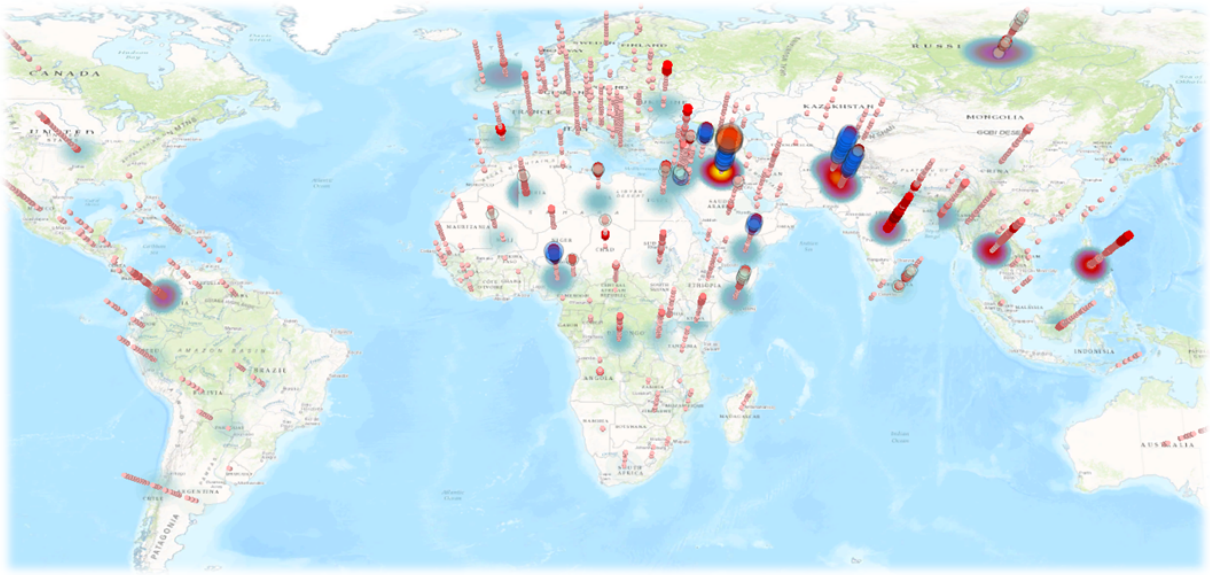
Geo-Spatial Visual Analytics of Global Terrorism

Saman Sadeghi Afgeh and Shakeel Raja



A Space-time cube drawn from data in Global Terrorism Database, highlighting some important variables covering all recorded events from 2000 to 2015. Each block in space-time shows the number of terrorism victims in each country for a year based on the contrast and graduation in the size of block. The heat-map raster on the ground corresponds to the total number of attacks in the country. Numbers of suicide attacks are shown as outlines to space-time blocks with intensity of color and graduating size reflecting an increased number of suicide attacks. (Generated by ESRI ArcGIS Pro)

Fig. 1: A Space-Time cube generated from Global Terrorism Database

Index Terms—GTD, Terrorism, Visual Analytics

1 Motivation, Data and Research Question

International and domestic terrorism is ubiquitous in today's world, and preventing terrorist attacks has become a priority for governments and law enforcement agencies around the globe [1]. The many studies that try to identify the causes of terrorism reach somewhat conflicting results due to the complex nature of the phenomenon and due to differences in perspectives, methods and data used for analysis. This project aims to add to literature by using data from the Global Terrorism Database (GTD) [2], the most comprehensive open data set of terrorism attacks, along with other geopolitical and economic indicators to identify the social and economic causes of terrorism. We began this exercise with the assumption that terrorism is a highly complex phenomenon [3], that a variety of different types of terror activities exist [4], and that a single model was unlikely to correctly identify the social and economic determinants of terrorism. Our assumption is based on the fact that different patterns of terrorism can be caused by different factors, and that the same idiosyncratic shocks could have different effects on the occurrence of terror attacks, depending on the prominent type of terrorism that was present in the specific country [5]. We aimed to identify the different social and economic determinants of terrorism, starting from the assumption that terror events occur in many forms with different characteristics and any predictive analysis exercise was likely to require different explanatory models.

To do so, we planned clustering and regression experiments. The GTD includes 137 variables and 156772 observations of all terrorist attacks happened between 1970 and 2015. Social, political and economic data were drawn from different datasets including political regime characteristics from the Polity IV dataset [6], data on attempted or successful coups from the CSP COUPS dataset [7] and data on political violence from the Major Episodes of Political Violence dataset [8]. Finally, we used economic data and population numbers from World Bank datasets [9].

Our research question were the following:

- Can we visually inspect the internal structure of data to measure prediction optimality?
- Can we visually identify groups of countries that exhibit similar characteristics based on event-specific data in order to study the particular factors that may govern the level of terrorism?
- Can we determine the social, political and economic determinants of terrorism for different groups of countries that may exhibit similar patterns of terrorist activity?

We used geo-spatial visual analytics along with standard statistical measures to study the data, identify possible groupings on countries, study the behavior of different countries in each group and finally to assess the impact of political and economic indicators on the frequency of terrorist attacks in the period 2000 to 2015. Some intermediate questions we faced dealt with evaluating the suitability of the GTD for clustering and with the optimality of building separate models for

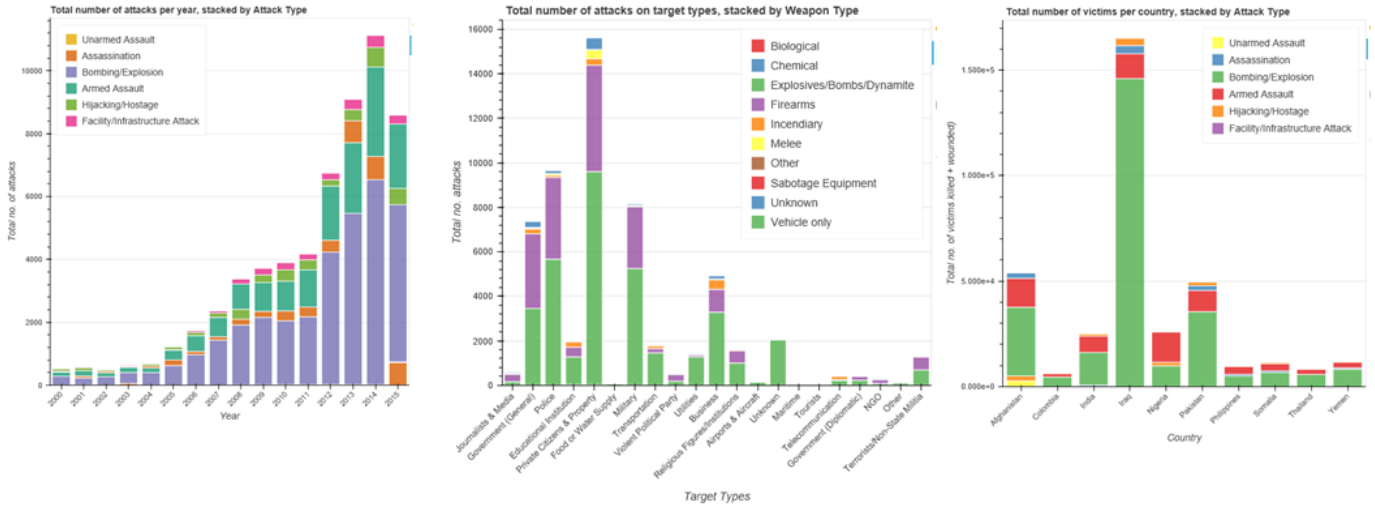


Fig. 2: Stacked bar plots showing mutual relationships of GTD variables

these different clusters. We addressed these questions by iteratively analyzing the data and refining our model according to our visual and statistical findings.

ArcGIS was selected for geo-visualizations due to its thorough documentation, a large online community, the stability of the system, the ability to deal with large data sets and the range of geographic statistics computational methods it provides.

The large scope of the project made it more suitable to be tried as a collaborative effort. Not only the review of the relevant literature on the social and economic determinants of terrorism required a vast effort, but integrating, wrangling and merging the various data sets and learning the necessary tools to perform the computational and visual analytics needed was indeed a large effort that couldn't have been completed in the given time by a single person. Furthermore, building the model, evaluating variables, preparing the visualisations and iteratively applying visual analytics methods to the data proved to be highly time consuming operations.

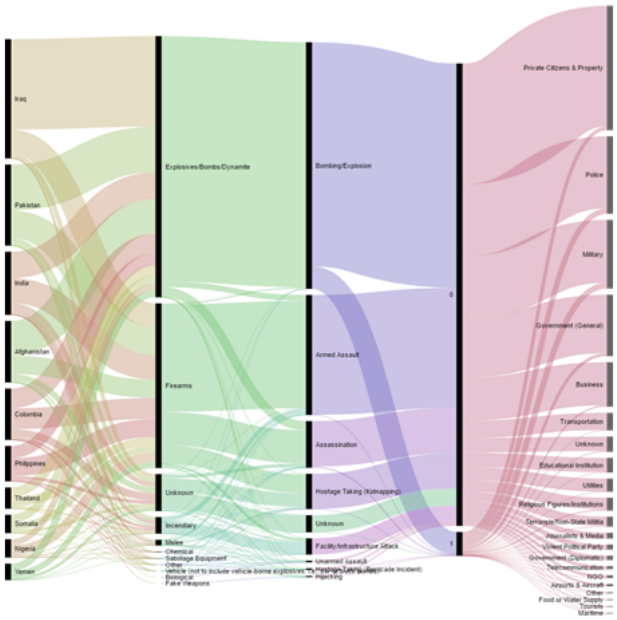
2 Tasks and Approach

2.1 Data pre-processing

We chose to limit our analysis to the years 2000 to 2015 mainly to provide control for changes in definitions of terrorism pre and post war on terror as highlighted in [28] and to cater for unavailability of socioeconomic and political data for more distant years. Most of GTD attributes had a large number of Null values and were excluded from the analysis, together with variables containing qualitative data as text summaries and local news items. Our reduced version of the GTD dataset contains variables including, for each terror event, geo-location (*longitude* and *latitude*), number of people killed and injured (*nkill* and *nwound*, integers combined together as 'victims' to give a measure of damage to human life), type of terror attack (*attack_type*, categorical), type of weapon used (*weap_type*, categorical), type of attack target (*target_type*, categorical) and a dummy identifying whether the attack was a suicide one (*suicide*, binary).

2.2 Visual inspection and evaluation of GTD structure

To inspect selected attributes of GTD and to visualise their interrelation, a number of graphs were generated. Due to the categorical nature of GTD attributes, we decided to use stacked bar graphs (Figure 2) and parallel co-ordinate diagrams (Figure 3). Initial visualizations helped identify the structure of the GTD dataset and the distribution of variable classes. In order to clearly visualize this relationship, we selected the top 10 countries that had been affected the most by terrorism for the given time period.



A parallel co-ordinate plot drawn from GTD for main categorical variables including attack, target and weapons type along with a count for suicide attacks. The diagram highlights the unbalanced nature of Global Terrorism Database with some classes in each variable having high frequencies than others.

Fig. 3: GTD attributes as parallel co ordinates

In order to spatially place the event data, we imported selected GTD attributes into the ArcGIS environment and spatially joined these with a polygon layer containing information on countries and their geographical boundaries. The result (Figure 4) showed an aggregated effect of terror attacks globally highlighting the countries that suffer more with a visual indication of number of victims. To further understand the temporal aspect of these attacks, point data was aggregated for every year to draw a space-time cube (Figure 1), highlighting a density based raster layer showing the number of attacks, the size and contrast of 3d blocks showing the number of victims and an outline to each block showing the number of suicide attacks from 2000 to 2015, in order

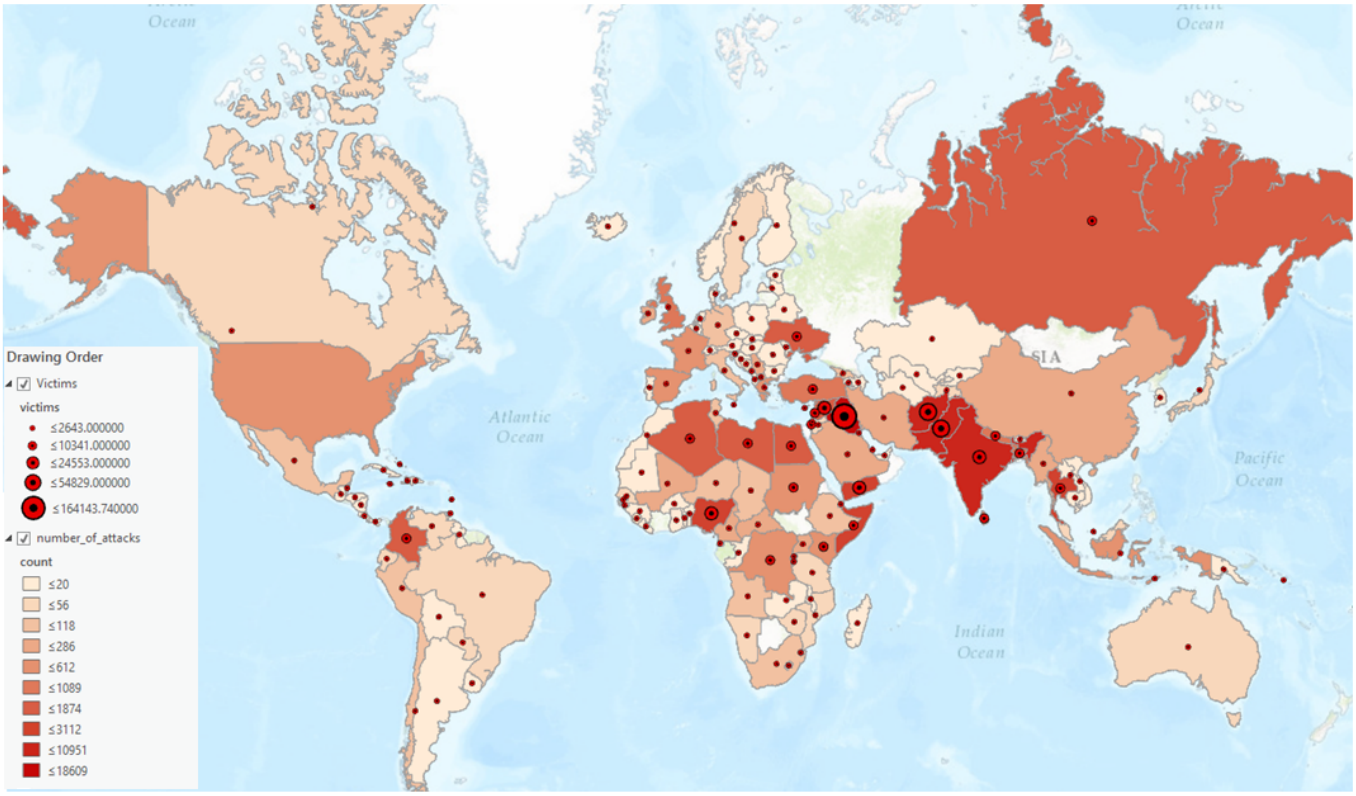


Fig. 4: Choropleth map showing GTD spatial data with attack count (contrast), and victims (graduating symbols)

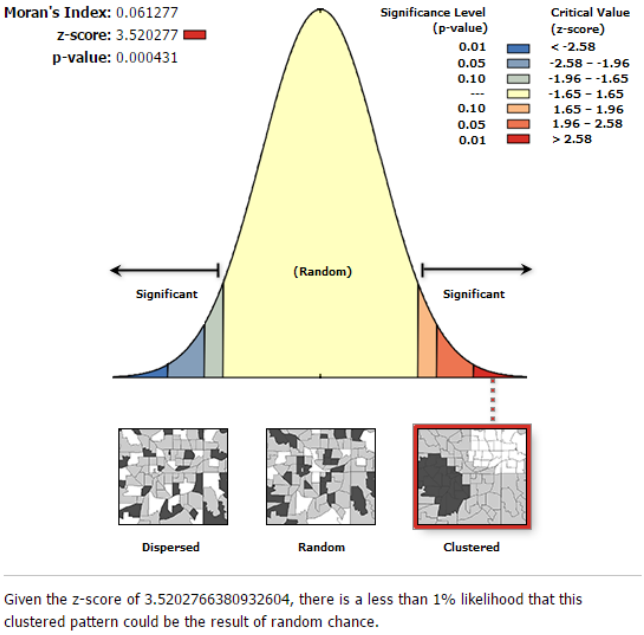


Fig. 5: A sample output from Global Moran's Index for attack count

to evaluate individually the temporal affect of terrorism in different countries.

2.3 Visually Identifying global spatial relationships between groups of countries based on event data

We identified the spatial relationship between affected countries using Moran's I spatial-auto correlation tests [12], using spatially joined attributes of the GTD dataset for each country . We calculated the correlation index for the number of attacks (example as shown in Figure 5), to measures the degree of spatial dependency among terrorism events in the global space, which was used to support clustering analysis. Other selected variables highlighting dominant attack, target and weapon types were also tested for spatial auto-correlation. These results were used to perform a global clustering exercise in ArcGIS's grouping analysis feature using uses k-means clustering algorithm to group input features into clusters based on defined attributes and elbow method to identify optimum number of clusters as shown in [29]. Contiguity was not enforced and no spatial constraints were applied to the clustering process at this stage. Clustering was performed based on attributes individually and in combined form. Choropleth based visualizations were obtained at each stage for visual analysis as shown in Figure 6.

2.4 Using local statistics to visually identify hot spots of terrorism activities

A Cluster and Outlier Analysis (Anselin Local Moran I) [24] test was performed in ArcGIS on country-wise aggregated data using attack count as the weighted feature to discover Local Indicators of Spatial Correlation (LISA) in terms of attack frequency. Anselin Local Moran I test was used to identify where high or low values of attack count clustered spatially, and features with values that were very similar or different from surrounding feature values. Results of this test are given in Figure 7.

To further explore the existence of hotspots with a temporal perspective, to measure the impact of cross-border terrorism for countries with a porous border (e.g. Pakistan and Afghanistan, Iraq and Syria) and to strengthen our decision for later regression with a view on factors like cross-border terrorism that go beyond geographical boundaries, point event data was used create a point based spacetime NetCDF cube

(a) Clustering with attack type attribute

(b) Clustering with weapon type attribute

(c) Clustering with target type attribute

(d) Clustering with all variables

Fig. 6: Global Clustering analysis of countries based on provided GTD attributes

in ArcGIS. An emerging hotspot analysis was performed in ArcGIS to identify temporal and spatial relationships of hot and cold spots of terror attacks. This analysis performed Getis-Ord Gi* statistic (Hot Spot Analysis) [25] and a Mann-Kendall trend test [26] each point in the space time cube to calculate z-score, p-value, and resulting hot spot bin classification. The results are shown in Figure 8.

2.5 Regression modeling of cluster heterogeneity

Expanding on the cluster analysis in the first part of our project, we built a model of the impact of social, political and economic factors on the number of attacks experienced by a country in a given year. In particular, we used our findings in the first part of our project to evaluate whether the different clusters we identified correspond to different patterns of terrorism activity, influenced by different social and economic factors.

To answer our research question, we first created a suitable model for the number of terrorist attacks conditional on a set of covariates, and we then fit it to the separate clusters we identified, to evaluate similarities and differences in the way terrorism activity is influenced by institutional arrangements, economic conditions and political shocks.

3 Analytical Steps

3.1 Exploratory analysis

Investigative visual analysis on GTD attributes' interlinking was visualised using stacked bar graphs Figure 2 along with a parallel plot in Figure 3. These plots helped identify the large class imbalance in the distribution of attribute classes in the GTD. This observation, supported by the literature identifying the GTD as 'non-granular' and 'poorly measured' [2], led us to consider the GTD as unsuitable for predictive modeling analysis. Lack of insightful variables made using data from GTD for our predictive modeling challenging. Furthermore, the point data contained within the GTD dataset is event specific and does not carry any explanatory variables. Clustering experiments were, however, performed to group similar events with this data. This helped us answer our analytical question.

The choropleth based geo-visualisation in Figure 4 shows total number of attacks as intensity of color, and number of victims as a graduated symbol for each country. It appears that Iraq suffered most in terms of number of attacks and victims. This was also confirmed by Figure 1. This gave us an initial insight into classification of countries with respect to damage they have faced as a result of terror attacks, an observation that was further confirmed by the clustering experiment as given in next section.

3.2 Geo-spatial cluster analysis for terrorist activity pattern identification

3.2.1 Global statistics

Clustering was applied to the selected features of the GTD dataset to identify and group countries that may exhibit similar characteristics globally and in close proximity. Moran's I tests confirmed the existence of high values spatial auto-correlation between the chosen GTD variables as shown in Fig. 5, displaying results of applying this test on the *count* variable, which provided a Moran's Index of 0.061277 indicating a high likelihood of clustering based on significance scores. Similar tests were performed at all the chosen variables which yielded similar results for correlation index. This indicated the presence of similar attack count values in close proximity to each other and hence provided a high potential for further clustering.

A number of repetitive clustering exercises (selected variables shown in Fig. 6) were performed at all chosen variables. Different values of K for K-means clustering were tried and tested and it was decided to continue with optimal number of clusters suggested by ArcGIS. Clustering on Attack type in Fig. 6(a) revealed the existence of two optimal clusters, one including Pakistan, Afghanistan, India and Iraq,

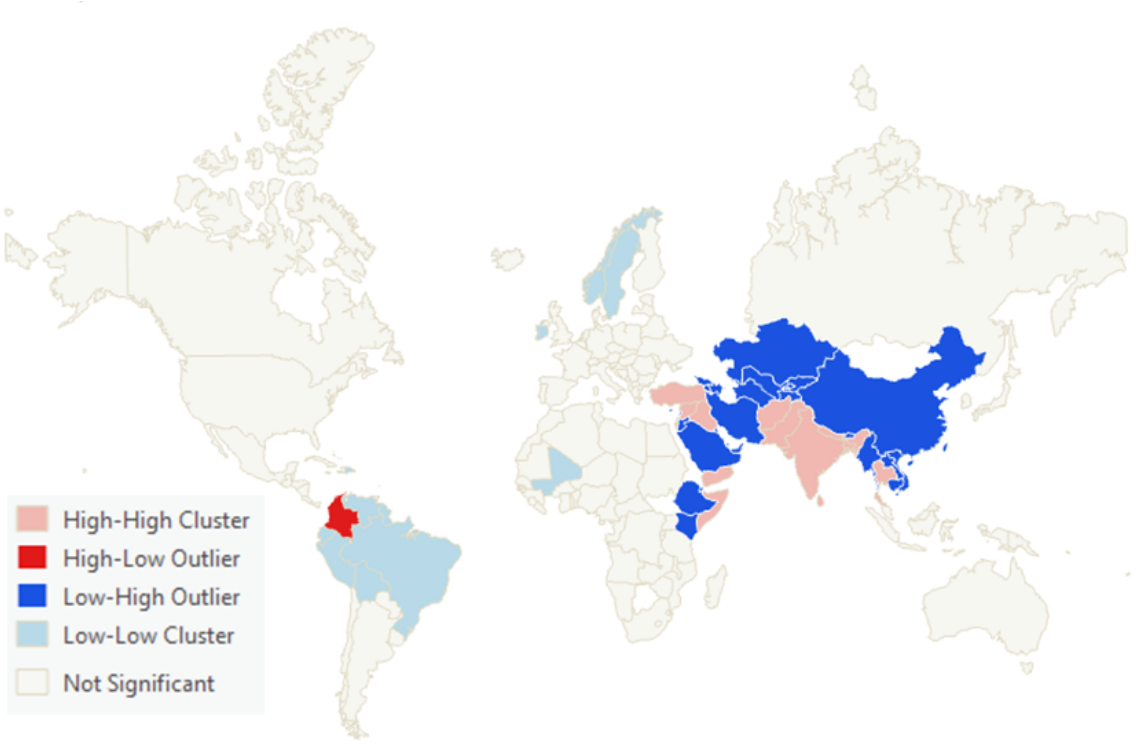


Fig. 7: Local Clustering Anselin Moran's Index

and the other with the rest of the world. We interpreted this as a reflection of heightened terrorism activities by Taliban in Afghanistan, Pakistan and India and ISIS in Iraq during the recent past. Clustering on weapon type in Fig. 6 (b) and target type in Fig. 6(c) showed three clusters. Our final clustering exercise using all event attributes appeared very similar to the first as shown in 6(d).

Having observed the 4 named countries in close proximity repeatedly being clustered, we decided to run further clustering exercise accounting for spatial relationships among geographical neighbors to study the local behavior of terror incidents.

3.2.2 Local statistics

Cluster and Outlier Analysis (Anselin Local Moran I) test on country-wise aggregated data (as shown in Figure 7), revealed few countries as high-high clusters (shown as light red in Figure 7). These included our previous cluster of 4 countries along with Yemen, Somalia, Thailand and Syria, and was reflection of the fact these countries had high attack counts having close neighbors of high attack counts. Other countries geographically close to this cluster were identified as low-high outliers (seen in dark blue, Figure 7), i.e. having low attack count with neighbors of high attack count. This was an initial indication of the fact that applying a unified regression model without spatial considerations may lead to poor predictability. Similarly a low-low cluster covering parts of Europe, south America and Africa was also seen, which was considered a normal behavior. Columbia was identified as a high-low outlier with high terrorism event count with respect to neighbors.

The results of emerging hotspot trend analysis, as shown in Figure 8, revealed a large number of points in Iraq, Pakistan, Afghanistan and some points in India highlighted as 'oscillating hotspots'. These were characterized as statistically significant hot spots of terrorism events following the time-step interval that also had a history of being a statistically significant cold spot during a prior time step. The pattern of these oscillating hotspots appears next the border areas of these countries in much higher densities than central regions. The porous borders between India, Pakistan and Afghanistan and also Syria and Iraq had a high concentration of ongoing terror events that break up periodically. Some 'new hotspots' and 'consecutive hotspots' were also present around oscillating hotspots indicating the spread of terrorism from borders to central regions of the countries. New hotspots indicated areas which may have been free from terror events in the past but recently started having a high count. Consecutive hotspots are indicators of points which have been statistically significant over time with repeated high counts of attacks.

This observation confirmed that using a global regression model to find socio-economic and political indicators affecting the number of terror events in these countries might not prove to be very fruitful as the terror attacks in these countries were not limited by geographical boundaries governing the values of those indicators. i.e. India may have high values of economic and political indicators when compared to Afghanistan but the terror events in these countries clearly follow spatial patterns. This fact must be taken into account when trying to identify the socioeconomic and political impact of terrorism. Similar observations can be associated with Iraq, Syria and Turkey neighborhood.

3.3 Model variables

Having identified clusters in patterns of terrorist activity, we proceeded to fit a model to the complete dataset and to each individual cluster, and we evaluated whether there are significant differences in the determinants of terrorism for different clusters. Our starting hypothesis was that different patterns of terrorist activity would be caused by different social, political and economic factors, and would react differently to political and economic shocks.

We based the choice of variables on three criteria: first, we inspected scatterplots and boxplot charts of the variable Number of Attacks (num_attacks) with each potential regressor, and excluded variables that showed no sign of correlation. Second, we selected variables based on their use in the literature and on their theoretical suitability for our research question. Finally, we inspected the Variance Inflation Factor for the chosen variables, and progressively eliminated the variables that showed strong signs of collinearity. This process made our model robust to issues of collinearity, as can be seen from the VIF scores for the chosen variables, repoted in Table 2.

The variables we included in our model fall into three broad categories. The first category included variables that relate to idiosyncratic

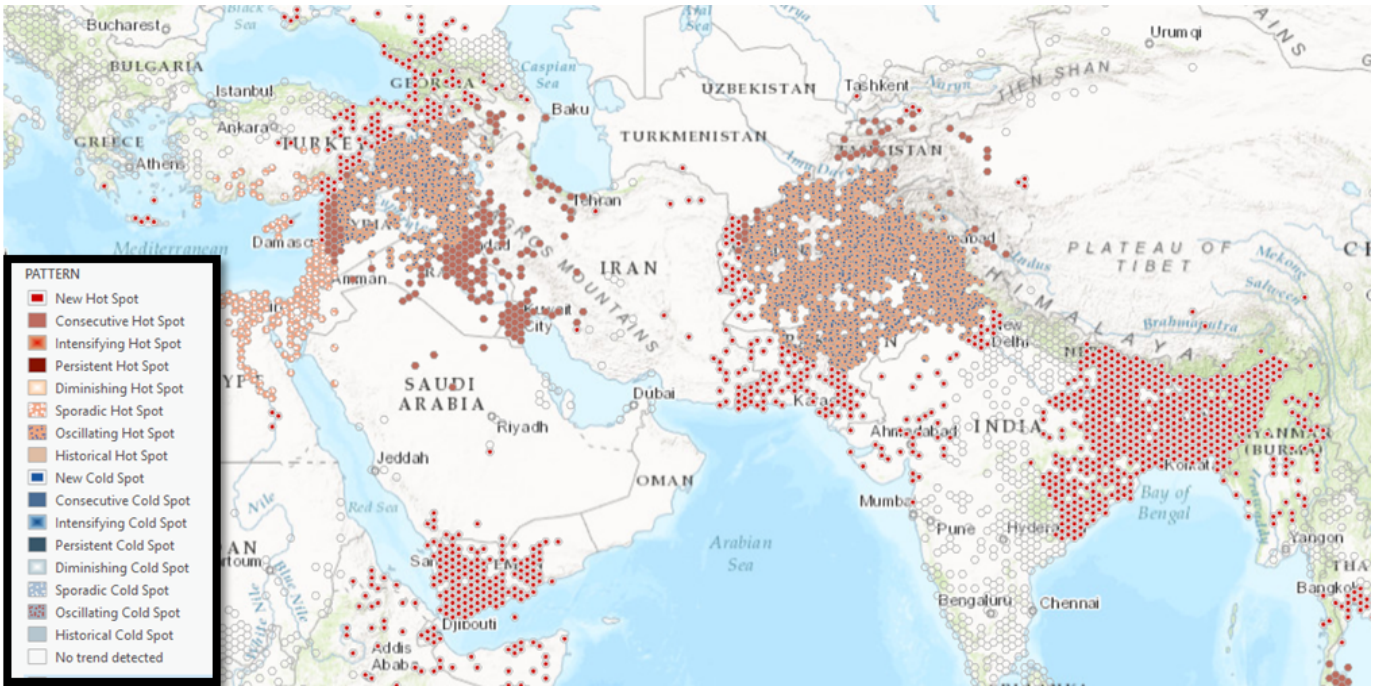


Fig. 8: Emerging hotspot analysis using spatio-temporal point event data from GTD with Getis-Ord Gi* and a Mann-Kendall trend test

political events, and were drawn from the CSP Coups dataset [7] and from the MEPV dataset [8]. A second category of variables contains country-level economic variables. In particular, GDP per capita is considered to be a good indicator of a countrys level of development and wealth, and it is widely used as a regressor in the literature on the determinants of terrorism (for example, in [14], [15] and in [17]). The third category includes institutional controls, that is, variables that control for the specific institutional and political arrangements of the country. A full description for all the variables included in the model is reported in Table 1.

3.4 Choice of model

From the GTD data we built a panel dataset with a row for each country-year. The resulting panel data set was unbalanced, as it had missing observations for some countries and years. Our dependent variable is the *number of attacks* recorded for each country and each year.

Linear regression is not suitable for count data (the estimates are inconsistent, inefficient and biased [21]) but it does provide an useful baseline model. The diagnostic plots for a regression on the complete dataset, presented in Figure 10, show strong evidence of deviation from normality and of fat tails in the distribution of the residuals: while residuals show no strong sign of nonlinearity and no observation has a very high leverage, the QQ plot shows that the distribution of the residuals is not normal, has a fat right tail and variance in excess of what would be expected from a normal distribution. The distribution of the number of attacks in the complete dataset (Figure 11) indicates the presence of fat tails and of non-normality. The histogram also shows the high count of zeros and the large overdispersion of the variable, displaying a variance that is much larger than mean. The negative binomial regression is the appropriate model for overdispersed count data. The model assumes a negative binomial distribution for the data, generalizing the Poisson regression model by relaxing the assumption of equality between mean and variance [21]. The fixed-effects negative binomial regression is widely used in the literature to evaluate the impact of social and economic variables on the number of terrorist attacks in a given country and year (for example, in [18], [16] and [19]). The coefficients for the negative binomial regression model, estimated on the complete dataset and on each cluster, are reported in Table 3.

4 Findings

Initial analysis of selected GTD data set revealed the presence of highly imbalanced data which could lead to poor predictability. This answered our first research question by establishing the unsuitability of the GTD dataset alone for prediction tasks.

Choropleth based visualizations revealed the obvious presence of clusters, and this observation was also in line with Global Moran's I spatial auto-correlation results. Clustering on different variables revealed somewhat similar patterns (as shown in Figure 6), highlighting Iraq, Afghanistan, Pakistan and India as countries showing similarity based on global statistics. The same countries also got identified as a hotspot of terrorism through local spatial statistics (Figure 7). This heterogeneity in patterns of terrorist activity for different countries highlighted the need for a modified approach for our regression modeling experiments.

We split our dataset based on the clustering results and proceeded to fit our model to the data. The two clusters identified were of very different sizes: cluster 2 includes 64 observations and four countries, namely Iraq, Afghanistan, India and Pakistan, while cluster 1 includes all the remaining countries in the dataset and a total of 2411 observations. In order to evaluate whether terrorist activity in these different clusters are driven by different causes, we fit a model to the complete dataset and to each cluster individually.

The negative binomial regression estimated on the complete dataset and on both clusters is reported in Table 3. The dispersion parameter theta is highly significant in all three models, showing that there is significant overdispersion of the *number of attacks* variable compared to a Poisson distribution. The coefficient estimates are relatively stable across the three models, with a few exceptions.

Figure 12 shows Pearson residuals for the negative binomial model fit on the complete dataset, plotted on a spacetime cube. We can see that while the residuals are low across the map, there appears to be both a spatial and temporal correlation in the residuals. A few countries, like Colombia, Thailand, Pakistan and the UK, have consistently high residuals across years, while we can identify a grouping of countries with high residuals in the Middle East. This spatial correlation is in accord with our choice of clustering, as the region with large spatially-correlated residuals roughly corresponds with our second cluster. However, this spatial and temporal correlation in the residuals can

Table 1: Model variables with description and source dataset

Variable	Type	Description	Source
State Failure	Binary	1 if state is in condition of "complete collapse of central authority" in that year	Polity IV
Succesfull Coups	Integer	"Number of successful coups dtat that occurred in the year of record"	CSP Coup
Attempted Coups	Integer	"Number of attempted (but ultimately unsuccessful) coups dtat that occurred in the year of record"	CSP Coup
Coup Plots	Integer	"Number of (thwarted) coup plots that were reported by government officials during the year of record"	CSP Coup
Auto Coups	Integer	"Auto-Coups: Indicator of the occurrence of subversion of the constitutional order by a ruling (usually elected) executive and the imposition of an autocratic regime during the year of record"	CSP Coup
Rebel Out Exec	Binary	"Ouster of Leadership by Rebel Forces: Indicator of the forced ouster of a ruling executive as a direct result of armed action by rebel forces fighting against forces loyal to the regime"	CSP Coup
Assassination Exec	Binary	"Assassination of Executive: Indicator of the assassination of the ruling executive during the year of record"	CSP Coup
Resignation Exec	Binary	"Resignation of Executive Due to Poor Performance and/or Loss of Authority"	CSP Coup
International Total	Integer	"Total summed magnitudes of all interstate MEPV - INTTOT = INTVIOL + INTWAR, where INTWAR/INTVIOL is: Magnitude score of episode(s) of international warfare/violence involving that state in that year. Scale: 1 (lowest) to 10 (highest) for each MEPV; Magnitude scores for multiple MEPV are summed; 0 denotes no episodes"	MEPV
Civil War	Integer	"Magnitude score of episode(s) of civil warfare involving that state in that year. Scale: 1 (lowest) to 10 (highest) for each MEPV; Magnitude scores for multiple MEPV are summed; 0 denotes no episodes"	MEPV
Ethnic Violence	Integer	"Magnitude score of episode(s) of ethnic violence involving that state in that year. Scale: 1 (lowest) to 10 (highest) for each MEPV; Magnitude scores for multiple MEPV are summed"	MEPV
Ethnic War	Integer	"Magnitude score of episode(s) of ethnic warfare involving that state in that year. Scale: 1 (lowest) to 10 (highest) for each MEPV; Magnitude scores for multiple MEPV are summed"	MEPV
GDP per capita growth	Numeric	GDP per capita growth	WB
GDP per capita growth Lagged	Numeric	GDP per capita growth lagged one period (one year)	From WB data
Log GDP	Numeric	Log of GDP per capita (2011 US$)	From WB data
Unemployment Rate	Numeric	Unemployment rate (all adult population)	WB
Number Borders	Integer	"Number of neighboring states sharing a border with the identified state"	MEPV
Durability	Integer	"The number of years since the most recent regime change (defined by a three-point change in the POLITY score over a period of three years or less) or the end of transition period defined by the lack of stable political institutions (denoted by a standardized authority score)"	Polity IV
Combpol Score	Integer	Equivalent of Polity score: "Combined Polity Score: The POLITY score is computed by subtracting the AUTOC score from the DEMOC score; the resulting unified polity scale ranges from +10 (strongly democratic) to -10 (strongly autocratic)". Polity scores of -66, -77 and -88 are set equal to 0.	From Polity IV
Foreign Intervention	Binary	Dummy equal to 1 if the country is under foreign intervention/occupation (Polity score -66)	From Polity IV
Transit	Binary	Dummy equal to 1 if country is undergoing a political transition (Polity score -77 or -88)	From Polity IV
Autocratic Democracy	Binary	Dummy equal to 1 if Combpol Score is equal to 0 and the ountry is not under Foreign Intervention or Transit	From Polity IV
Log Population	Numeric	Log of population size	From WB
Year	Factor	A series of dummy variables for each year between 2000 and 2015	GTD

potentially have biased our analysis, and further investigation would be needed to evaluate whether that is the case.

A tentative interpretation of our results can be made on the basis of our clustering results and on the difference in coefficient estimates. According to our findings, terrorist activity in the period 2000 to 2015 fits into two different patterns.

A first cluster, including the majority of countries, present a pattern of terrorist activity driven primarily by economic factors and by government stability. Countries in this cluster record a higher number of terrorist attacks associated with a negative growth rate of GDP per capita and with a higher unemployment rate. For these countries, the count of terrorist attacks decreases with the stability of a political arrangement (as measured by the *durable* variable), and are increased by a country's involvement in a foreign conflict, by the presence of a civil or ethnic war and by the occurrence of episodes of ethnic violence. Finally, a rebel group ousting an executive leader results in a large decrease in the number of terrorist attacks, suggesting that replacing an executive government is a big motivation behind terrorist attacks in these countries.

A second pattern of terrorist activity includes a few countries with extremely high number of attacks, namely those we grouped in cluster 2. Attacks following this pattern are primarily driven by ethnic conflicts and by response to international conflicts. For this group of countries, terrorist activity is not affected by idiosyncratic political shocks or by economic variables. We hypothesize that the peculiar pattern we identify in these countries is due to involvement in a foreign conflict and the subsequent rise in radical Islamic terrorist attacks, coupled with the presence of ethnic conflicts.

5 CRITICAL REFLECTION

5.1 Implications of findings for domain

While it might seem obvious that different patterns of terrorist attacks with different aims and motivations exist, this insight has been surprisingly absent in the literature we reviewed, and no study we know of uses both visual analytics techniques to identify these different patterns and statistical analysis to validate its findings.

Our finding that two different patterns of terrorist attacks, driven by different social, political and economic factors, can be identified can help understand one of the reasons for the widely differing results that are found in the literature. There is no consensus in the literature on the impact of social, economic and political variables on the frequency of terrorist attacks, with studies on the determinants of terrorism reaching often conflicting results (see Appendix A in [18] for an overview). Our study suggests that a potential source of this discrepancies is the heterogeneity in the patterns of terrorist activities, and suggests one way of identifying the different causes of terrorism for different types of terrorism.

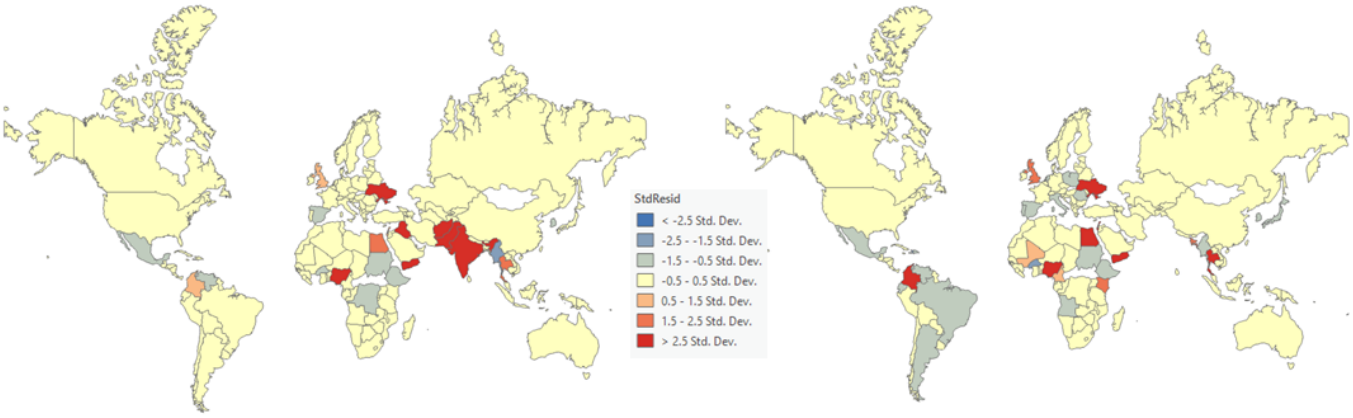


Fig. 9: Residual plots as choropleth maps highlighting the predictive capacity of the regression model

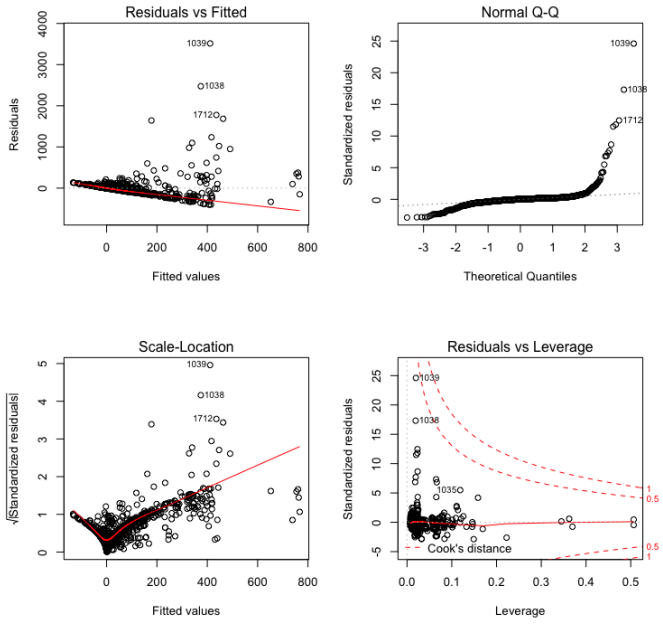


Fig. 10: Diagnostics plots - Linear Regression - Complete dataset

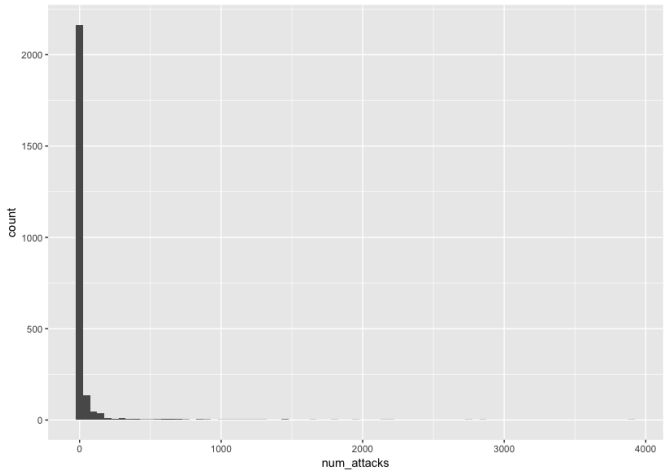


Fig. 11: Histogram of *number of attacks* for complete dataset

Variable	VIF - Whole dataset	VIF - Cluster 1	VIF - Cluster 2
sf	1.754666	1.780472	Nan
scoup1	1.028103	1.028490	Nan
atcoup2	1.077931	1.080128	Nan
pcoup3	1.050742	1.054257	Nan
agcoup	1.012918	1.013540	Nan
reboutex	1.179633	1.198036	Nan
assassex	1.049887	1.051084	Nan
resignex	1.050511	1.051651	Nan
inttot	1.239191	1.102762	39.107777
civwar	1.229763	1.168562	1121.716407
ethviol	1.105080	1.099606	13.977440
ethwar	1.119050	1.058513	57.029128
gdp_pc_gr	1.474227	1.600964	7.005285
gdp_pc_gr_lag1	1.306293	1.418044	2.655094
loggdp	1.955591	1.975374	478.115259
unemp_rate	1.133118	1.121028	110.261952
nborder	1.445965	1.438680	729.837350
durable	1.768213	1.847160	908.376179
combpol	1.198807	1.201489	49.940668
forinterv	1.271474	1.033092	12.594965
transit	1.809269	1.819129	NaN
autocrdemocr	1.042312	1.042576	NaN
logpop	1.595811	1.488658	398.786724
iyear	NA	NA	NA

Table 2: Variance Inflation Factor for model variables

5.2 How well the data and visual analytics approaches enabled answers to research questions

Our first research question about visually analyzing the the GTD to identify its suitability for advanced predictive analytics was answered in a negative way due to obvious class imbalance of variables. No exploratory attributes were found in the GTD that could have been used in regression modeling. This helped us identify the need to look for a number of external variables that could explain the count of terror attacks in specific countries.

The patterns of clusters seen as a result of visual clustering analysis successfully helped us answer our second research question, dealing with the optimality of clustering on GTD data. Two primary clusters helped us differentiate between areas of extreme terror attacks from the rest. Further hotspot and emerging hotspot analysis helped us look deeper into the spatial circumstances of these countries and to discover the possible existence of cross border terrorism in countries which have been traditionally and culturally close together i.e Afghanistan and Pakistan. Unfortunately, no further probing into this phenomenon was possible due to lack of data for responsible terrorist groups, nature of weapons used by these groups and cross-border movement of citizens of these countries. This helped us identify countries that behave similarly to neighboring countries and to visually observe dominant spatial factor in emerging hotspot analysis. This was a positive answer answer to the second second research question.

Our regression findings were inconclusive, and unfortunately our analysis was only partially successful.

First, the evaluation of our model's residuals indicated a possible spatial and temporal correlation. A few countries appear to have con-

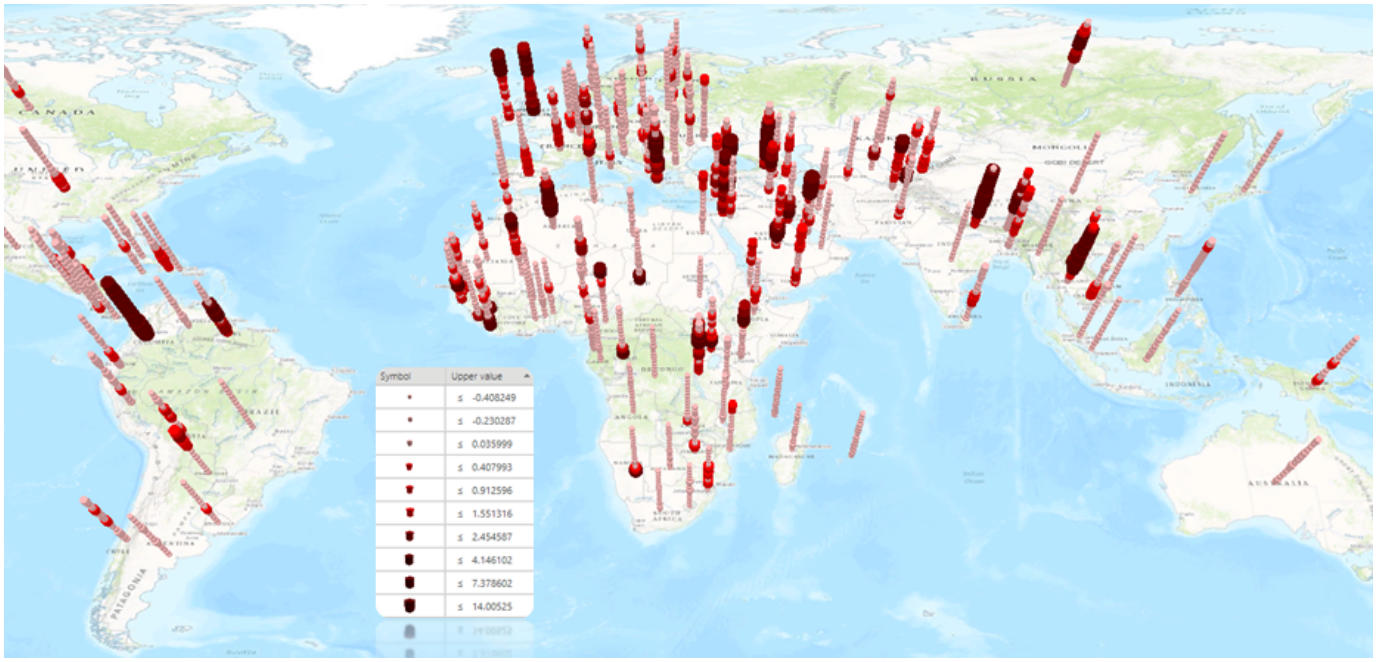


Fig. 12: Spacetime cube representation of Pearson residuals for complete dataset

sistently large residuals, and some countries with large residuals tend to group together. While the spatial correlation in the residuals is consistent with our choice of clustering, the temporal correlation is more troubling and would have to be taken into account in further analysis.

More importantly, the chosen clustering solution produced two unbalanced datasets for the clusters, and introduced severe limitations in our model. First, due to small sample size and the subsequent lack of variability in the dataset coefficients for most variables identifying idiosyncratic political shocks could not be estimated for cluster 2 (thus reducing the sample size for cluster 2 to only 52 observations). Most importantly, the small sample size could have biased our coefficient estimates and caused the variables in our model to be extremely collinear for cluster 2 (Table 2). This can potentially invalidate our results, as collinearity in a linear model inflates the standard errors and leads to unreliable coefficient estimates. Attempts to control for this issue were not successful, as our clustering clearly suggested that the four countries we identified should be included in a separate cluster. A possible solution was to increase the size of the dataset by including terrorist attacks for all year between 1970 and 2015 instead of limiting it to the period 2000-2015. This would have however introduced new issues, as not only it would have increased the heterogeneity in the distribution of attacks, but would have caused problems with data availability.

In our model, were including fixed effects for years but not for country, as were interested in evaluating the impact of institutional variables on frequency on terrorist attacks. Due to collinearity, including fixed effects for country would have prevented us from evaluating the impact of institutions and of political arrangements on the number of terrorist attacks. The decision not to include country fixed effects is supported by the literature in studies using an approach similar to ours [18]. This however implied that we couldn't control for time-invariant country effects, which could carry significant explanatory power. This again could have biased our coefficient estimates.

One crucial limitation of our analysis is that we weren't able to separate in our clusters domestic and international terrorist attacks. This is a serious limitation, as it is reasonable to assume that these different types of attacks are driven by different factors and react differently to idiosyncratic political and economic shocks. Unfortunately, we weren't able to separate the attacks based on this feature. Some studies (for example, [18] did that by using the variable *groupname*, which records the name of the group perpetrating the attack, and by marking the attack as international if the nationality of the perpetrators doesn't match the target country). While this is an interesting strategy, time limitations meant that we couldn't manually encode all group names with their nationality, so this line of inquiry had to be dropped. The heterogeneity between international and domestic terrorist attacks could potentially help explain a puzzling finding of our model, namely that countries in cluster 1 tend to experience more terrorist attacks when countries experience economic downturns (negative GDP per capita growth rate), while at the same time richer countries in that cluster (as measured by the log of the level of GDP per capita) also seem to experience more attacks than poorer countries. This could indicate a further clustering that could be operated on whether the terrorist attacks in a country are mainly domestic or international, as richer countries could be more often targets of international terrorist attacks, while countries experiencing economic downturns could have a higher count of domestic attacks.

There is also a possibility that our model is simply wrong and that the assumptions of the negative binomial regression are not appropriate for our data. In particular, it is possible that the assumption of independence fails, more specifically that process determining first attack not the same as the one determining the next attacks. In this case, different models are needed, such as Hurdle models [27]. Furthermore, due to the nature of our data, we have a very high count of countries with 0 attacks in a given year, and thus more than half of our observations have a *number of attacks* count of 0. Our assumption, based on theoretical grounds, was that these countries all follow the same negative binomial model. It is possible, however, that these countries could be divided into a set of 'real 0s', countries that could experience terrorist attack in a given year but do not, and 'inflated 0s', countries that could not experience a terrorist attack in a given year. If that was the case, a zero-inflated model would be more appropriate. A few studies in the literature on the social and economic determinants of terrorism fit a zero-inflated negative regression model, either as the main model [20] or as a robustness check [22] [23]). While the assumptions behind the zero-inflated model did not seem reasonable for our data, future work should include fitting such a model to the data, and evaluating the difference in fit between that and the standard negative binomial regression model.

Table 3: Negative Binomial Regression - Complete dataset and clusters

	Dependent variable:		
	Number of attacks		
	Complete Dataset	Cluster 1	Cluster 2
sf	0.611	0.872	
	(0.601)	(0.614)	
scoup1	0.977*	0.937*	
	(0.545)	(0.553)	
atcoup2	0.377	0.362	
	(0.323)	(0.329)	
pcoup3	0.242	0.217	
	(0.357)	(0.363)	
agcoup	0.166	0.145	
	(1.407)	(1.429)	
reboutex	−6.282***	−6.084***	
	(2.237)	(2.295)	
assassex	0.290	0.468	
	(1.207)	(1.219)	
resignex	0.936*	1.002**	
	(0.481)	(0.487)	
inttot	0.519***	0.379**	0.302***
	(0.108)	(0.172)	(0.093)
civwar	0.720***	0.554***	−0.200
	(0.075)	(0.079)	(0.776)
ethviol	0.658***	0.782***	−0.237*
	(0.091)	(0.097)	(0.141)
ethwar	0.763***	0.826***	0.401***
	(0.051)	(0.063)	(0.109)
gdp_pc_gr	−0.051***	−0.056***	0.0002
	(0.012)	(0.013)	(0.010)
gdp_pc_gr_lag1	−0.019*	−0.017	0.004
	(0.011)	(0.013)	(0.005)
loggdp	0.325***	0.310***	−0.728
	(0.038)	(0.039)	(0.728)
unemp_rate	0.047***	0.047***	0.016
	(0.008)	(0.008)	(0.057)
nborder	−0.097***	−0.112***	0.618
	(0.021)	(0.021)	(0.852)
durable	−0.016***	−0.015***	−0.054
	(0.002)	(0.002)	(0.034)
combpol	0.004	0.001	0.042
	(0.008)	(0.008)	(0.041)
forinterv	2.468***	1.133	0.157
	(0.468)	(0.842)	(0.215)
transit	−0.690	−0.752	
	(0.461)	(0.470)	
autocrdemocr	−0.694	−0.606	
	(0.537)	(0.539)	
logpop	1.007***	1.028***	0.237
	(0.038)	(0.039)	(0.373)
Constant	−17.042***	−17.175***	1.738
	(0.705)	(0.722)	(8.123)
Year fixed effects	Yes	Yes	Yes
Observations	2,238	2,186	52
Log Likelihood	−5,143.725	−4,715.181	−307.629
θ	0.271*** (0.011)	0.263*** (0.011)	25.451*** (5.253)
Akaike Inf. Crit.	10,363.450	9,506.363	671.258

*p<0.1; **p<0.05; ***p<0.01

5.3 Applications of approach to other domains

Our approach, which consists in using visual analytics techniques to identify meaningful subsamples for statistical analysis, can be used in any domain that can benefit from modeling heterogeneity in the distribution of the dependent variable. By using hotspot analysis and clustering, samples can be split according to relevant features, and different models can be fit to the reduced datasets and compared to identify differences and similarities between the clusters. We believe that domains would benefit from such an approach, from Sociology (modeling the impact of public programs on education in different schools), to Economics (modeling heterogeneity in the response to economic shocks), to Criminology (modeling heterogeneous patterns of crime between neighborhoods).

5.4 Conclusion

Despite all the limitations in our model and the issues we encountered in obtaining significant results, we believe that our approach is innovative and that that with a few necessary adjustments it could lead to interesting and accurate results. While our conclusions are only tentative, we believe that the interplay of visual analytics techniques and statistical analysis is a fruitful approach that can help researchers identify the heterogeneity in patterns of global terrorist activity.

REFERENCES

[1] Lum, C., Kennedy, L. W., and Sherley, A. (2006), *Are counter-terrorism strategies effective? The results of the Campbell systematic review on counter-terrorism evaluation research*, Journal of Experimental Criminology, 2(4), 489-516.

[2] LaFree, G. (2010), *The global terrorism database: Accomplishments and challenges*, Perspectives on Terrorism, 4(1).

[3] Suedfeld, P., and Leighton, D. C. (2002), *Early communications in the war against terrorism: An integrative complexity analysis*, Political Psychology, 23(3), 585-599

[4] Lizardo, O. A., and Bergesen, A. J. (2003), *Types of terrorism by world system location*, Humboldt Journal of Social Relations, 162-192

[5] Sandler, T., and Enders, W. (2008), *Economic consequences of terrorism in developed and developing countries*, Terrorism, economic development, and political openness, 17

[6] Center for Systemic Peace, POLITYTM IV PROJECT, *Political Regime Characteristics and Transitions, 1800-2015, Dataset Users Manual*, Retrieved from `http://www.systemicpeace.org/inscr/p4manualv2015.pdf`

[7] Marshall, M.G. and Ramsey Marshall, D. (2016), *Coup D'Etat Events, 1946-2015 Codebook*, Center for Systemic Peace, Retrieved from `http://www.systemicpeace.org/inscr/CSPCoupsCodebook2015.pdf`

[8] Marshall, M.G. (2016), *Major Episode of Political Violence and Conflict Regions, 1946-2015*, Center for Systemic Peace, Retrieved from `http://www.systemicpeace.org/inscr/MEPVcodebook2015.pdf`

[9] World Bank Open Data portal, Accessed online at: `http://data.worldbank.org/`

[10] Moten, A. R. (2010), *Understanding terrorism: Contested concept, conflicting perspectives and shattering consequences*, Intellectual Discourse, 18(1), 35

[11] Ikenberry, G. J. (2002), *America's imperial ambition*, Foreign Affairs, 44-60

[12] Bivand, R. (2015), *Spatial dependence: weighting schemes, statistics and models*, R-package, Accessed online at: `http://www.icesi.edu.co/CRAN/web/packages/spdep/`

[13] START, (2016), *Codebook: Inclusion Criteria and Variables*, Retrieved from `https://www.start.umd.edu/gtd/downloads/codebook.pdf`

[14] Abadie, A. (2006), *Poverty, political freedom, and the roots of terrorism*, American Economic Review 96, 5056

[15] Blomberg, B., Hess, G.D. (2008), *From (no) Butter to Guns? Understanding the Economic Role in Transnational Terrorism*, In: Keefer, P., Loyaza, N. (Eds.), Terrorism, Economic Development, and Political Openness. Cambridge University Press, Cambridge and New York, pp. 83115

[16] Burgoon, B. (2006), *On welfare and terror: social welfare policies and political-economic roots of terrorism*, Journal of Conflict Resolution 50, 176203

[17] Krueger, A.B., Laitin, D.D., (2008), *Kto Kogo? A Cross-Country Study of the Origins and Targets of Terrorism*. In: Keefer, P., Loyaza, N. (Eds.), Terrorism, Economic Development, and Political Openness. Cambridge University Press, Cambridge and New York, pp. 148173

[18] Kis-Katos, K., Liebert, H. and Schulze, G.G. (2011) *On the origin of domestic and international terrorism*, European Journal of Political Economy, 27, pp.S17-S36

[19] Campos, N. and Gassebner, M. (2009), *International Terrorism, Political Instability and the Escalation Effect*, IZA Discussion Paper, No. 4061, IZA Bonn

[20] Drakos, K., Gofas, A. (2006),*In search of the average transnational terrorist attack venue*, Defence and Peace Economics 17, 7393

[21] Long, J. S. (1997), *Regression models for categorical and limited dependent variables*, Thousand Oaks, CA: Sage

[22] Li, Q. (2005), *Does democracy promote or reduce transnational terrorist incidents?*, Journal of Conflict Resolution 49, 278297

[23] Piazza, J. (2008), *Incubators of terror: do failed and failing states promote transnational terrorism?*, International Studies Quarterly 52, 469488

[24] Anselin, L. (1995), *Local indicators of spatial associationLISA*, Geographical analysis, 27(2), 93-115

[25] Getis, A., and Ord, J. K. (1992), *The analysis of spatial association by use of distance statistics*, Geographical analysis, 24(3), 189-206

[26] McLeandod, A. I. (2005), *Kendall rank correlation and Mann-Kendall trend test*, R Package Kendall

[27] Cameron, A.C. and Trivedi, P.K. (2001),*Essentials of count data regression. A companion to theoretical econometrics*, 331

[28] Crenshaw, M. (2008), . *New vs. Old Terrorism: A Critical Appraisal*, in: *Jihadi Terrorism and the Radicalisation Challenge in Europe*, Burlington, VT:Ashgate

[29] Ming, M. and Chiang, T.,(2009), *Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads*, Journal of Classification 27