2017

Institutional Research with Public Data and Open Source Software

Vincent A. DiNoto, Jr.

GeoTech Center

10/10/2017

DUE1304591, 1644409, 1700496

Opinions expressed are those of the authors and not necessarily those of the National Science Foundation.

Getting Started

Institutional research based on student spatial location and local demographics requires the use of a mapping program, for data analysis. There are numerous software packages, which can be installed on the desktop computer, used online or a combination of both. For this demonstration, QGIS (Quantum GIS) an open source mapping program has been selected since it will work on multiple platforms and is free. Note: not all open source software is free. This 64-bit application is very fast on large data sets, not all commands are initially installed and there are numerous plugins that have been written by individuals. The methodology will be the same independent of the program. To download the software use the following hyperlink, http://www.qgis.org/en/site/forusers/download.html#, the file is relatively large so depending on the connection speed it may take several minutes. In this demonstration, the creating of mapping layers will be shown, but the actual analysis of student data will not be explored, the learner will construct a regional geography, geocode street addresses, query the data and symbolize the information.

Statement of the Problem

Clearly define the problem and state the hypothesis to be proved by the research (that involves the use of student data and demographics) is an important first step. The problem may be expanded or refined as the research proceeds, the initial problem defines the initial data requirements, but this may also evolve.

Student data must only contain attributes that are required for the research and names (student identifiers are generally not required). Addresses are required and should exclude post office boxes since they cannot properly be coded. Items such as number of credit hours enrolled, major, GPA, and total credit hours are useful parameters. Writing the query from the enrollment management system must be clear so the appropriate information is received, either in a comma-delimitated file (CSV) or as a spreadsheet. It is important when displaying data that care is taken so that no individual student can be identified. It is also equally important that all data be carefully protected and only used for institutional research purposes.

Demographical data, in general, will be obtained from the United States Census Bureau. There are many different ways in which the data can be collected from their online services. Census data is a vast data set, but generally, data used for this type of research is collected at the census tract level. A census tract is a polygon, which is contained within a single county. Census tracts were based on the concept of equal population per unit, thus the larger the area, the lower the population density. Tracts can be changed periodically (each 10 years), and as population has increased or declined they no longer contain equal populations. Every county in the United States has at least one census tract. Census information can also be obtained at the block and block group level, which are smaller divisions than the tract. The block and block group are generally too small of a division and do not have the depth of information contained within the census tract. Information is also available at the county and state level, which are too large of a division for individual college research. Information in some cases are available at the zip code

and congressional district level. Zip codes are not generally used for spatial research because they cross county boundaries and thus makes the data less valuable.

It is important to understand the region defined in the research problem. Creating maps, which contain rivers, major roads, city boundaries and the service area of the college, assists in this understanding. Overview maps showing the region compared to the entire state are useful. Written reports are also important, but are beyond the scope of this demonstration.

Some of the demographical parameters, which the author has used, include Population, Median Age, Median Income, Educational levels, African American, American Indian, Alaska Native, Asian, White, and Hispanic/Latino. In general the data contains both estimated head count has well as a percentage of the whole, it is important to utilize both parameters. Census data can be collected at the census tract level for the entire county and if the college district includes more than one county the counties can be combined together to create a regional geography.

The assumption made is that the students are reflective of the community in which they reside. For example, if a specific census tract has 85% of the population being of African American heritage, than it can be assumed that the students who live in this region are more likely to be African American. This type of assumption is not always possible, for example if 75% of the population of a census tract has less than a 12^{th} grade education, the students are not part of this assumed population, but exceptions can be equally important.

Basic Geospatial Understandings

While the geospatial technology field of study uses multiple terminology, a basic understanding of some terms is important.

Clipping is the extraction of data contained within a polygon boundary. For example, census tracts from a single county are to be extracted from a state level file, a polygon county file would be the clipping boundary. There are other ways to get the census tracts for a single county.

Merge is the process of combining multiple items of the same type together to form a regional geography map. For example, if the college service area is composed of five counties, combining these counties into a single geography is important. The combining must be of the same type of information, so only polygons can be combined with polygons and only census tracts with census tracts. You would not combine a county boundary file with a county census tract file. The polygons do not need to be contiguous.

Symbolism is an important concept that is used in all mapping applications. Symbolism is how information is displayed, such as colors, symbols, fills and widths of lines. Ramp colors, a group of colors to fill a polygon, should be selected to best understand the information being presented and care should be taken for those with visual color deficiencies, (do not use red and green on the same map). Care should be taken not to represent ethnic groups with a specific color. Symbolism allows the designer the ability to explain information discovered in the mapping documents.

A **tabular join** is the process of combining a table of data with a shape. For example, if a census tract shapefile of a single county and tabular income data for that county, needs to be combined so that location based incomes can be explored, a tabular join would be done. There must be at least one common field between the table and the shapefile.

A **spatial join** is the process of understanding data contained with a set polygon boundary. For example, if a census tract has the location of students plotted and the designer wants to know the number of students contained within each census tract than a spatial join would be done.

Geocoding is the process of giving address data a physical map location. This takes the street address and gives it an x,y position. To do this process a geocoder is required, which has a locator file. For example, student addresses are usually given with a street address, city, state and zip code; they need to have a geographic location so they can be used with the census demographic information.

Query is the process of asking questions of the data, for example, only display the locations that have median incomes less than $32,500. Queries that are more complex can be written that do multiple comparisons to reach a solution, for example, show those census tracts in which the median income is more than a certain amount, the number of people of Asian descent is a certain amount, and that 45% of the people have a college degree. To accomplish this process the individual must also understand basic logic and mathematical operations. A SQL type statement is used to search the database.

Labels for the information may be required to provide an understanding, but care must be taken that the labels do not obscure the data. Too many or too few labels can both create problems with displayed information. If too many labels are used data might be covered by the label and if too few are used not enough information is provided to the user.

There are multiple types of databases, in most mapping research the user creates a geodatabase, which is a non-relational database that means only a single individual can be connected at any one time. There are also relational databases such SQL or Oracle that can be used in which multiple users can be connected.

Generally, the files are referred to as shapefiles, vector-based information, and take the form of points, lines and polygons. A point has no size but a position in space. A line has no width but a starting location, a length and direction. A polygon is a closed figure surrounded by line segments.

Software

Numerous mapping platforms can be used to accomplish the goals set forth in the problem. Some of the software is free and others are commercial. So that the participants in this workshop can return to their home institution and repeat the process, it was decided to utilize free open source software. Note not all open source software is free. Open source software can have the same abilities as purchased software, but generally lack the customer support. In addition, users generally have the ability to write code that can be made available to other users. For this

exercise, it was decided that the QGIS platform would be utilized to meet the needs of the participants, since it runs on multiple platforms.

QGIS will operate on multiple platforms including Windows and Mac. The user will also be able to access files that were created in other application software. For example, a geodatabase created in Esri ArcMap Desktop can be used for the QGIS analysis.

Data Protection

As noted previously it is critical to receive only the data required to solve the problem and that the data contains only needed information. All data received should be protected in secure storage and not presented in any fashion that an individual can be identified by name, student ID or a location on a map.

Therefore, when projecting student location, they can be viewed as points on a map when a large area is displayed, but when zoomed to the street level the individual dots should be removed and only a summation of numbers presented for the area such as a census tract.

Getting Started

The first concept will be getting basic shapefiles (they maybe in a geodatabase) of the area to be studied. This is accomplished by visiting the U.S. Bureau website at: https://www.census.gov/geo/maps-data/data/tiger.html. **TIGER** stands for **T**opologically **I**ntegrated **G**eographic **E**ncoding and **R**eferencing.

Which product should I use?

Product	Best For...	File Format
TIGER/Line Shapefiles	Most mapping projects--this is our ***most comprehensive dataset***. Designed for use with GIS (geographic information systems).	Shapefiles (.shp) and database files (.dbf)
TIGER Geodatabases	Useful for users needing national datasets or all major boundaries by state. Designed for use in ArcGIS. Files are extremely large.	Geodatabase (.gdb)
TIGER/Line with Selected Demographic and Economic Data	Data from selected attributes from the 2010 Census, 2006-2010 through 2010-2014 ACS 5-year estimates and County Business Patterns (CBP) for selected geographies. Designed for use with GIS.	Shapefiles (.shp) and Geodatabases
Cartographic Boundary Shapefiles	Small scale (limited detail) mapping projects clipped to shoreline. Designed for thematic mapping using GIS.	Shapefiles (.shp)
KML - Cartographic Boundary Files	Viewing data or creating maps using Google Earth, Google Maps, or other platforms that use KML.	KML (.kml)
TIGERweb	Viewing spatial data online or streaming to your mapping application.	Interactive viewer, HTML data files, plus REST and WMS map services

Figure 1: TIGER Data

The table in Figure 1, shows the different boundary products that are available from the Census Bureau, generally the first two are used for this type of research. The geodatabases provides multiple layers, where the shapefile download provides only single layer. The information will download in a compressed format (zipped), before using the information it must be decompressed. The shapefiles and geodatabase requires a geospatial program to view. For this example the geodatabase for Kentucky was downloaded, roads, streams, and other parameters

are useful items, which can also be downloaded from the same site, but may be at the national level.

All images displayed in this lesson were taken from QGIS 2.18.13 using the Microsoft Windows operating system.

Loading a geospatial file:

1. Open QGIS
2. Open a geodatabase

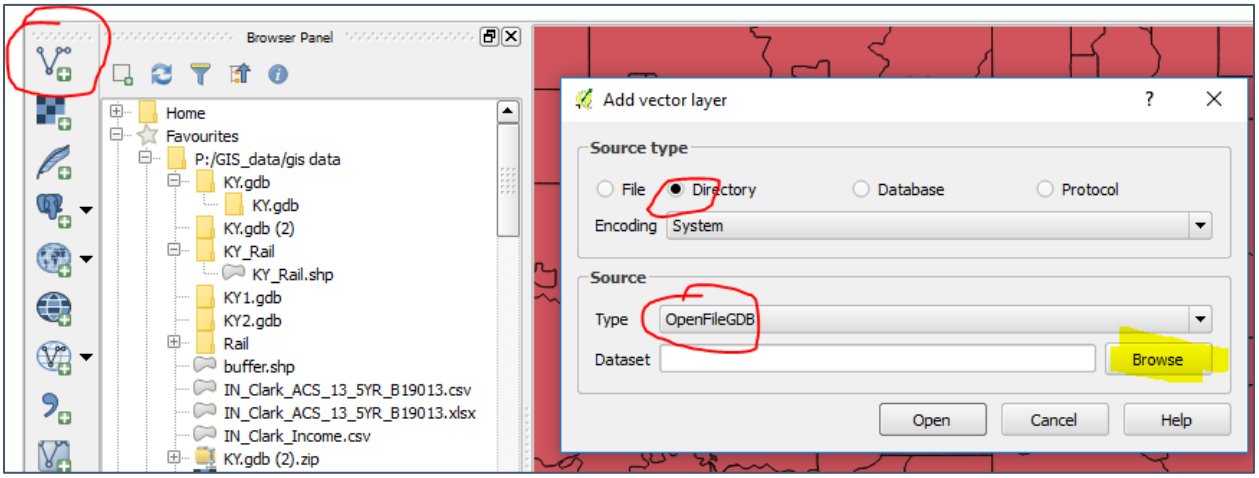

Figure 2: Opening a file geodatabase

   a. Select the icon on the upper left and a new window will appear,
   b. select Directory,
   c. use the pull-down to select OpenFileGDB,
   d. select Browse, browse to the location of the decompressed geodatabase from the downloaded from the Census Bureau. The geodatabase will look like a file folder but will have a .gdb extension,
   e. the final step is to select Open twice.
3. When the geodatabase opens a list of available geographies will be displayed. Select the Census tract one first.
4. Open the attribute table, this is done by right clicking on the loaded file, the type of data contained within the file should be visible. Note there are 1115 census tracts in the state of Kentucky.
5. Geodatabases from the U.S. Census Bureau cannot be directly edited when loaded into QGIS. To edit the information the format of the geodatabase must be changed. Select the geodatabase and right click to open the menu and select Save As. Save the geodatabase as a shapefile in the same folder as the original geodatabase. This saved file should be loaded into the operating system; the geodatabase file can be removed if the save was successful. To open the shapefile use the same icon as in Figure 2, but select file instead of directory. Browse to the saved location.

Understanding layers:

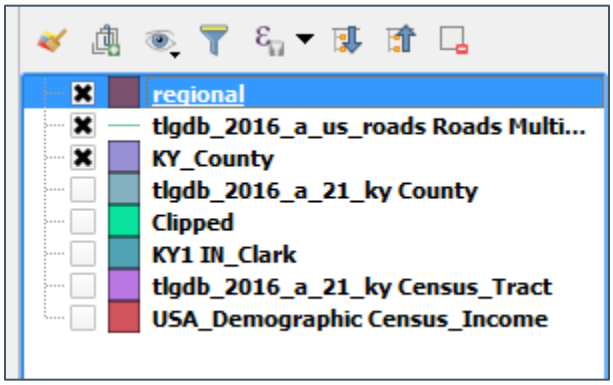

Figure 3: Layer Order

A layer of information is displayed on the bottom left corner. In this example, multiple layers have been loaded onto the map, those layers with an x are visible and the other layers are currently not visible, see figure 3. One layer is a line file, which contains roads for the entire country. The other layers are polygons and have an opaque fill color. The top three layers are turned on. Regional is the highest placed layer and since it is opaque, no information appearing below it can be seen. Therefore, the line file (second layer) for roads cannot be seen in the area in which the regional layer is visible. The roads will appear on top of the rest of the polygon counties in Kentucky. Layers work from top to bottom; generally, points are the highest layer, then lines and finally polygons. So that the road shapefile is on top of the regional shapefile, drag the road file above the regional file or the regional file below the road file, by clicking and holding. None of the information displayed in the unmarked layers is visible since they are not selected. When a layer is removed, it is not deleted from the storage site, but only removed from the map project.

Regional geography

Regional geographies can be created several different ways and different methods will be discussed, to some degree this will depend on the format of the geographical data.

Merge Method:

If the information is at the individual county level (each county individually downloaded at the census tract level), then the counties, need to be combined together. This is accomplished through merging the same type of shapefiles, i.e. census tract files.

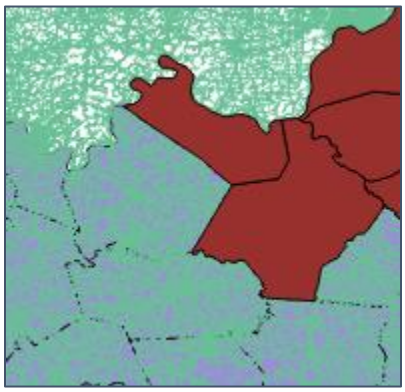

Figure 4: Opaque Layer

Individual counties of the same file type are selected and combined together using the merge command. For a merge to be effective, each file needs to be the same vector type, such as points, lines or polygons. The files must have the same type of information, for example, they all need to be at the census tract level. The concept of merging will be discussed in another section.

Regional Geography by Selection

Another method, if the data is statewide for example at the county level would be to select those counties to be eliminated from the study area and save the counties that remain as a new file, unfortunately it is a little more complicated than just selecting and deleting.

1. First, if the file is in a geodatabase it needs to be saved as a shapefile, this is accomplished by right clicking on the file and doing a save as. Once the file has been saved the original geodatabase can be removed.
2. Select the counties from the map, which are required for the study by clicking on them, hold the shift key while the selection is made.
3. Open the attribute table and select the invert button shown to the right. This will reverse the selection made, thus no longer selecting the counties chosen from the map, but all the other counties. Generally, fewer counties are selected to keep than those to be deleted. The appearance should be that most of the rows in the attribute table are highlighted.

4. The next process is turning on the editor. This is done by selecting the edit button on the attribute table that looks like a pencil. Once the editor is turned on press the delete key on the keyboard. What remains is the counties that were selected. This information needs to be saved with a new name before closing the attribute table, use a descriptive name, saving is accomplished by right clicking the file name and doing a save as. This method works well at the county or state level but not at the census tract level since there are so many tracts to select for an entire region.

Query Method

A query method could be used, but is not suggested as a method for this problem because the number of tracts that are used for most geographies. In rural area, this method would be effective. The query command will be discussed in another section.

Clipping Method

This method is suggested to be used in conjunction with the selection method for the census tracts. Once the county regional geography has been created, then this polygon file can be used to clip out the census tracts from statewide data. This method can also be used for regional roads, streams, etc.

1. To do a clip, a polygon boundary file is required, which for this case will be the regional county geography. A line, polygon or point file will be the data to be clipped. Make sure appropriate descriptive naming convention is used; a new file will be created and should be saved in an appropriate storage location.

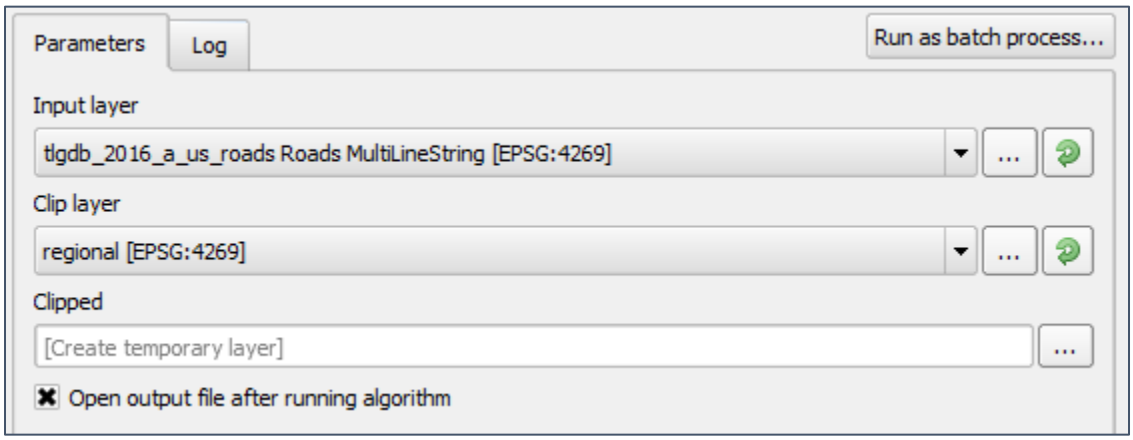

Figure 5: Clip

2. Under vector menu select geoprocessing tools and then select clip.
3. See Figure 5, the input layer is what is to be clipped, the clip layer is the polygon boundary layer and Clipped is the new file that was created. The clipped file will be saved after it is inspected. Note the check box to display the layer after the process has been completed. The larger the data set the longer this process will take.
4. The resulting output should be automatically loaded onto the map and can be used for the analysis. The larger area file can now be removed.
5. This process can be repeated for other files required for the maps, such as roads, rivers, points of interest, schools, etc. The same polygon file will be used as the boundary file.

Adding Demographic Data

Demographical data can be added to the map by joining a table to a shape. This requires two processes the first is getting the data that is required and the second is joining with a shapefile. The data set will generally come from the U.S. Census Bureau; currently the best source of this type of information is the American Community Survey (ACS). Using a five-year average of collected information is suggested, the most current information is the ACS 2011 to 2015. The annual ACS data generally does not give information at the census tract level. The following two publications from the U.S. Census Bureau will be helpful in understanding this data https://www.census.gov/library/visualizations/2015/comm/your-answer-your-future.html and https://www.census.gov/programs-surveys/acs/library/outreach-materials/Flyers/acs-how-it-works.html. It is also possible to download already mapped selected data, which might meet the needs for the project. If the already mapped features are used, generally, they are a larger area than specifically needed and the clipping method discussed previously will need to be employed. The most direct method is to locate the data set for the specific counties needed at the census tract level, download the data, and then join with the regional census tract geography.

Getting ACS Data

A good starting point is:

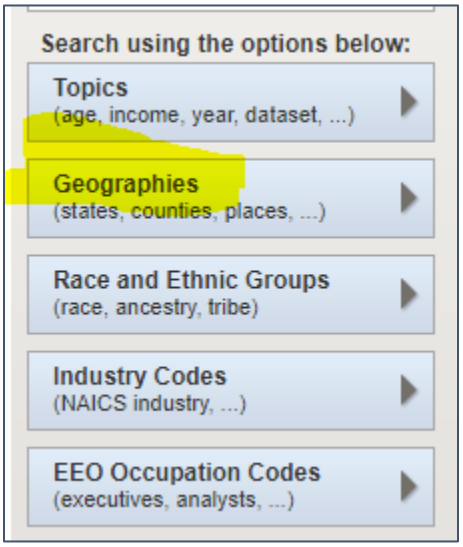

Figure 6: Setting the Geography

https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t first it is important to set the geography that will be used. This is accomplished on the left side by selecting geographies, see figure 6. Select the census tract level, than select the state, the county and highlight all census tracts. Repeat this process for all the counties, which will be part of your regional geography. Close the geography window.

See figures 6 & 8. The next component is to select the data for these counties, this is done by using the menu on the left side again, this time select the Topics area, select people, and then income. There are two choices and select the household and not the individual. Notice on the upper left corner, that the selections that have been made are shown. From this listing, the

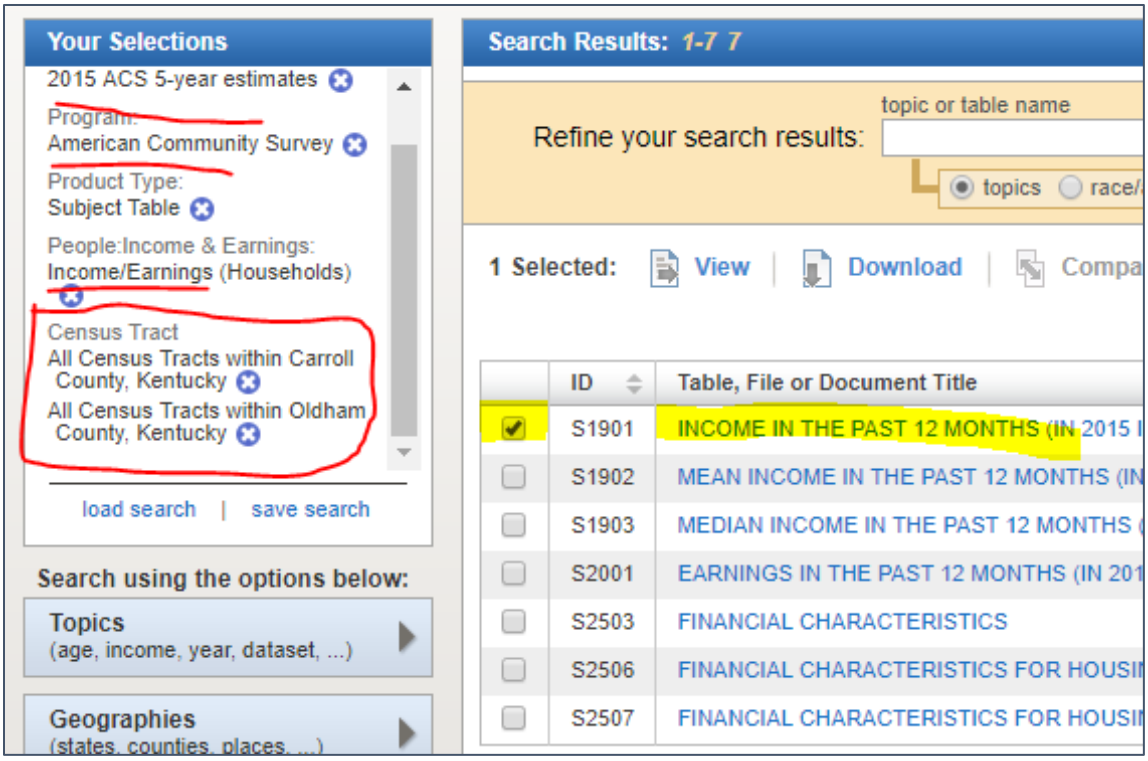

Figure 8: Data Selection

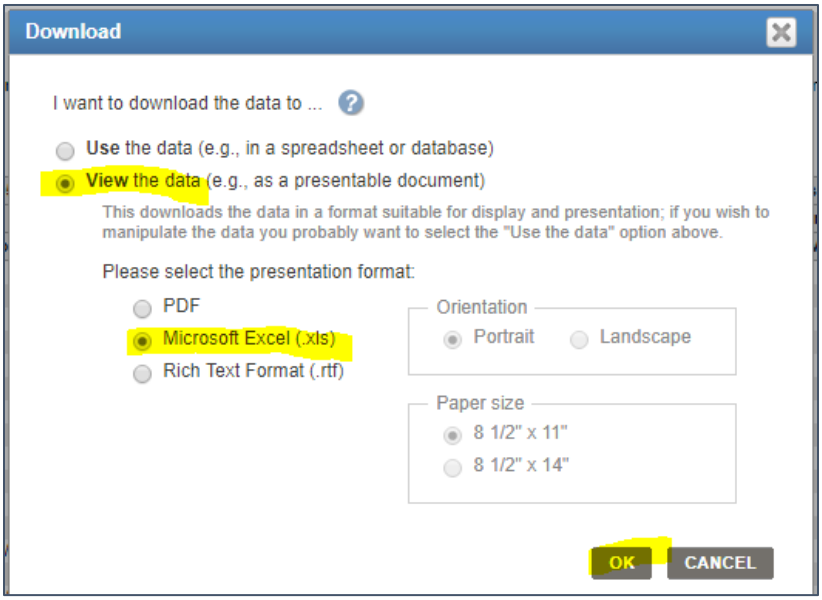

Figure 7: Data Download

appropriate table should be selected. Generally, the S1901 table is used for the demographic research we have done, select this table, see Figure 8. If the hyperlinked name is selected then the information will be visible, the next step is to select download. The download is generally done as a Microsoft Excel document that can be opened in spreadsheet programs. The file is downloaded in a compressed format, decompress and open the spreadsheet to review the information. Notice information for two counties is included in this example. All the counties in a regional geography can be downloaded at one time, which saves having to combine the spreadsheets together later. Note the amount of information contained within the spreadsheet, only one data column is required for this demographical research project.

Spreadsheet modifications

The next step is to do some modifications to the data, make sure the folder has been decompressed. The first two rows in the spreadsheet are headers and you can only have one header so delete the second row, this row is an explanation of the fields. Columns A, B and C are needed to do the join properly, Column D contains the number of households and that might be a useful attribute, the only other column really needed is column CN (for this example) labeled HC01_EST_VC13, this is the income data. The unneeded columns and row should be deleted, save the modified information with a different name, save it in a CSV format.

Loading the Spreadsheet

To create the spatial join first load the vector census tract regional file, next load the CSV file

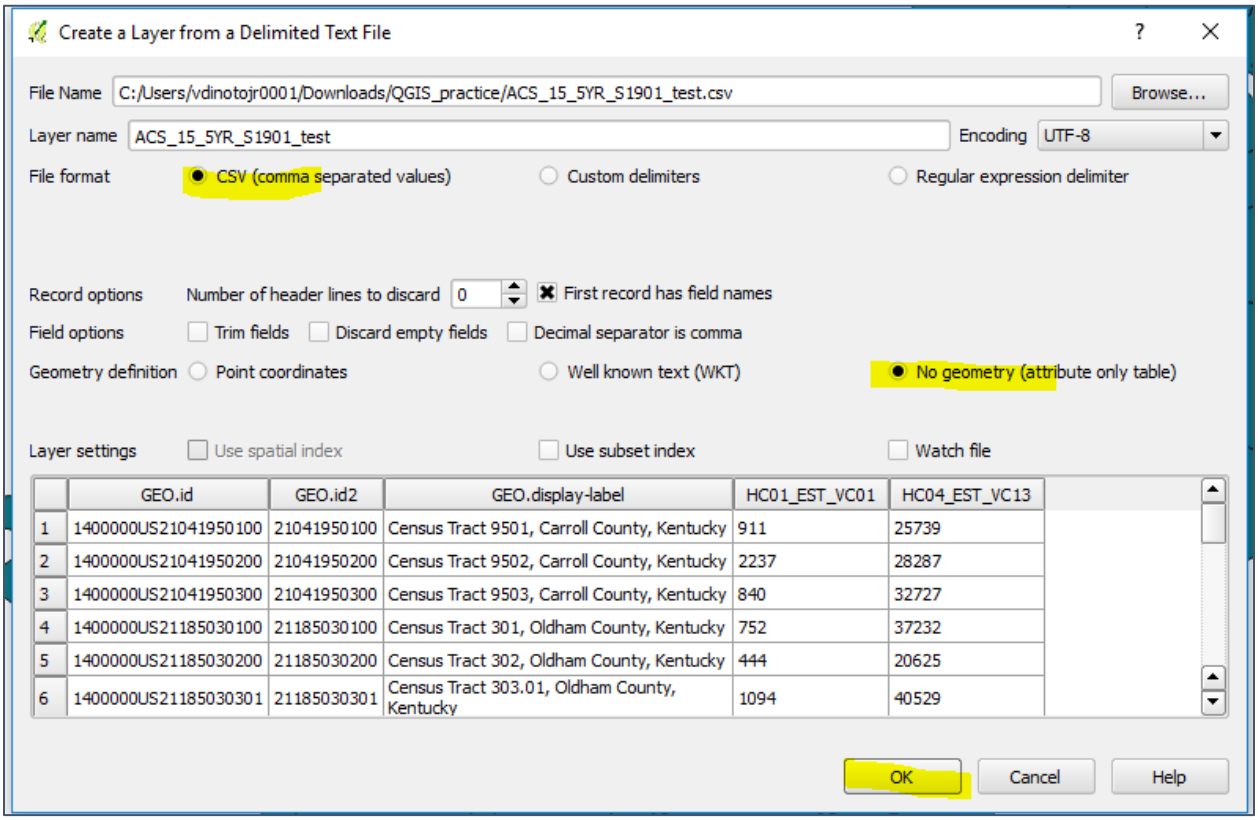

Figure 9: Loading a CSV file

using the CSV loader,  . When this tool, see figure 9, is used several selections should be made, after the file is located, make sure to select the CSV radio button and the No geometry radio button, then load the file, of course nothing will be displayed from this file, since there are no geographical locations associated with the data available.

Tabular Joining

Now that the data and the shapefile are opened in QGIS, the next process is to complete the join, for the join to be a success two fields; one in the CSV file and the other in the shapefile must have the same information. In the CSV table, it will be GeoID2 and in the shapefile GeoID. Open the properties window of the census tract regional shapefile. Find the Join Tab on the left side and select it. The next step will be to select the common fields from the two files (if more than two csv files are opened a selection must be made on which to join). Once this information is selected properly, the data from the CSV file will be placed into the shapefile and must be saved with a new name. Therefore, the created shapefile now contains demographical information. Repeat this process for other different demographic tables. See figure 10.

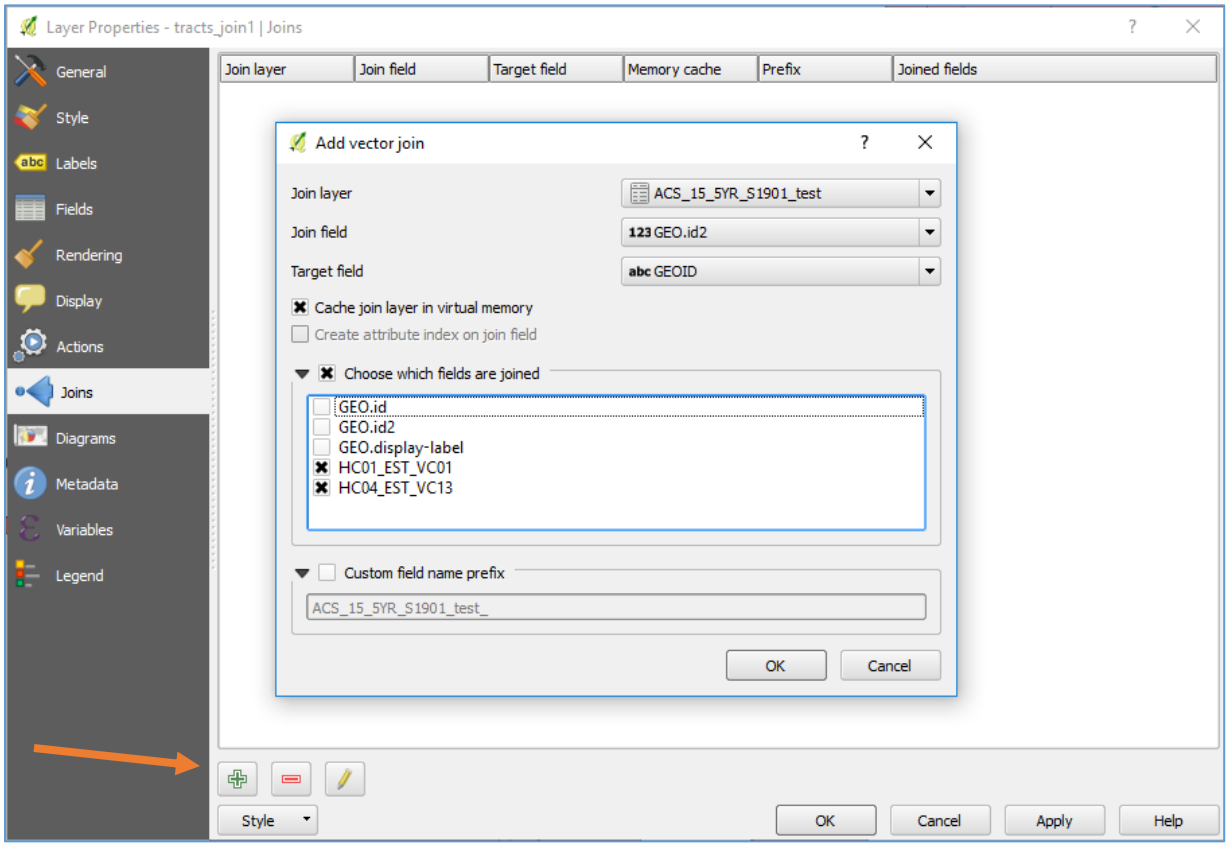

Figure 10: Tabular Join

Symbolism

Now that the shapefile contains demographical information, it is important to display the information in a useful format; this is done by selecting a ramp color, a type of classification and the field in which to be displayed. It is critical that others using the map can understand the

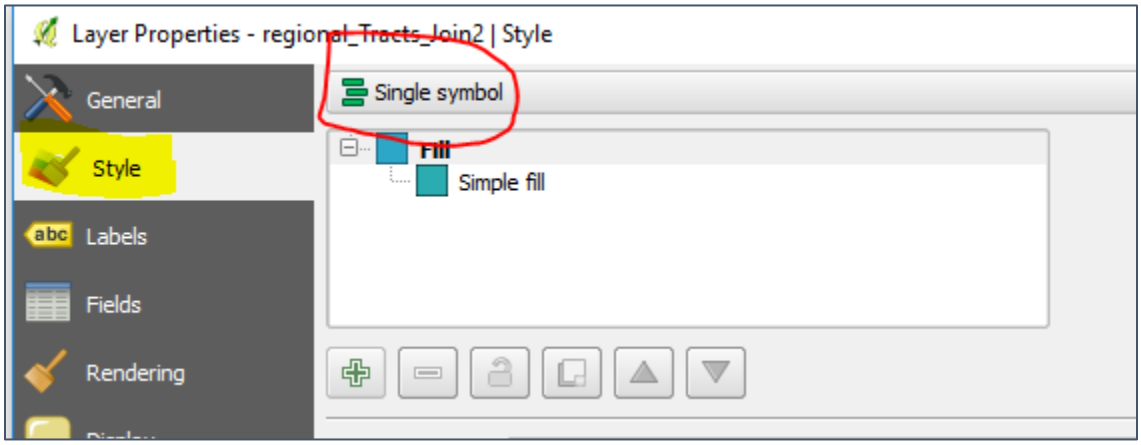

Figure 11: Style menu

information displayed. This is done in the style menu of the properties window. Open the property window and select Style, see Figure 11.

The default property is that a single symbol is used, thus every polygon has the same fill color. Since this is income data, it would be good to show the variations of income by doing a fill that uses different colors (shades) to represent different ranges of income. Use the pull-down on the single symbol and select graduated, which will open a new style window. Next, select the income field as the displayed column. Select a color ramp to be used. The next steps will specify how the data is classified.

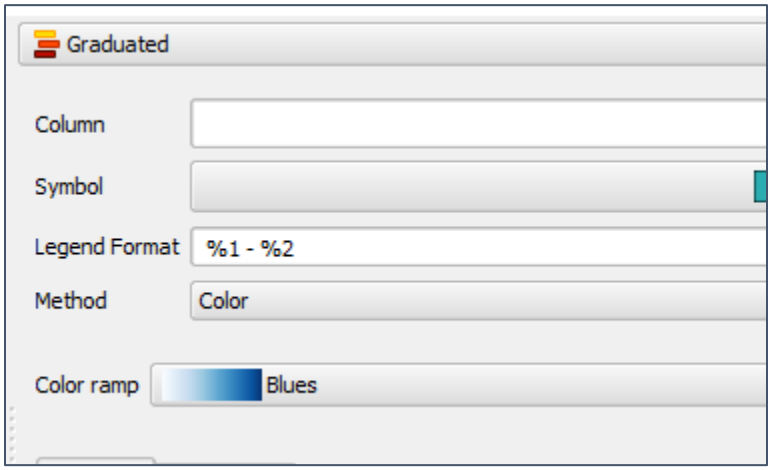

Figure 12: Graduated Fill

Figure 13: Classifying

There are several different modes of classification, equal interval is a good starting place, see figure 13. The number of classes that are being used should be enough to show variation, but not one for every different income level, for this example the default was five and was not changed. The next step is to select classify to see the different classes and the breakpoints; this can be changed as needed to highlight specific information. Once satisfied with the breakpoints and it is determined that the information is useful, click apply and okay.

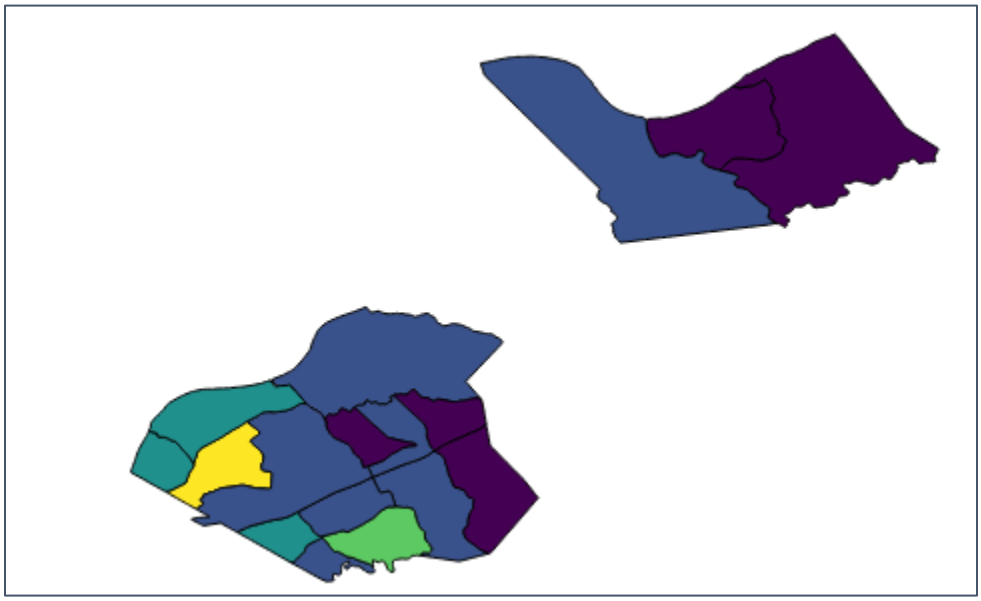

Figure 14: Two county classification

The result of the classification shows the two counties that form the regional geography, see Figure 14. The dark purple is the lower incomes and the yellow is the higher. The income breakdown with the extra decimal points needs to be removed since the income was in whole dollars.

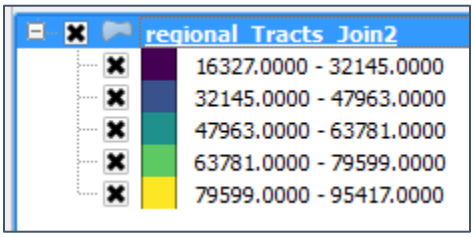

Figure 15: Classification breakpoints

As the two counties in Figure 14, show income levels the absence of a regional county map, makes the output appear incomplete. To solve this issue, load the regional county map and make the file color transparent, see figure 16, place this layer above the classified map. This will give a more refined appearance. Load the regional county map, used in the discussion of clipping. Then open the properties window and go to Style. Click on the simple

fill color and use the pull-down to select transparent. Further, down in the property window change the style of the boundary line, color, and width, see Figure 16.

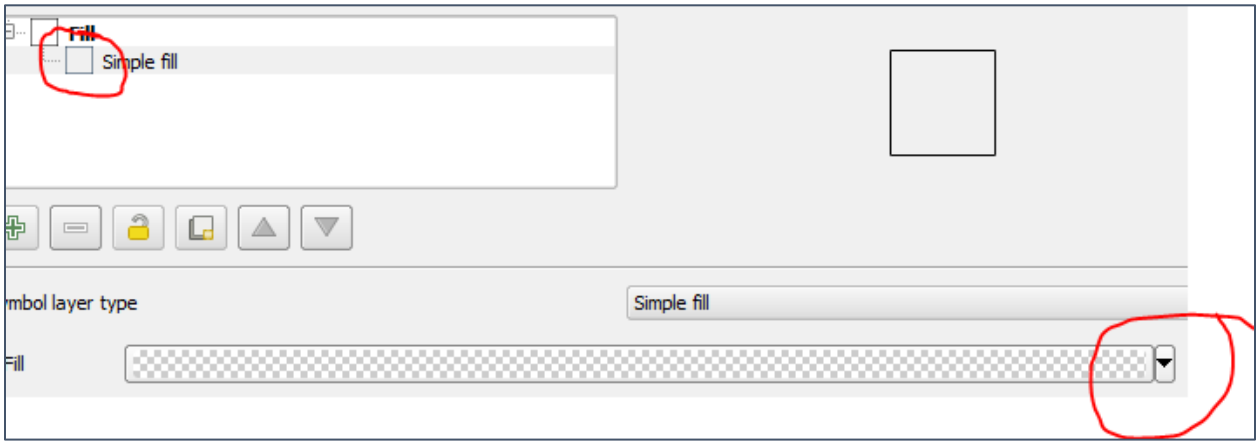

Figure 16: Selecting a Hollow Fill

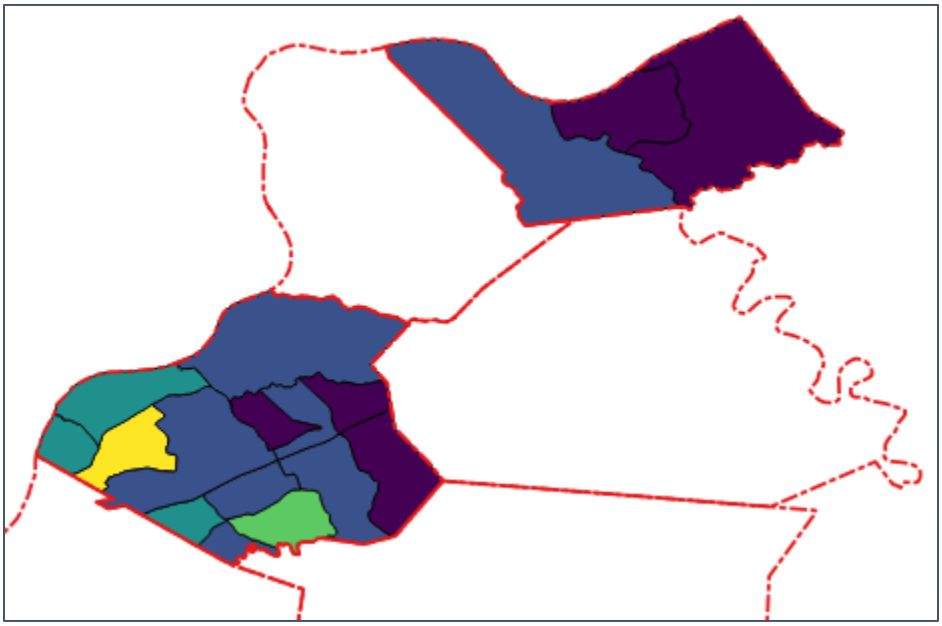

Figure 17: Results of adding county shapefile

Geocoding

Geocoding is the process of taking a street address and determining a coordinate location in two-dimensions. In general, the coordinate locations are an approximation. To complete the geocoding process a CSV file with street location, city, state and zip code will be required. A locator file is also needed to use with the addresses. The locator file provides location based addresses with coordinates. The degree of accuracy can be set in the geocoder. By changing the

accuracy of the geocoder, more or less data will be matched. It is important to understand the accuracy required to solve the problem and make appropriate selections in the geocoder.

Geocoding is a comparison of naming in the CSV and the locator file. One column in the locator will be used for zip code and the name in the CSV must be related. The same is true for state, city and address. Note, 9 digit zip codes can be a problem in some locator files and may require the user to split the zip code before attempting to geocode the results.

Generally the internal operation is to start with the entire data set of the locator and do a series of paring down, by first looking at only those addresses with a certain zip code, next additional addresses might be eliminated with state and the city filter. Finally, only addresses on the specified road are used, odd addresses should be on one side and even on the other. The road is specified as segments in the locator and the appropriate block is determined (in rural areas the blocks might be a large distance). For example, if there are eight houses in a block and the first house has an address number of 100 and the address number in the csv file is 102; this makes the address represent the second house on the right side of the street, a coordinate is given to this location. This coordinate can then be plotted on a map.

The key to finding good locations is:

1. A good locator file that is current and up to date. There are many different sources of locator files, which can be used within either QGIS or external. Many of these locator files will limit the number of records that can be batch coded without charge.
2. How well formatted is the data set (CSV), items that cause problems are apartment numbers so the address might look like 118 B Oak Lane, the B will cause an issue for the coding program and cause it not to be located. So depending on the size of the file, manually inspecting the data can be important.

Geocoding can be a slow process depending on the computer used and the locator. A few hundred records generally will go relatively quickly, but tens of thousands of records will be much slower and may exceed the free allocations. Most coders also allow records that did not match to be manually edited.

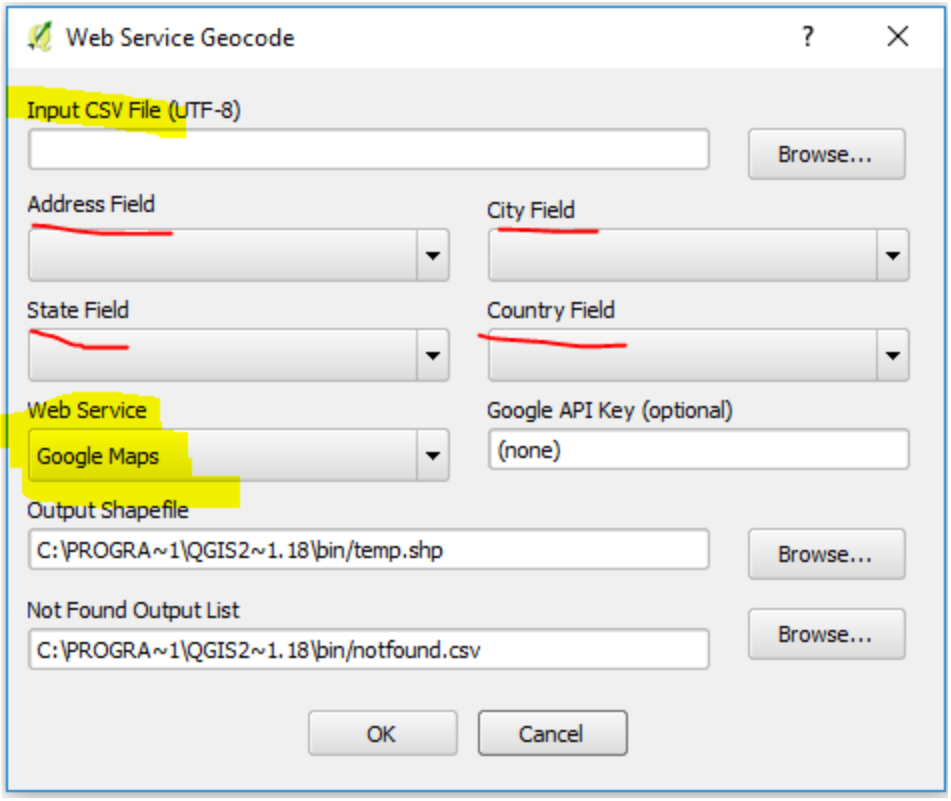

Figure 18: Geocoding

Use the plugin installer, load MMQGIS, this plugin has a geocoder, as well as other tools. A new menu will be created in QGIS and the geocoder is one of the selections (choose Geocode CSV). In the geocoder the first parameter, see Figure 18, will be the location of the address file, note it must be in a CSV format with only one header row. If the file is a spreadsheet, it will need to be exported into a CSV format, with commas being the separator. There are four items, which need to be compared; they are the address field name in the CSV, the city field in CSV, the state field in the CSV and the country field if it exists. These names are the header names in the CSV file. Other geocoders will ask for the zip code, leave any field not used blank, in this case, there will be no country field. The web service is the locator file that is being used, Google Maps, is currently selected but OpenStreetMap could be used. These map services usually have a size limitation of number of records they will process during a 24-hour time period, also they may slow the processing rate on the free version. The Google API Key only works for the Google Map Service https://developers.google.com/maps/documentation/geocoding/get-api-key , but is useful with some IP and firewall potential issues. The key is free to create and should specify geocoding, this is done through a web browser, and the key is copied and pasted in this location. The user can determine the location of the output matched and unmatched addresses or use the default settings. Once the geocoding has been completed, all successfully coded information in the CSV file will be added to a new shapefile.

The provided data set is a portion of a Kentucky Doctors file for the city of Louisville it contains approximately 500 records. While the majority of data geocoded, the map in figure 18 does not appear to have 500 dots. Many doctors are located at a single address, thus dots are on top of

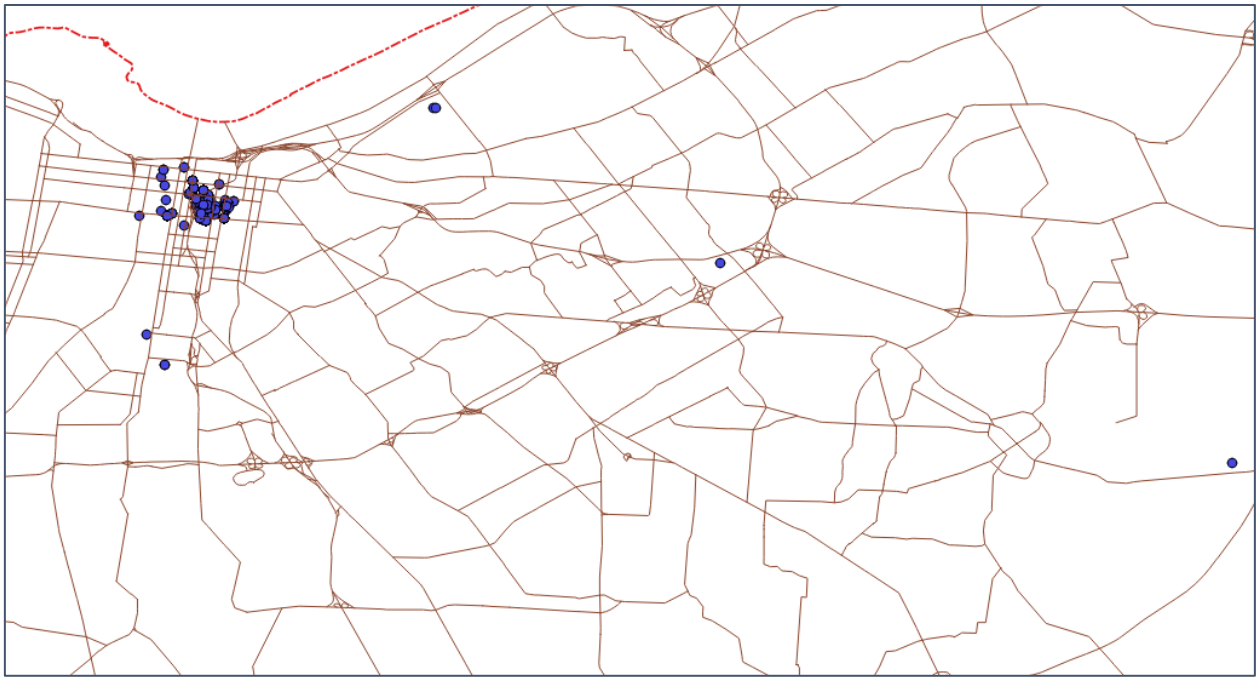

Figure 19: Geocoding Results

dots and does not appear representative of 500 doctors. There were more than 2370 doctors in the entire county.

Spatial Join

A spatial join is the process of combining objects contained within a polygon file, for this example, that would be counting the number of doctors that practice within a census tract of the regional geography. It is important when student demographic information is used to know the number of students in each census tract that will be compared with the demographical information. For example, a study might be made for a specific program of study to know if there are any general characteristics about where the students live compared with income and educational attainment.

This spatial join tool is under the vector menu, in the data management tools and is called, Join attributes by location.

See figure 21, the tool requires a polygon boundary, which in this case will be the regional geography at the census tract level. The file to be joined spatially is a point shapefile, which in this case is the doctors file from the geocoding process. For items to be joined they must be contained within the census tract, and the other parameters do not really play role. Selecting the summary will create a new column, which will be called count. For this case, data will only be present in a few census tracts. The result is the creation of a new shapefile. The results shown in figure 20 were symbolized and the labels turned on.

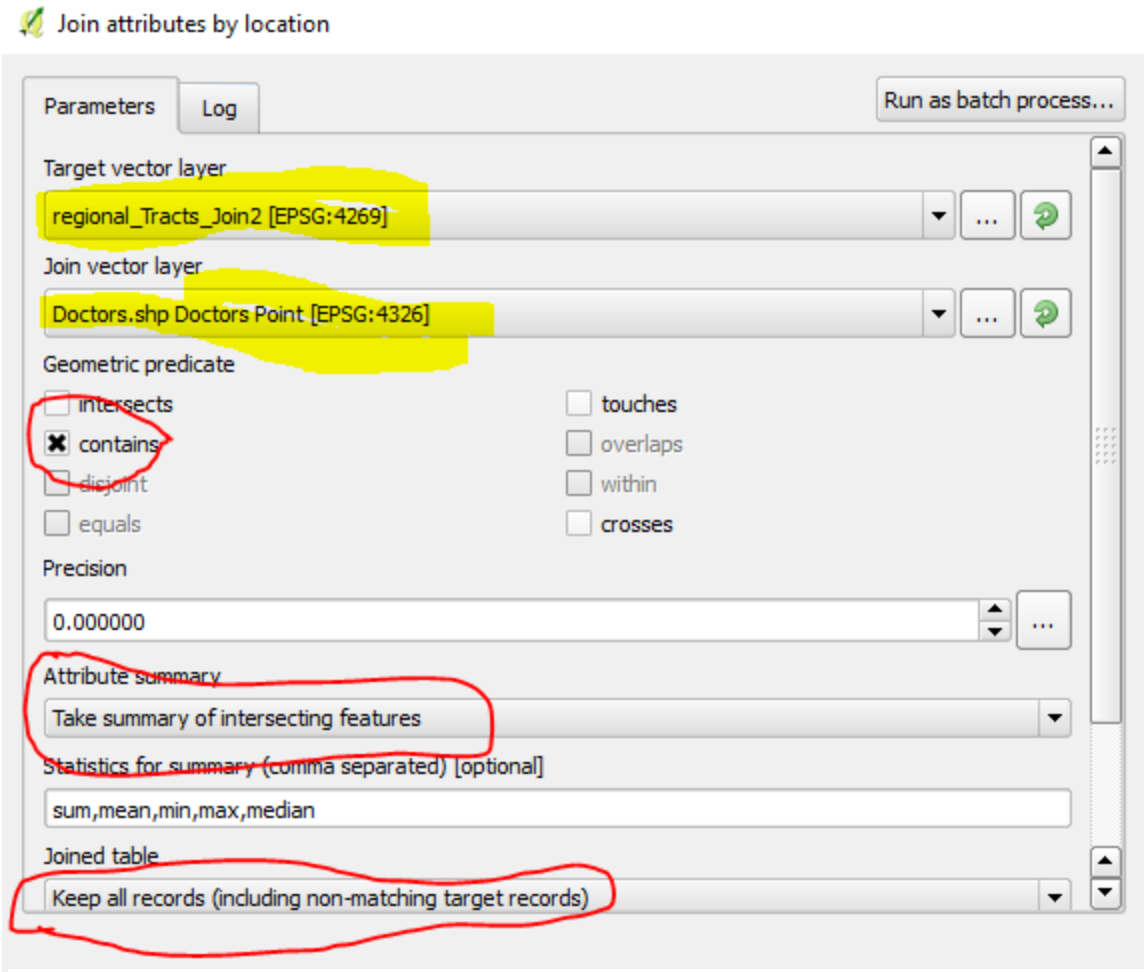

Figure 20: Spatial Join

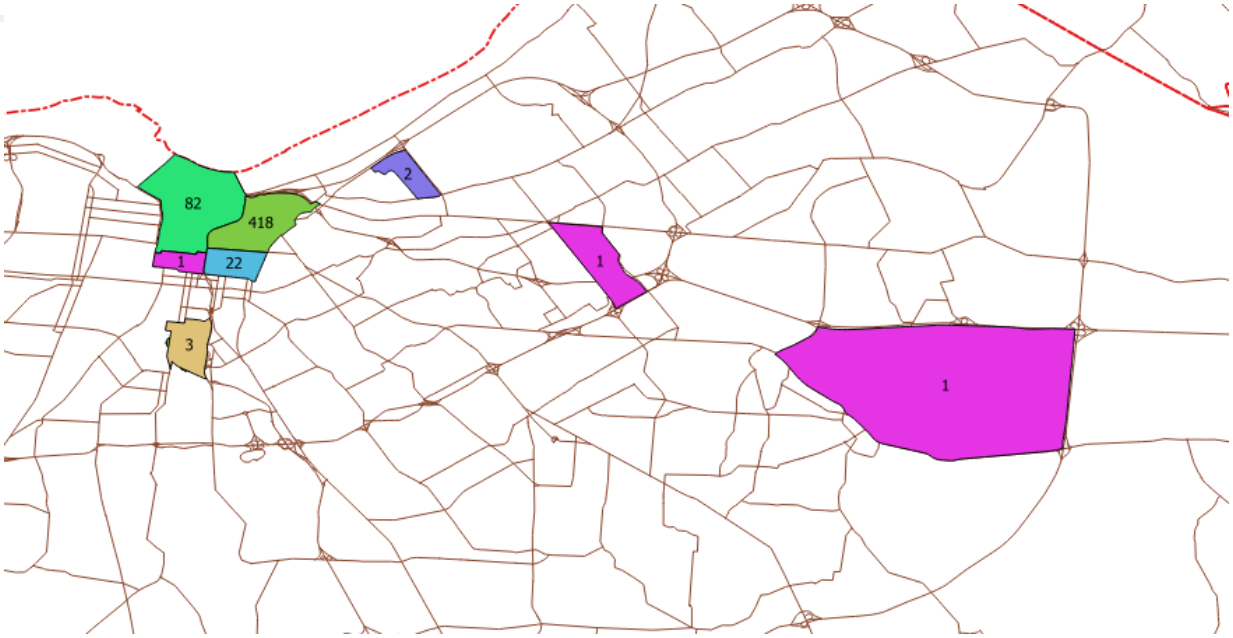

Figure 21: Spatial Join Results with Labels

Query

A query is the process of using logic to ask questions of the data, so that instead of displaying each census tract, only census tracts that meet certain criteria are displayed. The logical function can contain multiple parameters to see targeted locations.

The statistical analysis of the selected problem will generally utilize both geospatial concepts as well as basic statistics. The use of the query command is an important tool, for example, a researcher might look for those census tracts in which the median income is above the 'Living Wage' and that 50% of the population has a high school degree or less to compare with the percentage of the students which are retained between fall and spring semester of the same academic year. In addition, certain census tracts might be eliminated since the number of students living in the tract is too small to be statistically significant. For example, assumptions should not be made based on a small number of student population. There are numerous statistical tools that be used outside of geospatial analysis to make this determination.

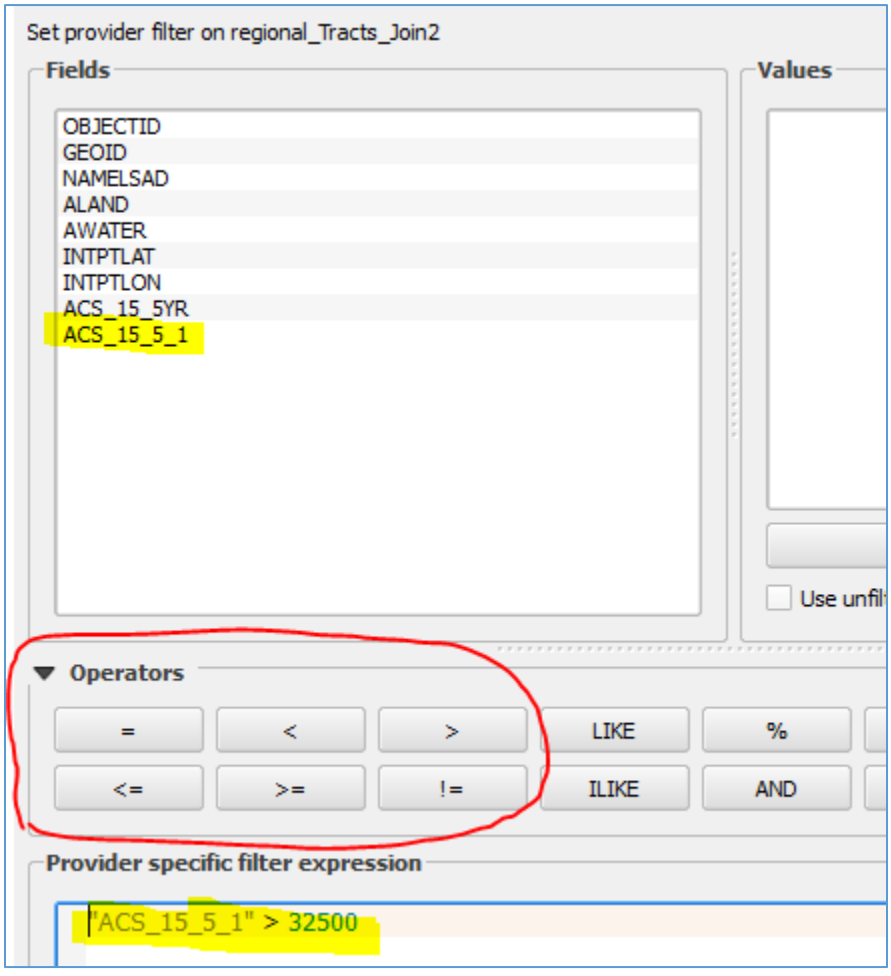

Figure 22: Query Window

The query function is part of the properties window, under the general tab and at the bottom right hand corner, scrolling to this location might be required. Click on Query Builder button to open the window, see Figure 22, to assist in writing the query.

The appropriate field for the query must be selected and multiple fields can be used in a single query. The field selected in this example is the income from the tabular join for two counties. Next, it is important to select the appropriate operation and use of the operator buttons is strongly suggested. For this simple example, the operation selected was greater than. The final component is to enter a value, note just a pure number is input which

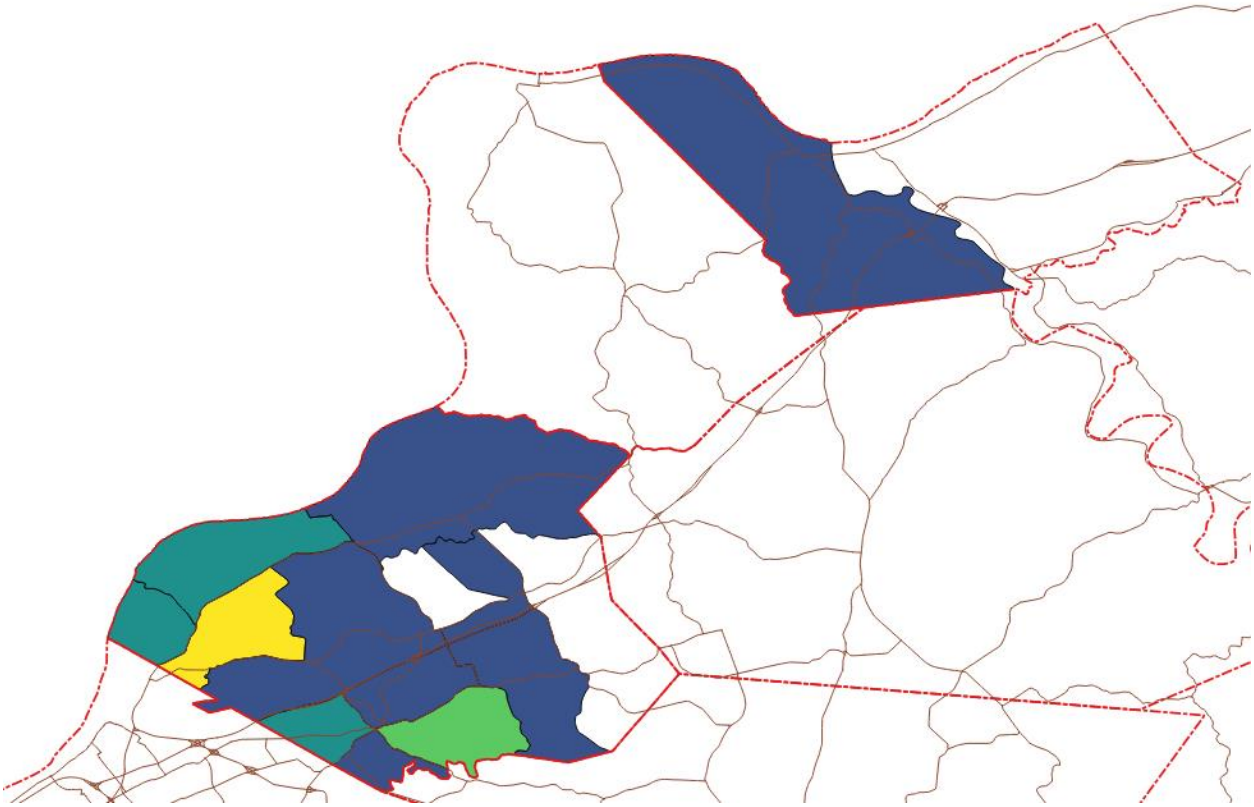

Figure 23: Query Results

corresponds to $32,500. The result of the query will show only those census tracts in which the meridian income is greater than $32,500, see Figure 23.

The query can be much more complex for other purposes, but all fields need to be in the single layer file. This include querying one field against another.

Conclusion

Concepts shown and discussed:

- Obtaining geographic and demographic information from the U.S. Census Bureau at the census tract level, which can be for a single county or multiple counties.
- Develop a regional geography which can be based upon the actual service area of the college or where the majority of the students come from which might not agree with the official region.
- The other critical factor is determining the location where the students live and then determining the number of students living in each census tract.

All data from the U.S. Census Bureau is based upon measuring a subset of the population and making determinations for the total population at the census tract level, the sample size is relatively small. It is suggested that the American Community Survey (ACS) five-year average be used, since it is believed that this is a more representative sample. ACS data does lack

specifics for the same time period as the students being studied, if current semesters are being used. The critical underlying principle is that the students mirror the region in which they live. Some of the concepts are historical trends for areas such as educational attainment, which might not be applicable for the learner. For example, a census tract in which a large majority of the residence over 25 do not have a high school education, would not be in agreement with a college student population, but would show how the students in this census tract are in the minority.

While a single semester of students can be studied, the accuracy of that information can be similar to the use of a single year ACS data, comparing students over multiple years gives a better picture of the whole. However, individual years are important to look for long-term trends, such as retention or enrollment declines. For example, a census tract that shows a student enrollment decline over multiple years needs to be examined to understand what is occurring. It is critical that students be compared in the same corresponding semesters (i.e. each fall) since it is known parameters that different semesters such as fall, spring and summer have different enrollments, retention, etc.