Pós-Graduação em Geografia Física
FLG 5153
Princípios de Cartografia e Análise Espacial aplicados à Geografia da Saúde

PROBLEMAS DE ESCALA (MAUP, falácia ecológica) e ESTIMADOR BAYESIANO

Ligia Vizeu Barrozo

Deptº. de Geografia

Faculdade de Filosofia, Letras e Ciências Humanas

USP

lija@usp.br

Referência bibliográfica

- DIAS, T.L., OLIVEIRA, M.P.G., CÂMARA, G., CARVALHO, M.S. Problemas de escala e a relação área-indivíduo em análise espacial de dados censitários. DPI-INPE.

Análise Espacial de Dados Geográficos

- Mensurar propriedades e relacionamentos, levando em conta a localização espacial do fenômeno em estudo de forma explícita (incorporar o espaço) – Lei de Tobler

Três tipos de dados

- Eventos ou padrões pontuais: fenômenos expressos através de ocorrências identificadas como pontos localizados no espaço, denominados processos pontuais. Ex.: localização de crimes, doenças, espécies vegetais;
- Superfícies contínuas: estimadas a partir de amostras de campo, regularmente ou irregularmente distribuídas;
- Áreas com contagens e taxas agregadas: dados associados a levantamentos populacionais, como censos e estatísticas de saúde. Estes dados são agregados em unidades de análise (setores censitários, CEP, municípios)

Análise de áreas

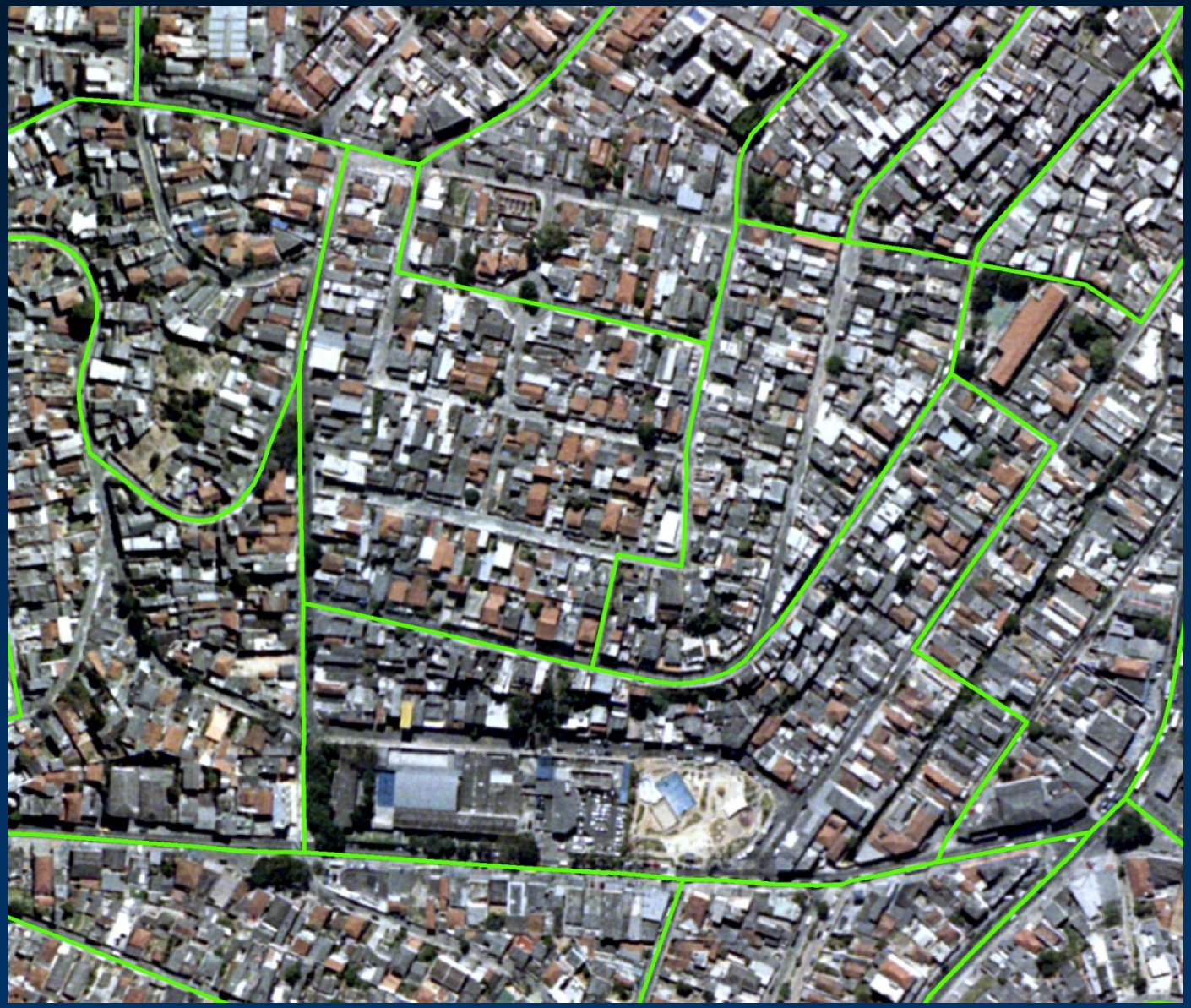
- dados oriundos de levantamentos populacionais
- supõe-se homogeneidade interna (agrupamentos aleatórios de indivíduos/moradias que tendem a ser semelhantes em relação a outras áreas)
- as unidades são definidas por critérios operacionais
- em diversas regiões, as unidades amostrais apresentam diferenças importantes em população e área

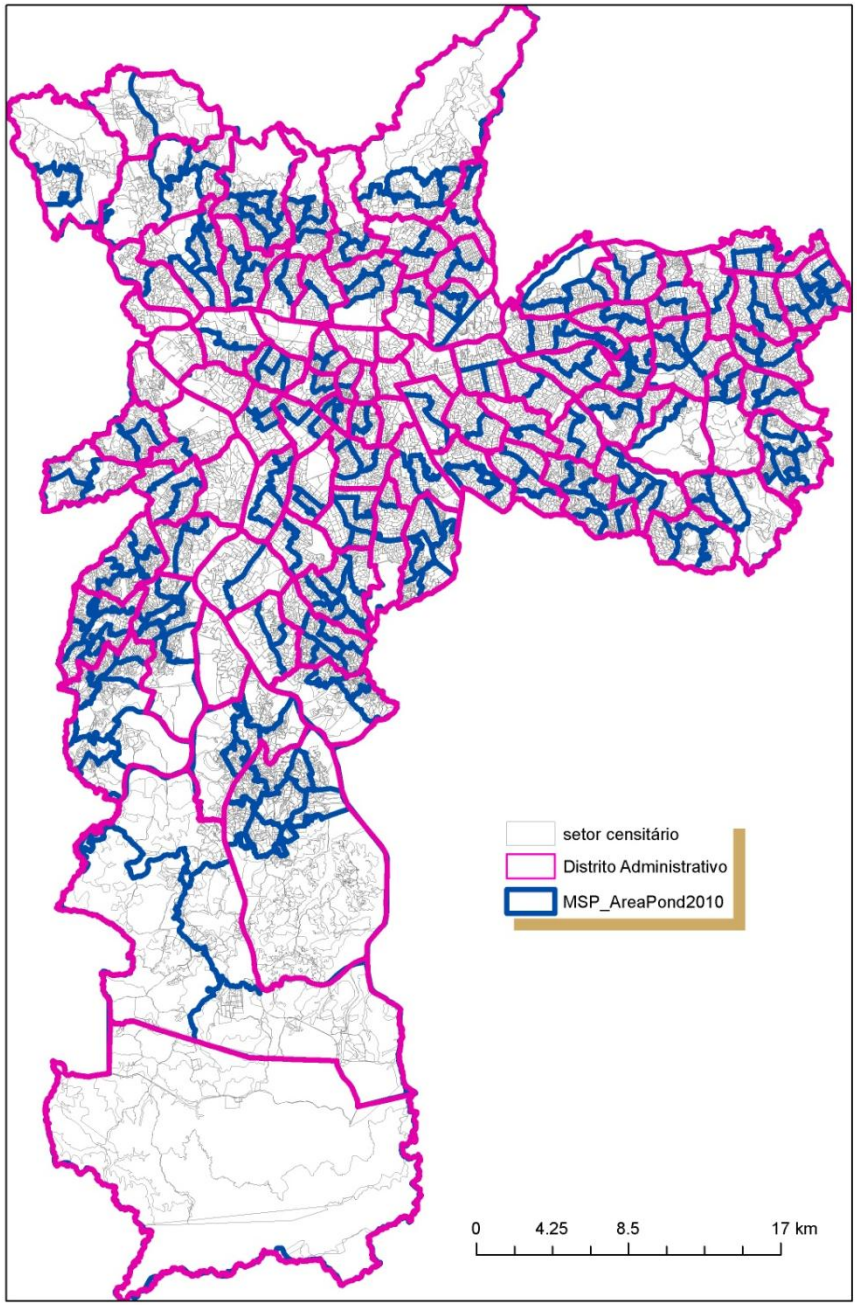
Efeitos de escala na análise de dados de área

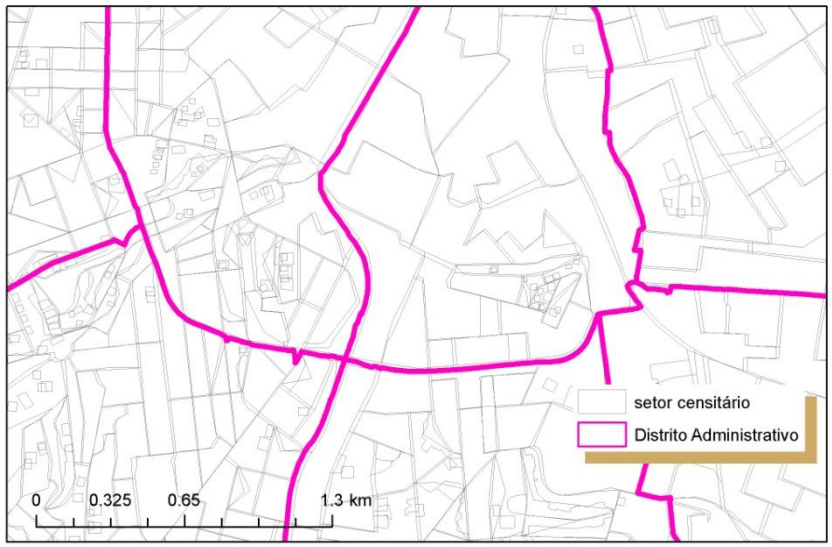
- Tendência de agregação geográfica para preservar a confidência de registros individuais
- No Censo, os dados são agregados por setor censitário
- cada setor corresponde à capacidade de levantamento do recenseador, variando em torno de 200 a 400 domicílios

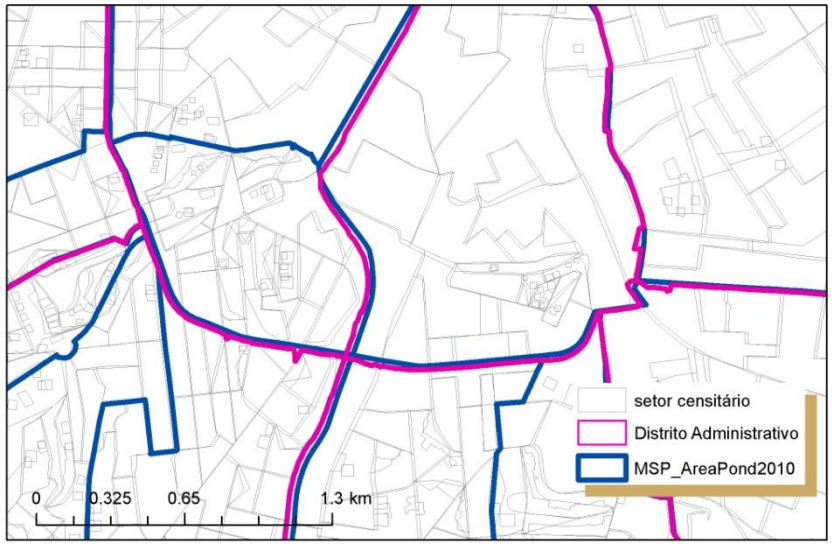

Spatial cluster
in São Paulo
preterm births
term births
0
25
50
100 m

SETOR CENSITÁRIO



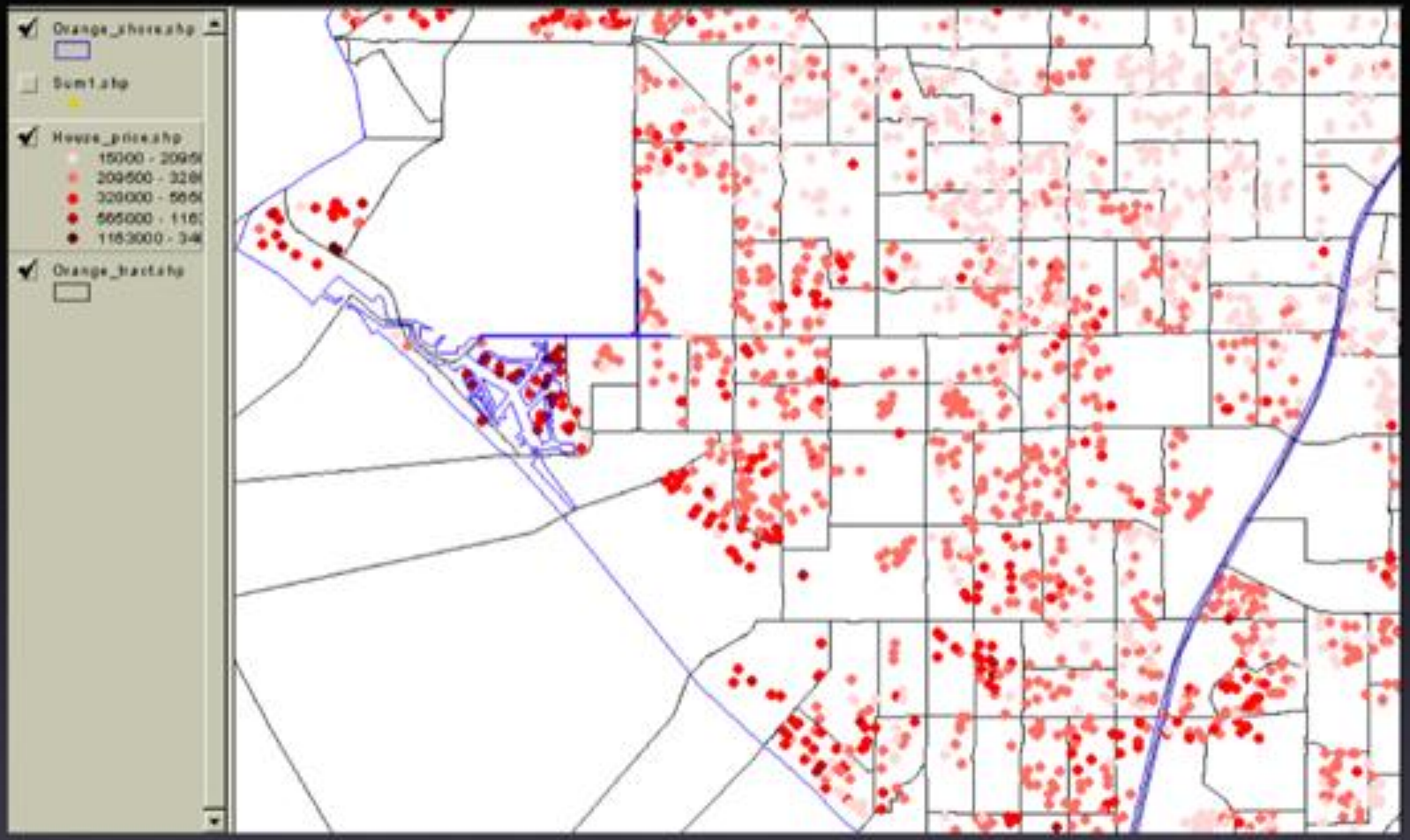

setor censitário
Distrito Administrativo
MSP_AreaPond2010
0
4.25
8.5
17 km


setor censitário
Distrito Administrativo
0
0.325
0.65
1.3 km


setor censitário
Distrito Administrativo
MSP_AreaPond2010
0
0.325
0.65
1.3 km

Problema com dados agregados por área

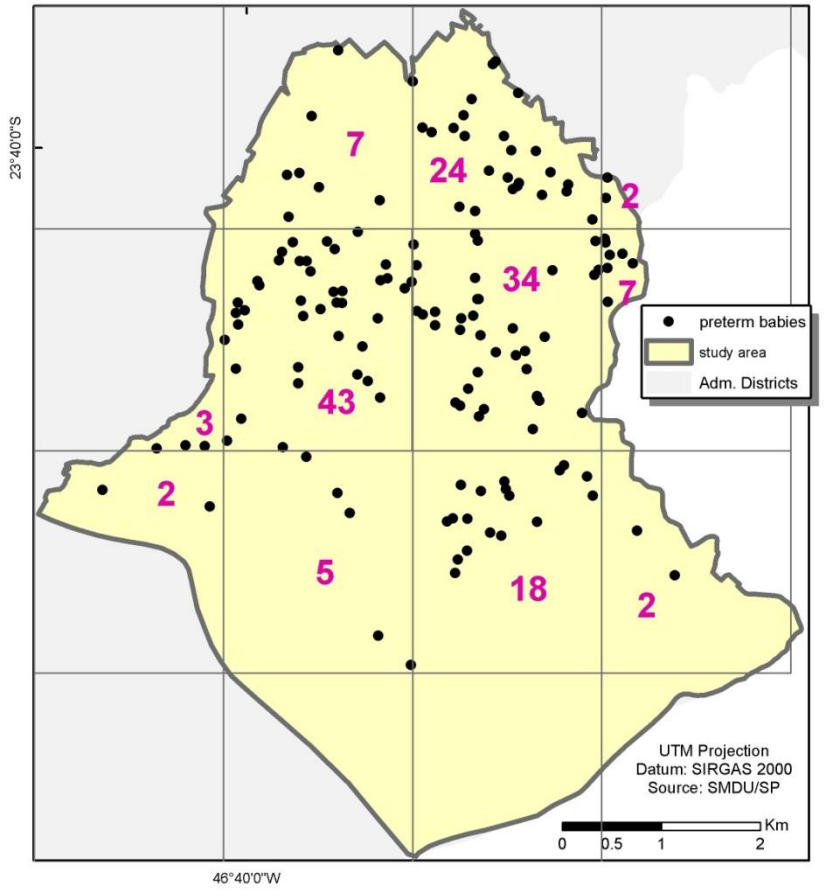
- Para uma mesma população estudada, a definição espacial das fronteiras afeta os resultados obtidos
- Pode-se obter resultados diferentes simplesmente alterando as fronteiras destas zonas – (MAUP – *modifiable areal unit problem*) "problema da unidade de área modificável"

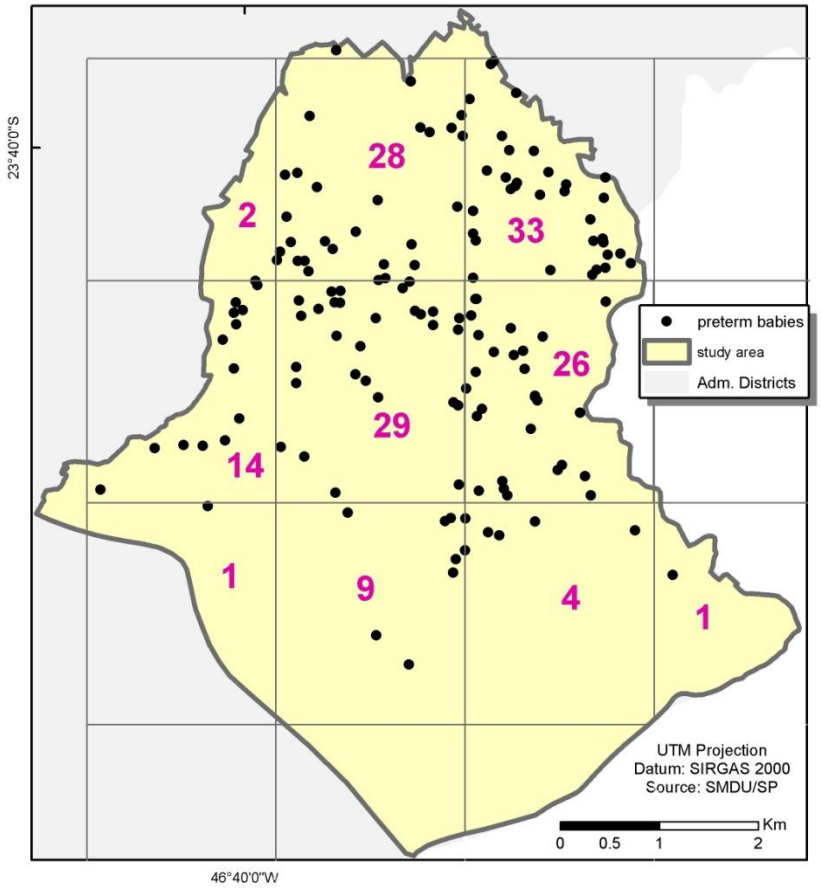


Spatial Heterogeneity of House Prices Within Tracts

Orange County, CA

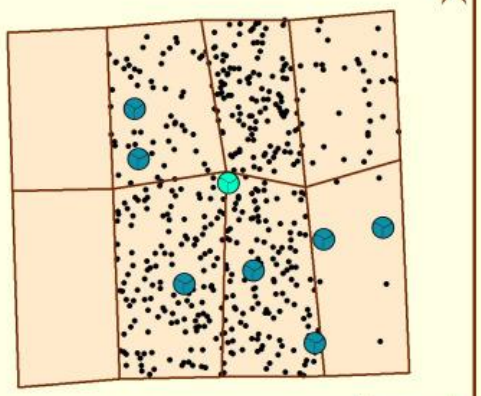
Zonation effect: Hypothetical divisions of administrative units and count of cases



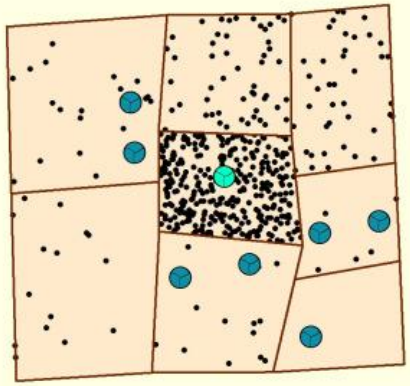


Revisiting John Snow's Cholera Map

Map 2



Map 3



Map 1

Cholera Deaths

- 1 - 2
- 3 - 5
- 6 - 10
- 11 - 15

Broad St. pump

Other pumps

Aggregated areas:

1 dot = 1 case

Dots do not show actual location

The map above (map 1) shows Cholera cases in London during an outbreak of the disease in 1854. The data was collected by Dr. John Snow (1813-1858) and in fact, the background map is his own map, that he prepared for the Board of Health. By mapping individual Cholera cases he made the connection between a contaminated water pump (on Broad Street) and the spread of the disease.

One of the challenges of mapping diseases is choosing the appropriate areal unit. These are usually census tracts or zip codes, where all cases that fall within that particular area are collected. Maps 2 & 3 show two types of areas where cases from map 1 have been counted within each area. Comparing maps 2 & 3 there is a striking difference in the number of cases and spatial distribution. This problem is referred to as the Modifiable Areal Unit Problem (MAUP).

John Snow

Ragnar Thrastarson - 5 Feb 2015 Lehman College - GEP 610 - LAB A

Sources: John Snow's map and data compiled by Robin Wilson 2011. Polygons for data aggregation created by Profs. Juliana Maantay and Andrew Maroko. Other imagery from Wiki Commons, 2015.

Problema com dados agregados por área

- Falácia ecológica: devido aos efeitos de escala e de agregação de áreas, os coeficientes de correlação podem ser inteiramente diferentes no indivíduo e nas áreas.

  Envolve a conclusão IMPRÓPRIA de relacionamentos a nível individual a partir de resultados agregados ao nível de unidade de área

FALÁCIA ECOLÓGICA

- Os resultados estatísticos têm validade dependente da unidade de área e do reconhecimento dos problemas existentes nas conclusões decorrentes de dados agregados;
- Propriedade inerente aos dados agregados por áreas;
- A agregação tende a AUMENTAR a correlação entre as variáveis e REDUZIR flutuações estatísticas.

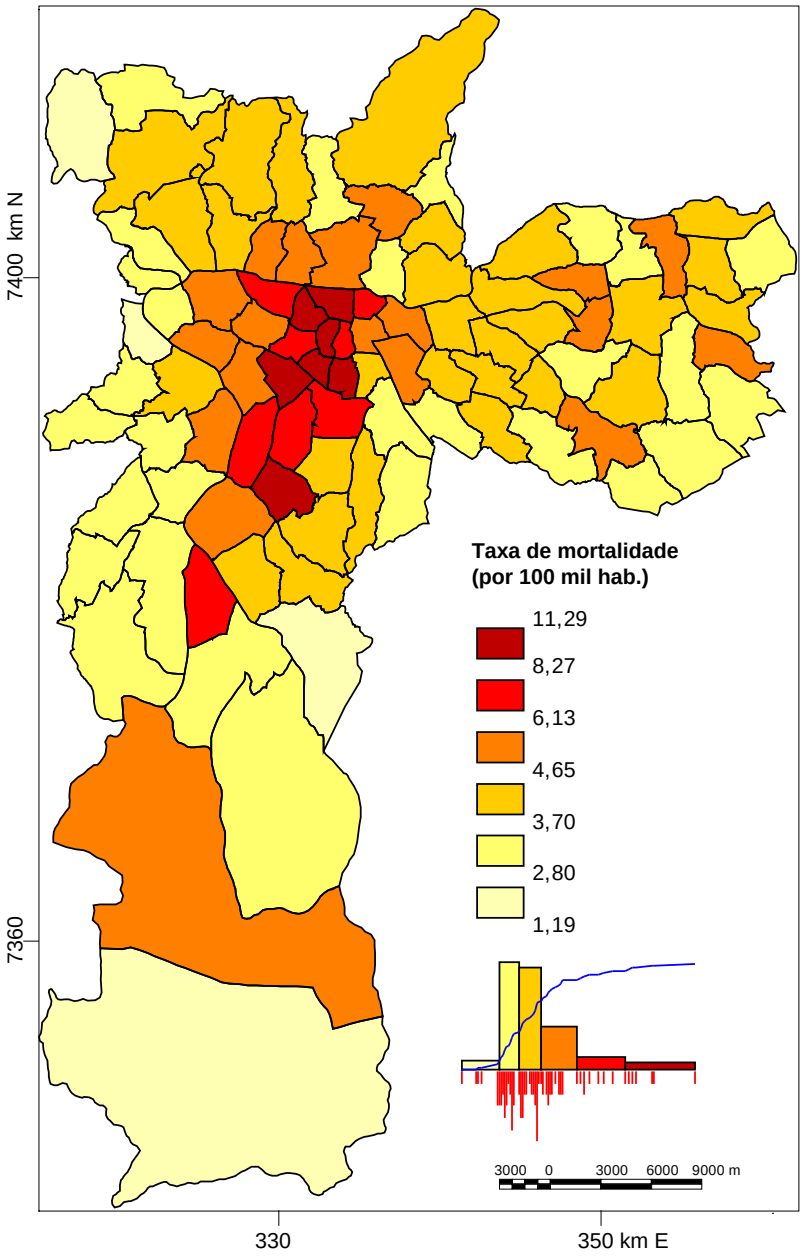
Bando *et al. BMC Psychiatry* 2012, **12**:124
http://www.biomedcentral.com/1471-244X/12/124

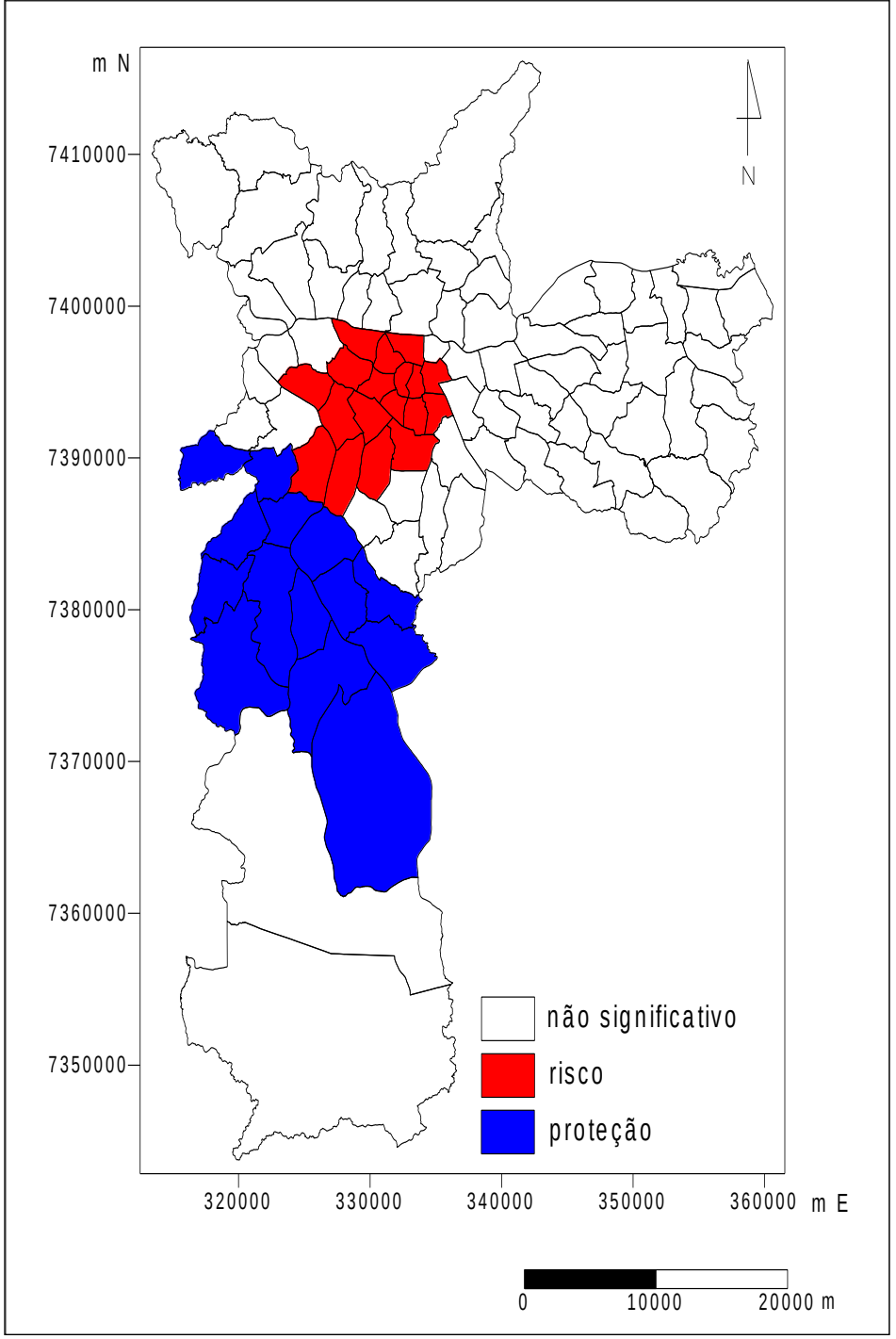
BMC Psychiatry

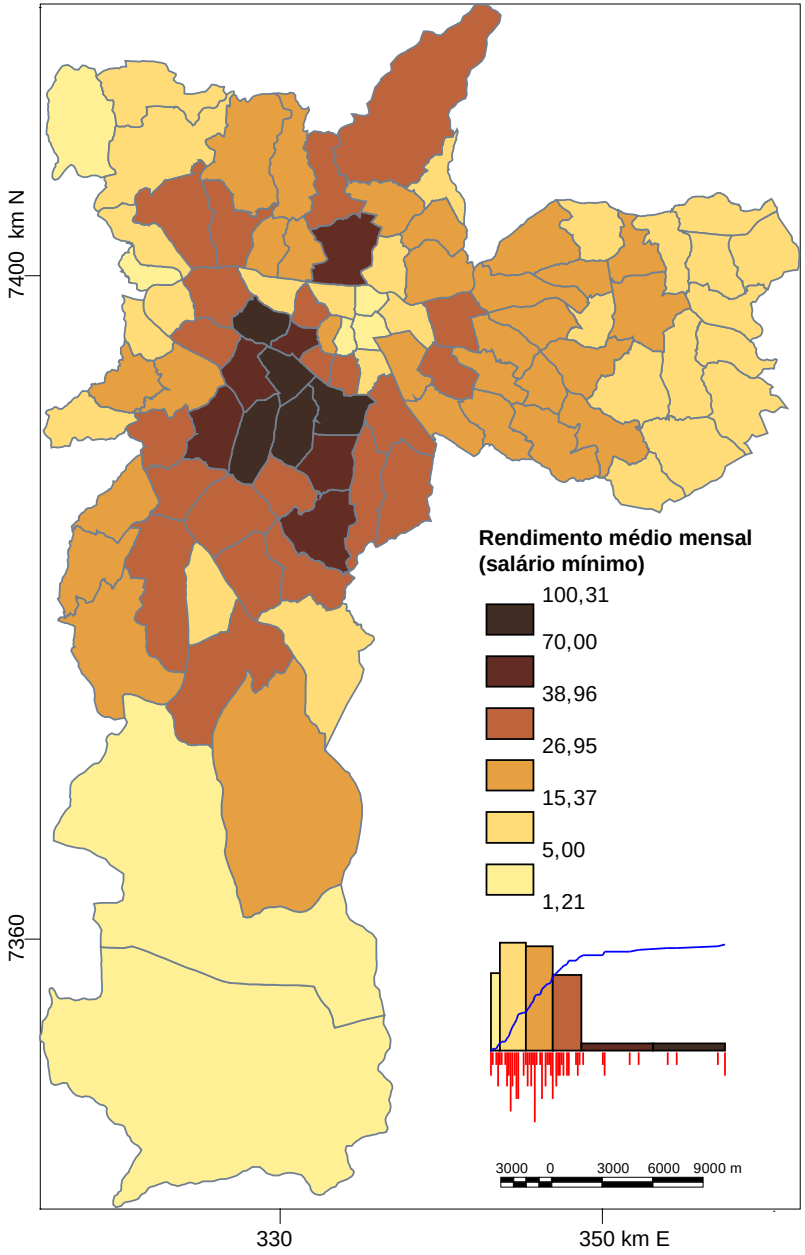
RESEARCH ARTICLE Open Access

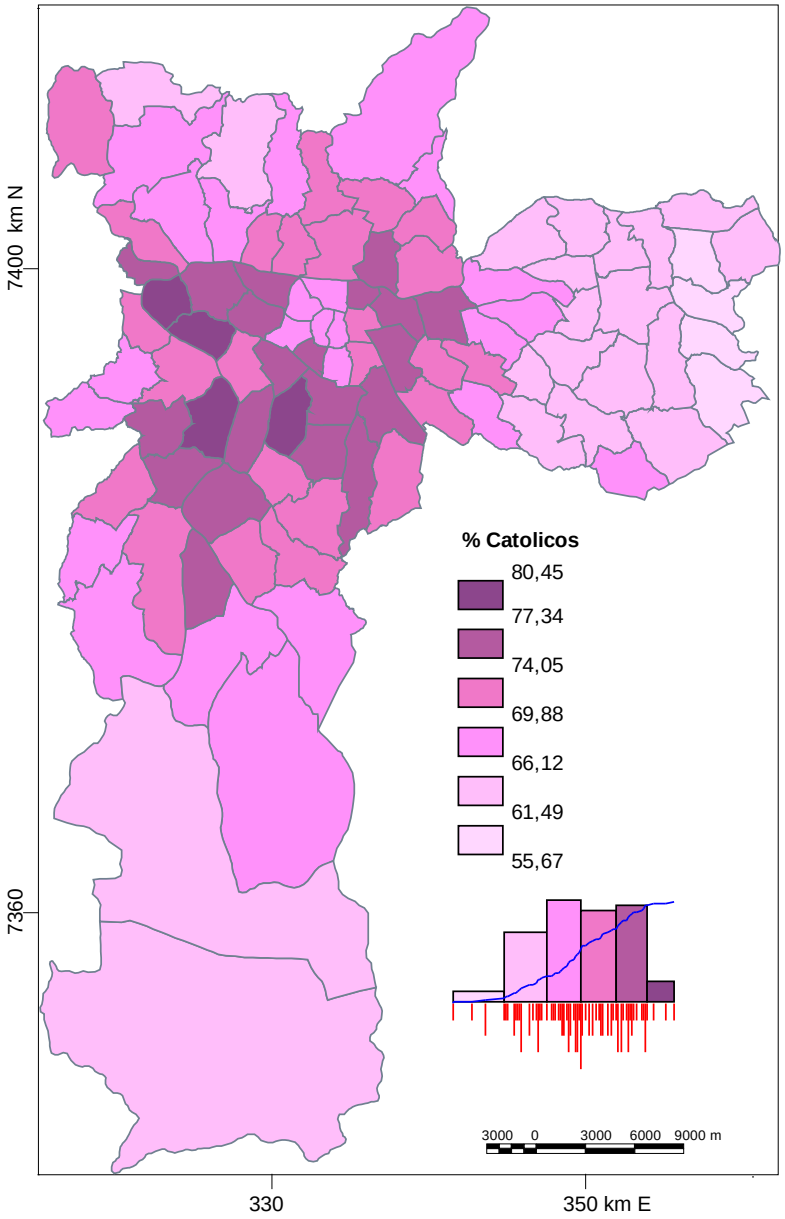
Spatial clusters of suicide in the municipality of São Paulo 1996–2005: an ecological study

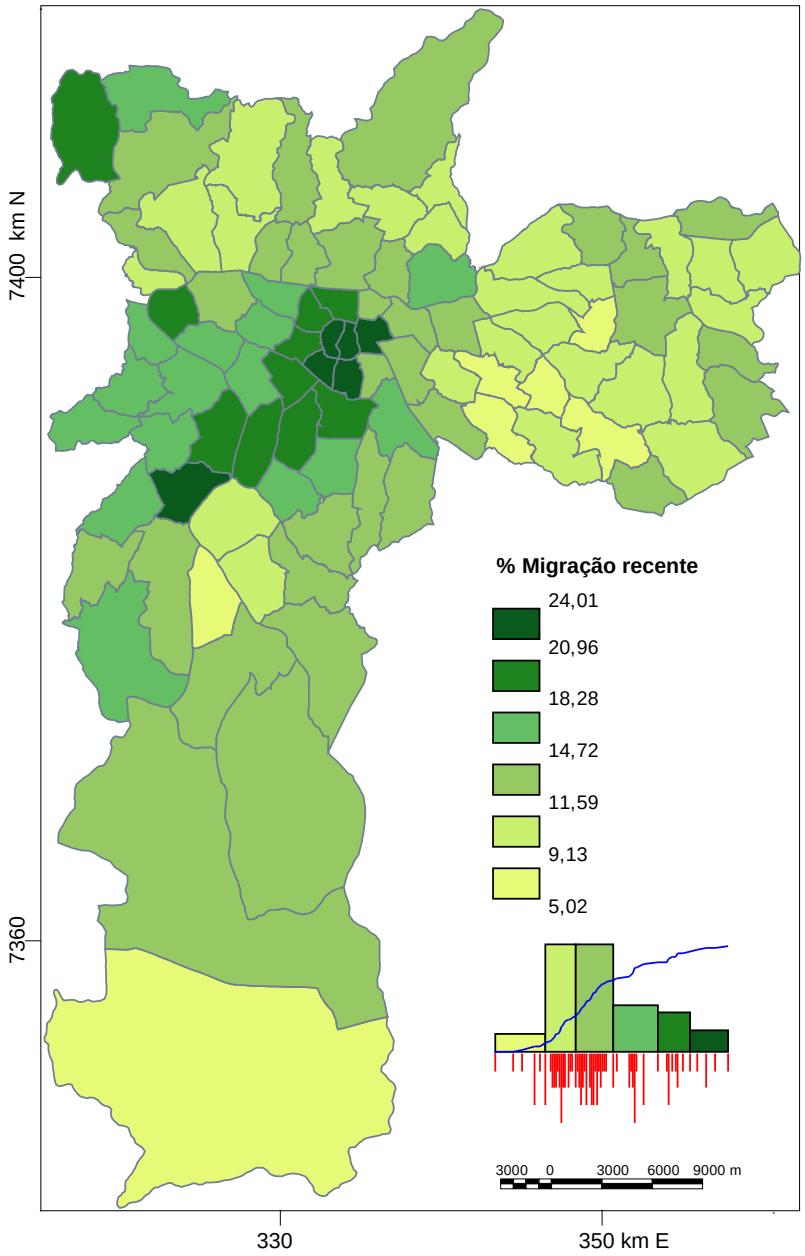
Daniel H Bando[1], Rafael S Moreira[2], Julio CR Pereira[3] and Ligia V Barrozo[4*]


7400 km N
7360
330
350 km E
Taxa de mortalidade
(por 100 mil hab.)
11,29
8,27
6,13
4,65
3,70
2,80
1,19
3000 0 3000 6000 9000 m


m N
7410000
7400000
7390000
7380000
7370000
7360000
7350000
N
não significativo
risco
proteção
320000
330000
340000
350000
360000 m E
0 10000 20000 m


7400 km N
7360
330
350 km E
Rendimento médio mensal
(salário mínimo)
100,31
70,00
38,96
26,95
15,37
5,00
1,21
3000 0 3000 6000 9000 m


7400 km N
7360
330
350 km E
% Catolicos
80,45
77,34
74,05
69,88
66,12
61,49
55,67
3000 0 3000 6000 9000 m


7400 km N
7360
330
350 km E
% Migração recente
24,01
20,96
18,28
14,72
11,59
9,13
5,02
3000 0 3000 6000 9000 m

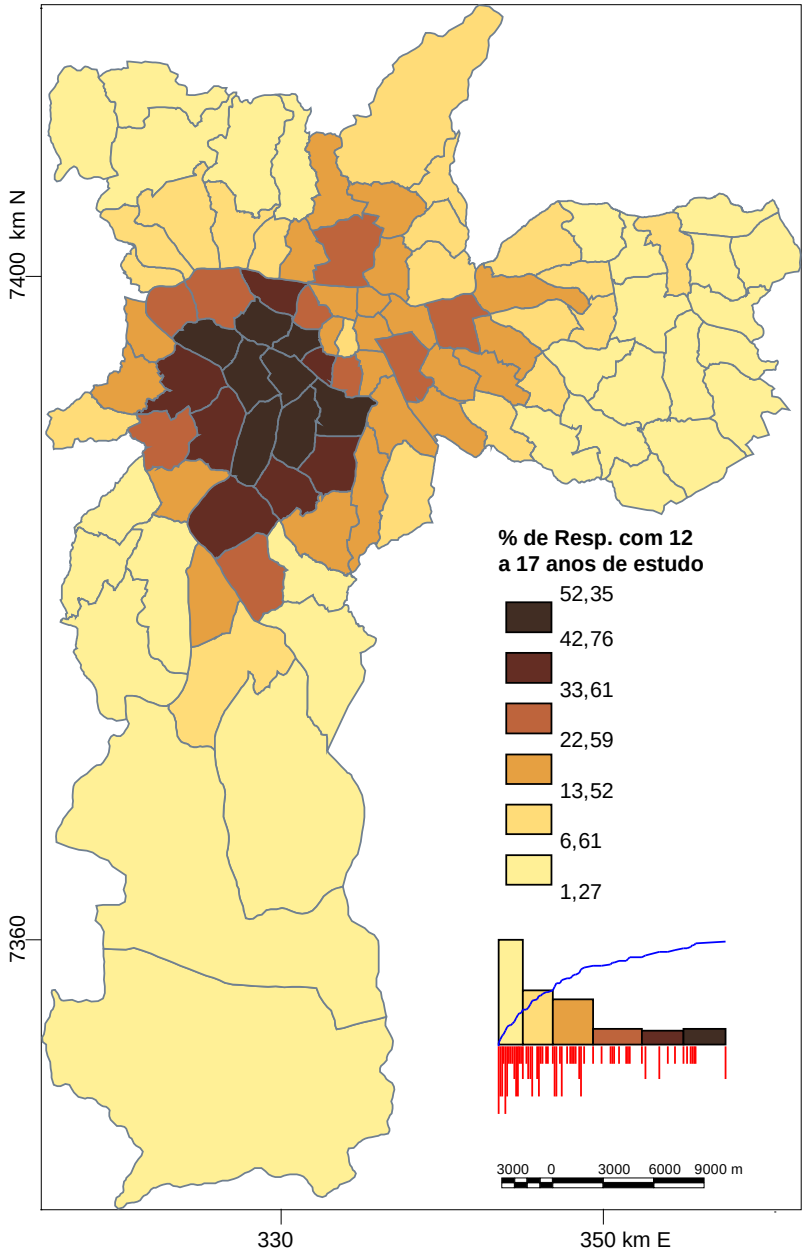

7400 km N
7360
330
350 km E
% de Resp. com 12
a 17 anos de estudo
52,35
42,76
33,61
22,59
13,52
6,61
1,27
3000 0 3000 6000 9000 m

Table 1 Multiple logistic regression measuring effects of risk factors to High Risk Suicide cluster

Variable	β	p	OR	95% CI
Marital status				
Single	0.859	0.031	2.36	1.081 – 5.150
All others			1	
Migrant				
Yes	0.403	0.002	1.497	1.156 – 1.937
No			1	
Religion				
Catholic	0.312	0.034	1.366	1.024 – 1.823
All others			1	
Income (n° of minimal wages)	0.054	0.026	1.056	1.007 – 1.107

Table 2 Multiple logistic regression measuring effects of protective factors to High Risk Suicide cluster

Variable	β	*p*	OR	CI 95%
Marital status				
Married	−0.72	0.003	0.487	0.302 – 0.786
All others			1	
Religion				
Evangelical	−0.507	0.013	0.603	0.404 – 0.899
All others			1	

De acordo com Morgenstern (1982), é importante reconhecer que os estudos ecológicos têm dois principais objetivos:

1) gerar ou testar hipóteses etiológicas, como por exemplo, para explicar a ocorrência da doença,
2) para avaliar a eficiência das intervenções na população.

O maior problema relacionado aos estudos ecológicos, a "falácia ecológica", resulta de fazer inferências *causais* sobre fenômenos individuais com base em observações de grupos. Nos dados de Durkheim, por exemplo, poderiam ter sido os católicos (não protestantes) que tinham cometido suicídio em províncias predominantemente protestantes. Esta explicação alternativa poderia ser possível porque nenhuma das províncias era inteiramente homogênea em relação à religião (MORGENSTERN, 1982).

As inferências causais sobre eventos individuais obtidas a partir de dados agrupados (ecológicos) são limitadas por problemas metodológicos (MORGENSTERN, 1982): vícios de especificação e de agregação, erros nas medidas, ambiguidade de causa e efeito, migração entre grupos e multicolinearidade.

Dadas as vantagens práticas dos dados ecológicos para gerar e testar hipóteses ecológicas, os problemas metodológicos podem ser minimizados:

1) ao se utilizar análise de regressão, no lugar de correlação, incluindo no modelo estatístico tantas variáveis quanto possível,
2) ao se usar dados agrupados na menor unidade geográfica possível, sujeito aos problemas de migração intergrupo e estimativa de taxa instável e,
3) ao se atentar em como os grupos são formados e analisá-los de acordo, o que na prática significa incluir todas as variáveis que possam estar relacionadas ao processo de agrupamento (MORGENSTERN, 1982).

Embora os estudos ecológicos sofram diversas críticas, um forte argumento justifica a contribuição de tais estudos na busca dos fatores de risco ao suicídio. A principal justificativa para a abordagem ecológica é estudar o suicídio no contexto ambiental. Compreender como o contexto afeta a saúde das pessoas e grupos através da seleção, distribuição, interação, adaptação e outras respostas não pode ser atingido por medidas de atributos individuais (SUSSER, 1994). De acordo com Rehkopf e Buka (2005), os fatores de risco associados apenas ao nível individual potencialmente mascaram causas fundamentais mais distantes. Por exemplo, enquanto o consumo de álcool está associado com maior probabilidade de o indivíduo cometer suicídio, o consumo de álcool também está associado com as características da vizinhança incluindo pobreza e densidade de estabelecimentos que vendem bebidas alcoólicas. Concorre nesse sentido a distribuição geográfica ímpar do suicídio que persiste ao longo do tempo, em diferentes países e dentro dos países, dando suporte às correlações ao nível de área.

Mas como o local de residência poderia influir nas taxas de suicídio? De acordo com Agerbo et al. (2006), há duas explicações principais para que as associações ocorram nestes estudos. As associações ecológicas podem resultar de altas concentrações de indivíduos em risco morando em áreas particulares (efeitos composicionais) e/ou do fato dos lugares de residência terem um efeito real nas taxas de suicídio (efeito contextual).

O efeito contextual pode não apenas influenciar as taxas totais de suicídio em localidades geográficas particulares, mas podem ter também efeitos diferenciais em residentes particulares que dependem de suas características pessoais, como seu estado civil ou classe social. Esses efeitos são chamados de "*cross-level effects*" (AGERBO et al., 2006).

EFEITOS DE ESCALA EM UNIDADES DE ÁREA

- Dados do Censo de BH para o ano de 1991, em duas escalas: setores censitários (1998 setores) e unidades de planejamento (UP's) (80 divisões)
- Os 1998 registros de setores censitários foram agregados em 80 unidades de planejamento
- Foram computadas 1000 correlações entre 40 pares de variáveis, utilizando os dados agrupados em setores e depois por UP

EFEITOS DE ESCALA EM UNIDADES DE ÁREA

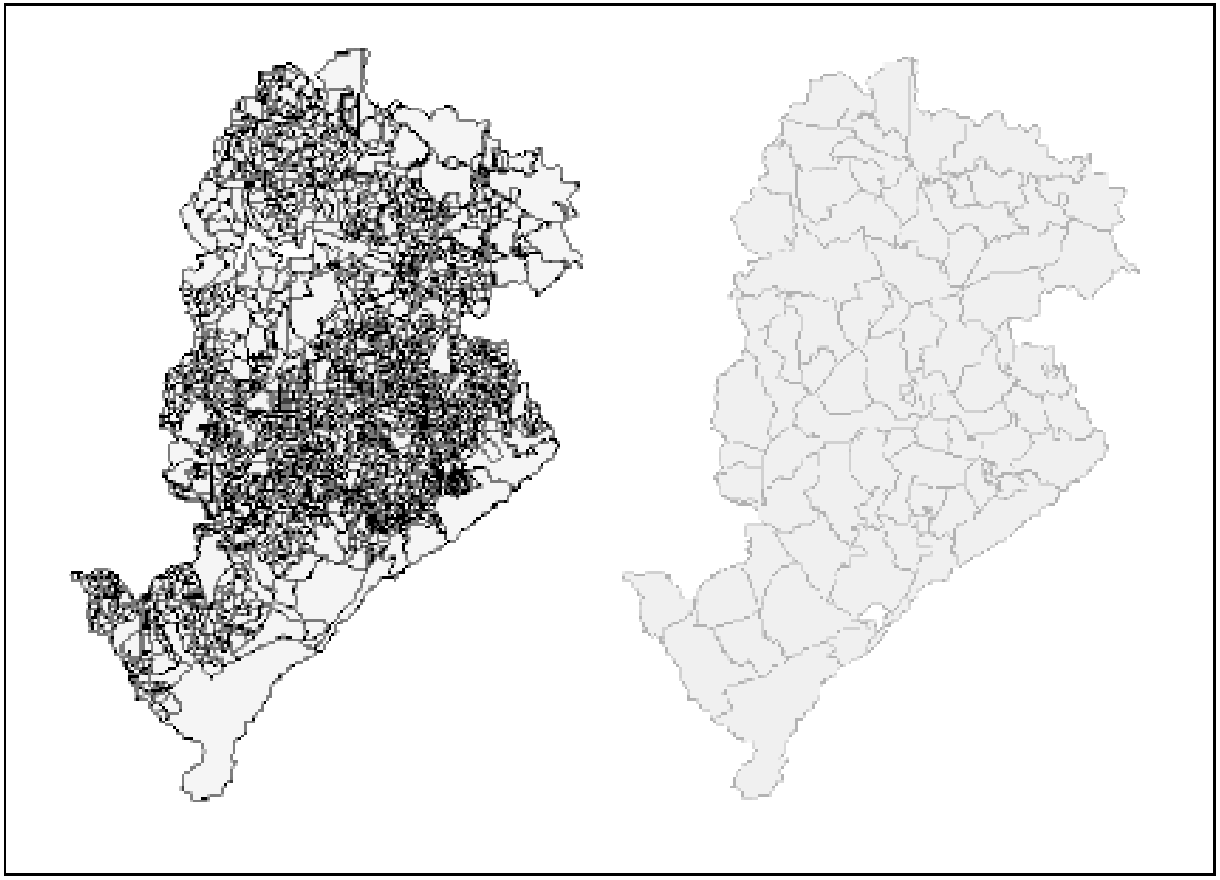


Figura 2 – Setores censitários (à esquerda) e Unidades de Planejamento (à direita) para o município de Belo Horizonte.

EFEITOS DE ESCALA EM UNIDADES DE ÁREA

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

EFEITOS DE ESCALA EM UNIDADES DE ÁREA

- Os resultados indicam que as correlações nos setores censitários são significativamente menores que as correlações por unidades de planejamento

- 802 correlações são menores para os setores censitários que para as Ups; apenas 40 têm o comportamento oposto

EFEITOS DE ESCALA EM UNIDADES DE ÁREA

- Teoricamente seria possível utilizar os dados individuais de coleta (ou amostra) para realizar correções nas correlações agregadas
- O problema de escala é um efeito inerente aos dados agregados por áreas. Não pode ser removido ou ignorado
- Para minimizar o impacto deve-se procurar utilizar a melhor escala de levantamento de dados disponível e tentar agregar os dados, de acordo com critérios relevantes para o fenômeno

EFEITOS DE ESCALA EM UNIDADES DE ÁREA

- Não existe uma escala certa, existem dois modelos:

  correlações mais fracas e maior flutuação aleatória, porém com mais homogeneidade interna

  correlações mais fortes com o viés ocasionado por desconsiderar a dispersão e a heterogeneidade em torno da média nas grandes áreas

- Quanto mais desagregado o dado, maior a flexibilidade na escolha de modelos (desagregar é impossível)

Estimação de taxas em áreas com pequenas populações

- Utilizar a melhor resolução espacial disponível requer tratamento adicional nos dados (universo populacional reduzido)

Estimação de taxas em áreas com pequenas populações

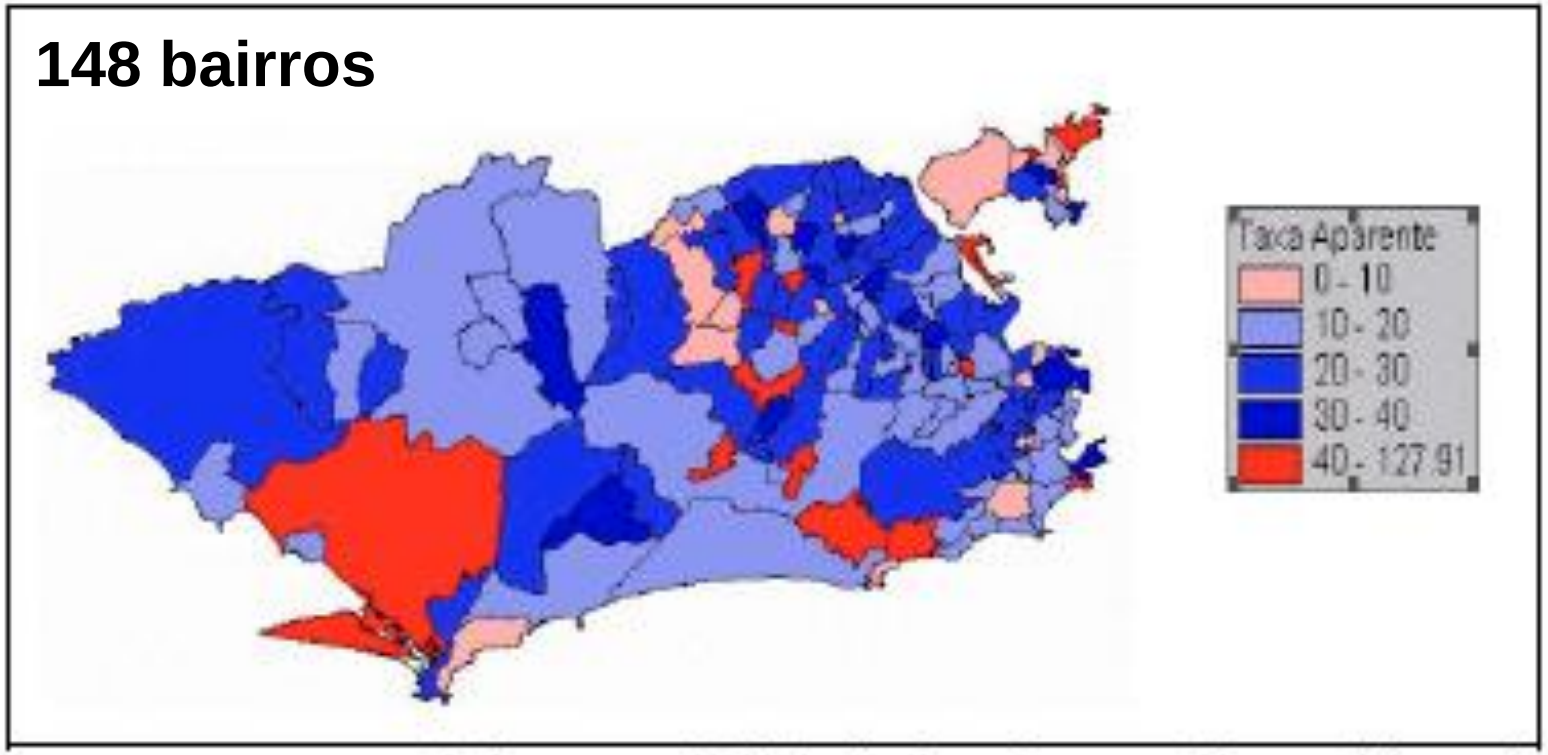


Figura 3– Taxa total de mortalidade infantil por mil nascidos vivos no Rio de Janeiro, em 1994.

valores extremos podem ser flutuações aleatórias!!

Estimação de taxas em áreas com pequenas populações

"efeito funil"

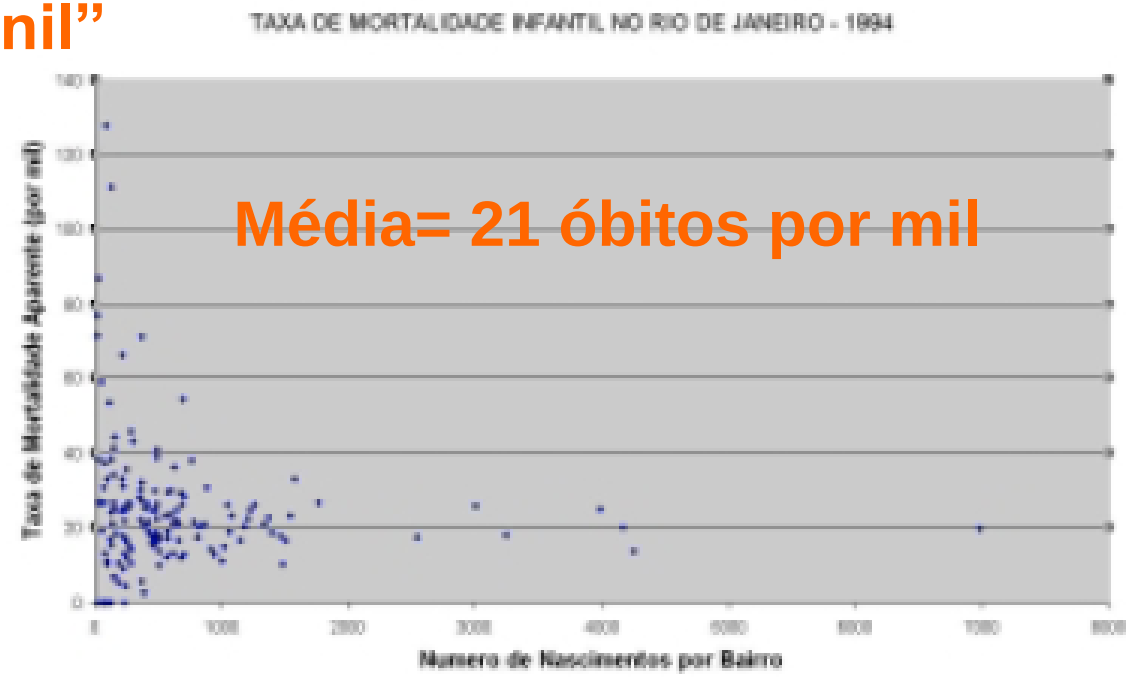


Figura 4– Taxa de mortalidade infantil no Rio de Janeiro em 1994 em função do número de nascimentos por bairro.

taxa em função da população em risco

Estimação de taxas em áreas com pequenas populações

- É razoável supor que as taxas das diferentes regiões estão autocorrelacionadas e poderia ser levado em conta o comportamento dos vizinhos para estimar uma taxa mais realista para as regiões de menor população
- Técnicas de estimação bayesiana

Estimação de taxas em áreas com pequenas populações

- A taxa "real" associada a cada área não é conhecida
- Dispõe-se de uma taxa observada

$$t_i = z_i / n_i$$

z_i é o número de eventos na i-ésima área

n_i é o número de pessoas observadas

Estimação de taxas em áreas com pequenas populações

- A idéia do estimador bayesiano é supor que a taxa "real" é uma variável aleatória, que possui uma média e uma variância

- O melhor estimador bayesiano é dado por uma combinação linear entre a taxa observada e a média

Estimador bayesiano

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

Fator *w*

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

w_i é o peso e 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 da média. Regiões com populações muito baixas terão uma correção maior e regiões populosas terão pouca alteração em suas taxas

Estimador bayesiano empírico

- A formulação bayesiana requer as médias e variâncias para cada uma das áreas
- este estimador parte da hipótese que a distribuição da variável aleatória é a mesma para todas as áreas; todas as médias e variâncias são iguais
- Pode-se estimar a média e a variância diretamente a partir dos dados

Medidas de tendência central

Média aritmética

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Soma de todos os valores (x) de um conjunto de dados dividida pelo número de valores (n)

Medidas de variabilidade ou dispersão – variância

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

Como os desvios são elevados ao quadrado, a variância é expressa em unidades quadradas

Média a partir das taxas observadas

$$\hat{\mu} = \frac{\sum y_i}{\sum n_i}$$

Variância a partir da variância das taxas observadas com relação à média estimada

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

Estimador bayesiano empírico

- 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

Estimador bayesiano empírico

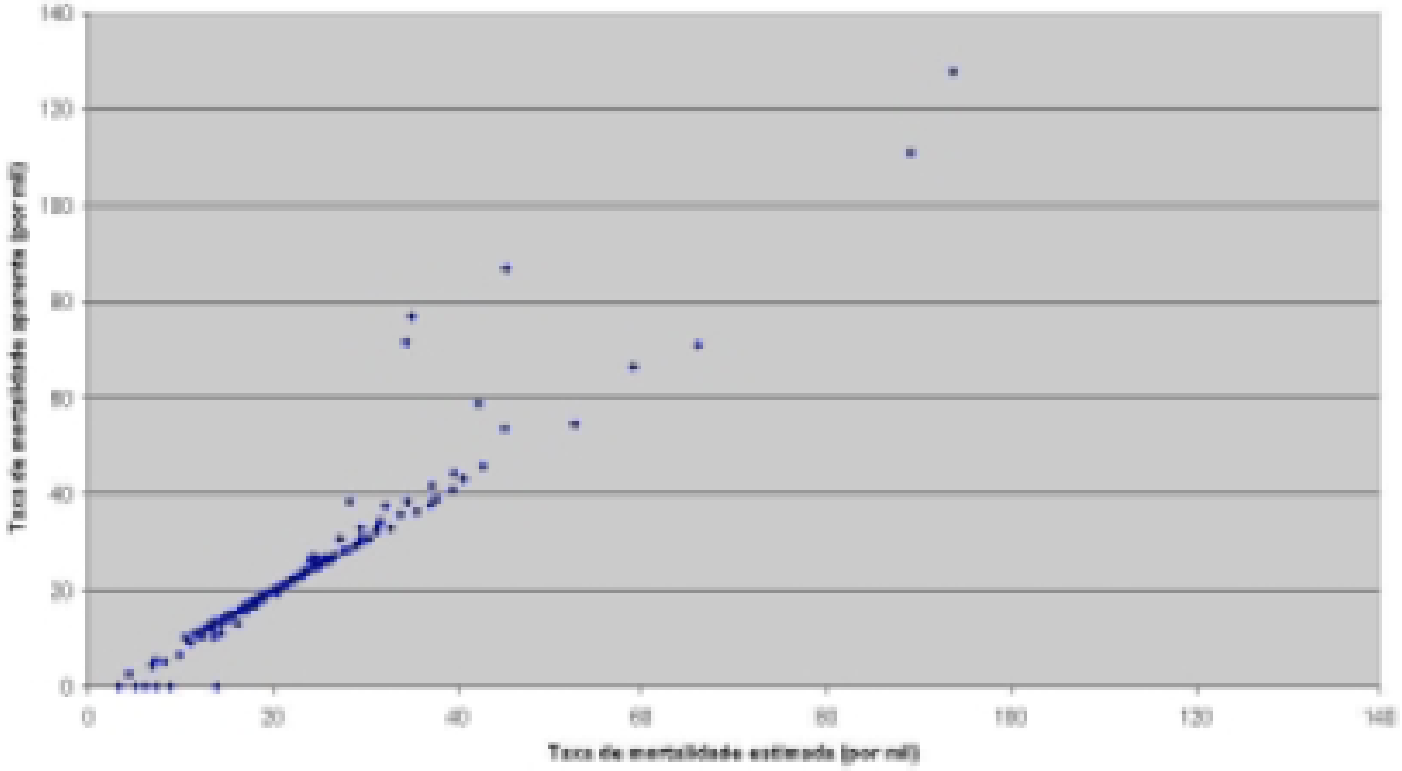


Figura 5– Comparação entre a taxa de mortalidade infantil observada e a taxa estimada pelo método *bayesiano* empírico.

Estimador bayesiano empírico local

- **O estimador bayesiano empírico pode ser generalizado para incluir efeitos espaciais. Pode-se fazer a estimativa bayesiana localmente, convergindo em direção a uma média local e não global (considerando como "região" a sua vizinhança)**

Estimador bayesiano empírico local

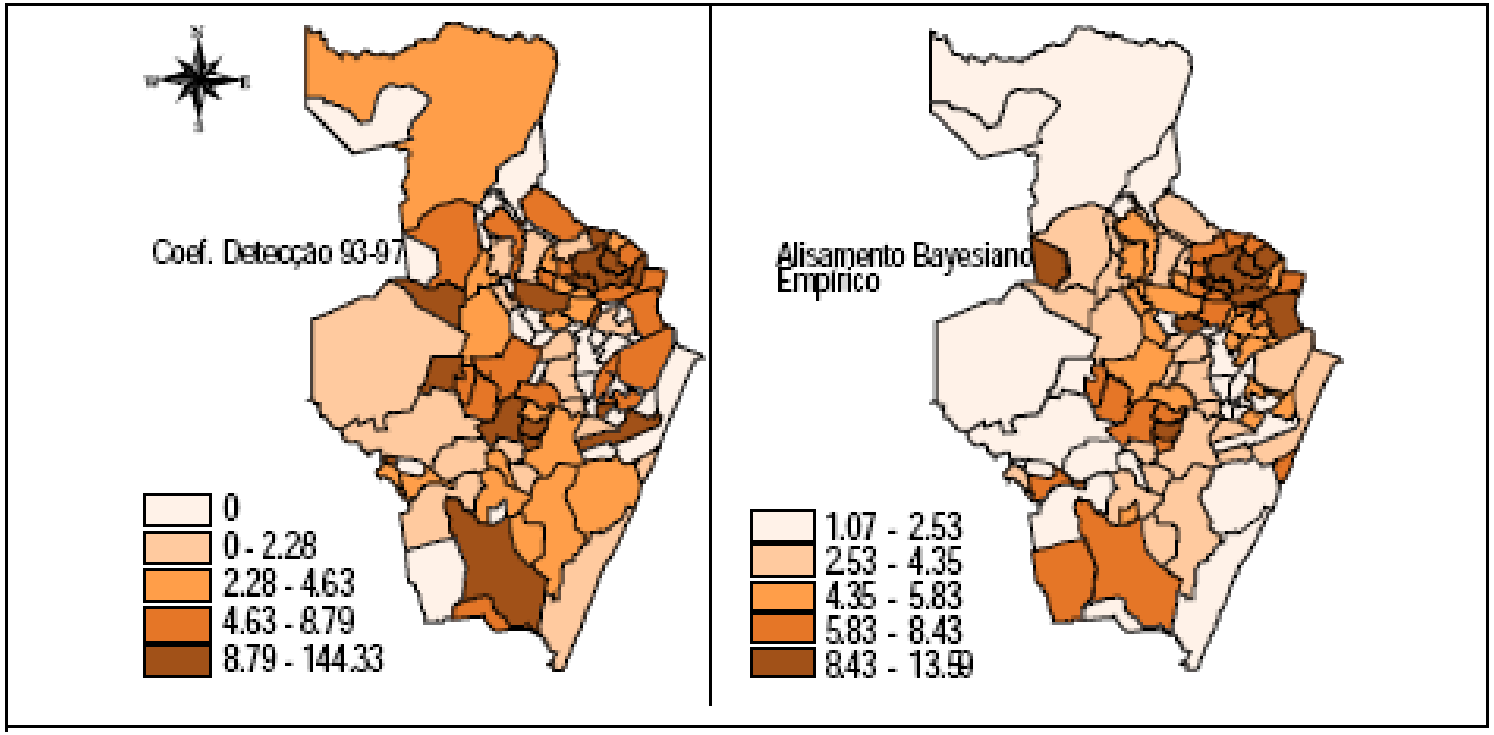


Figura 6- Taxas de detecção média de hanseníase em menores de 15 anos, período 1993-1997, por bairro do Recife, e taxas estimada através de alisamento bayesiano.

Estimador bayesiano empírico local

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

Referências bibliográficas

AGERBO, E., STERNE, J.A.C., GUNNELL, D.J. 2006. Combining individual and ecological data to determine compositional and contextual socio-economic risk factors for suicide. ***Social Science & Medicine***. doi:10.1016/j.socscimed.2006.08.043.

MORGENSTERN, H. Uses of ecologic analysis in epidemiologic research. ***American Journal of Public Health***, 1982, 72(12):1336-1344.

SUSSER M: The logic in ecological: I The logic of analysis. ***Am J Public Health*** 1994, 84:825–829