TERRORISM IN THE UNITED STATES

KOFFI PAULE BESSIN, George Mason University
PINZHOU LIANG, George Mason University
YUZHI SHENG, George Mason University
YIJING WANG, George Mason University

KEYWORDS
Terrorism, United States, Perpetrators, Attacks, Analysis, Trends, K-means Cluster, time-series analysis, Prediction, Results

1 Introduction

For decades, populations in the United States have been experiencing life-threatening and deadly attacks within each state of the country. Our research focuses on the analysis of the rate of terrorism in the United States over the years, study the factors impacting terrorist attacks-specifically perpetrators' involvement, predict and prevent future occurrences. It is really important to help solve this issue for a stronger economy, to bring stability, safety, prevent possible attacks, and reduce lethality.

2 Methodology and key components of our approach

Our leverage approach focuses on studying and analyzing perpetrators' organizations, understand their motives and goals, evaluate the damages these organizations occasioned, and study any predictors to prevent future attacks. For a steady analysis, we combined The Global Terrorism Database (GTD) and the Profiles of Perpetrators of Terrorism in the United States (PPT-US) datasets. Merging, cleansing, transforming, and analyzing both datasets was effective using R.
Several variables have been discarded, considering their low importance for the purpose of our analysis and a large amount of missing data. Statistical algorithms and models such as ggplot, plotly, wordcloud, k-means, ARIMA allowed us to analyze, visualize, and interpret discovered patterns, trends, and predict for decision making.

3 Metrics

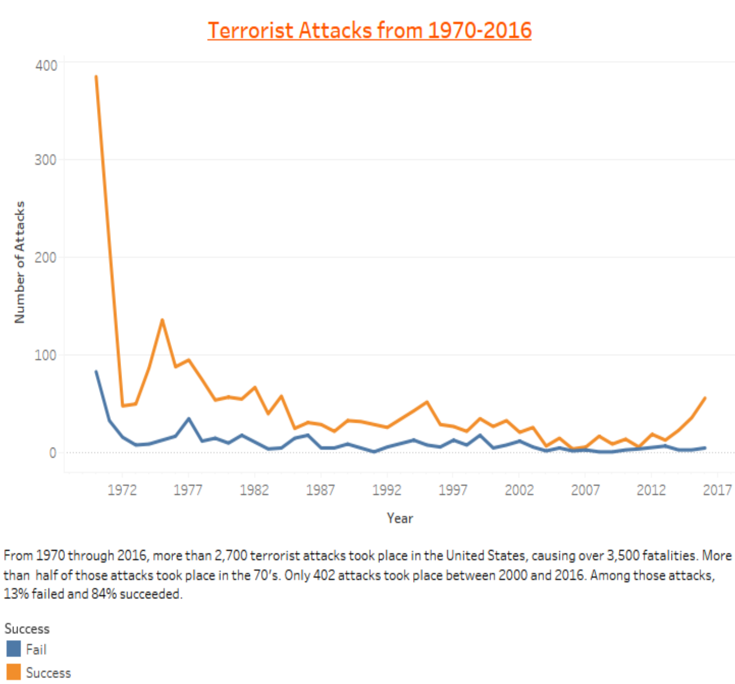


Figure 1: Terrorist Attacks from 1970-2016

3.1 Attack Trends

From 1970 through 2016, more than 2,700 terrorist attacks occurred in the United States. From those attacks, we observe a remarquable peak frequency of attacks in 1970. More than 380 successful attacks were carried out. From that standpoint, there is decline in the early 70s-under 100 attacks. After a slight increase in the mid 70s, attacks continue to decrease consistently. From 2000 to 2013, the frequency of terrorist attacks has considerably dropped to an average of 20 per year. Activities start picking back up in 2014 and keep increasing until 2016. Failed attacks- those that could not be accomplished either due to authorities' prevention or unsuccessful attempts- were also accounted. Almost 100 unsuccessful attacks occurred as well in 1970. There is a relatively low frequency of failed attacks as opposed to the successful ones.

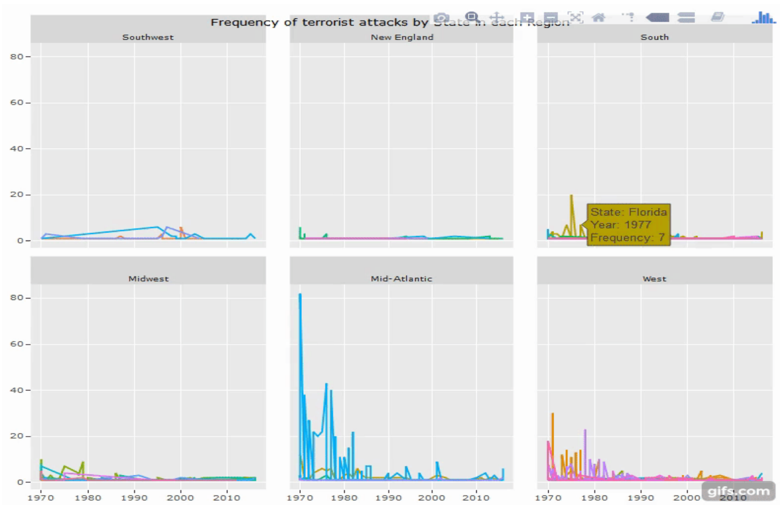


Figure 2: Frequency of Terrorist Attacks by State in Each Region

3.2 Attacks per Region

As observed above, most of attacks took place on the West, Mid-Atlantic, and some on the South regions.

Mid-Atlantic- From 1970 to 2016, New York has been one of the State where a lot of attacks occured. The peak of these attacks was in 1970 as mentioned before.New York by itself has over 480 attacks.

West- California, follows the trend with over 390 attacks.

South- Florida with 139 attacks. Puerto Rico accounts for 189 attacks.

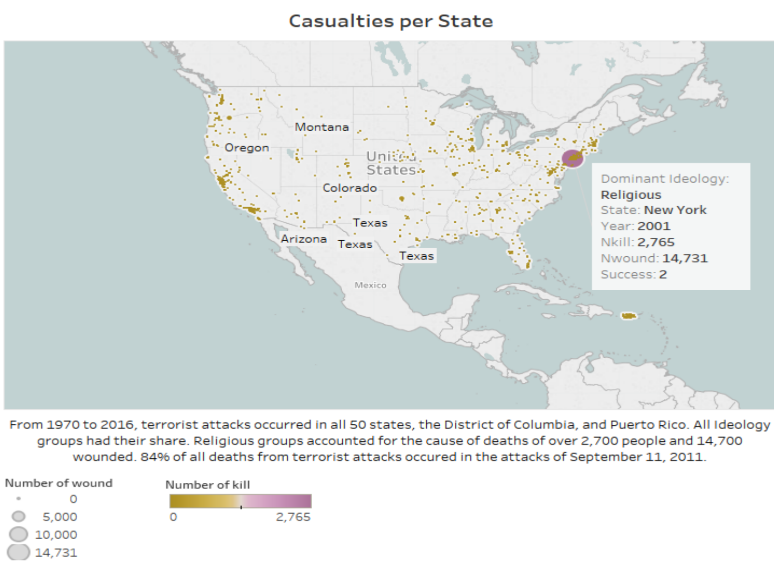


Figure 3: Lethal Attacks per Year and Dominant Ideology

3.3 Lethality

Overall, we account for over 3,500 deaths caused by these attacks from 1970 to 2016. The remarquable part is that 14,731 injuries and 84% of all deaths in the United States resulted in the attacks of September 11, 2001 perpetrated by Religious Ideology. All 50 states were victims of attacks including the District of Columbia and Puerto Rico.

As mentioned above, we observe clusters of many attacks on the West, Mid-Atlantic Regions, and Puerto Rico.

Figure 4: Main Motives of Organizations

3.4 Perpetrators' Motives

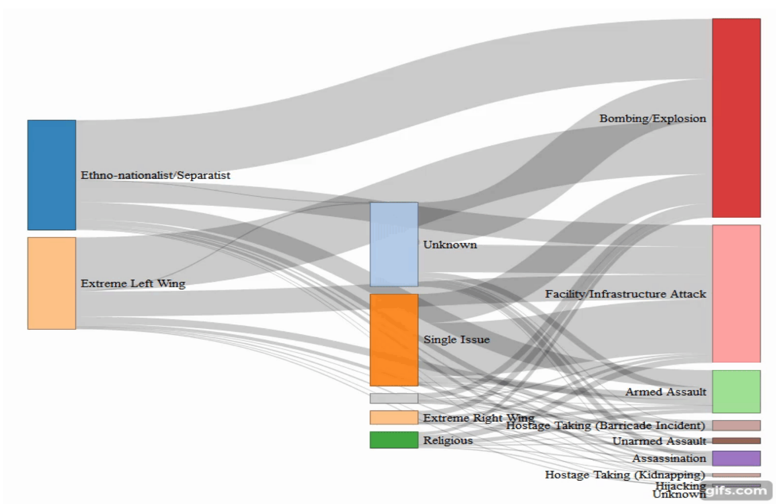
In the context of exploring perpetrators of terrorist attacks, we can infer that for more than 80 percent of the attacks that took place in the United States, 65% of these attacks were attributed to 162 named organizations.

To understand the logic behind these attacks, we examined the motives of these organizations using wordcloud algorithm. The most frequent reason and strongest motive is protest. Most of these organizations clearly intended to claim some values, policies, and more, followed by abortion. Abortion is the second strongest reason behind terrorist attacks. We observe a strong frequency of unknown reason to engage in terrorism. Sabotage is also one pertaining motive for terrorist attacks.

3.5 Recruitment Strategies

Our purpose after analyzing the motives of these organizations, is to study elements that could favor more attacks. Although our analysis revealed the recruitment processes for some organizations, the majority of the organizations' processes was unknown. Using the information available, 152 successful attacks-(~6% of total attacks)-in total were identified with a recruitment strategy. Looking at the strategies these organizations use to recruit adherents, it reveals that all ideologies groups adopt The Net process to get more recruits-74% (114) of these attacks identified. The Infection comes next. Looking for a pattern, it is too arbitrary to infer or deduce the impact of the recruitment strategy on the occurrence of an attack.

Figure 5: Measure of Successful Attacks based on Recruitment Strategies by Dominant Ideology



3.6 Dominant Ideologies- Attack Type and Targets

When we ignore the unknown ideologies, we observe three main ideologies that stood out the most. The sankey diagram shows the nodes of the ideologies most dominant type of attacks. Approximately 50% (327) of the attacks driven by Ethno-nationalist/Separatist were of bombing and explosion type. 60% of the attacks perpetrated by Extreme Left Wing were of bombing type as well, unknown ideologies had about 50% of their attack types be of type bombing/explosion. Single Issue on the other hand attacked mostly Facilities/Infrastructure. All other ideologies' attack types are spread out among armed and unarmed assaults, hijacking, kidnapping, barricade, etc.

Figure 6: The Ideologies Most Dominant Type of Attacks

It was unexpected to find out that those Bombing/Explosion and Facility/Infrastructure attacks caused by the top three dominant ideologies, targeted mainly businesses, and government. Abortion related, private properties, and citizens have been at stake as well during the attacks. Business targets were most commonly banks/commerce (32%), retail (25%), or multinational corporations (11%). About 60% of Government targets were federal, state, local, and tribal government buildings and facilities. They included the IRS, FBI, the Pentagon.

Figure 7: Main Target Type

3.7 Most Active Groups and the weapons used

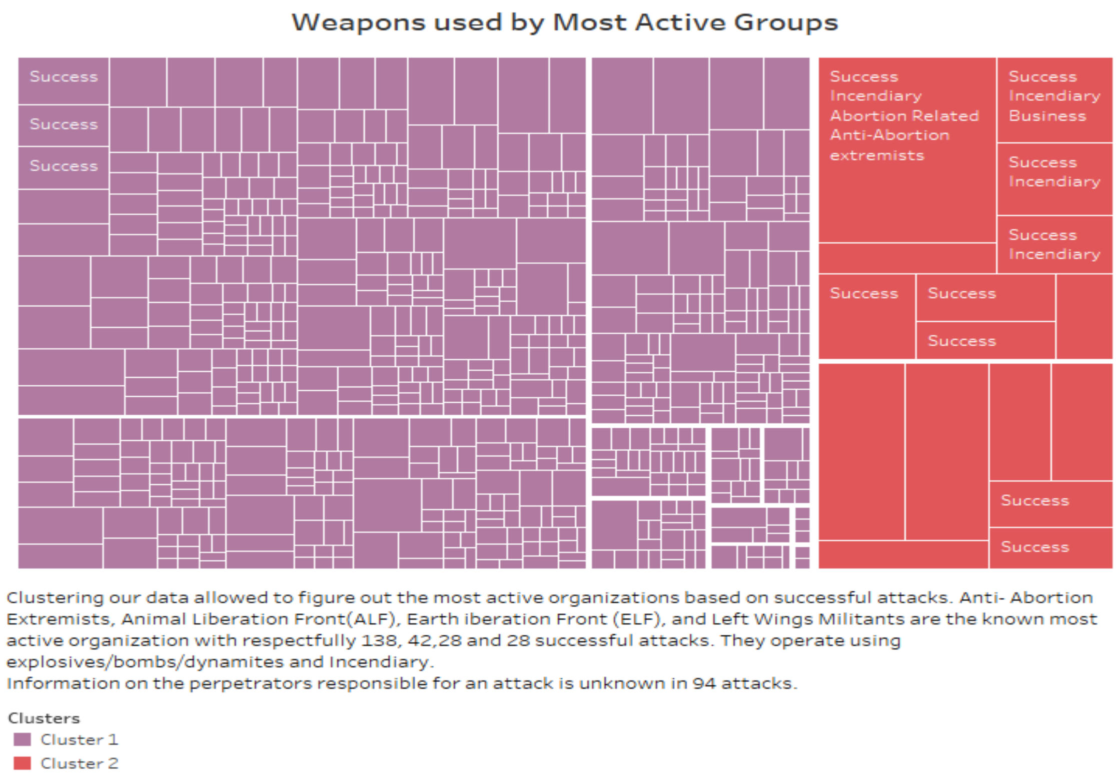


Figure 8: Weapons used by Most Active Groups

Using K-means, we were able to cluster our data to identify the most active based on the type of weapons and the total number of successful attacks.
Anti-Abortion Extremists, Animal Liberation Front, Earth Liberation Front, and Left Wings Militants groups are the most known active organization with respectively 138, 42, 28, and 28 attacks. Information on the perpetrators responsible for an attack is unknown for 94 attacks.

4 Model I – Build cluster model using "SimpleKMeans" in Weka.

4.1 Make a conjecture

Combining GTD and PPT datasets, we found five dimensions to describe the ideologies of terrorist organizations according to the codebook in PPT database. For instance, Extreme Right Wing, Extreme Left Wing, Religious, Racist, Single Issue. In addition, we found that there is a column for the value of property damage-only includes direct economic effects of the incident (i.e. cost of buildings, etc.)- after the terrorist attacks in the data of GTD. So, we wanted to find out any relationship between the ideologies of the terrorist organizations and the value of property damage.

4.2 Data selection and processing

Due to variables from different datasets, we merged two tables using "organization's name" and the "attack location" in GTD and PPT. In this process, we also removed the missing data and unitized data types. Therefore, we came up with a new dataset. With the new dataset, we observed a considerate amount of N/A in propvalue column. N/A does *not* indicate that there was no property damage but, rather, that no precise estimate of the value was available. We predicted that the existence of N/A would greatly affect our final result. So, we decided to build two datasets for our analysis. The first dataset is "NewData" and the second dataset is "NewData1" that deleted N/A in propvalue column.

The following is the metadata of variables that we choose from GTD and PPT.

Column name	Description	Data Type	Example
Religious	Religious perpetrator groups are those that use violence to achieve their goals and seek to smite the purported enemies of God and other evildoers, impose strict religious tenets or laws on society (fundamentalists), forcibly insert religion into the political sphere (i.e., those who seek to politicize religion, such as Christian Reconstructionists and Islamists), and/or bring about Armageddon (apocalyptic millenarian cults).	Categorical	Eg: 1= "Yes"; 0= "No" or "No indication based on available data" -99= Unknown/conflicting information exists in available data
Racist	Racist perpetrator groups are those that use violence to achieve their goals and are characterized by their belief that race is the primary determinant of human traits and that racial differences produce an inherent superiority of a particular race.	Categorical	Same as above.
Extreme right wing	Extreme right-wing perpetrator groups are those that use violence to achieve their goals and believe that one's personal and/or national "way of life" is under attack and is either already lost or that the threat is imminent (for some the threat is from a specific ethnic, racial, or religious group). They believe in the need to be prepared for an attack either by participating in paramilitary preparations and training or survivalism.	Categorical	Same as above.
Extreme left wing	Extreme left-wing perpetrator groups want to bring about change through violent revolution rather than through established political processes.	Categorical	Same as above.
Single issue	Single issue perpetrator groups are those that rely heavily on violence motivated by very specific or narrowly-defined causes of various sorts. This category includes groups from all sides of the political spectrum.	Categorical	Same as above.
propvalue	Value of property Damage (in USD)	Numeric	Eg:22500. $22500 of property has been damaged

4.3 Building model and Analysis

At first, we thought to use decision tree to build classification model to generate the relationship among organization types and the value of damaged property. The data set has only 6 variables: "propvalue" which is continuous that should be set as target, the rest five including "extreme left wing", "extreme right wing", "racist", "single", "religious" are all binary that should be set as input. We started to build decision tree model in R in different two ways. like: Fig 1. Fig 2. However, the result provided a bar chart when trying to plot tree using both ways.

```r
### second try building decision tree model
myFormula <- propvalue ~ religious + eleftwing + erightwing + racist + single

mytree <- ctree(myFormula, data=TreeData)
print(mytree)
plot(mytree)
```

```r
### first try building decision tree model
TreeData<- NewData[, c(8:13)]

rPartModel = rpart(propvalue~, method="anova", control=rpart.control(cp=0.01,maxdepth=30))
rpartTree = as.party(rPartModel)
dev.new()
plot(rpartTree)
```

Obviously, this is not the result expected. We, therefore decided to use cluster instead of decision tree- which is better model for continuous data and does not work well with binary data. Also, popular models of classification cannot work on such data set

either. Through cluster we were able to generate nodes by their similarity and by looking at those nodes, find the relationship among organization types and the value of damaged property.
This way we built a model with Waikato Environment for Knowledge Analysis (Weka). It is easy and free to use because it offers pre-designed data mining model, cluster included. There in Weka, we used “SimpleKMeans” cluster model. In “KMeans” cluster, the number of cluster has to be determined before model building and the default number of cluster in Weka is 2. Similar way has been tried to build cluster models on NewData which includes the value for property damaged of 0. The number of cluster was set as default which is 2 and starting points were chosen randomly by machine. The results are as followed.

```
Number of iterations: 2
Within cluster sum of squared errors: 107.83527269550794

Initial starting points (random):

Cluster 0: 0,0,1,0,0,1
Cluster 1: 4500,0,1,0,0,0

Missing values globally replaced with mean/mode

Final cluster centroids:
                                      Cluster#
Attribute                Full Data           0           1
                           (658.0)     (188.0)     (470.0)
==========================================================
propvalue              101642.0942 317605.0053  15256.9298
religious                        0           0           0
extreme left wing           0.8465      0.6968      0.9064
extreme right wing          0.0319      0.0106      0.0404
racist                      0.0106      0.0266      0.0043
political issue             0.2857           1           0
```

```
Number of iterations: 3
Within cluster sum of squared errors: 72.9404455999123

Initial starting points (random):

Cluster 0: 0,0,1,0,0,0
Cluster 1: 10000,0,1,0,0,0

Missing values globally replaced with mean/mode

Final cluster centroids:
                                      Cluster#
Attribute                Full Data           0           1
                           (434.0)     (305.0)     (129.0)
==========================================================
propvalue              133453.0046  14694.0033 414239.7907
religious                        0           0           0
extreme left wing           0.8456      0.9016      0.7132
extreme right wing          0.0323      0.0426      0.0078
racist                      0.0115      0.0066      0.0233
political issue             0.2972           0           1
```

Result NewData.

Two models have been built by randomly choosing two different starting points. There, the result is not that obvious because it is hard to say whether it can be rounded from 0.69 to 1. It is hard to say the value of “extreme left wing” is 0.69 means that this organization is extreme left wing or not. Therefore, we pursue the analysis by using “NewData1” which excludes the value for property damaged of 0.
In total NewData1 has 169 instances, and we chose 66% of that as training data set, and set the remaining as testing data set. The results are as followed. Weka has helped building two cluster models by random choosing two different starting points.

```
Number of iterations: 3
Within cluster sum of squared errors: 9.953298790646903

Initial starting points (random):

Cluster 0: 125000,0,1,0,0,0
Cluster 1: 2000,0,0,0,0,1

Missing values globally replaced with mean/mode

Final cluster centroids:
                                      Cluster#
Attribute                Full Data           0           1
                           (169.0)     (111.0)      (58.0)
==========================================================
propvalue              395742.5917  64601.4144 1029478.2931
religious                        0           0           0
extreme left wing           0.9408       0.982      0.8621
extreme right wing               0           0           0
racist                           0           0           0
political issue             0.3432           0           1
```

Result. 1

```
Number of iterations: 2
Within cluster sum of squared errors: 9.483886396901047

Initial starting points (random):

Cluster 0: 25000,0,1,0,0,0
Cluster 1: 193098,0,1,0,0,1

Missing values globally replaced with mean/mode

Final cluster centroids:
                                      Cluster#
Attribute                Full Data           0            1
                           (111.0)      (68.0)       (43.0)
===========================================================
propvalue              468300.7027  52743.6912 1125460.6279
religious                        0           0            0
extreme left wing           0.9099      0.9706        0.814
extreme right wing               0           0            0
racist                           0           0            0
political issue             0.3874           0            1
```

Result. 2

Result 1- two clusters have been built: cluster 0 includes 65.7% of training instances and cluster 1 includes 34.3% of training instances. Organizations with characteristic as extreme left wing stood out, without characteristics as religious, racist and single issue may cause $ 6,401 of property damaged. While organization with characteristic as extreme left wing and political issue, without racist and religious may increase the damage sharply to $ 1,029,478.
Result 2- cluster 0 contains 61.3% instances and cluster 1 contains 38.7% instances and the conclusions of two models are similar that organizations with characteristic as extreme left wing and political issue, without racist and religious may cause higher damage sharply to $ 1,125,460 than that caused by organizations with characteristic as extreme left wing, without characteristics as religious, racist and single issue which is $ 52,753.
Therefore, when hearing the advance notice of attacks coming from organizations fits in such two clusters, police and FBI should be prepared to deal with certain value loses of property which means such kind of organizations may cause such level of harm to the society and they should pay enough attention and be well prepared.

5 Model II – AR, MA, ARIMA, and ETS

In this part, we would like to conduct time series analysis models, which are Autoregressive model (AR), Moving average model (MA), Autoregressive integrated moving average (ARIMA) and Exponential Smoothing model (ETS) tested by evaluation matrix with minimizing five errors (ME, RMSE, MAE, MPE, MAPE) from 1994 to 2016 in order to predict terrorist attacks in the U.S. in the few years later, especially in 2017 and 2018.

	ME	RMSE	MAE	MPE	MAPE
AR	24.88954	26.96664	24.88954	47.51220	47.51220
MA	28.75043	30.78290	28.75043	55.33927	55.33927
ARIMA	26.48035	28.29770	26.48035	51.04098	51.04098
ETS	32.41145	34.22722	32.41145	63.03374	63.03374

According to five errors of each model, we are not hard to find that accuracy of all-time series models look like are pretty high, especially AR model (because range of errors from 0 to ∞).

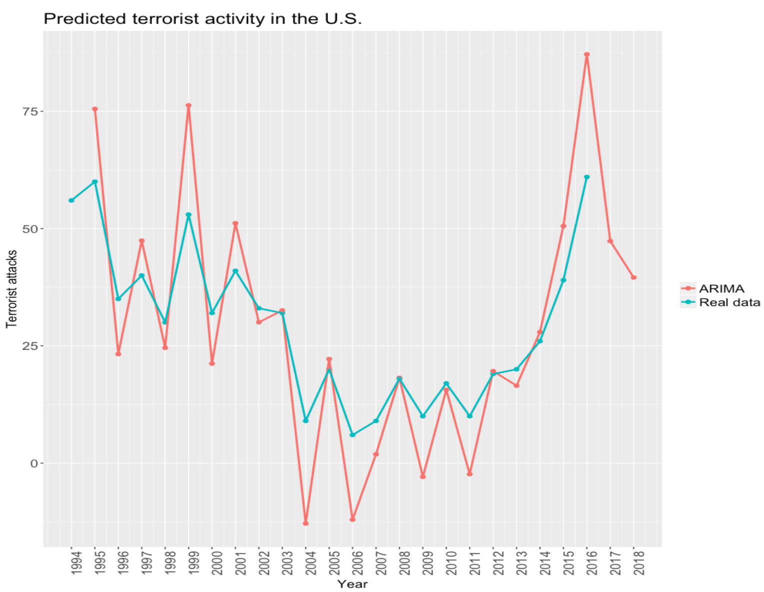


Figure 9: Predicted Terrorist Activity in the U.S.

However, due to these evaluation methods only can test error values, without taking into account the reason of the prediction error, therefore, it doesn't mean that the prediction performance is the best.

In this graph, we have two lines, ARIMA line and Real data line. Real data line based on our GTD and PPT datasets from 1994 to 2016, ARIMA line comes from ARIMA prediction model. We choose "state", "year" with "attack" to predict numbers of attacks happen in the U.S. 2017 and 2018. It indicates that terrorist attacks will decrease after getting a peak in 2016, it would be lower than 50 and more.

Although the downward trend of attacks is easy to find, but we need to increase our prediction performance, such as precise numbers of attacks in the future works, because we notice that there are gaps between Real data and ARIMA prediction data shown on the graph.

6 Limitations and Future works

There are several limitations in our project. Firstly, there are still some important data missed or unknown like dominant ideology, organization headquarters location information, and recruitment processes adopted. The missing data caused some bias about the final results and analysis. The limited information on events that processed that attacks affect our prediction, too. Secondly, there was not enough numerical to fit in PCA model. If we did, we could find which variable was affected mostly among all the main variables. PCA model didn't fit categorical data that well, so we analysis some variables individually. Some work that future research need to be done are building more models that can solve categorical data. Also, avoiding deleting data that contains N/A when aggregating variables.

7 Conclusion

As you can see, there were a lot of variables that caused terrorism in the United States from 1970 to 2016. From the data that we used in analysis were not enough to predict very accurate future. But still, there were some variables we need to pay more attention on than others. Different from our hypothesis, terrorist attacks were caused by inspired by people with protest and abortion motives against the government and business rather than in states where perpetrators have their organizations' headquarters. On the other hand, anyone who cares about terrorism issue should consider the religious and race events and the differences in political and economic policies or aspirations. By combining our time-series analysis, new technology, policy, and weapons, hopefully there might be less terrorism in 2017.