

Lab Exercise #6: Manipulating census data

Due Thursday, 10/10

CRP 4080: Introduction to Geographic Information Systems

Fall 2024

Prof. Wenzheng Li (wl563)

Lab TAs: Gauri Nagpal (gn247), Anika Sinthy (ats243), Shubham Singh (ss3736)

Location: Sibley 305, Barclay Gibbs Jones Computer Lab

Points Possible: 75

In this lab, you will:

- Explore how the Modifiable Areal Unit problem affects the distribution of a variable across space
- Download relevant boundary files and attribute information and join these in ArcGIS Pro.

NOTES:

1. If you export a table from Arc and wish to open it in Excel, Arc creates 2 files (a .dbf and a .xml). You want to be sure to open the .dbf file!
2. Please review the deliverable requirements at the end of the Lab Assignment to gain an understanding of what will need to be produced from the exercise.

Exploring Census Data

Understanding the Census Data Hierarchy

It is important to be able to analyze census information at the municipal, tract, block group and block level, particularly when looking at things like demographic change.

Census Tracts are comprised of several (or maybe one) *Block Groups*, depending on population density. Each Block Group is composed of *Blocks*. Blocks are the smallest scale for which census information is available. Blocks are generally coterminous with actual street blocks.

Due to confidentiality rules, some fields of census data are not available at the block level, but all the basic count information used for redistricting (a revision of Congressional Boundaries to reflect changes in population) is available at the block level. For a diagram of the standard census geographic hierarchy, go to

<https://www2.census.gov/geo/pdfs/reference/geodiagram.pdf?#>

To undertake our analysis, we need two things:

- a source for the boundary files (geographic data) and
- a source for the census (attribute/table) data.

We will join the appropriate attribute data tables to our geographical units to conduct any spatial analysis. We will discuss three sources of census data:

- New York State geospatial repository,
- The Federal census website, and

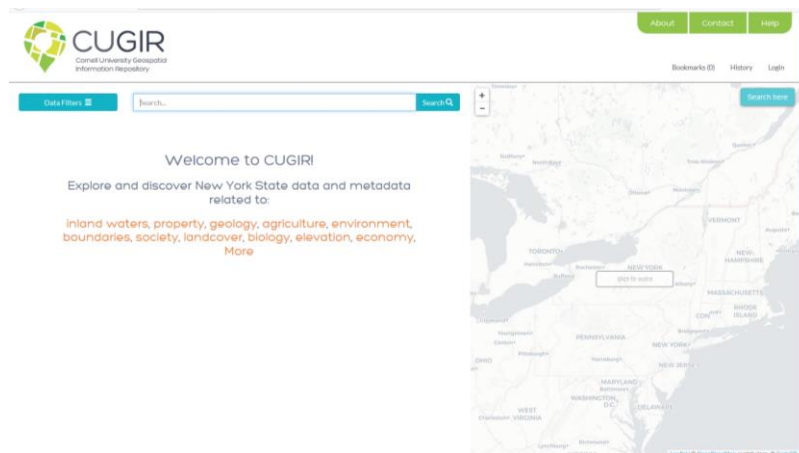


Figure 1. Cornell University Geospatial Information Repository (CUGIR)

- A private vendor.

Part 1: Accessing census data via CUGIR. We will first explore local data using the Cornell University Geospatial Information Repository (CUGIR), which has already combined both attribute information and boundaries for you. Please note that most states maintain similar geographic information repository with access to many states, county, and local datasets, usually hosted through a state agency, such as the Department of Environmental Conservation, or

other publicly funded institution. These are easy enough to find through googling.

A quick glance demonstrates that the Cornell University Geospatial Information Repository ([CUGIR](#)) hosts many datasets of interest to anyone working in the state. For this exercise, we will focus exclusively on accessing census data. Census blocks, block groups and tracts are available for 1990, 1995, and 2000.

Click “Data Filters” from the menu bar and then examine some of your options. We can explore the available data by category, year, author, collection, place, and data type. Familiarize yourself with some of the available options.

Under Theme, click “Boundaries” (See Figure 2). Note how the data is organized – if you click on a link with the spatial series identified as NY counties, you will have to further identify which county in particular (out of a total of 62). For our purposes, click on (or search for) **Census Tracts with Demographics, New York, 2000**. Note how the information is presented – the description contains some basic information about the data, the subject, the author, the year etc.

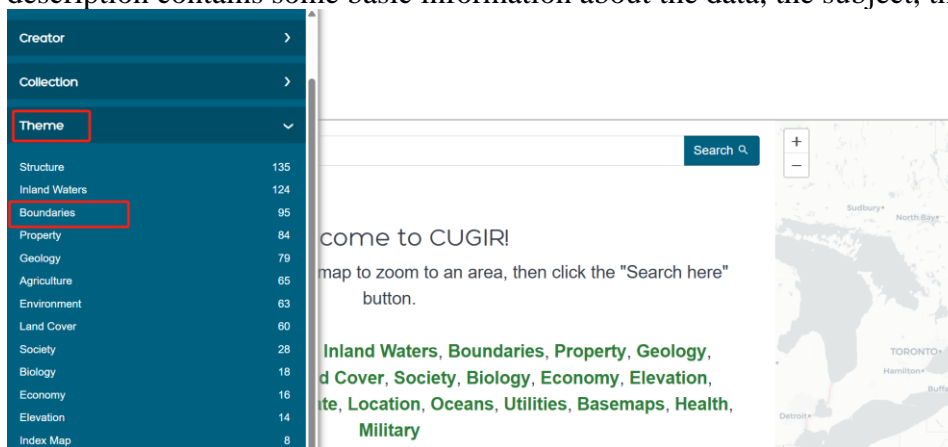


Figure 2. CUGIR Filter or Browse by Menu

Census Tracts with Demographics, New York, 2000

Author: U.S. Department of Commerce U.S. Census Bureau Geography Division

Description: CUGIR, the Cornell University Geospatial Information Repository, using US Department of Commerce, Bureau of the Census data, developed the GIS data set versions of 2000 Census TIGER/Line files for 12 coverages for each of the 62 counties in New York State. These files were developed to enable a user to see the geographic component from the 2000 Census. They can be used to join demographic data from the 2000 census for use in a geographic information system.

Collection: TIGER/Line Files

Place: New York

Category: boundaries and society

Subject: Census Tracts, demographic data, and Census

Year: 2000

File Size: 3.72 MB

More Details: Metadata

Download

Shapefile

Export Formats

KMZ (generated) **Export**

GeoJSON (generated) **Export**

Attribute	Value
Click on map to inspect values	

Figure 3. CUGIR Census Tracts with Demographics, New York, 2000

Under **Download** on the right-hand side, select ‘**Shapefile**’ (See Figure 3). Unzip the file into your lab 6 data folder. Briefly open up the html file named *fdgc* (federal geographic data committee). This is officially formatted metadata (data about the data).

Start ArcGIS Pro, add your new shapefile.

Opening the attribute table, we note the ‘STFID’ which is a unique identifier for each tract. It is a concatenation of the State (New York is 36 according to the [FIPS code](#)), County (Tompkins is 109 according to FIPS) and tract ID.

We will now build a query to select all census tracts of Tompkins County and export to a new shapefile.

- Right click the census tract layer and click “Data/Export Features.”
- Add a new expression.
- We can select by COUNTY. Build a query “COUNTY” = ‘109’
- We have now selected all census tracts within Tompkins County.
- Rename it as TC_censustract_2000 and Export it into a new shapefile

Export Features

Parameters Environments

Input Features: trt2000

Output Feature Class: TC_censustract_2000.shp

Filter

Expression

Load Save Remove

Where: COUNTY is equal to 109

+ Add Clause

Figure 4. Select by Attributes Dialog Box

Open the attribute table of *TC_Censustracts_2000* and take a look at the available information. This is the same array of data that was available in the thematic map lab. These are referred to as SF1 variables (more on this below).

You will note that the coordinates are decimal degrees, and that the data uses a Geographic Coordinate System (all Federal data uses GCS). Project the data on the fly into something more appropriate for New York state.

From CUGIR, download **Census Block Groups with Demographics, New York, 2000** and **Census Blocks with Demographics, New York, 2000**, then create shapefiles for Tompkins County using the same methods.

Create the following three projected maps (be sure to include projection information, data source, and a color classification as part of the notes on *each map*).

Compare and contrast them in terms of the MAUP (modifiable areal unit problem).

Map 1: Create a layout of normalized Black population (2000) for Tompkins County using graduated color with a classification of your choosing at the Census Tract level.


Map 2: Create a layout of normalized Black population (2000) for Tompkins County using graduated color with a classification of your choosing at the Block Group level.

Map 3: Create a layout of normalized Black population (2000) for Tompkins County using graduated color with a classification of your choosing at the Block level.

Question #1: Discuss how differences in the unit of analysis affects spatial patterns (5 points)

Creating a new field

Let's say we wish to normalize instead by 1000 population, in order to standardize our comparisons. This will eliminate the need for ratios, percentages, etc. and allow for easier whole number comparisons between states. Let's create new data by adding a new field.

- Open the attribute table for Tompkins County census tracts. Click on the "Add" button  to add a new field. Type "Blkper1000" in the *Field Name*, select "Double" in the *Data Type* field. Select "Numeric" and leave 2 decimal places in the *Number Format* field. Save and Close and to apply the changes.
- Now that we've created a field, we need to create data to populate the field (currently populated with '0').
- Back in the Tract attribute table, right click on the new field you created ('Blkper1000') and go to

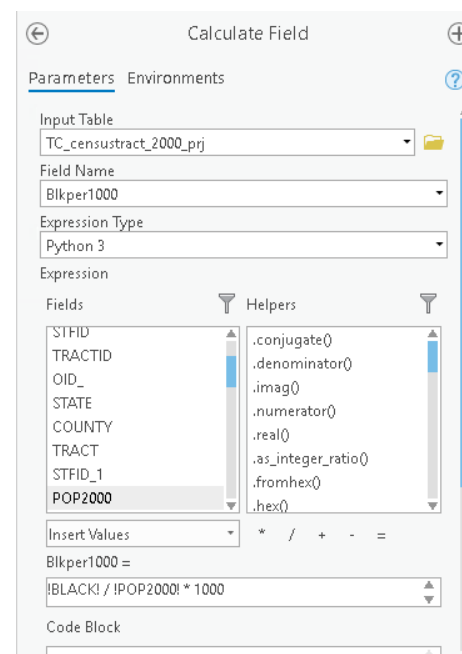


Figure 5. Field Calculator Dialog Box

‘Calculate Field’. Note that field calculator dialog box contains a listing of all the available fields, and below this is a query builder box. There are also a series of operations available.

- Enter the query (see Figure 5) to determine Black population per 1000. To enter the variables into the query box, double-click on the fields.

Map 4: Create a layout of Black population (2000) per 1000 for Tompkins County with a classification of your choosing at the Census Tract level. Note that the black population (2000) per 1000 should be shown in whole number in the legend.

Now, Select the census tracts with a majority (over 50%) of renter-occupied housing units.

Create a new field “renter_maj” depicting the percentage of renter occupied housing units per 100 housing units (in other words, a percentage).

Enter the appropriate formula in the “Calculate Field” to calculate the percentage of renter occupied housing in each census tract (Note: the field HSE_UNITS is the total number of housing units; RENTER_OCC is the number of housing units occupied by renters). We should know how to write the formula in the Expression.

Next, export to a new shapefile the census tracts with a majority (over 50%) of renter-occupied housing units. Save your project.

Map 5: Create a map layout that zooms into the census tracts with a majority (over 50%) of renter-occupied housing units. Include a context map as well. Make sure that each tract is labeled with its Tract_ID value.

Part 2: Joining boundary files and attribute data from the Census Bureau Website

An important part of spatial analysis is that you can collect additional data and join them with your shapefile data in ArcGIS. This means before we do the data join, we need to download the desired shapefile data and the tabulate data somewhere. In this section, we will join attributes from the 2020 Decennial census to Tompkins County tracts. We will first download the appropriate boundary files. The US Census Bureau website contains cartographic boundary files (including census tracts, block groups, and blocks, as well as many other sub-divisions at the state and county level).

Go [here](#) to download the TIGER/line shapefiles from the census. Note the available years. Although the census tract geography is unlikely to have changed much from any given year, select 2020 and click *Web Interface*.

You will be presented with a number of geographies, including many we mentioned in class, as well as several feature datasets (coastlines, etc.). Make sure the year is set to

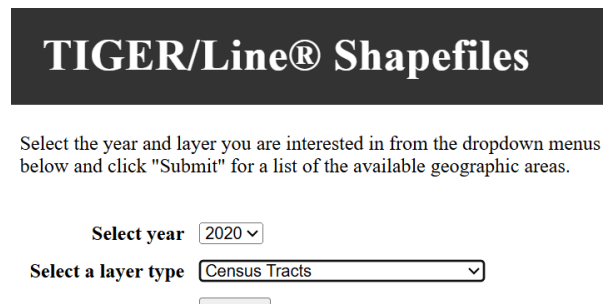
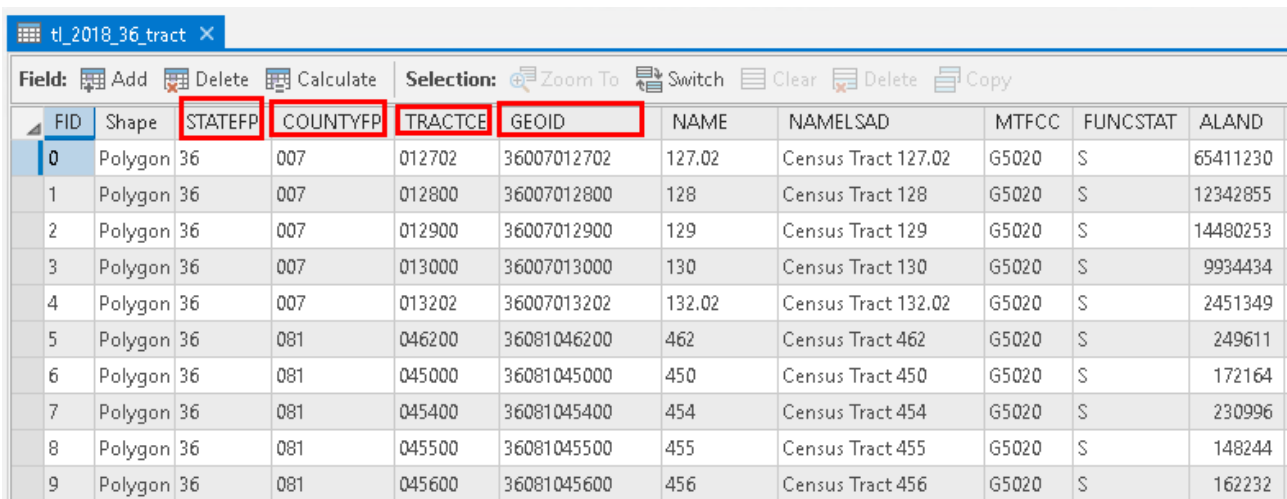


Figure 6 Downloading 2020 TIGER/Line Shapefile

2020, and under ‘Select a layer type’ select *Census Tracts* and click ‘Submit’ (Figure 6).

Next, select ‘New York’ and click ‘Download’.

- Save the zip file to your drive and unzip it. Open a new map in ArcGIS Pro and add this file.
- Open the attribute table and check out the attributes.
 - We see column entitled GEOID, which appears to be a unique identifier for each tract.
 - In the columns preceding it, we see the state identifier for New York (36), the county identifier (Tompkins is 109) and finally the individual tract identification.
 - GEOID is simply a concatenation of these identifier variables. If we were examining smaller units (Blocks, Block groups), we would simply add on to this existing number as the spatial aggregations are scalable. .



FID	Shape	STATEFP	COUNTYFP	TRACTCE	GEOID	NAME	NAMELSAD	MTFCC	FUNCSTAT	ALAND
0	Polygon	36	007	012702	36007012702	127.02	Census Tract 127.02	G5020	S	65411230
1	Polygon	36	007	012800	36007012800	128	Census Tract 128	G5020	S	12342855
2	Polygon	36	007	012900	36007012900	129	Census Tract 129	G5020	S	14480253
3	Polygon	36	007	013000	36007013000	130	Census Tract 130	G5020	S	9934434
4	Polygon	36	007	013202	36007013202	132.02	Census Tract 132.02	G5020	S	2451349
5	Polygon	36	081	046200	36081046200	462	Census Tract 462	G5020	S	249611
6	Polygon	36	081	045000	36081045000	450	Census Tract 450	G5020	S	172164
7	Polygon	36	081	045400	36081045400	454	Census Tract 454	G5020	S	230996
8	Polygon	36	081	045500	36081045500	455	Census Tract 455	G5020	S	148244
9	Polygon	36	081	045600	36081045600	456	Census Tract 456	G5020	S	162232

Figure 6. New York Census Tracts Attribute Table

For now, select the census tracts of Tompkins County using Select by attribute and export to a new shapefile with an appropriate name, e.g., TompTracts.

Finally, note the projection: Decimal Degrees. All census data is unprojected and utilizes a Geographic Coordinate System. Be sure to project it appropriately.

At this point, you probably notice that this boundary shapefile does **not** come with the demographic data or other information you need, meaning you will have to separately download the information you need and join it to the shapefile.

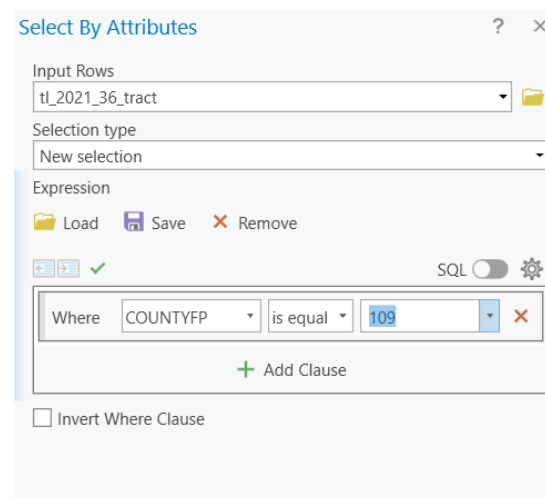


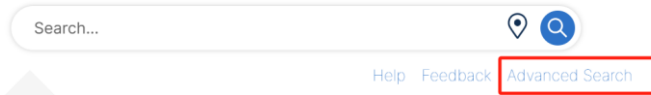
Figure 8. Select by attributes dialog

Joining attribute information to boundary files

Now that we have our properly projected boundary file, we will need our Decennial census information that joins to the boundary file. Go to [Explore Census Data](#) and click on ‘Advanced Search’.

Explore Census Data

Learn about America's People, Places, and Economy



Take note of how the site is organized. You can begin your selection (on the left side) by querying *Topics*, *Geography*, *Years*, *Surveys* and *Codes*. Take a moment to examine how each of these options is organized.

For our purposes, under ‘Years’ select 2020 (this filter will now be listed on the left hand side). Then go to ‘Geography.’ We see a listing of every possible census geography. For our purposes, select ‘Census Tract’/‘New York’/‘Tompkins County, New York’/‘All census tracts within Tompkins County’ (Figure 9).

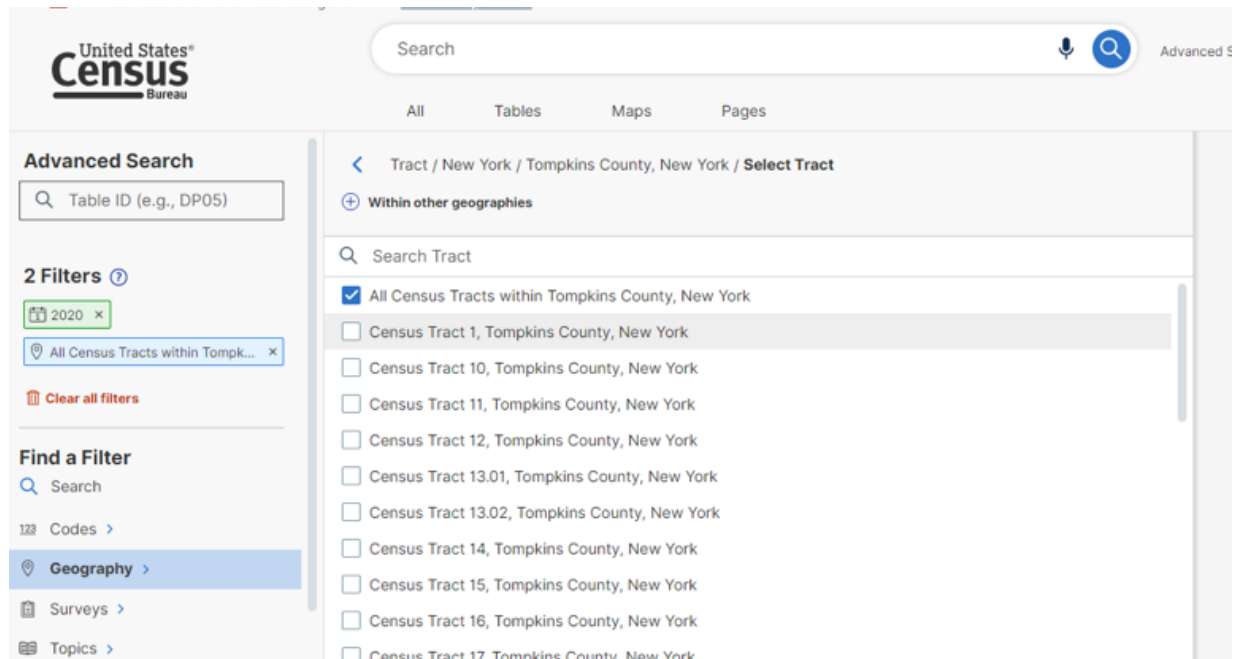


Figure 9. Selecting All Census Tract within Tompkins County, New York

The filter should now say “2020” and “All census tracts within Tompkins County.” Finally, under ‘Surveys’ select the ‘Decennial Census’ prompt, and click on ‘Demographic profile’ Click on the *Tables* tab at the top: This will open up a profile of the general population and housing characteristics (Table DP1) and includes both counts and percents (you’ll note if you click on the *Maps* tab, that a mapping features opens up. This has some geoprocessing functionality, but does not allow us to download any information).

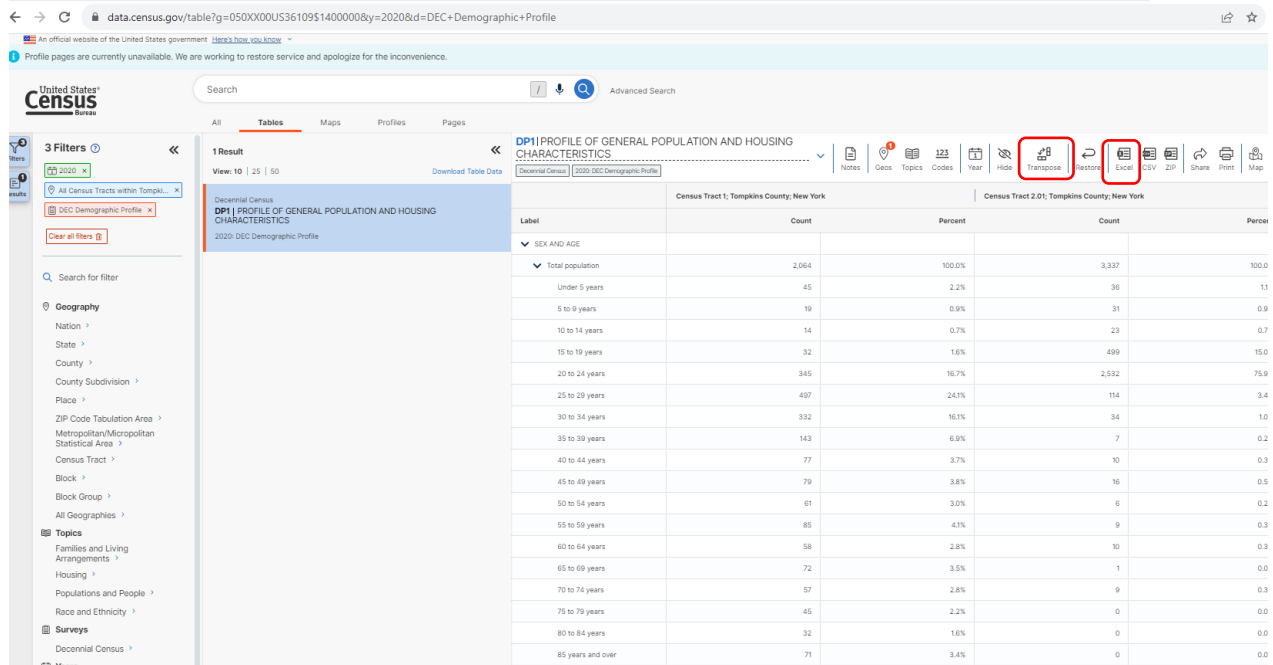



Figure 10. Downloading Data from the Census Bureau website

Before we download, click on ‘Transpose’ at the top . This will orientate your tracks as rows, and the accompanying variables as column headings. Finally select ‘Excel’ and download your attribute table.

Once downloaded, open up the data in Excel. The download consists of the metadata (under the Information tab) and the data table. You will immediately notice how unwieldy the raw census is! Tracts, uniquely identified by tract number, are in rows. Variables, identified by headers, are in columns. Ideally, your database will have field headings in the first row, and the unique identifier in the first column. Any extraneous information should be deleted (meaning any data which does not add to your analysis or is not unique).

Let’s manipulate and clean up the data. As we saw last time, we can create new fields and edit them in ArcGIS. However, it is preferable to do your manipulation, creating, editing, etc., in excel as it is generally less problematic.

Let’s say we are only interested in mapping the percent of the population that is Black or African American (this is column CH for me) This has data for the total count in each tract, as well as the percent of that tract that

	A	B	C	D
1		Black or African American	Renter-occupied housing units	
2	Census Tract 1; Tompkins County; New York			
3	Count	141	1,382	
4	Percent	6.8%	96.7%	
5	Census Tract 2.01; Tompkins County; New York			
6	Count	75	1,416	
7	Percent	2.2%	99.2%	
8	Census Tract 2.02; Tompkins County; New York			
9	Count	96	985	
10	Percent	3.8%	94.3%	
11	Census Tract 3; Tompkins County; New York			
12	Count	169	107	
13	Percent	6.1%	73.3%	
14	Census Tract 4; Tompkins County; New York			
15	Count	292	376	
16	Percent	5.9%	85.1%	
17	Census Tract 5; Tompkins County; New York			

are Black or African American. Similarly, let's pull data on the number and percent of renter occupied housing units (column FS—the last column). Delete all other columns for now! Now straighten up your column headings, so they are all aligned in the top row.

Now clean up the rows (sorry, I don't really know how to expedite this, let me know if anyone has any ideas!) and shorten the column headings so Arc will be able to read them better. Note that some of the tracts have decimals – this means that tracts have been split to accommodate population change, with the old tract number being retired.

Save your attribute table as a CSV (Comma Delimited) file. **Be sure to close out Excel!** (otherwise, you will have difficulty joining the table in Arc.

	A	B	C
1	ID	Per_Blk	Per_rent
2	1	6.8%	96.7%
3	2.01	2.2%	99.2%
4	2.02	3.8%	94.3%
5	3	6.1%	73.3%
6	4	5.9%	1%
7	5	6.4%	59.5%
8	6	2.7%	54.2%
9	7	4.3%	61.7%
10	8	10.5%	72.7%
11	9	7.4%	44.8%
12	10	13.7%	59.3%
13	11	5.8%	55.5%
14	12	7.4%	56.6%
15	13.01	2.9%	75.5%
16	13.02	3.4%	57.4%
17	14	2.4%	41.1%
18	15	3.9%	48.7%
19	16	1.3%	26.7%
20	17	3.0%	28.5%
21	18	3.2%	27.5%
22	19.01	2.2%	34.6%
23	12.02	2.7%	22.0%
24	20	5.3%	35.0%
25	21	2.4%	27.1%
26	22	1.0%	27.7%
27	23	1.8%	23.1%
28			
29			
30			

A .csv file can be understood as a clean, small, quick version of excel sheets, which is compatible with most programs. csv file does not save format information (font, font size, bold/italic styles, text color, or the multiple sheets). So, no matter how well you format the csv in Excel, once saved and re-opened, you only get the pure text/numeric content of only the FIRST Sheet.

Joining data

Now let's go back to the Map. If it's not already added, please add the Tompkins County tract boundaries shapefile you created (TompTracts). We will now join the tract boundaries with the new attribute information using a unique identifier as the common Field.

Add the CSV file. Open both attribute tables to make sure we know which variable we will use as our unique identifier for each individual tract. In the boundary layer, we found that the NAME field stores the Census Tract ID, which perfectly matches the ID field in the CSV we prepared.

Close both attribute tables.

	FID	Shape *	STATEFP	COUNTYFP	TRACTCE	GEOID	NAME	NAMELSAD	MTFCC	FUNCSTAT	ALAND	AWATER	INTPTLAT
11	5	Polygon	36	109	001800	36109001800	18	Census Tract 18	G5020	S	165139686	496155	+42.3421022
12	20	Polygon	36	109	001901	36109001901	19.01	Census Tract 19.01	G5020	S	153734340	297171	+42.3655698
13	21	Polygon	36	109	001902	36109001902	19.02	Census Tract 19.02	G5020	S	114308238	416043	+42.3392241
14	24	Polygon	36	109	000201	36109000201	2.01	Census Tract 2.01	G5020	S	154724	0	+42.4417310
15	25	Polygon	36	109	000202	36109000202	2.02	Census Tract 2.02	G5020	S	284518	1616	+42.4398340
16	19	Polygon	36	109	002000	36109002000	20	Census Tract 20	G5020	S	55093990	586578	+42.4758783
17	2	Polygon	36	109	002100	36109002100	21	Census Tract 21	G5020	S	87447965	609044	+42.5265868
18	8	Polygon	36	109	002200	36109002200	22	Census Tract 22	G5020	S	79314153	207523	+42.5886009
19	17	Polygon	36	109	002300	36109002300	23	Census Tract 23	G5020	S	169763397	19074296	+42.5786235

	ID	P_rent	Per_rent
1	1	6.8%	96.7%
2	2.01	2.2%	99.2%
3	2.02	3.8%	94.3%
4	3	6.1%	73.3%
5	4	5.9%	85.1%
6	5	6.4%	59.5%
7	6	2.7%	54.2%
8	7	4.3%	61.7%
9	8	10.5%	72.7%
10	9	7.4%	44.8%
11	10	13.7%	59.3%
12	11	3.8%	55.5%

In the Geoprocessing pane search for the ‘Join Field’ tool. Open the tool and enter the parameters. The Input table should be the boundary file, and the join table should be the attribute table of the CSV file. Specify the input join fields as shown in the figure below. Click Run.

Geoprocessing

Join Field

This tool modifies the input table.

Pending edits.

Parameters

Input Table: tomtracts

Input Join Field: NAME

Join Table: 2020census.csv

Join Table Field: ID

Transfer Fields: All Fields

Validate Join

Join spatial boundary with socio-economic information

Open the attribute table for our boundary files. If you scroll over, you’ll notice that the last few columns are exactly the socio-economic information we wanted. However, all the data is displayed as <NULL>, which indicates that our join has failed.

This is because, in our boundary layer, the data type for NAME is text (you can verify this yourself), while the column imported from the CSV file is of numeric data type. So, the unique identifier used in the join function not only **needs to have unique, non-null values but also matching data types**.

Now, let’s fix this problem.

- In the attribute table of the boundary layer, add a new field named CT_ID, set the data type to double, and the number format to numeric. Click the SAVE icon at the top.

- Go back to the attribute table, find the newly created CT_ID field, right-click -> calculate field, in the expression tab, double-click the NAME field. Click Apply and OK. We now have a newly created CT_ID field in double data format, which can perfectly match the ID field in the CSV file.

Now, open your boundary layer and scroll over to the last few columns, and you'll see that the socio-economic information has been matched.

After you complete the join operation, you must export the joined shapefile in order to permanently save your join. Export it as *Tompkins_CensusTract_2020* to your folder.

Map 6: Now create a map of percent population Black or African American and the percent of renters for Tompkins County, using whatever color classification scheme you wish. (10 points)

Part 3: Using Social Explorer to access census data:

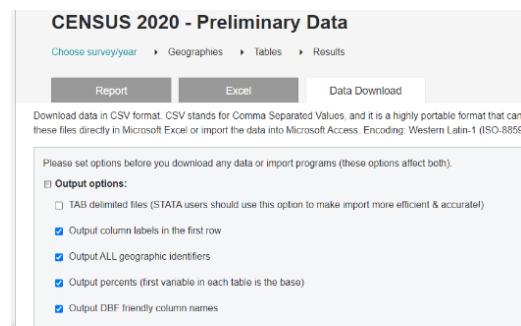
Both methods to retrieve census data above are useful and important to learn, but they are somewhat clunky or incomplete. A number of private vendors have been established to make the process of accessing, managing, and interpreting census data easier and generally more “gui-fied.” While Cornell has a license agreement for one of them (Social Explorer), many local governments, public agencies, community groups, Non-profits, and private firms may not, so it is important to understand how to access census data from different sources.

To access [Social Explorer](#), you are required to login using Cornell email and password. Social Explorer has readily available online maps that are quite easy to navigate (you can click on the ‘Explore Maps’ tab to do so).

Under ‘Tables’ on the left-hand side, you can specify a range of data products for download as an attribute table. For example, you can download the 2020 population from the decennial census from Social Explorer. Select U.S Decennial census, and the Census 2020-Preliminary (Begin Report).



Navigate to add all census tracts for Tompkins County as the Geography. There are not many variables available, but add ‘Total Population’. Select Data Download, and under output options select the following:



Download the CSV file, open the table and browse to the last column, you will find the percentage change of population (2010-2020) for each census tract has been calculated for you! Clean up the table as we have learned and join it to the 2020 boundary file for Tompkins County tracts.

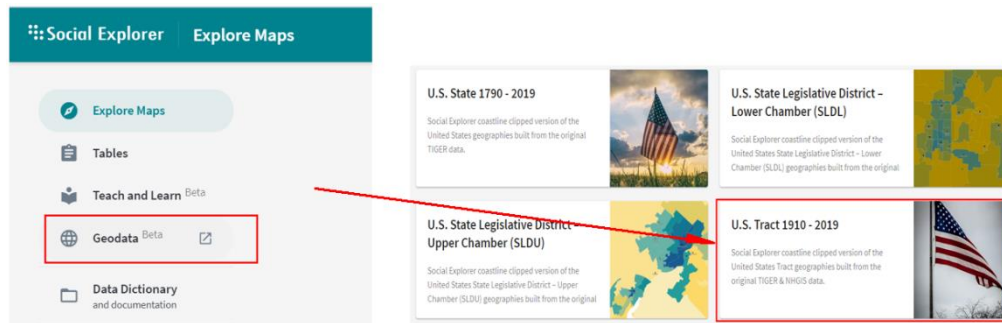
When cleaning the table, ensure that only the first row contains the column/field names. This means that

starting from the second row, it should record the information for each census tract.

Map 7: Create a Total Population Percent Change 2010 to 2020 map using census tracts for Tompkins County). Identify the data source you used (10 points).

Note: [here](#) is information on how to access and download data from Social Explorer

Please note: If you would like to download shapefiles from social explorer for your own project, click “Geodata” from the menu on the left side of the Homepage and take a look at the options. Note that they have many historical shapefiles as well. For example, browse to “U.S. Tract 1910-2019”, which stores all the historical boundaries at the census tract level. You have to download the entire country and use “Select by attribute” in ArcGIS Pro to select your target area.



Selecting Geospatial boundary file in Social Explorer

LAB 6 DELIVERABLES

Deliverables: *Note: All maps must have **a projection, data sources, and the classification system** (wherever applicable) should be mentioned. Also note that your data sources will no longer be Barclay Lab!*

1. Map of normalized Black population (2000) for Tompkins County with a classification of your choosing at the Census tract level (5 points)
2. Map of normalized Black population (2000) for Tompkins County with a classification of your choosing at the Block Group level (5 points)
3. Map of normalized Black population (2000) for Tompkins County with a classification of your choosing at the Block level (5 points)

Question #1: Discuss how differences in the unit of analysis affects spatial patterns (5 points)

4. Map of Black population (2000) per 1000 for Tompkins County with a classification of your choosing at the Census tract level. Note that the black population (2000) per 1000 should be in whole number in the legend. (10 points)
5. Map depicting only those census tracts with a majority (over 50%) of renter-occupied housing units. Include a context map as well. Make sure that each selected tract is labeled with its *tractid* value. (10 points)

6. Map of percent population Black or African American and the percent of renters for Tompkins County, using the 2020 Decennial census, using whatever color classification scheme you wish. (10 points)
7. Map 7: Create a Total Population Percent Change 2010 to 2020 map using census tracts for Tompkins County. Identify the data source you used (10 points).

Question #2: please include a discussion of the differences between the 5 year and 1 year ACS estimates, described [here](#) (10 points)