The Abdus Salam International Centre for Theoretical Physics

United Nations Educational, Scientific and Cultural Organization

IAEA International Atomic Energy Agency

2453-13

School on Modelling Tools and Capacity Building in Climate and Public Health

15 - 26 April 2013

Area Data Analysis

SA CARVALHO Marilia

PROCC FIOCRUZ
Avenida Brasil 4365
Rio De Janeiro 21040360
BRAZIL

Area Data Analysis

Marilia Sá Carvalho

Fundação Oswaldo Cruz

Outline

Outline

References

- Bailey,T.C. and Gatrell,A.C.. *Interactive Spatial Data Analysis*. Longman, 1996.
- Bivand. R.S; Pebesma, E.J.; Goméz-Rubio, V. *Applied Spatial Data Analysis with R*. Springer, 2008.
- Wood, S.N.. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC Texts in Statistical Science Series, 2006.

What is areal data

- The most common and available spatial data

Place	Cases	Population	Covariate
Rio das Flores	10	1200	5.34
Sao Pedro	25	2134	2.56
Botucatu	354	30405	10.45

- And a boundary!

Definition

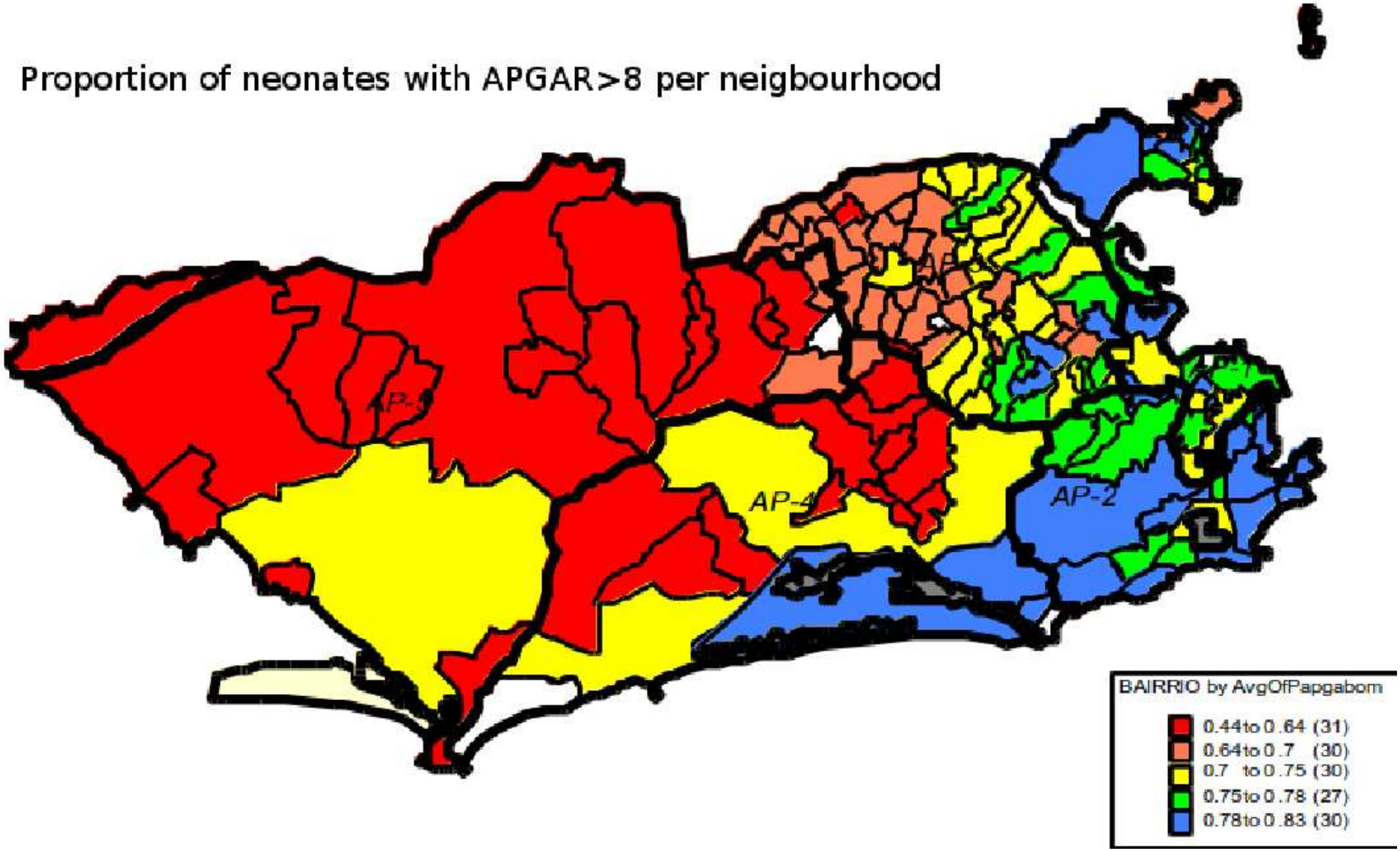
- An observed areal data x is a realisation of a random process X in a discrete space
  - The areas are fixed
  - The observations x associated with each area are random
- The aim is to estimate parameters of the distribution of X:
  - to describe
  - to explain based on covariates
- This structure is not "natural", but available, pragmatic

How can we use areal techniques here?

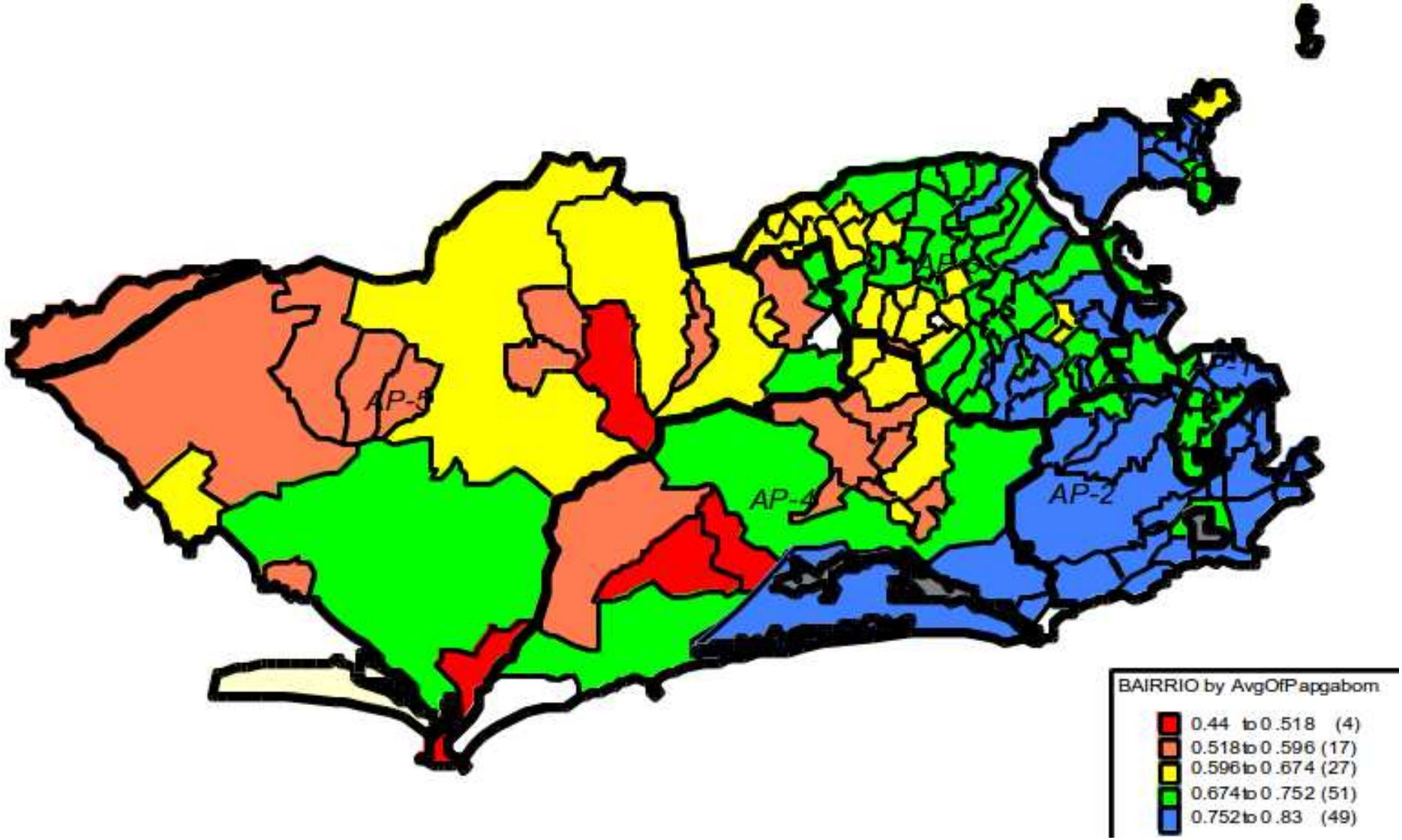
- Location of dengue fever cases
- Number of *Aedes aegypti* collected in a random sample of households
- Rainfall measures in 16 sites in Rio
- Results (Positive/Negative) of a dengue fever seroprevalence survey in a random sample of households
- Counts of leprosy by census tract

Outline

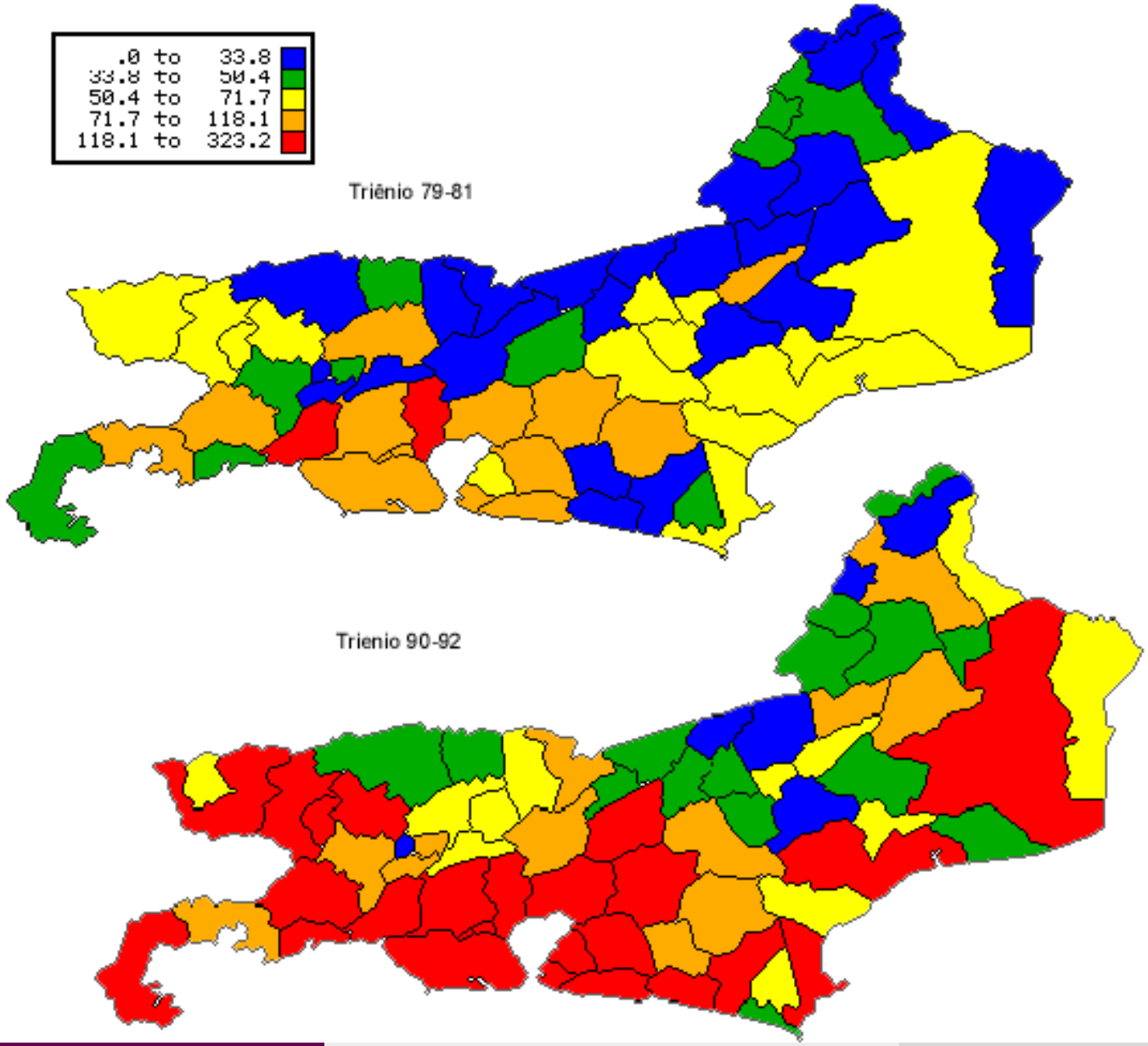
Easiest representation



Maybe...


AP-5
AP-4
AP-2
BAIRRIO by AvgOfPapgabom
0.44 to 0.518 (4)
0.518 to 0.596 (17)
0.596 to 0.674 (27)
0.674 to 0.752 (51)
0.752 to 0.83 (49)

Comparison


.0 to 33.8
33.8 to 50.4
50.4 to 71.7
71.7 to 118.1
118.1 to 323.2
Triênio 79-81
Trienio 90-92

Vicinity

- Any modelling of areal data has to specify how the areas are connected
- This requires a spatial weights matrix that reflects the intensity of the geographic relationship between observations in a neighbourhoods:

Vicinity

$$w_{ij} = \begin{cases} 1, & A_i \text{ centroid is the closest to } A_j \\ 0, & \text{otherwise} \end{cases}$$

$$w_{ij} = \begin{cases} 1, & A_i \text{ centroid is inside a buffer for} A_j \\ 0, & \text{otherwise} \end{cases}$$

$$w_{ij} = \begin{cases} 1, & A_i \text{ has a common border with} A_j \\ 0, & \text{otherwise} \end{cases}$$

$$w_{ij} = \begin{cases} 1, & A_i \text{ has a direct highway link with} A_j \\ 0, & \text{otherwise} \end{cases}$$

$$w_{ij} = \frac{l_{ij}}{l_i}, \text{where } l_{ij} \text{ is the length of the common border}$$

and l_i is the perimeter of A_i

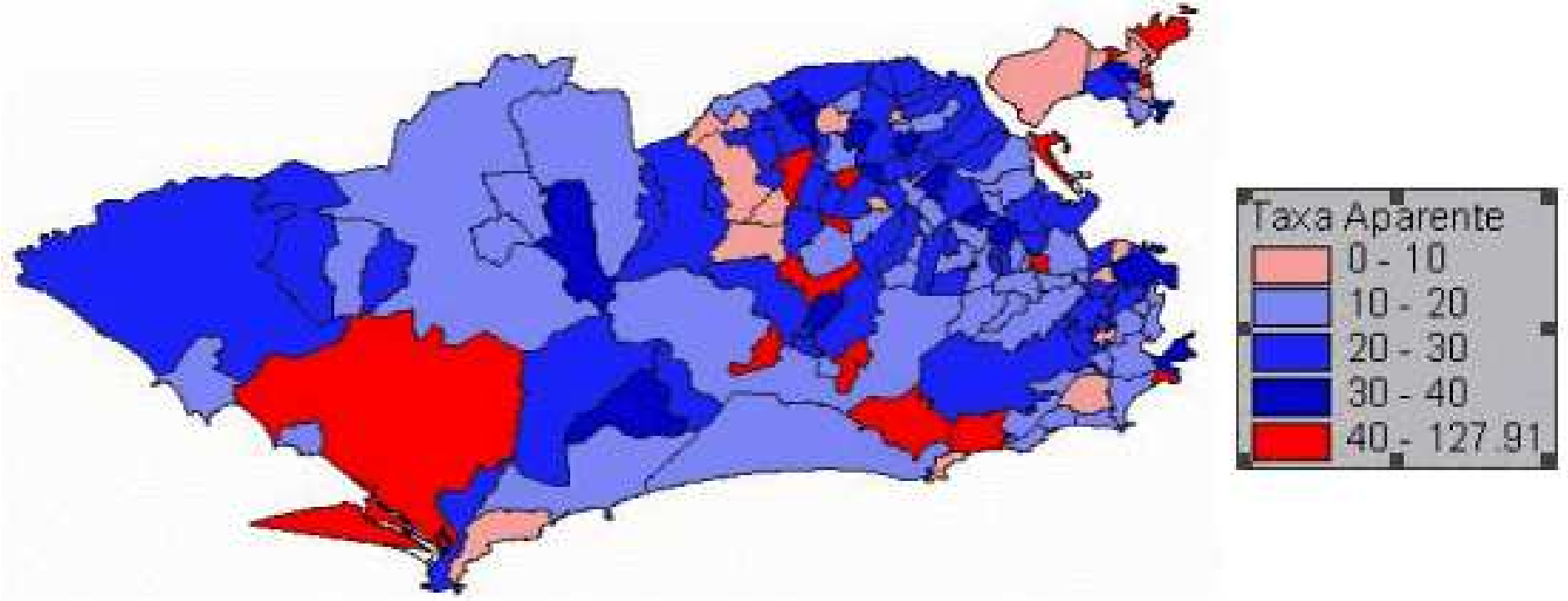
Rio has many beautiful mountains

And not many roads

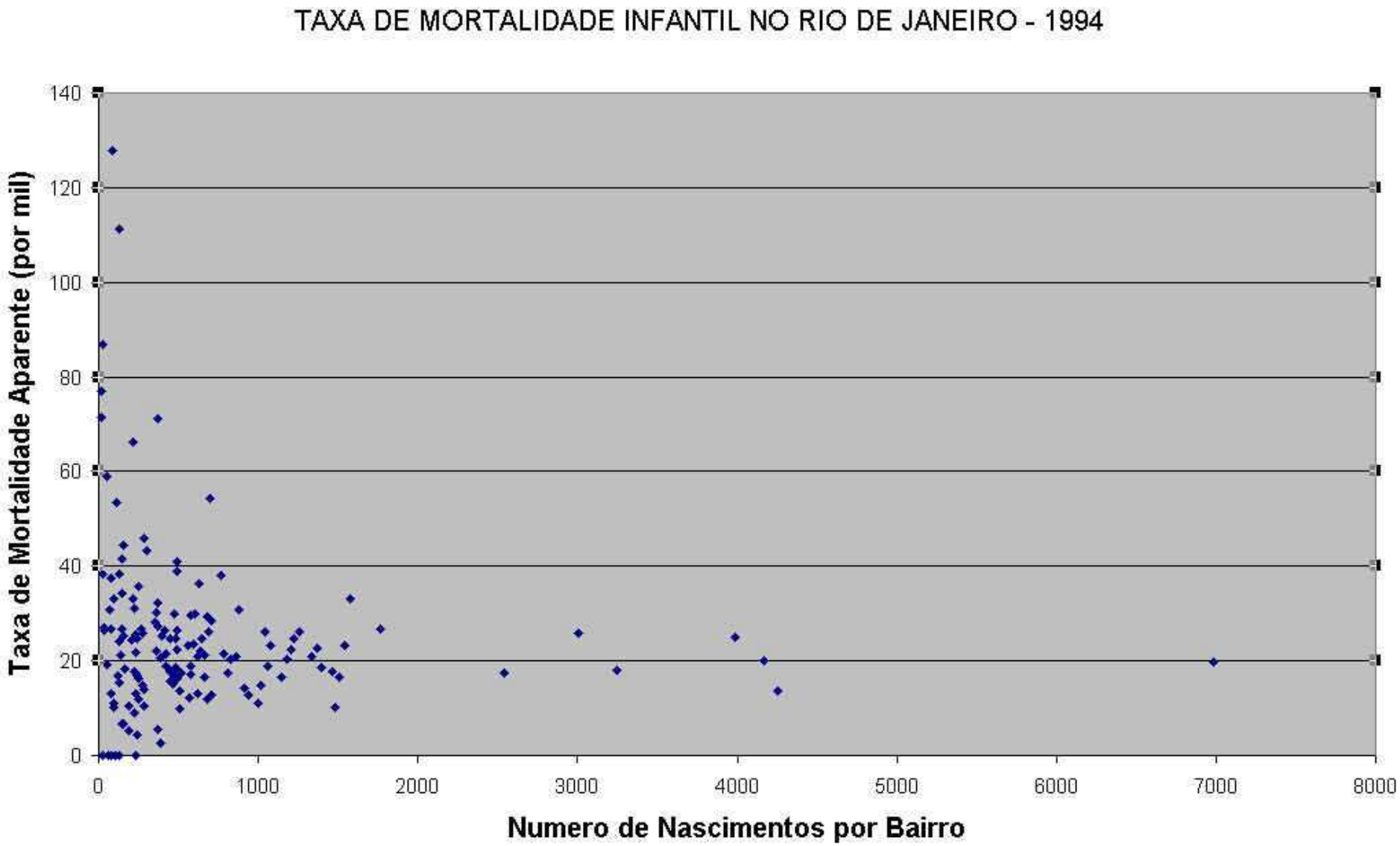
Mapping – Standardised Rates (SMR)

- Useful technique to allow comparison over time, and among areas with varying demographic structure
- Usually standardised by age and sex – direct and indirect standardisation
- Direct:
  - considering the overall rate r of a disease as: $r = \sum O_i / \sum Pop_i$
  - for each area i in region A calculate the number of expected cases as: $E_i = Pop_i \times r$
  - $SMR = O_i / E_i$

Random fluctuation



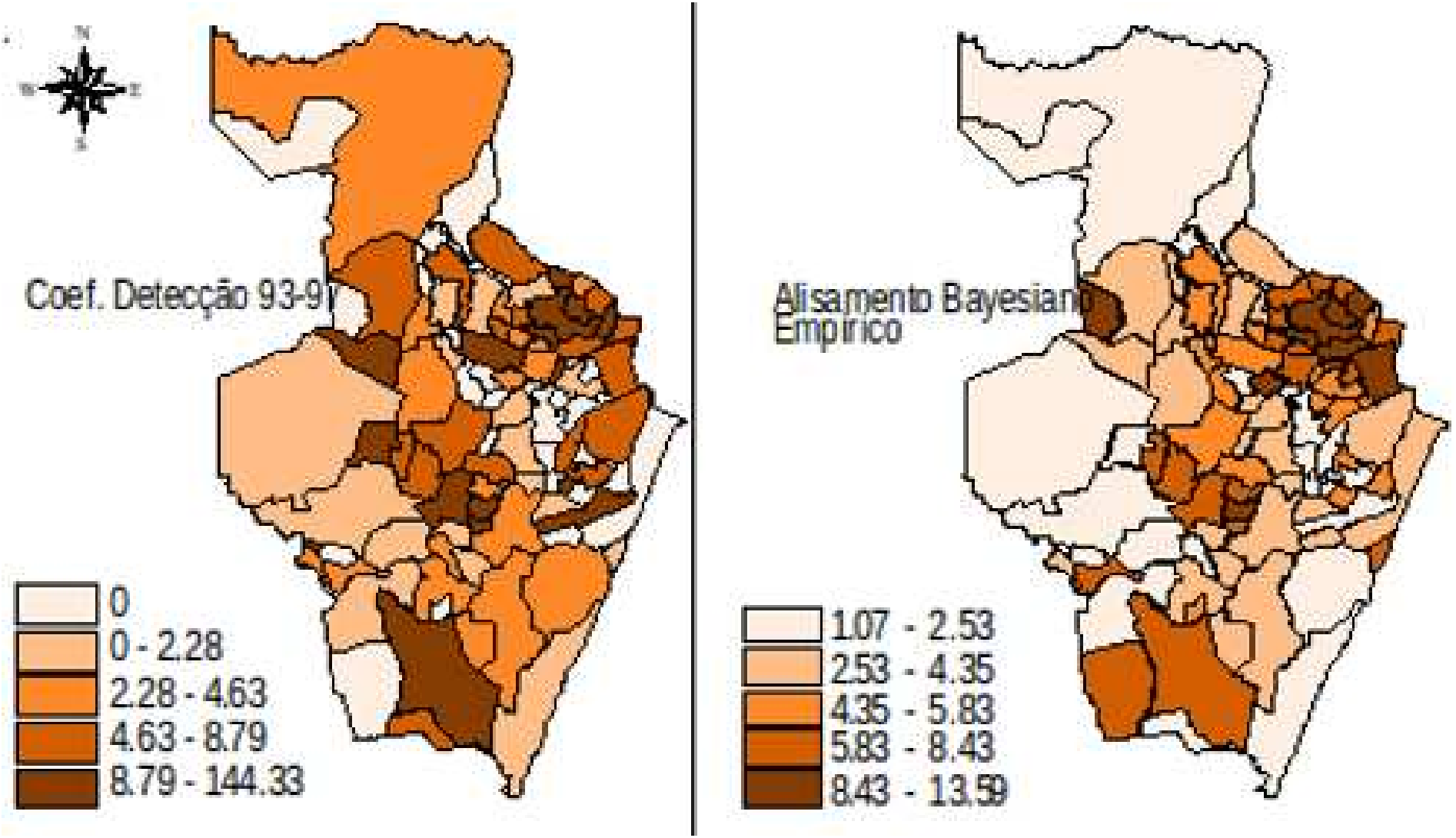
Random fluctuation



Smoothing areal data

- The same idea of any smoother: to eliminate random fluctuation
- We need a map to "see" the underlying pattern
- We do have prior information:
  - small population indicators fluctuate a lot!
  - neighbouring areas tend to be more similar to each other
- Empirical Bayes methods

Empirical Bayes Estimator



Empirical Bayes Estimator

- Consider the observed raw rate $r_i = y_i/n_i$, where y is the number of events in area i and n_i the population at risk
- Its variance will be: $s^2 = \sum n_i(r_i - \hat{\mu})^2 / \sum n_i$
- A better estimator of the underlying process is:

$$\theta_i = C_i r_i + (1 - C_i)\hat{\mu}$$

- Where the correction factor is:

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

- μ may be the overall mean or a local mean

Outline

Cluster detection

- A cluster among areas means that close by areas present much more similar indicator values than expected
- Causes of clusters: common source, contagion
- In general space and time concentrated
- But more often socioeconomic conditions!
- All tests depends upon the neighbourhood matrix
- Two types of tests:
  - generic – measure the overall degree of spatial autocorrelation
  - local – LISA and scan tests

Moran's I

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})}$$

- Values range from -1 (indicating perfect dispersion) to $+1$ (perfect correlation)
- Zero value indicates a random spatial pattern
- Be careful: under stationarity

Geary's C

$$C = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(x_i - x_j)^2}{\sum_i (x_i - \bar{x})}$$

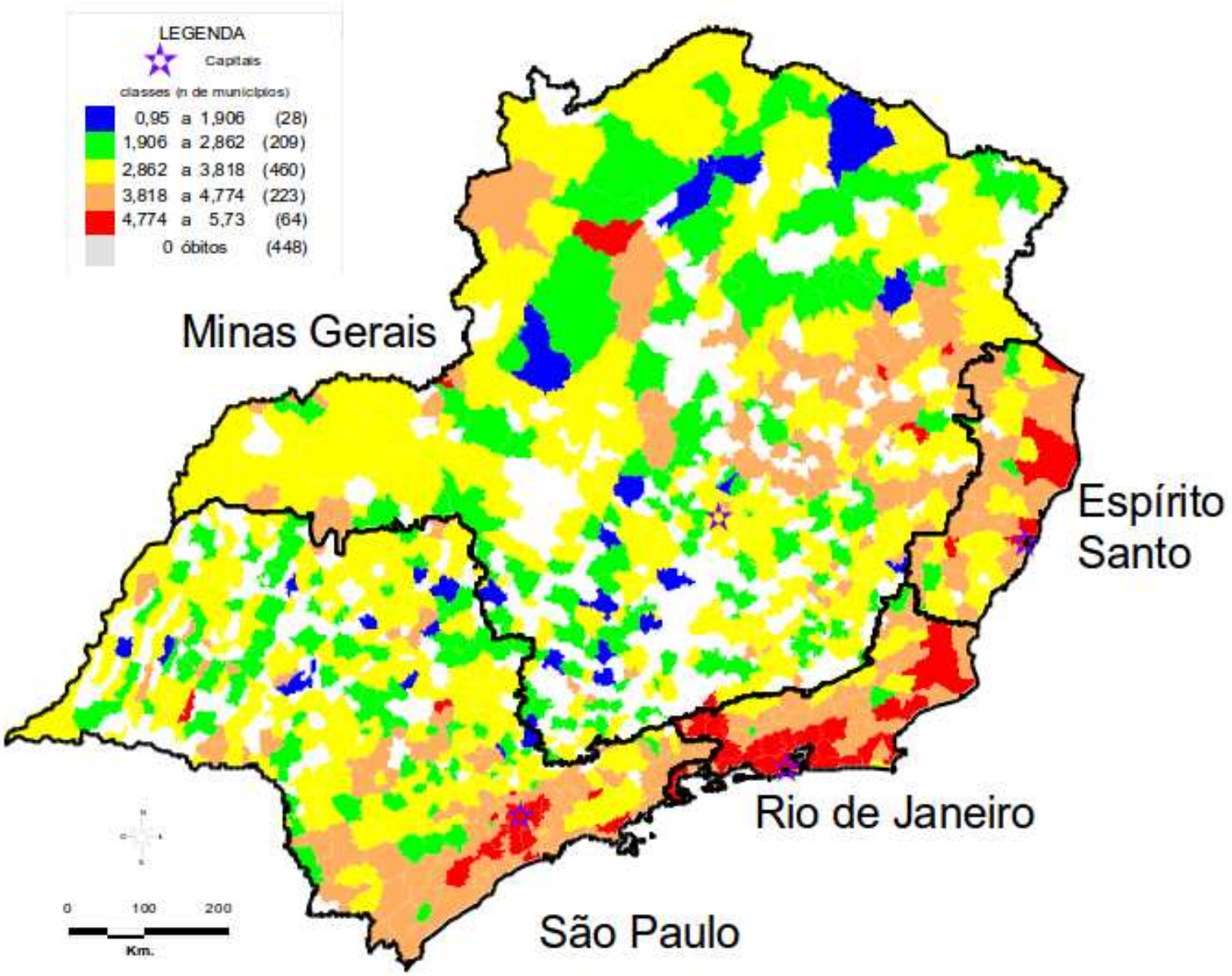
- Values range between 0 and 2
- 1 means no spatial autocorrelation
- Values lower the 1 – positive autocorrelation
- More sensitive to local spatial correlation

Correlogram

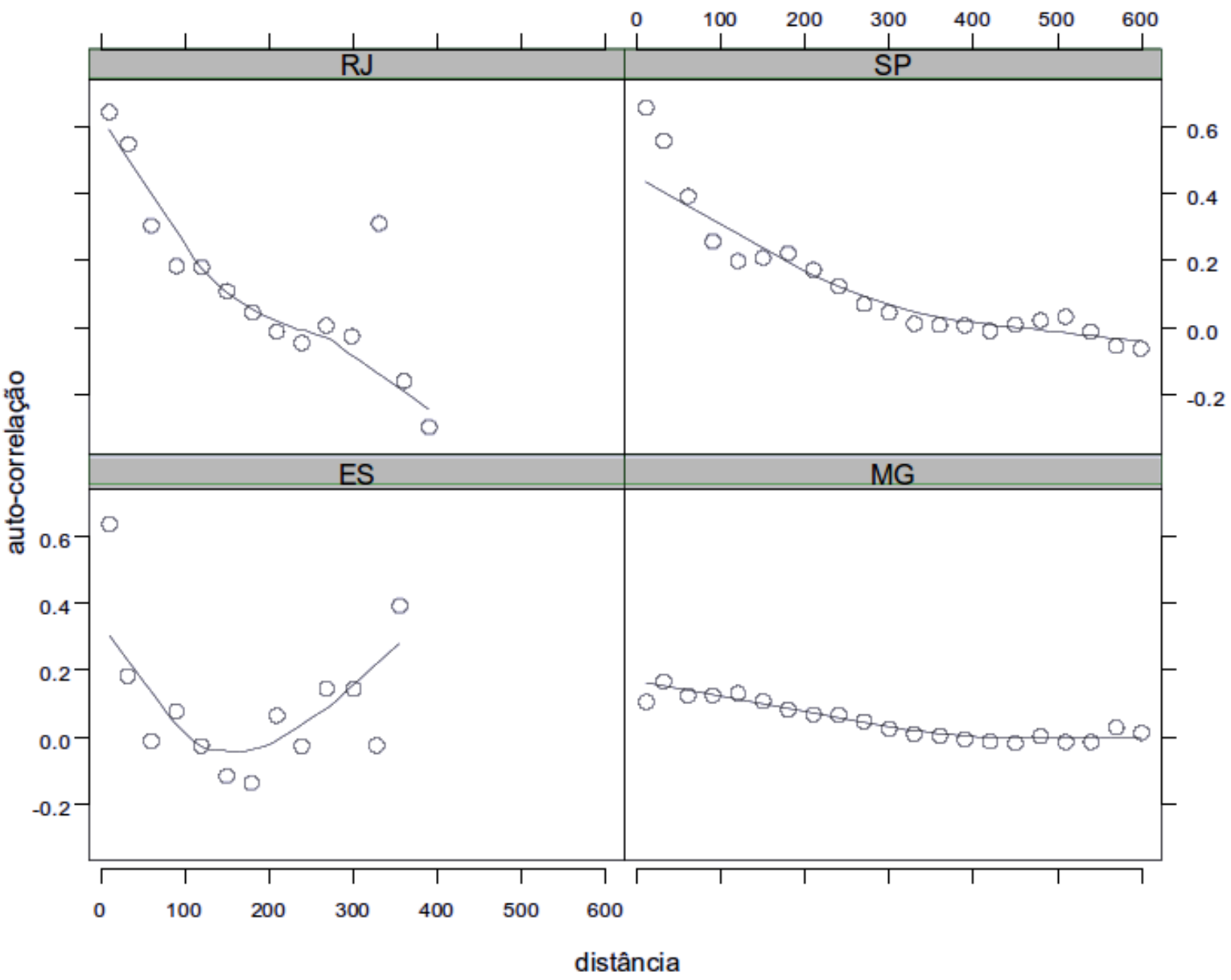
$$I^{(k)} = \frac{N}{\sum_i \sum_j w_{ij}^{(k)}} \frac{\sum_i \sum_j w_{ij}^{(k)} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})}$$

- k is the spatial lag
- Statistical significance – either permutation tests or Z test if x is normal

Correlogram



Correlogram



Local tests

- To find particularities in spatial pattern:
  - cluster
  - anomalous areas
  - more then one spatial underlying process
- LISA – Local indicators of spatial association
- Kulldorff scan statistics

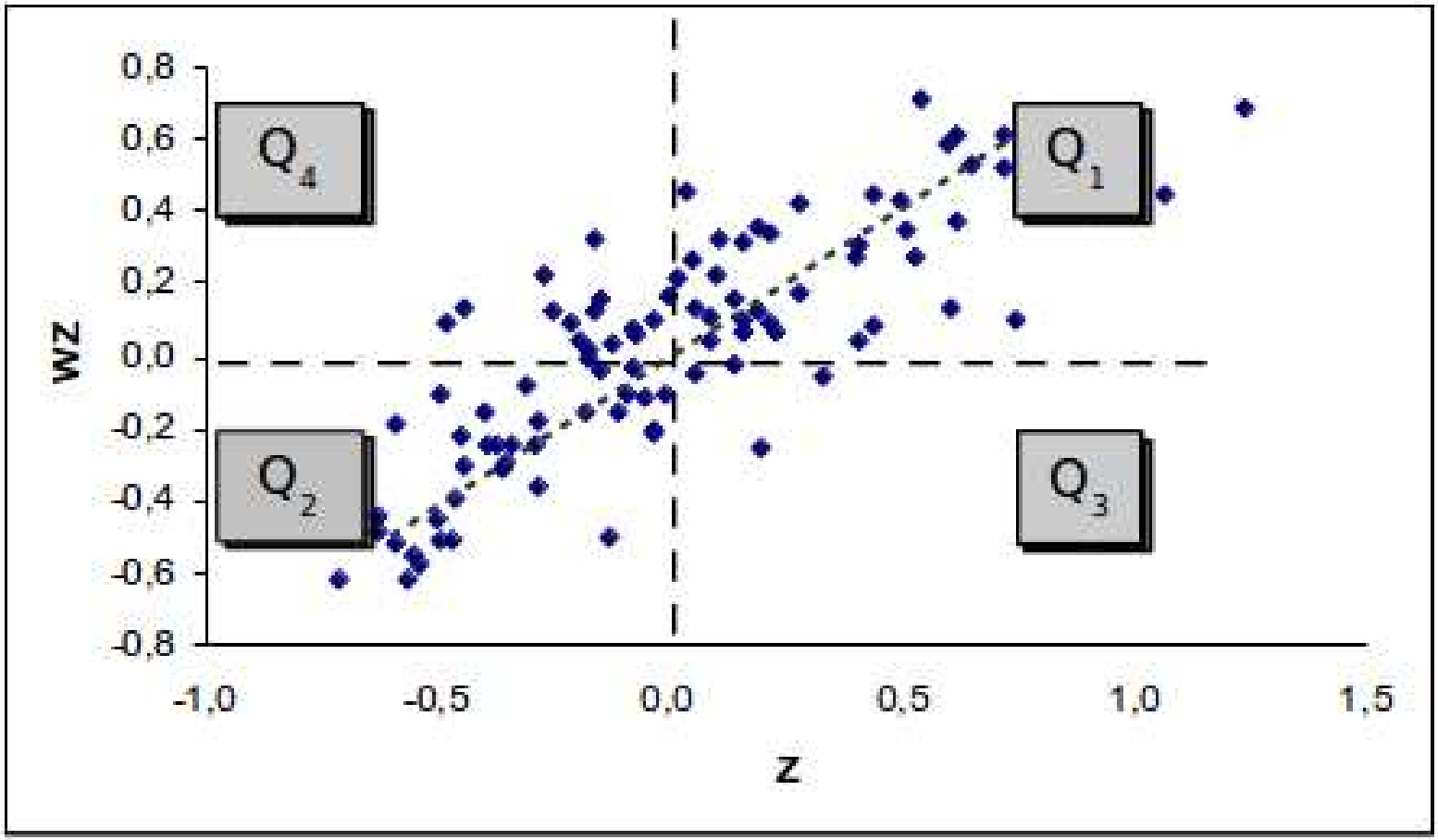
LISA

- The same idea as I, but instead of the global mean the statistics is based on local mean:

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

- $z_i \rightarrow$ the difference relative to the global mean
- If average of neighbours is similar to the area $i, I \rightarrow 1$
- Test of significance by permutation

Diagram LISA



Scan test – hypothesis

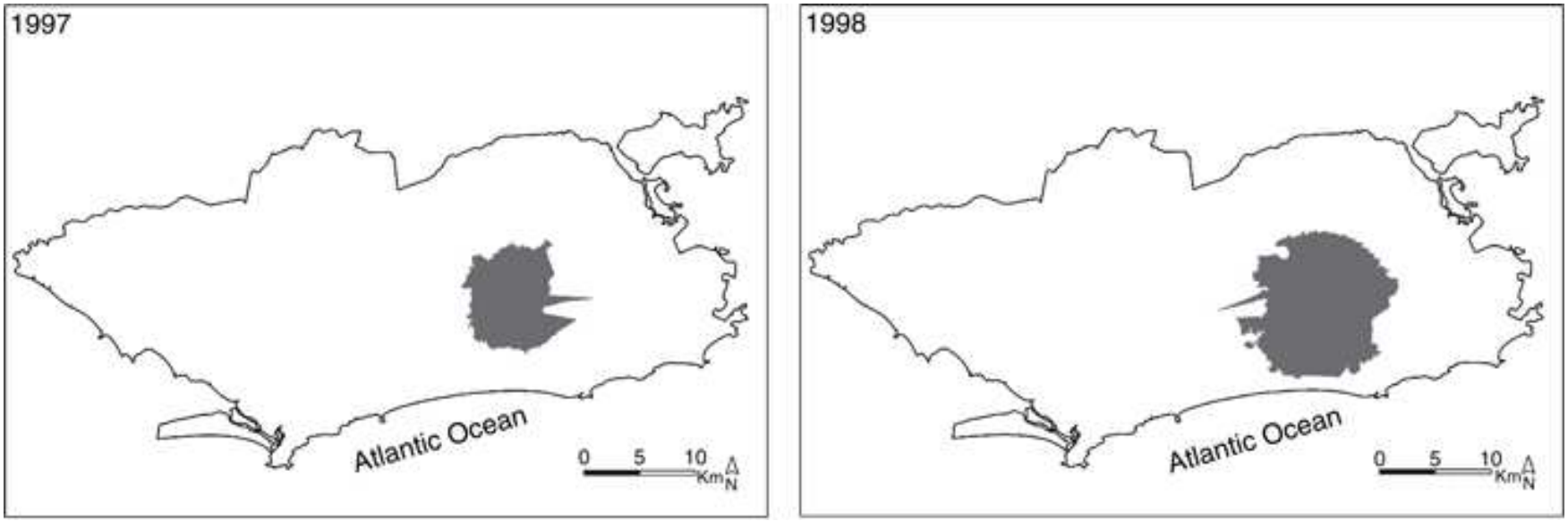
- Number of expected cases $= \lambda \times$ population at risk
- $\lambda \rightarrow$ rate, under H_0 constant over all region
- scan statistics $\rightarrow$ SaTScan[1]

[1] http://www.satscan.org/

Scanning

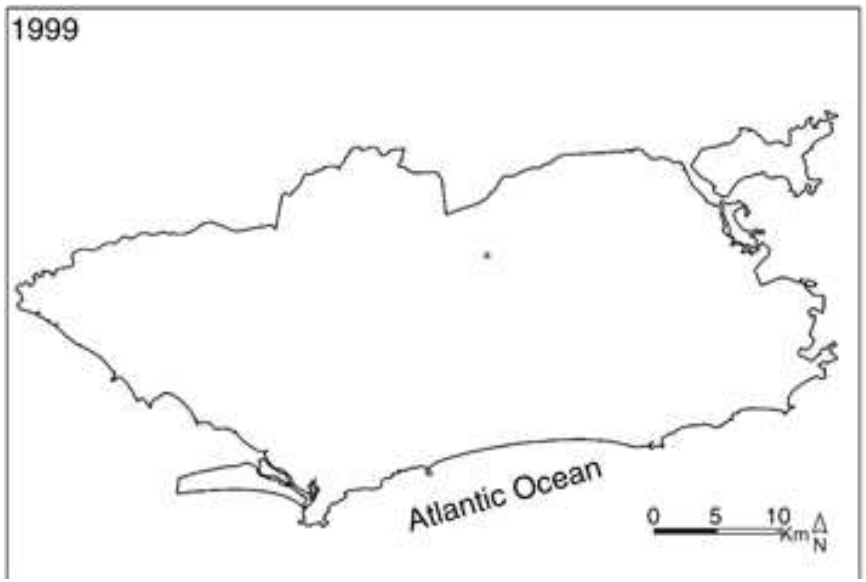
- For each area i, with population pop_i define a centroid
- Begin the search around each centroid, in circles
- Increase the radius of the circle, and at each step estimate the likelihood ratio between λ of the area inside the circle and the outside region
- The maximum is called $\rightarrow$ primary cluster
- The statistical significance is given by simulation
- The time can be included easily, considering instead of a circle around each centroid, a cylinder, the height increasing with time

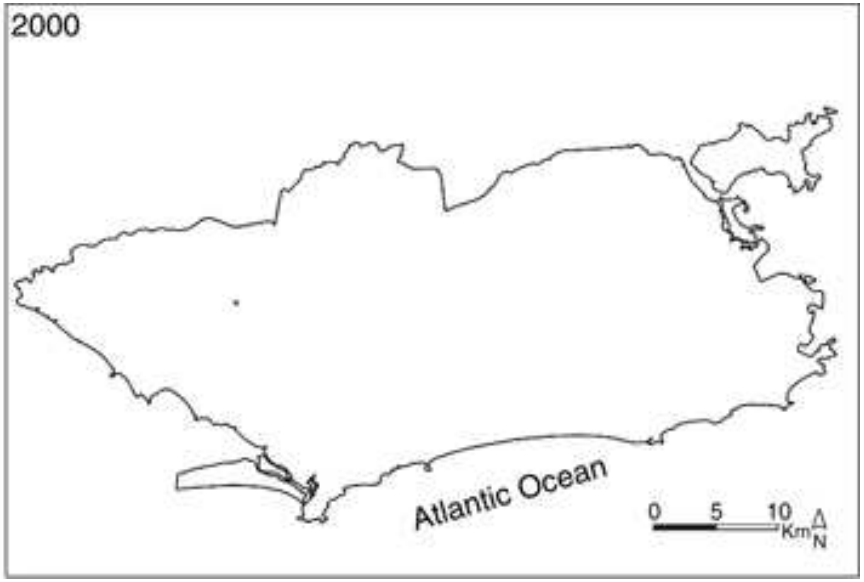
Leptospirosis no Rio – 1997-1998[1]



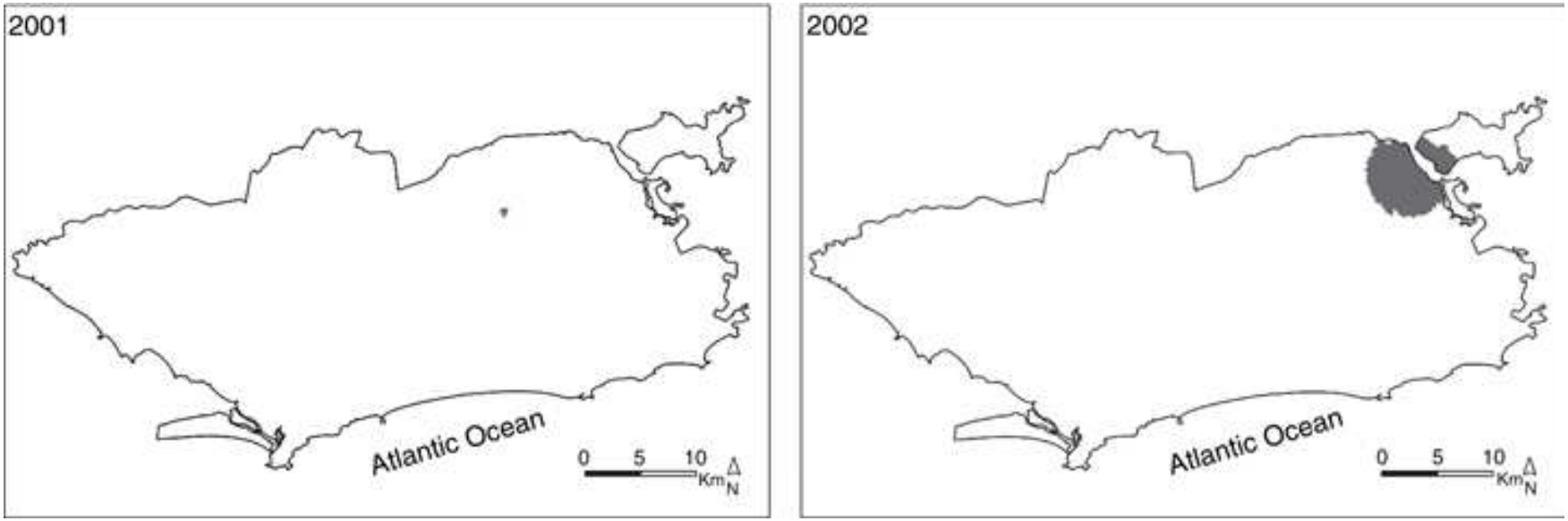
[1] Tassinari, WS; Pellegrini, DCP ; Sá, CBP ; Reis, RB ; Ko, A ; Carvalho, MS. Detection and modelling of case clusters for urban leptospirosis. TM & IH. 13:503-512, 2008.

Leptospirosis no Rio – 1999-2000





Leptospirosis no Rio – 2001-2002



Leptospirosis no Rio – Summary

Table 2. Characteristics of leptospirosis case clusters identified between 1997 and 2002

Cluster	1	2	3	4	5	6
Time span (days)	21	24	15	14	18	25
Time frame	04/01/97–28/01/97	07/01/98–30/01/98	04/03/99–20/03/99	23/09/00–06/10/00	28/04/01–25/05/01	03/01/02–27/01/02
Cluster area (km^2)	24.96	50.26	0.20	0.05	0.14	17.69
No. of cases	13	19	2	2	2	5
Population in cluster area	402 325	566 208	3361	1811	5906	478 952
Cluster attack rate (cases per 10 000 person-years)	5.10	5.62	144.80	287.92	68.67	1.52
Relative risk*	24.50	29.45	867.05	1393.24	446.42	12.68
P-value	0.001	0.001	0.291	0.161	0.590	0.973

*Relative risk was calculated as observed/expected ratio.

Outline

GAM & spatial models

- All distributions allowed, as in time series
- As usual: poisson, quasipoisson or negative binomial:
- Always use *offset*:
  - log of population at risk
  - log of expected number of cases, estimated using the overall mean
- Space coordinates are the centroids of the areas – either geometric or population's
- Counts are discrete, but underlying process is continuous

GAM

- 'NULL" model, only space distribution:

$$y_t = 1 * \log(pop) + s(coordX, coordY) + \varepsilon$$

- This model allows variation on mean (trend)
- Including covariates

$$y_t = 1 * \log(pop) + \beta_0 + s(coordX, coordY) + s(cov) + \varepsilon$$

- Covariates may be entered linearly or using a spline
- For each covariate included, see if there still exists any spatial structure
- The spatial part of the model can present interaction with another variable (i.e. sex, education)

When to use GAM

- Space is continuous and isotropic
- The spatial term models variation on the mean (trend)
- The places where the centroids is farther apart the confidence interval is larger.
- This mode does not take into account border effect – CI is larger, smooth spatial effect goes to areas outside the border
- As all spatial information is condensed at the areas centroid, any information on connection between areas is lost – only weight matrix is distance between centroids
- Spatial term is purely additive – it is summed to the effect of other covariates
- It is a very good exploratory approach to spatial modelling

Outline

The problem

- Data is collected aggregated in areas
- Autocorrelation tests are positive
- We want to model the association of some important variable on the number of cases of the disease, but we do know that spurious association are present if both present similar trend

Outline

Spatial Auto Regressive Models – SAR

- It is an extension of GLM that includes a spatial dependence term

$$Y \sim N(\mu, \sigma)$$
$$Y = \mu + e$$
$$\mu = \beta X$$

Y $\rightarrow$ mortality or morbidity rate

X $\rightarrow$ covariates

$\beta \rightarrow$ parameters to estimate

e $\rightarrow$ error

SAR

- To include a spatial term:

$$e_i = \sum_{i=1}^{m} b_{ij}\, e_i + \varepsilon_i$$

$i \rightarrow$ areas index

$b_{ij} \rightarrow$ parameter of dependence between neighbouring areas i and j

e $\rightarrow$ covariance between i e j

SAR

- If neighbourhood matrix is symmetric, the model can be re-parametrised as:

$$\sum_{\varepsilon_i} = \sigma^2 I$$

$$Var[Y] = \frac{\sigma^2}{(I - \lambda)^2}$$

$\sigma^2 \rightarrow$ residual variance

$I \rightarrow$ identity matrix

$\lambda \rightarrow$ spatial autocorrelation matrix

Estimating SAR

- Estimation:
  - First the autocorrelation parameters are estimated using Maximum Likelihood
  - The regression parameters ares estimated using least squares:
- Not available for GLMs
- Estimated parameters represent population averages, controlled for the effect of spatial dependence

Adaptations for SAR models

- To take into account the population size $\rightarrow$ include population as a weight
- When the dependent variables, conditioned by covariates, is not normal $\rightarrow$ variable transformation
- If log transformed $\rightarrow$ use $log(pop)$ for weights

Outline

CAR

- Similar to SAR, but for dependence structure

$$e_i | e_j \sim N\left(\sum_j \frac{c_{ij}\, e_j}{\sum_j c_{ij}}, \frac{\sigma^2}{\sum_j c_{ij}}\right)$$

j neighbour of i

$c_{ij} \rightarrow$ covariance between i and j, if neighbours

$\varepsilon_i \rightarrow$ residuals

CAR

- The neighbourhood matrix strongly affects both models (SAR and CAR)
- In R the function `nb2listw` allows exploring different neighbourhood matrix
- The default is standardise for each area, making the weight $= 1 \rightarrow$ total sum of weights $= n$ (number of areas)
- Observe that in this case the neighbourhood matrix is not symmetrical
- SAR models are OK, but CAR are not!

Steps to modelling

1. Estimate the `lm` as usual, verify residuals and results

```
mod <- lm(rate ~ cov, data=dataset)
summary(mod); plot(mod)
```

2. If residuals are not reasonably gaussian $\rightarrow$ transform the variable
3. Be careful with values zero $\rightarrow$ sum 1 before estimating the rate
4. If the size of population is very different among areas $\rightarrow$ include population as weight

```
modw <- lm(ratelog ~ cov, weigths=log(pop), data=dataset)
summary(modw)
plot(modw)
```

5. Assess dependence of residuals

```
moran.test(residuals(modw), neigh.weights)
```

Steps – going on

6. If residuals present spatial structure $\rightarrow$ model SAR or CAR

```
mod.sar <- spautolm(ratelog ~ cov, weights=log(pop),
           family = "SAR", listw=neigh.weights,
           data=dataset)
summary(mod.sar)
mod.car <- spautolm(ratelog ~ cov, weights=log(pop),
           family = "CAR", listw=neigh.weights,
           data=dataset)
summary(mod.car)
```

Steps – going on

7. There is no `plot` function to plot residuals
8. Do not worry about spatial dependence of residuals, the models do care about it
9. The choice between SAR and CAR can be based on AIC

Outline

Random effects

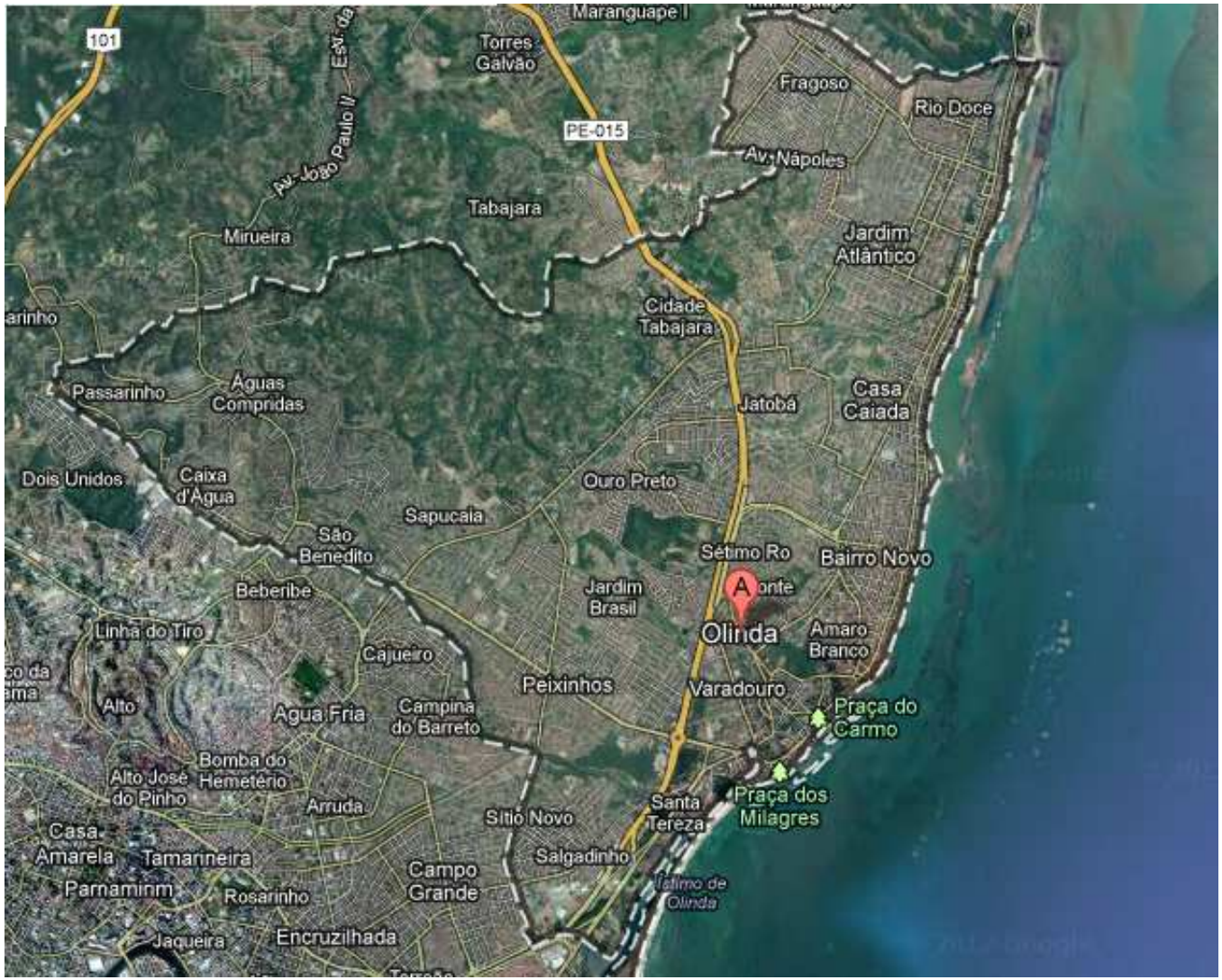
- The errors e_i from equation $y_i = \beta_0 + \beta_1 x_1 + \cdots + e_i$ can be treated as random effects, varying between areas
- Those e_i may present structures such as:
  - spatial weights, as in CAR and SAR
  - just random intercept
  - random intercept varying according to some specific region
  - a mix

How to fit

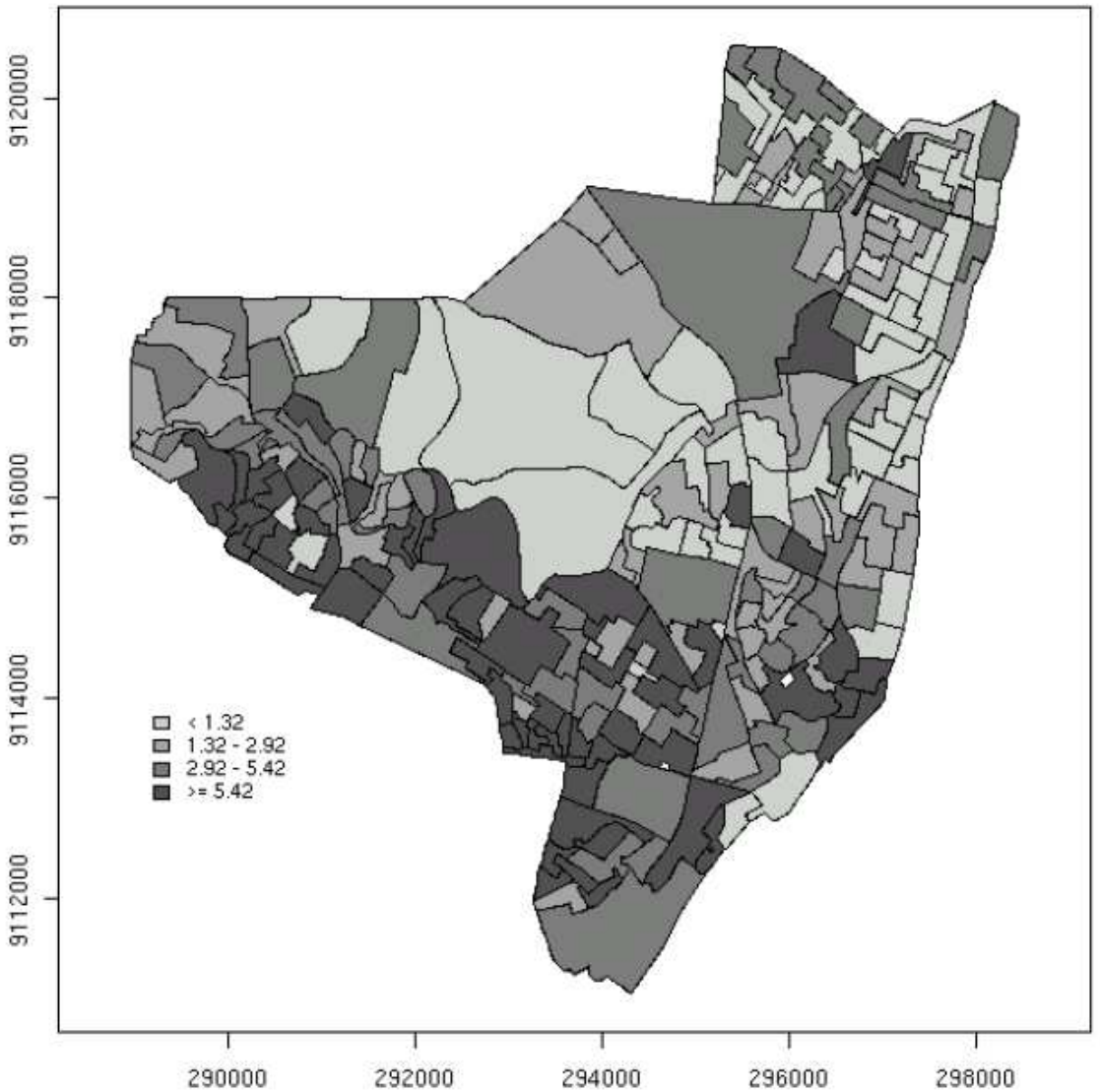
- library `lme4`
- Reference: Pinheiro, J. C.; Bates, D. M.. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000.

Outline

Olinda


101
Maranguape I
Torres Galvão
Fragoso
Rio Doce
PE-015
Av. Nápoles
Av. João Paulo II
Tabajara
Mirueira
Jardim Atlântico
Cidade Tabajara
Passarinho
Águas Compridas
Jatobá
Casa Caiada
Dois Unidos
Caixa d'Água
Ouro Preto
Sapucala
São Benedito
Sétimo Ro
Bairro Novo
Beberibe
Jardim Brasil
Olinda
Amaro Branco
Linha do Tiro
Cajueiro
Peixinhos
Varadouro
Praça do Carmo
Alto
Água Fria
Campina do Barreto
Bomba do Hemetério
Alto José do Pinho
Arruda
Sítio Novo
Santa Tereza
Praça dos Milagres
Casa Amarela
Tamarineira
Salgadinho
Parnamirim
Rosarinho
Campo Grande
Jaqueira
Encruzilhada

Leprosy in Olinda



Poverty in Olinda