

Introduction

The goal of these exercises is to give you a chance to put the concepts we have just discussed into practice. Keep in mind we have only a limited amount of time, so our focus today will be breadth rather than depth! Don't worry, we'll offer more workshops in the future and we are always available to schedule a consultation to work on your own questions in greater detail. Please feel free to **ask questions** as we proceed if something doesn't make sense, and certainly be vocal if find you've missed a step.

Text in **bold** refers to actual commands to be performed in order to complete the tasks of this lab. Text in `callout-boxes` form is meant to explain or develop the actions we are taking and may be most useful if you find yourself returning to these instructions at a later date.

To begin the exercise we first need to set up our computer with the correct data files. CSSCR's computers only allow you access to read and write files from the C:\temp directory, so we will start by copying our files there.

Exercise Setup

All of the materials for the course are available on the CSDE workshop web site at:

<http://csde.washington.edu/services/gis/workshops/ESDA.shtml>

- Navigate to this page and scroll down to the link for "All Workshop Materials (.zip)"
- Click through this link and Save it. A download box should appear. When the file is done downloading you can double-click on it to open the zip archive, showing the folder we need for our workshop, "csde_ESDA".
- Right-click the "csde_ESDA" folder and select copy, then navigate to C:\temp
- If there is already a folder with "CSDE" and "ESDA" in it's name here select and delete it.
- Paste the csde_ESDA folder into C:\temp.

Part 1 – Exploring Data Distributions

The purpose of this lab is to look at spatial data distributions in our data in much the way we would if we were just starting to work with a new data set in an aspatial context. Our data layer gives us political boundaries (census block groups) and has attached attribute data, but we want to get a sense of how those attributes are distributed in space, what outliers, if any, exist, and what kinds of broad patterns might be present. These are the precursors to our deciding whether there are potential errors in our data or whether the data requires some sort of transformation.

Aspatial Data Distributions

1. Open ArcMap on your computer and begin with a blank project
2. Add the layer **Seattle_blockgroups**.

Layers are added in ArcMap using the “Add Data” button



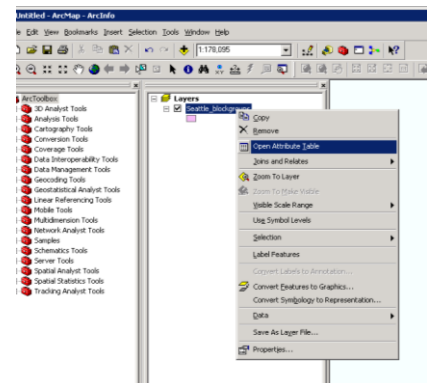
3. Tools->Graphs->Create

- a. Graph type = Histogram
- b. Value field: PctPov
- c. Number of bins: 20

In PctPov we have a continuous variable that ranges from 0 to 63% with a mean of 11%. It is not normally distributed, and has a long right tail. Most researchers will choose to make histograms (or scatterplots) in more specialized software, but we start here to demonstrate that ArcMap has this basic functionality and to make sure you recognize that if you need to do some basic aspatial tests on your data you don't have to switch back and forth between platforms.

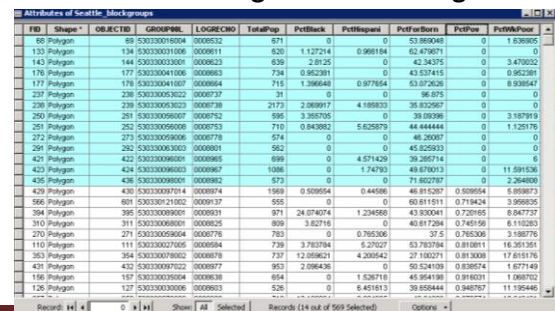
4. Right-click on “Seattle_blockgroups and select “Open Attribute Table”

Our spatial data file includes detailed information on where things are located, but also contains “attribute” data, in this case taken from the 2000 Census, associated with each polygon on our map. When we open the Attribute table we are telling the GIS to show us the census data, but certain actions we take here (like selecting) will show up on our map.



5. Navigate to the Column “PctPov” and Right-click on the header selecting “Sort Ascending”

- a. Hover your mouse pointer over the observation tabs on the far left of the screen. While holding the left mouse button down, drag the mouse down to select all of the “0” records for PctPov



FID	Shape	OBJECTID	GROUPFID	LOGRECNO	TotalPop	PctBlack	PctHispanic	PctForBorns	PctFlow	PctMPOuse
65	Polygon	65	530300016004	0000532	671	0	0	53.89048	0	1.63695
133	Polygon	134	530300033000	0000611	620	1.12724	0.960184	62.47671	0	0
143	Polygon	144	530300033000	0000623	636	2.8125	0	42.34375	0	3.47002
176	Polygon	177	530300041006	0000683	734	0.862881	0	43.537415	0	0.962881
177	Polygon	178	530300041007	0000684	715	1.366448	0.875654	63.272626	0	8.838447
237	Polygon	238	530300053022	0000737	31	0	0	96.875	0	0
238	Polygon	239	530300053023	0000738	2175	2.888917	4.888933	35.62567	0	0
260	Polygon	261	530300060007	0000752	695	3.365705	0	38.80386	0	3.187919
261	Polygon	262	530300060008	0000753	710	0.843882	5.625879	44.484444	0	1.125176
272	Polygon	273	530300060009	0000776	574	0	0	48.26027	0	0
286	Polygon	287	530300063003	0000801	662	0	0	45.625623	0	0
421	Polygon	422	530300060005	0000865	699	0	4.571429	38.26574	0	0
423	Polygon	424	530300060003	0000867	1086	0	1.74793	49.670013	0	11.591536
436	Polygon	438	530300060005	0000882	673	0	0	71.652797	0	2.264888
439	Polygon	430	53030007014	0000974	1569	0.509554	0.44586	48.815297	0.508654	5.888873
566	Polygon	601	530300121002	0003137	555	0	0	60.811511	0.719424	3.998835
568	Polygon	365	530300060007	0000931	971	24.574074	1.234565	43.530041	0.720455	6.847127
310	Polygon	311	530300060007	0000925	809	3.62716	0	40.817294	0.740156	6.115283
270	Polygon	271	530300059004	0000776	783	0	0.785306	37.5	0.785306	3.188776
110	Polygon	111	530300077005	0000584	738	3.783784	5.27527	63.783784	0.818811	16.361261
353	Polygon	354	530300076002	0000879	737	12.089621	4.200542	27.100271	0.813008	17.615176
431	Polygon	432	530300076022	0000977	953	2.896436	0	50.524108	0.838574	1.677149
156	Polygon	157	530300030504	0000638	694	0	1.536716	45.844188	0.914037	1.986702
126	Polygon	127	530300030006	0000603	528	0	6.451813	39.650444	0.948767	11.195448

- b. Repeat the process for the records at the bottom of the table, perhaps those where PctPov exceeds 50%**

By selecting the records this way we can learn more about our data than we could from the histogram alone. We can see some interesting things. The lowest poverty areas on our map are scattered around the city, but include a lot of waterside block groups. Note that one of the areas only has 31 residents (the industrial/maritime area south and east of the UW). Block groups this small tend to create problems in our analysis because a small number of individuals can create big changes in percent values and will likely contain imputed values as well to protect confidentiality. Switching to the high poverty areas we have three evident clusters including one right on the edge of the U district. Knowing where these extreme values are located will help us down the road as we are making sense of the data in context with our other variables.

Basic Maps: Choropleth

This is likely a review for many of you, but the most basic way in which we can understand data distributions in a spatial context is to create the most basic of maps: the Choropleth. Here we will look at our variable of interest based on its divergence from the mean 8%

1. Double-click on the Seattle_blockgroups layer to open the layer properties. Select Symbolology

- a. Select Quantities**
- b. Value = PctBlack**
- c. Select Classify**
 - i. Method = Standard Deviation**

From the histogram above and from a quick glance at the values in the table we know that our PctBlack variable is heavily skewed to the right. As such, the standard deviation map may not be the most appropriate. As an alternative try changing the classification scheme to Quartiles or the Value to LogBlack. Nevertheless, PctBlack is one of the most common demographic variables employed in poverty studies, and so it is instructive to see which observations stand out from the perspective of statistical difference and to have this image in our head as we begin to formulate our hypotheses about relationships in our data.

Geographic and Population Centers

The following lab section looks at the idea of centrality in our data. 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 of our data here 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 quickly begin to generate important information about the distribution of our data.

1. Central Feature

- a. **Spatial Statistics Tools>Measuring Geographic Distributions>Central Feature**
- b. **Input Feature Class = Seattle_blockgroups**

1a. Central Feature

- a. **Repeat the above, but use Weight = TotalPop**

Central Feature finds the single feature that minimizes the total distance traveled from the centroids of all other features in our map layer. The weighted version just counts each polygon k times where k is the population of that block group. 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.

2. Mean Center

- a. **Add Data = Financial_Points**
- b. **Spatial Statistics Tools > Measuring Geographic Distributions >Mean Center**
 - i. **Input Feature Class = Financial_Points**
 - ii. **Case Field = Type**

Mean Center gives nearly identical results to those above, but returns a point instead of a polygon. Mean Center is 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 Central Feature).

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 (normal, Poisson, etc) and to look for Skewness. We will accomplish this by examining the Standard Distance and Directional Distributions of our data.

1. Standard Distance

- a. **Spatial Statistics Tools>Measuring Geographic Distributions>Standard Distance**
 - i. **Input Feature Class =Financial_Points**
 - ii. **Case Field= Type**
- b. **Right click on the newly created layer and navigate to Properties and then Symbolology**
- c. **Under Categories Select Unique Values**
 - i. **Change the Value Field to "TYPE"**

- ii. **Add All Values**
- d. **For each value now listed (Bank, Check Cashing, etc) Click on the colored box and change the colors so that we are showing just an outline with a different color outline for each type**

Standard distance is equivalent to the standard deviation in aspatial diagnostics. Our 1 Standard Distance circle contains 68% of all businesses of a given type. By comparing the radii and location of the different circles we can get a sense of how clustered our different data types are. Note, for example, that money wiring services are heavily concentrated in the south, and that check-cashing operations are the most dispersed of any of the types.

2. Directional Distribution

- a. **Spatial Statistics Tools>Measuring Geographic Distributions>Directional Distribution**
 - i. **Input Feature Class = Financial_Points**
 - ii. **Case Field= Type**
- b. **Right click on the newly created layer and navigate to Properties and then Symbolology**

Note that, like the standard distance, this ellipse contains exactly 68% of our observations. The difference is that this function does not limit the shape to a circle. 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.

Part 2 –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

- a. **Spatial Statistics Tools>Analyzing Patterns>Spatial Autocorrelation (Moran’s I)**
 - i. **Input Feature Class = Seattle_blockgroups**
 - ii. **Input Field = PctBlack**
 - iii. **Display Output Graphically**
 - iv. **Conceptualization = Polygon Contiguity (First Order)**
 - v. **Standardization = Row**

Some elements of our selections here merit further explanation. The “Conceptualization” field is where we indicate the neighborhood we are assuming for the purposes of the calculation. Choosing Polygon Contiguity 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 some students may wish to take

the time to run the analysis both ways (results in a value of 0.69 for I). Selecting the “row standardization” option is also an important choice. Standardization refers to the choice of whether to scale the weights for each neighbor so that they sum to 1 (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.

When we look at the graphical output it shows unequivocally that our block groups are highly clustered in terms of their percentage black. To put our 0.74 value in context, the value we calculated for PctPov in the lecture was only 0.54. In fact, this is the most clustered result I have ever calculated outside of a simulation.

Local Indicators of Spatial Autocorrelation (LISA)

1. Calculate the Local Moran's I for the Percent Black Population

a. Spatial Statistics Tools > Mapping Clusters > Cluster and Outlier Analysis

- i. Input Feature Class = Seattle_blockgroups**
- ii. Input Field = PctBlack**
- iii. Conceptualization of Spatial Relationships = Polygon Contiguity (First Order)**
- iv. Standardization = Row**

The output we get from this operation is a map of the Local Moran's I Z scores classified by Standard Deviation. Unfortunately, this map is probably not what we are looking for or expecting to find. What does this map tell us? It tells us which block groups are more similar to their neighbors than might be expected as the result of a spatially random process (positive values—top right and bottom left quadrants on our Moran scatterplot from the lecture) and which are spatial outliers, that is to say more dissimilar from their neighbors than would be expected from a spatially random process (top left and bottom right from our Moran scatterplot). In other words, it tells us the block groups that contributed to the high positive value of our global Moran's I and those that contributed to the low, but doesn't differentiate between a cluster of high percent black block groups and a cluster of low percent black block groups.

2. Now we want to adjust our LISA output files to show the information we are interested in, not just the contributors, but the high/high vs. low/low values

a. Double-click on the LISA layer > Symbology

- i. Show = Categories, Unique values**
- ii. Value = COType**
- iii. Add All Values**
- iv. Deselect <all other values> (to remove from legend only)**
- v. Alter the colors for each category HH equals red, HL is pink LH is light blue LL is dark blue (this conforms to the default settings for other software packages that do LISA).**

So what happened here? We only have HH clusters? Does that mean that we have high percent black block groups clustered near one another but we don't have any low percent black block groups clustered near to one another? That can't be right....

3. Return to our Seattle block groups layer and make a quantile map showing PctBlack

- a. Double-click on Seattle_blockgroups>Symbology
- b. Quantities > Graduated Color
- c. Value = PctBlack
- d. Classify > Method = Quantile
- e. OK

What becomes obvious from looking at this quantile map is the high percentage of block groups where the percent black is zero or extremely low. The entire bottom 20% has zero and it is only in the top 20% of block groups that we even exceed 12% (percentage of the population as a whole). Since there is a lot of variation in the right tail of our distribution (things go as high as 65%) but not much movement at the bottom, our technique doesn't register similarities at the low end. Similar outcomes are possible in all sorts of data that doesn't conform to normal distributions, and as with many other forms of statistical analysis we can improve our results by transforming our variable of interest.

4. Calculate the Local Moran's I for the LOG of the Percent Black Population

- a. Spatial Statistics Tools > Mapping Clusters > Cluster and Outlier Analysis
 - i. Input Feature Class = Seattle_blockgroups
 - ii. Input Field = LogBlack
 - iii. Conceptualization of Spatial Relationships = Polygon Contiguity (First Order)
 - iv. Standardization = Row
- b. Double-click on the LISA layer > Symbology
 - i. Show = Categories, Unique values
 - ii. Value = COType
 - iii. Add All Values
 - iv. Adjust colors as above

Taking the log of our variable of interest reduces the impact of our outliers at the high end of the distribution and makes the slight variances at the low end seem more meaningful. As a result, our map now shows both HH and LL clusters, and significantly, does not show any spatial outliers. We can now see both the location of our clusters and their significance vis a vis a random spatial process.