Quantification and Prediction of Terrorist Attack Risk Under Geographic and Temporal Dimensions

Liu Wenting*
liuwt@shanghaitech.edu.cn
ShanghaiTech University
Shanghai, China

Chen Xinhe
chenxh1@shanghaitech.edu.cn
ShanghaiTech University
Shanghai, China

ABSTRACT

Global terrorism poses a severe threat to societal stability and human safety. Terrorist attacks result in casualties and property damage, and can also trigger political, economic, and humanitarian crises. Therefore, effective prediction and prevention of terrorist activities are crucial. To better predict terrorist attacks and understand the underlying patterns, we first utilize DBSCAN clustering to divide regions into different areas, allowing the model to operate in the clustered space. Subsequently, we construct models for time series prediction, designing a linear model as a baseline and training with LSTM and CNN-LSTM models, incorporating temporal attenuation effects to capture both temporal and geographical propagation. Finally, we map the results back to the grid space, aligning with international standards in scholarly papers.

Our advantage lies in training the model based on national geographic units rather than the coarse continental scale, making our model more targeted. Additionally, we are the first to propose a clustering method, clustering the problem before inputting it into the model. This approach allows us to focus on geographical properties rather than directly using the grid as a unit. Moreover, our model introduces equidistant time series encoding, enabling precise predictions on a fine temporal scale. Our model also emphasizes the modeling of the spread of terrorist attacks, integrating propagation into machine learning methods for the first time.

ACM Reference Format:
Liu Wenting and Chen Xinhe. 2024. Quantification and Prediction of Terrorist Attack Risk Under Geographic and Temporal Dimensions. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Terrorism poses a significant threat to global society. According to data from the Institute for Economics and Peace (IEP), in 2021 alone, terrorist incidents resulted in 7142 deaths and economic losses totaling billions of dollars worldwide. These activities predominantly occur in the Middle East and other parts of Asia, prompting many researchers to model the evolution process and spread of terrorism in these regions.

In previous studies, machine learning methods have been employed to address this issue. However, existing methods still have some shortcomings that need to be addressed. For example, some studies broadly consider data from entire continents, overlooking the uneven distribution of terrorist attacks; while others can only determine whether a terrorist attack will occur, lacking accurate predictions of its severity. Therefore, our optimization goal is to establish a more refined model and achieve a predictive system with warning capabilities. Additionally, there are currently no machine learning methods specifically designed to model the spread of terrorism, although some mathematical models attempt to model propagation, they are cumbersome and computationally inefficient. Therefore, we aim to first conduct intuitive validation and then construct machine learning methods specifically targeting propagation to more efficiently find optimal solutions.

We conducted two aspects of propagation validation: we found that propagation is closely related to time, with similarity decreasing gradually over time, indicating that the influence of an event diminishes over time. We found that the propagation is also related to the region. As within a cluster, the similarity of events is significantly higher internally than externally, indicating that events are more likely to propagate within the same cluster. Moreover, when an event occurs near a cluster, it exhibits stronger propagation compared to clusters that are farther away. These intuitive validation results lay a reliable foundation for the methods proposed in our subsequent work.

Specifically, this paper encompasses the following steps. Firstly, we employ the DBSCAN clustering algorithm to partition the region into different sub-regions. The objective is to aggregate similar results, thereby avoiding overly coarse divisions. Subsequently, we design five models for prediction, including a linear model, an LSTM model, an LSTM with time decay model, a CNN-LSTM model, and a CNN-LSTM with time decay model, aimed at capturing the temporal and geographical propagation patterns of terrorist attacks. The output of the models indicates the severity of terrorist attacks in each cluster for each day. Finally, we map the spatial clusters derived from the model outputs to a grid space, resulting in a gridded map that illustrates the occurrence of terrorist attacks in each cell.

2 RELATED WORK

In previous studies, efforts have been made to address the issue of terrorism detection. For instance, research has focused on categorizing pro-terrorism tweets and investigating attribution for terrorism [1].

*Both authors contributed equally to this research.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Currently, researchers have made promising initial progress in predictive analytics. For example, Python et al. [9] conducted a study where they utilized previous terrorism data (from GTD) along with other geographical and socio-economic features to train several machine learning models for predicting attacks with discrete spatiotemporal gaps. While their work demonstrated the feasibility of machine learning methods in prediction tasks and provided comprehensive analysis globally, their approach suffers from two fundamental issues. Firstly, they roughly evaluated a subcontinent region (such as West Africa), overlooking the imbalanced spatial distribution of attacks. Secondly, apart from autoregression, their models did not introduce other time-related inputs, making it challenging to make rigorous predictions (e.g., whether a terrorist attack will occur within a given number of days). Therefore, our optimization lies in: firstly, training models on a country geographical unit basis rather than roughly on a continent. Secondly, standardizing the time granularity, constructing equidistant time series encoding, and making precise predictions on a fine time scale.

Building upon the work of Python et al., Krieg et al.[6] proposed a novel variable-length moving average feature representation method and extensively integrated local news data from the United States. The study modeled the problem as a binary classification problem of 0/1 variables, indicating whether a terrorist attack would occur on a given date and state, and implemented various machine learning methods accordingly. We employed a similar time-series encoding approach, specifying time down to the day for prediction purposes. However, we also recognized certain issues. J. Krieg et al.[6] used states as prediction units. We aimed to further refine the granularity, thus dividing the map into 0.5-degree latitude and longitude grids.

Both teams' approaches utilized grids as the unit of analysis, overlooking the potential occurrence of terrorist attacks along highways or at the edges of deserts. Thus, rather than employing grids as the unit, we initially clustered terrorist attacks based on spatial information, thereby mitigating the loss of geographical information incurred by grid partitioning. Subsequently, we mapped the model's output clusters back onto a gridded map.

Furthermore, currently, there are no machine learning methods emphasizing modeling for spread. We have observed some papers in the mathematical domain have already modeled propagation with differential equations. Clark et al. utilized spatial temporal hierarchical modeling to model the spread of terrorism data in a given area and estimated diffusion parameters. Gill et al. [5] applied this model to actual scenarios of terrorist attacks with improvised explosive devices in Northern Ireland, achieving the first concrete application of terrorist attack scenarios. However, mathematical modeling methods are labor-intensive, solving equations is very cumbersome, reusability is poor, and manual modeling is incomplete, only able to input individual numerical variables. Therefore, we aim to intuitively verify the spread and apply it in machine learning methods more efficiently and concisely to automatically find the optimal function. In addition, the spread of differential equations is conducted with neighboring variables, which cannot fully simulate distant spread, whereas the advantage of our method is that each cluster can spread to more distant locations, corresponding to actual situations.

3 PROBLEM STATEMENT

Firstly, before addressing the modeling problem, let us restate the issue discussed earlier with mathematical variables. We begin by partitioning the region into p areas using DBSCAN clustering of terrorist attack locations. We want to determine whether terrorist attacks will occur in each area separately at time t, and if so, what level of harm they will cause.

We aim to train a model f with parameter θ:

$$f : \{H_t, \theta\} \longrightarrow Y$$

where H_t represents a series of features encoded in chronological order, obtained from a time series of m days ($h_{t-1}, h_{t-2}, ..., h_{t-m}$). Each h consists of p components, representing the terrorist attack situation in p areas on day t. Y represents the prediction results for all areas. When a terrorist attack does not occur in area i, $Y_i = 0$; otherwise, Y_i is a positive number in the range $(0, 1)$, indicating the degree of danger of the terrorist attack.

After that, we would use a map

$$g : Y \longrightarrow Y_{grid}$$

to align the results on the grid to make it more prescriptive

4 EXPERIMENT METHODOLOGY

4.1 Data exploration

4.1.1 Definition of Propagation. Propagation here refers to the effect that one event triggers the occurrence of other events. This effect decreases over time and diminishes with increasing distance between classes.

4.1.2 Feasibility of Temporal Propagation. First, we conducted validation of propagation observing a strong correlation between propagation and time.

We constructed event pairs for Syria from 2010 to 2022, excluding geographical factors, to calculate their similarity. The horizontal axis of the graph represents the time difference between event pairs, while the vertical axis represents the cosine similarity of the event pairs.

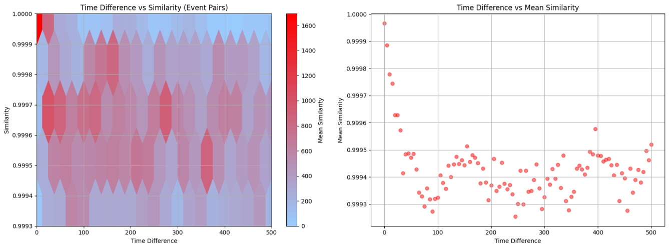


Figure 1: The left plot represents the distribution of terrorist attack event pairs composed from the Iraq region between 2013 and 2014. The horizontal axis denotes the time difference (days) between event pairs, while the vertical axis indicates the cosine similarity of the event pairs. The right plot shows the column-wise average of event pairs with the same time difference.

During the cosine similarity calculation process, we selected the following variables: 'attacktype1', 'targtype1', 'targsubtype1', 'ransom', 'nkill', 'nwound', 'property', and 'weaptype1'.These variables were chosen due to their importance and representativeness in describing terrorist attacks. They are all numerical, facilitating their processing and analysis within the model. Encompassing various aspects of terrorist attacks, these variables include the type of attack, target, ransom situation, casualties, property damage, and weapon type. By considering these aspects, we can gain a more comprehensive understanding and description of the nature, motives, and impacts of terrorist attacks. Therefore, the selection of these variables aims to provide as comprehensive a depiction as possible of terrorist events, thereby laying a more profound and accurate foundation for our understanding and analysis of the phenomenon of terrorism.

4.1.3 *Feasibility of Inter-Area Propagation.*

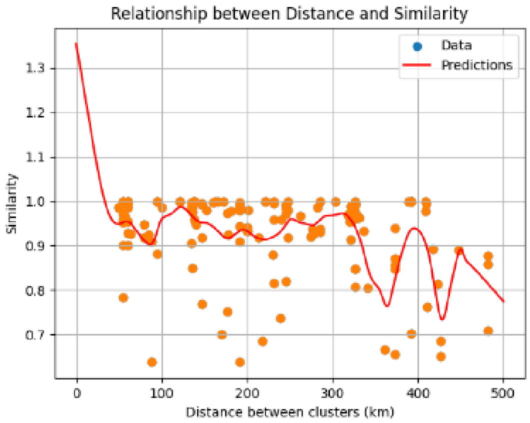


Figure 2: It examines the relationship between the average similarity of inter-cluster events in the Iraq region from 2012 to 2020 and the difference in distances between cluster centroids. The horizontal axis represents the difference in distances between cluster centroids (in kilometers), while the vertical axis represents the average similarity of events between clusters. The graph is fitted using Locally Weighted Linear Regression.

Validation 1: Within a cluster, events propagate more frequently internally than externally.

This validation aims to assess whether internal propagation of events within a selected cluster is more frequent than cross-cluster propagation. The process involves selecting a cluster with a significant number of events, identifying an early event within that cluster, and calculating the average similarity between the chosen event and internal nodes (i.e., other events within the same cluster) and external nodes (i.e., events not within the same cluster) over a 100-day time window.

Through comparing the average similarity between the reference event and events within and outside clusters, we found that the average similarity within clusters is significantly higher than that outside clusters. This indicates that events propagate more efficiently within clusters.

Validation 2: When an event occurs, it exhibits more propagation towards nearby clusters compared to distant clusters.

Through comparing the average similarity between the reference event and multiple other events from short-term clusters, it was observed that the average similarity between the reference event and events within clusters closer in distance is significantly higher than that with clusters farther away. Specifically, as the distance between the centroids of two clusters increases, the similarity decreases, with a probability of 0.7619047619047619 for this scenario. This indicates that terrorist attacks propagate more efficiently towards clusters that are closer in distance.

Table 1: Comparison of Average Internal and External Similarity

Average Internal Similarity	Average External Similarity
0.908	0.878

4.2 Preprocessing

4.2.1 *Clustering.*

The feasibility of employing clustering over grid methodology stems from the observation that grid methods may overlook certain spatial distributions of events. While grid methods are capable of identifying spatial patterns, they may fail to capture non-uniform distributions such as elongated or clustered patterns, as illustrated in 3.

As a result, we opt for clustering methodology, which offers greater flexibility in identifying spatial clusters and patterns. Clustering methods intuitively provide a more reliable approach for identifying spatial clusters, allowing for a nuanced understanding of the spatial dynamics of terrorist incidents.

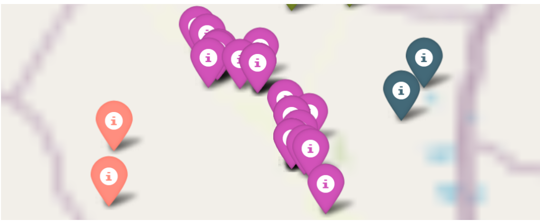


Figure 3: Some cluster distribution in Iraq, showing a long strip.

Therefore, we utilize the DBSCAN method, and the clustering results for the Iraq region are illustrated in 4.

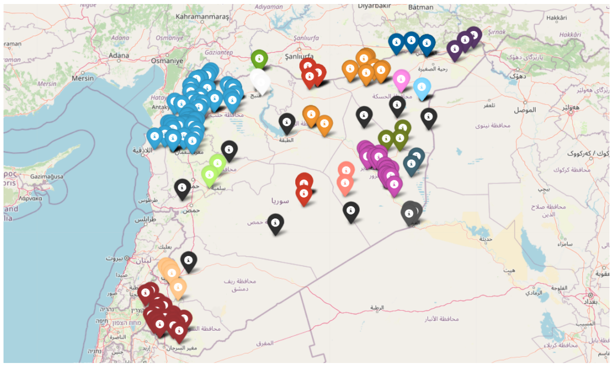
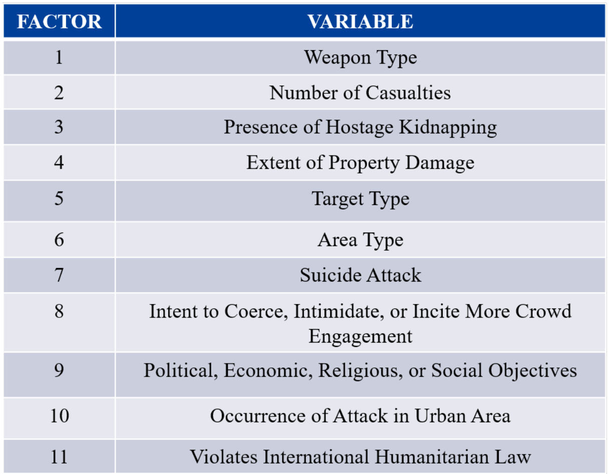


Figure 4: The results after applying DBSCAN to the Iraq region. The illustration depicts the clustering results of terrorist attacks in the Iraq region from 2013 to 2022 after executing DBSCAN (ϵ=0.15, min_samples=2). Black points represent noise points. The clustering performance metrics are as follows: Within-Cluster Sum of Squares (WSS) = 15687.298 and Between-Cluster Sum of Squares (BSS) = 33505.823 (measured in square degrees of latitude and longitude).

4.2.2 *Risk Quantification.*

We need a fixed numerical value to succinctly represent each terrorist attack event, making it convenient to input data into the model. Many papers use 0 and 1 to simply indicate whether a terrorist attack event will occur, which may result in loss of some information.

We adopted parameters from a certain person's paper and calculated the severity of each event using a weighted method. Additionally, we validated this method using the globally recognized top ten terrorist attacks, and the results proved to be reasonable.



FACTOR	VARIABLE
1	Weapon Type
2	Number of Casualties
3	Presence of Hostage Kidnapping
4	Extent of Property Damage
5	Target Type
6	Area Type
7	Suicide Attack
8	Intent to Coerce, Intimidate, or Incite More Crowd Engagement
9	Political, Economic, Religious, or Social Objectives
10	Occurrence of Attack in Urban Area
11	Violates International Humanitarian Law

Figure 5: The 11 variables used in Li's paper.

$$\text{Risk degree}_{\text{m}} = \text{w}_1\text{F}_{1\ \text{m}} + \text{w}_2\text{u}_{2\ \text{m}} + \ldots + \text{w}_\text{n}\text{u}_{\text{nm}}$$

Here Risk degree_m represents the severity or harm level of the m^{th} event. $\text{w}_1, \text{w}_2, \ldots, \text{w}_\text{n}$ are the parameters we are referring to or using in our model. $\text{F}_{1\ \text{m}}, \text{u}_{2\ \text{m}}, \ldots, \text{u}_{\text{nm}}$ represent the features of the m^{th} event. For example, $\text{F}_{1\ \text{m}}$ refers to the first feature of the m^{th} event.

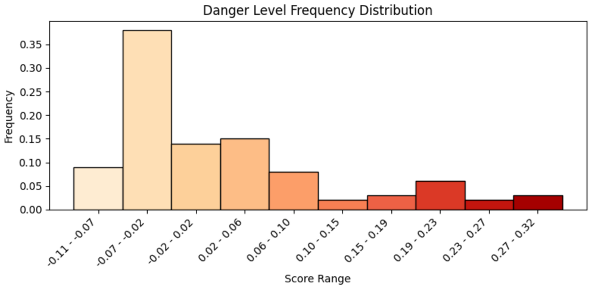


Figure 6: A frequency plot is generated for 100 randomly sampled terrorist attack events, calculated for severity using a weighted approach. The top ten globally recognized terrorist attacks are placed in the last cluster, and their values are distinguishable from those of the randomly sampled other terrorist attack events.

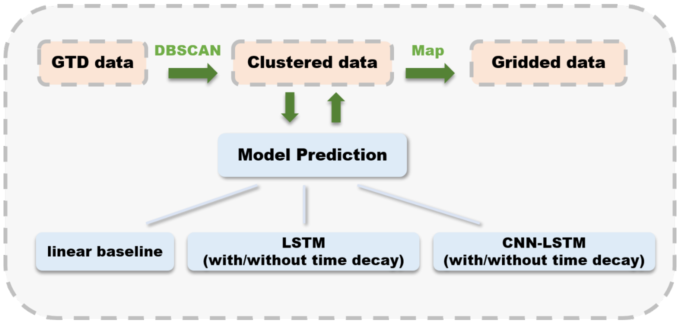


Figure 7: Our task flowchart. First, we apply the DBSCAN method to the GTD data to obtain clustered data. Next, we perform predictions based on this clustered data. Finally, we map the returned clustered data onto the grid.

4.3 Models

4.3.1 *Baseline.*

We linearly combine historical time series data to generate predictions for the following day's conditions. In this linear model, the input dimension is the length of the time series multiplied by the number of clusters, and the output dimension equals the number of clusters. The dimensions of input and output determine the size of the W matrix.

$$Y = WX + b$$

4.3.2 *LSTM.*

In this model training process, the batch size is set to 5, epochs to 100, and the size of the hidden layer is 64. A fully connected layer is employed at the end. During training, 10% of the validation set is used for validation. The loss function is implemented through a custom method, which calculates the absolute error between the predicted values and the true values for each sample. The total error is then computed as the weighted sum of errors for all samples, where the error for positive samples is taken as the original value and for negative samples as half of the original value.

4.3.3 *LSTM on time attenuation.*

We conducted tests on the LSTM model and found that the predicted locations tended to be more numerous than the actual ones. We attribute this to excessive propagation of distant historical events. Therefore, we aim to attenuate the weighting of historical events by multipling a time attenuation coefficient as follows:

$$attenuation\ parameter = e^{-a(t_{predict}-t)}, a > 0$$

The further away the date of an event is from the day to be predicted, the more it is attenuated.

We first apply attenuation processing to the dataset. Then, we input the data into a model with the same architecture parameters as in 4.3.2.

4.3.4 *CNN-LSTM.*

This model is a Sequential neural network architecture comprising a Conv1D layer with 64 filters and a kernel size of 3, activated by ReLU, followed by a MaxPooling1D layer with a pool size of 2. Subsequently, two LSTM layers are employed, the first with 64 units , and the second with 32 units. The output layer is a Dense layer with dimension p, representing predictions for p locations..The error function remains the same as before.

4.3.5 *CNN-LSTM on time attenuation.*

Similarly, we apply a attenuation method similar to that in section 4.3.3, and then input it into the CNN-LSTM model in section 4.3.4

4.4 Grain size alignment

In our model, computations are performed at the cluster level, and the outputs are also presented in the form of clusters. However, we aim to grid the model outputs to align with international standards, as mainstream journals typically adopt grid-based alignment.

Therefore, throughout the process, after obtaining the model output results based on clusters, we further map them onto the corresponding grid.

Our grid granularity is set to 0.5 degrees, with the longitude range from 35 to 43 and the latitude range from 32 to 38. This results in a total of 192 grid cells. We utilize a dictionary approach to establish the mapping from clusters to grids. Typically, one cluster may map to 1 to 4 grid cells.

5 RESULTS

5.1 Training and Testing Environment

We selected events from the Syrian region spanning from 2013 to 2022, excluding null and duplicate values. These events were then categorized into 86 distinct regions. Subsequently, sequences of length 30 were chosen as inputs, predicting the occurrence (0,1 variable) in the 86 regions for the following day. The dataset was split into training, validation, and testing sets in an 8:1:1 ratio.

5.2 Accuracy and Recall Table

Table 2: Comparison of Recall and Accuracy

Method	Precision	Recall
①baseline	0.428	0.575
②LSTM	0.163	0.695
③LSTM + time attenuation	0.169	0.715
④CNN-LSTM	0.131	0.75
⑤CNN-LSTM + time attenuation	0.159	0.77

The results of model testing show that the baseline method has the highest precision and the lowest recall because it predicts fewer positive samples each time, thus making fewer false positive errors. ③ compared to ②, and ⑤ compared to ④, both show a certain degree of improvement in precision, indicating that time decay can reduce the proportion of false positive samples to some extent.

④ compared to ②, and ⑤ compared to ③, both show an improvement in recall, but at the same time a decrease in precision. This indicates that CNN is a structure that can help capture positive samples but also increases the inaccuracy of positive predictions.

6 MAPREDUCE

6.1 Scenario for Utilizing MapReduce

The choice of employing MapReduce stems from the necessity to tackle a task that is both straightforward, parallelizable, thereby leveraging the inherent characteristics of MapReduce. Our current objective revolves around assessing the severity distribution of events within the GTD dataset. To achieve this, we have already assigned scores ranging from 0 to 1 to all analyzable events. Now, our focus shifts to computing the frequency of each interval, ranging from [0,0.1) to [0.9,1), in order to gain deeper insights into the distribution of event severity. The dataset encompasses a substantial size, consisting of 2,096,710 rows.

6.2 Logic

Firstly, we adopted the "map to range" approach, where we utilized the intervals as keys and assigned a value of 1 to each event, thereby constructing key-value pairs for all events. Subsequently, we aggregated these pairs based on their keys, employing a merging strategy wherein the values are summed together.

6.3 Running version and environment

Java version 11.0.22 and Scala version 2.12.15 are utilized. The execution process consists of two stages. Spark properties are configured with default settings.

6.4 Enhancement

Through comparing the performance with the regular method (without MapReduce), we observed a significant enhancement in efficiency. Following the implementation of MapReduce, the execution time decreased from 4.489758253097534 seconds to 3.724113702774048 seconds, resulting in a notable improvement of 17.05%.

REFERENCES

[1] Mohammad Fahim Abrar, Mohammad Shamsul Arefin, and Md. Sabir Hossain. 2019. A Framework for Analyzing Real-Time Tweets to Detect Terrorist Activities. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). 1–6. https://doi.org/science/article/abs/pii/S0306457316301571

[2] Arastoo Bozorgi, Hassan Haghighi, Mohammad Sadegh Zahedi, and Mojtaba Rezvani. 2016. INCIM: A community-based algorithm for influence maximization problem under the linear threshold model. Information Processing Management 52, 6 (2016), 1188–1199. https://doi.org/10.1016/j.ipm.2016.05.006

[3] Gian Maria Campedelli, Mihovil Bartulovic, and Kathleen M. Carley. 2021. Learning future terrorist targets through temporal meta-graphs. Scientific Reports 11, 1 (Apr 2021), 8533. https://doi.org/10.1038/s41598-021-87709-7

[4] Nicholas J. Clark and Philip M. Dixon. 2018. Modeling and Estimation for Self-Exciting Spatio-Temporal Models of Terrorist Activity. Annals of Applied Statistics 12, 1 (March 2018), 633–653. https://doi.org/10.1214/17-AOAS1112

[5] Paul Gill, John Lee, Karl R. Rethemeyer, et al. 2014. Lethal Connections: The Determinants of Network Connections in the Provisional Irish Republican Army, 1970–1998. International Interactions 40, 1 (2014), 52–78. https://doi.org/10.1080/03050629.2013.863190

[6] Steven J. Krieg, Christian W. Smith, Rusha Chatterjee, and Nitesh V. Chawla. 2022. Predicting Terrorist Attacks in the United States using Localized News Data. CoRR abs/2201.04292, 6 (2022), 1188–1199. arXiv:2201.04292 https://arxiv.org/abs/2201.04292

[7] Michael D. Porter and Gentry White. 2012. Self-exciting hurdle models for terrorist activity. The Annals of Applied Statistics 6, 1 (2012), 106 – 124. https://doi.org/10.1214/11-AOAS513

[8] Michael D. Porter and Gentry White. 2012. Self-exciting Hurdle Models for Terrorist Activity. The Annals of Applied Statistics 6, 1 (2012), 106–124. https://doi.org/10.1214/11-AOAS513 GeoEye Analytics and University of Queensland.

[9] A. Python, A. Bender, A. K. Nandi, P. A. Hancock, R. Arambepola, J. Brandsch, and T. C. D. Lucas. 2021. Predicting non-state terrorism worldwide. Sci Adv 7, 31 (Jul 2021), eabg4778. https://doi.org/10.1126/sciadv.abg4778