2
Spatial Statistics:
Introduction, spatial weights, clustering

GEOG-325: Applied Spatial Statistics and Urban Modelling

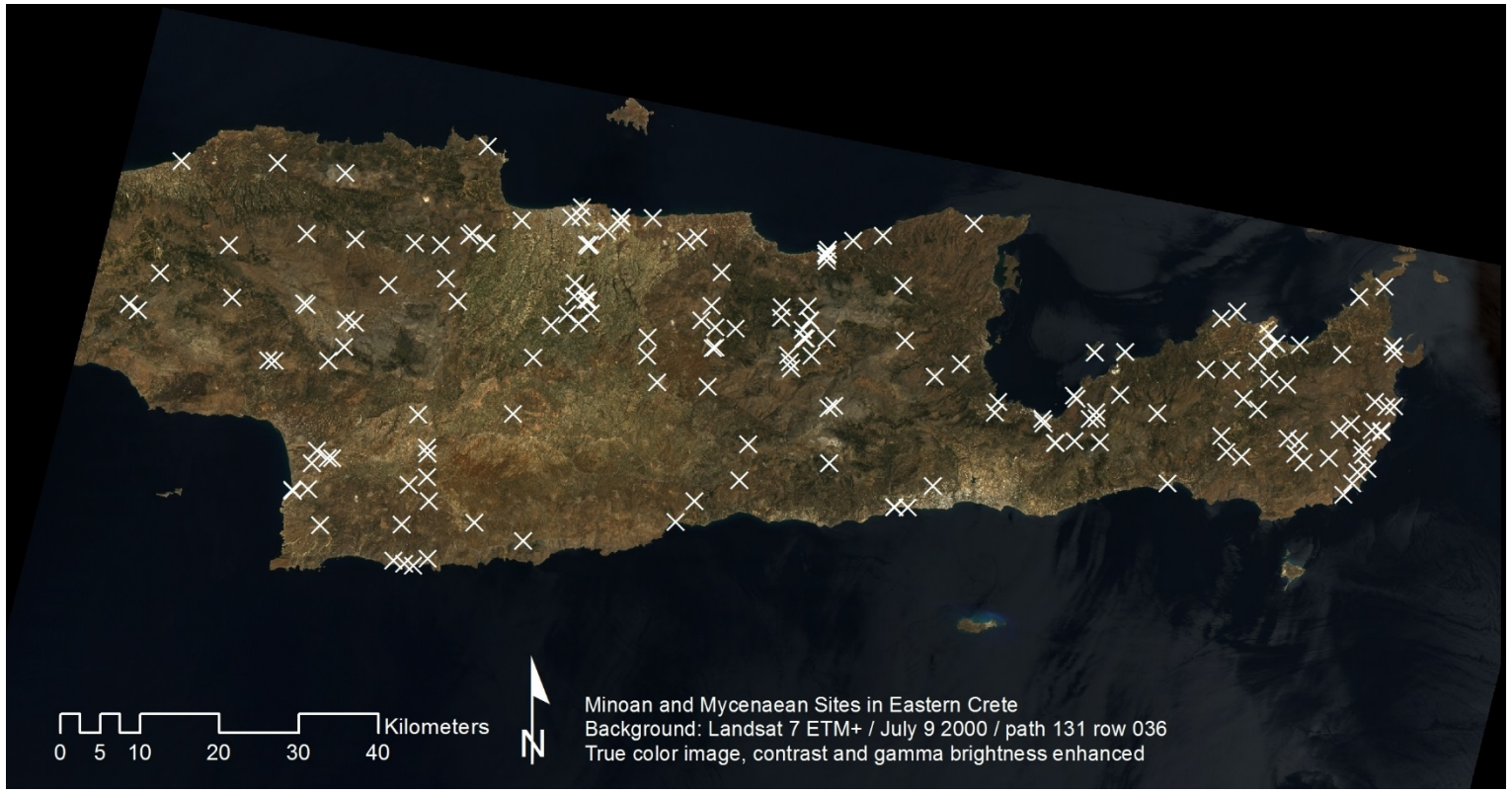
Athanasios Votsis

I

Spatial analysis

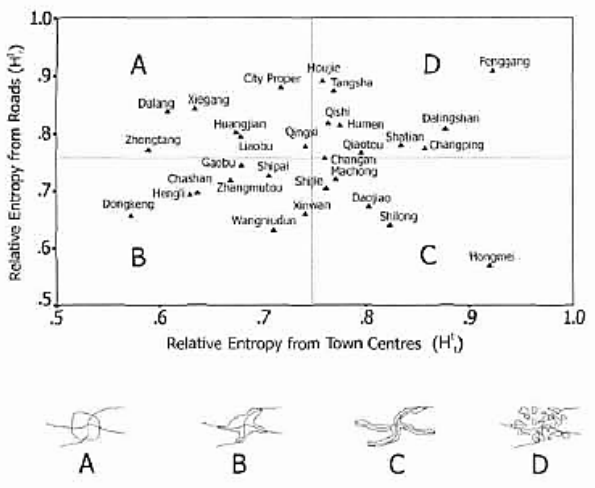
Spatial phenomena (1/4)

Consider the following image of a system of prehistoric settlements on a landscape:



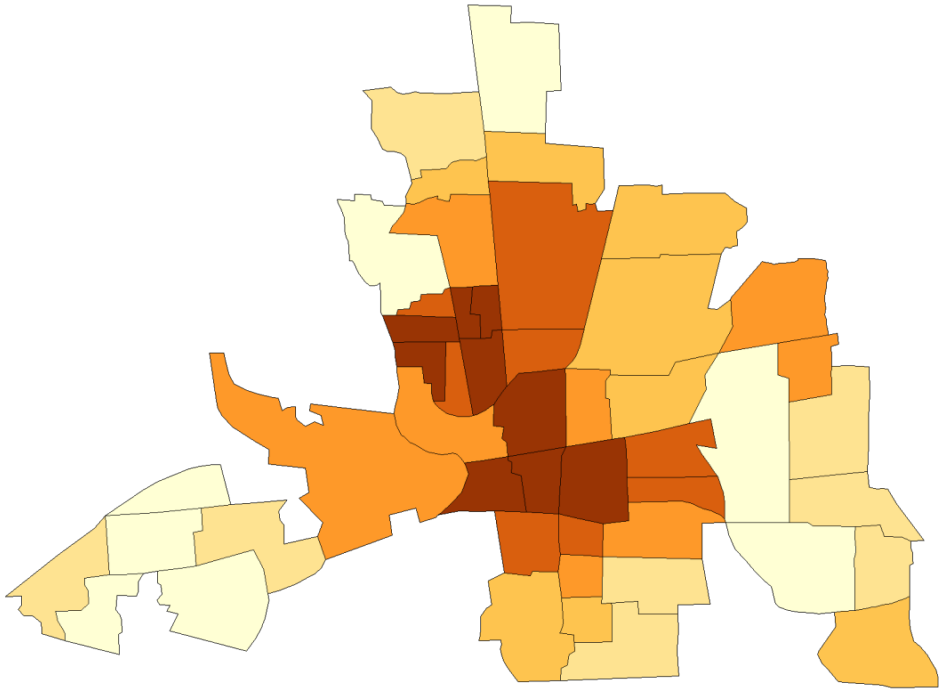
Or the architectural typology of spaces, spatial relationships between different types, formal organization:

Or the analysis of urban dynamics and urban growth:

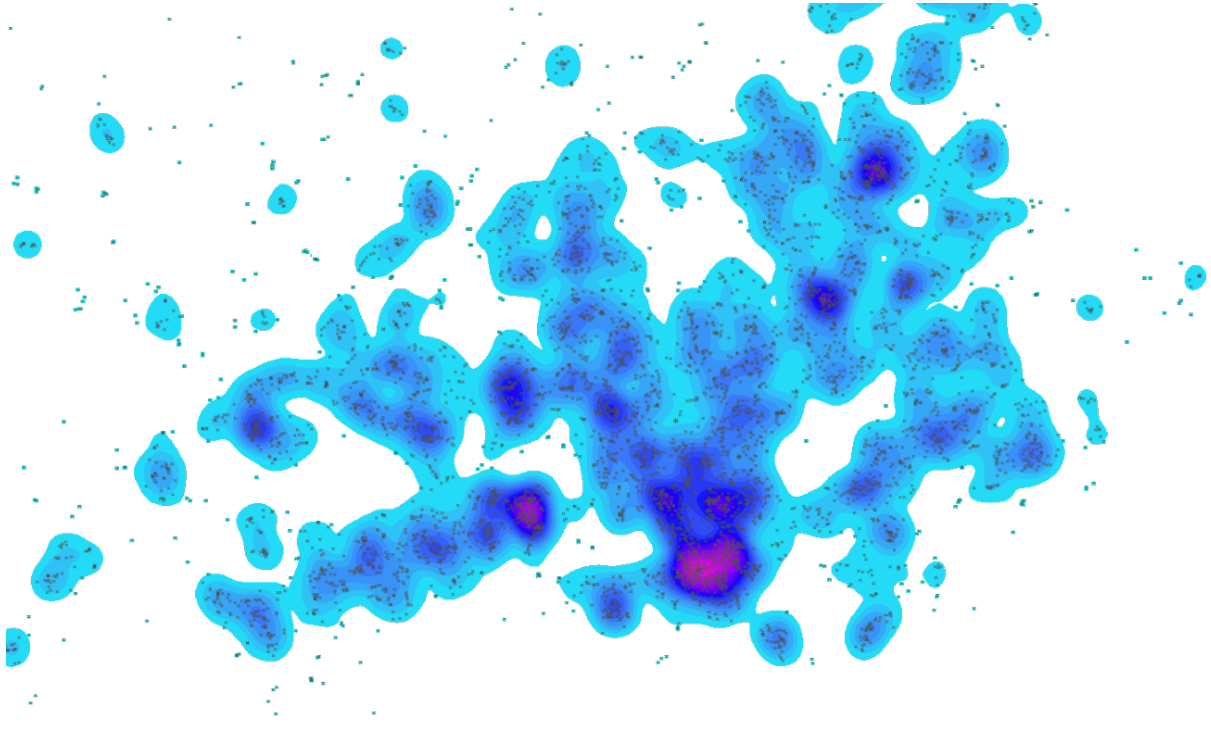


Spatial phenomena (2/4)

Or the distribution of crime in Columbus, Ohio:



Or the density of commercial buildings in Helsinki:



Spatial phenomena (3/4)

Task	Data	Example
Finding values for the locations we do not observe, based on locations we have observed	Surfaces	Temperature Earthquake acceleration
Investigating the complete distribution of values across a geographical domain	Lattices	Flood damage per district Housing prices per district
Investigating the occurrence of events in space	Events	Crime incidents Earthquakes

(adapted from Anselin 2011)

- In everyday practice, a combination of the three categories is the reality:
  - Combination of data;
  - Different ways to conceptualize the same phenomenon.

Spatial phenomena (4/4)

Several questions arise:

- Are some objects arranged in a certain way in space?
  - If so, why?
- Are some values dependent on location?
  - If so, what conditions/locations/attributes bring about such dependence?
- What is the behavior of a phenomenon in unobserved locations?
  - Now, or in the future?

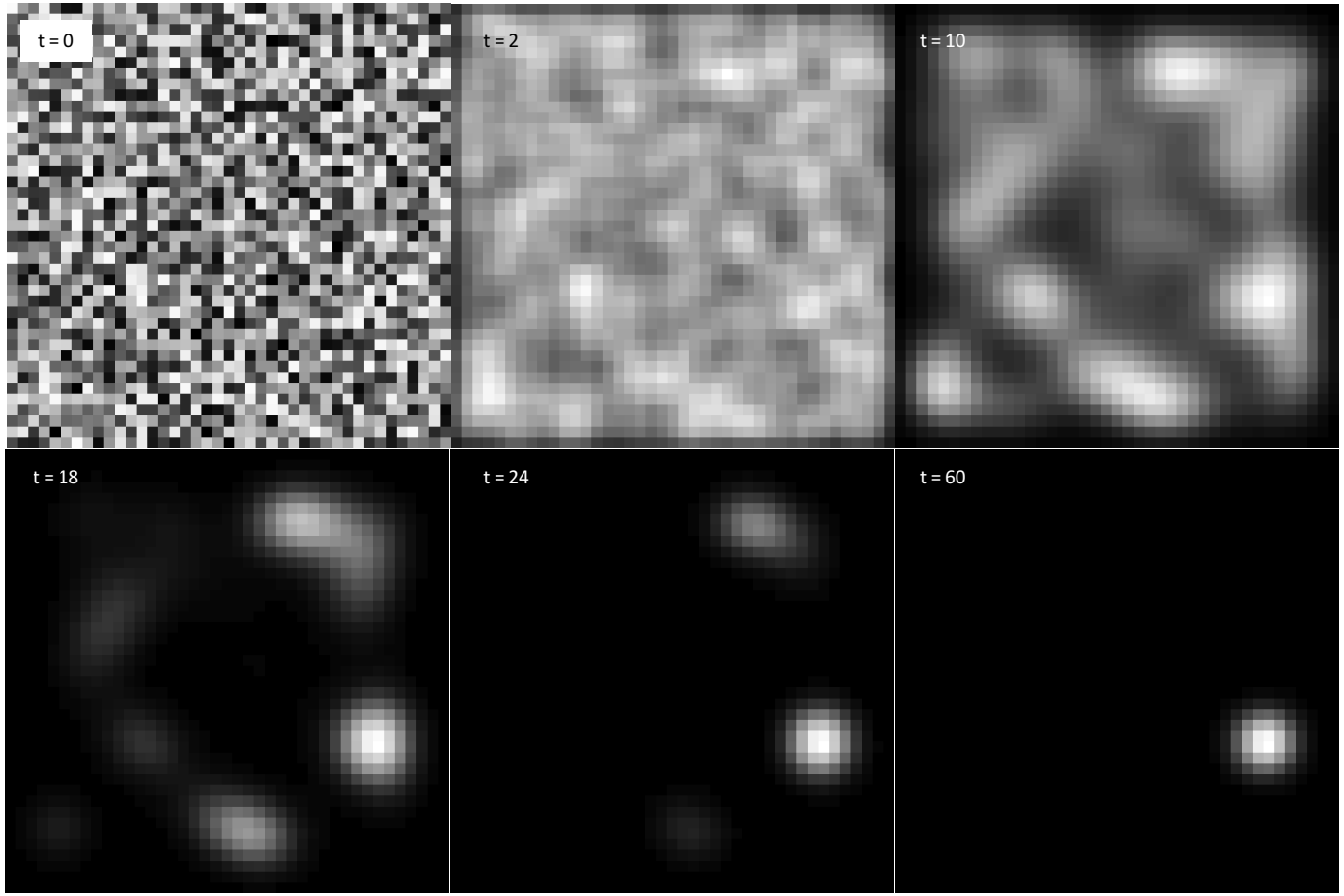
In other words, we see that **some phenomena are inherently spatial** (theoretical point of view) **or exhibit spatial behavior when measured** (empirical point of view).

A major implication

Disciplines that deal with human-made or natural systems are oftentimes confronted by research questions where **space cannot be “assumed irrelevant**”.

Illustration

- Consider a spatial averaging process.
- Explained by the value of each cell in time *t*, its geographical location, and type of spatial averaging.
- If no spatial information is available, then all kinds of erroneous theories might be developed to explain the consequence rather than the cause.



Approaching spatial questions: standard practice

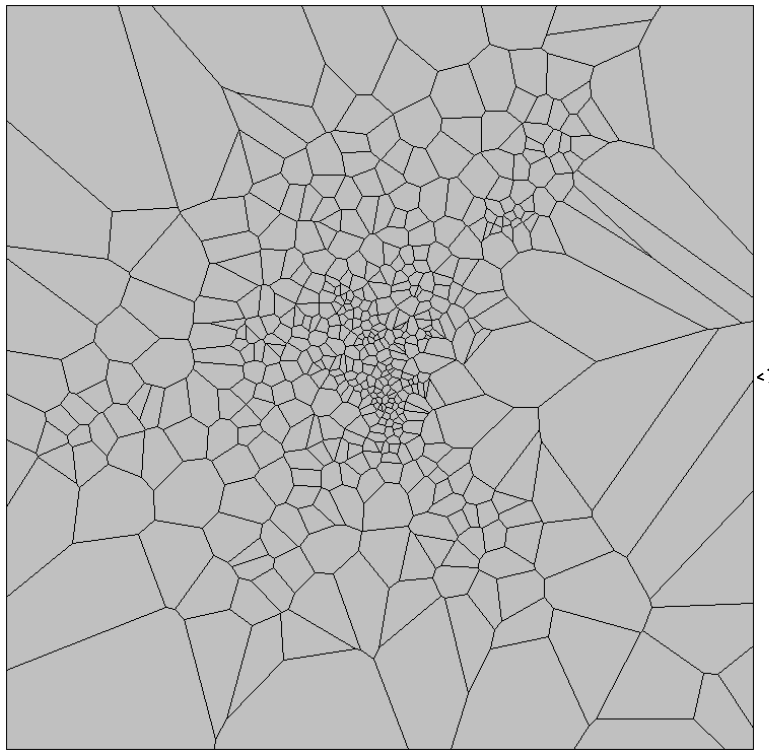
Two widely used ways to approach the spatial behavior of phenomena have been:

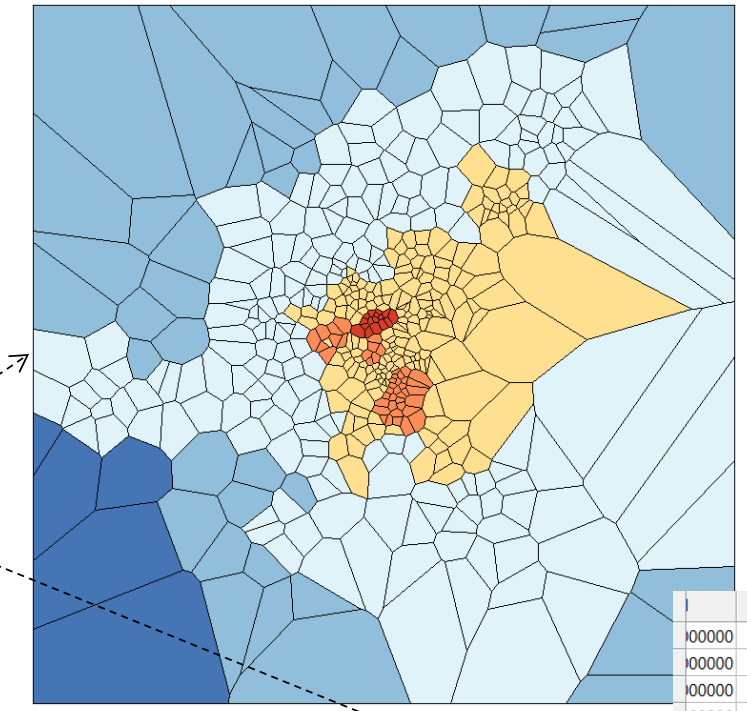
- Load-up the data in a GIS, calculate metrics, export to statistical software, analyze;
- Load-up the data in a GIS, make visualizations, make conclusions.

In these approaches:

- Space is assumed out of the picture and insufficiently formalized;
- Large bits of information about spatial behavior are left unexploited;
- Conclusions about spatial behavior cannot be formally verified or even communicated.

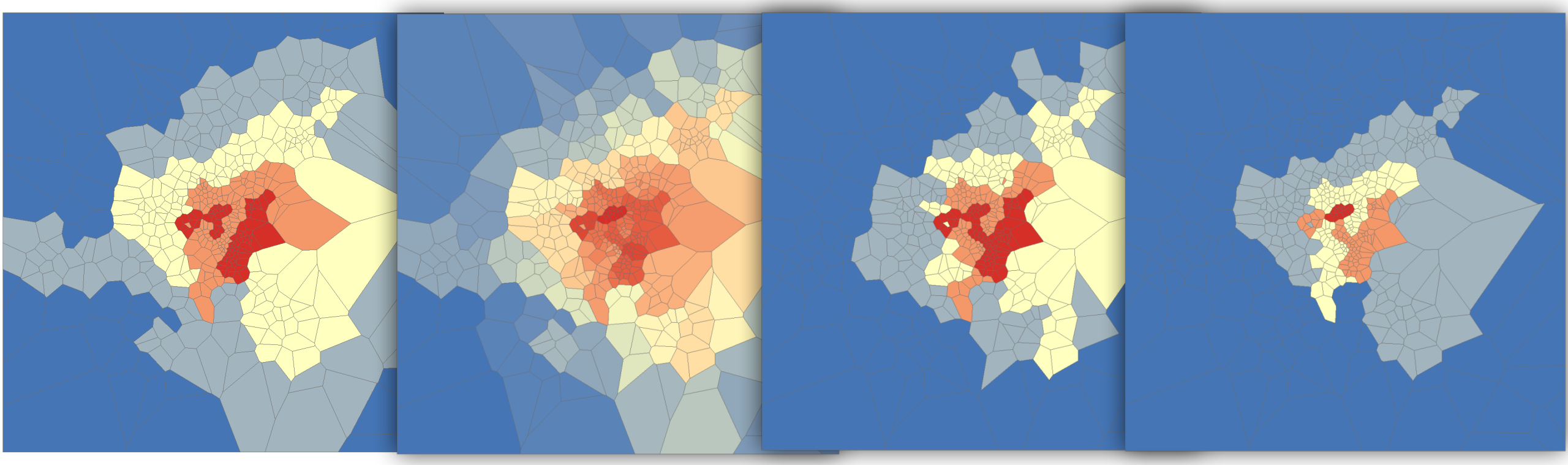
Example of standard practice





	INDUS	CHAS	NOX	RM	AGE	DI
00000	2.310000	0.000000	0.538000	6.575000	65.200000	4.
00000	7.070000	0.000000	0.469000	6.421000	78.900000	4.
00000	7.070000	0.000000	0.469000	7.185000	61.100000	4.
00000	2.180000	0.000000	0.458000	6.998000	45.800000	6.
00000	2.180000	0.000000	0.458000	7.147000	54.200000	6.
00000	2.180000	0.000000	0.458000	6.430000	58.700000	6.
00000	7.870000	0.000000	0.524000	6.012000	66.600000	5.
00000	7.870000	0.000000	0.524000	6.172000	96.100000	5.
00000	7.870000	0.000000	0.524000	5.631000	100.000000	6.
00000	7.870000	0.000000	0.524000	6.004000	85.900000	6.
00000	7.870000	0.000000	0.524000	6.377000	94.300000	6.
00000	7.870000	0.000000	0.524000	6.009000	82.900000	6.
00000	7.870000	0.000000	0.524000	5.889000	39.000000	5.
00000	8.140000	0.000000	0.538000	5.949000	61.800000	4.
00000	8.140000	0.000000	0.538000	6.096000	84.500000	4.
00000	8.140000	0.000000	0.538000	5.834000	56.500000	4.
00000	8.140000	0.000000	0.538000	5.935000	29.300000	4.

Ambiguities of standard GIS approaches



Approaching spatial phenomena: another way

Formalize/operationalize (mathematically) “space” and make it part of your analysis.

Everything else remains the same, but with the formalization of space the benefits of statistics come to the analysis:

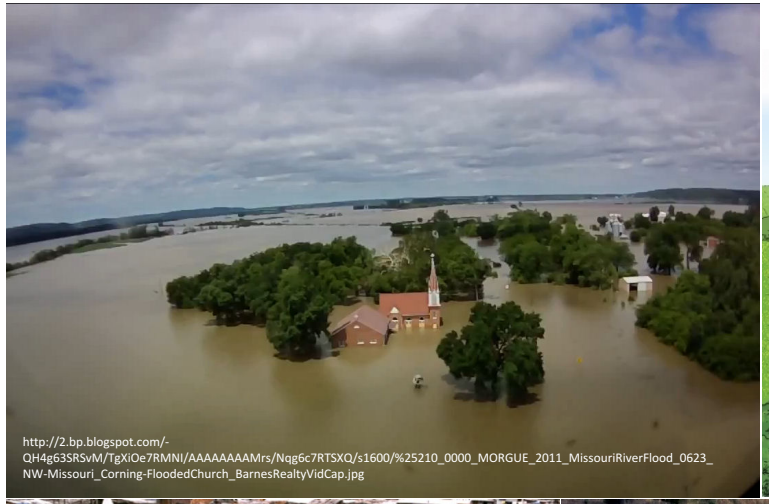
- Distinguish effects in a more detailed way: spatial **and** non-spatial;
- Be able to statistically assess the results we are getting.

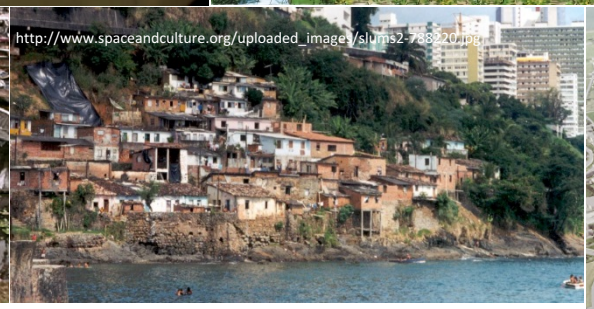
Spatial statistics: normal statistics, but with:

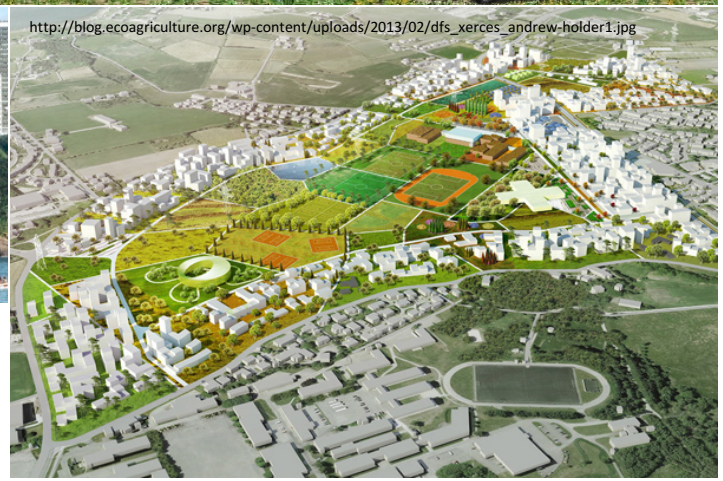
- formalized space;
- additional (statistical) tests;
- slightly modified models (“equations”);
- some additional data needs.

Why would we want to do this?

Complex challenges and complex visions require more than causes-effects







II

The first law of geography and spatial interaction:
Spatial dependence, spatial heterogeneity

1st law of geography

- In his 1970 paper "A computer movie simulating urban growth in the Detroit region" (included in your Literature folder), geographer Waldo Tobler set forth what he called the first law of geography:

- "[…] everything is related to everything else, but near things are more related than distant things." (Tobler 1970: 236).

- It is worth noting that the context of that paper was urban modelling and simulation (spatial growth mechanisms), which has a different mentality and scope than probabilistic statistics.

Key concepts

- In the empirical context, the 1st law of geography implies that an **underlying spatial process** causes a dependent variable to exhibit **spatial effects**.

- You can separate spatial effects into **spatial dependence** (spatial autocorrelation, clustering of values, spatial interaction) and **spatial heterogeneity** (the relationship between the dependent and independent variables is not constant in space; distinct data subgroups).

1. Spatial dependence is dealt with via clustering detection algorithms, and via spatial lag and error regression models
2. Spatial heterogeneity is dealt with by sub-setting the data, or by geographically weighted regression.

Dataset preparation

- Your observations need to be **georeferenced** (ideally in a projected coordinate system);
- Then, you need to model who is the neighbor of whom in your observations, how far they are apart, and what is understood as "**neighbors**" and as "**distance**";
- Based on your decisions on neighbors, a **spatial weights** file and subsequently matrix is constructed. It formalizes the spatial relationships of the observations;
- Recommended: have your data in a GIS system, convert to text or other file formats on demand.

Standard procedure of analysis

Create spatial weights
(technical requirement)

▼

Test for spatial effects
(exploratory analysis)

▼

Estimate the effects
(formal analysis)

▼

Simulate the effects
(what-if analysis)

All steps involve interpretation, but each step has specific roles and capabilities in your analysis.

Software

- Typical software includes:
  - **ArcGIS** or **QGIS** or **MapInfo**: for data management, descriptive tasks, visualization, meta-analysis
  - **GeoDa**: for spatial weights creation, simple analysis
  - **R** and **R-Studio**: for more advanced or flexible analysis

- GeoDa and R are both free and very good programs. Together with any GIS software that can save data as shapefiles (ESRI's format), they are sufficient for spatial econometrics.

- More recently, other econometric software have been including spatial algorithms: e.g. STATA.

III

Spatial weights matrices

Workflow

Create spatial weights

▼

Test for spatial effects

▼

Estimate the effects

▼

Simulate the effects

Similar to “setting time” in time series analysis.

Here, we “set space”.

Spatial connectivity & weights matrices

$$\mathbf{W}_{i\times j} = \begin{bmatrix} w_{11} & \cdots & w_{1j} \\ \vdots & \ddots & \vdots \\ w_{i1} & \cdots & w_{ij} \end{bmatrix}$$

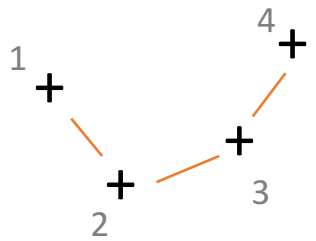
- In order to investigate spatial interaction, one needs to compute *which* observations are the neighbors of a given observation (implying that one needs to decide *what* the criteria for being a neighbor are).

- This purpose serves the **spatial weights matrix W**.
- **W** is an *i*j* matrix that formalizes the neighborhood structure of geographical entities.
- Usually a square matrix: its elements record whether observation *i* is a neighbor of observation *j*; $w_{ij} \neq 0$ for neighbors and 0 otherwise.
- **W** is utilized in several stages of spatial statistical tests and regressions.

- However, since there are several options to go about when computing a spatial weights matrix, it is more appropriate to say that the matrix ***imposes*** a neighborhood structure on the observations.

Examples of spatial weight matrices

Consider 4 observations:



row-unstandardized **W**:

$$\mathbf{W}_{4\times4} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

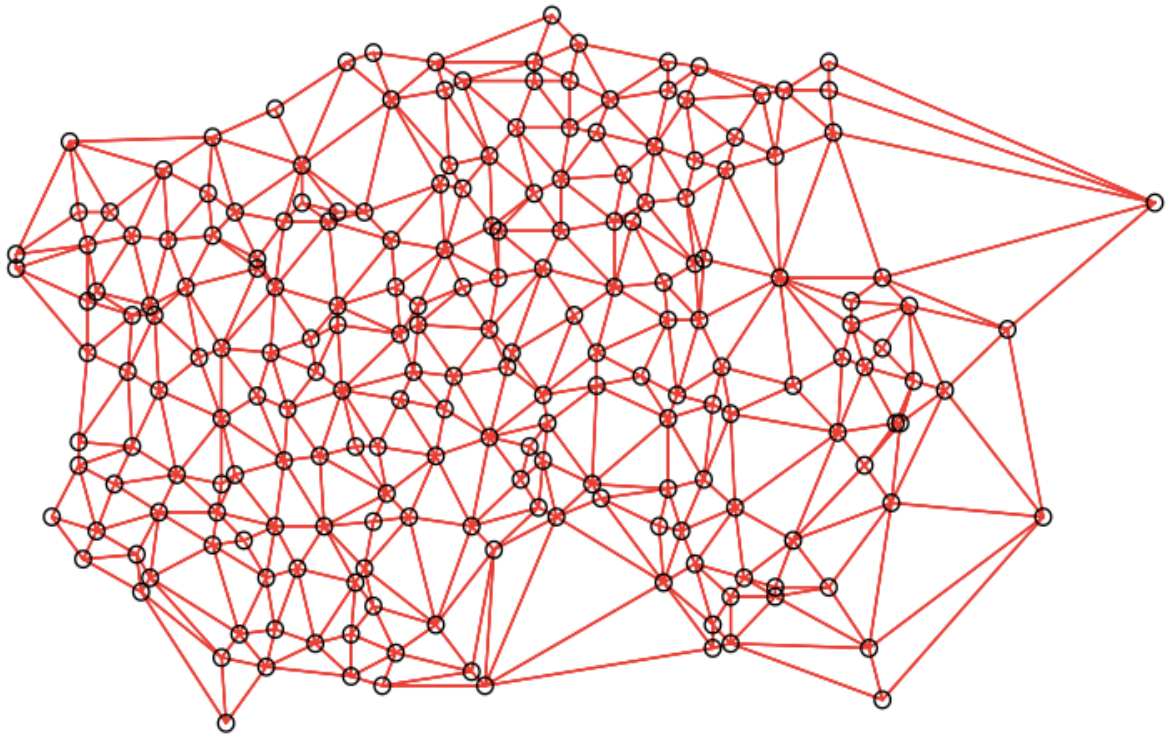
row-standardized **W** (for easier interpretation of lagged variables as spatial averages):

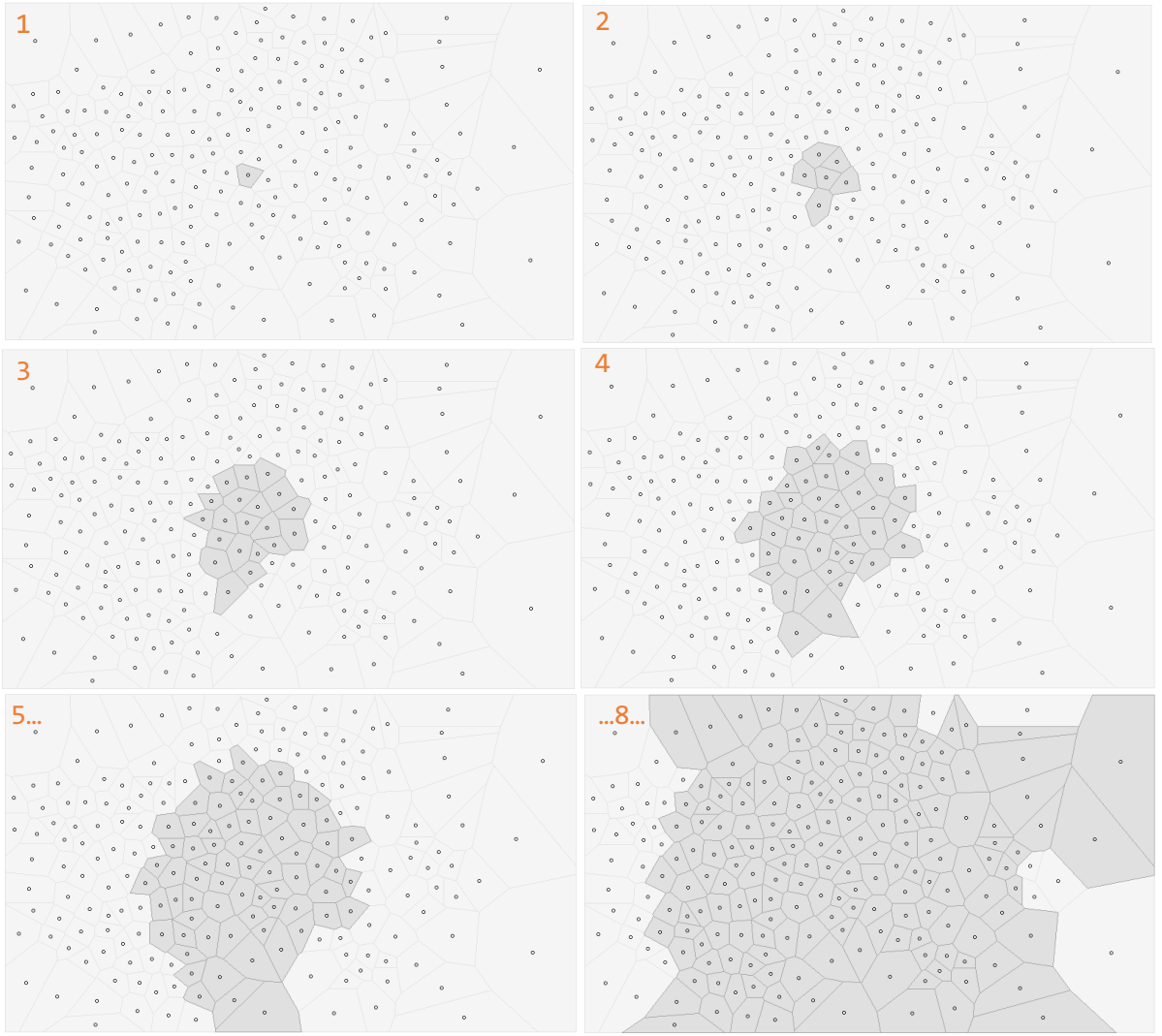
$$\mathbf{W}_{4\times4} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{matrix} = 1 \\ = 1 \\ = 1 \\ = 1 \end{matrix}$$

Imposed assumption of stronger spatial influence as neighbors become fewer

symmetric: $\mathbf{W} = \mathbf{W}^{\mathrm{T}}$; asymmetric: $\mathbf{W} \neq \mathbf{W}^{\mathrm{T}}$ (reciprocal neighboring is not honored)

Spatial weights: connectivity, spreading





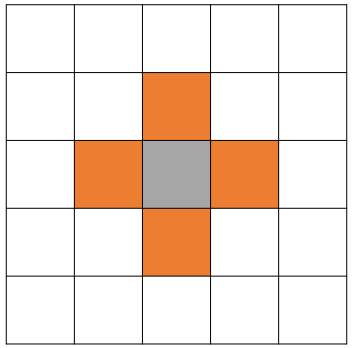
Types of neighbor definitions

There are two main ways for defining neighbors

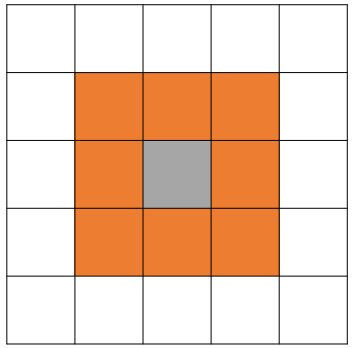
1. By **contiguity**:
   a. **Rook contiguity** (formally *von Neumann neighborhood*): common polygon edges
   b. **Queen contiguity** (formally *Moore neighborhood*): common polygon edges and vertices

2. By distance
   a. **K nearest neighbors**: the k (k = 1, 2, 3, …) closest entities
   b. **Distance bands**: entities within x meters

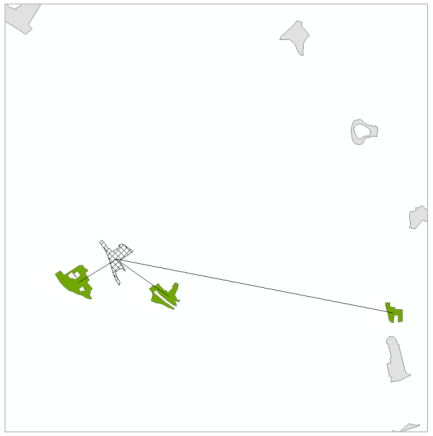
Examples of neighborhood definitions



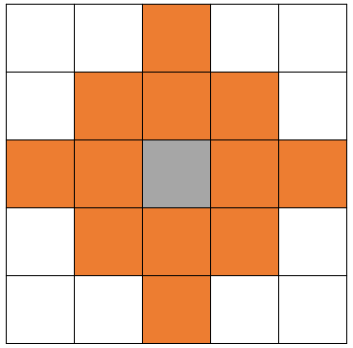
1st order Rook contiguity



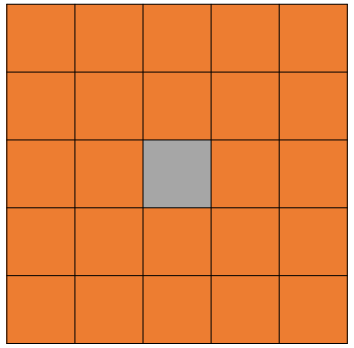
1st order Queen contiguity



3 nearest neighbors distance
(distance of *any* kind)



2nd order Rook contiguity



2nd order Queen contiguity

Points and polygons

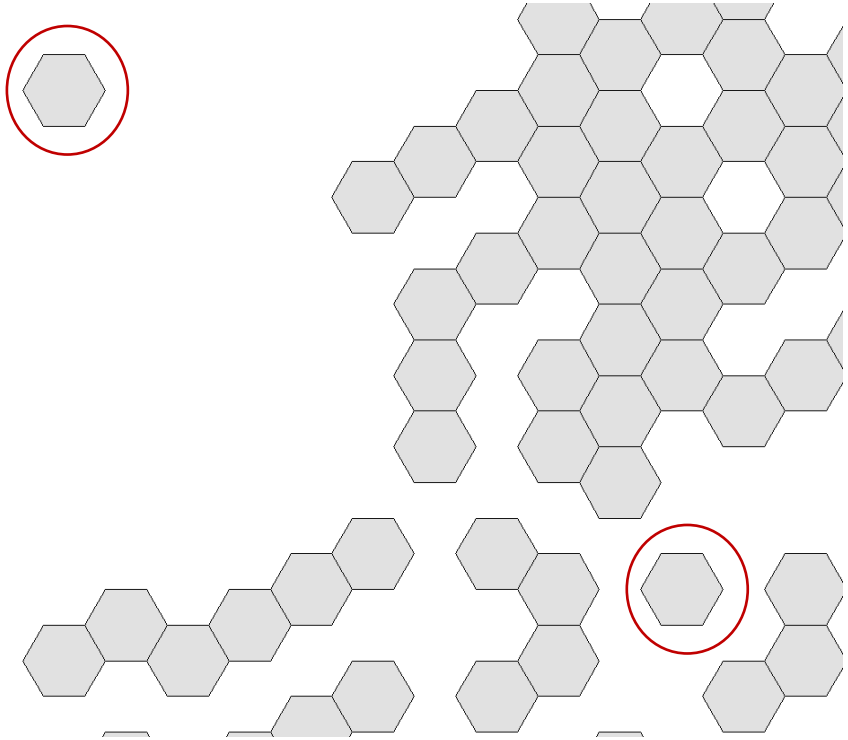
By definition, contiguity-based weights need polygons.

- For points you must employ a polygon grid and then aggregate your observations into polygons.
  - Problematic for some fields in microeconometrics (e.g. real estate analysis)
- Or you can assume that contiguity of points is borrowing by contiguity of their Thiessen polygons.
  - Imposes a few assumptions about the influence area of geographical entities

Distance-based weights work both with polygons and points.

- But you lose the symmetric property in **W**: GeoDa will not run regressions, R will issue warnings

Islands



- Islands are observations that have no contiguity neighbors. This can affect the analysis.
- Employing distance weights addresses this issue, but creates some other practical complications (e.g. with software).
- Another way is to manually remove the islands beforehand. But what if the removed observations are important?
- So, judgment and experimentation is needed. However, a rule is to think about the nature of the spatial process you are trying to reveal. If you have theoretical reasons to keep islands, then do so.

IV

Spatial clustering analysis

Workflow

✓ Create spatial weights

▼

Test for spatial effects → **Spatial (auto)correlation tests.**

Statistical test for clustering and identification of clusters.

▼

Estimate the effects

▼

Simulate the effects

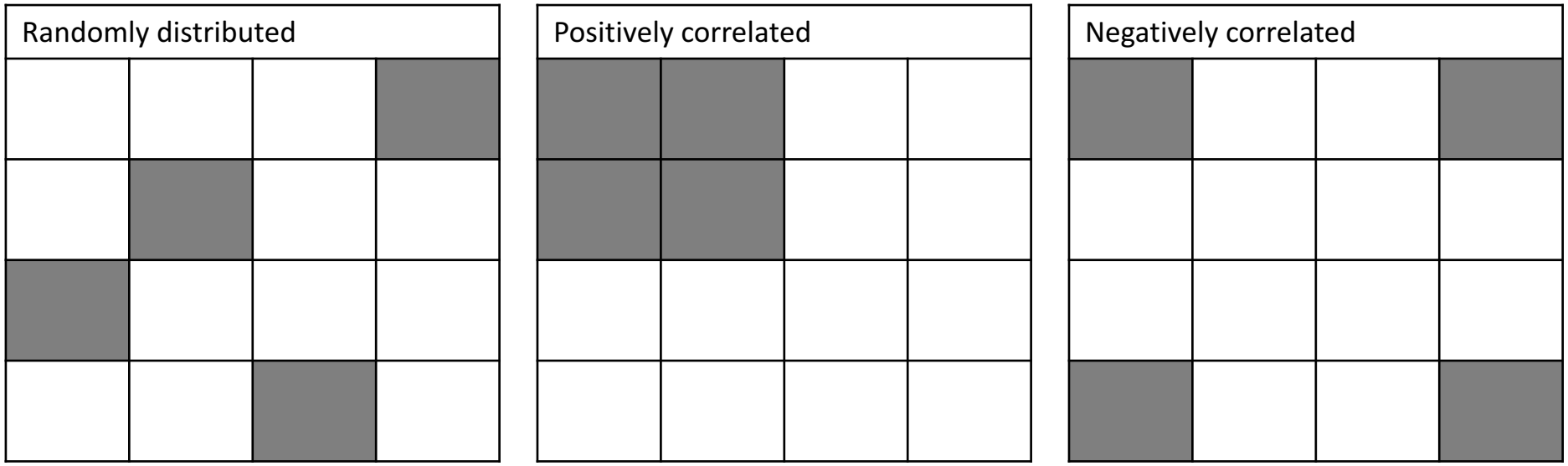
Spatial dependence and autocorrelation

- The effects induced by the *hypothesized* underlying spatial-temporal process are divided in spatial dependence and spatial heterogeneity.

- Spatial dependence is practically investigated through the concept of **spatial autocorrelation**.

- Auto (= self) + Correlation means that a variable ***Y*** at location ***i*** is correlated with the same variable ***Y*** at a nearby location ***j.***

Consider 4 geographical entities that try to organize inside a 4×4 grid



Variance, covariance, correlation

- Variance: the expected squared variation of X from its mean: $\boldsymbol{var(X) = E\left[(X - \mu_X)^2\right] = E[(X - \mu_X)(X - \mu_X)]}$
- Covariance measures the extent that two variables X and Y "move" together: $\boldsymbol{cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]}$.
- If the expected value of Y (μ_Y, its mean) depends on the value of X, then: $\boldsymbol{E(Y|X) \neq \mu_Y}$. This means that $\boldsymbol{cov(X, Y) \neq 0}$. *(however, it has difficult-to-interpret units: "X*Y")*
- So, we measure correlation: $\boldsymbol{corr(X, Y) = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}}$ *(no units: XY/XY = 1)*
- That in turn means that the variables are correlated: $\boldsymbol{corr(X, Y) \neq 0 \ \in \{-1, +1\}}$

Correlation in the spatial context

- Now, consider instead of X and Y, just one variable Y, but at two different locations *i* and *j*.

- If:
  - $\boldsymbol{cov}(\boldsymbol{Y_i}, \boldsymbol{Y_j}) = \boldsymbol{E}[\boldsymbol{Y_i Y_j}] - \boldsymbol{E}[\boldsymbol{Y_i}]\boldsymbol{E}[\boldsymbol{Y_j}] \neq \boldsymbol{0}$
  - and this nonzero covariance has a spatial structure,
  - then we have spatial covariance,
  - and Y has all the potential to exhibit spatial autocorrelation: "$\boldsymbol{corr}(\boldsymbol{Y_i}, \boldsymbol{Y_j}) \neq \boldsymbol{0}$" .

Spatially lagged variables

- The spatial weights matrix can produce the spatially lagged variable Y_L of a variable Y: $\boldsymbol{y}_L = \boldsymbol{y}\mathbf{W}$

- But **W** encodes if y_i is near y_j! It is a sort of a moving average – the surrounding *Y*'s of *Y* at location *i*, averaged and smoothed according to the spatial weights matrix.

- So with Y_L and *Y* in hand, you can formally test for spatial autocorrelation.

2 types of spatial autocorrelation tests

Global spatial autocorrelation

- Tests for cluster**ing** in the data
- Cannot tell you where the clusters are
- Global Moran's I statistic is one way to test for global SA

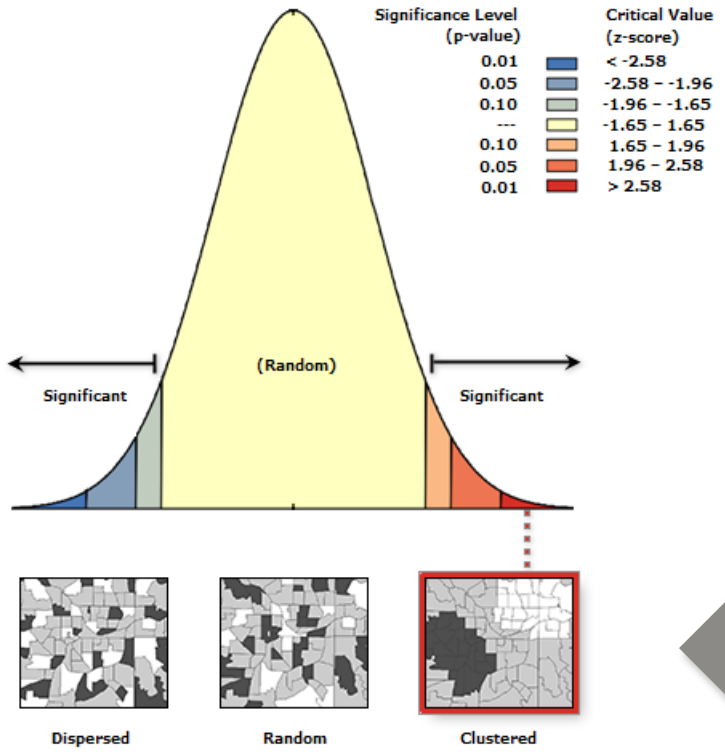
Local spatial autocorrelation

- Identifies cluster**s** and their statistical significance
- Local Indicators of Spatial Association (LISA) is one such test

Significance is assessed by a pseudo p-value

- Tells how likely it is that the identified autocorrelation is different than randomness
- Permutation-based: reshuffle observations many times, compare whether these new "maps" differ from your original data

Global Moran's I statistic
"global autocorrelation indicator"



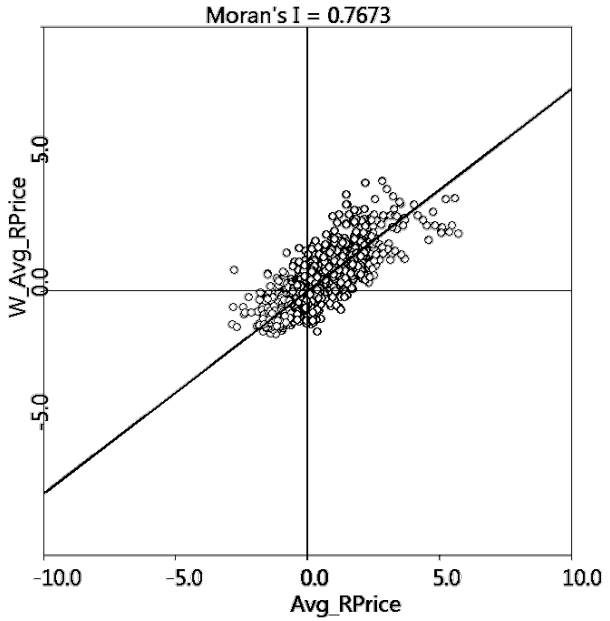
- Measures how dispersed or clustered the observations are, in general.

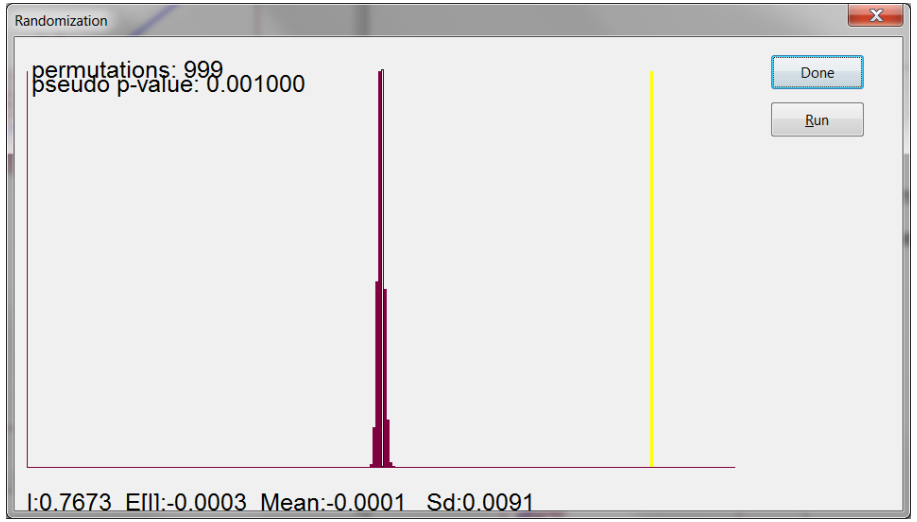
- Range of values {-1,+1}
- Negative values indicate dispersion
- Positive values indicate clustering
- The closer to -1 or +1, the stronger the effect

ESRI ArcGIS gives a nice visualization.

Moran’s I implementatioin in GeoDa

- A **Moran scatterplot** plots a variable Y against its lagged variable Y_L .
- Statistical significance is assessed by a **pseudo p-value**:
  - Tells how likely it is that the identified autocorrelation is different than randomness
  - Permutation-based: reshuffle observations many times, compare whether these new “maps” differ from your original data





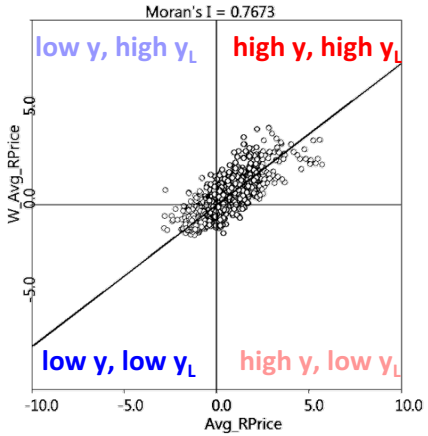
LISA (Anselin 1995)
"local indicators of spatial association"

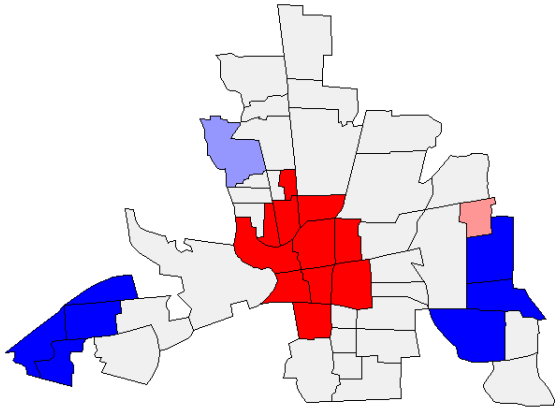
A local decomposition of global Moran's I.

- Indicates significant spatial autocorrelation for specific locations
- Indicates whether clustering in a particular location is significantly different than in its surrounding locations

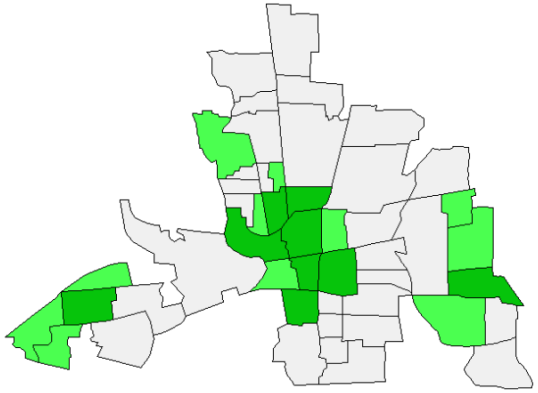
Statistical significance is assessed by a **pseudo p-value**:

- Permutation-based: holds value at location i fixed, permutes others e.g. 999 times, does the same for all locations





Clusters



Significances

Bivariate extensions & limitations

Extrensions

- Bivariate Moran's I, Bivariate LISA, Space-time correlation.
- The idea is that you investigate the correlation of variable Y at one location to a different variable X at another location.

Limitations:

- Keep in mind that this is **exploratory data analysis – you cannot infer causality, only correlations**!
- This framework of spatial clustering is biased toward **outliers**. Mid-range values are irrelevant.
- LISA is stricter than Getis-Ord G* hot-spot analysis (yields fewer clusters).

Questions for this session

1. What is a spatial weights matrix, what assumptions does it make, and how is it used?

2. What is spatial clustering and what are the different kinds of it?