Fundação Oswaldo Cruz
Escola Nacional de Saúde Pública
Departamento de Epidemiologia

Estatística espacial

Áreas

Áreas

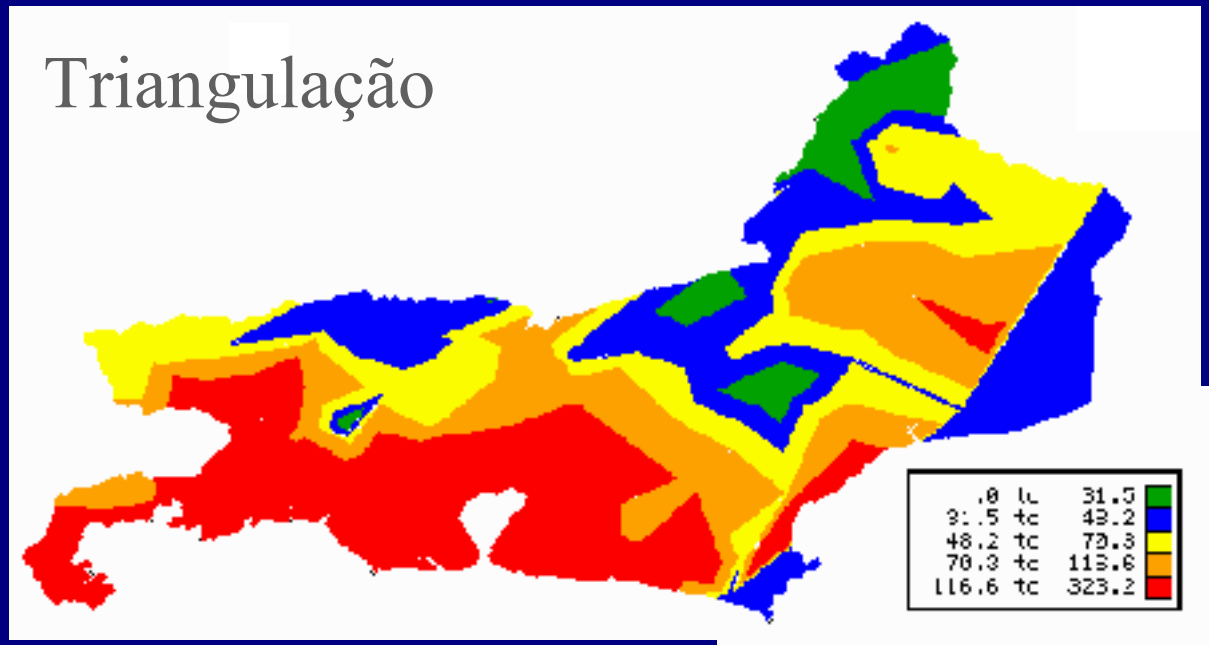
- Na análise de áreas o atributo estudado é em geral resultando de uma contagem ou um cálculo, apresentando valor constante: medida de síntese
- O objetivo não é a estimar a intensidade, mas a detecção e explicação de padrões e tendências observados entre áreas

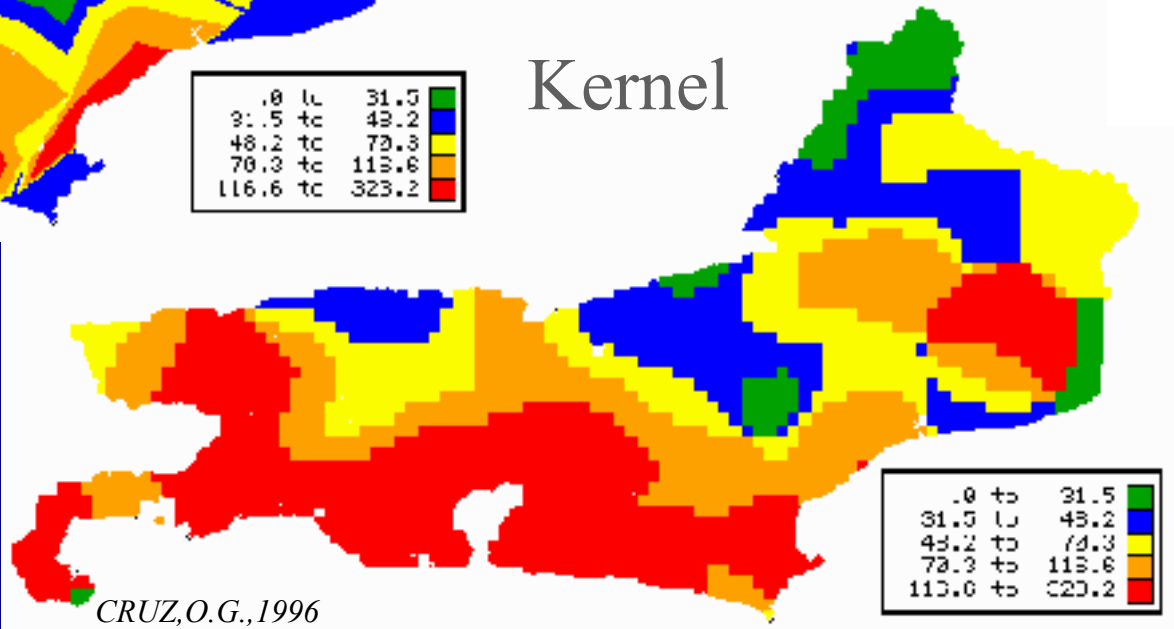
Áreas

- Área é definida por um polígono cuja forma e relações de vizinhança podem ser muito complexas
- O modelo básico do banco de dados:

Local	Casos	População	Médicos p/1000hab
Rio Bom	41	3209	5,4
Serra Verde	320	16897	2,6
Poço Fundo	67	2569	1,3

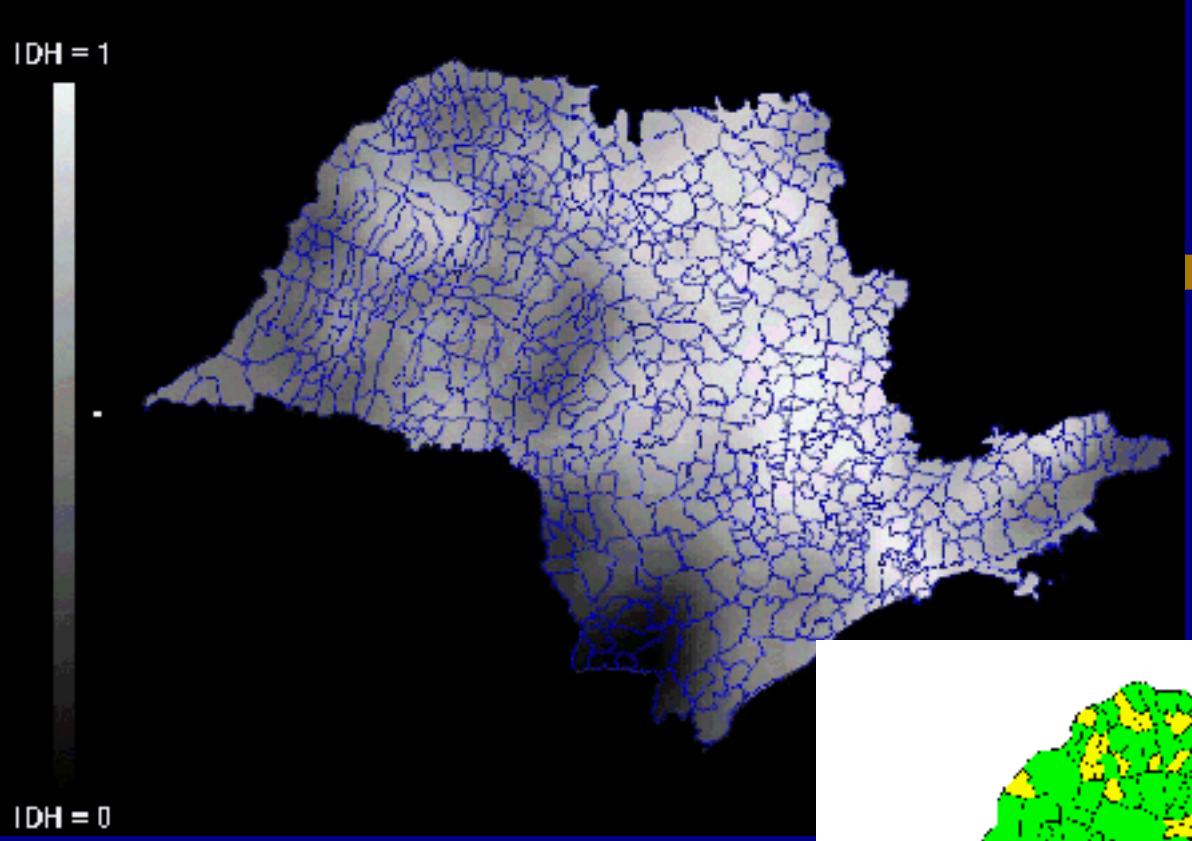
Interpolação em áreas



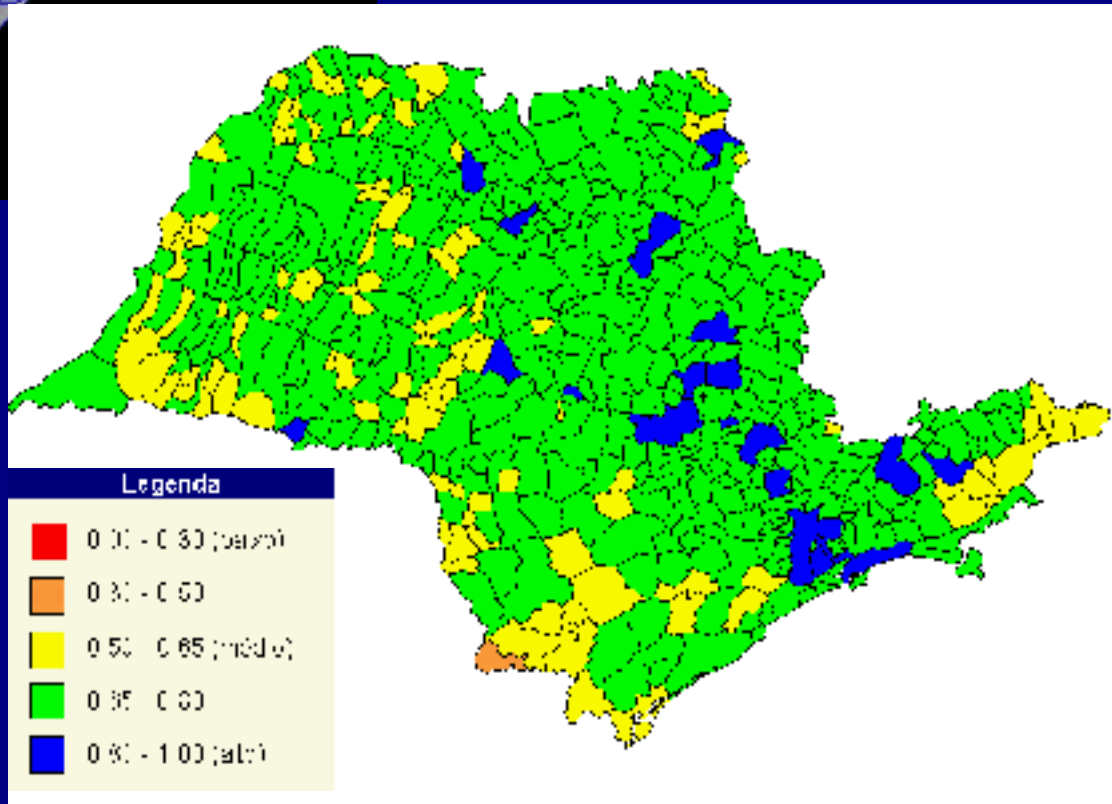


- O valor do indicador é atribuído a um ponto da área - centróide geométrico, populacional

CRUZ,O.G.,1996



Entidades ou superfícies



Kernel de áreas

- Utiliza-se para áreas alocando o valor do atributo a um ponto da área - centróide geométrico, populacional

$$\hat{p}_{\tau}(s) = \sum_{j=1}^{n} k\left(\frac{s - s_i}{\tau}\right) p_i$$

- Para o kernel de população, cada ponto receberá o atributo p_i (população) alisado pela função k, e largura de banda τ

Kernel de áreas

- No kernel de um atributo contínuo (por ex., indicadores), inclui-se no denominador o kernel da distribuição dos centróides das áreas

$$\hat{\mu}_{\tau}(s) = \frac{\sum_{j=1}^{n} k\left(\frac{s - s_i}{\tau}\right) y_i}{\sum_{j=1}^{n} k\left(\frac{s - s_i}{\tau}\right)}$$

- Obtém-se portanto a média do atributo na região e não uma contagem de eventos por unidade de área

Flutuação de pequenas áreas

- A taxa r_i estimada pela razão entre número de ocorrências e população a risco é um estimador de risco: $r_i = y_i / n_i$
- Entretanto onde a população é pequena os valores apresentam flutuação aleatória importante.
- Sendo objetivo estimar o risco nas áreas, nem sempre o estimador de máxima verossimilhança é o melhor.

Flutuação de pequenas áreas

- Técnica de alisamento pelo método Bayesiano empírico:
  - Global – encolhe a taxa das micro-áreas em direção à média global
  - Local – semelhante, porém considera apenas os vizinhos

Método Bayesiano Empírico (local)

- Seja: $r_i = y_i / n_i$

$\hat{m}_k$ a "taxa" média entre *k* vizinhos $\hat{m}_k = \dfrac{\sum_1^k y_i}{\sum_1^k n_i}$

s^2 a variância $s^2 = \dfrac{\sum n_i (r_i - \hat{m}_k)^2}{\sum n_i}$

Método Bayesiano Empírico

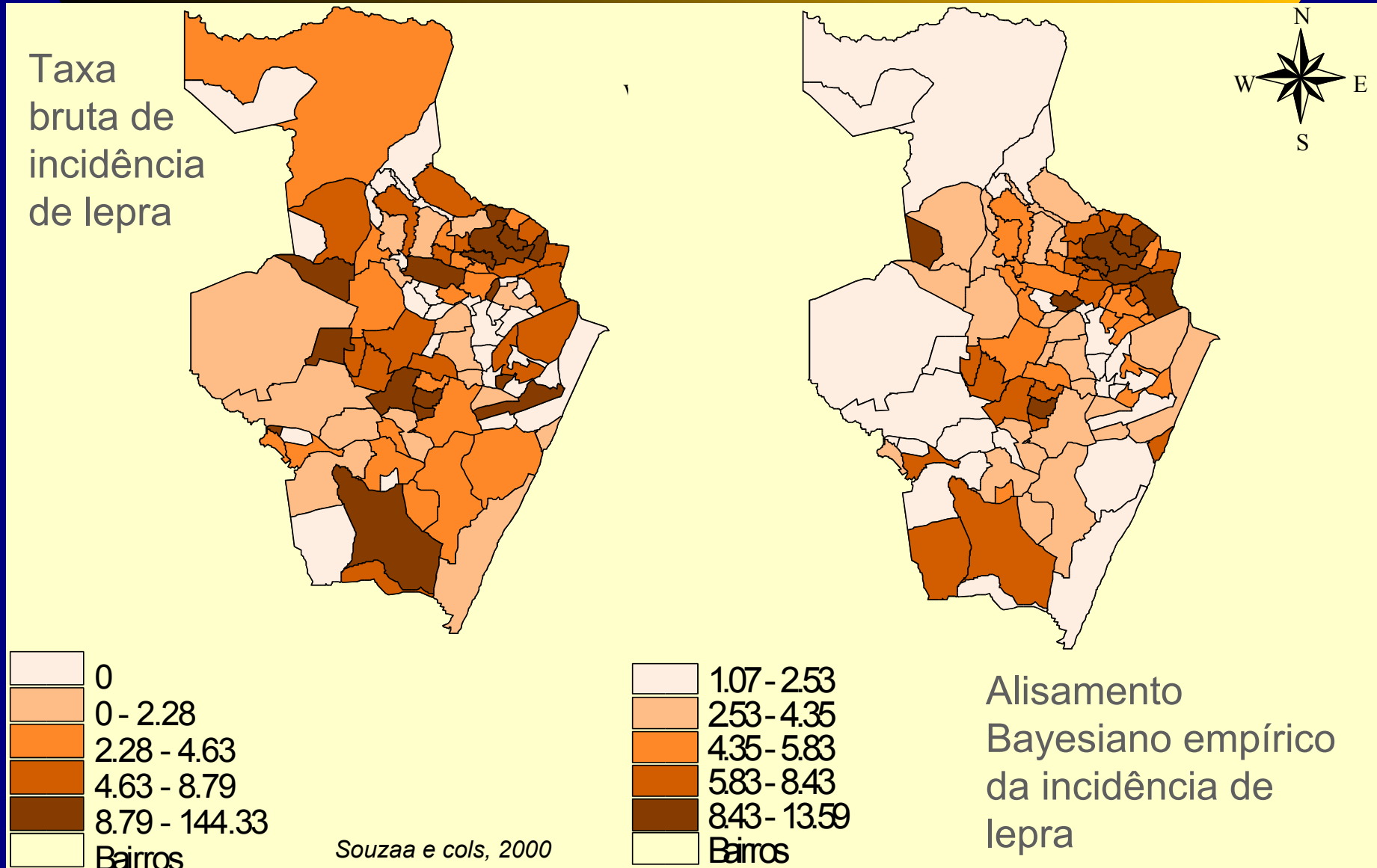
- A taxa corrigida será: $\theta_i = C_i r_i + \left(1 - C_i\right) \hat{m}_k$
- Onde C_i é fator de correção:

$$C_i = \frac{s^2 - \frac{\hat{m}_k}{\bar{n}_k}}{s^2 - \frac{\hat{m}_k}{\bar{n}_k} + \frac{\hat{m}_k}{n_i}}$$

Método Bayesiano Empírico

- C_i ↗ quando $\hat{m}_k / n_i$ ↘, ou seja, quando n_i ↗
- Quando $\hat{m}_k / \bar{n}_i > s^2$, então assume-se $C_i = 0$ e a taxa estimada para a área *i* será igual a média entre vizinhos: $\theta_i = \hat{m}_k$

Exemplo – Bayesiano empírico



Cluster em áreas

- Diz-se que existe um *cluster* entre áreas quando áreas com valores semelhantes ocorrem próximas no espaço;
- Ou quando existe uma quantidade "excessiva de eventos" na mesma área;
- São causas de cluster: fonte comum, contagiosidade, acaso.

Cluster em áreas

- Para testar se este agregado é acima de um valor esperado, existem diversos testes que procuram verificar a medida da autocorrelação espacial, testando se significativa
- Os resultados de qualquer destes métodos depende diretamente dos pesos da matriz de vizinhança.

Matriz de vizinhança

- utiliza-se matriz **W** , onde cada elemento w_{ij} representa medida de proximidade espacial entre as áreas A_i e A_j;
- a escolha de w_{ij} depende do tipo de dado, de região, dos mecanismos particulares da dependência espacial;
- vizinhos podem ser de primeira ordem, segunda até n.

Matriz de vizinhança

$$w_{ij} = \begin{cases} 1 \\ 0 \end{cases}$$

centróide de A_i é o mais próximo de A_j

caso contrário

$$w_{ij} = \begin{cases} 1 \\ 0 \end{cases}$$

centróide de A_i dentro de distância especificada de A_j (buffer)

caso contrário

$$w_{ij} = \begin{cases} 1 \\ 0 \end{cases}$$

A_i tem fronteira comum com A_j

caso contrário

$$w_{ij} = \frac{l_{ij}}{l_i}$$

l_{ij} é o comprimento da fronteira comum entre com A_i e A_j e l_i é o perímetro de A_i

Matriz de vizinhança

Testes de Cluster

Moran I

$$I = \frac{N \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\left(\sum_{i=1}^{N} (y_i - \bar{y})^2 \right) \left(\sum \sum_{i \neq j} w_{ij} \right)}$$

- W_{ij} é a matriz de vizinhança
- Relaciona-se à auto-correlação
- Média $\bar{y}$ suposta constante: processo estacionário

Testes de Cluster

Geary

$$C = \frac{(N-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - y_j)^2}{2\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)\left(\sum\sum_{i \neq j} w_{ij}\right)}$$

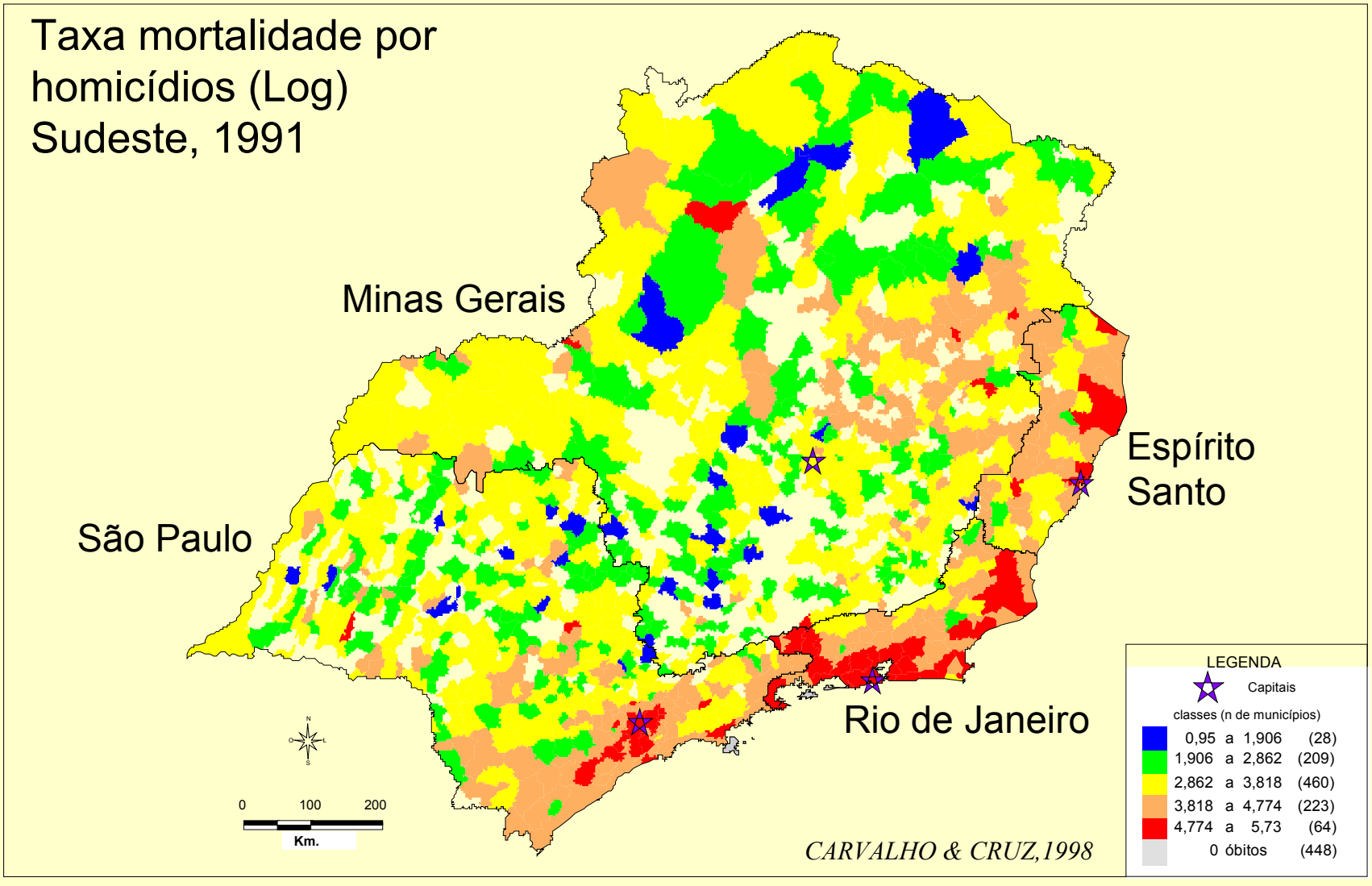
- W_{ij} é a matriz de vizinhança
- Relaciona-se ao variograma
- Média $\bar{y}$ suposta constante: processo estacionário

Função de autocorrelação

- Desta forma se constrói a função de autocorrelação para cada lag
- A significância estatística pode ser calculada por permutação ou, caso a variável tenha distribuição normal, por teste Z
- Moran no lag k

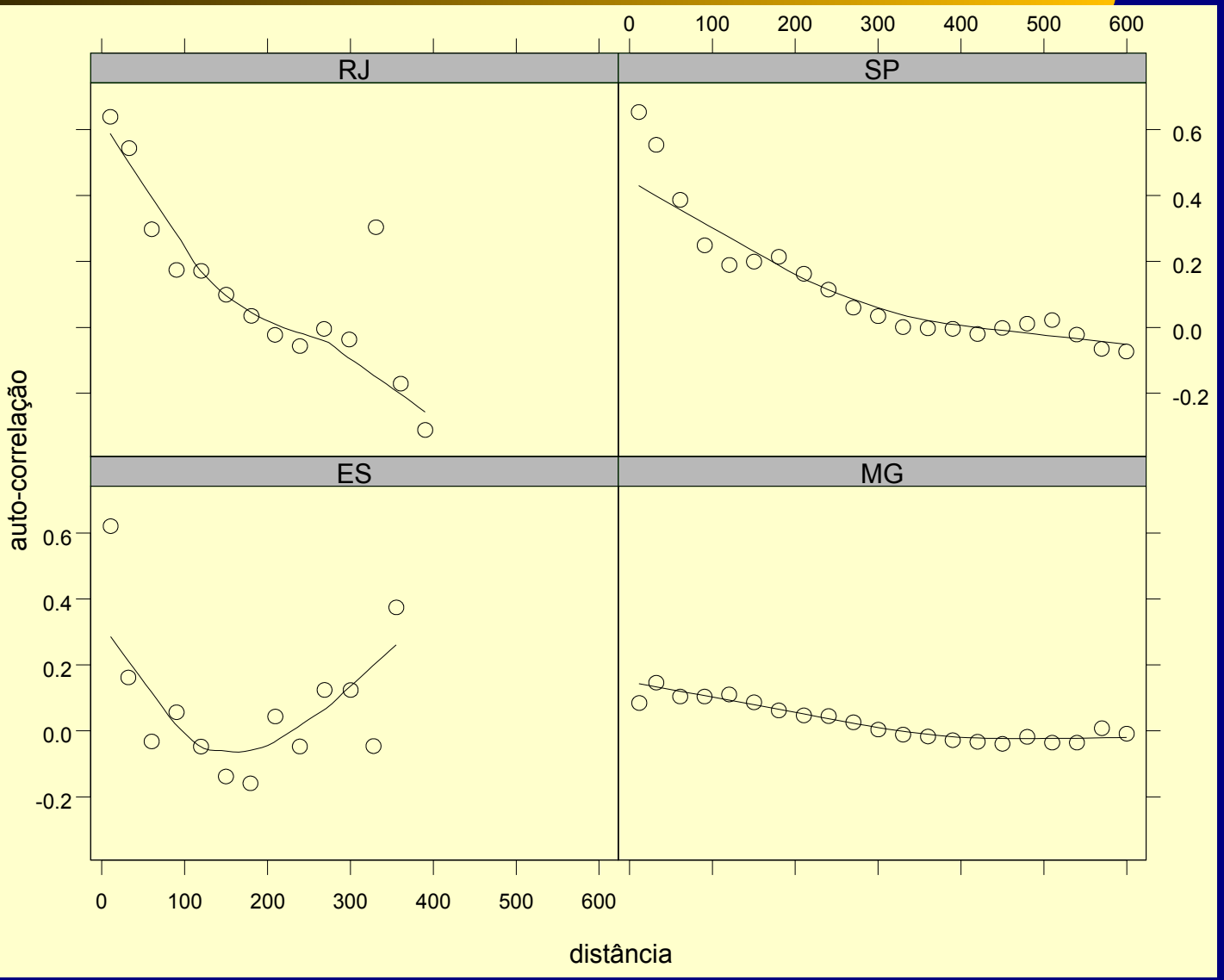
$$I^{(k)} = \frac{N \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}^{(k)} (y_i - \bar{y})(y_j - \bar{y})}{\left(\sum_{i=1}^{N} (y_i - \bar{y})^2 \right) \left(\sum \sum_{i \neq j} w_{ij}^{(k)} \right)}$$

Autocorrelação



Correlograma

Correlograma da taxa mortalidade por homicídios por UF



Indicadores locais

- Permitem encontrar os "bolsões" de dependência espacial não evidenciados nos índices globais
- Permitem identificar:
  - agrupamentos de objetos com valores semelhantes (*cluster*)
  - objetos anômalos
  - existência de mais de um processo espacial

Indicadores locais

- A significância estatística também é calculada por permutações e supõe-se normalidade da variável.
- Existem dois índices locais:
  - LISA (Anselin, 1996)
  - Índice Gi e Gi* (Getis e Ord, 1992)

Indicadores Locais

LISA - Indicador local de autocorrelação espacial

$$I_i = \frac{z_i \sum_j w_{ij} z_j}{\sum_{i=1}^{N} z_i^2}$$

Z_i - desvio de i em relação a média global

Z_j - média dos desvios dos vizinhos de i

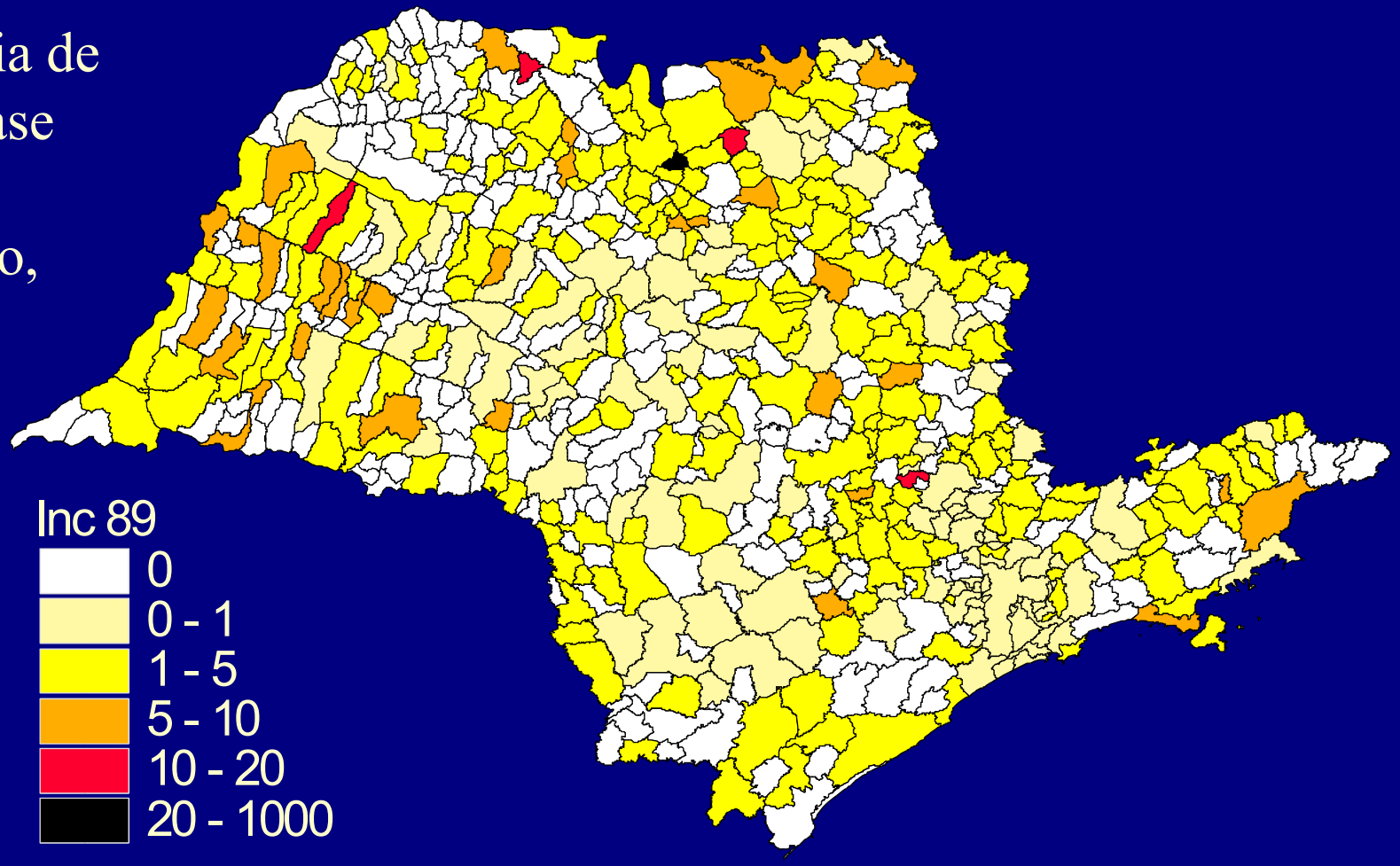
Média constante: processo estacionário

Significância semelhante a I - permutação ou normalidade

Mapa LISA

Incidência de Hanseníase

São Paulo, 1989

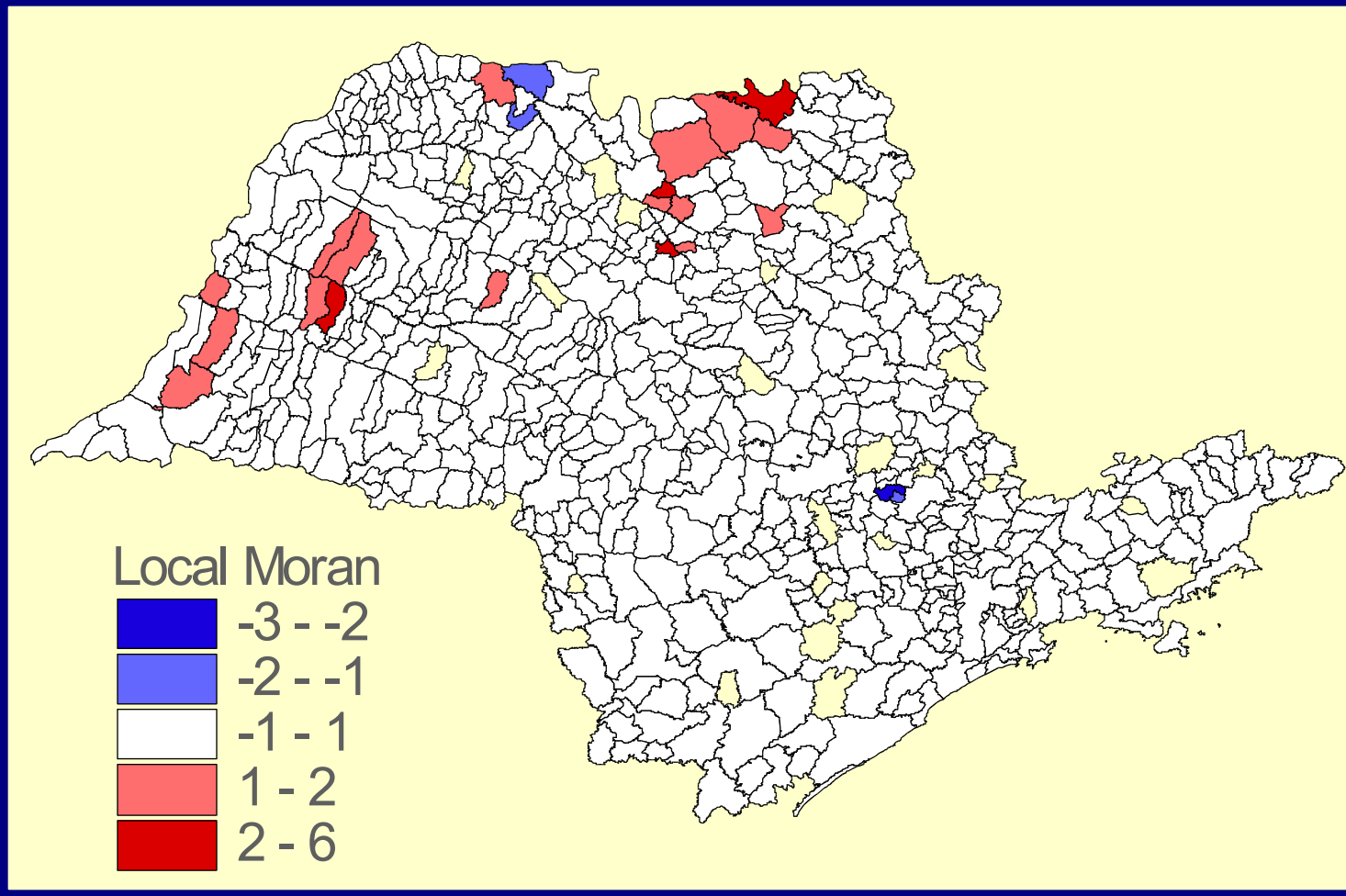


Moran I = 0.07185

Mapa LISA

Incidência de Hanseníase - LISA

São Paulo, 1989



Modelo de superfície de tendência

- Pode-se incluir no modelo de regressão comum as coordenadas geográficas de cada ponto como variáveis independentes, inclusive ao quadrado e seus produtos - neste caso se modela a superfície de tendência

$$\mu_i = \beta_{10} x_i + \beta_{20} x_i^2 + \beta_{11} x_i y_i + \beta_{01} y_i + \beta_{02} y_i^2 + \varepsilon_i$$

Modelos de regressão NÃO espacial

- Na investigação sobre causas de diferenças entre áreas é possível utilizar modelos multivariados não espaciais (estudos ecológicos clássicos).
- Embora úteis, se existir forte tendência ou correlação espacial, os resultados serão influenciados, apresentando associação estatística onde não existem (e vice-versa).

Modelos de regressão NÃO espacial

- As hipótese básicas deste modelo são:
  - As variáveis explicativas são linearmente independentes
  - $E(\varepsilon)=0$
  - $V(\varepsilon)=\sigma\varepsilon^2$

  $\varepsilon \sim (0, \sigma\varepsilon^2)$

$$y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon_i$$

- Embora úteis, se existir forte tendência ou correlação espacial, os resultados serão influenciados, apresentando associação estatística onde não existem (e vice-versa).

Modelos de regressão espacial

- Modelos CAR (Conditional AutoRegressive):

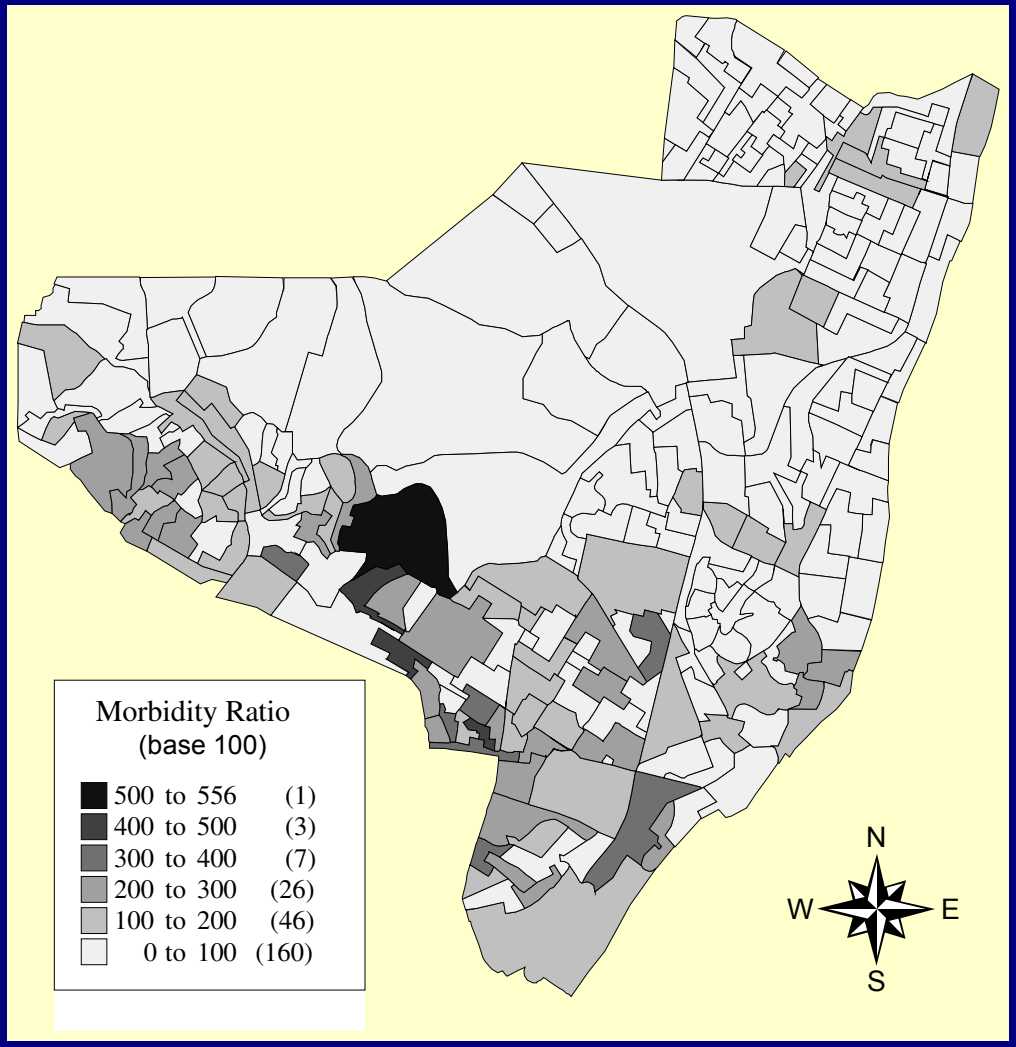
$$y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i} + u$$

- $$u = \lambda S(u) + \varepsilon$$
- Onde:
  - resíduos da regressão são espacialmente correlacionados
  - u tem matriz de covariância igual a $\sigma^2 V$
  - V é uma matriz não diagonal que descreve a dependência espacial

Modelagem Bayesiana

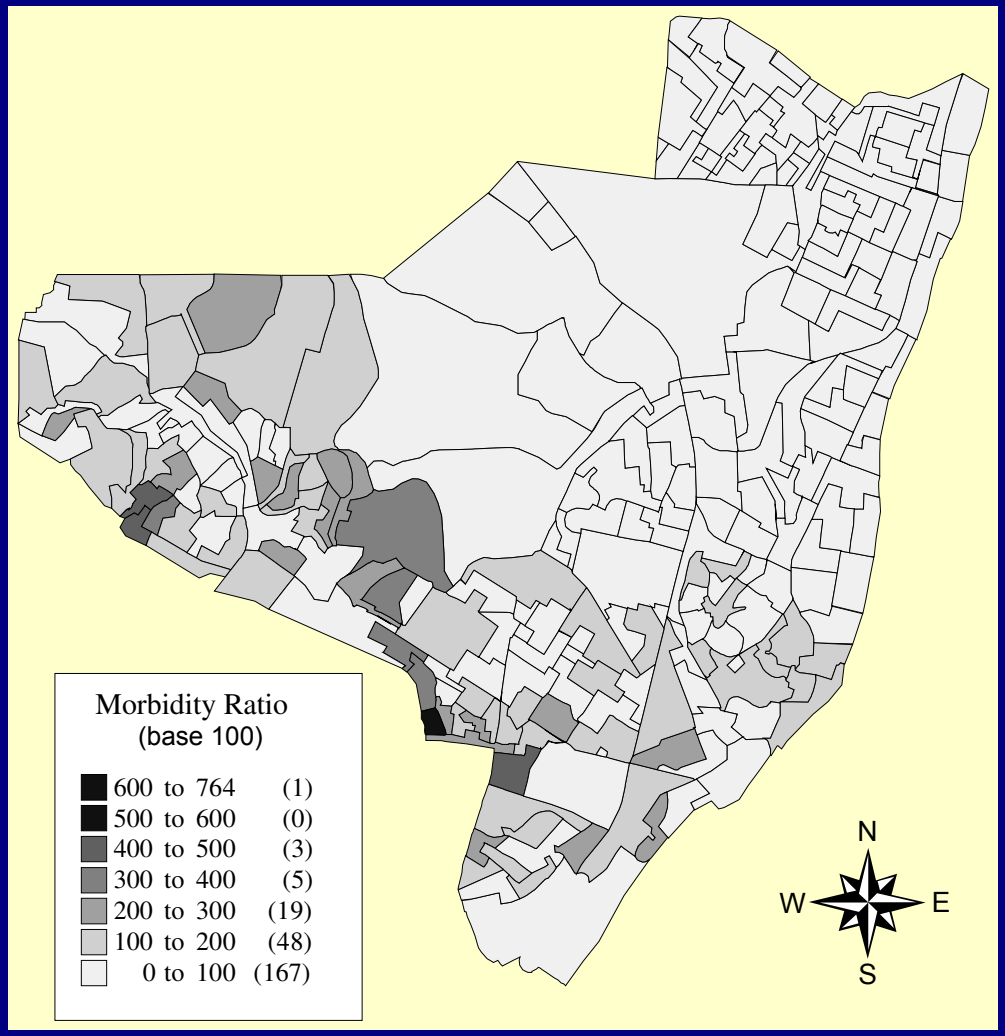
- Para estimar os parâmetros destes modelos, particularmente se incluir efeitos aleatórios simultaneamente ↙ inferência bayesiana.
- O mais utilizado método de estimativa - Markov Chain Monte Carlo (MCMC) - através de simulações permite estimar não só o valor esperado da variável estudada em cada área, mas outros parâmetros também.

Hanseníase em Olinda



Alisamento Bayesiano

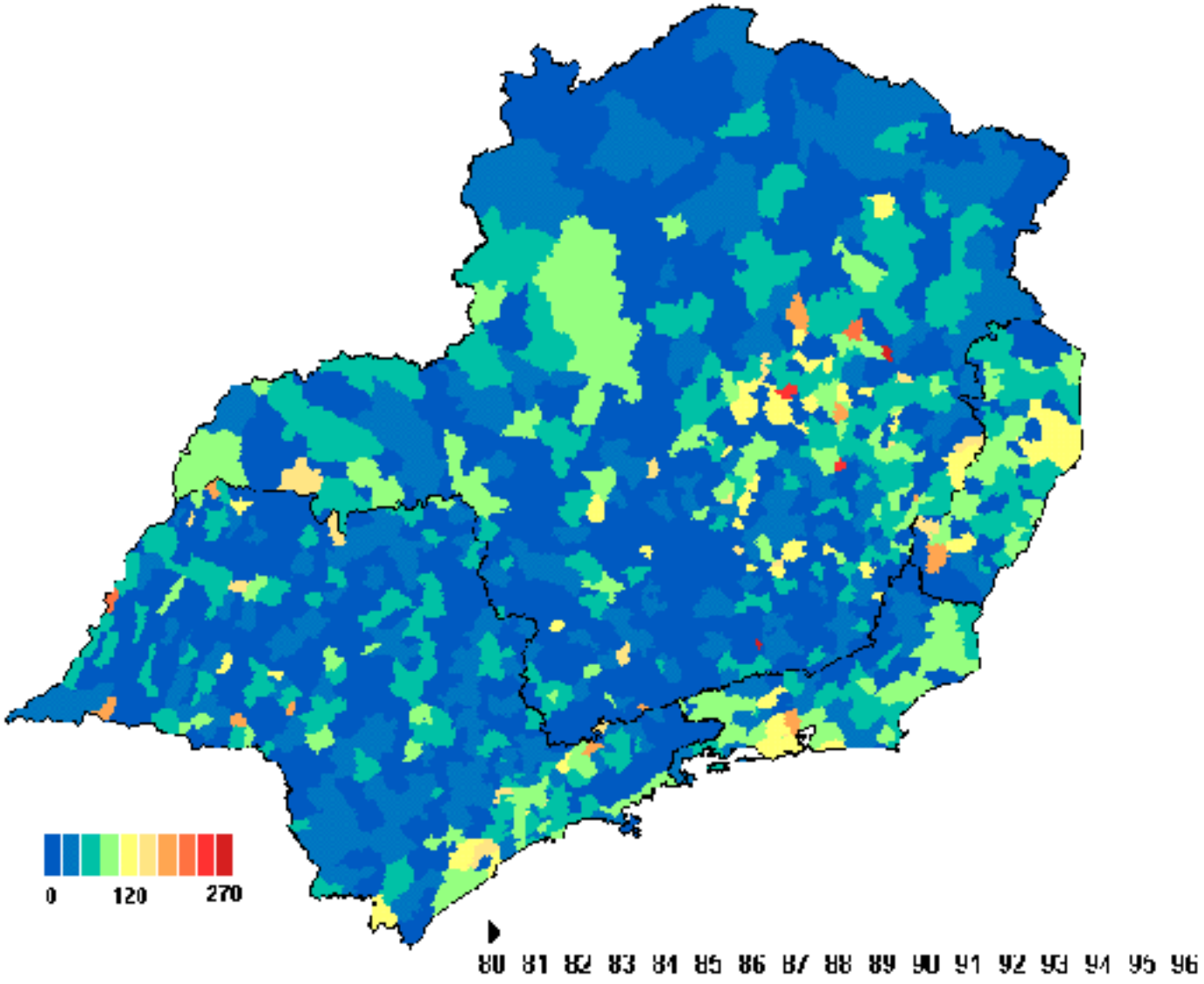
com correção de subregistro



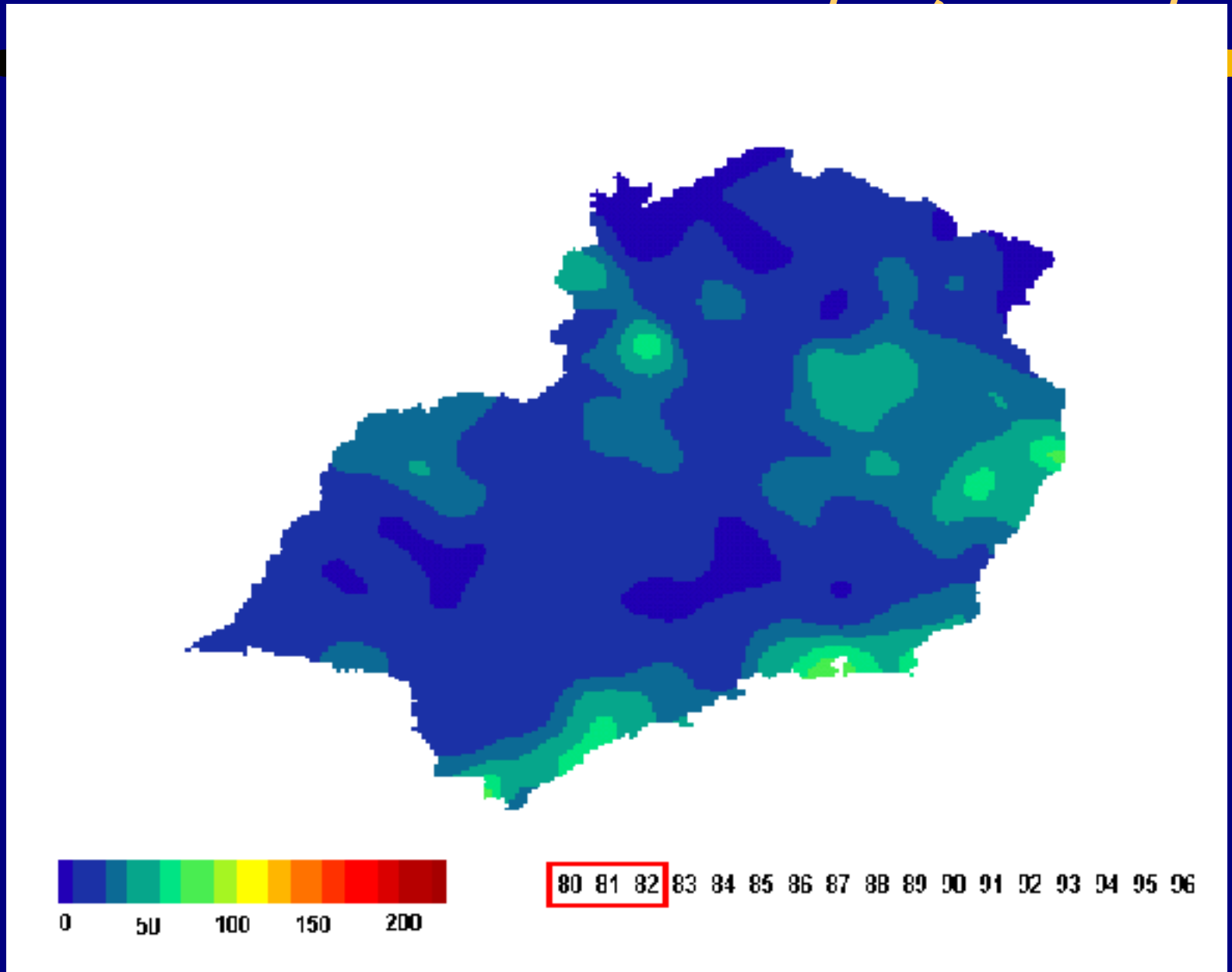
Modelagem Bayesiana

- Vantagens:
  - Mais flexível,
  - maior número de parâmetros,
  - problemas mais complexos.
- Problemas:
  - indicadores de ajuste,
  - sensibilidade,
  - ajuste fino – especialista.
- Softwares dedicados:
  - BayesX: http ://www.stat.uni-muenchen.de/~lang/bayesx/bayesx.html
  - WinBugs: http://www.mrc-bsu.cam.ac.uk/bugs/

Espaço-tempo



Espaço-tempo

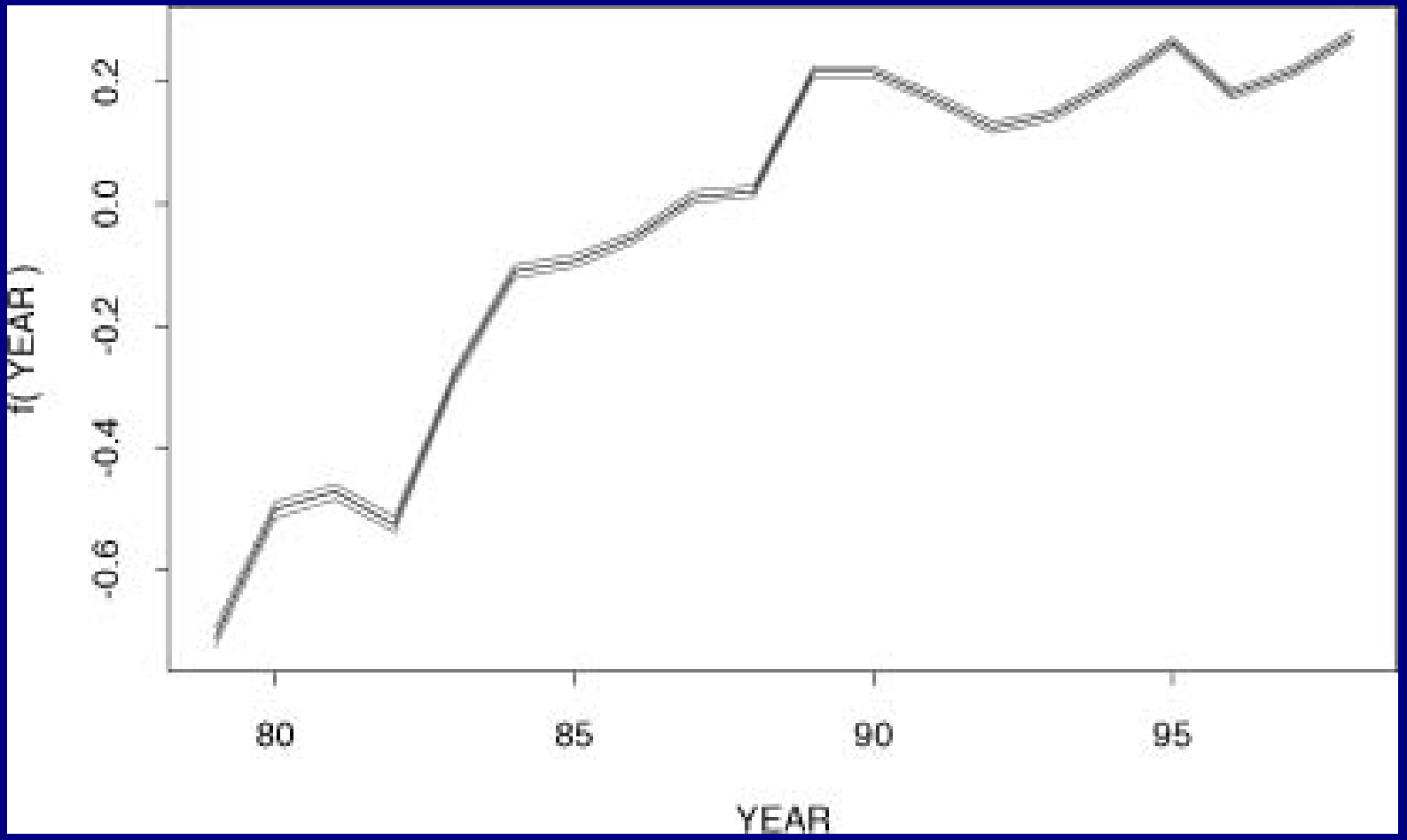


Modelagem espaço-temporal

- Permite detectar padrões: aplicações na vigilância epidemiológica ambiental
- Visa quantificar as alterações no padrão espaço-temporal e relacioná-las a fatores determinantes – socioeconômicos e ambientais
- Bastante complexa, é uma das áreas de desenvolvimento da estatística
- Inferência bayesiana

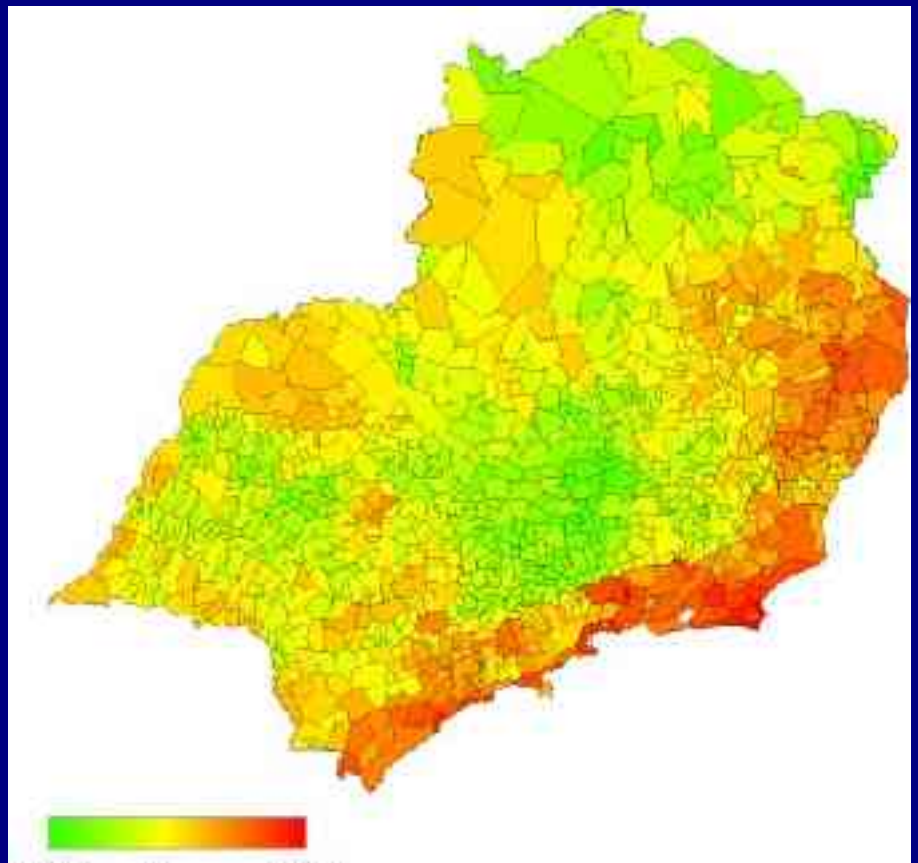
Modelo espaço-temporal

Componente Temporal

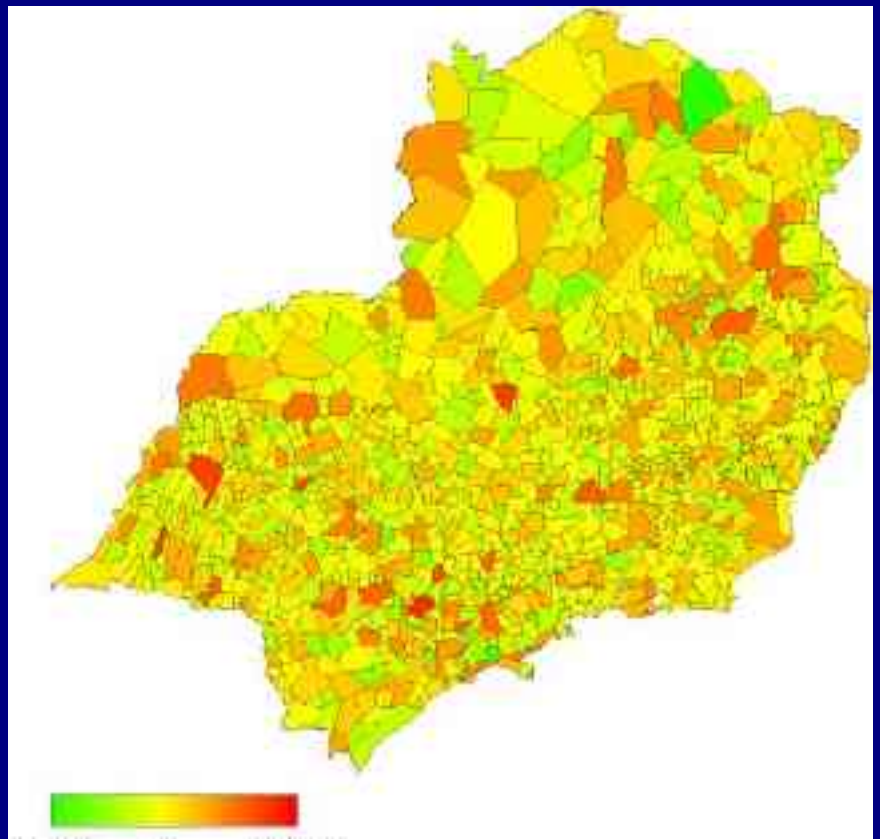


Modelo espaço-temporal

Componente espacialmente estruturado:



Componente aleatório

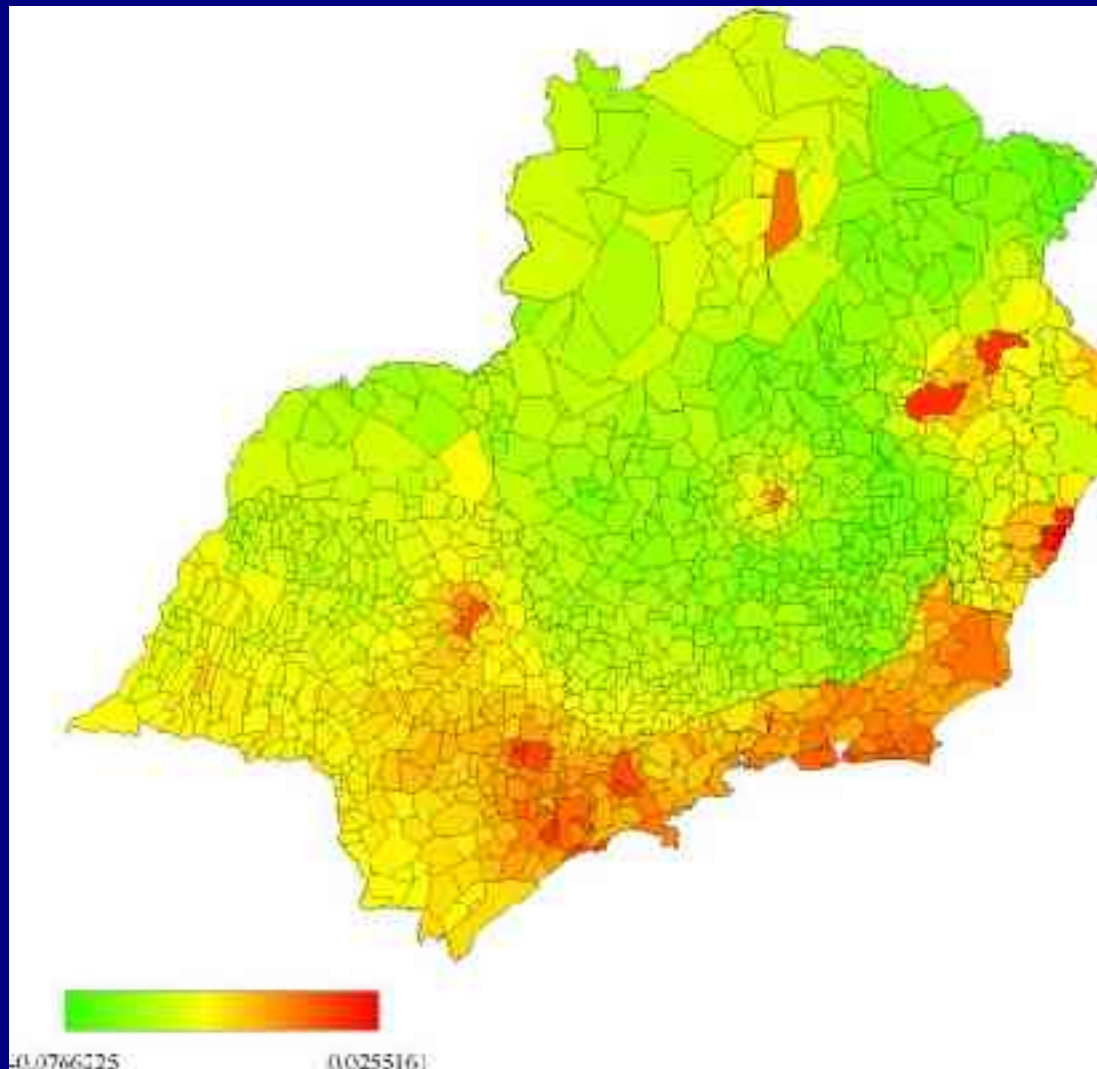


Modelo espaço-temporal

Componente interação Espaço-Tempo

Variable	mean	10% quant.	90% quant
Const	-6.5544	-7.16533	-6.12849
(efeito estados)			
RJ	0.373566	0.019446	0.75179
SP	-0.05851	-0.338679	0.17467
ES	0.87375	0.651499	1.1056
(efeito Região Metropolitana)			
MRJ	1.25247	0.94338	1.54339
MSP	1.34616	1.17254	1.54179
MES	-0.8428	-1.70334	-0.0225
MMG	0.1718	-0.03191	0.36108

Burnin: 4.000
Simulações 50.000



Conclusões

- Não é simples!
- É fundamental o desenvolvimento de ferramentas mais amistosas e integradas
- Mais que uma equipe multidisciplinar, é necessário um grupo de trabalho interdisciplinar, que crie um dialeto comum