USP/PLSI 492: Research Methods

Instructor: Pietro Calogero

Getting Census data and prepping it for use in GIS

The purpose of this exercise is to get some US Census data in a format that can be added into a geospatial database (in this case, a shapefile). After this procedure, we will get a Census TIGER shapefile that matches the geographic extent of our census data, and we will join this data to our shapefile.

1. Get the data.

Go to American FactFinder 2, Advanced tab, and set your filters to get data from the geographic area of your study, which whatever characteristics most closely correspond to your survey-design.

Select the table or tables that most closely correspond to the whole population (universe) that you are seeking to represent with survey data. Click the hyperlink of the table name itself so that you can preview the data.

Note: if you are getting ACS data, rather than straight-up decennial census data, the ACS data will be estimated values. And it will include margins of error, and percent margins of error for each data-point. That means that ACS data will show 4X the number of data-columns. To reduce this before downloading, click the “Modify Table” link in the Actions bar. Then, in the top left cell of the table itself, a small blue funnel will appear. Click on this funnel, and check only the “Estimates” column. This will hide the “Estimate Margin of Error”, “Percent”, and “Percent Margin of Error” columns, and exclude them from the data-download. That reduces the amount of data you will need to sift through by 75%.

Select “Download…” And then choose the .CSV option. That will be a text file with values separated by commas. **NOTE 1:** American FactFinder also offers an option to download the file in Excel format. That is convenient for a quickie, simple exercise. But the Census limits the size of the dataset you can download as an Excel spreadsheet. Worse, it strips off the geo-codes which you will need in order to import this data into ArcMap. So for many reasons, download as a .csv file, not an .xls file.

2. Prepare and arrange the data in Excel.

ArcMAP and QGIS can import .csv files. But the file you download from FactFinder will need to be cleaned up in two ways, which I call "winnowing" and "relabeling". Before we winnow and relabel, though, make sure you learn how to bring CSV data into Excel in a way that does not corrupt the data.

2a. Import a CSV into Excel without losing leading zeroes.

Importing CSV data into Excel is also a little tricky. If you open a FactFinder CSV file directly into Excel, it will strip off any leading zeros in any data. In California, the GEO.ID for census tracts begins with a zero, such as: 06075010100.

So: a) Open a **blank** Excel spreadsheet.
b) Under the Data tab > Get External Data subtab, Select "From text."
c) Choose the Path and file.
d) This will invoke the Data Import Wizard. Specify "Delimited" data, not fixed-width.
e) In the second pane, Choose "commas" as the separator and uncheck all the others.
f) In the third pane, specify the first three columns as Text (not General).
g) Finish the wizard. Check that the data imported cleanly.

2b. Winnowing the data: pulling the needle out of the haystack.

In the example I show in class, I download table DP04 for all the census tracts of San Francisco. For the row-data, I have eliminated Estimated margins of error, Percent, and Percent margins of error. But Table DP04 also includes more than 100 categories, including rates of occupancy, age of structure, number of rooms and bedrooms, year of occupant move-in, type of heating, presence of telephone, etc. I am going to use a tiny amount of this data. Likewise I expect that you will not be able to find data that exactly matches your research design, so you will need to grab "too much" data from FactFinder2 and in this step you will winnow out excess data. In my case, I will only keep columns A, B (geo-codes), C (census tract numbers), D (total units), CM (median house price), and EE (median gross rent).

Since I don't want to keep any of the data beyond EE, I select column EF by clicking the gray label of the column. Then I scroll to the end of the data—column EN. I hold down SHIFT and Left-click the gray column-label EN. That selects all the columns from EF to EN. Then I right-click the column label, and Delete. Delete columns CN to ED the same way, and likewise E through CL. And I delete column A.

This leaves only 5 columns of data: the GEO.id2, GEO.display-label, HC01_VC03, HC01_VC125, and HC01_VC185. Those headings in Row 1 are pretty obscure, and the explanation is in Row 2. In the final preparation-step we will change this into something more useful for us.

One more thing: Excel creates three-sheet Workbooks by default. Just to reduce confusion later, go to the bottom left of the window-border select Sheet2, R-click, Delete; and Sheet3, R-click, Delete. You can eve rename the Sheet1 as "Data" or something to remind you of which sheet to look for as you import this into other programs.

2c. Re-labeling the headers

In this third Excel step we need to do two things: edit the header-row (Row 1) so that ArcMAP can read the data, and also edit the header labels so that we ourselves can understand what data we are looking at.

To make the headers readable by ArcMAP, any punctuation-marks must be removed (except under_score). Rename “GEO.id2” to “GEOid2”. Rename “GEO.display-label” to “GEOdisplaylabel”.

To make the remaining three headers human-readable, rename them using the information in the cell below. I renamed them “totalunits”, “medvalue10”, and “medrent10”. I don’t use punctuation marks, I don’t use spaces in my “dehydrated” labeling. But I try to create the shortest possible name that I can intuitively understand.

Finally, I delete Row 2. Now, this table only has one header row (Row 1) and the rest of the rows are data-records for each census tract in San Francisco for 2010.

2d. Saving the Excel file

Save as an .xls file. If your computer system is hiding file-suffixes from you, then choose to Save As “Excel 97-2003 Workbook” format. That is .XLS format, and it predates the .XLSX format. ArcMAP10.0 can read .xls format, but we are not sure if it can read the newer .xlsx format, so to be conservative and maintain maximum file compatibility, let us be conservative and use the older, more widely-understood file format. Since we are going to join this file into another database, I like to choose a very short filename. I called this “housing10.xls”