



## GEOG 3105

### Laboratory 3: The Bivariate and Empirical Bayesian rates local Moran's I, k means partitioning

At the beginning of the lab please make sure to have these fields always filled out (for this lab documentation find an email I sent). In this lab you will be using 3 distinct datasets, so record them accordingly:

**Analysis author:** *You.*

**Title:** *Lab IV: As above*

**Dataset source:** *Where you got the data.*

**Dataset location:** *For this first analysis, the location would be “Nepal”.*

**Date(s) of analysis:** *When you worked on it.*

**Dataset time span:** *What dates do the data span?*

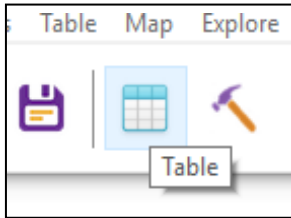
**Dataset Scale:** *For this first analysis, the scale would be “Districts in Nepal”.*

Please note: Location and Scale are not the same in this lab. Please consider your answer carefully. *There is a table at the end of the lab summarizing the deliverables. Anything in red is a deliverable.*

1.) In this lab we will continue where we finished in class, with a further exploration of GeoDa with three approaches, and three data sets:

- The Bivariate Local Moran's I:** International Aid to Nepal, consisting of some various demographic datasets, broken up by districts.
- Local Moran's with EB rates:** Various off-the-wall statistics on crime, suicide, literacy and other “moral statistics” in 1830s France.
- K means partitioning:**

2.) In your S: drive, load your Nepal shapefile into GeoDa. Under the table button,



, you will see variables listed, keyed below:

Contains development-related data for 75 districts in Nepal. Documentation for original data: AidData (1997-2014 with most projects from 2007-14) and [Open Nepal](#)

- Observations = 75
- Variables = 26
- Years = 1991-2013

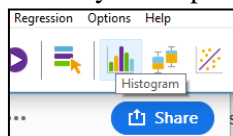
Variable name	Description	Source
DepEcProv	Deprivation in economic provisioning	<a href="#">Open Nepal</a>
PovIndex	Human Poverty Index	<a href="#">Open Nepal</a>
PCInc	Per Capita Income	<a href="#">Open Nepal</a>
PCIncPPP	Per Capita Income PPP	<a href="#">Open Nepal</a>
PCIncMP	Per capita income, Rs. at market price	<a href="#">Open Nepal</a>
MalKids	Percentage of children under age five who are malnourished	<a href="#">Open Nepal</a>
LI40	Percentage of People not expected to survive age 40	<a href="#">Open Nepal</a>
NoSafH20	Percentage without safe water	<a href="#">Open Nepal</a>
Population	Population	<a href="#">Open Nepal</a>
BoyG1_5	Number of Boys Enrolled in Grade 1-5 (2012-2013)	<a href="#">Open Nepal</a>
GrIG1_5	Number of Girls Enrolled in Grade 1-5 (2012-2013)	<a href="#">Open Nepal</a>
KIDS1_5	Number of Children Enrolled in Grade 1-5 (2012-2013)	<a href="#">Open Nepal</a>
SchoolCnt	Number of Schools (2012-2013)	<a href="#">Open Nepal</a>
SCHLPKD	Number of Schools per child (in thousands) (2012-2013)	<a href="#">Open Nepal</a>
SCHLPPOP	Number of Schools per population (in thousands) (2012-2013)	<a href="#">Open Nepal</a>
AD_ILIT	Adult literacy rate (2011)	<a href="#">Open Nepal</a>
AD_ILGT50	Dummy variable with value of 1 if adult literacy rate >50% (2011)	<a href="#">Open Nepal</a>
VotHum	Number of Voters (lunar years 2047-2063, approx. 1991 to 2006)	<a href="#">Open Nepal</a>
TotIEFMS	"Total economy including financial intermediation service indirectly measured (total value added)"	<a href="#">Open Nepal</a>
XXCAmt	Project Sector Committed Amount: XX = sector (see above)	Project Data from AidData Aggregated to District
XXDAmt	Project Sector Distributed Amount: XX = sector: Agriculture, Business Banking, Communication, Conflict Resolution, Budget Support + Finance, Education, Energy, Environment, Forestry, Gov + Civil Society, Health, Humanitarian Aid, Industry, Multi-Sector, Social Infrastructure, Tourism, Transport + Storage, Water Sanitation, Total	Project Data from AidData Aggregated to District

### 3.) Exploring your data

- I would like to continue to ask you to become comfortable with using GeoDa, in which case we will take a look at exploring your data, then performing an analysis on it.

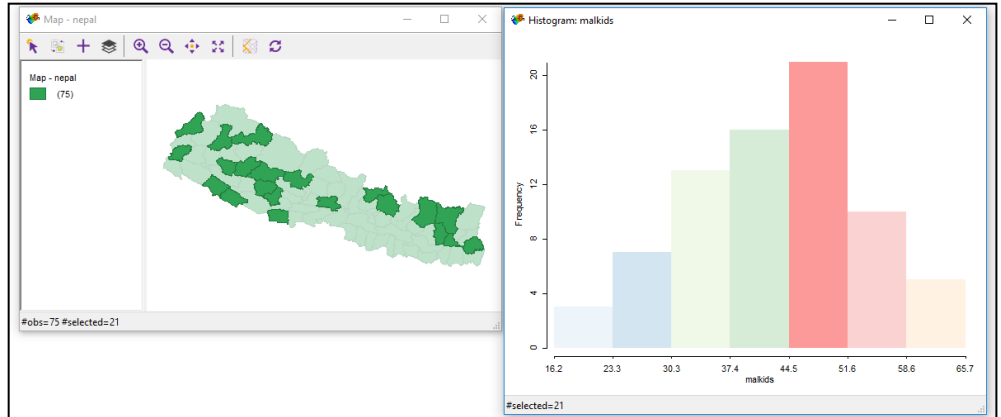
#### i. Histograms of your Nepal data.

- Within your Nepal file, click the Histogram button on your ribbon,

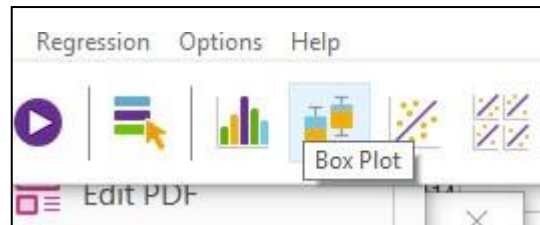


, and select **MalKids**, the Percentage of kids within

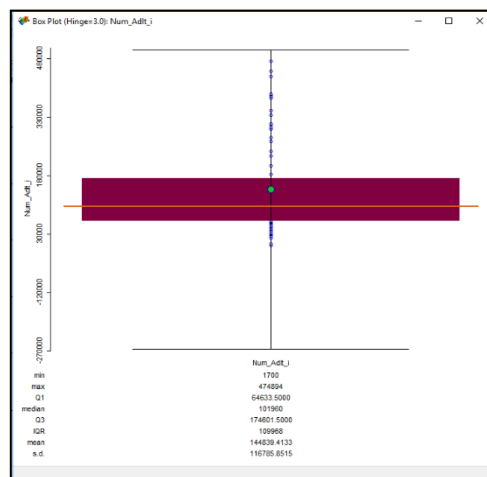
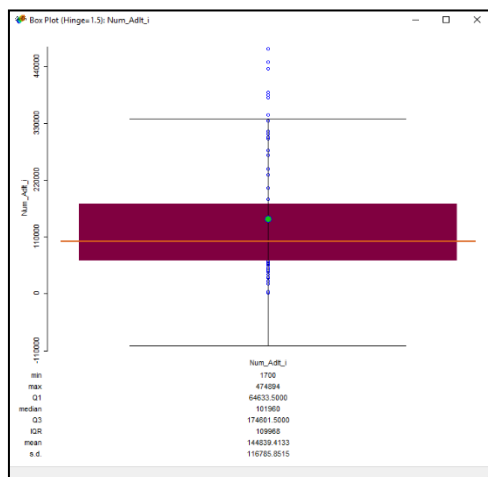
Nepal who are malnourished. You should see a rough normal distribution. **Use the interactive linkage function and click around your histogram. Please briefly explain what a histogram is, and provide a screenshot of this, including the spatial file. See below for what I mean:**



## ii. Boxplots of your Nepal data.

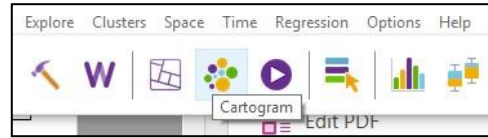


1. Select the boxplot button, and select *Num\_Adlt\_i*, or number of adults who are illiterate by Nepalese district. **Briefly explain what a boxplot is, screenshot your first boxplot, and then right-click, scroll to hinge, and change this to 3.0. Provide a screenshot of this boxplot as well, with an explanation of what the hinge is (this is in your in-class exercise on pages 10 – 12).**

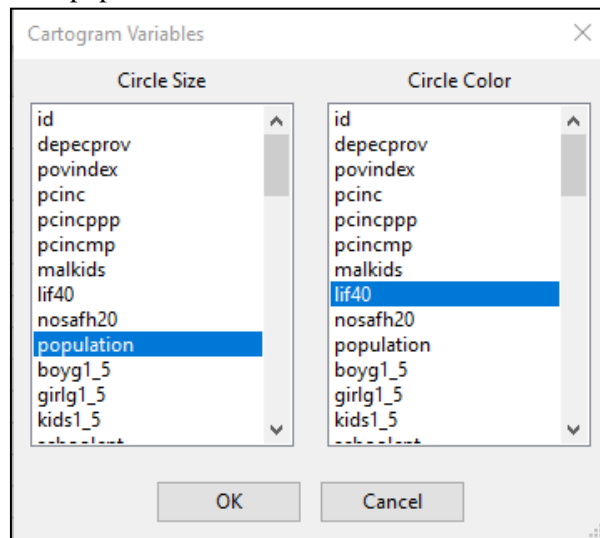


### iii. Cartogram of some Nepal data.

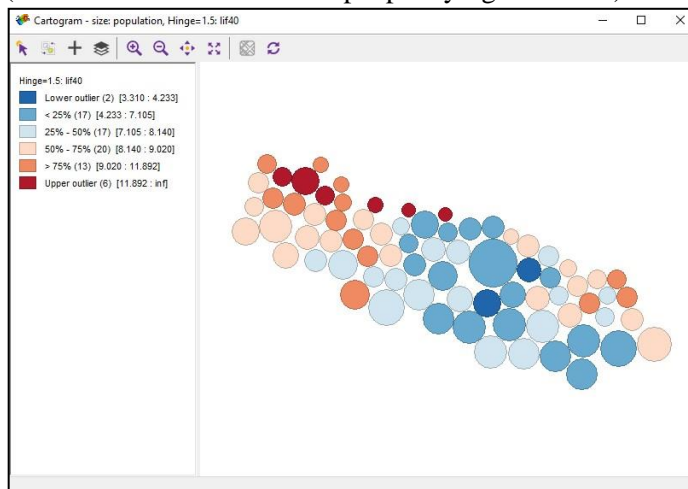
1. As explained in the in-class exercise, cartogram is a map type where the original layout of the areal unit is replaced by a geometric form (usually a circle, rectangle, or hexagon) that is proportional to the value of the variable for the location. We'd like to see, with Nepal, a cartogram of population sizes, and people not expected to live past 40 years old.



2. Click your cartogram button, And then select population for Circle Size, and lif40 for Circle Color:



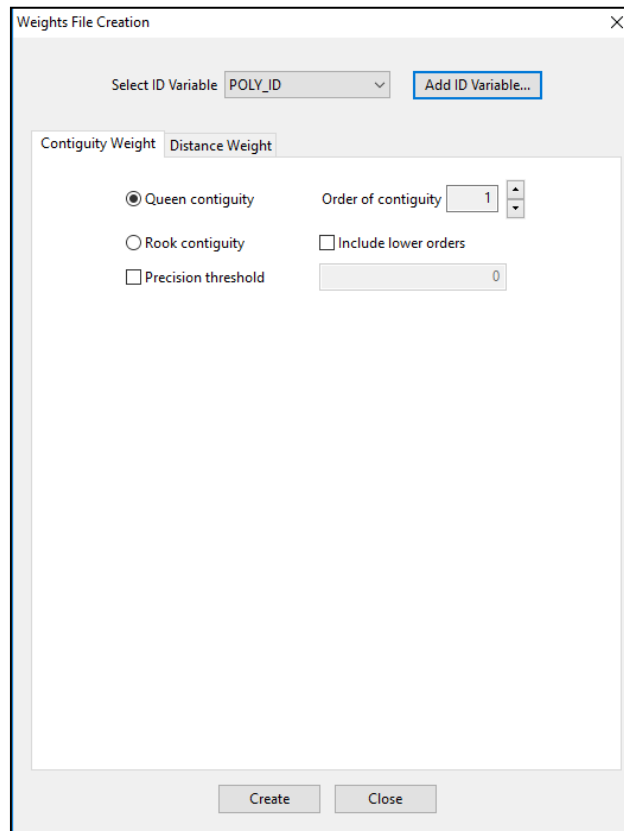
3. Your output should show contextually that higher areas of population (larger circles), show lower numbers of people expected to die before 40 (red is increased numbers of people dying before 40):



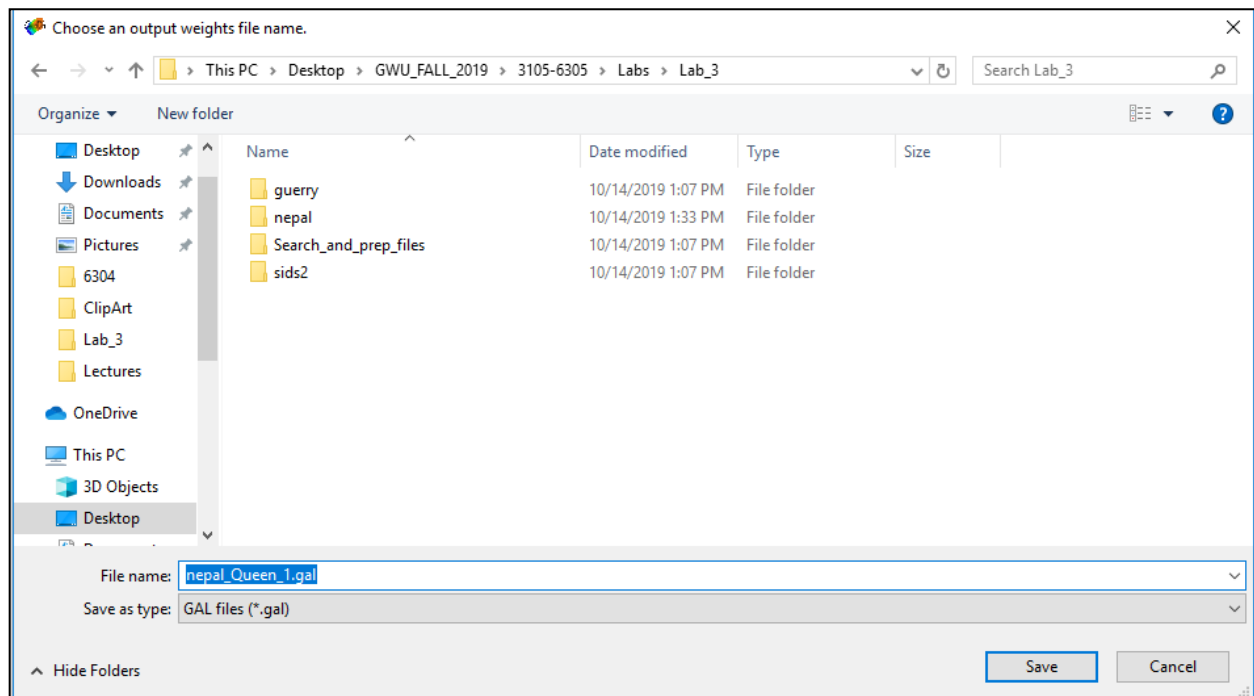
4. Provide a screenshot of your cartogram as above, as well as a short explanation of what it is showing.

#### 4.) The Bivariate Local Moran's I

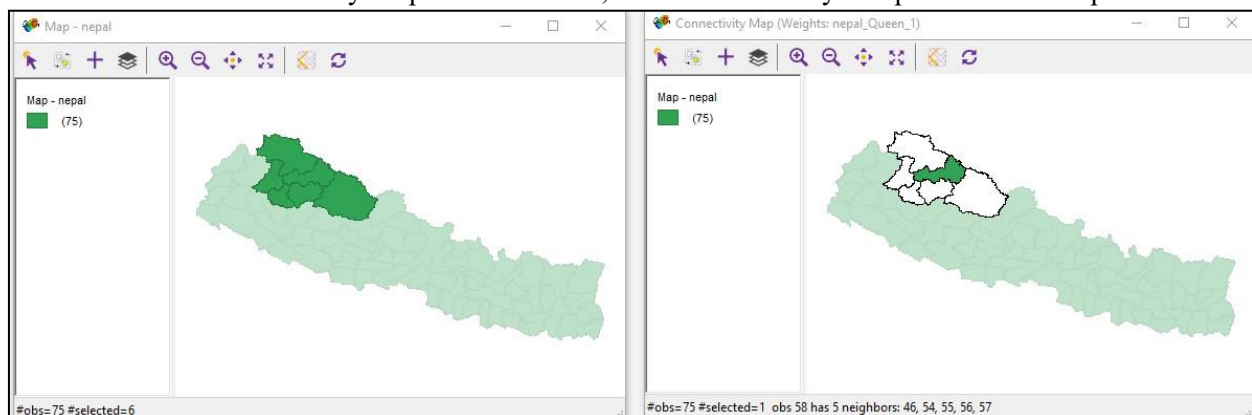
- In this section we will analyze whether there are similar clustering patterns for Number of Children Enrolled in Grade 1-5 (2012-2013) (Variable **KIDS1\_5**), and the number of adults who are illiterate (**Num\_Adlt\_i**). We will use a spatial relationship of Queen's case, first order, as usual.
- With your Nepal file still open, on the ribbon find **Tools → Weights Manager**. Click Create, and the Add Variable ID. Adding the variable ID ensures GeoDa can identify each individual district in Nepal, unambiguously. Poly\_ID is fine, and then you should see your weights manager turned on (*i.e.* it stops being grayed out):



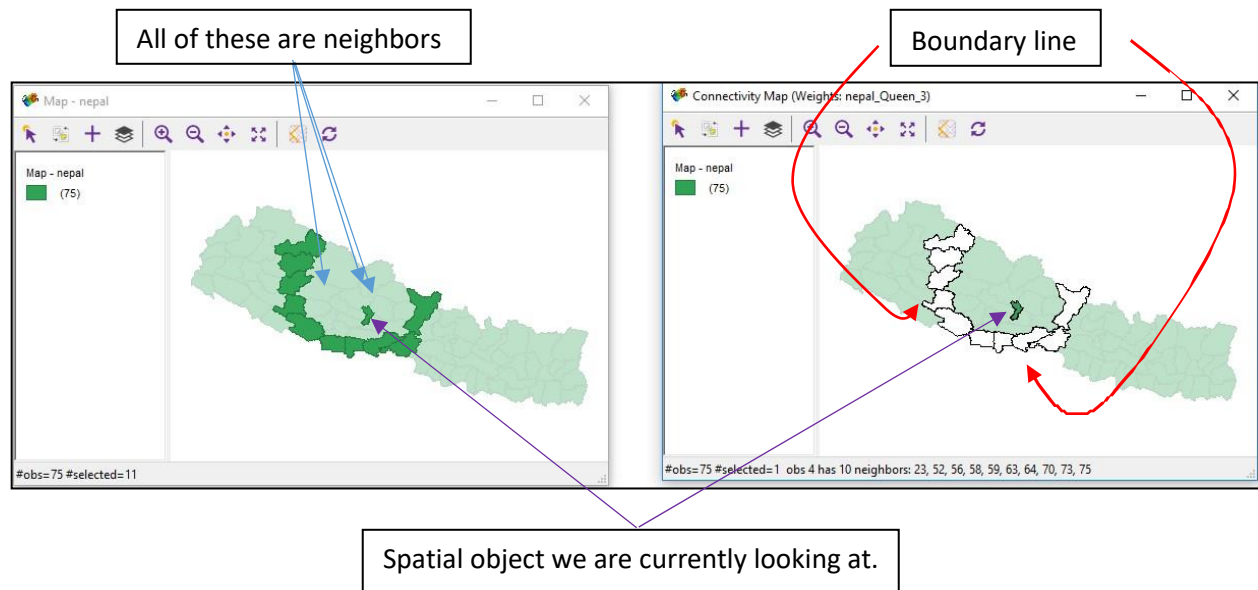
- Confirm **Contiguity Weight** is selected, as well as **Queen contiguity**, Order: 1. You will be prompted to save it. Name it **nepal\_Queen\_1.gal**. →



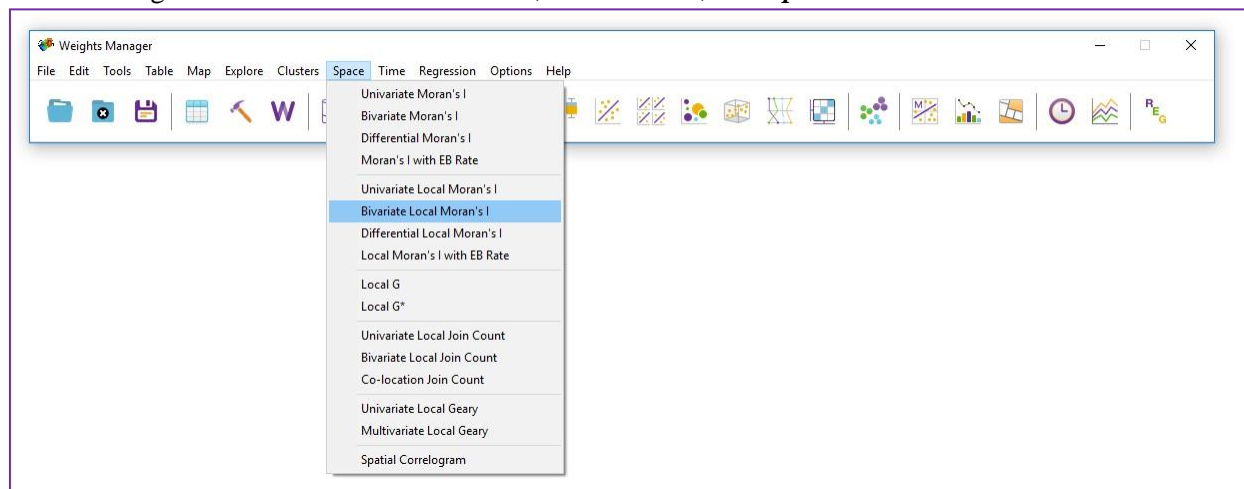
- d. Once created, you will see in the **Weights Manager** dialogue a button titled “Connectivity Map”. Click on this, and observe how your spatial relationship behaves:



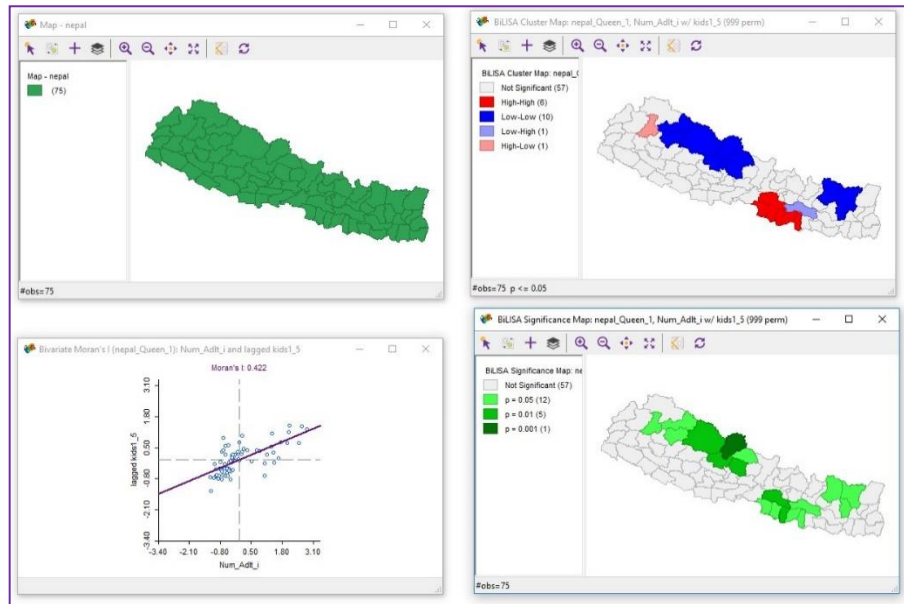
- e. **Please turn in a screenshot of your connectivity map as above.** As a self-exercise (*i.e.* I won't ask you to turn this in), you may want to try other, more complex spatial relationships and use the connectivity map to see how they play out. Below, observe what a Queen's Case 3<sup>rd</sup> order would look like (the outer highlighted line you see is the boundary line). In this case, we are asking GeoDa to show us the Neighbor, of the Neighbor, of the Neighbor (closest neighbor plus two more) of the object of interest:



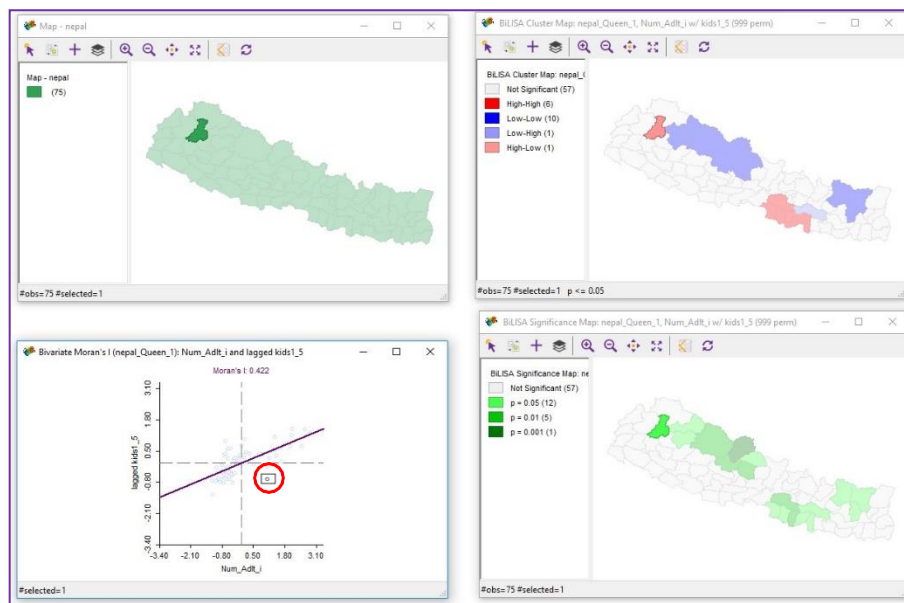
- f. Once you are done with your Weights matrix file, let's do a Bivariate Moran's I.
- g. As with the in-class exercise, in the ribbon, find *Space* → **Bivariate Local Moran's I**:



- h. A very important point to note here. In your variable settings, you will see listed a First Variable (X), and a Second Variable (Y). When your map is finished and you are looking at it, you may forget this and ask yourself, "Wait which variable is the neighbor and which is the center of interest?". **Here, you decide that. Your X is the variable of interest, and your Y is the values of the variable in the neighboring spatial features.**
- i. We want to look at whether Illiteracy in adults (*Num\_Adlit\_i*) has anything to do with patterns of children enrolled in school (*KIDS1\_5*). So our first (X) variable will be *Num\_Adlit\_i*, and our second (Y) variable will be *KIDS1\_5*.
- j. Please make sure to click on all three options: **Significance Map, Cluster Map, and Moran Scatter Plot**. Click **Ok**. See below for a check on your work:



- k. Pay close attention to the Moran's scatter Plot. I will reiterate the interpretation of these results here:

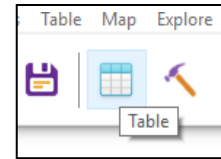


The district highlighted is pale pink (under the red circle in the scatterplot): ***This district shows a HIGH value of numbers of illiterate adults (x axis shows something like 1.5), and a LOW value of kids enrolled in school in neighboring locations (Y axis shows something like -0.70).***

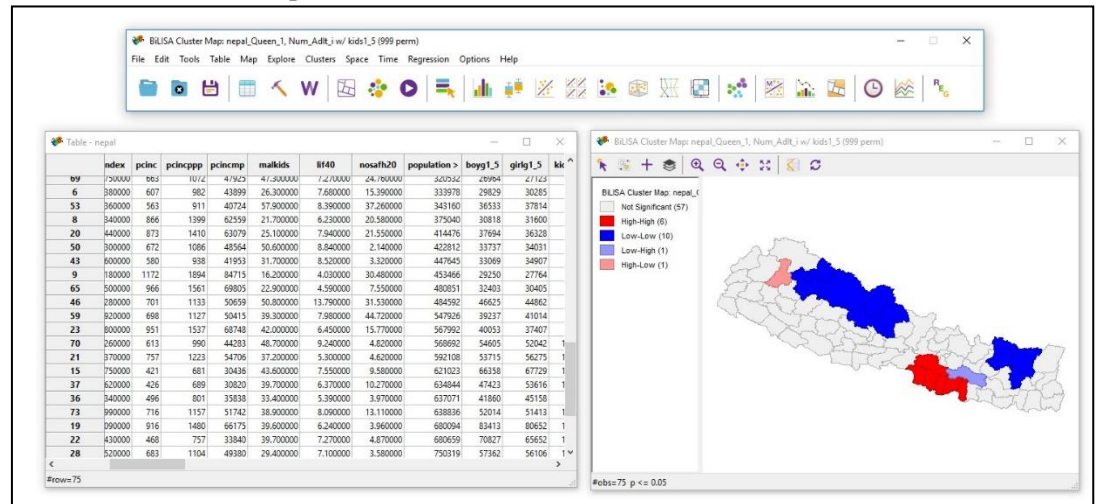
***This is a HIGH-LOW OUTLIER.***



1. The low-low cluster you may find interesting (the blue area in the upper leftish section of Nepal). Areas with LOW illiteracy and LOW enrollment? What's going on there? This may be partially answered by an examination of the shapefile table.

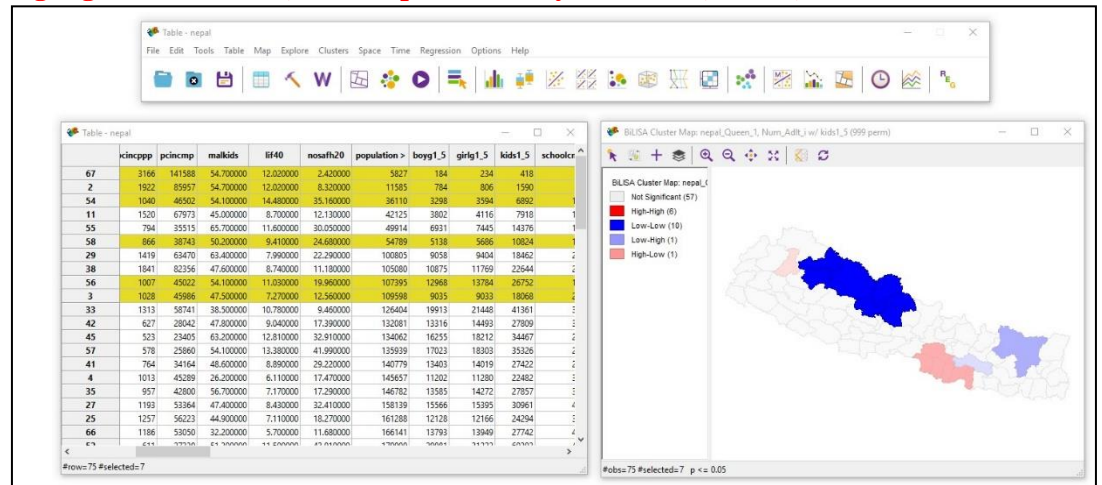


- i. While you have your Cluster Map up, click your table button,
- ii. You should have both open next to each other:



- iii. Holding **Shift**, Slowly click your blue districts. You will see them highlighted in your table as well. Once you have all of them highlighted, take a look at your table. Scroll your table to where you can see the column Population. If you double-click your population column until it appears like so: **population >** it will be sorted smallest to largest. What you should see as well is your highlighted districts suddenly all cluster at the bottom of the numbers (well, top of the table, but the numbers in Population are increasing as you go down the list). What we are seeing most likely here is that since these districts have low populations relative to all of Nepal, both numbers of school children enrolled in school, and numbers of illiterate adults are low. When we say LOW, we mean low compared to all of Nepal. There just are not a lot of people there.
- iv. **To turn in: An explanation of the bivariate local Moran's I; A HIGHLIGHTED example county with screenshot that helps explain spatial lag (see lecture VI); a screenshot of your Cluster Map, along with your**

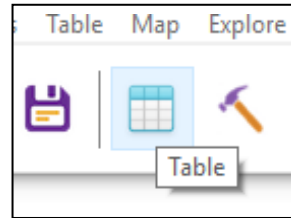
highlighted table as so, an interpretation of your results, and v., below:



v. Also, please answer this question: if we have somewhat misleading results with this analysis of absolute population numbers, what would be a better way to do this?

## 6.) The Local Moran's with EB Rates

- a. Once you have finished with your differential analysis, go ahead and save your project, and close GeoDa. After this, open GeoDa again, and navigate as before to your S drive, and your Lab 3 folder → Guerry, finding and loading Guerry.shp.



- b. Once loaded, examine your table as before, , keyed below:

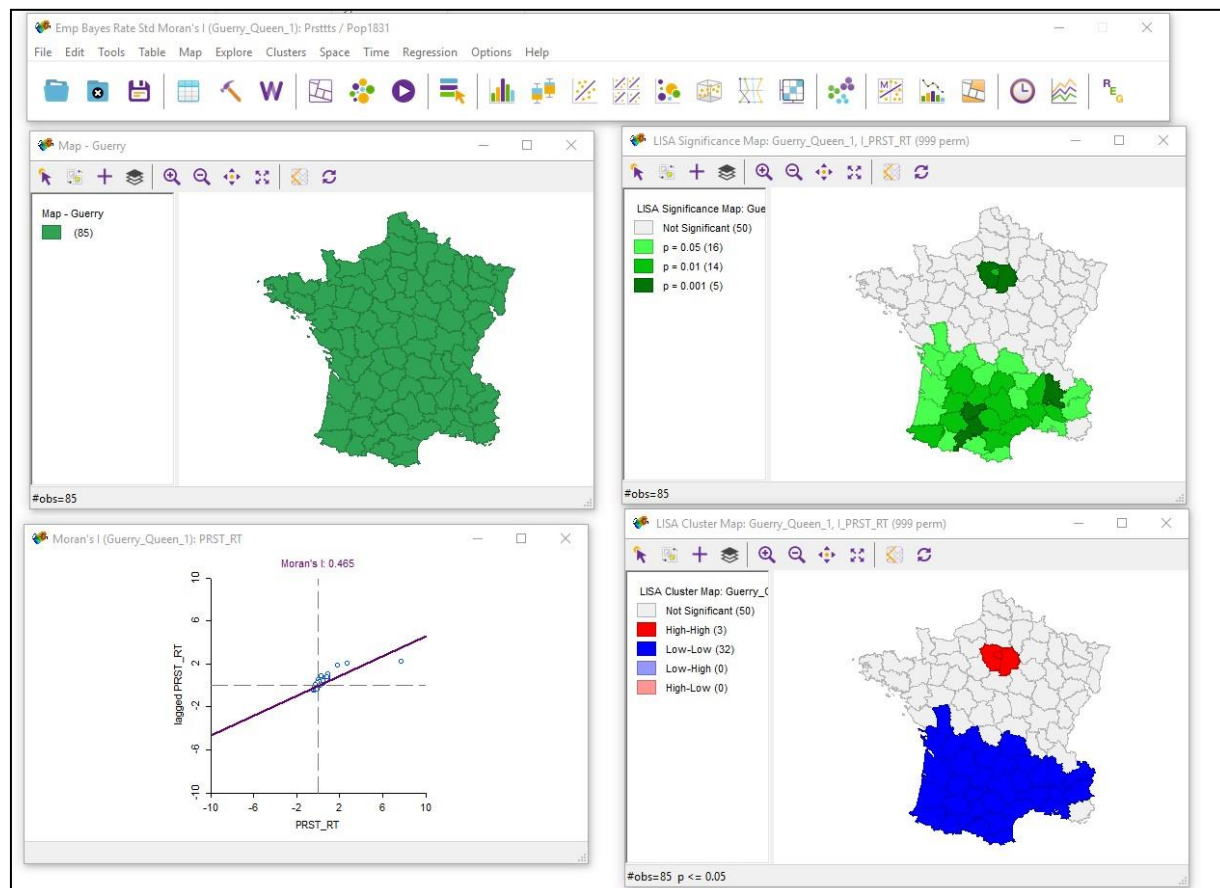
Classic social science foundational study by Andre-Michel Guerry on crime, suicide, literacy and other "moral statistics" in 1830s France. Data from the [R package Guerry](#) (Michael Friendly and Stephane Dray). See below for detailed sources.

- Observations = 85
- Variables = 23
- Years = 1815-1834

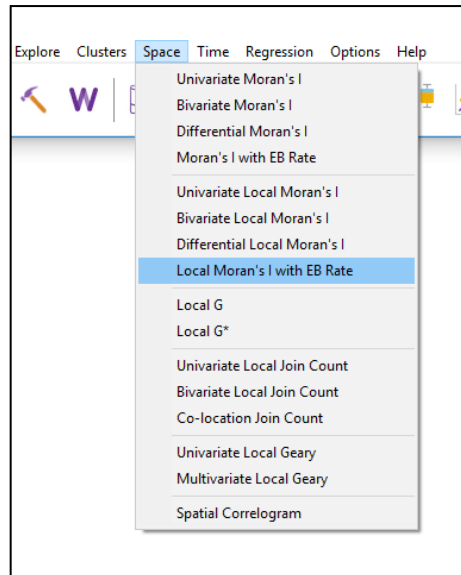
Variable	Description	Source
Variable	Description	Source
dept_code_de	Department ID: Standard numbers for the departments	
region	Region of France ('N'=North, 'S'=South, 'E'=East, 'W'=West, 'C'=Central). Corsica is coded as NA.	
deptnm	Department name: Departments are named according to usage in 1830, but without accents. A factor with levels Ain Aisne Aller ... Vosges Yonne	
crm_prs	Population per Crime against persons.	A2. Compte général, 1825-1830
crm_prp	Population per Crime against property.	Compte général, 1825-1830
litercy	Percent of military conscripts who can read and write.	A2
donatns	Donations to the poor.	A2. Bulletin des lois
infants	Population per illegitimate birth.	A2. Bureau des Longitudes, 1817-1821
suicids	Population per suicide.	A2. Compte général, 1827-1830
maincity	Size of principal city ('1.Sm', '2.Med', '3.Lg'), used as a surrogate for population density. Large refers to the top 10, small to the bottom 10; all the rest are classed Medium.	A1. An ordered factor with levels: 1.Sm < 2.Med < 3.Lg
wealth	Per capita tax on personal property. A ranked index based on taxes on personal and movable property per inhabitant.	A1
commerc	Commerce and Industry, measured by the rank of the number of patents / population.	A1
clergy	Distribution of clergy, measured by the rank of the number of Catholic priests in active service population.	A1. Almanach officiel du clergé, 1829
crm_prm	Crimes against parents, measured by the rank of the ratio of crimes against parents to all crimes – Average for the years 1825-1830.	A1. Compte général
infntcd	Infanticides per capita. A ranked ratio of number of infanticides to population – Average for the years 1825-1830.	A1. Compte général
drtn_cl	Donations to the clergy. A ranked ratio of the number of bequests and donations inter vivos to population – Average for the years 1815-1824.	A1. Bull. des lois, ordonn. d'autorisation
lottery	Per capita wager on Royal Lottery. Ranked ratio of the proceeds bet on the royal lottery to population – Average for the years 1822-1826.	A1. Compte rendu par le ministre des finances
deserth	Military desertion, ratio of number of young soldiers accused of desertion to the force of the military contingent, minus the deficit produced by the insufficiency of available billets – Average of the years 1825-1827.	A1. Compte du ministère du guerre, 1829 état V
instrct	Instruction. Ranks recorded from Guerry's map of Instruction. Note: this is inversely related to Literacy	
Prsttts	Number of prostitutes registered in Paris from 1816 to 1834, classified by the department of their birth	Parent-Duchatelet (1836), De la prostitution en Paris
distanc	Distance to Paris (km). Distance of each department centroid to the centroid of the Seine (Paris)	Calculated from department centroids
area	Area (1000 km <sup>2</sup> ).	Angeville (1836)
pop1831	Population in 1831, in 1000s	Taken from Angeville (1836), Essai sur la Statistique de la Population français

- a. We will be examining a comparison of pre-calculated rates of donations to the poor/population in 1831 France (*DNTNS\_RT*) vs. allowing the Moran's I with Empirical Bayesian Rates to do the calculation; and a comparison of pre-calculated rates of prostitutes/population (*PRST\_RT*) around France (1816 – 1834) vs. again allowing the Moran's with Empirical Bayesian rates to do the calculation. Two things to note:
  - i. In these examples the time lines for population, and data analyzed with it do not line up perfectly (for example the population numbers are from 1831 and suicide numbers are aggregated from 1827 – 1830). This should be fine for a HISTORICAL analysis (it's the best we have for the time period), but be careful when doing this with modern data (*i.e.* make sure your time periods line up).

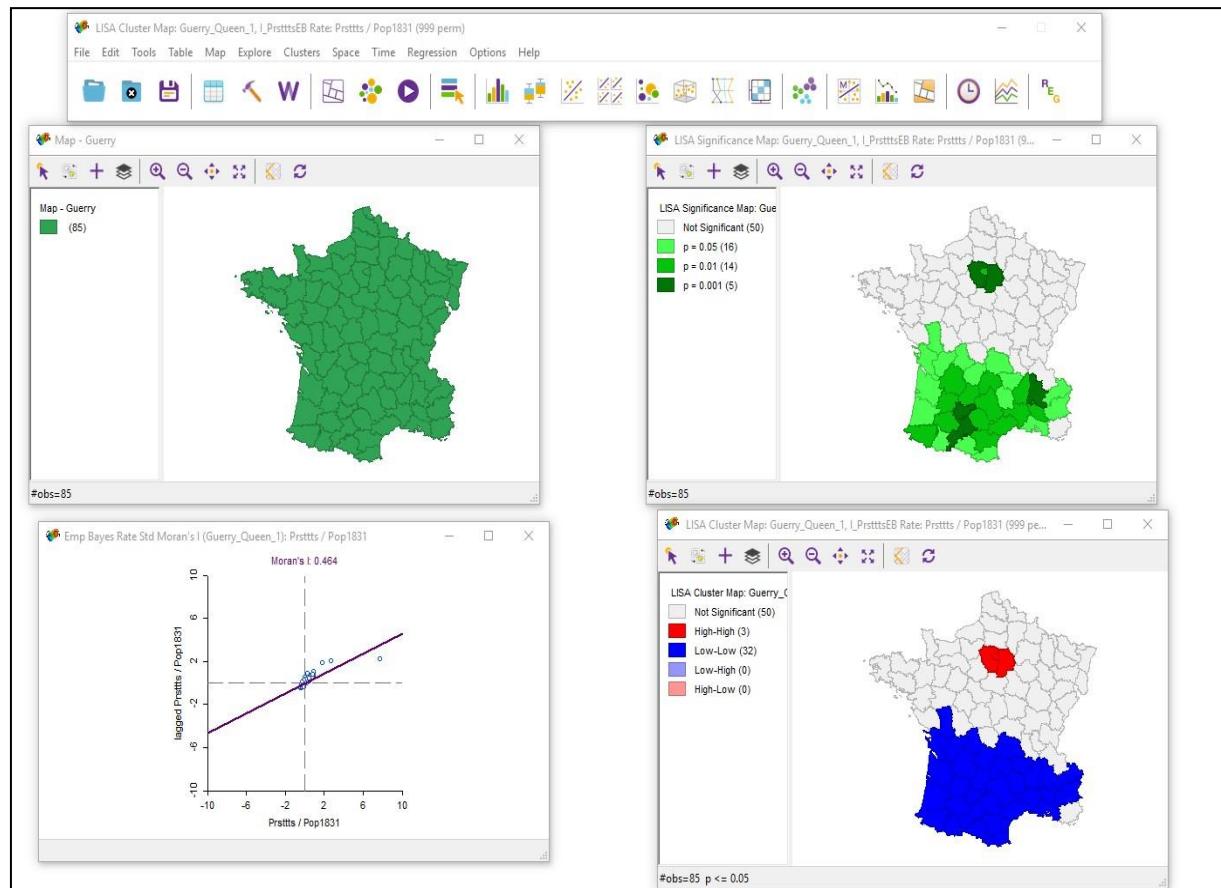
- ii. Following from above, some of the numbers here may be a bit off, all-in-all. Just roll with it!
- b. From your readings: Rates have an intrinsic variance instability, which may lead to the identification of *spurious* outliers. In order to correct for this, we can use smoothing approaches (also called shrinkage estimators), which improve on the precision of the crude rate by *borrowing strength* from the other observations. The formal logic behind the idea of smoothing is situated in a Bayesian framework (Empirical Bayesian in this sense, EB), in which the distribution of a random variable is updated after observing data. In essence, the EB technique consists of computing a weighted average between the raw rate for each county and the state average, with weights proportional to the underlying population at risk. Simply put, small counties (*i.e.*, with a small population at risk) will tend to have their rates adjusted considerably, whereas for larger counties the rates will barely change. Shown in the following equation for rate ( $r_i$ ):  $r_i = \frac{O_i}{P_i}$ , Where  $O_i$  is the number of cases (counts) in area  $i$ , and  $P_i$  is the overall population of that area ( $i$ ), the equation begins to break down (become unstable) where  $P_i$  becomes very small (as above, counties or areas with really, really small populations). EB Rates attempts to correct for this.
- c. As before, go ahead and create a spatial weights file in the usual way, defining it as Queen's case first order as before, named Guerry\_Queen\_1.gal
- d. In the first case, load a *normal univariate local Moran's I*. Using Guerry\_Queen\_1 as your weights file, select *PRST\_RT*, and check Significance Map, Cluster Map, and Moran Scatter Plot:



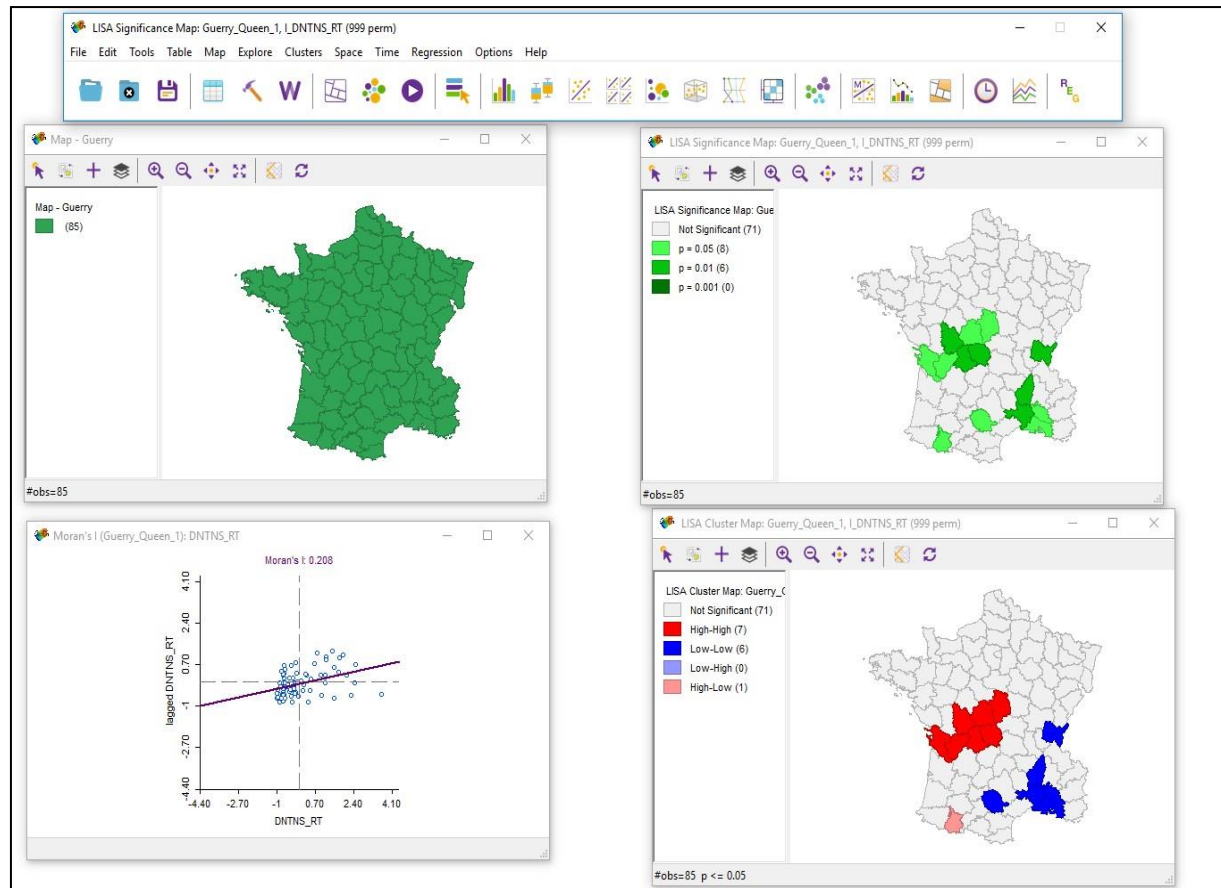
- e. For comparison to the above file, using the same weights file, now try the Moran's with EB Rates:



- f. As your event variable, load *Prsttts*, and for your Base Variable, load *Pop1831*. See below for self-check:

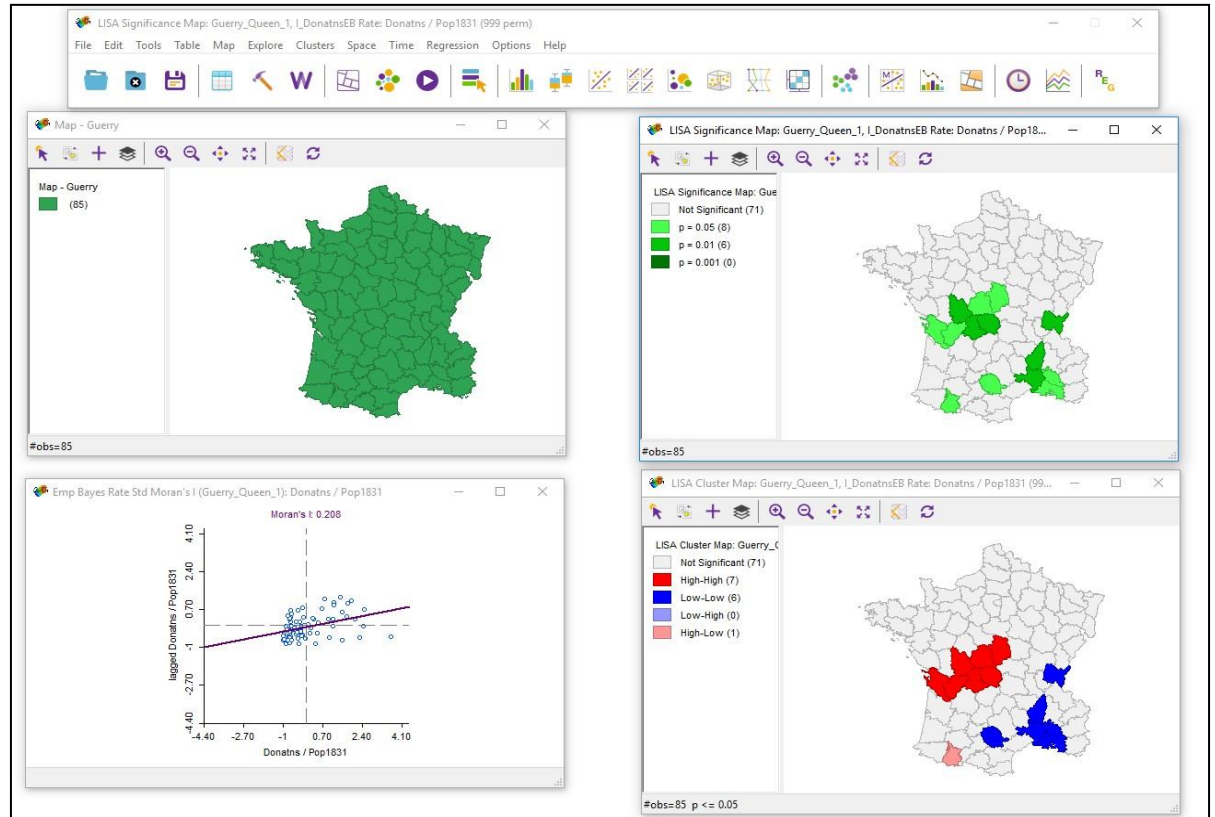


- g. You will notice they are pretty much the same, except the slope of the Moran's line is different (0.465 in the pre-calculated analysis, and 0.464 in the EB rate-corrected analysis).
- h. **For your report, please explain what the Moran's with EB Rates does; a screenshot as above (both analyses) in your final report, and what your results could mean (see next question for what I am looking for): Why do you think the results are almost identical?**
- i. Next, try the pre-calculated donations rate (*DNTNS\_RT*), with the same spatial weights file:



- j. And finally, try the EB Rates-corrected version:
  - i. Event Variable: *Donatns*
  - ii. Base Variable: *Pop1831*
- k. See next page for self-check:





- I. Now, these are exactly the same. **Why do you think the results are exactly identical (answer should be the same as above)?**

## 7.) K means clustering & gerrymandering in Texas

### a. A little background

1. When redistricting for congressional districts, state legislatures or redistricting commissions are provided certain criteria with which to draw the lines. These criteria are intended to make the districts easy to identify and understand, and to ensure fairness and consistency.

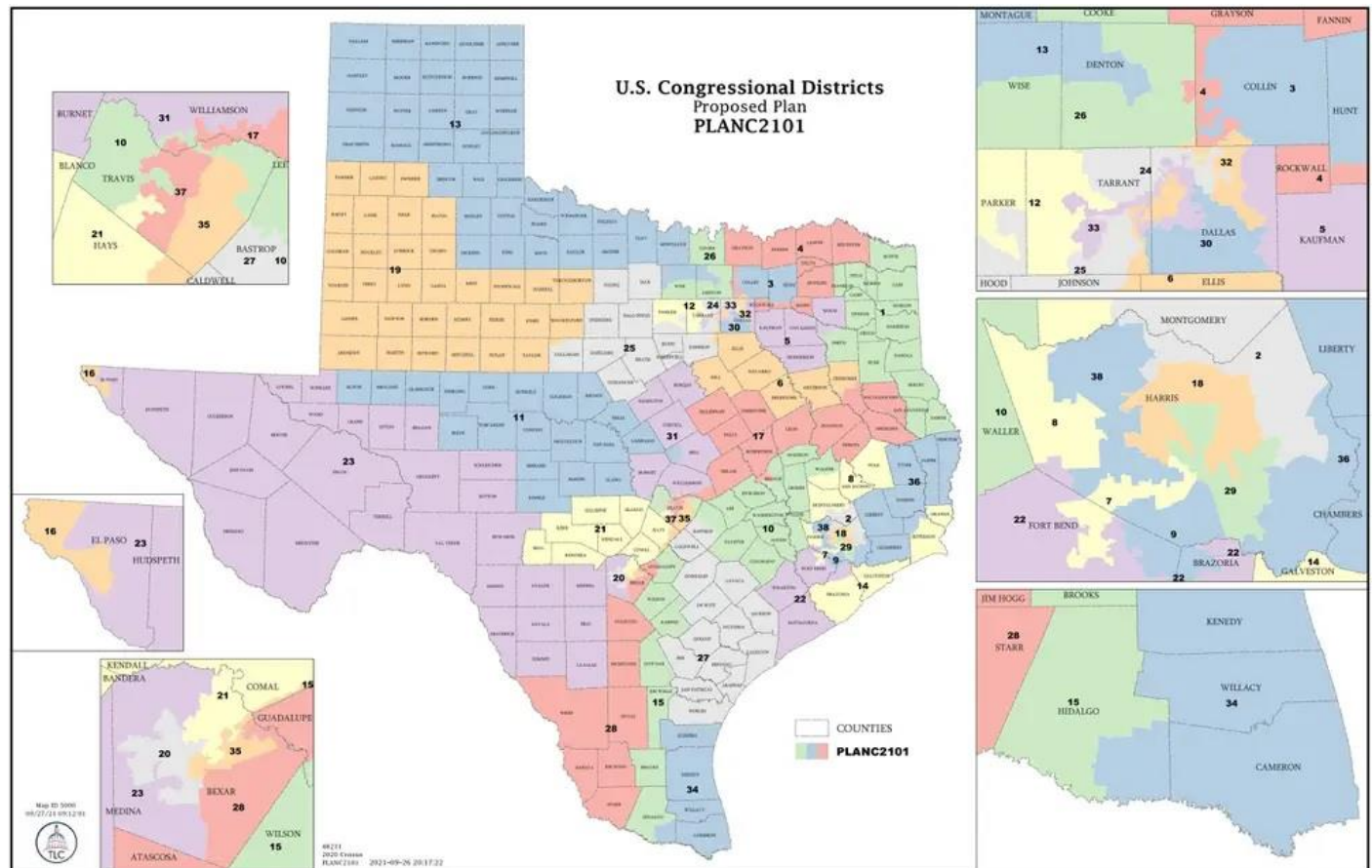
All states must comply with the federal constitutional requirements related to population and anti-discrimination. ***For congressional redistricting, the Apportionment Clause of Article I, Section 2, of the U.S. Constitution requires that all districts be as nearly equal in population as practicable, which essentially means exactly equal.*** For state legislative districts, the Equal Protection Clause of the 14<sup>th</sup> Amendment to the U.S. Constitution requires that districts be substantially equal.

2. These **traditional districting principles (or criteria)** have been adopted by many states:

- **Compactness:** Having the minimum distance between all the parts of a constituency (a circle, square or a hexagon is the most compact district).

- **Contiguity:** All parts of a district being connected at some point with the rest of the district.
- **Preservation of counties and other political subdivisions:** This refers to not crossing county, city, or town, boundaries when drawing districts.
- **Preservation of communities of interest:** Geographical areas, such as neighborhoods of a city or regions of a state, where the residents have common political interests that do not necessarily coincide with the boundaries of a political subdivision, such as a city or county.
- **Preservation of cores of prior districts:** This refers to maintaining districts as previously drawn, to the extent possible. This leads to continuity of representation.
- **Avoiding pairing incumbents:** This refers to avoiding districts that would create contests between incumbents.

For reference, here is a map of current Texas districts:



b. K means



1. *“The objective of k means is simple: group similar data points together and discover underlying patterns. To achieve this, k means looks for a fixed number (given as k) of clusters in a dataset.” ~ Andrey Bu*

2. To analyze a given data set, spatial or otherwise, the K-means algorithm starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids.

3. Those positions are optimized when it classifies objects in multiple groups (i.e., clusters), such that objects within the same cluster are as similar as possible (i.e., high intra-class similarity), whereas objects from different clusters are as dissimilar as possible (i.e., low inter-class similarity). In k-means clustering, each cluster is represented by its center (i.e., centroid) which corresponds to the mean of points assigned to the cluster.

It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their optimization because the clustering has been successful.
- The defined number of iterations has been achieved.

In this final exercise, we'll see if we can make a contiguity-based map of Texas census tracts that balances for population.

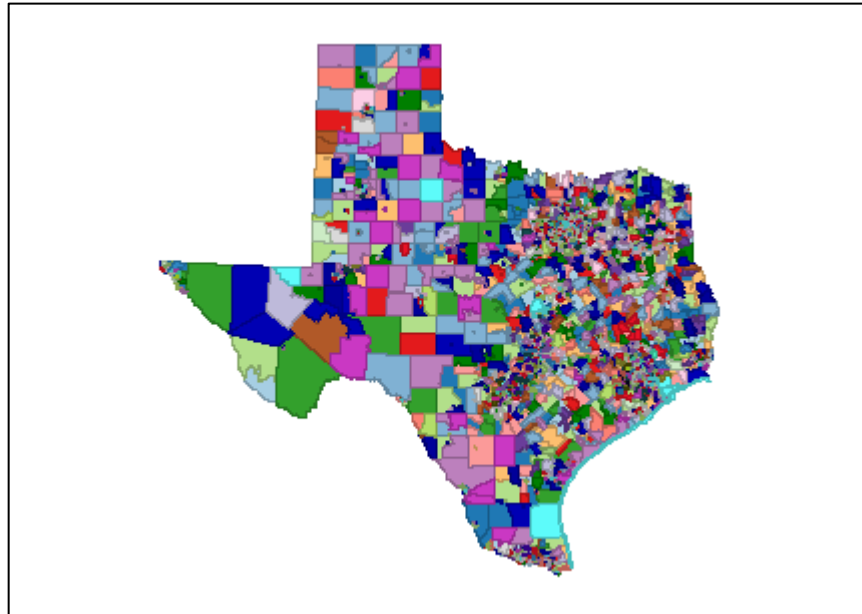
2. Open GeoDa, and load your “Texas\_Census\_tracts” shapefile from your “Texas” folder.
3. We need to create a spatial weights matrix before beginning.
  - a. Open Tools → Weights Manager, and a Queen's case first order matrix.
4. Once created, find “clusters” in the ribbon, and find “K means”.
5. Our first run will be for “total” as in, an attempt to cluster total persons in each district. We will come back to geometric centroids later.
6. We will shoot for 36 districts, the current number that Texas has been allotted. Set “**Number of Clusters**” to 36.
7. **Minimum bound** can be used for if you'd want to cluster a variable above some sort of bound. Example: you have cities with populations, and want to find clusters of those cities with populations above 50,000. Then run another with populations above 100,000, etc. We won't need this setting.

8. Set your **Transformation** option is set to Standardize (Z), which converts all variables such that their mean is zero and variance one, i.e., it creates a z-value as

$$z = \frac{(x - \bar{x})}{\sigma(x)}$$

with  $\bar{x}$  as the mean of the original variable  $X$ , and  $\sigma(X)$  as its standard deviation.

9. **Initialization method** is where you'd want k means to drop the first centroids to calculate from. Since its sort of random, and the computer learns from the first run, it can *be set* to random (see drop down) or a more careful selection known as k-means ++. We will use k-means ++.
10. **Initialization** re-runs allow the computer to “re-drop” the initialization as many times as you set it. **We'll set it to 1000**. Normally tis will not change things that much (especially in this data set), but it may help tighten up clusters.
11. **Specified seed** is a bit complicated but think of it as a “record” of this run. If you click change seed, a new run may change the outcome (since we are running a learning algorithm here). If you keep the same seed, you will see the same results. This defaults to **keep same seed** and that is fine here.
12. **Maximum iterations** is the setting to allow for a halt scenario. Basically, “find a convergence answer or stop after X iterations.” We will leave this as 1000.
13. Go ahead and run. Save your mapping and text output results, noting that they are pretty bad 😞. This is without any geographic contiguity added though.



14. Text output results:

- a. Your first section is a summary of your inputs, this should look pretty familiar since YOU set them.
- b. Cluster centers is the found value of the center point for each cluster. This does not necessarily mean it is a point that exists.
- c. Sum of squares are the values each cluster has for the sum of all values in each cluster, squared.
- d. Within cluster SOS (sum of squares): a summation of all cluster values.
- e. Between cluster SOS: value found by subtracting within grouping SOS from total SOS.
- f. Ratio: total SOS / between cluster SOS. Here, a **close** example to your output should be:  
total SOS: 6895;  
within-cluster SOS: 15.74;  
between cluster:  $6895 - 15.74 = 6879.25$ .  
ratio is then:  $6879.25 / 6895 = .997716$ .

```

-----
Method:    KMeans
Number of clusters:    36
Initialization method: KMeans++
Initialization re-runs: 1000
Maximum iterations:    1000
Transformation: Standardize (Z)
Distance function:      Euclidean

```

Cluster centers:

```

|  |total  |
|---|-----|
|C1 |3942.71|
|C2 |2654.12|
|C3 |2393.17|
|C4 |3194.2  |
|C5 |3693.65|
|C6 |3449.46|
|C7 |2931.92|
|C8 |4201.34|
|C9 |4434.92|
|C10|4665.9  |
|C11|4902.25|
|C12|2114.69|
|C13|5171.6  |
|C14|5436.7  |
|C15|1818.91|
|C16|5697.35|
|C17|6335.69|
|C18|5996.82|
|C19|1456.55|
|C20|6747.74|
|C21|7184.82|
|C22|7609.14|
|C23|8115.84|
|C24|991.698|
|C25|42.3051|
|C26|8810.9  |
|C27|9665.28|
|C28|10537.2|
|C29|11464.3|
|C30|12419.8|
|C31|13821.3|
|C32|15306.1|
|C33|17480  |
|C34|25935.5|
|C35|20539  |

```

```
|C36|30199 |
```

```
The total sum of squares: 6895
```

```
Within-cluster sum of squares:
```

```
| Within cluster S.S. |
```

```
|---|-----|
```

```
|C1 |0.548494 |
```

```
|C2 |0.605844 |
```

```
|C3 |0.553161 |
```

```
|C4 |0.497538 |
```

```
|C5 |0.478947 |
```

```
|C6 |0.474321 |
```

```
|C7 |0.530381 |
```

```
|C8 |0.465103 |
```

```
|C9 |0.41793 |
```

```
|C10|0.344206 |
```

```
|C11|0.395503 |
```

```
|C12|0.461054 |
```

```
|C13|0.406659 |
```

```
|C14|0.318169 |
```

```
|C15|0.502522 |
```

```
|C16|0.362242 |
```

```
|C17|0.505303 |
```

```
|C18|0.364205 |
```

```
|C19|0.656543 |
```

```
|C20|0.617795 |
```

```
|C21|0.484319 |
```

```
|C22|0.424031 |
```

```
|C23|0.577044 |
```

```
|C24|0.460523 |
```

```
|C25|0.150048 |
```

```
|C26|0.508388 |
```

```
|C27|0.647468 |
```

```
|C28|0.484273 |
```

```
|C29|0.277068 |
```

```
|C30|0.388036 |
```

```
|C31|0.38246 |
```

```
|C32|0.345039 |
```

```
|C33|0.306319 |
```

```
|C34|0.804326 |
```

```
|C35|3.94865e-005 |
```

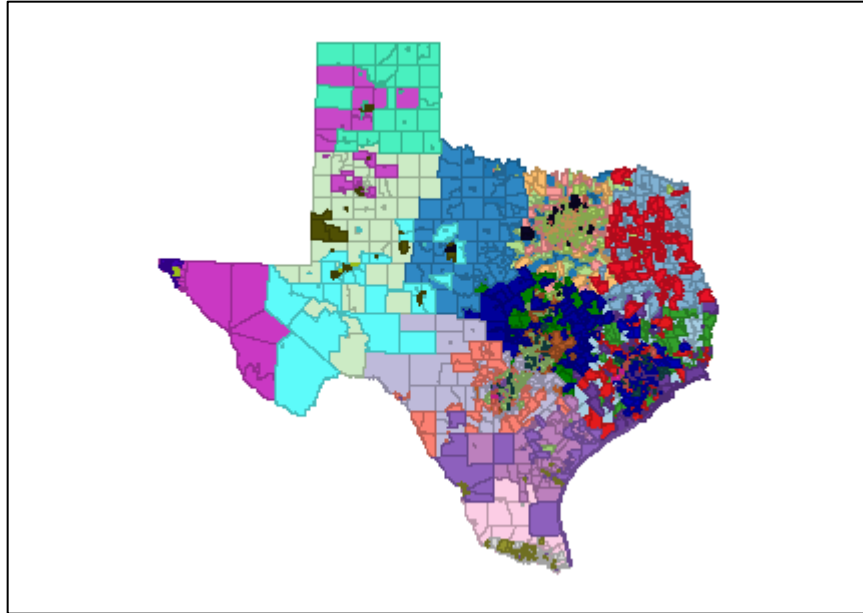
```
|C36|0 |
```

```
The total within-cluster sum of squares: 15.7453
```

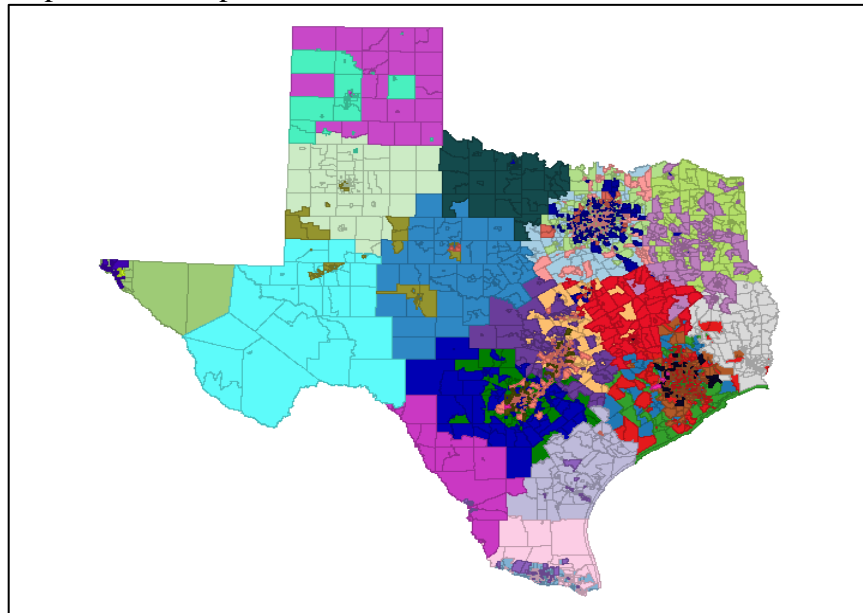
```
The between-cluster sum of squares: 6879.25
```

```
The ratio of between to total sum of squares: 0.997716
```

- g. The ratio value is what we are going to use for the overall evaluation of the clustering. Since we didn't really set any geographic rules here, it will be impossible to beat this ratio. Make a note of this and be ready to explain out as above your first ratio value.
- h. Let's do another, this time with weighting.
- i. Turn on weighting, load your Queen's case weights matrix and click Auto Weighting. Make appoint of watching it auto weight the data. It will iterate and move the lever toward the value best for allowing geography to play a role and keeping values clustered in data space. What's its doing explicitly is seeing if it can group BOTH data space clusters AND keep the data contiguous. This may take some time, but you will see it set the wright to 1.000000. What this means is it *can't* satisfy both constraints. Let's set it to 0.65 anyway and see what happens:




- j. It's beginning to cluster! Now, our ratio score has decreased, but what can you do? We need groupings! Save your map and text output.
- k. Let's try 0.85. It's getting closer but still can't match the present districts. Save your map and text output.

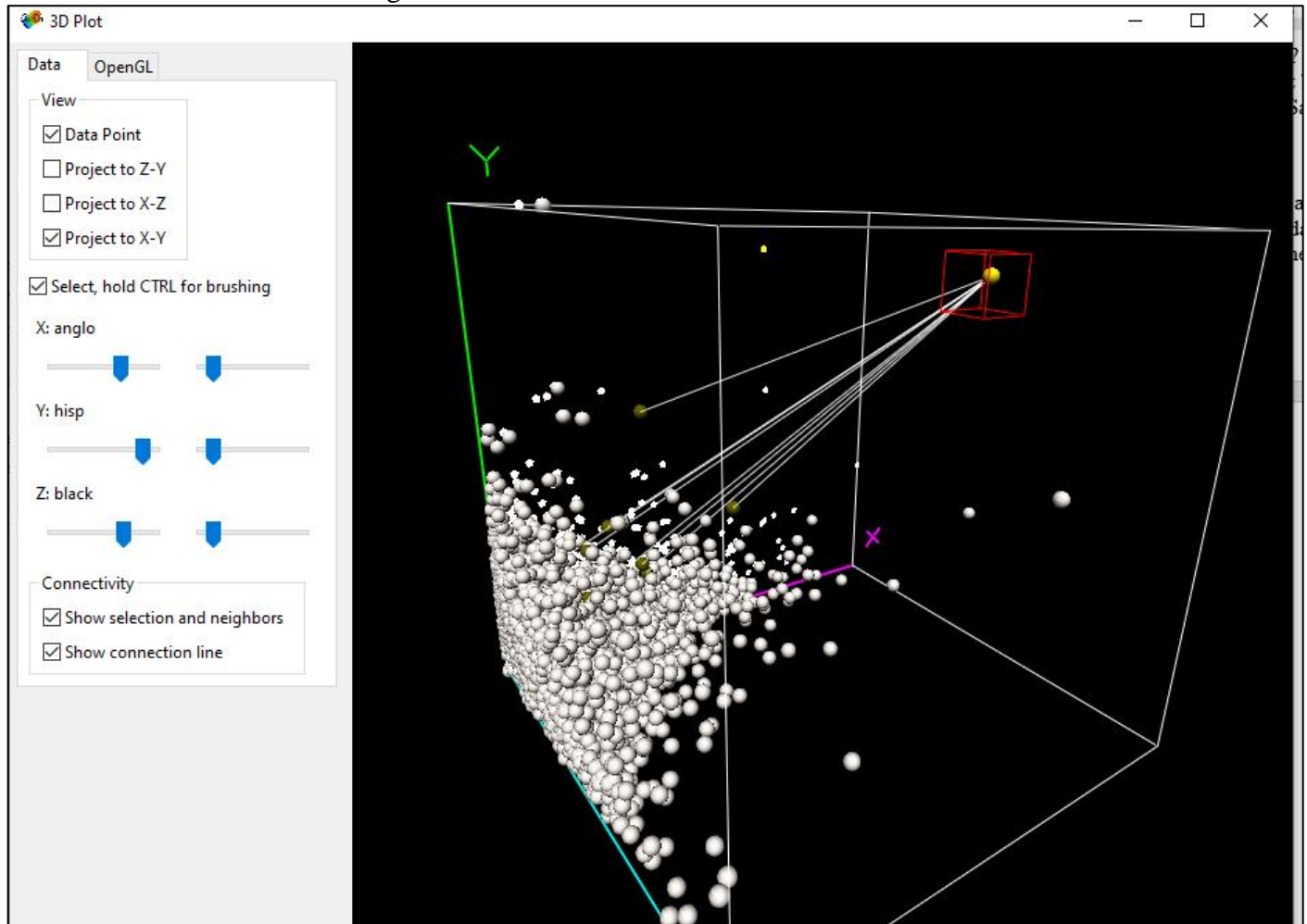


- l. Let's switch gears and see if we can group better by ethnicity?
- m. Back on your cluster settings page, UN-select total, and select "algo", "asian", "hisp", and "black". Try weighting at 0.65 again. Go ahead and run. Save your map and text output.



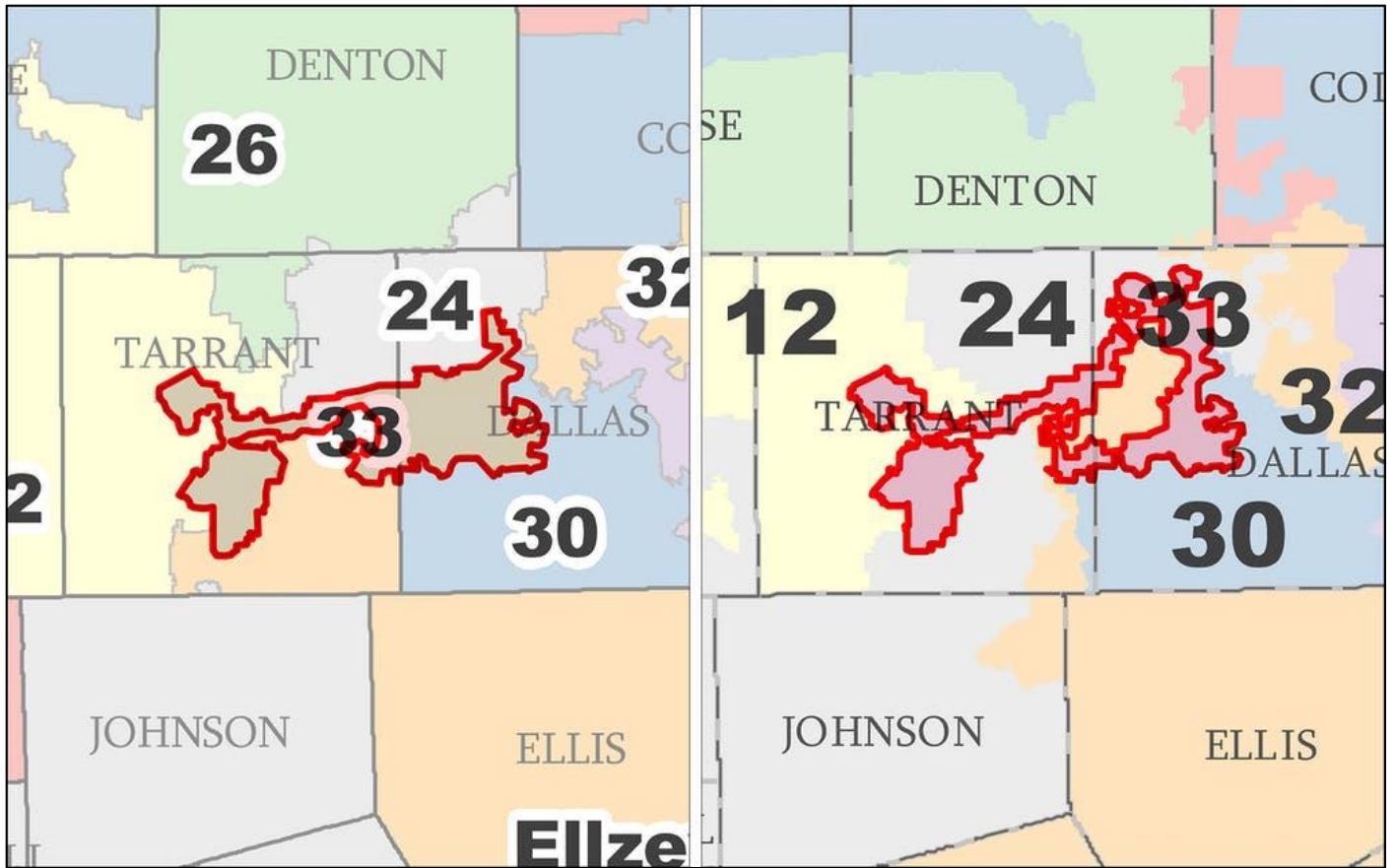
- n. Open your 3d plot visualizer , and load "anglo", "hisp" and "black". Check the box next to "Select", and navigate to the high up point in the data cloud (as

below). You will see “rays” coming from the point showing which other data points are its neighbors.

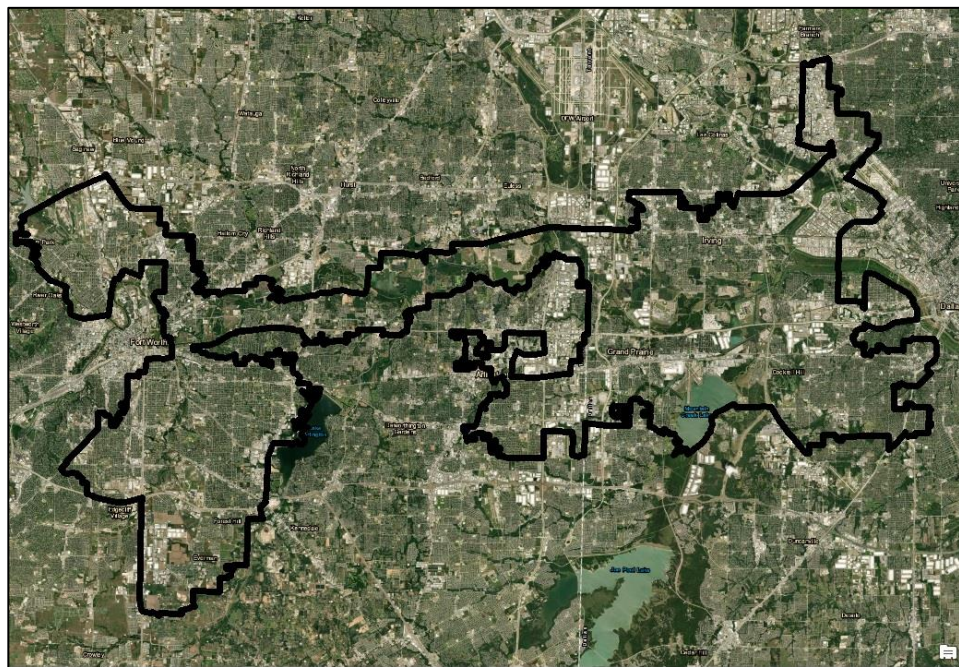


- o. Final thoughts:
  - i. We've seen it's probably not possible to create clean districts in Texas based on any required metric. That said, here is the new Texas Congressional district 33 (old on left, new on right):





Old district on imagery file (note the middle section following an area of low development):



Part of the “solution” is to tenuously connect via highways or other means areas that satisfy demographic rules.

**Deliverables listed next page.**



Deliverables for Lab 3:

Section	What to turn in
Introduction	Metadata
	Explanation of what a histogram is, and a screenshot.
	Explanation of what a boxplot is including what a "hinge" is, and two screenshots.
	Explanation of what a cartogram is, and a screenshot .
Bivariate Moran's I	Screenshot of your connectivity map for a Queen's Case first order.
	Explanation of the bivariate Moran's I.
	A highlighted county that helps explain spatial lag (with screenshot), and an explanation from you.
	Screenshot of your cluster map, along with your highlighted table, and an interpretation of your results.
	Answer to Question: if we have somewhat misleading results with this analysis of absolute population numbers, what would be a better way to do this?
Local Moran's with EB Rates	Explanation of what the Moran's with EB rates is doing.
<i>Prostitutes in Paris: 1830's</i>	A screenshot of the regular univariate local Moran's I (like on page 18).
<i>Prostitutes in Paris: 1830's</i>	A screenshot of the EB rate-corrected local Moran's I (like on page 19).
<i>Prostitutes in Paris: 1830's</i>	Explanation of your results.
<i>Prostitutes in Paris: 1830's</i>	Answer to the question (asked on page 20): Why do you think the results are almost identical?
Donations to the Poor	A screenshot of the regular univariate local Moran's I (like on page 20).
Donations to the Poor	A screenshot of the EB rate-corrected local Moran's I (like on page 21).
Donations to the Poor	Explanation of your results.
Donations to the Poor	Answer to the question (asked on page 21): Why do you think the results here actually <i>are</i> identical?
K means clustering	Your first map based on total population without geometric centroids, and your printout. Explain your printout as on pages 18 and 19 here.
	Your second map with centroids weighted at 0.65, and printout. What did your ratio of between to total sum of squares do in comparison to the first map?

	Your third map with centroids weighted at 0.85, and printout. What did your ratio of between to total sum of squares do in comparison to the second map? Are we better overall (keep in mind our main goal)?
	Your final map of clusters by ethnicity weighted at 0.65, and printout. Did you find this map was closer to the districts as seen in the reference image?