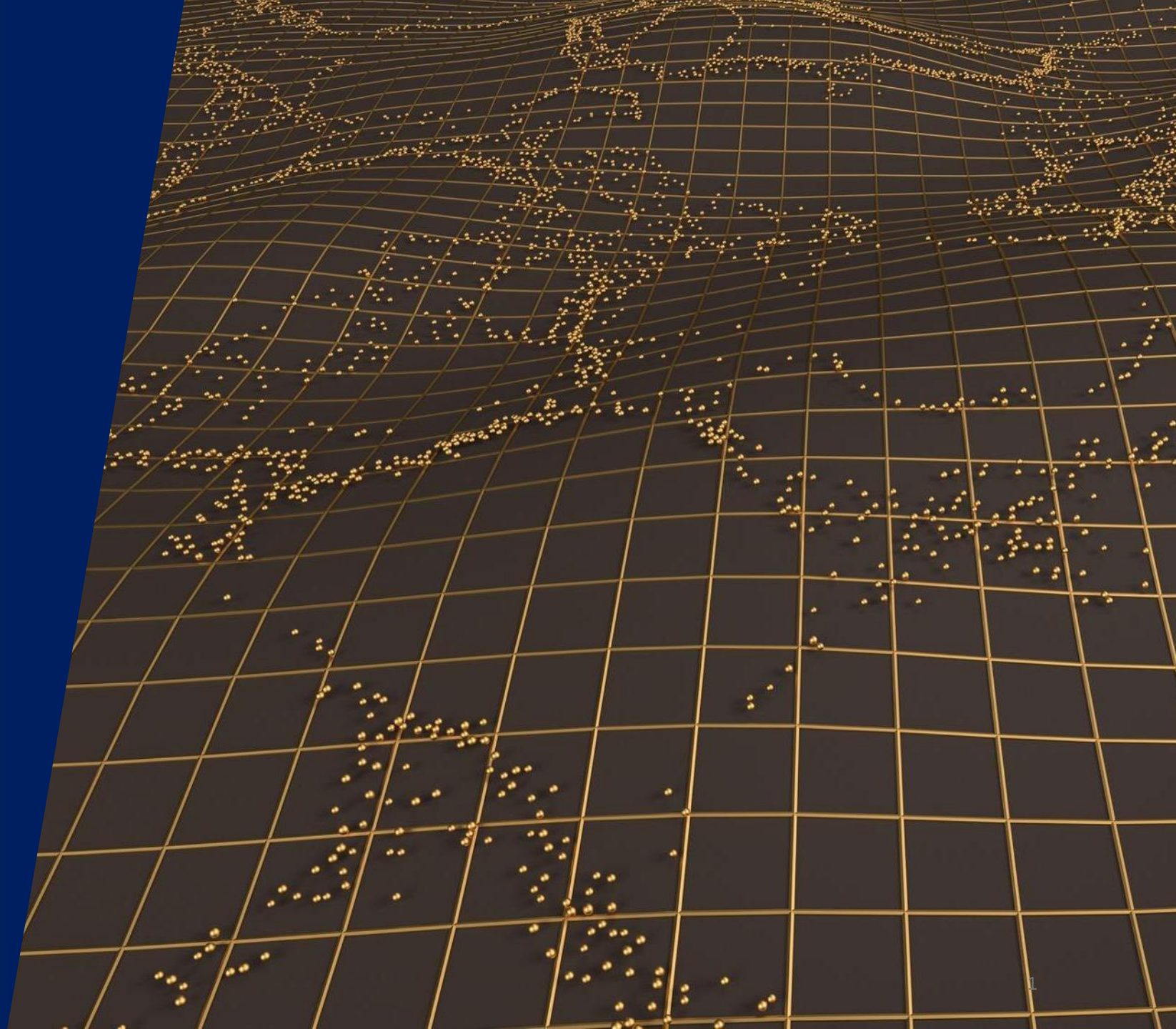
SAINT LOUIS UNIVERSITY.

Data Exploration and Spatial Statistics

Sourav Bhadra, Ph. D.



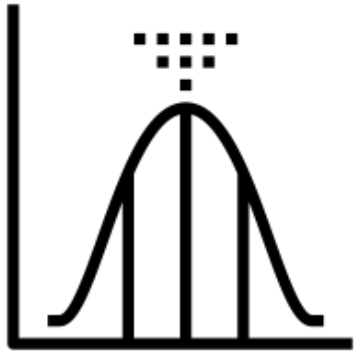
GIS4120/5120 Geospatial Analytics
9/26/2024

Contents we will cover

- Descriptive statistics
- Spatial sampling
- Exploratory spatial data analysis
- Grid-based statistics
- Point sets and distance statistics

Descriptive Statistics

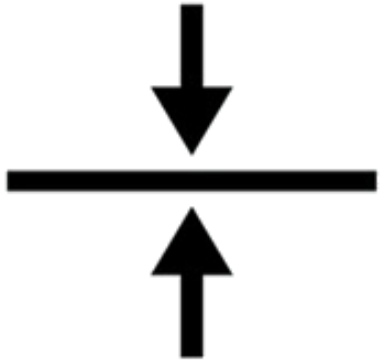
Measures of central tendency



Mean

Sum of scores divided by the number of scores

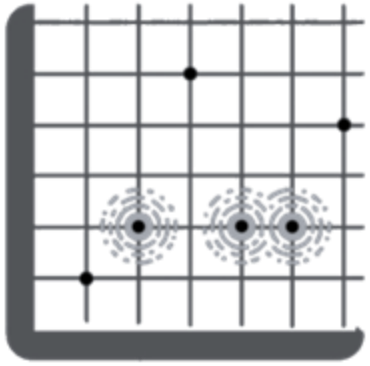
Biased Estimator



Median

The middle value when all values are ordered from smallest to largest.

Unbiased Estimator



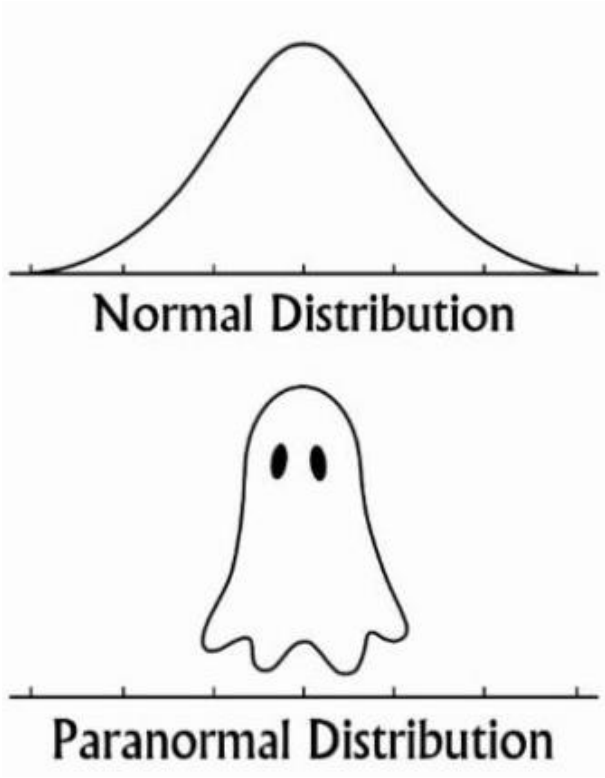
Mode

The value that appears most frequently in a dataset.

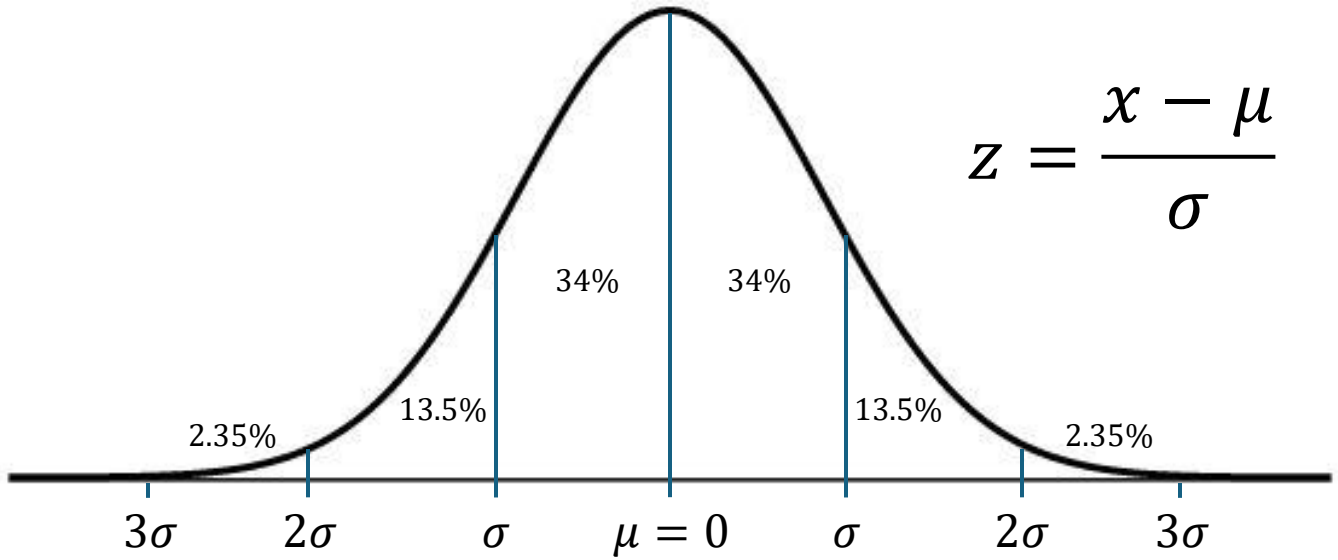
Measures of variability

Range	The difference between the highest and lowest values in a dataset.	$Range = Max\ Value - Min\ Value$
Variance	The average of the squared differences from the mean, measuring data spread.	Population: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$ Sample: $s^2 = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$
Standard Deviation	The square root of the variance, indicating the average distance from the mean.	$\sigma = \sqrt{\sigma^2}$ $s = \sqrt{s^2}$
Interquartile Range	The difference between the 75th and 25th percentiles, showing the spread of the middle 50% of the data.	$IQR = Q_{75\%} - Q_{25\%}$
Coefficient of Variation	The ratio of the standard deviation to the mean, expressed as a percentage, indicating relative variability.	$CV = \frac{\sigma}{\mu} \times 100$

A normal distribution



A normal distribution is a symmetric, bell-shaped distribution where most values cluster around the mean.



Standard Normal Distribution

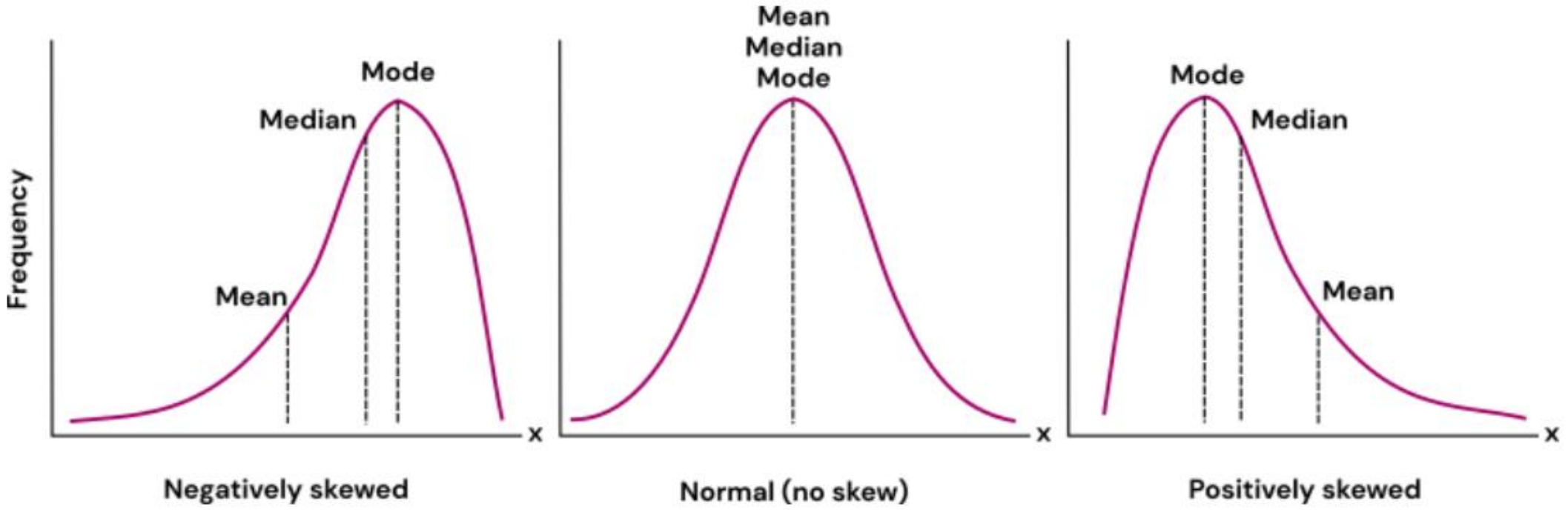
“Abnormal” distribution is also common

Skewness measures the asymmetry of a distribution.

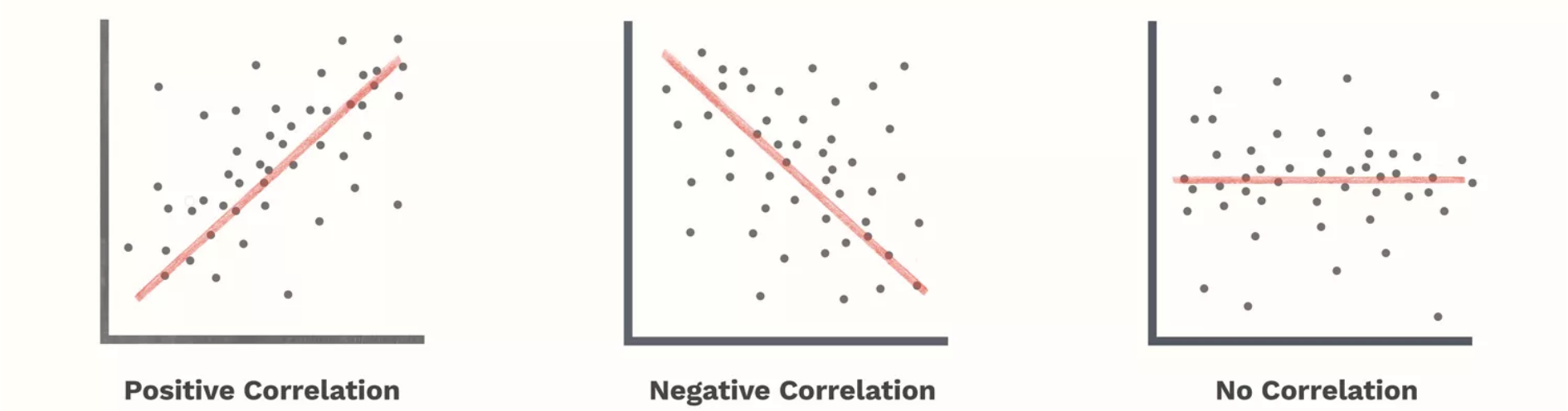
$$\frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i-\bar{x}}{s}\right)^3$$

Kurtosis the sharpness of the peak in a distribution.

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{x_i-\bar{x}}{s}\right)^4-\frac{3(n-1)^2}{(n-2)(n-3)}$$



Correlation between two variables



Pearson's Correlation (r)

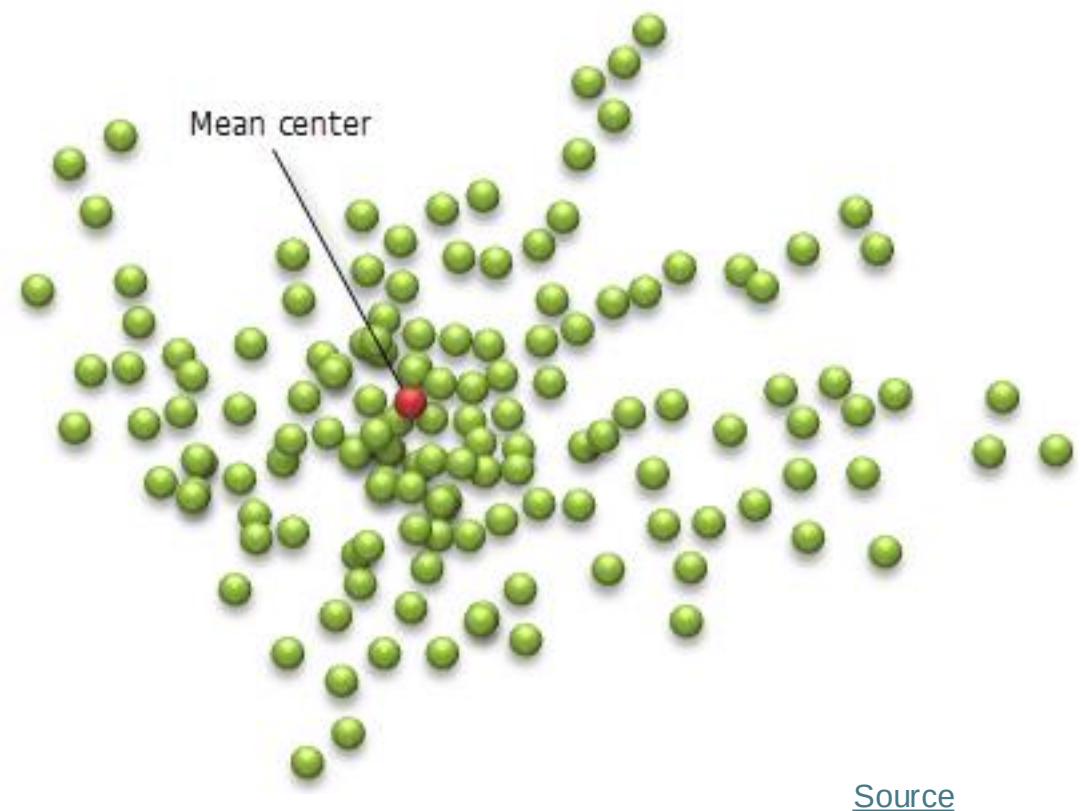
- Pearson's correlation measures the strength and direction of the linear relationship between two continuous variables.
- $r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$, if r is +1, perfect correlation and vice-versa.

Spearman's Rank Correlation (ρ)

- Spearman's a rank-based correlation assesses the relationship by converting values to ranks.
- $\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$, if ρ is +1, perfect monotonic relationship and vice-versa.

Descriptive statistics in terms of spatial data

Mean Center



Source

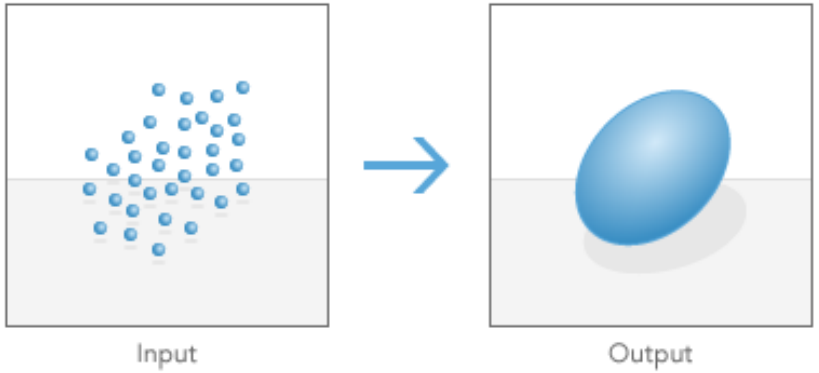
$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

We can add weight to the points:

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

Descriptive statistics in terms of spatial data

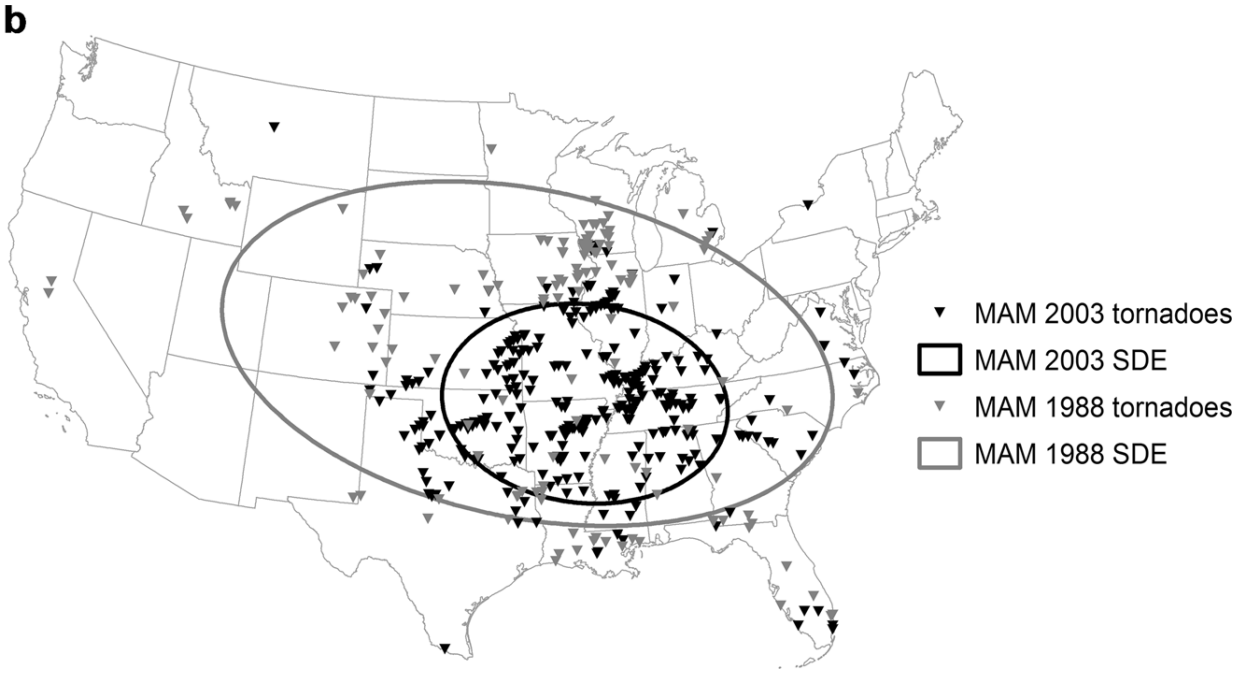
Standard Deviational Ellipse



Source

- Summarizes the geographic distribution of point features by showing their orientation, dispersion, and concentration in an elliptical shape.
- Contains center, major axis, minor axis, rotation, standard deviations along major and minor axis.

Using the standard deviational ellipse to document changes to the spatial dispersion of seasonal tornado activity in the United States

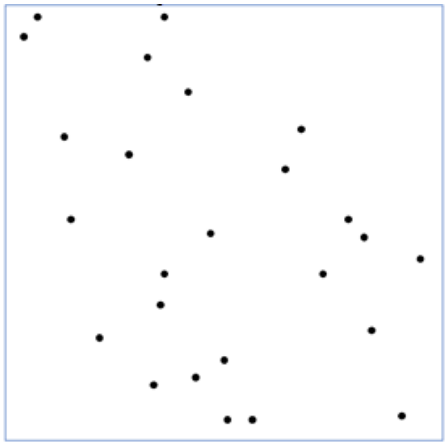


Source

Spatial Sampling

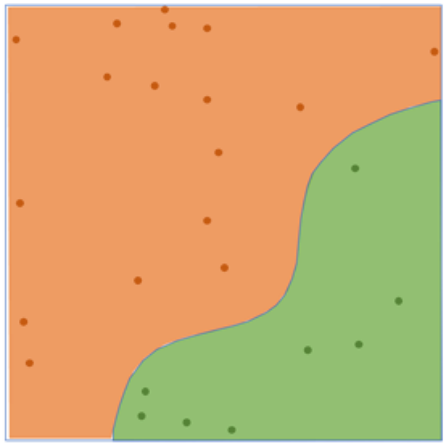
Four sampling strategies for points

Simple Random



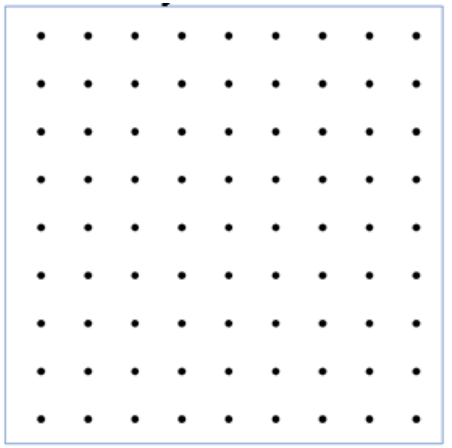
If the study area is a dense forest where every location can be assumed to have a tree.

Stratified Random



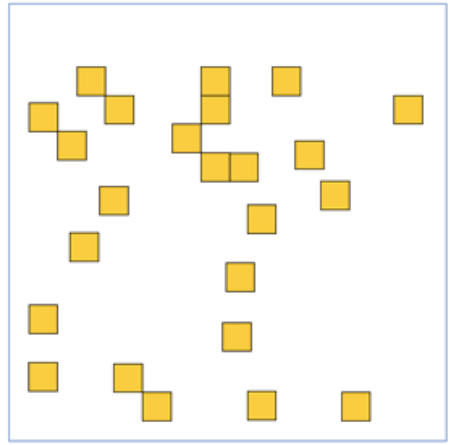
If a national park is divided into elevation classes, stratified random sampling can be used to collect soil samples separately for each elevation class.

Systematic



To study the ocean floor in a marine area, you can create a hexagonal grid of sample locations to sample marine plant species.

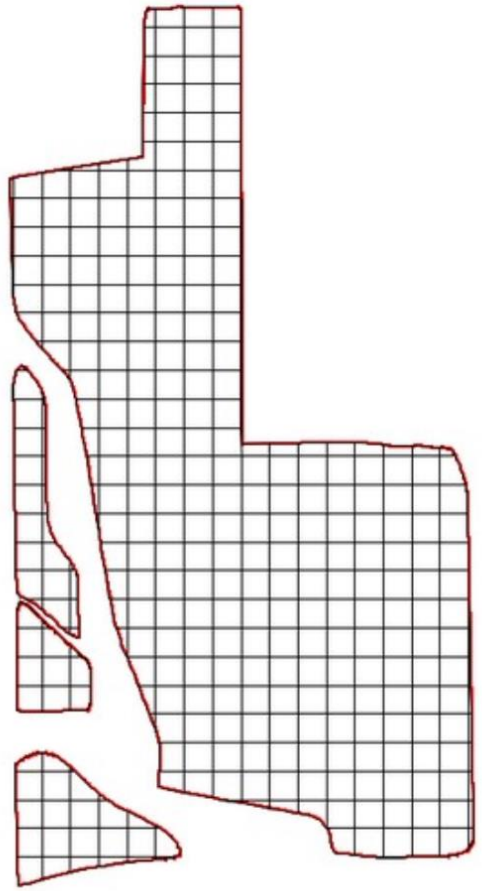
Cluster



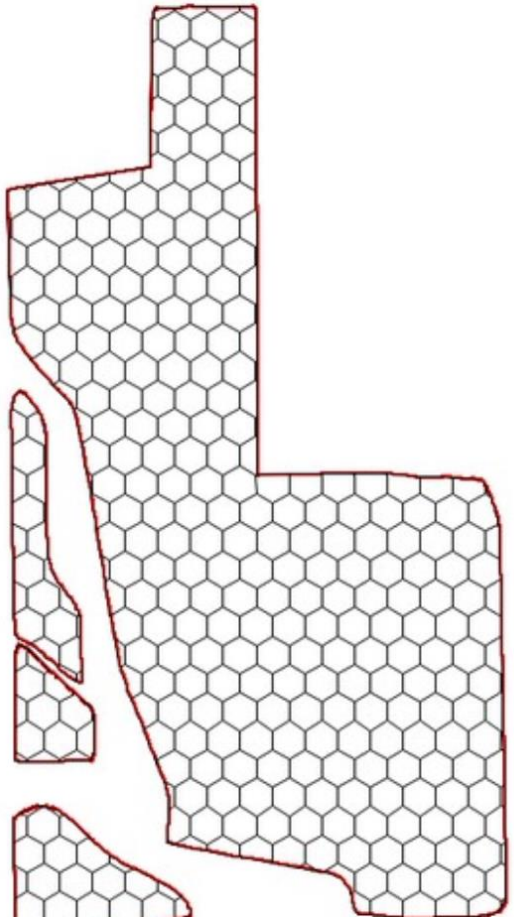
When sampling insect colonies, cluster sampling can be used to create small areas of a plot, and all insect colonies within the clusters will be sampled.

Grids are useful for defining the strata

Square Grid

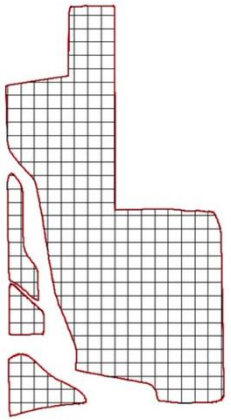


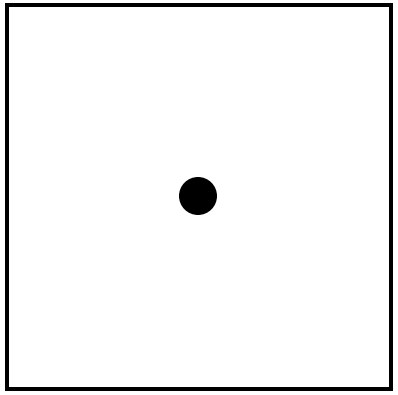
Hexagon Grid



Grids are useful for defining the strata

Square Grid

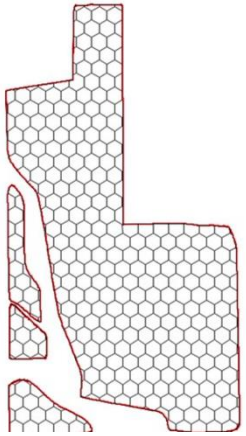


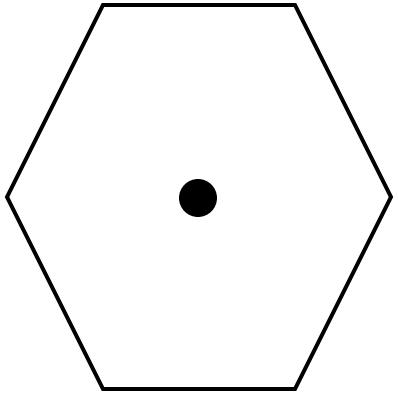


- Simple to implement
- Aligns well with raster data formats
- Facilitates straightforward spatial analysis
- Works well for regular-shaped, man-made environments.

Sampling soil properties (e.g., moisture or nutrient levels) across uniform agricultural fields.

Hexagon Grid



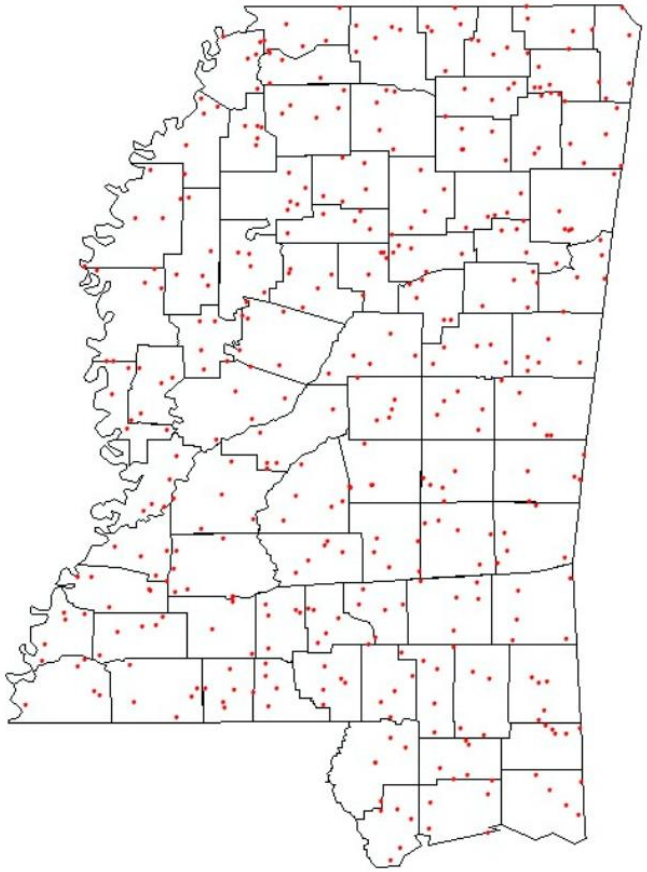


- Reduces sampling bias
- Provides equal distances between centers
- Ideal for natural terrains.

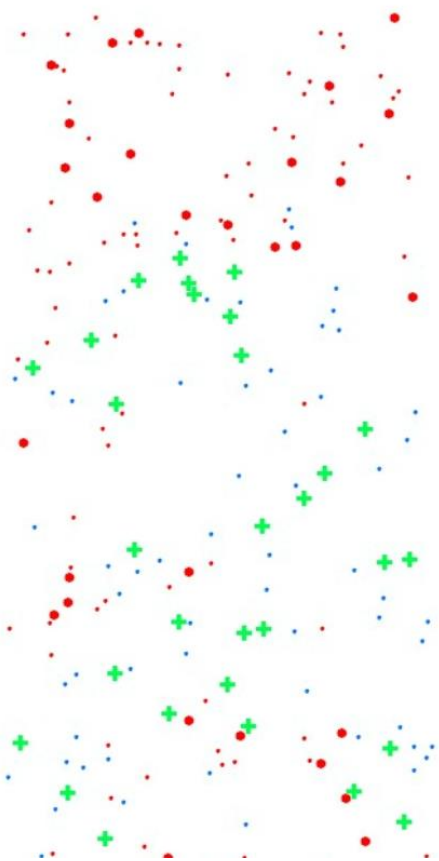
Sampling wildlife habitats, such as tracking species diversity or population density across large ecosystems

Some examples of sampling

Selecting 5 samples per county



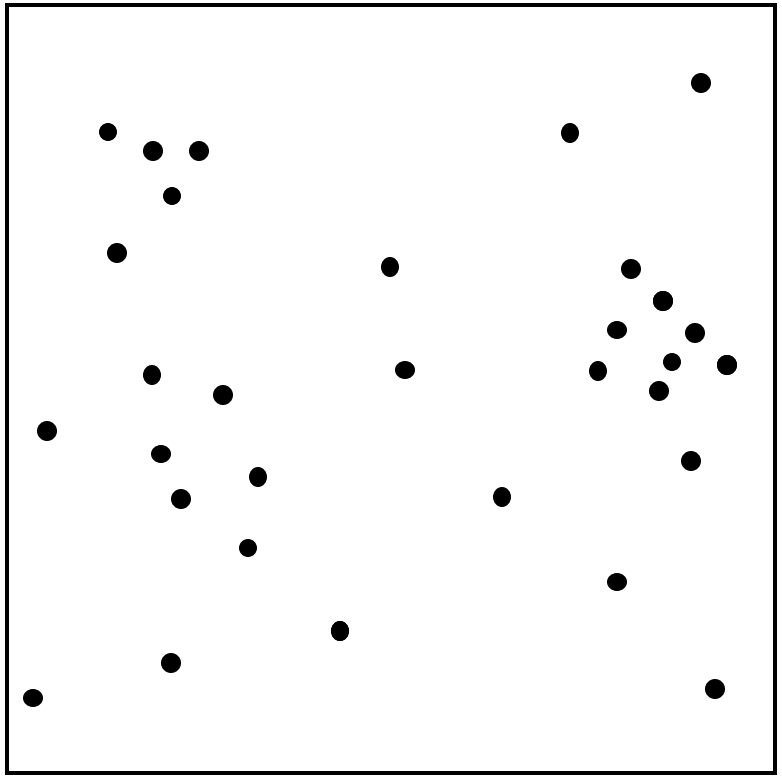
Sampling within a selected points



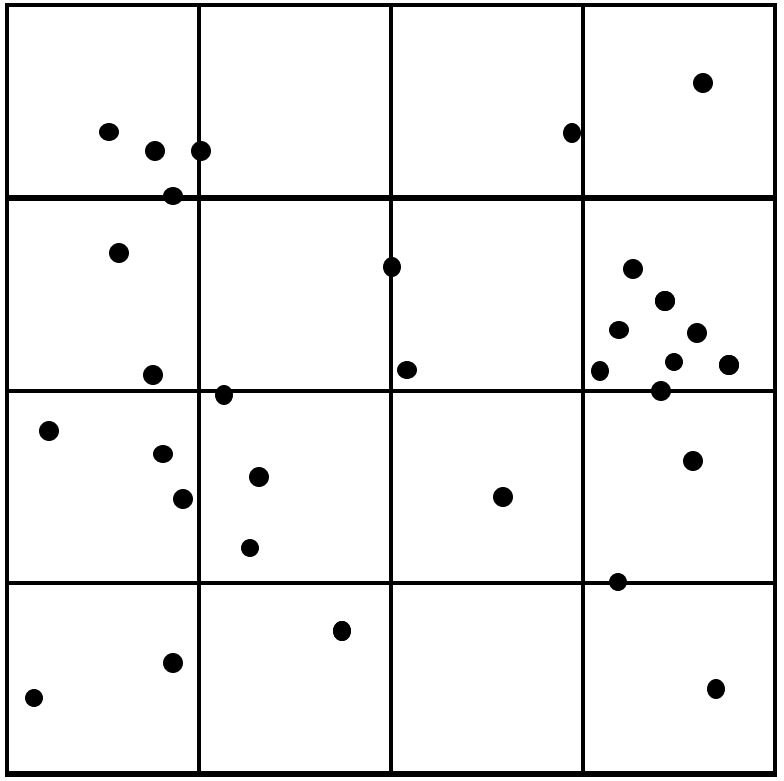
Sampling points along lines

Spatial declustering can be necessary in cases

Geological and hydrological surveys often involves intensive data collection in localized areas, resulting in clusters



Grids can be used to decluster the data and it is very necessary technique as part of data preprocessing



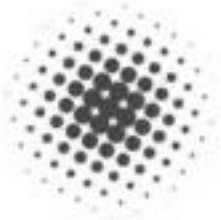
1. Cells with many points can be reduced to one by calculating mean or median value of the attribute.
2. The number of points per cell can be used as weight of the grid if having higher sample value matters.
3. Weight can also be applied by creating Voronoi polygons.

Exploratory spatial data analysis

What is Exploratory Data Analysis or EDA?

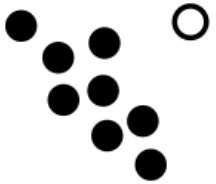
- The process of analyzing and visualizing data as a *preliminary step* before applying more complex statistical or machine learning models.
- The goals are:

Maximize insights into a dataset



Uncover underlying patterns

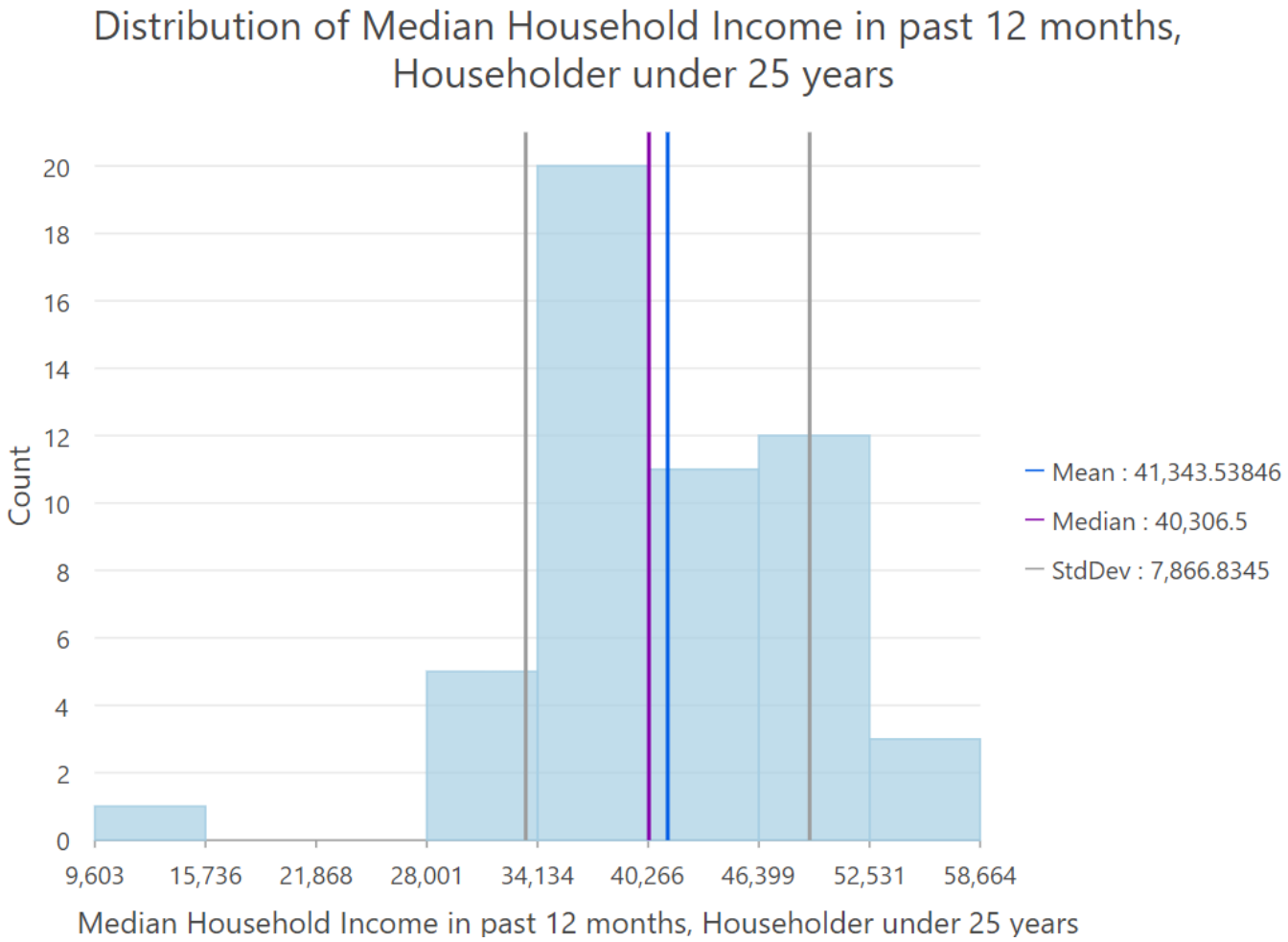
Extract important variables



Detect outliers and anomalies

Test underlying assumptions

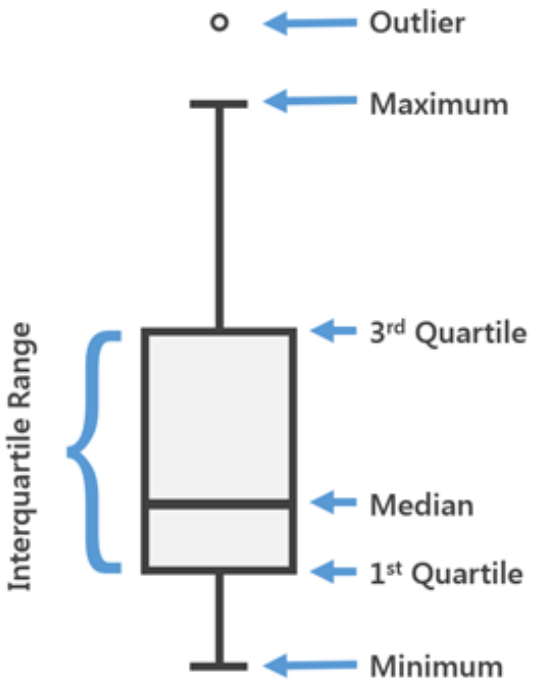
Step 1: Examine the distribution of your data

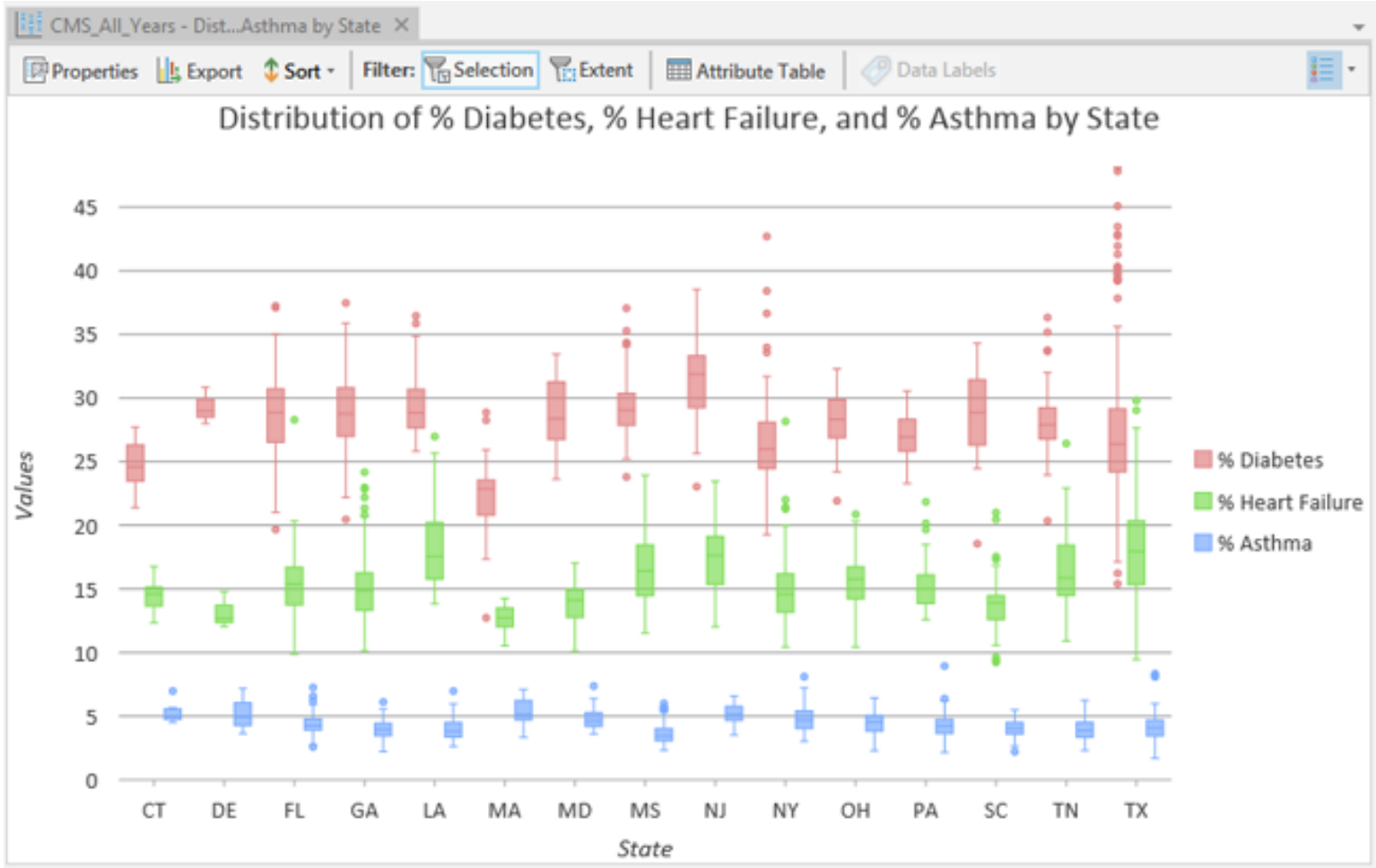


- Always check the histogram of each continuous variables (attributes) in your data before any analysis
- Perform any necessary data transformation
  - Logarithmic
  - Square root
  - Inverse
  - Box-cox
- Also check for the mean, median, and standard deviation

Step 1: Examine the distribution of your data

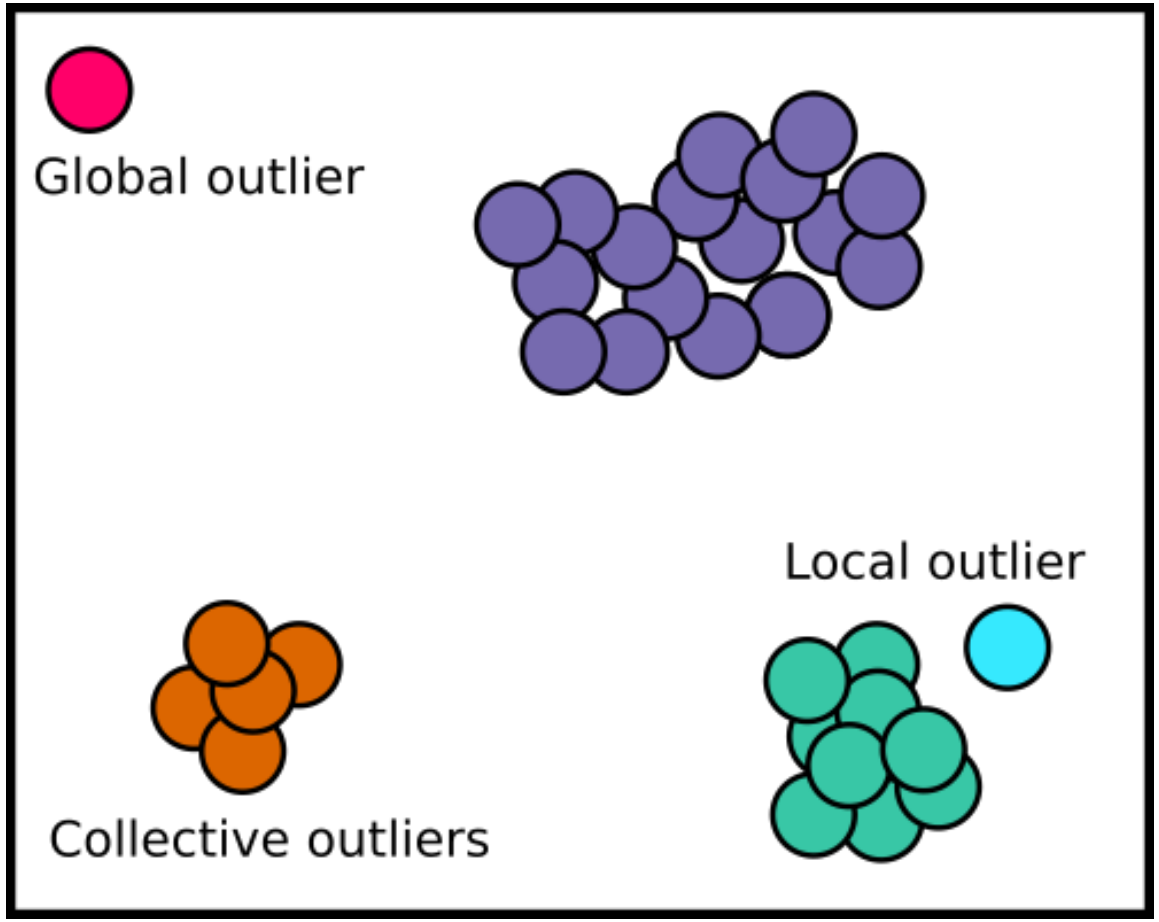
Use box plots for comparisons across variables





Source

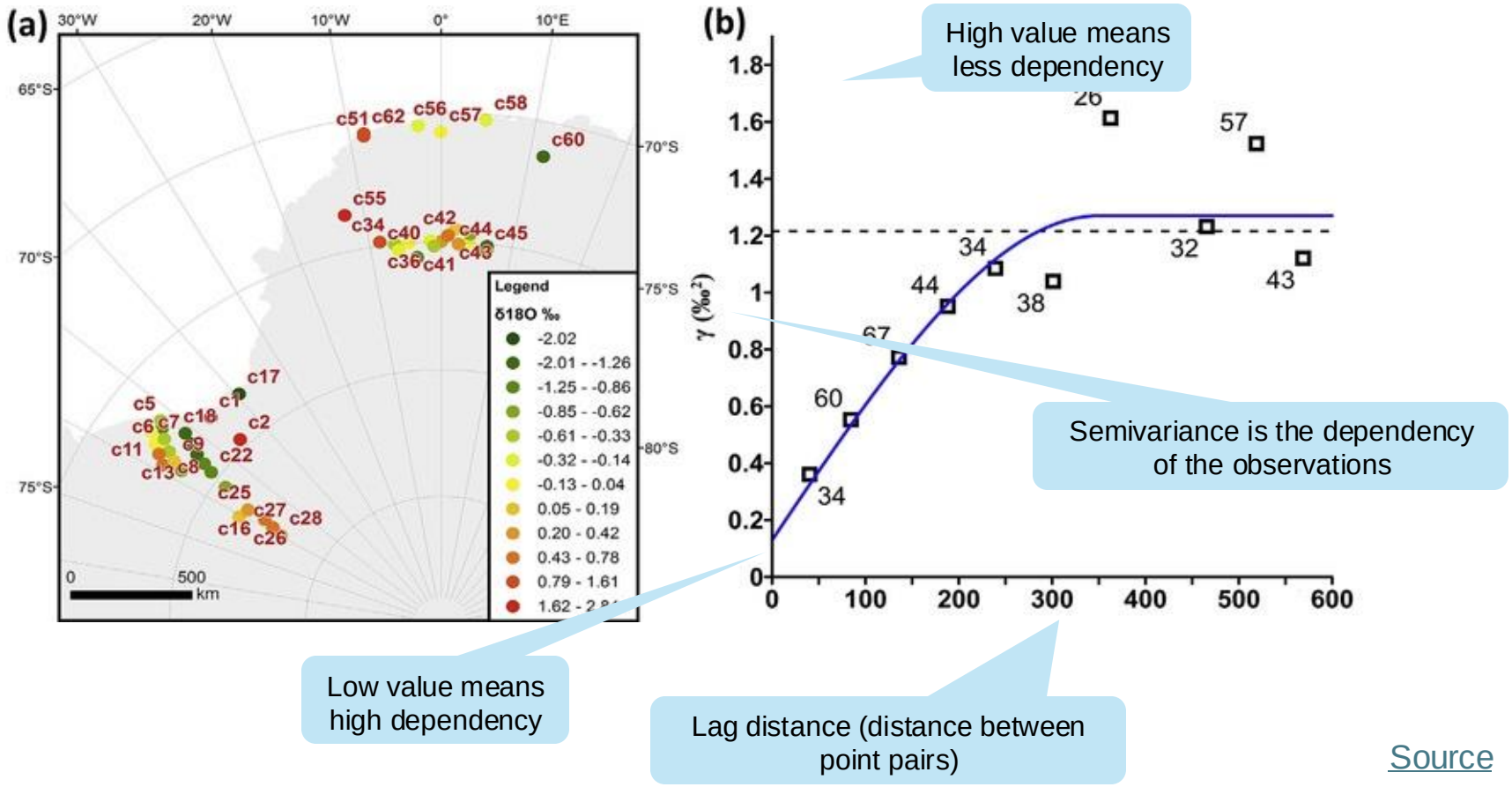
Step 2: Look for local and global outliers



Source

- **Local outliers**
  - These are points that are close to groups of data but don't belong to any cluster.
- **Global outliers**:
  - These are the data points that are completely off on their own and are far away from other data points
- **Collective outliers**
  - These are groups of outlying points which may have some underlying pattern.

Step 2: Semivariograms can be used to identify spatial outliers



High value means less dependency

Semivariance is the dependency of the observations

Low value means high dependency

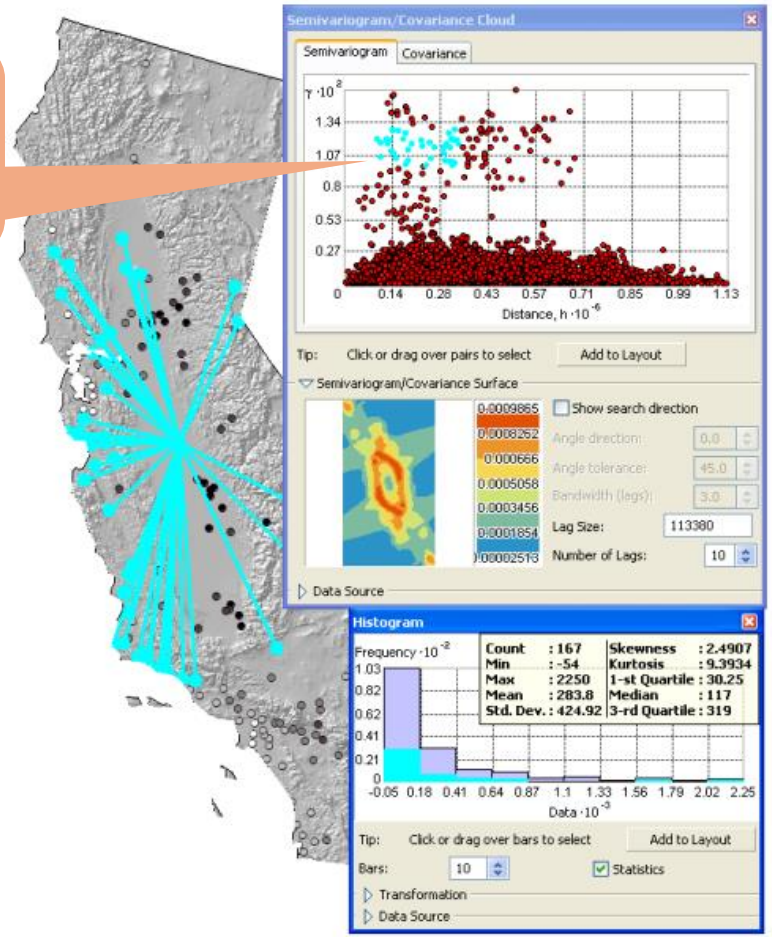
Lag distance (distance between point pairs)

Source

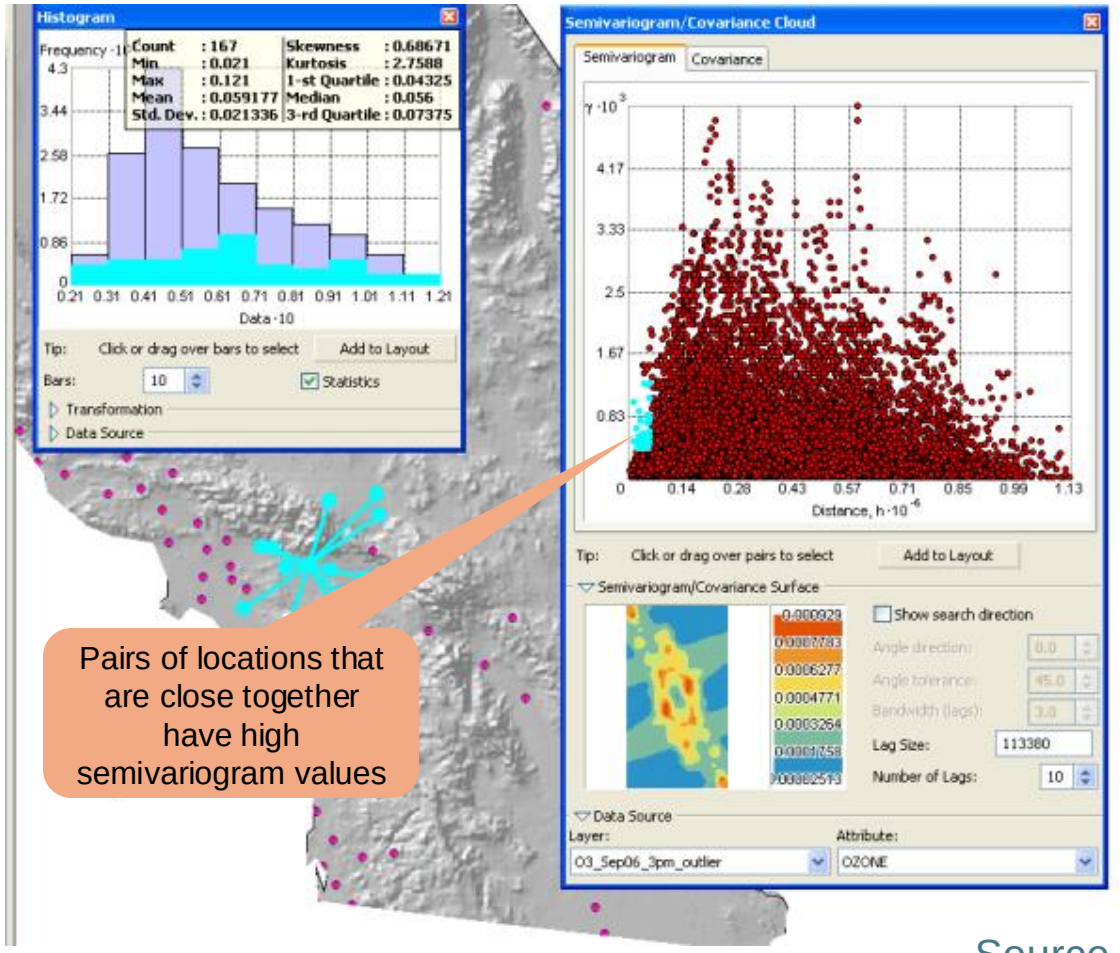
Step 2: Semivariograms can be used to identify spatial outliers

Global Outlier

No matter the lag distance, always high semivariance value



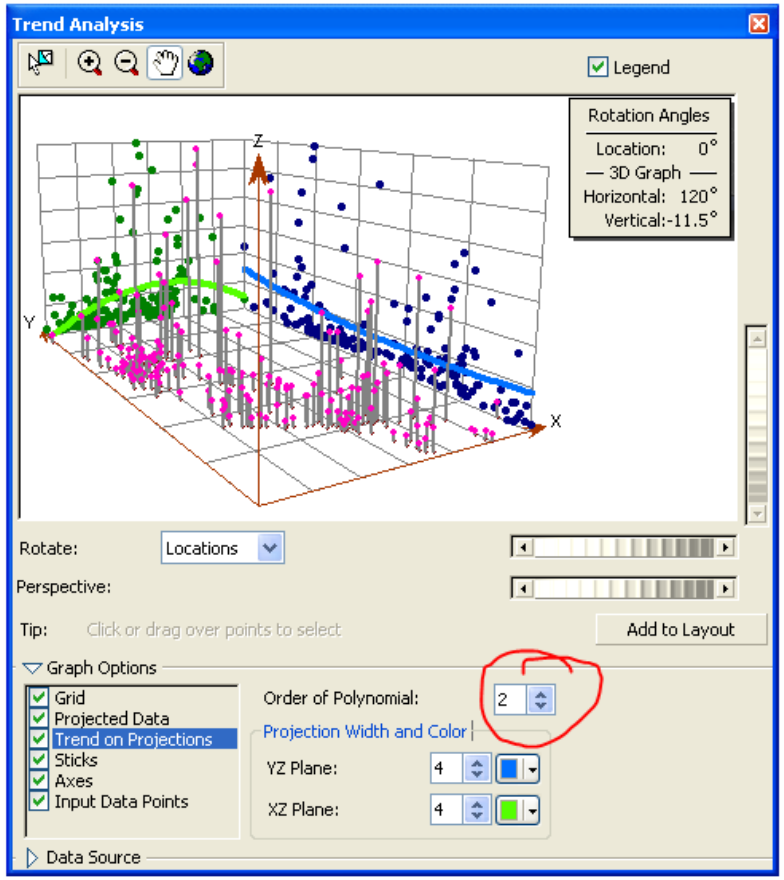
Local Outlier

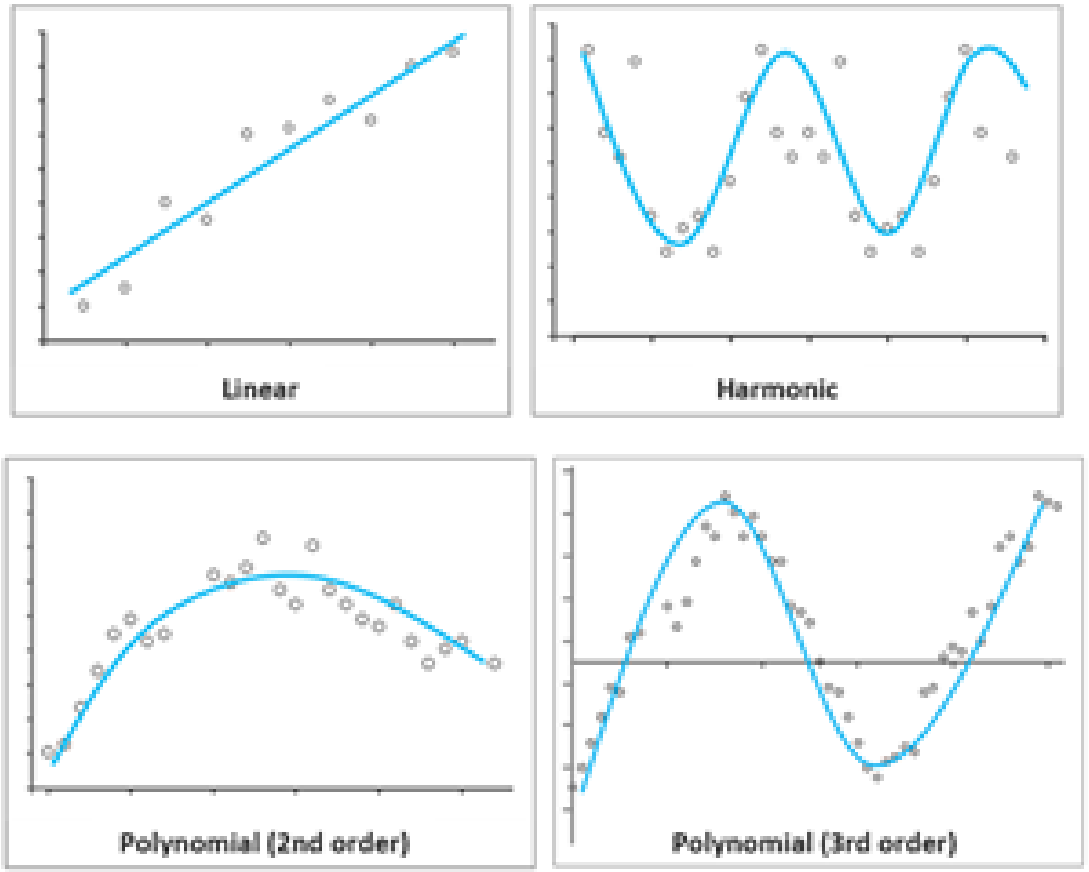


Pairs of locations that are close together have high semivariogram values

Source

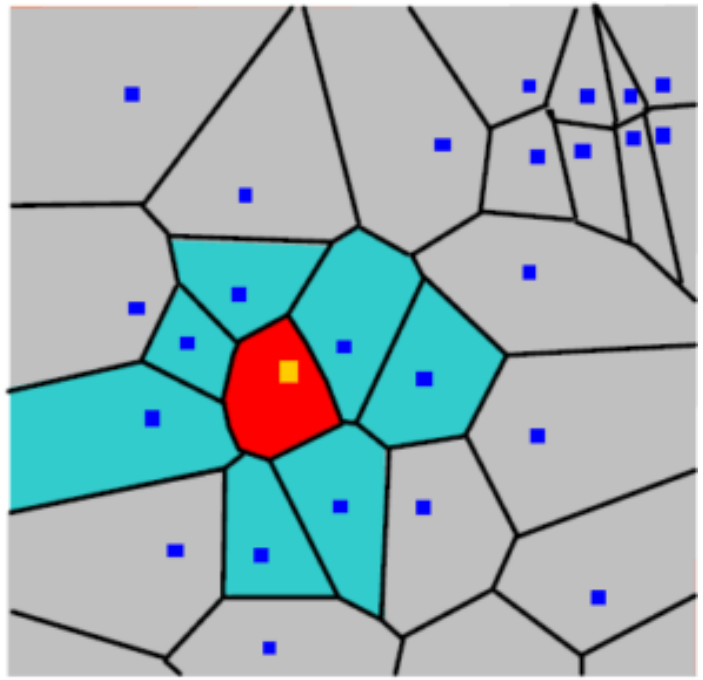
Step 3: Looking for global trends





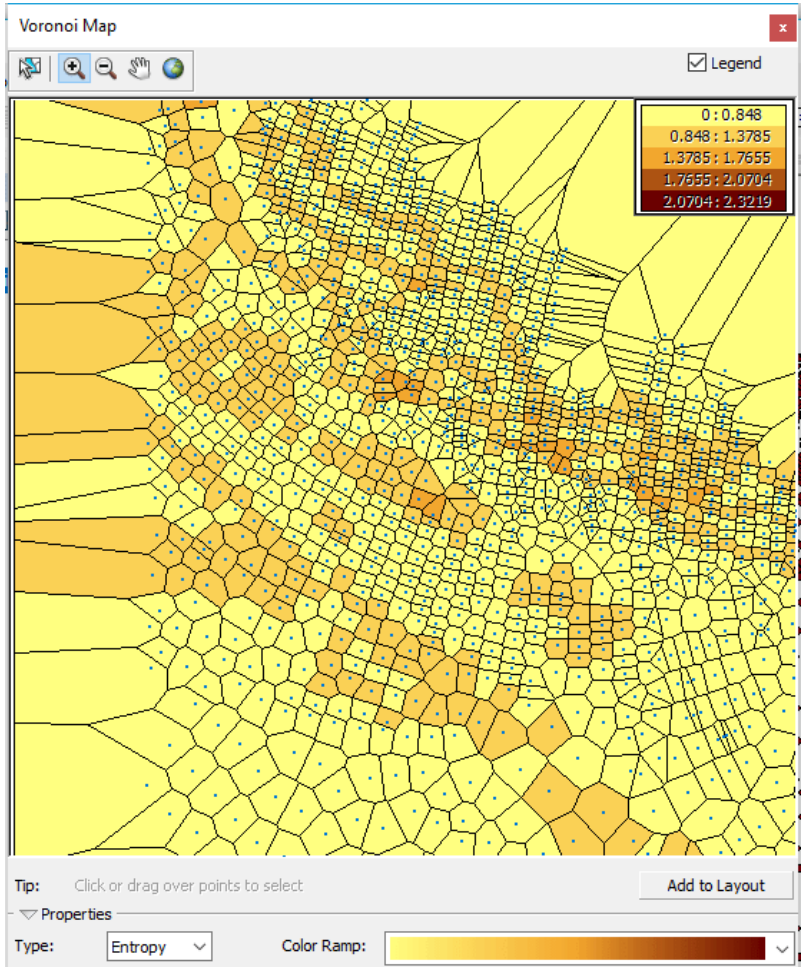
Source

Step 4: Examine local variation using Voronoi polygons



- Voronoi polygons are created so that every location within a polygon is closer to the sample point in that polygon than any other sample point.
- For the yellow point, everything within red belongs to yellow.
- All other skyblue polygons that touches the red polygon's edge are the yellow points neighboring area.
- We can compute different statistics for the neighboring points to understand local variation:
  - Standard deviation
  - Interquartile range
  - Entropy

Step 4: Examine local variation using Voronoi polygons



Source

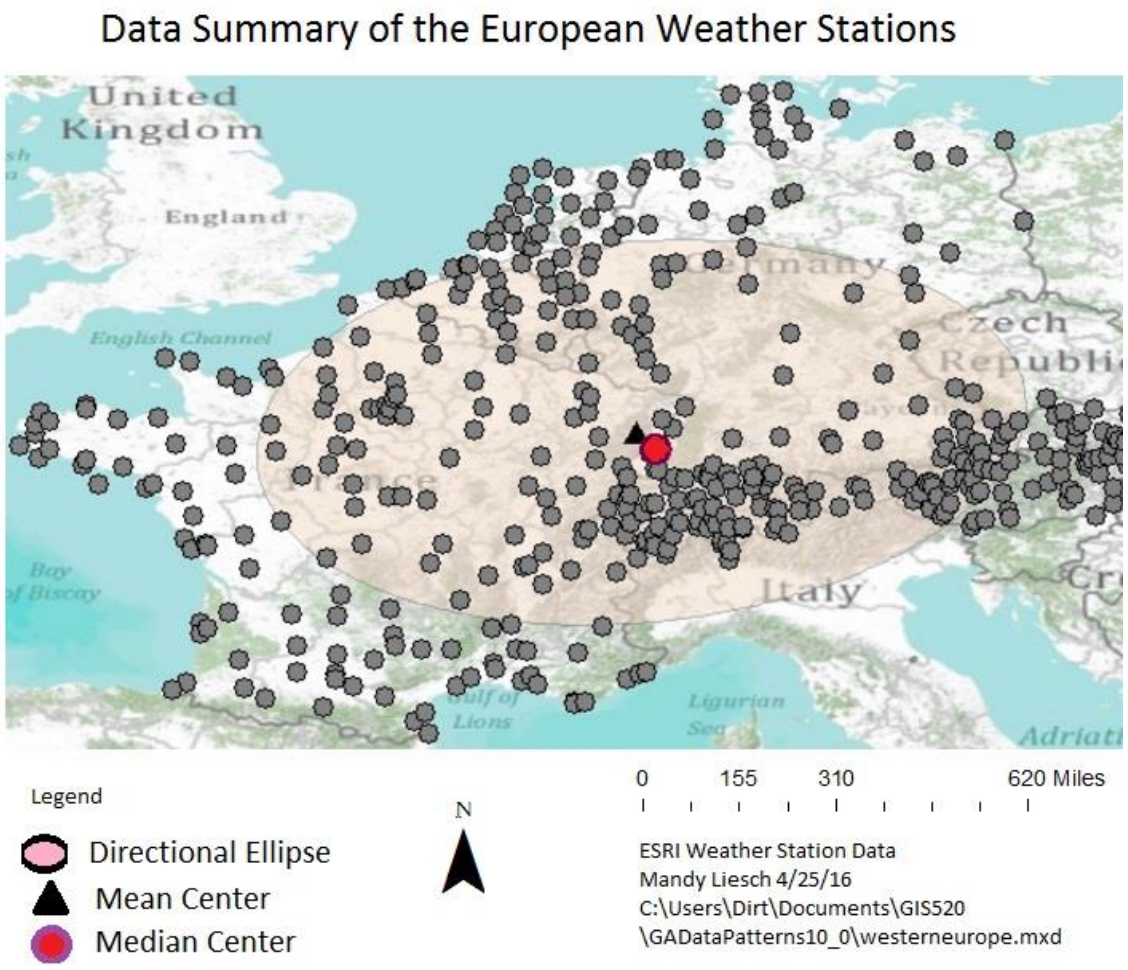
Entropy in Voronoi polygon values is calculated by analyzing the distribution of values within the polygons, typically using the Shannon entropy formula

$$H = -\sum p(x)\log(p(x))$$

- **High entropy** means the values are more uniformly distributed or unpredictable, indicating greater diversity or randomness
- **Low entropy** indicates more uniformity or predictability, where certain values dominate.

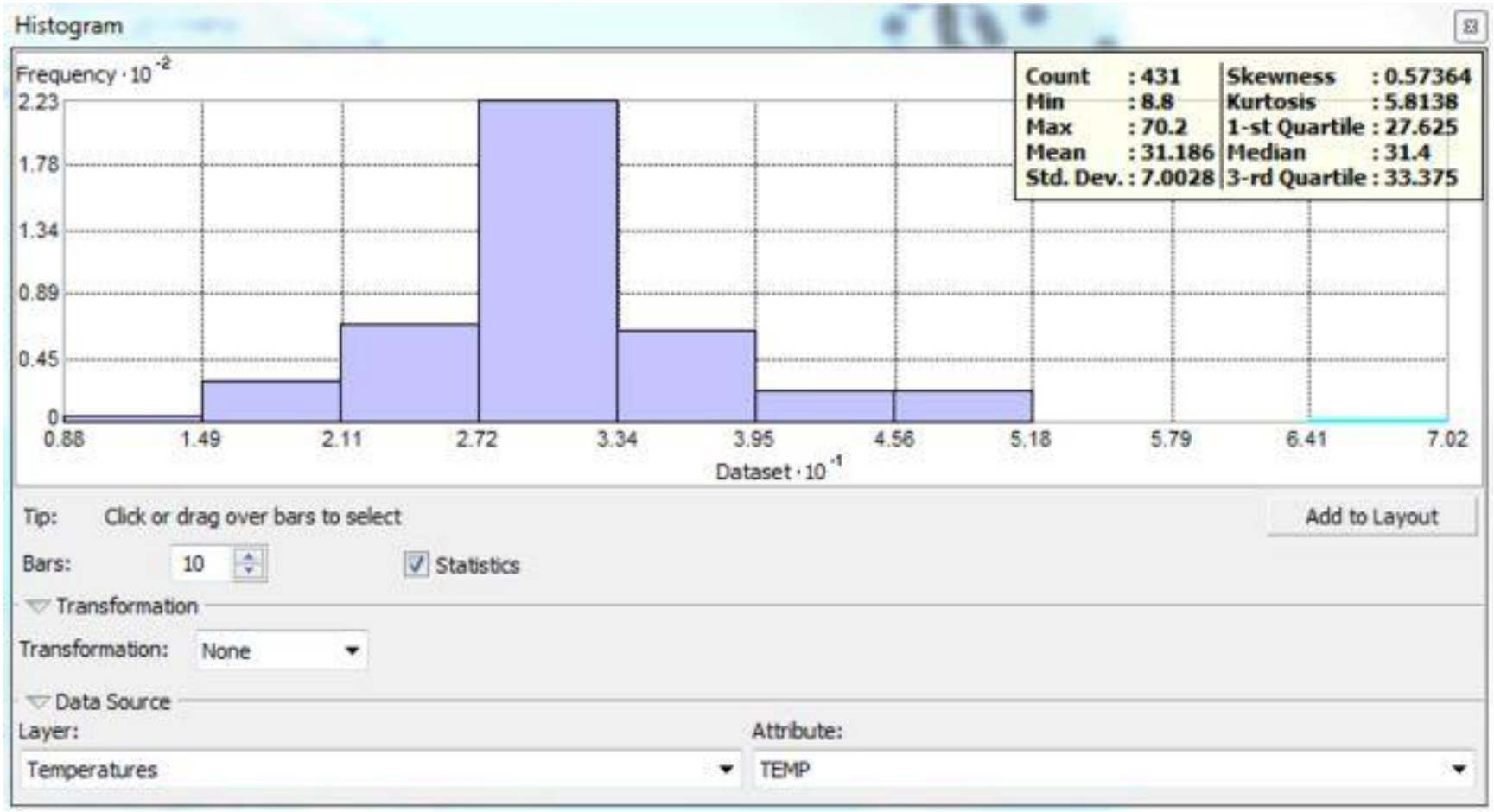
Case: Data summary of European weather stations

Locating the centrality and directional bias



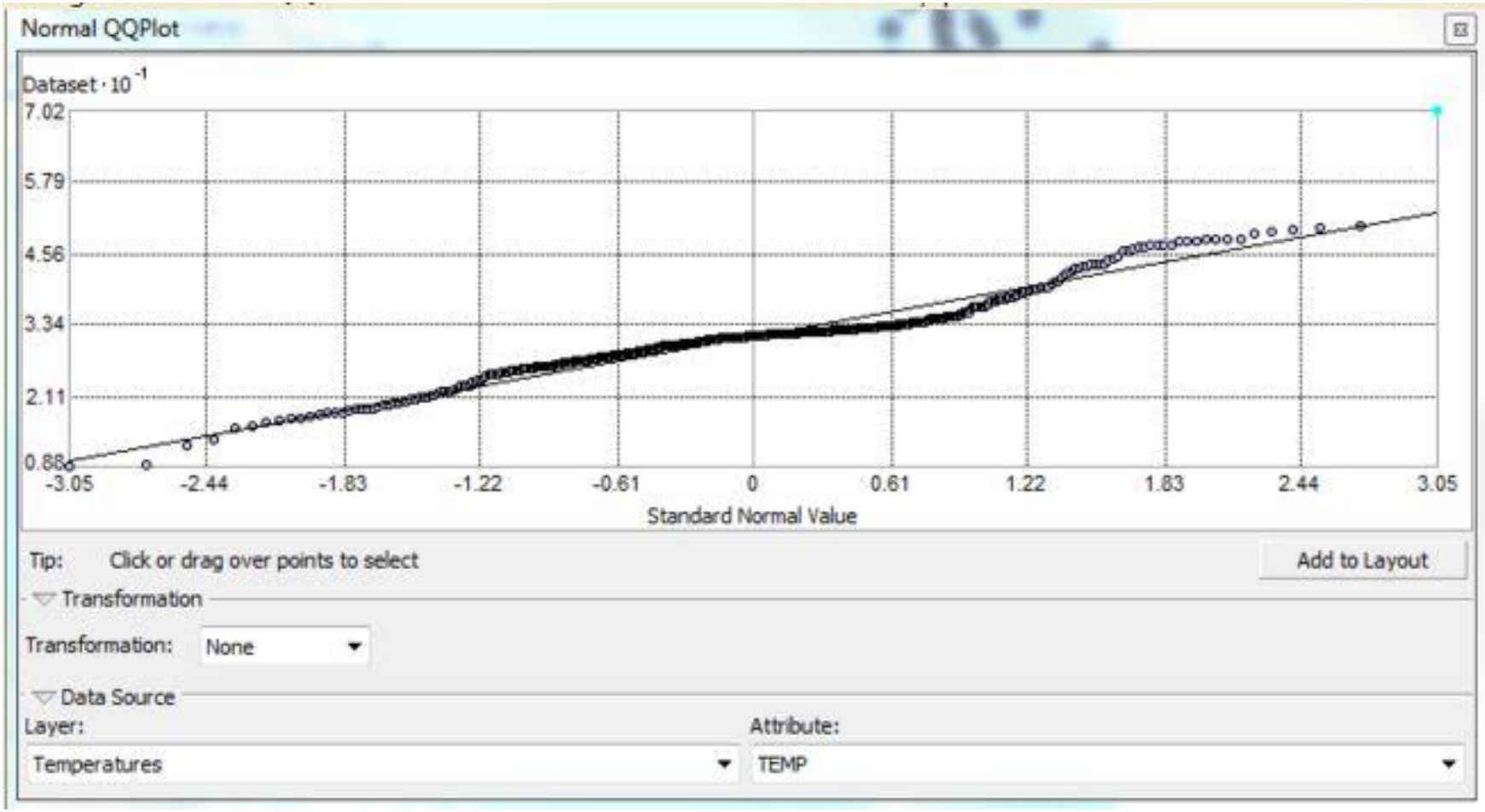
Source

Histogram Analysis of the potential outlier



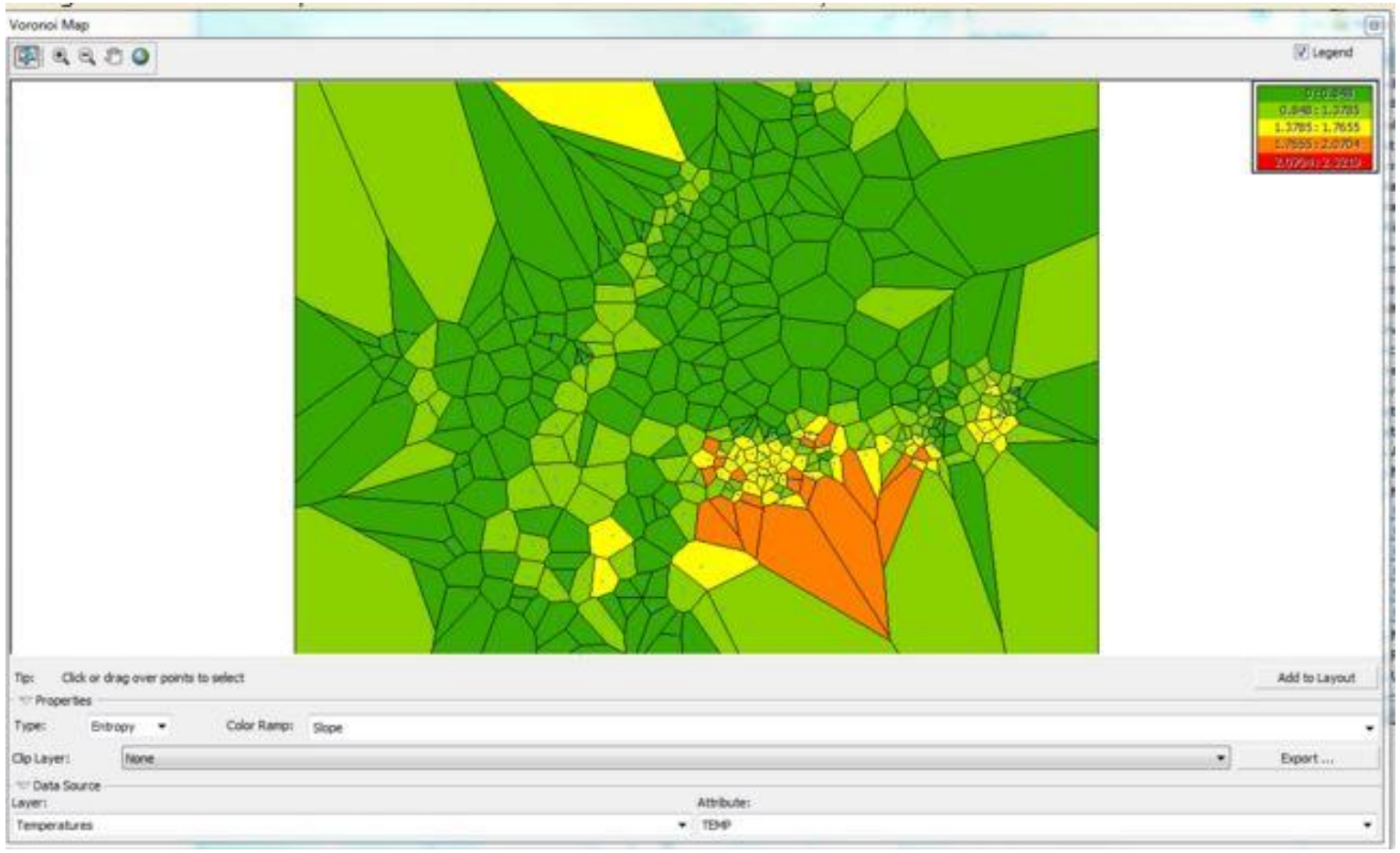
Source

Normalized QQ Plot of the data



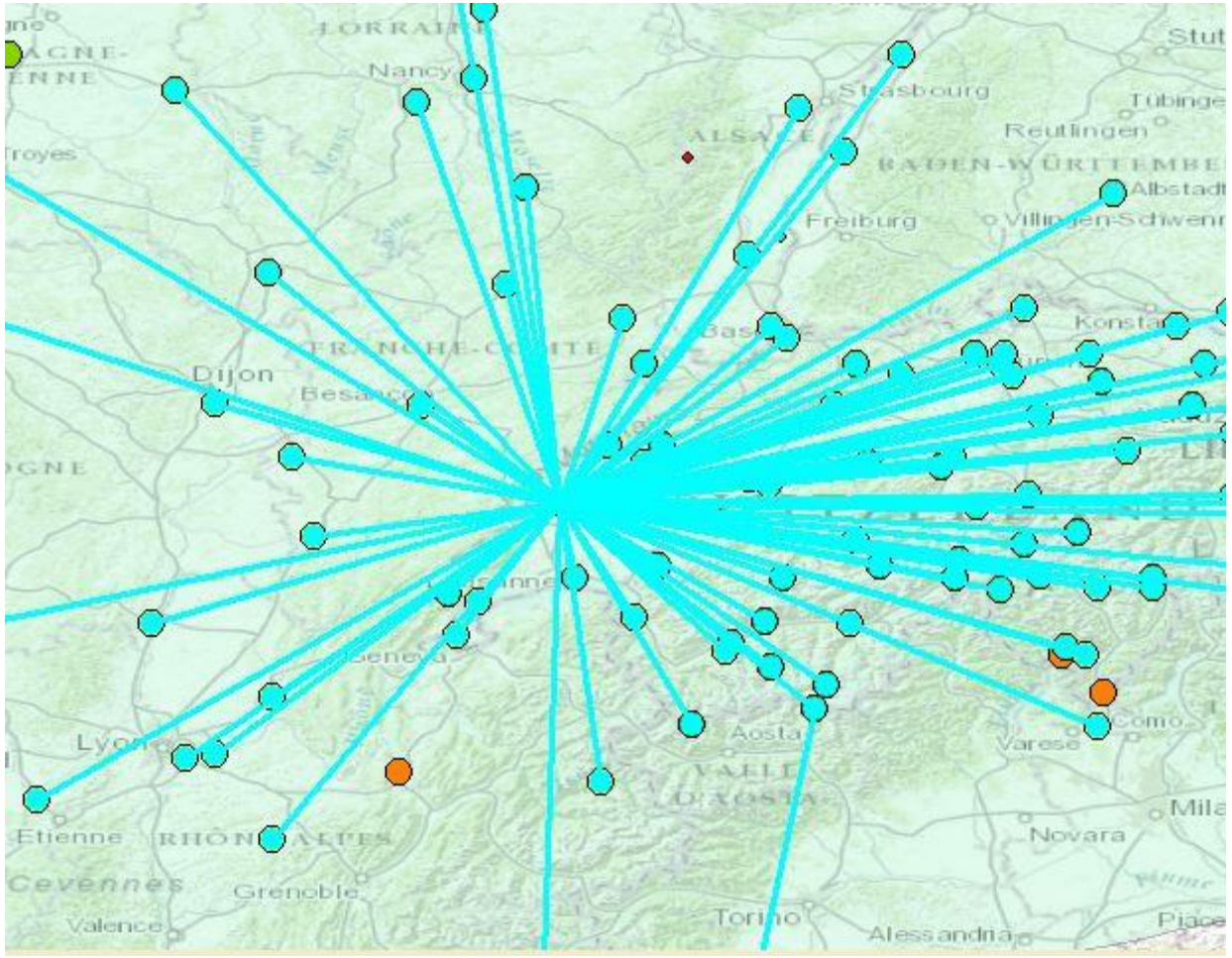
Source

Voronoi histogram showing stationarity of the data



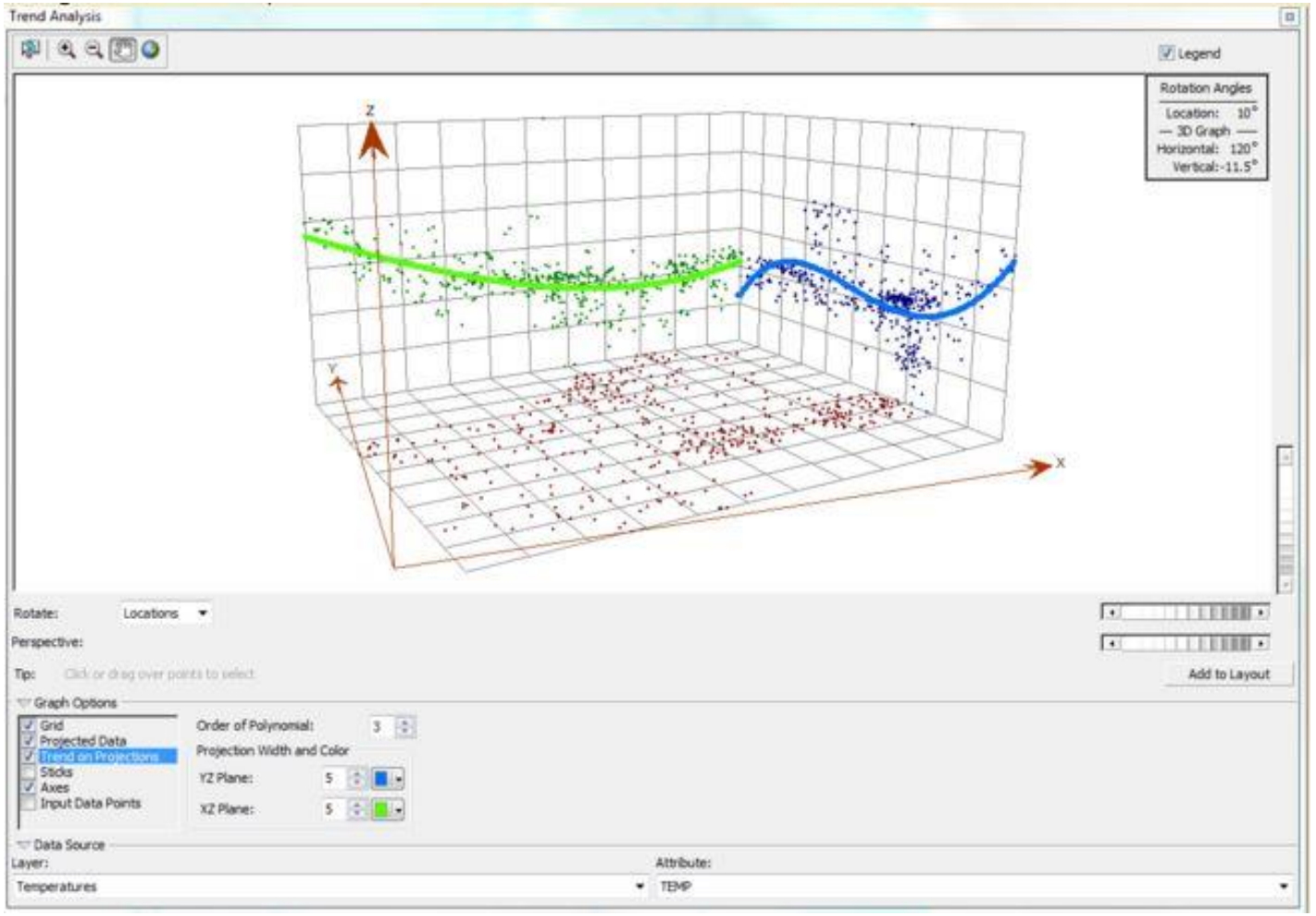
Source

Visualization of the outlier by the semivariogram



Source

Trend analysis for the data showing the third order polynomial

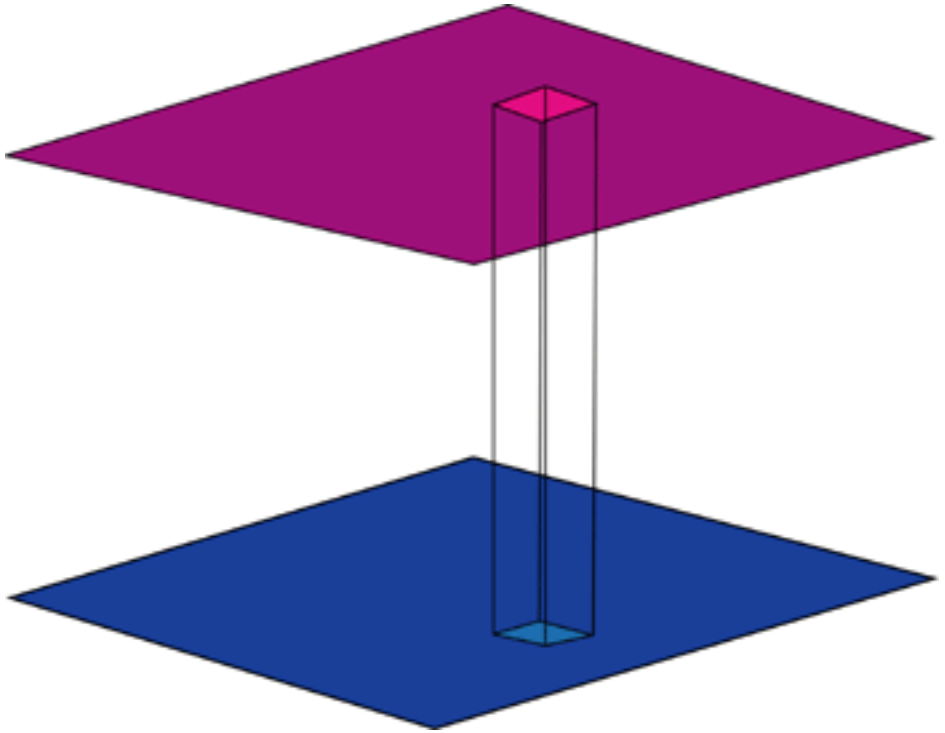


Source

Grid-based statistics

Five types of operations for grids or rasters

- Local
- Focal
- Zonal
- Global
- Application



- Operations applied to each grid cell individually, based on its value.
- Reclassifying land cover types (e.g., assigning new codes to different vegetation classes) for each cell in a raster dataset.

Five types of operations for grids or rasters

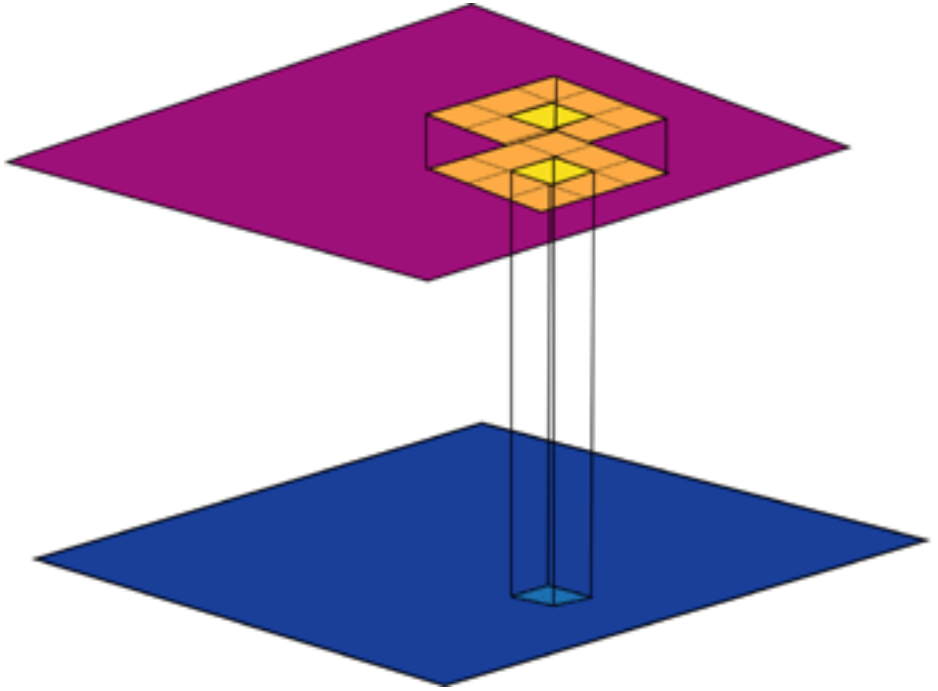
Local

Focal

Zonal

Global

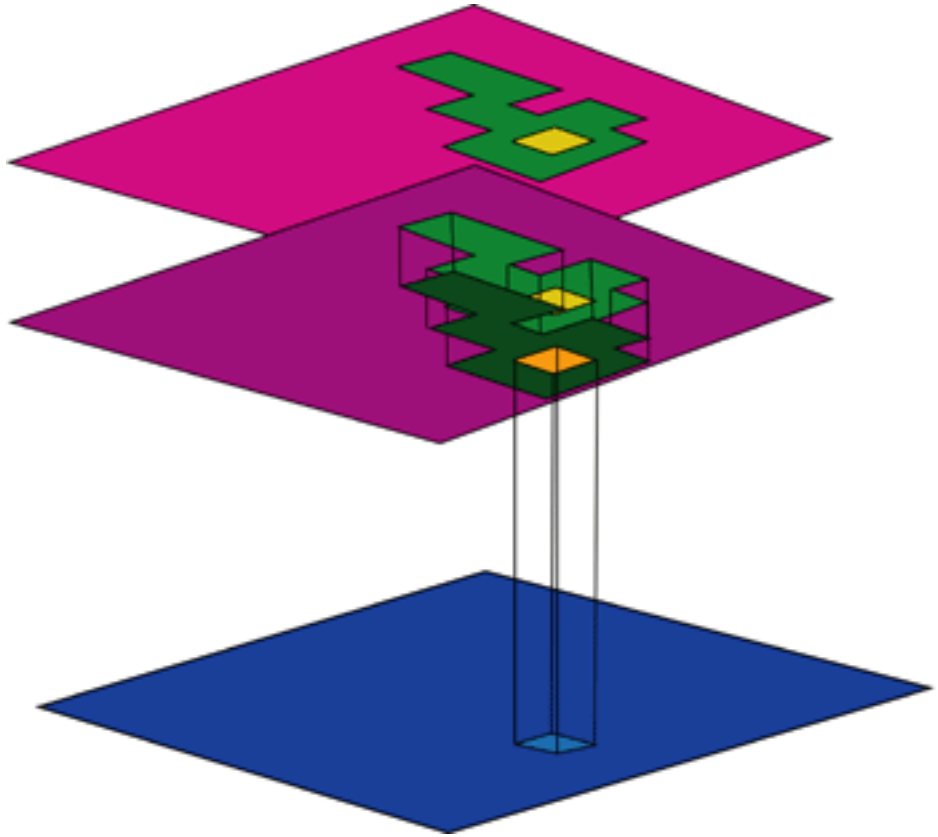
Application



- Operations applied to a cell and its neighbors, using a moving window (e.g., 3x3 or 5x5 grid).
- Smoothing elevation data by calculating the average elevation in a 3x3 window around each cell for noise reduction in a DEM (Digital Elevation Model).

Five types of operations for grids or rasters

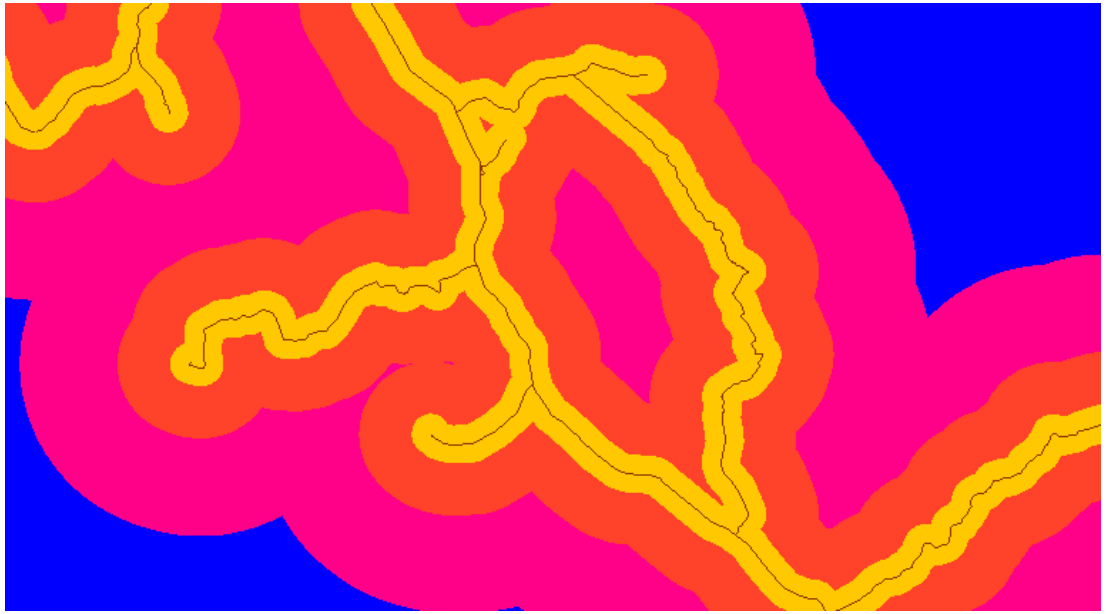
- Local
- Focal
- Zonal
- Global
- Application



- Operations applied to groups of cells that share the same zone or category, aggregating values within those zones.
- Calculating the average temperature for different land-use zones (e.g., forests, urban areas) using a temperature raster.

Five types of operations for grids or rasters

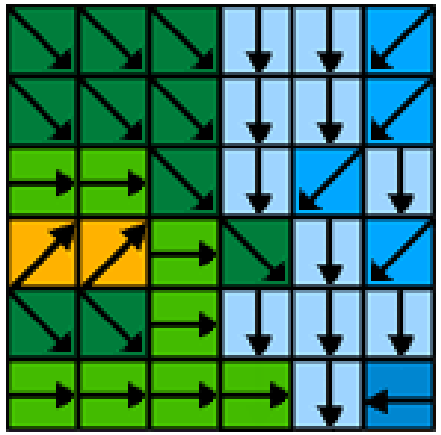
- Local
- Focal
- Zonal
- **Global**
- Application



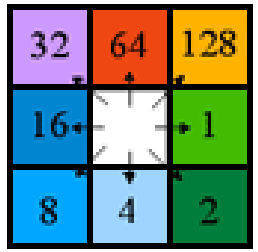
- Operations applied to all grid cells in the dataset, treating the grid as a whole.
- Euclidean distance global operations assign to each cell in the output raster dataset its distance from the closest source cell.

Five types of operations for grids or rasters

- Local
- Focal
- Zonal
- Global
- **Application**



Flow direction



Direction coding

- Combining multiple grids or layers to perform complex calculations, derive new outputs, or use a combination of local, focal, zonal and global operations.
- Flow direction calculation in hydrologic modeling.

Linear spatial filtering: Low pass and high pass

Low pass filter

This filter smooths data by allowing low-frequency components (gradual changes) to pass through while reducing or eliminating high-frequency components (sharp variations). It is used to reduce noise or smooth out fluctuations.

1	1	1
1	1	1
1	1	1

High pass filter

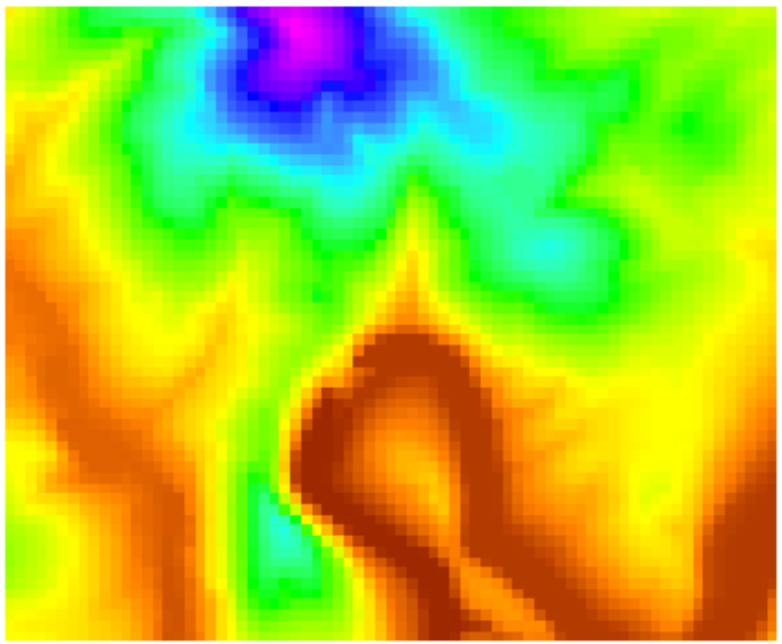
This filter emphasizes rapid changes in values by allowing high-frequency components (edges, sharp transitions) to pass through while reducing or eliminating low-frequency components (smooth, gradual changes). It is used to highlight fine details or edges in images.

-1	-1	-1
-1	16	-1
-1	-1	-1

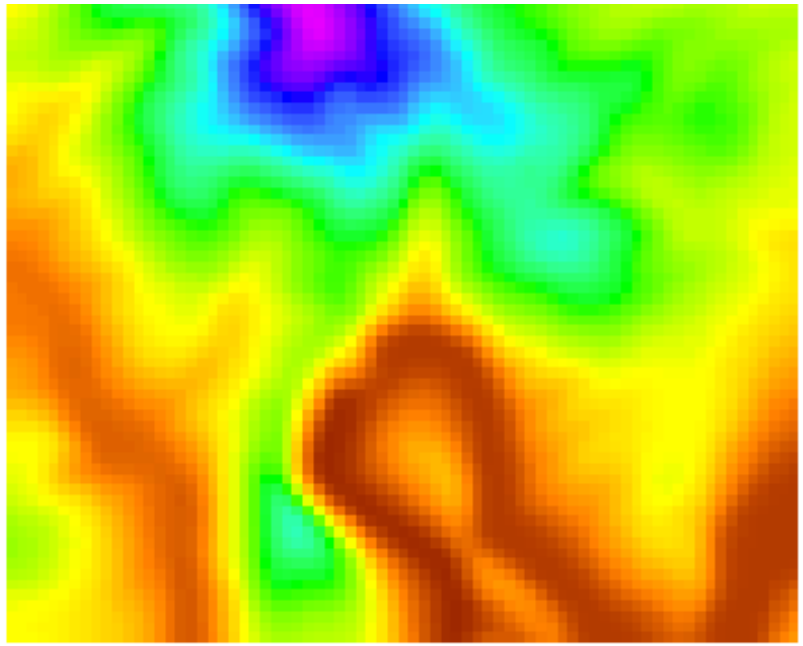
Linear spatial filtering: Low pass and high pass

Smoothing out edge effects

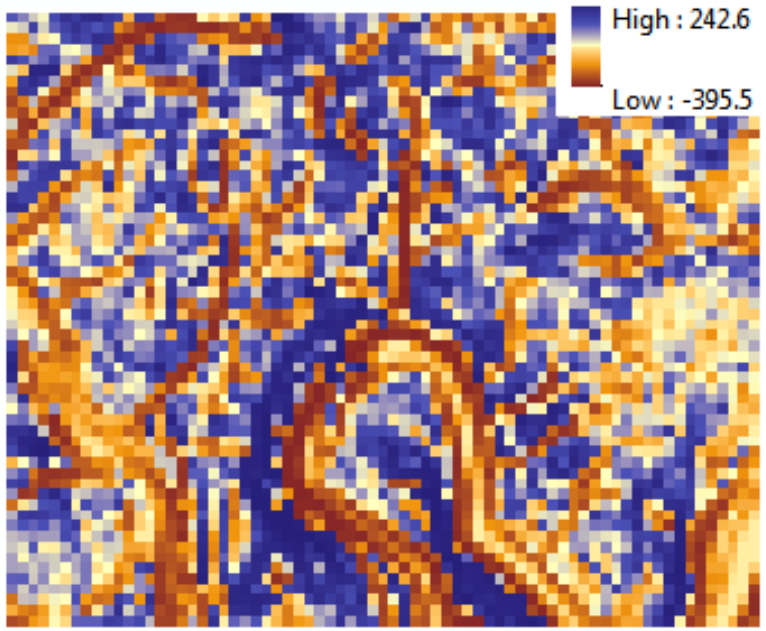
Enhancing edges of streams



Original DEM



low-pass filter 3x3

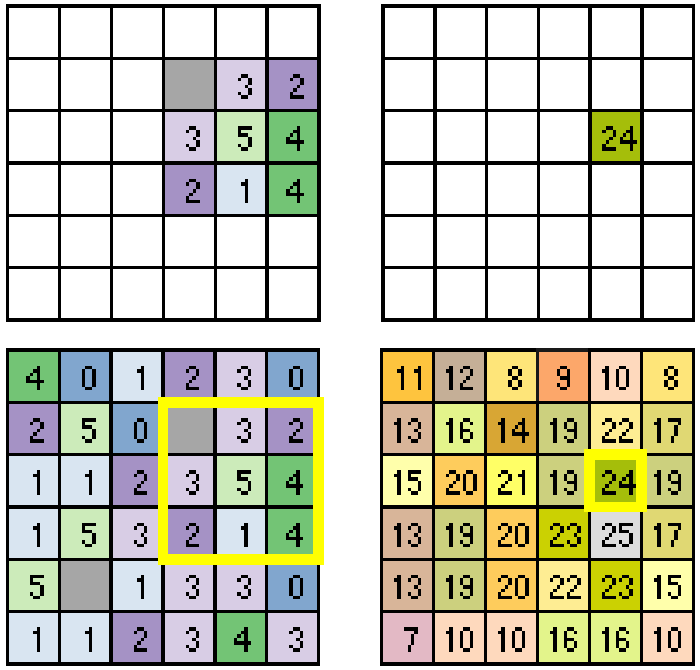


high-pass filter 3x3

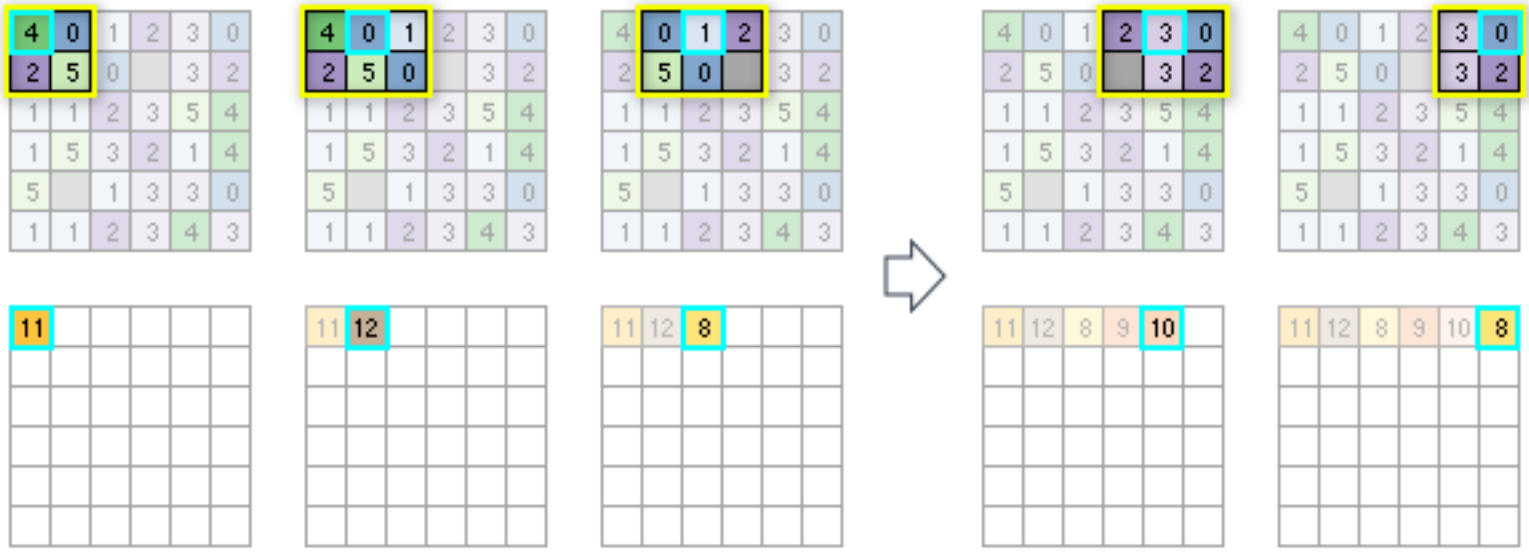
Source

Nonlinear spatial filtering: Focal statistics

Applying a mathematical operation to each cell in a raster based on the neighboring values within a moving window.



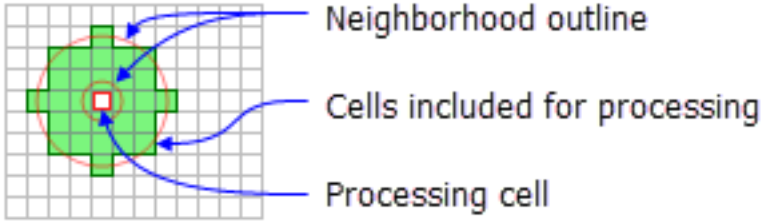
Figuring out the values in edge involves special consideration. Multiple strategies are available to handle the corner and/or edges of grids.



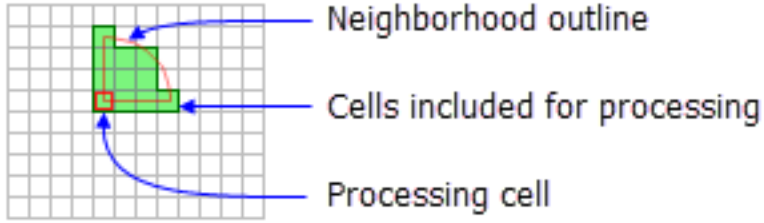
Source

Nonlinear spatial filtering: Focal statistics neighborhood concepts

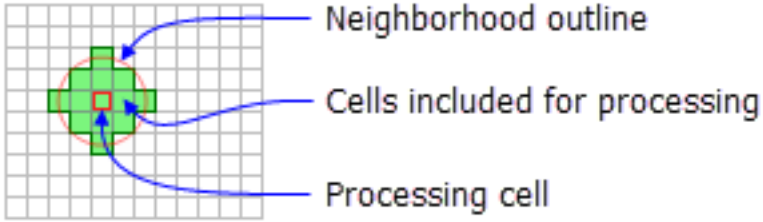
Annulus



Wedges



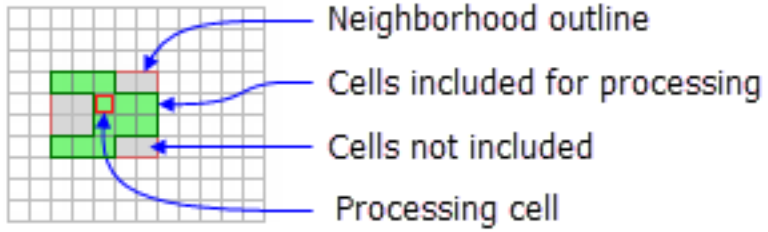
Circle



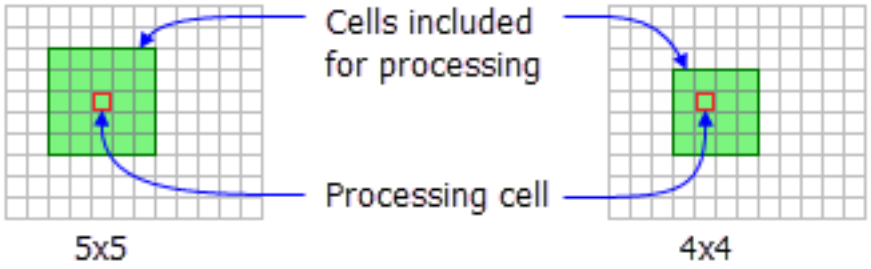
Irregular

Irregular kernel

```
5 4
1 1 1 0 0
0 0 1 1 1
0 0 1 1 1
1 1 1 0 0
```



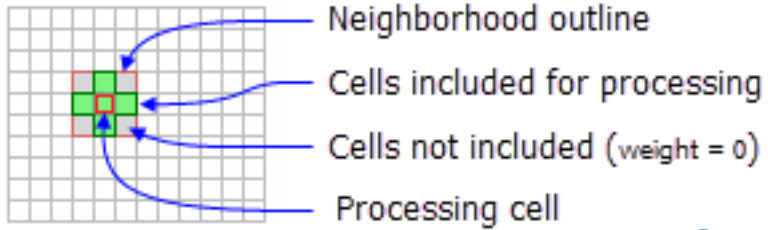
Rectangle



Weight

Weighted kernel

```
 3 3
 0    -0.25  0
-0.25  2.00 -0.25
 0    -0.25  0
```

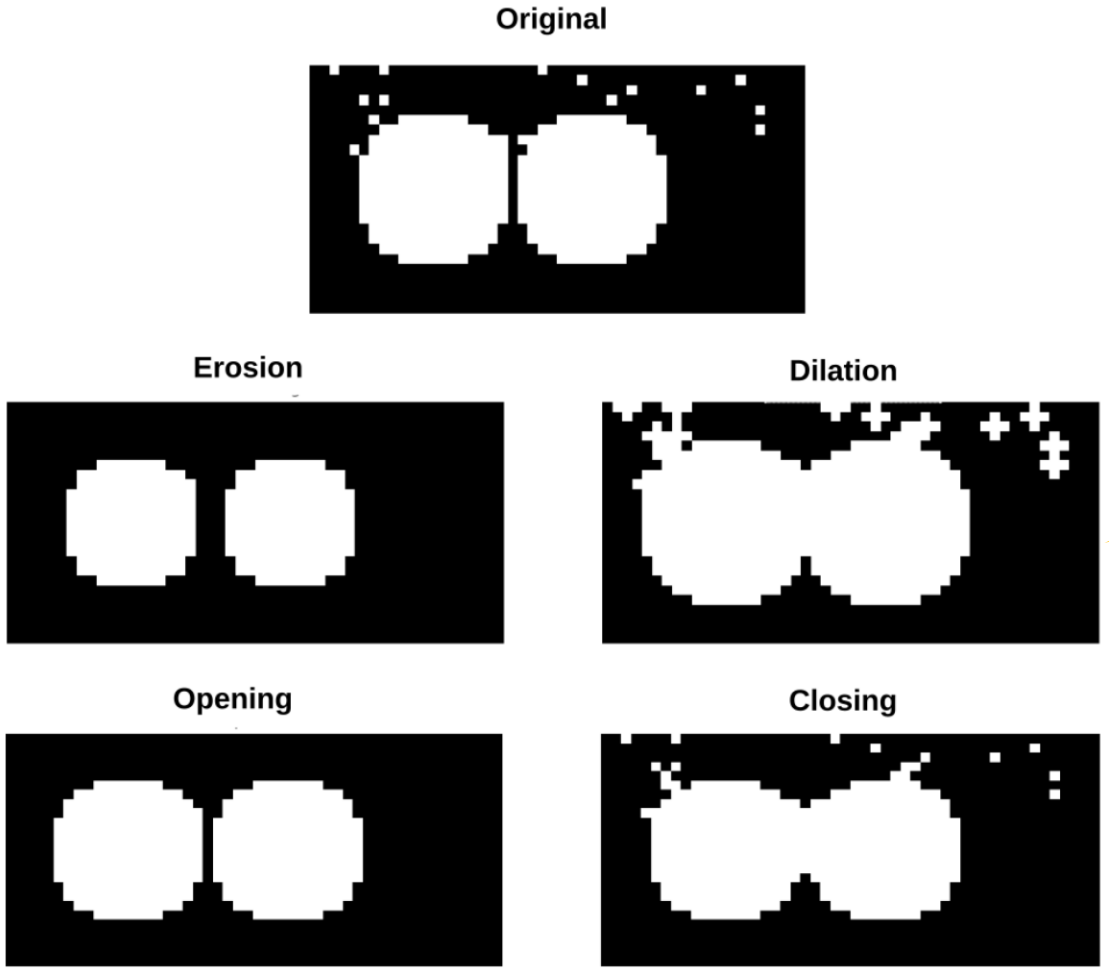


Source

Nonlinear spatial filtering: Different statistics can be calculated in focal analysis

Statistics	Input Data Type	Output Data Type
Majority	Integer	Integer
Maximum	Integer/Float	Integer/Float
Mean	Integer/Float	Float
Median	Integer/Float	Float
Minimum	Integer/Float	Integer/Float
Minority	Integer	Integer
Percentile	Integer/Float	Float
Range	Integer/Float	Integer/Float
Standard Deviation	Integer/Float	Float
Sum	Integer/Float	Integer/Float
Variety	Integer	Integer

Morphological operations



Shrinks or thins objects in a binary image by removing pixels at the boundaries.

Expands or thickens objects in a binary image by adding pixels to the boundaries.

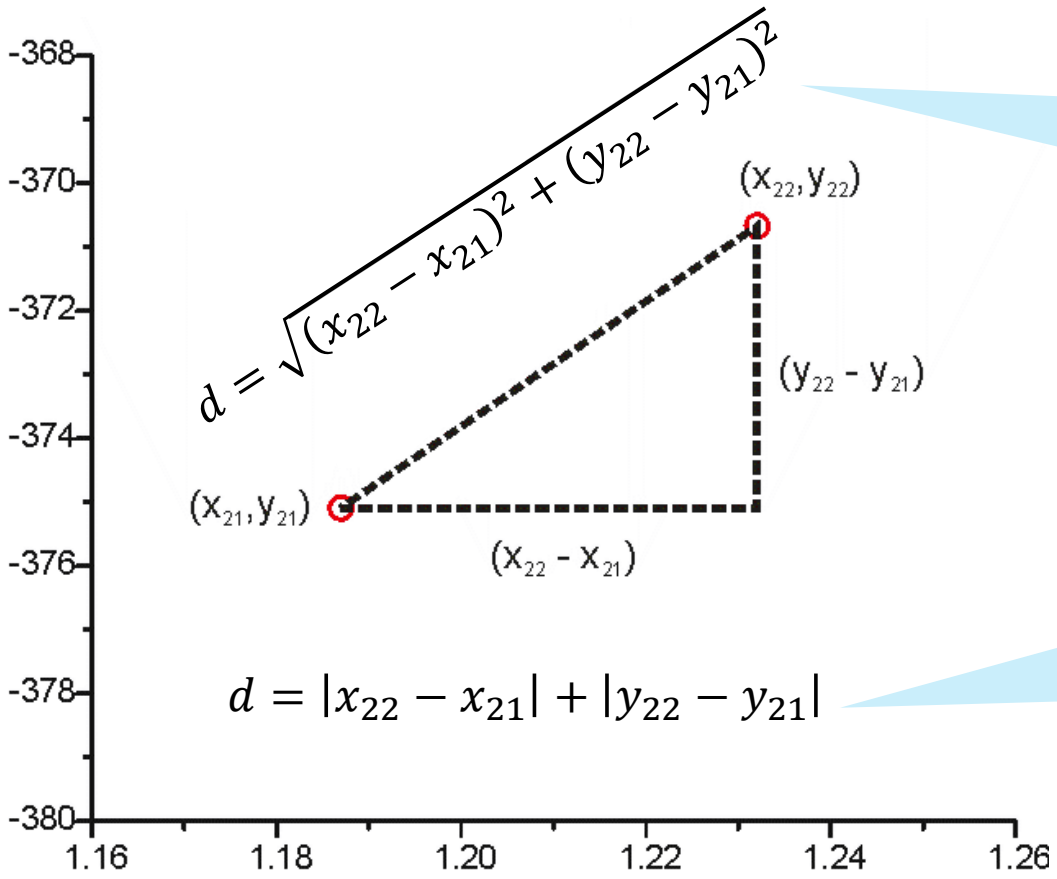
A combination of erosion followed by dilation, used to remove small objects or noise while preserving the overall shape.

A combination of dilation followed by erosion, used to fill small gaps or holes in objects while maintaining their structure.

Source

Point Sets and Distance Statistics

Basic distance derived statistics



Euclidean Distance (L2 Norm)
The straight-line distance between two points in space, calculated using the Pythagorean theorem.

Manhattan Distance (L1 Norm)
The distance between two points measured along the axes at right angles. It sums the absolute differences of their coordinates.

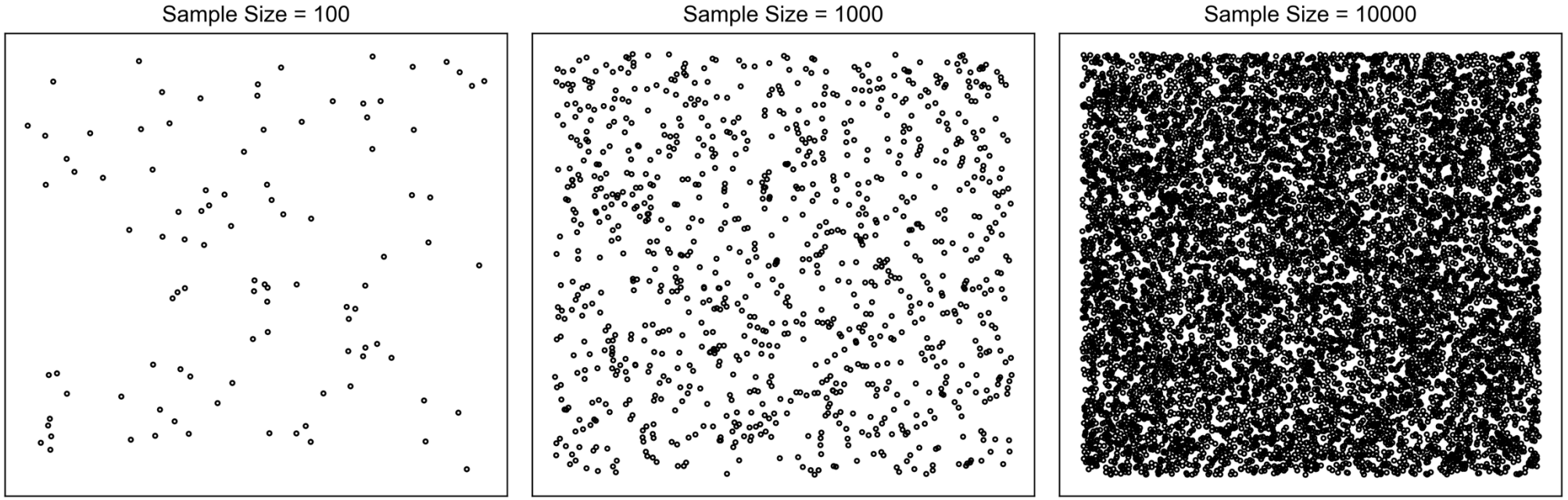
Source

Other less common distance measures

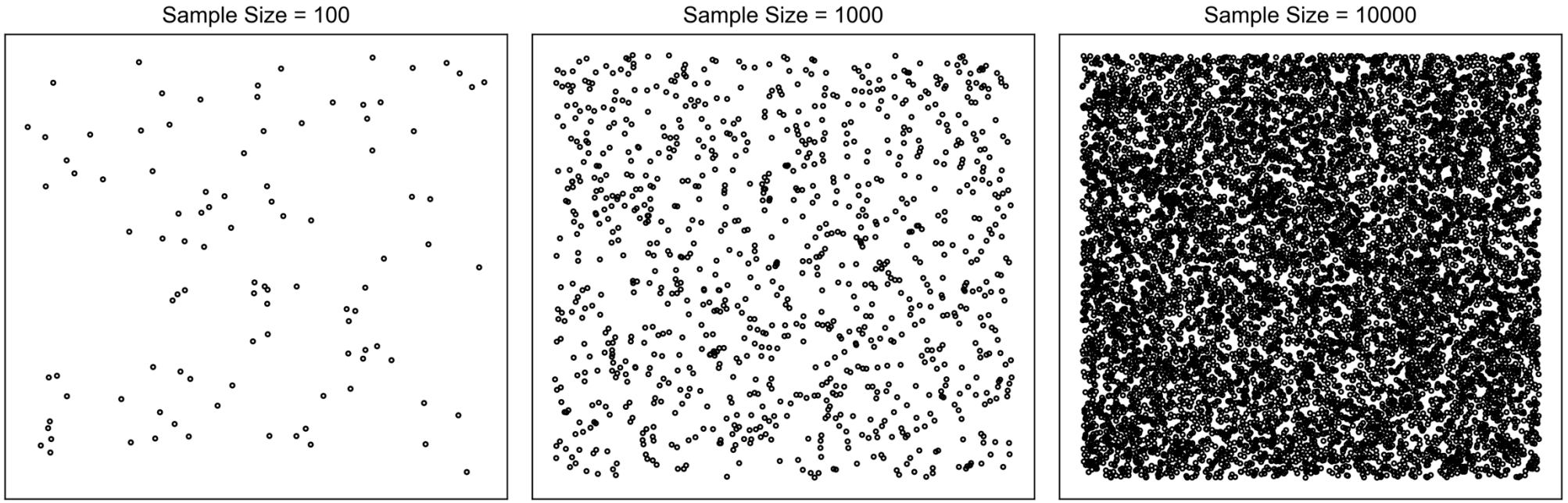
Distance Metrics	Explanation	Formula
Minkowski Distance	A generalization of both Euclidean and Manhattan distances, controlled by a parameter p. It is Euclidean when $p = 2$ and Manhattan when $p = 1$.	$\left(\sum \lvert x_i - y_i \rvert^p\right)^{1/p}$
Haversine Distance	Used to calculate the shortest distance between two points on the Earth's surface, considering the curvature of the Earth.	$2r \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\theta}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\Delta\lambda}{2}\right)}\right)$
Mahalanobis Distance	Measures the distance between a point and a distribution, considering correlations between variables, often used in multivariate analysis.	$\sqrt{(x-\mu)^T S^{-1}(x-\mu)}$
Canberra Distance	A weighted version of Manhattan distance, giving more importance to smaller differences.	$\sum \frac{\lvert x_i - y_i \rvert}{\lvert x_i \rvert + \lvert y_i \rvert}$
Cosine Similarity	Measures the cosine of the angle between two vectors, often used to compare directional similarity in high-dimensional spaces.	$1 - \frac{x \cdot y}{\lVert x \rVert \cdot \lVert y \rVert}$
Jaccard Distance	Measures dissimilarity between two sets by comparing the size of their intersection to their union.	$1 - \frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$

Complete Spatial Randomness

- CSR refers to a spatial distribution where point events occur independently of each other, and their locations are equally likely to occur anywhere within a given study area.
- The number of events in any sub-region follows a Poisson distribution: $P(N = k) = \frac{(\lambda A)^k e^{-\lambda A}}{k!}$
- Here, $P(N = k)$ is the probability of observing k events; λ is the average number of events per unit area; A is the area under consideration.



Complete Spatial Randomness



If you know the number of events (points) in a defined area with fixed resolution, you can calculate the expected location of the event in an area assuming the hypothesis of Complete Spatial Randomness or CSR.

Different point pattern analysis methods

- Average nearest neighbor
- Spatial autocorrelation
- High/Low clustering (Getis-Ord)
- Incremental spatial autocorrelation
- Multi-distance spatial cluster analysis (Ripley’s K)

Average Nearest Neighbor

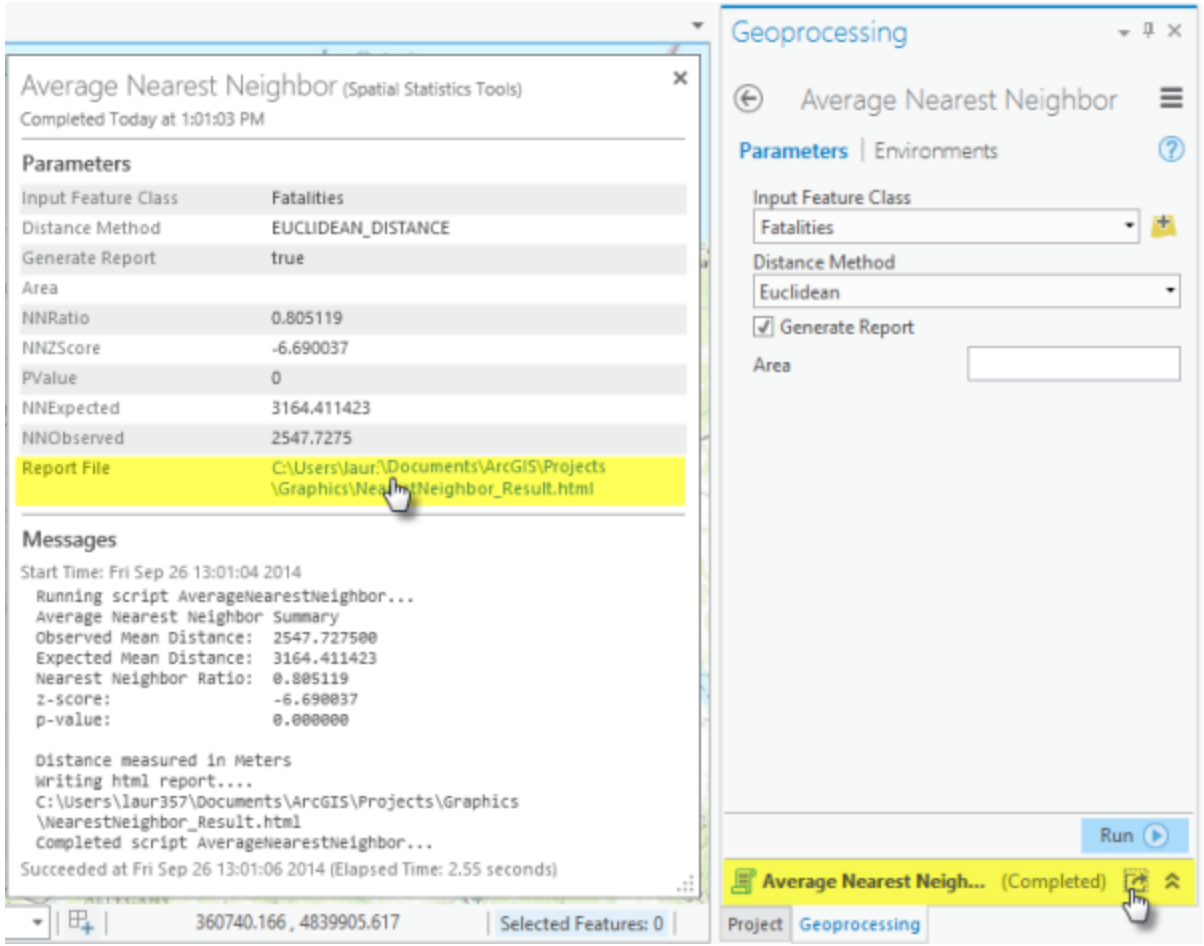
Measures the average distance between each point in a dataset and its closest neighboring points. Calculated as Nearest Neighbor Ratio (NNR).

$$NNR = \frac{\overline{D}_o}{\overline{D}_E} \qquad \overline{D}_o = \frac{\sum_{i=1}^{n} d_i}{n} \qquad \overline{D}_E = \frac{0.5}{\sqrt{n/A}}$$

$\overline{D}_o$ is the observed mean distance, $\overline{D}_E$ is the expected mean distance in a random pattern, d_i is the distance between feature i and its nearest neighboring feature, n corresponds to the total number of features and A is the area of a minimum enclosing rectangle around all features.

$$z = \frac{\overline{D}_o - \overline{D}_E}{SE} \qquad SE = \frac{0.26136}{\sqrt{n^2/A}}$$

Average Nearest Neighbor

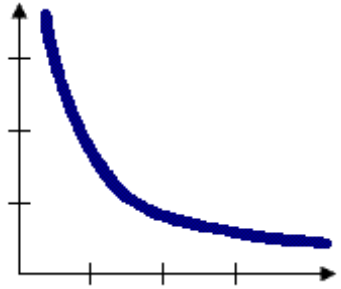


- $NNR < 1$: Indicates clustering.
- $NNR \approx 1$: Suggests a random distribution.
- $NNR > 1$: Indicates dispersion.

- z score tells us how far the observed pattern deviates from randomness.
- p value indicates if the deviation is statistically significant or not.

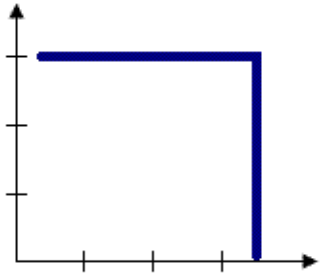
Understanding the spatial weights in patterns: Conceptualization of spatial relationships

Inverse Distance



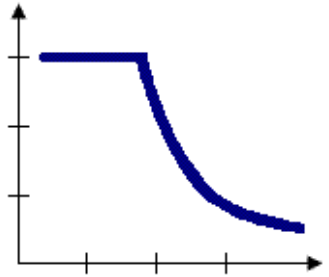
- All features impact or influence all other features, but the farther away something is, the smaller the impact it has.
- Appropriate for modeling continuous data.

Distance Band



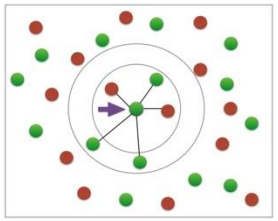
- Impose a sphere of influence or moving window conceptual model of spatial interactions.
- If you know that the average journey to work is 15 miles, then pick 15 miles as distance band.

Zone of Indifference



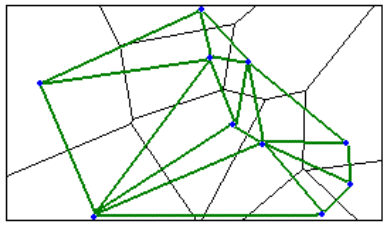
- Combination of Inverse distance and distance band.
- Features within the distance band are included.
- Once the distance is exceeded, the level of influence drops off.

K nearest neighbors



- If K (the number of neighbors) is 8, the eight closest neighbors to the target feature will be included in computations for that feature.

Delaunay Triangulation

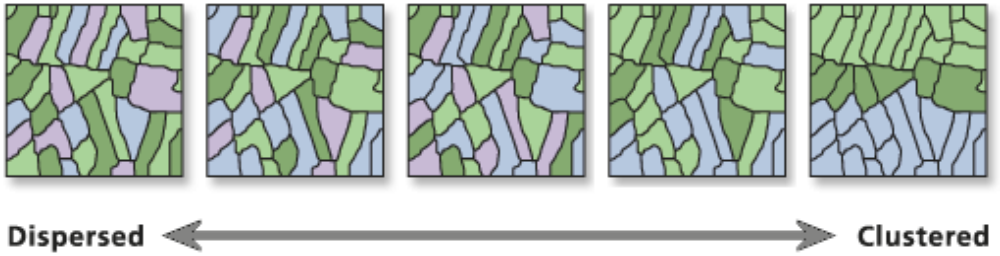


- Constructs neighbors by creating Voronoi triangles from point features or from feature centroids such that each point or centroid is a triangle node.

Source

Global Moran's I

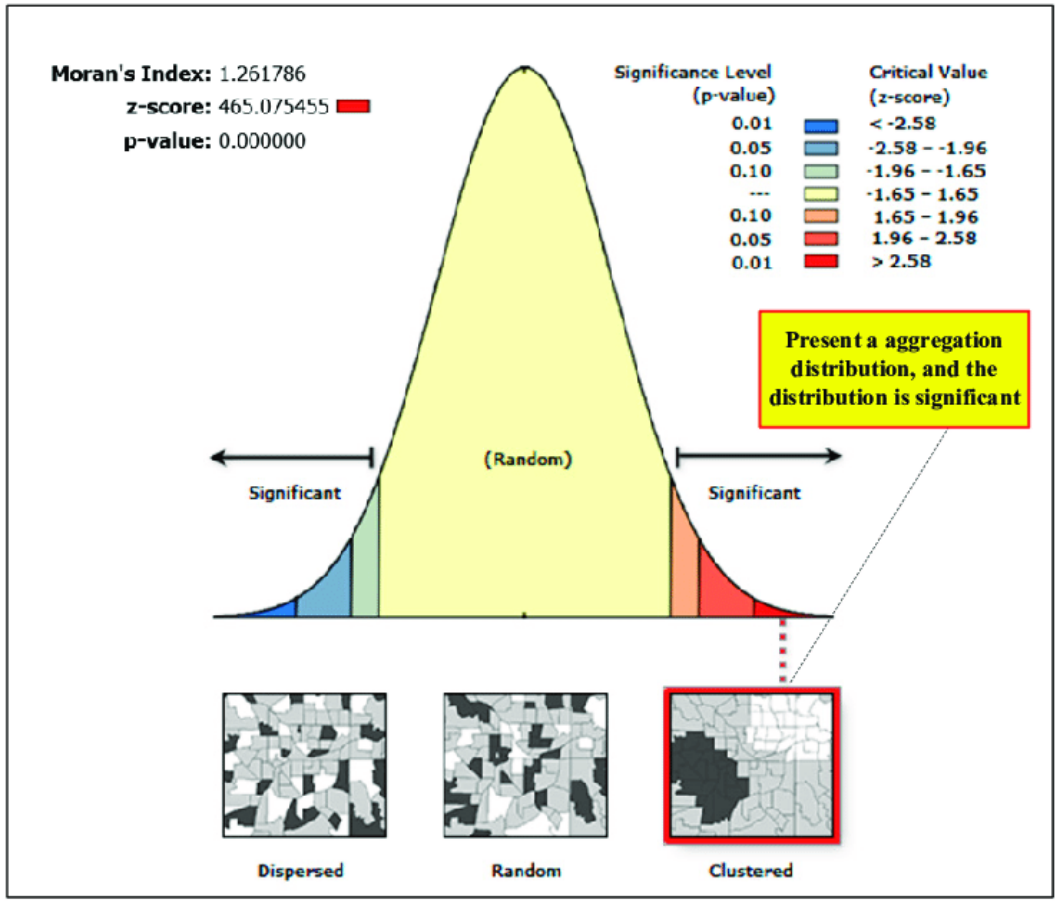
Measures spatial autocorrelation based on feature locations and attribute values using the Global Moran's I statistic.



$$I = \frac{N \sum_i \sum_j w_{ij}(x_i - \bar{x})\left(x_j - \bar{x}\right)}{W \sum_i (x_i - \bar{x})^2}$$

N is the total number of spatial units; x_i and x_j are the attribute values at location i and j; $\bar{x}$ is the mean of the attribute values; w_{ij} is the spatial weight, W is the sum of all spatial weights ($\sum_i \sum_j w_{ij}$)

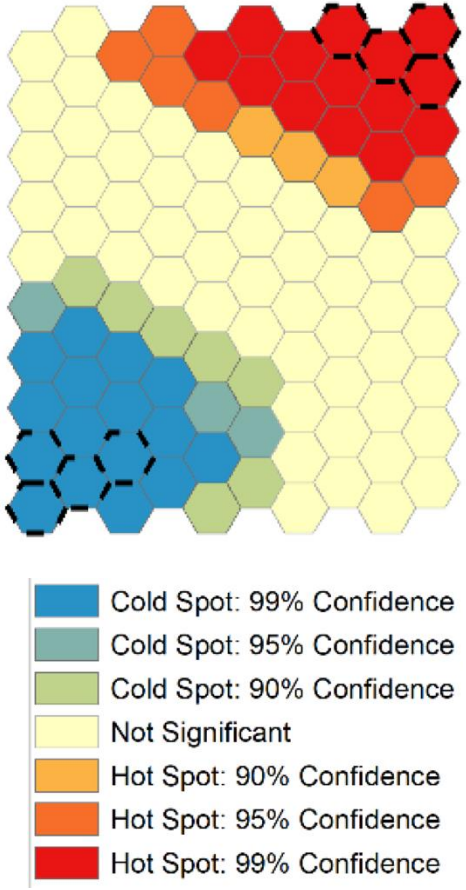
Global Moran's I: Interpretation



Source

- Moran's I value
  - $I > 0$: Positive spatial autocorrelation, similar values are clustered together.
  - $I < 0$: Negative spatial autocorrelation, dissimilar values are clustered together, dispersed pattern.
  - $I \approx 0$: Random spatial distribution
- z score and p value
  - p value insignificant, forget about z score, random
  - p value significant
    - z score positive: clustered
    - z score negative: dispersed

High/Low clustering or Getis-Ord



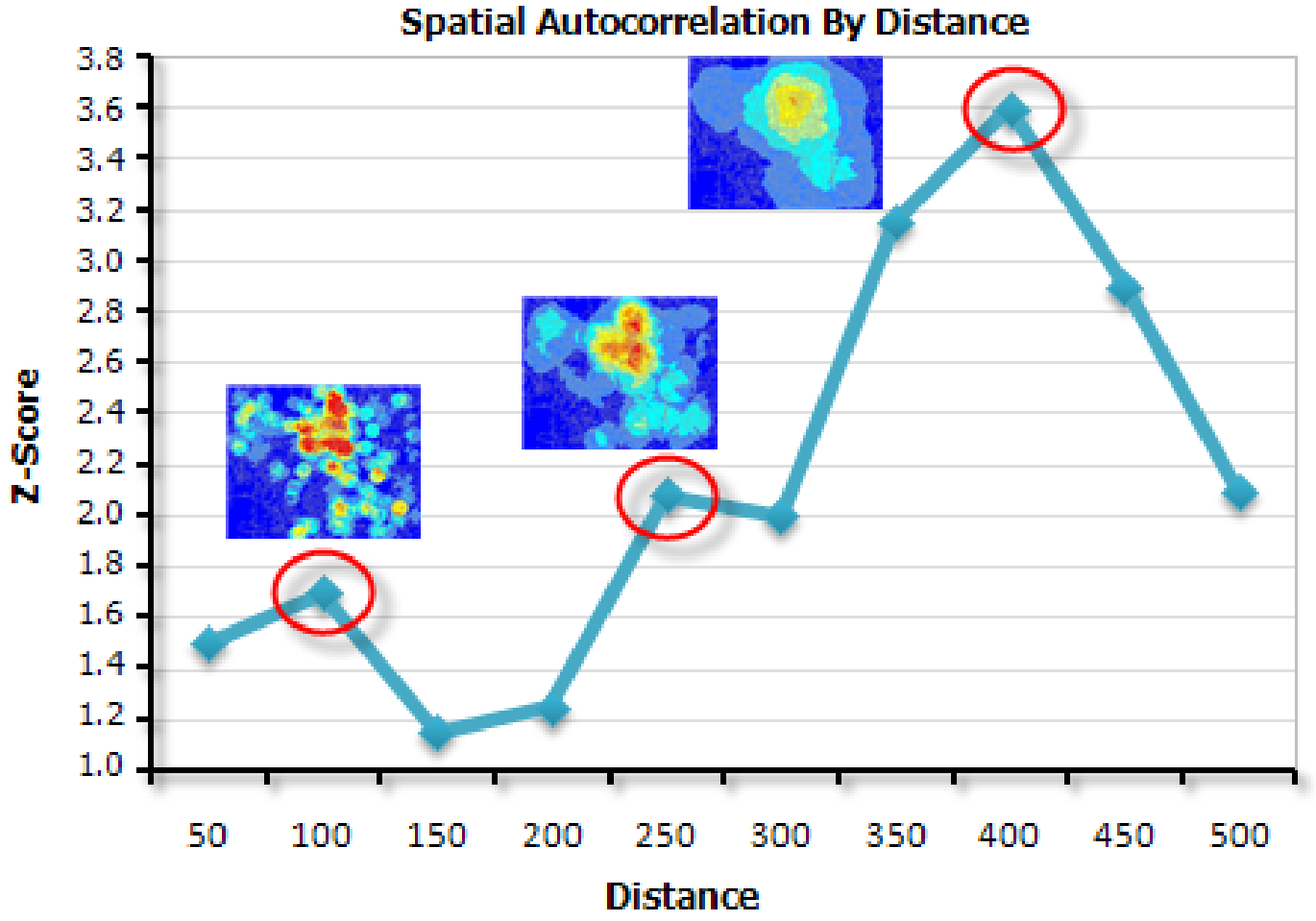
Source

- Getis-Ord G_i^* (or G-statistic) identifies **hot spots** (cluster of high values) or **cold spots** (cluster of low values) in geographic data.
- It measures the degree of clustering for either high or low values, revealing areas with unusually high or low concentrations relative to the overall dataset.
- $$G_i^* = \frac{\sum_j w_{ij}x_j - \bar{x}\sum_j w_{ij}}{S\sqrt{\frac{\left(N\sum_j w_{ij}^2\right) - \left(\sum_j w_{ij}\right)^2}{N-1}}}$$
- Here
  - i is the point of interest
  - x_j is the attribute value at location j
  - w_{ij} is the spatial weight between i and j
  - N is the total number of spatial units
  - $\bar{x}$ is the mean and S is the standard deviation of the attribute values

$$Z(G_i^*) = \frac{G_i^* - E(G_i^*)}{Std(G_i^*)}$$

- (+)ve Z: Hot Spots
- (-)ve Z: Cold Spots

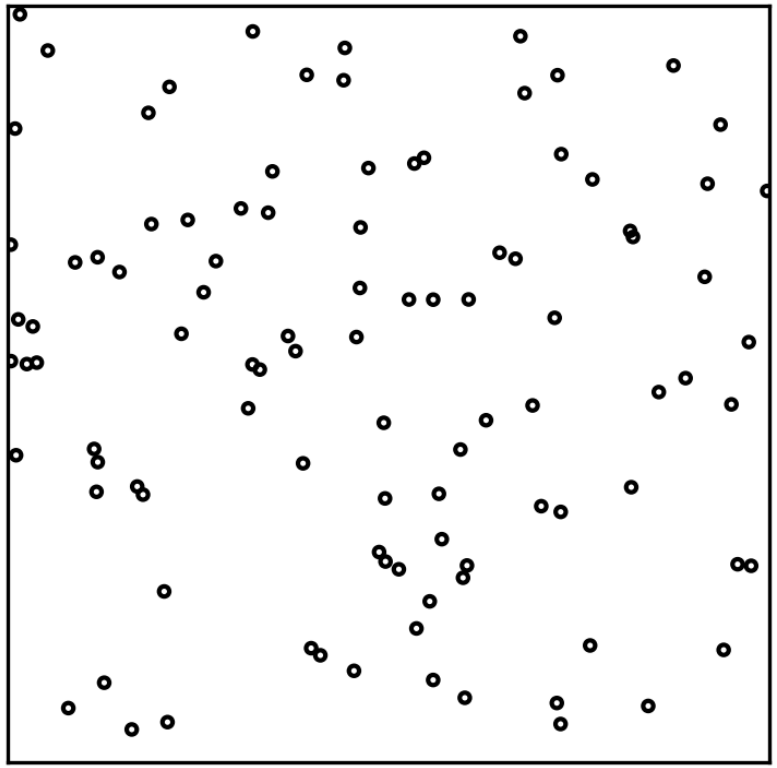
Incremental Spatial Autocorrelation



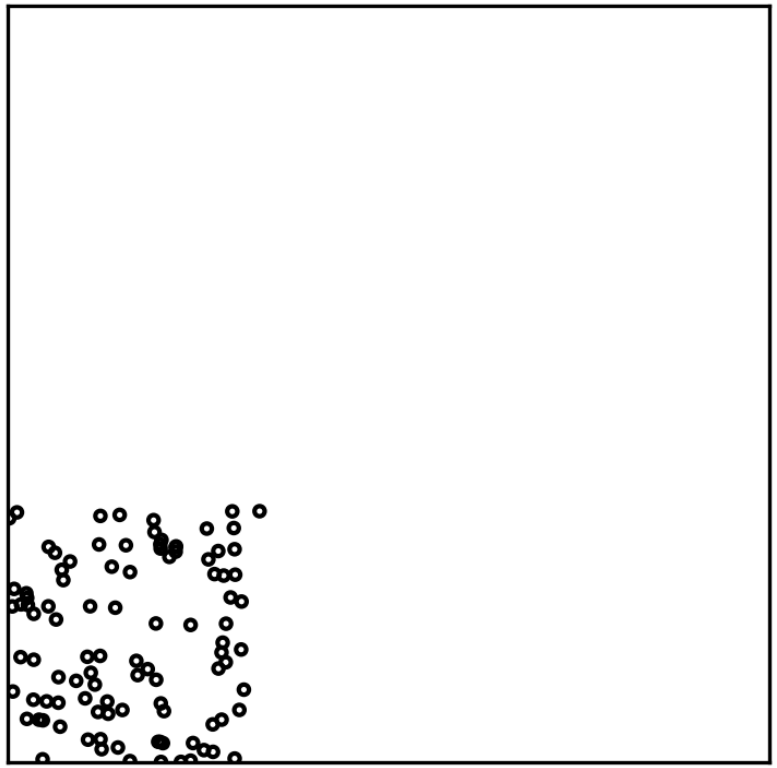
- Measures spatial autocorrelation for a series of distance increments and reports, the associated **Moran's Index**, **Expected Index**, **Variance**, **z-score** and **p-value**.
- Useful to find an appropriate **Distance Threshold** or **Radius parameter** value.
- When more than one statistically significant peak is present, clustering is pronounced at each of those distances. Select the peak distance that best corresponds to the scale of analysis you are interested in; often this is the first statistically significant peak encountered.

Multi-Distance Spatial Cluster Analysis (Ripley's K Function)

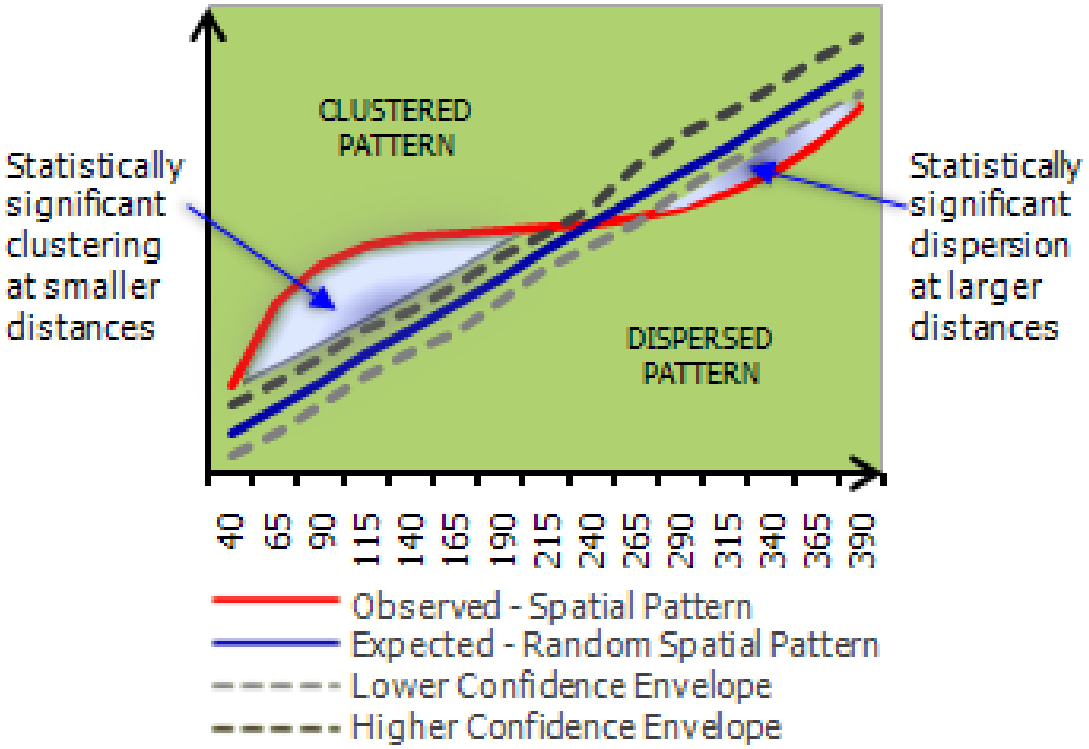
Points look random at this scale



If we increase our scale, the points look clustered



Multi-Distance Spatial Cluster Analysis (Ripley's K Function)



Source

- Ripley's K-Function
- $L(d) = \sqrt{\frac{A \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} k_{i,j}}{\pi n(n-1)}}$
- Here, d is the distance, n is the number of features, A is the total area of features, $k_{i,j}$ is weight

SAINT LOUIS UNIVERSITY.

Thank You