Introduction to Spatial Autocorrelation

Jenny Palomino
March 24, 2017

Berkeley UNIVERSITY OF CALIFORNIA
College of Natural Resources
INTERNATIONAL AND EXECUTIVE PROGRAMS

GEOSPATIAL INNOVATION FACILITY
Cutting-Edge Mapping Technology at UC Berkeley

Objectives of Workshop

1. Define spatial autocorrelation and understand how it is measured

2. Learn about methods to quantify the degree to which it is significant for a dataset/attribute (i.e. Moran's I)

3. Explore datasets and analysis functions in a nice Graphical User Interface (GUI) called GeoDa

4. Write Python code in Jupyter Notebook to run functions for Global and Local Moran's analysis

   *Complete bonus activities in each Jupyter Notebook or explore any of the exercises for additonal data available in the Bonus Data folder

Why do we care about spatial patterns?

"Everything is related to everything else, but near things are more related than distant things" (Tobler's First Law of Geography)

How does this translate to data?

"Data from locations near one another in space are more likely to be similar than data from locations remote from one another" (O'sullivan & Unwin)

In other words:

Spatial dependence exists, and spatial pattern analysis methods can help us identify the degree to which this is significant for a dataset.

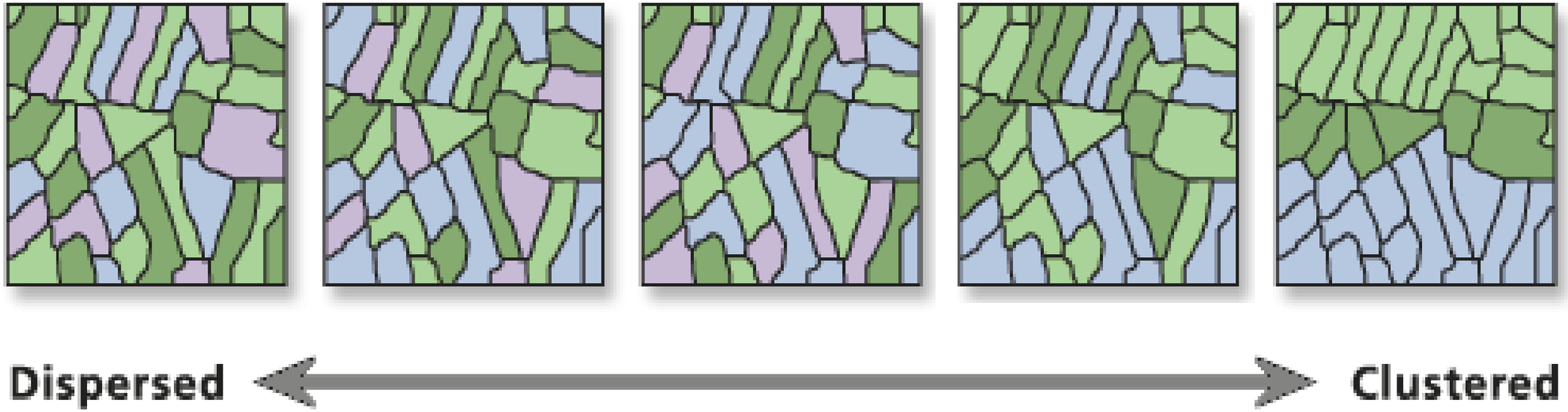
What is Spatial Dependence?

"Spatial dependency is the co-variation of properties within geographic space: characteristics at proximal locations appear to be correlated, either positively or negatively." (Wikipedia)

"Spatial dependence exists when the value associated with one location is dependent on those of other locations." (GeoDa)

What is Spatial Autocorrelation?

A measure of spatial dependency of an attribute/value:
“when the observed spatial pattern is different from what would be expected under a random process operating in space” (PySAL website)



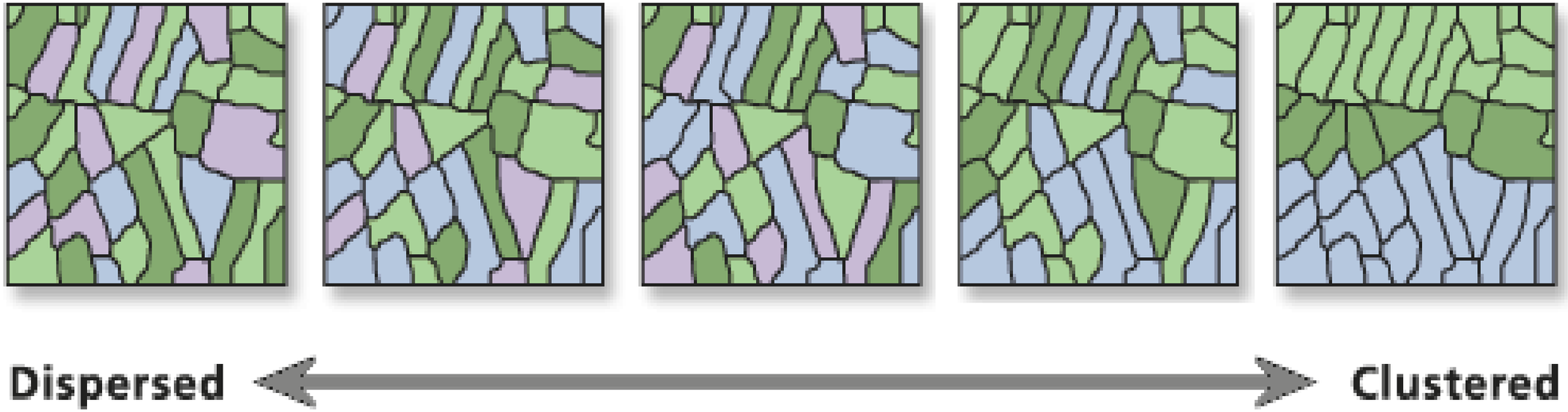
ESRI Help

How do you describe Spatial Autocorrelation?

Positive: similar values are closer together in space (clustered pattern)

Negative: Dissimilar values are closer together in space (dispersed pattern)

None: no structured spatial pattern is observed (random)



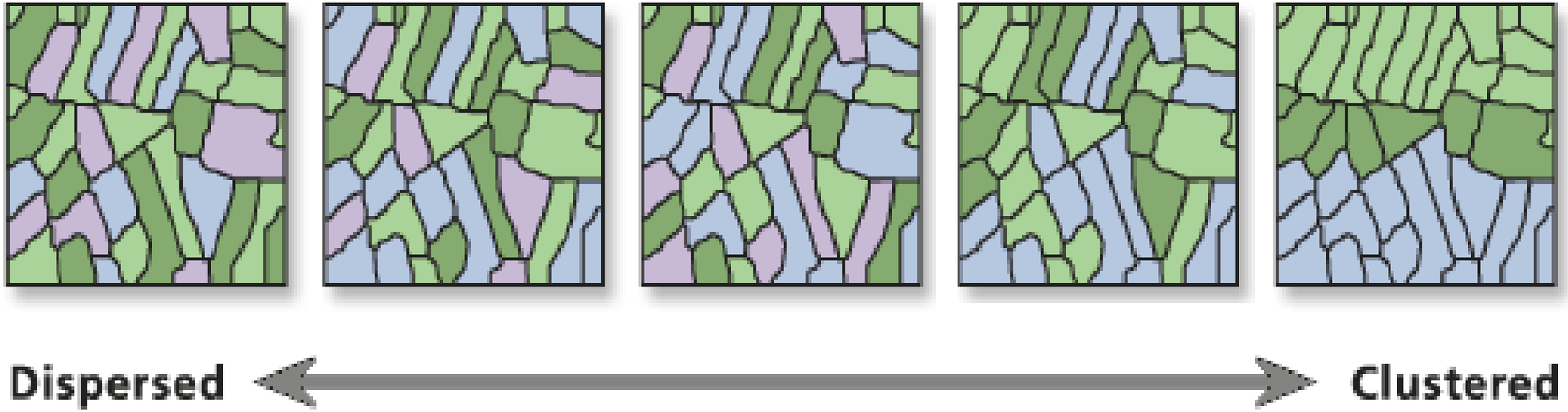
ESRI Help

At what scales do you measure Spatial Autocorrelation?

Two Measures of Spatial Autocorrelation:

1. Global - quantifies clustering or dispersion across a dataset

2. Local - identifies hot or cold-spots within the dataset (individual features are compared to their defined neighbors)



ESRI Help

How do you quantify Spatial Autocorrelation?

Many commonly used tests are available in GeoDa:

Global Autocorrelation

- Moran's I
- Gamma Index of Spatial Autocorrelation
- Join Count Statistics
- Geary's C
- Getis and Ord's G

Local Autocorrelation

- Local Moran's I
- Local G and G*

How do you quantify Spatial Autocorrelation?

The hands-on exercises focus on two metrics: one for global and one for local.

Global Autocorrelation

- **Moran's I**
- Gamma Index of Spatial Autocorrelation
- Join Count Statistics
- Geary's C
- Getis and Ord's G

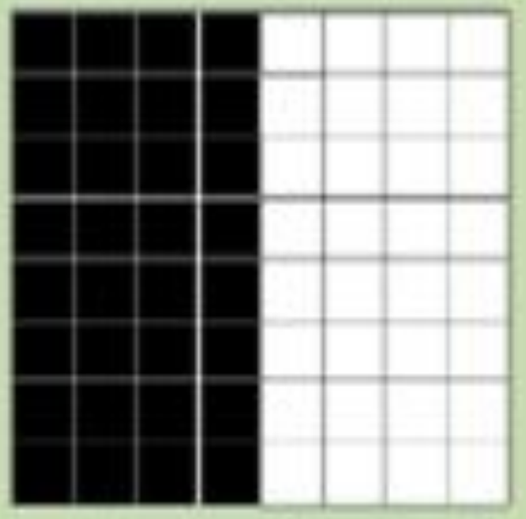
More details on the others available from UT Dallas and R documentation.

Local Autocorrelation

- **Local Moran's I** → **Local Indicators of Spatial Autocorrelation (LISA)**
- Local G and G*

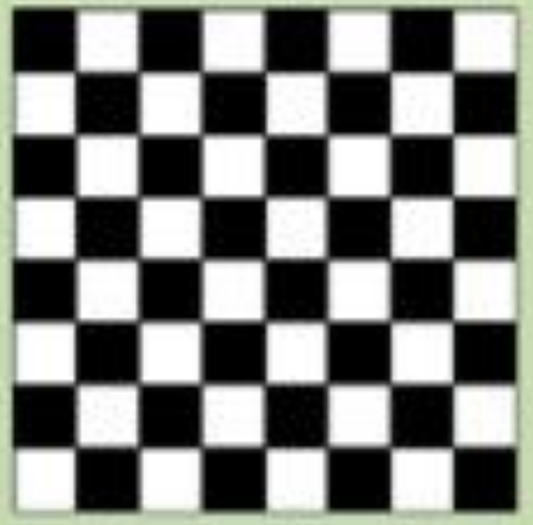
Global Moran's I

Moran's I ~ 1.0
(Clustered Pattern)



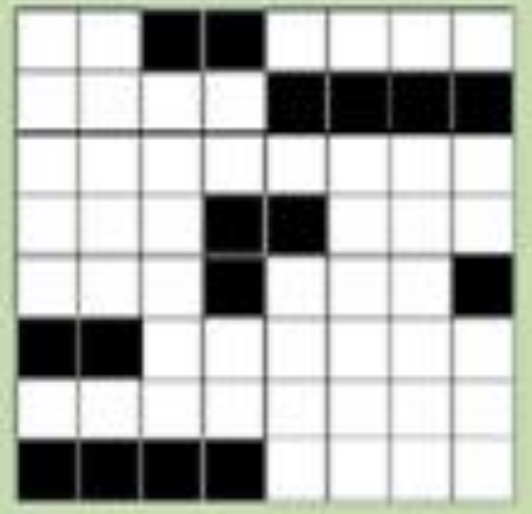
Positive
Autocorrelation

Moran's I ~ -1.0
(Dispersed Pattern)



Negative
Autocorrelation

Moran's I ~ 0.0
(Random)



No Autocorrelation

(O'Sullivan and Unwin, 2002)

Global Moran's I: analysis

Set up a null hypothesis: the chosen attribute exhibits complete spatial randomness across the map (no spatial autocorrelation in the attribute values)

Global Moran's I: analysis

Set up a null hypothesis: the chosen attribute exhibits complete spatial randomness across the map (no spatial autocorrelation in the attribute values)

Test the null hypothesis by creating random versions of your dataset and comparing Moran's I values (i.e. Does the original dataset exhibit more spatial autocorrelation than the random landscapes?)

Global Moran's I: analysis

Set up a null hypothesis: the chosen attribute exhibits complete spatial randomness across the map (no spatial autocorrelation in the attribute values)

Test the null hypothesis by creating random versions of your dataset and comparing Moran's I values (i.e. Does the original dataset exhibit more spatial autocorrelation than the random landscapes?)

Accepting the null hypothesis means that there is no spatial autocorrelation in the dataset (complete randomness in the location of the attribute values)

Rejecting the null hypothesis means that there is spatial autocorrelation in the dataset (clustered if value closer to 1.0 or dispersed if value closer to -1.0)

Global Moran's I: output

I = actual observed value of spatial autocorrelation in the dataset, ranging from -1.0 (dispersed) to 1.0 (clustered); ~0.0 indicating a random pattern

Global Moran's I: output

I = actual observed value of spatial autocorrelation in the dataset, ranging from -1.0 (dispersed) to 1.0 (clustered); ~0.0 indicating a random pattern

E[I] = expected value of spatial autocorrelation (= -1/n-1, approx. ~ 0.0)

Simulated I = calculated I value(s) based on random distribution(s) of the dataset (always ~0.0) (after multiple permutations, there is a mean and standard deviation value for simulated I)

Global Moran's I: output

I = actual observed value of spatial autocorrelation in the dataset, ranging from -1.0 (dispersed) to 1.0 (clustered); ~0.0 indicating a random pattern

E[I] = expected value of spatial autocorrelation (= -1/n-1, approx. ~ 0.0)

Simulated I = calculated I value(s) based on random distribution(s) of the dataset (always ~0.0) (after multiple permutations, there is a mean and standard deviation value for simulated I)

p-value = statistical significance of difference between I and E[I] or simulated I after multiple runs ($p < 0.05$ is the default minimum criteria for significance)

Interpreting the Z Score

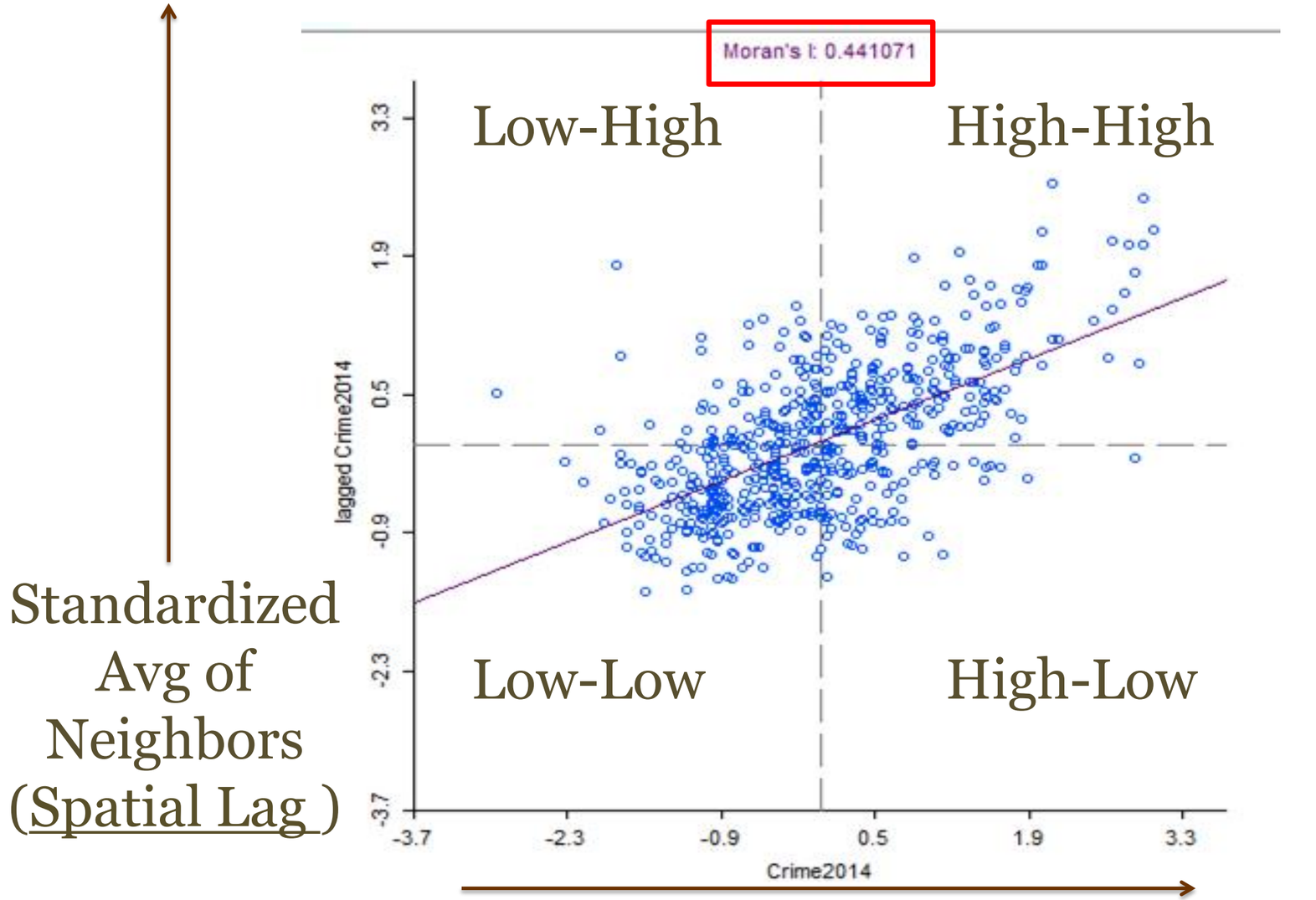
z = (observed I – expected I) / (standard deviation of simulated I)

Standardized statistical score of difference between I and simulated I.

Positive z value indicates clustered global pattern (higher than 1.96 is statistically significant), while negative z value indicates dispersed global pattern (less than -1.96 is statistically significant).

Based on a 95% confidence level, if the z score is between -1.96 and 1.96, you CANNOT reject your null hypothesis; the pattern exhibited is very likely random (i.e. less than 5% chance that pattern is not random).

Global Moran's I: Moran Scatter Plot



Standardized Avg of Neighbors (Spatial Lag)

Standardized Value at Starting Location

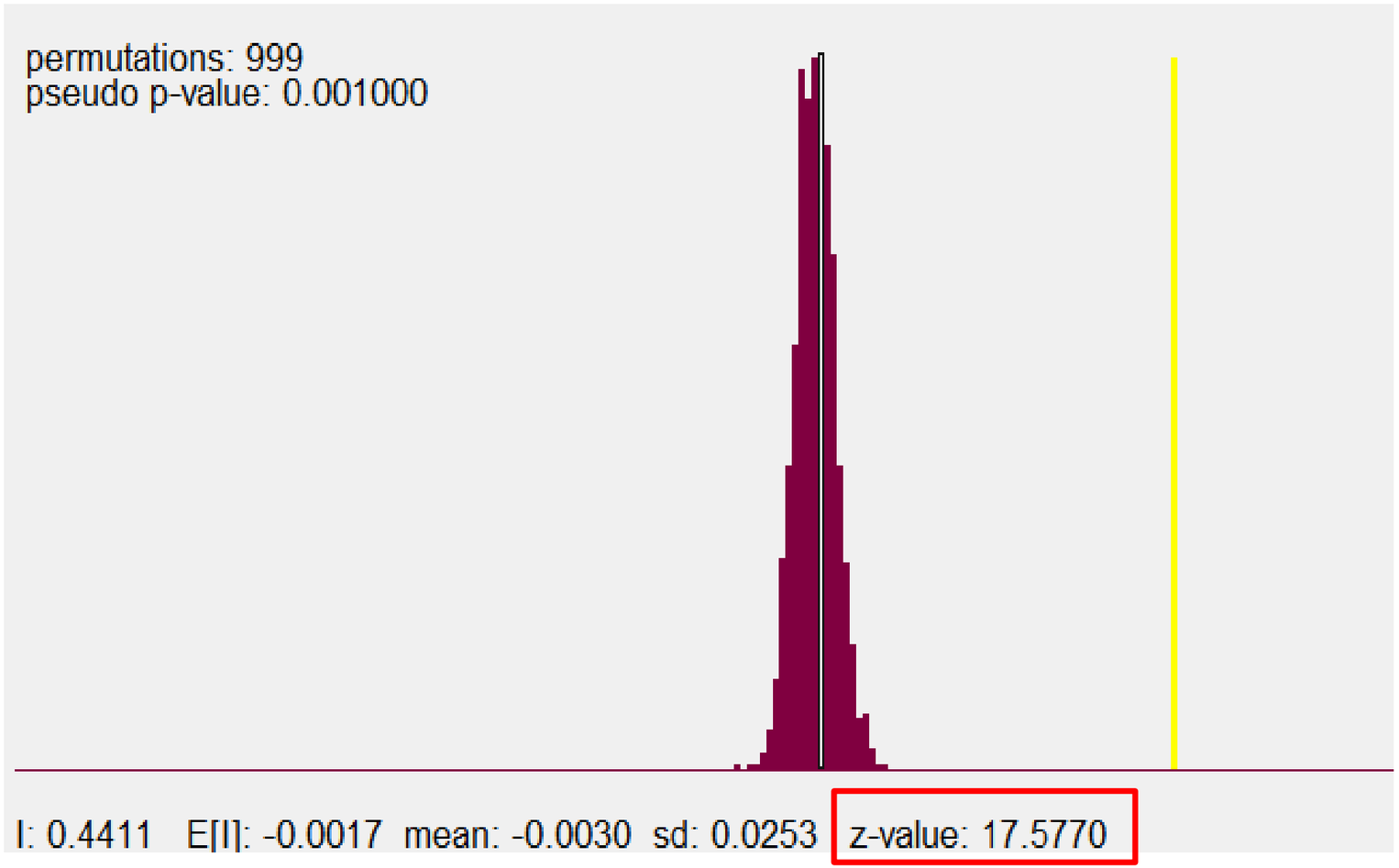
Both Moran's I **(=0.44)** and slope of scatter plot indicate positive spatial autocorrelation.

Moran's I is dependent on this dataset's spatial weights matrix (neighbors).

This means you CANNOT compare Moran's I value as is across datasets.

GeoDa

Global Moran's I: Z Score



Yellow line is the observed I of the dataset, as compared to the distribution of I from random runs.

Z value higher than 1.96 indicates significant positive spatial autocorrelation…

AND because it is standardized, you CAN compare this z value to another dataset.

GeoDa

Local Indicators of Spatial Autocorrelation (LISA) – analysis and output

Accept or reject null hypothesis (whether due to complete spatial randomness, a location has a significantly higher or lower value compared to its neighbors)

Local Indicators of Spatial Autocorrelation (LISA) – analysis and output

Accept or reject null hypothesis (whether due to complete spatial randomness, a location has a significantly higher or lower value compared to its neighbors)

Outputs of LISA analysis **for each location**:

I = actual observed value of its dispersion/clustering compared to its neighbors

p-value = statistical significance of difference between I and simulated I based on its neighbors ($p < 0.05$ is the default minimum criteria for significance)

Local Indicators of Spatial Autocorrelation (LISA) – analysis and output

Accept or reject null hypothesis (whether due to complete spatial randomness, a location has a significantly higher or lower value compared to its neighbors)

Outputs of LISA analysis **for each location**:

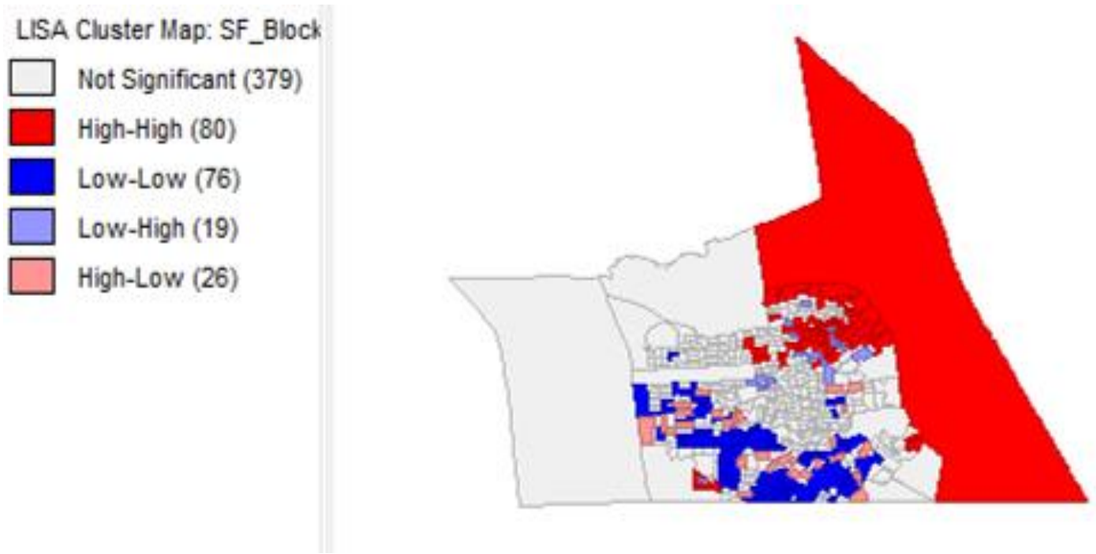
I = actual observed value of its dispersion/clustering compared to its neighbors

p-value = statistical significance of difference between I and simulated I based on its neighbors ($p < 0.05$ is the default minimum criteria for significance)

Quadrant (q) = the type of local cluster it is (i.e., high value next high values, high value next to low values, etc)

LISA: Cluster Types

Cluster Map:
identifies significant clusters
and spatial outliers



High-High = a high value surrounded by other high values (local hot-spot)

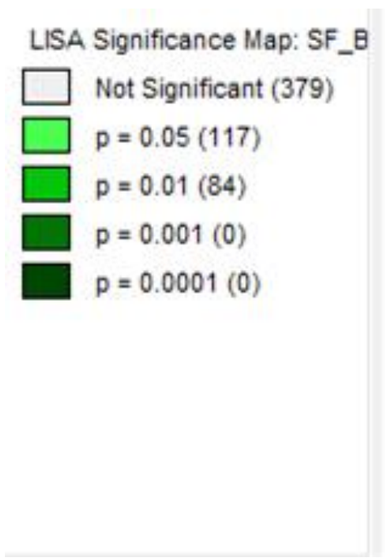
Low-Low = a low value surrounded by other low values (local cold-spot)

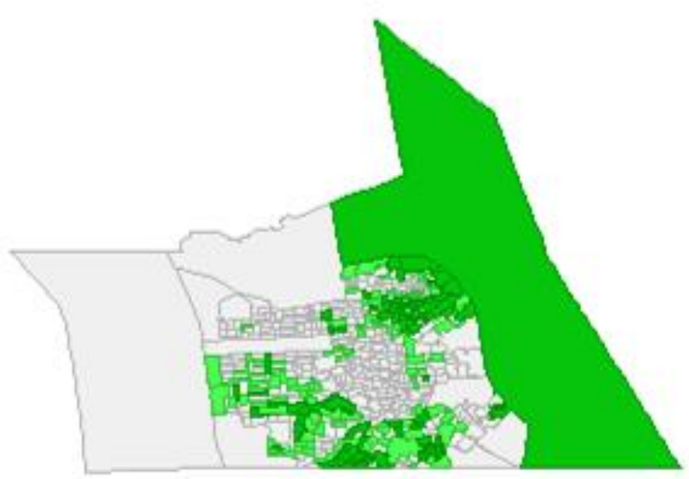
Low-High = a low value surrounded by high values (local low outlier)

High-Low = a high value surrounded by low values (local high outlier)

LISA: Significance Levels

Significance Map:
displays significance levels of each location's cluster type

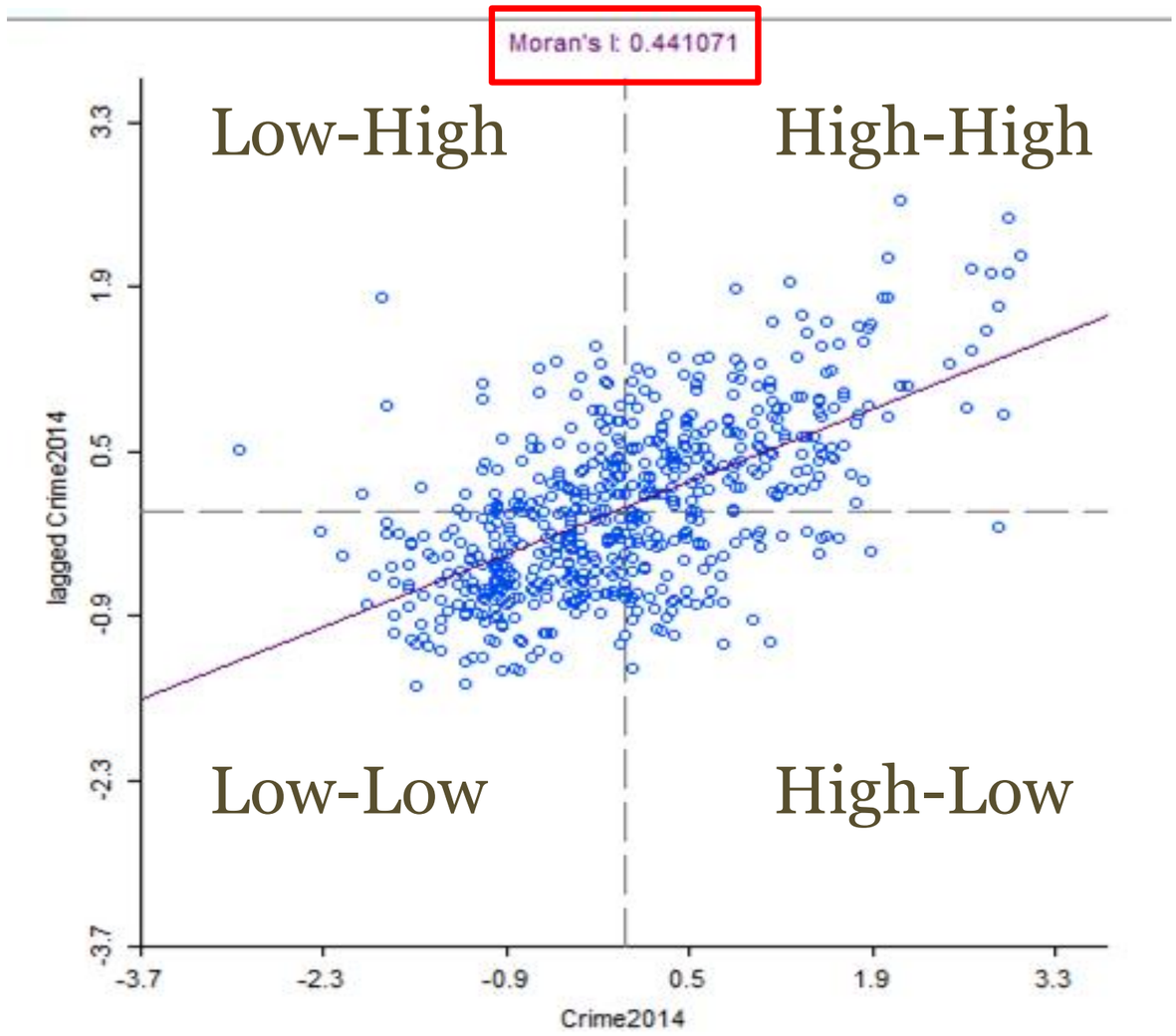




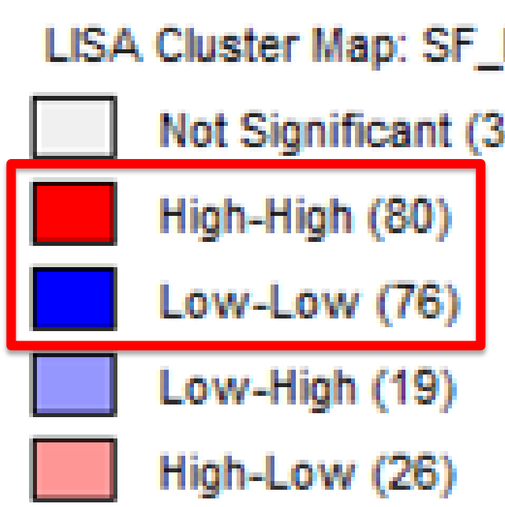
GeoDa shows you only Cluster Types with significant p-values (i.e. statistically significant clusters).

In other software, Significance and Cluster Type Maps need to be combined to find statistically significant clusters.

Global and Local Moran's Together



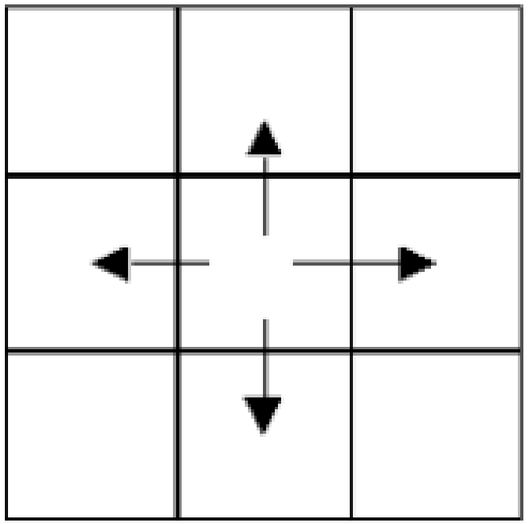
Moran's I = 0.44



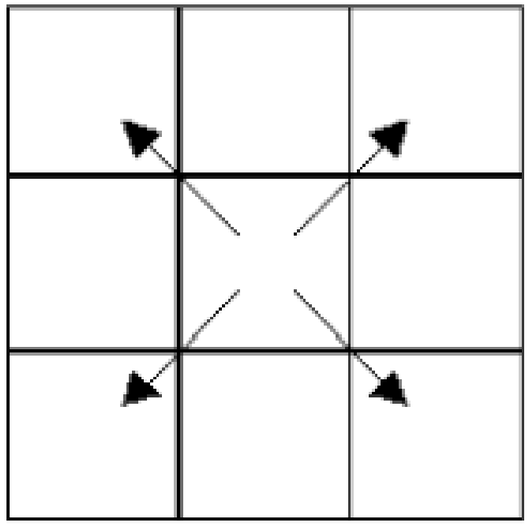
156 signficant clusters of hot-spots and cold-spots

Defining Neighbors for Spatial Weights Matrix

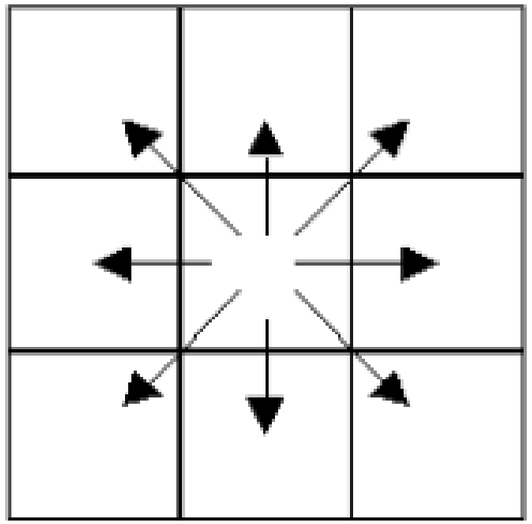
Rooks Case



Bishops Case



Queen's (Kings) Case



University of Ottawa

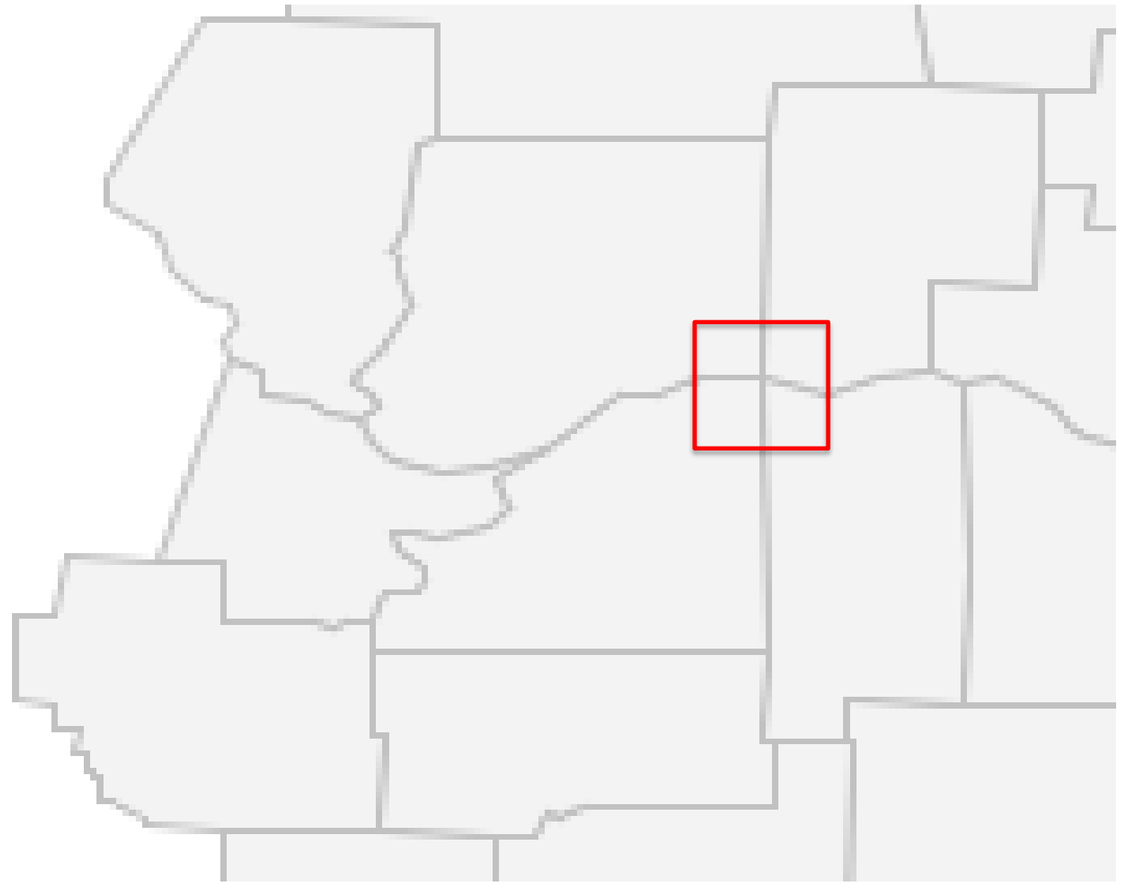
Defining Neighbors for Spatial Weights Matrix

Rook: neighbors have to share a border

- North and South neighbors
- East and West neighbors

Queen: neighbors only have to share a vertex (corner) point

- same as rook PLUS:
- Northeast, Northwest, Southeast, and Southwest neighbors



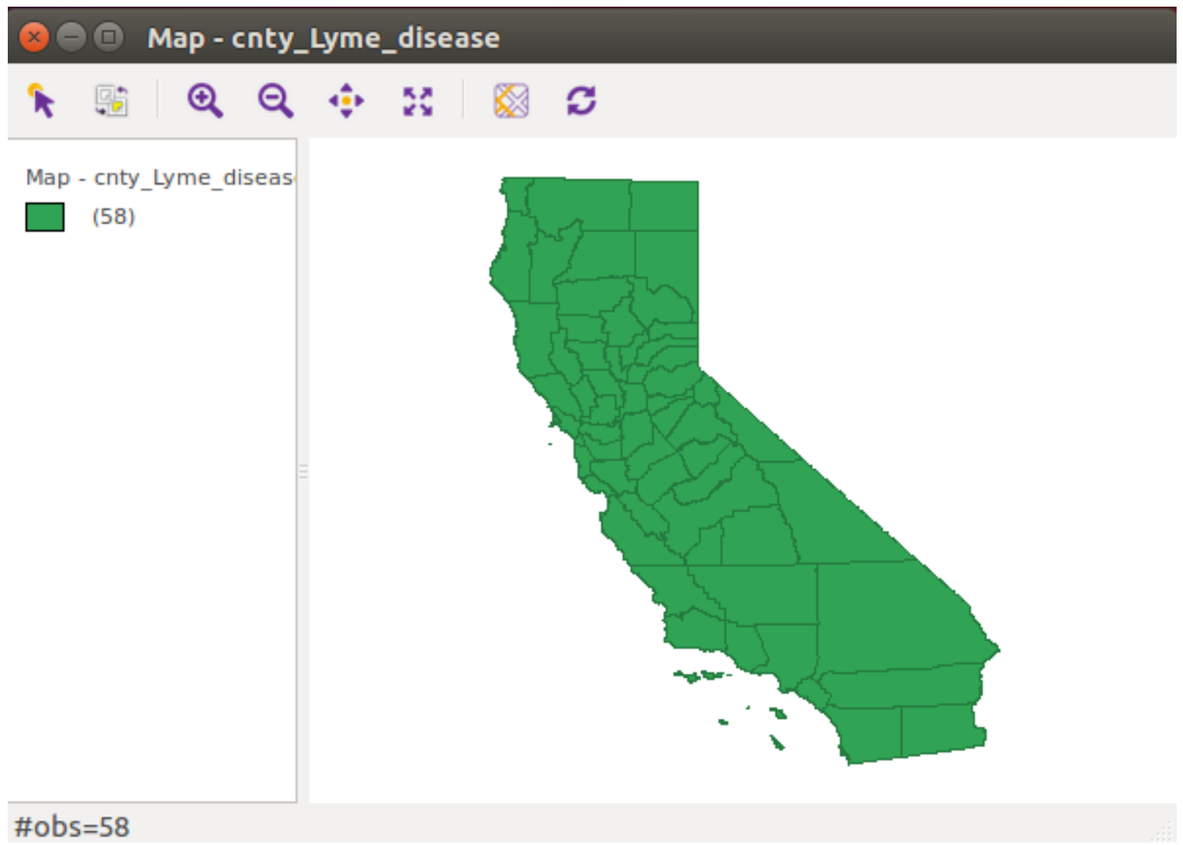
Other Tools for Spatial Autocorrelation Analysis

- ArcGIS

- R (i.e. spdep package)

- Stata (i.e. spatgsa package)

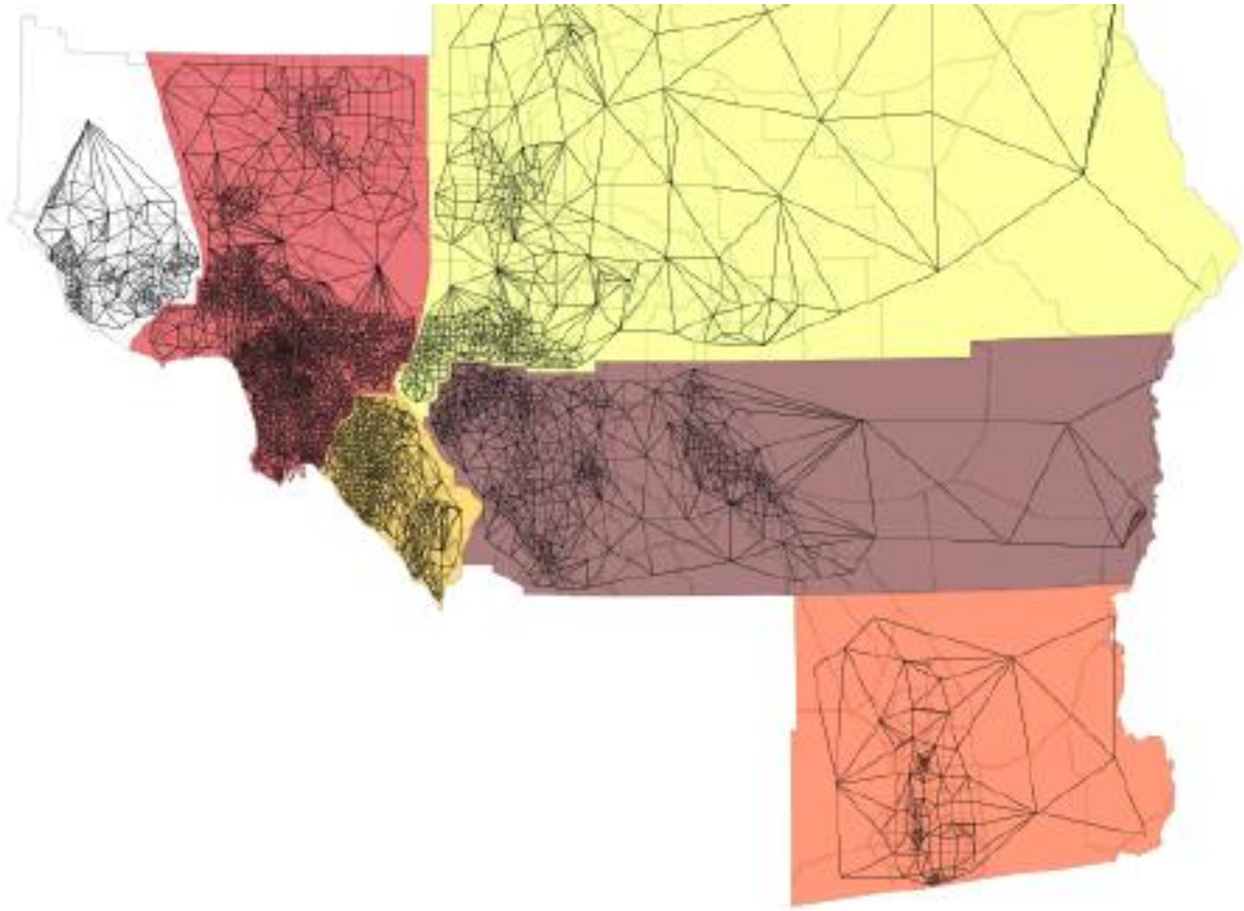
GeoDa Graphical User Interface (GUI)



PySAL - Python Spatial Analysis Library

Vector and Raster Data Analysis

- Spatial Autocorrelation
- Spatial Econometrics
- Spatial Smoothing
- Regionalization
- Markov Chains



Requires Python 2.6, 2.7, or 3.4

- numpy (1.3 or later)
- scipy (0.11 or later)
- add shapely for more options

Why code in PySAL when we can use an easy tool like GeoDa?

Reproducible, Standardized, and Collaborative

1. Run the process in an automated way for multiple years, datasets, attributes
2. No need to save/backup output files (simply store or share code that can be run to create new results as needed)
3. Easily share a standardized process with research team or others
4. Integrate code into a web app or connect it easily to a spatial database

Additional Resources: Spatial Autocorrelation

PySAL Plug-in for QGIS

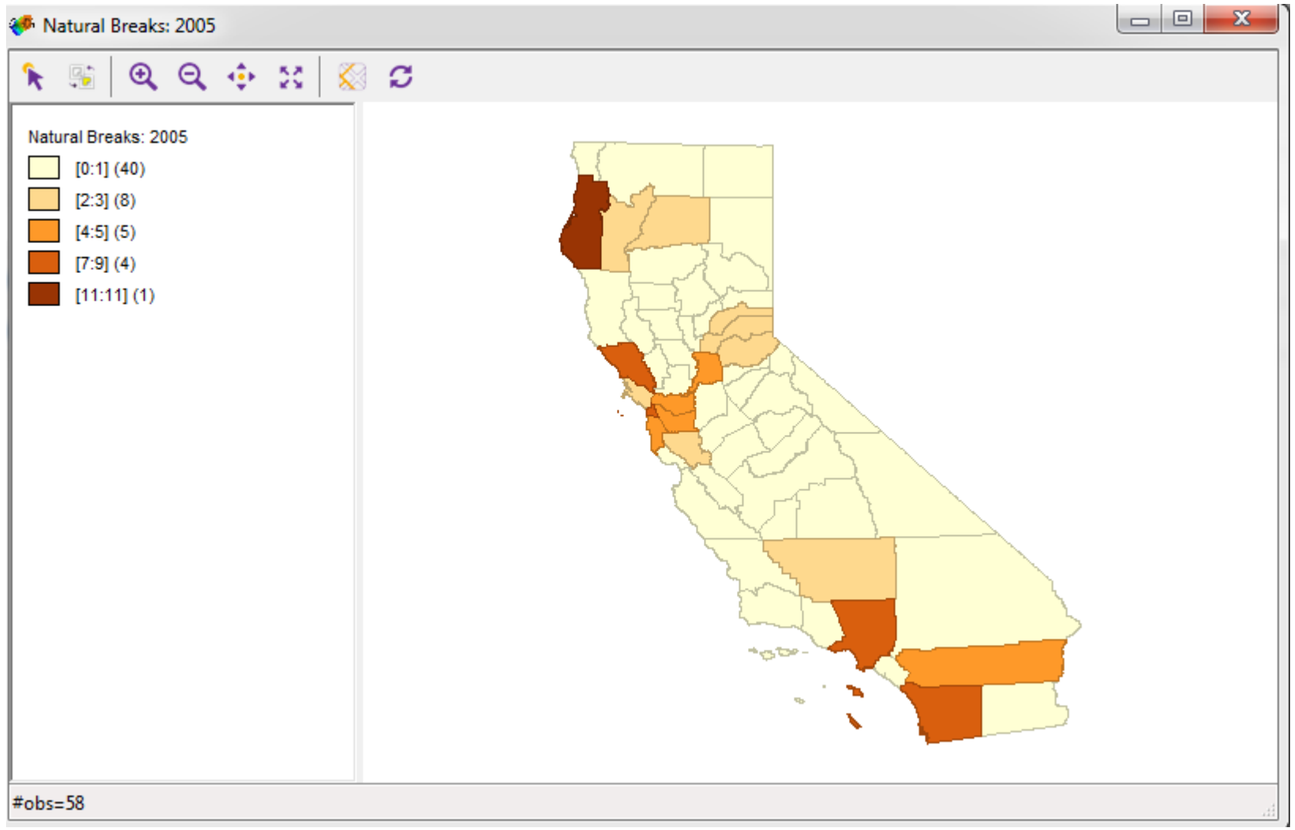
Presentations on Spatial Patterns on GeoDa project website

PDF of powerpoint from USC lecture - very good explanations of key terms

pg. 199- 210 in O'Sullivan and Unwin - overview of Moran's I

Hands-on Exercise: Data

- cnty_Lyme_disease.shp
- Yearly counts of Lyme Disease by county for 2005 - 2014
- Population
- Incidence Rate for Lyme Disease



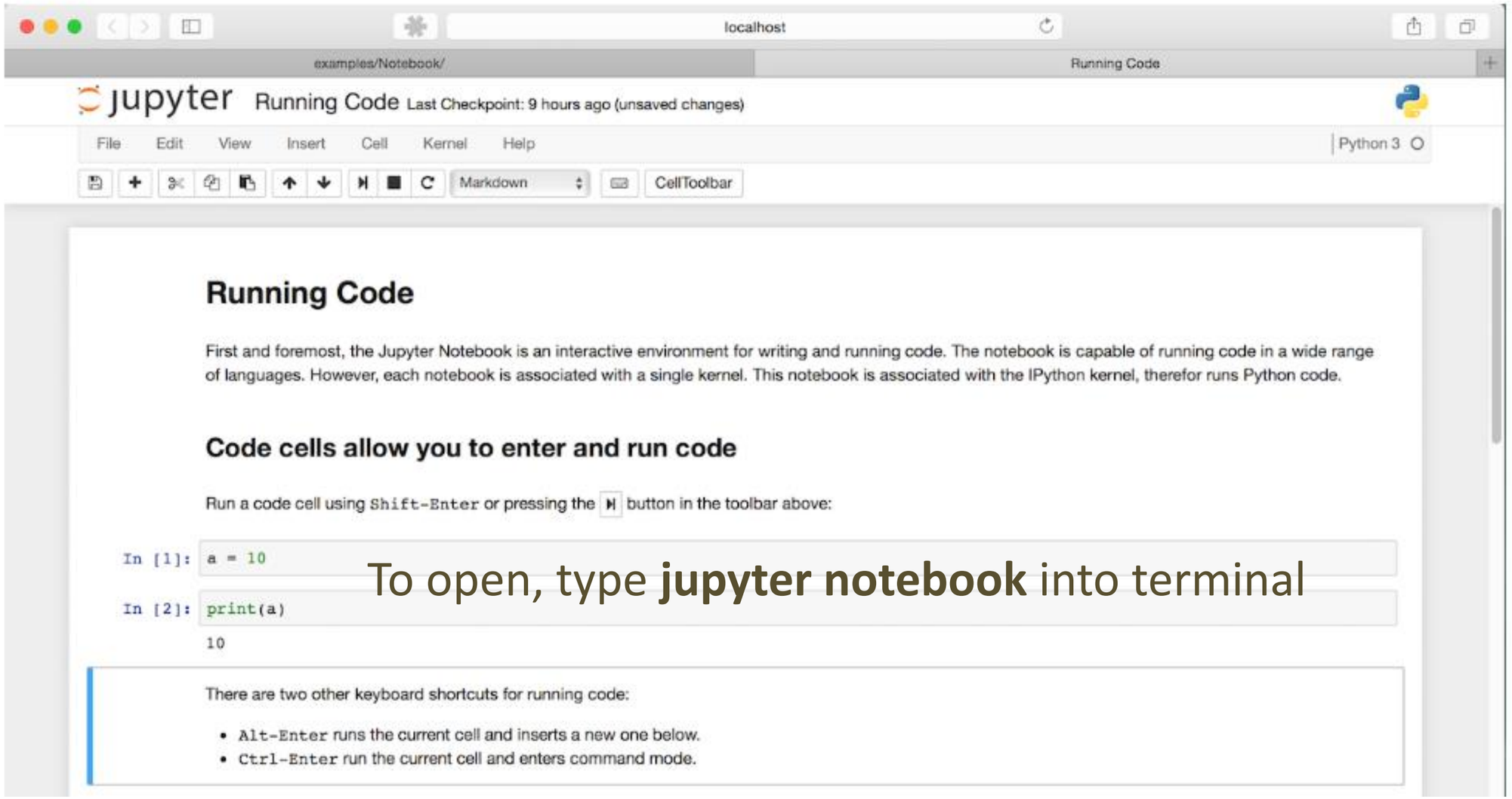
Hands-on Exercise: Outline

1. Run Global and Local Moran's analysis using the GeoDa user interface
2. Write Python code in Jupyter Notebook to calculate Global Moran's I
3. Write Python code in Jupyter Notebook to calculate Local Indicators of Spatial Autocorrelation (LISA)
4. Explore ways to query and visualize data in Python using pandas, geopandas, and folium

Hands-on Exercises: Files

1. GeoDa_Exercise.pdf: instructional guide for exercise in GeoDa GUI
2. PySAL - Spatial Autocorrelation - Global Moran's I - Start with this script.ipynb: Jupyter notebook guiding you through Global Moran's I analysis
3. PySAL - Spatial Autocorrelation – LISA - Start with this script.ipynb: Jupyter notebook guiding you through Local Moran's I analysis (LISA)
4. Resources (folder): Python cheat sheet and PySAL documentation
5. Final Scripts (folder): final Jupyter Notebooks and output files
6. Bonus Data (folder): extra data to play with

Jupyter Notebook Environment



To open, type **jupyter notebook** into terminal

Github

Additional Resources for Python

Online (and free!) Resources for Python

Code Academy (Python programming tutorials)

Coursera (online courses using Python)

Python Beginner's Guide

Python Resources compiled by Berkeley

FREE! Python Books

List of Python Cheatsheets

Python Training Course

ESRI Python tutorial from ESRI

The Hacker Within – Berkeley Chapter

General Useful Jupyter Notebooks for Python

Jupyter Notebooks for Visualization in Python

Other Useful Jupyter Notebooks

Pre-packaged (and FREE) Python Distributions that run PySAL

WinPython for Python 3.x:

- numpy
- scipy
- pandas
- Windows only

Enthought Canopy for Python 3.x:

- numpy
- scipy
- PySAL (in academic option)
- pandas
- shapely (in academic option)
- Windows, Mac, Linux

Anaconda for Python 3.6:

- numpy
- scipy
- PySAL
- pandas
- Virtual Machine images
- Windows, Mac, Linux

Other Python distribution options listed at: http://www.scipy.org/install.html