Global Terrorism Database Case Study:

Understanding the history of Hostage & Kindapping Terror Attacks?

By Taher Darwesh

***Problem Statement* —** In this paper we look at exploring hostage taking and kidnapping terror attack incidents from around the world from 1970 to 2019. We look to understand these unfortunate incidents by analysing its history and determining the worst effected regions hit by this tragedy and during which decades. We also look at how we can classify the severity of these incidents into groups and can they show us patterns or shifts in terror attacks through the decades. For each group can we explain the main motives and target types. We use The Global Terrorism Database which is an open data source in a standard object data structure. Information about domestic and international terror attacks are available such as the date and the location of the attack, the motive behind the attack, the target of the attack and third source news on the attack. This database is perfectly suitable for answering our questions due to its rich variety of variables and quantity, up to 129 different variables to choose from and data spanning nearly 50 years.

I. STATE OF THE ART

A. . New Frame Work that uses patterns and relations to undestand terrorist behaviours..

This paper [1] looks at useful patterns of suicide attacks to analyses the terrorist activity, to understand and prevent future moves due to the random nature of the event.
This paper uses The Global Terrorism Data Base as the main dataset but integrates other datasets like Pinkerton Global Intelligence Studies database to improve the quality of the overall data. The paper uses the method of an evolutionary simulating annealing lasso logistic regression model to select relevant features for the similarity function. The weighted heterogeneous similarity function is proposed to define hazardous places for the outbreak of violence. I have learned that GTD alone may not be enough quality to answer specific questions. Predicting a logistic based output for future outbreaks would be suitable piece of analysis for my case study. The similarity function is used to define popularity and outliers to determine future attacks. Calculating relations results in finding patterns to predict future attacks. The assumption that rate of successful attacks is based on select attractive attacks, if the success rate of the event is less than 0.2 or greater than 0.8 the event is considered attractive. In terms of results the city Baghdad and surrounding areas were highlighted as high risk of future attacks. For outbreaks, the network showed attack behaviors in the surrounding areas come from Baghdad due to its high degree of centrality. Future attacks can be controlled and prevented if the behavior in Baghdad is controlled.

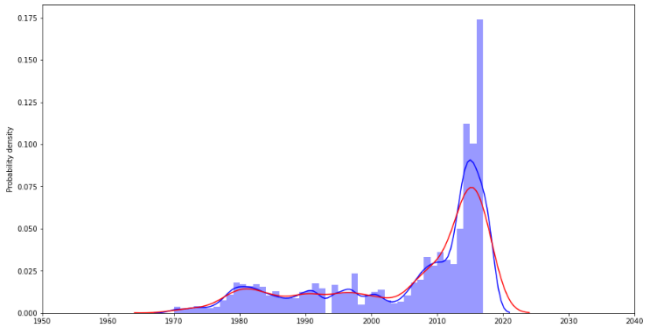
B. Space, Time and Visual Analytics

This Paper [2] investigates analyzing spatio-temporal data and solving spatio-temporal problems. The paper looks at identify these attributes to allow the undertaking analysis of information in time and space and in doing so using different approaches to deal with the complexities of the data and make it accessible to the wider community. The specifics of time such as hierarchical system of granularities is discussed alongside the commodities between time and space. Time and space in some disciplines are visualized dynamically which wouldn't be suitable for this report while time can be visualised using plots such as line graphs which show trends and outliers but also spiral visualisations which show cyclic aspects of time related data. The assumption is that current visual analytics doesn't consider the complexity of time but instead treats it as an ordinary numerical variable. The paper also discussed the problem of trying to visualise every row for large datasets, a very useful technique to use to overcome this issue is space-time density, where volume is produced in a three-dimensional space-time continuum. This is a great technique to use to visualize the GTD as it considered as a large dataset. Another technique is self-organizing map which may require data mining techniques prior to aggregate the data before visualizing it. This could be another good technique to use in my case study as clustering could be applied as a data mining technique before using SOM.

I. PROPERTIES OF THE DATA

The Global Terrorism originated from handwritten records from the pinkerton global intelligence service. The GTD contains information on more than 200,000 terror attacks around the world from 1970 through to 2019. Techniques such natural language processing and machine learning models to identify information about terror attacks from 200 million articles. GTD uses an object data structure with variables stored as columns and each incident as a row. Different data types such as strings and integers are present. For example, "country_txt" contains the name of the country as a string and "ishostkid" is a binary column which indicates "1" kidnap or hostage true and "0" for false and "-9" for unknown and "scite3" contains information as text. As for temporal, date is split into columns year, month and day. Some months between 1970 and 2011 are missing and are given the value "0" and likewise for day. This was over come by replace the "0" value with "1" so we could convert the column to datetime stamp. In terms of spatial data latitude and longitude values of the city in which the incident occurred. In terms of the precision of these values the column "specificity" identifies the geospatial resolution with the most specific resolution throughout the dataset is the center of

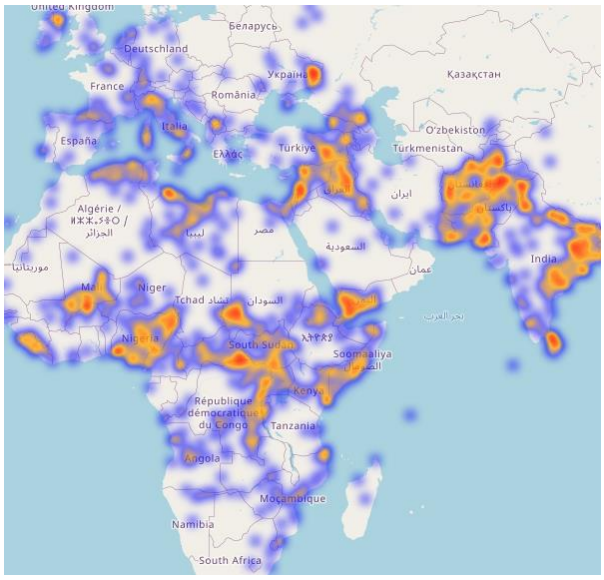
town, village or city in which the attack occurred. We need to investigate the quality of the data, so we start by checking the shape of the data and we have 139 columns and 192212 rows. We filter down to 49 columns which will help us answer our questions. We filter the rows to the incidents that resulted to a kidnap or a hostage only. We then check the data information viewing data types and computing basic statistic.



(Figure 1)

The next step is to concatenate the date columns into one and convert to a timestamp so we can see the temporal distribution. We then look at the temporal pattern using a nicer plot and a kernel density estimate as seen above. We apply smoothing at bandwidth a year in blue and 2 years in red. As we can see the distribution is left skewed with minor peaks through the years but with a dramatic increase from around 2010s and huge peak around 2015. This visual is informing us to potentially split the data to isolate this huge increase to stop it effecting the rest of the data, potentially we can split the tragedies in decades to get a better understanding.

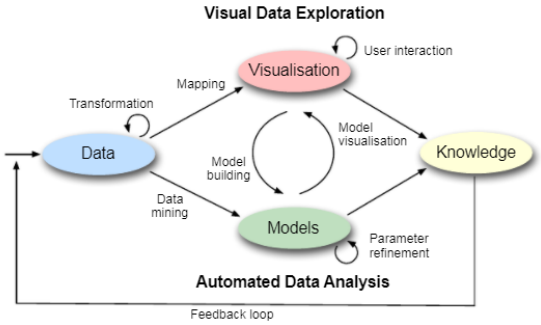
For the spatial variable we build a map seen below which represents each hostage incident using an operation that represents the spatial distribution as a smooth continuous surface. This as a visual can allow us to see the most hit areas around the world such as middle east, Pakistan & India and across central Africa. What we can learn from this is the distribution isn't equally spread across, so potentially splitting the data by regions will stop highly targeted areas influencing the low targeted areas.



(Figure 2)

I. ANALYSIS

A. *Analysis Aproach*



(Figure 3)

We will be adopting a visual analytics process which includes human interaction alongside computational methods to help aid the process. As a result, we will adopt Kiem et al. [4] visual analytics process as seen (figure 3.) which is characterised through interaction between data, visualisations, models, and the users in order to discover knowledge.

The first step in the process is to transform the data, common steps such as cleaning the data, normalising and grouping are part of this step. The next step the analyst can either apply visuals or apply a model to the original data, if the model is created first, visualisations can help the analyst evaluate and refine the model by adjusting parameters or opting for a different algorithm. Model visuals can be used to evaluate the results of a model and alternating between visuals and automatic methods is what characterises the visual analytical process, which leads to a continuous refinement and verification of preliminary results.
If visualization is applied before modelling, the user must confirm the hypotheses by an automated analysis. The user interaction at this step reveals information using visuals which gives insight and knowledge to which helps steer model building. This approach was used where we used temporal distribution plot to give us insight into how data is distributed though time in our data, we then used a spatial distribution map to also display the spread of data over a map. This knowledge was then used to separate the data to stop highly influential periods of time and space to influence the rest of the data. Insight can be gained from exploring using visualisations before and after modelling using human interaction.

The next step is to apply data transformation to prepare the data for analysis. One of the first steps is to aggregate to yearly level to avoid zero, missing and unknown values. As a result of viewing the figure 1 my interaction with the data is to engineer a new feature called "decade" which will aggregate the years for each decade and allow analysis to run on each decade instead of looking at the overall data as whole do to the trend seen already.

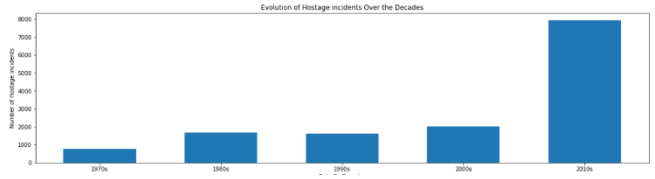
B. *Analysis Process*

The first step is to look at the data and apply transformation if needed if not to visualise the data and see if transformation is needed and repeat these steps to see if we gain insight. Some of these points have been mentioned previously but will iterate as they are part of the analysis process.

Initially we look at the data realise we have too many columns so we view the columns and the variables and decide what columns will help us answer the questions. I have selected that the main variables I will use to help me understand the history is “nkills” which is the number of people killed in the incident and “hostnum” which is the number of hostages or kidnapped in the incident, but I will also use the number of incidents occurred in a year to help me understand the data.

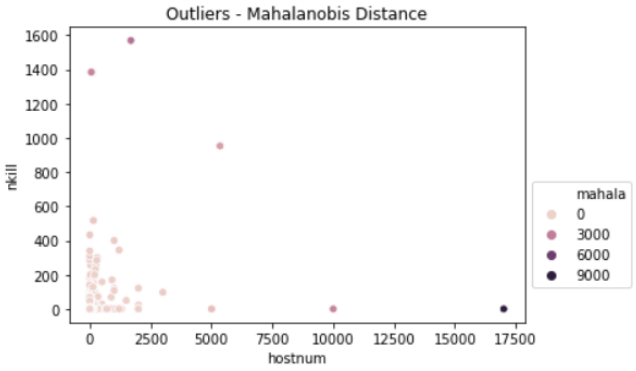
We are visualizing our data as a table to see its content and information about its data types to gain a better understanding, as a result of this interaction we create a new date column concatenating year, month and day using a to_datetime function. The next step is to visualise the null values by summing for each column to gain an overview of how much missing data is present. We then start the process of dealing with the null values by replacing them and filtering to incidents that resulted in a hostages or kidnappings, again an example of using visuals to step back to data transformation.

The next example is moving from data to visualisation to knowledge back to data for transformation. Now that the data is somewhat prepared, I start to visualise it is using graphs to gain a better understanding. The first step is to visualise the temporal and spatial distribution of my data and this is where I used figure 1 & 2 as discussed before in the data preparation section. Like Figure 4 I visualised the temporal distribution of time histograms showing the count of incidents by year time intervals. The insight gained was the later years of the data are significantly greater than the rest of the data so creating a separation will allow us to analyse each period separately without any potential influence. As a result, we go back to data transformation and create a new variable called decade which simply stats which decade the date of the row falls in. We then display this in a time histogram (Figure 4).



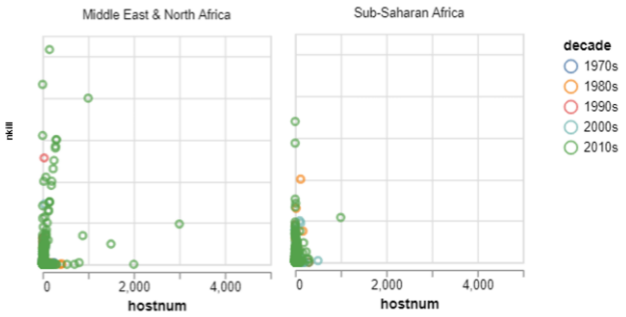
(Figure 4)

Now we have managed to separate all the large values into the decade of the 2010s and can view each decade separately. The next step is to tackle outliers, first we calculate the Mahalanobis distance between our two main variables which are “nkills” and “hostnum”. We then visulaise this in a scatter graph to be able to see the outliers which will be distinguished by being coloured by the Mahalanobis distance scale where the darker the colour the further away the point is and more it is seen as an outlier as in Figure 5. This is where my interaction comes as its my decision which data points I consider as outliers based on the visual and then filter them out.



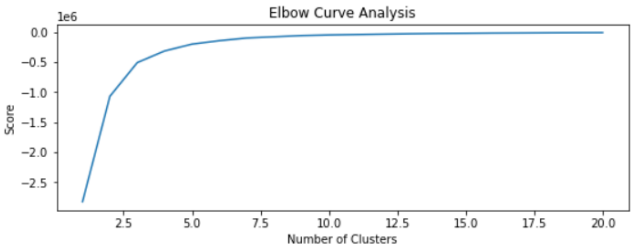
(Figure 5)

Now, the next step for us is to understand the history and the development over the years through the regions. To do this we plot our two variables in scatter plot series by region coloured by decade (Figure 6). The idea is to see if we can see the worst regions and worst decades and if can we group them together. Through my interaction with the visual Figure 7 only shows two regions which have been specifically handpicked as they seem to be the worst hit areas, but we cannot see much except the decade 2010s, which is completely dominant in these regions. As for the other regions it was hard to see which decade was the worst as all the data patterns overlapped and no clear distinction appeared between the regions.



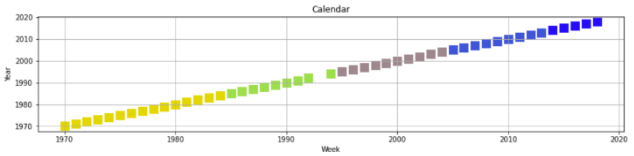
(Figure 6)

It’s through my interaction with the figure above that I decided the visuals don’t fully answer my questions and its best to implement a machine learning model to group the data points. I will be using K-means clustering model which is a distance-based algorithm. Before applying the model, we must scale our data as its distance based. After applying the model to 6 clusters, I found an uneven spread of the data especially in one cluster that was significantly smaller than the others. This led me to believe the number of clusters was not suitable for our data. I decided to produce an elbow curve which would help me visualise the optimum number of clusters and this was found to be 5 clusters as seen in Figure 7.



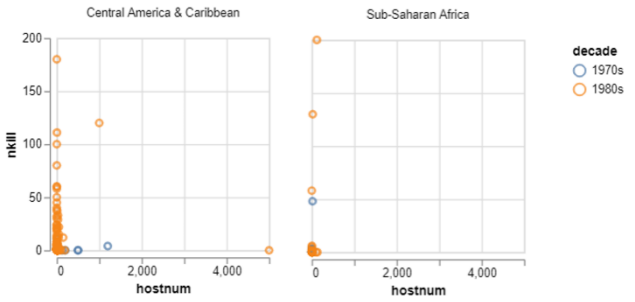
(Figure 7)

Figure 8 was plotted to help account for the cyclic organisation of the time and clusters in our data.
We create a 2D plot where both dimensions correspond to the years. Each year interval is represented by a mark painted in the colour of the cluster this interval belongs to. This plot beautifully shows how the algorithm has grouped the data points through time and indeed shows that the idea of using just decade to explain the history wouldn't of been the best way to group the data as some cluster have data from two different decades except one which has some data from the decade 2010s, coloured in dark blue.



(Figure 8)

I go back and refilter the data for each cluster and viuaslise it is using the same method as for Figure 6. As we can see in Figure 9, we start to unravel new insight which was not seen in the previous iteration of plots for each region. In this iteration we can see that in the 1970s & 1980s the worst hit regions were Central America & Caribbean and Sub-Saharan Africa, which was insight we gained from looking closely at cluster 0.



(Figure 9)

Now that we are happy with insight from the clusters in which they answer questions of the worst hit regions during what decades and being able to cluster their respective incidents, we turn our attention to answer the most common motives and target type for each cluster. This is achieved by using a word cloud on each column to produce the most common text, the size of the word amounts to the frequency it appears, Figure 10 is the example of target type for cluster 0.

C. Analysis Results

(Figure 10)

The summary of our results shows that for:

- Cluster 0 - Worst hit areas were Central America & Caribbean and the Sub-Saharan Africa for the decades 1970s and 1980 with the most common motives as "Specific Motive" and "Government" related, and the most common target type was "Private Citizen".
- Cluster 1 – Worst areas were Middle East & North Africa and Sub-Saharan Africa for the decades 2000s and 2010s with the most common motive as "Targeted" attack and the most common target was ""Citizens Property".
- Cluster 2 – Worst areas were Middle East & North Africa and Sub-Saharan Africa for the decade 2010s with the most common motive as "Islamic State" and "Levant ISIL" and the most common target was "Property Private".
- Cluster 3 – Worst areas were South America and Southeast Asia for the decades 1980s and 1990s with the most common motive as "Protest" and "Abortion" and the most common target was "Government General".
- Cluster 4 – Worst areas were Middle East & North Africa and Eastern Europe for the decades 1990s and 2000s, most common motive as "Government" and the most common target was "Property Business"

II. Critical Reflection

Kiem et al visual analytics process is well structured general methodology that can be applied to any project or answering any questions especially with a question is which you are discovering patterns. The process is versatile in allowing the analyst to be able to iterate through the same steps to uncover more insight to help steer the analysis in the right direction to uncover the answers. For example, the feature engineering process of creating a new variable called decade allowed us to aggregate data but to also represent it in a more familiar way to the end user which makes for a more intriguing read.
We were successfully able to answer the questions of the project, but as the results unfolded, I found myself wishing to further investigate the fascinating results. For example, investigating the clusters in more depth to be able to profile the severity of each cluster.
The visuilisation played a key role in helping steer the direction of the project and were fundamental in every decision I took; without them I don't believe the results I achieved would have been as insightful.

The data was a great source but had some flaws, after exploring I found a significant amount of missing data, sometimes entire columns. This can be a huge issue especially in reporting a topic such as terrorism as you do not wish to paint an inaccurate perception. Not only did we have to deal with missing values, but most columns had a value which represented unknown, such as "-99". This was a slight issue as it reduced the amount of data and would have affected the numbers if not noticed initially. I did not know how to deal with replacing this data as I didn't wish to affect the final numbers, I simply dropped these rows which may have not been the best approach.

Selecting only two variables to look at I realise now that I limited my approach in the analysis. I could of have included a lot of more useful variables to enrich the outcome of my results and used more advance techniques such as MDS or even used existing variables to calculate a more advance variable. Looking back, I would have loved to have taken a different direction with this analysis and maybe investigated density-based clustering and created terror zones, I feel I didn't take full advantage of the spatial variables. I do feel my approach can be applied to any other project or discipline as it uses generic methods and the analyst's interaction. In terms of this dataset I would advise future users to focus on maximizing the use of all the data variables and minimizing the loss of null values and unknown values.

I. REFERENCES

[1] S. Tutun, M. T. Khasawneh, and J. Zhuang, "New framework that uses patterns and relations to understand terrorist behaviors," Expert Systems with Applications, vol. 78, pp. 358–375, 2017.

[2] Andrienko G, Andrienko N, Demsar U, Dransch D, Dykes J, Fabrikant S I, Jern M, Kraak M J, Schumann H, Tominski C. Space, time and visual analytics. International Journal of Geographical Information Science, 2010, 24(10): 1577-1600.

[3] https://www.start.umd.edu/gtd/

[4] Keim D A, Kohlhammer J, Ellis G, Mansmann F. Mastering the Information Age: Solving Problems with Visual Analytics. Florian Mansmann, 2010.

II. WORD COUNT

- Overall - 3512
  - ✓ Problem Statement - 176
  - ✓ State of the Art - 499
  - ✓ Properties of the Data- 499
  - ✓ Analysis – 1798
    - Analysis Approach - 476
    - Analysis Process - 1283
    - Analysis Results - 198
  - ✓ Critical Reflection - 456