Preprint Manuscript: Bridgelall, R. (2022). Applying Unsupervised Machine Learning to Counterterrorism. *Journal of Computational Social Science*, Springer Nature, DOI: 10.1007/s42001-022-00164-w

Applying Unsupervised Machine Learning to Counterterrorism

Raj Bridgelall, Ph.D.
Assistant Professor, Transportation, Logistics & Finance, College of Business
North Dakota State University
PO Box 6050, Fargo ND 58108-6050
Email: raj@bridgelall.com
ORCiD: 0000-0003-3743-6652

Abstract

To advance the agenda in counterterrorism, this work demonstrates how analysts can combine unsupervised machine learning, exploratory data analysis, and statistical tests to discover features associated with different terrorist motives. A new empirical text mining method created a "motive" field in the Global Terrorism Database to enable associative relationship mining among features that characterize terrorist events. The methodology incorporated K-means co-clustering, three methods of non-linear projection, and two spatial association tests to reveal statistically significant relationships between terrorist motives, tactics, and targets. Planners and investigators can replicate the approach to distill knowledge from big datasets to help advance the state of the art in counterterrorism.

Keywords: Data Wrangling; Local Indicators of Spatial Association; Nonlinear Projection; Statistical Learning; Text Mining

1 Introduction

Terrorism has a long-term impact on the psyche of affected populations and can damage their economies, infrastructures, and the environment. It is difficult for humans to predict and prepare to act upon terror threats (Krasmann & Hentschel, 2019). While there are many motivations for terror threats, racially and ethnically motivated terrorism is on the rise (Miller, 2020). Consequently, researchers have been turning to emerging techniques such as machine learning (ML) for insights to help counterterrorism efforts (Lu, Zhang, Li, Chen, & Yang, 2020).

The **goal** of this work is to further the research agenda in counterterrorism by exploring the potential of combining exploratory data analysis (EDA) and ML to help produce insights from large datasets. The **objective** of this research is to mine a large dataset to reveal any associative relationships between perpetrator motives, tactics, and targets that are statistically significant. The most comprehensive open-source database of terrorism worldwide is the Global Terrorism Database (GTD) (LaFree, 2010). The National Consortium for the Study of Terrorism and Responses to Terrorism (START), a Department of Homeland Security Emeritus Center of Excellence, maintains the GTD®.

ML techniques are divided into supervised ML (SML) and unsupervised ML (UML) methods (James, Witten, Hastie, & Tibshirani, 2013). SML classification models are easier to train because they require a class label for each observation in a dataset. However, class labels are not generally available for unprocessed data or they may be expensive to add. UML models can classify unlabeled data, but it is more difficult to interpret and assess the significance of their outputs. Hence, most researchers applied SML to the GTD. This research applied less-explored UML models to the GTD. Section 2 further reviews related work.

The **contributions** of this work are:

1. Derivation of perpetrator motive categories (PMCs), which are not present in the GTD, by empirical text mining (ETM) of various narrative fields (Sections 3.1 and 4.1)
2. Demonstration of how the UML technique of k-means co-clustering (KMC) associated motives and behavioral characteristics of perpetrators (Sections 3.2 and 4.2)
3. Visualization of associative relationships among the key features by using non-linear projections (NLP) (Sections 3.3 and 4.3)
4. Introduction of an unconventional but relevant method to assess the significance of clustered features by spatial autocorrelation (Sections 3.3 and 4.3)

The analysis focused on the U.S. subset of the GTD so that motives and behaviors would be comparable based on local situations. Examples of local situations are reactions to changes in taxation, immigration, government parties, and rulings on the constitutional rights of U.S. citizens.

2 Literature Review

The recent proliferation of violent extremism and the popularity of data mining has catapulted an interest in ML techniques to help with insights (Wall, 2021). Social science research found that theoretically informed models has the potential to predict and explain complex forms of political violence (Python, et al., 2021). For example, Pruyt and Kwakkel (2014) used three system simulation models to examine the dynamics of radicalization under deep uncertainty (Pruyt & Kwakkel, 2014). However, Mashechkin et al. (2019) cautioned against using databases of completed terrorist attacks to predict future attacks because such databases do not contain information about how the attackers maximized the significance of their propaganda, which is often their main aim (Mashechkin, Petrovskiy, Tsarev, & Chikunov, 2019).

ML techniques are among the least understood of technical concepts used to analyze terrorism (Ammar, 2019). Early analysis examined trends and cycles using time-series analysis (Enders, Parise, & Sandler, 1992), spatio-temporal visualizations (Guo, Liao, & Morgan, 2007), and panel data (Bayar & Gavriletea, 2018). More recently, the SML subset of ML applied to the GTD predicted either the attack location, incidence type, perpetrator group, or casuality levels. The model types used were Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), Naïve Bayes (NB), Logistic Regression (LR), K-Nearest Neighbor (KNN), Iterative Dichotomiser (ID3), Adaboost (ADB), and Extreme Gradient Boosting (XGB) (James, Witten, Hastie, & Tibshirani, 2013).

For location prediction applications, Huamaní et al. (2020) found that DT and RF provided a predictive accuracy of 75.5% and 89.5%, respectively (Huamaní, Alicia, & Roman-Gonzalez, 2020). Similarly, Ding et al. (2017) used RF, ANN, and SVM to predict attack locations in 2015 and found that RF provided the highest prediction accuracy of 96.6% (Ding, Ge, Jiang, Fu, & Hao, 2017). Focusing more locally on the Indochina Peninsula, Hao et al. (2019) found that RF coupled with a geographical information system (GIS) provided good location prediction performance (Hao, Jiang, Ding, Fu, & Chen, 2019). Using simplier methods, Clauset and Wiegel (2010) found that the frequency of severe terrorist attacks follows a power-law form and discussed ways that others can test it empirically to understand terrorism (Clauset & Wiegel, 2010).

For incidence type prediction, Uddin et al. (2020) found that an ANN with five layers provided significantly better performance than LR, SVM, and NB (Uddin, et al., 2020). Nizamani and Memon (2012) found that using a simple decision tree classifier to text mine news summary data in the GTD narratives provided better performance than NB and SVM (Nizamani

& Memon, 2012). However, a limitation of mining news summary is that terrorist groups claim credit for just a small portion of their attacks. (Abrahms & Conrad, 2017).

For perpetrator group prediction, Oketch et al. (2019) combined NB, KNN, DT, SVM and ANN to create a hybrid classifier that provided a slight boost in the predictive accuracy from 77.4% to 77.8%, without stratefied resampling (Opiyo, Mukisa, & Ratemo, 2019). Tolan and Soliman (2015) compared NB, KNN, DT, ID3, and SVM on the Egypt subset of the GTD and found that SVN and KNN provided perpetrator group predictive accuracies of 75.4% and 73%, respectively (Tolan & Soliman, 2015). Hung et al. (2018) used a dynamic graph pattern matching technique to detect radicalization trends of terrorism in the United States to inform law enforcement and intelligence (Hung, Jayasumana, & Bandara, 2018). Campedelli et al. (2019) analyzed the GTD with a multi-partite network to detect latent clusters of similar terrorist groups (Campedelli, Cruickshank, & M. Carley, 2019). The group also used temporal meta-graphs and deep learning to forecast future terrorist targets based on temporal similarities (Campedelli, Bartulovic, & Carley, 2021). In additional work, the group used multi-modal networks to reveal patterns of behavioral similarity among terrorist organizations based on their deployed tactics, attacked targets, and utilized weapons (Campedelli, Cruickshank, & Carley, 2021).

For casualty level prediction, Feng et al. (2020) found that XGB outperformed LR, ADB, RF, DT, and SVM with an accuracy of 86.3% for the China subset of the GTD (Feng, Wang, Yin, Li, & Hu, 2020). With the intent of improving surveillance methods, Mishra et al. (2020) mined the GTD to develop a directed graph that exposed an interconnected network of terrorist activities (Mishra, Swagatika, & Singh, 2020).

As mentioned previously, there were far fewer investigations that used UML approaches. Various methods of clustering data were the most popular application because of its simplicity and ease of understanding. Loia and Orciuoli (2019) used rough set theory to build a temporal sociogram of perpetrator groups based on similarity measures to help understand how terrorism networks evolved over time (Loia & Orciuoli, 2019). Aleroud and Gangopadhyay (2018) found that co-clustering can simultaneously cluster heterogeneous activities in terrorist networks to reveal multimode structures (Aleroud & Gangopadhyay, 2018). Adnan and Rafi (2015) used a heatmap to visualize the clustering of terrorist attacks by location and perpetrator group over time (Adnan & Rafi, 2015). Salem and Naouali (2016) used clustering to determine locations and target types of the most active perpetrator groups that conducted armed attacks (Salem & Naouali, 2016). For a similar purpose, Naouali et al. (2020) developed an improvement of the k-modes clustering algorithm by incorporating density (Naouali, Salem, & Chtourou, 2020). Curia (2020) found that there was very good agreement with attack success when using ANN based autoencoder and k-modes clustering to rank terrorist behaviors (Curia, 2020). Similarly, Atsa'am et al. (2020) used agglomerative hierarchical clustering to rank terrorist activities into four clusters of casualty and financial consequence levels (Atsa'am, Wario, & Okpo, 2020).

Only a few researchers used text mining techniques to extract intelligence. For example, Sun et al. (2003) extracted features from text documents in the terrorism domain and used SVM to classify the information for knowledge discovery (Sun, Naing, Lim, & Lam, 2003). Conlon et al. (2015) demonstrated a text mining system that could extract intelligence from the GTD such as the results of the incidents (Conlon, Abrahams, & Simmons, 2015). Strang and Sun (2017) applied data mining tools to Google News and found that there was a significant relationship between group ideology and attack type (Strang & Sun, 2017).

3 Method

The next subsections describe the methodological framework for the analysis. The framework includes data preparation, exploratory data analysis (EDA), and the UML methods. Figure 1 shows the logic flow of procedures in the framework. The next subsections provide a detailed description of each procedure whereas a separate results section discuss the outcome.

3.1 Data Wrangling

The series of methods used to prepare the dataset for EDA and ML was data cleaning, feature selection, feature extraction by empirical text mining (ETM), and variable transformation.

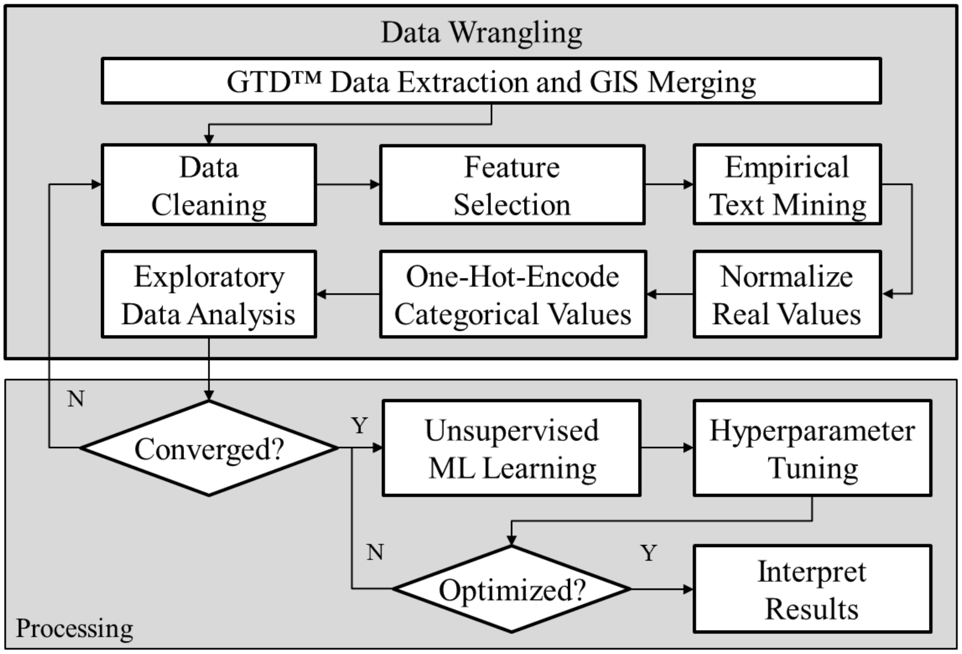

Fig 1 The analytical framework for applying UML to the GTD.

EDA then helped to discover any errors remaining in the data through outlier analysis and other methods of data visualization. The data cleaning process repeated upon discovery of new errors until the process converged. The next subsections discuss details of the various data wrangling procedures.

3.1.1 Data Extraction

The GTD data release at the time of this writing contained records of worldwide terrorist attacks from 1970 to 2018. However, the 1993 records were separate, so the workflow needed to align and merge those datasets. The explanation for the separated data was that prior to compilation by START, the 1993 subset was missing but was later reconstructed. This research limited the geographical span to the continental United States (CONUS) to enable comparability of features.

3.1.2 Data Cleaning

Using a geographical information system (GIS) tool, a spatial join procedure combined geospatial data from the Topologically Integrated Geographic Encoding/Line (TIGER/Line™) database of the U.S. Census Bureau (USCB, 2019). This procedure merged associated data tables containing information about each county, such as their names, state, Federal Information Processing Systems (FIPS) codes, and centroid geospatial coordinates. The spatial join procedure provided several means to detect errors. First, it enabled the detection of misspelled, duplicated, and incorrect city or county names by correlating features of unmerged instances. Second, it provided a means to fill in blank data such as latitude (LAT) and longitude (LON) values that were missing in the GTD records, especially the 1993 portion. Third, it provided a means to detect erroneous geospatial coordinates, such as positive LON values that should have been negative.

3.1.3 Feature Selection

Feature selection removed attributes that were redundant or not useful, such as text descriptions of the categorical variables and other meta data. A measure of variable dispersion indicated the relative amount of their variability or the amount of information contributed. For numeric variables, the measure of dispersion was the coefficient of variation for the value distribution.

For categorical variables, the measure of dispersion was the entropy of the value distribution. The procedure removed variables with dispersion near zero because they were essentially constants and did not add information to the models. The results section provides examples of the low dispersion variables removed.

3.1.4 Empirical Text Mining

An ETM algorithm invented for this research added a new field to the GTD that defined a unique perpetrator motive category (PMC). The ETM used fields that contained information to discern or infer motives such as the perpetrator group name (PGN), publication sources on the event, and a series of short text descriptions about possible motives, tactics, and associated events elsewhere. Standard text mining tools alone were not appropriate to implement the ETM because a substantial amount of contextual background, historic knowledge, subject matter expertise, and associative reasoning was necessary to accurately classify the motive category for many of the events. It was also important to recreate any misspellings in the text narratives to maximize the completeness of the classifications. Hence the process employed human cognition to minimize the number of categories while still covering the entire motive space.

Figure 2 shows the logic flow of the ETM algorithm that assembled a keyword list to assign the appropriate PMC to all observations. The ETC first applied a filter to the narratives of each observation by removing duplicate and insignificant words. Next, a word cloud provided helpful visualization to identify keywords associated with a PMC. The program then isolated unassigned PMCs to focus human cognition on assigning appropriate keywords from the various non-empty text narrative fields until all observations received a PMC label. The process prioritized non-empty text narratives to discern or infer motives and then to the PGN. For cases where more than one PGN was involved, the procedure selected the first mentioned group. For cases where both the PGN and the motive narrative was missing, it was necessary to examine other text narratives and fields to suggest a PMC.

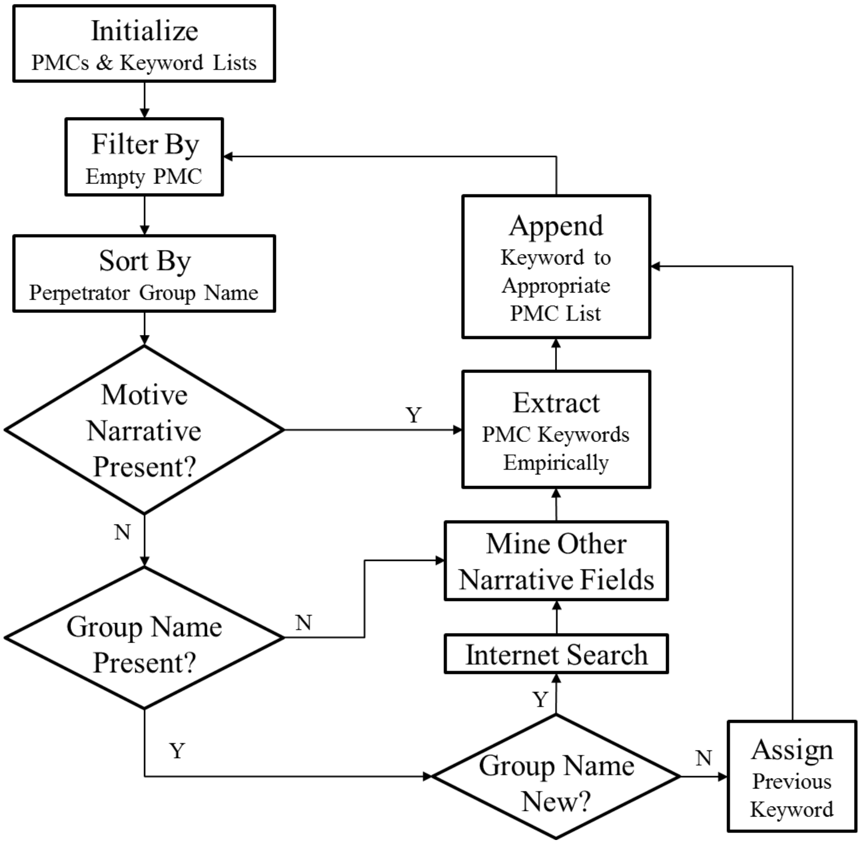

Fig 2 The Empirical Text Mining algorithm.

For cases where the PGN was present, but not previously associated with a PMC, an internet search was necessary to determine the PMC. Some keywords such as "Government" were associated with multiple categories. Hence, those category labels needed to be correlated with additional fields to determine the appropriate PMC.

3.1.5 Variable Transformation

Most ML methods require numerical variables. Therefore, the framework converted all categorical variables to binary variables by using a procedure known as *one-hot encoding*. The encoding created one new feature per category such that it received a value of 1 if present and 0 otherwise.

3.2 Exploratory Data Analysis

The approach to Exploratory data analysis (EDA) in this work was primarily empirical. The best methods available to summarize and visualize the main characteristics of data depends on the types of features analyzed. Multivariate categorical bar plots can visualize the global distribution of categories for each feature. For example, the ranked distribution of events by PMC and ATC can be plotted as a stacked bar chart to show the ATC proportion of each PMC. However, one limitation of stacked bar charts is that they can quantify the proportions of only two variables at a time. Therefore, separate charts are needed to visualize a select combination of variables.

For continuous variables, a related tool is the box plot which allows for the simultaneous visualization of six summary variables, namely the two extremes, the median, the mean, the standard deviation, and the quartiles. The relative positions of the means and the medians also provides an indication of the amount of skew in the distribution.

3.3 Unsupervised Machine Learning

This subsection provides a succinct review of the UML methods used. Each method reduced the dimensionality of the dataset to enhance interpretability while preserving the statistical information. The subsections describe the KMC method and three non-linear methods to project feature space for visualization and cluster analysis.

3.3.1 k-Means Co-clustering

In data mining, there is no single best method of UML that can reveal inherent patterns in data (Agresti, 2018). The workflow selected the base method of k-means clustering because it is one of the most popular and effective UML methods (Aggarwal, 2015). The general method of k-means clustering groups observations by assigning each to the cluster of the nearest mean or centroid (James, Witten, Hastie, & Tibshirani, 2013). Clusters have the property that they minimize the within-cluster distance variance and maximize cluster separation, with the Euclidian distance as a common measure. Co-clustering is an application of k-means clustering to both the observations (rows) and features (columns) of the observations. The method identifies groups of observations that express feature similarity. The method is also called “bi-clustering” and has gained popularity in the field of bioinformatics to visualize gene expression data (Madeira & Oliveira, 2004).

A hyperparameter sets the number of clusters desired; the algorithm is very sensitive to the selection of *k*. A common method to pick the best value for *k* is to plot a function of the *reduction* in total within-cluster variance against *k* and pick the value that is at an elbow in the curve. Visually, the best point is where the reduction decreases more slowly than those of the previous trials.

3.3.2 Nonlinear Projections

Methods of nonlinear projections (NLP) map the similarity relationship among observations in multidimensional feature space to a lower dimension for visualization and inspection. The various methods NLP optimize the spatial arrangement of observations in a lower dimension, for example a planar map of two-dimensions, to preserve the neighborhood relationship in high dimensional feature space. The next subsections describe the three algorithms that are most often cited for their effective visualizations (Becht, et al., 2019). Those algorithms are the t-distribution stochastic neighbor embedding (t-SNE), multiple dimensional scaling (MDS), and uniform manifold approximation and projection (UMAP).

<u>t-SNE</u>

This method is recognized as the current state of the art for dimension reduction and serves as a benchmark to compare performance with newer developments such as UMAP (Becht, et al., 2019). The method measures feature similarity by using a local probability density function of distance (Maaten & Hinton, 2008). Hence, t-SNE preserves the *local* relationships among observations, possibly at the expense of their global relationship. The t-SNE algorithm first builds a transformed similarity matrix in the original feature space. The transformation associated with each observation converts distances to a similarity score based on a Gaussian distribution centered at the observation and spanning a set of neighbors. Hence, the similarity score is proportional to a conditional probability, based on a Gaussian probability density, that an observation is related to another. The spread of the Gaussian probability density is determined by a hyperparameter called the perplexity. The similarity score is normalized for comparability across neighborhoods. The calculation is repeated to build a pairwise similarity matrix for every observation, each with its own set of neighbors.

Subsequently, the algorithm distributes the observations on a planar map by sampling points randomly from an isotropic Gaussian centered at the origin. A *gradient descent* method then iteratively adjusts the position of each observation in small increments until the pairwise proximity relationship matches the original. The direction of incremental movements for an observation depends on a “resultant force” that reflects the original arrangement.

In each iteration, the algorithm constructs a new pairwise similarity matrix from the planar t-SNE map. The distance transformation for that similarity matrix uses a t-distribution instead of a Gaussian distribution to produce a new normalized similarity score. Points that are very close in the original feature space is spread out a bit in the planar map because the t-distribution is a slightly “flattened” version of the Gaussian distribution. Hence, points that bunch up in the original feature space produce equal similarity scores when they are still slightly spread apart in the planar map, making them easier to see.

MDS

Unlike t-SNE, the method of MDS aims instead to preserve the *global* relationship among all observations in feature space (Heiser, 1985). The algorithm does so by solving a numerical minimization problem for *M* observations as

$$\min_{x_1,\cdots,x_M} \sum_{i<j} \left(\|x_i - x_j\| - d_{i,j} \right)^2 \tag{1}$$

where x_{ij} is the distance between the i^{th} and j^{th} observation in the reduced dimension space and d_{ij} is the distance between the corresponding observations in the original dimension feature space. The measure of similarity can be any common measure of distance in feature space, such as the Euclidean, Manhattan, Hamming, or log-fold-change distances. Selecting the best measure of distance is part of the "art" of using these algorithms. A disadvantage of using direct distances in the transformation is that there will be non-unique solutions. However, using direct distances eliminates hyperparameter settings such as the t-SNE perplexity.

UMAP

A more recently developed method is UMAP, which has gained popularity in the field of biotechnology due to its speed, reproducibility, and meaningful organization of clusters (Becht, et al., 2019). According to the inventor, UMAP works by assuming that the data is uniformly distributed on a locally connected *manifold* where distances and angles can be measured. Then, a non-linear optimization algorithm minimizes the cross-entropy between the original and the reduced dimension feature spaces. Although the manifold assumption leads to tighter clustering, the application requires several hyperparameter settings, to which the output is relatively sensitive.

3.3.3 Performance Measures

The next sections describe two measures that provide a level of confidence about the arrangement of observations in the visualized feature space: trust level and spatial autocorrelation.

Thrust Level

The trust level expresses a proportion of the local neighborhood that the transformation preserved in lower dimensional feature space (Venna & Kaski, 2001). The trust level function produces values within the range [0, 1] and its definition is

$$T(k) = 1 - \frac{2}{nk(2n-3k-1)} \sum_{i=1}^{n} \sum_{j \in \mathbb{N}_i^k} \max\big(0, (r(i,j) - k)\big) \tag{2}$$

where k is a parameter that defines the number of nearest neighbors to consider. The set $\mathbb{N}_i^k$ is the k nearest neighbors of observation i in the output space, and $r(i, j)$ is the nearest neighbor rank between observation i and neighbor j in the input space. The max function produces zero or a positive penalty when the rank is smaller or greater than k in the input space, respectively. Note that if the sum of the penalty is zero, then $\mathrm{T}(k) = 1$. For non-zero penalties, the term before the summation normalizes the measure within [0, 1]. Intuitively, the output is proportional to the amount of rank match between the input and output spaces.

Spatial Autocorrelation

Humans tend to see patterns in spatial representations of data even when there may be none. Hence, a statistical test is necessary to determine if the pattern is real, that is, different from randomness. Such a test can use a *local indicator of spatial association* (LISA) to quantify the statistical significance of clusters based on both their spatial and attribute similarities (Anselin, 1995). The Moran's I statistic is a popular LISA for evaluating the extent of significant spatial clustering of similar attributes on a map. This research uses the LISA in an unconventional manner to evaluate spatial clustering of nonlinearly projected observations with feature similarity.

The Moran's I statistic measures feature correlation among neighboring observations by

$$I_i = z_i \sum_{j \in J_i} w_{ij} z_j \tag{3}$$

where the z_i and z_j are deviations from the mean feature value for observation i and its neighbor j, respectively. J_i is the set of all neighbors for observation i and w_{ij} is the matrix of weights that define the neighbors of an observation.

The null hypothesis of the statistical test is that the cluster formation is random. A random permutation of the attributes for the neighbors of an observation produced a distribution of I_i values to test for extremeness based on a p-value. The general practice is to reject the null hypothesis (spatial randomness) when the p-value for the observed statistic is less than 0.05. However, a more conservative threshold of 0.01 was used to compensate for the tendency of a small number of common elements in adjacent neighborhoods to increase the correlation between statistic.

Observations that have statistically significant neighbors and similar features are considered to be part of that cluster. An observation is considered to be a *contaminant* of a cluster if statistically significant neighbors have different feature values. Observations that do not have statistically significant neighbors are considered *orphaned.*

4 Results

This section presents the results of the EDA and the three UML methods, KMC, and NLP. In general, the EDA methods highlight the global distribution of events by feature type whereas the UML methods provide insights about the relationships among observations and their features.

4.1 Data Wrangling

The next subsections provide a detailed chronicle of the data cleaning, feature selection, and feature extraction.

4.1.1 Data Cleaning

Table 1 provides a detailed chronicle of the GTD data cleaning activities and results. Examples of features with more than 90% missing values include those labeled as “targsubtype3,” “weapsubtype3,” “claimmode2,” “nreleased,” “nhours,” and “attacktype2.”

The 21 low dispersion variables removed included those labeled as “eventid,” “propextent,” “region,” “doubtter,” and “INT_LOG.” The GTD metadata document provides an extended description of all the labeled variables (Miller, 2020).

Table 1. GTD Data Wrangling Actions and Results

Procedures	Features	Records
Extract GTD 1970-2018 data for “country” = 217 (US)	135	2926
Load GTD 1993 (all locations)	135	748
Extract GTD 1993 based for “country” = 217 (US)	135	37
Append GTD 1993 to GTD 1970-2018 (US locations)	135	2963
Remove features with >90% missing values	73	2963
Remove features with >25% missing values	58	2963
Remove features with >4% missing values	53	2963
Remove unimportant features, and those with very low dispersion	32	2963
Remove records with both an unknown city and unknown state (1 record)	32	2962
Remove off-continental US state instances (Alaska, Hawaii, Virgin Islands, Puerto Rico)	32	2706
Fill in missing LAT/LON values for 36 records (mostly found in the isolated 1993 data)	32	2706
Replace incorrect LON values for 4 records; values were positive instead of negative	32	2706
Remove observations with feature values in an “unknown” category	32	2565
Add two letter postal code for US states	34	2565
Spatial join adds cities, counties, FIPS codes, and a unique key	37	2565

4.1.2 Feature Selection

Table 2 summarizes the final feature set and statistics after cycles of the data wrangling procedures converged. Early cycles of the EDA indicated that the latitude and longitude coordinates did not contribute to attack characteristic descriptions, so the process removed those. The final set of key features of were weapon type category (WTC), attack type category (ATC), and target type category (TTC). Table 3 enumerates the categories of the three descriptive features. The workflow eliminated categories such as "Other" and "Unknown" because they did not contribute information to the UML process.

Table 2. Features Selected or Extracted for ML.

Variable	Description	Type	Dispersion
City	City of attack	Text	-
State	State of attack	Text	-
County	County of attack	Text	-
YR	Incident Year	Ordinal	0.008
DY	Incident Day	Ordinal	0.599
MO	Incident Month	Ordinal	0.541
WTC	Weapon Type	Cat: Table	1.240
ATC	Attack Type	Cat: Table	1.310
TTC	Target Type	Cat: Table	2.350

Table 3. GTD Categorical Codes for Tactic Variables

WTC	ATC	TTC
1. Biological	1. Kill	1. Business
2. Chemical	2. Shoot	2. Government (general)
3. Radiological	3. Explode (bomb)	3. Police
4. Nuclear	4. Hijack	4. Military
5. Firearms	5. Imprison (hostage)	5. Abortion (related)
6. Explosives	6. Kidnap (hostage)	6. Aviation
7. Fake (weapons)	7. Vandalize (facility)	7. Diplomats
8. Fire (incendiary)	8. Assault (unarmed)	8. Educational Institutions
9. Melee (fight)	*9. Unknown*	9. Sustenance (food/water)
10. Vehicle		10. Media (News outlets)
11. Sabotage (equipment)		11. Maritime
12. Other		12. Non-Gov. Organization
13. *Unknown*		*13. Other*
		14. Person (individuals)
		15. Worship (related)
		16. Telecommunication
		17. Militias
		18. Tourists
		19. Transportation (non-aviation)
		20. Unknown
		21. Utilities
		22. Extremists (political parties)

4.1.3 Feature Extraction

Ten PMCs were defined by cognition after observing the word cloud of keywords. Table 4 summarizes the PMCs and their associated keywords. All misspellings and unusual words or names were retained to maximize the completeness of the category assignment. The largest category of “rule” labeled events that were associated with sentiments reacting to authoritative actions from legislative, military, law enforcement, and political entities. The second largest category of “race” labeled events that were associated with racial prejudices or xenophobic ideologies. The other categories were much less frequent, and their sentiments were more homogeneous. In cases where the category assignment was ambiguous, for example “Jewish” as race or religion, it was necessary to use human interpretation (empirics) of the word relevance by examining the context from the full text narrative.

Table 4. Keywords Separating the Perpetrator Motivation Categories

PMC	Keywords
Race	African, Afrika, Airport, Airports, Al-Qaida, Arab, Armenia, Armenian, Aryan, Black, Brotherhood, Chicano, Chinese, Croatian, George Jackson, Immigrant, Incel, Indian, Iranian, Iraqi, Islamic, Islamist, Jewish, Jihadi, Klux, Lebanese, Muslim, Nazi, Order II, Otpor, Pakistan, Qaddafi, Rajneeshees, Semitic, Sikh, Skinheads, White, Zebra, nationalist, white
Rule	Anarchists, Armenians, BAY Bombers, Castro, Charles Manson, Communist, Comrades, Confederate, Conspiracy, Court Reform, Cuba, Cypriot, Drug Cartel, East Side, FIN, Fascism, Fifth Battalion, Fred Hampton, Freedom, Gaddafi, Gestapo, Government, Guerilla, Gun, Hatikvah, ISIS, Iran, Israel, Jackson Brigade, Judicial, June Organization, Kahane, Kim Jong, LAACA, Liberal, Liberation, Luis Boitel, MEK, Maccabee, Macoute, Militant, Military, Minutemen, NGO, NWLF, New Year, Nuclear, Omega-7, Palestine, Palestinians, Phineas, Phong, Police, Political, Posse, Puerto Rican, Putin, Quartermoon, Regulators, Republic of Texas, Republican, Revolutionary, Scorpion, Secret Army, Secret Organization, Serbian, Socialist, Sovereign, Squad, Star, Student, Symbionese, Trump, Veterans, Vietnamese, WUFI, War, Weathermen, Zion, legislation, political, wing
Sex	Gay, LGBT, Male, Female, Woman
Environment	Anti-gentrification, EMETIC, Earth, Environment, Genetic, Oil, Preserves, Trees, Utilities, construction, housing, river-saving, transmission
Religion	Christian, Church, God, Jamaat, Lord, Mormon, Religious
Abortion	Abortion, Christian Liberation
Money	Liberty, Liberacion, IRS, Justice, Business
Animal	Animal, bacon, Strikers, Veterinary
Modernization	Technology, George Jackson, Unaffiliated Individual
Labor	Union, Teamster

4.2 Exploratory Data Analysis

The next subsections examine the frequency and temporal distribution of each feature for the U.S. subset of the GTD from 1970 to 2018.

4.2.1 Multivariate Distributions

Only a few categories within each feature dominated the U.S. subset of the GTD. Figure 3a shows the frequency of each target type, proportioned by attack type. Figure 3b shows the frequency of each attack type, proportioned by PMC. Figure 3c shows the frequency of each PMC, proportioned by attack type.

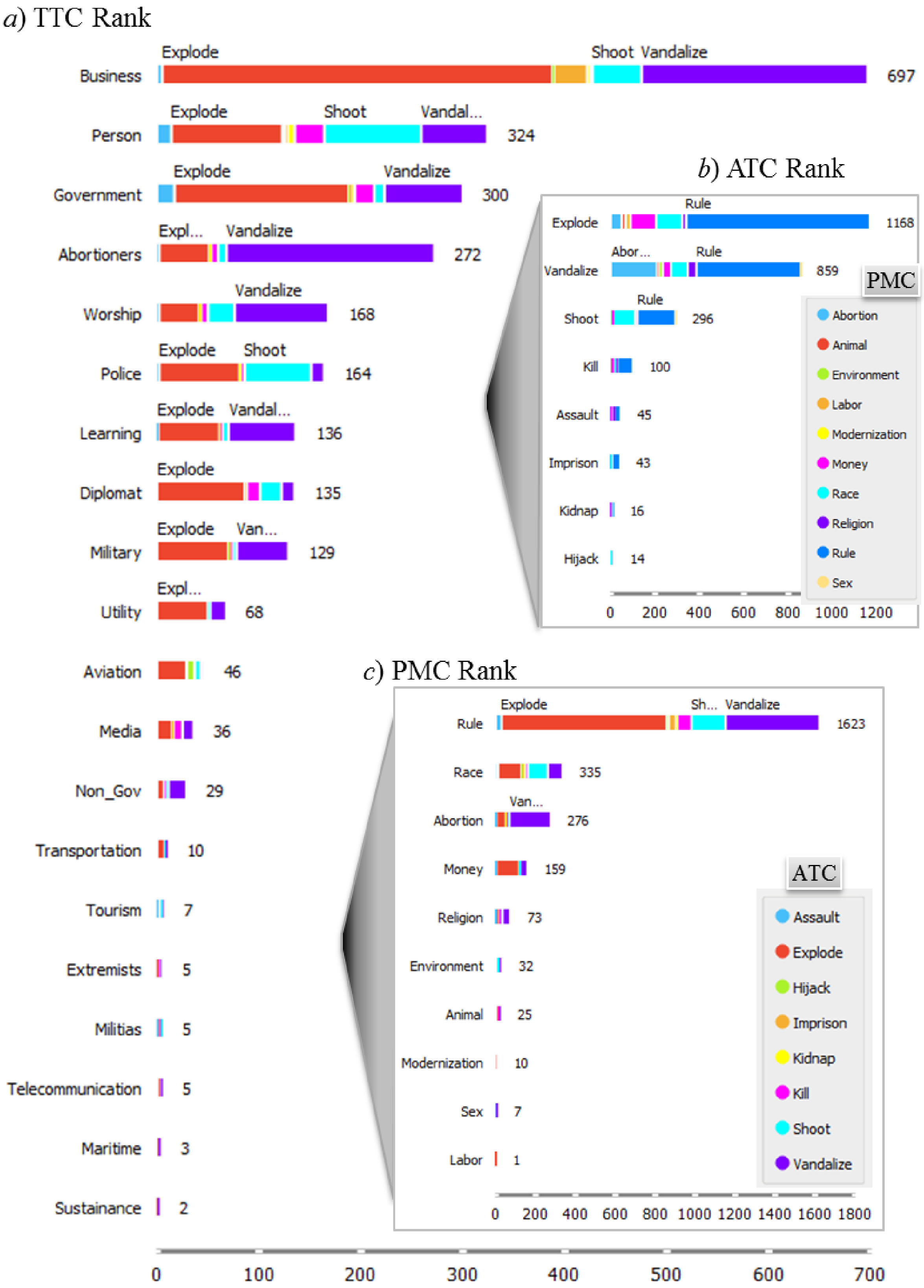

Fig 3 Ranked distribution of a) TTC, b) ATC, and c) PMC in the GTD.

The top three PMCs were rule, race, and abortion. They accounted for 88% of the terrorist activities that occurred in the United States. Therefore, the UML focused on analyzing relationships among features of only the top three PMCs.

The cross-sectional proportions of pairwise features are not easy to see in Figure 3, so Figure 4 and Figure 5 show them as proportions.

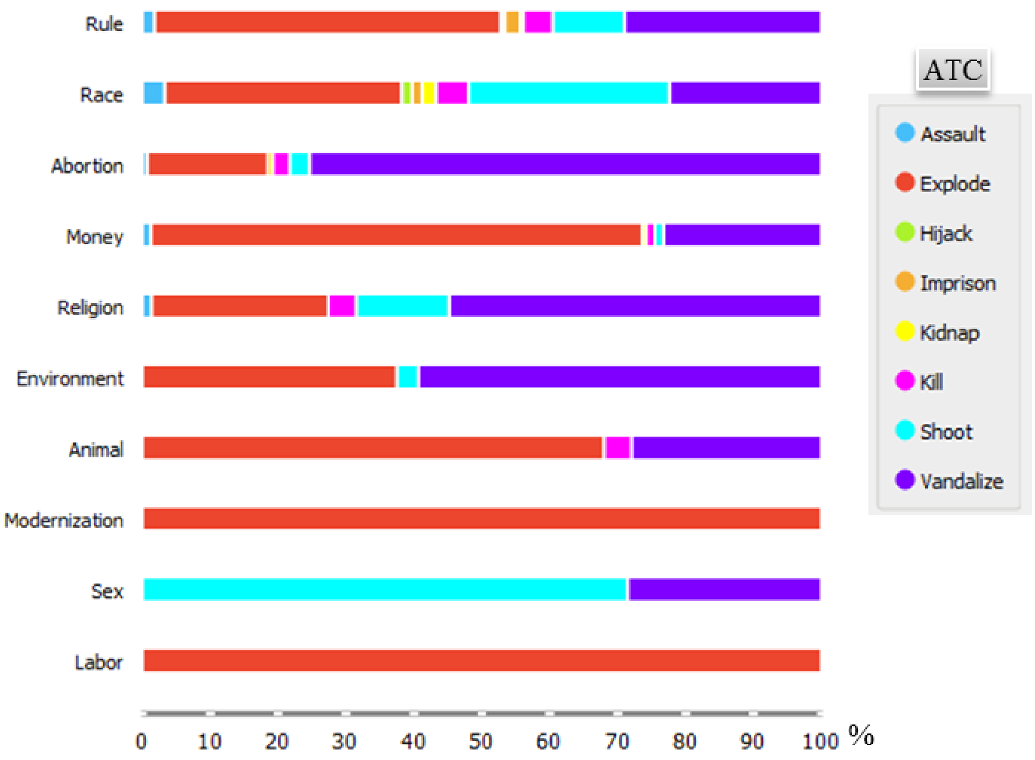

Fig 4 Proportional distribution of tactic by perpetrator motive category.

Explosives accounted for the largest proportion of attacks associated with the PMC categories of rule and race. The attack proportions for rule and race were 51% and 35% of the observations, respectively (Figure 4). Attacks associated with abortion-related sentiments used incendiary devices to vandalize facilities in more than 75% of the cases (Figure 4). As expected, attacks associated with abortion-related sentiments almost exclusively targeted abortion-related entities, collectively labeled as “abortioners” for lack of a better word (Figure 5).

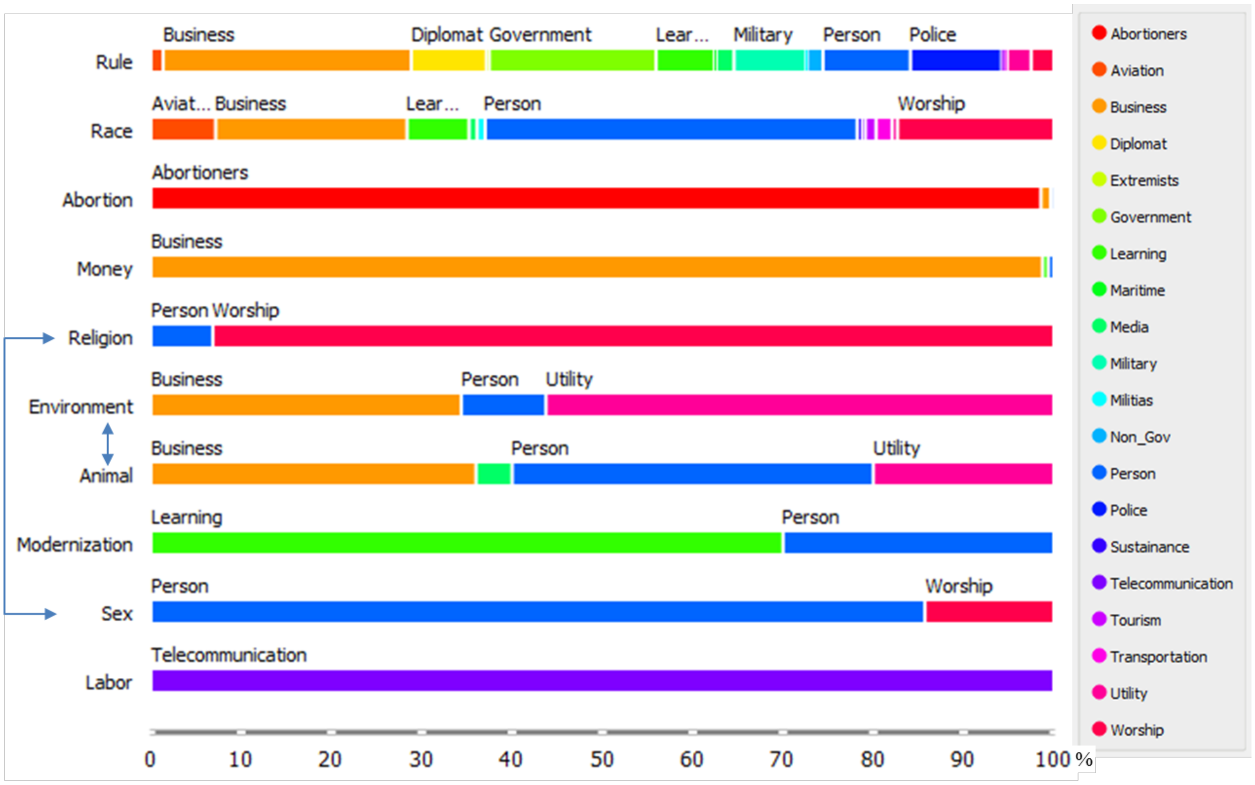

Fig 5 Proportional distribution of attack targets by perpetrator motive category.

The largest proportion of attacks associated with rule and race sentiments targeted businesses (28%) and people (41%), respectively (Figure 5). Table 5 summarizes the proportions of attacks within the top three categories of the PMC feature, and their associated top three categories within the ATC and TTC features.

Table 5. Proportions of the Top Three Categories in the PMC, ATC and TTC features

			ATC			**TTC**	
PMC	Proportion	Explode	Vandalize	Shoot	Business	Person	Abortion
Rule	**0.639**	**0.510**	0.290	0.105	**0.275**	0.097	0.000
Race	0.132	**0.349**	0.224	0.105	0.212	**0.412**	0.000
Abortion	0.109	0.178	**0.754**	0.029	0.011	0.004	**0.986**
Data %	0.879	0.460	0.338	0.116	0.274	0.128	0.107

There were some similarities and differences in the distribution of ATC categories across the PMC categories. Based on a chi-squared test, the distribution was statistically different within the top three PMC categories. However, the distributions were statically similar among five of the less-frequent PMCs, namely environment, religion, abortion, animal, and money. Figure 6 compares the proportional distributions for three pairs of PMCs that had statistically similar ATC distributions.

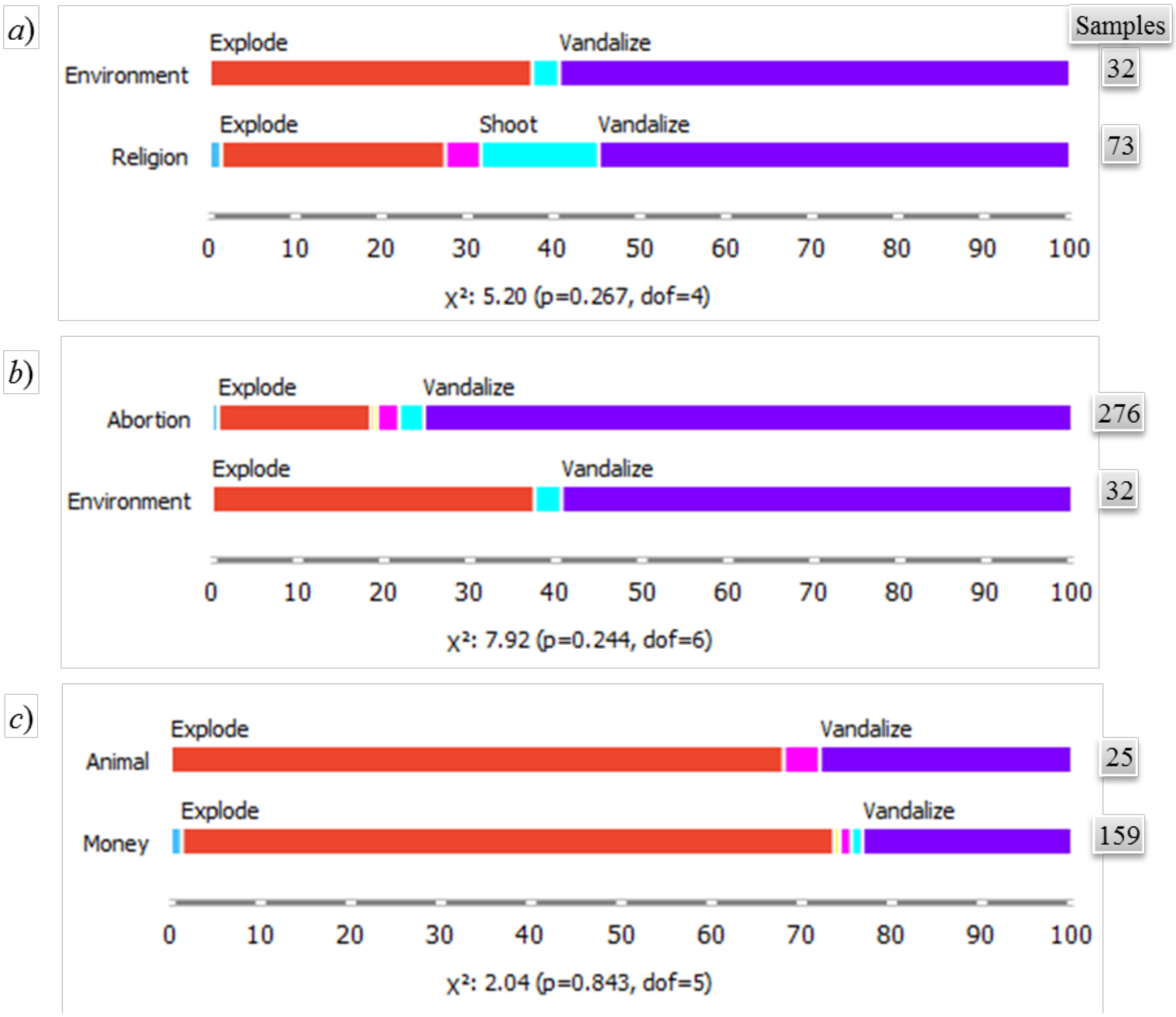

Fig 6 Similarity measures of ATC distribution across significant PMC.

The bottom of each chart is annotated with the number of observations per category, the degrees-of-freedom (dof) for the chi-squared distribution, the value computed for the chi-squared statistic, and the p-value of the chi-squared statistic. The standard practice is to reject the hypothesis that the values distribute similarly if the p-value is less than 0.05 or otherwise accept the alternative hypothesis that there is no significant difference their distributions. The results are summarized as follows:

1. The attack types distribute similarly for sentiments related to environment and religion with a statistical significance of p = 0.267 (Figure 6a).
2. The attack types distribute similarly for sentiments related to abortion and environment with a statistical significance of p = 0.244 (Figure 6b).
3. The attack types distribute similarly for sentiments related to animal and money with a statistical significance of p = 0.843 (Figure 6c).

4.2.2 Temporal Waves of Sentiments

The statistical distribution of PMCs for attacks within the United States show that there were distinct waves of sentiments over time. Figure 7 illustrates the trends with a boxplot that shows several pieces of statistical information:

1. The blue boxes span the first (25%) to third (75%) *quartile*.
2. The short vertical blue line within each box marks the *mean* of the distribution.
3. The horizontal blue line spans the *standard deviation*.
4. The horizontal dotted blue line spans the *extent* of observations (minimum to maximum).
5. The long vertical gray lines that intersect the horizontal axis are the *median* values.

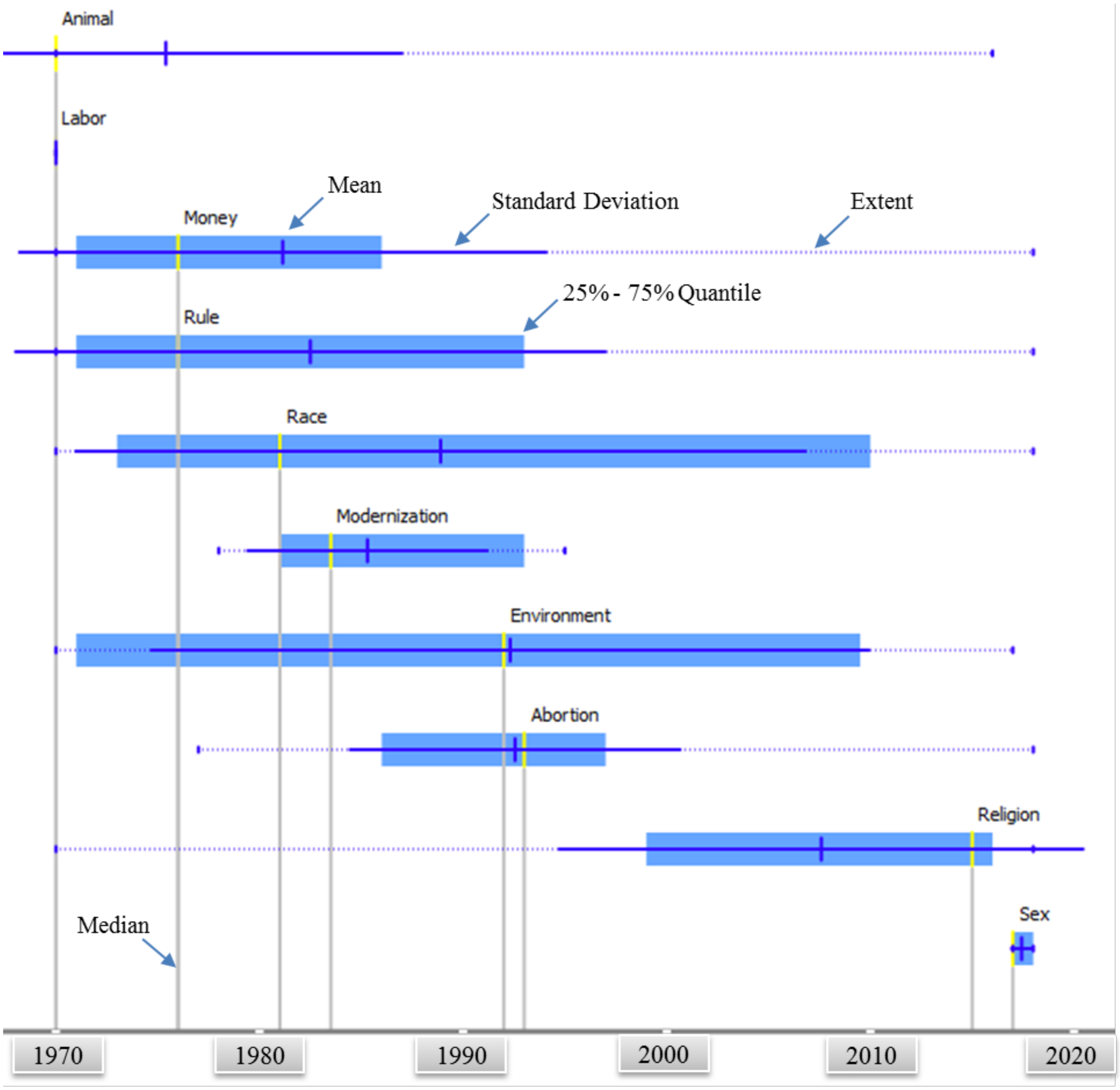

Fig 7 Temporal shifts in sentiments.

4.3 Unsupervised Machine Learning

This section summarizes the results of applying the UML techniques of KMC and NLP to explore relationships among the features.

4.3.1 K-Means Co-clustering

Associative relationships among the dominant features can be inferred from the co-clustering of observations and their features. Figure 8 shows a co-clustered heatmap using the spectral color scheme of a rainbow. That is, the highest and lowest frequency groupings are colored red and blue, respectively, with intermediate colors shown as green and yellow. The average from 10-means clustered observations produced the three observation clusters R1 to R3. Clustered features appear in the five column groups of C1 to C5.

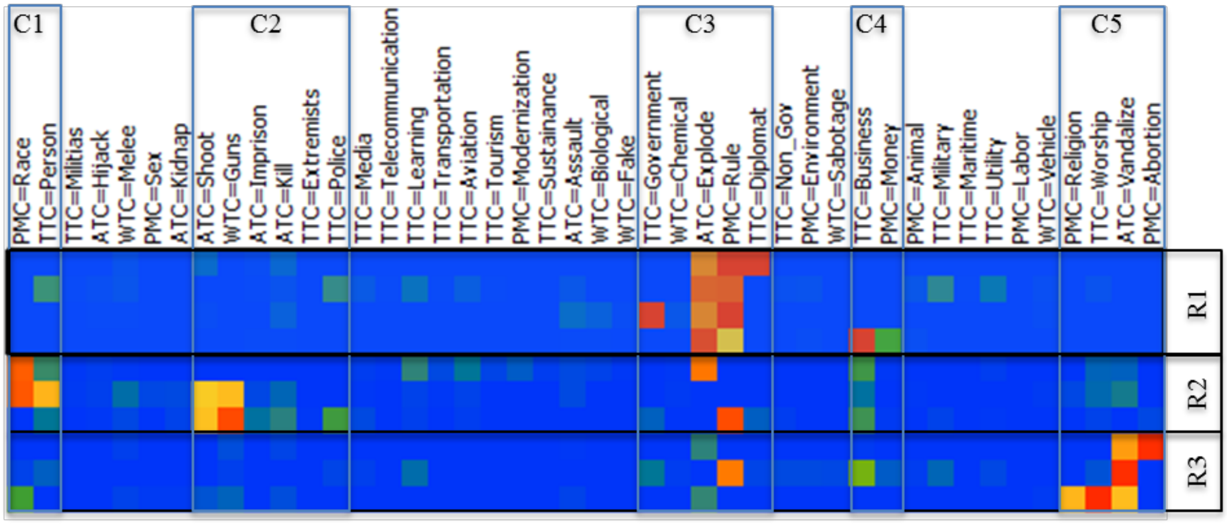

Fig 8 Heatmap with k-means clustering of the rows and columns.

Table 6 summarizes the dominant associative relationships across features, within each of the observation clusters. The PMC of “Rule” and the TTC of “Business” were common across the three observation clusters. Hence, the other categories served to distinguish the clusters. Figure 9 shows two sets of charts for each of the observation clusters. The chart on the left is the cross-sectional distribution of PMC features, split by ATC. The chart on the right is the corresponding ATC frequency distribution. The figure does not show the corresponding TTC distributions, but Table 6 summarizes them.

Table 6. Data (row) and Feature Cluster (column) Characteristics

	PMC	**ATC**	**TTC**
R1	Rule, Money	Explode	Business, Diplomat, Government
R2	Rule, Race	Shoot, Explode, Vandalize	Business, Person, Police
R3	Rule, Abortion, Religion	Vandalize	Business, Abortion, Worship

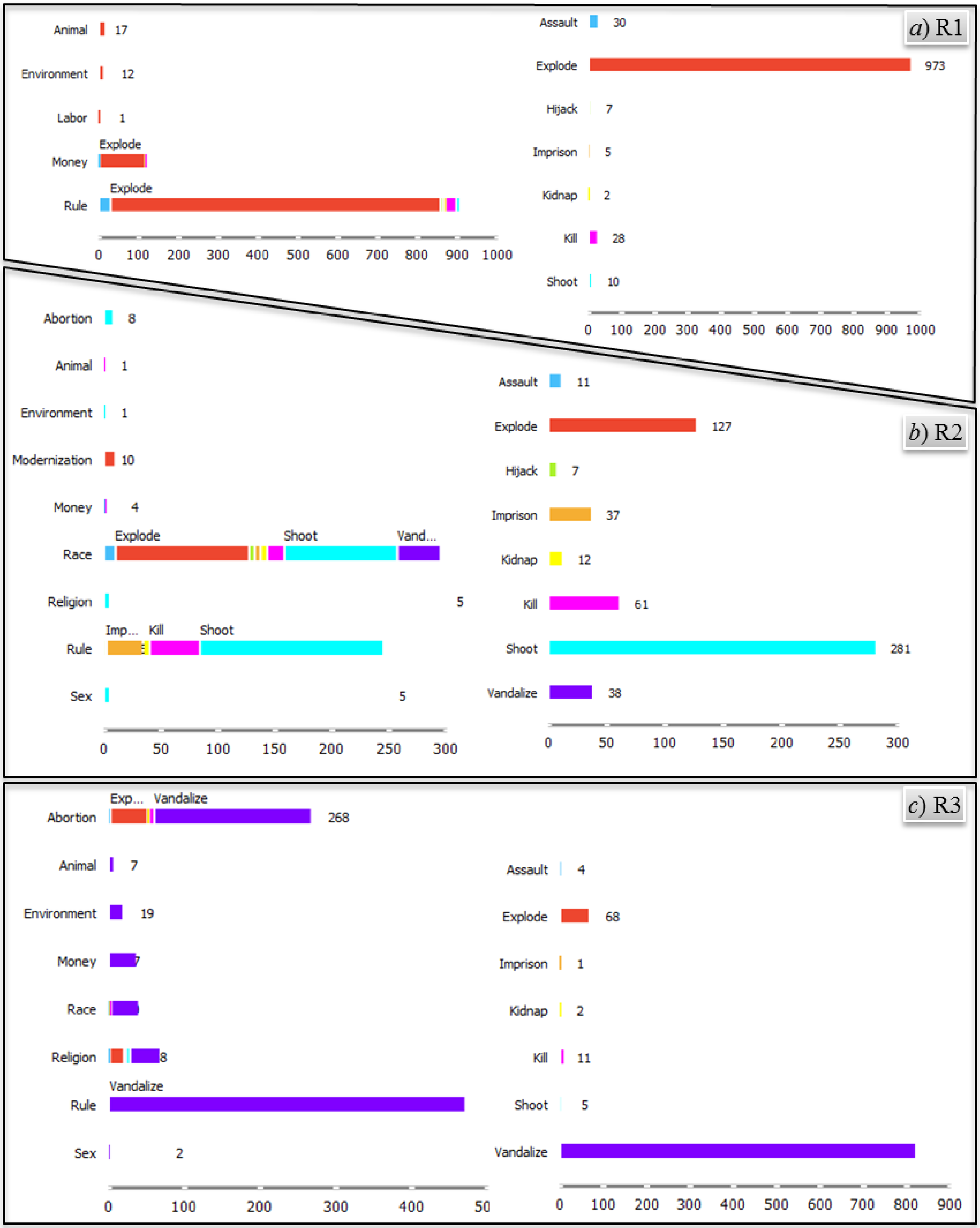

Fig 9 Cluster characteristics.

4.3.2 Nonlinear Projections

The EDA methods highlighted that the dominant PMCs (rule, race, abortion) accounted for most of the terrorist attacks but could not show clustered relationships between their attack and target types. In contrast, the UML methods revealed associative relationships between the dominant PMCs, ATCs, and TTCs. Figure 10 and Figure 11 are NLP visualizations for all observations and the four categorical features of the dataset. The rows of charts in those two figures provide visualizations for the four categorical features, including WTC. The three columns in those two figures show visualizations based on each of the three NLP methods. Table 7 summarizes for each NLP method the proportion of trust in neighborhood preservation, the computational time required, and the optimum hyperparameter settings.

Table 7. NLP Performance Measures and Hyperparameter Settings

Method	**Trust (%)**	**Time (s)**	**Hyperparameters**
tSNE	98.0	81.4	P=95, EE=12, Method=Exact, Init=PCA, Iterations=300, Learn Rate=200
MDS	96.9	49.5	Max Iteration=300
UMAP	97.3	35.0	Neighbors=39, Min Distance=1

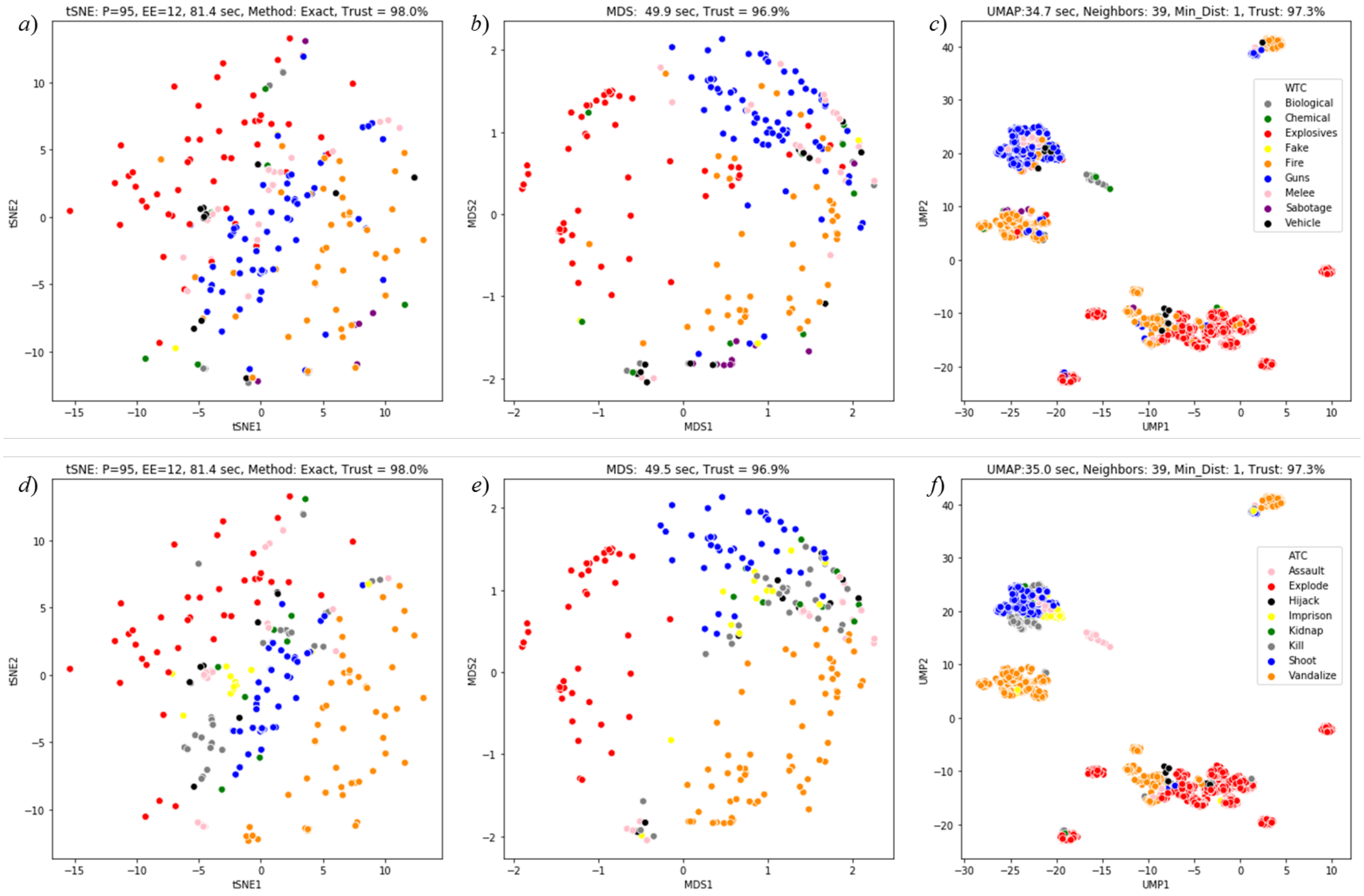

Fig 10 NLP visualizations for WTC (a-c) and ATC (d-f) using tSNE, MDS, and UMAP.

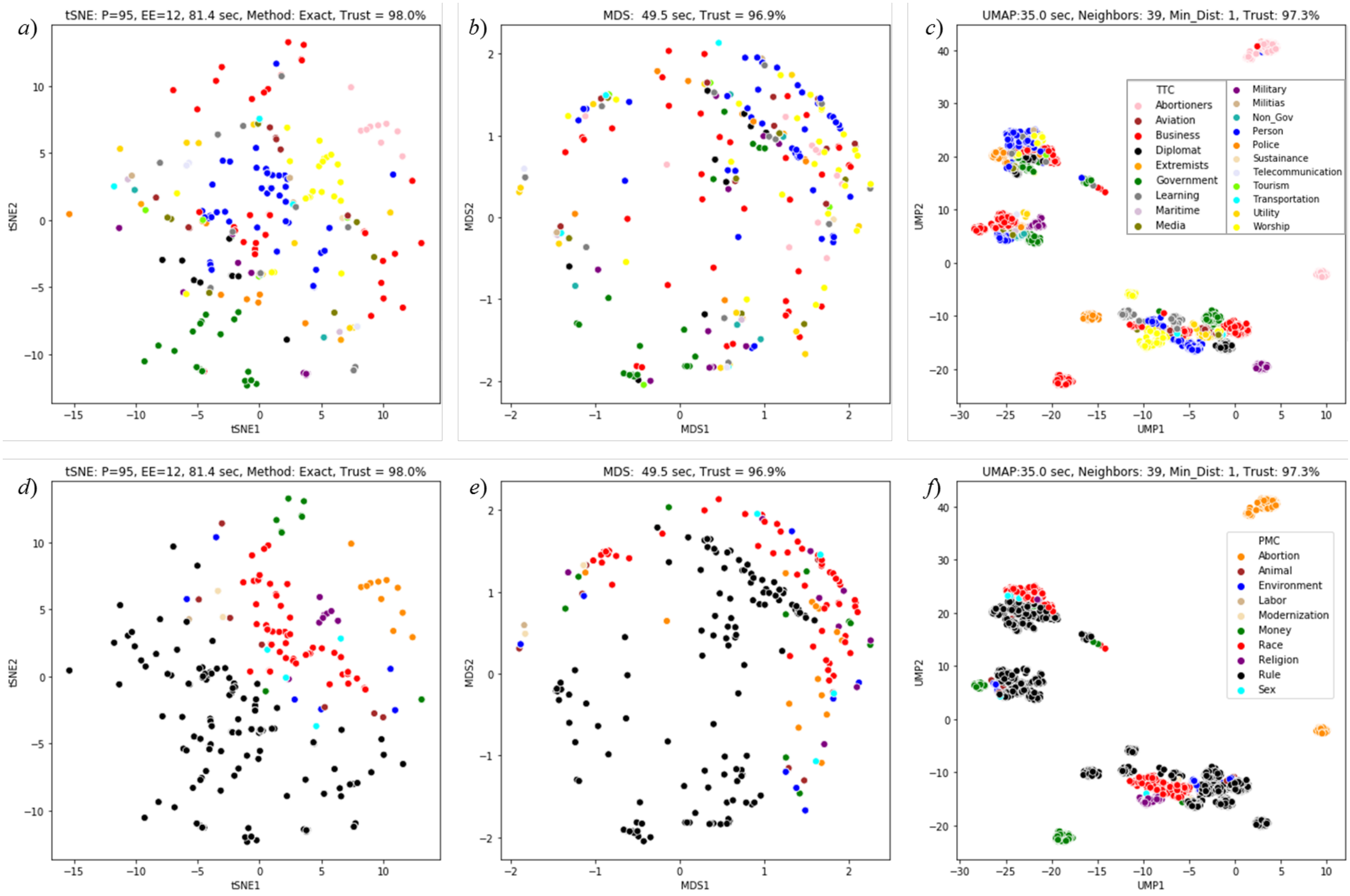

Fig 11 NLP visualizations for TTC (a-c) and PMC (d-f) using tSNE, MDS, and UMAP.

The two dominant values of ATC, TTC, and PMC were selected to focus the statistical tests for clustering by using the three NLP methods. The reasons were as follows:

1. The EDA results showed that the top three PMCs accounted for a majority (88%) of the dataset.
2. The EDA results showed that the third dominant PMC (abortion-related) had a nearly exclusive associative relationship with the target type of abortion-related facilities, and attack types involved mostly incendiary devices (75%) or explosives (18%).
3. The TTC category of "Abortion" was removed due to a 99.2% correlation with the PMC feature of "Abortion"
4. The WTC categories of "Explosives" and "Fire" were removed because of their high correlation (92%) with the ATC categories of "Explode" and "Vandalize."
5. The first three PCs load the top two PMCs and the top three ATCs, accounted for nearly half (48%) of the variance in the dataset.

Table 8 summarizes the clustering identified by the three NLP methods for the two dominant categories in the three features of PMC, ATC, and TTC. The numbers in the table are the proportions of all the data that formed significant clusters (Clu.) for each category, were orphaned (Orp.), and contaminated (Con.) clusters. The proportions are based on a measure of spatial autocorrelation by the Moran's I statistic. The two dominant and statistically significant clusters within the PMC and ATC features accounted for an average of 74.6% and 79.2% of the GTD, respectively. The two dominant and statistically significant clusters within the TTC feature accounted for an average 37.5% of the GTD.

Table 8. NLP Performance Measures by Cluster Significance

	tSNE			MDS			UMAP		
Variable	Clu.	Orp.	Con.	Clu.	Orp.	Con.	Clu.	Orp.	Con.
ATC=Explode	0.452	0.009	0.015	0.458	0.013	0.000	0.459	0.020	0.003
TTC=Business	0.253	0.057	0.017	0.253	0.089	0.015	0.264	0.086	0.011
PMC=Rule	0.627	0.018	0.002	0.621	0.027	0.002	0.600	0.061	0.003
ATC=Vandalize	0.338	0.011	0.012	0.335	0.018	0.001	0.335	0.029	0.006
TTC=Person	0.118	0.110	0.077	0.116	0.123	0.067	0.121	0.302	0.022
PMC=Race	0.132	0.072	0.013	0.129	0.094	0.019	0.129	0.269	0.010
ATC	0.790	0.020	0.027	0.793	0.031	0.001	0.794	0.049	0.009
TTC	0.371	0.167	0.094	0.369	0.213	0.082	0.385	0.388	0.033
PMC	0.759	0.091	0.014	0.750	0.121	0.021	0.729	0.329	0.013

All methods identified statistically significant clusters for a similar proportion of the top two categories of each feature with an average correlation of 99%. Feature mapping transformations by tSNE and UMAP orphaned the least and most proportion of the observations, respectively. However, UMAP transformed the least proportion of observations as cluster contaminants.

The fastest method was UMAP, which took 35 seconds on a 64-bit Windows 7 machine with 32 GB of RAM and an Intel® Core i7-4790 CPU clocked at 3.6 GHz. MDS and tSNE took 41.4% and 132.6% more time than UMAP, respectively.

5 Discussion

The next subsections discuss the results of the ETM, EDA, and the three methods of UML.

5.1 Empirical Text Mining

The main objective of creating the PMCs was to minimize the number of categories while making the sentiments distinct. This strategy resulted in a few dominant categories and several smaller ones. The PMC of "rule" was the largest category in the U.S. subset of the GTD. The "rule" category reflected reactions to authority and policies enacted by all governments or departments in the United States, including the U.S. Congress, Executive Branch, Supreme Court, Internal Revenue Service, military, state governments, and local governments. The "rule" category also reflected reactions to enforcement of "rules" by those with power or authority.

Examples include sanctions against foreign nations and related military deployments. The “rule” category also reflected anti-government sentiments or political extremism.

The next largest PMC was “race” which reflected mostly racial and xenophobic sentiments. In cases where there was some overlap with “religion” the empirical approach more heavily weighed sentiments associated with sovereign nation states towards the race category. The “religion” category also had some overlap with the “abortion” and “rule” categories but was more heavily weighted towards “abortion” when the target type was abortion related. There were relatively few observations in the remaining categories such as “labor” and “modernization,” which simplified the analysis.

5.2 Exploratory Data Analysis

The next two subsections discuss the observations from the multivariate distributions of features, and the temporal analysis of sentiment span.

5.2.1 Multivariate Distributions

The main observation from the multivariate distributions was that the top three PMCs (rule, race, abortion) accounted for most (88%) of the terrorist attacks in the United States between 1970 and 2018. Nevertheless, some additional insights can be gleaned for the remaining categories within each feature by jointly observing the following patterns across Figures 4 through 6:

1. The targets of persons and places of worship were similarly homogeneous for sentiments related to sex and religion. This suggests a hypothesis that those sentiments are more related to each other than to the others.
2. Sentiments linked to either the environment or animals were more associated with attacks on utilities than other sentiments (Figure 5). This suggests a hypothesis that there is a

relationship between sentiments linked to environmental stewardship and to animal rights.

3. Sentiments linked to modernization were more associated with targeting institutions of learning than others (Figure 5). This suggests a hypothesis that heightened sentiments about advancements in technology can increase the risk of attacks on institutions of learning.
4. Sentiments linked to money were almost exclusively associated with the use of explosives or incendiary devices to target businesses. This suggests a hypothesis that escalating financial issues can increase the risk of attacks on businesses.
5. Sentiments linked to abortions were almost exclusively associated with targets that are abortion related, which is consistent with expectations.

Future work will explore statistical tests on the hypothesis identified above, for the entire database, to evaluate any regional differences in associations and outcomes.

5.2.2 Temporal Waves of Sentiments

Some of the temporal changes in sentiment, such as money- and abortion-related, were more pronounced than others. Their means and medians coincided with the historical heights of those sentiments in the United States. For example, during the 1970s, which was a time of high inflation and high interest rates, there were widespread protests of the local property tax (Martin, 2008). Similarly, during the 1980s, the rapid evolution and commercialization of personal computers confirmed Moore’s law that predicted a sustainable doubling of transistor density every two years (Schaller, 1997). This corresponded with an increase of attacks associated with modernization and automation sentiments. Religion related sentiments ramped around the year 2000, which was punctuated by the September 2001 attacks. Abortion-related attacks centered around the 1992 landmark U.S. Supreme Court case on abortion. There is a large standard deviation for sentiments linked to race and the environment, which is not surprising based on their long-standing prominence throughout U.S. history. Future work will examine how similar sentiments emerged and dissipated in time for other regions of the world.

5.3 Unsupervised Machine Learning

EDA provided a global view of how the features of terrorist attacks distributed and revealed those that were dominant. It was also possible to hypothesize global relationships based on the similarity of feature distributions. However, beyond a summary of the data and some hypothesis to explore, EDA alone could not provide insights about underlying relationships among observations and features based on local structure and proximity in feature space. The next three subsections discuss how the three UML methods provided such insights.

5.3.1 K-Means Co-Clustering

The intersection of the three observation clusters (R1 to R3) and the five feature clusters (C1 to C5) revealed the following associative relationships, also summarized in Table 6:

- The R1/C3 cluster characteristics highlighted similarities between sentiments linked to rule and the targeting of governments and diplomats (government representatives) with explosives.
- The R1/C4 cluster associated sentiments linked to money with attacks on businesses.
- The R2/C1 cluster associated sentiments linked to race with attacking people.
- The R2/C2 and R2/C3 clusters associated sentiments linked to rule with attacking police and businesses with guns and explosives.
- The R3/C5 cluster associated sentiments linked to religion and abortion with attacking places of worship using incendiary devices.

- The R3/C3 cluster associated sentiments linked to rule with attacking government interests using explosives.

Sentiments related to rule and business targets were common in each observation cluster (R1, R2, R3.) The PMCs that separated the observation clusters were money, race, and abortion (coupled with religion.) The ATCs that separated the observation clusters were explode, shoot, and vandalize. The TTCs that separated the observation clusters were government (coupled with diplomat), person (coupled with police), and abortion (coupled with worship.) In conclusion, the differentiating themes across observation clusters were that sentiments linked to money, race, and religion (coupled with abortion) were associated with attacks on businesses, persons (coupled with police), and places of worship (coupled with abortion-related facilities), respectively. The corresponding attack types on governments, persons, and places of worship used explosives, guns, and incendiary devices, respectively.

The clusters and their associative relationships, highlighted by the co-clustering, were qualitative. The NLP methods, discussed later, provided a means to visualize the intersections of clustered observations and features. The also enabled quantitative statistical tests to assess whether or not the clusters observed were real.

5.3.2 Nonlinear Projections

The EDA and NLP methods provided complementary benefits. The EDA methods were useful for summarizing the global patterns in the data and to formulate hypothesis based on them. The NLP methods produced insights into the relationships among features based on clustered observations. Given an NLP method, the associative relationships among features (PMC, WTC, ATC, TTC), as revealed by the KMC method, corresponded to the intersections of those feature clusters across the NLP maps. The two dominant and statistically significant clusters within the key feature categories accounted for the main structure or “signal” in the dataset. The NLPs also revealed some additional structure for the less dominant features. For example, the attack types of assault, imprison, and kill clustered near or within the shoot cluster. Another example was observed in the WTC maps where the “Vehicle” weapon type clustered near the “Fire” weapon type, suggesting a closer relationship than with other weapon types.

Although local clustering could be observed visually, it was important to determine if they were real by applying a statistical test. The Moran’s I statistic was used with a conservative p-value to determine the proportion of observations that clustered with statistical significance into the top two categories of the three features of PMC, ATC, and TTC. The three methods of tSNE, MDS, and UMAP agreed on cluster significance with slight differences in the proportion of orphaning and contamination. There was an apparent tradeoff between orphaning and contamination because tSNE had the highest proportion of contaminants but the least proportion of orphans whereas UMAP expressed the converse result.

The relative positions of the larger clusters in feature space indicated a similarity or opposing relationship among them. For example, the attack types of “Explode” and “Vandalize” tended to be opposite but adjacent in the maps of MDS and UMAP, and the attack type of shoot tended to bifurcate them at one end of the tSNE map. Isolated TTC clusters were more evident with UMAP than with the other two methods. For example, the “Abortion” and “Military” clusters were isolated at opposite ends of the map, and they were homogeneous.

Overall, the three NLP methods provided complementary performance in visualization, clustering, and speed. Although UMAP was the fastest algorithm, it was more sensitive to the “Neighbors” setting than tSNE was sensitive to the “Perplexity” or P setting. MDS had the fewest hyperparameters to tune. Principle component analysis (PCA) initialization provided the most stable performance for tSNE. The trust level was higher than 97% for each NLP method. Therefore, the confidence was high that each method equally preserved the similarity relationship of local neighborhoods in high dimensional feature space. UMAP clusters were generally more compact and separated than those of either tSNE or MDS. Each method tended to focus cluster formation on the ATC values where even the less-dominant categories such as "assault" and "imprison" were distinguishable. The methods of tSNE and UMAP tended to better cluster the TTC values than MDS. With both tSNE and MDS, the PMC of "rule" clustered to one side of the map whereas UMAP fragmented that cluster.

5.4 Limitations

Limitations of this work stem from manual interventions to adjust parameters of the UML methods and to interpret their output. The methodology requires substantial human intervention to converge the cyclical process of the workflow and analysis framework. This reliance on the intuition and experience of experts helps but it also increases the difficulty of eventually automating a process that can be tedious and time consuming, especially when applied to exceptionally large databases. Another example of the required manual intervention is the resolution of ambiguous word meaning. In those cases, a human needed to further examine the word usage context in the entire text narrative. For the uncommon cases that involved more than one PGN, the procedure selected the first mentioned group. This decision was based on the empirical observation that each group reflected a similar or related sentiment. Hence, selecting one group over the other did not change the PMC label assigned.

There are numerous of methods of UML and no single method works best on any given dataset (Aggarwal, 2015). This analysis selected co-clustering because it is based on k-means clustering, which is one of the simplest yet most powerful method of UML. An exhaustive comparison of all known UML methods would have distracted from the main purpose and contribution of this paper. In this social science application of UML, the combination of co-clustering with non-traditional methods of spatial analysis like nonlinear projections and statistical tests of spatial association (Moran's I) was enough to reveal the patterns of data associations discussed.

6 Conclusions

It is difficult to predict terrorism with mathematical models because of the complex and adaptive behaviors of perpetrators. Therefore, counterterrorism research can benefit from machine learning (ML) techniques that can mine large multidimensional datasets to advance knowledge about terrorism. With the rapid evolution of storage and computing capacity, ML techniques continues to evolve along its main branches of supervised ML (SML) and unsupervised ML (UML) methods. Unlike the former, the latter does not require a labeled dataset, which can be tedious and expensive to create. However, there are still far fewer UML model types and their outputs are more difficult to interpret. Consequently, there has been relatively few investigations of UML to derive insights from the Global Terrorism Database (GTD®).

This research created a new categorical variable for the GTD by applying empirical text mining to the narrative fields. The result was a new categorical feature defined as the perpetrator motive category (PMC) that contained ten categories. The subsequent application of k-means co-clustering revealed an associative relationship between the PMCs, tactics, and targets for 88% of the terrorist attacks that occurred in the continental United States from 1970 to 2018. Three methods of nonlinear projection (NLP) showed that the relationships clustered in multidimensional feature space. The Moran's I test for spatial autocorrelation validated that the observed clustering of the dominant features was statistically significant. Clusters of the two dominant categories in PMC, attack type, and target type accounted for 74.6%, 79.2%, and 37.5% of the GTD, respectively. The main relational associations observed was that attacks motivated by sentiments associated with rules targeted business and government interests with explosives or incendiary devices. Conversely, the dominant attack type was shooting for sentiments associated with race.

These results demonstrated the effectiveness of using UML techniques to benefit counterterrorism research by exposing statistically significant patterns in large multidimensional databases. Such patterns could inform counterterrorism strategies in surveillance, intelligence, and deterrents. Future work will utilize the presented framework of UML techniques to explore worldwide behavioral patterns for a subset of target types, for example, transportation. An investigation of the hypothesis previously discussed about the similarities in feature distribution for some of the less frequent motive categories has begun for the international dataset.

7 Declarations

Funding: None

Conflicts of interest/Competing interests: None

Availability of data and material: Publicly available.

8 References

Abrahms, M., & Conrad, J. (2017). The Strategic Logic of Credit Claiming: A New Theory for Anonymous Terrorist Attacks. *Security Studies, 26*(2), 279-304. doi:10.1080/09636412.2017.1280304

Adnan, M., & Rafi, M. (2015). Extracting Patterns from Global Terrorist Dataset (GTD) Using Co-Clustering Approach. *Journal of Independent Studies and Research, 13*(1), 7. doi:10.31645/jisrc/(2015).13.1.0002

Aggarwal, C. C. (2015). *Data Mining.* New York, New York, United States of America: Springer International Publishing.

Agresti, A. (2018). *Statistical Methods for the Social Sciences* (5th ed.). Boston, Massachusetts, U.S.: Pearson.

Aleroud, A., & Gangopadhyay, A. (2018). Multimode co-clustering for analyzing terrorist networks. *Information Systems Frontiers, 20*(5), 1053-1074. doi:10.1007/s10796-016-9712-4

Ammar, J. (2019). Cyber Gremlin: social networking, machine learning and the global war on Al-Qaida-and IS-inspired terrorism. *International Journal of Law and Information Technology, 27*(3), 238-265. doi:10.1093/IJLIT/EAZ006

Anselin, L. (1995). Local Indicators of Sspatial Association—LISA. *Geographical Analysis, 27*(2), 93-115. doi:10.1111/j.1538-4632.1995.tb00338.x

Atsa'am, D. D., Wario, R., & Okpo, F. E. (2020). A New Terrorism Categorization Based on Casualties and Consequences using Hierarchical Clustering. *Journal of Applied Security Research*, 1-16. doi:10.1080/19361610.2020.1769461

Bayar, Y., & Gavriletea, M. (2018). Peace, terrorism and economic growth in Middle East and North African countries. *Quality & Quantity, 52*(5), 2373-2392. doi:10.1007/S11135-017-0671-8

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., . . . Newell, E. W. (2019). Dimensionality Reduction for Visualizing Single-Cell Data using UMAP. *Nature Biotechnology, 37*(1), 38-44. doi:10.1038/nbt.4314

Campedelli, G., Bartulovic, M., & Carley, K. (2021). Learning future terrorist targets through temporal meta-graphs. *Scientific Reports, 11*(1). doi:10.1038/s41598-021-87709-7

Campedelli, G., Cruickshank, I., & Carley, K. (2021). Multi-modal Networks Reveal Patterns of Operational Similarity of Terrorist Organizations. *Terrorism and Political Violence*, 1-20. doi:10.1080/09546553.2021.2003785

Campedelli, G., Cruickshank, I., & M. Carley, K. (2019). A complex networks approach to find latent clusters of terrorist groups. *Applied Network Science, 4*(1). doi:10.1007/s41109-019-0184-6

Clauset, A., & Wiegel, F. W. (2010). A Generalized Aggregation-Disintegration Model for the Frequency of Severe Terrorist Attacks. *Journal of Conflict Resolution, 54*(1), 179–97. doi:10.1177/0022002709352452

Conlon, S., Abrahams, A., & Simmons, L. (2015). Terrorism Information Extraction from Online Reports. *Journal of Computer Information Systems, 55*(3), 20-28. doi:10.1080/08874417.2015.11645768

Curia, F. (2020). Unsupervised Hybrid Algorithm to Detect Anomalies for Predicting Terrorists Attacks. *International Journal of Computer Applications, 176*(35), 975: 8887. doi:10.5120/ijca2020920432

Ding, F., Ge, Q., Jiang, D., Fu, J., & Hao, M. (2017). Understanding the Dynamics of Terrorism Events with Multiple-Discipline Datasets and Machine Learning Approach. *PLoS ONE, 12*(6), e0179057.

Enders, W., Parise, G. F., & Sandler, T. (1992). A time-series analysis of transnational terrorism: Trends and cycles. *Defence and Peace Economics, 3*(4), 305-320. doi:10.1080/10430719208404739

Feng, Y., Wang, D., Yin, Y., Li, Z., & Hu, Z. (2020). An XGBoost-Based Casualty Prediction Method for Terrorist Attacks. *Complex & Intelligent Systems*, 1-20.

Guo, D., Liao, K., & Morgan, M. (2007). Visualizing patterns in a global terrorism incident database. *Environment and Planning B: Planning and Design, 34*(5), 767-784. doi:10.1068/b3305

Hao, M., Jiang, D., Ding, F., Fu, J., & Chen, S. (2019). Simulating Spatio-Temporal Patterns of Terrorism Incidents on the Indochina Peninsula with GIS and the Random Forest Method. *ISPRS International Journal of Geo-Information, 8*(3), 133.

Heiser, W. J. (1985). *Multidimensional Scaling by Optimizing Goodness of Fit to a Smooth Hypothesis.* Leiden, Netherlands: University of Leiden.

Huamaní, E. L., Alicia, A. M., & Roman-Gonzalez, A. (2020). Machine Learning Techniques to Visualize and Predict Terrorist Attacks Worldwide using the Global Terrorism Database. *International Journal of Advanced Computer Science and Applications, 11*(4). doi:10.14569/IJACSA.2020.0110474

Hung, B., Jayasumana, A., & Bandara, V. (2018). INSiGHT: A system to detect violent extremist radicalization trajectories in dynamic graphs. *Data & Knowledge Engineering*, *118*, pp. 52-70. doi:10.1016/J.DATAK.2018.09.003

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R* (Vol. 112). New York: Springer. doi:10.1007/978-1-4614-7138-7

Krasmann, S., & Hentschel, C. (2019). 'Situational awareness': Rethinking security in times of urban terrorism. *Security Dialogue, 50*(2), 181-197. doi:10.1177/0967010618819598

LaFree, G. (2010). The Global Terrorism Database (GTD) Accomplishments and Challenges. *Perspectives on Terrorism, 4*(1), 24-46. Retrieved from http://www.jstor.org/stable/26298434

Loia, V., & Orciuoli, F. (2019). Understanding the Composition and Evolution of Terrorist Group Networks: A Rough Set Approach. *Future Generation Computer Systems, 101*, 983-992. doi:10.1016/j.future.2019.07.049

Lu, P., Zhang, Z., Li, M., Chen, D., & Yang, H. (2020). Agent-Based Modeling and Simulations of Terrorist Attacks Combined with Stampedes. *Knowledge-Based Systems, 205*, 106291. doi:10.1016/j.knosys.2020.106291

Maaten, L. v., & Hinton, G. (2008). Visualizing Data Using t-SNE. *Journal of Machine Learning Research, 9*(Nov 2008), 2579-2605.

Madeira, S. C., & Oliveira, A. L. (2004). Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1*(1), 24-45. doi:10.1109/TCBB.2004.2

Martin, I. W. (2008). *The Permanent Tax Revolt: How the Property Tax Transformed American Politics.* Stanford: Stanford University Press.

Mashechkin, I. V., Petrovskiy, M. I., Tsarev, D. V., & Chikunov, M. N. (2019). Machine Learning Methods for Detecting and Monitoring Extremist Information on the Internet. *Programming and Computer Software, 45*(3), 99-115. doi:10.1134/S0361768819030058

Miller, E. (2020). *Global Terrorism Overview: Terrorism in 2019.* College Park, Maryland: University of Maryland. Retrieved from https://www.start.umd.edu/pubs/START_GTD_GlobalTerrorismOverview2019_July2020.pdf

Mishra, N., Swagatika, S., & Singh, D. (2020). An Intelligent Framework for Analysing Terrorism Actions Using Cloud. In P. S., I. A., T. M., & J. V. (Eds.), *New Paradigm in Decision Science and Management. Advances in Intelligent Systems and Computing* (Vol. 1005, pp. 225-235). Singapore: Springer. doi:10.1007/978-981-13-9330-3_21

Naouali, S., Salem, S. B., & Chtourou, Z. (2020). Uncertainty Mode Selection in Categorical Clustering using the Rough Set Theory. *Expert Systems with Applications*, 113555. doi:10.1016/j.eswa.2020.113555

Nizamani, S., & Memon, N. (2012). Detecting Terrorism Incidence Type from News Summary. In T. K. (Ed.), *Advanced Information Technology in Education. Advances in Intelligent and Soft Computing* (Vol. 126, pp. 95-102). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-25908-1_14

Opiyo, O. P., Mukisa, M. T., & Ratemo, M. C. (2019). An Evaluation of Hybrid Machine Learning Classifier Models for Identification of Terrorist Groups in the aftermath of an Attack. *International Research Journal of Engineering and Technology, 6*(9), 1856-1864.

Pruyt, E., & Kwakkel, J. (2014). Radicalization under deep uncertainty: A multi-model exploration of activism, extremism, and terrorism. *System Dynamics Review, 30*(1-2). doi:10.1002/sdr.1510

Python, A., Bender, A., Nandi, A., Hancock, P., Arambepola, R., Brandsch, J., & Lucas, T. (2021). Predicting non-state terrorism worldwide. *Science Advances, 7*(31). doi:10.1126/sciadv.abg4778

Salem, S. B., & Naouali, S. (2016). Pattern Recognition Approach in Multidimensional Databases: Application to the Global Terrorism Database. *International Journal of Advanced Computer Science and Applications (IJACSA), 7*(8). doi:10.14569/IJACSA.2016.070838

Schaller, R. R. (1997). Moore's Law: Past, Present and Future. *IEEE Spectrum, 34*(6), 52-59. doi:10.1109/6.591665

Strang, K., & Sun, Z. (2017). Analyzing Relationships in Terrorism Big Data Using Hadoop and Statistics. *Journal of Computer Information Systems, 57*(1), 67-75. doi:10.1080/08874417.2016.1181497

Sun, A., Naing, M., Lim, E., & Lam, W. (2003). Using support vector machines for terrorism information extraction. In C. H., M. R., Z. D.D., D. C., S. J., & M. T. (Eds.), *Intelligence and Security Informatics. ISI 2003. Lecture Notes in Computer Science* (Vol. 2665). Berlin: Springer. doi:10.1007/3-540-44853-5_1

Tolan, G. M., & Soliman, O. S. (2015). An Experimental Study of Classification Algorithms for Terrorism Prediction. *International Journal of Knowledge Engineering-IACSIT, 1*(2), 107-112.

Uddin, M. I., Zada, N., Aziz, F., Saeed, Y., Zeb, A., Shah, S. A., . . . Mahmoud, M. (2020). Prediction of Future Terrorist Activities Using Deep Neural Networks. *Complexity*. doi:10.1155/2020/1373087

USCB. (2019). *TIGER/Line Shapefiles Technical Documentation.* Washington, D.C.: United States Census Bureau (USCB).

Venna, J., & Kaski, S. (2001). Neighborhood Preservation in Nonlinear Projection Methods: An Experimental Study. In D. G., B. H., & H. K. (Ed.), *International Conference on Artificial Neural Networks. 2130*, pp. 485-491. Berlin, Heidelberg: Springer. doi:10.1007/3-540-44668-0_68

Wall, C. (2021). The (Non) Deus-Ex Machina: A Realistic Assessment of Machine Learning for Countering Domestic Terrorism. *Studies in Conflict and Terrorism*. doi:10.1080/1057610X.2021.1987656