Testes de conglomerados para dados de área

Renato Assunção
Departamento de Estatística
UFMG

Material da aula

- Introdução através de um exemplo
- Testes genéricos de conglomerados
  - Whittemore, GAM, Besag e Newell,
- Testes focados de conglomerados
  - Bithell e Stone
- Conclusões

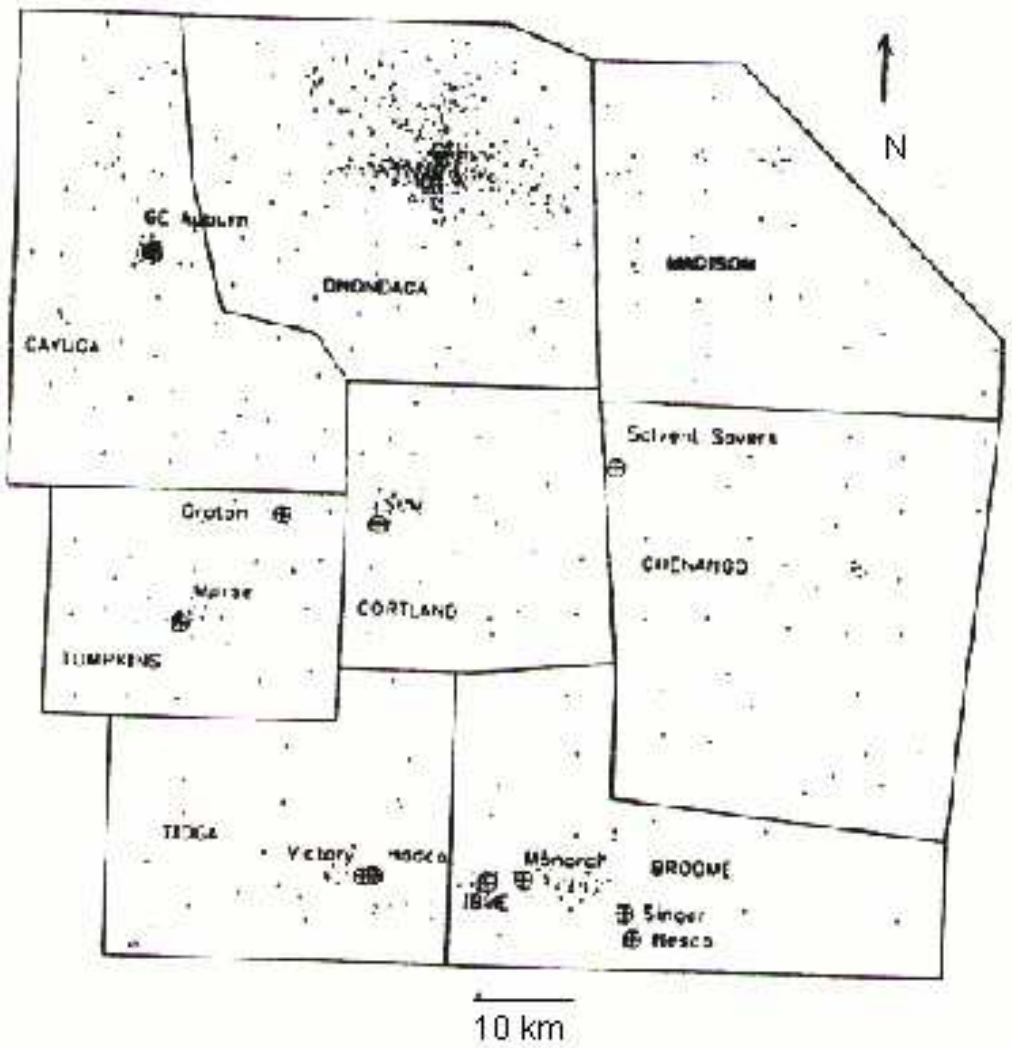
Estudo de Caso: Câncer em NY

- Lagatos et alli (1986) registraram um elevado número de casos de leucemia de 0 a 19 anos ao redor de depósitos de lixo tóxico contaminados com um composto orgânico volátil chamado tricloroethyleno.
- O Depto de Saúde do estado de NY iniciou uma série de estudos para averiguar o caso e ciar uma política pró-ativa de saúde (em contraste com uma política reativa).

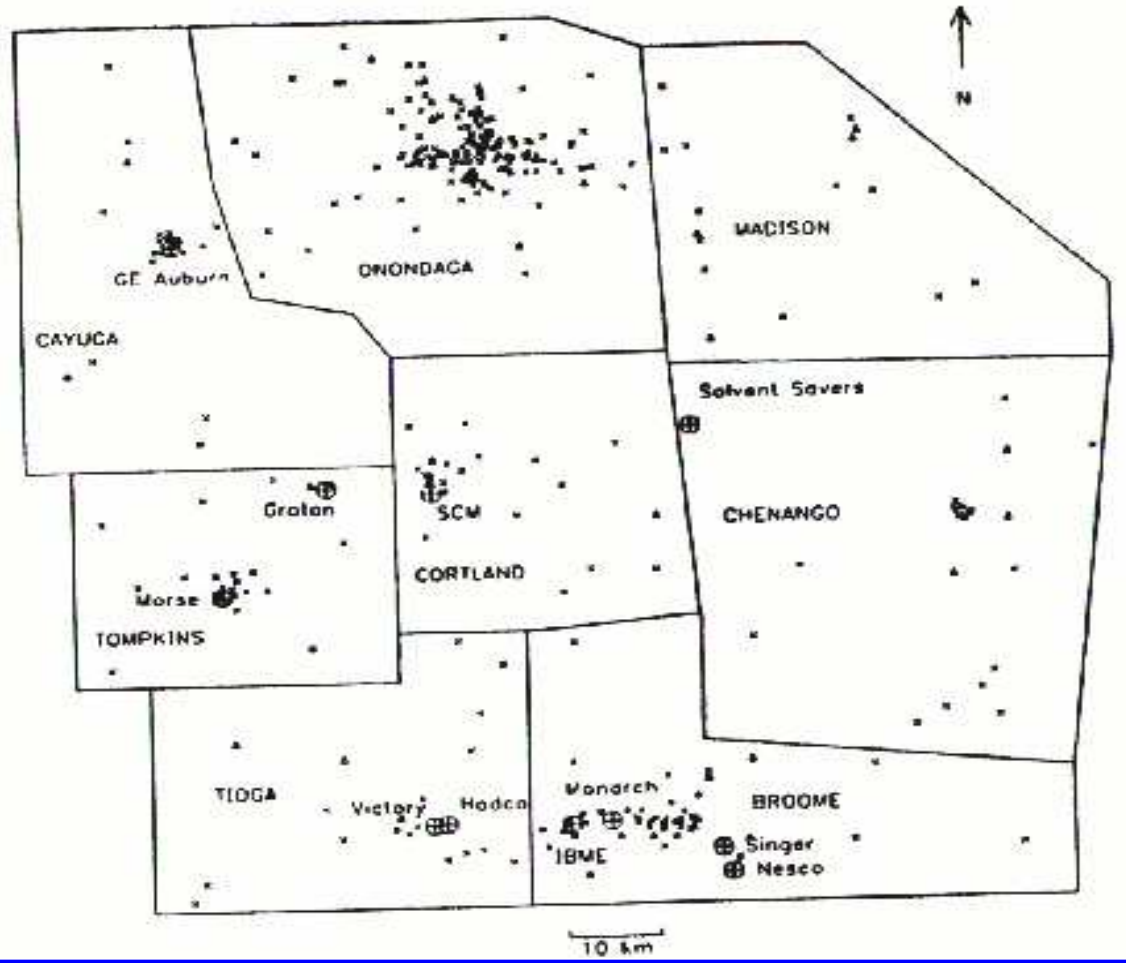
Região e População de Estudo

- Estudou-se o período de 1978-1982 numa região com pouco mais de 1 milhão de pessoas entre 0 e 19 anos.
- Região dividida em 790 áreas.
- 592 casos de leucemia no período.
- Apenas a área de residência do caso era conhecida.

	Total	Min	Max	Média
Casos	592	0	9	0.75
Pop	1.06 milhão	3	12221	1339



Mapa de região ao redor de Nova Iorque com posições dos centróides das áreas indicadas por . e sítios de processamento de lixo contendo trichloroethylene indicado por x (Waller et al. 1992)



Mapa da região de Nova Iorque com a localização dos casos de leucemia, 1978-1982. Centróides de células sem casos não são mostradas, com dois casos são indicadas por Δ e com 3 ou mais, por * (Turnbull et al. 1990).

Características do problema

- Dados da região: população e 592 casos agregados por área, 790 ao todo.
- 10% dos casos não tinham sua área de residência conhecida. Sabia-se apenas que pertenciam a uma dentre algumas áreas possíveis que formavam uma agregação maior. Estes casos foram repartidos entre suas possíveis áreas proporcionalmente às suas populações.
- Casos e população foram ***concentrados*** nos centróides das áreas.

Testes genéricos de conglomerados

- Hipótese nula é que todas as pessoas possuem o mesmo risco de contrair a doença independentemente dos outros casos e da posição de sua residência.
- Hipótese alternativa: ***hipótese nula não é válida***
- Número observado O*i* de casos na área *i* é variável aleatória
- Hipótese comum: O*i* ~ Poisson($\lambda\, i$)

$\lambda\, i$ = número esperado na área *i*

Se risco é constante na região, então

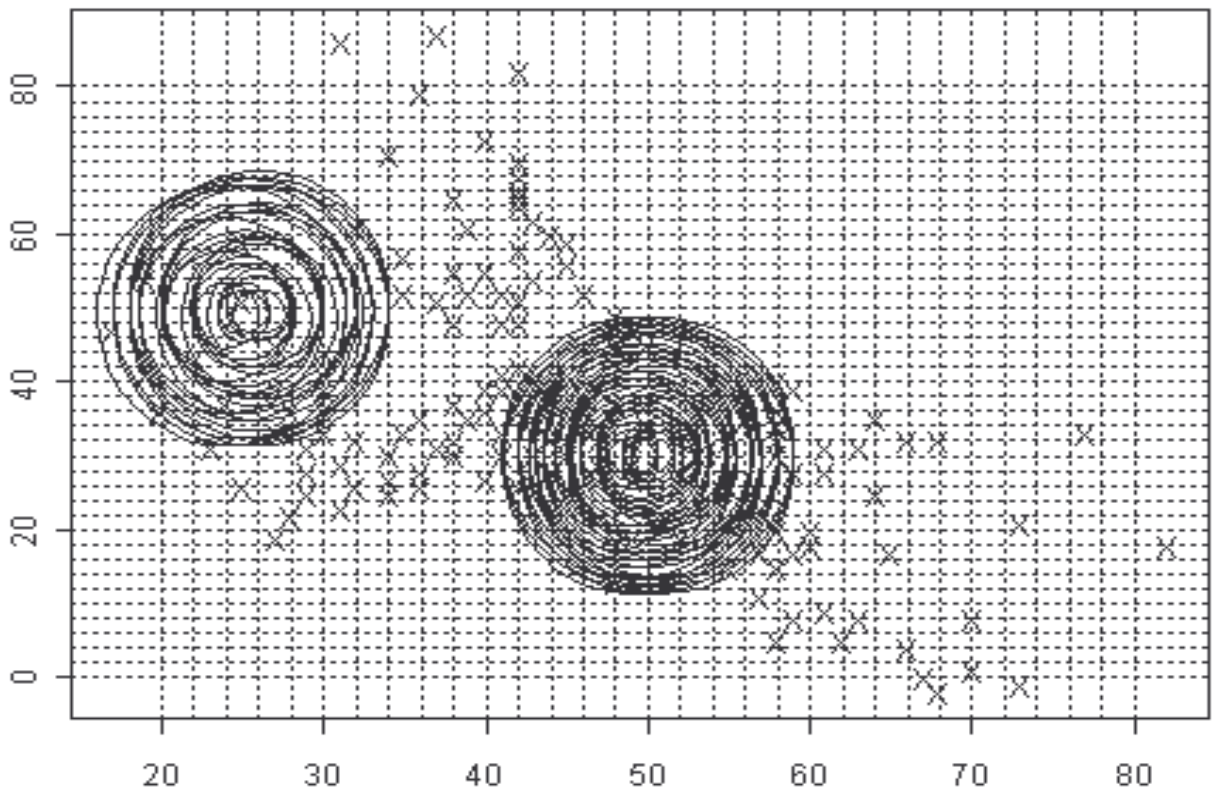
$$\lambda_i = r * Pop_i$$

Teste de Whittemore et al. (1987)

- Estatística T de teste é a distância média entre todos os pares de casos
- É assintoticamente normal sob H_o
- Para os dados: T observado é 60.24 Km e valor esperado é 59.01 Km com DP=0.96 produzindo z=1.28, não significativo.
- Desvantagens: teste global sem indicar onde estão os conglomerados; não possui controle de população subjacente

GAM: Geographical Analysis Machine de Openshaw (Openshaw et al., 1988, em Lancet)

Resultado Visual



GAM de Openshaw

- Crie grade bem fina sobreposta ao mapa com as posições dos centróides
- Associe os valores das áreas aos centróides
- Em cada nó da grade, fixe um círculo de raio r e calcule o p-valor associado com o número de eventos dentro do círculo
- Desenhe APENAS os círculos com p-valor < 0,002
- Refaça o procedimento acima aumentando o raio r do círculo
- Áreas de risco são identificadas por emaranhado de círculos

RESULTADO

Círculos de raios 1,2 e 4 que foram significativos ao nível nominal de 0,2% pelo método GAM de Openshaw et al. (1988)

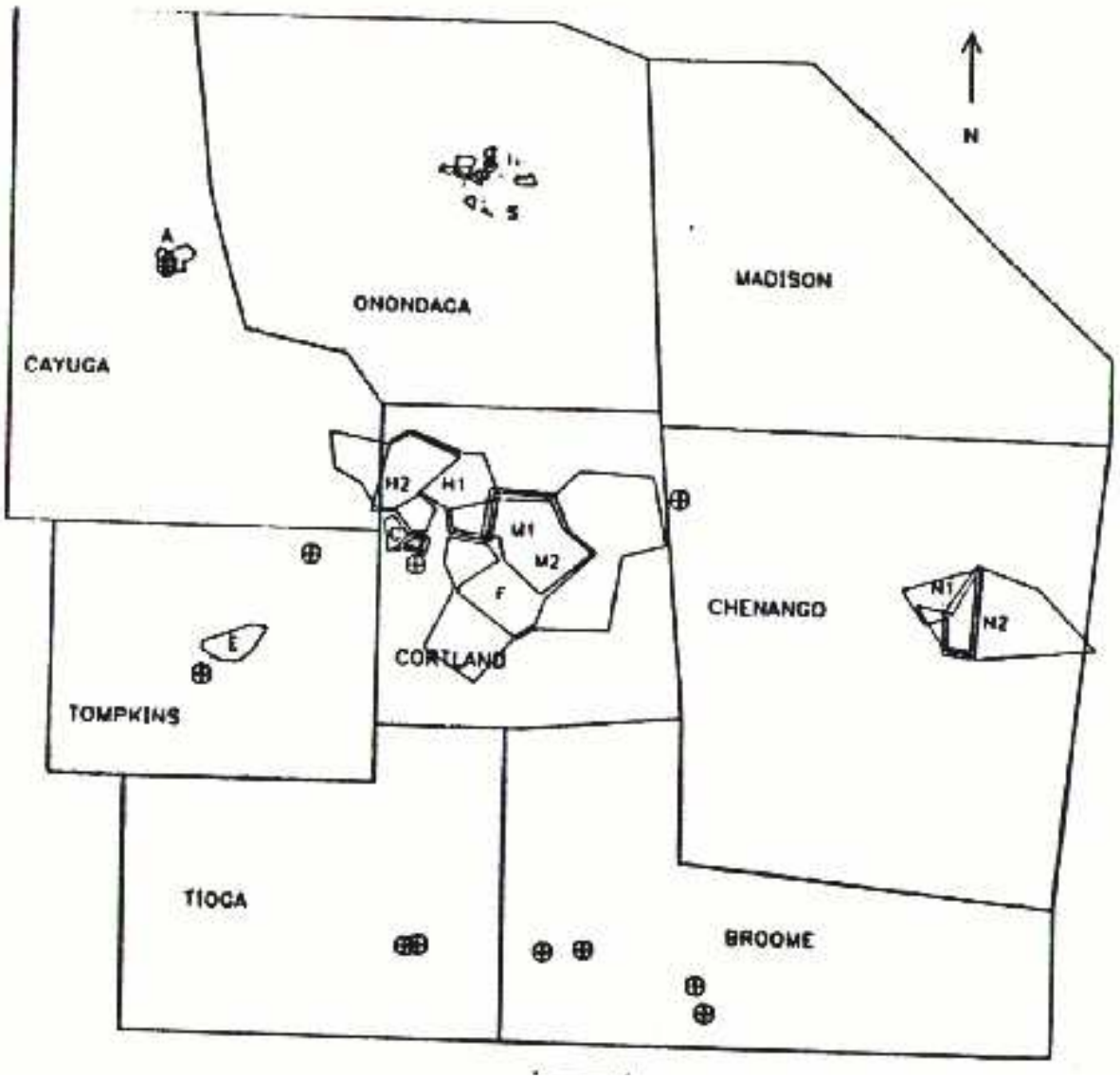
Vantagens e Desvantagens

- Intuitivo e visualmente impressionante
- Exploratório. Comparações múltiplas dão caráter nominal ao nível de significância
- Intensivo computacionalmente
- Círculos não são inteiramente comparáveis pois variáveis aleatórias possuem diferentes distribuições. Isto pode ser corrigido fixando-se o numerador ou o denominador

Método de Besag e Newell

- Método GAM de Openshaw fixa o raio do círculo e calcula P-valor do que encontra dentro do círculo
- Método de Besag e Newell fixa o número *k* de eventos que devem ser buscados e calcula o raio necessário para englobá-los. No círculo resultante, calcula p-valor.
- Desenha apenas os círculos significativos (p-valor < 0.005)
- Varia *k*

RESULTADO



Conglomerados de k=8 casos que foram significativos ao nível nominal de 5% usando método de Besag e Newell (1991).

Como é calculado o p-valor ?

- Em toda região, existem C casos e M pessoas em risco
- A área i possui pelo menos um caso.
- Seja L a variável aleatória que conta o número de outras áreas necessárias para acumular os k primeiros casos + próximos de i
- Seja l o valor observado de L seja m o número de pessoas em risco nessas l áreas.
- O número esperado de casos nessas l áreas é dado por

$$m \frac{C}{M}$$

P-valor para Besag e Newell

- O número de casos nas primeiras l áreas possui distribuição de Poisson com valor esperado $m\frac{C}{M}$
- Assim, $P(L \le l) = 1 - P(L > l + 1)$
- Mas isto é 1 - Probabilidade de que as l primeiras áreas possuam menos que k casos
- Desse modo, o p-valor é dado por

$$1 - \sum_{j=1}^{k-1} P(N = j) = 1 - \sum_{j=1}^{k-1} \frac{(mC/M)^j}{j!} e^{-mC/M}$$

Vantagens e desvantagens

- Estabiliza mais as estatísticas de teste locais
- Continua o problema de comparações múltiplas.
- Exploratório.
- Visualmente agradável.
- Como GAM, identifica os conglomerados.

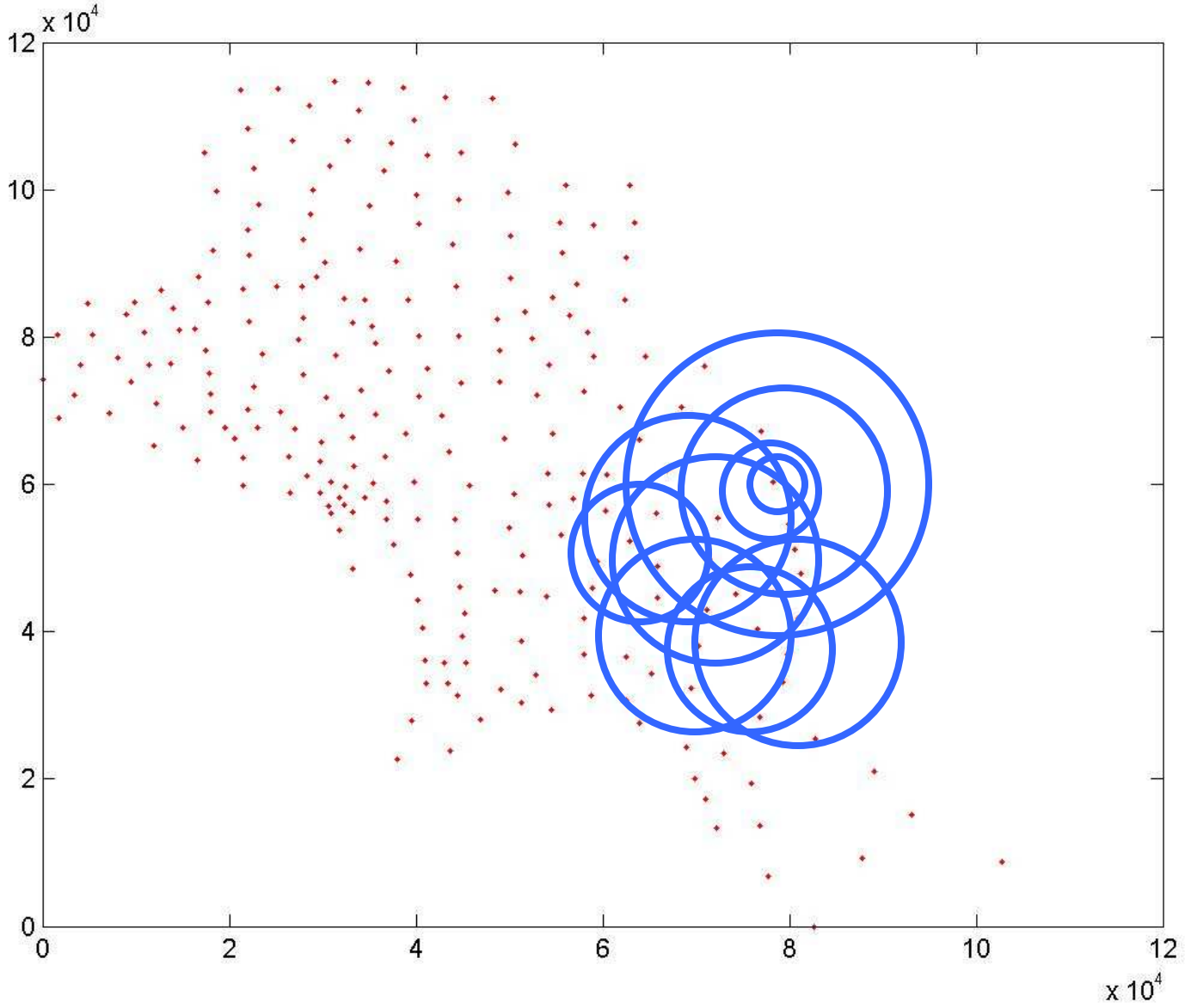
Método de Cuzick e Edwards

- Fixa número de eventos *k*
- Em torno do centróide de cada área *i* onde exista algum caso, faça círculo que vai expandindo até conter *k* eventos *ESPERADOS*
- Verifique quantos eventos *observados* O_i existem dentro desse círculo
- Calcule $$U_k = \sum_{i=1}^{c} O_i - ck$$
- Faça teste de hipótese com esta estatística U_k (Cuzick e Edwards derivaram as fórmulas)

Método de Kulldorff (1997)

- Coloque janela circular de raio fixo na região
- Hipótese: risco dentro e fora da janela são iguais
- Varie posição da janela e tamanho do raio
- Ache aquela que possui menor P-valor
- Calcule P-valor levando em conta a multiplicidade de testes efetuados (esta é a contribuição de Kulldorff)
- P-valor recalculado $< 0{,}05 \Rightarrow$ evidência de clustering e identificação do círculo
- Testes e p-valores associados com clusters secundários não são rigorosamente corretos

O Método Scan de KULLDORFF - Funcionamento



3. O Método SCAN

É fundamentado no método de máxima verossimilhança. O parâmetro é definido por (z,p,r)

z : representa o círculo em Z
p : é a probabilidade de que um indivíduo qualquer dentro de z seja um caso
r : é a probabilidade de que um indivíduo qualquer fora de z seja um caso

$$\hat{p} = \frac{c_z}{n_z}$$

$$\hat{r} = \frac{(C - c_z)}{(M - n_z)}$$

c_z : número observado de casos em z
n_z : número de indivíduos na região z (população em risco)

Função de Verossimilhança – Método SCAN

$$L(z, p, r) = p^{c_z}(1-p)^{(n_z - c_z)} r^{(C-c_z)}(1-r)^{(M-n_z-C+c_z)}$$

Um possível candidato a conglomerado é definido por:

$$L(z, p(z), r(z)) = \sup_{z \in Z,\, p>r} p^{c_z}(1-p)^{(n_z - c_z)} r^{(C-c_z)}(1-r)^{(M-n_z-C+c_z)}$$

É realizada uma varredura sobre todos os candidatos a conglomerados definido em Z, o conglomerado com máxima verossimilhança é a região *z,* para a qual *L*(*z*, *p(z)*, *r(z)*) é maximizada.

$$L(\hat{z}, p(\hat{z}), r(\hat{z})) \geq L(z, p(z), r(z))$$

Ao conglomerado verossímil é atribuída uma estatística do teste da razão de verossimilhança

$$\kappa = \frac{L\left(\hat{z}, p(\hat{z}), r(\hat{z})\right)}{L_o}$$

$$L_o = \frac{C^C \left(M - C\right)^{M-C}}{M^M}$$

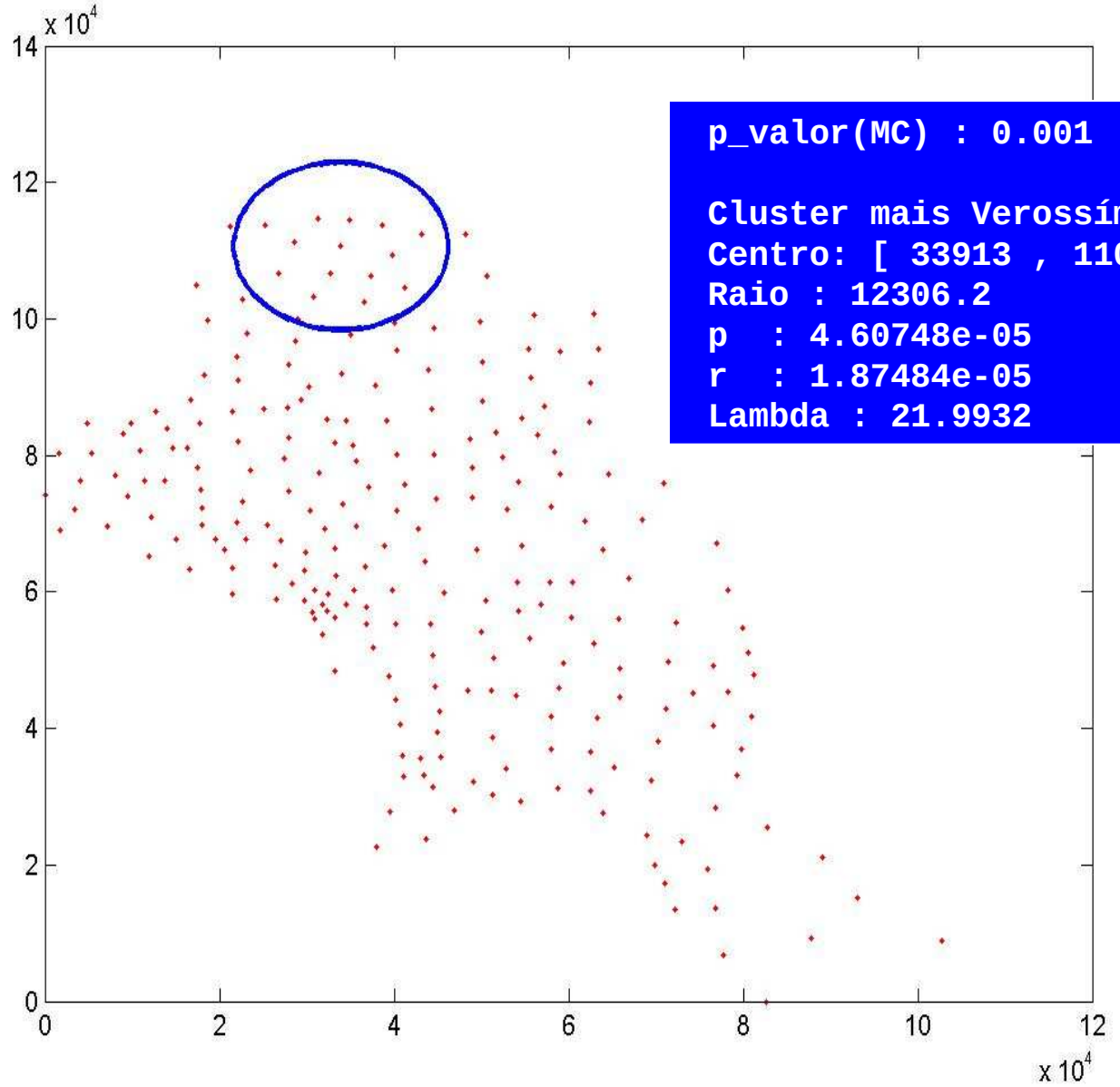
Obtenção da distribuição de κ via Simulação Monte Carlo

1. Gerar S conjuntos independentes, possuindo o mesmo número de casos C (distribuição multinomial proporcional à população de cada área). Para cada conjunto calcula-se a estatística κ. (κ_1, ..., κ_S).

2. Ordene os valores de κ. Comparar o valor original de κ, se o mesmo estiver entre os maiores 100(1-α)% valores, rejeitar H_o ao nível de significância α.

3. Uma vez rejeitada H_o, o conglomerado de máxima verossimilhança é o conglomerado mais verossímil.

O Método SCAN - Funcionamento



p_valor(MC) : 0.001

Cluster mais Verossímill
Centro: [33913 , 110719]
Raio : 12306.2
p : 4.60748e-05
r : 1.87484e-05
Lambda : 21.9932

Teste focados de conglomerados

- Ganha-se poder definindo a alternativa de forma mais precisa.
- Como antes, a hipótese nula é que todas as pessoas possuem o mesmo risco de contrair a doença independentemente dos outros casos e da posição de sua residência. Ou ainda, Oi ~ Poisson($\lambda\, i$), independentes e com $\lambda_i = r * Pop_i$
- A hipótese alternativa agora é que o risco aumenta com a proximidade de locais ***pré-especificados***

De volta ao exemplo

- 11 locais que são possíveis fontes de contaminação
- Comparar círculos obtidos pelos métodos GAM e Besag/Newell com locais suspeitos. Algumas fontes são cobertas, outras não...
- É preciso quantifivcar essa “correlação”
- Além disso, as posições dos locais suspeitos não foi usada nos métodos acima.

Adaptando Besag/Newell e GAM

- Pode-se calcular as estatísticas de teste como antes mas considerando apenas círculos que possuam como centro as fontes suspeitas.
- Simultaneidade pode ser controlada agora pelos métodos usuais (Bonferroni, etc.)
- Por Besag/Newell, apenas dois focos são ligeiramente significativos e com k's diferentes (resultados mais a frente).

Teste de Stone (1988)

- Stone (1988) propôs teste para uma única fonte suspeita.
- Ordene as *n* áreas pela distância à fonte de modo que *i=1* é a mais próxima e *i=n* é a mais distante.
- Estatística de teste é o máximo do risco relativo estimado para uma região em torno da fonte suspeita:

$$T_{Stone} = \max_{1 \le j \le n} \frac{\sum_{i=1}^{j} Casos_i}{\sum_{i=1}^{j} Esp_i}$$

Ainda Stone (1988)

- Suponha que as observações são as contagens

$$Casos_i \sim \text{Poisson}\left(r_i \; E_i\right)$$

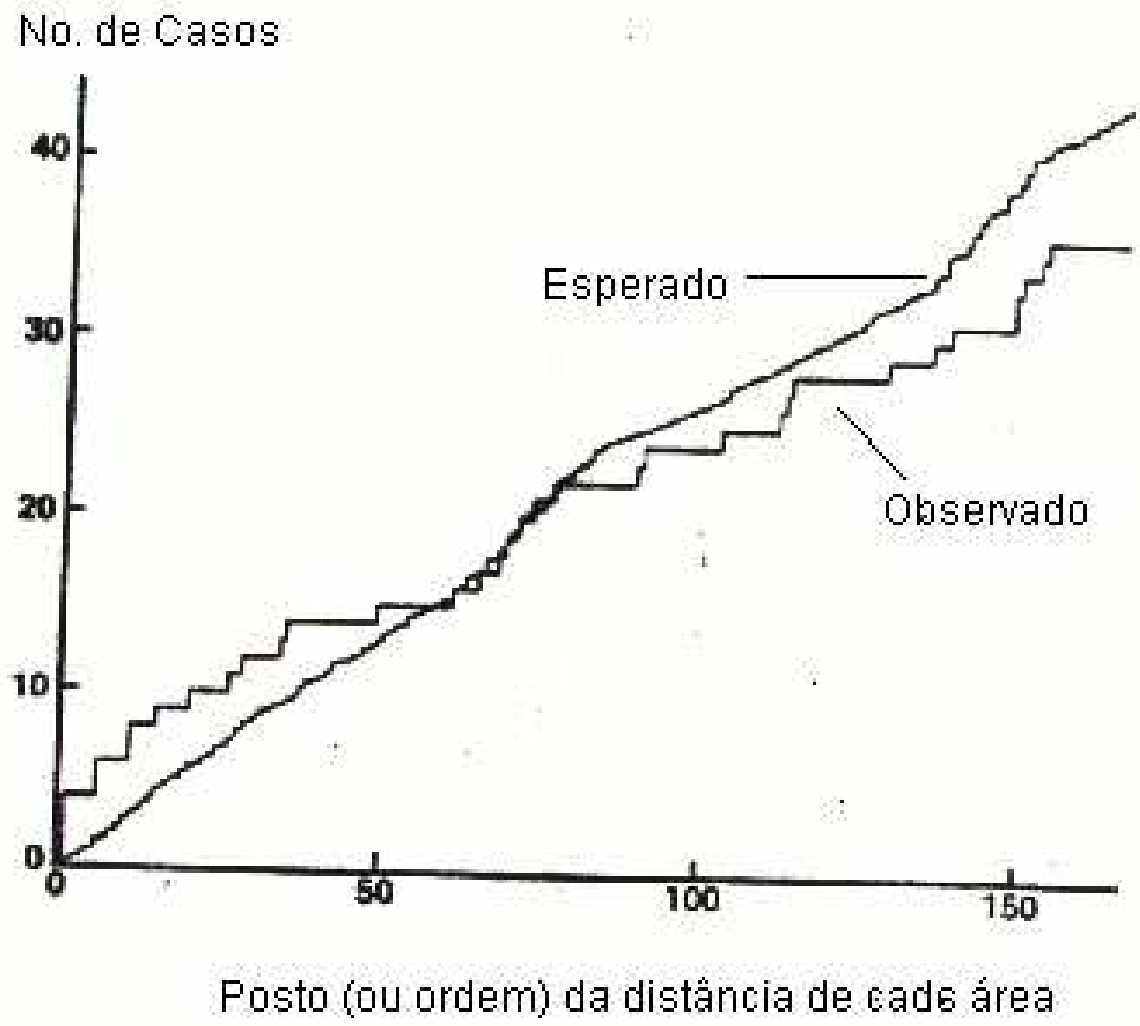
- Nesse teste, a hipótese nula e a alternativa são as seguintes:

$$H_0 : r_1 = r_2 = \ldots = r_n$$
$$H_A : r_1 \geq r_2 \geq \ldots \geq r_n$$

- Stone calculou a distribuição da estatística do teste de razão de máxima verossimilhança considerando um passeio aleatório no plano

Aplicação a Sellafield

- Usina nuclear em Sellafield, Cumbria, England
- 36 casos de leucemia infantil entre 0 e 15 anos entre 1968 e 1982.
- O teste produziu uma região acumulada com 4 casos contra um número esperado de 0,196 e p-valor 0,00005



Número de casos esperados e observados em pequenas áreas à medida que elas são acumuladas pela ordem de distância a Sellafield

Results of Focused Tests for Leukemia in Upstate New York, 1978–1982[a]

Focus	Besag and Newell's (1991) *p*-Value[b] $k = 6$	$k = 8$	$k = 10$	T_{Stone} (*p*-Value[c])	U^* (*p*-Value[d])
Monarch Chemicals	0.069	0.072	0.013	2.50 (0.011)	4.12 (<0.001)
IBM Endicott	0.227	0.056	0.072	2.53 (0.036)	3.39 (<0.001)
Singer	0.665	0.725	0.470	1.48 (0.245)	2.47 (0.007)
Nesco	0.665	0.725	0.470	1.48 (0.245)	2.46 (0.007)
GE Auburn	0.108	0.133	0.143	2.13 (0.156)	2.06 (0.020)
Solvent Savers	0.551	0.254	0.168	1.77 (0.502)	0.29 (0.386)
Smith Corona	0.337	0.258	0.120	2.13 (0.162)	2.56 (0.005)
Victory Plaza	0.013	0.492	0.237	3.37 (0.102)	1.97 (0.024)
Hadco	0.769	0.490	0.284	1.52 (0.572)	1.87 (0.031)
Morse Chain	0.843	0.726	0.879	1.19 (0.784)	−0.78 (0.782)
Groton	0.847	0.875	0.722	1.60 (0.385)	0.81 (0.209)
Combined foci	0.315	0.602	0.334	1.88 (0.207)	3.54 (<0.001)

[a] No adjustments are made for multiple foci in any of the three methods.
[b] Unadjusted *p*-value obtained for each focus following (1).
[c] Analytic *p*-value from (3) and (4).
[d] Asymptotic *p*-value obtained from standard normal distribution (see text).