Image from NASA:
https://commons.wikimedia.org/wiki/File:Earth_Western_Hemisphere_transparent_background.png#filelinks

GIScience

Spatial Statistics

WV VIEW

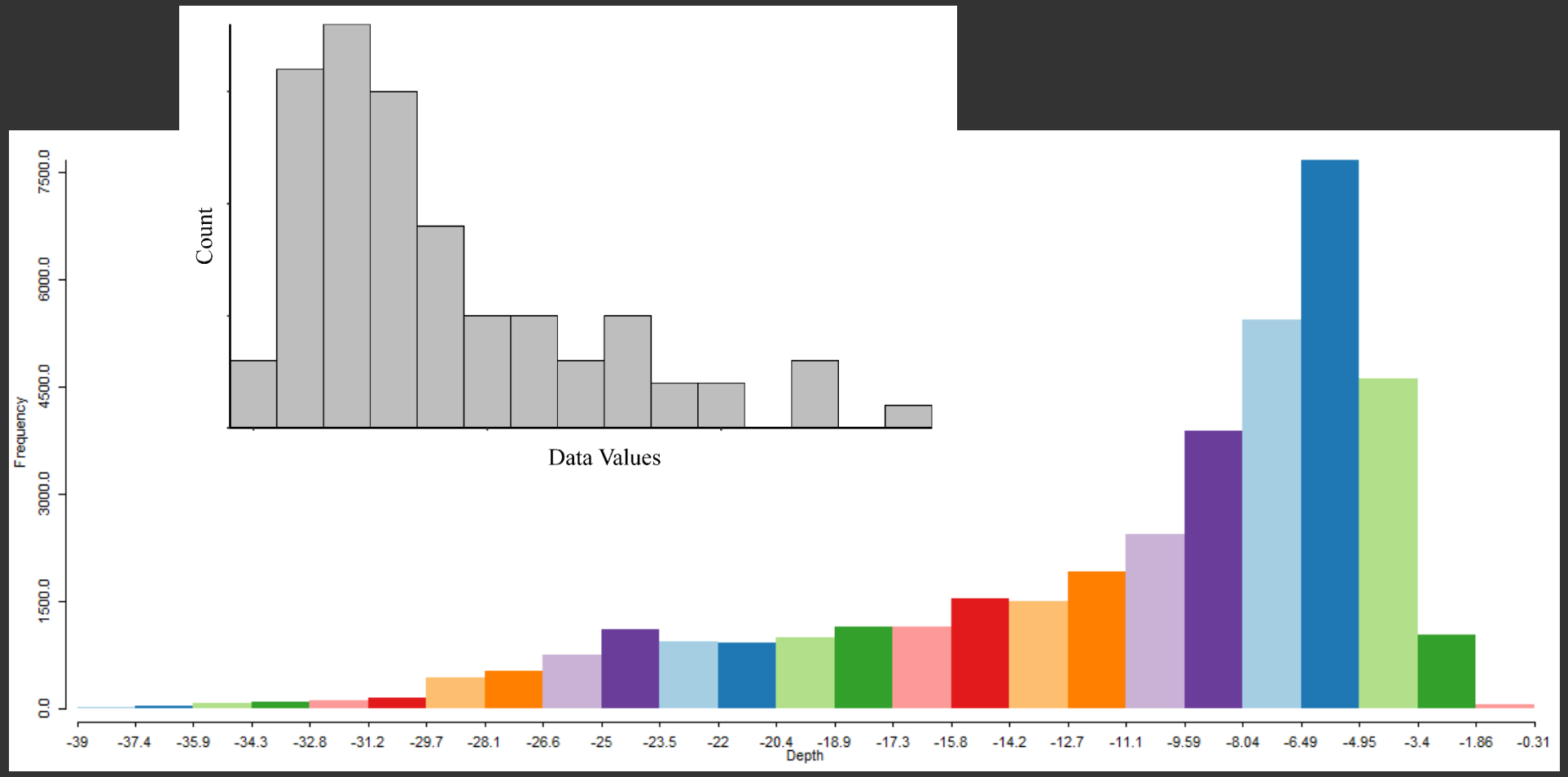
AmericaView℠
Empowering Earth Observation Education
AmericaView.org - Est. 2003

What is unique about geographic data?

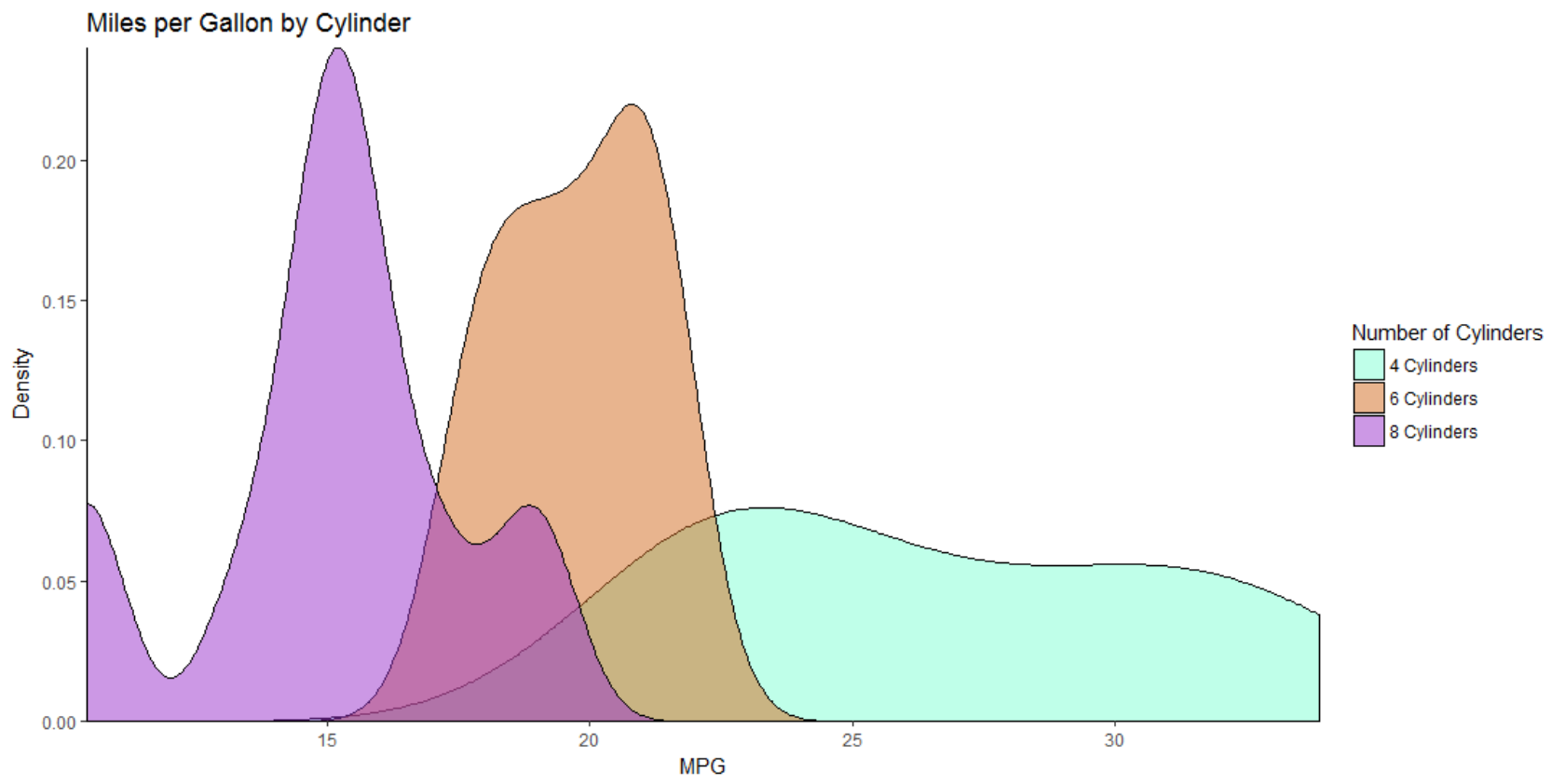
1. Occurs over a map surface
2. Spatial heterogeneity/spatial autocorrelation
3. Temporal heterogeneity/temporal autocorrelation
4. Modifiable areal unit problem

Exploring Your Data

Data Distribution: Histograms



Data Distributions: Kernel Density Plots



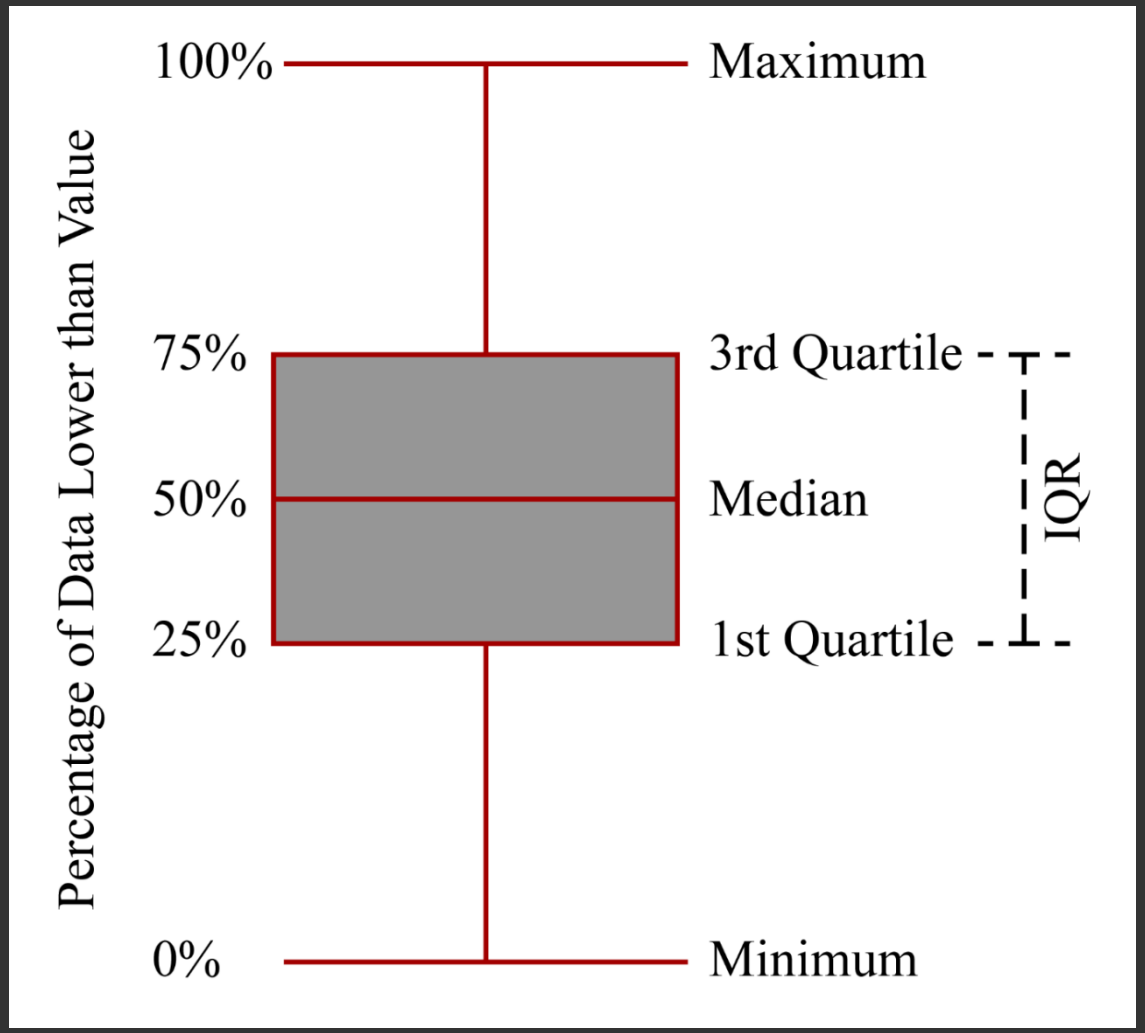
Measures of Central Tendency

- Population Mean = $\mu = (\Sigma X_i) / N$
- Sample Mean = $x_bar = (\Sigma x_i) / n$
- Median = Most middle number, 50% percentile, 2nd quartile
- Mode = Most common number

Measure of Dispersion or Variability

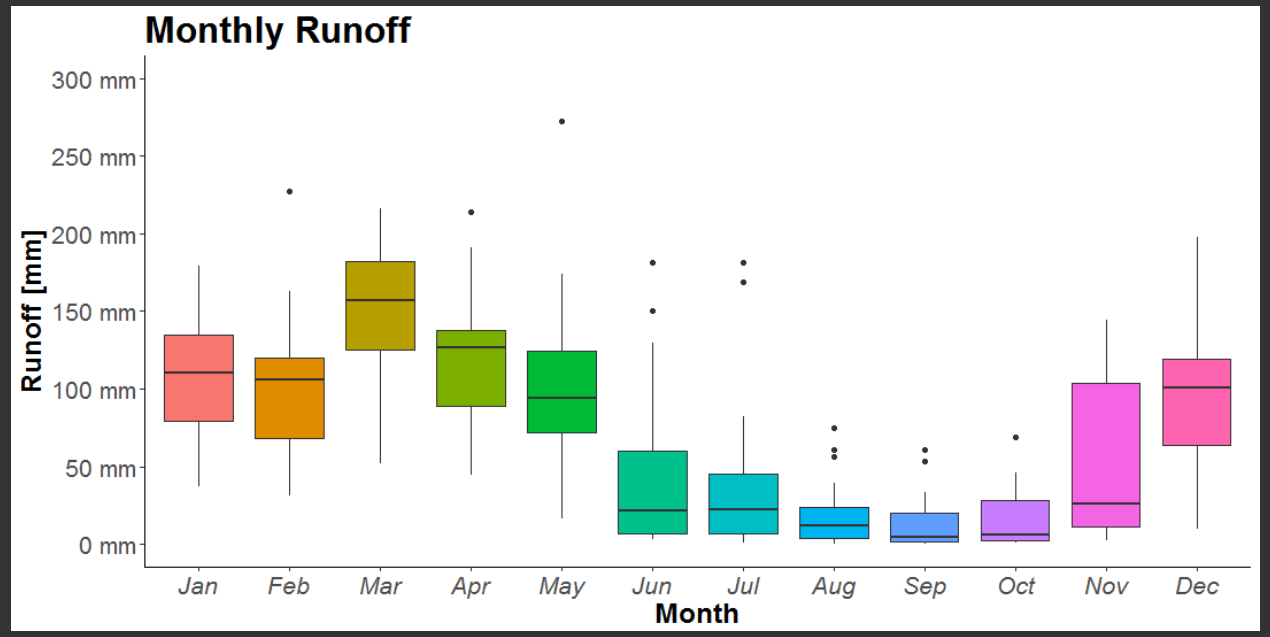
- Range = Maximum – Minimum
- Population Variance: = $\sigma^2 = \Sigma (X_i - \mu)^2 / N$
- Sample Variance = $s^2 = \Sigma (x_i - x_bar)^2 / (n - 1)$
- Population Standard Deviation = $\sigma = sqrt[\Sigma (X_i - \mu)^2 / N]$
- Sample Standard Deviation = $s = sqrt [\Sigma (x_i - x_bar)^2 / (n - 1)]$

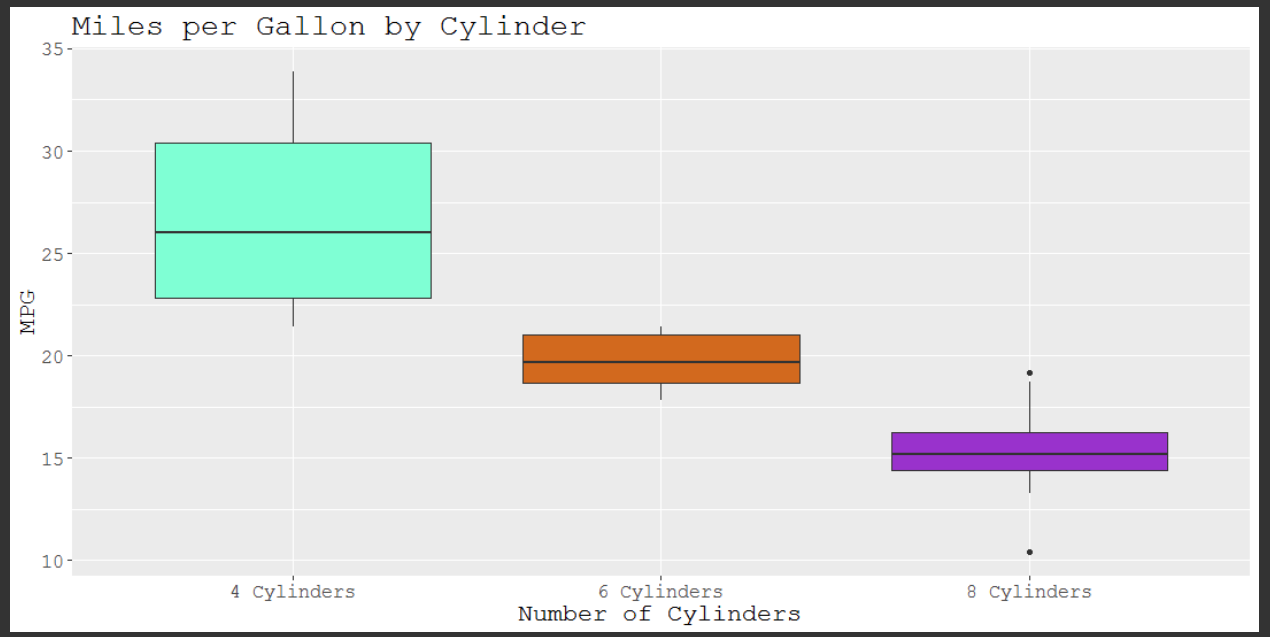
Box Plot



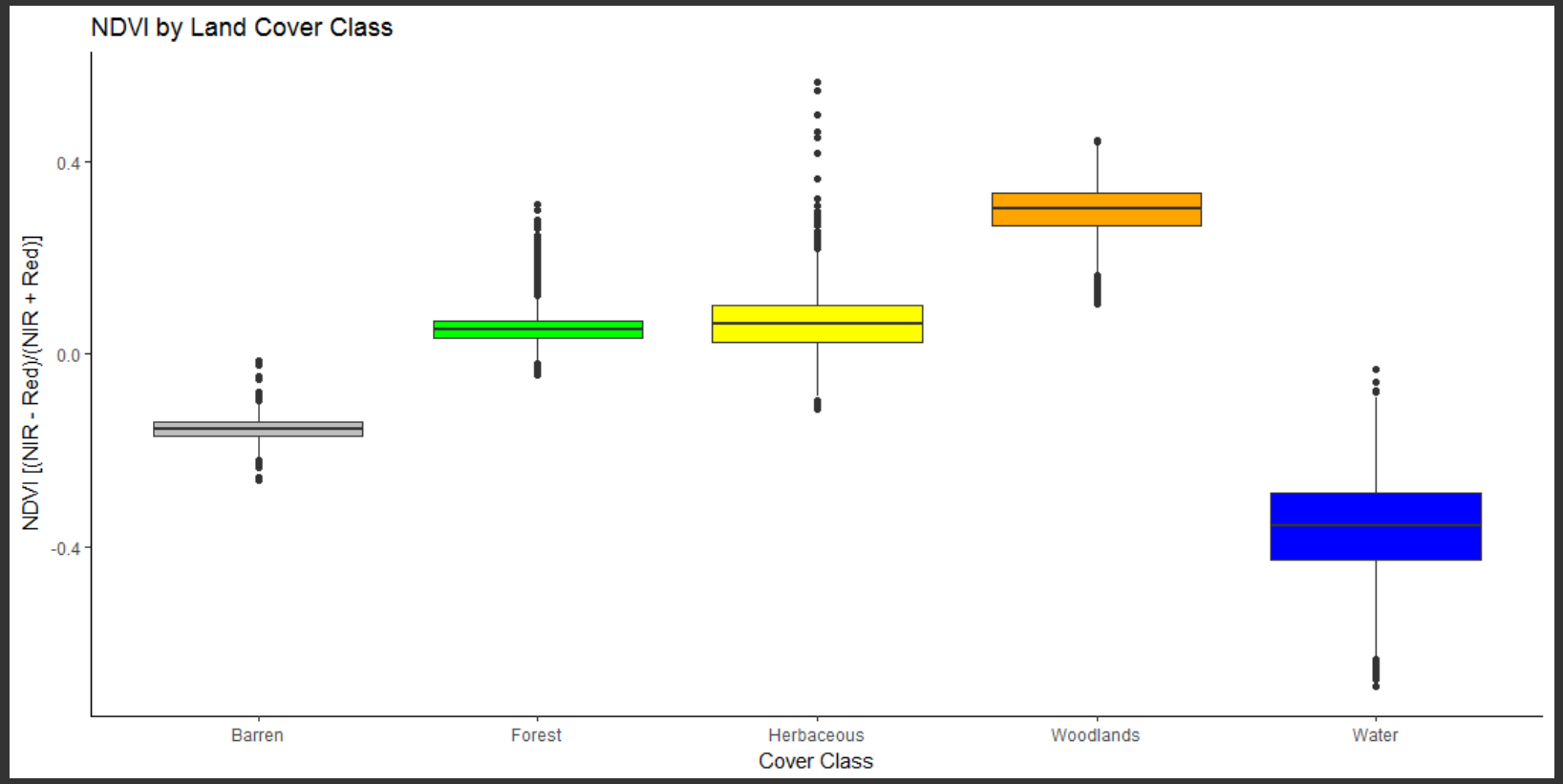
- Minimum = lowest value
- Maximum = highest value
- 1st Quartile = 25% of data below, 75% above
- 3rd Quartile = 75% of data below, 25% above
- Interquartile Range (IQR) = range of values between 1st and 3rd quartile

Box Plot

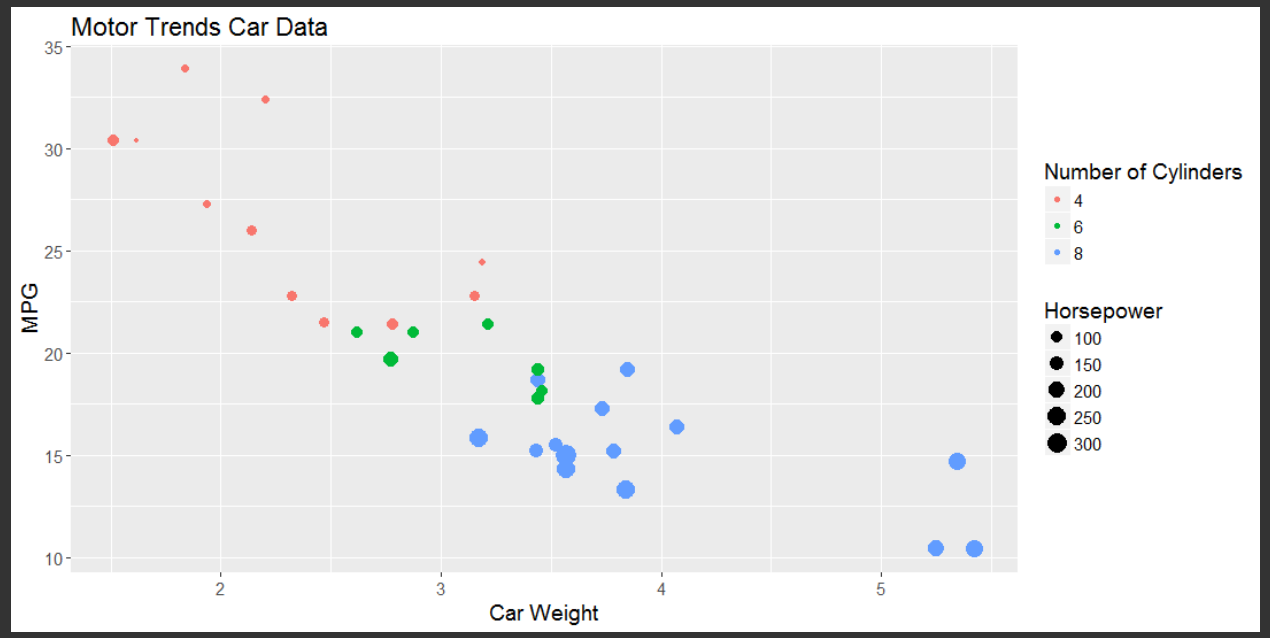

Monthly Runoff
300 mm
250 mm
200 mm
150 mm
100 mm
50 mm
0 mm
Runoff [mm]
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

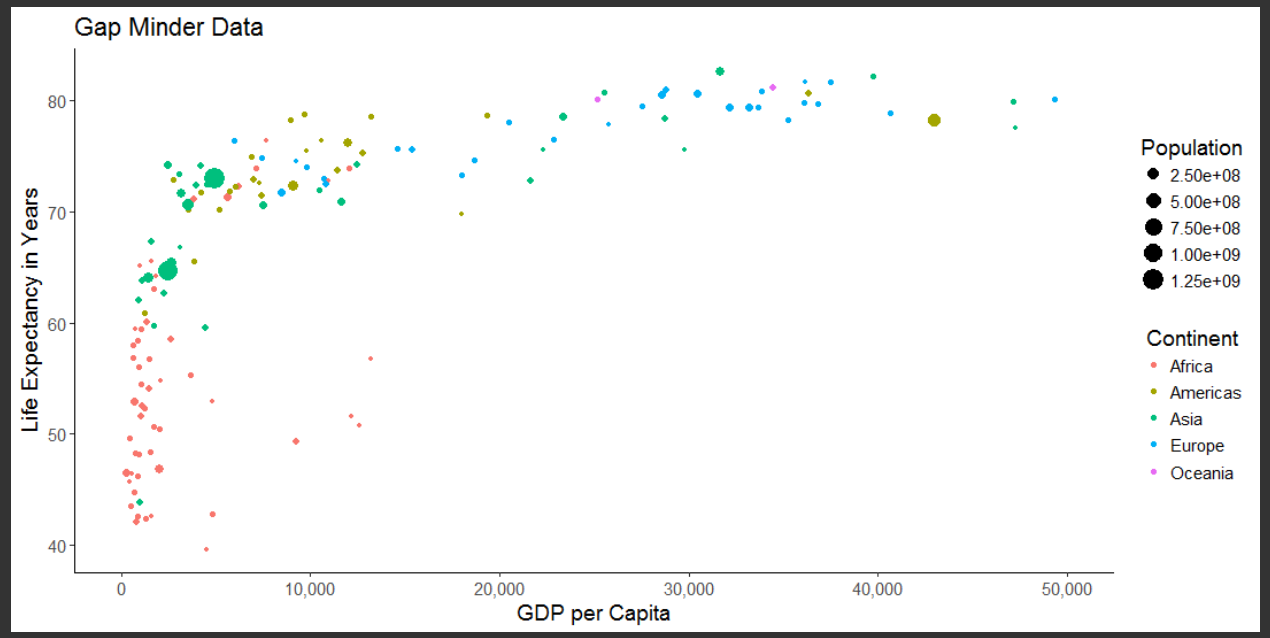

Miles per Gallon by Cylinder
35
30
25
20
15
10
MPG
4 Cylinders
6 Cylinders
8 Cylinders
Number of Cylinders

Box Plot

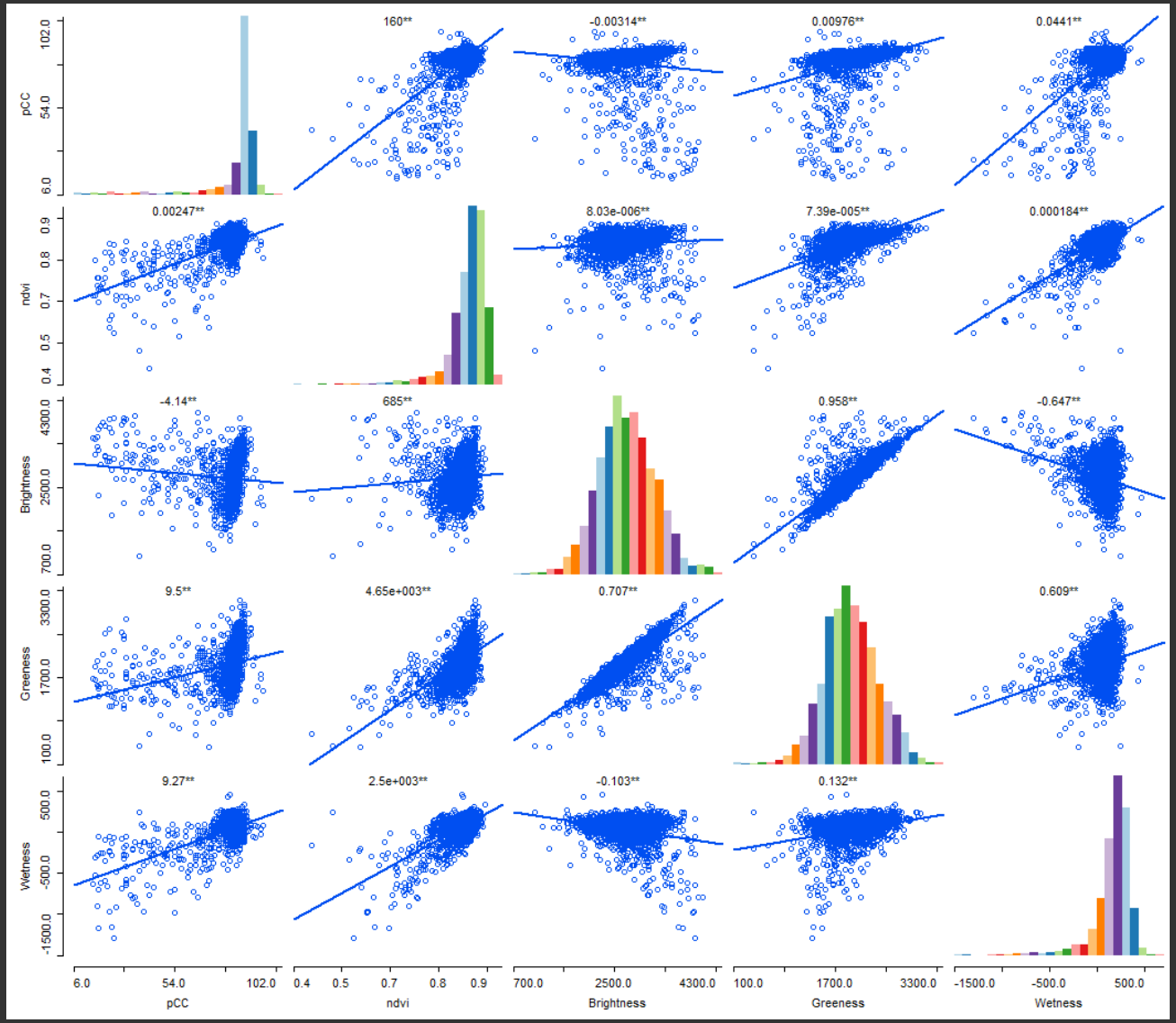

NDVI by Land Cover Class
NDVI [(NIR - Red)/(NIR + Red)]
0.4
0.0
-0.4
Barren
Forest
Herbaceous
Woodlands
Water
Cover Class

Relationships Between Variables: Scatterplot



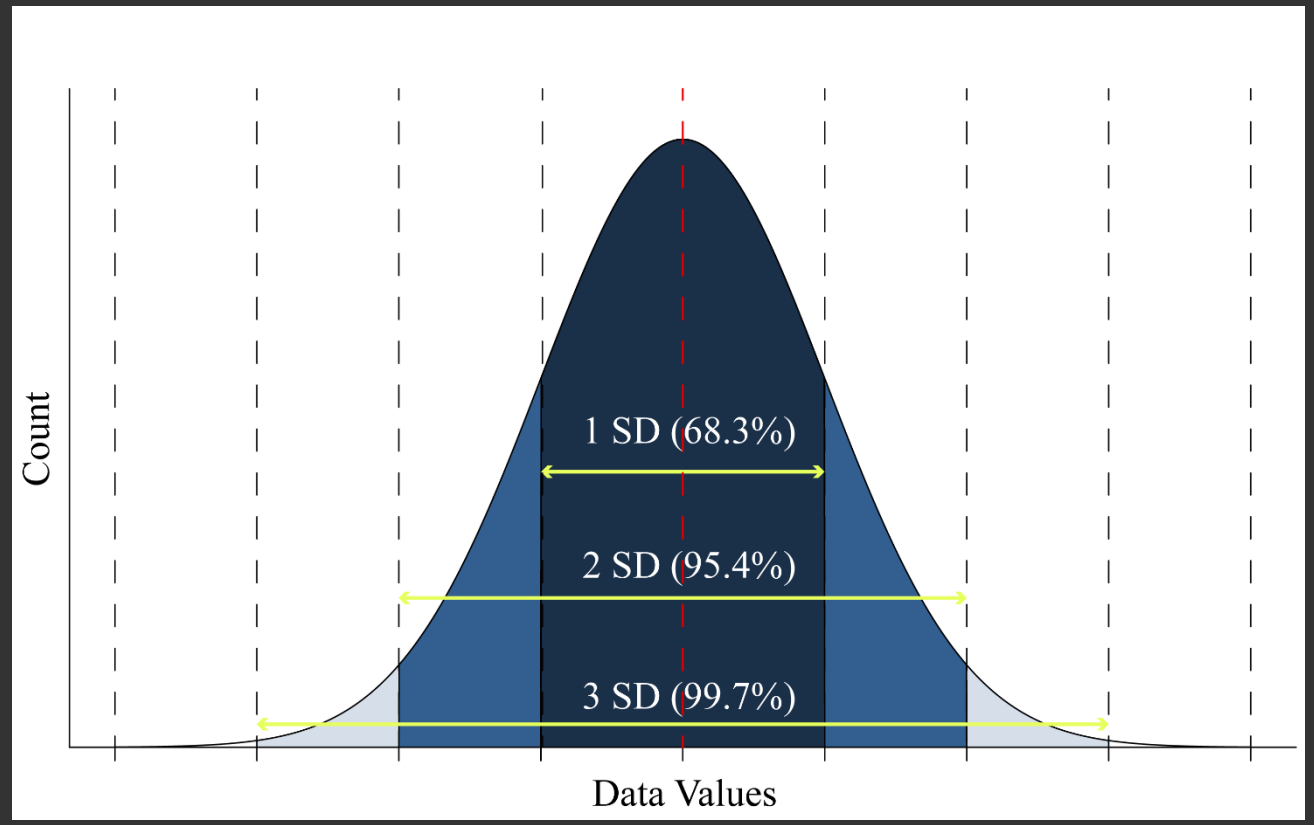


Relationships Between Variables: Scatterplot Matrix



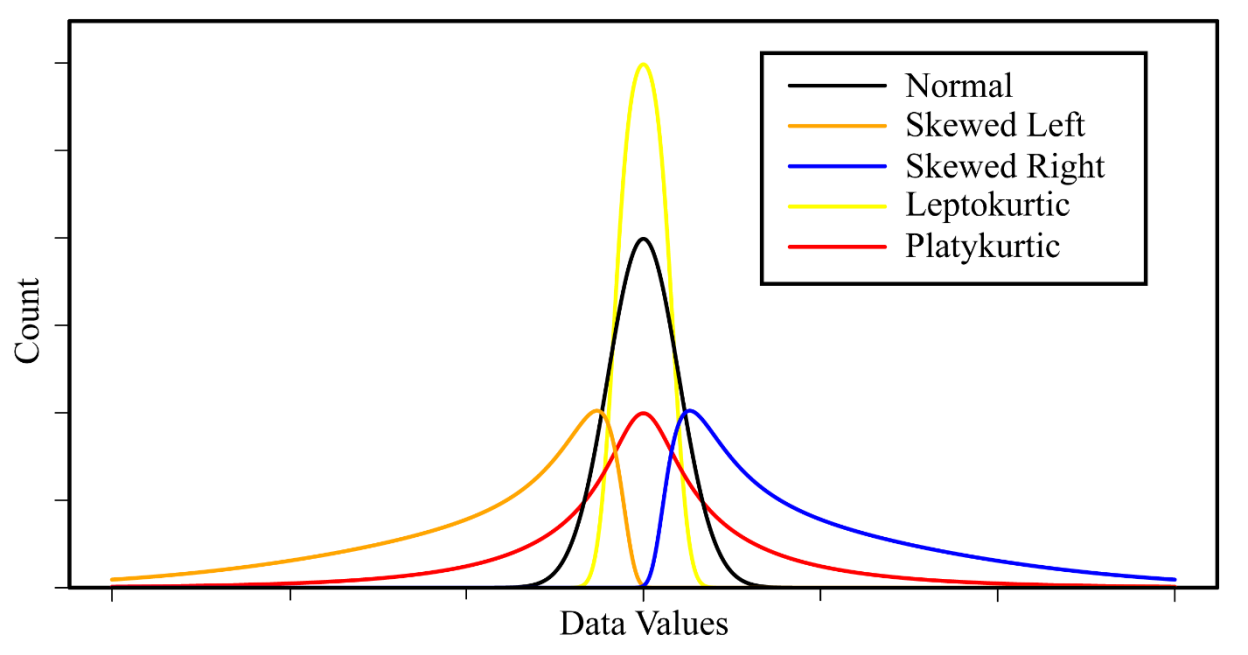
Normal Distribution

- Bell-shaped
- Mean, median, and mode are equivalent
- Curve is symmetrical about the mean
- Shape of the normal curve varies based on the mean and standard deviation of the data
- 68.3% of the data are within one standard deviation of the mean
- 95.4% of the data lie within two standard deviations of the mean
- 99.7% of the data lie within three standard deviations of the mean

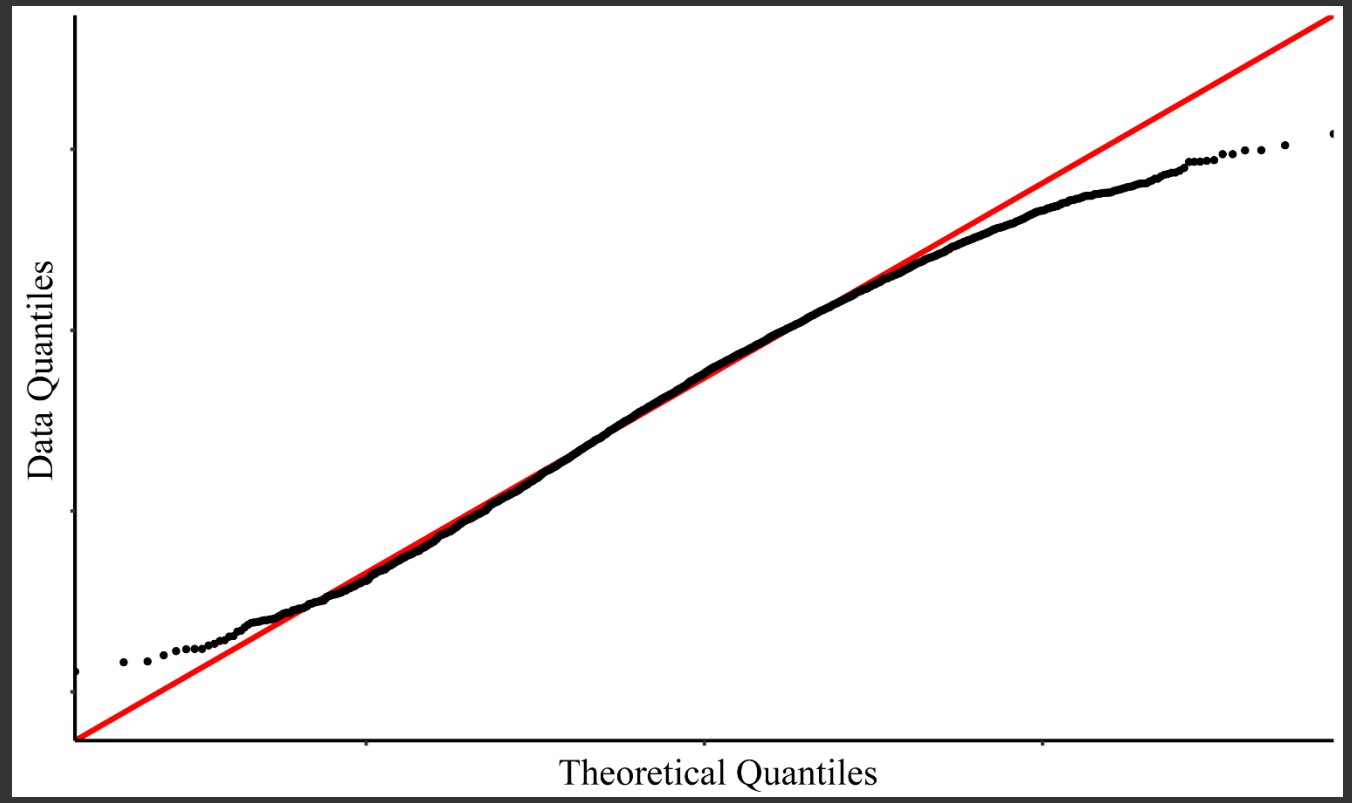


Skewness and Kurtosis

- Skewness = measure of symmetry
  - - = skew to left
  - + = skew to right
  - 0 = near normal
- Kurtosis = heavy-tailed or light-tailed compared to a normal distribution
  - - = light-tailed (leptokurtic)
  - + = heavy-tailed (platykurtic)
  - 0 = near normal (mesokurtic)



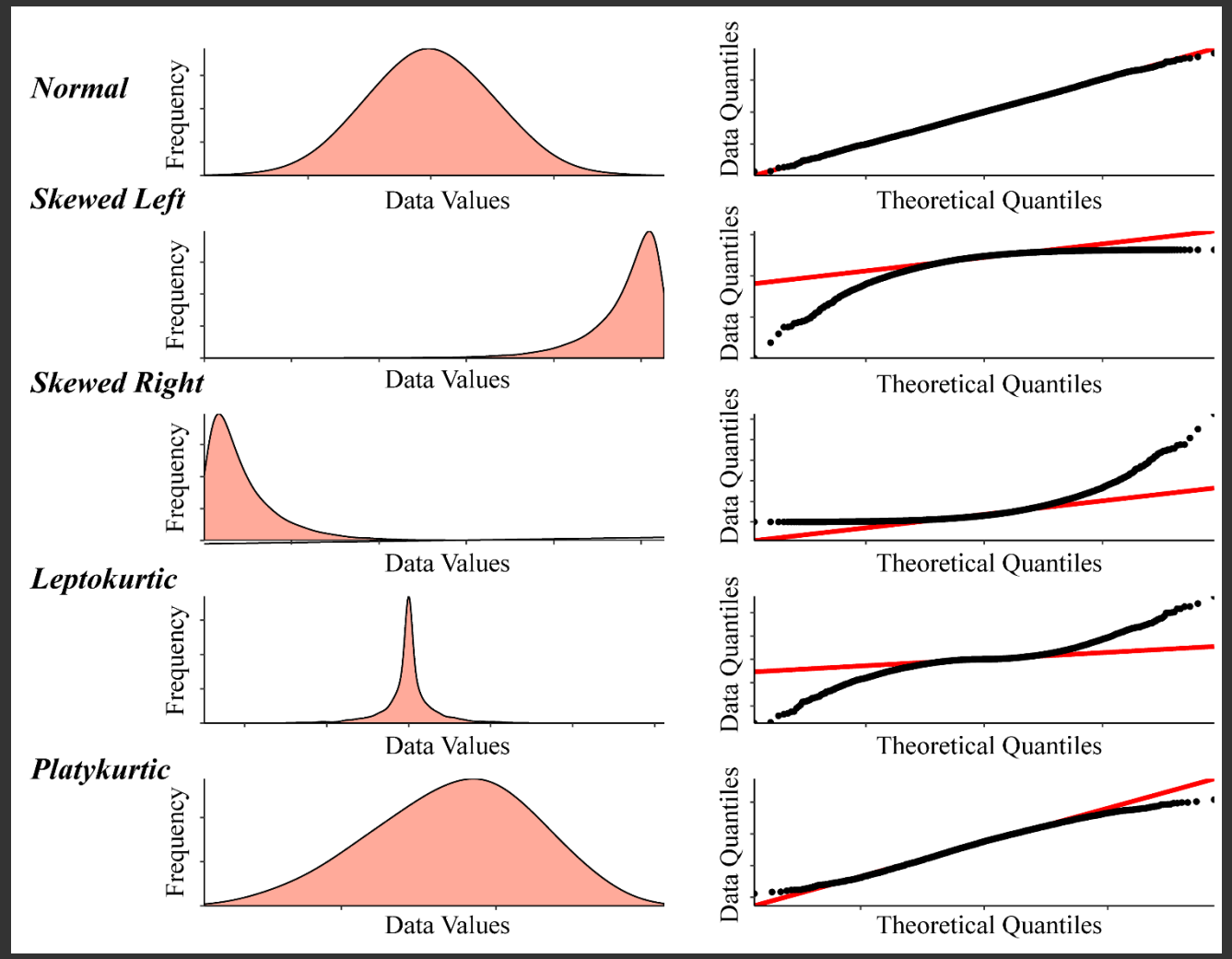
Are your data normally distributed?



Q-Q Plot

- On line = normal
- Concave = skewed to right
- Convex = skewed to the left
- High/low off line = kurtosis

Are your data normally distributed?

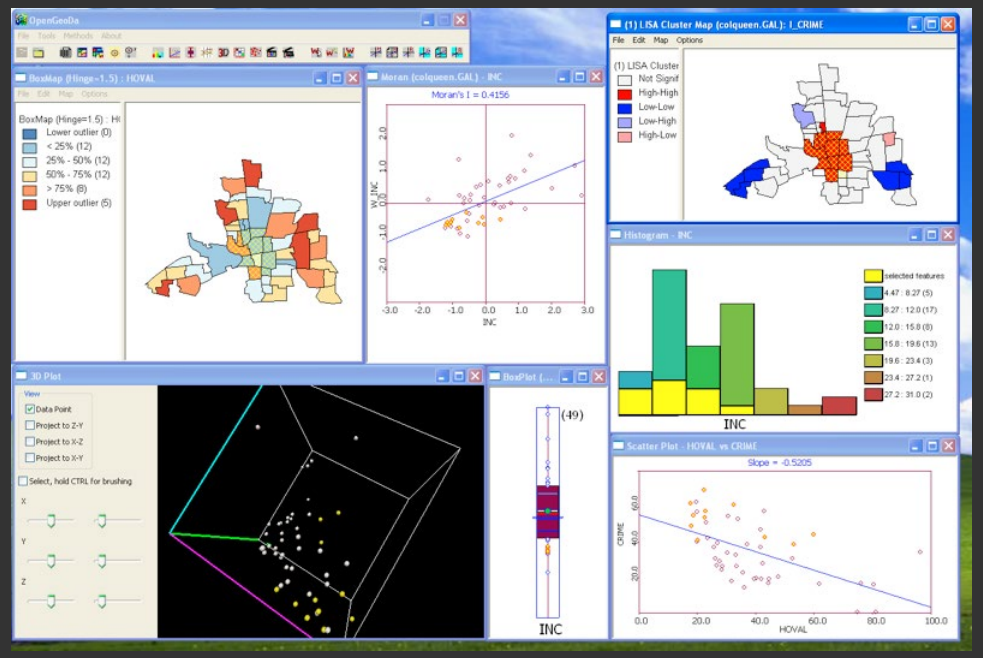


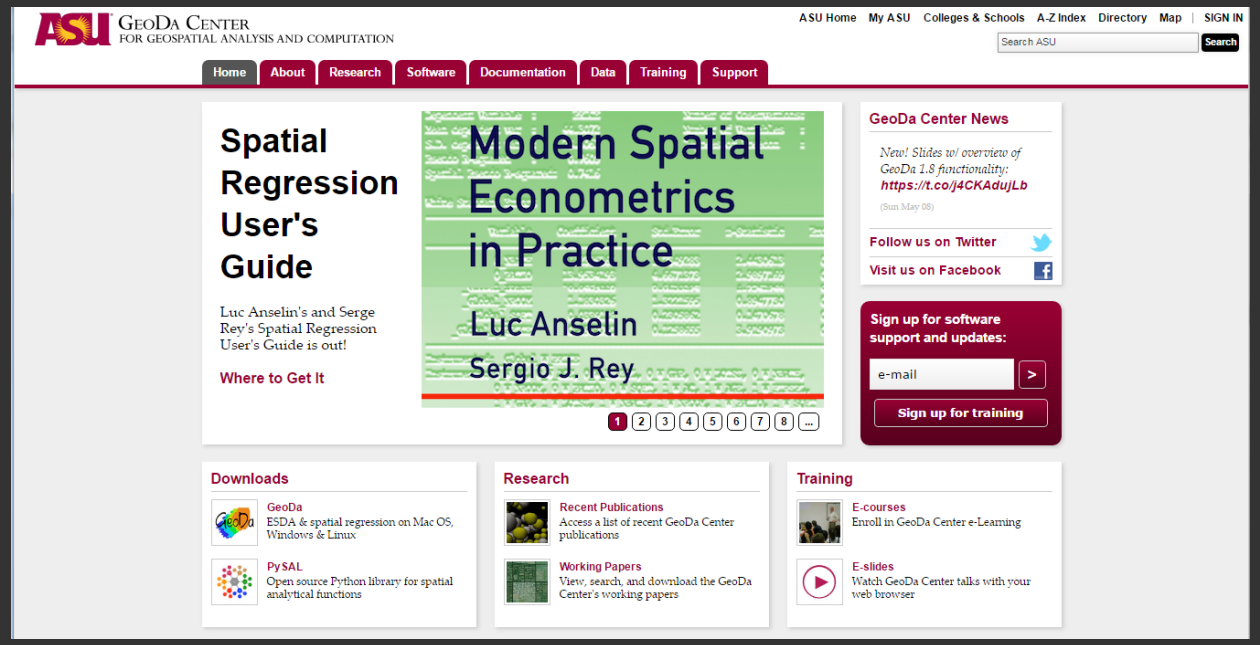
- On line = normal
- Concave = skewed to right
- Convex = skewed to the left
- High/low off line = kurtosis

Q-Q Plot

GeoDa

- A valuable to for geographic statistical and exploratory data analysis
- Also, its free!



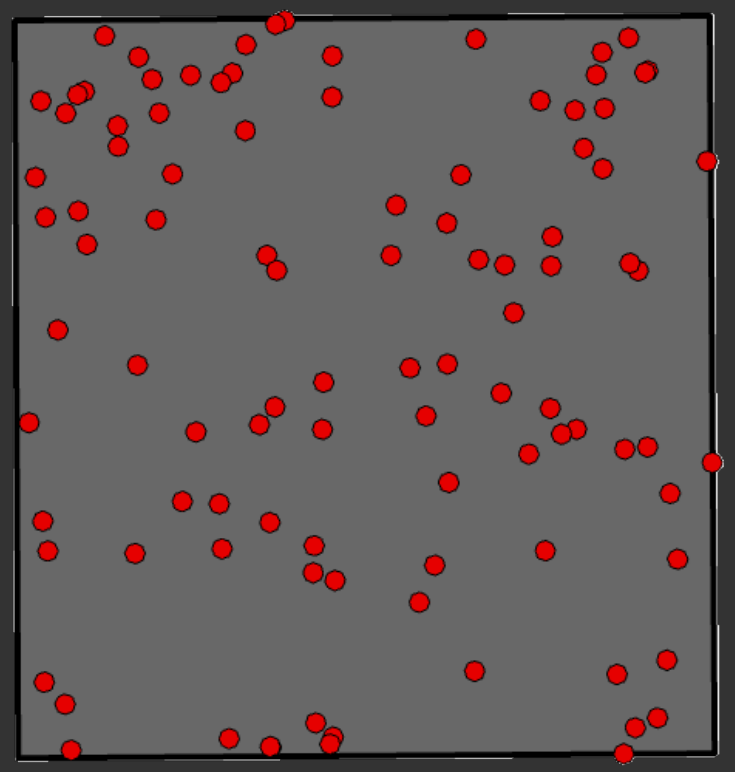


https://geodacenter.asu.edu/

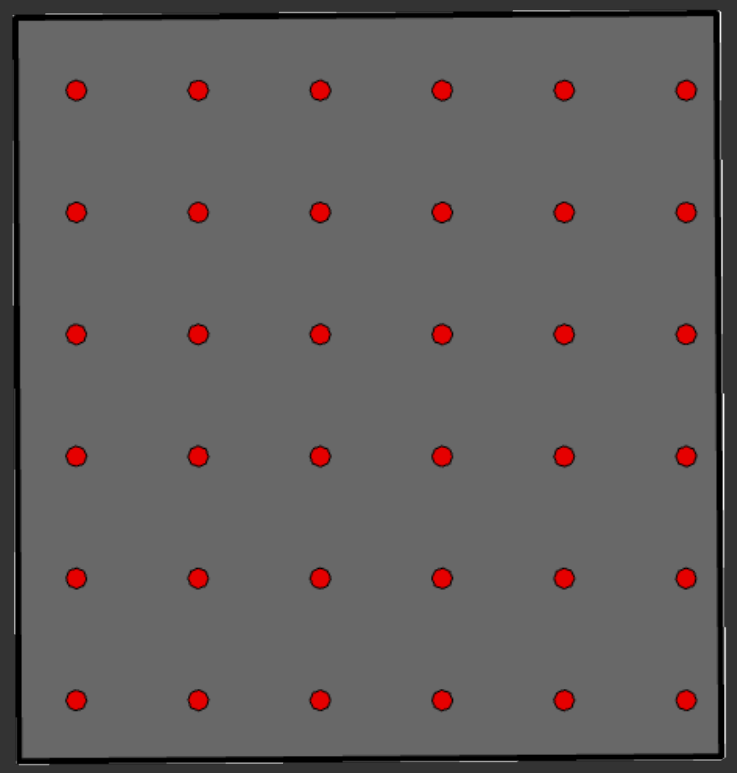
Sampling

- A sample is a set of measurements taken from a larger group or population.
- Sample means and variances can serve as estimates for their population.
- Easier to measure with samples, then draw conclusions about entire population.

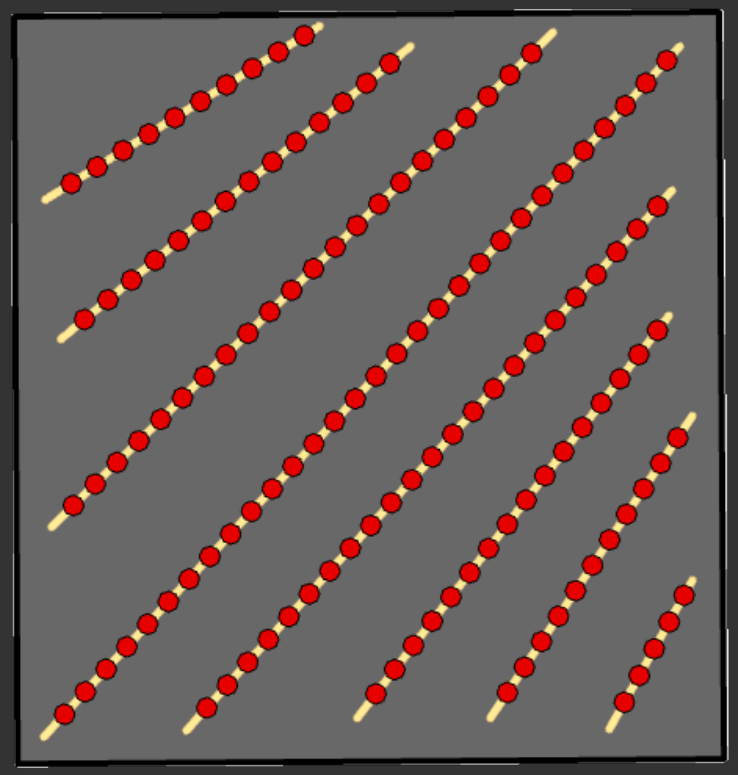
Types of Sampling



Simple Random Sampling

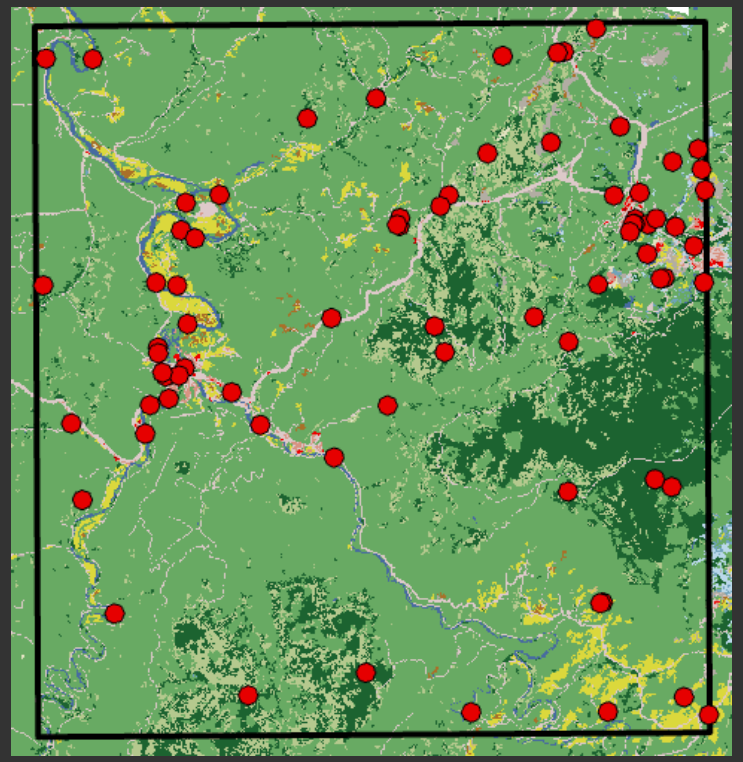


Systematic Sampling

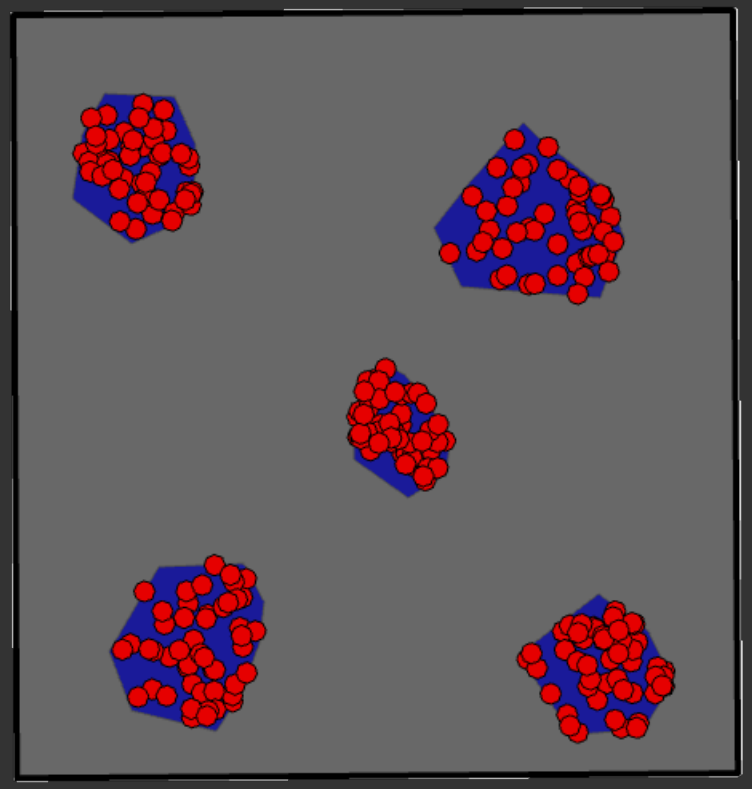


Systematic Sampling Along Transects

Types of Sampling



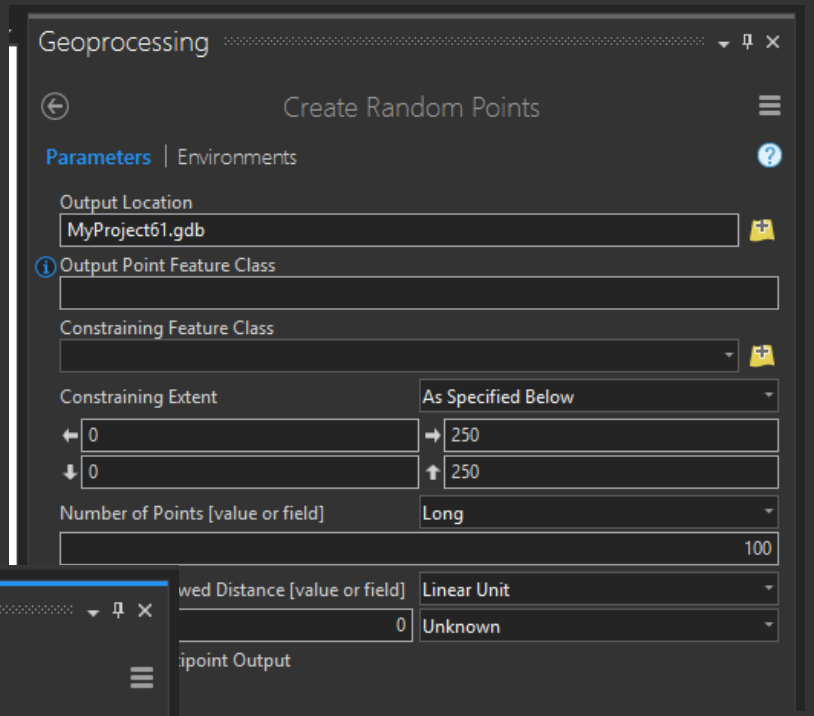
Stratified Random Sampling
(By land cover type)

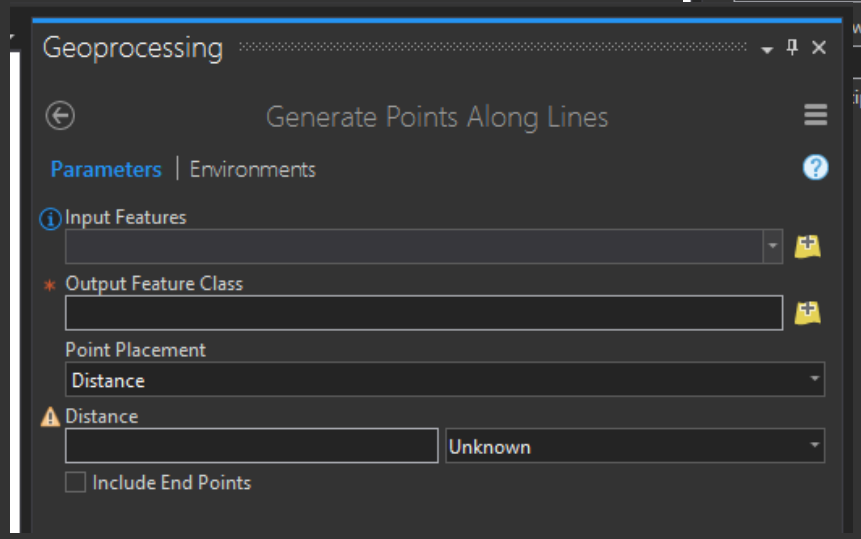


Random Sampling
within Clusters

Tools for Sampling

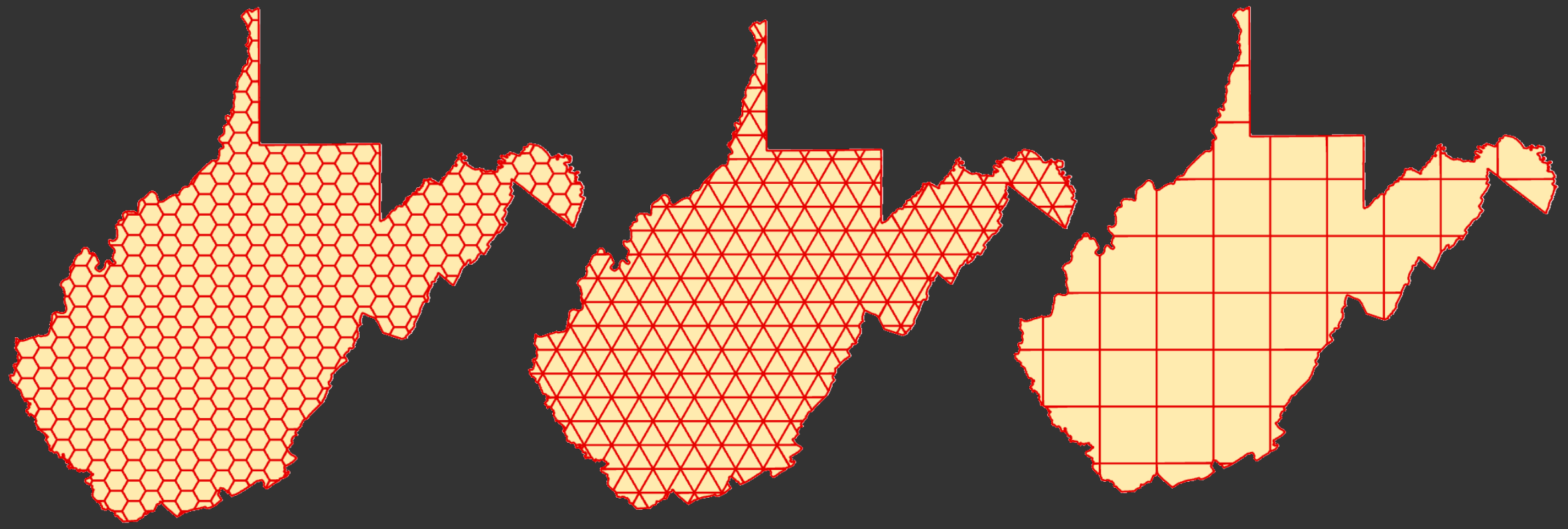
- QGIS
- R
- ArcGIS Pro
  - Generate Random Points
  - Generate Points Along Line
  - Generate Tessellation
  - Create Fishnet





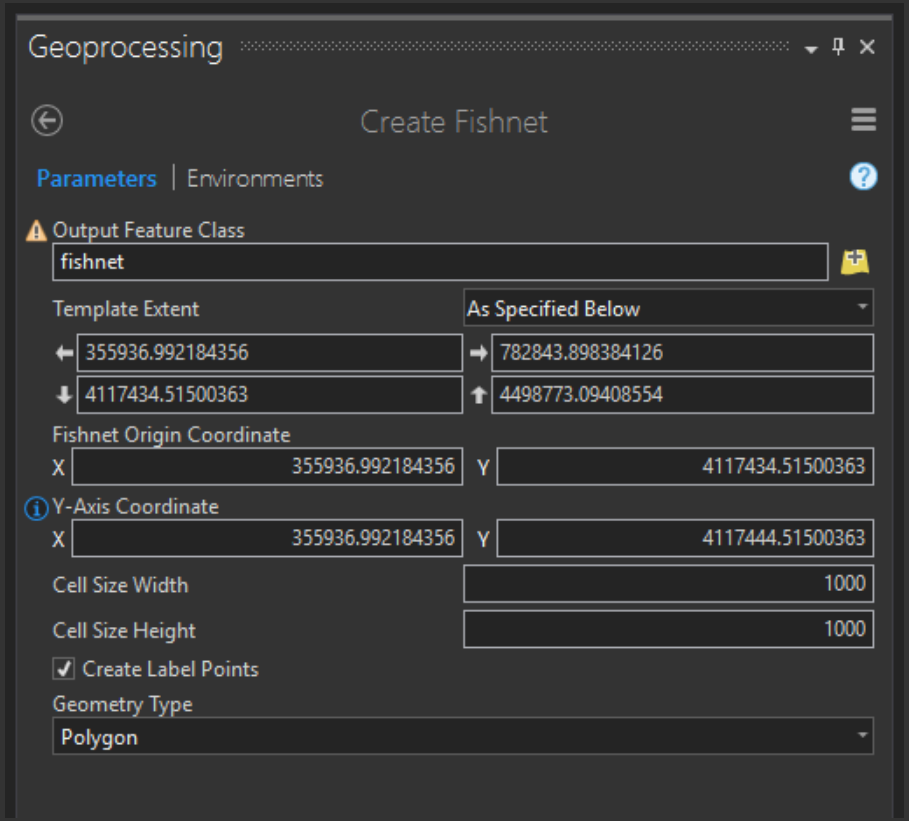
Tessellation

- Filling a plane completely with non-overlapping geometric shapes
- Can use generated features as sampling or aggregation units



Fishnets

- Creates a fishnet of regular cells

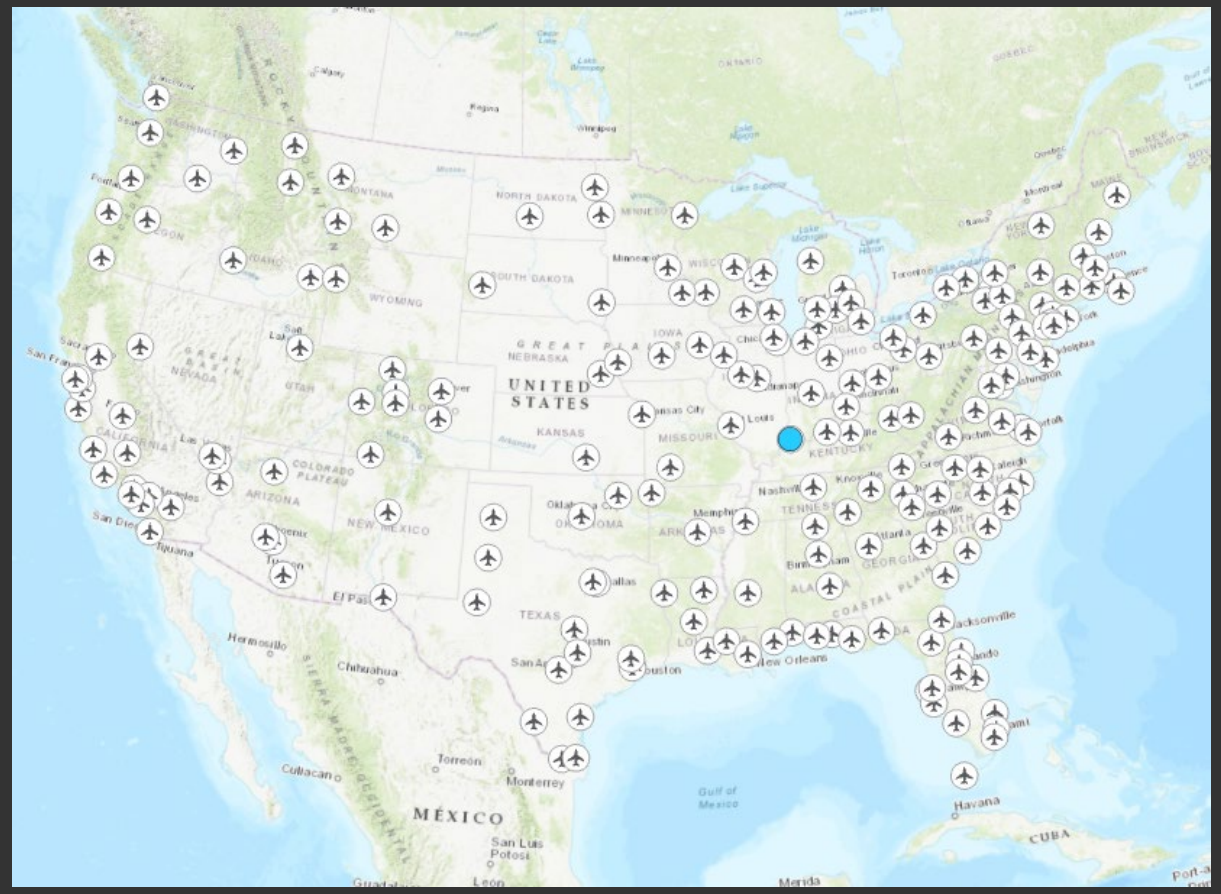


Imagery from ESRI Basemap

Spatial Central Tendency and Dispersion

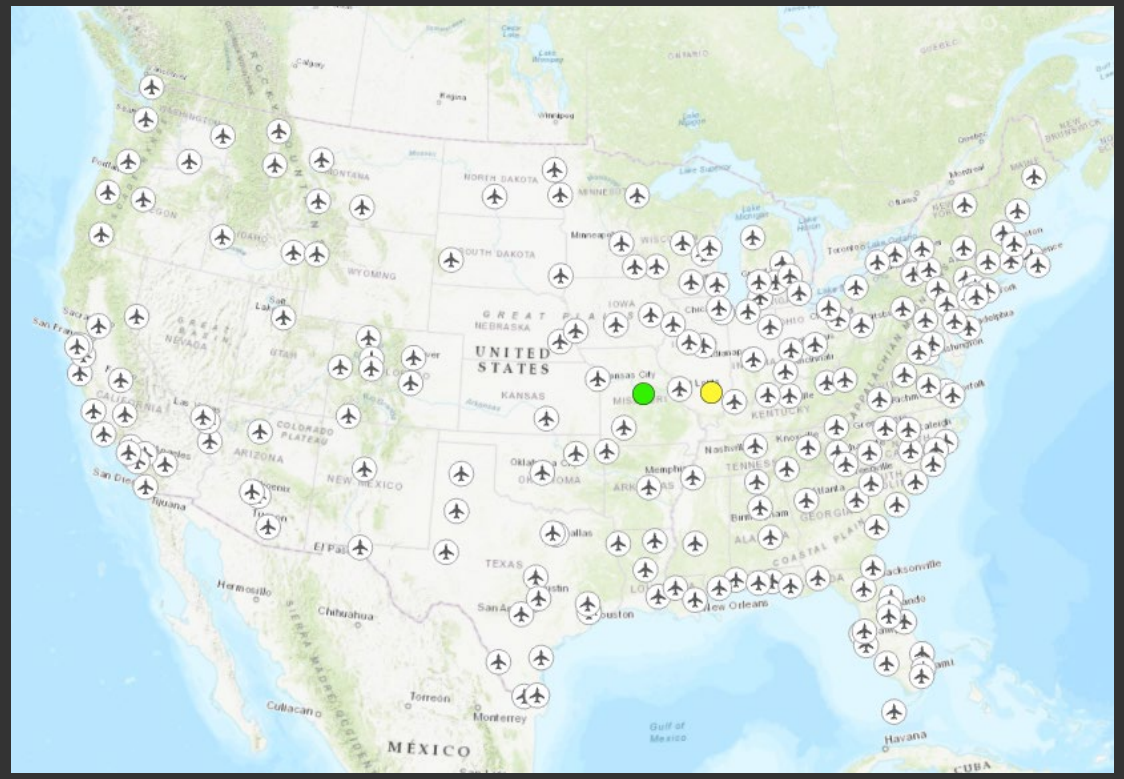
Central Feature

- Identifies the most central feature
- The central feature is defined as the feature in the dataset with the smallest accumulated distance from all other features
- Euclidean Distance = straight-line distance
- Manhattan Distance = sum of distance in x and y directions
- Can also apply a weight
- Can stratify by category



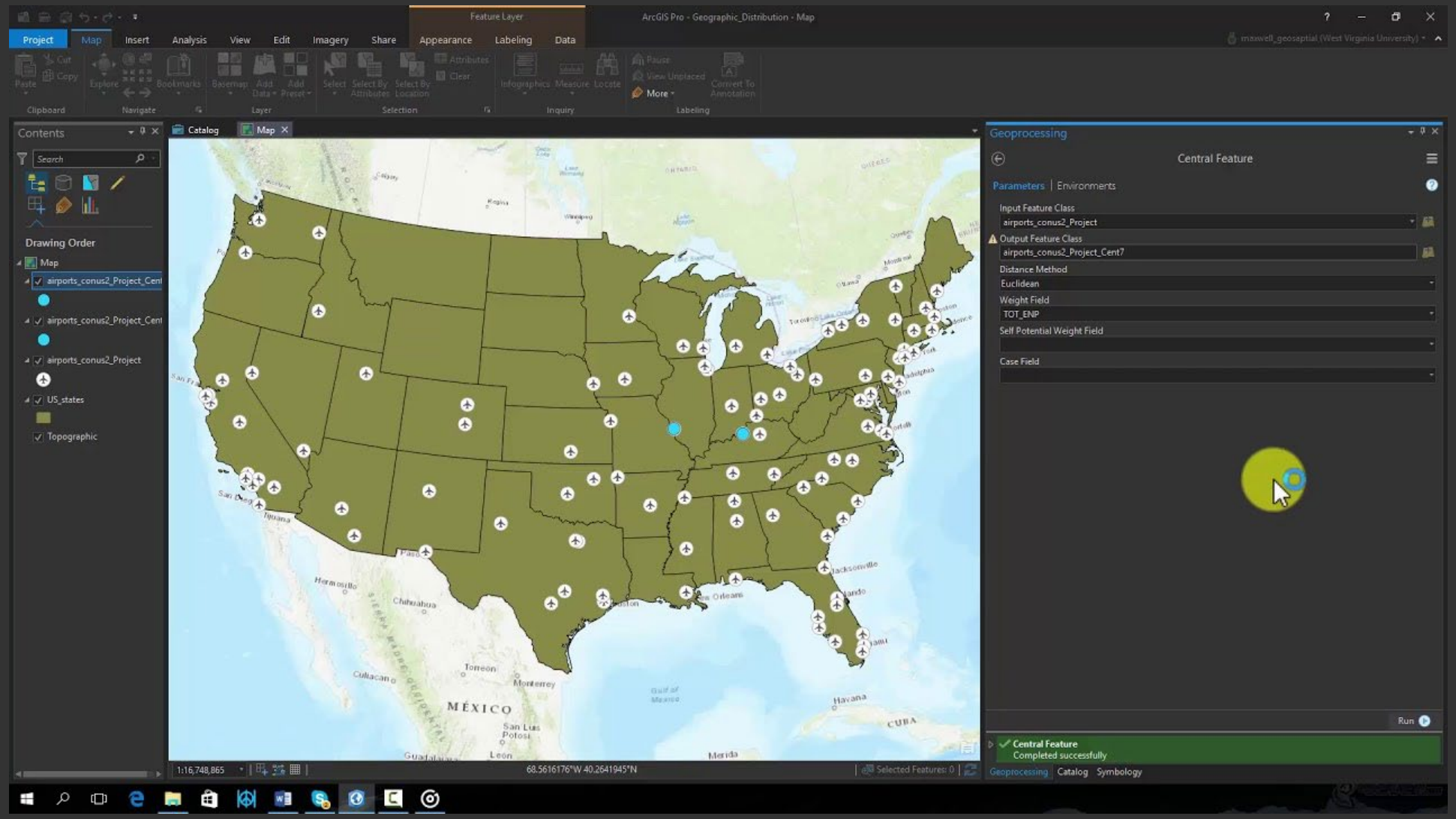
Mean and Median Center

- Mean Center = the average x and y coordinate from the features
- Median Center = the median x and y coordinate from the features
- A new location, one of the input features will not be selected
- Can apply a weight
- Can stratify by category



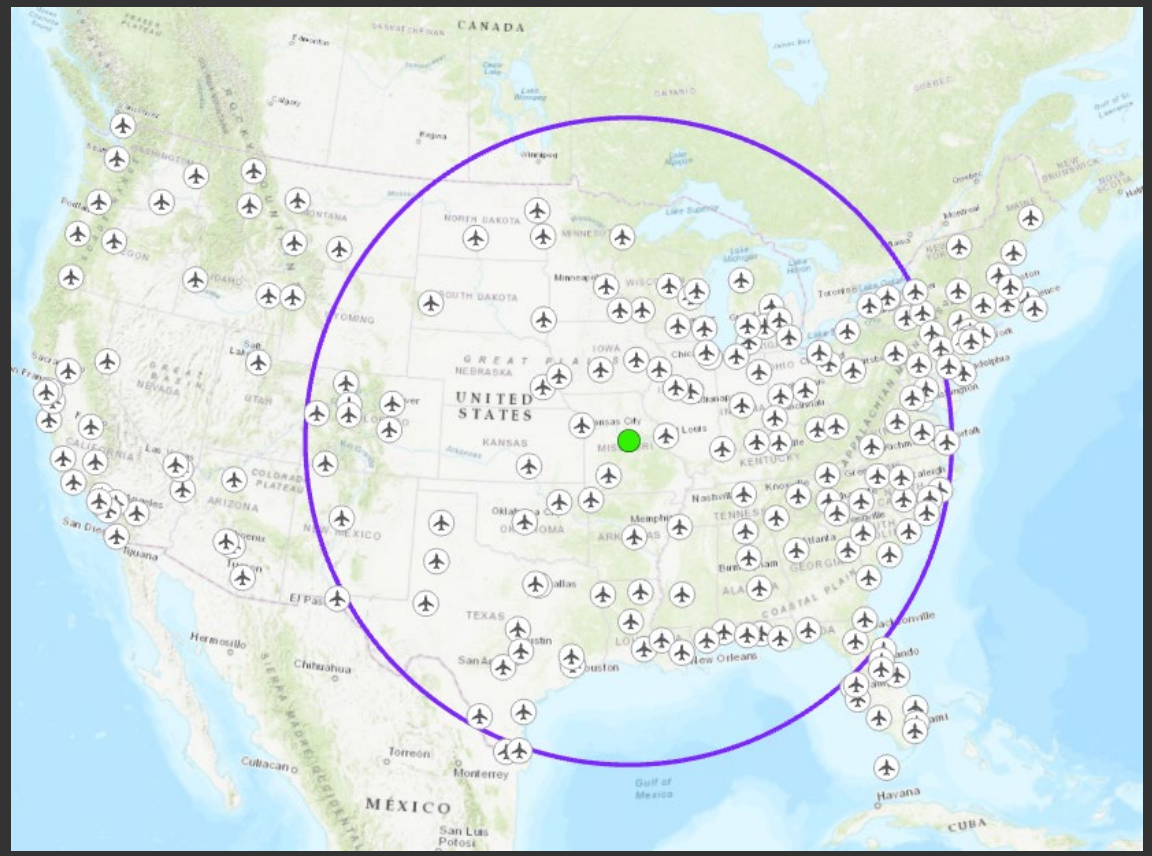
Green = Mean Center
Yellow = Median Center

Video: Spatial Central Tendency



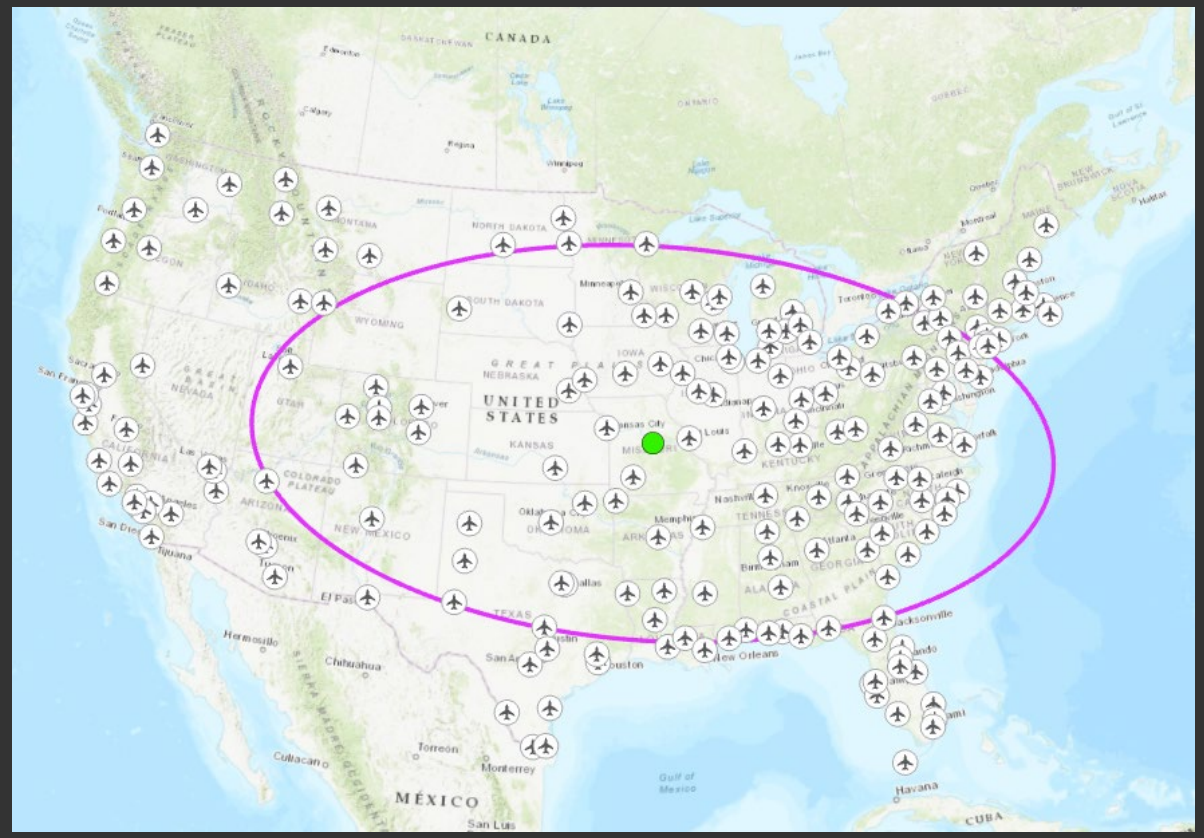
Standard Distance

- Measures the degree to which features are concentrated or dispersed around the geometric mean center
- Larger = more spatial dispersion
- Smaller = less spatial dispersion
- Center = mean center
- Can specify the number of standard deviations
- Can apply a weight
- Can stratify by category



OBJECTID	Shape	Shape_Length	Shape_Area	CenterX	CenterY	StdDist
1	Polygon	9079239.746236	6559610913522.605469	319484.463441	-36783.762592	1445024.225863

Directional Distribution (Standard Deviation Ellipse)



- Creates and ellipse that summarizes spatial central tendency, spatial dispersion, and directional trends
- Can specify number of standard deviations
- Can apply a weight
- Can stratify by category

OBJECTID	Shape	Shape_Length	Shape_Area	CenterX	CenterY	XStdDist	YStdDist	Rotation
1	Polygon	8844397.929445	5190266364815.936523	319484.463441	-36783.762592	1834356.420375	900736.67056	93.898536

Spatial Statistics

How to Interpret a Statistical Test

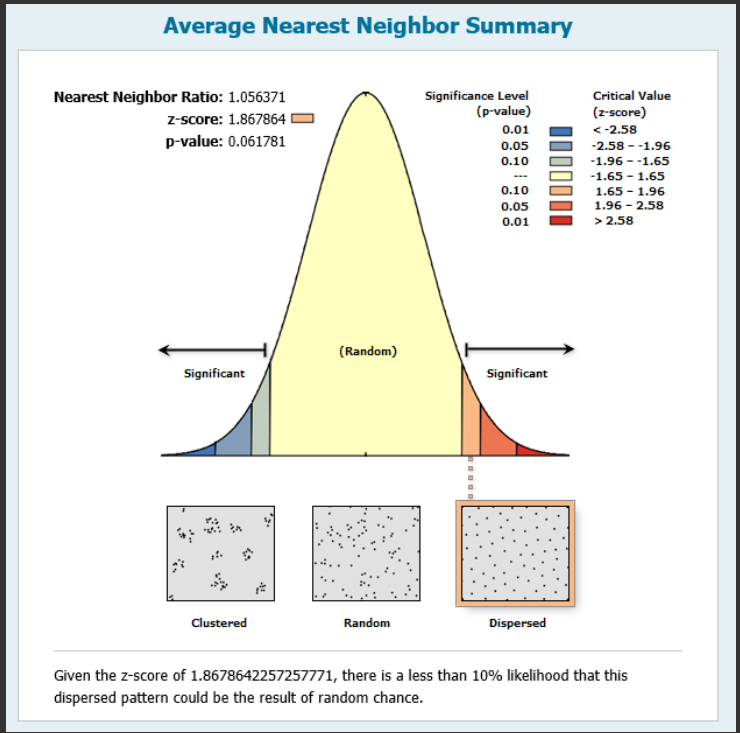
- Null Hypothesis (H_0) = there is no statistical significance or difference
- Alternative Hypothesis (H_a) = there is statistical significance or difference
- If the statistical test does not show statistical significance, we fail to reject the null hypothesis
- A specific statistical test will commonly produce a statistic or index
- This is then compared to a distribution, such as a z-distribution or normal distribution
- From this, you can get a *p*-value
- A p-value of 0.05 suggest that there is less than a 5% change of obtaining the result by random chance (this is equivalent to a 95% confidence interval)

Analyzing Point Patterns

Average Nearest Neighbor

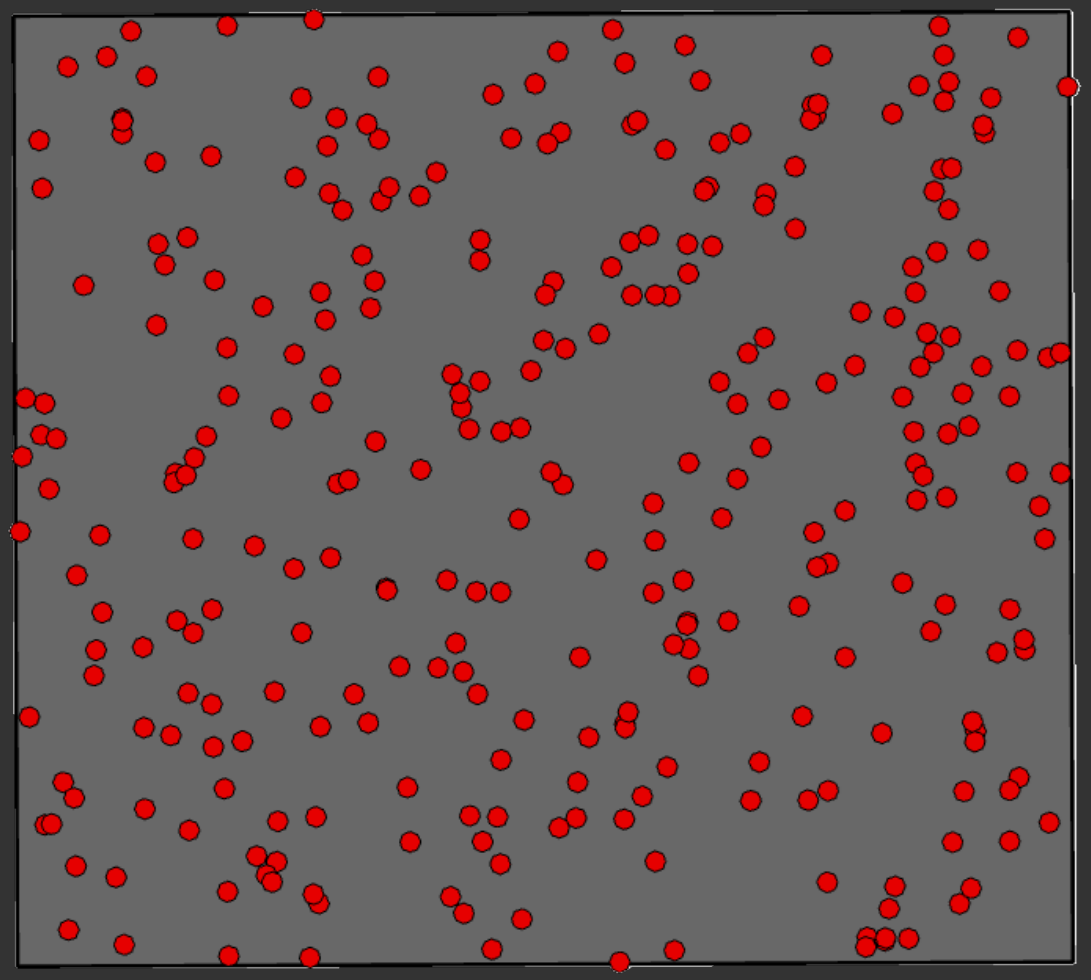
- Test to determine if a point pattern/distribution is random, clustered, or dispersed
- Null Hypothesis = point pattern is random
- Alternative Hypothesis = point pattern is not random (clustered or dispersed)
- Only looks at the spatial distribution of points, not the attributes stored on the points
- Can calculate distance using the Euclidean or Manhattan method

Average Nearest Neighbor

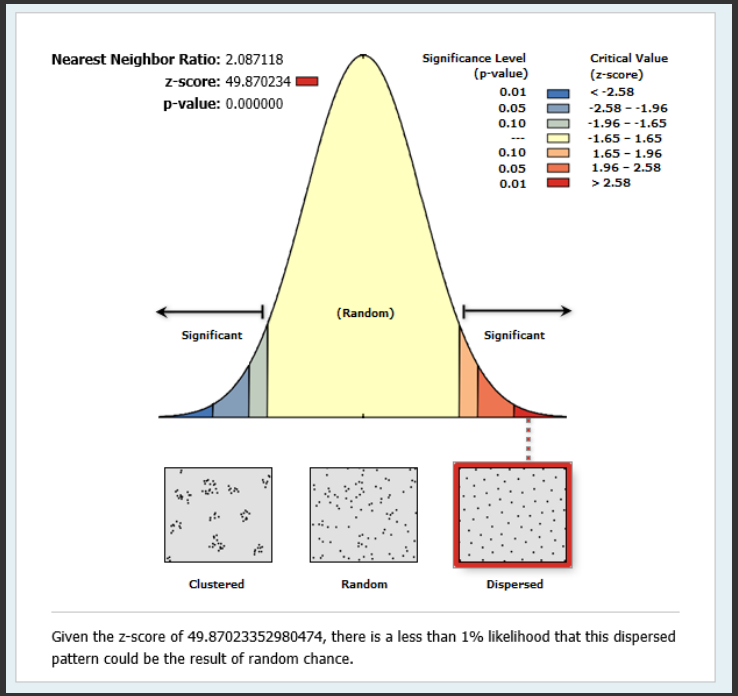


Average Nearest Neighbor Summary

Observed Mean Distance:	7332.7008 Meters
Expected Mean Distance:	6941.4084 Meters
Nearest Neighbor Ratio:	1.056371
z-score:	1.867864
p-value:	0.061781

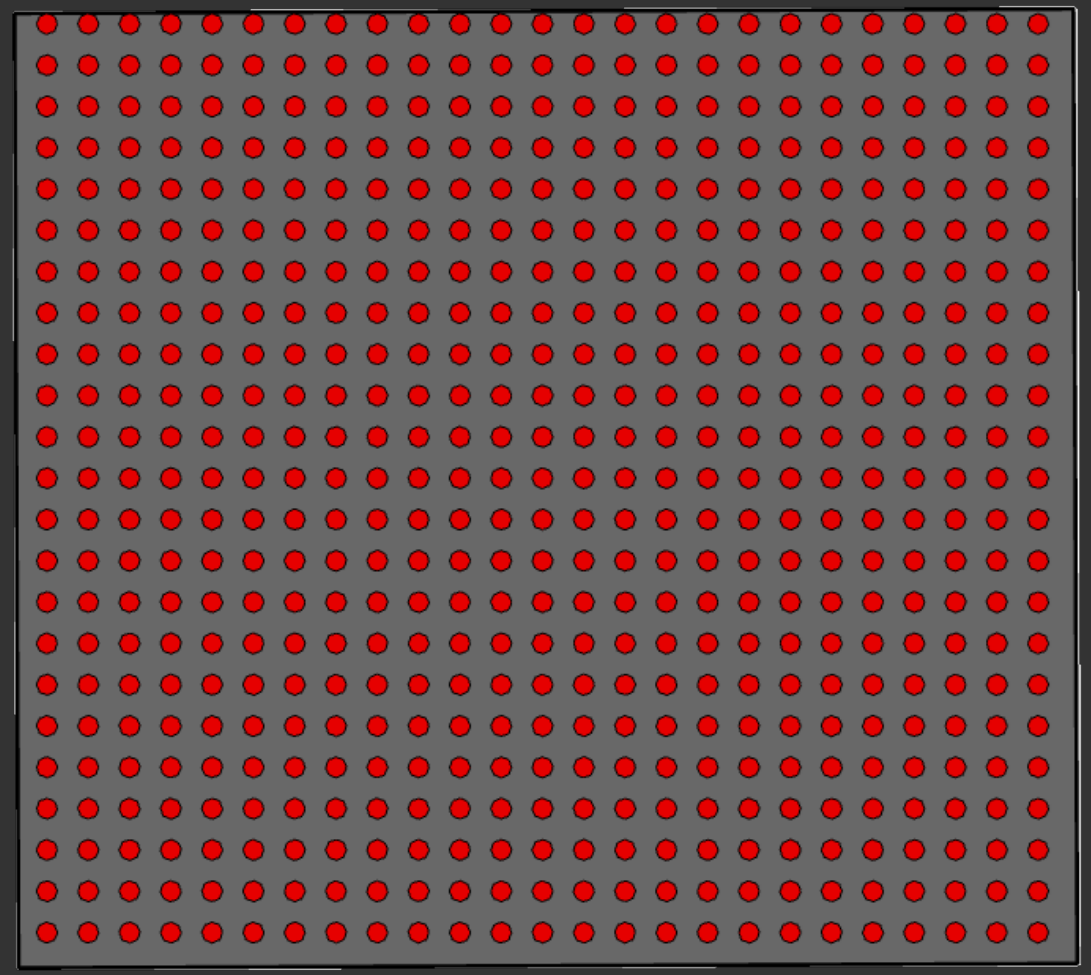


Average Nearest Neighbor

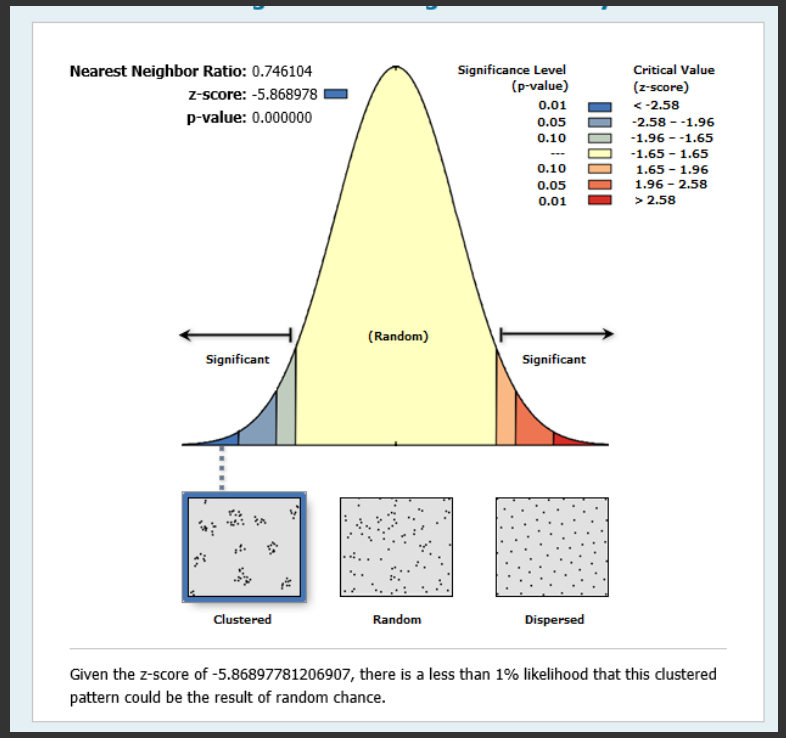


Average Nearest Neighbor Summary

Observed Mean Distance:	10000.0000 Meters
Expected Mean Distance:	4791.2965 Meters
Nearest Neighbor Ratio:	2.087118
z-score:	49.870234
p-value:	0.000000

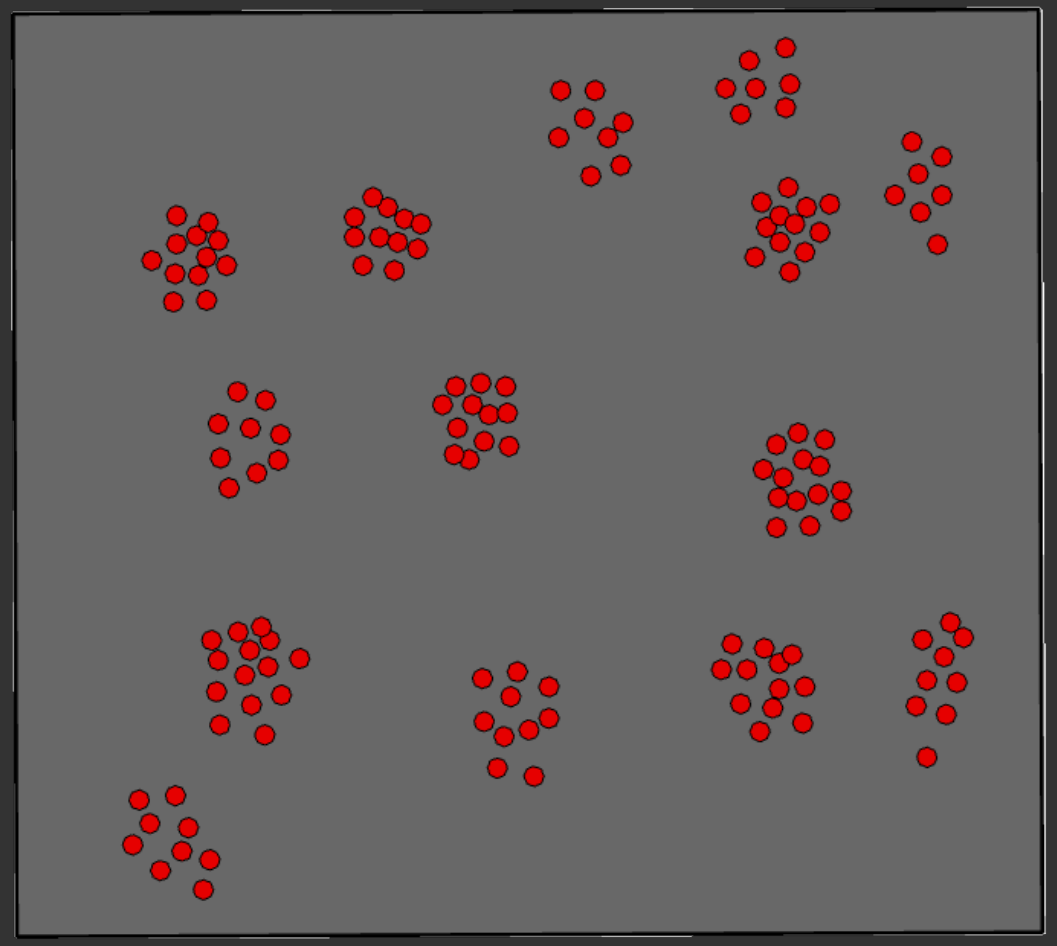


Average Nearest Neighbor

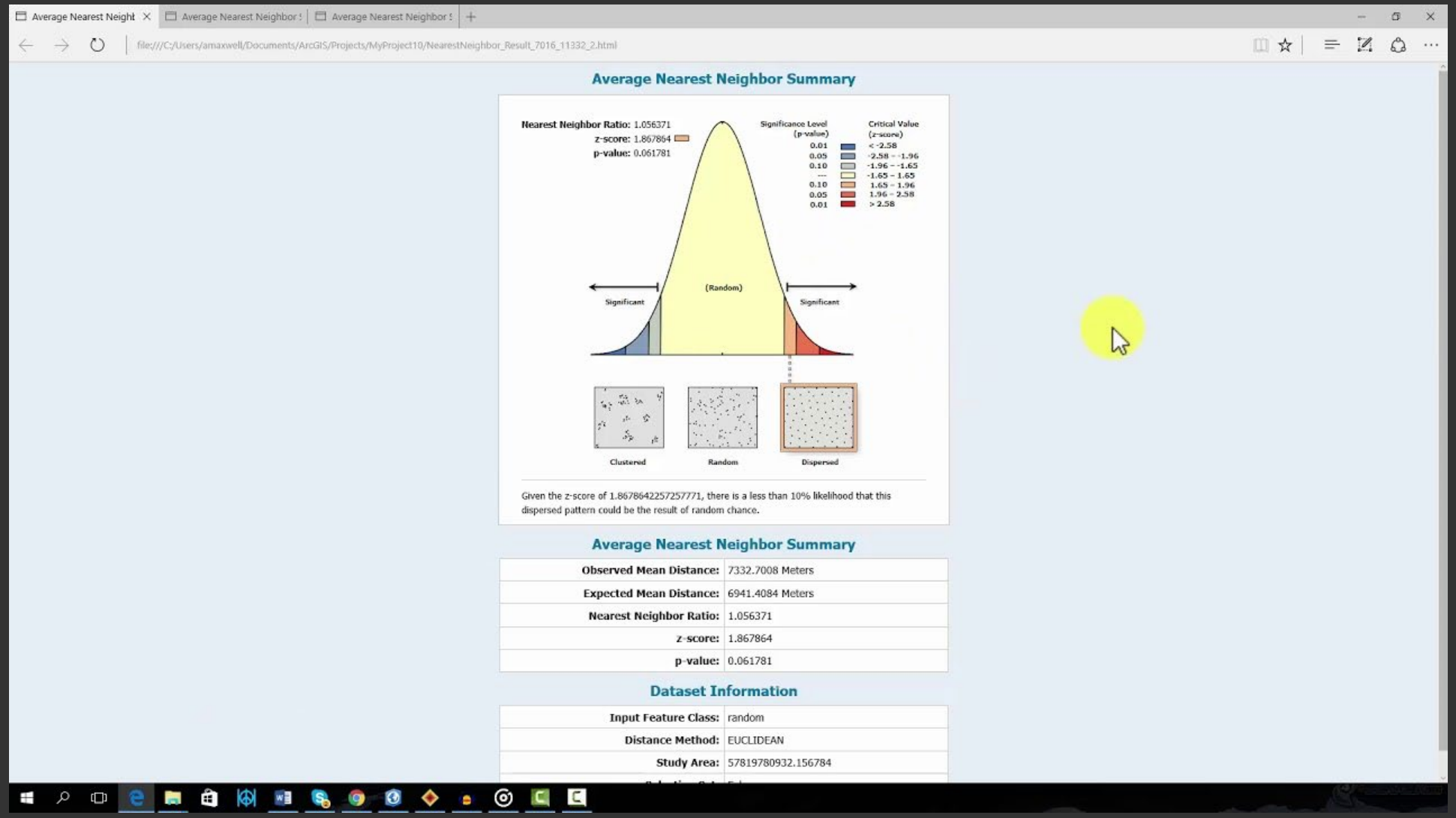


Average Nearest Neighbor Summary

Observed Mean Distance:	6302.1403 Meters
Expected Mean Distance:	8446.7271 Meters
Nearest Neighbor Ratio:	0.746104
z-score:	-5.868978
p-value:	0.000000



Video: Nearest Neighbor



Ripley's K

- Assess patterns at multiple scales or distances
- When the observed K value is larger than the expected K value for a particular distance, the distribution is more clustered than a random distribution at that distance (scale of analysis).
- When the observed K value is smaller than the expected K value, the distribution is more dispersed than a random distribution at that distance.

ExpectedK	ObservedK	DiffK
108958.883418	136149.67867	27190.795253
217917.766835	280286.368455	62368.60162
326876.650253	404103.709895	77227.059642
435835.533671	514082.560518	78247.026848
544794.417088	615369.007216	70574.590127
653753.300506	707204.895619	53451.595114
762712.183924	798759.51621	36047.332287
871671.067341	884449.103728	12778.036387
980629.950759	968118.086686	-12511.864072
1089588.834177	1046882.223296	-42706.61088

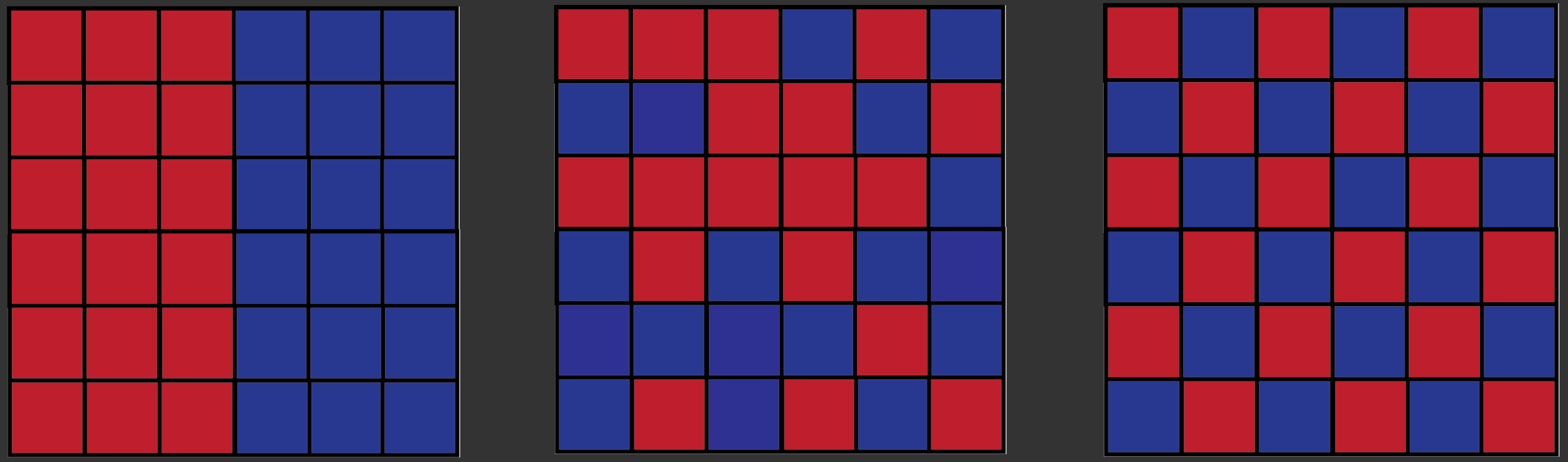
Global Spatial Autocorrelation

Tobler's First Law of Geography

Everything is related to everything else, but near things are more related than distant things.

Tober, W. (1970) "A computer movie simulating urban growth in the Detroit region." *Economic Geography*, 45(2): 234-240.

Tobler's First Law of Geography



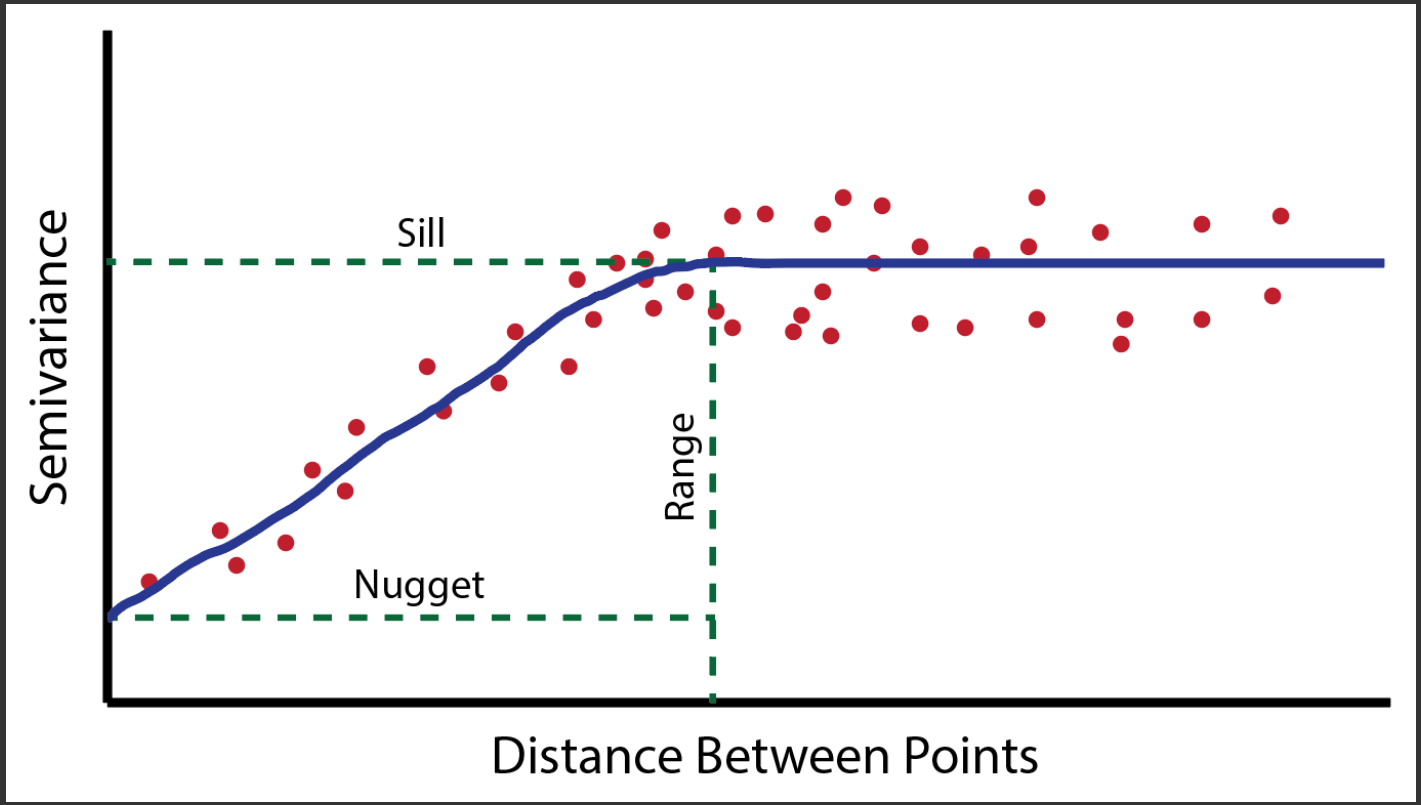
Clustered

Random

Dispersed

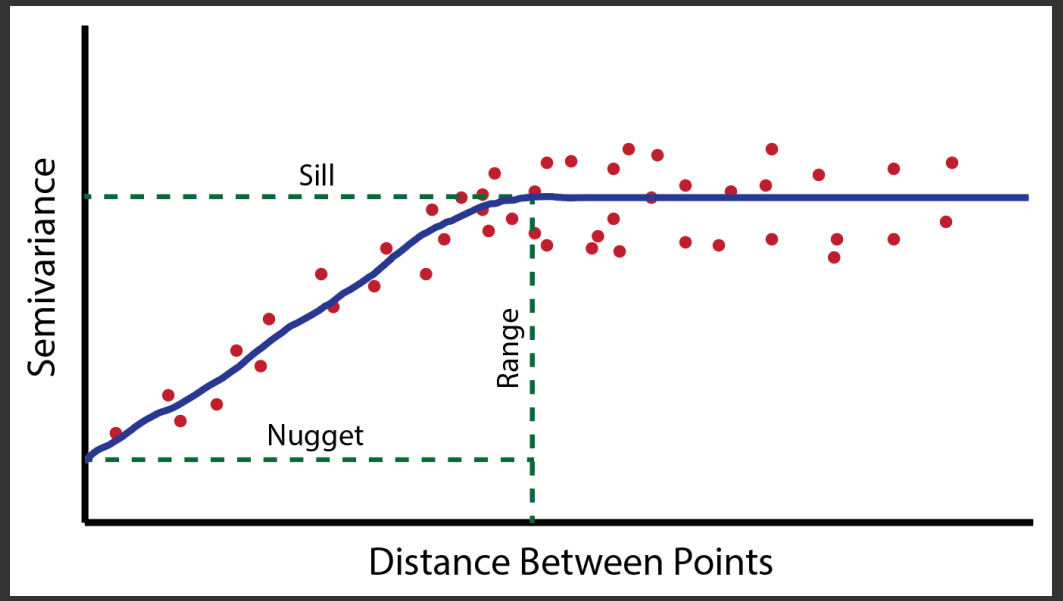
Semivariogram

- Semivariance = measure of the spatial dependence between two observations as a function of the distance between them
- = squared difference of the two values
- Semivariogram = a graph of how semivariance changes as the distance between observations changes

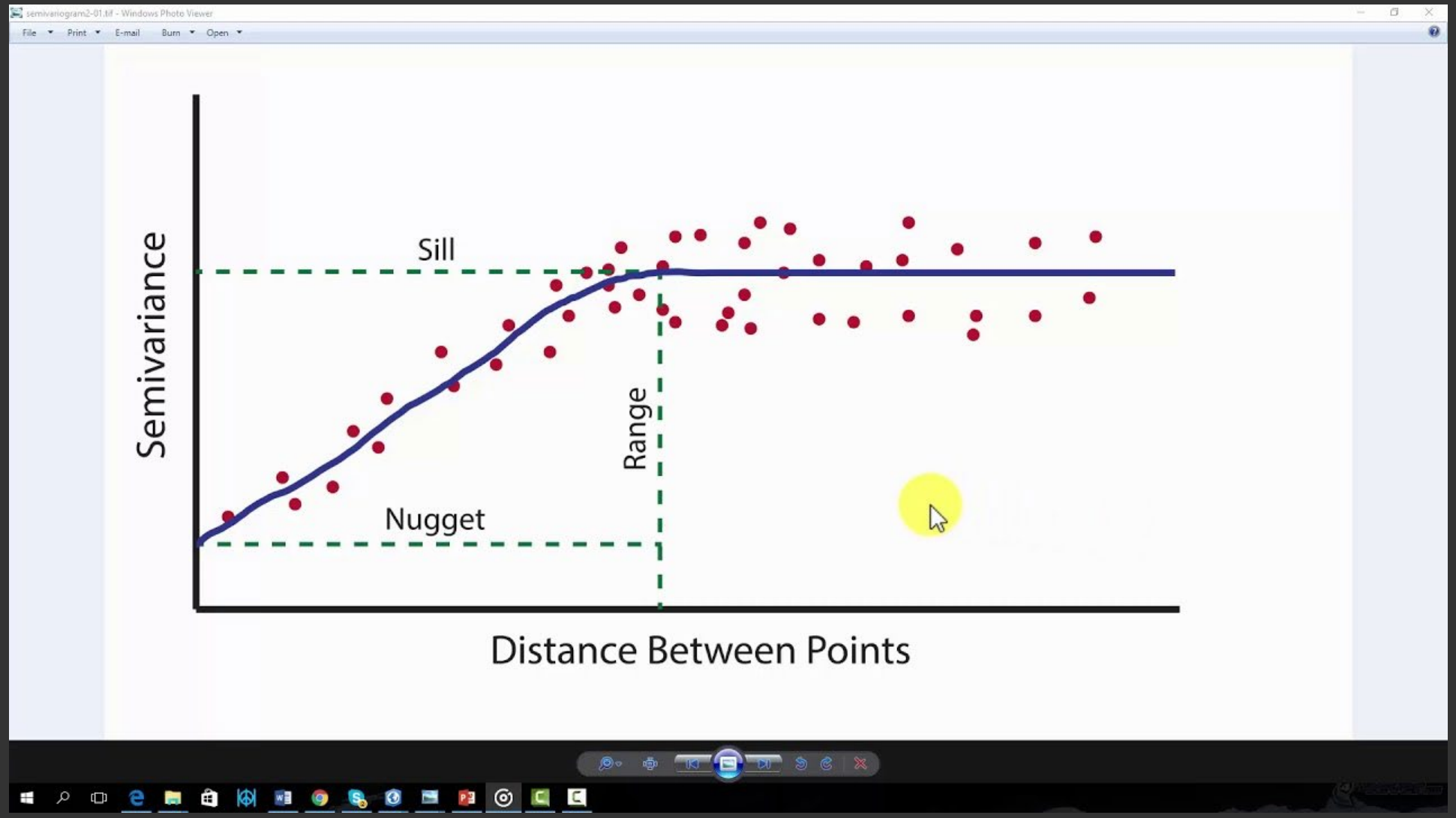


Semivariogram

- Range = distance where model first flattens out
- Sill = value of semi-variance at range
- Nugget = semi-variance at distance = zero
- Partial Sill = sill minus nugget
- Theoretically, at zero separation distance (lag = 0), the semivariogram value is 0. However, at an infinitesimally small separation distance, the semivariogram often exhibits a nugget effect, which is some value greater than 0.
- The nugget effect can be attributed to measurement errors or spatial sources of variation at distances smaller than the sampling interval or both. Measurement error occurs because of the error inherent in measuring devices. Natural phenomena can vary spatially over a range of scales. Variation at microscales smaller than the sampling distances will appear as part of the nugget effect. Before collecting data, it is important to gain some understanding of the scales of spatial variation.

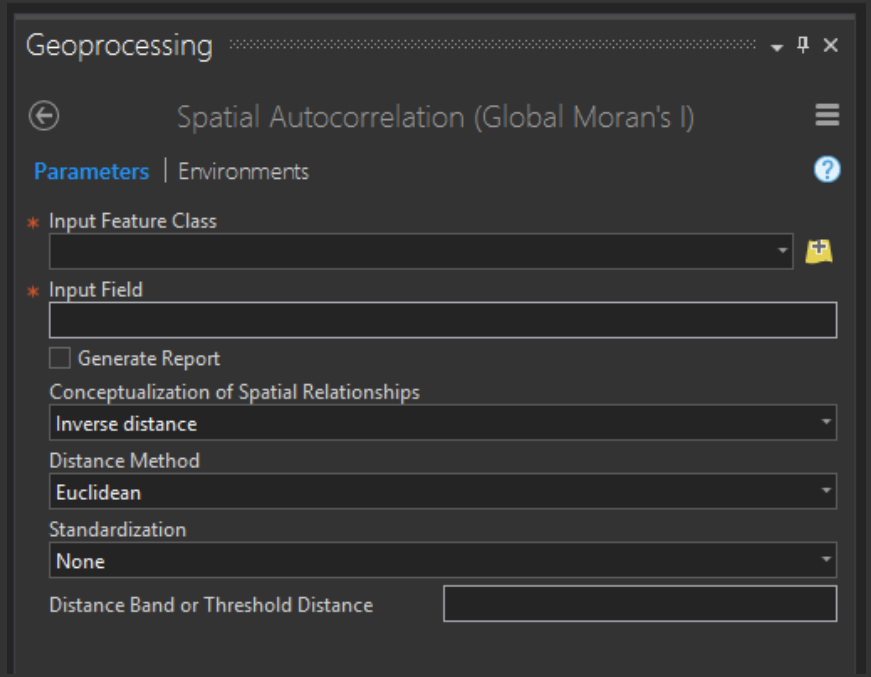


Video: Semivariogram



Moran’s I (Spatial Autocorrelation)

- Measures spatial autocorrelation based on feature locations and attribute values using the Global Moran's I statistic
- Null Hypothesis = spatial pattern of attributes is random
- Alternative Hypothesis = spatial patter of attributes is not random (spatially clustered or dispersed)
- Can use Euclidean Distance or Manhattan Distance
- Must specify a conceptualization of spatial relationships
- Positive Moran’s I = clustering
- Negative Moran’s I = dispersion

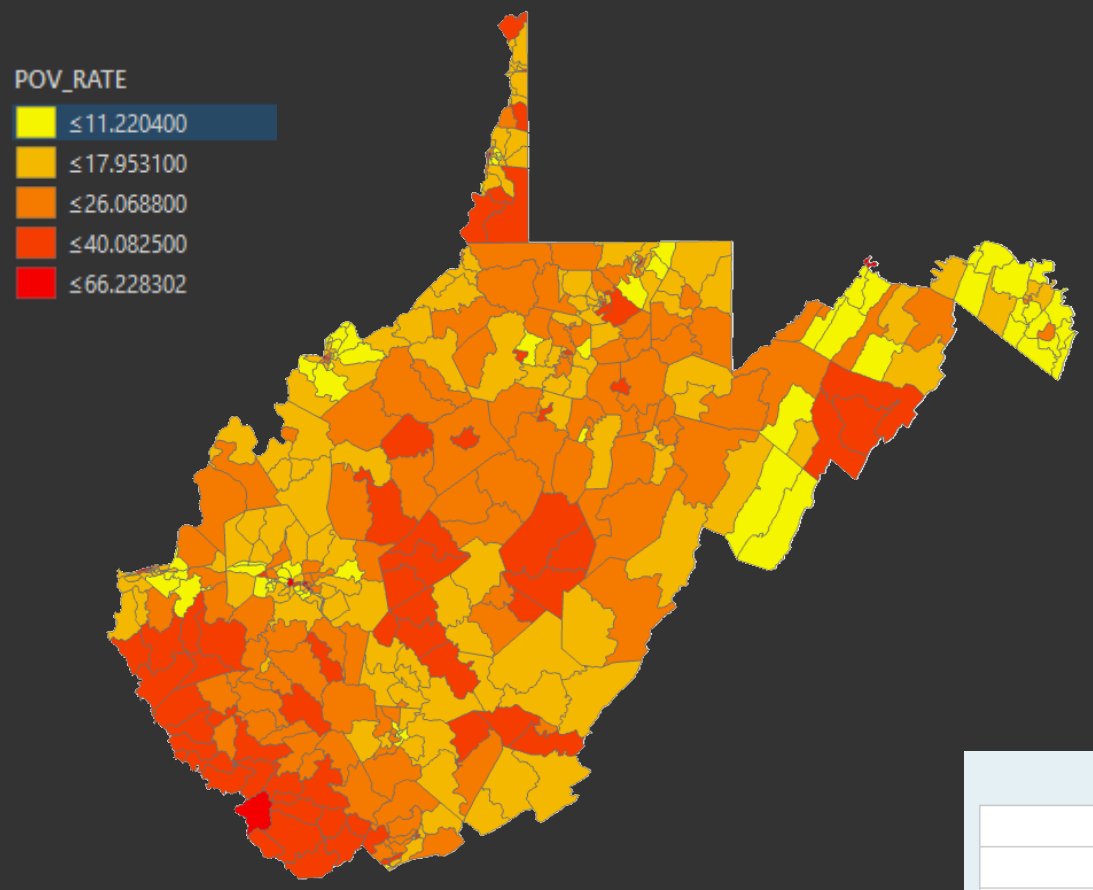


Conceptualization of Spatial Relationships

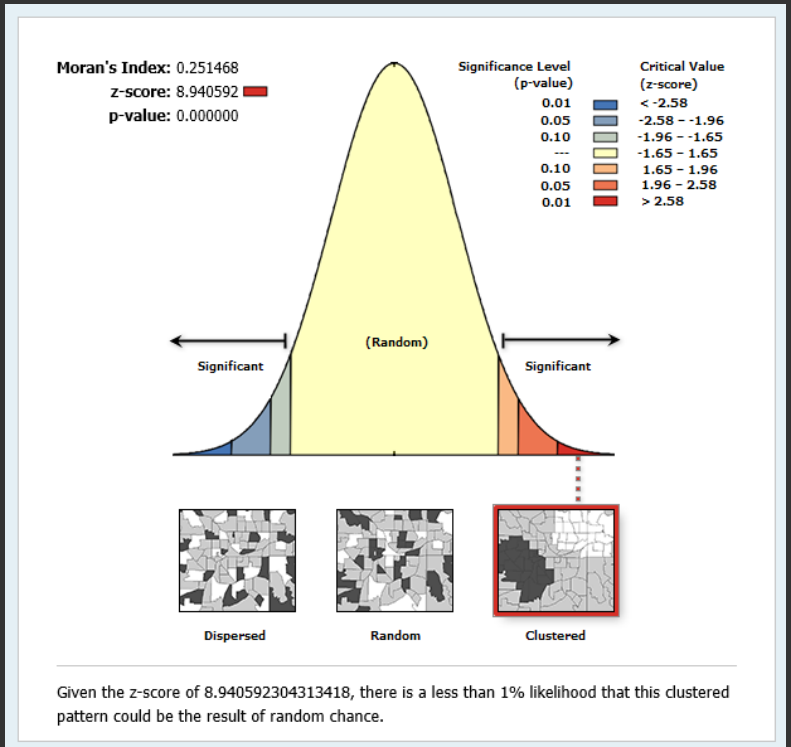
- INVERSE DISTANCE = Nearby neighboring features have a larger influence on the computations for a target feature than features that are far away.
- INVERSE DISTANCE SQUARED = Same as INVERSE_DISTANCE except that the slope is sharper, so influence drops off more quickly, and only a target feature's closest neighbors will exert substantial influence on computations for that feature.
- FIXED DISTANCE BAND = Each feature is analyzed within the context of neighboring features. Neighboring features inside the specified critical distance (Distance_Band_or_Threshold) receive a weight of one and exert influence on computations for the target feature. Neighboring features outside the critical distance receive a weight of zero and have no influence on a target feature's computations.
- ZONE OF INDIFFERENCE = Features within the specified critical distance (Distance_Band_or_Threshold) of a target feature receive a weight of one and influence computations for that feature. Once the critical distance is exceeded, weights (and the influence a neighboring feature has on target feature computations) diminish with distance.
- CONTIGUITY EDGES ONLY = Only neighboring polygon features that share a boundary or overlap will influence computations for the target polygon feature.
- CONTIGUITY EDGES CORNERS = Polygon features that share a boundary, share a node, or overlap will influence computations for the target polygon feature.
- GET SPATIAL WEIGHTS FROM FILE = Spatial relationships are defined by a specified spatial weights file. The path to the spatial weights file is specified by the Weights_Matrix_File parameter.

http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/spatial-autocorrelation.htm

Moran's I (Spatial Autocorrelation)



Percent Poverty by Census Tract

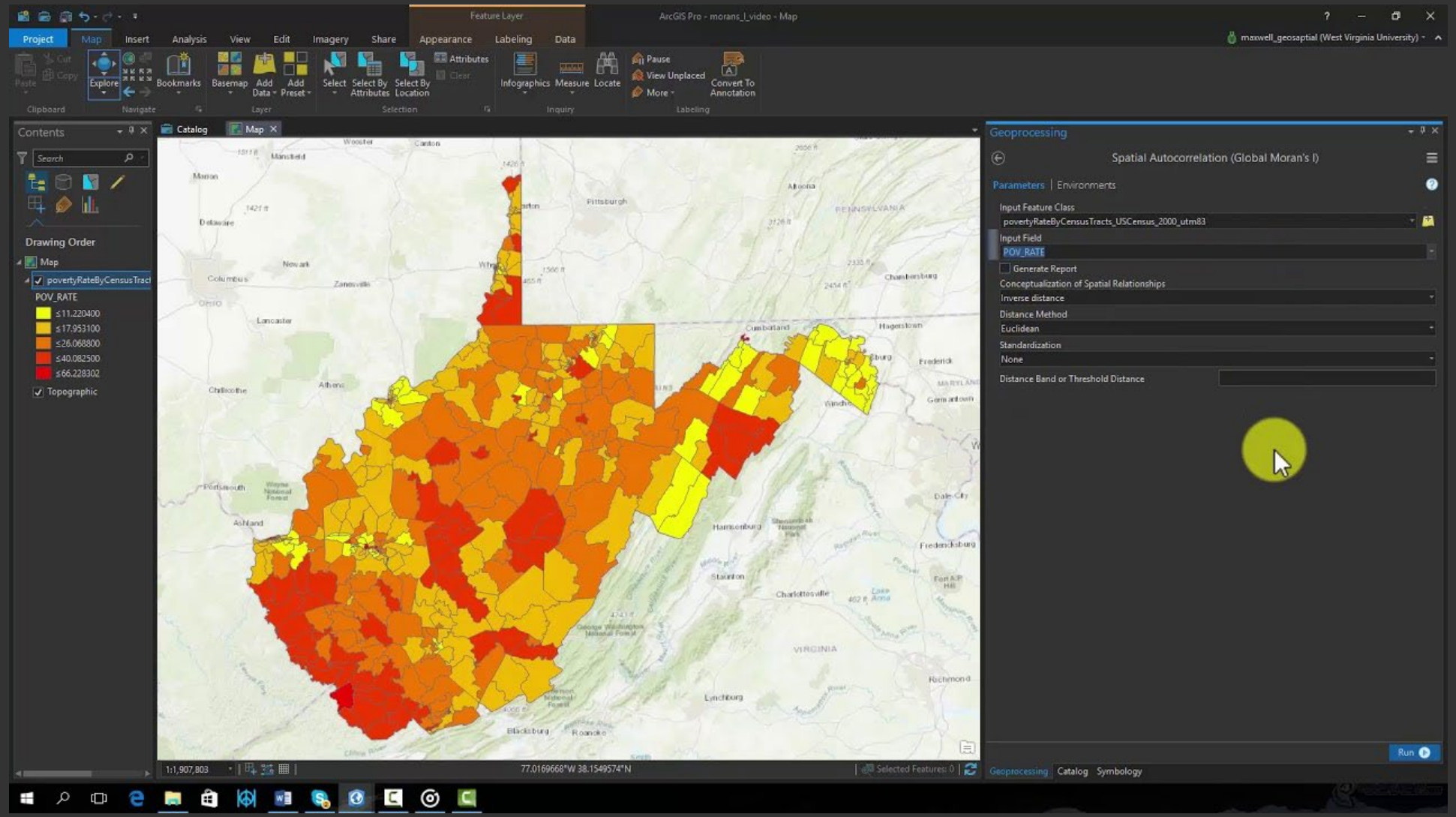


Global Moran's I Summary

Moran's Index:	0.251468
Expected Index:	-0.002151
Variance:	0.000805
z-score:	8.940592
p-value:	0.000000

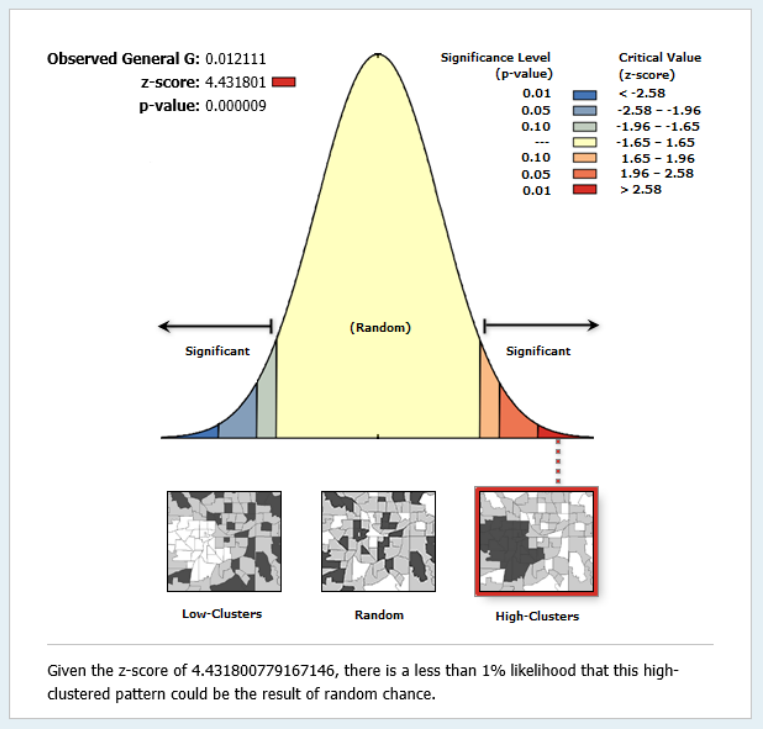
Used contiguity edges only

Video: Moran's I



Getis-Ord General G

- Moran's I assesses the clustering of similar values
- This statistic specifically looks for the clustering of high and low values globally
- Null Hypothesis = high values or low values aren't clustered
- Alternative Hypothesis = high and/or low values are clustered
- Can use Euclidean Distance or Manhattan Distance
- Must specify a conceptualization of spatial relationships
- Positive G = High clusters
- Negative G = Low clusters



Observed General G:	0.012111
Expected General G:	0.011223
Variance:	0.000000
z-score:	4.431801
p-value:	0.000009

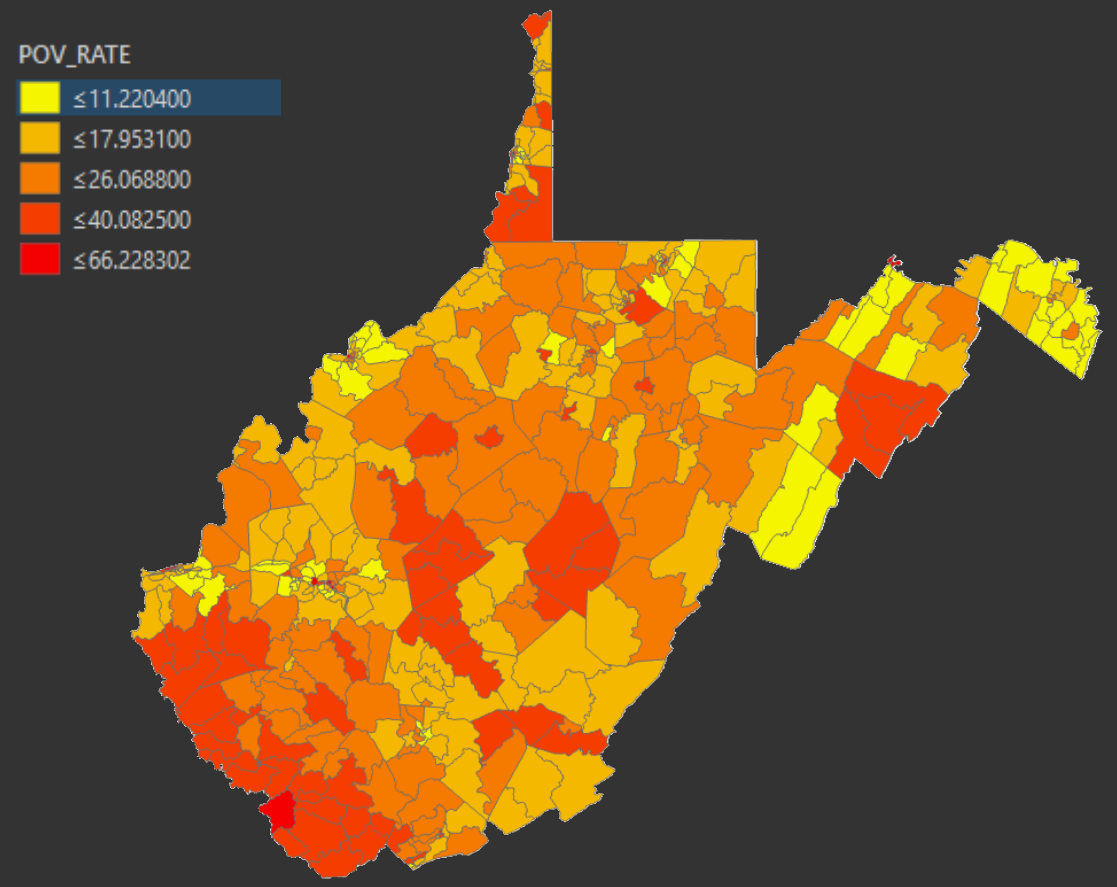
Used contiguity edges only

Local Spatial Autocorrelation

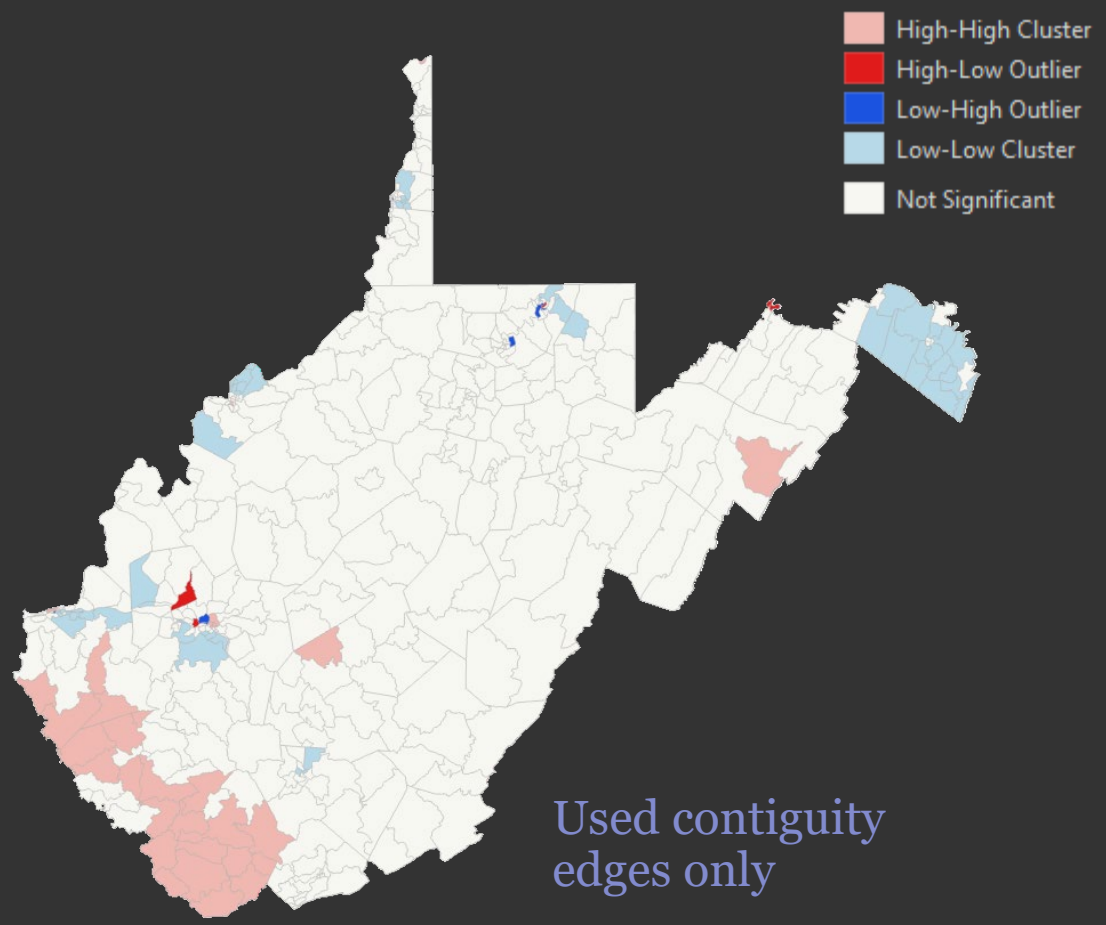
Local Indicators of Spatial Autocorrelation (LISA)

- A local version of Moran's I
- Looks for local patterns in the values associated with map features
- Null Hypothesis = local random pattern
- Alternative Hypothesis = local spatial autocorrelation (dispersion or clustering)
- Can use Euclidean Distance or Manhattan Distance
- Must specify a conceptualization of spatial relationships
- Positive I = feature has neighboring features with similarly high or low attribute values
- Negative I: feature has neighboring features with dissimilar values
- HH = cluster of high values
- LL = cluster of low values
- HL = high value surrounded by low values
- LH = Low value surrounded by high values

Local Indicators of Spatial Autocorrelation (LISA)



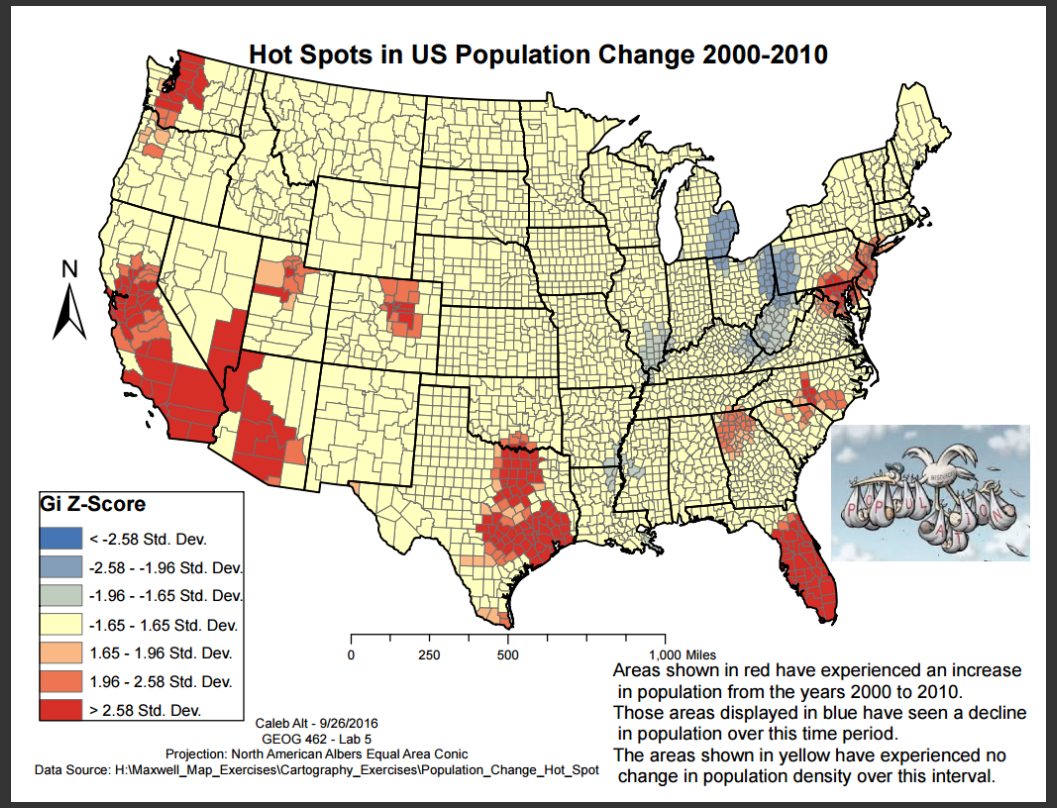
Percent Poverty by Census Tract



Used contiguity edges only

Hot Spot Analysis (Getis-Ord Gi*)

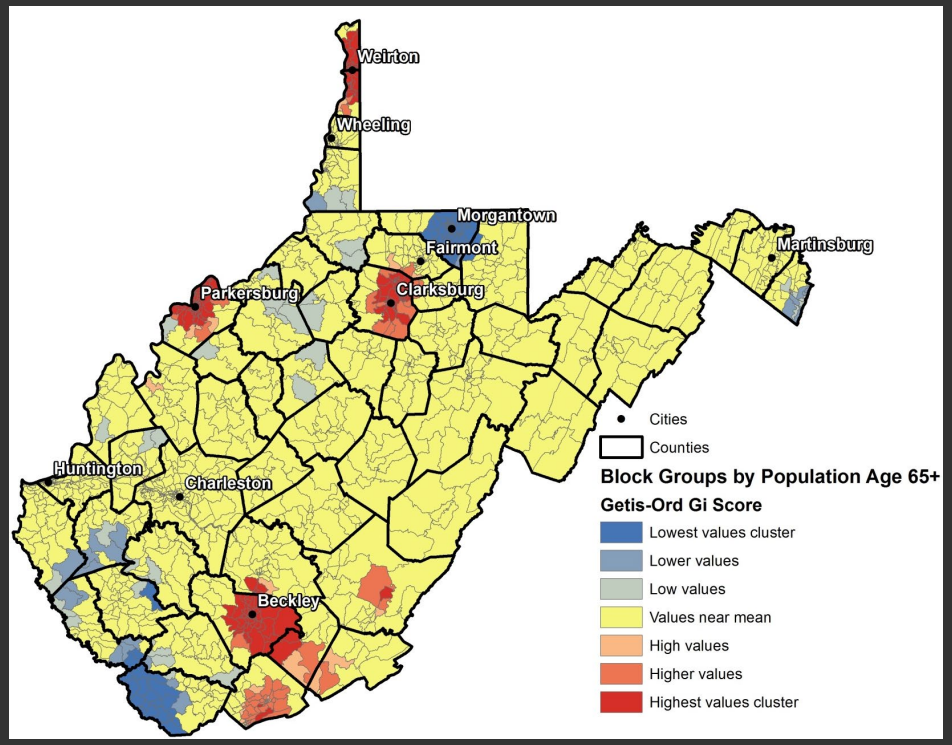
- This is a local version of the Getis-Ord General G
- It looks for local clustering of high or low values
- Hot Spot = Cluster of high values
- Cold Spot = Cluster of low values
- Null Hypothesis = No local cluster of high or low values
- Alternative Hypothesis = Local cluster of high or low values
- Can use Euclidean Distance or Manhattan Distance
- Must specify a conceptualization of spatial relationships
- Larger z-score = clustering of high values
- Smaller z-score = clustering of low values



Map by Caleb Alt

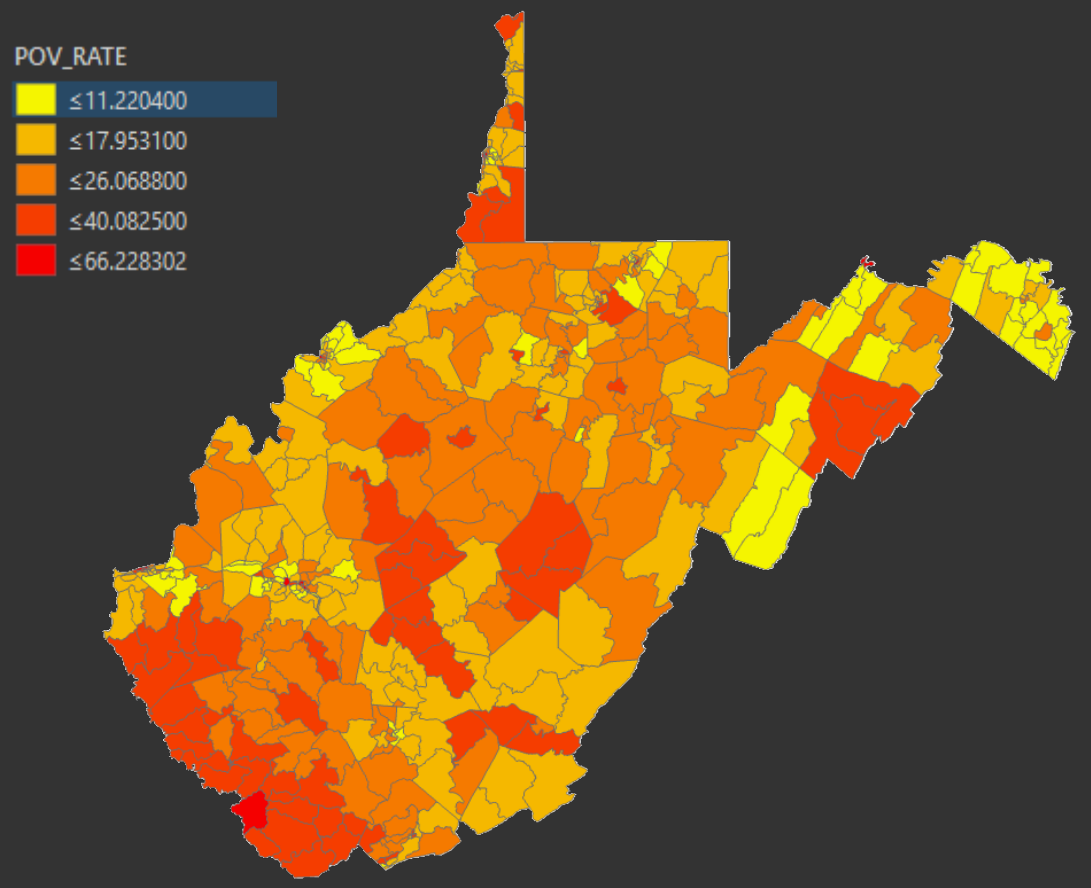
Hot Spot Analysis (Getis-Ord Gi*)

- This is a local version of the Getis-Ord General G
- It looks for local clustering of high or low values
- Hot Spot = Cluster of high values
- Cold Spot = Cluster of low values
- Null Hypothesis = No local cluster of high or low values
- Alternative Hypothesis = Local cluster of high or low values
- Can use Euclidean Distance or Manhattan Distance
- Must specify a conceptualization of spatial relationships
- Larger z-score = clustering of high values
- Smaller z-score = clustering of low values

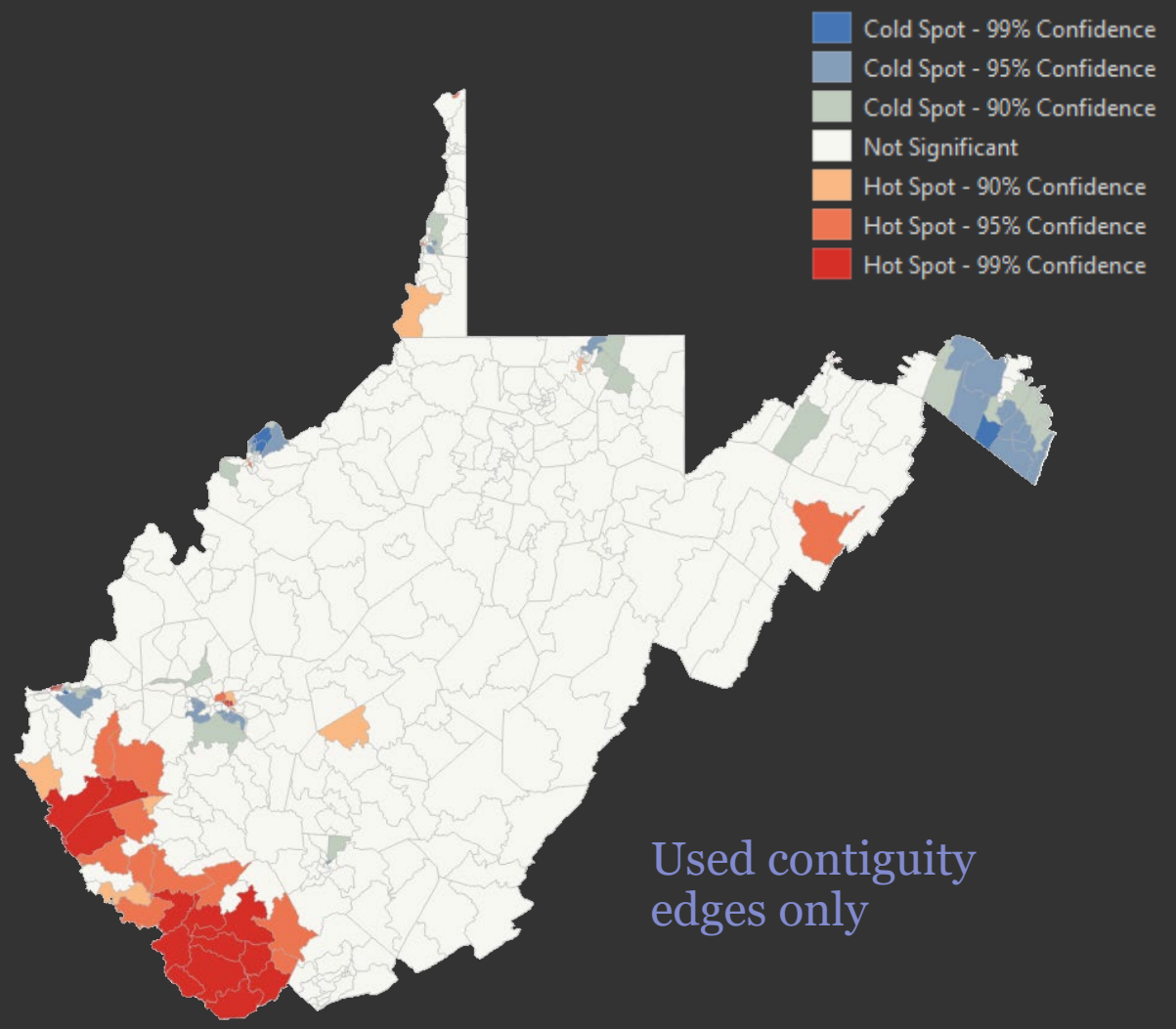


Example provided by
Mike and Jackie Strager

Hot Spot Analysis (Getis-Ord Gi*)



Percent Poverty by Census Tract



Used contiguity edges only

Regression

What is regression used for?

- Regression can be used to explore the relationship between variables
- You can predict a continuous variable using other variables
- Dependent Variable = what you are trying to predict
- Independent Variable(s) = what is used to predict the independent variable

Single Regression

$y = mx + b + \varepsilon$

$\hat{y} = mx + b$

y = true value being predicted

$\hat{y}$ = estimated value for y

m = slope

x = predictor or independent variable

b = y-intercept

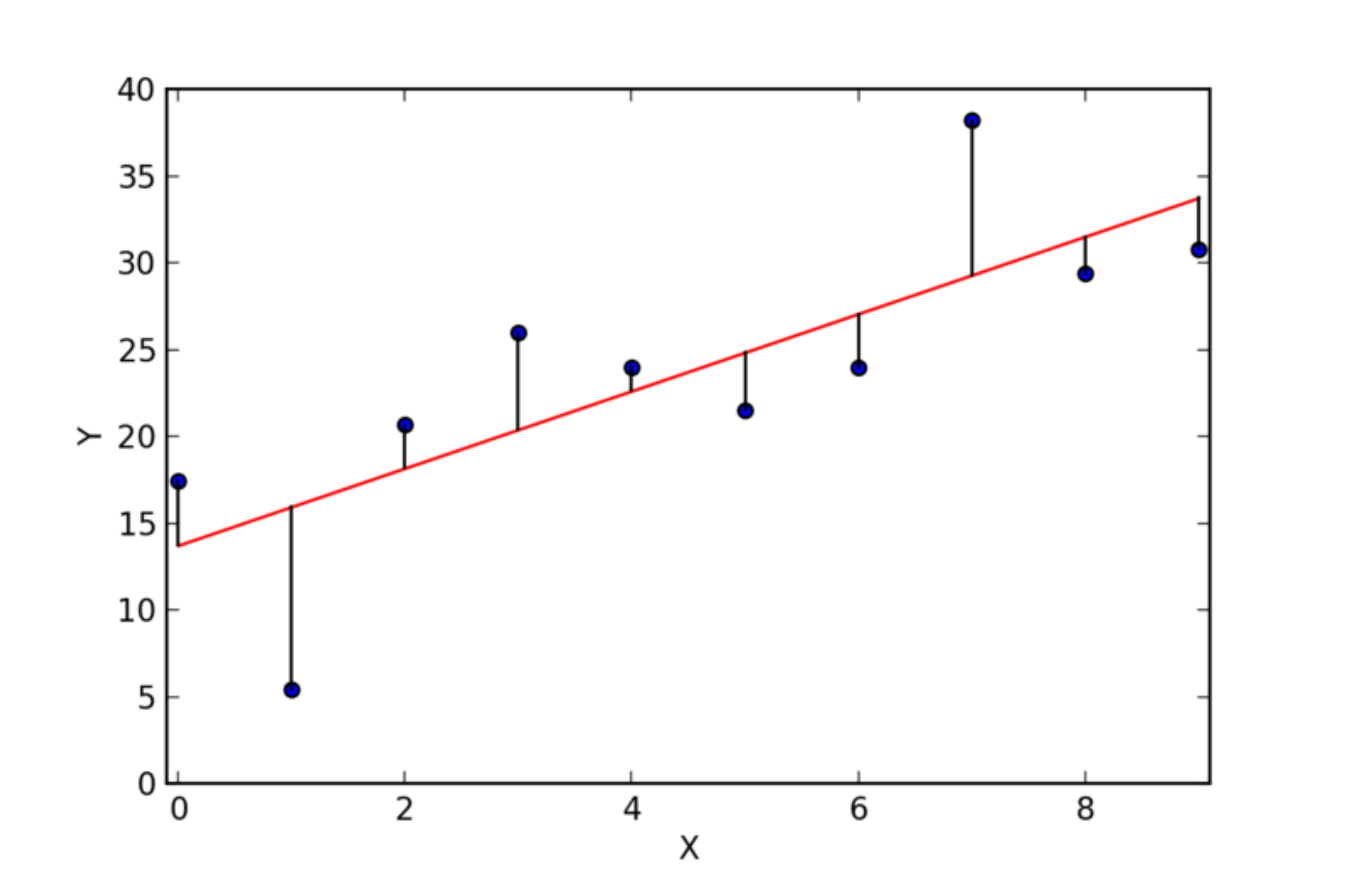
ε = the part of y not explained by x (residual)

Multiple Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\ldots\ldots\ldots \beta_n x_n + \varepsilon$$

- Predict Y using more than one independent variable

Residuals



$$\varepsilon = y - \hat{y}$$

https://commons.wikimedia.org/wiki/File:Residuals_for_Linear_Regression_Fit.png

Assessing Models: RMSE

- $RMSE = \sqrt{\Sigma(y - \hat{y})^2/n}$

- Root Mean Square Error (RMSE) will be in the units of y

- Mean Square Error (MSE) = $\Sigma(y - \hat{y})^2/n$

Assessing Models: R^2

$$R^2 = \frac{Total\ Sum\ of\ Squares - Residual\ Sum\ of\ Squares}{Total\ Sum\ of\ Squares}$$

Total Sum of Squares = $\Sigma(y - \bar{y})^2$

Residual Sum of Squares = $\Sigma(y - \hat{y})^2$

- Proportion of variance in y explained by the model

Assessing Models: Adjusted R^2

- When you have more than one x (multiple regression) you have to use adjusted R^2
- Adjusted R^2 = $1 - \frac{(1-R^2)(N-1)}{(N-p-1)}$
- N = sample size, p = number of variables

More on R^2

1. R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression
2. The percentage of the response variable variation that is explained by a linear model
3. R-squared = Explained variation / Total variation
4. R-squared is always between 0 and 100%
5. R-squared cannot determine whether the coefficient estimates and predictions are biased, which is why you must assess the residual plots
6. R-squared does not indicate whether a regression model is adequate. You can have a low R-squared value for a good model, or a high R-squared value for a model that does not fit the data!

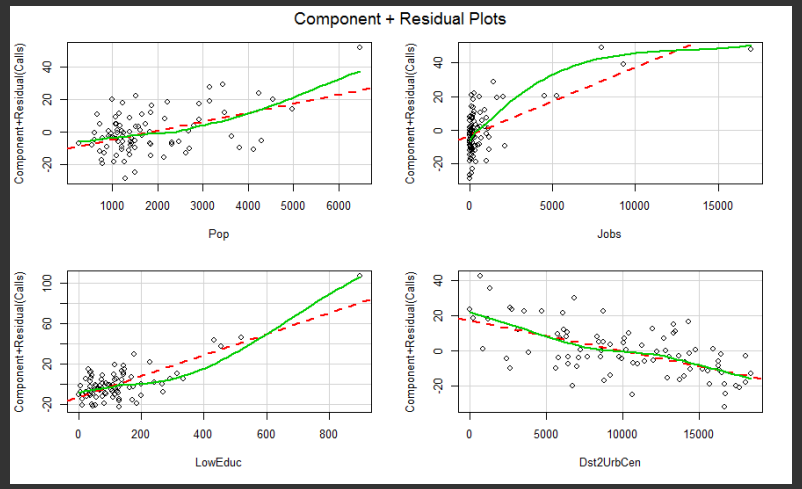
More on Adjusted R^2

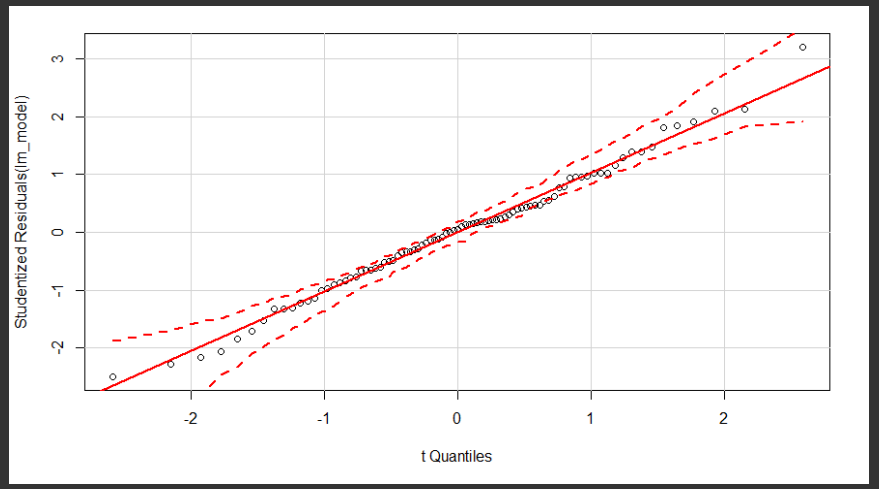
1. One major difference between R-squared and the adjusted R-squared is that R-squared supposes that every independent variable in the model explains the variation in the dependent variable. It gives the percentage of explained variation as if all independent variables in the model effect the dependent variable
2. Whereas the adjusted R-squared gives the percentage of variation explained by only those independent variables that in reality affect the dependent variable

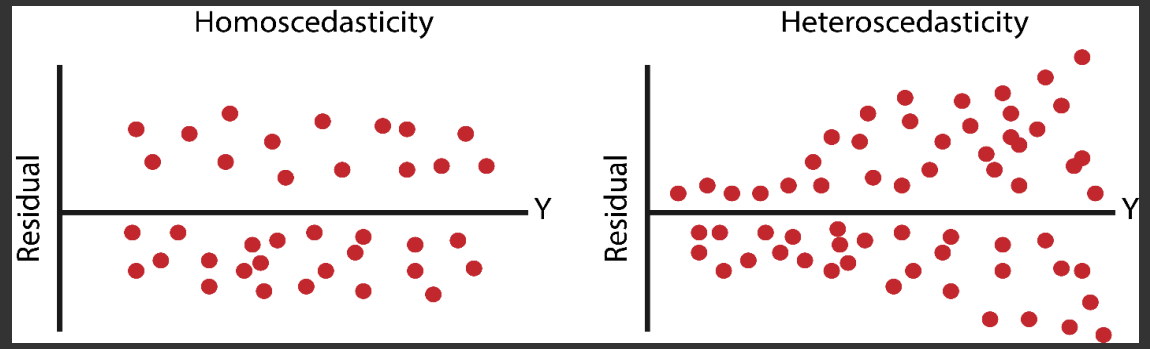
https://www.youtube.com/watch?v=KjRrdb2x6dA

Regression Assumptions

- Linear relationships
- Residuals are normally distributed (use QQ Plot)
- No or little multicollinearity (Variance Inflation Factor (VIF))
- No spatial autocorrelation (Moran's I)
- Homoscedasticity







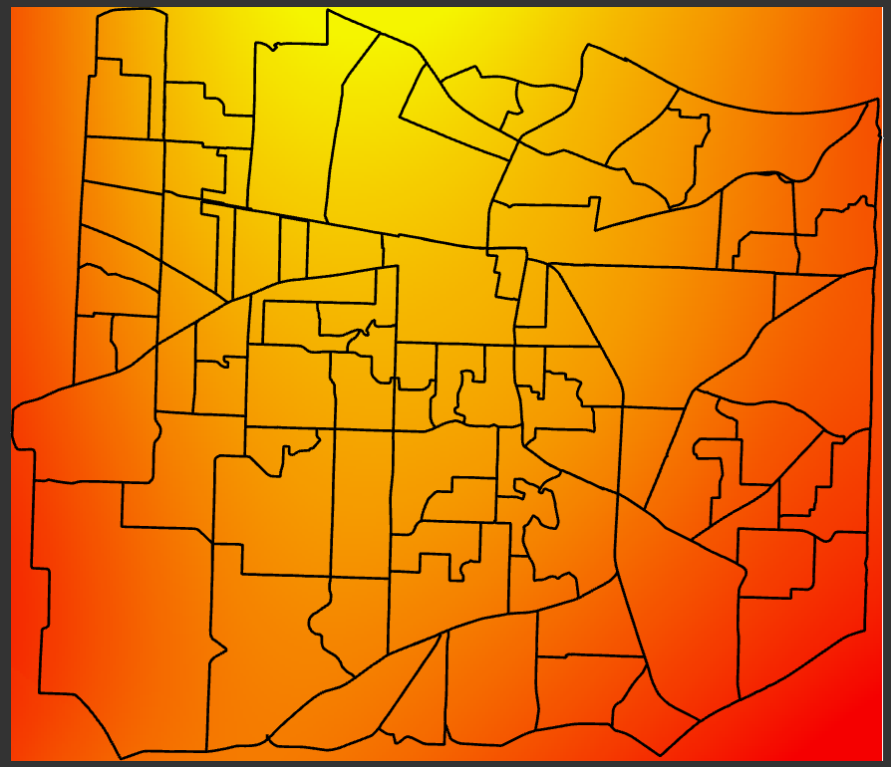
Geographically Weighted Regression (GWR)

- What if the relationship between y and x vary across space?
- In this case, a single regression equation could be misleading or weak
- It is possible to create regression equations that change over space
- There will be a different intercept (β_o) and coefficients (β_n) across the map space

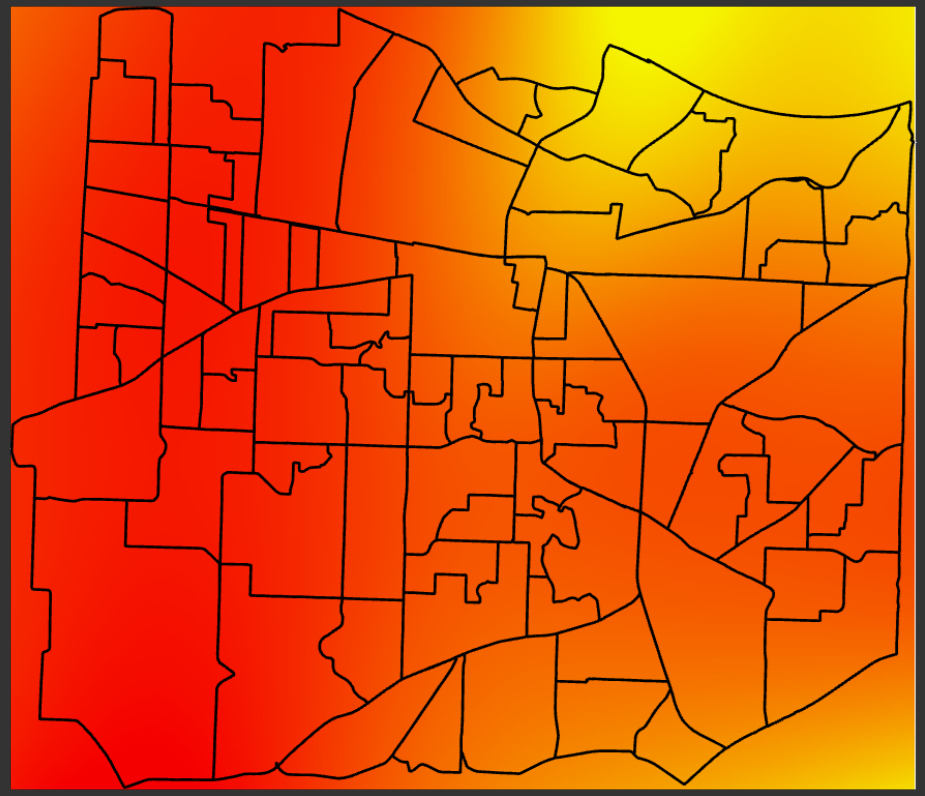
Predicted	Coefficient Intercept	Coefficient #1 Jobs	Coefficient #2 LowEduc	Coefficient #3 Dst2UrbCen	Coefficient #4 Pop	Residual
15.599533	18.887107	0.005548	0.081487	-0.001834	0.006131	-9.599533
18.917235	17.867772	0.005726	0.082911	-0.001712	0.00568	11.082765
10.805351	17.1108	0.005914	0.088301	-0.001523	0.004417	-2.805351
38.410381	19.789016	0.005427	0.080452	-0.001887	0.006279	-7.410381
37.225689	17.83196	0.005089	0.089096	-0.001761	0.006477	-1.225689
27.186792	15.908554	0.004816	0.093942	-0.001731	0.007013	11.813208
29.434545	14.380091	0.004465	0.095351	-0.001768	0.007796	-12.434545
17.755479	16.160131	0.004922	0.093266	-0.00176	0.006902	-6.755479
47.370895	13.371033	0.004103	0.095197	-0.001829	0.008492	-22.370895
25.859232	13.265688	0.004124	0.097559	-0.001804	0.008158	10.140768

https://www.youtube.com/watch?v=plfCMZhROeQ

Geographically Weighted Regression (GWR)



Y Intercept



Coefficient for Distance from Urban Center

https://www.youtube.com/watch?v=plfCMZhROeQ

This is the end of this lecture module.

Please return to the West Virginia View Webpage for additional content.