

Fonte de Variação	g. l.	SQ	QM	F
Regressão de grau p	k	SQReg(p)	QMReg(p)	<u>QMReg(p)</u> (1)
Desvios do grau p	n-k-1	SQRes(p)	QMRes(p)	QMRes(p)
Regressão de grau p+1	l	SQReg(p+1)	QMReg(p+1)	<u>QMReg(p+1)</u> (2)
Desvios de grau (p+1)	n-l-1	SQRes(p+1)	QMRes(p+1)	QMRes(p+1)
Regressão devido ao incremento de p para p+1	l-k	SQI = SQReg(p+1) - SQRes(p)	QMI	<u>QMI</u> (3) QMRes(p+1)

(1) Teste de significância da superfície de grau p.

(2) teste de significância da superfície de grau p+1.

(3) Teste de significância da melhoria de ajuste da superfície p+1 em relação à superfície p.

↳ respondendo a

to: a contribuição do incremento polinomial para o ajuste dos dados é nula.

Exercício:

X	Y	Z
10	17	-665
21	89	-613
33	38	-586
35	20	-440
27	58	-544
60	18	-343
65	74	-455
82	93	-437
89	60	-354
97	15	-142

- Verifique se os modelos polinomiais:
 linear, quadrático e bilinear
 $(z = \beta_0 + \beta_1 x_1 + \beta_2 y + \beta_3 xy)$
 são adequados para descrever z.
 Qual é o melhor dos três?

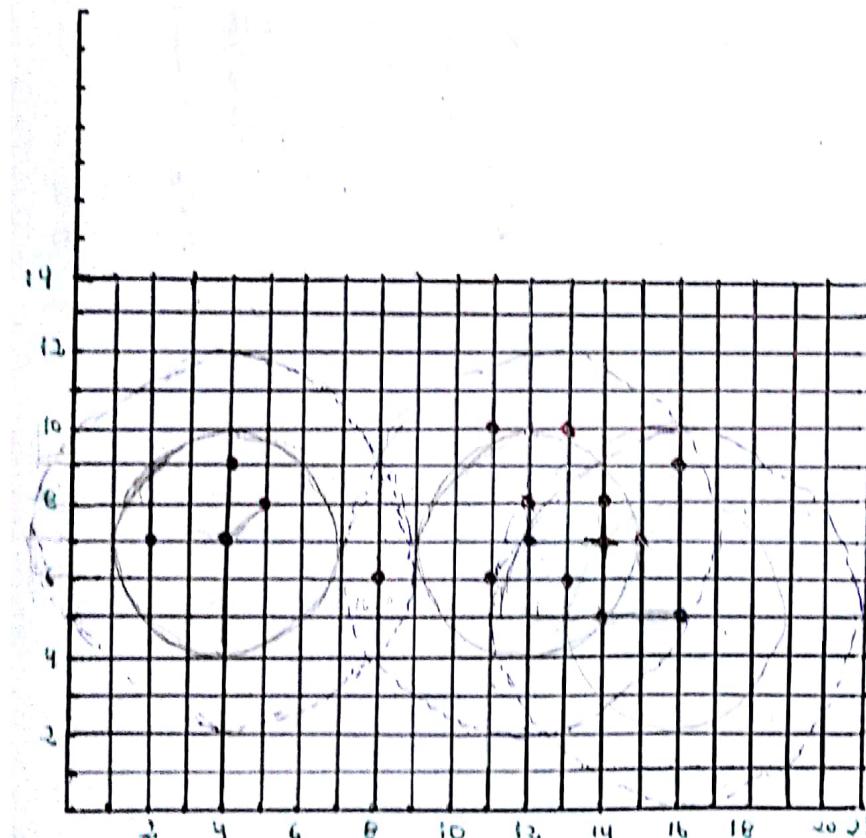
25/10/2010

→ Dados de área →

- médias móveis
- medida de autocorrelação espacial
- I de Moran (local e global)
- C de Geary
- G e G' de Getis Ord.

→ Dados Pontuais

- Método Quadrat
- Medidas de I
- Kernel
 - Vizinho mais próximo (G e G0(d))
 - K de Ripley



$$\hat{s}_k(12,7) = \frac{1}{\pi} \frac{8}{9} \frac{1}{\pi} \frac{8}{9} \left[\left(1 - \frac{4}{9}\right)^2 + \left(1 - \frac{2}{9}\right)^2 + \left(1 - \frac{2}{9}\right)^2 + \left(1 - \frac{8}{9}\right)^2 + \left(1 - \frac{5}{9}\right)^2 + \left(1 - 1\right)^2 \right] = \frac{1}{9\pi^2} \left[\frac{479}{81} \right] = 0,708411$$

1) Use a função Kernel quârtico e raio 3 para estimar os valores nos pontos (4,7), (12,7) e (16,5). Use a dist. euclidiana.

$$\hat{s}_k(z) = \frac{1}{\delta_k(z)} \sum_{i=1}^{n_k} \frac{1}{\pi} \frac{1}{\delta_i^2} K\left(\frac{z - z_i}{\delta_i}\right)$$

$$K(k) = \frac{3}{\pi} (1 - k^2)^2$$

$$\hat{s}_k(4,7) = \frac{1}{\pi} \frac{1}{9} \frac{1}{\pi} \frac{1}{9} \left[\left(1 - \frac{4}{9}\right)^2 + \left(1 - \frac{4}{9}\right)^2 + \left(1 - \frac{2}{9}\right)^2 \right]$$

$$\hat{s}_k(12,7) = \frac{1}{\pi} \frac{8}{9} \frac{1}{\pi} \frac{8}{9} \left[\left(1 - \frac{4}{9}\right)^2 + \left(1 - \frac{2}{9}\right)^2 + \left(1 - \frac{2}{9}\right)^2 + \left(1 - \frac{8}{9}\right)^2 + \left(1 - \frac{5}{9}\right)^2 + \left(1 - 1\right)^2 \right] = \frac{11}{81\pi^2} = 0,1847$$

$$\hat{s}_k(16,5) = \frac{1}{\pi} \frac{8}{9} \frac{1}{\pi} \frac{8}{9} \left[\left(1 - \frac{4}{9}\right)^2 + \left(1 - \frac{2}{9}\right)^2 \right] = \frac{1}{9\pi^2} \left(\frac{61}{81} \right) = 0,2397$$

② Repita para o raio igual a 5.

$$\lambda(4,7) = \frac{1}{0,962} \cdot \frac{3}{5\pi} \left[\left(1 - \frac{2}{25}\right)^2 + \left(1 - \frac{4}{25}\right)^2 + \left(1 - \frac{6}{25}\right)^2 + \left(1 - \frac{17}{25}\right)^2 \right]$$
$$= \frac{3}{4,81\pi} \left(\frac{1475}{625} \right) = 0,489,$$

$$\lambda(12,7) = \frac{3}{5\pi} \left[\left(1 - \frac{1}{25}\right)^2 + \left(1 - \frac{2}{25}\right)^2 + \left(1 - \frac{3}{25}\right)^2 + \left(1 - \frac{8}{25}\right)^2 + \left(1 - \frac{9}{25}\right)^2 + \left(1 - \frac{5}{25}\right)^2 + \left(1 - \frac{10}{25}\right)^2 + \left(1 - \frac{11}{25}\right)^2 + \left(1 - \frac{17}{25}\right)^2 + \left(1 - \frac{20}{25}\right)^2 + \left(1 - \frac{21}{25}\right)^2 \right]$$
$$= \frac{3}{5\pi} \cdot \frac{3190}{625} = 0,975$$

$$\lambda(16,5) = \frac{3}{5\pi} \left[\left(1 - \frac{4}{25}\right)^2 + \left(1 - \frac{5}{25}\right)^2 + \left(1 - \frac{10}{25}\right)^2 + \left(1 - \frac{16}{25}\right)^2 + \left(1 - \frac{13}{25}\right)^2 + \left(1 - \frac{20}{25}\right)^2 \right]$$
$$= \frac{3}{5\pi} \left[\frac{1916}{625} \right] = 6,402211$$

População

Economicamente Ativa: (PEA, empregados e desempregados)

Economicamente Inativa (PENA, menor círculo)

Amostragem Aleatória

é o procedimento básico da amostragem científica, consistindo em atribuir a cada elemento da população um número único para, então, selecionar alguns desses elementos de forma casual, para se garantir que isso seja mesmo aleatório.

Amostragem Sistemática: é uma forma de seleção de unidades de amostragem com características semelhantes às da amostragem aleatória, mas com procedimentos mais rápidos e mais simples.

1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16

Amostragem Estratificada: Nessa técnica, a população é dividida em estratos, que devem ser homogêneos internamente, heterogêneos entre eles, segundo alguma variável de interesse. Em cada estrato, obtém-se uma amostra casual simples, após construção da lista de referência de cada estrato, sendo que a amostra total é a união das amostras de cada estrato.

- Estratificada proporcional ao tamanho do estrato
- Estratificada uniforme (tamanhos iguais ind. estrato)

Amostragem por Conglomerado: agrupamento de elementos ou unidades de amostragem de uma população.

Exemplo

Bairro, Quartelão, ...

Sortear os bairros, depois os quartelões, depois os quadras, depois as residências. (amostragem por conglomerados em 4 estágios).

Amostragem Não-Probabilística

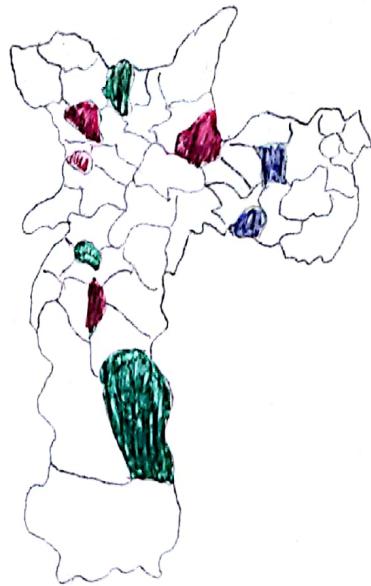
Por Cotas - Amplamente utilizada em pesquisa social.

Consiste em classificar a população em propriedades relevantes para o estudo, determinar a proporção da população pesquisada e determinar uma cota a cada pesquisador.

Por Tipicidade.

Por Acessibilidade.

Estatística Espacial



- floresta
- área
- água
- áreas

Tópicos

Análise Espacial

Análise de Padrões em Dados de Área

Análise de Padrões Pontuais

Análise de Dados Espacialmente Contínuos

"Se onde é importante para o seu estudo, então a Análise Espacial é a sua ferramenta"

Tipos de Problemas

Epidemiologia

Estudo de uma determinada doença:

- Distribuição pelo estado?
- Regiões com maior ocorrência?
- Associação com fontes de poluição ou hábitos alimentares?

Agricultura

Mapear a região de acordo com a produtividade.

Será que a produção é igual em toda a propriedade?

Há partes da propriedade que tenha necessidade de mais fertilizantes?

Geologia

Dado um conjunto, qual a extensão de um depósito mineral?

Comércio

Analisar espacialmente a possibilidade de abrir novas lojas.

Exemplo:

A epidemia de Colera em Londres - John Snow (1854)

Análise Espacial

Estatística Convencional - independência

Estatística Espacial - "Todas as coisas se parecem, coisas mais próximas são mais parecidas do que aquelas que são distantes".
correlação

Estuda métodos científicos para a coleta, visualização e análise de dados que possuem coordenadas geográficas.

Enfase: mensurar propriedades e relacionamentos, levando em conta a localização espacial.

Tipos de Dados

Dados de Superfície Aleatória

- v.a. contínua (temperatura)
- n pontos de coleta de dados em localizações
- gerar uma superfície ($z(x)$) - descrição do fen. de interesse

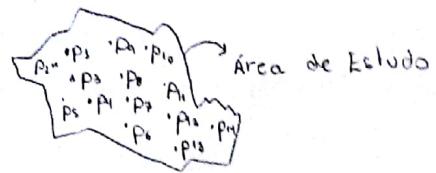
Krig (1951) - considerar a distância entre observações

No início, geoestatística - agricultura de precisão
geoestatística - considerações da localização e dependência espacial.

Dados de Processos Pontuais

Dados em que o principal interesse está no conjunto de coordenadas geográficas representando as loc. exatas dos eventos.

ex: ocorrência de dengue



Verificar se existe agrupamento ("cluster") ou regularidade

- Um método: Estimador de Kernel



abota

Dados de Área

localização está associada a áreas delimitadas por polígonos

Informações disponibilizadas por Ministério e Secretaria - geralmente contagens

- Índices ou taxas, proporções, médias, etc.

Ex. votos para Lula em 2006.

Matriz de Vizinhança

Dados de Intereração Espacial

Migração, fluxo de passageiro

Interação Espaço-Tempo

Softwares:

ESTATCART

GEODA

INFOMAT

TABWIN

SAS

CRIMESTAT

TERRAVIEW

R

Correlação: mede a relação entre as v.a. (X, Y)

$$\rho_{xy} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\text{D.P.}(X) \text{D.P.}(Y)}$$

Covariância - variância conjunta entre as v.a. X e Y.

Variância - variação (espalhamento, dispersão) de uma variável em torno de sua média

O cálculo de ρ_{xy} pode ser simplificado para

$$\rho_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)}}$$

pois

$$\rho_{xy} = \frac{\text{Cov}(X, Y)}{\text{D.P.}(X) \text{D.P.}(Y)} = \frac{I}{II}$$

$$I = \text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i y_i - \bar{x}\bar{y} - \bar{x}y_i + \bar{x}\bar{y}) =$$

$$= \sum x_i y_i - \bar{y} \sum x_i - \bar{x} \sum y_i + \bar{x}\bar{y} =$$

$$= \sum x_i y_i - \bar{y} n \sum x_i - n \bar{x} \sum y_i + n \bar{x} \bar{y} =$$

$$= \frac{1}{(n-1)} \left[\sum x_i y_i - 2n \bar{x} \bar{y} + n \bar{x} \bar{y} = \sum x_i y_i - n \bar{x} \bar{y} \right]$$

e

$$II = \text{D.P.}(X) \cdot \text{D.P.}(Y) = \sqrt{\frac{1}{(n-1)} \sum (x_i - \bar{x})^2} \sqrt{\frac{1}{(n-1)} \sum (y_i - \bar{y})^2} =$$

$$= \frac{1}{(n-1)} \sqrt{\sum (x_i^2 - 2x_i \bar{x} + \bar{x}^2) \sum (y_i^2 - 2y_i \bar{y} + \bar{y}^2)} =$$

$$= \frac{1}{n-1} \frac{[\sum x_i^2 - n\bar{x}^2][\sum y_i^2 - n\bar{y}^2]}{[(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)]}$$

Logo

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)}}$$

Para uma única variável, temos ^{MCORRELAÇÃO}
_{AUTOCORRELAÇÃO}

Autocorrelograma \rightarrow gráfico de autocorrelações

Variograma \rightarrow gráfico de variações

Semivariograma \rightarrow " " " particionados por d_u .

Análise de Dados de Área

A localização está delimitada por polígonos (geralmente contagem).

Polígonos \rightarrow quadras, setores censitários, município, estado, país, etc.

Forma usual de apresentação - uso de mapas coloridos com padrão espacial do fenômeno

Após o mapa, ao deparar com algum padrão espacial.

\hookrightarrow ele é aleatório

\hookrightarrow tem agrupamento definido

\hookrightarrow pode estar associado a causas mensuráveis?

\hookrightarrow os valores obs. são suficientes para analisar o fen. espacial de interesse?

\hookrightarrow Existem grupos de áreas c/ diferentes padrões na área do estudo?

Modelo de distribuição de dados de área

Seja Y_i uma v.a que descreve a contagem, indicador ou taxa \rightarrow área A_i .

Temos um $O_i = y_i$.

Hipótese mais comum: $Y \sim \text{Poisson}(\lambda)$

Outras dist. podem ser mais adequadas, dependendo da v.a de interesse. As taxas podem ter dist. normal.

Problemas de Unidade de Área



maior nº de áreas \rightarrow menor nº de obs

\rightarrow baixa qualidade.

A redução de escala (áreas maiores) tende a homogeneizar os dados, reduzir a flutuação aleatória e reforçar correlações que, assim, aparentam ser mais fortes que áreas menores.

Não se pode afirmar qual escala está certa mas sim qual serve melhor

Proximidade Espacial

Outro tipo de dados \rightarrow geral... distância euclidiana.

A principal diferença p/ dados de área está na formalização da proximidade espacial.

A) Proporção da Fronteira pelo Perímetro

B) Distância linear entre os centroides dos objetos
 { o para $d > 1$ limiar } c.c

c) Inverso da Distância Linear

D) Existência de fronteira comum

↓ se faz, o c.c

Dai, constrói-se a matriz de proximidade espacial

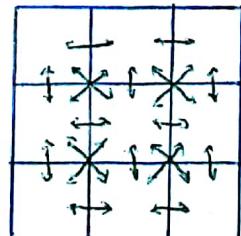
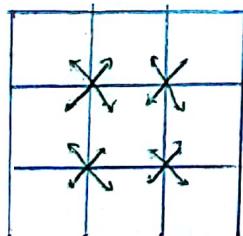
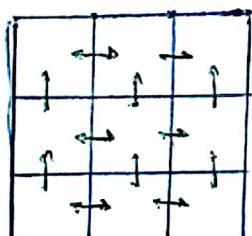
$$W = \begin{bmatrix} w_{11} & w_{12} & w_{13} & \dots & w_{1n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & w_{n3} & \dots & w_{nn} \end{bmatrix}$$

w_{ij} : "distância" do objeto i ao objeto j .

Padronizando W

$$W^* = \begin{bmatrix} \frac{w_{11}}{\sum w_{1i}} & \frac{w_{12}}{\sum w_{1i}} & \dots & \frac{w_{1n}}{\sum w_{1i}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{w_{n1}}{\sum w_{ni}} & \frac{w_{n2}}{\sum w_{ni}} & \dots & \frac{w_{nn}}{\sum w_{ni}} \end{bmatrix}$$

Contiguidade



Medidas Básicas de Autocorrelação

Testes para dados qualitativos nominais

Classificação binária:

$$z_i = \begin{cases} 1, & \text{se a região é P} \\ 0, & \text{se a região é B} \end{cases}$$

P - possui a característica de interesse

B - semelhante

As contiguidades possíveis são: (BP, BB, PP, PB) → considerando

BP - PB

Quando tiver n grande de vizinhos,

PP → agrupamento ou cluster (corr. positiva)

Quando n grande de vizinhos,

PB → padrões alternados - (corr. negativa)

A estatística do produto geral é dado por:

$$\alpha = \sum_i \sum_j w_{ij} y_{ij}$$

w = {w_{ij}} é a med. de proximidade esp. dos locais i e j.

y = {y_{ij}} é a medida de proximidade de i e j em alguma outra dimensão.

Upton e Fingleton (1985) \Rightarrow 98 permutações aleatórias

Distribuição de τ obtida.

τ	8	10	12	14	16	18	20	22	24	Total
freq	8	11	25	25	7	10	0	2	99	

A tabela mostra que 8 é um valor extremo raramente ocorrendo apenas 8 vezes. Portanto, não suficiente para ser julgado significativo.

Cliff e Ord (1981) - Aproximar Normal

$$E(\tau) = \frac{S_0 T_0}{n(n-1)}$$

$$\text{Var}(\tau) = \frac{S_1 T_1}{2 n^{(2)}} + \frac{(S_2 - 2S_1)(T_2 - 2T_1)}{4 n^{(3)}} + \frac{(S_0^2 + S_1 - S_2)(T_0^2 + T_1 - T_2)}{n^{(4)}} - E(\tau)^2$$

$$S_0 = \sum_{i=3} \sum w_{ij}$$

$$S_1 = \frac{1}{2} \sum_{i=3} \sum (w_{ij} + w_{ji})^2 = \frac{1}{2} (w_{12} + w_{21})^2 + (w_{13} + w_{31})^2 + (w_{14} + w_{41})^2 + (w_{23} + w_{32})^2 +$$

$$S_2 = \sum (w_{ij} + w_{ji})^2 = (\text{diagonal} + \text{col 1})^2 + (\text{linha} + \text{col 2})^2 + \dots + (\text{linha} + \text{col } n)^2$$

$$w_{10} = \sum_i w_{ij} \quad w_{0i} = \sum_j w_{ji}$$

$$n^{(2)} = n(n-1)$$

$$n^{(3)} = n(n-1)(n-2)$$

$$n^{(4)} = n(n-1)(n-2)(n-3)$$

$$z = \frac{|r - E(r)|}{\sqrt{\text{Var}(r)}} - 1$$

A estatística apresentada anteriormente vale para todos as situações (não só categorias binárias)

No caso de dados binários

p = nº de caselas pretas

b . " " " brancas

n = nº total de caselas

$$T_0 = 2pb$$

$$T_1 = 2T_0$$

$$T_2 = 4npb = nT_1$$

$$S_0 = 2(2lc - l - c)$$

$$S_1 = 2S_0$$

$$S_2 = 8(8lc - 7l - 7c + 4)$$

l : nº de linhas

c = nº de colunas

Contagem de Juncões

Caso particular: resposta binária

W	W	
W	B	
	B	
W	B	

continua \Rightarrow 2 categ.

Ex: acima da mediana (1) - Preto

abaixo da " (0) - Branco

Discreta - transforma em duas categorias \Rightarrow a de maior interesse (1) e o restante (0)

Dessa forma, juncões possíveis são: PP, PB e BB

Quando n' grande de vizinhos PP \rightarrow corr. positiva

" " " " " PB \rightarrow padrões alternados (corr. negativa)

O nº de junções PB é $\pi/2$, de acordo com $n = \sum_{i,j} w_{ij}$ obtido da matriz W e Y, este último definido como $y_{ij} = (x_i - x_j)^2$.

O nº de junções PP é $r^*/2$, r^* é o valor de r quando $y_{ij} = \infty$.

Se o total de junções no sistema for J \Rightarrow nº BB = J - PB - PP.

PB \rightarrow mais informativo.

Desde que nº PB = $r^*/2$, $E(PB) = E(r^*)/2 = V(PB) = V(\frac{r}{2}) \cdot \frac{\sqrt{r}}{2}$

Logo

$$E(PB) = \frac{E(r)}{2} = \frac{1}{2} \cdot \frac{S_0 T_0}{n(n-1)}$$

No caso $l=c$ (área quadrada)

$$E(PB) = \frac{1}{2} \frac{S_0 T_0}{n(n-1)}$$

$$S_0 = 2(l(2lc - l - c)) = 2(2lc \cdot l - 2l) = 2(2l^2 - 2l) = 4l(l-1)$$

e

$$T_0 = 2pb$$

assimut ob propriedad

\Rightarrow

$$E(PB) = \frac{1}{2} \frac{4l(l-1)2pb}{l^2(l^2-1)}, \text{ pois } n = lc = ll = l^2$$

$$\Rightarrow E(PB) = \frac{4l pb}{l^2(l+1)} = \frac{4pb}{l(l+1)}$$

Da mesma forma

$$E(PP) = \frac{2p(p-1)}{l(l+1)}$$

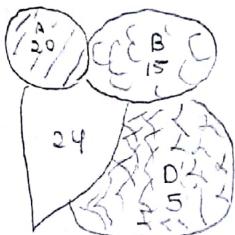
Média Móvel Espacial

Explorar variação da tendência espacial - cálculo da média com valores dos vizinhos de 1^a ordem.

Considerando w_i , a estimativa da média móvel espacial pode ser:

$$\hat{\mu}_i = \frac{\sum_{j=1}^n w_{ij} y_j}{\sum_{j=1}^n w_{ij}}, i=1, 2, \dots, n$$

Exemplo



Antos

$$\text{Amplitude} = 20 - 5 = 15$$

até 10 - < 2

10 à 15 - CCC

15 à 20 - III

> de 20 - *

$$\mu_A = \frac{20 + 15 + 24}{3} = \frac{59}{3} = 19,67$$

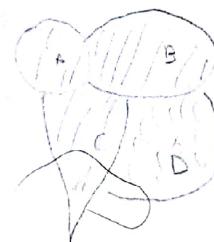
$$\mu_B = \frac{15 + 20 + 24 + 5}{4} = 16$$

$$\mu_C = 6^o / 4 = 16$$

$$\mu_D = 4^o / 5 = 14,67$$

Ou por matrizes

$$\hat{\mu} = W Y = \underbrace{\quad}_{\text{padronizada}} \begin{bmatrix} \mu_A \\ \mu_B \\ \vdots \\ \mu_n \end{bmatrix}$$



Índices Globais de Autocorrelação

Mede o nível de interdependência geográfica entre as variáveis e a natureza e a força desse relacionamento.

Medidas Mais Utilizadas

Índice global de Moran

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{[\sum_{i=1}^n (y_i - \bar{y})^2] [\sum_{i,j} w_{ij}]}$$

n = áreas em estudo

y_i = valor do atributo considerado na área

\bar{y} = valor médio do " na região

w_{ij} = pesos atribuídos conforme a conexão

Testa se as áreas conectadas apresentam maior semelhança quanto ao indicador estudado do que o esperado num padrão.

Índice C de Geary

$$C = \frac{(n-1) \sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - y_j)^2}{2 [\sum_{i=1}^n (y_i - \bar{y})^2] [\sum_{i,j} w_{ij}]}$$

Quando independência esp., I próximo de 0.

Qnd há similaridade entre localidades próximas, I positivo e I negativo

Quando n autocorrelação

$$E(I) = \frac{-1}{n-1} \quad \text{e} \quad E(C) = 1$$

Quando autocor. máxima ($C \rightarrow 0$ e $I \rightarrow 1$)

$$I \approx p_{\text{aus}}$$

Sob H_0 (não existe dependência espacial entre as localidades)

$$I \sim N(E(I), \text{Var}(I))$$

$$\text{Var}(I) = \frac{n\{(n^2 - 3n + 3)s_1 - ns_2 + 3s_0^2\} - R\{n(n-1)s_1 + 2ns_2 + 6s_0^2\}}{(n-1)^2 s_0^2}$$

$$= \frac{1}{(n-1)^2}$$

$$R = \frac{m_4}{m_2^2}$$

$$m_n = \frac{1}{n} \sum (x_i - \bar{x})^n$$

notar se o resultado é de amparo

Fazendo a padronização $\tilde{y}_i = \underline{y}_i - \bar{y}$ temos

$$\text{Var}(\tilde{y}_i)$$

$$I = \frac{\sum \sum w_{ij} \tilde{z}_i \tilde{z}_j}{\sum \tilde{z}_i^2}$$

$$(AB, AC)$$

Usando matrizes

$$W \cdot (z; z_j) = W \cdot z z^T, \text{ em que}$$

$$z = \begin{bmatrix} z_A \\ z_B \\ z_C \\ z_D \\ \vdots \end{bmatrix}$$

onde

zz^T = multiplicação de matrizes = m

$W \cdot m$ = multiplicar wiggis (elemento por elemento)

Outra medida de autocorrelação espacial

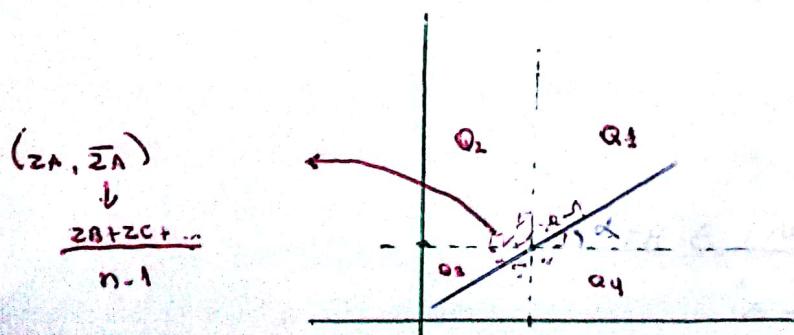
C de Geary

$$C = \frac{(n-1) \sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - \bar{y}_j)^2}{2 \left[\sum_{i=1}^n (y_i - \bar{y})^2 \right] \left[\sum_{i,j} w_{ij} \right]}$$

$C \rightarrow 0$, autocorrelação espacial

$C \rightarrow 1$, não existe "

Diagrama de Espalhamento de Moran



$$I = \frac{Z'WZ}{Z'Z} \Rightarrow \begin{array}{l} \text{coeficiente} \\ \text{de reg. linear} \end{array}$$

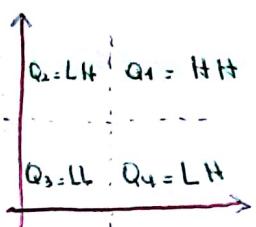
Q_1 (valores +, médias +) e Q_3 (valores -, médias -)

pontos associação espacial positiva \Rightarrow uma localização possui vizinhos com valores semelhantes

Q_4 (valores +, médias -) e Q_2 (valores -, médias +)

pontos associação espacial negativa \Rightarrow uma localização possui vizinhos com valores distintos

O diagrama de espalhamento de Moran é mapa ceroplético.



- $Q_1 = \text{alto-alto}$
- $Q_2 = \text{baixo-alto}$
- $Q_3 = \text{baixo-baixo}$
- $Q_4 = \text{alto-baixo}$

Índices Locais de Associação Espacial

O Índice global fornece um único valor para toda região de estudo, mas muitas vezes, temos interesse em examinar padrões em uma escala maior

Os Índices Locais

permitem avaliar diferentes regimes espaciais existentes na área de estudo;

medem a associação espacial entre uma observação i e a sua vizinhança.

Requisitos:

Soma dos índices locais prop. índice global

Deve indicar a significância da associação espacial para cada observação

$$I \propto \sum_{i=1}^n I_i \quad \left(I = \frac{\sum_{i=1}^n I_i}{n} \right)$$

$$I_i = \frac{(y_i - \bar{y}) \sum_{j=1}^n (y_j - \bar{y}) w_{ij}}{\sum_{k=1}^n (y_k - \bar{y})^2}, \quad i=1, 2, \dots, n$$

Se padronizados

$$I_i = \frac{g_i \sum_{j=1}^n g_j w_{ij}}{\sum_{k=1}^n g_k^2 / n}$$

$I_i > 0$: "cluster de valores similares (altos ou baixos)

$I_i < 0$: " " " " distintos (ex: um área com valor alto rodeada por uma vizinhança com valores baixos)

Indicadores Locais G_i e G_i^* (Getis e Ord)

$$G_i = \frac{\sum_{\substack{j=1 \\ j \neq i}}^n w_{ij} y_j}{\sum_{\substack{k=1 \\ k \neq i}}^n y_k}, \quad j \neq i$$

$$G_i^* = \frac{\sum_{\substack{j=1 \\ j \neq i}}^n w_{ij} y_j}{\sum_{\substack{k=1 \\ k \neq i}}^n y_k}, \quad j \neq i$$

G_i - Soma dos dados da vizinhança de i relativa a soma de todos os dados exclusivos y_j

G^* - idem a G_i , considerando todos os valores (inclusive y_i) no denominador.

Valores altos da estatística G indicam alta concentração espacial (agrupamento).

G_i e G^* tem aprox. dist. Normal

$$E(G_i) = \frac{w_i}{n-1} \quad \text{e} \quad E(G^*) = \frac{\bar{w}_i}{n}$$

sendo

$$w_i = \sum_{\substack{j=1 \\ j \neq i}}^n w_{ij} \quad \text{e} \quad \bar{w}_i = \frac{1}{n} \sum_{j=1}^n w_{ij}$$

$$\text{Var}(G_i) = \frac{w_i(n-1-w_i)}{(n-1)^2(n-2)} \left[\frac{s(i)}{\bar{x}(i)} \right]^2,$$

$$\text{Var}(G^*) = \frac{\bar{w}_i(n-\bar{w}_i)}{n^2(n-1)} \left[\frac{s}{\bar{x}} \right]^2$$

$$\bar{x}(i) = \frac{\sum_{\substack{j=1 \\ j \neq i}}^n x_j}{n-1}, \quad s^2(i) = \frac{\sum_{\substack{j=1 \\ j \neq i}}^n x_j^2 - [\bar{x}(i)]^2}{n-1}$$

Também função da distância d entre as localidades (com tróides)

$$G(d) = \frac{\sum_{\substack{j=1 \\ j \neq i}}^n w_{ij}(d) y_j}{\sum_{\substack{k=1 \\ k \neq i}}^n y_k}$$

$$G^*(d) = \frac{\sum_{\substack{j=1 \\ j \neq i}}^n w_{ij}(d) y_j}{\sum_{k=1}^n y_k}$$

Gráfico de $d \times G(d)$ ou $d \times G^*(d)$

↳ vários artigos - espec. em sens. remoto.

A estatística geral G da associação espacial global é dada por

$$G = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} y_i y_j}{\sum_{i=1}^n \sum_{j=1}^n y_i y_j} \quad , \forall i \neq j$$

$G \sim N(E(G), V(G))$

$$E(G) = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}}{n(n-1)}, \quad i \neq j$$

$V(G)$ - muitos cálculos

Problemas de Estimação em Áreas Pequenas

Áreas menores - pop. peq. populações → valores extremos - menos confiável

Como resolver?

agregar áreas

Desvantagem - perder inf. localizada

Mapas de Probabilidade - estimar melhor o risco de uma área i, com abordagens bayesianas

- empírica - fácil implementação
- puramente bayesiana - requer mais esf. comput.

Abordagem Bayesiana

Ex. depósito do mineral

$$p(D|A) = p(D) \cdot \frac{p(A|D)}{p(A)}$$

$$p(D|A) = \frac{p(D \cap A)}{p(A)} \quad (I)$$

$$p(A|D) = \frac{p(A \cap D)}{p(D)} \quad \text{como } p(A \cap D) = p(D \cap A)$$

$$p(A|D) = \frac{p(D \cap A)}{p(D)} \Rightarrow p(D \cap A) = p(A|D) \cdot p(D)$$

Sub em I

$$p(D|A) = \frac{p(A|D) \cdot p(D)}{p(A)} = p(D) \cdot \frac{p(A|D)}{p(A)}$$

Bayesiana Empírica

θ : o parâmetro desconhecido

$x = \underline{x_i}$, a taxa obs.

n_i

Clássica $\Rightarrow \tilde{\theta}_i = \bar{x}_i$

$\theta_i \sim \text{distribuição} (\text{média} = \bar{x}_i, \text{var} = \phi_i)$

$$\Rightarrow \tilde{\theta}_i = w_i \bar{x}_i + (1-w_i) \bar{x}_i$$

sendo $w_i = \frac{\phi_i}{\phi_i + \bar{x}_i}$

Processos Pontuais

Estacionariedade e Isotropia

Estacionários - momentos cte em R

Se média é cte e var. dif ao longo do processo, estacionariedade de 1^a ordem.

Se var é cte e média varia ao longo do processo, estacionariedade de 2^a ordem.

Isotropia quando além estacionário covariância invariante a direção, comportamento igual em todas as direções.

Caso a Cov, além de variar com a dist. variar simul. em função da direção, ela é considerada anisotrópica.

Dados de Área

$$O_i = Y_i \rightarrow \sim \text{Poisson}(\lambda)$$

tasas \approx normal

Proximidade Espacial (Contiguidade)

Medida Básica de Autocorrelação

$$\mu = \sum \sum_{ij} w_{ij} y_{ij}$$

Clifford e Ord (1981)

$$E(\mu) = \frac{S_0 T_0}{n(n-1)}$$
$$\text{Var}(\mu) = \frac{S_1 T_1}{2n(n-1)} + \frac{(S_2 - 2S_1)(T_2 - 2T_1)}{4n(n-1)}$$
$$\frac{(S_0^2 + S_1 - S_2)(S_0^2 + T_1 - T_2) - [E(\mu)]^2}{n(n-1)}$$

3
W
W
" " " " S

$$\sin^{-1} \sum_{i=1}^n (\sin x_i + \cos y_i)^2$$

$$W = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$$

$$\begin{array}{r} 5 \\ \times 3 \\ \hline 15 \end{array}$$

$$(1-\alpha) \in \bar{a}^{\perp}$$

$$a = \log_{10}(1 - g)$$

$$(h-u)(x-u)(v-u) \cup = (h,u)$$

$$S = \frac{|x - E(x)| - 1}{\sqrt{\text{Var}(x)}}$$

Dados Binários

9
9
8
11
0
F

10
8
11
5

$$T_2 = 4npb = nT_1$$

$$S_0 = 2(\alpha_0 - \alpha - c)$$

၁၀၈

$$S_2 = \frac{1}{4} \sin(8\pi x - 7\varphi - 7c + h)$$

$$\text{Junges } PB = \pi/2$$

$$E(PB) = \frac{E(n)}{2} = \frac{1}{2} \frac{S_0 T_0}{n(n-1)}$$

no caso $\lambda = 0$

$$\begin{aligned} &= \frac{1}{2} \frac{2\lambda(2\lambda - \lambda - \lambda)T_0}{n(n-1)} = \frac{1}{2} \frac{(4\lambda^2 - 4\lambda)T_0}{n(n-1)} = \frac{4\lambda(\lambda - 1)T_0}{2\lambda^2(\lambda - 1)} = \frac{2\lambda T_0}{\lambda^2} \\ &= \frac{4\pi b}{\lambda(\lambda + 1)} \end{aligned}$$

$$E(PB) = \frac{2P(\beta-1)}{\lambda(\lambda+1)}$$

Média Móvel

$$\hat{\mu}_i = \frac{\sum_{j=i}^{i+3} w_j y_j}{\sum_{j=i}^{i+3} w_j} \quad i=1, 2, \dots, n$$

wii 70

ou

$$\hat{\mu}_i, \hat{w}_i y_i = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$$

padron.

Índice Globais de Autocorrelação

$$I = \frac{n}{n-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})$$

onde w_{ij} pesos
 $\max(p) \Rightarrow p \rightarrow 1 \Rightarrow c \rightarrow 0$

$$C = \frac{(n-1)}{2} \sum_{i=1}^{n-1} \sum_{j=i+1}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})$$

Padronização:

$$I = \frac{\sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_i w_{ii}}$$

$$W \cdot (Z; \bar{z}) = W \cdot Z'$$

multiplicação matriz

multipl. elementos

$$\bar{E}(I) = \frac{-1}{n-1}$$

Índices Locais de Associação Espacial

$$I \propto \sum_i I_i$$

$$I_i = \frac{(g_{ii} - \bar{g}) \sum_{j \neq i} (w_{ij} - \bar{w}_{ij})}{\sum_k (w_{ik} - \bar{w})^2}$$

Padronizado:

$$I_i = \frac{\sum_{j \neq i} w_{ij} z_{ij}}{\sum_{k \neq i} w_{ik}}$$

$I > 0$ cluster similar
 $I < 0$ clusters distintos (forte redistribuição)

$$\begin{aligned} \text{Indicadores Locais } G_i \text{ e } G_i^* & (\text{Getis e Ord}) \\ G_i &= \frac{\sum_{j \neq i} w_{ij} g_{ij}}{\sum_{k \neq i} w_{ik}}, \quad j \neq i \\ G_i^* &= \frac{\sum_{j \neq i} w_{ij} g_{ij}}{\sum_{k \neq i} w_{ik}}, \quad j \neq i \end{aligned}$$

$$E(G_i) = \frac{w_i}{n-1}$$

$$E(G_i^*) = \frac{w_i^*}{n}$$

$$w_i^* = \sum_{j \neq i} w_{ij}$$

$$w_j^* = \sum_{i \neq j} w_{ij}$$

$$\text{Var}(G_i^*) = \frac{w_i^*(n-w_i^*)}{n^2(n-1)} \left[\frac{n}{\bar{x}} \right]^2$$

$$s^2(\omega) = \frac{\sum_{j=1}^n w_j^* - \bar{x}(\omega)^2}{n-1}$$

$$\text{Var}(G_i) = \frac{w_i(n-1-w_i)}{(n-1)^2(n-2)} \left[\frac{\bar{x}(\omega)}{\bar{x}(\omega)} \right]^2$$

$$\bar{x}(\omega) = \frac{\sum_{j=1}^n w_j}{n-1}$$

TABELA 1.1: Dados de umidade de solo, em porcentagem de volume, medidos com sonda de nêutrons na profundidade de 25 centímetros.

Tubo nº	Faixa			Tubo nº	Faixa		
	1	2	3		1	2	3
1	39.0	39.0	37.6	33	41.2	39.3	41.8
2	40.3	37.8	39.1	34	38.9	38.3	44.4
3	38.9	37.6	36.8	35	38.2	38.8	42.5
4	37.3	36.9	38.6	36	36.6	38.1	43.4
5	38.4	37.5	37.9	37	41.2	37.1	42.0
6	39.5	35.0	39.5	38	39.3	38.4	39.5
7	41.8	36.4	37.6	39	41.2	37.8	40.2
8	39.8	38.9	38.2	40	38.3	42.5	41.9
9	37.0	40.3	38.3	41	43.9	40.1	41.8
10	37.5	36.5	34.9	42	44.5	39.3	42.5
11	40.7	37.5	38.2	43	39.5	41.4	44.7
12	41.3	37.6	39.7	44	44.6	45.9	43.4
13	36.4	39.0	37.0	45	43.9	43.8	45.9
14	42.9	37.5	37.3	46	41.6	43.5	44.1
15	39.1	37.5	41.4	47	46.4	46.7	44.6
16	39.1	37.0	39.4	48	44.6	45.2	45.4
17	40.6	36.9	39.6	49	45.1	44.8	44.7
18	40.3	39.0	39.7	50	43.8	40.5	45.6
19	40.4	37.4	41.2	51	44.1	45.7	49.3
20	41.2	40.1	42.6	52	42.6	44.0	47.1
21	43.6	40.6	42.8	53	44.7	45.8	46.7
22	41.5	39.4	40.4	54	48.2	46.9	46.1
23	42.1	43.2	40.8	55	45.1	47.3	46.0
24	42.5	39.8	39.9	56	48.3	41.6	48.0
25	42.5	34.8	39.1	57	47.6	44.4	45.6
26	39.4	37.6	40.3	58	47.8	42.4	47.9
27	34.6	38.7	39.0	59	42.3	42.1	45.9
28	34.9	38.6	42.5	60	41.6	39.5	48.0
29	34.8	39.3	43.9	61	45.0	45.7	42.2
30	35.6	38.2	42.5	62	39.5	43.2	38.3
31	38.9	39.7	44.7	63	40.0	42.7	38.3
32	34.7	37.8	40.5	64	42.7	42.2	40.8

Fonte: Reichardt et al. (1984).

(1,3)

média =

TABELA 1.1: Dados de umidade de solo, em porcentagem de volume, medidos com sonda de nêutrons na profundidade de 25 centímetros.

Tubo nº	Faixa			Tubo nº	Faixa		
	1	2	3		1	2	3
1	39.0	39.0	37.6	33	41.2	39.3	41.8
2	40.3	37.8	39.1	34	38.9	38.3	44.4
3	38.9	37.6	36.8	35	38.2	38.8	42.5
4	37.3	36.9	38.6	36	36.6	38.1	43.4
5	38.4	37.5	37.9	37	41.2	37.1	42.0
6	39.5	35.0	39.5	38	39.3	38.4	39.5
7	41.8	36.4	37.6	39	41.2	37.8	40.2
8	39.8	38.9	38.2	40	38.3	42.5	41.9
9	37.0	40.3	38.3	41	43.9	40.1	41.8
10	37.5	36.5	34.9	42	44.5	39.3	42.5
11	40.7	37.5	38.2	43	39.5	41.4	44.7
12	41.3	37.6	39.7	44	44.6	45.9	43.4
13	36.4	39.0	37.0	45	43.9	43.8	45.9
14	42.9	37.5	37.3	46	41.6	43.5	44.1
15	39.1	37.5	41.4	47	46.4	46.7	44.6
16	39.1	37.0	39.4	48	44.6	45.2	45.4
17	40.6	36.9	39.6	49	45.1	44.8	44.7
18	40.3	39.0	39.7	50	43.8	40.5	45.6
19	40.4	37.4	41.2	51	44.1	45.7	49.3
20	41.2	40.1	42.6	52	42.6	44.0	47.1
21	43.6	40.6	42.8	53	44.7	45.8	46.7
22	41.5	39.4	40.4	54	48.2	46.9	46.1
23	42.1	43.2	40.8	55	45.1	47.3	46.0
24	42.5	39.8	39.9	56	48.3	41.6	48.0
25	42.5	34.8	39.1	57	47.6	44.4	45.6
26	39.4	37.6	40.3	58	47.8	42.4	47.9
27	34.6	38.7	39.0	59	42.3	42.1	45.9
28	34.9	38.6	42.5	60	41.6	39.5	48.0
29	34.8	39.3	43.9	61	45.0	45.7	42.2
30	35.6	38.2	42.5	62	39.5	43.2	38.3
31	38.9	39.7	44.7	63	40.0	42.7	38.3
32	34.7	37.8	40.5	64	42.7	42.2	40.8

Fonte: Reichardt et al. (1984).

média, Desv.

corr. Z e 2

media = 40,28

Iar = 10,03

Desvio = 3,16

1) 40 pares - $r = 0,272$

$$r = 0,075$$

1º par: (39, 37.8)

40º par: (42.5, 40.1)

(39, 36.8)

(42.5, 41.4)

De acordo com os dados fornecidos em aula, temos:

Leste - Oeste

$$h = 300$$

$$\delta(h) = \frac{1}{2N(h)} [(40-42)^2 + (42-40)^2 + (40-39)^2 + (39-37)^2 + (37-36)^2 + (43-42)^2 + (42-39)^2 + (39-37)^2 + (37-41)^2 + (41-40)^2 + (40-38)^2 + (38-37)^2 + (37-35)^2 + (35-38)^2 + (38-37)^2 + (37-37)^2 + (37-38)^2 + (38-39)^2 + (35-38)^2 + (35-37)^2 + (35-36)^2 + (36-37)^2 + (37-38)^2 + (38-36)^2 + (36-35)^2 + (35-34)^2 + (34-33)^2 + (33-32)^2 + (32-29)^2 + (29-28)^2 + (38-37)^2 + (37-35)^2 + (29-30)^2 + (30-32)^2] =$$
$$= \frac{1}{2.36} \cdot 109 = \frac{109}{72} = 1,51889$$

$$h = 200$$

$$\delta(h) = \frac{1}{2N(h)} [(44-40)^2 + (40-40)^2 + (42-39)^2 + (40-37)^2 + (37-36)^2 + (42-43)^2 + (43-39)^2 + (42-39)^2 + (39-41)^2 + (39-40)^2 + (41-38)^2 + (37-37)^2 + (37-35)^2 + (37-38)^2 + (35-37)^2 + (38-37)^2 + (37-33)^2 + (37-34)^2 + (38-35)^2 + (35-38)^2 + (37-36)^2 + (38-35)^2 + (36-36)^2 + (35-35)^2 + (37-33)^2 + (35-33)^2 + (34-32)^2 + (33-29)^2 + (32-28)^2 + (38-35)^2 + (35-30)^2 + (30-29)^2 + (29-32)^2] =$$
$$= \frac{1}{2.33} \cdot 224 = \frac{224}{66} = 3,54545$$

$h = 300$

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \cdot [(44-42)^2 + (40-39)^2 + (42-37)^2 + (40-36)^2 + (42-42)^2 + (43-39)^2 + (42-41)^2 + (39-40)^2 + (39-38)^2 + (37-35)^2 + (37-38)^2 + (37-37)^2 + (35-37)^2 + (38-33)^2 + (37-34)^2 + (35-35)^2 + (38-37)^2 + (35-36)^2 + (37-35)^2 + (36-35)^2 + (35-34)^2 + (36-33)^2 + (35-32)^2 + (34-29)^2 + (33-28)^2 + (37-30)^2 + (30-30)^2] =$$

$$= \frac{1}{2 \cdot 27} (233) = 4,3148$$

$h = 400$

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \cdot [(44-40)^2 + (40-37)^2 + (42-36)^2 + (42-39)^2 + (43-41)^2 + (42-40)^2 + (39-38)^2 + (37-38)^2 + (37-37)^2 + (37-37)^2 + (35-33)^2 + (38-34)^2 + (35-37)^2 + (38-38)^2 + (35-35)^2 + (36-34)^2 + (35-33)^2 + (36-32)^2 + (35-29)^2 + (34-28)^2 + (38-30)^2 + (35-29)^2 + (30-32)^2] =$$

$$= \frac{1}{2 \cdot 23} (304) = 6,6087$$

$2N(h)$

$h = 500$

$$\begin{aligned}\hat{\sigma}(h) &= \frac{1}{2N(h)} \left[(44-39)^2 + (40-36)^2 + (42-39)^2 + (43-40)^2 + (42-38)^2 + (37-37)^2 + (37-37)^2 + \right. \\ &\quad (37-33)^2 + (35-34)^2 + (35-38)^2 + (38-36)^2 + (36-38)^2 + (35-32)^2 + (36-29)^2 + \\ &\quad \left. (35-28)^2 + (37-29)^2 + (35-30)^2 \right] \\ &= \frac{1}{2 \cdot 14} (310) = 9,117.6.\end{aligned}$$

$h = 600$

$$\begin{aligned}\hat{\sigma}(h) &= \frac{1}{2N(h)} \left[(44-37)^2 + (42-41)^2 + (43-38)^2 + (37-37)^2 + (37-33)^2 + (37-34)^2 + (35-36)^2 + \right. \\ &\quad (38-35)^2 + (36-32)^2 + (35-29)^2 + (36-28)^2 + (38-29)^2 + (37-30)^2 + (35-32)^2 \\ &= \frac{1}{2 \cdot 14} \cdot 333 = 11,892.9\end{aligned}$$

$h = 700$

$$\begin{aligned}\hat{\sigma}(h) &= \frac{1}{2N(h)} \left[(44-36)^2 + (42-40)^2 + (37-33)^2 + (37-34)^2 + (35-35)^2 + (36-29)^2 + (35-28)^2 + (38-30)^2 + (37-32)^2 \right] \\ &= \frac{1}{2 \cdot 9} (280) = 15,555.6\end{aligned}$$

$$h = 800 \quad \hat{\sigma}(h) = \frac{1}{2N(h)} \left[(42-38)^2 + (37-34)^2 + (36-28)^2 + (38-32)^2 \right] = \frac{1}{8} [125] = 15,625$$

De acordo com os dados fornecidos em anexo, temos:

Norte - Sul

$$h = 100$$

$$\hat{\sigma}(n) = \frac{1}{2N(n)} \left[(44-42)^2 + (42-37)^2 + (37-35)^2 + (35-36)^2 + (36-38)^2 + (37-38)^2 + (38-35)^2 + (35-37)^2 + (40-43)^2 + (43-37)^2 + (36-35)^2 + (42-42)^2 + (42-35)^2 + (35-35)^2 + (35-35)^2 + (40-39)^2 + (39-38)^2 + (38-37)^2 + (36-35)^2 + (42-37)^2 + (42-35)^2 + (35-35)^2 + (35-35)^2 + (40-39)^2 + (39-38)^2 + (37-38)^2 + (38-33)^2 + (37-41)^2 + (38-37)^2 + (37-34)^2 + (34-30)^2 + (39-39)^2 + (39-37)^2 + (37-38)^2 + (38-33)^2 + (35-29)^2 + (41-37)^2 + (37-36)^2 + (36-32)^2 + (32-27)^2 + (36-40)^2 + (40-33)^2 + (33-35)^2 + (35-29)^2 + (29-30)^2 + (38-34)^2 + (28-22)^2 \right] =$$

$$= \frac{1}{2 \cdot 36} (40) = 5,5694$$

$$h = 200$$

$$\hat{\sigma}(n) = \frac{1}{2N(n)} \left[(44-37)^2 + (42-35)^2 + (37-36)^2 + (35-36)^2 + (37-35)^2 + (38-37)^2 + (40-37)^2 + (37-36)^2 + (39-37)^2 + (38-34)^2 + (37-30)^2 + (42-35)^2 + (42-35)^2 + (35-35)^2 + (40-38)^2 + (39-37)^2 + (38-34)^2 + (37-30)^2 + (39-37)^2 + (39-38)^2 + (37-33)^2 + (37-37)^2 + (41-36)^2 + (37-32)^2 + (36-29)^2 + (39-37)^2 + (39-38)^2 + (37-33)^2 + (37-37)^2 + (41-36)^2 + (37-32)^2 + (36-29)^2 + (36-33)^2 + (40-35)^2 + (33-29)^2 + (35-30)^2 + (34-28)^2 \right]$$

$$= \frac{1}{2 \cdot 27} [525] = 9,7222$$

$h = 300$

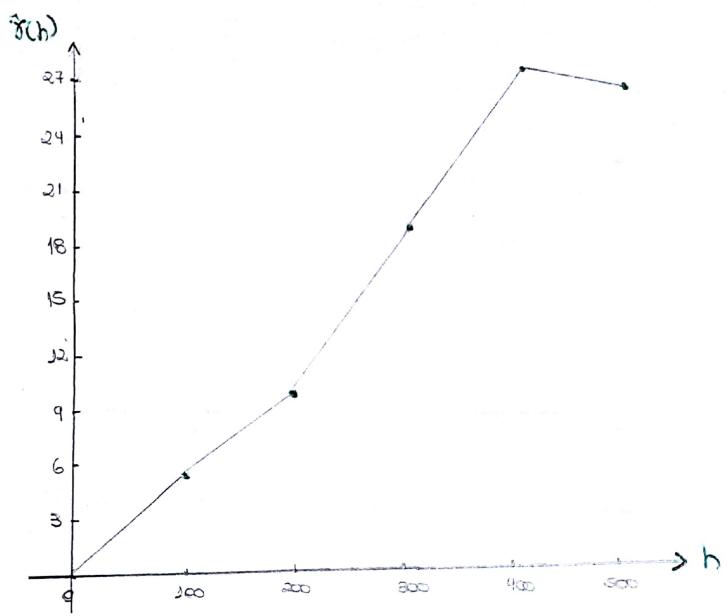
$$\hat{\sigma}(h) = \frac{1}{2N(h)} \left[(44-35)^2 + (42-36)^2 + (37-38)^2 + (37-37)^2 + (48-36)^2 + (37-35)^2 + (42-35)^2 + (42-35)^2 + (40-37)^2 + (39-34)^2 + (38-30)^2 + (39-38)^2 + (39-33)^2 + (37-36)^2 + (41-32)^2 + (37-29)^2 + (36-35)^2 + (40-29)^2 + (33-30)^2 + (38-28)^2 + (34-32)^2 \right]$$
$$= \frac{1}{2 \times 21} (785) = 18,6905,$$

$h = 400$

$$\hat{\sigma}(h) = \frac{1}{2N(h)} \left[(44-36)^2 + (42-38)^2 + (40-36)^2 + (43-35)^2 + (42-35)^2 + (40-34)^2 + (39-30)^2 + (39-33)^2 + (37-32)^2 + (41-29)^2 + (36-29)^2 + (40-30)^2 + (38-32)^2 \right]$$
$$= \frac{1}{2 \cdot 13} (716) = 27,5385,$$

$h = 500$

$$\hat{\sigma}(h) = \frac{1}{2.5} \left[(44-38)^2 + (40-35)^2 + (40-30)^2 + (37-29)^2 + (36-30)^2 \right] = \frac{1}{10} \cdot 261 = 26,1$$



$\tilde{f}(h) \times h$ para o caso Norte-Sul.

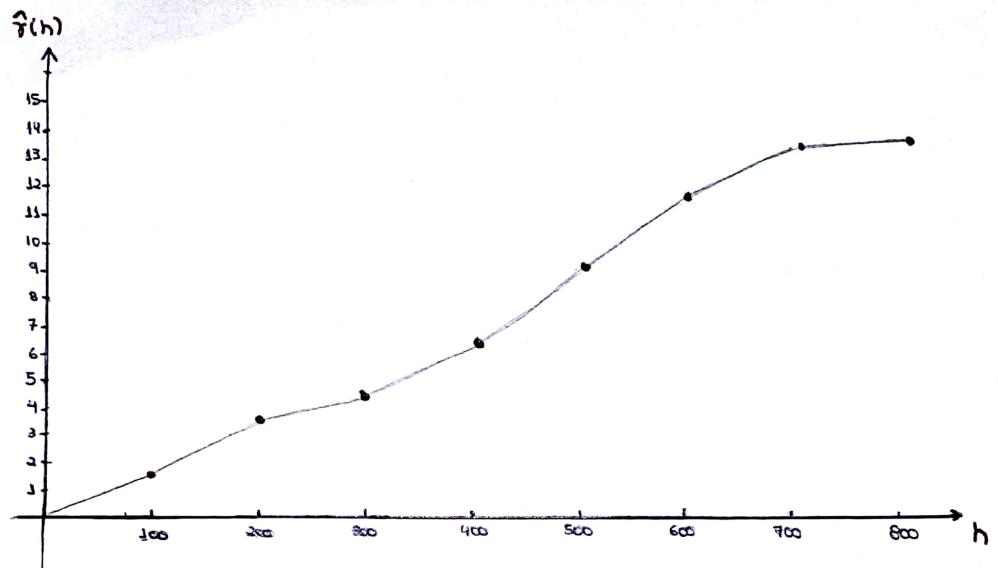
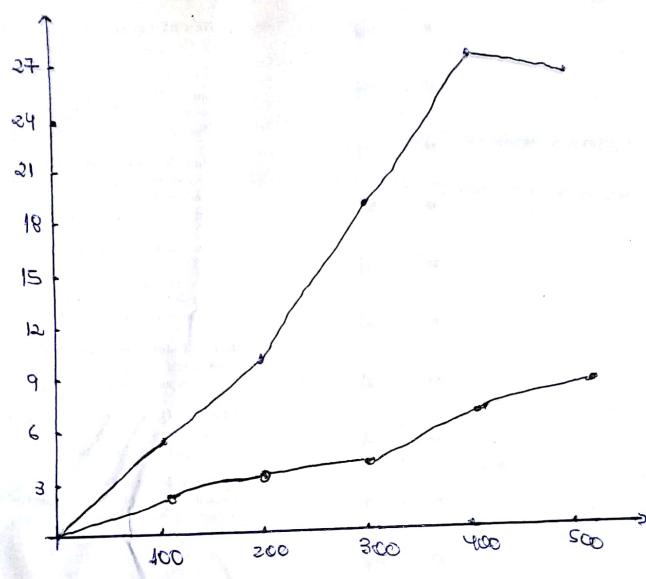
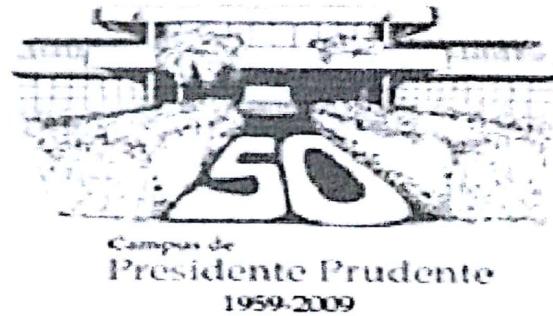


Gráfico $\hat{f}(h) \times h$ para o estudo Leste-Oeste



Maicon Aparecido Pinheiro

**Análise de dados de área e de Processos Pontuais: Uma aplicação
nos dados do Primeiro turno das eleições de 2010.**



PRESIDENTE PRUDENTE

2010



UNIVERSIDADE ESTADUAL PAULISTA
Faculdade de Ciências e Tecnologia
Câmpus de Presidente Prudente

Maicon Aparecido Pinheiro

**Análise de dados de área e de Processos Pontuais: Uma aplicação
nos dados do Primeiro turno das eleições de 2010.**

Trabalho apresentado à disciplina
Estatística Espacial, ministrada pela Profª
Dra. Vilma Mayumi Tachibana.

PRESIDENTE PRUDENTE

2010

1 INTRODUÇÃO

De acordo com o que foi solicitado em sala de aula, o objetivo deste trabalho se resume em calcular o índice Local de Moran para os estados Pará e Paraná em relação aos percentuais de votos recebidos pela candidata à presidência da república Marina da coligação do partido Verde tratando então cada estado como um polígono. Além disso, utilizar o método do vizinho mais próximo para verificar se há agrupamentos em relação aos estados em que a candidata Dilma do PT obteve mais de cinquenta por cento dos votos.

2 ÍNDICE LOCAL DE MORAN

Dado que o objetivo desta pequena atividade é aplicar estes dois métodos em dados de eleição no intuito de reforçar o que foi apresentado em aulas, separa-se este tópico apenas para o cálculo do Índice Local de Moran para os estados de Pará e Paraná.

Para executar tal tarefa, precisamos entender a situação geográfica destes dois estados e verificar quantos e quais são os estados com que fazem fronteiras e devido a isso segue a figura com o mapa do Brasil:



Figura 1: Mapa do Brasil.

Por este cartograma pode-se perceber que Paraná faz fronteira com Santa Catarina, São Paulo e Mato Grosso do Sul, enquanto que Pará com Amapá, Amazonas, Roraima, Mato Grosso, Tocantins, Maranhão. Os dados para estes estados são dados nos quadros abaixo:

% Marina	
	Pará
Amapá	29,71
Amazonas	25,71
Pará	37,69
Roraima	18,77
Tocantins	20,56
Maranhão	13,59
Mato Grosso	12

Quadro 1

% Marina	
	Paraná
São Paulo	20,77
Paraná	15,91
Mato Grosso do Sul	16,88
Santa Catarina	13,99

Quadro 2

Dessa forma, dado que os percentuais obtidos pela candidata Marina são apresentados nos quadros 1 e 2, o Índice Local de Moran, que é dado por

$$I_i = \frac{\left(y_i - \bar{y} \right) \sum_{j=1}^n w_j \left(y_j - \bar{y} \right)}{\left[\sum_{k=1}^n \left(y_k - \bar{y} \right)^2 \right] / n}, \quad i = 1, 2, 3, \dots, n,$$

para os estados de Paraná e Pará ficam dados por:

$$I_{\text{Paraná}} = \frac{(15,91 - 19,973) \frac{1}{3} [(20,77 - 19,973) + (16,88 - 19,973) + (13,99 - 19,973)]}{[1637,29]/27} = +0,185 \approx 0,185$$

$$I_{\text{Pará}} = \frac{(37,69 - 19,973) \frac{1}{6} [(29,71 - 19,973) + (25,71 - 19,973) + (18,77 - 19,973) + (20,56 - 19,973) + (13,59 - 19,973) + (12 - 19,973)]}{[1637,29]/27} = 0,02444$$

O que nos indica que não há relação entre os resultados tanto para Paraná como para Pará em relação a seus vizinhos.

3 MÉTODO DO VIZINHO MAIS PRÓXIMO

Considerando agora apenas os estados em que a candidata à presidência da república Dilma do PT obteve mais de cinquenta por cento dos votos, temos o seguinte mapa:

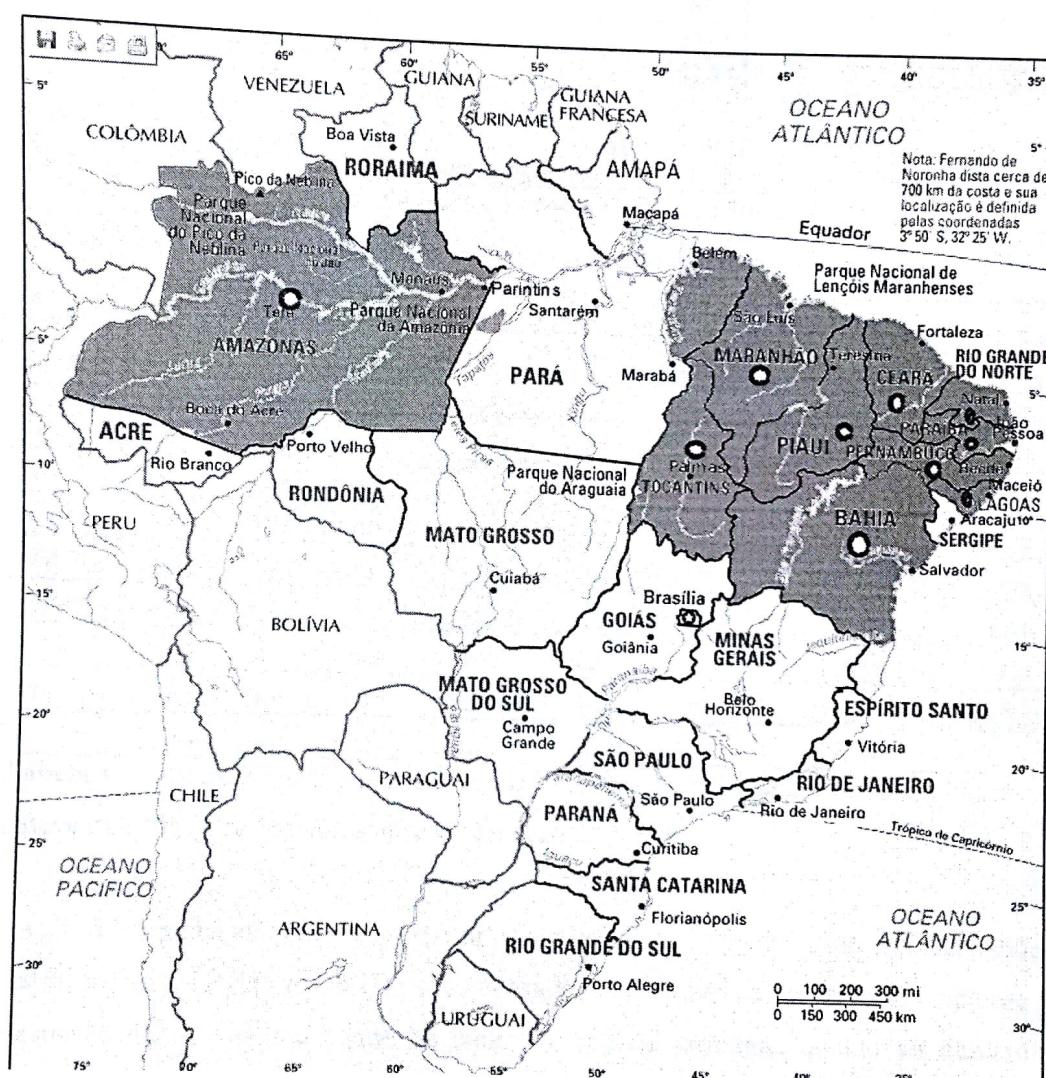


Figura 2 – Estados do Brasil onde Dilma obteve mais de 50% da votação.

Para responder se há ou não agrupamentos, utilizaremos aqui o método do vizinho mais próximo, considerando a distância entre as capitais de tais estados. Os percentuais obtidos pela candidata Dilma nestes estados são dados no quadro 3, logo abaixo:

Estados	% Dilma
Amazonas	64,98
Tocantins	50,98
Alagoas	50,92
Bahia	62,61
Ceará	66,3
Maranhão	70,65
Paraíba	53,21
Pernambuco	61,74
Piauí	67,09
Rio Grande do Norte	51,76

Quadro 3 – Estados % Dilma>50

As distâncias entre as capitais destes estados é dado na tabela abaixo:

	Manaus	Palmas	Maceió	Salvador	Fortaleza	São Luiz	João Pessoa	Recife	Teresina	Natal
Manaus		1509	5491	2605	5763	1746	5808	2833	1921	2765
Palmas	1509		1851	1114	2035	964	2253	1498	835	2345
Maceió	5491	1851		475	1075	1234	395	202	929	434
Salvador	2605	1114	475		1389	1323	949	839	994	1126
Fortaleza	5763	2035	1075	1389		1070	555	800	634	537
São Luiz	1746	964	1234	1323	1070		1660	1573	329	1607
João Pessoa	5808	2253	395	949	555	1660		120	1224	185
Recife	2833	1498	202	839	800	1573	120		934	297
Teresina	1921	835	929	994	634	329	1224	934		1171
Natal	2765	2345	434	1126	537	1607	185	297	1171	

Tabela 1 – Distância entre capitais dos estados onde a candidata a presidência Dilma

obteve mais de 50% dos votos apurados no primeiro turno.

Com base na tabela, pode-se verificar qual é o vizinho mais próximo de cada capital e assim construir $G^*(D)$ e $G_0(D)$, e compará-las no intuito de verificar a hipótese de aleatoriedade. A verdade é que há aqui, um grande problema quanto ao número de quadrantes e, em consequência, na determinação do lambda. Como aqui tratamos de capitais, decidiu-se nesse trabalho dividir o Brasil em áreas iguais a de 1528 Km^2 , valor correspondente a área da cidade de São Paulo de acordo com o site <http://www.cidados.com.br/cidade/sao_paulo/004784.html> no dia 02 de novembro de 2010. Dessa maneira, temos que

$$\lambda = \frac{10 \cdot 1.528}{8.511.925} = 0,001795$$

onde o denominador representa a área territorial brasileira segundo o site

<http://www.portalbrasil.net/brasil.htm> no dia 02 de novembro de 2010. Assim o sendo, temos a seguinte tabela para os valores assumidos pelas funções $G^*(D)$ e $GO(D)$.

	$G^*(D)$	$GO(D)$
1,2	0,2	0,000811709254869
1,85	0,3	0,001928140847609
2,02	0,4	0,002298357085818
3,29	0,6	0,006085291049379
4,75	0,7	0,012642752424633
5,37	0,8	0,016130080088314
8,35	0,9	0,038554718453209
15,09	1	0,120505715946297

Tabela 2 – Valores para $G^*(D)$ e $GO(D)$

O que nos indica que existe um agrupamento espacial em relação a localização dos estados em que a candidata Dilma obteve mais de 50% dos votos apurada, o que de fato, pelo mapa anteriormente apresentado é uma afirmação verdadeira, já que existe grande concentração no nordeste.

4 CONCLUSÃO

Através deste trabalho pode-se compreender um pouco mais sobre o que significa o Índice Local de Moran e entender o que o número resultante significa. Além disso, houve um outro grande aprendizado que é o em relação ao método do vizinho mais próximo, que se mostra fortemente influenciado pela forma com que a dividimos a área em estudo.

Anexos (Valores Para todos os estados)

	V Dilma	V Marina
Região Norte		
Acre	23,92	23,45
Amapá	47,38	29,71
Amazonas	64,98	25,71
Pará	47,93	37,69
Rondônia	40,73	12,71
Roraima	28,72	18,77
Tocantins	50,98	20,56
Região Nordeste		
Alagoas	50,92	11,5
Bahia	62,61	15,75
Ceará	66,3	16,36
Maranhão	70,65	13,59
Paraíba	53,21	20,3
Pernambuco	61,74	20,3
Piauí	67,09	11,41
Rio Grande do Norte	51,76	19,16
Sergipe	47,67	13,26
Centro Oeste		
Distrito Federal	31,74	41,96
Goiás	42,23	17,18
Mato Grosso	42,94	12
Mato Grosso do Sul	39,86	16,88
Sudeste		
Espírito Santo	37,25	26,26
Minas Gerais	46,98	21,25
Rio de Janeiro	43,76	31,52
São Paulo	37,31	20,77
Sul		
Paraná	38,94	15,91
Rio Grande do Sul	46,95	11,33
Santa Catarina	38,71	13,99

Geo Da.

1º passo

Abrimos o Geo Da.

2º Abrimos o arquivo Columbus se Poly kD a segunda opção.

3º - Tool

 └ weight

 └ Create

para criar uma matriz dos pesos

Logo após

Space

↳ Univariate Moran

L crime

~~botão~~ seleciona a matriz.

Gera um gráfico com o valor I de Moran.

Clicando com o direito sobre o gráfico, há a opção aleatorização, a qual nos permitirá ver o valor p.

Ainda em

Space

↳ Univariate LISA

↳ Crime

↳ Selecionar o mapa de interesse

www.dpi.inpe.br

↓

pós-graduação

↓

análise espacial

↓

Laboratório

↳

Lab. de Áreas

↳ Lab-areas BD (spring)

abin.

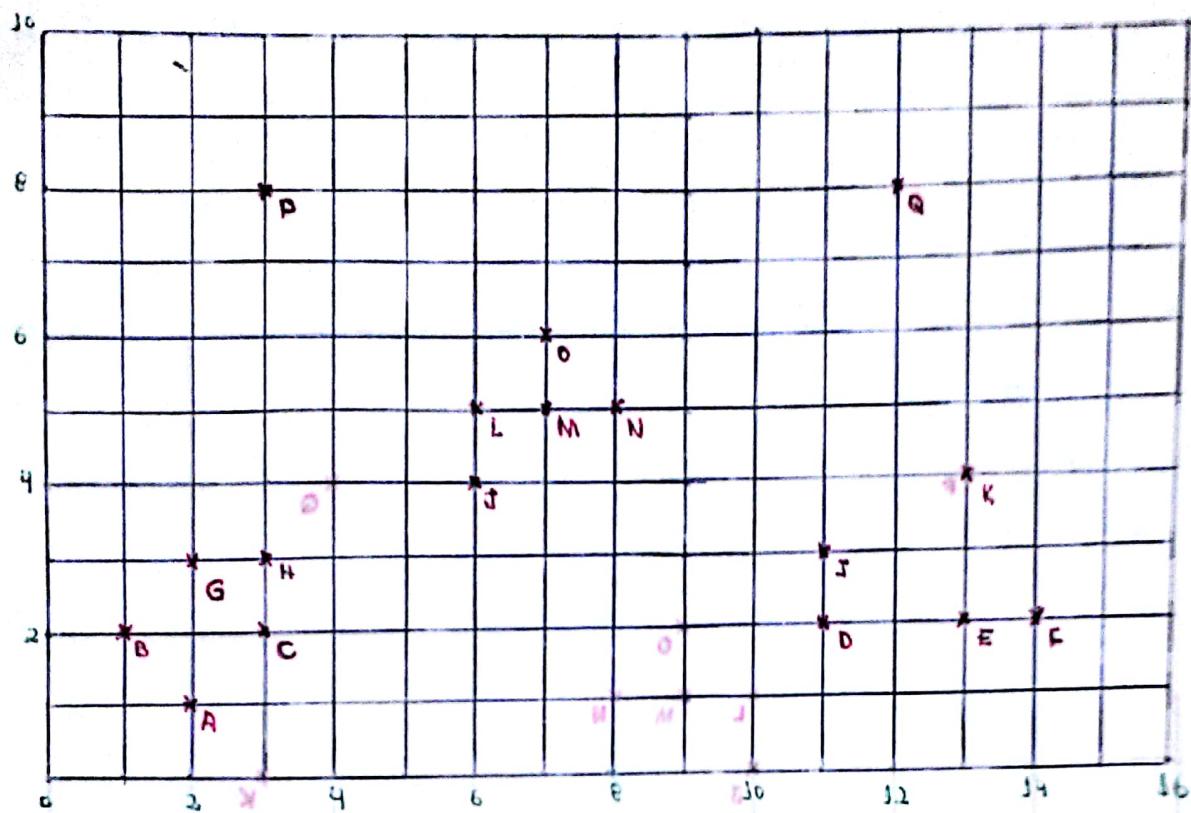
profissional técnico de inteligência

Atividade

Lab-Areas-Spr.pdf - Análise de padrões de área

nome: Maicon Aparecido Pinheiro

Atividade Estatística Espacial



De acordo com o que foi solicitado, para o ponto N^o Temps

$$d_{NA}^2 = (8-2)^2 + (5-1)^2 = 6^2 + 4^2 = 36 + 16 = 52$$

$$d_{NB}^2 = (8-1)^2 + (5-2)^2 = 7^2 + 3^2 = 49 + 9 = 58$$

$$d_{NC}^2 = (8-3)^2 + (5-2)^2 = 5^2 + 3^2 = 25 + 9 = 34$$

$$d_{ND}^2 = (8-3)^2 + (5-8)^2 = 5^2 + (-3)^2 = 25 + 9 = 34$$

$$d_{NE}^2 = (8-13)^2 + (5-2)^2 = (-5)^2 + 3^2 = 25 + 9 = 34$$

$$d_{NF}^2 = (8-14)^2 + (5-2)^2 = (-6)^2 + 3^2 = 36 + 9 = 45$$

$$d_{NG}^2 = (8-2)^2 + (5-3)^2 = 6^2 + 2^2 = 36 + 4 = 40$$

$$d_{NH}^2 = (8-3)^2 + (5-3)^2 = 5^2 + 2^2 = 25 + 4 = 29$$

$$d_{NI}^2 = (8-11)^2 + (5-3)^2 = (-3)^2 + 2^2 = 9 + 4 = 13$$

$$d_{NJ}^2 = (8-6)^2 + (5-4)^2 = 2^2 + 1^2 = 4 + 1 = 5$$

$$d_{NK}^2 = (8-11)^2 + (5-4)^2 = (-3)^2 + 1^2 = 9 + 1 = 10$$

$$d_{NL}^2 = (8-6)^2 + (5-5)^2 = 2^2 + 0^2 = 4$$

$$d_{NM}^2 = (8-7)^2 + (5-5)^2 = 1^2 + 0^2 = 1$$

$$d_{NO}^2 = (8-7)^2 + (5-6)^2 = 1^2 + (-1)^2 = 1 + 1 = 2$$

$$d_{NP}^2 = (8-3)^2 + (5-8)^2 = 5^2 + (-3)^2 = 25 + 9 = 34$$

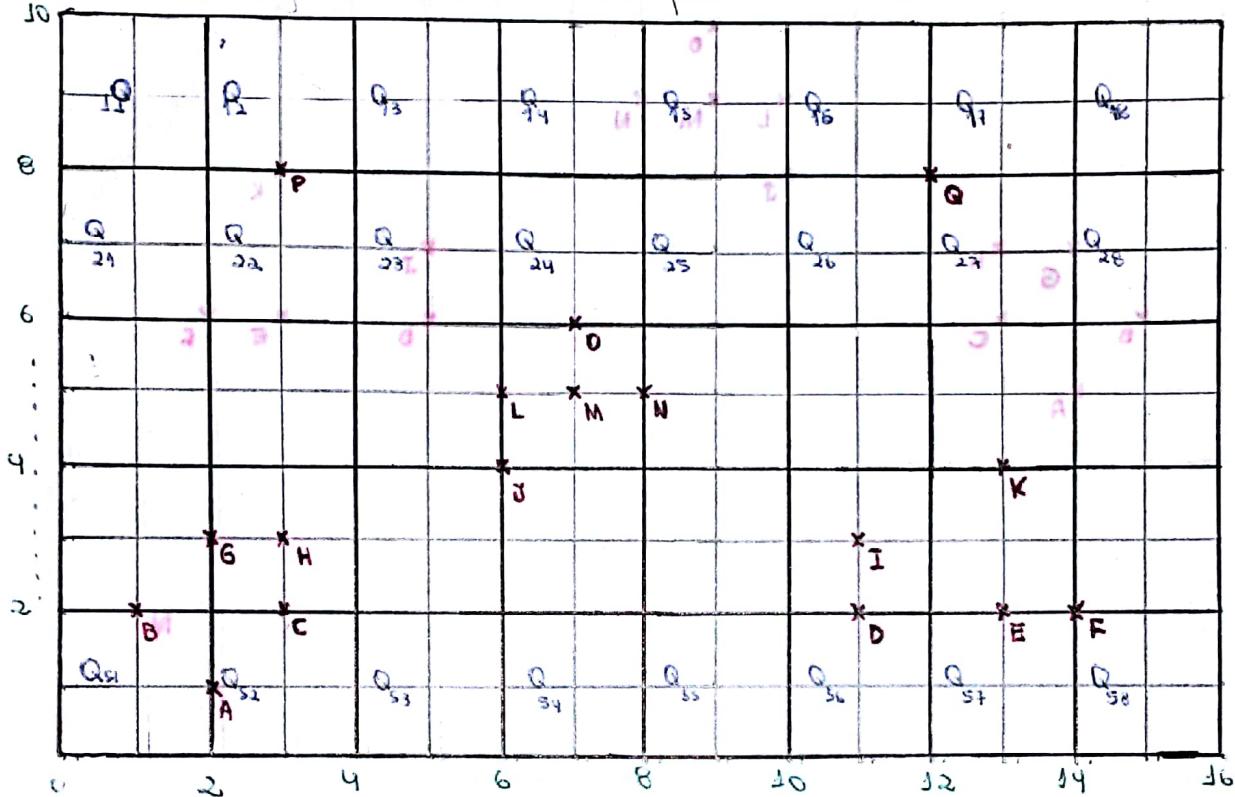
$$d_{NQ}^2 = (8-12)^2 + (5-8)^2 = (-4)^2 + (-3)^2 = 16 + 9 = 25$$

Assim, organizando as distâncias obtidas em um quadro, temos:

d^2	A	B	C	D	E	F	G	H	I	J	K	L	M	O	P	Q
N	52	58	34	34	34	45	40	29	13	5	10	4	1	2	34	25

onde $d^2 = d_i^2 N_i$ e $i = A, B, C, \dots, L, M, O, P, Q$.

Dividindo agora a área em quadrantes (2×2) vem



Contando então o número de eventos por Q_{ij} (quadrante da fila i e coluna j), tem-se a seguinte v.a. X^2 medida.

x_{ij}	0	1	2	3	Total
o_{ij}	27	10	2	1	40
e_{ij}	26,15	11,11	2,36	0,38	40
$P(X=x_{ij})$	0,6537	0,2778	0,059	0,0095	

Como parece seguir uma Poisson($0,425$) pois $\lambda = \frac{\text{total de pontos}}{\text{nº de quad.}} = \frac{17}{40}$

tem-se que estes pontos são aleatórios. Nesse Caso $\chi^2_{(r-1)(c-1)-1} \approx 0,91$

onde e_{ij} = número de eventos em Q_{ij} .

Obs: Para tal contagem usou-se a convenção: pontos no limite contar para a área inferior e para esquerda.

Tomando agora, $Q_{11}, Q_{12}, Q_{13}, \dots, Q_{18}, Q_{21}, Q_{22}, \dots, Q_{28}, \dots, Q_{58}$ como uma população, uma amostra aleatória de tamanho 20 desta é a seguinte:

$$\{Q_{21}, Q_{34}, Q_{11}, Q_{41}, Q_{24}, Q_{18}, Q_{55}, Q_{25}, Q_{37}, Q_{35}, Q_{43}, Q_{24},$$

$$\{Q_{21}, Q_{38}, Q_{45}, Q_{34}, Q_{48}, Q_{24}, Q_{28}, Q_{46}, Q_{32}\}$$

Associando a esta amostra o nº de eventos em cada Q_{ij} amostrado, vem, com auxílio da matriz E :

$$\{0, 3, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 3, 0, 0, 0, 1, 0\}$$

Supondo X : nº de eventos por Q_{ij} , a dist. de prob. de X fica

x	0	1	3
$p(X=x)$	$\frac{15}{20}$	$\frac{3}{20}$	$\frac{2}{20}$