The Abdus Salam
International Centre
for Theoretical Physics

2453-12

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

15 - 26 April 2013

Point Event Analysis

SA CARVALHO Marilia

PROCC FIOCRUZ
Avenida Brasil 4365
Rio De Janeiro 21040360
BRAZIL

Point Event Analysis

Marilia Sá Carvalho

Fundação Oswaldo Cruz

Outline

1. Introduction
2. Exploratory Analysis
3. Hypothesis tests
4. Modelling with location
5. Dengue fever

Outline

1 Introduction

References

- Bailey,T.C. and Gatrell,A.C.. *Interactive Spatial Data Analysis*. Longman, 1996.

- Baddeley, A and Turne, R. Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1-42, 2005.

- Baddeley, A. *Analysing spatial point patterns in R*. Workshop Notes, Version 4.1, 2010. Available at http://www.spatstat.org/spatstat/.

- Wood, S.N.. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC Texts in Statistical Science Series, 2006.

What is point data

- The simplest spatial data

Event	Coord X	Coord Y
1	3.5	0.34
2	1.6	0.56
3	9.2	1.45

Definition

- An observed point pattern x is a realisation of a random point process X in two-dimensional space:
  - The number of points is random
  - The locations of the points is random
- The aim: to estimate parameters of the distribution of X
- Another: to estimate the effect of a given covariate on the observed pattern
- Or: to model the point process

Point pattern state-of-art

- Techniques to fit realistic models to point pattern data are new (2000's)
- Most applied work is based on hypothesis testing, to detect whether the point pattern is completely random
- We will try to cover both the classical tests and a bit of modelling
- The main reason to include this topic in this course is the availability of data (GPS!), that is generally poorly analysed

Is this a point process?

- Location of dengue fever cases
- Number of *Aedes aegypti* collected in a random sample of households
- Results (Positive/Negative) of a dengue fever seroprevalence survey in a random sample of households
- Original dataset is the counts of leprosy by census tract. As those areas area very small, can we use the centroid?

Marked point process

- Results (Positive/Negative) of a dengue fever seroprevalence survey
- Counts of leprosy by census tract

$$y = (x_1, m_1), ..., (x_n, m_n), x_i \in W, m_i \in M$$

Marks and Covariates

- Marks are “response” variable, integrating the pair of plane coordinates
  - Time
  - Positive/negative results of tests
  - Counts
  - Size of trees
- Covariates are explanatory variables
  - Income
  - Education
  - Rainfall
  - Temperature

Intensity

- Average density of points per unit area
- May be constant → uniform or homogeneous
- May vary from location to location → inhomogeneous
- First step in analysing a point pattern

Theory

- If $\boldsymbol{X}$ is homogeneous $\rightarrow$ in any sub-region B of two-dimensional space the expected number of points in B is proportional to the area of B:

$$E[N(\boldsymbol{X} B)] = \lambda \text{area(B)}$$

- The constant of proportionality λ is the intensity
- If a point process is homogeneous, then the empirical density of points is:

$$\hat{\lambda} = \frac{n(x)}{area(W)}$$

- $\hat{\lambda}$ is an unbiased estimator of the true intensity λ

Complete Spatial Randomness

- The basic reference model of a random point pattern is the uniform Poisson point process in the plane with intensity $\lambda \rightarrow$ Complete Spatial Randomness (CSR)
- Properties:
  - the number of points falling in any region A has a Poisson distribution with mean λ
  - the n points inside A are uniformly distributed inside A
  - the contents of two regions A and B are independent
- Uniform Poisson process are the *null model* in a statistical test

Stationary & Isotropic

- If the process is stationary:
  - $\lambda(x) =$ constant
  - Invariant to translation
- If the process is isotropic:
  - $\lambda(x, y) = \lambda|h|$ (h=distance between x and y)
  - Invariant to rotation
- Most hypothesis tests assume a stationary and isotropic process
- It is a global feature

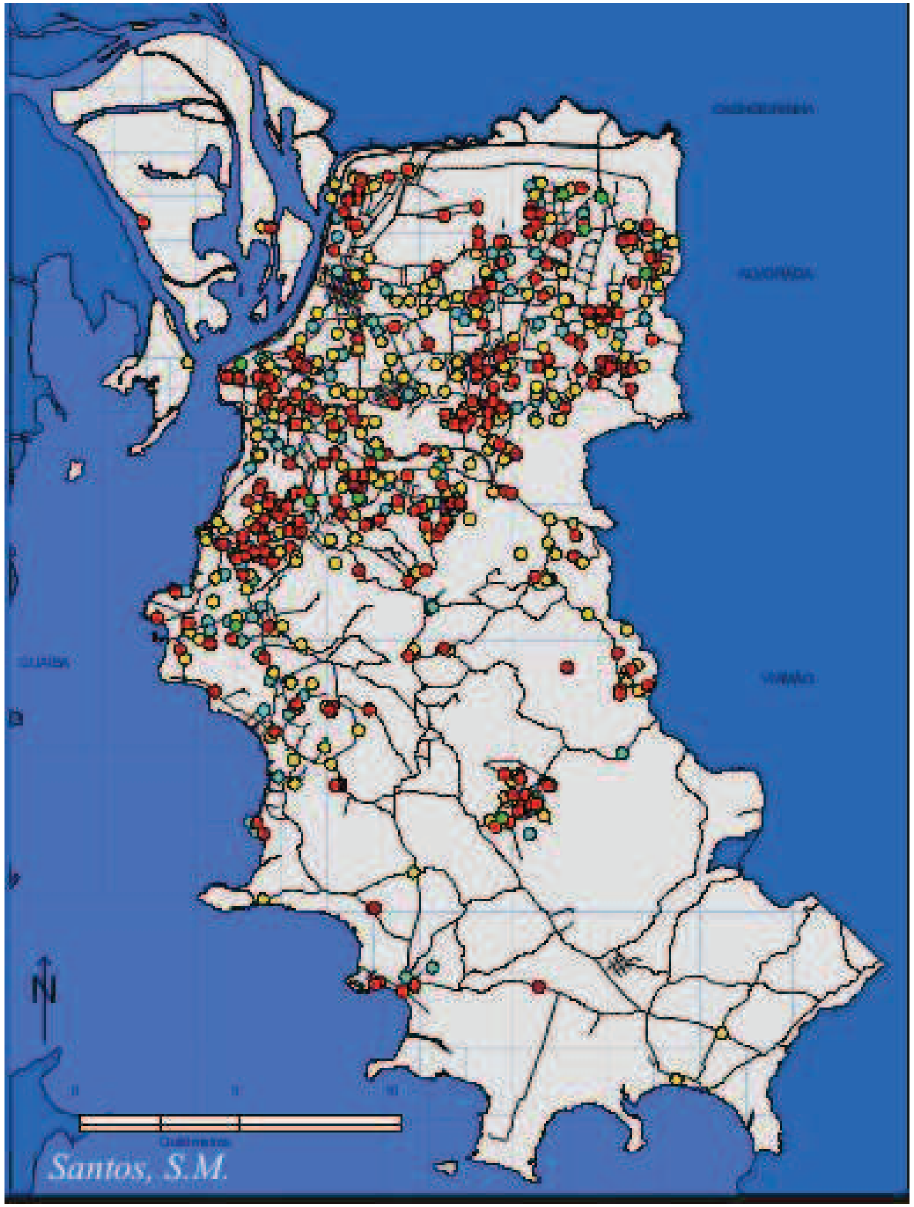
Interaction

- Stochastic dependence between the points
- Interaction generates either clusterisation or repulsion
- Distance between points are used to investigate interaction
- It is a local feature

Outline

2 Exploratory Analysis

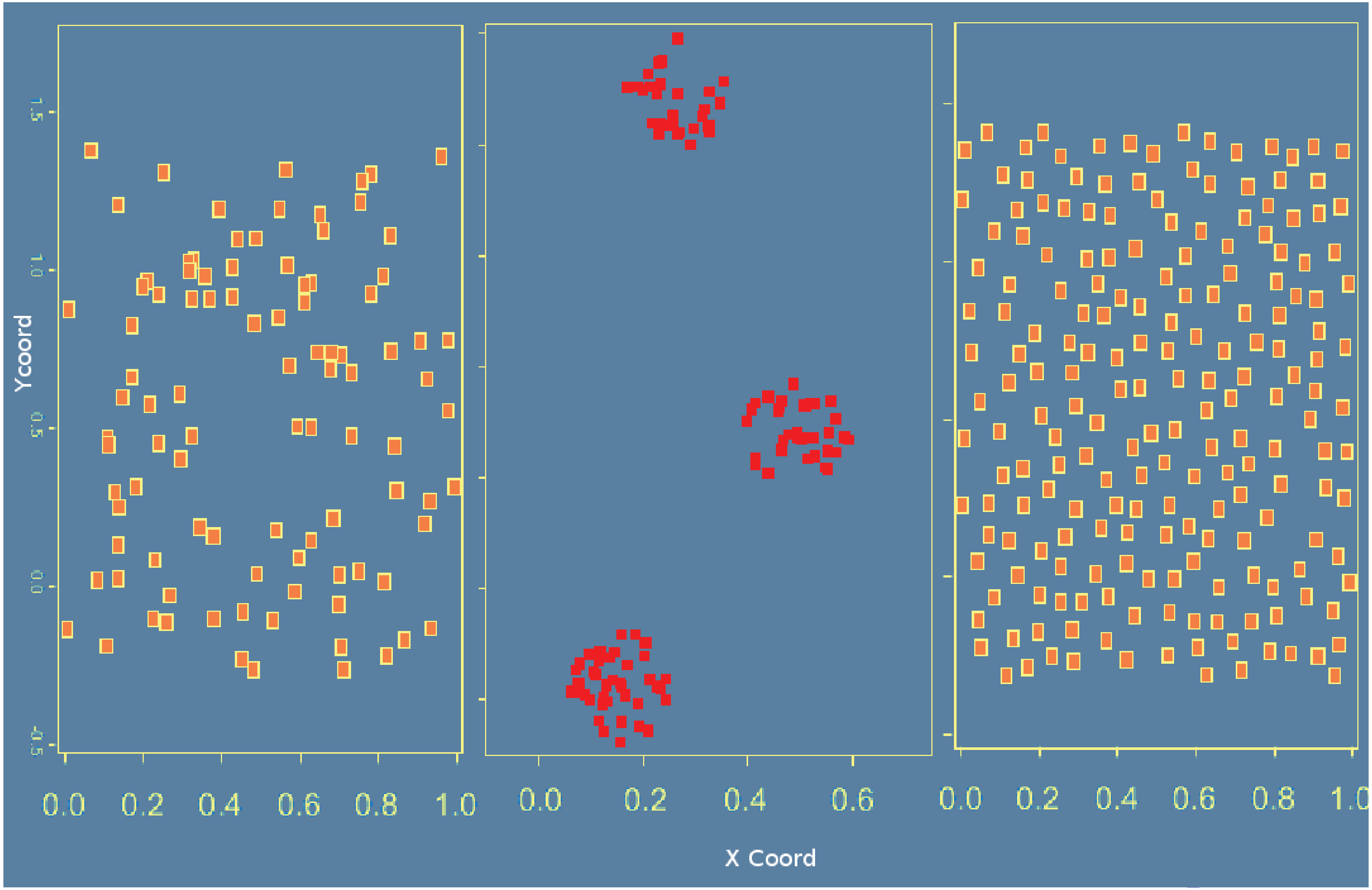
Map of Points: External Causes Mortality



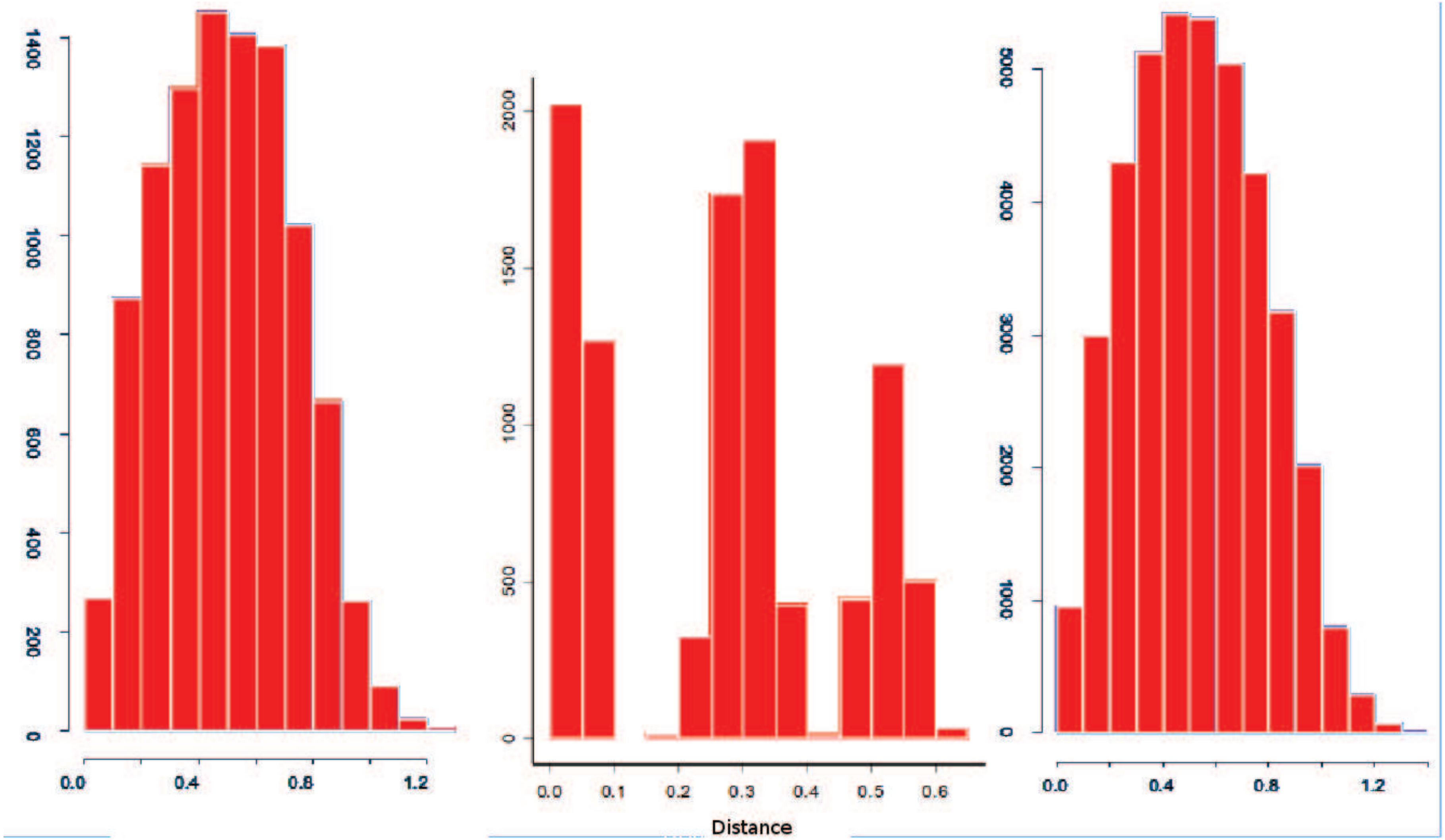
Map of Points

- The simplest!
- It allows brief inspection of spatial patterns
- Different types of events can be depicted
- However... events in human populations follow the demographic pattern

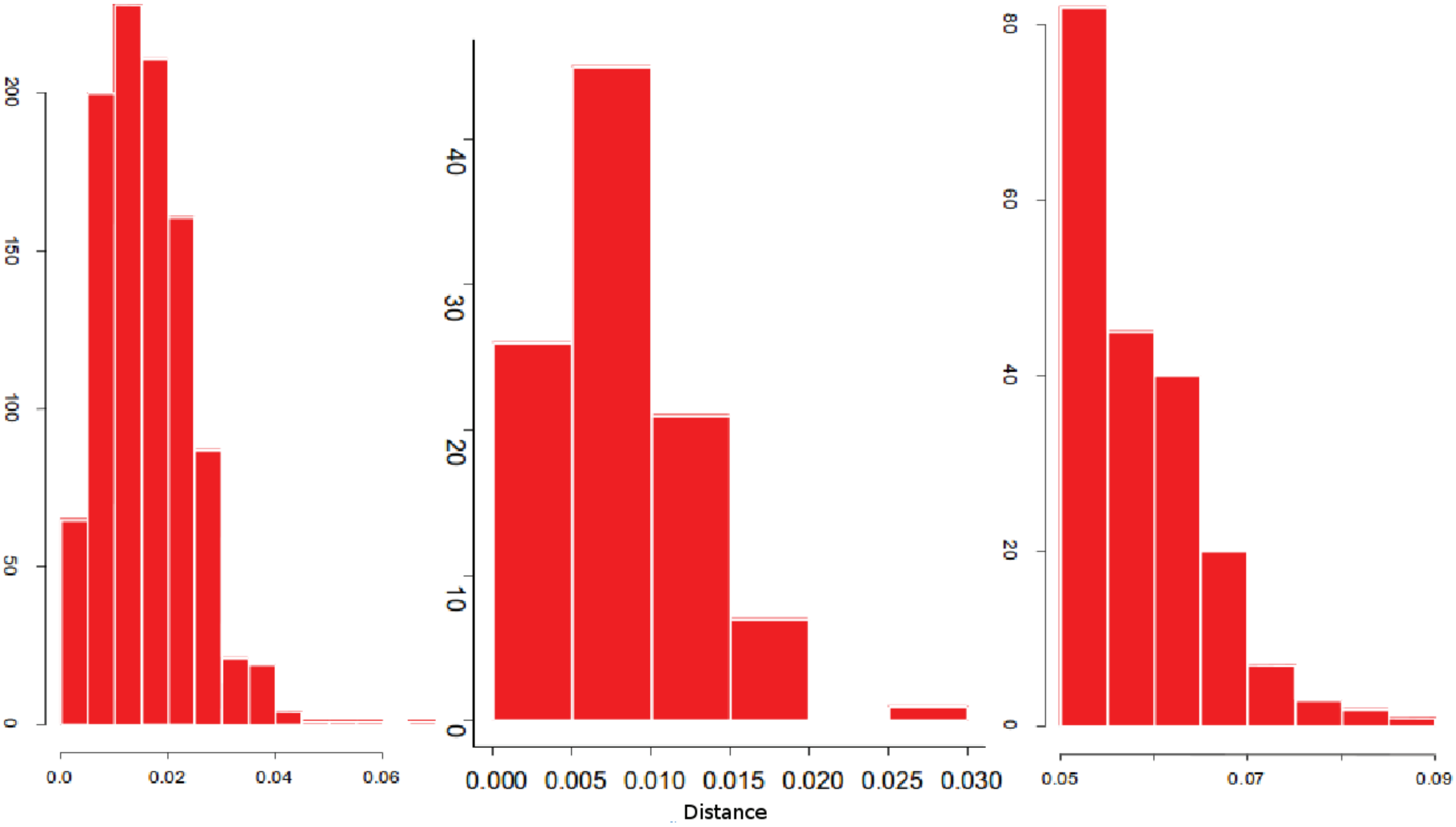
Different spatial patterns



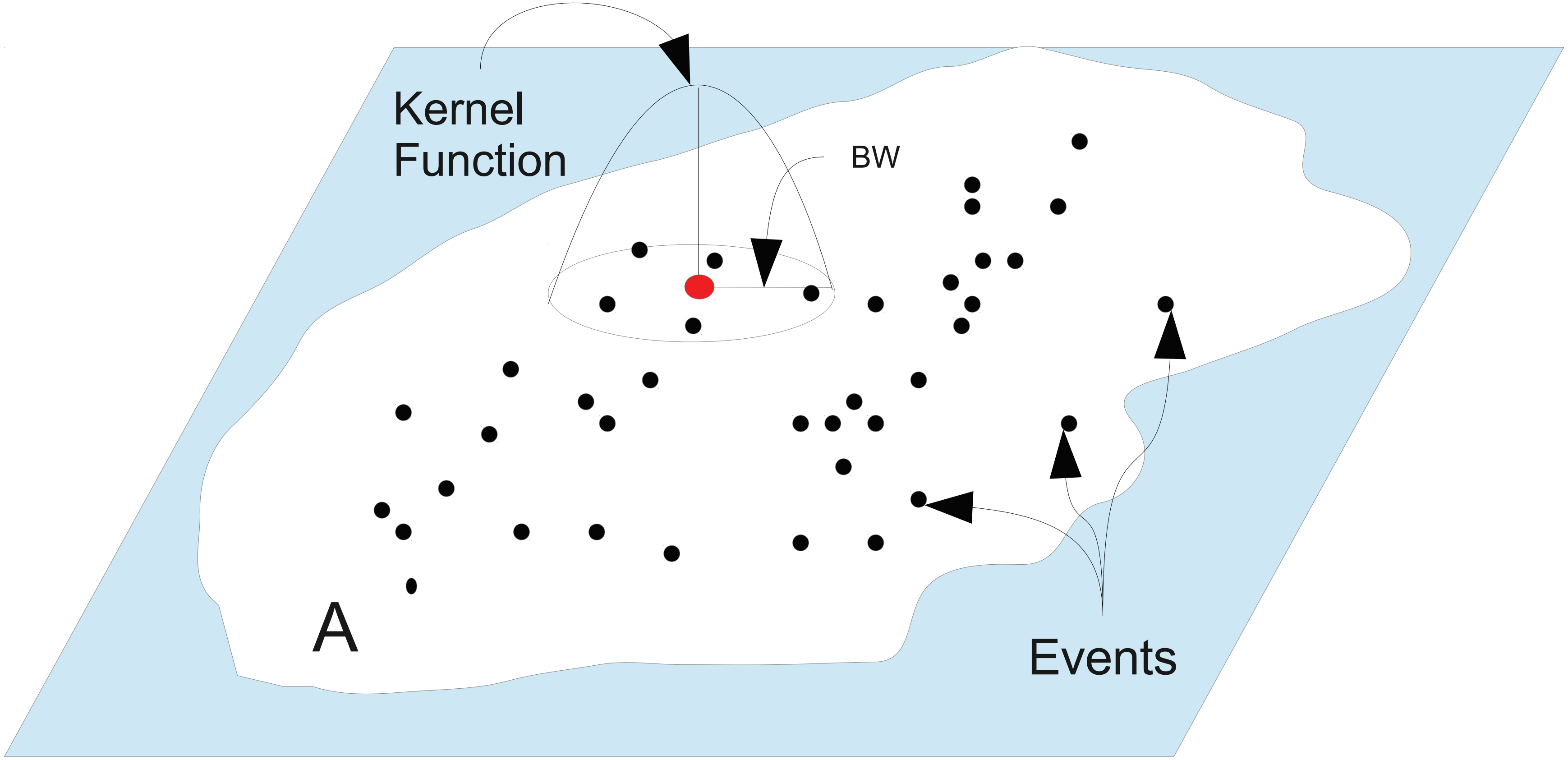
Different spatial patterns



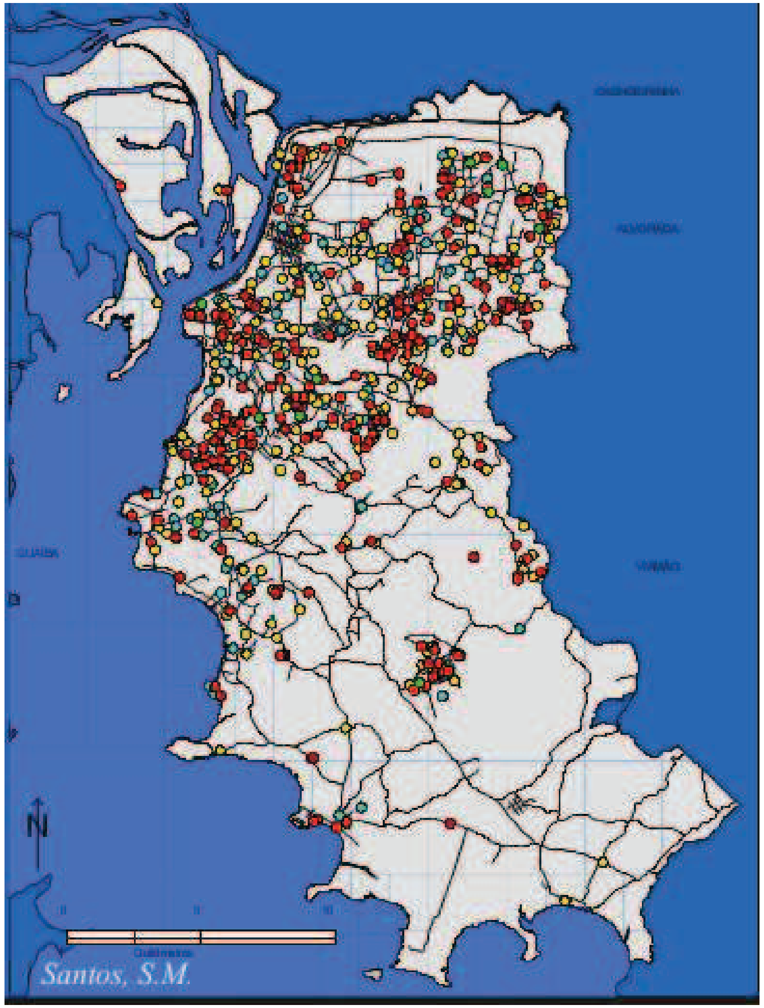
Different spatial patterns

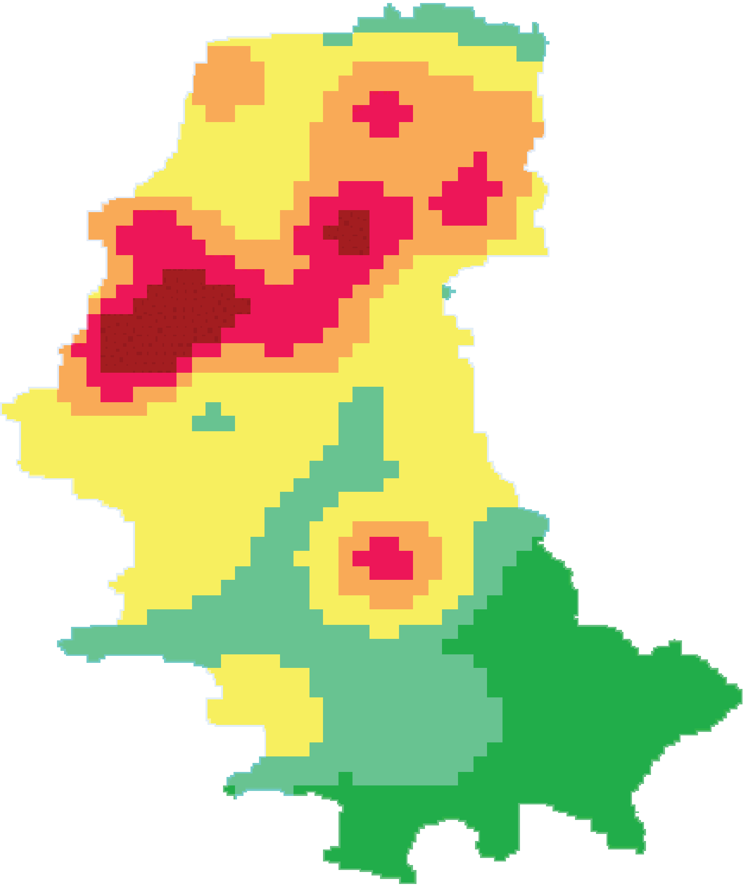


Kernel



Kernel Map: Homicides





Outline

3 Hypothesis tests

Cluster detection

- A cluster is a group of events geographically limited in size and density that is improbable to be due to randomness (Knox)
- Causes of clusters: common source, contagion
- In general space and time concentrated
- To take into account:
  - Various risk factors – age, population density
  - Place of living and place of working
  - Latency period
- Two types of tests:
  - focused – around a suspected source
  - generic

Many tests for CSR

- χ^2 for quadrats
- Kolmogorov-Smirnov
- Maximum likelihood for Poisson processes
- Nearest neighbour distances:
    - G function
- Ripley's K-function with Monte Carlo envelope

Pairwise distances and K-function

- The observed pairwise distances $r_{ij} = \|x_i - x_j\|$ in the data is a biased sample of pairwise distances in the point process, in favour of small distances
- Observed K-function:

$$\hat{K}(s) = \frac{1}{\lambda n} \sum_{i \neq j} I(d_{ij} < s)$$

- If CRS → $\hat{K}(s) = \pi s^2$
- Envelope – dimulation

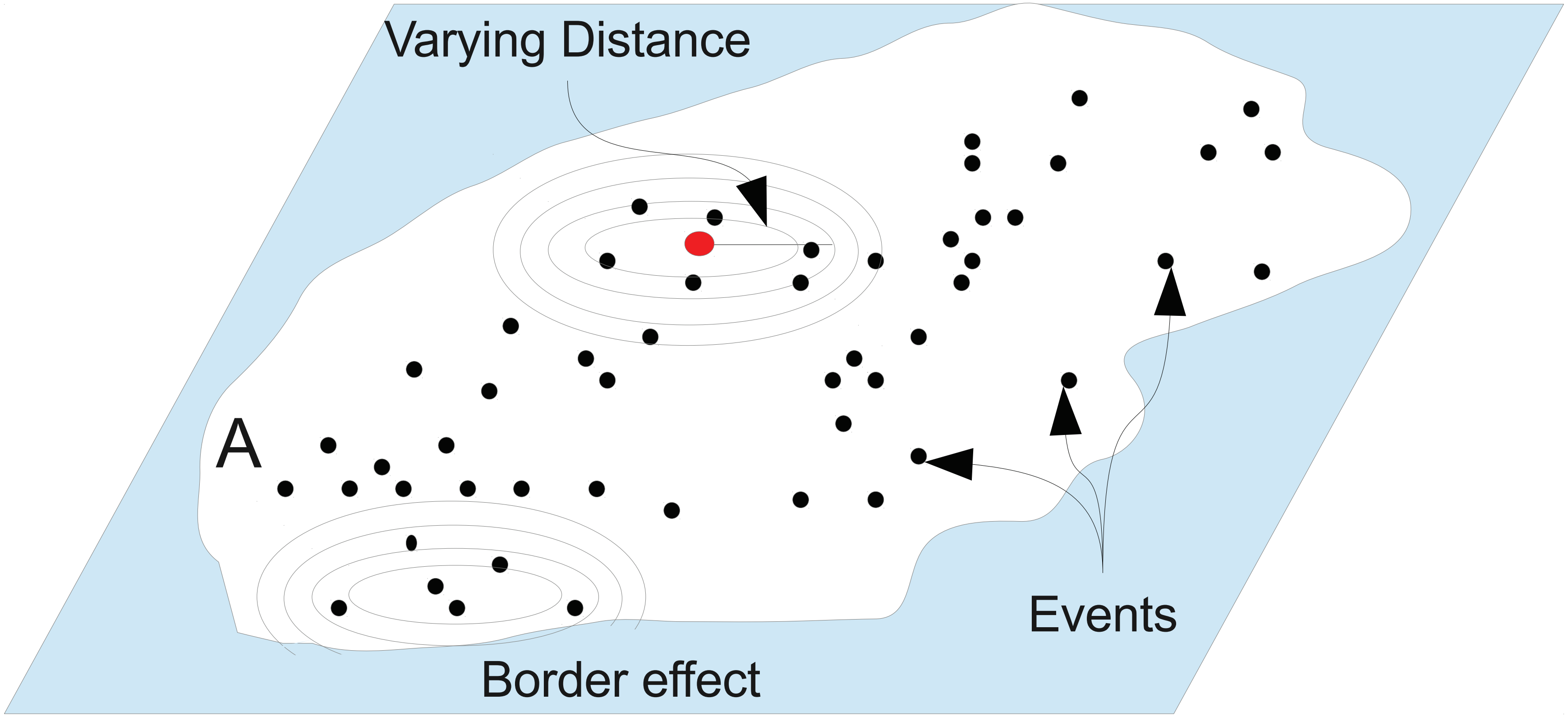
L-function

- Variance stabilised:
- L-function:

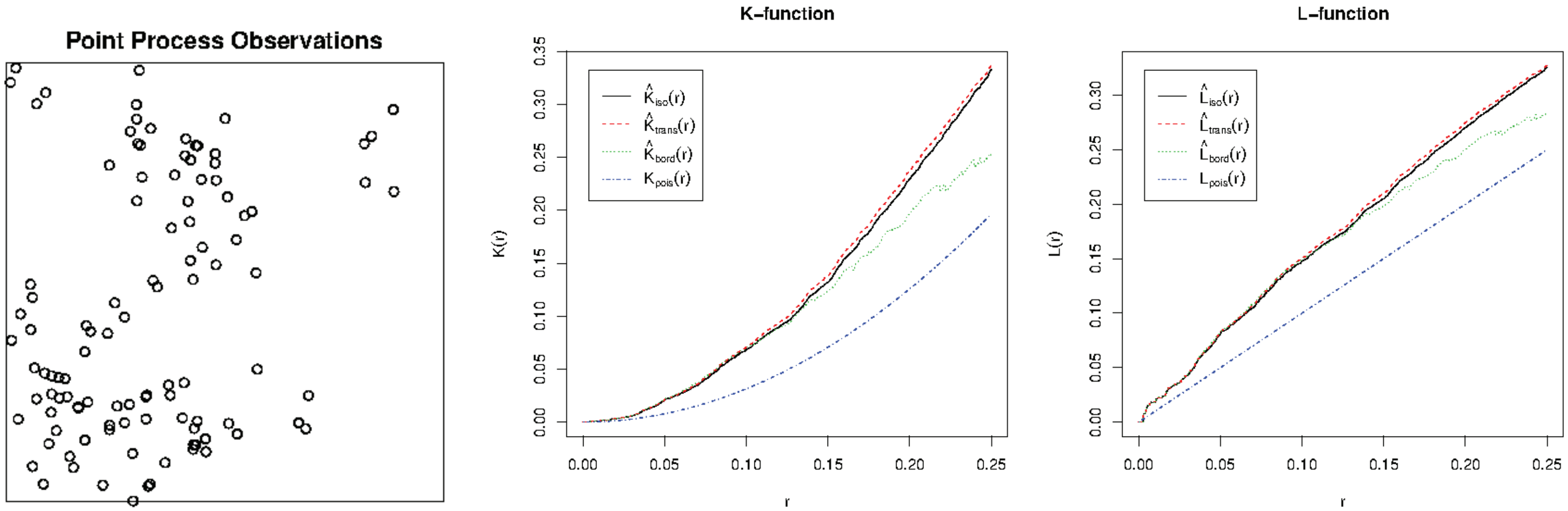
$$\hat{L}(s) = \sqrt{\frac{\hat{K}(s)}{\pi}}$$

- If CSR $E(L) = s$

K and L-function



K and L-function



Be careful

- K-function (and other tests – F, G) assume the process is stationary
- Difference between the empirical and theoretical functions are not evidence of cluster, but may be just the variation in intensity over the large scale

Focused tests

- The cluster is around a point or a line
- Is there an excess number of cases as compared to some control?
- Tests include a function of distance to the source

Focused tests

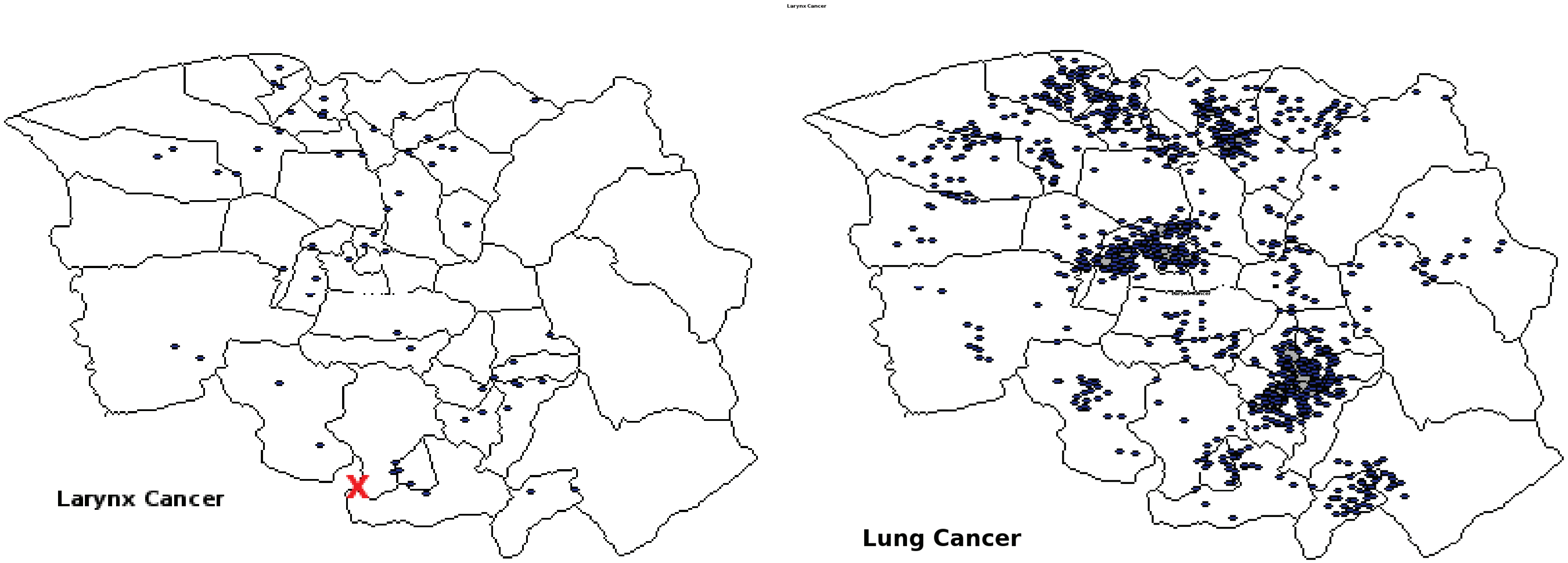


Fig.: Larynx Cancer: is it due to the incinerator?

Finding where

- Local Indicators of Spatial Association (LISA)
- Scan tests
- Both will be discussed in Areal data class

Outline

Statistical models

- Define a statistical distribution so that:
  - functional form reflects some properties of interest
  - terms of the probability distribution have an interpretation
  - introduction of covariates is possible
- Gibbs point processes

GAM Models

- For point pattern process we need cases and controls
- Just the distribution of points could me modelled only as a Gibbs point processes
- For human diseases, controls can be negative serology or samples of demographic census, but we do need controls
- It reduces to logistic regression, with a spatial term

$$y_s = s(coordX, coordY) + covs_s \varepsilon$$

GAM Models

- Consider a model with just the spatial term
- The spatial distribution of points could be modelled as a Gibbs point processes
- But in GAM setting we use another approach, comparing the spatial distribution of cases to controls
- For human diseases, controls can be negative serology or samples of demographic census, but we do need controls
- It reduces to logistic regression, with a spatial term

$$Y \sim Bernoulli(p_i)$$
$$logit(p_i) = s(coordX, coordY) + \varepsilon$$

GAM Models

- Each covariate included could “explain” the spatial distribution
- If “explained”, the map goes flatter
- The spatial term may interact with some factor covariate

$$logit(p_i) = s(coordX, coordY) + covs_s + \varepsilon$$

Interaction Time-Space

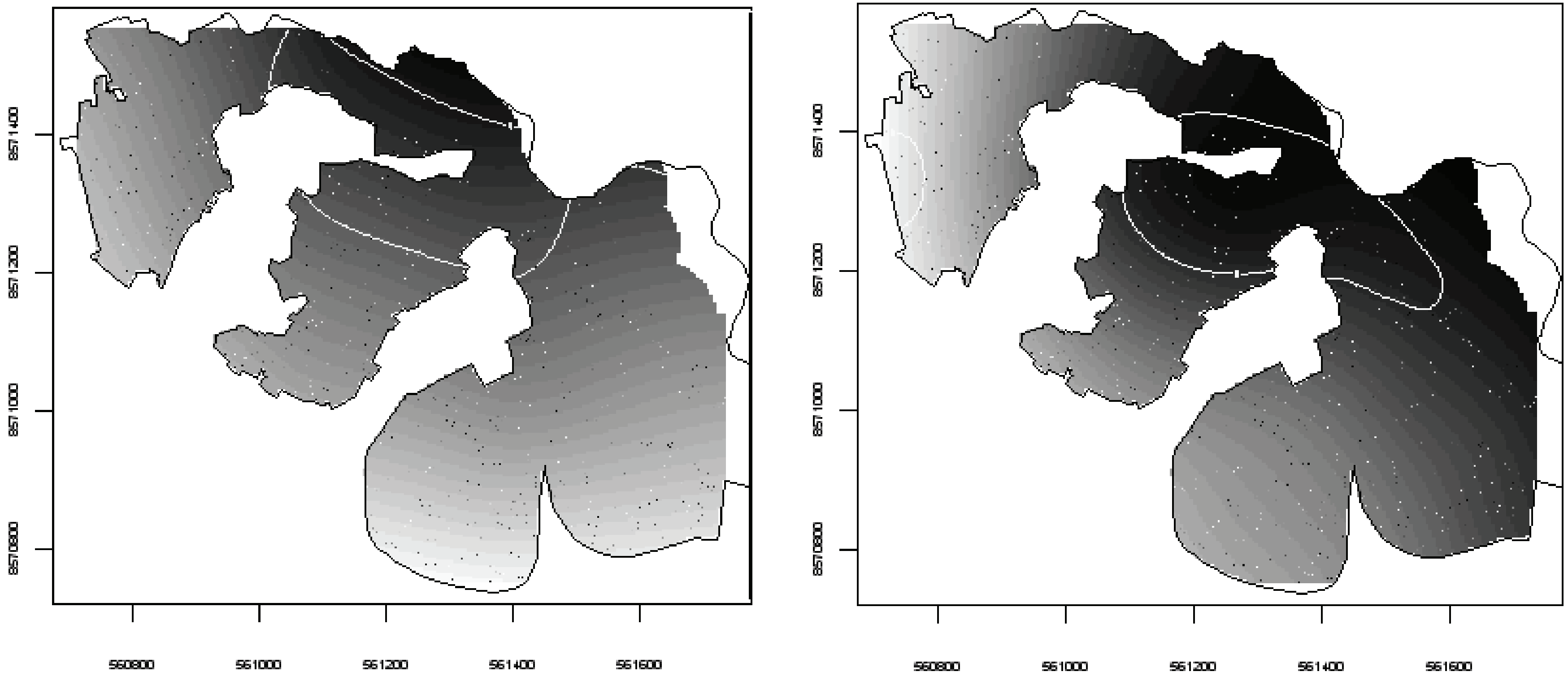


Fig.: Seroprevalence for Leptopirosis, Pau da Lima/Ba, two years follow-up

Outline

5 Dengue fever

Dengue fever

- Dengue is a mosquito-borne viral infection
- Four serotypes, no cross imunogenicity
- High frequency of asymptomatic infection, as shown by seroprevalence studies:
  - 45.5% of schoolchildren in Niterói in 1987
  - 29.2% of schoolchildren in Paracambi in 1997
  - 26.6% in pre-schoolchildren Salvador in 1998 and 33.2% in 2000
- No easy serotype identification based on IgG

Dengue fever transmission

- *Aedes aegypti* is the most important dengue vector worldwide
- It is domestic – it mates, feeds, rests, and lays eggs in and around human habitation
- Vector population is sensitive to rainfall and temperature
- Virus transmission varies with temperature

Intra-city variation

- Seroprevalence varies among different areas inside the same city:

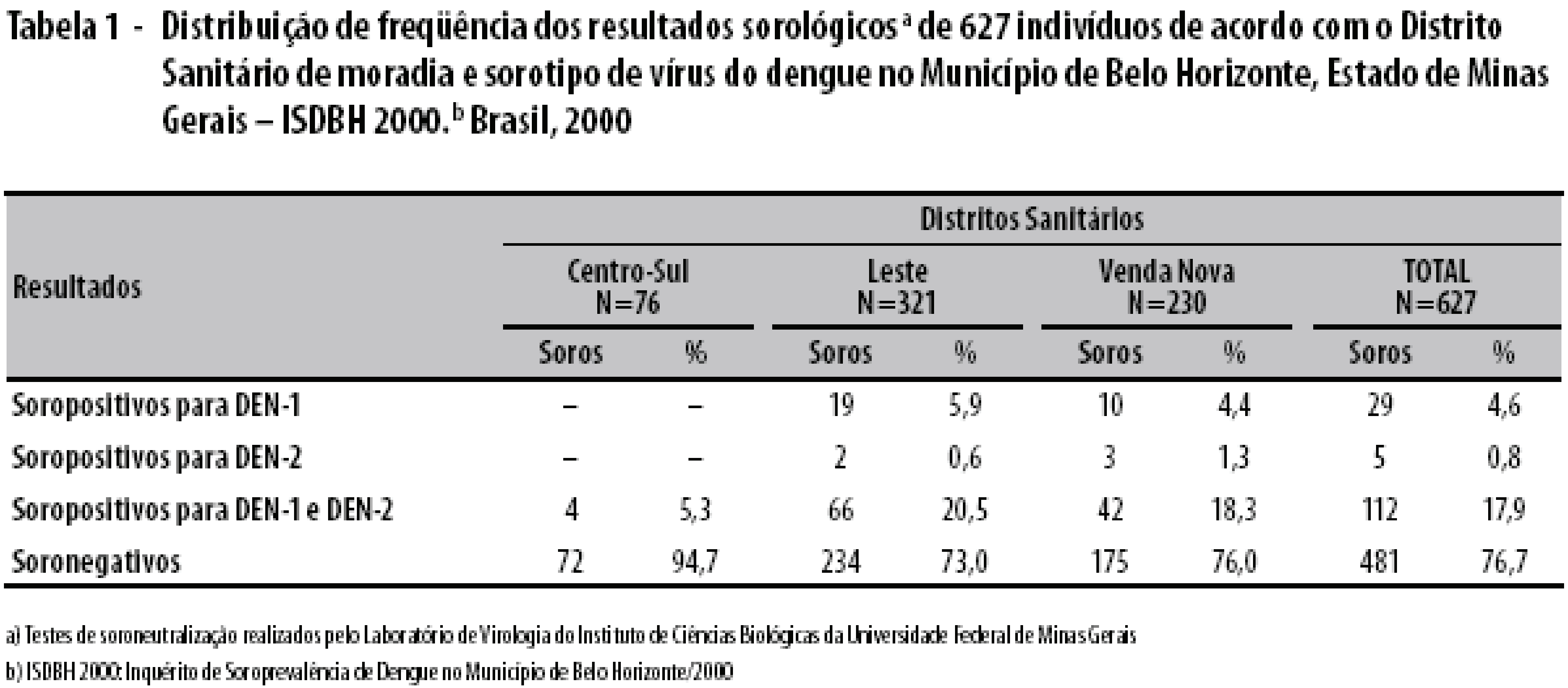


Tabela 1 - Distribuição de freqüência dos resultados sorológicos[a] de 627 indivíduos de acordo com o Distrito Sanitário de moradia e sorotipo de vírus do dengue no Município de Belo Horizonte, Estado de Minas Gerais – ISDBH 2000.[b] Brasil, 2000

Resultados	Distritos Sanitários							
	Centro-Sul N=76		Leste N=321		Venda Nova N=230		TOTAL N=627	
	Soros	%	Soros	%	Soros	%	Soros	%
Soropositivos para DEN-1	–	–	19	5,9	10	4,4	29	4,6
Soropositivos para DEN-2	–	–	2	0,6	3	1,3	5	0,8
Soropositivos para DEN-1 e DEN-2	4	5,3	66	20,5	42	18,3	112	17,9
Soronegativos	72	94,7	234	73,0	175	76,0	481	76,7

a) Testes de soroneutralização realizados pelo Laboratório de Virologia do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais

b) ISDBH 2000: Inquérito de Soroprevalência de Dengue no Município de Belo Horizonte/2000

Aim of the study[1]

To identify potential high-risk intra-urban areas of dengue, using data collected at household level from survey

[1] Siqueira-Junior,JB; Maciel, IJ; Barcellos, C; Souza, WV; Carvalho, MS; Nascimento, NE; Oliveira, RM; Morais-Neto, O; Martelli, CMT. Spatial point analysis based on dengue surveys at household level in central Brazil. *BMC Public Health* 8:361, 2008.

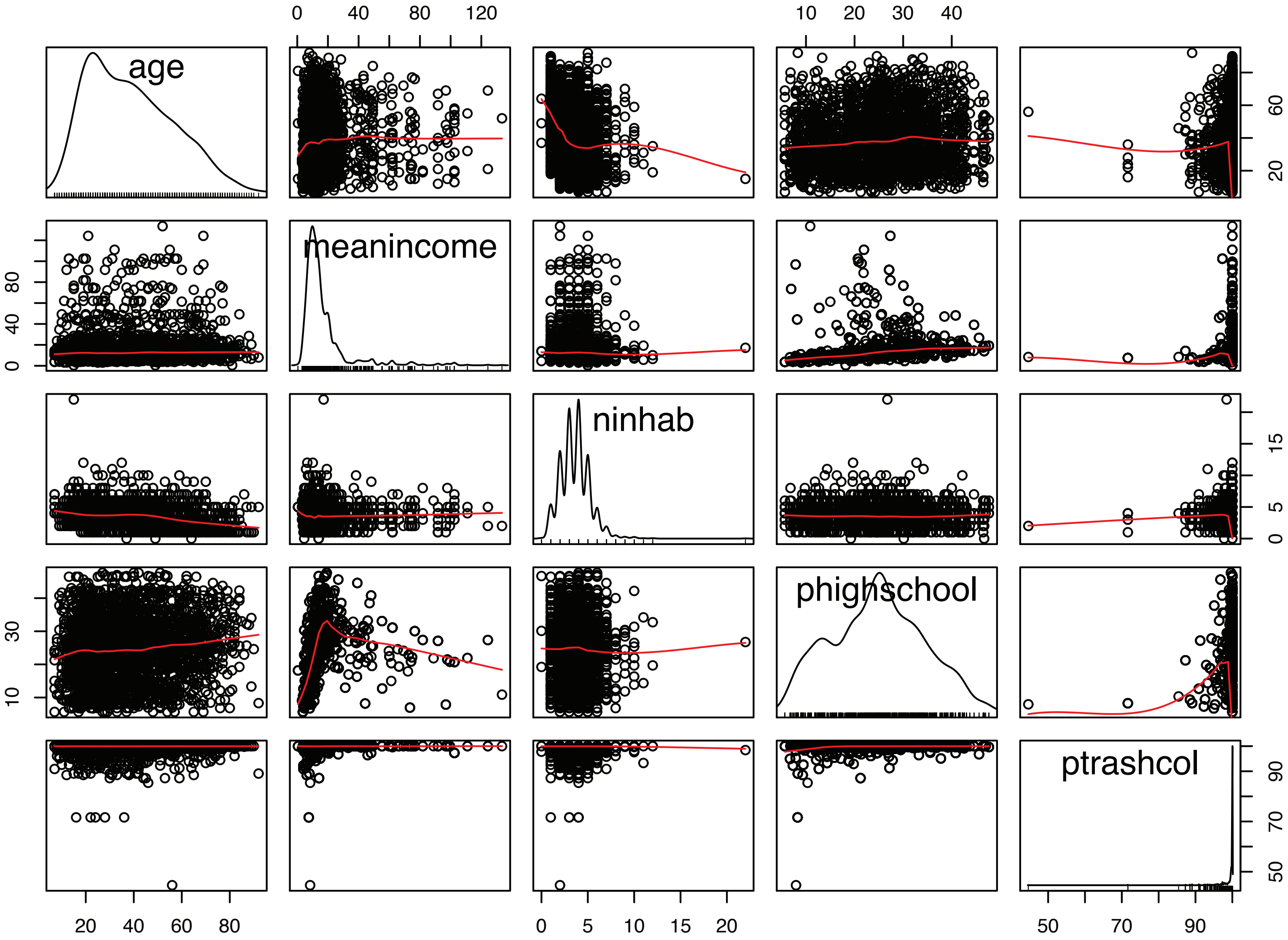
The data

- 2581 participants of the 2002 survey
- Individual data:
  - household coordinates (UTM) – eventually more then one person in the same location
  - positive or negative serology
  - age
  - sex
  - school:
    - Incomplete Basic – less then 8 years at school
    - Basic – complete 8 years
    - High School – 11 years
    - College or university
  - nrooms – number of rooms in the household
  - ninhab – number of people living in the same household

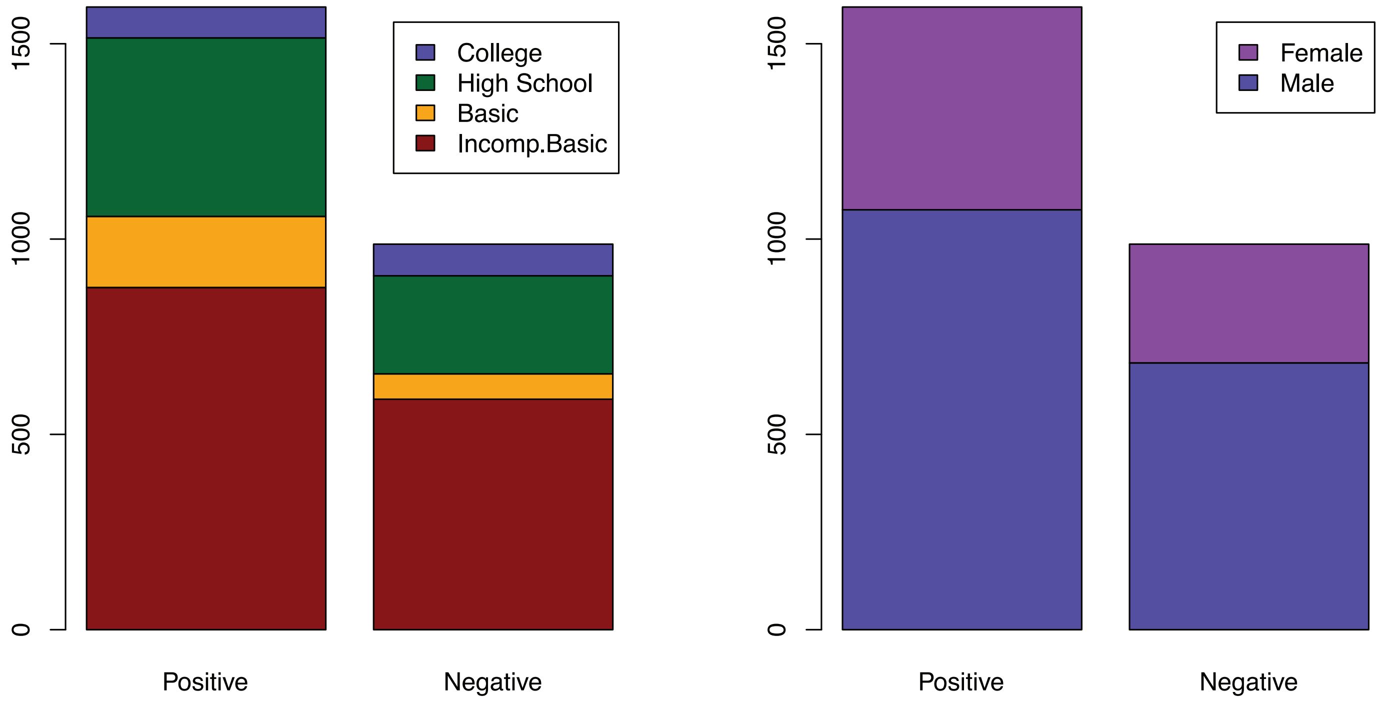
The data

- Census tract level data:
  - pop2000 – population count
  - ptrashcol – % of households with regular trash collection
  - phighschool – % of head of household with complete high school
  - meanincome – average household income in “minimum wages” of census tract

The data


age
meanincome
ninhab
phighschool
ptrashcol
0
40
80
120
10
20
30
40
60
20
80
40
0
15
5
0
30
10
90
70
50
20
40
60
80
0
5
10
15
20
50
70
90

The data


0
500
1000
1500
College
High School
Basic
Incomp.Basic
Positive
Negative
0
500
1000
1500
Female
Male
Positive
Negative