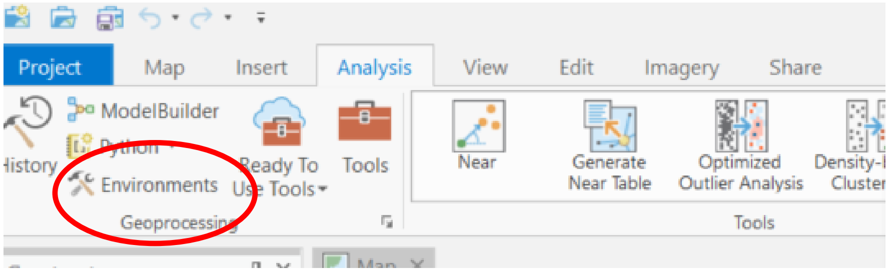
Lab 9: Introduction to Spatial Statistics: *Descriptive spatial statistics and spatial autocorrelation*

CRP 4080: Introduction to GIS **Due:** Nov 8, 2024
Fall 2024

Prof. Wenzheng Li (wl563)
Lab TAs: Gauri Nagpal (gn247), Anika Sinthy (ats243), Shubham Singh (ss373)
Location: Sibley 305, Barclay Gibbs Jones Computer Lab
Total points: 60

Environmental setting

Note: *Before we begin, please create an 'output' folder which you will set as your workspace*
These settings allow you to ensure that geoprocessing is performed in a controlled environment where you decide things such as the processing extent that limits processing to a specific geographic area, a coordinate system for all output geographic datasets, and the cell size of output raster datasets. You can specify geoprocessing environment settings once for your project using the **Environments** window. These settings are saved with your project and will be automatically used by all tools that honor the environments. Open this window by clicking **Environments** on the **Analysis** ribbon tab.



Under the "workspace" tab, set the current (and scratch) workspace as the 'output' folder you just created. You can navigate to your output folder using the (📁) button.

Part 1: Geographic and Population Centers

Centrality*:* Centrality is important in all sorts of mapping applications where we want to know the location that minimizes distances travelled. For our purposes this may just be another way of describing our data or it may be the first step in understanding the spatial structure of our data. Measuring centrality by itself is perhaps not the most informative measure—in the case we will explore this week we could probably make a pretty good guess by just eyeballing our map. However, when we conduct our tests on a subset of our data and/or weight our selection based on some attribute of interest, we can begin to generate statistical information about the spatial distribution of our data.

1. Central Feature

The “Central Feature” tool in ArcGIS Pro works on vector data sets to find the single feature that minimizes the total distance from the centroids of all other features.

Under the Map tab, use the “add data” button (+) to add “Seattle_blockgroups.shp” to the map. The file is found in this lab’s “data” folder and shows all census block groups in the city of Seattle. If you examine the attribute information, you can see that the data has already been cleaned and normalized.

1A. Under the Analysis tab, select Tools. In the Geoprocessing window, type “Central Feature” and select the appropriate tool. The Central feature dialog box should now be visible. For the *Input Feature Class* specify “Seattle_blockgroups.” The *output feature class* will now default to the workspace you previously set up (that’s why we did this!). Leave the Distance Method as Euclidean. Click Run. A new shapefile is created indicating the central block group.

1B. Now, repeat the above, but under *Weight Field* specify “TotalPop.” The weighted test just counts each polygon k times where k is the population of that block group. The result is very similar. What does this suggest?

One of the key reasons we may be interested in this information at this stage is as a basis for identifying our study area. When using spatial statistics, it is often the case that we will introduce bias into our results because observations on the edge of our study area may not have the same number of neighbors, and consequently less detailed information about their environs, as equivalent observations that are centrally located within our study area. As such, a careful understanding of the shape and centrality characteristics of our data will be necessary to help interpret later findings.

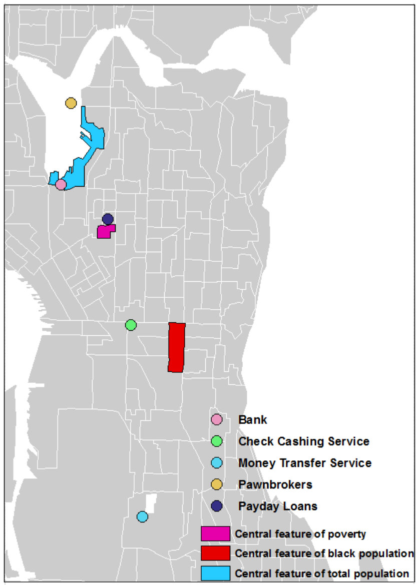
1C. Now, repeat the above, but in the *Weight Field* specify “PctBlack” instead of “TotalPop.”

Question #1: Interpret your results by comparing the three central features calculated in steps 1A, 1B, and 1C. 5 points.

2. Mean Center

The “Mean Center” tool gives nearly identical results to those above, but returns a point instead of a polygon. “Mean Center” is often more appropriate than “Central Feature” when we have event data since we do not have any expectation that the point of interest lies on top of one of our observations (a characteristic that is enforced by the “Central Feature” tool).

2A. Add “Financial_Points.shp” to ArcGIS Pro. If you open the attribute table, you will note that the data is organized according to the type of financial institution (Banks, Payday loans, Check cashing services, pawnbrokers, and money transfer services).

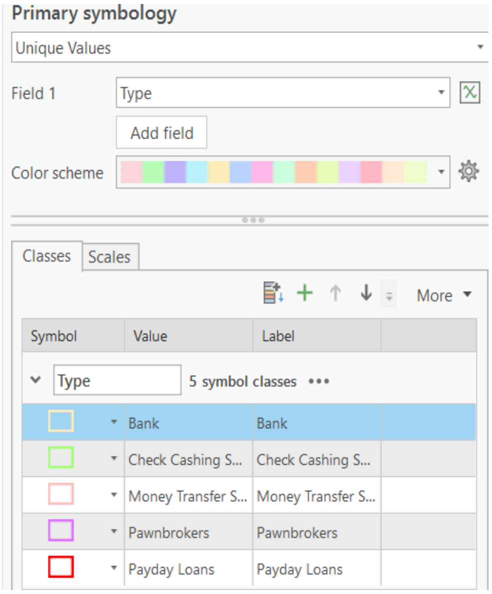

Bank
Check Cashing Service
Money Transfer Service
Pawnbrokers
Payday Loans
Central feature of poverty
Central feature of black population
Central feature of total population

2B. Under the Analysis tab, select Tools. In the Geoprocessing window, type "Mean Center" and select the appropriate tool. The Mean Center dialog box should now be visible. For the "*Input Feature Class"*, specify "Financial_Points." A default "Output Feature Class" will appear in your workspace. You may wish to change the default file name. For the "*Case Field*," select "Type."

Question #2: Discuss the relationship between financial institutions and both the Black population and population in poverty (use the variable: PctPov) using central feature analysis. 5 points

Part 2: Dispersion and Directionality

In this section we are going to start looking at compactness and direction in our data. The equivalent in aspatial analysis would be to look at the shape of the distribution and to look for Skewness. We will accomplish this by examining the Standard Distance and Directional Distributions of our data.



Primary symbology

Unique Values

Field 1 Type

Add field

Color scheme

Classes Scales

More

Symbol	Value	Label
	Type	5 symbol classes
□	Bank	Bank
□	Check Cashing S...	Check Cashing S...
□	Money Transfer S...	Money Transfer S...
□	Pawnbrokers	Pawnbrokers
□	Payday Loans	Payday Loans

Figure 1: Categorizing financial services based on unique colors

1. Standard Distance:

Standard distance is equivalent to the standard deviation in aspatial diagnostics. If all the distances of each point to the mean center are normally distributed, our 1 Standard Distance circle should contain 68% of all businesses of a given type.

1A. In the Geoprocessing window, type "Standard distance" and select the appropriate tool. The Standard distance dialog box should now be visible. For the *Input Feature Class*, specify "Financial_Points," and for the *Case Field*, specify "type." Leave the value of 1 standard deviation. Run the tool.

1B. Based on the output from step 1A, create a unique value map according to the "type" of financial institution: In the table of contents, right click on the layer and select "symbology." Under "Primary Symbology' select 'Unique Values.' Under Field 1 select "type," Hit OK.

You will now have several large multi-colored standard distance circles for each financial institution type. You can edit the symbology for each financial institution type by clicking on the colored boxes. Change the symbology so that each circle only has a colored outline. This will allow you to easily compare the standard distance circles.

The standard distance works the same way as a regular standard deviation. Recall that +/- standard deviation from the mean in a normal distribution encompasses 68% of the observations. There are 199 banks in our sample. If the banks are normally distributed, then there should be approximately 135 (68% of 199) within the standard distance. How many payday loans institutions are within the 1 standard distance for payday loans? If the numbers are far from 68%, this suggests that the distances are not normally distributed. We will learn about GIS tools to check normality for existing variables later in this lab.

2. Directional Distribution

Directional Distribution measures the orientation and spatial spread of a dataset, often represented by an ellipse, to show the overall trend and dispersion direction of geographic features.

2A. In the Geoprocessing window, type "Directional Distribution" and select the appropriate tool. For the *Input Feature Class* specify "Financial_Points," and for the *Case Field*, specify "Type." Run the tool.

2B. Using a similar process to step 1B above, create a unique values map (based on "TYPE"), and for each value (Bank, Check Cashing, etc.) create a unique hollow outline.

Note that, unlike the standard distance, directional distribution returns ellipses rather than circles. (The numbers of observations contained in the ellipses may not be the same as those contained in corresponding circles created by standard distance due to different calculation approaches.) In practice this is a much better fit for our data than what we did in the previous step given the shape of our map layer, but we really learn very little additional information than what we had from the previous effort. When employed on a map layer with a less pronounced oblong form, this technique can actually help to indicate important factors, particularly corridor effects.

Question #3: Interpret your results by pointing out some findings of your dispersion analysis. 5 points

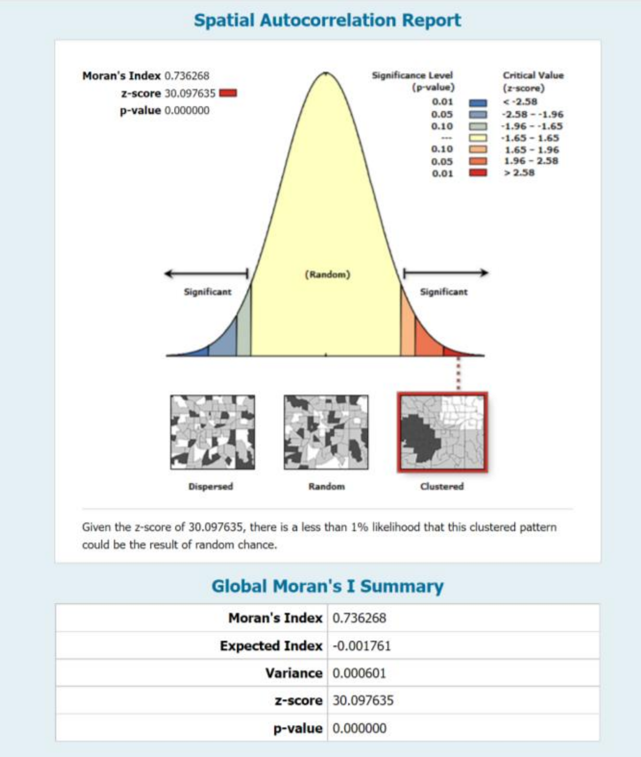
---

Part 3: Global and Local Clustering

In this section we will test out the methods for quantifying the degree of clustering in our data. Specifically, we will look at the degree to which financial institutions and Seattle's black neighborhoods are clustered (independently of one another). In subsequent sections we will decompose these measures to try and relate the clustering patterns to one another.

1. Calculate the Global Moran’s I for the Percentage Black Population

1A. In the Geoprocessing window, type “Spatial Autocorrelation” and select the appropriate tool. The Spatial Autocorrelation (Moran’s I) dialog box should now be visible. For the *Input Feature Class*, specify “Seattle_blockgroups.” For the *Input Field*, specify “PctBlack.” Check the *Generate Report* box.



Spatial Autocorrelation Report

Moran's Index 0.736268
z-score 30.097635
p-value 0.000000

Given the z-score of 30.097635, there is a less than 1% likelihood that this clustered pattern could be the result of random chance.

Global Moran's I Summary

Moran's Index	0.736268
Expected Index	-0.001761
Variance	0.000601
z-score	30.097635
p-value	0.000000

Figure 2: Checking the spatial autocorrelation of the percent of black population using Moran's I

1B. Under *Conceptualization of Spatial Relationship* specify “Contiguity edges corners”. The *Conceptualization of Spatial Relationship* field is where we indicate the neighborhood we are assuming for the purposes of the calculation. Choosing Contiguity edges corners is the same as choosing Queen’s 1st order and is the most common choice in demographic research. Inverse Distance, (the default option) is also a good choice and we will eventually run the analysis both ways.

1C. Under *Standardization*, specify “Row.” Standardization refers to the choice of whether to scale the weights for each neighbor so that they sum to 1 (as we did not define specific weights). If an observation has two neighbors, they will each be given a weight of 50%. If it has three neighbors, then each will be assigned a weight of 33%. In general, we will choose row standardization unless our data is quite uniform in terms of the number of neighbors.

1D. Run the tool by clicking “OK.”

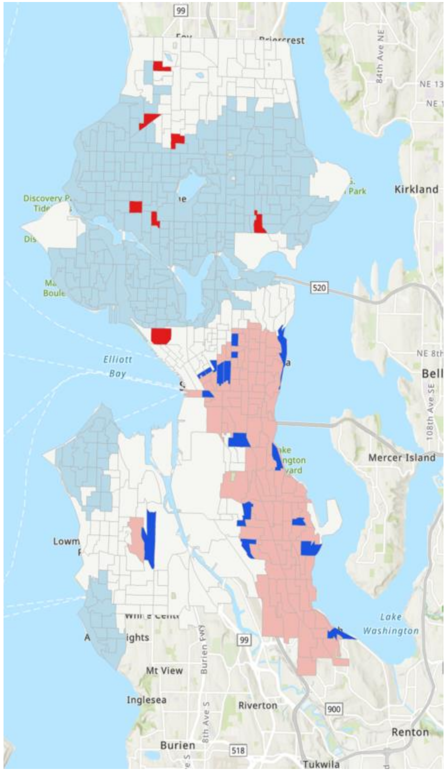
Select “View Details’ at the bottom and Open the Report file by clicking the link in the View Details dialog. The details box will include some of the basic diagnostics.

When we look at the graphical output, it shows that our block groups are highly clustered in terms of the percentage of their populations that is black. The Moran’s Index is 0.73, indicating positive spatial autocorrelation. Given the z score of 30.09, the p-value indicates that there is less than 1% likelihood that this clustered pattern could be the result of random chance (the null hypothesis).

Question #4: Run the Global Moran's I using inverse distance: a) don't specify a threshold distance (use the default), 2) specify a threshold distance of 500 ft. This specifies a cutoff distance: Features outside the distance are ignored in analyses. Compare the Moran's I value for each of these outputs and discuss any differences. 10 points

2. Calculate Local Indicators of Spatial Autocorrelation (LISA)

2A. In the Geoprocessing window, type "Cluster and Outlier Analysis" and select the appropriate tool. The Cluster and Outlier Analysis (Anselin Local Moran's I) dialog box should now be visible. Under *Input Feature Class*, specify "Seattle_blockgroups." Under *Input Field*, specify "PctBlack." Under *Conceptualization of Spatial Relationship* specify "Inverse distance". Under *Standardization*, specify "Row."

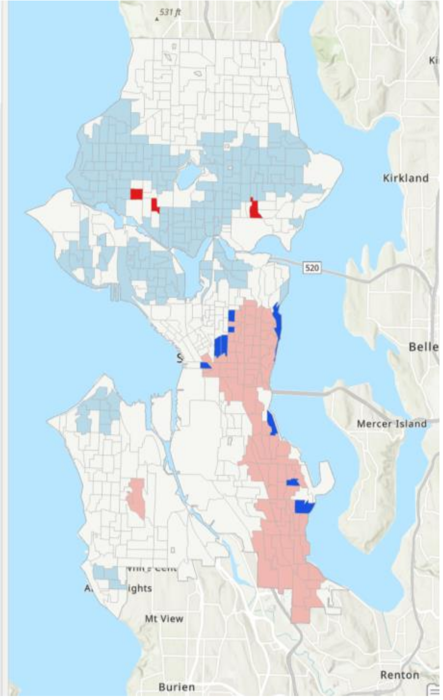


This tool creates a new output with the following attributes for each feature in the input feature class: local Moran's I index, z-score, p-value. The output we get from this operation is a map of the Local Moran's I Z scores classified by Standard Deviation. What does this map tell us? It tells us which block groups are *more similar and dissimilar* to their neighbors than might be expected as the result of a spatially random process. In other words, it tells us the block groups that contributed to the high positive (and low) value of our global Moran's *I.*

You will note the overall pattern (generally High-High in the south, and Low-Low in the north), with several spatial outlier block groups located throughout. The High – High cluster indicates census tracts with relatively high percent black proximate to tracts also with relatively high percent black. Open the attribute table. You will note columns indicating the local Moran's' I value, the relevant Z score and p-value, and the CO Type (Cluster Outlier type) – many of these are blank, indicating they are not significant (the p-value will confirm this), but some are labeled with the appropriate Cluster (HH, LL) or Outlier (LH, HL).

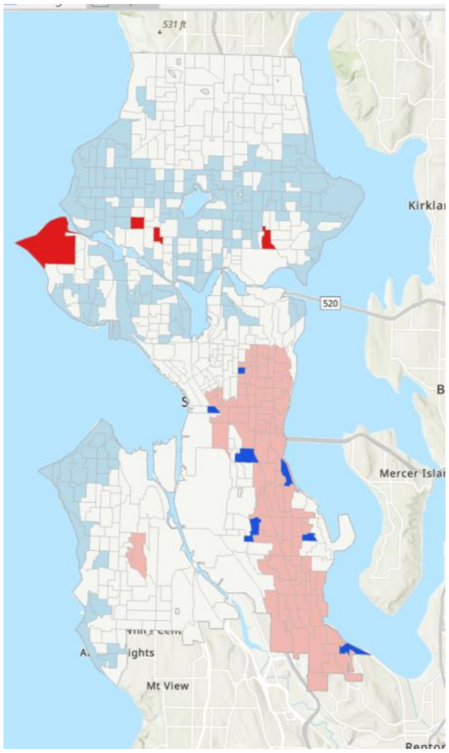
We can also open the Moran's Scatterplot, which indicates the overall Morans I, but so allows us to identify tracts according to where they fall (the x axis is the z score for the Percent Black, and Y-axis is the value of the spatial lag (according to the conceptualization of spatial relationship we identified)

Remember, selecting ‘Inverse distance’ makes every blockgroup a neighbor of every other block group. Is that an accurate reflection of the variable in question? Many American cities are highly segregated, with potentially much more granular spatial patterns.



Let’s run the Moran’s I again, but this time we set a threshold distance to 3000 feet.

Now our cluster are more define, although the overall pattern remains the same. We find a high-high spatial cluster in the south west part of the city and several low-low spatial clusters throughout the northern part of the city (these are the top right and bottom left quadrants on our Moran scatterplot from the lecture). We also note several outliers (high-low and low-highs) – more dissimilar from their neighbors than would be expected from a spatially random process.
We can validate our results. Using ‘Contiguity edges corners’ as our conceptualization of spatial relationships (remember this identifies neighbors based on contiguous borders). We find a similar result:



Question 5. Calculate both the global Moran’s I as well as the Local Indicators of Spatial Autocorrelation using the PctPov (percent of population living in poverty) field. Be sure to include a discussion and justification of the parameters you included. Include a map and discuss your results. 10 points

What if we would like to know whether there is a particular spatial clustering pattern to crime in the City of Seattle?

Open the Seattle_crime attribute table. We need locations and a numeric variable to calculate spatial autocorrelation. Scroll through the variables and note there is information about crimes, but no numerical variable to calculate a Moran’s I. In this situation, how could we create a numeric variable to test spatial autocorrelation?

One option is to create counts using the spatial join tool. Join the seattle_crime (join feature) to the seattle_blockgroup (target) data sets. You know how to do this from previous lab work. The outcome should have a variable titled

‘Join_count’, this is the number of crimes for each block group.

Question 6. What kind of join is needed to create a count (one-to-one or one-to-many)? Calculate both the global Moran’s I as well as the Local Indicators of Spatial Autocorrelation using the Join_Count (number of crimes per blockgroup) field. Include a discussion and justification of the parameters you included. Include a map. Are there similarities between the spatial distributions of poverty and crime? 10 points

Part 4: Deliverables

1. Answer to questions 1 – 6 above. (45 points)

2. Homework:
We will use some of the spatial statistics tools to better understand the pattern of Dengue Fever for Pennathur, a village in Southern India. This village is one of 44 villages that are part of a Dengue Fever study. Dengue Fever is a painful, potentially fatal illness that is spread by a tiny mosquito, and unfortunately it is quite common in Southeast Asia and Central America. It is estimated that as many as 100 million people contract this disease each year...and health agencies are still years away from finding a vaccine.

From the “\homework” folder, add IndCases, Households, and Village boundary. Open their respective attribute table to better understand the dataset.

1. Calculate several measures of Geographic central tendency and dispersion. If appropriate, select a weight field. Include a map and discuss and analyze some of the relationships. (5 points)

2. Calculate both the global Moran’s I as well as the Local Indicators of Spatial Autocorrelation using the hhrate field for the household data set. Be sure to include a discussion and justification of the parameters you included. Include (keep in mind, these are points, not polygons!) a map and discuss your results. (10 points)