Università degli Studi di Napoli Federico II

Geostatistics with R

Dr. Francesco Carotenuto, Ph.D.

DISTAR

(Dipartimento di Scienze della Terra, dell'Ambiente e delle Risorse)

Course programme

First part:

- Vector and raster data
- Simulating and managing vector data
- Simulating and managing raster data
- Geographic projections

Second part:

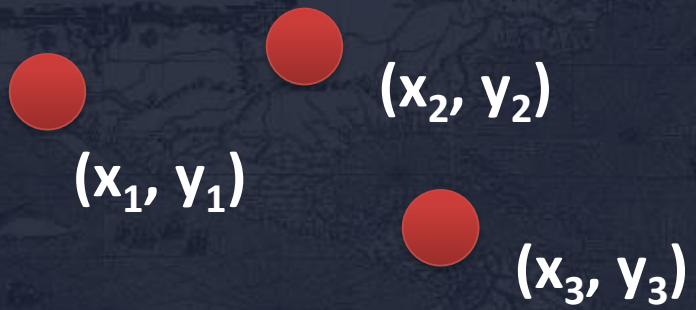
- Spatial Autocorrelation
- Semivariogram and fitting of..

Geostatistics

Statistics applied to georeferenced data: Long, Lat and other variables….

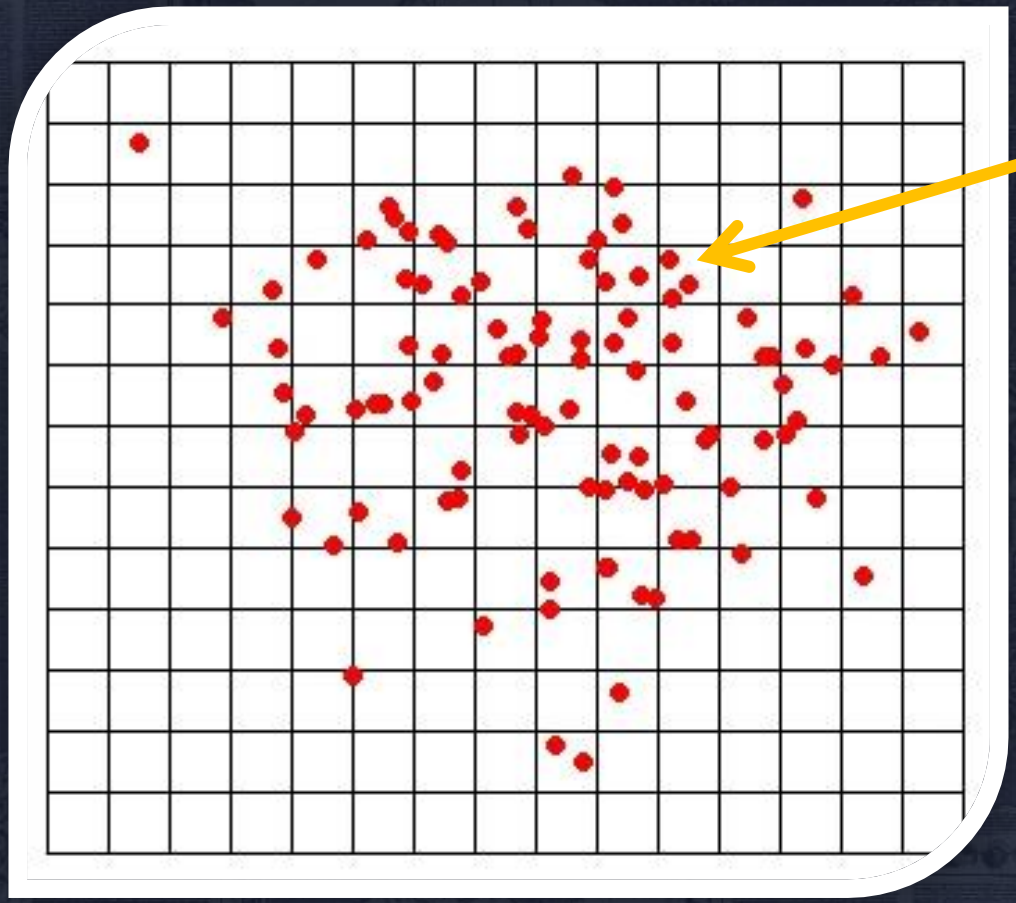
Vector data

Vector data: points, lines and polygons



Streets, buildings, weather stations, sampling locations on the field..

A simple spatial analysis: grid-based statistics



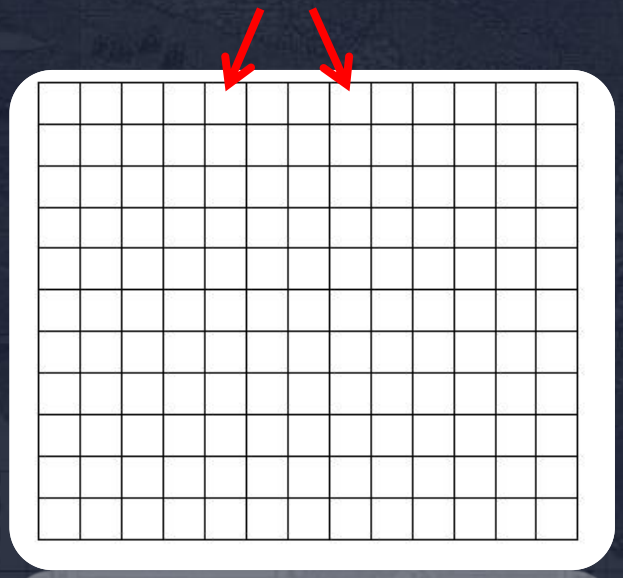
For each cell
Mean;
Min;
Max;
Sd.....

Raster data

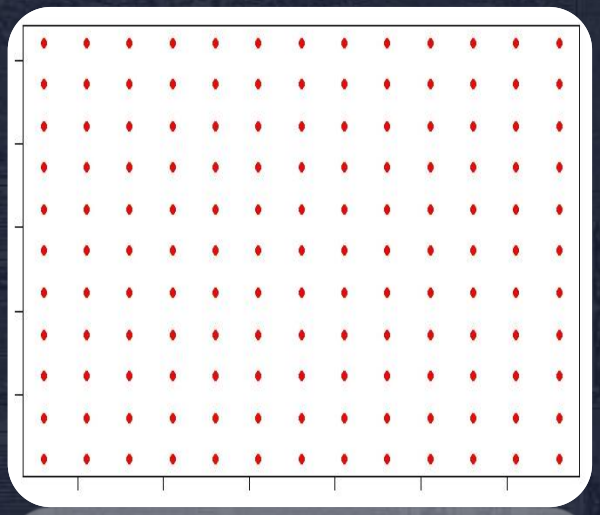
Raster data: grids whose cells include geographical coordinates plus an additional z variable

Equally spaced cells

The same edge lenght in one dimension



Equivalent to regularly distributed points

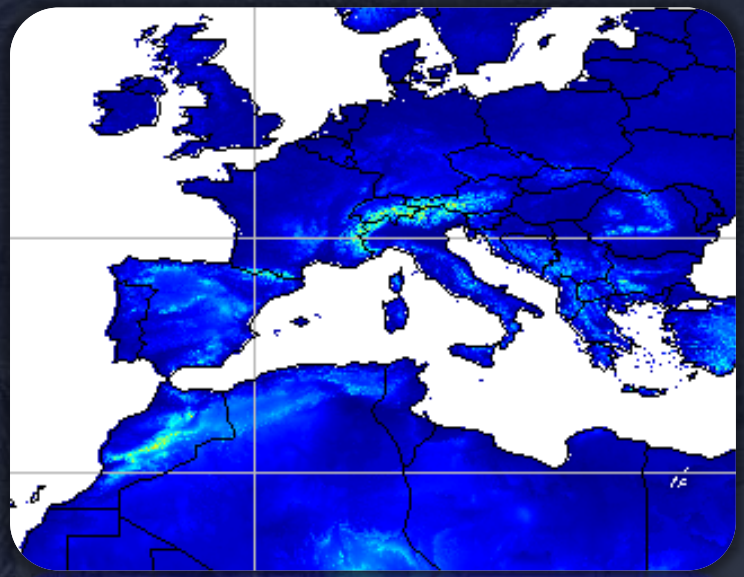


Raster data

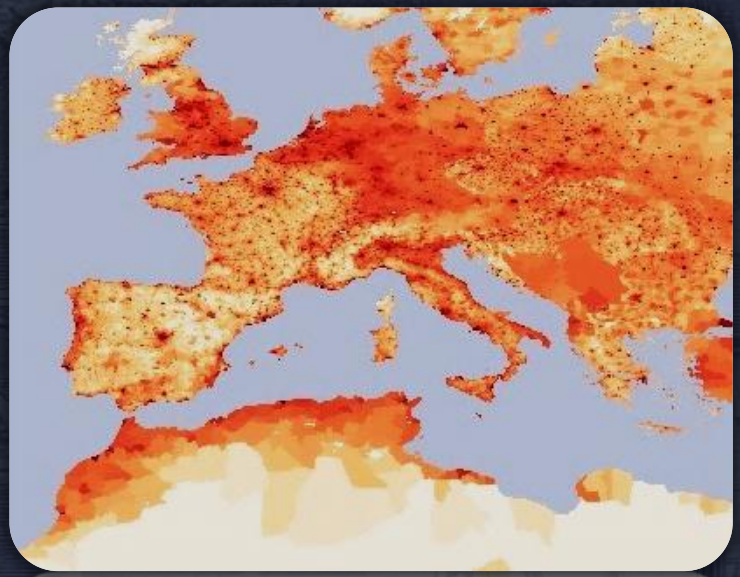
Raster data: grids whose cells include geographical coordinates plus an additional z variable

Different kinds of z values!

Altitude



Population density map



The spatial autocorrelation

The first law of geography says: "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). This phenomenon is called spatial autocorrelation by statisticians.

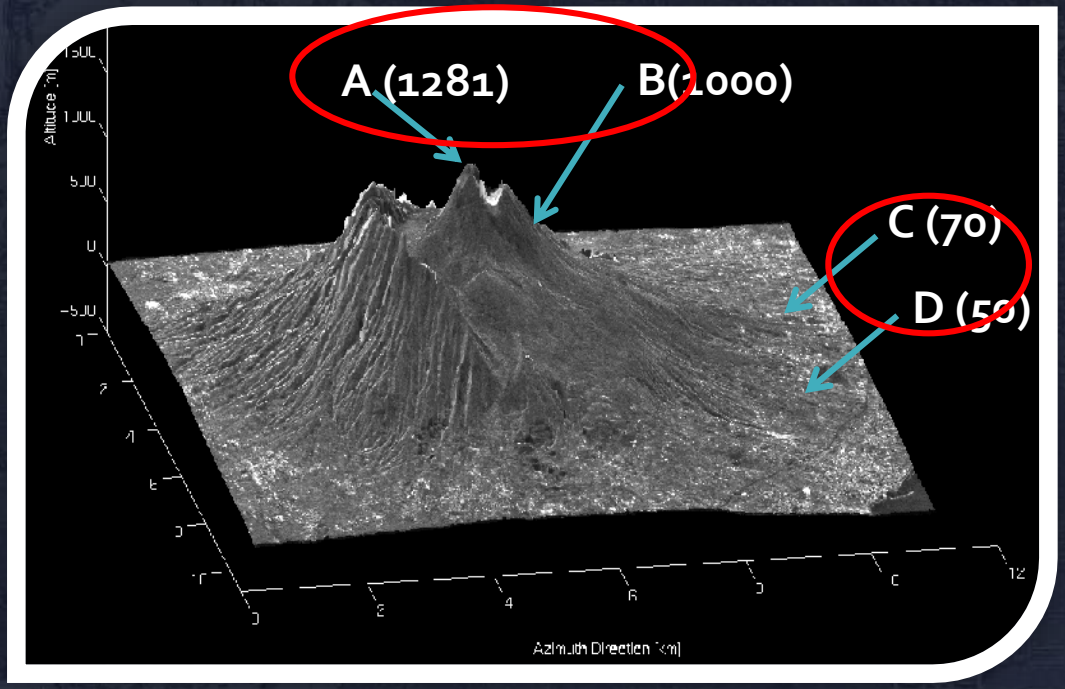
Spatial autocorrelation is a feature characterizing natural phenomena but it could be exacerbated by the way we sampled variables over the landscape (artificial increase of the spatial autocorrelation)

The spatial autocorrelation

The first law of geography says: "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). This phenomenon is called spatial autocorrelation by statisticians.

This happens in nature!

The profile shown by my data



Mean A,B
1140.5 m

Mean A,B
60 m

We have to take into account this phenomenon when sampling data on the landscape
In any statistical analysis the values of A,B,C and D form two different and separate groups!

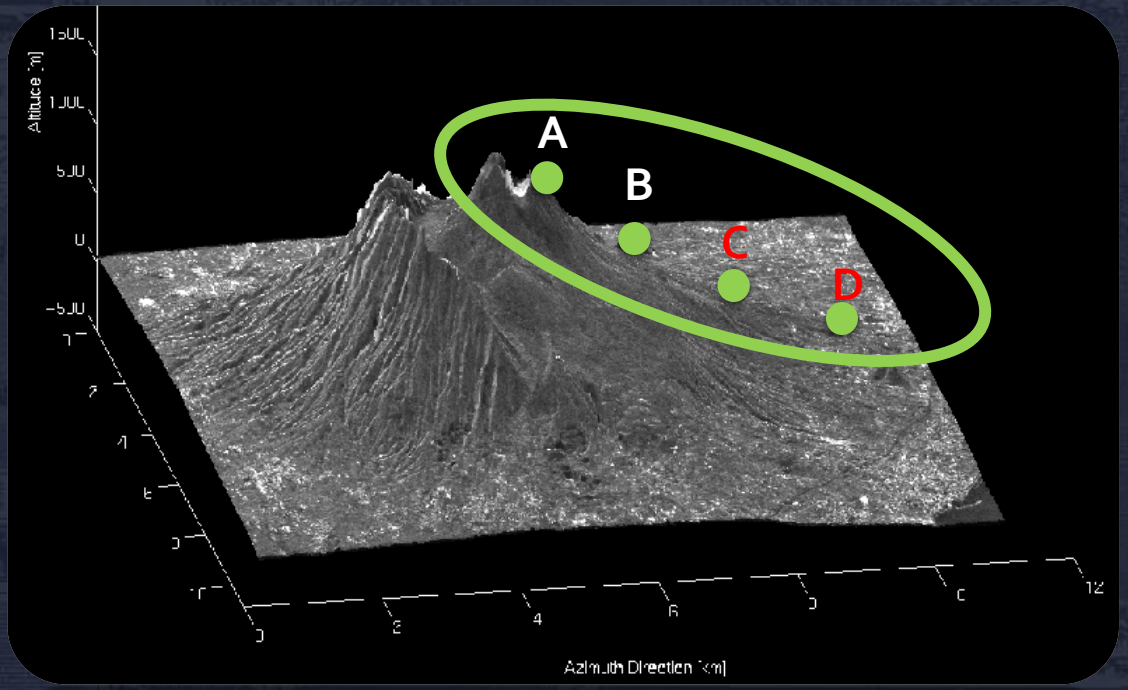
Our sample could highlight differences due to the non random distribution of points. Our sampled data describe a landscape with only a tip and a valley, without intermediate values: the filed is a staircase!

The spatial autocorrelation

This is a better sample to represent what really happens

Differences still exist but are less extreme!

A better profile!



Mean A,B
850 ,m

Mean C,D
350 ,m

The spatial autocorrelation

When we have a very clustered sample…..

An observation in close proximity to another one does not substantially increase the quality information in the data because it is similar to the one already measured.

This kind data collection produces pseudo-replicates: i.e. a lot of data but a poor representation of the landscape!

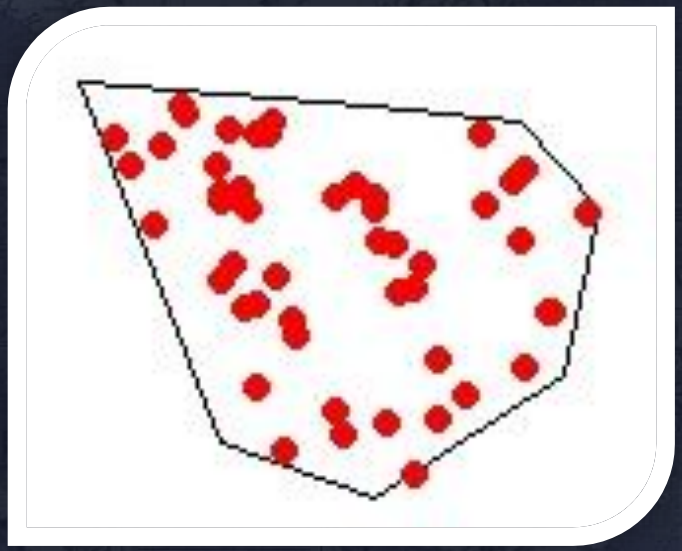
Such measurements only artificially increase sample size without any useful contribution in terms of information. Autocorrelation (either natural or artificial) will lead to an overestimation of the precision of the results (e.g. the confidence intervals will be too small, the detection of groups that don't exist in nature…).

Highlighting differences even when they don't exist!

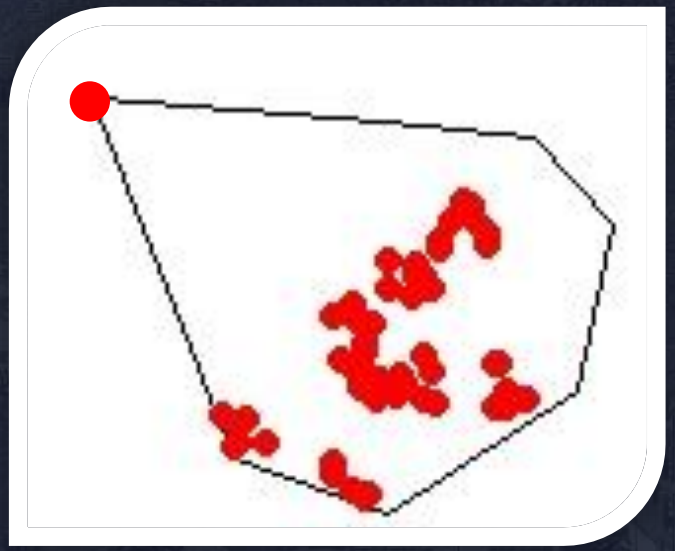
Spatial Autocorrelation bias in regression models

In many cases we have a little control on the experimental design as the spatial distribution of the sample is determined by natural phenomena (landslides, earthquakes…)

The ideal condition!



The real condition!



Spatial bias

Non-Independence of data!

The spatial autocorrelation

If we can to take into account the spatial relatedness of data

Independent (random) data with no or less bias in the results!

If we can have control of sampling on the landscape we MUST prefer randomly distributed sampling localities!

By this way we can't still take into account natural spatial autocorrelation but our sample will be a better representation of the whole landscape

We still have to take into account spatial bias in our analyses!

Spatial Autocorrelation

One of the most important measure of spatial autocorrelation is the Moran's Index (Moran, 1950)

A measure of how values are similar (or different) at some very near locations

Distance between two neighbouring localities

Product of deviations of two neighbouring localities from the mean value

The number of localities

Moran's Index

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

Sum of distances of neighbouring localities

i is a location;
j is its nearest neighbour

Values significantly >0: Positive Spatial Autocorrelation: observations in close proximity are more similar than distant ones

Values significantly <0: Negative Spatial Autocorrelation: observations in close proximity are more different than distant ones

The structure of data when there is no spatial autocorrelaion

Concept of regression

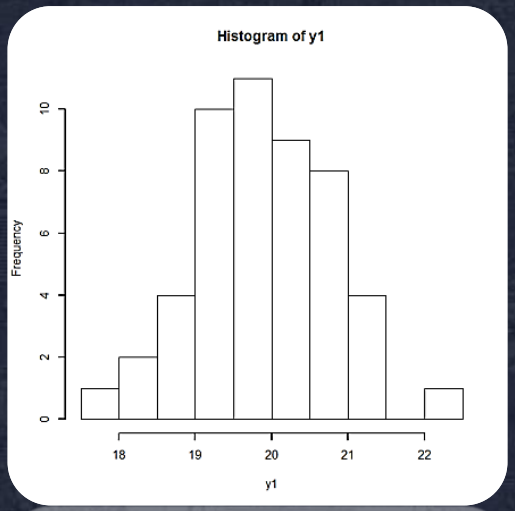
Minimun Least Squares

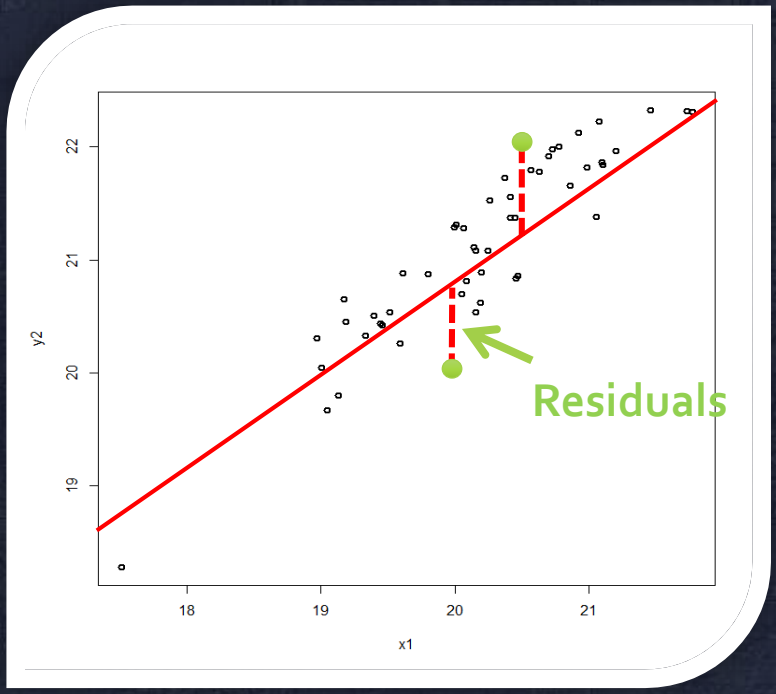
$$y = b * x + a + e$$

b = slope

a = intercept

e = errors or model's residuals

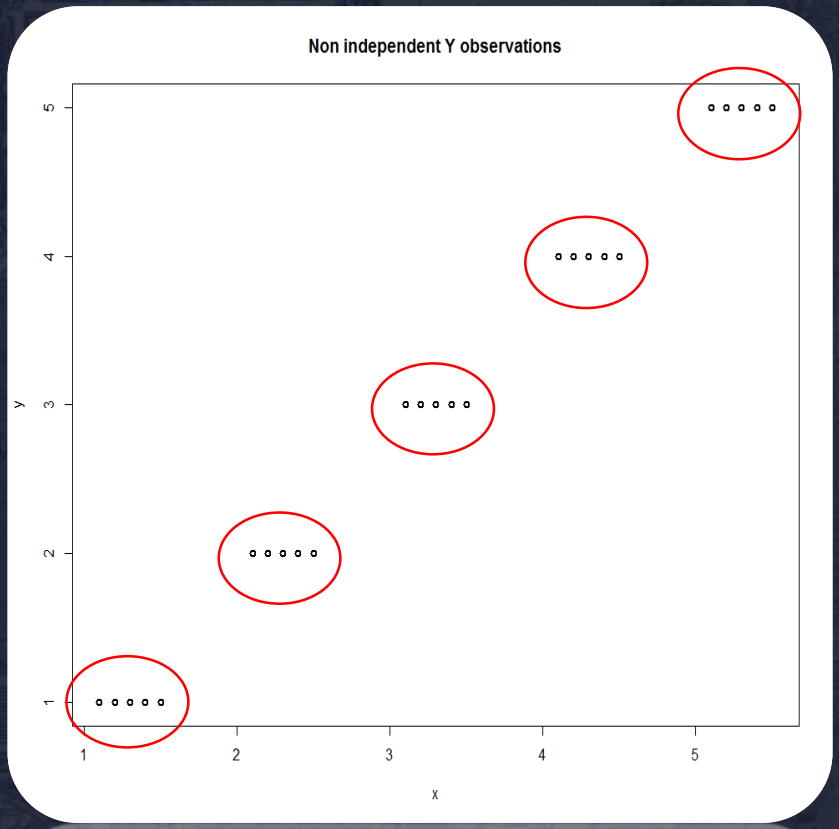




To find the best model describing the data, we need errors are normally distributed

The model is a good representation of the sampled data

The structure of data when there is significantly spatial autocorrelation



Y data show some relatedness and seem clustered

No Gaussian distribution of the residuals

Inaccurate model estimations!!!

Spatial structure can explain The non Gaussian distribution of residuals!

How to take into account spatial autocorrelation?

To find the best model describing the data, we have to make model's errors normal

We first compute Moran's Index of the residuals

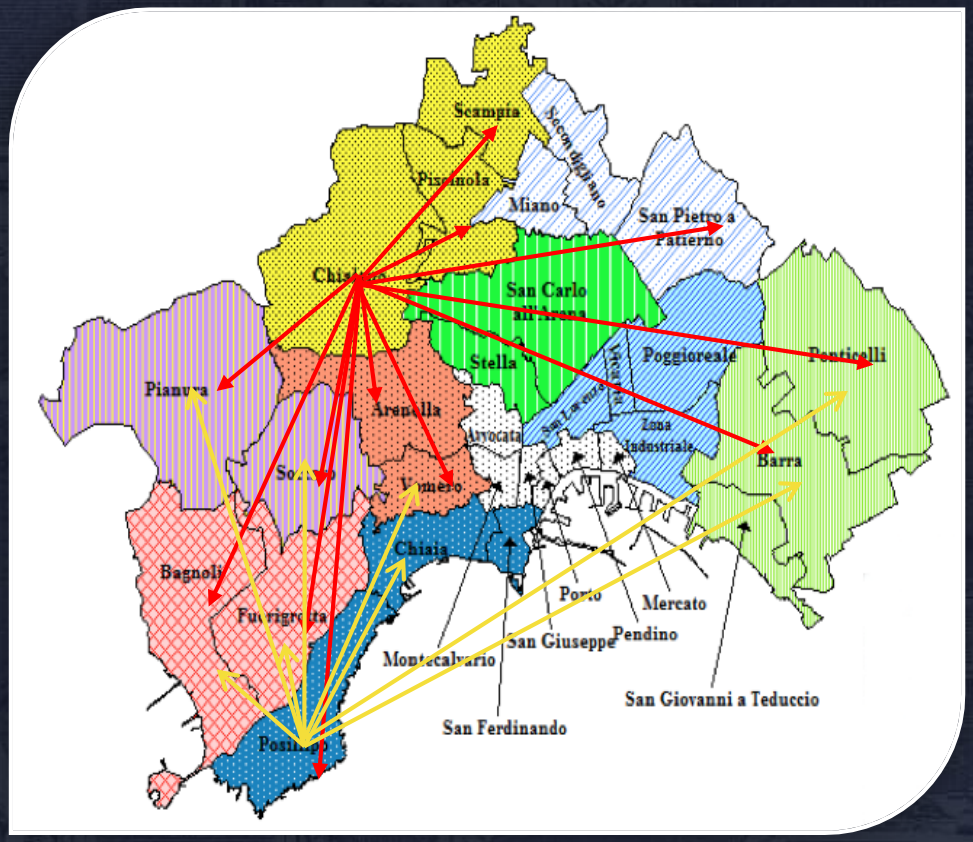
If we detect spatial autocorrelation, may be spatial distribution of residuals have determined the non-normality of residuals, we have to find a model describing the spatial structure of residuals

We have to find a model describing the spatial structure of residuals

Then, we can include this model in the regression equation to make residuals normally distributed

The spatial autocorrelation

Comparing pupils test's scores in Naples' districts schools and detecting the social factors



	School
Scampia	A,B,C
Chiaiano	D,E,F
Stella	G,H,I
Pianura	J,K,L
Vomero	M,N,O
Chiaia	P,Q,R
Posillipo	R,S,T

The test could tell us that richer pupils are smarter than poorer ones. But this could be an artefact of spatial autocorrelation!!

Accounting for spatial autocorrelation in regression: Mixed effect models

$$Y_{ij} = \mu + \beta_1 \mathrm{Sex}_{ij} + \beta_2 \mathrm{Race}_{ij} + \beta_3 \mathrm{ParentsEduc}_{ij} + U_i + W_{ij},$$

Fixed effects (predictors) and their slopes

Random effects
The bias due to the spatial structure of the data

Y_{ij} = score of the *j*th pupil at the *i*th school

μ is the average test score for the entire population

U_i = the cluster-specific random effect: it measures the difference between the average score at school i and the average score in the entire country

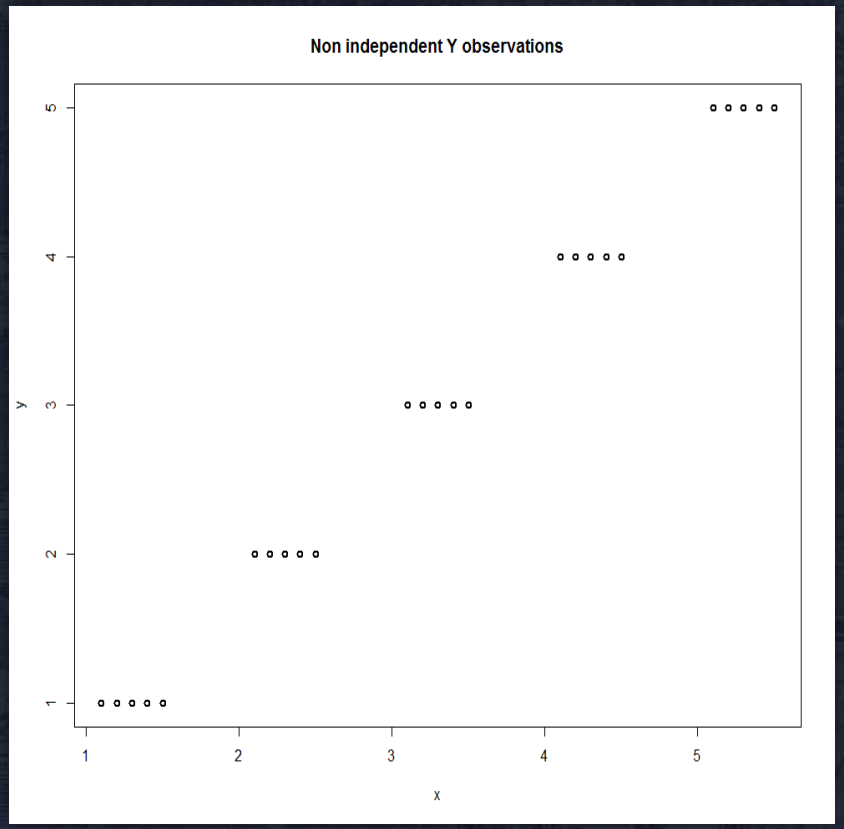
W_{ij} is the individual-specific random effect, i.e., it's the deviation of the *j*-th pupil's score from the average for the *i*-th school

Accounting for spatial autocorrelation in regression models

Generalized Least Squares is a special case of Mixed model that allows including the spatial structure of model's residuals (the random effect)

$$y = b * x + a + \varepsilon$$

ε = error term (model's residuals)

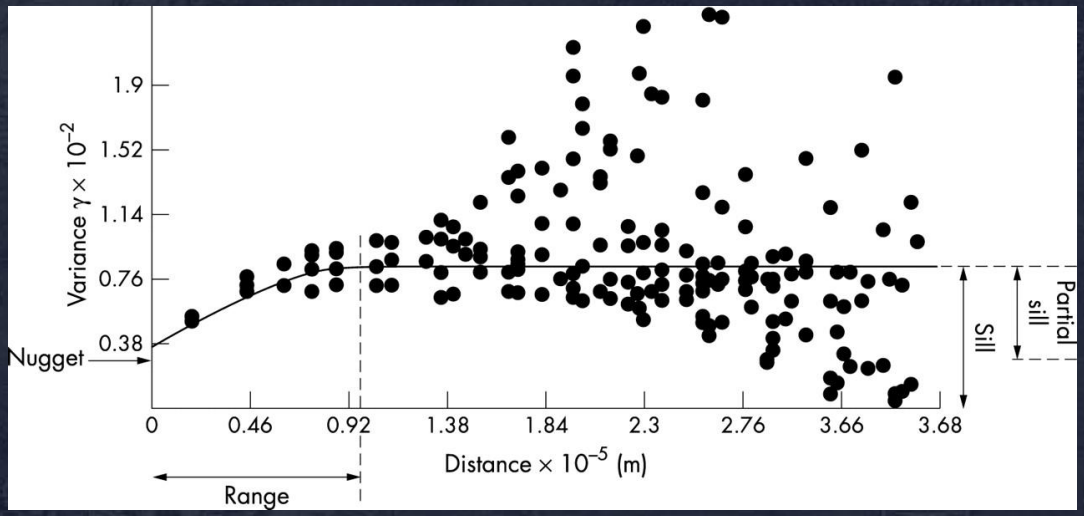


Average parents' education degree for each school

How to model the spatial structure of the residuals?

We need to find a model that can describe the spatial structure of the model' residuals.

Semivariogram: describes the variation of model's residuals by increasing the distance between schools

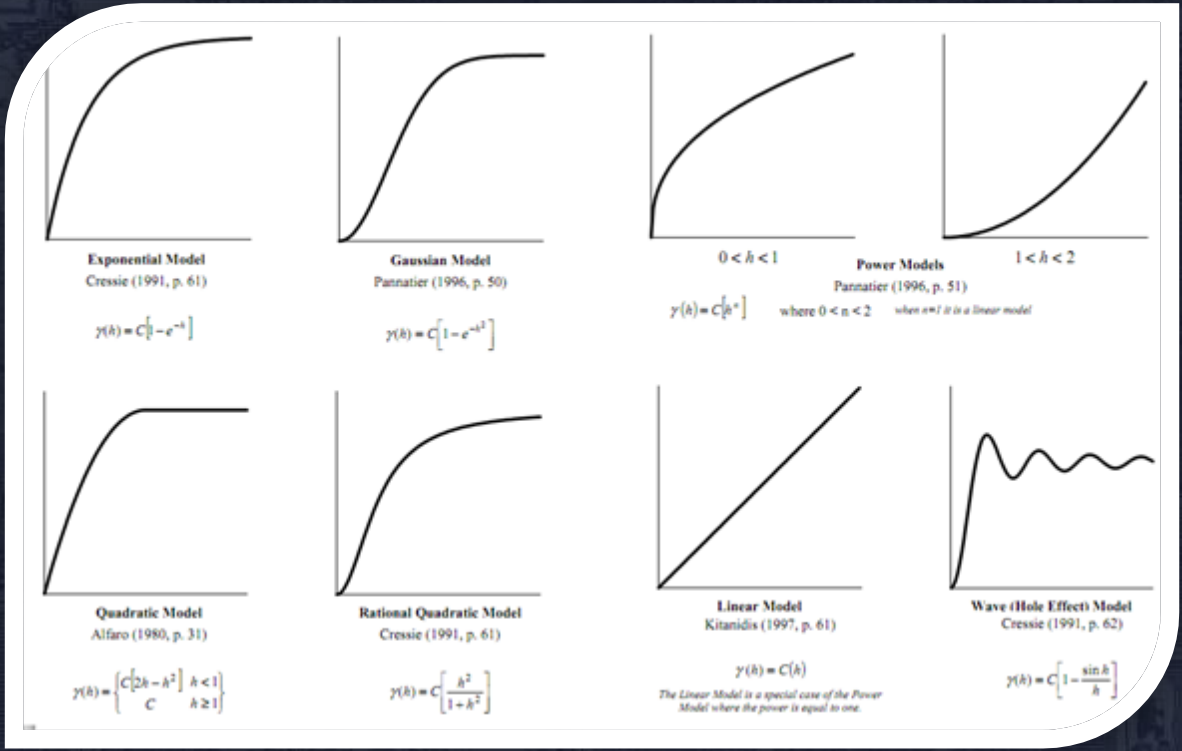


How to model the spatial structure of the residuals?

Semivariogram: describes the relationship between the differences of values at sampling localities and their spatial distances

This variation can be described by many mathematical models

We have to find the model that best describes the data!



The best model will be used to describe the spatial structure of the data and to correct the spatial bias of the residuals

Accounting for spatial autocorrelation in regression models

Simple model
Linear regression

$$Y_{ij} = \mu + \beta_1 \mathrm{Sex}_{ij} + \beta_2 \mathrm{Race}_{ij} + \beta_3 \mathrm{ParentsEduc}_{ij}$$

predictors

Mixed model with spatial structure of data

$$Y_{ij} = \mu + \beta_1 \mathrm{Sex}_{ij} + \beta_2 \mathrm{Race}_{ij} + \beta_3 \mathrm{ParentsEduc}_{ij} + U_i + W_{ij},$$

predictors

The models describing the spatial structure of data

Accounting for spatial autocorrelation in regression models

Selecting the best model

Finally we will compare ALL the models and will choose the one best describing the data

How to choose the best model?

AIC (Akaike Information Criterion), is an estimator of the RELATIVE quality of models in describing the observed data.

The model that better describes the data but has a low degree of complication is assigned a low score.

Low AIC scores are better than high AIC scores, thereby we have to consider the model with the lowest AIC score!

The reason why simple models are preferred to complex ones

Suppose we want to use the computed model to predict the Y value at a new X