A Visual Analytics Approach to Explore and Understand Terrorism

Abilash Nair

Abstract—The Global Terrorism Dataset (GTD) compiles significant data concerning terrorist activity on a per incident basis spanning from 1970 to 2018. The dataset maintained by the University of Maryland in the USA has been used by leading academics all over the world to understand terror groups and terrorist phenomena. This coursework will aim to investigate the factors influencing terrorism and the temporal changes in terrorist activity spanning the last decade (2010 – 2018). The visualisations are generated via Python and the analysis will utilise visualisation techniques to investigate terrorist phenomena. The paper will focus on three main research questions. The first dealing with the influence of several factors on terrorism and investigating effective visual clustering techniques. The second analysing the temporal changes in terrorist activity and its impact over time. And lastly, the influence of terror groups across the globe and their modus operandi.

1 Problem Statement

The Global Terrorism Database (GTD)[1], created and maintained by the University of Maryland in the USA, has been a great source of information regarding terrorist activity the past few decades. The meticulously collected and maintained data is an important source to understand and analyse terror group motivations as well as terrorist activity trends across the past few years. This coursework will aim to investigate the factors surrounding terrorism as well as the progression of terrorist activities and evolution of terror groups from 2010 to 2018.

A lot of data scientists in academia has analysed the data and have created their own software tools to visualise the data. The research problems investigated in this coursework will be strictly limited to visualisations that can be generated using Python. There are three main research areas that this analysis will focus on.

1. Global Terrorist Activity:
   How are the factors provided in the dataset helping us understand terrorist activity taking place globally? What sort of analysis techniques can effectively group terrorist activity?

2. Terrorist Activity Trends:
   How was terrorism changed across the last decade? Can the analysis point towards trends occurring after this decade?

3. Terrorist Groups and their Impact:
   Which terrorist groups cause the most social and economic devastation? How has their influence changed across the last decade? What is the modus operandi of major terror groups?

2 State of the Art

As mentioned earlier, many of the research papers analysed the GTD data using proprietary or publicly available visualisation software. Since temporal trends and activity was of importance, a survey of available literature yielded two publications[4][3] that analysed the dataset using primarily the 'GTD Explorer' and cross compared the visualisations using other software such as 'ThemeRiver', etc. The publications analysed data from different time periods[4] and/or focused on movement of terrorist activity across geographical locations and time[3]. For instance, the 'GTD Explorer' explored terrorist activity from 1970 to 1996 and both publications used it as part of their analysis. In Lee's paper[4], colour, grouping, user interaction, and other visual cues are explored to understand the effectiveness of the 'GTD Explorer'. The analysis reveals important temporal activity patterns over a 30-year span and primarily explores activity in the USA and Europe. In Wang's paper[3], the social and geographic causes are evaluated through the 'GTD Explorer' and 'ThemeRiver'. Their analysis plots terrorist activity on geographical maps quite like Bahgat's paper[2]. The intention of this coursework is to extend that analysis to a more recent period, namely the last decade, to understand modern terrorist activity. Although, movement of terrorist activity will not be analysed here, the general analysis approach, results, and the visualisation criteria can be very useful.

The use of clustering techniques can be extremely useful in visualisation analysis. For this purpose, a data analysis paper published in a 'Journal of Physics' conference[5] utilised Analytical Hierarchy Process (AHP) and DBSCAN to cluster the data using primarily categorical variables. The authors were able to organise the data using AHP, cluster the data, and analyse the effect of epsilon score on the clusters. The process utilised by the authors will be used as a reference technique in favour of another clustering technique known as K-mode clustering. K-mode clustering is described by Huang[6] who effectively extended the K-means clustering technique so that it is more effective when it comes to categorical data. Since most factors in the GTD dataset are categorical, K-mode clustering could serve as an effective tool. The K-mode analysis will be compared with the DBSCAN method to understand which can better explain factors affecting terrorist activity.

Finally, a geo-spatial analysis of terrorist activity was carried out in Bahgat's paper[2]. The paper explores the origin and movement of terrorist activity and attempts to explain the underlying causes of terrorism using geographical data. Geographical limitations to terrorist activity are explored and attack strategies of different groups are also evaluated. Although the analysis techniques utilised were not implemented in this coursework, the conclusions and results of the paper were very relevant to this analysis. The paper provides key root causes of terrorism that can be used to verify the analysis results in this coursework.

3 Properties of the Data

Access to the GTD data is controlled by the University of Maryland, however educational uses of the data is allowed. The GTD data is massive, containing terrorist activity data that spans from 1970 to 2018. The dataset has over 100 columns and 190,000 rows. Since this coursework aims to investigate activity in the last decade, data from 1970 to 2009 was discarded. The resulting data is very messy with several null values across the dataset, categorical variables with 'Unknown' as a category, and text as well as numerical coding that requires significant parsing. A document known as the 'GTD Codebook'[1] is extremely valuable to understand inclusion criteria of variables/data as well as the assumptions involved when recording data.

The dataset is heavily segmented and required the careful splitting of the dataset into several smaller datasets. For instance, analysis that revolves around property damage can be severely biased if the data points that focused on kidnappings are included. Any data points that were not verified to be 'terrorism proper' are discarded. The criteria that classify an event as terrorist activity can be either political, social, ideological, public coercion or intimidation, or must be outside humanitarian law. If there is significant doubt as to whether the event fits any of the above criteria, that data point is removed. Several columns were dropped as they did not have a significant influence on the research questions. Columns that dealt with news sources, headlines, claim modes, victim nationalities, etc. were dropped. Important columns in the dataset include the date, geographical information, terrorism criteria, type of attack, weapon types, target nature, casualties, and economic damages.

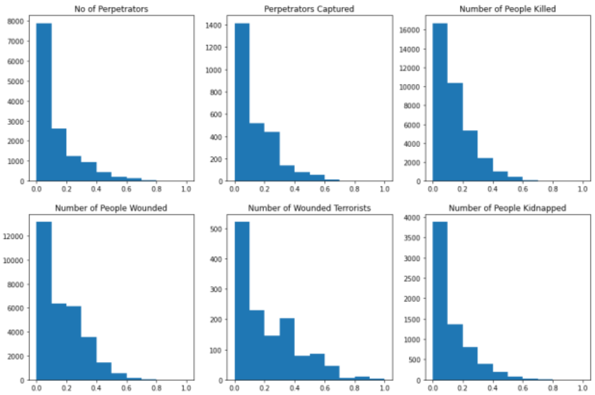

Fig. 1. Histogram Plots

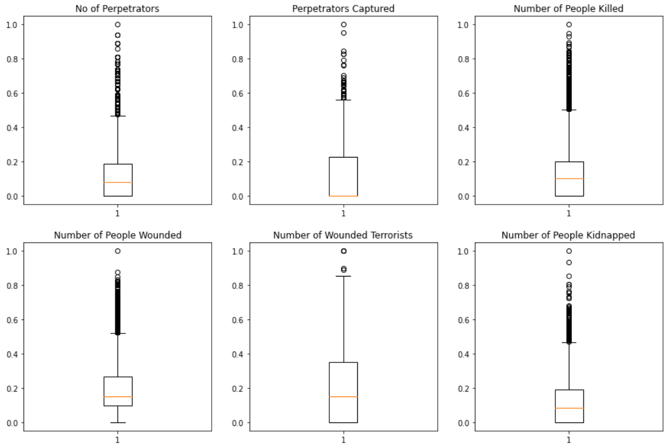

Fig. 2. BoxPlots

The NaN data points are extensive throughout the dataset. Some NaN data points in columns such as latitude and longitude were discarded as there is no approach that could impute coordinates. The approach avoids geographical uncertainty. However, other columns such as number of people killed, or wounded are continuous numerical data types that with very careful handling can be useful. The numerical columns of interest were plotted on histograms to evaluate the distributions. Almost all were heavily skewed to the left. A log transformation was used to get a better distribution and all variables were normalised to ensure consistency. The histograms and boxplots are given in figure 1 and figure 2. The NaN points in the columns were mean imputed. However, after splitting the data into focused datasets meant that the NaN points were not that extensive. The transformed values were added as new columns rather than modifying the original. Outliers were evaluated, however removing outliers from the smaller datasets would have meant a significant reduction in the volume of data. Although extreme outliers were dropped (5 sigma), most of the data points remain intact. Most of the numerical columns were transformed to better understand the data, organise the data, and prepare it for any models or clustering techniques.

4 Analysis

4.1 Approach

All the visualisations will be created through Python and its respective libraries. Excel is used to clean up the data on a macro level while Python will be used to fine tune the cleaning process. The approach will involve using three main visualisation techniques to answer the three research areas. For the first research question, clustering technique involving categorical variables will used. For second question which primarily deals with the temporal nature of terrorist activity, line plots, trend plots will be utilised. The third question which deals with the terrorist groups and their behaviour will be analysed using tree maps.

Clustering

K-mode clustering will be used to cluster the data points as it is an excellent clustering tool to classify categorical data.

For the K-mode technique, the relevant categorical columns will be included in a new dataset and one-hot encoding technique is used to generate pure binary categorical columns. An iterative technique will used to find the optimal number of clusters. Using the cost function and the silhouette scores, plots will be generated, and a visual inspection will be used to assess the optimal number of clusters. Once the clustering algorithms are run, the clusters will be visualised using a latitude and longitude scatter plot to observe the clusters. The K-mode clustering technique will be contrasted with the DBSCAN technique covered in Wang's paper[5]. However, the exact process will not be used to implement the clustering. AHP will not be used and instead the data will be significantly filtered, and the location coordinates will be run through a distance function before using DBSCAN. Although imperfect, to simplify the DBSCAN clustering, the epsilon value of 0.08[5] recommended by the group will be used.

Temporal Analysis

Terrorist activity will be grouped on a yearly basis and the social and economic damage will be assessed over the time span. To better visualise the extent of the destruction or damage, the plots will share the same horizontal axis and grouped together. Trends such as the number of people killed or wounded, the number of perpetrators, economic damage, rates of kidnapping and ransom paid will be analysed. A visual inspection of the overall trend will be used to project the trends to the next decade.

Terrorist Group Analysis

Tree maps will be extensively used to study the influence of terrorist groups and the damage and havoc they cause globally. Two sets of tree map analysis will be conducted. The first analysis will analyse the behaviour over the entire decade and an evaluation of the number of terrorist incidents committed by groups will be contrasted with the damage they cause over the same time. The second analysis will look at the temporal variations of the influences of the major terrorist groups which will be then cross referenced with the clustering and temporal analysis conducted earlier. The tree map analysis will tie into the earlier analysis and provide a lot of context for the results.

4.2 Process

The process is segmented into three parts each dealing with a single research area.

Research Question 1

Initially, the dataset was pruned to include only terrorist activity that were successful in their objectives. Further filtering was done to ensure that at least one individual was killed, or one individual wounded, or the event is related to kidnapping, or there was some form of property damage. This reduced the dataset to approximately 60,000 rows and the relevant columns such as whether it was a suicide mission, the attack type, the target information, extent of property damage(categorical), etc. were included. Once the dataset for clustering was assembled, one-hot encoding technique was used to binarize the categorical variables. Once the dummy variables were obtained the host column was discarded and eventually a dataset comprising of over 40 columns and 60,000 rows was obtained.

The next stage was to use an iterative process to find the optimal number of clusters. The K-mode algorithm generated a model with number of clusters ranging from 2 to 11. The dataset was fitted to the model and the cost as well as the silhouette score was obtained and stored in arrays. The iterative process yielded two plots, the first plot in figure 3 charts the cost vs cluster relationship while the second plot in figure 3 charts the silhouette score vs cluster relationship. For the cost vs cluster figure the 'Elbow' techniques was used to visually determine the clusters. However, the plot seemed to be quite smooth with no defining elbow. In fact, the 'Elbows' seemed to occur at clusters 4 and 7. To better determine the number of clusters the second plot containing silhouette scores was used. The higher the silhouette score the better and, in this case, there were two noticeable peaks at clusters 4 and 7. Although either choice would have been suitable, number of clusters was chosen to be 4 for an initial analysis. Seven number of clusters was chosen afterwards for a comparison.

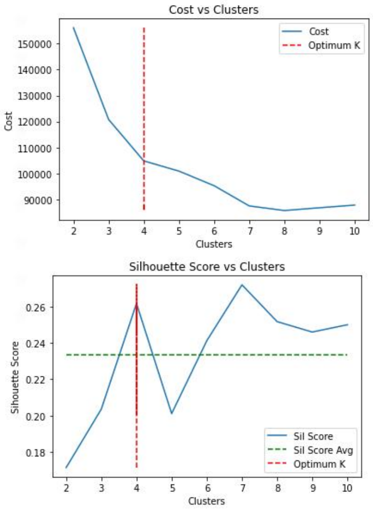

Fig. 3. Cost and Silhouette Scores

The labels generated from the clustering process was added back to the dataset. Next, the clusters corresponding to different attack types were plotted to observe which attack types were associated with cluster groups as shown in figure 4. A scatter plot was generated using matplotlib. Each diagram corresponds to one single cluster and the 4 together provides an opportunity to observe how the clusters vary across geographical locations. Another clustering process was run with seven clusters and the scatter plot is given in figure 5. Comparing the two clustering plots, it is quite difficult to observe any clearly delineated clusters forming across geographical locations. However, when the clusters are seven in total, it is relatively better to observe the differences in cluster spread across the plot.

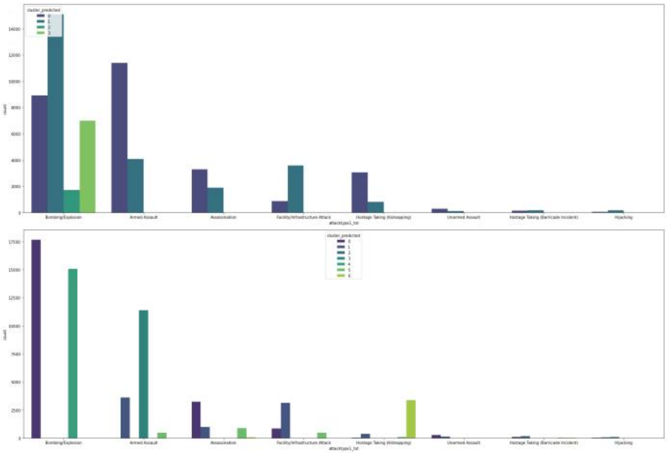

Fig. 4. Cluster and Attack Types

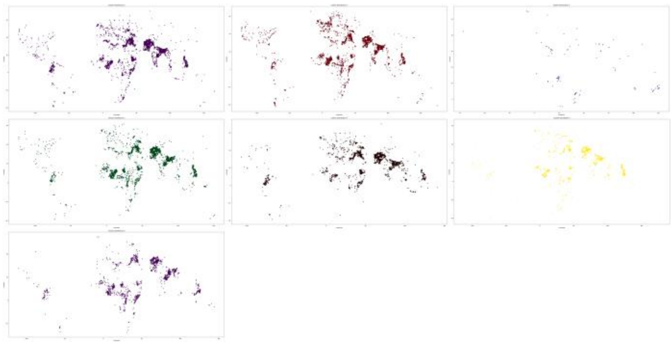

Fig. 5. K-Mode 7 Clusters

Assuming that the number of optimal clusters is seven, it can be observed that the certain attack modes are popular in certain geographical regions. To start off, clusters 1 and 2 point to armed assault and bombings which are extremely popular with terrorists across the globe. In the scatter plots, the spread of this attack strategy is present on all continents. Alternatively, hostage taking is very popular in parts of Africa, India, Afghanistan, Pakistan, Middle East, SE Asia, and few parts of central Asia and Europe. It is almost absent in the west, in countries such as the USA, UK, and Australia. Finally, cluster 3 which prominently points to facility and infrastructure attacks, is the least popular across the world with its presence predominately seen across parts of Africa and Asia. This could be because most of such facilities are guarded which disincentivises most terrorist activities. A few attack strategies are lumped together in clusters and therefore not so easy to determine its popularity.

A quick clustering analysis was done using DBSCAN with an epsilon score of 0.08 as a quick reference. The clusters are better delineated here, and a scatter plot is shown in figure 6. The clusters seem to indicate concentrations of terrorist activity. The clusters are very prominent in South Asia as well as parts of Africa.

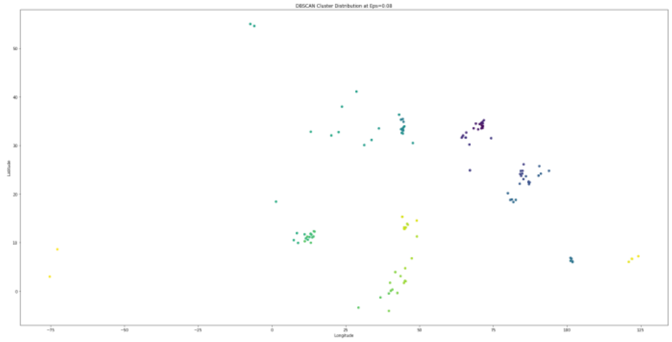

Fig. 6. DBSCAN Clustering Eps = 0.08

Research Question 2

The data was aggregated and plotted on a yearly scale to observe the changes in activity occurring across the decade. The figures were plotted together to make comparisons across factors easier. They share the same horizontal scale marked by individual years while the vertical axis is of different units. The temporal plot is shown in figure 7.

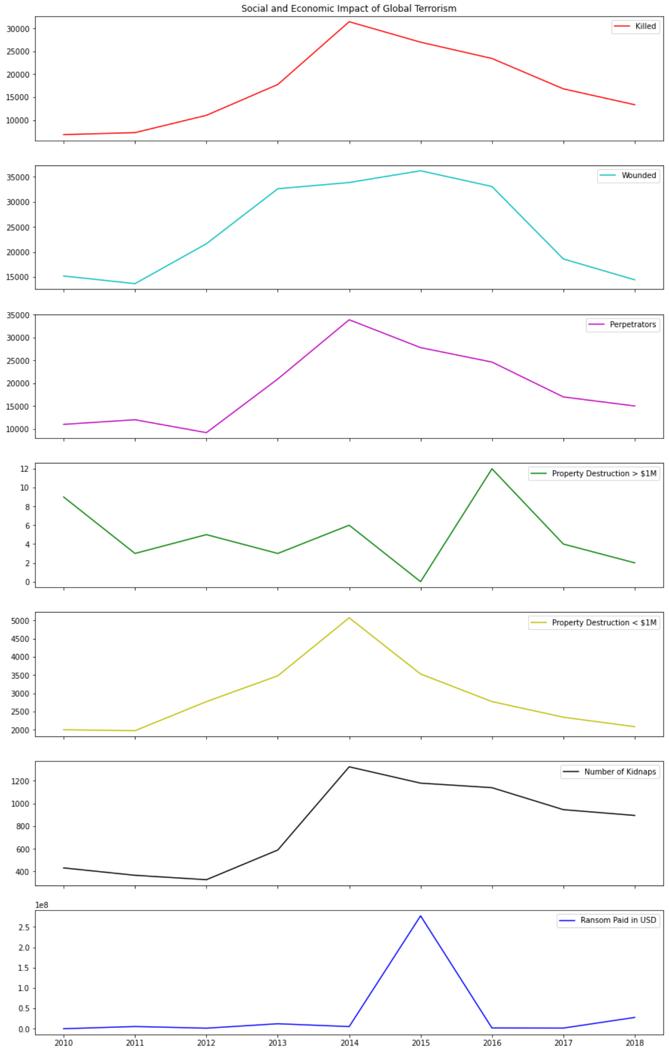

Fig. 7. Temporal Plots

It can be clearly observed from the plot that there was a massive influx of terrorist activity peaking at the year of 2014. Across all factors, be it the number of individuals killed, wounded, kidnapped, and property destroyed peaks in 2014. It can also be noticed that this was not a sudden unexpected event as there is a gradual slope pointing towards the rise in deaths, wounded, and destruction till the year 2014. The number of perpetrators also sharply increase. This will be further corroborated by the tree map analysis in the next section, which will give much needed context for this peak. It can also be observed that there is a declining slope pointing to lower incidents of terrorist activity after 2014.

Perhaps, the blowback from governmental enforcement authorities across the world could have contributed to the decline after 2014. However, this cannot be conclusively stated as there is not enough data to support it. When it comes to kidnappings, its popularity grew from 2014 and declined slightly till 2018. Interestingly, the amount of ransom paid remains relatively low throughout the decade only peaking at 2015. The peak falls sharply after that year, perhaps contributing to the slightly declining popularity of kidnappings across the world. It could indicate the unwillingness of governments and individuals across the world to pay ransoms to terrorists.

Property damage is a factor that is segmented into two sections. One plot describes economic damage that is less than 1 million US dollars while the other plot describes economic damage exceeding the million-dollar figure. It is obvious from the plots, that terrorists prefer to inflict damage to relatively low-profile targets than focusing on high profile buildings/facilities worth millions. It can also be observed that damages to low profile properties comes to around more than 2 billion US dollars from 2014 to 2018 and peaking in 2014 to a staggering $5 billion. The low amount of destruction on high profile targets can perhaps be explained by the difficulty in attacking targets that are guarded better than low profile properties.

Research Question 3

For the questions in this research area, tree maps were chosen to illustrate the influence of terrorist groups and their behaviour. This section is segmented into two parts. The first looking at the question from the entire temporal period and the other looking at snapshots of activity across only 3 years. It is to be noted that the data included in this group is rigidly filtered. The veracity of a group's involvement is vetted before including them in the dataset. So, for instance any terrorist incidents that cannot be attributed to a group with some credibility will be omitted. This is to prevent biasing the results with false claims and false attributions.

The first group of tree maps is shown in figure 8. The four diagrams illustrate the number of incidents, casualties inflicted, property damage caused, and kidnappings attributed to each group. Looking at the past decade, it can be observed that the top 10 terrorist groups are predominantly centred in Asia and Africa. By number of attacks, the Communist Party of India (CPI) leads the pack followed by the Taliban and Boko Haram. Interestingly, ISIL is a relatively minor player in the big ten.

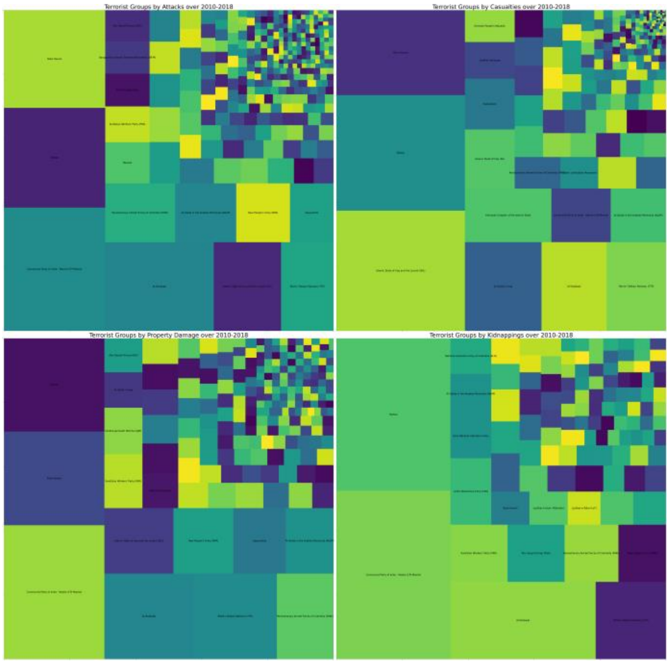

Fig. 8. Tree Map Terror Groups 2010 - 2018

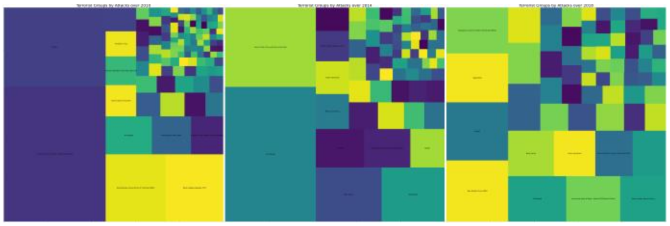

Fig. 9. Tree Map 2010, 2014, and 2018

However, when grouping by number of casualties, ISIL has an outsized influence over others. The Taliban's and Boko Haram's impact is clearly seen across both plots with a massive number of attacks with severe casualties attributed to them. Interestingly, the CPI is a minor player here, but has a strong influence over kidnappings and property damage. The preferred modus operandi of the CPI seems to be kidnappings followed by economic destruction. In contrast, the Taliban leads on all fronts perhaps being one of the most destructive groups across the world.

Looking at temporal maps seen in figure 9, the influence of terrorist groups is constantly fluctuating. As seen in the temporal plots before, it is now obvious why there was a peak in 2014. The rise of Al Shabab and ISIL is a major contributor to the social and economic destruction faced by people across Africa, Middle East, Asia, and other parts of the world. The popularity of the groups followed by significant backlash from governments across the world seem to explain the temporal data.

4.3 Results

The K-mode clustering techniques served as a useful visual technique to understand attack types across the globe. The technique has the potential to describe the effect of other factors influencing terrorism although such an in-depth investigation was not done in this coursework. The DBSCAN on the other hand tended to form well defined clusters indicating terrorist hotspots across the world. The results from the clusters are in good agreement with Bahgat's[2] and Wang's[3] results.

The temporal analysis yielded very good representations of terrorist impact across the last decade. The peak in 2014 across most temporal plots shows the massive increase in terrorist activities that resulted in massive casualties of over 60,000 and economic devastation in the billions of dollars. Overall, the trend of terrorist activity and its adverse impacts can be projected to the future. As seen from the plots, the impact of such activities is increasing, and if conditions do not change the next decade will also see damages from such activities.

The tree maps were very useful visualisations illustrating group dynamics. The maps were in good agreement with the temporal plots. Taliban, ISIL, and Boko Haram tend to be the worse perpetrators overall.

5 Critical reflection

K-mode was quite effective as a clustering technique however, the technique clearly has its limitations. The severe overlapping of the clusters across the scatter longitude vs latitude plot seems to be a poor visualisation choice. The clusters were not the most appropriate for a geo-spatial visualisation. The clusters had to be separated from each other to tease out information. Alternatively, a quick implementation of DBSCAN proved to be effective. However, its ability to explain terrorist strategies, or social impact of terrorism was not assessed and should be explored further. The clustering technique only included categorical data which also proved to be limiting. In future coursework, techniques such as K-Prototype[6] clustering can be explored which can effectively include both categorical and numerical variables for clustering. OPTICS clustering can serve as a good alternative clustering technique and its effectiveness can be compared to DBSCAN. Ultimately, the choice of clustering technique should depend on how well it can describe or explain terrorism and therefore a trial-and-error technique should be used to select the most appropriate visual.

Another limitation of this analysis, that is obvious, is the absence of choropleth maps or map visualisations. Scatter plots are rarely good substitutes for good map visualisations. However, due to library and hardware compatibility issues, map visualisations had to be discarded. In future coursework, such visualisations should be included.

Temporal maps tended to describe the overall phenomena well. However, it is lacking in richness of data which can be obtained by aggregating data points on finer temporal scales such as a monthly or daily basis. Such refinements can provide a lot more information and context to activities happening across a long-time span.

Finally, the tree maps proved to be very effective in describing the influence of terrorist groups and their modus operandi. However, the tree map could not sufficiently capture the relationships between terrorist groups and the relationship between terrorist groups and geographic locations. In the future, network graph analysis can be implemented to capture this area of detail. However, care must be taken to ensure that the visualisations are not overly complex preventing visual discerning of information.

Table of word counts

Problem statement	224/250
State of the art	458/500
Properties of the data	485/500
Analysis: Approach	488/500
Analysis: Process	1496/1500
Analysis: Results	195/200
Critical reflection	351/500

References

[1] Start.umd.edu. 2021. GTD | Global Terrorism Database. [online] Available at: <https://start.umd.edu/gtd/> [Accessed 10 January 2021].

[2] K. Bahgat and R. M. Medina, "An Overview of Geographical Perspectives and Approaches in Terrorism Research," Perspectives on Terrorism, vol. 7, no. 1, pp. 38–72, 2013.

[3] Xiaoyu Wang, Erin Miller, Kathleen Smarick, William Ribarsky, and Remco Chang. 2008. Investigative visual analysis of global terrorism. In Proceedings of the 10th Joint Eurographics / IEEE - VGTC conference on Visualization (EuroVis'08). The Eurographs Association & John Wiley & Sons, Ltd., Chichester, GBR, 919–926. DOI:https://doi.org/10.1111/j.1467-8659.2008.01225.x. Accessed 26 Dec. 2020.

[4] Joonghoon Lee. 2008. Exploring global terrorism data: a web-based visualization of temporal data. XRDS 15, 2 (December 2008), 7–14. DOI:https://doi.org/10.1145/1519390.1519393.

[5] Wang, S., Wang, G. and Zhang, J., 2019. Data Analysis Method of Terrorist Attacks Based on AHP-DBSCAN Method. Journal of Physics: Conference Series, 1168, p.032029.

[6] Huang, Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery 2, 283–304 (1998). https://doi.org/10.1023/A:1009769707641.

[7] Scikit-learn.org. 2021. Scikit-Learn: Machine Learning In Python — Scikit-Learn 0.24.0 Documentation. [online] Available at: <https://scikit-learn.org/stable/> [Accessed 10 January 2021].

[8] Matplotlib.org. 2021. Matplotlib: Python Plotting — Matplotlib 3.3.3 Documentation. [online] Available at: <https://matplotlib.org/> [Accessed 10 January 2021].