# Environmental Statistics

## Slides - Part 1

Fedele Greco

Department of Statistical Sciences
University of Bologna

fedele.greco@unibo.it

# Contents

1. Course contents
2. Environmental Statistics
3. Introduction to Areal Data Analysis
4. Global measures of spatial autocorrelation

# Environmental Statistics - Part I

1. Course contents
2. Environmental Statistics
3. Introduction to Areal Data Analysis
4. Global measures of spatial autocorrelation

# Course contents

**Part I - 26 hours** (Fedele Greco)

1. Analysis of areal data
2. *Analysis of geostatistical data*

**Part II - 10 hours** (Aldo Gardini)

1. Analysis of point process data

# Contents of Part I

**Analysis of areal data**

1. Global measures of spatial association
2. Local Indicators of Spatial Association (LISA)
3. Hypothesis testing for global and local spatial association
4. Spatial linear regression

***Analysis of geostatitical data***

1. Geostatistics: descriptive measures of spatial dependence
2. Variogram and covariogram estimation
3. Spatial random fields
4. Spatial prediction: Kriging

# Teaching Material

**Slides and R code**:
can be downloaded at the course website

**Textbook - Part I**:
Bivand R.S., Pebesma E., Gómez-Rubio V. (2013) Applied Spatial Data Analysis with R. Springer.

Lovelace R., Nowosad J., Muenchow J., Geocomputation with R (can be downloaded from the Bookdown website)

# Exam

The final exam aims at evaluating the achievement of the following educational targets:

- deep knowledge of the topics covered along the course
- ability to analyse spatial data
- ability to implement statistical methods suited for spatial analysis in R.

The exam consists of a **practical test in the computer lab** and an **optional oral** exam. The practical test will include theoretical questions.

# Environmental Statistics - Part I

1. Course contents
2. Environmental Statistics
3. Introduction to Areal Data Analysis
4. Global measures of spatial autocorrelation

# Is Environmental Statistics a discipline?

If the question is cast as follows:

"Are there statistical methods that:

- are typical of environmental statistics?
- have been developed in the context of environmental statistics and have been adopted in other fields?"

then the answer is definitely **yes**.

# Environmental Statistics - History

In 1971, Philip Cox used for the first time the word **Environmetrics**

What does it mean?

“Problems in the atmospheric, ecological, geological, toxicological, biomedical and economic sciences, and concerns in public health, risk managment and social policy, have provided rich data for quantitative analyses, collectively called environmetrics.”

(Piegorsch et al. (1998) “Statistical advances in Environmental Sciences”, 13 186-208)

# Environmental Statistics - History

In 1989 during a meeting in the city of Cairo, the International Environmetric Society (**TIES**) was founded.

The society aims at fostering the development of statistical methods for environmental sciences, environmental engineering and for environmental monitoring and protection.

The society promotes the interaction among statisticians, mathematicians, engineers, physicians, meteorologists, (...), in environmental studies, highlighting the need for a **multidisciplinary approach**.

The Italian counterpart of TIES is the **GRASPA** (Research Group for Statistical Applications to Environmental Problems)

# Institutions

Italy: ISPRA (national environmental agency), ARPA (regional environmental agencies), ISTAT (Italian National Statistical Institute)

Europe: European Environment Agency (http://www.eea.europa.eu/)

United States: Environmental Protection Agency (EPA, www.epa.gov )

United Nations: http://unstats.un.org/unsd/environment/

# Environmental Phenomena - Challenges

Environmental Phenomena are characterized by the interaction of several factors, they are **complex phenomena**

Spatial and temporal **heterogeneity**

Very often, data are collected on different temporal and spatial scales (**misalignment problem**) and they come from **different sources**

# Environmental studies: few examples

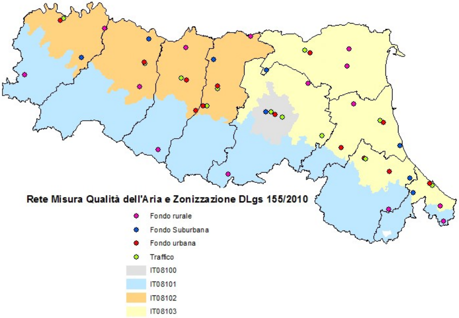
- Air quality, water quality, soil pollution... (with a spatio-temporal perspective)
- Effects of pollution on public health
- Climate modelling
- ...

These studies are relevant (among other things) because they contribute to environmental monitoring and they allow to identify areas subject to environmental risks. Moreover they provide valuable information for designing environmental and reclamation policies.

# How statisticians contribute to Environmental studies

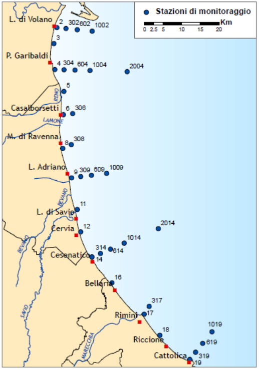
- Design of sampling schemes for environmental monitoring
- Environmental indexes
- Modeling environmental thresholds exceedance (thresholds are established by law)
- Models for **spatial**, temporal and spatio-temporal data
- Hierarchical models
- Methods for merging data from different surces
- Evaluation of uncertainty in deterministic models
- Methods for data reduction
- Extreme events modeling
- Environmental epidemiology
- Estimation of animal populations for the evaluation of ecological risk

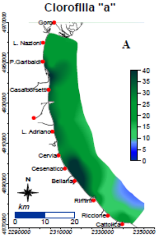
# Some example of environmental study - Air Quality



# Some example of environmental study - Water Quality

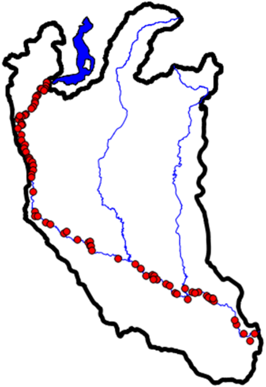
Nutrients in the North Adriatic Sea





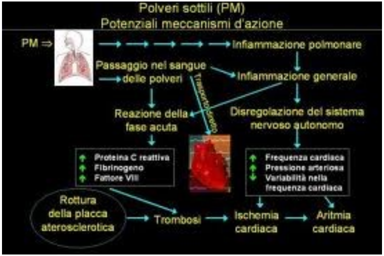
# Some example of environmental study - Water Quality

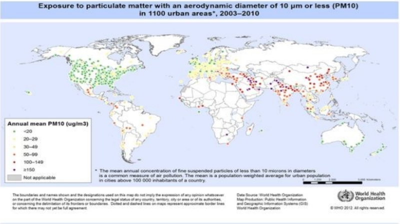
Modelling pollution in a river



# Some example of env. study - Spatial Epidemiology

Exposure to Particulate Matter (PM) and adverse health effect





Exposure to particulate matter with an aerodynamic diameter of 10 µm or less (PM10) in 1100 urban areas*, 2003–2010

* The mean annual concentration of fine suspended particles of less than 10 microns in diameters is a common measure of air pollution. The mean is a population-weighted average for urban population in cities above 100 000 inhabitants of a country.

The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

Data Source: World Health Organization
Map Production: Public Health Information and Geographic Information Systems (GIS)
World Health Organization

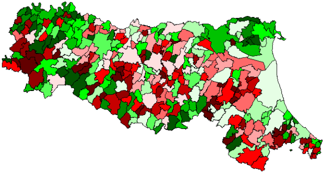
World Health Organization

© WHO 2012. All rights reserved.

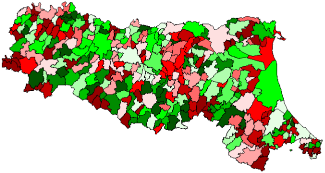
# Some example of env. study - Disease Mapping

## Smoothing Standardised Mortality Ratios

SMR - Disease 1 (70000 counts)



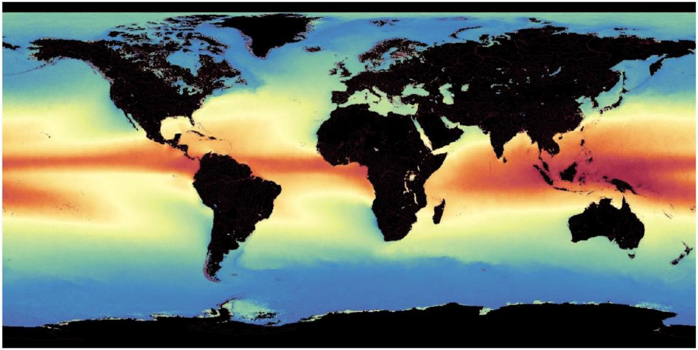
SMR - Disease 2 (3000 counts)



# Some example of env. study - Climate modelling

Detecting the Intertropical Convergence Zone

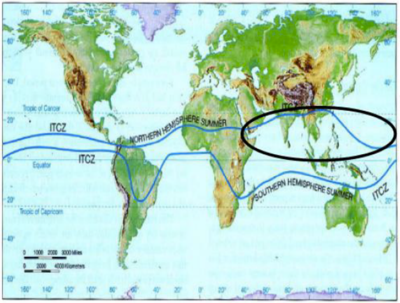
AIRWAVE dataset of TCWV (Total Column Water vapour)



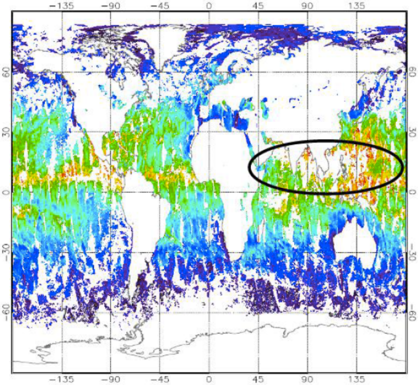
1991-2012 average TCWV global field at coarse resolution

# Some example of env. study - Climate modelling

Detecting the Intertropical Convergence Zone



AATSR August 2008 (5 days)



# Environmental Statistics - Part I

1. Course contents

2. Environmental Statistics

3. Introduction to Areal Data Analysis

4. Global measures of spatial autocorrelation

# Spatial Data

According to the classification proposed in Cressie (1993), there are three main types of spatial data.

Each data type requires appropriate statistical methodologies.

- **areal data** (also known as lattice data)
- **geostatistical data**
- point process data

Part I of this course is devoted to areal and geostatistical data.

From the methodological point of view, the main difference between these data is the **spatial support**.

# Areal Data

The spatial domain $D$ is partitioned in $n$ areas $\{A_1.A_2, ..., A_n\}$.

Areas $\{A_1.A_2, ..., A_n\}$ are a lattice of $D$ if

$$D = \bigcup_{i=1}^{n} A_i$$

and

$$A_i \cap A_j = \emptyset \qquad \forall i \neq j.$$

The variable under study is not observable “between” areas, i.e. the **domain is discrete.**

# Areal Data

One or more variables are observed in each area.

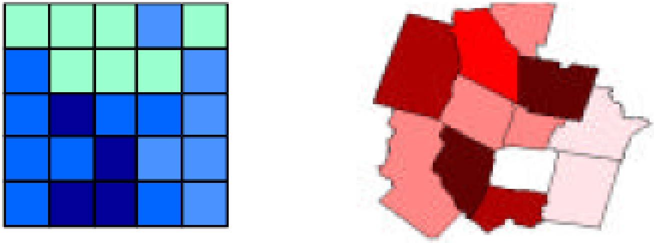
The domain partition is usually determined by administrative criteria (municipalities, counties, zip codes, provinces, ...).

The information for each area are usually aggregated measures: averages or counts are observed.

The observed value does not refer to a specific point in the study region: it refers to the entire area.

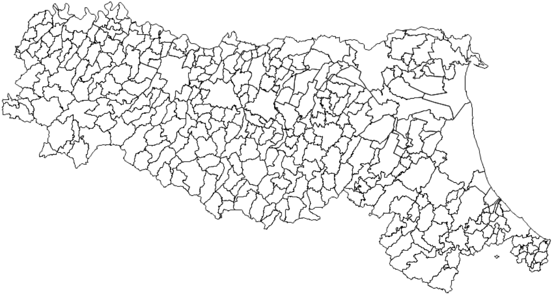
# Areal Data

Data collected over regular (left) and irregular (rigth) lattices



# An Irregular Lattice

Municipalities of the Emilia Romagna Region



# Adjacency Matrix

The spatial structure of the study region is described by the **adjacency matrix.**

This can be specified following different criteria. We will always build adjacency matrices according to the following criterion:

*two areas are neighbors if they share a common boundary*

Remember: this is a subjective choice, other choices are possible and the choice of the neigboring structure has effect on all statistical analyses.

# Adjacency Matrix

The adjacency matrix will play a prominent role both for describing and modeling areal data.

The adjacency matrix is denoted as $\boldsymbol{W}$, it is an $n \times n$ symmetric matrix, it defines conditional dependence relationships among areas.

By definition $w_{ii} = 0$, i.e. each area is not neighbor to itself.

On the other hand, $w_{ij} = w_{ji} = 1$ iff areas $i$ and $j$ are considered as neighbors (e.g. if they share a common boundary)

# Adjacency Matrix - Some notation

- $N(i)$: **set** of neighbors of $A_i$
- $w_{i+}$ : number of neighbors of $A_i$

Note that, $w_{ii} = 0$ implies that $A_i \notin N(i)$.

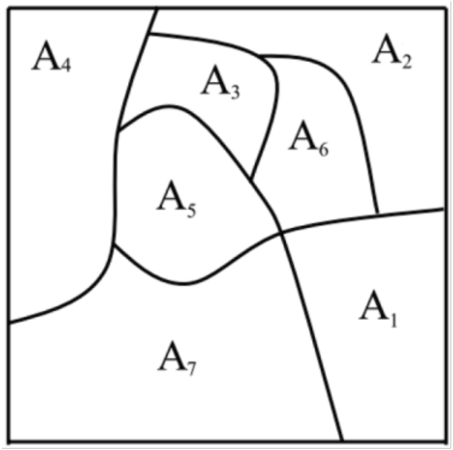
Summarizing:

$$w_{ij} = \begin{cases} 1 & \text{if} \quad j \in N(i), \quad j \neq i \\ 0 & \text{otherwise} \end{cases}$$

$$w_{i+} = \sum_{j=1}^{n} w_{ij}$$

# Building the Adjacency Matrix - A toy example

Consider the following region *D* partitioned in 7 areas.



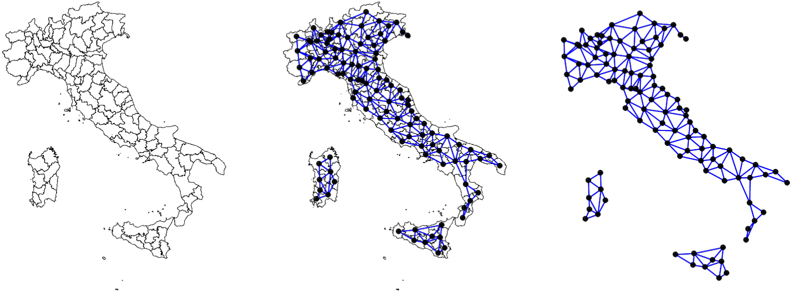
# Building the Adjacency Matrix - A toy example

| $A_i$ | $N(i)$ | $w_{i+}$ |
|---|---|---|
| 1 | 2,5,6,7 | 4 |
| 2 | 1,3,4,6 | 4 |
| 3 | 2,4,5,6 | 4 |
| 4 | 2,3,5,7 | 4 |
| 5 | 1,3,4,6,7 | 5 |
| 6 | 1,2,3,5,7 | 5 |
| 7 | 1,4,5,6 | 4 |

# Building the Adjacency Matrix - A toy example

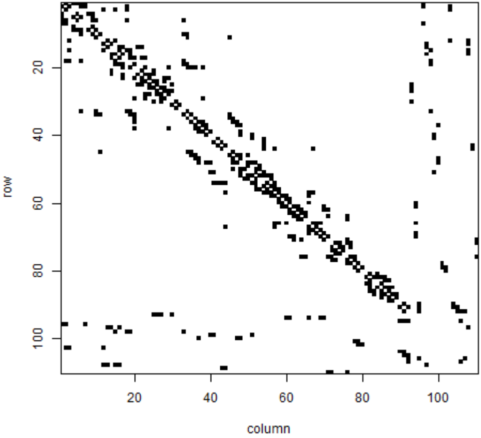
$$\boldsymbol{W} = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

# Adjacency Matrix - Italian Provinces



# Adjacency Matrix - Italian Provinces

A **sparse** matrix



# Adjacency Matrix - Observations

By building the adjacency matrix we are actually building a graph.

Graphs "are mathematical structures used to model pairwise relations between objects. A graph in this context is made up of vertices, nodes, or points which are connected by edges, arcs, or lines." Wikipedia

Adjacency matrices are usually sparse. Sparse matrices allow for faster computations with respect to dense matrices.

# Environmental Statistics - Part I

1. Course contents

2. Environmental Statistics

3. Introduction to Areal Data Analysis

4. Global measures of spatial autocorrelation

# Global measures of spatial autocorrelation

Tobler’s First Law of Geography: "everything is related to everything else, but near things are more related than distant things." (Tobler, 1970)

Spatial autocorrelation measures the correlation of a variable with itself through space. (You already know that temporal autocorrelation measures the correlation of a variable with itself through time).

# Global measures of spatial autocorrelation

Suppose that a variable ($y(A_i)$) has been observed at each area $A_i$ of the domain $D$. We will use the notation $y_i = y(A_i)$, the data vector is denoted as

$$\boldsymbol{y} = (y_1, \ldots, y_i, \ldots y_n)^T$$

We will consider the following measures of spatial correlation:

- Moran's Index
- Ord's Index
- Approximate Profile Likelihood Estimator (APLE)

# Global measures of spatial autocorrelation

Measures of spatial autocorrelation are built as weighted averages of similarity measures of observed data with weights accounting for spatial proximity.

$$\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} s_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}}$$

# Global Moran's *I*

By choosing the following similarity measure

$$s_{ij} = n\frac{(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

one obtains a well-known measure of spatial autocorrelation, **the Global Moran's *I*.**

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

# Global Moran's *I*

With no further assumptions on spatial weights $w_{ij}$, Moran's $I$ is not easily interpretable as it depends on $n$ and on the sum of weights.

It is customary to **row-standardize the adjacency matrix $\boldsymbol{W}$**. The row-standardized adjacency matrix is denoted as $\widetilde{\boldsymbol{W}}$.

$$
\begin{aligned}
\widetilde{w}_{ij} &= \frac{w_{ij}}{w_{i+}} = \frac{1}{w_{i+}} && \text{if areas } i \text{ and } j \text{ are neighbors} \\
\widetilde{w}_{ij} &= 0 && \text{otherwise} \\
\sum_{i=1}^{n} \widetilde{w}_{ij} &= 1 && \forall i = 1, \ldots, n
\end{aligned}
$$

i.e. all row sums are equal to 1.

# Spatial lag

The spatial lag is defined as:

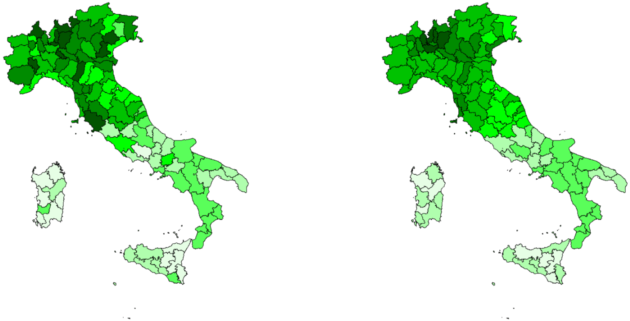
$$\widetilde{y}_i = \sum_{j=1}^{N} \widetilde{w}_{ij} y_j = \frac{1}{w_{i+}} \sum_{j \in N(i)} y_j$$

Note that $\widetilde{y}_i$ is the mean of the variable in $N(i)$.

Introducing the spatial lag allows a more intuitive discussion of global (and local) spatial autocorrelation.
In time series analysis $L(y_t) = y_{t-1}$, but no obvious ordering is available in space.

# Spatial lag

Observed data $\boldsymbol{y}$ (left) and spatial lag $\widetilde{\boldsymbol{y}}$ (right):



The map of $\widetilde{\boldsymbol{y}}$ is *smoother* than the map of $\boldsymbol{y}$. This is an expected result: indeed, $\widetilde{\boldsymbol{y}}$ is a (spatial) moving average.

# Global Moran's $I$

Using the row-standardized adjacency matrix $\widetilde{\boldsymbol{W}}$ one obtains:

$$I = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \widetilde{w}_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

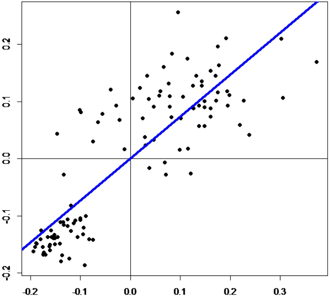
$$I = \frac{\sum_{i=1}^{n}(y_i - \bar{y}) \sum_{j=1}^{n} \widetilde{w}_{ij}(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\widetilde{y}_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$I = \frac{\text{Cov}(\boldsymbol{y}, \widetilde{\boldsymbol{y}})}{\text{Var}(\boldsymbol{y})}$$

i.e. Moran's $I$ is the regression coefficient of $\widetilde{\boldsymbol{y}}$ on $\boldsymbol{y}$

# Spatial correlation and spatial lag: Moran scatter plot

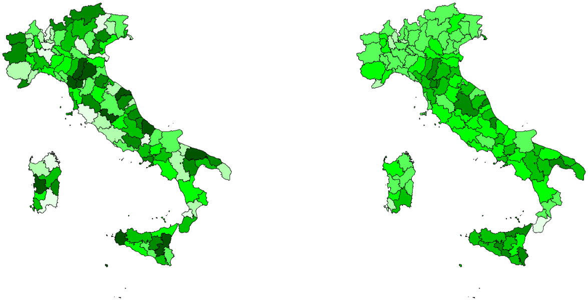
Scatter plot of $\widetilde{\boldsymbol{y}} - \bar{y}$ vs $\boldsymbol{y} - \bar{y}$:



Moran's $I$, ($I = 0.735$) is the slope of the regression line (blue).

# Data simulated from an i.i.d. process

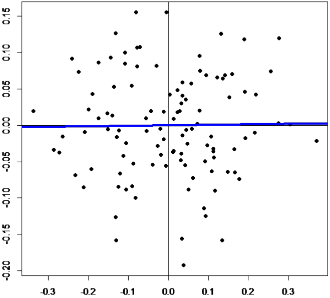
Simulated data $\boldsymbol{y}$ (left) and spatial lag $\widetilde{\boldsymbol{y}}$ (right):



Since the data generating process is i.i.d., the map of $\widetilde{\boldsymbol{y}}$ is **far** *smoother* than the map of $\boldsymbol{y}$.

# Data simulated from an i.i.d. process

Scatter plot of $\widetilde{\boldsymbol{y}} - \bar{y}$ vs $\boldsymbol{y} - \bar{y}$:



Moran's $I$, ($I = 0.007$) is the slope of the regression line (blue).

# Moran's I and Ord Index - Matrix form

If $\boldsymbol{y}$ and $\widetilde{\boldsymbol{y}}$ are centered, Moran's $I$ can be expressed in matrix form as:

$$I = \frac{\boldsymbol{y}^T \widetilde{\boldsymbol{W}} \boldsymbol{y}}{\boldsymbol{y}^T \boldsymbol{y}} = \frac{\boldsymbol{y}^T \widetilde{\boldsymbol{y}}}{\boldsymbol{y}^T \boldsymbol{y}}$$

The Ord Index is just Moran's $I$ with a different denominator, namely, it is the regression coefficient of $\boldsymbol{y}$ on $\widetilde{\boldsymbol{y}}$.

$$O = \frac{\boldsymbol{y}^T \widetilde{\boldsymbol{W}} \boldsymbol{y}}{\boldsymbol{y}^T \widetilde{\boldsymbol{W}}^T \widetilde{\boldsymbol{W}} \boldsymbol{y}} = \frac{\boldsymbol{y}^T \widetilde{\boldsymbol{y}}}{\widetilde{\boldsymbol{y}}^T \widetilde{\boldsymbol{y}}}$$

# Toward the APLE index

In exploratory analyses of spatial data, a formal statistical model may not be explicitly assumed. However, we argue that in these situations the informal notion of spatial dependence, or spatial autocorrelation, is often implicitly based on a Spatial Autoregressive (SAR) framework where the goal is to assess the predictive ability of neighboring values of the data. In order for informal assessments of the strength of spatial autocorrelation to translate into this implied formal statistical modeling framework, exploratory spatial data analysis (ESDA) tools should be based on estimators of the spatial autocorrelation parameter in the SAR model. (Lee et al., 2007)

# The Spatial Autoregressive process

The spatial autoregressive (SAR) model is commonly used to analyze spatial processes on a lattice.

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_i, \ldots, Y_n)^T$ be an $n$-dimensional **random vector**, the SAR model is specified as:

$$\boldsymbol{Y} = \rho \widetilde{\boldsymbol{W}} \boldsymbol{Y} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$$

$$\boldsymbol{Y} = \rho \widetilde{\boldsymbol{Y}} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$$

$$Y_i = \rho \widetilde{Y}_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

Parameter $\rho$ captures the strength of spatial association.

# The Spatial Autoregressive process

Moments of the SAR process

$$\begin{aligned}
\boldsymbol{Y} &= (\boldsymbol{I}_n - \rho\widetilde{\boldsymbol{W}})^{-1}\boldsymbol{\epsilon} \\
E(\boldsymbol{Y}) &= \boldsymbol{0} \\
V(\boldsymbol{Y}) &= \sigma^2\left((\boldsymbol{I}_n - \rho\widetilde{\boldsymbol{W}})^T(\boldsymbol{I}_n - \rho\widetilde{\boldsymbol{W}})\right)^{-1}
\end{aligned}$$

Thus the joint distribution of $\boldsymbol{Y}$ is:

$$\boldsymbol{Y} \sim N_n\left(\boldsymbol{0}, \sigma^2\left((\boldsymbol{I}_n - \rho\widetilde{\boldsymbol{W}})^T(\boldsymbol{I}_n - \rho\widetilde{\boldsymbol{W}})\right)^{-1}\right)$$

# The APLE Index

A closed-form estimator of the spatial correlation parameter $\rho$ is not available.

Li et al. (2007) proposed the Approximate Profile Likelihood Estimator (APLE):

$$APLE = \frac{\boldsymbol{Y}^T \widetilde{\boldsymbol{W}} \boldsymbol{Y}}{\boldsymbol{Y}^T \widetilde{\boldsymbol{W}}^T \widetilde{\boldsymbol{W}} \boldsymbol{Y} + \boldsymbol{Y}^T \left(\boldsymbol{\lambda}^T \boldsymbol{\lambda} \boldsymbol{I}_n / n\right) \boldsymbol{Y}} = \frac{\boldsymbol{Y}^T \widetilde{\boldsymbol{W}} \boldsymbol{Y}}{\boldsymbol{Y}^T \left(\widetilde{\boldsymbol{W}}^T \widetilde{\boldsymbol{W}} + \boldsymbol{\lambda}^T \boldsymbol{\lambda} \boldsymbol{I}_n / n\right) \boldsymbol{Y}}$$

where $\boldsymbol{\lambda}$ is the vector of eigenvalues of $\widetilde{\boldsymbol{W}}$.
This estimator is obtained by maximizing a first-order approximation of the profile likelihood of $\rho$.

# The APLE Index

The likelihood function $L(\rho, \sigma^2)$ is proportional to:

$$(\sigma^2)^{-\frac{n}{2}} |(\boldsymbol{I}_n - \rho \widetilde{\boldsymbol{W}})^T (\boldsymbol{I}_n - \rho \widetilde{\boldsymbol{W}})|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \boldsymbol{Y}^T (\boldsymbol{I}_n - \rho \widetilde{\boldsymbol{W}})^T (\boldsymbol{I}_n - \rho \widetilde{\boldsymbol{W}}) \boldsymbol{Y} \right\}$$

To obtain the APLE estimator, we keep $\rho$ fixed and obtain the maximum likelihood estimator of $\sigma^2$:

$$\hat{\sigma}^2_{ML} = \frac{\boldsymbol{Y}^T (\boldsymbol{I}_n - \rho \widetilde{\boldsymbol{W}})^T (\boldsymbol{I}_n - \rho \widetilde{\boldsymbol{W}}) \boldsymbol{Y}}{n}$$

# The APLE Index

Substituting $\hat{\sigma}^2_{ML}$ in the likelihood function, one obtains the profile likelihood for $\rho$.

It turns out that a closed form estimator is not yet available. For this reason the profile likelihood is approximated (first-order Taylor approximation), then the APLE is the maximizer of the *approximated* profile likelihood.

We will discuss the effect of this approximation in a *simulation study*.

# Global Indices - A simulation study

The aim of this simulation study is to compare Moran's *I*, Ord and APLE indices with respect to their performances as estimators of $\rho$.

Among other things, we are interested in the Mean Squared Error of the estimators as a function of $\rho$.

Why do we need to perform a simulation study?
Because the sampling distribution of the indices, as well as their expected value and variance, are not known in closed form.

# Global Indices - A simulation study

Simulation from SAR models with different $\rho$ values. (See the R script `GlobalIndices-Simulation.R`)

$$\rho = (0, 0.11, 0.21, 0.32, 0.42, 0.53, 0.63, 0.74, 0.84, 0.95)$$

For each value of $\rho$, we draw $K = 5000$ samples from a SAR process. Note that we are drawing 5000 $n$-dimensional vectors that we denote as:

$$\boldsymbol{y}_k^*, \quad k = 1, \ldots, 5000$$

*For each* simulated vector, and for each vlue of $\rho$, we compute Moran's $I$, Ord and APLE indices.

# Global Indices - A simulation study

As an example, consider Moran's $I$ (of course the same rationale applies to all indices).
For each value of $\rho$, we obtain $K$ values

$$\boldsymbol{I}^* = (I_1^*, I_2^*, \ldots, I_k^*, \ldots, I_K^*)$$

where

$$I_k^* = \frac{(\boldsymbol{y}_k^* - \bar{y})^T \widetilde{\boldsymbol{W}} (\boldsymbol{y}_k^* - \bar{y})}{(\boldsymbol{y}_k^* - \bar{y})^T (\boldsymbol{y}_k^* - \bar{y})}$$

We are drawing samples from the sampling distribution of an estimator. (Remember that we want to assess the performances of Moran's $I$ as an estimator of the strength of spatial autocorrelation, i.e. as an estimator of $\rho$).

# Global Indices - A simulation study

Once we have samples from the sampling distribution of the index, we can both view and summarise it.
The quantities below are **Monte Carlo (MC) approximations** of the moments of the estimators.

Expected value (MC approximation)

$$E(I|\rho) \cong \bar{I}^* = \frac{1}{K}\sum_{k=1}^{K} I_k^*$$

Variance (MC approximation)

$$V(I|\rho) \cong \frac{1}{K}\sum_{k=1}^{K}(I_k^* - \bar{I}^*)^2$$

# Global Indices - A simulation study

Thus we can compute MC approximations of the Bias, Variance and Mean Squared error of the estimator.

Bias

$$B(I;\rho) = E(I;\rho) - \rho$$

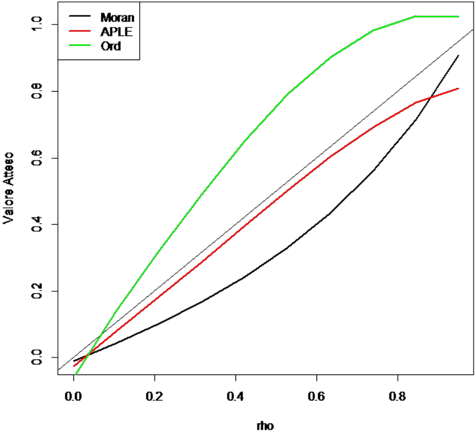
Variance

$$V(I;\rho) = E\left[(I - E(I;\rho))^2 ; \rho\right]$$

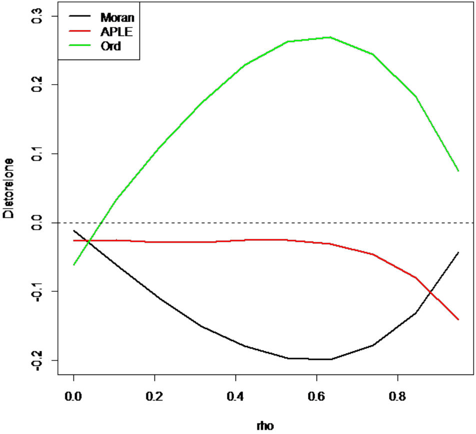
Mean Squared Error

$$MSE(I;\rho) = E\left[(I-\rho)^2;\rho\right] = V(I;\rho) + [B(I;\rho)]^2$$

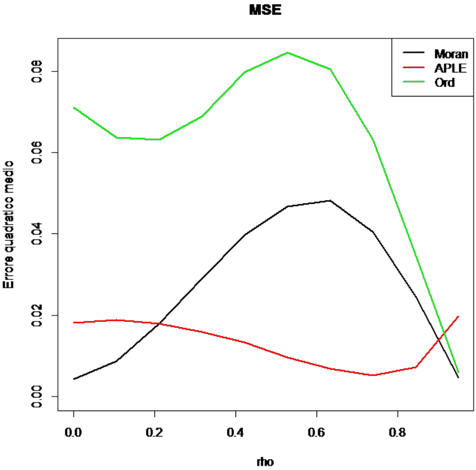
# The Expected Value as a function of $\rho$



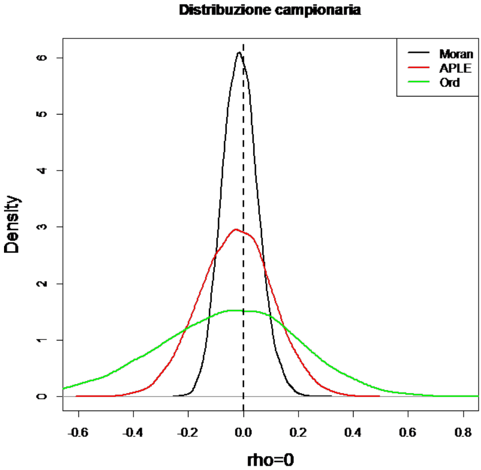
# The Bias as a function of $\rho$



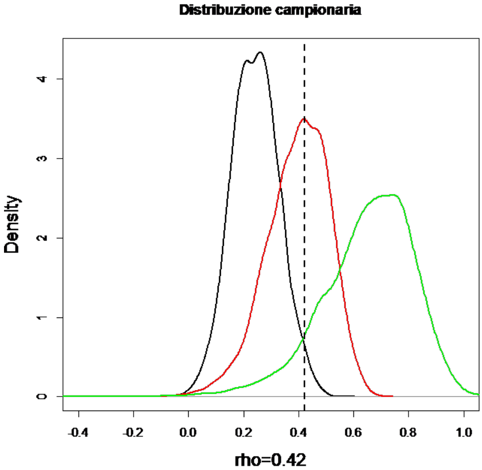
# The Mean Squared Error as a function of $\rho$



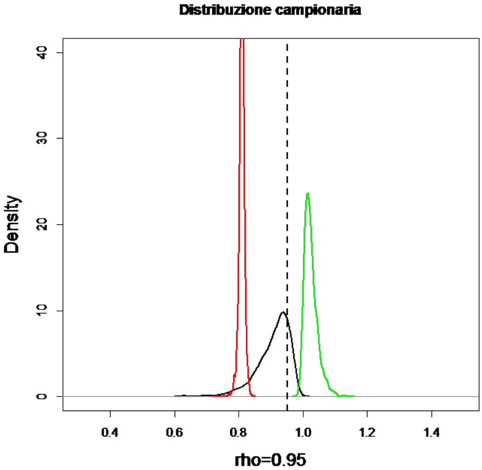
# Sampling distributions - $\rho = 0$



# Sampling distributions - $\rho = 0.42$



# Sampling distributions - $\rho = 0.95$


Distribuzione campionaria
Density
rho=0.95