Narrative

Madeleine Stevens

12/8/2019

Brief substantive background / goal

I intended to merge three datasets and collect my own data in order to address the following questions: does the proportion of female combatants in an armed group impact violence against civilians, particularly sexual violence? Additionally, does groups' publicity vary based on proportion of female membership and levels of violence against civilians? I intended for this project to use many of the skills we learned in the course (merging data, visualizing data, collecting data from the web, and perhaps some basic statistics) as well as some new skills (geospatial and potentially interactive charts).

To explore this question with existing data, I intended to merge the Women in Armed Rebellion Dataset (WARD, 1964-2014), the Sexual Violence in Conflict dataset (SVAC, 1989-2015), and the UCDP One-sided Violence Dataset (1989-2018). The WARD is organized by actor, while the SVAC and the UCDP datasets are organized by actor-year, so it was necessary to create compatibility between them. However, both the WARD and the SVAC dataset are offshoots of the broader UCDP dataset, so there are several variables that made appropriate keys. To address the publicity element, I intended to search the New York Times API for the groups in the merged dataset. After I successfully merged the data, I intended to create both descriptive charts and charts that visualize the relationships between my main variables of interest. I intended for that to involve geospatial data, but also potentially interactive maps of the relationship between some of the variables.

My intentions when beginning this project were very ambitious, so ultimately I scaled some things back. For instance, given I know the tiniest bit of statistics, I opted not to run any statistical models on my data. I instead opted to put my efforts into visualizing aspects of the data. Additionally, I found that in order to do what I was interested in doing with interactive maps, it was necessary to have data that was more geospatially detailed, so I added the UCDP Georeferenced Event Dataset to my project. I also opted against using the New York Times API for three reasons: 1) group names are frequently too long for headlines, 2) group names are often spelled in a myriad of different ways, and 3) it seemed like a tacked on part of the project that was not really related to the rest of what I was interested in.

The biggest challenges were related to saving the interactive maps I created in a way that would enable me to use them in my presentation. Unfortunately, my computer was not up to the task. Despite successfully using the "mapshot" function on my smallest interactive map and saving it as an HTML file, trying the same things on the larger files continually dramatically slowed down my computer and froze R, forcing me to quit and lose my changes. Instead, I have done an unconventional workaround: I video-captured my screen as I interacted with each map I created and have embedded the resulting videos into my powerpoint.

Collecting data

In order to explore this topic, I first merged the Sexual Violence in Armed Conflict dataset (SVAC, 1989-2015), and the UCDP One-sided Violence Dataset (1989-2018). Later, I added in the UCDP Georeferenced Event Dataset (UCDP GED, 1988-2018, minus Syria) and the Women in Armed Rebellion Dataset (WARD, 1964-2014).

All four datasets are easily downloadable from their authors' websites (links in documentation).

Cleaning / pre-processing data

First, I filtered both the SVAC and the UCDP one-sided violence datasets so that I only had information on nonstate actors, as I am not interested in state violence at this time. Then, I made them compatible time-wise by limiting the UCDP data, which originally went from 1989 to 2018, to the same range as the SVAC (1989 to 2015). I also made the region variable in the UCDP dataset numeric to match the SVAC. Then I merged the two with the following keys: Actor ID, Actor Name, Year, Region, and Location. I then replaced all -99 values (what the SVAC uses to indicate there is no information) with NAs. I also went through and made sure that all other ways of saying NA (n/a, etc) were transferred to NA. I also then dropped several irrelevant or superfluous columns: I dropped "version" because I said what version of the UCDP data I used in my README, I dropped the "interm" and "postc" columns (from the SVAC dataset) because I already opted to only use conflict years, and I dropped the "is_government_actor" column because I already filtered for nonstate actors. Additionally, I opted to use the UCDP location data, so I dropped superfluous columns "gwnoloc", "gwnoloc2", "gwnoloc3", and "gwnoloc4" (from the SVAC).

I then renamed certain columns in the merged dataset in order to be more informative. The "type" column from the SVAC, which says what type of conflict it is, became "conflict_type", "incomp" from the SVAC, which says what the conflict was over, became "conflict_issue", and "form" from the SVAC, which says what form of sexual violence was perpetrated, became "form_sv". I also replaced region numbers in the dataframe with region names, to make plotting easier. I also made an actor-year column for this same reason.

Next, I created an actor dataframe out of my actor-year dataframe. I also added a "sexual violence prevalence" variable made up of the combined scores across the State Department, Amnesty International, and Human Rights Watch. Here, I ran into the problem that transnational actors (the Islamic State and al-Qaeda) appeared multiple times in the dataset, but decided that for visualization purposes (rather than statistical purposes) this wasn't a huge problem.

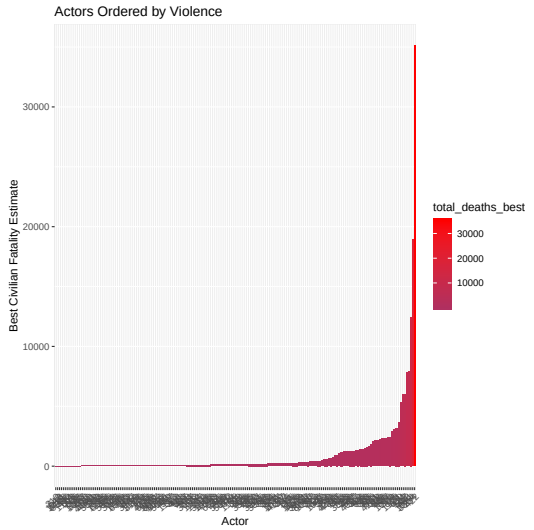
Next, I loaded data from the WARD and combined it with my merged, actor dataset, as it is also actor data rather than actor-year data. I then filtered this dataset so that it only included actors for which there was information on their levels of female members. I opted to use the "cat4_prevalence_high" estimate to maximize the number of groups in this new dataset. It is a high estimate of the number of female group members (not exclusive to combatants).

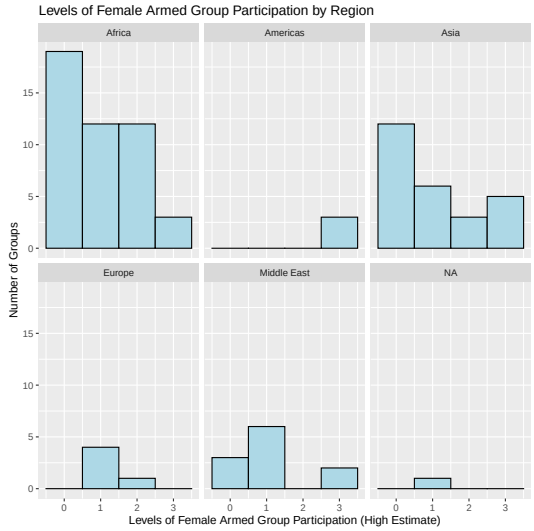
Finally, I moved on to preparing the data for maps. I decided to limit my data to the top five most violent groups (by the UCDP one-sided violence dataset's best_fatality_estimate variable) by region for ease of mapping (and so my computer wouldn't explode from the effort). I created a function that extracted the top 5 most violent groups from each region, and then merged all of them into a dataframe. I then loaded the UCDP GED, filtered for violence against civilians, and merged it with the top5 dataset. I unfortunately had to drop the Syrian insurgents from the merged data because the UCDP GED does not contain data on Syria (presumably because the ongoing war involves so many little groups and so much violence), so I ended up with 4 groups for the Middle East and 5 for everywhere else.

I additionally made one dataset just of geocoded data on the Middle East for the fun of making a regional map.

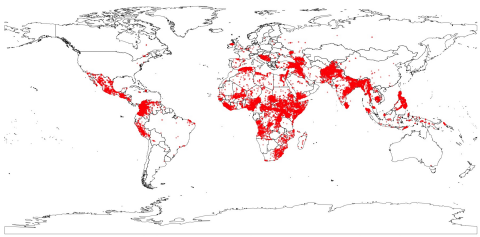
Analysis and visualization

I used the various datasets I created to make a series of bar charts visualizing different aspects of the data. For instance, I found the top 10 most violent actor-years, figured out how many groups each region had, and mapped group violence by best fataliies estimate across every group in the dataset. While that graph is basically unreadable, it demonstrates something interesting: there are a few groups with death counts in the thousands, but most groups just cause a few hundred (or less) fatalities here and there. I also made histograms about what percentage of the groups' membership was women, and faceted it by region. Africa had the most robust data on this, which was not terribly surprising because Africa also has the most groups by far.





I then made several maps. One was a static map of all of the event data worldwide, which is pretty much unreadable, but I provided it anyway as evidence of my learning process. The other static map was just of the events in the Middle East, and was more readable. Then, I created three interactive maps: one of the top four groups by civilian fatalities in the Middle East, one of all violent events perpetrated on civilians by the top 5 armed nonstate actors in each region (colored by actor), and one of all violent events perpetrated on civilians by the top 5 armed nonstate actors in each region (colored and sized by number of fatalities).



Future work

Upon collecting the necessary information for the README file, I became aware of two facts: 1) the version of the SVAC dataset I used had a flaw in it that has since been corrected by its creators, and a new version was posted as of this month. Unfortunately, it was posted too late for me to use the updated dataset in my analyses. 2) there is a geocoded version of the SVAC dataset, which I did not know about until the day I wrote the README file (12/8/19). While I regret not having the time to engage with either the corrected SVAC dataset or the geocoded SVAC dataset due to my other class obligations (12/9 I have a mock comprehensive exam so 12/8 is the last day I can do anything big on this project), I hope to use them in future projects.

Additionally, in the future I hope to conduct statistical analyses related to this topic. While I began with questions concerning the way the presence of female combatants in nonstate armed groups influences violence - particularly sexual violence - against civilians, the available data is limited. The WARD contains far fewer groups than any of the other datasets I used, and there is not data available for every group in that dataset. Likewise, there are a lot of groups in the SVAC for which there is no data concerning sexual violence prevalence, and the data that is present is based on State Department and NGO reports, which is less than ideal. Particularly NGO reports are often constructed in a way that will gain them the most publicity for the cause, which calls the data into question. However, I do think this is the best way to get at this data, because it is such a traumatic and sensitive subject that there is no real way to survey survivors accurately. This is just to illustrate a limitation.

In order to conduct statistical analyses related to this topic, I also need to take more classes on statistics and the use of R for such work. While I aspire to have the methodological skills for this activity, I do not have them yet.

An additional avenue for future work relates to a new dataset based on the SVAC dataset that is being produced by Elisabeth Wood. It will contain more detailed information on the types of sexual violence used in different conflicts and by different armed groups. When I spoke with her about it, she indicated to me that it is not meant to be used to run regressions or anything like that (the n is likely too small), but it could be a worthy source for case selection on this topic, as well as data visualization.