Urban Computing

Dr. Mitra Baratchi

Leiden Institute of Advanced Computer Science - Leiden University

February 21, 2019

Universiteit Leiden
The Netherlands

Third Session: Urban Computing - Processing Spatial Data

Agenda for this session

- Part 1: Preliminaries
  - What is spatial data?
  - How do we represent it?
- Part 2: Methods for processing spatial data
  - Spatial auto-correlation
  - Neighborhoods
  - Spatial regression and auto-regressive models

Part 1: Preliminaries

Spatial data?

- Data with spatial location associated with variables
- Spatial data analysis takes the locations in data into account.
- Spatial statistics is a particular kind of spatial data analysis in which the observations or locations (or both) are modeled as random variables.
- Geostatistics considers Geo-spatial knowledge discovery and not only mapping
- Geographic information systems (GIS)
- Spatial data
- Geo-spatial data

Spatial versus geo-spatial

- **A spatial database**: is a database optimized for storing objects defined in a geometric space.
  - Geometric objects:
    - points
    - lines
    - polygons
- **A geo-database**: is a database of geographic data, such as countries, administrative divisions, cities, and related information.

Geodesic features

Figure: Point data

Figure: line data

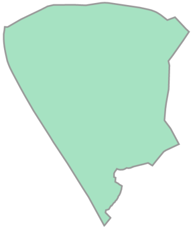


Figure: polygon data

What can you do with spatial data?

What can you do with spatial data?

- Understanding where things are happening?
- Find spatial patterns?
  - clustering
  - where is the clustering happen?
- Predicting the unknown values over space?

What is the approach you take to solve this case?

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:** $\rightarrow$ GPS coordinates, temperature

Different between classical and spatial statistics

Key difference:

- Assumption: 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
  - In many practical urban applications this is not true

Limitation of traditional statistics

Classical statistics:

- Data samples are independent and identically distributed (i.i.d)
- Simplified mathematical ground (Example: Linear Regression)

Spatial statistics:

- Data are non-iid distributed.
- 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.
- Time flows in one direction only (past to present).

Many statistical indicators designed for non-spatial data is not valid for spatial data.

iid and spatial correlation

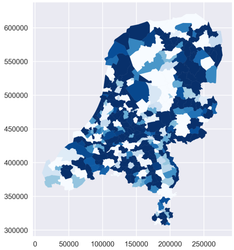


Figure: Randomly distributed data

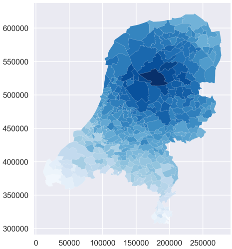


Figure: Data distributed with correlation over space

Spatial data

First law of geography:

[1]https://en.wikipedia.org/wiki/Waldo$_{R}.{}_{T}obler$

Spatial data

First law of geography:
All things are related, but nearby things are more related than distant things. [Tobler70]

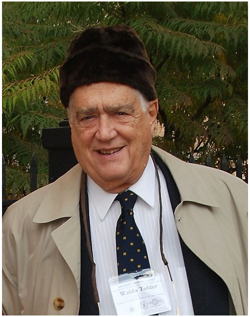


Figure: Waldo Tobbler [1]

[1]https://en.wikipedia.org/wiki/Waldo$_R.{}_T obler$

How do we represent data?

How do we represent data?

Points to consider

- What is a variable's nature?
  - Discrete, continuous
- What is the location data nature?
  - Can you say something about it within the space of its neighboring points?
  - Is location also happen at random?

How to represent data over space?

In general there are three classic approaches for dealing with spatial data: [CW15]

- Geostatistical process
- Lattice process
- Point process

Geo-statistical process

- **Fixed station** observations with a continuously varying quantity; a spatial process that varies continuously being observed only at few points
- Spatial random process $D_s \subset \mathbb{R}^d$
- Examples:

Geo-statistical process

- **Fixed station** observations with a continuously varying quantity; a spatial process that varies continuously being observed only at few points
- Spatial random process $D_s \subset \mathbb{R}^d$
- Examples: rainfall, wind speed, temperature
- Main concern is building models of spatial dependence and predicting the spatial process optimally
- **Gaussian data** model and Gaussian process model
- Parameters are defined based on **mean**, **variance** and **covariance**
- Methods:
  - **Variogram**: measures how similarity decreases with distance
  - **Kriging**: spatial interpolation
- Not suitable for binary or count data

Kriging [CW15]

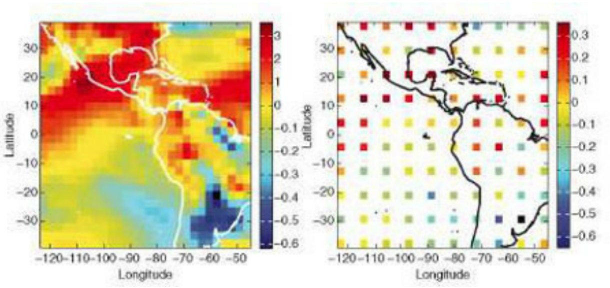


Figure: simple geo-statistical data and recovering through simple kriging predictor

Lattice process

- Counts or spatial averages of a quantity over regions of space; aggregated unit level data.
- $\{Y(s) \in D_s\}$ defined on a finite and countable subset D_s of $\mathbb{R}^d$
- Examples:

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

Lattice process

- Counts or spatial averages of a quantity over regions of space; aggregated unit level data.
- $\{Y(s) \in D_s\}$ defined on a finite and countable subset D_s of $\mathbb{R}^d$
- Examples: aggregate data of census, income, number of residents
- Discrete spatial units (grid cells, regions, pixels, areas)
- Markov type models
- Methods: spatial autocorrelation

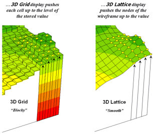


Figure: 3D Grid and Lattice [2]

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

Lattice process

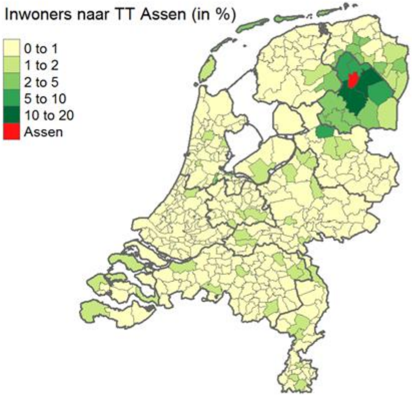


Figure: People who went to TT Assen from other cities

Point process

- Locations and number of events are **both random**. The spatial process is observed at a set of locations and the locations are interesting as well
- Random location of event $\{s_i\}$ in some set $D_s \subset \mathbb{R}^d$ where the number of events in D_s are also random
- Examples:

Point process

- Locations and number of events are **both random**. The spatial process is observed at a set of locations and the locations are interesting as well
- Random location of event $\{s_i\}$ in some set $D_s \subset \mathbb{R}^d$ where the number of events in D_s are also random
- Examples: location of wildfires, earthquakes, accidents, burglaries
- Data is represented by arrangement of points on a region
- Poisson process in space
- Methods: K-function, considers the distance between points in a set

Point process

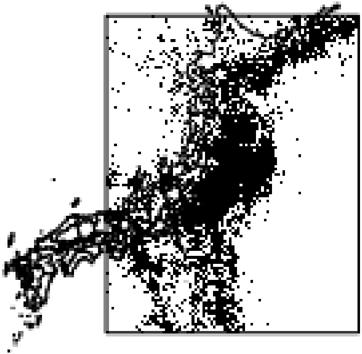


Figure: The Japan Earthquake data contained earthquake locations and magnitudes from 2002 to 2011[3]

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

Various statistical indicators and methods for different representation

- **Geo-statistics:** kriging, variogram, etc.
- **Point Processes:** point patterns, marked point patterns, K-functions, etc.
- **Lattice Data:** cluster and clustering detection, spatial autocorrelation, 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

Part 2: Methods for processing spatial data

Spatial auto-correlation

Spatial auto-correlation, does spatial correlations exist?

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

- Exploration
- Spatial randomness $\rightarrow$ equal probability of every point in space
- No spatial randomness $\rightarrow$ spatial structure exists. Later we can exploit this structure in prediction of values, etc

Spatial Auto-correlation

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

- Positive

Spatial Auto-correlation

What does +1, 0, -1 spatial auto-correlation 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

Spatial Auto-correlation

What does +1, 0, -1 spatial auto-correlation 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
  - i,i,d
  - Randomly arranged data over space
  - No spatial pattern
- Negative

Spatial Auto-correlation

What does +1, 0, -1 spatial auto-correlation 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
  - i,i,d
  - Randomly arranged data over space
  - No spatial pattern
- Negative
  - Not very typical in Urban data, still possible, hard to interpret
  - Dissimilar values happen in neighboring locations (High, Low), (Low, High)
  - Checker board pattern
  - Closer values are more dissimilar to each other than further ones
  - Typically a sign of spatial competition

Spatial auto-correlation key factors

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?

[4] T is used in circular autocorrelation

[5] max value of $\tau canbesmaller$

The different between temporal and spatial auto-correlation

What do you remember about temporal auto-correlation?

- **Temporal:** Previous data instances determine future data instances

[4] T is used in circular autocorrelation
[5] max value of $\tau canbesmaller$

The different between temporal and spatial auto-correlation

What do you remember about temporal auto-correlation?

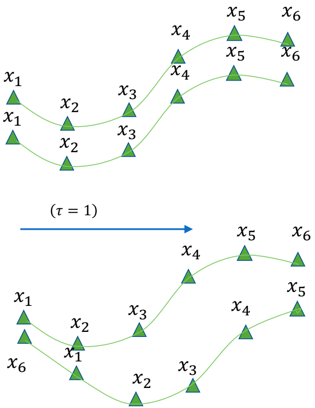
- **Temporal:** Previous data instances determine future data instances
- $ACF_{\tau} = \frac{1}{T} \sum_{t=1}^{t=T-\tau (or T)}$ [4] $(x_t - \overline{x})(x_{t+\tau} - \overline{x}), \tau = 0, 1, 2, ..., T$[5]
- **Spatial:** Neighboring data instances determine each other
- ?

-

[4] T is used in circular autocorrelation

[5] max value of $\tau can be smaller$

Temporal auto-correlation



$ACF(0) \rightarrow (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots.$

$ACF(1) \rightarrow (x_1 - \bar{x})(x_6 - \bar{x}) + (x_2 - \bar{x})(x_1 - \bar{x}) + \ldots.$

How did we capture attribute and neighborhood similarity?

Spatial auto-correlation

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

- Moran's I
- $I(d) = \frac{N}{W}\frac{\sum_i \sum_j w_{i,j}(x_i - \overline{x})(x_j - \overline{x})}{\sum_i (x_i - \overline{x})^2}$
  - $I(d)$= Moran's I correlation coefficient as a function of distance d, $d \in \{1, 2, ...\}$
  - x_i is the value of a variable at location i
  - W_{ij} is a matrix of weighted values
  - W is sum of the values of W_{ij}
  - N is the sample size

Global and location spatial autocorrelation

Clusters versus clustering ....

- **Global spatial autocorrelation:**
  - A measure of the overall clustering of the data.
  - Moran's I
- **Local spatial autocorrelation:**
  - Are there any local clusters?
  - We can still find clusters at a local level using local spatial autocorrelation even if there is no global clustering
  - Local cluster detection involves:
    - Identifying the location of clusters
    - Determining the strength of clusters
    - Local indicators of spatial association
    - Local significance map

How to show spatial dependence over neighborhoods?

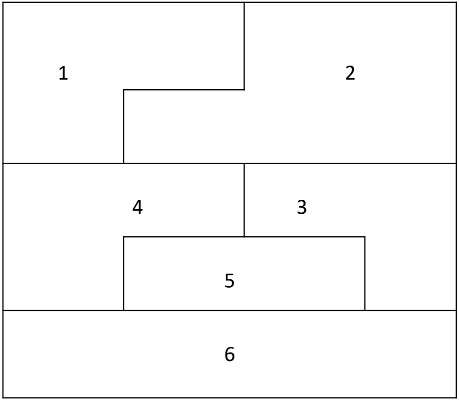
- We need some representation of dependence and interactions over space
- The most common way people have came up with is using Spatial Weights Matrices $W_{i,j}$
  - N$\times$ N positive matrix containing the strength of interactions between spatial point i and j
- Many spatial algorithms rely on them

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.)
  - ...

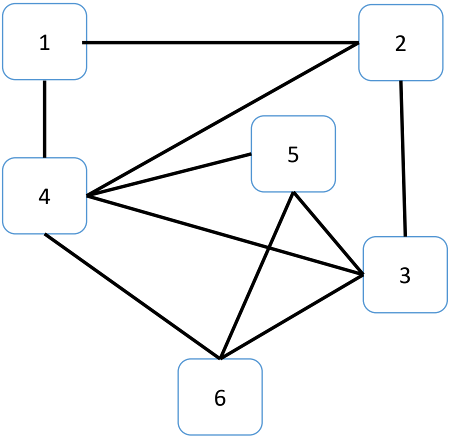
Weights matrix

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



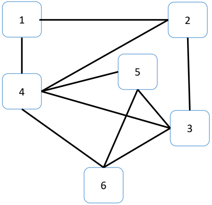
Weights matrix

Create a graph representation...



Graph representation and adjacency matrix

Adjacency matrix



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

Neighbors

How do we define neighborhood? What neighbors do we care about? (i.e. select non-zero elements of $W_{i,j}$):

Neighbors

How do we define neighborhood? What neighbors do we care about? (i.e. select non-zero elements of $W_{i,j}$):

- **Contiguity-based**: Having a common border

Neighbors

How do we define neighborhood? What neighbors do we care about? (i.e. select non-zero elements of $W_{i,j}$):

- **Contiguity-based**: Having a common border
- **Distance-based**: Being in the vicinity

Neighbors

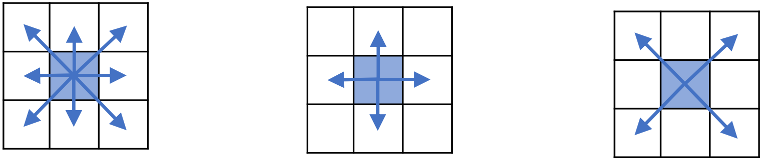
How do we define neighborhood? What neighbors do we care about? (i.e. select non-zero elements of $W_{i,j}$):

- **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: How can you move to a neighboring cell?

Contiguity-based weights



Queen's case Rook's case Bishop's case

Figure: neighborhood cases

Queen's case

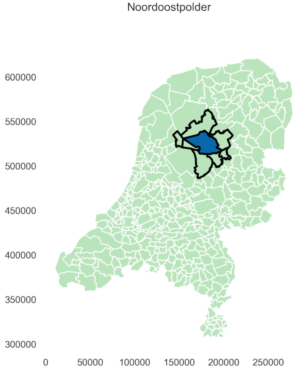


Figure: Queen's case

Rook's case

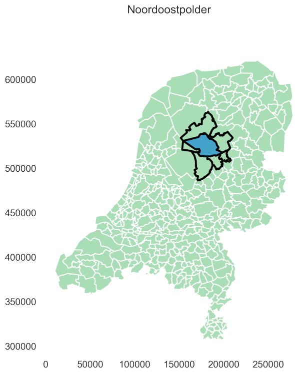


Figure: Rook's case

Bishop's case

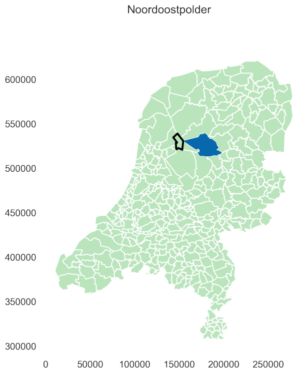


Figure: Bishop's case

Distance-based

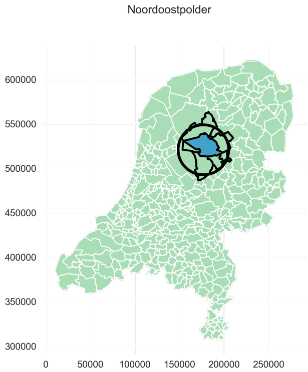


Figure: distance-based neighborhoods

Block neighborhood

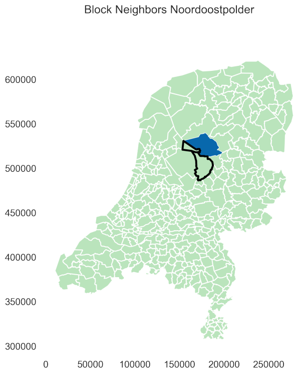


Figure: Block neighborhood based on province (Flevoland)

What neighborhood to choose from

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

What neighborhood to choose from

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

- **Contiguity weights:** Processes propagated geographically (e.g. weather, disease spread)
- **Distance weights:** Accessibility
- **Block weights:** Effects of provincial laws

[AB17]

Spatial auto-regressive models

Regressive models over space

Problem: given Y_n a vector of dependent variables what is the value of y_j

- Auto-regressive models (for time)
- Auto-regressive models (for space)
- Key factors to consider:
  - How the phenomenon diffuses in space? (spatial lag model)
  - Local and Global effect

Autoregressive models

- Spatial (synchronous) autoregressive model (SAR)
  - $Y_n = W_n Y_n \lambda + E_n$,
- Regression model with SAR disturbance
  - $Y_n = X_n \beta + U_n$, $U_n = \rho W_n U_n + E_n$,
  - U_n Captures the effect of variables that we do not have in our data
- Mixed regressive, spatial autoregressive model (MRSAR)
  - $Y_n = W_n Y_n \lambda + X_n \beta + E_n$,

$W_n Y_n$ is referred to as the spatial lag term in the models
How we use W_n determines global and local effect
[6]

[6] X_n and Y_n are vectors of independent and dependent variables of size n. λ and β are model parameters. E represents the noise term. W_n is the spatial weights matrix

End of theory!

References I

Dani Arribas-Bel, *Geographic data science'16*, 2017.

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