Spatial Statistics and Analysis Methods

(for GEOG 104 class).

- Acknowledgement: Part of the content is contributed by Dr. An Li, San Diego State University.

Types of spatial data

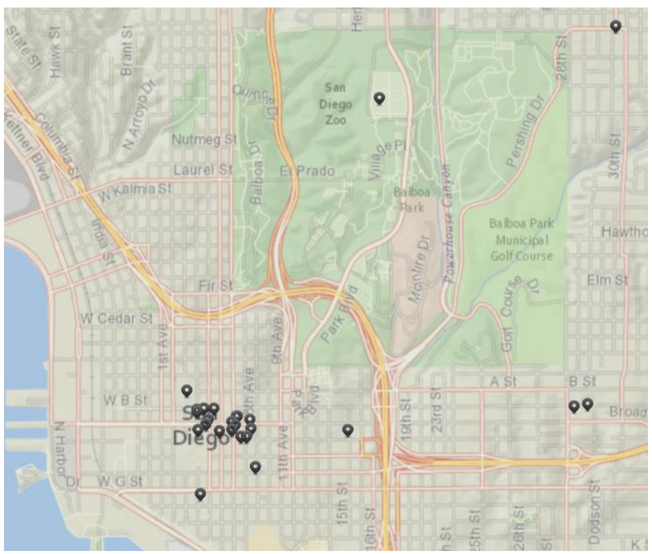
Three ways to represent and thus to analyze spatial data:

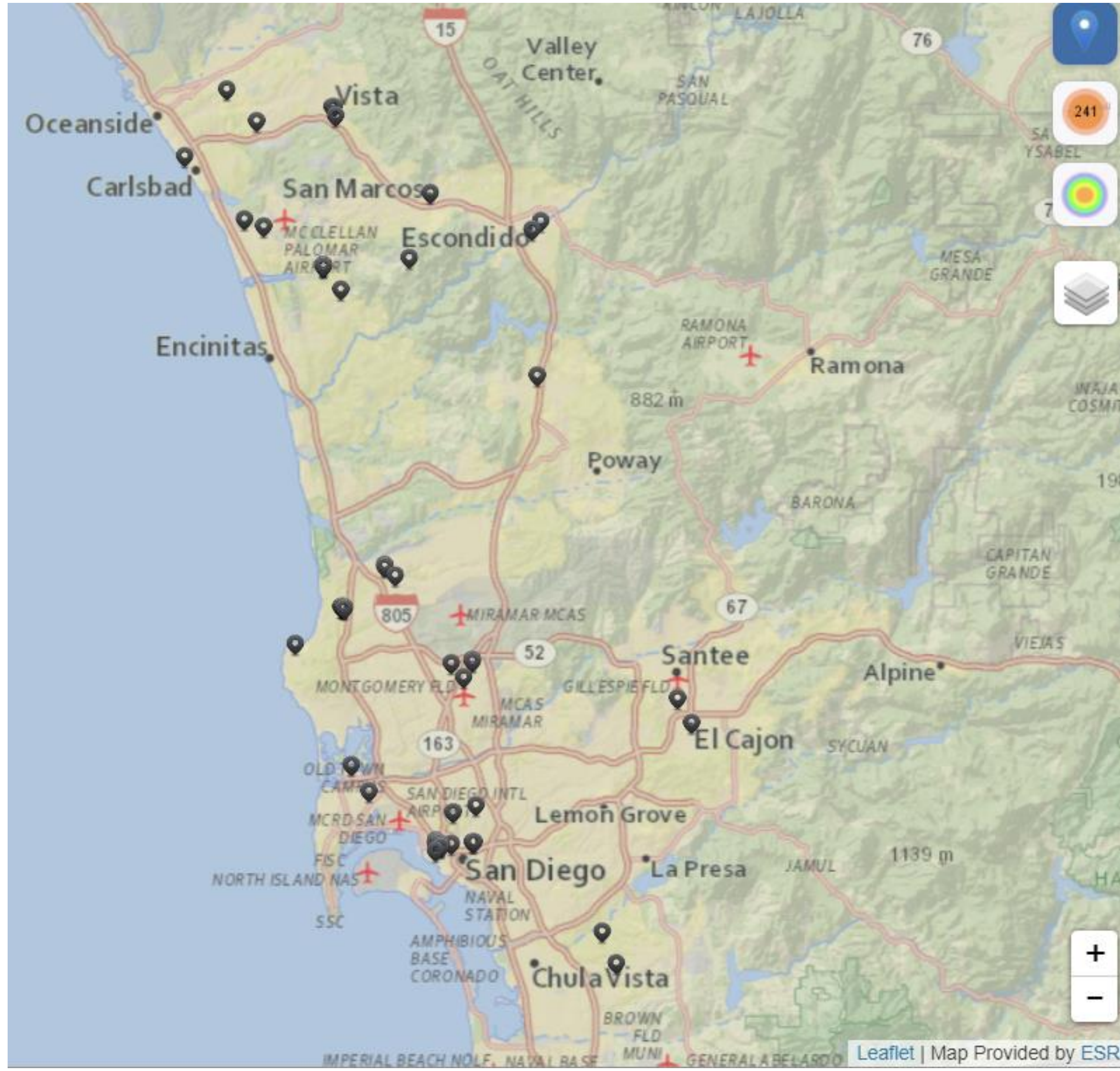
- Points
  - **Point pattern analysis (PPA**; such as nearest neighbor distance, quadrat analysis)
  - Statistic indexes: **Moran's I, Getis G***
- Areas
  - Area pattern analysis (such as join-count statistic)
  - Switch to PPA if we use centroid of area as the point data
- Lines
  - Network analysis

Point pattern analysis (PPA)

Twitter (Geo-tagged) messages distribution pattern in San Diego. (4/24/2018).

http://vision.sdsu.edu/ec2/geoviewer/sanDiego#



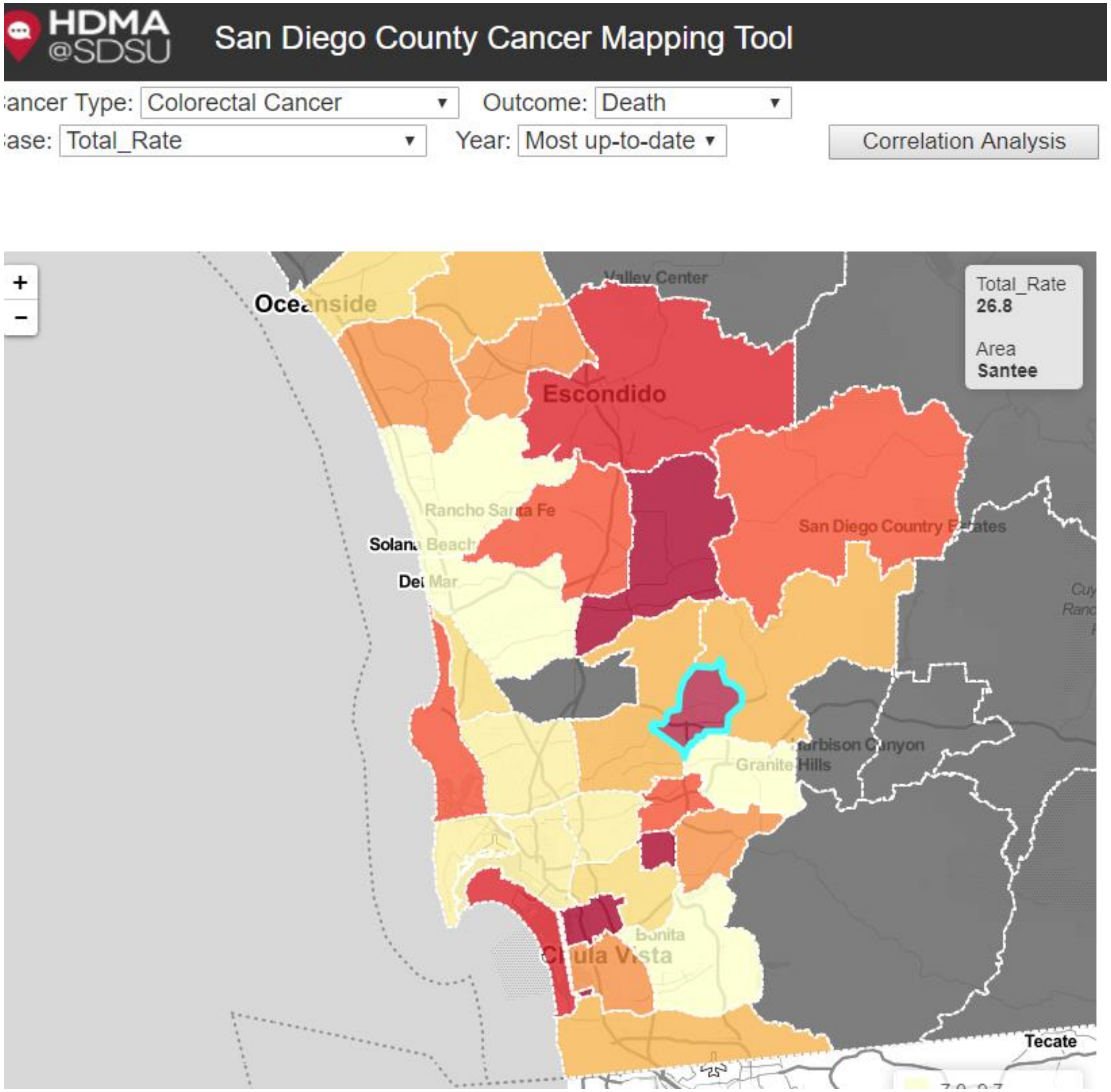


Area pattern analysis (such as join-count statistic)

Switch to PPA if we use centroid of area as the point data.

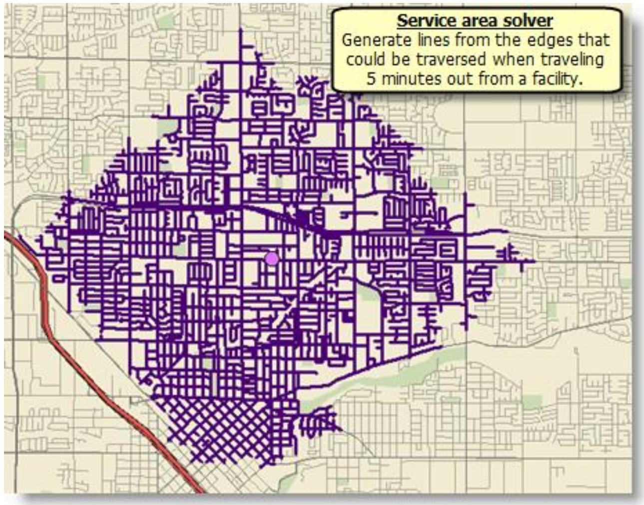
Example:

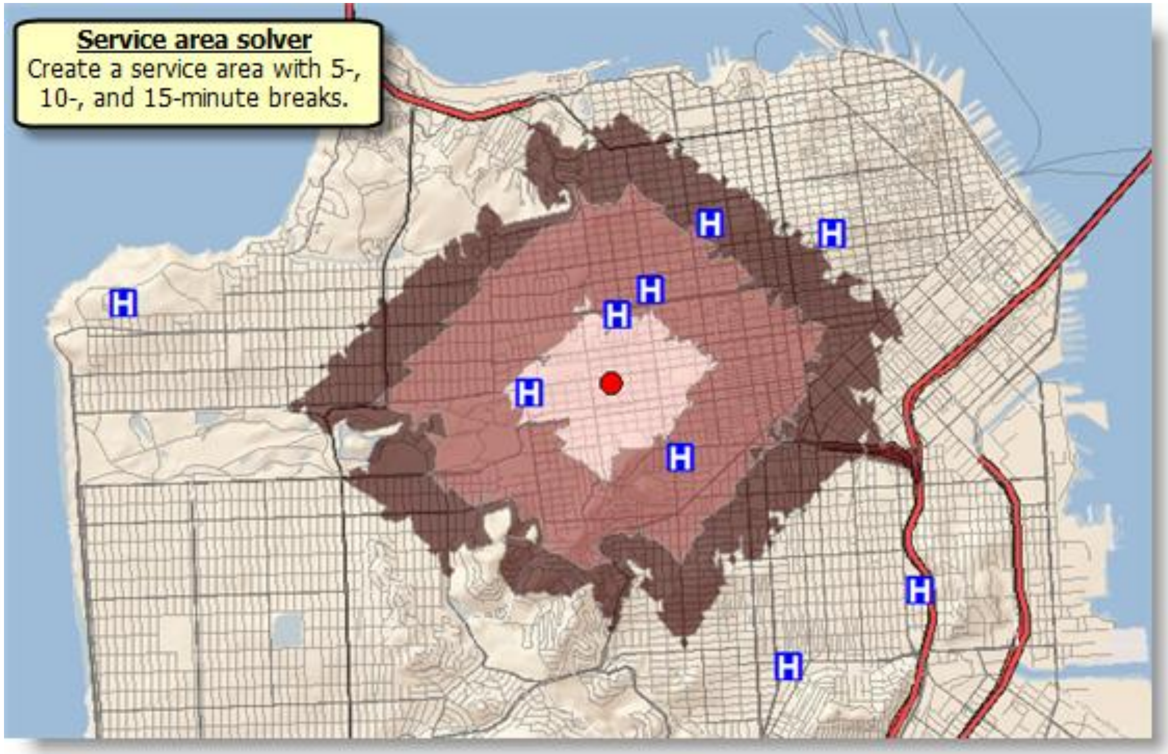
Cancer Death Rate (Colorectal Cancer in San Diego)



Network analysis (Accessibility)

Source: http://desktop.arcgis.com/en/arcmap/latest/extensions/network-analyst/types-of-network-analyses.htm





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** — Everything is related to everything else, but near things are more related than distant things.
  - Positive **spatial** **autocorrelation**

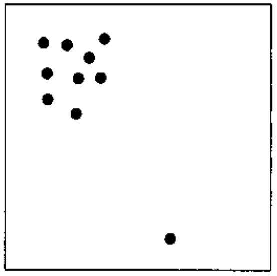
→**Three basic ways in which points or areas may be spatially arranged**

Spatial Distribution with R value

standardized nearest neighbor index (R)

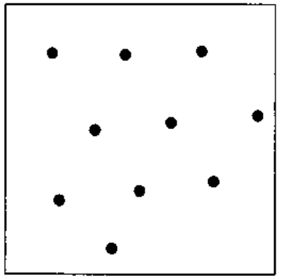
Case 1:
Clustered

χ^2 is
large



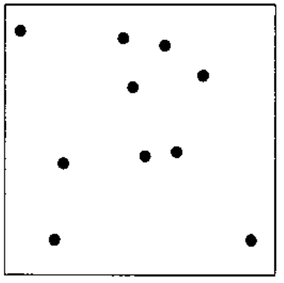
Case 3:
Dispersed

χ^2 is
small



Case 2:
Random

χ^2 is
intermediate



Nearest neighbor distance (NND)

- 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]

Nearest neighbor distance (NND)

- 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

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

Nearest neighbor distance (NND)

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

$$\overline{NND_R} = \frac{1}{2\sqrt{Density}} = \frac{1}{2\sqrt{0.068}} = 1.92$$

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 <u>dispersed</u> than random

Perfectly dispersed — 2.15

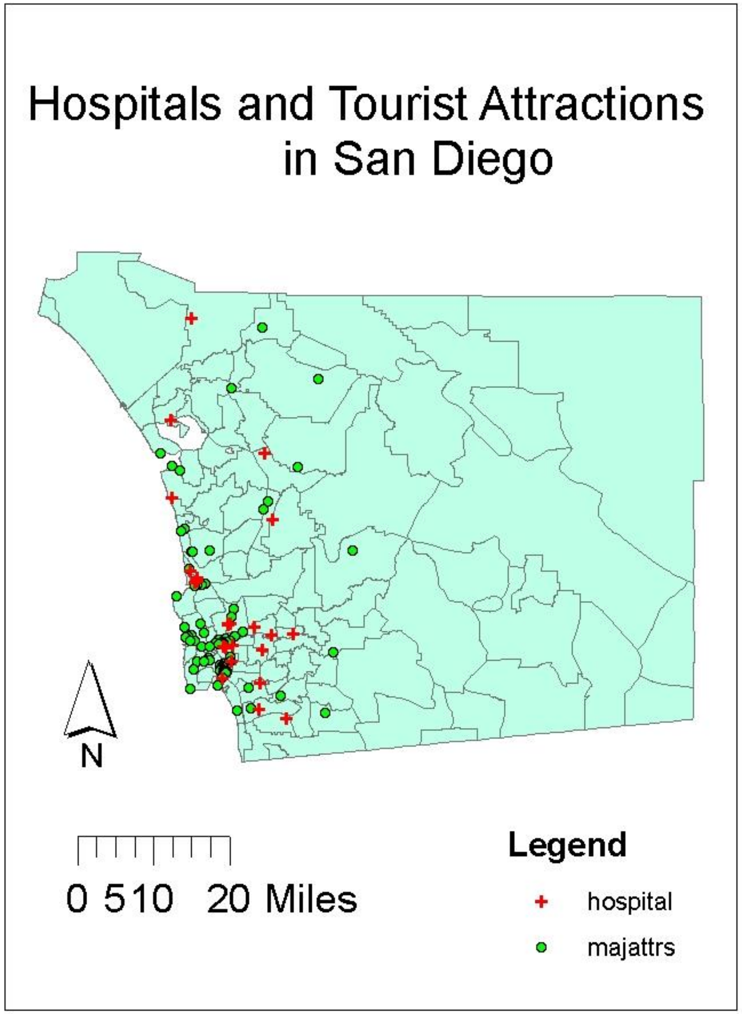
More dispersed than random

Totally random — **1**

More clustered than random

Perfectly clustered — 0

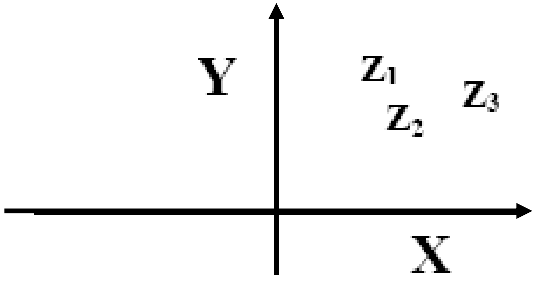
Hospitals & Attractions in San Diego



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

Spatial Data (with X, Y coordinates)

- Any set of information (some variable ‘z’) for which we have locational coordinates (e.g. longitude, latitude; or x, y)



- Point data are straightforward, unless we aggregate all point data into an areal or other spatial units
- **Area data require additional assumptions** regarding:
  - Boundary delineation
  - Modifiable areal unit (states, counties, street blocks)
  - Level of spatial aggregation = scale

Pancreatic Cancer (brown) and Liver Cancer (red) Death Age-Adjusted Rates 2013 in San Diego County

Population Density

low

high

Pancreatic Cancer Death Rates

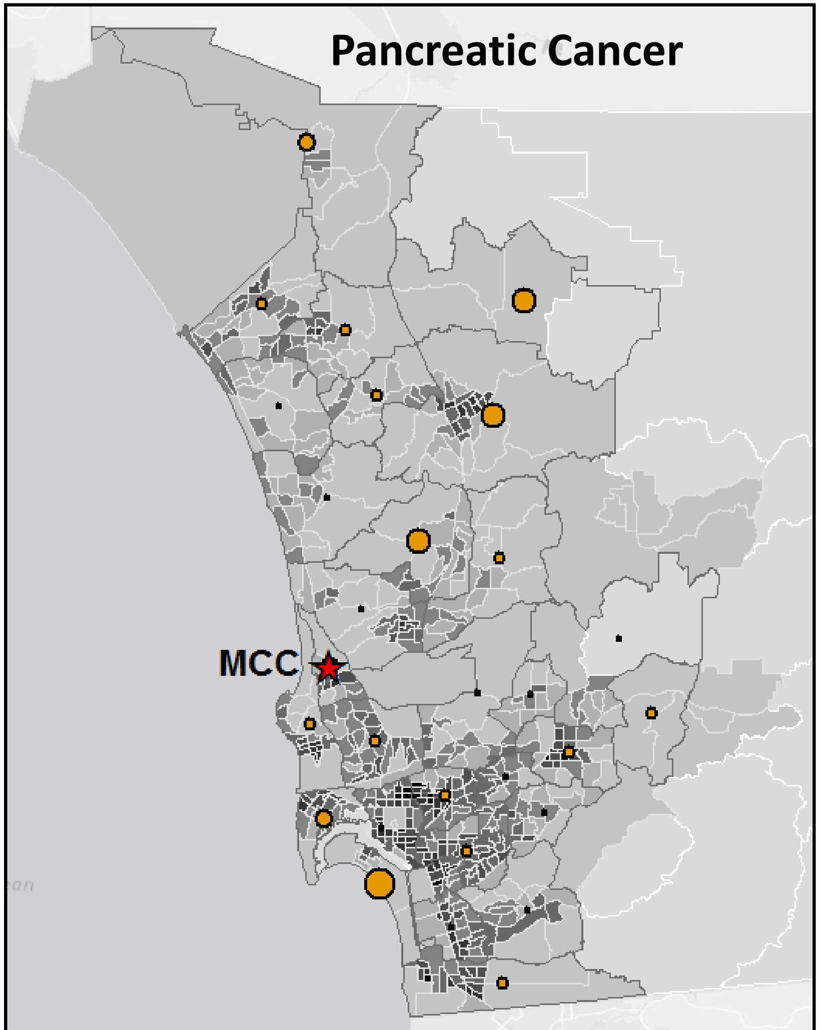
low

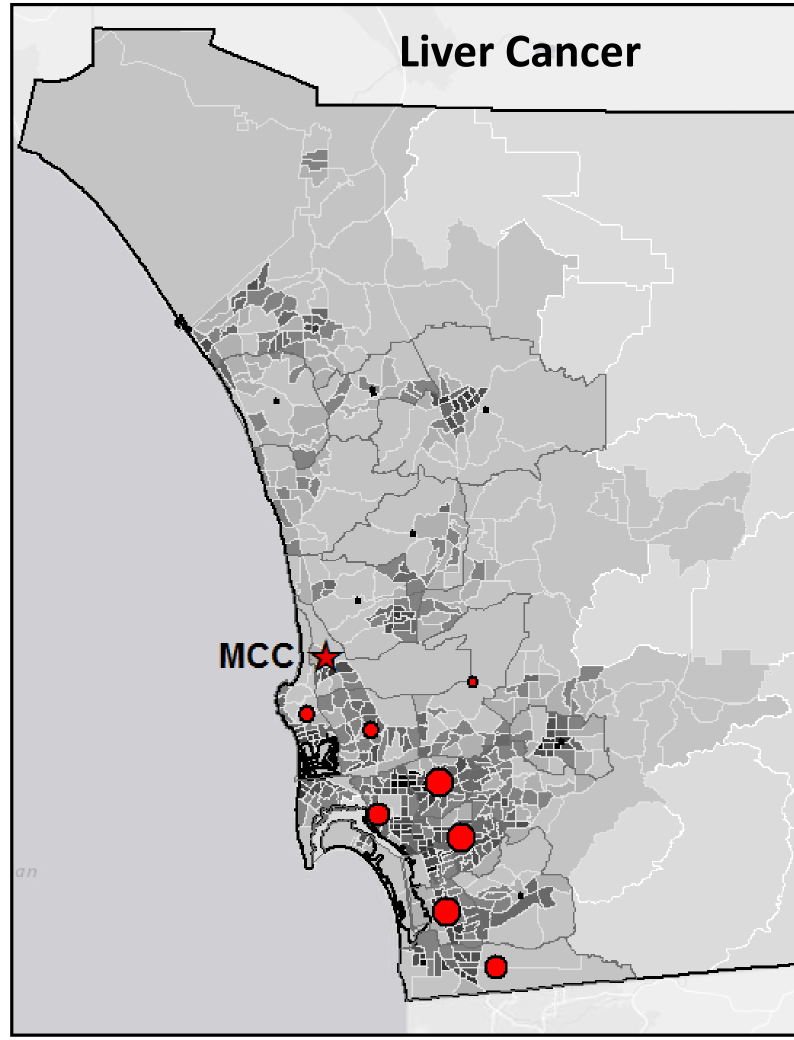
high

Liver Cancer Death Rate

low

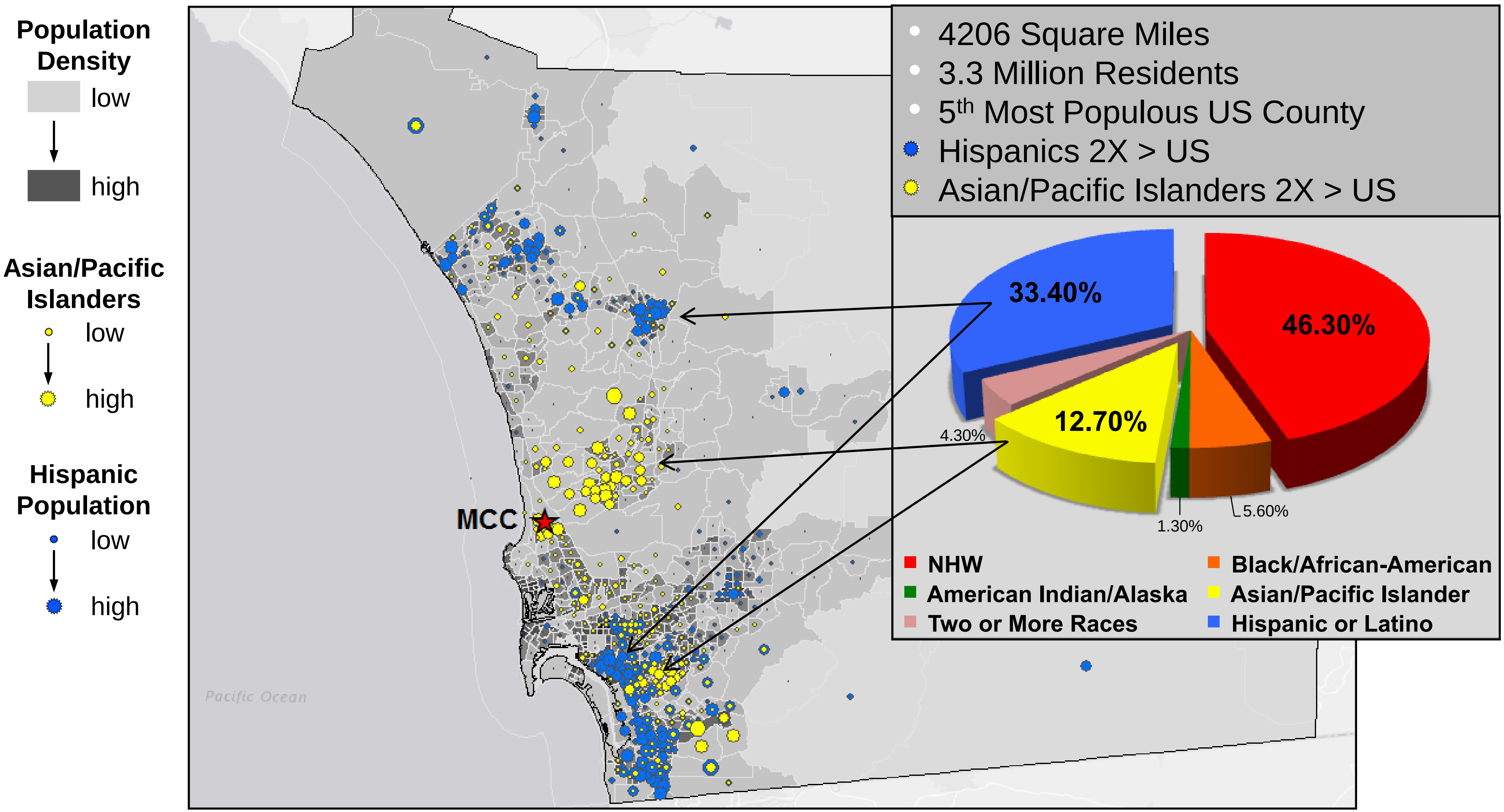
high





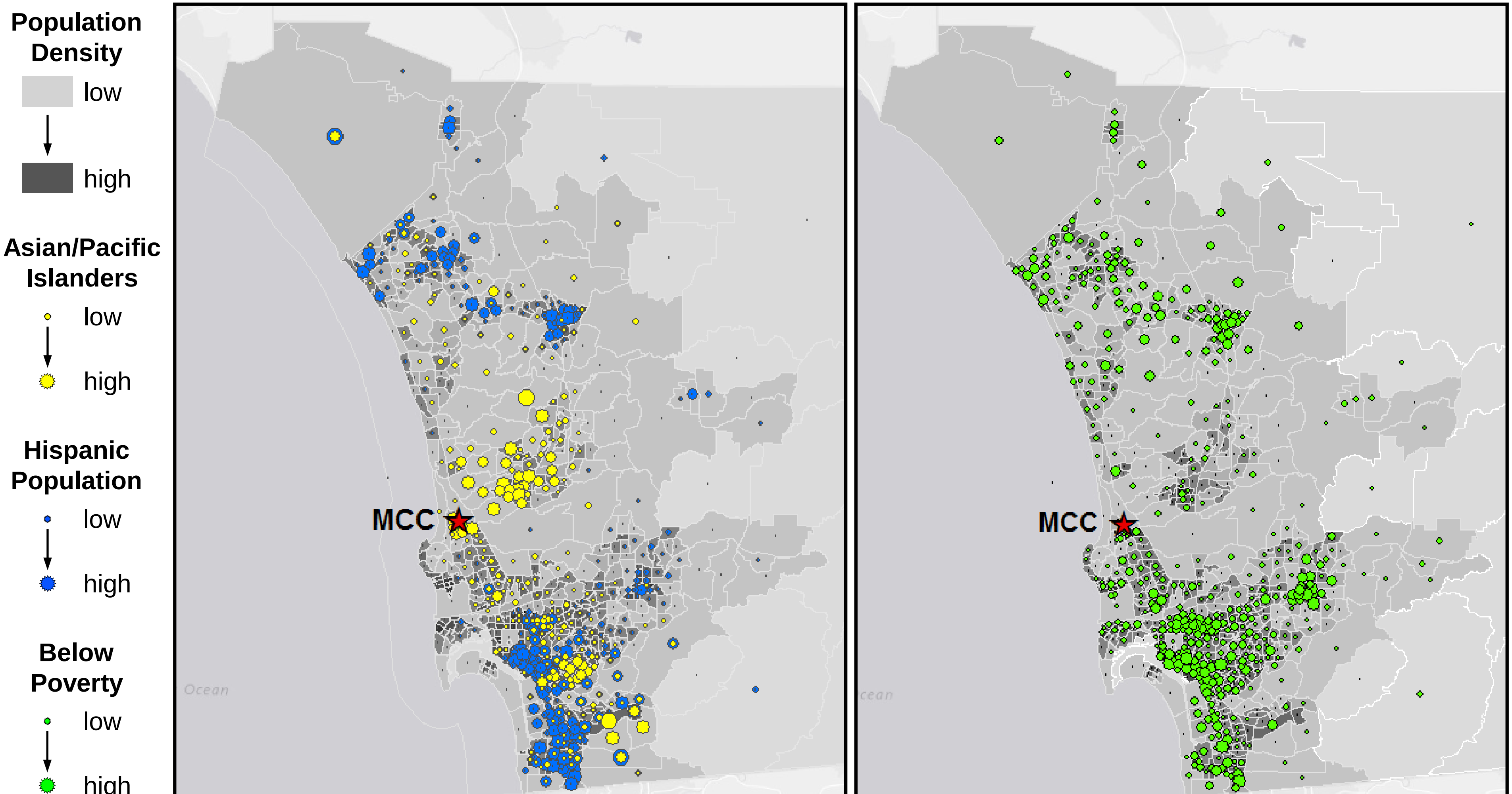
MCC Catchment:
San Diego County Population Demographics

UC San Diego
MOORES CANCER CENTER



MCC Catchment: San Diego County

UC San Diego
MOORES CANCER CENTER

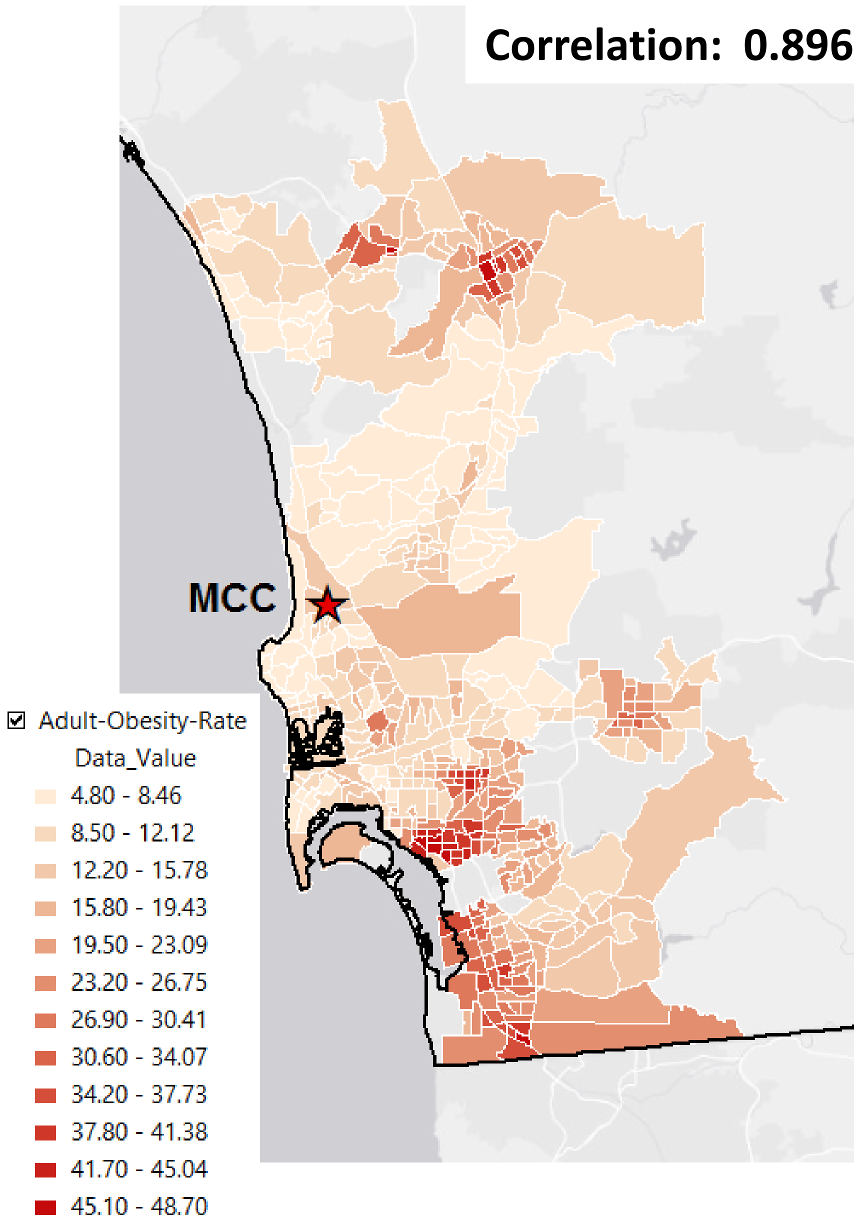


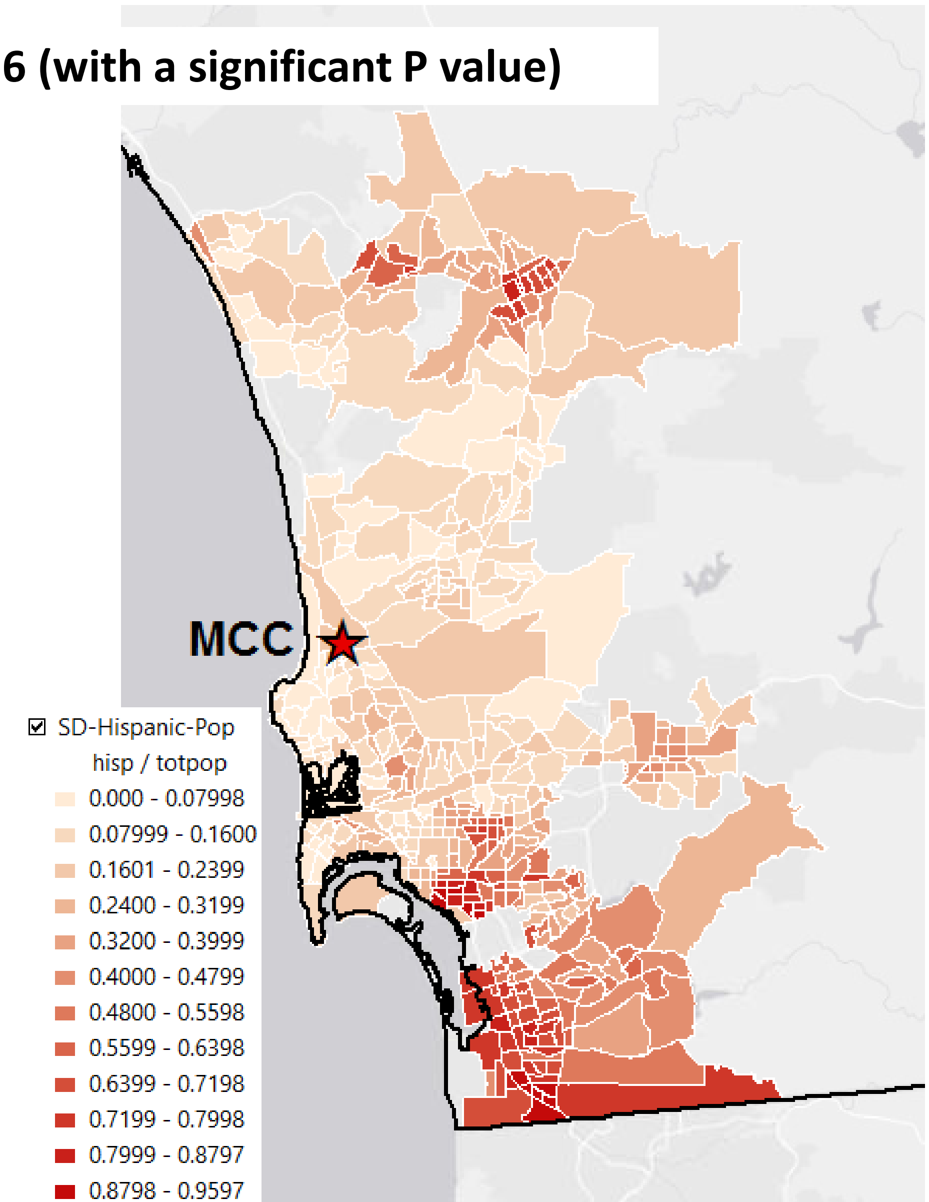
	SDC	US
Non-English Speakers	43.5%	21.0%
Below Poverty Level	15.3%	13.5%

Adult Obesity Rates (%)

Hispanic Population Rates (%)

Correlation: 0.896395136 (with a significant P value)



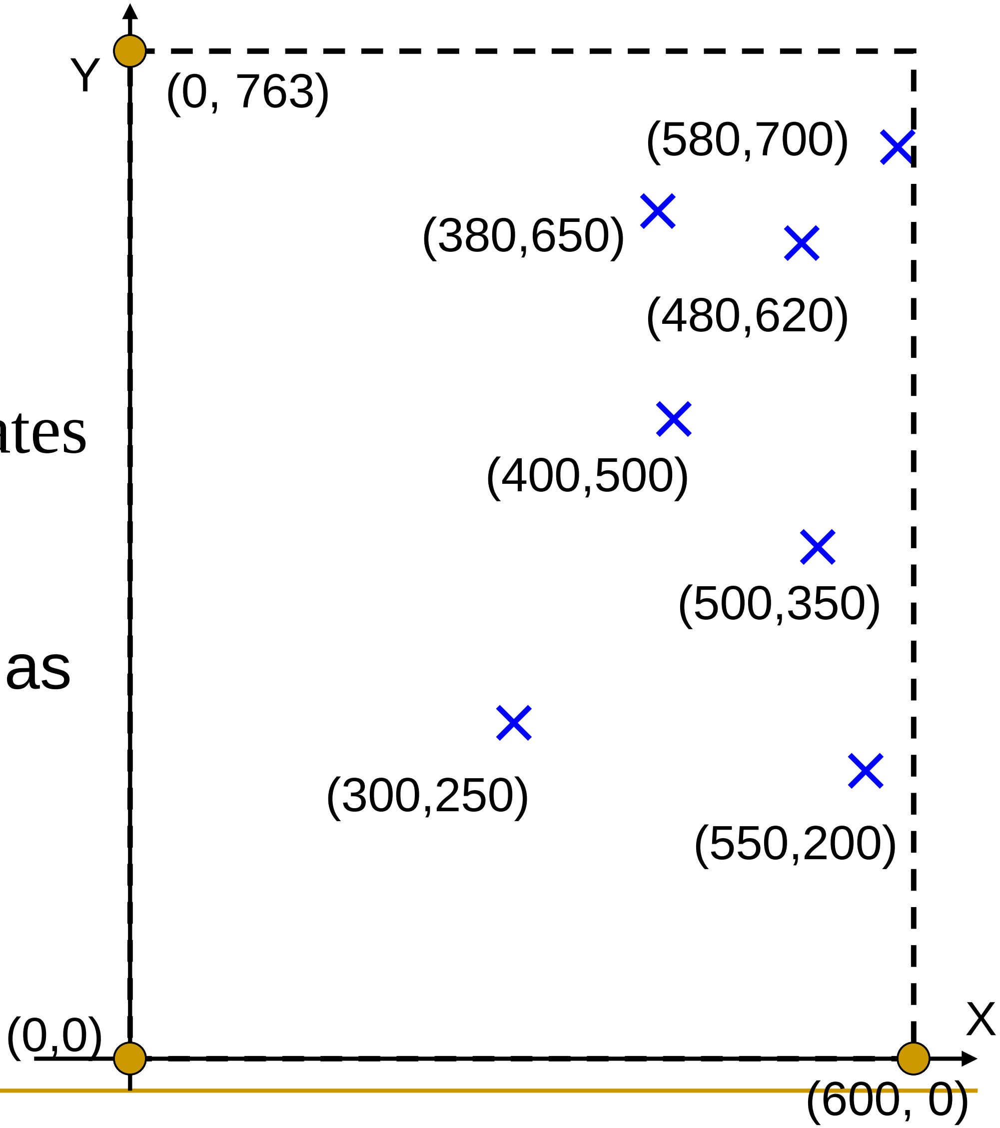


Area Statistics Questions

- 2003 forest fires in San Diego
- Given the map of SD forests
  - What is the average location of these forests?
  - How spread are they?
  - Where do you want to place a fire station?

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: $\bar{X}_C = \frac{\sum x}{n}$

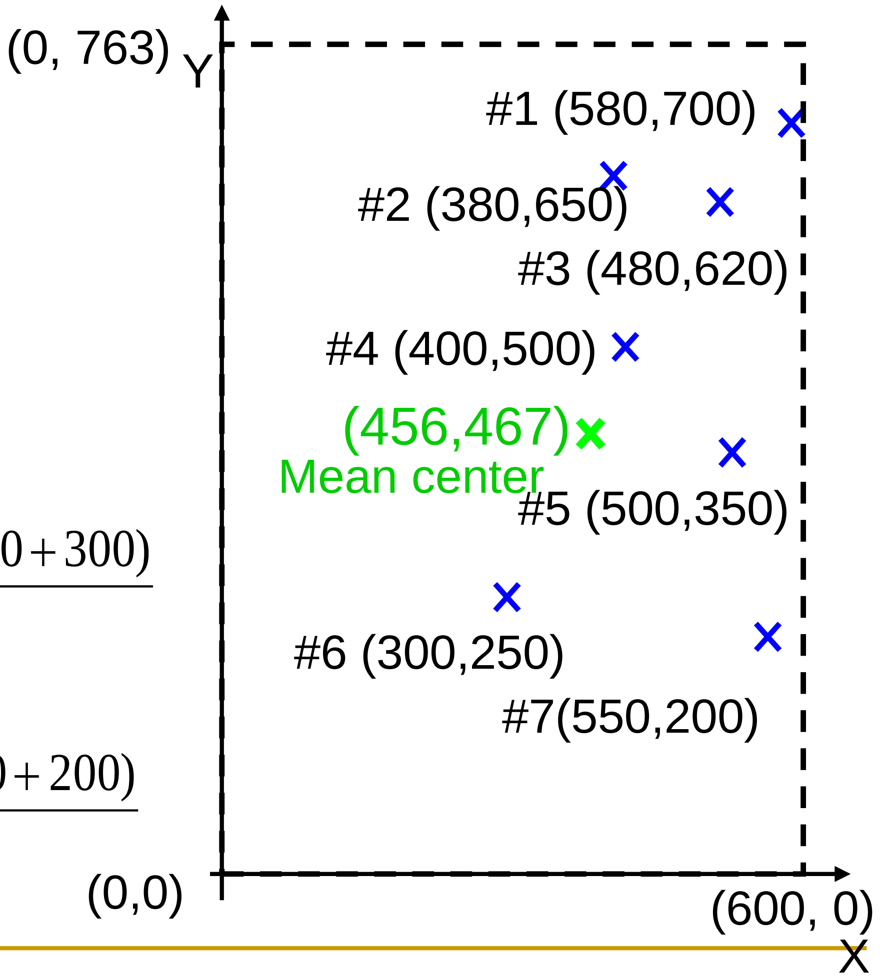
  Mean center of Y: $\bar{Y}_C = \frac{\sum y}{n}$

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

$$= 455.71$$

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

$$= 467.14$$



Standard distance

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

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

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

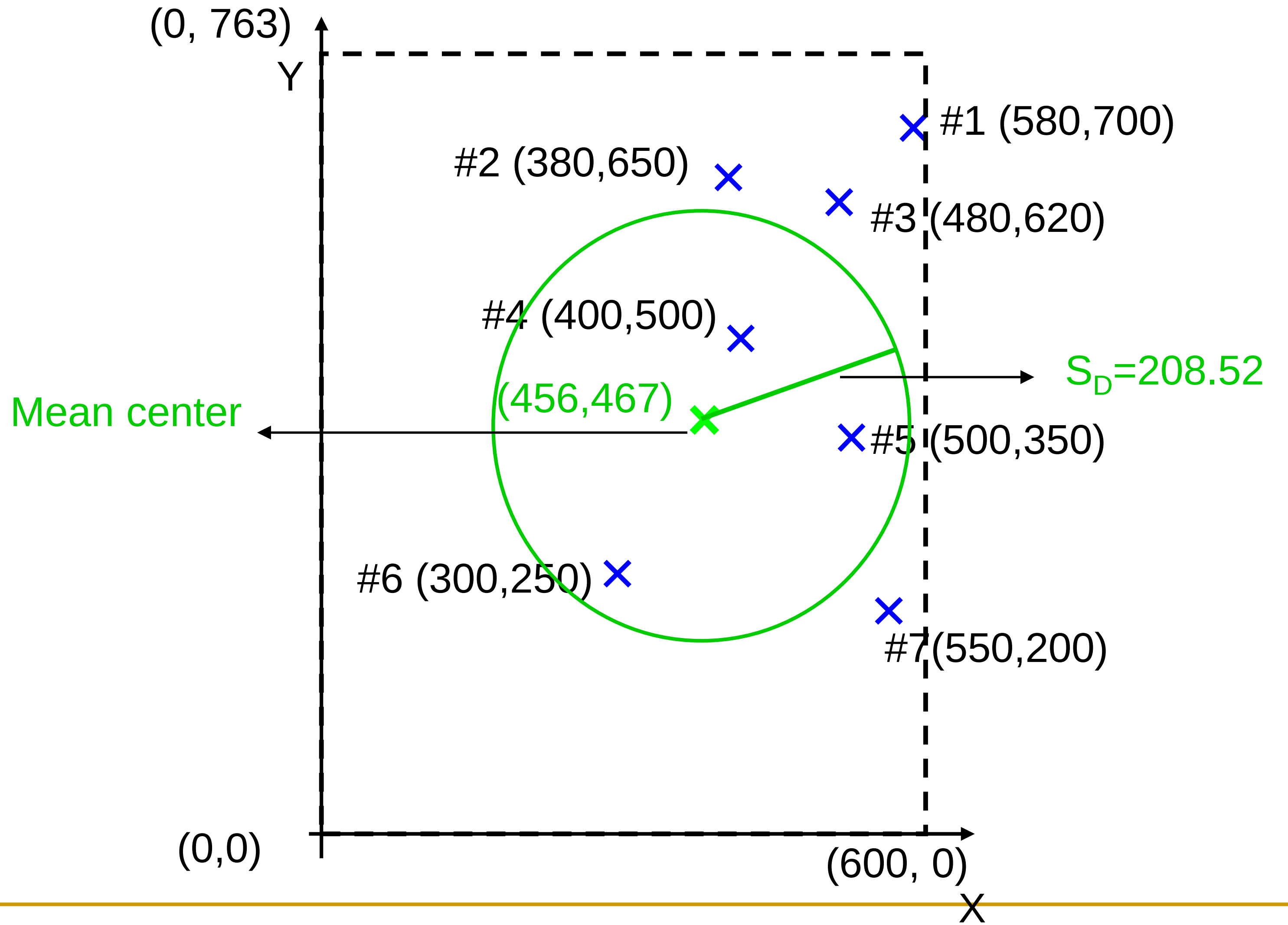
Standard distance

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 X^2	1771900
	$\bar{X}_C = 455.71$		$\bar{Y}_C = 467.14$	

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

$$= \sqrt{\left(\frac{1513700}{7} - 455.71^2\right) + \left(\frac{1771900}{7} - 467.14^2\right)} = 208.52$$

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?

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

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

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

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

Calculation of weighted standard distance

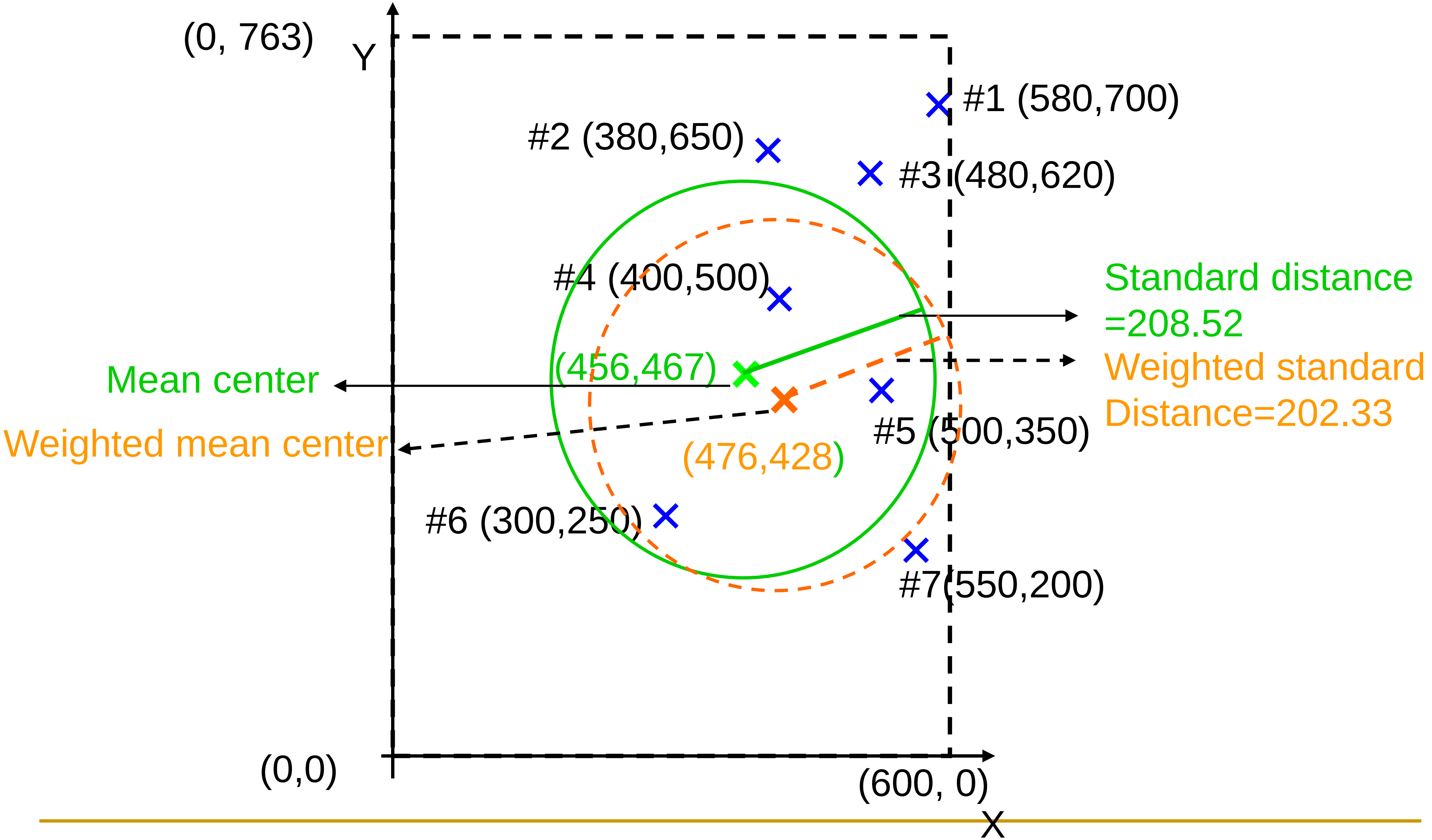
- What if the forests with bigger area (the area of the smallest forest as unit) should have more influence?

Forests	f_i(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

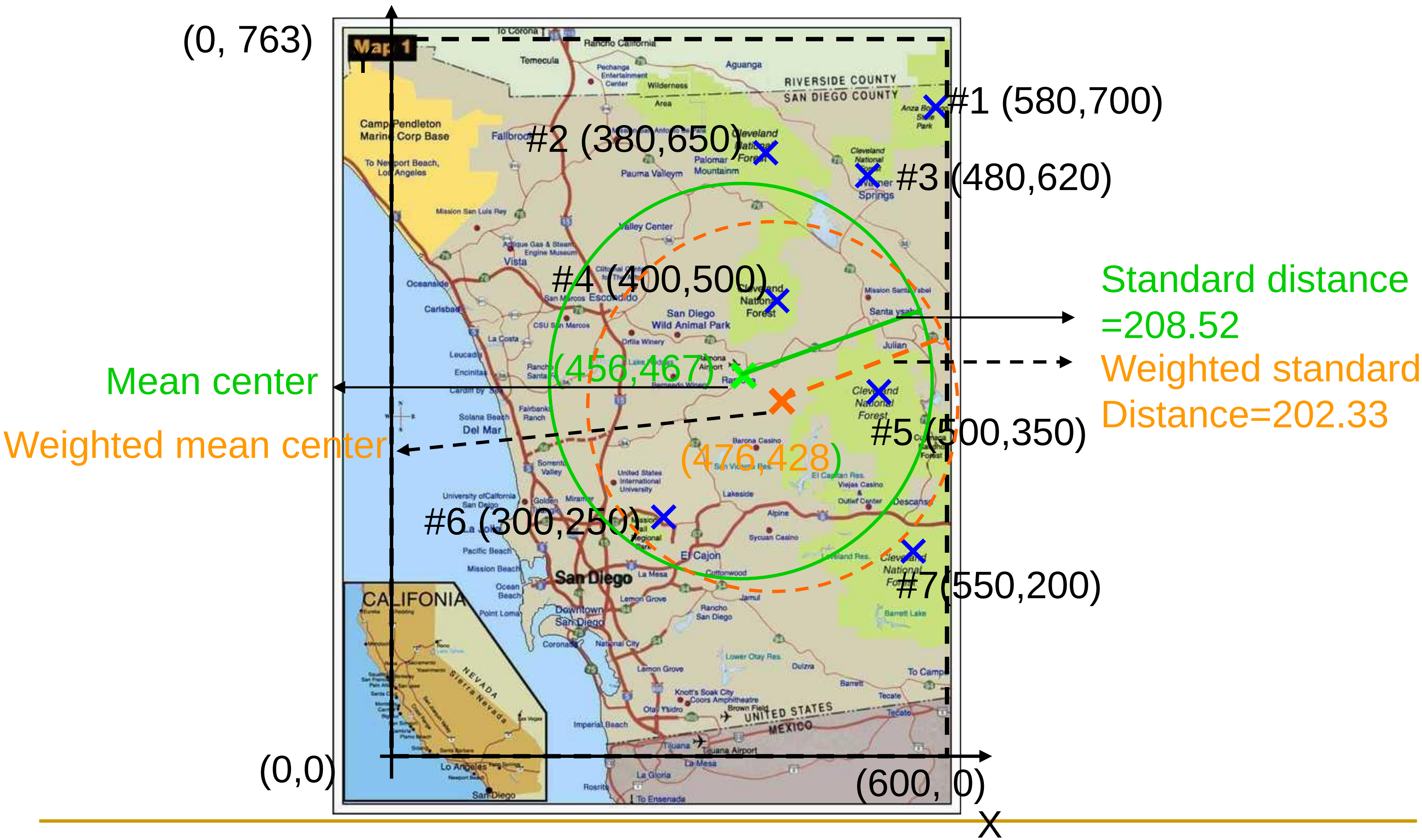
$$S_{WD} = \sqrt{\left(\frac{\sum f_i X_i^2}{\sum f_i} - \overline{X}_{wc}^2\right) + \left(\frac{\sum f_i Y_i^2}{\sum f_i} - \overline{Y}_{wc}^2\right)}$$

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

Standard distance

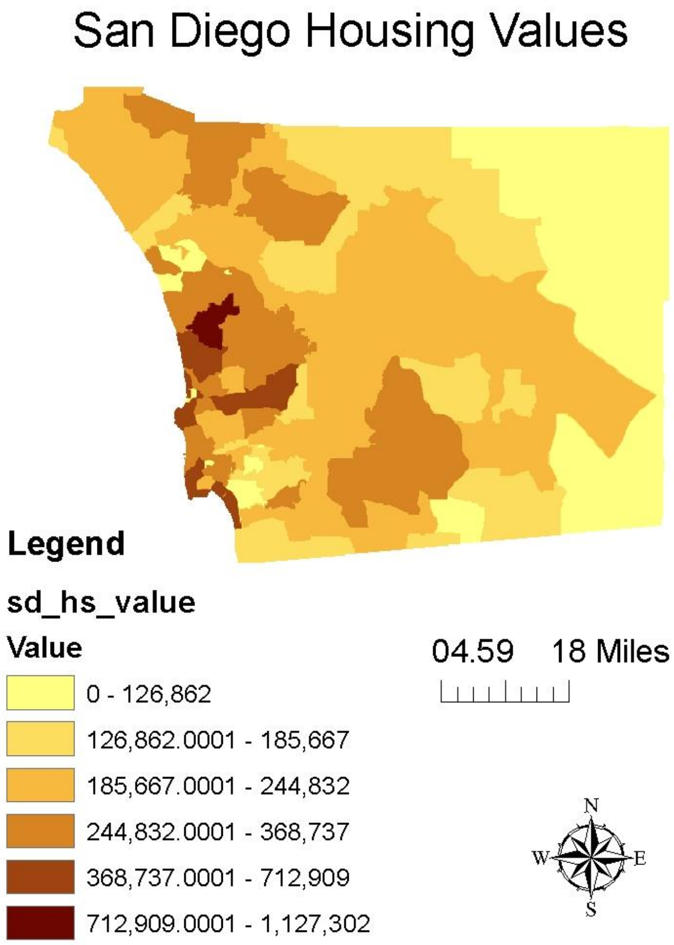


Standard distance



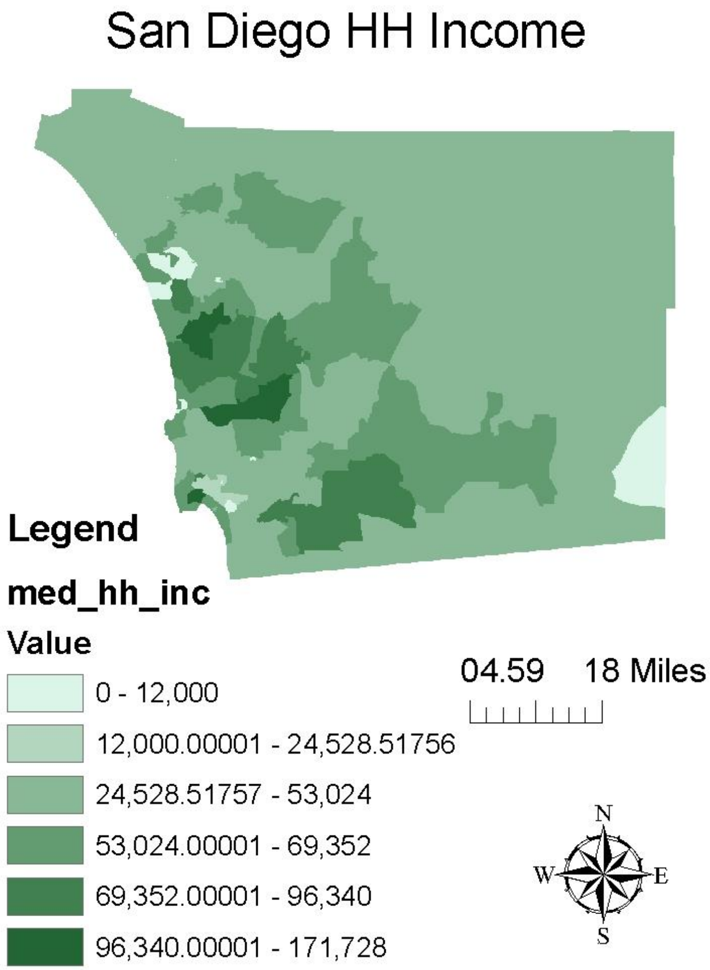
Spatial clustered?

- 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?



Moran's I statistic

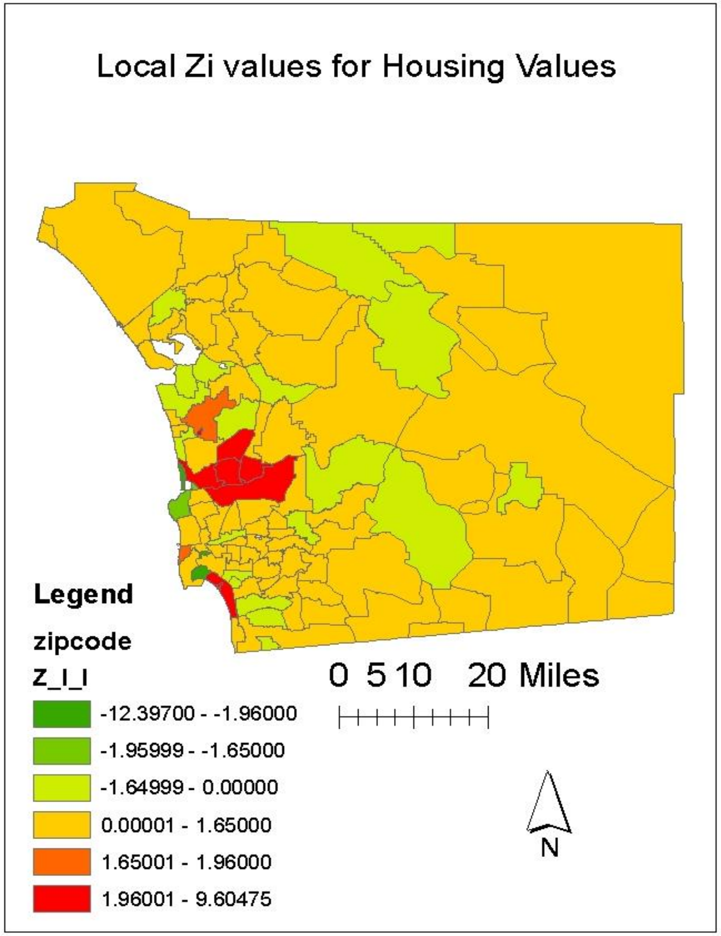
- Global Moran's I
  - Characterize the overall spatial dependence among a set of areal units

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

Covariance

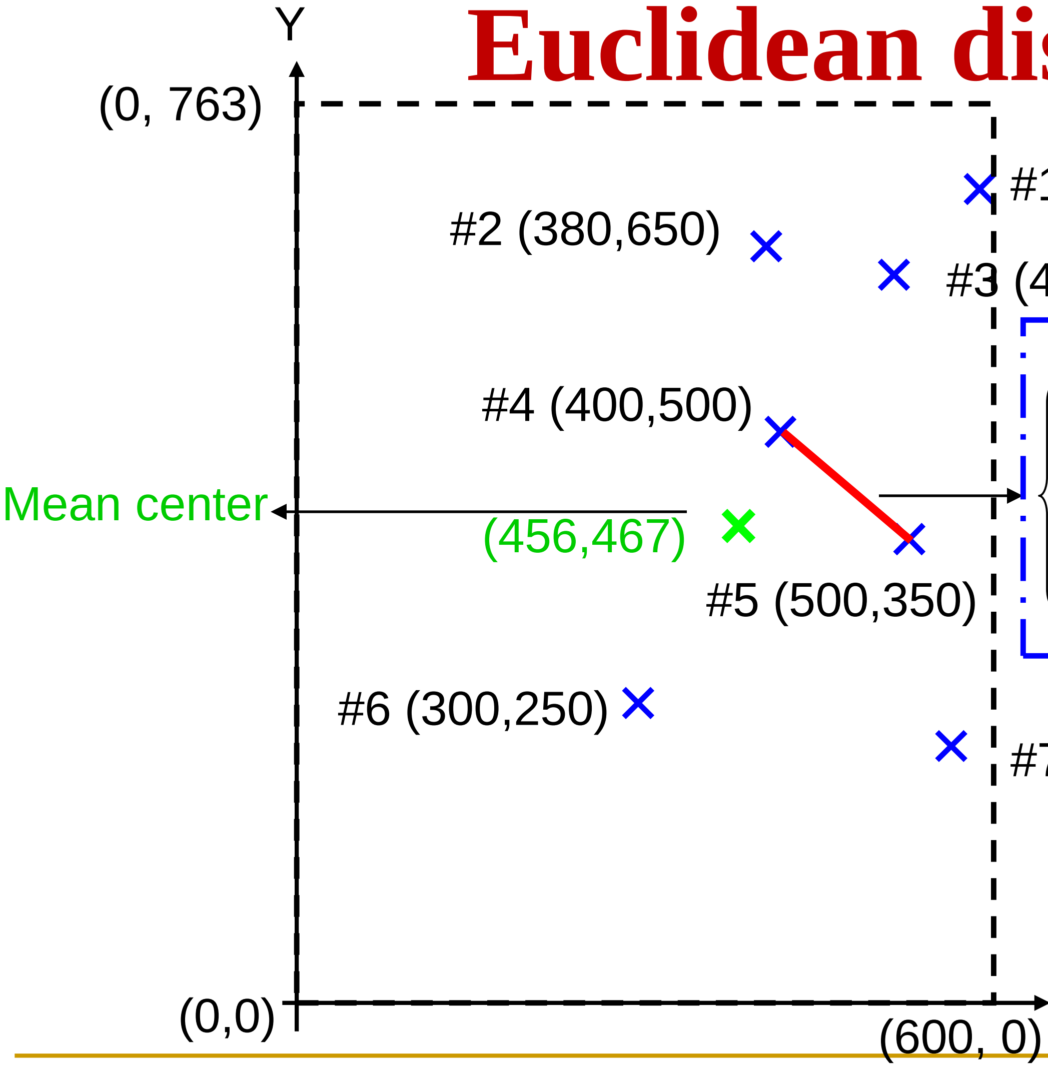
Summary

- Global Moran's I and local I_i have different equations, one for the entire region and one for a location. But for both of them (I and I_i), or the associated scores (Z and Z_i)
  - **Big positive values → positive spatial autocorrelation**
  - **Big negative values → negative spatial autocorrelation**
  - **Moderate values → random pattern**



Network Analysis: Shortest routes

Euclidean distance



$$
\begin{cases}
d_i = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2} \\
= \sqrt{(500-400)^2 + (350-500)^2} \\
= 180.28
\end{cases}
$$

Manhattan Distance

- Euclidean median
  - Find (X_e, Y_e) such that

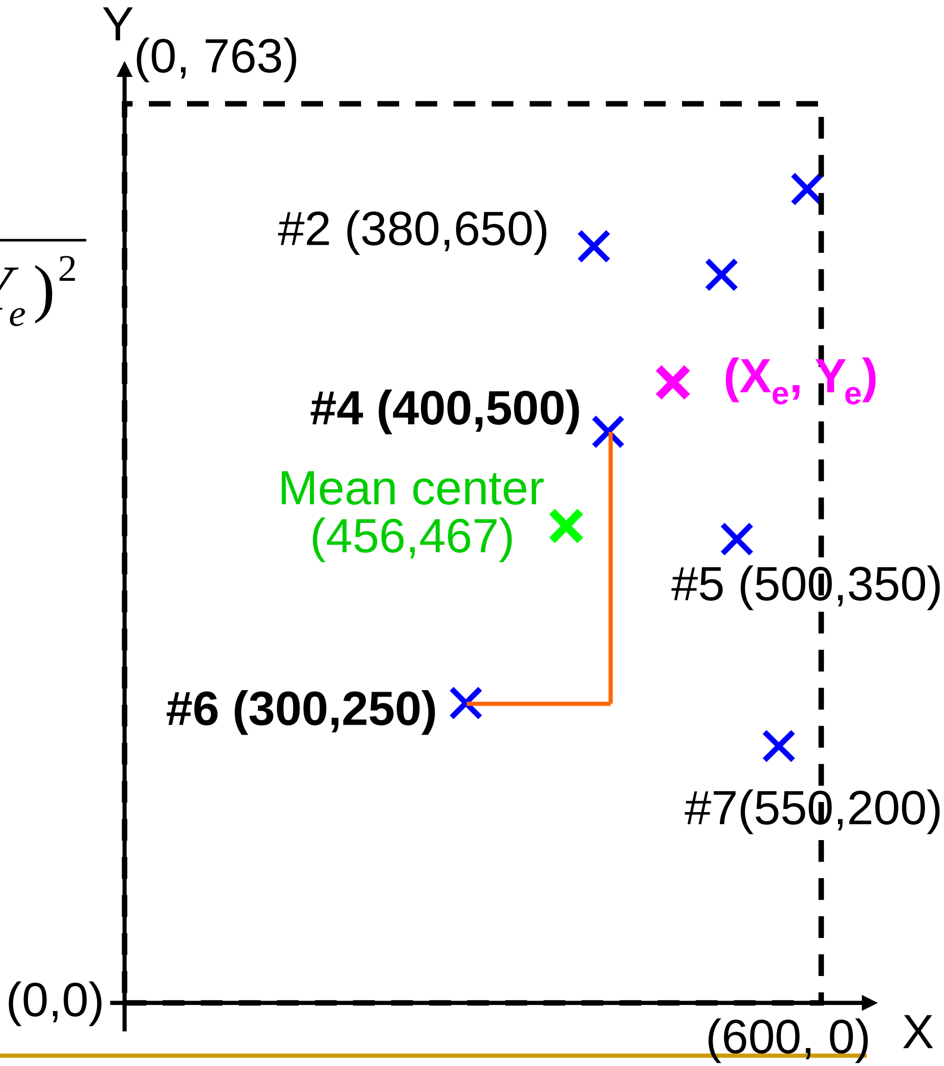
$$d_e = \sum \sqrt{(X_i - X_e)^2 + (Y_i - Y_e)^2}$$

    is minimized
  - Need iterative algorithms
  - Location of fire station
- Manhattan median

$$d_{ij} = |X_i - X_j| + |Y_i - Y_j|$$

$$= |400 - 300| + |500 - 250|$$

$$= 350$$



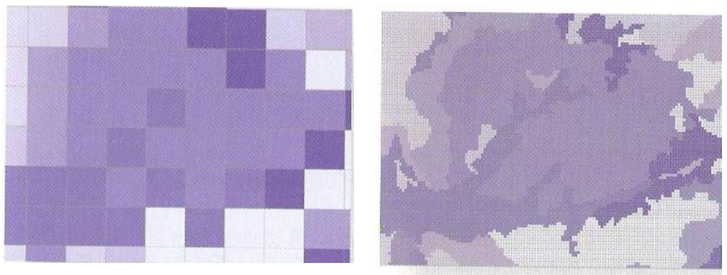
Summary

- What are spatial data?
- Mean center
- Weighted mean center
- Standard distance
- Weighted standard distance
- Euclidean median
- Manhattan median

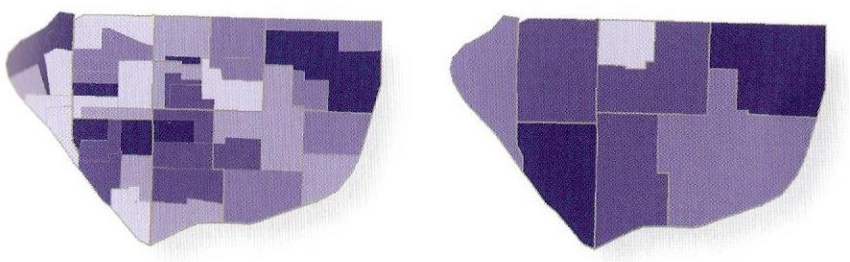
} **Calculate in GIS environment**

Spatial resolution

- Patterns or relationships are scale dependent
  - Hierarchical structures (blocks → block groups → census tracks…)
  - Cell size: # of cells vary and spatial patterns masked or overemphasized
- How to decide
  - The goal/context of your study
  - Test different sizes (Weeks et al. article: 250, 500, and 1,000 m)



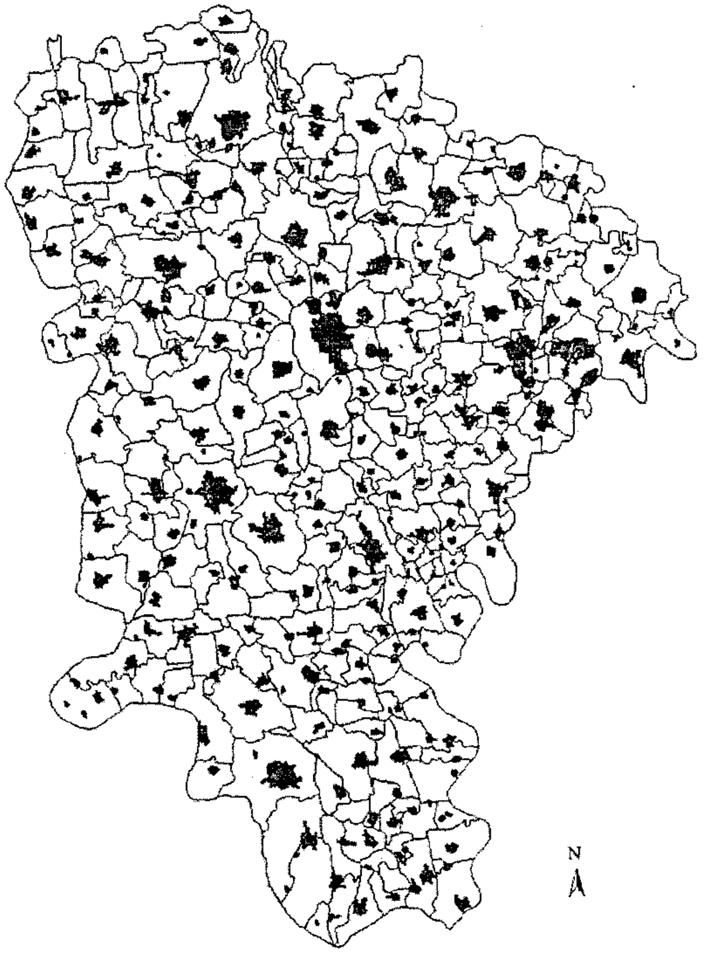
Vegetation types at large (left) and small cells (right)

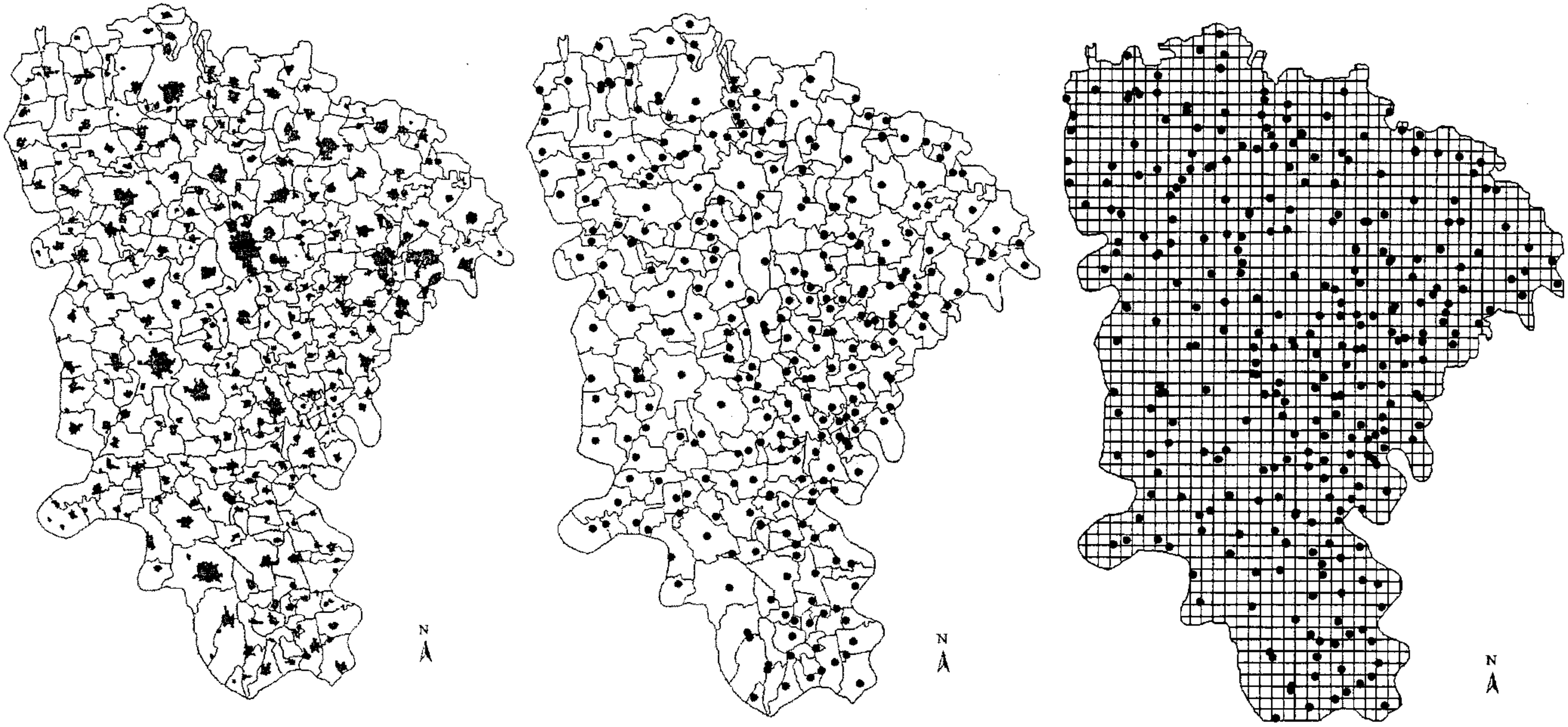


% of seniors at block groups (left) and census tracts (right)

Administrative units

- Default units of study
  - May not be the best
  - Many events/phenomena have nothing to do with boundaries drawn by humans
- How to handle
  - Include events/phenomena outside your study site boundary
  - Use other methods to “reallocate” the events /phenomena (Weeks et al. article; see next page)





A. Locate human settlements using RS data

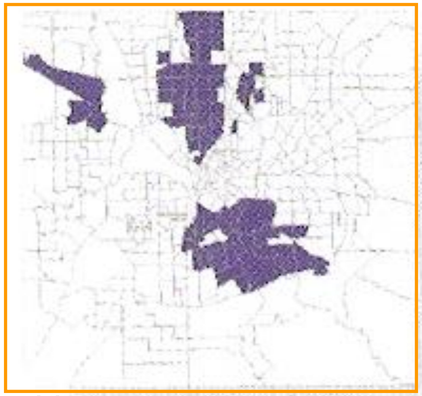
B. Find their centroids

C. Impose grids.

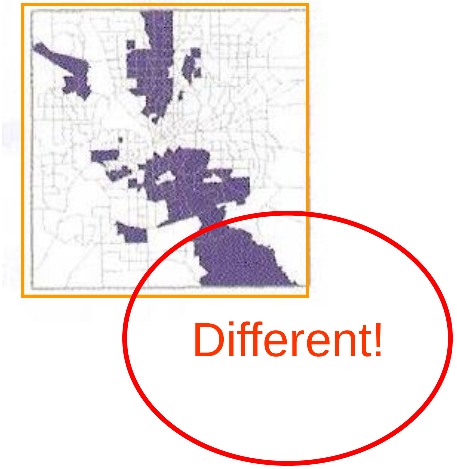
Edge effects

- 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_i)



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

d. More areas are significant

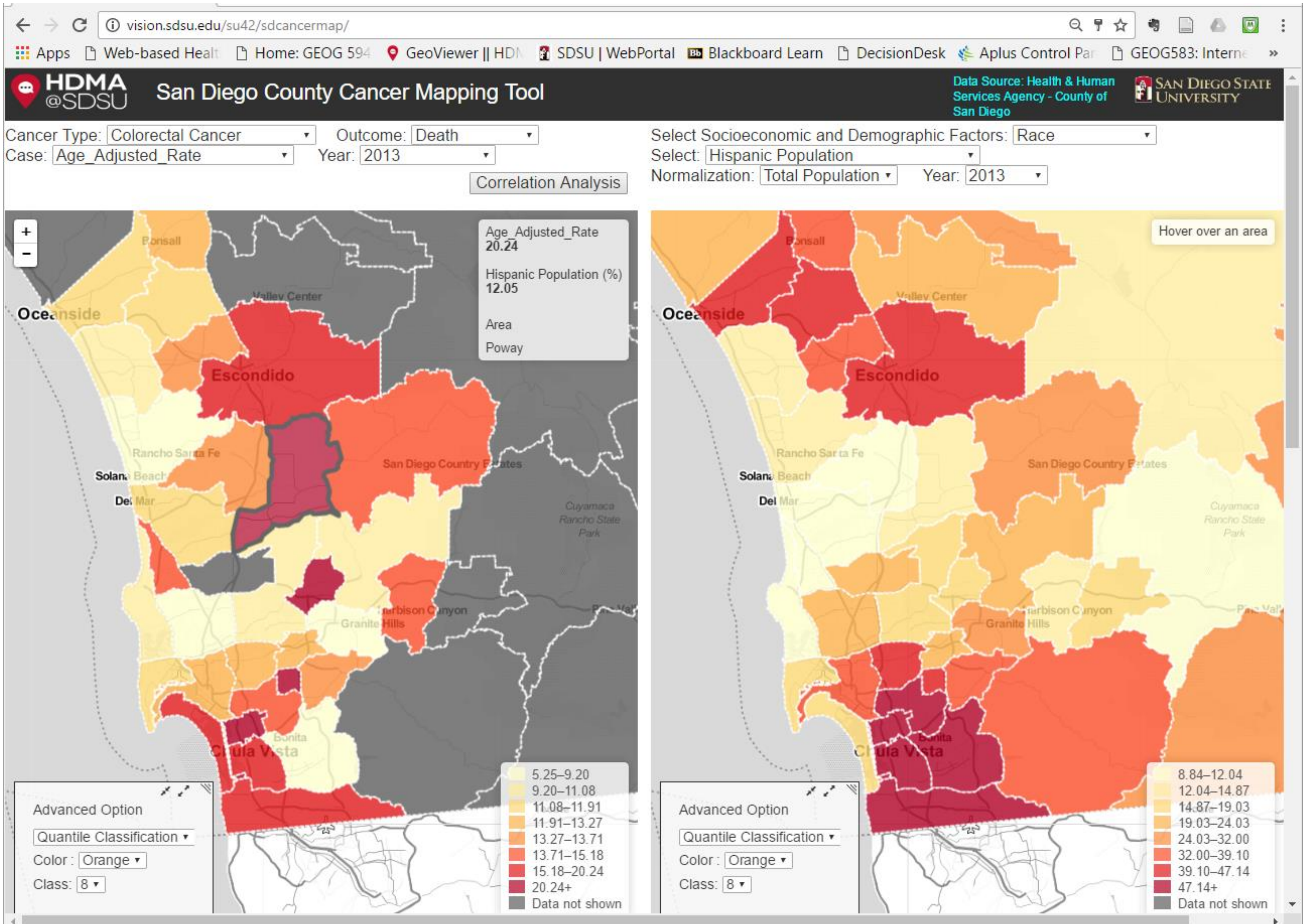
Applying Spatial Statistics

- **Visualizing spatial data (Software: GeoDa)**
  - 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

HDMA @SDSU THE CENTER FOR HUMAN DYNAMICS IN THE MOBILE AGE

Visualizing Cancer Disparities (left map) with Socioeconomic Variables (right map)

SAN DIEGO STATE UNIVERSITY



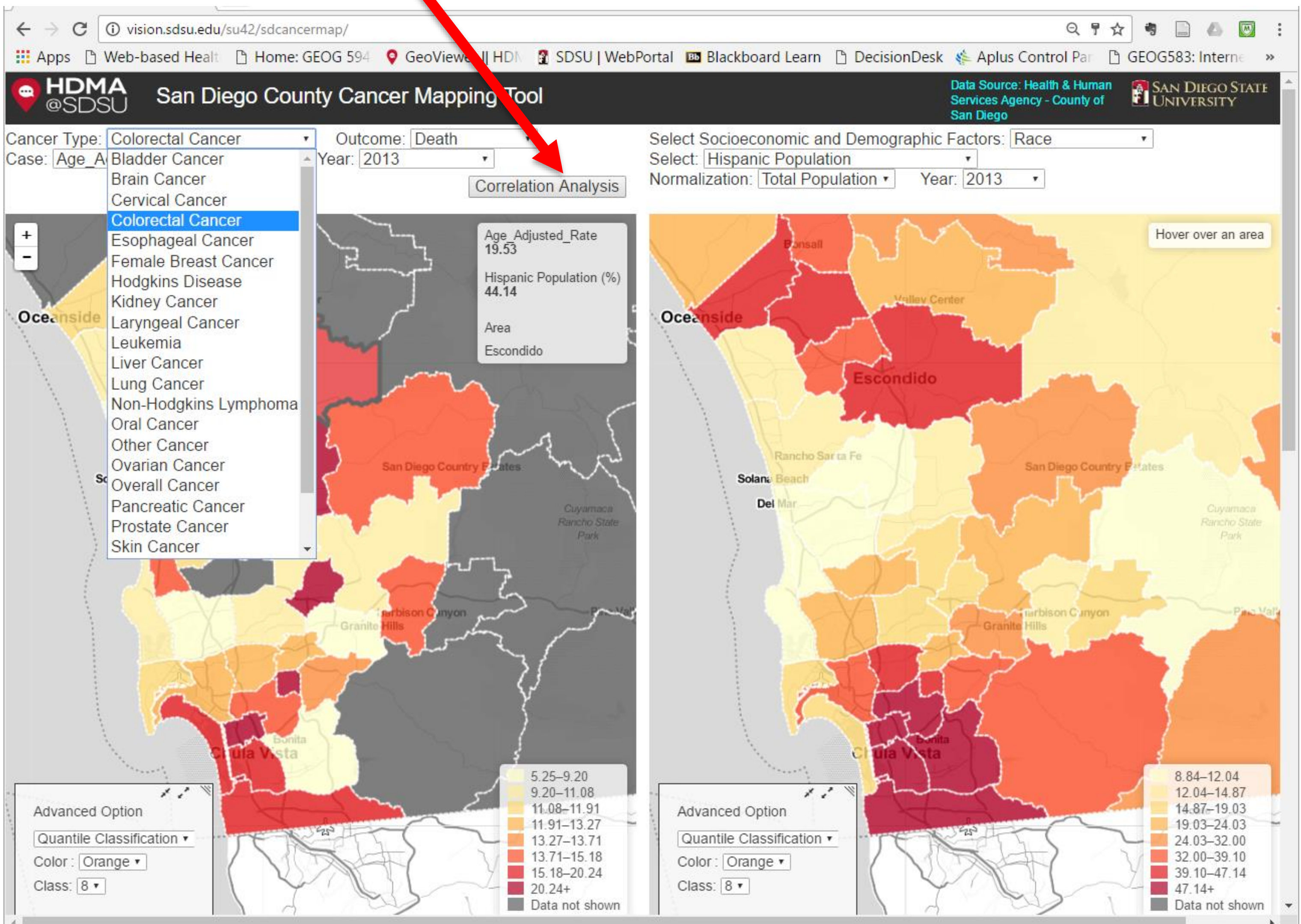
HDMA @SDSU THE CENTER FOR HUMAN DYNAMICS IN THE MOBILE AGE

Correlation Analysis

SAN DIEGO STATE UNIVERSITY

The correlation analysis is performed after this button click

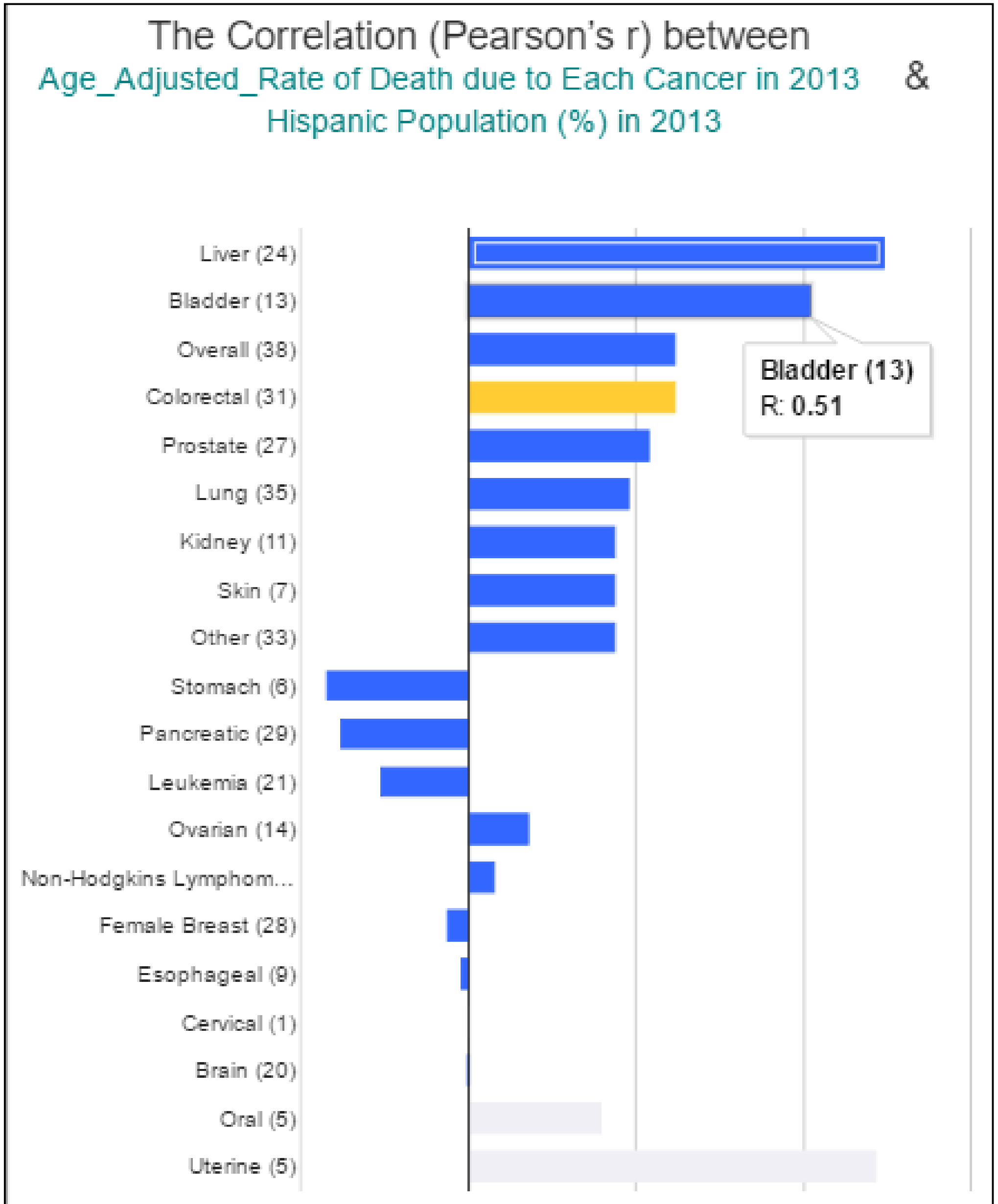
Side-by-Side Synchronous Visualization of Cancer Mortality and Socioeconomic & Demographic Variables



HDMA @SDSU THE CENTER FOR HUMAN DYNAMICS IN THE MOBILE AGE

Compare ALL Cancers with Hispanic Population (at the SRA level)

San Diego State University



- **Yellow bars** represent the correlation that users have selected.
- The **grey bars** represent that the correlation result might not be reliable because the sample size is too small (fewer than 6).
- The number in the parentheses represents **the sample size**
- Pearson's r ranges from -1 to 1.

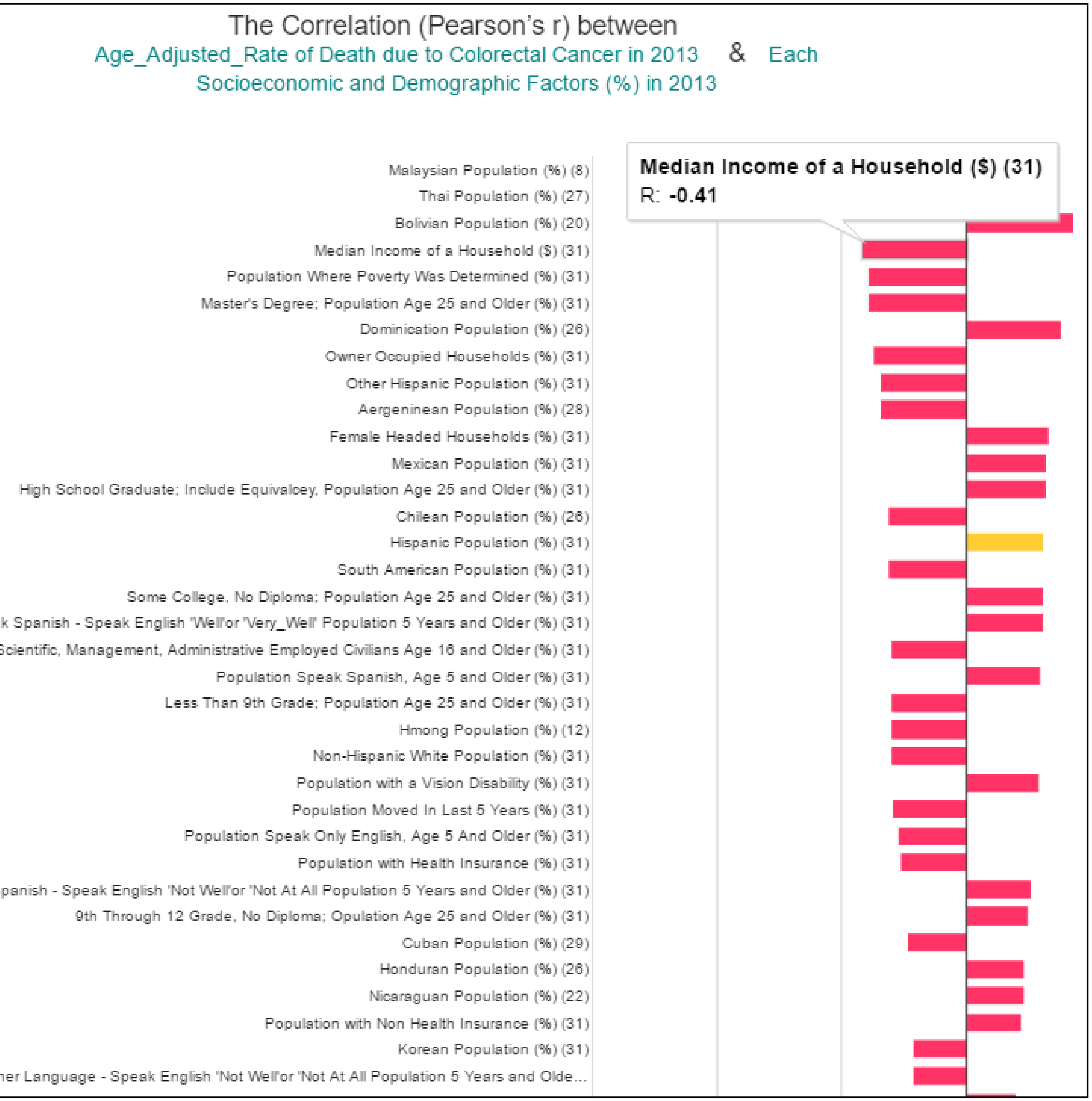
- **Hispanic** Population has **higher correlation** in **Liver, Bladder, and Colorectal Cancer Mortality Rates (Age_Adjusted).**

- **Negative correlation** with **Stomach, Pancreatic and Leukemia** Cancers.

HDMA @SDSU | THE CENTER FOR HUMAN DYNAMICS IN THE MOBILE AGE

Compare Colorectal Cancer Mortality Rate with All Socioeconomic Variables (from census data).

SAN DIEGO STATE UNIVERSITY



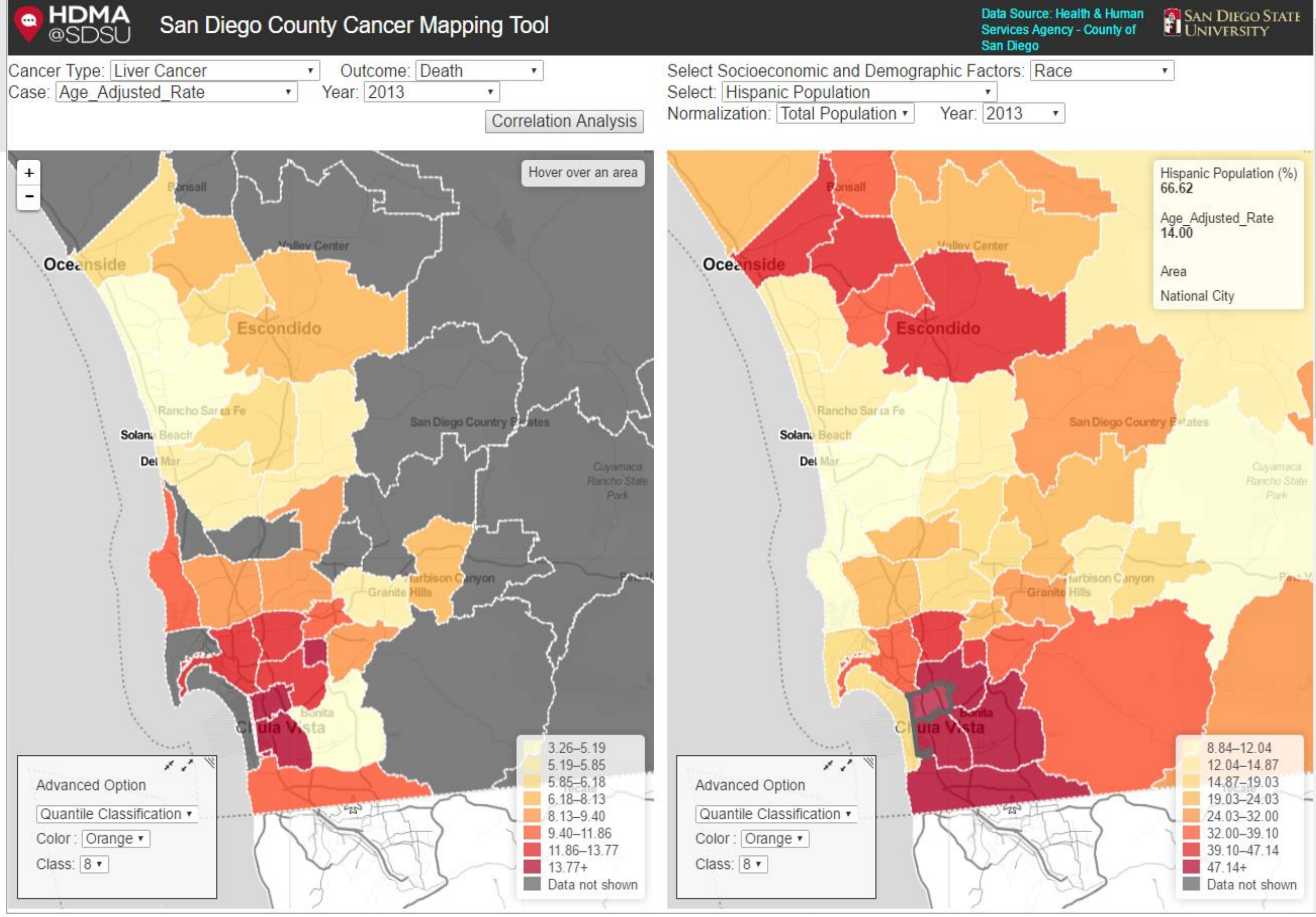
- **Yellow bars** represent the correlation that users have selected.
- The **grey bars** represent that the correlation result might not be reliable because the sample size is too small (fewer than 6).
- The number in the parentheses represents **the sample size**
- Pearson's r ranges from -1 to 1.

- **Colorectal Cancer Mortality Rate** has **higher correlation** with **lower education** levels and, **% of female headed households**.

- **Negative correlation** with **higher median income households**, and **% with masters degree**.

HDMA @SDSU | THE CENTER FOR HUMAN DYNAMICS IN THE MOBILE AGE

Correlation is not equal to **Causality**



Pearson's r = 0.62

(One Example of Liver Cancer with Hispanic population)

One Case Study Example: Mortality rate of **liver cancer** is likely to be high in the region where the density of population with **Hispanic population**, but it does not necessarily mean that ethnicity is the only factor to increase the risk of death from liver cancer. Other factors that correlate with the racial composition of a particular area could also be confounding the observed relationship between Hispanic race and liver cancer mortality, such as limited access to health care or poor health behaviors like drinking alcohol, both of which are known to increase the risk of death from liver cancer.

HDMA @SDSU THE CENTER FOR HUMAN DYNAMICS IN THE MOBILE AGE

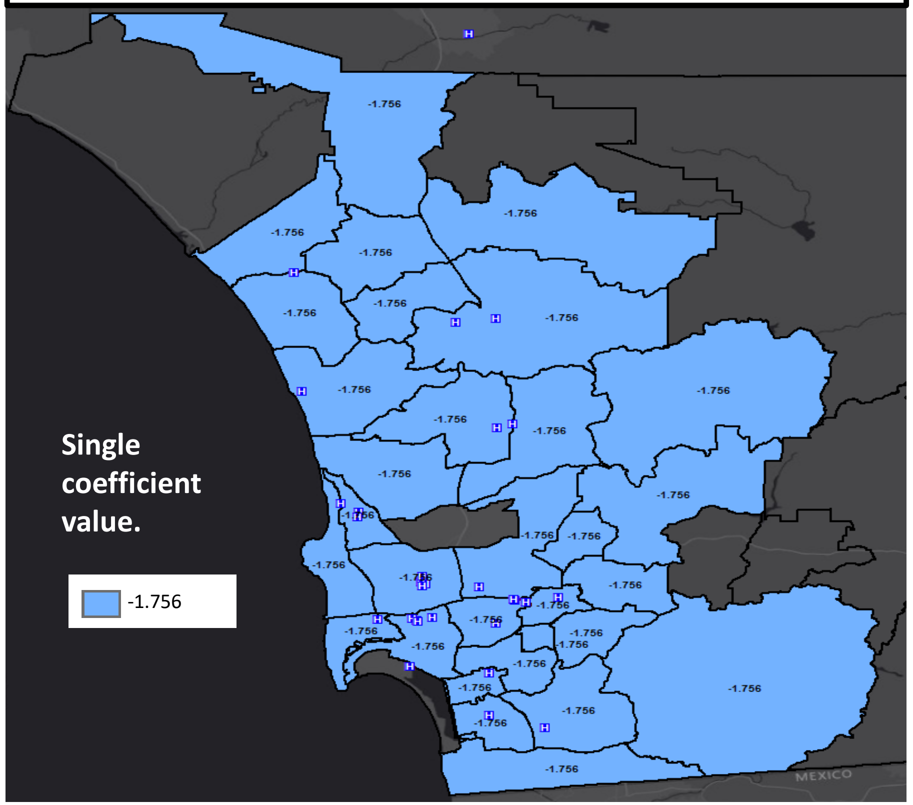
Geographic Weighted Regression

SAN DIEGO STATE UNIVERSITY

Dependent variable: Female Breast Cancer Age Adjusted Mortality Rate
Independent variables: % black, % Asian, %Hispanic, affluence score, disadvantage score, stability score, **% insured**, % foreign born

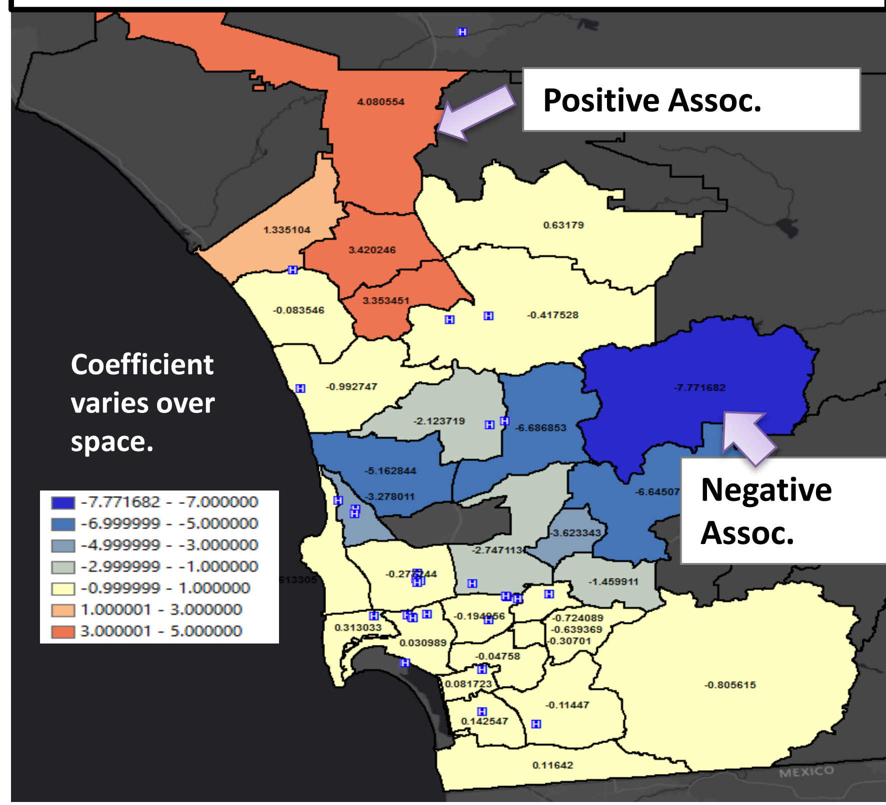
Map created by Nara

GLM: % Insured coefficient



Generalized Linear Model (Poisson Regression)

GWR: % Insured coefficient



Generalized GWR (Poisson GW Regression)