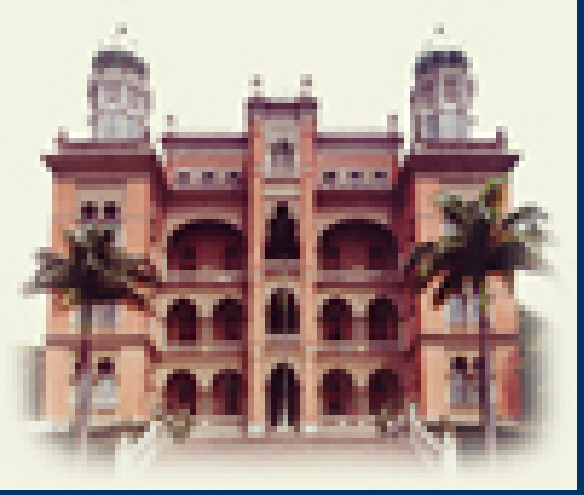


Fundação Oswaldo Cruz
Escola Nacional de Saúde Pública
Departamento de Epidemiologia

Estatística espacial

Padrão Pontual

Padrão de Pontos

- A análise de padrão de pontos, é o tipo mais simples de análise de dados espaciais. Baseia-se na localização dos eventos em determinada área a partir das coordenadas. O objetivo é estudar a disposição espacial dos pontos, a partir de suas coordenadas;
- Os processos pontuais são definidos como um conjunto de pontos cuja localização em $\Re^2$ foi gerada por um mecanismo estocástico.

Padrão de Pontos

- O modelo básico do banco de dados neste tipo de análise é:

Evento	Coord X	Coord Y
1	4,30	2,45
2	5,39	3,35
3	4,10	3,50

Conceito – 1ª ordem

- Os efeitos de primeira ordem, considerados globais ou de larga escala, correspondem a variações na média do processo no espaço. Neste caso, procuramos interessados na intensidade do processo, isto é, no número de eventos por unidade de área.

Conceito – 2ª ordem

- Efeitos de segunda ordem, denominados locais ou de pequena escala, representam a dependência espacial no processo, proveniente da estrutura de correlação espacial.

Completa Aleatoriedade Espacial

- A análise estatística dos padrões de distribuições de pontos requer um modelo teórico de referência, base para o desenvolvimento de métodos formais que checam a significância dos resultados exploratórios.

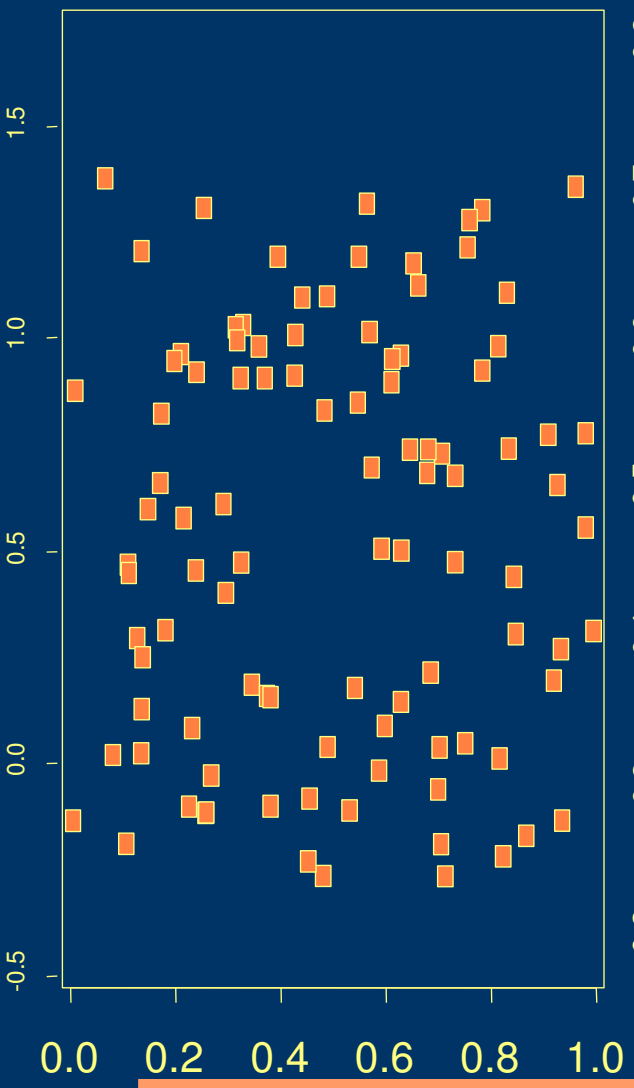
Completa Aleatoriedade Espacial

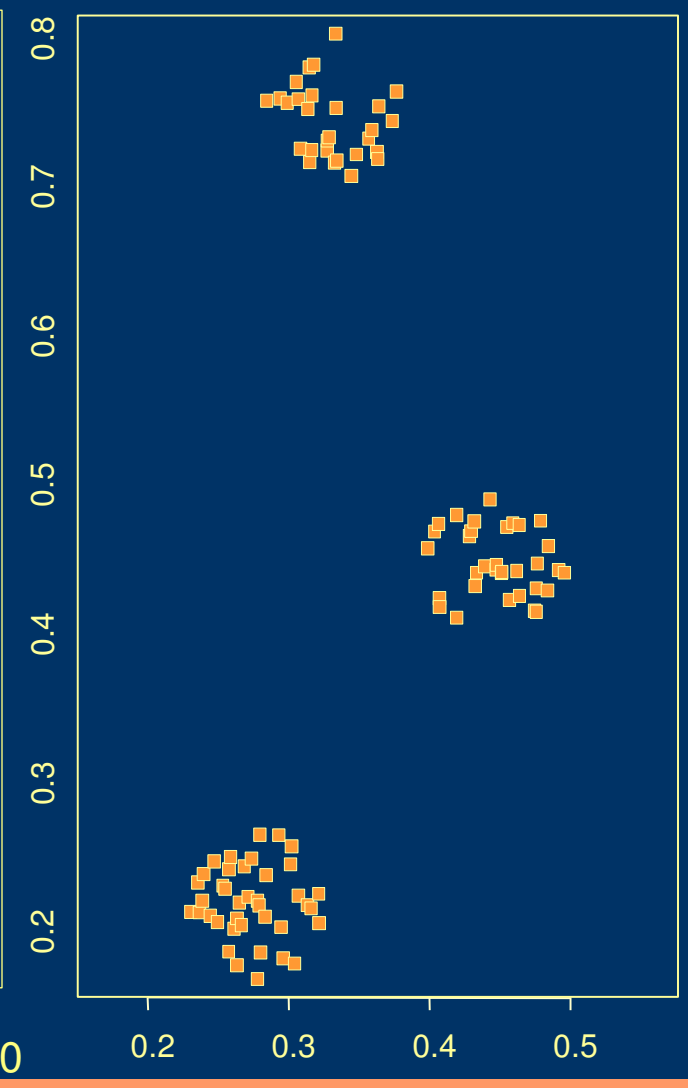
- O modelo teórico mais simples (e bastante aplicado na prática) é conhecido como aleatoriedade espacial completa ("*complete spatial randomness - CSR*").

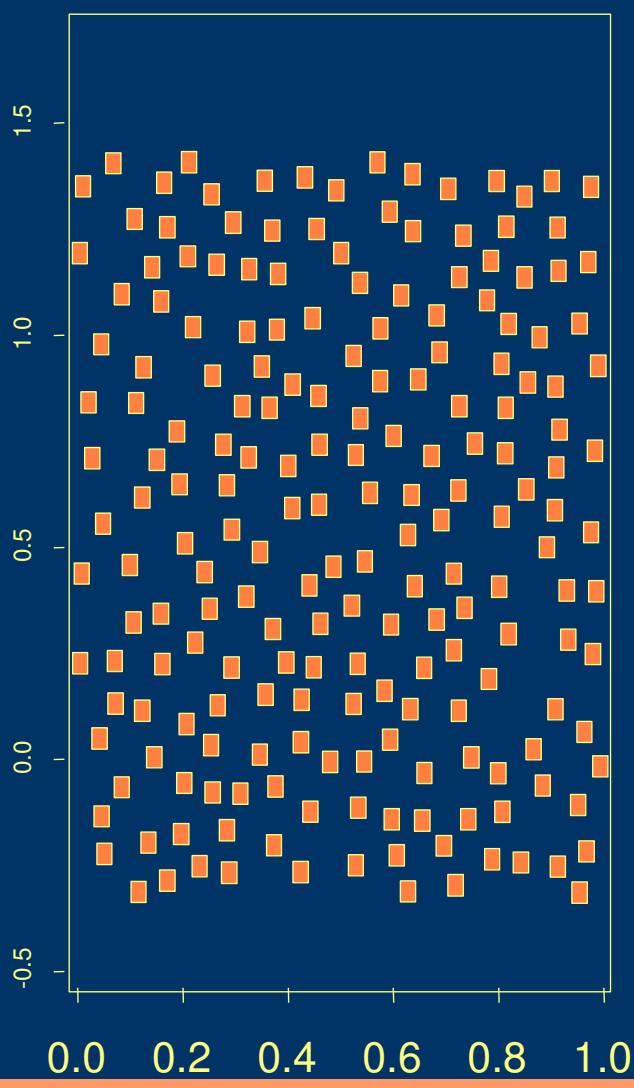
Completa Aleatoriedade Espacial

- A hipótese de CSR consideramos que as ocorrências em cada sub-área (S_i) são não-correlacionadas e homogêneas, e estão associadas à mesma distribuição de probabilidade de Poisson.
- Intuitivamente: eventos que ocorrem de forma independente uns dos outros têm igual probabilidade de ocorrência em toda a região.

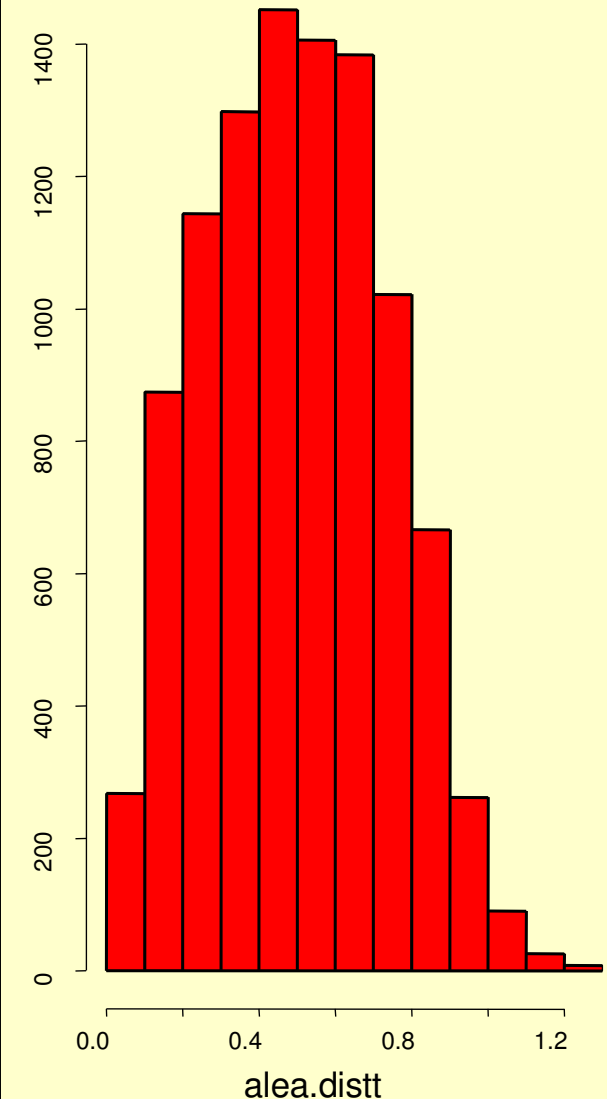
Padrões

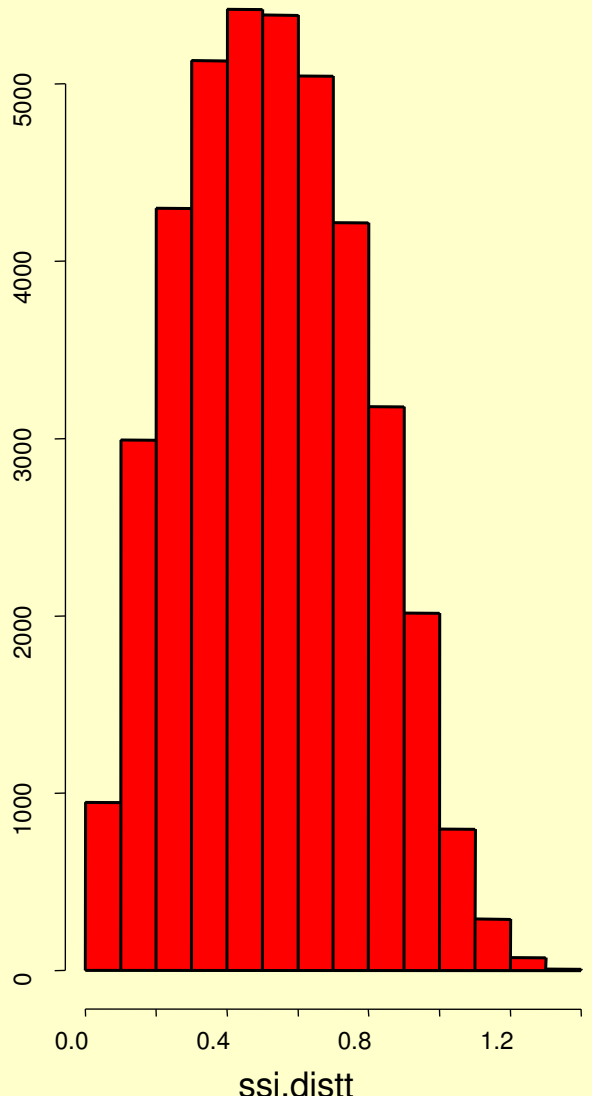

-0.5
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0


0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.2
0.3
0.4
0.5

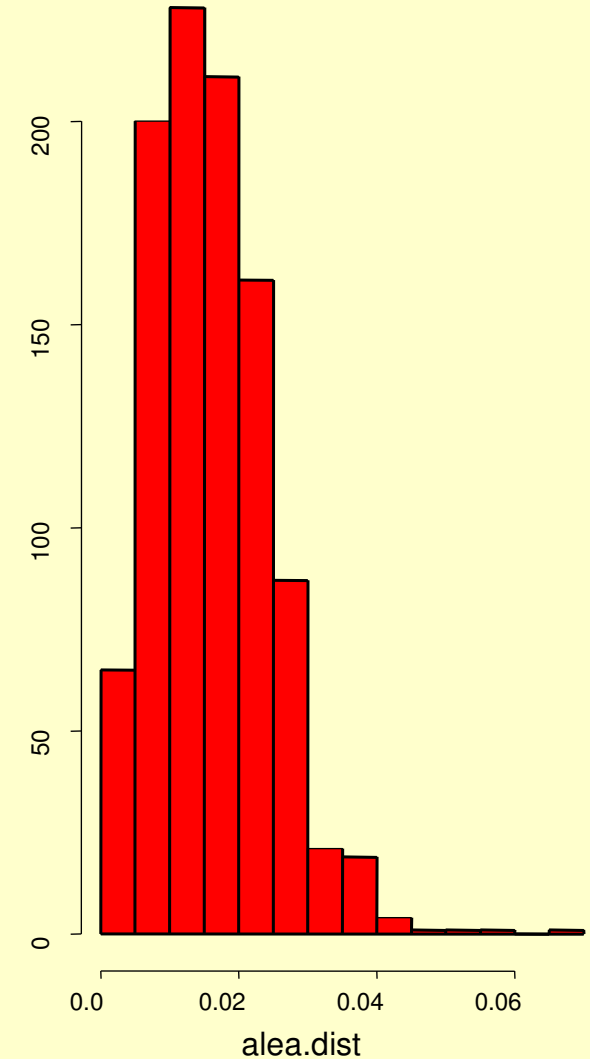

-0.5
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0

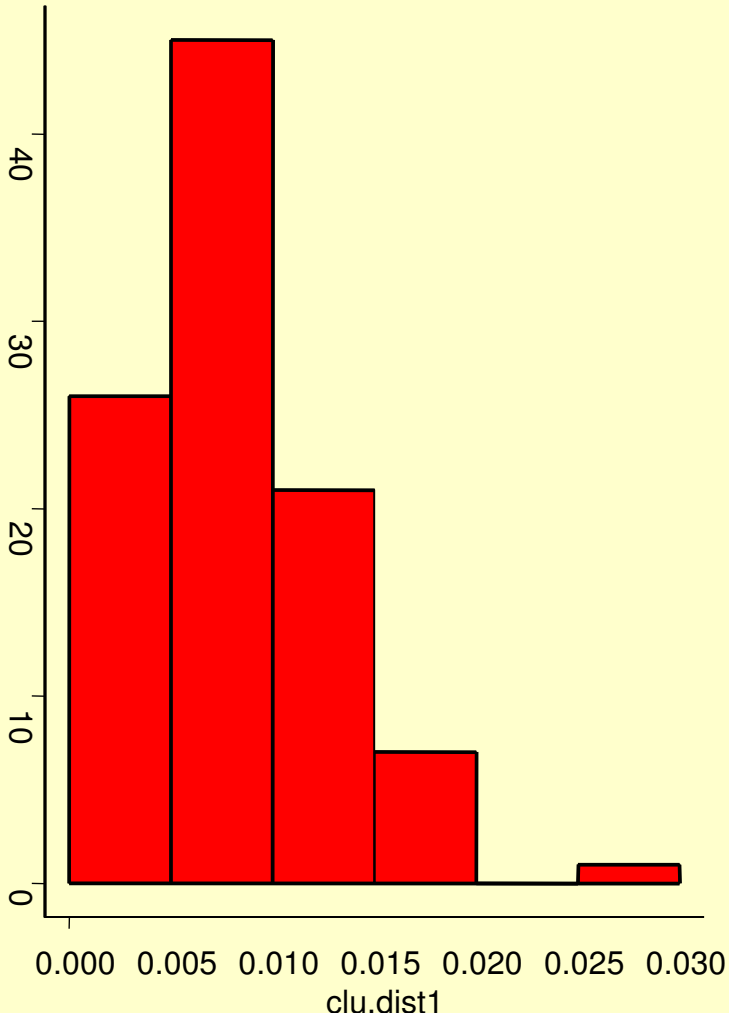
Distâncias entre todos

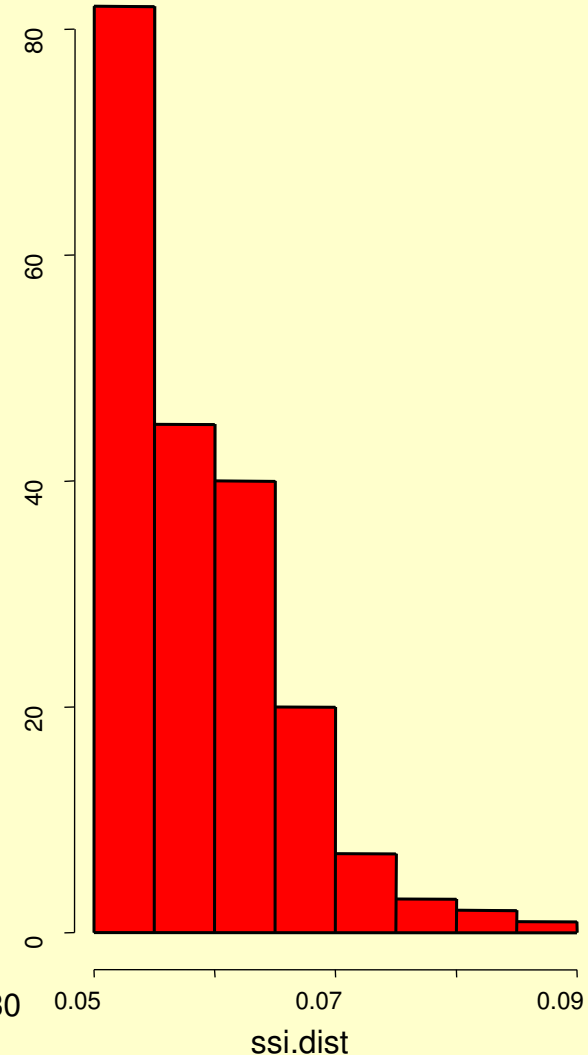




Distâncias – 1º vizinho



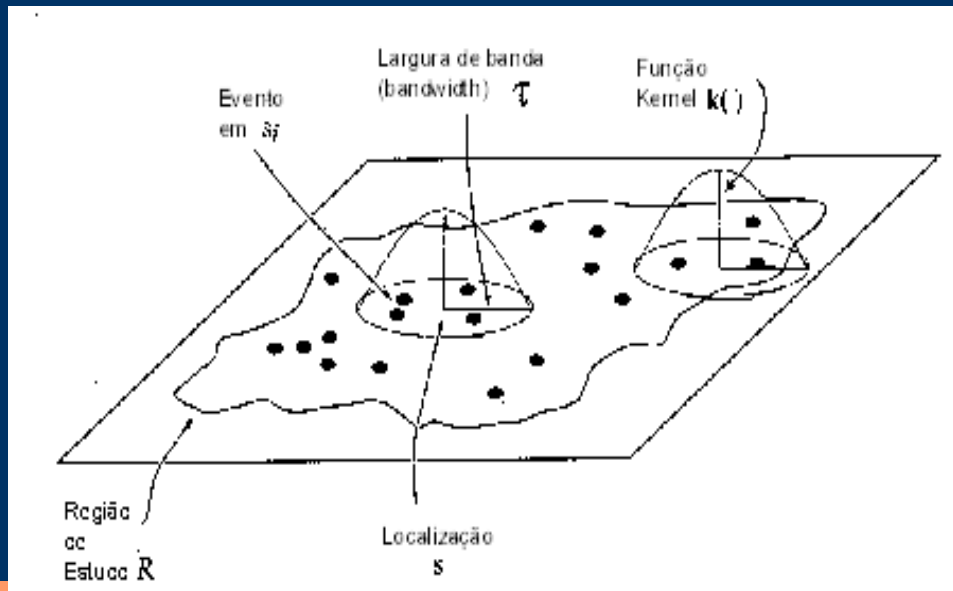




Kernel

- Técnica de alisamento que utiliza janela móvel e função que dá a cada área um peso variável conforme a distância.

- Estimar a intensidade de pontos dispostos no espaço é semelhante a estimar uma densidade de probabilidade bivariada.



Kernel

$$\hat{\lambda}(s) = \sum_{i-1}^{n} \frac{1}{\tau^2} k\left(\frac{(s - s_i)}{\tau}\right)$$

τ

$\hat{\lambda}(s)$ - valor estimado por área;

τ - largura da banda (fator de alisamento);

$k()$ - função de ponderação *kernel*;

s - centro da área;

s_i - local do ponto.

Kernel

$$\delta_\tau(s) = \int_R \frac{1}{\tau^2} k\left(\frac{(s-u)}{\tau}\right) d\,\boldsymbol{u}$$

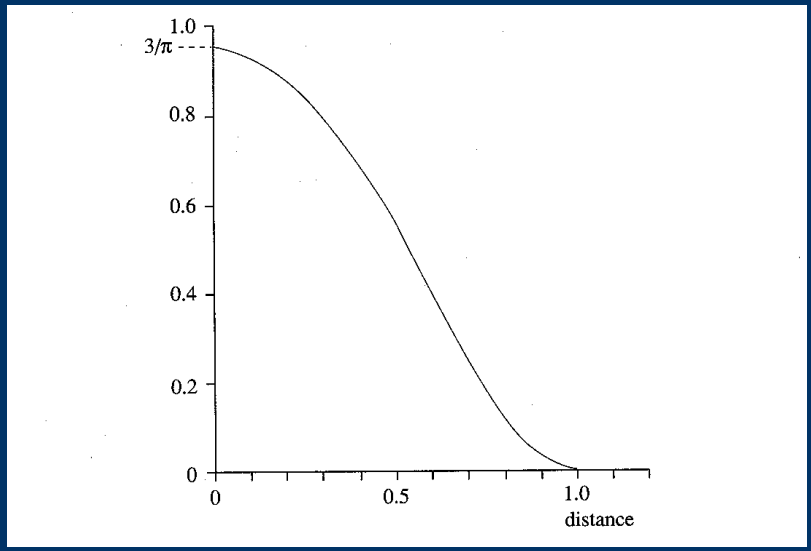
- Aplicando a correção das bordas obtém-se um estimador corrigido

- Deve-se fazer correção para as bordas
- Calcula-se o volume sob o Kernel que está de fato dentro da região de estudo

$$\hat{\lambda}(s) = \frac{1}{\delta_\tau(s)} \sum_{i-1}^{n} \frac{1}{\tau^2} k\left(\frac{(s-s_i)}{\tau}\right)$$

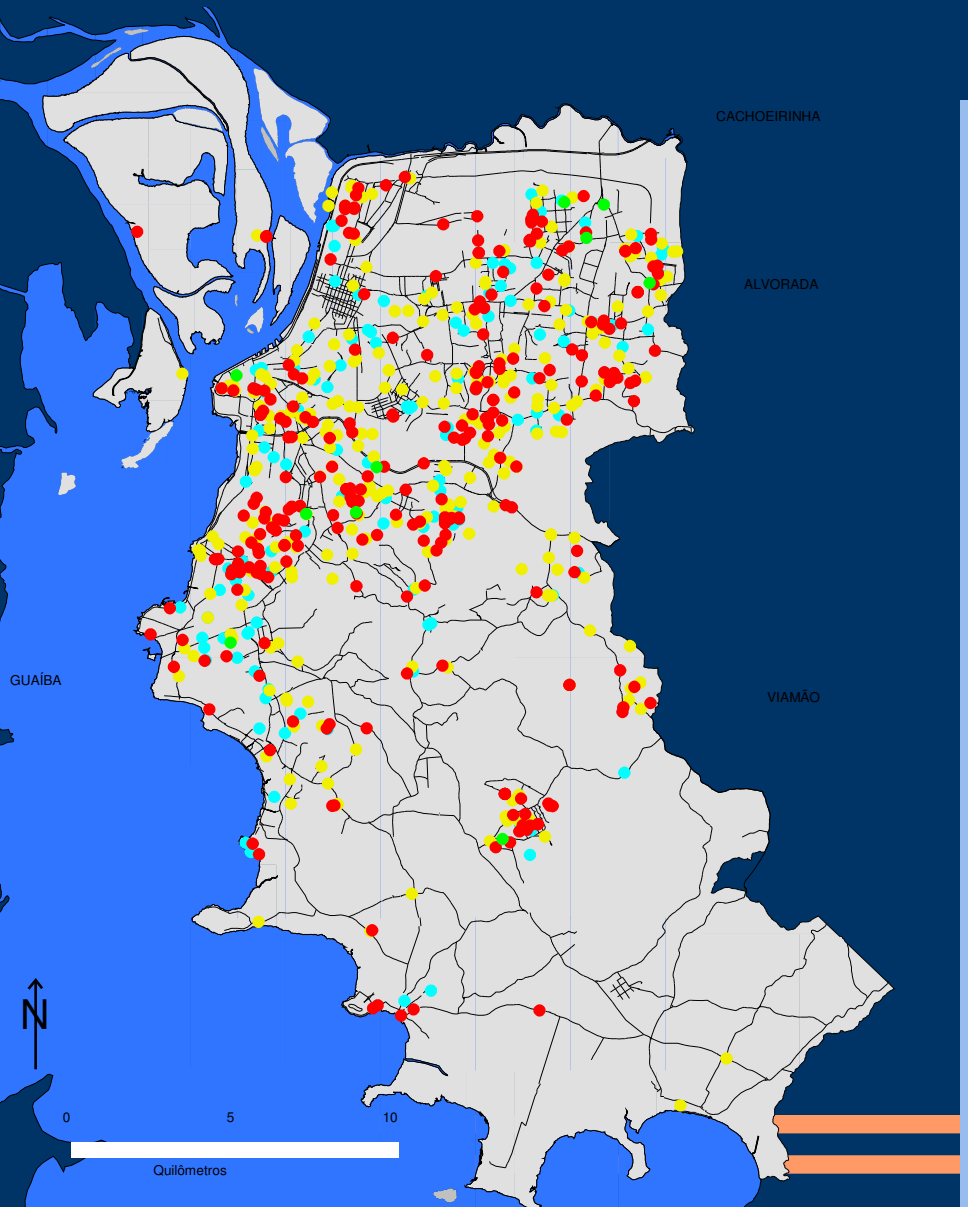
Kernel

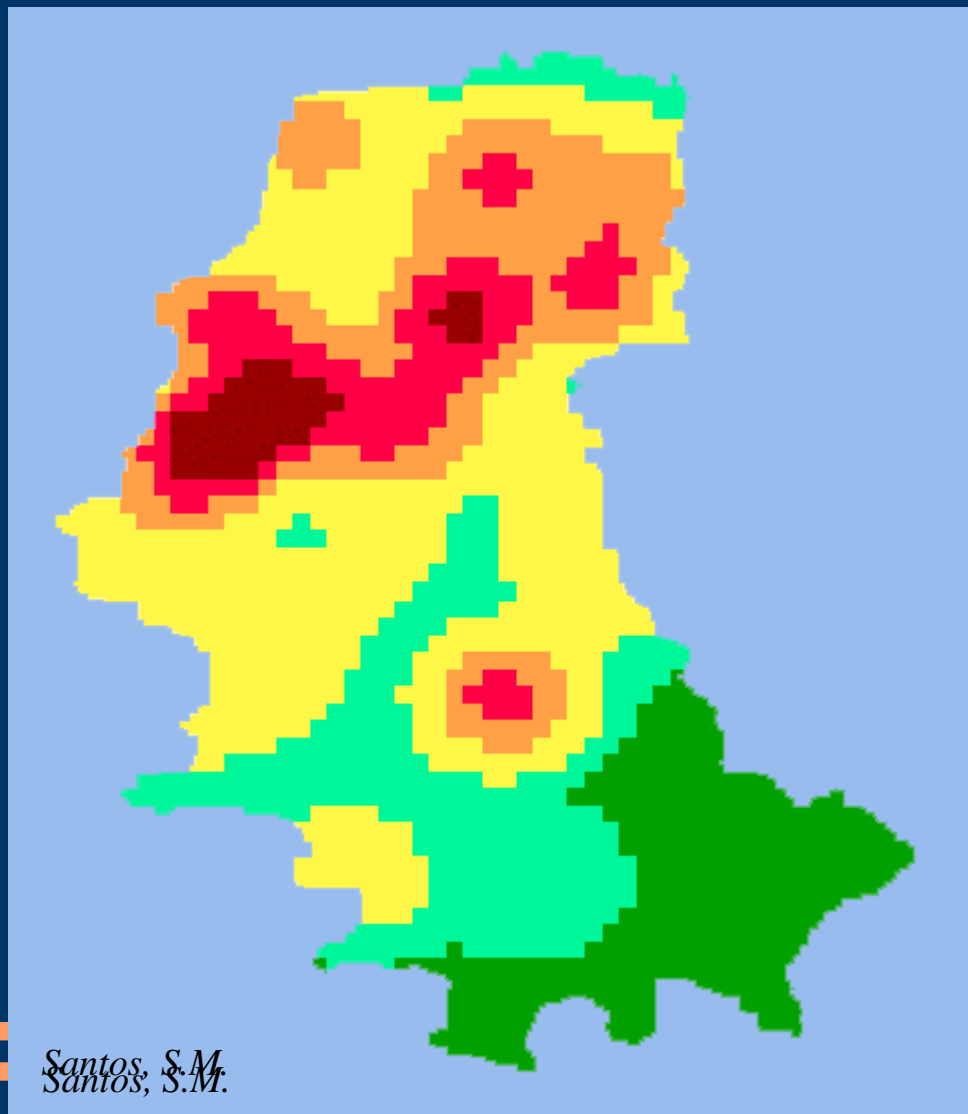
- A função de alisamento escolhida - *Kernel* - deve ser simétrica à origem
- Ex: Kernel quártico



- É possível estimar uma largura de banda ótima, por mínimos quadrados

Kernel





Santos, S.M.

Vizinho mais próximo

- Kernel e quadrat permitem explorar a variação da média do processo na região de estudo - propriedade de primeira ordem
- Para investigar propriedade de segunda ordem é necessário observar as distâncias entre os eventos

Vizinho mais próximo

- Dois tipos de distâncias: evento-evento (W) e ponto aleatório-evento (X)
- O resultado desta função empírica é o histograma das distâncias para o vizinho mais próximo - cada classe do histograma é uma contagem de eventos que ocorrem até aquela distância

Função K

- A funções anteriores somente permitem analisar a distribuição do vizinho mais próximo - pequena escala
- A função *K* permite analisar as propriedades de segunda ordem de um processo isotrópico

Função K - estimativa

- A função *K(h)* é, para cada distância *h*, o somatório do total de pares cuja distância é menor de que *h*, vezes o inverso do total de pares ordenados existente na região *R*.

Detecção de cluster

- Definição (Knox): grupo de ocorrências geograficamente limitado em tamanho e concentração tais que seja improvável ocorrer por mero acaso.
- São causas de cluster:
  - fonte comum,
  - contagiosidade.

Detecção de cluster

- Clusters são em geral espaço-temporais.
- É importante considerar:
  - Demais fatores de risco – sexo, idade;
  - Residência X outros locais;
  - Latência.

Detecção de cluster

- Dois tipos básicos de testes:
  - Focados – testa-se a hipótese de excesso de casos ao redor de fonte suspeita, identificada antes de observar os dados;
  - Genéricos – busca identificar áreas quentes, sem especificar quais e quantas.

Testes de Cluster

- H_0 é ausência de cluster: completa aleatoriedade espacial.
- *CSR*:

$$H_0 : y_i \sim \text{Poisson}\left(E_i = \lambda N_i \right), \text{ independentes, } i = 1, \cdots, n$$

Onde: n são subdivisões da região do estudo,

y_i nº de casos observados e E_i esperados,

λ eventos por unidade de área (e tempo)

Testes de Cluster

- Hipótese Alternativa:
  - Focados – λ varia com distância da fonte
  - Genéricos – existe regiões onde λ é mais elevado

Testes genéricos de Cluster

- Knox: testa um número acima do esperado de pares de casos excessivamente próximos (segundo critério pré-estabelecido) no espaço e no tempo.
- Mantel: $\sum\sum_{i \neq j} x_{ij} y_{ij}$

  distância no tempo e distância no espaço, se *x* for 1 e *y* for 1, equivale ao teste de Knox

Testes genéricos de Cluster

- Cuzick-Edwards - caso-controle onde a coincidência de casos vizinhos aumenta o peso, e a junção controle-controle ou caso-controle tem peso zero; este teste permite considerar a variação populacional.

Fonte específica

- Cluster ao redor de um ponto ou uma linha
- Compara-se a ocorrência de nº excessivo de “casos” em relação à população a partir de uma função de decaimento em relação à possível fonte

Fonte específica

$$\lambda(s) = \rho\lambda'(s) f(h;\theta)$$

$$f(h;\theta) = 1 + \theta_1 e^{\theta_2 h^2}$$

$\lambda(s)$ - estimativa do evento p/ unidade de área

ρ - parâmetro que indica a razão entre "casos" e "controles"

$\lambda'(s)$ - estimativa população p/ unidade de área

f - função da distância para a fonte

θ - parâmetros a estimar que descrevem como a incidência varia em torno da fonte

Variação da população

- O alisamento Kernel permite estimar eventos por unidade de área, sem considerar a população
- Pode-se estimar população por unidade de área, e fazer a razão dos dois obtendo uma estimativa alisada de eventos por população
- Pode-se usar outro evento como "estimador da população a risco"

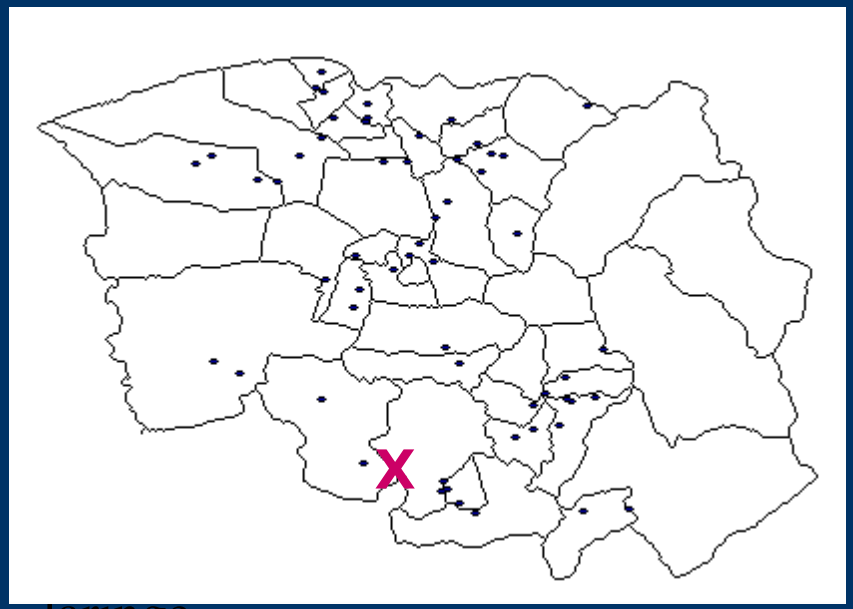
Variação da população

- A criação da taxa é a divisão dos alisamentos:

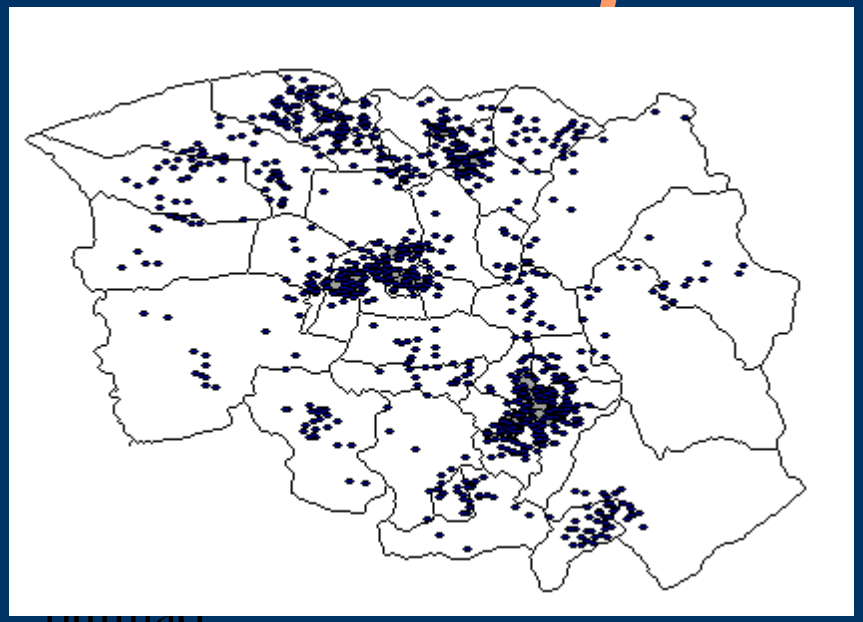
$$\frac{\text{eventos p/ unidade de área}}{\text{população p/unidade de área}}$$

$$\rho_\tau(s) = \frac{\sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{(s - s_i)}{\tau}\right)}{\sum_{j=1}^{m} \frac{1}{\tau^2} k\left(\frac{(s - s'_j)}{\tau}\right) y_j}$$

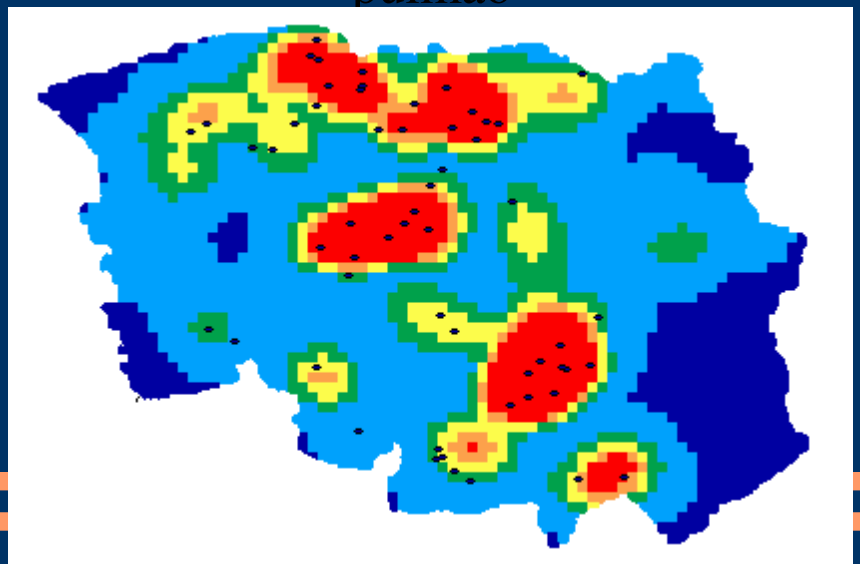
Razão de Kernel - exemplo



laringe
câncer de
laringe



câncer de
pulmão



kernel câncer de pulmão
casos de câncer de laringe

BAILEY & GATRELL, 1995

Caso-controle espacial

- O interesse não é estimar a variação da intensidade do processo na região, mas modelar a razão de risco entre casos e controles visando:
  - controlar fatores conhecidos,
  - identificar a variação espacial de determinação desconhecida

Caso-controle espacial

- Conjunto de pontos $x_i \in A : i = 1, \cdots, N$, onde n são casos e $m = N - n$ são controles.
- O interesse não é estimar a variação da intensidade do processo na região, mas modelar a razão de risco entre casos e controles, levando em conta fatores conhecidos, identificar a variação espacial de determinação desconhecida

Caso-controle espacial

- Uma forma natural são modelos logísticos:
  - Resposta é 0 ou 1 (casos e controles)
  - Co-variáveis individuais são incluídas
  - As coordenadas de casos e controles são incluídas no modelo através de uma kernel
  - Os parâmetros são estimados iterativamente
  - Testa-se por simulação se a variação espacial no risco é significativa

Caso-controle espacial

$$\text{logit}(y_i) = \beta x_i + g(s_i)$$

y_i é a variável resposta (sim/não, zero/um, casos/controles) e a função de ligação da regressão é o *logit*, como usual para dados binomiais,

$\mathbf{x}_i$ é o vetor de co-variáveis,

β é o vetor de parâmetros estimado pelo modelo, que no caso da regressão logística é a razão de chances (odds ratio) relacionada a cada co-variável,

$g(s_i)$ é a razão do estimador de intensidade kernel de casos e controles.

Caso-controle espacial

- Estima-se iterativamente:
  - Parâmetros da regressão logística;
  - Sobre os resíduos – kernel;
  - Inclui-se os valores do kernel no modelo, reestima-se os parâmetros das co-variáveis;
  - Repete-se até que não haja mais variação nos parâmetros estimados;

Caso-controle espacial

- A largura de banda pode ser definida pelo pesquisador ou estimada por validação cruzada.
- Testa-se se a variação espacial é significativa, H_0: $g(s)=0$ utilizando simulação.

Exemplo

- Mortalidade infantil em Porto Alegre

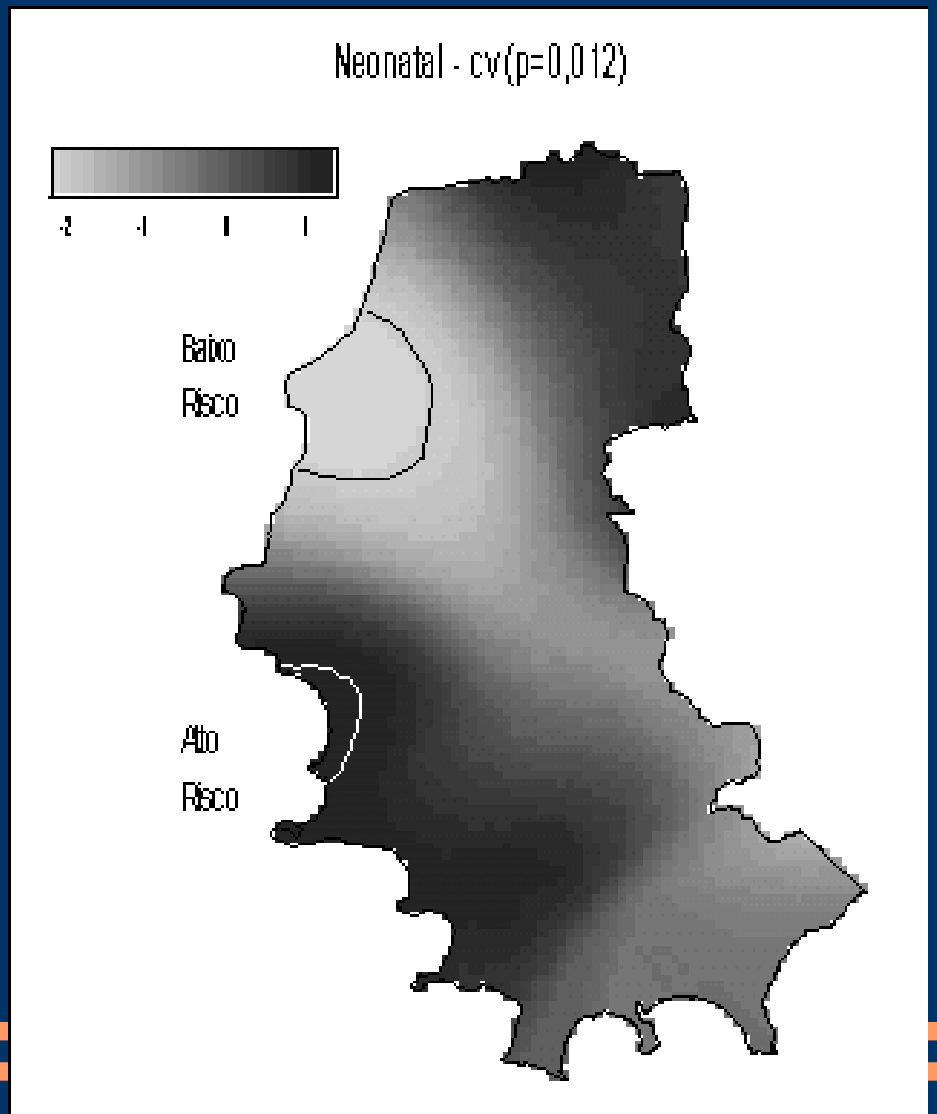
$$\log\left\{\frac{p(s,x)}{1-p(s,x)}\right\}=\beta_0+\beta_1 sexo+\beta_2 peso+\beta_3 idade+\beta_4 instr+\beta_5 ges+\beta_6 grav+\beta_7 parto+g(s)$$

Estimativas dos efeitos de covariáveis utilizando o valor da banda obtido por validação cruzada

Fator	Estimativa	Erro padrão	P-valor
Intercepto	40 717	0,9487	**0,0000**
Sexo	-0,3674	0,2713	0,1761
Peso ao nascer	-0,0018	0,0002	**0,0000**
Idade da mãe	-0,0131	0,0197	0,5059
Instrução da mãe	0,0718	0,2753	0,7942
Duração da gestação	11 685	0,3737	**0,0018**
Tipo de gravidez	-0,2006	0,6558	0,7598
Tipo de parto	-0,5320	0,2838	0,0613

Exemplo

- Mapa de risco para a mortalidade infantil, controlando para fatores individuais, com a largura de banda estimada por validação cruzada



Shimakura e cols. Cad Saúde Pública, 17(5):1251-61

www.maths.lancs.ac.uk/dept/stats/techabstracts02.html