Individual Assignment Report

Mathilda Ahr

17th January 2020

Contents

1 Introduction

This report will answer two research questions. It will first find what the total number of fatalities was in Indonesia in year 2004. Then, to build on this it will also look into how the number of fatalities evolved from 1989 to 2015 in Indonesia ?

This will be done by using a the most aggregated data (Croicu & Sundberg, 2015) from The Uppsala Conflict Data Program. This data set is in the json format and contains data for all years since 1989 for almost every country. Before conducting any coding and analysis, the data set will be explored through the shell.

After having filtered the file for Indonesia and in parallel for Indonesia specifically for year 2004, the first big step will be to make them into two csv files with tidy data that would be much easier to work with in python and R. This step will be done through python. The csv files should have each line representing the information for one observation of violent event and each column representing one characteristic or variable about the events. When creating the two csv files, only the relevant variables for the research will be kept: id, year, conflict_name, country, date_start, best (estimation of fatalities). These tidy data sets will then be opened in R. R will be used to count the total number fatalities for year 2004 and make a bar-graph displaying the fatalities for different conflicts for the first question. For the second questions, it will be used to count fatalities for all years since 1989 and plot them in a bar-graph.

2 Methods

2.1 Exploring data

This sections displays the first steps of the exploration of the data conducted through the shell.The data set used for this research is provided in json format. It is a list of dictionaries where each is indicated by curly brackets within square brackets. There are 377 lines in the sample data (bash: 'wc UCDP_conflict_sample.json'). Reading the first few lines enabled me to find a few characteristics of the first event:

- "country": events took place in Turkey
- "best": the best estimate number of fatalities is 33
- "type_of_violence": type of violence 1 which is state-based conflict
- "year": took place in 1997

This step enabled me to understand the structure of the data set: it is composed of numerous dictionaries that are each composed of 48 keys. The 4 most relevant keys for this research are "country", "best", "type_of_violence" and "year".

2.2 Choosing country and year

When the format of the data set was clear, I used the command 'grep' and 'wc' in order to pick a country and a year with a sufficient data quantity. The manual for each command (e.g. 'man wc') helped checking what the different numbers given by 'wc' represented.

I first selected the occurrences of violence in Indonesia recorded in the data set and counted them using this command: grep '"country": "Indonesia"' conflict_data_full_lined.json — wc. There are 1615 cases of violence for Indonesia since 1989 in the data set. I considered that 1615 cases of violence in a span of 36 (average of 44 events per year) years was enough to answer both research questions.

I then added the condition that the events had taken place in the year 2004 using the command: grep '"country": "Indonesia"' conflict_data_full_lined.json — grep '"year": 2004' — wc. There are 315 entries for Indonesia for the year 2004 which is also enough to answer the first research question that related specifically to this year.

2.3 Data pre-processing

The first step in the data pre-processing was to create to two subfiles with for the first one the data only for Indonesia and for the second one the data specifically for Indonesia in 2004. The first and second new file contain 1615 ad 315 dictionaries respectively that each represent the cases of violence for Indonesia and those for Indonesia in 2004.

The second step to be taken was to convert the smaller json files into csv files that could be read and be interpreted by R to ultimately make plots. Loading and reading the json files was done through python. Because it was best to keep the files intact, I created two csv files that would become the csv files to conduct the data analysis. In the python code I made it type the name of the columns that I wanted: id, year, conflict_name, country, date_start, best as the first line. Then, I created a fore-loop that would go through each dictionary and print as a string the variables listed above. I also had to add "backslash n" symbol in order for the values of each dictionary to be printed on separate lines. These steps resulted into two tidy csv data sets with 1615 and 315 entries and both with 6 variables that were ready to be directed to R to be analysed.

2.4 Data analysis

In a R notebook in RStudio I first downloaded the tidyverse, read the second csv file (for Indonesia in 2004) and put it into a variable. I then created a new variable in which I summarized the data set by summing all fatalities in one entry and calculating the average fatalities per day in the second entry. This table is displayed in Table 1. I then created a table with the number of fatalities for each different conflicts (with conflict name). Lastly, by using ggplot I created a bar graph that displayed the number of fatalities for each conflict see Figure 1. Due to very long conflicts names, the labels of the graph were overlapping and made them illegible. I managed bypassing this issue by both shortening the conflict names in a new column and by putting an angle to the labels.

In order to answer the second research question, I made R read the csv file that contained all observations for Indonesia. Instead of grouping the data by name of conflict, I grouped it by year and summarize it in a table with the sum of fatalities. The last step was to use ggplot to plot the results of the table in a bar graph that can be found in Figure 2.

Here is the link to my GitHub repository:
https://github.com/TillyAhr/Individual-Assignment-GITHUB.git

3 Results

4 Discussion

The number of fatalities in Indonesia in the year 2004 is as high as 925, this is equivalent to an average of 2,5 deaths per day. Those fatalities are divided

Total fatalities	Average fatalities per day
925	2,5342

Table 1: Total and average number of fatalities for year 2004

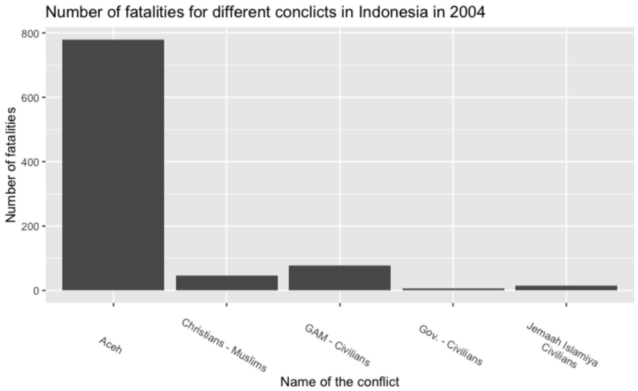

Figure 1: Fatalities per conflict

between five different conflicts. The conflict called 'Aceh' is the one that is the cause of the greatest majority of fatalities: 780 out of 925. It is classified as a type of violence 1, that is, a state-based conflict.

The number of fatalities per year has been very irregular since 1989. Indeed, from 1989 to 2000, the number of fatalities was medium low: ranging between 20 and 300 per year. However, in the years from 2000 to 2005, there was an gigantic upsurge of occurrences of fatalities: no less that 750 per year. This could partially be explained by the traumatic earthquake of the year 2000 that could have increased poverty and therefore crimes.

The study could have been improved by distinguishing the number of fatalities between the type of violence. It would have shed light onto the big increase that took place in the early 2000s.

References

Croicu, M. & Sundberg, R. (2015). Ucdp georeferenced event dataset codebook version 4.0. *Journal of Peace Research*, *50*(4), 523–532.

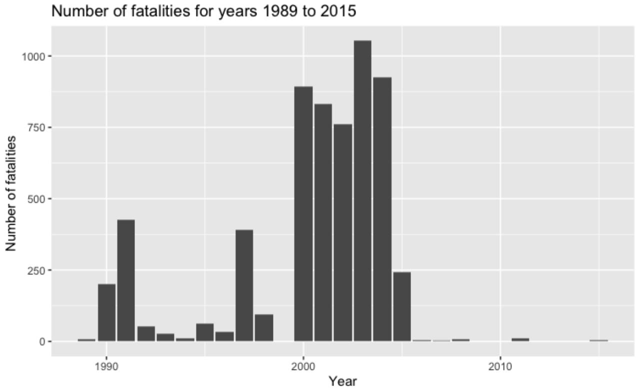

Figure 2: Fatalities per year since 1989