MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

INPE-8557-PRE/4301

ANÁLISE ESPACIAL DE EVENTOS

Gilberto Câmara
Marilia Sá Carvalho

ANÁLISE ESPACIAL DE DADOS GEOGRÁFICOS□
Instituto Nacional de Pesquisas Espaciais – INPE, São José dos Campos, SP, Brazil.

INPE
São José dos Campos
2002

2 ANÁLISE ESPACIAL DE EVENTOS

Gilberto Câmara
Marilia Sá Carvalho

2.1 INTRODUÇÃO

Neste capítulo serão estudados os fenômenos expressos através de ocorrências identificadas como pontos localizados no espaço, denominados processos pontuais. São exemplos: localização de crimes, ocorrências de doenças, e localização de espécies vegetais. O objetivo destas análises é estudar a distribuição espacial destes pontos, testando hipóteses sobre o padrão observado: se é aleatório, se apresenta-se em aglomerados ou se os pontos estão regularmente distribuídos. O objeto de interesse é a própria localização espacial dos eventos em estudo.

O tipo de dado nestes estudos consiste em uma série de coordenadas de pontos (p_1, p_2, ...) dos eventos de interesse dentro da área de estudo. O termo *evento* refere-se a qualquer tipo de fenômeno localizável no espaço que, dentro de nossa escala de investigação, possa estar associado a uma representação pontual. Exemplos incluem:

- Epidemiologia: residência de casos de doenças
- Sociologia: local de ocorrência de ofensas criminais
- Demografia: localização de cidades
- Biologia: localização de espécies vegetais de interesse

Para ilustrar estes conceitos, considere a figura 2.1, que apresenta a distribuição de 299 óbitos de menores de um ano, registrados no ano de 1998, de crianças nascidas no mesmo ano na cidade de Porto Alegre, Rio Grande do Sul, divididos em neonatais (menores de 28 dias de nascidos) e posneonatais (entre 28 dias e um ano). A análise de padrões neste tipo de dado pode ser utilizada como uma forma de identificação de possíveis áreas com maior concentração de mortes infantis, de comparação entre os óbitos nos dois grupos de idade, e de identificação de fatores de risco associados a esta ocorrência.

Os dados de distribuições pontuais têm as seguintes características:

- A área dos eventos não é uma medida válida apesar de em muitos casos ocuparem espaço. Mesmo na análise do padrão de distribuição de cidades estas são consideradas como um ponto no espaço do estudo.

- Os pontos em geral não estão associados a valores, mas apenas à ocorrência dos eventos considerados.
- Em alguns estudos os pontos podem estar associados a atributos de identificação, como no exemplo acima, em óbitos neonatais e posneonatais. Quando este atributo é elemento do estudo, através da comparação da distribuição espacial destes atributos, denomina-se processo pontual marcado.

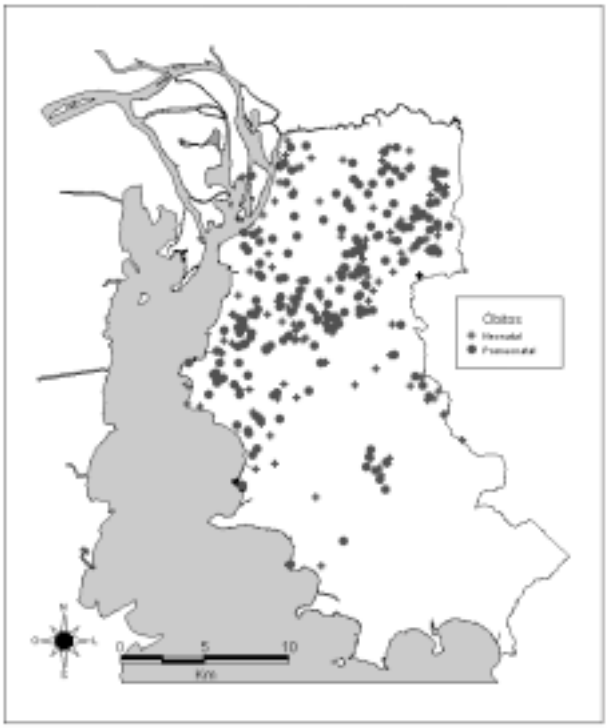

Figura 2-1 - Distribuição espacial de mortalidade infantil – neonatal e posneonatal - em Porto Alegre em 1998.

Nosso interesse primário ao analisar padrões de distribuição de pontos é determinar se os eventos observados exibem algum padrão sistemático, em oposição à uma distribuição aleatória. Busca-se detectar a existência de padrão de conglomerados espaciais (*cluster*), através da constatação de um número acima do esperado de casos excessivamente próximos, considerando uma distribuição estocástica, usualmente um processo de Poisson. Se um padrão de eventos pontuais apresentar desvios significativos do comportamento esperado para uma distribuição de Poisson, isto indica a existência de uma distribuição espacial diferente da completa aleatoriedade, que merece ser objeto de maior análise.

2.2 CARACTERIZAÇÃO DE DISTRIBUIÇÕES DE PONTOS

Numa visão estatística, processos pontuais são definidos como um conjunto de pontos irregularmente distribuídos em um terreno, cuja localização foi gerada por um mecanismo estocástico. Para sua caracterização, este processo estocástico pode ser descrito em termos dos *efeitos de primeira ordem* e *efeitos de segunda ordem*.

Os efeitos de primeira ordem, considerados globais ou de larga escala, correspondem a variações no valor médio do processo no espaço. Neste caso, estamos interessados na *intensidade* do processo, isto é, no número de eventos por unidade de área. Efeitos de segunda ordem, denominados locais ou de pequena escala, representam a *dependência espacial* no processo, proveniente da estrutura de correlação espacial. Para medir a dependência espacial, procuramos estimar o relacionamento entre pares de eventos (por unidade de área) no espaço, o que corresponde a uma aproximação do cálculo da covariância entre as variáveis aleatórias que representam cada evento[1].

Considera-se um conjunto de pontos $(u_1, u_2, ...)$ numa determinada região *A* onde ocorreram eventos. O processo pontual é modelado considerando subregiões *S* em *A* através de sua esperança $E[N(S)]$ e a covariância $C[N(S_i), N(S_j)]$, onde $N(S)$ denota o número de eventos em *S*. Sendo o objetivo da análise estimar as localizações prováveis de ocorrência de determinados eventos, essas estatísticas devem ser inferidas considerando o valor limite da quantidade de eventos por área. Este valor limite corresponde à esperança de $N(S)$ para uma pequena região du em torno do ponto u, quando essa tende a zero. Essa esperança é denominada *intensidade* (propriedade de primeira ordem), sendo definida como

$$\lambda(u) = \lim_{|du| \to 0} \left\{ \frac{E[N(du)]}{|du|} \right\}, \tag{2.1}$$

Propriedades de segunda ordem podem ser definidas da mesma forma, considerando a intensidade conjunta $\lambda(u_i, u_j)$ entre duas regiões infinitesimais $|du_i|$ e $|du_j|$ que contém os pontos u_i e u_j.

$$\lambda(d(u_i), d(u_j)) = \lim_{du_i, du_j \to 0} \left\{ \frac{C[N(du_i), N(du_j)]}{du_i, du_j} \right\} \tag{2.2}$$

Quando o processo é *estacionário*, $\lambda(u)$ é uma constante, ou $\lambda(u) = \lambda$; se também é *isotrópico*, $\lambda(u_i, u_j)$ se reduz à $\lambda(|h|)$, sendo $|h|$ a distância entre os dois pontos. Quando o processo é não estacionário, ou seja, a intensidade média varia

[1] Vale relembrar a discussão do seção 1, onde caracterizamos os eventos no espaço por um processo estocástico, onde cada ocorrência é uma realização de uma variável aleatória distinta.

na região *A*, a modelagem da estrutura de dependência $\lambda(u_i, u_j)$ deve incorporar a variação de $\lambda(u)$. A maior parte das técnicas de análise de distribuição de pontos supõe, explícita ou implicitamente, um comportamento estacionário e isotrópico do processo aleatório subjacente aos eventos analisados.

No exemplo acima da mortalidade infantil, a ocorrência dos óbitos está condicionada pela distribuição dos nascimentos. Além disso, características individuais da criança, tais como prematuridade e peso, são importantes condicionantes do óbito. É possível, entretanto, modelar estes eventos e detectar áreas de sobre-risco, considerando simultaneamente o padrão de distribuição dos nascimentos e óbitos, e verificando a variação da intensidade do evento na região e a estrutura de correlação local.

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. O modelo teórico mais simples (e bastante aplicado na prática) é conhecido como *aleatoriedade espacial completa* ("complete spatial randomness - CSR"). Este modelo divide a região de estudo *A* em subáreas S_i e modela a distribuição de eventos pontuais como um processo aleatório

$$\{Z_i(u_i),\ u_i\ \varepsilon\ S_i : i = 1,...,n\} \tag{2.3}$$

Neste caso, consideramos $Z_i(u_i)$ como o número de eventos que ocorrem na sub-área S_i. No modelo CSR, consideramos que as ocorrências em cada sub-área são não-correlacionadas e homogêneas, e estão associadas à mesma distribuição de probabilidade de Poisson. Numa visão intuitiva, pode-se considerar que a posição dos eventos é independente e de que os eventos tem igual probabilidade de ocorrência em toda a região *A*.

Esta formulação nos permite estabelecer uma base de comparação entre uma distribuição completamente aleatória (que seria gerada por um processo de Poisson) e os dados coletados em campo. O procedimento mais usual para estimar a probabilidade associada ao padrão encontrado será produzir uma *simulação* do processo aleatório na região de estudo. Dado um número fixo de eventos medidos em campo (denotado por *n*), determinamos o retângulo envolvente da região *A* (seja $\{(x,y) : x_1 \le x \le x_2,\ y_1 \le y \le y_2\}$). Os eventos são gerados a partir de abscissas *x*, obtidas de uma distribuição uniforme em (x_1,x_2) e de ordenadas *y*, obtidas de uma distribuição uniforme em (y_1,y_2). Pontos que caem fora da região são rejeitados. Este processo é repetido até que *n* eventos tenham sido obtidos na região.

Podemos gerar um conjunto de simulações, para que possamos obter uma base de comparação entre o comportamento de um processo aleatório e a distribuição dos eventos medidos. Os conceitos de CSR são utilizados para

caracterizar os efeitos de segunda ordem em distribuição de pontos, utilizando os métodos do *vizinho mais próximo* e da função K, descritos a seguir. São também utilizados para avaliação em vários métodos de detecção de aglomerados (*clusters).*

2.3 ESTIMADOR DE INTENSIDADE ("KERNEL ESTIMATION")

Uma alternativa simples para analisar o comportamento de padrões de pontos é a estimar a intensidade pontual do processo em toda a região de estudo. Para isto, pode-se ajustar uma função bi-dimensional sobre os eventos considerados, compondo uma superfície cujo valor será proporcional à intensidade de amostras por unidade de área. Esta função realiza uma contagem de todos os pontos dentro de uma região de influência, ponderando-os pela distância de cada um à localização de interesse, como mostrado na Figura 2-2.

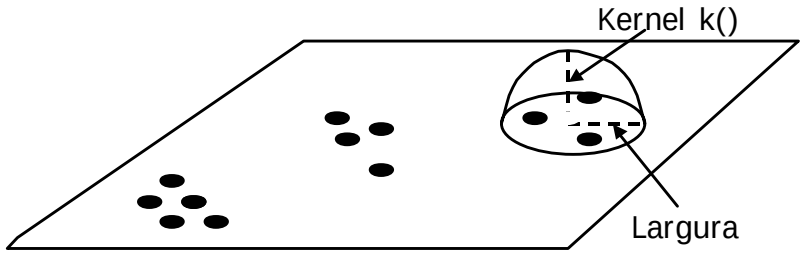

Figura 2-2 - Estimador de intensidade de distribuição de pontos.

A partir dos conceitos apresentados, suponha e $u_1,...,u_n$ são localizações de n eventos observados em uma região A e que u represente uma localização genérica cujo valor queremos estimar. O estimador de intensidade é computado a partir dos m eventos $\{u_i,...u_{i+m-1}\}$ contidos num raio de tamanho τ em torno de u e da distância d entre a posição e a i-ésima amostra, a partir de funções cuja forma geral é:

$$\hat{\lambda}_\tau(u) = \frac{1}{\tau^2}\sum_{i=1}^{n} k(\frac{d(u_i,u)}{\tau}) \text{ , } d\ (u_i,u) \le \tau \qquad (2.4)$$

Este estimador é chamado *kernel estimator* e seus parâmetros básicos são: (a) um raio de influência ($\tau \ge 0$) que define a vizinhança do ponto a ser interpolado e controla o "alisamento" da superfície gerada; (b) uma função de estimação com propriedades de suavização do fenômeno. O *raio de influência* define a área centrada no ponto de estimação u que indica quantos eventos u_i contribuem para a estimativa da função intensidade λ. Um raio muito pequeno irá gerar uma superfície muito descontínua; se for grande demais, a superfície poderá ficar muito amaciada. No caso da função de interpolação *k()*, é comum usar funções de terceira ou quarta ordem, como

$$k(h) = \frac{3}{\pi}(1-h^2) \tag{2.5}$$

ou o *kernel gaussiano*

$$k(h) = \frac{1}{2\pi\tau}\exp\left(-\frac{h^2}{2\tau^2}\right) \tag{2.6}$$

Nestes estimadores, h representa a distância entre a localização em que desejamos calcular a função e o evento observado. Com o uso desta função de quarta ordem (equação 2.5), o estimador de intensidade pode ser expresso como:

$$\hat{\lambda}_\tau(u) = \sum_{h_i \le \tau} \frac{3}{\pi\tau^2}\left(1-\frac{h_i^2}{\tau^2}\right)^2 \tag{2.7}$$

O estimador de intensidade é muito útil para nos fornecer uma visão geral da distribuição de primeira ordem dos eventos. Trata-se de um indicador de fácil uso e interpretação. A figura 2.3 ilustra a aplicação do estimador de intensidade para o caso de mortalidade por causas externas em Porto Alegre, com os dados de 1996. A localização dos homicídios (vermelho), acidentes de trânsito (amarelo) e suicídios (azul) esta mostrada na figura 2.3 à esquerda e o estimador de intensidade dos homicídios é apresentado na figura 2.3. A superfície interpolada mostra um padrão de distribuição de pontos com uma forte concentração no centro da cidade e decrescendo em direção aos bairros mais afastados.

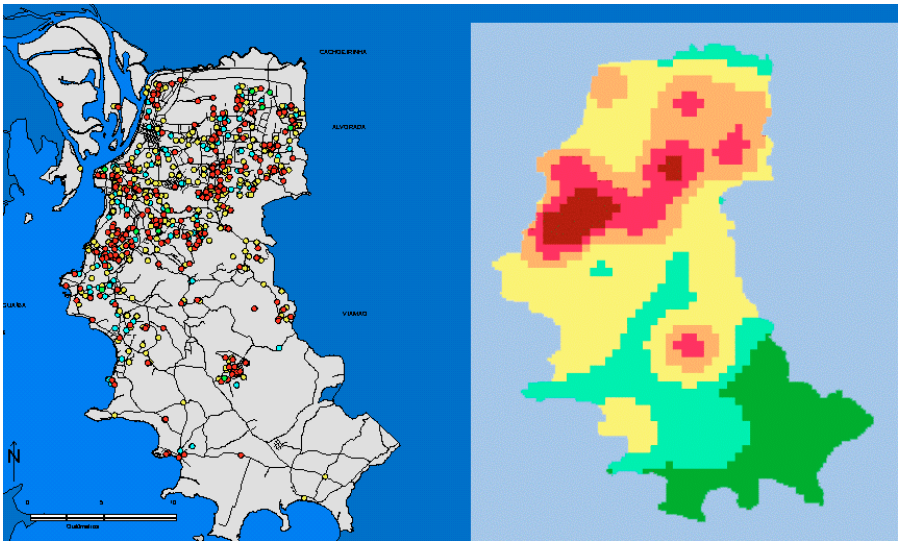

Figura 2.3: Distribuição de casos de mortalidade por causas externas em Porto Alegre em 1996 e estimador de intensidade.

2.4 ESTIMADORES DE DEPENDÊNCIA ESPACIAL

Para a estimação de propriedades de segunda ordem do processo pontual, as técnicas mais utilizadas são o *vizinho mais próximo* e a *função K*, descritos a seguir.

Método do Vizinho Mais Próximo

O método do vizinho mais próximo estima a função de distribuição cumulativa $\hat{G}(h)$ baseado nas distâncias *h* entre eventos em uma região de análise. Esta função de distribuição pode ser estimada empiricamente da seguinte forma:

$$\hat{G}(h) = \frac{\#(d(u_i, u_j) \leq h)}{n} \tag{2.8}$$

onde o valor normalizado acumulado para uma distância *h* corresponde à soma dos vizinhos mais próximos de cada evento cuja distância é menor ou igual a *h*, dividido pelo número de eventos na região.

A plotagem dos resultados desta função de distribuição cumulativa empírica $\hat{G}(h)$ pode ser usada como um método exploratório para se verificar se existe evidência de interação entre os eventos. Se esta plotagem apresentar um crescimento rápido para pequenos valores de distância, esta situação aponta para interação entre os eventos caracterizando agrupamentos nestas escalas. Se esta plotagem apresentar valores pequenos no seu início, e só crescer rapidamente para valores maiores de distância, esta situação aponta para uma distribuição mais regular. A Figura 2-4 mostra a função $\hat{G}(h)$ para os dados de mortalidade infantil de Porto Alegre (figura 2.1), com distância mínima de 0 km e distância máxima de 1 km. Verifica-se que a curva mostra um crescimento acentuado para distâncias até 500 m para depois se estabilizar, o que caracteriza agrupamento nesta faixa de distâncias.

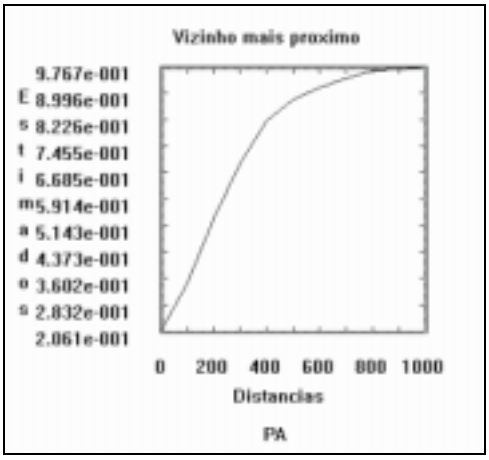

Figura 2-4 – Função vizinho-mais-próximo para mortalidade infantil neonatal em Porto Alegre.

A análise de vizinhança pode ser usada como método formal para se comparar estatísticamente a distribuição dos eventos observados com o que se esperaria na hipótese da aleatoriedade espacial completa (CSR). Esta metodologia consiste em se criar envelopes de simulação para a distribuição CSR, a fim de se acessar a significância dos desvios. Na hipótese de CSR, a função de distribuição G(w) seria dada por um processo de Poisson

$$G(h) = 1 - e^{-\lambda \pi h^2} \quad h \geq 0 \tag{2.9}$$

A estimação simulada para a distribuição *G(w)* assumindo-se CSR é calculada como

$$\overline{G}(h) = \frac{\sum_{i}^{k} \hat{G}_i(h)}{k} \tag{2.10}$$

onde $\hat{G}_i(h)$, $i = 1,2..,k$ são funções de distribuição empíricas, estimadas a partir de *k* simulações independentes dos *n* eventos, na hipótese de CSR (*n* eventos independentes e uniformente distribuídos). Para verificar a condição de aleatoriedade, calculamos ainda os envelopes de simulação superior e inferior, definidos como se segue:

$$\begin{aligned} U(h) &= \max\{\hat{G}_i(h)\}, \quad i = 1,...,k \\ L(h) &= \min\{\hat{G}_i(h)\}, \quad i = 1,...,k \end{aligned} \tag{2.11}$$

A plotagem da distribuição estimada $\hat{G}(h)$ versus a distribuição simulada $\overline{G}(h)$, com a adição dos envelopes inferior e superior, permite medir a significância dos desvios relativo a aleatoriedade. Se a condição CSR for válida para os dados observados, o gráfico da curva de $\hat{G}(h)$ versus $\overline{G}(h)$ deve ser praticamente linear com um ângulo de 45 graus. Se o dado apresenta tendências para agrupamentos, os traçados no gráfico estarão acima da linha de 45 graus, ao passo que para padrões de regularidade os traçados ficarão abaixo da linha de 45 graus.

A Figura 2-5 mostra um exemplo de gráfico mostrando o posicionamento da distribuição e dos envelopes com relação a linha de 45 graus, para os dados referentes mortalidade infantil neonatal em Porto Alegre. Neste caso percebe-se a posição dos envelopes e da distribuição acima da linha de 45 graus, o que caracteriza agrupamento para as distâncias em análise.

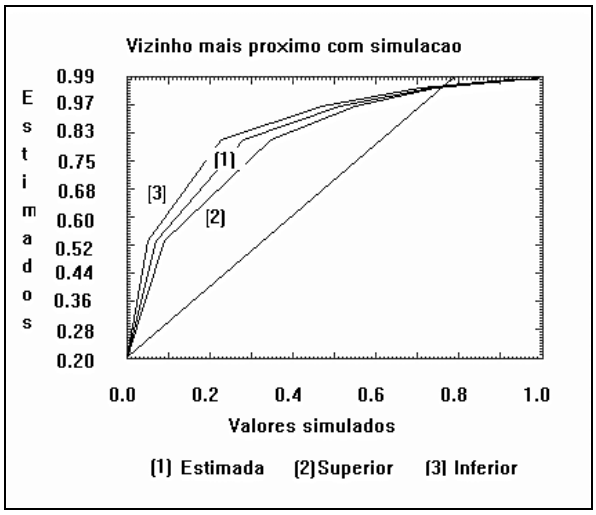

Figura 2-5 – Gráfico de $\hat{G}(h)$ (estimado) versus $\overline{G}(h)$ (CSR), com envelopes superior e inferior, para os dados de mortalidade neonatal em Porto Alegre

Embora o método do vizinho mais próximo forneça uma indicação inicial da distribuição espacial, ele considera apenas escalas pequenas. Para se ter informação mais efetiva para o padrão espacial em escalas maiores, o melhor método a ser utilizado é o da função K.

Função K

A função K, também denominada medida de *momento de segunda ordem reduzido*, é definida para o processo univariado como:

$\lambda K(h) = E(\#$ *eventos contidos a uma distância h de um evento arbitrário)* (2.12)

onde # está associado ao número de eventos, E() é o operador de estimativa, e λ é a intensidade ou número médio de eventos por unidade de área, assumida constante na região. Uma estimativa de $K(h)$ é:

$$\hat{K}(h) = \frac{A}{n^2} \sum_{i}^{n} \sum_{j,i \neq j}^{n} \frac{I_h(d_{ij})}{w_{ij}} \qquad (2.13)$$

onde A é a área da região, n é o número de eventos observados, $I_h(d_{ij})$ é uma função indicatriz cujo valor é 1 se $(d_{ij}) <= h$ e 0 em caso contrário, e w_{ij} é a proporcão da circunferência do círculo centrado no evento i que está dentro da região (correção devido ao efeito de borda).

A função K é usada como ferramenta exploratória na comparação entre estimativa empírica — $\hat{K}(h)$ — e a resultante de um processo de padrão de pontos espacial aleatório — $\overline{K}(h)$. Para um processo aleatório $K(h)$ seria πh^2. Portanto, uma forma de comparar a estimativa K de um conjunto de dados observados com πh^2 seria plotar a função $\hat{L}(h)$ definida como:

$$\hat{L}(h) = \sqrt{\frac{\hat{K}(h)}{\pi}} - h \qquad (2.14)$$

O gráfico de $\hat{L}(h)$ em função da distância *h* indica atração espacial entre eventos ou agrupamentos para valores positivos, sendo o agrupamento mais forte em picos positivos, e indica repulsão espacial ou regularidade em pontos de valores negativos. Uma abordagem similar à do vizinho mais próximo pode ser feita para se estimar a significância dos desvios da distribuição $\hat{L}(h)$ em relação à condição de aleatoriedade (CSR). Os envelopes inferior e superior são construídos a partir de *k* simulações independentes de *n* eventos na região *A*. Na análise do gráfico com a distribuição e os envelopes, picos positivos na função estimada $\hat{L}(h)$ que estão acima do envelope superior evidenciam ocorrência de agrupamento na escala considerada, portanto, se todos os valores da função $\hat{L}(h)$ estiverem acima do envelope superior e com valores positivos, teremos agrupamentos em todas as escalas. Depressões negativas na função estimada $\hat{L}(h)$ que estiverem abaixo do envelope inferior, evidenciam regularidade nessa escala, portanto, se todos os valores de $\hat{L}(h)$ estiverem abaixo do envelope inferior e com valores negativos, tem-se regularidade em todas as escalas.

A Figura 2-6 mostra o gráfico da função $\hat{L}(h)$ e dos envelopes de simulação para o dado de Porto Alegre (Figura 2-1). Verifica-se valores positivos para a função *L*, estando os mesmos acima dos envelopes, o que caracteriza agrupamento em todas as escalas de distância.

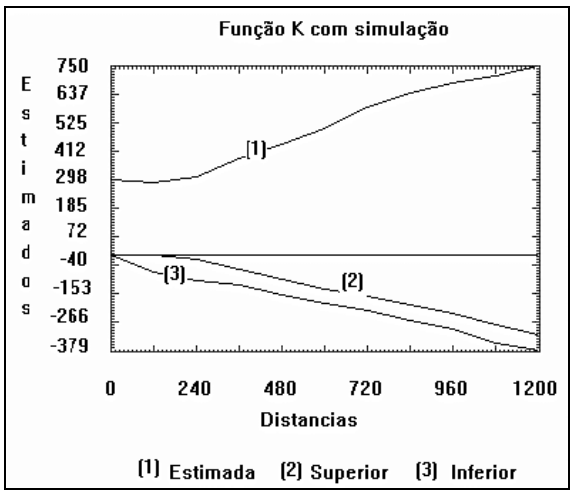

Figura 2-6 – Função K com simulação para os dados de mortalidade neonatal em Porto Alegre.

2.5 PROCESSO PONTUAL MARCADO

Um das situações mais importantes na análise espacial de pontos é a possibilidade de comparação entre dois processos espaciais. Tipicamente, um dos processos representa os casos em estudo, e o outro os casos de um processo pontual que representa um conjunto de casos de controle. Por exemplo, num estudo realizado por Peter Diggle na Inglaterra sobre câncer de laringe, foi utilizado dados de câncer de pulmão como indicadores da distribuição espacial da população. Esta situação pode ser generalizada supondo dois processos pontuais, o primeiro cujos casos localizam-se em $(u_1,u_2,...,u_1)$ e o segundo cujos casos estão nos pontos $(u_{n+1},u_{n+2},...,u_{n+m})$. Cada tipo de evento pode ser modelado como uma distribuição de Poisson, I e II, com intensidades $\lambda_1(u)$ e $\lambda_2(u)$. Define-se o risco na localidade u como uma função da razão entre λ_1 e λ_2. O objetivo da análise é investigar a variação espacial desta razão na região.

Se estimarmos a intensidade de cada processo através de uma função *kernel*, a razão entre as duas funções será a intensidade do risco. E cada uma das funções estudadas anteriormente pode ser adaptada para verificar a relação entre os pontos do processo I com os pontos do processo II. Por exemplo, visando estudar a dispersão de duas espécies vegetais pode-se verificar a relação de cada ponto com o vizinho mais próximo da outra espécie.

2.6 ESTUDOS CASO-CONTROLE

Considere-se um tipo de estudo onde temos dois tipos de eventos, por exemplo recém-natos que morrem antes de completar um ano e os que sobrevivem a esta idade. Sendo esta variável do tipo binomial a resposta do estudo, dependente de diversas covariáveis tais como prematuridade, existência de doenças na gravidez, escolaridade da mãe, e incluindo sua localização no espaço, pode-se modelar o processo utilizando o método clássico de regressão logística, próprio para este tipo de distribuição. O que particulariza o contexto espacial é a forma de se incluir a localização dos pontos no modelo. Diversas formas de estimar este risco em cada localidade são possíveis, entre as quais utilizar o mesmo kernel da razão como um dos termos da regressão, que toma uma forma semi-paramétrica abaixo:

$$\text{logit}(y_i) = \beta \mathbf{x}_i + g(s_i), \qquad (2.15)$$

onde:

- y_i é a variável resposta, e tem a forma sim/não, zero/um (óbitos/nascimentos),
- a função de ligação da regressão é o logit, como usual para dados binomiais,
- $\boldsymbol{x}_i$ é o vetor de covariáveis,

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

O ganho deste tipo de modelagem é possibilitar a estimativa da variação espacial do risco, controlando pelos fatores conhecidos de variação de risco. Os procedimentos de estimação dos parâmetros destes modelos baseia-se em métodos iterativos usuais de modelos aditivos generalizados, onde se estima os parâmetros da regressão, e sobre os resíduos estima-se a função *kernel*, e assim sucessivamente até que as estimativas não mais se alterem. O método permite identificar áreas de sobre ou sub risco significativamente diferente da média global. A largura de banda a ser utilizada é importante, e pode ser definida através de métodos automáticos ou selecionada pelo pesquisdor visando ajustar a uma conhecida estrutura espacial. No estudo da mortalidade infantil em Porto Alegre (figura 2-1) os dados foram analisados segundo esta proposta, incluindo como fatores de risco individuais: (a) peso ao nascer, (b) semanas gestacionais, (c) sexo da criança, (d) (e) idade da mãe, (f) grau de instrução da mãe, (g) tipo de gravidez e (h) tipo de parto, numa regressão logística cuja expressão é:

$$\log\left\{\frac{p(s,\boldsymbol{x})}{1-p(s,\boldsymbol{x})}\right\} = \beta_0 + \beta_1\, sexo + \beta_2\, peso + \beta_3\, idade + \beta_4\, inst + \beta_5\, ges + \beta_6\, grav + \beta_7\, parto + g(s). \quad (2.16)$$

A interpretação dos resultados é razoavelmente direta: os parâmetros β indicam a razão de chances estimada pelo modelo (Quadro 2-1), da forma usual da regressão logística, e no mapa são apresentadas as áreas onde a probabilidade de obter o valor do *kernel* estimado está "significativamente" diferente da intensidade média do processo. O algoritmo para estimar a largura de banda ótima para os dados utiliza validação cruzada de mínimos ponderados para o passo de regressão não-paramétrica. No passo de suavizamento (Eq. 2.15) escolhe-se o valor de *h* que minimiza:

$$CV(h) = \frac{\sum_{i=1}^{n} w_i \left\{z_i - \hat{g}^{-1}(s_i)\right\}^2}{n}, \quad (2.17)$$

onde $\hat{g}^{-1}(s_i)$ é a estimativa de $g(s_i)$ construída com o valor de banda *h* usando todos os dados com exceção do par (s_i, z_i). Testa-se diferentes valores de h, sendo escolhido o que minimiza o somatório.

Quadro 1: Estimativas dos efeitos de covariáveis utilizando o valor da banda obtido por validação cruzada			
Fator	**Estimativa**	**Erro padrão**	**P-valor**
Intercepto	4,0717	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	1,1685	0,3737	**0,0018**
Tipo de gravidez	-0,2006	0,6558	0,7598
Tipo de parto	-0,5320	0,2838	0,0613

A figura 2-7 mostra os mapas de risco para a mortalidade infantil após, incluídas as co-variáveis individuais da criança e da mãe. É interessante observar que no centro da cidade de Porto Alegre existe uma região onde o risco da mortalidade é significativamente menor e outra onde é maior. Quanto às variáveis individuais, somente foram significativas o peso ao nascer, que é reconhecidamente a variável mais associada à mortalidade neo-natal, e a duração da gestação, indicativo de prematuridade. Além de mapeamento do risco, é importante avaliar se a superfície estimada varia significativamente na região, ou seja, se existem evidências estatísticas suficientes para rejeitar a hipótese nula de risco constante na região, tendo-se controlado os fatores individuais de risco. Em termos do modelo, isso equivale ao teste da hipótese H_0: $g(s) = 0$. Também é de interesse a construção de contornos de tolerância que auxiliam na identificação de áreas onde o risco é significativamente superior (ou inferior) à média global. Ou seja, reconhecendo o papel de um dado fator como um preditor importante da mortalidade infantil e controlando-o, deseja-se identificar áreas onde o risco é significativamente mais elevado, buscando orientar a intervenção.

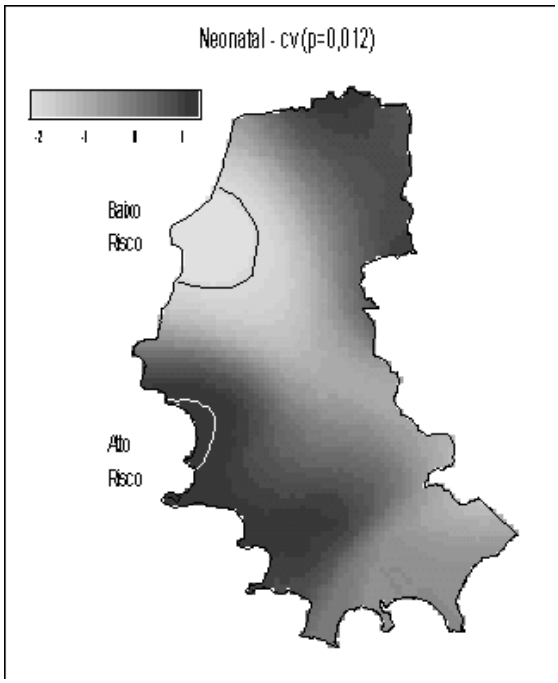

Figura 2-7. Mapas de risco para a mortalidade infantil, controlando para fatores individuais, com a largura de banda estimada por validação cruzada, Porto Alegre, 1998

O teste global do risco e a identificação de áreas de baixo e alto risco podem ser feitos utilizando o método de simulação Monte Carlo, seguindo os passos do algoritmo abaixo:

1. Ajustando-se um modelo de regressão logística convencional, para cada evento – caso ou controle – estima-se a probabilidade ajustada $\hat{p}_i$. Ou seja, dadas as covariáveis daquele registro, qual é a probabilidade ser um caso.
2. Fixando-se as localizações de cada ponto, amostra-se *m* dos *n* indivíduos (sem reposição) com probabilidade proporcional a $\hat{p}_i$ e estes são rotulados como casos e os *n-m* restantes como controles.
3. Calcula-se uma nova estimativa de *g(s)*, $\hat{g}_1(s)$, a estimativa centralizada em torno da média $\tilde{g}_1(s) = \hat{g}_1(s) - \bar{g}_1$, onde $\bar{g}_1 = n^{-1}\sum_{i=1}^{n}\hat{g}_1(s_i)$ e a estatística $t_1 = n^{-1}\sum_{i=1}^{n}\left(\tilde{g}_1(s_i)\right)^2$.
4. Repete-se os passos 1 e 2 *m* vezes.
5. Constrói-se uma superfície de p-valores que para cada *s* fornece a proporção dos valores de $\tilde{g}_j(s)$, $j=1,\ldots,m$, menores do que a estimativa original, digamos $\tilde{g}_0(s)$.

6. Adiciona-se os contornos de 0.05 e 0.95 da superfície de p-valores ao mapa de $\tilde{g}_0(s)$ como contornos de 90% de confiança para indicar áreas de alto/baixo risco.

7. Para o teste de hipótese, define-se k o número de $t_j > t_0$ (obtida a partir de $\tilde{g}_0(s)$) e o nível de significância correspondente por $p = (k+1)/(m+1)$.

2.7 REFERÊNCIAS

A referência das técnicas mais básicas apresentadas neste capítulo é o livro de Trevor Bailey, "*Spatial Data Analysis by Example*" (Bailey and Gattrel, 1995). As técnicas de caso-controle espacial foram desenvolvidas por Peter Diggle e colaboradores, e a maior parte das rotinas e algoritmos está disponível na página da do Departamento de Matemática e Estatística da Universidade de Lancaster (http://www.maths.lancs.ac.uk). O relatório técnico "An S+ library on risk estimation and cluster detection in case-control studies", de Jarner, M. F. and Diggle, P. J., mostra as funções desenvolvidas e como usá-las. Está disponível em http://www.maths.lancs.ac.uk/dept/stats/techabstracts02.html.

Os modelos aditivos generalizados, que servem de base para a extensão espacial podem ser melhor estudados em HASTIE, T. J.; TIBSHIRANI, R. J., 1990, *Generalized Additive Models*. London:Chapman and Hall. Um excelente livro para estudar modelos de regressão é o HOSMER, D. W.; LEMESHOW, S., 1989, *Applied Logistic Regression*. New York:Wiley.

Os trabalho sobre mortalidade infantil em Porto Alegre foi publicado no número especial dos Cadernos de Saúde Pública sobre o tema de estatísticas espaciais em saúde (volume 17(5), outubro-novembro 2001, 1251-1261), disponível na Internet (www.scielo.br).

1. DIGGLE, P. J., 1992. **Point process modelling in environmental epidemiology. Relatório Técnico** MA92/70, Lancaster: Department of Mathematics and Statistics, Lancaster University.

2. KELSALL, J. E.; DIGGLE, P. J. , 1995b. Non-parametric estimation of spatial variation in relative risk. **Statistics in Medicine**, 14:2335-2342.

3. KELSALL, J. E.; DIGGLE, P. J., 1998. Spatial variation in risk of disease: a nonparametric binary regression approach. **Applied Statistics**, 47:559-573.