3600 Morrissey Hall Access Code: 1553

Course philosophy

* The result of using the scientific method is NOT absolute truth.
* More interested in your position and opinion rather than just the correct answer.
* Quality of work is more important that arbitrary due dates.
* Course synthesizes statistics and demography.

Trends

* A lot of traditional social science assumes that space doesn’t matter.
* Spatial demography is becoming computational.

Spatial demography

* Creating maps is a prerequisite for developing the spatial model.

Statistics concepts

* Goal is to create dependent variables that are interval or ratio whenever possible.
  + Allows use of the most sophisticated spatial statistics (i.e., high statistical power).
* Distributions
  + Methods are based on normal distribution
  + Real world data is rarely normally distributed
  + Income distribution is an example of positively skewed data (right skewness)
  + High school education is an example of negatively skewed data (left skewness)
* Ideal situation
  + Normal distribution
  + Small standard deviation
  + 68-95-99 rule
* Errors
  + Type I error is error of commission
  + Type II error is error of omission
  + Type II error is preferred to Type I error
  + Type I errors lead to retractions of journal articles
  + U.S. legal system
    - H0: person is NOT guilty (not the same as innocent)
      * Evidence is insufficient to reject the null hypothesis
    - HA: person is guilty
* Generally, cannot achieve a BLUE model.

Demography concepts

* The scientific study of human populations
* Elements of demography
  + Mathematical knowledge of populations
  + General movement of populations
  + The physical, civil, intellectual, and moral state of populations

The Importance of Space

* Y indicates dependent variable
* X indicates independent variable
* The goal is integrating space as part of the model
* More interdisciplinary research
* Many disciplines are afraid of integration of social sciences

Software and Infrastructure

* ArcMap will be going away; will be replaced by ArcPro and ArcGIS online
* QGIS is open source

Spatial Perspectives

* Functional distance
  + Time and effort required to move from point A to point B
  + Often must create additional data
* Spatial position
  + Comparative advantage and disadvantage of one space to another space
  + e.g., location of new MLS stadium
    - St. Charles residents will be the primary users
    - St. Louis City residents will absorb the costs
* Spatial order
  + Spatial hierarchy within a region
  + Relevant to class and stratification

Spatial Data

* Geodatabases are easier to use than shapefiles
* Quality of shapefile is important
* Can produce errors in spatial analysis
* Dr. Sandoval recommends obtaining shapefiles from U.S. Census

Data Sources

* U.S. Census ([www.census.gov](http://www.census.gov))
* Social Explorer ([www.socialexplorer.com](http://www.socialexplorer.com))

Homework

* Choose an MSA region other than St. Louis MSA or choose variables for St. Louis MSA that are different from the lab.
* Dependent variable should be interval/ratio or dichotomous.

Lab Procedure

* Download data from SocialExplorer.com
  + Click on Tables menu item
  + Click on the ACS 5-year estimates (2014-2018)
  + Click on Begin Report
  + For the geographic type, select Census Tract (140)
  + Select the states of interest
  + For geographic area, select All Census Tracts in the state
  + Click Add
  + Click Proceed to Table
  + Select tables to download
    - A02001 Sex
    - A01001 Age
    - A04001 Hispanic or Latino by Race
    - A14028 Gini index of income inequality
  + Click Add
  + Click Show Results
  + Download for files STATA
    - Tab delimited data
    - STATA .dct file (i.e., dictionary for how to read text files)
    - STATA .do file (i.e., program)
    - Data dictionary .txt file
  + Download for RStudio
    - Comma separated data .csv file
    - Data dictionary .txt file
  + Treat the category Hispanic or Latino as a race
* Creating ratio variables using STATA
  + Double click .do file (i.e., R12432906) to open it
  + \* used to add comments
  + Add folder file path in front of .txt filename in code
  + Add code before infile line code
    - Add line code capture log close
    - Add line code set more off
    - Add line code clear
  + Add code after infile line code
    - Add line code log using analysis.log, replace
  + Copy code for calculating ratio variables from lab .do file on Blackboard
  + Paste code for calculating ratio variable into .do file from data download
    - After line code log using analysis.log, replace
  + Change folder file path in
    - line code for exporting to Excel file
    - line code for creating new STATA database
  + Run program
    - Control-A to highlight the program
    - Click Execute selection(do) menu button
  + Creates new files in lab01/data folder
    - part01.xls
    - stl\_part01.dta
* Download base shapefiles
  + Download shapefiles from U.S. Census using TIGER/line shapefiles database
    - 2019 ACS 5-year estimate
  + Export ZIP files into lab01/data subfolders named for each shapefile
    - Illinois Census tract 17 (tl\_2019\_17\_tract)
    - Missouri Census tract 29 (tl\_2019\_29\_tract)
    - Entire USA (tl\_2019\_us\_cbsa)
* Create new shapefile in ArcGIS
  + Open new map in ArcMap
  + Add each shapefile as a layer
    - tl\_2019\_17\_tract (i.e., Illinois)
    - tl\_2019\_29\_tract (i.e., Missouri)
    - tl\_2019\_us\_cbsa (i.e., entire USA)
  + Create new file geodatabase in lab01 folder
    - Right Click on folder 🡪 New 🡪 File Geodatabase
    - Rename geodatabase (e.g., lab01)
  + Merge Illinois and Missouri shapefiles
    - Select layers for Illinois and Missouri to activate them
    - Deactivate layer for Entire USA
    - Geoprocessing 🡪 Merge
    - Drag and drop Illinois and Missouri layers
    - Set output file path to newly created file geodatabase
    - Enter name for new shapefile (e.g., states01)
      * Use feature class file type
  + Select Entire USA shapefile layer to activate it
    - Deactivate other layers
  + Export St. Louis MSA shapefile
    - Select tract
    - Right click on layer 🡪 Data 🡪 Export Data
    - Export to file geodatabase
    - Name shapefile (e.g., stl\_msa)
      * Use file and personal geodatabase feature class file type
  + Clip census track shapefile to St. Louis MSA shapefile
    - Select stl\_msa shapefile layer
    - Select states01 shapefile layer with census tracts
    - Geoprocessing 🡪 Clip
    - Input feature is states01
    - Clip feature is stl\_msa
    - Output to file geodatabase
    - Name shapefile (e.g., stl\_msa\_ct\_00)
      * Use feature class file type
      * GEOID is unique to each case
  + Merge ratio variables data created using STATA into shapefile for St. Louis MSA
    - Select St. Louis MSA census tracts shapefile layer (stl\_msa\_ct\_00) to activate it
      * Deactivate all other layers
    - Add ratio variables data as a new layer
      * Click on Add Data button
      * Select part01.xls file
      * Select Sheet1$
    - Right click on St. Louis MSA census tracts shapefile layer (stl\_msa\_ct\_00)
      * Join and Relates 🡪 Join
      * Choose GEOID as the basis for the shapefile layer
      * Choose Sheet1$ as the table
      * Use FIPS as the basis for the table
      * Validate join (optional)
      * Click OK
  + Convert to permanent shapefile
    - Right click on St. Louis MSA census tracts shapefile layer (stl\_msa\_ct\_00)
    - Data 🡪 Export Data
    - Output to file geodatabase
    - Name shapefile (stl\_msa\_ct\_01)
    - Select type “File and Personal Geodatabase feature classes”
    - Click OK

ArcMap

* Fragile software; has integrity issues.
* Restart the computer if the software becomes unstable and starts crashing a lot.

Education Attainment

* The amount of education attainment of a region.
* Flexibility with how trade school education is handled.
* Individuals still seeking education are not counted in the index.

Income Inequality

* Can apply concept to other topics such as education.

Gini Concentration Ration

* Measure area of inequality in Lorenz curve.
* Calculated as part of Social Explorer data.

Theil Index

* More powerful than the Gini Concentration Ratio
  + Better mathematical properties.
  + Additive across different subgroups or regions.
  + Local scores add up to global score.
* Not used very often.
* Difficult to interpret.
* Part of General Entropy class which is a family of measures.

Creating Index

* Rescale individual components (i.e., variables) on 0 to 1 scale.
* Default is that each variable gets the same weight.

Class Project

* Creating index for project is optional.

Lab Procedure

* NOTE: Create data dictionary for newly generated shapefiles.
* Creating Excel file (part02.xls) with additional demographic data
  + Download folder lab02
  + Double click the downloaded .do file with the downloaded tract level demographic data (i.e., R12438420) in STATA to open it.
  + Change the file path in front of the .txt filename in the infile line code (line 11).
  + Change the file path in front of the .xls filename in the export function line code (line 152).
  + Change the file path in front of the .dta filename in the save function line code (line 154).
  + Run program
    - Control-A to highlight the program
    - Click the Execute(do) menu button
* Add additional demographic data to shapefile stl\_msa\_ct\_01
  + Open stl\_msa\_ct\_01 shapefile in ArcMap and select it as an active layer.
    - Deactivate all other layers.
  + Click the Add Data button.
  + Add Sheet$1 of part02.xls as a layer and select it as an active layer.
  + Join the Excel data to the shapefile.
    - Highlight and right click on stl\_msa\_ct\_01
    - Select Join and Relates
    - Select Join
    - Select GEOID as the shapefile field on which to base the join.
    - Select Sheet$1 as the data layer for joining.
    - Select FIPS as the data field on which to base the join.
    - Click on Validate Join to check (optional).
    - Click OK
  + Save as new permanent shapefile to the file geodatabase.
    - Highlight and right click on stl\_msa\_ct\_01
    - Select Data
    - Select Export Data
    - Select the file geodatabase (lab01.gdb) as the output location.
    - Name the shapefile stl\_msa\_ct\_02
    - Select type “File and Personal Geodatabase feature classes”
    - Click OK
* Remove census tracts that will skew results.
  + Add stl\_msa\_ct\_02 as a new layer in ArcMap.
    - Deactivate all other layers.
  + Add a new variable named Keep in the attributes table.
    - Right click on the layer.
    - Select Open Attributes
    - Select Add Field from the drop down menu
    - Name the field “Keep”
  + Set default value for the Keep variable
    - Highlight the Keep variable column in the attributes table.
    - Right click and select Field Calculator
    - Enter “1” in the Keep = formula box (where 1 means “yes”)
    - Click OK
  + Identify census tracts to eliminate and remove them
    - Activate the shapefile layer stl\_msa\_ct\_02
    - Select Customize drop down menu
    - Select Toolbars
    - Select Editor
    - Select Start Editing
    - Select the census tracts to eliminate with the selector
    - Right click on the census tract
    - Select Delete Polygon
    - Click Save
    - Select Stop Editing
  + Save the new shapefile to the gis folder.
    - Highlight and right click on modified stl\_msa\_ct\_02 layer
    - Select Data
    - Select Export Data
    - Select the Labs/data/gis folder as the output location.
    - Name the shapefile stl\_ct
    - Select type “Shapefile”
    - Click OK
* Create indexes for statewide data
  + Open gis.do file from the Labs/data/gis folder in STATA
  + Change the file path in line code 3 to the Labs/data/gis folder
  + Run program
    - Control-A to highlight the program
    - Click the Execute(do) menu button
  + Open merge\_final.do file from the Labs/data/lab02 folder in STATA
  + Change the file path in line code 4, 8, 10, and 14
  + Run program lines 1-31
    - Highlight the program code
    - Click the Execute(do) menu button
  + Insert the table from the output after line code 31
  + Replace the minimum and maximum values for the Theil index in the merge\_final.do file
  + Change the file path in the remainder of the program
    - Line code 66
    - Line code 68
  + Run the remainder of the program
    - Highlight the program code
    - Click the Execute(do) menu button
* Add indexes to modified shapefile
  + Open stl\_ct shapefile in ArcMap and select it as an active layer.
    - Deactivate all other layers.
  + Click the Add Data button.
  + Add Sheet$1 of final.xls as a layer and select it as an active layer.
  + Join the Excel data to the shapefile.
    - Highlight and right click on stl\_ct
    - Select Join and Relates
    - Select Join
    - Select GEOID as the shapefile field on which to base the join.
    - Select Sheet$1 of final.xls as the data layer for joining.
    - Select FIPS as the data field on which to base the join.
    - Click on Validate Join to check (optional).
    - Click OK
  + Save the new shapefile to the file geodatabase.
    - Highlight and right click on stl\_ct
    - Select Data
    - Select Export Data
    - Select the file geodatabase (lab01.gdb) as the output location.
    - Name the shapefile stl\_msa\_ct\_03
    - Select type “File and Personal Geodatabase feature classes”
    - Click OK
* Add projection to the shapefile data
  + Open stl\_msa\_ct\_03 shapefile in ArcMap and select it as an active layer.
    - Deactivate all other layers.
  + Click on the ArcToolbox menu icon
  + Select Data Management Tools
  + Select Projections and Transformations
  + Select Project
  + Select stl\_msa\_ct\_03 shapefile as the input
  + Select the file geodatabase (lab01.gdb) as the output location
  + Name the shapefile stl\_msa\_ct\_04
  + Select UTM-15 for the output coordinate system.
  + Click OK
* Activate projection in current ArcMap project
  + Go to layout view
  + Right click and select Properties
  + Select Coordinate System tab
  + Change projection coordinate system to NAD 1983 UTM-15
  + Save changes

Homework

* Select city with at least 250,000 people or MSA with at least 500,000 people.
* Only include census tracks with at least 100 people.
  + All other variables should be okay.
* Additional social data available at PolicyMap
  + Access through SLU Pius XII Memorial Library under databases

Lecture on Geographic Distribution

* Technique of exploratory spatial analysis.
* Spatial variation
  + Non-spatial models may not be valid.
  + Models that apply in one locality are unlikely to be applicable in another locality.
* Why geographic distribution is examined.
  + Identify the center
  + Determine shape and orientation of data
  + Understand the dispersion of the features
* Three Kinds of Center
  + Mean 🡪 average x-y coordinate for all features
  + Median 🡪 x-y coordinate with shortest distance to all features
  + Central Feature 🡪 feature that has the shortest total distance to all other features
* Weighted mean center
  + Not every case should be treated the same
    - land value
    - distinguishing between degrees of poverty
    - different types of crime
  + Sensitive to outliers
* Median center
  + No single equation to calculate median center
  + Approximated by software through iterative calculation
  + Tends to gravitate toward areas with the most features
* Central feature
  + Sum the distance to all other features for each feature.
  + Select the feature with the shortest distance.
* Measuring concentration and dispersion
  + Default is to use one standard deviation when the boundary is known.
  + Use three standard deviation when we don’t know the boundary.
  + Standard distance
  + Standard deviational ellipse (SDE)
* A central city is a heavily populated city at the center of a large metropolitan area [*Central City*. (n.d.). Oxford Dictionary. Retrieved February 6, 2020 from https://www.google.com/search?client=firefox-b-d&q=what+is+a+central+city].
  + Defined and designated by the U.S. Census bureau.
  + There are three central cities (cc) in the St. Louis region.
    - St. Charles, MO
    - City of St. Louis, MO
    - East St. Louis, IL

Lab Procedure

* Ensure that spatial analysis extensions in ArcMap are activated
  + Select Customize dropdown menu
  + Select Extensions
  + Select the extensions you want to activate
  + Click Close
* Create file geodatabase subfolder to organize data from exploratory spatial analysis (ESA)
  + Right click on file geodatabase (i.e., lab.gdb)
  + Select New
  + Select Feature Dataset
  + Name feature dataset (e.g., “esa”)
  + Click Next
  + Select the coordinate system (e.g., NAD 1983 UTM-15)
  + Click Next
  + Click Next to skip the z-coordinate settings
  + Modify XY tolerance settings if necessary (generally, should not be necessary)
  + Click Finish
* Access functions for measuring geographic distribution
  + Select Geoprocessing dropdown menu
  + Select ArcToolbox
  + Select Spatial Statistics Tools
  + Select Measuring Geographic Distributions
* Create shapefiles for mean center
  + Add and activate base layer (i.e., stl\_msa\_ct\_04) for new map in ArcMap
  + Select Geoprocessing dropdown menu
  + Select ArcToolbox
  + Select Spatial Statistics Tools
  + Select Measuring Geographic Distributions
  + Select Mean Center
  + Select stl\_msa\_ct\_04 as the input feature class
  + Select the esa subfolder in the file geodatabase (i.e., lab.gdb/esa) as the output feature class
  + Enter name for the new shapefile (e.g., mc\_wht)
  + Set type to Feature classes
  + Click Save
  + Select the variable upon which to base the calculation (e.g., wht) in Weight Field (optional)
  + Click OK
  + Repeat for other groups as necessary
  + Obtain XY coordinates from attribute table
* Create shapefiles for standard distance
  + Add and activate base layer (i.e., stl\_msa\_ct\_04) for new map in ArcMap
  + Select Geoprocessing dropdown menu
  + Select ArcToolbox
  + Select Spatial Statistics Tools
  + Select Measuring Geographic Distributions
  + Select Standard Distance
  + Select stl\_msa\_ct\_04 as the input feature class
  + Select the esa subfolder in the file geodatabase (i.e., lab.gdb/esa) as the output feature class
  + Enter name for the new shapefile (e.g., sd\_wht)
  + Set type to Feature classes
  + Click Save
  + Select desired standard deviation for Circle Size (e.g., 1\_Standard\_Deviation)
  + Select the variable upon which to base the calculation (e.g., wht) in Weight Field (optional)
  + Click OK
  + Repeat for other groups as necessary
  + Obtain area value from attribute table
  + Create new variable to convert shape area to desired units (optional)
    - Open attribute table
    - Click drop down menu
    - Select Add Field
    - Enter name of field (e.g., area-km)
    - Select the kind of value for Type (e.g., Double)
    - Click OK
    - Right click on the newly added field (variable)
    - Select Calculate Geometry…
    - Select the value to calculate in Property
    - Select the units to use in Units
    - Click OK
* Create shapefiles for standard deviational ellipses (i.e., standard distance taking into consideration geographic features such as rivers and lakes)
  + Add and activate base layer (i.e., stl\_msa\_ct\_04) for new map in ArcMap
  + Select Geoprocessing dropdown menu
  + Select ArcToolbox
  + Select Spatial Statistics Tools
  + Select Measuring Geographic Distributions
  + Select Directional Distribution (Standard Deviational Ellipse)
  + Select stl\_msa\_ct\_04 as the input feature class
  + Select the esa subfolder in the file geodatabase (i.e., lab.gdb/esa) as the output feature class
  + Enter name for the new shapefile (e.g., sde\_wht)
  + Set type to Feature classes
  + Click Save
  + Select desired standard deviation for Circle Size (e.g., 1\_Standard\_Deviation)
  + Select the variable upon which to base the calculation (e.g., wht) in Weight Field (optional)
  + Click OK
  + Repeat for other groups as necessary
  + Obtain area value and rotation from attribute table
    - Rotation is measured counterclockwise
  + Create new variable to convert shape area to desired units (optional)
    - Open attribute table
    - Click drop down menu
    - Select Add Field
    - Enter name of field (e.g., area-km)
    - Select the kind of value for Type (e.g., Double)
    - Click OK
    - Right click on the newly added field (variable)
    - Select Calculate Geometry…
    - Select the value to calculate under Property
    - Select the units to use under Units
    - Click OK

Lecture Notes

Announcements

* No class next Thursday, February 20, 2020
* Can replicate lab for a different city for Homework 01
* Poverty rate can be used as a dependent variable for Homework01 and Class Project

Spatial Interpolation with Grids

* Interpolation is an estimate.
* Administrative boundaries don’t match social boundaries.
* Administrative boundaries may change over time.
* Used to standardize variables.

Modifiable Areal Unit Problem

* Recommendation to no longer use ZIP code data.
  + Will likely result in a Type I error.
* Problems
  + Scale (i.e., aggregation)
  + Grouping (i.e., Zones)
  + Ecological fallacy
* Generally, the larger the spatial units the stronger the relationship among variables.
* Spatial units are arbitrarily defined; different definitions may introduce biases in the analysis.
* Ecological fallacy is that aggregated results are applied to individuals.
* Solutions
  + Normalize boundaries to a specific administration definition of spatial units (e.g., year 2000 census tract boundaries to year 2010 census tract boundaries).
  + Normalize boundaries to standard space (e.g., grid)

Making a Grid

* The smaller the spatial unit the more accurate the data.
* General rule is to use the most accurate data available
  + e.g., census block data are preferred to census tract data if they contain the data for the variable of interest.

Lab Procedure

Create file geodatabase subfolder to organize workflow for grid interpolation

* Right click on file geodatabase (i.e., lab.gdb)
* Select New
* Select Feature Dataset
* Name feature dataset (e.g., “grid”)
* Click Next
* Select the coordinate system (e.g., NAD 1983 UTM-15)
* Click Next
* Click Next to skip the z-coordinate settings
* Modify XY tolerance settings if necessary (generally, should not be necessary)
* Click Finish

Create subfolder in file geodatabase for workflow

* Right click on file geodatabase (i.e., lab.gdb)
* Select New
* Select Feature Dataset
* Name feature dataset (e.g., “grid”)
* Click Next
* Select the coordinate system (e.g., NAD 1983 UTM-15)
* Click Next
* Click Next to skip the z-coordinate settings
* Modify XY tolerance settings if necessary (generally, should not be necessary)
* Click Finish

Define study area

* Start new project in ArcMap
* Add base layer shapefile (e.g., stl\_msa\_ct\_04)
* Right click on layer
* Select Attribute Table
* Sort by COUNTYFP
* Select rows for the City of St. Louis (e.g., COUNTYFP 510)
* Right click on layer
* Select Properties
* Select Defintion Query tab
* Click on Query Builder
* Double click COUNTYFP
* Set COUNTYFP = ‘510’
* Click OK
* Click on Geoprocessing drop down menu
* Select Dissolve
* For Input Feature Class, select the shapefile for the study area (e.g., stl\_msa\_ct\_04)
* For Output Feature Class, select the subfolder in the file geodatabase (e.g., lab/grid)
* Enter name for the new shapefile (e.g., stl\_city)
* Click Save

Create grid

* Add new shapefile for study area as the active layer (e.g., stl\_city)
  + Deactivate any other layers
* Click on Geoprocessing drop down menu
* Select ArcToolbox
* Select Cartography Tools
* Expand Data Driven Pages
* Double click Grid Index Features
* For Output Feature Class, select the subfolder in the file geodatabase (e.g., lab/grid)
* Enter name for the new shapefile (e.g., stl\_grid01)
* Click Save
* For the unit of measure (Polygon Width and Polygon Height), select Meters
* For Polygon Width, enter the desired dimension (e.g., 1000)
* For the Polygon Height, enter the desired dimension (e.g., 1000)
* Click OK

Clip grid to study area boundary

* Click on Geoprocessing drop down menu
* Select Clip
* For Input Feature Class, select the shapefile for grid (e.g., stl\_grid01)
* For Clip Features, select the shapefile for the study area (e.g., stl\_city)
* For Output Feature Class, select the subfolder in the file geodatabase (e.g., lab/grid)
* Enter name for the new shapefile (e.g., stl\_grid02)
* Click Save

Clean partial grids to remove from analysis

* Add and activate base layer shapefile (e.g., stl\_grid02)
  + Deactivate all other layers
* Select the Customize drop down menu
* Select Toolbars
* Select Editor
* Select Start Editing
* Right click on layer
* Select Open Attributes Table
* Sort on the Shape\_Area variable
* Select rows with partial grids less than designated threshold (e.g., 50,000 square meters)
* Right click on selected rows
* Select Delete Selected
* Select Save Edits
* Select Stop Editing
* Click on Geoprocessing drop down menu
* Select Intersect
* For Input Feature Class, add the features to intersect
  + The shapefile for grid (e.g., stl\_grid02)
  + The shapefile for the census tracts (e.g., stl\_msa\_ct\_04)
* For Output Feature Class, select the subfolder in the file geodatabase (e.g., lab/grid)
* Enter name for the new shapefile (e.g., stl\_int01)
* Click Save

Prepare to interpolate census tract data to grid

* Add and activate base layer shapefile (e.g., stl\_int01)
  + Deactivate all other layers
* Right click on layer
* Select Open Attributes Table
* Click on drop down menu
* Select Add Field
* Enter name for new variable (e.g., area)
* Select Double for type
* Click OK
* Highlight the column for the new variable
* Right click on the column
* Select Calculate Geometry
* Verify that it defaults to the correct coordinate system
* Select the desired unit of measure (e.g., square kilometers)
* Click OK
* Click on drop down menu
* Select Add Field
* Enter name for second new variable (e.g., area\_wgt)
* Select Double for type
* Click OK
* Click on drop down menu
* Select Add Field
* Enter name for third new variable (e.g., pop\_new)
* Select Double for type
* Click OK
* Decide on variable to use that is unique to the census tracts (e.g., GEOID)
* Highlight the column for the unique variable
* Right click on the column for the unique variable
* Select Summarize
* Verify that the selection field defaulted to the unique variable (e.g., GEOID)
* Select the summary statistics
  + First new variable (i.e., area)
    - Sum
* For Output Table, you can leave the default (Sum\_Output)
* Click OK
  + Add the result table to the map
* Right click on the working layer (e.g., stl\_int01)
* Select Join and Relates
* Select Join Data
* Select GEOID as the field in the layer upon which the base the join
* Select the result table (e.g., Sum\_Output) as the table for joining
* Select GEOID as the field in the table upon which to base the join
* Click Validate Join (optional)
* For Join Options, select keep all records
* Click OK
  + Sum\_Area variable has been added to the working layer (e.g., stl\_int01)
* Right click on layer
* Select Data
* Select Export Data
* For Output Feature Class, select the subfolder in the file geodatabase (e.g., lab/grid)
* Enter name for the new shapefile (e.g., stl\_int02)
* Click Save

Finishing interpolating census tract data to grid

* Add and activate base layer shapefile (e.g., stl\_int02)
  + Deactivate all other layers
* Right click on layer
* Select Open Attributes Table
* Highlight the column for the second new variable created (e.g., area\_wgt)
* Right click on the column for the variable
* Select Field Calculator
* Enter formula
  + area\_wgt = area / Sum\_area
* Click OK
* Highlight the column for the third new variable created (e.g., pop\_new)
* Right click on the column for the variable
* Select Field Calculator
* Enter formula
  + pop\_new = area\_wgt \* tot
* Click OK
* Check data
  + Sort on GEOID
  + Highlight rows for one GEOID
  + Right click on area\_wgt header
  + Select Statistics
  + Verify that area\_wgt for census tract (e.g., same GEOID) sums to one
  + Right click on pop\_new header
  + Select Statistics
  + Verify that pop\_new variable for census tract (e.g., same GEOID) sums to same value for tot variable of the census tract
* Highlight the PageNumber variable column
* Right click on the column
* Select Summarize
* Verify that the selection field defaulted to the unique variable (e.g., PageNumber)
* Select the summary statistics
  + i.e., pop\_new variable
    - Sum
* For Output Table, you can leave the default (Sum\_Output\_2)
* Click OK
  + Add the result table to the map
* Activate desired layer for interpolation (e.g., stl\_grid02)
* Right click on the layer
* Select Join and Relates
* Select Join Data
* Select PageNumber as the field in the layer upon which to base the join
* Select the result table (e.g., Sum\_Output\_2) as the table for joining
* Select PageNumber as the field in the table upon which to base the join
* Click Validate Join (optional)
* For Join Options, select keep all records
* Click OK
  + Sum\_pop\_new variable has been added to the layer (e.g., stl\_grid02)
* Right click on layer
* Select Data
* Select Export Data
* For Output Feature Class, select the subfolder in the file geodatabase (e.g., lab/grid)
* Enter name for the new shapefile (e.g., stl\_grid03)
* Click Save

Lecture

Spatial Statistics for Spatial Data

* The assumption of spatial randomness is almost never true of the data that has been used in classical statistical analysis.
* Classical statistical modeling answers the question how much of a relationship?
  + Assumes spatial randomness.
* Spatial statistical modeling answers the question how much of a relationship where?
  + Identify spatial patterns.

Spatial Autocorrelation

* The correlation of a variable with itself through space.
  + Must be able to see it on a map.
* Four distinct approaches to defining spatial autocorrelation:
  + Geography
    - Tobler’s first law of geography
    - Near things are more related than distant things.
      * e.g., gentrification
  + Similarity
    - Positive spatial autocorrelation manifests as a clustering pattern.
    - Negative spatial autocorrelation manifests as a checker board pattern.
  + Probability
    - The occurrence of an event in one geographic location makes the occurrence of a similar event in a neighboring geographic location more or less probable.
  + Correlation
    - The correlation between the same attribute at two locations.

Importance

* Evidence of the potential existence of a spatial process.
* Invalidates most traditional statistical inference tests.
  + Introduces bias in classical statistical analysis.
  + Increases the probability of making a Type I error.
* Impact on classical statistical inference.
  + Over-estimates the degree of correlation (i.e., value of coefficients).
  + More likely to find a statistically significant relationship where there is none.

Defining nearness of observations

* Conceptual approaches
  + Common borders
  + Radial boundary
* Must be able to defend methodological choices.
* Contiguity
  + First order 🡪 nearest neighbor
  + Second order 🡪 next nearest neighbor

Null Hypothesis When Testing for Spatial Autocorrelation

* H0: The values associated with the geographic features in one location in the study area DO NOT depend on values associated with geographic features in other locations in the study area.
* Failing to reject null hypothesis means that there is no spatial autocorrelation.
  + Simplifies workflow.
  + Can continue with classical statistical inference tests.

Measuring Spatial Autocorrelation

* Moran’s I is a first regression analysis.
* Regressing a variable on itself.
  + Does the dependent variable predict itself?
* The slope of the regression is the Moran’s I value.
* Variable must be interval-ratio
  + Cannot use dichotomous variable.

Lab Procedure

Method 1: ArcMap

* Add and activate map layer (e.g., ct\_04)
* Click on Geoprocessing menu item
* Select ArcToolbox
* Select Spatial Statistics Tools
* Select Analyzing Patterns
* Double click Spatial Autocorrelation (Moran’s I)
* For the Input Feature Class, add the shapefile (e.g., ct\_04)
* For the Input Field, select the desired variable for analysis (i.e., interval-ratio variables)
* Check Generate Report option box
* Select the Distance Method (e.g., Zone of Indifference, Contiguity\_Edges\_Corner)
* For Standardization, select Row
* Click OK
  + Output describes positive spatial autocorrelation only.

Method 2: GeoDa

* Click on Input File
* Select ESRI File Geodatabase
* Select geodatabase (e.g., lab.gdb)
* Click Select Folder
* Select layer (e.g., ct\_04)
* Open Weights Manager to begin creating weight file
* Click on Create
* Select the ID variable (e.g., GEOID)
* Decide on using Contiguity Weighting or Distance Weighting
* Contiguity Weight tab
  + Select Contiguity Weight (e.g., Queen contiguity)
  + Select Order of Contiguity (e.g., 1)
* Click Create
* Select where to save the file (e.g., the file geodatabase lab.gdb)
* Enter filename for .gal file (e.g., ct\_04\_Moran\_queen)
* Click Histogram to view the distribution
  + Should resemble Normal distribution for irregular polygons (e.g., census tracts)
* Click Connectivity Map for
* Click Connectivity Graph for
* Select Space menu
* Select Univariate Moran’s I
* Select the variable
* For Weights, select the weight file
* Click OK
* Graph is produced
  + Slope of line is the Moran’s I value.
  + Points in the lower left and upper right quadrants indicate positive spatial autocorrelation.
  + Points in the upper left and lower right quadrants indicate negative spatial autocorrelation.
* Evaluate spatial outliers as necessary
* Run Monte Carlo simulation
  + Right click on graph
  + Select Randomization
  + Select permutation
  + Resulting graph displays the pseudo p-value and z-value

Lecture Notes

Difference between Global and Local Spatial Autocorrelation

* Local disaggregates the Global statistic.
* Local has multi-valued statistics that can take on different values at different locations.
* Global statistics cannot be mapped; typically communicated in tables.
* Local is used to identify anomalies to theory.

Moran Scatter Plots

* Moran’s I can be thought of as a plot of X and lag-X (i.e., variable and spatial lag of the variable).
* An adjustment needs to be made for rate-based data (e.g., number of crimes per 1,000 population, number of infant deaths per 1,000 births, etc.)
  + Particularly if the rate denominator varies greatly.
  + Called the EB adjustment (Empirical Bayes Standardization)

Local Indicator of Spatial Association (LISA)

* Also called Cluster Analysis or Outlier Analysis
* Identifies which local values are statistically significant
  + Can reject null hypothesis that there is no spatial autocorrelation.

Lab Procedure

General Comments

* Dr. Sandoval recommends producing LISA maps in GeoDa
  + Calculations can be replicated manually.
* Use ArcMap to create more professional appearing map for presentation and publication.

Method 1: GeoDa

* Prepare for analysis
  + For input file, select shapefile from the file geodatabase (e.g., ct\_04)
  + Select Tools menu item
  + Select Weights Manager
  + Click Load
  + Select the weights file (e.g., ct\_04queen.gal)
  + Click Open
  + Click Histogram to view histogram
  + Close the window
* Create thematic map
  + Select the Maps menu item
  + Select Standard Deviation Map
  + Select variable (e.g., pblk)
  + Click OK
* Create spatial lag variable
  + Select Table menu item
  + Select Add Variable
  + Enter the name of the variable (e.g., slag\_pblk)
  + Select variable attributes
  + Click Add
  + Open the Attribute Table
    - Verify that the spatial lag variable was created
  + Select Table menu item
  + Select Calculator
  + Select Spatial Lab tab
  + Select the weight file (e.g., ct\_04queen)
  + Select the newly created spatial lag variable under Add Variable
  + Select the original variable (e.g., pblk)
    - Equation field should populate as “slag\_pblk = ct\_04queen \* pblk”
  + Click the use row standardized weights option
  + Click Apply
  + Open the Attribute Table
    - Verify that the calculation was successful
* Create scatter plot of spatial lag variable against the variable for Global Moran’s I
  + Select the Explore menu item
  + Select Scatter Plot
  + Select variable for Independent Variable X (e.g., pblk)
  + Select variable for Depended Variable Y (e.g., slag\_pblk)
  + Click OK
    - Scatter plot is created with best fit line (e.g., regression line)
    - Slope of the best fit line is the Moran’s I
* Calculate univariate Local Indicator of Spatial Association (LISA)
  + Select the Space menu item
  + Select Univariate Local Moran’s I
  + Select the variable (e.g., pblk)
  + Select the Weights file (e.g., ct\_04queen)
  + Click OK
  + Select the windows to open
    - Significance Map
    - Cluster Map
    - Moran Scatter Plot
  + Click OK
* Evaluate the degree to which two variables are correlated with each other globally
  + Select the Space menu item
  + Select Bivariate Moran’s I
  + Select the first variable (e.g., pblk)
  + Select the second variable (e.g., edtot)
  + Select the Weights file (e.g., ct\_04queen)
  + Click OK
  + Obtain significance levels
    - Right click on a window
    - Select Randomization
    - Select number of permutations (e.g., 999 permutations)
    - Click Run to re-run the permutations
* Evaluate the degree to which two variable correlate with one another locally
  + Select the Space menu item
  + Select Bivariate Local Moran’s I
  + Select the first variable (e.g., pblk)
  + Select the second variable (e.g., edtot)
  + Select the Weights file (e.g., ct\_04queen)
  + Click OK
  + Select the windows to open
    - Significance Map
    - Cluster Map
    - Moran Scatter Plot
  + Click OK
  + Obtain significance levels
    - Right click on a window
    - Select Randomization
    - Select number of permutations (e.g., 999 permutations)
    - Click Run to re-run the permutations
* To copy maps
  + Right click on window
  + Select Copy Image to Clipboard
* Export results (e.g., LISA) to process in ArcMap (for publication)
  + Right click on any window
  + Select Save Results
  + Select results to save
  + Enter names for the new variables to add to the attribute table
  + Click OK
  + Select the File menu item
  + Select Save As
  + For file path, select ESRI shapefile
  + Select the location to save the shapefile (e.g., file geodatabase, shapefile folder)
  + Enter the name for the shapefile (e.g., geoda\_stl.shp)
  + Change field names if necessary
  + Click OK

Method 2: ArcMap

* Add shapefile as map layer (e.g., ct\_04)
* Select Geoprocessing menu item
* Select ArcToolbox
* Expand Spatial Statistics Tools option
* Expand Mapping Clusters option
* Double click Cluster and Outlier Analysis (Anselin Local Moran’s I)
* For Input Feature Class, add the shapefile (e.g., ct\_04)
* For Input Field, select the variable (e.g., pblk)
* For Output Feature Class, select the file geodatabase (e.g., lab.gdb)
* Enter name for the new shapefile (e.g., lisa\_pblk)
  + Click Save
* For Conceptualization of Spatial Relationships, select desired contiguity (e.g., contiguity\_edges\_corners)
* For Standardization, select Row
* Click OK
  + LISA map added as a new layer
* Export the LISA map layer as a new shapefile
  + Right click on layer
  + Select Data
  + Select Export Data
  + For Output Feature Class, select the subfolder in the file geodatabase (e.g., lab/lisa)
  + Enter name for the new shapefile (e.g., stl\_lisa\_pblk)
  + Click Save

Lecture Notes

Class Projects

* Presentations for individual projects will be conducted on April 30, 2020.
* Presentations for group projects will be conducted on May 7, 2020.
* Only required to attend the presentation session for individual projects.

Spatial Dependence

* There is spatial autocorrelation.
* Assumes the relationship is the same throughout the study area.
* BLUE models (best linear unbiased error)
  + Spatial dependence violates the assumptions for BLUE models.
* Statistics are biased upwards
  + Correlation coefficients appear larger than they are.
  + Coefficients of determination appear larger than they are.
* Precision of the model is exaggerated
  + Standard errors appear smaller than they are.
* Variables may be independent in theory but not in practice for several reasons:
  + Structure
  + Policy
  + Racism
* There are numerous types of spatial regression models.
  + This lecture focuses on two of them.
  + Future lectures will discuss other types of spatial regression models.

Ordinary Lease Squares (OLS)

* Assumes that space does not matter
  + No relationship occurring spatially
* No relationship between yi and yj

Spatial Lag Models

* Theoretical model (i.e., substantial issue)
  + Has a theoretical interpretation.
  + Preferred to spatial error model.
* Something is happening that needs to be included in the model
* There is a relationship between yi and yj
* Introduces new spatial lag variable.
* Assumes no correlation between the unexplained component (i.e., error component).

Spatial Error Models

* Estimation problem that must be fixed (i.e., nuisance).
  + More mathematical in nature.
  + Does NOT have a theoretical interpretation (i.e., doesn’t change the underlying theory of the model).
* Assumes errors from the model are spatially correlated.
* Assumes no relationship between yi and yj
* Introduces new error term.

Goodness of Fit Statistics

* Difficult to use R2 in spatial regression models
  + Adjusted R2 is a pseudo-R2; should not be used in articles for publications
* Log likelihood (LL) and Akaike Information Criteria (AIC) are recommended
  + For LL, higher values indicate better fit.
  + For AIC, lower values indicate better fit.

GeoDa Decision Tree

* First perform OLS regression
  + Assumes that the null hypothesis will not be rejected.
  + No need to perform more sophisticated spatial regression if OLS regression is not significant.
  + Check the residuals for spatial autocorrelation (i.e., Moran’s I)
* Perform both spatial lag (SL) model and spatial error (SE) model.
* Compare goodness of fit of spatial lag model and spatial error model.
* When Rho (ρ) is not statistically significant but SL model is better than OLS model, it indicates multicollinearity between the errors (residuals) and the independent variable.

Spatial Weights

* Analysis is based on the assumptions of the spatial weight matrix.
* Must consider how neighbors interact with each other.
* Not possible to model every single possible interaction.
* Spatial weights impose structure on the data by considering how connected neighbors are.
  + Quantify spatial similarity
* There is no optimal weight matrix
  + Weight matrix is a methodological choice.
  + Must be able to defend choice.
  + Goodness of fit used to help justify methodological choice.
* K-nearest neighbor (KNN) method
  + Method is asymmetric
  + GeoDa can’t estimate spatial lag and spatial error models with asymmetric weights.
    - Use GeoDaSpace

Lab Procedure

Data for lab

* Download files from Blackboard.
* Homicide data is just for lab; don’t use in paper for publication.

Create new shapefile for study area use ArcMap

* Add and activate base layer (e.g., ct\_04).
* Double click on shapefile layer to open Properties
* Enter a Definition query for City of St. Louis census tracts
  + County code = 510
* Click OK
* Add homicide data table as layer
* Join homicide data table with shapefile
* Export as a new permanent shapefile

Create new variables

* Open Attribute table
* Select Add Field from dropdown menu
  + Enter variable name (e.g., hom\_rate01)
  + Select type of variable
* Repeat for second variable (e.g., hom\_rate02)
* Calculate value for first new variable
  + Select Field Calculator
  + Highlight variable column
  + Enter formula (i.e., hom\_rate01 = Sum\_count / tot)
  + Click OK
* Calculate value for second new variable (per 1,000)
  + Select Field Calculator
  + Highlight variable column
  + Enter formula (i.e., hom\_rate02 = hom\_rate01 \* 1000)
  + Click OK
* Export as a new permanent shapefile

Evaluate spatial autocorrelation using GeoDa

* Create a weight matrix for the new shapefile.
  + OBJECTID is the unique identifier for each case.
* Evaluate spatial autocorrelation for the variables.
  + Calculate global univariate Moran’s I for variables.
* Evaluate bivariate LISA for the variables.

Conduct OLS regression without weights matrix using GeoDa

* Select Regression menu item
* Enter variables
* Select Classic model
* Check the Pred. Val. And Res. option
* Check the Coeff. Var. Mat. option
* Check the White Test option
* Click Run
* Click Save to Table
* Rename variables if necessary
* Click OK

Conduct OLS regression with weights matrix using GeoDa

* Select Regression menu item
* Enter variables
* Select the weight matrix file
* Select Classic model
* Check the Pred. Val. And Res. option
* Check the Coeff. Var. Mat. option
* Check the White Test option
* Click Run
* Click Save to Table
* Rename variables if necessary
* Click OK
* Calculate the global univariate Moran’s I for the residuals variable to determine if spatial regression model is necessary.
* Check if the residuals value (i.e., error variable) is significant to determine if spatial error model is necessary.

Conduct spatial lag regression using GeoDa

* Select Regression menu item
* Enter variables
* Select the weight matrix file
* Select Spatial Lag model
* Check the Pred. Val. And Res. option
* Check the Coeff. Var. Mat. option
* Click Run
* Click Save to Table
* Rename variables if necessary
* Click OK
* Calculate the global univariate Moran’s I for the residuals (i.e., errors) variable
  + It should indicate negligible spatial autocorrelation

Conduct spatial error regression using GeoDa

* Select Regression menu item
* Enter variables
* Select the weight matrix file
* Select Spatial Error model
* Check the Pred. Val. And Res. option
* Check the Coeff. Var. Mat. option
* Click Run
* Click Save to Table
* Rename variables if necessary
* Click OK
* Calculate the global univariate Moran’s I for the residuals variable
  + It should indicate negligible spatial autocorrelation

Lecture Notes

Spatial autoregressive models with GeoDaSpace

* Spatial econometric models
* Economics paradigm

Building Models

* Process
  + Research question
  + Theoretical framing
    - Intellectual (i.e., theoretical contribution)
      * New knowledge
      * Demonstrating flaws in previous approaches
    - Test at least one hypothesis
    - Most papers test two additional hypotheses
  + Types of models
    - Baseline
    - Reduced model (i.e., model with the hypothesis)
      * Focus on primary variables
      * The intellectual contribution
    - Full model
      * Baseline plus the reduced model.
* Factors for evaluating model
  + Theory
  + Complexity
    - Other models that could have been used
    - Why other models were not used
  + Parsimony and power
    - R2 is used as a proxy for power
    - Most powerful models are not necessary the best model
    - Fewest variables to produce the highest power
* Approaches to model building.
  + Evaluate variables of interest separately then aggregate.
  + Add variables to baseline separately then aggregate (Sandoval’s preference)
  + Use hierarchical modeling.

Model Assumptions

* Spatial heterogeneity versus spatial dependency
  + Intensity equally distributed (i.e., heterogeneity).
    - Magnitude of relationship between dependent and independent variables (i.e., coefficient values).
  + Intensity at one location influences intensity at neighboring locations (i.e., dependency)
    - Spatial autocorrelation
* Heterogeneity
  + Influencing factors
    - Spatial structure
    - Spatial process
  + Assumes coefficients are fixed throughout study
  + Types
    - Structural stability
    - Heteroscedasticity

Spatial Regime Models

* Spatial fixed effects models
* Subsets in data correspond to regions or spatial clusters
* For each regime added you lose a degree of freedom
* Sample size lower limit rules of thumb
  + Two-thirds of degrees of freedom not used
  + At least 40 observations for spatial regression model with up to 2 variables

Higher order spatial regression

* Endogenous variables
* Exogenous variables
* Instrumental variable

Software packages

* GeoDa
  + Capable of producing visualizations (i.e., maps)
  + No heterogeneity present in the data.
  + Lag and error models are mutually exclusive.
* GeoDaSpace
  + No visualization (i.e., maps)
  + Heterogeneity is present in the data.
  + Can combine lag and error models.
  + Use maximum likelihood (ML) as the default method.

GeoDaSpace

* Components
  + Spatial data
  + Spatial weight matrices
  + Spatial regression
* Handling multicollinearity
  + Too many variables are alike.
  + Variables may be statistically significant in either SLM or SEM but becomes statistically insignificant in SLM and SEM combined.
  + One or more of the independent variables may be mathematically identical to spatial lag or spatial error variable.

Lab Procedure

Code regimes using theory (i.e., manually)

* Open the Attribute Table
* Select Add variable from the menu
* Add new variable to code the regimes (e.g., region)
* Close the Attribute Table
* Click the Select Feature tools
* Select the tracts for the region
* Open Attribute Table
* Code highlighted tracts
  + Regimes need to be contiguous unless using distance for spatial weight matrix)

Code regimes using statistical technique

* Calculate univariate LISA for variable
  + Select the Geoprocessing menu item
  + Select ArcToolbox
  + Expand the Spatial Statistics Tool option
  + Click on Cluster and Cluster Analysis (Anselin Moran’s I)
  + Select the shapefile for the Input Feature (e.g., stl\_city03)
  + Select the dependent variable for the Input Field (e.g., pov)
  + Select the file geodatabase for the Output Feature Class (e.g., lab.dbf)
* Assign regimes based on LISA clusters
  + Regimes need to be contiguous unless using distance for spatial weight matrix)

Code regimes using mathematical criteria

* Standard deviational ellipses at one standard deviation for a variable (e.g., pct Black population)
* Census tracts with a minimum value for variable (e.g., at least 90% Black population)
* Regimes need to be contiguous unless using distance for spatial weight matrix)

Calculate Regression Models Using GeoDaSpace

* Select the file geodatabase for the Data File (e.g., stl\_city03.dbf)
* Load spatial weights matrix in the Model Weights field
* Drag the dependent variable in the Y field
* Drag the independent variables into the X field
* Drag the endogenous variables into the YE field
* Drag the instrumental variable into the Instruments field
* Drag the regimes variable into the Regime field
* Create first model
  + Select Standard for Model Type
  + Select OLS for Method
  + Click Run
* Create second model
  + Select Standard for Model Type
  + Select ML for Method
  + Click Run
* Iterate through options for each model needed for analysis

Iterating models

* Variables might not be statistically significant because of heteroscedasticity.
  + Correct for by trimming standard error.
  + Use GMM method.