Appendix B: Tutorial-Creating a State Interest Group Variable from IRS Business Master Files

Data Source:
NCCS Data Archive: http://nccs-data.urban.org/data.php?ds=bmf

Before you begin, you should identify relevant organization names through relevant publications such as the *Encyclopedia of Associations*, or through common key words used in the name. You will search the IRS Business Master Files for these names. You should also familiarize yourself with the National Taxonomy of Exempt Entities (NTEE) and the codes used to classify organizations. See http://nccs.urban.org/classification/national-taxonomy-exempt-entities. Also, review the full list of codes http://nccs.urban.org/sites/all/nccs-archive/kbfiles/324/NTEE%20Two%20Page_2005.DOC.[1] You will search for relevant NTEE codes in the IRS Business Master Files. For our research, we used P88 (LGBT Centers), R26 (Lesbian and Gay Rights), G81 (AIDS) and H81 (AIDS Research). As noted in Appendix A, we also searched for a large number of related key words.

1. See IRS Business Master Files http://nccs-data.urban.org/data.php?ds=bmf
2. Select a start year and download the last csv file for the year. Note that these files are very large given the number of observations.
3. Import that file to Stata or the statistical or database package of your choice. You will get a large number of variables. The most important ones for this exercise are employment identification number (ein), name, state, NTEE code (nteecc), assets and income.
4. Generate an identifier variable so that you select relevant organizations. In Stata, this is *gen identifier=0*
5. Replace the value of the identifier variable to 1 for relevant organizations. You can do this through text searches of the organization name variable (name). In Stata, this is *replace identifier=1 if strpos(name, "SEARCH STRING")*. Note that the name variable uses capital letters. Repeat this for all relevant name searches. In all likelihood, you will pick up unrelated organizations. For instance, we searched on the term "gay" but also got hits on "gaylord." You will be cleaning this data later.
6. Replace the value of the identifier variable=1 through text searches of the relevant NTEE classification code variable (nteecc). In Stata, this is *replace identifier=1 if strpos(nteecc, "SEARCH STRING")*. Note that the nteecc variable uses capital letters. Repeat this for all relevant NTEE code searches.
7. Drop observations where the identifier variable =0. In Stata, *drop if identifier ==0*
8. Generate a year variable that denotes the origin of the IRS Business Masterfile. In Stata, *gen year=XXXX*
9. Save your file and repeat the process for the next year of IRS Business Master Files. Continue this process until you have sufficient observation years.
10. Combine all of your datasets. In Stata, you can "append" the data. See *data>combine datasets>append*. Copy and pasting observations is also possible depending on the size of

[1] Note that some users may have trouble connecting to this document directly due to security settings on their computer. It can be found indirectly at http://nccs.urban.org/classification/national-taxonomy-exempt-entities and choose the link for "Full list of NTEE codes."

your dataset. Be sure to save the dataset as a new name. You may desire or need to covert variables such as the employer identification number (ein) or other variables from string to numeric data. In Stata, this command is *destring* and it can be found under Data >Create or change data> Other variable-transformation commands.

11. Sort your data by employer identification number (ein) and year. In Stata, *sort ein year*.
12. You must now clean your data to remove unrelated organizations that were picked up in your search of the name field. You should review these manually and this should be done after you have combined datasets because you can drop multiple observations from the same organization at a time. While there will be some obvious organizations to remove, this will take a substantial amount of work. It might require looking at tax forms held by Guidestar or a similar organization. Web searches on the organization are also a useful strategy. You might find it helpful to a create a drop variable to identify observations to be dropped. In Stata, *gen drop=0*. Reorder the drop and year variables to facilitate data cleaning. In Stata, *order year drop, after(ein).* Change the value of *drop* to 1 for each observation to be dropped. After all observations have been reviewed save the file and then remove the offending observations. In Stata, *drop if drop==1*. Save the file under a new name.
13. After all data cleaning, you can collapse your data by state and year. This will allow you for instance to look at the combined income or assets in a state for a given year. In Stata, *collapse (sum) income assets, by (state year).* Save the file under a new name.
14. Sort the collapsed data by year and state. In Stata, *sort year state*. Because you are using financial data over time, you must put this nominal data in real terms. To accomplish this, obtain a price index from the Bureau of Economic Analysis (BEA) or Bureau of Labor Statistics (BLS). In our analysis, we used the annualized Price Indexes for Gross Domestic Product that is produced by BEA (see https://www.bea.gov/iTable/index_nipa.cfm). This price index looks at all goods and services produced domestically (imports not included) rather than just things bought by consumers (as is the case in the Consumer Price Index computed by BLS). Things like lobbying expenses, professional services, office furniture, donor tracking software and the like would not be in BLS's Consumer Price Index. After navigating to the BEA link, choose "Begin using the data" and then select "Section 1-Domestic Product and Income." We used Table 1.1.4 Price Indexes for Gross Domestic Product. Be sure to modify the table (see the upper right corner of the webpage) with the desired year range and select "annual" for the series. Download your data in Excel or CSV format. In your Stata dataset, all states should have the same price index for a given year but this price index will vary across years. Add this variable (priceindex) to your dataset by carefully copying and pasting the data from Excel. The basic formula for converting nominal to real is real= nominal/(priceindex/100). Generate a new variable (realincome) based on this formula for income. In Stata, *gen realincome=income/(priceindex/100).* Generate a similar variable the formula for real assets.
15. Obtain state population data from the Census Bureau for each year and state (see https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml). This variable should not be held constant over time because your revenues and assets are not held constant. For efficiency, one could interpolate population estimates between Decennial Census or American Community Survey years. Note that Census Bureau's data for the 1990 Census is often plagued with outdated links. One source that can give 1990 population estimates to use for interpolation between 1990 and 2000 can be found at

https://www.census.gov/population/www/censusdata/pop1790-1990.html. Download this data into an Excel or CSV format. To interpolate between years, take the state population in the newer year and subtract the state population from the earlier year. This provides the total state population change between years. Then, divide total state population change by the number of periods between the newer and earlier years. Add that value to the state population for the earlier year and then do it for each intervening year between the earlier and newer years.

16. Add this variable, *population* to your dataset for each state/year by copying and pasting or other process.
17. Divide your real income and real assets variables by state population so that you have real per capita income and assets for each state over time. In Stata, *gen realincomepercap=realincome/population*. Generate a similar variable for real per capita assets.
18. Congratulations. You now have real per capita income and asset data for state interest groups over time.