A machine learning approach to the analysis of terrorism

Peter Brennan
ITT
Department of Computer Science

M.Sc

October 22,2016

Abstract

Terrorism though having been present since antiquity gained widespread adoption as a perceived method to accomplish a political goal, ideological or religious change in the latter part of the 20th century and into the 21st century. Being able to detect changes in of behaviour, particularly changes in intensity in terrorism is important for economic and societal reasons. Being able to detect changes in intensity is therefore an important task. This thesis investigates a number of methodologies capable of being able to detect changes in behaviour using online open source terrorist incident databases. A number of methodologies are evaluated specifically for modelling count time series data. These include count regression techniques, HMM's, time series aberration detection techniques (RAD and SURUS) and syndromic surveillance methods. The different methodologies are evaluated and the strengths and weaknesses of the different approaches are discussed. Using open source software the demonstration of the application of multiple techniques is illustrated and specifically the power of using multiple techniques in the diagnosis of changes in intensity of terrorism.

Contents

1 What is Terrorism? 2

1.1 The taxonomy of terrorism 2
1.2 Creating the conditions for terrorism? 3
1.3 Why terrorist groups commit terrorist acts? 4
1.4 The strategic aims of terrorism 4
1.5 Classifying the different types of terrorism 5
1.6 Terrorist tactics 7
1.7 Does terrorism work? 9
1.8 Countering terrorism and countering the terrorist narrative . . . 10
1.9 Research Question to be addressed in the study 12
1.10 Thesis overview 14
1.11 Discussion 15
1.12 Conclusion 18

2 Datamining and statistical approaches to research in terrorism and counter terrorism 20

2.1 Terrorism informatics 21
2.2 Data mining approaches for counter terrorism 22
2.3 Terrorism research using data mining and statistical techniques . 23
2.4 The history of terrorism informatics, a German, US and Israeli perspective 26
2.5 Problems associated with data mining and counter terrorism? . . 29
2.5.1 The dangers of mass surveillance and its misuse 30
2.5.2 The problem of class imbalance in counter terrorism datamining 30
2.5.3 The problem of low base rates of terrorism in counter terrorism datamining 31
2.5.4 Low observable signal for terrorism and the use of unsupervised methods 31
2.6 Discussion 32
2.7 Conclusion 34

3 Electronic terrorist incident databases 36
3.1 Introduction to terrorism incident databases 36
3.2 The synthesis of the Global Terrorism Databases, the data collection process 39
3.3 Measurement issues with using incident terrorist databases . . . 41
3.4 The coding of the GTD database 41
3.4.1 GTD admittance metrics 42
3.5 The database variables 43
3.6 The use of the GTD in terrorism research 43
3.7 Discussion 45
3.8 Conclusion 46

4 Count time series aberration techniques 48
4.1 Introduction 48
4.2 Outbreaks, syndormes and anomalies in time series count data . 49
4.3 Algorithms used for modelling time series count data and their methodology 50
4.3.1 Online time series outbreak and anomaly detection algorithm 50
4.3.2 Surveillance algorithms 52
4.3.3 Farrington's Method 52
4.3.4 EarsC3 53
4.4 Experimental 54
4.4.1 Modelling anomalies and outbreaks in Iraqi Death counts due to terrorism 54
4.4.2 The use of syndromic surveillance methods (EARSC3 and Farrington surveillance methods) to analyse Iraqi Death counts due to terrorism 56
4.5 Surveying multiple countries or regions at the same time using purrr 58
4.6 Discussion 60
4.7 Conclusion 65

5 Final discussion and conclusion 66
5.1 Detecting changes in intensity and behaviour in terrorism 66
5.2 The study of statistical and data mining methodologies for the detection of changes in intensity of terrorism 67
5.3 Future work 68

Appendices 70

A A Preliminary understanding of the global terrorism database 71
A.1 Introduction to terrorism incident databases 71
A.2 Data mining techniques used in the preliminary understanding of the GTD database data 73

A.3 Data mining techniques used in the preliminary understanding of the GTD database data 74
A.3.1 The evolution of terrorism over time (year and month) by region and world wide 75
A.3.2 The evolution of terrorism over time (month) by region . 75
A.3.3 The evolution of terrorism over time (year) by region . . . 75
A.3.4 The evolution of terrorism over time (year) by region using stacked bar charts 76
A.3.5 Visualizing deaths by attack vector and weapon type type 77
A.3.6 Visualizing deaths by country and decade type 77
A.3.7 Visualizing deaths by rural urban categorization using K-means and Kernel Density estimation and interactive maps to create heatmaps 78
A.4 The use of dimension reduction techniques and clustering in understanding the GTD 83
A.4.1 Correspondence analysis of weapon type by year 84
A.4.2 Correspondence analysis of attack type by year 85
A.4.3 Multiple Correspondence analysis of terrorist incidents the GTD by year 85
A.5 Preliminary modelling techniques to gain a temporal understanding of terrorism 85
A.6 Poisson, Quasi Poisson, Negative Binomial, Linear and Robust (of log transformed) regression of deaths due to terrorism in Iraq by month 86
A.7 Using Hidden Markov Models to analyse the number of daily death due to terrorism 91
A.8 Discussion 94
A.9 Conclusion 97

B Chapter 4 regression estimates and two way contingency table results 128

C Chapter 5 outbreak and outlier detection detected aberrations 137

D Code repos 141

List of Figures

4.1 Time series plot interval(week) of deaths 54
4.2 Time series plot interval(week) of deaths and application of twitter outbreak detection algorithm. The detected outbreaks are shown indicated as vertical red lines 55
4.3 Time series plot interval(week) of deaths and application of SURUS RAD algorithm. 56
4.4 Time series plot interval(week) of deaths and application of EARSC3 algorithm 57
4.5 Alarm plot for Farringtons and EarsC3 syndromic surveillance methods. 58
4.6 Alarm plot for Farringtons, EarsC3, SURUS and RAD method syndromic and aberration detection methods. 59
4.7 Alarm plot for Farringtons, EarsC3, SURUS and RAD method syndromic and aberration detection methods, overlayed with the time series plot. 60
4.8 Time series plot interval(week) of deaths and application of Farrington algorithm 61
4.9 Applying the SURUS model functionally by mapping it across different countries and extracting the dataframe locations of the outbreak. 61
4.10 Processing the output of functional program to mark the dates across multiple countries the outbreak occurred. 62
4.11 Plotting the output of the SURUS accross multiple countries using ggplot. 62
4.12 Time series plot interval(week) of deaths and application of SURUS algorithm across multiple countries in the mid-east 63

A.1 Time series plot interval (month) of deaths by region 76
A.2 Time series plot interval(month) of deaths 77
A.3 Time series plot interval(year) of deaths 78
A.4 Time series plot interval(year) of deaths 79
A.5 Stacked bar plot interval(year) of deaths 80
A.6 Stacked bar plot interval(year) of deaths by regions 81
A.7 Stacked bar plot interval(year) of deaths by weapon type in the Middle East and North Africa 82

A.8 Stacked bar plot interval(year) of deaths by attack type 83
A.9 A stacked choropleth showing deaths by decade due to terrorism. 100
A.10 Clustering of Terrorist incidents in Syria and Iraq 101
A.11 Clustering of Terrorist incidents in Syria and Iraq, centred on the Sunni triangle 102
A.12 Heatmap of terrorist incidents in Syria and Iraq 102
A.13 Biplot CA of contingency table of deaths by weapons and year. . 103
A.14 Biplot CA of contingency table of deaths by weapons and year, created using Plotly, zoomed in. 104
A.15 Biplot CA of contingency table of deaths by weapons and year, created using Plotly, zoomed in. 104
A.16 Biplot CA of contingency table of deaths by attack type and year, created using Plotly, zoomed in to show associations between explosions and years they occurred. 105
A.17 Biplot CA of contingency table of deaths by attack type and year, created using Plotly, zoomed in to show associations between assassinations and years they occurred. 106
A.18 Biplot CA of contingency table of deaths by attack type and year, created using Plotly, zoomed in to show associations between hostage taking and years they occurred. 107
A.19 Biplot of MCA of incidents by attack type, weapon type, target type, group name and year, created using Plotly, zoomed in to show associations between hostage taking and years they occurred.108
A.20 Timeline of Iraq post and pre invasion (2003 108
A.21 Timeline of Iraq and US and British presidents and major events since the Iraq invasion. 109
A.22 Cook distance diagnostic plot 110
A.23 QQ plot diagnostic plot 111
A.24 Fitted residuals diagnostic plot 112
A.25 Scale location diagnostic plot 113
A.26 Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2003-2004. 114
A.27 Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2004-2005. 115
A.28 Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2005-2006. 116
A.29 Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2006-2007. 117
A.30 Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2007-2008. 118

A.31 Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2008-2009. 119
A.32 Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2009-2010. 120
A.33 Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2010-2011. 121
A.34 Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2011-2012. 122
A.35 Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2012-2013. 123
A.36 Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2013-2014. 124
A.37 Faceted plot of daily death count and epoch classification due to terrorism from Iraq modelled using a HMM, 2006. 125
A.38 Faceted plot of daily death count and epoch classification due to terrorism from Iraq modelled using a HMM, 2007. 125
A.39 Faceted plot of daily death count and epoch classification due to terrorism from Iraq modelled using a HMM, 2008. 126
A.40 Faceted aggregate plot of monthly death count and epoch classification due to terrorism from Iraq modelled using a HMM with the daily death count for the same period, 2008. 127

List of Tables

B.1 Two way contingency table of Deaths by year and Weapon type. 129
B.2 Two way contingency table of Deaths by year and Attack type. . 130
B.3 Poisson Regression estimates 131
B.4 quasi-Poisson Regression estimates 132
B.5 Negative binomial regression estimates 133
B.6 Linear regression model estimates. 134
B.7 Robust regression estimates. 135
B.8 Robust regression robust F-test results. 136

C.1 Identified breakouts of deaths due to terrorism in Iraq post invasion137
C.2 Output of RAD detection algorithm 138
C.3 A table of alarms raised by the EARSC3 algorithm 139
C.4 A table of alarms raised by the Farrington algorithm 140

Chapter 1

What is Terrorism?

1.1 The taxonomy of terrorism

From such high profile incidents as the September 11th attacks on the US eastern coast, to Beslan in Russia in 2004 to the attacks on the World trade centre in 2001, to more recent attacks in San Bernadino and the Nice attacks in 2016, terrorism is now perceived as a major threat to many countries. Terrorism though is not a new phenomenon and has existed since ancient times from the Sicarii (Horsley, 1979) to the Assassins (Sloan and Anderson, 2009). Later terrorism took on more sophisticated methodologies from the gunpowder plot in the 16th century (Fraser, 2010), through to the nationalist and politically motivated groups of the 20th century , on to the religiously motivated groups of the twenty first century.

Terrorism even from the small number of examples cited previously can be characterized as being adaptive and has adopted new technologies, tactics, aims and ideologies. Due to the large scale increase in the adoption of advanced communications the means of disseminating it's message to propagandise recruits, disseminate technology and aid in the exchange of information between members and between allied groups, its prevalence has grown considerably.

Terrorism's commonly accepted definition is the 'utilization and/or the perceived menace of violence whose aim is to further a political agenda'. While there is no universally agreed definition of terrorism (Ruby, 2002), with various governments, institutions, law enforcement agencies and legal entities having different definitions of terrorism. For instance the U.S department of defence defines terrorism as (Pub, 1998): "politically motivated violence perpetrated against non combatant targets by sub-national groups or clandestine agents." While NATO defines terrorism as (Chase, 2013): "The unlawful use or threatened use of force or violence against individuals or property in an attempt to coerce or intimidate governments or societies to achieve political, religious or ideological objectives" The most commonly agreed tenets of terrorism can be summarized as such:

1. It is the utilization of violence in order to coerce a government into a political, religious or ideological shift in policy (Chase, 2013).
2. It is committed by non-state actors (or in the case of state sponsored terrorism) or the indirect or direct usage of a states military or para-military forces.
3. It is aimed at influencing society as a whole besides the immediate victims of an act of terrorism. The immediate victims can be seen as a surrogate for the state, ideology or religion under attack, through the use of media to trumpet their (the terrorists) message,(El-Nawawy, 2014).
4. Another important tenet of terrorism is that when considered in the terms of international law as outlined by the Geneva and Hague conventions, they are considered to be 'mala prohibita' (a crime that is illegal according to legislation) and 'mala in se' (a crime that is morally wrong), (Ganor, 2002).
5. To the terrorist, their acts are considered to be 'Jus ad bellum' the idea of a just war, being either legitimized from a religious, political or constitutional point of view (Kennedy, 1999).

1.2 Creating the conditions for terrorism?

For terrorism to occur, a number of criteria must be met (Crenshaw, 1981). The first of these is that an identifiable grievance must exist amongst a detectable section of the populace. This can often be an ethnic minority who feel discriminated against by the majority as in the case of Irish nationalists during the troubles in Northern Ireland (Moxon-Browne, 1983), Basques in Spain (Llera et al., 1993), or dissident Quebecois (Ross, 1995) who wanted to separate from Canada. However a dissatisfied minority is often not a sufficient condition to establish a terrorist group. Not everyone who is subjected to deprivation and discrimination will turn to terrorism or neither all persons who turn to terrorism will be from a persecuted background.

Numerous cases of people turning to terrorism who were not subject to deprivation exist, examples being the red brigades of Germany, Italy and Japan of the 1970's (Zwerman et al., 2000). Therefore it is not a case of actual deprivation or discrimination but a case of perceived deprivation, with the government to blame for this deprivation or discrimination that is responsible for radicalization.

Context is another important causal factor, especially when it affects an elite and not the population as a whole. The elite constituting a well educated, middle class and young disaffected cohort of the population see the only chance of affecting change through drastic measures such as terrorism (Ronchey, 1979).

A third factor necessary for (the exclusive use of as opposed to the use of terrorism as part of an overall insurgency) terrorism is the intersection of not only elite discontent and dissatisfaction but with indifference amongst the population.

The final causal factor identified by Crenshaw (Crenshaw, 1981) is the idea of a precipatory event involving a demonstrative and impulsive use of force that occurs directly preceding the outbreak of terrorism. Examples of this would be the killing of Beno Ohnesorg at the hands of the West German police in 1968 led to the development of the RAF (Red Army Faction) (Jazić, 2013).

1.3 Why terrorist groups commit terrorist acts?

Terrorists commit terrorist act's as part of wider struggle or insurgency against a state. As such terrorist acts can be used along with guerilla warfare within an insurgency against a government as a tactic. Were terrorism diverges from guerilla warfare is that, were guerilla warfare targets military forces, terrorism targets civilians. Often insurgents will use both and legitimize the use of terrorism by considering it to be a form of politically motivated violence (Ganor, 2002). Terrorists also see terrorism as a mechanism of radicalising the populace. They believe this to either occur through the use of sensational acts of terrorism to alight an insurgent spirit in the people or through over zealous responses of governments to terrorist acts which will cause large discontent in the populace and sympathy for the terrorists (Jenkins, 1985). A proponent of such a tactic has been the FARC group in Colombia who has utilized terrorism as part of a wider insurgency strategy (Wickham-Crowley, 1990),(Marks, 2002).

Terrorists may also aim to impose an economic cost on a country. Acts of terrorism have a direct economic effect by discouraging DFI (Direct Foreign Investment) through redirecting public funds to the security forces or through the impediment of foreign trade (Sandler and Enders, 2008). While these consequences have a local economic effect terrorism also can have global consequences. In the aftermath of the September 11th attacks, saw both a substantial transitory lowering of demand but also a long term negative need or demand (Ito and Lee, 2005) on US airline use.

Other motivating factors for terrorists to carry out terrorist acts include the weakening of key infrastructure to create mistrust in government amongst the populace to provide for its citizens or to put pressure on a government to release prisoners (terrorism research). In doing this, they disrupt the processes of government. Doing so they demoralize officials of the state without impacting directly upon the public.

1.4 The strategic aims of terrorism

Related to types of terrorism are the strategic goals which lie at the aim of a terrorist campaign (Kydd and Walter, 2006), these are:

1. Attrition. This is a strategy employed by a group to persuade an opponent to give up due to the considerable cost of continuing their current policy. It is also based on the premise that the terrorist organization can sustain losses at higher rates than their opponents.

2. Intimidation. Intimidation is a control strategy, where the terrorists aim to persuade the populace that the terrorist group is of ample strength so as to punish a lack of loyalty to the group and also the government is not of sufficient strength to protect them, thereby subjugating the people, to assert control over the people.

3. Provocation. This is the use of terrorist groups to provoke responses with an indiscriminate unmeasured response, thereby instilling a sense of injustice in the populace which leads to a radicalization of the populace and support for the terrorist group.

4. Spoiling. A spoiling strategy is an attempt to dissuade a government from dialogues with groups whose cause may be allied with their own (the terrorist group), by launching spoiling attacks which may sabotage a peace/reconciliation process.

5. Outbidding. This is the process of the use of violence to convince the general population that they posses greater steadfastness and will than their opponents and rivals.

While these terms provide a useful categorization of strategies, it should be remembered that they are not mutually exclusive and groups will often indulge in usage of multiple different strategies to obtain their goals.

1.5 Classifying the different types of terrorism

Beginning in the 1970's researchers began to distinguish between the different types of terrorism based upon the act in order to better understand the subject, due to the perceived threat of terrorism. A clear consensus has developed on the different types, these are:

1. State terrorism. State terrorism can be defined as the use of force (or threat of force) with the aim of coercion and intimidation of a populace by state actors. This is either carried out directly by state actors (police, military) or carried out indirectly by agents acting on behalf of the state, which can be seen as a proxy for the state. A very early example of state terrorism would be revolutionary France in the late 18th century, where a revolutionary dictatorship who seized power after the fall of the monarchy, slaughtered tens of thousands who they thought may oppose the regime. Similarly the strongly authoritarian regimes of Nazi Germany (Gibbs, 1989), Maoist China and Stalinist Russia (Blakeley, 2009) deployed state terrorism but this time on an industrial scale and killed millions. Later in the 20th century, Military juntas utilized state terrorism as part of strategy to suppress populist left leaning or communist liberation movements, who themselves often employed terrorism.

   This use of state terrorism as a method of affecting a counter insurgency strategy was often openly supported by liberal democracies. This was done by various means which ranged from simply condoning acts, disregarding or indifference to acts of state sponsored terrorism, to the supplying of weapons and training to the providing of military (particularly special forces troops) aid and intelligence advisor's for training, through to large scale deployment of military force.

Examples of this would be US policy in central and South America through out the 1970's (Gareau, 2004), where everything from providing training (particularly through the use of the school of the America's) to the provision of advisor's and weapons (Koonings and Krujit, 1999) could be labelled as terrorism. It should be remembered though that the use of labelling acts as state terrorism is extremely contentious, as it can often be attributed incorrectly and is often not considered terrorism. The GTD for instance does not consider state terrorism as a form of terrorism.

2. The new terrorism. This type of terrorist activity came about at the end of the the 20th Century. It is characterized by its wish to carry out large casualty type attacks, these groups are defined as having innovative command and control structure, a pan-national religious or political shared belief and their own moral norms for the justification of political violence. Al Qa'ida would be the prime example of new terrorism, with the attacks on September the 11th being the most publicized attack of this type (Burke, 2004). Another well known example of such a group would be Shinrikyo cult, (Morgan, 2004) who carried the Sarin gas attacks in Tokyo in the mid 1990's.

3. Non-state terrorism or dissident terrorism. This is terrorism that is carried out by non-state actors against either the government or different ethnic groups. Examples of dissident terrorist groups would be the IRA in Ireland or ETA in Spain (Lutz and Lutz, 2009).

4. Religious Terrorism. This can be defined as a type of terrorism where the group carrying out the terrorist acts believes they are being allowed to do this through their faith in a higher being. Religious terrorism is often accompanied by literal interpretation and a strict adherence to a persons holy book (Pratt, 2015).

5. Ideological terrorism is a type of terrorism inspired by a belief system or political ideology. The central theme of this type of terrorism is a blind belief in a political or religious ideology. On a political spectrum terrorists come from both the left and the right.

6. International terrorism are terrorist acts which have very clear global repercussions or impact in particular countries. These acts aim to focus world attention on a particular cause. One of the most prominent examples of this were the Munich attacks, the kidnapping and murder of Israeli athletes at the 1972 Munich Olympics would be one of the prime examples of international terrorism. The aim of international terrorism is

to gain both attention but also to gain both acknowledgement and compromise from the people they are in conflict with. Abu Iyad, the architect of the Munich attacks commenting on the attacks, concluded on the outcome of the Munich attacks: *"The sacrifices made by the Munich heroes were not entirely in vain. They did not bring about the liberation of any of their comrades imprisoned in Israel ... but they did obtain the operation's other two objectives: World opinion was forced to take note of the Palestinian plight, and the Palestinian people imposed their presence on an international gathering that had sought to exclude them"* (Iyād and Rouleau, 1981).

1.6 Terrorist tactics

Terrorists must be adaptive in their tactics (Bennett, 2007) as the security apparatus maintained by government can quickly implement countermeasures. A terrorist can deploy a number of types of attacks:

1. Arson. The legal definition of Arson, defines it as the purposeful destruction of buildings, land or assets through the use of fire. Arson attacks are carried out against infrastructure necessary for the proper functioning of society. While arson attacks are less sensational than other types of attacks. They are easy to carry out, requiring little technological know how, no access to advanced weaponry as the materials required to build an incendiary device are cheap and readily available. One of the most notable recent uses of arson would be the the attack on the US mission in Benghazi in 2012 by Ansar al-Shari'a and al-Qa'ida in the Islamic Maghreb (AQIM) to kill Ambassador Christopher Stevens (Maldonado, 2015).

2. Assassination. This is the unexpected, audacious murder of persons who are strategically critical to a regime or group for doctrinal (both political or religious) reasons. Assassination does require a certain amount of technical expertise and may require access to advanced weapons and opportunity to get close to the target. The aim of assassinations are to strike at the heart of regime, instilling fear and chaos. The use of assassination by al-Jama'a's terrorists in 1981 to murder President Sadat of Egypt while attending a military review remains one of the most striking uses of the tactic (Haykal, 1983). Another example of the use of assassination as a terrorist tactic to remove key leaders of opposition would be the murder of the Ahmad Shah Massoud, the leader of the United front, main opponents to the Taliban and AL Qa'ida regime in Afghanistan (Wolf, 2003). This had the effect of removing the only person capable of offering a credible alternative to the Taliban (Rashid, 2001) in Afghanistan.

3. Cyber. Cyber terrorism is the wilful and malevolent disturbance and upheaval of nationwide or global computer networks by an individual or a group of individuals. This is achieved through the use of computer

viruses, DDOS (Distributed denial of service) attacks, computer hacking and damage to critical infrastructure (Golandsky and Dombe, 2016). High value targets such as high profile commercial buildings or developments, historic or government buildings and buildings of religious significance. While the potential of cyber terrorism has received considerable attention and caused much consternation, very few instances of cyber terrorism have been recorded and its use has been limited to attacking websites and communication networks. For example Hezbollah were able to hack into the Israeli Defence Forces secured mobile network ”Vered Harim” in 2006 (Golandsky and Dombe, 2016). Islamic Jihad have also developed specialist software to gain access to CCTV systems and track aircraft. However successful attempts to gain access to critical defence or public infrastructure have been limited, though in 2008 Russian sponsored attack on the Ukranian power grid (Perez, 2016) occurred. The benefits to using this type of terrorist attacks are that it is relatively easy to carry out and inexpensive.

4. Economic attacks. These are attacks designed to cause financial hardship or loss and adverse economic affects (Drake, 1998). For example terrorist attacks against ports and shipping cause the introduction of extra costs to pay for increased security and safety. Piracy can also be used to increase be as a tactic of economic terrorism, causing shipping companies and governments to increase spending in insuring maritime safety and also the diversion of military forces to protect shipping. This can lead to an increase in cost of importing and exporting goods.

5. Environmental. Environmental terrorism is the purposeful introduction of toxic or hazardous material into the environment with the aim of causing pollution. An example of this would be the poisoning of a countries water supply (Gleick, 2006).

6. Explosives. Through the use of conventional explosives such as RDX (a well known commercial military formulation is Semtex) and HMX sourced commercially has been used widely to build IED’s (Improvised explosive devices) (Kopp, 2008). Alternatively explosives can be acquired through the manufacture of explosives. Homemade explosives used by terrorists are often based upon inorganic salts and/or peroxides and can be synthesized from legally acquired chemicals, whose purchase would not cause suspicion (Johns et al., 2008). An example of usage of such compounds to such devastating affect where the Bali Bombings of 2006 (Royds et al., 2005).

7. Hijacking. Hijaacking is the illegal and illicit seizure of a mass transit vehicle by a group or persons with the aim of creating a mass media spectacle were the hostage situation can be used as a mechanism to force concessions (release of prisoners) or elucidate money from a government or organisation.

While aircraft hijacking is the most well known form of this type of attack, occasionally other mass transport types are targeted, notably ships with the seizure of the Achille Lauro (Halberstam, 1988).

8. Hoaxes or threats are used to by terrorist groups who have established a past record and capability of carrying out attacks as a weapon of intimidation. Hoax's can serve to not only intimidate but by keeping security forces in an almost constant state of alert and deployed, degrades their effectiveness and readiness through exhaustion from being deployed for extended periods (Nagl et al., 2008).

9. Stochastic terrorism. Stochastic terrorism is the utilization of information media (previously television and radio but now the internet) to radicalise and exhort random members of the population to participate in terrorism. The stochastic terrorist is considered the instigator of the violence and the perpetrator of the violence is considered to be the lone wolf. The term stochastic is used because the acts may be statistically predictable but individually random. Stochastic terrorism can not only be used to trigger individuals with views sympathetic to the terrorists world view but also people suffering from mental illness, personality disorders etc. In this way it makes the use of targeted surveillance extremely difficult. The idea of stochastic terrorism is problematic, concerning what role certain media played in a persons radicalization, as how do we know what specific method of mass communique triggered the terrorist attack. This has fundamental effects for civil liberties, as if dissemination of radical media is leading to triggering of terrorist attacks, then it would make sense to control this media, but this would violate freedom of speech and thought. Also to what extent the media had in triggering the attack as compared to the mental well being of the person is difficult to ascertain.

1.7 Does terrorism work?

In Alan Dershowitz's seminal work, *Why terrorism works*, he argues that the advancement of the Palestinian cause since the early 1970's (Dershowitz, 2002) suggests that terrorism does work and is therefore a completely rational tactic to opt for or employ to accomplish a political aim. Numerous case studies and studies based on game theoretic model have cited the example of Hezbollah whose sustained terrorist actions (most notable of these the 1983 attack on the marine barracks) after the US and French deployment of peace keeping troops in the early 1980's during the Lebanese civil war, forced their withdrawl (Atran, 2004).

A more recent example of this phenomenon would be the 2004 Madrid bombings, which resulted in Spain withdrawing its forces from Iraq (Rose et al., 2007). These studies have suffered from focussing on particular cases and being selective in nature, lacking the scope to establish a more universal truth (they are difficult to generalize).

While terrorism has often been studied using case studies or game theoretic's, very few studies have taken a data-centric approach focussing on outcomes of terrorist activities. One of the first such empirical studies, carried out by the US state department found that rarely did terrorist groups achieve all their aims (approximately only 7 percent of the time). The tactic of terrorism is strongly correlated with a group not meeting its objectives (Cronin et al., 2004). In Abrams empirical study on whether terrorism achieved its goals, very little evidence was obtained to support the hypothesis that terrorism was an effective strategy, infact the opposite was found, very rarely did terrorism achieve its goals. (Abrahms, 2006).

1.8 Countering terrorism and countering the terrorist narrative

Counter terrorism is the unified application of theory, practice, specific military techniques and approach (including specialist units), government strategy which can be deployed to impede and eventually defeat terrorism (Jackson, 2005). Just as terrorism tactics can take many different forms so can counter terrorist strategies. From a top down view an anti-terrorist strategy starts with a policy direction dictated by a government. This policy dictates the (political) direction and how the counter terrorist policy will be enacted as laws and what powers these laws will give military, police and (para military) intelligence agencies in combating terrorism.

They may also bring about structural changes to the forces who carry out counter terrorism by changing the structure of these entities, changing the focus of the organisation or merging or forcing more collaborative work (Jamwal, 2003). Governments may also pioneer the use of new technology in countering terrorism, one of the most visible of these technological developments, was the development of mass data collection and data mining to help process the information for investigation that emanated from the war on terror. Finally governments will also deploy military, police or para military (the meaning of the use of paramilitary forces refers to intelligence operatives) forces against terrorist which may be either domestic or in some cases (the war on terror, the Israeli response to the Munich terror attacks) foreign (Calahan, 1995). The type of military / para military / police actions taken by a government range in their scale and breadth. These operations can range from surveillance, propaganda, arrests and raids on terrorist locations with the aim of interrupting their activities, large scale deployment of troops, assassinations or targeted killings upto full scale invasions of territories where terrorists operate from (Conetta, 2002). Governments may also seek diplomatic solutions to terrorism, from building coalitions to fight international terrorism or 'The new terrorism', to changes in international law to make it easier to investigate or prosecute terrorism. Diplomatic solutions to the prosecution of terrorism often cause liberal western democracies particular problems as they require cooperation with states they would otherwise be strongly opposed to (Jarvis and Lister, 2014). An example of this would be Saudi Arabia, a country whose citizens have provided much material and human capital to Al Qa'ida (Abuza, 2003). They may also require cooperation with groups who could themselves be regarded as terrorists. This has the effect of opening these governments up to criticisms of hypocrisy or imperialism as they are allying themselves with people who are themselves despotic or are seen as an invading alien (from a cultural and political point of view) force. During the Colombian response to FARC terrorism in the 1980's and 1990's the war on FARC by government police, paramilitary and military forces was supplemented with right wing anti-communist militias. These groups can aid governments through the provision of additional manpower. However, Groups such as the United self defence forces were responsible for a number of vicious campaigns against FARC forces (Duffy). They were also responsible for the killing of many civilians and people whose ideologies they opposed particularly trade unionists, communists and democratic socialists (Peceny and Durnan, 2006). Such groups actions can act as a obstacle to peace with terrorist groups unwilling to negotiate with governments while such groups exist (Alsema, a).

In circumstances were force is used, it is important that it is used as judiciously as possible. The rules of international law should be followed, while at an operational level certain practices may be unique when compared to war (Roberts, 2002). A measured response is required with civilian casualties being kept to a minimum. Over zealous counter terrorist responses can also act as a propaganda tool for the terrorist themselves (as they are seen as an opposing force to perceived imperialists or forces who are allied with undemocratic regimes), by framing the counter terrorist activities as being a disproportionate use of force especially when civilian casualties are high. However minimization of casualties can sometimes be difficult especially when countering terrorists operating in densely populated areas (Graham, 2009).

Another concern for governments is how they frame a counter terrorist strategy. For instance the US response to the attacks on September the 11th 2001 was to declare a 'war on terror'. This terminology frames the terrorist campaign instigated against US which was tantamount to mass murder, as not a criminal act but as an act of war and the terrorists as not criminals but as soldiers. This legitimization of terrorism serves to further the cause of the terrorists through dignifying it and terrorists as combatants (Moeller, 2009). One successful counter terrorism strategy employed by governments is that based upon the British experience in Malaysia and the troubles in Northern Ireland, that is that terrorism is a political problem and requires a political solution. This was set out in a counter terrorism framework by Robert Thompson (Hamilton, 1998) who created five fundamentals of a counter terrorism (Thompson, 1966), these being:

1. The government has a well defined political goal.
2. The governments intelligence services, police and military forces must operate within the bounds of the law.

3. The government, it's intelligence services, police and military forces have a clear strategic framework for defeating terrorism with all actors working together to achieve this and not in competition with each other.

4. The governments focus must be on collapsing the terrorists attempted disruption and overthrow of the political system and not defeating the group militarily.

5. A government must first secure its hinterland or initial area of operation before moving into the areas affected by terrorism or insurgency.

   While this strategy proved successful in Malaysia, a successive British administration in dealing with the troubles in Northern Ireland, initially followed almost exclusively a military strategy before moving to a more successful strategy modelled on Thompson's doctrine (LaFree et al., 2009). During the 1970's IRA (British) mainland bombing campaign, the 1974 prevention of terrorism act introduced the term 'suspect community' by directing laws targeting specifically the Irish community (Hillyard, 1993), a direct contravening of the counter terror axioms developed by Thompson. This only served to make the Irish community in Britain more vociferous in their support of the PIRA and more removed from the British political system. It is also been argued that the Muslim community in Britain since September the 11th has also been under the law, categorized as suspect, leading to discontent and isolation of the Muslim community and creating the environment for radicalization of the Muslim community (Pantazis and Pemberton, 2009).

This argument has been countered by the narrative that "the new terrorism" is a more severe form of terrorism and is in-fact becoming more warlike, being capable of destroying western liberal states, it is important that this type of terrorism be fought on a war footing, (Bobbitt, 2008).

Governments face a challenge in structurally changing their intelligence services and military to fight a counter terrorist campaign instead of a more traditional and conventional conflict. Often the technology they have put in place for use by their military is not fit for purpose and neither is the structure of the military whose forces are setup to fight a conventional war and not a counter terrorist campaign (Gazette, 1989). This position can be further complicated as a government must maintain a significant conventional force as a deterrent to conventional wars and in case an actual conventional conflict breaks out (Gates, 2009).

1.9 Research Question to be addressed in the study

Businesses (particularly insurance), governments and risk management professionals face particular challenges with regard to terrorism. Being able to diagnose changes in intensity of terrorism is crucial so as to take action to take account for this change in behaviour. The most obvious changes in intensity would be increased occurrences of terrorist incidents or deaths due to terrorism. Other changes would be for instance a change in the attack vector (i.e. assassination, bombing, attacks on infrastructure) or weapon type (i.e. biological, chemical, explosives). Others would be has there been a change in geographic location of attacks (or what is the current geographical base of attacks is there a rural urban divide).

These changes in behaviours may then prompt a different response from business or government. For example, for the purpose of re-insurance, a re-insurer may want to be alerted as soon as a change an increase in deaths due to terrorism or terrorist incidents is detected so it can account for the higher risk associated with the increase in terrorism, by increasing reinsurance costs. However it would also want to be sure that there has been a change in behaviour and not just a 'one-off' anomalous attack. For a government, a behavioural change in terrorism (due to change in incidents or deaths due to terrorism) would provide an impetus to carry out an intervention to counteract the change. Governments may also want to examine the effects of a specific intervention and has it had the desired effects (i.e. a reduction or stabilization in levels of terrorism in terms of either events or deaths due to terrorism).

The benefits of using an electronic terrorist incident database is that they expand the range of studies that can be carried out to determine whether a change in behaviour with regards to terrorism has taken place. Terrorist incident databases also ensure the data is of higher quality and reliability as it has already been processed to improve quality and classified (or encoded) under certain criteria to aid in its use. They also offer the benefit in that they are easily accessible and can be accessed by businesses, governments or researchers free of charge.

Particularly they allow the application of statistics and data mining techniques to the study of terrorism.

The research question directly addressed in this thesis is whether one can detect changes in behaviour associated with terrorism. These would primarily be changes in the intensity of terrorism in terms of deaths or counts of incidents, but also have there been changes in terms of temporal,temporal spatial changes, changes in attack or weapon vector. Temporal spatial effects would be if there has been a change in the precedence of regions or countries in terms of terrorism over time. To examine whether changes in behaviour have occurred, a number of modelling techniques which are appropriate to modelling count and time series count data are utilized. The methods specifically used to model the count of deaths and incidents in Iraq are count regression modelling techniques and count time series analysis. The latter techniques are used to identify 'interesting' time count events, which can take the form of mean or gradual shifts, time series outliers or outbreaks (either using a E-Divisive with Means or syndromic surveillance based methods) or 'epochs of high and low terrorism' and the probability of being in one or the other, detected using Hidden Markov Models (HMM's).

1.10 Thesis overview

The thesis will take the form of:

1. Chapter 1, serves as an introduction to the topic of terrorism and includes the following; what defines terrorism, terrorism in an historical context, what is counter terrorism and how one counters terrorism. This chapter also details the particular research question addressed in the thesis, can one detect changes in behaviour in terrorism using the terrorist incident database.

2. Chapter 2, will be an introduction to data and data mining approaches to research in terrorism and in investigation of terrorism along with its application to counter terrorism. The chapter also covers the failures and challenges faced concerning the use data mining when applied to terrorism or counter terrorism research.

3. Chapter 3, will be an introduction to electronic terrorism databases, open and closed and their use in analysis of terrorist incidents. The chapter also covers the use of data held within the database, how the data is encoded and what problems arise when using the GTD. A review of the use of the GTD in terrorism research is also carried out.

4. Chapter 4, will address the application of machine learning techniques to the investigation of terrorism incidents. Both Chapter 4 and 5 address the specific research question at the centre of this thesis can one detect a change in behaviour (in terms of incidents or deaths) of terrorism associated with a particular country. In this chapter a number of methods appropriate to the analysis of time series count data are applied to the time series dataset of post invasion count of deaths in Iraq created from the GTD. A number of algorithms are applied to the dataset, two of these are online time series detection algorithms (the twitter outbreak detection algorithm and Netflix's SURUS algorithm) and two medical syndromic surveillance algorithms (EARSC based method and Farrington's method).

   Application of outbreak detection algorithms using purrr over multiple countries is also demonstrated.

5. Chapter 5 will serve as conclusion to the study, detailing both the usefulness of data mining techniques but also any insights uncovered from the analysis.

6. Appendix A, contains an exploratory analysis of electronic terrorism databases and ancillary datasets such as polity, transparency international or Fas.org. A number of steps were carried out in this section including exploratory analysis using visualizations, generation of summary statistical data and exploratory analysis. Other steps in this section deal with cleaning and validating data and the methodology used through the study (CRISP-DM). The exploratory analysis carried out in this chapter was directed

with the motives of gaining an understanding of both the data held in GTD, to uncover any changes in behaviour in terms of terrorism on both a global and regional level and to discover appropriate techniques for modelling the data for the purpose of being able to detect changes in behaviour. Preliminary modelling of the data is carried out using count regression modelling and time series count data with HMM's. The output of these models use is discussed along with problems with their usage particularly model specification (in case of count regression modelling) and difficulties in generalizing the models (in the case of HMM's).

7. Appendix B contains preliminary modelling regression coefficients and contingency table s supporting the CA used in the preliminary modelling.

8.

1.11 Discussion

Terrorism while often being portrayed as a recent phenomenon, it has ancient roots demonstrated by presence of terrorist groups such as the Sicarii and the Assassins. While having its roots in ancient times, the tactic of terrorism became highly popularized firstly by national liberation movements and left wing groups in the second half of the 20th century. Then by religious / pan nationalist groups (in the early 21st century) which are now the most common perpetrators of terrorism. Terrorism is most commonly defined as the use or the threat of violence against the populace, the agents of the government and government with the aim of advancing their political/ideological/religious beliefs.

Another key component of terrorism is that it is carried out by 'non state' actors. This last point can be problematic as the use of terrorism by government agents such as sections of the security forces or government backed paramilitary forces is a reality. This second attribute of terrorism has proved often to be a point of contention with academics, with many of them disagreeing with this second component. Many examples exist of state sponsored terrorism from the Nazi's use of terrorism in occupied countries in WW2, to the use of state sponsored paramilitary forces to terrorize the populace in central America in the 1980's, exemplified by the murder of Arch Bishop Romero of San Salvador in the 1980's (Romero and Brockman, 1998).

While state terrorism is one particular form of terrorism (or in the strictest sense of the word not) other forms of terrorism include religious terrorism, dissident terrorism, idealogical and international terrorism, but the most prescient of these types of terrorism is what is known as 'new terrorism'. New terrorism has become the most prevalent form of terrorism and is exemplified by attacks which create the most casualties and are carried out by groups which are characterized as being pan national and/or religious, such as Al-Qa'ida or the Shinrikyo cult. It also employs innovative command and control structures and particularly innovative communication techniques.

New terrorism has become strongly associated with the tactic of stochastic terrorism a term used to define the way in which mass media methods utilizing the internet like youtube, facebook, twitter, online manuals are used to disseminate their methods, their core message and their attacks with the hope of encouraging random people to carry out terrorist acts inspired by the material (Margulies, 2016).

Terrorism can be seen to have a number of over-arching tenets besides the use of violence to coerce a change and it being carried out by non-state actors, these are by the standards of international law as laid out by the Geneva or Hague conventions, they are illegal. However to the terrorist they are considered to be a form of 'just war' being permitted through either a religious, political or ideological perspective.

When looking at why terrorism occurs it is often wrongly argued that a legitimate grievance must exist for the conditions to be created for a terrorist group to flourish. This is not always true often terrorist groups are constituted from wealthy, well educated middle class and young disaffected people who see the only chance of affecting political change through terrorism. Examples of this behaviour would be the leadership of the red brigades in the 1970's in Europe and Japan to the leadership of Al Qa'ida who emerged and flourished throughout the 1990's. Both Al Zawahiri (Al Qa'ida's operational brains behind the attacks on the world trade centre) and Osama Bin Laden (the leader of Al Qa'ida at the time) were both from wealthy middle class backgrounds (Henzel, 2005).

The strategic and tactical aims of terrorism are disparate and diverse. The strategic aims of how a group attains a political or ideological change against a government is achieved through many strategic aims. These range from attrition by making the cost of continuing to oppose the terrorism impractical and inordinate, intimidation where the terrorist aim to subjugate the populace to its cause by demonstrating an ability to punish lack of perceived loyalty of the populace and also to show a lack of ability on the governing bodies part to protect its citizens.

Another common strategic aim of terrorism is provocation which is the use of terrorist acts to provoke actions against the populace. Some of the most evident examples of this would be the bloody Sunday killings by the British army in Derry in 1970 where a overtly heavy handed response to a peace March in Derry which ended in the killings of 14 people (Dawson, 2005). The subsequent invasions of Afghanistan and Iraq following the September the 11th attacks could also be argued to have been an example of a provocation strategy. The last two strategies associated with terrorism and can largely be seen as competitive strategies amongst terrorist groups to establish themselves as the dominant group, these are outbidding and and spoiling.

Spoiling is where groups try to dissuade dialogues between rival terrorist groups and the government, while outbidding is where the terrorist group tries to demonstrate it being the dominant force of opposition to the government by carrying out more and more audacious attacks. Examples of spoiling would be the continuity IRA's terrorist campaign post the peace process to try and reignite the conflict in Northern Ireland (Whiting, 2015). An example of the outbidding strategy would be the 'IN AMENAS' attack by a Mokhtar Belmokhtar led group of terrorists who broke away from Al Qa'ida on the Tigantourine gas refinery, which was carried out by Belmokhtar and his followers had become disillusioned with AlQa'ida's campaign in North Africa and wanted to up the ante (Watling, 2015). The tactics employed by terrorism are also diverse ranging from more traditional methods such as arson, assassination, hijacking of mass transit vehicles to more modern forms of terrorism such as environmental to cyber attacks. Cyber terrorism is of particular concern to governments as not only can it damage critical infrastructure it can also undermine confidence of the public in the government (Gross et al., 2016). It is also difficult to counter and cyber terrorist operations are inexpensive to launch though do require technical expertise (Blakemore, 2016). Why terrorism tactics and strategies are important to understand is that they inform both the counter terrorist approach at an operational level and at a governmental level. For a government this may involve making a political, military intervention to remediate this change in behaviour. At the key of any counter terrorist strategy is government policy which will direct and inform all counter terrorist activities and whether they be military/paramilitary or political in nature. Successful frameworks as laid down by successful practitioners such as Robert Thompson dictate that a government must have a clear political goal, it must operate within the bounds of the law, its focus should be focussed disrupting the terrorist groups operations and particularly winning support of the local people.

It also requires securing first the hinterland before moving into areas affected by terrorism. However this strategy has strongly been contested especially in the light of what is referred to as 'new terrorism', defined by being a more extreme, more violent and having a more advanced command and control structure(which is characterized as being loosely interconnected and having very limited command and control structure) and employ 'stochastic terrorist' methods. This requires a change in the way governments counter the threat posed from terrorism, this necessitates governments to view terrorism not as a criminal act but as war. This compels government to make changes to its military and security apparatus and calls for enhanced inter-agency and multinational cooperation between governments. Central to a counter terrorism strategy is an effective counter cyber offensive measure (particularly countering their communications networks or methods of disseminating information), better terrorism informatics, air power and the use of special forces troops and a move away from more conventional forces. (Lesser et al., 1999). However this strategy has evolved, (Lesser et al., 1999) to a more wide ranging strategy that encompasses strengthening and deepening deterrence, limiting the likelihood of the use of weapons of mass destruction to avoid acts of terrorism of unprecedented magnitude and a number of other initiatives, which include:

- Increase the capacity and willingness of states to act against terrorism, when it occurs.

- To place a microscope on terrorist groups and show what they are doing and how it affects the population.

- To decrease the amount of chaos and instability that the terrorist base is located in. To codify counter terrorism strategies as part of formal international treaties and alliances.
- To limit a countries exposure to terrorism by limiting as much as possible by hardening a countries infrastructure and to possible effects of terrorism, More controversial it may also require governments to target non state actors and individuals.

1.12 Conclusion

Terrorism is a difficult subject to define with no universally agreed definition of terrorism. It does have traits or tenets which define it, these include; the utilization of violence to gain an ideological gain whether that be political, religious or ideological aim. It is carried out by non-state actors (though they can be state sponsored in the case of state sponsored terrorism) and its aim is to influence the public as well as the government. While the conditions for the rise of terrorist groups has often been wrongly argued to be that of a strongly disaffected ethnic or politico-economic minorities, this is not always the case as is the example of the red brigades of Europe and Japan of the 1970's and 1980's. Groups carry out terrorist acts for a number of reasons; to intimidate the populace, to inspire the populace (to rise against the current ruling entity) by attacking the government.

Terrorism also serves to undermine the government by attempting to disrupt its supply of critical services to the populace or to cause economic effects by discouraging DFI (Direct Foreign Investment) there by underpinning its ability to govern. These acts while at tactical level do contribute to the strategic aims of terrorism these include attrition, intimidation, spoiling, provocation and outbidding.

Terrorism is also diverse in the number of vectors that are employed to carry out acts of terrorism, these include, Arson, assassination, cyber-terrorism, environmental, hijacking, use of explosives, hoaxes or threats and economic attacks.

Counter terrorism efforts are techniques used to impede and eventually defeat terrorism, a counter terrorist strategy is policy directed, by a government, that dictates both the direction of counter terrorism policy and the tactics of how the counter terrorist policy will be enacted. A number of successful counter terrorist strategies have been proposed and been shown to work and frame works have been created for their implementation. The most notable of these are those of Robert Thompson, who employed them in the successful Malaysia campaign against Maoist terrorists in the 1950's and 1960's. Key to Thompson's framework was the implementation of a well defined political goal, the operation of all state actors (military, police etc.) within the bound of the law, the securing of an initial base of operations in the hinterland before expanding operations to areas outside this.

The key objectives of the research to be carried out in this thesis is to be able to detect changes in behaviour regarding terrorism these changes may be changes in intensity of attacks, changes in attack vector weapon type or geographical location of attacks (for example whether there is a rural urban divide). For example, changes in intensity of attacks are monitored by detecting changes in counts of deaths or incidents and being able to classify the type of change detected, whether it is a persistent change, or is it signalling an outbreak or a non persistent anomaly such as a one-ff spike or dip in terrorist attacks. Being able to detect such a change would be advantageous to businesses or governments as detection of such events could allow these entities to take remedial action to counter such changes. For a business such as a re-insurer this may involve accounting for these changes by increasing re-insurance rates once a change in behaviour has been detected, or even not offering re-insurance till a change in conditions (to a lower intensity) is detected.

Chapter 2

Datamining approaches to research in terrorism

To counter terrorism not only requires that governments research new methods to collect and assemble intelligence but also to scrutinize it so as produce actionable insights. During the cold war much intelligence could be gained by utilizing a limited number of sources which were rich in information and analysis of such, could gleam a large amount of understanding into an enemies effectiveness and competencies in different areas, their disposition and stratagem. Traditional methodologies used in intelligence gathering include (Tanner, 2014):

1. HUMINT (human intelligence) is intelligence collected from actors on the ground. These can include specialist troops, intelligence operatives, captured combatants.

2. GEOINT (Geospatial intelligence) Geo-spatial analysis carried out by satellites, drones and dedicated spy planes.

3. MASINT (Measurement and signature intelligence). This is a methodology used in intelligence gathering to identify, trace and recognize characteristics of source targets. Sources of MASINT would include radar and acoustic intelligence. The most famous of these would be SURTASS and SOSUS (Noles Jr, 2003) which are collections of hydrophones used to track surface and sub-surface maritime traffic.

4. SIGINT (Signal intelligence), this is the analysis of intercepted signals (Aid, 2003).

5. TECHINT (Technical intelligence), is intelligence gathered from the technical analysis of equipment.

6. FININT (Financial intelligence) is the generation of intelligence from analysis of financial records.

7. CYBINT/OSINT (Cyber intelligence and Open Source intelligence) (ŞAHİN) is information gathered from the web or open source intelligence and would be considered a new form of intelligence. Automated data mining and data analysis can be applied to both old and new intelligence methodologies but has been applied most notably to CYBINT and OSINT and has in-fact largely enabled these fields.

   However for all methodologies data integration is important, enabling the analysis of disparate types of data such as text, audio, biometric information, technical plans and imagery.

2.1 Terrorism informatics

Terrorism informatics is the utilization of a number of different techniques for combining disparate data from different sources and analysis of the data to provide insights to intelligence specialists (Chau et al., 2015). This requires a number of methodologies and system sourced from a number different disciplines which include computer science, business informatics, statistics, machine learning, computational linguistics and social network analysis. The analysis needed to produce actionable insights is usually produced in two ways by either data analysis and/or data mining. A framework for data mining for counter terrorism includes problem description, data gestation, data mining, model appraisal and illustrating or reporting the knowledge gained in a succinct form.

Data mining is the automated extraction of hidden patterns from databases (Shmueli et al., 2016). Established data mining approaches include supervised methods such as classification, regression/prediction or forecasting. Unsupervised methods include such methodologies as association rules mining, cluster analysis and outlier detection. More recent developments for identification of patterns within data include entity extraction. This is used to analyse text imagery, audio recordings and has been utilized to automatically identify suspects, whereabouts, cars and personnel traits from police reports. Data mining began to seriously grab the attention of the US government in 2002 (as a response to the September 11th of 2001 attacks) as a means to counter terrorism. Data mining efforts applied to the investigation of terrorism had pre-empted September 11 by a number of years with the building of 'Able Danger', a counter terrorist data mining tool. 'Able Danger' (Keefe, 2006) had identified a number of the 9/11 terrorists prior to the attacks but the intelligence services had failed to act on it (Lance, 2006). This has been superseded by other data mining tools (particularly network access tools) such as the NSA's special access program which builds a social web from a suspect person, examining their links to people, through emails, phones, address and other information used to link people. This type of network based on links is one means to carry out social network analysis, another means of using network analysis to detect suspect behaviour is to use affiliation networks. These are networks in which members of the network are connected by having common interests and has been shown to be effective in corporations detecting fraud (Ben-Itzhak, 2009).

Automated data analysis involves the automation of data analysis through the use of subject based queries or pattern based queries. Subject based queries start with a specific subject and expand out from this subject by finding entities linked to this person. These could be other people, places, vehicles, phone numbers, emails or activities. This type of analysis is often referred to as link analysis (Berry and Linoff, 1997) and has not only been used for terrorism research but has found commercial usage within fraud investigation work in the financial sector particularly in banking and insurance. The benefit of using this approach is that it can be used to find leads to other person(s), places, vehicles, phone numbers or emails. Commercial based systems such as I2 (IBM, 2016) (which is also used by military, intelligence and law enforcement) is used within fraud prevention units to investigate fraudulent claims and financial transactions.

Pattern based query automated data analysis involves the use of a pattern or a predictive model to identify a particular type of behaviour in a dataset. Again this type of analysis has seen considerable adoption, for example Detica NetReveal (Oatley and Ewart, 2011) which is used for finding patterns of credit card fraud amongst other applications in the financial services.

Before analysis of the data takes place, the data must first be collated, the quality assessed and then if necessary be cleansed. At it's simplest, this process may involve removal of identical records, normalizing the data, removing unneeded data (not only because it may be unneeded but because it may be illegal to hold such data due to privacy laws). Other typical tasks would be the consideration of missing data and how to allow for it (for example through imputation of missing values) and the regulation of data types. More advanced pre-processing would include integration of disparate data through extraction of metadata from the original data (DeRosa, 2004). Examples of this would include extraction of time and place from a temporal and geo-tagged photograph. Also the heterogeneous nature of data used in counter terrorism have necessitated the adoption of new types of database technology which allow for the storage and querying of heterogeneous data.

Data stored in a NOSQL database does not require data to be cleansed to the extent it must be to be useful, when stored in a non NOSQL (typically relational database) database. NOSQL also allows the storage of more data and is extremely important were a lot of data is generated such as analysis of logs, sensor data (or large arrays of sensors in the case of IOT) or social media (Jeberson and Sharma, 2015).

2.2 Data mining approaches for counter terrorism

Due to the transnational nature of modern terrorism, the emergence of the internet to enable rapid communication and the decentralized nature of these groups, automatic collection and analysis of data allows examining text for words or phrases synonymous with terrorism. The DHS (Department of Homeland Security) uses a system to carry out entity extraction for terrorist specific terms such as 'Sarin', 'Anthrax', and different types of improvised explosive devices. These systems if they work coupled with information from financial details, social network information and geodesic data can give an investigator a much richer set of data to investigate and also give more context to alerts.

Though complex investigations involving large complex multiple data and data type investigations remains a huge challenge for the intelligence services. In the 2013 Boston bombings 10 tb of data were generated within 24 hours of the crime by the FBI regarding the terrorist incident (Jeberson and Sharma, 2015). The data came from a number of disparate sources including phone tower cell call logs, sms's, social media data, images posted to instagram and camera phone and CCTV footage. Advanced data pre-processing was used to carry out automatic face recognition and other biometric identification. Although not designed specifically for counter terrorism but for criminal pattern analysis, systems such as NYPD's 'Domain Awareness' system are capable of similar feats and carry out integration of CCTV and traffic camera imagery along with processing the imagery to extract number plates, speed and location data, being able to trace completely a vehicles journey throughout New York (Coscarelli, 2012). The capabilities of the system can be further enhanced through the integration of data from arrays of IOT devices consisting of radiation and biochem sensors along with integration of more traditional data sources.

2.3 Terrorism research using data mining and statistical techniques

Research into terrorism has also been enabled through machine learning and big data technologies. By integrating data from Google News with other information located in terrorist incident databases researchers were able to show a relationship between particular types of attack and individual groups ideology (Strang and Sun, 2015).

Researchers have also been able to observe a power law between number of people killed and the frequency of terrorist incidents. To explain this phenomena, researchers built a model to explain why this happens, about how terrorist groups dissolve and how often they can commit major atrocities and how easy it is for them to enlist willing volunteers. The model almost perfectly reproduced the power distribution (Clauset and Young, 2005).

An investigation has also been carried out into why some groups persist over others. Using a dataset partially sourced from the global terrorist incident database (GTD) controlling for well established theories regarding the longevity of terrorist group and theory around terrorism itself. A explanation was constructed from the study to account for the effect of inter terrorist group rivalry has on the length of time a terrorist group will persist. That is, the larger the number of terrorist group in a country the less likely the group will survive.

The capacity and ability of a group to carry out transnational attacks, carry out different type of attacks, carry out attacks with higher levels of mortality the likelier it is to persist (Young and Dugan, 2014). Regression methods (particularly survival) are quite common to the study of counter terrorism. Particularly, intervention analysis to quantify the effect of US interventions have on the number of attacks due to terrorism (Enders and Sandler, 1993).

The application of social network analysis by researchers to terrorism has a long history dating back to the 1980's, when pioneering work by (Sterling, 1981) described links between different terrorist groups and their backers (the KGB), the IRA and Palestinian terrorist organisations. The application of network analysis has been shown to be one of the most fruitful means for carrying out terrorism research. A number of reasons have been cited for this, these include:

1. Social ties and social influence have been contended to be key in the course of someone's radicalization (Hegghammer, 2006). The reason for this being that most peoples company they keep and social connections where in place prior to some-ones radicalization.

2. Network methods allow for a true representation of the internal structure of terrorist organizations without introducing incorrect beliefs about how a terrorist group should behave. Instead it allows the true composition and form of an organisations structure to emerge, often giving rise to the discovery of unexpected conclusions from the data such as the discovery of central actors, or channels of communications (Morselli).

3. The task of mapping terrorist networks is extremely useful from a counter terrorist perspective as it can potentially improve the capabilities of counter terrorist efforts. That is because it allows the pinpointing of key members whose removal would cause the most disarray to a terrorist network (Joffres et al., 2011).

The application of SNA to terrorism started in earnest after the September the 11th attacks in the US. Social graphs constructed after the event (Krebs, 2002b), (Krebs, 2002a) were used to visualize the connections between the hijackers and their support network who supplied financing, training and intelligence support to aid the in the attacks. The two central actors in the network Nawaf Alhazmi and Khalid Almihdha were never more than two links from any of the other nodes (in this case the terrorists and their support in the network). it was stated by (Krebs, 2002b) that if closer attention had been paid to disrupting the activities of key actors in the network, individuals with high connectivity and a distinctive skill set, the network once it had been found it could have been easily disrupted. (Sageman, 2004) utilised the profile and history of 366 members of the "global Salafi network" to construct a network based upon a number of characteristics including extended and immediate family ties, acquaintanceship, religious beliefs and work associations and affiliations. Four large clusters were identified around Al Qa'ida central staff, north African Arabs (Arab Shamal Ifriqiya), middle eastern Arabs and south east Asia (Indonesia and Malaysia).

Analysis of terrorist cells found a large density of ties between 17 members of a large Jemaah Islamiyah cell that carried out the Bali bombings in 2002 (Koschade, 2006). Both the operational leader and organizational and support structure leader had the largest score on a number of different centrality measures, which indicated they were key actors in the network. Due to the high connectivity of the cell and the high centrality of key figures in the network, discovery of one member would have led to the detection of the whole network. Network theory has also shown the changing or evolving structure of terrorists from a top down corporate like structure, to a coupled network structure to a loosely coupled network (Jackson, 2006). SNA has been also used in conjunction with other tools such as information extraction using text mining to automate the elucidation of data germane to the discovery of the network structure of a terrorist network. Such an approach has been used to create terrorist behavioural activity model (**?**).

Unsupervised methods such as HMM's and clustering have also seen use within terrorism research. HMM's have been used to model the temporal evolution of suspicious behavioural patterns in terms of financial records, intelligence communique's, media (print and web) articles and email communications. HMM's have been used with this data to identify atypical behaviour of terrorist activities from the signal associated with normal behaviour(Allanach et al., 2004). HMM's have been applied to the study of the activity patterns of terrorist groups, that is spurt in activity or sudden drops in activities or an end to the groups activity. To do this a n-state hidden markov model (HMM) is developed which captures the inherent states underpinning the the dynamics of a terrorist organization and from this an activity profile for the group can be developed.

An alternative to the use of HMM's for the study of terrorist groups organisational health and ability to rebound as well as its level of organisation is to use a method based on the detection of large increase in the activity profiles of groups. This approach involves binning the count data associated with terrorist activity and then analysing them in comparison to one another. These observation vectors of terrorist data are then transformed via a series of functionals inspired by the use of majoritization theory schema to deliver a spurt classification. This methodology was found to give a small number of incorrect classifications when compared to the parametric methods such as HMM (Raghavan and Tartakovsky, 2016).

Clustering has been widely applied to spatial analysis of crime patterns. Particularly clustering has been used to construct crime motifs (Nath, 2006). Applied spatial analysis has also been applied for spatial forecasting terrorist events. Through the use of non-smooth demographic prediction to enhance the spatial prediction of terrorist events, substantial increase in forecasting power is achieved over the base model (utilizing past locations to predict future locations of attack) (Brown et al., 2004).

2.4 The history of terrorism informatics, a German, US and Israeli perspective

Terrorism informatics (and in particular data mining and data analysis) has been applied to not only study of terrorism but also to investigative analysis of terrorism. One of the first known instances of this was in West Germany in the 1970's. The German state was faced with a dedicated campaign of attacks by far left terrorist groups such as the RAF (Red Army Faction) and the red brigades aided and supported by Warsaw Pact countries (Leighton, 2014) and the acquiescence and support of these groups to aid in terrorist attacks by other groups such as Black September (Nacos, 2016). The West German states response was the adoption of the ideal of militant democracy (streitbare Demokratie), the giving of a comprehensive set of powers to defend a liberal democracy against those who wish to get rid of it and establish a totalitarian state (Rosenfeld, 2014). To lesson the overreaching effects of such sweeping powers the West German state was one of the first to adopt data mining techniques to group or cluster suspected members of terrorist groups so as to aid in targeting the correct group of individuals, through development of dragnet (Rasterfahndung). Dragnet was the integration of data from a number of different data sources to essentially act as a filter, so as to narrow the number of individuals to investigate (Weinhauer, 2014). This technique had a number of benefits not only to aiding in the investigation but also massive societal benefits, these were:

1. An efficient, targeted search for RAF operatives, which resulted in the defeat, elimination, arrest of the leadership (Hauser, 1997) and destruction of the first generation of the RAF (Weinhauer et al., 2006). Though it did disrupt the RAF it did not cause their elimination and the actions of the RAF did continue up until the early 1990's.

2. It was a targeted intelligence led investigation, not targeting innocent members of the public, which had been evident in previous uses of the concept of militant democracy, (De Graaf, 2010).

Israel due to its unique geo-political position has led to it developing advanced counter terrorism data mining capabilities. Israel faces a terrorist foe who now not only use the web to propagandise terrorism, fundraise, but also use it as a strategic tool to direct terrorism. Terrorists have also used widely available geodesic data as an intelligence tool to plan to carry out attacks. Handheld GPS or devices (such as smart phones) with GPS sensors can be used to coordinate attacks. Israel has also been victim to some of the more advanced or new tactis of the terrorist including cyber terrorism, one of the most high profile cases of cyber terrorism against Israel was the breach of an IDF spokesmans Twitter which was used to spread malicious misinformation regarding a leak at the nuclear facilities at Dimona (Williams, 2014). As part of Israel's overall counter intelligence strategy is counter terrorist informatics, with the five elements of its strategy being (Tucker, 2003):

1. Data collation, analysis and investigation.
2. Military, intelligence and paramilitary operations to target terrorist groups.
3. Biometrics and security measure to protect passengers on mass transit systems particularly the airline industry.
4. Prevention of chemical, biological and nuclear attacks. This includes the use of arrays of sensors, the integration of this information and the analysis of this information.
5. Improving the moral of the people and strengthen the resolve of the population in coping with a sustained campaign of terrorism operated against it. Israel has affected a policy of marginal gains to affect this particular strategy. This is the aggregation of small changes delivered by the methods above but also structural or environmental changes that led to significant changes in the security of the nation. This system of marginal gains has been recently popularized by David Brailsford who promoted the technique with the general public (Durrand et al., 2014) and who used it to massively improve the performance of the British cycling team. Similarly Israel has seen the use of everything from data mining to structural changes to its city to its pioneering use of airport scanners and pre flight screening to dramatically reduce the number of fatalities from terrorism. For instance when in 2014 a Palestinian terrorist attempted to crash his car into a number of pedestrians queueing at a bus stop, he was prevented from doing so by a concrete bollard (Proser, 2016) used to cordon off pedestrian areas from motor vehicles and potential motor based terrorist attacks. Such devices may have prevented the nice attacks of 2016. These attacks employed a 16 tonne cargo truck driven into large crowds pf people gathered on a promenade(Nesser et al., 2016).

As a response to terrorism, Israel has developed sophisticated information technology based tools to support counter terrorism. These tools are focussed on securing Israel's communication networks from cyber attack and the gathering of information from disparate sources such as intelligence reports and national and international anti terrorist databases to give a detailed collected, concise and complete assessment of threats and highlights the most high risk of these. This system is currently operated by numerous countries through out the world. This system came to large scale public attention in 2015 due to leaks within the French security services about its non adoption (JPOST.COM, 2016). The security services blamed its non adoption on political pressures. The non-adoption of the system and lack of similar capabilities was directly attributed by the same person as one of the failures in counter intelligence that led to the dereliction to detect the Paris attacks. The Israeli company Verint specializes in software designed to integrate and collate data from disparate sources providing analysis to security specialists (Zureik et al., 2010).

Israel's response to terrorism has also prompted a large interest in IOT, especially computerized perimeter security systems. These systems use large arrays of sensors, CCTV cameras, trip wire detection, IR (infra-red) detection which is then collected. The information is then processed and automatically analysed so as to detect intrusions. Companies such as IDSST and Orad Group (Gordon, 2011) have commercialized such systems and implemented solutions throughout the world for everything from protecting industrial plants, key infrastructure to helping securing the US Mexico border.

Since the September the 11th attacks of 2001 to 2011 it is estimated that the US has spent 1 trillion dollars on approaches and polices to fight against terrorism (Roche and Blaine, 2015). The rise of big data technologies has been correlated with the US's increase in use of data mining technology in counter terrorism. As previously stated the US had been deploying data mining for counter terrorist purposes pre 9-11, in the form of Able danger which was purported to have identified some of those involved in the attacks. Other systems which predated 9-11 were link detection tools such as EELD (Mooney et al., 2002) which was an information elicitation and link investigation tool which later became part of NSA's TIA (Total Information Awareness) program (DEIBEL et al., 2016). TIA was a program that was established by DARPA, which was focussed on the application of both surveillance and gathering of information to the automated or semi automated analysis and identification of terrorists and other asymmetric actors who may pose a threat to the US. The US has embraced data mining technologies and while mass surveillance systems were officially defunded by congress in 2003 a number of these projects were later reclassified under different names and were continued to be run. These programs gave rise to such systems as PRISM (SIGAD US-984XN).

PRISM is a mass surveillance system which collects a large amount of information relating to internet traffic including searching emails, search history, file hosting service providers, instant messaging, video chat and voice calls over the web etc. Any information that then matches court approved search terms (under the FISA amendments act of 2008) are handed over to the NSA and are collected and analysed by PRISM. While PRISM remains somewhat controversial it's importance to counter terrorism is reported to be significant and in the words of NSA Director (in 2013) General Keith Alexander before congress as being responsible for generating "uniquely valuable intelligence".

However others have criticized the system and it's usefulness stating that claims where the system played a pivotal role the information uncovered could have been found through other means.

This was most apparent in the case of Najibullah Zazi where the NSA claimed the use of PRISM and its unique capabilities had been key in investigating and the capture of the budding New York city bomber. These claims were quickly refuted and a number of systems that could have provided the same information were identified (Ohlheiser, 2013). Other less known systems are those developed for US customs and Border Protection forces who created ATS (Automatic Targeting System) and secure flight used by the TSA (Transportation Security Administration). Secure flight is a flight screening system currently in service with the TSA (SPEAR, 2015). While ATS is a screening system employed by US custom and Border protection (Jizba et al., 2015). ATS was attributed with recognising Rael Al-Banna as a potential terrorist and barred his passage to the US in 2003.

The investment in these technologies has also been followed by the commercialization of these technologies for non security purposes. This has also had the effect much in a similar way to Israel in the creation of successful commercial entities based on the technologies pioneered by the security systems. An example of this would be the commercialization of sophisticated platforms such as Palantir which consolidates different types of data from different systems (Soklakova et al., 2016), then creates a number of graph models based on the different sources of data and also provides additional advanced analysis capabilities.

2.5 Problems associated with data mining and counter terrorism?

So far the potential uses of informatics or data mining for counter terrorism purposes has been shown to apply to a number of use cases, these are:

1. Risk assessment. This can be either classification of a person as a terrorist or scoring risk for a person associated with them being a potential threat.

2. Generation of profiles. This involves the collection and collation of information on person providing this in a easy to digest format, providing simple self discovery tools to the user.

3. Discovery of networks. Link analysis, graph theory and graph visualization to allow discovery of terrorist networks.

4. Data collection and collation. This is the process of collection of data and data about that data (meta data) from disparate sources, cleaning the data, treatment of missing values and the storage of this data in readily available repository for querying by analysts.

Each of these poses problems from a technical, moral and legal perspective. These problems are discussed below but can be restricted to the following areas:

1. The dangers of misuse of mass surveillance, to target, minorities, people of differing politic or religious persuasion.

2. The problem of class imbalance when building a predictive model of having relatively few samples to train on.

3. The related problem of the 'base rate fallacy' (Bar-Hillel, 1980). This is common in a number of areas where the base rate of occurrences of what a statistical test or a machine learning algorithm is quite low, including fraud detection and intrusion detection (Axelsson, 2000).

4. The use of unsupervised methods where there is a very low frequency of a signal for terrorist behaviour.

2.5.1 The dangers of mass surveillance and its misuse

While data mining can be an important counter terrorism tool its use especially when used in tandem with mass surveillance techniques is extremely effective and important for counter terrorism operations. However care must be taken when using such systems so as to guard against their misuse. There are many examples of these systems being misused in both democratic regimes and non democratic regimes. The misuse of mass surveillance techniques pre-date the digital computer. Examples of this would be the attempted mass conscription of Norwegian males in WW2 by occupying German forces. On learning of the proposed plan Norwegian resistance fighters instead of destroying the files which were going to be used to base the conscription orders of the men, they destroyed the machines used to sort the data. Without the necessary information to aggregate the population data, a draft of the Norwegian populace was too difficult to instantiate without access to properly aggregated by age cohort data (Bignami, 2007). Abuses of these techniques are not just limited to totalitarian regimes but also to democratic ones. During the civil rights era in the US, the FBI frequently wire-tapped Dr. Martin Luther King with his colleagues in the civil rights movement (Garrow, 2015). The wire-taps were done to try and establish whether Dr. King had any ties with the Soviet Union. Other notable examples of this are the Lyndon Johnston administrations use of mass surveillance to investigate if any of Barry Goldwaters staff were homosexual's (for the purpose of discrediting them) during the 1964 presidential elections (Sales, 2014).

2.5.2 The problem of class imbalance in counter terrorism datamining

One of the criticisms of data mining (for terrorism or terrorists) is that very few training patterns exist in terrorism due to the relatively few cases of it. This low frequency of it occurring (or people perpetrating acts of terrorism) has made it very difficult to detect. When carrying out a classification type supervised data mining task to build a model, this problem is referred to as class imbalance (Wang and Novik, 2015b). If this problem is not extremely severe, certain actions which involve either sampling (over or under sampling of minority/majority class), algorithmic means (using algorithms that are less susceptible to class imbalance), using cost based classification techniques or using synthesized artificial minority classes as surrogates for the minority class can be used to treat the class imbalance. However if the class imbalance is very severe, the above methods will not be effective especially over/under sampling, which will only serve to train a model to identify specific examples (Jonas and Harper, 2006).

2.5.3 The problem of low base rates of terrorism in counter terrorism datamining

This problem (of classification of terrorism) is exasperated by the fact even if you can train a model to accurately predict terrorist behaviour, the base rate of terrorism is so low the model may not be useful (Jensen et al., 2003). The base rate fallacy can be best explained with the following example. If the frequency of terrorists in a fictional city is very low, 100 terrorists in 1,000,000 people. Say if we build a system which has a true positive rate of 99% (sensitivity, also referred to as the recall, this measures the proportion of positives are are rightly classified) and a true negative rate of 99% (specificity, the proportions of negatives that are classified rightly) then in a city of 1,000,000 inhabitants where there are 100 terrorists. 99 out of 100 terrorists will will be classified correctly. Of the 999,900 of the non terrorists 9999 will be wrongly identified as terrorists. Therefore the detection rate will be approximately 99/(9999+99), which is 0.009803922 correctly identified as terrorists, this at best could function as a filter and at worst would either overwhelm the security apparatus of a state or a city with too many cases (due to the security services being overwhelmed of 'drowning' in false positives) to investigate.

Bayes theorem states that:

$$p(A|B) = P(B|A)P(A)/P(B) \tag{2.1}$$

Where A and B are observed incidents, P(A) and P(A) are the probabilities of said events, and P(A|B) is a conditional probability of observing incident A given that B is true. P(B|A) is the conditional probability of observing incident B given that A is true. from this we can get the equation for terrorist events occurring .

$$p(terrorist|T) = p(T|terrorist)p(terrorist)/p(T) \tag{2.2}$$

Where T means the algorithm indicates the person as a terrorist.

The base rate fallacy is not uncommon and exists in other fields, particularly in other related criminal investigation fields such as fraud detection. Other problems arise to do with data mining approaches to counter terrorism and may cause it to fail these include; bad data quality due to a myriad of reasons, missing data, misinterpreted field reports or inconsistent intelligence reports (Thuraisingham, 2004).

2.5.4 Low observable signal for terrorism and the use of unsupervised methods

To overcome this difficulty data mining for terrorism has instead tried to detect anomalous information, but again anomalous behaviour may not be indicative of terrorist behaviour just behaviour which is outside the norm (Thuraisingham, 2004). Therefore if incorrectly used data mining can exacerbate a problem of trying to identify terrorists, however if used to support investigators in collating information acting as a filter to narrow down a list of suspects, data mining can be effective.

2.6 Discussion

Traditional inteligence gathering methodologies used by the intelligence services include HUMINT, GEOINT, MASINT, SIGINT, TECHINT, FININT and CYBINT and OSINT. While all methodologies are of use in combating terrorism CYBINT and OSINT have received in particular alot of attention with mass intelligence gathering and the use of data mining techniques. The most notable example of this would be systems would be terrorist informatics systems such as PRISM. Terrorism informatics is the use of both data analysis and data mining for the provision of actionable insights to intelligence specialists. Data analysis is very much different to data mining in the context of terrorism informatics, data mining is the extraction of hidden patterns algorithmically from databases, while data analysis involves the use of subject based or pattern based queries to extract information from databases. Subject based queries start from a specific entity of interest (a person, a vehicle) and expand out to related items based on this entity. While pattern based queries uses data-mining or SNA to uncover previously unseen patterns in the data. Both data analysis and data mining are a key components of terrorism informatics, which is the science of management and analysis of data terrorism related information to aid in the investigation and prevention of terrorism.

Terrorism informatics, while its use has only recently been popularized, its use goes back to at least the 1970s. The use of terrorism informatics is reviewed from 3 different perspectives, ranging from West Germany in the 1970s, to Israeli and US uses of terrorism informatics currently. The West Germans were one of the first countries to use data analysis in the investigation of terrorism. The Dragnet system they developed was used to locate members of the Baader Meinhof terrorist group and disrupt their operations. The Dragnet system developed by the West German government was one of the first instances of the use of informatics to the mass surveillance of a population and involved collating personnel information about West German citizens . Then through the recursive use of pattern based queries they were able to identify a cohort of people who fitted the characteristics of members of the Baader Meinhof terrorist organization. Through the targeting of these specific individuals and not through mass targeting of (left leaning) individuals, they were able to successfully pursue the Baader Meinhof organization without the use of a heavy handed counter terrorism response and without falling victim to one of the key tactics of terrorism, provocation (see section 1.4).

The US has seen wide adoption of terrorist informatics particularly since the September the 11th attacks of 2001 and the resulting war on terror. The US has expanded the use of terror informatics from data collection and collation to data mining to extract useful patterns in the detection of terrorists. The US has used a plethora of tools and technologies to data mine large repositories of information held on citizens, the most famous of these being the NSAs PRISM and TSAs ATS. PRISM is the NSAs mass surveillance system and the ATS is the the TIAs flight screening system. While the ATS has had some success most notably in its detection of Rael Al-Banna, the use of PRISM has been more controversial. The use of PRISM is controversial for a number of reasons, these are its constitutionality (Park and Wang, 2013) and equally important, its effectiveness.

The dangers of mass surveillance is if used inappropriately is that it can be used not for its intended purpose (detection of terrorists), but for more nefarious uses such as targeting groups or people with different political beliefs. Cases like this are not just limited to just totalitarian regimes, but there are examples of these techniques being used wrongly in democratic states. Just as importantly is the effectiveness of terrorism informatics. The utility and effectiveness of PRISM has been questioned amongst many commentators, stating that the information provided by PRISM offers nothing over what could be provided by more traditional systems and investigative methods. One of the most difficult problems when applying data mining to terrorism and in particularly in use of classification for predicting someone as a suspect terrorist is the low base rate of terrorism and also the low instances of terrorists compared to normal innocent civilians (class imbalance). This results in a large number of false positives even for very accurate models. This could end up with an analyst drowning in false positives and such a system being useless or worse still exasperating the problem of detecting terrorism.

Israels use of terrorist informatics is one of pragmatism. Israels use of terrorist informatics has seen wide adoption of terrorist informatics and particular the integration of IOT and and terrorist informatics to provide security analysts with large amounts of information from disparate sources in an easy to digest and analyse manner. It has also adopted data mining but it has not used it for problems which may over reach from a technical point of view (for instance the detection of terrorism due to the low base rate problem) its capabilities. Instead Israel has adopted data mining approaches which provide additional information to investigators by being able to extract further information to aid an investigation. An example of such a task would be the Israeli security services use of data mining to classify vehicles or armed individuals as potential threats for further analysis.

Both statistics and machine learning have also been applied to the study of terrorism. One of the first methodologies applied to the study of terrorism was SNA and they have been used to study particular groups. One of the earliest studies being the study of the interaction between the KGB and the IRA undertaken in the 1980s. More recently SNA has been applied to new terrorism groups and cells or components within those groups. Particular incidents that have received a large amount of analysis are the September 11th 2001 attacks on the world trade centre and the Bali bombings of 2004. SNA offers a number of advantages to a researcher, as they allow the true representation of a terrorist network, allowing discovery of key actors within the group, channels of communication and the true structure of the network to be ascertained. This can also be useful from a counter terrorist perspective either during an active investigation or post-hoc. During an active investigation key actors can be possibly found and removed from a network causing disruption to the network.

Post-hoc analysis is useful as it can provide investigators useful information regarding structures within groups. However as new terrorism moves more towards employing stochastic terrorism methods its usefulness may become less important as no formal network exists. Instead only a deeply fragmented network may exist and people who carry out acts of terrorism only use the group as point of inspiration or use technical information disseminated by the terrorist group to plan their attack.

Supervised methods such as survival analysis has been applied to the study of terrorsim through the use of electronic terrorist incident databases to allow the study of how terrorist groups persist. Survival analysis has been applied extensively to study the persistence and longevity of terrorist groups using regression methods. Through the use of survival analysis and data sourced partially from the GTD, researchers (Young and Dugan, 2014) were able to ascertain what attributes contribute to a groups longevity. Unsupervised methods have also seen widespread adoption in the study of terrorism. Clustering has been used to enhance the predictive power to which analysts can ascertain the location of terrorist events. By spatio-behavioural clustering, similar attacks together and using these past locations to help predict future attacks (Townsley et al., 2008),(Brown et al., 2004).

HMMs (Allanach et al., 2004) have also been widely applied to the study of longevity of terrorist groups by being used to model the activity of a terrorist groups, is to use it to model activity of a particular group. Similarly they also been used to model activity of groups and used to model large increases (spurts) in activity of these groups.

2.7 Conclusion

Data mining has evolved as a powerful supplemental to traditional intelligence methods such as sigint and humint and with newer intelligence methods as an enabler such as cyber (CYBINT) and open source (OSINT) intelligence. Data mining has been used to study both its application to counter terrorism as well as its application to the study of terrorism itself. A number of different data mining types (supervised and unsupervised methods) have been applied to both the study of terrorism and counter terrorism. A related field to data mining for the study of terrorism is the field of terrorism informatics, which also encompasses the collation, collection and cleaning of data along with the presentation of the data in a simple to interact with manner. Terrorism informatics systems may be also required to ingest data from many disparate sources, for instance IOT banks of sensors or social media data.

Problems arise to applying data mining approaches to counter terrorism, bad data quality due to a myriad of reasons, missing data, misinterpreted field or inconsistent report. If incorrectly used data mining can cause more problems than it solves to predict terrorism or terrorists, for the following reasons:

- Rare occurrences and low base rates of terrorism make application of structured data mining techniques impossible to use for counter terrorism purposes of classifying terrorists or predicting where and when a terrorist incident.
- Rare occurrence of acts and characteristics of each terrorist act which makes the act unique, makes the application of data mining techniques difficult whether unstructured or structured and can be a waste of time as the signal for terrorism is very low.

US experience with PRISM and other TIA related projects have so far proven that mass surveillance when used in combination with data mining does not work or when it does work can over burden an analyst with the amount of false positives it generates or else can cause legal issues or can be abused and misused. Big data technologies are not a panacea for the underlying problems of low base rates of terrorism and the massive class imbalance in terrorists/non-terrorists (Wang and Novik, 2015a). However data mining, if used as a a support to an analyst in a directed investigation and employed in a wider top down approach to counter terrorism (as the Israeli experience) utilising data mining and auto collection of data has shown to be an effective counter terrorist tool. US experience with mass surveillance technologies like PRISM has thought us that due to the underlying mathematics (Bayes rule and the base rate fallacy) (Schneier, 2015) that accurate prediction of individual terrorists or terrorist incidents is extremely difficult, though detection of changes in behaviour such as an increase in intensity, change in attack vector, geo spatial trends or weapon type is possible and extremely valuable in identifying underlying trends.

Data mining is an effective means to carry out research into terrorism especially after an incident has occurred. Of particular interest is graph theory and the use of unsupervised methods (Hidden Markov Models) to try to understand the behaviour of groups. Data mining has also been successful in the application of supervised techniques particularly variance autoregressive regression and time series analysis (modelling the effectiveness of interventions) along with survival analysis in the analysis of what factors effect the longevity of a terrorist group.

Chapter 3

Electronic terrorist incident databases

3.1 Introduction to terrorism incident databases

Most of the initial research on terrorism was done by utilizing ”small-n” qualitative case studies, with the emergence of large n-scale trans-national database of terrorist events, these large n-scale databases allow ”large-n” quantitative studies. Since the publication of Geddes’s paper, ’How the cases you choose affect the answer you get: Selection Bias in Comparative politics’ (Geddes, 1990) on how selection bias influences your results, particularly in small n studies, effort has been made to guard against selection bias and ’cherry-picking’ particular cases for the purpose of confirming a particular theory or point of view. Geddes’s paper has resulted in a large body of work directed at creating a standard to allow maximum levels of clarity, fidelity and substance in the data (Geddes, 1990). At the same time there has been a realization that ”large-n” studies involving data on political incidents such as terrorism suffer from the same problems that curse ”small-n” studies. Through the use of large freely available databases for the purpose of academic research and through the use of reproducible research delivered through technology like Jupyter notebooks (Kluyver et al., 2016) or git code repositories, selection bias can be identified easier and guarded against. The same techniques of open reproducible research are also particularly useful in a business context and allow sharing and easy testing/proofing of research and work.

Terrorism event databases are standardized analytical record sets which are mostly sourced from media articles. These databases can be joined with other data sources such as world economic forum (WEF) to examine the causes and repercussions of terrorism. Through their existence, terrorism incident databases have been utilized to carry out time series analysis of terrorist incidents to see how they effect such diverse topics as tourism (van Niekerk and Pizam, 2015) and foreign investment. But also they have been used to show the effects of particular interventions such as how the use of metal detectors in airports had an effect on hijacking of planes and also how increasing the toughness of police responses on the level of violence in Ireland.

The five best known terrorism event databases are:

1. International Terrorism: Attributes of Terrorist Events ITERATE.
2. Rand Database of Worldwide Terrorism Incidents ,RDWTI.
3. Global Terrorism Database, GTD.
4. World Incident Tracking System, WITS.
5. Terrorism in Western Europe: Events Data, TWEED.

One of the main benefits with the collection of terrorist incidents in databases is that they make the hitherto application of statistics to the study of terrorism which has seldom been utilized in the past, now possible. (Silke, 2001) carried out a survey of research carried out on terrorism between 1995 and 2000 and (Lum et al., 2006) audit of terrorism literature between 1971 and 2003 both concluded that only a tiny minority of research studies (approximately 3 percent) employed statistical analysis. This was much lower than allied research fields such as criminology. Criminologists have a wealth of official data available to the them including the FBI's UCR (Uniform Crime Reports) and census data which holds the NCVS (National Crime Victimization Survey).

In comparison to the collation of other criminal statistics, the collection of information on terrorist activity is especially difficult. With typical criminal behaviour, a number of formal sources exist for recording this type of information. These include official police statistics, governments, news media (in open democracies) and international organizations. Primary sources submitted by international organizations tasked with the gathering of criminal statistics include the International Police Organization (INTERPOL) (Bresler, 1992), United Nations (UN) criminal activity reports (on Drugs and Crime, 2013) and the World Health organization (WHO) who record homicide rates. Other secondary sources would include in the US, the national crime victimization survey and for an international form of the data the International Crime Victims Survey (ICVS). All these sources however fail to completely capture terrorist incidents or completely omit incidents or fail to capture significant detail. While some governments do collect information on acts of terrorism, such as the United States Department of State and the British Home Office (Office, 2016). However this data can often contain a large amount of bias due to political sensitivity of reporting terrorism. Secondly, while certain countries do produce statistics on terrorism (such as those listed above) they are in the minority, and very few countries produce data around terrorist attacks. This has limited the research of terrorism to the collation of data from secondary sources primarily media sources, rather than on primary sources (military/paramilitary or police incident), however over time the collation of these secondary sources have led to the synthesis of ever increasing more complete and inclusive data sets for the study of terrorism.

Since the 1970's, terrorism databases have began to evolve for the collation of terrorist incidents. The most notable of these is the Pinkerton Global Intelligence Services (PGIS) database (Dugan et al., 2006). This is not only the most well known of these terrorist incident databases, it also has the largest number of events recorded. PGIS database was an initiative started by the PGIS (Sheehan, 2012), when they began to instruct analysts how to recognize terrorist incidents from wire reports or world radio news reports and record them. The PGIS characterizes terrorism as "the threatened or actual use of illegal force to attain a political, economic, religious or social goal" (PGIS). The data collection process excluded insurgency activity (even when terrorist acts carried out) by state sponsored military and paramilitary groups and state sanctioned terrorist activities. This was due to the fact that the databases primary function was to serve as a risk assessment tool for corporate clients. The PGIS was assembled by specially trained analysts who were mainly ex US air force personnel (fivethirtyeight.com, 2015), listening to news wire stories and recording the incidents in the database if they met the PGIS criteria of terrorism

Also in the 1970's ITERATE (International Terrorism: Attributes of Terrorist Events) began to record terrorist incidents and classified the terrorist incidents according to a number of variables including;

1. The date of the terrorist attack.
2. The country the attack occurred in.
3. The objective of the attack.
4. The nature of the terrorist attack carried out.
5. The number of casualties.
6. The identity of the terrorist group.
7. The origin of both terrorists and the victims.
8. A number of negotiation variables.

ITERATE was again created through the analysis of news media. One particularly useful encoding held within the database was the negotiation variables proving highly useful in studies that involve the kidnapping of hostages (Gaibulloev et al., 2012). The ITERATE dataset which was initially worked on by Edward Mickolus (Mickolus and Flemming, 2013). ITERATE has been the most extensively utilized data source for research in terrorism. While an international database , the data only records incidents from 1968-2000. The RAND corporation also collates a international (originally they planned to collect data upto 1998) terrorist dataset, which is created and maintained in conjunction with the Oklahoma Institute for the prevention of terrorism to create an additional dataset from 1998 on (LaFree and Dugan, 2007). The RAND database records transnational terrorism from 1968-2009 and domestic (US) terrorism from 1998-2009 (Sandler, 2013). The ITERATE database has been shown to offer a higher coverage of coded variables than the RAND database and also has a wider scope of coverage of transnational events than the RAND database (Enders and Sandler, 2011). TWEED (Terrorism in Western Europe Event Database) is limited to acts of terrorism carried out in Western Europe and is the only database created and maintained outside the US. TWEED is also limited to to terrorist acts that originated from groups that are based in Western Europe and not imported from elsewhere (Engene, 2007). The TWEED database also differs from the GTD in that it does include acts of state terrorism. The TWEED database is also different from the other databases in that it is sourced from a single source (as compared to the others, which are sourced from the news media), '*Keesings Record of world Events*' (East, 2016).

3.2 The synthesis of the Global Terrorism Databases, the data collection process

University of Maryland through the Global terrorism Database initiative acquired the PGIS data and began not only digitizing these records but also began to qualitatively check the entries against the cases recorded in the RAND and ITERATE databases. Data collection beyond 1997 is carried out by the University of Maryland through sponsorship by the Department of Homeland Security (DHS) (LaFree, 2011). The data collection for the GTD2 is currently conducted by the START (start.umd.edu, 2016b) consortium at the University of Maryland and is led by the Start consortium. The database post 1998 is referred to as GTD2 (LaFree, 2010).The GTD team then assembled the GTD in a similar manner to the PGIS database and through the use of a team of analysis who are fluent in a number of languages including Arabic, Spanish, Mandarin etc. These analysts then began analysing data in open sources of data such as Lexis Nexis (Professional) and Opensource.gov. These were then reviewed and those that were deemed to be terrorist acts were then passed to supervisors for review and then on subsequent review are added to the database, unless they are borderline. In this case they are referred to a criteria council to submit a verdict on whether they should be added to the database or not (LaFree, 2012).

The collection process has now been somewhat revised, the analysts now use an advanced boolean search algorithm to filter the number of articles. The articles are now also sourced through the MetaBase API (LexisNexis) and supplemented with information from Opensource.gov. These articles are then processed using both NLP and machine learning to narrow the number of articles to review by the analyst. A further algorithm is then run on the data to remove equivalent entries. After this a machine learning algorithm based upon a pattern recognition algorithm is then used to decide the likelihood of the news story being related to terrorism (fivethirtyeight.com, 2015). The stories are then analysed by the GTD team and are then sorted manually and coded as described above (Ben Salem and Naouali, 2016). The pattern recognition is based on the clustering method utilizing the Khi^2 distance derived from multiple correspondence analysis. Multiple correspondence analysis is a development of correspondence analysis which allows the investigation of the relationships and affinity (or strength of association) of a number of different categorical variables. Multiple correspondence analysis is achieved by carrying out a regular correspondence analysis on an indicator matrix. The percentage of variance that can be accounted for (by the different components) are then corrected and further adaptation of the distance between points is carried out. This can be explained in the following manner. Taking K number of variables and each variable has J_k number of levels and the total of the levels (J_k) is equal to J. As we have I number of observations, the I x J indicator matrix is expressed as X. Carrying out correspondence analysis on the indicator matrix provides two groups of factor scores, one which corresponds to the rows and the other corresponds to the columns, these factor scores are then scaled in such a manner so that the variance equals their corresponding eigenvalues. The complete total of the table made of the rows and columns described above is expressed as N, from this the probability matrix (Z) can be calculated as $Z = N^{-1}X$. r the vector of row totals of the probablity matrix is given by $r = Zl$ and c is the vector of column totals. The diagonal matrices (for the row and column vectors) are given by Dc=diag(c) and Dr=diag(r). This is achieved through the use of the below singular value decomposition 3.1.

$$D_r^{-1/2}(Z - rc^T)D_c^{-1/2} = P\Delta Q^T \tag{3.1}$$

The principal coordinates of the rows and columns as they pertain to the principal axis can then be derived as 3.2:

$$F = D_r^{-1/2}P\Delta \quad \text{and} \quad G = D_c^{-1/2}Q\Delta \tag{3.2}$$

Where Δ is the diagonal matrix of the singular values and $\Lambda = \Delta^2$ is the matrix of the eigenvalues.

The adapted Khi^2 formula is shown below by the formula 3.3.

$$D^2(x, x') = \frac{1}{p}\sum_{\mu=1}^{\alpha}\frac{(x_i - x'_i)2}{m_{\mu/n}} = \frac{n}{p}\sum_{\mu=1}^{\alpha}\frac{(x_i - x'_i)2}{m_\mu} \tag{3.3}$$

Where:

x and x' are two different observations from the dataset

p is the amount of data dimensions in each row

m_μ is the number of times the modality μ

α is the amount of modalities present in each of the dimensions.

The Khi^2 distance was tested in conjunction with Euclidean distance and the Khi^2 distance measurement was found to be superior, delivering more distinguishable results.

3.3 Measurement issues with using incident terrorist databases

One of the main drivers for the increase in utilization of statistics to study data is the rise of open source data sources on terrorism, which can be downloaded via the internet. However quantitative terrorism involving these open source widely available datasets often suffer from problems, particularly the failure of most studies to utilize control groups. An additional measurement problem that faces researchers using open source data sources is the issue of source type reliability. The support materials used to add an incident will often contain inconsistent details. (Ackerman and Pinson, 2016) have also raised issues related to the legitimacy and authenticity of the open sources used to gather the incidents from. (Ackerman and Pinson, 2016) propose the inclusion of validity and credibility metrics for open sources used to gather the terrorist incident information from. (Ackerman and Pinson, 2016) contest that enabling an unambiguous analyses of the terrorist incident database. Related to this problem is coder reliability which is due to the use of machine learning and different people to encode variables, thus testing the validity of encoding if items in these incident databases is a problem. Other problems associated with the use of terrorist incident databases are that of missing values, though some open source databases have had a large amount of success in overcoming this problem, not by imputing missing data but by carrying out additional directed probes or explorations of publicly available information to limit missing data (Parkin, 2012). The GTD also makes efforts to update incident detail if new information becomes available on historic incidents.

3.4 The coding of the GTD database

The data management system (DMS) employed by the GTD combines the functions of source material management and assessment with incident diagnosis and selection of incident data for inclusion. The subsequent process of incident coding and disseminating the information through the internet as a downloadable file is also managed by the team responsible for the GTD. The tools developed for the GTD data collection makes this process seamless and allows each team to encode all the different data variables relating to an attack. The coding strategy employed by the GTD involves one of the six teams in the generation of the GTD concentrating on one specific area. These include;

1. Where the incident occurred.

2. The group or individual who carried out the attack.

3. The target of the attacks.
4. The weapons utilized in the incident by the attackers.
5. The tactics used by the attackers.
6. The number of people injured and killed and the ramifications of the attack.

This methodology ensures that each piece of information is encoded by a team that is familiar with it, for example the team responsible for identification of the groups responsible for an attack in a specific geographical region, as this group due to working exclusively in this area will have greater knowledge working in this area, familiar with the relationships between different (terrorist) groups, naming conventions, name alternatives or aliases for different groups, factions and groups who have evolved from terrorist groups.

3.4.1 GTD admittance metrics

For inclusion into the GTD an incident is defined as a terrorist act if it meets certain criteria. While the GTD definition of terrorism is similar to the most widely held definitions of terrorism. The GTD data collection teams have imposed strict criteria for inclusion of incidents into the database, these are:

1. The event must have been intentional in nature.
2. The event must include some use or threat of violence.
3. The groups who carried the attacks must be sub-national in nature, terrorism carried out by states are not included.
4. The event must have been carried out with aim of achieving an economic, religious or social aim.
5. There must be some indication that the event was in the pursuit of sending a message to a wider audience than the immediate victims.
6. The event must be considered outside the bounds or norms of rightful or what is deemed to be reasonable actions under the rules of war.

Additional filters which may discount the act from inclusion are if it is part of a wider insurgency, it is a criminal act, it is part of an inter or intra power struggle. Other problems with inclusion into the database of incidents, are the inclusion of the same incident twice. To overcome this the analysts involved in collation of the incidents only record a single incident at a specific location and time, if the location or time of the events are different they are counted as separate incidents.

3.5 The database variables

The database variables included in the database are described below . They can be grouped under the following areas:

1. GTDID (an id) and date, a unique identifier and year month date
2. Extended incident, did the incident last longer than 24 hours.
3. Incident information, a summary of the incident, what incident criteria the incident met to be included, is there a doubt the incident was terrorism and if there is doubt what is the alternative designation of the incident.
4. Was it a component incident of an attack that was an multiple incident. What are the related incidents to the event.
5. The incident location, including country, region, province.
6. Data related to the attack, including attack type, the success of the attack, whether the attack was a suicide attack.
7. Data on the weapons used in the attack, including weapon type.
8. The target or victim type including nationality.
9. Attacker information. The name of the group, number of attackers, number of perpetrators captured. Did the group claim responsibility and how was this claim made.
10. The number of casualties (including number or people killed and injured), the number of US injured and killed. How much property damage occurred.
11. Information regarding kidnapping or hostage taking. This includes the number of people kidnapped, the duration of time the kidnapping lasted, the origin country of the kidnappers, the resolution of the kidnapping, the amount of ransom paid, the ransom note details, the outcome of the event i.e rescue, attempted rescue, release of hostages, etc.
12. Additional notes on the event, containing miscellaneous information on whether coordination of a number of attacks within a specific locale and time. Unusual circumstances such as sudden changes in tactics.

3.6 The use of the GTD in terrorism research

The GTD has seen wide application to research into terrorism, particularly its ability to enable statistical based research on terrorism. The type of research it has enabled has been broad in scope from enabling statistical data visualizations to explain changes in terrorist specific behaviour such as attack vector, regio-specific changes or weapon types.

A number of different techniques have been applied to enable knowledge discovery on the GTD. These range from, data visualizations, statistical analysis to machine learning techniques such as network analysis or supervised learning. To the use of survival analysis to assess the affect of particular interventions on the level of violence in a particular region.

Novel use of graphics have made the analysis and comprehension of terrorism to better enable the understanding of complex patterns of patterns, possible (Wang et al., 2008). By building a visualization tool that addresses the basic tenets of investigative analysis, the five W's of of investigative study, the who, what, why, where and where. Understanding of the GTD has been further enhanced by the creation of web based tools that work in conjunction with the GTD. These allow for the use of interactive visualizations to carry out an exploratory understanding of terrorism data (Lee, 2008).

The benefit of using such a tool is that can provide an initial understanding of terrorism data by providing the analyst with search, aggregation, filter and terrorism feature tables consoles and does not require the user to have detailed understanding of data visualization tools such as tableau (Chabot et al., 2003) or programming tools such as D3 (Bostock, 2012) or ggplot (Wickham, 2016c).

The GTD has also been applied to more subtle types of study particularly the use of statistics or machine learning to investigate whether terrorists act as rational actors and follow rational choice theory.

Rational choice theory is a methodology utilized to both comprehend and to represent or explain social or economic activity. The central tenets of rational choice theory is that the overall behaviour of the group is a result of the behaviour of a persons making their own specific decisions in a rational manner as to why they do something by taking account of their motivations or impetus and purpose or goal. (Hepworth, 2013) used a mix of summary statistics, statistical tests (chi-squared test) to analyse the type of target attacked and regression modelling (negative binomial model) to investigate the relationship between the use of suicide attacks and target types and the lethality of attacks. The analysis found that the attack behaviour was consistent with the beliefs of the groups. For instance it was found that Al Qa'ida were found to target civilians more often than other groups, particularly those they consider to be unbelievers, which is completely consistent with their beliefs and also is necessary to allow.

(LaFree et al., 2009) has used the GTD to analyse the effects of a particular counter terrorist intervention in Northern Ireland from 1969 to 1992 and whether the counter terrorist intervention resulted in a backlash by the terrorist groups, that is they were immediately followed by an increase in risk of future attacks. Only one of the interventions carried out by the British government (operation Motorman) provided strong evidence for a preventative reaction on terrorism. The particular intervention which saw a resulting decrease in terrorist incidents was the military intervention, operation Motorman. Operation Motorman began on 31st of July 1971 in response to the increased levels of violence in Northern Ireland (Edwards, 2011) and was a large scale reinforcement of the British army in Northern Ireland which were deployed to occupy catholic 'no-go' areas in Belfast and Derry which had been seen as a haven for Irish republican organisations. (Neumann, 2003).

The GTD has also allowed the application of survival analysis to be utilized in the study of terrorism for the purpose of discerning what characteristics of a group or underlying aspects of their theatre of operation are correlated with their longevity, (Young and Dugan, 2014). The research also provided evidence for the process of outbidding (see section 1.4), the use of elevated levels of terrorism to garner support from the general populace. Also the more the group carries out attacks the longer the group will survive.

Machine learning has been applied to analysis of the GTD to aid in the application of complex networks to aid in the discovery of communities in multidimensional analysis, i.e. these are collections of nodes which are highly networked through many connections across many dimensions. (Berlingerio et al., 2011) utilized the GTD to identify communities by creating a group to group network based on the criteria that if a terrorist group committed an act within the same country within the same year, with the dimensions of the network being defined as the attacked country. The authors then developed a framework or approach that is used transform data from a multidimensional to a mono-dimensional community and then applied a number of multidimensional algorithms to it based on p scores, which represent the redundancy, which represents the idea that a community based on a single dimension have the tendency to represent a community in other dimensions. This serves as a measure of the redundancy of the connections, with the more dimensions networking a pair of nodes within a network, the higher the level of redundancy.

Other applications of machine learning to the GTD in the prediction of terrorist attacks involved in an attack. Using Weka (Hall et al., 2009) datamining software, (Khorshid et al., 2015) found that support vector machines (SVM's) proved to be the most accurate learners in developing classification models for prediction of terrorist. Weka was utilised to not only perform the modelling but also to carry out pre-processing tasks such as feature selection.

3.7 Discussion

Electronic terrorist incident databases have aided in terrorism research by enabling the use of quantitative research to terrorism and moving the study of terrorism away from small n studies (case studies) to large n quantitative studies. Terrorism incident databases have existed since the 1970's with the establishment of the PGIS. The PGIS was a database developed to aid companies who had to deal with risk management issues regarding terrorism, and was collated from analysis of news wire reports regarding terrorism that were reviewed by former intelligence analysts. However this was not used for terrorism research but for risk management by being able to supply companies and governments information on terrorist incidents. Also in the 1970's the development of the ITERATE database was started. Since then, a number of similar initiatives have followed including the RAND corporations terrorist incident database and the TWEED database. All the database posses different definitions of terrorism and different collection methods. For example the TWEED database only covers terrorism that took place within Western Europe, includes acts of state terrorism and only uses a single source (Keesings Record of world events) (Ravndal, 2016). The GTD, the database used in this study does not include acts that would be considered acts of state terrorism (LaFree and Dugan, 2016). However the GTD is a widely used electronic incident database for research and has the widest coverage in its breadth of incidents covered.

Another issue with the use of the electronic incident terrorist databases is that primary sources of information that are often used in allied fields of research such as criminology (where crime statistics regarding particular types of crime are made available by police or paramilitary organisations) are not available to terrorist researchers, instead incident databases must be sourced from secondary sources such as news reports.

As these incidents are recorded from secondary sources source type reliability issues can occur, to overcome this issue the team responsible for delivering the GTD has developed a system that incorporates both machine learning to identify perspective events, but also expert teams who review the filtered incidents and add them to the GTD if they are valid. The analysis team are trained in a specific area or terrorism type and they also handle the data encoding of the incident. In this way the GTD have addressed the problem of source type reliability through this hybrid system, by its combination of machine learning and expert analysts.

The PGIS has since morphed into the GTD which is used in this study. The benefit of using the GTD for research is that it is available freely as well as the code book and the methodology of collection is freely available. The data for the GTD is sourced from media accounts which is processed using both NLP and machine learning before being reviewed by an expert panel of analysts at the university of Maryland, (fivethirtyeight.com, 2015). Before being coded added to the GTD. The analysis of stories is based upon the use of (K-Means) clustering and MCA. The coding of the data within the GTD includes information relating to the date and time, incident detail, casualties, geographic location, the success of the attack, the intended target, the number of casualties. Other benefits of the GTD is that it is open source, readily available and regularly updated and financially supported by the DHS (department of homeland security). The GTD has been used extensively in terrorism research, the research ranging from examining the effects of counter terrorist interventions, the longevity of groups and what makes them persist and has been used to discover connections between groups.

3.8 Conclusion

While terrorist incident terrorism databases have facilitated large scale n quantitative studies they do suffer from some problems. These include the conceptualization of what and what is not terrorism as well the recording and the encoding of these conceptualizations. Greater transparency is also required around the context and sources of information used in the compilation of terrorist incident databases. This has the effect of increasing confidence a researcher has in utilizing the data. Some open source providers have made great strides in this area particularly START in their synthesis of GTD who always quote citations for each event. Issues also exist around the coding of variables and how coding conflicts if they occur are resolved. The GTD serves as a good model for this, using teams with specific expertise in a particular area to encode those specific variables. Other issues with the synthesis of a terrorist incident database is the inclusion of variables indicating doubt in whether the incident was a terrorist event. For example the WITS database includes a confidence variable around whether the event was a terrorist event. Further issues related to the use of terrorist incident databases for the study of terrorism include the disclosure of those who backed/funded/payed for the research and who controls the information. By allowing full disclosure this allows the appraisal of any potential conflicts of interest that may occur as is in the use of the data held in the original PGIS database. The PGIS for instance discounted state terrorism due to the Pinkerton agency not believing logging these incidents would be of use to their corporate clients. The GTD also does not consider state terrorism, only considering terrorism to be only carried out by non-state actors and have made full disclosure of their sources of funding (due to its partial funding through the DHS).

The GTD has enabled a wide body of research to be undertaken into terrorism. A broad body of research enabled by GTD ranges from the use of survival analysis to investigate the longevity of terrorist groups to the study of specific counter-terrorist interventions.

Chapter 4

Count time series analysis of the GTD

4.1 Introduction

A preliminary understanding of the GTD was gained through examining the data through the use of descriptive statistical visualization (using time series plots, geo-spatial plots, stacked bar plots), dimensional reduction techniques and unsupervised learning techniques and supervised techniques. An insight was gained into the underlying trends around temporal, spatial temporal, attack, weapon type and regio-specific significance (rural/urban divide) of terrorism.

From the various descriptive visualizations it could be seen that post 2000, the Middle East and North Africa (particularly Iraq and Syria), along with South Asia (particularly Afghanistan) as being the pre-eminent regions in terms of deaths due to terrorism. As these were the predominant areas and countries in terms of deaths due to terrorism, they were selected for further analysis using supervised and unsupervised methods. This information is held in Apendix A.

In the preliminary modelling work the number of deaths due to terrorism in Iraq was modelled using count regression to examine the underlying correlation or relationships between a number of explanatory variables. These were

- Major US and NATO strategic actions, i.e. invasion of Iraq, the troop surge in 2007, the pull out, the launch of inherent resolve to combat ISIS.
- Iraqi presidential reign.
- Month Post 1970.

However the count modelling suffered from the models not being correctly specified, with count specific modelling techniques suffering from over dispersion. While linear and robust regression techniques applied to the count data required the data to be log transformed and again the models appeared to be specified incorrectly.

The number of counts (of deaths due to terrorism in Iraq) were also modelled using HMM's which proved to be very useful in terms of time series analysis by being able to ascertain different epochs (regions in time time of high probability of high or low terrorism counts of deaths due to terrorism) of counts of deaths of terrorism. Through the use of a simple transformation these transition state probabilities can be transformed into classes of 'non terror' or terror epochs, making the HMM easier to interpret (see figure A.39). Such a model would be useful for modelling outbreaks of terrorism and would have applications in a number of areas from risk management to enabling counter terrorism/counter insurgency policies. However the models as they were unsupervised required the number of states to be decided before hand or be empirically derived. Therefore the use of these algorithms for detecting algorithms on a country by country (or even the preference of a particular analyst for a particular country) basis would be difficult.

4.2 Outbreaks, syndormes and anomalies in time series count data

Anomaly detection, outbreak detection and medical surveillance methods can act as an intelligent filter to highlight interesting events in terms of count time series data. These filters are necessary to prevent overloading an analyst with information and serve to place in the spotlight the observations that are the most different (and least like) the main population and focus attention on the most surprising or unanticipated results. These anomalies can be interesting or unexpected in two ways, they can be vertically or horizontally uncharacteristic. Vertically uncharacteristic observations would be sudden spikes, while a horizontally uncharacteristic observations would be a stepwise change (a typical stepwise change would be a mean shift) or a ramp up.

A breakout is defined by 2 steady states and an intermediate transition period. Breakout detection is usually accomplished using two means. These are:

- Mean shift. This is where an abrupt change in a what is being assessed by the modelling technique (in the context of this research this would either be deaths due to terrorism or terrorist incidents).

- Ramp up. This is where a slow and steady increase in the variable being studied from one state to another occurs.

Breakout detection algorithms have a wide use and have been applied to everything from monitoring disease outbreaks to emerging trends on social media.

Medical syndromic surveillance methods (such as the Farrington or EARSC algorithm used in this research) are used in syndromic medical surveillance. A syndrome can be defined as a collection of symptoms or conditions that happen concurrently and correlate highly with the presence of a disease or a higher probability of catching a disease. Syndromic surveillance is aimed at detecting changes in time series counts of certain symptoms for the early warning and detection of a possible outbreak of disease. These algorithms are designed to provide an alarm of a possible outbreak not to confirm that an outbreak has occurred or why it has.

Time series outlier detection can be used to detect time series events which appear different from the the overall trend, these may manifest themselves as (positive) sudden spikes or troughs in count data.

The benefits to using these methodologies in the detection of changes in intensity of terrorist activity is that they are highly generalizable as the numbers of tunable parameters are small and the resulting outbreaks, outliers and syndromic outbreaks can be correlated to real world events. Another benefit of using these methods is that they act as a compliment of each other identifying different time-count events. For instance an outlier may be detected in the absence of an outbreak (or syndromic surveillance outbreak) which would indicate it to be an aberration not a step change or a time event which may indicate a more serious sustained outbreak.

Providing this data to an analyst allows him to marry objective criteria with his own expert (and subjective) knowledge as to whether an aberration has occurred or if something far more serious is taking place, for example a break out in terrorism which is symptomatic of major insurgent activity. In this way, using the algorithms in conjunction with each other they can act as an intelligent filter to an analyst only highlighting the 'most interesting' events.

Finally syndromic surveillance and outbreak detection can detect the effect of a particular interventions and if it has a positive effect (in the context of this research this would translate to a mean or gradual shift down in terrorist deaths, or a drop off in outbreaks) in this way it provides an analyst with greater situational awareness about the effectiveness of particular interventions.

4.3 Algorithms used for modelling time series count data and their methodology

Four models were initially trialled, two were anomaly detection algorithms (twitter outbreak detection algorithm SURUS and RAD) and two were outbreak detection algorithms used in disease outbreak monitoring, the EARSC and Farrington algorithm and are classed as medical surveillance algorithms.

4.3.1 Online time series outbreak and anomaly detection algorithm

The underlying algorithm utilized by the twitter outbreak detection algorithm (Vallis et al., 2014b) is E-Divisive with Medians (James et al., 2014) which makes use of energy statistics to figure out when a large departure from the median has occurred. EDM has a number of advantages for outbreak detection, such as, it is robust when anomalies are present in the data. EDM has the following characteristics:

1. EDM as stated above utilizes robust statistical measures, the median and approximates the statistical importance of a breakout through availing of a permutation test.

2. EDM is non-parametric, the underlying distribution of count data does not have to be known.

The twitter outbreak detection algorithm makes use of a weighted $L^2 distance$. The $L^2 distance$ of of X and Y (which are independent random variables) and X' and Y' which are their relevant i.i.d copies, is given by equation 4.1.

$$\varepsilon(X, Y) = 2E|X - Y| - E|X - X'| - |Y - Y'| \tag{4.1}$$

The $L^2 distance$ between the cumulative distribution functions of X and Y which are represented by F and G is given by equation 4.2.

$$2\int_{-\infty}^{\infty} (F(x) - G(x))^2 dx = \varepsilon(X, Y) \tag{4.2}$$

For X, its characteristic function is described by equation 4.3

$$\phi_x(t) = E(exp(iXt)) \tag{4.3}$$

Based on equation 4.3 the energy distance can be written in terms of characteristic functions as equation 4.4

$$\varepsilon(X, Y) = \int_{-\infty}^{\infty} \frac{|\phi_x(t) - \phi_y(t)|^2}{\pi t^2} dt. \tag{4.4}$$

Since the cumulative function similarly to the characteristic function (equation 4.4) specifies a random variable, a class distance measure can be defined as equation ,

$$D(X, Y; \alpha) = \int_{-\infty}^{\infty} |\phi_x(t)^2 - \phi_y(t)|^2 \epsilon\omega(t : \alpha) dt \tag{4.5}$$

Where $\omega(t : \alpha)$ is a weight function described by the indexing parameter α. α is utilized to account for the distance intermediate between the distributions, where $D(X, Y; \alpha) < \infty$

Time series outlier detection can be seen as a complementary methodology to outbreak detection. To carry out outlier detect on our data Robust PCA (RPCA) is used (Zhou et al., 2010). Matrix decomposition has been previously used in the exploratory analysis section, in chapter 4 matrix decomposition through the use of MCA was used to examine the associations between different categorical variables. Upon performing RPCA on a dataset, a decomposition is calculated of the form (equation 4.6, where x is the decomposition, L is the low-rank matrix, S a sparse matrix and E is a dense error matrix)

$$X = L + S + E \tag{4.6}$$

In this analysis, Netflix's RAD (Robust Anomaly Detection) (start.umd.edu, 2016a) is used to identify the least related time series count events within a larger dataset. In the RAD algorithm an array of features is determined for each time series, which calculate different components of the series, including lag correlation (the correlation between the series and a series with a lag of a certain period derived from the same series), the amount of seasonality and spectral entropy (which describes the complexity of a series). PCA is then preformed on the features followed by the application of a number of bivariate outlier detection methods based on highest density regions and α hulls applied to the first two principal components to identify outliers. Again, the benefit to using Netflix's RAD algorithm is its application can be generalized very easily, as very few parameters are required to tune the algorithm.

4.3.2 Surveillance algorithms

In the field of medical statistics univariate time series of count data are readily available methodologies for for surveying outbreak of medically related incidents, these form the basis of (medical) surveillance problems. Two of the most common surveillance methods Farrington's (Farrington et al., 1996) and the EARSC algorithms (Fricker et al., 2008) were applied to the problem of modelling count of deaths due to terrorism to test their suitability for determining outbreaks of terrorist activity (counts of deaths due to terrorism).

The two methodologies used in this analysis available in the surveillance package (Vallis et al., 2014a) were:

- Farrington's Method, the algorithm farrington or farringtonFlexible (surveillance package).
- EARSC Method (EARSC3), the algorithm EARSC3 (surveillance package).

4.3.3 Farrington's Method

Farrington's methodology (Salmon et al., 2016) is used to predict the actual value of counts of incidents y_n using a unique collection of reference values taken from the observed values $y_1..y_{n-1}$, to account for far reaching trends and seasonality, values are only selected from a window of size $2w+1$ at time n for b years in the past. This has the effect of only selecting a set of recent values with comparable properties at a time and can be stated as (equation 4.7) where r is the period from which the observations were taken from. Poisson regression accounting for over dispersion is then used to model the $(2w+1)B$, reference sample, (Robertson and Nelson, 2010).

$$R(w,b) = (\bigcup_{i=1}^{b} \bigcup_{j=-w}^{w} y_{n-i+r+j}) \tag{4.7}$$

From the model a one sided $(1 - k) \cdot 100\%$ prediction interval for y_n can be taken. To account for the skewness in the Poisson distribution a power transformation is utilized to normalize the data. The resulting upper limit of the prediction interval for y_n is given in equation 4.8

$$U_n = \hat{\mu}_n \left\{ 1 + \frac{2}{3} z_{1-k} \cdot \sqrt{\frac{\phi \hat{\mu}_n}{\hat{\mu}_n^2}} \right\}^{3/2} \tag{4.8}$$

where the mean $\hat{\mu}_n = exp(\hat{\alpha} + n\hat{\beta})$,$z_{1-k}$ is the 100(1-k) quantile and resulting in the U_n, upper limit. Once the upper limit is exceeded an alarm is sounded.

4.3.4 EarsC3

The EARS (Early Aberration Detection System) is based on the C_{t0} statistic and described by equation 4.9, this statistic is used to establish the baseline to which observations are compared to. Explaining the EarsC3 algorithm (Stacey et al., 2007), take y_t as the count of incidents at a particular timepoint. C3 utilizes the C2 statistic calculated from timepoint t and a 2 timepoint lag. C2 is calculated as equation 4.10, where $\bar{Y}_3(t)$ is the moving sample mean and is defined by equation 4.11 and $S_3(t)$ the standard deviation is given by equation 4.12.

$$C_{t0} = \frac{(y_{to} - \bar{y}_{to})}{s_{t0}} \tag{4.9}$$

$$C_2(t) = \frac{Y(t) - \bar{Y}_3(t)}{S_3(t)} \tag{4.10}$$

$$\bar{Y}_3(t) = \frac{1}{7} \sum_{i=t-3}^{t-9} Y(i) \tag{4.11}$$

$$S_3^2(t) = \frac{1}{6} \sum \sum_{i=t-3}^{t-9} [Y(i) - \hat{Y}_3(i)]^2 \tag{4.12}$$

The C3 statistic is then calculated from the C2 statistic from time period t to the previous two time periods, the C_3t statistic calculation is shown in equation 4.13. The EARS C3 algorithm is then triggered when $C_3(t) \geq z_{1-\alpha}$, where the C3 statistic is greater than the $(1 - \alpha)$th quantile of the standard normal distribution.

$$C_3(t) = \sum_{i=t}^{t-2} max[0, C_2(i) - 1] \tag{4.13}$$

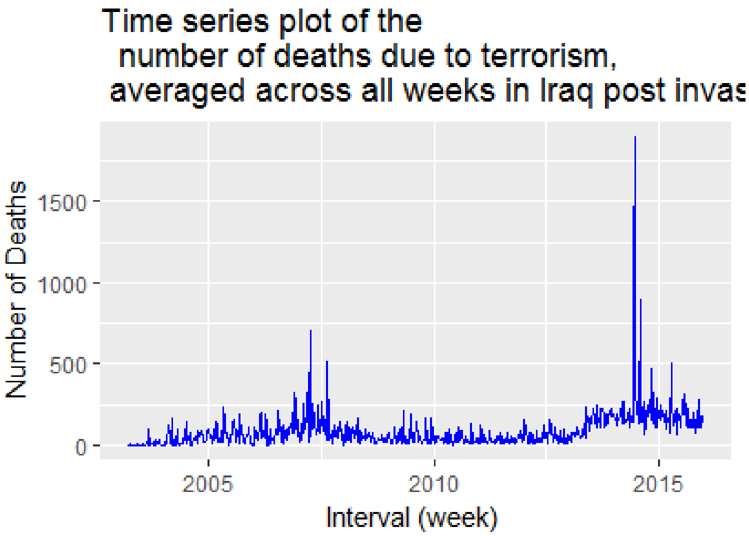

Figure 4.1: Time series plot interval(week) of deaths

4.4 Experimental

The two surveillance algorithms and the outlier detection and methodology were applied to weekly counts of data by country (initially Iraq). Weekly counts were used as this is required for use by the Farrington algorithm and also weekly would seem to be a reasonable time to monitor for outbreaks. To enable this, counts were assembled by grouping counts of deaths by week using dplyr (Wickham and Francois, 2015), which was used previously in section A.3. The workflow to do so can be described so forth. The dataset was assembled by filtering by country, grouping by week, the data (assembled via lubridate by (Grolemund et al., 2011) the death counts due to terrorism by week were calculated. The data was then filtered to only include data after the US led invasion of Iraq in 2003.

4.4.1 Modelling anomalies and outbreaks in Iraqi Death counts due to terrorism

After the dataset was assembled the time series plot was created. This plot is shown in figure 4.1. This plot is similar to those previously seen except the interval being used is weeks as opposed to days (HMM's) or months (count regression) that were previously used.

Initial Application of the Twitter outbreak algorithm was applied to terrorist deaths in Iraq, over death count data in the GTD pertaining to Iraq, from the Post invasion period upto the pullout and post-pullout . Using the twitter breakout detection algorithm, the algorithm was applied using a minimum num-

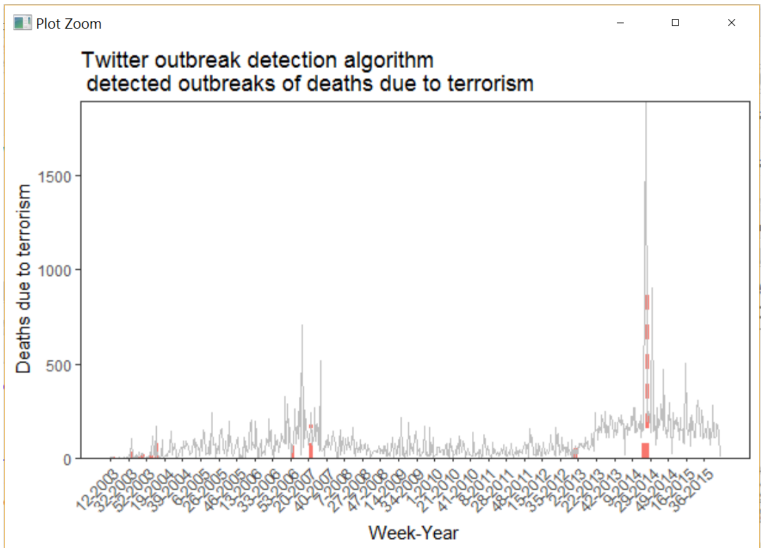

Figure 4.2: Time series plot interval(week) of deaths and application of twitter outbreak detection algorithm. The detected outbreaks are shown indicated as vertical red lines

ber of transitions between change points of four weeks (approximately month). This seemed to be a reasonable time period so as to survey counts and compare to a minimum of four weeks previously. The time series plot over-layed with the detected outbreaks is shown in figure 4.2

The breakout detection algorithm showed a number of breakout of incidents upto the surge, break outs are seen approximately 6 months after the invasion, these breakouts become frequent after this initial period with a break out occurring more frequently upto the surge and after and post-pullout at the time of the major ISIS offensive. It should be noted that no outbreak is detected post surge till the last week in 2012. A table of identified breakouts in Iraq is shown in table C.1.

Outbreaks are detected almost immediately after the invasion, however they seem to be extremely intermittent. Outbreaks are again detected at the end of 2003 and at the start of 2004. Outbreaks again are seen at the start 2007, which would correlate with the surge. An outbreak again is detected midway through 2007 which would correlate with the period of the US troop surge in Iraq. It is interesting to note that no outbreak (corresponding to mean or gradual shift downward trend in violence) is detected in the period post surge (late 2007 to 2008).

The same dataset was examined using the SURUS RAD algorithm. The detected outliers are shown in Figure 4.3. The low rank signal is represented in blue and is a pretty decent approximation of the underlying trend bar the outliers, which are represented as red circles with the size representing the mag-

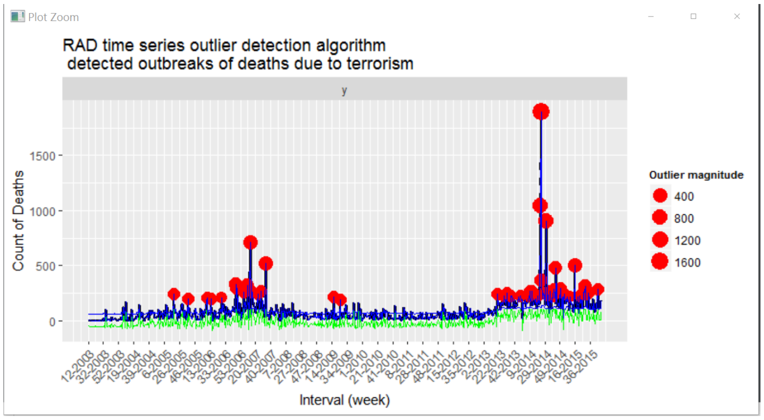

Figure 4.3: Time series plot interval(week) of deaths and application of SURUS RAD algorithm.

nitude of the outlier. The green line represents the underlying noise in the signal. The table detailing the outliers are in table C.2. The magnitude of the outlier is denoted in the S_ Transform column. A high frequency of outliers are detected throughout 2006, this is correlated with the heightening of a bitter sectarian war in Iraq and the coming to the fore of AL-Qa'ida in Iraq as the pre-eminent Sunni armed opposition to the US led invasion (Fearon, 2007). The US troop surge (Ricks, 2009) is also correlated with a large number of outliers in deaths due to terrorism. There is also a high frequency of outliers at the start of 2013. What is particularly interesting is that the number of outliers increases in 2013 to 2014 and decreases in 2015. While there is no outliers detected from mid 2007 to 2009 and mid 2009 to 2014, which would indicate their was no extremes in activity over this period.

4.4.2 The use of syndromic surveillance methods (EARSC3 and Farrington surveillance methods) to analyse Iraqi Death counts due to terrorism

As stated previously both EARSC3 and Farrington's syndromic surveillance algorithm were applied to the Iraq data. In both cases, the only data preparation required was minimal, with the dataframe of counts (of deaths due to terrorism) by week requiring casting as an sts (surveillance time series) object before applying the surveillance algorithm. The output though as with the previous plots (of the SURUS outlier detection algorithm and the twitter outbreak detection algorithm) were enhanced by plotting them with ggplot, as the plotting functionality provided by the base package lacked clarity and the resulting plots were difficult to understand. For the EARSC3 algorithm, a baseline of 18 weeks (4 months roughly) was chosen (which roughly responds to a quadrimestral review period, this was also used for the Farrington algorithm).

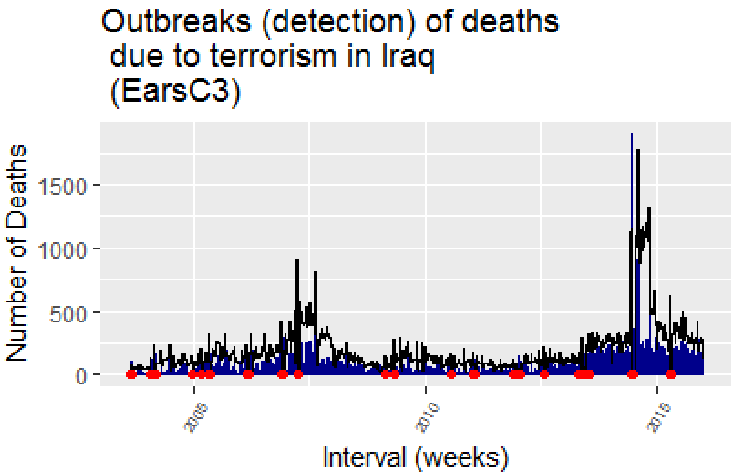

Figure 4.4: Time series plot interval(week) of deaths and application of EARSC3 algorithm

The detected syndromic surveillance outbreaks using the EARSC3 method are shown in figure 4.4. A table showing the alarms raised by algorithm EARSC3 are shown in C. The plot shows the alarms raised (indicating an outbreak) as red circles, the count of deaths is shown as the black line and the blue line represents the threshold. Similar to the behaviour noted with the outlier detection algorithm, post the US troop surge in 2007 there is a relatively small amount of outbreaks detected till 2009, only one in 2010 and after that a steady increase in outbreaks till 2014 and into 2015 when the number of detected outbreaks begins to decrease.

The output of the Farrington algorithm is shown in figure 4.4.2 and the table of outbreaks is shown in table C. The output of the Farrington algorithm is quite similar to that of the EARSC3 algorithm except the algorithm fails to observe the outbreak that was noted at the end of April/ start of May 2013 and also no outbreaks are observed to occur in 2015. The Farrington algorithm detects far less outbreaks than are detected using the EARSC3 methodology, while both use the same review period. The Farrington algorithm was implemented using a B parameter of 0 so as to include only 0 years back in time, i.e the current window so as to allow a direct comparison with the EARSC3 method which only uses data specified in the current window, allowing a like with like comparison of detected outbreaks. This would suggest that the Poisson regression over estimated the baseline values (from which the outbreaks are ascertained) or else the C statistic used in the EARSC3 method is not sensitive enough to noise and is being exceeded too often because of this. This would be due to the methodology not taking account of specific characteristics of count data such as over dispersion (**?**), which is accounted for by the Farrington model.

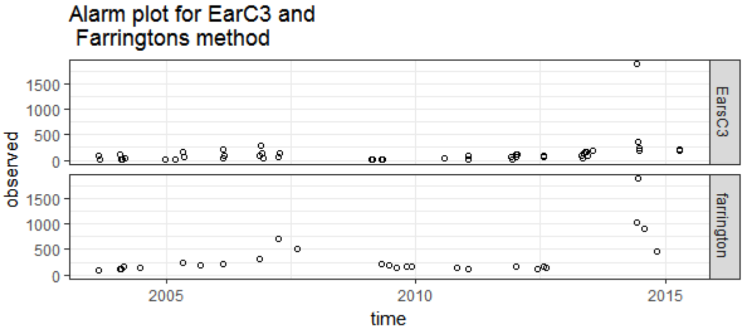

Figure 4.5: Alarm plot for Farringtons and EarsC3 syndromic surveillance methods.

Also as the methodology is an unsupervised learning problem (the detection of aberrations in time series data), multiple methods are often used together and if multiple algorithms are tripped their is stronger evidence for an outbreak. An alarm plot, figure 4.4.2 shows the performance of both algorithms, showing alarms raised by both the EarsC and Farrington method. Figure 4.4.2 shows the output of all models (EarsC, Farrington, SURUS and RAD) overlayed on top of each other this allows the deconstruction of the time series by the comparing the alarms raised by the different methods. For instance if and alarm is raised which detects a syndromic outbreak coincides with an alarm raised by RAD, this can be further classified as an 'outbreak due to an outlier'. While ,if and alarm is raised which detects a syndromic outbreak coincides with an alarm raised by SURUS, this can be further classified as an 'outbreak due mean or gradual shifts'. Its also interesting to note that during the post surge period while a number of outbreaks are detected no outliers or EDM detected outbreaks are detected (see figure **??**).

Comparison to real world events (such as major insurgent offensives) and/or by an expert in the field (of terrorism or counter terrorism) could also be used to judge the effectiveness of the methods in detecting outbreaks of terrorism.

4.5 Surveying multiple countries or regions at the same time using purrr

Using the different methodologies on the Iraq dataset it was possible to detect 'interesting time events' whether they were outbreaks or time series outliers. As previously stated (see section 4.1), the benefit of using these methodologies is there generalizability, as they do not have to be supplemented with additional data or (as is the case with HMM's) require extensive analysis to make sense of. Instead the outbreak and anomaly detection algorithms used in this chapter highlight where 'interesting time events' occur. The models also have the benefit

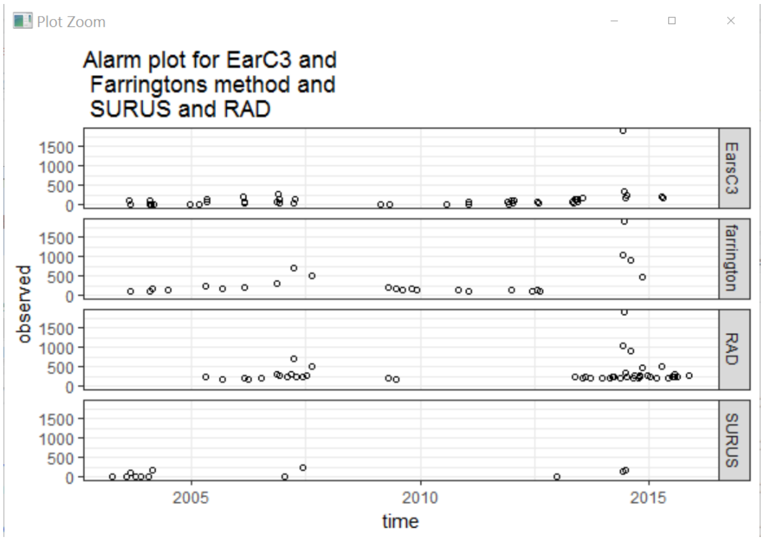

Figure 4.6: Alarm plot for Farringtons, EarsC3, SURUS and RAD method syndromic and aberration detection methods.

in that they can be further generalized by using functional programming to apply them to individual countries or regions. Using this approach, it is possible to carry out a mass survey of regions, locations and highlight outbreaks of terrorism in these areas. This is achieved through the use of purrr package (Wickham, 2016a) in R.

The purrr package (Wickham and Grolemund, 2016) increases R's functional programming capabilities by implementing a set of capabilities for working with functions, vector. Of particular use is the map function which allows instead of looping over a vector, carrying out an operation and saving the results. This minimizes the amount of code that needs to be written and allows the code to be generalized further. In the analysis code map() is used to split a dataframe by country and fit a outlier detection model to each subset of data and extract the output of where the outbreak events occurred. This is achieved through code similar to that shown in listing **??**:

Explaining listing **??** the sbsetmideast dataframe is segmented by country (the time series count data deaths due to terrorism is partitioned by country) using the split function. The map function is then applied to each different partition and location (the location of the outbreaks in the time series) are outputted to a list and a dataframe of countries and the resulting list of outbreaks is outputted. This simple code listing illustrates how generalizable the methodologies are.

Once the model is applied the locations of the locations can be extracted and alarm dataframe for all countries created using very simple functional and procedural code which could be easily placed into a simple function to make

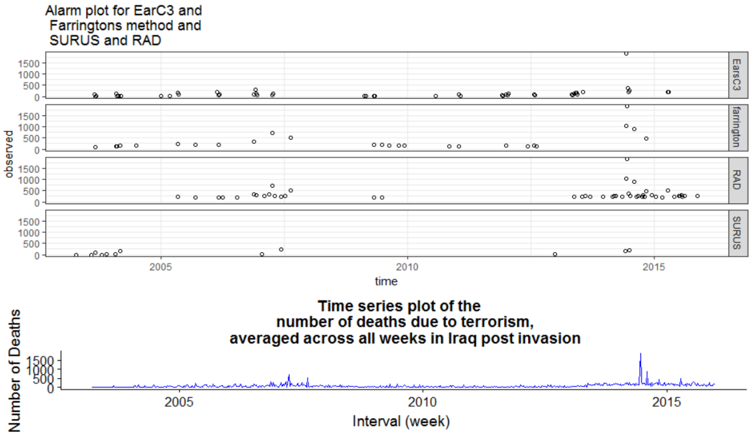

Figure 4.7: Alarm plot for Farringtons, EarsC3, SURUS and RAD method syndromic and aberration detection methods, overlayed with the time series plot.

application simpler. This is achieved through code similar to that shown in listing 4.10 and listing 4.11:

This code simply unlists the dates of the outbreaks by country and creates a dataframe based off these. Then using a for loop, creates a column with a marker if an outbreak is detected. Finally the results are plotted using ggplot faceted by country. This faceted plot is shown in figure 4.5.

4.6 Discussion

Preliminary modelling of count of deaths due to terrorism either using count regression techniques or count time series modelling techniques proved problematic. Regression models suffered from incorrect specification. For the count regression models the models suffered from over dispersion. Also the data held within the GTD was not sufficient on its own and had to be integrated with additional data regarding presidential reigns of US and Iraqi presidents along with coalition troop deployment and withdrawls. HMMs were also experimented with for modelling count (deaths due to terrorism) time series data.

HMMs proved very useful in defining periods or epochs of high terrorism or low terrorism, however they suffered from lack of generalizability as the number of transition states would have to be empirically derived per country or region (or the preferences of an analyst) being analysed.

To overcome this problem a number of different methods were used which are used to model interesting count time series events. Typical events of this type include:

- Mean shifts. These are step changes are characterized by sudden shifts,

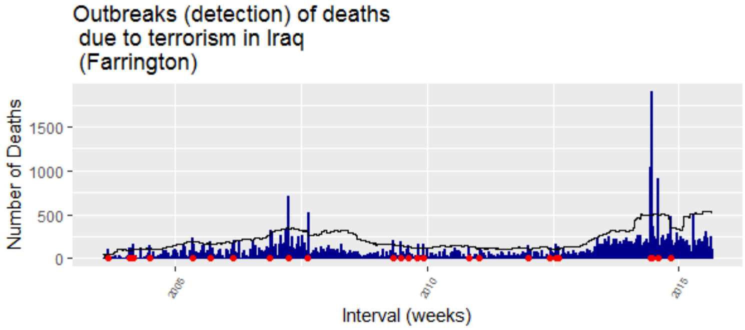

Figure 4.8: Time series plot interval(week) of deaths and application of Farrington algorithm

```
subset_models <- sbsetmideast %>%
  split(
    .$country_txt
  ) %>%
  map(~ breakout(.$sum_kill, min.size=12,
                  method='multi', percent=.1,
                  degree=1, plot=FALSE)) %>%
    map("loc")
```

Figure 4.9: Applying the SURUS model functionally by mapping it across different countries and extracting the dataframe locations of the outbreak. fig:purrrlabel)

which result in large continuous change when compared to previous time series events.

- Gradual shifts or ramp ups which are steady and slow changes between two levels.
- Outliers, these are data points which are far from other observations in the dataset.
- Outbreaks, these are the sustained occurrences (of an indefinite time period) of terrorist events (in this case deaths due to terrorism) over what would normally be expected in a specific region, country or community.

Statistical methods are used to uncover all of the above defined terms. To uncover, mean shifts or gradual shifts the twitter outbreak detection algorithm (SURUS) is used. SURUS uses E divisive with means (EDM), which is a methodology which makes use of energy statistics to detect large departures from the median. A benefit of using this methodology is that it makes no

```
na.pad <- function(x,len){
  x[1:len]
}

makePaddedDataFrame <- function(l,...){
  maxlen <- max(sapply(l,length))
  data.frame(lapply(l,na.pad,len=maxlen),...)
}

locs.df<-subset_models %>% map(unlist) %>% makePaddedDataFrame()

listcountries<-unique(sbsetmideast$country_txt %>% as.character())

pl.holder.dfx<- data.frame()

for(i in listcountries){
  tsti<-sbsetmideast  %>% filter(country_txt==i)
  country_outbreaks<-tsti[locs.df[,i],] %>% filter(!is.na(year))
  tsti$outbreak<-ifelse(tsti$weekstrdate %in% country_outbreaks$weekstrdate,"outbreak","no o
  pl.holder.dfx<-rbind(pl.holder.dfx,tsti)
  }
```

Figure 4.10: Processing the output of functional program to mark the dates across multiple countries the outbreak occurred.

```
ggplot(pl.holder.dfx, aes(x=weekstrdate, y=sum_kill)) +
  geom_line(color = "blue")+geom_text(data=pl.holder.dfx %>% filter(outbreak=="outbreak"),ae
  ggtitle("Time series outbreak (SURUS) plot of the  \n
  number of deaths due to terrorism, \n
  averaged across all weeks in
  Iraq, Syria, Yemen, Jordan post Iraq invasion")+
  xlab("Interval (week) ")+
  ylab("Number of Deaths")
```

Figure 4.11: Plotting the output of the SURUS accross multiple countries using ggplot.

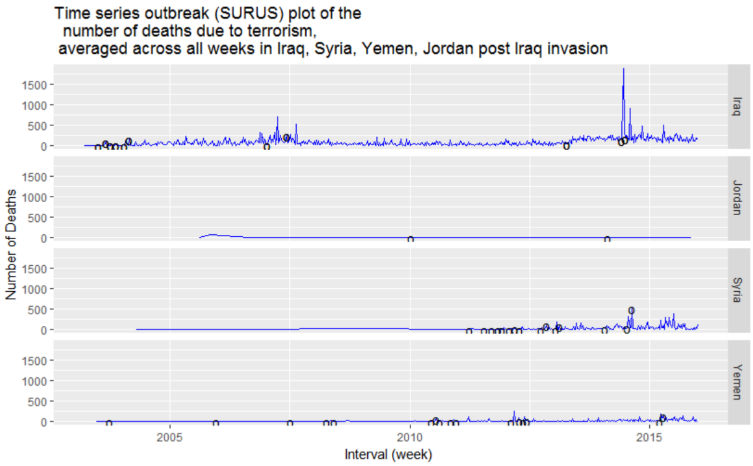

Figure 4.12: Time series plot interval(week) of deaths and application of SURUS algorithm across multiple countries in the mid-east

underlying assumptions about the underlying distributions as it makes use of robust statistics.

To elucidate the presence of outliers, Netflixs RAD (Pylypenko) algorithm is used to detect time series outliers. RAD works in two components, firstly it creates an array of features composing of the time series lag correlation, seasonality and spectral entropy, from this PCA is then performed to detect outliers (these are points that are for removed from the highest density regions). Another benefit of using this type of methodology is that similarly to the EDM method, it is non-parametric.

Lastly syndromic surveillance methods which are primarily used to detect disease outbreaks are applied to the study of outbreaks in terrorism. An outbreak in this sense is the manifestation of terrorist events or deaths which exceed what would be normally expected. An outbreak can last for an indefinite period and for that reason is particularly addressing the research question of interest in this thesis, the detection of changes in intensity associated with terrorism particularly an increase in intensity of terrorism (due to deaths or incidents). As changes in behaviour are ill defined one would not want them to be restricted to certain time series events but be able to detect sustained changes or short lived changes, this makes surveillance methods specifically suited to detecting outbreaks in terrorists deaths. Both Farringtons and the EarsC3 methods were used to monitor outbreaks of terrorism. Both methods function by establishing a baseline and if this is exceeded an outbreak is detected. The C3 method is particularly useful when working with data that does not have alot of historic values as only data is required from the recent past to function (Stacey et al., 2007).

Both methods differ in their approach to outbreak detection. Farringtons method for every time point in the series predicts the number of deaths due to terrorism using a GLM, which is then compared to the actual count of deaths and if this observation exceeds an analyst specified quantile of the prediction interval an alarm is triggered by the algorithm. The EarsC3 method works differently by comparing each point to a threshold calculated from time points in a period in the immediate past (the length of this period is determined by the analyst).

The threshold is then compared to the observed value and as with the Farrington method, if the observed value is greater than an arbitrary (again determined by the analyst) quantile of the prediction interval, an alarm is triggered.

The application of these surveillance methods requires little tuning (with the exception of setting specific alpha levels), are easy to understand and can be generalized easily. Also as a syndromic outbreak is defined not as a mean shift, gradual shift or an outlier but as an observation outside that of above an expected baseline value under normal conditions it is better suited to answering the specific research question is it possible to detect a change in intensity of terrorism (a rather non-specific term), in terms of increases of incidents or deaths due to terrorism.

However when used in conjunction with EDM and time series outlier detection it allows further classification of outbreaks if detected. A further benefit of using these methods is that they are not only highly generalizable when used with specific functional programming paradigms within R they can be applied across many different regions. This makes them particularly useful for the task of surveying many different countries or regions at once for changes in intensity of terrorism in an automated (due to the methods being highly generalizable) fashion. After the outbreaks are detected they can either be assessed by a subject matter expert or additional analysis carried out and combined with the surveillance analysis to give a better understanding of the nature of the outbreaks. By combining with other count time series anomaly detection methods outbreaks can be further classified as mean shifts, gradual shifts or outbreaks. For example EarsC3 algorithm detected an Outbreak from 2014-06-09 to 2014-06-30, during the same period at weeks 25,27,31 and 34 of 2014 the RAD algorithm detected outbreaks occurring, while the SURUS (Kelly and Ahmad, 2015) outbreak detection methods detected a mean shift occurring during week 22 and 26 of 2014. Almost immediately after the surge outbreaks are detected using the and time series anomalies are detected following the US troop surge in 2007.

It should be noted that the compared to the syndromic surveillance methods the EDM based methods fail to detect a number of outbreaks detected by the syndromic surveillance methods. This may be due to the use of robust statistical methods by the SURUS algorithm which do not posses the necessary sensitivity to detect outbreaks detected by the syndromic surveillance methods.

While the methods were easy to generalize and apply across multiple countries or regions, there are weaknesses to their implementation particularly in R, that is the visualization methods associated with them are implemented in Base R graphics and are sparse and not very clear to implement. While the data could be visualized in ggplot, it must be done so by writing custom visualization code and not using a native visualization function within the surveillance package.

The second problem associated with the use of the surveillance methods is the count time series data must be cast as a sts (surveillance time series object) which consists of subsets of slots within in it containing observed, population and state (Höhle, 2007). The sts object is quite complex in structure and its hierarchical nature can be difficult to interact with for analysts more used to working with flatter data structures more common in data analysis such as data frames, data tables, or time series objects (such as Rs ts or xts data structure).

4.7 Conclusion

A number of methods for exploring the detection of 'interesting' count time series events for the purpose of detecting changes in behaviour (particularly intensity) of terrorism. 4 methodologies were tested, that tested for particular time series aberration events; outbreaks, outliers, gradual and mean shifts.

These methods proved easier to both apply and interpret than the methods originally explored in the preliminary analysis. Also unlike the previous methods they did not suffer from problems associated with incorrect specification (in the case of the regression models). The methods were also more easy to generalize with little if any parameters having to be tuned and using functional programming paradigms which exist in R, specifically the purrr package, their application to multiple countries or even administrative regions is relatively easy.

The methods used in this chapter are also complimentary to each other, as when a an outbreak is detected using the syndromic surveillance methods they can be further classified using the other methods as either outliers or mean shifts or gradual shifts.

Chapter 5

Final discussion and conclusion

5.1 Detecting changes in intensity and behaviour in terrorism

Terrorism while often being portrayed as a recent phenomenon, occurrences have been present since antiquity. While having its roots in ancient times, it has recently come to prominence especially since the 1970's. Terrorism is most commonly defined as the use or the threat of violence against the populace, the agents of the government and government with the aim of advancing their political/ideological/religious beliefs.

Data mining has evolved as a powerful additional tool to the traditional intelligence tool-kit of signal intelligence (SIGINT) and human intelligence (HUMINT) and newer intelligence methods such as OSINT or CYBINT (particularly with these fields datamining can be viewed as an enabler of these technologies).

Data mining has been utilized to both study it's application to counter terrorism as well as its application to the study of terrorism itself. A number of different data mining types(supervised and unsupervised methods) have been applied to both the study of terrorism and counter terrorism. An allied field to data mining for the study of terrorism is the field of terrorism informatics, which also includes the collation, collection and cleaning of data along with the presentation of the data in a simple to interact with and understand manner. Terrorism informatics systems may be also required to ingest data from many disparate sources, for instance IOT banks of sensors or social media data. Predicting acts of terrorism or terrorists though is an extremely difficult task from a technical standpoint.

While the prediction of the terrorist acts or classification of a suspect as a terrorist is extremely difficult due to the low signal of the actors involved in terrorism from the general populace (class imbalance) is particularly difficult.

The application of data mining techniques to predicting terrorism due to the low occurrence of terrorists amongst the general populace results in high numbers of false positives (base rate fallacy) (Axelsson, 2000). However data mining and terrorist informatics are particularly useful in aiding the analysis of terrorism.

Electronic terrorist incident databases have assisted terrorism research by enabling the use of quantitative research to terrorism and moving the study of terrorism away from small n qualitative studies (case studies) to large n quantitative studies. Terrorism incident databases have existed since the 1970's with the establishment of the PGIS (Enders et al., 2011) and have morphed into electronic terrorist incident databases such as the GTD, which is open source, freely available, has clearly defined encoding standards, is regularly updated and maintained, utilizes open collection and collation methodologies and is freely disseminated through the internet.

These types of databases are ideal for carrying out terrorism research for academic, economic and governmental motivations.

Being able to detect changes of intensity and particularly the type or classification of change (is an outlier, a mean shift etc.) is an extremely important task from both a political and economic viewpoint. From a social point of view early detection of changes of intensity in terrorism may influence a change in government anti terror policy or a military/para military response. From an economic viewpoint being able to detect the early onset of changes in intensity is equally important. For example being able to detect the early onset of increases in intensity of terrorism would be of particular interest to risk management professionals or insurers who would want to on the detection of changes in intensity of terrorism want to revise insurance costs upwards or withdraw from particular markets till such time as the risk due to terrorism has dissipated. Risk management agents may want to advice clients to withdraw employees from a particular country or region, till increases in intensity of terrorism subside or the situation stabilizes.

In the field of terrorism research being able to detect changes in intensity are also of use especially in the fields of research of counter terrorism intervention analysis or backlash modelling, which are active fields of research which are aimed at discovering the outcomes of particular counter terrorism strategies on terrorism (did they affect an increase, decrease or neutral effect on the intensity of terrorism).

5.2 The study of statistical and data mining methodologies for the detection of changes in intensity of terrorism

Initial investigation of detecting changes in intensity centred on using count regression techniques or HMM's. Both methods suffered from problems, the regression methods suffering from being incorrectly specified and the HMM's suffered from problems of generalizability and understandability. The problems found with the count regression models (Poisson, Quasi-Poisson, Negative Binomial) was that they suffered from incorrect specification. Using linear regression and robust regression on log transformed also suffered from the same problem of incorrect specification. Also the data for the regression models had to be enriched by joining to datasets capturing coalition troop levels, major insurgent activity and US and Iraqi government detail (reign of Leaders of country). Due to the unspecified nature of the HMM's the correct number of transition states to model must be determined empirically. These problems of incorrect specification and lack of generalizability, make their application difficult across multiple countries or regions.

Instead a number of time series count aberration techniques were tried to address the very weaknesses that were found in the preliminary modelling. One of the key learning outcomes from the research was that these techniques are not only highly generalizable, but are more pertinent to the specific research question being addressed (the detection in change in intensity of terrorism or resulting changes in behaviour). Four methods for detecting count time series aberrations were trialled, two syndromic surveillance methods (EarsC3, Farrington's method), EDM (SURUS) and outlier detection (RAD). These methods all detect different type of aberration:

- SURUS, uses EDM and detects mean or gradual shifts.
- RAD, detects using PCA time series outliers.
- Syndromic surveillance methods detect values which exceed a threshold which would normally be expected. If this value is exceeded an outbreak is detected.

These methods while valuable in their own right when used together can provide further classification of a count time-series aberration. Also they can be applied across multiple regions relatively easy using R's functional programming paradigms particularly purrr. These qualities make their use for the detection of 'interesting' count time-series events particularly useful. As far as can be determined none of these methods have been applied to the study of terrorism before and this makes their use for detecting changes in intensity of terrorism novel.

5.3 Future work

Aberration detection methods proved both easy to apply, highly interpretable and are highly generalizable when compared to typical methods for analysing count time series data. They are of particular use for detecting changes in intensity or behaviour in terrorism, with being particularly suited to the early detection of outbreaks or shifts in intensities of terrorism.

Possible future work would be to apply the methodologies to the large n quantitative methods applied to backlash modelling/intervention analysis, which both study the effects of counter terrorist interventions. Backlash modelling is the study of counter terrorist initiatives having the opposite effect to the one desired by causing an increase in terrorism activity (Argomaniz and Vidal-Diez, 2015). While intervention analysis studies how successful a particular counter terrorist initiative was. A possible application of outbreak detection to backlash modelling/intervention analysis is to assess the success of a particular counter terrorist intervention would be to measure the number of outbreaks before and after a particular intervention.

Appendices

Appendix A

Data understanding of the GTD

A.1 Introduction to terrorism incident databases

As part of a larger initiative to apply machine learning techniques to the study of terrorism a initial data understanding was taken as part of the data mining process. CRISP-DM (Chapman et al., 2000) was used which consists of six phases, these are:

1. Business understanding. This introductory step of the analysis is aimed at getting an understanding of the purpose of the project and resulting requisites from a business point of view and transforming these into data mining problem statement. In the context of this project, the data mining problem statement would be to investigate of the application of data mining techniques to the study of terrorism. Through the use of on-line terrorist incident databases, to determine if its possible to determine changes in intensity or behaviour using machine learning. Particularly this is the application of machine learning techniques to investigate underlying spatial, temporal, regional and attack vector type relationships or associations. With the aim of gaining an understanding of the data and the methodologies that are most appropriate to the analysis of the data. Specifically, to see what insights can be gained from these methodologies, particularly the discovery of unusual time series count events. A number of count time series analysis techniques were investigated so as to determine their usefulness in modelling terrorist incident data.

2. Data understanding. This stage begins with a commencing assessment of the data involving a compiling of the data necessary to carry out the analysis after which a number of activities are carried out to gain a familiarity with the data and to establish if data quality issues exist within

the data. Secondly to determine initial insights or to find compelling subpopulations within the data on which to construct a number of hypothesis on veiled or obfuscated information which may exist within the data. In the context of this investigation the data is provided by the GTD. The data quality problems associated with the dataset are highlighted in section 3.3. However due to collection process and data encoding process (detailed in the GTD codebook) employed by the GTD, the data understanding was a relatively easy task.

3. Data preparation. This stage would traditionally deal with assembly of the full final dataset, however as the data was already pre-assembled this stage was not required. Instead various aggregation operations utilizing R's data aggregation capabilities were carried out.

4. Model building. During this stage a number of different modelling techniques are applied. To make the data applicable to the specific data mining techniques, a number of data preparation tasks may be required. Therefore an iterative approach, returning back to the data preparation step was required. In the context of this analysis this involved both joining the GTD to other datasets or aggregating or sub-setting the data. While in the context of modelling or visualizing time series data experimenting with different time intervals would be a typical task undertaken in the analysis. This was achieved by aggregating the data at different time intervals (week, day, month, year).

5. Model evaluation. During this step the model construction and any insights gained from the model are evaluated. Data modelling was carried out in two stages firstly an initial exploratory modelling phase was carried which evaluated some preliminary modelling techniques. After evaluation of the preliminary modelling techniques additional modelling techniques were evaluated which overcame specific deficiencies in the preliminary techniques. Model evaluation was carried out in a number of ways:

   - Checking model specification, this involves the determination of the suitability of a modelling technique to a particular task by determining if the correct iid (independently and identically distributed variables) have been included and if the model meets any underlying assumptions of the model.
   - Checking the accuracy of the model. This involves a determination of the accuracy of a particular modelling technique. This is a specifically difficult task for unsupervised techniques.

6. Model Deployment. During this step the model is deployed in a production setting, however this stage is not applicable in the current setting.

A.2 Data mining techniques used in the preliminary understanding of the GTD database data

Preliminary analysis was used to gain an initial understanding of the GTD through the use of a number of geo-spatial, regio-specific and regio-temporal patterns this was done using a mix of descriptive visualization. Dimension reduction and preliminary modelling were also used to gain an understanding of the GTD but also to see which modelling techniques might applicable to its study. A broad range of different techniques were applied, these techniques included:

1. Data visualization. A number of data visualization techniques were used to gain an initial understanding of evolution of terrorism from a temporal, attack vector type and regional perspective. Unsupervised techniques, both hierarchical and K-Means clustering, were utilized to gain a better understanding of spatio temporal terrorist patterns (particularly regio-specific attack types). These techniques were used in conjunction with both static and interactive visualization to enhance and ease analysis, particularly with the spatio clustering of terrorist incidents using leaflet (Cheng and Xie, 2016). To accomplish this, interactive spatial visualization in conjunction with clustering is used to show the concentration of terrorist incidents in urban areas. This section can be seen as a an extension of descriptive analytics using visualizations. Descriptive analytics can be seen as the most basic form of analytics as they allow alot of data to be condensed into smaller more effective and insightful packets of information. The purpose of the descriptive analytics is to provide more purposeful and useful information in a more digestible form. The descriptive analytics are then visualized using an appropriate visualization technique or provided in tabular format.

   Dimension reduction along with visualization offers a level of analysis over more traditional descriptive analytics and can be seen as providing a level of reduction in complexity of data and offering more insightful information. Dimension reduction allows higher dimension to be summarized in fewer dimensions while maintaining the variation of the original data in these new dimensions. Both the creation of summary statistics and simple data visualization and dimension reduction can be seen as a form of descriptive analytics. Both forms of descriptive analytics allow the analyst to uncover underlying patterns within the data.

2. Dimension reduction techniques. Dimension reduction techniques (in the context of this project this can be considered an unsupervised learning technique as opposed to a data pre-processing step) are used to gain a better understanding of the interaction between multiple data dimensions. With the use of data visualization techniques and the descriptive statistics

used as described above, they are limited to analysis of a small number of dimensions (three to four dimensions). Dimension reduction techniques used in conjunction with interactive data visualization techniques allow the researcher to gain a better insight into the data by allowing the ability to analyse multiple dimensions with many levels and drill deep into the data. This would not be possible with static data visualizations of the data as the visualizations would become crowded and difficult to understand. In this section/appendix correspondence (CA) and multiple correspondence analysis (MCA) (Lê et al., 2008) are utilised to carry out dimension reduction on a number of categorical dimensions (CA and MCA can be seen as analogous to principal component analysis but for categorical data) and the results are then visualized using Plotly (Sievert et al., 2016).

3. A number of preliminary modelling techniques are to used to gain a further understanding of the temporal spatial and attack vector types. A number of supervised methods particularly count regression techniques and Hidden Markov Models (HMM's) for the analysis of time series data of deaths due to terrorism.

A.3 Data mining techniques used in the preliminary understanding of the GTD database data

A number of different visualization techniques were used to explore the temporal and spatio temporal relationships between deaths due to terrorism. These ranged from simple time series plots to stacked bar charts to various types of choropleths or thematic maps. Stacked bar plots were of particular use as they allowed the visual encoding of a third data dimension along with year and deaths due to terrorist incidents including region attack vector and weapon type. The static plots were created with ggplot2 (Wickham, 2009), Hadley Wickham's own implementation of the grammar of graphics (Wilkinson, 2006). Dplyr (Wickham and Francois, 2016), reshape2 (Wickham, 2007) and tidyr (Wickham, 2016b) were used to manipulate and transform the data before visualizing the data. Typical transformations used are aggregation of data and conforming data to the correct shape through wide to long or vice-versa transformations.

Googlevis (which is an R wrapper around google's visualization library), Plotly (Sievert et al., 2016) and leaflet (Cheng and Xie, 2016) were used to create a number of interactive plots. Using leaflet and Googlevis to create interactive choropleths offers a number of advantages over traditional static methods as more information can be stored through the use of interactive layers and captions which can display further information in call out boxes. This allows further number of data dimensions to be visually encoded allowing more subtle understanding of the data.

A.3.1 The evolution of terrorism over time (year and month) by region and world wide

Plotting the monthly totals of deaths due to terrorism by regional and by world wide figures, using these two plots , temporal and regio-temporal relationships emerging. The relationships are shown in figures A.2 and A.1and can be summarized as:

1. Since records have started to be recorded in 1970, deaths due terrorism is on the rise. From the time series world plot we see that during the 1980's there was a sharp rise in terrorism. This flattened out with end of the cold war and appears to be declining up to the end of 1990's, while after September the 11th 2001 there is an increase in terrorism. This increases dramatically with the invasions of Iraq and Afghanistan, it steadily increases upto 2007, before sharply rising after 2010, specifically after the rise of ISIS in Syria and Iraq after 2012.

2. When examining the time series plot by month and by region, a more refined pattern is clear. A spike in terrorism can be seen in the 1980's due to a sharp rise in terrorism in Central America and later in the decade in South America. This fell at the end of the 1980's and start of the 1990's. Terrorism fell across all regions before beginning to rise in the 2000's especially in sub-Saharan Africa, the Middle East and North Africa and south Asia (particularly Afghanistan and Pakistan).

The time series plot for the different regions is shown below. The temporal, regional relationships described above are clearly visible (see figure A.1).

A.3.2 The evolution of terrorism over time (month) by region

While the world wide time series plot (by time interval month) quite clearly shows the rise in terrorism worldwide since 1970 (see figure A.2). Plotting by region shows the specific temporal regional relationships are clearly visible, these are the sharp rise in the 1980's to the start of the 1990's of deaths due to terrorism in central and South America. Again a regional shift of deaths due to terrorism has shifted to the Middle East, this rose sharply after the September 11th attacks and the subsequent invasions of Afghanistan and Pakistan (see figure A.1).

A.3.3 The evolution of terrorism over time (year) by region

The rise in deaths to due to terrorism is more clearly observed (when using interval year) when the global rise in deaths due terrorism by year since 1970, using a year interval instead of month interval (which was used in the previous section A.3.1). The pattern is even more clear when the time series of deaths by

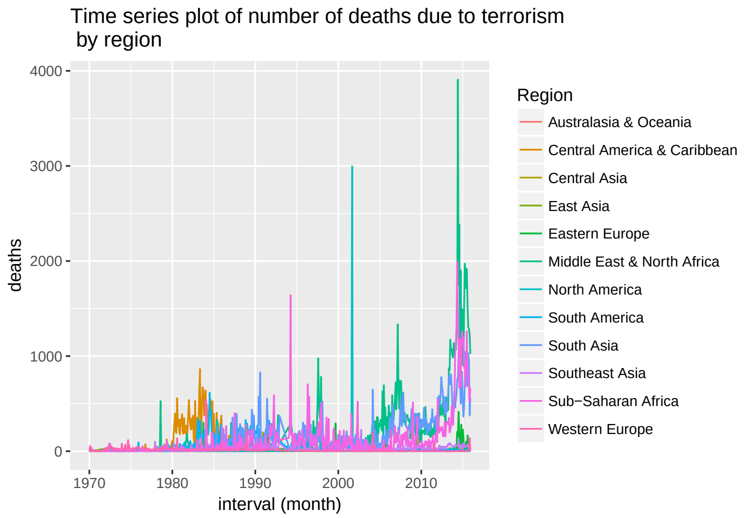

Figure A.1: Time series plot interval (month) of deaths by region

years and deaths by year and region is plotted. Figure A.3 shows the number of deaths globally (not broken out by region). A rise in the late 1970's and 1980's followed by a period of relative stabilization which persisted till the early 2000's followed by a large increase thereafter is seen.

When viewing these deaths by year and region. A clear pattern emerges the rise in number of deaths due to terrorism in Central America in the 1980's which fell sharply towards the end of the cold war. A rise in terrorism is also clearly visible post the September the 11th attacks in North Africa and the Middle East and in South Asia, Figure A.4.

A.3.4 The evolution of terrorism over time (year) by region using stacked bar charts

The regio-temporal shifts are even more evident when visualizing the data as a stacked bar chart, Figure fig:stackbaryear1. The stacked bar chart allows the assessment of the contributing fraction of total deaths for a particular time interval by region. When viewed as a stacked bar plot (with the region represented as portion of the bar). It is clear to see the fall off in terrorism in South America and Central America (in the 1990's), corresponding with a rise in deaths due to terrorism in the Middle East and South Asia. For the last decade the number of deaths that are due to terrorism has been dominated by South Asia, the Middle East and Sub Saharan Africa. Note the gap at 1993 due to the loss of data from the GTD when transferring data from the Pinkerton agency to the GTD.

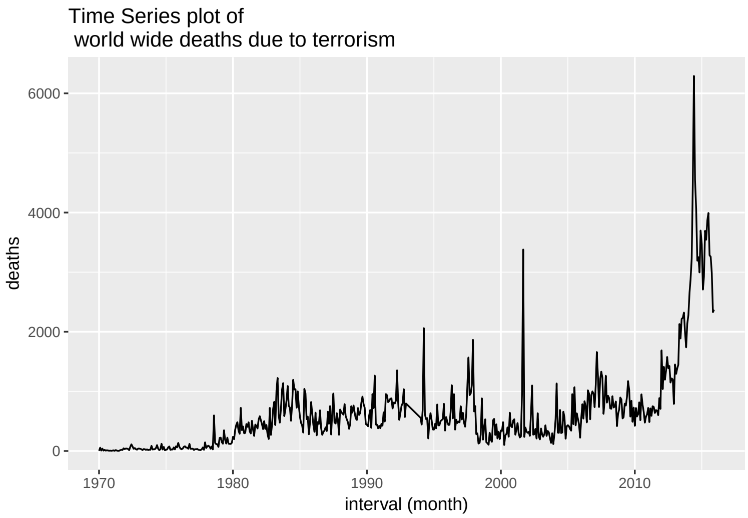

Figure A.2: Time series plot interval(month) of deaths

A.3.5 Visualizing deaths by attack vector and weapon type type

Examining the plot of deaths due to attack vector (armed assault, hijacking, hostage taking etc.) is shown in figure A.6. Again a temporal relationship exists, since the 2000's emergence of bombings and explosives as the prominent attack vector. Also the emergence of hostage taking (barricade incidents and Kidnapping) is observed.

A similar trend is observed when we just look at the Middle east, with the dominating attack vector type being bombings and explosions figure A.7.

When visualizing by attack type by interval year a clear pattern emerges. Up until the early 2000's the dominant weapon type used was firearms, however after this period explosives began to be the pre-eminent attack type (see figure A.8).

A.3.6 Visualizing deaths by country and decade type

Previously the relationship by year and region had shown a regio-temporal shift in death due to terrorism (see section A.3.1). By creating a stacked choropleth (figure A.9), the number of Deaths by decade (1970's, 1980's, 1990's, 2000's, 2010's) are visualized as a stacked choropleth, see figure A.9. Again the regio-temporal shift in deaths can be seen. In the 1970's terrorism was not particularly associated with no one particular region. Britain, Spain, Italy, USA, Nicaragua, Colombia, Philippines and Argentina, show the largest number of deaths due to terrorism. In the 1980's a clear regio-specific pattern can be seem concentrated in Central and South America, particularly in Nicaragua,

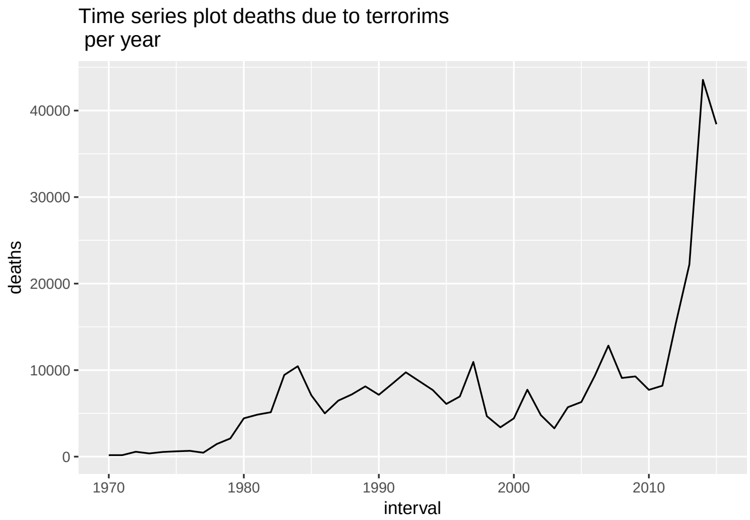

Figure A.3: Time series plot interval(year) of deaths

Guatemala and Colombia.

In the 1990's again no clear regio-specific pattern is seen with deaths due to terrorism being widely dispersed through out the World, though large scale deaths due to terrorism persisting in Colombia and Ecuador, Algeria emerging as a prominent country along with Southern Africa particularly (Angola, Mozambique and South Africa) and Turkey and Russia.

The 2000's sees a large increase in deaths in the US, Iraq, Russia, Afghanistan and Pakistan, but shows large decreases in India and Algeria. As well as an almost total disappearance of terror from Southern Africa and Columbia and Ecuador.

A.3.7 Visualizing deaths by rural urban categorization using K-means and Kernel Density estimation and interactive maps to create heatmaps

In the previous sections(**??**section:viewing-deaths-by-attack-vector-type and A.3.1) it could be quite clearly seen that South Asia (particularly Afghanistan) and the Middle East (particularly Iraq and Syria) were the predominant regions in terms of terrorist incidents and deaths. Iraq was chosen for further analysis as it has been (since the invasion of Iraq in 2003), the country which consistently ranks first in terms of deaths due to terrorism. Similarly, Syria since the Arab spring and the ensuing rebellion against the Assad regime of 2012 ((Dabashi, 2012)) which led to the rise of ISIS and the establishment of a caliphate across certain parts of Iraq and Syria, has also been a pre-eminent country in terms of deaths due to terrorism and incidents.

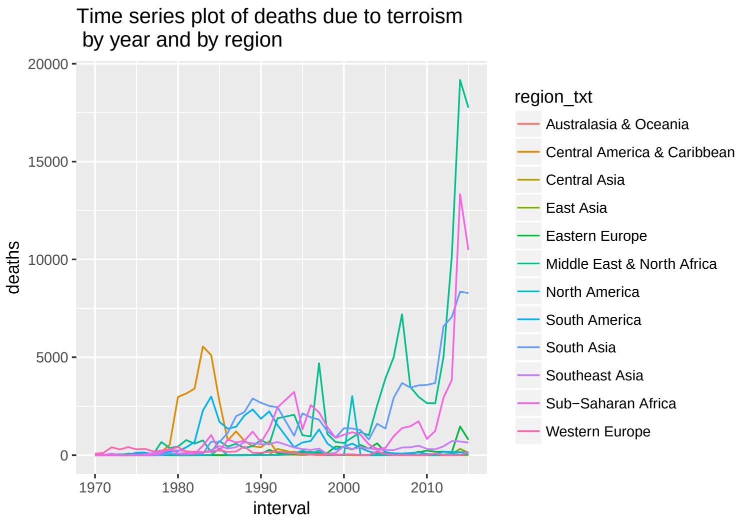

Figure A.4: Time series plot interval(year) of deaths

These countries due to their predominance of terrorist activities, led to their selection for further analysis. Analysis at a finer grain (particularly spatial analysis) as opposed to the previous section which looked at quite broad trends can be carried out by using a combination of K-Means clustering, kernel density estimation and interactive visualization. Iraq and Syria were further examined and one of the experiments undertaken was to see if a rural urban divide (are attacks focussed in one specific area or/and are they associated with rural or urban areas) exists. To do this, only incidents from Iraq and Syria were chosen and the data was also limited to incidents that occurred in Syria and Iraq for 2015. Density estimation creates a fundamental estimate of the probability density function, kernel density estimation creates a smooth kernel function for all data points then aggregating these to get a density estimate. The fundamental kernel estimator is given by the equation A.1.

$$\hat{f}_{kde} = 1/n \sum_{i=1}^{n} K((x - x_i)/h) \tag{A.1}$$

$$MISE(\hat{f}) = E[\int (\hat{f}(x) - f(x))^2 \ dx] \tag{A.2}$$

Where:

Where K is the kernel which is a symmetric, positive function which sums to one. Typical kernel function are uniform triangle, normal and cosine.

Where h is the bandwidth which is a smoothing parameter, large bandwidths create extremely smooth estimates while the converse is true for

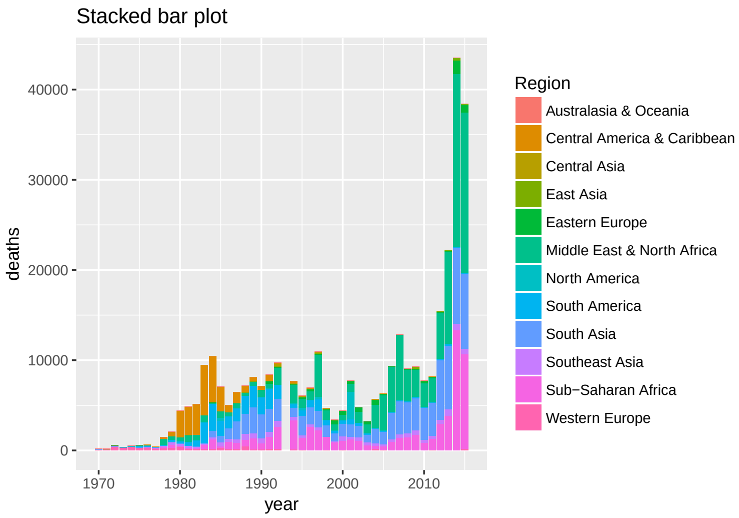

Figure A.5: Stacked bar plot interval(year) of deaths

small bandwidths which produce noisy estimates, as the smoothing tends to impact the estimates much more so than the choice of kernel, it is imperative to determine the optimum bandwidth. The bandwidth is usually chosen by reducing to a minimum the mean integrated square error, given by the equation A.2.

The bkde2d function of the Kernsmooth package is used to calculate the 2d kernel density estimate based upon the WGS84 coordinates of the terrorist incidents, linear binning is then used to create a series of bin counts and a 'Fast Fourier Transform (FFT)' is then utilized to carry out a a series of discrete convolutions. A bivariate Gaussian kernel concentrated at the specific location and the heights of the kernel scaled by the bandwidths are then calculated and aggregated at every datapoint. This allows the creation of heatmaps by overlaying the density estimates on a map created using leaflet. The resulting map which overlays the points on to of the density estimates shows that n Iraq the incidents are largely concentrated around Baghdad while in Syria the incidents appear to be concentrated around the city of Allepo, figure A.12. To gain a greater understanding of the of spatial distribution of incidents K-Means clustering was also used.

K-means clustering aims to divide a number of into k number of clusters and is one of the easiest to understand and popular implementations of clustering. The number of clusters is decided apriori, metrics can be used to guide to the choice of K which minimize inter observation distance within clusters and maximize the distance between clusters. Such metrics include Davis Bouldin index (Davies and Bouldin, 1979). Alternately useful values for K can be chosen on whether the clusters make sense to the analyst based on expert knowledge.

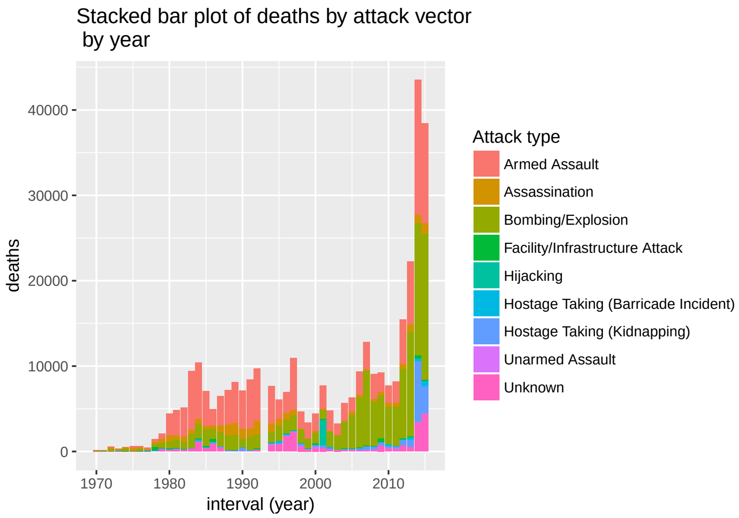

Figure A.6: Stacked bar plot interval(year) of deaths by regions

The algorithm can be summarized using the subsequent stages:

1. K points are arranged into the space characterized by the observations that are being clustered. These points are defined as the first cluster centroids.

2. For all observations they are assigned to the cluster with the nearest centroid.

3. Upon all observations being designated to a particular cluster, the location of the cluster centroids are determined.

4. Stages 2 and 3 are repeated until the centroids no longer change position. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Succinctly put, the K-Means algorithm goal is to minimize an objective function a squared error function given by the formula A.3,

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^j - c_j \right\|^2 \tag{A.3}$$

Where the formula A.3 represents the distance metric between a specific observation and the cluster centre.

$$\|x_i^j - c_j\|^2 \tag{A.4}$$

In the case of clustering a set of longitude/latitude locations are utilized to calculate a distance matrix. This distance matrix is calculated by use of

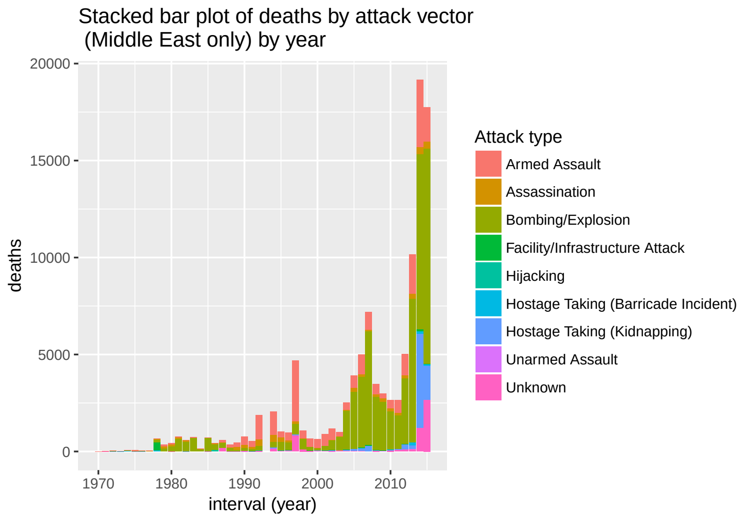

Figure A.7: Stacked bar plot interval(year) of deaths by weapon type in the Middle East and North Africa

the great circle distance which is calculated for all couples or pairings of longitude/latitude (which locate a specific terrorist incident) using the haversine method (equation A.5) where h is haversine (equation A.6). Solving equation A.5 is done by applying the inverse of the haversine equation A.7) is derived.

$$d = rhav^{-1}(h) - 2r\arcsin(\sqrt{h}) \tag{A.5}$$

$$hav(\frac{d}{r}) = hav(\varphi_2 - \varphi_1) + cos(\varphi_2)hav(\lambda_2 - \lambda_2) \tag{A.6}$$

$$d = 2r\arcsin(\sqrt{(}sin^2(\frac{\varphi_2 - \varphi_1}{2}) + cos(\varphi_2)sin^2(\frac{\lambda_2 - \lambda_1}{2})) \tag{A.7}$$

Where:

hav is the haversine function given by $hav(\theta) = sin^2(\frac{\theta}{2}) = \frac{1-cos(\theta)}{2}$

r gives the radius of the sphere.

φ_1, φ_2: represent latitude of observation 1 and latitude of observation 2, in radians, note in fields the rdist.earth function used to calculate the distance matrix in the fields package **??**, carries out a conversion from WGS84 (Misra et al., 1996) to radians.

λ_1, λ_2: represent the longitude of observation 1 and observation of point 2, in radians, note again in fields the rdist.earth function carries out a conversion from WGS84 (Misra et al., 1996) to radians.

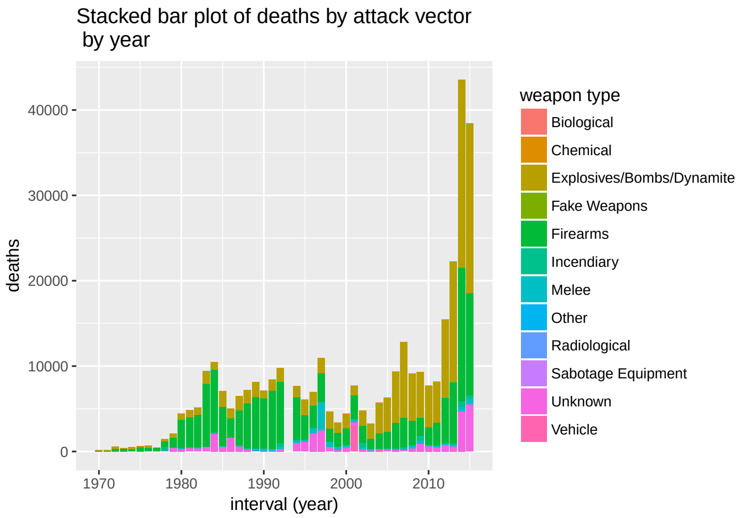

Figure A.8: Stacked bar plot interval(year) of deaths by attack type

Contrasting the analysis from the kernel density estimate based visualization and the clustering exercise provides some interesting insights.

From figure A.12, it can be seen from the overlaying of the binned kernel density of the probability of incidents (estimated from latitude and longitude of the incidents) that the incidents in Iraq are concentrated in Western Iraq specifically around Baghdad and the Sunni triangle (Rand and Heras, 2015) of Tikrit, Ramadi and Baghdad which are the most densely populated areas of Iraq and inhabited mostly by Sunni Muslims, which became the centre for armed resistance Sunni opposition to the 2003 of Iraq (Hashim, 2005), figure A.11. In Syria the incidents are concentrated around Aleppo, Homs, Latakia and Damascus. The clusters relating to these highly urbanized areas are categorized as being tightly grouped together as opposed to the clusters covering the Iraqi Western Desert and the Syrian Eastern Desert which are sparsely populated, figure A.10.

A.4 The use of dimension reduction techniques and clustering in understanding the GTD

While the previous section largely utilised simple descriptive data visualizations or kernel density estimation and cluster analysis to uncover underlying patterns in the data, these visualizations are limited in the number of data dimensions they can encode. To overcome this dimension reduction techniques (particularly Correspondence Analysis (CA) and Multiple Correspondence Analysis (MCA)) are used to reduce the data down from n dimensions to a more lower level while still maintaining the underlying structure in the data. These techniques can also be considered descriptive analytical techniques but are not limited to a low number of dimensions. CA (and MCA) is used in conjunction with D3 (Bostock, 2012) based interactive visualizations, using plotly (Sievert et al., 2016) accessed through the R plotly package to create the bi-plots of the output of the analysis. This is done to aid navigation of the data and to ease understanding of many level categorical variables, which may appear crowded otherwise and difficult to understand. The reason correspondence analysis is so appropriate to the study is the type of data, which is largely categorical in nature.

CA is a multivariate statistical technique which is notionally similar to the more well known dimension reduction technique, principal component analysis (PCA). It is used as a statistical visualization technique for envisaging the associations (note the phrase correspondence comes from the french 'Analyses des Correspondances' where correspondence signifies a "system of associations" which exist between the different items which make up the two sets of data) or relationships between the different levels of a two way cross-tab table (Husson et al., 2010). Two-way contingency tables are utilized in statistics to show how the perceived associations of two attributes (represented as rows and columns) and displayed as cell frequencies of a matrix. An archetypical task within inferential statistics is to ascertain the level of association between levels of one variable and levels of the other variable. CA transforms the data held within rows and columns of a two way contingency table into a lower dimensional space so that the points of the rows and columns in the lower dimensional space are representative of their associations in the table. The mathematics of CA is briefly explained previously in section 3.2.

A.4.1 Correspondence analysis of weapon type by year

CA was used understand better both the relationship year and weapon type. The contingency table of number of deaths by years and weapon type is shown in table B.1.

Correspondence analysis is carried out and visualized in the form of a bi-plot using a standard static plot created using using ggplot, however the static nature of the plot does not aid in their interpretation due to over-crowded nature of visualization, this is illustrated in figure A.13. When Plotly (D3 based interactive visualization tool) is used to create the visualizations, the resulting visualizations can be configured to zoom in and out (the zoomed out visualization is shown in figure A.15) and particular regions (of the visualization) can be highlighted and it is possible to clearly ascertain what associations if any exist (see figure A.14). From the 2000's on explosives bombs and dynamite have been the dominant, while vehicular attacks do not appear to be strongly associated with any particular year or epoch. While the use of Firearms seem to be particularly popular in the 1980's and very strongly associated with 1996.

A.4.2 Correspondence analysis of attack type by year

CA was used to understand better both the temporal relationship between year and attack type. The contingency table of number of deaths by years and weapon type is shown in table B.2. From the Biplot , it can be seen that bombings and explosions are strongly associated with 2005 and 2006. Hostage taking or barricade incidents are strongly associated with 1974 and 1998 infrastructure attacks are strongly associated with 2002 and assassinations.

A.4.3 Multiple Correspondence analysis of terrorist incidents the GTD by year

Multiple correspondence analysis (MCA) is an expanded form of correspondence analysis which accommodates the analysis of the relationships between a number of categorical variables. MCA is also known by a number of synonyms including optimal scaling or appropriate scoring. Methodologically MCA is carried out by carrying out a atypical CA of an indicator matrix (this is a matrix were the values of individual cells take on values or either 0 or 1). Upon carrying out the CA, the percentages of the explained variance are revised and the inter-point distances which result from the CA are adapted to account for this. Multiple correspondence is an extension of correspondence analysis that allows analysis of multiple variables. Instead of analysis counts of deaths, analysis of terrorist incidents is carried out, the data is also filtered to include only incidents from 'North Africa and the Middle East'. The benefit of using MCA is that the relationship between multiple levels of multiple categorical variables can be examined. A number of incites can be gained from examining the bi-plot (see figure A.19), examining the bi-plot it can be seen that:

- There is strong association between 1974 and terrorist attacks on non state militia or terrorists.
- Unarmed assaults are strongly associated with chemical weapons use.
- 2003 are strongly associated with attacks against government, while 2009 shows a strong association with attacks targeting police.
- 1994 is strongly associated with attacks using weapon types sabotage equipment and mellee weapons against facilities and infrastructure.

A.5 Preliminary modelling techniques to gain a temporal understanding of terrorism

Both Poisson regression and HMM's (Hidden Markov Models) are employed to gain a better understanding of temporal nature of terrorism. While the models utilized are relatively simple they are used aid in the understanding of the temporal nature of terrorism and also to ascertain how events not held in the GTD database can be used to explain the events held within. To do this two separate preliminary studies are carried out, these are:

- Poisson regression is used model the count of deaths due to terrorism in Iraq in terms of months and major events over the last the period 1970 to 2015. These events are represented in the time-line chart shown in figure A.20. The aim of carrying out the Poisson regression analysis is to identify what events are affecting the count of deaths due to terrorism.
- HMM's are used to model the number of deaths as a time series and the resulting transition state probabilities delivered by the model are used to determine whether the country is in a state of insurgency (transition probability of being in a period of large numbers of daily deaths is high) or in a state of 'relative calm' (transition probability of being in a period of low numbers of daily deaths is high).

A.6 Poisson, Quasi Poisson, Negative Binomial, Linear and Robust (of log transformed) regression of deaths due to terrorism in Iraq by month

Iraq has had a Tumultuous history and in recent years has seen a massive increase in terrorism since the US invasion to overthrow Saddam Hussein's Ba'athist regime, this is visualized in figure A.20. From the time-line plot, one can see a number of key events in Iraq and when the event, both on the time-line but also in terms of the reign of the different governing bodies in Iraq, US and Great Britain. From the time line one can see the following:

1. The invasion and the US Troop surge took place under Bush regime. The Post surge period and initiation of the draw down of US and NATO troops from Iraq took place.
2. Under the Obama administration, the Post surge period continuing into the drawdown of US troops to the eventual pullout of US and NATO troops took place. The Obama administration was also in place during the rise of ISIS and also the replacement of the Nouri Al Maliki government, which was largely viewed as ineffective (both in terms of effectively managing the country and also the fight against Al Qa'ida), (Simon, 2008) and (Kuoti, 2016).

Over the same period the deaths due to terrorism is shown in figure A.21. From the plot of deaths by years for Iraq not only it is clear that only after the invasion of Iraq, did the number of deaths due to terrorism rise. It's also clear from the bar chart that the outbreak of deaths only occurred increased sharply from 2003 to 2007, peaking during the year of the US troop surge. The preceding years upto 2012 sees a fall off in the number of deaths over this period followed by rise of ISIS (Sekulow and Sekulow, 2015) and the resulting civil war in neighbouring Syria, which sees a large increase.

Poisson regression was first used to model the monthly death totals due to terrorism in Iraq. Poisson regression is used to model count data, counts are positive integers which represent the number of events which have occurred (in this case it represents the number of deaths due to terrorism). The density distribution for a Poisson distribution is given by the following equation (equation A.8).

$$f(y:\mu) = \frac{exp(-\mu)\cdot\mu^{y}}{y!} \tag{A.8}$$

In modelling such events, a Poisson distribution is appropriate as the mean is always greater than 0. Both the mean and variance of the distribution can be shown to be equal. As the variance and the mean are equal anything that affects the mean will also affect the variance. In Poisson regression the logarithm of the dependent variable is linked or connected to a linear function of the explanatory variables in such way that it is given by (equation A.9).

$$log(y) = Intercept + b_1x_1 + b_2x_2...b_nx_n \tag{A.9}$$

Specifically a Poisson regression model, estimates the response variable, the log outcome rate as a linear function of a group of independent variables. Poisson regression makes a number of assumptions about the data. These are:

- The variable you are trying to predict, the response variable is count data and must be positive.
- There are one or more predictor variables that are continuous, ordinal or categorical.
- Independence of observations should be observed in the data, in other words, i.e one observation cannot inform another observation.
- The distribution of counts being modelled should follow a Poisson distribution. This can be observed by checking the dispersion (for over and under dispersion) of the model. Over dispersion is were a large amount of variability is observed in the response variable than is seen in the predicted response using the statistical model. Under dispersion is where there is less variability in the response variable than than in the predicted response.

The data was modelled using US and Iraqi president term data, US interventions, terrorist interventions and month since 1970 (when the precursor to the GTD, the PGIS began). The dataset was created from the GTD by joining it to the CRS (Congressional Research Service) reports for congress on us troop levels (which were used to assign the Surge, Post surge, Pullout and Post pull-out time periods) along with CRS reports (Peters et al., 2016), (Belasco,

2009) and (O'Bryant and Waterhouse, 2007), regarding inherent resolve (US led coalition against ISIS), (Fischer, 2015). US and Iraqi presidential terms were scraped from Wikipedia. The dataset is created by aggregating the monthly death count due to terrorism in Iraq aggregated by Month, Iraqi President, US President, US

The summary of table shows that the model may suffer from a number of a problems. The first parameter of note is the residual deviance of the model. This can be used to calculate a goodness of fit test by carrying out a goodness-of-fit chi-squared test. The null hypothesis of the test is that the model has been specified in a correct manner, however when the test is run, a low probability is obtained approximately 0, therefore rejecting the null hypothesis that the model is correctly specified. On discovering the Poisson model (and the related Quasi-Poisson and Negative binomial to less an extent) were incorrectly specified, a linear model (as well as a robust linear regression model) on log transformed count data was trialled instead (these suffered from their own problems due to the transformation on the data required to meet the specifications of the model). Also a dispersion test is ran and the test clearly shows that the model is strongly over-dispersed.

As stated in the assumptions of Poisson GLMS, for a Poisson GLM the variance and mean are equal, the disperion test (Kleiber and Zeileis, 2008) examines whether this assumption (the data is equidispersed) is true or not against the alternative hypothesis that the variance is equal to the mean A.10.

$$VAR[y] = \mu + \alpha * trafo(\mu) \tag{A.10}$$

Where alpha is greater than 0 the model is said to be over-dispersed. This test confirmed that the .As the model suffered from over dispersion. Quasi-Poisson regression was used, this methodology does not restrict the dispersion parameter *phi* to be 1, but is instead calculated from the data. This has the effect that the coefficient estimates that are calculated from the Poisson model being the same but the inference from the model is changed to take account for over dispersion. Examining the summary of the Quasi-Poisson, we can see that the residual deviance has not changed. The dispersion parameter which was forced to be 1 when fitting a Poisson regression model is now estimated to be 77.30, indicating over-dispersion. The summary information for both models are shown in tables B and B.

Another alternative tried was the negative binomial model which is another methodology for modelling count data (Ver Hoef and Boveng, 2007). This methodology presumes a negative binomial distribution which comes about as a gamma mixture of different Poisson distributions. Its probability density can be estimated using the function A.11:

$$f(y : \mu, \theta) = (\frac{\Gamma(y+\theta)}{\Gamma(\theta) \cdot y!}) \cdot \frac{\mu^y \cdot \theta^\theta}{(\mu+\theta)^{y+\theta}} \tag{A.11}$$

where:

μ = mean

θ = shape parameter

$\Gamma\cdot$ = is the gamma function

The negative binomial model is available in the MASS package (Venables and Ripley, 2002). The negative binomial models makes the assumption that the conditional means and the conditional variances are the same. This inequality is accounted for by evaluating a dispersion parameter which is held constant when using Poisson regression. In the case of the negative binomial model the theta value is estimated to be 1.035329979.

From the data (table B and B) it can be seen that, twice the difference between log likelihoods of the Poisson model and the negative binomial model is observed of 14643 with a difference in degrees of freedom of 22, df=22-21=1. The large chi-squared value estimated from the difference in log likelihood would tend to suggests the negative binomial model, which estimates the dispersion parameter, is a more appropriate choice than the Poisson model. Also from examining the summary model of the Poisson (table B), quasi-Poisson (table B) and negative binomial model (table B) shows that the model with the lowest AIC is the negative binomial model (3011.1). However a goodness of fit test for the negative binomial model performed by comparing the residual deviance compared to the maximum deviance of a perfect model where the predicted values match exactly the observed values, gives a statistically significant result, due to the large residual difference. This would indicate that the model does not fit the data perfectly well (though it would appear to be a better fit than either the Poisson or quasi-Poisson model), so while the model may be wrong it may still be useful. Examining the table of regression coefficients for the negative binomial model (table B), allows us to see which events were correlated (and having a statistically significant effect on the independent variable) with positive or negative effects on the numbers killed per month in Iraq. The post invasion the surge and the ISIS campaign 'soldiers Harvest' have positive regression coefficients which are statistically significant, indicating that these events are correlated with increases in deaths due to terrorism.

Finally a linear model (and a robust linear model) was fitted using log transformed count outcomes and analysed using OLS regression (UCLA).Before carrying out the linear regression model the response variable is log transformed. The model is then created and the specification is checked by examining the fit output. The general linear f-test applied to our model tests the assumption that the fit of the intercept only model is decidedly less than that of the fit of the model trained. From the summary table the calculated F-statistic is 72.3 (on 20 and 255 degrees of freedom) and the resulting p value is $< 2.2e - 6$, suggesting we reject the null hypothesis and accept the alternative hypothesis that the model fit is significantly better than that of the intercept only model. The R^2 value obtained from the model is 0.8501 and the residual standard error which is the positive square root of the mean square error, this is calculated to be 0.9778 on 255 degrees of freedom. The model output is shown in table B.6.To check if our model is specified correctly, a series of diagnostic plots are created to check for the following criteria :

- Linearity. The relationship between the predictor variables and the response is linear.
- Homoescadicity. There should be no relationship between the variation in the predictor variables and the error.
- Independence. The error values are not a caused by any neighbouring values.
- The error term is normally distributed.

To check if the assumptions made by the linear model are met a number of diagnostic plots were used to judge if the assumption of the linear model were met. First a plot of the residuals versus the fitted values was plotted to check the linearity assumption, figure A.24. While no clear patterns are visible from the residuals versus fitted plot and while the residual are not evenly spread around the line but instead they are clustered together around the midpoint and either extremes of the horizontal line. However a u-shape indicative of a non linear relationship is not observed.

Secondly a Q-Q plot is plotted to see if the residuals are normally distributed. A Q-Q plot is a plot used to check for normality, (see figure A.23) . By plotting the theoretical quantiles (for a normal distribution) against the actual one can see how well the residuals approximate the normal distribution. While very few points deviate from the line very much. Thirdly a scale location plot (figure A.25) is created to check for homoescadicity, i.e. that the residuals are spread out evenly across the different values of the predictors (Neter, 1996). While the spread of residuals do not spread out as much along the x-axis the horizontal red line is fairly straight and the residuals do appear to be randomly deposited. No Clear pattern is observed in the data and a sloped line either up or down indicating heteroescadicity is not observed. However at lower observed values there appears to be greater variation in our residuals, which may indicate heteroescadicity. Finally a residual versus leverage plot is run to check if any influential cases which may be present in the data are affecting the fit of our model and as such are extremely influential (Sall, 1990). If these points exist they are located in the upper and lower right hand corners. Cooks distance (Chatterjee and Hadi, 2009) which is used to check for the presence of outliers is denoted in Figure A.22. From the leverage plot it would appear that some of the observations did posses a high amount of leverage which could effect our regression estimates. Instead of removing the points, a regression method was used which could cope better with any violations of the assumptions of the linear model, robust regression B.

A robust regression analysis (Rousseeuw and Leroy, 2005) was then run as a final step in case any of the the linear model was correctly specified, (which from the diagnostic plots there did appear to be some concerns regarding the presence of outliers). The model coefficients are shown in table B.

Robust regression methods are created to be not overly sensitive by violations of assumptions by the underlying data-generating process (i,e outliers). The residual standard error (0.4984) on 255 degrees of freedom and the weighted R^2 value of 0.9590327. A robust F-test is computed for each predictor (to show the significance of each predictor) in the robust regression model and the output is shown in table B.

A robust F-Test is then performed on each of the coefficients of the model to test the importance of the predictor variables. The values are listed in table B. From the table we can see that the variables that are significant level (at p 0.05) are from Period 1 (which refers to periods of major US activity in Iraq) the surge and the post invasion period. Period 2 which (relates to periods of major activity by Insurgent groups) is to ISIS soldiers harvest campaign which was launched in August 2013 by ISIS. From the coefficients in table B, we can see that the US troop surge of 2007 is correlated with a major increase in deaths due to terrorism as is post invasion period as the ISIS soldiers harvest campaign, while the pullout is correlated with a major drop in deaths due to terrorism. This is the same as what has been seen previously, with the negative binomial model. What is counter intuitive is that the periods of AL Maliki's government instituting sectarian policies and the founding of ISI (Al Qa'ida in Iraq) are correlated with a drop in terrorism (negative coefficients).

The effect of different British counter terrorist initiatives has been evaluated by LaFree, (LaFree et al., 2009). Using the GTD the researchers investigated the effects of different initiatives by the British government throughout the troubles to stop terrorist activity. Six different prominent counter terrorist strategies utilized by the British government from 1969 to 1992 were evaluated using statistical tests to evaluate whether upon applying the intervention did the future risk of attacks rise, reduce or remain the same. Only one of the major counter terrorist military interventions was found to have a negative effect on terrorist activity (Operation Motorman). Similarly from the preliminary regression studies the effects of the specific interventions can be examined, major offensives by insurgent groups are correlated (Soldiers harvest) with increase in deaths due to terrorism as are some offensives by military powers in support of the civilian government (Post invasion occupation and the Surge in troops in 2007) are also accompanied by increases in deaths due to terrorism. However the period post the surge is identified using robust regression as having a statistically significant negative effect on the number of deaths due to terrorism. From the analysis of the regression analysis (both count regression and regression using log transformed counts), both major military interventions by both sides would appear to be highly correlated with deaths due to terrorism.

A.7 Using Hidden Markov Models to analyse the number of daily death due to terrorism

Hidden Markov Models (HMM's) are a common machine learning approach for modelling time series data. They have seen wide usage and have application in everything from robotics (Ladd et al., 2005), speech recognition (Gales, 1998), genetics for sequence modelling (Sonnhammer et al., 1998) and financial applications for modelling financial markets (Gales, 1998) and (Park et al., 2009). They can be be viewed as a certain type of dependent mixture model were $X^(t)$ the process subordinate to the state and $C^(t)$ the unobserved parameter process which meets the Markov process. A process is said to to meet the Markov property if conditioning on the previous states in a process up to a time t is the same as conditioning only upto the last value of C_t (the transition state probabilities is shown in equation A.12).

$$P_r(C_t + 1|C_t, ..., C_t) = Pr(C_{t+1}|C_t) \tag{A.12}$$

In a HMM the state dependent process (represented by $X^(t)$) the distribution of X_t is only dependent upon the on the present state of C_t and is not related to earlier or preceding states. These relationships can be summarized as (the transition state probabilities are given by equations A.13,A.14).

$$Pr(C_t|C^{(t-1)}) = Pr(C_t|C_{t-1}) \tag{A.13}$$

$$Pr(X_t|X^{(t-1)}, C^{(t)}) = Pr(X_t|C_t) \tag{A.14}$$

However the state at time t of the hidden state is dependent upon previous states. The probability mass function of X_t of the Markov chain being in state i at time t is then given by the formula A.15

$$p_i(x) = Pr(X_t = x|C_t = t) \tag{A.15}$$

Where p_i represents the probability of Xt of the Markov chain (extracted from the HMM) in state i at time t.

In the exploratory analysis HMM's were built utilizing the depmix S4 package (Visser et al., 2012). Depmix S4 is an R package and infrastructure which allows for the specification and building of HMM's along with the decoding of the models. Depmix S4 carries out optimization either using using Expectation maximization or alternately through Rdonlp2, which is an R interface to the DONLP2 (Do nonlinear Programming) (Spellucci) software which is used to solve non linear programming problems. When using depmixS4 a model is firstly specified using the depmixS4 function. The Depmix S4 package also allows for a number of distributions to be utilised for the state depending process, those incorporated into the package include binomial, gamma and normal distributions. Depmix S4 offers the user a number of advantages to the analyst including allowing them to utilize inclusion of covariates in state and state dependent process and also the ability to produce synthesized data from the model.

The daily death counts due to terrorism was modelled and a 2 state HMM was trained on the data to predict to different insurgency epochs or regimes (one of relative calm and small numbers of deaths due to terrorism and the second an epoch were there is a high number of deaths due to terrorism). The forward backward algorithm is then used to determine the probability of being in a particular state at any particular moment in time (Austin et al., 1991). The forward backward algorithm is an inference algorithm used to determine the posterior marginals of all the hidden state based upon a succession of observations.

Regime detection is often used in financial time series modelling to aid in deciding a particular strategy to use. These models are used to detect whether markets are in particular periods such as bull or bear (a bull market is where prices are expected to rise and bear markets are periods or epochs which are characterized by pessimism and falling prices are expected) periods. HMMs are often used to detect such market periods or regimes, (Alsema, b) and (Bae et al., 2014). The use of HMM's for regime detection would be considered a form of unsupervised learning. Regime detection of epochs of high intensity of terrorism and low terrorism would be of particular use when either investigating or studying terrorism. For instance, regime detection would be of particular use when deciding which particular anti-terrorist strategy to deploy, does the current situation require deployment of troops or is a draw down of troops required as they are no longer needed to bolster the civilian powers. To this end a HMM was fitted to the terrorism death count data on a dataset derived from GTD pertaining to Iraq. A HMM was specified (using a Gaussian distribution) and fitted on the daily death counts. Two states are specified in the model. The transition state probabilities are then extracted from the model and joined to the death count data and the resulting dataframe is then transformed from wide to long format. Creating a faceted plot of the number of deaths, transition state PS1.

The transition state probabilities are plotted faceted against the daily death counts due to terrorism (see figure A.26). From figure A.26, it can be seen that state 1 (PS1) that periods with a high probability of being in PS1 are in a regime or epoch of low terrorism. While periods with a high probability of being in state 2 (PS2) refers to a regime or epoch of high terrorism. Looking at particular time periods the plot become more useful. Examining the plot for 2003 - 2004 (figure A.26), the probability being in a state was very low, up until July of that year when the probability of it being in this state was high, after that time the probability of PS2 was rare but appeared to become more frequent as the year continued. Looking at 2004-2005 (figure A.27), as the year went on the frequency of a high probability of PS2 (relating to a terror state) increases. Looking at 2005-2006 (figure A.28), as the year went on the frequency of a high probability of PS2 (relating to a terror state)again increases, again this trend continues into 2006-2007 (figure A.29). During the period of the US troop surge (2007 to 2008, figure A.30 and A.31) again the frequency of PS2 again increases. However from 2009 to 2011 (figures A.32, A.33 and A.34), the frequency of being in PS2 becomes less and less. This begins to change in August 2012, when the frequency of being in PS2 begins to increase again. This trend starts to increase dramatically in 2013 and by the end of 2013 and 2014 (figures A.35 and A.36) the frequency of being in PS2 is very frequent.

Using the state probabilities its possible to discern change points, where you are seeing a major shift from one observed state to another. The methodology shows great potential a a means of identifying outbreaks of terror and possible sudden shifts in (temporal) regime, from a state of 'low terror' or relative calm to a state of 'high terror'. However the methodology also suffers from some difficulties. The idea of a transition state probability can be a difficult for a non statistical worker to understand. Also the method suffers from being hard to generalize as the number of transition states must be specified, this number of states must be derived empirically and the number of states must be determined by what makes sense? These points would make wide-scale largely automated deployment of these models across a number of regions difficult.

A.8 Discussion

The aim of this appendix which was a preliminary examination of the data was two fold:

- Firstly to gain a understanding of underlying trends in the data and the relationships between the different data dimensions.
- Secondly to ascertain which modelling techniques may be appropriate to the data and which would need further consideration.

A mix of descriptive data visualizations, regio and regio temporal visualizations, dimension reduction and the resulting visualization of the dimensions. Preliminary modelling of count of deaths due to terrorism (country specific to Iraq) was carried out. Time series plots of Descriptive visualizations of temporal relationships between different weapon, attack vector and regions were carried out using stacked bar charts, this allowed visual encodings of 3 data dimensions. From the resulting data visualization it was clear to see the resulting regio-temporal relationships and attack and weapon type relationships. A clear regio-specific trend can be seen, where Western Europe was the predominant region in the 1970's, Central America and South America in the 1980's, Africa in the 1990's and the 2000's dominated by South Asia and the Middle East.

This view was supplemented by creating a faceted regio-temporal sums of deaths due to terrorism by decade and the individual countries which are the foremost countries are easy to distinguish. There is also a clear temporal relationship between weapon type vector(the foremost attack vectors now being armed assault and bombings and explosions) and attack type (bombings and explosions).

The understanding gained via the various visualization techniques described previously was further enhanced using dimension reduction techniques (CA and MCA) and interactions between multiple categorical variables were evaluated. Specifically certain associations can be seen that are not discernible from the descriptive data visualizations. The dimension reduction techniques when used with interactive visualization techniques based upon the D3 framework made this possible as it allowed analysis to be focussed very deeply onto the data and uncover underlying associations between the different levels of the categorical data being examined. The interactive visualization packages plotly package proved adept at this. Using these techniques it was possible to ascertain particularly relationships between attack types and weapons, specifically chemical weapons and unarmed assaults and time periods with attack types, 2009 and attacks against the police while 2003 was strongly associated with attacks against governments. These techniques were useful at showing a change in behaviour in terrorism but did not give any particular insight into uncovering a change in intensity in terrorism.

Cluster analysis was also carried out on terrorist incidents in Iraq and Syria and the incidents (using both kernel density estimation and k-means clustering) appeared to be largely clustered in urban areas. Iraq and Syria were chosen as the descriptive visualization had shown these countries along with Afghanistan to be predominant in terms of terrorist activity. Cluster analysis (and kernel density estimation) was able to identify (by overlaying the the clusters of the incidents on an interactive map) a rural-urban divide in terms of where incidents occur. This method suffered from a number of a problems, as an unsupervised method, the number of clusters must be empirically derived so its country wide application to look for a global trend of rural urban divide in terms of where terrorist incidents occur would be difficult to do as the optimal k value would have to be ascertained per country. Kernel density estimation used to create thematic maps showed similar information to the spatial clustering, and again illustrated that attacks in Iraq and Syria were largely concentrated in urban areas.

After carrying out the descriptive statistical visualizations, cluster analysis and dimension reduction techniques a number of preliminary modelling techniques were applied to countries in specific regions based on their pre-eminence in terms of terrorism activity (both death count and incident count). These techniques and the choice of preliminary modelling technique used and their use case was informed by the literature review of current research into terrorism research using electronic event databases concerning terrorism. Either using terrorist data to predict or forecast future trends in incidents or else the use of the GTD to help explain the effects of specific strategies. The aim of the preliminary studies was also to see how universal the application of a technique would be and what reach it would have (could it identify changes in intensity in terrorism and the factors causing them?), how easy would the technique would be to apply across multiple regions? Another aim of the analysis was to challenge (by confirming or denying) existing conceptions held about terrorism.

Examples of such conceptions would be what was the effect of the US surge in Iraq and what, if any were the underlying reasons for ISIS's rapid rise, was it a inaction by the West?, what were the effects of the interventions of the or could the GTD even answer this question. What were the limitations of the data or the type of data held in the GTD and how easy it to augment the GTD with other data sources?

Preliminary modelling of count data was done with both regression (and alternative methodologies) analysis on Iraq's counts of deaths due to terrorism to explain the counts of deaths and the interdependencies with particular US, Iraqi or insurgent actions along with the months (since 1970, when records of terrorism began in the PGIS, the precusros of the GTD).

HMM's were used to carry out a preliminary time series analysis of counts of deaths due to terrorism In Iraq. From the analysis of the emitted state probabilities and the frequency of the states it is possible to see particular epoch and a change of epochs from one of relatively low levels of terrorism to another of a high levels of terrorism. The aim of the modelling was two fold to evaluate how effective the modelling tool. Both modelling count data using regression and using using HMM's proved problematic.

When modelling count data, the count regression models were found to suffer from issues with them being not specified correctly or else assumptions regarding the modelling technique can be violated. When using Poisson regression (and its related techniques of quasi-Poisson and negative binomial regression) to model the count data, the models were found to suffer from being either over dispersed or having a poor goodness of fit. Transformations can be particularly useful when trying to increase the ability of a particular modelling technique to be able to fit to the data before applying simple multivariate linear regression or by steadying variance. However log transforms does suffer from a number of known shortcomings, including its inability to deal with zero observations requiring the addition of one to a value before it is log transformed and these methods have performed poorly in the past when compared to using models established using Poisson, Quasi-Poisson or negative binomial distributions (Ohara and Kotze, 2010). Ascertaining which model performed best though also proved difficult as different 'goodness of measurements' are used across lm's and glm's.

While HMM's proved useful at identifying different epochs (an epoch of high numbers of deaths due to terrorism, or an epoch of low deaths due to terrorism) of terrorist activity and the onset of these epochs, it did suffer from problems of generalizability and understandability. Such a model would have undoubted usefulness from a number of standpoints, such a model would be capable informing risk management and reinsurance decisions (from an insurance underwriting perspective) or from a governmental point of view identifying when to deploy or mobilise armed forces. However these models due to their unsupervised nature would also prove problematic to deploy globally on a country wide basis as much like clustering the number of transition states would have to be empirically derived. However models that would be able to ascertain similar changes in epoch or outbreaks in terrorist activity would be useful in serving as an early warning system or alarm system to analysts in modelling terrorism. The problems of understandability is due to the output being difficult to understand for non statistically trained analyst as the output which is a transition state probability is difficult to understand. This problem can be addressed by turning the transition probability to a class if it exceeds a probability threshold (this is commonly set to 50% in machine learning, but could be tuned), if this threshold is exceed the specific period as class as being in that transition state. The facet class /count deaths plots are shown in figure and A.37,A.38 and A.39. It is quite clear from these epoch plots the effect of the 2007 surge. In 2006 there for a large part of the year the predominant epoch is one of being in a 'terror state' (encoded as state 1 the blue coloured points). In 2007 there for a large part of the year the predominant epoch is one of being in a ’terror state’. While in 2008 the opposite is true with the predominant epoch being one of a ’non terror state’ (encoded as state 0 the red coloured points), which would give some support for the surge having a positive effect and infact reducing terrorism as the post surge period is correlated with a reduction in number of days being in an epoch days in a ’terror state’.

Understanding of the effects of US troop surge in 2007 can be further enhanced by aggregating up the terror epoch days (per month) (see figure A.40) and the deaths per month and creating a faceted plot of both of these with the daily death count plot for the period 2006-2009. It is clear from the plot that the monthly number of ’terror epoch days’ was higher pre compared to post surge and the surge did see an initial increase in ’terror epoch days’. This again would give some support to the notion that the troop surge did have a positive effect on reducing terrorism in post invasion Iraq.

A.9 Conclusion

This section/appendix focussed on carrying out an exploratory analysis. The reason for this was twofold, to gain an initial understanding of the data and also to gain an understanding of what methods maybe appropriate to analyse the data. To this end, a number of data visualization techniques were used to explore the data along with a combination of data visualization techniques and unsupervised machine learning techniques were used to gain an insight into the data. To analyse time series of count of incidents, time series plots were appropriate to visualize the data using a number of different time intervals (year, month). The time series plots revealed that the intensity of terrorism (in terms of counts of terrorism and counts of deaths due to terrorism) was increasing. Since records began to be recorded in the 1970s terrorism has been on the increase, though in the 1990s the trajectory appeared to slow down or level off. In the 2000s post September 11th 2001 again increased followed by a much more pronounced increase post the US invasion of Iraq. The emergence of ISIS has also seen a large increase in the intensity of terrorism. By stratifying the time series by region and plotting the time series plot, a regio-temporal relationship is also apparent. That is, before the 1980s no one region is seen to be dominant in terms of terrorism, while in the 1980s central America and South America is the dominant region. Again in the 1990s no particular no one region appears to be dominant though, South Asia, South America , Middle East and North Africa and Sub-Saharan Africa appear to be the prevalent regions. In the 2000s and 2010s the Middle East and North Africa, Sub-Saharan Africa and South Asia are the predominant areas. By utilizing a stacked choropleths (a thematic map of deaths due to terrorism) by decade, an enhanced understanding of the regional time-series plots could be attained.

One can see from the choropleth that in the 1970s terrorism appears to be dominated by the United Kingdom and Iran. The 1980s by Central and South American countries, particularly Columbia, El Salvador, Nicaragua and Peru and India. During the 1990s the pattern becomes more diffuse with terrorism being dominant across a number of countries Algeria, South Africa, India and Turkey. Russia also starts to become prevalent in the 1990s in terms of deaths due to terrorism and this trend continues into the 2000s. The 2000s sees terrorism almost disappear from countries were it was previously prevalent in South America (Colombia and Peru). South Asia (Afghanistan, Pakistan and India) and Iraq (terrorism only starts to increase after the US led invasion of Iraq). In the 2010s terrorism becomes dominated by the Middle East (Iraq and Syria), South Asia (Afghanistan and Pakistan) and Nigeria. The use of stacked choropleths proved particularly at showing the regio-temporal shift in terrorism over the decades though is limited by the number of epochs that the stacked choropleth can accommodate. If too many epochs are used the display of the information becomes difficult, as one can only (even with the use of interactive graphics) process a limited number of stacked choropleths. While using too few epochs the visualization suffers from lack of detail and no patterns become discernible. The use of analysis libraries (the r googleVis library) allowed the creation of very rich data visualizations, free of charge that offer a large amount of insight into spatio-temporal patterns of terrorism. Stacked barplots also proved useful at exploring how changes in attack vector or weapon type changed with time. It is quite clear to see that since the 1970s armed assault and bombings become the prevalent attack types. While since the 1970s explosives/bombs/dynamite has emerged as the dominant weapon type. While this data visualization methodology proved useful it was limited to the number of dimensions that could be visualized. To overcome this, a number of unsupervised techniques used in conjunction with interactive data visualization techniques were used. Correspondence analysis, particularly multiple correspondence analysis were used to understand interactions and associations between multiple data dimensions. The use of interactive data visualizations in tandem with MCA through the use of interactive bi-plots allowed associations between different categorical levels to be made. Bi-plots are a type of scatterplot which allows data on both the transformed dimensions, original variables and samples to be displayed together. By displaying the output in an interactive form if the visualization becomes too dense the analyst can either zoom in a particular are of interest or else can temporarily remove particular variables to allow a clearer understanding of the association between variables. From the use of MCA a number of associations could be observed these were:

- There is a vigorous association between the year 1974 and terrorist attacks on non state militia or terrorists.
- Unarmed assaults are largely associated with chemical weapons use.
- The year 2003 is strongly associated with attacks against government, while the year 2009 shows a strong association between attacks targeting police.
- The year 1994 is strongly associated with attacks using weapon types sabotage equipment and mellee weapons against facilities and infrastructure.

These methods allowed uncovering of visualizations that be difficult to detect by other methods but their use with the GTD did pose some problems. Particularly trying to analyse the associations between terrorist groups with particular epochs due to the large number of groups or more apt the numerous designations for groups.

Geo-spatial analysis was another exploratory analysis method that was possible by combining either clustering or kernel density estimation. Such techniques in conjunction with interactive data visualizations allow an analyst to determine centres of activity for terrorism or the spatial dispersion of incidents whether they are occurring in rural or urban areas. Spatial K-means clustering of terrorist events showed large concentrations of terrorist incidents. create a spatio temporal plot of incidents in Iraq and Syria The preliminary modelling consisted of two distinct parts Regression modelling of count of deaths due to terrorism in Iraq and time series count modelling using HMMs.

Preliminary modelling was carried out on the GTD and enhanced dataset sourced from the GTD. The enhanced dataset was created by appending data sourced on Iraqi and US presidential terms and US troop activity regarding the invasion along with information regarding periods of major insurgency activity within Iraq. US troop activity would be concerned with major actions by Allied and US forces concerning the invasion of Iraq , the 2007 surge to combat the Iraqi insurgency and the subsequent drawdown and eventual pull out and withdrawl of Allied and US troops. While count models were used they proved to be problematic and suffered from being incorrectly specified.

HMMs were also used to model the count data using two different states (one representing a state of high level terrorism and the other one of low level terrorism). While this model worked well at being able to classify a certain period of time being in one of the above epochs it proved difficult to generalize and would be of limited use in modelling count terrorism data. For this reason alternative methodologies were employed in the modelling chapter (chapter 5).

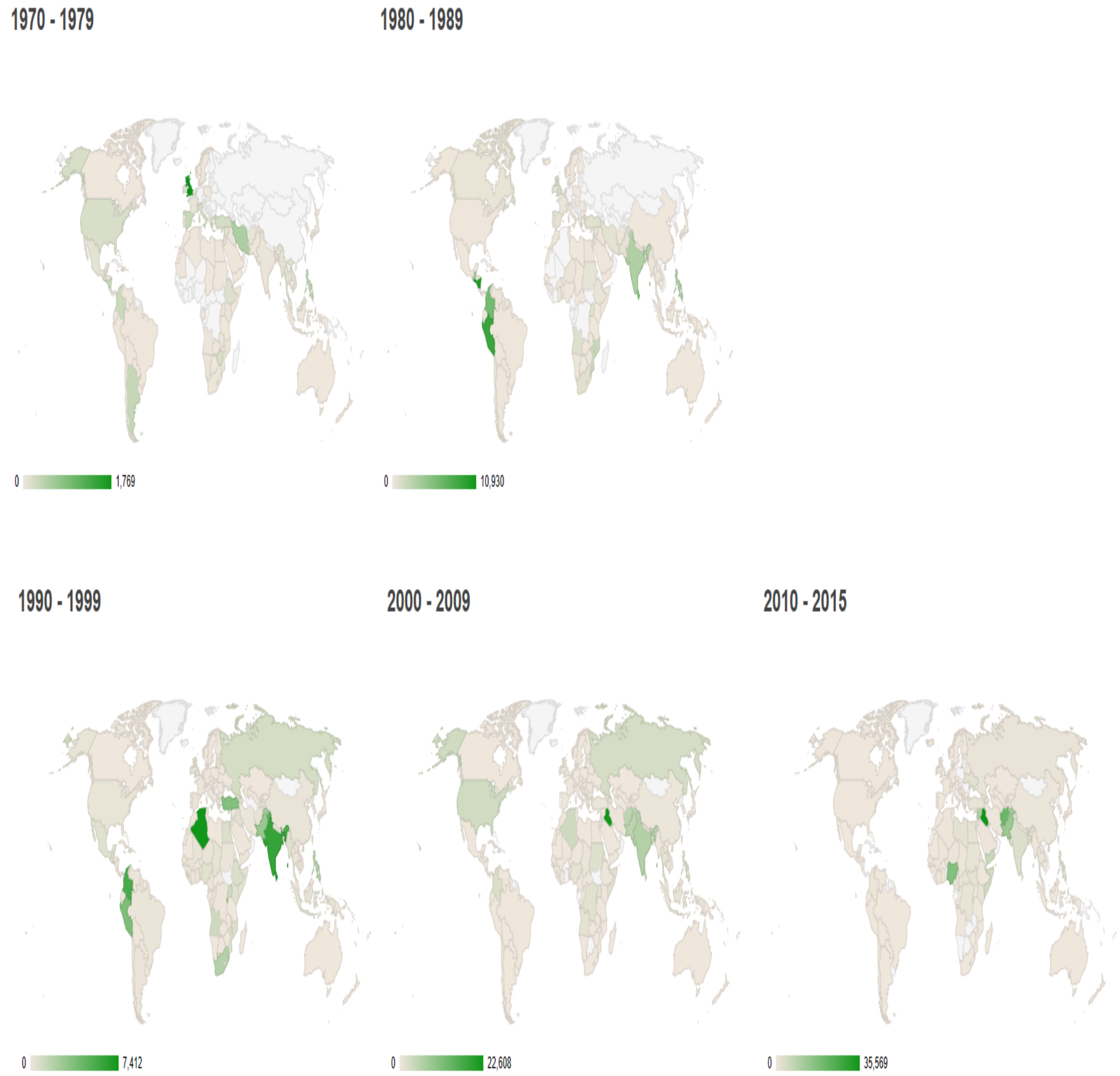

Figure A.9: A stacked choropleth showing deaths by decade due to terrorism.

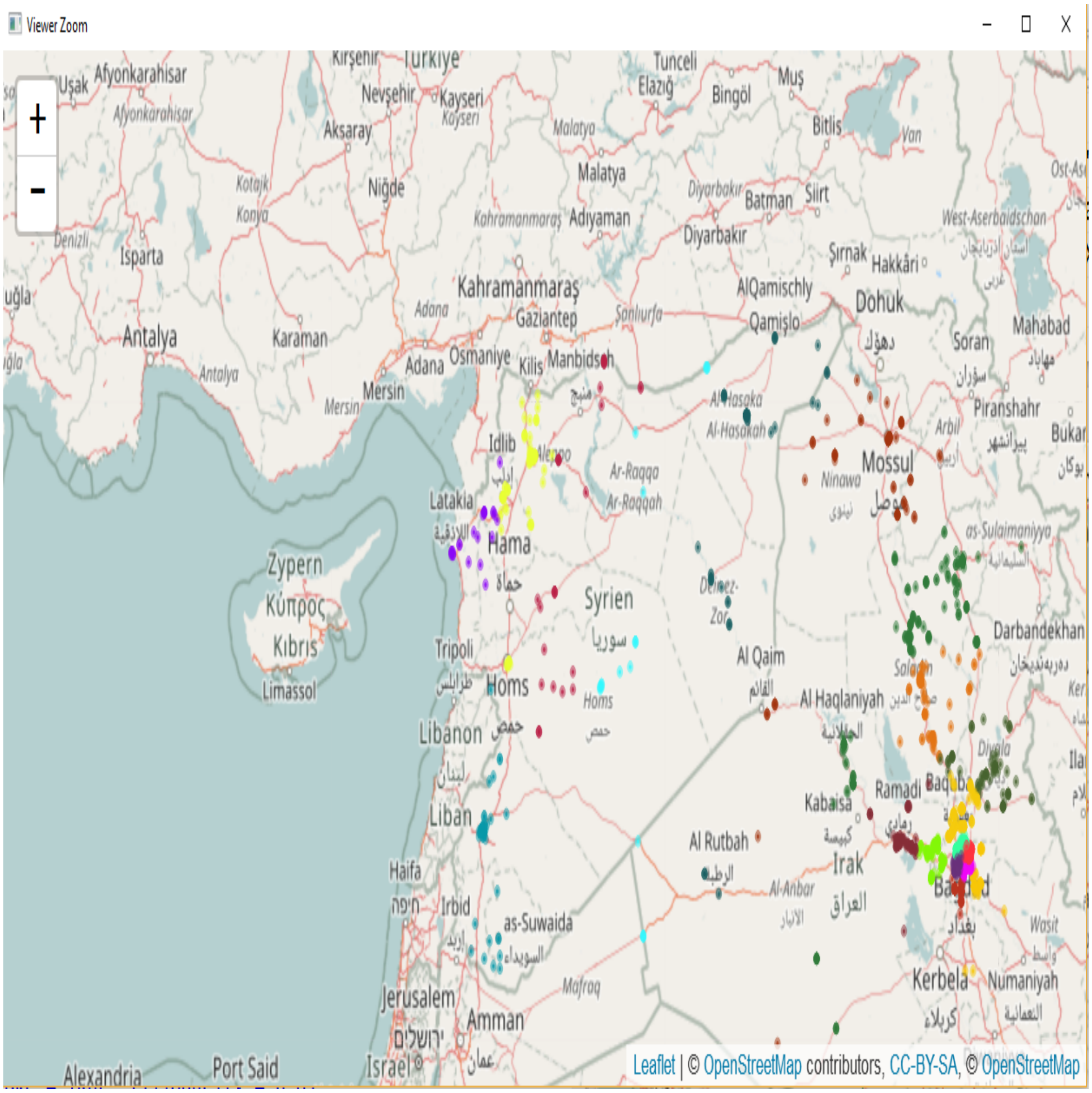

Figure A.10: Clustering of Terrorist incidents in Syria and Iraq

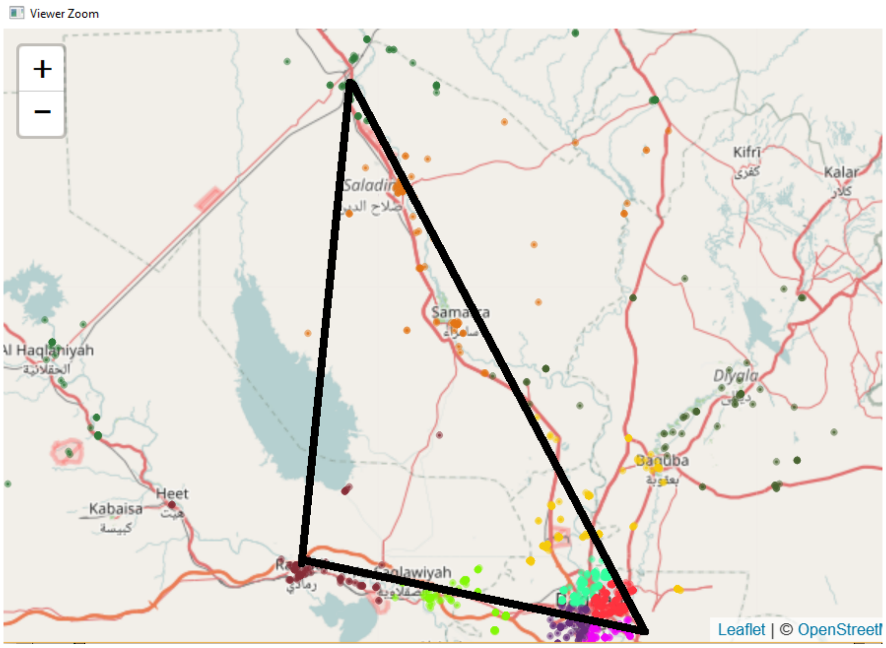

Figure A.11: Clustering of Terrorist incidents in Syria and Iraq, centred on the Sunni triangle

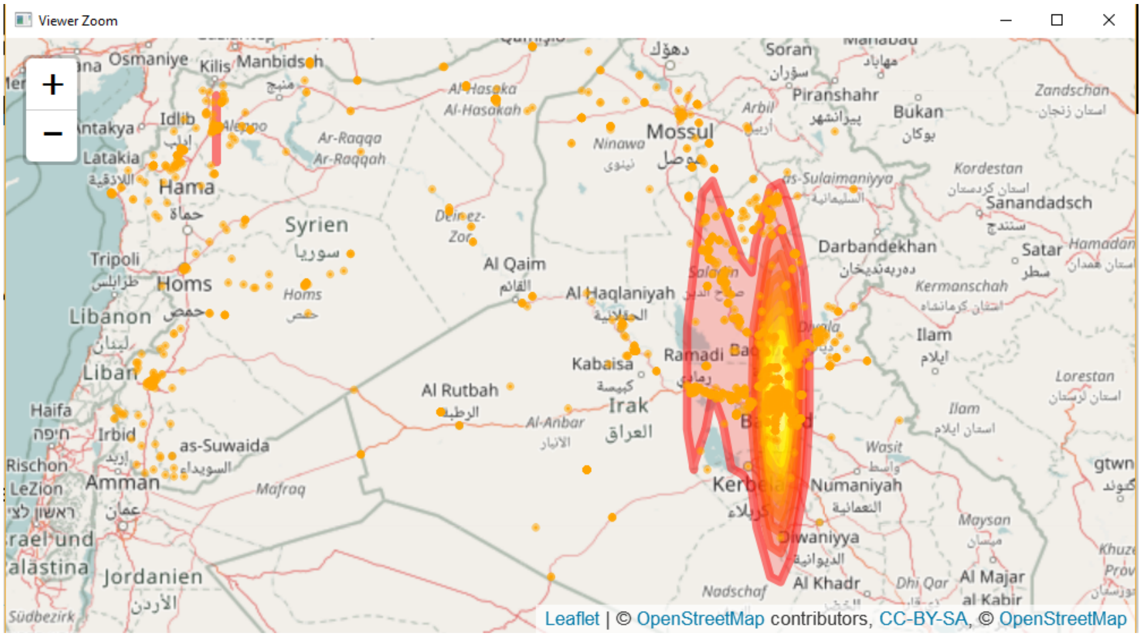

Figure A.12: Heatmap of terrorist incidents in Syria and Iraq

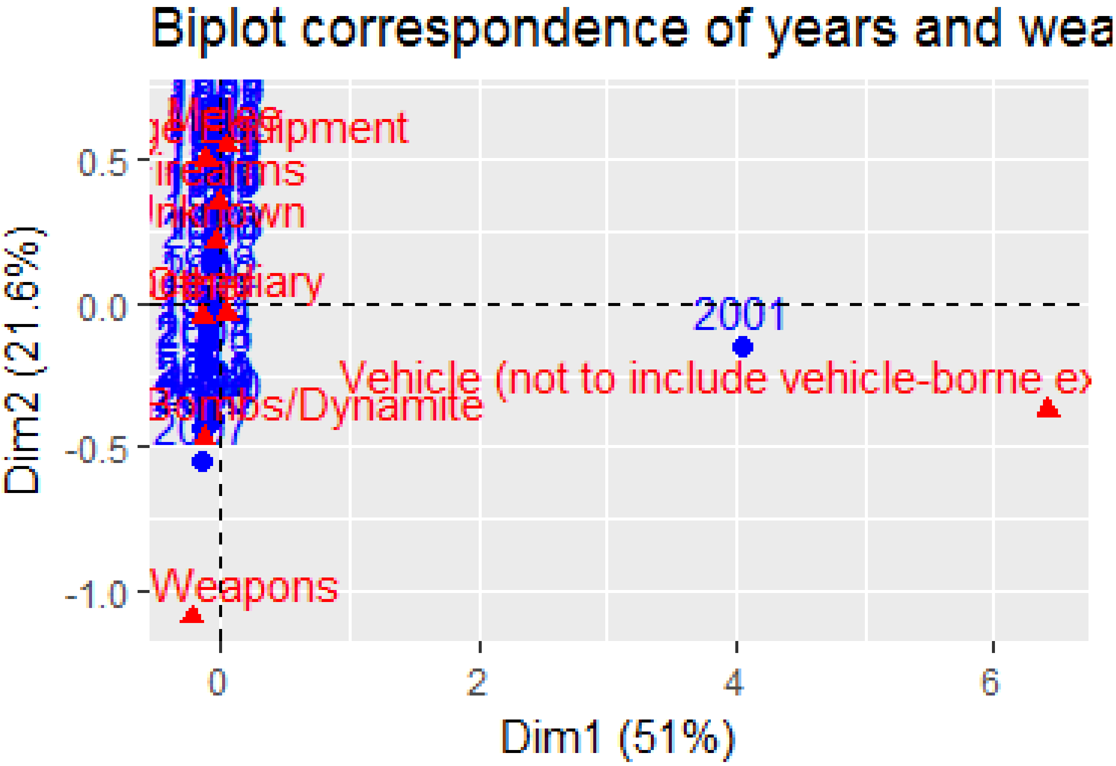

Figure A.13: Biplot CA of contingency table of deaths by weapons and year.

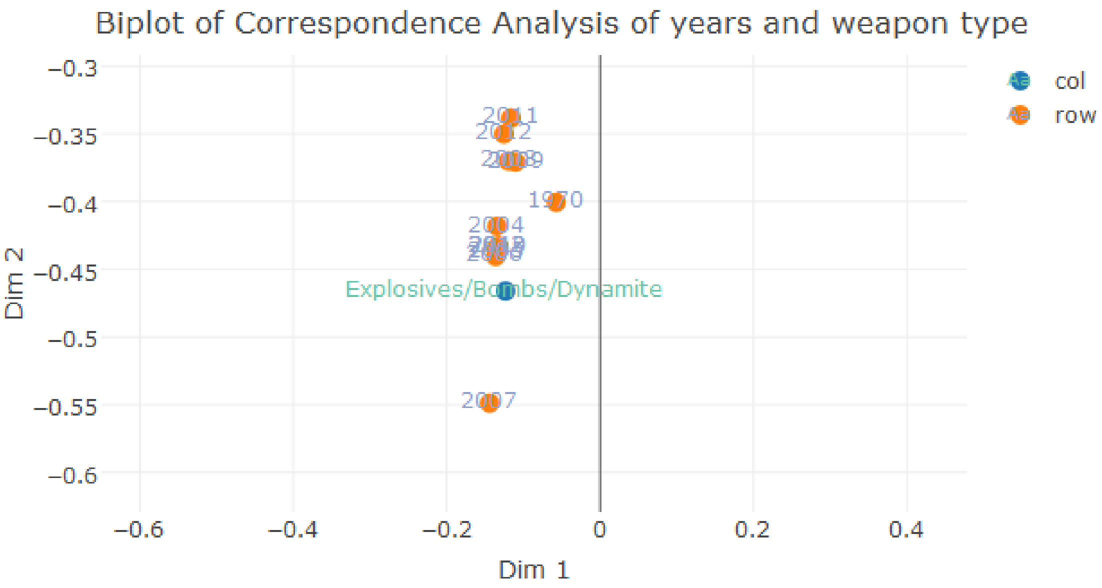

Figure A.14: Biplot CA of contingency table of deaths by weapons and year, created using Plotly, zoomed in.

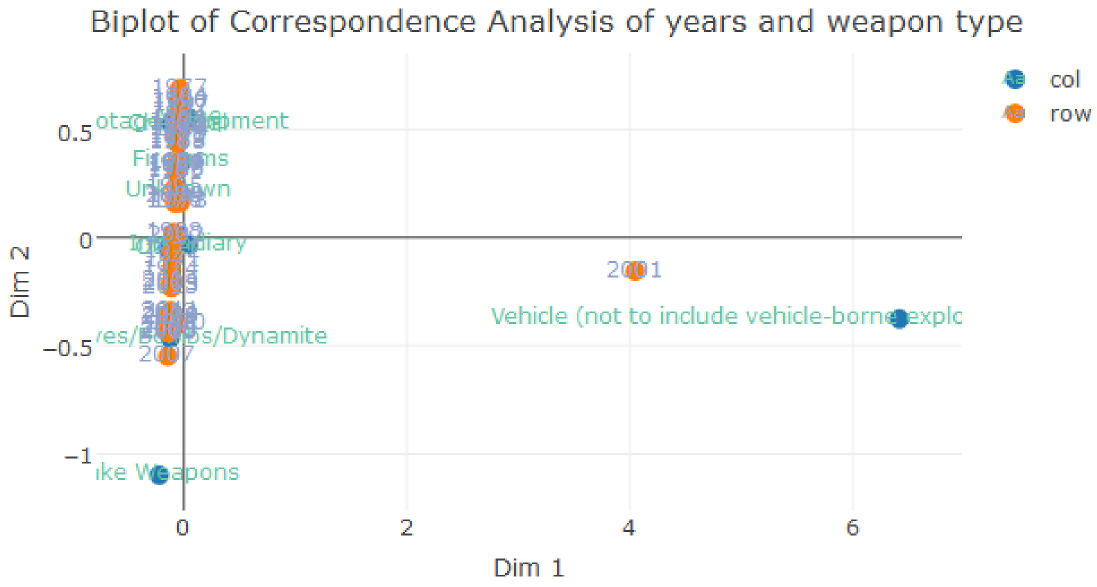

Figure A.15: Biplot CA of contingency table of deaths by weapons and year, created using Plotly, zoomed in.

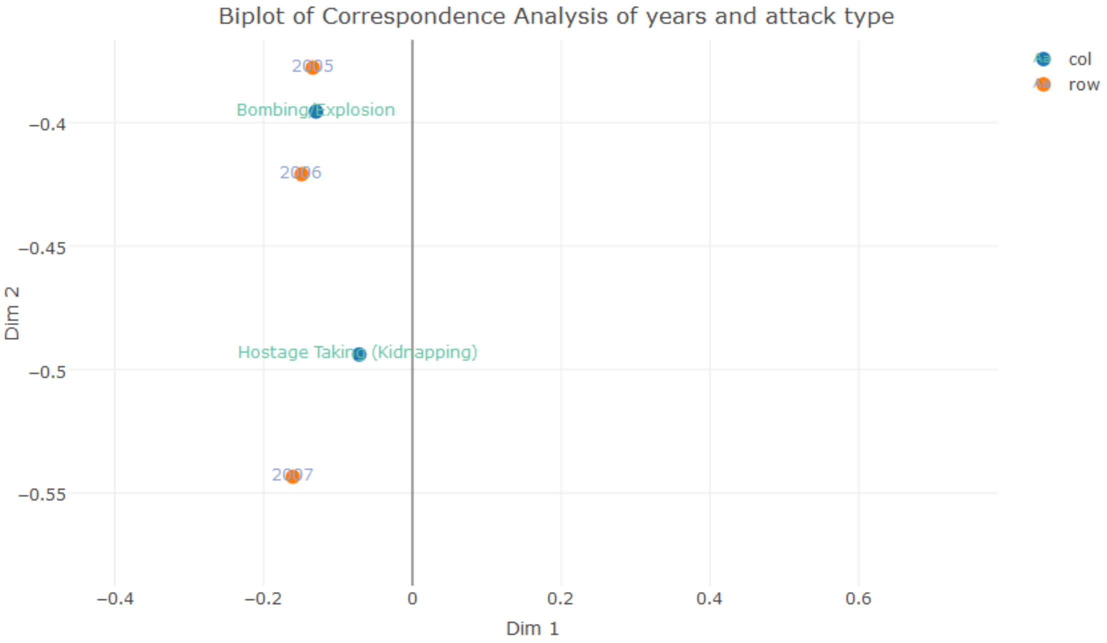

Figure A.16: Biplot CA of contingency table of deaths by attack type and year, created using Plotly, zoomed in to show associations between explosions and years they occurred.

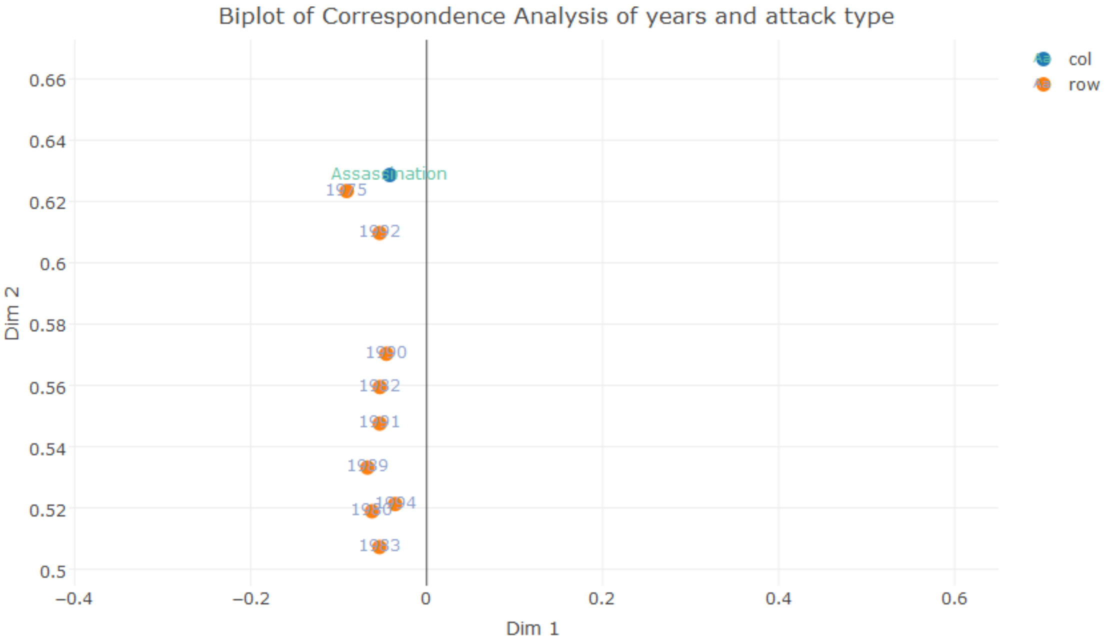

Figure A.17: Biplot CA of contingency table of deaths by attack type and year, created using Plotly, zoomed in to show associations between assassinations and years they occurred.

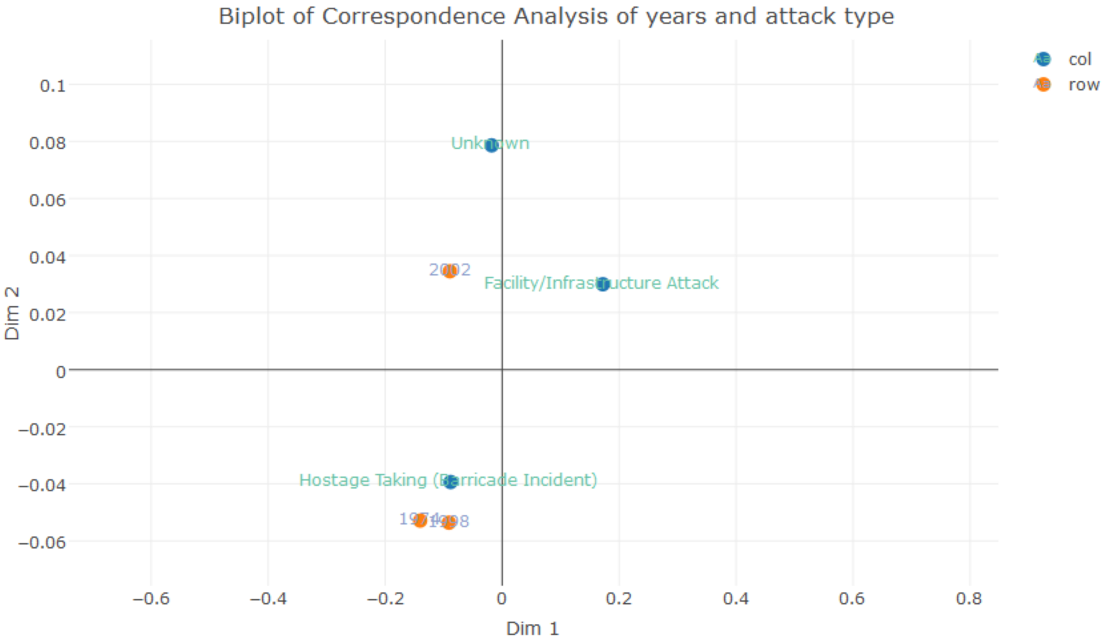

Figure A.18: Biplot CA of contingency table of deaths by attack type and year, created using Plotly, zoomed in to show associations between hostage taking and years they occurred.

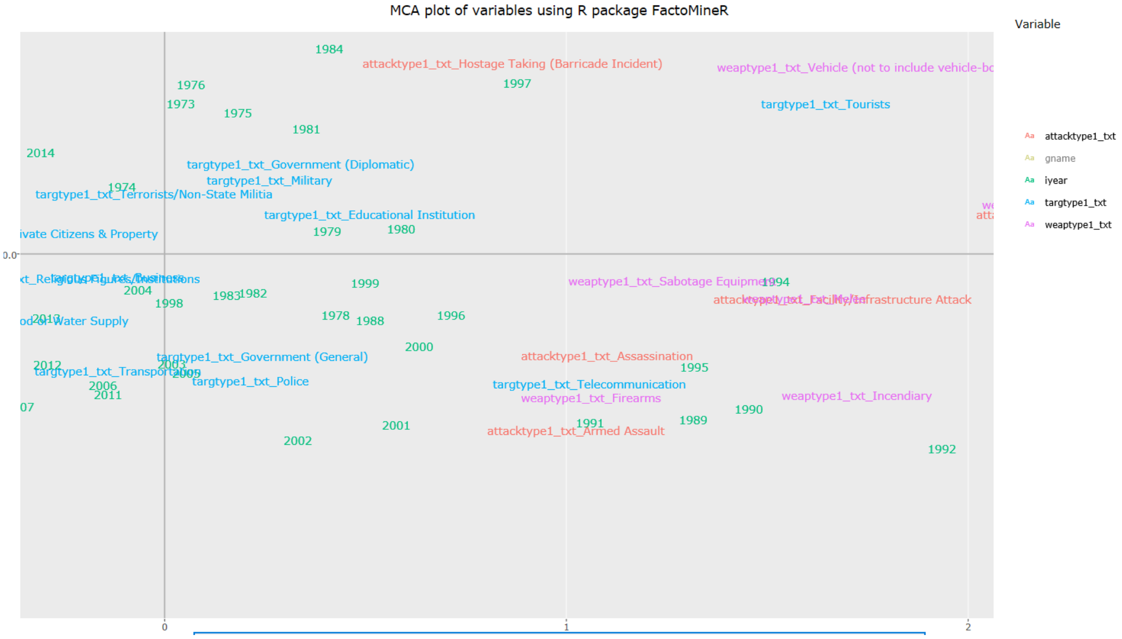

Figure A.19: Biplot of MCA of incidents by attack type, weapon type, target type, group name and year, created using Plotly, zoomed in to show associations between hostage taking and years they occurred.

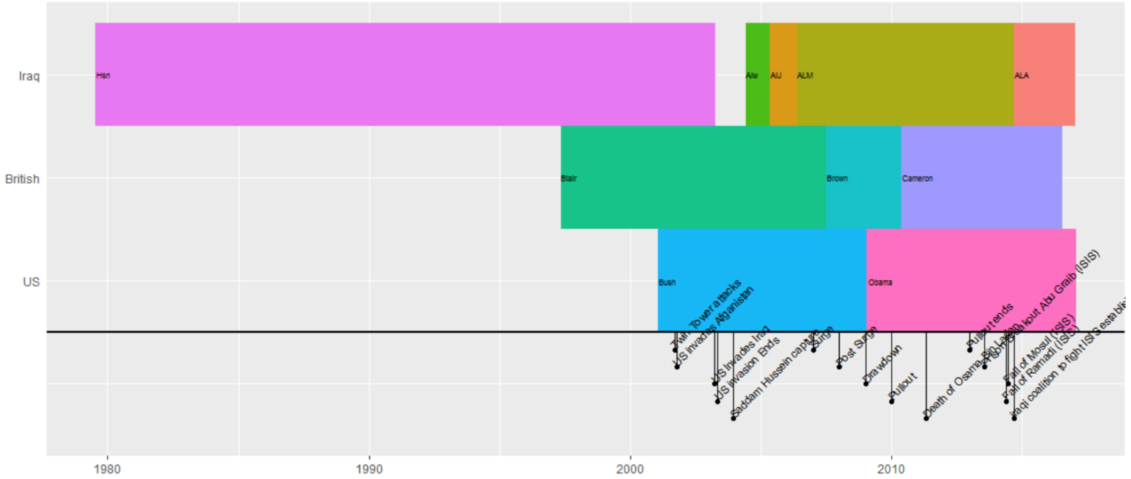

Figure A.20: Timeline of Iraq post and pre invasion (2003

.

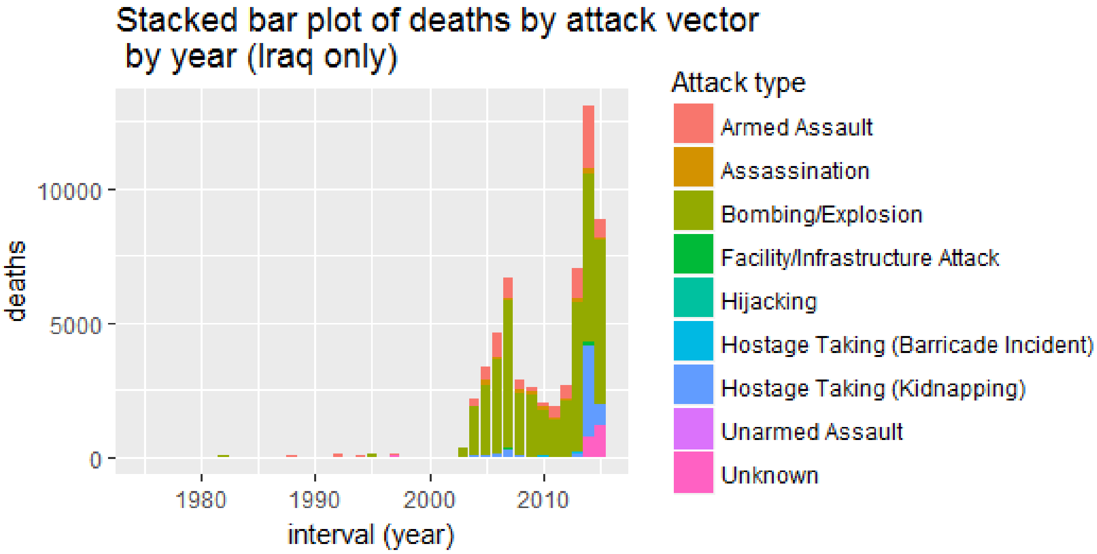

Figure A.21: Timeline of Iraq and US and British presidents and major events since the Iraq invasion.

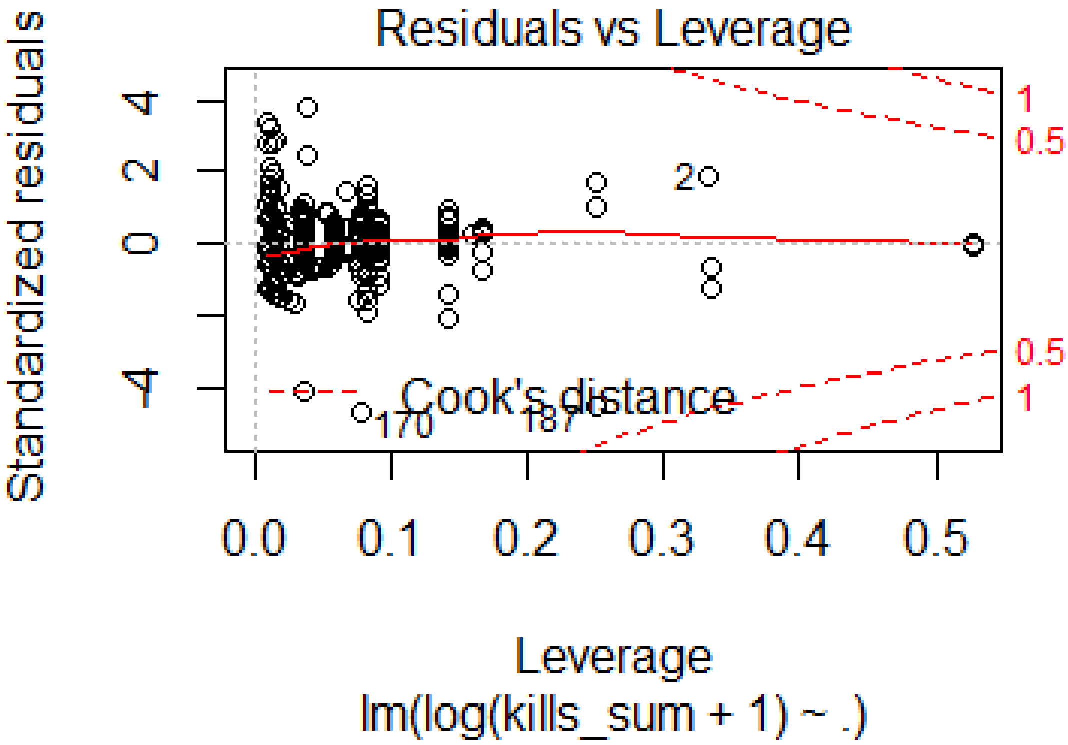

Figure A.22: Cook distance diagnostic plot

.

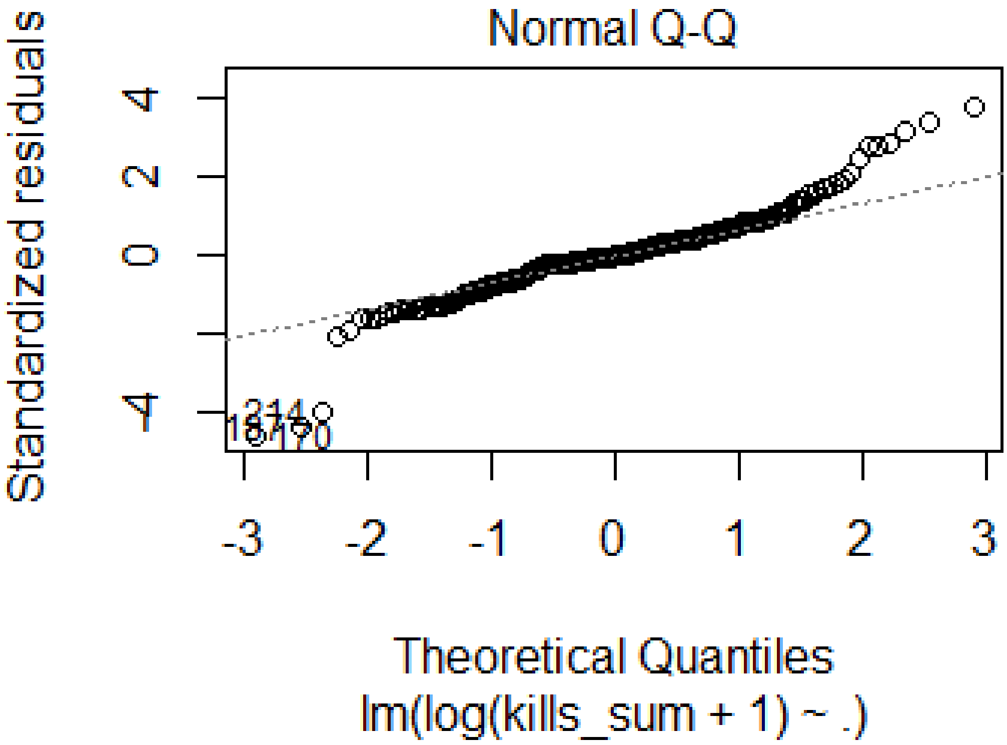

Figure A.23: QQ plot diagnostic plot

.

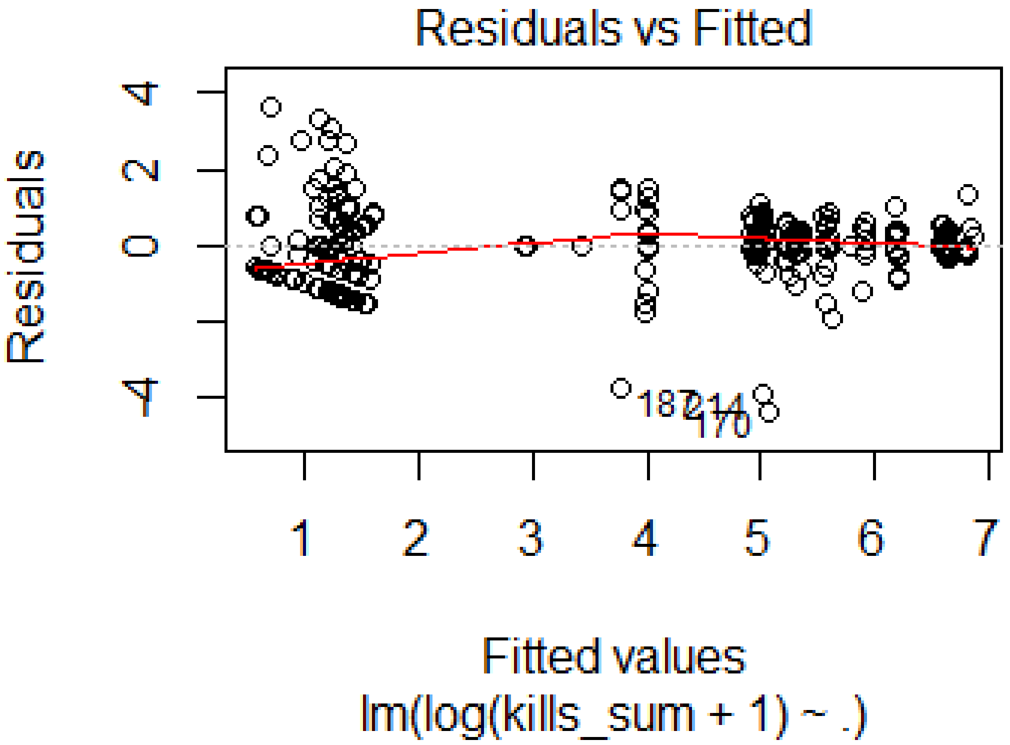

Figure A.24: Fitted residuals diagnostic plot

.

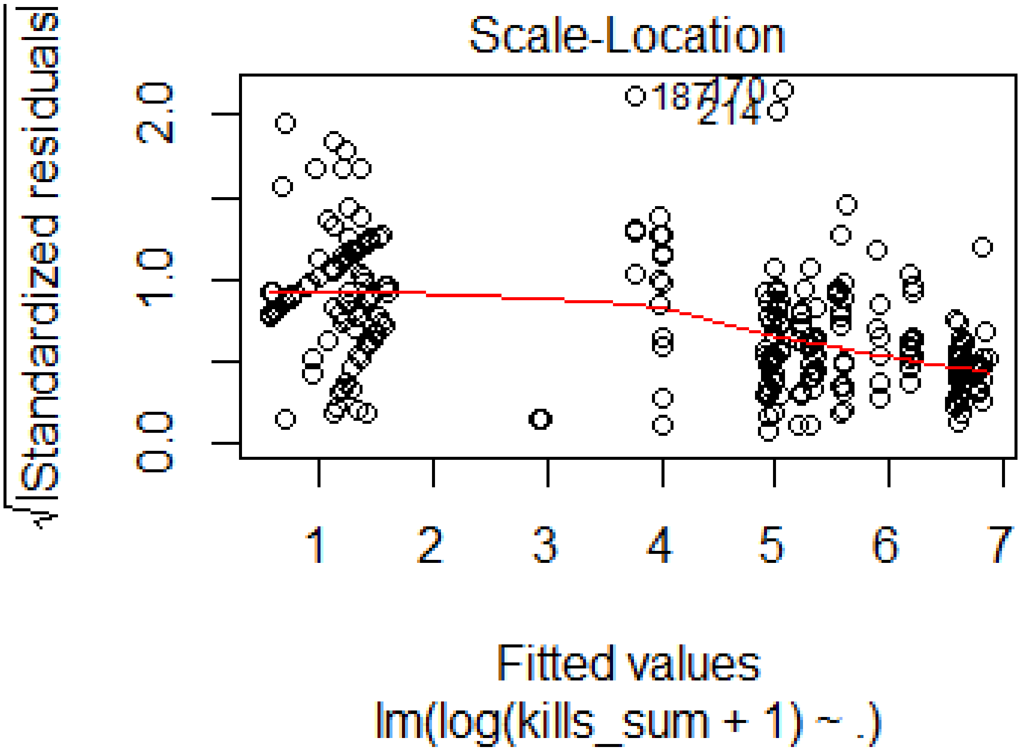

Figure A.25: Scale location diagnostic plot

.

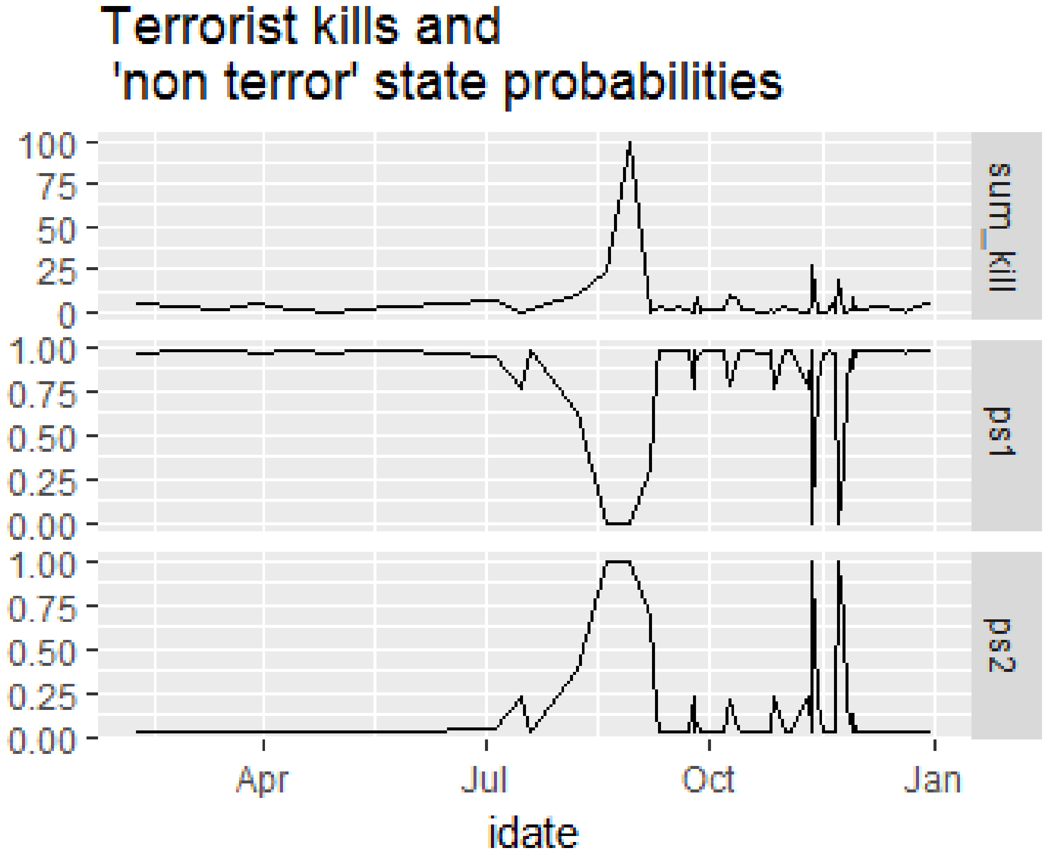

Figure A.26: Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2003-2004.

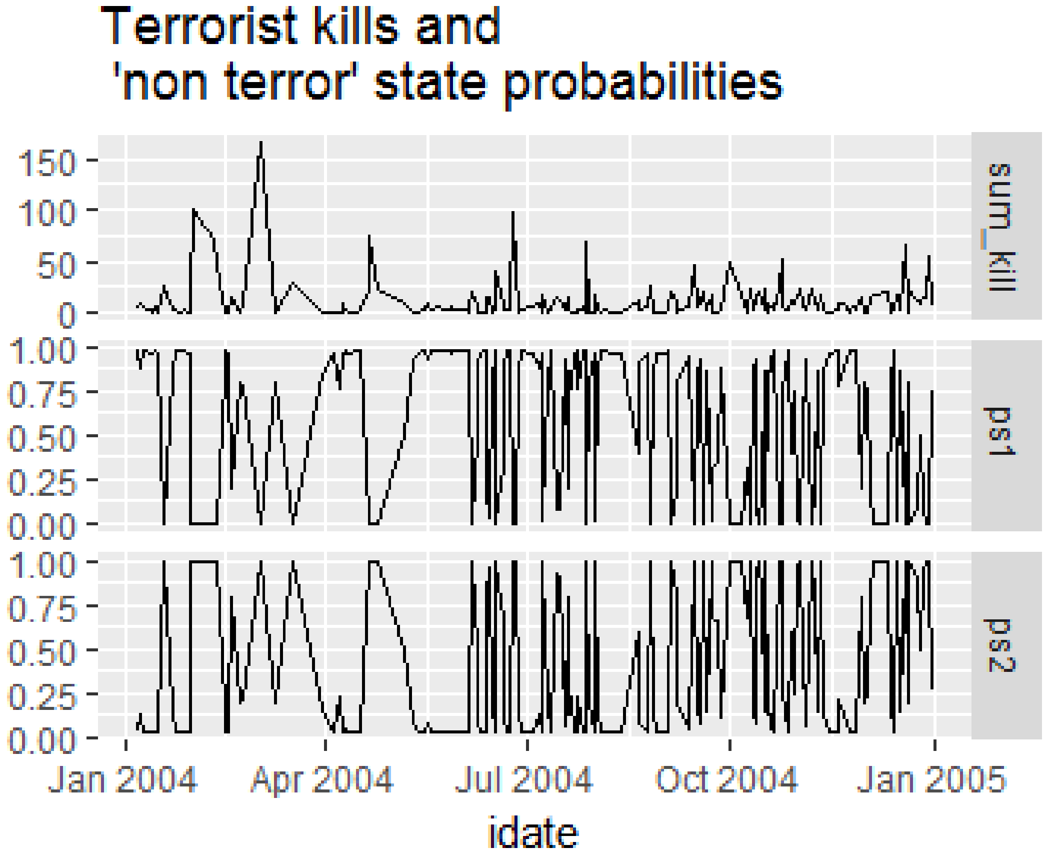

Figure A.27: Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2004-2005.

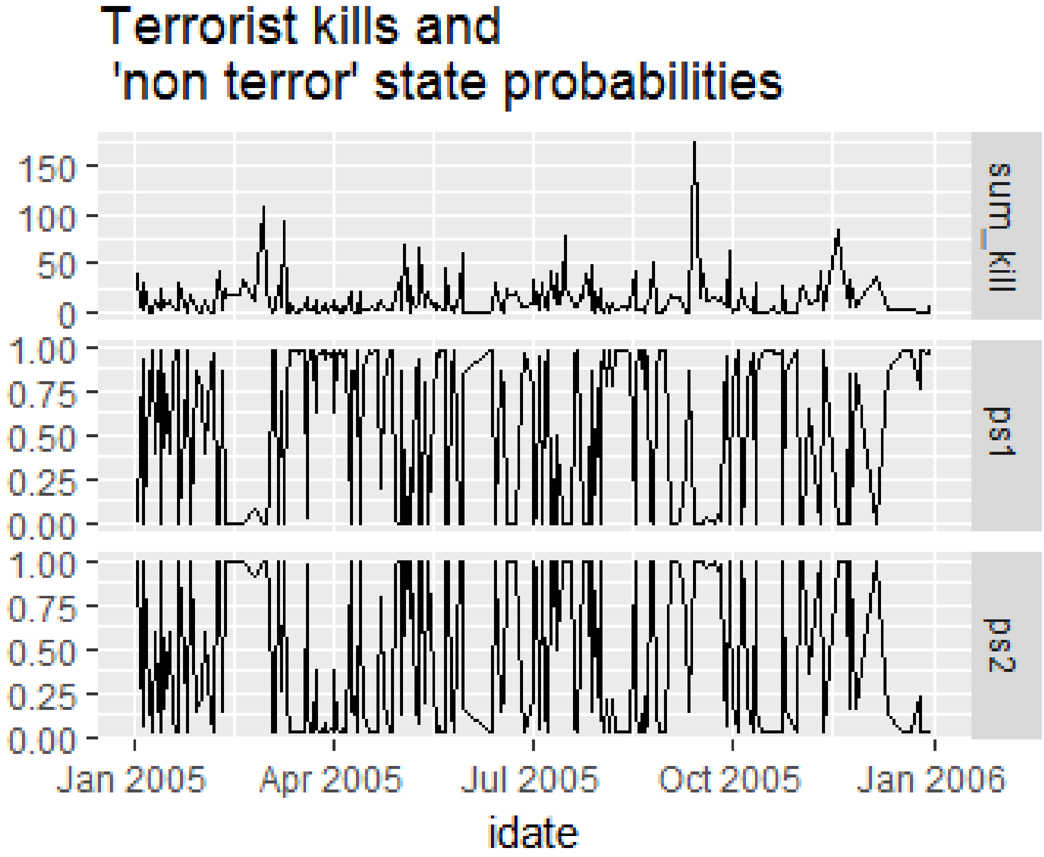

Figure A.28: Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2005-2006.

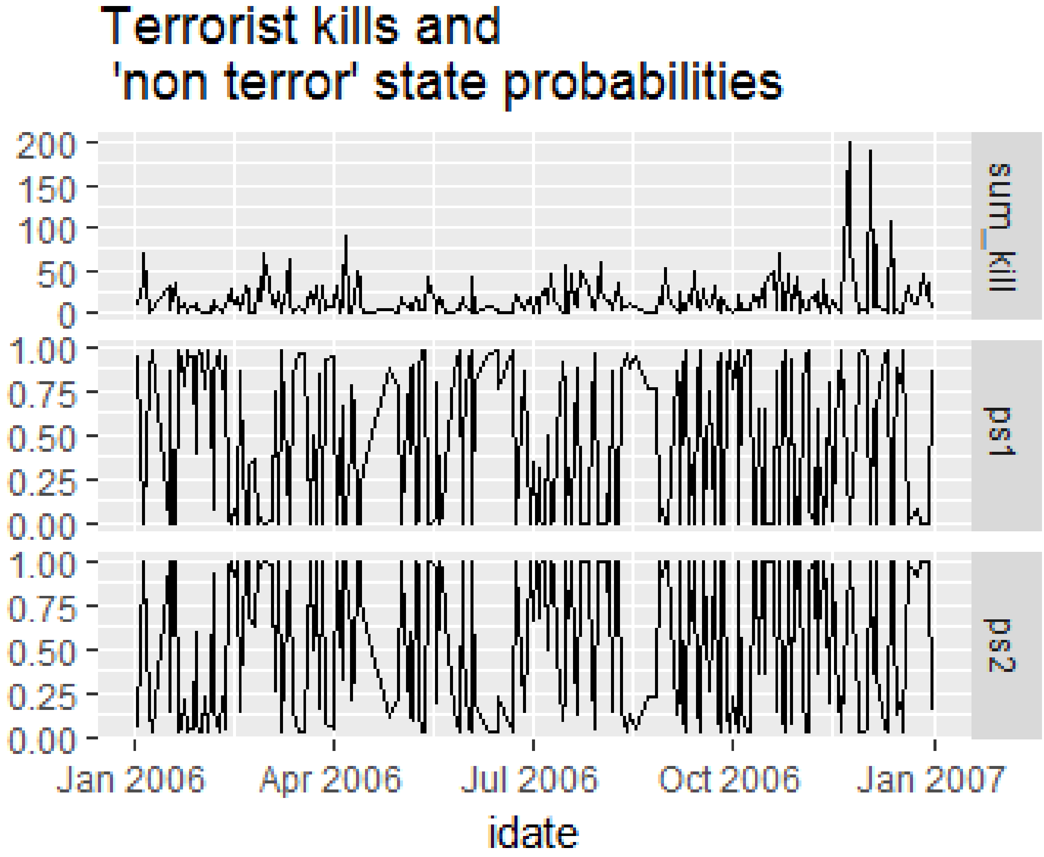

Figure A.29: Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2006-2007.

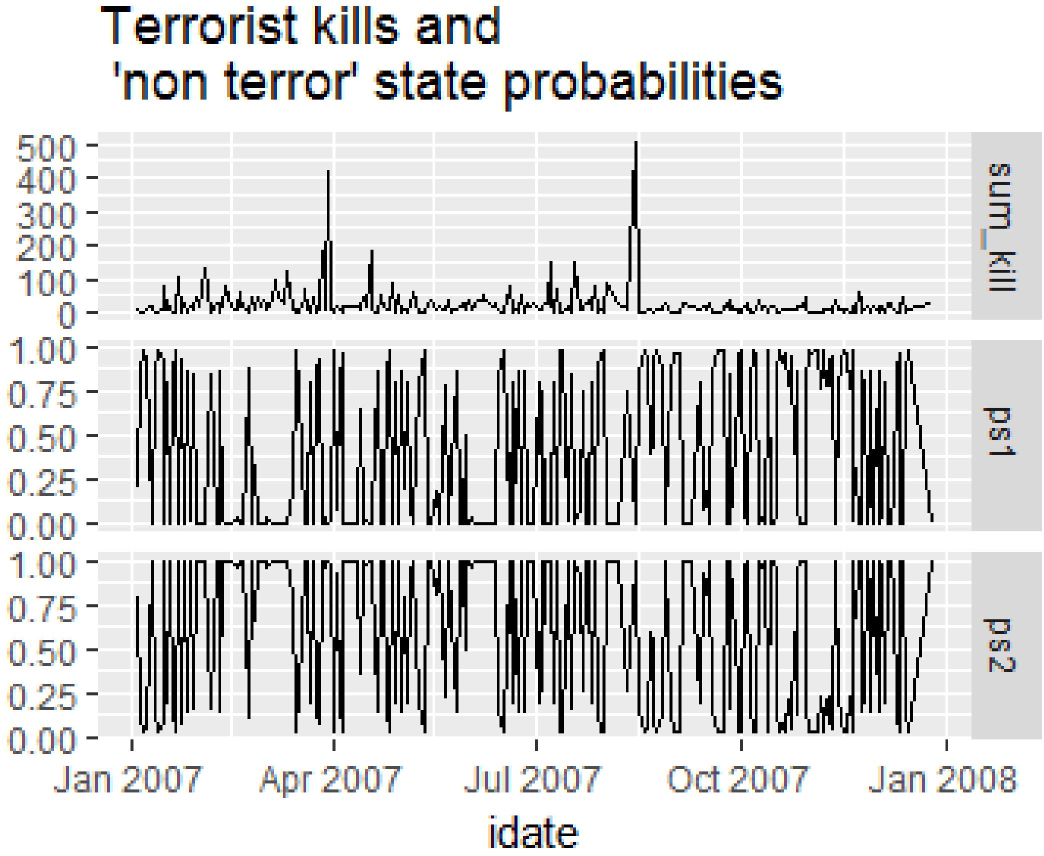

Figure A.30: Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2007-2008.

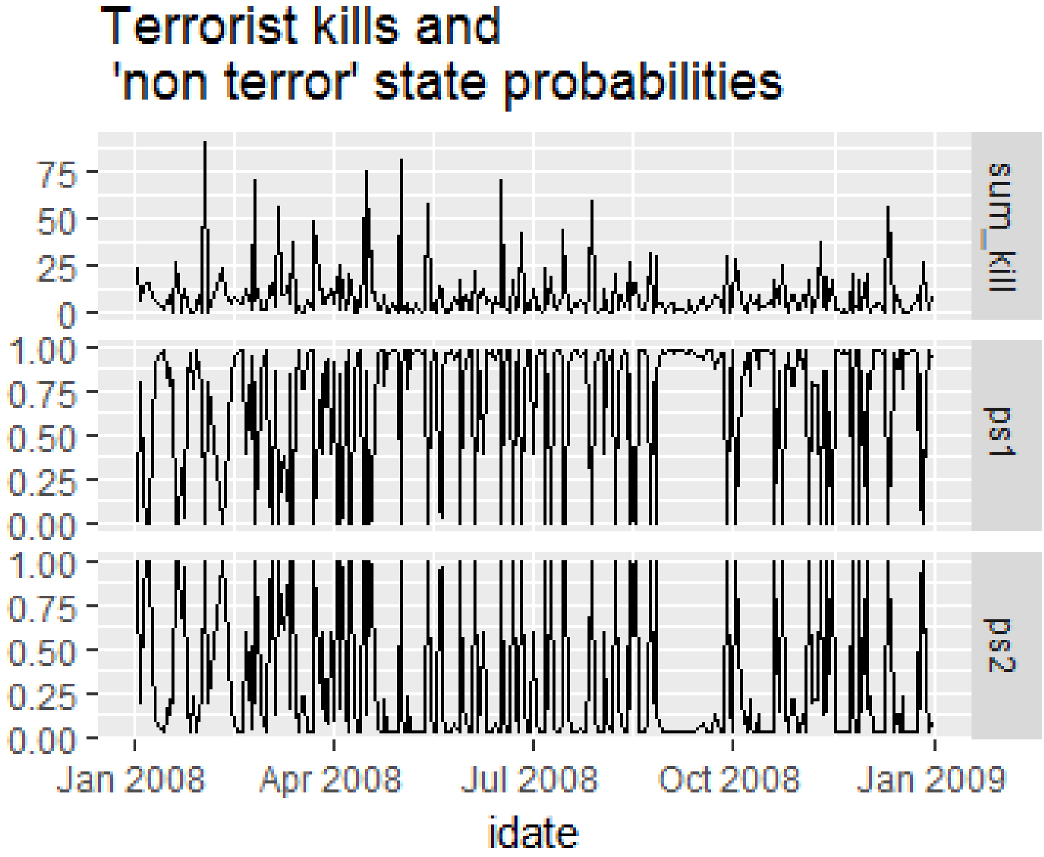

Figure A.31: Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2008-2009.

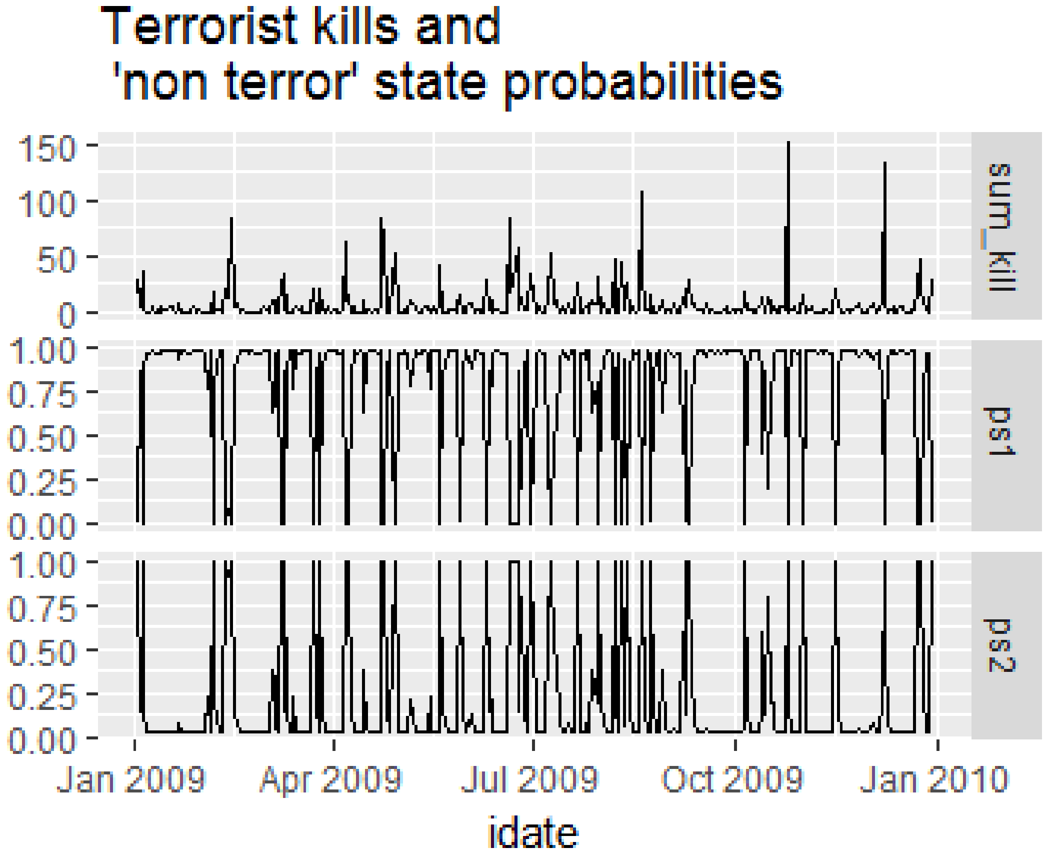

Figure A.32: Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2009-2010.

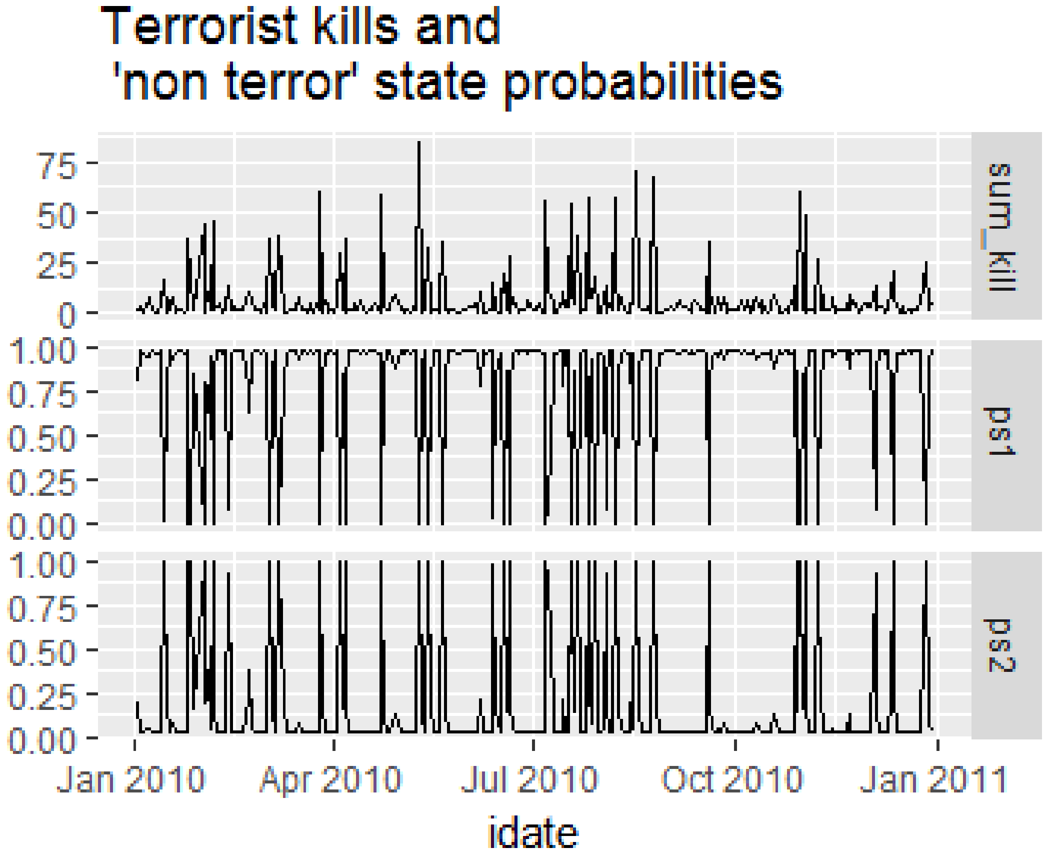

Figure A.33: Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2010-2011.

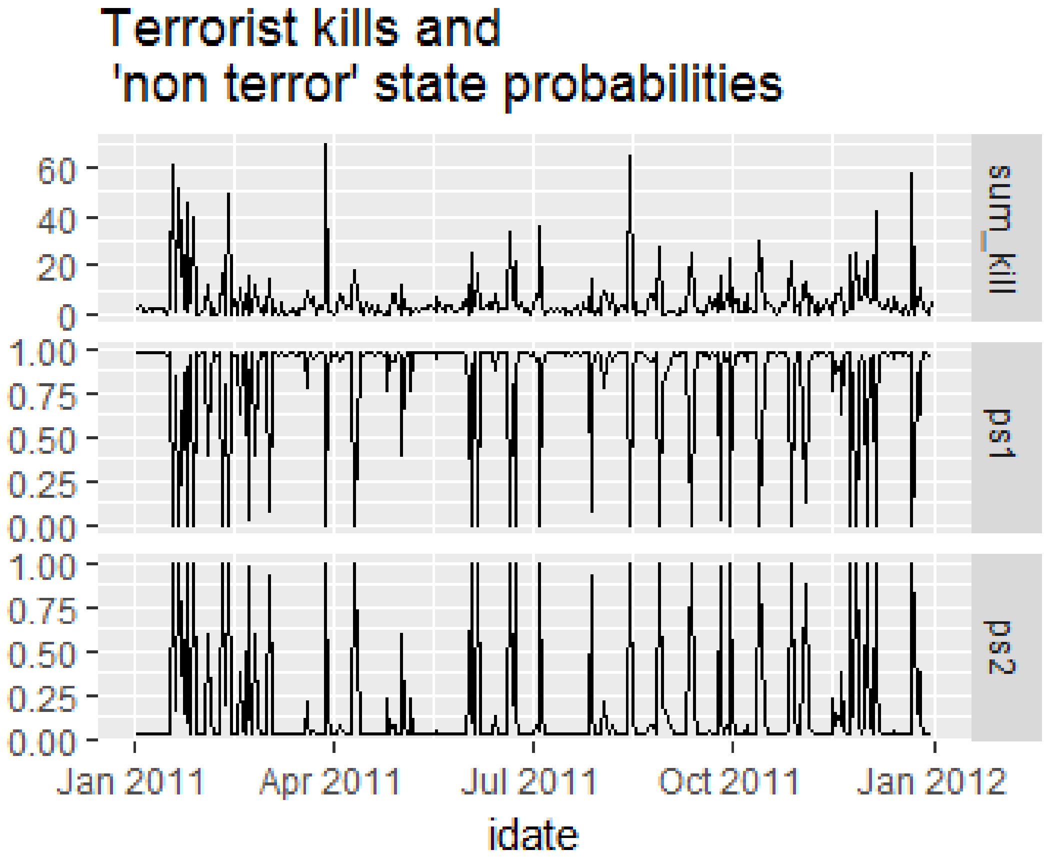

Figure A.34: Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2011-2012.

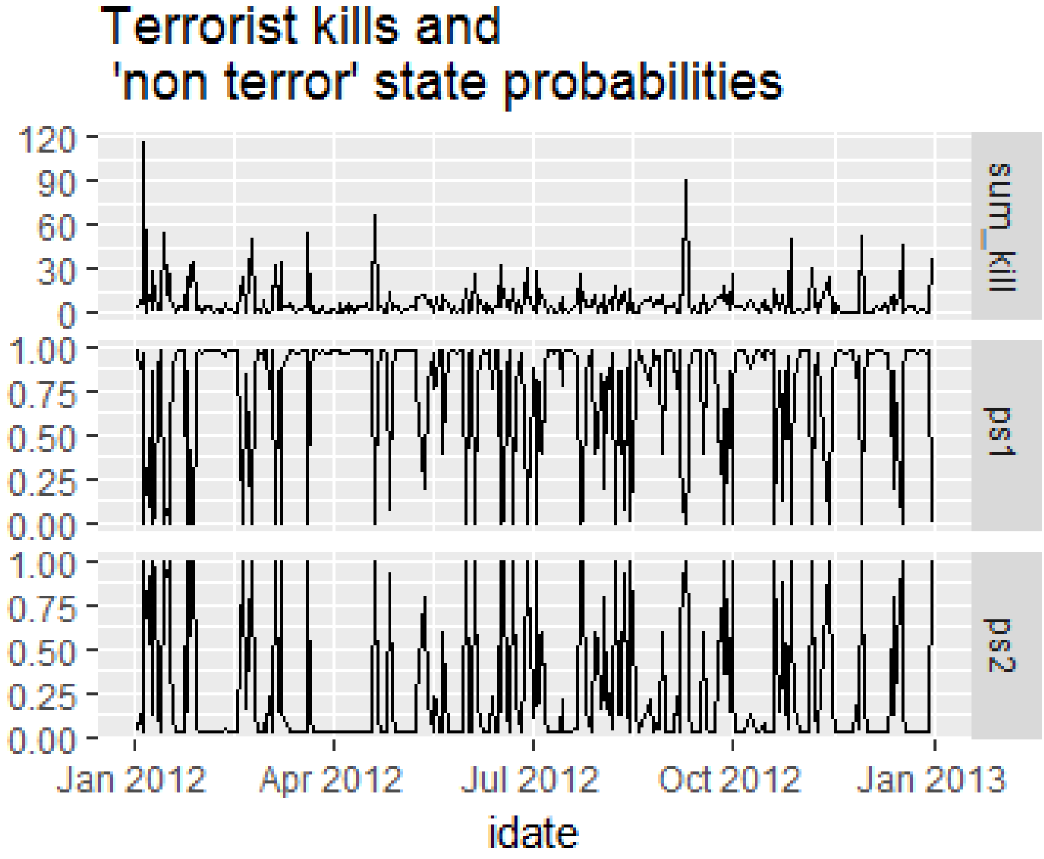

Figure A.35: Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2012-2013.

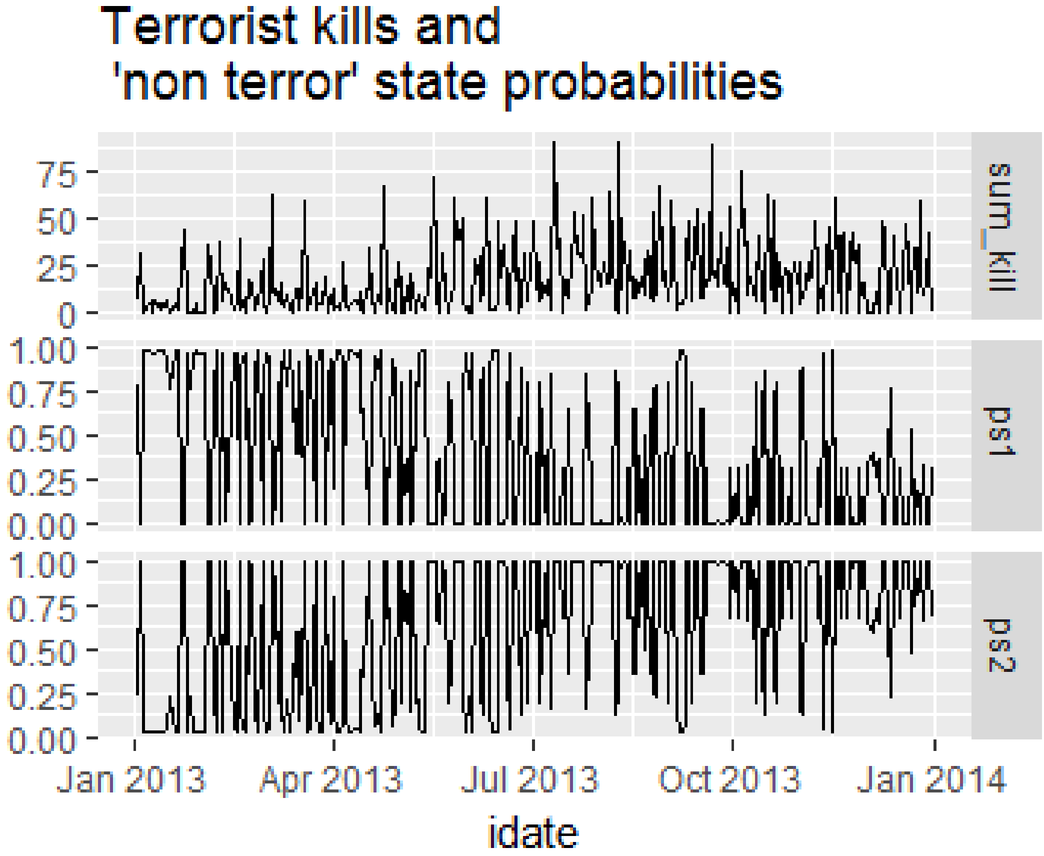

Figure A.36: Faceted plot of daily death count and probability of emission states due to terrorism from Iraq modelled using a HMM, 2013-2014.

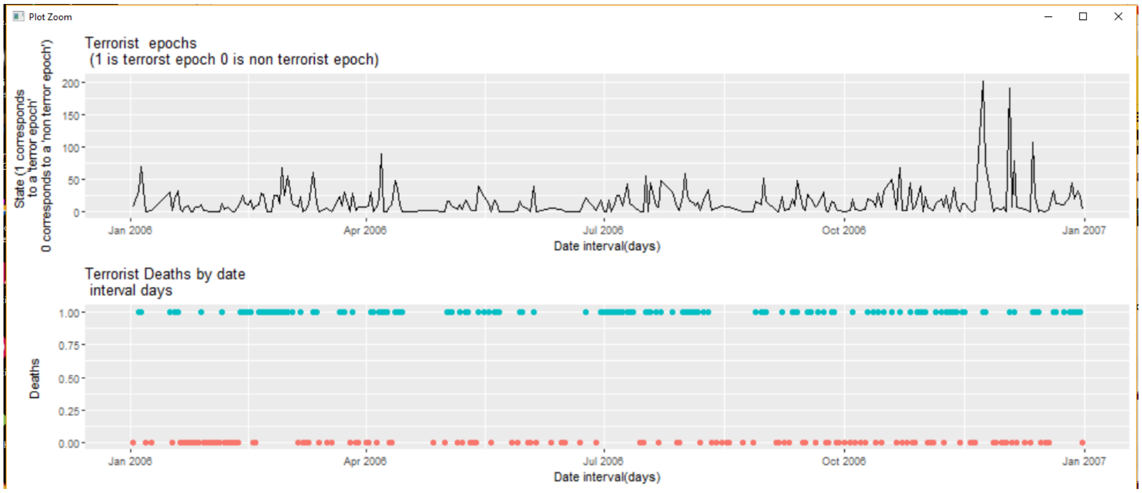

Figure A.37: Faceted plot of daily death count and epoch classification due to terrorism from Iraq modelled using a HMM, 2006.

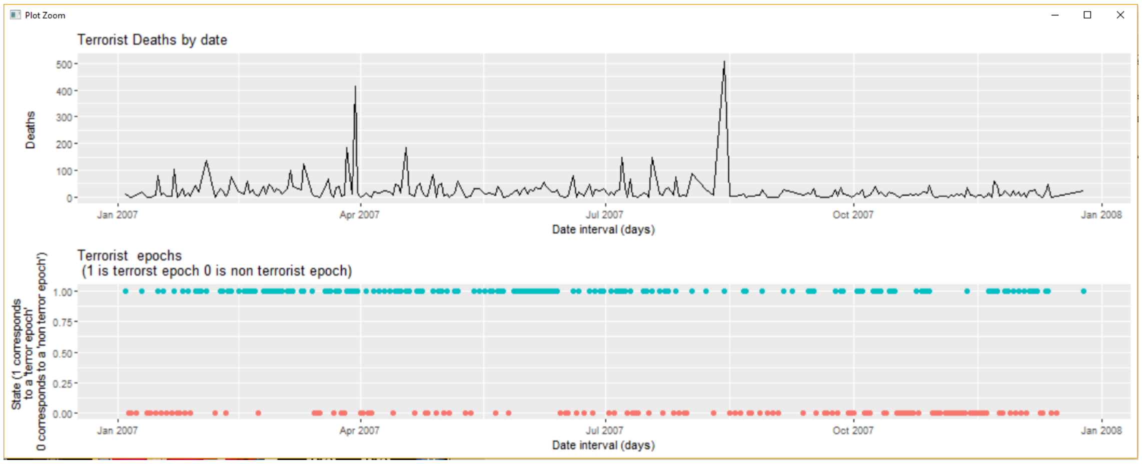

Figure A.38: Faceted plot of daily death count and epoch classification due to terrorism from Iraq modelled using a HMM, 2007.

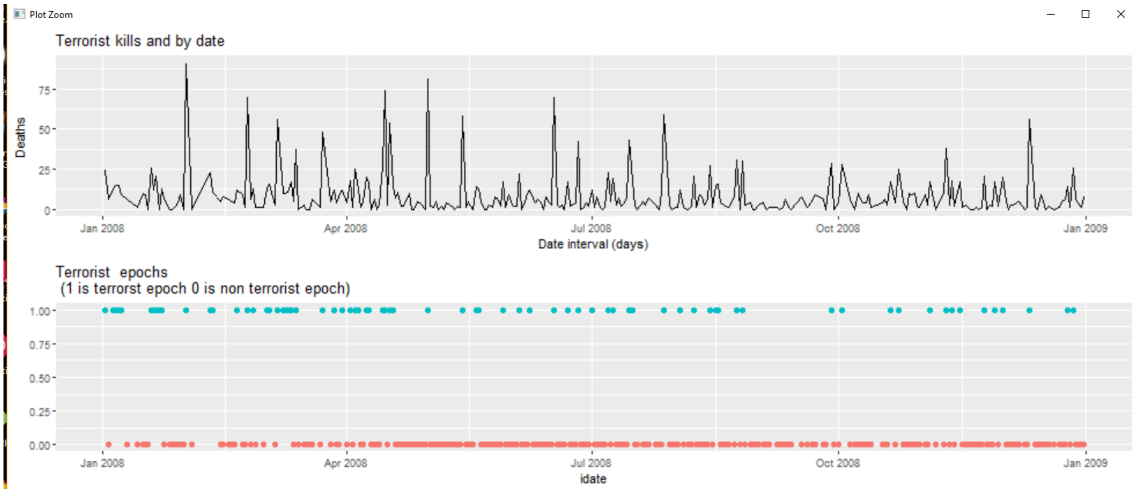

Figure A.39: Faceted plot of daily death count and epoch classification due to terrorism from Iraq modelled using a HMM, 2008.

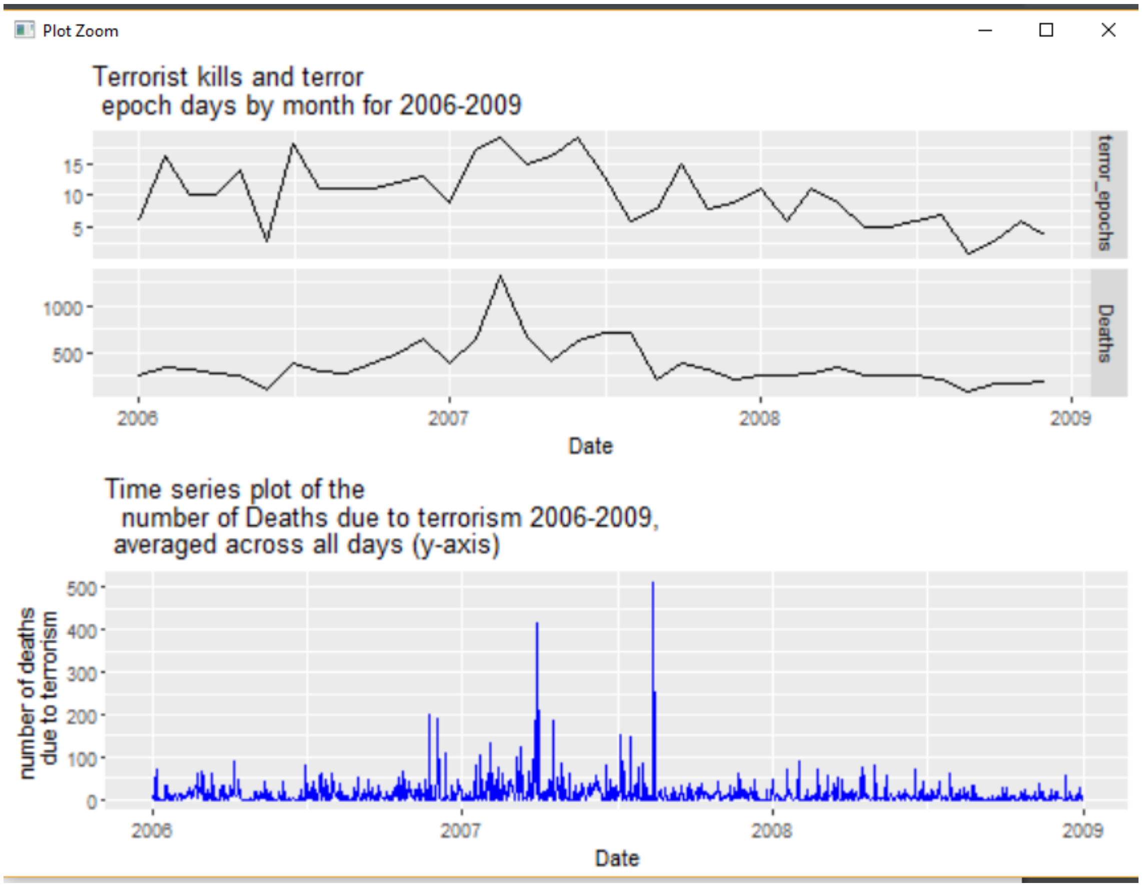

Figure A.40: Faceted aggregate plot of monthly death count and epoch classification due to terrorism from Iraq modelled using a HMM with the daily death count for the same period, 2008.

Appendix B

Chapter 4 regression estimates and two way contingency table results

	Biological	Chemical	Explosives/Bombs/Dynamite	Fake Weapons	Firearms	Incendiary	Melee	Other	Radiological	Sabotage Equipment	Unknown	Vehicle
1970	0.00	0.00	97.00	0.00	45.00	14.00	0.00	0.00	0.00	0.00	14.00	1.00
1971	0.00	0.00	82.00	0.00	80.00	1.00	2.00	0.00	0.00	0.00	8.00	0.00
1972	0.00	0.00	260.00	0.00	289.00	3.00	5.00	0.00	0.00	0.00	9.00	0.00
1973	0.00	1.00	80.00	0.00	270.00	1.00	8.00	0.00	0.00	0.00	10.00	0.00
1974	0.00	0.00	270.00	0.00	246.00	0.00	4.00	0.00	0.00	0.00	22.00	0.00
1975	0.00	0.00	125.00	0.00	481.00	0.00	3.00	0.00	0.00	0.00	8.00	0.00
1976	0.00	0.00	231.00	0.00	395.00	9.00	12.00	0.00	0.00	0.00	25.00	0.00
1977	0.00	1.00	36.00	0.00	374.00	3.00	3.00	0.00	0.00	0.00	39.00	0.00
1978	0.00	0.00	292.00	0.00	618.00	436.00	11.00	0.00	0.00	0.00	102.00	0.00
1979	0.00	1.00	474.00	0.00	1150.00	39.00	14.00	0.00	0.00	0.00	422.00	0.00
1980	0.00	0.00	684.00	0.00	3293.00	96.00	17.00	0.00	0.00	0.00	338.00	0.00
1981	0.00	0.00	864.00	0.00	3495.00	21.00	4.00	0.00	0.00	3.00	464.00	0.00
1982	0.00	0.00	824.00	0.00	3832.00	15.00	82.00	0.00	0.00	0.00	382.00	0.00
1983	0.00	0.00	1504.00	0.00	7380.00	10.00	9.00	0.00	0.00	0.00	540.00	0.00
1984	0.00	1.00	852.00	0.00	7367.00	52.00	95.00	2.00	0.00	0.00	2077.00	3.00
1985	0.00	0.00	1845.00	0.00	4571.00	114.00	17.00	0.00	0.00	0.00	547.00	0.00
1986	0.00	0.00	1106.00	0.00	2240.00	40.00	34.00	0.00	0.00	25.00	1552.00	6.00
1987	0.00	19.00	1642.00	0.00	4107.00	37.00	64.00	0.00	0.00	0.00	609.00	0.00
1988	0.00	0.00	1530.00	0.00	5305.00	22.00	70.00	0.00	0.00	2.00	262.00	1.00
1989	0.00	1.00	1732.00	0.00	5997.00	73.00	209.00	1.00	0.00	0.00	108.00	0.00
1990	0.00	1.00	888.00	0.00	5891.00	65.00	192.00	0.00	0.00	0.00	110.00	1.00
1991	0.00	3.00	1273.00	0.00	6848.00	66.00	171.00	3.00	0.00	0.00	62.00	3.00
1992	0.00	8.00	1552.00	0.00	7250.00	359.00	237.00	2.00	0.00	0.00	337.00	0.00
1994	0.00	48.00	1299.00	0.00	5004.00	56.00	237.00	7.00	0.00	0.00	1021.00	19.00
1995	0.00	30.00	1782.00	0.00	2906.00	18.00	199.00	1.00	0.00	1.00	1157.00	0.00
1996	0.00	0.00	1559.00	0.00	2621.00	65.00	583.00	0.00	0.00	0.00	2121.00	4.00
1997	0.00	0.00	1757.00	0.00	3375.00	115.00	3217.00	0.00	0.00	0.00	2483.00	1.00
1998	0.00	2.00	1987.50	0.00	1547.00	26.00	558.50	0.00	0.00	0.00	557.00	0.00
1999	0.00	67.00	1161.00	0.00	1572.00	89.00	300.00	0.00	0.00	0.00	198.00	1.00
2000	2.00	200.00	1465.00	0.00	1981.00	92.00	200.00	0.00	0.00	12.00	470.00	0.00
2001	7.00	4.00	1148.00	0.00	2744.00	138.00	232.00	0.00	0.00	0.00	461.00	3004.00
2002	0.00	10.00	1789.00	0.00	1965.00	569.00	188.00	0.00	0.00	0.00	277.00	1.00
2003	0.00	3.00	1762.00	0.00	1180.00	22.00	175.00	0.00	0.00	0.00	129.00	0.00
2004	0.00	0.00	3576.00	0.00	1831.00	23.00	38.00	0.00	0.00	0.00	245.00	0.00
2005	0.00	0.00	4000.00	1.00	1925.00	74.00	81.00	0.00	0.00	0.00	230.00	0.00
2006	0.00	0.00	6005.00	0.00	3009.00	34.00	100.00	0.00	0.00	0.00	215.00	0.00
2007	0.00	0.00	8859.00	0.00	3487.00	168.00	111.00	2.00	0.00	0.00	209.00	0.00
2008	0.00	0.00	5490.00	0.00	2857.00	105.00	256.00	2.00	0.00	1.00	375.00	7.00
2009	0.00	1.00	5338.00	0.00	2053.00	535.00	400.00	7.00	0.00	0.00	929.00	8.00
2010	0.00	0.00	4851.00	0.00	2080.00	48.00	133.00	0.00	0.00	0.00	606.00	2.00
2011	0.00	0.00	4806.00	0.00	2666.00	46.00	179.00	2.00	0.00	1.00	490.00	8.00
2012	0.00	1.00	9125.46	0.00	5323.21	80.00	133.00	2.00	0.00	0.00	763.33	4.00
2013	0.00	0.00	14086.41	0.00	7191.08	183.01	127.83	0.00	0.00	4.00	632.67	1.00
2014	0.00	20.00	21992.78	0.00	15660.18	638.00	488.00	7.00	0.00	0.00	4737.04	7.00
2015	0.00	8.00	19841.00	0.00	11973.00	442.00	603.00	8.00	0.00	0.00	5527.00	20.00

Table B.1: Two way contingency table of Deaths by year and Weapon type.

	Armed Assault	Assassination	Bombing/Explosion	Facility/Infrastructure Attack	Hijacking	Hostage Taking (Barricade Incident)	Hostage Taking (Kidnapping)	Unarmed Assault	Unknown
1970	36.00	15.00	96.00	9.00	1.00	0.00	10.00	0.00	4.00
1971	14.00	73.00	79.00	1.00	0.00	0.00	2.00	0.00	4.00
1972	58.00	208.00	281.00	1.00	5.00	0.00	12.00	0.00	1.00
1973	41.00	171.00	75.00	0.00	34.00	38.00	11.00	0.00	0.00
1974	34.00	161.00	289.00	0.00	1.00	35.00	14.00	1.00	7.00
1975	225.00	211.00	136.00	0.00	0.00	35.00	8.00	0.00	2.00
1976	160.00	231.00	222.00	6.00	27.00	16.00	8.00	0.00	2.00
1977	141.00	140.00	33.00	2.00	107.00	1.00	22.00	0.00	10.00
1978	318.00	255.00	317.00	434.00	0.00	52.00	25.00	7.00	51.00
1979	599.00	505.00	582.00	44.00	0.00	20.00	81.00	3.00	266.00
1980	2524.00	714.00	829.00	66.00	0.00	9.00	110.00	5.00	171.00
1981	2946.00	465.00	1045.00	39.00	9.00	8.00	51.00	1.00	287.00
1982	3374.00	601.00	862.00	24.00	2.00	26.00	31.00	2.00	213.00
1983	6902.00	482.00	1624.00	19.00	0.00	1.00	29.00	0.00	386.00
1984	6599.00	598.00	1725.00	85.00	28.00	204.00	50.00	0.00	1160.00
1985	4093.00	400.00	1964.00	90.00	83.00	20.00	30.00	5.00	409.00
1986	1920.00	461.00	1144.00	312.00	99.00	2.00	146.00	3.00	916.00
1987	3442.00	704.00	1730.00	9.00	1.00	6.00	68.00	21.00	497.00
1988	4012.00	1255.00	1710.00	28.00	12.00	1.00	79.00	1.00	94.00
1989	4799.00	1400.00	1741.00	16.00	3.00	21.00	101.00	2.00	38.00
1990	4463.00	1178.00	1015.00	10.00	10.00	0.00	395.00	22.00	55.00
1991	5712.00	992.00	1510.00	39.00	5.00	11.00	116.00	5.00	39.00
1992	6154.00	1564.00	1649.00	116.00	3.00	6.00	46.00	5.00	202.00
1994	4375.00	1005.00	1256.00	32.00	18.00	60.00	116.00	80.00	749.00
1995	2256.00	852.00	1707.00	36.00	11.00	236.00	83.00	34.00	879.00
1996	2425.00	785.00	1627.00	16.00	5.00	147.00	95.00	11.00	1842.00
1997	6039.00	696.00	1742.00	48.00	0.00	4.00	125.00	8.00	2286.00
1998	2023.50	40.00	1607.50	39.00	1.00	0.00	338.00	101.00	528.00
1999	1833.00	77.00	1110.00	30.00	7.00	0.00	81.00	68.00	182.00
2000	2008.00	214.00	1256.00	172.00	0.00	0.00	215.00	203.00	354.00
2001	2723.00	145.00	1019.00	133.00	3000.00	17.00	262.00	20.00	419.00
2002	2393.00	110.00	1754.00	1.00	6.00	170.00	106.00	10.00	249.00
2003	1295.00	142.00	1638.00	2.00	0.00	6.00	57.00	13.00	118.00
2004	2103.00	170.00	3086.00	28.00	46.00	26.00	59.00	0.00	195.00
2005	1697.00	346.00	3922.00	16.00	12.00	0.00	141.00	4.00	173.00
2006	2752.00	243.00	5887.00	67.00	2.00	5.00	249.00	2.00	156.00
2007	3183.00	190.00	8717.00	117.00	1.00	10.00	450.00	7.00	161.00
2008	2878.00	365.00	5211.00	37.00	19.00	0.00	336.00	16.00	231.00
2009	2263.00	415.00	5064.00	448.00	2.00	0.00	364.00	9.00	706.00
2010	2014.00	484.00	4361.00	34.00	1.00	141.00	272.00	7.00	406.00
2011	2466.00	481.00	4575.00	46.00	2.00	7.00	276.00	6.00	339.00
2012	5171.29	555.33	8177.30	64.00	0.00	65.00	647.75	8.00	743.33
2013	7331.00	829.00	12258.83	139.00	12.00	262.00	849.50	2.00	542.67
2014	15790.68	992.00	15529.47	351.01	27.00	276.00	7113.00	24.00	3446.84
2015	11702.00	1216.00	17113.00	99.00	56.00	586.00	3175.00	32.00	4443.00

Table B.2: Two way contingency table of Deaths by year and Attack type.

	Dependent variable:
	kills_sum
month_since_1970	−0.0001
	(0.0004)
PeriodInherrentresolve	−0.556***
	(0.081)
PeriodPost_Invasion	2.122***
	(0.249)
PeriodPostPullout	−0.582***
	(0.080)
PeriodPre_Invasion	−1.641**
	(0.751)
PeriodPullout	−1.011***
	(0.072)
PeriodSurge	2.394***
	(0.250)
Period_2ISI driven out of Iraq	−0.773***
	(0.077)
Period_2ISI founded in Iraq	−2.231***
	(0.260)
Period_2ISIS breaking wall cmpgn Start	0.279***
	(0.032)
Period_2ISIS cmpgn Soldier Harvest starts	1.368***
	(0.039)
Period_2Maliki sectarian Policies garnishes support for ISIS	−0.373***
	(0.049)
Period_2None	−0.826***
	(0.077)
Iraq_PresAyad Allawi	0.193
	(0.772)
Iraq_PresHaider al-Abadi	1.976**
	(0.812)
Iraq_PresIbrahim Al Jaafari	0.552
	(0.773)
Iraq_PresIraqi Transition Council	−0.734
	(0.771)
Iraq_PresNot stated	−1.785**
	(0.794)
Iraq_PresNouri al-Malaki	2.267***
	(0.811)
Iraq_PresSaddam Hussein	0.553*
	(0.291)
Constant	3.936***
	(0.804)
Observations	276
Log Likelihood	−8,805.310
Akaike Inf. Crit.	17,652.620
Note:	*p<0.1; **p<0.05; ***p<0.01

Table B.3: Poisson Regression estimates

	Dependent variable:
	kills_sum
month_since_1970	−0.0001
	(0.004)
PeriodInherrentresolve	−0.556
	(0.711)
PeriodPost_Invasion	2.122
	(2.193)
PeriodPostPullout	−0.582
	(0.703)
PeriodPre_Invasion	−1.641
	(6.599)
PeriodPullout	−1.011
	(0.636)
PeriodSurge	2.394
	(2.198)
Period_2ISI driven out of Iraq	−0.773
	(0.679)
Period_2ISI founded in Iraq	−2.231
	(2.283)
Period_2ISIS breaking wall cmpgn Start	0.279
	(0.278)
Period_2ISIS cmpgn Soldier Harvest starts	1.368***
	(0.346)
Period_2Maliki sectarian Policies garnishes support for ISIS	−0.373
	(0.429)
Period_2None	−0.826
	(0.678)
Iraq_PresAyad Allawi	0.193
	(6.792)
Iraq_PresHaider al-Abadi	1.976
	(7.138)
Iraq_PresIbrahim Al Jaafari	0.552
	(6.799)
Iraq_PresIraqi Transition Council	−0.734
	(6.777)
Iraq_PresNot stated	−1.785
	(6.981)
Iraq_PresNouri al-Malaki	2.267
	(7.129)
Iraq_PresSaddam Hussein	0.553
	(2.563)
Constant	3.936
	(7.070)
Observations	276
Note:	*p<0.1; **p<0.05; ***p<0.01

Table B.4: quasi-Poisson Regression estimates

	Dependent variable:
	kills_sum
month_since_1970	0.001
	(0.001)
PeriodInherrentresolve	−0.571
	(1.271)
PeriodPost_Invasion	2.160*
	(1.299)
PeriodPostPullout	−0.589
	(1.230)
PeriodPre_Invasion	−1.595
	(1.799)
PeriodPullout	−1.013
	(1.101)
PeriodSurge	2.426*
	(1.380)
Period_2ISI driven out of Iraq	−0.748
	(1.151)
Period_2ISI founded in Iraq	−2.228
	(1.751)
Period_2ISIS breaking wall cmpgn Start	0.267
	(0.444)
Period_2ISIS cmpgn Soldier Harvest starts	1.348**
	(0.633)
Period_2Maliki sectarian Policies garnishes support for ISIS	−0.361
	(0.528)
Period_2None	−0.810
	(1.156)
Iraq_PresAyad Allawi	−0.055
	(1.494)
Iraq_PresHaider al-Abadi	1.718
	(1.924)
Iraq_PresIbrahim Al Jaafari	0.295
	(1.494)
Iraq_PresIraqi Transition Council	−0.974
	(1.431)
Iraq_PresNot stated	−2.037
	(1.775)
Iraq_PresNouri al-Malaki	2.015
	(1.933)
Iraq_PresSaddam Hussein	0.402
	(0.694)
Constant	3.857*
	(2.193)
Observations	276
Log Likelihood	−1,484.564
θ	1.035*** (0.101)
Akaike Inf. Crit.	3,011.127
Note:	*p<0.1; **p<0.05; ***p<0.01

Table B.5: Negative binomial regression estimates

	Dependent variable:
	log(kills_sum + 1)
month_since_197503	0.004^{***}
	(0.001)
PeriodInherrentresolve	−1.772
	(1.262)
PeriodPost_Invasion	2.293^{*}
	(1.268)
PeriodPostPullout	−1.800
	(1.221)
PeriodPre_Invasion	−0.587
	(1.627)
PeriodPullout	-2.017^{*}
	(1.093)
PeriodSurge	2.539^{*}
	(1.350)
Period_2ISI driven out of Iraq	−1.811
	(1.142)
Period_2ISI founded in Iraq	-3.152^{*}
	(1.723)
Period_2ISIS breaking wall cmpgn Start	0.342
	(0.440)
Period_2ISIS cmpgn Soldier Harvest starts	1.513^{**}
	(0.629)
Period_2Maliki sectarian Policies garnishes support for ISIS	-1.133^{**}
	(0.523)
Period_2None	−1.663
	(1.148)
Iraq_PresAyad Allawi	0.130
	(1.269)
Iraq_PresHaider al-Abadi	1.873
	(1.733)
Iraq_PresIbrahim Al Jaafari	0.340
	(1.269)
Iraq_PresIraqi Transition Council	−1.142
	(1.197)
Iraq_PresNot stated	−1.779
	(1.577)
Iraq_PresNouri al-Malaki	2.128
	(1.744)
Iraq_PresSaddam Hussein	−0.637
	(0.625)
Constant	3.247
	(2.030)
Observations	276
R^2	0.850
Adjusted R^2	0.838
Residual Std. Error	0.978 (df = 255)
F Statistic	72.304^{***} (df = 20; 255)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table B.6: Linear regression model estimates.

	Dependent variable:
	log(kills_sum + 1)
month_since_197503	0.004^{***}
	(0.001)
PeriodInherrentresolve	-1.772
	(1.262)
PeriodPost_Invasion	2.293^{*}
	(1.268)
PeriodPostPullout	-1.800
	(1.221)
PeriodPre_Invasion	-0.587
	(1.627)
PeriodPullout	-2.017^{*}
	(1.093)
PeriodSurge	2.539^{*}
	(1.350)
Period_2ISI driven out of Iraq	-1.811
	(1.142)
Period_2ISI founded in Iraq	-3.152^{*}
	(1.723)
Period_2ISIS breaking wall cmpgn Start	0.342
	(0.440)
Period_2ISIS cmpgn Soldier Harvest starts	1.513^{**}
	(0.629)
Period_2Maliki sectarian Policies garnishes support for ISIS	-1.133^{**}
	(0.523)
Period_2None	-1.663
	(1.148)
Iraq_PresAyad Allawi	0.130
	(1.269)
Iraq_PresHaider al-Abadi	1.873
	(1.733)
Iraq_PresIbrahim Al Jaafari	0.340
	(1.269)
Iraq_PresIraqi Transition Council	-1.142
	(1.197)
Iraq_PresNot stated	-1.779
	(1.577)
Iraq_PresNouri al-Malaki	2.128
	(1.744)
Iraq_PresSaddam Hussein	-0.637
	(0.625)
Constant	3.247
	(2.030)
Observations	276
R^2	0.850
Adjusted R^2	0.838
Residual Std. Error	0.978 (df = 255)
F Statistic	72.304^{***} (df = 20; 255)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table B.7: Robust regression estimates.

	Variable	F-Statistic	p value
1	(Intercept)	2.02359823810467	0.156093879216133
2	month_since_197503	36.0776198110256	6.50364162802437e-09
3	PeriodInherrentresolve	0.414977642935234	0.52003281066751
4	PeriodPost_Invasion	7.98158258115534	0.00509935808610932
5	PeriodPostPullout	0.495448791030915	0.482148984041935
6	PeriodPre_Invasion	0.423624669339769	0.515719967204547
7	PeriodPullout	1.82680821883129	0.177702751846541
8	PeriodSurge	7.82192873483884	0.00555489425458082
9	Period_2ISI driven out of Iraq	0.408322169955938	0.523395889899178
10	Period_2ISI founded in Iraq	3.33519486191744	0.0689816032713713
11	Period_2ISIS breaking wall cmpgn Start	0.515202644885829	0.473552347608327
12	Period_2ISIS cmpgn Soldier Harvest starts	7.00367404446488	0.00864025159758107
13	Period_2Maliki sectarian Policies garnishes support for ISIS	0.219122961166581	0.64010883757994
14	Period_2None	0.796198164498034	0.373073733563684
15	Iraq_PresAyad Allawi	0.0264225397673897	0.871001402559836
16	Iraq_PresHaider al-Abadi	1.78888890793696	0.182253313611277
17	Iraq_PresIbrahim Al Jaafari	0.012027010991077	0.912759148623237
18	Iraq_PresIraqi Transition Council	2.493262194502	0.115574154436575
19	Iraq_PresNot stated	3.9611969609461	0.0476282486970668
20	Iraq_PresNouri al-Malaki	2.34609447269646	0.126836955725164
21	Iraq_PresSaddam Hussein	3.01154659984654	0.0838814672952424

Table B.8: Robust regression robust F-test results.

Appendix C

Chapter 5 outbreak and outlier detection detected aberrations

year	week	sum_kill	weekstrdate	wkyr
2003.00	15.00	2.00	1050274800.00	15-2003
2003.00	31.00	0.00	1059951600.00	31-2003
2003.00	35.00	100.00	1062370800.00	35-2003
2003.00	42.00	0.00	1066604400.00	42-2003
2003.00	47.00	26.00	1069632000.00	47-2003
2004.00	4.00	10.00	1075075200.00	4-2004
2004.00	9.00	168.00	1078099200.00	9-2004
2007.00	2.00	28.00	1168819200.00	2-2007
2007.00	23.00	242.00	1181516400.00	23-2007
2012.00	52.00	6.00	1356307200.00	52-2012
2014.00	22.00	147.00	1401663600.00	22-2014
2014.00	26.00	195.00	1404082800.00	26-2014

Table C.1: Identified breakouts of deaths due to terrorism in Iraq post invasion

X_transform	L_transform	S_transform	E_transform	wkyr
240.00	101.21	37.59	101.20	18-2005
194.00	84.37	8.43	101.20	37-2005
203.00	90.13	11.67	101.20	9-2006
195.00	88.31	5.49	101.20	14-2006
207.00	102.28	3.52	101.20	29-2006
325.00	94.50	129.30	101.20	47-2006
289.00	105.05	82.75	101.20	49-2006
259.00	110.48	47.32	101.20	5-2007
323.00	100.38	121.42	101.20	10-2007
707.00	109.40	496.40	101.20	13-2007
255.00	109.82	43.98	101.20	16-2007
242.00	100.38	40.42	101.20	23-2007
268.00	97.75	69.05	101.20	27-2007
517.00	82.64	333.16	101.20	33-2007
212.00	82.62	28.18	101.20	17-2009
187.00	78.79	7.01	101.20	25-2009
239.00	96.05	41.75	101.20	20-2013
227.00	118.54	7.26	101.20	28-2013
248.00	113.07	33.73	101.20	32-2013
225.00	109.61	14.19	101.20	37-2013
228.00	110.01	16.79	101.20	50-2013
223.00	114.77	7.03	101.20	8-2014
264.00	122.33	40.47	101.20	10-2014
264.00	119.12	43.68	101.20	12-2014
227.00	114.71	11.09	101.20	19-2014
1045.00	123.45	820.35	101.20	23-2014
1898.00	120.26	1676.54	101.20	24-2014
363.00	136.28	125.52	101.20	25-2014
248.00	134.34	12.46	101.20	27-2014
903.00	112.38	689.42	101.20	31-2014
231.00	110.92	18.88	101.20	34-2014
266.00	108.71	56.09	101.20	36-2014
226.00	114.03	10.77	101.20	40-2014
284.00	131.53	51.27	101.20	41-2014
247.00	129.93	15.87	101.20	42-2014
472.00	126.06	244.74	101.20	44-2014
291.00	118.22	71.58	101.20	50-2014
243.00	105.15	36.65	101.20	2-2015
212.00	108.51	2.29	101.20	9-2015
503.00	99.53	302.27	101.20	15-2015
225.00	113.87	9.93	101.20	22-2015
256.00	107.25	47.55	101.20	27-2015
262.00	130.61	30.19	101.20	28-2015
309.00	129.04	78.76	101.20	29-2015
237.00	128.88	6.92	101.20	30-2015
252.00	111.99	38.81	101.20	33-2015
282.00	107.73	73.07	101.20	46-2015

Table C.2: Output of RAD detection algorithm

observed	epoch	state	alarm	upperbound	population	freq	epochInPeriod	time
100.00	1062370778.00	FALSE	TRUE	81.26	1.00	52.00	0.90	2003-08-25
1.00	1062975578.00	FALSE	TRUE	0.00	1.00	52.00	0.67	2003-09-01
6.00	1063580378.00	FALSE	TRUE	0.00	1.00	52.00	0.44	2003-09-08
118.00	1076284778.00	FALSE	TRUE	34.16	1.00	52.00	0.83	2004-02-02
19.00	1076889578.00	FALSE	TRUE	0.00	1.00	52.00	0.60	2004-02-09
10.00	1077494378.00	FALSE	TRUE	0.00	1.00	52.00	0.37	2004-02-16
28.00	1079308778.00	FALSE	TRUE	0.00	1.00	52.00	0.67	2004-03-08
7.00		FALSE	TRUE	5.99	1.00	52.00		2004-12-27
19.00	1110758378.00	FALSE	TRUE	0.00	1.00	52.00	0.67	2005-03-07
158.00	1115593178.00	FALSE	TRUE	102.79	1.00	52.00	0.60	2005-05-02
73.00	1116197978.00	FALSE	TRUE	0.00	1.00	52.00	0.37	2005-05-09
203.00	1140998378.00	FALSE	TRUE	152.51	1.00	52.00	0.13	2006-02-20
46.00	1141603178.00	FALSE	TRUE	0.00	1.00	52.00	0.90	2006-02-27
83.00	1142207978.00	FALSE	TRUE	16.50	1.00	52.00	0.67	2006-03-06
80.00	1164585578.00	FALSE	TRUE	0.00	1.00	52.00	0.13	2006-11-20
289.00	1165190378.00	FALSE	TRUE	196.96	1.00	52.00	0.90	2006-11-27
137.00	1165795178.00	FALSE	TRUE	55.46	1.00	52.00	0.67	2006-12-04
50.00	1166399978.00	FALSE	TRUE	0.00	1.00	52.00	0.44	2006-12-11
60.00	1176073178.00	FALSE	TRUE	0.00	1.00	52.00	0.52	2007-04-02
137.00	1176677978.00	FALSE	TRUE	0.00	1.00	52.00	0.29	2007-04-09
5.00	1235347178.00	FALSE	TRUE	0.00	1.00	52.00	0.13	2009-02-16
21.00	1235951978.00	FALSE	TRUE	8.75	1.00	52.00	0.90	2009-02-23
19.00	1241391578.00	FALSE	TRUE	0.00	1.00	52.00	0.60	2009-04-27
10.00	1241996378.00	FALSE	TRUE	0.00	1.00	52.00	0.37	2009-05-04
29.00	1280703578.00	FALSE	TRUE	16.91	1.00	52.00	0.60	2010-07-26
94.00	1295827178.00	FALSE	TRUE	90.38	1.00	52.00	0.06	2011-01-17
17.00	1296431978.00	FALSE	TRUE	0.00	1.00	52.00	0.83	2011-01-24
67.00	1323043178.00	FALSE	TRUE	62.86	1.00	52.00	0.67	2011-11-28
24.00	1323647978.00	FALSE	TRUE	0.08	1.00	52.00	0.44	2011-12-05
106.00	1326067178.00	FALSE	TRUE	51.38	1.00	52.00	0.52	2012-01-02
61.00	1326671978.00	FALSE	TRUE	0.00	1.00	52.00	0.29	2012-01-09
109.00	1327276778.00	FALSE	TRUE	91.72	1.00	52.00	0.06	2012-01-16
92.00	1343602778.00	FALSE	TRUE	70.18	1.00	52.00	0.60	2012-07-23
63.00	1344207578.00	FALSE	TRUE	21.38	1.00	52.00	0.37	2012-07-30
86.00	1367794778.00	FALSE	TRUE	84.48	1.00	52.00	0.37	2013-04-29
40.00	1368399578.00	FALSE	TRUE	35.73	1.00	52.00	0.13	2013-05-06
140.00	1369609178.00	FALSE	TRUE	45.56	1.00	52.00	0.67	2013-05-20
151.00	1370213978.00	FALSE	TRUE	6.22	1.00	52.00	0.44	2013-05-27
157.00	1370818778.00	FALSE	TRUE	59.79	1.00	52.00	0.21	2013-06-03
94.00	1371423578.00	FALSE	TRUE	48.96	1.00	52.00	0.98	2013-06-10
188.00	1375052378.00	FALSE	TRUE	185.63	1.00	52.00	0.60	2013-07-22
1898.00	1402873178.00	FALSE	TRUE	1161.16	1.00	52.00	0.98	2014-06-09
363.00	1403477978.00	FALSE	TRUE	0.00	1.00	52.00	0.75	2014-06-16
195.00	1404082778.00	FALSE	TRUE	0.00	1.00	52.00	0.52	2014-06-23
248.00	1404687578.00	FALSE	TRUE	0.00	1.00	52.00	0.29	2014-06-30
201.00	1429484378.00	FALSE	TRUE	41.23	1.00	52.00	0.83	2015-04-13
185.00	1430089178.00	FALSE	TRUE	16.89	1.00	52.00	0.60	2015-04-20

Table C.3: A table of alarms raised by the EARSC3 algorithm

	observed	upperbound	alarm	time
1	100.00	47.13	1.00	2003-08-31
2	118.00	89.24	1.00	2004-02-02
3	118.00	88.41	1.00	2004-02-09
4	168.00	105.34	1.00	2004-03-01
5	145.00	112.50	1.00	2004-06-27
6	240.00	179.97	1.00	2005-05-01
7	194.00	164.46	1.00	2005-09-11
8	203.00	190.05	1.00	2006-02-27
9	325.00	271.64	1.00	2006-11-20
10	707.00	338.27	1.00	2007-04-01
11	517.00	268.43	1.00	2007-08-19
12	212.00	161.05	1.00	2009-04-26
13	187.00	148.97	1.00	2009-06-21
14	143.00	135.01	1.00	2009-08-16
15	166.00	121.27	1.00	2009-10-26
16	163.00	115.78	1.00	2009-12-07
17	138.00	133.96	1.00	2010-11-01
18	123.00	113.05	1.00	2011-01-17
19	155.00	112.83	1.00	2012-01-02
20	125.00	111.53	1.00	2012-06-10
21	161.00	114.48	1.00	2012-07-22
22	128.00	112.69	1.00	2012-08-12
23	1045.00	490.19	1.00	2014-06-08
24	1898.00	491.25	1.00	2014-06-15
25	903.00	499.67	1.00	2014-08-03
26	472.00	349.86	1.00	2014-11-03

Table C.4: A table of alarms raised by the Farrington algorithm

Appendix D

Code repos

All code is maintained through github at :`https://github.com/brennap3/thesis_2`

Bibliography

M. Abrahms. Why terrorism does not work. *International Security*, 31(2):42–78, 2006.

Z. Abuza. Funding terrorism in southeast asia: the financial network of al qaeda and jemaah islamiya. *Contemporary Southeast Asia*, pages 169–199, 2003.

G. A. Ackerman and L. E. Pinson. Speaking truth to sources: Introducing a method for the quantitative evaluation of open sources in event data. *Studies in Conflict & Terrorism*, 2016.

M. M. Aid. All glory is fleeting: Sigint and the fight against international terrorism. *Intelligence and National Security*, 18(4):72–120, 2003.

J. Allanach, H. Tu, S. Singh, P. Willett, and K. Pattipati. Detecting, tracking, and counteracting terrorist networks via hidden markov models. In *Aerospace Conference, 2004. Proceedings. 2004 IEEE*, volume 5. IEEE, 2004.

A. Alsema. The farcs biggest fear: Colombias paramilitary groups. `http://colombiareports.com/the-farcs-biggest-fear-colombias-paramilitary-groups/`, a. Accessed: 2016-10-22.

A. Alsema. Regime detection. `https://systematicinvestor.wordpress.com/2012/11/01/regime-detection/`, b. Accessed: 2016-10-22.

J. Argomaniz and A. Vidal-Diez. Examining deterrence and backlash effects in counter-terrorism: The case of eta. *Terrorism and Political Violence*, 27(1): 160–181, 2015.

S. Atran. Trends in suicide terrorism: Sense and nonsense. *World Federation of Scientists Permanent Monitoring Panel on terrorism, Erice, Sicily*, 204, 2004.

S. Austin, R. Schwartz, and P. Placeway. The forward-backward search algorithm. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pages 697–700. IEEE, 1991.

S. Axelsson. The base-rate fallacy and the difficulty of intrusion detection. *ACM Transactions on Information and System Security (TISSEC)*, 3(3):186–205, 2000.

G. I. Bae, W. C. Kim, and J. M. Mulvey. Dynamic asset allocation for varied financial markets under regime switching framework. *European Journal of Operational Research*, 234(2):450–458, 2014.

M. Bar-Hillel. The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3):211–233, 1980.

A. Belasco. Troop levels in the afghan and iraq wars, fy2001-fy2012: Cost and other potential issues. DTIC Document, 2009.

Y. Ben-Itzhak. Organised cybercrime and payment cards. *Card Technology Today*, 21(2):10–11, 2009.

S. Ben Salem and S. Naouali. Pattern recognition approach in multidimensional databases: Application to the global terrorism database. *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, 7(8):280–286, 2016.

B. T. Bennett. *Understanding, assessing, and responding to terrorism: Protecting critical infrastructure and personnel.* John Wiley & Sons, 2007.

M. Berlingerio, M. Coscia, and F. Giannotti. Finding and characterizing communities in multidimensional networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 490–494. IEEE, 2011.

M. J. Berry and G. Linoff. *Data mining techniques: for marketing, sales, and customer support.* John Wiley & Sons, Inc., 1997.

F. Bignami. European versus american liberty: a comparative privacy analysis of anti-terrorism data-mining. *Boston College Law Review*, 48:609, 2007.

R. Blakeley. *State terrorism and neoliberalism: The north in the south.* Routledge, 2009.

B. Blakemore. *Policing Cyber Hate, Cyber Threats and Cyber Terrorism.* Routledge, 2016.

P. Bobbitt. Terror and consent, wars for the 21st century. *Allen Lane, London*, 2008.

M. Bostock. D3. js. *Data Driven Documents*, 2012.

F. S. Bresler. *Interpol.* Vintage, 1992.

D. Brown, J. Dalton, and H. Hoyle. Spatial forecast methods for terrorist events in urban environments. In *International Conference on Intelligence and Security Informatics*, pages 426–435. Springer, 2004.

J. Burke. Al qaeda. *Foreign Policy*, pages 18–26, 2004.

A. B. Calahan. Countering terrorism: The israeli response to the 1972 munich olympic massacre and the development of independent covert action teams. *The Marine Corps Command and Staff College*, 1995.

C. Chabot, C. Stolte, and P. Hanrahan. Tableau software. *Tableau Software*, 2003.

P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. Crisp-dm 1.0 step-by-step data mining guide. 2000.

E. Chase. Defining terrorism: A strategic imperative. *Small Wars Journal, January*, 24, 2013.

S. Chatterjee and A. S. Hadi. *Sensitivity analysis in linear regression*, volume 327. John Wiley & Sons, 2009.

M. Chau, G. A. Wang, and H. Chen. Intelligence and security informatics: Pacific asia workshop, paisi 2015, ho chi minh city, vietnam, may 19, 2015. proceedings. *Lecture Notes in Computer Science*, 2015.

J. Cheng and Y. Xie. *leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library*, 2016. URL `https://CRAN.R-project.org/package=leaflet`. R package version 1.0.1.

A. Clauset and M. Young. Scale invariance in global terrorism. *arXiv preprint physics/0502014*, 2005.

C. Conetta. Strange victory: a critical appraisal of operation enduring freedom and afghanistan war. *Afghan Digital Libraries*, 2002.

J. Coscarelli. Nypds domain awareness system is watching you. *New York*, 2012.

M. Crenshaw. The causes of terrorism. *Comparative politics*, 13(4):379–399, 1981.

A. K. Cronin, H. Aden, A. Frost, and B. Jones. Foreign terrorist organizations. DTIC Document, 2004.

H. Dabashi. The arab spring: the end of postcolonialism, 2012.

D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

G. Dawson. Trauma, place and the politics of memory: Bloody sunday, derry, 1972–2004. In *History Workshop Journal*, volume 59, pages 151–178. Oxford Univ Press, 2005.

B. De Graaf. Counter-narratives and the unrehearsed stories counter-terrorists unwittingly produce. *Perspectives on Terrorism*, 3(2), 2010.

C. L. DEIBEL et al. Nsa data collection program: The challenge of assesing effectiveness. 2016.

M. DeRosa. *Data mining and data analysis for counterterrorism.* CSIS Press, 2004.

A. M. Dershowitz. *Why terrorism works: Understanding the threat, responding to the challenge.* Yale University Press, 2002.

C. J. Drake. The role of ideology in terrorists target selection. *Terrorism and Political Violence*, 10(2):53–85, 1998.

K. Duffy. Colombias right-wing terror. `https://www.jacobinmag.com/2016/05/colombia-peace-farc-urabenos-santos-uribe-up/`. Accessed: 2016-10-22.

L. Dugan, G. LaFree, and H. Fogg. A first look at domestic and international global terrorism events, 1970–1997. In *International Conference on Intelligence and Security Informatics*, pages 407–419. Springer, 2006.

J. Durrand, A. Batterham, and G. Danjoux. Pre-habilitation (i): aggregation of marginal gains. *Anaesthesia*, 69(5):403–406, 2014.

R. East. *Keesing's Record of World Events 2016.* Longman, 2016.

A. Edwards. *The Northern Ireland Troubles: Operation Banner 1969–2007.* Bloomsbury Publishing, 2011.

M. El-Nawawy. Terrorist or freedom fighter? the arab media coverage of terrorism or so-called terrorism. *Global Media Journal*, 2004, 2014.

W. Enders and T. Sandler. The effectiveness of antiterrorism policies: A vector-autoregression-intervention analysis. *American Political Science Review*, 87 (04):829–844, 1993.

W. Enders and T. Sandler. *The political economy of terrorism.* Cambridge University Press, 2011.

W. Enders, T. Sandler, and K. Gaibulloev. Domestic versus transnational terrorism: Data, decomposition, and dynamics. *Journal of Peace Research*, 48 (3):319–337, 2011.

J. O. Engene. Five decades of terrorism in europe: The tweed dataset. *Journal of Peace Research*, 44(1):109–121, 2007.

C. Farrington, N. J. Andrews, A. Beale, and M. Catchpole. A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 547–563, 1996.

J. D. Fearon. Iraq's civil war. *Foreign Aff.*, 86:2, 2007.

H. Fischer. A guide to us military casualty statistics: Operation inherent resolve, operation new dawn, operation iraqi freedom, and operation enduring freedom. 2014. *Congressional Research Service http://www. fas. org/sgp/crs/-natsec/RS22452. pdf Google Scholar*, 2015.

fivethirtyeight.com. Counter terrorism statistics. ,`http://fivethirtyeight.com/features/the-paris-attacks-are-just-a-few-of-125000-entries-in-the-global-terrorism-database/`, 2015. Accessed: 2016-16-11.

A. Fraser. *The gunpowder plot: Terror and faith in 1605*. Hachette UK, 2010.

R. D. Fricker, B. L. Hegler, and D. A. Dunfee. Comparing syndromic surveillance detection methods: Ears'versus a cusum-based methodology. 2008.

K. Gaibulloev, T. Sandler, and C. Santifort. Assessing the evolving threat of terrorism. *Global Policy*, 3(2):135–144, 2012. ISSN 1758-5899. doi: 10.1111/j.1758-5899.2011.00142.x. URL `http://dx.doi.org/10.1111/j.1758-5899.2011.00142.x`.

M. J. Gales. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer speech & language*, 12(2):75–98, 1998.

B. Ganor. Defining terrorism: Is one man's terrorist another man's freedom fighter? *Police Practice and Research*, 3(4):287–304, 2002.

F. H. Gareau. *State terrorism and the United States: from counterinsurgency to the war on terrorism*. Zed Books, 2004.

D. J. Garrow. *The FBI and Martin Luther King, Jr.: From" Solo" to Memphis*. Open Road Media, 2015.

R. M. Gates. A balanced strategy: Reprogramming the pentagon for a new age. *Foreign Affairs*, pages 28–40, 2009.

M. C. Gazette. The changing face of war: Into the fourth generation william s. lind, colonel keith nightengale (usa), captain john f. schmitt (usmc), colonel joseph w. sutton (usa), and lieutenant colonel gary i. wilson (usmcr). *Marine Corps Gazette*, pages 22–26, 1989.

B. Geddes. How the cases you choose affect the answers you get: Selection bias in comparative politics. *Political analysis*, 2(1):131–150, 1990.

J. P. Gibbs. Conceptualization of terrorism. *American Sociological Review*, pages 329–340, 1989.

P. H. Gleick. Water and terrorism. *Water policy*, 8(6):481–503, 2006.

Y. Golandsky and A. R. Dombe. A review and analysis of the world of cyber terrorism. Technical report, CybeRisk, 2016.

N. Gordon. Israels emergence as a homeland security capital. *Surveillance and Control in Israel/Palestine: Population, Territory and Power*, pages 134–153, 2011.

S. Graham. The urban battlespace. *Theory, Culture & Society*, 26(7-8):278–288, 2009.

G. Grolemund, H. Wickham, et al. Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3):1–25, 2011.

M. L. Gross, D. Canetti, and D. R. Vashdi. Cyber terrorism: Its effects on psychological well being, public confidence and political attitudes. 2016.

M. Halberstam. Terrorism on the high seas: the achille lauro, piracy and the imo convention on maritime safety. *American Journal of International Law*, pages 269–310, 1988.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

D. W. Hamilton. *The art of insurgency: American military policy and the failure of strategy in Southeast Asia.* Greenwood Publishing Group, 1998.

A. Hashim. *Insurgency and Counter-insurgency in Iraq.* Cornell University Press, 2005.

D. Hauser. *Baader und Herold: Beschreibung eines Kampfes.* Alexander Fest, 1997.

M. . Haykal. *Autumn of fury: the assassination of Sadat.* Random House (NY), 1983.

T. Hegghammer. Terrorist recruitment and radicalization in saudi arabia. *Middle East Policy*, 13(4):39, 2006.

C. Henzel. The origins of al qaeda's ideology: implications for us strategy. Technical report, DTIC Document, 2005.

D. P. Hepworth. Analysis of al-qaeda terrorist attacks to investigate rational action. *Perspectives on Terrorism*, 7(2), 2013.

P. Hillyard. *Suspect community: people's experience of the Prevention of Terrorism Acts in Britain.* Pluto Pr, 1993.

M. Höhle. : An r package for the monitoring of infectious diseases. *Computational Statistics*, 22(4):571–582, 2007.

R. A. Horsley. The sicarii: Ancient jewish" terrorists". *The journal of religion*, 59(4):435–458, 1979.

F. Husson, S. Lê, and J. Pagès. *Exploratory multivariate analysis by example using R.* CRC press, 2010.

IBM. Intelligence analysis in a dedicated environment for information fusion and sharing. ,`http://www-03.ibm.com/software/products/en/i2-analyze`, 2016. Accessed: 2016-16-11.

H. Ito and D. Lee. Assessing the impact of the september 11 terrorist attacks on us airline demand. *Journal of Economics and Business*, 57(1):75–95, 2005.

A. Iyād and E. Rouleau. *My home, my land: a narrative of the Palestinian struggle.* Times Books (NY), 1981.

B. A. Jackson. Groups, networks, or movements: a command-and-control-driven approach to classifying terrorist organizations and its application to al qaeda. *Studies in Conflict & Terrorism*, 29(3):241–262, 2006.

R. Jackson. *Writing the war on terrorism: Language, politics and counter-terrorism.* Manchester University Press, 2005.

N. A. James, A. Kejariwal, and D. S. Matteson. Leveraging cloud data to mitigate user experience from" breaking bad". *arXiv preprint arXiv:1411.7955*, 2014.

N. Jamwal. Counter terrorism strategy. *Strategic Analysis*, 27(1):56–78, 2003.

L. Jarvis and M. Lister. *Critical Perspectives on Counter-terrorism.* Routledge, 2014.

A. Jazić. Rise and fall of left wing terrorism. *Medjunarodni problemi*, 65(2): 238–269, 2013.

W. Jeberson and L. Sharma. Survey on big data for counter terrorism. 2015.

B. M. Jenkins. International terrorism: The other world war. Technical report, DTIC Document, 1985.

D. Jensen, M. Rattigan, and H. Blau. Information awareness: A prospective technical assessment. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 378–387. ACM, 2003.

M. Jizba, R. L. Cheu, T. Horak, and H. Binova. Analysis of screening checkpoint operations for transatlantic container transportation. *Journal of Transportation Security*, 8(3-4):79–97, 2015.

K. Joffres, M. Bouchard, R. Frank, and B. Westlake. Strategies to disrupt online child pornography networks. In *Intelligence and Security Informatics Conference (EISIC), 2011 European*, pages 163–170. IEEE, 2011.

C. Johns, R. A. Shellie, O. G. Potter, J. W. OReilly, J. P. Hutchinson, R. M. Guijt, M. C. Breadmore, E. F. Hilder, G. W. Dicinoski, and P. R. Haddad. Identification of homemade inorganic explosives by ion chromatographic analysis of post-blast residues. *Journal of Chromatography A*, 1182(2):205–214, 2008.

J. Jonas and J. Harper. *Effective counterterrorism and the limited role of predictive data mining.* Cato Institute, 2006.

JPOST.COM. Report: Israel bias suspected behind axed tech deal that could have flagged paris attackers. `http://www.jpost.com/Israel-News/Report-Anti-Israel-views-suspected-in-nixed-terror-tech-deal-before-Paris-at`, 2016. Accessed: 2016-10-27.

P. R. Keefe. Can network theory thwart terrorists? *The New York Times*, 12, 2006.

S. Kelly and K. Ahmad. Propagating disaster warnings on social and digital media. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 475–484. Springer, 2015.

R. Kennedy. Is one person's terrorist another's freedom fighter? western and islamic approaches to just warcompared. *Terrorism and Political Violence*, 11(1):1–21, 1999.

M. Khorshid, T. Abou-El-Enien, and G. Soliman. A comparison among support vector machine and other machine learning classification algorithms. *IPASJ International Journal of Computer Science*, 3(5):26–35, 2015.

C. Kleiber and A. Zeileis. *Applied Econometrics with R.* Springer-Verlag, New York, 2008. URL `https://CRAN.R-project.org/package=AER`. ISBN 978-0-387-77316-2.

T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, et al. Jupyter notebooksa publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, page 87, 2016.

K. Koonings and D. Krujit. *Societies of fear: the legacy of civil war, violence and terror in Latin America.* Zed Books, 1999.

C. Kopp. Technology of improvised explosive devices. *Defence Today*, 4649: 46–49, 2008.

S. Koschade. A social network analysis of jemaah islamiyah: The applications to counterterrorism and intelligence. *Studies in Conflict & Terrorism*, 29(6): 559–575, 2006.

V. Krebs. Uncloaking terrorist networks. *First Monday*, 7(4), 2002a.

V. E. Krebs. Mapping networks of terrorist cells. *Connections*, 24(3):43–52, 2002b.

Y. Kuoti. Exclusion and violence in post-2003 iraq. *Journal of International Affairs*, 69(2):19, 2016.

A. H. Kydd and B. F. Walter. The strategies of terrorism. *International Security*, 31(1):49–80, 2006.

A. M. Ladd, K. E. Bekris, A. Rudys, L. E. Kavraki, and D. S. Wallach. Robotics-based location sensing using wireless ethernet. *Wireless Networks*, 11(1-2): 189–204, 2005.

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

G. LaFree. *Building a global terrorism database.* DIANE Publishing, 2011.

G. LaFree. Generating terrorism event databases: Results from the global terrorism database, 1970 to 2008. In *Evidence-based counterterrorism policy*, pages 41–64. Springer, 2012.

G. LaFree and L. Dugan. Introducing the global terrorism database. *Terrorism and Political Violence*, 19(2):181–204, 2007.

G. LaFree and L. Dugan. Global terrorism and the deadliest groups since 2001. *Peace and Conflict 2016*, page 67, 2016.

G. LaFree, L. Dugan, and R. Korte. The impact of british counterterrorist strategies on political violence in northern ireland: Comparing deterrence and backlash models. *Criminology*, 47(1):17–45, 2009.

P. Lance. Triple cross: How bin laden\'s master spy penetrated the cia, the green berets, and the fbi–and why patrick fitzgerald failed to stop him. 2006.

S. Lê, J. Josse, and F. Husson. FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18, 2008. doi: 10.18637/jss.v025.i01.

J. Lee. Exploring global terrorism data: a web-based visualization of temporal data. *Crossroads*, 15(2):7–14, 2008.

M. K. Leighton. Strange bedfellows: The stasi and the terrorists. *International Journal of Intelligence and CounterIntelligence*, 27(4):647–665, 2014.

I. Lesser, J. Arquilla, B. Hoffman, D. F. Ronfeldt, and M. Zanini. *Countering the new terrorism.* RAND corporation, 1999.

F. J. Llera, J. M. Mata, and C. L. Irvin. Eta: From secret army to social movement–the post-franco schism of the basque nationalist movement. *Terrorism and Political Violence*, 5(3):106–134, 1993.

C. Lum, L. W. Kennedy, and A. Sherley. 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, 2006.

J. M. Lutz and B. J. Lutz. How successful is terrorism? In *Forum on Public Policy*, volume 11, 2009.

D. E. Maldonado. *Leading and Learning: Understanding and Reducing Intelligence Leadership Failures*. PhD thesis, 2015.

B. Margulies. Trump doesnt really mean what he says. hes just trying to change the subject. *USApp–American Politics and Policy Blog*, 2016.

T. Marks. Colombian army adaptation to farc insurgency. Technical report, DTIC Document, 2002.

E. F. Mickolus and P. A. Flemming. Iterate international terrorism: Attributes of terrorist events. 2013.

P. N. Misra, R. I. Abbot, and E. Gaposcbkin. Integrated use of gps and glonass: Transformation between wgs 84 and pz-90. In *Proceedings of the 9th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 1996)*, pages 307–314, 1996.

S. Moeller. *Packaging terrorism: Co-opting the news for politics and profit.* John Wiley & Sons, 2009.

R. J. Mooney, P. Melville, L. R. Tang, J. Shavlik, I. d. Castro Dutro, D. Page, and V. S. Costa. Relational data mining with inductive logic programming for link discovery. Technical report, DTIC Document, 2002.

M. J. Morgan. The origins of the new terrorism. Technical report, DTIC Document, 2004.

C. Morselli. *Inside criminal networks.* Springer.

E. Moxon-Browne. *Nation, class, and creed in Northern Ireland.* Gower Publishing Company, 1983.

B. L. Nacos. *Terrorism and counterterrorism.* Routledge, 2016.

J. A. Nagl, J. F. Amos, S. Sewall, D. H. Petraeus, et al. *The US Army/Marine Corps Counterinsurgency Field Manual.* University of Chicago Press, 2008.

S. V. Nath. Crime pattern detection using data mining. In *Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on*, pages 41–44. IEEE, 2006.

P. Nesser, A. Stenersen, and E. Oftedal. Jihadi terrorism in europe: The is-effect. *Perspectives on Terrorism*, 10(6), 2016.

J. Neter. *Applied linear regression models*, volume 1. Richard D Irwin, 1996.

P. R. Neumann. The myth of ulsterization in british security policy in northern ireland. *Studies in Conflict and Terrorism*, 26(5):365–377, 2003.

J. L. Noles Jr. Judge enjoins testing of naval surveillance technology (continued). *Nat. Resources & Env't.*, 18:57, 2003.

G. Oatley and B. Ewart. Data mining and crime analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(2):147–153, 2011.

J. O'Bryant and M. Waterhouse. Us forces in iraq. DTIC Document, 2007.

B. H. Office. Counter terrorism statistics. ,https://www.gov.uk/government/collections/counter-terrorism-statistics, 2016. Accessed: 2016-16-11.

A. Ohlheiser. The nsas best defense of prism didn't even last a week. ,http://www.thewire.com/national/2013/06/nsas-only-terroristdefense-prism-didnt-even-last-week/66143/, 2013. Accessed: 2016-11-05.

U. N. O. on Drugs and Crime. *Global study on homicide 2013: trends, contexts, data*. 2013.

R. B. Ohara and D. J. Kotze. Do not log-transform count data. *Methods in Ecology and Evolution*, 1(2):118–122, 2010.

C. Pantazis and S. Pemberton. From the oldto the newsuspect community examining the impacts of recent uk counter-terrorist legislation. *British Journal of Criminology*, 49(5):646–666, 2009.

C. Park and T. Wang. Big data and nsa surveillance–survey of technology and legal issues. In *Multimedia (ISM), 2013 IEEE International Symposium on*, pages 516–517. IEEE, 2013.

S.-H. Park, J.-H. Lee, J.-W. Song, and T.-S. Park. Forecasting change directions for financial time series using hidden markov model. In *International Conference on Rough Sets and Knowledge Technology*, pages 184–191. Springer, 2009.

W. S. Parkin. *Developing theoretical propositions of far-right ideological victimization*. 2012.

M. Peceny and M. Durnan. The farc's best friend: Us antidrug policies and the deepening of colombia's civil war in the 1990s. *Latin American politics and society*, 48(2):95–116, 2006.

E. Perez. U.s. official blames russia for power grid attack in ukraine. http://edition.cnn.com/2016/02/11/politics/ukraine-power-grid-attack-russia-us/, 2016. Accessed: 2016-10-23.

H. M. Peters, M. Schwartz, and L. Kapp. Department of defense contractor and troop levels in iraq and afghanistan: 2007-2016. Technical report, Congressional Research Service Washington United States, 2016.

D. Pratt. Terrorism and religious fundamentalism: Prospects for a predictive paradigm. *Marburg Journal of Religion*, 11(1), 2015.

R. Proser. Israels counterterrorism lessons for europe. `http://www.wsj.com/articles/israels-counterterrorism-lessons-for-europe-1468870438`, 2016. Accessed: 2016-10-27.

J. Pub. Pub 3-07.2. *Joint Tactics, Techniques, and Procedures for Antiterrorism*, 1998.

S. Pylypenko. Cognitive alerting for performance measures.

V. Raghavan and A. G. Tartakovsky. Tracking changes in resilience and level of coordination in terrorist groups. *arXiv preprint arXiv:1604.02051*, 2016.

D. Rand and N. Heras. Iraqs sunni reawakening. *Foreign Affairs*, 16, 2015.

A. Rashid. The fires of faith in central asia. *World Policy Journal*, 18(1):45–55, 2001.

J. A. Ravndal. Right-wing terrorism and violence in western europe: Introducing the rtv dataset. *Perspectives on Terrorism*, 10(3), 2016.

T. E. Ricks. *The gamble: General David Petraeus and the American military adventure in Iraq, 2006-2008*. Penguin, 2009.

A. Roberts. Counter-terrorism, armed force and the laws of war. *Survival*, 44 (1):7–32, 2002.

C. Robertson and T. A. Nelson. Review of software for space-time disease surveillance. *International Journal of Health Geographics*, 9(1):16, 2010.

E. M. Roche and M. J. Blaine. The intelligence gap: What the multinational enterprise can learn from government and military intelligence organizations. *Thunderbird International Business Review*, 57(1):3–13, 2015.

O. A. Romero and J. R. Brockman. *The violence of love*. Plough Publishing House, 1998.

A. Ronchey. Guns and gray matter: terrorism in italy. *Foreign Affairs*, 57(4): 921–940, 1979.

W. Rose, R. Murphy, and M. Abrahms. Does terrorism ever work? the 2004 madrid train bombings. *International Security*, 32(1):185–192, 2007.

A. Rosenfeld. Militant democracy: The legacy of west germanys war on terror in the 1970s. *The European Legacy*, 19(5):568–589, 2014.

J. I. Ross. The rise and fall of québécois separatist terrorism: A qualitative application of factors from two models. *Studies in Conflict & Terrorism*, 18 (4):285–297, 1995.

P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*, volume 589. John wiley & sons, 2005.

D. Royds, S. W. Lewis, and A. M. Taylor. A case study in forensic chemistry: the bali bombings. *Talanta*, 67(2):262–268, 2005.

C. L. Ruby. The definition of terrorism. *Analyses of social issues and public policy*, 2(1):9–14, 2002.

M. Sageman. *Understanding terror networks*. University of Pennsylvania Press, 2004.

C. Ö. ŞAHİN. The new discipline of intelligence world and its critical component: Cybint and humint1.

N. A. Sales. Domesticating programmatic surveillance: Some thoughts on the nsa controversy. *ISJLP*, 10:523, 2014.

J. Sall. Leverage plots for general linear hypotheses. *The American Statistician*, 44(4):308–315, 1990.

M. Salmon, D. Schumacher, and M. Höhle. Monitoring count time series in r: Aberration detection in public health surveillance. 2016.

T. Sandler. The analytical study of terrorism taking stock. *Journal of Peace Research*, page 0022343313491277, 2013.

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

B. Schneier. *Data and Goliath: The hidden battles to collect your data and control your world*. WW Norton & Company, 2015.

J. Sekulow and J. Sekulow. *Rise of ISIS: A threat we can't ignore*. Simon and Schuster, 2015.

I. S. Sheehan. Assessing and comparing data sources for terrorism research. In *Evidence-based counterterrorism policy*, pages 13–40. Springer, 2012.

G. Shmueli, N. R. Patel, and P. C. Bruce. *Data Mining for Business Analytics: Concepts, Techniques, and Applications in XLMiner*. John Wiley & Sons, 2016.

C. Sievert, C. Parmer, T. Hocking, S. Chamberlain, K. Ram, M. Corvellec, and P. Despouy. *plotly: Create Interactive Web Graphics via 'plotly.js'*, 2016. URL `https://CRAN.R-project.org/package=plotly`. R package version 4.5.2.

A. Silke. The devil you know: Continuing problems with research on terrorism. *Terrorism and political violence*, 13(4):1–14, 2001.

S. Simon. The price of the surge-how us strategy is hastening iraq's demise. *Foreign Aff.*, 87:57, 2008.

S. Sloan and S. K. Anderson. *Historical dictionary of terrorism.* Scarecrow Press, 2009.

T. Soklakova, I. Iemelianov, T. B. Amer, and I. Hahanov. Technological culture of big data. In *2016 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET)*, pages 549–552. IEEE, 2016.

E. L. Sonnhammer, S. R. Eddy, E. Birney, A. Bateman, and R. Durbin. Pfam: multiple sequence alignments and hmm-profiles of protein domains. *Nucleic acids research*, 26(1):320–322, 1998.

H. E. SPEAR. Secure flight program airline passenger screening efforts. 2015.

P. Spellucci. Donlp: Do nonlinear programming, 1993. *Obtained via Xnetlib server.*

D. A. Stacey, B. Chrusczc, and D. Calvert. Comparison of aberration detection algorithms for syndromic surveillance. *Advances in Disease Surveillance*, 2: 69, 2007.

start.umd.edu. Surus-time series anomaly detectio. ,`https://github.com/Netflix/Surus/blob/master/resources/R/RAD/man/AnomalyDetection.rpca.Rd`, 2016a. Accessed: 2017-05-02.

start.umd.edu. Start, the study of terrorism and responses to terrorism. ,`http://www.start.umd.edu/`, 2016b. Accessed: 2016-16-11.

C. Sterling. *The terror network: the secret war of international terrorism.* Weidenfeld and Nicolson, 1981.

K. D. Strang and Z. Sun. Analyzing relationships in terrorism big data using hadoop and statistics. *Journal of Computer Information Systems*, 57(1):67–75, 2015.

A. Tanner. Examining the need for a cyber intelligence discipline. *Journal of Homeland and National Security Perspectives*, 1(1):38–48, 2014.

terrorism research. Goals and motivations of terrorists. `http://www.terrorism-research.com/goals/`. Accessed: 2016-10-15.

R. G. K. Thompson. *Defeating communist insurgency: The lessons of Malaya and Vietnam.* Number 10. FA Praeger, 1966.

B. Thuraisingham. Data mining for counter-terrorism. *Data Mining: Next Generation Challenges and Future Directions*, pages 157–183, 2004.

M. Townsley, S. D. Johnson, and J. H. Ratcliffe. Space time dynamics of insurgent activity in iraq. *Security Journal*, 21(3):139–146, 2008.

J. B. Tucker. Strategies for countering terrorism: Lessons from the israeli experience. *Journal of Homeland Security*, 2003.

UCLA. Poisson regression. `http://www.ats.ucla.edu/stat/r/dae/poissonreg.htm`.

O. S. Vallis, J. Hochenbaum, and A. Kejariwal. *AnomalyDetection: Anomaly Detection Using Seasonal Hybrid Extreme Studentized Deviate Test*, 2014a. R package version 1.0.

O. S. Vallis, J. Hochenbaum, and A. Kejariwal. *AnomalyDetection: Anomaly Detection Using Seasonal Hybrid Extreme Studentized Deviate Test*, 2014b. R package version 1.0.

M. van Niekerk and A. Pizam. How do terrorism and tourism coexist in turbulent times?: Introduction to a conflicting relationship. *Terrorism and the Economy: Impacts of the Capital Market and the Global Tourism Industry*, pages 109–125, 2015.

W. N. Venables and B. D. Ripley. Random and mixed effects. In *Modern applied statistics with S*, pages 271–300. Springer, 2002.

J. M. Ver Hoef and P. L. Boveng. Quasi-poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, 88(11):2766–2772, 2007.

I. Visser, M. Speekenbrink, and M. I. Visser. Package depmixs4. 2012.

R. Wang and L. Novik. Why mass surveillance can't, won't, and never has stopped a terrorist. `,http://digg.com/2015/why-mass-surveillance-cant-wont-and-never-has-stopped-a-terrorist`, 2015a. Accessed: 2016-12-11.

R. Wang and L. Novik. A case study: Improve classification of rare events with sas enterprise miner. `,https://support.sas.com/resources/papers/proceedings15/3282-2015.pdf`, 2015b. Accessed: 2016-11-05.

X. Wang, E. Miller, K. Smarick, W. Ribarsky, and R. Chang. Investigative visual analysis of global terrorism. In *Computer Graphics Forum*, volume 27, pages 919–926. Wiley Online Library, 2008.

J. Watling. Ctc sentinel. volume 8, issue 5. Technical report, DTIC Document, 2015.

K. Weinhauer. Terror and democracy in west germany. *German History*, 32(3): 509–511, 2014.

K. Weinhauer, J. Requate, and H.-G. Haupt. *Terrorismus in der Bundesrepublik*. Campus Verlag, 2006.

S. A. Whiting. *Spoiling the Peace?: The Threat of Dissident Republicans to Peace in Northern Ireland*. Oxford University Press, 2015.

H. Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20, 2007. URL `http://www.jstatsoft.org/v21/i12/`.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL `http://ggplot2.org`.

H. Wickham. *purrr: Functional Programming Tools*, 2016a. URL `https://CRAN.R-project.org/package=purrr`. R package version 0.2.2.

H. Wickham. *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*, 2016b. URL `https://CRAN.R-project.org/package=tidyr`. R package version 0.6.0.

H. Wickham. Programming with ggplot2. In *ggplot2*, pages 241–253. Springer, 2016c.

H. Wickham and R. Francois. dplyr: A grammar of data manipulation. *R package version 0.4*, 1:20, 2015.

H. Wickham and R. Francois. *dplyr: A Grammar of Data Manipulation*, 2016. URL `https://CRAN.R-project.org/package=dplyr`. R package version 0.5.0.

H. Wickham and G. Grolemund. R for data science, 2016.

T. P. Wickham-Crowley. Terror and guerrilla warfare in latin america, 1956–1970. *Comparative Studies in Society and History*, 32(02):201–237, 1990.

L. Wilkinson. *The grammar of graphics*. Springer Science & Business Media, 2006.

D. Williams. Hackers declare 'nuclear leak' on israeli twitter account. `http://www.reuters.com/article/us-cybersecurity-israel-idUSKBN0F91G520140704`, 2014. Accessed: 2016-10-31.

P. Wolf. The assassination of ahmad shah massoud. *Pincourt, QC: Center for Research on Globalization*, 2003.

J. K. Young and L. Dugan. Survival of the fittest: Why terrorist groups endure. *Perspectives on Terrorism*, 8(2), 2014.

Z. Zhou, X. Li, J. Wright, E. Candes, and Y. Ma. Stable principal component pursuit. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 1518–1522. IEEE, 2010.

E. Zureik, D. Lyon, and Y. Abu-Laban. *Surveillance and control in Israel/-Palestine: Population, territory and power*. Routledge, 2010.

G. Zwerman, P. Steinhoff, and D. Porta. Disappearing social movements: Clandestinity in the cycle of new left protest in the us, japan, germany, and italy. *Mobilization: An International Quarterly*, 5(1):85–104, 2000.