Terrorism Dataset analysis

Benjamin Vadurel[1] and Emilien Komlenovic[1]

[1]Université Claude Bernard Lyon 1

November 17, 2024

Abstract

This study performs data mining techniques on the University of Maryland's comprehensive terrorism database covering events from 1970 to 2021. We employ multiple analytical approaches, including tetrachoric correlation analysis, clustering algorithms, and temporal pattern recognition, to uncover underlying relationships between terrorist events and their geopolitical context. Our methodology encompasses preprocessing categorical data through one-hot encoding and applying various clustering techniques (K-Means, Agglomerative Clustering, K-Modes, and DBSCAN) to identify distinct attack patterns. We obtained preliminary results, such as data correlations, and established links with the events in the real-world.

Keywords: correlation; pattern recognition; cluster analysis; temporal anomaly detection

1 Introduction

The dataset describes each terrorist event that occurred from 1970 to 2021, it gives data about the date, the place and the individuals involved in the event. The dataset (University of Maryland, 2024) is made by the University of Maryland and the data come from various organizations :

- The Pinkerton Global Intelligence Service (PGIS) : a private security agency.
- Center for Terrorism and Intelligence Studies (CETIS) : a research center dedicated to the study of terrorist groups, political and religious extremism, and various types of covert intelligence operations.
- Institute for the Study of Violent Group (ISVG) : a research center focused on insurgency, terrorism, and transnational organized crime.

2 Data Description

2.1 Fields and content

Each row of the dataset aim to describe a terrorist event with 135 columns. Using the information contained in these columns we can get :

- The date the event occurred
- The place the event occurred
- Some descriptors for the attack (has the attack succeeded, has the attacker killed itself, etc..)
- The attack type representing the method used (ex : Assassination, Bombing, etc..)
- 3 Inclusion criteria :
  crit1 means that the attacks was made for political, economic, religious or social goal instead of personal motive (like profit).
  crit2 concerns the attacks that were made to coerce, intimidate or publicize large audiences.
  crit3 represent if the attacks respect the international humanitarian law.
- The weapon used during the attack
- The target of the attack

- The name of the terrorist group
- If the attack was claimed
- The material and human damages the attack caused

A complete documentation for the data can be found in the *Codebook*.

2.2 Preprocessing

The dataset was originally encoded for human understanding and sharing but unfit for data analysis. For the categorical data, each field are duplicated having the category label and the associated category in text. For those, we decided to remove the duplicated fields and one-hot encode the categories. In the dataset, unknown data are usually represented by -9, we replaced those by NaN.

2.3 Visualization

As we are working with spatial data, we can visualize it by building a map representing the attacks such as Figure 1 where each red dot represent an attack. Most of the attacks are located on populated areas of the world.

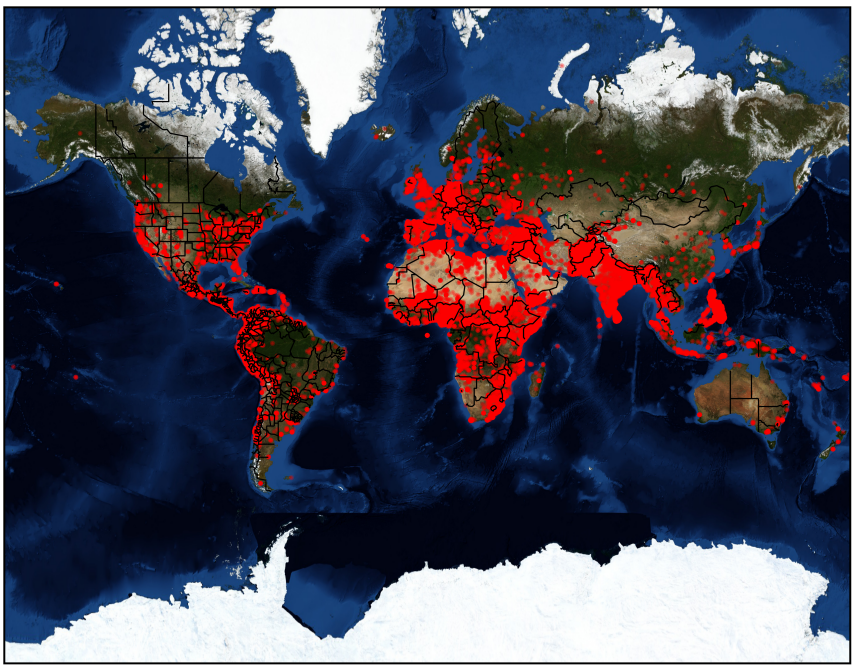


Figure 1: Map displaying each positions where a terrorist event happened from 1970 to 2021

3 Problems

Our goal is to correlate data on terrorist attacks with the corresponding geopolitical context. To achieve this, we have established several sub-questions:

- Match population data by year for each country
- Find attacks that can be associated together
- Identify frequent patterns in the data: Uncovering recurring trends or associations to better understand the underlying dynamics.

- Detect anomalies in the data: Highlight irregularities that may indicate geopolitical problems.
- Analyze temporal relationships: Studying the evolution of events over time.

4 Data processing

4.1 Variable correlations

To process the correlation among the variables, we'll use the tetracoric correlation which is efficient in our case because the dataset contains a lot of boolean variables.

$$v = a \cdot \frac{d}{\frac{b}{c}}$$

a, b, c, d Corresponding to the factor of the 2 variables in the confusion matrix

$$corr = \cos\left(\frac{\pi}{1+\sqrt{v}}\right)$$

Using this coefficient, we produce the heatmap given in Figure 7 and using it we can make several observations.

We can first focus on the inclusion criterion : we observe that for each criterion we find a correlation with military targeting. The criterion 3 indicating that the attacks doesn't follow the international humanitarian laws is often correlated with targets like educational institution, citizen, properties, businesses and religious figures. Moreover, criterion 3 attack seems to have a correlation with the claiming method such as letters.

Hijacking attacks are often correlated to targets such as maritime ports and airports which is coherent to the nature of the attack.

Finally, abortion-related attacks are correlated to attack on infrastructure which is also coherent as abortion-related attacks often happen in abortion clinics as specified in the *Codebook*

4.2 Clustering

The objective here is to highlight the presence of groups in the attacks. As we're dealing with a large dataset, we'll split the dataset based on the time periods.

All the clustering we built, were based on the *attack_type*, *the target_type*, *the claim mode* and some additional descriptors. Using these fields, we first performed clustering using 4 different algorithms : K-Means(Wikipedia contributors, 2024e), Agglomerative Clustering(Wikipedia contributors, 2024d) and K-Modes(Huang, 1998). Implementing the automatic k-selection, we evaluated the silhouette score for each cluster size represented in Table I.

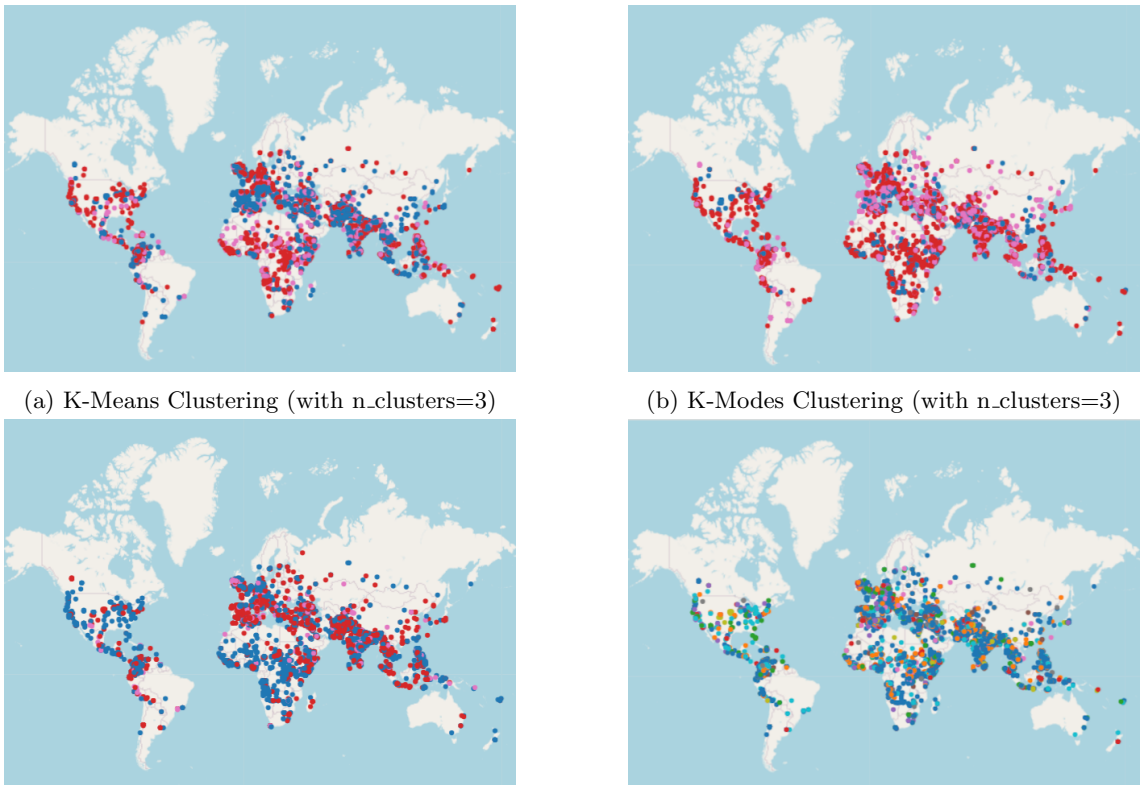
Cluster number	Agglomerative clustering	K-Means	K-Modes
2	0.187756	0.262278	0.197518
3	0.194116	0.21552	0.206285
4	0.215864	0.187088	0.20739
5	0.192268	0.218445	0.197518
6	0.209013	0.21552	0.197518

Table I: Computed silhouette(Wikipedia contributors, 2024f) score for each algorithm depending on the number of cluster

We observe in Table I that for the 3 methods, silhouette score nears 0,20. As the cluster size does not seems to impacts the quality of the cluster, we will choose arbitrarily 3 clusters. As we're working with spatial data, we can ensure spatial continuity with the agglomerative clustering using a connectivity matrix.

We also tried DBScan algorithm for the clustering, it gives a clustering with 387 clusters and a silhouette score of 0.881. Even if we get a decent silhouette score for this clustering, it's hard to determine a meaning for these clusters.

Figure 2 represents the obtained clustering for the attacks between 2000 and 2010.



(a) K-Means Clustering (with n_clusters=3)

(b) K-Modes Clustering (with n_clusters=3)

(c) Agglom. Clustering (with n_clusters=3)

(d) DBScan Clustering (with eps=0.5)

Figure 2: Comparison of Clustering Methods

To better understand the produced clusters, we can observe the plots 8, 9, 10 representing the proportion of each features present in the clusters. The plot is build from the results obtains by the algorithms with 3 clusters.

The k-modes algorithms has its first cluster containing bombing attacks, the second armed assaults on the police or military and the third contains armed assaults on citizens or properties.

The k-means algorithms has its first cluster containing every attacks targeting citizens such as armed assaults or kidnapping, the second cluster contains bombing attacks on citizen and the third contains bombing on the police which also includes suicide attacks.

The agglomerative clustering has its first cluster containing all kinds of attacks on citizens (assassination, kidnapping and armed assaults), the second mainly contains bombing attacks on citizens and properties and the third contains attacks on the police.

The 3 algorithms produced clustering separating attacks on citizen from the attacks on institutions (such as police and military) and bombing type attack from armed attack or kidnapping.

We won't perform this analysis on the clustering produced by DBScan algorithm as it produced a high number of cluster.

4.3 Frequent Patterns

Initially, we search for frequent patterns on the boolean data in the dataset. Our primary objective was to discover association rules by applying the Apriori algorithm(Raschka, 2024). However, we observed that no significant frequent patterns emerged.

After preprocessing the dataset, the results changed significantly. The preprocessing allowed us to reveal strong association rules that were not apparent in the raw data.

The criteria *crit1*, *crit2*, and *doubtterr* play a central role in defining armed incidents against military targets. Successful incidents involving firearms have a very high probability of involving military targets. The very strong relationships (high lift, high conviction) suggest that these patterns are significant in the context of terrorism, as illustrated in Figure 5.

We aimed to identify patterns between terrorist groups and attack types. However, based on the metrics, we found that there is no consistent pattern.

4.4 Temporal Analysis

4.4.1 Anomaly Detection

To detect temporal anomalies, we counted the number of attacks per year by country. However, we encountered an issue related to scaling, as countries like Afghanistan and Iraq report a significantly higher number of attacks in certain years, which creates challenges with the scaling of the data.

To address this issue, we computed the deviations from the average for each countries and then visualized the results in *heatmap* Figure 6. The heatmap included in this report represents data for a subset of 50 countries. For the complete version, you can refer to the repository on GitHub. We can try to make some analysis based on our results.

The years 2010 to 2020 in Afghanistan were marked by the intensification of the Taliban insurgency, the rise of other extremist groups such as ISIS-K, and a series of political and military events that contributed to the resurgence of terrorist attacks in the country.(Wikipedia contributors, 2024a)

Between 1983 and 1997, Colombia was engulfed in a multifaceted conflict involving leftist guerrillas, right-wing paramilitary groups, and drug cartels. Violence, drug trafficking, and political instability made this period ripe for increased terrorist acts.(Wikipedia contributors, 2024b)

From 1977 to 1986, France experienced a pivotal shift in governmental policy, particularly with the election of François Mitterrand, which may have contributed to the rise in violence during this period.(Wikipedia contributors, 2024c)

4.4.2 France over years

To examine the evolution of the number of attacks in France Figure 3 in relation to its population(World Bank, 2024), we observed that while the total number of attacks remained relatively stable, the population has significantly increased over time. Consequently, the ratio of attacks per capita has markedly decreased.

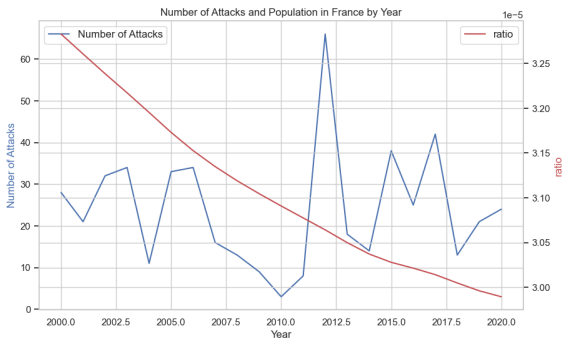


Figure 3: Number of attacks and ratio in France between 2000-2020

By highlighting the anomalies in Figure 4 in the number of attacks each months, we can reveal terrorism crisis in France.

As mentioned before, the period from 1978 to 1986 represents the pivotal shift in the governmental politic. During the period going from 1990 to 1997 a lot of attacked were made by the Armed Islamic Group of Algeria more precisely between 1994 and 1996. The attacks in 2011 and 2012 mainly comes from the Corsica region because of political issues.

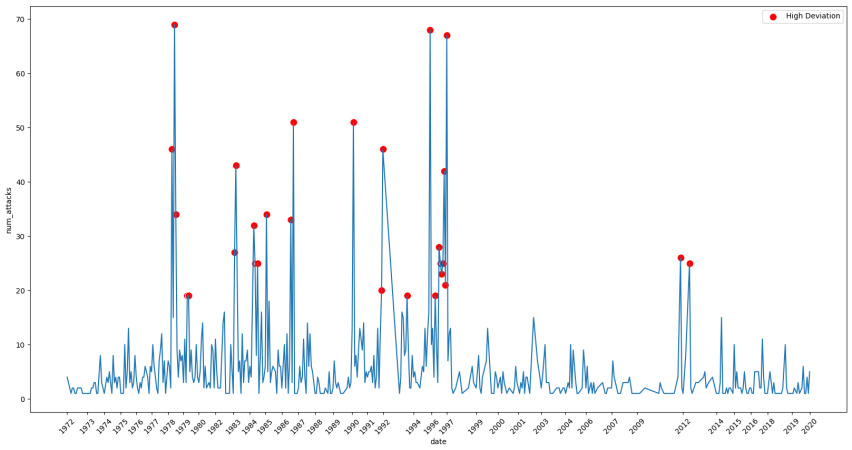


Figure 4: Evolution of the number of attack in France from 1972 to 2020

5 Contributions

We both worked on data description and the problem statement, as well as data processing. Benjamin focused on variable correlation and clustering, while Emilien worked on frequent pattern analysis and temporal analysis.

6 Annexe

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
360022	(crit1, crit2, doubtterr, weapon_unknown)	(attack_unknown, target_military)	0.066734	0.062097	0.055847	0.836858	13.476674	0.051703	5.748999	0.991998
360071	(attack_unknown, target_military)	(crit1, crit2, doubtterr, weapon_unknown)	0.062097	0.066734	0.055847	0.899351	13.476674	0.051703	9.272451	0.987093
192020	(crit2, doubtterr, weapon_unknown)	(attack_unknown, target_military)	0.066935	0.062097	0.055847	0.834337	13.436082	0.051690	5.661525	0.991972
360047	(crit2, doubtterr, weapon_unknown)	(attack_unknown, crit1, target_military)	0.066935	0.062097	0.055847	0.834337	13.436082	0.051690	5.661525	0.991972
360046	(attack_unknown, crit1, target_military)	(crit2, doubtterr, weapon_unknown)	0.062097	0.066935	0.055847	0.899351	13.436082	0.051690	9.270447	0.986854
192037	(attack_unknown, target_military)	(crit2, doubtterr, weapon_unknown)	0.062097	0.066935	0.055847	0.899351	13.436082	0.051690	9.270447	0.986854
170445	(attack_unknown, target_military)	(crit1, doubtterr, weapon_unknown)	0.062097	0.067137	0.055847	0.899351	13.395733	0.051678	9.268444	0.986615
170428	(crit1, doubtterr, weapon_unknown)	(attack_unknown, target_military)	0.067137	0.062097	0.055847	0.831832	13.395733	0.051678	5.577175	0.991946
360041	(crit1, doubtterr, weapon_unknown)	(attack_unknown, crit2, target_military)	0.067137	0.062097	0.055847	0.831832	13.395733	0.051678	5.577175	0.991946
360052	(attack_unknown, crit2, target_military)	(crit1, doubtterr, weapon_unknown)	0.062097	0.067137	0.055847	0.899351	13.395733	0.051678	9.268444	0.986615
63241	(doubtterr, weapon_unknown)	(target_military, attack_unknown)	0.067339	0.062097	0.055847	0.829341	13.355626	0.051665	5.495784	0.991920
360066	(doubtterr, weapon_unknown)	(attack_unknown, crit1, crit2, target_military)	0.067339	0.062097	0.055847	0.829341	13.355626	0.051665	5.495784	0.991920
192034	(doubtterr, weapon_unknown)	(attack_unknown, crit2, target_military)	0.067339	0.062097	0.055847	0.829341	13.355626	0.051665	5.495784	0.991920
192023	(attack_unknown, crit2, target_military)	(doubtterr, weapon_unknown)	0.062097	0.067339	0.055847	0.899351	13.355626	0.051665	9.266441	0.986376
170442	(doubtterr, weapon_unknown)	(attack_unknown, crit1, target_military)	0.067339	0.062097	0.055847	0.829341	13.355626	0.051665	5.495784	0.991920
170431	(attack_unknown, crit1, target_military)	(doubtterr, weapon_unknown)	0.062097	0.067339	0.055847	0.899351	13.355626	0.051665	9.266441	0.986376

Figure 5: Frequent Pattern Ordered by lift

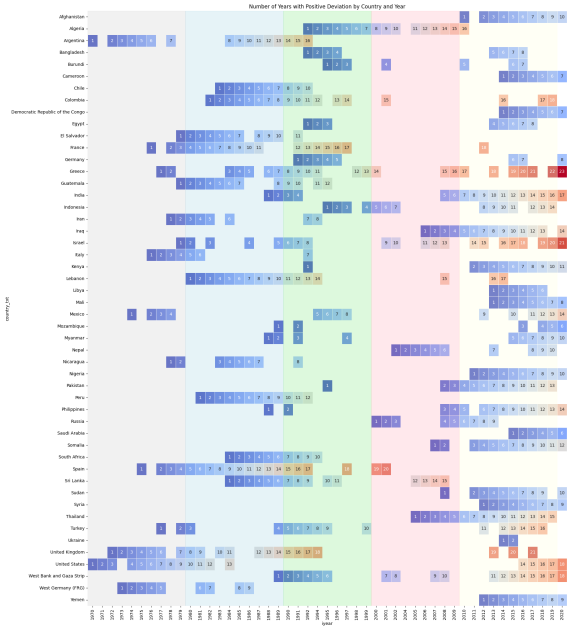


Figure 6: Heatmap Anomaly Detection

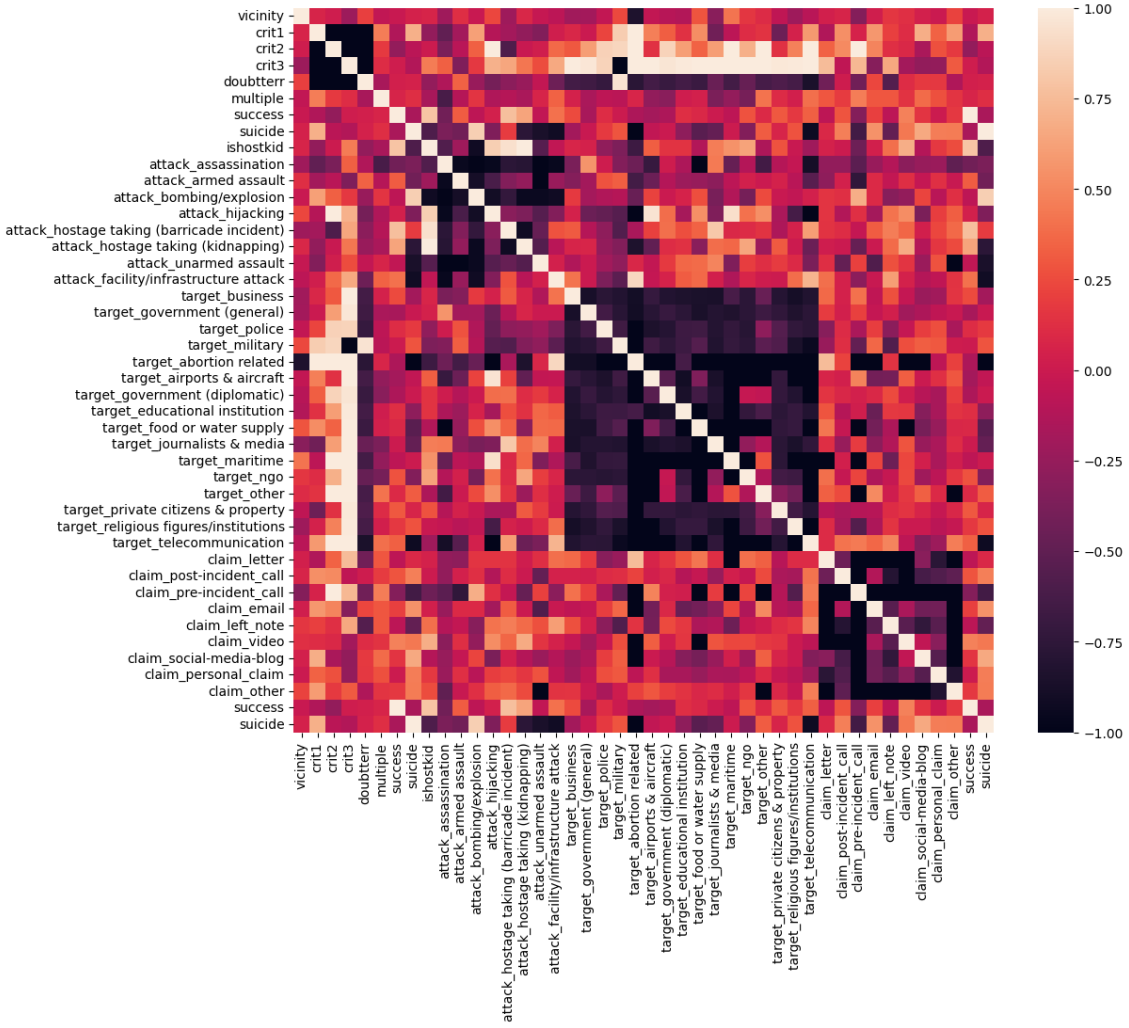


Figure 7: Tetracoric correlation heatmap

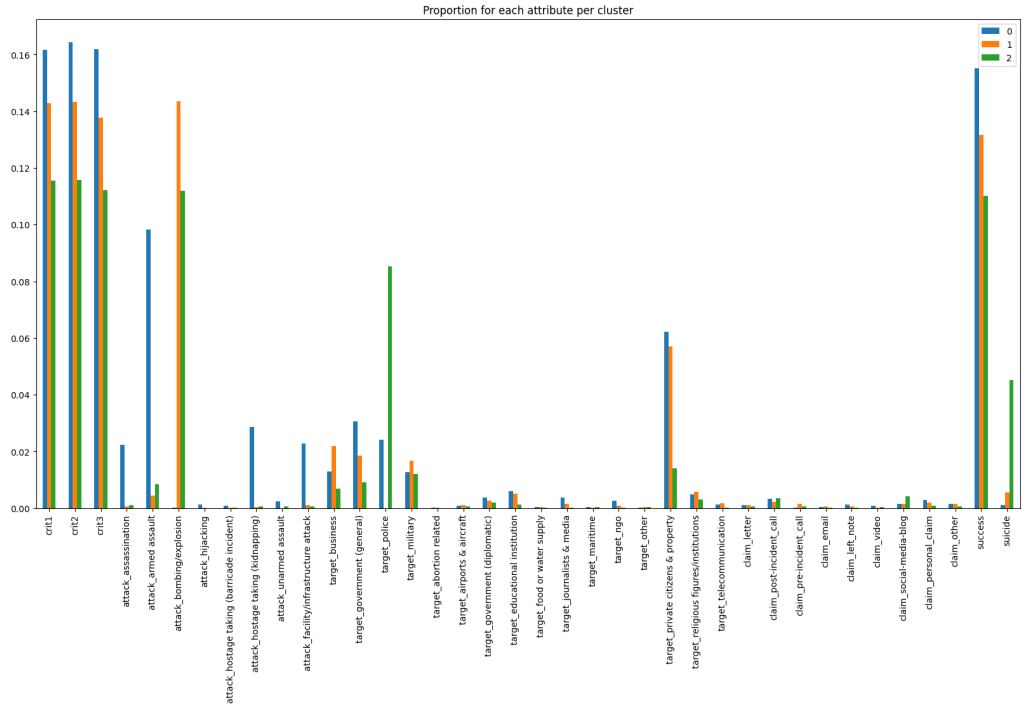


Figure 8: Attribute proportion per cluster for the K-Means algorithm

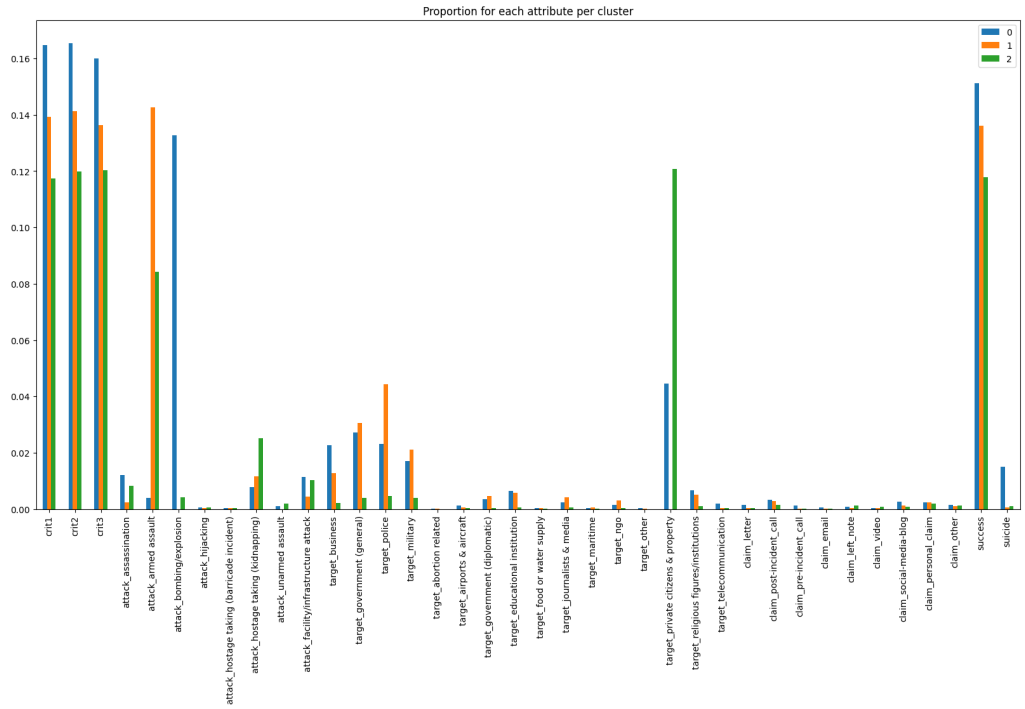


Figure 9: Attribute proportion per cluster for the K-Modes algorithm

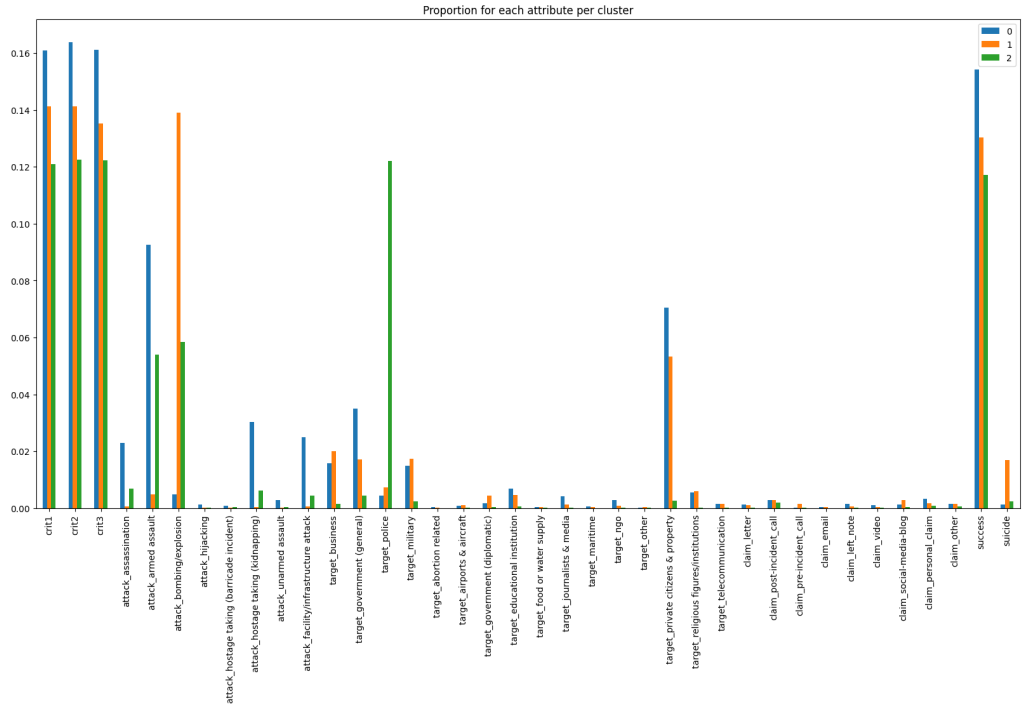


Figure 10: Attribute proportion per cluster for the Agglomerative clustering algorithm

References

Huang, Z (1998). *Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values*, [Accessed 11-11-2024]. **available at**: `https://home.cse.ust.hk/~qyang/Teaching/537/Papers/huang98extensions.pdf`.

Raschka, Sebastian (2024). *mlxtend: Machine Learning Extensions*, [Accessed 17-11-2024]. **available at**: `https://rasbt.github.io/mlxtend/`.

University of Maryland (2024). *Global Terrorism Database (GTD)*, [Accessed 28-09-2024]. **available at**: `https://www.start.umd.edu/gtd/`.

Wikipedia contributors (2024a). *Afghanistan*, [Accessed 16-11-2024]. **available at**: `https://en.wikipedia.org/wiki/Afghanistan`.

— (2024b). *Colombia*, [Accessed 16-11-2024]. **available at**: `https://en.wikipedia.org/wiki/Colombia`.

— (2024c). *François Mitterrand*, [Accessed 17-11-2024]. **available at**: `https://fr.wikipedia.org/wiki/Fran%C3%A7ois_Mitterrand`.

— (2024d). *Hierarchical clustering*, [Accessed 11-11-2024]. **available at**: `https://en.wikipedia.org/wiki/Hierarchical_clustering`.

— (2024e). *K-means clustering*, [Accessed 11-11-2024]. **available at**: `https://en.wikipedia.org/wiki/K-means_clustering`.

— (2024f). *Silhouette (clustering)*, [Accessed 17-11-2024]. **available at**: `https://en.wikipedia.org/wiki/Silhouette_(clustering)`.

World Bank (2024). *Total Population (SP.POP.TOTL)*, [Accessed 30-09-2024]. **available at**: `https://data.worldbank.org/indicator/SP.POP.TOTL`.