Problemas de Escala e a Relação Área-Indivíduo em Análise Espacial de Dados Censitários

Taciana de Lemos Dias[1]
Analista de Sistemas da Empresa de Informática e Informação do Município de Belo Horizonte - PRODABEL
Doutoranda em Computação Aplicada do Instituto Nacional de Pesquisas Espaciais – INPE
Áreas de interesse: Modelos para representação espaço-temporais urbanos, ontologias, banco de dados, gestão de informação, análise espacial e geoprocessamento

Maria da Piedade Gomes de Oliveira[2]
Analista de Sistemas da Empresa de Informática e Informação do Município de Belo Horizonte - PRODABEL
Doutoranda em Computação Aplicada do Instituto Nacional de Pesquisas Espaciais – INPE
Áreas de interesse: Análise espacial, geoestatística, mineração de dados espaciais, ontologias e geoprocessamento

Gilberto Câmara[3]
Coordenador Geral de Observação da Terra do INPE – Instituto Nacional de Pesquisas Espaciais
Doutor em Computação Aplicada pelo INPE
Professor do Curso de Pós-Graduação em Computação Aplicada do INPE
Áreas de interesse: Tecnologia de Sistemas de Informação Geográfica, bancos de dados geográficos, análise espacial e estatística espacial, modelagem espaço-temporal de informação e processamento de imagens de sensores remotos

Marilia Sá Carvalho[4]
Pesquisadora Titular da Escola Nacional de Saúde Pública- ENSP e Fundação Oswaldo Cruz – FIOCRUZ
Doutora em Engenharia Biomédica, COPPE/UFRJ
Pós-doutorado em Estatística na Universidade de Lancaster/Reino Unido
Áreas de interesse: Métodos de análise de dados espaciais e modelagem estatística de dados dependentes em saúde pública e epidemiologia

Palavras-Chave

Análise espacial – Geoestatística – Dados censitários – SIG – Sistemas de Informações Geográficas

[1] E-mail:taciana@pbh.gov.br
[2] E-mail:mpiedade@pbh.gov.br
[3] E-mail:gilberto@dpi.inpe.br
[4] E-mail:carvalho@procc.fiocruz.br

RESUMO

Este artigo apresenta os problemas relacionados com a manipulação de dados agregados por área e sua interpretação em diferentes subdivisões de unidades de áreas. A granularidade da subdivisão territorial interfere nos resultados, podendo gerar conclusões impróprias sobre o fenômeno estudado. Assuntos relevantes para a análise desses dados, tais como agregação e zoneamento, além de estimativas de taxas em áreas de pequenas populações, são discutidos através de exemplos. Soluções no campo da análise espacial são propostas para reduzir as distorções causadas pela agregação dos dados em áreas.

1. INTRODUÇÃO

Compreender a distribuição espacial de fenômenos constitui hoje um grande desafio para a elucidação de questões centrais em diversas áreas do conhecimento, tais como saúde, meio ambiente, geologia, agronomia, e várias entre tantas outras. Tais estudos vêm se tornando cada vez mais comuns devido à crescente democratização das informações, à evolução e redução dos custos das tecnologias e à difusão de Sistemas de Informação Geográfica (SIG) com interfaces amigáveis. As informações estão mais facilmente acessíveis devido aos avanços tecnológicos como Internet, redes e meios de armazenamento com maior capacidade.

Os SIG permitem a apresentação espacial de variáveis como população de indivíduos, índices de qualidade de vida e vendas de empresas numa região, através de mapas. Para tanto, basta dispor de um banco de dados e de uma base geográfica contendo alguma forma de divisão territorial em unidades espaciais de referência (como um mapa de municípios), e qualquer SIG torna-se capaz de apresentar um mapa colorido (coroplético) que permite a visualização do padrão espacial do fenômeno. Esses mapas são construídos através de valores que correspondem a uma combinação de propriedades das áreas geográficas ou que consideram uma propriedade específica à qual é associada uma cor [LGMR01].

Além da percepção visual da distribuição espacial do problema, é muito útil traduzir os padrões existentes com considerações objetivas e mensuráveis, como nos seguintes casos:

- Epidemiologistas coletam dados sobre ocorrências de doenças. A distribuição dos casos de uma doença forma um padrão no espaço? Existe associação com alguma fonte de poluição? Existe evidência de contágio? Houve variação no tempo?
- Policiais desejam investigar se existe alguma concentração espacial na distribuição de crimes. Roubos que ocorrem em determinadas áreas estão correlacionados com características socioeconômicas dessas áreas?

- Geólogos desejam estimar a extensão de um depósito mineral em uma região a partir de amostras. Pode-se usar essas amostras para estimar a distribuição do mineral na região?
- Planejadores desejam analisar uma região para fins de zoneamento agrícola. Como escolher as variáveis explicativas – solo, vegetação, geomorfologia – e determinar qual a contribuição de cada uma delas para definir em que local o tipo de cultura é mais adequado?

Todos esses problemas fazem parte da *análise espacial de dados geográficos*. A ênfase da análise espacial está em mensurar propriedades e relacionamentos, levando em conta a localização espacial do fenômeno em estudo de forma explícita. Ou seja, a idéia central é incorporar o espaço à análise que se deseja fazer, levando-se em consideração "a primeira lei da geografia" de Waldo Tobler [LGMR01]: *"todas as coisas são parecidas mas coisas mais próximas se parecem mais que coisas mais distantes"*.

A taxonomia mais utilizada [BaGa95] para caracterizar os problemas de análise espacial considera três tipos de dados:

- *Eventos ou Padrões Pontuais* - 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.
- *Superfícies Contínuas* - estimadas a partir de um conjunto de amostras de campo, que podem estar regular ou irregularmente distribuídas. Usualmente, este tipo de dado é resultante de levantamento de recursos naturais, e que incluem mapas geológicos, topográficos, ecológicos, fitogeográficos e pedológicos.
- *Áreas com Contagens e Taxas Agregadas* - trata-se de dados associados a levantamentos populacionais, como censos e estatísticas de saúde, e que originalmente se referem a indivíduos localizados em pontos específicos do espaço. Esses dados são agregados em unidades de análise, usualmente delimitadas por polígonos fechados (setores censitários, distritos censitários e municípios).

As origens dos dados geralmente utilizados em *análise de áreas* são, em grande parte, oriundas de levantamentos realizados por órgãos públicos, tais como os populacionais do censo, os estatísticos de saúde e cadastramento de imóveis dos municípios. Essas áreas usualmente possuem uma delimitação onde se supõe haver homogeneidade interna, ou seja, as áreas são compostas de agrupamentos aleatórios de indivíduos/eventos/moradias que tendem a ser semelhantes em relação a outras áreas. A probabilidade dessa semelhança pode ocorrer, por exemplo, nas características socioeconômicas, demográficas, de saúde e morfologia do solo [WHST96]. Evidentemente, esta premissa nem sempre é verdadeira e não há qualquer garantia de que a distribuição do evento seja homogênea dentro dessas unida-

des, visto que freqüentemente as unidades de levantamento são definidas por critérios operacionais (setores censitários), políticos (municípios), ou podem refletir o modo com que os cartógrafos ou ferramentas de SIG interpolam um limite entre pontos amostrais, como na criação de mapas isopléticos.

No caso de áreas, deve-se ainda considerar que, em países com grandes contrastes sociais como o Brasil, é freqüente que estejam agregados em uma mesma região de coleta grupos sociais distintos – favelas e áreas nobres –, resultando em indicadores calculados que representam a média entre populações diferentes. Adicionalmente, em diversas regiões, as unidades amostrais apresentam diferenças importantes em população e área [Mart95]. Neste caso, tanto a apresentação em mapas coropléticos quanto os cálculos simples de taxas populacionais podem levar a distorções nos indicadores obtidos, e será preciso utilizar técnicas de ajuste de distribuições. O inverso ocorre em áreas com pequenas populações.

Este artigo apresenta um conjunto de procedimentos para responder a esses desafios. Pretende-se auxiliar os interessados a estudar, explorar e modelar processos que se expressam através de uma distribuição no espaço, aqui chamados de *fenômenos geográficos*.

2. Efeitos de Escala na Análise de Dados de Área

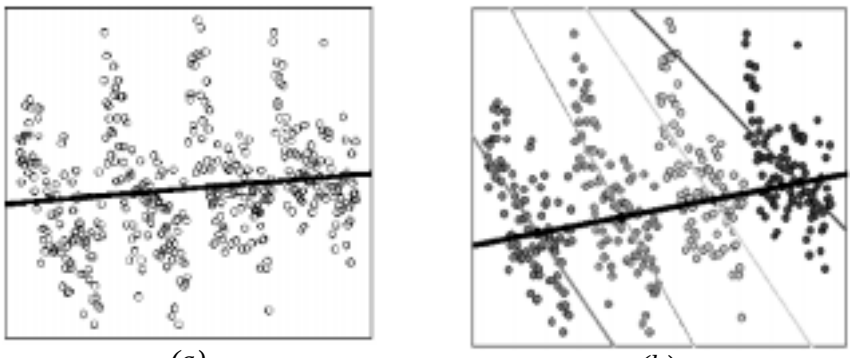
Em muitos dos estudos envolvendo dados de área, existe a necessidade de preservar o que há de confidencioso nos registros individuais. Os processos de disseminação de dados são projetados para evitar que informações que possibilitem a identificação dos indivíduos sejam disponibilizadas. E a alternativa disponível para essa preservação é a agregação geográfica [Mart00]. Isso ocorre no caso do Censo, onde os dados já agregados por setores censitários são o menor nível de agregação a que a comunidade em geral tem acesso para vários tipos de análises. Alguns desses estudos procuram estabelecer relações de causa-efeito entre diferentes medidas com o uso de modelos de regressão; um exemplo clássico é correlacionar anos de estudo do chefe de família e sua renda, dados que usualmente apresentam forte correlação. Um setor censitário, no Brasil, corresponde à capacidade de levantamento do recenseador, variando por região em torno de 200 a 400 domicílios.

Um dos problemas básicos apresentados em 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 segundo as quais essas unidades podem ser agrupadas. Pode-se obter resultados diferentes simplesmente alterando as fronteiras entre essas áreas. Este problema é conhecido como "problema da unidade de área modificável" (*Modifiable Areal Unit Problem*- MAUP) [FBC00,LoBa96]. Por exemplo, Openshaw e Taylor, em [OpWy97], descrevem como obter correlações completamente diferentes entre comportamento eleitoral e

idade no Estado americano de Iowa, apenas modificando a agregação de seus condados.

Devido aos efeitos de escala e de agregação de áreas, os coeficientes de correlação podem ser inteiramente diferentes no nível individual do nível de áreas. O *efeito de escala* é a tendência, dentro de um sistema de unidades de área modificáveis, de prover resultados estatísticos diferentes para o mesmo conjunto de dados quando a informação se agrupa em níveis diferentes de resolução espacial (por exemplo, setores censitários, unidades de planejamento, bairros, distritos e regiões) [WHST96]. O conceito de escala, neste trabalho, não corresponde à noção tradicionalmente usada em cartografia, e sim a diferentes níveis de resolução espacial**.** Este fenômeno, nas ciências sociais e na epidemiologia, é denominado *falácia ecológica*, envolvendo conclusões impróprias em nível individual a partir de resultados agregados por unidades de área [StHo96]. Sendo assim, os resultados estatísticos têm validade dependente da unidade de área e do reconhecimento dos problemas existentes nas conclusões decorrentes de dados agregados. Deve-se observar que a chamada *falácia ecológica*, a rigor, nem é uma falácia nem é ecológica. Trata-se de uma propriedade inerente aos dados agregados por áreas**.**

A agregação de indivíduos em áreas tende a aumentar a correlação entre as variáveis e a reduzir as flutuações estatísticas. Por exemplo, considere-se um conjunto de indivíduos, dos quais são medidas duas características, conforme indicado na Figura 1 (a). Uma regressão linear considerando todos os indivíduos (linha negra do diagrama à esquerda) resulta em um coeficiente positivo de 0,1469. Esses indivíduos pertencem a grupos distintos indicados pela tonalidade do ponto no diagrama da direita. Com isso, passa-se a obter correlação negativa, variando entre – 0,5 e – 0,8. Utilizando as médias de cada grupo (linha negra do diagrama à direita), o coeficiente vai a 0,99. No primeiro caso, pode-se dizer que sem informações que permitam separar os indivíduos nos grupos, as variáveis se relacionam positivamente. No segundo, 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.



(a) *(b)*

Figura 1 – Modelos de regressão: (a) indivíduos, (b) indivíduos em estratos e grupos diferentes

Para ilustrar os efeitos de escala em unidades de área, tomaram-se os dados oficiais do Censo 1991 em Belo Horizonte, em duas escalas: os setores censitários e as unidades de planejamento (UPs), mostradas na Figura 2. Os setores censitários foram utilizados pelo IBGE para o Censo de 1991, totalizando 1998 setores, e as unidades de planejamento correspondem aos agregamentos de áreas utilizados pela Prefeitura de Belo Horizonte. As UPs são 81 divisões político-administrativas do município, adotadas para os estudos básicos do Plano Diretor de BH em 1995. Os limites de cada UP foram definidos considerando: divisão limites das Regiões Administrativas da PBH; grandes barreiras físicas, naturais ou construídas; continuidade de ocupação; padrão de ocupação. Os grandes aglomerados de favelas e conjuntos habitacionais de BH foram considerados unidades independentes. As favelas menores, incorporadas às UPs próximas [SMPL96].

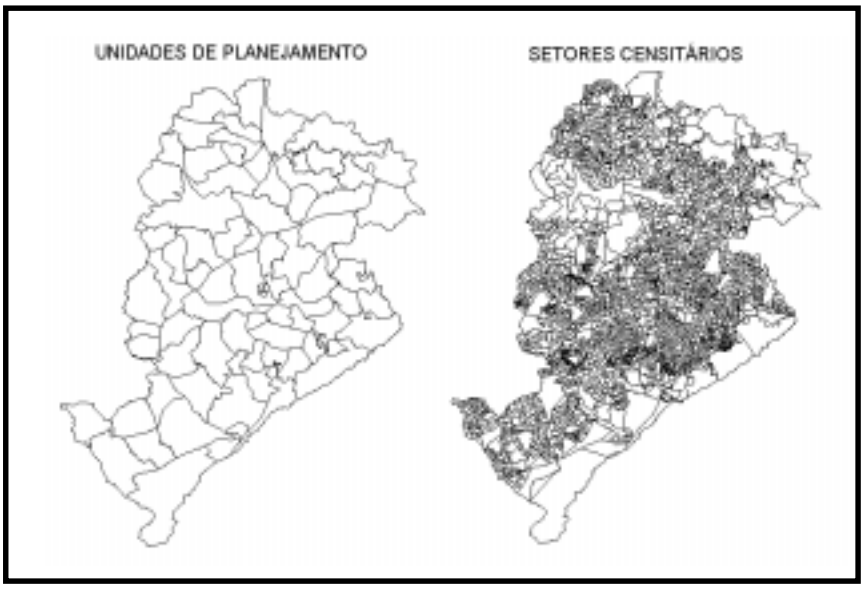


Figura 2 – Mapas do município de Belo Horizonte dividido em Unidades de Planejamento e em Setores Censitários

Para avaliar os efeitos da falácia ecológica foram computadas 1000 correlações entre 40 pares de variáveis do Censo, primeiramente utilizando os dados agrupados em setores censitários e posteriormente agrupados por UP. Foram definidos sete intervalos de valores de correlação (de – 0,4 a + 1,0) nos quais se enquadraram os valores encontrados. A Tabela 1 mostra o cruzamento dos coeficientes de correlação por setor censitário com as correlações por UP. Nas linhas da tabela representam-se os valores absolutos de correlação dos setores censitários e nas colunas os níveis de correlação por UP.

Tabela 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

		1	2	3	4	5	6	7
		Estudo1A3	Estudo4A7	Estudo Mais 15	Ocupa Própria	Ag Sem Can Inter	Sanea NãoTem	San Com Rede AE
Salário0,5A1	Setor Censitário	0,793	0,664	-0,500	0,477	0,535	0,506	0,388
	UP	0,969	0,907	-0,146	0,753	0,777	0,732	0,801
Salário2A3	Setor Censitário	0,557	0,829	-0,482	0,438	0,126	0,053	0,286
	UP	0,874	0,981	0,076	0,869	0,392	0,345	0,711
Salário3A5	Setor Censitário	0,073	0,466	-0,145	0,286	-0,157	-0,189	0,029
	UP	0,690	0,879	0,317	0,887	0,228	0,186	0,552

Os resultados da Tabela 1 indicam que as correlações nos setores censitários são significativamente menores que as correlações por unidades de planejamento. Nada menos que 802 correlações entre as 1000 são menores para os setores censitários que para as UPs. Apenas 40 (4%) têm o comportamento oposto. Em algumas situações, ocorre inclusive mudança de sinal, isto é, variáveis correlacionadas negativamente entre setores censitários passam a ser correlacionadas positivamente entre UPs.

Para melhor exemplificar, apresenta-se a Tabela 2 com sete variáveis nas colunas correlacionadas com três variáveis (linhas) de rendimentos em salários mínimos do chefe de família variando em três faixas, de 0,5 a 5. Nessa tabela pode-se observar a mudança de sinal e a diferença de valores nos dois níveis de subdivisões, como no caso em que foram tomadas as variáveis "número de chefes de família com 1 a 3 anos de estudo" e "número de chefes de família com rendimento entre 0,5 e 1 salário mínimo" e computou-se a correlação; caso de setores censitários em 0,793 e para o caso de UP aumentou para 0,969. Para os seguintes pares de variáveis o sinal da correlação mudou: o par "número de chefes de família com mais de 15 anos de estudo" e "número de chefes de família com 2 a 3 anos de estudo" e o par "não possui saneamento" e "número de chefes de família com rendimento entre 3 e 5 salários mínimos".

Tabela 2 – Demonstrativo das Correlações de Variáveis por Setor Censitário x Unidade de Planejamento

		1	2	3	4	5	6	7
		Estudo 1A3	Estudo 4A7	Estudo Mais 15	Ocupa Própria	Ag Sem Can Inter	Sanea Não Tem	San Com Rede AE
Salário0,5A1	Setor Censitário	0,793	0,664	-0,500	0,477	0,535	0,506	0,388
	UP	0,969	0,907	-0,146	0,753	0,777	0,732	0,801
Salário2A3	Setor Censitário	0,557	0,829	-0,482	0,438	0,126	0,053	0,286
	UP	0,874	0,981	0,076	0,869	0,392	0,345	0,711
Salário3A5	Setor Censitário	0,073	0,466	-0,145	0,286	-0,157	-0,189	0,029
	UP	0,690	0,879	0,317	0,887	0,228	0,186	0,552

Legenda: 1-"número de chefes de família com 1 a 3 anos de estudo", 2- "número de chefes de família com 4 a 7 anos de estudo", 3-"número de chefes de família com mais de 15 anos de estudo", 4-"domicílio ocupado é próprio", 5-"possui água mas sem canalização interna", 6-"não possui saneamento", 7-"possui saneamento com rede água e esgoto".

Teoricamente, seria possível lidar com esse problema conhecendo os dados individuais de coleta (ou pelo menos uma amostra deles). Neste caso, Wrigley et al [WHST96] indicam como utilizar os dados não-agregados para realizar correções nas correlações agregadas. Porém, na prática os dados individuais raramente estão disponíveis. Uma possibilidade é trabalhar com os dados mais desagregados possíveis (i.e., setores censitários no caso de censo) e utilizar técnicas de *clustering* ou de otimização combinatória para obter áreas mais agregadas, mas que preservem o fenômeno estudado da melhor forma possível.

Deve-se também adotar modelos que capturem as características de uma população composta em grupos geograficamente definidos. Wrigley et *al.* [WHST96] apresentam três modelos:

- *modelos de agrupamento*, em que os indivíduos não são escolhidos aleatoriamente e são utilizadas restrições de semelhança para pertencerem ao mesmo grupo/área;
- *modelos grupo-dependentes*, para o mesmo grupo/área são consideradas as influências externas semelhantes que afetam todo o grupo;
- *modelos de feedback*, considera-se a interação e influência entre os indivíduos, e esta se torna mais intensa entre indivíduos de um mesmo grupo/área.

Nos recentes censos no Reino Unido, o *Ordnance Survey* inglês[5] produz os dados agregados em *output areas* (áreas de agregação), distintas dos setores censitários, considerados apenas como unidades de suporte à coleta de dados [Mart98]. A agregação dos dados para a geração de *output areas* depende da definição de uma propriedade a ser estudada e da aplicação de um algoritmo de otimização [OpAl99]. Essencialmente, o algoritmo proposto por Openshaw maximiza as correlações das variáveis escolhidas, dentro das novas áreas agregadas, com restrições de forma dos polígonos resultantes. Como resultados, produz regiões mais homogêneas com relação ao critério escolhido.

Openshaw criou uma metodologia de procedimentos de divisão em zonas automatizados para uma maior padronização de modelos existentes de agregação geográfica para censo. E de acordo com Openshaw [Open84], é necessário projetar um esquema próprio de divisão em zonas, mas isto apenas minimiza em lugar de remover os problemas genéricos associados com geografias zonais sobre as quais foram esboçadas. Openshaw e Alvanides [OpAl99] desenvolveram uma rotina para divisão em zonas que oferece um número de funções de desenho de zonas genéricas, o Sistema de Desenho de Zona (ZDES) como um módulo adicional para o software ARC/INFO[6].

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 [OpWy97]. Para minimizar seu impacto com relação a estudos socioeconômicos, deve-se procurar utilizar a melhor subdivisão de área para o levantamento de dados disponíveis e utilizar técnicas semelhantes às de Openshaw *et al.* [OpAl99] para agregar os dados, de acordo com critérios relevantes para o fenômeno a ser estudado.

Os resultados acima indicam que não se pode afirmar que qualquer subdivisão de área 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 grandes áreas. Como regra geral, quanto mais desagregado o dado, maior a flexibilidade na escolha de modelos, pois agregar em unidades de área (regiões) maiores é fácil, mas desagregar é impossível.

[5] http://www.ordsvy.gov.uk

[6] http://www.geog.leeds.ac.uk/research/ccg.html

3. Estimação de Taxas em Áreas com Pequenas Populações

As seções anteriores apresentaram o problema de agregação de contagem 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 a Figura 3, que apresenta um mapa temático com a mortalidade infantil dos bairros do Rio de Janeiro, em 1994. Nesse mapa, o Rio de Janeiro está dividido em 153 bairros, e a taxa de mortalidade infantil anual para cada bairro expressa o número de óbitos de menores no primeiro ano de vida por mil nascidos vivos [CCB00] .

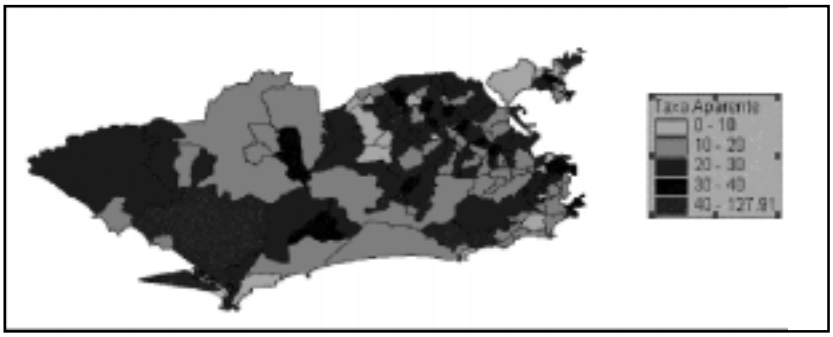


Figura 3 – Taxa total de mortalidade infantil por mil nascidos vivos no Rio de Janeiro, em 1994.

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 dois casos com taxas acima de 100 por mil nascidos. Um observador desatento poderia concluir que todos esses bairros apresentam um grave problema social. Na realidade, muitos desses valores extremos ocorrem nos bairros com pequenas populações, pois a subdivisão da cidade utilizada esconde enormes diferenças na população em risco, variando de 15 até 7.500 crianças nascidas 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 nesse 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 da população em risco. Vários estudos têm mostrado que as 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 [LoBa96]. Por isso mesmo, os valores extremos freqüentemente são os que mais chamam a

atenção num mapa temático de taxas, muitas vezes são resultado de um número reduzidíssimo de observações sendo, portanto, menos confiáveis, 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 é apenas uma realização de um processo não observado, e que é tanto menos confiável quanto menor é 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 4.

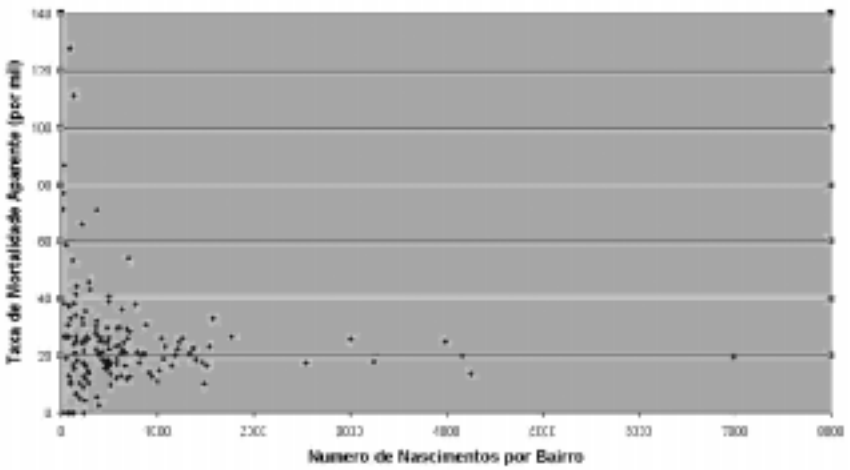


Figura 4 – Taxa de mortalidade infantil no Rio de Janeiro em 1994 em função do número de nascimentos por bairro

Nesse caso, 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. À medida que diminui a população em risco, aumenta muito a flutuação da taxa medida, formando o que é denominado *efeito funil* [BaGa95]. 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 [Anse92,Anse95,Anse96]. Esta formulação sugere o uso de técnicas de *estimação bayesiana* [Mars91]. 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, e z_i é o número de eventos na i-ésima área.

A idéia do estimador bayesiano [Bail01] 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:

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

O peso w_i é dado por:

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

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.

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*. Esse 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:

$$\hat{\mu} = \frac{\sum y_i}{\sum n_i} \qquad (3)$$

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

$$\sigma^2 = \frac{\sum n_i (ti - \hat{\mu})^2}{\sum n_i} - \frac{\hat{\mu}}{n} \qquad (4)$$

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 se aplicar esta correção às taxas de mortalidade infantil do Rio de Janeiro, observa-se 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 compara-

ção entre a taxa primária e o valor estimado está apresentada na Figura 5. Em resumo, é necessário tomar extremo cuidado ao produzir mapas temáticos, especialmente em casos onde são apresentadas taxas calculadas sobre populações com valores reduzidos.

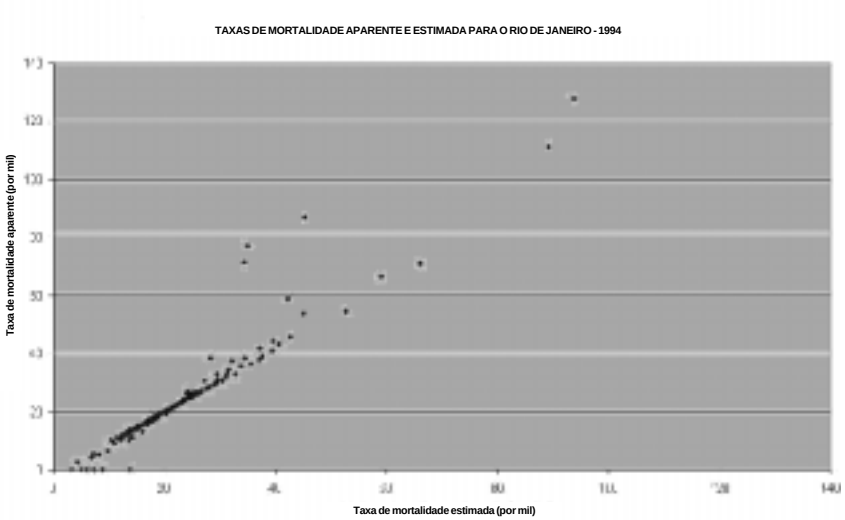


Figura 5 – Comparação entre a taxa de mortalidade infantil observada e a taxa estimada pelo método bayesiano empírico.

O estimador bayesiano empírico pode ser generalizado para incluir efeitos espaciais. Neste caso, a idéia é fazer o cálculo da 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 no Recife (Figura 6) onde foi utilizado esse método local para estimar a taxa da doença nos bairros da cidade [SBB+01].

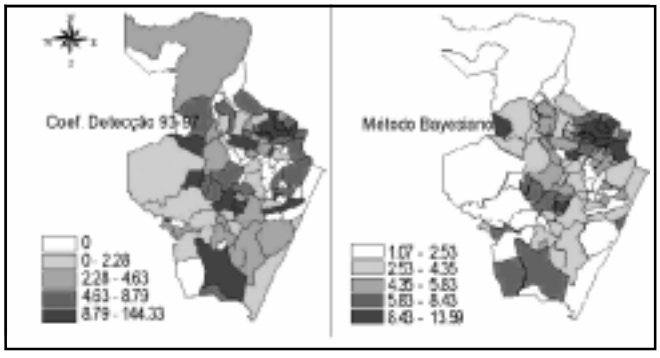


Figura 6 – Taxas de detecção média de hanseníase em menores de 15 anos, período 1993-1997, por bairro do Recife, e taxas estimadas através do método bayesiano

Através do mapa "corrigido" foi possível indicar bairros prioritários para a atuação da vigilância epidemiológica por apresentarem valores altos mesmo após suavização do indicador.

4. Considerações Finais

Este artigo 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. No estudo realizado foram discutidas algumas das principais fontes dos problemas advindos dos efeitos de escala e de agregação e apresentados métodos de estimação bayesiana para taxas que permitem a correção de efeitos associados a pequenas populações. Em resumo, estudiosos de dados socioeconômicos podem se beneficiar substancialmente das técnicas apresentadas.

Keywords
Spatial analysis – Geostatistics – Census data – Geographic Information Systems

Abstract
This paper presents problems related to the manipulation of spatial data consolidated by areal units and their interpretation in various scales. The granularity of the territorial subdivision interferes in the results, possibly leading to inappropriate conclusions about the phenomenon under study. Other themes that are relevant in the analysis of these data, such as aggregation and zoning, along with the estimation of rates in sparsely populated areas are discussed through examples. Solutions in the realm of spatial analysis are proposed for the reduction of the distortions caused by the data aggregation according to areas.

Referências Bibliográficas

[Anse92] 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.

[Anse95] ANSELIN, L. Local Indicators of Spatial Association - LISA. *Geographical Analysis,* v.27, p.91-115, 1995.

[Anse96] 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, p.111-126, 1996.

[Bail01] BAILEY, T. C. Spatial Statistics Methods in Health. *Cadernos de Saúde Pública*, v.17, n.5, 2001.

[BaGa95] BAILEY, T.C., GATRELL, A.C. . *Interactive spatial data analysis*, 1. ed. Essex. Longman Scientific & Technical, 1995.

[CCB00] CAMPOS, T.P.; CARVALHO, M.S.; BARCELLOS, C. *Áreas de risco e trajetória dos pacientes aos serviços: uma discussão da mortalidade infantil no município do Rio de Janeiro*. Revista Panam Salud Publica (*Panam. J. Public Healht*), Washington, v. 8, n. 3, p. 164-171, 2000.

[FBC00] FOTHERINGHAM A . S., BRUNSDON C, e CHARLTON M. . *Quantitative Geography: perspectives on spatial data analysis.* Londres: Salva, 2000.

[HSTW96] HOLT, D., STEEL, D., TRANMER, M., WRIGLEY, N. *Aggregation and ecological effects in geographically based data*. Geographical Analysis, 1996.

[LoBa96] LONGLEY, P., BATTY, M... *Spatial analysis: modelling in a GIS environment*, John Wiley & Sons, 1996.

[LGMR01] LONGLEY, P. A., GOODCHILD, M. F., MAGUIRE, D. J. RHIND, D. W. *Geographic information systems and science.* John Wiley & Sons, 2001.

[Mars91] MARSHALL, R. Mapping disease and mortality rates using empirical Bayes estimators. *Applied Statistics*, v.40, p.283-294, 1991.

[Mart00] MARTIN, D. Census 2001: making the best of zonal geographies. Paper presented at *The Census of Population: 2000 and Beyond*, University of Manchester 22-23. June, 2000.

[Mart98] MARTIN, D. Optimizing census geography: the separation of collection and output geographies. *International Journal of Geographical Information Science, v.*12, p. 673-685, 1998.

[Mart95] MARTIN, D. *Geographic Information Systems: Socioeconomic Applications.* London, Routledge, 1995.

[OpAl99] OPENSHAW, S., ALVANIDES, S. Applying geocomputation to the analysis of spatial distributions In: LONGLEY, P. A., GOODCHILD, M. F., MAGUIRE, D. J., RHIND, D. W. (eds) *Geographical Information Systems: Principles, Techniques, Applications and Management* Chichester: Wiley, v. 1, 267-282, 1999.

[Open84] OPENSHAW, S.. *Ecological fallacies and the analysis of areal census data*. Environment and Planning, 1984.

[OpWy97] OPENSHAW, S., WYMER, C.. *Artificial Intelligence in Geography*. Chichester, John Wiley,1997.

[SBBC 01] SOUZA, W. V., BARCELLOS, C., BRITO, A. M., CARVALHO, M. S., *et al*. Aplicação de modelo bayesiano empírico na análise espacial da ocorrência de hanseníase. *Revista de Saúde Pública*, São Paulo, v. 35, n. 5, p. 474-480, 2001.

[SMPL96] SECRETARIA MUNICIPAL DE PLANEJAMENTO – PBH, 1996 – O IQVU – Índice de qualidade de vida urbana. http://www.pbh.gov.br/smpl/iqvu/.

[Stee85] STEEL, D. *Statistical analysis of populations with group structure*. Unpuplished PhD dissertation available from Department of Social Sciences, University of Southampton, Southampton, UK *apud* Spatial Analysis: modelling in a GIS environment. John Wiley & Sons, 1996.

[StHo96] STEEL, D., HOLD, T.. Analysing and adjusting aggregation effects: the ecological fallacy revisited. *International Statistical Review,* 1996.

[WHST96] WRIGLEY, N., HOLD, T., STEEL, D., TRANMER, M. *Analysing, modelling, and resolving the ecological fallacy* In: LONGLEY, P. BATTY, M.. *Spatial analysis:* m*odelling in a GIS environment*. John Wiley & Sons, 1996.