# LAB 2: Detecting Clusters with Health Datasets

In this lab, we will use a GISystem named GeoDa to identify spatial clusters of disease and undesirable health outcomes. We will begin by working through the software (as it's different from ArcGIS), work with an example dataset of cancer rates in the southeast US, and then you will conduct an independent analysis on the Toronto health profiles data.

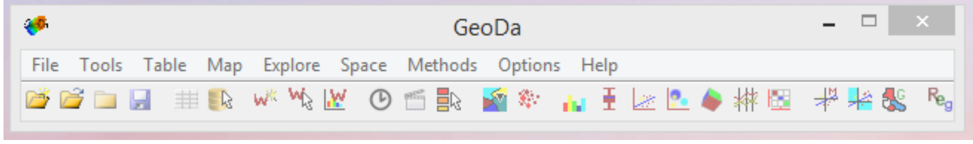
*Software*: GeoDa

*Data:* US Cancer Rates 1970-1994 & Toronto Health Profiles

To Hand In: Three responses to lab questions (with corresponding figures) and your independent work with the Toronto Health Profiles data. **Include all figures, statistics, and maps in your lab write up and turn in all lab report questions and the final independent work in the same document with clearly labeled headers.**

## PART 1: Loading Spatial Data

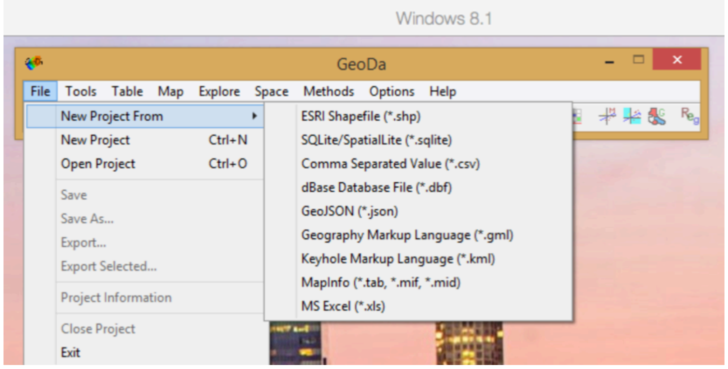
All of GeoDa's user interface is accessed through this menu:



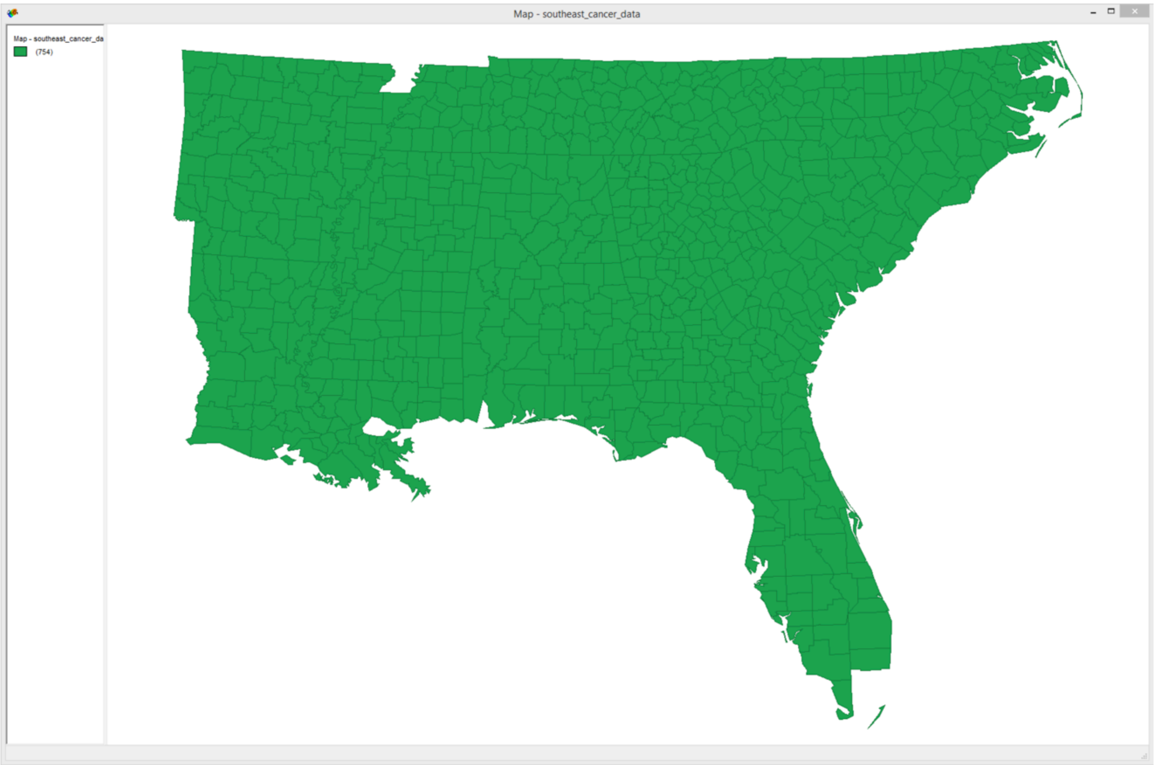
There are a number of "short cut" icons – but all of these functions are also available via the pull down menus that appear after clicking the text labels – e.g. "File," "Tools," etc.

Generally, GeoDa deals with one spatial file at a time. Unlike ArcGIS, this GIS does not specialize in layering spatial files. Instead, it is intended for indepth analyses of one particular layer, usually point or polygon files.

To load a spatial file into GeoDa, click on "File"-> "New Project From" -> <the type of spatial file you want to load>.


Windows 8.1
GeoDa
File Tools Table Map Explore Space Methods Options Help
New Project From
New Project Ctrl+N
Open Project Ctrl+O
Save
Save As...
Export...
Export Selected...
Project Information
Close Project
Exit
ESRI Shapefile (*.shp)
SQLite/SpatialLite (*.sqlite)
Comma Separated Value (*.csv)
dBase Database File (*.dbf)
GeoJSON (*.json)
Geography Markup Language (*.gml)
Keyhole Markup Language (*.kml)
MapInfo (*.tab, *.mif, *.mid)
MS Excel (*.xls)

Let’s start a new project from an ESRI Shapefile. Navigate to where you downloaded the ”southeast_cancer_data” shapefile and double click. A basic map will pop up that looks like this:



The data are now loaded. Like ArcGIS, we have a map *and* an attribute table. To find the attribute table, either click on the table icon in the menu: ▦. The attribute table will pop up:

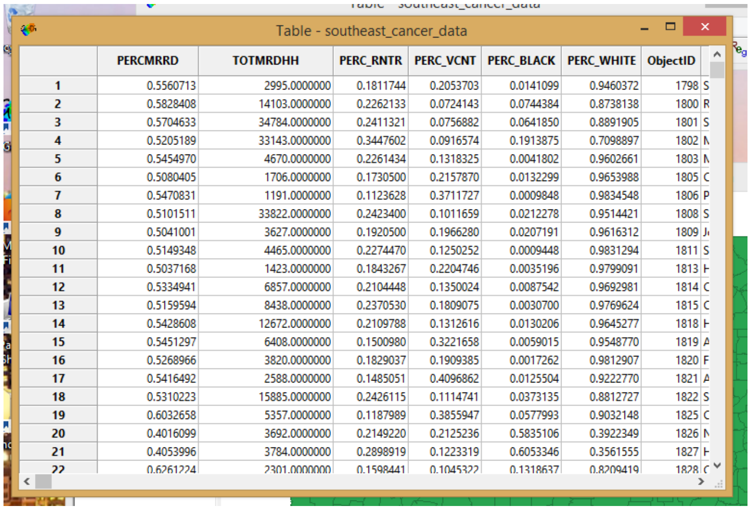
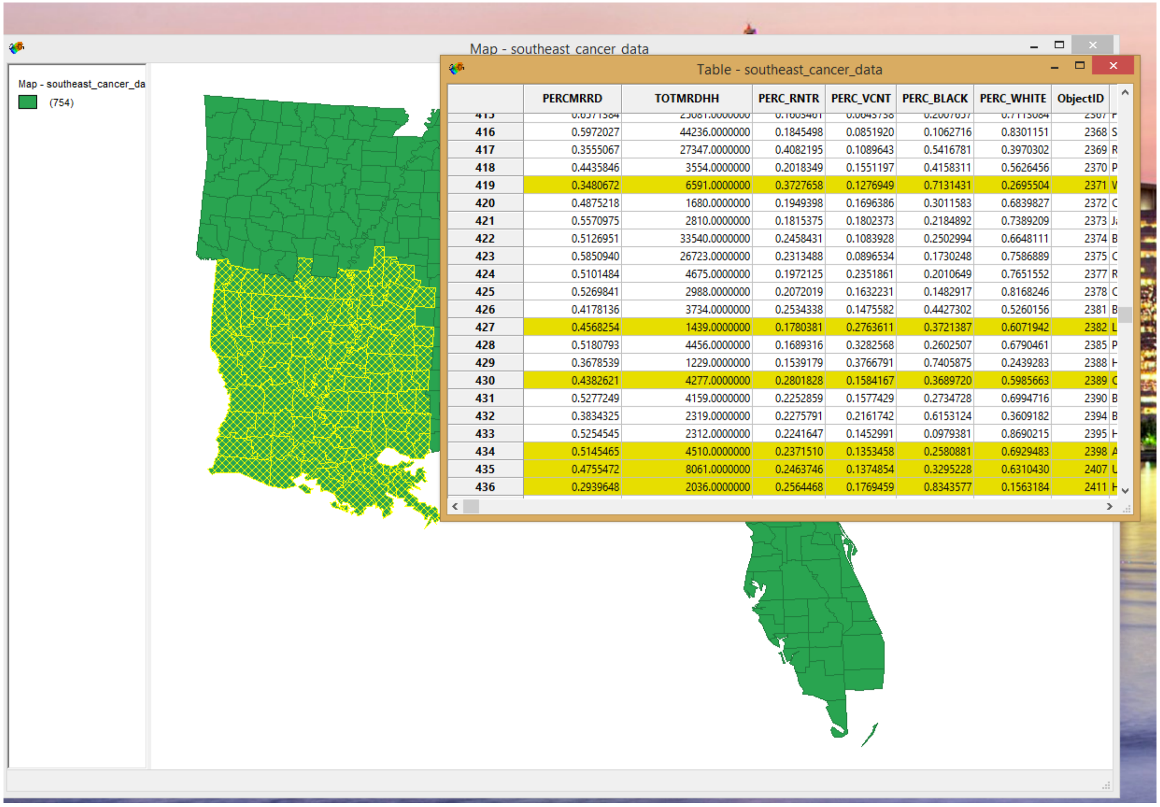


Table - southeast_cancer_data

| | PERCMRRD | TOTMRDHH | PERC_RNTR | PERC_VCNT | PERC_BLACK | PERC_WHITE | ObjectID | |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.5560713 | 2995.0000000 | 0.1811744 | 0.2053703 | 0.0141099 | 0.9460372 | 1798 | S |
| 2 | 0.5828408 | 14103.0000000 | 0.2262133 | 0.0724143 | 0.0744384 | 0.8738138 | 1800 | R |
| 3 | 0.5704633 | 34784.0000000 | 0.2411321 | 0.0756882 | 0.0641850 | 0.8891905 | 1801 | S |
| 4 | 0.5205189 | 33143.0000000 | 0.3447602 | 0.0916574 | 0.1913875 | 0.7098897 | 1802 | N |
| 5 | 0.5454970 | 4670.0000000 | 0.2261434 | 0.1318325 | 0.0041802 | 0.9602661 | 1803 | N |
| 6 | 0.5080405 | 1706.0000000 | 0.1730500 | 0.2157870 | 0.0132299 | 0.9653988 | 1805 | C |
| 7 | 0.5470831 | 1191.0000000 | 0.1123628 | 0.3711727 | 0.0009848 | 0.9834548 | 1806 | P |
| 8 | 0.5101511 | 33822.0000000 | 0.2423400 | 0.1011659 | 0.0212278 | 0.9514421 | 1808 | S |
| 9 | 0.5041001 | 3627.0000000 | 0.1920500 | 0.1966280 | 0.0207191 | 0.9616312 | 1809 | J |
| 10 | 0.5149348 | 4465.0000000 | 0.2274470 | 0.1250252 | 0.0009448 | 0.9831294 | 1811 | S |
| 11 | 0.5037168 | 1423.0000000 | 0.1843267 | 0.2204746 | 0.0035196 | 0.9799091 | 1813 | H |
| 12 | 0.5334941 | 6857.0000000 | 0.2104448 | 0.1350024 | 0.0087542 | 0.9692981 | 1814 | C |
| 13 | 0.5159594 | 8438.0000000 | 0.2370530 | 0.1809075 | 0.0030700 | 0.9769624 | 1815 | C |
| 14 | 0.5428608 | 12672.0000000 | 0.2109788 | 0.1312616 | 0.0130206 | 0.9645277 | 1818 | H |
| 15 | 0.5451297 | 6408.0000000 | 0.1500980 | 0.3221658 | 0.0059015 | 0.9548770 | 1819 | A |
| 16 | 0.5268966 | 3820.0000000 | 0.1829037 | 0.1909385 | 0.0017262 | 0.9812907 | 1820 | F |
| 17 | 0.5416492 | 2588.0000000 | 0.1485051 | 0.4096862 | 0.0125504 | 0.9222770 | 1821 | A |
| 18 | 0.5310223 | 15885.0000000 | 0.2426115 | 0.1114741 | 0.0373135 | 0.8812727 | 1822 | S |
| 19 | 0.6032658 | 5357.0000000 | 0.1187989 | 0.3855947 | 0.0577993 | 0.9032148 | 1825 | C |
| 20 | 0.4016099 | 3692.0000000 | 0.2149220 | 0.2125236 | 0.5835106 | 0.3922349 | 1826 | N |
| 21 | 0.4053996 | 3784.0000000 | 0.2898919 | 0.1223319 | 0.6053346 | 0.3561555 | 1827 | H |
| 22 | 0.6261224 | 2301.0000000 | 0.1598441 | 0.1045322 | 0.1318637 | 0.8209419 | 1828 | C |

If you click your mouse across the row index (the numbers on the left hand side of the table) the row you click will become highlighted. Notice, that the corresponding spatial area will also light up on your map. Similarly, if you drag your mouse across the map and select a range of spatial areas, they will become highlighted in your table.


Map - southeast_cancer_data
Map - southeast_cancer_da
🟩 (754)
Table - southeast_cancer_data

| | PERCMRRD | TOTMRDHH | PERC_RNTR | PERC_VCNT | PERC_BLACK | PERC_WHITE | ObjectID | |
|---|---|---|---|---|---|---|---|---|
| 415 | 0.6371384 | 23081.0000000 | 0.1605461 | 0.0043738 | 0.2007657 | 0.7113084 | 2367 | F |
| 416 | 0.5972027 | 44236.0000000 | 0.1845498 | 0.0851920 | 0.1062716 | 0.8301151 | 2368 | S |
| 417 | 0.3555067 | 27347.0000000 | 0.4082195 | 0.1089643 | 0.5416781 | 0.3970302 | 2369 | R |
| 418 | 0.4435846 | 3554.0000000 | 0.2018349 | 0.1551197 | 0.4158311 | 0.5626456 | 2370 | P |
| 419 | 0.3480672 | 6591.0000000 | 0.3727658 | 0.1276949 | 0.7131431 | 0.2695504 | 2371 | V |
| 420 | 0.4875218 | 1680.0000000 | 0.1949398 | 0.1696386 | 0.3011583 | 0.6839827 | 2372 | C |
| 421 | 0.5570975 | 2810.0000000 | 0.1815375 | 0.1802373 | 0.2184892 | 0.7389209 | 2373 | J |
| 422 | 0.5126951 | 33540.0000000 | 0.2458431 | 0.1083928 | 0.2502994 | 0.6648111 | 2374 | B |
| 423 | 0.5850940 | 26723.0000000 | 0.2313488 | 0.0896534 | 0.1730248 | 0.7586889 | 2375 | C |
| 424 | 0.5101484 | 4675.0000000 | 0.1972125 | 0.2351861 | 0.2010649 | 0.7651552 | 2377 | R |
| 425 | 0.5269841 | 2988.0000000 | 0.2072019 | 0.1632231 | 0.1482917 | 0.8168246 | 2378 | C |
| 426 | 0.4178136 | 3734.0000000 | 0.2534338 | 0.1475582 | 0.4427302 | 0.5260156 | 2381 | B |
| 427 | 0.4568254 | 1439.0000000 | 0.1780381 | 0.2763611 | 0.3721387 | 0.6071942 | 2382 | L |
| 428 | 0.5180793 | 4456.0000000 | 0.1689316 | 0.3282568 | 0.2602507 | 0.6790461 | 2385 | P |
| 429 | 0.3678539 | 1229.0000000 | 0.1539179 | 0.3766791 | 0.7405875 | 0.2439283 | 2388 | H |
| 430 | 0.4382621 | 4277.0000000 | 0.2801828 | 0.1584167 | 0.3689720 | 0.5985663 | 2389 | C |
| 431 | 0.5277249 | 4159.0000000 | 0.2252859 | 0.1577429 | 0.2734728 | 0.6994716 | 2390 | B |
| 432 | 0.3834325 | 2319.0000000 | 0.2275791 | 0.2161742 | 0.6153124 | 0.3609182 | 2394 | B |
| 433 | 0.5254545 | 2312.0000000 | 0.2241647 | 0.1452991 | 0.0979381 | 0.8690215 | 2395 | H |
| 434 | 0.5145465 | 4510.0000000 | 0.2371510 | 0.1353458 | 0.2580881 | 0.6929483 | 2398 | A |
| 435 | 0.4755472 | 8061.0000000 | 0.2463746 | 0.1374854 | 0.3295228 | 0.6310430 | 2407 | U |
| 436 | 0.2939648 | 2036.0000000 | 0.2564468 | 0.1769459 | 0.8343577 | 0.1563184 | 2411 | H |

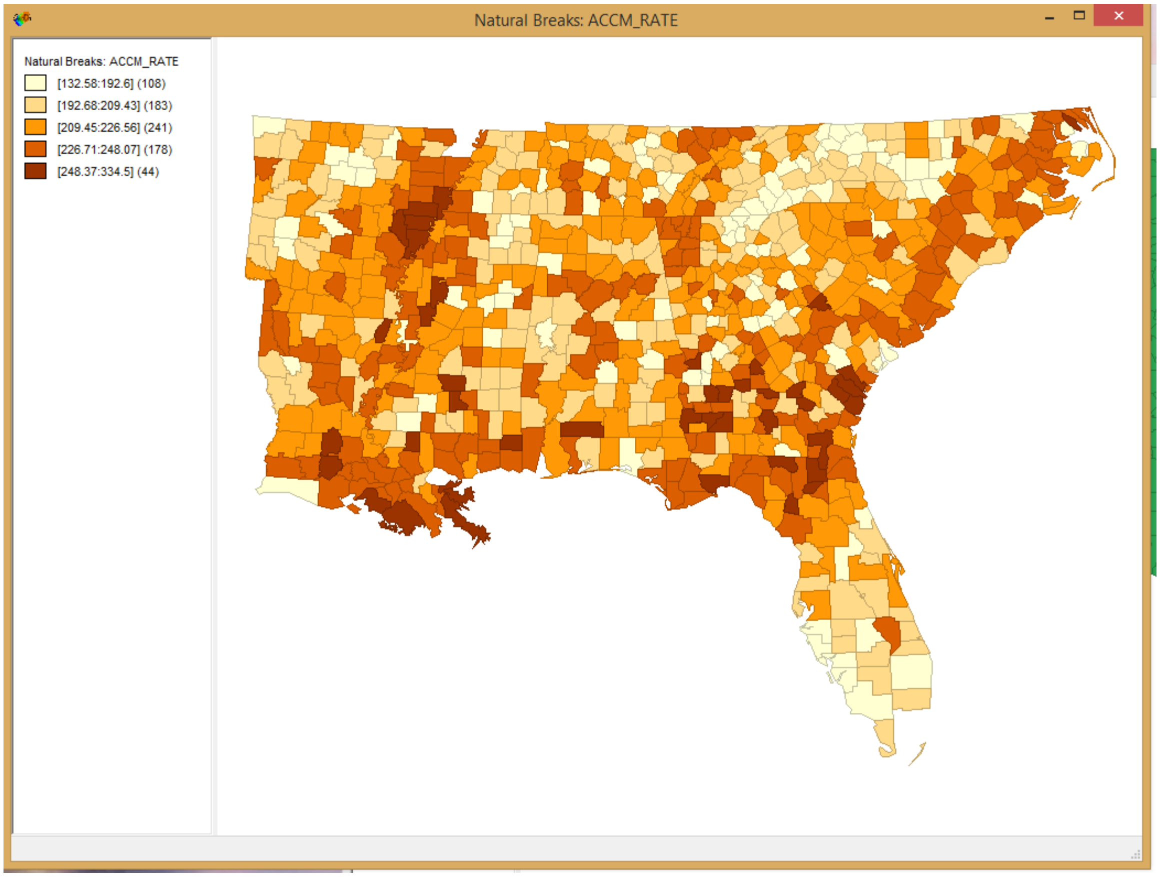
Right clicking on the map or table will provide access to a menu that allows for zooming/panning or selection mode (for the map), or access to a selection tool (for the table). The latter allows you select rows (related to spatial objects) based on their attributes. The equivalent in ArcGIS is "select by attribute."

## Developing Simple Maps

In GeoDa it is very easy to construct choropleth maps. Simply click the "Map" menu and then select the type of map you are interested in developing. Note, the default map that appears after loading spatial data is known as the "Themeless" map.

To develop a map where data are categorized using "natural breaks" (like we learned about in class), click on "Map"->"Natural Breaks Map"->choose the number of categories you would like to have. Let's try 5. Next you'll see a menu pop up asking you to choose a variable to map. Let's choose "ACCM_RATE," which is the (age-adjusted) rate of all types of cancer for males per 100,000 people across the county-level geography in our map.

Once you click "OK" a new window will pop up, showing a map of the 5 categories. On the left side of the window you will see the legend, which automatically displays the category groupings as well as the number of spatial objects within each category.



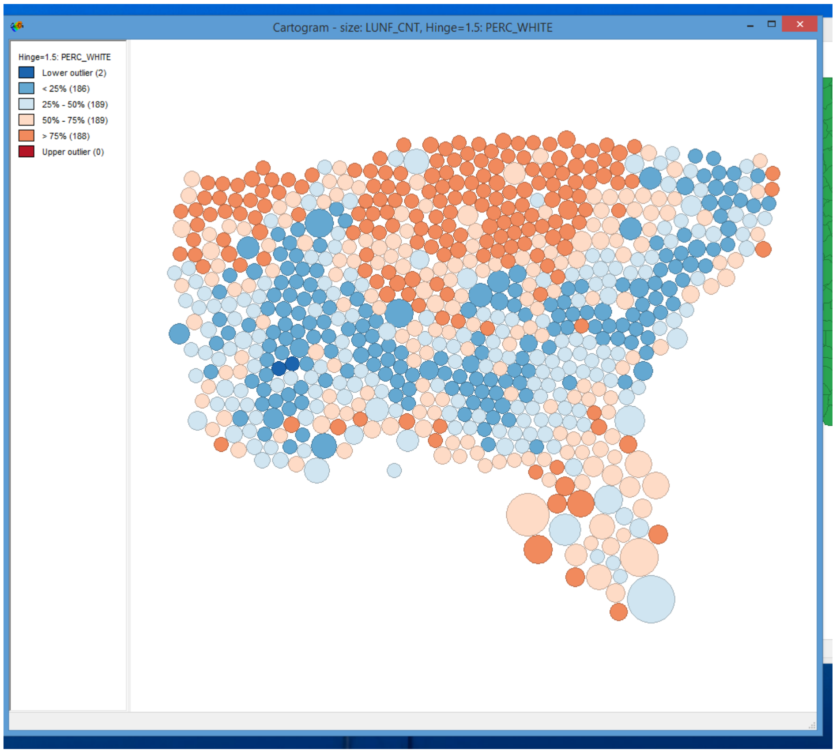
Unfortunately, GeoDa is not very good about allowing you to change color schemes. You can manually change colors for each category by right clicking on the legend color and assigning a new color – but I recommend simply accepting the default. Right clicking on the map, you can change a few features (e.g. remove polygon outlines), but again the visualization of spatial objects is not very complex.

QUESTION 1 FOR LAB REPORT (1 paragraph): Read through the included "cancerp020.txt" document. This is a well done metadata file that describes the data you are working with. Note the naming conventions (for example, see the key with different cancer types and the gender flag – control-F "XXXY" to jump directly to this point). Pick a different cancer type for males and produce a map of the rate of that cancer across the southeast US. Describe what cancer you chose, the name of the variable you mapped, and how the spatial distribution of the rate differs from the distribution of ACCM_RATE. Export the map you produced, and include it in your report beneath your paragraph. The export a map choose "Options" from the top menu, then "Save Image As" to produce a file, or "Copy Image to Clipboard" to copy-paste the image directly into your report.

One nice feature in GeoDa is the cartogram mapper – which allows you to create cartograms with relative ease. The software converts the polygons into circles (in this case counties will be turned into circles) and then arranges them in a way that emphasizes the magnitude of some variable of interest. Try making a cartogram by clicking on "Map" -> "Cartogram." A menu will pop up allowing you to select two variables. One of the variables will be represented by the size of the circle and the second variable will be represented by color.

For circle size, let's use "LUNF_RATE" (the age adjusted lung cancer rate among females) and for circle color let's choose "PERC_WHITE" (percent of the population that identifies as white).

Notice the map is not very remarkable. Because we used a rate to determine circle size, all of the circles are about the same size! Let's try again, but this time use "LUNF_CNT" for the circle size variable. Here, we will see that many of the urban areas (e.g. Memphis, New Orleans, etc.) have larger circles, mostly due to the larger populations of these areas.



QUESTION 2 FOR LAB REPORT (2-4 sentences): Why is it more appropriate to use counts (instead of rates) for sizing the geographic areas in a cartogram?

## PART 2: Establishing Spatial Relationships

The mapping features of GeoDa are useful, if not a bit limited, but the analysis features of GeoDa are very powerful. However, before we can use any of these analysis functions, we first have to establish spatial relationships between the polygons in our spatial dataset.

Recall from lecture how spatial weights matrices are constructed. This can be a time consuming task to complete by hand – computers are much better at it!

In GeoDa we can build a range of spatial matrices by clicking on "Tools" -> "Weights" -> "Create". A menu will appear that allows us to create a spatial weights matrix using a binary 1/0 weight determined by contiguity (e.g. rook/queen) or a distance based weight that is a function of how far away two spatial objects are from each other.

First, select the "ObjectID" as the "Weights File ID Variable." This ensures that the weights matrix we generate can be joined back to the spatial data we're using.

Next, let's select "Queen contiguity" with an order of contiguity of 1. This just means each polygon will assign a weight of "1" to any other polygon that it touches.

To finish, click "Create" and save this weights matrix file as "queen_weights_cancer_data" in the same folder as the cancer data shapefile. Notice it will be a "*.gal" file – this is a common weights format that can be used in different spatial analysis software like R.

*(Navigate to the newly created .gal file, and open it with a text editor. Notice how the file works by - after the initial row that shows a place holder 0, the number of polygons, the name of the shp file, and the name of the attribute acting as an ID - identifying the ID of a polygon and then stating the number of neighbours it has. Then, the next row lists the IDs of the neighbours.)*

Once you get a message saying the creating of the file was a success, go back to "Tools"->"Weights"->"Connectivity Histogram" to inspect the spatial relationships we just constructed. Notice, most counties in our dataset have 5-7 neighbours. If you click on one of the histogram bars, you will see the counties that have that number of neighbours on your map.

Try constructing a k-nearest neighbors weights matrix by selecting "k-Nearest Neighbors" in the weights file creation menu, selecting 10 nearest neighbors. Save this file as "knbr_weights_cancer_data" in the same folder as the cancer data shapefile. This file saves as a "*.gwt" format, which can also be used by spatial analysis software like R. (Open up the .gwt file in a text editor and notice how it differs from the .gal file.)

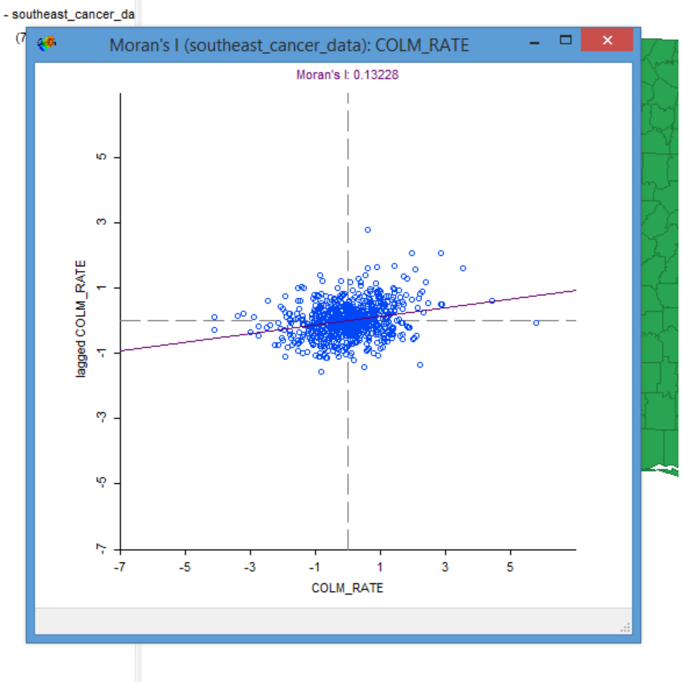
Look at the connectivity histogram for this file. Is it different? Yes! All counties have 10 neighbors, by design.
If you want to switch back and forth between the two different weights matrices, simply click "Tools"->"Weights"->"Select" and then navigate to the weights matrix you're interested in using. **Let's stick with the queen weights matrix for now.**

## PART 3: Global Clusters

Now that we have our spatial weights matrix constructed, we can perform various spatial analyses. Let's calculate the Moran's I value for colon cancer among males across the South Eastern U.S.

To do this, click on "Space" and then "Univariate Moran's I." From the variable list select "COLM_RATE."

A graph will pop up that looks like this:



The Moran's I value appears in the top of the graph (~0.13). This score indicates mild spatial autocorrelation.

The chart shows the relationship between each county's z-statistic (for the variable COLM_RATE) and the spatially lagged z-statistic value (usually just the average of z-statistics for our county of interest's neighbors).

The slope of this relationship happens to be the Moran's I score!

From this graph you can use your mouse to select points on the graph, and they will highlight the corresponding spatial objects in your map.

## PART 4: Local Clusters

The Moran's I statistic is useful in letting us know if spatial autocorrelation exists … but it does not tell us *where* it exists. For this we need to use a local clustering statistic.

We will use the Univariate Local Moran's I statistic (also known as LISA, which stands for local indicators of spatial autocorrelation). Click on "Space" and then "Univariate Local Moran's I" and select "COLM_RATE" again.

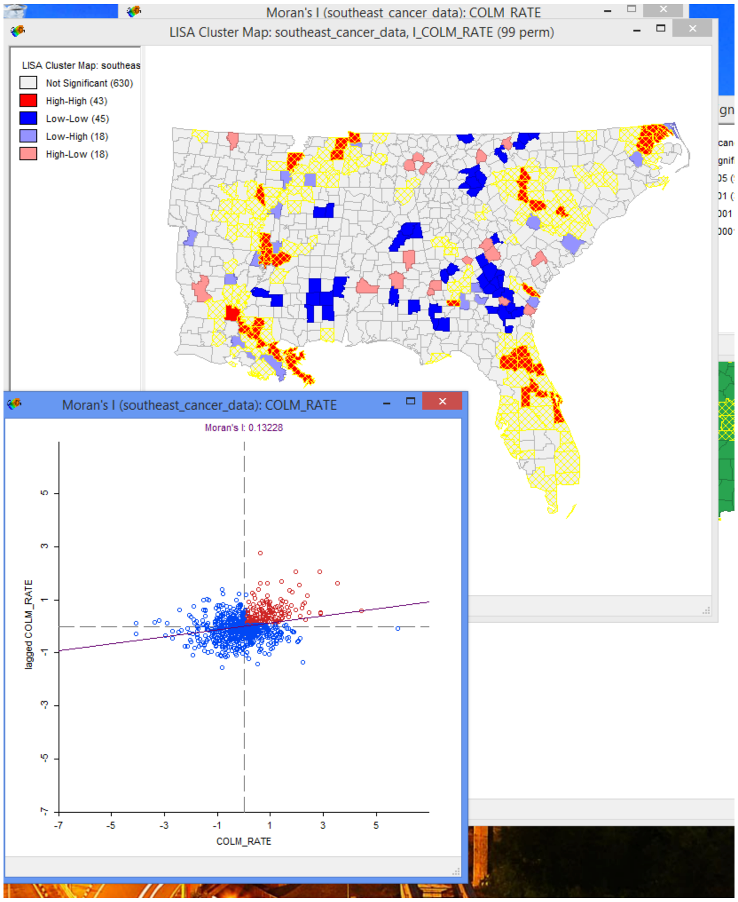
GeoDa will ask if you want to see a significance map, cluster map, and moran scatter plot.

- The significance map shows where there is a statistically significant spatial relationship.
- The cluster map colors the spatial objects (e.g. polygons) to show if they belong to one of four different categories.
    - High-High – a polygon with a significantly high value surrounded by neighbors that also have significantly high values.
    - Low-Low – a polygon with a significantly low value surrounded by neighbors that also have significantly low values.
    - High-Low – a polygon with a significantly high value surrounded by neighbors that have significantly low values.
    - Low-High – a polygon with a significantly low value surrounded by neighbors that have significantly high values.
- The Moran Scatter Plot is the same graph that appears when you compute the Univariate Moran's I statistic.

Check all of these options and then press OK.

The significance map shows you how significant local clusters are using p-values (the lower the value, the more we would expect this cluster to exist in reality).

The Moran's I plot is the exact same as what we saw in the previous section. However, if you select all of the data points in one quadrants (say the data above 0,0 on both axes), you will see that each quadrant corresponds to different LISA categorization (e.g. High-High, Low-Low, etc.). Not all of these data points are significant, so not all of them will have a color in the cluster map.


Moran's I (southeast_cancer_data): COLM_RATE
LISA Cluster Map: southeast_cancer_data, I_COLM_RATE (99 perm)
LISA Cluster Map: southeas
Not Significant (630)
High-High (43)
Low-Low (45)
Low-High (18)
High-Low (18)
Moran's I (southeast_cancer_data): COLM_RATE
Moran's I: 0.13228
lagged COLM_RATE
COLM_RATE
-7
-5
-3
-1
1
3
5

Inspect the cluster map to identify where the hot spots are (High-High clusters) and cold spots are (Low-Low clusters).

QUESTION 3 FOR LAB REPORT (2-4 sentences): Generate a new choropleth map using a categorization of your choosing and note whether or not the locations of these clusters correspond well to the simpler choropleth visualization. Include the map in your report, along with an explanation of why we should use both maps to communicate the locations of places that have a high rate of colon cancer.

## PART 5: Miscellaneous Data Visualization

In addition to this very short tutorial, GeoDa has a wide range of other visualization tools. Try playing around with "Explore" menu to generate histograms, 3D scatter plots, etc.

Additionally, I recommend repeating the previous exercises with different weights matrix schemes to see how cluster detection computations change with varying weights matrices.

Ultimately, GeoDa is a very powerful GIS that specializes in spatial statistics. We have only scratched the surface on what the software can do. If you would like to learn more, I recommend checking out the 244 page tutorial available here: https://geodacenter.asu.edu/og_tutorials

**INDEPENDENT WORK (3-4 paragraphs):** Pick a disease from the Toronto Neighbourhood adult health file (High Blood Pressure, Asthma, Diabetes, Mental Health, or COPD).

Calculate a new column by clicking on Table -> Variable Calculation. Click on "Add Variable" and name this new attribute the disease you chose, plus "rate" (e.g. total_dbts_rate). Accept the defaults and click "add." Select the "Rates" tab at the top of the menu, and select the count variable corresponding to your chosen disease for "Event Variable" and the relevant population attribute for the "Base Variable." This will leave you with a rate attribute that you can use in cluster calculations.

You are tasked with writing a brief summary about the health outcome of your choosing, backed up with maps and clustering statistics, to Toronto Public Health. Use GeoDa to present choropleth maps, cartograms, Moran's I statistics, LISA maps, etc. to explain if and where clustering for your health variable occurs in Toronto. Do some research to understand if the clusters relate to relevant socioeconomic variables, and present a hypothesis on potential causal factors.

*A reminder: Include all figures, statistics, and maps in your lab write up and turn in all lab report questions and the final independent work in the same document with clearly labeled headers.*