
Spatial Statistics and Analysis Methods

(for GEOG 104 class).

- Provided by Dr. An Li, San Diego State University.

Types of spatial data

■ Points

- Point pattern analysis (PPA; such as nearest neighbor distance, quadrat analysis)
- Moran's I, Getis G*

■ Areas

- Area pattern analysis (such as join-count statistic)
- **Switch to PPA** if we use centroid of area as the point data

■ Lines

- Network analysis
- Also switch to PPA

→ Three ways to represent and thus to analyze spatial data:

Spatial arrangement

- Randomly distributed data
 - The assumption in “classical” statistic analysis
- Uniformly distributed data
 - The most dispersed pattern—the antithesis of being clustered
 - **Negative spatial autocorrelation**
- Clustered distributed data
 - **Tobler’s Law** — all things are related to one another, but near things are more related than distant things
 - **Positive spatial autocorrelation**

→ Three basic ways in which points or areas may be spatially arranged

Spatial Distribution with p value

Case 1:
Clustered

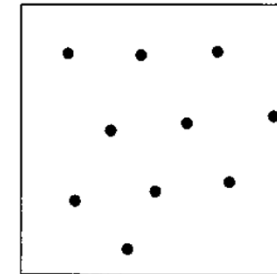
χ^2 is
large



p is low
(close to 0)

Case 3:
Dispersed

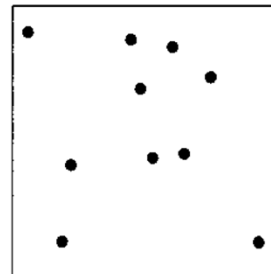
χ^2 is
small



p is high
(close to 1)

Case 2:
Random

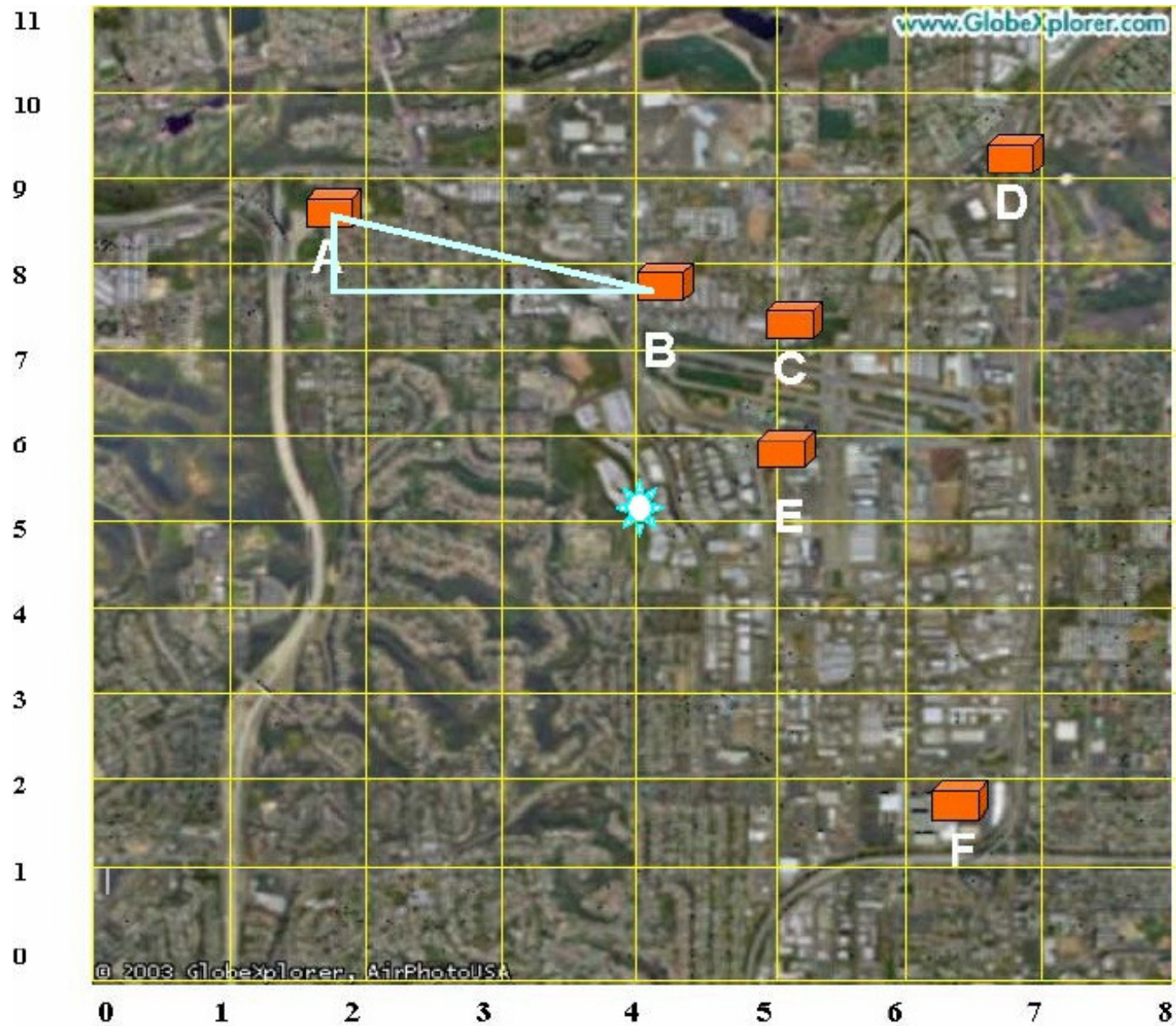
χ^2 is
intermediate



p is intermediate
(not close to
either 0 or 1)

Nearest neighbor distance

- Questions:
 - What is the pattern of points in terms of their nearest distances from each other?
 - Is the pattern random, dispersed, or clustered?
- Example
 - Is there a pattern to the distribution of toxic waste sites near the area in San Diego (see next slide)? [hypothetical data]



- Step 1: Calculate the distance from each point to its nearest neighbor, by calculating the hypotenuse of the triangle:

$$NND_{AB} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

Site	X	Y	NN	NND
A	1.7	8.7	B	2.79
B	4.3	7.7	C	0.98
C	5.2	7.3	B	0.98
D	6.7	9.3	C	2.50
E	5.0	6.0	C	1.32
F	6.5	1.7	E	4.55
				13.12

$$\overline{NND} = \frac{\sum NND}{n} = \frac{13.12}{6} = 2.19$$

-
- Step 2: Calculate the distances under varying conditions
 - The average distance if the pattern were random?

$$\overline{NND}_R = \frac{1}{2\sqrt{Density}} = \frac{1}{2\sqrt{0.068}} = \frac{1}{2 \times 0.261} = 1.92$$

Where density = n of points / area = 6/88 = 0.068

- If the pattern were completely clustered (all points at same location), then:

$$\overline{NND}_C = 0$$

- Whereas if the pattern were completely dispersed, then:

$$\overline{NND}_D = \frac{1.07453}{\sqrt{Density}} = \frac{1.07453}{0.261} = 4.12$$

(Based on a Poisson distribution)

- Step 3: Let's calculate the standardized nearest neighbor index (R) to know what our NND value means:

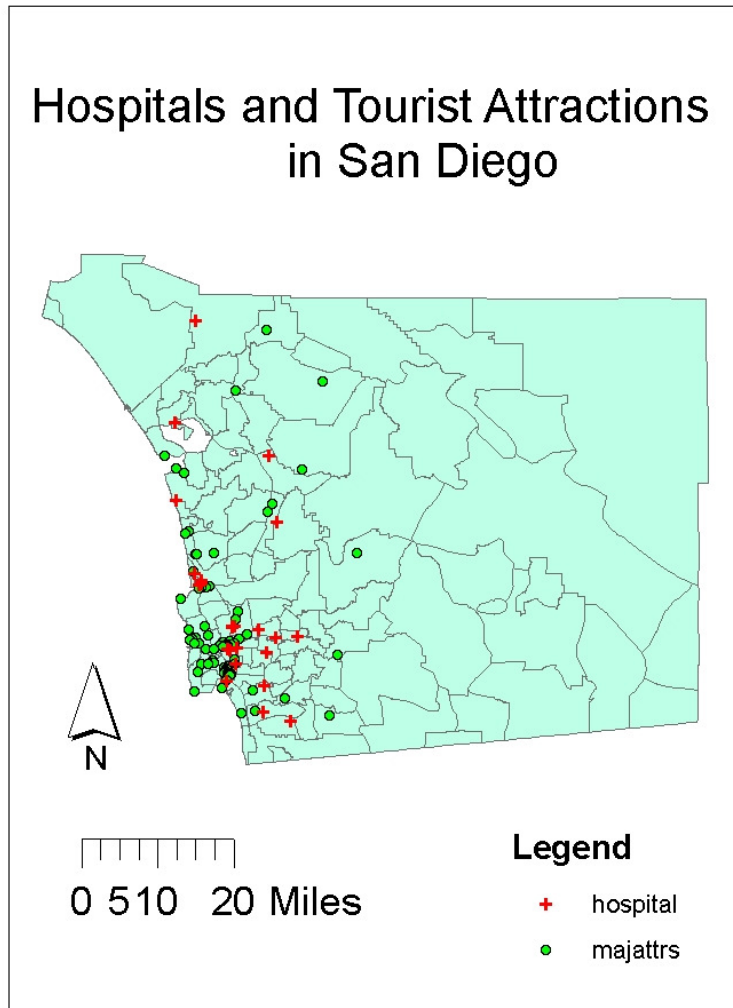
$$R = \frac{\overline{NND}}{\overline{NND}_R} = \frac{2.19}{1.92} = 1.14$$

= slightly more dispersed than random

cf. Perfectly dispersed: $R = \frac{\overline{NND}_D}{\overline{NND}_R} = \frac{4.12}{1.92} = 2.15$



Hospitals & Attractions in San Diego

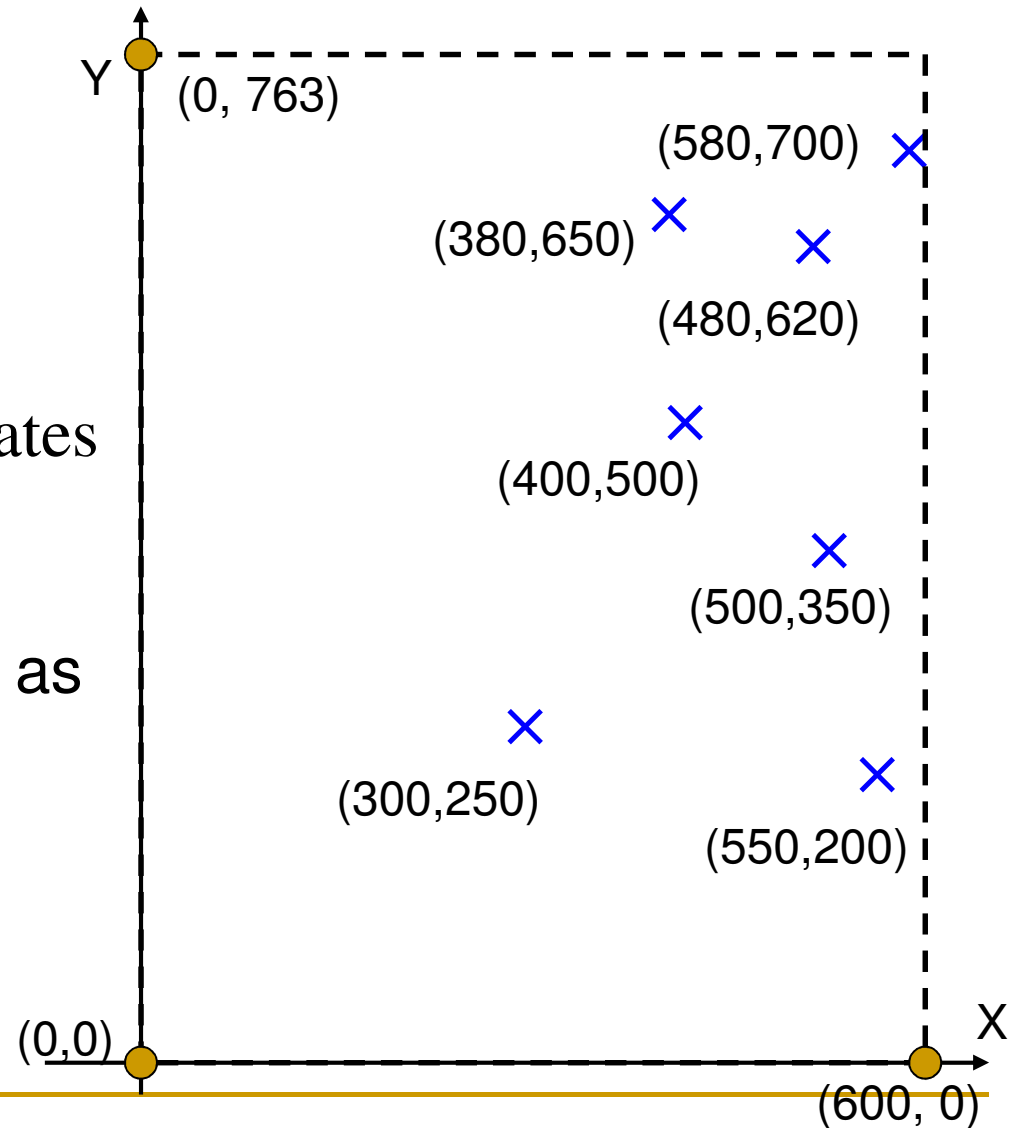


- The map shows the locations of hospitals (+) and tourist attractions (●) in San Diego
- Questions:
 - Are hospitals randomly distributed
 - Are tourist attractions clustered?

What can we do?

■ Preparations

- Find or build a coordinate system
- Measure the coordinates of the center of each forest
- Use **centroid of area** as the point data



Mean center

- The mean center is the “average” position of the points

- Mean center of X: $\bar{X}_C = \frac{\sum x}{n}$

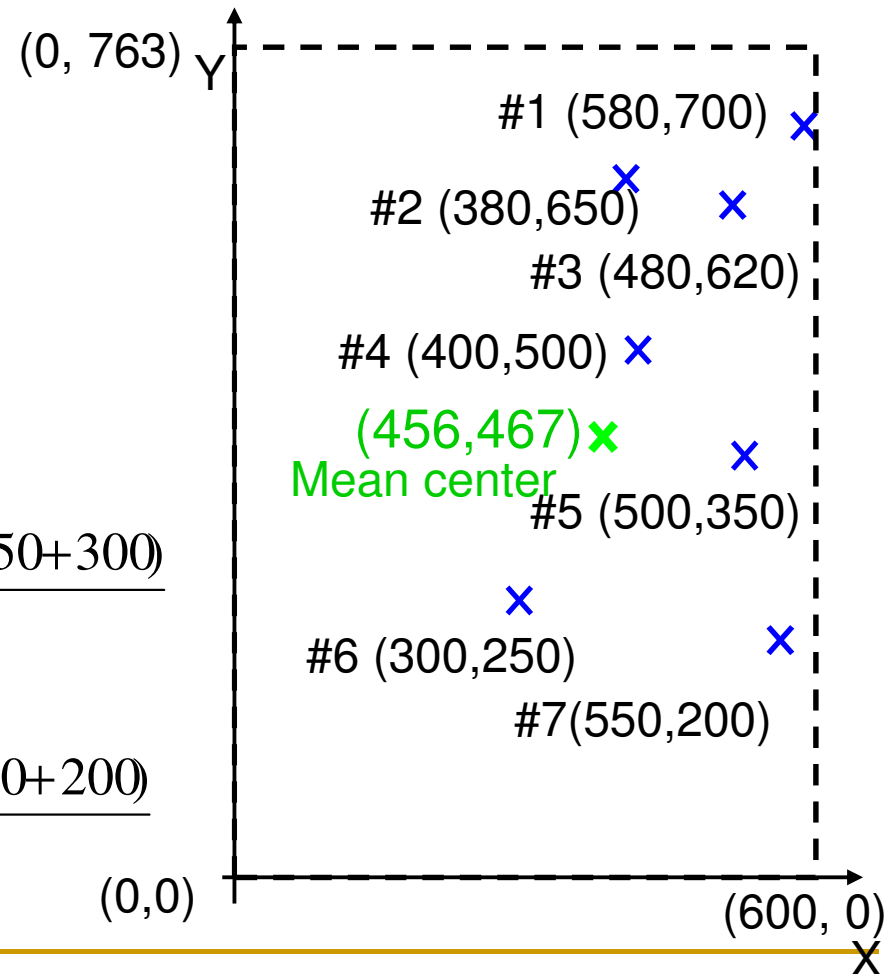
Mean center of Y: $\bar{Y}_C = \frac{\sum y}{n}$

$$\bar{X}_C = \frac{(580+380+480+400+500+550+300)}{7}$$

$$= 455.71$$

$$\bar{Y}_C = \frac{(700+650+620+500+350+250+200)}{7}$$

$$= 467.14$$



Standard distance

- The standard distance measures the amount of dispersion
 - Similar to standard deviation
 - Formula

$$S_D = \sqrt{\frac{\sum (X_i - \bar{X}_c)^2 + \sum (Y_i - \bar{Y}_c)^2}{n}} \quad \leftarrow \text{Definition}$$

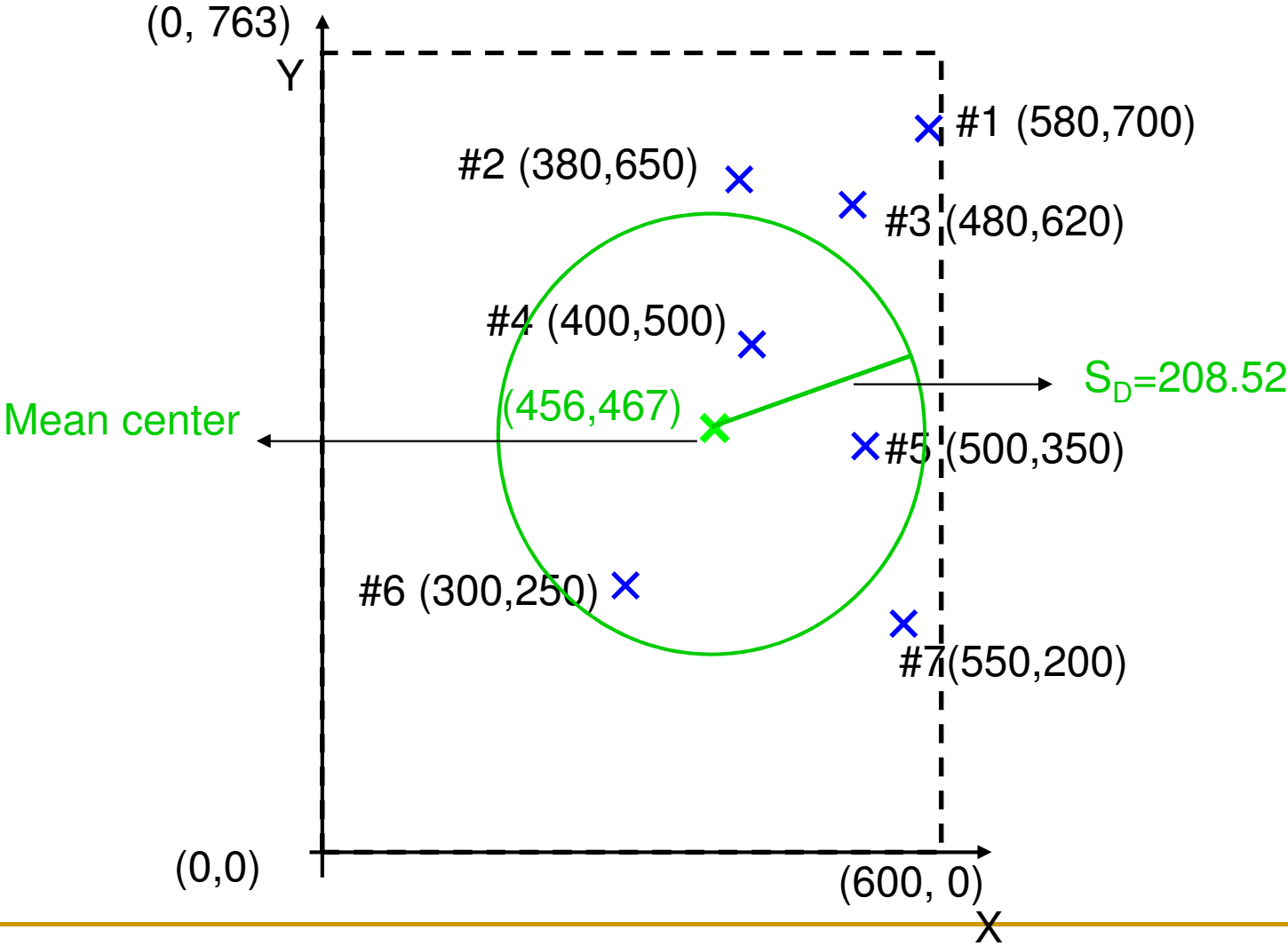
$$S_D = \sqrt{\left(\frac{\sum X_i^2}{n} - \bar{X}_c^2\right) + \left(\frac{\sum Y_i^2}{n} - \bar{Y}_c^2\right)} \quad \leftarrow \text{Computation}$$

Standard distance

Forests	X	X ²	Y	Y ²
#1	580	336400	700	490000
#2	380	144400	650	422500
#3	480	230400	620	384400
#4	400	160000	500	250000
#5	500	250000	350	122500
#6	300	90000	250	62500
#7	550	302500	200	40000
	Sum of X ²	1513700	Sum of Y ²	1771900
	$\bar{X}_c = 455.71$		$\bar{Y}_c = 467.14$	

$$\begin{aligned}
 S_D &= \sqrt{\left(\frac{\sum X_i^2}{n} - \bar{X}_c^2\right) + \left(\frac{\sum Y_i^2}{n} - \bar{Y}_c^2\right)} \\
 &= \sqrt{\left(\frac{1513700}{7} - 455.71^2\right) + \left(\frac{1771900}{7} - 467.14^2\right)} = 208.52
 \end{aligned}$$

Standard distance



Definition of **weighted** mean center standard distance

- What if the forests with **bigger area** (the area of the smallest forest as unit) should have more influence on the mean center?

$$\bar{X}_{wc} = \frac{\sum f_i X_i}{\sum f_i} \quad \bar{Y}_{wc} = \frac{\sum f_i Y_i}{\sum f_i}$$

$$S_{WD} = \sqrt{\frac{\sum f_i (X_i - \bar{X}_{wc})^2 + \sum f_i (Y_i - \bar{Y}_{wc})^2}{\sum f_i}} \quad \leftarrow \text{Definition}$$

$$S_{WD} = \sqrt{\left(\frac{\sum f_i X_i^2}{\sum f_i} - \bar{X}_{wc}^2 \right) + \left(\frac{\sum f_i Y_i^2}{\sum f_i} - \bar{Y}_{wc}^2 \right)} \quad \leftarrow \text{Computation}$$

Calculation of weighted mean center

- What if the forests with bigger area (the area of the smallest forest as unit) should have more influence?

Forests	f(Area)	X_i	$f_i X_i$ (Area*X)	Y_i	$f_i Y_i$ (Area*Y)
#1	5	580	2900	700	3500
#2	20	380	7600	650	13000
#3	5	480	2400	620	3100
#4	10	400	4000	500	5000
#5	20	500	10000	350	7000
#6	1	300	300	250	250
#7	25	550	13750	200	5000
$\sum f_i$	86	$\sum f_i X_i$	40950	$\sum f_i Y_i$	36850

$$\bar{X}_{wc} = \frac{\sum f_i X_i}{\sum f_i} = \frac{40950}{86} = 476.16 \quad \bar{Y}_{wc} = \frac{\sum f_i Y_i}{\sum f_i} = \frac{36850}{86} = 428.49$$

Calculation of weighted standard distance

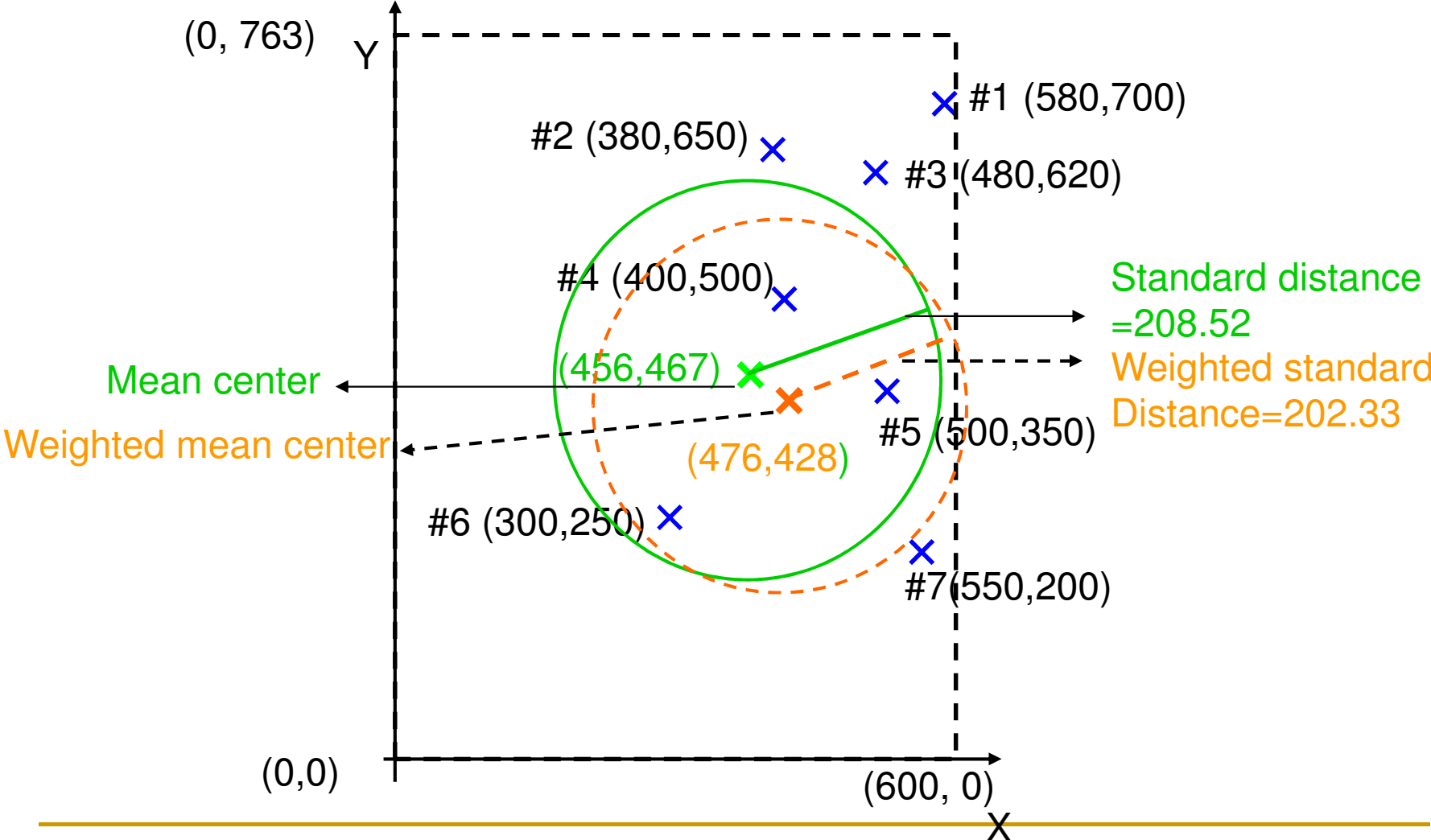
- What if the forests with bigger area (the area of the smallest forest as unit) should have more influence?

Forests	$f_i(\text{Area})$	X_i	X_i^2	$f_i X_i^2$	Y_i	Y_i^2	$f_i Y_i^2$
#1	5	580	336400	1682000	700	490000	2450000
#2	20	380	144400	2888000	650	422500	8450000
#3	5	480	230400	1152000	620	384400	1922000
#4	10	400	160000	1600000	500	250000	2500000
#5	20	500	250000	5000000	350	122500	2450000
#6	1	300	90000	90000	250	62500	62500
#7	25	550	302500	7562500	200	40000	1000000
$\sum f_i$	86		$\sum f_i X_i^2$	19974500		$\sum f_i Y_i^2$	18834500

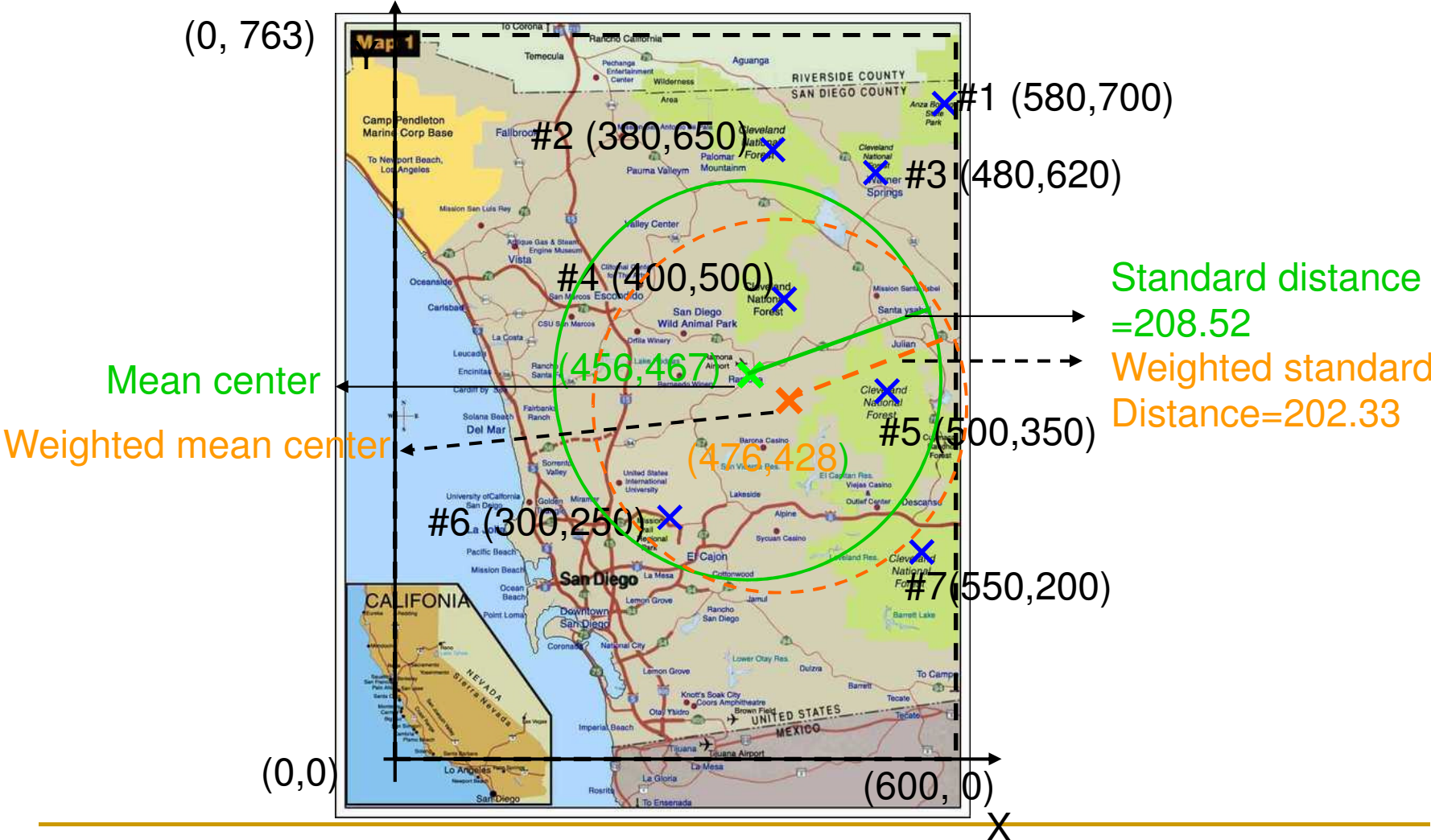
$$S_{WD} = \sqrt{\left(\frac{\sum f_i X_i^2}{\sum f_i} - \bar{X}_{wc}^2 \right) + \left(\frac{\sum f_i Y_i^2}{\sum f_i} - \bar{Y}_{wc}^2 \right)}$$

$$= \sqrt{\left(\frac{19974500}{86} - 476.16^2 \right) + \left(\frac{18834500}{86} - 428.49^2 \right)} = 202.33$$

Standard distance



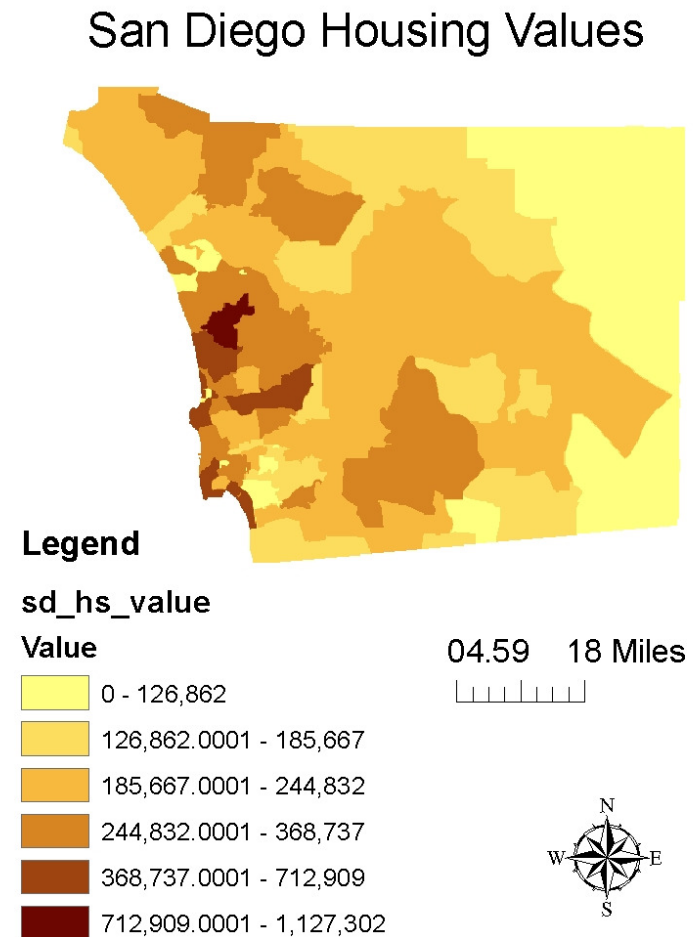
Standard distance



Spatial clustered?

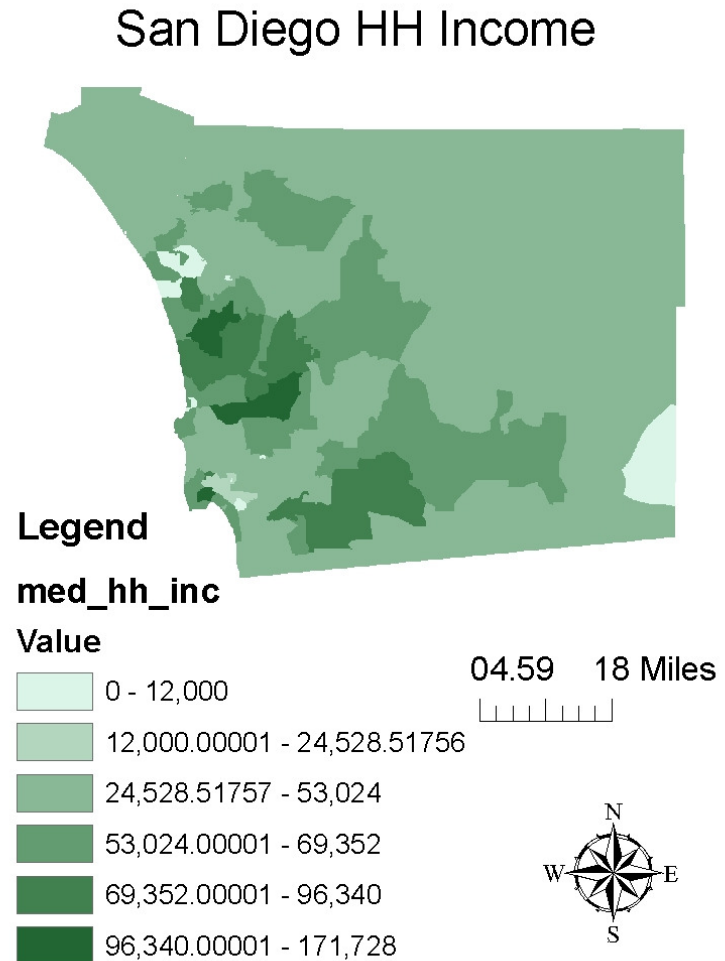
Tobler's First Law

- Given such a map, is there strong evidence that housing values are clustered in space?
 - Lows near lows
 - Highs near highs



More than this one?

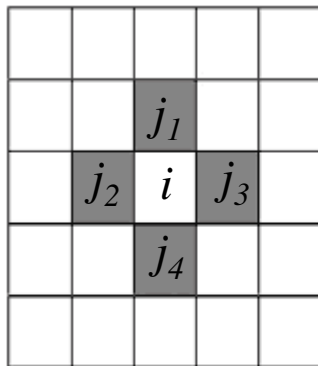
- Does household income show more spatial clustering, or less?



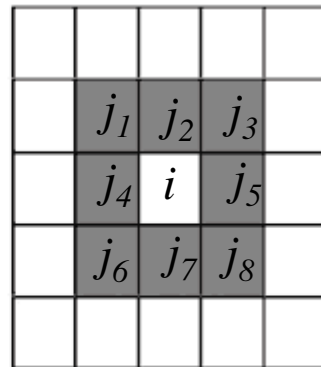
Neighborhood?

i : Center Cell

j : Neighborhoods

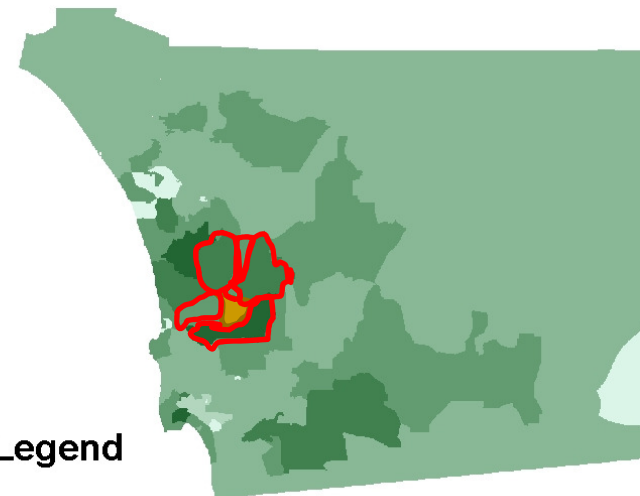


The von Neumann
neighbourhood



The Moore
neighbourhood

San Diego HH Income



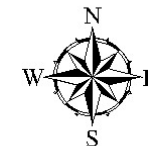
Legend

med_hh_inc

Value



04.59 18 Miles



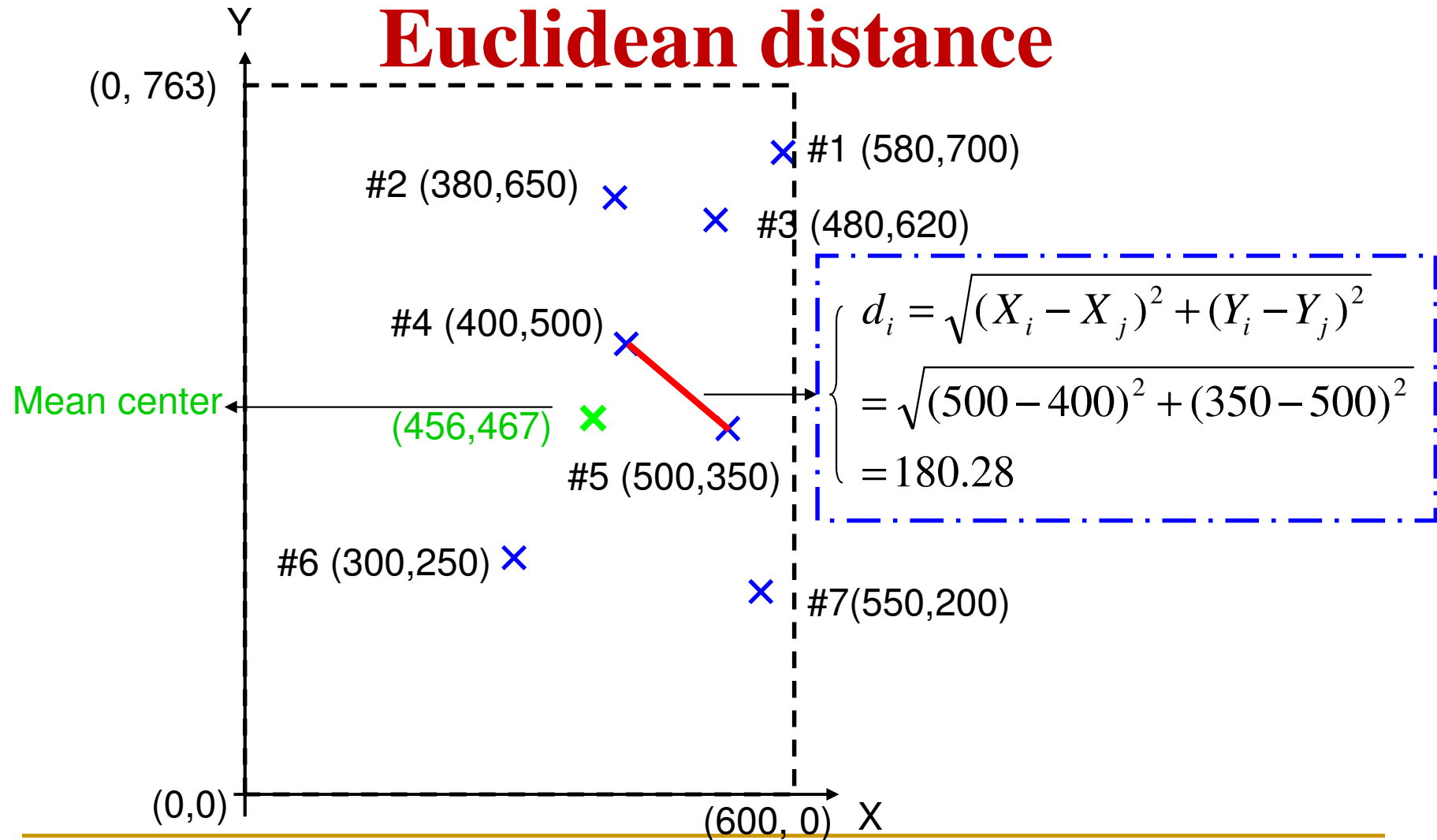
Moran's I statistic

- Global Moran's I
 - Characterize the overall spatial dependence among a set of areal units

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

Covariance

Network Analysis: Shortest routes



Manhattan Distance

■ Euclidean median

- Find (X_e, Y_e) such that

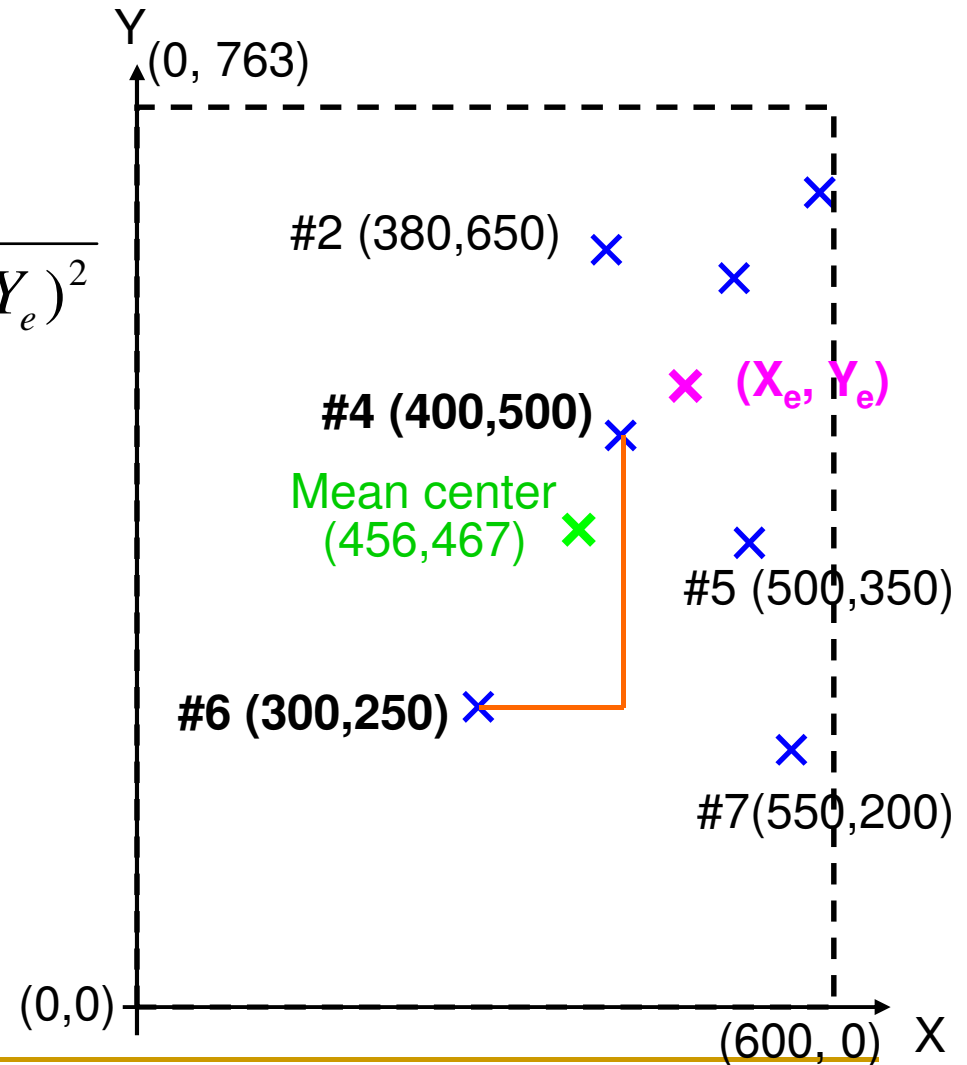
$$d_e = \sum \sqrt{(X_i - X_e)^2 + (Y_i - Y_e)^2}$$

is minimized

- Need iterative algorithms
- Location of **fire** station

■ Manhattan median

$$\begin{aligned} d_{ij} &= |X_i - X_j| + |Y_i - Y_j| \\ &= |400 - 300| + |500 - 250| \\ &= 350 \end{aligned}$$



Summary

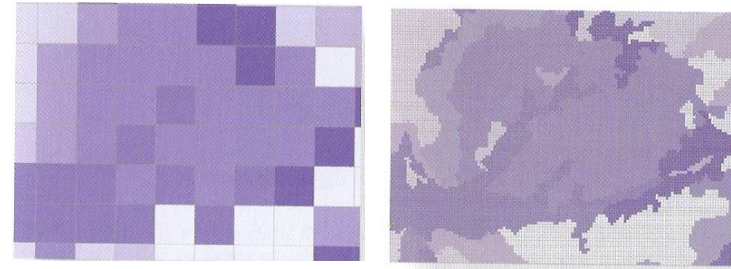
- What are spatial data?
 - Mean center
 - Weighted mean center
 - Standard distance
 - Weighted standard distance
 - Euclidean median
 - Manhattan median
- } **Calculate in GIS environment**

Spatial resolution

- Patterns or relationships are

scale dependent

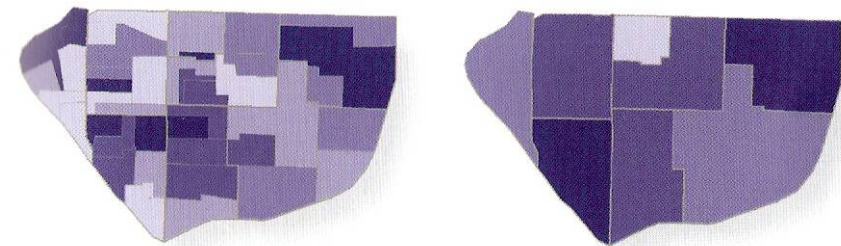
- Hierarchical structures (blocks → block groups → census tracts...)
- Cell size: # of cells vary and spatial patterns masked or overemphasized



Vegetation types at large cell (left)
and small cells (right)

- How to decide

- The **goal/context** of your study
- Test different sizes (Weeks et al. article: 250m, 500m, and 1,000 m)



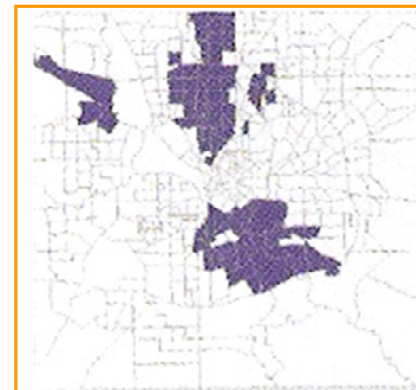
% of seniors at block groups (left)
and census tracts (right)

Edge effects

- What it is
 - Features near the boundary (regardless of how it is defined) have **fewer neighbors** than those inside
 - The results about near-edge features are usually **less reliable**
- How to handle
 - Buffer your study area (outward or inward), and include more or fewer features
 - Varying weights for features near boundary



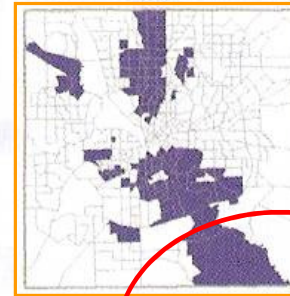
a. Median income by census tracts



b. Significant clusters (Z-scores for I_j)



c. More census tracts within the buffer (between brown and black boxes) included



Different!

d. More areas are significant

Applying Spatial Statistics

- Visualizing spatial data
 - Closely related to GIS
 - Other methods such as Histograms
- Exploring spatial data
 - Random spatial pattern or not
 - Tests about randomness
- Modeling spatial data
 - Correlation and χ^2
 - Regression analysis