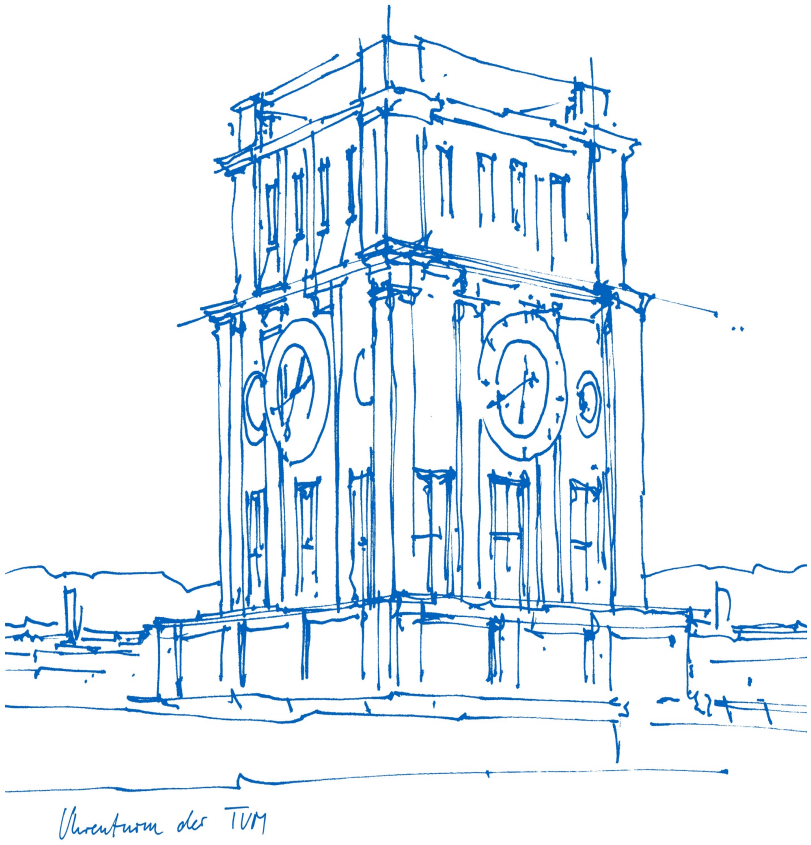
Geoinformation

Prof. Dr.-Ing. Liqiu Meng

Chair of Cartography

WS 24/25



Contents of Lectures

L1: Introduction

L2: Spatiotemporal representations and databases

L3-L4: Spatial data analysis

L5: Cartographic techniques

Spatial data analysis

I. Data quality

II. Pitfalls of spatial data analysis

III. Types of spatial data analysis

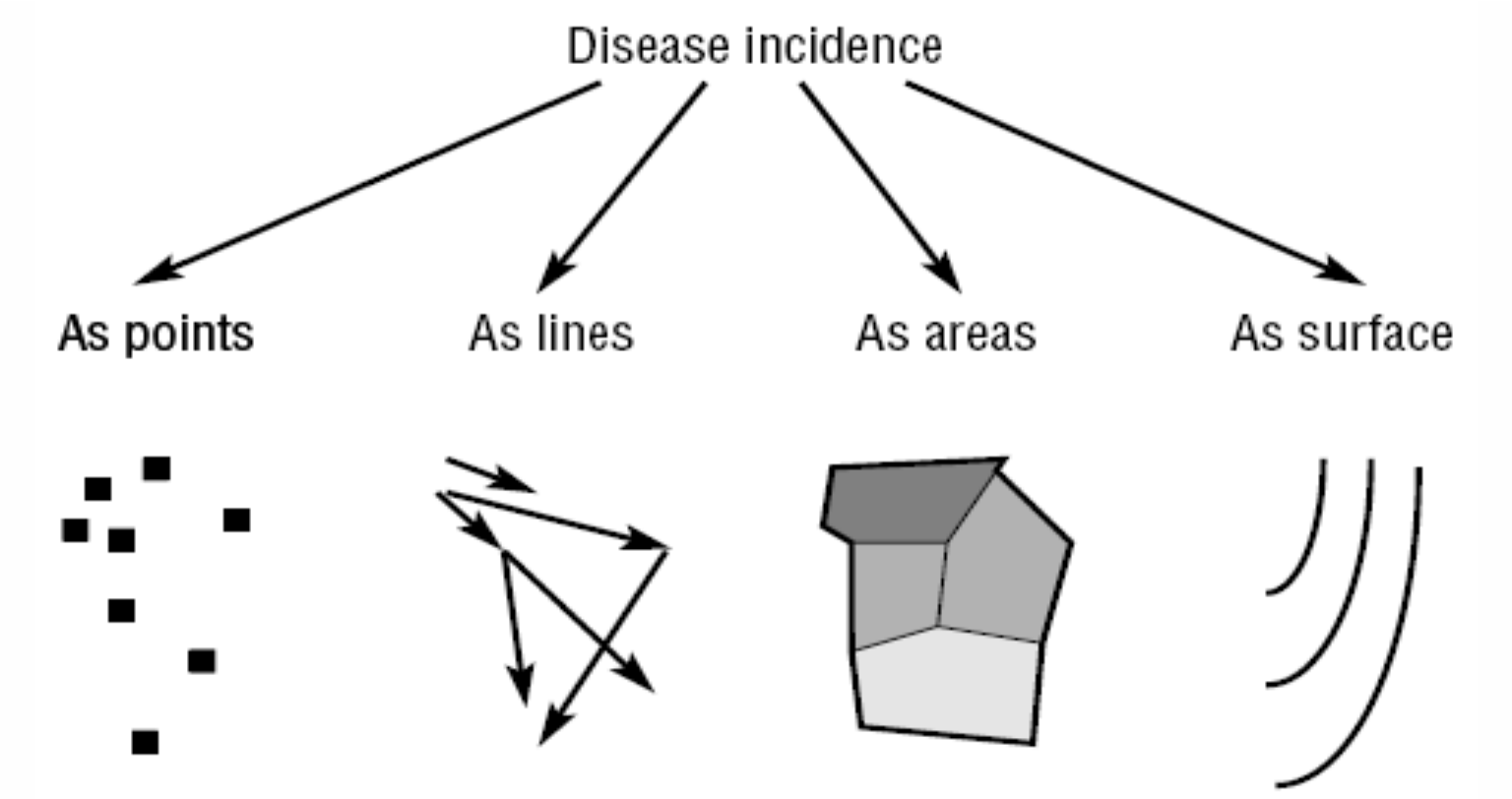
Purpose of spatial data analysis

Find, explore and express stored and hidden dependences between spatiotemporal observations

$$\{Y(s_1), Y(s_2), \ldots Y(s_i), \ldots\}$$

observations at spatiotemporal location s_1, s_2, …s_i

The spatiotemporal locations can be points, lines, areas or continuous surfaces.

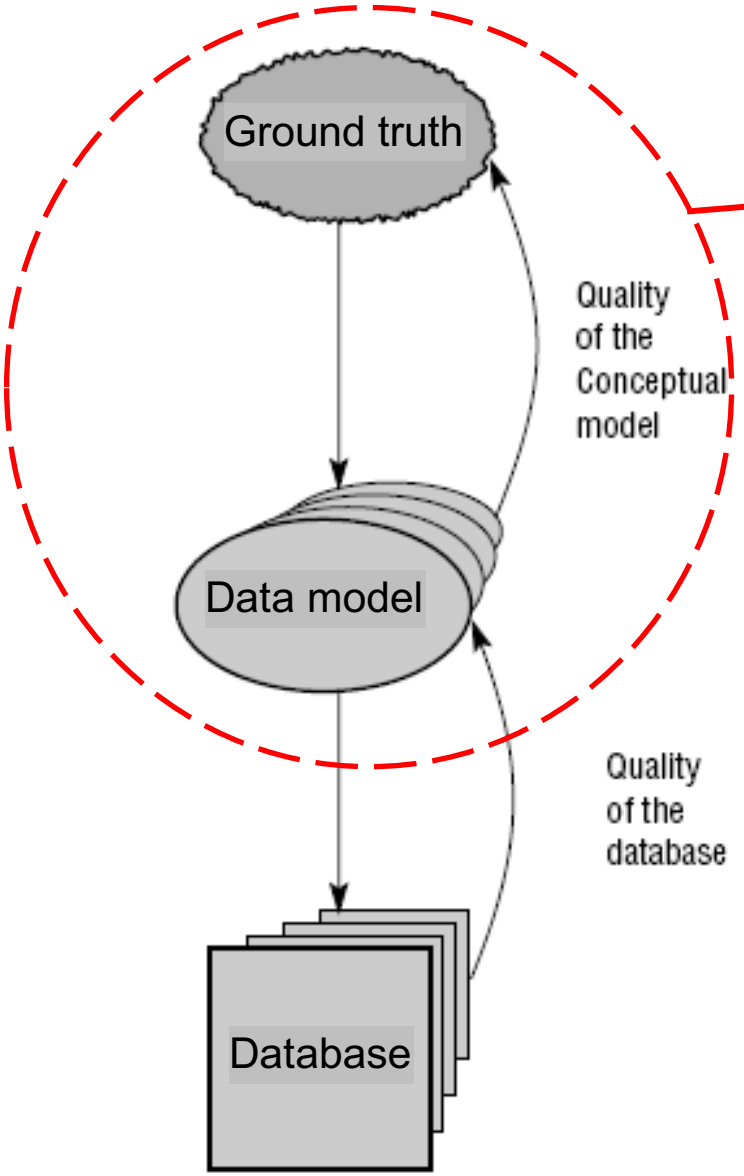

Disease incidence
As points
As lines
As areas
As surface

I. Data quality

Quality assurance serves to facilitate dissemination and sharing of data. It provides a good starting point for the spatial data analysis.

Data quality can be measured separately for space, time, and theme, although these measurements are not independent.

Data quality is assessed against a reference model (representing the ground truth). Different models may be used.



Data model provides standard definitions of data elements, aggregation level, data transfer format and metadata about database contents.

Components of data quality

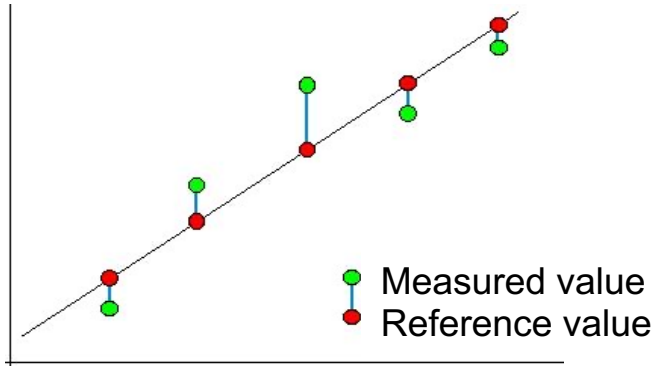
1. Accuracy

It is the agreement between measured values and the reference value.

- Spatial accuracy
  Positional accuracy can be reflected by the Euclidean distance between observed and reference value.

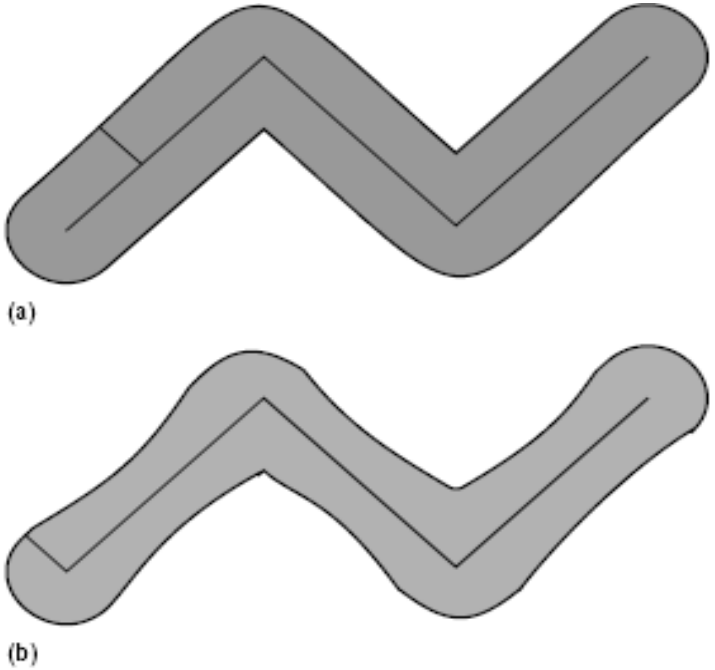
Root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$



Errors of a line can be defined by the errors of points on the line, or by a buffer around a reference line within which the observations are made.

The error distribution along a line might be non-uniform in shape.



- Temporal accuracy
  The agreement between measured and reference temporal value, i.e. the degree to which a database is up-to-date.

- Thematic accuracy
  For quantitative values, it can be calculated in a similar way to spatial accuracy. For categorical values, a comparison of the observed category with the reference category can be made.

  Errors of omission (not assigning a location to any category)

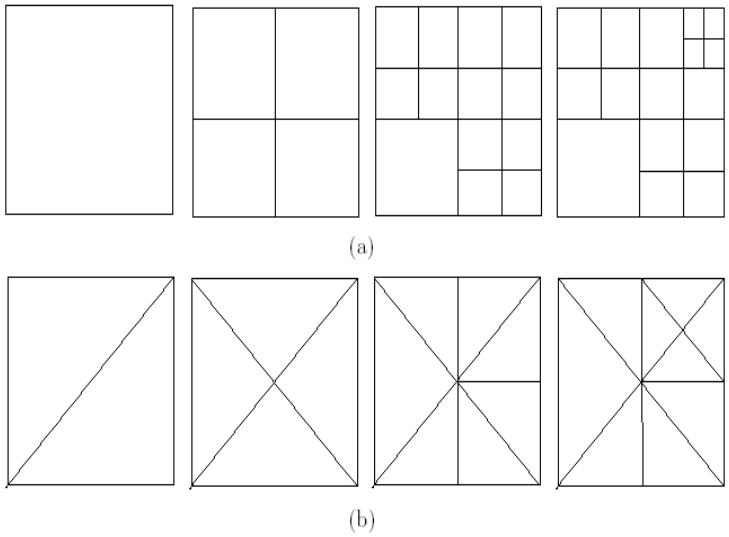
  Errors of commission (assigning a location to a wrong category)

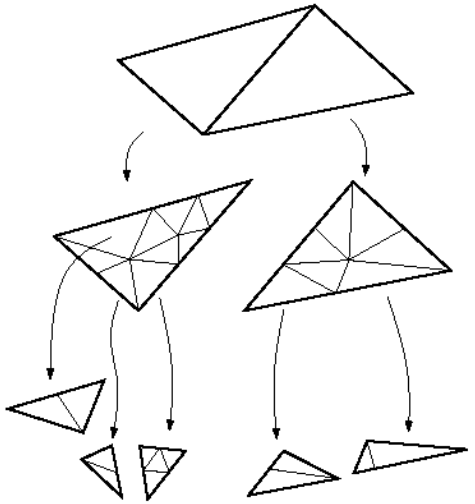
2. Precision / resolution / level of detail

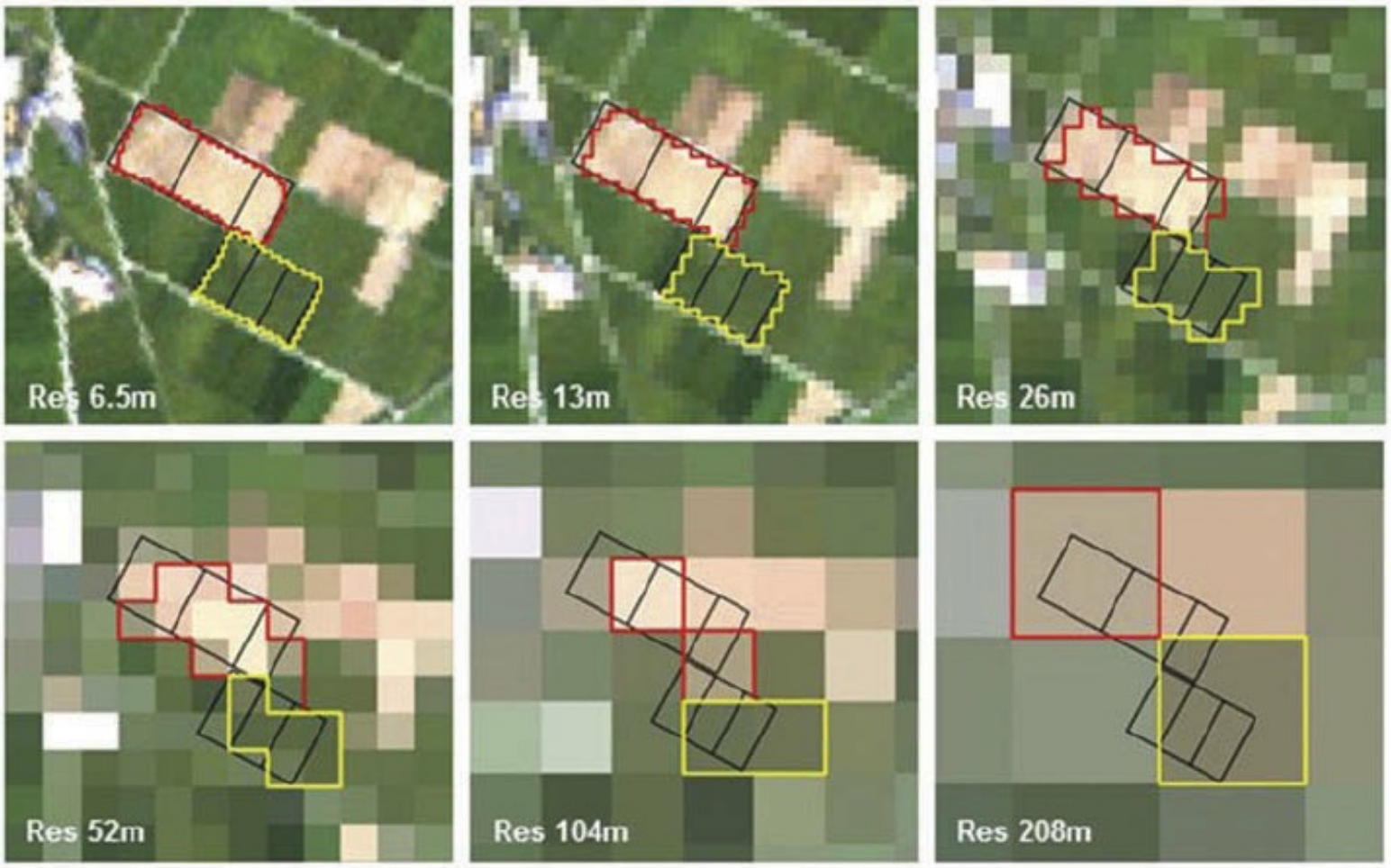
It refers to the amount of detail that can be discerned. Low resolution does not have the same negative connotation as low accuracy.

- Spatial resolution

  The minimum size of objects on the ground that can be discerned. For vector data, the smallest discernable feature is related to map scale.





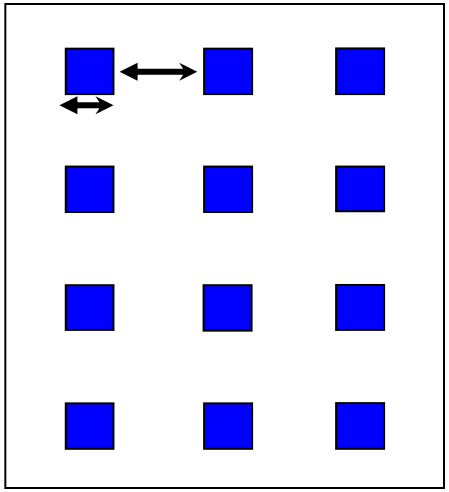


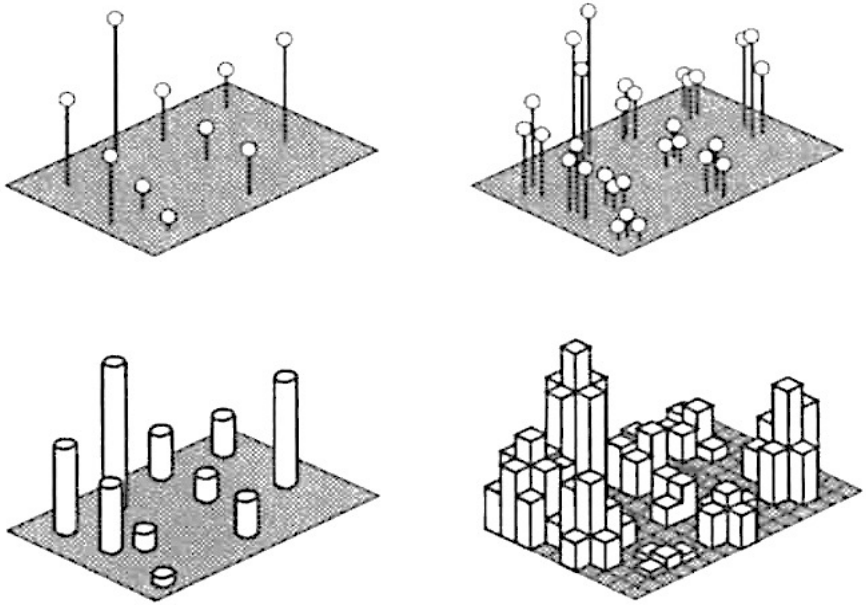
Different image resolutions

Finer resolutions may be obtained from redundant coarser resolutions

Spatial resolution refers to the pixel size

Spatial sampling rate refers to the spaces between pixels





- Temporal resolution

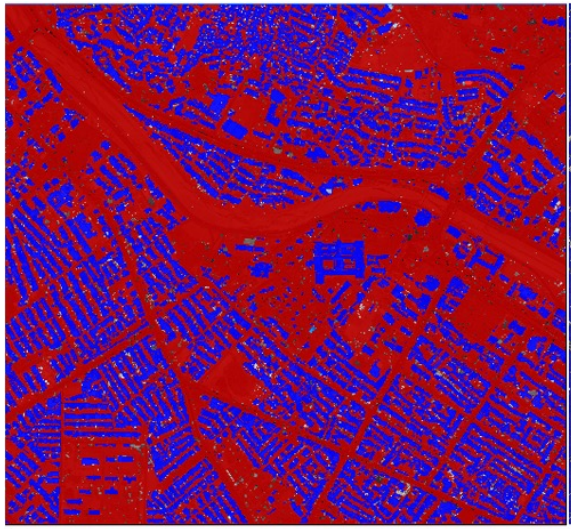
It refers to the minimum duration for the recording of the spectral reflection of a frame, e.g. 1/10s for one frame.

Temporal sampling rate refers to the frequency of repeat coverage, e.g. 24 frames per second.

- Thematic resolution

For quantitative values, it is determined by the precision of the measurement device. For categorical value, it is defined as the fineness of category.

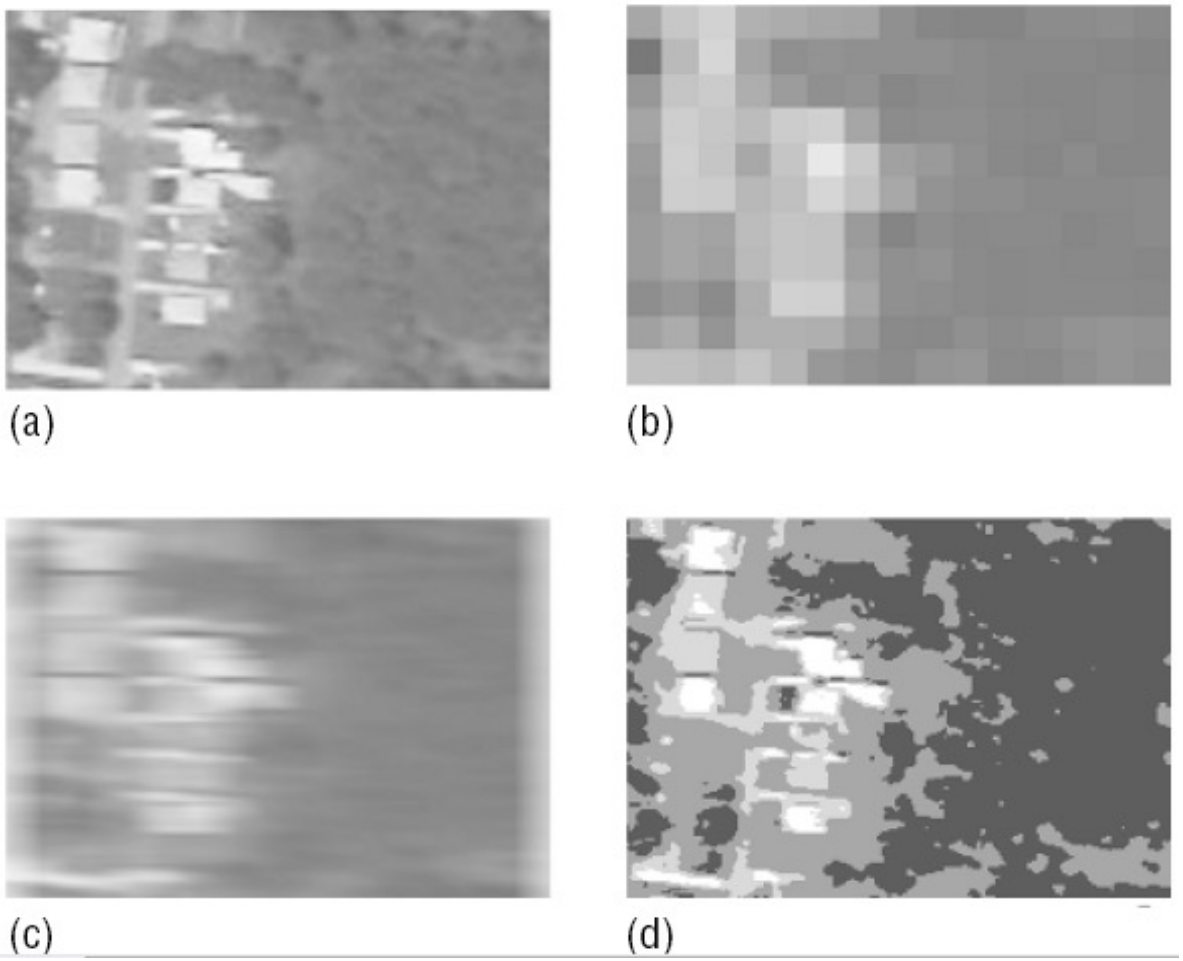
For example, land cover classification systems can illustrate fine or coarse boundaries of land cover classes on the spectral reflectance data.



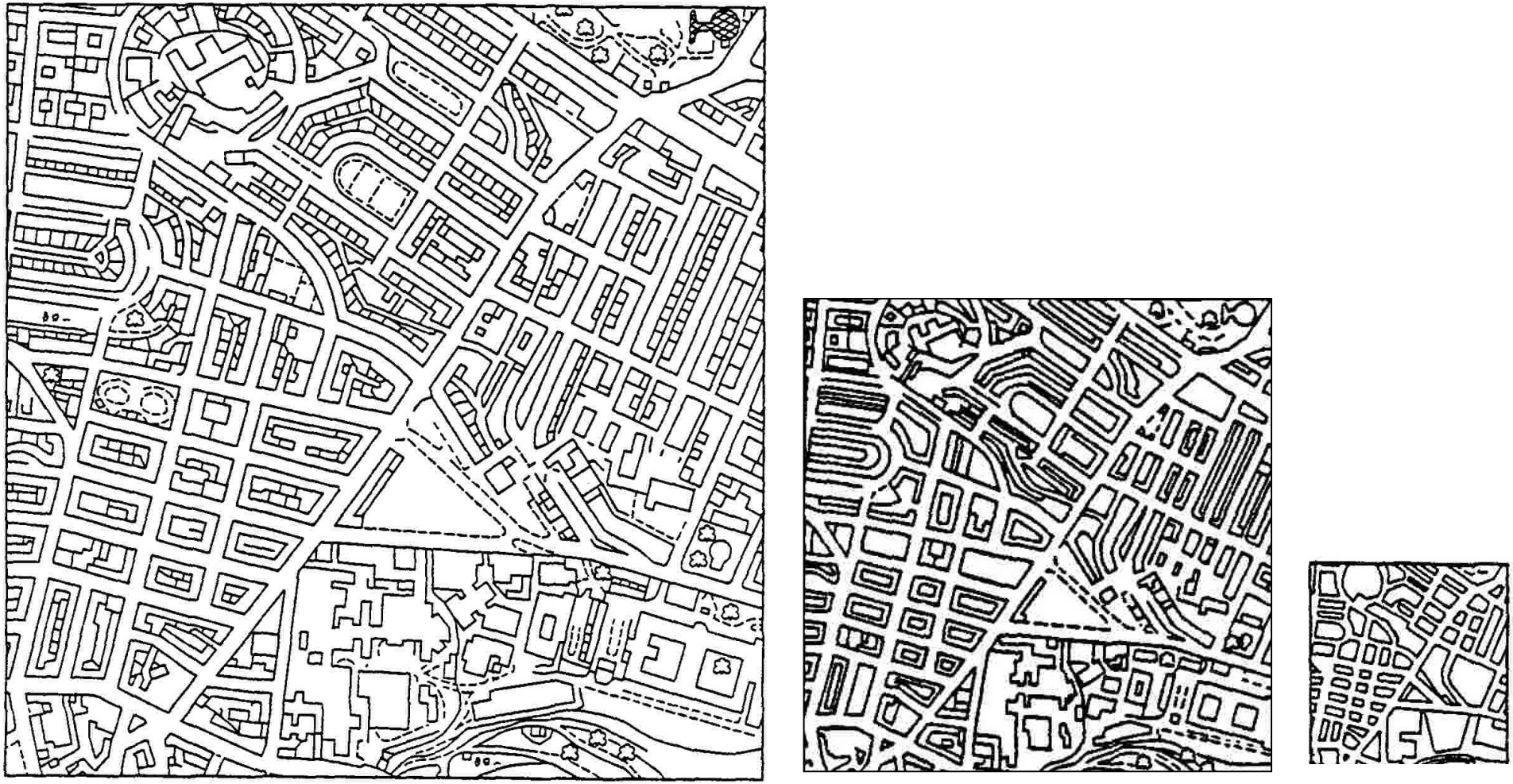
building / non-building

building / vegetation / road / open space

finer classes of building / vegetation / road / open space



(a) an image with spatial resolution of 1m, temporal resolution of 1/30 s, and thematic resolution of 8 bits
(b) the effect of degraded spatial resolution
(c) the effect of degraded temporal resolution
(d) the effect of degraded thematic resolution



Both spatial and thematic resolutions are reflected in map scales

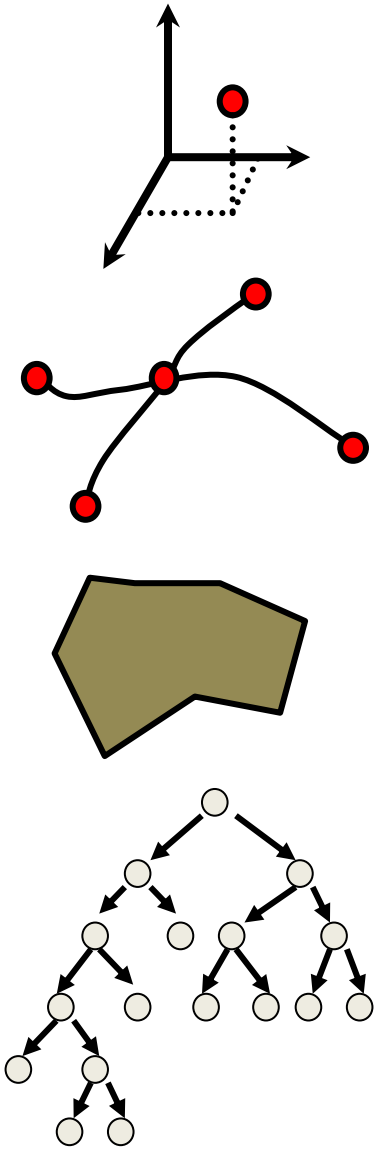
3. Consistency

It is a degree at which the data are free of conflict.

Spatial consistency, e.g. one class at a given location, lines must intersect at nodes, polygons are bounded by lines

Temporal consistency, e.g. a seat is occupied only by one person at one time

Thematic consistency, e.g. population P, mean household size S_m, and total number of households T have the relation $P = S_m \cdot T$



4. Completeness

Data completeness: a measurable error of omission between the observation and the reference.

Attribute completeness: the degree to which all relevant attributes of a feature have been encoded.

Value completeness: the degree to which values are present for all attributes.

Geospatial data specifications of Germany

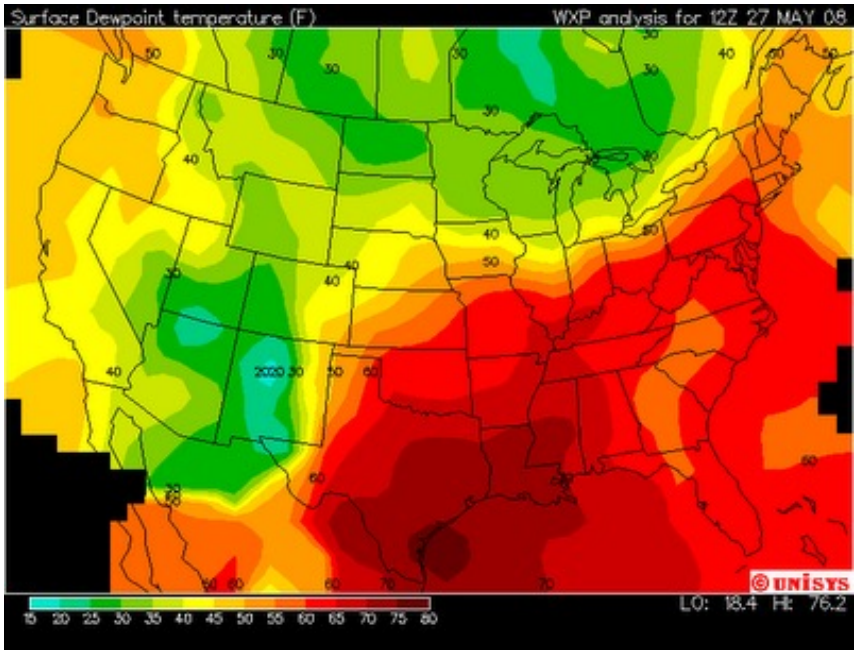
II. Pitfalls of spatial data analysis

Methods from classic statistics are not directly applicable to spatial data analysis due to the pitfalls

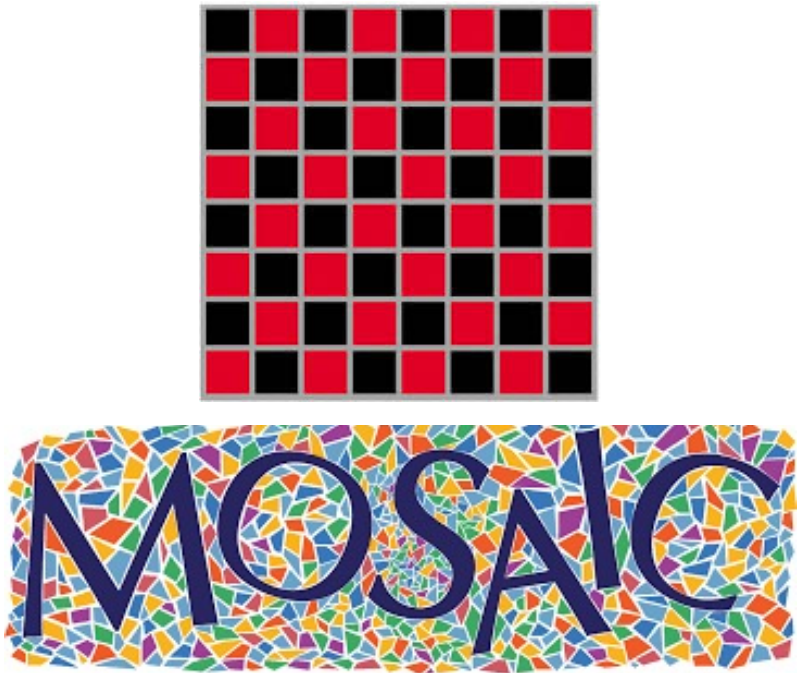
1. Spatial autocorrelation
2. Modifiable areal unit problem
3. Ecology fallacy
4. Scale
5. Non-uniformity of space
6. Edge effect

Pitfall 1 - Spatial autocorrelation

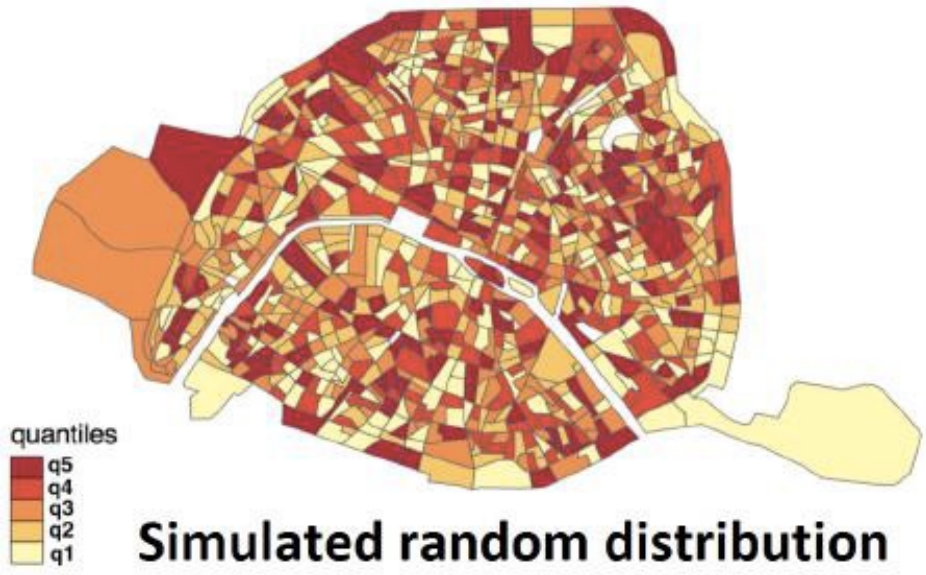
Tobler's first law of geography - Everything is related to everything else. But near things are more related than distant things.

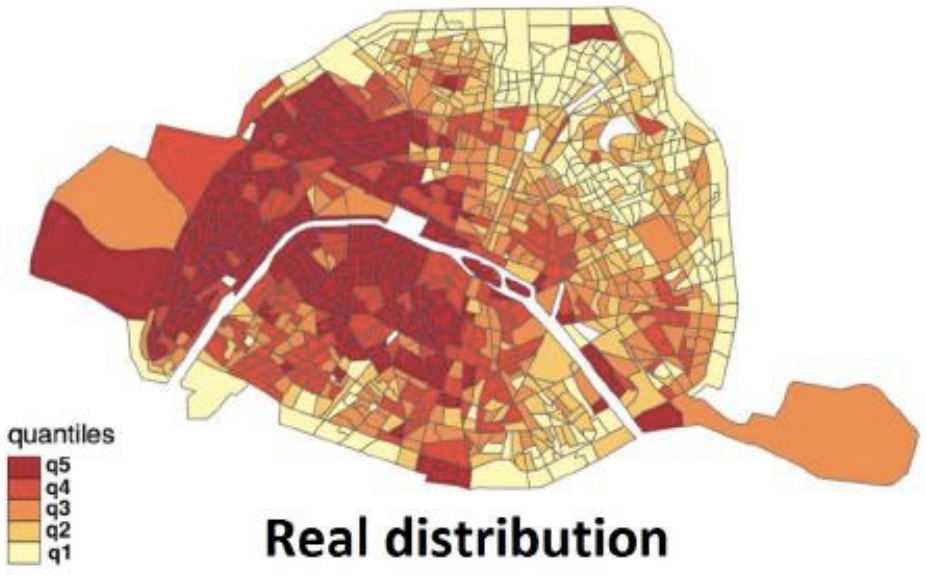


effect of autocorrelation



no autocorrelation





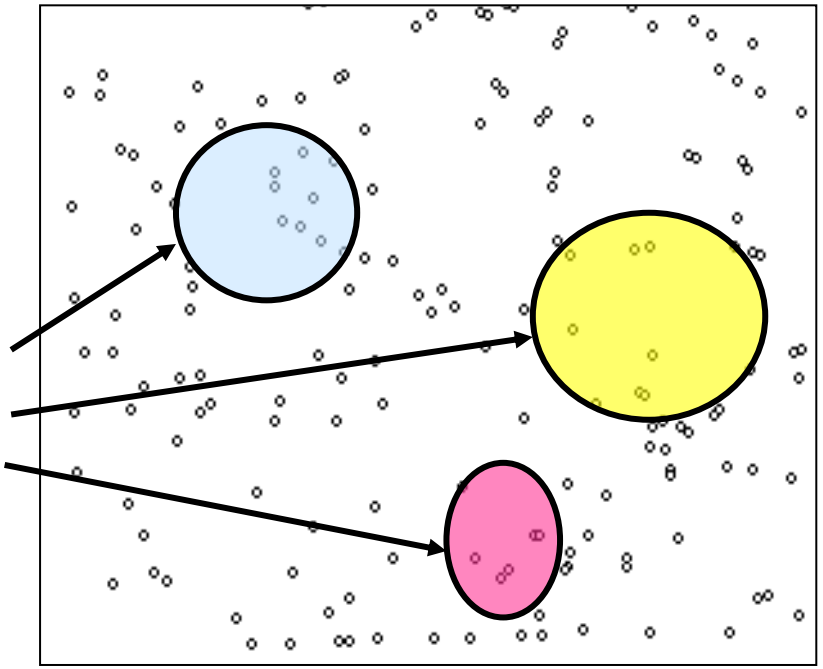
Source: INSEE, Localised Tax Revenues System (RFL)

Parisian census districts: random distribution vs autocorrelated distribution

Spatial variations are caused by

- local conditions at different places
- autocorrelations at the same place

Local conditions with varying spillover boundaries



How do you explain similar environments at very different places?

Spatial autocorrelation measures

Moran's index

I = 0 (independent)

I > 0 (positive autocorrelation)

I < 0 (negative autocorrelation)

$$I = \frac{N\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

numerical values of object at location i and j

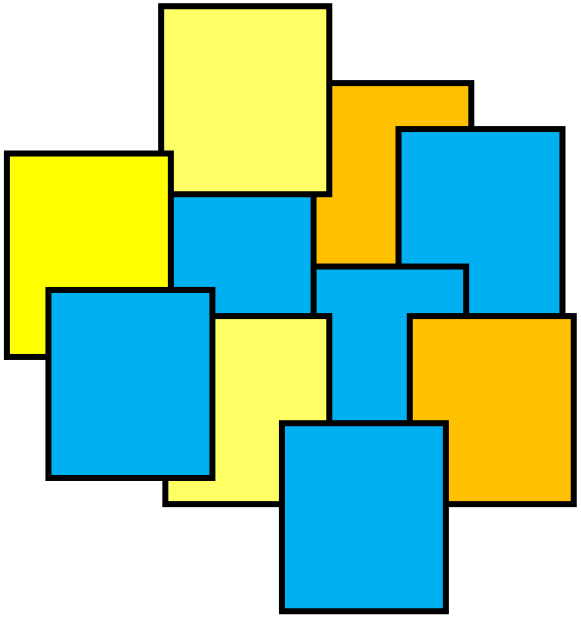
weight (e.g. spatial similarity)

Geary's Coefficient

C = 1 (independent)
C < 1 (positive autocorrelation)
C > 1 (negative autocorrelation)

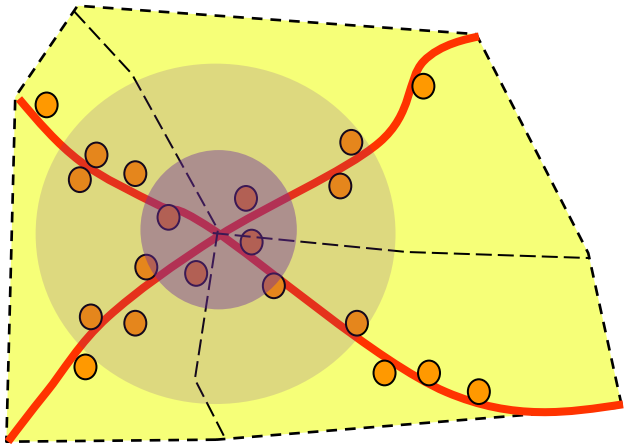
$$C = \frac{(N-1)}{2(\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij})} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i - x_j)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Join-count statistics for nominal variables - They are based on counting the numbers of same-quality and different-quality joins between neighbors.



Pitfall 2 - Modifiable Areal Unit Problem

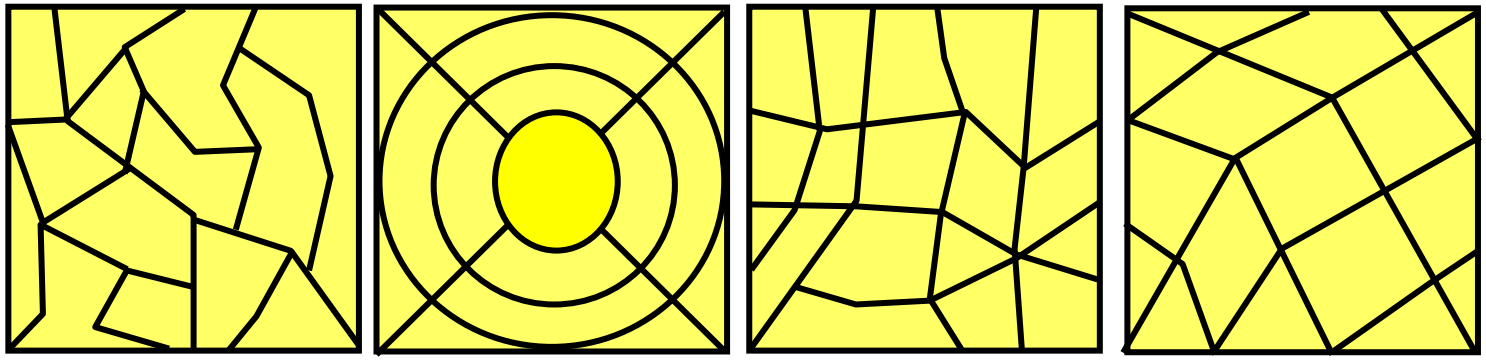
Geographic data are often aggregates of observations within spatial units of varying shape and size. Units such as census, traffic analysis buffer zone, school district are forced instead of being natural.



National census, for example, is conducted at the household level but aggregated into various administration units (streets, neighborhoods, risk zones…)

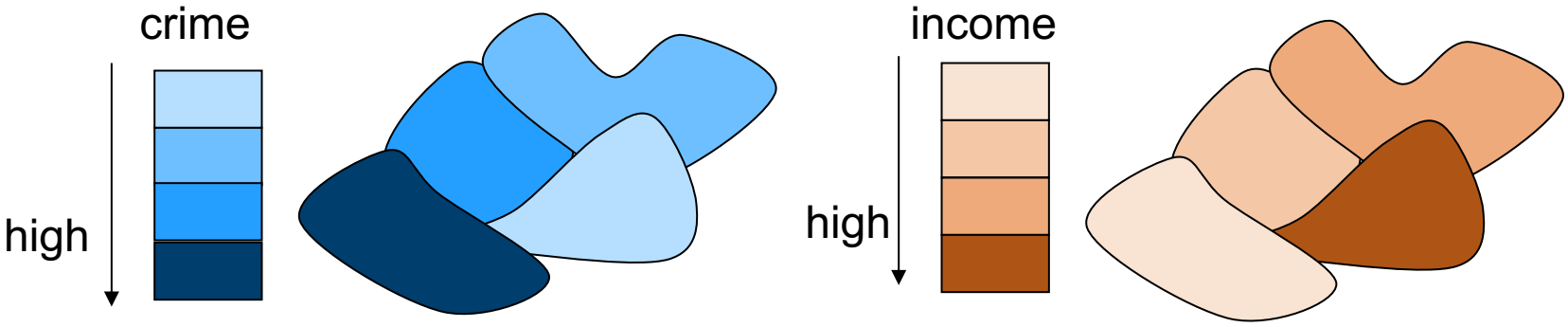
– Modified areal units lead to modified analytical results.

– The larger the modification, the larger the deviation.



Pitfall 3 - The Ecological Fallacy

An error that can occur when we make an inference about an individual based on aggregated data for a group. This fallacy assumes that all members of a group exhibit characteristics of the group at large.



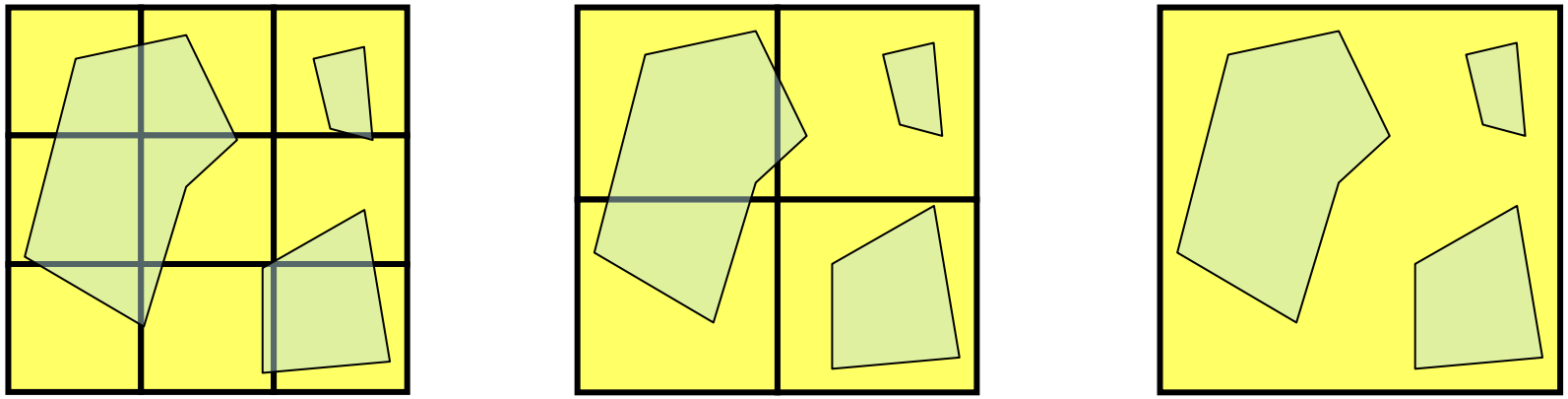
e.g., we observe a strong correlation between income and crime at the county level. If we conclude that a lower-income person is more likely to commit crime, we are falling for the ecological fallacy, also termed as confirmation bias (stereotyping, prejudice).

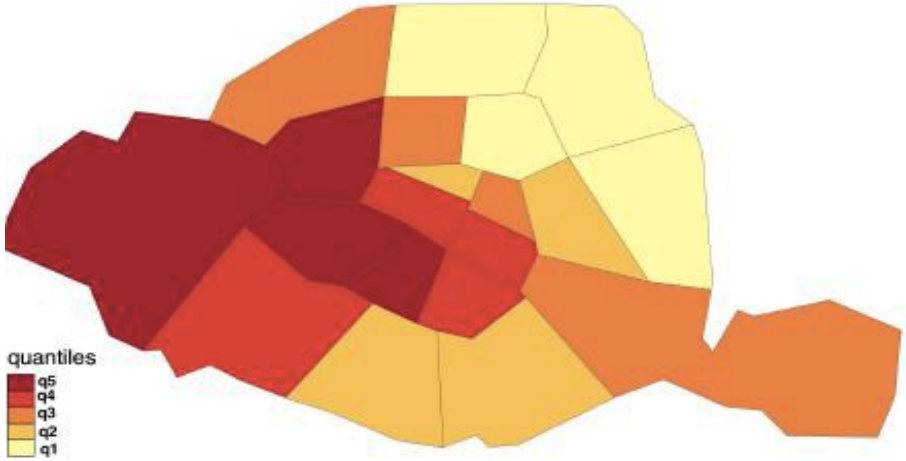
Related references:

- Daniel Kahneman: Thinking, fast and slow.
- Daniel Kahneman, Olivier Sibony, Cass R. Sunstein: Noise

Pitfall 4 – Scale

Spatial entities can be best observed in a certain scale range.
Objects may only be partially observed at a too fine scale.
Objects may be aggregated at a too coarse scale.
Domain knowledge helps to determine an optimal scale range.



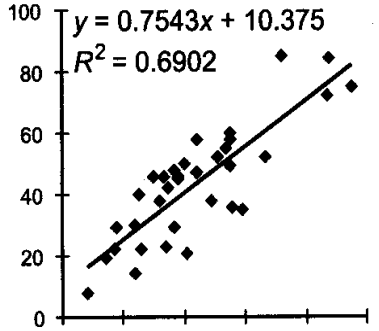


Source: INSEE, Localised Tax Revenues System (RFL)

Spatial units at different hierarchical levels

Independent variable					
87	95	72	37	44	24
40	55	55	38	88	34
41	30	26	35	38	24
14	56	37	34	8	18
49	44	51	67	17	37
55	25	33	32	59	54

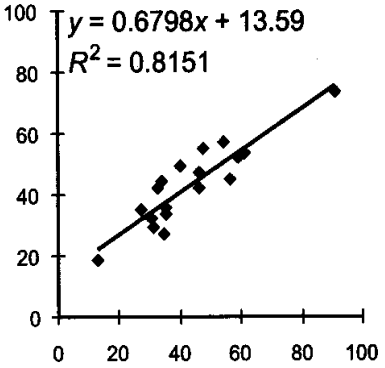
Dependent variable					
72	75	85	29	58	30
50	60	49	46	84	23
21	46	22	42	45	14
19	36	48	23	8	29
38	47	52	52	22	48
58	40	46	38	35	55



Aggregation scheme 1

91	54.5	34
47.5	46.5	61
35.5	30.5	31
35	35.5	13
46.5	59	27
40	32.5	56.5

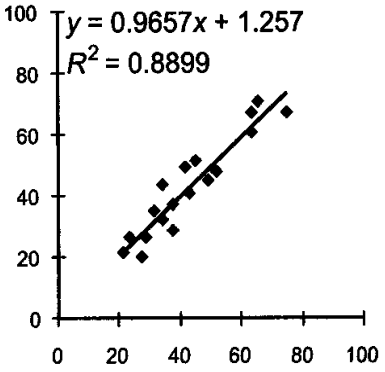
73.5	57	44
55	47.5	53.5
33.5	32	29.5
27.5	35.5	18.5
42.5	52	35
49	42	45



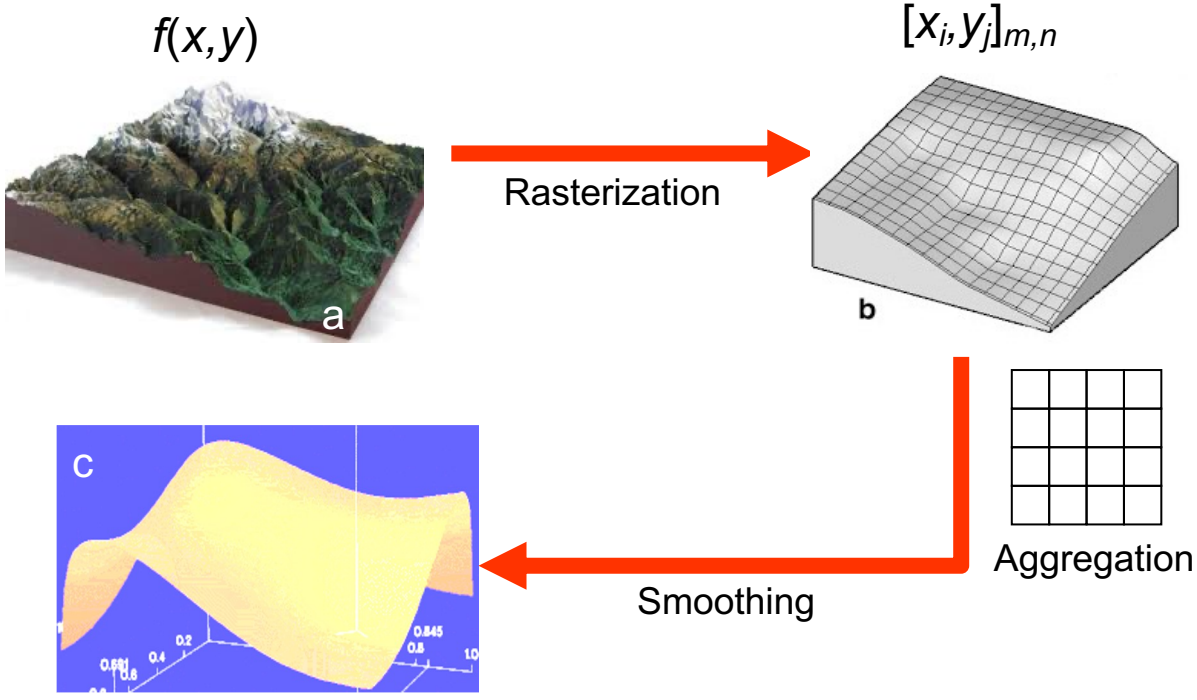
Aggregation scheme 2

63.5	75	63.5	37.5	66	29
27.5	43	31.5	34.5	23	21
52	34.5	42	49.5	38	45.5

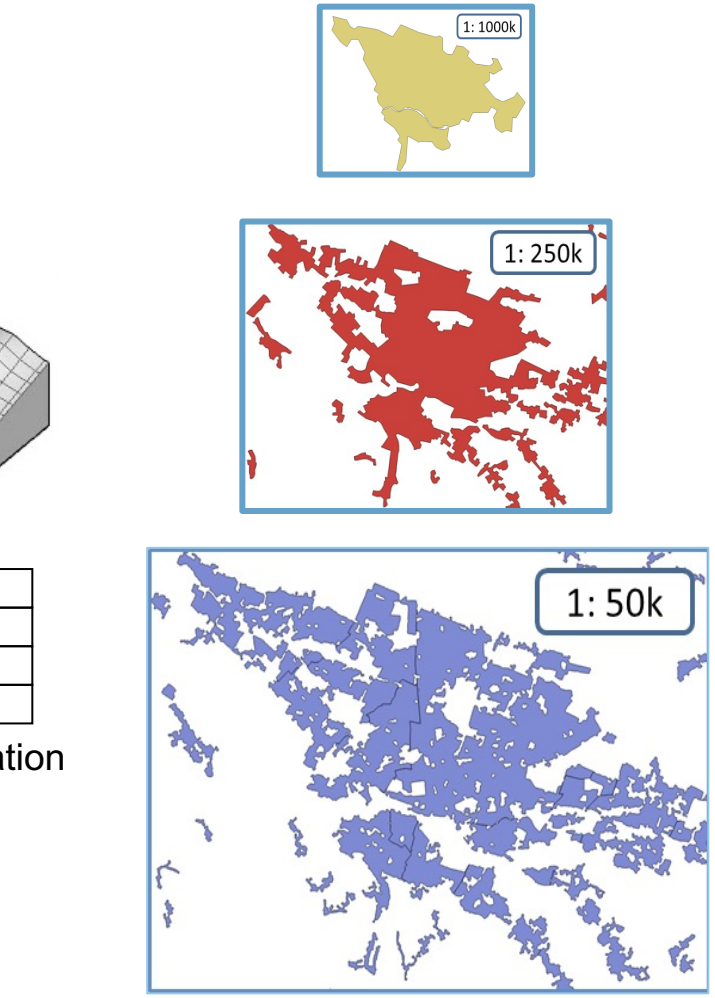
61	67.5	67	37.5	71	26.5
20	41	35	32.5	26.5	21.5
48	43.5	49	45	28.5	51.5



Distributions get smoother at more aggregated levels.



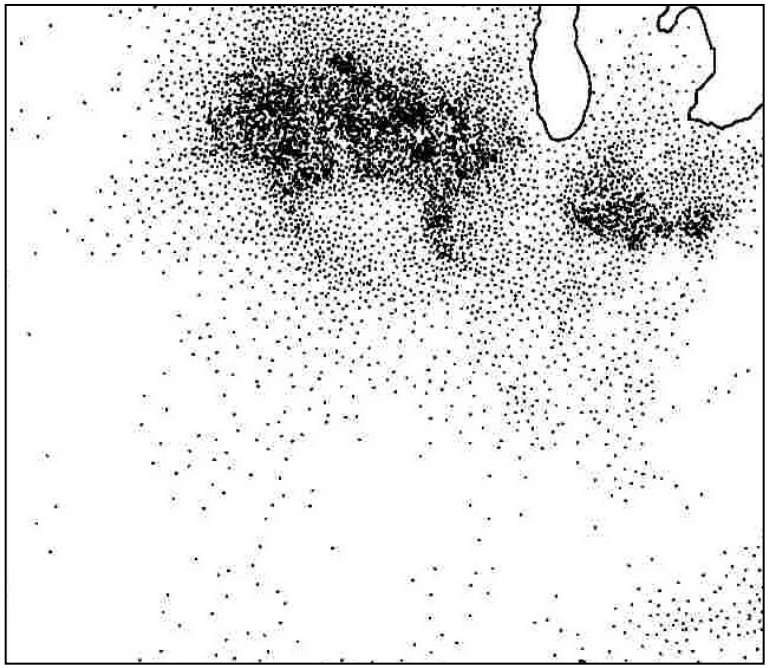
Smoothing effect of image processing

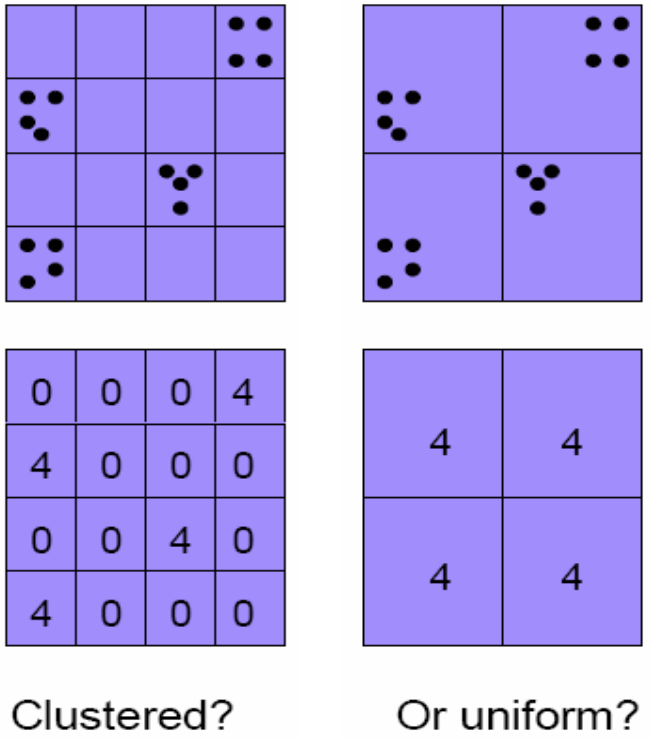


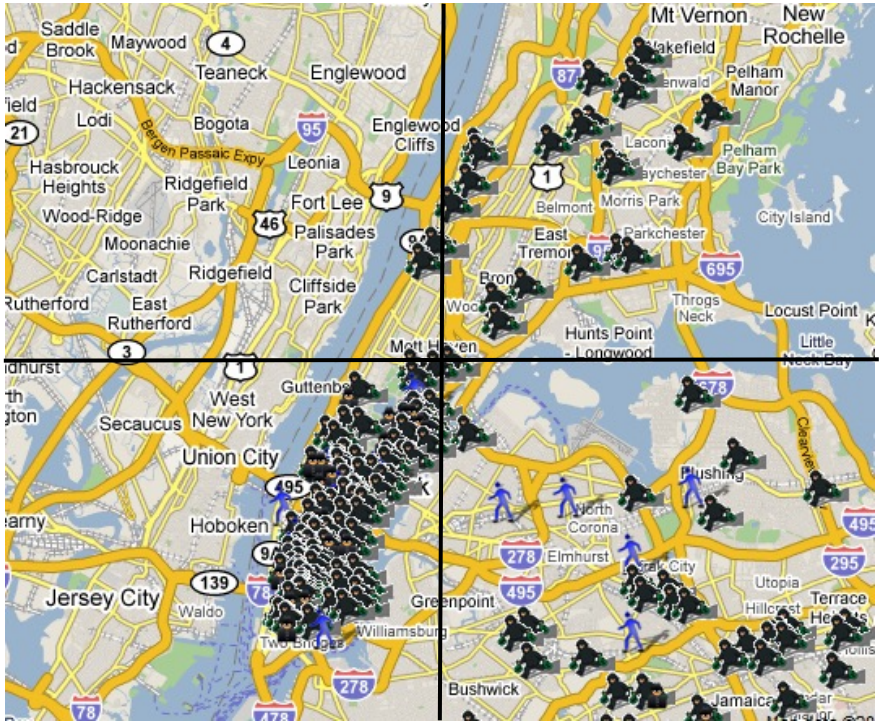
Smoothing effect of map generalization

Pitfall 5 – Nonuniformity

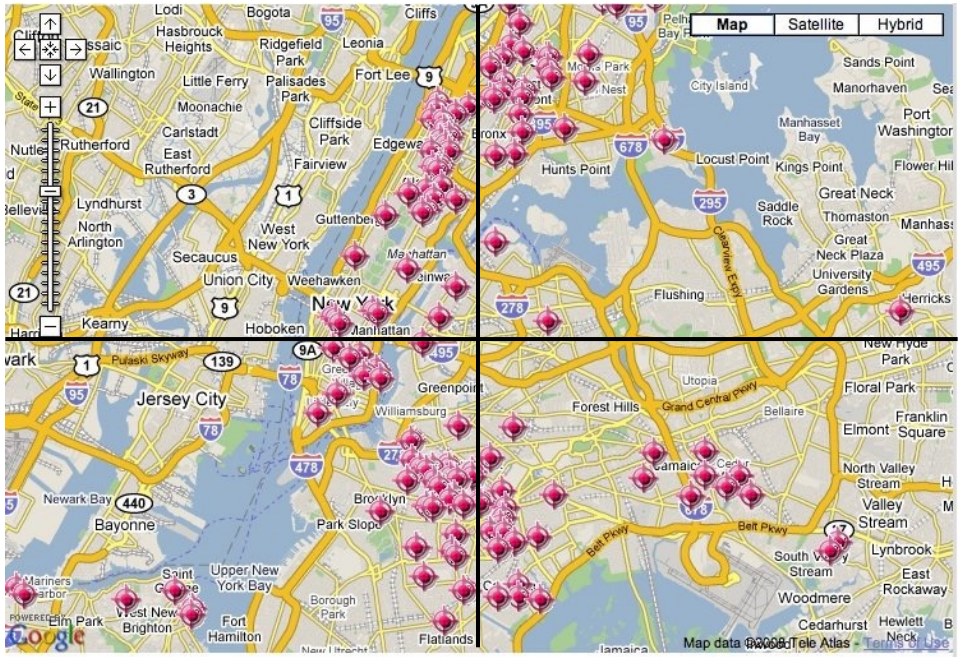
Spatial information is distributed with different concentrations, which can cause redundant or missing samples in different places. The hypothesis that any of the events could have occurred anywhere in the study area does not hold.







Robberies, Theft and Burglary in NYC

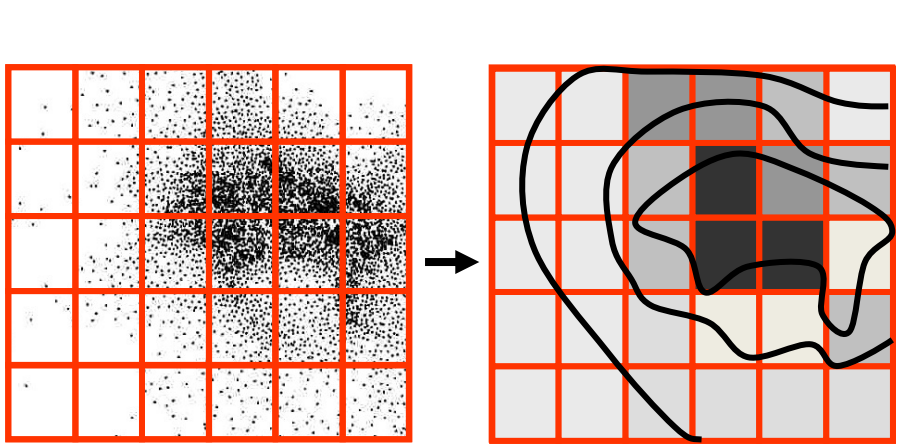


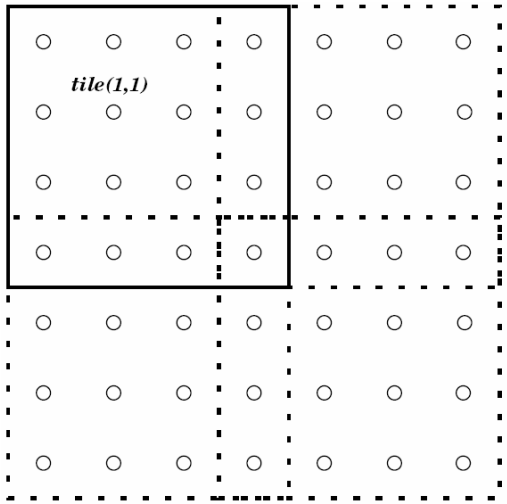
Shootings in NYC

Pitfall 6 - Edge effects

They arise where an artificial boundary is imposed, often just to keep it manageable. A real distribution does not stop at the boundary.

To reduce the edge effect, it is possible to allow the tiles to overlap each other to some extent at their borders. For points where tiles overlap the prediction will be calculated as arithmetic mean of all their results from different tiles.





III. Types of spatial data analysis

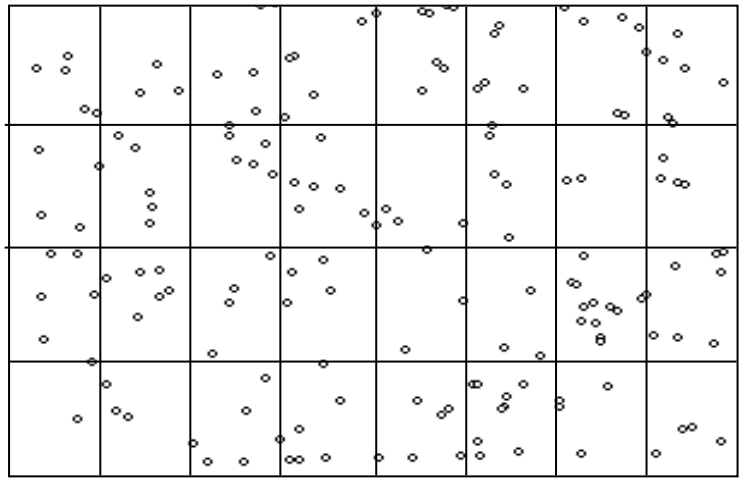
1. The analysis of point patterns
2. Vertical analysis
3. Network analysis

III.1 The analysis of point patterns

The first order analysis: density variations via the mean number of items per unit area

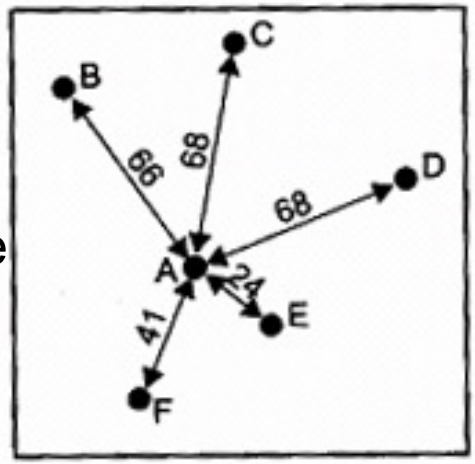
The second order analysis: spatial relationship between the values in pairs of arbitrarily selected areas within the study area.

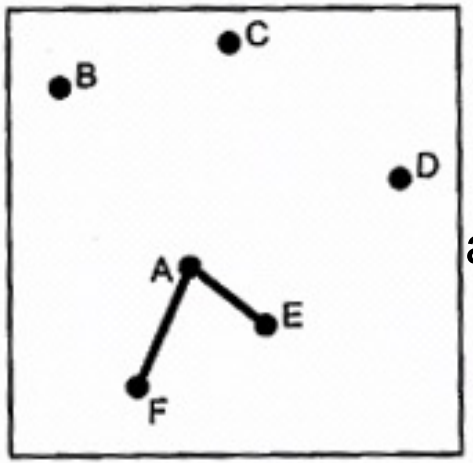
This relationship depends on the distance and direction between the pair of areas; if the relationship depends on distance alone, the distribution is termed isotropic.



Some analytical measures based on distances

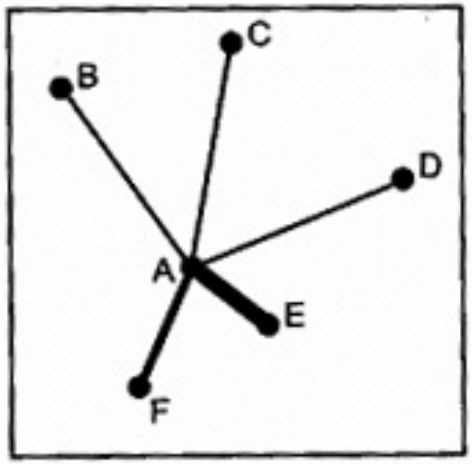
distance

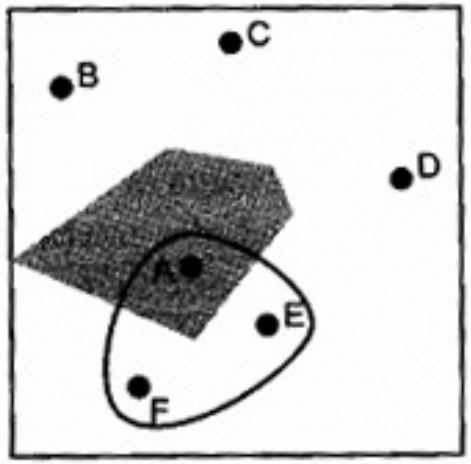




adjacency

interaction



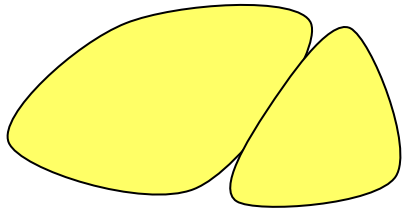


neighborhood

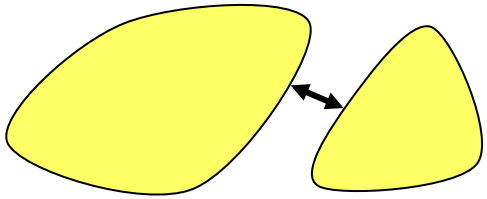
Adjacency

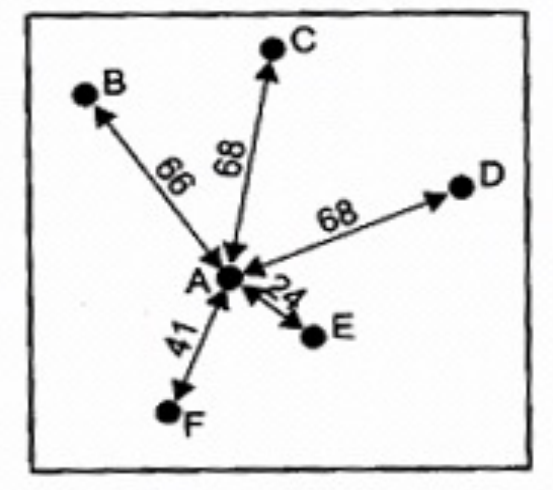
It is the binary equivalent of distance. Two spatial entities are either adjacent or not.

Example 1: two entities are adjacent if they share a common boundary



Example 2: two entities are adjacent if they are within a specified distance





Distance matrix

$$\begin{bmatrix} 0 & 66 & 68 & 68 & 24 & 41 \\ 66 & 0 & 51 & 110 & 99 & 101 \\ 68 & 51 & 0 & 67 & 91 & 116 \\ 68 & 110 & 67 & 0 & 60 & 108 \\ 24 & 99 & 91 & 60 & 0 & 45 \\ 41 & 101 & 116 & 108 & 45 & 0 \end{bmatrix}$$

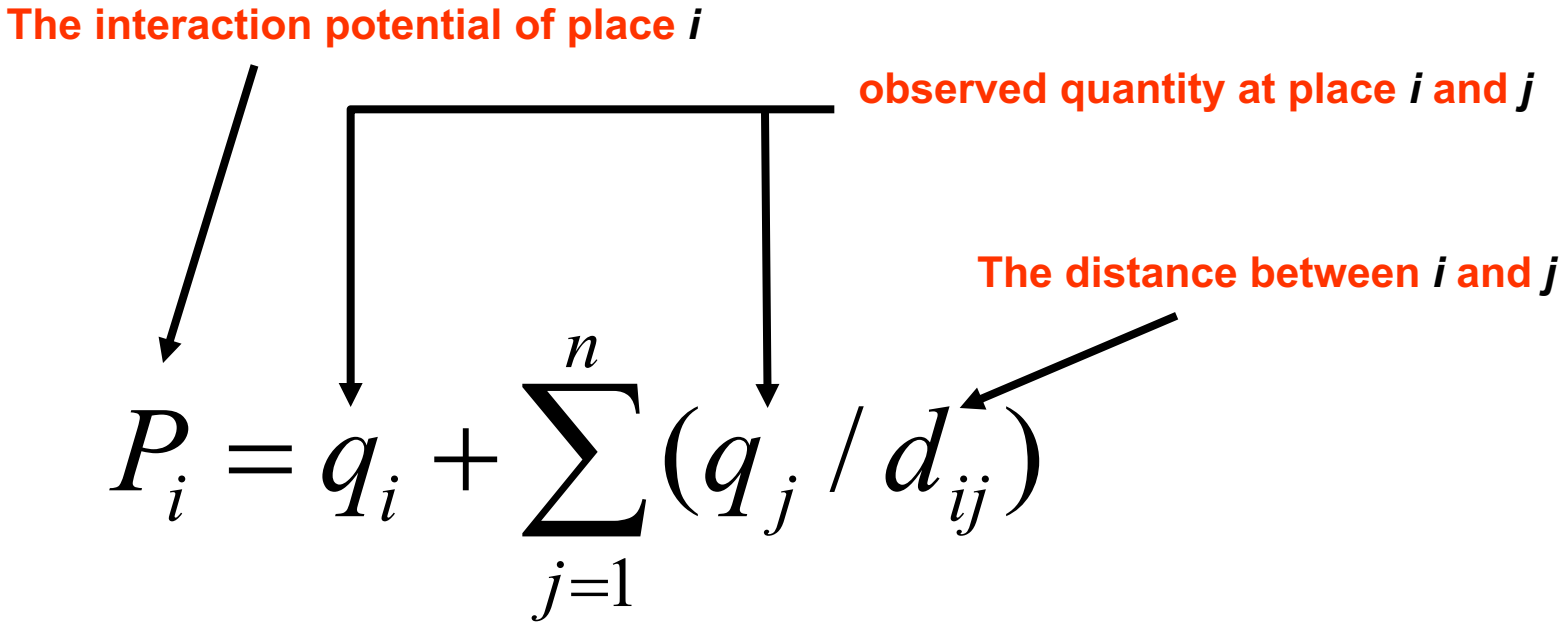
Adjacency matrix

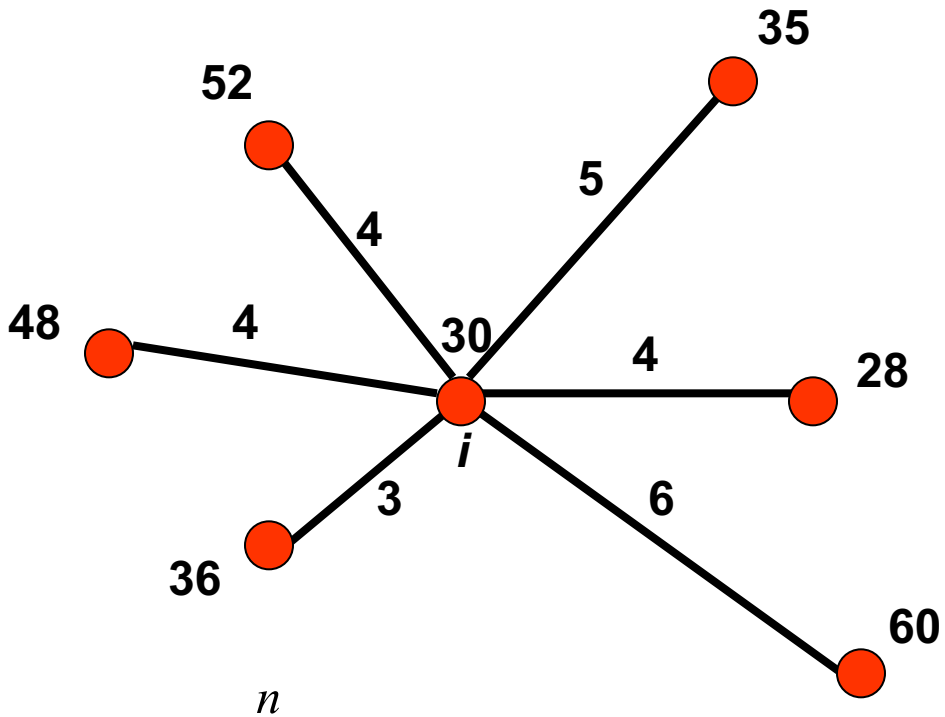
Adjacency $_{d<=50}$

	A	B	C	D	E	F
A	*	0	0	0	1	1
B		*	0	0	0	0
C			*	0	0	0
D				*	0	0
E					*	1
F						*

Interaction potential

is a function of measured value and distance.





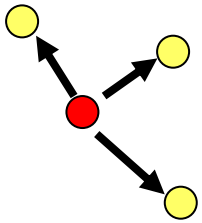
$$P_i = q_i + \sum_{j=1}^{n} (q_j / d_{ij})$$

= 30 + (35/5 + 28/4 + 60/6 +36/3+48/4+52/4)

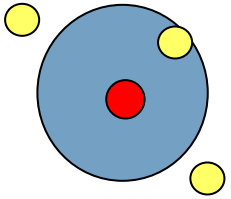
= 91

Neighborhood

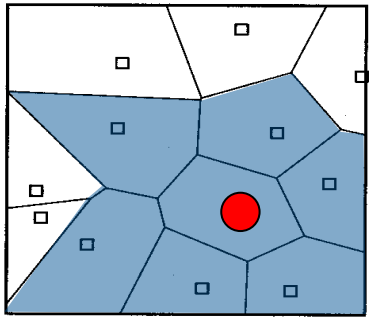
Example 1: a spatial entity and its adjacent entities



Example 2: a region of space around an entity



Example 3: a polygonal area bordering an entity



To be continued

1. What is the goal of spatial data analysis?
2. Use examples to explain the various components of data quality – accuracy, precision, consistency, completeness.
3. What is the difference between resolution and sampling rate?
4. What does it mean by spatial autocorrelation?
5. What are Moran's index and Geary's Coefficient?
6. What is the Modifiable Areal Unit Problem?
7. What is the Ecological Fallacy?
8. What are the differences between scale, nonuniformity and edge effect?
9. Explain the analytical measures - adjacency, neighborhood and interaction.
10. Do you know further significant pitfalls of spatial data analysis?