Poster Project

Ignacio Cabezudo

December 13, 2017

Project Topic and Relevance (~ 300 words) Describe what interested you in the project you've chosen to work on this quarter. What is sociologically relevant about the topic you've chosen and why do you think this is an important area of investigation?

The purpose of this study is to search for a correlation between causalities caused by terrorists and economic growth. The region of focus is the Lake Chad Basin which lies at the meeting point of the boarders of 4 countries: Cameroon, Chad, Niger, and Nigeria. This region was chosen because it has been under siege by the terrorist group Boko Haram since the summer of 2009. After their initial major attacks, they have continued to grow and build power in the region. More than just responsible for mass-death and destruction, they are also responsible for the displacement of more than two million people. The latest numbers reported by the United Nations show Boko Haram has killed more than 35,000 people in total.

I believe that his is sociologically relevant as the field of the sociology of terrorism is relatively knew and has many avenues of exploration. By looking for, and studying the macro-effects of terrorism, a better understanding of how people are impacted, not only by specific events, but the on-going impact of terrorist activities. Specifically, exploring how terrorism impacts developing countries might help us create better systems by which to respond to, and manage the effects of terrorism.

Finding Data (~ 350 words) Describe how you found the data used for your project. This should include a discussion of the resources you used and the reasons you selected the dataset you did. If you changed datasets at any point in the quarter explain why you made the decision to switch and what effect(s) you think this had on your final analysis. What theoretical questions did you set out to answer with your data? How did these change as the quarter progressed?

To accurately capture the intensity of terrorist impact in the region, data from the Global Terrorism Database maintained by the Study of Terrorism and Response to Terrorism (START) at the University of Maryland was used. The total casualty number used is derived from the sum of victims who were killed and those wounded by terrorist attacks between the years 2000 and 2016. The economic indicator used is the real gross domestic product (RGDP) for each country in the time period taken from the International Monetary Fund's October 2017 World Economic Outlook.

Originally, the study was using the Armed Conflict Location & Event Data Project (ACLED) database was being used. The reason for the switch from one to another, was the nature of the database, and the specific question being asked. The Global Terrorism database ONLY includes data from events deemed terrorist attacks using 5 of 7 specific criteria. While the ACLED does include some terror attacks, the methodology used was not in line with what I was searching for.

Data Structure & Data Munging (~ 500 words) Reflect on the original structure of your data. Explain the ways in which you had to clean, transform, or merge the data to do the following: • Initially organize your data • Create meaningful visualizations • Run statistical analyses What challenges did you encounter in any or all of the aforementioned activities? Did anything surprise you about this aspect of research? Explain.

The Global Terrorism Database (GTD) consisted of 170,350 terror events, each containing a possible 135 variables. After agreeing to their terms of use, and downloading it, the database was loaded using read_csv. It contained several parsing errors. The initial organization phase of the project was more of a study phase, using the codebook provided to gain an understanding of what all the available data meant. After doing this, I decided that the nkill variable was one of the most important factors to highlight about the database. The rationale behind this seems obvious enough: the purpose of terrorist events is to terrorize a population, what could be more terrifying than a measure of how much death was caused by an attack.

Because the variable was already a quantification of death in itself, I had to find a way to show what the quantification means on a macro level. My previous knowledge of the Lake Chad region of Africa told me that searching for data about the impact of terrorism there would yield results that should be on going and extremely impactful. It is for this reason that I decided to begin by filtering events that occurred in Cameroon, Chad, Niger, and Nigeria.

The data gathered by this database is done so through a heavy vetting process that uses two automated searches and filter, then a manual read by a researcher. Even still, there are many accounts of terrorism that are lacking in total impact numbers. In an effort to keep the event count accurate, all NA's were replaced with 0's. Finally, the data was broken down into just 3 variables: Year, NKill & NWound. These three variables were then mutated to create TOTALCAS, which gives a single count of casualties, per year. This was done as it was the best way to relate the data to the IMF's Real GDP Data.

The Real GDP data was gathered from their database, which was last updated October, 2017. It was downloaded in Excel format and imported to R as such. Because of its native formatting, it was grouped by year and country and finally, merged with the data cleaned from the GTD. The process to do this took several weeks and was very confusing for me. I found it to be one of the most frustrating parts of the entire project.

Finally, linear regressions were run on the RGDP & TOTALCAS variables, as well as TOTALCAS & Year and RGDP & Year. Looking back, I believe I was chasing the shape of the RGDP curve, as I knew that when graphed, the nkill data had a similar structure. I was surprised that there was no statistically significant correlation.

Reproducible Research (~ 350 words) As you understand it, what is the value of reproducibility for research? Include a description of the ways you employed principles of reproducibility in your own project. Which techniques seem most useful (i.e. what might you want to learn more about/incorporate into your research methods going forward)?

I believe that reproducibility should be the keystone component of any research. Without it, the integrity of the entire research process is put in jeopardy. While working on this project I started with databases that are both free, and publicly available. I do understand the importance of data collection and creation of your own data, however in the context of this project, it was neither feasible nor reasonable to do so. Backups of the code were made often, and GIT was used as a method of storing a fully reproducible project. Finally, packrat was employed in an attempt to make the project portable.

The mistake that I made was in doing my work on a machine I do not have direct control over. The bulk of work was done on the CSDE server, which went through a full operating system upgrade during the quarter. Despite numerous backups, data was still lost as I misunderstood the staggered upgrade schedule. The decision was made to move all of the data to the university's shared drive, however this caused numerous problems. First and foremost, the nature of packrat makes accessing anything in the project that is on a shared drive with high network traffic extremely difficult. Everything takes longer, package installs started to fail, packages started disappearing between saves. Everything went wrong.

In the end, the entire project was migrated to a laptop and finished there. The final GIT pushes were made by transferring files to the CSDE server AFTER they had been knitted on the laptop. I cannot stress enough, how much of a mistake it was to use the CSDE server for a quarter long project. The ability to start a new project from an existing GIT was a life-saver. Without GIT, I might not have had any bit of hope left in me to attempt to finish the quarter. Packrat, I could probably do without. I can't think of a time when I'm going to need to work in R without an internet connection. I'll just go ahead and install required packages as needed in the future.

Finally, I need to do more work on understanding how to best utilize the statistical methods available to me in R. I'm simply not knowledgeable enough to even begin to really get what I can out of the language. Moving forward I plan on exploring the GTD in more depth, I am not convinced that there is no measurable correlation between the RGDP of this region and terrorism. It is much more likely that user error is at fault, and I'd like to solve that