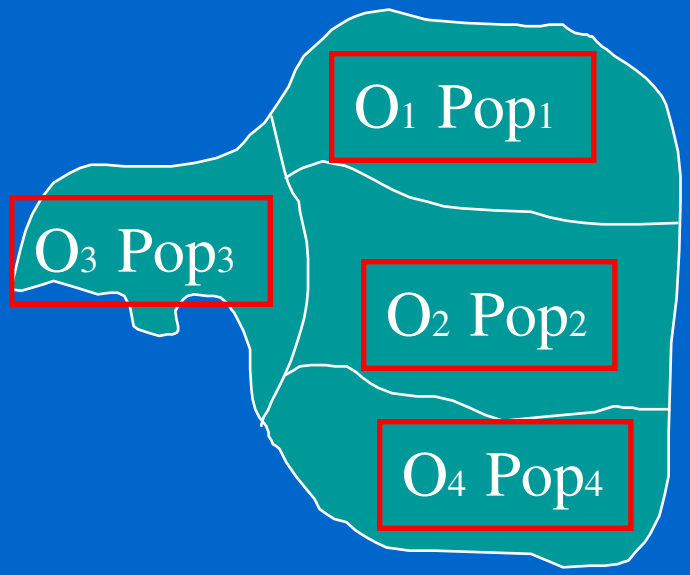
Estatística Espacial: Dados de Área

- Distribuição do número observado de eventos
- Padronização e SMR
- Mapas de Probabilidades
- Mapas com taxas empíricas bayesianas

Padronização

- Para permitir comparações entre diferentes populações no espaço ou no tempo, variáveis devem ser padronizadas.
- Padronizar as população de risco por tamanho, estrutura etária e sexo é o mais comum.
- Padronização pode ser também por área, por tempo de exposição, etc.

Padronizando os tamanhos de população



- i = índice das áreas

  Em cada área i :
- O_i = número de eventos em i
- Pop_i = pop sob risco em i
- $r_i = O_i / Pop_i$ = taxa em i
- Às vezes, usa-se $t_i = 100000 * r_i$, taxa por 100 mil em i

Taxa de Morbidade Padronizada

- Terminologia: região é composta de áreas
- É comum trabalhar com medidas de risco relativo. Mas relativo a quê ??? Relativo ao padrão global da região
- Em inglês: standardized mortality ratio (SMR)
- Em cada área *i*, calcular o número esperado de eventos caso risco na área *i* seja igual ao risco na região total

  $$E_i = Pop_i * r$$

  onde $r = \Sigma O_i / \Sigma Pop_i$

  Isto é, r = número total de eventos na região dividido pela população total na região
- Compare o número observado de eventos em *i* com o número esperado: $SMR = O_i / E_i$
- É comum SMR ser multiplicada por 100

Aleatoriedade da contagem de eventos

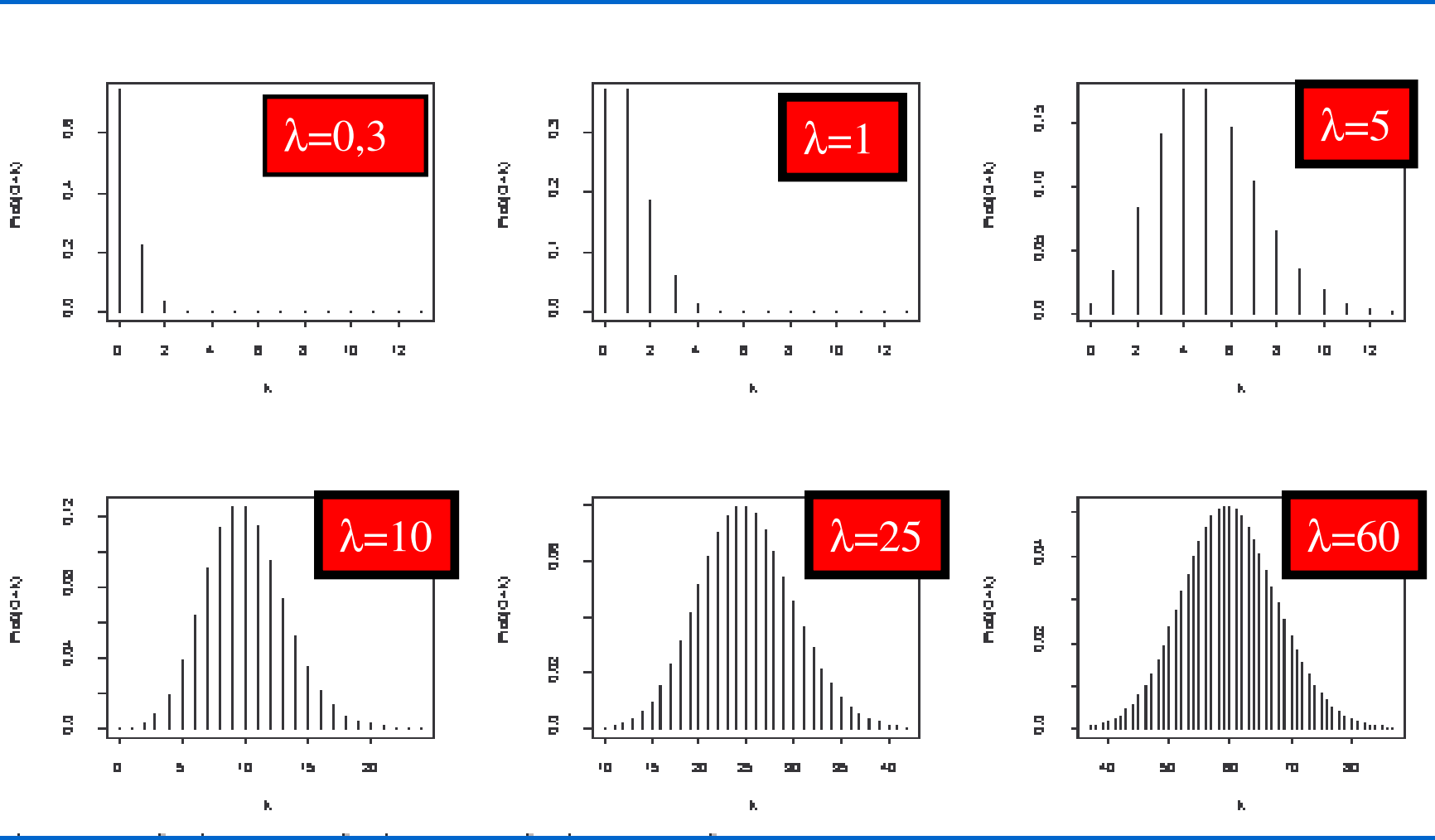
- Número observado O_i de eventos na área *i* é variável aleatória
- Isto implica que O_i possui distribuição de probabilidade, valor esperado, variância, etc.
- Hipótese comum: $O_i \sim \text{Poisson}(\lambda_i)$

  λ_i = número esperado na área *i*

Se risco é constante na região, então $\lambda_i = E_i$

onde $E_i = r * Pop_i$

Função de Probabilidade da Poisson



Padronizar a estrutura etária/sexo das populações

- Quando risco varia com idade ou sexo, não é suficiente apenas a padronização do tamanho da população
- Exemplos: Mortes violentas, câncer, doenças cardíacas, AIDS, etc
- i = índice da área

  j = índice da classe de idade-sexo

  $j=1$ indica MASC de 0 a 4 anos de idade

  $j=2$ indica MASC de 5 a 9 anos de idade, etc...

Número esperado de eventos na área *i* e classe *j*

- Fixar atenção numa classe *j*. Por exemplo, *j=5* que significa homens de 20 a 24 anos de idade
- O_{ij} = número de eventos que ocorreram entre pessoas da classe *j* na área *i*
- A taxa global na classe de idade-sexo *j* é dada por

$$r_j = \frac{\sum_i O_{ij}}{\sum_i Pop_{ij}} = \frac{\text{total de eventos na classe } j}{\text{população total na classe } j}$$

- Então $E_{ij} = Pop_{ij} * r_j$ = número esperado se risco na classe *j* fosse constante no espaço

SMR padronizada por idade/sexo

- Número esperado na área *i* se risco é constante no *espaço* é dado por $E_i = \sum_j E_{ij}$ = soma dos números esperados nas classes de idade-sexo
- Comparar número observado com o esperado: $SMR = \frac{O_i}{E_i}$
- Hipótese comum sobre a distribuição de probabilidade dos valores observados: $O_i \sim \text{Poisson}(\lambda_i)$
- Hipótese adicional de risco constante no espaço em cada classe de idade-sexo $\rightarrow \lambda_i = E_i$ onde E_i como acima
- $O_i \sim \text{Poisson}(E_i)$ hipótese de risco constante por indivíduo

Mapas de SMR

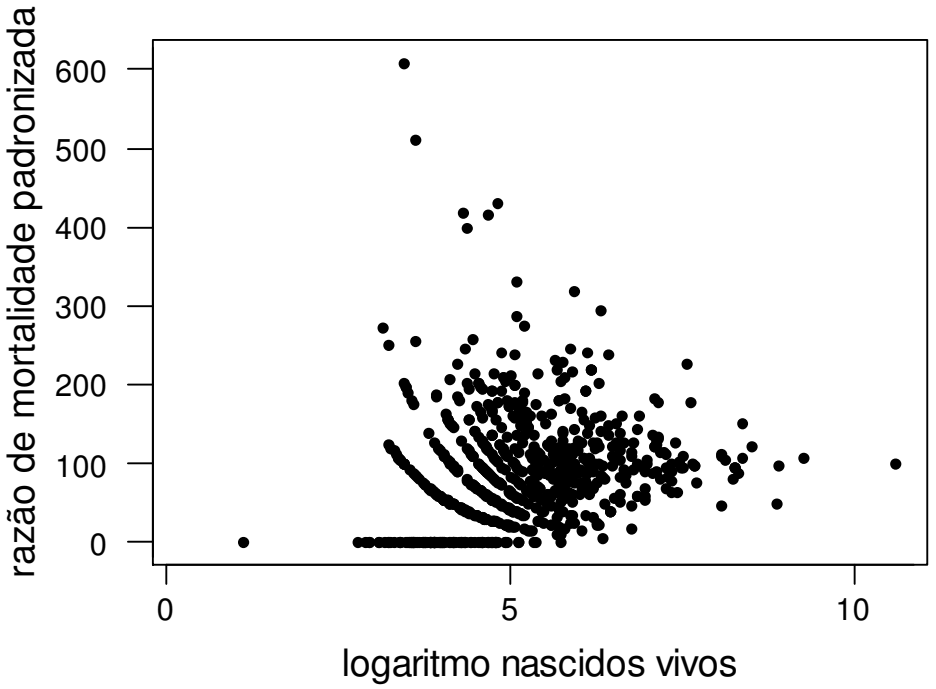
- Vantagens: escala de risco RELATIVO permite comparar diferentes variáveis se adotamos uma escala escala única em vários mapas. Por exemplo, taxas de morbidade de doenças com incidência muito diferentes.
- Padronização direta pode ser usada (com população mundial como padrão) quando queremos oferecer o estudo para possíveis comparações internacionais. Estas taxas tendem a ter mais variância.

Problemas de Estimação em Áreas Pequenas

- Valores extremos ocorrem nas áreas com pequenas populações
- O que mais chama a atenção num mapa (os valores extremos), é o menos confiável !
- As maiores oscilações não estarão, em geral, associadas com variações no risco subjacente; serão apenas flutuação aleatória casual.

Mortalidade Infantil em MG

- Taxas municipais em Minas Gerais em 1994.
- Existiam 756 municípios em MG.
- RMP variam de 0 a 600 !
- Observe a forma de funil.



EFEITO DA INSTABILIDADE

- Exemplo de mortalidade infantil por município em MG
- 15 municípios com: 0 mortes e < 30 nascidos vivos.
- Se uma única morte é registrada, taxas passam de 0 para valores entre 116 e 1048!!!
- O valor extremo anterior era 608.9

Como resolver ?

- Agregar áreas para formar áreas maiores. Desvantagem é perder informação localizada.
- Mapas de probabilidade (a seguir)
- Estimar melhor o risco localizado de uma área *i*. Pode-se obter grande redução do problema abordagens bayesianas.
- Abordagens bayesianas:
  - empírica: fácil de implementar
  - puramente bayesiana: preferível mas requer mais esforço computacional.
- Só veremos abordagem bayesiana empírica

Mapas de probabilidade

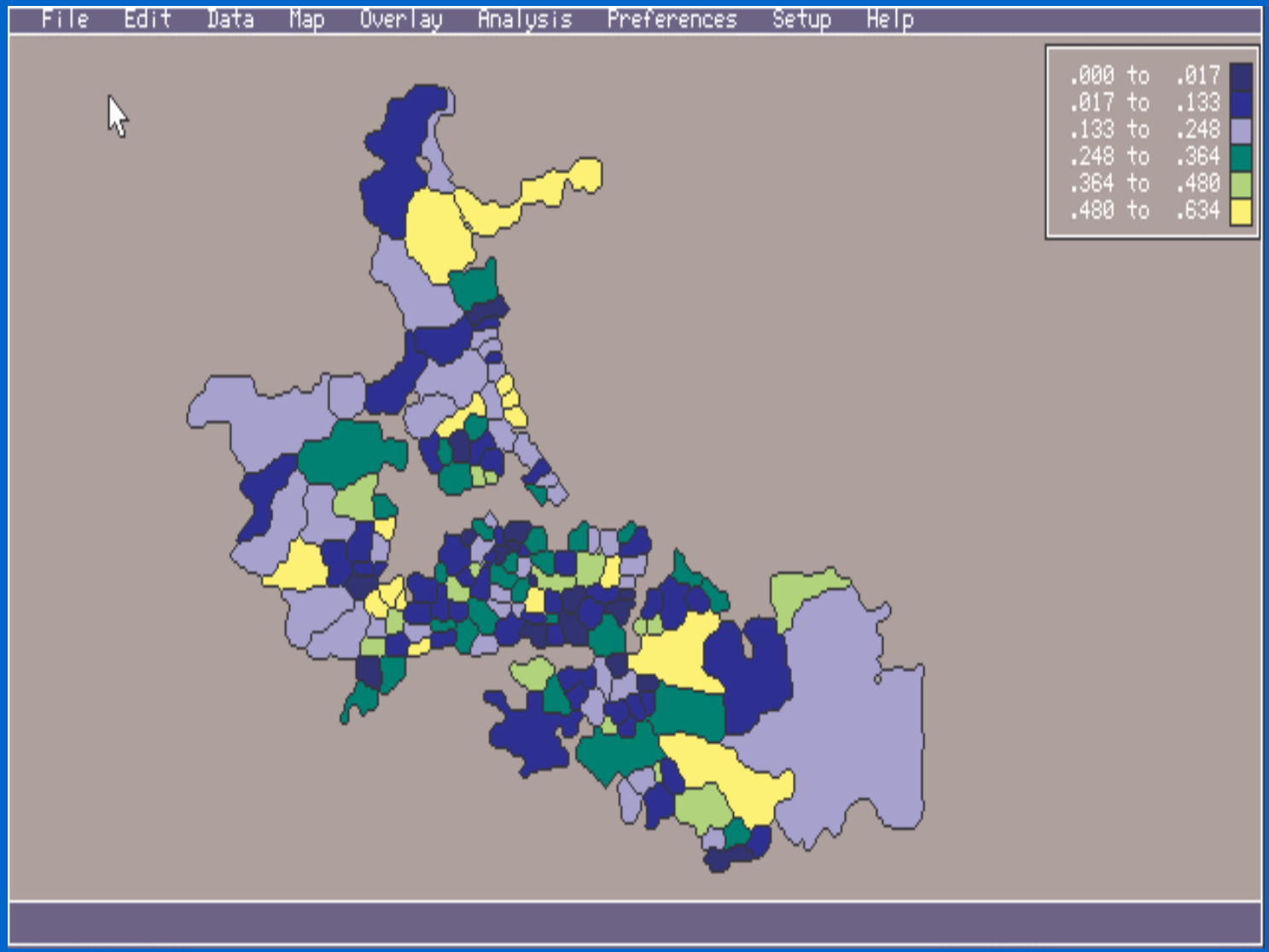
- Choynowski (JASA, 1959) propôs fazer mapas com P-valores:
- Suponha O_i com distribuição Poisson(E_i) (risco constante no espaço). Defina então:

  $\rho_i = P(X \geq O_i)$ se $O_i \geq E_i$

  $P(X \leq O_i)$ se $O_i < E_i$

  onde X tem distribuição Poisson(E_i).
- $\rho_i \approx 0$ indica taxa muito alta ou muito baixa.

Mapa de probabilidades com mortalidade infantil em Auckland, NZ



Problemas

- P-valores muito próximos de zero se N_i é muito grande (mais a seguir).
- Não levam em conta similaridade espacial.
- Não são medidas fáceis de interpretar.

P-valor é influenciado por E_i

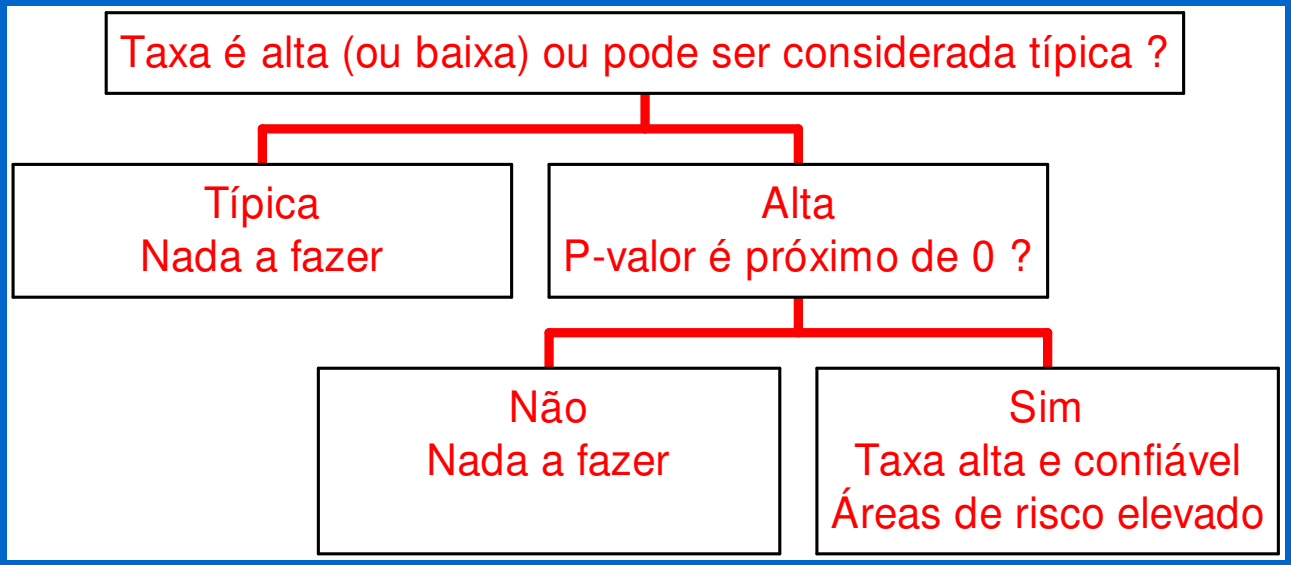
- P-valor é influenciado ao revés pelo tamanho da população de risco! Pequenas diferenças entre O_i e E_i terão `pi` próximo de zero se E_i é grande.
- Exemplo: Com $E_i = 1000$, se $O_i > 1052$ então `pi` < 0.05. No entanto, a diferença entre esses valores é de apenas 5,2% do valor esperado.
- Explicação mais técnica: Suponha que $O_i - E_i = 1.05\, E_i$. Usando Teorema Central do Limite,

$$P(X > O_i) = P\left(\frac{X - E_i}{\sqrt{E_i}} > \frac{O_i - E_i}{\sqrt{E_i}}\right) \approx P\left(N(0,1) > \frac{1.05\, E_i}{\sqrt{E_i}}\right) = P\left(N(0,1) > 1.05\sqrt{E_i}\right) \to 0$$

- quando $E_i \to \infty$

Como usar o P-valor ?

- O melhor é NÃO fazer mapas de probabilidades baseadas nos ρ_i
- Um procedimento possível em duas etapas:

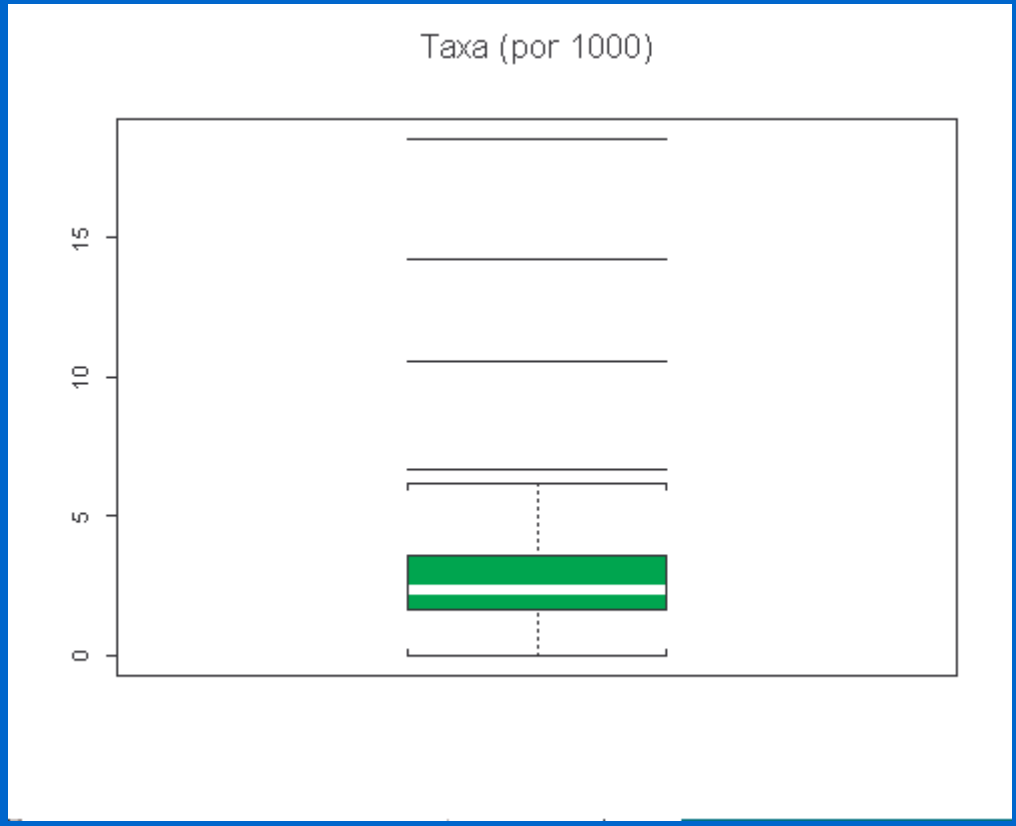


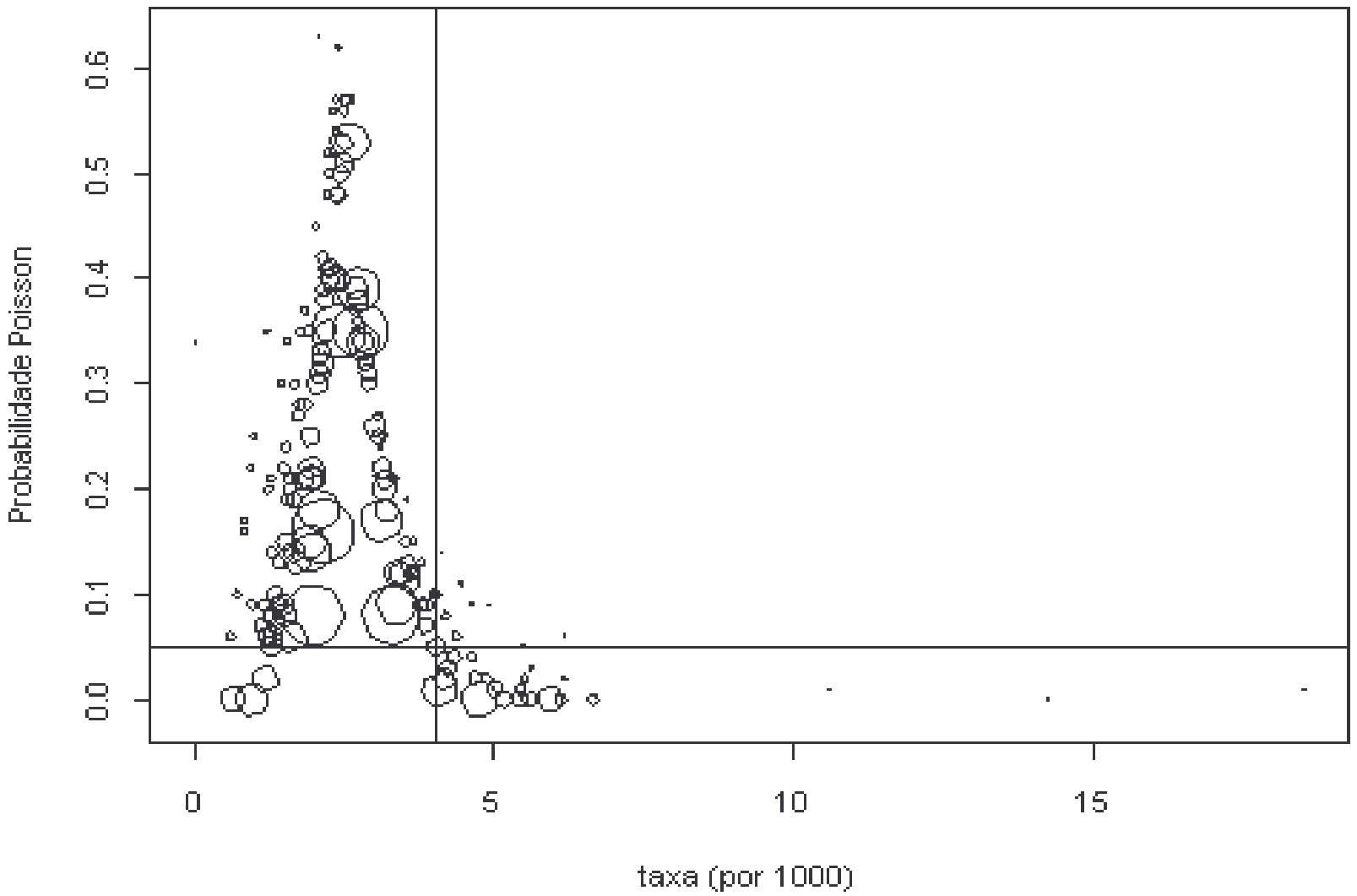
Voltando ao exemplo da NZ

Mediana = 2.4 (por 1000)
percentil 75% = 3.6
percentil 80% = 4.0

Taxa “alta” é > 4.0

Destas áreas, quais possuem p-valor < 0.05 ?




Probabilidade Poisson
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
5
10
15
taxa (por 1000)

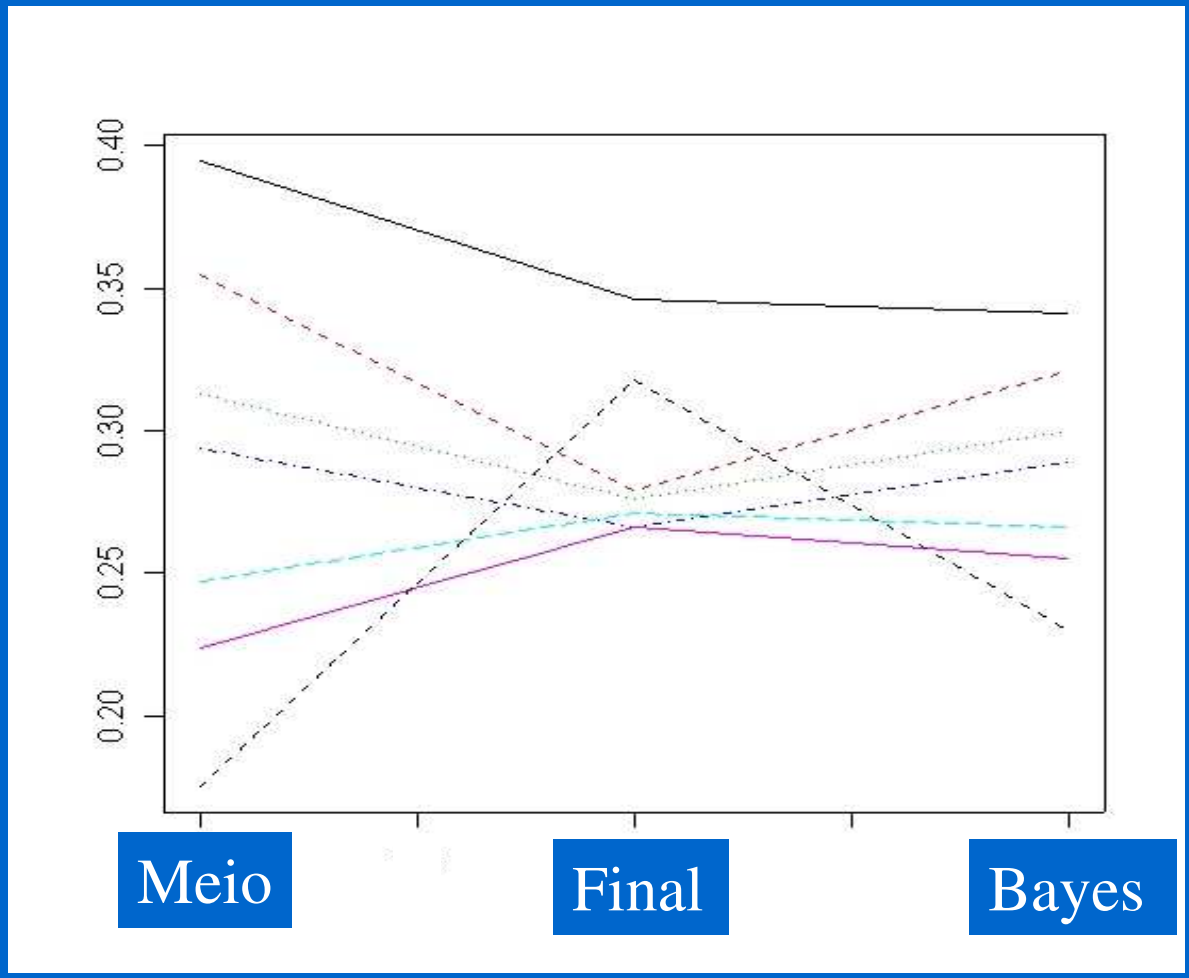
Abordagem Bayesiana Empírica

- Histórico em atlas: Manton, Tsutakawa, etc.
- Assumir que riscos das diferentes áreas não são totalmente "desconectados" e assim pedir uma força pros vizinhos (to borrow strength from the neighbors)
- Idéia: contrair taxa em direção à média global. Fator de contração depende da população da área.

Exemplo Clássico: Efron

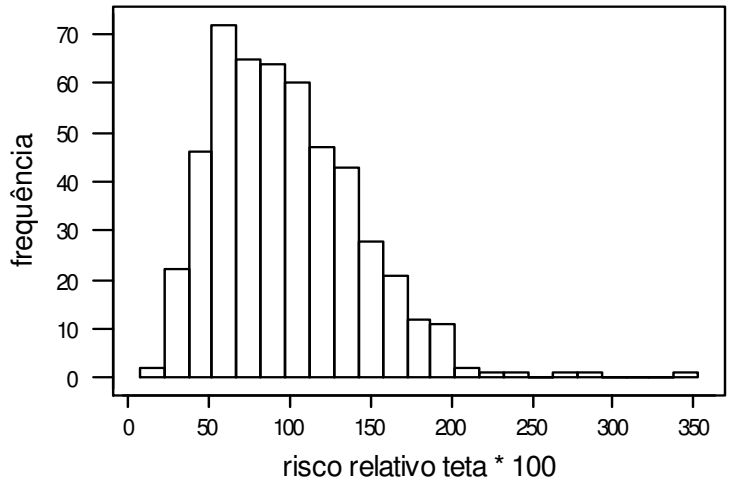
- Dados de 18 batedores de baseball: nº de vezes no taco e nº de acertos
- Dados até a metade do campeonato.
- Qual a melhor predição para o desempenho final de cada batedor?
- Simplesmente a proporção de acertos, certo?
- Ou não? Usar todos os outros jogadores para estimar um dado jogador.
- Mas como jogador A pode ajudar a prever o desempenho do jogador B??

Estimação e resultado final



Proposta de Marshall (1991)

- Fácil de ser implementada (pode usar excel) e produz resultados similares ao de métodos mais sofisticados
- Idéia: cada área *i* possui um taxa subjacente (por 1) θ_i desconhecida. Embora diferentes, estas taxas possuem certa estrutura.
- Se pudéssemos fazer um histograma desses riscos subjacentes, deveríamos observar algo semelhante a quê ?



Objetivo: recuperar θ

- Numa área, observa-se um número aleatório O_i de casos.
- NÃO assumimos risco constante: O_i tem distribuição de Poisson com número esperado de casos igual a $Pop_i\ \theta_i$
- Assume-se que as taxas θ_i possuem distribuição com média *m* e variância *V*.
- Qual é a melhor estimativa $\hat{\theta}_i$ possível dos θ_i ? Melhor em que sentido ?
- Melhor no sentido de minimizar a soma dos erros de estimação de todas as áreas: $\sum_i \left(\hat{\theta}_i - \theta_i\right)^2$

Simplificar o problema

- Buscar estimativa ótima APENAS DENTRE as estimativas que podem ser escritas como médias ponderadas de *m* e da taxa observada na área *i*
- *Solução:*

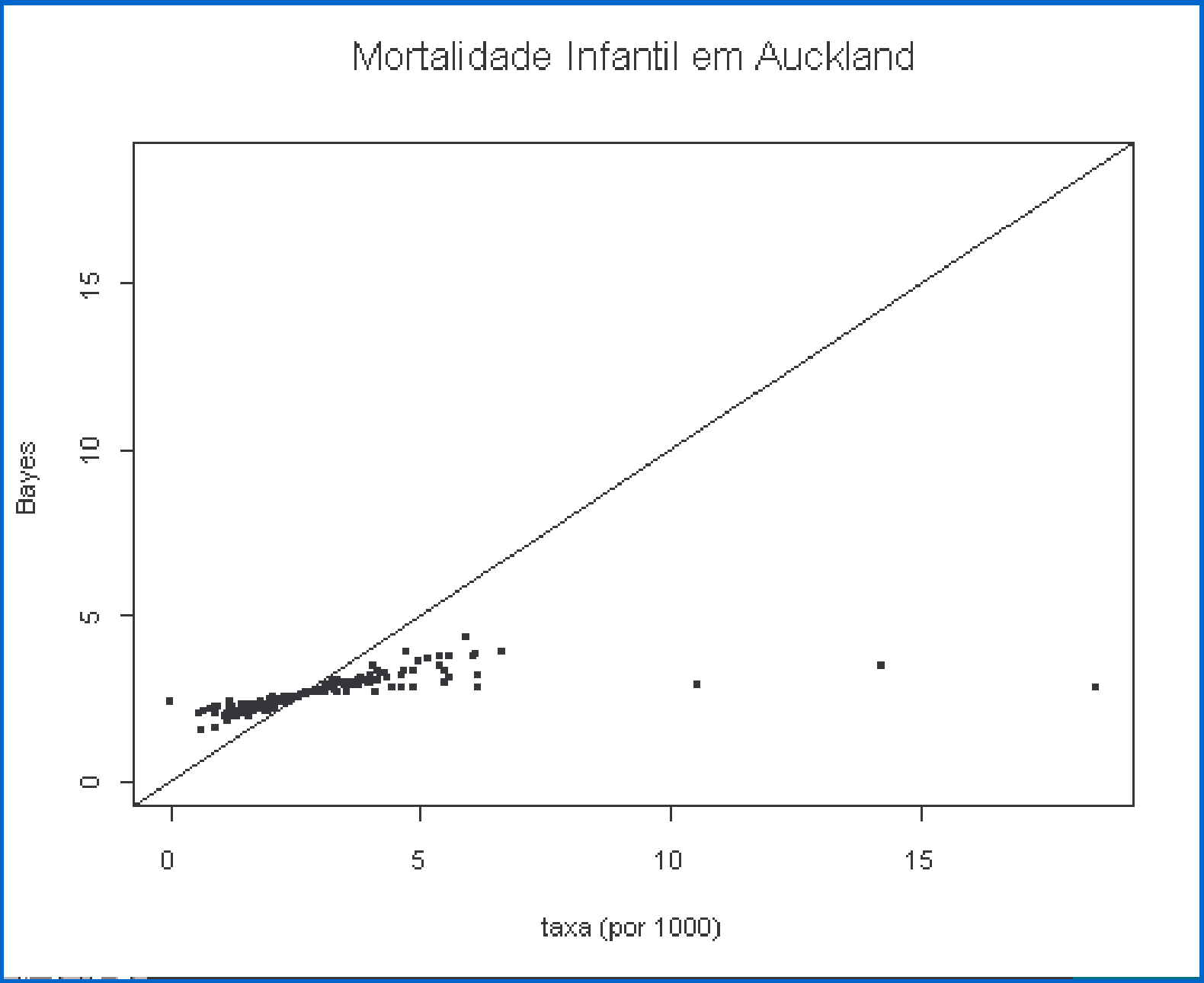
$$\hat{\theta}_i = w_i\, r_i + (1 - w_i)m \quad \text{onde} \quad w_i = \frac{V}{V + \dfrac{m}{Pop_i}}$$

- *Problema: V* e *m* não são conhecidos.
- Bayes empírico *estima* estes valores a partir dos dados (daí vem o nome *empírico*)

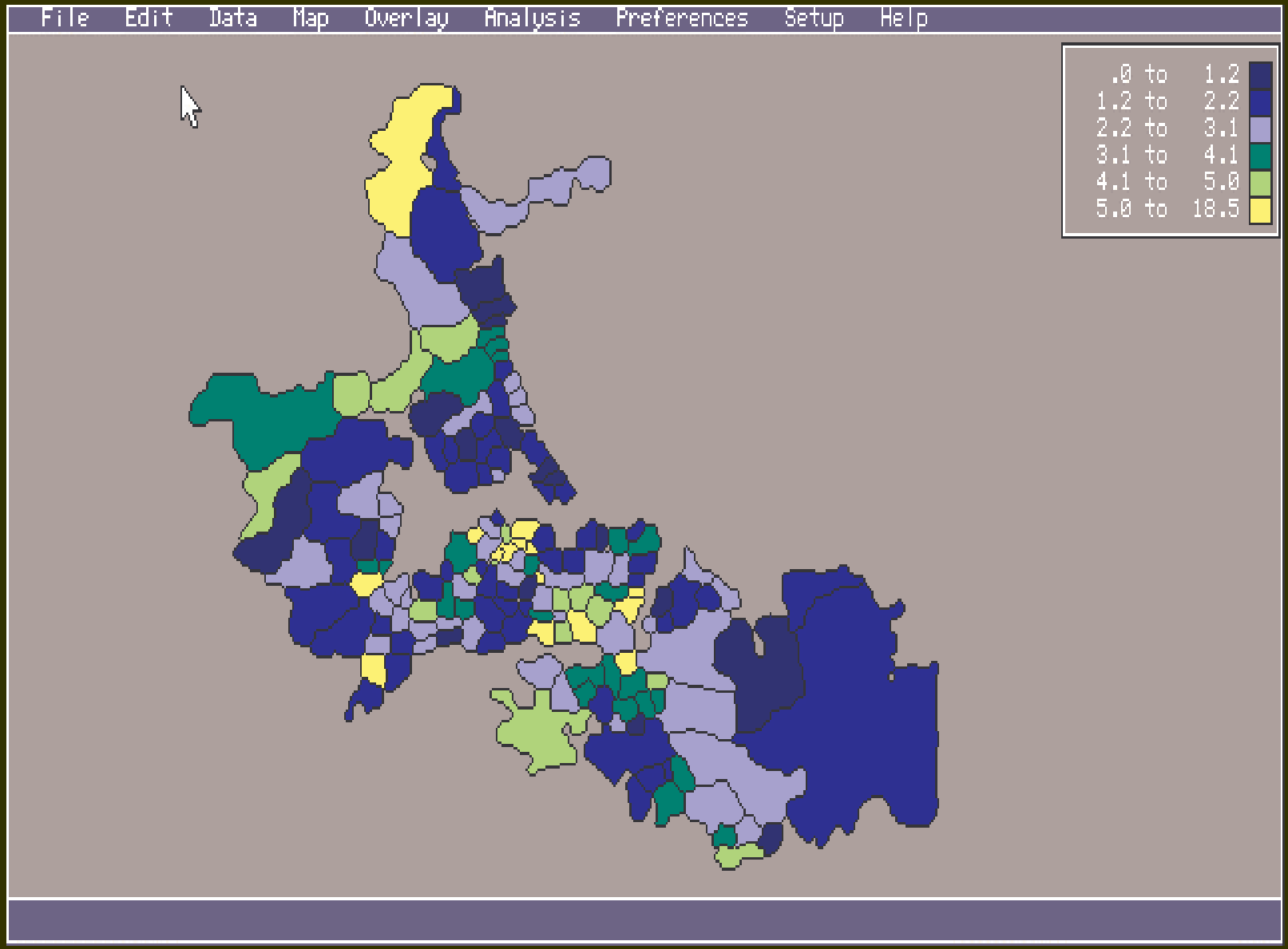
Estimando m e V

$$m = \frac{\sum_i O_i}{\sum_i Pop_i} = \text{taxa global}$$

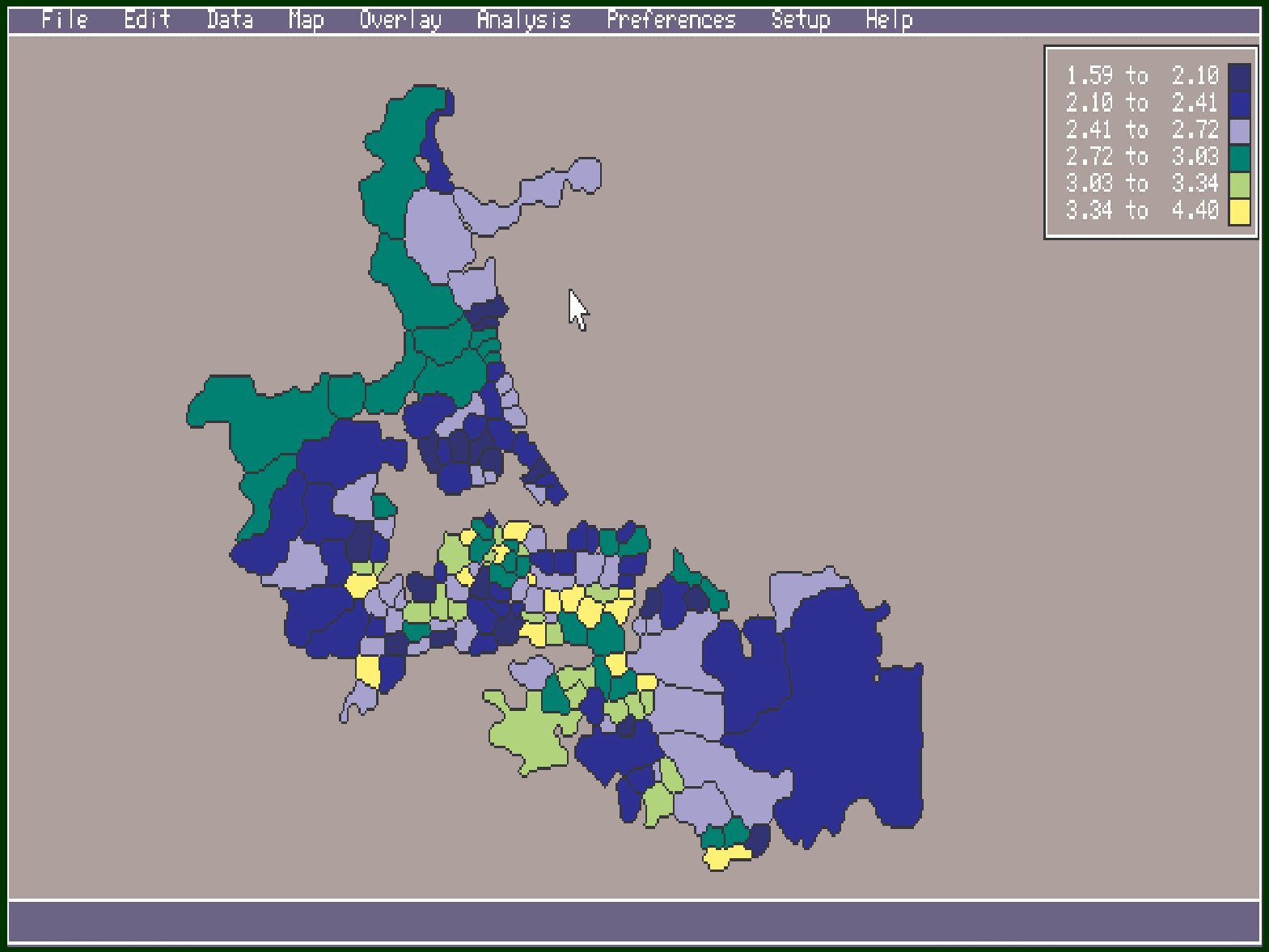
$$V = \frac{\sum_i Pop_i (r_i - m)^2}{\sum_i Pop_i} - \frac{m}{Pop\ média}$$


Mortalidade Infantil em Auckland
Bayes
0
5
10
15
0
5
10
15
taxa (por 1000)

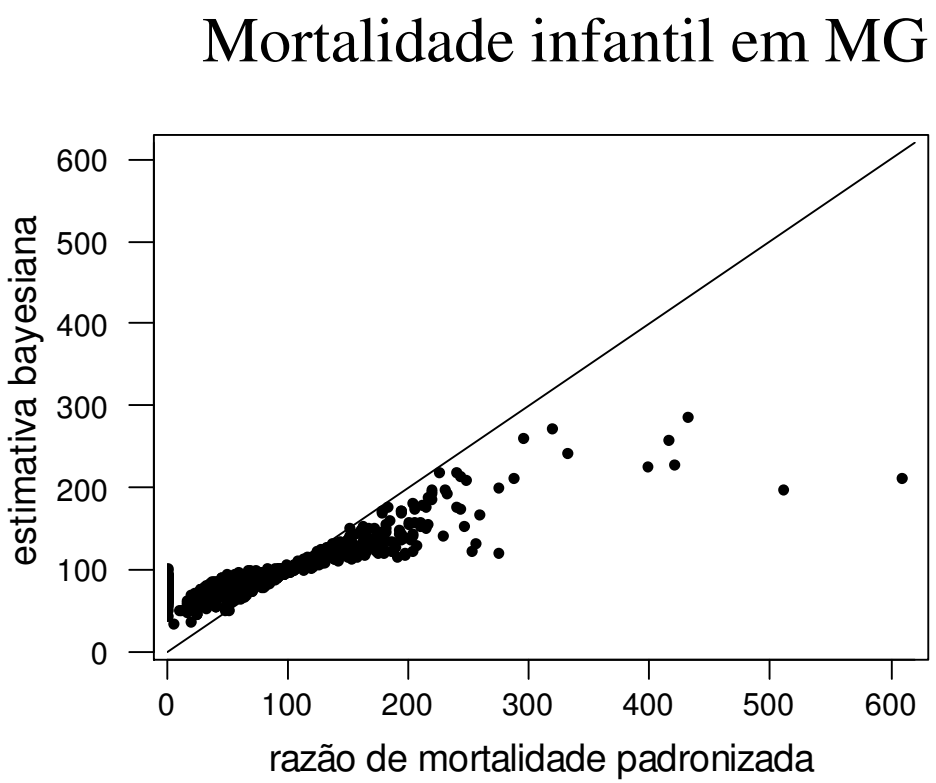
Taxas usuais (por 1000)



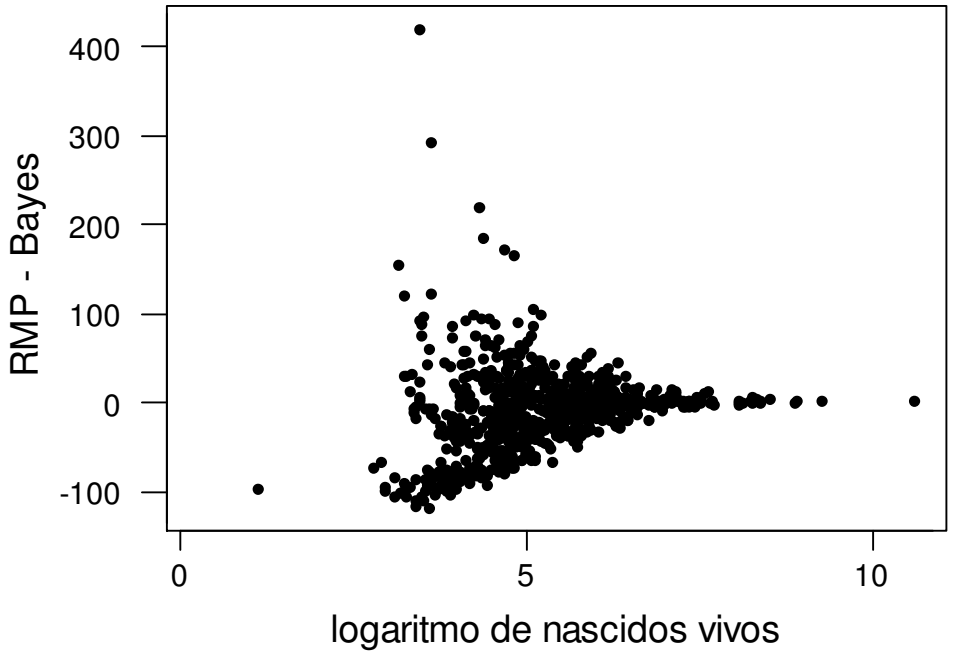
Taxas Bayesianas Empíricas



Estimativa contrai em direção à média regional

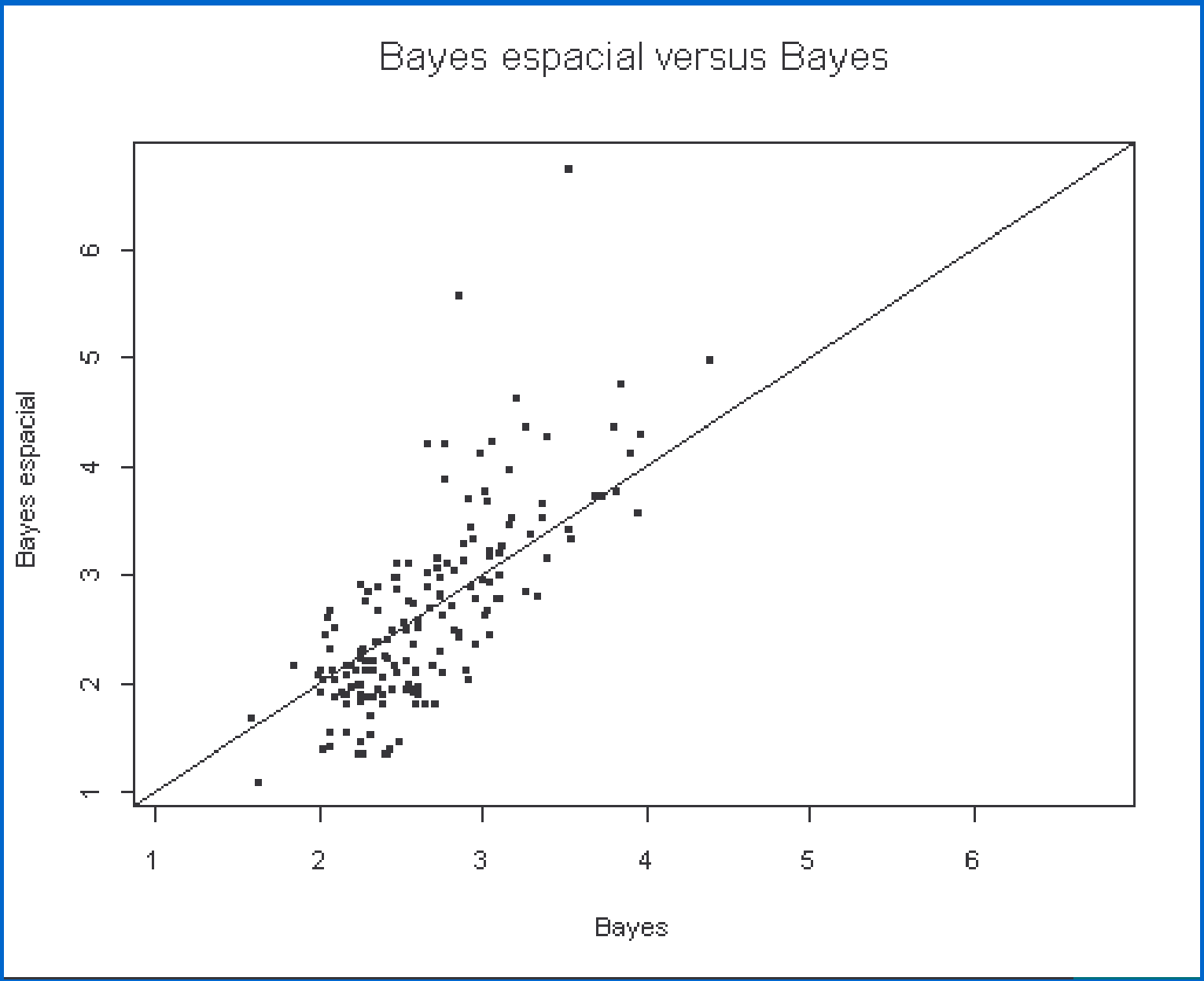


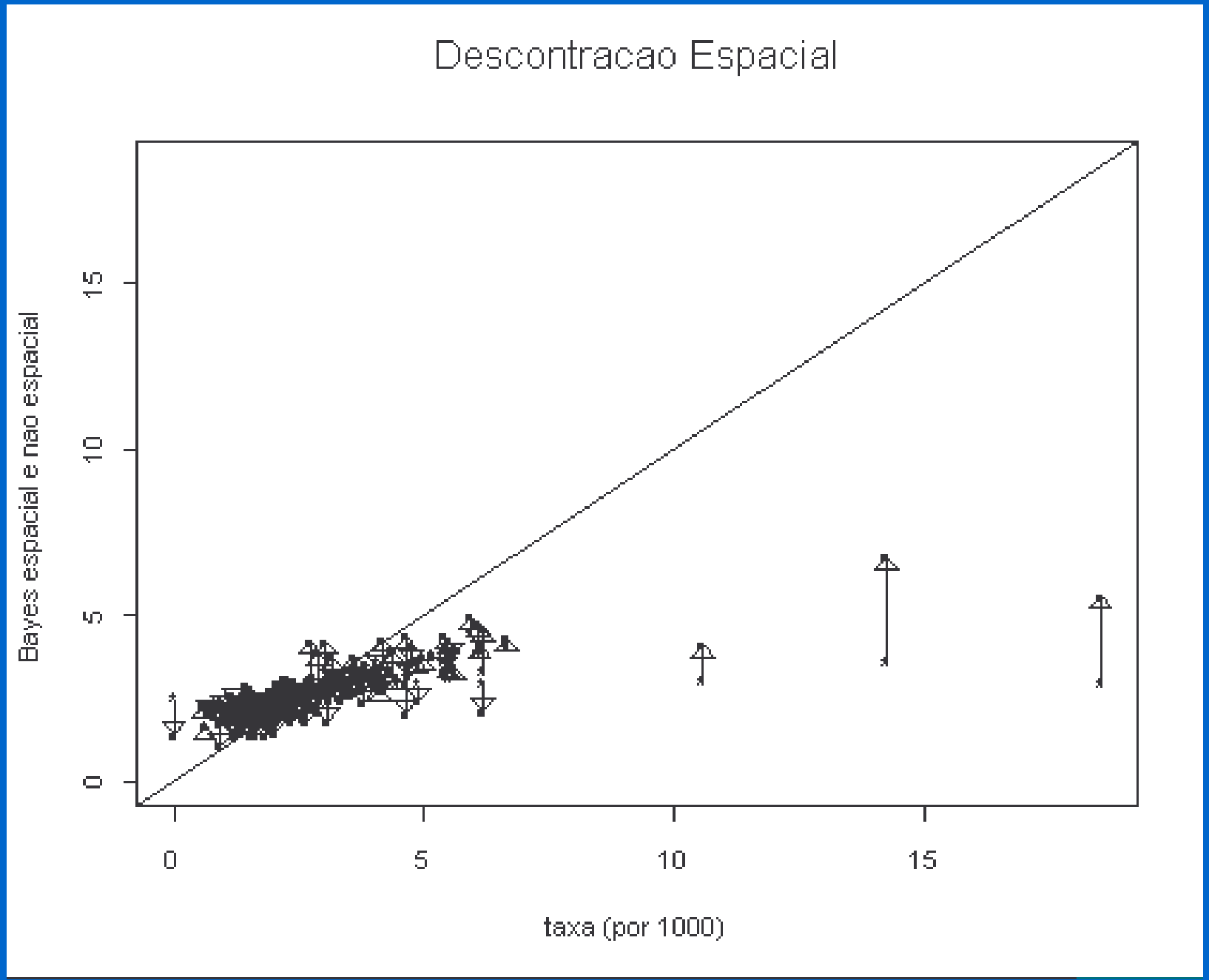
Efeito de contração é maior nos municípios menores



Estimativa Bayesiana Empírica Espacial

- Fazer estimativa bayesiana localmente: contrair 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 m_i e variância V_i


Bayes espacial versus Bayes
Bayes espacial
Bayes
1
2
3
4
5
6


Descontracao Espacial
Bayes espacial e nao espacial
taxa (por 1000)
0
5
10
15