Lab #5 Census Data Lab

Note: Use Firefox NOT Internet Explorer for this lab. The U.S. Census website and Internet Explorer do not get along well.

In this lab, we're going to work with United States census data. **If there is one data set that is nearly universally applicable for GIS users, no matter what their subject focus (i.e. politics, environment, business), it's census data**. Almost every country has a census. We're going to be working with United States data because, well, that's where we are and where most of us are from, but most of the skills you learn in this section are likely transferrable to most other censuses (as well as many other data sets!)

First, a little bit about the United States census. It has been mandated by United States constitution since the beginning. The mandate appears in this particularly nasty passage in article I, section II:

> "Representatives and direct taxes shall be apportioned among the several states which may be included within this union, according to their respective numbers, which shall be determined by adding to the whole number of free persons, including those bound to service for a term of years, and excluding Indians not taxed, three fifths of all other Persons. The actual Enumeration shall be made within three years after the first meeting of the Congress of the United States, and within every subsequent term of ten years, in such manner as they shall by law direct."

Of course, the incredibly racist provisions about Native Americans and African Americans (as well as indentured servants) were removed after the civil war in the Fourteenth Amendment. That said, racism and racial ignorance pervaded the U.S. census for a long time thereafter. Until the mid-twentieth century, a person's race was determined by the census taker, who made his judgement entirely by the appearance of the person. Only in the 2000 census were people able to mark more than one race and ethnicity. Before then, folks who were of mixed ancestry had to choose one over the others!

The census is completed with two different forms. The "short form" of the census, whose results are coded into "Summary File 1" or "SF1", has a 67 percent final response rate, even though it is sent to nearly every household in the United States. This means that, based on this 67 percent sample, the census calculates the total *population statistic.* The "short form" contains only basic questions, such as inquiries about race, sex, age, ethnicity, housing units, and population counts. The long-form, coded in Summary File 3 or "SF3", is sent to 1/6 of households, and has much more detailed questions such as those about income, rent, kitchen facilities, poverty, ancestry, commutes, armed forces status, etc. The only subject not covered on the long form is religion. The census views collecting data about religion as violating the separation of church and state.

Census data is aggregated over two types of regions: legal and statistical. Legal regions include congressional districts, counties, city boundaries. The available statistical areas include metropolitan statistical areas (MSAs), census tracts, and block groups. I would argue that there are also other *functional* region types. ZIP codes, for example, do not perform an inherent legal or statistical function, but rather a postal one.

Now, given what you now know about the census, what do you think are some challenges in comparing 2000 census data with historical census data? Think about both attribute and spatial data (**Question 2.1: 5 points)**

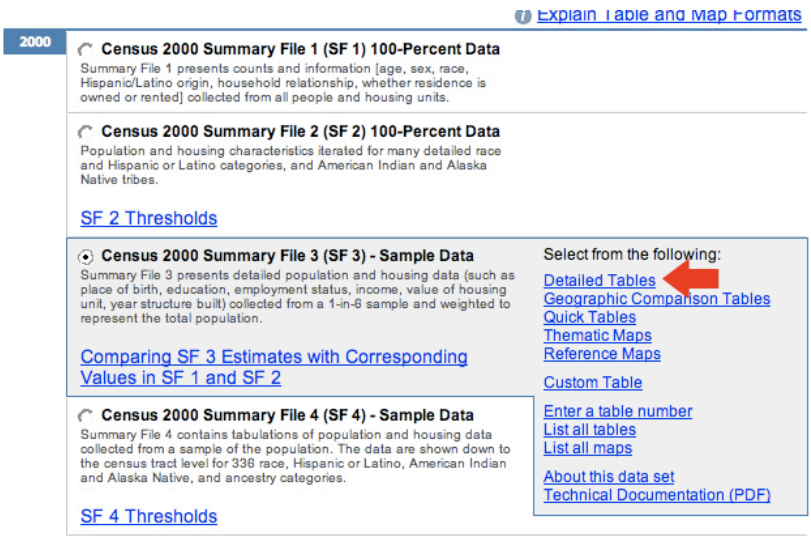
1. *Download U.S. Census Spatial Data (3-Digit Zip Code Areas) and Add It to ArcMap*

Go to http://www.census.gov/ and click on "Maps" under the "Geography" heading. Click on "Boundary Files" and then "Download Boundary Files". Click on "3-Digit ZIP Code Tabulation Areas (ZCTAs): 2000". Zip code regions are used frequently by businesses and marketers for research. Download the file in shapefile format. Shapefiles are usually directly below the "ARC/INFO Export .e00" format files. Choose the link next to "All 50 States, D.C., and Puerto Rico". Download and unzip the file in your "Lab2" folder. Add the shapefile to your new ArcMap project. Do not worry about the "Unknown Spatial Reference" dialog.

Did you note all the other boundary files available for download? These can be extremely useful. Be warned however, some of the other boundary files can be very hard to attach other census data to until you learn better how to use advanced join functionality. However, if you intend to manually enter data or enter data from another source, census boundary files can be a gold mine! (and since your taxes paid to prepare them, they're free!)

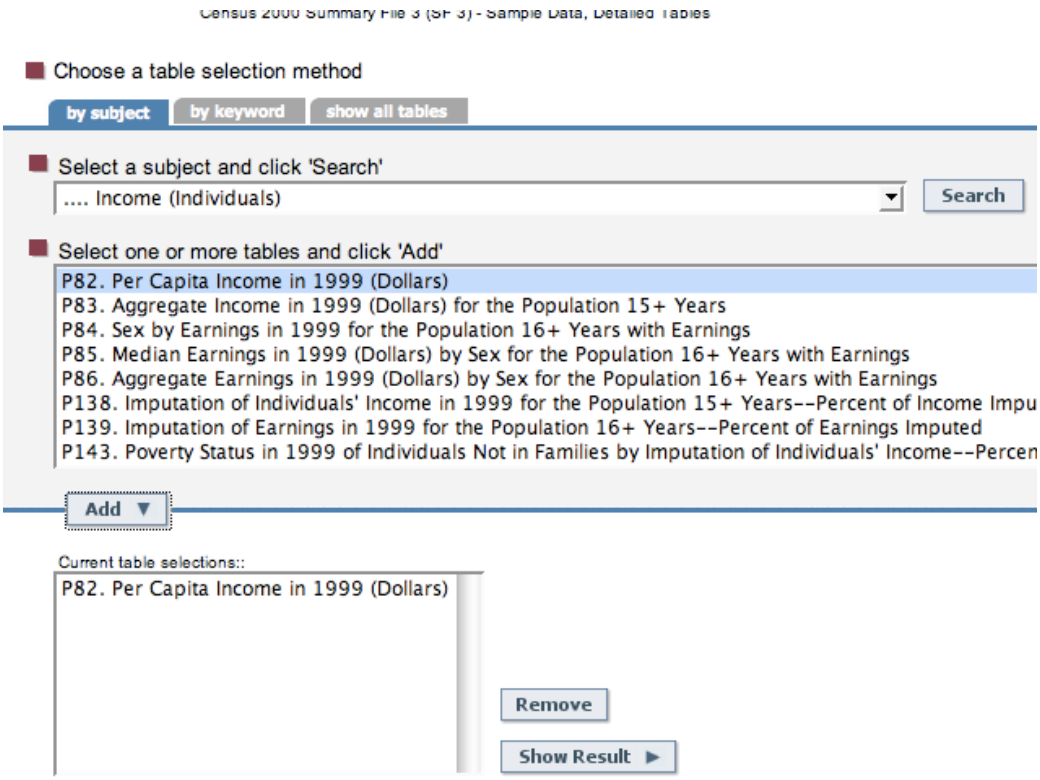
2. *Download U.S. Census Attribute Data and Add It To ArcMap*

Go back to the main census website at http://www.census.gov/. Click on "American Factfinder". In the American Factfinder, choose "Data Sets -> Decennial Census". Choose either SF 1 or SF 3 (SF 3 is better) and click on "Detailed Tables" (see figure below).



In the subsequent screen, choose “3-Digit ZIP Code Tabulation Area”, click on “All 3-Digit ZIP Code Tabulation Areas” and click “Add”. Finally, click “Next” when all the ZIP areas have been added to “Current geography selections”.

Here comes the fun part. In this next screen, you get to choose which attribute values you want to analyze spatially. I suggest clicking on the “by subject” tab so you can view the plethora of possible attributes in a more organized fashion. (Remember to hit the “Search” button after you choose something in the pop-up menu.) Choose two or three attribute fields and add them to the “Current table selections” by hitting the “Add” button. You can add attributes from any category, mixing and matching as you please. Click the “Show Result” button when you’re done. (see figure below for an idea of how this all should look).



A web-based table should come up next. Choose the “Print/Download” menu and hit “download”. Choose the “Microsoft Excel (.xls)” button near the bottom of the screen and hit “OK”. Save the file to your “Lab2” folder. Unzip the census data you just downloaded. Inside the resulting folder, you should see two Excel files. Open the file labeled “dt_dec_2000_sf3_u_data1.xls” in Excel.

We need to format this Excel worksheet so that ArcMap can understand it. Doing so used to be more difficult, but these days we simply need to make sure of the following:

1) that the top row contains a terse description of the column/field below it.
2) that there are NO “special” characters in the top row (i.e “+”, “-”), no spaces in the top row, and no titles in the top row that start with a number. Underscores (“_”) are allowed, as long as they are not the first character.
3) that there are no rows other than the first one that do not contain actual data for a spatial feature. Having such rows will seriously confuse ArcMap.

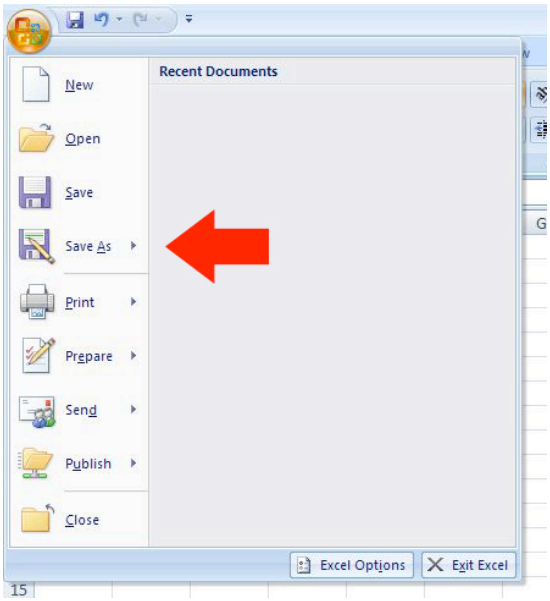
In this case, satisfying these two requirements amounts to changing the cryptic headings like “P082001”, etc. to something more meaningful, as well as deleting the second row. Here is an example of what the top rows of the should file look like before (see figure below):

GEO_ID	GEO_ID2	SUMLEVEL	GEO_NAME	P082001
Geography Identifier	Geography Identifier	Geographic Summary Level	Geography	Total population: Per capita income in 1999
85000US006	006	850	006 3-Digit ZCTA	6716

Here's what it should look like afterwards (see figure below):

GEO_ID	GEO_ID2	SUMLEVEL	GEO_NAME	AVG_INCOME
85000US006	006	850	006 3-Digit ZCTA	6716

Note that I have changed "P082001", which is described as "Total population: Per capita income in 1999" to "AVG_INCOME". "AVG_INCOME" is both descriptive and short. ArcGIS will not accept any column header greater than around ~15 characters. This is another holdover from the Chuck Norris era. In your "Lab2" folder, save your Excel file as a "Excel 97 - 2003 Workbook" NOT an "Excel Workbook file" file (see below image for a hint about how to do this) and name it "censusdata" or something like this.



Normal "Excel Workbook" files in Office 2007 are exclusive to Office 2007 and are not well-supported by other programs yet. You can't even open these files in older versions

of Excel without downloading special software. (Note: this is important to know both inside and outside of the context of GIS).

Now, we need to load the Excel file into ArcGIS. Click the “Add Data” button and navigate to the Excel workbook file. Double-click on the Excel file, DO NOT click “Add” when you have selected it. ArcMap does not support multiple worksheets (like Excel does), so ArcMap treats each worksheet as a file within an Excel workbook file. It doesn’t tell you this and will give a cryptic message if you try to add a workbook file. Choose the “Sheet0$” worksheet (ArcGIS adds a “$” to the end of every worksheet).

Support for Excel files is new and limited in ArcGIS, so we’re going to convert our table to a “.dbf” file in order to enable full support and functionality, and avoid any and all bugs that inherently occur in new ArcGIS features. We can do this by right-clicking on the table and choosing “Data->Export...” In the subsequent window, use the “Browse” button to navigate to your “Lab2” folder and save the exported data as “censusdata.dbf”. When ArcMap asks you if you want to add the table to the current map, choose “Yes”.

Finally, remove the Excel table from the map by right-clicking on it and choosing “Remove”.

4. ***Joining*** *the spatial and non-spatial data (in matrimony)*

In the first part of this lab, you learned about spatial joins. Now, you’re going to learn about the equally important database/attribute (a.k.a. non-spatial) join. Attribute joins are essential in many GIS data input tasks. They allow you to attach to spatial features data you find on the Internet, enter yourself in Excel, given to you by a coworker, or retrieved some other way.

In formal terms, a join works as follows: Let there be two tables, table *A* and table *B*. Each table *A* and *B* has a “key” or “id” field called *IDa* and *IDb*, as well as many other fields *FieldA1....10* and *FieldB1....10* respectively. If you perform a join of *B* onto *A*, the join operation will search for rows in *A* and *B* that have matching “key” or “id” fields and will *append* all of *FieldsB1...10* to table A for all matching rows. Martin has a wonderful graphic that describes this operation (see figure below and enjoy the German):

KatNr#	Kategorie
1	Bohrstock
2	GPS-Gerät
3	Topogr. Karte

ArtNr#	KatNr#	Artikel
A1	2	Garmin
A2	3	TK25
A3	3	DGK5

➔

KatNr#	Kategorie	ArtNr#	Artikel
2	GPS-Gerät	A1	Garmin
3	Topogr. Karte	A2	TK25
3	Topogr. Karte	A3	DGK5

Let's get a more concrete idea of what's going on. Open the attribute table of the ZIP code *spatial data* layer and look at the field called "NAME". Now, open the table for the "censusdata.dbf" file by right-clicking on it and choosing "Open" . Look at the field called "GEO_ID2". Although it may not seem so from their names, these fields are the same in each table. In other words, we can use this field to match the two tables. If a zip code polygon in the spatial data layer has a "NAME" of 931, we can match it with a row in the census data table that has the same value in "GEO_ID2" field and copy all the values in the census data table to the spatial data table. Doing a join operation does this for every row in the spatial data table.

Let's see this in action. Right click on the zip code spatial data file and choose "Joins and Relates->Join..." Make sure the top pop-up menu reads "Join attributes from a table". In step one, "Choose the field in this layer that the join will be based on:", choose the "NAME" field. In step two, indicate that you want to join the "censusdata" table to join to the spatial data layer. Finally, in step three, "Choose the field in the table to base the join on:", select "GEO_ID2", of course. Hit "OK".

Now, horiztonally scroll through the spatial data layer's attribute table. Look what's happened! It's like you've had this data in the table all along! The one difference is that each field is preceded by a prefix and a dot separator. The prefix tells you from which file the field originally came.

Make at least two choropleth maps of different variables that you downloaded from the census and export them to pdf, naming them Map2a.pdf and Map2b.pdf. Each are worth 15 points **(Map 2a, Map 2b: 15 points each.)**

Note: You can always remove joins by right-clicking on the spatial layer and choosing "Joins and Relates->Remove Joins(s)". Also, to make a join permanent, you have to

export the shapefile data by right-clicking on the spatial layer and choosing Data->Export Data...

5. *Learn more about joins from ArcGIS Desktop Help*

In the main menu, choose "Help->ArcGIS Desktop Help". In the resulting window, do a search for "Join" by choosing the "Search" tab. Find out about "one-to-one" relationships and "many-to-one" relationships, both of which work well for join operations. Describe these types of joins and how they might be used. **(Question 2.2: 5 points).**

6. *Think harder about joins*

Here's a tough question. What happens when a row in the spatial data layer has a value in the "key" field that does not occur in the "key" field of the census data table? Make a hypothesis and test it, either through documentation (i.e. the help file) or practice (i.e. seeing what happens). **(Question 2.3: 5 points)** *You already have all the skills you need to test your hypothesis through practice. Hint: Think about your newfound editing skills.*