MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

INPE-8560-PRE/4304

ANÁLISE ESPACIAL DE ÁREAS

Gilberto Câmara
Marilia Sá Carvalho
Oswaldo Gonçalves Cruz
Virginia Correa

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

INPE
São José dos Campos
2002

5 ANÁLISE ESPACIAL DE ÁREAS

Gilberto Câmara
Marilia Sá Carvalho
Oswaldo Gonçalves Cruz
Virginia Correa

5.1 INTRODUÇÃO

Este capítulo discute métodos de análise de dados espaciais cuja localização está associada a áreas delimitadas por polígonos. Este caso ocorre com muita freqüência quando lidamos com eventos agregados por municípios, bairros ou setores censitários, onde não se dispõe da localização exata dos eventos, mas de um valor por área. Alguns desses indicadores são contagens, como é o caso da maior parte das variáveis coletadas no censo: por exemplo, o IBGE fornece, para cada setor censitário, o número de chefes de família em cada uma das faixas de renda consideradas. Diversos indicadores de saúde também são deste tipo: o Ministério e Secretarias de Saúde organizam e disponibilizam dados de óbitos, partos, doenças transmissíveis por município. Utilizando duas contagens – óbitos e população, por ex. – taxas de densidade de ocorrência, como taxas de mortalidade ou incidência são estimados. Outros indicadores bastante úteis são: (a) proporções, como percentual de adultos analfabetos; (b) médias, como renda média do chefe da família por setor censitário; e (c) medianas, como mediana etária em homens.

A forma usual de apresentação de dados agregados por áreas é o uso de mapas coloridos com o padrão espacial do fenômeno. A Figura 5-1 mostra a distribuição espacial do índice de exclusão social[1] para os 96 distritos da cidade de São Paulo, para os dados do censo de 1991. Verifica-se que 2/3 dos 96 distritos de São Paulo estavam abaixo dos índices mínimos de inclusão social em 1991. Uma forte polarização centro-periferia é claramente perceptível no mapa, que apresenta duas grandes regiões de exclusão social, as zonas Sul e Leste da cidade. Na zona Leste, nota-se um gradiente do índice de exclusão/inclusão social, que piora à medida que nos afastamos do centro. Na zona Sul, a descontinuidade do índice é mais

[1] O índice de exclusão/inclusão social é uma medida agregada das disparidades socioeconômicas, que varia de –1 a +1, onde o valor 0 (zero) indica o um nível básico de inclusão social.

abrupta, e verificamos a existência de distritos com altos índices de exclusão/inclusão social próximos a áreas excluídas.

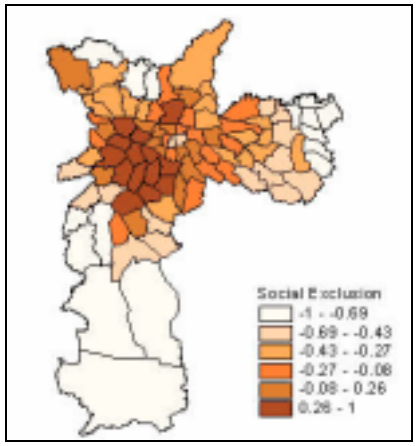


Figura 5-1– Índice de Exclusão/Inclusão Social dos Bairros da Cidade de São Paulo para os dados de 1991, com 96 distritos agrupados por sextis.

Grande parte dos usuários limita seu uso de SIG a essas operações de visualização, tirando conclusões intuitivas. Mas é possível ir muito além. Quando visualizamos um padrão espacial, é muito útil traduzi-lo em considerações objetivas: o padrão que observamos é aleatório ou apresenta uma agregação definida? Esta distribuição pode ser associada a causas mensuráveis? Os valores observados são suficientes para analisar o fenômeno espacial a ser estudado? Existem agrupamentos de áreas com padrões diferenciados dentro da região de estudo?

Para abordar estas questões, este capítulo apresenta um conjunto de técnicas de análise espacial de dados agregados por áreas. O primeiro passo é escolher o modelo inferencial a ser utilizado. A hipótese mais comum é supor que as áreas são diferenciadas, e que cada uma delas possui uma "identidade" própria. Do ponto de vista estatístico, isto implica em que cada área apresenta uma distribuição de probabilidade distinta das demais, o chamado *modelo espacial discreto*. A alternativa é supor que o fenômeno estudado apresenta continuidade espacial, formando uma superfície, o chamado *modelo espacial contínuo* estudado no capítulo anterior. Neste caso, as áreas são consideradas apenas um suporte para coleta de dados, e o modelo inferencial desconsidera os limites de cada área. A produção de superfícies a partir de dados de área será discutida no final deste capítulo.

A questão de agregação de contagens em áreas levanta ainda problemas conceituais importantes: Pode-se estimar comportamentos individuais a partir de dados agregados? Em que medida a comportamento dos

agregados reflete mais do que a soma dos indivíduos? Qual o erro cometido ao estimar indicadores onde as contagens são muito pequenas? Neste capítulo, após a apresentação dos modelos adequados à análise de dados agregados por áreas serão abordados os conceitos básicos da análise espacial, para dados agregados por área.

5.2 MODELOS DE DISTRIBUIÇÃO DE DADOS EM ÁREAS

O modelo de distribuição mais utilizado para dados de área é *o modelo de variação espacial discreta*. Considere-se a existência de um processo estocástico Z_i, $i=1,...,n$, onde Z_i é a realização do processo espacial na área i e n é o total de áreas A_i. O objetivo principal da análise é construir uma aproximação para a distribuição conjunta de variáveis aleatórias $Z=\{Z_1,...,Z_n\}$, estimando sua distribuição.

De forma semelhante ao modelo de eventos pontuais discutido no capítulo 2, considere-se Z_i como a variável aleatória que descreve a contagem, indicador ou taxa associada à área A_i. Dispomos de um valor observado z_i , correspondente à contagem na *i*-ésima área. A hipótese mais comum é supor que a variável aleatória Z_i, que descreve o número de ocorrências em cada área pode ser associada a uma distribuição de probabilidade de Poisson. Tal hipótese justifica-se por ser esta a distribuição estatística mais adequada a fenômenos que envolvem contagens de eventos, como é o caso na maioria dos dados agregados por áreas. Evidentemente outras distribuições podem ser mais adequadas, dependendo da variável a ser analisada. Taxas podem ser modeladas utilizando a distribuição normal, pois ainda que esta admita valores negativos, evidentemente impossíveis neste tipo de indicador, as propriedades da distribuição normal podem ser adequadas.

A alternativa à hipótese de *variação espacial discreta* é supor que os dados apresentam *variação espacial contínua*. Considera-se um processo estocástico $\{Z(x), x\in A, A\subset\Re^2\}$, cujos valores podem ser conhecidos em todos os pontos da área de estudo. Neste caso, as contagens agregadas devem ser transformadas em taxas ou indicadores, pois o que varia continuamente no espaço são as taxas e não as contagens. A estimação deste processo estocástico pode ser feita como descrito nos capítulos 3 e 4 deste livro. O uso de modelos espaciais contínuos será discutido na seção 5.8.

5.3 PROBLEMAS DE ESCALA E A RELAÇÃO ÁREA-INDIVÍDUO

Um dos problemas básicos com dados agregados por área é que, para uma mesma população estudada, a definição espacial das fronteiras das áreas afeta os resultados obtidos. As estimativas obtidas dentro de um sistema de unidades de área são função das diversas maneiras que estas unidades podem ser agrupadas; pode-se obter resultados diferentes simplesmente alterando as fronteiras destas zonas. Este problema é conhecido como “problema da unidade de área modificável”.

Em muitos dos estudos envolvendo dados de área, o dado agregado é a única fonte disponível, porém o objeto de estudo diz respeito a características e relacionamentos individuais. Alguns destes estudos procuram estabelecer relações de causa-efeito entre diferentes medidas, como o uso de modelos de regressão; um exemplo clássico é correlacionar anos de estudo do chefe de família e sua renda, que usualmente apresenta forte correlação. Note-se, no entanto, que devido aos efeitos de escala e de agregação de áreas, os coeficientes de correlação podem ser inteiramente diferentes no indivíduo e nas áreas. Este fenômeno, nas ciências sociais e na epidemiologia, é chamado de “falácia ecológica”.

Considere um conjunto de indivíduos onde são medidas duas características de cada um dos indivíduos, conforme estimado na Figura 5-2. Uma regressão considerando todos os indivíduos (linha negra do quadro à esquerda) resulta em coeficiente positivo de 0,1469. Esses indivíduos pertencem a grupos distintos, separando cada grupo conforme o atributo cor, obtém-se correlação negativa, variando entre –0,5 e –0,8. Utilizando as médias de cada grupo (linha negra do quadro à direita), o coeficiente vai a 0,99. É importante observar que cada modelo mede um aspecto diferente e que não há modelo correto. No primeiro caso, pode-se dizer que sem informações que permitam separar os indivíduos nos grupos coloridos, as variáveis se relacionam positivamente. No último exemplo, o interesse do estudo é o efeito da variação na média de uma variável sobre a média da outra, nos grupos. São perguntas diferentes, e modelos diferentes.

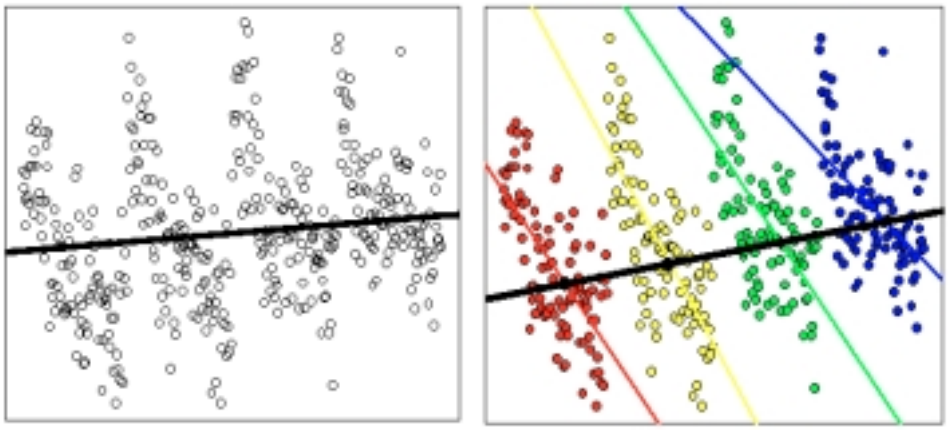


Figura 5-2 – Modelos de regressão: indivíduos, indivíduos em estratos diferentes e grupos.

Para ilustrar o problema das mudanças de unidade de análise, estudou-se os dados de censo de Belo Horizonte para o ano de 1991, em duas escalas: os setores censitários e as unidades de planejamento (UP), mostradas na Figura 5-2. Os setores censitários foram utilizados pelo IBGE para o censo de 1991, e as unidades de planejamento correspondem a agregamentos de áreas utilizados pela prefeitura de Belo Horizonte.

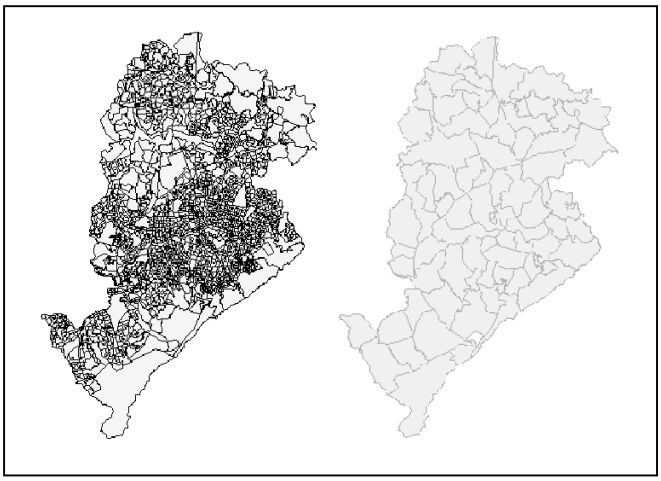


Figura 5-3. Setores censitários (à esquerda) e Unidades de Planejamento (à direita) para o município de Belo Horizonte.

A partir das variáveis do censo, foram computadas 1000 correlações entre pares de variáveis, tanto por setor censitário como por UP. Por exemplo, tomou-se as variáveis "número de chefes de família com rendimento entre 0,5 e 1 salário mínimo" e "número de chefes de família com 1 a 3 anos de estudo" e computou-se a correlação para o caso de setores censitários (0,79) e para o caso de UP (0,96). Os resultados, mostrados na Tabela 5-1, indicam que as correlações nos setores censitários são significativamente menores que as correlações por unidades de planejamento. Nada menos que 773 correlações são menores para os

setores censitários que para as UPs. Apenas 40 (4%) tem o comportamento oposto. Em algumas situações, ocorre inclusive mudança de sinal, isto é, variáveis correlacionadas negativamente no nível dos setores censitários passam a ser correlacionadas positivamente. 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 em áreas menores.

Os resultados acima indicam que não se pode afirmar que qualquer escala seja a "certa", mas apenas qual dos modelos melhor serve ao que se deseja esclarecer: correlações mais fracas e maior flutuação aleatória, porém com mais homogeneidade interna, ou mais fortes com o viés ocasionado por desconsiderar a dispersão e a heterogeneidade em torno da média nas grande áreas. Como regra geral, quanto mais desagregado o dado, maior a flexibilidade na escolha de modelos; pois agregar em regiões maiores é fácil, mas desagregar impossível.

Tabela 5-1

CORRELAÇÕES ENTRE PARES DE VARIÁVEIS SEGUNDO DIFERENTES UNIDADES DE ÁREAS – SETOR CENSITÁRIO E UNIDADE DE PLANEJAMENTO - PARA O CENSO DE 1991 EM BELO HORIZONTE

Correlações por Unidade de Planejamento

Correlação por Setor Censitário	-0,4/-0,2	-0,2/0,0	0,0/0,2	0,2/0,4	0,4/0,6	0,6/0,8	0,8/1,0	Pares
-0,8/-0,6	0	0	1	1	1	0	2	**5**
-0,6/-0,4	2	11	7	4	2	7	0	**33**
-0,4/-0,2	**3**	23	14	11	10	3	6	**70**
-0,2/0,0	3	**5**	9	27	34	13	21	**112**
0,0/0,2	0	1	**2**	42	75	32	55	**207**
0,2/0,4	0	2	0	**17**	44	50	68	**181**
0,4/0,6	0	2	3	1	**10**	42	110	**168**
0,6/0,8	0	0	2	7	8	**9**	75	**101**
0,8/1,0	0	0	0	4	4	3	**112**	**123**
Totais	**8**	**45**	**38**	**114**	**187**	**159**	**449**	**1000**

Na prática, por razões de confidencialidade, os dados individuais muito raramente estão disponíveis. O que fazer então? Uma possibilidade é trabalhar com os Uma possibilidade é trabalhar com os dadosna maior escala

espacial possível, usualmente denominadas micro-áreas, por exemplo, setores censitários. E utilizar técnicas de agregação ou de otimização combinatória para obter regiões mais agregadas, mas que preservem o fenômeno estudado da melhor forma possível. Deste modo, deve-se reconhecer que o problema da escala é um efeito inerente aos dados agregados por áreas. Ele não pode ser removido e não pode ser ignorado. Para minimizar seu impacto com relação a esses estudos, deve-se procurar utilizar a melhor escala de levantamento de dados disponível e utilizar técnicas que permitam tratar a flutuação aleatória, sempre buscando critérios de agregação dos dados que sejam consistentes com os objetivos do estudo.

5.4 ANÁLISE EXPLORATÓRIA

As técnicas de análise exploratória aplicadas a dados espaciais são essenciais ao desenvolvimento das etapas da modelagem estatística espacial, em geral sensível ao tipo de distribuição, à presença de valores extremos e à ausência de estacionariedade. As técnicas empregadas são, em geral, adaptações das ferramentas usuais. Assim, se na investigação de valores extremos se utiliza ferramentas gráficas como histogramas ou *boxplots*, na análise espacial é importante também investigar *outliers* não só no conjunto dos dados mas também em relação aos vizinhos. Além disso, a não-estacionariedade do processo espacial na região de estudo também deve ser investigada, nos seus vários aspectos: variação na média (primeira ordem), na variância e na covariância espacial.

Visualização de Dados

A forma mais simples e intuitiva de análise exploratória é a visualização de valores extremos nos mapas. Vale ressaltar que o uso de diferentes pontos de corte da variável induz a visualização de diferentes aspectos. Os SIGs dispõem usualmente de três métodos de corte de variável: intervalos iguais, percentis e desvios padrões. No caso de *intervalos iguais*, em que os valores máximo e mínimo são divididos pelo número de classes. Se a variável tem um distribuição muito concentrada de um lado, este corte deixa apenas um número muito pequeno de áreas nas classes da perna mais longa da distribuição; como resultado, a maior parte das áreas será alocada a uma ou duas cores. O uso de *percentis* para definação de classes obriga a alocação dos polígonos em quantidades iguais pelas cores; isto pode mascarar diferenças significativas em valores extremos e dificultar a identificação de áreas críticas. Finalmente, o uso de desvios padrões, no qual a distribuição da variável é apresentada em gradações de cores diferentes para valores acima e abaixo da média, faz a suposição da normalidade da distribuição da variável; esta hipótese é pouco realista no caso de variáveis censitárias em países de

grande desigualdade social com o Brasil.” Em resumo, é parte importante da análise exploratória experimentar diferentes pontos de corte da variável na visualização dos mapas.

As diferentes técnicas de visualização estão ilustradas no exemplo a seguir, em que mostramos a distribuição espacial do indicador que mede a proporção de recém-natos que nasce em boas condições de saúde (Índice de APGAR) para os bairros do Rio de Janeiro, no ano de 1994. Foram geradas duas visualizações, ambas com 5 pontos de corte e 5 cores. Na Figura 5-4, utilizou-se quintis; na Figura 5-5, cinco classes de igual tamanho. Como a distribuição da variável não é simétrica, quando se divide em classes de amplitudes iguais as de valores mais baixos (ou piores), assinaladas em vermelho ficam reduzidas a poucas áreas, enquanto que na divisão em quintis, por definição, um quinto das áreas ficará em cada classe. A pergunta então é: o que se deseja mostrar? Certamente o responsável pela assistência peri-natal da região não ficará satisfeito visualizando um quinto dos bairros como sendo de “alto” risco. Por outro lado, como as áreas onde o índice é mais baixo têm população pequena, a confiabilidade dos valores encontrados pode ser efeito apenas da flutuação aleatória descrita anteriormente. Vale a pena então olhar mapas? Claro que sim, da mesma forma como olhamos histogramas e box-plots, e procurando sempre ver a distribuição utilizando diferentes pontos de corte. Os SIGs em geral tem uma forma padrão, mas dezenas de possibilidades podem e devem ser exploradas.

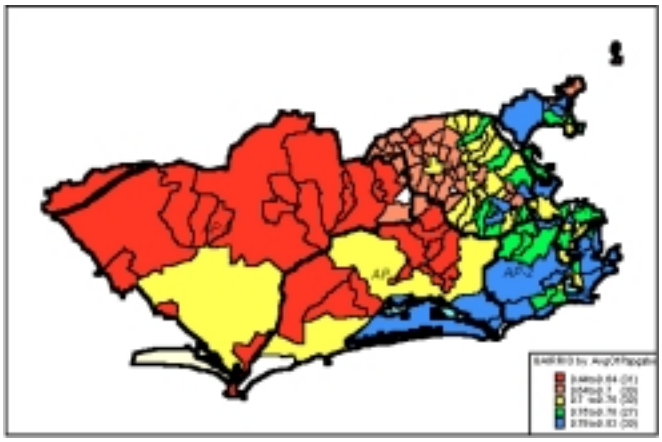


Figura 5-4– Distribuição do índice de APGAR, agrupada em quintis.

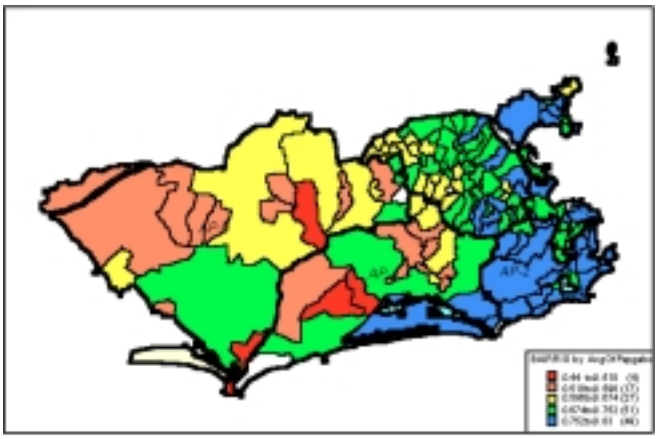


Figura 5-5 - Distribuição do índice de APGAR, agrupada em classes de igual amplitude.

Outra questão interessante é a comparação de mapas. Supondo a distribuição espacial de um indicador em diferentes anos: como visualizar a evolução temporal? Certamente os pontos de corte da variável nos diferentes períodos devem ser os mesmos. Observe na Figura 5-4 a evolução temporal da mortalidade por homicídios para os triênios 79-81 e 90-92, no Estado do Rio de Janeiro. A apresentação dos quintis da *distribuição conjunta* dos indicadores permite visualizar bem o espalhamento desta "doença".

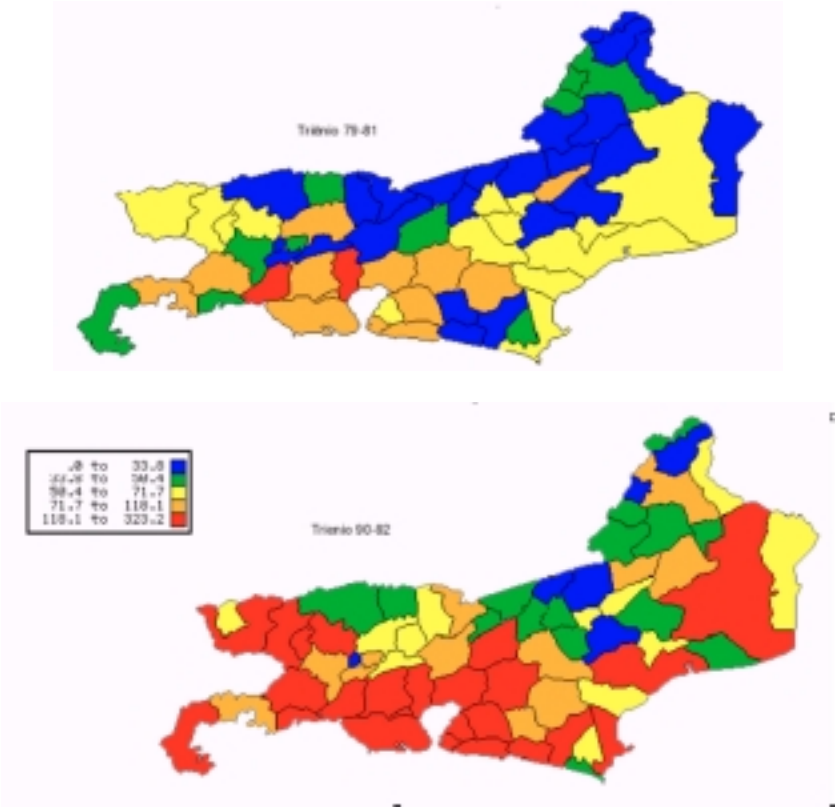


Figura 5-6– Mortalidade por homicídios no Rio de Janeiro, para os triênios 79-81 e 80-92.

Gráficos de Médias e Medianas

Os gráficos de médias e medianas segundo linhas e colunas permitem explorar simultaneamente a presença de tendência (não-estacionariedade de primeira ordem), e não-estacionariedade de segunda ordem, onde a variância e a covariância entre vizinhos não se mantém constante. Para construir estes gráficos, utiliza-se as coordenadas dos centróides das áreas, aproximando-as para um espaçamento regular de forma a montar uma matriz. Calcula-se então as médias e as medianas do indicador ao longo das linhas (eixo Leste-Oeste) e colunas (eixo Norte-Sul) desta matriz. Esta técnica permite identificar a flutuação das medidas ao longo de duas direções, sugerindo a presença de valores discrepantes quando a diferença entre estas é grande, e a tendência ao longo de uma direção quando os valores variam suavemente.

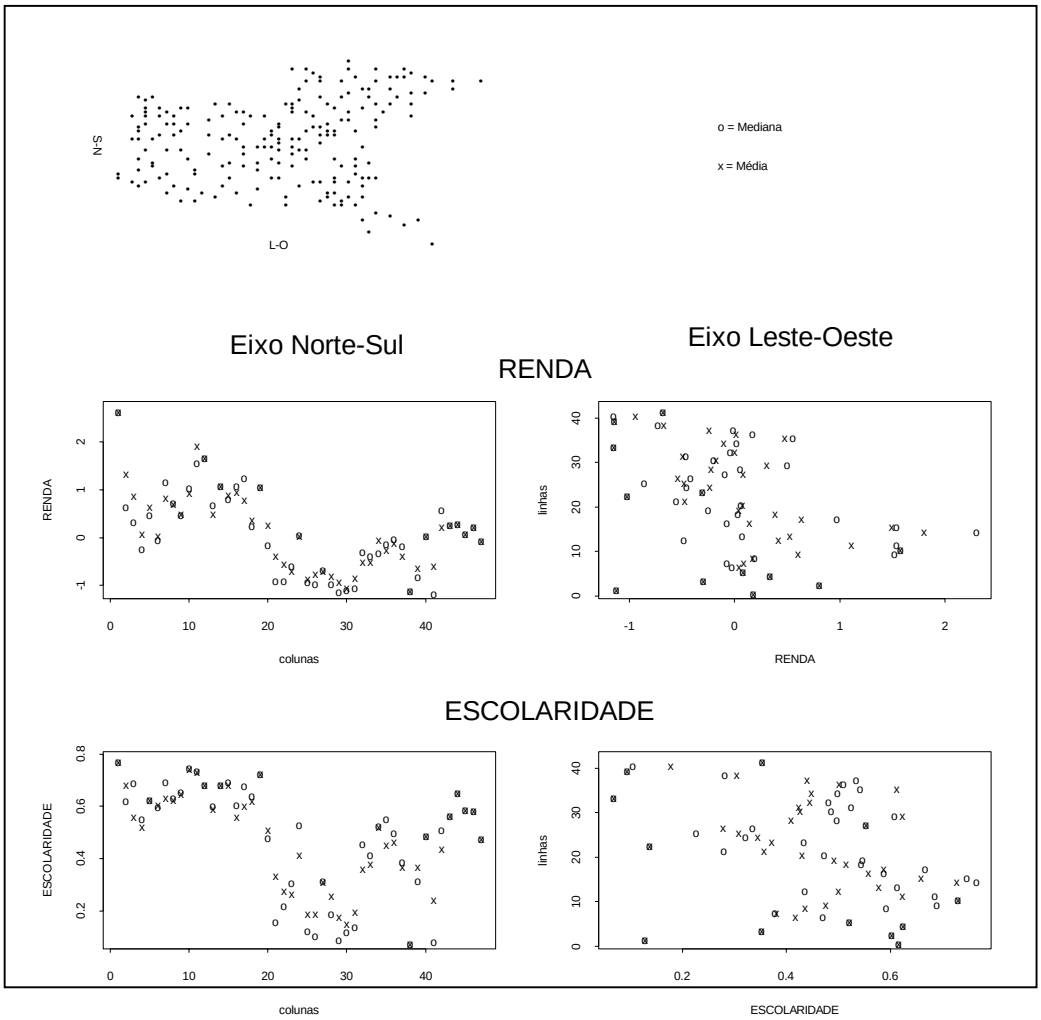


Figura 5-7 – Médias e medianas para escolaridade e renda na Ilha do Governador.

Na Figura 5-7, apresenta-se o resultado desta técnica aplicada a dois indicadores socioeconômicos do censo 1991 – renda média do chefe da família e proporção de chefes de família com escolaridade igual ou superior ao segundo grau – para setores censitários da Ilha do Governador, no Rio de Janeiro. Esta é composta por 225 setores censitários, cujos centróides estão assinalados no primeiro quadro da figura: observe que nas extremidades do "mapa" a quantidade de pontos é muito pequena, e, consequentemente, qualquer medida nesta área será pouco robusta.

No eixo Norte-Sul (colunas) pode-se observar que a renda média do chefe da família apresenta tendência variável, bem menor no centro da região. A mesma coisa acontece para escolaridade, embora com maior flutuação. No eixo Leste-Oeste (linhas), também parece haver algum deslocamento para valores mais altos no sentido leste, mas o descolamento de médias (x) e medianas (o) sugere a presença de valores extremos dos indicadores. A variação na média dos indicadores na região está, aparentemente, dividida entre as duas direções analisadas, e pode-se explorar melhor a tendência através da rotação dos eixos de referência.

Análise de Autocorrelação Espacial

Outra etapa da análise exploratória visa identificar a estrutura de correlação espacial que melhor descreva os dados. A idéia básica é estimar a magnitude da autocorrelação espacial entre as áreas. Neste caso, as ferramentas utilizadas são o índice global de Moran, o índice de Geary e o variograma. Quando se dispõe de grande número de áreas, resultantes por exemplo de escalas espaciais detalhadas, a natureza dos processos envolvidos é tal que é muito provável a existência de diferentes regimes de correlação espacial em diferentes sub-regiões. Para evidenciar estes regimes espaciais, pode-se utilizar os indicadores locais de autocorrelação espacial e o mapa de espalhamento de Moran, descritos também nesta seção. Todas estas estatísticas dependem da definição de vizinhança adotada, discutida a seguir.

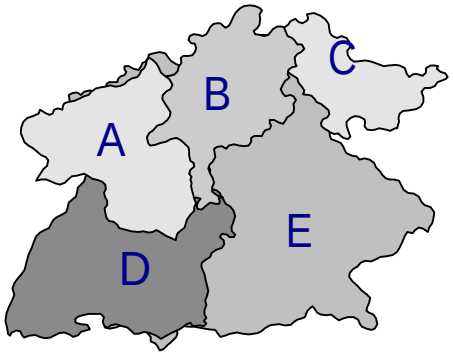
Matrizes de Proximidade Espacial

Para estimar a variabilidade espacial de dados de área, uma ferramenta básica é a *matriz de proximidade espacial*, também chamada matriz de vizinhança. Dado um conjunto de $\boldsymbol{n}$ áreas $\{A_1,..,A_n\}$, construímos a matriz $\mathbf{W^{(1)}}$ (n x n), onde cada um dos elementos w_{ij} representa uma medida de proximidade entre A_i e A_j. Esta medida de proximidade pode ser calculada a partir de um dos seguintes critérios:

- w_{ij} = 1, se o centróide de A_i está a uma determinada distância de A_j; caso contrário w_{ij} = 0

- w_{ij} = 1, se A_i compartilha um lado comum com A_j , caso contrário w_{ij} = 0
- w_{ij} = l_{ij}/l_i, onde l_{ij} é o comprimento da fronteira entre A_i e A_j e l_i é o perímetro de A_i

Como a matriz de proximidade é utilizada em cálculos de indicadores na fase de análise exploratória, é muito útil normalizar suas linhas, para que a soma dos pesos de cada linha seja igual a 1. Isto simplifica muito vários cálculos de índices de autocorrelação espacial, como se verá a seguir. A Figura 5-8 ilustra um exemplo simples de matriz de proximidade espacial, em que os valores dos elementos da matriz refletem o critério de adjacência e foram normalizados.



	A	B	C	D	E
A	0	0,5	0	0,5	0
B	0,25	0	0,25	0,25	0,25
C	0	0,5	0	0	0,5
D	0,33	0,33	0	0	0,33
E	0	0,33	0,33	0,33	0

Figura 5-8- Matriz de proximidade espacial de primeira ordem, normalizada pelas linhas.

A idéia da matriz de proximidade espacial pode ser generalizada para vizinhos de maior ordem (vizinhos dos vizinhos). Com critério análogo ao adotado para a matriz de vizinhança de primeira ordem, pode-se construir as matrizes $\mathbf{W}^{(2)}$, ..., $\mathbf{W}^{(n)}$. Por exemplo, na Figura 5-6, as áreas A e C são vizinhas na matriz de proximidade espacial de ordem 2. No que segue, por simplicidade, os coeficientes da matriz de primeira ordem são designados simplesmente por w_{ij}, e os das matrizes de ordem k por $w_{ij}^{(k)}$ e que essas matrizes estão normalizadas por linhas.

Média Móvel Espacial

Uma forma simples de explorar a variação da tendência espacial dos dados é calcular a média dos valores dos vizinhos. Isto reduz a variabilidade espacial, pois a operação tende a produzir uma superfície com menor flutuação que os dados originais. A média móvel $\hat{\mu}_i$ associada ao atributo z_i, relativo à i-ésima área, pode ser calculada a partir dos elementos w_{ij} da matriz normalizada de proximidade espacial $\mathbf{W}^{(1)}$, tomando-se simplesmente a média dos vizinhos:

$$\hat{\mu}_i = \sum_{j=1}^{n} w_{ij} z_i \qquad (5.1.)$$

A Figura 5-9 ilustra o uso do estimador de média móvel para o percentual de idosos (mais de 70 anos) para os 96 distritos da cidade de São Paulo. Estes dados são indicadores da grande disparidade social da cidade, com uma grande variação entre o centro (onde a proporção de idosos chega a 8%) com a periferia (onde há várias regiões com menos de 1%). O valor máximo do percentual de idosos é de 8,2% e o mínimo de 0,8%, com um desvio padrão de aproximadamente 2%. Com a média local, há um alisamento: o valor mínimo é de 1% e o máximo é reduzido a 6,8%. Pode-se notar, ao comparar os dois mapas da Figura 5-9, que a média móvel local fornece uma visão das grandes *tendências* do fenômeno em estudo e no caso do percentual de idosos, mostra um forte gradiente centro-periferia.

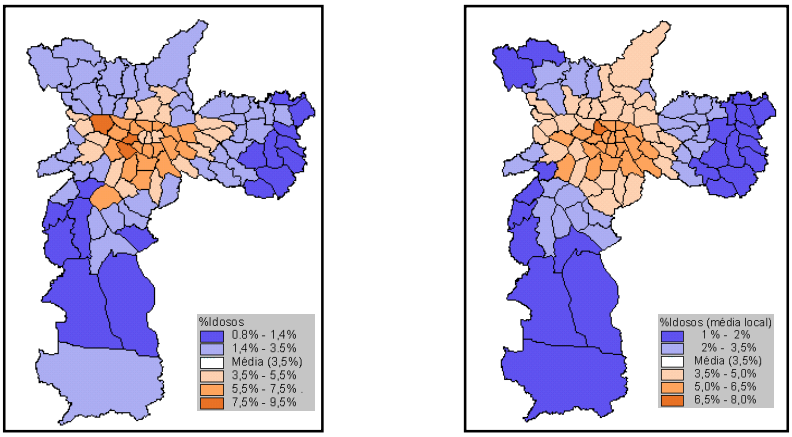


Figura 5-9- Distribuição dos idosos na cidade de São Paulo (censo de 1991). À esquerda, apresentação dos valores por distribuição estatística. À direita, média móvel local.

Indicadores Globais de Autocorrelação Espacial: Índices de Moran e Geary

Um aspecto fundamental da análise exploratória espacial é a caracterização da dependência espacial, mostrando como os valores estão correlacionados no espaço. Neste contexto, as funções utilizadas para estimar quanto o valor observado de um atributo numa região é dependente dos valores desta mesma variável nas localizações vizinhas são a *autocorrelação espacial* e o *variograma*. O índice global de Moran *I*, é a expressão da autocorrelação considerando apenas o primeiro vizinho:

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n}(z_i - \bar{z})^2} \qquad (5.2.)$$

Na equação acima, n é o número de áreas, z_i o valor do atributo considerado na área i, $\bar{z}$ é o valor médio do atributo na região de estudo e w_{ij} os elementos da matriz normalizada de proximidade espacial. Neste caso a correlação será computada apenas para os vizinhos de primeira ordem no espaço, conforme estabelecido pelos pesos w_{ij}. O mesmo cálculo feito para matrizes de proximidade de maior ordem permite estimar a função de autocorrelação para cada ordem de vizinhança (ou "*lag*").

$$I^{(k)} = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^{(k)}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{N}(z_i - \bar{z})^2} \qquad (5.3.)$$

De uma forma geral, o índice de Moran presta-se a um teste cuja hipótese nula é de independência espacial; neste caso, seu valor seria zero. Valores positivos (entre 0 e +1) indicam para correlação direta e negativos, (entre 0 e –1) correlação inversa. Uma vez calculado, é importante estabelecer sua validade estatística. Em outras palavras, será que os valores medidos representam correlação espacial significativa? Para estimar a significância do índice, será preciso associar a este uma distribuição estatística, sendo mais usual relacionar a estatística de teste à distribuição normal. Outra possibilidade, sem pressupostos em relação à distribuição, e abordagem mais comum é um *teste de pseudo-significância*. Neste caso, são geradas diferentes permutações dos valores de atributos associados às regiões; cada permutação produz um novo arranjo espacial, onde os valores estão redistribuídos entre as áreas. Como apenas um dos arranjos corresponde à situação observada, pode-se construir uma distribuição empírica de *I*, como mostrado na Figura 5-10. Se o valor do índice *I* medido originalmente corresponder a um "extremo" da distribuição simulada, então trata-se de valor com significância estatística.

No caso do índice exclusão/inclusão social em São Paulo, apresentado na Figura 5-1, o índice global de Moran medido é 0,642. Uma pseudo-distribuição com 100 valores está mostrada na Figura 5-10. Neste caso, o valor de significância associado é de 5,23, o que nos leva a rejeitar a hipótese nula (não correlação entre as regiões), com significância de 99,5%. Pode-se dizer então que a exclusão social em São Paulo apresenta forte estrutura espacial,

parte variação ampla, ou tendência, parte dependência espacial entre vizinhos.

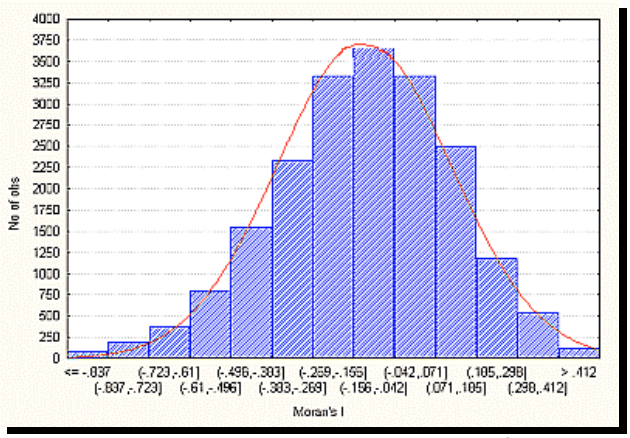


Figura 5-10– Exemplo de distribuição simulada para o índice de Moran.

A hipótese implícita do cálculo do índice de Moran é a estacionariedade de primeira e segunda ordem, e o índice perde sua validade ao ser calculado para dados não estacionários Quando existir não-estacionariedade de primeira ordem (tendência), os vizinhos tenderão a ter valores mais parecidos que áreas distantes, pois cada valor é comparado à média global, inflacionando o índice. Da mesma forma, se a variância não é constante, nos locais de maior variância o índice será mais baixo, e vice-versa. Quando o dado é não-estacionário, a função de autocorrelação continua decaindo mesmo após ultrapassar a distância onde há influências locais. Algumas variações deste modelo são o teste *C* de Geary e o teste *Ipop*. O primeiro (*C* de Geary) difere do teste *I* de Moran por utilizar a diferença entre os pares, enquanto que Moran utiliza a diferença entre cada ponto e a média global. Assim, o indicador *C* de Geary assemelha-se ao variograma, e o *I* de Moran ao correlograma.

$$C = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i - z_j)^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\sum_{i=1}^{n} z_i^{\,2}} \qquad (5.4.)$$

O teste *Ipop* também é utilizado para detectar desvios de uma distribuição espacial aleatória, porém incorpora a variação da população nas áreas. Assim, é sensível à ocorrência de aglomerado intra-área – ou seja, a ocorrência de elevado número de casos numa pequena população de um único município – além dos aglomerados entre áreas, onde municípios com muitos casos são adjacentes. Portanto o índice *Ipop* pode ser decomposto em

um componente intra-áreas e outro inter-áreas, que podem ser apresentados sob forma percentual nos resultados. A hipótese nula (H_0) assume que a variação geográfica do número de casos segue a variação geográfica do tamanho da população, sendo particularmente útil quando a população das áreas não é homogênea.

$$Ipop = \frac{N^2 \sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}(e_i - d_i)(e_j - d_j) - N(1-2\bar{b})\sum_{i=1}^{m} w_{ij}e_i - N\bar{b}\sum_{i=1}^{m} w_{ii}d_i}{\left(X^2 \sum_{i=1}^{m}\sum_{j=1}^{m} d_i d_j w_{ij} - X\sum_{i=1}^{m} d_i w_{ii}\right)\bar{b}(1-\bar{b})} \qquad (5.5.)$$

onde: $\tilde{a}$ → Número de áreas
k → Número total de casos em todas as áreas.
$å_{i}$ → Número de casos na área i
$É_{i}$ → Proporção de casos na área i $(É_i = å_i / k)$
u → População total em todas as áreas
$\tilde{n}_{i}$ → Tamanho da população na área i
$Ç_{i}$ → Proporção de população na área i $(Ç_i = \tilde{n}_i / k)$
w_{i} → Diferença entre a taxa u_i e a média de u
$ï_{ij}$ → Pesos atribuídos conforme a conexão entre as áreas i e j
$Ä$ → Prevalência média (k/u)

A tabela 5.2 apresenta os resultados dos testes de aglomerado espacial para a mortalidade por homicídios no Estado do Rio. Observe que o grau de significância do teste *Ipop* é maior que o Moran, e que aproximadamente metade da agregação deve-se a fatores intra-municipais. Ou seja, além de municípios próximos apresentarem padrões semelhantes, existe um excesso de casos dentro dos municípios violentos, que ultrapassa o esperado em função da população.

TABELA 5.2

RESULTADOS DOS TESTES DE AGLOMERADOS ESPACIAIS:
HOMICÍDIOS NO RIO DE JANEIRO, 90-92

	Moran I	Ipop
Indicador	0,5861	0,00015
p-valor	7,5091	88,9238
% entre áreas	-	54,3
% intra áreas	-	45,7

Variograma

De maneira análoga ao apresentado no capítulo 3, podemos utilizar o variograma como indicador da dependência espacial. Para tanto, associamos o valor único do atributo de cada área a um ponto, usualmente o centro geométrico ou populacional do polígono. Com base nestas localizações, calcula-se a função variograma. Note-se quando o dado é não-estacionário, também o variograma não se estabiliza, mas continua crescendo sempre com a distância. Como exemplo de uso do variograma para dados de área, a Figura 5-11 ilustra o Índice de Desenvolvimento Humano – IDH – para o estado de São Paulo, calculado pelo IPEA, com base no censo de 1991. A Figura 5-12 apresenta o variograma do IDH, computado a partir dos centróide de cada município.

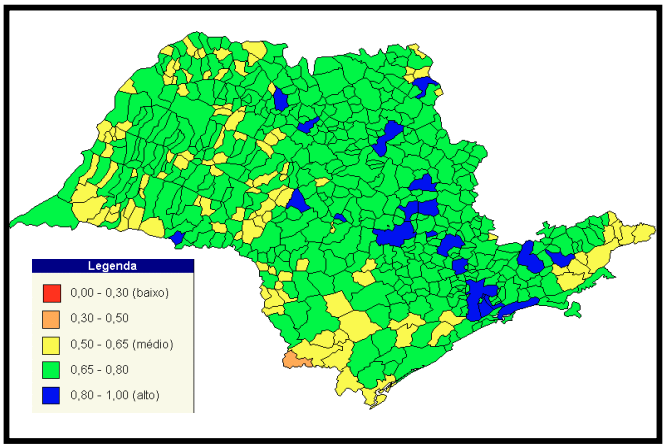


Figura 5-11- IDH para São Paulo (censo de 1991)

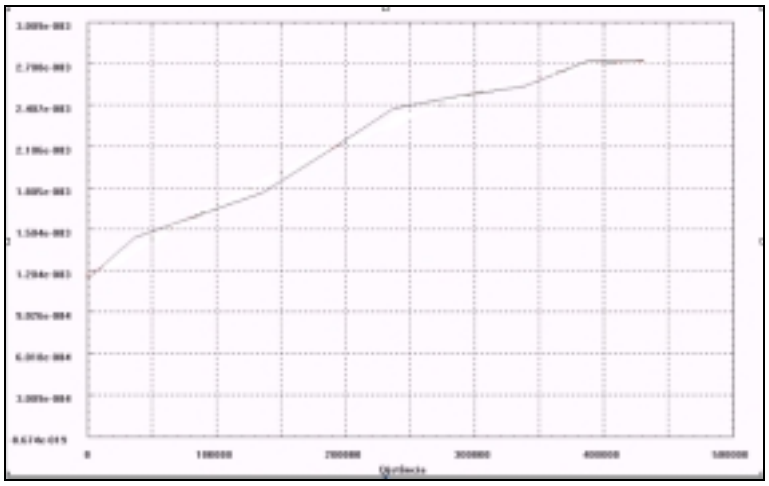


Figura 5-12 Variograma experimental do IDH para São Paulo (censo de 1991). Passo de amostragem: 40 km (tolerância : 20 km).

O que mostra o variograma da Figura 5-10? No eixo dos X, apresentam-se as distâncias entre os municípios, e no eixo Y, a média do quadrado das diferenças do IDH, para municípios separados por faixas de distância, com intervalos de 40 km e tolerância de 20 km. Assim, o primeiro ponto calcula a diferença de IDH entre os municípios cuja distância entre os centros seja de 20 a 60 Km, e assim por diante, até a distância de 400 km. O gráfico evidencia uma forte dependência espacial entre os indicadores de qualidade de vida dos municípios de São Paulo. Trata-se de um resultado dos processos de ocupação do estado, que seguiram perspectivas regionais. A partir da lógica de expansão do café do século XIX, observa-se hoje uma região de forte produção agrícola situada ao longo do eixo da rodovia Anhanguera, a predominância da pecuária na região do Oeste Paulista, e uma forte concentração industrial na região metropolitana de São Paulo, no ABC e no médio Vale do Paraíba. Assim, todos os processos históricos apontam para uma dependência espacial no desenvolvimento econômico no estado.

Para considerar um exemplo adicional, considere-se o estudo sobre mortalidade por homicídios na região Sudeste. que são a causa de mais de 20% dos óbitos dos homens entre 15 e 45 anos no Brasil. A Figura 5-13 ilustra a distribuição espacial da mortalidade por homicídios, usando como indicador o logaritmo do coeficiente de mortalidade específico , por 100.000 residentes do mesmo grupo etário. Entendendo o processo da violência como o de uma "epidemia" da modernidade, que se "propaga" no espaço, uma simples observação visual permite identificar uma elevada ocorrência de mortes violentas no RJ, com uma tendência espacial capital-interior. No caso de ES e SP, há uma concentração próxima da capital e grandes cidades. No entanto, em MG, as áreas mais violentas situam-se longe das regiões metropolitanas, o que indica um padrão espacial distinto. Adicionalmente, há uma marcada transição na fronteira entre MG e RJ, indicando uma mudança nas condições de disseminação da "epidemia da violência". Cabe lembrar que foi utilizado o logaritmo do indicador, dado ser a distribuição do mesmo bastante concentrada em torno de valores muito baixos, com uma grande cauda a direita.

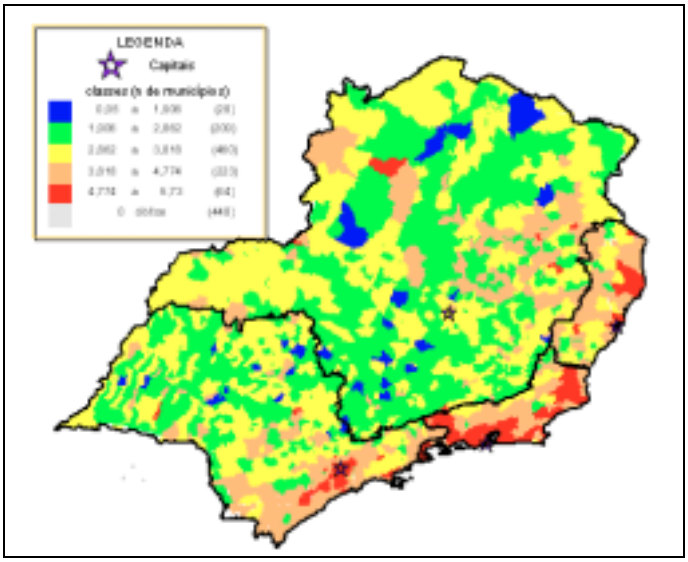


Figura 5-13 - Mortalidade por homicídos, região Sudeste do Brasil

O correlograma da Figura 5-14 apresenta a autocorrelação espacial entre os municípios de cada estado, expressa através da função definida pela equação 5.3. O gráfico indica a existência de uma forte tendência espacial no RJ, pois a função de autocorrelação não se estabiliza com a distância, mas continua decrescente, ao contrário de MG, que não apresenta dependência espacial marcante. Em outras palavras, no RJ, se o município vizinho ao seu é violento, é altamente provável que a sua cidade também o seja; todo o estado apresenta uma estrutura de violência regionalizada, e a violência decai no interior do estado. Em MG, esta padrão não é observado: a violência parece flutuar aleatoriamente.

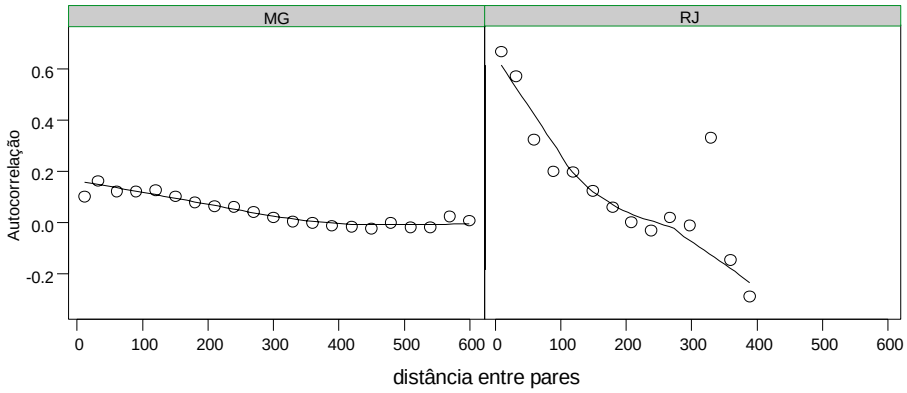


Figura 5-14. Correlograma da mortalidade por homicídios nos estados do Sudeste.

Diagrama de Espalhamento de Moran

O diagrama de espalhamento de Moran é uma maneira adicional de visualizar a dependência espacial. Construído com base nos valores

normalizados (valores de atributos subtraídos de sua média e divididos pelo desvio padrão), permite analisar o comportamento da variabilidade espacial. A idéia é comparar os valores normalizados do atributo numa área com a média dos seus vizinhos, construindo um gráfico bidimensional de *z* (valores normalizados) por *wz* (média dos vizinhos), que é dividido em quatro quadrantes, como mostrado na Figura 5-15 para o índice de exclusão/inclusão social de São Paulo, censo de 1991. Os quadrantes podem ser interpretados como:

- Q1 (valores positivos, médias positivas) e Q2 (valores negativos, médias negativas): indicam pontos de associação espacial positiva, no sentido que uma localização possui vizinhos com valores semelhantes.
- Q3 (valores positivos, médias negativas) e Q4 (valores negativos, médias positivas): indicam pontos de associação espacial negativa, no sentido que uma localização possui vizinhos com valores distintos.

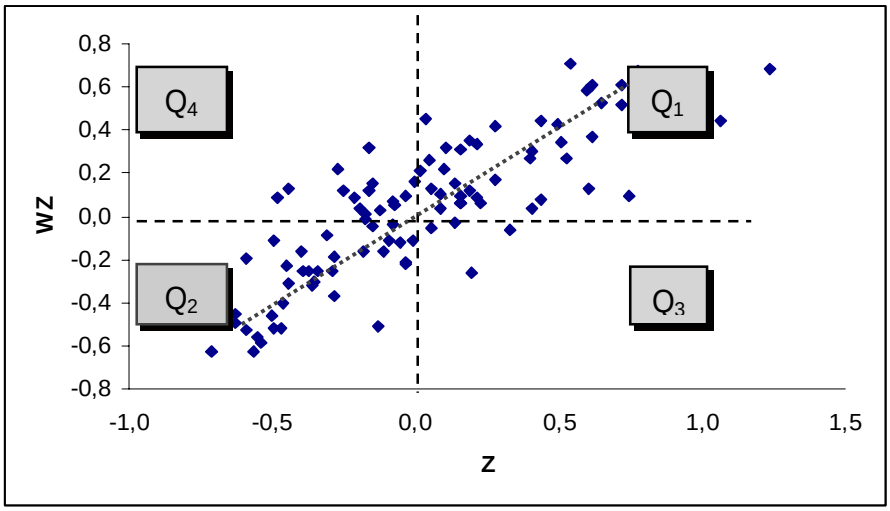
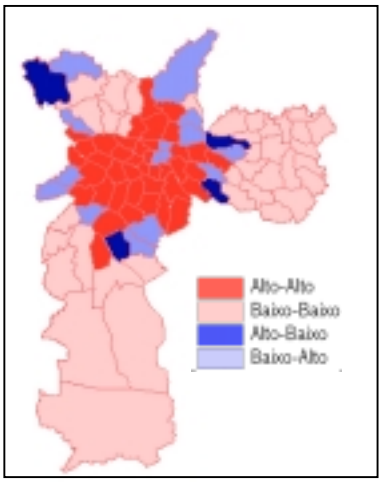


Figura 5-15 Diagrama de Espalhamento de Moran para o índice de exclusão/inclusão social de São Paulo, censo de 1991.

O diagrama de espalhamento de Moran corrobora os resultados apresentados, onde indicamos que o índice global de Moran para o indicador de exclusão/inclusão social para os distritos de São Paulo apresentava valor estatisticamente significativo. Como mostrado na Figura 5-15, a maior parte dos distritos de São Paulo está localizado nos quadrantes Q1 e Q2, que apresentam associação espacial positiva. Os pontos localizados nos quadrantes Q3 e Q4 podem ser vistos como regiões que não seguem o mesmo processo de dependência espacial das demais observações. Evidentemente, o diagrama reflete a estrutura espacial nas duas escalas de análise: vizinhança e tendência.

O índice de Moran *I* é equivalente ao coeficiente de regressão linear que indica a inclinação da reta de regressão (α) de ***wz*** em ***z***. Para o caso dos dados apresentados na Figura 5-15, este coeficiente é igual a 0,642, o mesmo valor calculado aplicando-se a fórmula da equação 5.3. O diagrama de espalhamento de Moran também pode ser apresentado na forma de um mapa temático bidimensional, no qual cada polígono é apresentado indicando-se seu quadrante no diagrama de espalhamento, como ilustra a Figura 5-16, em que mostramos o mapa do espalhamento do índice de Moran para o índice de exclusão/inclusão social da cidade de São Paulo em 1991. Nesta figura, "Alto-Alto", "Baixo-Baixo", "Alto-Baixo" e "Baixo-Alto" indicam, respectivamente, os quadrantes Q1, Q2, Q3 e Q4, mostrados na Figura 5-15. Nota-se uma forte polarização centro-periferia e observa-se que os distritos localizados nos quadrantes Q3 e Q4 (indicados pela cor azul) podem ser entendidos como regiões de transição entre o centro da cidade (que tende a apresentar valores positivos do índice de exclusão/inclusão social) e as duas grandes periferias de São Paulo (zona Sul e zona Leste).

Figura 5-16 Mapa de Espalhamento de Moran para o índice de exclusão/inclusão



social da cidade de São Paulo, censo 1991

Indicadores Locais de Associação Espacial

Os indicadores globais de autocorrelação espacial, como o índice de Moran, fornecem um único valor como medida da associação espacial para todo o conjunto de dados, o que é útil na caracterização da região de estudo como um todo. Quando lidamos com grande número de áreas, é muito provável que ocorram diferentes regimes de associação espacial e que apareçam máximos locais de autocorrelação espacial, onde a dependência espacial é ainda mais pronunciada. Assim, muitas vezes é desejável examinar padrões em maior detalhe. Para tanto, é preciso utilizar indicadores de associação espacial que possam ser associados às diferentes localizações de uma variável distribuída espacialmente. Os indicadores locais produzem um

valor específico para cada área, permitindo assim a identificação de agrupamentos. O índice local de Moran pode ser expresso para cada área *i* a partir dos valores normalizados z_i do atributo como:

$$I_i = \frac{z_i \sum_{j=1}^{n} w_{ij} z_j}{\sum_{j=1}^{n} z_j^2} \qquad (5.6.)$$

A significância estatística do uso do índice de Moran local é computada de forma similar ao caso do índice global. Para cada área, calcula-se o índice local, e depois permuta-se aleatoriamente o valor das demais áreas, até obter uma pseudo-distribuição para a qual possamos computar os parâmetros de significância.Uma vez determinada a significância estatística do índice local de Moran, é útil gerar um mapa indicando as regiões que apresentam correlação local significativamente diferente do resto do dados. Estas regiões podem ser vistas como "bolsões" de não-estacionariedade, pois são áreas com dinâmica espacial própria e que merecem análise detalhada. Para o caso do índice de exclusão/inclusão social da cidade de São Paulo (censo de 1991), esse mapa (Figura 5-17) mostra claramente os agregados de pobreza e de riqueza na cidade. Na zona Leste e na zona Sul de São Paulo há regiões críticas, onde o agravamento das condições sociais resulta numa degradação significativa das condições de vida.

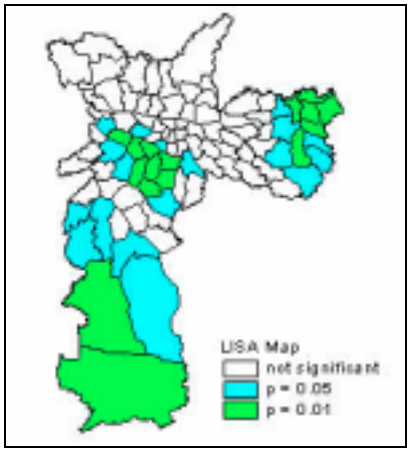


Figura 5-17 – Indicador de autocorrelação espacial para o índice de exclusão/inclusão social de São Paulo (censo de 1991). Apenas os valores com significância maior que 95% estão mostrados.

5.5 ESTIMAÇÃO DE INDICADORES:

A seção 5.3 apresentou o problema de agregação de contagens em áreas, com a recomendação final de utilizar a melhor resolução espacial disponível. Na prática, o uso desta estratégia requer um tratamento adicional nos dados, principalmente nos casos de pequenas áreas em que calculamos taxas sobre um universo populacional reduzido. Para entender melhor o problema, considere-se a Figura 5-18 que apresenta um mapa temático com a mortalidade infantil dos bairros do Rio de Janeiro, em 1994. Neste mapa, o Rio está dividido em 148 bairros, e a taxa de mortalidade infantil anual para cada bairro, expressa o número de óbitos de menores de 1 ano, por mil nascidos vivos.

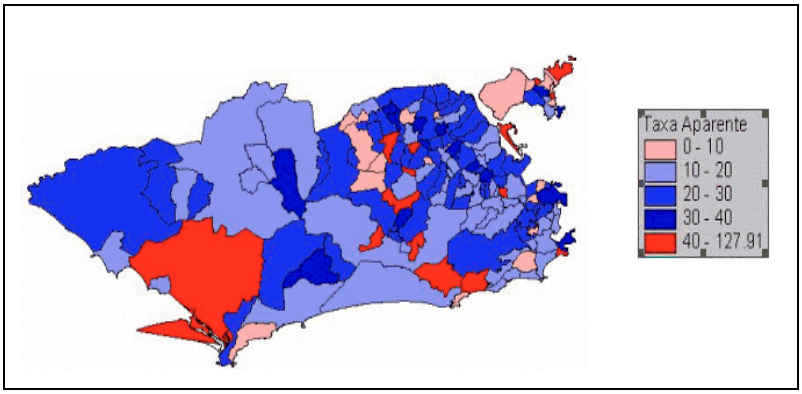


Figura 5-18 - Taxa total de mortalidade infantil por mil nascidos vivos no Rio de Janeiro, em 1994.

Numa primeira leitura, este mapa choca pelas altas taxas de mortalidade de vários bairros, com 15 bairros apresentando uma taxa maior que 40 óbitos por mil nascidos, e 2 casos com taxas acima de 100 por mil nascidos. Um observador desatento poderia concluir que todos estes bairros apresentam um grave problema social. Na realidade, muitos destes valores extremos ocorrem nos bairros com pequenas populações, pois a divisão da cidade utilizada esconde enormes diferenças na população em risco, variando de 15 até 7500 crianças por bairro. Por exemplo, considere uma região com 15 crianças nascidas e nenhuma morte, o que aparentemente indicaria uma situação ideal. Se apenas uma criança morre neste ano, a taxa passa de 0 por mil para 66 por mil !

Tais problemas são típicos de recobrimentos espaciais sobre divisões político-administrativas, onde se analisam áreas com valores muito distintos de população em risco. Vários estudos têm mostrado que em divisões políticas como bairros e municípios apresentam relações inversas de área e população, isto é, os maiores bairros em população tendem a ter menores

áreas, e vice-versa. Por isso mesmo, freqüentemente o que mais chama a atenção num mapa temático de taxas, que são os valores extremos, muitas vezes são resultado de um número reduzidíssimo de observações, sendo portanto menos confiável, ou seja, apenas flutuação aleatória .

Para suavizar a flutuação aleatória, considera-se que a taxa estimada pela divisão simples entre contagem de óbitos e de população – taxa observada – é apenas **uma** realização de um processo não observado, e que é tanto menos confiável quanto menor a população. Assim, propõe-se re-estimar uma taxa mais próxima do risco real ao qual a população está exposta. A primeira providência é fazer um gráfico que expresse a taxa em função da população em risco, como mostrado na Figura 5-19.

Figura 5-19 Taxa de mortalidade infantil no Rio de Janeiro em 1994 em função do número de nascimentos por bairro.

No caso do Rio, a taxa média de mortalidade infantil da cidade, em 1994, foi de 21 óbitos por mil nascidos. Neste gráfico, observa-se que os bairros com maior população apresentam taxas próximas da média da cidade. Conforme diminui a população em risco, aumenta muito a flutuação da taxa medida, formando o que já foi denominado de “efeito funil”. Nos bairros de menor população, esta variação oscilou de 0 a quase 130 por mil.

É razoável supor que as taxas das diferentes regiões estão autocorrelacionadas, e levar em conta o comportamento dos vizinhos para estimar uma taxa mais realista para as regiões de menor população. Esta formulação sugere o uso de técnicas de estimação bayesiana. Nesse contexto, considera-se que a taxa “real” θ_i associada a cada área não é conhecida, e dispomos de uma taxa observada $t_i = z_i / n_i$, onde n_i é o número de pessoas observadas, é z_i é o número de eventos na i-ésima área.

A idéia do estimador bayesiano é supor que a taxa θ_i é uma variável aleatória, que possui uma média μ_i e uma variância σ_i^2. Pode ser demonstrado que o melhor estimador bayesiano é dado por uma combinação linear entre a taxa observada e a média μ_i:

$$\hat{\theta}_i = w_i t_i + (1 - w_i)\mu_i \ , \qquad (5.7.)$$

O fator w_i é dado por:

$$w_i = \frac{\sigma_i^2}{\sigma_i^2 + \mu_i / n_i} \qquad (5.8.)$$

O peso w_i é tanto menor quanto menor for a população em estudo da *i*-ésima área e reflete o grau de confiança a respeito de cada taxa. Para o caso de populações reduzidas, a confiança na taxa observada diminui e a estimativa da taxa se aproxima de nosso modelo a priori (ou seja, se aproxima de μ). Regiões com populações muito baixas terão uma correção maior, e regiões populosas terão pouca alteração em suas taxas. Logo θ_i será estimado, quando n for pequeno, com maior peso da média da vizinhança.

Neste ponto, deve-se observar que a formulação bayesiana requer as médias e variâncias μ_i e σ_i^2 para cada uma das áreas. A abordagem mais simples para tratar a estimação destes parâmetros é o chamado *estimador bayesiano empírico*. Este estimador parte da hipótese que a distribuição da variável aleatória θ_i é a mesma para todas as áreas; isto implica que todas as médias e variâncias são iguais. Pode-se então estimar μ_i e σ_i^2 diretamente a partir dos dados. Neste caso, calcula-se μ_i a partir das taxas observadas:

$$\hat{\mu} = \frac{\sum y_i}{\sum n_i} \qquad (5.9.)$$

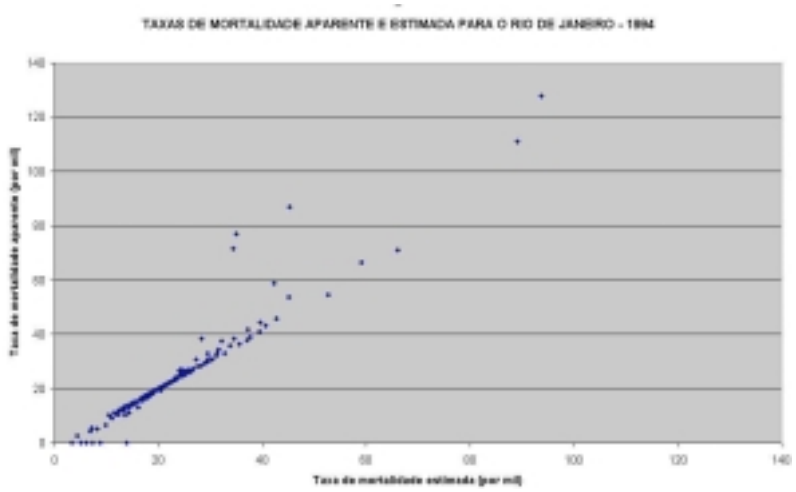
E estima-se a variância σ_i^2 a partir da variância das taxas observadas com relação à média estimada:

$$\sigma^2 = \frac{\sum n_i (t_i - \hat{\mu})^2}{\sum n_i} - \frac{\hat{\mu}}{\bar{n}} \qquad (5.10.)$$

As regiões terão suas taxas re-estimadas aplicando-se uma média ponderada entre o valor medido e a taxa média global, em que o peso da média será inversamente proporcional à população da região. Ao aplicarmos esta correção às taxas de mortalidade infantil do Rio de Janeiro, observamos que há uma redução significativa nos valores extremos. Por exemplo, a Cidade Universitária (Ilha do Fundão), onde nasceram 13 crianças em 1994, apresentou uma taxa aparente de 76 por mil nascidos vivos e uma taxa

corrigida de 36 por mil. Bairros com pouca população no grupo de risco apresentaram reduções semelhantes, enquanto que bairros mais populosos mantiveram as taxas originalmente medidas. A comparação entre a taxa primária e o valor estimado está apresentada na Figura 5-18. Em resumo, é preciso extremo cuidado ao produzir mapas temáticos, especialmente em casos onde apresentamos taxas medidas sobre populações com valores reduzidos.

Figura 5-18. Comparação entre a taxa de mortalidade infantil observada e a taxa estimada



pelo método bayesiano empírico.

O estimador bayesiano empírico pode ser generalizado para incluir efeitos espaciais. Neste caso, a idéia é fazer a estimativa bayesiana localmente, convergindo em direção a uma média local e não a uma média global. Basta aplicar o método anterior em cada área considerando como "região" a sua vizinhança. Isto é equivalente a supor que as taxas da vizinhança da área *i* possuem média μ_i e variância σ_i^2 comuns. Neste caso, pode-se falar em *estimativa bayesiana empírica local*. A seguir, apresenta-se a detecção de hanseníase em Recife (Figura 5-20) onde foi utilizado esse método local para estimar a taxa da doença nos bairros da cidade. Através do mapa "corrigido" foi possível indicar bairros prioritários para a atuação da vigilância epidemiológica por apresentarem valores altos mesmo após suavização do indicador.

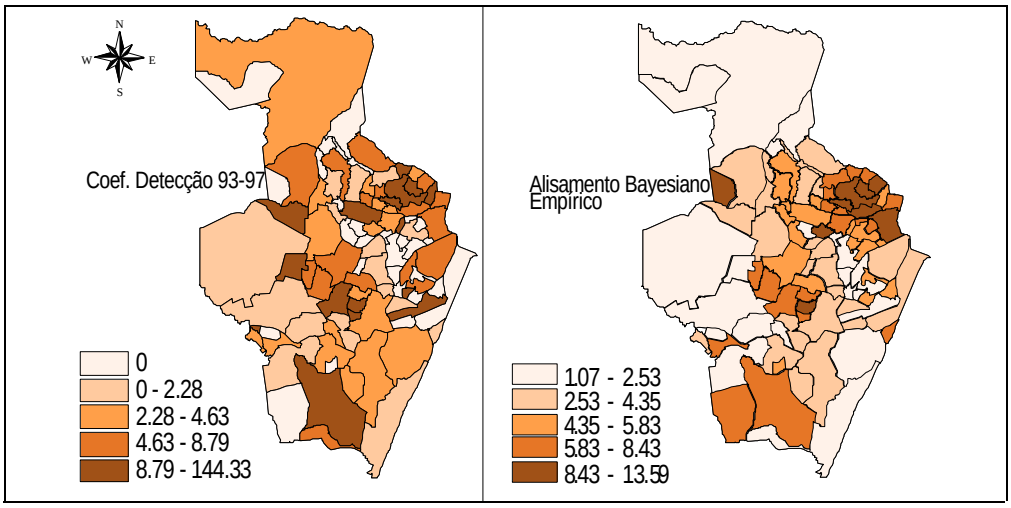


Figura 5-20 - Taxas de detecção média de hanseníase em menores de 15 anos, período 1993-1997, por bairro do Recife, e taxas estimadas através de alisamento bayesiano.

Como apresentado acima, o *estimador bayesiano empírico* parte da hipótese que a distribuição da variável aleatória θ_i é a mesma para todas as áreas e que as médias e variâncias μ_i e σ_i^2 para cada uma das áreas são iguais. Deve-se lembrar que esta hipótese nem sempre é realista, pois em estatísticas socioeconômicas (como no caso dos dados de saúde discutidos) as características das populações estudadas são muito heterogêneas. Deste modo, em muitos casos é desejável fazer a hipótese de que cada área tem seu próprio padrão (e os μ_i e σ^2_i são distintos); isto implica em estimar a distribuição conjunta $Z = \{Z_1, ..., Z_n\}$ das variáveis aleatórias.

À primeira vista, a estimativa da distribuição conjunta pode parecer impossível, dado que está disponível para análise apenas uma amostra de cada uma das variáveis aleatórias, ou seja, sabe-se apenas o valor coletado em cada unidade de área. Entretanto, os *estimadores bayesianos completos (full Bayes)* tornaram possível resolver o problema, através da utilização de técnicas de simulação baseadas em MCMC – *Markov Chain Monte Carlo* – para a inferência dos parâmetros de interesse. Em função da complexidade de formulação, este livro não aborda os estimadores bayesianos baseados em MCMC. O leitor deve referir-se à bibliografia no final do capítulo para maiores detalhes.

5.6 MODELOS DE REGRESSÃO

Um dos tipos de estudos mais comuns com dados de área é o uso de modelos de regressão. Um modelo de regressão é uma ferramenta estatística que utiliza o relacionamento existente entre duas ou mais variáveis de maneira que uma delas possa ser descrita ou o seu valor estimado a partir das demais. Na situação dos dados espaciais, quando está presente a autocorrelação espacial, as estimativas do modelo devem incorporar esta estrutura espacial, uma vez que a dependência entre as observações altera o poder explicativo do modelo. A significância dos parâmetros é usualmente superestimada, e a existência de variações em larga escala pode até mesmo induzir a presença de associações espúrias.

Neste livro, não será feita uma descrição detalhada dos modelos tradicionais de regressão, disponível em diversos livros consagrados, mas apenas será apresentado um breve resumo, necessário ao entendimento dos modelos de regressão espacial. O objetivo geral de uma análise de regressão linear é quantificar a relação linear entre uma variável dependente e um conjunto de variáveis explicativas, conforme expresso na equação matricial:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \varepsilon,\ \varepsilon \sim \mathrm{N}(0, \sigma^2) \text{ ou} \qquad (5.11.)$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ .. \\ .. \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & .. & X_{1k-1} \\ 1 & X_{21} & .. & X_{2k-1} \\ .. & .. & .. & .. \\ .. & .. & .. & .. \\ 1 & X_{n1} & .. & X_{nk-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ .. \\ .. \\ \beta_{k-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ .. \\ .. \\ \varepsilon_n \end{bmatrix} \qquad (5.12.)$$

onde $\boldsymbol{Y}$ é a variável dependente, composta de um vetor *(n* x *1)* de observações tomadas em cada um das *n* áreas, $\boldsymbol{X}$ é uma matriz *(n* x *k)* com *k-1* variáveis explicativas também tomadas nas *n* áreas, $\boldsymbol{\beta}$ é vetor *(k* x *1)* com os coeficientes de regressão, e ε é um vetor *(n* x *1)* de erros aleatórios, ou resíduos.

Tipicamente, quando se faz uma análise de regressão, procura-se alcançar dois objetivos: (a) encontrar um bom ajuste entre os valores preditos pelo modelo e os valores observados da variável dependente; (b) descobrir quais da variáveis explicativas contribuem de forma significativa para este relacionamento linear. Para tanto, a hipótese padrão é que as observações não são correlacionadas, e, consequentemente, que os resíduos ε_i do modelo também são independentes e não-correlacionados com a variável dependente, tem variância constante, e apresentam distribuição normal com média zero.

No entanto, no caso de dados espaciais, onde está presente a dependência espacial, é muito pouco provável que a hipótese padrão de observações não correlacionadas seja verdadeira. No caso mais comum os resíduos continuam apresentando a autocorrelação espacial presente nos dados, que pode se manifestar por diferenças regionais sistemáticas nas relações do modelo, ou ainda por uma tendência espacial contínua.

A investigação dos resíduos da regressão em busca de sinais de estrutura espacial é o primeiro passo em uma regressão espacial. As ferramentas usuais de análise gráfica e o mapeamento de resíduos, podem dar as primeiras indicações de que os valores observados estão mais correlacionados do que seria esperado sob uma condição de independência. Neste caso, utilizar os testes de autocorrelação espacial – Moran e Geary – nos resíduos da regressão informa sobre sua presença. Em caso de existir autocorrelação, deve-se especificar um modelo que considere a interferência causada pela mesma.

No restante desta seção, apresentamos vários tipos de modelos de regressão que permitem incorporar efeitos espaciais, desde aqueles que tratam a estrutura espacial de forma global (como um único parâmetro) até modelos em que os parâmetros variam continuamente no espaço.

Modelos com Efeitos Espaciais Globais

A inclusão explícita de efeitos espaciais em modelos de regressão pode ser feita de diferentes formas. A classe de modelos de regressão espacial mais simples, chamados de *modelos com efeitos espaciais globais*, supõe que é possível capturar a estrutura de correlação espacial num único parâmetro, que é adicionado ao modelo de regressão tradicional. Neste caso, tem-se duas alternativas para tratar a autocorrelação global em um modelo de regressão. Na primeira, a autocorrelação espacial ignorada é atribuída à variável dependente *Y*. Esta abordagem é denominada como *modelo espacial autoregressivo misto* ("*Spatial AutoRegressive*– SAR" ou ainda como "*spatial lag model*"), dado que se considera a dependência espacial através da adição ao modelo de regressão de um novo termo na forma de uma relação espacial para a variável dependente. Formalmente isto é expresso como:

$$\boldsymbol{Y} = \rho \boldsymbol{WY} + \boldsymbol{X\beta} + \varepsilon, \qquad (5.13.)$$

onde $\boldsymbol{W}$ é a matriz de proximidade espacial, e o produto $\boldsymbol{WY}$ expressa a dependência espacial em $\boldsymbol{Y}$ e ρ é o *coeficiente espacial autoregressivo*. A hipótese nula para a não existência de autocorrelação é que $\rho = 0$. A idéia básica neste modelo é incorporar a autocorrelação espacial como componente do modelo. Em termos de componentes individuais, este modelo pode ser expresso como

$$y_i = \rho\left(\sum_j w_{ij} y_j\right) + \sum_{i=1} x_i \beta_i + \varepsilon_i \qquad (5.14.)$$

O segundo tipo de modelo de regressão espacial com parâmetros globais considera que os efeitos espaciais são um ruído, ou perturbação, ou seja, fator que precisa ser removido. Neste caso, os efeitos da autocorrelação espacial são associados ao termo de erro ε e o modelo pode ser expresso por:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} = \lambda \boldsymbol{W} + \xi, \qquad (5.15.)$$

onde $\boldsymbol{W\varepsilon}$ é a componente do erro com efeitos espaciais, λ é o coeficiente autoregressivo e ξ é a componente do erro com variância constante e não correlacionada. A hipótese nula para a não existência de autocorrelação é que $\lambda = 0$, ou seja, o termo de erro não é espacialmente correlacionado. Este modelo é também chamado de modelo do *erro espacial* ("*spatial error model*" ou ainda "*Conditional AutoRegressive*" - CAR).

A partir da equação 5.15, pode-se mostrar que o modelo de erro espacial pode também ser expresso como:

$$\boldsymbol{Y} - \lambda \boldsymbol{WY} = \boldsymbol{X\beta} - \lambda \boldsymbol{WX\beta} + \xi \qquad (5.16.)$$

ou ainda como

$$(\boldsymbol{I} - \lambda \boldsymbol{W})\boldsymbol{Y} = (\boldsymbol{I} - \lambda \boldsymbol{W})\boldsymbol{X\beta} + \xi \qquad (5.17.)$$

o que pode ser visto como uma regressão não-espacial nas variáveis "filtradas"

$$\boldsymbol{Y}^* = (\boldsymbol{I} - \lambda \boldsymbol{W})\boldsymbol{Y}, \quad \boldsymbol{X}^* = (\boldsymbol{I} - \lambda \boldsymbol{W})\boldsymbol{X} \qquad (5.18.)$$

Na prática, a distinção entre os dois tipos de modelos de regressão espacial com parâmetros globais é difícil pois, apesar da diferença nas suas motivação, eles são muito próximos em termos formais. Estes modelos estão incluídos em ambientes de estatística espacial avançados, como nos softwares SpaceSat™, S-Plus™ e R, esse de domínio público. Nas referências no final do capítulo, o leitor poderá encontrar indicações sobre como tais modelos podem ser estimados e sobre testes de hipóteses sobre seu comportamento.

Os modelos de regressão espacial com efeitos globais partem do princípio de que o processo espacial subjacente aos dados analisados é estacionário. Isto implica que os padrões de autocorrelação espacial existentes nos dados podem ser capturados num único parâmetro. Na prática, para conjuntos de dados censitários de médio e grande porte, a natureza dos processos espaciais é tal que diversos padrões de associação espacial podem estar presentes. Esta hipótese, que pode ser verificada, por

exemplo, pelos indicadores locais de autocorrelação espacial, está na origem aos modelos cujos parâmetros variam no espaço, discutidos a seguir.

Modelos de Regressão com Efeitos Espaciais Locais

(a) *Caso Discreto – Modelos de Regressão com Regimes Espaciais*

Quando o processo espacial é não-estacionário, os coeficientes de regressão precisam refletir a heterogeneidade espacial. Para tanto, há duas grandes alternativas: (a) modelar a tendência espacial de forma contínua, com parâmetros variantes no espaço; (b) modelar a variação espacial de forma discreta, ao dividir o espaço em sub-regiões estacionárias, chamadas de *regimes espaciais*.

A idéia de *regimes espaciais* é dividir a região de estudo em sub-regiões, cada uma com seu padrão espacial próprio, e realizar regressões em separado, uma para cada região. As observações são classificadas em dois ou mais subconjuntos, a partir de uma variável por indicação, a saber:

$$Y_1 = X_1\beta_1 + \varepsilon_1 \text{, } ind = 1 \qquad (5.19.)$$

$$Y_2 = X_2\beta_2 + \varepsilon_2 \text{, } ind = 2 \qquad (5.20.)$$

Apesar de cada regime possuir os seus próprios valores de coeficientes, estes valores são estimados conjuntamente, ou seja, todo o conjunto de observações disponível é utilizado na regressão. Para a determinação dos regimes espaciais, as técnicas de análise exploratória apresentadas no início do capítulo são muito úteis, especialmente o mapa de espalhamento de Moran e os indicadores locais de autocorrelação espacial.

Na prática, para os dados sócio-econômicos típicos de cidades brasileiras, o modelo de regimes espaciais tende a apresentar resultados melhores que os modelos de regressão simples ou de regressão espacial com efeitos globais. Isto ocorre em função das fortes desigualdades sociais no Brasil, que ocasionam descontinuidades abruptas nos fenômenos estudados, como no caso do recorte entre favelas e áreas ricas, como é freqüente nas em nossas grandes cidades.

Modelos de Regressão com Efeitos Espaciais Locais

(b) *Modelos de Regressão com Efeitos espaciais contínuos*

Esta classe de modelos procura modelar fenômenos não-estacionários. Diferentemente do modelo por regimes espaciais, os efeitos espaciais são modelados de forma contínua, com duas hipóteses: (a) a existência de uma variação suave em larga escala, sem efeitos locais significativos ou (b) a existência de variações locais contínuas, sem uma forte tendência global. O primeiro caso corresponde às *superfícies de tendência*, descritas no capítulo 3 deste livro, resumidas no que segue para conveniência de leitura. O modelo

de *superfícies de tendência* considera um processo espacial onde o valor da variável é uma função polinomial de sua posição no espaço. O modelo de regressão múltipla utilizando notação vetorial é:

$$\boldsymbol{Y}(s) = \boldsymbol{X}(s)\beta + \varepsilon(s) \qquad (5.21.)$$

onde, $\boldsymbol{Y(s)} \rightarrow$ variável aleatória representando o processo no ponto $\boldsymbol{s}$,
$\boldsymbol{X(s)\beta} \rightarrow$ tendência (ou seja, o valor médio $\boldsymbol{\mu(s)}$),
$\boldsymbol{\varepsilon(s)} \rightarrow$ erro aleatório com média zero e variância $\boldsymbol{\sigma^2}$

O vetor $\boldsymbol{x(s)}$ consiste em $\boldsymbol{p}$ funções das coordenadas espaciais $\boldsymbol{(s_1, s_2)}$, do ponto amostrado $\boldsymbol{s}$. Para uma superfície de tendência linear é apenas $\boldsymbol{(1, s_1, s_2)}$, para quadrática é $\boldsymbol{(1, s_1, s_2, s_1^2, s_2^2, s_1.s_2)}$, e assim sucessivamente. $\boldsymbol{\beta}$ é o vetor $\boldsymbol{(p+1)}$ de parâmetros a ser ajustado. O pressuposto básico deste modelo supõe que os erros têm variância constante e são independentes em cada local, consequentemente, a covariância é zero: não há efeitos de segunda ordem presentes no processo. Neste contexto, é feito o ajuste do modelo por mínimos quadrados ordinários. O modelo de *superfícies de tendência* é útil sobretudo como uma primeira aproximação do fenômeno, pois na prática, são limitados os casos em que a variação espacial pode ser expressa desta forma. No entanto, os resíduos destes modelos são muito informativos sobre a natureza das variações locais.

No caso de *modelos de variações locais contínuas*, é idéia é ajustar um modelo de regressão a cada ponto observado, ponderando todas as demais observações como função da distância a este ponto. Desta forma, serão feitos tantos ajustes quantas observações existirem e o resultado será um conjunto de parâmetros, sendo que cada ponto considerado terá seus próprios coeficientes de ajuste. Estes parâmetros podem ser apresentados visualmente para identificar como se comportam espacialmente os relacionamentos entre variáveis. Esta técnica é denominada *geographically weighted regression* (GWR ou regressão ponderada espacialmente). Para aplicar o modelo GWR, o modelo padrão de regressão é reescrito na forma:

$$Y(s) = \beta(s)X + \varepsilon\,, \qquad (5.22.)$$

onde, $\boldsymbol{Y(s)}$ é a variável aleatória representando o processo no ponto $\boldsymbol{s}$, e $\boldsymbol{\beta(s)}$ indica que os parâmetros são estimados no ponto $\boldsymbol{s.}$ Para estimar os parâmetros deste modelo, a solução padrão por mínimos quadrados para o caso não-espacial, dada por

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \qquad (5.23.)$$

é generalizada usando um método de ajuste local:

$$\beta(s) = (\boldsymbol{X}^T W(s)\,\boldsymbol{X})^{-1}\boldsymbol{X}^T W(s)\boldsymbol{Y} \qquad (5.24.)$$

O ajuste local é feito de forma a garantir uma influência maior dos pontos mais próximos, de forma semelhante aos estimadores de densidade por *kernel*, discutidos no capítulo 2 do livro. Um exemplo é o uso de uma função gaussiana, do tipo

$$w_{ij}(s,\tau)=\frac{1}{2\pi\tau}\exp\left(-\frac{{d_{ij}}^2}{2\tau^2}\right) \qquad (5.25.)$$

onde τ representa o raio de influência considerado, e d_{ij} a distância entre a localização considerada e o *j*-ésimo ponto. Pode-se fazer testes de hipóteses para verificar se as variações espaciais têm significado estatístico ou são aleatórias. Para maiores detalhes sobre o modelo GWR, o leitor deve referir-se à bibliografia no final do capítulo.

Diagnóstico de Modelos com Efeitos Espaciais

A análise gráfica dos resíduos é o primeiro passo para avaliar a qualidade do ajuste da regressão. Mapear os resíduos é uma etapa importante no diagnóstico do modelo, buscando indícios de ruptura dos pressupostos de independência. Uma alta concentração de resíduos positivos (ou negativos) numa parte do mapa é um bom indicador da presença de autocorrelação espacial. Para um teste quantitativo, o mais comum é utilizar o índice I de Moran sobre os resíduos.

Como os estimadores e os diagnósticos tradicionais de regressão não levam em conta os efeitos espaciais, as inferências, como por exemplo as indicações de qualidade de ajuste baseadas em R^2(coeficiente de determinação), serão incorretas. Estas conseqüências são similares às que acontecem quando uma variável explicativa significativa é omitida do modelo de regressão. Quando se quer comparar um ajuste obtido por um modelo de regressão padrão, com um ajuste obtido por um dos modelos cuja especificação considera a autocorrelação espacial, uma medida como o R^2 não é mais confiável.

O método mais usual de seleção de modelos de regressão baseia-se nos valores de *máxima verossimilhança* dos diferentes modelos, ponderando pela diferença no número de parâmetros estimados. Nos modelos com estrutura de dependência – espacial ou temporal - utilizam-se os *critérios de informação* onde a avaliação do ajuste é penalizada por uma função do número de parâmetros. Cabe observar que é necessário ainda levar em conta o número de parâmetros independentes ao se incluir funções espaciais nos modelos. Para cada nova variável em modelo de regressão, acrescenta-se um parâmetro.

Usualmente a comparação de modelos é feita utilizando o logaritmo da máxima verossimilhança, que é o que possui melhor ajuste para os dados observados. O critério de informação de Akaike (AIC) é expresso por:

$$AIC = -2 * LIK + 2k \qquad (5.\ 26.)$$

onde LIK é o log de verossimilhança maximizado e *k* é o número de coeficientes de regressão. Segundo este critério, o melhor modelo é o que possui menor valor de AIC. Diversos outros critérios de informação estão disponíveis, a maior parte dos quais são variações do AIC, com mudanças na forma de penalização de parâmetros ou observações.

Exemplo Ilustrativo

Como exemplo ilustrativo das técnicas de regressão espacial, estudou-se o relacionamento entre renda e longevidade na cidade de São Paulo, para os dados do Censo de 1991. Tratam-se de duas das três variáveis utilizadas para compor o IDH (índice de desenvolvimento humano) da ONU. A variável dependente a ser explicada é denotada por PERIDOSO (percentual de pessoas com mais de 70 anos por distrito de São Paulo) e a variável independente é indicada por PERREN20 (percentual de chefes de família com renda de mais de 20 salários mínimos mensais). A distribuição espacial destas variáveis está mostrados na Figura 5-21.

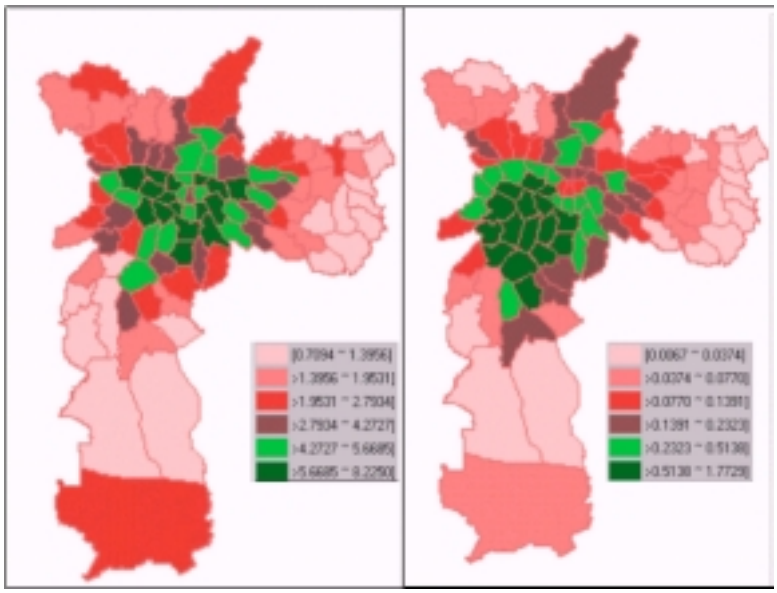


Figura 5-21. Percentual de idosos (à esquerda) e de chefes de família com renda maior que 20 SM mensais (à direita) para os distritos de São Paulo (1991).

Foram comparados três modelos de regressão: o modelo padrão não-espacial, o modelo autoregressivo (*spatial lag*) e o modelo em regimes espaciais. No caso dos regimes espaciais foram consideradas três regiões da

cidade (centro, periferia e a transição centro-periferia). O modelo padrão é expresso como:

$$PERIDOSO = \beta_0 + \beta_1 PERREN20 + \varepsilon \qquad (5.\ 27.)$$

Utilizando-se a matriz de vizinhança W dos distritos, o modelo "spatial lag" pode ser expresso como:

$$PERIDOSO = \beta_0 + \beta_1 PERREN20 + \rho W(PERIDOSO) + \varepsilon \qquad (5.\ 28.)$$

Considerando-se três regiões da cidade, o modelo de regimes espaciais pode ser expresso como

$$PERIDOSO\_1 = \beta^1_0 + \beta^1_1 PERREN20\_1,\ reg = 1 \qquad (5.\ 29.)$$

$$PERIDOSO\_2 = \beta^2_0 + \beta^2_1 PERREN20\_2,\ reg = 2 \qquad (5.\ 30.)$$

$$PERIDOSO\_3 = \beta^3_0 + \beta^3_1 PERREN20\_3,\ reg = 3 \qquad (5.\ 31.)$$

Os resultados destes modelos de regressão são apresentados na Tabela 5-3. No modelo de regressão tradicional, a relação entre renda e longevidade em São Paulo é muito reduzida, o que dá suporte a idéia do IDH de que tratam-se de dimensões complementares da desenvolvimento humano. No entanto, quando os efeitos espaciais são levados em conta, verifica-se que a existência de real dependência entre os dois fatores. Na Figura 5-22, apresenta-se a distribuição espacial dos resíduos da regressão para os modelos de mínimos quadrados e *spatial lag*. Uma análise visual dos resíduos da regressão tradicional indica uma prevalência de resíduos positivos no centro da cidade e resíduos negativos na periferia, principalmente nas Zonas Leste e Sul. Os resultados numéricos confirmam esta análise, pois o índice de Moran dos resíduos é altamente significativo. Com relação ao desempenho global, as medidas R^2 são indicadores limitados e devem ser encaradas com cuidados, e deve-se preferir as medidas baseadas em verossimilhança (LIK, AIC). Neste caso, o modelo *spatial lag* teve um desempenho muito superior ao modelo padrão. Este efeito é esperado, pela existência de um índice de Moran significativo nos resíduos, que é capturado no coeficiente de efeito espacial (ρ).

Os regimes espaciais escolhidos para São Paulo são mostrados na Figura 5-23, bem como os resíduos da regressão considerando estes regimes. Da análise visual dos resíduos, verifica-se a não-existência de forte tendência espacial, o que é evidenciado pelo baixo índice de Moran dos mesmos, indicado na Tabela 5-3. No geral, o modelo de regimes espaciais apresentou o melhor desempenho, por qualquer dos critérios (R^2, LIK e AIC). O resultado reflete a forte polarização centro-periferia da cidade de São Paulo, e é compatível com estudos que mostram os resultados da violência urbana nas taxas de mortalidade, especialmente de homens dos 15 aos 25 anos.

Tabela 5-3

Resultados da Regressão para Longevidade e Renda em São Paulo, 1991

	Regressão MMQ	Spatial Lag	Regimes Espaciais
R^2 ajustado	0,280	0,586	0,80
Log verossimilhança	-187,92	-150,02	-124,04
AIC (Críterio de Inf. Akaike)	379,84	306,51	260,09
Índice de Moran dos resíduos	0,620	-	0,020

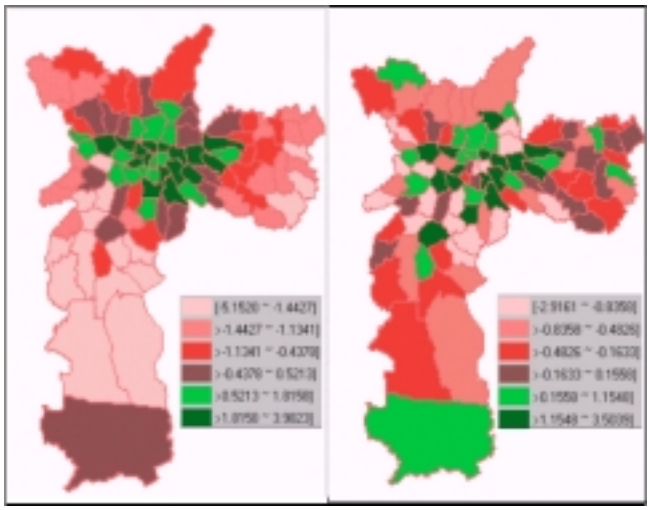


Figura 5-22- Resíduos da regressão por mínimos quadrados (à esquerda) e resíduos da regressão com o modelo *spatial lag* (à direita).

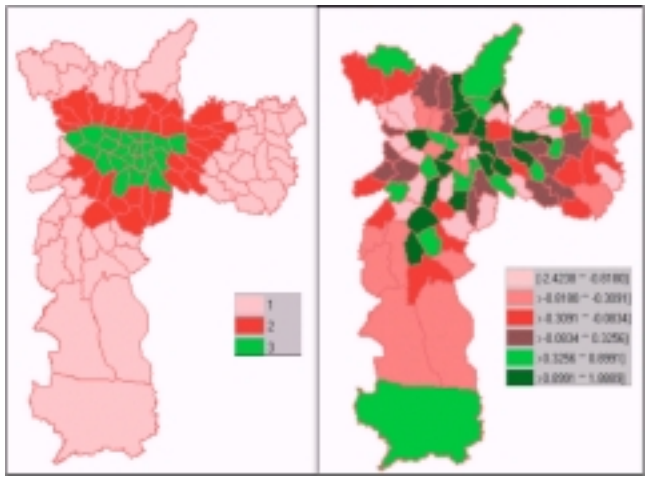


Figura 5-23 Regimes espaciais para os distritos de São Paulo (à esquerda) e resíduos da regressão por regimes espaciais (à esquerda).

5.7 ESTIMAÇÃO DE MODELOS CONTÍNUOS A PARTIR DE DADOS DE ÁREA

As seções anteriores apresentaram técnicas de análise espacial de dados de área tomando por base o modelo de *variação espacial discreta*, onde cada área é modelada respeitando seus limites, adjacências e vizinhança. Nesta seção, considera-se o modelo de *variação espacial contínua*, que supõe um processo estocástico $\{Z(x), x \in A, A \subset \Re^2\}$, cujos valores podem ser conhecidos em todos os pontos da área de estudo. A idéia de modelos contínuos para dados socioeconômicos decorre do fato que os levantamentos censitários muitas vezes impõem limites de áreas a partir de critérios puramente operacionais, que não têm relação direta com o fenômeno modelado. Este fato leva à idéia de dissolver os limites das áreas em superfícies contínuas, de forma a modelar melhor a real continuidade de, por exemplo, setores censitários em regiões urbanas densamente povoadas.

No caso de estimadores de superfícies, as principais alternativas são o uso de técnicas não-paramétricas e o uso de interpoladores geoestatísticos, descritos nos capítulos 3 deste livro e que são brevemente resumidos no que segue.

Estimador de Intensidade Não-Paramétrico

De forma similar como no caso de superfícies, podemos utilizar o estimador de intensidade (*kernel estimator*) para nos fornecer uma primeira aproximação da distribuição espacial do fenômeno ou variável. Neste caso, quando os valores observados representam uma medida "média" como taxa de mortalidade ou renda per capita, podemos utilizar um estimador que nos permitiria calcular o valor do atributo por unidade de área. Para toda posição ***(x;y)*** cujo valor queremos estimar, o estimador de intensidade será computado a partir dos valores $\{z_1,...,z_n\}$ contidos num raio de tamanho τ, a partir da equação

$$\hat{z}_i = \frac{\sum_{j=1}^{n} k(\frac{d_{ij}}{\tau}) z_j}{\sum_{j=1}^{n} k(\frac{d_{ij}}{\tau})}, \; d_{ij} \leq \tau \qquad (5.\ 32.)$$

Na equação acima, a função $\kappa()$ é um interpolador não-paramétrico, que pode ser, por exemplo, um *kernel* gaussiano, como apresentado nos capítulos 2 e 3 deste livro, onde o leitor poderá encontrar uma discussão mais aprofundada sobre os estimadores de intensidade não-paramétricos. Um exemplo do estimador de intensidade para taxas pode ser visto na Figura 5-22, onde são apresentados os dados de mortalidade por homicídios para o Estado do Rio de Janeiro, para o triênio 90-92 interpolados pelo estimador

de intensidade, que nos dá uma idéia da distribuição espacial da variável estudada. Na Figura 5-24(a) é apresentado um mapa com os valores de indicadores de taxa de mortalidade, agregados por município. Na Figura 5-24(b), apresentamos o resultado do estimador de intensidade, que nos dá uma idéia melhor da distribuição espacial da variável estudada.

Quando as observações nas áreas representam contagens, como as obtidas pelo censo, o estimador de kernel apresentado acima não é apropriado. Um valor "médio" de um atributo como "número de domicílios precários" não faria sentido, e deve-se pensar em termos de "número de domicílios precários por unidade de área". Neste caso, pode-se utilizar o numerador da equação (5.32), dividido pela área do círculo definido pelo raio de busca:

$$\hat{z}_i = \frac{1}{\pi\tau^2}\sum_{j=1}^{n} k(\frac{d_{ij}}{\tau}) z_j \ , \ d_{ij} \ \leq \tau \qquad (5.33.)$$

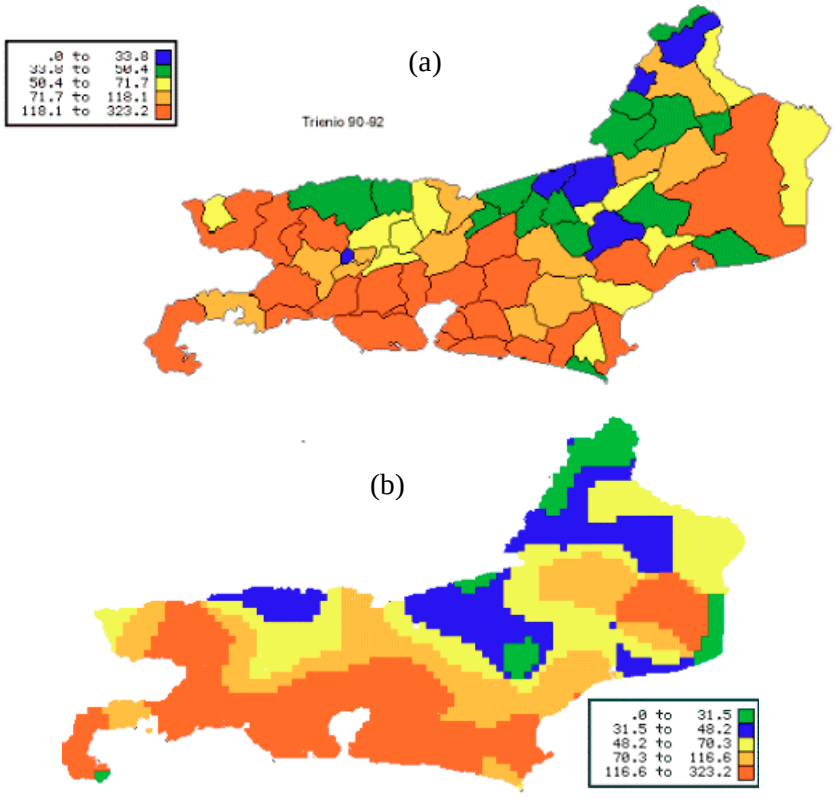


Figura 5-24 (a) Mortalidade por homicídios no RJ (1990-1992). Mapa temático com valores por município. (b) Superfície obtida por estimador de intensidade não-paramétrico

Uso de Interpoladores Geoestatísticos

No capítulo 3, apresenta-se a teoria básica da geoestatística, cuja motivação tradicional está associada a dados do meio físico como medidas de teor mineral ou de poluição. No caso da krigagem ordinária, a hipótese subjacente é que os dados apresentam distribuição gaussiana, e neste caso as propriedades ótimas dos estimadores (como a mínima variância do resultado) são garantidas. Para o caso de dados socioeconômicos ou de saúde coletiva, a hipótese da normalidade dos dados muito raramente é realista, sendo mais comum supor uma distribuição de Poisson, por se tratar de contagens de eventos. No entanto, as propriedades ótimas do estimador de krigagem e sua ampla disponibilidade em diferentes sistemas de informação geográfica fazem com que seja importante investigar seu uso para dados socioeconômicos. Neste caso, a primeira providência é investigar quão aproximados da distribuição normal se apresentam os dados; se for necessário, pode-se aplicar transformações apropriadas (com a transformação logarítmica) para "simetrizar" a distribuição empírica e assim aproximar-se da distribuição normal. Para considerar uma situação concreta, Figura 5-25 apresenta a distribuição da taxa de homicídios por 100 mil habitantes, para os 96 distritos de São Paulo em 1996, acompanhada do gráfico de probabilidade normal, que indica o quanto estes dados se aproximam de uma distribuição gaussiana. Da análise dos dois dados, e considerando-se ainda que a média (43,6) é suficientemente próxima da mediana (39,3), e como o teste de normalidade de Shapiro-Wilk indica um valor de 0,9653 (p-valor de 0,012), a hipótese de normalidade não pode ser rejeitada e permite aplicar uma interpolador de krigagem.

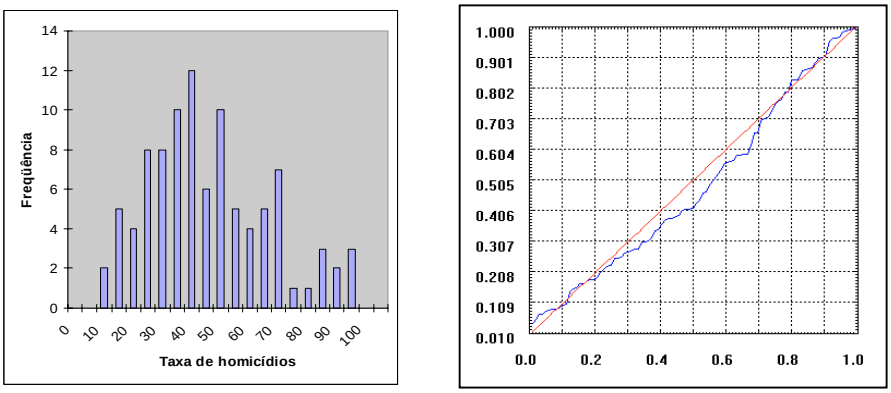


Figura 5-25. Distribuição da taxa de homicídios por 100 mil habitantes para São Paulo em 1996. À direita: frequência relativa; à esquerda: gráfico de probabilidade normal.

Com base nestas hipóteses, e com o objetivo de entender os padrões espaço-temporais em São Paulo, utilizou-se a krigagem ordinária para produzir superfícies das taxas de homicídio para os 96 distritos de São Paulo para os anos de 1996 e 1999 (a distribuição de taxas de 1999 apresentou padrões semelhantes que a de 1996). Para tal, o conjunto de pontos obtido pela associação do valor do parâmentro de cada área, ao seu centróide, foi tomado como uma amostra, usada para computar um variograma que modelou a estrutura de correlação espacial. A superfície obtida está apresentada na Figura 5-26 e mostra uma queda significativa nas áreas com as menores taxas de homicídios (menos que 30 mortes por 100,000 pessoas) em 1999 com relação a 1996. Como as áreas de menor taxa de homicídio correspondem às áreas mais ricas da cidade (compare com as figuras 5.1), o resultado mostra um espalhamento espacial do crime, com a violência ocupando progressivamente toda a cidade.

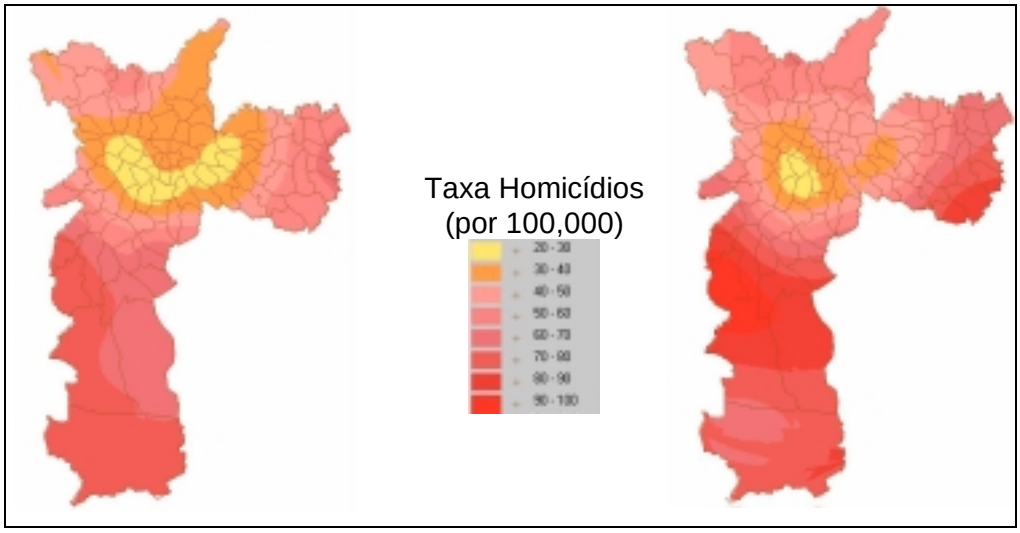


Figura 5-26. Superfícies estimadas para as taxas de homicídio em São Paulo em 1996 (esquerda) e 1999 (direita).

5.8 COMENTÁRIOS FINAIS

Este capítulo mostrou que as técnicas de análise espacial podem ampliar consideravelmente a capacidade de compreender os padrões espaciais associados a dados de área, especialmente quando se trata de indicadores sociais, que apresentam autocorrelação espacial global e local. Técnicas exploratórias como os indicadores de Moran e os mapas de espalhamento de Moran são muito úteis para mostrar as agregações espaciais e indicar áreas prioritárias em termos de política pública. Métodos de estimação bayesiana

para taxas permitem a correção de efeitos associados a pequenas populações. Modelos de regressão espacial permitem estabelecer as relações entre as variáveis, levando em conta os efeitos espaciais; neste caso, o poder explicativo dos modelos pode ter ganhos significativos. A geração de superfícies é um maneira eficiente de apreensão visual dos padrões espaciais. Em resumo, estudiosos de dados sócio-econômicos podem se beneficiar substancialmente das técnicas deste capítulo.

5.9 REFERÊNCIAS

A referência básica para a maior parte das técnicas apresentadas neste capítulo é o livro de Trevor Bailey, "*Spatial Data Analysis by Example*" (Bailey and Gattrel, 1995) e uma discussão geral sobre os modelos de distribuição para dados espaciais é apresentada em Diggle (2001). A homepage de Peter Diggle (www.maths.lancs.ac.uk/~diggle) contém material relevante sobre estatística espacial.

No caso dos modelos de regressão espacial, o software SpaceStat de Luc Anselin, e a documentação associada (Anselin, 1992) apresenta em detalhe os modelos de regressão com efeitos globais (*spatial lag e spatial error)*, e o modelo de regimes espaciais. O SpaceStat foi utilizado para computar os modelos no exemplo apresentado no capítulo. Os trabalhos de Luc Anselin no campo de indicadores locais de autocorrelação espacial (Anselin, 1995; Anselin, 1996) também são referências importantes. O sítio do SpaceStat é www.spacestat.com.

O modelo de regressão GWR (*geographically weighted regression)* foi idealizado por A.Stewart Fotheringham, e está descrito em seu livro *Quantitative Geography* (Fotheringham et al., 2000) e outros trabalhos (Fotheringham et al., 1996) (Brunsdon et al., 1996). Maiores informações podem ser encontradas no sítio http://www.ncl.ac.uk/~ngeog/GWR/.

A discussão sobre o problema dos efeitos de escala e a chamada "falácia ecológica" deve muito aos trabalhos de Stan Openshaw; como exemplo, veja-se Openshaw (1997). Seus trabalho sobre o uso de técnicas de otimização combinatória para obter regiões mais agregadas, também são muito importantes (Openshaw and Alvanides, 1999).

A questão da geração de superfícies a partir de dados socioeconomicos deve muito aos trabalhos de David Martin, em seu livro "*Geographic Information Systems: Socioeconomic Applications*" (Martin, 1995) e seus trabalhos sobre os dados censitários no Reino Unido (Martin, 1996; Martin, 1998).

Os estimadores bayesianos empíricos foram inicialmente propostos em (Marshall, 1991). Uma discussão geral sobre o assunto, incluindo uma discussão sobre os estimadores bayesianos completos, pode ser encontrada no excelente trabalho de Renato Assunção (Assunção, 2001) ou na revisão abrangente de Trevor Bailey, publicada nos Cadernos de Saúde Pública (Bailey, 2001).

Os dados de São Paulo do censo de 1991 foram extraídos do trabalho "Mapa de Exclusão/Inclusão Social na Cidade de São Paulo", coordenado pela prof. Aldaíza Sposati, da PUC/SP (Sposati, 1996). As taxas de homicídio para os distritos de São Paulo em 1996 e 1999 foram produzidas pela Fundação SEADE e a geração de superfícies por krigeagem foi feita por José Luiz Rodriguez Yi.

Os dados do censo de Belo Horizonte para o ano de 1991 foram cedidos pela PRODABEL, e o estudo do problema das mudanças de unidade de análise foi realizado por Taciana Dias e Maria Piedade Oliveira.

Os dados de mortalidade infantil para a cidade do Rio de Janeiro foram organizados pela FIOCRUZ e estão apresentados no trabalho de Eleonora D'Orsi e Marilia Carvalho (D'Órsi & Carvalho, 1998). Os dados do estudo sobre mortalidade por homicídios na Região Sudeste também foram publicados pela equipe da FIOCRUZ, e podem ser acessados nas páginas pessoais dos autores: http://www.procc.fiocruz.br/~marilia/ e www.procc.fiocruz.br/~oswaldo/.

O número especial dos Cadernos de Saúde Pública sobre o tema de estatísticas espaciais em saúde (volume 17(5), outubro-novembro 2001), disponível na Internet (www.scielo.br) representa um bom ponto de partida sobre o tema, com vários estudos relevantes.

1. ANSELIN, L. **SpaceStat tutorial: a workbook for using SpaceStat in the analysis of spatial data**. Santa Barbara, NCGIA (National Center for Geographic Information and Analysis), 1992.
2. ANSELIN, L. Local indicators of spatial association - LISA. **Geographical Analysis** v.27, p.91-115, 1995.
3. ANSELIN, L. The Moran scatterplot as ESDA tool to assess local instability in spatial association. In: M. Fisher, H. J. Scholten and D. Unwin (ed). **Spatial Analytical Perspectives on GIS**. London, Taylor & Francis, 1996. v., p.111-126.
4. ASSUNÇÃO, R. **Estatística Espacial com Aplicações em Epidemiologia, Economia e Sociologia**. São Carlos, SP, UFScar, 2001. Disponível na homepage www.est.ufmg.br/~assuncao.

5. BAILEY, T. Spatial Statistics Methods in Health. **Cadernos de Saúde Pública** v.17, n.5,, 2001.

6. BAILEY, T. and A. GATTREL. **Spatial Data Analysis by Example**. London, Longman, 1995.

7. BRUNSDON, C. A.S. FOTHERINGHAM AND M.E. CHARLTON, Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. **Geographical Analysis**, 28(4), 281-298, 1996.

8. CRUZ, O. C. **Homicídios no Estado do Rio de Janeiro: análise da distribuição espacial e sua evolução**. Dissertação de mestrado/Faculdade de saúde Pública-USP, 1996.
http://malaria.procc.fiocruz.br/~oswaldo/publi/ogc-diss.pdf

9. DIGGLE, P. **Spatial statistics in the biomedical science: future directions**. Lancaster, Lancaster University, 2001.

10. D'ÓRSI, E. and M. S. CARVALHO. Perfil de Nascimentos no Município do Rio de Janeiro - Uma Análise Espacial. **Cadernos de Saúde Pública** v.14, n.1, p.367-379, 1998.

11. FOTHERINGHAM, A.S., C. BRUNSDON AND M.E. CHARLTON, 2000, **Quantitative Geography**, London: Sage

12. FOTHERINGHAM, A.S., M.E. CHARLTON AND C. BRUNSDON, The Geography of Parameter Space: An Investigation into Spatial Non-Stationarity. **International Journal of Geographic Information Systems**, 10: 605-627, 1996.

13. GELMAN, A., CARLIN, J.B., STERN, H.S., RUBIN, D.B. (1995) **Bayesian Data Analysis** Chapman & Hall/CRC.

14. GILKS, W.R., RICHARDSON, S., SPIEGELHALTER, D.J. (orgs) (1998), *Markov Chain Monte Carlo in Practice*, Chapman & Hall.

15. MARSHALL, R. Mapping disease and mortality rates using empirical Bayes estimators. **Applied Statistics** v.40, p.283-294, 1991.

16. MARTIN, D. **Geographic Information Systems: Socioeconomic Applications.** London, Routledge, 1995.

17. MARTIN, D. An assessment of surface and zonal models of population. **International Journal of Geographical Information Systems** v.10, p.973-989, 1996.

18. MARTIN, D. Optimizing census geography: the separation of collection and output geographies. **International Journal of Geographical Information Science** v.12, p.673-685, 1998.

19. OPENSHAW, S. Developing GIS-relevant zone-based spatial analysis methods. In: P. Longley and M. Batty (ed). **Spatial Analysis: Modelling in a GIS Environment**. New York, John Wiley, 1997. v., p.55-73.

20. OPENSHAW, S. and S. ALVANIDES. Applying Geocomputation to the analysis of spatial distributions. In: P. A. Longley, Goodchild, M. F., Maguire, D. J. and Rhind, D. W (ed). **Geographical Information Systems: Principles, Techniques, Management and Applications**. Chichester, Wiley, 1999. v., p.267-282.

21. SPOSATI, A. **Mapa de Exclusão/Inclusão Social de São Paulo**. São Paulo, EDUC, 1996.