Geospatial Problem Solving

Spatial Statistics

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

This module will provide an introduction to spatial statistics. My goal here is to provide a broad overview of how we summarize and visualize data and also offer a conceptual introduction to spatial statistical methods.

This is simply an introduction. If you are interested in spatial statistics, I would suggest taking a full spatial statistics course. This skillset is very valuable for anyone wanting to work with and analyze spatial data and spatial patterns.

Module Topics

1. Exploratory data analysis
2. Data summarization
3. Data summarization by group
4. Spatial sampling
5. Tessellation
6. Measures of spatial central tendency and dispersion
7. Point pattern analysis
8. Global spatial autocorrelation
9. Local spatial autocorrelation
10. Regression and geographically weighted regression

2

These are the topics covered in this module.

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

3

Unique concerns arise when we want to explore and analyze spatial data statistically.

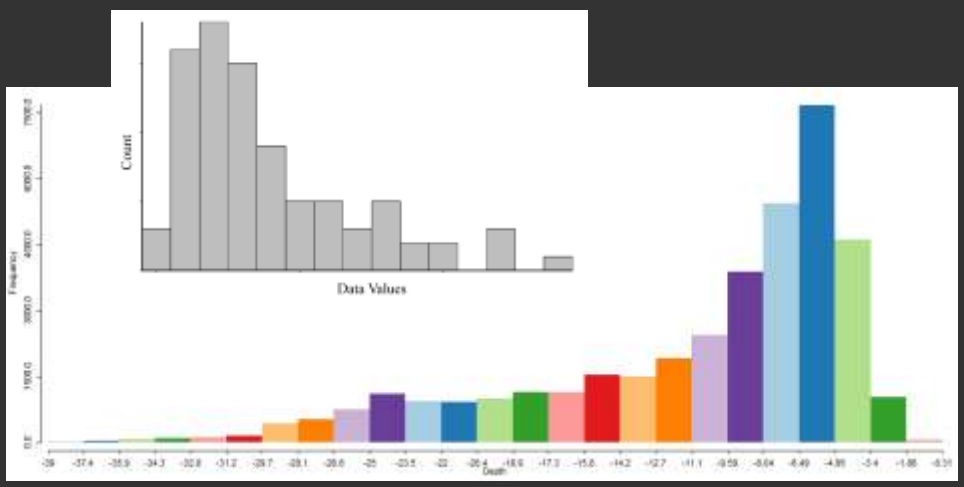
First, they occur over a map space, so issues of spatial heterogeneity and spatial autocorrelation arise, which can make it difficult to apply traditional statistical methods that assume samples are independent of one another.

Although this issue is not necessarily unique to spatial data, there can also be issues with temporal heterogeneity and autocorrelation.

As mentioned in previous modules, the scale of the data aggregation can impact the analysis results, which we term the Modifiable Areal Unit Problem or MAUP.

Exploring Your Data

Data Distribution: Histograms



5

We will begin with a discussion of visualizing and describing a single variable graphically or statistically. Note that we are not making use of the spatial information here. This only relies on the tabulated data.

Histograms are one means to explore the distribution of a single variable and are an example of a univariate graph.

The x-axis represents the attribute or quantity of interest, and the y axis will be the frequency or count.

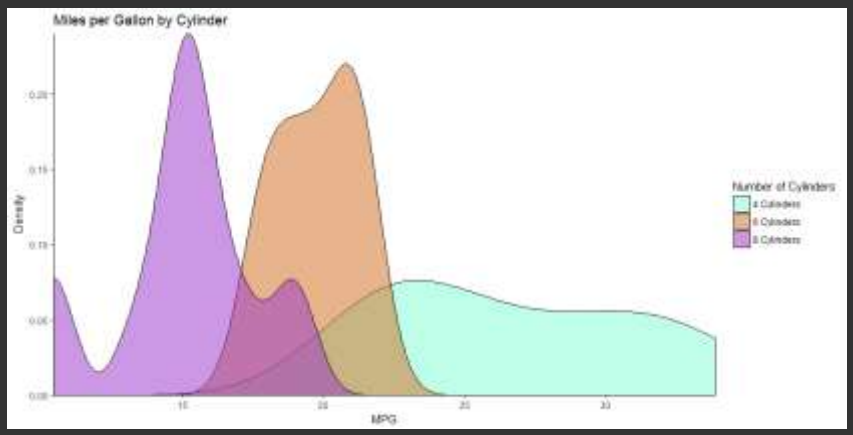
Data ranges are binned and the number of samples within each data range are counted.

This allows us to visualize the distribution of the data to assess central tendency, dispersion, skewness, or the presence of outliers.

How the data are binned can have a large impact on how patterns are observed. Smaller bins tend to highlight more local patterns whereas larger bins will represent a generalized pattern.

There is not necessarily a correct bin width. This depends on the patterns of interest and the data being graphed.

Data Distributions: Kernel Density Plots



6

Similar to histograms, kernel density plots are used to show the distribution of a single variable, or are univariate graphs.

The x-axis will be the variable of interest, and the y-axis will be density, count, or frequency.

The curves are created using a kernel density function. Peaks indicate common values and troughs indicate less frequent values.

You can also use these graphs to compare data distributions for different categories, as shown in the example graph.

Since histograms and kernel density plots are showing the same patterns and can have the same axes, they can be plotted in the same graph space.

The kernel density function can be adjusted to show more local or generalized patterns, similar to changing the bin width for histograms.

Measures of Central Tendency

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

7

Several statistics are available to describe the central tendency of a continuous variable.

The mean or average is calculated by summing all the samples then dividing by the number of samples. The mean for a population can be approximated using the mean calculated from a sample of that population.

The median represents the most middle number, the 50% percentile, or the 2nd quartile. 50% or half of the values will be lower than it and 50% will be larger than it.

The mode is the most commonly occurring value.

The mean tends to be impacted by outliers or extremely large or small values. So, if this is an issue in a dataset, the median is sometimes used as opposed to the mean to describe central tendency.

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 = $s = sqrt [\Sigma (x_i - x\_bar)^2 / (n - 1)]$

8

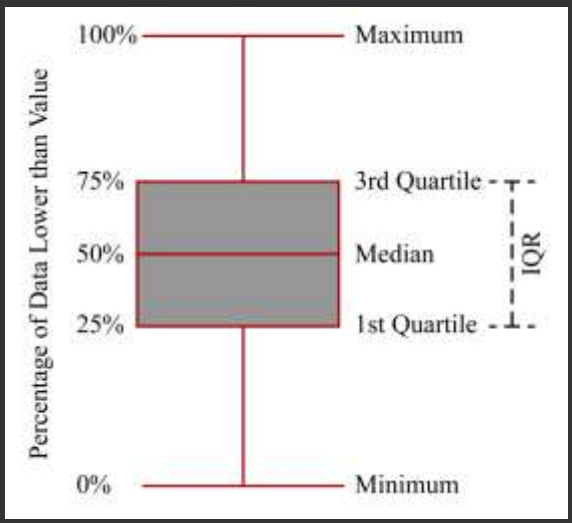
Central tendency only explains one component of your data. It is also important to describe the variability or dispersion of the data. This can be done using a variety of measures.

For example, you can calculate the largest value in the dataset as the maximum and the smallest as the minimum. Subtracting the minimum value from the maximum value yields the range.

You can calculate population variance (σ^2) or approximate is using sample variance (s^2). For population variance, the mean is subtracted from each value. These differences are then squared, summed, and divided by the number of features. Note that variance will be reported in the square of the units of the variable. For example, if concentration data are measured in parts per million (ppm), then the units of the variance would be ppm^2.

The population standard deviation (σ) is simply the square root of the variance. Taking the square root will return the original units of measurement for the variable. Larger standard deviations suggest that the data are more dispersed or variable. The population standard deviation can be approximated from a sample as the sample standard deviation (s).

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

9

Another useful tool for describing data distribution is the box plot.

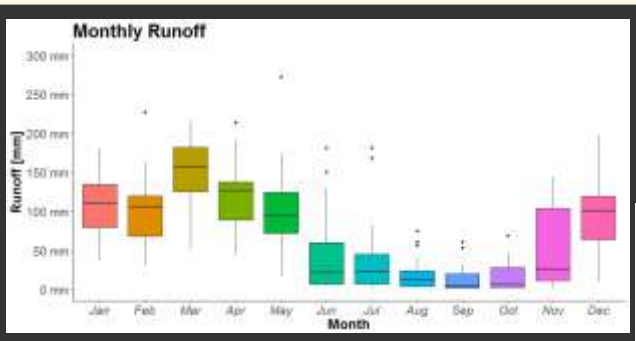
Using this method, we graph the median, 1st quartile, 3rd quartile, minimum, and maximum values as defined and shown on the slide. Note that the mean is generally not included on the boxplot.

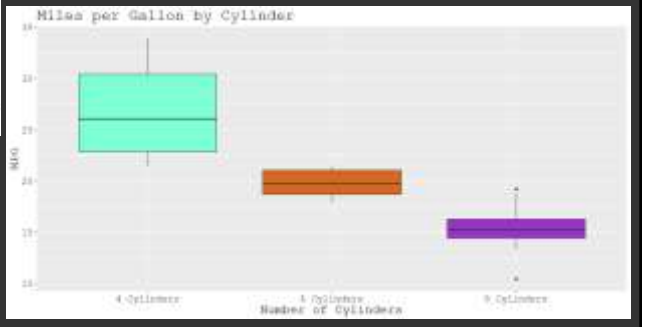
The 1st quartile is defined as the value for which 25% of the data are below it and 75% are above. In contrast, the 3rd quartile has 75% of the data below it and 25% above. The difference between the 3rd and 1st quartile is called the interquartile range or IQR.

You can also graph outliers as points. The definition of outliers can vary; however, one commonly used criteria is that large values that are more than 1.5 IQR from the 3rd quartile and small values that are more than 1.5 IQR from the 1st quartile are outliers.

Similar to histograms and kernel density plots, boxplots can also be useful in visually assessing the central tendency, dispersion, and normality of your data.

Box Plot





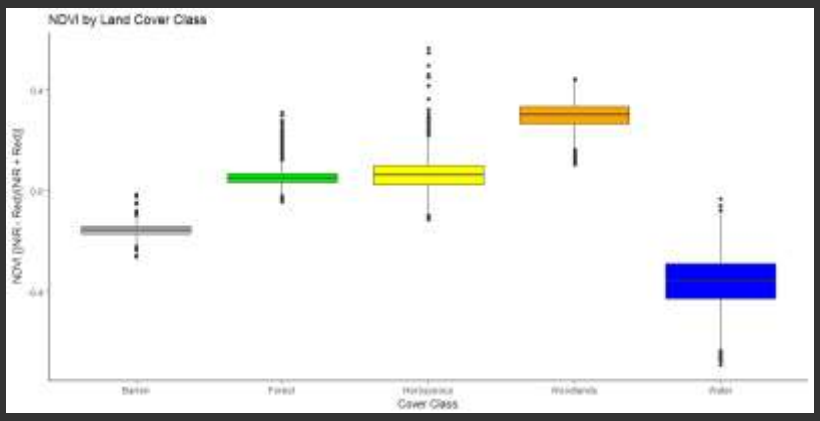
10

Boxplots can be used to compare groups.

For example, the first graph shown here compares the central tendency and variability in stream runoff by month.

The second graph compares gas mileage based on the number of engine cylinders.

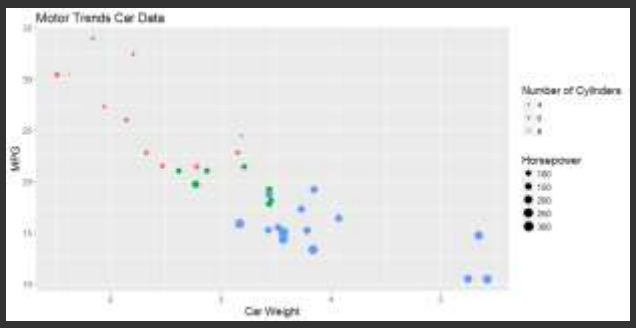
Box Plot

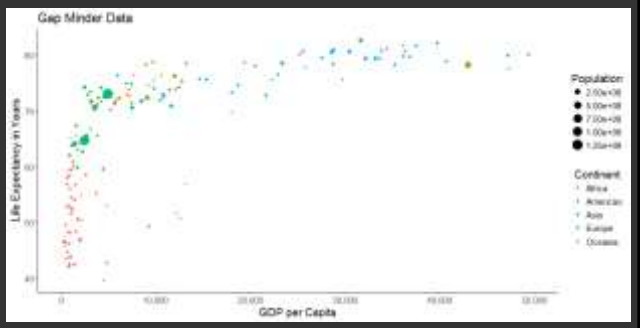


11

Here is another example. In this example, the normalized difference vegetation index, or NDVI, is compared for different land cover types.

Relationships Between Variables: Scatterplot





12

Now we will move on to bivariate or multivariate graphs in which more than one variable is shown, generally to allow for comparisons.

A common example is the scatterplot.

Scatterplots allow you to explore the relationship or correlation between two variables, one plotted to the x-axis and the other plotted to the y-axis. Generally, the independent variable is plotted to the x-axis and the dependent variable is plotted to the y-axis. For example, if it is assumed that a person's weight is partially dependent on or related to the person's height, then height would be graphed to the x-axis and weight would be graphed to the y-axis.

For time series graphs the x-axis is generally time while the y-axis is a variable that can change over time, such as temperature or barometric pressure at a location.

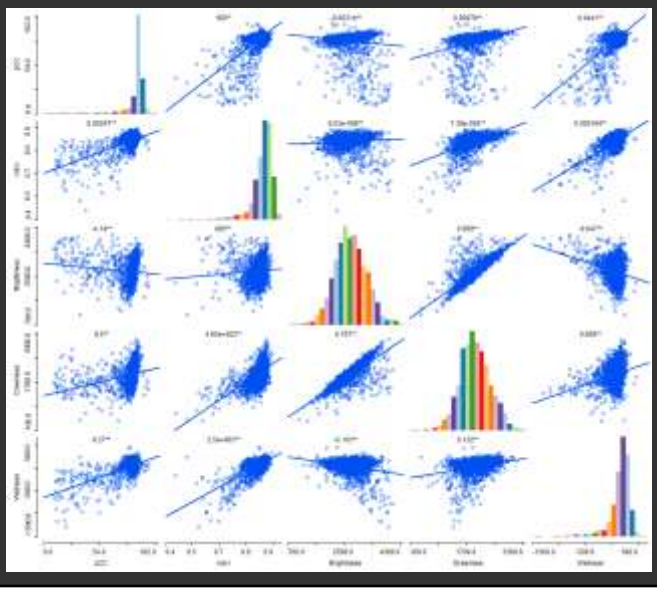
Note that it is not always clear which variable is the dependent and which is the independent variable.

If one variable tends to increase as the other increases, this is known as a positive or direct relationship. In contrast, if one variable decreases as the other increases this would be a negative or indirect relationship.

In the first example, car weight and fuel efficiency are being compared, and the graph suggests a negative or indirect relationship: fuel efficiency tends to decrease with car weight. Car weight was plotted on the x-axis and fuel efficiency on the y-axis because it is assumed that fuel efficiency is being impacted by weight. So, fuel efficiency is the dependent variable and car weight is the independent variable. It is also possible to show additional variables using other graphical parameters. For example, in this graph color represents the number of engine cylinders and the dot size represents the engine's horsepower.

The second graph provides a comparison of gross domestic product per capita and life expectancy for countries. Also, color represents the continent and size represents the country population. Here, we see a correlation between GDP per capita and life expectancy. However, it does not appear to be a linear relationship.

Relationships Between Variables: Scatterplot Matrix

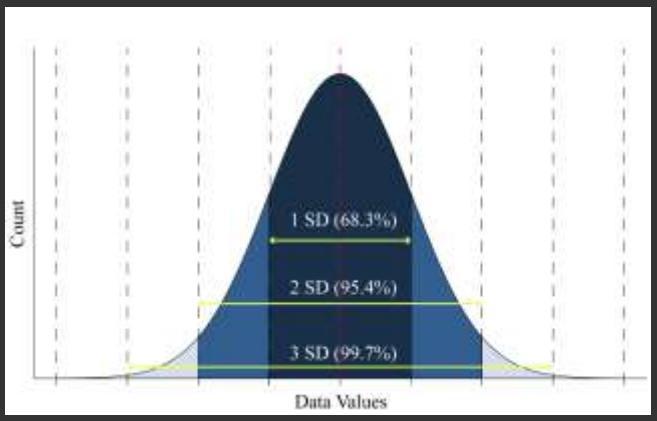


13

It is also possible to compare multiple variables using a scatterplot matrix. Using this method, all variables are compared as pairs within a matrix where rows and columns are defined by the variables.

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



14

Sometimes we compare our data distribution to a theoretical distribution, such as a normal distribution. Further, some statistical tests assume that data follow a defined distribution in order for the test to be valid. Such tests are called parametric tests.

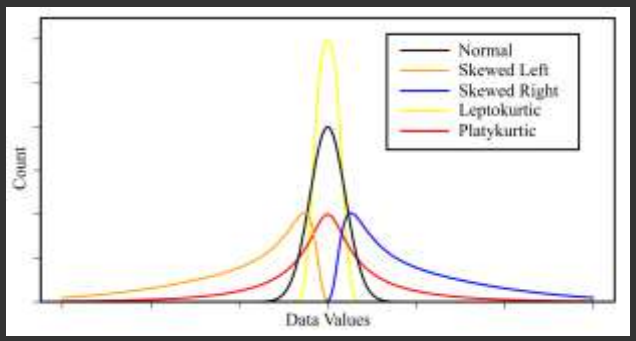
One example is the normal distribution, which has defined characteristics.

It is bell-shaped, the mean, median, and mode are equivalent, and it is symmetrical about the mean, which is shown in the example as a red dotted line.

The shape of the normal curve varies based on the mean and standard deviation of the data. However, 68.3% of the data are within one standard deviation of the mean, 95.4% of the data are within two standard deviations of the mean, and 99.7% of the data are 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)



15

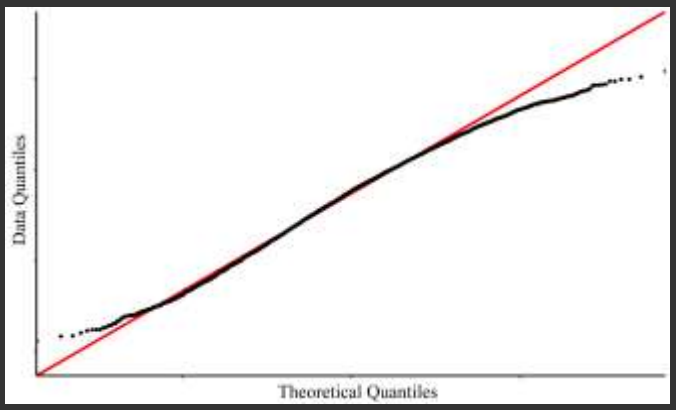
We can compare our data distribution to a normal distribution using skewness and kurtosis.

Skewness measures the symmetry of the data distribution. Negative values indicate that the data are skewed to the left or have a tail to the left, representing low values, whereas positive values indicate that data are skewed to the right or have a tail to the right, presenting high values. Skewness values near zero indicate symmetrical data and is one indication of normality.

Kurtosis measures how variable or dispersed the data are relative to a normal distribution. Kurtosis exists when the percentage of the data occurring within a specified standard deviation of the mean differs from the percentage defined for a normal distribution. Data are considered leptokurtic if they have less variability or are less dispersed than a normal distribution, platykurtic if they are more variable or dispersed than a normal distribution, and mesokurtic if the variability approximates that of a normal distribution.

So, data that are normally distributed will have a skewness near zero and be 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

16

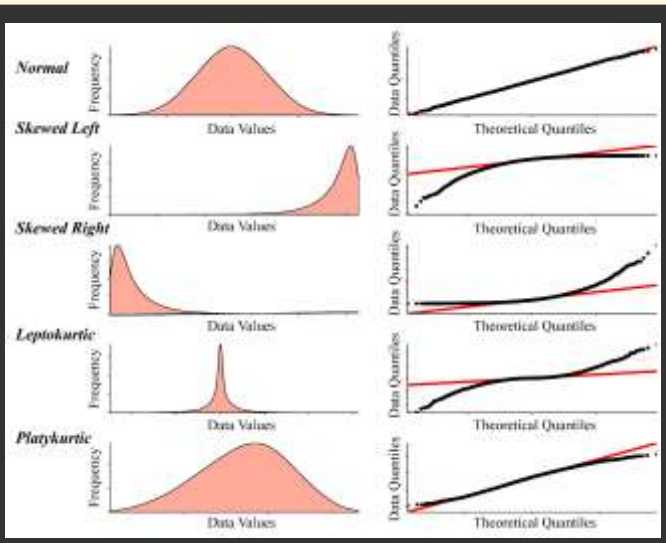
Another way to assess normality is by using a QQ Plot.

A QQ Plot compares the quantiles of the data to theoretical quantiles that represent a normal distribution of the data.

Features that fall along the reference line have quantiles that approximate those of a normal distribution.

If the points fall off of the reference line, this suggests issues in normality. The data may be skewed, have kurtosis issues, or be multimodal.

Are your data normally distributed?



Q-Q Plot

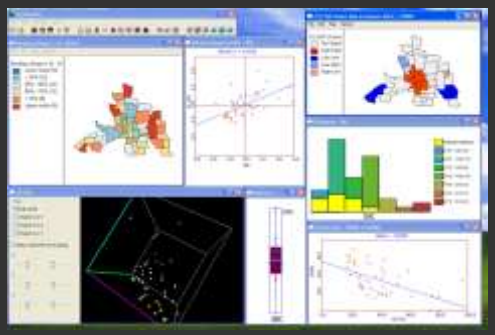
- On line = normal
- Concave = skewed to right
- Convex = skewed to the left
- High/low off line = kurtosis

17

Generally, QQ plots will take on a convex shape if the data are skewed to the left, a concave shape if skewed to the right, and will diverge from the reference line at high and low values if there are kurtosis issues.

GeoDa

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



https://geodacenter.asu.edu/

18

If you would like to visualize and explore your data, I would recommend looking into the free and open-source GeoDa tool made available by the Center for Spatial Data Science at the University of Chicago.

Sampling

Sampling

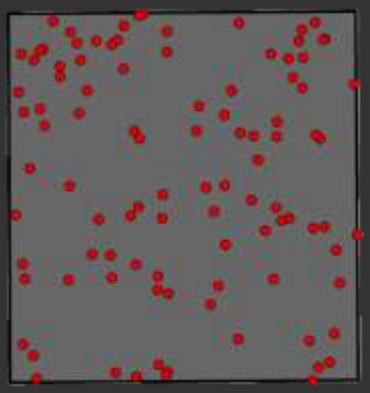
- A sample is a set of measurements taken from a larger group or population.
- Sample means and variances can serve as estimates for their population.
- Easier to measure with samples, then draw conclusions about entire population.

20

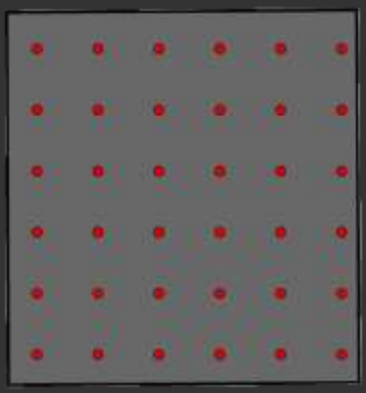
It is generally not possible to measure all members of a population. For example, if you wanted to conduct a study in which all undergraduates at a university are considered the population, it would be very difficult to collect data for every single student. As additional examples, it would not be feasible to measure every tree in a forest or every fish in a river.

So, we must collect a sample from the larger population. In order to draw unbiased conclusion about the population from the sample, it must be an unbiased representation of the population. This is accomplished using random sampling. Using appropriate sampling methods will allow you to approximate the mean and variance of the population with the sample.

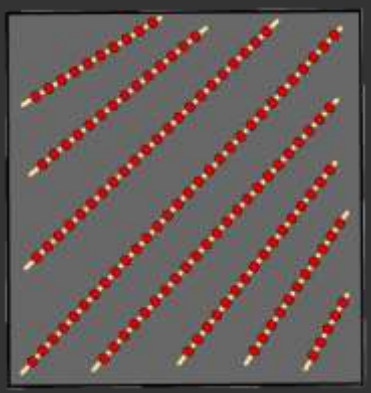
Types of Sampling



Simple Random Sampling



Systematic Sampling



Systematic Sampling Along Transects

21

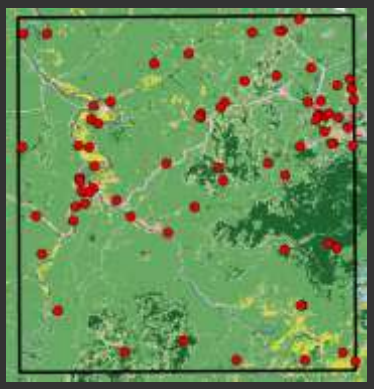
Different types of sampling methods are available. Here, I will focus on sampling methods used over map space.

When using simple random sampling, random locations are selected across a study area extent as random x, y coordinate pairs. So, this is purely random.

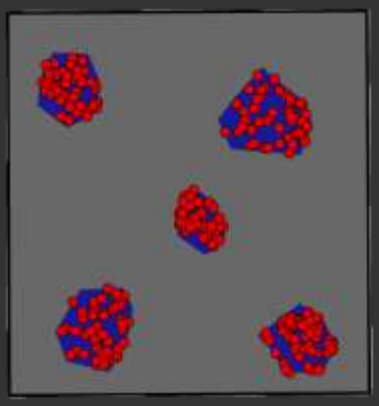
In contrast, systematic sampling involves sampling using a fixed pattern or spacing. For example, a sample could be collected every 50 meters.

Random or systematic sampling can also be performed along transects. This is a common method used by ecologists and foresters.

Types of Sampling



Stratified Random Sampling
(By land cover type)



Random Sampling
within Clusters

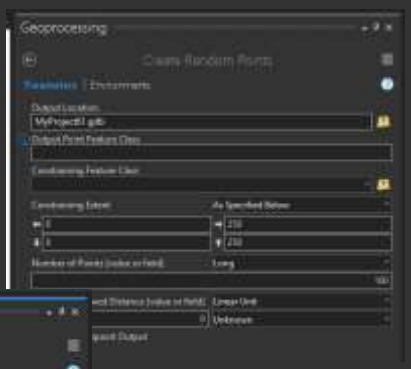
22

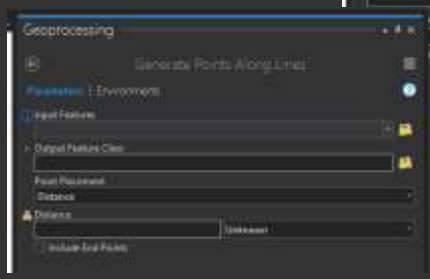
Stratified random sampling involves collecting a defined number of random samples per category. For example, a fixed number of samples could be collected for each land cover class.

Random sampling within clusters involves defining clusters or smaller extents within the larger extent of interest then selecting random samples within these extents. This method is commonly used when it is too difficult or infeasible to collect random samples across the entire extent of interest.

Tools for Sampling

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





23

Many tool are available in geospatial software packages to perform random sampling.

For example, ArcGIS Pro has a tool to generate random points and points along lines or transects.

Tessellation

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

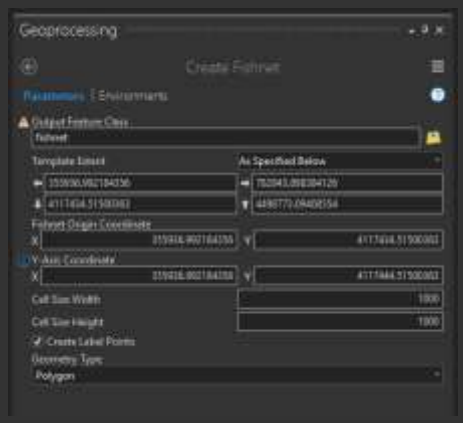
24

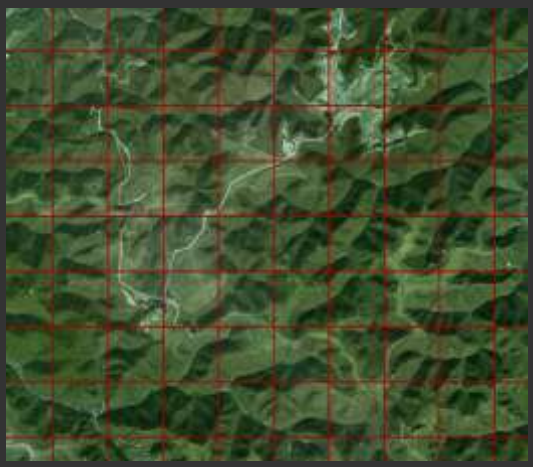
Tessellation involves filling a plane completely with non-overlapping geometric shapes. Commonly used shapes include hexagons, triangles, and squares.

These units can then be used to define sampling strategies or aggregate data.

Fishnets

- Creates a fishnet of regular cells





Imagery from ESRI Basemap

25

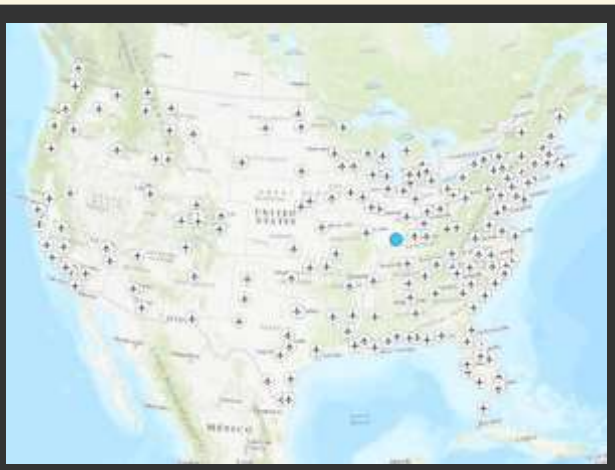
Fishnets can be used to generate grids over spatial extents. They can be useful in planning or defining sampling strategies.

I have also found them to be useful to aid in digitizing over large areas, as they can act as tiles to subset the task and as guides while digitizing.

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



27

We will now look at measures of spatial central tendency and variability. The summary statistics that we have already discussed relate to attribute information. Now, we are interested in measuring the spatial central tendency in the map space as opposed to the distribution of the attribute values.

One method to explore spatial central tendency is the central feature technique. This feature is defined as the feature in the dataset with the smallest accumulated distance from all other features.

Distance can be defined using the Euclidean or Manhattan methods. Euclidean distance is the straight-line distance between 2 points while Manhattan distance is the sum of the distance between points in the x and y directions, similar to walking city blocks. The Manhattan and Euclidean distance can be the same length if the two features are exactly north to south or east to west of each other. However, the Manhattan distance is generally longer.

This process will identify one of the input features as the central feature. You can also incorporate weights into the calculation. For example, for the airport example provided here you could include a weight as the number of passengers per year. You can also obtain different central features by category as defined by an attribute. For example, you could get different central features for international and regional airports.

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

28

As an alternative to central feature, you can also calculate the mean or median center.

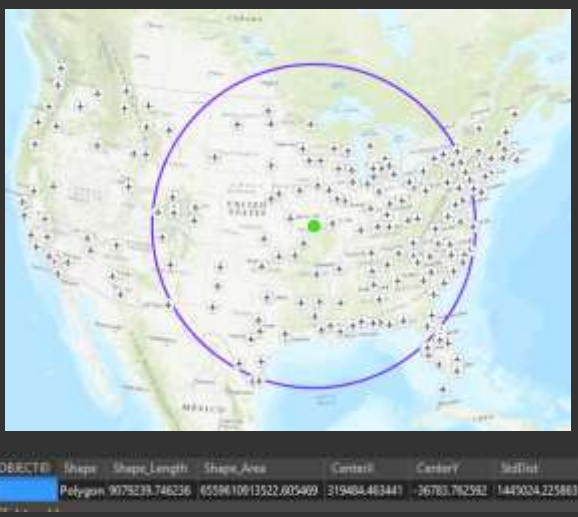
The mean center is the average x and y coordinate for the features while median center represents the median.

Note that this tool will return a new coordinate as opposed to one of the input features.

Similar to central feature, you can apply weights relative to an attribute or get different coordinates by category.

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	1449024.225863

29

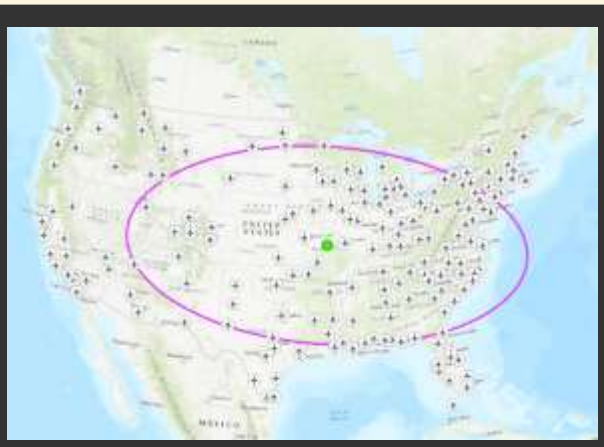
Central feature, mean center, and median center, provide a measure of spatial central tendency. They do not represent dispersion.

Standard distance can be used to obtain a measure of dispersion. It provides a measure of the degree to which features are concentrated or dispersed around the geometric mean center using standard deviation.

Smaller circles indicate less dispersion while larger circles indicate more dispersion. Standard distance for different datasets can be compared to get a sense of how they differ in regards to spatial dispersion.

Similar to the central tendency, weights can be applied and separate standard distances can be obtained by category or group.

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

30

Standard distance does not provide any measure of directional trends.

Directional distribution or the standard deviation ellipse summarizes the spatial central tendency, spatial dispersion, and directional trends.

The center of the ellipse will be the mean center, the size of the ellipse relates to spatial dispersion, and the long and short axis relate to directional trends.

Similar to the central tendency, weights can be applied and separate standard deviation ellipses can be obtained by category or group.

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)

32

We will now discuss statistical tests that can be used to address a spatial question.

First, we will review how to interpret statistical tests.

If you fail to reject the null hypothesis, this suggests there is no statistical significance or that the results were not significant. In contrast, if the null hypothesis is rejected, this suggests statistical difference or that there is statistical significance.

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.

Comparison to this distribution will allow for the calculation of 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

33

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

34

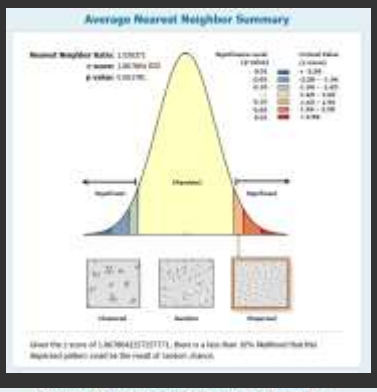
The first statistical method that we will explore is average nearest neighbor. This test is used to determine if a point pattern or distribution is random, clustered, or dispersed.

The null hypothesis is that the pattern is random while the alternative hypothesis is that the points are not random, or that they are clustered or dispersed.

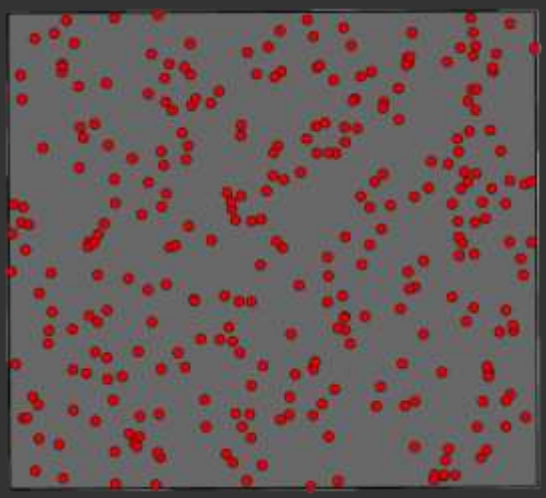
This is based on the calculation of the nearest neighbor ratio, which is based on a comparison of the actual point distribution to a hypothetical one. Distances in this test can be calculated using Euclidean distance or Manhattan distance.

Note that this test does not consider the actual attributes stored on the points. It only looks at the point pattern.

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



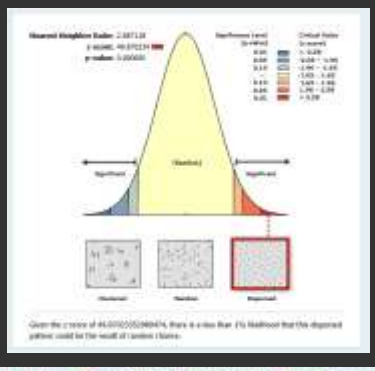
35

Here is an example output for average nearest neighbor for the pictured point distribution.

The nearest neighbor ratio is greater than 1 and the *p*-value is 0.062. This suggest slight dispersion. However, this is not statistically significant at the 95% confidence interval.

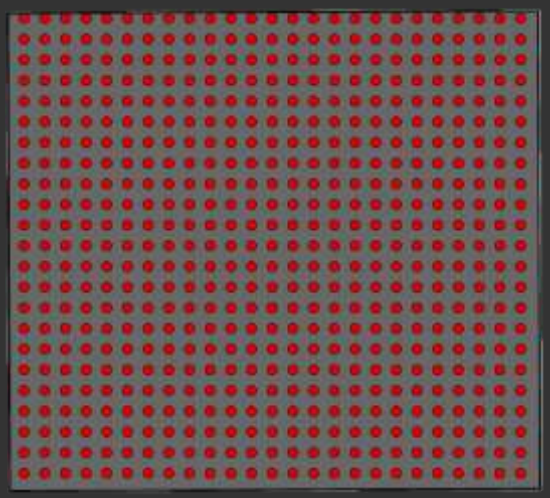
So, this would not be statistically significant, and the pattern would be considered to be random.

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

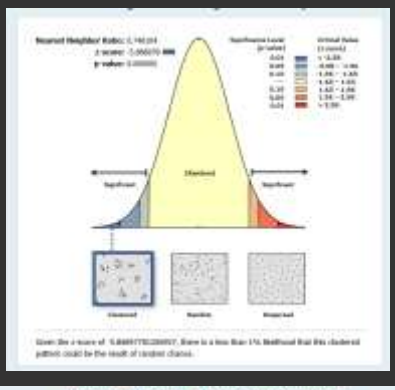


36

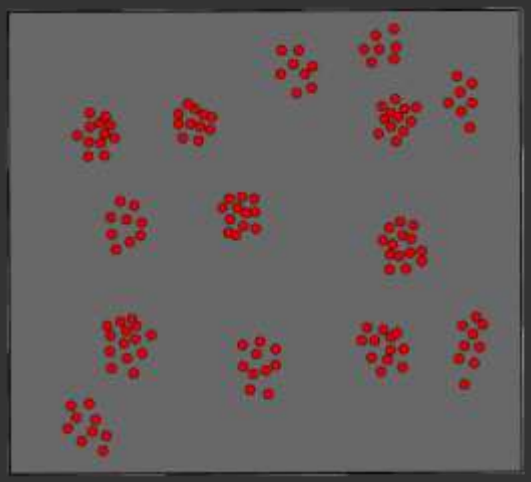
This is a perfectly dispersed pattern. The nearest neighbor ratio is greater than 1 and the *p*-value is 0, suggesting strong statistical significance.

So, this pattern is suggested to be dispersed.

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



37

For this result, the nearest neighbor ratio is less than 1 and the *p*-value is zero, suggesting strong statistical significance for a clustered pattern.

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

38

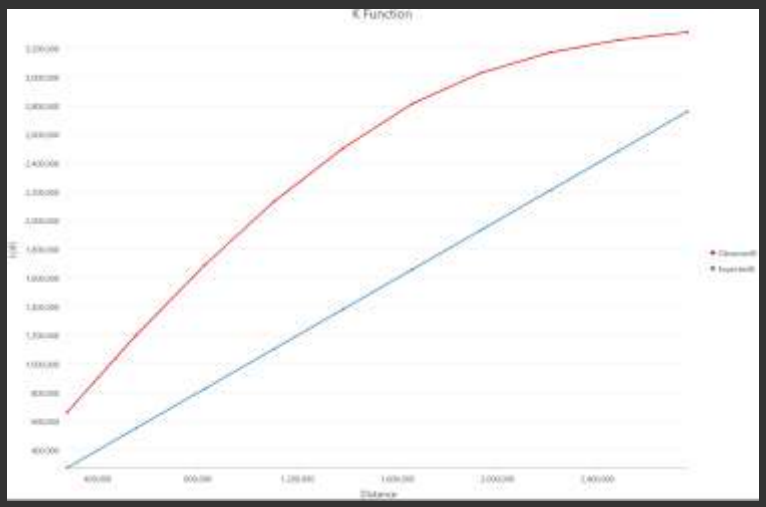
Average nearest neighbor offers a global assessment of the point pattern. However, different patterns could exist at different spatial scales.

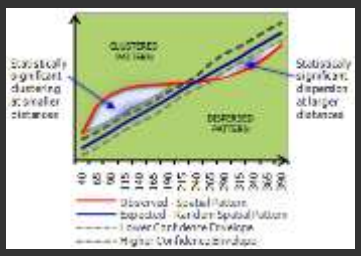
Ripley's K provides this comparison at different scales. This test will be performed at a variety of different distances.

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.

When the observed K value is smaller than the expected K value, the distribution is more dispersed than a random distribution at that distance.

Ripley's K





https://pro.arcgis.com/en/pro-app/3.5/tool-reference/spatial-statistics/multi-distance-spatial-cluster-analysis.htm

39

This slide shows an example output from a Ripley's K analysis (left) and a conceptualization from ESRI (right). Generally, the results suggest clustering across scales or distances.

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.

41

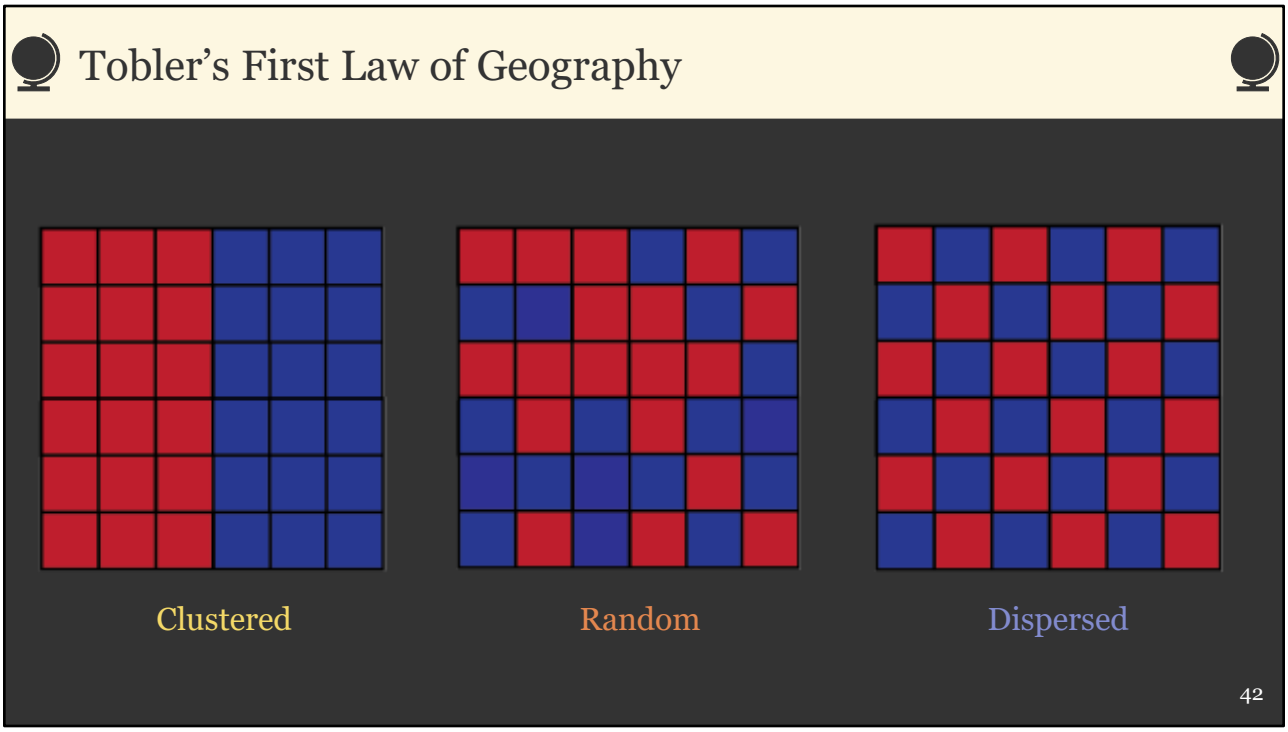
One key concept in geography is Tobler's First Law of Geography that states that everything is related to everything else, but near things are more related than distant things.

For example, you would expect individuals that have clustered into communities to have more similar beliefs and customs than individuals in different communities.

As a second example, you would expect soil samples that were collected close together to have more similar physical and chemical characteristics than samples that are separated by a greater distance.

This suggests that there are patterns across geographic space that can be explored and studied.

This is the concept of spatial autocorrelation.



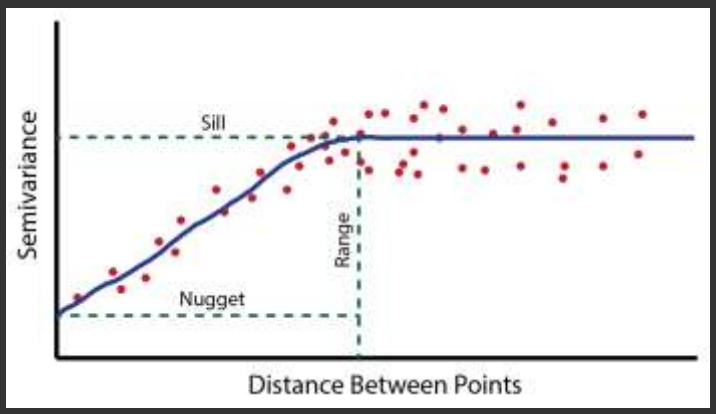
Clustering or positive spatial autocorrelation suggests that near features are more similar than features that are more separated. The first example provided here is an example of clustering since all the red cells are on one side and all the blue cells are on the other.

Dispersion or negative spatial autocorrelation suggests that more similar values are actually more separated over space. For example, the dispersed example above is dispersed because the different colors are as spread out as they can be.

Random implies no patterns across space.

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



43

A semivariogram is a graphical representation of the concept of spatial autocorrelation.

On a semivariogram, the x-axis represents distance between sample points, or lag. In the graph space, each point represents a pair of samples or observations. The x-coordinate is the lag or distance between the samples. The y-axis is the semivariance, which is a measure of the spatial dependency between two samples or observations. High semivariance suggests that the two observations or samples have very different values whereas low values of semivariance suggest they have very similar values.

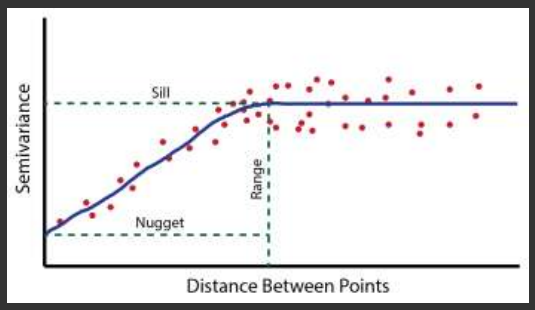
Based on Tobler's First Law of Geography, we would expect near objects to have more similar values and further objects to have less similar values. Or, near objects are expected to have low semivariance whereas samples that are more separated are expected to have high semivariance. This is the concept of spatial autocorrelation.

The semivariogram shown here represents simulated data. Real data are likely to show a more complex or noisy pattern.

The next slide will explain key features of the semivariogram.

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.



44

First, the distance at which the semivariance stops increasing is the range. This value can be determined by extrapolating from the fitted semivariance curve to the x-axis. Samples that are closer together then this distance are spatially autocorrelated. Samples that are further than this distance are not spatially autocorrelated.

The sill is the semivariance at the range: the semivariance at the distance where spatial autocorrelation is no longer present.

You would expect the semivariance at a distance or lag of zero to be zero. Or, you would expect samples at the same location to have the same value. However, this is not always the case for real data. This is known as the nugget effect. This arises due to measurement error or spatial sources of variation at distances smaller than the sampling unit.

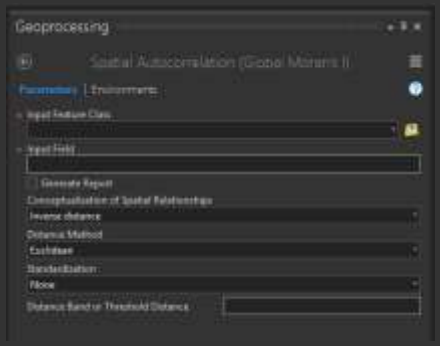
The partial sill is the semivariance obtained when the nugget is subtracted from the sill.

Again, the semivariogram can be thought of as a graphic representation of spatial autocorrelation. In this example, near features tend to have a lower semivariance, or are more correlated, whereas more distant features tend to have a greater semivariance, or are less correlated. At a certain distance, the

range, samples are no longer spatially autocorrelated and the semivariance stops increasing. So, near features have more similar values than observations that are more separated. In other words, spatial autocorrelation is present.

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



45

Moran's I provides a statistical assessment of global spatial autocorrelation. Note that this is different from average nearest neighbor because we are now assessing how the values associated with features are correlated based on distance, not simply the point pattern itself.

The null hypothesis for this test is that the pattern is random or that there is no correlation between the attribute of interest and distance between samples. In other words, there is no spatial autocorrelation.

If the test is statistically significant, then this suggests that there is spatial autocorrelation.

A positive Moran's I suggests clustering or positive spatial autocorrelation while a negative value suggests dispersion or negative spatial autocorrelation.

Distance in this test can be defined based on Euclidean or Manhattan distance. It also requires a conceptualization of spatial relationships, which will be explained in the next slide.

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 46

Since Moran's I requires that values associated with a feature be compared to values for nearby features, you must define what you mean by neighboring or near features.

Different methods are available, as described on this slide.

As an example, when using inverse distance nearby neighbors have a larger influence on the computations for a target feature than features that are further away.

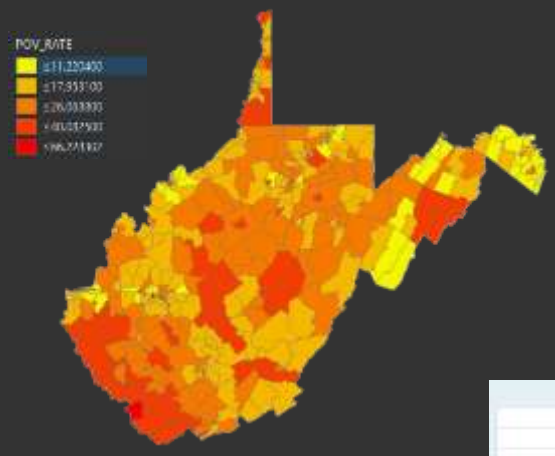
When a fixed distance band is used only features within a defined distance are considered.

For zone of indifference, features within a critical distance are assigned a weight of 1. The weight diminishes with distance past the critical distance.

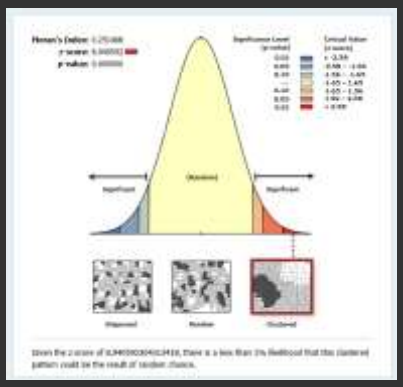
When contiguity edges only is used only features that share a boundary or overlap will be considered in the computation for the target feature.

There is not necessarily a correct method to use, as this is case specific. However, you should make and attempt to justify your decision based on the nature of the data and the question being investigated.

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:	0.000800
z-score:	8.940930
p-value:	0.000000

Used contiguity edges only

47

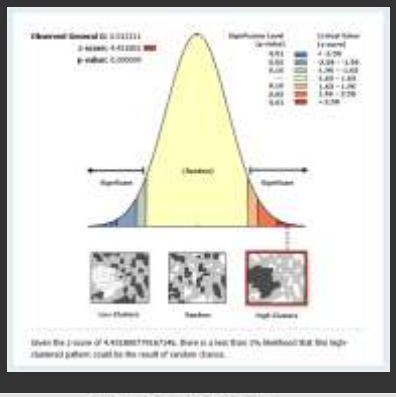
Here is an example result for Moran's I. Specifically, we are investigating spatial autocorrelation for percent poverty at the Census tract scale.

Here, a positive Moran's I of 0.25 is obtained and a *p*-value of 0. This indicates that there is clustering or positive spatial autocorrelation globally for percent poverty as summarized on Census tracts.

Near Census tracts tend to have more similar percent poverty rates.

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:	0.000008
z-score:	4.431801
p-value:	0.000009

Used contiguity edges only

48

Another test to explore or assess spatial autocorrelation is the Getis-Ord General G.

This test is similar to Moran's I; however, it specifically explores the clustering of high or low values.

The null hypothesis is that there is no clustering of high or low values, or that the pattern is random. The alternative hypothesis is that there is clusters of high or low values.

A positive General G suggests clustering of high values while a negative General G suggests clustering for low values.

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

50

The Moran's I and Getis-Ord General G tests assess global spatial autocorrelation. A single result is provided for all features or the entire map.

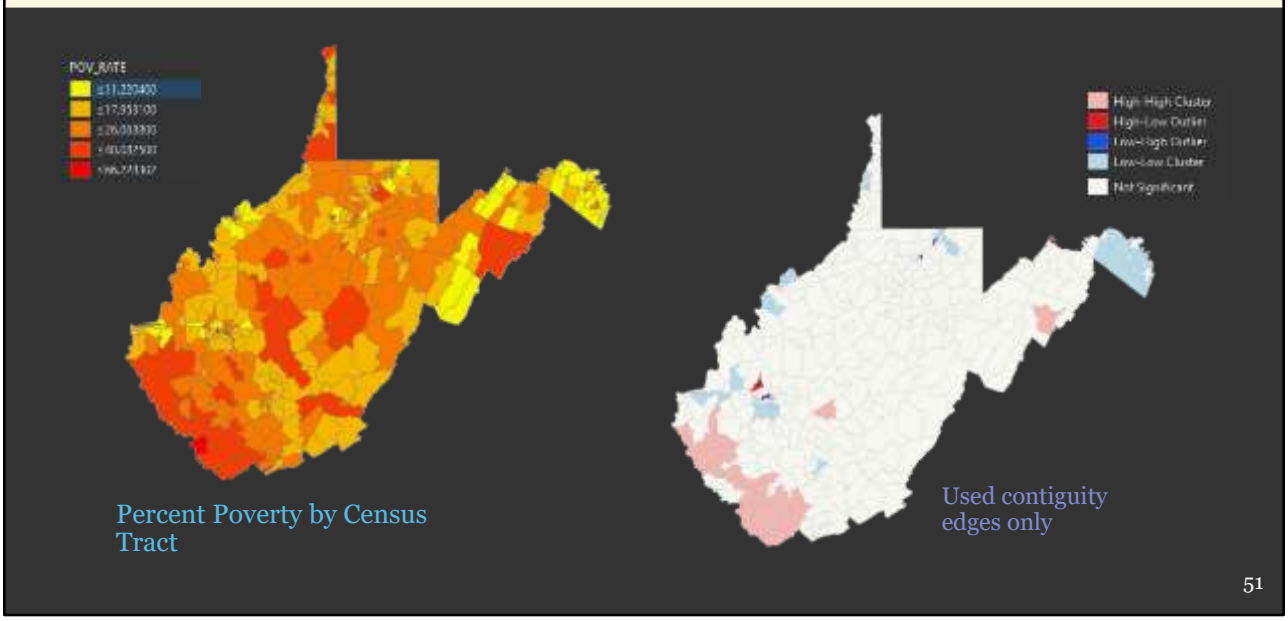
However, patterns could be investigated locally. In other words, a result could be obtained for each map feature.

Local Indicators of Spatial Autocorrelation, or LISA, provides a local version of Moran's I.

The null hypothesis is that there is no pattern or randomness while the alternative hypothesis is that there is local spatial autocorrelation, or that dispersion or clustering exists locally.

This test is able to differentiate local patterns as clusters of high values, clusters of low values, high values surround by low values, and low values surround by high values.

Local Indicators of Spatial Autocorrelation (LISA)



Here is an example result for percent poverty summarized by Census tract.

Light red indicates a high value surround by other high values while dark red indicates high values surround by low values.

Light blue indicates low values surround by other low values while dark blue indicated low values surround by other low values.

White indicates no pattern or no statistical significance.

Again, this is a local measure, so a result is provided for each map feature individually as opposed to a global measure.

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

52

The Getis-Ord Gi* provides a local version of the Getis-Ord General G. This test is also known as hot spot analysis.

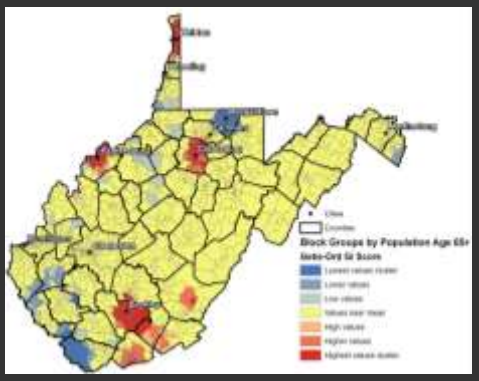
It is used to find local clusters of high values, called hot spots, or local clusters of low values, called cold spots.

The null hypothesis is that there are no local clusters of high or low values while the alternative is that there are hot and/or cold spots.

The provided example represents an assessment of population change between 2000 and 2010. Red indicates hot spots, or clusters of counties that have seen population growth during this time period. Blue areas indicate clusters of counties that have seen population decreases during this time period. Tan indicates no pattern.

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



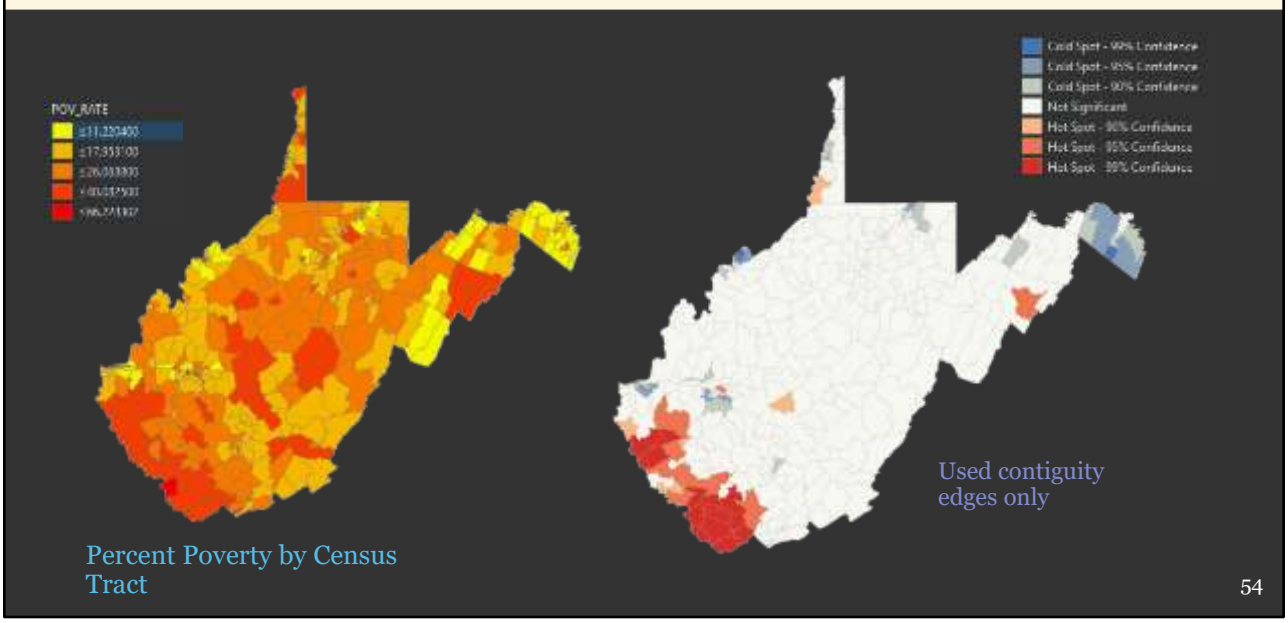
Example provided by
Mike and Jackie Strager

53

The example provided here shows results for a hot spot analysis by Census block group for the number of people over 65.

Red areas indicate clusters of block groups with a large number of people over 65, or hot spots. Blue indicates clusters of counties with lower numbers of people over 65, and tan indicates no pattern.

Hot Spot Analysis (Getis-Ord Gi*)



As another example, this slide represents results for a hot spot analysis of percent poverty at the Census tract scale.

Red indicates clusters of high poverty while blue indicates clusters of low poverty. White suggests no pattern.

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

56

We will end this module with a brief introduction to regression and geographically weighted regression.

Regression techniques can be used to explore the relationship or correlation between variables and also predict a continuous variable using other variables.

There are forms of regression designed for probabilistic or classification problems, like logistic regression, but we will not discuss those here.

Regression can be used to predict a dependent variable using one or more independent variables.

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)

57

When using single regression, a single independent variable is used to predict a dependent variable.

The result will be a linear equation that relates the two variables. The equation will include a slope term, or coefficient, and a y-intercept.

$\hat{y}$ is the predicted y value whereas y represents the correct y value.

The difference between y and $\hat{y}$ is the residual or error term. So, y is equal to the linear equation or model plus the error term.

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

58

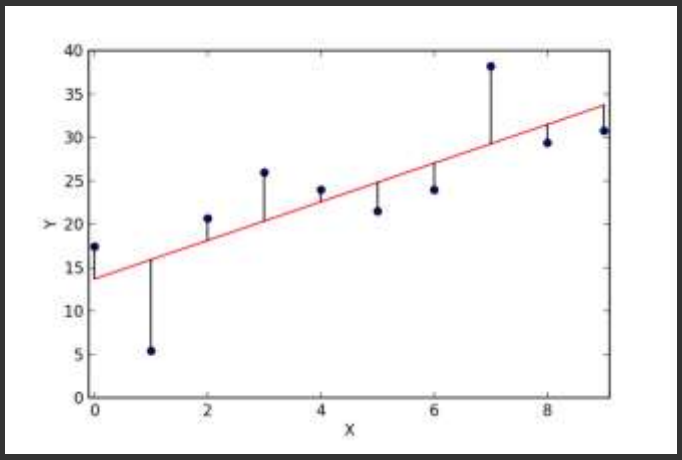
It is also possible to predict the dependent variable using more than one independent variable.

This is known as multiple regression.

In this case a linear equation will be produced that includes a y-intercept and coefficients for all x variables.

β_0 is the y intercept while β_n represents the coefficient or slope for the n^{th} x variable.

Residuals



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

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

59

Again, the residual, or error term, represents the difference between y and the predicted value for y, or ŷ.

In the example graphic, the linear model is represented by the red line. The residual for each point would be the distance from the line to the sample point along the y-axis.

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

60

As already mentioned in previous modules, continuous predictions can be assessed using root mean square error or RMSE.

MSE, or Mean Square Error, is simply RMSE without the square root applied. RMSE will be reported in the units of the variable being predicted while MSE will be in the square of the units.

For example, if you are predicting chemical concentration in parts per million, then RMSE will be reported in parts per million and MSE will be reported in parts per million squared.

RMSE is generally calculated by comparing the predicted value to the value provided by the validation data.

The values are subtracted to obtain a residual or error. The residuals are then squared, summed, and divided by the number of samples. This will provide MSE. The square root is taken to obtain RMSE.

Lower RMSE and MSE suggests a better model.

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

61

Another means by which to assess continuous predictions is R^2. This value represents the proportion of variance in the quantity being predicted explained by the model.

It is scaled from 0 to 1. 0 indicates that no variance is explained, or this is a poor model, while 1 indicates that all variance is explained. So, higher values are better.

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

62

When more than one predictor variable is used, R^2 must be modified to obtain adjusted R^2 to deal with inflation when multiple predictor variables are included.

The equation requires altering R^2 based on the sample size and number of predictor variables.

Note that this is not required when calculating R^2 using withheld data.

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!

63

This slide provides some additional information relating to R^2.

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

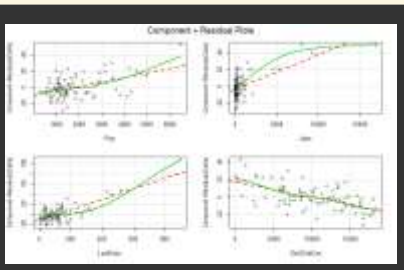
https://www.youtube.com/watch?v=KjRrdb2x6dA

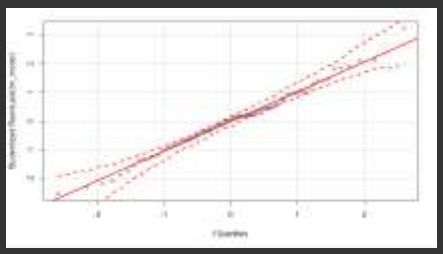
64

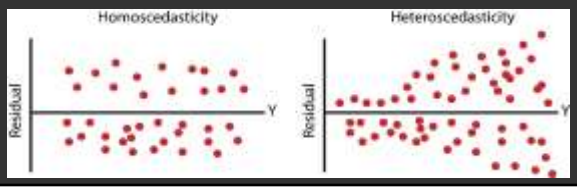
This slide provides some additional information relating to adjusted R^2.

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







65

Linear regression is a parametric method that has assumptions. If these assumptions are not met, then the result may be misleading or invalid.

When performing linear regression, it is assumed that there is a linear relationship between the variables. If the relationship is not linear then linear regression is not valid. In other words, the variables may need to be transformed so that the relationship becomes more linear in the transformed space.

The residuals should be normally distributed. This can be assessed using the QQ plot as described above.

There should also not be a lot of correlation between the multiple independent variables in multiple regression. If they are highly correlated, this is called multicollinearity. Although some correlation can be tolerated, this makes it difficult to interpret the coefficients. This can be assessed using the Variance Inflation Factor, or VIF.

It is also assumed that the samples are independent of each other. If spatial autocorrelation exists, then it cannot be assumed that the samples are independent. This can be assessed using Moran's I. This is a common issue when performing regression analysis for spatial data. If spatial autocorrelation

exists, geographically weighted regression may be more appropriate.

Lastly, the model should be able to estimate similar values of y with similar error or residual terms. In other words, the variance in the residuals should not vary with values of y. This is known as homoscedasticity. If the variance of the residuals does vary with values of y, this is known as heteroscedasticity.

Interpreting Coefficients

Single Linear Regression

- y-intercept often meaningless
- Amount of change in y with one unit change in x
- + = y increases as x increases
- - = y decreases as x increases
- Near 0 = week relationship
- Values of coefficients will depend on units of measurement for x and y
- Can test for statistical significance
- Can be difficult to interpret if variables are transformed

Multiple Linear Regression

- y-intercept often meaningless
- Average effect of specific predictor variable on y when all other predictor variables are held fixed
- Coefficients are impacted by collinearity
- Must also consider interaction terms
- Can test for statistical significance
- Can be difficult to interpret if variables are transformed

66

Interpreting coefficients for single linear regression is generally simpler than interpreting them for multiple linear regression. For both single and multiple linear regression, the y-intercept can be thought of as an adjustment or offset factor; however, it is not generally interpretable or meaningful.

In single linear regression, the coefficient represents the estimated change in y with one unit change in x. For example, in the single linear regression model discussed above in which fuel efficiency was predicted using weight, a coefficient of roughly -6.6 was obtained for the weight predictor variable. This suggests that for a unit change in weight, you would expect a 6.6 miles per gallon decrease in fuel efficiency. Note again that this coefficient would change if the units of y and/or x changed.

Positive coefficients for the predictor variable in single linear regression suggests a positive relationship: as x increases, y increases. In contrast, a negative coefficient suggests a negative relationship: as x goes up, y goes down. In the weight example, a negative coefficient was obtained, which suggests that fuel efficiency decreases with vehicle weight, as expected. If the coefficient is near zero, this generally indicates that there is no relationship between y and the predictor variable.

Note that it is possible to determine if the coefficient is statistically significant,

or if there is a statistically significant correlation between x and y. This is represented by the p-value included in the example output tables presented above.

Again, coefficients will change if the unit of measurement change or if the variable(s) are transformed in some way, such as using a log transformation.

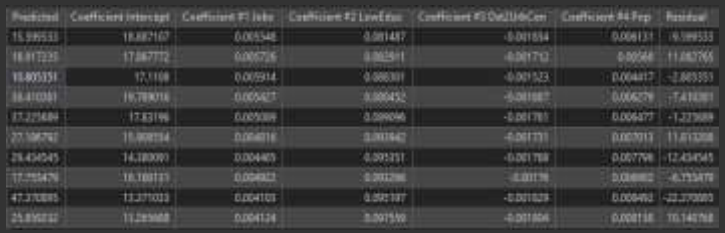
For multiple linear regression, the coefficients are generally interpreted as the average effect of a specific predictor variable on y when all other predictors are held fixed. If there is no collinearity, or relationship between predictor variables, then the coefficients in single and multiple linear regression for the same variable will be the same. However, multicollinearity will impact the coefficients. This means that interpretation of the coefficients, or relationships between each predictor variable and y, is complicated when predictor are correlated with each other, which is a common occurrence.

The key component of understanding coefficients in linear regression is the phrase “when all other predictor variables are held fixed.” For example, if two data points have all the same measures for each predictor variable but one, you would expect the difference between the predictions for y for these two data points to be summarized by the coefficient for the variable that is different. However, if multiple measures for multiple predictor variables change, this simple interpretation is not valid. In short, you should be careful when interpreting coefficients in multiple regression equations.

Another issue is that interactions may be present in which the impact of one predictor variable on y depends on another predictor variable included in the model. Note that it is possible to include interaction terms directly in the model.

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



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

67

What if the relationship between the dependent and independent variables vary over space? Or, what if there is spatial heterogeneity in relationships? In this case, it may not be appropriate to generate one regression equation for all features or the entire mapped extent.

Instead, it might be more appropriate to allow the equation to vary over space to capture the changing relationships. This is the concept of geographically weighted regression or GWR. Note that there are different forms of GWR.

When using GWR, the regression equation will vary across the map space or for each mapped feature. Specifically, different coefficients will be calculated, as demonstrated in the provided table.

This is a powerful technique for improving the suitability of linear regression for making predictions over map space where relationships can vary.

Geographically Weighted Regression (GWR)

Y Intercept

Coefficient for Distance from Urban Center

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

68

As highlighted in the provided examples, the coefficients vary over the map space when using GWR as opposed to the same coefficients or a single equation across the map space.

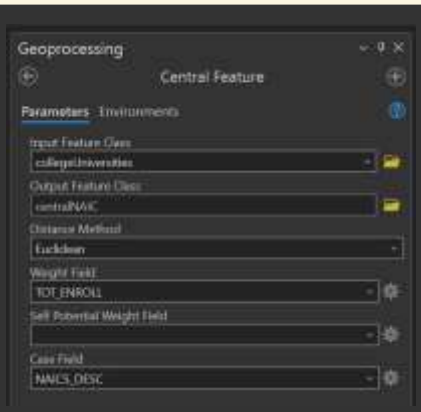
Geostatistical Analysis In ArcGIS Pro

69

The following exercises relate to using ArcGIS Pro for spatial statistics.

Spatial Central Tendency

1. Use the spatialStats map.
2. Use the Central Feature Tool to find the most central feature using the following settings:
   - **Input Feature Class** = collegeUniversities
   - **Distance Method** = *Euclidean*
   - **Weight Field** = "TOT_ENROLL"
   - **Class Field** = "NAICS_DESC"



What is the most central feature for the "COLEGES, UNIVERSITIES, AND PROFESSIONAL SCHOOLS" NAIC group?

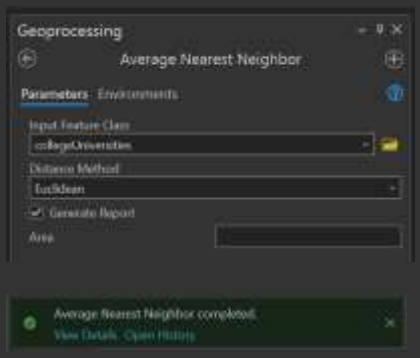
Spatial Statistics Tools →
Measuring Geographic Distributions →
Central Feature

70

Average Nearest Neighbor

1. Use the spatialStats map.
2. Use the Average Nearest Neighbor Tool to explore whether the distribution of the collegeUniversities points is clustered, random, or dispersed. Use the following settings:
   - **Input Feature Class** = collegeUniversities
   - **Distance Method** = *Euclidean*
   - Click the Generate Report radio button.
   - Do not set an **Area**.
3. Open the report by selecting it in the View Details Window for the tool execution.

Does the statistical test suggest that the pattern is clustered, random, or dispersed?



Moran's I

1. Use the spatialStats map.
2. Use the Spatial Autocorrelation (Global Moran's I) Tool to explore whether population density by county is clustered, random or dispersed using the following settings:
   - **Input Feature Class** = usDataAgg
   - **Input Field** = "POP_SQM"
   - Click the Generate Report radio button.
   - Conceptualization of Spatial Relationships = *Inverse distance*
   - Use the default for all other settings
3. Open the report by selecting it in the View Details Window for the tool execution.

Does the statistical test suggest that population density by county is clustered, random, or dispersed?

Hot Spot Analysis

1. Use the spatialStats map.
2. Use the Hot Spot Analysis (Getis-Ord Gi*) Tool to explore spatial clusters of median income.
3. **Input Feature Class** = usDataAgg
   - **Input Field** = "med_ncm"
   - Conceptualization of Spatial Relationships = *Inverse distance*
   - Use the default for all other settings

Is this the pattern you expected?

Spatial Statistics Tools →
Mapping Clusters →
Hot Spot Analysis (Getis-Ord Gi*)

This is the end of this lecture module.

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

Thanks! Hope you found this useful.