

An Algorithm for Identification of Terror Events and Hotspots Using  
K-Means and Discriminant Analysis Approach

By  
John Kelvin Ndambuki

A Research Thesis submitted to the School of Computing and Engineering  
Sciences in Partial Fulfilment for the Requirement of the Degree of Master of  
Science in Information Technology at Strathmore University

Strathmore University

December, 2021

### **Declaration**

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

© No part of this thesis may be reproduced without the permission of the author and Strathmore University.

Name: John Kelvin Ndambuki

Signature.....

Date: .....

### **Approval**

The thesis of John Kelvin Ndambuki was reviewed and approved by the following:

Dr. Vincent Omwenga,  
Senior Lecturer, School of Computing and Engineering Sciences,  
Strathmore University

Dr Julius Butime,  
Dean, School of Computing and Engineering Sciences,  
Strathmore University

Dr. Bernard Shibwabo,  
Director of Graduate Studies,  
Strathmore University

## Abstract

This study aimed at developing an algorithm for the identification of terror events and hotspots using K-means clustering and discriminant analysis, with pre-terror recruitment, planning and preparatory activities as the determinant factors. Rules and logic for quantifying the risk state of a pre-terror event based on the values of the constituent determinants of that pre-terror event, for example recruitment as determined by factors such as age of recruits, location and terror organization involved, were developed. K-means clustering was used to come up with two clusters based on the risk value combination of the pre-terror event activities. The two clusters represented the outcome of having a terrorist event happening and a terrorist event not happening, and were duly labelled. Discriminant analysis was used on the now labelled clustered dataset to come up with two identification functions, one for terror event happening and another for a terror event not happening. Unseen possible values for pre-terror event activities were fed into the developed algorithm and an identification of whether a terror event would take place and possibly where was accomplished. The purpose of this identification as per the aim of the study was to offer insights to the organizations dealing with counterterrorism activities. The general public would benefit from this effort once a possible terror attack was prevented before it actually took place. The main research question that guided this study was how accurately a possible terrorist attack incident could be identified before it happened. This study used data from the Global Terrorism Database (GTD) retrieved from the National Consortium for the Study of Terrorism And Responses of Terrorism (START) to come up with geospatial datasets as part of a geodatabase with spatial and temporal information on areas that have been attacked before and the risk values of their constituent pre-terror events.

**Keywords:** *Terrorism, Hotspot, Identification, Spatial-temporal, Mapping*

## Table of Contents

Declaration.....	ii
Abstract.....	iii
List of Figures .....	vii
List of Equations.....	viii
List of Tables .....	ix
Abbreviations and Symbols .....	x
Acknowledgements .....	xi
Dedication .....	xii
Chapter 1: Introduction.....	1
1.1    Background of the study.....	1
1.2    Problem Statement.....	2
1.3    Objectives .....	4
1.3.1    General Objective .....	4
1.3.2    Specific Objectives.....	5
1.4    Research Questions.....	5
1.5    Justification .....	5
1.6    Scope and Limitation.....	6
Chapter 2: Literature Review .....	8
2.1    Introduction .....	8
2.2    Theoretical Framework .....	8
2.3    Spatio-temporal activities of pre-terrorism events and terror hotspots.....	9
2.4    Existing techniques used to identify terrorism events and terror hotspots using a spatio-temporal approach.....	14
2.4.1    Using GIS and Random Forest Method .....	14
2.4.2    Prospective Space – Time Scan Statistics .....	17
2.4.3    Risk Terrain Modelling Framework .....	19
2.4.4    Hawkes processes .....	20
2.4.5    New framework that uses patterns and relations to understand terrorist behaviors .....	21
2.4.6    K-means clustering .....	22
2.4.7    Discriminant Analysis .....	23
2.5    Proposed Algorithm and Conceptual Framework.....	25
Chapter 3: Research Methodology .....	27

3.1	Introduction .....	27
3.2	Agile Software Development Methodology .....	27
3.2.1	Requirements.....	28
3.2.2	Design.....	28
3.2.3	Development.....	28
3.2.4	Testing .....	28
3.2.5	Implementation and Deployment .....	28
3.2.6	Review .....	28
3.3	Research Design .....	29
3.3.1	System Analysis .....	29
3.3.2	System Design .....	29
3.4	Target Population and Sampling/Experimental setup .....	30
3.5	Data Collection .....	30
3.6	Data Analysis.....	31
3.7	Research Quality .....	31
3.8	Ethical Consideration .....	32
Chapter 4: System Analysis, Design and Architecture.....		33
4.1	Introduction .....	33
4.2	System Analysis .....	33
4.2.1	Requirements Gathering.....	33
4.2.2	Functional Requirements .....	34
4.2.3	Non-Functional Requirements .....	35
4.3	System Architecture.....	35
4.4	System Design .....	36
4.4.1	Context Diagram.....	36
4.4.2	Data Flow Diagram (DFD).....	36
4.4.3	Entity Relation Diagram (ERD).....	37
4.4.4	Wireframes of the system .....	39
Chapter 5: System Testing and Implementation .....		40
5.1	Introduction .....	40
5.2	System Implementation .....	40
5.3	System Testing.....	46
5.4	System Validation/Deployment.....	49
Chapter 6: Discussion.....		51
6.1	Introduction .....	51

6.2	Identifying pre-terror activities.....	51
6.3	The existing tools for identification.....	52
6.4	K-means and Discriminant analysis combined algorithm.....	52
6.5	Validating the k-means and Discriminant analysis combined algorithm ....	53
Chapter 7: Conclusion and Recommendation .....		55
7.1	Conclusion .....	55
7.2	Recommendations.....	56
7.3	Future Work.....	56
References .....		57
Appendix A: Originality Report .....		62
Appendix B: Ethical Approval from SU-IERC .....		63
Appendix C: Ethical Approval certificate from NACOSTI .....		64
Appendix D: Code Snippets of the Identification Model.....		65

## List of Figures

Figure 2- 1: Distribution of the terror attacks according to locality on the map in 2018 .....	10
Figure 2- 2 Flow chart of terrorist group activity .....	11
Figure 2- 3 Temporal Patterns of terrorist group activities.....	12
Figure 2- 4 Using Random Forest to simulate a terrorist act. ....	15
Figure 2- 5 Spatial distribution of potential terrorist attack risk.....	16
Figure 2- 6 Cylinders showing clusters of incidents in space and time. ....	18
Figure 2- 7 Illustration of Hawkes Process algorithm visually .....	21
Figure 2- 8: Conceptual Framework.....	26
 Figure 3- 1 : Phases of Agile Methodology.....	 27
 Figure 4- 1 System architecture.....	 35
Figure 4- 2 Context Diagram .....	36
Figure 4- 3 Data Flow Diagram.....	37
Figure 4- 4 ERD Diagram.....	38
Figure 4- 5 System Wireframe.....	39
 Figure 5- 1: ArcGIS Interface showing the pre-terror events in Kenya, Nigeria, Somalia and Iraq as these were the selected areas in the dataset used. ....	 43
Figure 5- 2: Data Input form for user .....	44
Figure 5- 3 A warning message to tell the user not to leave out blank fields.....	47
Figure 5- 4 Results from k-Means clustering algorithm .....	48
Figure 5- 5 Results from Discriminant analysis algorithm.....	49
Figure 5- 6 Input form with values from user.....	50
Figure 5- 7 The input instance of pre-terror values will result to a Y as predicted by the discriminant analysis tool. ....	50
 Figure AA- 1: Originality Report.....	 62
 Figure AB- 1 : Research permit letter from SU-IERC as internal ethical approval.....	 63
 Figure AC- 1 : Ethical approval certificate from NACOSTI.....	 64
 Figure AD- 1 : The tool_exec ( ) function and some of the expected input and output parameters.....	 65
Figure AD- 2: Calculating Recruitment activity risk value logic .....	66
Figure AD- 3 : Calculating Planning activity risk value logic .....	67
Figure AD- 4 : Calculating Preparatory risk value logic .....	68
Figure AD- 5 : Logic for aggregating risk values. ....	68
Figure AD- 6 : K-means Clustering in the algorithm.....	69
Figure AD- 7 : logic for outputting results for KMEANS clustering from R script to ArcGIS .....	69
Figure AD- 8 : Discriminant analysis in the algorithm.....	69
Figure AD- 9 : Writing the results from Discriminant analysis algorithm to ArcGIS as a table.....	70
Figure AD- 10 : Exception handling for negative values of age.....	70

## List of Equations

Equation 2- 1 k-means clustering Equation .....	22
Equation 2- 2 Discriminant Analysis equation .....	24
Equation 5- 1 Application of the K-Means clustering equation to the study variables.....	41
Equation 5- 2 : Discrimination functions from values of the determinant variables.....	42



## **List of Tables**

Table 2- 1: Best specification model for the Risk Terrain Modelling framework .....	20
Table 5- 1 A : List of Different Test cases and Results done on the system. ....	47
Table 5- 1 B : (Cont..)List of Different Test cases and Results done on the system. ....	48

## Abbreviations and Symbols

<b>ANN</b>	-	Artificial Neural Network
<b>CSV</b>	-	Comma Separated Value
<b>DFD</b>	-	Data Flow Diagram
<b>ERD</b>	-	Entity Relation Diagram
<b>ESALLOR</b>	-	Evolutionary Simulating Annealing Lasso Logistic Regression
<b>GIS</b>	-	Geographical Information System
<b>GTD</b>	-	Global Terrorism Database
<b>GTI</b>	-	Global Terrorism Index
<b>IDE</b>	-	Integrated Development Environment
<b>ISIS</b>	-	Islamic State of Iraq and the Levant
<b>K-NN</b>	-	K- Nearest Neighbor
<b>NACOSTI</b>	-	National Commission for Science, Technology & Innovation
<b>NB</b>	-	Naïve Bayes
<b>RDWTI</b>	-	RAND Database of Worldwide Terrorism Incidents
<b>START</b>	-	Study of terrorism and Responses of Terrorism
<b>SSA</b>	-	Structured System Analysis
<b>SSD</b>	-	Structured System Design
<b>SU-IERC</b>	-	Strathmore Institutional Ethics Review Committee
<b>UML</b>	-	Unified Modeling Language

## **Acknowledgements**

I would like to thank my supervisor Dr Vincent Omwenga for his guidance and insights throughout the research period. I also appreciate the rest of the School of Computing and Engineering Sciences (SCES) faculty members, both administrative and teaching staff for their various roles in my research journey. I acknowledge Prof. David Gichoya's role in informing me what the research field was all about. Lastly I appreciate my course mates for their invaluable feedback during our in-class presentation sessions.

## **Dedication**

This work is dedicated to my aunts Lucy Ndambuki and Veronica Nduva who advised me to start this journey. To my friends: Florence, Nahashon, John, Grace, Michael and Esther and the rest of the “sacco” members, be blessed for your support.

## **Chapter 1: Introduction**

### **1.1 Background of the study**

Terrorism is the “premeditated use or threat to use violence by individuals or subnational groups to obtain a political or social objective through the intimidation of a large audience beyond that of the immediate victims”, (Enders & Sandler, 2012, as cited in Sandler, 2014). Terrorism as analyzed by Otiso (2009), has multiple facets to it, the first being that it can be viewed from a legal standpoint as a violation of laws of a nation, from a moral standpoint as the violence being inhumane and morally wrong, and finally from a behavioral approach, in which case terrorism is characterized by its nature in that it is destructive to both property and lives. For instance according to statistics from the Global Terrorism Database (GTD), more than 98,773 terrorist attacks were reported between 2001 and 2016, which resulted in approximately 238,808 deaths. These incidents are spatially aggregated in the Middle East, South Asia, and North Africa, which are considered geopolitically vulnerable regions. There are certain areas, regions, countries or continents that have a high risk of experiencing frequent terror attacks (Braithwaite et al., 2007). These are called terror hotspots. For an area to be considered a terror hotspot, it ought to have either been attacked before, is in close proximity to an area that has been attacked before, or has similar characteristics to an area that has had a terror attack before.

Terror events and hotspot identification refers to the process of trying to locate areas, regions, countries or continents that are more susceptible to attacks by terrorists. Various techniques and tools have been utilized to try and identify these hotspots. Braithwaite et al. (2007) used spatial statistics in their research to identify terrorism hotspots while Nemeth et al. (2014) identified terrorist hotspots using Geographic Information Systems (GIS). Another technique that has been used to identify an area as a hotspot indirectly by predicting future terror events is the Hawkes Process (Tench et al., 2016; Porter & White, 2012).

Hasisi et al. (2020) did their research on spatial clustering and distance decay of lone terrorist vehicular attacks, to check whether there is a correlation between terror hotspots and “hot-routes”, the transport route used by attackers. Machine Learning techniques such as Support Vector Machines (SVM) as used by Mo et al. (2017), K-Nearest Neighbor (K-NN) and Naïve Bayes (NB) (Abou-El-Enien et al., 2015) and Random Forest (RF) (Saha et al., 2017), have been used in research related to terrorist attacks in a hotspot area. Discriminant analysis is a statistical technique for grouping observations into distinct groups based on the combination of select individual characteristics of the observation variables (Dikko & Osi, 2014). Research done by Wang et al. (2013) utilized Discriminant Analysis by using geospatial discriminative patterns to understand spatial distribution of crime. K-means algorithm is an unsupervised machine learning technique that is used to optimize outcome through finding  $K$  groups in the data (Li & Wu, 2012). This research uses the discriminant analysis and k-means tools as part of a novel algorithm to analyze pre-terrorism events: recruitment, planning and preparation activities and their constituent sub-activities: radicalization, terrorist cells creation, communication both online and offline, funds request and transfer, training, espionage, assembling weapons, travelling to target locations, surveying the target locations and any other form of ancillary crime such as killing informants and accomplices. The pre-terrorism activities and sub-activities can be presented spatially and temporally as soon as they are detected and reported to visually show linear trends and patterns leading to an actual terrorist event.

## **1.2 Problem Statement**

The identification of possible terrorist activities and hotspot areas continues to be a challenging task for researchers and government agencies concerned. This is due to the fact that parameters involved in the identification are not always well defined with a direct correlation to an area being a possible terror hotspot. Examples of these parameters include: age and gender demographics of an area, economic status, religion, marital status, population density of an area,

proximity of an area to a capital city, political and leadership style and the type of terrain of that area. These parameters involved in the identification are usually based on a specific area and region of a country and even continent and are very biased and stereotypical to a large extent, for example profiling areas where certain religions and ideologies are practiced or assuming that areas within and in the outskirts of capital cities are the likely hotspot areas. Different motives from different terrorist groups also introduce a lot of ambiguity in terms of which factors, whether economic, political or social and religious, are the real triggers of the attacks. The area that will be attacked at times does not necessarily have a direct correlation with the motive. Therefore there is a lot of variability in the determination of a common end goal, which is a possible terrorist attack in a certain location.

Terrorism has a complexity dimension to it comprising of a myriad of known and unknown factors. This makes it difficult to predict, especially at an early stage (Gao et al., 2013). According to Afsar et al. (2014), a lot of research has been done explaining root causes of terrorism, but there is a gap when it comes to identifying patterns of terrorist attacks, including the hotspots over space and time. Some factors such as demographics and economic factors can add some perspective to the present and past states of terrorist's activities in an area, but they cannot be used to identify future terrorist activity related events accurately without some bias.

Research on application of technology to counter terrorism generally has been conducted. In predicting terrorism related variables like the group that was involved, weapons used and casualties, Machine learning models have been applied widely. These include SVM, as used by Mo et al. (2017) in their research on prediction of terrorist events based on revealing data. Artificial Neural Networks (ANN) are also being used in research (Samantha & Elliot, 2011; Uddin et al., 2020), K-NN and NB as applied by Abou-El-Enien et al. (2015) in their research and RF as used by Saha et al. (2017) to predict weapon types used in an attack, the actual attack type and the types of organizations as

targets. Hawkes process is used to predict future terror events as shown in the research done by Tench et al. (2016) and also in another research by Porter and White (2012).

This study proposes the development of an algorithm that considers a neutral and unbiased dataset of an area that depicts terrorism pre-attack activities as they take place over space and time. Pre-terrorism event activities: recruitment, planning and preparatory, and their sub-activities as determinants such as radicalization, terrorist cells creation, communication both online and offline, funds request and transfer, training, espionage, assembling weapons, travelling to target locations, surveying the target locations and any other form of ancillary crime such as killing informants and accomplices are considered because they are objective and unbiased. Through the unsupervised machine learning technique using k-means, two clusters are achieved of pre-terror events that lead to a terror event and pre-terror events that do not lead to a terror event. Discriminant analysis tool is used on the dataset to create two identification functions: one for terror event happening and another for a terror event not happening based on the two clusters from the k-means algorithm, by statistically analyzing values that characterize a cluster. When these discriminating functions are given values of the pre-terrorism events activities and sub-activities as they are detected, an identification of whether an actual terrorism event will result or not and possibly where based on the entries so far is accomplished. The algorithm is used to develop a terrorist attack event and hotspot identification system.

### **1.3 Objectives**

#### **1.3.1 General Objective**

The aim of this research is to develop an algorithm that can be used for the identification of terrorism events and hotspots, to provide awareness and aid counterterrorism, using k-means and discriminant analysis approach on pre-terrorism activities and sub-activities.



### **1.3.2 Specific Objectives**

- i. To analyze the spatio-temporal activities and sub-activities of pre-terrorism events and terror hotspots.
- ii. To examine existing techniques used to identify terrorism events and hotspots using a spatio-temporal approach.
- iii. To design and develop an algorithm for the identification of terrorism activities and terror hotspots using k-means and discriminant analysis approach.
- iv. To validate the reliability of the proposed algorithm in identifying terrorism activities and terror hotspots.

### **1.4 Research Questions**

- i. What are the spatio-temporal activities and sub-activities of pre-terrorism events and terror hotspots?
- ii. How are existing techniques used to identify terrorism events and hotspots using a spatio-temporal approach?
- iii. How can an algorithm for identifying terrorism activities and terror hotspots using k-means and discriminant analysis approach be designed and developed?
- iv. How can the reliability of the proposed algorithm to identify terrorism events and hotspots be validated?

### **1.5 Justification**

Terrorism is a vice that usually brings a lot of suffering: physically, mentally and psychologically for those affected. Innocent victims end up losing their lives. When this happens, the effects are felt by the entire community, especially where those victims are the breadwinners for their families or skilled workforce such as doctors and teachers. Those who survive the attacks end up having disabilities caused by the violence involved in the attack, either losing limbs to explosives or bullets lodged in their bodies that cannot be removed.

This study dealt with how to use past patterns and trends from the pre-terrorism activities conducted by terrorists as they prepared for the actual attack in the cases where the attack occurred, to identify a possible attack before it happens based on current pre-terrorism activities as they happen. An algorithm developed using k-means clustering technique and discriminant analysis was used for this trend and pattern analysis and identification. This boosts counterterrorism efforts through identifying possible terror events accurately. In the field of academics, the algorithm developed adds new knowledge to the domain of classification and identification algorithms.

## **1.6 Scope and Limitation**

This study focused on identifying the occurrence of a possible terrorist attack and the likely location it could occur. This is based on three general pre-terror events: recruitment, planning and preparatory activities, and their determinant characteristics such as where they take place, duration of time they take place, age demographic of those involved, the group name of the terror organization, any legal and illegal ancillary activities such as raising and transfer of funds, training and weapon movement, travelling to target locations and surveillance activities.

The study worked with the incidents that were present in the developed geodatabase from terrorist incidents extracted from the GTD database for four purposive sampled countries: Kenya, Nigeria, Somalia and Iraq. Despite this sample of countries considered, the tool is not specific to any region of the world, it could be applied globally on any pre-terror activities to determine if they can lead to a terror event. The tool would be used on reported or detected pre-terror activities as done by any person from the entire population in an area and not just a particular group of interest.

The approach used to create an association of the occurrence of a terror event or lack of from the independent pre-terrorism activities and sub-activities

considered the risk quantification of that activity to the overall determination of the terrorism event in a combined state. This means that the pre-terror sub-activities' characteristics: where they take place, duration of time they take place, age demographic of those involved, the group name of the terror organization, any legal and illegal ancillary activities such as raising and transfer of funds, training and weapon movement, travelling to target locations and surveillance activities' risk values, would determine their specific associated pre-terrorism activities': recruitment, planning and preparatory risk values when combined. This would in turn determine the overall risk value of the final terror event happening or not when also combined for the three pre-terrorism activities. The pre-terror activities and sub-activities were identified but they did not have a consistent data type and value that could be used to model the risk associated with them when it came to overall determining the possibility of a terror event. The research involved an improvisation technique of imputing a range of values to represent risk factors of High, Medium and Low, based on a scale of the range of values from 1.0 – 10.0. Values between 1.0 – 4.0 represented Low risk, 4.1 – 5.9 represented Medium risk and 6.0 – 10.0 represented High risk. The researcher worked with the column on Country, Town, Terrorist Organization, Latitude and Longitude data from the GTD dataset, the other columns on a custom eventID, Recruitment, Planning and Preparatory risk value columns and Terror\_Event were created based on the risk value imputation discussed earlier. A geospatial dataset was created from this dataset.

## **Chapter 2: Literature Review**

### **2.1 Introduction**

This chapter provides a review of relevant literature related to the study with the aim of analyzing the spatio-temporal activities and sub-activities of pre-terrorism events and terror hotspots and examining the existing techniques being used to identify possible terrorism events and hotspots. The gaps in the existing literature are pointed out and the chapter looks at how a specific research gap can be addressed by the proposed algorithm.

### **2.2 Theoretical Framework**

The Crime Pattern theory explains why certain areas are hotspots for crime. It states that crimes do not occur randomly in space, time or society. According to Brantingham and Brantingham (1999, as cited in Wortley & Mazerolle, 2008), there are eight rules of crime pattern theory that can be analyzed to conclude on the general formation of hotspots.

One of the rules is that as individuals move through a series of activities they make decisions. When activities are repeated frequently, the decision process becomes regularized. This regularization creates an abstract guiding template. For decisions to commit a crime this is called a crime template.

Second, most people do not function as individuals, but have a network of family, friends and acquaintances. These linkages have varying attributes and influence the decisions of the others in the network.

Third, when individuals are making their decisions independently, individual decision making processes and crime templates can be treated in a summative fashion, that is, average or typical patterns can be determined by combining the patterns of individuals. The fourth rule says that individuals or network of individuals commit crimes when there is a triggering event and a process by which an individual can locate a target or a victim that fits within a crime template. Criminal actions change the bank of accumulated experience and future actions.

The fifth rule states that individuals have a range of routine daily activities. Usually these occur in different nodes of activity such as home, work, school, shopping, and entertainment or time with friends, and along the normal pathways between these nodes.

On the sixth rule, people who commit crimes have normal spatio-temporal movement patterns like everyone else. The likely location for a crime is near this normal activity and awareness space.

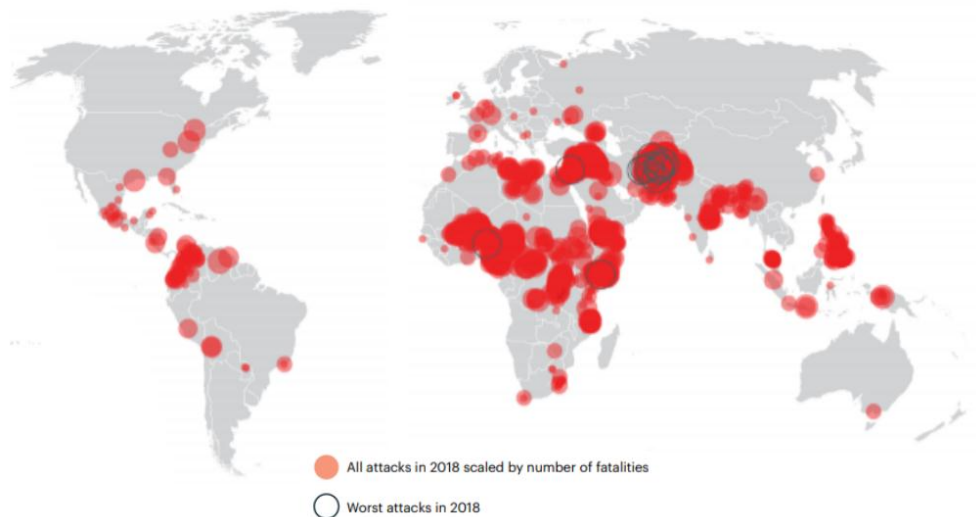
The seventh one states that potential targets and victims have passive or active locations or activity spaces that intersect the activity spaces of potential offenders. The potential targets and victims become actual targets or victims when the offenders willingness to commit a crime has been triggered and when the potential target or victim fits the offender's crime template.

Lastly, the prior rules operate within the built urban form. Crime generators are created by high flows of people through and to nodal activity points. Crime attractors are created when targets are located at nodal activity points of individuals who have a greater willingness to commit crimes.

### **2.3 Spatio-temporal activities of pre-terrorism events and terror hotspots**

Terror hotspots are locations that are more susceptible to terror attacks. Research has been done to understand what factors, characteristics or activities can make an area to be a potential terror hotspot, and in what magnitude and correlation level. According to Smith et al. (2006), terrorists operate within the constraints and boundaries of both space and time.

The Global Terrorism Index (2019) shows the distribution of the terror attacks according to locality on the map in 2018 as illustrated in the Figure 2-1 below:



*Figure 2- 1: Distribution of the terror attacks according to locality on the map in 2018*

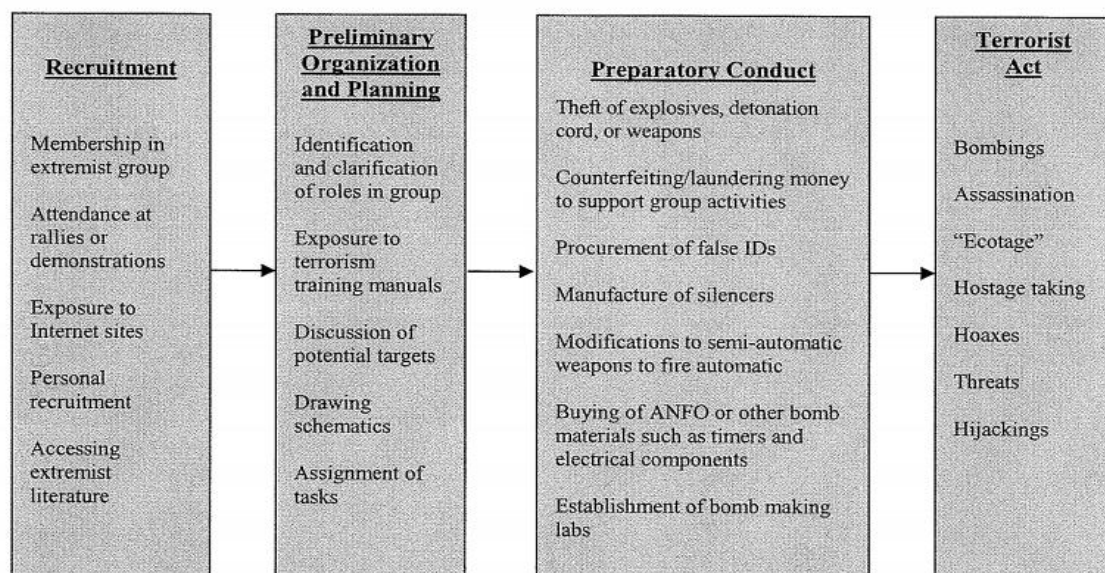
*Source: (GTI, 2019)*

According to the statistics from the GTD, more than 98,773 terrorist attacks were reported between 2001 and 2016, which resulted in approximately 238,808 deaths. These incidents are spatially aggregated in the Middle East, South Asia, and North Africa, which are considered geopolitically vulnerable regions (Hao et al., 2019).

Guo et al. (2016) in their research on Visual analysis methods of terrorism events, talked of how uncertainty in spatio-temporal distribution of events made it difficult to forecast a single event. However with a wide range of events data leading up to a terror activity, spatio-temporal patterns and trends could be extracted. The argument herein was that inasmuch as terrorism events happened “suddenly”, there was usually a relationship between internal attributes item such as event time, event location, the manipulating organization and target type. From the report, the researchers identified the terror group Islamic State of Iraq and the Levant (ISIS) to be operating mainly in Iraq and Syria, with the main targets being civilians, police and military. The attacks were most frequent in June to October in a given year.

Smith et al. (2006) published a report on the identification of Behavioral, Geographic and Temporal patterns of terrorism preparatory conduct, as pre-incident indicators of terrorism incidents in the United States from 1980 to 2002. A comparison was made between traditional criminality and terrorism activities, where it was noted that the former could happen abruptly without prior excessive planning as long as the offender's willingness to commit a crime had been triggered when the potential target or victim fitted the offender's crime template. For terrorism events, they were usually preceded by a lot of planning activities, some within legal lines while others were illegal like ancillary and preparatory crimes. The report sought to find if there was a substantial open source spatio-temporal data on terrorist activities before a terrorism event and whether a pattern or trend could be identified from it, so that an intervention could be made prior to a terrorist attack. Smith et al. (2006) also identified 4 major categories of pre-terrorism event activities – Recruitment, Preliminary Organization and Planning, Preparatory conduct and then the actual Terrorist act.

The Figure 2-2 illustrates these pre-terrorism events activities and their constituent sub-activities:



*Figure 2- 2 Flow chart of terrorist group activity*

*Source: (Smith et al., 2006)*



Still in the report by Smith et al. (2016), it's noted that terrorist incidents were committed slightly over 3 months from the time the planning activities and consequently preparatory activities took place. A large percentage of the planning and preparatory activities took place closer to the target location, however their relationship was not linear to avoid easy detection. On average for international terrorists, more than 50% of terrorist attacks took place within a 0-45km radius from the planning and preparation locations.

The Figure 2-3 below shows the pre-terrorism event activities with respect to time:

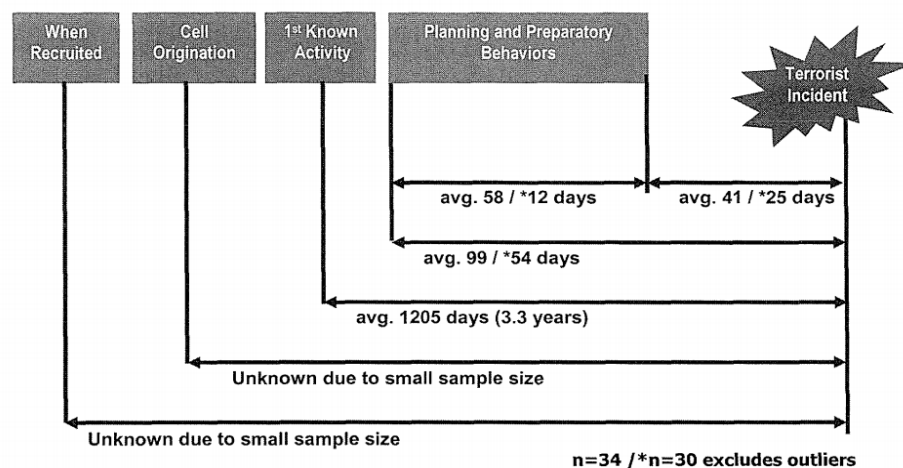


Figure 2- 3 Temporal Patterns of terrorist group activities

Source: (Smith et al., 2006)

O'Neil (2007) conducted research on pre-terrorism events, and discussed their characteristics—some being legal, while most were illegal and criminal, right before a terrorist event, and proceeded to discuss why they were on the rise. These activities were committed because there were needs and these included: funding, security, operatives/support/preparatory, propaganda and means and/or appearance. These activities, especially the ones that were of a criminal nature took place geographically away from the terrorist target location. These activities included: fraud schemes, petty crime, identity theft and immigration crimes, counterfeiting money and goods, import/export violations narcotics trade and illegal weapons procurement, incitement, training, bribery and



phone scams and cell phone activity. O'Neil argued that most pre-cursor activities related to raising funds for a terror cause, were conducted in wealthy western nations. The choice between a legal or a criminal pre-cursor activity for future terrorist event depended on the risk level of detection.

Khalsa (2005) presented a terrorism forecasting methodology using indicators that encompassed terrorists's intentions, capability and target vulnerability. Indicators according to McDevitt (2002, as cited in Khalsa, 2005, How to Forecast Terrorism ) were defined as " those [collectable] things that would have to happen and those that would likely happen as [a] scenario unfolded". These included terrorist travel, weapons movement, terrorist training, target surveillance, and tests for security.

Smith et al. (2013) and Piazza (2014) identified characteristics of communities where terrorism perpetrators lived and also looked at pre-incident activities prior to an attack. The researchers checked whether attackers lived close to where they conducted pre-incident activities prior to a terrorist event. The purpose of the research was to identify characteristics of communities where terror planning events occurred – the researchers used census data of a specific residential location where terrorists and their sympathizers lived. The characteristics included the socioeconomic status, type of politics and leadership style in that country where the residential location above is in i.e democracy or dictatorship, whether there had been any form of military intervention, the laws set by the government with respect to human rights, freedom to speak out and air grievances without fear of oppression, and finally how the country treated religious and ethnic minority. These characteristics on their own could not be used to predict future terrorist events as they would stereotype some places and people. However they can be used to provide insight and magnitude to the pre-terror activities like recruitment, planning and preparatory activities. For example, a certain age group and people facing poor economic situations such as high unemployment can be vulnerable to

recruitment, the ethnic diversity of a place can enable terrorists to camouflage well without detection.

In a study done by Polo (2020) on how Terrorism spread due to emulation, it was theorized that groups would emulate the terrorist choice of others if they happened to share a similar political grievances and spatial networks. The researchers tested that theory on a new and original group-level data set of ethnic and ethno religious terrorism collected between 1970 and 2009 using geospatial analysis and spatial econometric models-with the conclusion that indeed emulation did influence a terrorist's tactic choice. This tactic choice involved their preparatory behavior/characteristics and activities and eventual the actual terror act.

#### **2.4 Existing techniques used to identify terrorism events and terror hotspots using a spatio-temporal approach**

The various degrees of some of the pre-terror determinants above could be analyzed in combination through statistical models such as Multivariate Analysis of Variance (MANOVA), Logistic regression, Factor Analysis, Multi-Dimensional Scaling, Discriminant function analysis, Log-Linear Analysis, Cluster Analysis and Structural Equation analysis to explain crime occurrences in an area (Dikko and Osi, 2014). The following section looks at various existing techniques in literature:

##### **2.4.1 Using GIS and Random Forest Method**

Hao et al. (2019) developed a geospatial statistic tool based on Kernel Density to analyze the spatio-temporal evolution of terrorist attacks on the Indochina peninsula .Using Random Forest algorithm, 15 variables were used in prediction of potential risk of terrorist attacks in the Indochina Peninsula. The researchers followed the following steps:

- 1: Extracted the terrorist attacks on the Indochina Peninsula from the GTD and using the ArcGIS software to spread terrorist attacks on the map;

- 2: The “Kernel Density” function in ArcGIS and OriginLab were used to analyze the evolution of terrorist attacks from a time and space perspective;
- 3: Preparation of spatial geographic data and corresponding raster data of the terrorist attack;
- 4: Construction of the RF algorithm to predict terrorist attacks at the spatial scale on the Indochina Peninsula

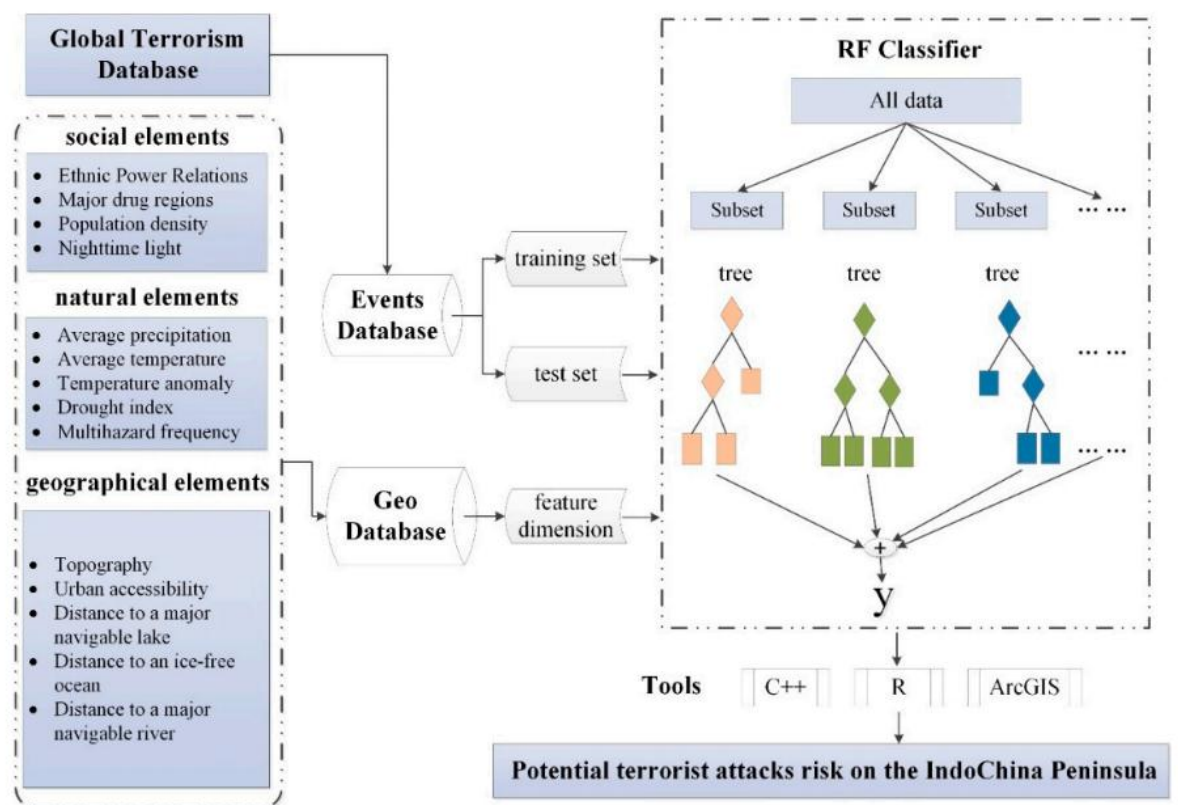
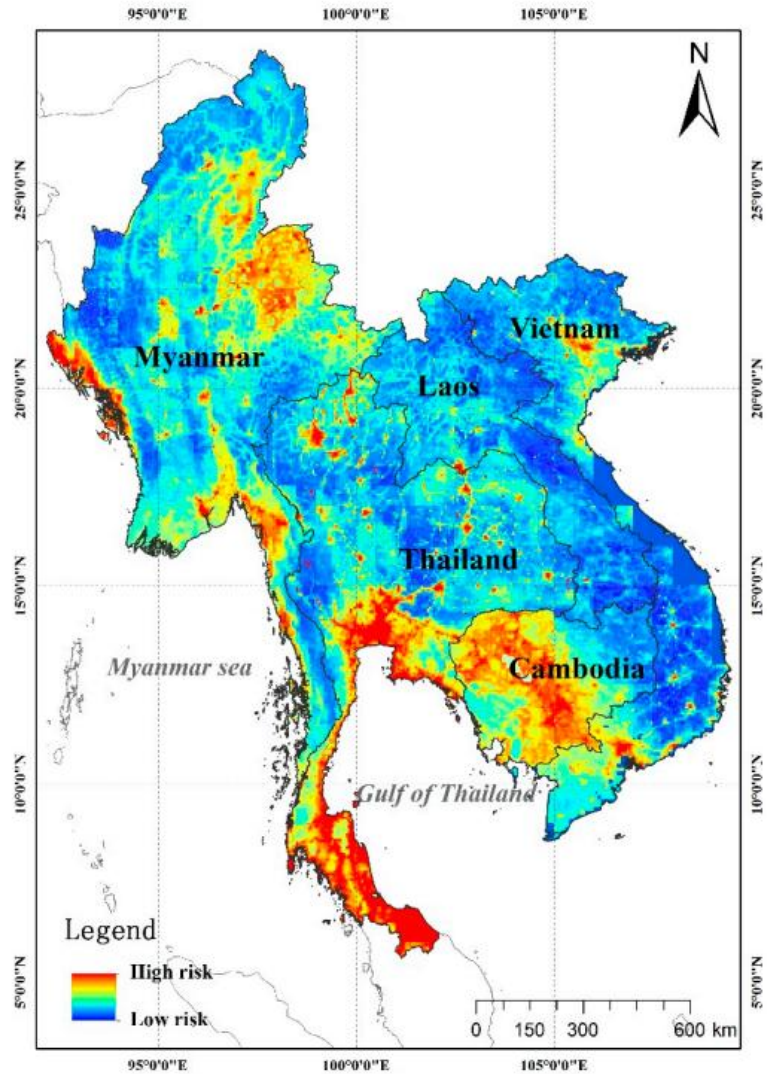


Figure 2- 4 Using Random Forest to simulate a terrorist act.

Source (Hao et al., 2019)

The figure 2-4 above shows how the RF algorithm was used to simulate a terrorist attack, whereby multiple element types were introduced into a RF classifier that was used to predict potential terrorist threats.



*Figure 2- 5 Spatial distribution of potential terrorist attack risk.*

Source (Hao et al., 2019)

The figure 2-5 above shows hotspots and coldspots when it comes to potential terrorism events happening in a certain region. From the research findings, some high risk hotspot areas were Southern Thailand, Bangkok and its surrounding cities, Middle Cambodia, southern parts of Myanmar.

The research does not consider temporal aspects of a terrorist event. The aspect of estimating when an attack will occur is very vital in the accuracy of the prediction model, which will properly guide counterterrorism.

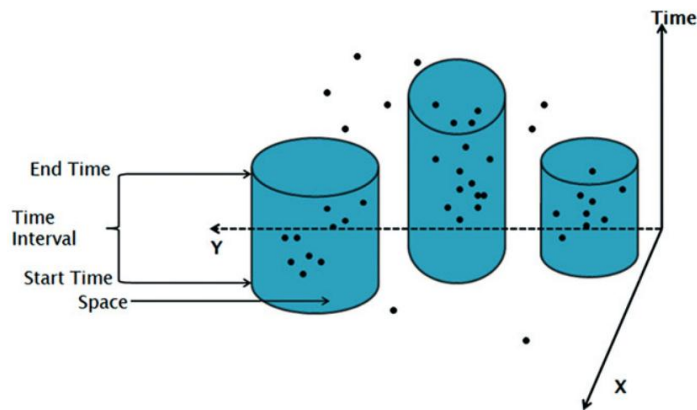
#### 2.4.2 Prospective Space – Time Scan Statistics

Gao et al. (2013) worked on an approach to model terrorism activity for detection at an early stage that involved monitoring and detecting space-time clusters of terrorist incidents using prospective space-time scan statistics. These clusters provided indicators of potential outbreaks of terrorist incidents.

The researchers used the GTD dataset for documented terrorist incidents from 1998 to 2004, which included a total of 7,231 incidents. These incidents had attributes such as date of terrorist attack, name of attack place, type of weapons used, target type, number of casualties and terrorist organization name. Gao et al. (2013) considered only the date of terrorist attack and name of the attacked place – to obtain the number of incidents per each location. The researchers proceeded to create a geocoded dataset, using attributes such as Country name, province name, city name and specifically provided longitude and latitude values.

The technique used involved doing a Space-time scan statistics on the geocoded dataset, using software tool for spatial, temporal and space-time scan statistics called *SaTScan*. A Space-time scan statistics uses 3D cylinders to represent two spatial dimensions and one temporal dimension. The base of a cylinder represents a 2-D circular window in geographic space, and the height of the cylinder represents a time range.

The Figure 2-6 below shows incidents of terror related activities over space and time, with the base representing space and height representing time.



*Figure 2- 6 Cylinders showing clusters of incidents in space and time.*

*Source : (Gao et al., 2013)*

These researchers demonstrated that clusters detected at an earlier time could predict clusters detected at a later time, hence could detect terrorism outbreaks at an early stage after analysis. The researches stated that what differentiates their work from existing terrorist event forecasting tool is that it has a low data requirement and secondly, it considered both spatial and temporal information.

This research by Gao et al. (2013) provides a way to forecast an incoming terror event in a select area, and the likely time period, based on past terror events. This early detection of terrorism outbreak can help in counterterrorism measures through allocating a good percentage of resources to the select region.

This tool however has its weaknesses. The regions forecasted are too general, most likely due to lack of refined data on locations like street addresses and even specific targeted structures like hospitals, schools, military base etc. The time forecasted in months is also general to offer an exact counter date and time. The research uses a final outcome, in this case a terror event in space and time to predict a future terror event in space and time, without considering the indicators of these terror events which can make the tool more refined and offer multiple points of counterterrorism intervention. Another weakness is the assumption that terrorists will always follow their usual criminal template during an attack.

### **2.4.3 Risk Terrain Modelling Framework**

Onat (2019) did research whose aim was to identify the correlates of terrorism in space. The physical structures in areas where terrorist events took place were examined and patterns identified to check if there was a correlation between the terrain and the terror incidence. The research focused on terrorist events committed from 2008 to 2012 in Istanbul, Turkey. The researcher argued that there was a terrorism exposure risk in urban centers infrastructure where people lived, worked and recreated, based on their social routine. The technique of Risk Terrain Modelling sought to identify places that were ideal for terrorism to occur, given the existing environmental contexts.

Using Risk Terrain Modelling framework on data on police incidents and infrastructure (for example buildings like schools, hospitals), (Onat, 2019) used GIS analysis techniques together with an event count model. Once data was aggregated and combined for different buildings, it led to the identification of higher risk infrastructure in relation to terrorist attacks.

The researcher obtained the data from the Counter Terrorism Department at the main headquarters of the Turkish National Police, cleaned up and geocoded it. The dependent variable in this case, terrorism was identified, and the independent variables, the predictor, Infrastructure was also identified. It was concluded that regardless of terrorists' intent, the significantly associated infrastructure increased the risk in the surrounding areas where these features were located.

The Table 2-1 below illustrates the best model specification for the Risk Terrain Modelling Framework.



<b>Risk Factor</b>	<b>Operationalization</b>	<b>Spatial Influence (meters)</b>	<b>Coefficient</b>	<b>Relative Risk Value</b>
Bakery	Proximity	440	1.2931	3.6441
Religious Facility	Proximity	440	1.0127	2.7530
Bar/Club	Density	660	0.9171	2.5021
Grocery Store	Density	660	0.8725	2.3930
Franchise	Proximity	660	0.7864	2.1956
Office Block	Proximity	660	0.5741	1.7755
NGO	Proximity	440	0.5735	1.7744
Eatery	Density	220	0.5245	1.6896

*Table 2- 1: Best specification model for the Risk Terrain Modelling framework*

*Source: (Onat, 2019)*

This research goes a level deeper in analyzing the spatial environments where terror events take place, to prove that depending on the different types of buildings and the activities that take place inside, a risk magnitude can be identified.

#### **2.4.4 Hawkes processes**

Hawkes process is a stochastic process in which past events have the potential of temporarily raising the probability of future event, on the assumption that the excitation is “ positive, additive over past events and exponentially decaying with time” (Hawkes, 1971; Liniger, 2009, as cited in Mei & Eisner, 2017). In another study by Laub et al. (2015), it was also pointed out that one of the characteristic of the Hawkes processes was that they self-excited, to mean that given a certain period of time, each arrival of an event increased the rate of future arrivals.



The Figure 2-7 below illustrates the Hawkes Process visually, showing events as they triggered by previous events as they happen.

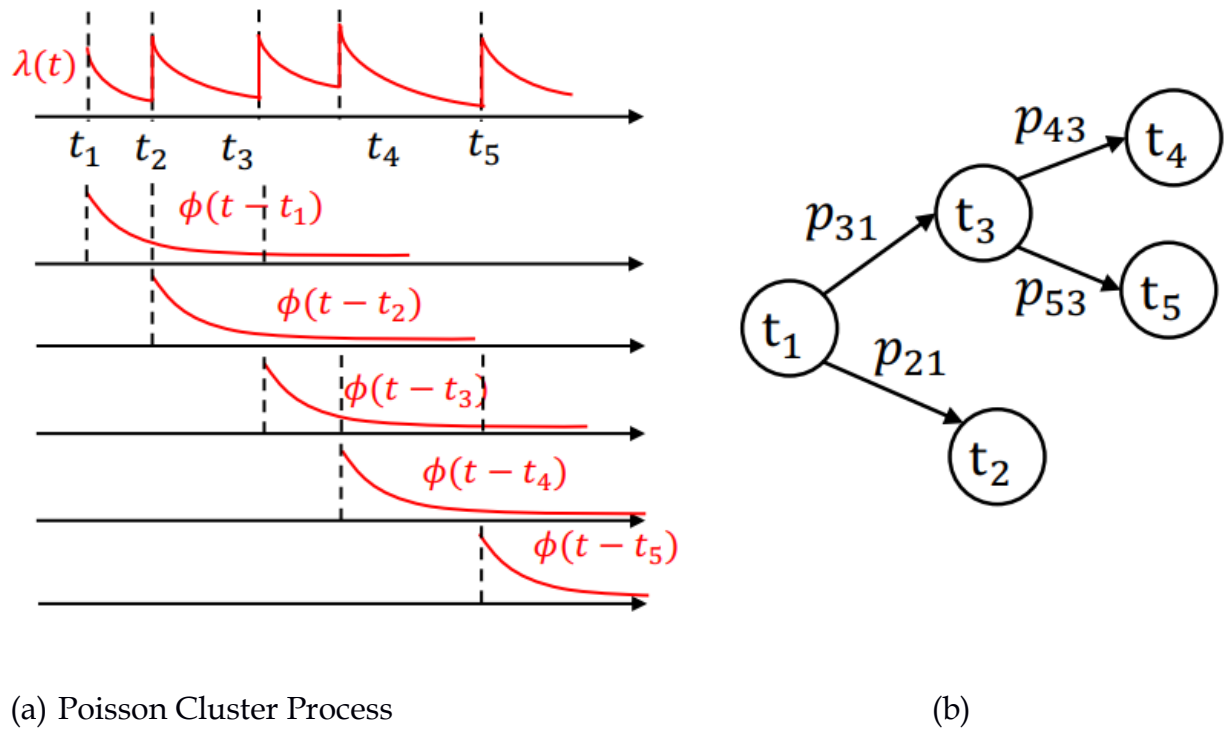


Figure 2- 7 Illustration of Hawkes Process algorithm visually

Source: (Zhang et al., 2018)

Hawkes process was used to predict future terror events, in the research done by Tench et al. (2016) and also another by Porter and White (2012). Other areas of application include, analysis and prediction of earth quakes and also used in financial analysis and stock market prediction.

#### 2.4.5 New framework that uses patterns and relations to understand terrorist behaviors

Research by Tutun et al. (2017) aimed at coming up with a new framework to analyze and understand terrorist activity patterns, specifically suicide attacks, and be able predict future events, leading to prevention of a potential terror attack. Among the features of the tool were a model to select features for similarity function called an Evolutionary Simulating Annealing Lasso Logistic

Regression (ESALLOR) model, a new weighted heterogeneous similarity function to estimate the relationships among attacks and finally a graph-based out-break detection to define hazardous places for the outbreak of violence.

Data was obtained from GTD and RAND Database of Worldwide Terrorism Incidents (RDWTI) .The researchers reported that the experimental results from the framework found patterns in terrorism data with up to more than 90% accuracy.

This research focused a lot on preventing suicide attacks, and this limits its full usefulness in detecting terrorist activities in general, where suicide as a tactic has been ruled out.

#### 2.4.6 K-means clustering

K-means algorithm is an unsupervised machine learning technique that is used to optimize outcome through find  $K$  groups in the data (Li & Wu, 2012). It is mostly applied in data mining and pattern recognition.

The k-means clustering equation is illustrated in Equation 2-1 shown below:

$$j = \sum_{j=1}^k \sum_{i=1}^n ||x_i^{(j)} - c_j ||^2$$

*Equation 2- 1 k-means clustering Equation*

Where,  $j$  is the objective function,

$\sum_{j=1}^k$  Represents the number of clusters,

$\sum_{i=1}^n$  Represents the number of cases,

$||x_i^{(j)} - c_j ||^2$  represents the distance function, with  $x_i^{(j)}$  being case

$i$  and  $c_j$  the centroid for cluster  $j$

The following algorithm describes how k-means works, as analyzed by (Gill, 2013).

**Input:** Set of  $N$  items to cluster

$$D = \{d1, d2, d3, d4...dN\}$$

Select a number of clusters desired,  $k$  where  $k \leq N$

$$K = \{k1, k2, k3, k4...kN\}$$

$$C = \{c1, c2, c3, c4...cN\}$$

Where  $k$  is a subset of  $D$  as a temporary cluster and  $C$  is the centroid of those clusters.

And therefore  $k1 = \{d1\}$ ,  $k2 = \{d2\}$ ,  $k3 = \{d3\}$ ,  $k4 = \{d4\}...kn\{dN\}$ ,  
 $c1 = d1$ ,  $c2 = d2$ ,  $c3 = d3$ ,  $c4 = d4...cN=dN$

**Output:**  $K = \{k1, k2, k3, k4...kN\}$

$$C = \{c1, c2, c3, c4...cN\}$$

Here  $K$  is set of subset of  $D$  as final cluster and  $C$  is set of centroids of these cluster.

K-means algorithm has a computational complexity of  $O(TKn)$ , where  $n$  is the number of input patterns,  $K$  is the desired number of clusters, and  $T$  is the number of iterations needed to complete the clustering process (Pakhira, 2014).

#### 2.4.7 Discriminant Analysis

Discriminant Analysis is a statistical technique for grouping observations into distinct groupings based on the combination of select individual characteristics of the observation variables (Dikko & Osi, 2014). The type of data for the independent variables, i.e the individual characteristics is continuous in nature while that of the final group categories or the dependent variables is discrete. The main idea behind Discriminant Analysis is looking what combination of variables can be used to identify group membership. A linear combination of the discriminating variables is formed and once a variable set which provides

very high accuracy in discriminating group or category membership, it can be used in classification of new cases where the new values of the variables are available, but the final grouping or categorization is unknown (Yekinni & Ayogu, 2015).

The following equation 2-2 illustrates discriminant analysis:-

$$Z_{jk} = a + W_1X_{1k} + W_2X_{2k} \dots + W_nX_{nk}$$

Equation 2- 2 Discriminant Analysis equation

Source (Ramayah et al., 2010)

Where,  $Z_{jk}$  is the discriminant Z score of discriminant function j for object k

$a$  is the intercept,

$W_1$  is the Discriminant coefficient for the independent variable i.

$X_j$  Is the independent variable i for object k.

The following algorithm describes how discriminant analysis tool works:

1. Compute the global mean (M) using all the observations from all groups
2. Compute Mean,  $m$  for the observations in a distinct group.eg  $m_1, m_2 \dots m_N$  for N groups
3. Compute covariance matrix  $c$  for each of the groups ,  $c_1, c_2 \dots c_N$ , for N groups
4. Compute within class scatter matrix C
5. Create discriminant functions  $f_1, f_2 \dots f_N$  for N groups.

The Linear Discriminant analysis algorithm has a computational complexity time of  $O(mnt + t^3)$ , where m is the number of samples, n is the number of features and  $t = \min(m, n)$  (Cai et al., 2008).

## 2.5 Proposed Algorithm and Conceptual Framework

This research proposed a novel algorithm to identify a terrorist attack event and an estimate hotspot location where this event was likely to occur. The algorithm combined the unsupervised learning method of k-means clustering and discriminant analysis.

Steps of the proposed algorithm:

- i. Using an existing dataset that has pre-terror events, and final outcome from past terrorist events that happened and those that did not, omit the outcome value column and only retain the pre-terror events data.
- ii. Apply k-means clustering on the data, with a value of  $k$  as 2, to obtain two clusters.
- iii. Once this data has been clustered, check the cluster value the pre-terror activities are placed in.
- iv. Compare this cluster value and check if they represent items that have the same terror event output value from the column that was omitted in step i.
- v. Count the number of values correctly clustered versus the one not correctly clustered to get the level of accuracy of the k-means algorithm in this clustering.
- vi. If the data from step i did not have an outcome value column to begin with, now that the clusters have been identified this data can be labelled so as to proceed and use it in a discriminant analysis tool.
- vii. Use Discriminant analysis to train a percentage of this dataset, include the outcome variable, as it is. Test the model using the remaining dataset. Confirm model accuracy vs from the one retrieved through k-means.
- viii. Test the model using a new single user input on pre-terror events and check the category it was classified in by the discriminant analysis model. This category is the final outcome of whether a terror event is likely or not.

The art of combining the two tools above in the proposed algorithm meant that the computational complexity of the general k-means algorithm and the computational complexity of the discriminant analysis was aggregated.

The Figure 2-8 below illustrates the conceptual framework of the system.

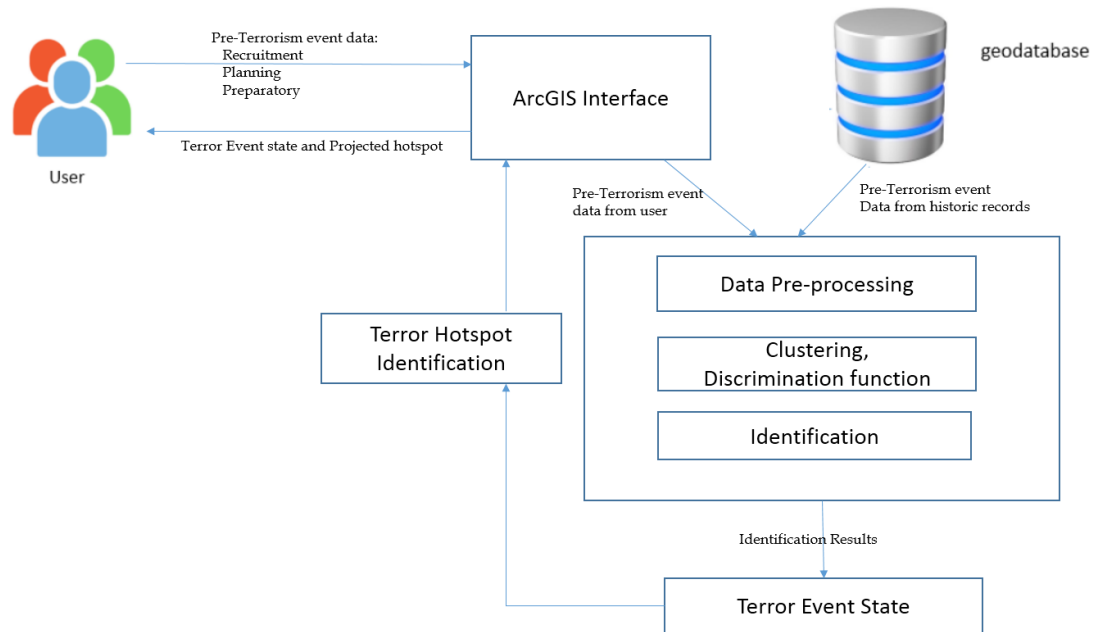


Figure 2- 8: *Conceptual Framework*

## Chapter 3: Research Methodology

### 3.1 Introduction

A research methodology is a systematic way to conduct research. This research was aimed at developing an algorithm for identification of terrorist attack events and hotspots using k-means clustering algorithm and discriminant analysis Approach. This chapter describes the research methodology and research methods that were used. In the system development methodology, all steps of the appropriate methodology are analyzed. The research design appropriate for this type of research is looked at as well as data collection and data analysis procedures.

### 3.2 Agile Software Development Methodology

The Agile Software development is characterized by being iterative, with evolving requirements, regular feedback, client involvement and fast output of a prototype. This prototype can always be improved upon in another iteration process of the agile methodology(Kumar & Bhatia, 2012). This system development methodology was considered for this project because of the iterative nature of the development life cycle. When utilized in this project it allowed for repeated improvements on the algorithm being developed. The Figure 3-1 below illustrates the various phases of the agile software development methodology.

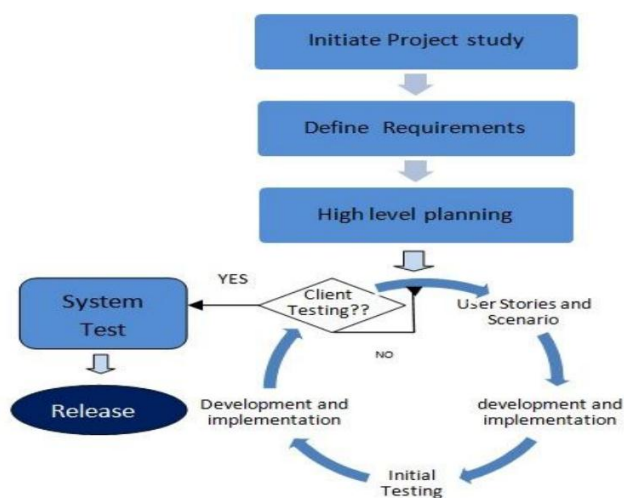


Figure 3- 1 : Phases of Agile Methodology

Source: (Thakur and Kaur, 2013)

### **3.2.1 Requirements**

This step involved collecting the necessary prerequisite data for conducting the project from existing literature and various stakeholders involved in the research for each iteration in the agile methodology.

### **3.2.2 Design**

The design stage involves defining the architecture and design of the system. This was achieved through a well-researched and thought out novel algorithm.

### **3.2.3 Development**

The third step called development stage involves actual implementation of the algorithm developed in the design stage. This stage involved Programming the proposed system by implementing the proposed algorithm. R programming language was used to achieve this.

### **3.2.4 Testing**

Testing involved verifying if the identification algorithm designed and implemented actually performs well, and whether successfully deployed in the application area of concern, which is counterterrorism.

### **3.2.5 Implementation and Deployment**

This step involved releasing the final product of an iteration for actual use in another iteration until the final aggregated tool was achieved. In this study, this involved integrating the developed model together with ESRI's ArcGIS desktop software on a specially prepared geodatabase to visually show areas spatially on a map.

### **3.2.6 Review**

This stage involved gathering feedback and new ideas for the product that has just been deployed and is in use. Any problems that arose were noted. With this new insight, the next iteration could always start.



### **3.3 Research Design**

Research design is concerned with how the research is conducted. The research objectives and consequently source of data, data collection methods and tools and finally data analysis guided the research design. This was a constructive research that aimed at developing an algorithm to identify terrorist attack events and hotspots. The model developed using this algorithm was evaluated using a specially prepared geodatabase. The research incorporated both qualitative and quantitative methods of research.

#### **3.3.1 System Analysis**

Structured System development approach was adopted for this research. In Structured System approach a list of instructions would be executed in the order they were written by default. The development process involved writing of R programming language scripts that would run on the back-end of the system and be executed whenever the system was being used for terrorism event identification. R uses generic functions to process data, which is normally passed as arguments.

Structured System Analysis (SSA) was done to identify functions that to be designed to handle data in the R scripts. Unified Modelling Language (UML) was used to design the Entity Relation Diagrams (ERD), Context diagram and Data Flow Diagrams (DFD).

#### **3.3.2 System Design**

Structured System design (SSD) techniques will be used to transform the output from the System Analysis stage into actual designs techniques for the specific objects identified. It involves Logical data modelling, data flow modelling and entity behavior modelling. A context diagram was used to show the interaction of the system with external entities. DFD were used to describe how data flows through the processed in the algorithm, in this case the functions and logic points

### **3.4 Target Population and Sampling/Experimental setup**

The target population is the specific group of focus for the study. This study considered the characteristics of events that have happened in preparation of a terrorist event as perpetrated by the terrorists. These include recruitment and radicalization, planning meetings, training, preparatory activities like smuggling weapons and illegal immigration to a target location. Towns that have been hit by terrorists were the point of focus, as well as those which have not but appear to have the preparatory events recorded. The GTD dataset provided by START was used, and data on the locations attacked by terrorists was obtained using purposive sampling. The study worked with the incidents that were present in the developed geodatabase from terrorist incidents extracted from the GTD database for four purposive sampled countries: Kenya, Nigeria, Somalia and Iraq.

Despite this sample of countries considered, the tool was not made to be specific to any region of the world, it could be applied globally on any pre-terror activities to determine if they can lead to a terror event.

### **3.5 Data Collection**

Data collection involves collecting information from the relevant sources to guide the researcher in solving the research questions.

Data retrieved from the Global Terrorism Database (GTD) provided the locations that have experienced terror related activities. This dataset was retrieved from START's official website. The study worked with the incidents that were present in the developed geodatabase from terrorist incidents extracted from the GTD database for four purposive sampled countries: Kenya, Nigeria, Somalia and Iraq. The researcher worked with the column on Country, Town, Terrorist Organization, Latitude and Longitude data from the GTD dataset, the other columns on a custom eventID, Recruitment, Planning and Preparatory risk value columns and Terror\_Event were created.

Pre-terrorism activity: recruitment, planning and preparatory events data was retrieved from research that had already been done to identify them. The pre-terror activities' determinant sub-activities were identified but they did not have a consistent data type and value that could be used to model the risk associated with them when it came to overall determining the possibility of a terror event. The research involved an improvisation technique of imputing a range of values to represent risk factors of High, Medium and Low, based on a scale of the range of values from 1.0 – 10.0. Values between 1.0 – 4.0 represented Low risk, 4.1 – 5.9 represented Medium risk and 6.0 – 10.0 represented High risk. A new geodatabase that supports plotting on a map characteristics variables of an area that has previously been hit by terrorism was made.

### **3.6 Data Analysis**

The analysis of data in this research study involved the process of cleaning the data to remove any missing values, inspecting and transforming the data into a format that can be analyzed to get insight on the objectives of the study. This format is the geospatial data. This research utilized both qualitative and quantitative data analysis methods. For the qualitative data analysis, the study relied on deductive approach because there was inadequate time and data resources to extensive research.

### **3.7 Research Quality**

Research Quality involves check against two criteria: reliability and validity. Reliability involves the ability to replicate processes and achieve consistent results while validity is the extent to how appropriate the tools, techniques and data used are for the research problem.

The geospatial dataset developed from the GTD dataset was divided into areas that have experienced terrorism before, and areas that have not, as determined by pre-terror activities perpetrated by the terrorists in preparation, through the k-means algorithm. After identification, the two categories were labelled as

“Yes” and “No” in terms of experiencing terrorism. This newly labelled dataset was divided into two, a training set to be used to train a discriminant analysis model so that it can learn patterns in the data, and a testing dataset to evaluate the accuracy of terror event detection and consequently hotspot location identification.

### **3.8 Ethical Consideration**

The study did not require collection of data from the field, therefore the ethical approval that was sought was through the Strathmore Institutional Ethics Review Committee (SU-IERC). Research permit from the National Commission of Science, Technology & Innovation (NACOSTI) was also secured. The study involved identification of future terror events, and the cost of misidentification had to be considered. If a terrorist attack was to be identified as imminent and does not take place, there will be a big relief, but people always live in fear. On the other hand the cost of misidentification is greater when an imminent attack is dismissed off and it actually ends up happening.

## **Chapter 4: System Analysis, Design and Architecture**

### **4.1 Introduction**

This chapter describes the analysis of requirements and design of the proposed system based on the algorithm designed. UML diagrams were used to show the system design and architecture, various actors and their interaction with the system.

### **4.2 System Analysis**

System analysis involves coming up with the requirements of the system; what the system should do. This study was aimed at developing an algorithm that will identify a potential terrorist attack event and allow an area to be labelled as a terror hotspot. This analysis section outlines the functional and non-functional requirements that were provided in the implemented solution.

#### **4.2.1 Requirements Gathering**

Data collection involves collecting information from the relevant sources to guide the researcher in solving the research questions.

Data on pre-terror activities such as recruitment and radicalization, planning meetings, training, preparatory activities like smuggling weapons and illegal immigration to a target location was retrieved from various existing research literature done by researchers working with similar data for different implementation and tools. This specific information guided the research on the aspect of defining credible rules for the algorithm developed.

Data retrieved from the GTD provided the locations that have experienced terror related attacks in the past. This dataset was retrieved from START's official website. This dataset was transformed into a geospatial dataset that can be loaded as a geodatabase in Geographic Information System like ESRI's ArcGIS. A new dataset of places that have not been hit by terrorism, yet appear to have had some similar pre-terror activities was also developed by the researcher and also transformed into a geospatial dataset.

The analysis of data in this study involved the process of cleaning the data, inspecting and transforming the data into a geospatial format that could be analyzed to get insight on the objectives of the study. This research utilized both qualitative and quantitative data analysis methods.

#### **4.2.2 Functional Requirements**

Functional requirements are characteristics that the system must have; what it is expected to do when being used. The functional requirements gathering involved looking at the specific functions of the system and methods of operations it would have, and its responses to user inputs and any exceptions encountered. The main categories of system actors were regular users and administrators. The following were the functional requirements:

- i. User can launch the system to commence use.
- ii. Check safety state of an area – Users can use the system to key in the required variable values that explain pre-terror event such as recruitment of young people in an area for example age and their employment status.
- iii. Generate reports inform of statistical data and visual elements like map images that has data on them – Users can generate reports based on the information generated after analysis of whether a terror event is imminent based on pre-terror events values.
- iv. Manage datasets and retrain model – Administrators can update the datasets that the model is learning from and retrain it.
- v. Manage system use and scaling and maintenance issues – the administrators should be able to assist with system crashes or any form of help as requested by a user.
- vi. Users and administrators can save their progress and close the system after use.

### 4.2.3 Non-Functional Requirements

Non-functional requirements deal with the design, quality features and constraints and external factors that may influence the system's performance. The following were the non-functional requirements:

- i. Availability – The system should always be available for use.
- ii. Reliability – The system should give reliable and consistent results after analysis and processing.
- iii. Accuracy – The system will be providing sensitive output, therefore it should have a very minimal error rate.
- iv. Usability – The system should be designed and developed with a layman user in mind, a GUI interface is preferred.
- v. Maintainability – The system should be easily corrected and fixed in case of an error detection.
- vi. Scalability – The system should be able to cover more geographical areas, and their characteristics incorporated.

### 4.3 System Architecture

The system architecture shown in the Figure 4-1 below illustrates the layout of the application implementing the algorithm.



*Figure 4- 1 System architecture*

## 4.4 System Design

System Design involved creating diagram models of the system to be developed based on the analysis report, using the UML notation. This study adopted the Structured System Design methodology.

### 4.4.1 Context Diagram

A context diagram defines the boundaries of the system to be developed. It identifies the flow of information between the system and external entities. The entire software system is shown as a single process. It is also known as a level 0 data-flow diagram. There were two major categories of system users identified for the system; the normal users and the administrator. Both interact with the system providing various data items and instructions and getting results from the system. The Figure 4-2 below shows a context diagram for the system.

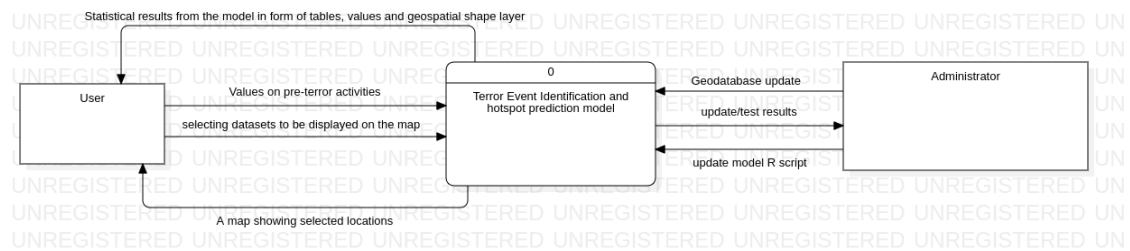


Figure 4- 2 Context Diagram

### 4.4.2 Data Flow Diagram (DFD)

The DFD models the flow of data through the system and activities that process the data. Data flow diagrams have levels, starting with level 0 (context diagram), level 1, level 2 and so on. As the level number increases, the complexity and detail of the system represented by the flow diagram also increases. Sub-processes begin to appear, as well as data stores. For the design of a system model developed from the proposed algorithm, the Figure 4-3 shows the level 1 data flow diagram designed.



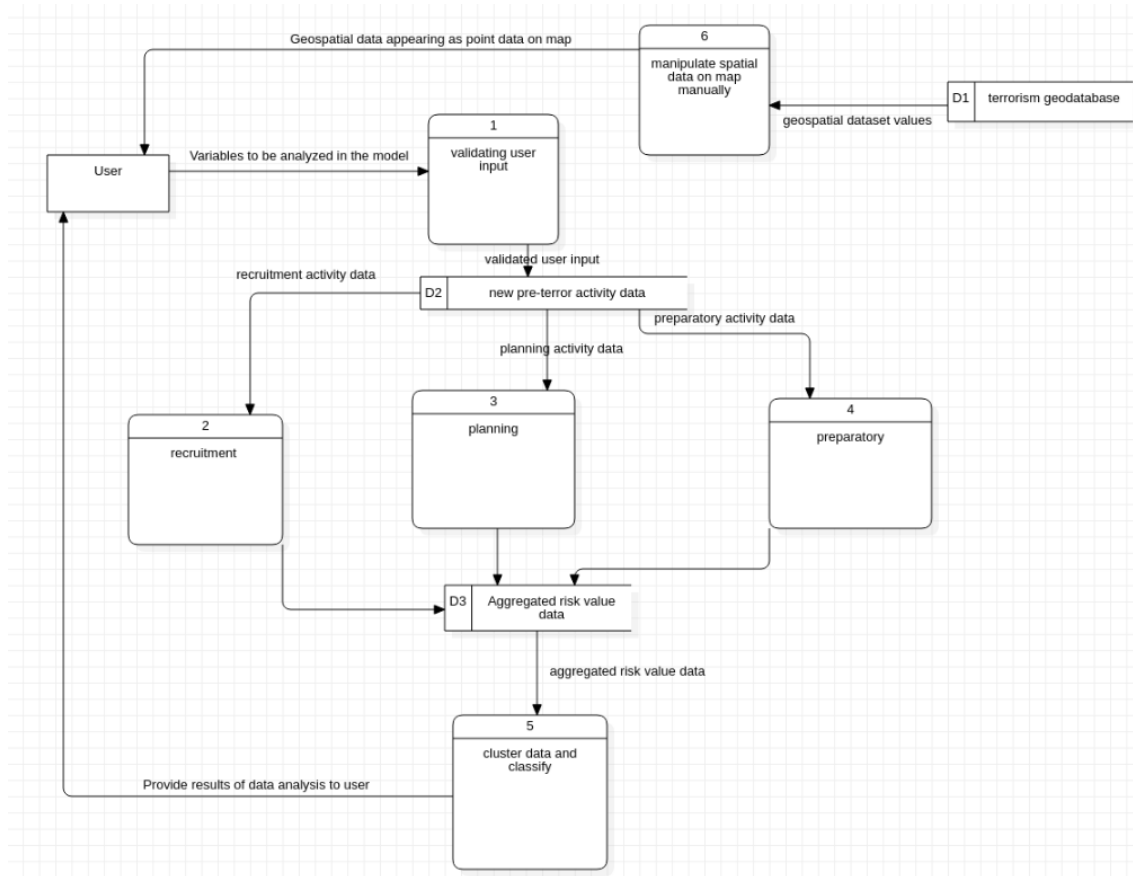


Figure 4- 3 Data Flow Diagram

#### 4.4.3 Entity Relation Diagram (ERD)

An ERD is a data modeling technique that graphically illustrates an information system's entities and the relationships between those entities. An ERD is a conceptual and representational model of data used to represent the entity framework infrastructure.

Elements of an ERD include: entities, relationships and attributes. The Figure 4-4 shows the designed ERD for the system.

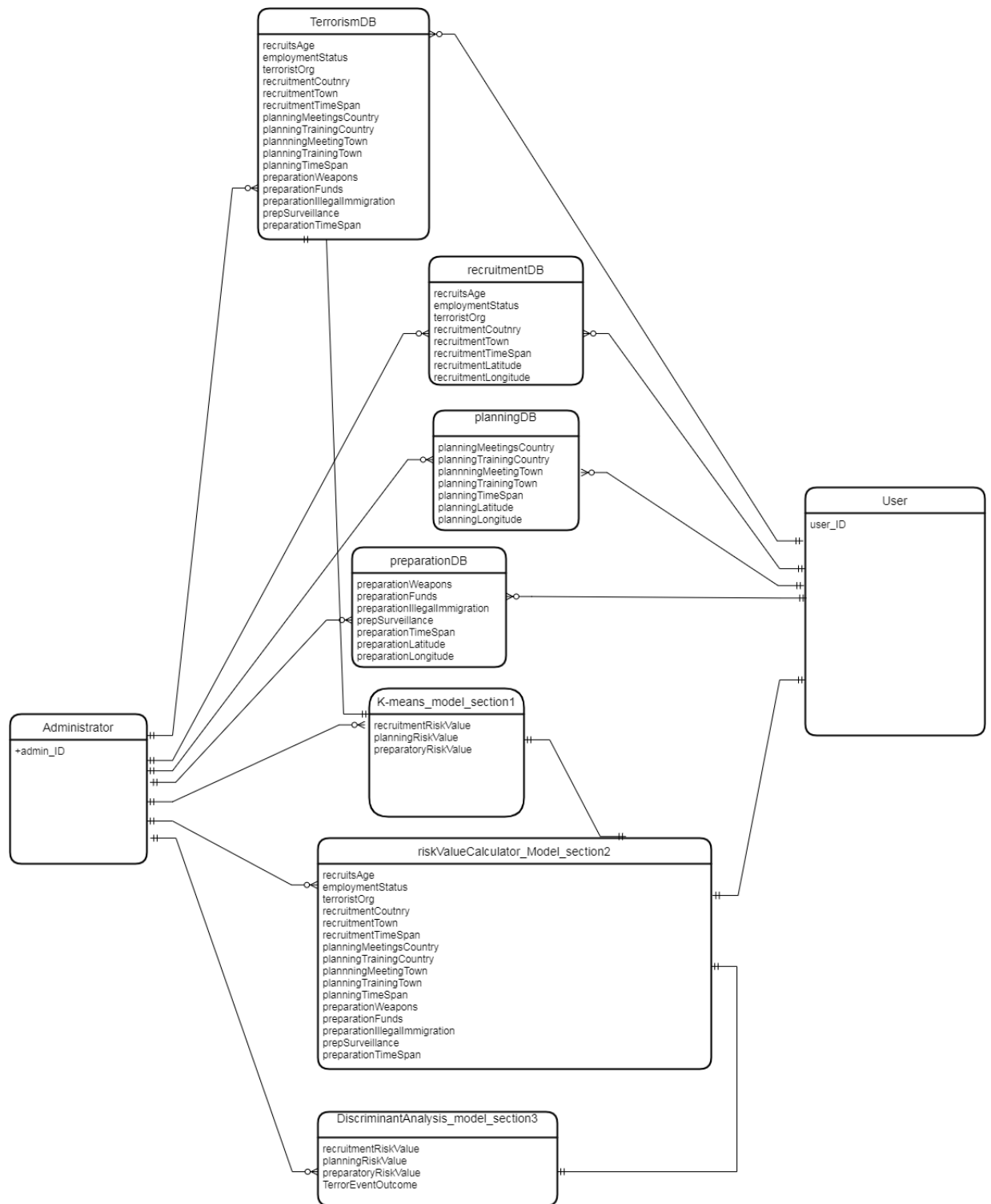


Figure 4- 4 ERD Diagram

#### 4.4.4 Wireframes of the system

A wireframe is used to design the desired layout of the system to be developed. It establishes a basic structure of the main interfaces before full styling, content and logic functionality is implemented. Figure 4-5 below shows the designed system wireframe for the system implemented using the proposed algorithm.



Figure 4- 5 System Wireframe.

## **Chapter 5: System Testing and Implementation**

### **5.1 Introduction**

This chapter describes how the algorithm was defined and used to develop an identification model. How the model was tested and validated is also covered. The process of building a geodatabase of geospatial datasets was the first step in the process of implementing the algorithm. The second step involved building the model using R programming language and testing it for identification accuracy.

The final section of this chapter describes the use of the model in identifying future terrorist attack events and showing highlighted hotspot areas.

### **5.2 System Implementation**

The system implementation started with preparation of a geospatial dataset from the existing terrorism data. It involved data cleaning and imputing missing values.

Using Microsoft Excel software, a comma separated value (.csv) file was created. The csv file was loaded with data selected from the GTD dataset retrieved in purposive sampling technique. The columns that were of interest to the researcher from the GTD dataset were the Country, Town, Terrorist Organization, Latitude and Longitude data. A column was created for denoting whether the terrorist attack event happened and this was populated with the character value "Y", to show that it did. The following columns were added, a custom eventID, Recruitment, Planning and Preparatory risk value columns which represents values that can be analyzed to explain the outcome, which is Terror\_Event.

The researcher also created and imputed extra rows of data for risk values on Recruitment, Planning and Preparatory that would lead to there being No terror event, ie "N".

For the specific predictor variables: Recruitment, Planning and Preparatory activities, datasets were also created based on the risk values of the characteristics of the sub-activities that constitute them: where they took place, duration of time they took place, age demographic of those involved, the group name of the terror organization, any legal and illegal ancillary activities such as raising and transfer of funds, training and weapon movement, travelling to target locations and surveillance activities' risk values. When these risk values are combined, they determine the final risk value of the predictor variable, which would in turn be used to produce the final risk value that exists in the first dataset mentioned earlier.

The .csv dataset was transformed to a geospatial dataset, a shape file (.shp) for showing point data using ArcGIS, which is a GIS owned by a company called Esri.

The proposed algorithm was interpreted as represented in the equations below, where the data items were the exact ones used for the study. The variables used in the study were selected from the basis of what existing research reports on their effect on the final outcome of a planned terrorist attack.

### **The K-Means Clustering applied equation**

Two functions for the two clusters were represented as shown in the equation 5-1 below:

$$\text{YesTerrorEvent} = \sum_{i=1}^n ||x_i^{(j)} - c_j||^2$$

$$\text{NoTerrorEvent} = \sum_{i=1}^n ||x_i^{(j)} - c_j||^2$$

*Equation 5- 1 Application of the K-Means clustering equation to the study variables*

$\sum_{i=1}^n$  Represented the number of variables that were considered for the K-Means, in this case 3, recruitment, planning and preparative risk values.

$||x_i^{(j)} - c_j||^2$  represented the distance function, with  $x_i^{(j)}$  being case One of the 3 variables recruitment, planning and preparative risk values, and  $c_j$  the centroid for either of the clusters YesTerrorEvent or NoTerrorEvent.

### **Discriminant analysis applied equation**

The equation 5-2 below shows the determination of the two discriminant functions that would be used to identify a terror event or lack of, where a is the intercept,  $W_i$  is the discriminant coefficient for the independent variable i, in this case recruitment, planning and preparatory variables.

DAfuncYesTerrorEvent

$$= a + W_1(\text{recruitment}) + W_2(\text{planning}) + W_3(\text{preparatory})$$

DAfuncNoTerrorEvent

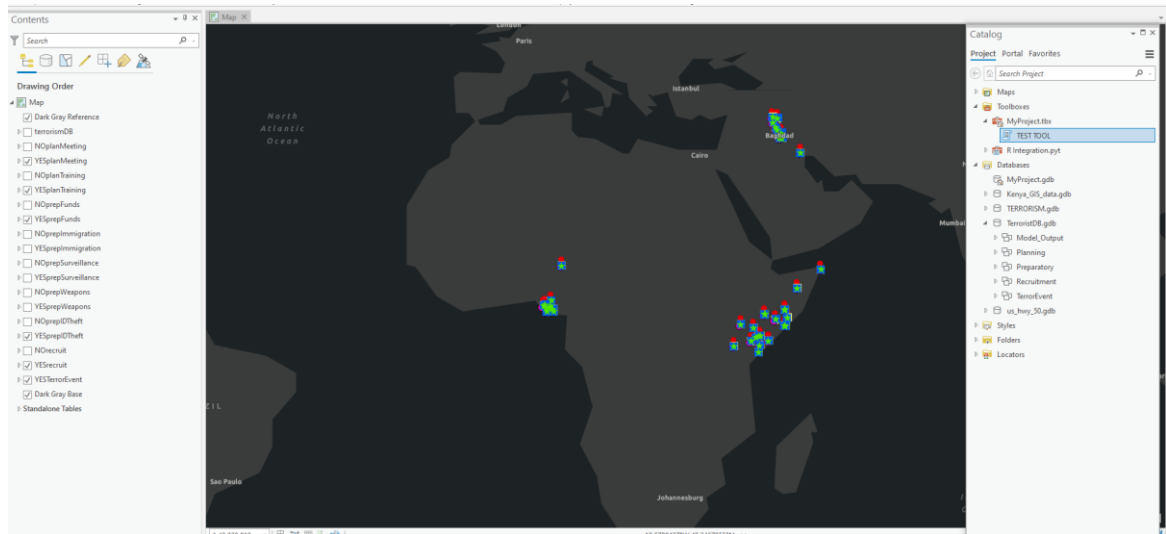
$$= a + W_1(\text{recruitment}) + W_2(\text{planning}) + W_3(\text{preparatory})$$

*Equation 5- 2 : Discrimination functions from values of the determinant variables*

After the necessary data was prepared, the second step which is the implementation of the proposed algorithm as a model commenced. R programming language was selected as the implementation tool and R Studio as the Integrated Development Environment (IDE). R was selected as it offers a lot of open source packages are very useful in statistical analysis and machine learning. The model was developed to run as an R script being linked and called from the earlier mentioned Arc GIS program. This linkage is possible because of a tool known as the R-arcGIS-Bridge. The model of the computer used to implement and run the model was a HP Probook 4230s computer, running windows 10 Operating System.

The ArcGIS provided the User Interface. There is a map interface that shows the user the location of pre-terror activities, from the recruitment, planning, preparatory and the actual terrorist event location. This is a visual representation of the transformed .csv file dataset into a geospatial dataset. A

user hides or displays a particular pre-terror event indicator on the left pane, in the map as shown in Figure 5-1 below.



*Figure 5- 1: ArcGIS Interface showing the pre-terror events in Kenya, Nigeria, Somalia and Iraq as these were the selected areas in the dataset used.*

A user can interact with the data input section by launching and filling the required data for the variables that are needed to give the three pre-terror activities: recruitment, planning and preparation events a risk value. The user can select the geospatial dataset they want to run the variables against, and also specify the output folders of the statistical analysis plots and tables produced by the model, as shown below in Figure 5-2.

**Select Dataset**

\* Age

\* Employment Status

\* Terror Group

\* Recruitment Country

\* Recruitment Town

\* Recruitment Time Span

\* planning Meeting Country

\* plan Meeting Town

\* plan Training Country

\* plan Training City

\* planning Time Span

\* prep Weapons

\* prep Funds

\* prep Illegal Immigration

\* prep Surveillance

\* prep Time Span

\* Discriminant Analysis Table

\* DA Predicted

\* KMEANS Predicted

\* KMEANS Analysis Table

**Run**

*Figure 5- 2: Data Input form for user*

Once the data was submitted to the R script for analysis when user clicked the “Run” button, it was passed to a function in the R script that was expecting the exact given set of input parameters, and output links to folders. This function



had to have the name `tool_exec( in_params, out_params)`, see the figure AD1 from Appendix D.

Immediately the input and output parameters were received, they got stored in local variables and were passed down to a system of logical statements that determined whether their value posed a high, low or medium risk, in relation to that particular pre-terror activity pointing to an actual terror activity, as illustrated in the figures AD-2, AD-3 and AD-4 from Appendix D. The final risk values from all the three pre-terror activities determinants were aggregated using a simple count algorithm, see the figure AD-5 from Appendix D, to determine which risk level appeared frequently, and this was be the overall risk state of a particular area.

After the final risk value had been evaluated, it had to be checked against the existing dataset values, which were used to place the new unseen risk values of recruitment, planning and preparatory in one of two clusters, either a terror event happens or it does not, using the k-means algorithm as implemented in the R script. This is shown in the figure AD-6 from Appendix D. The algorithms second part was a discriminant analysis implementation to predict the final `Terror_event` column value from the existing dataset which was subdivided into training and testing datasets in magnitudes of 66% and 34% respectively. It was used to predict the `Terror_event` column value of any provided pre-terror events risk values. Refer to the figure AD-7 from Appendix D.

The results from both the k-means clustering and discriminant analysis four tables of analysis results, which were sent to the ArcGIS for view by the user.

In the case where data is missing for a specific variable such as in the case of communication and meetings being virtual hence lacking a spatial characteristic, or as in case of the terrorists not fully following their crime template by either skipping some pre-terror activities or not following them in a linear order, their risk values can be calculated by altering the logic to expect

a binary input value. For example during data input, the presence of detection of communication and meeting without a spatial aspect could be given a medium risk value in the logic, as opposed to a high risk value when the locations are known. In the case of terrorists not following their crime template in conducting the pre-terror events, the independently detected activities could be given a medium risk value because it infers lack of order in the strategy, as opposed to a scenario where activities are taking place one after another, the previous one triggering the next and a trend is visible, which would be a high risk.

### **5.3 System Testing**

The system aimed at predicting the identification of a possible terror event, and a highlighted hotspot location. The model was given a set of pre-terror activity events and their constituent values, and from this it was able to run a k-means clustering instance and obtain 2 clusters, one for when a terror event resulted from those values of the pre-terror events, and another for when there was no terror event given the values of the pre-terror events. These clusters could be compared to the actual states of whether the terror event happened or not, so as to check the accuracy of the identification process.

Unit testing was implemented in the input form from ArcGIS internally. When the user provides a new unseen set of pre-terror event values, they cannot leave a particular field blank. The user is not also able to field a wrong data type value in a field, as illustrated in the Figure 5-3 below:

\* plan Training City

\* planning Time Span

\* prep Weapons

\* prep Funds

\* prep Illegal Immigration

\* prep Surveillance

\* prep Time Span

\* Discriminant Analysis Table

\* DA Predicted

\* KMEANS Predicted

\* KMEANS Analysis Table

21 Parameters are missing or invalid

Run

Figure 5- 3 A warning message to tell the user not to leave out blank fields.

In the R script code internally, some logic was put in place to make sure that the system does not run when the keys in invalid data. For example a user can field a negative value as age or duration of a particular event, the model should not process that as illustrated in figure AD-8 from Appendix D.

Table 5-1 A and B below shows system testing done:

Test Case	Importance Level	Results
Can the user launch the system to use?	High	The system launches once the user clicks on it.
Can the user generate reports from the system?	High	The system will generate any desired report output and displayed or stored.
Does the system allow the administrator to update datasets or the logic in the model?	High	The system accepts any new updates from the administrator.

Table 5- 1 A : List of Different Test cases and Results done on the system.

Test Case	Importance Level	Results
Can the system fail safely	Medium	The model was integrated into the ArcGIS software which fails safely in case of a crash.
Can the system allow user to save current progress and resume later?	Medium	The model was integrated into the ArcGIS software which allows saving of progress on analysis.

*Table 5- 2 B : (Cont..)List of Different Test cases and Results done on the system.*

### Testing System Accuracy

The accuracy of the system in correctly clustering pre-terror events was looked at, and below are the results per the incidents looked at versus what is the actual state of events. The dataset that was used for testing the model developed had 61 countries, 41 of which are positive for terrorist attacks and 20 did not have any attack despite having detected some characteristics similar to that of the ones that led to a terror attack, but in different magnitudes.

The Figure 5-4 below shows the results from the k-means algorithm from the analyzed 61 countries:

OBJECTID *	Var1	Var2	Freq
1	1	N	19
2	2	N	1
3	1	Y	0
4	2	Y	41

*Figure 5- 4 Results from k-Means clustering algorithm*

The results were that 19 items were placed in cluster 1, and this represented the pre-terror events that did not lead to a terrorist event. 41 items belonged to

cluster 2 which represented pre-terror events that led to a terror event. 1 item was misclassified into cluster 2, yet it belonged to cluster 1 as per the var2 label shown in figure 5-4 above.

The model therefore demonstrated good accuracy level.

The figure 5-5 below shows the output from the testing dataset of the discriminant analysis tool. There were 8 items correctly classified under no terror event happened, and 12 events classified correctly as happened. 1 item was misclassified into the category terror event happened. This could be interpreted as an outlier because the k-means algorithm above showed the same behavior in output.

OBJECTID *	pred_class	Var2	Freq
1	N	N	8
2	Y	N	1
3	N	Y	0
4	Y	Y	12

*Figure 5- 5 Results from Discriminant analysis algorithm*

#### 5.4 System Validation/Deployment

The system was used to validate a user entry of unknown and unseen values of pre-terror event activities and sub-activities' characteristics. The results are shown in the figures 5-6 and 5-7.

**Select Dataset**  
 terrorismDB

Age: 23

Employment Status: 1

Terror Group: Unknown

Recruitment Country: Kenya

Recruitment Town: Nairobi

Recruitment Time Span: 20

planning Meeting Country: Uganda

plan Meeting Town: Nairobi

plan Training Country: Kenya

plan Training City: LAMU

planning Time Span: 12

prep Weapons: 1

prep Funds: 1

prep Illegal Immigration: 0

prep Surveillance: 1

prep Time Span: 10

**Discriminant Analysis Table**  
 c45\_TESTTOOL

**DA Predicted**  
 c45\_TESTTOOL1

**KMEANS Predicted**

**Run**

Figure 5- 6 Input form with values from user

OBJECTID *	pred_class	Var2	Freq
1	N		0
2	Y		1

Figure 5- 7 The input instance of pre-terror values will result to a Y as predicted by the discriminant analysis tool.

## **Chapter 6: Discussion**

### **6.1 Introduction**

This chapter analyzes the results from the study relative to the objectives that were to be met. The purpose of this research was to develop an algorithm that can be used for the identification of terror events and hotspots, to provide awareness and aid counterterrorism, using k-means and discriminant analysis approach on pre-terrorism activities and sub-activities

### **6.2 Identifying pre-terror activities**

During literature review stage, it was discovered that minimal research had been conducted on the spatial and temporal characteristics of a potential terror event and hotspot. Many of the research papers were looking at how to utilize the already identified characteristics in various novel tools. However for the existing studies that do look at identifying the characteristics, they ended up with the same set. This gave confidence in the characteristics selected for this research. These included but not limited to terrorist cells creation, recruitment activities and radicalization, planning activities such as making a lot of phone calls to areas known to be a terrorist harbor, training the terrorists on using weapons and espionage and preparatory activities like assembling weapons like guns and bombs, getting funding, travelling to target location, surveying the target location, and ancillary crimes to maintain order such as murder of informants.

It is evident that the characteristics are not very many and in most cases, there tends to be a thin line between a normal criminal activities such as moving weapons from one gang member to another for normal robbery crimes versus moving weapons for terrorism purposes. This therefore means that a pre-terror activity must be put in context, so as to be analyzed together with other pre-terror activities to determine the overall pattern of leading to a terror event. This study maintained the same variables that represent the characteristics of a terror event happening, which is presence of recruitment, planning and preparatory activities.

### **6.3 The existing tools for identification**

During the research process it was discovered that research has been done on the process of identification of terror hotspots, using spatial and temporal datasets. The techniques looked at in the literature review part include: Hawkes process, Risk terrain modeling framework, Prospective Space - Time Scan Statistics and the Random Forest Method. Some research material sought to highlight the final concentration of the terror events as the likely location of a potential attack because of the concept on repeat victimization, which states that an area that has had a terrorist attack is more likely to experience another terrorist attack, than an area that has not been attacked before. Such techniques are very useful in explaining how the characteristics of such hotspots areas make it a suitable environment for attacks. Factors such as economy of an area, demographics like number of youth and population density, type of target location - military base and terrain of an area for example can explain why it was possible for that terror event to take place, but cannot be used to identify another terrorism event based on similar characteristics. This leads to stereotyping and bias for example policing heavily a certain region because people of a particular ethnic group or religion reside there. Based on this finding it was therefore important to look at tools and techniques that analyze data that actually can be objective in its prediction of terrorist attack events.

### **6.4 K-means and Discriminant analysis combined algorithm**

The research sought to identify terror events before they happened based on past pre-terror events. This means that in order to get a valid and reliable output, the algorithm had to identify patterns in activities that could potentially lead to a terrorism event and similar ones that would not. Hence it was important to first implement a tool that first places these activities into clusters based on their values. The k-means algorithm was used for that grouping purpose. From this clusters, labels of whether a terror events happens given a set of values for the pre-terror activities, were assigned. From this stage the statistical method of discriminant analysis could be used to confirm the end



results of the cluster the k-means algorithm would assign to a new set of pre-terror activity values. Another thing to note is that the terrorism identification process had to be general enough to be used in any other area in the world.

### 6.5 Validating the k-means and Discriminant analysis combined algorithm

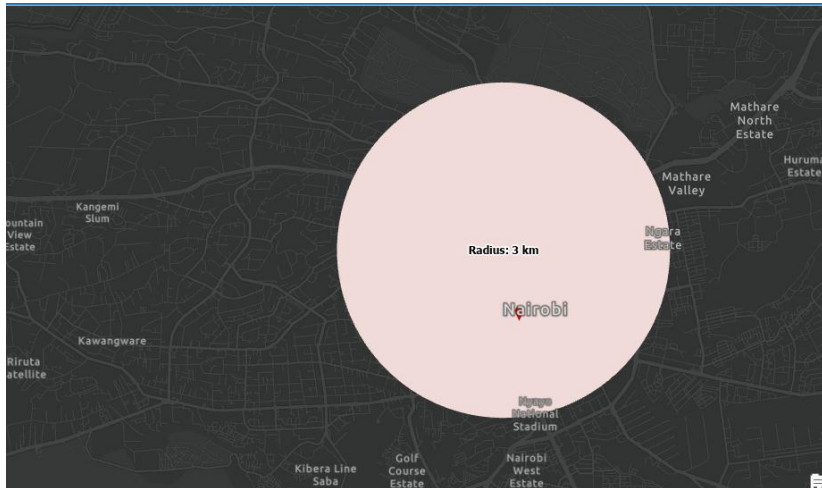
The tool developed was tested for reliability using a specially prepared dataset. This dataset was first created as a .csv file and later geocoded into a geospatial dataset so that it could be plotted on a map as points. The dataset had a field called Terror\_Event that had a value of either be “Yes” or “No”. This field was firstly hidden from the model, and it was up to the model to find the patterns that could lead to a terrorist attack by analyzing the risk factor values in the pre-terror activities field of recruitment, planning and preparatory events. The model was tested for performance by comparing what was the actual values in the Terror\_Event, versus the cluster the values appeared to be in after the k-means algorithm. This gave the confidence level of the model, as shown in the Figure 6-1 below:

OBJECTID *	Var1	Var2	Freq
1	1	N	19
2	2	N	1
3	1	Y	0
4	2	Y	41

*Figure 6- 1 Results from k-Means clustering algorithm*

In the case of hotspot area identification, based on the combination of ideas from the research done by Smith et al.(2006), where it is discovered that over 50% of terrorist attacks took place within a radius of 0-45km from the planning and preparation location and from Onat (2019), who showed that specific buildings ‘ concentration over space and even proximity to such buildings can

raise the risk value of an area to a terrorist attack, a radius from the last reported known planning or preparation location can be used to draw the outline of a circumference which would enclose the hotspot areas. The Figure 6-2 below shows this.



*Figure 6- 2 An area highlighted as a possible terror hotspot.*

## **Chapter 7: Conclusion and Recommendation**

### **7.1 Conclusion**

Predicting exactly where and when a terrorist attack will occur is a very difficult task, given that a lot of varying factors go into planning and preparing for the attack. This fact however does not deter or demoralize any counterterrorism measures being put in place. In fact there have been successful reports of terrorism being stopped at one of its planning or preparatory activity stages. Some of the solutions that exist in trying to identify a terror attack before it happens end up just giving an estimate based on where the last event occurred and variables that were in play then, and if these variables were to be used to predict a terror event, it would lead to a biased model. This research aimed at using data collected on pre-terrorism activities and attempt to identify whether the final terrorism event happens or not, how soon and even where. The pre-terror activities included: recruitment, planning and preparatory events. The research set to achieve four objectives as discussed below:

The first objective set to analyze the spatio-temporal characteristics/activities of pre-terrorism attack events and terror hotspots. Through extensive literature review, the following were identified: recruitment, planning and preparatory events. In detail they include terrorist cells creation, recruitment activities and radicalization, planning activities such as making a lot of phone calls to areas known to be a terrorist harbor, training the terrorists on using weapons and espionage and preparatory activities like assembling weapons like guns and bombs, getting funding, travelling to target location, surveying the target location, ancillary crimes to maintain order such as murder of informants

The second objective was to examine existing techniques used to identify terror events and hotspots using a geospatial dataset and a time variable. These techniques were analyzed and looked at extensively in chapter two, literature review. They include but not limited to Random Forest model, Space-time scan statistic tool and Hawkes process.

The third objective was to design and develop an algorithm for the identification of terror events and hotspots using k-means clustering and discriminant analysis approach. Agile software development methodology was considered. The design part was done using Structured System Analysis and Design. Implementation was done through developing the tool as an R script to run on ArcGIS, a system for analyzing and processing geospatial datasets.

The final objective of the study was to test the performance of the system; its reliability in identifying a terror event right before it happens. The model based on the validation dataset evaluated achieved a percentage of above 95% in its prediction of whether a terror event happens based on the dataset used.

## **7.2 Recommendations**

Based on the findings from the study, the following recommendations are made:

The system can be tested with a wide variety of datasets to from different parts of the world so as for the patterns of pre-terror activities to be visually visible globally.

An inclusion into the dataset and analysis of the type of infrastructure found in an area as a variable can be used to highlight buildings with a high risk in that identified hotspot, narrowing down the scope of a possible risk spot.

Have the model integrated as a server side program connecting to clients that are web based and mobile based for easier use – Google maps integration.

## **7.3 Future Work**

For the logic part of the algorithm, more combinations of the logic rules can achieve a stronger risk identifier. A statistical model that actually generates a value based on a specific risk value can be identified, as opposed to the random value generator, given a range, as used in the current study.

## References

- Afsar, M., Aslam, K., Minhas, A., & Iqbal, J. (2014, November 12). *Spatio-Temporal Analysis of Terrorism Incidents & Law Enforcement Operations in Pakistan (2008-12)*.
- Braithwaite, A., Li, Q., & Associate, L. (2007). Transnational Terrorism Hot Spots: Identification and Impact Evaluation. *Conflict Management and Peace Science - CONFLICT MANAG PEACE SCI*, 24.  
<https://doi.org/10.1080/07388940701643623>
- Cai, D., He, X., & Han, J. (2008). Training Linear Discriminant Analysis in Linear Time. *2008 IEEE 24th International Conference on Data Engineering*, 209–217. <https://doi.org/10.1109/ICDE.2008.4497429>
- Dikko, H. G., & Osi, A. A. (2014). Discriminant Analysis as an Aid to the Classification and Prediction of Safety across States of Nigeria. *International Journal of Statistics and Applications*, 4(3), 153–160.
- From Extremist to Terrorist: Identifying the Characteristics of Communities Where Perpetrators Live and Pre-Incident Activity Occurs* | [START.umd.edu](http://START.umd.edu).  
(n.d.). Retrieved March 19, 2021, from  
<https://www.start.umd.edu/publication/extremist-terrorist-identifying-characteristics-communities-where-perpetrators-live-and>
- Gao, P., Guo, D., Liao, K., Haney (Webb), J., & Cutter, S. (2013). Early Detection of Terrorism Outbreaks Using Prospective Space–Time Scan Statistics\*. *The Professional Geographer*, 65.  
<https://doi.org/10.1080/00330124.2012.724348>

- Gill, S. S. A. N. S. (2013). Analysis and Study of K-Means Clustering Algorithm. *International Journal of Engineering Research & Technology*, 2(7). <https://www.ijert.org/research/analysis-and-study-of-k-means-clustering-algorithm-IJERTV2IS70648.pdf>, <https://www.ijert.org/analysis-and-study-of-k-means-clustering-algorithm>
- Guo, W., Liu, H., Yu, A., & Li, J. (2016). RESEARCH ON VISUAL ANALYSIS METHODS OF TERRORISM EVENTS. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B2, 191–196. <https://doi.org/10.5194/isprs-archives-XLI-B2-191-2016>
- Hao, M., Jiang, D., Ding, F., Fu, J., & Chen, S. (2019). Simulating Spatio-Temporal Patterns of Terrorism Incidents on the Indochina Peninsula with GIS and the Random Forest Method. *ISPRS International Journal of Geo-Information*, 8(3), 133. <https://doi.org/10.3390/ijgi8030133>
- Hasisi, B., Perry, S., Ilan, Y., & Wolfowicz, M. (2020). Concentrated and Close to Home: The Spatial Clustering and Distance Decay of Lone Terrorist Vehicular Attacks. *Journal of Quantitative Criminology*, 36. <https://doi.org/10.1007/s10940-019-09414-z>
- Khalsa, S. K. (2005). Forecasting Terrorism: Indicators and Proven Analytic Techniques. In P. Kantor, G. Muresan, F. Roberts, D. D. Zeng, F.-Y. Wang, H. Chen, & R. C. Merkle (Eds.), *Intelligence and Security Informatics* (pp. 561–566). Springer. [https://doi.org/10.1007/11427995\\_59](https://doi.org/10.1007/11427995_59)

- Laub, P., Taimre, T., & Pollett, P. (2015). *Hawkes Processes*.
- Li, Y., & Wu, H. (2012). A Clustering Method Based on K-Means Algorithm. *Physics Procedia*, 25, 1104–1109.  
<https://doi.org/10.1016/j.phpro.2012.03.206>
- Mei, H., & Eisner, J. (2017). The Neural Hawkes Process: A Neurally Self-Modulating Multivariate Point Process. *ArXiv:1612.09328 [Cs, Stat]*.  
<http://arxiv.org/abs/1612.09328>
- Mo, H., Meng, X., Li, J., & Zhao, S. (2017). Terrorist event prediction based on revealing data. 239–244. <https://doi.org/10.1109/ICBDA.2017.8078815>
- Nemeth, S. C., Mauslein, J. A., & Stapley, C. (2014). The Primacy of the Local: Identifying Terrorist Hot Spots Using Geographic Information Systems. *The Journal of Politics*, 76(2), 304–317. JSTOR.  
<https://doi.org/10.1017/s0022381613001333>
- Onat, I. (2019). An analysis of spatial correlates of terrorism using risk terrain modeling. *Terrorism and Political Violence*, 31, 277–298.  
<https://doi.org/10.1080/09546553.2016.1215309>
- O'Neil, S. (2007). Terrorist Precursor Crimes: Issues and Options for Congress. *Undefined*. /paper/Terrorist-Precursor-Crimes%3A-Issues-and-Options-for-O%27Neil/b0fa17e76af1ceeab659134c6d2fc35cfda2f2a9
- Pakhira, M. K. (2014). A Linear Time-Complexity k-Means Algorithm Using Cluster Shifting. *2014 International Conference on Computational Intelligence and Communication Networks*, 1047–1051.  
<https://doi.org/10.1109/CICN.2014.220>

(PDF) *Application of Discriminant Function Analysis in Agricultural Extension Research*. (n.d.). Retrieved August 15, 2020, from [https://www.researchgate.net/publication/321245298\\_Application\\_of\\_Discriminant\\_Function\\_Analysis\\_in\\_Agricultural\\_Extension\\_Research?enrichId=rgreq-503dfd9d493a35540a07b800dfe61781-XXX&enrichSource=Y292ZXJQYWdlOzM4MTI0NTI5ODtBUzo1NjM4NDQzNjQ5MzUxODRAMTUxMTQ0MjM4MDc5Ng%3D%3D&el=1\\_x\\_2&\\_esc=publicationCoverPdf](https://www.researchgate.net/publication/321245298_Application_of_Discriminant_Function_Analysis_in_Agricultural_Extension_Research?enrichId=rgreq-503dfd9d493a35540a07b800dfe61781-XXX&enrichSource=Y292ZXJQYWdlOzM4MTI0NTI5ODtBUzo1NjM4NDQzNjQ5MzUxODRAMTUxMTQ0MjM4MDc5Ng%3D%3D&el=1_x_2&_esc=publicationCoverPdf)

(PDF) *Impact of Agile Methodology on Software Development Process*. (n.d.). ResearchGate. Retrieved August 15, 2020, from [https://www.researchgate.net/publication/255707851\\_Impact\\_of\\_Agile\\_Methodology\\_on\\_Software\\_Development\\_Process](https://www.researchgate.net/publication/255707851_Impact_of_Agile_Methodology_on_Software_Development_Process)

Polo, S. (2020). How Terrorism Spreads: Emulation and the Diffusion of Ethnic and Ethnoreligious Terrorism. *Journal of Conflict Resolution*, 64, 002200272093081. <https://doi.org/10.1177/0022002720930811>

Porter, White: *Self-exciting hurdle models for terrorist activity*. (n.d.). Retrieved July 8, 2020, from <https://projecteuclid.org/euclid.aoas/1331043390>

Ramayah, T., Ahmad, N., Abdul-Halim, H., Rohaida, S., mohamed zainal, S., & Lo, M. chiun. (2010). Discriminant analysis: An illustrated example. *African Journal of Business Management*, 4, 1654–1667.

Saha, S., Aladi, H., Kurian, A., & Basu, A. (2017). *Future Terrorist Attack Prediction using Machine Learning Techniques*. <https://doi.org/10.13140/RG.2.2.17157.96488>



- Sandler, T. (2014). The analytical study of terrorism: Taking stock. *Journal of Peace Research*, 51(2), 257–271.  
<https://doi.org/10.1177/0022343313491277>
- Smith, B., Damphousse, K., & Roberts, P. (2006). *Pre-Incident Indicators of Terrorist Incidents: The Identification of Behavioral, Geographic, and Temporal Patterns of Preparatory Conduct*.
- Tench, S., Fry, H., & Gill, P. (2016). Spatio-temporal patterns of IED usage by the Provisional Irish Republican Army. *European Journal of Applied Mathematics*, 27(3), 377–402.  
<https://doi.org/10.1017/S0956792515000686>
- Tutun, S., Khasawneh, M. T., & Zhuang, J. (2017). New framework that uses patterns and relations to understand terrorist behaviors. *Expert Systems with Applications*, 78, 358–375.  
<https://doi.org/10.1016/j.eswa.2017.02.029>
- Wang, D., Ding, W., Lo, H., Morabito, M., Chen, P., Salazar, J., & Stepinski, T. (2013). Understanding the spatial distribution of crime based on its related variables using geospatial discriminative patterns. *Computers, Environment and Urban Systems*, 39, 93–106.  
<https://doi.org/10.1016/j.compenvurbsys.2013.01.008>
- Zhang, R., Walder, C., Rizoïu, M.-A., & Xie, L. (2018). *Efficient Non-parametric Bayesian Hawkes Processes*.

## Appendix A: Originality Report



### Document Information

<b>Analyzed document</b>	An Algorithm for Identification of Terror Events and Hotspots Using K-Means and Discriminant Analysis Approach.docx (D112324743)
<b>Submitted</b>	9/10/2021 5:02:00 AM
<b>Submitted by</b>	
<b>Submitter email</b>	Kelvin.John@strathmore.edu
<b>Similarity</b>	4%
<b>Analysis address</b>	library.strath@analysis.orkund.com

### Sources included in the report





<b>W</b>	URL: <a href="http://unigis.sbg.ac.at/files/Mastertheses/Full/103522.pdf">http://unigis.sbg.ac.at/files/Mastertheses/Full/103522.pdf</a> Fetched: 7/19/2021 10:10:09 AM	 <b>1</b>
<b>W</b>	URL: <a href="https://researchoutput.csu.edu.au/files/71300591/Ryan_Prox_thesis.pdf">https://researchoutput.csu.edu.au/files/71300591/Ryan_Prox_thesis.pdf</a> Fetched: 6/24/2021 11:24:10 AM	 <b>1</b>
<b>W</b>	URL: <a href="http://scholarworks.csun.edu/bitstream/handle/10211.3/172162/Martinez%20Reyes-Efren-thesis-2016.pdf?sequence=1">http://scholarworks.csun.edu/bitstream/handle/10211.3/172162/Martinez%20Reyes-Efren-thesis-2016.pdf?sequence=1</a> Fetched: 6/10/2021 8:05:12 PM	 <b>1</b>
<b>W</b>	URL: <a href="https://www.ncjrs.gov/pdffiles1/nij/grants/214217.pdf">https://www.ncjrs.gov/pdffiles1/nij/grants/214217.pdf</a> Fetched: 9/10/2021 5:03:00 AM	 <b>2</b>

Figure AA- 1: Originality Report

## Appendix B: Ethical Approval from SU-IERC



27<sup>th</sup> September 2021

Mr Ndambuki Kelvin,  
kelvin.john@strathmore.edu

Dear Mr Ndambuki,

### **RE: An Algorithm for Identification of Terror Hotspots Using Discriminant Analysis Approach**


This is to inform you that SU-IERC has reviewed and **approved** your above **SU- master's** research proposal. Your application reference number is **SU-IERC0940/20**. The approval period is **27<sup>th</sup> September 2021 to 26<sup>th</sup> September 2022**.

This approval is subject to compliance with the following requirements:

- i. Only approved documents including (informed consents, study instruments, MTA) will be used
- ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-IERC.
- iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-IERC within 48 hours of notification
- iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-IERC within 48 hours
- v. Clearance for export of biological specimens must be obtained from relevant institutions.
- vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
- vii. Submission of an executive summary report within 90 days upon completion of the study to SU-IERC.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology and Innovation (NACOSTI) <https://research-portal.nacosti.go.ke/> and also obtain other clearances needed

Yours sincerely,

  
for: Prof Fred Were,  
**Chairperson; SU-IERC**



Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000  
Email [admissions@strathmore.edu](mailto:admissions@strathmore.edu) [www.strathmore.edu](http://www.strathmore.edu)

*Figure AB- 1 : Research permit letter from SU-IERC as internal ethical approval*

## Appendix C: Ethical Approval certificate from NACOSTI

 <p>REPUBLIC OF KENYA</p>	 <p><b>NATIONAL COMMISSION FOR SCIENCE, TECHNOLOGY &amp; INNOVATION</b></p>
Ref No: <b>433772</b>	Date of Issue: <b>19/October/2021</b>
<b>RESEARCH LICENSE</b>	
	
<p><b>This is to Certify that Mr., Kelvin Ndambuki John of Strathmore University, has been licensed to conduct research in Nairobi on the topic: An Algorithm for Identification of Terror Events and Hotspots Using K-Means and Discriminant Analysis Approach for the period ending : 19/October/2022.</b></p>	
License No: <b>NACOSTI/P/21/13543</b>	
433772	
Applicant Identification Number	Director General
	<b>NATIONAL COMMISSION FOR SCIENCE, TECHNOLOGY &amp; INNOVATION</b>
	Verification QR Code
	
<p><b>NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.</b></p>	

Figure AC- 1 : Ethical approval certificate from NACOSTI

## Appendix D: Code Snippets of the Identification Model

```
tool_exec <- function(in_params, out_params) {  
  # Load required packages  
  arc.progress_label('Loading required R packages...')  
  arc.progress_pos(25)  
  pkgs = c('wakefield', 'MASS', 'factoextra', 'clue', 'dplyr')  
  load_pkgs(pkgs)  
  
  # Get parameters - INPUT and specify OUTPUT  
  source_data <- in_params[[1]]   ###Dataset to run inputs against  
  
  ***Recruitment Input  
  recruitsAge <- as.integer(in_params[[2]])  
  employmentStatus <- as.integer(in_params[[3]])  
  terrorGroup <- in_params[[4]]  
  recruitCountry <- in_params[[5]]  
  recruitmentCity <- in_params[[6]]  
  recruitmentTimespan <- as.integer(in_params[[7]])  
  
  *****Planning Input  
  planMeetingsCountry <- in_params[[8]]  
  planMeetingsCity <- in_params[[9]]  
  planTrainingCountry <- in_params[[10]]  
  planTrainingCity <- in_params[[11]]  
  planningTimespan <- as.integer(in_params[[12]])  
  
  *****Preparatory Input  
  prepweapons <- as.integer(in_params[[13]])  
  prepFunds <- as.integer(in_params[[14]])  
  prepIllegalImmigration <- as.integer(in_params[[15]])  
  prepsurveillance <- as.integer(in_params[[16]])  
  prepTimespan <- as.integer(in_params[[17]])  
  
  ## Output##STATISTICAL RESULTS eg.ANOVA etc  
  # kMeansTable <- out_params[[1]]  
  # predictionTable <- out_params[[1]] ##KMEANS  
  predictionByDiscriminantAnalysis <- out_params[[1]]  
  DAPredictingUserCase <- out_params[[2]]  
  KMEANSpredictinguserCase <- out_params[[3]]  
  predictionByKMEANS <- out_params[[4]]  
  # estimateAreaOfFocus <- out_params[[3]] ## if time allows
```

Figure AD- 1 : The tool\_exec ( ) function and some of the expected input and output parameters

```

##RECRUITMENT DETAILS LOGIC

#1. AGE
if(!(recruitsAge <= 0)) {
  if(recruitsAge > 15 && recruitsAge <= 25)
  {
    ageRiskval <- "H"
  }else if(recruitsAge > 25)
  {
    ageRiskval <- "M"
  }else {
    ageRiskval <- "L"
  }
}
ageRiskval

#2. EMPLOYMENT STATUS
if(employmentStatus == TRUE)
{
  empRiskval <- "L"
}else{
  empRiskval <- "H"
}
empRiskval

#3. COUNTRY AND TOWN
countryUnique <- unique(data_df$Country)
cityUnique <- unique(data_df$Town)
# countryUnique <- unique(x)
# cityUnique <- unique(w)
testFlag <- TRUE
for (country in countryUnique)
{
  if(country == recruitCountry)
  {
    for(town in cityUnique)
    {
      if (town == recruitTown)
      {
        recruitLocationRiskval <- "H"
        testFlag <- FALSE
        break
      }else
      {
        if(testFlag == FALSE){

```

*Figure AD- 2: Calculating Recruitment activity risk value logic*

```

##PLANNING DETAILS LOGIC
#MEETINGS
#1. COUNTRY AND TOWN
countryUnique <- unique(data_df$Country)
cityUnique <- unique(data_df$Town)
# countryUnique <- unique(x)
# cityUnique <- unique(w)
testFlag <- TRUE
for (country in countryUnique)
{
  if(country == planMeetingsCountry)
  {
    for(town in cityUnique)
    {
      if (town == planMeetingsCity)
      {
        meetingLocationRiskVal <- "H"
        testFlag <- FALSE
        break
      }else
      {
        if(testFlag == FALSE){
          # print("Do nothing")
        }
        else{
          meetingLocationRiskVal <- "M"
          testFlag = FALSE
        }
        break
      }
    }
  }
}
else
{
  if(testFlag == FALSE){
    #print("Do nothing")
  }
  else
  {
    meetingLocationRiskVal <- "L"
  }
}

```

*Figure AD- 3 : Calculating Planning activity risk value logic*

```

##3. PREPARATORY DETAILS LOGIC
#1. Weapons movement and purchases
if(prepareWeapons == TRUE)
{
  weaponsRiskVal <- "H"
}else{
  weaponsRiskVal <- "M"
}
weaponsRiskVal

#2 Funds
if(prepareFunds == TRUE)
{
  fundsRiskVal <- "H"
}else {
  fundsRiskVal <- "L"
}
fundsRiskVal

#3 Illegal Immigration crime
if (prepareIllegalImmigration == TRUE)
{
  immigrationRiskVal <- "H"
}else {
  immigrationRiskVal <- "M"
}
immigrationRiskVal
#4 Surveillance

if(prepareSurveillance == TRUE)
{
  surveillanceRiskVal <- "H"
}

```

Figure AD- 4 : Calculating Preparatory risk value logic

```

if(prepareCountH >= prepareCountM)
{
  if(prepareCountH >= prepareCountL)
  {
    finalPrepRiskLevel <- "H"
    finalPrepRiskValue <- runif(1, 6.0, 10.0)
  } else
  {
    finalPrepRiskLevel <- "L"
    finalPrepRiskValue <- runif(1, 0.0, 4.0)
  }
} else if (prepareCountM >= prepareCountL)
{
  finalPrepRiskLevel <- "M"
  finalPrepRiskValue <- runif(1, 4.1, 5.9)
} else
{
  finalPrepRiskLevel <- "L"
  finalPrepRiskValue <- runif(1, 0.0, 4.0)
}

```

Figure AD- 5 : Logic for aggregating risk values.



```

###KMeans Logic
#* Extract only fields needed for clustering
testData <- data_df[, 2:5]
dataToCluster <- data_df[, 2:4]
clusteredData <- kmeans(dataToCluster, 2)

clusteredData
clusteredData$cluster
clusteredData$centers

clusteredData$size

table(clusteredData$cluster, testData$Terror_Event)    #OUTPUT 1

```

Figure AD- 6 : K-means Clustering in the algorithm

```

if (!is.null(predictionByKMEANS) && predictionByKMEANS != 'NA') {
  arc.write(predictionByKMEANS, td)
}

```

Figure AD- 7 : logic for outputting results for KMEANS clustering from R script to ArcGIS

```

#* Discriminant Analysis for prediction/Classification
#* dim(testData)

train = sample(1:nrow(testData), nrow(testData)/3 * 2)
testData_train = testData[train, ]
testData_test = testData[-train,]

# fit = lda(Terror_Event ~ Recruitment + Planning + Preparatory, data = testData_train)
fit = qda(Terror_Event ~ Recruitment + Planning + Preparatory, data = testData_train)
pred = predict(fit, testData_test)

pred_class = pred$class

table(pred_class, testData_test$Terror_Event)    ### POTENTIAL OUTPUT
t <- table(pred_class, testData_test$Terror_Event)
t
summary(t)
t <- as.data.frame(t)

if (!is.null(predictionByDiscriminantAnalysis) && predictionByDiscriminantAnalysis != 'NA') {
  arc.write(predictionByDiscriminantAnalysis, t)
}

mean(pred_class == testData_test$Terror_Event)

###Testing
####TESTING MODEL WITH NEW UNSEEN DATASET -- User Input
## discriminant analysis
T1 <- data.frame("Recruitment" = finalRiskValue, "Planning"= finalPlanRiskValue, "Preparatory"= finalPrepRiskValue, "Terror_Event" = " ")
pred = predict(fit, T1)

pred_class = pred$class

table(pred_class, T1$Terror_Event)    ### POTENTIAL OUTPUT
mean(pred_class == T1$Terror_Event)    ### POTENTIAL OUTPUT

```

Figure AD- 8 : Discriminant analysis in the algorithm

```

##Sending output 2
t1 <- table(pred_class, T1$Terror_Event)
t1
summary(t1)

t1 <- as.data.frame(t1)

if (!is.null(DAPredictingUsercase) && DAPredictingUsercase != 'NA') {
  arc.write(DAPredictingUsercase, t1)
}

```

*Figure AD- 9 : Writing the results from Discriminant analysis algorithm to ArcGIS as a table*

```

#1. AGE
if (!(recruitsAge <= 0)) {
  if(recruitsAge > 15 && recruitsAge <= 25)
  {
    ageRiskval <- "H"
  }else if(recruitsAge > 25)
  {
    ageRiskval <- "M"
  }else {
    ageRiskval <- "L"
  }
}

```

*Figure AD- 10 : Exception handling for negative values of age*