GEOGRAPHY 5303
Exercise 7 – 100 points
Due: April 18, 2020, by 5 PM
Spatial Autocorrelation Statistics

Purpose of Exercise

Here we explore spatial autocorrelation (SAC) from two perspectives. First (Part A), discrete point patterns are studied in an attempt to identify statistically-significant evidence of clustering or dispersion, using two complementary methods (e.g. nearest neighbor and quadrat analyses). These are useful, though limited, techniques in spatial analysis for detecting SAC.

Second, interval-ratio tests applied to point and areal data raise the level of sophistication. The ability to identify patterns for point and areal data, with measured quantities for point locations or for areal units, adds to the geographer's toolbox. To facilitate analysis in Parts B-D, you will be using the free software package GeoDa (see exercise supplement for installation details and other tips). Data files in various formats, and some maps, are available in Canvas. Be sure to insert all assigned maps and other graphics into your Word document when indicated and where appropriate to the discussion.

A. Binary Point Pattern Analyses

This analysis examines the locations of the 73 earthquakes of magnitude 2.5 or greater that occurred in Oklahoma and southern Kansas between March 19 and April 16, 2017. The uptick in earthquakes in Oklahoma and Kansas was been the topic of much debate and research at the time, and we will jump on the bandwagon with our own analyses of the phenomenon. Data were downloaded in CSV format from the US Geological Survey (USGS), and you can explore what all the fields mean if you wish (although here we are solely interested in the latitudes and longitudes) from: https://earthquake.usgs.gov/earthquakes/feed/v1.0/csv.php

These data were imported into GIS for the purpose of making maps (available on Canvas) showing the earthquake locations. Clearly at the national scale (Map 1), there was a cluster in the central Plains states, but we are going to zoom in and see if there was clustering within our region (Map 2). Recently earthquake activity has been much lower, which is why we are studying the "hot spot" of spring 2017.

Note that both questions A.1 and A.2 require the same basic steps, which are listed a-e. below instead of being repeated each time (PLEASE use a-e. labeling in your Word document):

a. Pre-analysis discussion: What pattern (if any) do you think is most evident from a visual inspection of the map that you perform prior to analysis?
b. Hypotheses: State appropriate null and alternative hypotheses, explaining the rationale/format behind both and what they mean in concrete terms (i.e. don't just list a generic hypothesis without elaboration connecting to the specific problem).

c. Computations: Clearly display your computations of the appropriate statistic (show and identify all sub-components) and the p-value/significance in Excel. **Report final test statistics and p-values in your Word document**.
d. Decision: Explain what decision you reach regarding the pattern of the phenomenon under study, provide an interpretation of this result (i.e. the bigger picture), and discuss whether or not this result matched your prediction and why this happened.
e. Post-analysis assessment: Discuss the theoretical and practical issues associated with each analysis method. Provide all pertinent information, decisions, and rules regarding your research methodology with respect to each analysis.

1. Quadrat Analysis (QA) – Impose quadrats on Map 2 and determine if any statistically significant pattern is evident. Your map should show the cells clearly marked, and consecutively number each cell included in your analysis. Insert this cell map into Word, whereas Excel should include a table listing each numbered cell and its tally of earthquakes (in adjacent columns) in addition to your computational work. It will be particularly important that you provide a detailed answer for part e. of this question as you will need to explain your rationale for your choice of quadrat size (cite pertinent "rules" governing optimum quadrat sizes), shape, orientation, and boundary rules (internal and external). Also, be sure to use the correct statistical test depending on how many quadrats you have.

   Note: you will have difficulty tabulating all 73 earthquakes from Map 2, so Map 3 provides counts for small clusters. Feel free to import the raw data (Excel file provided) into GIS and impose quadrats to tabulate your counts, if your skills permit.

2. Nearest Neighbor Analysis (NNA) – Re-examine the earthquake location problem using the nearest neighbor technique (see instructions in Appendix for easily identifying each earthquake's nearest neighbor). Part e. of this problem will be more related to the limitations and issues of the test itself, since you have very little control over the geography or set-up of this problem like you did in Question A.1. For this problem, assume a rectangular box covering exactly 4.0 square decimal degrees for the study area. This is a box just large enough to contain all 73 earthquakes; basically, the extent shown in Map 2 can be considered the study area.

3. Summarize the overall results of your study of Oklahoma and Kansas' earthquakes; compare/contrast outcomes, discuss the implications of your findings, and point out any shortcomings (big picture stuff, not the details you may have mentioned in part e. of Questions A.1-2) of what this analysis does and does not tell you. Overall, is there any pattern to earthquakes *within* our defined study area? Why or why not?

B. SAC Analysis of Earthquake Intensity (Point Data)

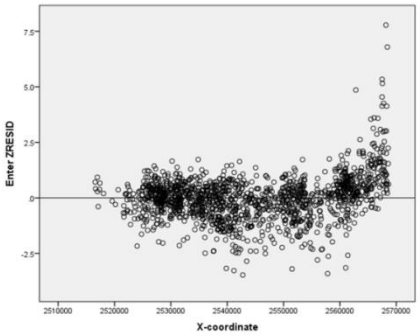
We will study the earthquakes a bit more here, now focusing on the intensity as measured by magnitude ("mag") and determine if there is any global SAC of earthquakes based on this attribute. You will need to create a shapefile from Excel data (KSOK Earthquakes.xls) as outlined in the exercise GeoDA supplement. Since we have already studied the earthquake pattern with NNA, create a weights matrix based on 6 nearest neighbors for this part.

1. Run global Moran's I on the earthquake magnitudes and discuss the results, being sure to include or do the following:
   a. Discuss the value of I
   b. Insert the Moran scatterplot in your write-up
   c. Right click on the scatterplot, choose "Randomization", and choose "Other" and enter 10000 permutations. Once the screen and histogram come up, hit "Run" several times and then report a representative z-score and pseudo p-value and interpret its meaning.

2. Evaluate/compare/contrast these results with those obtained in Part A. Does it matter if we only look at location, or does magnitude change the findings?

C. SAC Analysis of Regression Residuals for Milwaukee Home Sales, 2012 (Point Data)

Recall on Exercise 4 (B.2) you ran a kitchen-sink regression on all 11 analyzable variables (including dummy variables) pertaining to home sale prices in Milwaukee in 2012. Specifically, you mapped the residuals and possibly noticed some patterns and hypothesized about those patterns. Besides a spatial pattern, the residual plot (versus the X (E-W) coordinate, at right) definitely looks bad and we are likely violating at least two regression assumptions.



You will examine the residuals in this section, but also conduct a sensitivity analysis under different weighting schemes. For this part of the assignment you must also create a shapefile from an Excel file (Milwaukee_Sales_2012_withXY.xls) in order to redo the regression analysis originally done in Exercise 4 in SPSS.

1. Create three weights files and name them appropriately (GeoDa will supply the correct extensions so only type the parts before the periods of the file names):
   a. First-order queen's contiguity (QUEEN.gal), found under the "Contiguity Weight" tab.
   b. K-Nearest neighbors (NN#.gwt), found under the "Distance Weight" tab and "K-nearest neighbors" as the method. You can choose any NN value but be sure to indicate this in your answers and replace # in the file name with your NN choice.
   c. Threshold distance (THRESHOLD.gwt), found under the "Distance Weight" tab and "Distance band" as the method. The default value of 4853.001470 (meters)

that pops up is the *longest* distance between *nearest* neighbors, ensuring every observation has at least one neighbor. Change this to an even 5000.

d. Create (and insert in Word) the Connectivity histogram for each weighting scheme and discuss the differences that exist in the distributions of the numbers of neighbors each county has, as well as any other oddities you may uncover.

While visually searching for patterns in regression residuals (Exercise 4, Question B.2.d) is valuable, simply classifying the values into categories and evaluating the result does not tell us whether any patterns represent a *significant* deviation from randomness. Furthermore, a residual of -0.02 is not much different than a residual of 0.02 as they both represent extremely good predictions for those observations, but failure to account for *magnitudes* (and signs) limits the strength of the analysis. Moran's I, on the other hand, provides a tool for evaluating the pattern of measured values of observations.

2. Run a basic regression on the dataset replicating the analysis done on Exercise 4, Part B. Click "Methods", "Regression", and then put SalePrice in the "Dependent Variable" box and all 11 *analyzable* (interval-ratio and dummy, but NOT Alderman district) variables in the "Covariates" box. "Classic" (OLS) should already be checked, but also check "Pred. Val. and Res.", then "Run".

   A Regression Report window will open, and now that you have run the regression, you can choose "Save to Table" and "Run" again. This option allows you to save the regression predicted values and/or residuals to your data table. We only need the residuals; I recommend adding them *after* your last variable instead of before the first variable. Note that these are raw residuals, not standardized ones (z scores) like we are used to working with, but GeoDa eventually computes standardized results. We *could* create our own standardized residuals in GeoDa under "Table", "Variable Calculation" if we needed to transform a variable for some reason – just saying. Also, you can choose "Save to File" and the results of the report window can be saved to a text file for future reference. Also just saying.

   Review/discuss the important regression statistics available in the top of this output, and compare them to the results you got for Exercise 4, Question B.2.a and B.2.c.

3. Compute *global* Moran's I for your regression residuals by choosing "Space", "Univariate Moran's I" and produce a Moran scatterplot for your residuals (OLS_RESIDU) using each of your three weighting files. Also produce a 10000 permutations histogram for each weighting scheme (click "Run" multiple times but just include one representative histogram). Panel each scatterplot and its histogram side-by-side in Word (3 pairs, one for each weighting file) and be sure to label them according to weighting scheme.
   a. Assess the amount and direction of spatial autocorrelation in the residuals.
   b. Assess the significance of I values based on the Z scores in the histograms.
   c. Is the level of SAC particularly affected by the choice of weighting scheme?
   d. Does the regression assumption of non-autocorrelated residuals hold?

D. SAC Analysis of SES Variables for Oklahoma (Areal Data)

The full Oklahoma SES dataset has been converted into a shape file for your use, and then put into a single ZIP file (5303_EX7.zip). Unzip these files and then open them in GeoDa; since you're opening a shape file GeoDa is ready to go (woo-hoo!).

On Exercise 6, Part B, you chose 4-6 (out of your original 10 assigned) variables to undertake a cluster analysis. Here, choose the *three* variables (which could include Pct_Repub, the dependent variable on the regression analyses – it's up to you) that interest you the *most*, either from their intrinsic properties or else how they performed in regression, factor, and/or cluster analyses. Be sure to clearly identify each of the three variables you are using as you discuss each one. For this section, you are assigned first-order Queen contiguity.

Create I_i *and* G_i^* cluster and significance maps for *each variable*. For I_i, click "Space", "Univariate Local Moran's I", choose a variable and click "OK", check all three boxes in the "What windows to open" dialog, and "OK". While you have your Moran scatterplot on screen, right-click, choose "Randomization", and "10000 Permutations". Make sure you screen capture and/or save all these graphics.

For G_i*, click "Space", "Local G*"', choose a variable and click "OK", and select both boxes in the "What windows to open" dialog, and leave row-standardized weights selected before hitting "OK".

For each variable, discuss the presence of SAC clusters, their significances, and the similarities/contrasts for both I_i and G_i^*. Also, discuss the Moran's scatterplot and 10000 permutation histogram to discuss the overall level of SAC for each variable. Panel all these graphics in an efficient fashion in Word (*i.e.,* do NOT devote one whole page to each individual graphic). Don't comingle discussions about each variable.

1. Discussion and graphics for first variable.
2. Discussion and graphics for second variable
3. Discussion and graphics for third variable.

APPENDIX: OBTAINING A DISTANCE MATRIX IN SPSS

The dataset for Part A contains information on 73 earthquakes in the area in spring 2017. Import the Excel file into SPSS, then go to *Analyze, Correlate, Distances,* put **latitude** and **longitude** in the "Variables:" box, put **Quake** in the "Label Cases by:" box, and then click *OK*. This will produce a distance matrix that can be exported back to Excel. Once back in Excel, find the distance from each quake to its nearest neighboring quake (pay attention to units!). Be aware that the shortest distance in each row or column is 0.000 (on the diagonal) which is the distance between each quake and itself; you need to determine the minimum (next-smallest) value, which should be non-zero. The Excel function =small will be exceedingly helpful.