PROBLEMAS DE ESCALA E A RELAÇÃO ÁREA-INDIVÍDUO EM ANALISE ESPACIAL DE DADOS CENSITÁRIOS

Taciana de Lemos Dias[1]
Analista de Sistemas da Prodabel
Doutoranda em Computação Aplicada do INPE – Instituto Nacional de Pesquisas Espaciais
Áreas de Interesse: modelos para representação espaço-temporais urbanos, Ontologias, Modelos conceituais e temporalidade em Banco de dados, Gestão e Recuperação da Informação e Geoprocessamento.
Maria da Piedade Gomes de Oliveira[2]
Assessora do Centro de desenvolvimento e da Prodabel -CDE
Doutoranda em Computação Aplicada do INPE – Instituto Nacional de Pesquisas Espaciais
Áreas de Interesse: Analise Espacial, Geoestatistica, 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 do 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]
Professora e Pesquisadora titular da Escola Nacional de Saúde Pública-ENSP e Fundação Oswaldo Cruz – FIOCRUZ
Doutora em Engenharia Biomédica, COPPE/UFRJ
Áreas de Interesse: Métodos de análise de dados espaciais, Modelagem estatística de microáreas, Saúde Pública e Epidemiologia.

PALAVRAS –CHAVE
Analise espacial – Falácia ecológica - Estimação de Taxas – Dados censitários — GIS – Sistemas de Informações Geográficas – Planejamento urbano e de cidades Modelagem estatística de microáreas.

[1] E-mail;taciana@dpi.inpe.br
[2] E-mail;piedade@dpi.inpe.br
[3] E-mail;Gilberto@dpi.inpe.br
[4] E-mail; carvalho@procc.fiocruz.br

RESUMO

A falácia ecológica ocorre quando se realiza analises com resultados derivados de agregação de valores por unidade de área, inferindo que estes valores correspondem ao nível individual. Em geral os resultados apresentam diferenças e podem dar margem a análises incorretas sobre determinado fenômeno. Este artigo analisa a agregação de dados e exemplifica os problemas de analise advindos desta. E, usando de estatística faz uma analise para avaliar a falácia ecológica e a estimação de taxas com o objetivo de alertar os planejadores de cidades e analistas quanto aos efeitos desses problemas.

Este artigo discute os problemas de estatística espacial associados ao uso de dados censitários, com ênfase

1. INTRODUÇÃO

Compreender a distribuição espacial de fenômenos constitui hoje um grande desafio para a elucidação de questões centrais em diversas áreas do conhecimento, seja em saúde, em ambiente, em geologia, em agronomia, entre tantas outras. Tais estudos vêm se tornando cada vez mais comuns, devido à crescente democratização das informações, aos avanços tecnológicos e seu baixo custo e a 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´s permitem a apresentação espacial de variáveis como população de indivíduos, índices de qualidade de vida ou vendas de empresas numa região, através de mapas. Para tanto, basta dispor de um banco de dados e de uma base geográfica (como um mapa de municípios), e grande parte dos SIG´s é capaz de apresentar um mapa colorido (coropléticos) permitindo a visualização do padrão espacial do fenômeno. Esses mapas são construídos através de valores que correspondem às propriedades das áreas geográficas ou considera o valor de uma propriedade especifica a qual é associada a uma cor [LGMR,01].

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 ocorrência de doenças. A distribuição dos casos de uma doença forma um padrão no espaço? Existe

associação com alguma fonte de poluição? Evidência de contágio? Variou no tempo?

- Deseja-se investigar se existe alguma concentração espacial na distribuição de roubos. Roubos que ocorrem em determinadas áreas estão correlacionados com características sócio-econômicas 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?

- Deseja-se 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 é 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 [LGMR,01] :" *todas as coisas são parecidas mas coisas mais próximas se parecem mais que coisas mais distantes"*.

A taxonomia mais utilizada 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 regularmente 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* - tratam-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. Estes dados são agregados em unidades de análise, usualmente delimitadas por polígonos fechados (setores censitários, zonas de endereçamento postal e municípios).

As origens dos dados geralmente utilizados em *analise de áreas* são, em grande parte, oriundos de levantamentos populacionais tais como

censos, estatísticas de saúde e cadastramento de imóveis. Estas áreas usualmente possuem uma delimitação onde se supõe haver homogeneidade interna, ou seja, as áreas são compostas de agrupamentos aleatórios de indivíduos/moradias que tendem a ser semelhantes em relação a outras áreas. A probabilidade dessa semelhança pode ocorrer, por exemplo, no campo sócio-econômico, demográfico, de variáveis de saúde e morfologia do solo [WHST,96]. Evidentemente, esta premissa nem sempre é verdadeira e não há qualquer garantia de que a distribuição do evento seja homogênea dentro destas unidades, 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 GIS interpolam um limite entre pontos amostrais, como na criação de mapas isopleticos..

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 [Mart,95]. Neste caso, tanto a apresentação em mapas coropléticos como 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 estes 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 a confidência de registros individuais e estes são projetados para evitar que informações que possibilitem a identificação dos indivíduos sejam disponibilizadas, e a agregação geográfica é a única forma disponível [Mart,00]. Isso ocorre no caso do Censo, onde os dados já agregados por setores censitários são o menor tipo de agrupamento a que a comunidade em geral tem acesso para vários tipos de analises. Um setor censitário corresponde à capacidade de levantamento do recenseador, variando por região em torno de 200 a 400 domicílios. 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.

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 funções 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" (MAUP - *Modifiable Areal Unit Problem*) [FBC,00] [LB,96]. Openshaw e Taylor [OW,97] 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 indivíduo e nas áreas [WHST,96]. Este fenômeno, nas ciências sociais e na epidemiologia, é chamado de "falácia ecológica" que envolve a conclusão imprópria de relacionamentos a nível individual a partir de resultados agregados ao nível de unidade de área.. 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 reduzir flutuações estatísticas.

Por exemplo, em um conjunto de indivíduos onde são medidas duas características de cada individuo, conforme estimado na Figura 1 (a). Uma regressão considerando todos os indivíduos (linha negra do quadro à esquerda) resulta em um coeficiente positivo de 0,1469. Esses indivíduos pertencem a grupos distintos, separando cada grupo conforme o atributo tons de cinza, 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 existe modelo correto. No primeiro caso, pode-se dizer que sem informações que permitam separar os indivíduos nos grupos tons de cinza, 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.

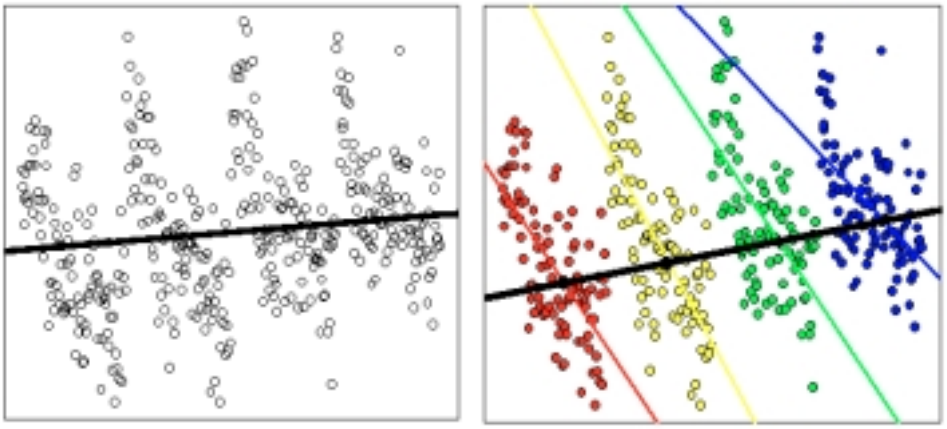

Figura 1 – Modelos de regressão: (a) indivíduos, (b) indivíduos em extratos diferentes e grupos.

Para ilustrar os efeitos de escala em unidades de área, tomou-se os dados oficiais do censo de Belo Horizonte para o ano de 1991, em duas escalas: os setores censitários e as unidades de planejamento (UP´s), 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 UP´s são 80 divisões político administrativas do município, definidas em 1996, que levaram em consideração fatores como topologia, agrupamentos sociais e outros.

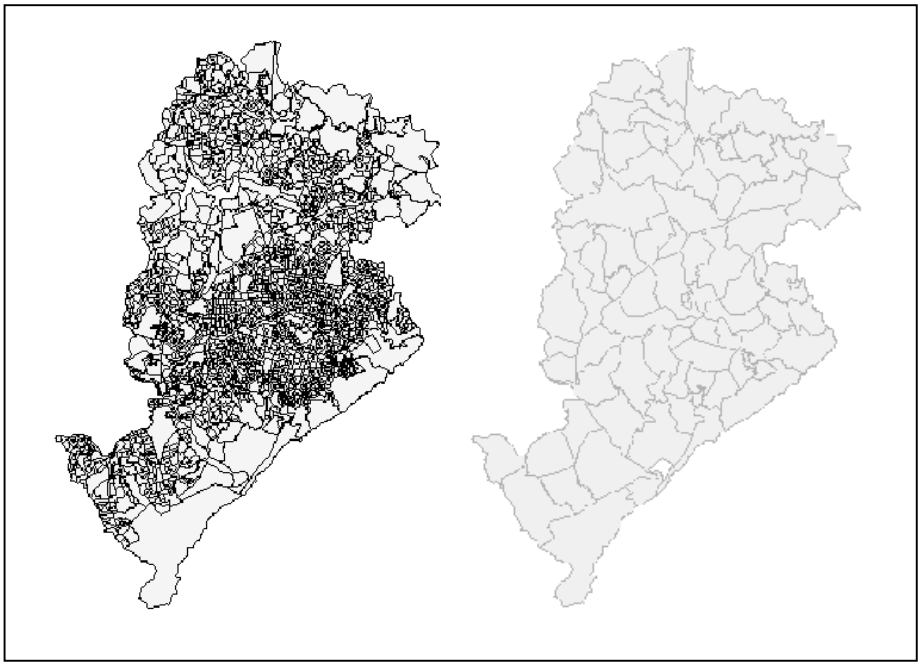

Figura 2 – Setores censitários (à esquerda) e Unidades de Planejamento (à direita) para o município de Belo Horizonte.

Para avaliar os efeitos da *falácia ecológica*, os 1998 registros de setores censitários foram agregados em 80 unidades de planejamento. A partir das variáveis do censo, foram computadas 1000 correlações entre 40 pares de variáveis, primeiramente utilizando os dados agrupados em setores censitários e posteriormente agrupados por UP. Foram definidos 7 intervalos de valores de correlação (de –0,4 à +1,0) nos quais foram enquadrados 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.

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**

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

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 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 no nível dos setores censitários passam a ser correlacionadas positivamente.

Para melhor exemplificar apresenta-se a Tabela 2 com algumas variáveis (“número de chefes de família com 1 a 3 anos de estudo”, “número de chefes de família com 4 a 7 anos de estudo”, “número de chefes de família com mais de 15 anos de estudo”, “domicílio ocupado é próprio”, “possui água mas sem canalização interna”, “não possui saneamento”,

"possui saneamento com rede água e esgoto") correlacionadas com as variáveis de "número de chefes de família com rendimento entre 0,5 e 1 salário mínimo" até "número de chefes de família com rendimento entre 3 e 5 salários mínimo". Nessa tabela se pode observar a mudança de sinal e a diferença de valores das duas escalas. Como no caso em que se tomou 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 para o caso de setores censitários (0,79) e para o caso de UP (0,96). 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 mínimo"

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

		Estudo 1A3	**Estudo 4A7**	**Estudo Mais 15**	**Ocupa Própria**	**AgSem Can Inter**	**Sanea Não Tem**	**SanCom RedeAE**
Salário 0,5A1	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
Salário 2A3	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
Salário 3A5	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

Teoricamente, seria possível lidar com este problema conhecendo os dados individuais de coleta (ou pelo menos uma amostra deles). Neste caso, Wrigley et al [WHST,96] indicam como utilizar os dados não-agregados para realizar correções nas correlações agregadas. Na prática os dados individuais muito raramente estão disponíveis. O que fazer então? Uma possibilidade é trabalhar com os dados na escala espacial mais desagregada possível (menores) (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 [WHST,96] apresentam tres modelos :

- *modelos de agrupamento*, quando os indivíduos não são escolhidos aleatoriamente e são utilizadas restrições de semelhança para pertencerem ao mesmo grupo/área;
- *modelos grupo-dependentes*, quando para o mesmo grupo/área são consideradas as influências externas semelhantes que afetam todo o grupo;
- *modelos de feedback*, quando se considera 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 *Ordinance Survey* inglês (htpp://www.ordsvy.gov.uk) produz os dados agregados em "output areas" (áreas de agregação), distintas dos setores censitários, considerados apenas como unidades de suporte à coleta de dados [Mart,98]. 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 [Open,99]. 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 [LGMR,01] criou uma metodologia de procedimentos de divisão em zonas automatizados (AZP) para uma maior padronização de modelos existentes de agregação geográfica para censo. E de acordo com Openshaw [Open,84], 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 Rao, 1995; Alvanides, 1995) desenvolveram uma rotina para divisão em zonas que oferece um número de funções de desenho de zona genéricas, o Sistema de Desenho de Zona (ZDES) como um modulo adicional para o Arc/Info (http://www.geog.leeds.ac.uk/research/ccg.html).

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 [OW,97]. Para minimizar seu impacto com relação a estudos sócio-econômicos, deve-se procurar utilizar a melhor escala de levantamento de dados disponível e utilizar técnicas semelhantes às de Openshaw et al [OA,99] 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 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 grandes 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 é impossível.

3. ESTIMAÇÃO DE TAXAS EM ÁREAS COM PEQUENAS POPULAÇÕES

As seções anteriores apresentaram 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 **Erro! A origem da referência não foi encontrada.** Figura 3 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 no primeiro ano de vida, por mil nascidos vivos. [DC, 98]

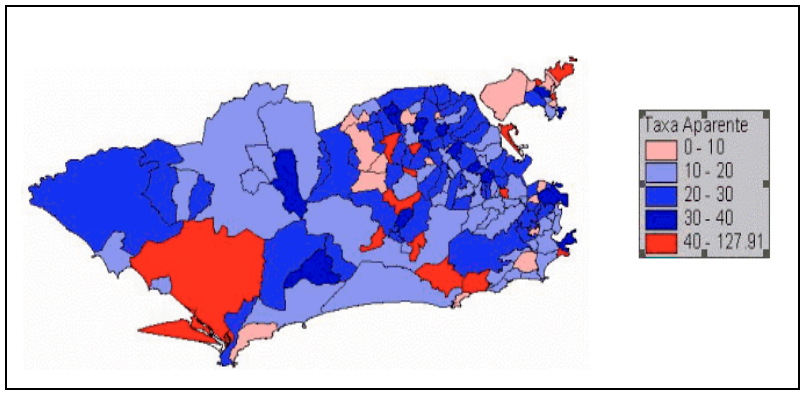

Figura 3– Taxa total de mortalidade infantil por mil nascidos vivos no Rio de Janeiro, em 1994.

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 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 da 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 [LB,96] **.** 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 4.

Figura 4– 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 [CCN,95]. Conforme diminui a população em risco, aumenta muito a flutuação da taxa medida, formando o que já foi denominado de "efeito funil" [BG,95] **.** Nos bairros de menor população, esta variação oscilou de 0 a quase 130 por mil. [CCN,96]

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 [Anse,92,95,96]. Esta formulação sugere o uso de técnicas de estimação bayesiana. [Mars,91] [Carv,97]. 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 [Bail,01] 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 :

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

O fator w_i é dado por:

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

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*. 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:

$$\hat{\mu} = \frac{\sum y_i}{\sum n_i} \qquad (4)$$

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

$$\sigma^2 = \frac{\sum n_i (t_i - \hat{\mu})^2}{\sum n_i} - \frac{\hat{\mu}}{\bar{n}} \qquad (5)$$

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

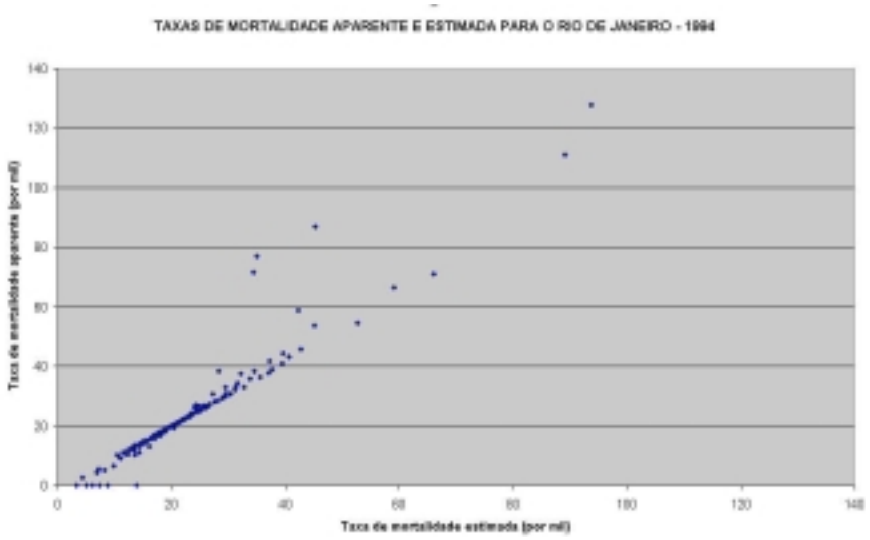

Figura 5– Comparação entre a taxa de mortalidade infantil observada e a taxa estimada pelo método *bayesiano* empírico.

O estimador bayesiano empírico pode ser generalizado para incluir efeitos espaciais. Neste caso, a idéia é fazer 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 6) onde foi utilizado esse método local para estimar a taxa da doença nos bairros da cidade.

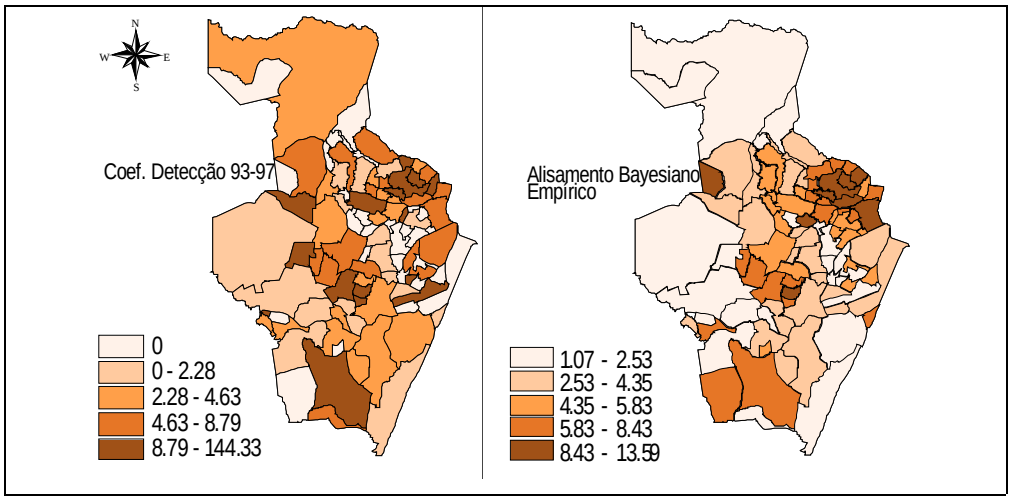

Figura 6- Taxas de detecção média de hanseníase em menores de 15 anos, período 1993-1997, por bairro do Recife, e taxas estimada através de alisamento bayesiano.

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

No estudo realizado, são discutidas algumas das principais fontes dos problemas advindos dos efeitos de escala e de agregação, que alteram os resultados obtidos e acarretam conclusões impróprias. São ressaltados as

preocupação dos pesquisadores e o reconhecimento de que não existe uma solução, mas sim, possíveis caminhos capazes de minimizar esses problemas.

Os exemplos dados permitiram perceber os problemas sobre a falácia ecológica como também da estimação de taxas e a facilidade dos investigadores em definir e tomar decisões sobre unidades de área, nas quais, os efeitos de escala e zoneamento degradam a qualidade do dado.

A geografia regional busca delinear zonas uniformes, com homogeneidade interna dentro de um esquema zonal que maximiza heterogeneidade entre zonas, utilizando variáveis como clima, desenvolvimento econômico, uso de terra agrícola, ou distribuição populacional. Para obter zonas razoavelmente homogêneas foram apresentados estudos avançados de técnicas estatísticas multivariadas e análise de agrupamentos.

KEYWORDS

ABSTRACT

REFERÊNCIAS BIBLIOGRÁFICAS

[Anse,92] 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.

[Anse,95] ANSELIN, L. Local indicators of spatial association - LISA. *Geographical Analysis* v.27, p.91-115, 1995.

[Anse,96] 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.

[Bail,01] BAILEY, T. Spatial Statistics Methods in Health. *Cadernos de Saúde Pública*, v.17, n.5,, 2001.

[BG,95] BAILEY,T.C., GATRELL,A.C. . *Interactive spatial data analysis,* 1 ed. Essex. Longman Scientific & Technical.1995.

[Carv,97] CARVALHO, Marilia Sá. *Aplicação de métodos de análise espacial na caracterização de áreas de risco à saúde* . Tese defendida na Universidade Federal do Rio de Janeiro, COPPE.

[CCN,95] CARVALHO, M.S., CRUZ, O.G., NOBRE, F.F.. Análise multivariada do censo 1991 por setores censitários - Região Metropolitana do Rio de Janeiro/Brasil. In: *Resumos do III Congresso Brasileiro de Epidemiologia*, pp.18, Salvador, Jun.1995.

[CCN,96] CARVALHO, M.S., CRUZ, O.G., NOBRE, F.F., 1996, *Spatial partition using multivariate cluster analysis and contiguity algorithm: application to Rio de Janeiro, Brazil.* Statistics in Medicine, v.15, pp.1885-1894.

[DC, 98] D'ÓRSI, E. and Marília 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.

[FBC,00] FOTHERINGHAM.2000 A . S., BRUNSDON C, e CHARLTON M. . *Quantitative Geography: Perspectives on spatial data analysis.* Londres: Salva. 2000.

[HSTW,96] HOLT, D., STEEL,D., TRANMER, M.,WRIGLEY, N. *Agreegation and ecological effects in geographically based data*. Geographical Analysis. 1996.

[LB,96] LONGLEY, Paul, BATTY, Michael. *Spatial Analysis: Modelling in a GIS Environment.* John Wiley & Sons, 1996.

[LGMR,01] LONGLEY, Paul A., GOODCHILD, Michael F., MAGUIRE, David J. RHIND, David W. *Geographic information systems and science.* John Wiley & Sons, 2001.

[Mars,91] MARSHALL, R. Mapping disease and mortality rates using empirical Bayes estimators. *Applied Statistics* v.40, p.283-294, 1991.

[Mart,00] MARTIN, David. 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.

[Mart,95] MARTIN, D. *Geographic Information Systems: Socioeconomic Applications.* London, Routledge, 1995.

[Mart,98] MARTIN, D. Optimizing census geography: the separation of collection and output geographies. *International Journal of Geographical Information Science.* 12, 673-685. 1998.

[OA,99] OPENSHAW, S., ALVANIDES, S. Applying geocomputation to the analysis of spatial distributions In: Longley, P. A., Goodchild, M. F.,

Maguire, D. J. and Rhind, D. W. (Eds) *Geographical Information Systems: Principles, Techniques, Applications and Management* Chichester: Wiley, Vol 1, 267-282.1999.

[Open,84] OPENSHAW, Stan.*Ecological fallacies and the analysis of areal census data*. Environment and Planning. 1984.

[OW,97] OPENSHAW, S., WYMER, C. *Artificial Intelligence in Geography*. Chichester, John Wiley,1997.

[SH,96] STEEL, David, HOLD, Tim. Analysing and adjusting aggregation effects: the ecological fallacy revisited. *International Statistical Review* .1996.

[Stee,85] 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.

[WHST,96] WRIGLEY, Neil, HOLD, Tim, STEEL, David, TRANMER, Mark. *Analysing, modeling, and resolving the ecological fallacy* In: LONGLEY, Paul, BATTY, Michael. *Spatial Analysis: Modelling in a GIS Environment*. John Wiley & Sons, 1996.

AGRADECIMENTOS