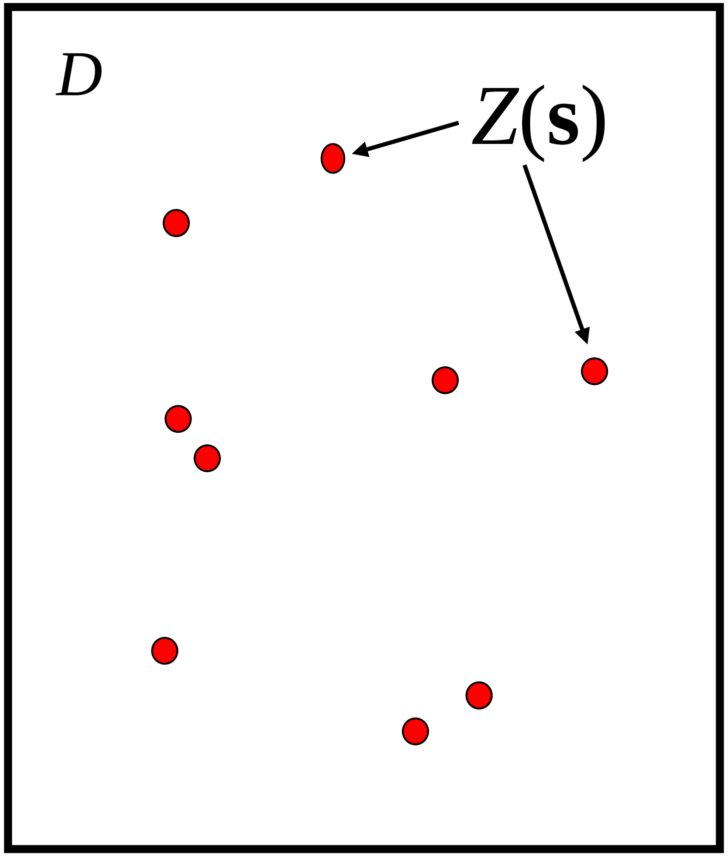
Lattice Data

- The neighborhood matrix
- Exploratory data analysis

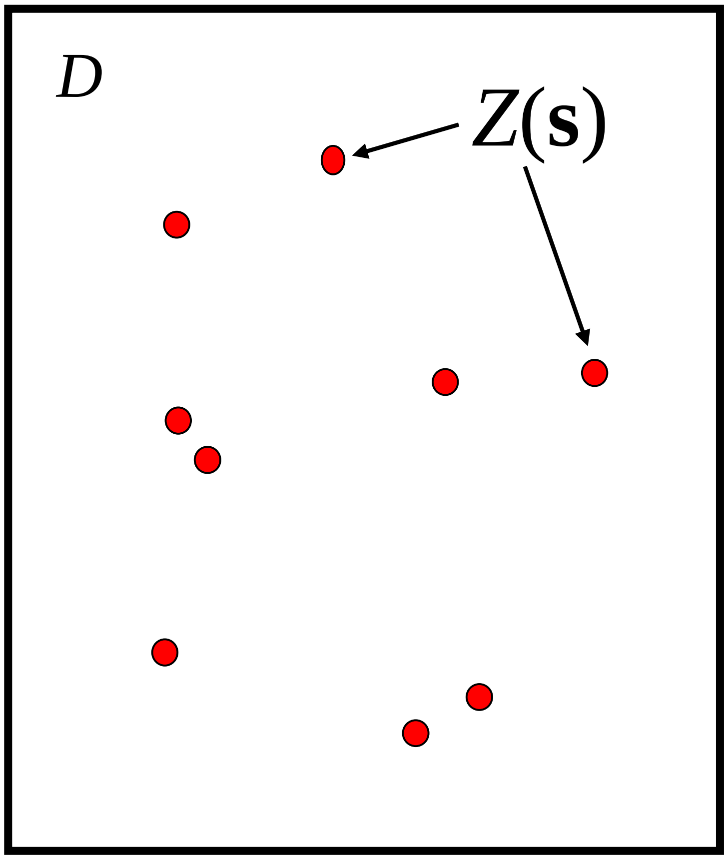
Dra. ROSA ABELLANA

DPT. Fonaments Clínics(UB)

Notation



- *D:* Spatial domain or interest area
- **s:** spatial coordinates
- *Z: study variable*



$\{Z(\mathbf{s}) : \mathbf{s} \in D\}$

Sample n :

i=(1...n)

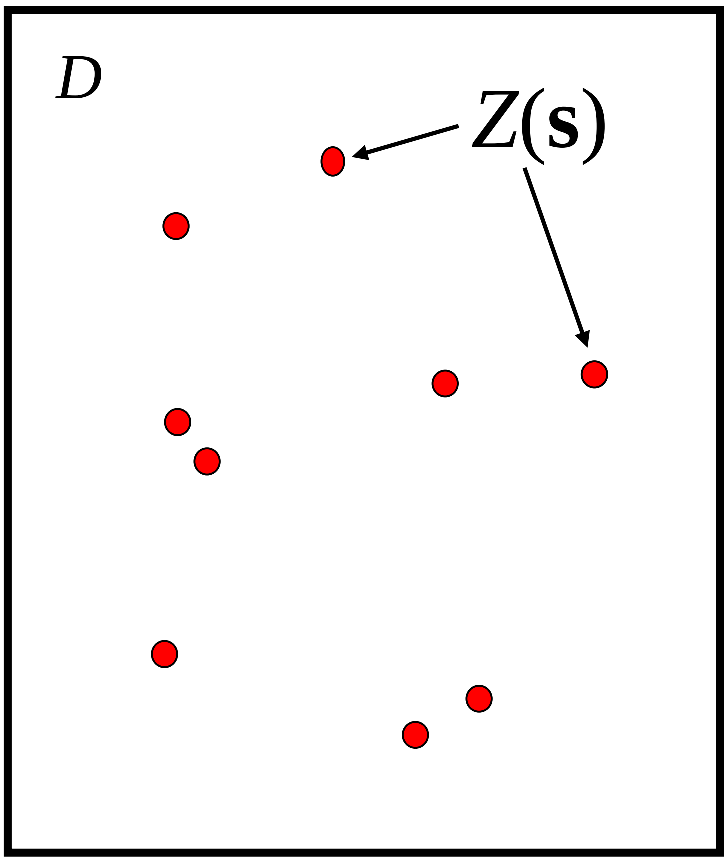
Locations of the events

$s=(s_1,s_2,...s_n)$ $s_i=(x_i,y_i)$

Interest variable

$Z(s_i)$ $z(s_1),....z(s_n)$

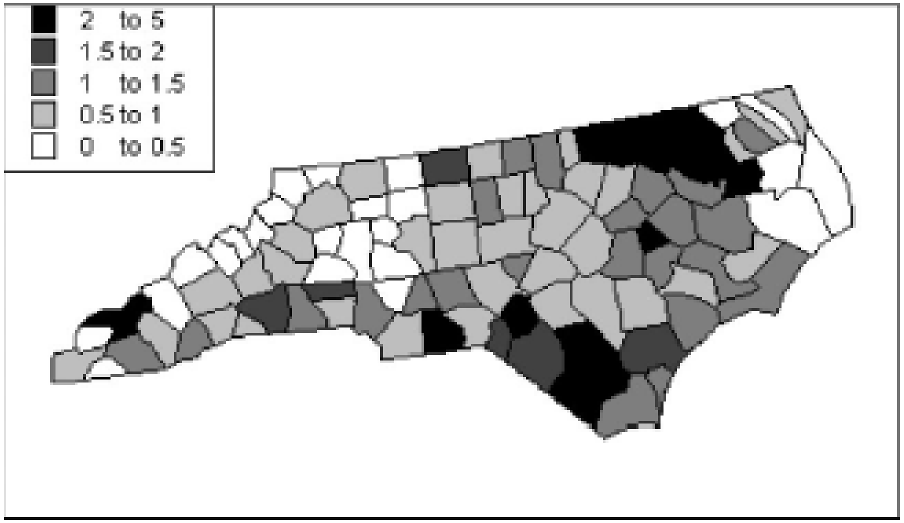
Type of spatial data



$$\{Z(\mathbf{s}) : \mathbf{s} \in D\}$$

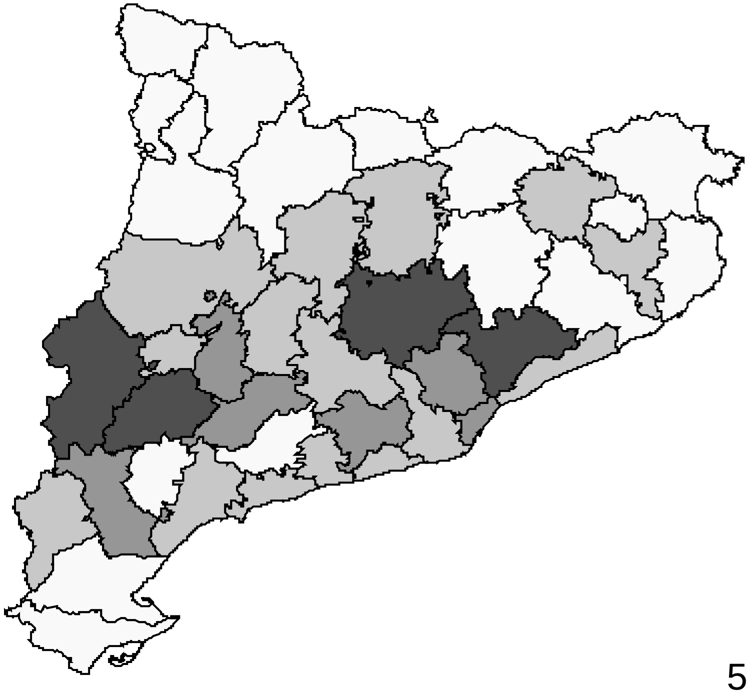
- **Geo statistics**: Z random; D fixed, infinity,
- **Lattice Models**: Z random; D fixed, finite, (ir)regular grilled
- **Point Patterns**: D random, finite
  - **Spatial point pattern**: $Z=1$
  - **Marked point pattern**: Z random

Examples Lattice Data



North Carolina: Sudden infant death standardized mortality

Catalonia: SMR Diabetes type I



Examples (II)

Lattice data

- Number of patients by region in a county.
- Number of fruits for each tree in a field
- Number of accidents for road section.
- Number of fishes by river section.

Neighborhood

Modelling of data at an areal level, it is necessary to define

$\boxed{\text{Adjacency matrix}}$

that characterizes the neighbourhood structure of the data being analysed

Adjacency matrix

1. CHOOSE A NEIGHBORHOOD CRITERION

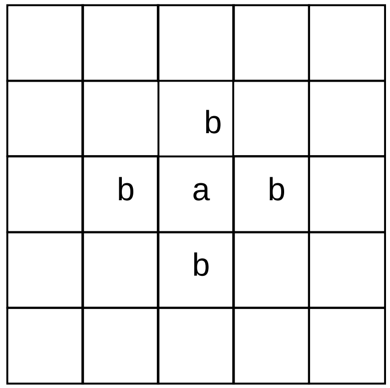
a. Contiguity based neighbors
b. Distance based neighbors k nearest neighbors
c. Distance based neighbors

2. DEFINE SPATIAL WEIGHTS MATRICES

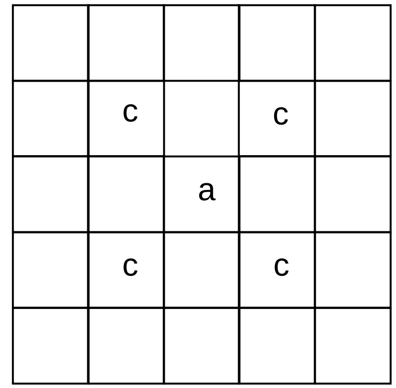
a. Binary weights (0/1)
b. Row-standardized weights
c. According the distances between regions
other criteria

a. Contiguity based neighbors

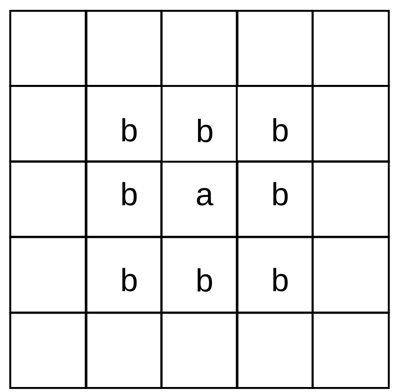
Regular grid



Common margin



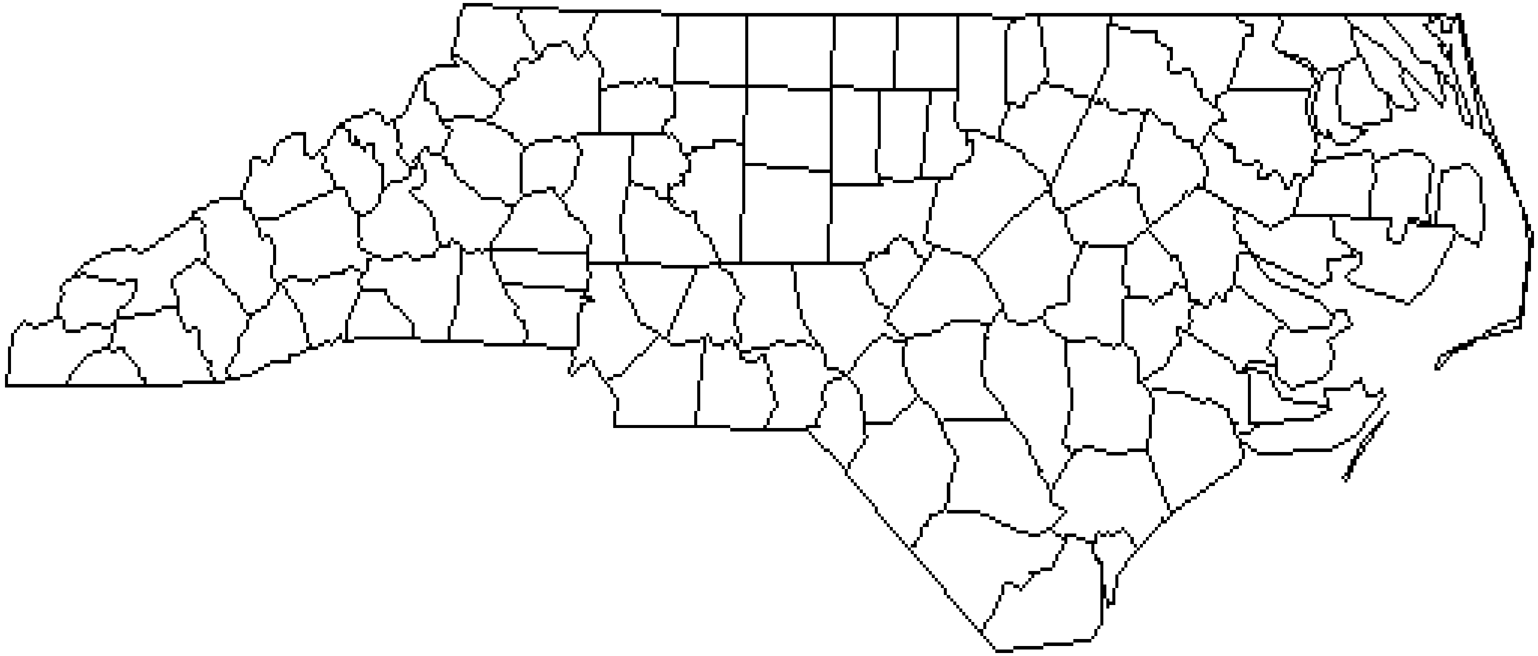
Common vertex



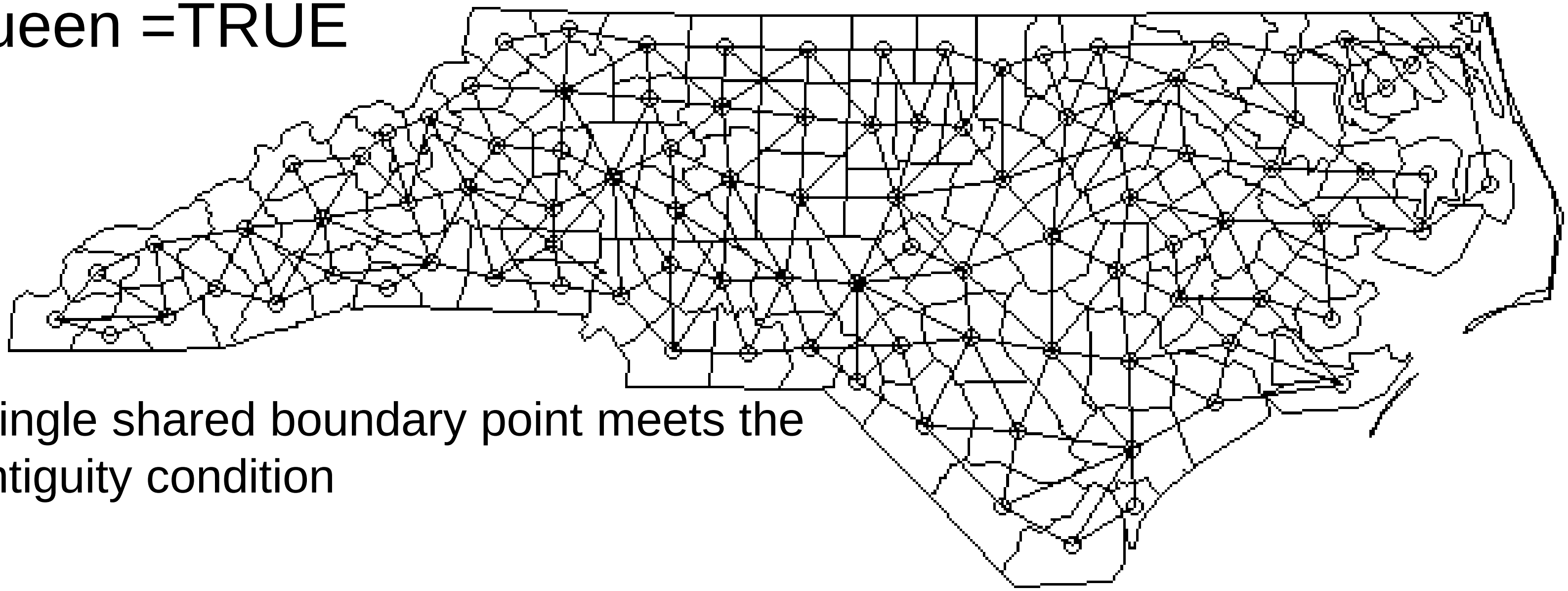
Combination

a. Contiguity based neighbors

Irregular grid

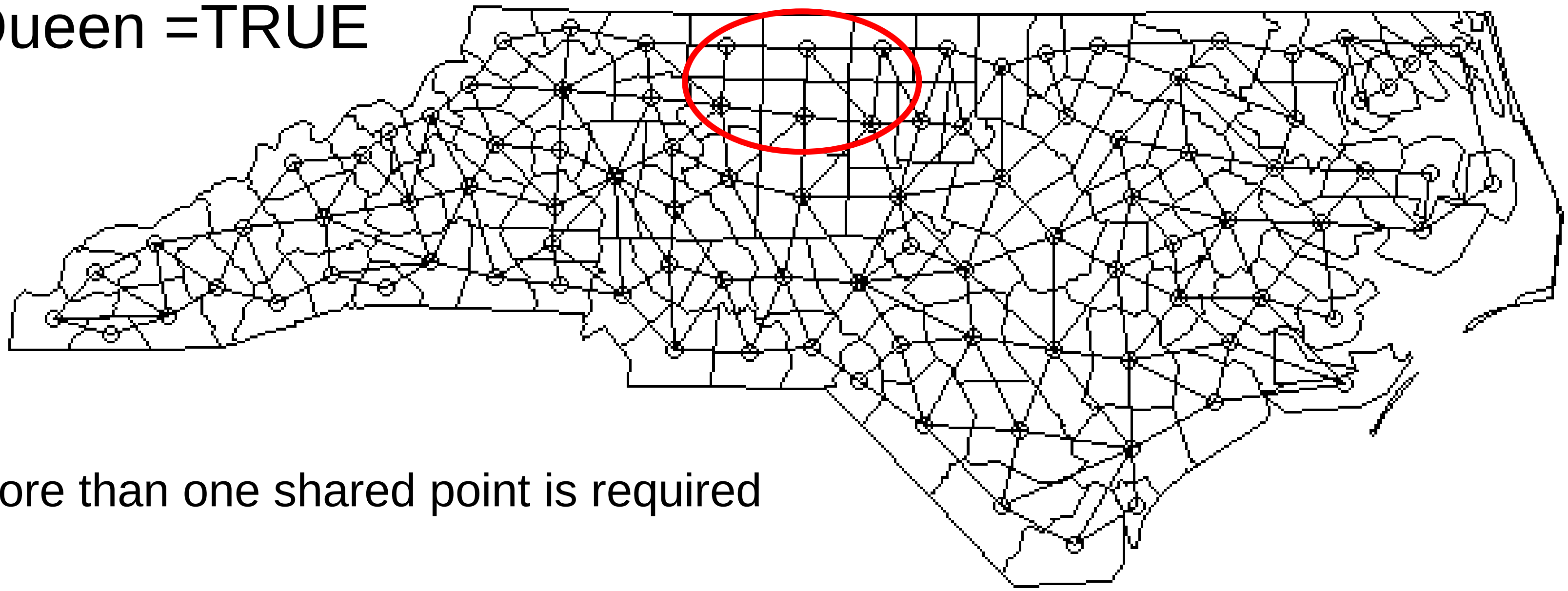


Queen =TRUE



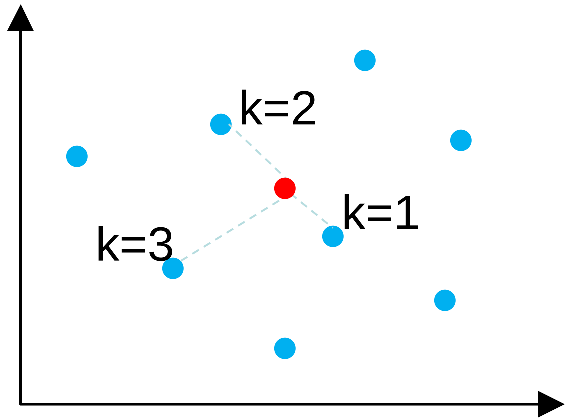
A single shared boundary point meets the contiguity condition

Queen =TRUE

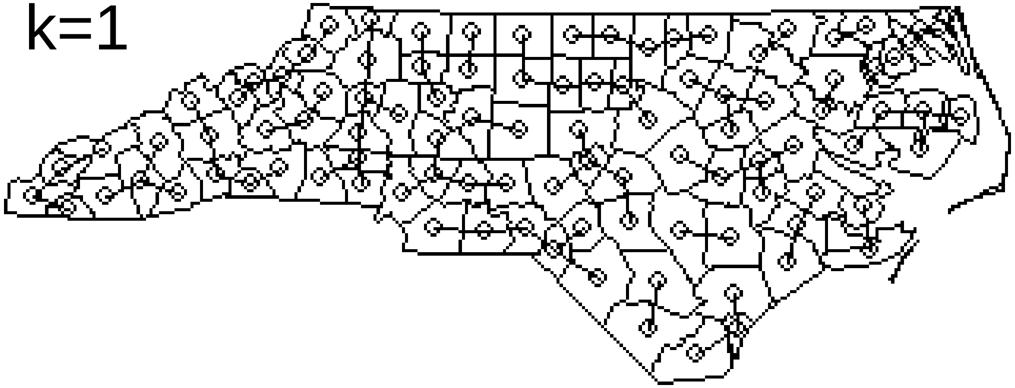


More than one shared point is required

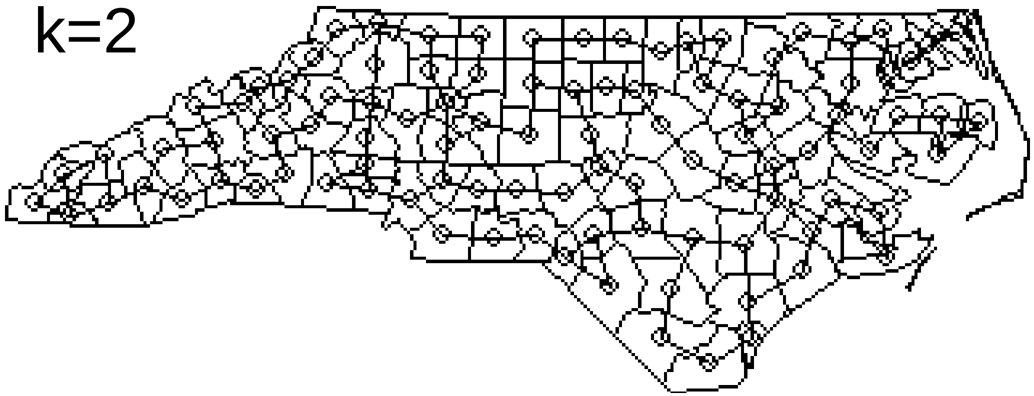
b. Distance based neighbors k nearest neighbors



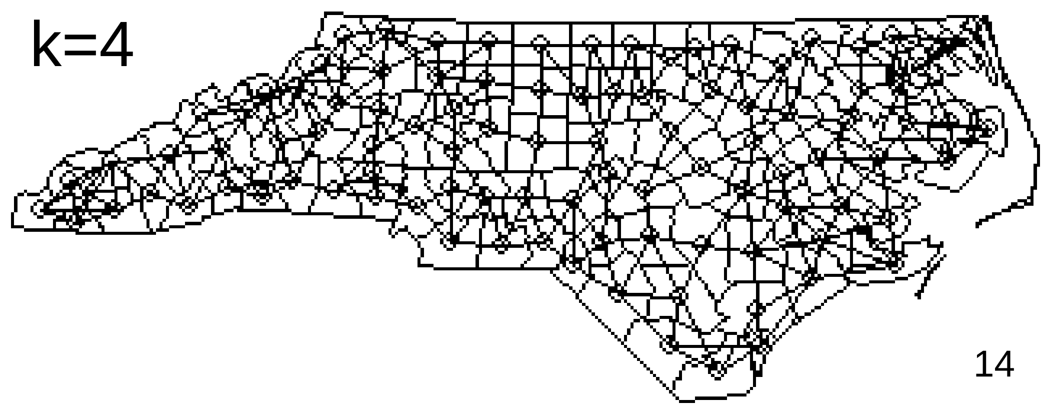
k=1



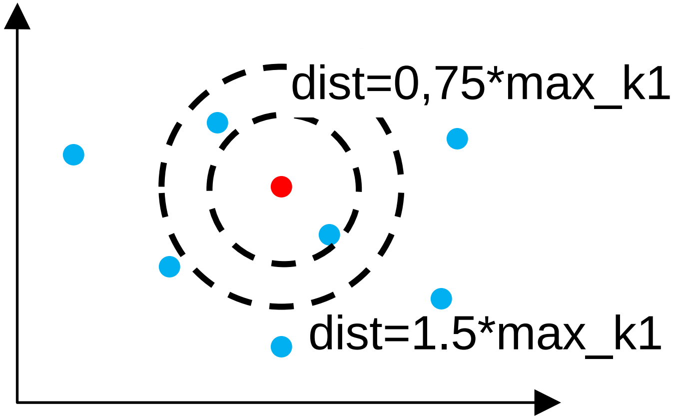
k=2



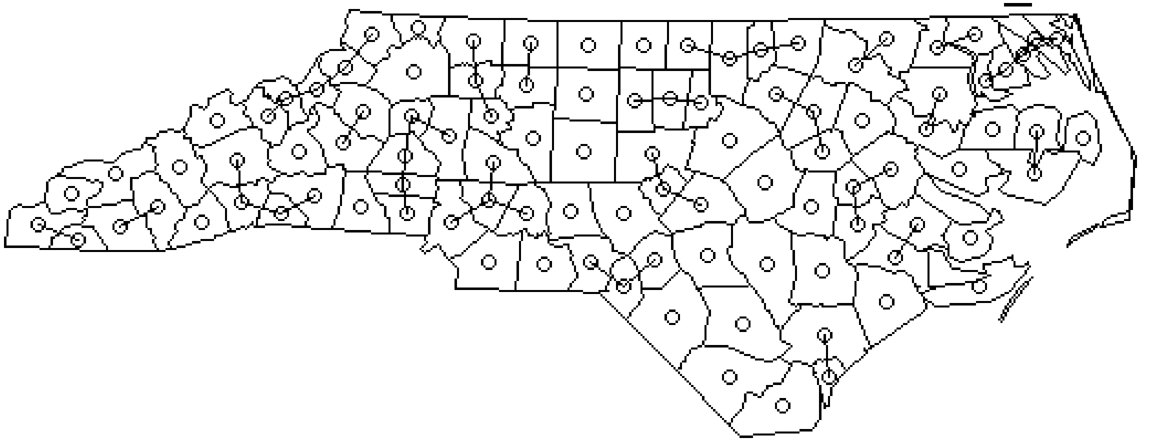
k=4



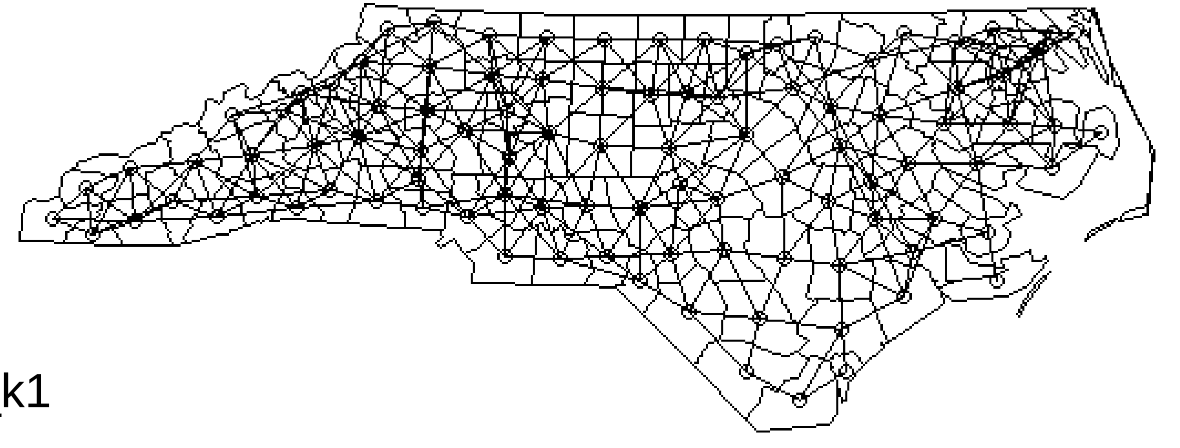
c. Distance based neighbors

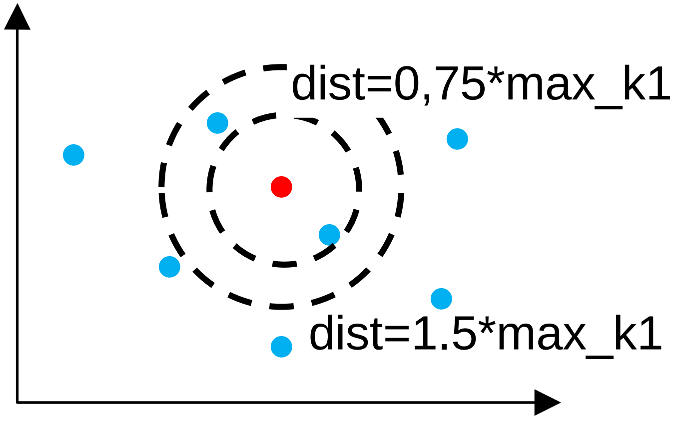


Max_k1=distance maximum between centroids dist=0.75*max_k1



dist=1.5*max_k1





2. DEFINE SPATIAL WEIGHTS MATRICES

a. Binary weights (0/1)

Z_1	Z_2
Z_3	Z_4

$$W = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Properties:

Matrix of 0's and 1's, where:

- 1 indicates that region j^{th} is neighbor of the region i^{th}
- 0 indicates that region j^{th} is not neighbor of the region i^{th}
- $w_{ii}=0$
- $w_{ij}=w_{ji}$ (Symmetry matrix)

Grid of 3x3 = 9 regions

1	2	3
4	5	6
7	8	9

$$W = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

b. Row-standardized weights

Spatial weights style of row standardization $$\frac{w_{ij}}{\sum_{j \in J} w_{ij}}$$

$$W = \begin{bmatrix} 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ \ldots\ldots\ldots\ldots\ldots & & & & & & & & \end{bmatrix}$$

c. According the distances between regions or other criteria

Degree of relationship between two regions can depend on:

- distance among their geographic centroids
- length of the share boundaries
- frequency of public transport services

Definitions

Anselin (1993) $\longrightarrow$ $$w_{ij} = d_{ij}^{-2}$$

Where d_{ij} Euclidean distance between region i*th* and j*th* i

Cliff i Odr (1973) $\longrightarrow$ $$w_{ij} = d_{ij}^{-a}\left[\beta_{i(j)}\right]^{b}$$

where β proportion of the length of the share boundaries between region i*th* and j*th* (a,b>0)

Bodson and Peeters (1975), proposed accessibility criteria [0,1], that combines by the logistic function the influence of several communication channels between regions.

$$w_{ij} = \sum_{n=1}^{N} K_n \left\{ \frac{a}{1 + b \cdot \exp\left(-c_j d_{ij}\right)} \right\}$$

K_n: relative importance of the channel of communication (n) ,
d_{ij} : la distance between region i th and j th
a, c_j and b : parameters to be estimate

A few papers justify the selection of weighted matrix. The most used it is the binary definition (share boundaries)

Eranest Arul (2007). Evaluated the influence for different weighted matrix. Data number of births with problems of malformations in New South Wales (Austràlia)

Exploratory data analysis

Exploratory analyses of space data consists in a group of technical that allows to describe the spatial distribution, identify spatial outliers and discover spatial clusters

Definition Spatial autocorrelation

Types of tests

- Global
- Local

GLOBAL TEST

1. Moran's Test (Moran 1948)

$$I = \frac{N\sum_{i}\sum_{j} w_{ij}\left(z_i - \bar{z}\right)\left(z_j - \bar{z}\right)}{\left(\sum_{i \neq j} w_{ij}\right)\sum_{i \neq j}\left(z_i - \bar{z}\right)^2}$$

where w_{ij} are the spatial weights between region i*th* and *jth*

Interpretation

$$I \begin{cases} < E(I)\ \textit{negative spatial correlation} \\ = E(I)\ \textit{spatial independence} \\ > E(I)\ \textit{positive spatial correlation} \end{cases}$$

Null hypothesis , Z_i , are i.i.d **I** is asymptotically normal with

$$E(I) = {-1}/{(n-1)}$$

$$\mathrm{Var}(I) = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2}$$

on $$S_0 = \sum_{i \neq j} w_{ij} \quad S_1 = \frac{1}{2}\sum_{i \neq j}\left(w_{ij} + w_{ji}\right)^2 \quad S_2 = \sum_k \left(\sum_j w_{kj} + \sum_i w_{ik}\right)^2$$

Interpretation $$z_I = {I - E(I)}/{SD(I)}$$

$$z_I \begin{cases} < 0 \; negative\; spatial\; correlation \\ = 0 \; spatial\; independence \\ > 0 \; positive\; spatial\; correlation \end{cases}$$

Alternatives of assumption of normality

The variance of I calculated under the assumption of normality or assumption of randomisation.

Randomisation introduce a correction term bases on the kurtosis of the variable of interest

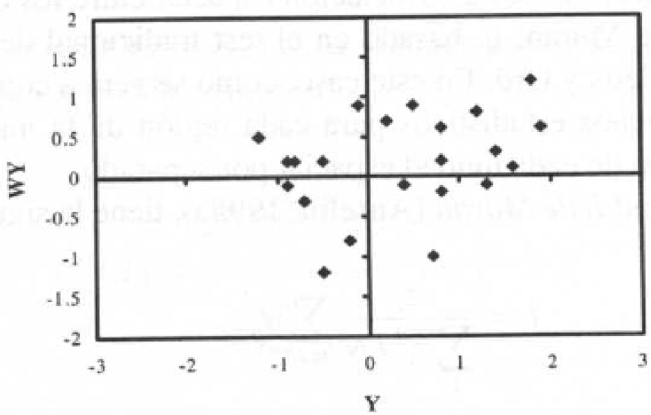
Test Moran- Boostrap permutation based test

The values of the variable of interest are randomly assigned to entities

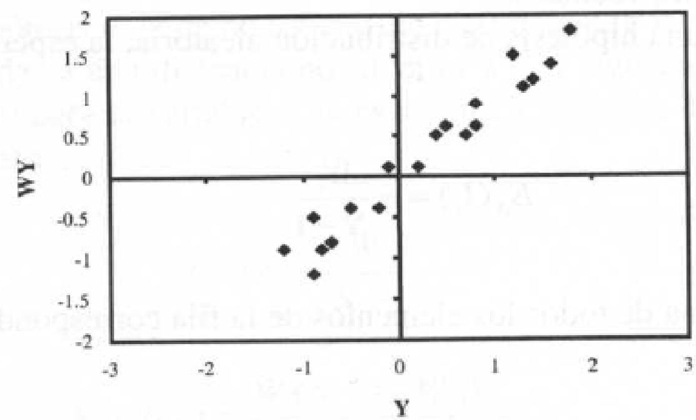
Establish position occupies the observed I Moran test in relation to the values of I Moran simulated under the hypotheses of spatial independence

MORAN'S SCATTERPLOT

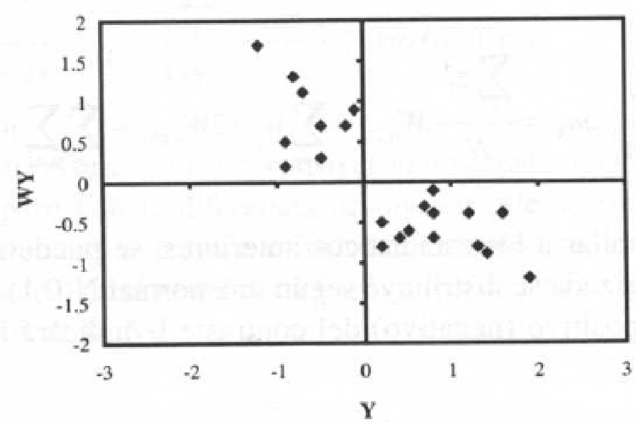
Variable of interest on the x-axis and the spatially weighted sum of values of neighbors- the spatially lagged values- on the y-axes.



Independent



Positive spatial dependence



Negative spatial dependence

2. Geary's Test (Geary 1954)

$$c = \frac{(n-1)\sum_{i=1}^{n}\sum_{j}^{n} w_{ij}\left(z_i - z_j\right)^2}{\left(\sum_{i \neq j} w_{ij}\right)\sum_{i}\left(z_i - \bar{z}\right)^2}$$

where w_{ij} are the spatial weights between region i*th* and j*th*

Normal assumptions (i.i.d)

$$\mathrm{E}(\mathrm{c}) = 1$$

$$\mathrm{Var}(\mathrm{c}) = \frac{\left(2\mathrm{S}_1 + \mathrm{S}_2\right)\left(\mathrm{n}-1\right)\mathrm{S}_2 - 4\mathrm{S}_0}{2(\mathrm{n}+1)\mathrm{S}_0^2}$$

Interpretation $z_c = {c - E(c)}/{SD(c)}$

$$z_c \begin{cases} <1 \; positive \;\; spatial \; correlation \\ =1 \; spatial \; independence \\ >1 \; negative \; spatial \; correlation \end{cases}$$

Same interpretation than Moran’s test

$$c' = 1 - \frac{\sum_{i=1}^{N}\sum_{j}^{N} w_{ij}\left(z_i - z_j\right)^2}{\left(\sum_{i \neq j} w_{ij}\right)\sum_{i}\left(z_i - \bar{z}\right)^2}$$

$$Z_c = \frac{E(c) - c}{SD(c)}$$

Moran's *I* is a measure of global spatial autocorrelation, while Geary's *C* is more sensitive to local spatial autocorrelation.

INDICADORS LOCALS

Local Moran I_i and G_i.

Global tests for spatial autocorrelation are calculated from the local relationship between the values observed at a spatial entity and its neighbors, for the neighbor definition chosen.

Break global measures down into their components and by extension construct localizes tests to detect clusters and hotspots.

Local de Moran I_i

$$I_i = \frac{\left(z_i - \bar{z}\right)\sum_{j}^{n} w_{ij}\left(z_j - \bar{z}\right)}{\frac{\sum_{i=1}\left(z_i - \bar{z}\right)^2}{n}}$$

I_i (positive) detect clusters: observations with very similar neighbors

I_i (negative) detect hotspots: observations with very different neighbors.

Maptools Package (for use with R - WIN, MacOS, Linux) - Freeware
http://cran.r-project.org/web/packages/maptools/index.html

Set of tools for manipulating and reading geographic data, in particular ESRI shapefiles; C code used from shapelib. It includes binary access to GSHHS shoreline files.

The package also provides interface wrappers for exchanging spatial objects with packages such as PBSmapping, spatstat, maps, RArcInfo, Stata tmap, WinBUGS, Mondrian, and others.

sf

.Support for simple features, a standardized way to encode spatial vector data. Binds to 'GDAL' for reading and writing data, to 'GEOS' for geometrical operations, and to 'PROJ' for projection conversions and datum transformations.

spdep: Spatial Dependence (for use with R - WIN, MacOS, Linux) - Freeware
http://cran.r-project.org/src/contrib/Descriptions/spdep.html or
http://spatial.nhh.no/R/spdep/

A collection of functions that run in the R language to:

–Create spatial weights matrix objects from polygon contiguities, from point patterns by distance and tesselations, for summarising these objects, and for permitting their use in spatial data analysis.

–Tests for spatial autocorrelation, including global Moran’s I, Geary’s c, Hubert/Mantel general cross product statistic, Empirical Bayes estimates and Assun¸c˜ao/Reis Index, Getis/Ord G and multicoloured join count statistics, local Moran’s I and Getis/Ord G, saddlepoint approximations for global and local Moran’s I.

– Estimate spatial simultaneous autoregressive (SAR) models.

Bibliografia:

Banerjee S Carlin BP, Gelfrand A.E. (2004) Hierarchical Modelling and Analysis for Spatial Data. Chapman & Hall /CRC

Bibliografia addicional:

Cliff A and Odr J (1973). Spatial autocorrelaction, London:Piom