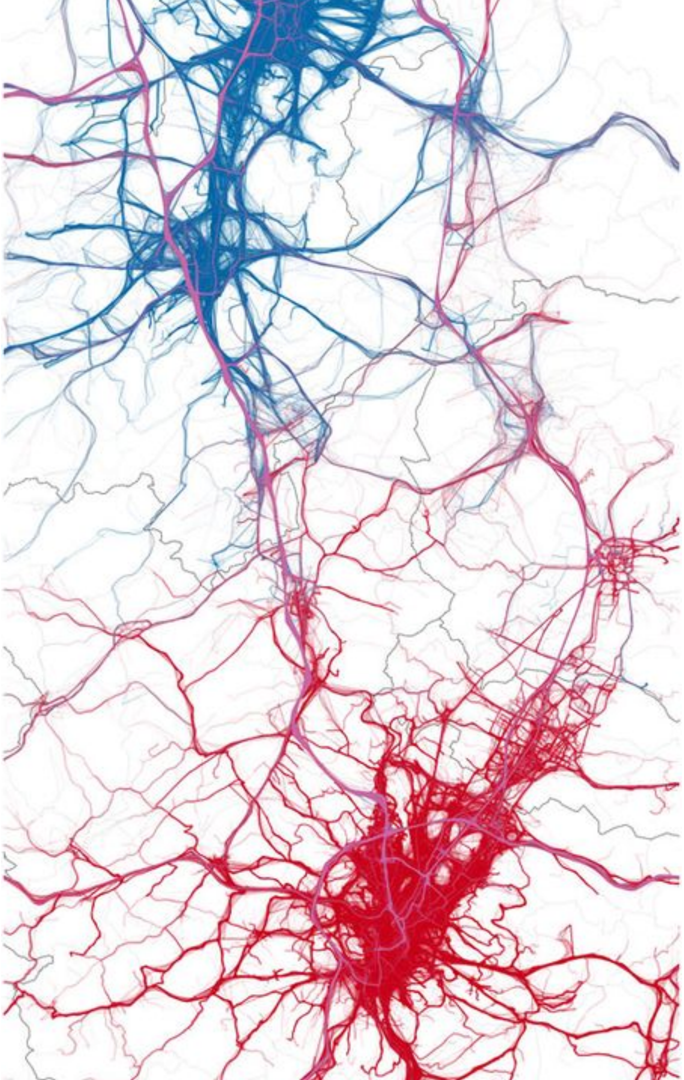


IN SUPREMÆ DIGNITATIS 1343

Consiglio Nazionale delle Ricerche

SCUOLA NORMALE SUPERIORE

03bis Spatial Data Analysis

Todays contents

- Spatial interpolation
  - Thiessen polygons
  - IDW
  - Kriging
- Spatial regression

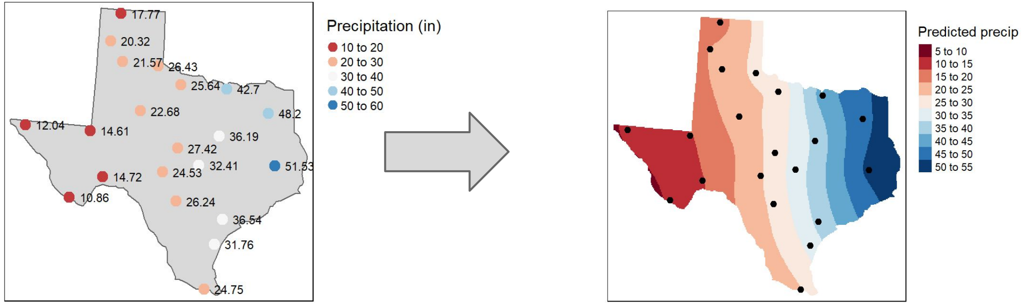
- Spatial associations: co-location patterns

- Spatial trends

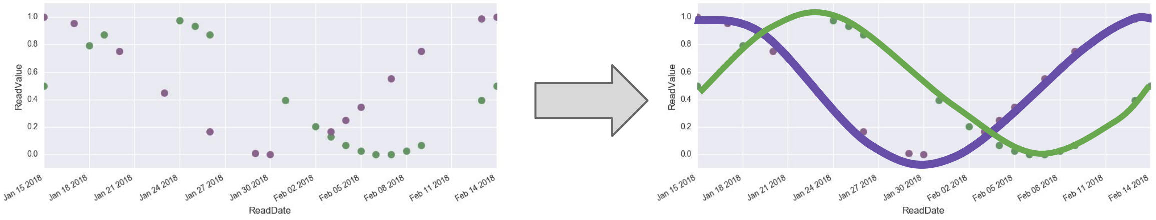
Spatial interpolation

Spatial interpolation: definition

- Given the value of an attribute for a set of spatial points,
  Compute the value of the attribute for all the points in space



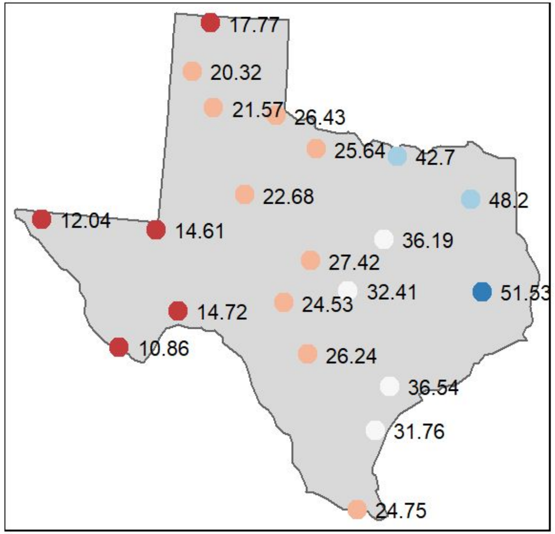
- Analogy with time series: given a subset of points, reconstruct the whole time series

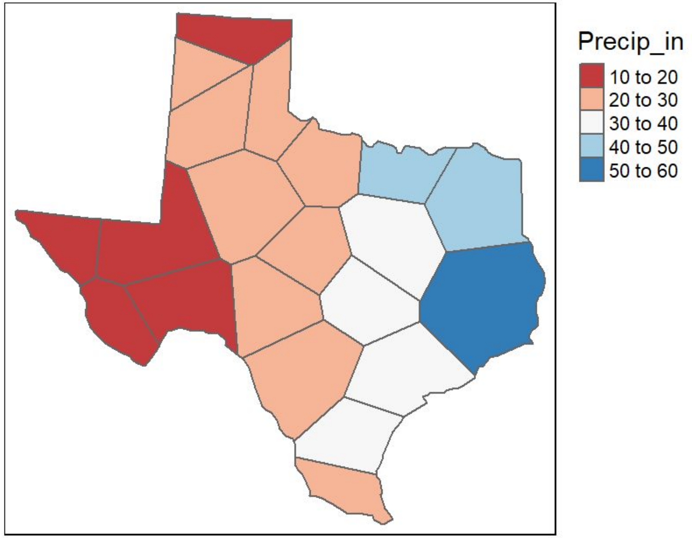


Spatial Interpolation

Thiessen polygons or Proximity interpolation

- The value of a point is the same as the closest input sample
    - Each point is associated to its Nearest Neighbor
    - That yields a Voronoi tessellation around each input point





Spatial Interpolation

Inverse Distance Weighted (IDW) interpolation

- The value of a point is computed as a weighted average of the other points
- Weights are defined as inverse distance:

$$\hat{Z}_j = \frac{\sum_i Z_i/d_{ij}^n}{\sum_i 1/d_{ij}^n}$$

Z_j = value at location "j"
d_{ij} = distance between locations
n = power (input parameter)

- Basic features:
  - Closer points influence more the value estimate
  - All estimates are between *min* Z_i and *max* Z_i

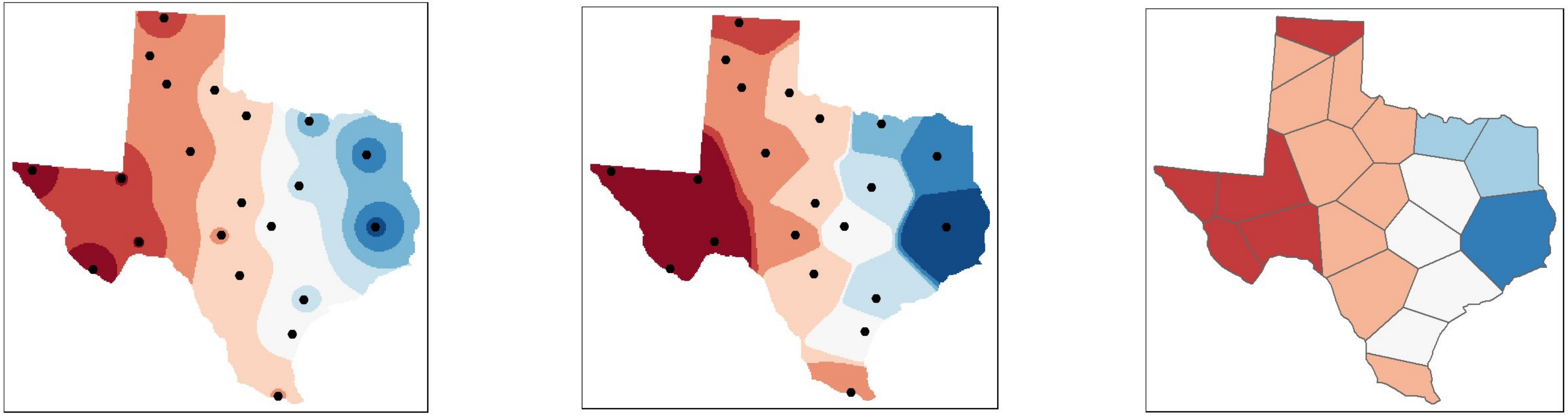
Spatial Interpolation

Inverse Distance Weighted (IDW) interpolation

- The effect of *n*
  - The higher, the more emphasis to closer points

$$\hat{Z}_j = \frac{\sum_i Z_i / d_{ij}^n}{\sum_i 1/d_{ij}^n}$$

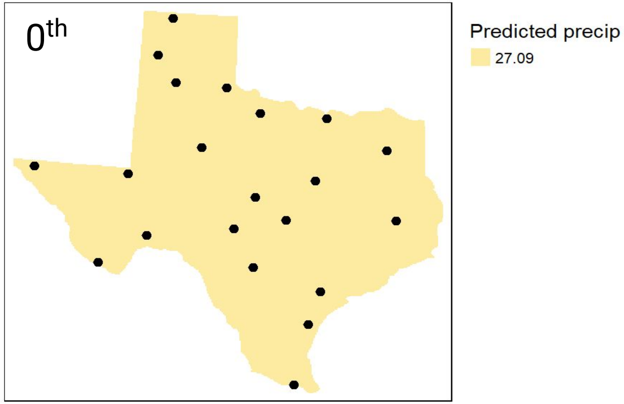
n = 2 n = 15 n → ∞

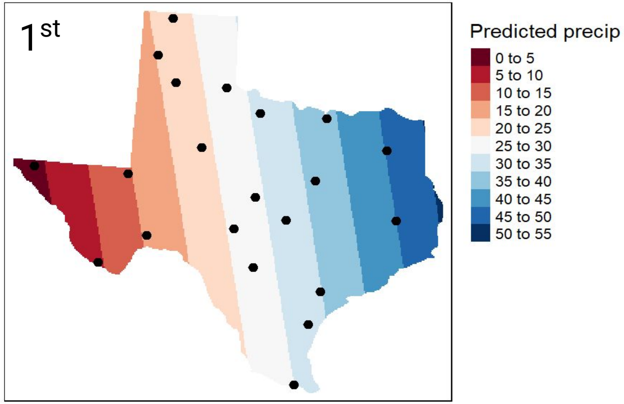


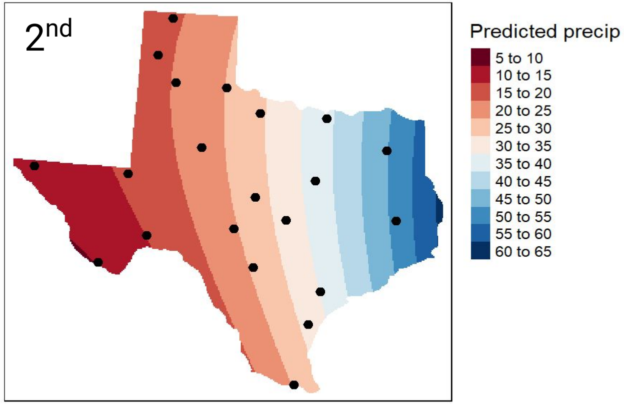
Spatial Interpolation
Surface interpolation

- Basic statistical approach: fit an equation expressing Z_i as function of its coordinates
  - The "trend surface"

- 0^{th} order trend surface: $Z = const.$
- 1^{st} order trend surface: $Z = aX + bY + const.$
- 2^{nd} order trend surface: $Z = aX^2 + bY^2 + cXY + dX + eY + const.$







Spatial Interpolation
Kriging

- As for IDW, the value of Z_i is estimated as weighted average of its neighbors
  - The difference is in how to compute weights…
- Kriging assigns to each point a different set of weights, based on local conditions
- It is a 4-step process:
  - de-trend (if needed)
  - experimental variogram
  - variogram model (inferred from experimental values)
  - interpolation

Spatial Interpolation

Kriging - step 1: de-trend

- An assumption of kriging is that data should have no trend (= constant mean)
- We can apply any trend model seen before and “subtract” it from data

original data | 1st order trend | residuals

- From now on, we work on residuals
  - At the end, Prediction = Trend + Predicted residuals

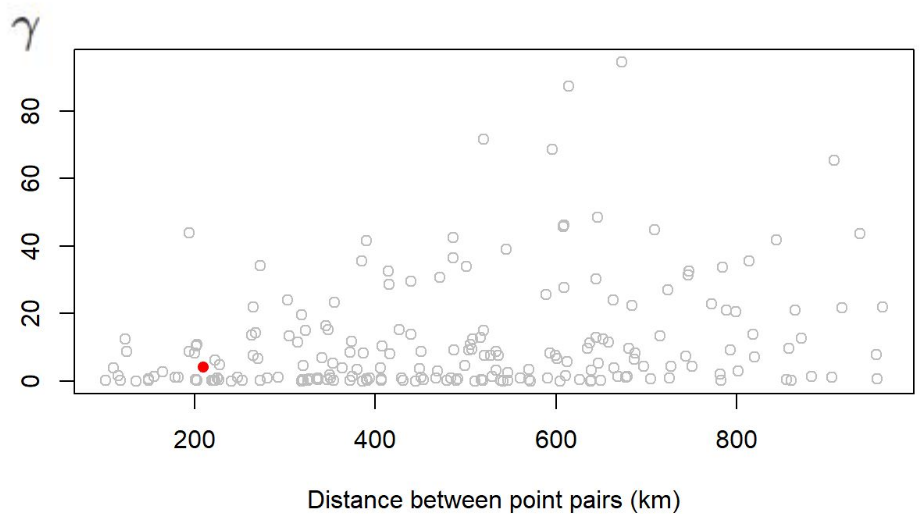
Spatial Interpolation

Kriging - step 2: experimental (semi)variogram

- For each pair of points Z_1,Z_2 in the dataset compute (semi)variance γ :

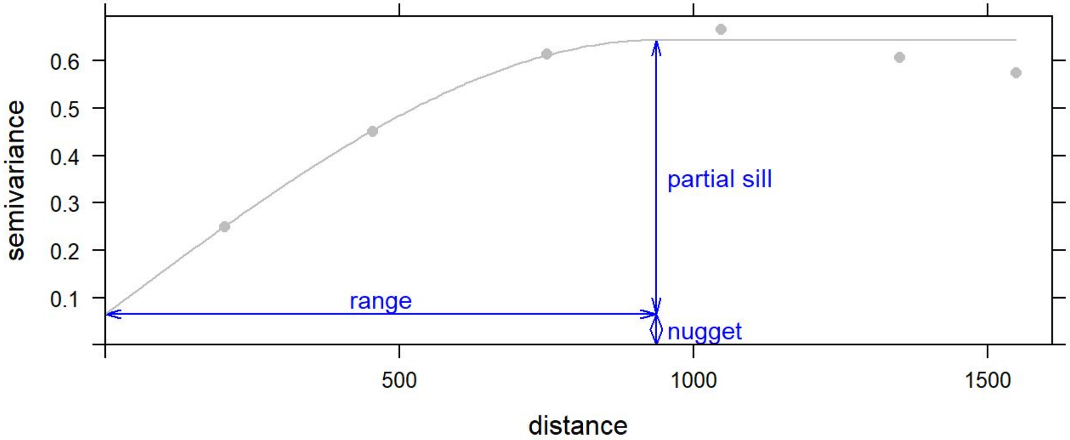
$$\gamma = \frac{(Z_2 - Z_1)^2}{2}$$

- Collecting all pairs < distance(Z_1,Z_2), γ > we obtain the experimental (semi)variogram
    - Usually simplified by binning and computing average γ in each bin



Spatial Interpolation
Kriging - step 3: (semi)variogram model

- The empirical variogram is modelled by a simple function γ(h)   (h = distance)
  - Most variograms have a general common shape



- Several variants exist, main ones: Gaussian, Linear, Spherical

Spatial Interpolation

Kriging - step 4: interpolation

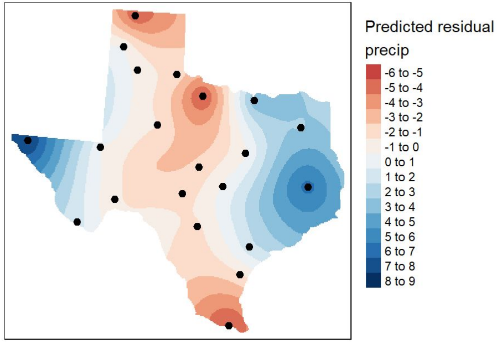
- Key question: how to compute the weights between a point and its neighbors?
- Complex approach:
  - based on solving a set of "n" equations, if we have "n" neighbors
  - Example for n=3, computing the value for point "0" from neighbors "1", "2", "3"

$$W_1\gamma(h_{11}) + W_2\gamma(h_{12}) + W_3\gamma(h_{13}) + \lambda = \gamma(h_{10})$$
$$W_1\gamma(h_{21}) + W_2\gamma(h_{22}) + W_3\gamma(h_{23}) + \lambda = \gamma(h_{20})$$
$$W_1\gamma(h_{31}) + W_2\gamma(h_{32}) + W_3\gamma(h_{33}) + \lambda = \gamma(h_{30})$$
$$W_1 + W_2 + W_3 + 0 = 1.0$$

h_{ij} = distance between points "i" and "j"

- Then, the (residual) for point "0" will be:

$$z_0 = z_1W_1 + z_2W_2 + z_3W_3$$

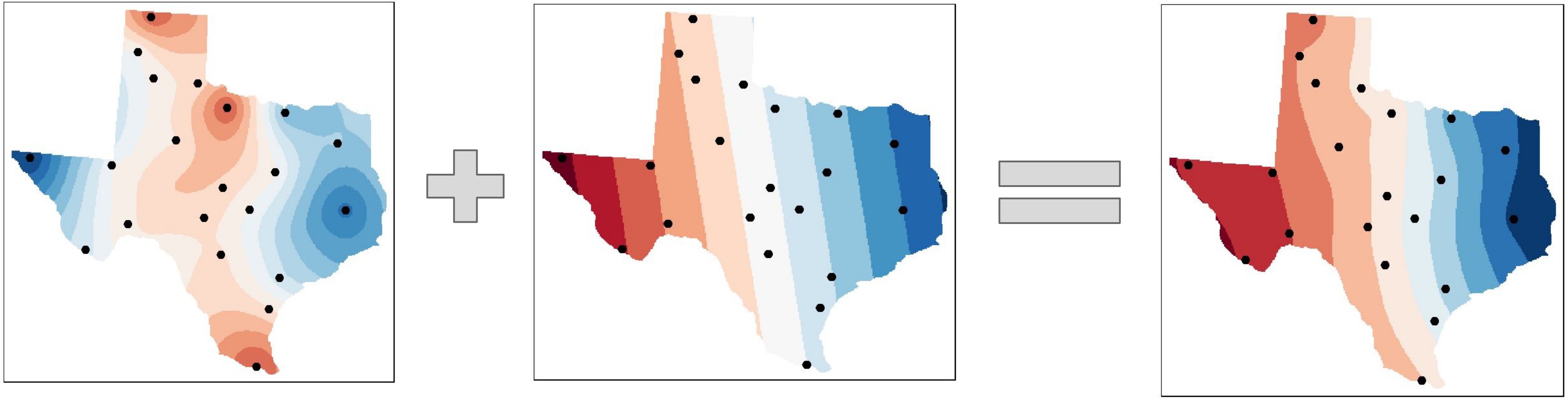


Spatial Interpolation

Kriging - step 4: interpolation

- Predicted residuals and the trend are summed up:

predicted residuals | 1st order trend | value prediction



Spatial Interpolation

Kriging: Remarks

- Most of the interpolation methods studied today have a common schema: compute a weighted average of neighbors

- They only differ by how the weights are computed:
  - Thiessen polygons: w=1 for the NN, w=0 for the others
  - IDW: 1/distance (normalized)
  - Kriging: complex system based on local variance

- Many other methods exist, mainly playing on the weights
  - e.g. Minxing Zhang, Dazhou Yu, Yun Li, and Liang Zhao. Deep geometric neural network for spatial interpolation. In SIGSPATIAL 2022. https://doi.org/10.1145/3557915.3561008
    - Machine learning approach: learns to estimate weights through an MLP based on distance and direction of neighboring points

Spatial regression

Regression vs. Interpolation

- Generalization of objectives

- The (non-spatial) attribute values of points are predicted based on
  - Predictive attributes (regressors) of the point
  - Predictive attributes of the neighbors

- The spatial component is used to link points and share their information
  - Interpolation, instead, makes predictions directly from location (the coordinates)

- Extend common regression tasks:

$$(P_i) = \alpha + \beta X_i + \epsilon_i \quad \Rightarrow \quad (P_i) = \alpha + \beta X_i + \delta \sum_j w_{ij} X'_j + \epsilon_i$$

(Standard regression model) (Spatial regression model)

Spatial Regression

Spatially lagged exogenous regressors

- “Spatial lag” of variable X w.r.t. point “i” : (weighted) average of X over “i”s neighbors
    - Basically, a spatial interpolation of X

summary information from “i” ’s neighbors

$$(P_i) = \alpha + \beta X_i + \delta \boxed{\sum_j w_{ij} X'_j} + \epsilon_i$$

target variable at point “i”

local predictive variable at “i”

exogeneous predictive variable

- Weights w_{ij} are a parameter, as in interpolation methods

Spatial Regression

Spatially lagged endogenous regressors

- Integrate the spatial dependence among target values of neighboring points
  - Basically, a spatial interpolation component over the target variable

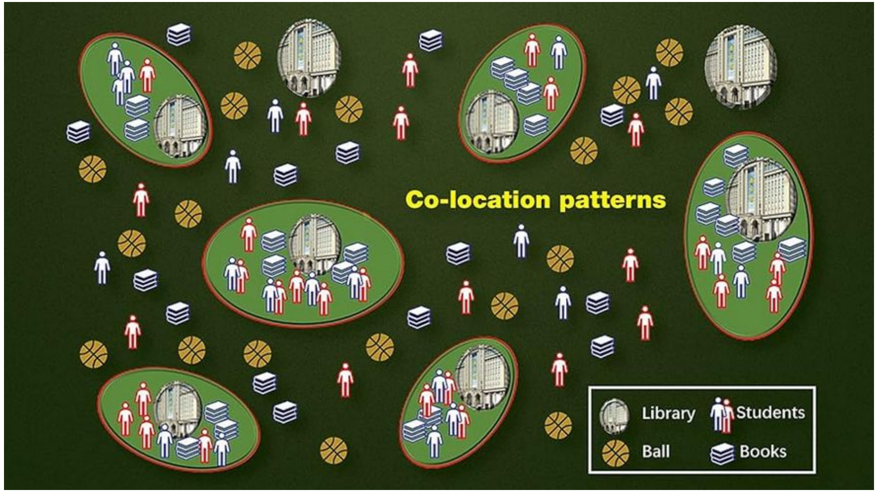
$$\underbrace{(P_i)}_{\text{target variable at point “i”}} = \alpha + \lambda \sum_j \underbrace{w_{ij}}_{\text{dependence w.r.t. neighbors’ targets}} \cdot (P_j) + \underbrace{\beta X_i}_{\text{local predictive variable at “i”}} + \epsilon_i$$

- IMPORTANT: it means predictions are inter-related, it is not a recursive function
  - We cannot apply simple interpolation → more complex methods are needed (omitted here)

Spatial associations: co-location patterns

Co-location Patterns

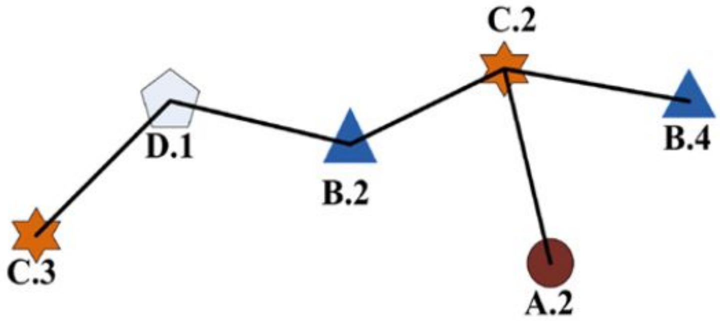
- Similar to frequent pattern analysis
  - “find sets of items that occur together in several transactions”
- Items are replaced by spatial points
  - Issue: what is a “transaction” ?
  - Answer: any set of points that are close to each other
- The concept of frequency needs to be revisited
  - a point/item might participate to multiple instances of the same pattern



“Data Mining for Co-location Patterns: Principles and Applications”, Guoqing Zhou.
https://www.amazon.it/Data-Mining-Co-location-Patterns-Applications/dp/0367688662

Co-location Patterns

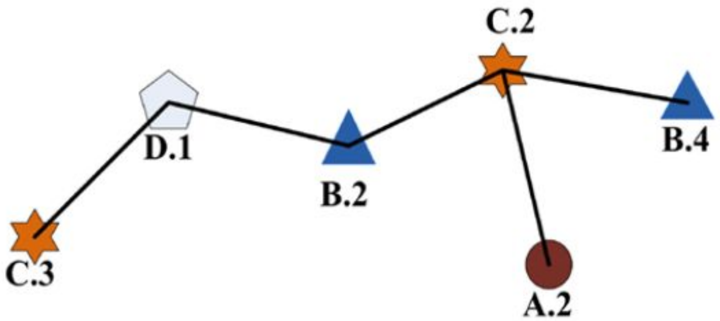
- Given
  - A set of **features (types) F**= $\{f_1, .., f_m\}$
    - E.g. F = { restaurant, bar, hotel, barber, supermarket }
  - A set of **spatial instances O** = $\{o_1, …, o_n\}$ of features F
    - E.g.: O = { restaurant#1, bar#1, bar#2, bar#3, hotel#1, barber#1, barber#2, supermarket#1, supermarket#2}
  - A **neighbor relation R** between pairs of instances
    - I.e. $R(o_1,o_2) \Leftrightarrow distance(o_1,o_2) < threshold$ (or equivalent)



F = {circle, triangle, pentagon, star}
O = {A.2, C.2, C.3, … , D.1}
R(C.3, D.1) = True
R(D.1, B.2) = True
R(C.3, B.2) = False
…

Co-location Patterns

- Definitions
  - A **co-location pattern CL** = $\{f_1, \ldots, f_k\}$ is a subset of features, i.e. CL $\subseteq$ F
    - The aim is to find those where the features appear together very often
  - An **instance I** = $\{o_1, \ldots, o_k\}$ of pattern CL is a subset of O (namely, I $\subseteq$ O) such that
    - for each feature f $\in$ CL there is exactly one instance o $\in$ I of type f, and viceversa
    - I forms a clique w.r.t. R, i.e. $o_1, o_2 \in I \Rightarrow R(o_1, o_2)$



Example:
- CL = {triangle, star}
- I = {B.2, C.2} or {B.4, C.2}

I = {B.2, C.3} is not an instance
I = {C.3, D.1} is not an instance

Co-location Patterns

- Pattern quality measures
  - **Participation Ratio** of a feature within a pattern

how many different instances of f_i appear in the set of instances of Cl

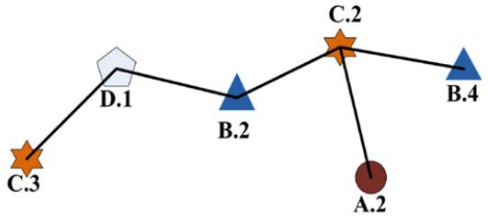
$$PR(Cl, f_i) = \frac{|\pi_{f_i}(table\_instance(Cl))|}{|table\_instance(\{f_i\})|}$$

how many different instances of f_i are in the whole dataset

  - **Participation Index** of a pattern

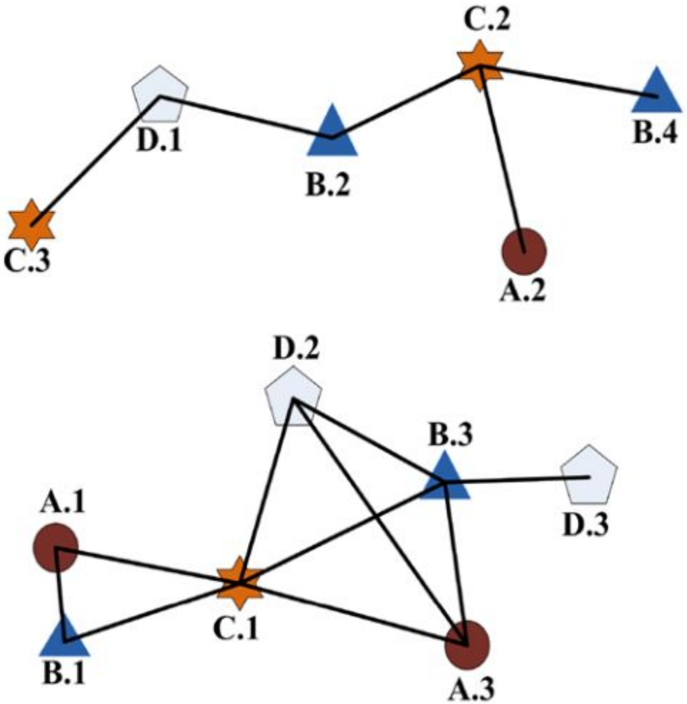
$$PI(Cl) = \min_{i=1}^{k}\{PR(Cl, f_i)\}$$

Example:
- PR({B,C}, B) = 2 / 2
- PR({B,C}, C) = 1 / 2
- PI({B,C}) = 1/2



Co-location Patterns

A larger example



Size-2

A	B
1	1
3	3
2/3	2/4

PI({A,B})

A	C
1	1
2	2
3	1
3/3	2/3

A	D
3	2
1/3	1/3

B	C
1	1
2	2
3	1
4	2
4/4	2/3

B	D
2	1
3	2
3	3
2/4	3/3

C	D
1	2
3	1
2/3	2/3

Size-3

A	B	C
1	1	1
3	3	1
2/3	2/4	1/3

A	B	D
3	3	2
1/3	1/4	1/3

A	C	D
3	1	2
1/3	1/3	1/3

Size-4

A	B	C	D
3	3	1	2
1/3	1/4	1/3	1/3

Spatial Trends Detection

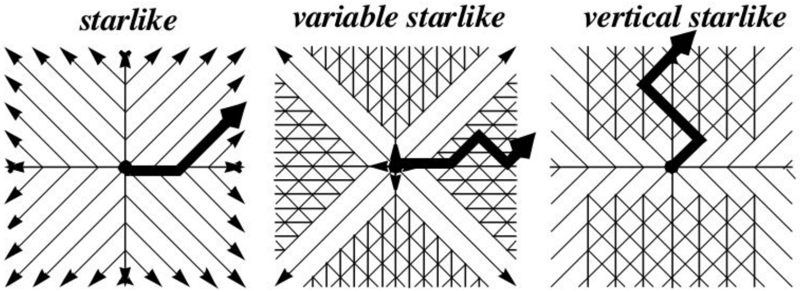
Spatial Trends

- Idea: extend the concept of trends in time series
  - A sequence of points having a (non-spatial) attribute that changes following a trend

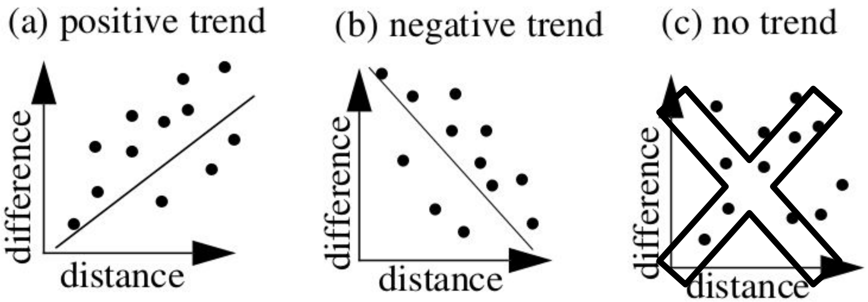
- The linear direction of time is replaced by many possible paths in space

- Focus on paths that
  - Start from a common location (e.g. the center of a city)
  - Have meaningful shapes (e.g. quasi-straight lines, not random walks)
  - Show a (statistically) significant trend

Sample path shapes allowed



Trends need an high correlation



Spatial Trends
Definition

- Let **g** be a **neighborhood graph**
  - paths move from an object to one of its neighbors

- Let **o** be an object in g
  - this is the starting point of paths

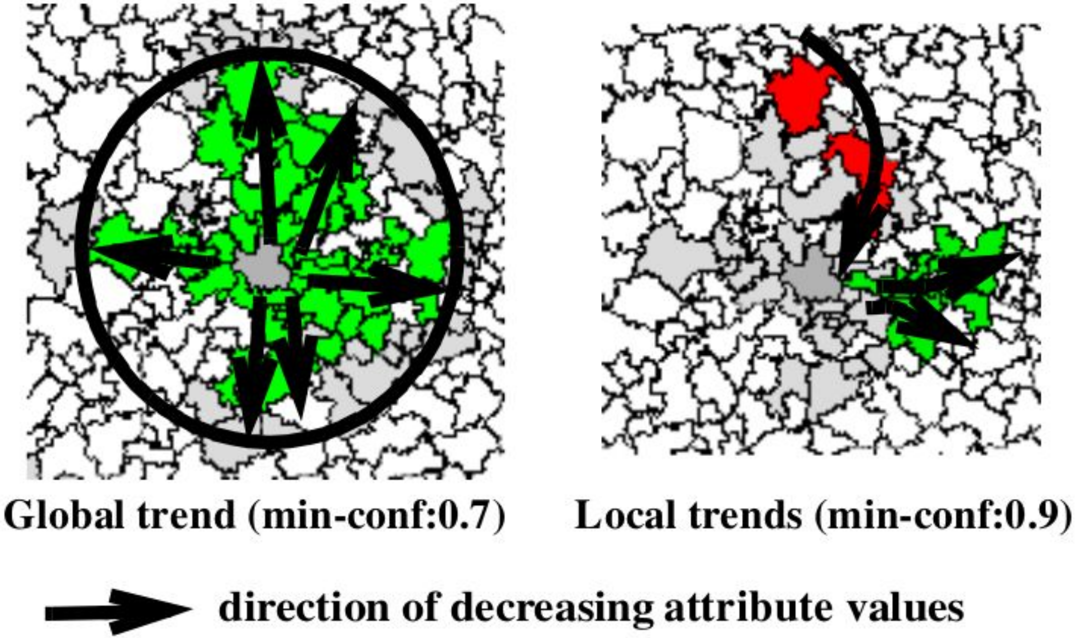
- Let **a** be a subset of all non-spatial attributes
  - this is where we search for trends

- Let **t** be a type of function, e.g. linear or exponential, used for the regression

- The task of **Spatial Trend Detection** is to discover the set of all neighborhood paths in g starting from o and having a trend of type t in attributes a with a correlation of at least *min-conf*

Spatial Trends
Examples

- Global negative trend of variable "average rent" from the Regensburg city center
- A few other single, local trends
  - Less surprising, yet correlation is higher



Global trend (min-conf:0.7) Local trends (min-conf:0.9)

Food for thought

- **Co-location Patterns**: why not to use just the frequency of patterns? (Namely, number of instances of the colocation pattern)

- **Spatial Trends**: why not to just take the peak values – maybe after a spatial interpolation, if needed? The other values around them will obviously follow a decreasing trend

- **Spatial Classification** [yes, it is outside the program of this course]: let say we have a training set of polygons of buildings, each associated to the class “public building” or “private building”. Then we have a set of polygons of other buildings of unknown label, which we would like to classify as public or private. How would you do that?

to study for the exam

Material

- [book chapter] Introduction to geographic information systems, Kang-Tsung Chang, McGraw-Hill
  - Chapter 15: Spatial Interpolation

- [book chapter] Intro to GIS and Spatial Analysis, Manuel Gimond, online: https://mgimond.github.io/Spatial
  - Chapter 14: Spatial Interpolation

- [book chapter] Spatial data science for sustainable development, Henrikki Tenkanen, online: https://sustainability-gis.readthedocs.io/en/latest/
  - Tutorial 3: Spatial Regression in Python

to study for the exam

Material

- [paper] A MapReduce approach for spatial co-location pattern
- mining via ordered-clique-growth, Yang-Wang-Wang, 2020 https://doi.org/10.1007/s10619-019-07278-7
    - Section 3.1: Co-location pattern mining

- [paper] Algorithms for Characterization and Trend Detection in Spatial Databases, Ester-Frommelt-Kriegel-Sander https://www.lri.fr/~sebag/Examens/Ester_KDD98.pdf
    - Section 4: Spatial Trend Detection
    - Have a quick look also to the rest of the paper