Global Terrorism Database Analysis for the Prevention of Terrorist Activities

Liam Keyek
Department of Computer Science
University of Colorado Boulder
Boulder, Colorado, US
like4684@colorado.edu

ABSTRACT

The Global Terrorism Database (START, 2022)[1] provides data to analyze the dynamics of terrorist attacks. This project applies advanced data mining techniques and methodologies to extract understandings and generate insights into terrorism trends. Specifically, the project focuses on spatial and temporal patterns and network analysis among various terrorist entities. By applying statistical tools and data visualization techniques, the project aims to inform the development of new technologies and strategic defense adjustments to prevent terrorist activities targeted at the United States.

Temporal analysis of terrorism incidents is one of the core objectives of the project. This section will investigate the frequency and nature of terrorist activities over time. Identifying specific times of heightened activities will determine correlation with significant dates, events, or periods of political unrest.

The spatial analysis will focus on geographical patterns of terrorism. Spatial analysis will support discerning whether terrorist events are isolated or if there are clusters.

Additionally, the project will utilize the data about tactics and weapon use to understand the temporal change of terrorist operations. This is a critical area of study, as understanding the mode of operation of terrorist entities can significantly aid in US counter-terrorism efforts and inform US defense capabilities. The analysis will examine how these tactics have adapted overtime to increased surveillance and security measures by allied nations.

In the network analysis component, the goal is to understand relationships between terrorist groups. This analysis aims to inform future strategies for combating and mitigating the impacts of terrorism towards the United States.

LITERATURE REVIEW

Researchers from the National University of Defense Technology in China (Thakur, 2014)[2] explored the role of data science in counter-terrorism efforts. The team utilized data from three primary sources, including open-source internet data, social media content, and manually collected data. Data preprocessing was applied, using techniques like data cleaning, data integration, data transformation, and classification. The study used Social Network Analysis (SNA) to find patterns and insights within terrorist networks. SNA identified nodes and connections in these networks. They utilized degree centrality, closeness centrality, and betweenness centrality to find influential terrorist organizations. The researchers also focused on surveillance and forecasting, applying dynamic social network analysis for real-time insights and predictive analytics. Methodologies from longitudinal social network studies were adapted, using tools like Markov models and Multi-agent simulation models, which were proven effective in monitoring parameter shifts in terrorist groups.

Researchers from Delft University of Technology (Verhelst, 2020)[3] discuss the complexities and challenges of using machine learning (ML) in counterterrorism efforts. ML relies on identifying patterns within large data sets. The uniqueness of each terrorist attack makes it difficult to train algorithms effectively. They highlight training is challenging due to mathematical complications like class imbalance, the curse of dimensionality, and spurious correlations. The researchers underscore that ML can theoretically enhance security measures by predicting terrorist activities, but its practical efficacy is limited. There is a risk of misclassification, leading to false positives. Secondly, the unpredictability of terrorist attacks means that available data is often insufficient for training purposes.

PROPOSED WORK

The first phase is the data cleaning process, which is crucial for maintaining the integrity and reliability of the analysis drawn from the GTD. This phase will address several key issues within the dataset. Firstly, I will tackle missing values in various fields (e.g. approxdate, resolution, latitude and longitude) by applying appropriate imputation methods or comprehensively documenting these instances. I will also focus on outlier detection and management in continuous variables (e.g., nkill, nwound, propvalue). I'll use strategies such as transformation or binning to ensure these data points don't change future analysis. Performing consistency checks on categorical variables such as country_txt, region_txt, and attacktype1_txt will be completed to maintain uniformity.

The data preprocessing stage of the project is focused on preparing the GTD dataset for more advanced analytical and modeling tasks. I'll begin this process by encoding categorical variables and transforming them into a format that is conducive to modeling. I will also conduct feature engineering to develop variables relevant to the project. For instance, introducing binary variables that indicate whether incidents resulted in U.S. citizens

being killed or injured. Additionally, to capture temporal aspects, feature engineering reflecting the time between incidents will be developed. Conducting normalization of continuous variables will allow me to improve effectiveness when used for modeling. Finally, I will address class imbalance particularly between binary variables. This will help prevent a skewed analysis and create a more representative dataset.

The data integration phase of my project seeks to augment the GTD by incorporating additional datasets. The major step is integrating country-level socio-economic indicators. This will support contextual information about the underlying factors of terrorist activities within various geographical regions. Furthermore, I will incorporate geospatial data to create spatial analysis of terrorist events. This integration allows for the creation of spatial features that add depth to the analysis, such as the proximity of incidents to critical locations like nearest capital city or international borders. Another crucial integration component is the related attribute within the GTD, which will support networking related incidents. The related variable will support forming networks and clusters showing organized terrorist activities.

DATA SET

The GTD is a dataset that offers the user longitudinal data on domestic and international terrorism incidents, which supports enhancing the understanding of terrorism dynamics. The dataset encompasses a variety of etiological and situational variables related to individual terrorist events. There are 120 distinct attributes per incident and approximately 75 coded variables. These attributes include the date, geographical coordinates, involved perpetrator groups, attack tactics, detailed target nature and identities, weapon types used, and the attack's success status. The dataset also captures specific details like claims of responsibility, extent of damage (including that pertinent to the United States), and the total number of fatalities and injuries, delineating between persons, U.S. nationals, and terrorists. The dataset is compiled from public, open-source materials like electronic news archives, books, journals, and legal documents. Some example data attributes include:

- incident date
- region
- country
- state/province
- city
- latitude and longitude (beta)
- perpetrator group name
- tactic used in attack
- nature of the target (type and sub-type, up to three targets)
- identity, corporation, and nationality of the target (up to three nationalities)
- type of weapons used (type and sub-type, up to three weapons types)
- whether the incident was considered a success
- if and how a claim(s) of responsibility was made
- amount of damage, and more narrowly, the amount of United States damage
- total number of fatalities (persons, United States nationals, terrorists)
- total number of injured (persons, United States nationals, terrorists)
- indication of whether the attack is international or domestic

EVALUATION METHODS

I will employ several quantitative methods, each tailored to the specific type of analysis undertaken. For the temporal analysis, the project will visualize yearly frequencies of terrorist incidents using line plots. These trends will be substantiated with statistical trend tests to confirm the significance and consistency of observed patterns, which will provide quantitative support of temporal shifts in terrorism incidents.

In the spatial analysis, I will utilize choropleth maps to visually highlight regions with high concentrations of incidents. I'll employ clustering algorithms, like K-means clustering, which will identify hotspots and discern geographical patterns where terrorist incidents are more prevalent, giving statistical weight to the spatial distribution observed.

For the tactic and weapon analysis, I will leverage bar plots to visually represent the tactics employed and weapon types preferred in terrorist incidents, offering a clear view of shifts and trends over the selected period. A quantitative layer will be added through the combination of the temporal analysis previously discussed. I will examine the correlation between different periods and shifts in tactics and weaponry. This method will help understand whether changes are random or part of an evolving trend.

Network analysis will involve the construction of network graphs. I will attempt to represent the relationships and collaborations between different terrorist groups. Centrality measures will be utilized to quantitatively identify the most connected nodes within the terrorism network. The network analysis will incorporate temporal features to observe the evolution of these networks over time, providing quantitative insights into how these relationships have strengthened, weakened, or altered.

TOOLS

For this project, I've selected a number of tools to facilitate data cleaning, preprocessing, visualization, and various forms of analysis. The foundational stages of data cleaning and preprocessing will use pandas for data manipulation capabilities and numpy for numerical operations. Data visualization will utilize matplotlib and seaborn for plotting functions. Ploty will be used for interactive graphs and folium for dynamic maps. I will also incorporate statistical analysis with statsmodels for exploring data and applying statistical models. To understand the complex network of relationships within the terrorism landscape, networkx will be used for network analysis and visualization. Lastly, for geospatial analysis I will utilize Geopandas for spatial data operations and shapely for manipulation and analysis of geometric

objects. I've chosen these tools to support the temporal, spatial, tactic/weapon, and network analysis of the GTD dataset.

MILESTONES

Week 1 (10/30 - 11/5): Project Planning and Review

Week 2 (11/6 - 11/12): Data Cleaning and Preprocessing

Week 3 (11/13 - 11/19): Data Analysis - Initial Phase

Week 4 (11/20 - 11/26): Advanced Data Analysis

Week 5 (11/27 - 12/3): Compilation and Review

REFERENCES

[1] START (National Consortium for the Study of Terrorism and Responses to Terrorism). (2022). Global Terrorism Database 1970 - 2020 [data file]. https://www.start.umd.edu/gtd

[2] Naman Thakur, Satnam Singh Saini, Abhishek Kumar Pathak, "Data Mining Model Framework for GTD (Global Terrorism Database)", 2022 International Conference on Cyber Resilience (ICCR), pp.1-5, 2022.

[3] Verhelst HM, Stannat AW, Mecacci G. Machine Learning Against Terrorism: How Big Data Collection and Analysis Influences the Privacy-Security Dilemma. Sci Eng Ethics. 2020 Dec;26(6):2975-2984. doi: 10.1007/s11948-020-00254-w. Epub 2020 Jul 21. PMID: 32696430; PMCID: PMC7755624.