Universiteit Leiden
The Netherlands

Urban Computing

Dr. Mitra Baratchi

14 September 2020

Leiden Institute of Advanced Computer Science - Leiden University

Third Session: Urban Computing - Processing Spatial Data

Table of Contents

Preliminaries on spatial data

Table of content

What is spatial data?

- What is spatial data?
- Spatial datasets?
- Spatial statistics versus classical statistics?

What is spatial data?

- Data that associates locations to each data instance
  - $\mathbf{X} = \{x_{s_1}, x_{s_2}, ..., x_{s_n}\}$ where $s_i \in \mathbb{R}^{2\times 1}$
- Examples:
  - Temperature values for different cities
  - GDP values for countries
  - Number of crimes happening across a city
  - Pixel values in a grayscale image
  - Frequency band values of remote sensing images
  - ...
- Spatial versus geo-spatial $\rightarrow$ Any image versus geo-spacial images

Spatial databases

- **A spatial database**: is a database optimized for storing and querying objects defined in a geometric space.
  - Geometric objects:
    - Points
    - Lines
    - Polygons

Geometric feature

Vector data structures that represent specific features on the Earth's surface, and assign aributes to those features. Example: Geopandas

Figure 1: Point data

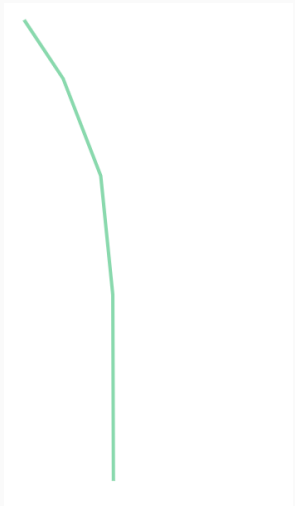


Figure 2: line data

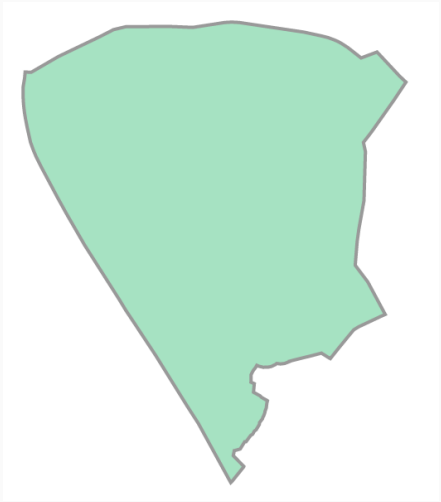


Figure 3: polygon data

Spatial statistics versus classical statistics

- **Case:** You have the data on the amount of rainfall in different locations in the Netherlands and you want to predict the value of temperature in Leiden
  - **Data you have:** $\rightarrow$ temperature, wind power, rainfall, GPS coordinates
- How can you define a regression task to solve this? (dependent value, independent value)

Spatial statistics versus classical statistics

Key difference:

- **The assumption in classical statistics:** Data samples are Independent and identically distributed (i.i.d. or iid or IID)
    - Each random variable has the same probability distribution as the others and all are mutually independent

iid versus spatial correlation

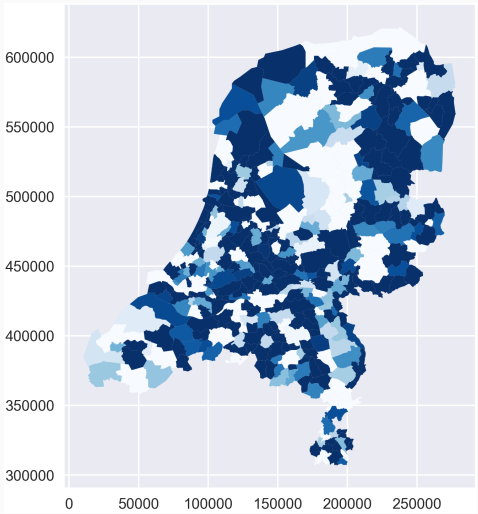


Figure 4: Independent and identically distributed data

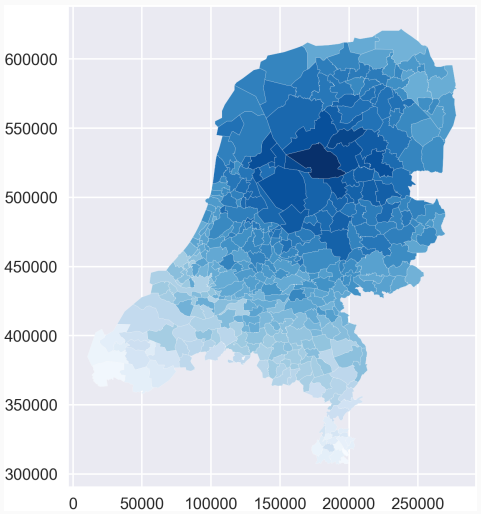


Figure 5: Data distributed with correlation over space

iid versus spatial correlations

X

Independent

ean_hotel_id	name	star_rating	high_rate	low_rate
269955	Hilton Garden Inn Albany/SUNY Area	3	154.0272	124.0216
113431	Courtyard by Marriott Albany Thruway	3	179.01	134
108151	Radisson Hotel Albany	3	134.17	84.16
254756	Hilton Garden Inn Albany Medical Center	3	308.2807	228.4597
198232	CrestHill Suites SUNY University Albany	3	169.39	89.39
125200	The Desmond Hotel Albany	3.5	189.0266	153.0644
109728	Ramada Plaza Albany	3	158.6321	89.036

X (s)

Spatially dependent

ean_hotel_id	name	star_rating	high_rate	low_rate	city	latitude	longitude
269955	Hilton Garden Inn Albany/SUNY Area	3	154.0272	124.0216	Albany	42.68751	-73.81643
113431	Courtyard by Marriott Albany Thruway	3	179.01	134	Albany	42.68971	-73.82021
108151	Radisson Hotel Albany	3	134.17	84.16	Albany	42.7241	-73.79822
254756	Hilton Garden Inn Albany Medical Center	3	308.2807	228.4597	Albany	42.65157	-73.77638
198232	CrestHill Suites SUNY University Albany	3	169.39	89.39	Albany	42.68873	-73.81854
125200	The Desmond Hotel Albany	3.5	189.0266	153.0644	Albany	42.72874	-73.79807
109728	Ramada Plaza Albany	3	158.6321	89.036	Albany	42.68031	-73.78444

Spatial data

First law of geography:

All things are related, but nearby things are more related than distant things. [Tobler70]

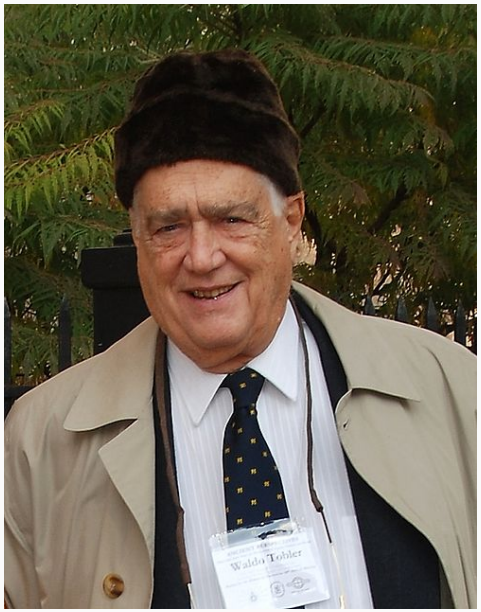


Figure 6: Waldo Tobler [1]

[1] https://en.wikipedia.org/wiki/WaldoR.Tobler

Spatial statistics versus classical statistics

Classical statistics: Data samples are IID

- Simplified mathematical ground (Example: Linear Regression)

Spatial statistics: Data samples are non-IID distributed.

- Methods should be able to capture spatial affects:
  - **Spatial correlation:** What happens north, south east, and west of here depends is very likely to be dependent on what is happening here.
  - **Spatial heterogeneity:** Different concentration of events, etc over space. Similarity of values decay with distance.

Temporal statistics: Data are non-IID

- **Temporal correlation:** What happens now determines what happens next (one directional flow from past to present)
- **Temporal heterogeneity**: Non-stationarity over time

Table of content

How do we represent spatial data to algorithms?

- How do you represent each of these examples (space domain):
    - Crime events and coordinates
    - Rainfall and coordinates
    - Population and coordinates

How do we represent spatial data to algorithms?

What points should you consider:

- What is a variable's nature?
    - Discrete, continuous
- What is the location data nature?
    - Discrete, continuous
        - To answer this question we need to know about the nature of the underlying process

How to represent data over space?

In general there are three classic approaches for dealing with spatial data. This depends on the underlying process:[2]

- Geostatistical process
- Lattice process
- Point process

[2] Noel Cressie and Christopher K Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2015.

Geo-statistical process

- **Fixed continuous location:** observations with a continuously varying quantity; a spatial process that varies continuously being observed only at few points
- Examples: rainfall, wind speed, temperature, coordinates
- Statistical methods based on geo-spatial data:
    - **Gaussian process regression (Kriging)**: spatial interpolation

Kriging[3]

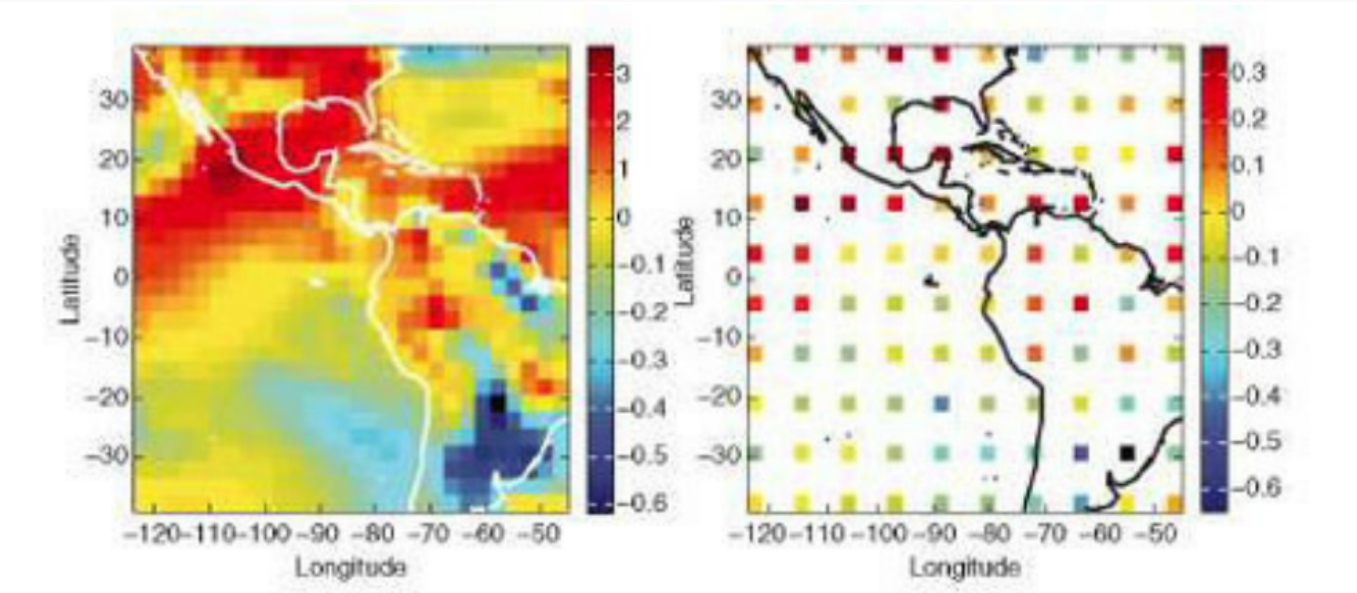


Figure 7: simple geo-statistical data and recovering through simple kriging predictor

[3] Cressie and Wikle, *Statistics for spatio-temporal data*.

Lattice process

- **Fixed discrete location**: Counts or spatial averages of a quantity over regions of space; aggregated unit level data.
- Examples: aggregate data of census, income, number of residents
- Data is represented in discrete spatial units (grid cells, regions, pixels, areas)
- Statistical methods designed based on lattice processes:
    - **Spatial auto-correlation:** Is there a correlation between neighboring units?

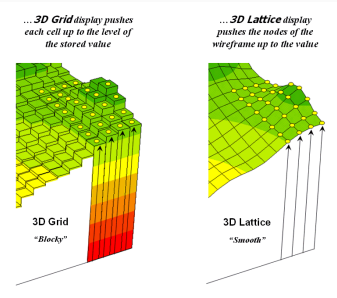


Figure 8: 3D Grid and Lattice [4]

[4] https://blogs.ubc.ca/advancedgis/schedule/slides/spatial-analysis-2/lattices-vs-grids/

Lattice process

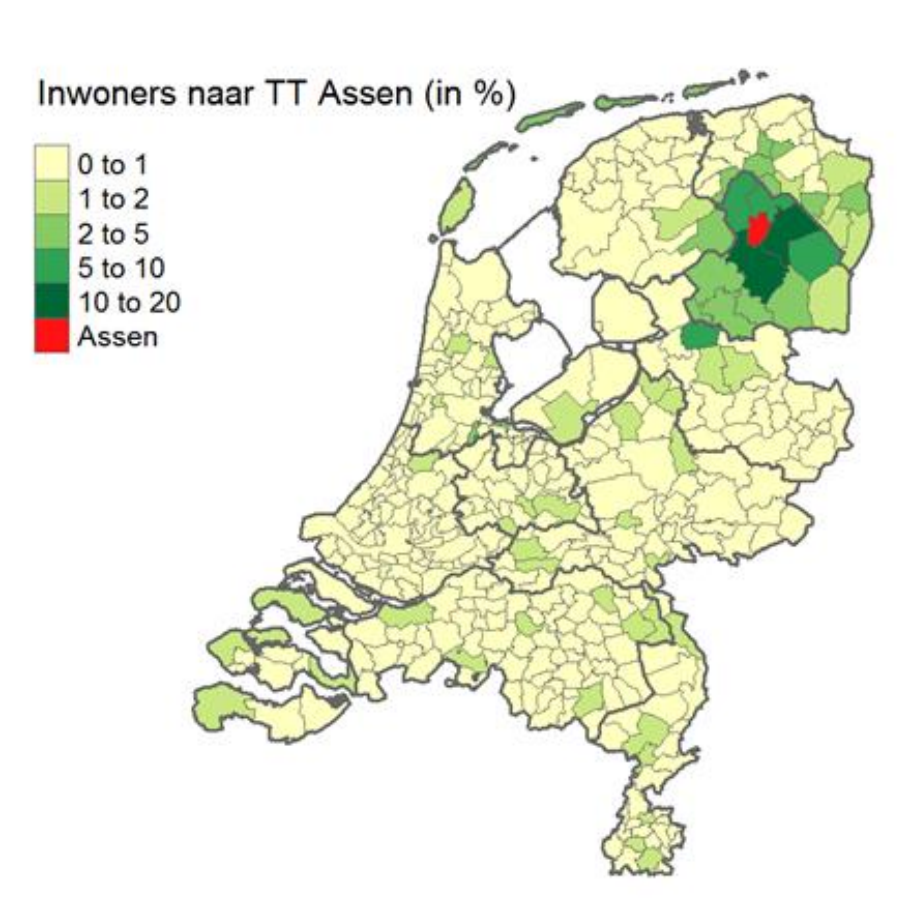


Figure 9: People who went to TT Assen from other cities

Point process

- **Random continuous location:** the spatial process is observed at a set of locations; the locations are interesting as well
- Examples: location of wildfires, earthquakes, accidents, burglaries
- Data is represented by arrangement of points on a region
- Methods designed based on point processes:
    - **K-function**: considers the distance between points in a set

Point process

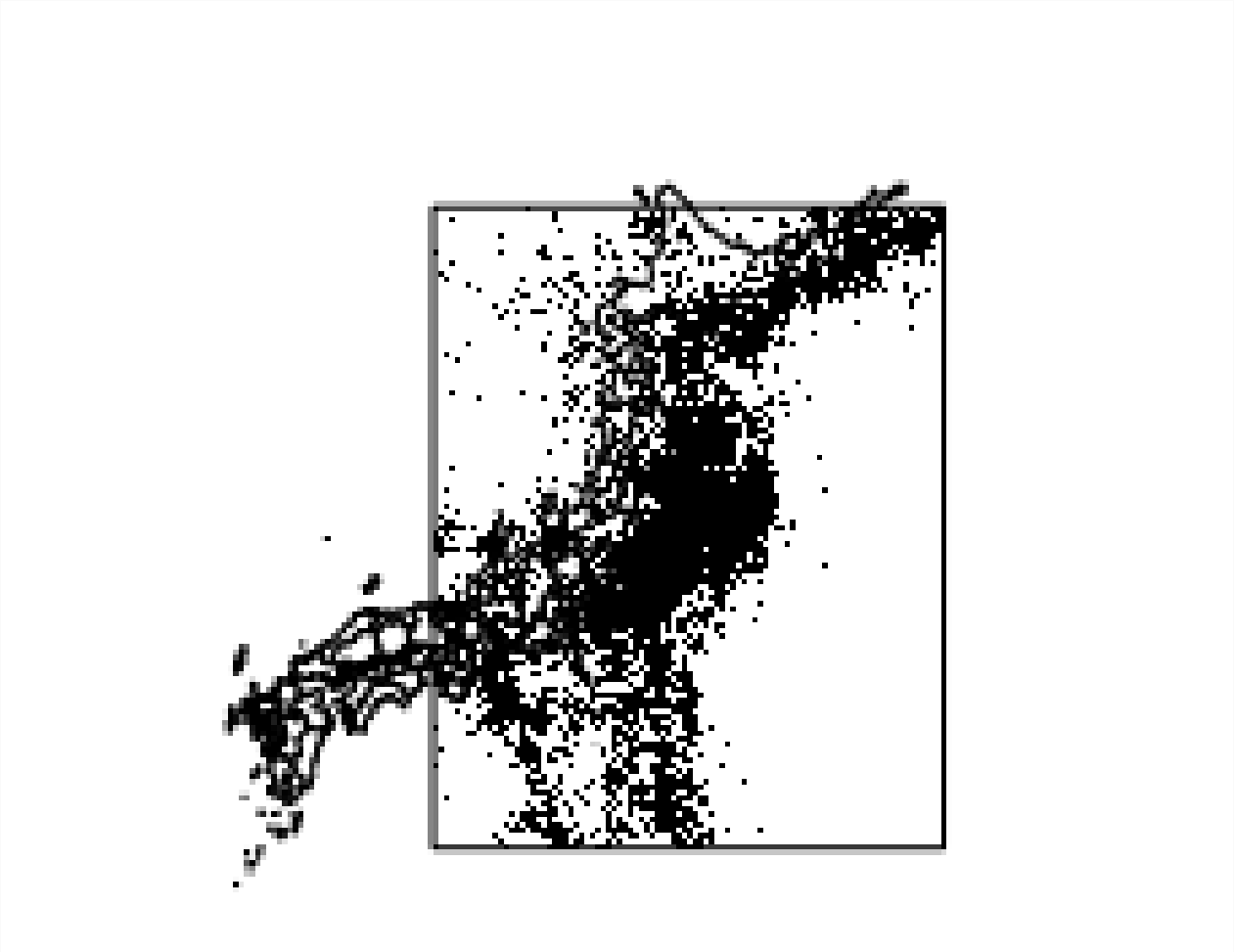


Figure 10: The Japan Earthquake data contained earthquake locations and magnitudes from 2002 to 2011[5]

[5] http://www.stat.purdue.edu/ huang251/pointlattice1.pdf

Various statistical indicators and methods for different representation

- **Geo-statistical process:** kriging, variogram, etc.
- **Lattice process:** cluster and clustering detection, spatial autocorrelation, etc.
- **Point processes:** point pattern analysis, marked point patterns, K-functions, etc.

We can't take a look at all of them but we will look at some

Other ways to represent data

- Space domain (point, geo-spatial, lattice)
- Alternative domains (out of the scope of this session):
  - Applying Fourier, Wavelet transform on the Lattice representation
  - Inspired from the image processing literature
  - Convolutional neural networks: (convolutions are multiplication of signals in frequency domain)

Methods for processing spatial data

Table of content

Spatial auto-correlation, does spatial correlations exist?

Problem: Are the data instances IID or non-IID? Does spatial correlation exist?

- A question to answer during the data exploration phase

Spatial auto-correlation

What does +1, 0, -1 spatial auto-correlation value mean when observed in data?

- Positive
  - Typical in Urban data
  - Similar values happen in neighboring locations. (High, High), (Low, Low)
  - Closer values are more similar to each other than further ones
- Zero
  - IID
  - Randomly arranged data over space
  - No spatial pattern
- Negative
  - Dissimilar values happen in neighboring locations (High, Low), (Low, High), Checker board pattern
  - Typically a sign of spatial competition

How spatial auto-correlation function is designed:

We learned about the temporal auto-correlation. How should be implement spatial auto-correlation?

- We need to capture
  - Attribute similarity
  - Neighborhood similarity

The different between temporal and spatial auto-correlation

What do you remember about temporal auto-correlation?

- **Temporal:** Self-similarity of data over time, Previous data instances determine future data instances
- $ACF_\tau = \frac{1}{T}\sum_{t=1}^{t=T-\tau(or T)}$ [6]$(x_t - \overline{x})(x_{t+\tau} - \overline{x}), \tau = 0, 1, 2, ..., T$ [7]
- **Spatial:** Self-similarity over space, Neighboring data instances determine each other
- ?

[6] T is used in circular autocorrelation

[7] max value of τ can be smaller

Spatial auto-correlation

What is the equivalent of temporal lag in space? $\rightarrow$ Distance

Moran's I (spatial auto-correlation)

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

- x_i is the value of a variable at location i
- W is a matrix of weighted values
  - Each w_{ij} in W represents the the effect of element x_i on element x_j
- $|W|$ is sum of the values of w_{ij} and N is the sample size

Table of content

How to show spatial dependence over neighborhoods?

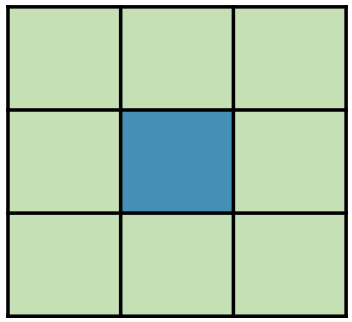
- We need some representation of dependence and interactions over space
- The most common way people are considering these effects is by using Spatial Weights Matrices W
  - N$\times$ N positive matrix containing the strength of interactions between spatial point i and j
- Many algorithms designed for spatial data make use of weight matrices
  - Spatial auto-correlation
  - Spatial regression
  - Spatial clustering

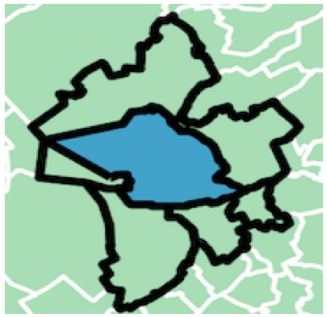
How to assign weights to neighbors?

- N variables and N^2 comparisons to make to consider all neighbors $\rightarrow$ for the sake of efficiency some can be ignored (the interaction can be set to zero)
- Ignored neighbors: $w_{ij} = 0$
- Important neighbors:
  - $w_{ij} = 1$
  - $w_{ij} = 0 < wij < 1$
- Non-binary weights can be a function of:
  - Distance
  - Strength of interaction (e.g. commuting flows, trade, etc.)
  - ...

Weights matrix

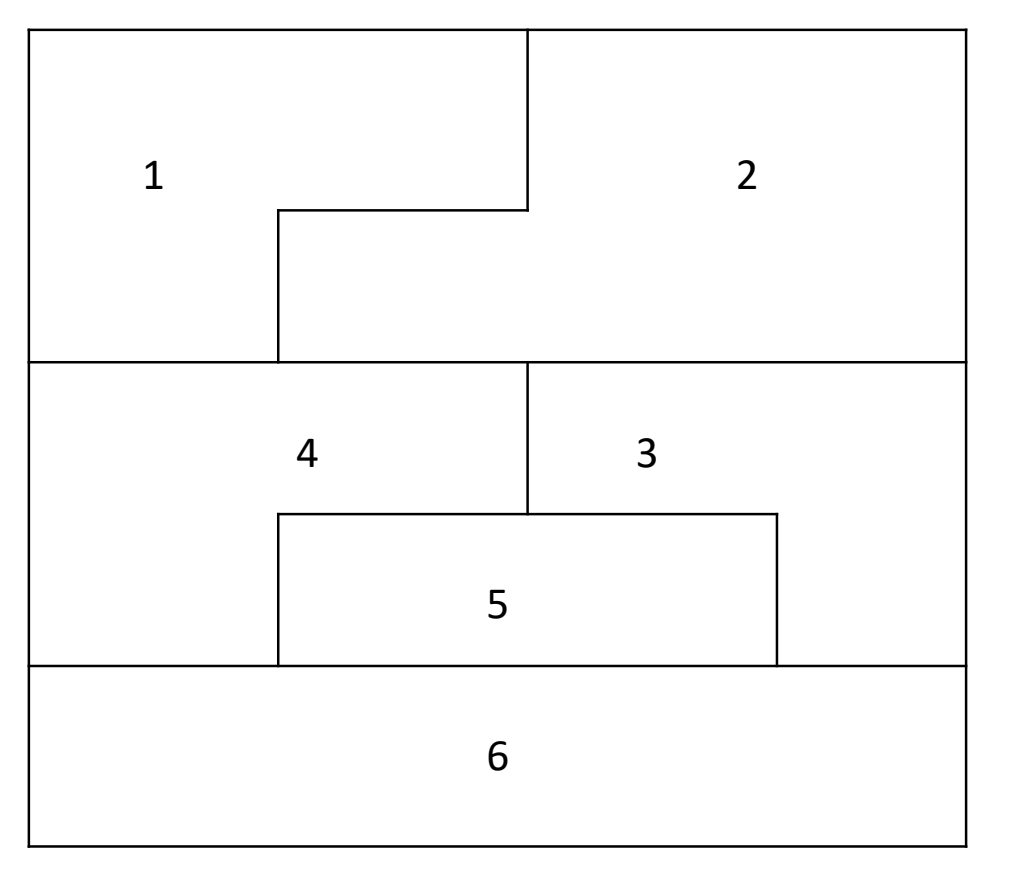
How do we represent interactions from raster and polygon data in a matrix?





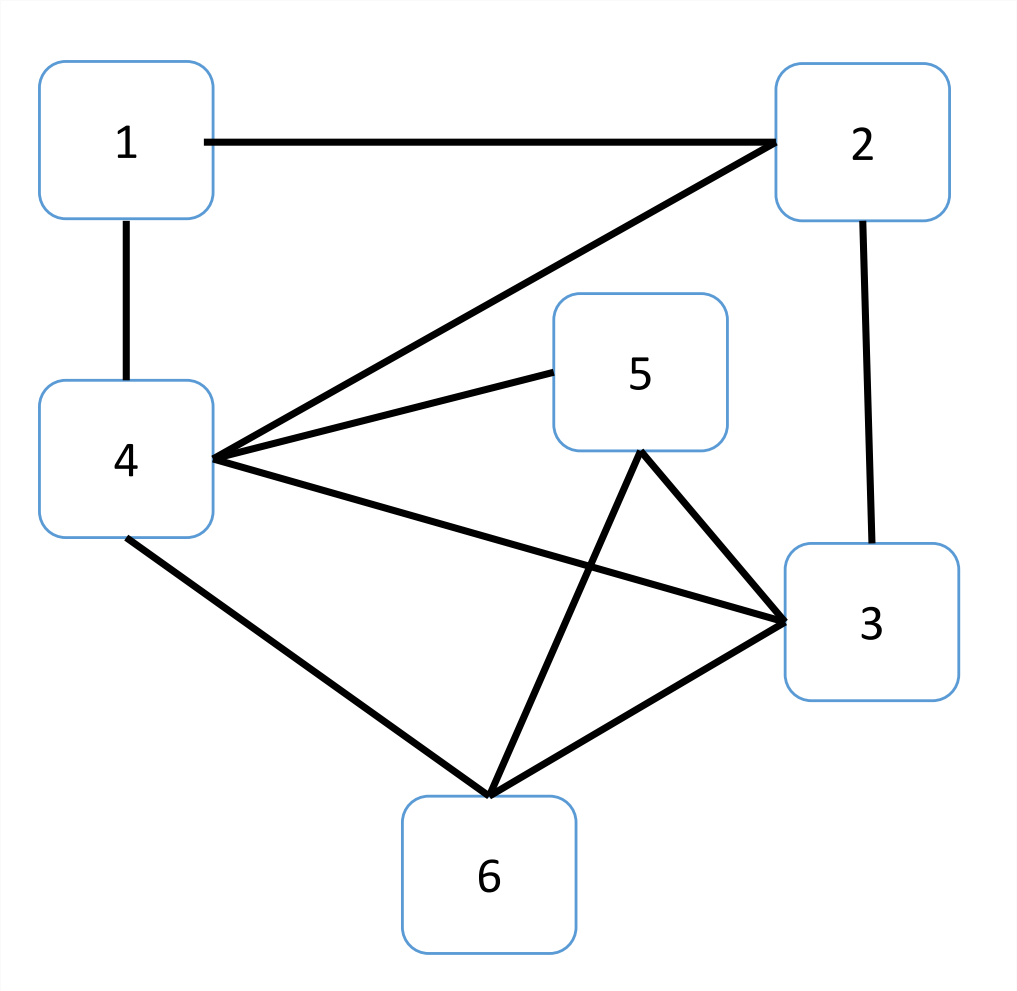
Weights matrix

How do we represent interactions from raster and polygon data in a matrix?



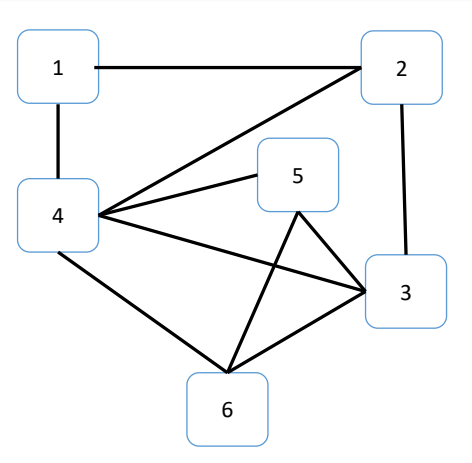
Weights matrix

Create a graph representation showing neighboring cells based on having a common border



Graph representation and adjacency matrix

Use the adjacency matrix of the graph to create the weight matrix:



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

Is there a solution?

This way we can only show neighbors that have common edges. What if we cared about the physical distance? or two-hop away neighbors?

Neighbors

How do we define neighborhood? What neighbors do we care about? (i.e. select non-zero elements of W):

- **Contiguity-based**: Having a common border
- **Distance-based**: Being in the vicinity
- **Block-based**: Being in the same place based on an official agreement
  - Provinces
  - Cities and countries
  - ..
- ...

Contiguity-based weights

Figure 11: How can you move to a neighboring cell?

Contiguity-based weights

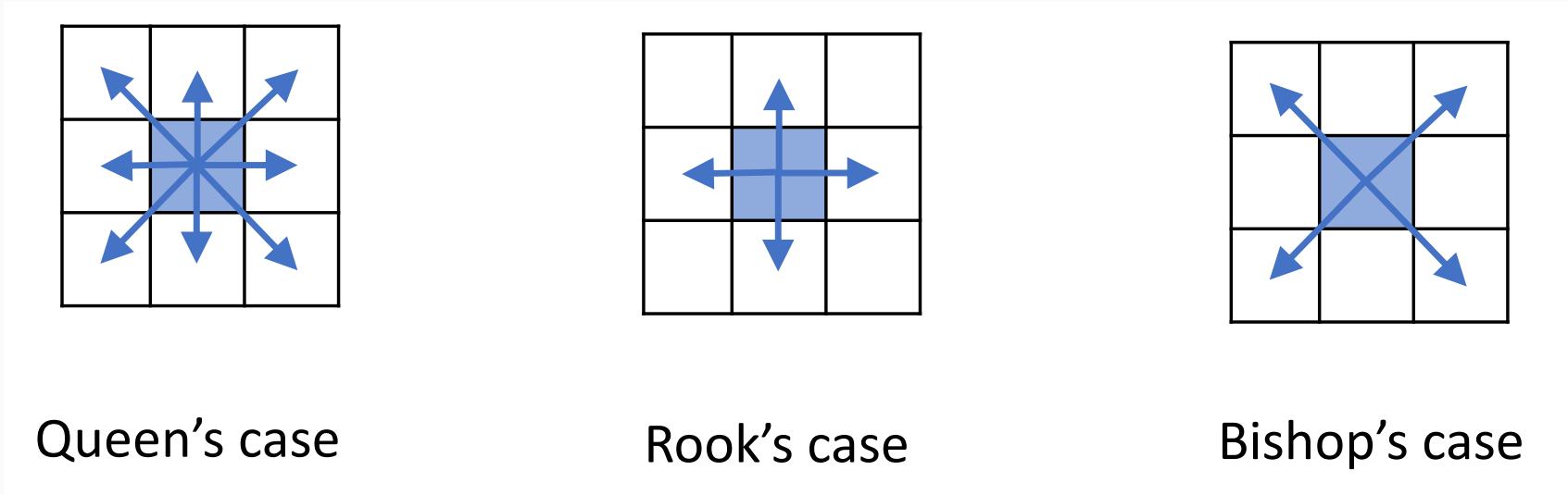


Figure 12: neighborhood cases

Queen's case

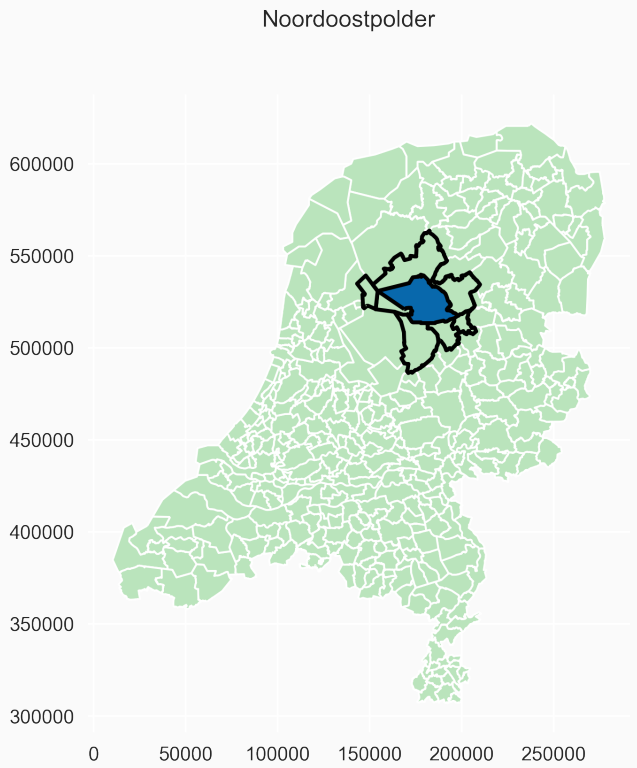


Figure 13: Queen's case

Rook's case

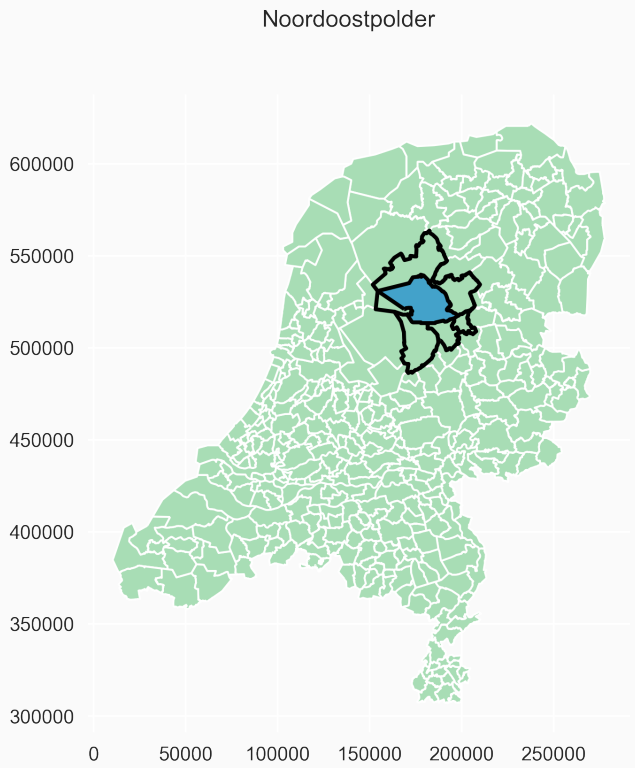


Figure 14: Rook's case

Bishop's case

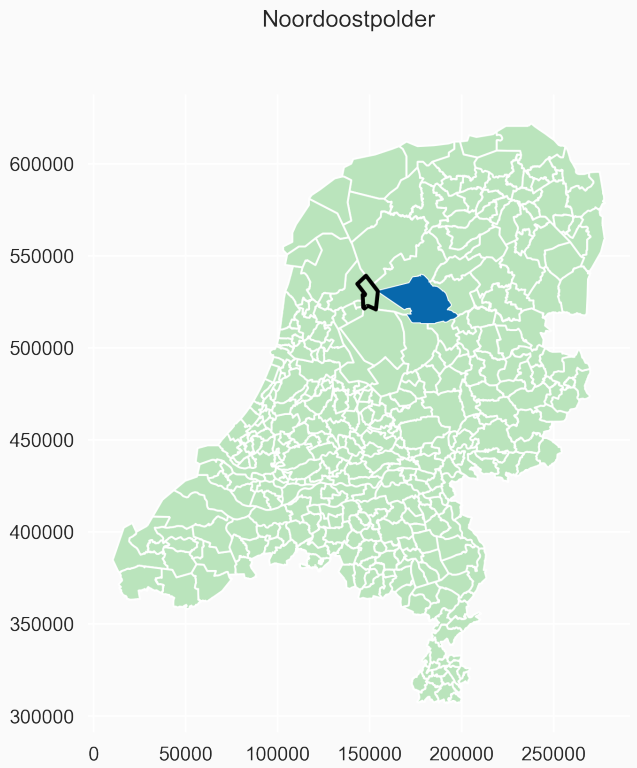


Figure 15: Bishop's case

Distance-based

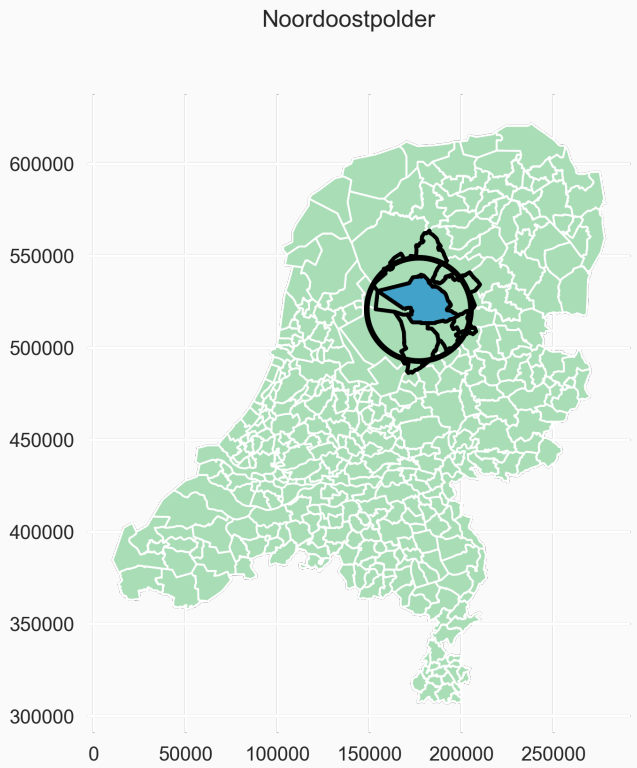


Figure 16: distance-based neighborhoods

Block neighborhood

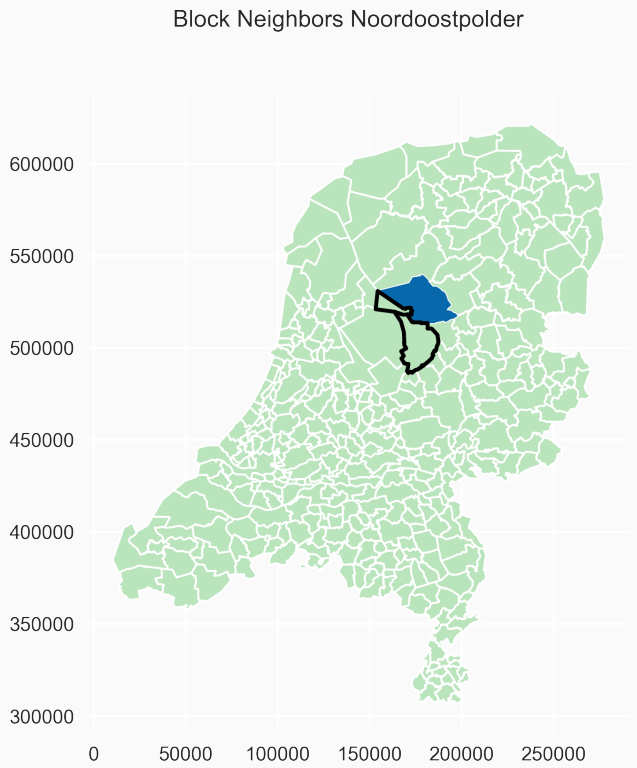


Figure 17: Block neighborhood based on province (Flevoland)

What neighborhood to choose from

Neighborhood should reflect how interaction happens for the question at hand.

- **Contiguity weights:** Processes that propagate geographically from borders
- **Distance weights:** Accessibility
- **Block weights:** Effects of provincial laws

8

[8] Dani Arribas-Bel. *Geographic Data Science'16*. 2017. DOI: 10.5281/zenodo.290239. URL: http://darribas.org/gds16 (visited on 02/10/2017).

Table of content

Regressive models over space

Problem: A regression model for predicting the value of a dependent variables (represented in a vector Y_n)

- **Regression model** (no temporal and spatial effect)
- **Auto-regressive models** (temporal effect)
- **Auto-regressive models** (spatial effect)
  - Key factors to consider:
    - How the phenomenon diffuses in space? (spatial lag model)
    - Local and Global effect

Remider:Regression, Auto-regressive, Moving average

$\rightarrow$ c is constant, ϕ is model parameter, ϵ is white noise

- **Regression**
  - $Y_i = c + \phi X_i + \epsilon_i$
- **Auto-regressive**
  - $X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t$
- **Moving average**
  - $X_t = c + \sum_{i=1}^{q} \phi_i \epsilon_{t-i}$
  - Literally moving average, (i.e.) average value of previous values of the time-series
- **Auto-Regressive Moving Average (ARMA)**
  - $X_t = c + \sum_{i=1}^{q} \phi_i \epsilon_{t-i} + \sum_{i=1}^{p} \phi_i X_{t-i}$

Auto-regressive models

$\rightarrow X_n$ and Y_n are vectors of independent and dependent variables of size n. ϕ, λ, ρ are model parameters. c is a constant. ϵ represents the noise term. W_n is the spatial weights matrix

- **Regression**
  - $Y_n = c + \phi X_n + \epsilon_n$
- **Spatial Auto-Regressive model (SAR)**
  - $Y_n = c + \lambda W_n Y_n + \epsilon_n$,
  - $W_n Y_n$ is referred to as the spatial lag term in the models
  - How we use W_n determines global and local effect
- **Spatial Moving Average (SMA)**
  - $Y_n = c + U_n$, $U_n = \epsilon_n - \rho W_n \epsilon_n$
  - U_n captures the effect of variables that we do not have in our data
- **Mixed Regressive, Spatial Auto-Regressive Moving Average model (MRSARMA)**
  - $Y_n = c + \phi X_n + \lambda W_n Y_n + U_n$,

Lessons learned

- Spatial statistics versus classical statistics
  - Spatial correlation effect $\rightarrow$ many statistical indicators designed for non-spatial data are not valid for spatial data
- Before working with data we need to represent it in points or polygons based on the underlying process of data
  - **Geo-statistics**: locations are fixed and continuous, numbers are random values
  - **Point Processes**: location and numbers are both random
  - **Lattice Data**: locations are fixed and discrete, numbers are random aggregate values
- Spatial auto-correlation
- Neighborhoods and spatial weights for capturing the effects
  - Contiguity
  - Distance
  - Block

Lessons learned continued...

- Spatial auto-regressive models
  - **SAR:** value of neighboring points as predictive value
  - **SMA:** Noise on the neighboring values as predictive values
  - **MRSARMA:** combination of independent predictive values, neighboring values, and noise from neighboring values as predictive values

End of theory!