Urban Terror: Data Gathering and Cleaning

Sascha Schuster, Lukas Bretzinger, Cameron Reed

Thursday, November 13, 2014

Contents

Introduction

Data Collection Process

Data Categories & Sources

We use three main categories of data, which stem from a different number of sources and serve different purposes.

Global Terrorism Database

We have introduced the [GTD] (http://www.start.umd.edu/gtd/) (Study of Terrorism and Terrorism (START) 2013) extensively in the last assignment. It gives qualitative data on about over 120k terrorist attacks, including (in about 2/3 of the observations) information that can be used to georeference the attack.

Geolocated City-Data

We used two open-source datasets of city level data that we need to establish a relation between the place of the attacks and their urbanity.

a. "world.cities" from the R package 'maps'. The database "is primarily of world cities of population greater than about 40,000. Also included are capital cities of any population size, and many smaller towns." (Richard A. Becker and Ray Brownrigg. Enhancements by Thomas P Minka <tpminka@media.mit.edu> 2014) The variables include the city name, country name, approximate population (as of January 2006), latitude, longitude, and capital status indication.

b. "worldcities2013" from MaxMind Inc.(Inc. 2008). This data set provides similar information, but is updated more regularly.

c. "Urban Centers" from wikipedia. In the absence of a free data set on urban centers, we scraped a list with around 500 urban centers (>1 million inhabitants) of the [respective Wikipedia page] (http://en.wikipedia.org/wiki/List_of_urban_areas_by_population)(Wikipedia 2014). It draws from seven different types of sources and is put together in terms of defining urban space and urban centers. We added a hand-coded "coastal city" variable to indicate if a city is close to the coastline and has a port.

Note: All three datasets are time invariant. Since we did not find comprehensive data containing city-level data over the past years (which is a crucial requirement for our analysis), we need another data set for country-level data.

Country-Level Data

Our source for country-level data is the set of World Development Indicators (WDI) provided by the World Bank. We download them using the WDI package for R, a shortcut to the World Bank's API that provides data already formatted in long country-year format (Arel-Bundock 2013).

Additional Data

In the future, we plan to include additional data that helps us control for phenomena affecting our analysis.

a. For example, we are working on including civil war dummy variables because civil wars are likely to exponentially increase the amount of terror attacks in a given year and city. It comes from the Correlates of War project and is called the Intra-State War Database 4.0 (Sarkees and Wayman 2010).

Data Cleaning

Challenges in All the Data Sets

Missing Information

None of the datasets used can be considered complete with regard to the individual observations. In fact, they contain a huge number of NAs. The subset of the GTD the we use for our analysis (containing only 18 of the original 123 variabels, and only successful terror attacks) has 107143 NA values, summing up to a total of 5.2% of all values. We aim not to have a drastically increased share of NA values in the dataset used for the final analysis. All datasets are very comprehensive and stem from sources with high reputation. An extensive cleaning process was necessary nonetheless.

Spelling Inconsistencies

The main challenge across and within all datasets is the huge variation in spelling of countries and cities, which triggered an extensive hand recoding process. We developed a standardized style for country and city names and applied that to all data sets.

- GTD ~ 120k rows
- cities.a. ~ 50k rows
- cities.b. ~ 50k rows
- WDI ~ 10k rows
- Urban Centers ~ 500 rows
- War ~ 500 rows

Coding Gaps, Information Inconsistencies, and Lack of Detail

All datasets containing georeferenceable data contained this information on varying scales and for different time periods. For example, while some attacks in the GTD were probably geolocated using GPS guidance, others lack their own geoposition and are only presented using the central point of the city or district. When possible, we tried to define position data.

A huge gap existed between the WDI data and the GTD. The GTD assigns attacks to the countries they took place in at the time they happened. However, these countries (Soviet Union, Yugoslavia, GDR,etc.), in some cases, do not exist anymore. The WDI, on the other hand, contains country-level data back to 1960 in the form of countries as they are today.

Data Cleaning Process

We brought all country names to the standard of the World Bank data as a point of reference and because we will draw most of our country level data from there.

Although we combine the two world city datasets, we decided not to bring the city names to the same standard before merging them into the GTD. This has to do with the sort and amount of inconsistencies mentioned above: The more (even inconsistent, wrong, or outdated) city names we have in the world city datasets, the higher our chances to match them with cities mentioned in the GTD (even if by the coincidence of matching typos that we may have overlooked).

Because of the abysmal quality of the city_txt variable in the GTD, at least 750 lines of code were necessary to bring the ~2,5k unique city names to a level in which we could work. Codings like "somewhere at the border" or up to 10 typos (from "Buen%%s Eir$" to "Buenos Aires") for a heavily targeted city are not unusual.

Merging Process and Current Status

First, we merged the WDI country level data into the GTD by country and year. These indicators contain information on population sized in different settings (living in largest city, living in urban environment, etc.) per country and year.

Second, we merged the two city data sets. We eliminated duplicates, keeping either the city entry that was truthfully coded as capital or the one with the higher population (we ended with ~50k rows + ~50k rows = ~80k rows). As we use them to merge with the cleaned GTD city_txt variable, the more cities in our dataset, the better.

The third step is the most computing intensive one so far: We merged the urban center dataset with the now combined city dataset, assigning each city to its nearest urban center. The reasoning behind this step is that, while we have around 50k different cities in our GTD, only a share of them fulfills the requirement of being "urban" the way we understand it. A small or big distance between the city the attack took place and its closest urban center may serve as a rudimentary indicator for an intent to attack urbanity.

Therefore, we include lat/long data for each urban center using the google maps API. Then, the distance from each urban center to each city was calculated. The merged dataset assigns the closest urban center to each city (and the respective distance). The necessity comes from the way cities are coded in the GTD. While an attack on Tokyo, which is rarely attacked, is usually coded using "Tokyo", attacks in often targeted cities are usually localized more precisely - assigned to districts. Good examples for this phenomenon are Lima, or the urban area around Tel Aviv. Both are attacked often and the GTD delivers predominantly the sub-municipality as the place of attack.

With the new dataset, we can set a parameter of distance around each urban center (as a place holder we currently use $2 * sqrt(urban - centers - area/pi)$, and later decide to count any attack that falls into that parameter as an attack on the urban center itself. If the GTD codes "New York City", it finds both the urban center and the city - but as the GTD sometimes codes "Manhattan", we now have a match on the urban center "New York City" as the distance between the two falls within our parameter.

Finally, we merge the GTD and the combined city-urbancenter dataset. We use a merging varaiable which is a clean character string of the form of *countrynamecityname*, in order to avoid false positives of similar city names across countries. Thanks to our previously unified country and city coding in all datasets, we find a city (thus, population size and also closest urban center) for around 60% of all 120k terror attacks in the GTD. As the GTD often lacks any city name and has "unknown" or area codings (e.g. "District xyz"), 60% is a satisfying result given complexity and resource constraints.

Currently, We are working on increasing the robustness of our observations by cleaning further. The google maps API might provide for further analysis over lat/long calculated distances to cities within, e.g., Arabic-speaking countries with rivaling city names in the latin alphabet.

To Do Before Analysis

1. Include the 0-1 war variable in the country level data.
2. Continue cleaning process in order to increase usable observations (e.g. Drive up the matches between our combined city-urbancenter dataset and the GTD by analysing "messy" coded countries like India, Sri Lanka, and Arab countries. We did this already with Iraq and had promising results.)
3. Defend assumptions on choice of urban center radius, population growth on the city level, aggregation or disaggregation of variable values, etc.
4. Look for further helpful sources for control variables and other phenomena impeding our analysis.
5. Include population data 1970 - ~2000 into the GTD by combining WDI with city data. This could happen in the following form.

Preliminary Analysis

We have a large amount of information on each incident in the GTD already. This includes:

- Time
- Country
- WDI Data for county and year
- City
- Data on the City for 60% of the incidents including

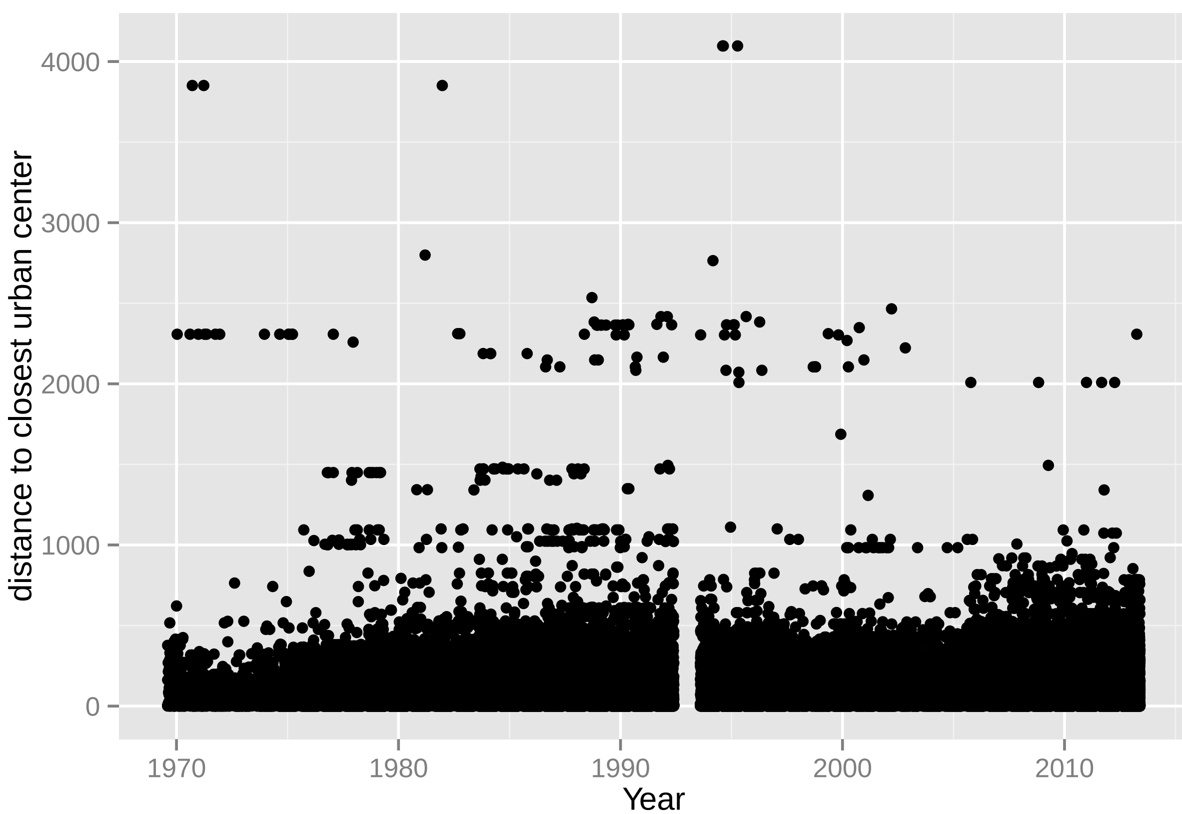
- Population estimate
    - Capital City or not
    - Distance to Urban Center
    - Population of that Urban Center
    - Area and Population Density of that Urban Center
    - Coastal Location of the Urban Center
- Attack Type (Bomb, Assault, Hostage Taking etc.)
- Target Type (eg. Restaurant, Electricity Grid, Military Installation)
- Number of Killed and Wounded
- Economic Damage from the Attack

Examples of what we can say (tables, figures)

With our preliminary GTD (PreGTD), we can already look at a lot of different information that helps us understand the distribution of terror attacks across either time *or* space (time and space is not possible since we do not have population data available on a city level across time so far). We use this information to reveal pecularities in the data that we need to investigate. We can use this to either explain features or direct us towards further data cleaning.

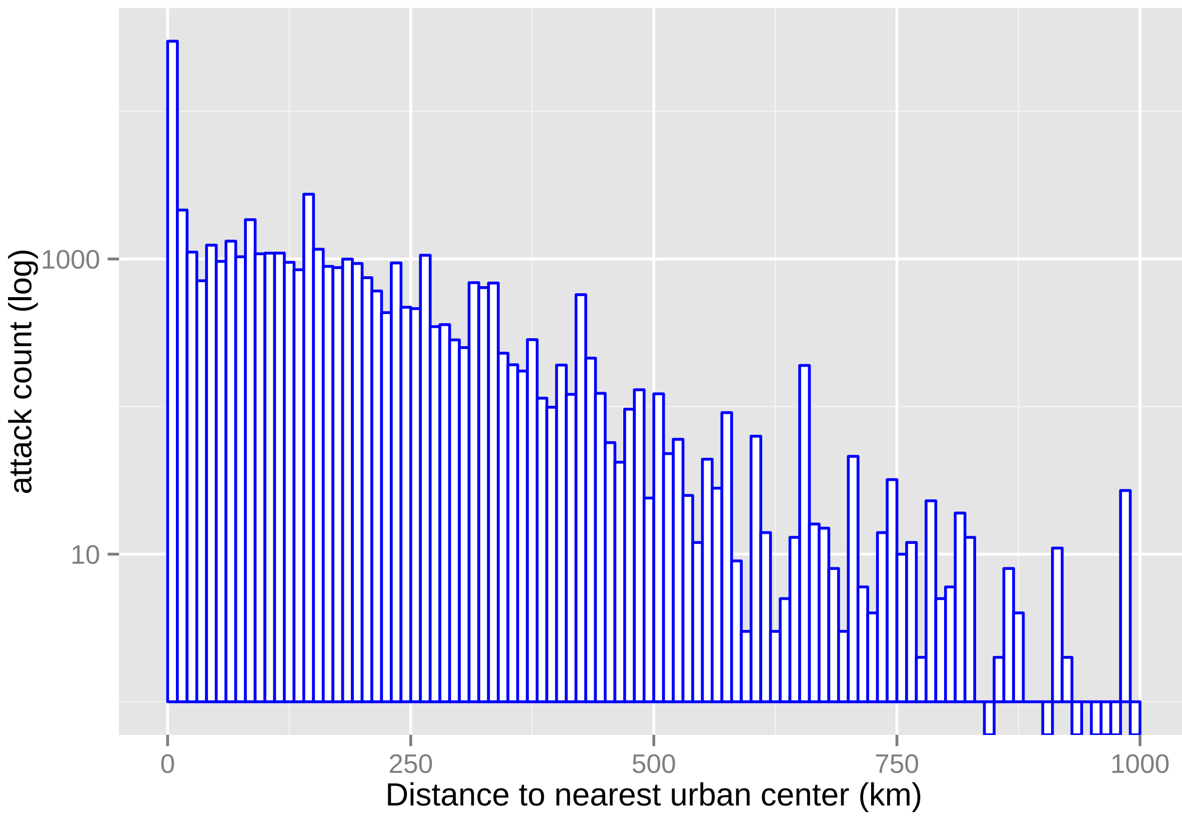
So far, our analysis has given us results along the line that we suspected to find or that is in line with other resarch. For us, this means that our way of approaching the problem and conducting our research is suitable.

Table 1: Attack's Distance to closest urban center over time



This table shows that there the average distance of attacks to the closest urban center varies over time. Overall, there seem to move closer. Here, we do not control for any variable. We also discovered that there are no attacks in 1993. We assume this to be an error, but we do not now where it comes from right now. We found it as well in the original GTD.

Table2: Histogram of attacks over increasing distances from nearest Urban Center (binning attacks to 10km)



We see that the is strongly decreasing number of attacks with increasing distance to the nearest urban center. That is supporting our initial idea. However, we do not control for time, what the target is, or the urbanity of the target.

NOTE *Because of the necessarily intensive computation process in for example gathering data from google and the WDI and loading the GTD, we have used the write.csv function in several intermediate steps. Reproducing the code is possible however, If you want to do that, the DataCleaning.R script is the main file. Different levels and the respective cleaning processes can be found in their folders (City Data, Country Data…). Also, we are thinking about creating a new organizational structure for the repo, which will probably happen over the next days.*

References

Arel-Bundock, Vincent. 2013. *WDI: World Development Indicators (World Bank)*. http://CRAN.R-project.org/package=WDI.

Inc., MaxMind. 2008. *World Cities Database [Data File]*. http://download.maxmind.com/download/geoip/database/LICENSE_WC.txt.

Richard A. Becker, Original S code by, and Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka <tpminka@media.mit.edu>. 2014. *Maps: Draw Geographical Maps*. http://CRAN.R-project.org/package=maps.

Sarkees, Meredith Reid, and Frank Wayman. 2010. *Resort to War: 1816 - 2007*. http://www.correlatesofwar.org/.

Study of Terrorism, National Consortium for the, and Responses to Terrorism (START). 2013. *Global Terrorism Database [Data File]*. http://www.start.umd.edu/gtd/using-gtd/CitingGTD.aspx.

Wikipedia. 2014. "List of Urban Areas by Population — Wikipedia,The Free Encyclopedia." http://en.wikipedia.org/w/index.php?title=List_of_urban_areas_by_population&oldid=632678337.