Spatial analysis in public health applications

Siddhi Munde, MSc

Research Data Analyst

PhD student – MSIA

Department of Environmental, Agricultural & Occupational Health, COPH,UNMC

University of Nebraska Medical Center

Topics

- Introduction to GIS – data types and mapping
- Hotspot analysis methods
  - Local Moran’s I statistics
  - Getis Ord G* statistics
- Space-time analysis in ArcGIS
- Spatial correlation in R

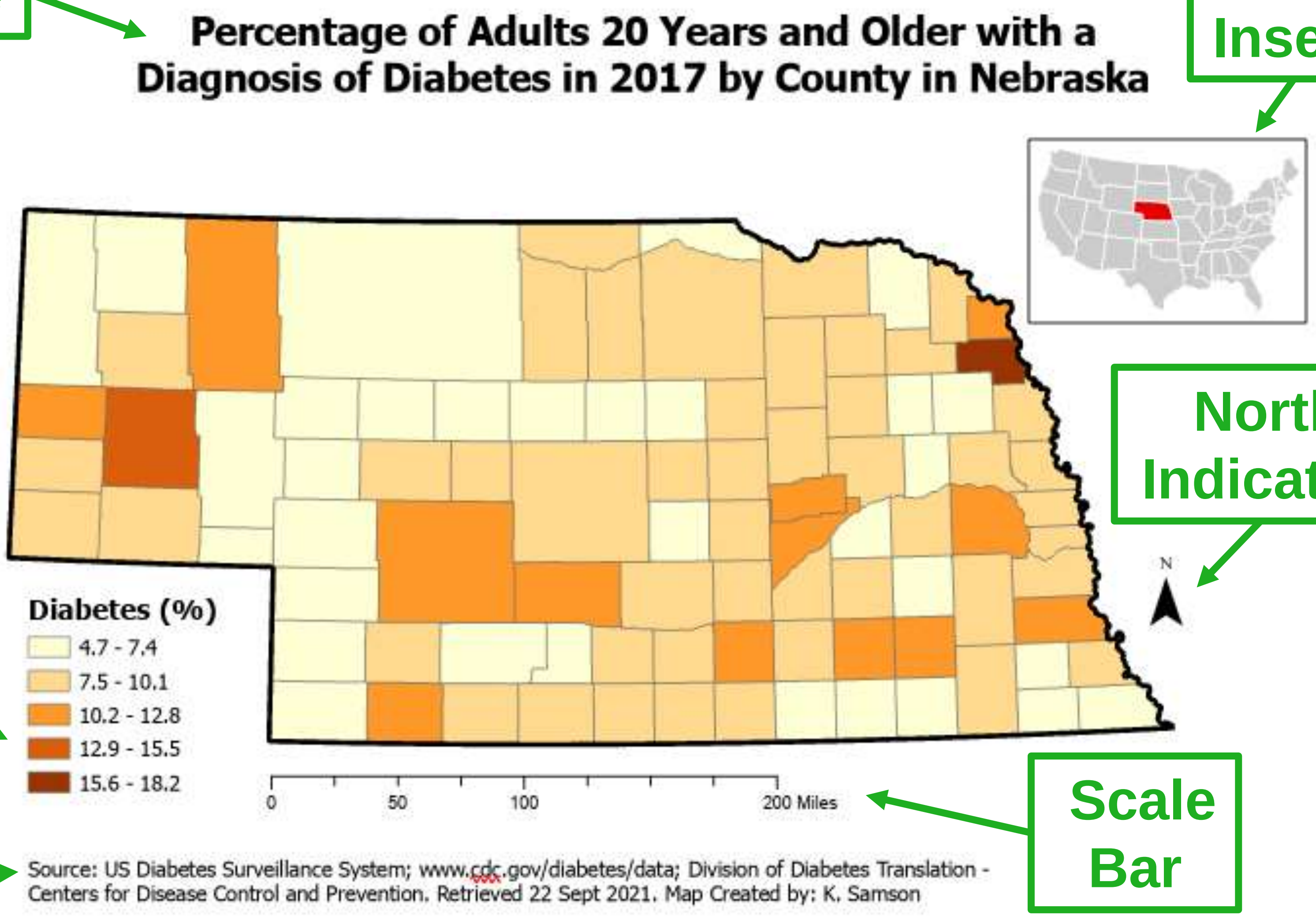
What is GIS?

“GIS, or **geographic information systems**, are computer-based tools used to store, visualize, analyze, and interpret geographic data. Geographic data (also called spatial, or geospatial data) identifies the geographic location of features.”

-CDC

Source: https://www.cdc.gov/gis/what-is-gis.htm

Mapping - Map Layout Elements



Percentage of Adults 20 Years and Older with a Diagnosis of Diabetes in 2017 by County in Nebraska

Source: US Diabetes Surveillance System; www.cdc.gov/diabetes/data; Division of Diabetes Translation - Centers for Disease Control and Prevention. Retrieved 22 Sept 2021. Map Created by: K. Samson

Title

Inset

North Indicator

Legend

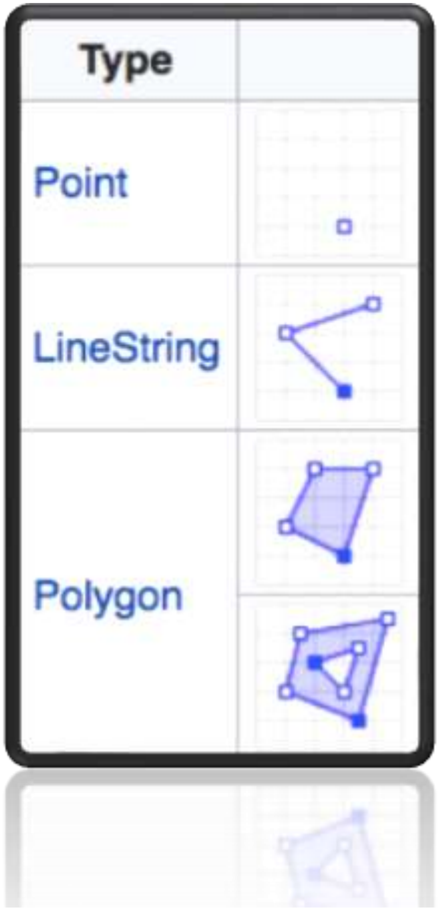
Scale Bar

Data Source

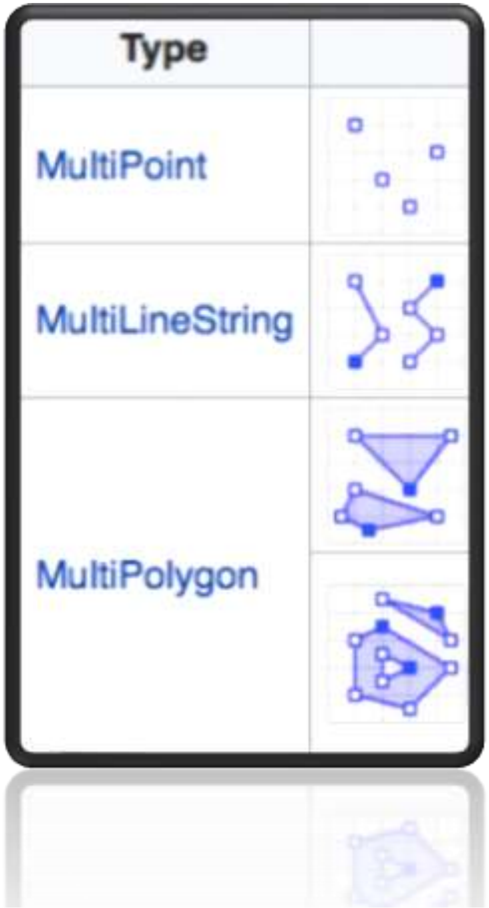
Slides: Dr.Azar Abadi , Kaeli Samson, UNMC

GIS and data type

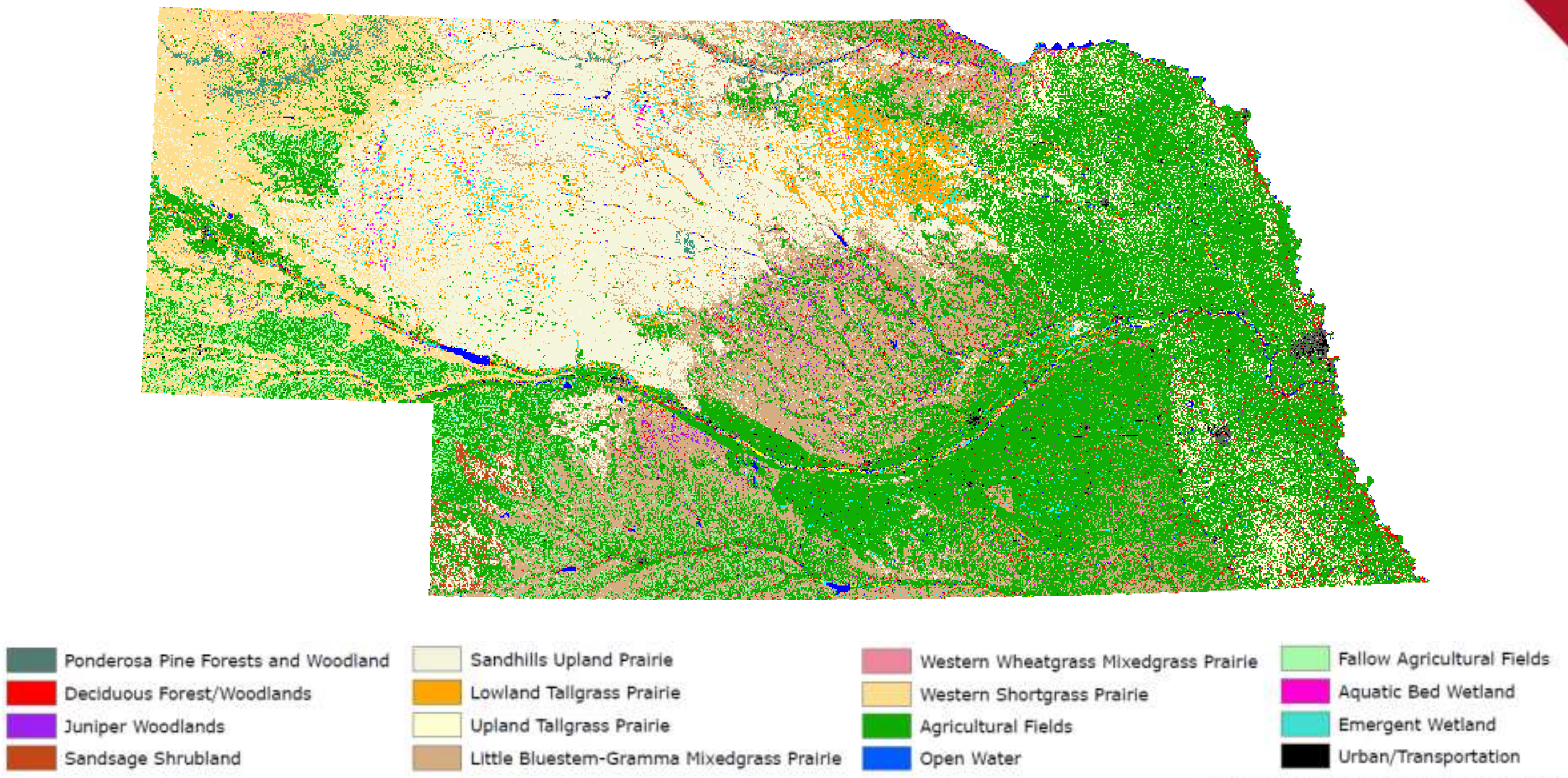
- Vector data



Type	
Point	
LineString	
Polygon	



Type	
MultiPoint	
MultiLineString	
MultiPolygon	


Ponderosa Pine Forests and Woodland
Deciduous Forest/Woodlands
Juniper Woodlands
Sandsage Shrubland
Sandhills Upland Prairie
Lowland Tallgrass Prairie
Upland Tallgrass Prairie
Little Bluestem-Gramma Mixedgrass Prairie
Western Wheatgrass Mixedgrass Prairie
Western Shortgrass Prairie
Agricultural Fields
Open Water
Fallow Agricultural Fields
Aquatic Bed Wetland
Emergent Wetland
Urban/Transportation

For land cover class descriptions visit http://calmit.unl.edu/gap/

Source: https://calmit.unl.edu/land-cover

Slides: Dr.Azar Abadi , Kaeli Samson, UNMC

Spatial Analysis

Hotspot analysis

- Analysis for identification of clustering of spatial phenomenon represented as points on map
- Points can represent an event or object
- Hotspot defined as concentration of events

Public health research

- Where did the disease occur?
  - Spatial analysis

- When and where did the disease occur?
  - Space-time analysis

Spatial autocorrelation

- Spatial autocorrelation refers to the **correlation of a variable with itself** in a space with neighbor

- Most common interpretation is in terms of trends, gradients, or patterns across a map

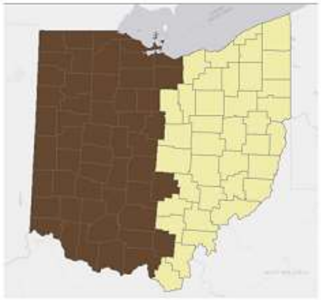
- Commonly used in hotspot analysis

Common Measures of spatial autocorrelation

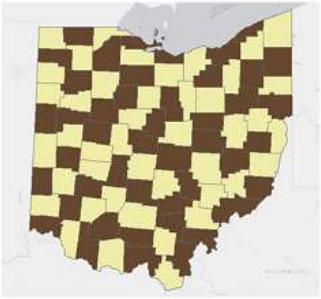
- LISA (Local Moran’I) statistics
- Getis Ord G* statistics

Global measures of spatial association

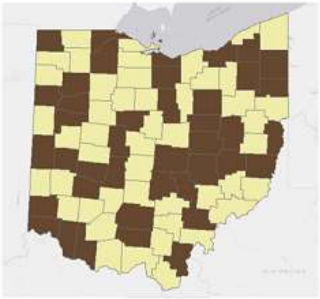
- Gives single statistic **that summarizes whether the values (of the single variable) are similar to their neighbors** across the entire study region.
- It, however, does not tell us where the similarity or dissimilarity occurs.
- Example : Global Moran’s I statistics



Positive autocorrelation



Negative autocorrelation



No spatial autocorrelation

Local measures of spatial association

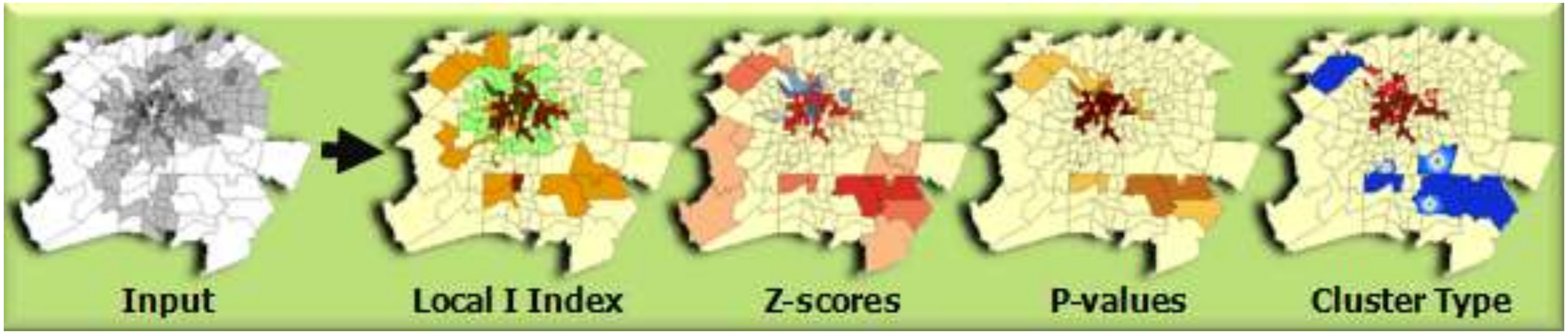
- Local measures call on the principle of spatial heterogeneity, which assumes that the relationships between locations are not constant over the study area

- Provide means of measuring **local variation**

- Example : Local Moran’s I statistics, Getis-Ord G* statistics

Local Moran’s I

- Local Moran’s *I* is the most widely used LISA statistic to locate local clusters and spatial outliers
- For each observation and neighboring observations *j*, the equation for local Moran’s *I* incorporates **deviations from the mean**
- Aim is to **quantify how similar or different the variable values for each observation and their neighbors are when compared to the global average.**

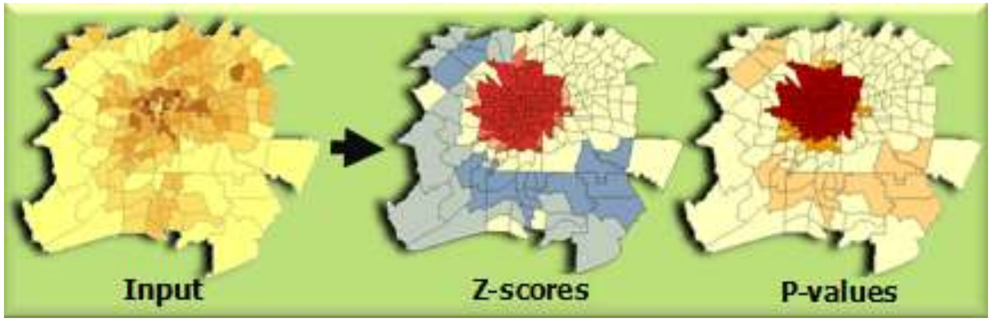


Source :https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/cluster-and-outlier-analysis-anselin-local-moran-s.htm

Getis Ord G* statistics

- Measures overall concentration of similar or dissimilar values located within a specified distance of one another.

- Identifies local spatial clustering patterns, namely “hot spots” (spatial clusters of high values) and “cold spots” (spatial clusters of low values).



Source : https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/hot-spot-analysis.htm

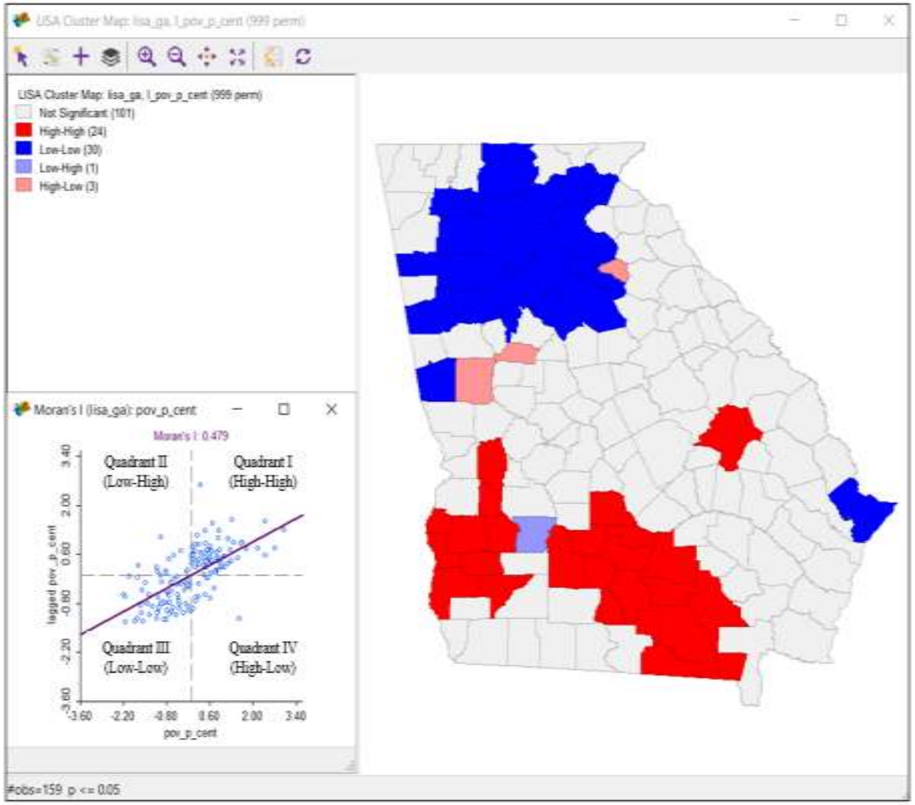
Difference between Local Moran's I and Getis Ord * statistics

- Moran's I statistics measures similarity of nearby features with respect to global mean

- Getis Ord * statistics indicate whether high or low values are concentrated over the area

- Hence, when question is
  - Is data clustered(autocorrelated) -> Local Moran's I statistics
  - Is data cluster of high/low values ? -> Getis Ord G* statistics

Analysis of county-level poverty rate in Georgia



LISA Cluster Map: lisa_ga, I_pov_p_cent (999 perm)

LISA Cluster Map: lisa_ga, I_pov_p_cent (999 perm)

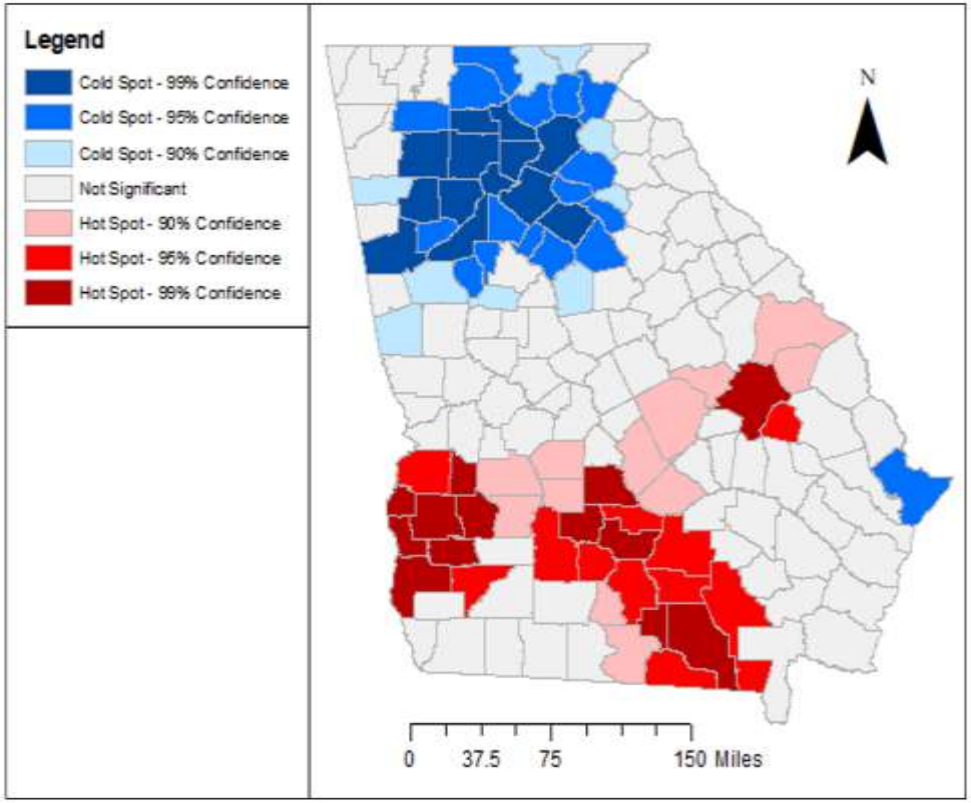
- ⬜ Not Significant (101)
- 🟥 High-High (24)
- 🟦 Low-Low (30)
- 🟪 Low-High (1)
- 🟧 High-Low (3)

Moran's I (lisa_ga): pov_p_cent

Moran's I: 0.479

#obs=159 p <= 0.05

Figure 1. A Moran scatterplot (lower left) and corresponding map (right) showing counties that are spatial clusters (High-High, Low-Low) or spatial outliers (High-Low, Low-High) of poverty rates in the state of Georgia.



Legend

- 🟦 Cold Spot - 99% Confidence
- 🟦 Cold Spot - 95% Confidence
- 🟦 Cold Spot - 90% Confidence
- ⬜ Not Significant
- 🟥 Hot Spot - 90% Confidence
- 🟥 Hot Spot - 95% Confidence
- 🟥 Hot Spot - 99% Confidence

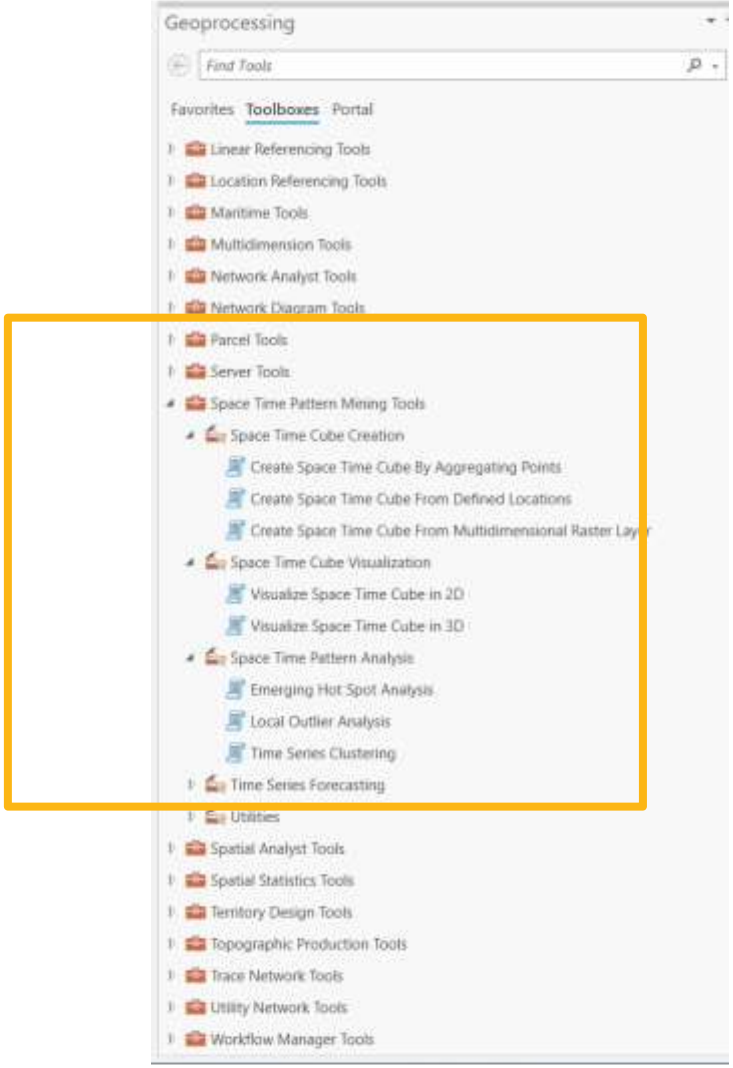
Figure 2. A hot spot map of county-level poverty rate in the state of Georgia generated using Getis-Ord Gi.*

Source : https://gistbok.ucgis.org/bok-topics/local-measures-spatial-association

Space time pattern mining in ArcGIS

Space-time Analysis

- Space time pattern mining toolbox in ArcGIS pro



Geoprocessing

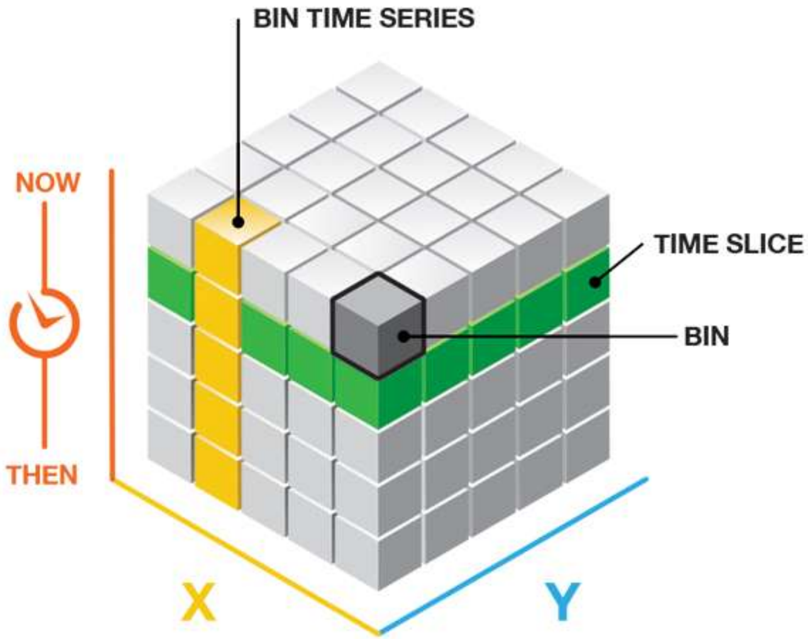
Find Tools

Favorites Toolboxes Portal

- Linear Referencing Tools
- Location Referencing Tools
- Maritime Tools
- Multidimension Tools
- Network Analyst Tools
- Network Diagram Tools
- Parcel Tools
- Server Tools
- Space Time Pattern Mining Tools
  - Space Time Cube Creation
    - Create Space Time Cube By Aggregating Points
    - Create Space Time Cube From Defined Locations
    - Create Space Time Cube From Multidimensional Raster Lay
  - Space Time Cube Visualization
    - Visualize Space Time Cube in 2D
    - Visualize Space Time Cube in 3D
  - Space Time Pattern Analysis
    - Emerging Hot Spot Analysis
    - Local Outlier Analysis
    - Time Series Clustering
  - Time Series Forecasting
  - Utilities
- Spatial Analyst Tools
- Spatial Statistics Tools
- Territory Design Tools
- Topographic Production Tools
- Trace Network Tools
- Utility Network Tools
- Workflow Manager Tools

Space time cube by aggregating points

- A space time cube is created from space- time dataset
- Summarizes a set of points into a netCDF data structure by aggregating them into space-time bins.
- Within each bin, the points are counted, and specified attributes are aggregated.



Source : https://pro.arcgis.com/en/pro-app/latest/tool-reference/space-time-pattern-mining/create-space-time-cube.htm

Application to thyroid cancer incidence

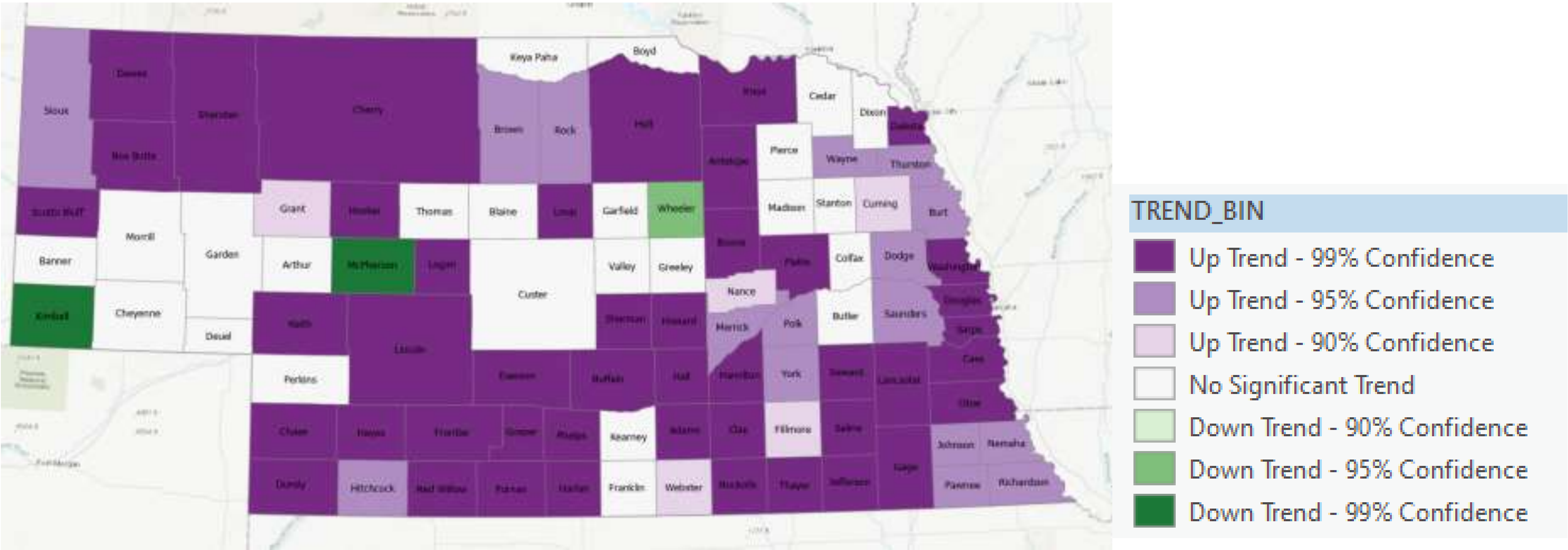
- Data
  - Thyroid cancer data from Nebraska cancer registry
  - Calculated 5 year rolling average of age adjusted thyroid cancer incidence rate by county for Nebraska state
  - Time period – 1990-2004

Trend analysis in space time cube

- The **Mann-Kendall trend test** is performed on every location with data as an independent bin time-series test.

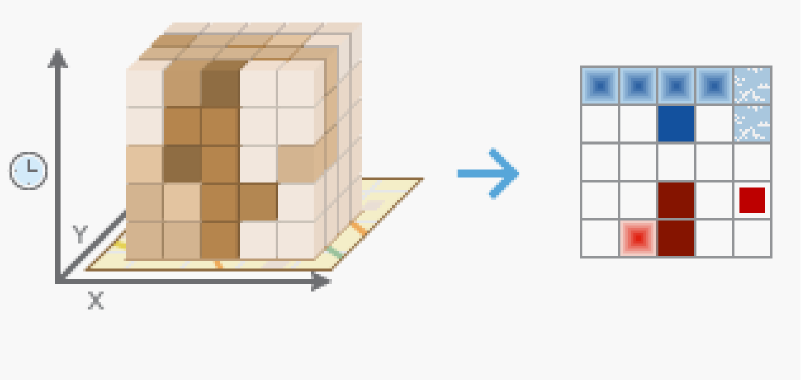
- The Mann-Kendall statistic is a rank correlation analysis for the bin count or value and their time sequence.

Trend analysis

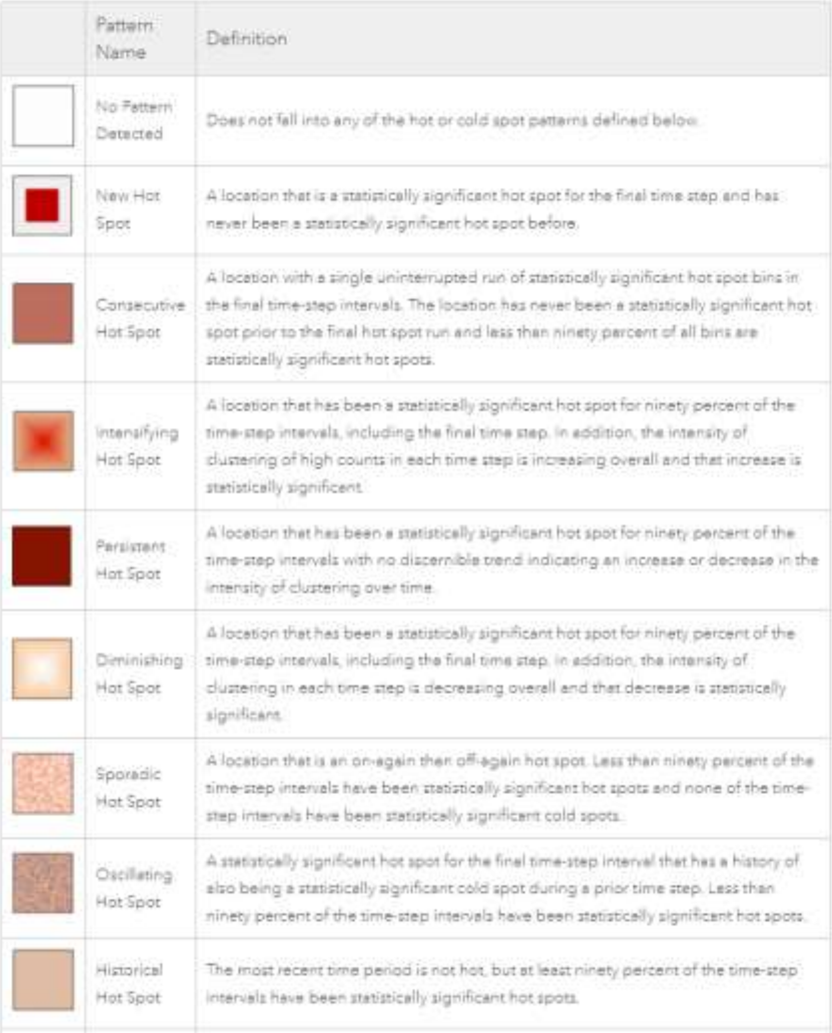


Emerging hotspot analysis - Identifies patterns

- Calculates the **Getis-Ord Gi* statistic** for each bin in the cube and generates z-score, p-value, and hot spot bin classification
- Identifies patterns of **statistically significant new, intensifying, diminishing, and sporadic hot and cold spots**
- Next, these trends are evaluated using the **Mann-Kendall trend test.**
- With the trend z-score and p-value for each location and the hot spot z-score and p-value for each bin, the Emerging Hot Spot Analysis tool categorizes each study area location into patterns

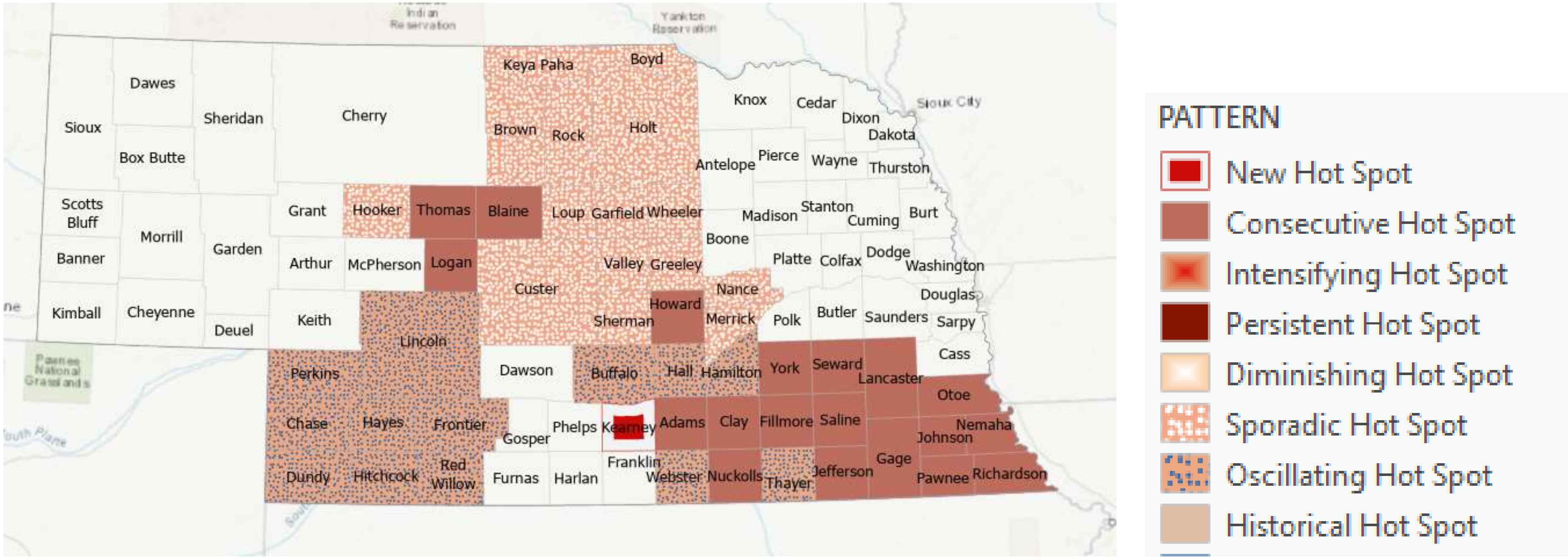


Patterns from Emerging hotspot analysis

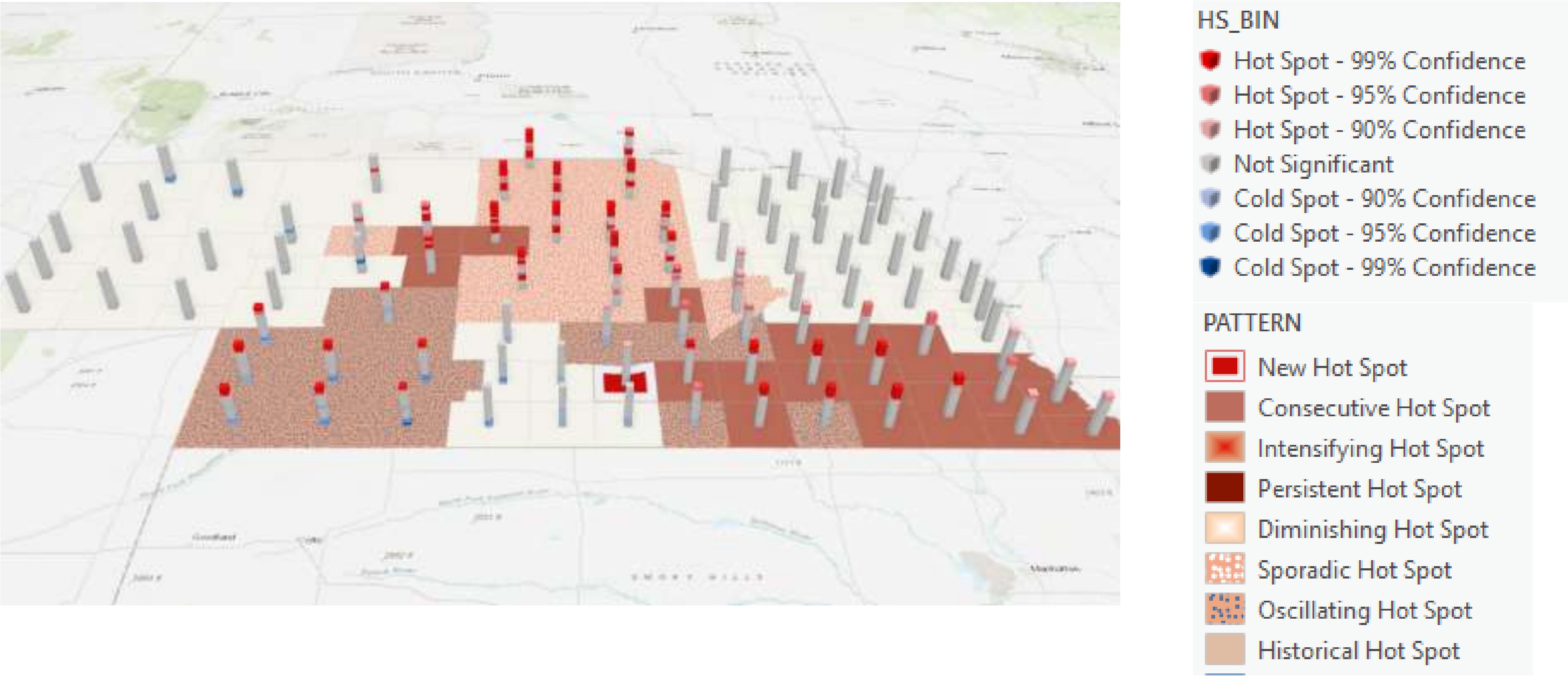


	Pattern Name	Definition
□	No Pattern Detected	Does not fall into any of the hot or cold spot patterns defined below.
🟥	New Hot Spot	A location that is a statistically significant hot spot for the final time step and has never been a statistically significant hot spot before.
🟥	Consecutive Hot Spot	A location with a single uninterrupted run of statistically significant hot spot bins in the final time-step intervals. The location has never been a statistically significant hot spot prior to the final hot spot run and less than ninety percent of all bins are statistically significant hot spots.
🟧	Intensifying Hot Spot	A location that has been a statistically significant hot spot for ninety percent of the time-step intervals, including the final time step. In addition, the intensity of clustering of high counts in each time step is increasing overall and that increase is statistically significant.
🟥	Persistent Hot Spot	A location that has been a statistically significant hot spot for ninety percent of the time-step intervals with no discernible trend indicating an increase or decrease in the intensity of clustering over time.
🟧	Diminishing Hot Spot	A location that has been a statistically significant hot spot for ninety percent of the time-step intervals, including the final time step. In addition, the intensity of clustering in each time step is decreasing overall and that decrease is statistically significant.
🟧	Sporadic Hot Spot	A location that is an on-again then off-again hot spot. Less than ninety percent of the time-step intervals have been statistically significant hot spots and none of the time-step intervals have been statistically significant cold spots.
🟫	Oscillating Hot Spot	A statistically significant hot spot for the final time-step interval that has a history of also being a statistically significant cold spot during a prior time step. Less than ninety percent of the time-step intervals have been statistically significant hot spots.
🟧	Historical Hot Spot	The most recent time period is not hot, but at least ninety percent of the time-step intervals have been statistically significant hot spots.

Emerging hotspot analysis thyroid cancer incidence



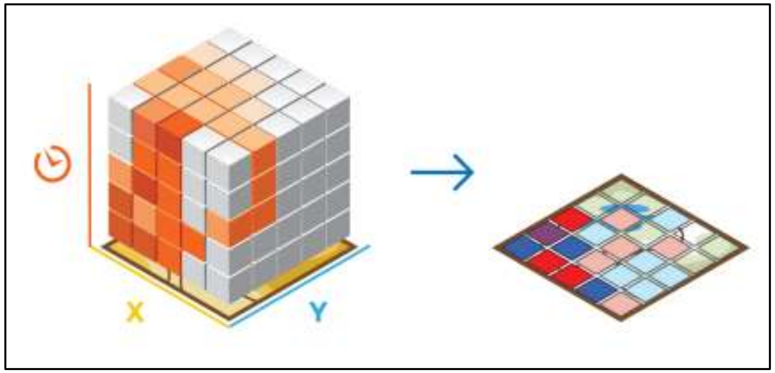
3D Emerging hotspot analysis results



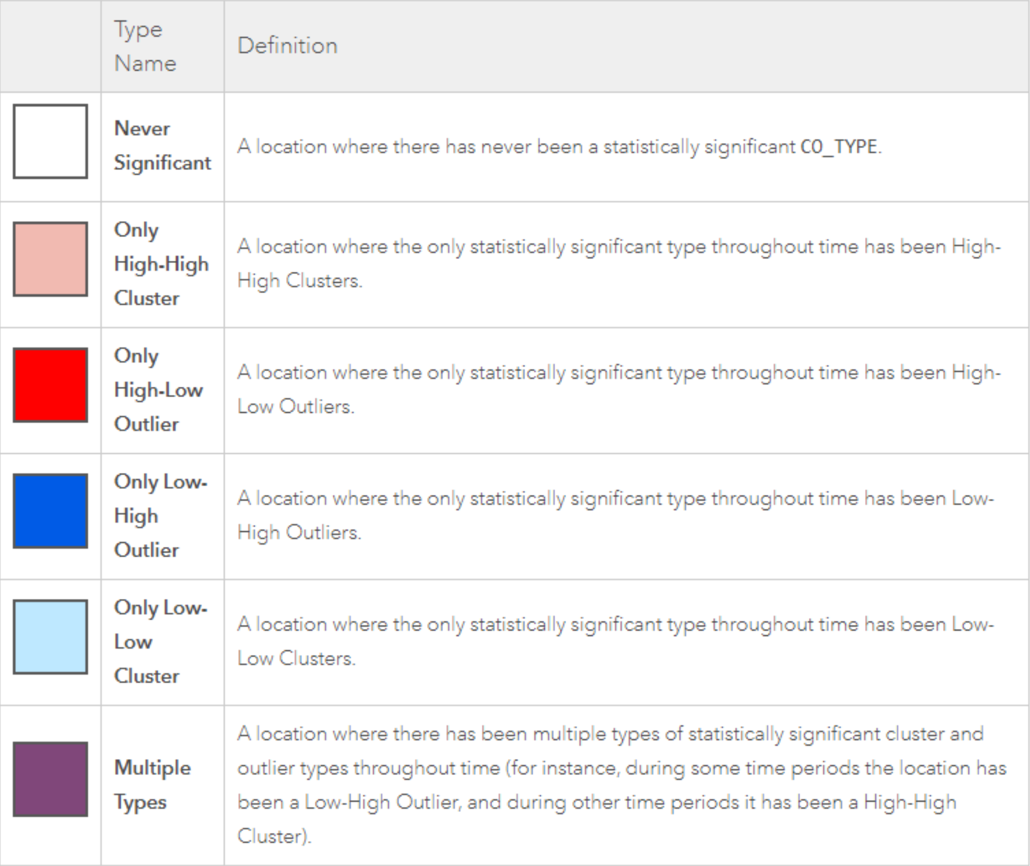
Local outlier analysis –

Identify clusters and outliers

- **Find locations in your study area that have been statistically different than their neighbors in both space and time.**
- Space-time implementation of the Anselin Local Moran's I statistic
- Tool calculates a Local Moran's I index, a pseudo p-value and a type code representing the cluster or outlier category type for each statistically significant bin in the input Space Time Cube
- An index with a positive value indicates that a bin has **neighboring bins with similarly high or low attribute** values; this bin is part of a **cluster**.
- An index with a negative value indicates that a bin has **neighboring bins with dissimilar values**; this bin is an **outlier**.

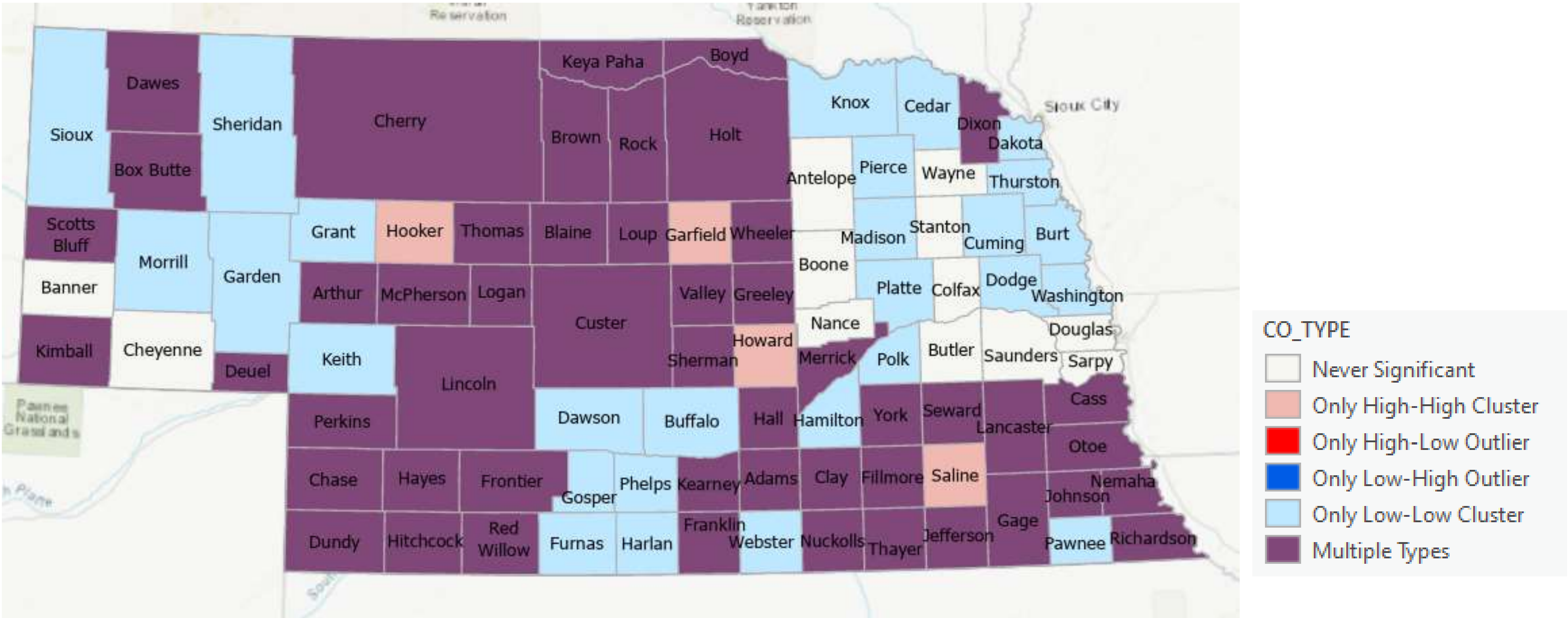


Patterns in Cluster and outlier analysis



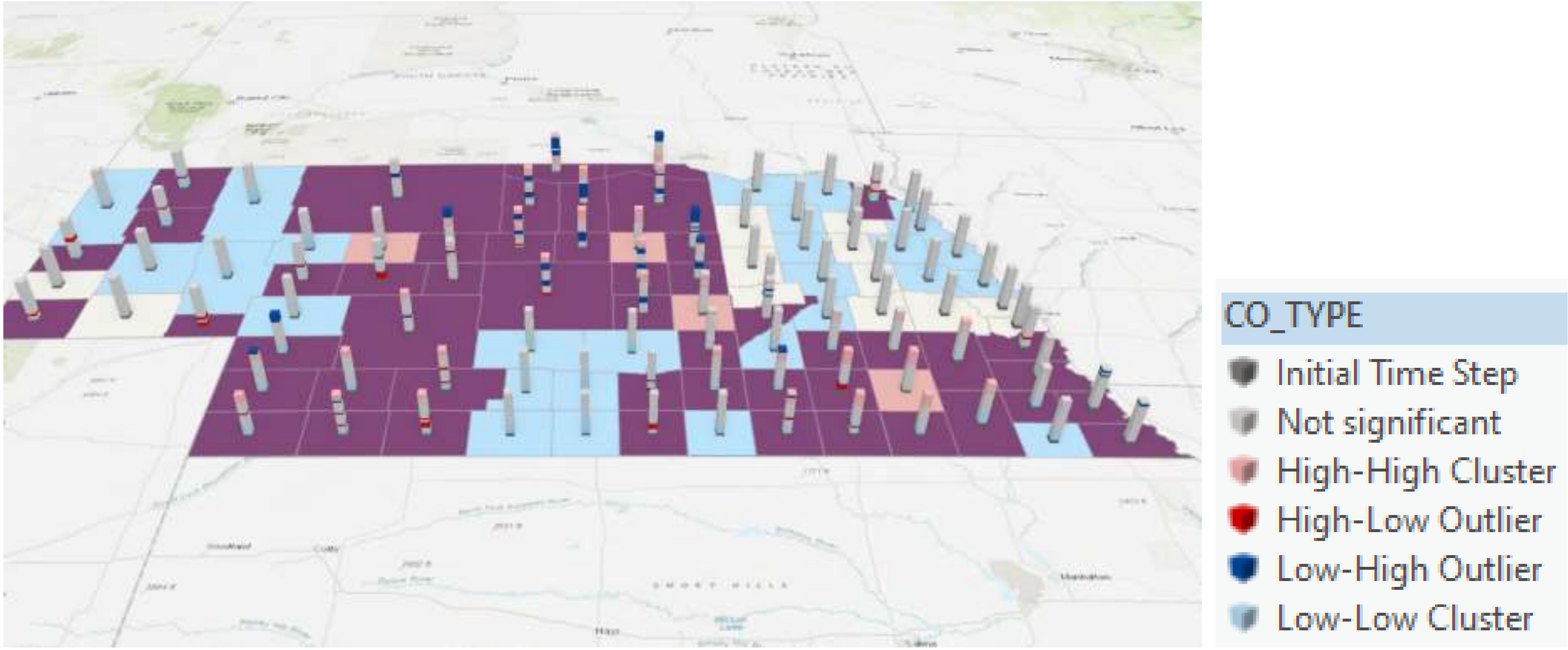
	Type Name	Definition
⬜	**Never Significant**	A location where there has never been a statistically significant CO_TYPE.
🟥	**Only High-High Cluster**	A location where the only statistically significant type throughout time has been High-High Clusters.
🟥	**Only High-Low Outlier**	A location where the only statistically significant type throughout time has been High-Low Outliers.
🟦	**Only Low-High Outlier**	A location where the only statistically significant type throughout time has been Low-High Outliers.
🟦	**Only Low-Low Cluster**	A location where the only statistically significant type throughout time has been Low-Low Clusters.
🟪	**Multiple Types**	A location where there has been multiple types of statistically significant cluster and outlier types throughout time (for instance, during some time periods the location has been a Low-High Outlier, and during other time periods it has been a High-High Cluster).

Clusters and outlier analysis results



- **High-High or Low-low cluster** represents high risk or low risk of thyroid cancer cases in space and time
- **High-Low outlier** is high risk county surrounded by low-risk county and thus immediate attention required to that county
- **Low-High outlier** represents county with low risk but surrounded by high-risk county. Thus, prevention and monitoring required for these counties

3D visualization of outlier analysis



Conclusion

- Trend analysis shows that thyroid cancer incidence has been increasing in Nebraska from 1990-2014.
- Kearney is a new emerging hotspot for thyroid cancer.
- Counties in central Nebraska has been sporadic hotspot throughout the time period. Southern Nebraska counties are consecutive hotspot for thyroid cancer

Next steps

- Bivariate analysis to identify association between thyroid cancer and pesticide application by county
- Are there space time trends between pesticide application and emergence of thyroid cancer ?

Spatial correlation in R

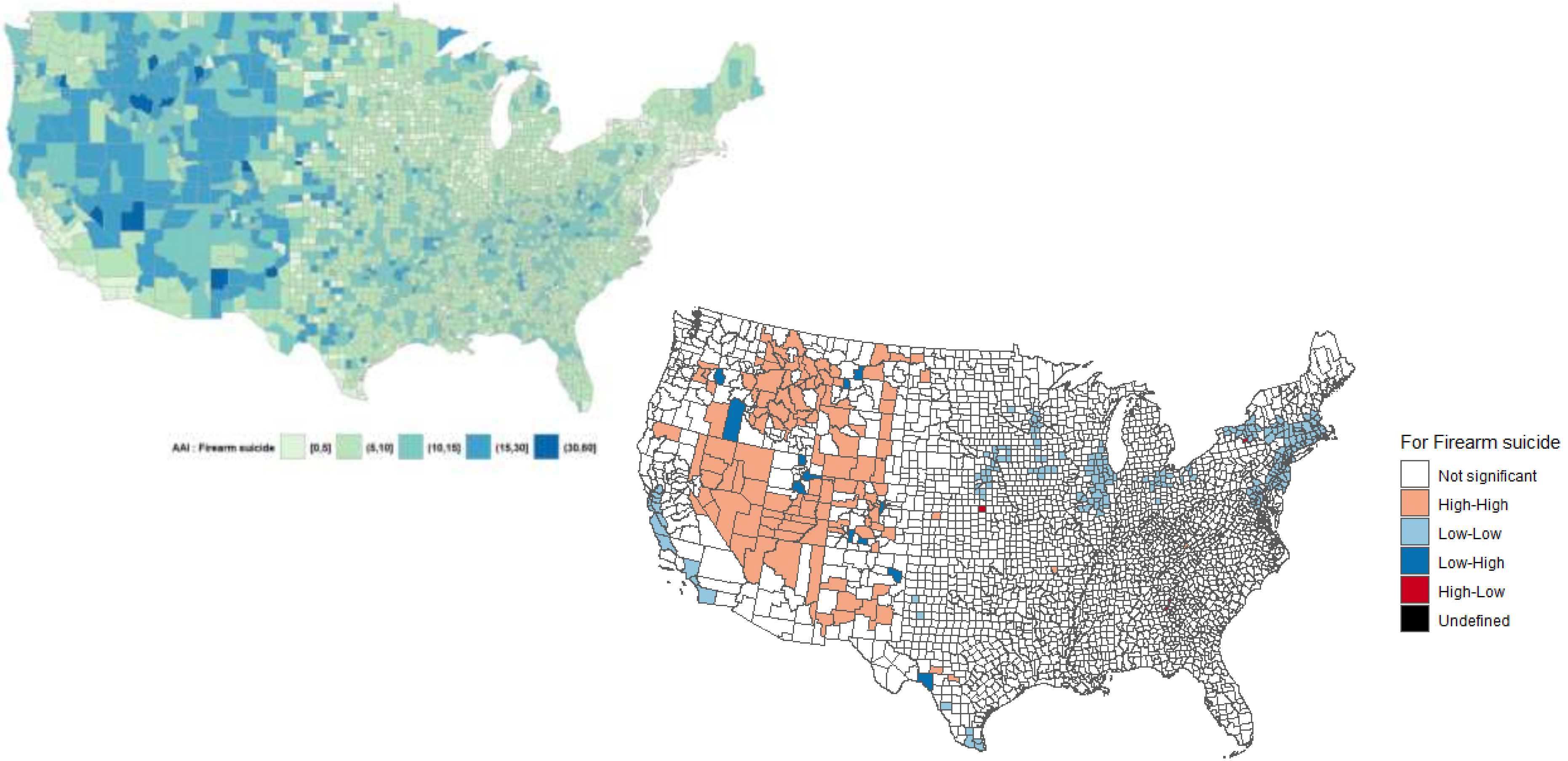
R package - rgeoda

- This package includes functions for spatial data analysis in R
- Includes various **spatial association measures and spatial clustering**
- Resources : https://geodacenter.github.io/rgeoda/

https://geodacenter.github.io/rgeoda/articles/rgeoda_tutorial.html

LISA in R

Spatial clusters and outliers for firearm suicide AAI rate by county in USA



Take away

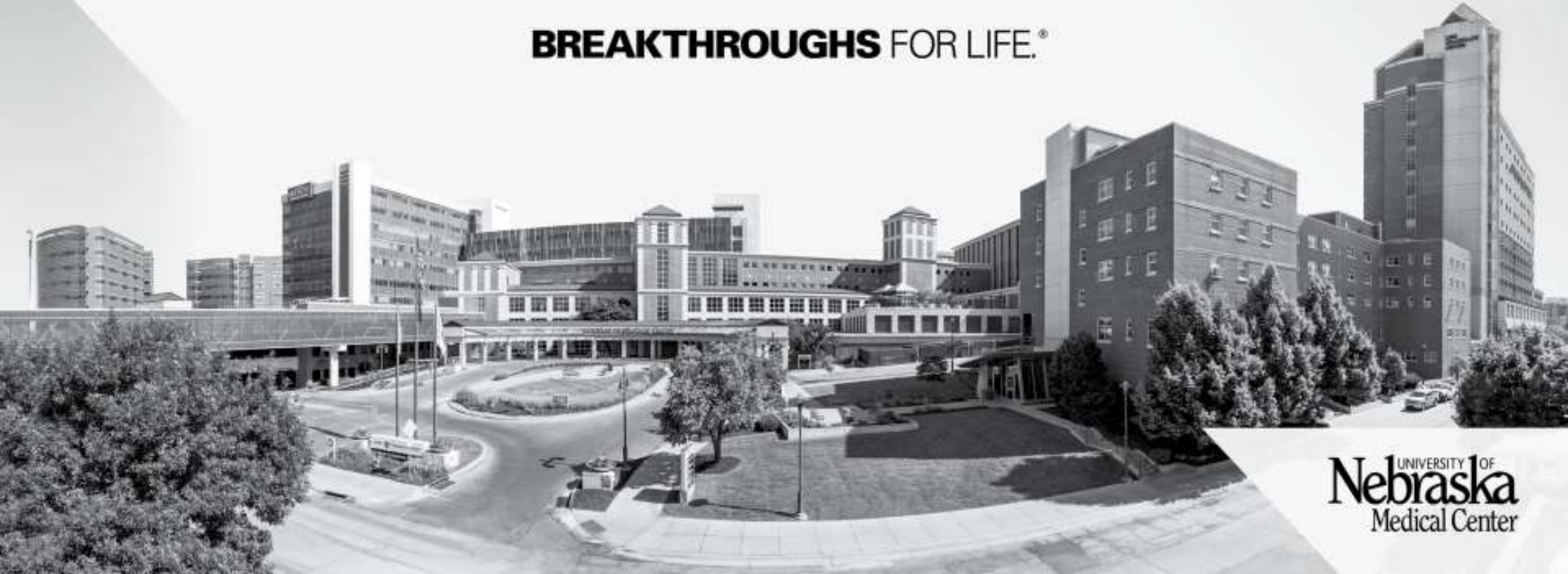
- Spatial autocorrelation methods for hotspot analysis have many applications in public health research studies

- However, it is important to relate the outcome of the statistics with our research question

- The space time pattern mining toolbox from ArcGIS shows strong potential to go beyond just spatial analysis

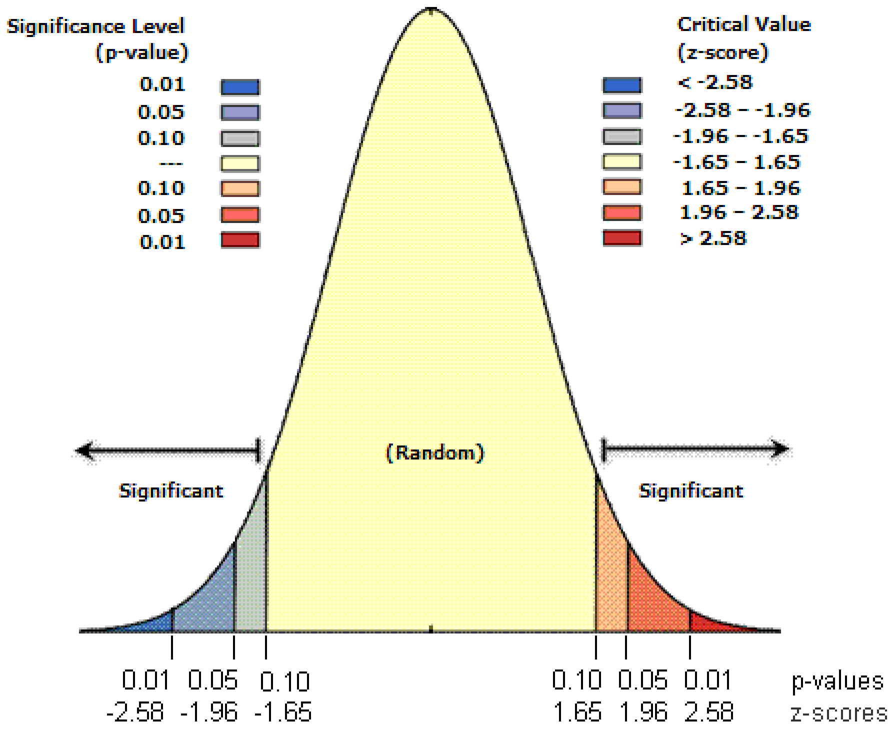
University of Nebraska Medical Center℠

BREAKTHROUGHS FOR LIFE.®



University of Nebraska Medical Center

Extra


Significance Level
(p-value)
0.01
0.05
0.10

0.10
0.05
0.01
Critical Value
(z-score)
< -2.58
-2.58 - -1.96
-1.96 - -1.65
-1.65 - 1.65
1.65 - 1.96
1.96 - 2.58
> 2.58
(Random)
Significant
Significant
0.01 0.05 0.10
-2.58 -1.96 -1.65
0.10 0.05 0.01
1.65 1.96 2.58
p-values
z-scores

Spatial association

- Describes how values observation or samples are related in space
- Based upon Tobler’s First Law of Geography – “Everything is related to everything else, but near things are more related than distant things”
- Can be measured locally or globally

About me

- Background:
  - Bachelor in Engineering (IT) from India
  - MSc Geoinformatics - Netherlands
- Research Interest
  - Spatial analysis
  - Machine learning