D. Gupta

Assignment I: Comparing Ethnic Conflict Datasets

Assignment Purpose

As you have already noticed from our readings, studies of ethnic conflict increasingly draw on several well-known datasets like the Correlates of War or Minorities at Risk. These large datasets allow us to examine trends in conflict onset and duration across many different countries and time periods, making it possible to form generalized statements about how and when and where conflict is likely to happen.

At the same time, these datasets can be critiqued for taking highly complicated concepts and assigning them simple numerical values—a process that is never truly objective, but instead requires scholars to make decisions about how to capture qualitative variables in quantitative form. Because political scientists often deal with large, abstract, and hard-to-define variables like "democracy" or "ethnicity," we have to be aware of how these datasets code such variables, how datasets might differ in their coding, and how these differences might affect the inferences that we draw.

Instructions

1. Go to the course website and open the "Datasets" folder. There, you will find links to several datasets that are commonly used in studies of ethnic conflict.

2. Pick two datasets from this list. Use the codebooks to explore how the datasets were put together, the geographic/period coverage, and the data sources. Note differences in the structure and focus of the dataset and think about why these differences might exist (e.g., audience, focus, author's affiliation or intent, etc.)

3. Next, use the codebooks to identify a variable related to "ethnicity" OR "democracy" OR "ethnic war" that exists in both datasets. You may find that there are multiple measures related to these concepts in both datasets. Use the codebooks to decide which of these variables are most comparable. For example, "ethnicity" could be measured by an ethnic fractionalization index or the number of ethnic groups or the size of the second largest ethnic group as a percent of the population. It is your responsibility to find variables that "match" as closely as possible. NOTE: not all the datasets will contain variables related to all three concepts.

4. Once you have selected your variables, use the codebooks to compare how they were defined and measured. Note any differences and think about why these differences might exist and the advantages/disadvantages to constructing the variable in a particular way.

5. Next, use a statistics program (Stata and SPSS are recommended, but Excel will work too) to dig a bit deeper into the data:

   - How many observations are there for this variable? What is the geographic range and, for time-series datasets, the chronological range? Are there missing values? How many? Are these random or is there a pattern to the missing values?

- What are the minimum and maximum values of the variable in each dataset? The mean? Are the data points tightly clustered or dispersed?
- Using a histogram, check whether the data are normally distributed. Are the histograms for the two variables similar?
- Select some observations that appear in both datasets (make sure you match for both country and year if both are specified). Check the level of correlation between the two sets of variables. How would you interpret the results? Plot the observations from one dataset against the other; are there observations that deviate significantly from the diagonal? Is there a systematic pattern to the deviation (e.g., are certain types of cases off the diagonal)? If yes, what might account for this?

6. Write up your findings in a 5-6 page (double-spaced) paper. In your write-up, discuss the similarities and differences in your datasets and the specific variable you have chosen. Include the results of your exploratory data analysis and provide summary statistics and graphs where appropriate (all papers should include at least one table of descriptive statistics and a histogram for each dataset, properly labeled and referenced in the discussion itself). Your paper should also answer the following questions:

    - What are the strengths and weaknesses of these two datasets?
    - Does one do a better job of measuring ethnicity, democracy, or ethnic war? Why or why not?
    - If you do think one dataset is superior, are there cases for which the other dataset might be more appropriate?