Geographic Information System

Spatial Statistics I

Dr. Chan, Chun-Hsiang
Department of Geography
National Taiwan Normal University

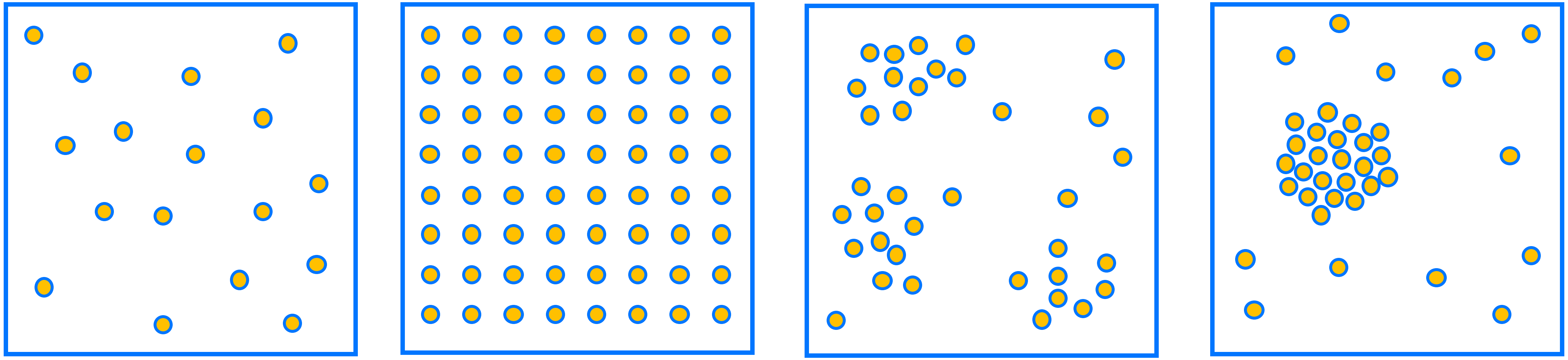
Photo credit: ChatGPT

Outline

- Spatial Data Distribution
- Mean Center
- Median Center
- Standard Distance
- Central Feature
- Directional Distribution
- Central Tendency Problems
- What's a z-score? What's a *p*-value?
- Average Nearest Neighbor
- High/Low Clustering (Getis-Ord General G)
- Spatial Autocorrelation (Global Moran's I)
- Incremental Spatial Autocorrelation
- Repley's *k*-function
- Lab#01

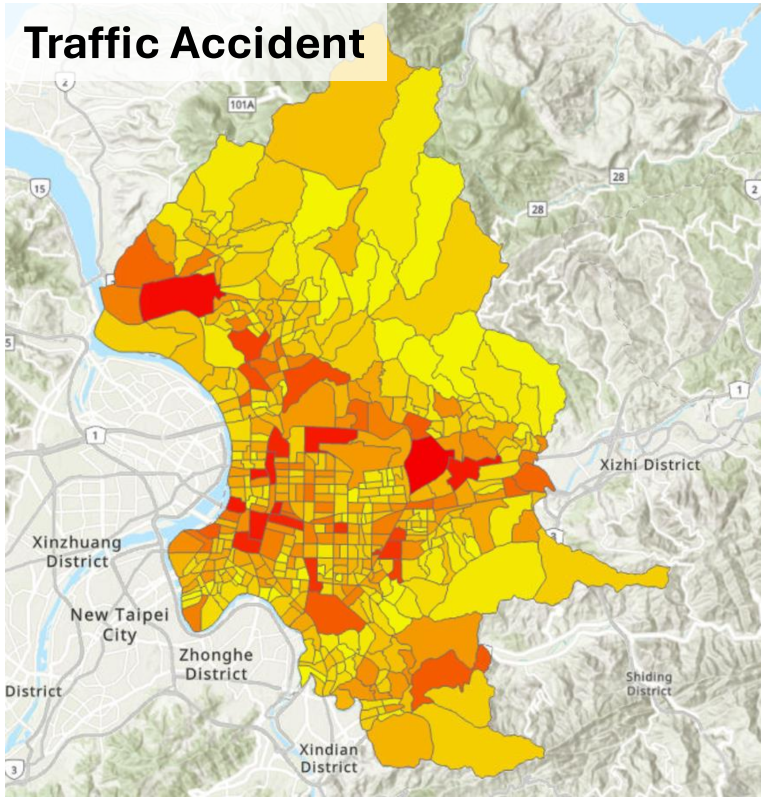
Spatial Data Distribution

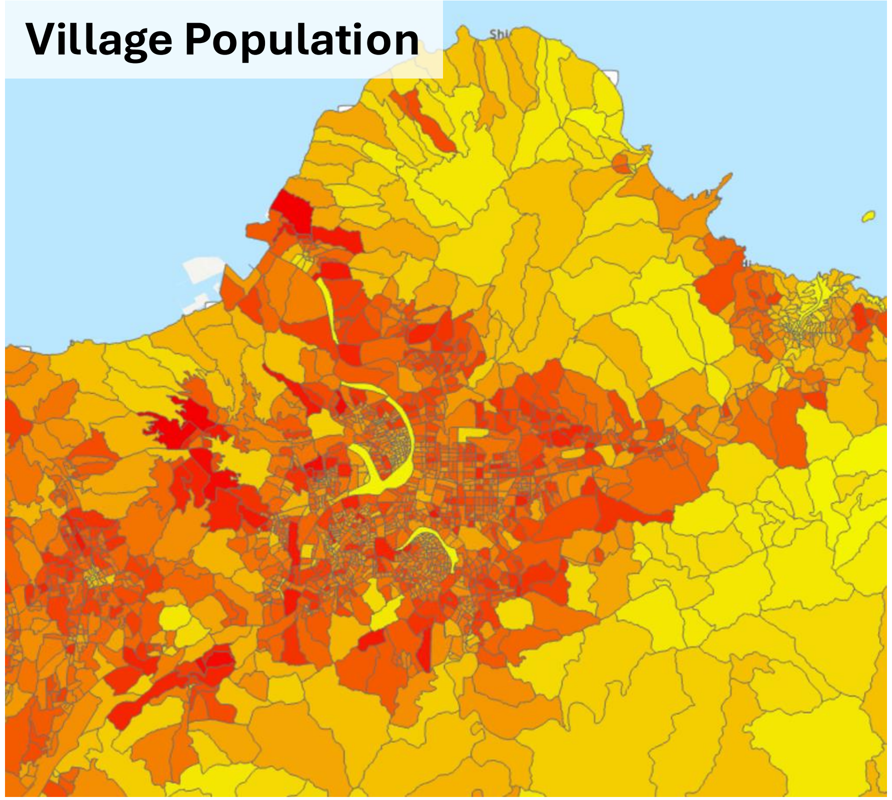
- Thinking about a situation! When you have to observe and report the spatial data distribution, how do you describe the following three spatial data patterns?



Spatial Data Distribution

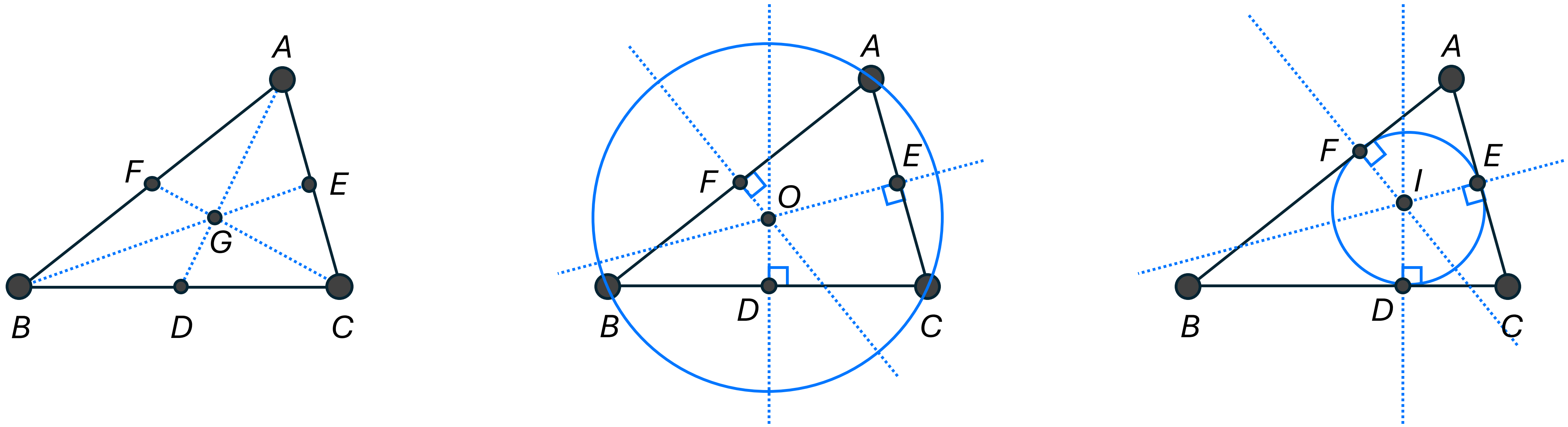
- But, in reality, … how do you define … the following patterns?





Central Tendency Measurement

- From a geometric perspective, we may adopt various types of methods to quantify the center of a triangle ...

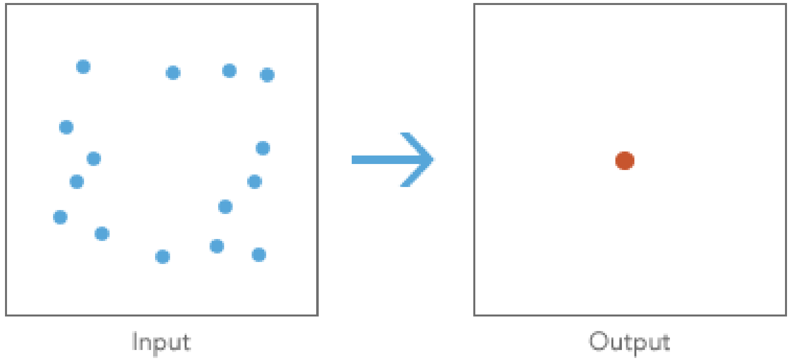


Central Tendency Measurement

- Central tendency measurement quantifies the central point or feature of a set of points or features from a geometric perspective.
- In ArcGIS Pro, we may conduct some functions to describe the characteristics of spatial distribution …

Functions	Definition
Mean center	Identifies the geographic center (or the center of concentration) for a set of features.
Median center	Identifies the location that minimizes overall Euclidean distance to the features in a dataset.
Standard distance	Measures the degree to which features are concentrated or dispersed around the geometric mean center.
Central feature	Identifies the most centrally located feature in a point, line, or polygon feature class.
Directional distribution	Creates standard deviational ellipses or ellipsoids to summarize the spatial characteristics of geographic features: central tendency, dispersion, and directional trends.

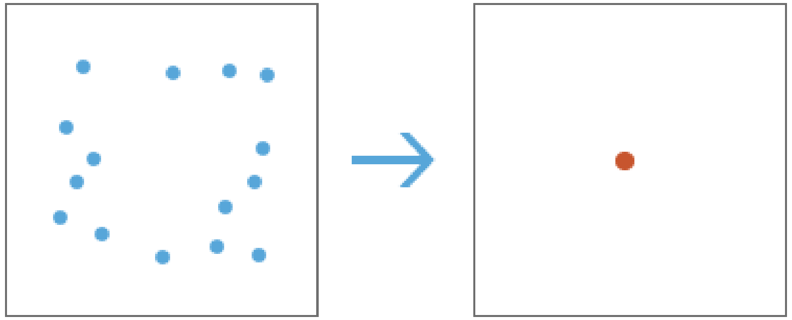
Mean Center



- Identifies the geographic center (or the center of concentration) for a set of features.
- The mean center is a point constructed from the average x, y and if available, z values for the input feature centroids.
- This tool requires **projected data** to accurately measure distances.
- The Case Field is used to group features for separate mean center computations. When a Case Field is specified, the input features are first grouped according to case field values, and then a mean center is calculated for each group.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/mean-center.htm

Mean Center



- The mean center is given as:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}, \bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

where x_i and y_i are the coordinate for feature i, and n is equal to the total number of features.

- The weighted mean center extends to the following:

$$\bar{X}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \bar{Y}_w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

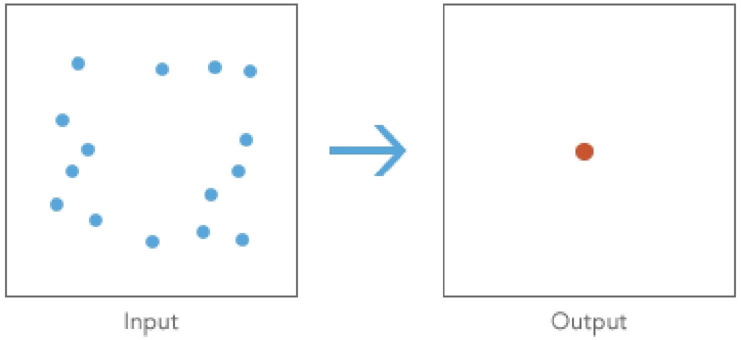
where w_i is the weight at feature i.

- The tool also calculates the center for a 3rd dimension if a z attribute exists for each feature:

$$\bar{Z} = \frac{\sum_{i=1}^{n} z_i}{n}, \bar{Z}_w = \frac{\sum_{i=1}^{n} w_i z_i}{\sum_{i=1}^{n} w_i}$$

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/mean-center.htm

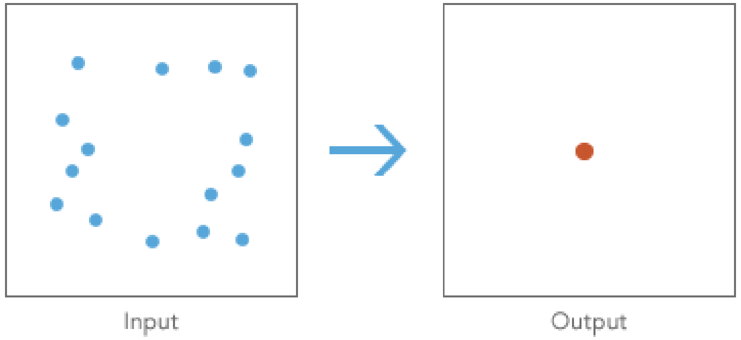
Median Center



- Identifies the location that minimizes overall Euclidean distance to the features in a dataset.
- While the Mean Center tool returns a point at the average X, average Y and, if applicable, average Z coordinate for all feature centroids, the Median Center tool uses an iterative algorithm to find the point that minimizes Euclidean distance to all features in the dataset.
- Both the Mean Center and Median Center are measures of central tendency. The algorithm for the Median Center tool is less influenced by data outliers (**why?**).
- This tool requires **projected data** to accurately measure distances.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/median-center.htm

Median Center



- The Case Field is used to group features for separate median center computations. When a case field is specified, the input features are first grouped according to case field values, and then a median center is calculated for each group.
- The median center is given as:

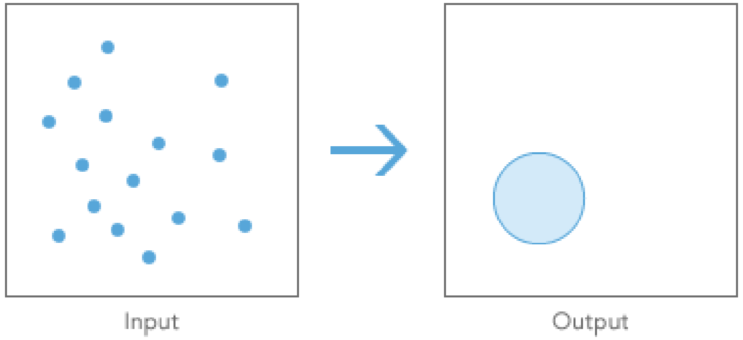
$$d_i^t = \sqrt{(x_i - x^t)^2 + (y_i - y^t)^2 + (z_i - z^t)^2}$$

where x_i, y_i, z_i are the coordinates for feature i at step t.

- At each step (t) in the algorithm, a candidate Median Center is found (x_t, y_t) and then refined until it represents the location that minimizes the Euclidean Distance d to all features (or all weighted features) (i) in the dataset.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/median-center.htm

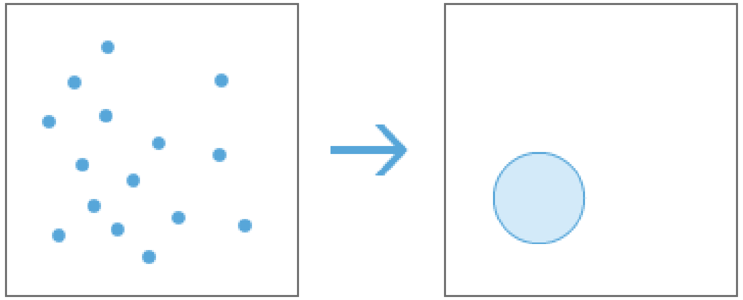
Standard Distance



- Measures the degree to which features are concentrated or dispersed around the geometric mean center.
- The standard distance is a useful statistic as it provides a single summary measure of feature distribution around their center (similar to the way a standard deviation measures the distribution of data values around the statistical mean).
- This tool requires **projected data** to accurately measure distances.
- The Case Field parameter is used to group features prior to analysis. When a Case Field is specified, the input features are first grouped according to case field values. Then a standard distance circle is computed for each group.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/standard-distance.htm

Standard Distance



- The standard distance is given as:

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n} + \frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{n} + \frac{\sum_{i=1}^{n}(z_i - \bar{Z})^2}{n}}$$

where x_i, y_i, and z_i are the coordinates for feature i, $\{\bar{X}, \bar{Y}, \bar{Z}\}$ represents the Mean Center for the deatures, and n is equal to the total number of features.

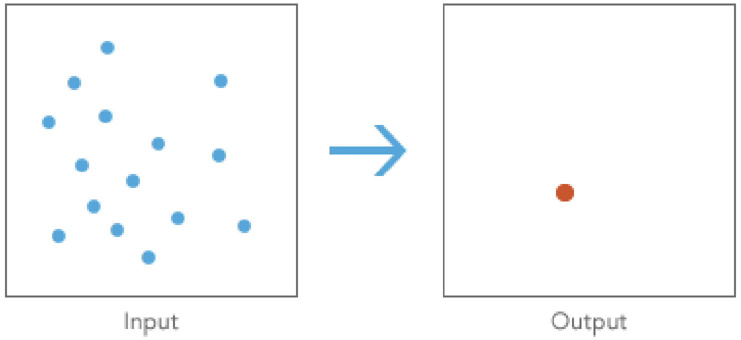
- The weighted standard distance extends to the following:

$$SD_w = \sqrt{\frac{\sum_{i=1}^{n} w_i(x_i - \bar{X}_w)^2}{\sum_{i=1}^{n} w_i} + \frac{\sum_{i=1}^{n} w_i(y_i - \bar{Y}_w)^2}{\sum_{i=1}^{n} w_i} + \frac{\sum_{i=1}^{n} w_i(z_i - \bar{Z}_w)^2}{\sum_{i=1}^{n} w_i}}$$

where w_i is the weight at feature i and represents the weighted Mean Center.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/standard-distance.htm

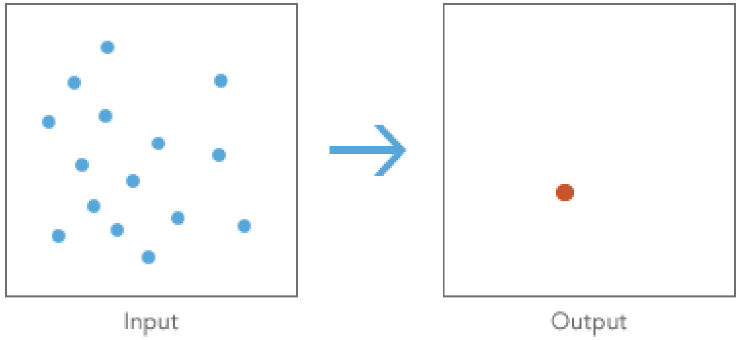
Central Feature



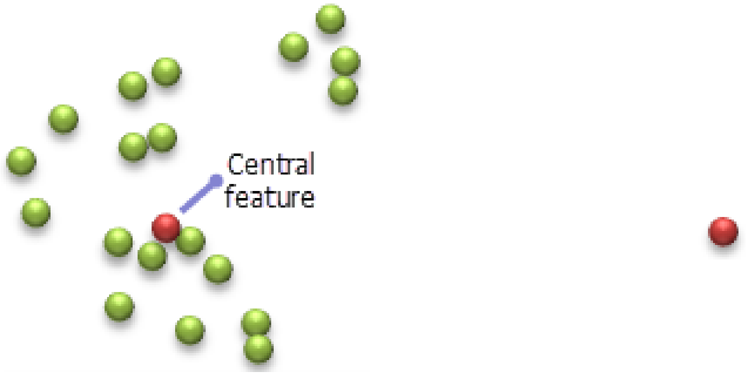
- Identifies the most centrally located feature in a point, line, or polygon feature class.
- Accumulated distances are measured using Euclidean distance or Manhattan distance , as specified by the Distance Method parameter.
- For line and polygon features, feature centroids are used in distance computations.
- For multipoints, polylines, or polygons with multiple parts, the centroid is computed using the weighted mean center of all feature parts.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/central-feature.htm

Central Feature



- The weighting for point features is 1, for line features is length, and for polygon features is area.
- The Case Field is used to group features for separate Central Feature computations.

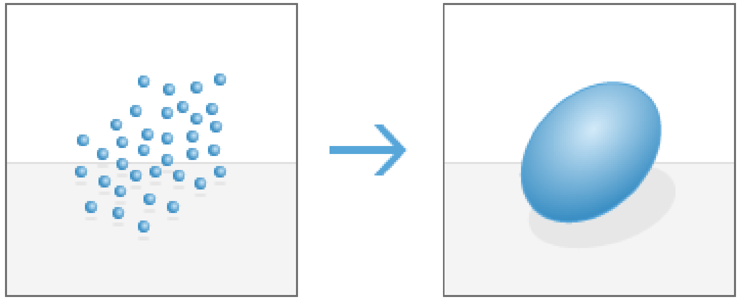


OBJECTID	Shape	Name
1	Point	A
2	Point	B
3	Point	C
4	Point	D
5	Point	E
6	Point	F
7	Point	G
8	Point	H
9	Point	I
10	Point	J
11	Point	K
12	Point	L

OBJECTID	Shape	Name
1	Point	G

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/central-feature.htm

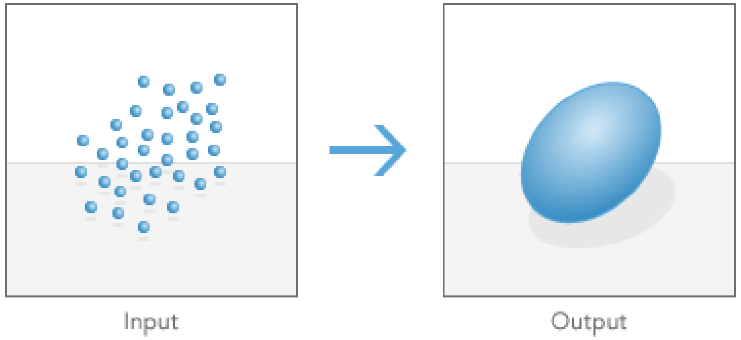
Directional Distribution

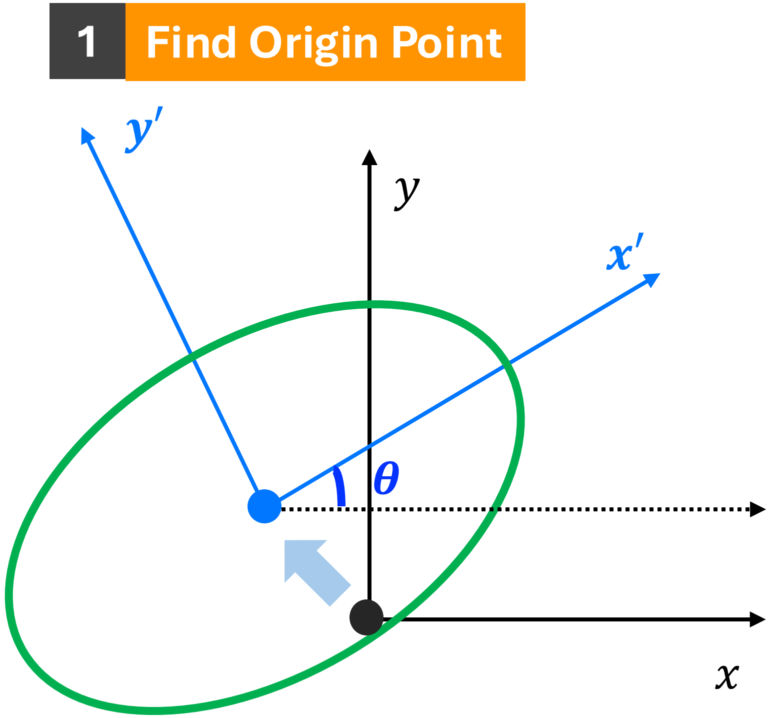


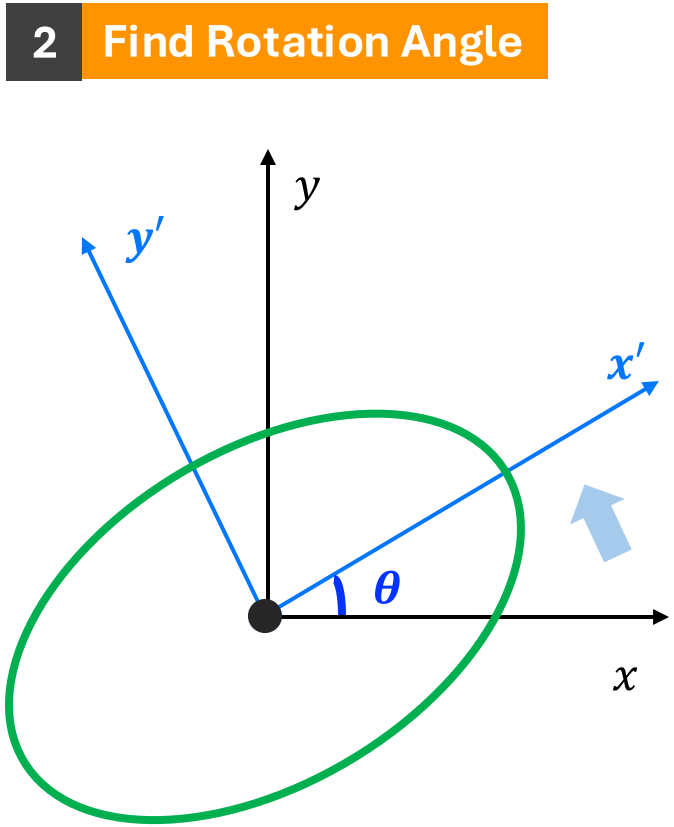
- Creates standard deviational ellipses or ellipsoids to summarize the spatial characteristics of geographic features: central tendency, dispersion, and directional trends.
- The Standard Deviational Ellipse tool creates a new Output Ellipse Feature Class containing elliptical polygons or 3D ellipsoidal multipatches, one for each case if the Case Field parameter is used. The attribute values for these elliptical polygons include x and y coordinates for the mean center, two standard distances (long and short axes), and the orientation of the ellipse.
- Calculations require projected data to accurately measure distances.

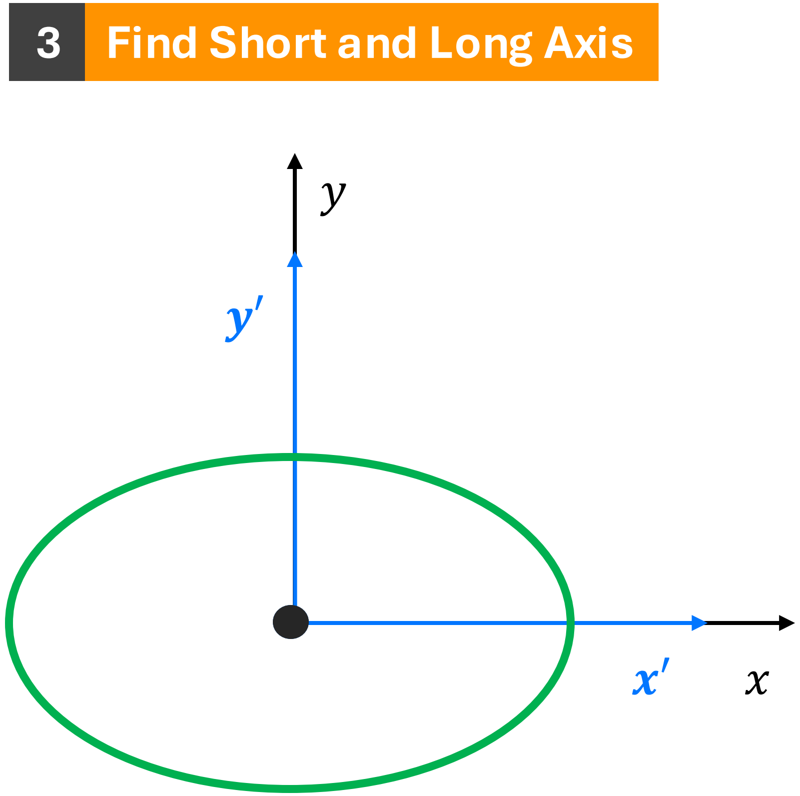
Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/directional-distribution.htm

Directional Distribution

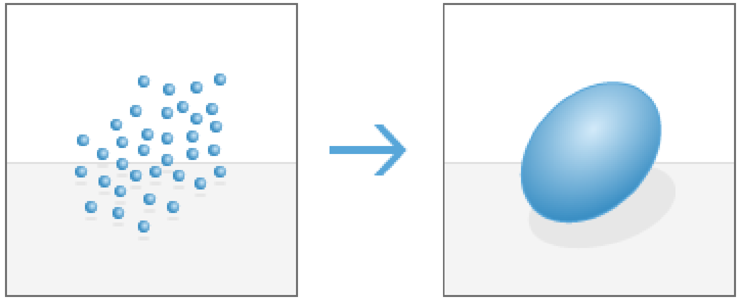








Directional Distribution



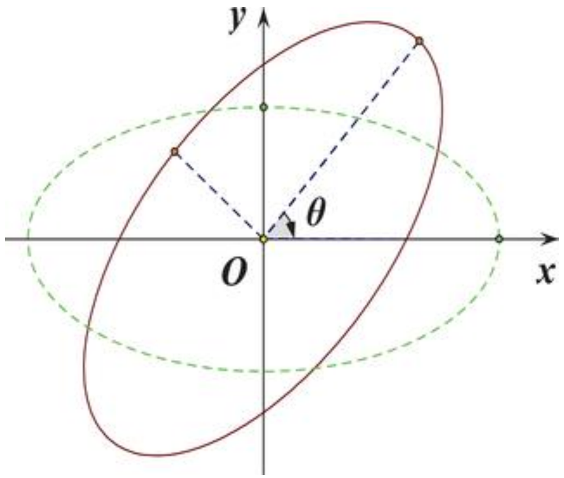
1 Find Origin Point

- The Standard Deviational Ellipse is given as:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,; \begin{pmatrix} \tilde{x}_i \\ \tilde{y}_i \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}$$

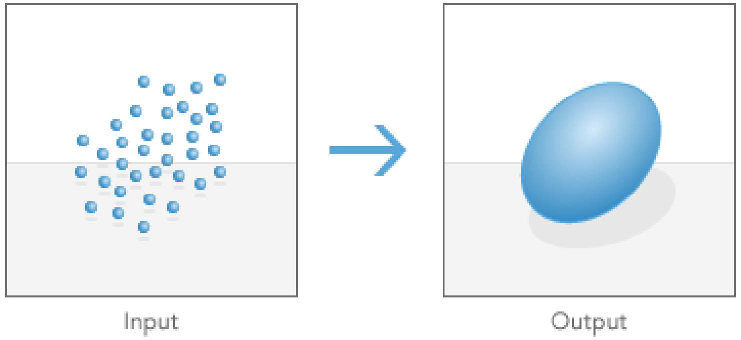
2 Find Rotation Angle

- Next, we introduce a rotation matrix $G = \begin{pmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{pmatrix}$ with an angle θ in clockwise direction.
- $\begin{pmatrix} x_i{}' \\ y_i{}' \end{pmatrix} = G\begin{pmatrix} \tilde{x}_i \\ \tilde{y}_i \end{pmatrix} = \begin{pmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{pmatrix}\begin{pmatrix} \tilde{x}_i \\ \tilde{y}_i \end{pmatrix}$
- $= \begin{pmatrix} \tilde{y}_i sin\theta + \tilde{x}_i cos\theta \\ \tilde{y}_i cos\theta - \tilde{x}_i sin\theta \end{pmatrix}$



Source: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118537

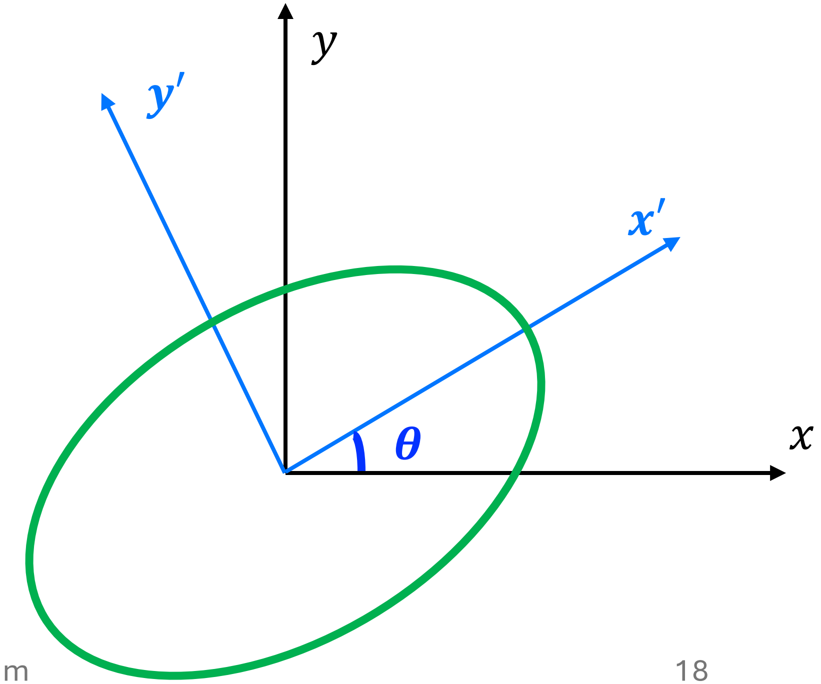
Directional Distribution



2 Find Rotation Angle **3 Find Short and Long Axis**

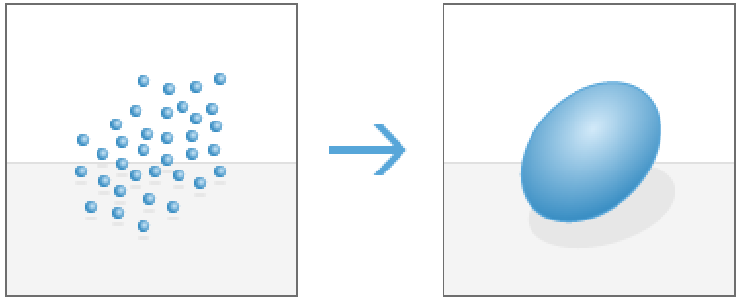
- The maximum likelihood estimator of the rotated samples' variance yields,

$$\begin{cases} \sigma_{x'}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i{}')^2 = \frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i cos\theta + \tilde{x}_i sin\theta)^2 \\ \sigma_{y'}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i{}')^2 = \frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i cos\theta - \tilde{x}_i sin\theta)^2 \end{cases}$$



Source: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118537

Directional Distribution



2 Find Rotation Angle

- Consequently, corresponding angles for producing the maximum and minimum standard deviations can be obtained by equating any derivative of the above variance estimators w.r.t. θ to be zero

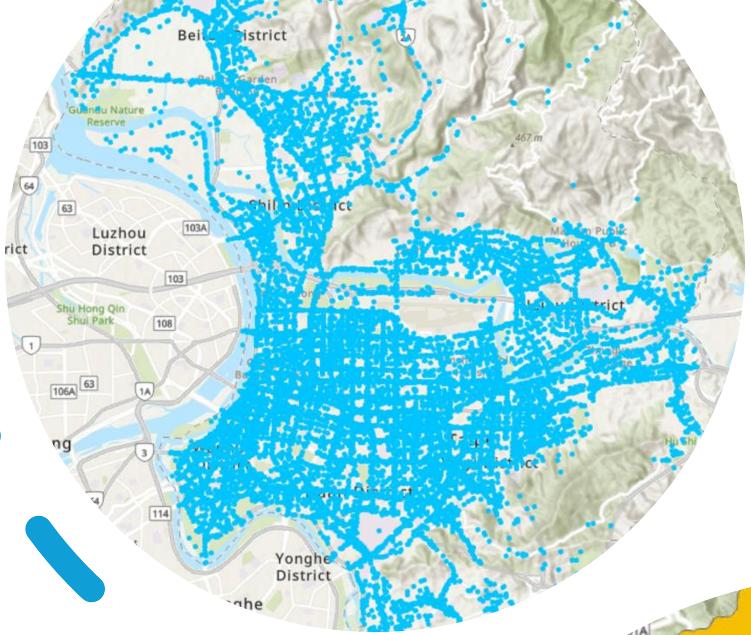
$$\frac{d\sigma_{x'}^2}{d\theta} = \frac{2}{n}\sum_{i=1}^{n}\left(\tilde{y}_i^2 sin\theta cos\theta + \tilde{x}_i\tilde{y}_i(\cos^2\theta - \sin^2\theta) - \tilde{x}_i^2 sin\theta cos\theta\right) = 0$$

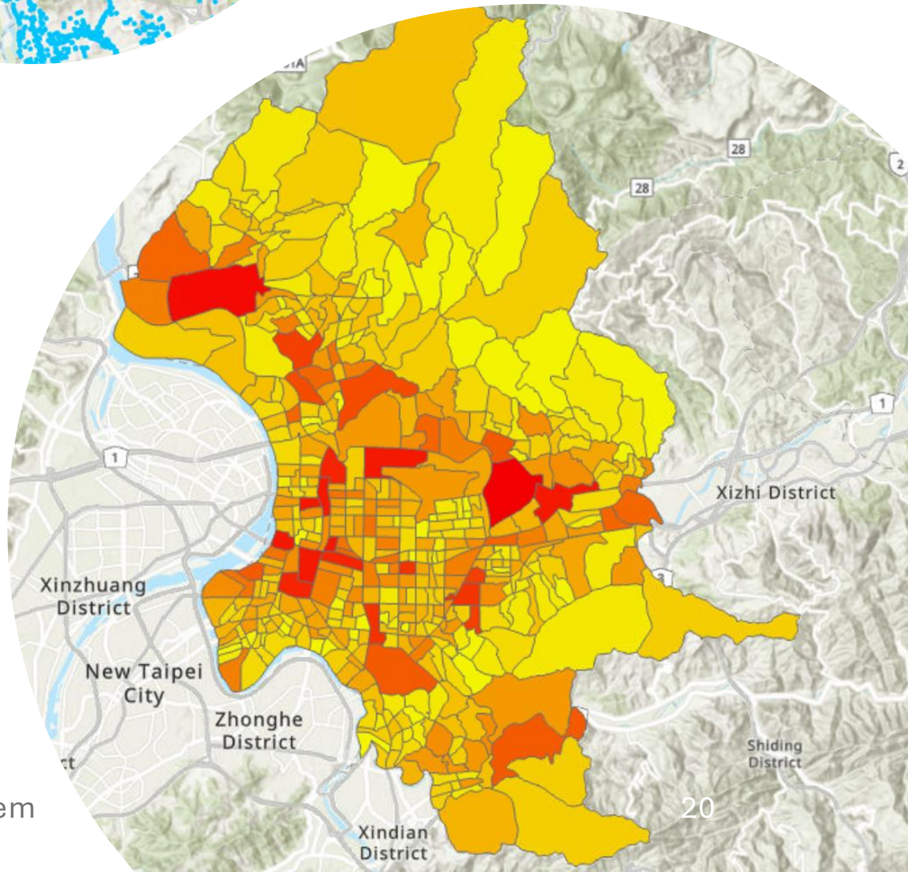
- According to Vieta's formulas, general solution to the above quadratic equation is

$$tan\theta = \frac{\left(\sum_{i=1}^{n}\tilde{x}_i^2 - \sum_{i=1}^{n}\tilde{y}_i^2\right) \pm \sqrt{\left(\sum_{i=1}^{n}\tilde{x}_i^2 - \sum_{i=1}^{n}\tilde{y}_i^2\right)^2 + 4\left(\sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i\right)^2}}{2\sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i}$$

Source: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118537

Central Tendency Measurement





- The abovementioned central tendency measurements truly illustrate the characteristics of central location or tendency of given spatial data.
- However, is it sufficient to describe a spatial data distribution?
- What kinds of spatial distribution characteristics that we have not well depicted yet?

Spatial Pattern Analysis

- From a pattern analysis perspective, the objective of the following analysis tools/ functions is to quantitatively describe the spatial data tendency based on the relationship between features (points/ polylines/ polygons/ multipoints).

Functions	Definition
Average nearest neighbor	Calculates a nearest neighbor index based on the average distance from each feature to its nearest neighboring feature.
High/ low clustering	Measures the degree of clustering for either high or low values using the Getis-Ord General G statistic.
Spatial autocorrelation	Measures spatial autocorrelation based on feature locations and attribute values using the Global Moran's I statistic.
Incremental spatial autocorrelation	Measures spatial autocorrelation for a series of distances and optionally creates a line graph of those distances and their corresponding z-scores.
Repley's *k*-function	Determines whether features, or the values associated with features, exhibit statistically significant clustering or dispersion over a range of distances.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/an-overview-of-the-analyzing-patterns-toolset.htm

What is a z-score? What is a *p*-value?

- Most statistical tests begin by identifying a null hypothesis. The null hypothesis for the pattern analysis tools (Analyzing Patterns toolset and Mapping Clusters toolset) is Complete Spatial Randomness (CSR), either of the features themselves or of the values associated with those features.
- **CSR ~ independent random process (IRP)** must satisfy:
  1) Any event has equal probability of being in any location, a 1^{st} order effect, e.g., intensity measurement.
  2) The location of one event is independent of the location of another event, a 2^{nd} order effect, e.g., dependency measurement.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/what-is-a-z-score-what-is-a-p-value.htm

What is a z-score? What is a *p*-value?

Z-score definition:

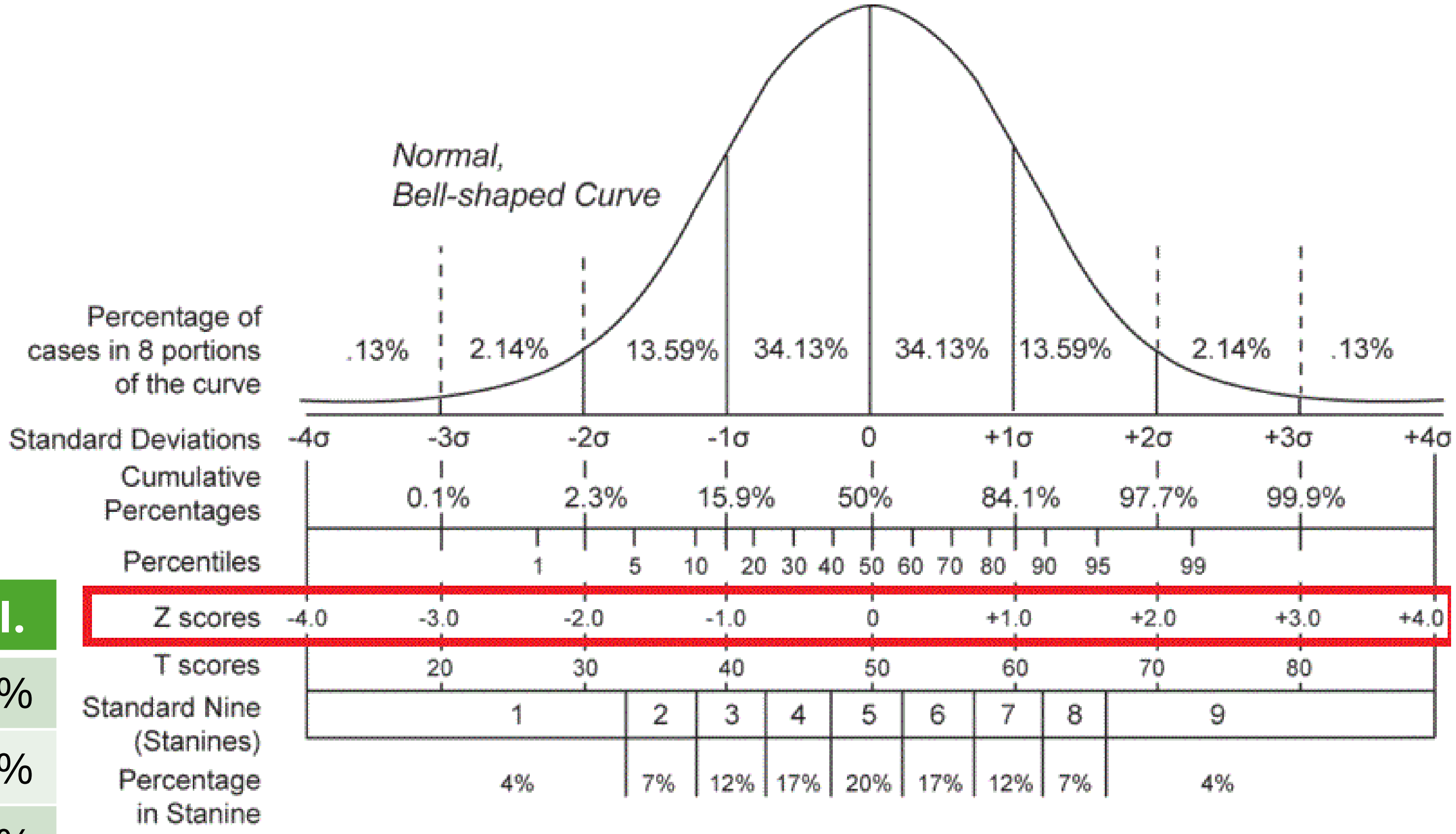
$$z = \frac{x_i - \mu}{\sigma}$$

x_i: observations

μ: mean

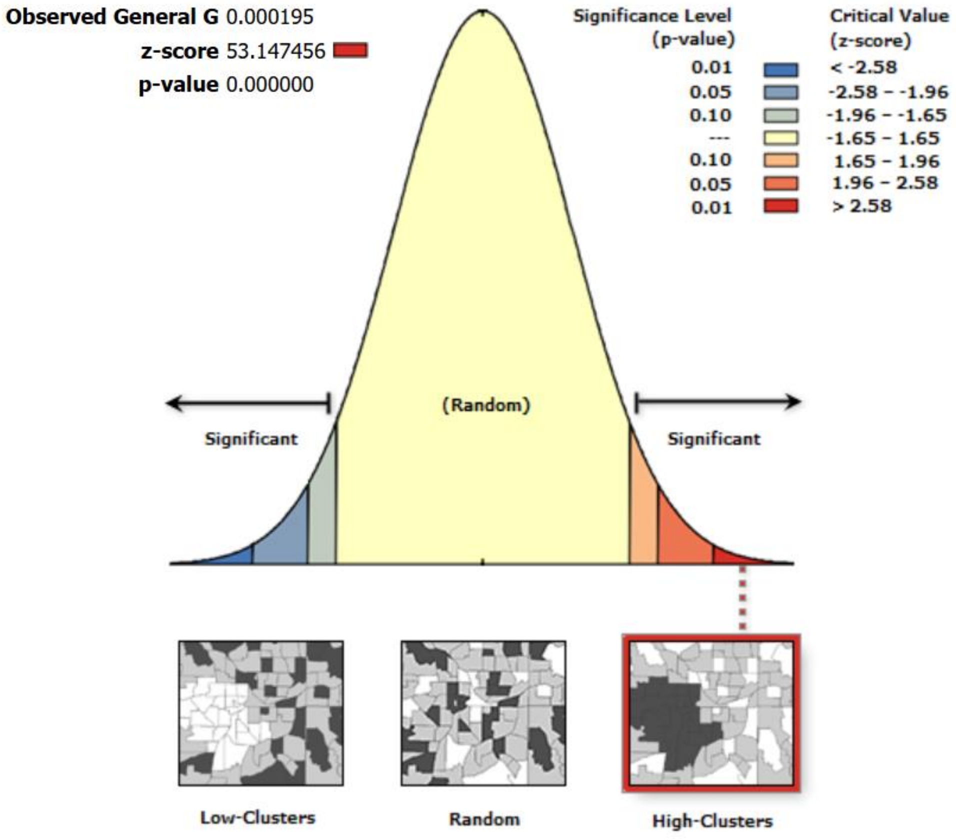
σ: standard deviation

Z-score(SD)	*p*-value	C.I.
< -1.65 or $> +1.65$	<0.10	90%
< -1.96 or $> +1.96$	<0.05	95%
< -2.58 or $> +2.58$	<0.01	99%



What is a z-score? What is a *p*-value?

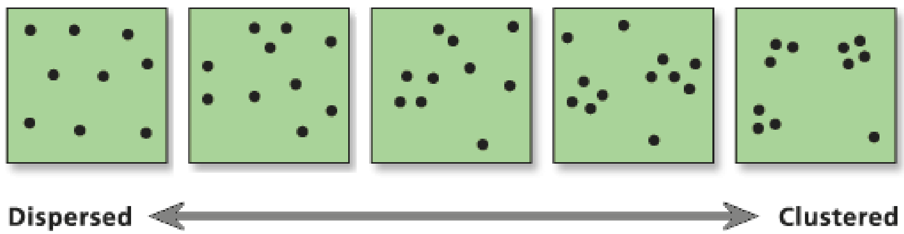
- The z-scores and p-values returned by the pattern analysis tools tell you whether you can reject that null hypothesis or not.
- Z-scores are standard deviations. If, for example, a tool returns a z-score of +2.5, you would say that the result is 2.5 standard deviations. Both z-scores and p-values are associated with the standard normal distribution as shown below.



Given the z-score of 53.147455873097456, there is a less than 1% likelihood that this high-clustered pattern could be the result of random chance.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/what-is-a-z-score-what-is-a-p-value.htm

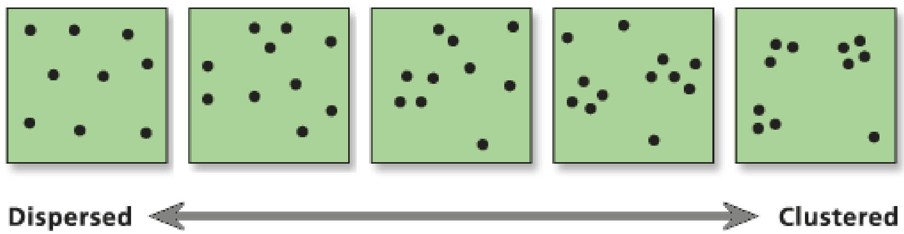
Average Nearest Neighbor



- Calculates a nearest neighbor index based on the average distance from each feature to its nearest neighboring feature.
- The Average Nearest Neighbor tool returns five values: Observed Mean Distance, Expected Mean Distance, Nearest Neighbor Index, z-score, and p-value.
- The z-score and p-value results are measures of statistical significance which tell you whether or not to reject the null hypothesis. However, that the statistical significance for this method is strongly impacted by study area size (**why?**). For the Average Nearest Neighbor statistic, the null hypothesis states that features are randomly distributed.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/average-nearest-neighbor.htm

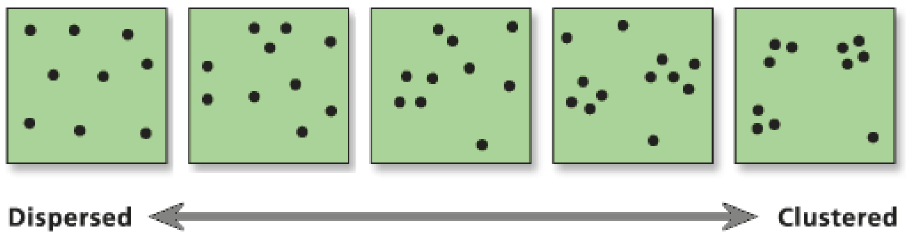
Average Nearest Neighbor



- The Nearest Neighbor Index is expressed as the ratio of the Observed Mean Distance to the Expected Mean Distance.
- The expected distance is the average distance between neighbors in a hypothetical random distribution.
- If the index is less than 1, the pattern exhibits clustering; if the index is greater than 1, the trend is toward dispersion or competition.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/average-nearest-neighbor.htm

Average Nearest Neighbor



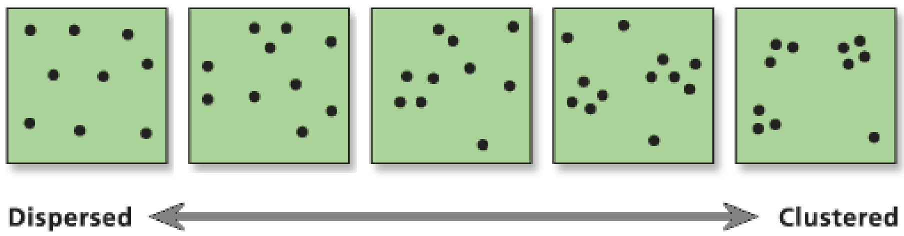
- The average nearest neighbor ratio is given as:

$$ANN = \frac{\bar{D}_o}{\bar{D}_E}, \bar{D}_o = \frac{\sum_{i=1}^{n} d_i}{n}, \bar{D}_E = \frac{0.5}{\sqrt{n/A}}$$

where $\bar{D}_o$ is the observed mean distance between each feature and its nearest neighbor. And $\bar{D}_E$ is the expected mean distance for the features given in a random pattern. d_i equals the distance between feature i and it is nearest neighboring feature, n corresponds to the total number of feature, and A is the area of a minimum enclosing rectangle around all features, or it's user-specified area value.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/average-nearest-neighbor.htm

Average Nearest Neighbor



$$
\begin{aligned}
ANN &= 0 \Rightarrow completely\ clustered \\
ANN &= 1 \Rightarrow random \\
ANN &= 2.149\ completely\ dispersed
\end{aligned}
$$

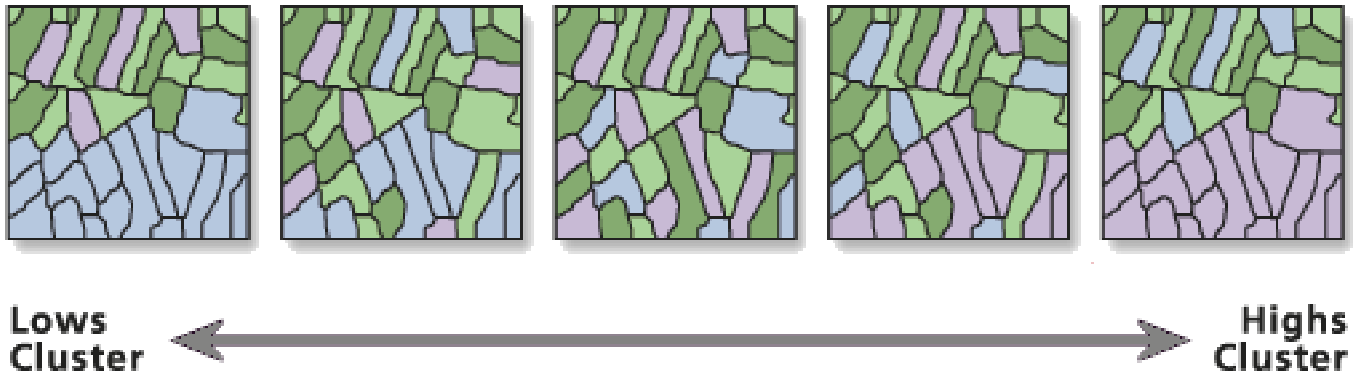
- The average nearest neighbor z-score for the statistic is calculated as:

$$z = \frac{\bar{D}_o - \bar{D}_E}{SE}, SE = \frac{0.26136}{\sqrt{\frac{n^2}{A}}}$$

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/average-nearest-neighbor.htm

High/ Low Clustering (Getis-Ord General G)

- Measures the degree of clustering for either high or low values using the Getis-Ord General G statistic.



- The High/Low Clustering tool returns four values: Observed General G, Expected General G, z-score, and *p*-value.
- When chordal distances are used in the analysis, the Distance Band or Threshold Distance parameter value, if specified, should be in meters.

High/ Low Clustering (Getis-Ord General G)

- The Conceptualization of Spatial Relationships parameter should reflect inherent relationships among the features you are analyzing. The more realistically you can model how features interact with each other in space, the more accurate your results will be.
- **Fixed distance band:** The Distance Band or Threshold Distance parameter will ensure that each feature has at least one neighbor. This is important, but often the calculated default will not be the most appropriate distance to use for your analysis.
- **Inverse distance or Inverse distance squared:** When zero is entered for the Distance Band or Threshold Distance parameter, all features are considered neighbors of all other features; when this parameter is left blank, the default distance will be applied.

High/ Low Clustering (Getis-Ord General G)

- The General G statistic of overall spatial association is given as:

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} x_i x_j}{\sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j}, \forall j \neq i$$

  where x_i and x_j are attribute values for features i, j, and $w_{i,j}$ is the spatial weight between feature i and j. n is the number of features in the dataset and $\forall j \neq i$ indicates that features i and j cannot be the same feature.

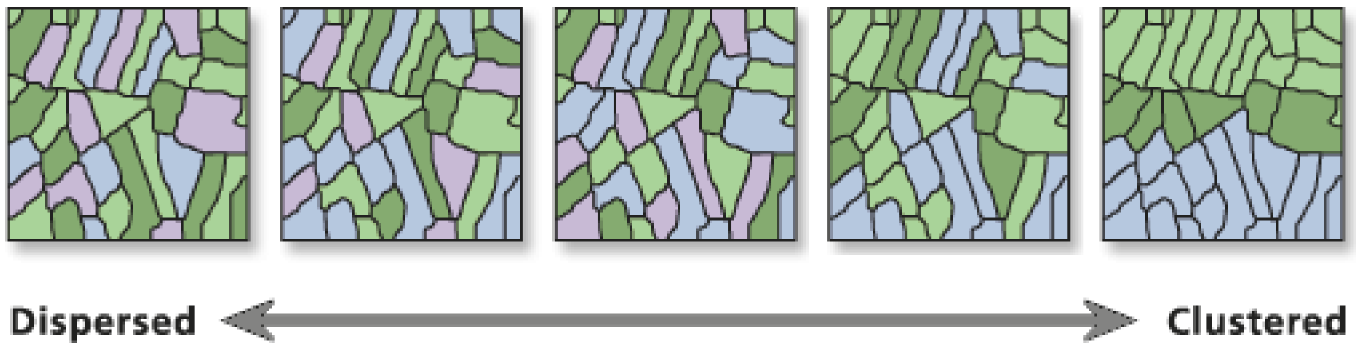
- The $z_G - score$ for the statistic is computed as:

$$z_G = \frac{(G - E[G])}{\sqrt{V[G]}}, where\ E[G] = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}}{n(n-1)}, \forall j \neq i$$

$$V[G] = E[G^2] - E[G]^2$$

Spatial Autocorrelation (Global Moran’s I)

- Measures spatial autocorrelation based on feature locations and attribute values using the Global Moran's I statistic.



- The Spatial Autocorrelation tool returns five values: the Moran's I Index, Expected Index, Variance, z-score, and p-value.

Spatial Autocorrelation (Global Moran’s I)

- For a set of features and an associated attribute, this tool evaluates whether the pattern expressed is clustered, dispersed, or random. When the z-score or p-value indicates statistical significance, a positive Moran's I index value indicates tendency toward clustering, while a negative Moran's I index value indicates tendency toward dispersion.
- When using the Fixed distance band option, the default Distance Band or Threshold Distance parameter value will ensure that each feature has at least one neighbor. This is important, but often this default will not be the most appropriate distance to use for an analysis.

Spatial Autocorrelation (Global Moran's I)

- The Moran's I statistic for spatial autocorrelation is given as:

$$I = \frac{n}{W} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2} = \frac{n}{W} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

  where z_i is the deviation of an attribute for feature i from its mean $(x_i - \bar{X})$, $w_{i,j}$ is the spatial weight betweeen feature i and j, n is equal to the total number of features, and W is the aggregate of all the spatial weights:

$$W = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}$$

Spatial Autocorrelation (Global Moran's I)

- The $z_I - score$ for the statistic is computed as:

$$z_I = \frac{(I - E[I])}{\sqrt{V[I]}}$$

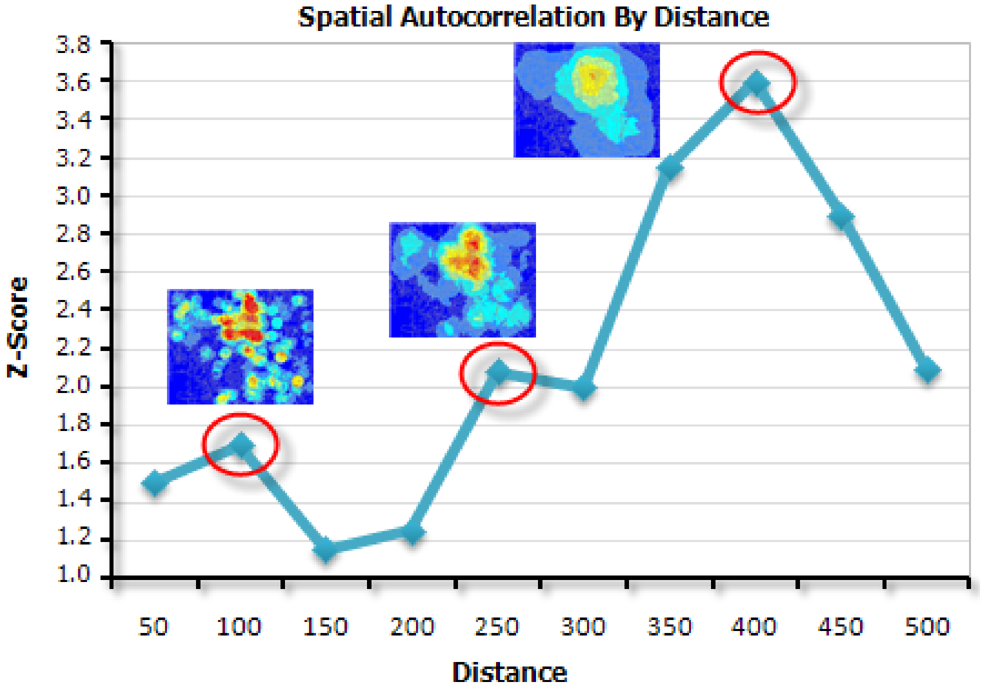
$$where: \quad E[I] = -\frac{1}{n-1}$$

$$V[I] = E[I^2] - E[I]^2$$

Incremental Spatial Autocorrelation

Measures spatial autocorrelation for a series of distances and optionally creates a line graph of those distances and their corresponding $z - scores$.

$Z - scores$ reflect the intensity of spatial clustering, and statistically significant peak $z - scores$ indicate distances where spatial processes promoting clustering are most pronounced.

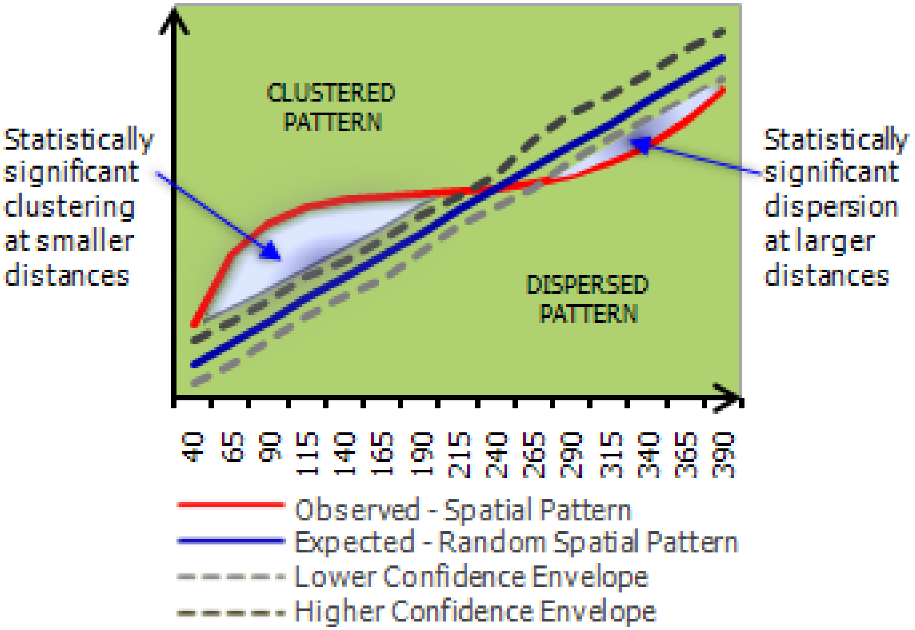


These peak distances are often appropriate values to use for tools with a Distance Band or Distance Radius parameter.

Incremental Spatial Autocorrelation

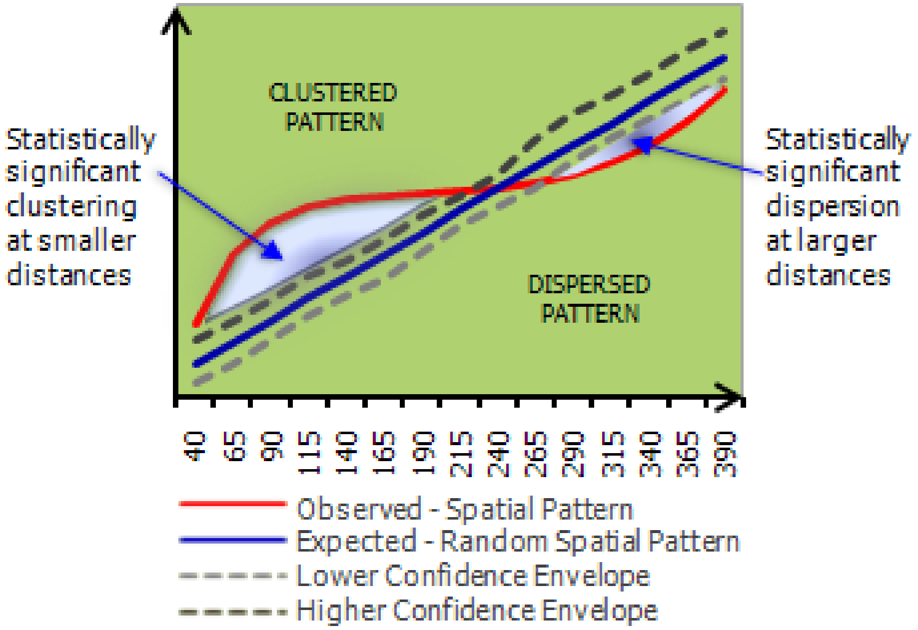
- Use this tool to specify an appropriate Distance Threshold or Radius parameter value for tools that have these parameters, such as Hot Spot Analysis or Point Density.
- When chordal distances are used in the analysis, the Beginning Distance and Distance Increment parameter values, if provided, should be in meters.
- For line and polygon features, feature centroids are used in distance computations. For multipoints, polylines, or polygons with multiple parts, the centroid is computed using the weighted mean center of all feature parts. The weighting for point features is 1, for line features is length, and for polygon features is area.

Repley's *k*-function



- Determines whether features, or the values associated with features, exhibit statistically significant clustering or dispersion over a range of distances.
- This tool requires projected data to accurately measure distances.
- Tool output is a table with fields: ExpectedK and ObservedK containing the expected and observed K values, respectively. The Weight Field is most appropriately used when it represents the number of incidents or counts.

Repley's *k*-function



- The following explains how the confidence envelope is c
- **No Weight Field:** When no Weight Field is specified, constructed by distributing points randomly in the stu for that distribution. Each random distribution c "permutation". If 99 permutations is selected, for exa distribute the set of points 99 times for each iteration. After distributing the points 99 times the tool selects, for each distance, the Observed k value that deviated above and below the Expected k value by the greatest amount; these values become the confidence interval.
- **Including a Weight Field:** When a Weight Field is specified, only the weight values are randomly redistributed to compute confidence envelopes; the point locations remain fixed. In essence, when a Weight Field is specified, locations remain fixed and the tool evaluates the clustering of feature values in space. On the other hand, when no Weight Field is specified the tool analyzes clustering/dispersion of feature locations.

Repley's *k*-function

- The k-function is given as:

$$L(d) = \sqrt{\frac{A \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} k_{i,j}}{\pi n(n-1)}}$$

where d is the distance, n is equal to the total number of features, A represents the total area of the features and $k_{i,j}$ is a weight. If there is no edge correction, then the weight will be equal to one when the distance between i and j is less than d, and will equate to zero otherwise.

Lab #01 Central Tendency Algorithm ...

- How does the central tendency algorithm process the following data types in the Mean Center, Median Center, Standard Distance, Central Feature, Directional Distribution?

Algorithm	Point	Polygon	MultiPoint	Polylines	Projected?*
Mean Center					
Median Center					
Standard Distance					
Central Feature					
Directional Distribution					

* Projected?: Does the central tendency algorithm require the projected data?

The End

Thank you for your attention! |

Email: chchan@ntnu.edu.tw
Web: toodou.github.io