Spatial Analysis and Modeling

04 - ESDA

Instructor: Rossano Schifanella

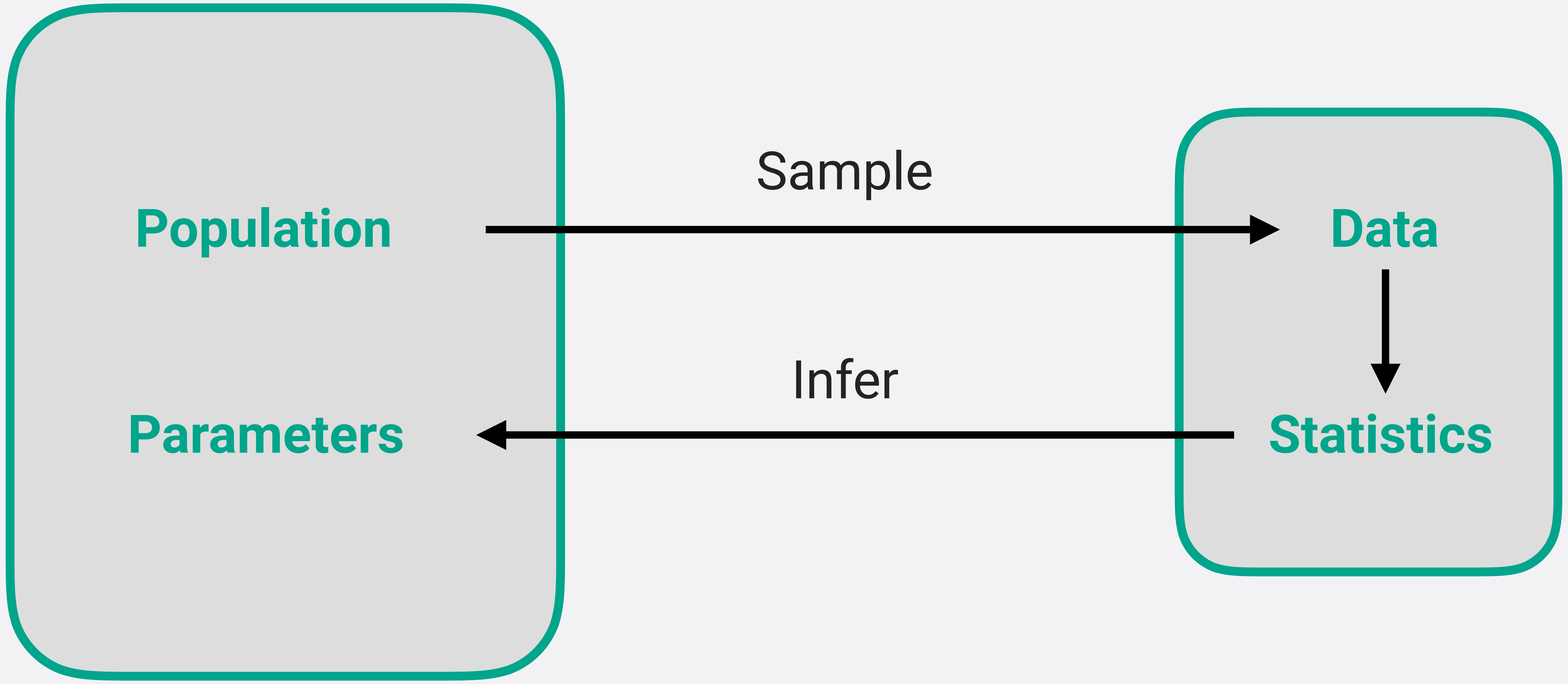
@UNITO

spatial statistical inference

Statistical inference

Statistical inference is the act of **generalizing from a sample to a population with calculated degree of certainty.**

we want to learn about population parameters



using statistics calculated in the sample

Statistical inference
main concepts

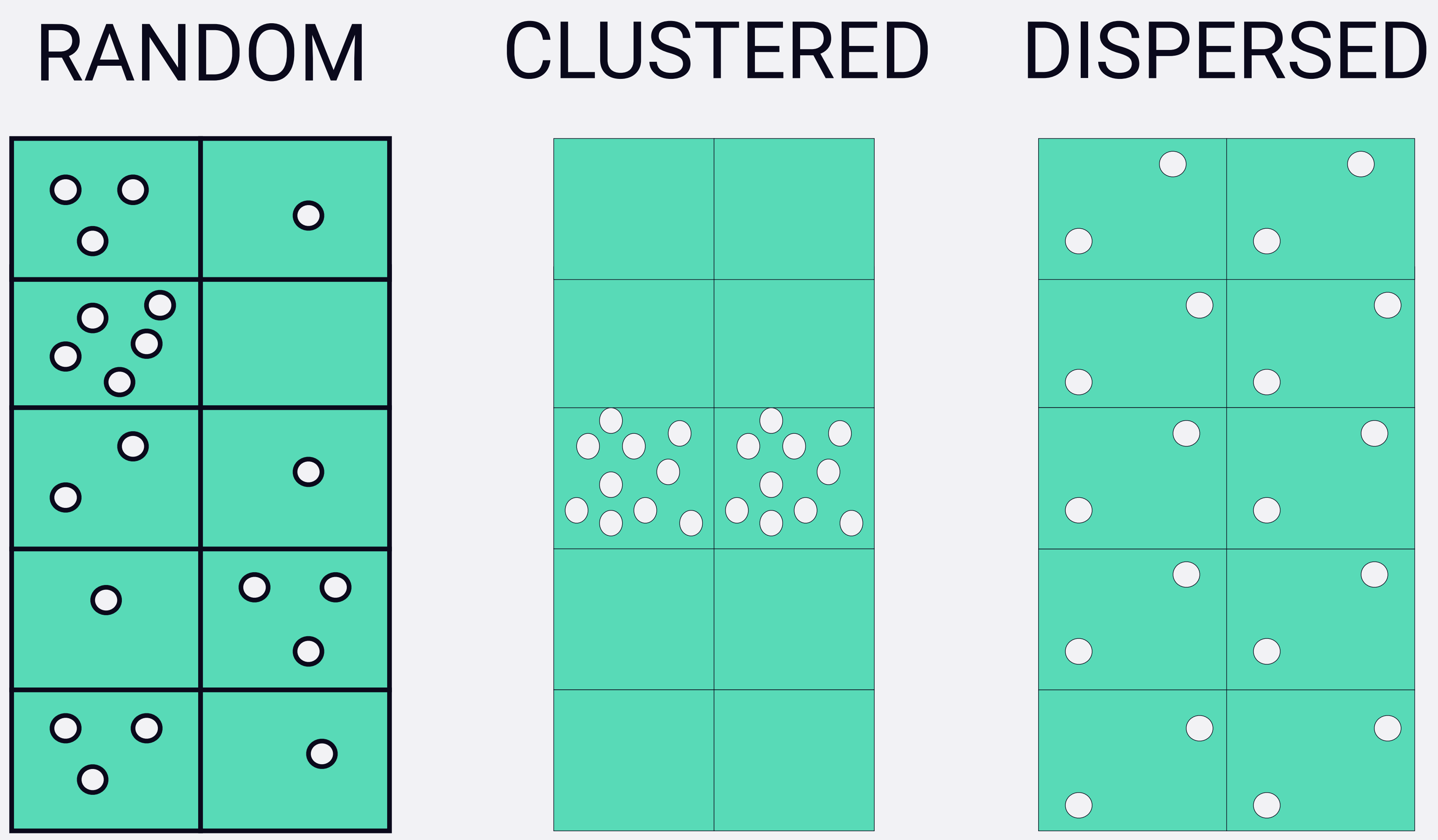
- Population
- Population parameters
- Sample
- Sample statistics
- Hypothesis testing
- Sampling distribution of a statistic
- Statistical significance test

Spatial Statistical Inference:
Null and Alternative Hypotheses

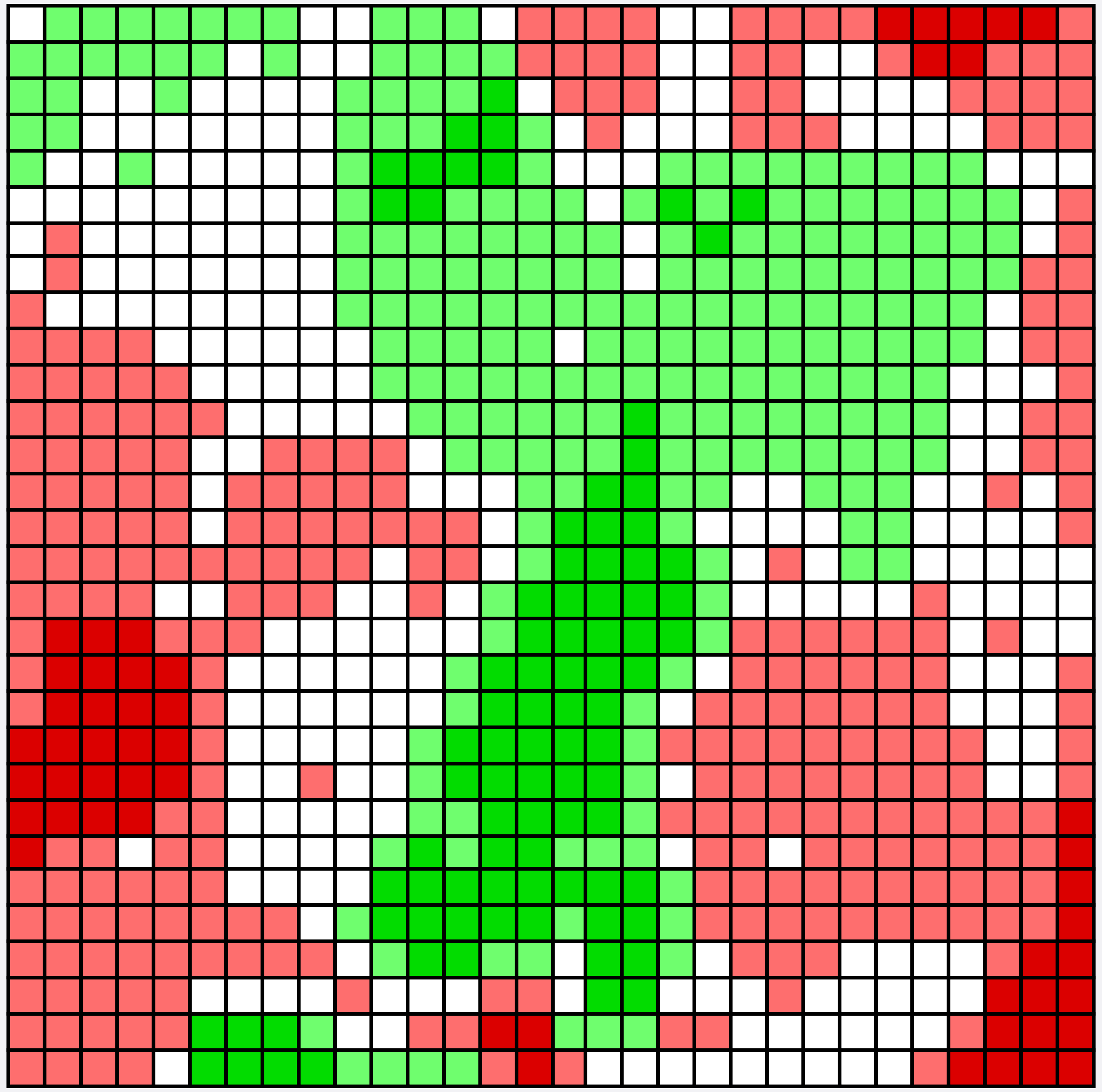
- **Null Hypothesis:**
  - The observed spatial pattern is random
  - Usually called **Complete Spatial Randomness** (CSR) hypothesis
  - Not very interesting!
- **Alternative Hypothesis:**
  - The spatial pattern is not random
  - It may be **clustered** or **dispersed**

What do we mean by spatially random?

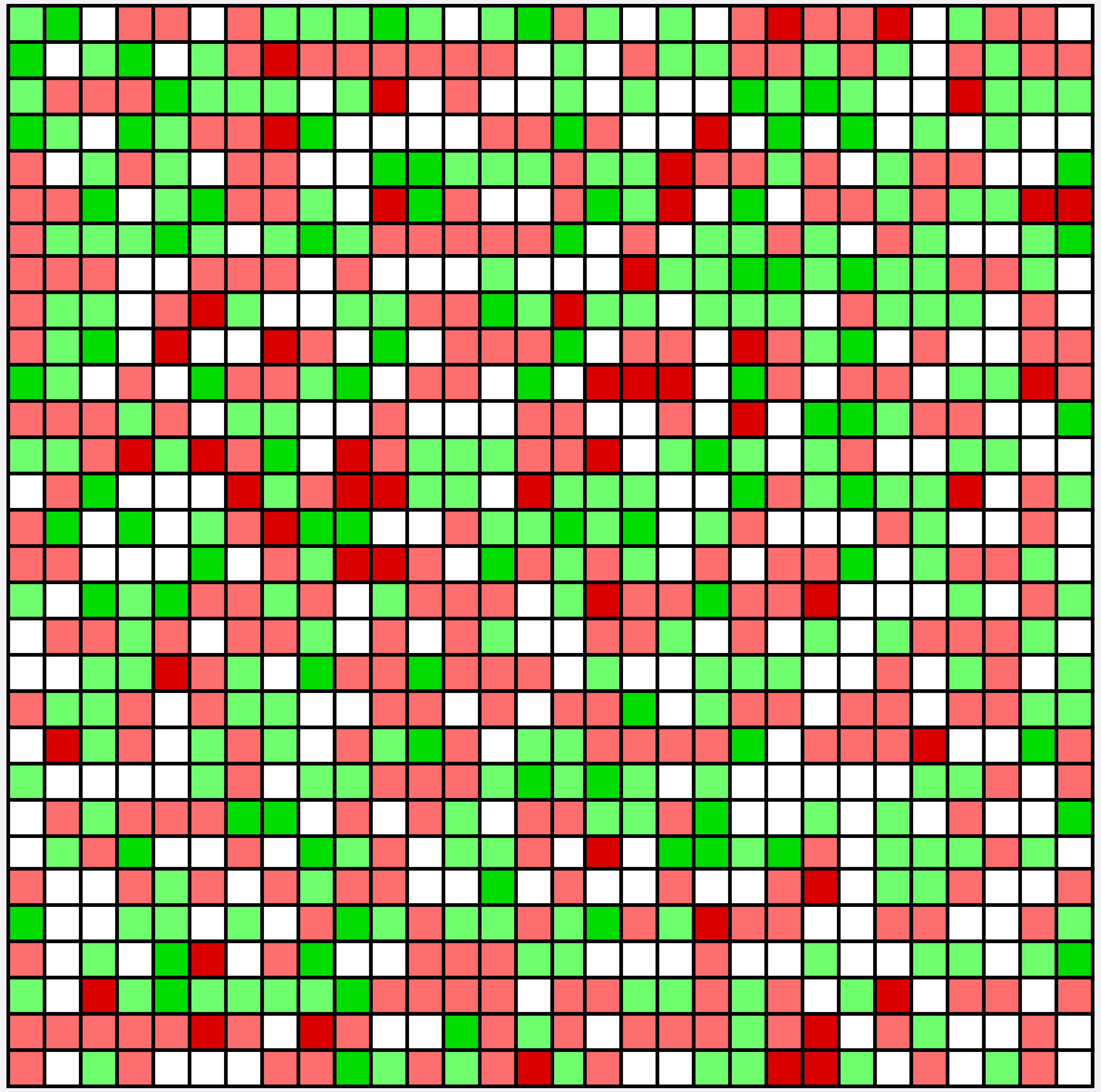
- **Random**
  - An event is equally likely to occur at any location
  - The position of an event is not affected by the position of any other event
- **Clustered**
  - every event is close to other event
- **Dispersed/Uniform**
  - every event is as far from other events as possible



High Peak district biomass index:
ratio of remotely sensed data spectral bands B3 and B4



SPATIALLY CLUSTERED



GEOGRAPHICALLY RANDOM

first order or second order effect?
example of bank robberies

- **Bank robberies are clustered**
    - **First order**
        - because banks are clustered
        - we call this the effect of “non-uniformity of space”
    - **Second order**
        - because one robbery influences nearby robberies
- **In practice, it is very difficult to distinguish these two effects merely by the analysis of spatial data**

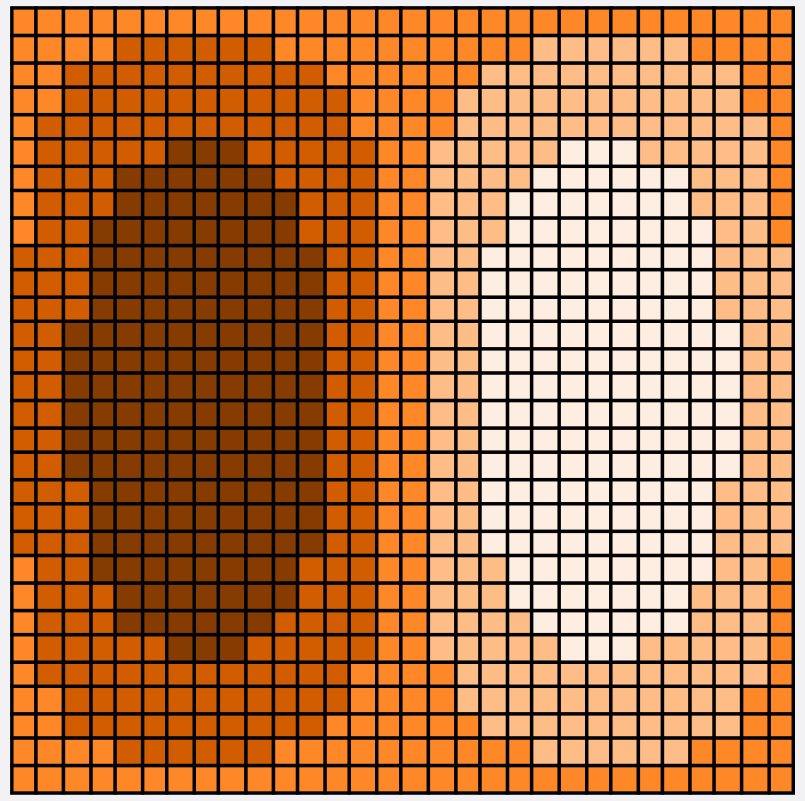
spatial statistical hypothesis testing:
simulation approach

- **Because of the complexity of spatial processes, it is often difficult to derive theoretically a test statistic with known probability distribution**
- **Instead, we often use computer simulations**
    - We take multiple samples from a random spatial pattern, the spatial statistic we are using is calculated for each sample, and then a frequency distribution is drawn
- **This simulated sampling distribution is used to measure the probability of obtaining our actual observed spatial statistic**

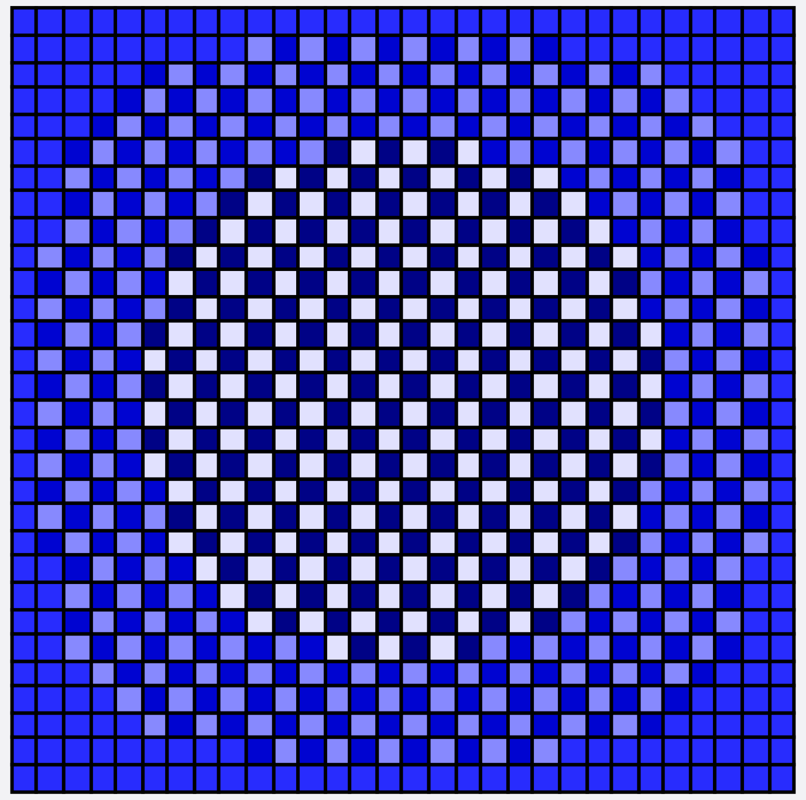
spatial autocorrelation

spatial autocorrelation

- **Measures the correlation of a variable with itself through space**
- **Related to Tobler's first law of geography**
  - Everything is related to everything else, but near things are more related than distant things.



positive = clustered



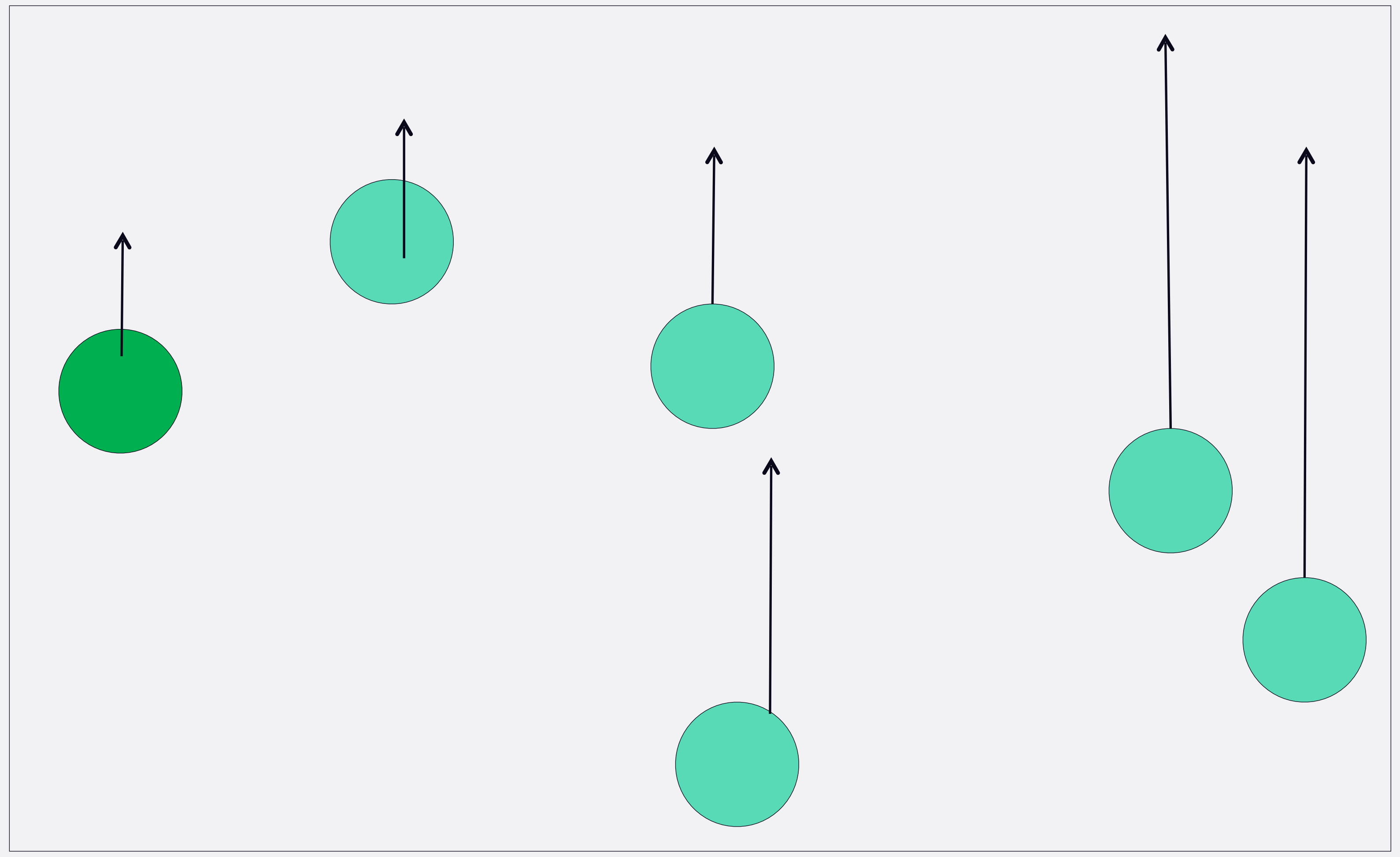
negative = dispersed

why is spatial autocorrelation important?

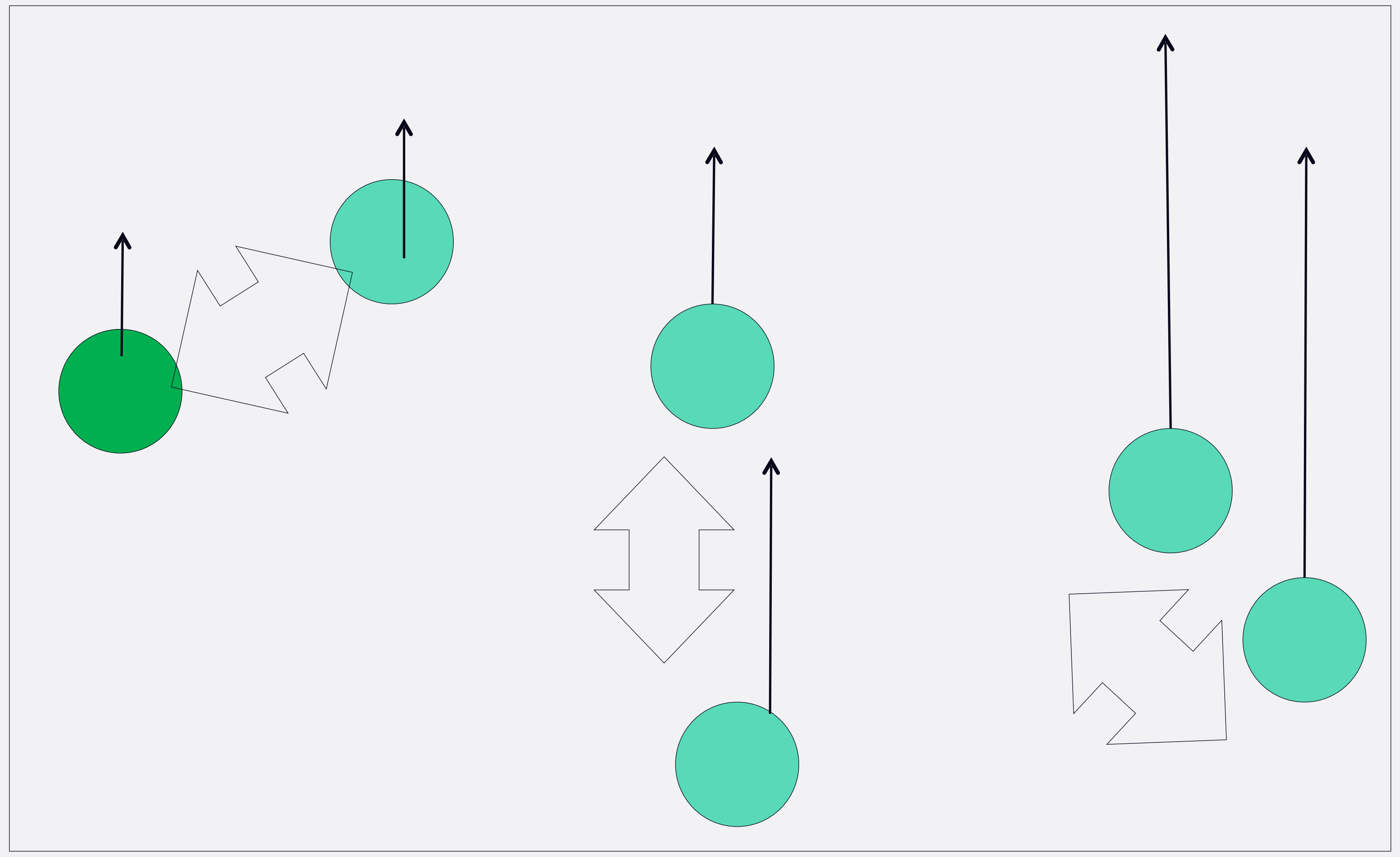
- **It implies the existence of a spatial process**
    - Why are near-by areas similar to each other?
    - Why do high income people live close each other?
    - These are geographical questions.
        - They are about location
- **It invalidates most traditional statistical inference tests**
    - If spatial autocorrelation exists, the results of standard statistical inference tests may be incorrect
    - We need to use spatial statistical inference tests
- **For example**
    - You are more likely to incorrectly conclude a relationship exists when it does not
    - You believe that the relationship is stronger than it really is

Why are standard statistical tests wrong?

- **Statistical tests are based on the assumption that the values of observations in each sample are independent of one another**
- **spatial autocorrelation violates this**
  - samples taken from nearby areas are related to each other and are not independent



Values near each other are <u>similar in magnitude.</u>

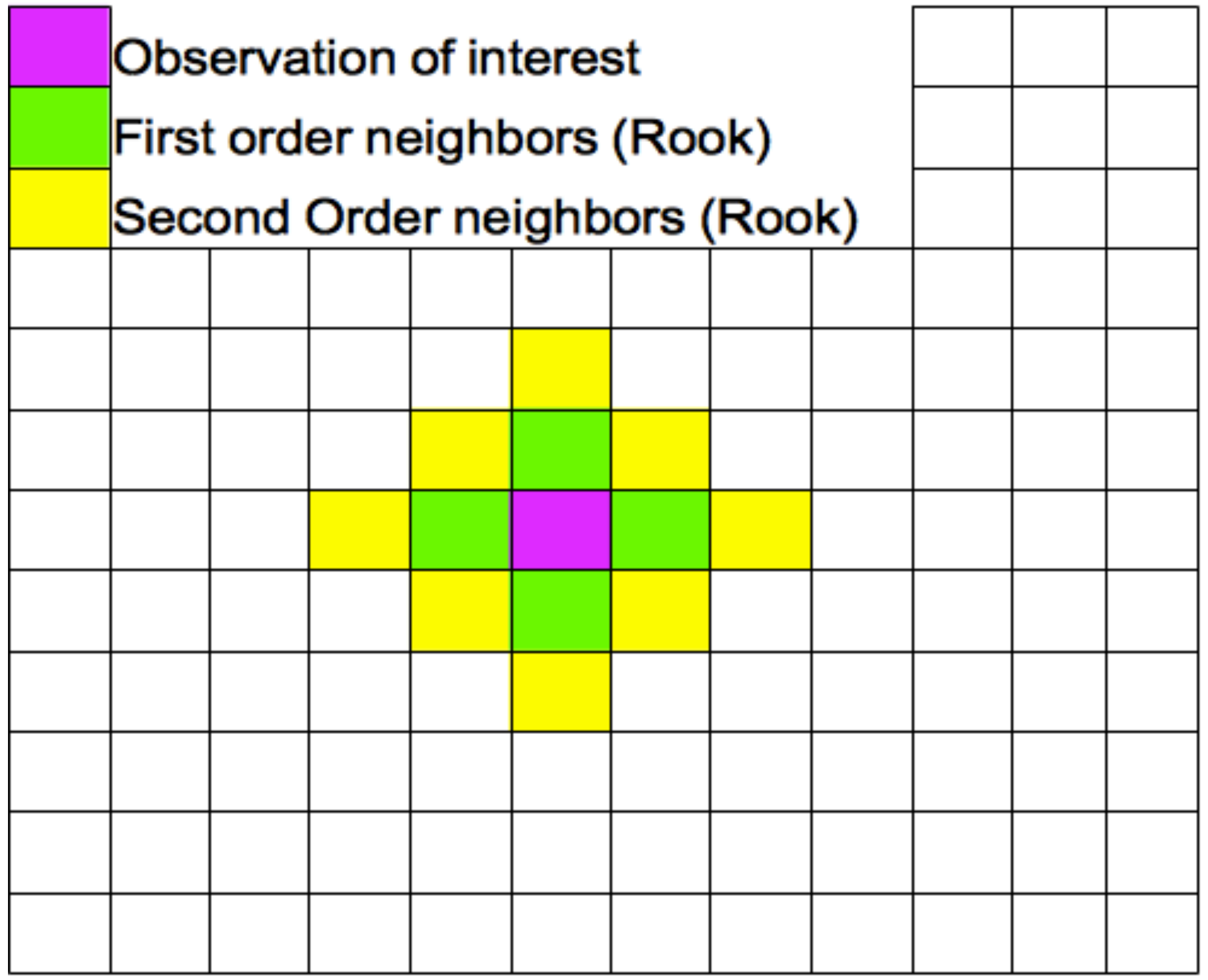


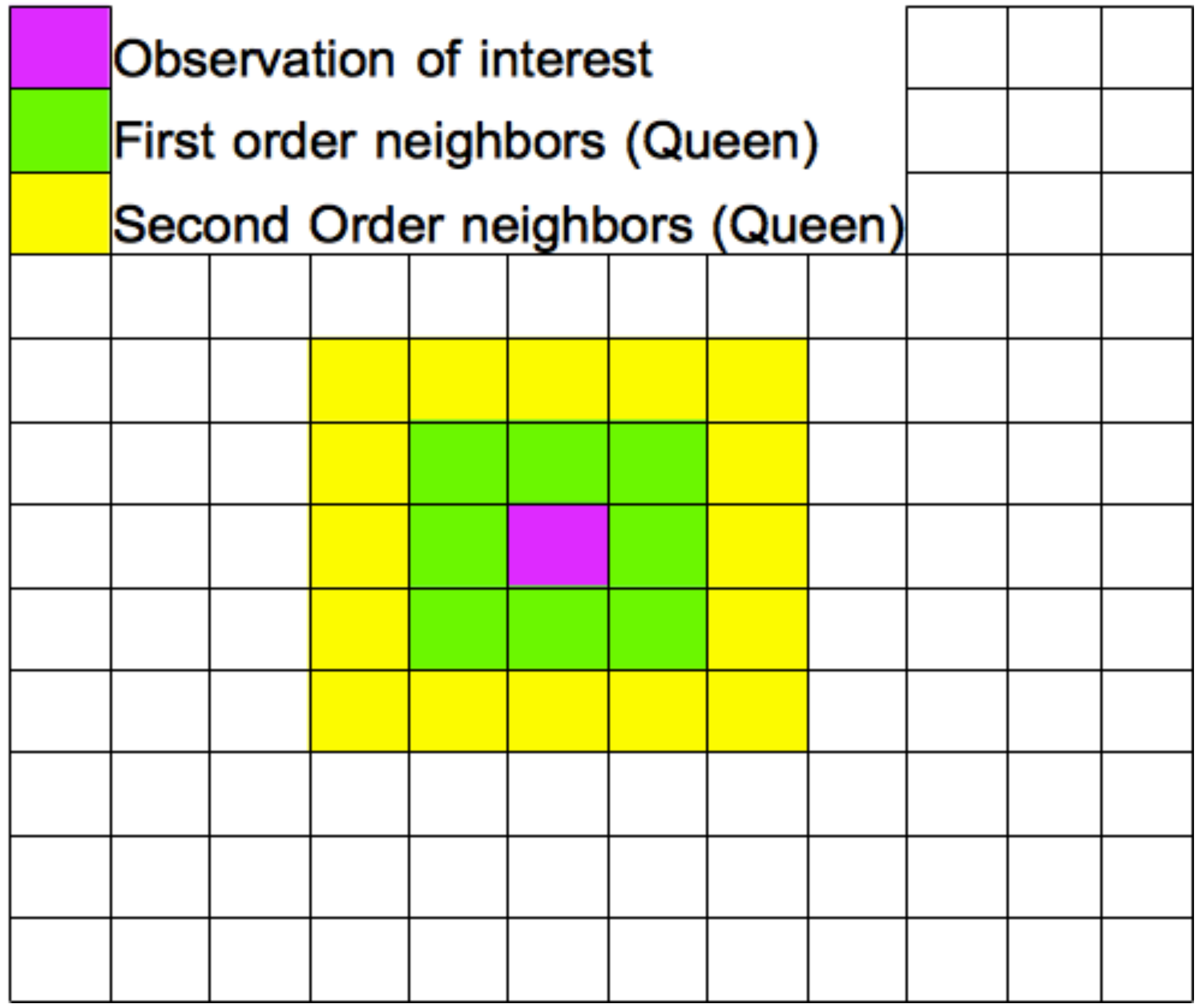
Implies a <u>relationship</u> between nearby observations

spatial autocorrelation: the problem of measuring proximity

contiguity

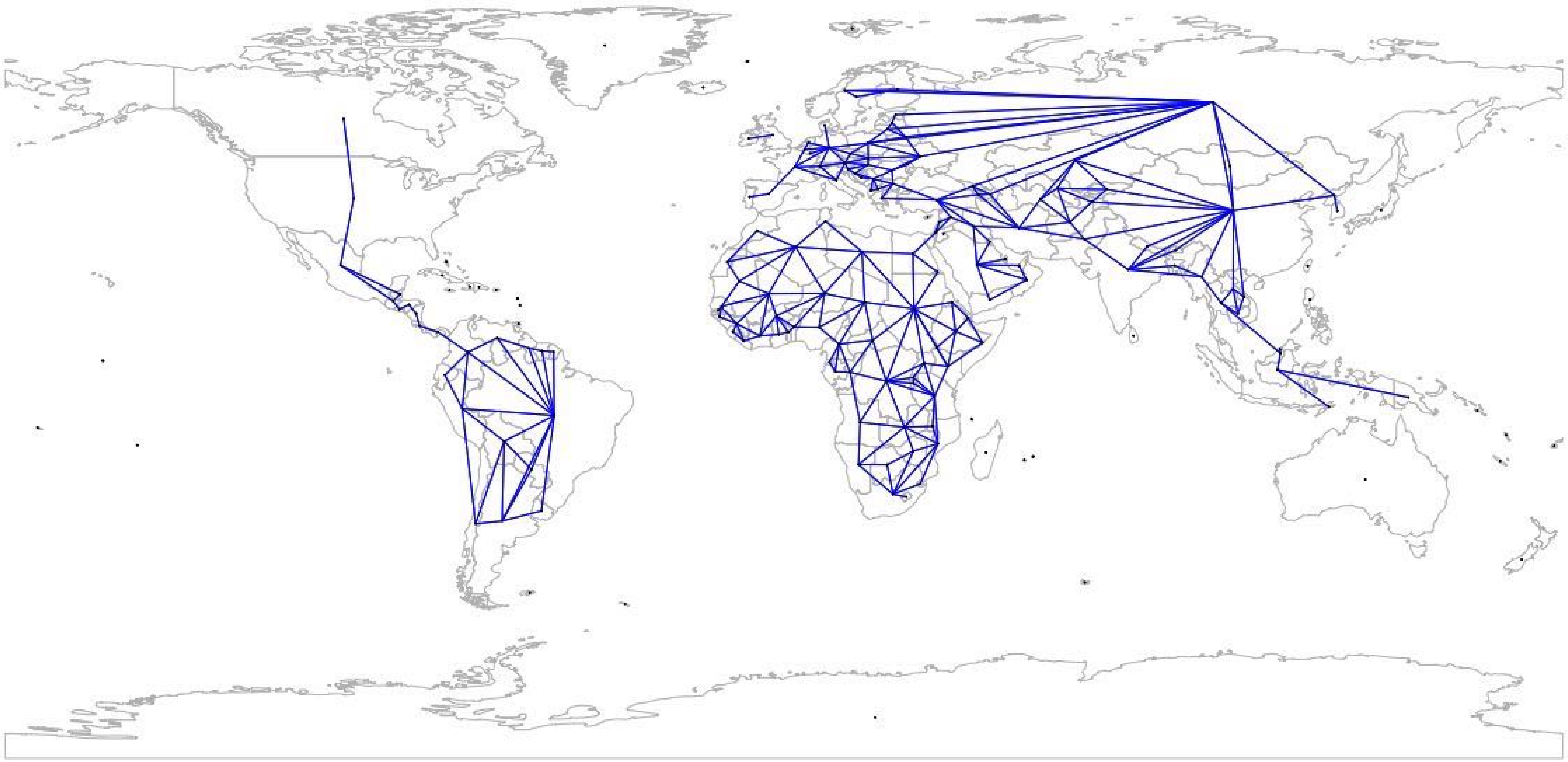
REGULAR GRID





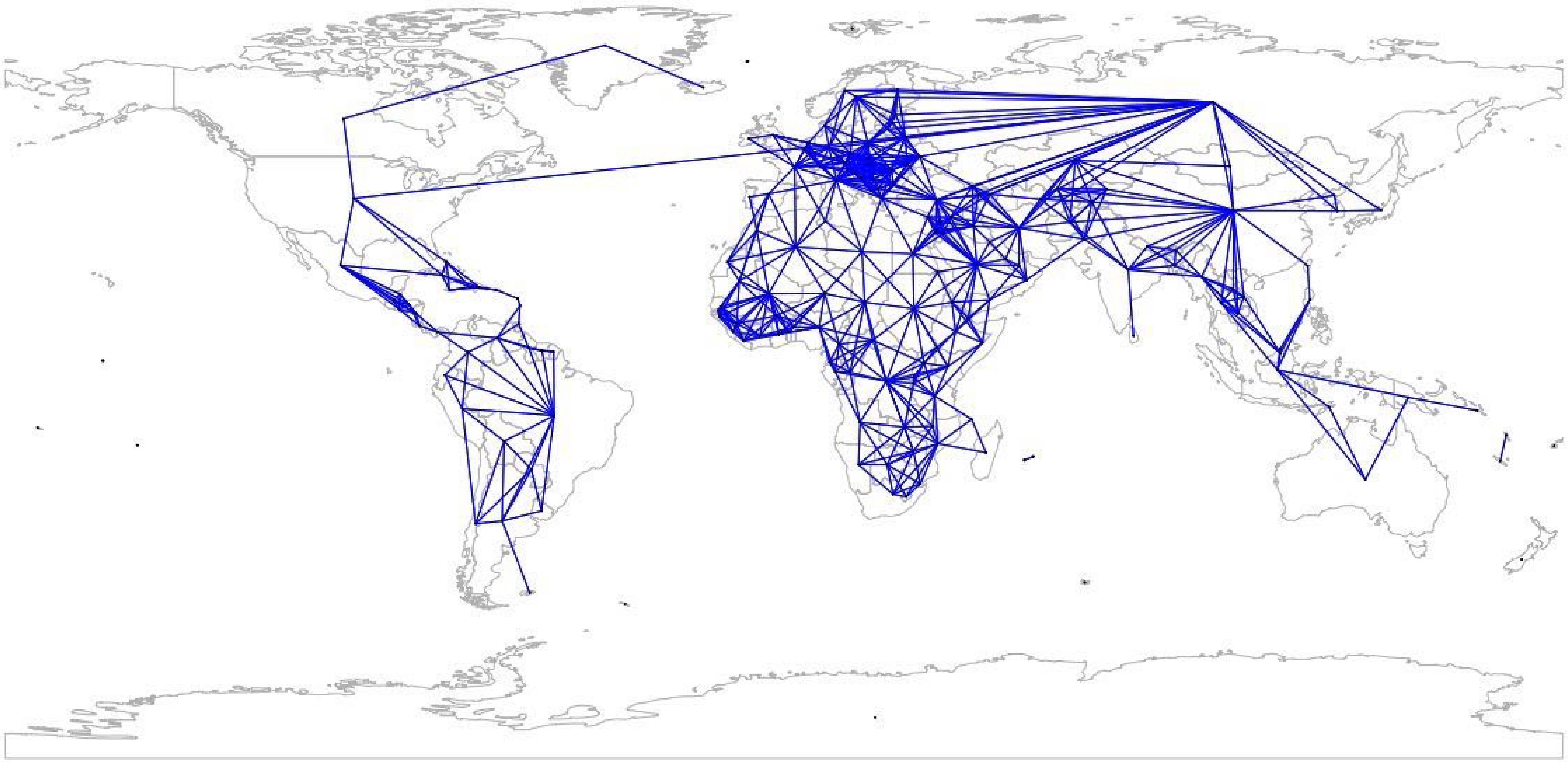
spatial autocorrelation: the problem of measuring proximity

Contiguity
Polygons



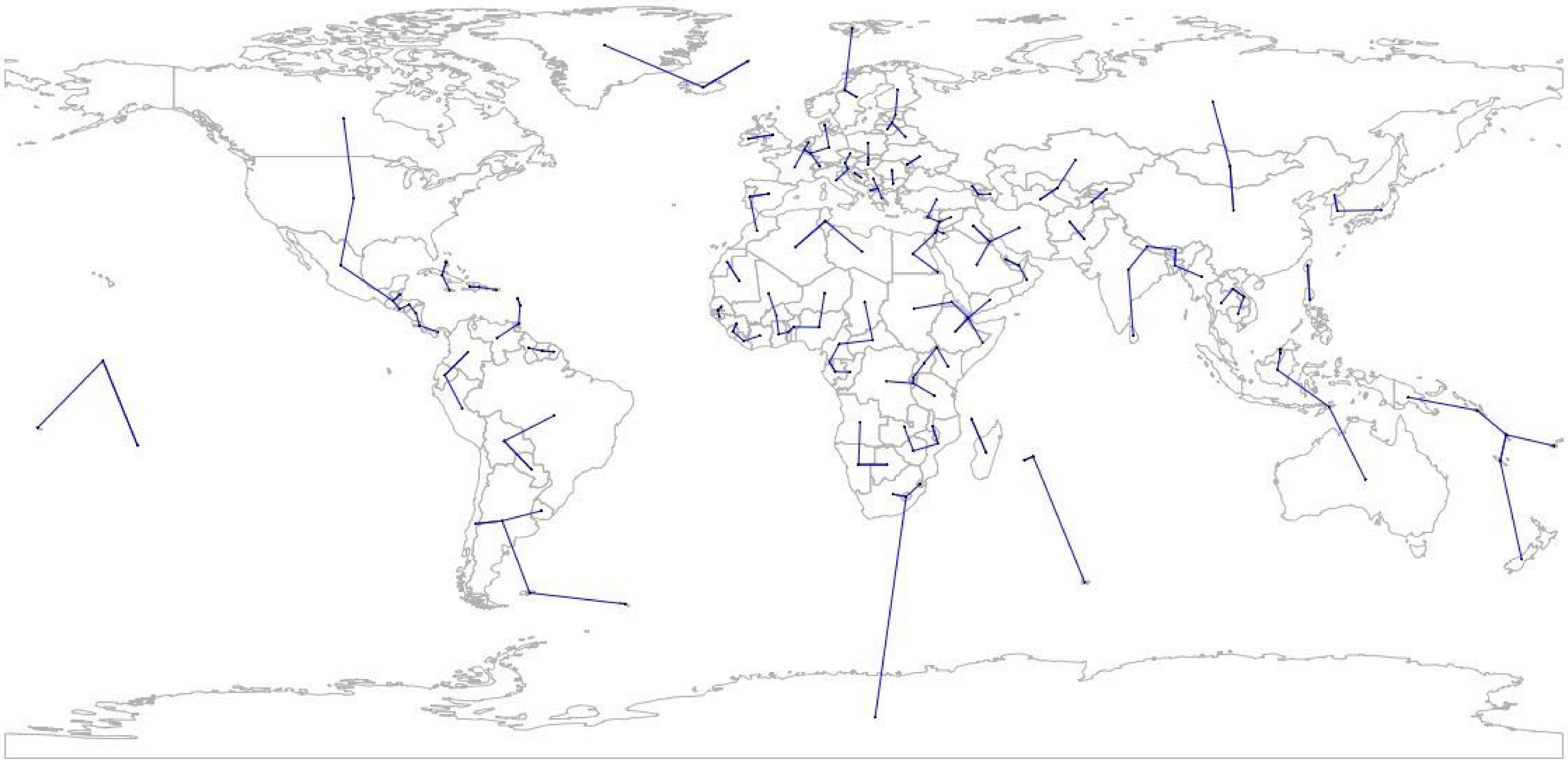
spatial autocorrelation:
the problem of measuring proximity

Contiguity + Distance
Polygons



spatial autocorrelation: the problem of measuring proximity

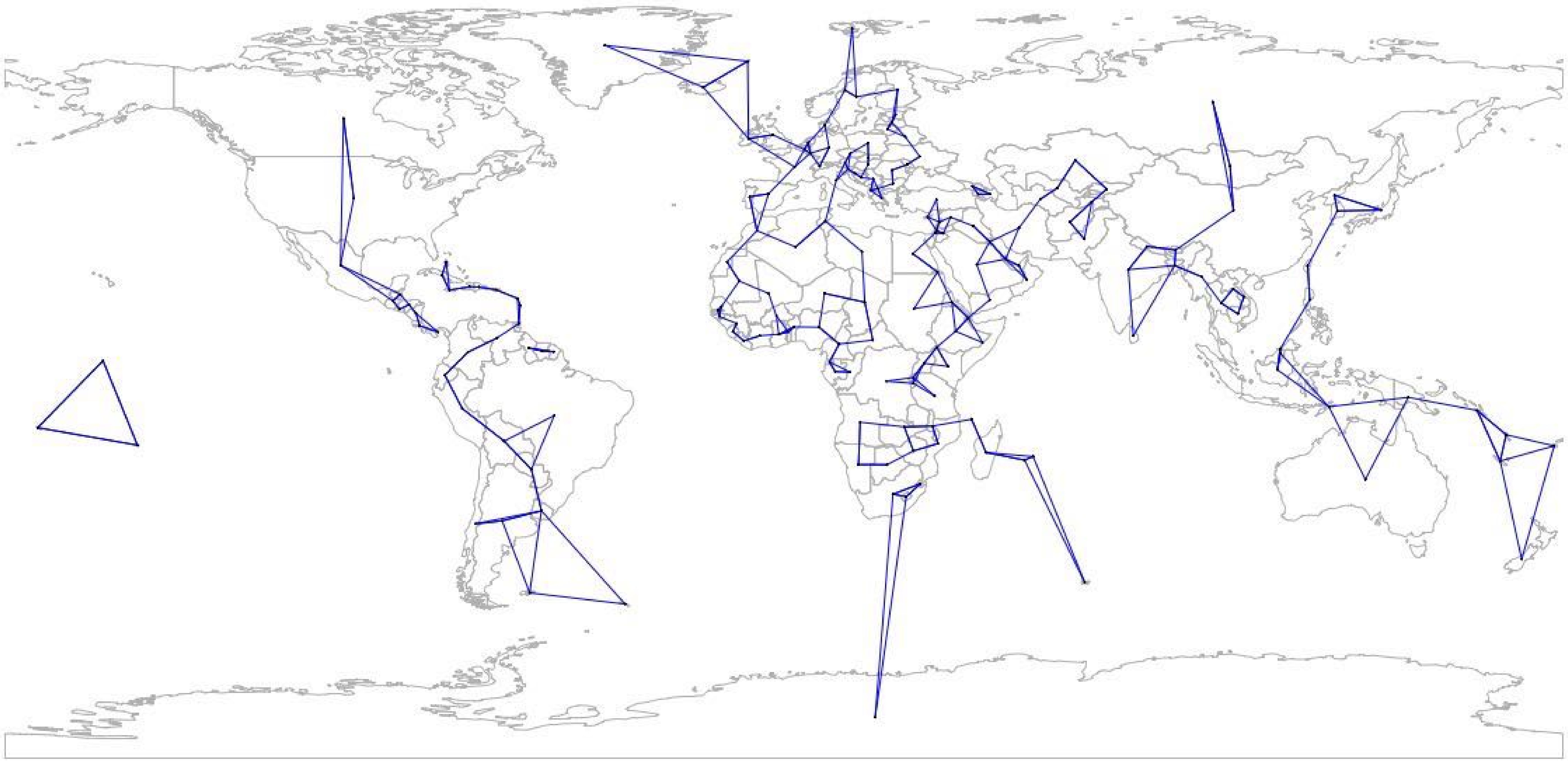
K-Nearest Neighbors
N=1



spatial autocorrelation:
the problem of measuring proximity

K-Nearest-Neighbors

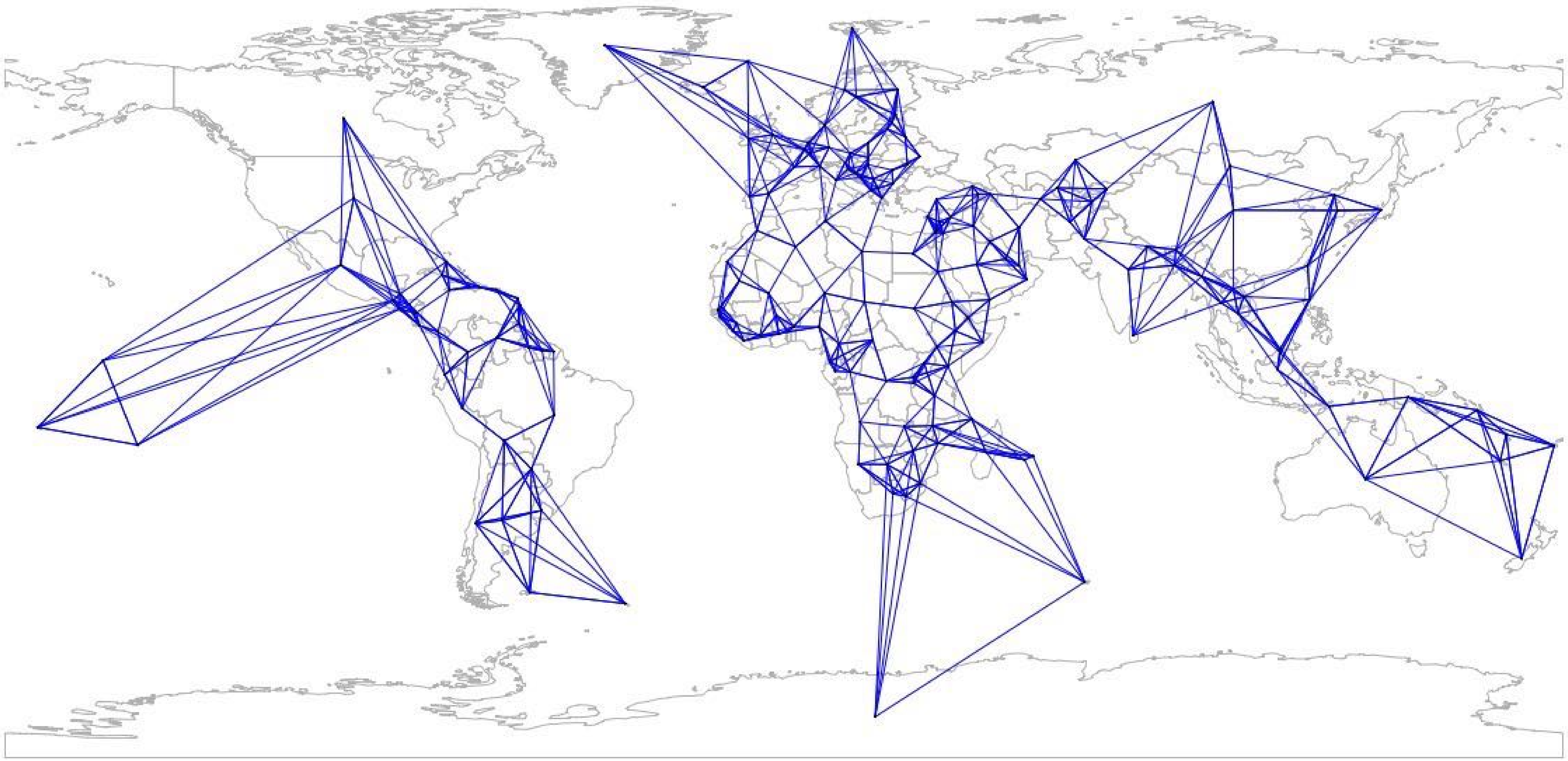
N=2



spatial autocorrelation: the problem of measuring proximity

K-Nearest-Neighbors

N=5



spatial autocorrelation:
spatial weights matrix

$$
W = \begin{array}{|c|c|c|c|c|c|c|}
\hline
 & & & w_{0j} & & & \\
\hline
 & & & w_{1j} & & & \\
\hline
 & & & \dots & & & \\
\hline
w_{i0} & w_{i1} & \dots & w_{ij} & \dots & \dots & w_{in} \\
\hline
 & & & \dots & & & \\
\hline
 & & & \dots & & & \\
\hline
 & & & w_{nj} & & & \\
\hline
\end{array}
$$

- **Different methods of calculating w_{ij} can result in different values for autocorrelation and different conclusions from statistical significance tests!**
- **Often we use row standardized weights ($\sum_i \sum_j w_{ij} = 1$)**
- **Problematic situations for irregular polygons**

global measures of spatial autocorrelation

Global Measures and Local Measures

- **Global Measures**
  - A single value which applies to the entire data set
    - The same pattern or process occurs over the entire geographic area
    - An average for the entire area
- **Local Measures**
  - A value calculated for each observation unit
    - Different patterns or processes may occur in different parts of the region
    - A unique number for each location

Joins Count Statistic

- **Polygons only**
- **binary (1,0) data only**
    - Polygon has or does not have a characteristic
    - For example, a candidate won or lost an election
- **Based on examining polygons which share a border**
    - Do they have the same characteristic or not?
- **Requires a contiguity matrix for polygons**
- **Measures the number of borders (“joins”) of each type (1,1), (0,0), (1,0 or 0,1) relative to total number of borders**

Moran's I statistic

- **The most common measure of spatial autocorrelation**
- **Use for points or polygons**
    - Join Count statistic only for polygons
- **Use for a continuous variable (any value)**
    - Join Count statistic only for binary variable (1,0)

Moran's I statistic: formulation

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

N # observations

x_i variable of interest at i

$\bar{x}$ mean of the variable of interest

If W_{ij} is row standardized

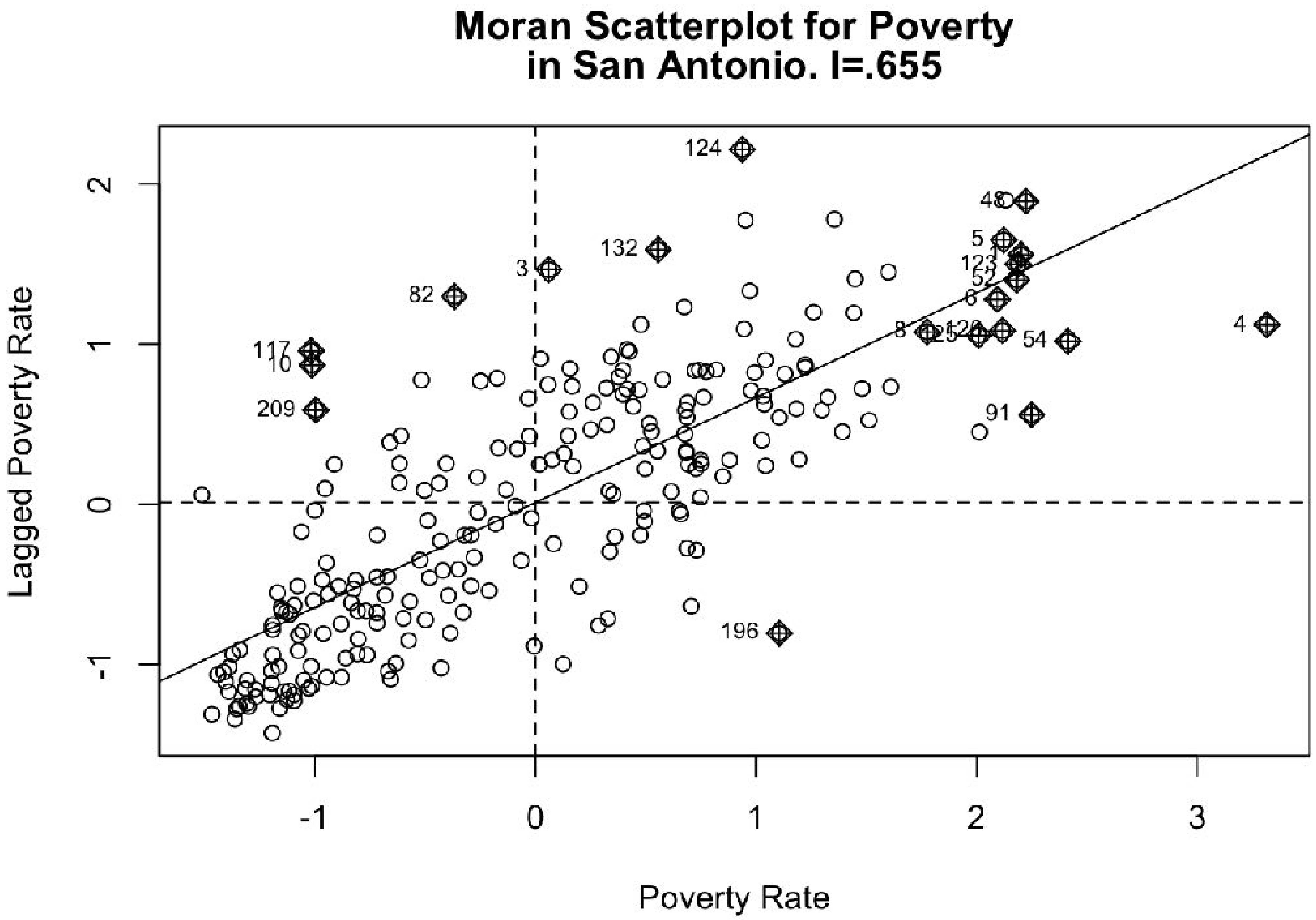
$$\sum_i \sum_j w_{ij} = N$$

then

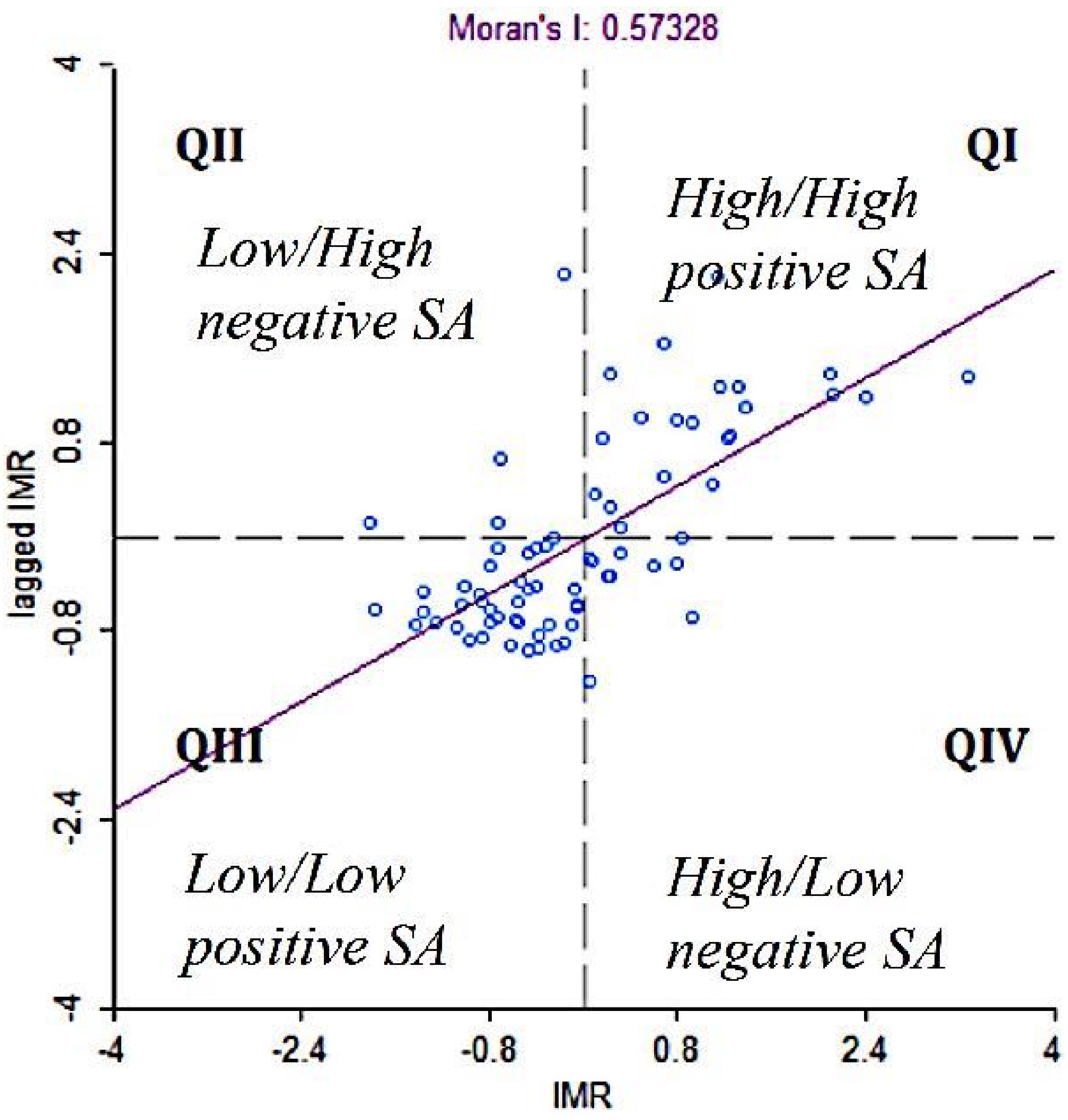
$$I = \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Need adjustment for short or zero distances

Moran Scatterplot



Moran Scatterplot
bivariate



Moran's I: 0.57328
QII
Low/High
negative SA
QI
High/High
positive SA
QIII
Low/Low
positive SA
QIV
High/Low
negative SA
lagged IMR
IMR
4
2.4
0.8
-0.8
-2.4
-4
-4
-2.4
-0.8
0.8
2.4
4

Statistical significant tests for Moran's I

- **How to assess whether computed value of Moran's I is significantly different from a spatially random distribution?**
- **Analytically**

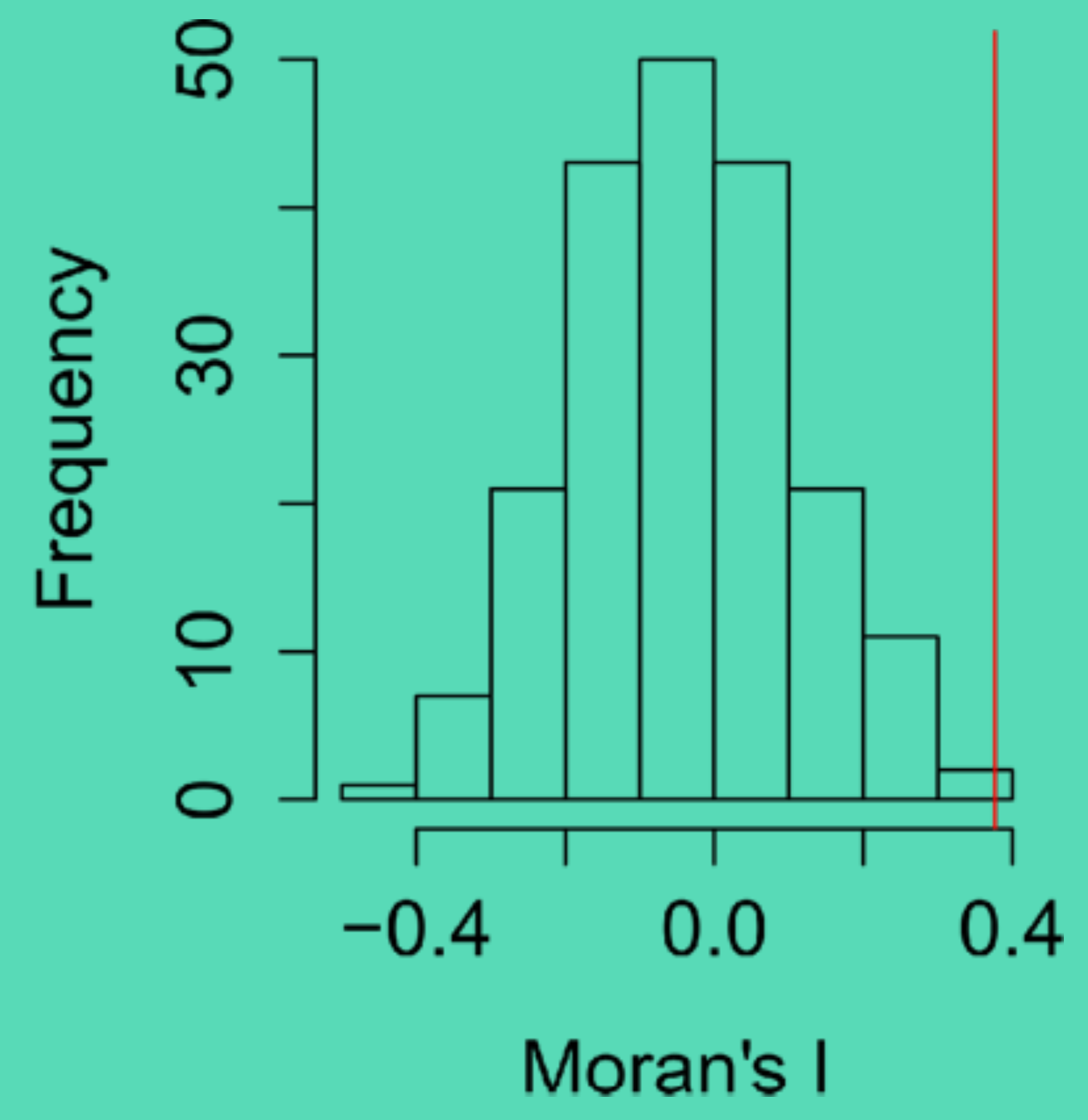
$$E[I] = \frac{-1}{N-1} \qquad V[I] = E[I^2] - E[I]^2 \qquad z_I = \frac{I - E[I]}{\sqrt{V[I]}}$$

In many cases is difficult!

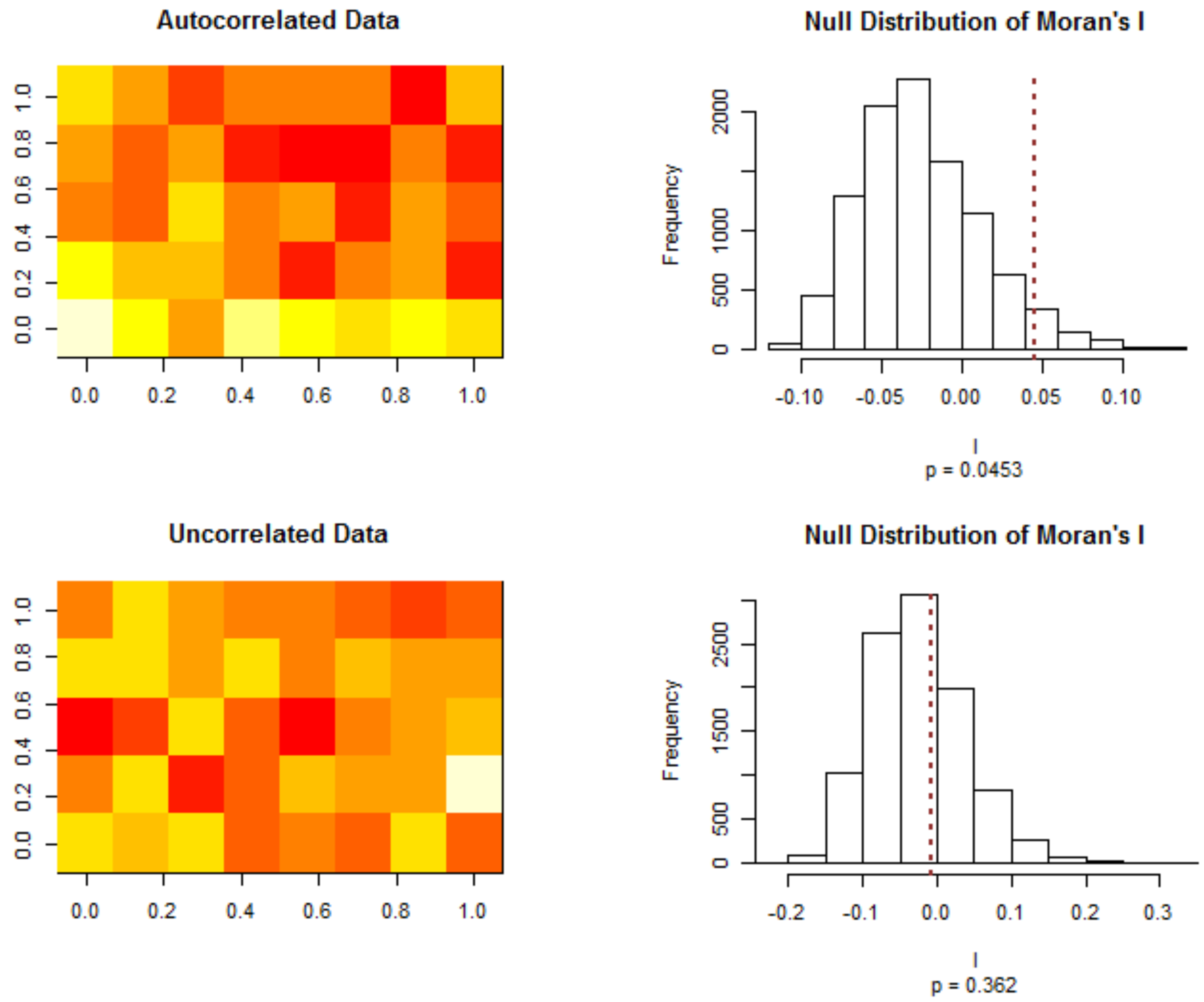
- **Computationally**
  - Randomly reshuffling observations and recomputing Moran's I each time (building an empirical **reference distribution**)
  - Compare observed value to the reference distribution

Montecarlo test
permutation bootstrap test

- pseudo p-value $$\frac{N_{extreme} + 1}{N + 1}$$
- $N_{extreme}$ is the number of simulated Moran's I values more extreme than our observed statistic and N is the total number of simulations
- This is interpreted as "*there is a 1% probability that we would be wrong in rejecting the null hypothesis H_o*"



example


Autocorrelated Data
0.0
0.2
0.4
0.6
0.8
1.0
Null Distribution of Moran's I
Frequency
0
500
1000
2000
-0.10
-0.05
0.00
0.05
0.10
I
p = 0.0453
Uncorrelated Data
Null Distribution of Moran's I
Frequency
0
500
1500
2500
-0.2
-0.1
0.0
0.1
0.2
0.3
I
p = 0.362

Moran's I:
tips

- **Raw Moran's I are not comparable across variables and spatial weights**
  - Use standardized z-values instead
- **Other measures**
  - Geary's C
    - inversely related to Moran's I
    - more sensitive to local spatial autocorrelation
  - High/Low Clustering (Getis-Ord General G)
    - detect clusters of **low** or **high** values **(hot/cold spots)**

local measures of spatial autocorrelation

LISA
Local Indicators of Spatial Association (Anselin 1995)

- **The statistic is calculated for each areal unit**
- **For each polygon, the index is calculated based on neighbouring polygons with which it shares a border**
- **Can be mapped to indicate how spatial autocorrelation varies over the study region**
- **Each index has an associated test statistic**
- **Local version of**
    - Moran’s I
    - Geary’s C
    - Getis-Ord G

Calculating LISA

The local Moran statistic for areal unit **i** is:

$$I_i = z_i \sum_j w_{ij} z_j$$

where **z_i** is the original variable **x_i** in standardized form

$$z_i = \frac{x_i - \bar{x}}{SD_x}$$

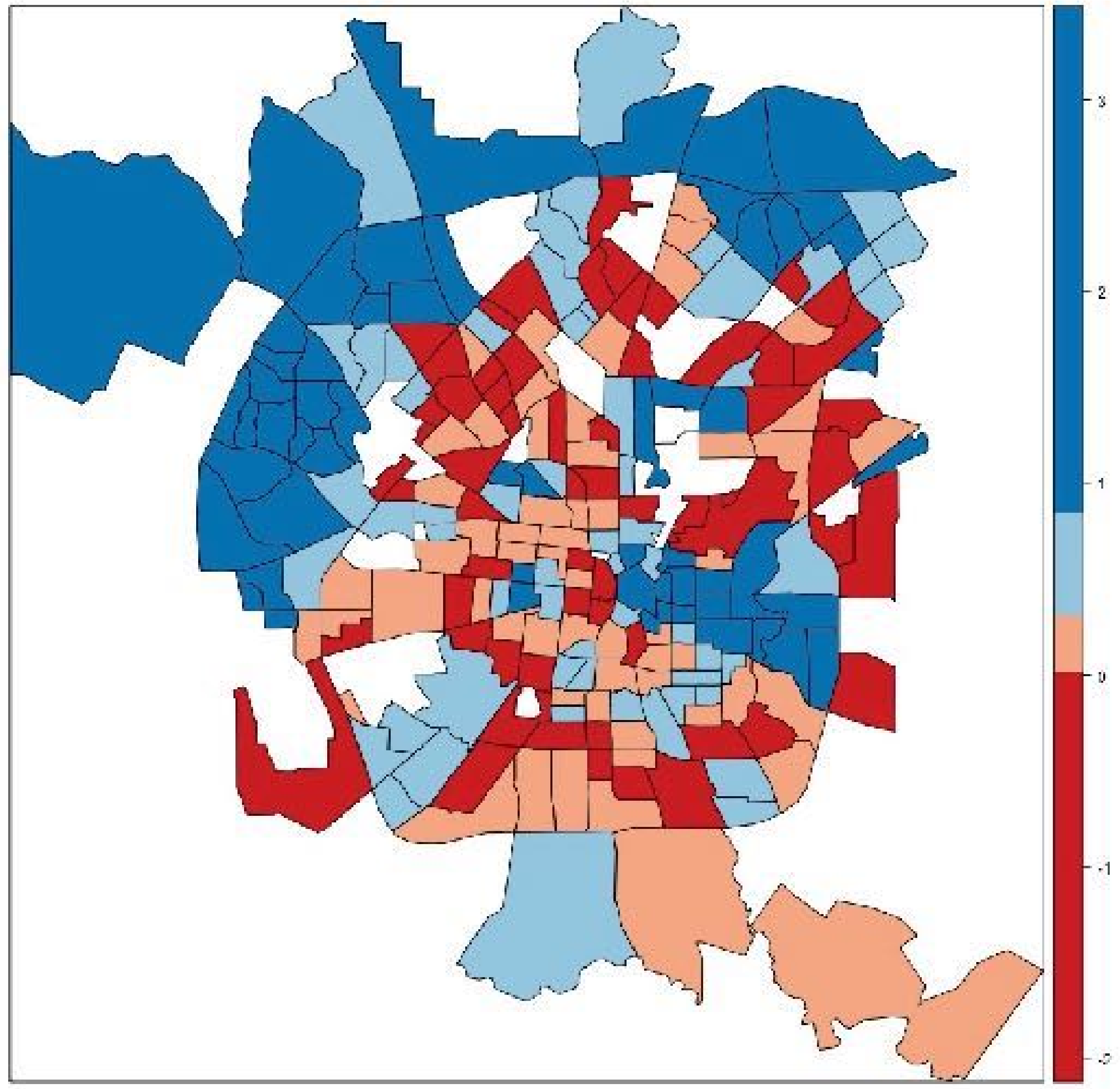
or it can be in deviation form

$$x_i - \bar{x}$$

and **w_{ij}** is the spatial weight

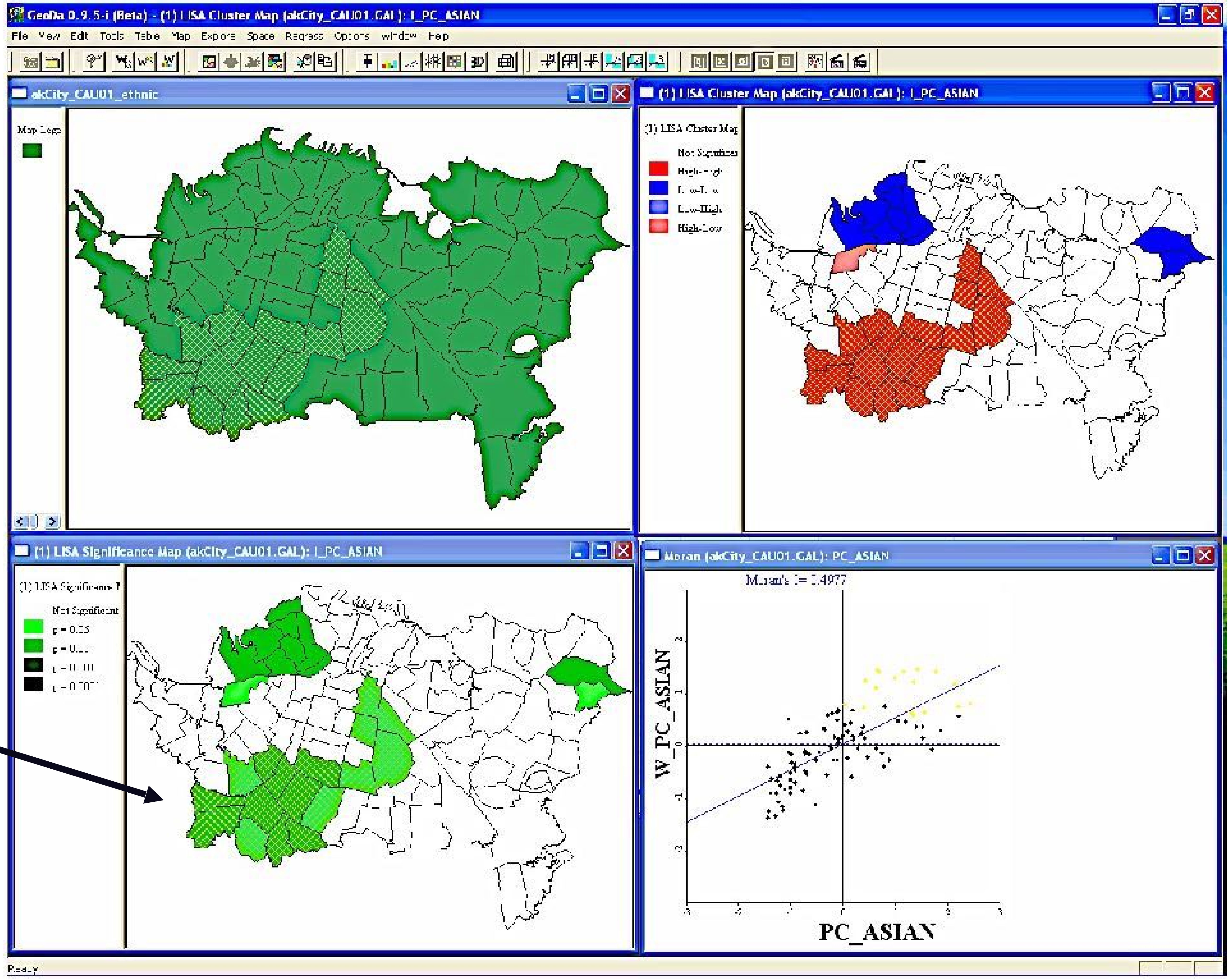
San Antonio poverty rates

Local Moran's I



Local Moran Clusters (red)

example
percentage of Asian people in Auckland City (geoDa)



LISA Significance Map
map the statistical significance level and use it as a measure of the **strength** of the spatial autocorrelation

@rschifan

schifane@di.unito.it

http://www.di.unito.it/~schifane