Project report (Group 2)

Manuel Rickli and Lukas Stöckli

University of Basel
Databases (CS244) course
Fall Semester 2018

1 Introduction

This report covers the work on following projects tasks:

- Select 2-3 interesting data sources in various size, format and source.
- Analyze sources and create integrated schema.
- Integrate sources into unified and normalized database
- Use data for analysis queries and visualizations to achieve analysis goals.

The datatesets used for this project are Global Terrorism from 1970-2017, Metal Bands, World Population Data from 1960-2015 and Global Weather Data. The following analysis goals are posed:

- Are terror events dependent on the weather?
- Do terror attacks influence founding/splitting of metal bands and vice versa?
- Does the population influence the number of existing metal bands?
- Do terror attacks have an influence on the population?
- Which main genre has the most terror events?

To achieve this, the single sources are analyzed and combined to an integrated schema. This allows an overview of the existent data and it's derived relations. The data is then cleaned, normalized and persisted to a database to further process it. For analysis, index structures on the foreign keys as well as views are created. Analysis queries and visualization are then used to take an insight in the posed questions. The report is concluded with the results and lessons learned.

2 Sources

Four data sources were chosen for this project. The global terrorism data source deemed to be interesting and was selected at first, followed by metal bands which strangely also had the population data (different file, but same source). The idea of weather having an influence on terrorism occurred and a fitting weather source was selected in addition to the others. The next part provides a short introduction to the single data sources.

2.1 Global Terrorism 1970 - 2017

- URL: `https://www.kaggle.com/START-UMD/gtd`
- Dimensions: 181'691 rows x 135 columns
- Size: 162.8 MB
- Format: CSV

This data set contains a list of global terror events. The tuples state a location, time and descriptions of attack groups, targets, weapons used, etc..

2.2 Metal Bands 1964 - 2016

- URL: `https://www.kaggle.com/mrpantherson/metal-by-nation#metal_bands_2017.csv`
- Dimensions: 5000 rows x 7 columns
- Size: 264 KB
- Format: CSV

Here, 5000 metal bands are listed with the country they originate from, when they were formed and when they split up. Additionally, one or multiple metal styles are given for each band.

2.3 World Population 1960 - 2015

- URL: `https://www.kaggle.com/mrpantherson/metal-by-nation#world_population_1960_2015.csv`
- Dimensions: 264 rows x 57 columns
- Size: 125 KB
- Format: CSV

This data set contains the yearly population of 264 countries.

2.4 Weather Data

- URL: `ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/`
- Inventory
  - Dimensions: 65236 rows x 6 columns
  - Size: 26.9 MB
  - Format: TXT
- Daily
  - Dimensions: ~10M rows x 35 columns
  - Size: 2.9 GB
  - Format: DLY

The inventory file contains information about the time span of measurements, location and id of all weather stations from the National Oceanic and Atmospheric Administration. After the relevant stations are found, their measured data can be found in the `dly` file, which has the same name as the station ID. The data consists of daily means of multiple elements (such as temperatures, precipitation, snow fall, etc.) for the whole time span the station was active.

2.5 ER Diagram

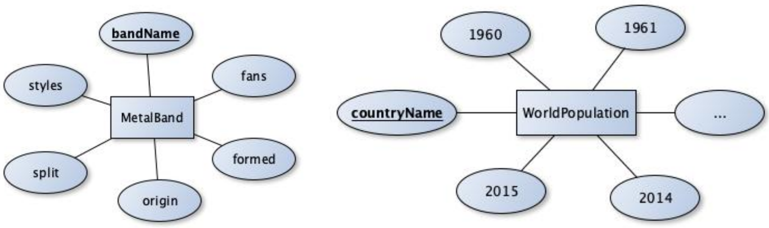


Fig. 1: Metal and Population

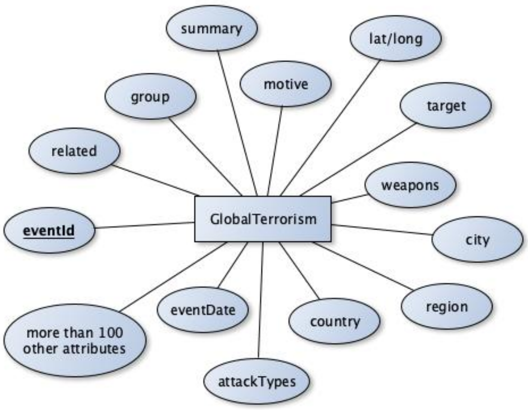


Fig. 2: Terrorism

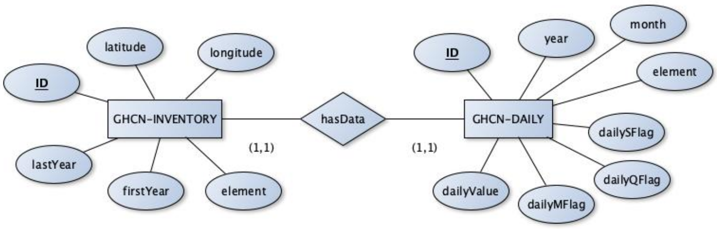


Fig. 3: Weather

3 Integrated Schema

A new population entity replaces the year attribute in *Country*. A new MetalStyle entity replaces the styles attribute in *MetalBand*. Some of the terror data is outsourced to new entities, as some attributes are listed data points (same colunm name with different index), as well as the *related* attribute that has e new entity *TerrorRelation*. A new entity *TerrorLocation* with unique locations is created to simplify relations with countries and weather data.

3.1 Logical Schema

- Country (countryName)
- MetalBand (bandName, formed, origin, split)
- MetalStyle (SID, bandName, style)
- Population (PID, country, year, population)
- TerrorAttack (AID, EID, attackTypeID, attackType)
- TerrorEvent (EID, eventDate, approxDate, extended, resolution, LID, summary, crit1, crit2, crit3, doubtterr, alternativeID, alternative, multiple, success, suicide, nkill, nkillus, nkillter, nwound, nwoundus, nwoundte, property, propextentID, propextent, propvalue, propcomment, addnotes, weapdetail, gname, gsubname, gname2, gsubname2, gname3, gsubname3, motive, guncertain1, guncertain2, guncertain3, individual, nperps, nperpcap, claimed, claimmodeID, claimmode, claim2, claimmode2ID, claimmode2, claim3, claimmode3ID, claimmode3, compclaim, ishostkid, nhostkid, nhostkidus, nhours, ndays, divert, country, ransom, ransomamt, ransomamtus, ransompaid, ransompaidus, ransomnote, hostkidoutcomeID, hostkidoutcome, nreleased, scite1, scite2, scite3, dbsource, INT_LOG, INT_IDEO, INT_MISC, INT_ANY)
- TerrorLocation (LID, countryID, country, regionID, region, provstate, city, latitude, longitude, specificity, vicinity, location)
- TerrorRelation (RID, EID, related)
- TerrorTarget (TID, EID, targTypeID, targType, targSubtypeID, targSubtype, corp, target, nationalityID, nationality)
- TerrorWeapon (WID, EID, weapTypeID, weapType, weapSubtypeID, weapSubtype)
- Weather (LID, weatherDate, rain, temperature, station)

3.2 ER

see next page

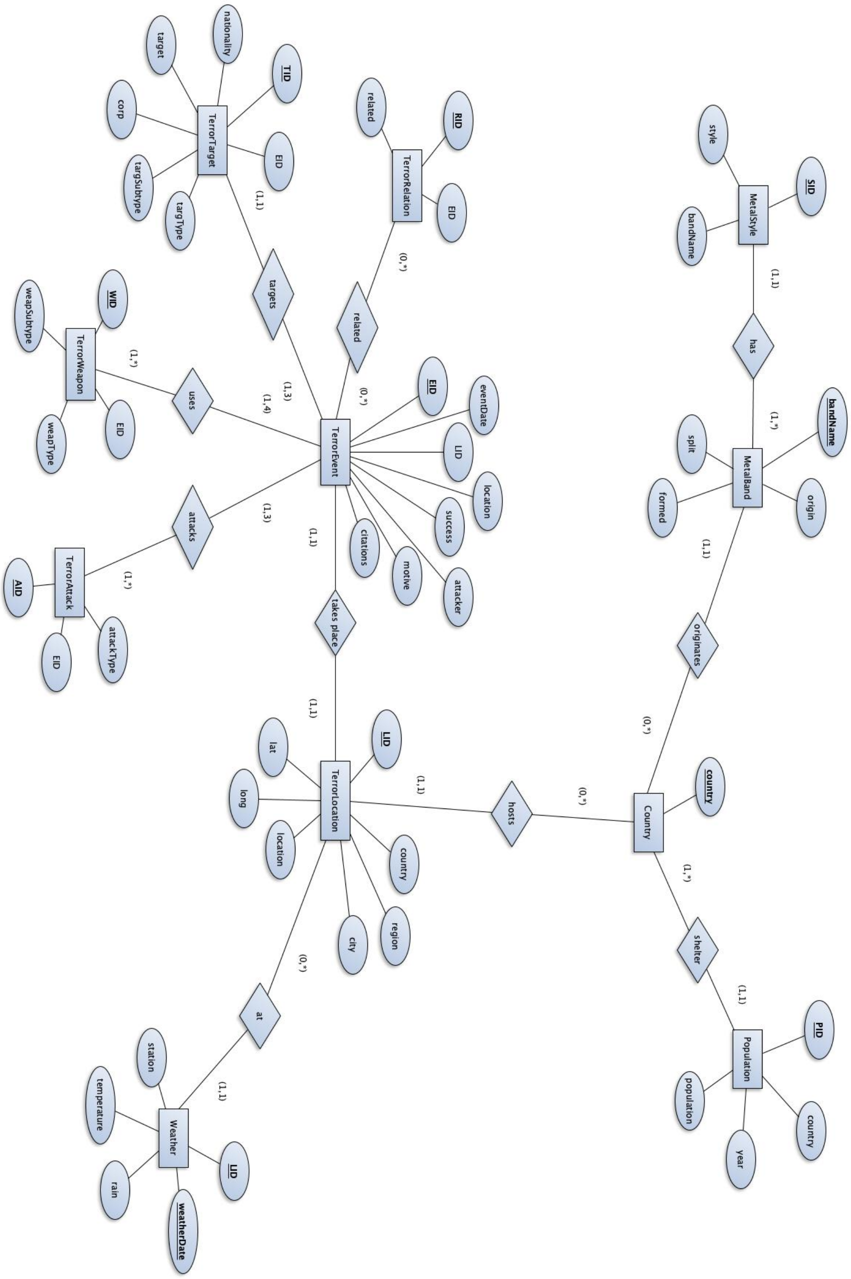


Fig. 4: Integrated schema

4 Methods

The chosen datasets are transformed and uploaded to the database where views are created. The data is then used to answer and visualize given questions by plots and maps.

4.1 Data Integration

To retrieve the integrated entities from the raw sources, *R* is used as it allows to create data frames that can be pre-analyzed and written to **.csv*. The R script mostly just readjusts the columns to new frames but also cleans and splits some raw data. Apart from some character level import fixes, the following changes are made to the raw datasets:

- **Country** The country column represents the new *Country* entity. The population columns are split up and added to the new *Population* entity along with country and year.
- **Metal** Everything but styles represents the new *MetalBand* entity. The styles are split and added to the new *MetalStyle* entity, along with the band name and an autoincrement primary key.
- **Terror** The distinct entries of the location columns form the new *TerrorLocation* entity. These have an ID and are mapped back to the new entity *TerrorEvents*. The up to four weapon types from raw data are organized in the new entity *TerrorWeapon* along with the event ID. The same procedure is applied to up to three attack and target types to yield the new entities *TerrorAttack* and *TerrorWeapon*. The column *related* from raw data is split and added to the new entity *TerrorRelation* along with the event ID and an ID. Finally, for the table *TerrorEvent*, that holds most of the raw data, the columns *year*, *month* & *day* are combined to a new column *eventDate*.
- **Weather** For each terror location, the closest ($<$ 50km) weather station and then the data for the event dates is retrieved and added to the new *Weather* entity along with the location ID and the weather date.

Each entity's file is then uploaded to a *MySQL* database by a *Python* script, using *mysql.connector*, that reads and inserts the rows.

4.2 Analysis

The analysis part is written in *Python*, using *matplotlib* to visualize the results. The scripts execute predefined queries, again with the use of *mysql.connector*, plots it and saves the plots in a directory. Three kinds of plots were used:

- **Bar plot** Used for Genre vs Terrorism.
- **Star plot** Used to demonstrate influence of weather on terrorism. The star plot has the advantage of showing changes, in proportions and not only total numbers, nicely.

- **Graph** Used to compare all other questions. The graphs always show two lines, one for each aspect that is inspected, to enable a direct comparison between them.

Since *MySQL 5.7* doesn't support *LIMIT* in subqueries, a view with the top (most bands) metal countries and one with the top target types is created for further analysis. To answer the questions, the following data was gathered:

- **Are terror events dependent on the weather?** For the chosen aspects of terror events (attack type, target type and weapon type) the significant entries are selected and represented in a star plot. The query matches all events which occurred at the same time and location for which there is weather data available. The result set is then further decreased by selecting only events in the predefined weather condition. The number of events is then counted for each chosen entry.
- **Do terror attacks influence founding/splitting of metal bands and vice versa?** A list of countries that have both band and terrorism data is retrieved, then for each country the terrorism and band data is retrieved. To get the mean a list is used where formed gets added and split gets subtracted.
- **Does the population influence the number of existing metal bands?** A list of countries that have both band and population data is retrieved, then for each country the population and band data is retrieved. To get the mean, a list is used where *formed* gets added and *split* gets subtracted.
- **Do terror attacks have an influence on the population?** A list of countries that have both attack and population data is retrieved. Then for each country the yearly attack and population data is retrieved.
- **Which main genre has the most terror events?** To get somewhat significant data, the top 30 countries with most (over 15) metal bands are selected. Then the country's main genre is selected along with the number of attacks. Group by genre, sum attacks and mean attacks by countries yields the result.

5 Results

The posed questions are answered by creating visualizations of the integrated data. Each visualization is designed to show if there is a correlation between the inspected attributes.

5.1 Are terror events dependent on the weather?

Here, the influence of the weather on terror events is inspected. The visualizations show the number of events that took place under the specified conditions. The weather influence is measured by observing the distribution of terror events for different conditions.

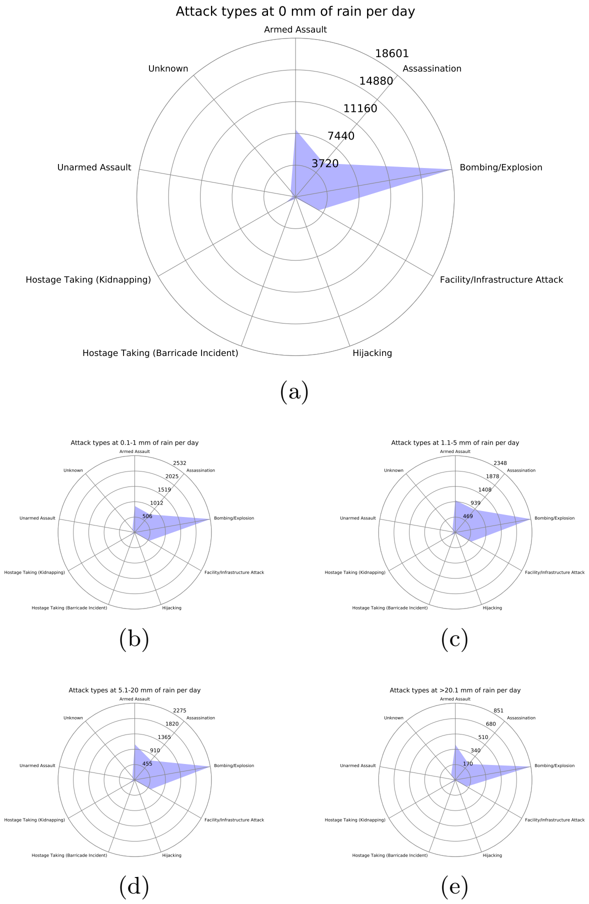
The weather data contains information about the daily mean temperature and daily precipitation. The temperature is split into intervals of 10°C beginning with $< -10°$C and ending with $> 30°$C. The daily precipitation is mapped to types of rain, namely: no rain, light rain, moderate rain, heavy rain and very heavy rain.

For the terror events, three aspects are chosen:

- Types of terror attacks
- The targets of attacks
- The used weapons in the attacks

These three aspects are represented by the tables `TerrorAttack`, `TerrorTarget` and `TerrorWeapon`, for which only the most significant attributes are chosen, if the total number of them is too large.

Weather - Attack Types There are nine distinct attack types, which are all displayed in the visualization. Bombing is the most frequent one for each weather condition, followed by armed assault.



(a)

(b)

(c)

(d)

(e)

Fig. 5: Influence of rain on terror attack types

The influence of rain on attack types is very low, as the different types are proportionally similar for each type of rain.

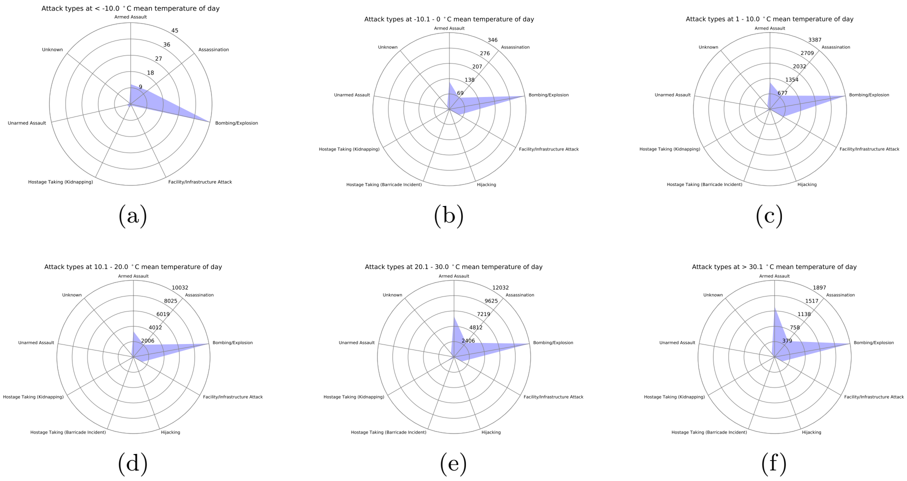
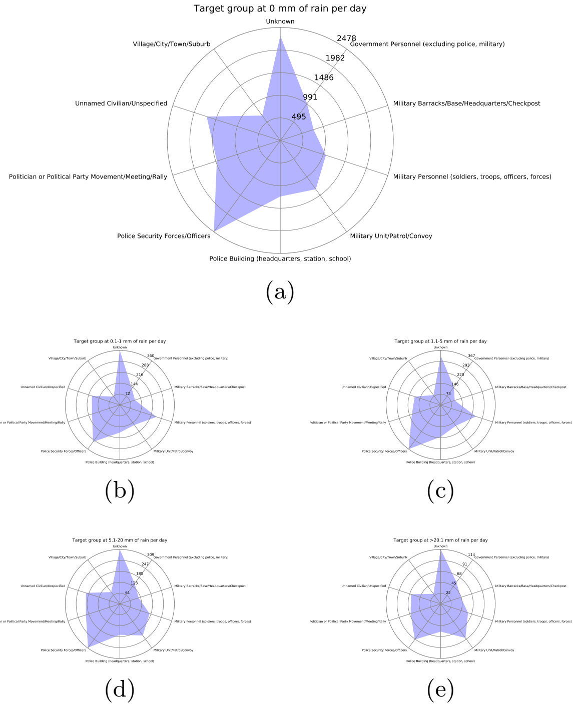


Fig. 6: Influence of temperature on terror attack types

The temperature has a greater influence on attack types. It can be observed that more armed assaults take place when the temperature rises.

Weather - Targets Types The number of distinct target types is quite large, since they can be very specific (e.g. Priest). For the analysis, the ten most representative attributes have been chosen. Different to the attack type, the target types vary more for the different conditions.



(a)

(b)

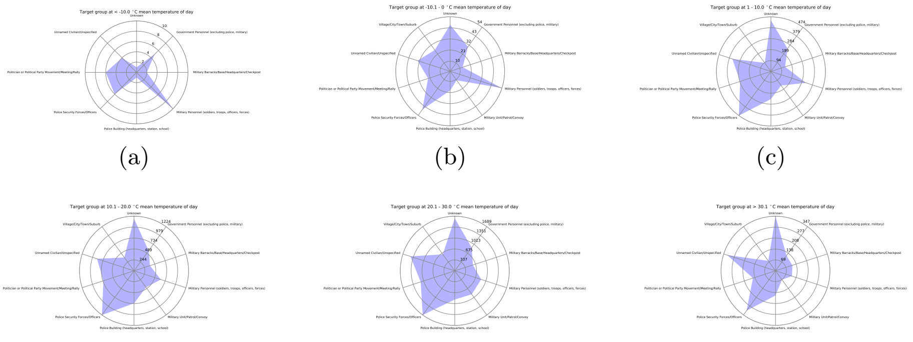
(c)

(d)

(e)

Fig. 7: Influence of rain on attack targets

It can be observed, that heavier rain results in a bigger number of attacks on military units, patrols and convoys.



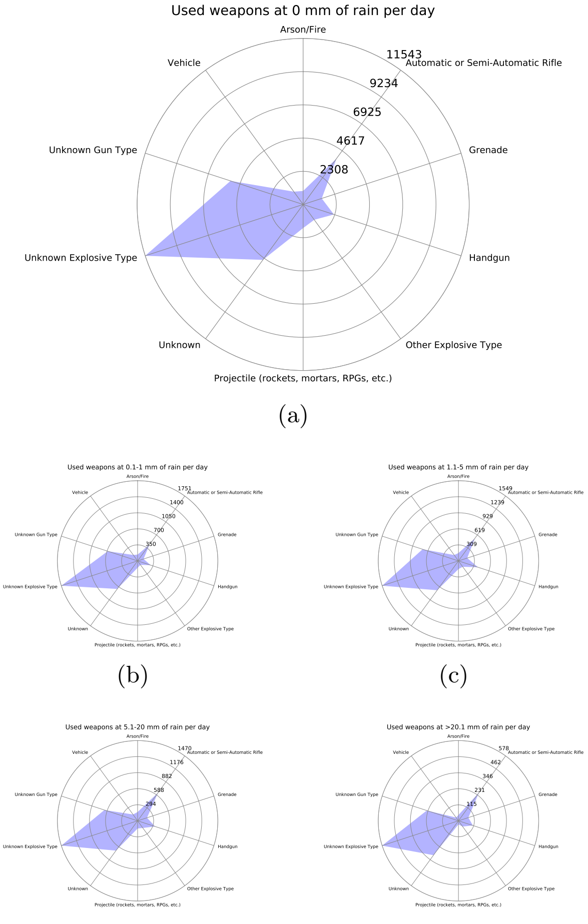
(a) (b) (c)

(d) (e) (f)

Fig. 8: Influence of temperature on attack targets

Lower temperatures have a high number of attacks on military personnel. With increasing temperature, this shifts towards civilians. Therefore, with higher temperature, more civilians but less military personnel are attacked.

Weather - Weapon Types There are, like attack targets, many distinct attack weapons. Again, the ten most representative attributes have been chosen. The weapons have a high correlation to the attack types, seen by the attributes `Bombing/Explosion & Unknown explosive type` and `Armed assault & Unknown gun type`.



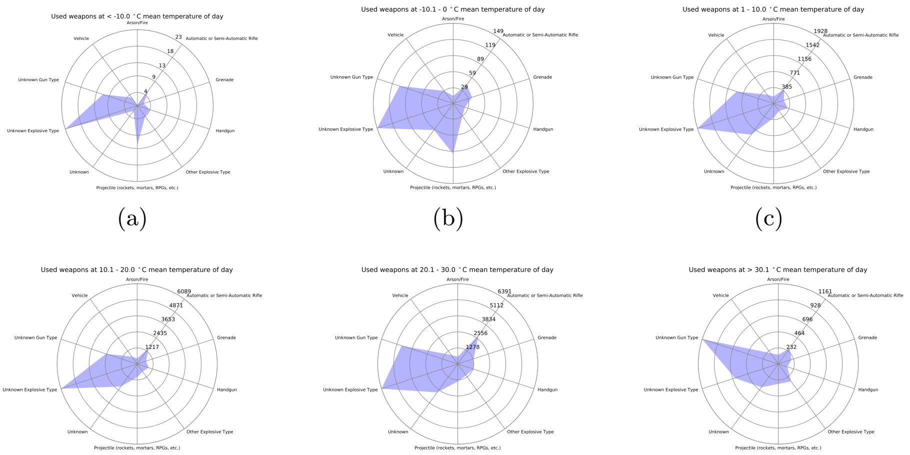
(a)

(b) (c)

(d) (e)

Fig. 9: Influence of rain on terror attack weapons

Similar to the attack types, the influence of rain on the used weapons can hardly be seen.



(a) (b) (c)

(d) (e) (f)

Fig. 10: Influence of temperature on terror attack weapons

As the armed assaults increase with temperature, the number of guns used increases as well.

5.2 Are acts of terrorism related to the number of metal bands?

These plots show the number of formed, split and existing metal bands vs the number of attacks per year.

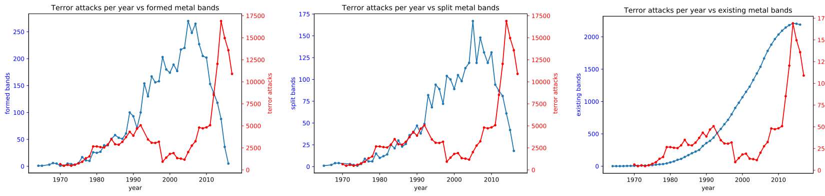


Fig. 11: formed, split and existing | l,blue: bands / r,red: attacks

The curves of formed and split metal bands show the same evolution. First the data is aligned, then it diverges. To have another view, the mean data (existing bands) was also plotted. One could say that terrorism triggered the creation of metal, but then metal went all in. A peak of terrorism then stopped the rapid growth. But no, the number of terror attacks is probably not related to the number of metal bands.

5.3 Is population growth related to number of metal bands?

To answer this question, a country's population was plotted against the number of existing metal bands in this country.

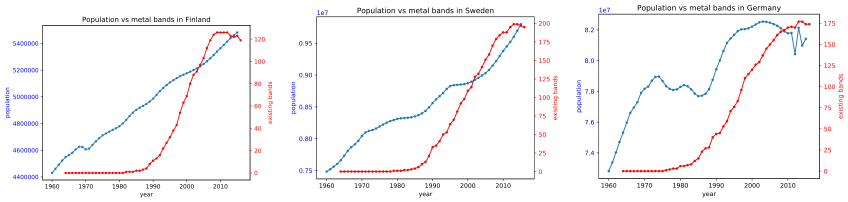


Fig. 12: Finland, Sweden and Germany | l,blue: population / r,red: bands

There are a lot of different patterns for the population, while the band curve pretty much remains the same, but no interesting correlation. What was found, though, were countries that show an opposite behaviour for population and metal bands.

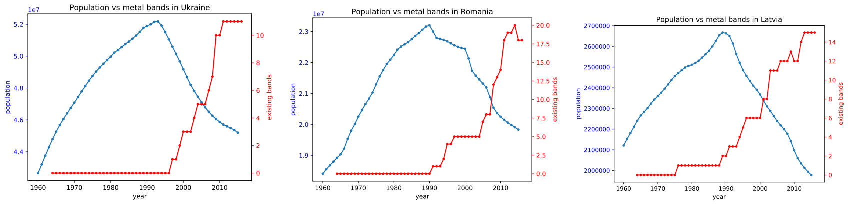


Fig. 13: Ukraine, Romania and Latvia | l,blue: population / r,red: bands

It seems as after the rise of metal the population crashed. As it turned out, after the opening of borders due to the fall of communism, there was a lot of emigration. So population growth and metal bands probably aren't related either.

5.4 Are acts or terror related to population?

The number of terror attacks are plotted against the population by country.

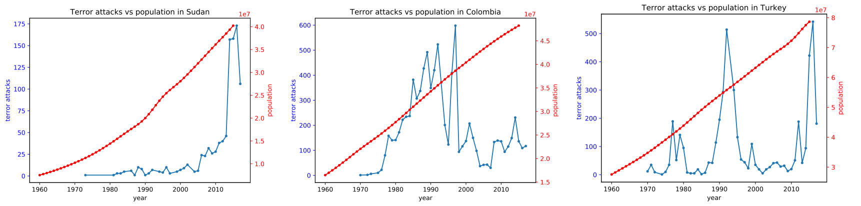


Fig. 14: Sudan, Colombia and Turkey | l,blue: attacks / r,red: population

Obviously population (growth) doesn't care about terror attacks. A lot of different patterns were found, both for population and terrorism. So terrorism and population probably aren't related either, although some countries show interesting figures.

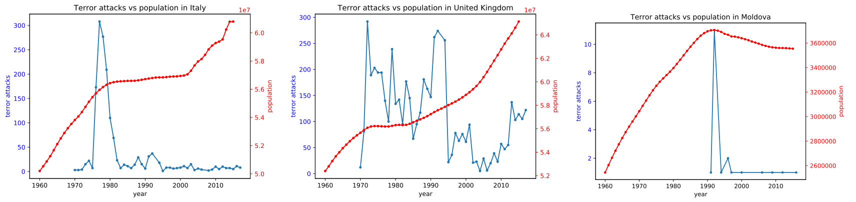


Fig. 15: Italy, United Kingdom and Moldova l,blue: population / r,red: attacks

Some countries have a stagnating or even decreasing population after or during the presence of terrorism. Moldova is probably also communism related, the behaviour in Italy and the UK remain unknown to the authors.

5.5 Does a country's main (most represented) metal genre have influence on terror?

How do you compare metal genre with terrorism? Comparing band names with types and subtypes of attacks, targets and weapons brought two results: There is a metal band and a terror group called *Condor* and there is a metal band called *Suffocation*, which may or may not be related to 17 incidents in the past years. Number of attacks vs main genre gives us:

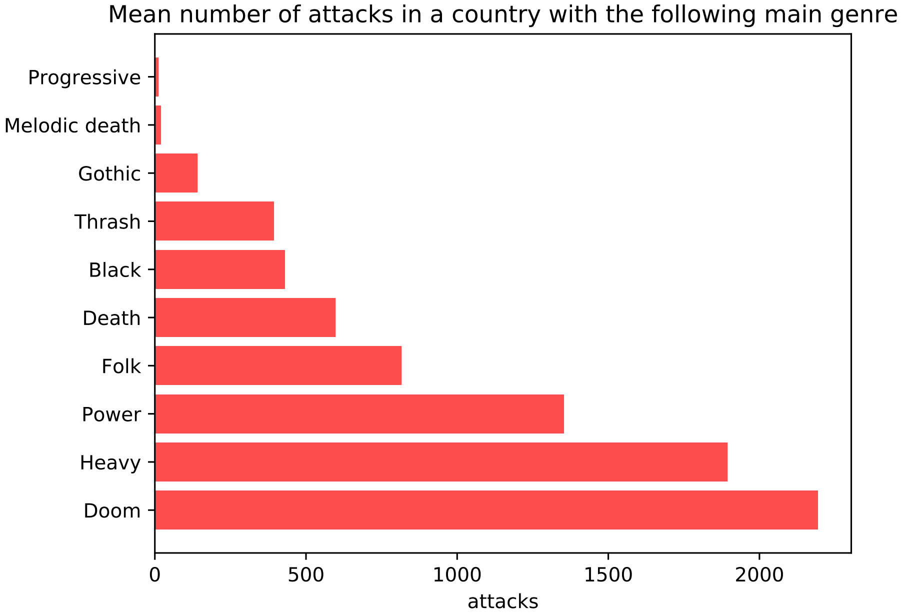


Fig. 16: Main genre vs terrorism

So if your country's main genre is progressive you should be good, if it is doom you should probably leave.

5.6 Terror Map

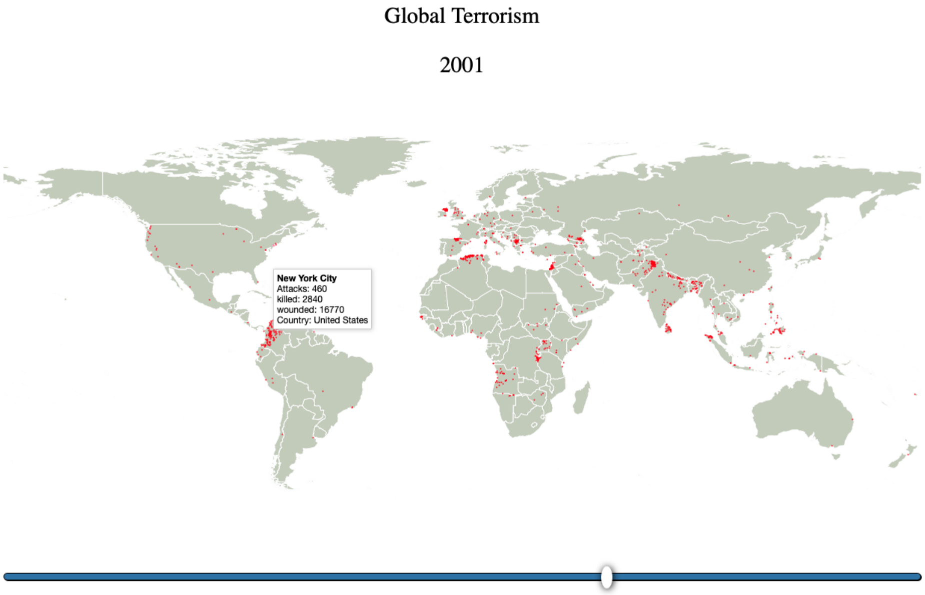


Fig. 17: Global Terror Map

The terror data points were visualized with *Datamaps*. For each year, the distinct locations were retrieved and plotted. Mouseover the location shows the city, country, number of attacks and the number of killed and wounded people.

6 Lessons Learned

In the course of this project were some lessons to be learned for us:

- **Working with large data sets** Getting the desired information from large datasets proved to take longer than expected. Especially the required steps before integrating the weather data take a long time.
- **Data may consist of errors or have missing parts** The assumption that data from a legitimate source is flawless is definitely wrong. For instance, we found out that there are no terror entries for the year 1993.
- **Some formats are difficult to work with** The lesser known format `dly` used for storing weather measurements posed some difficulties because it had to be read manually. Therefore a documentation of the formats' structure was necessary and had to be studied before being able to access the contained data.