MASTER’S THESIS

TERRORISM AND ITS IMPACT ON HOSPITALITY INDUSTRY PATTERNS: A DATA MINING APPROACH

Ljubljana, May 2018
PEDRO RAMOS
AUTHORSHIP STATEMENT

The undersigned Pedro Ramos, a student at the University of Ljubljana, Faculty of Economics, (hereinafter: FELU), author of this written final work of studies with the title Terrorism and Its Impact on Hospitality Industry Patterns: A Data Mining Approach, prepared under supervision of Jurij Jaklič, PhD and co-supervision of Roberto Henriques, PhD,

DECLARE

1. this written final work of studies to be based on the results of my own research;
2. the printed form of this written final work of studies to be identical to its electronic form;
3. the text of this written final work of studies to be language-edited and technically in adherence with the FELU’s Technical Guidelines for Written Works, which means that I cited and / or quoted works and opinions of other authors in this written final work of studies in accordance with the FELU’s Technical Guidelines for Written Works;
4. to be aware of the fact that plagiarism (in written or graphical form) is a criminal offence and can be prosecuted in accordance with the Criminal Code of the Republic of Slovenia;
5. to be aware of the consequences a proven plagiarism charge based on the this written final work could have for my status at the FELU in accordance with the relevant FELU Rules;
6. to have obtained all the necessary permits to use the data and works of other authors which are (in written or graphical form) referred to in this written final work of studies and to have clearly marked them;
7. to have acted in accordance with ethical principles during the preparation of this written final work of studies and to have, where necessary, obtained permission of the Ethics Committee;
8. my consent to use the electronic form of this written final work of studies for the detection of content similarity with other written works, using similarity detection software that is connected with the FELU Study Information System;
9. to transfer to the University of Ljubljana free of charge, non-exclusively, geographically and time-wise unlimited the right of saving this written final work of studies in the electronic form, the right of its reproduction, as well as the right of making this written final work of studies available to the public on the World Wide Web via the Repository of the University of Ljubljana;
10. my consent to publication of my personal data that are included in this written final work of studies and in this declaration, when this written final work of studies is published.

Ljubljana, May 15th, 2018

Author’s signature: _________________________
TABLE OF CONTENTS

INTRODUCTION ......................................................................................................................... 1

1 THEORETICAL FRAMEWORK ............................................................................................... 4
   1.1 Bibliographic Analysis ........................................................................................................ 4
   1.2 Terrorism Framework .......................................................................................................... 6
      1.2.1 Defining Terrorism ...................................................................................................... 6
      1.2.2 Brief History of Modern Terrorism ............................................................................. 7
      1.2.3 Terrorism in the Twenty-First Century ....................................................................... 8
      1.2.4 Risk Perception and Terrorist Threat Awareness ......................................................... 9
      1.2.5 The Economic Impact of Terrorism ........................................................................... 10
   1.3 Tourism Economics Framework ....................................................................................... 11
      1.3.1 Defining Tourism ....................................................................................................... 11
      1.3.2 The Importance of a Statistical Definition of Tourist ................................................. 12
   1.4 Data Mining Framework ................................................................................................... 14
      1.4.1 Definition of Cluster Analysis .................................................................................... 15
      1.4.2 Proposing a Framework for Cluster Analysis .............................................................. 15
      1.4.3 Cluster Methods and Algorithms .............................................................................. 16
      1.4.4 Cluster Validity and Evaluation Criteria ...................................................................... 20
      1.4.5 Longitudinal Data Clustering Methods .................................................................... 25
      1.4.6 Text Clustering Methods and Algorithms .................................................................... 27

2 DATA COLLECTION AND PREPARATION .......................................................................... 29
   2.1 Terrorism Data: Global Terrorism Database ...................................................................... 29
      2.1.1 Origin and Purpose .................................................................................................... 29
      2.1.2 Data Collection: Methodology and Limitations .......................................................... 30
      2.1.3 Metadata and Data Cleansing Codebook .................................................................. 31
   2.2 Tourism Data – World Tourism Indicators ....................................................................... 36
      2.2.1 Origin and Purpose ................................................................................................... 37
      2.2.2 Data collection: Methodology and limitations ............................................................ 38
      2.2.3 Metadata and Data Cleansing Codebook .................................................................. 40
   2.3 Unifying Data Sources ...................................................................................................... 41

3 EXPLORATORY ANALYSIS ................................................................................................. 46
   3.1 Feature Selection ............................................................................................................... 47
   3.2 Cluster Analysis ............................................................................................................... 50
      3.2.1 Country-aggregated Cluster Analysis ....................................................................... 50
      3.2.2 “Pooled” Dataset Cluster Analysis .......................................................................... 56
      3.3.3 Longitudinal Dataset Cluster Analysis ....................................................................... 58

4 RESULTS DISCUSSION ....................................................................................................... 60

CONCLUSION .......................................................................................................................... 67

REFERENCE LIST ................................................................................................................. 69

APPENDIXES
LIST OF TABLES

Table 1. GTD Collection Phases ................................................................. 29
Table 2. Political Changes in Europe .......................................................... 33
Table 3. Terrorism Dataset: Attributes and Metadata Summary ...................... 36
Table 4. World Development Indicators ....................................................... 37
Table 5. World Development Indicators: Data Cleansing ................................. 41
Table 6. Merged Dataset: Countries/Region Breakdown .................................. 42
Table 7. Terrorism Geographical Distribution: Region Aggregates (1970-2014) ..... 43
Table 8. Terrorism Geographic Distribution: Region Aggregates (1995-2014) ....... 45
Table 9. Cross-correlation Analysis ............................................................ 49
Table 10. Optimal Partition Schemes per HAC Method and Distance Measure Employed 53
Table 11. Cluster Country List: PAMEU4 ................................................... 62
Table 12. Cluster Country List: PAMMN6 .................................................. 63
Table 13. Cluster Country List: LDKM4 ..................................................... 65

LIST OF FIGURES

Figure 1. Initial Bibliographic Network ...................................................... 5
Figure 2. Refined Bibliographic Network .................................................... 5
Figure 3. The Traveler and Visitor Concepts ............................................... 13
Figure 4. Typical Cluster Analysis Workflow .............................................. 15
Figure 5. Cluster Similarity Notions in HAC .............................................. 19
Figure 6. Terrorism Geographic Distribution: Included Attacks ....................... 44
Figure 7. Terrorism Geographic Distribution: Excluded Attacks ...................... 44
Figure 8. The Impact of the Dataset Merging Process: Countries per Region ....... 46
Figure 9. Aggregated Dataset: HAC Solutions (Compactness versus Separation) 54
Figure 10. Aggregated Dataset: Partitioning Solutions (Compactness versus Separation) 56
Figure 11. Evaluation of PAM for Pooled Dataset: Compactness, Separation and Classification 57
Figure 12. Longitudinal Data Clustering: Validity Criterion (Calinski-Harabatz) .... 59
Figure 13. Cluster Solution Overview: PAMEU4 ....................................... 61
Figure 14. Cluster Solution Overview: PAMMN6 ...................................... 63
Figure 15. Cluster Solution Overview: LDKM4 ......................................... 64
Figure 16. Final Solution: Geographic Overview ......................................... 65
INTRODUCTION

Terrorism is commonly described as a strategy embraced by an extensive variety of ideologists and for a similarly extensive variety of political purposes (Cronin, 2002; Drake, 1998; Hoffman, 2006; Reich, 1998; Rubin, 1989). It is not indispensably connected to any ideology, belief or religion even though these can be potential reasons to trigger terrorist attacks. The choice for this strategy is empirically related to the potential ability of the strategy to successfully attain the perpetrator’s objective. Eradicating terrorism is far from being a simple task. Before eliminating terrorism, is crucial to de-legitimize terrorism at a global level. For such de-legitimization process to occur, it is fundamental to disprove the ‘false sociologies’ advanced as milestones for backing-off terrorism, exhibiting their foundation on fallacious or deceptive reasoning (Schmid, 2013). The growing pace and reach of technology, particularly social media, brought terrorism to the spotlight as internet and social networks became vehicles for news spreading and raising awareness for such events, playing a crucial role in both educating but also deceiving publics: mass media was at first a useful communication tool to spread alerts and inform endangered, threatened communities. On the other hand, and through means of misleading information, spreading sentiments like fear and panic, that might evolve into a violence snowball effect. The success of terroristic activities, from the perpetrators’ worldview, may not be measured by the same criteria that are held by steady societies and state systems. It is important to notice that terrorism is not always completely aimed to a specific target, permanently impacting several publics at once, even individuals who acknowledge or spread the underlying ideologies leading such attacks. Such victims might be accidental or collateral damages from the terrorist practices, but usually they are affected as part of high potential plans carried out by the attack perpetrators. Simply put, and intuitively, the more severe are the damages (casualties, injuries, property damage, etc.) accounted for an attack, the higher its social impact and thus its success from the perpetrators perspective. Although, this hypothesis is not always valid as some authors have been investigating, terrorist attacks of different magnitude and heinousness are considered to impact distinct publics in different manners. It is a fact that violence-related events affect societies and countries development and growth in multiple fields – economies are impacted by wars and continued crime scenes, cultures are oppressed, demographic pyramids might get skewed and several other areas are vulnerable to such events. Even presenting quite insignificant number of casualties and damages estimate when compared to military warfare or other (natural) catastrophes that occurred in the past, terrorist attacks are still a class of events that impact several industries, markets and economies. One of the economic areas that has been suffering the most from terrorism impact during the last decades is the tourism (and the whole hospitality industry). As the awareness for the terrorist threat is becoming more and more present on the quotidian life from the peoples of many countries around the globe, travelers and tourists develop their risk perception mindset by processing information disseminated by official (e.g., travel alerts and warnings from governmental institutions such as the U.S. Department of State) and non-official sources (e.g., generalist media or social media). The growing perception of the terrorism threat mostly due to the magnitude of the attacks – in terms of casualties and destruction – but also to the
frequency of the occurrences and to the press covering on the aftermath of terrorist attacks, given the
reach that mass media has been acquiring in the last few decades. These reasons are just some of the
factors that prompt changes in tourists’ travel intention, often affecting their travel agenda
convenience and, most likely, causing financial harm. Considering travelers as being customers in
the context the hospitality industry, that same prejudice spreads throughout the several chained
businesses, producing impacts of different levels on a bullwhip effect that can generate severe
tourism crises. Examples are usually hotels and resorts, airline companies, travel agencies and
general tourist attractions.

Abadie and Gardeazabal (2003), Blomberg, Hess and Orphanides (2004), Frey, Luechinger and
Stutzer (2007) and a variety of other authors (with further mention in this written work) employ
different data analysis techniques into studying the macroeconomic impacts of terrorism. Also a
variety of authors engage on the impact of terrorism in financial markets and how terrorism impacts
focused the scope of their studies into the reaction of stock markets to different periods of terrorist
activity. Carter and Simkins (2004) and Drakos (2004), specifically study the effects of terrorism
over the price of stocks from hospitality-related companies, in particular airlines as those are clearly
one of the first business types to suffer immediate – but also long term – effects of terrorism and
similar conflicts. However those studies are connected to the topic hereby addressed, there was very
little research done in terms of applying data mining techniques – and especially cluster analysis – to
provide better understanding of the relationship between terrorism and tourism in different
countries. That said, this study is somehow a pioneer of its kind, mostly when considering the vast
work done in terms of data pre-processing and visualization.

The purpose of this thesis is to combine several descriptive data mining techniques to identify
similarities between a set of countries according to the nature of their hospitality industry and the
unavoidable occurrence of terrorist activity over time. The ultimate goal of this written work is to
collect, prepare and make an exploratory work of two-fold kinds of data: the previous pertains to
terrorist events; the second belongs to the specific hospitality industries from the countries under the
scope of the study, describing facts like tourist (and travelers) arrivals for the different existing
segments. The second goal of this research work is to identify which countries are of high relevance
for the analysis by understanding existing patterns of terrorist activity, both regarding geographies
and nature of the event. Descriptive data mining techniques such as cluster analysis are used to
investigate and group countries presenting similarities both regarding their hospitality industries but
also the existing terroristic activity registered in their territory. The consistent combination of such
databases generates well-structured sets of data that can be employed in further studies as sample for
testing empirical models for studying the impact of terrorism over tourism and economics in
different country groups. In addition, the methodologies employed in this research allow for easy
replicability when feeding the algorithms with more structured and complete initial datasets. One
clear example of a further application of this methodology is to explore how terrorist attacks in
different cities within a country affect the inflow of tourist for that same country.
Before defining the methodology to be carried out, it is important to understand what kinds of data are available and their properties regarding the quality for achieving the purpose of this research work. Therefore, the data to be used is divided into two groups: the terrorism event data and the hospitality industry-related data. The former is extracted from the Global Terrorism Database (hereinafter: GTD), a comprehensive database of more than hundred thousand terrorism events that occurred between 1970 and 2014. This database comprises qualitative and quantitative data to describe each attack from where it is important to highlight certain attributes as described in Section 2. The second kind of data, about the hospitality industry, it is a compilation of indicators from arrivals and revenues from tourism from a large set of countries at the global level. Such dataset results from a laborious data examination and manipulation work to overcome certain constraints and inconsistencies from the data sources at the collection stage. The choice of variables at use is based both on the sources available but also on the extension of the previous researches supporting this approach. Within the critical literature, several ways of estimating the economic impact of a certain phenomenon namely terrorism were identified. Previous research works focused on the financial and stock markets response to terrorism, both for general and specific industries, such the hospitality and tourism. The present research follows a different scope by focusing mostly on detailed data from terrorist activity, exposing relationships with tourist arrivals data, collected for the respective periods as the existing data from GTD. The methodology selected to carry out this research project consists into the application of the previously stated data mining techniques, focusing over descriptive methods to perform an exploratory analysis that aims to enrich the knowledge from the data that is combined and prepared in previous stages.

The remainder of this written work has five primary sections. Section 1 (Theoretical Framework) is dedicated to present the relevant literature identified during the bibliographic research phase, enabling the establishment of a theoretical background to support discussions considered to be pertinent along with the scope of the research problem(s) here addressed. Additionally, and before introducing the theoretical framework itself - where definitions and proven empirical results are presented – this section embeds a preliminary subset designed to illustrate the methodology carried out during the scientific literature analysis stage. Section 2 (Data Collection and Preparation) describes the data used as input for the statistical analysis carried out while approaching the research problem. Here all the collected data are described regarding its characteristics regarding the collection process and all inherent assumptions that are crucial to anticipate prior to any further analytical steps, such as the data preparation stage. Section 3 (Exploratory Analysis) embeds the application of descriptive data mining techniques, such as cluster analysis, to provide a clearer understanding of the data at use, enabling a simplifying description of the dataset at a lower complexity level. Section 4 (Results Discussion) presents and compares the solutions from various clustering analysis approaches employed, providing meaningful interpretation of such solutions in the context of describing how tourism and terrorism are connected for the geographical and temporal timeframes in analysis.
1 THEORETICAL FRAMEWORK

The objective of this initial section is to provide the reader with relevant concepts and insights that are vital regarding the purpose of the thesis. Here the main definitions are presented and examined in accordance with prior research works carried out within the various fields covered in this thesis.

1.1 Bibliographic Analysis

To brace for the study of a given socio-economic phenomenon, it is crucial to examine prior scientific research works to have a clear understanding of the documented knowledge available for use in the purposed exploration. In the case of this written work, when studying terrorism patterns and their impacts in the hospitality industry, there is an ample variety of literature that addresses this topic in broader terms. It is important to notice that, although most of the sources addressed in this master thesis are generally concerned to terrorist events, some of them points into different directions, presenting different purposes, goals, methodologies and hence, results. This subsection is meant for providing an overview of the bibliographic research process and the methods selected for carrying it out. This research work relies on a set of scientific articles and textbooks concerned to the study of terrorism as a global phenomenon, the socioeconomic study of terrorism, tourist crisis management (those of terrorist related nature), etc. In the other hand, there is the need to browse through sources of more technical knowledge to identify appropriate research methodologies: those methods are essentially of statistical nature, with special emphasis on exploratory and descriptive methods, especially data mining techniques such as several methods for performing cluster analysis. If on one hand combining all those terms keywords for the purpose of literature research produces a scarce amount of relevant results (perhaps due to the novelty of the research question), on the other hand when searching each keyword separately, the amount of materials found is too broad for a single researcher to analyze through human inspection. The importance of understanding the quality and reliability of such resources dictated the implementation of a social network analysis of nearly 1,500 articles. Such articles relate to each other through means of citation of authors work that motivated each specific article in first place. The following analysis is obtained by first collect a list of potentially relevant articles together with the respective citations mapped out in Web of Science platform (see http://www.webofscience.com). Citation data is pre-processed by using HAMMER, a web-based server for automating the network analysis interface for literature studies scripts, providing the user with the tables that represent the nodes and the ones that represent the edges, also a set of relevant summary statistics regarding the results found; as a final step, the aforementioned tables are used to design the graph network (with resource to Gephi), that is then filtered in terms of in-degree metric. It is expectable that the most reliable articles that can potentially add superior quality and relevance to a new research project are those being cited more often among the scientific community. In the context of a directed social network, where nodes represent articles, and unidirectional edges represent citations, articles cited more often are said to have a higher in-degree, a centrality measure that counts the number of headends adjacent to a node (Biswal, 2013).
The strategy then used is to set up a certain bottom in-degree threshold, through the application of filters, to significantly reduce the dimension of the network capturing only articles that meet that criteria (at the time). In the particular case of this network, the initial 1,495 extracted articles pointed into 50,540 nodes, from which 64,259 connections arose (see Figure 1) due to the fact that each article cites a specific variety of sources, resulting in an exponential growth of the citations (edges) universe; when a citation is not common to other articles it originates a new node as well. To reduce the dimension of the resources universe, the minimum threshold for the in-degree was set up to 20 simplifying the network to only 101 nodes and 83 edges.
Also, to understand the relationships established within the simplified network, it is possible to cluster the nodes into different modularity classes, identifying the community structure (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). By iteratively increasing the resolution limit of the modularity computed using Gephi, the lowest number of communities (or clusters) was of 47 different modularity classes. This complexity reduction (depicted above in Figure 2) enables a more efficient inspection of the bibliographic resources. Note that additional sources are then added through ad-hoc research of more specific topics upon need during further research phases to support assumptions that fall out of the scope of the bibliographic network analysis initially performed.

1.2 Terrorism Framework

This section presents the various resources that result from previous research works, dividing them into logically constructed and connected matters that are used as a basic support for the understanding of the terrorism phenomenon and its impacts both at a global and industry-level.

1.2.1 Defining Terrorism

Despite the nonexistence of a universal consensus regarding the definition of terrorism, governments have been striving to accurately define these phenomena to enable the authorities responsible for law enforcement and homeland security to have a similar perception of the terrorist acts of different natures. It is crucial that those who are empowered to prevent and minimize the impact of terrorist events can agree in a definition that fully covers the main types of terrorism practices to develop and plan counter-terrorist strategies in effective and efficient manners, using similar criteria for acting upon them. From a simplistic point of view, terrorism is just a form of coercion using violence that goes far behind the achievement of a direct and immediate goal like in another kind of crime. “Without violence or its threat, terrorists cannot make a political decision maker respond to their demands. Moreover, in the absence of a political/social motive, a violent act is a crime rather than an act of terrorism.” (Enders & Sandler, 2012). In the absence of a consensual definition, there is the need to adopt – or combine – several existing definitions to best describe the terrorism phenomena. Although studies have found that there are more than two hundred definitions of terrorism, Schmid and Jongman, (2005) define terrorism as “an anxiety-inspiring method of repeated violent action, employed by (semi-)clandestine individual, group, or state actors, for idiosyncratic, criminal, or political reasons, whereby — in contrast to assassination — the direct targets of violence are not the main targets. The immediate human victims of violence are generally chosen randomly (targets of opportunity) or selectively (representative or symbolic targets) from a target population, and serve as message generators”. Nonetheless, the previous definition is supported on a wide theoretical background, resulting from the examination of over a hundred narrower definitions, it may be convenient for this written work to adopt the definition considered during the data collection process instead. That being said, terrorism is hereinafter defined as “the threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal.
through fear, coercion, or intimidation."
(National Consortium for the Study of Terrorism and Responses to Terrorism [START], 2016)

1.2.2 Brief History of Modern Terrorism

Terrorism has been put into practice since more than two thousand years already, evolving over times and gaining different proportions and importance: one of the earliest recorded events was the Zealot struggle against the Roman occupation of Palestine during 66-73 A.D. (Laqueur, 1978). In spite of the elderliness of terrorism as guerrilla warfare practice, the use of the term “terrorism” arose from French terrorism (from Latin terror) to describe the State of Terror (1793-1794), as an onset of the French Revolution, period when the post-radical government of France slaughtered the French aristocracy and associates (Primoratz, 2004). Some examples of the most noteworthy terrorist movements happening during modern and contemporary eras, are as follows, according to White (2001), Hoffman (2006) and Enders and Sandler (2012):

- the Russian anarchist faction Narodnaya Volya (hereinafter: People’s Will), active between 1878 and 1881, perpetrating the assassination of Czar Alexander II and becoming one of the most remarkable influences in the modern-day terrorism;
- also in Russia, Lenin and Trotsky led the Bolsheviks revolution of October 1917, using terrorist tactics such as bombings and assassinations to throw apart the Menshevik government, targeting their officials and its middle-class electorate;
- the Irish Republican Army (hereinafter: IRA), commanded by Michael Collins refined the terrorist tactics of People’s Will during Ireland’s 1919–1921 campaign to struggle for the independence from British control; from 1930, until the end of the century, the battle for Irish unification was coordinated by the IRA and the Provisional IRA (hereinafter: PIRA);
- in the Middle East, during the period of 1947–1948, and in order to attain the independence of Israel, two terrorist groups – Irgun Zvai Leumi and the Stern Gang – imported IRA bombing and assassination tactics targeting British occupants to emphasize the cost of denial Jewish demands for statehood, escalating their campaigns into urban guerrilla warfare and proving the effectiveness of urban terrorism;
- the Algerian rebellion against French regime between 1954 and 1962 was helmed by the Front de Libération Nationale (FLN), borrowing its methods from those of the Jewish terrorists in Palestine;
- similarly, from 1956 to 1959, Cypriot radicals followed the same methods, also inspired by the Jewish insurgents’ tactics successfully applied during the 1940s to struggle against the British dominance;
- the slaughter for the Irish unification expanded significantly in 1969, when the British armed force was conveyed to Northern Ireland, due to the developing PIRA’s urban guerrilla fighting strategies, making the strikes against police and British troops seem arbitrary but persistent in Belfast, Londonderry, and other urban focuses in Northern Ireland;
• the *Tupamaros*, active terrorist faction in Latin America (mostly in Uruguay) during 1968-1972 refined their urban guerilla methods by including kidnappings and bank robberies to fund their activities;

• founded in 1959 with the purpose of fighting for a sovereign Basque state ruling the autonomous Spanish region of Basque Country, the separatist group *Euskadi Ta Askatasuna* (Basque for “Basque Country and Freedom”, hereinafter: ETA) just adopted sizable terrorist attacks during the mid-1970s, having its bloodiest period during 1978–1980, totalizing 235 victims for that same period. Along with assassinations, and at a similarly growing pace, ETA perpetrated various fund-raising crimes such as kidnapping-for-ransom, coercion and, occasionally, assaults targeting Basque wealthy entrepreneurs, resulting in an exodus of those from the region, until the year of 2000, date when the group activity was officially ceased (Abadie & Gardeazabal, 2003);

• Lastly, a remarkable ongoing case, which emerged after the 1967 Arab-Israeli War is the resistance for a Palestinian state against the Israeli regime, undertaken by the Palestinian Liberation Organization (hereinafter: PLO) and respective dissident factions. These studied other aforementioned terrorist group terror-spread techniques emphasizing the internationalization of their attacks to attain superior recognition for their cause against Israel, embracing a global struggle that shaped terrorism as it happens nowadays.

The assessment of the abovementioned timeline of remarkable continued terrorist campaigns, enabled Enders and Sandler (2012) to conclude empirically:

1. terrorists have been learning how to execute and enhance their techniques to achieve their immediate and – every so often, due to persistence – long-run goals, borrowing the knowledge of allied – or even rival – factions’ operations in seeking for more effective results;

2. terrorists manage to orchestrate their operations within metropolitan areas pursuing the accessible media covering, increasing the exposure of the attacks throughout the communication channels, and hence the wider spread of their message;

3. the steady increasing globalization of the terrorist activity from certain groups has been a successful strategy for capturing the world attention for their causes, spreading global fear and threat among potential target publics;

4. several are the cases of fortunate separatist movements that, due to the inaptitude of authorities to combat their campaigns, succeed to sustain a constituency, and thus becoming more likely to achieve their long-run goals.

1.2.3 Terrorism in the Twenty-First Century

A new era was born for terrorism with the events on 11 September, 2001 (hereinafter: 9/11): the sequence of four airplane hijackings culminated in a worldwide unprecedented death toll in the context of transnational terrorism: roughly 3,000 lives were lost as a result of the attacks, exceeding the
analogous number of deaths during the period of 1988-2000 (Sandler, 2003). The 9/11 attacks
confirmed the conclusions above regarding the evolution of terrorist campaigns sophistication,
particularly their search for media exposure and the internationalization of the tragedy. In fact, that
was a day that changed the world, both for terrorists but for their targets as well. Some of the
changes induced in the aftermath of 9/11 are as follows:

- since 9/11, innumerable businesses are not able to afford a terrorism coverage given the change
  of paradigm for insurance companies: according to Kunreuther, Michel-Kerjan and Porter
  (2003) and Kunreuther and Erwann (2004), roughly half of the associated losses ascending to
  $80 billion were covered by insurance institutions, which rapidly announced their unwillingness
  to provide coverage for terrorist-induced damages;
- suicidal terrorism was proven able to achieve maximal casualties, outstanding financial losses,
  and superior outspread panic by recurring to relatively low-budget plans and involving supposed
  inoffensive technologies;
- countries rapidly redesigned their homeland security enforcement strategies by reallocating
  resources from other areas of expenditure over additional measures to attain safety and combat
  terrorism in their own soil (Enders & Sandler, 2012). One of the most illustrative examples is the
  US Department of Homeland Security which benefited from successive budget reallocation over
  the upcoming years, dedicating roughly 60% of their expenditure to fight terrorism on US
  territory.

1.2.4 Risk Perception and Terrorist Threat Awareness

The perception of risk arising from the terrorist threat plays a fundamental role in the behavior of
people as individuals and/or as part of groups within society. Several studies have been made to
identify relationships between risk awareness and reaction to such risks, depending on several
characteristics of individuals and the way risk situations are perceived. Roehl and Fesenmaier
(1992) used cluster analysis to classify tourists into three - neutral, functional risk and place risk
groups –according to the way they react to certain risk scenarios when traveling. Tourists classified
as being neutrally risk takers have shown a surprisingly higher willingness to engage in riskier
activities and destinations, however at the cost of more in-depth planning methods. Reisinger and
Mavondo (2005) through means of a path analysis “intentions to travel internationally were
determined by travel anxiety levels and level of perceived safety”, concluding that “terrorism and
sociocultural risk emerged as the most significant predictors of travel anxiety”. Different individuals
might perceive terrorist acts differently, which give perpetrators strong advantages towards the
achievement of their goals, as attacks at random might spread more severe awareness of the terrorist
imminent risk. As described by Enders and Sandler (2012) “terrorists broaden their audience beyond
their immediate victims by making their actions appear to be random, so that everyone feels anxiety.
In contrast to a drive-by shooting on a city street, terrorist acts are not random but well-planned and
often well-executed attacks where the terrorists account for risks and associated costs as well as possible gains”.

According to Chang and Zeng (2011) the hospitality industry is notably responsive to the occurrence of terrorist attacks and the risk perception from stakeholders in particular areas such as tourism. The major reasons the authors point out for justifying such fragility of this sector are three:

- for some people travelling and leisure are both superfluous expenditures, yet more when aligned with augmented fear and risk arising from terrorist activity – considering this proposition is expectable that risk adverse travelers retreat their travelling intentions or choose to travel to substitute destinations instead;
- given the dimension and spread of the hospitality industry, when addressing specific developing countries of high dependence on the tourism sector, any terrorist attack might understandably impact the broad economy since many industries profit on economies of scale created by tourism and travelling sectors;
- finally, since hotels have no ability to store capacity nor inventory, the incidence of terrorist episodes easily leads immediate effects that might be hard to flatten over time.

1.2.5 The Economic Impact of Terrorism

There are several previous studies that address the long-term macroeconomic effects of terrorism, most of them referring to cases of countries (or regions) under enduring conflicts. Abadie and Gardeazabal (2003) conclude that the Basque GDP substantially underperforms that of a similar synthetic region (in Spain) without continued terrorist attacks. Blomberg et al. (2004) note that terrorism has a negative impact on growth, especially in developing countries. Likewise, Frey et al. (2007) find that the net effects of terrorism impacts negatively the total utility and go far beyond losses of pure economic nature. Also a variety of authors engage on the impact of terrorism in financial markets and how terrorism impacts stock returns. Chen and Siems (2004) study the impact of fourteen terrorist and military attacks prior to 1915 on U.S. capital market, and the response of global capital market to later events such as the Iraq’s invasion of Kuwait, in 1990, and the 9/11 attacks to World Trade Center, concluding that in spite of the generality of stock market returns are negative, U.S. capital markets got more volatile in the later years recovering faster than other capital markets. This study corroborates the findings from Burch et al. (2003) showing that closed-end mutual discount funds aggravated seriously following the 9/11 attacks – indubitably as an echo of the negative sentiment-based response of investors – but consecutively recovered apace with the broader market. Additionally, Karolyi and Martel (2010) investigate the most impactful seventy-five events occurring worldwide during the prior decade, noticing that terrorism “attacks in countries that are wealthier and more democratic are associated with larger negative share price reactions, and […] human capital losses, such as kidnappings of company executives, are associated with larger negative stock price reactions than physical losses, such as bombings of facilities or buildings.”
The hospitality industry has often referred to terrorism as one of its biggest operating risks because a single isolated attack may affect interactions between customers and businesses, interrupting a whole previously scheduled logistic chain. Carter and Simkins (2004) scrutinize the 9/11 attacks finding extremely unusual and negative returns for airline stocks following the markets reopening on September 17th of that same year. Similarly, and according to Drakos (2004), it is also possible to verify this financial phenomenon given the abnormal increase of the volatility of the airline stocks in the post-9/11 era, predominantly due to increments in systematic risk. Looking through a slightly different scope, there is a variety of studies addressing the problematic of the terrorist attacks targeting tourists and tourism facilities. Richter and Waugh (1986) make an overview of a set of drivers that potentiate the apetite of terrorists on targeting tourists depending on the – direct and indirect – objectives terrorist organizations want to achieve which such actions. The problem hereby addressed is the fact that small work has been done regarding the impact of terrorist attacks over tourists from a descriptive data mining perspective.

1.3 Tourism Economics Framework

The present section aims to make use of literature resources to define milestone concepts regarding the tourist activity, enabling a clear understanding of the target object within the hospitality industry: tourists.

1.3.1 Defining Tourism

Like terrorism, the concept of tourism caresses of a unique unanimous definition. Proposals have been made by authors and institutions, changing and adapting over the years to different realities, often varying across geographies and from each tourism stakeholders. McIntosh and Goeldner (1986) refer the four perspectives to take in consideration prior to present a tourism definition:

- the tourist, an individual seeking “various psychic and physical experiences and satisfactions”;
- the hospitality industry, the group of “businesses providing tourism goods and services” in order to satisfy the tourist needs while generating wealth;
- the hosting government, presenting an important role in both regulating tourism as a business activity and tourism conflict management;
- the host community, resident people from tourist destinations playing both economic and cultural roles, respectively in terms of the labor force and source of cultural interaction between the local culture and the tourist, being a crucial public in order to attain the engagement of existing tourists and use their relative satisfaction to generate future tourism opportunities (e.g., through word-of-mouth, although this can also have negative effects on the tourist destination development).
Additionally, Vanhove (2005) discern between “conceptual and statistical (operational or technical) definitions” of tourism identified in the literature. The first group addresses the tourism phenomenon from a higher abstraction level, broadly defining tourism as:

- the “sum of the phenomena and relationships arising from the travel and stay of non-residents, in so far as they do not lead to permanent residence and are not connected with any earning activity.” (Hunziker & Krapf, 1942);
- a travel event consisting of the sum of five fundamental characteristics, according to Burkart and Medlik (1974): “tourism is an amalgam of phenomena and relationships rather than a single one […] these phenomena and relationships arise from a movement of people to, and a stay in, various destinations; there is a dynamic element (the journey) and a static element (the stay) […] the journey and stay are to and in destinations outside the normal place of residence and work, so that tourism gives rise to activities which are distinct from those of the resident and working populations of the places through which tourists travel and of their destinations […] the movement to the destinations is of a temporary, short-term character […] destinations are visited for purposes not connected to paid work – that is, not to take up employment”;
- the “processes, activities, and outcomes arising from the relationships and the interactions among tourists, tourism suppliers, host governments, host communities, and surrounding environments that are involved in the attracting and hosting of visitors.” (McIntosh & Goeldner, 1986);
- “one part of recreation which involves travel to a less familiar destination or community, for a short-term period, to satisfy a consumer need for one or a combination of activities.” (Gilbert, 1990).

1.3.2 The Importance of a Statistical Definition of Tourist

Although the previously stated definitions provide quite an extensive characterization of the tourism phenomenon, it is important to reinforce the “need for exact definitions of tourism and the tourist (as an individual)” highlighted by Mieckzowski (1990) to enable superior quantitative measurements for statistical and operational purposes. One of the most common difficulties arising from the lack of a precise definition of tourist (and tourism) is the case of business travelers, usually included together as the remainder of the tourists, disregard of the purpose of the trip. To fulfil such conceptual breaches Burkart and Medlik (1974) suggest that a statistical definition of tourism must specifically cover the various travel and visit categories that are (and are not) included, outlining the time requirements regarding the duration of the stay in precise units, and finally identifying exceptions and special cases (e.g., transitory air passengers). There is a variety of statistical definitions that fulfill such requirements, characterizing not only tourists but a visitor in general. As a result of the United Nations Conference on Travel and Tourism (Rome, 1963), a suitable definition was agreed, defining visitor as “any person visiting a country other than that in which he has usual place of residence, for any reason other than following an occupation remunerated from within the
country visited”. Additionally, the resulting definition defined boundaries for distinguishing between:

- tourists, defined as “temporary visitors staying at least 24 hours in the country visited and the purpose of whose journey can be classified under the headings of Economic characteristics of the tourism sector either leisure (recreation, holiday, health, study, religion, and sport) or business, family, mission, meeting”;
- and excursionists, defined as “temporary visitors staying less than 24 hours in the country visited (including air travelers on cruises)”.

Although this definition of visitor enables a more rigorous coverage of the different possible cases, there was a place for successive minor adjustments, as it is an example the replacement of the term “24 hours” length by “overnight” that, in spite of being a broader expression, allows the inclusion of overnight stays lasting for periods smaller than one complete day. Furthermore, and in order to create a superior limit for the possible stay of visitors, the International Network on Regional Economics, Mobility and Tourism (hereinafter: INRouTE) and the World Tourism Organization (hereinafter: UNWTO) reformulated the previous definition, considering a visitor as “a traveler taking a trip to a main destination outside his/her usual environment, for less than a year, for any main purpose (business, leisure or other personal purpose) other than to be employed by a resident entity in the country or place visited. A visitor (domestic, inbound or outbound) is classified as a tourist (or overnight visitor) if his/her trip includes an overnight stay, or as a same-day visitor (or excursionist) otherwise” (INRouTE & UNWTO, 2013). For a clearer understanding of the scope of this definition, Figure 3 breaks down the concept of (inbound) traveler and the underneath defined classes.

![Figure 3. The Traveler and Visitor Concepts](source: Adapted after Department of Economic and Social Affairs, United Nations, International Recommendations for Tourism Statistics, 2008.)
According to the previously mentioned set of guidelines suggested by UNWTO, tourism must be perceived as containing three distinct kinds, partitioned as follows:

1. domestic tourism, including all the activities of resident visitors within the country of reference, either as part of a domestic tourism trip or part of an outbound tourism trip’’;
2. inbound tourism, which encompasses the “activities of non-resident visitors within the country of reference on an inbound tourism trip’’;
3. outbound tourism, comprehending the “activities of a resident visitor outside the country of reference, either as part of an outbound tourism trip or as part of a domestic tourism trip’’.

For the case of the present study, and given the nature of the data and respective limitations (see 2.2.2), only the inbound tourism figures are considered, leaving domestic and outbound tourism as opportunities for further studies.

1.4 Data Mining Framework

Han, Kamber and Pei (2012) define the term data mining as a “synonym for knowledge discovery from data”, a set of steps to undertake while taking valuable insights from an initial set of data (that can be the combination of data from several different sources). This knowledge discovery is described as a process including 5 major steps, as follows:

1. data cleansing and integration: eliminating existent noise and inconsistencies and merging data if numerous sources are identified;
2. data selection and transformation: identifying and fetching data considered as relevant for the research problem from the database (or data warehouse, if employed), transforming data through means of summary, reduction and/or aggregative procedures (when applicable, depending on the aim and precision of the analysis);
3. data mining: employing comprehensive statistical and machine learning methods to derive patterns and similarities on data;
4. pattern evaluation: identifying patterns of interest based the research scope and aimed knowledge, comparing them against predefined theoretical measurements;
5. knowledge presentation: employing visual representation techniques for the knowledge derived from the previously stated stages to attain concise understanding from final users.

This section is meant to establish a broad framework to support the data mining related tasks addressed in the execution of the research work. The following subsections present specific knowledge from several fields of data mining and machine learning areas in application to clustering problems, supported by previous research works found upon bibliographic research.
1.4.1 Definition of Cluster Analysis

A clustering problem is broadly defined in the literature as:

- the “unsupervised process of grouping data patterns into clusters so that patterns within a cluster bear strong similarity to one another but are very dissimilar to patterns in other clusters.” (Xiong & Yeung, 2002);
- to “identify structure in an unlabeled data set by objectively organizing data into homogeneous groups where the within-group-object similarity is minimized and the between-group-object dissimilarity is maximized.” (Liao, 2005);
- an “unsupervised learning task aimed to partition a set of unlabeled data objects into homogeneous groups or clusters (...) in such way that objects in the same cluster are more similar to each other than objects in different clusters according to some defined criterion.” (Montero & Vilar, 2014).

1.4.2 Proposing a Framework for Cluster Analysis

Defining clusters is not an exact science, rather than the application of computational and statistical techniques of different nature that provide different solutions. There are no right or wrong solutions for a clustering problem: from all the potential solutions to be encountered some are more “meaningful” in order to resolve a specific research question (Norušis, 2010).

Figure 4. Typical Cluster Analysis Workflow

Cluster solutions can vary significantly among the constraints of the problem, depending upon the selection of clustering variables, the clustering procedure (i.e., the set of techniques involved) and the selected number of clusters. In fact, the selection of the variables to cluster is one of the main stages to attain a set of solutions of enhanced accuracy for the cluster problem solving. The above Figure 4 presents the main steps recommended by Halkidi, Batistakis and Vazirgiannis (2001) to perform while formalizing and resolving a cluster analysis problem. Following this methodology, it is crucial to have an extensive understanding of the data prior to employing any clustering task.

1.4.3 Cluster Methods and Algorithms

The input data can be of various types, for which different algorithms apply, each one presenting its particular strengths and weaknesses from the perspective of the required inputs, computational complexity and results performance. Han and Kamber (2001) categorize the existent static data clustering methods into five main classes: partitioning, hierarchical, density-based, grid-based and model-based methods. A brief overview of the previously stated methods follows next, presenting their main features, advantages and limitations.

1.4.3.1 Partitioning Methods

Considering a set of \( n \) records of data, a partitioning method aims to divide the dataset into \( k \) partitions (with \( k \leq n \)), each one representing a cluster containing at least one element. Although their straightforward implementation, Halkidi et al. (2001) denote partitioning clusters major shortcomings as they:

- are “applicable mainly to numerical datasets”, although implementations such as the \( k \)-Modes algorithm have been purposed for handling categorical data;
- such methods are “unable to handle noise and outliers”;
- are limited in terms of “discover clusters with non-convex shapes”;
- often require aprioristic information regarding the number of partitions to obtain; such information might be \( k \), number of clusters, or some other parameter (usually intimately linked to the number of clusters).

The employed partition might be crisp (in case each element strictly belongs to a single cluster) or fuzzy (in case elements are possible to be included within several clusters). Crisp partitioning algorithms are often generically defined as

\[
E = \sum_{i=1}^{c} \sum_{x \in C_i} d(x, m_i),
\]

(1)
where \( m_i \) is the representative of the cluster \( C_i \), and \( d(x, m_i) \) is the distance measure between a point \( x \) and \( m_i \).

One of the most illustrious methods used to perform crisp partitions is \( k \)-Means algorithm. It relies on a quite simple clustering criteria: “to minimize the distance of the objects within a cluster from the representative point of this cluster” (Halkidi et al., 2001) where such representative is given by “the mean value of the elements within that cluster” (MacQueen, 1967).

The second most popular method within the category of partitioning cluster methods is \( k \)-Medoids, popularized through implementation of the Partitioning Around Medoids (hereinafter: PAM) algorithm. PAM “minimizes a sum of dissimilarities instead of a sum of squared Euclidean distances” (allows for the use of an arbitrary non-Euclidean distance measure) for which “each cluster representative is the most centrally located object in a cluster.” (Kaufman & Roussseuw, 1990). The same does not hold for \( k \)-Means (based on pairwise Euclidean distances between data points): in fact, implementations of \( k \)-Means using non-Euclidean distance measures do not span Euclidean space. The authors describe PAM as being more robust to the presence of noise and outliers than \( k \)-Means. However, PAM carries an important drawback since its associated computational cost is quite demanding. Although this issue is becoming less crucial given the improvements in terms of processing power, it is still a limitation when trying to scale it for large datasets (Salvador & Chan, 2004).

Clustering for Large Applications (hereinafter, CLARA) is built on top of the same clustering criterion as PAM but it operates on samples of the data set, applying PAM throughout each one of those samples and returning the best partition as the final output. The fact that the iterations are performed over samples of the dataset allows for superior scalability than PAM. On the other hand, its performance depends on the sample size, producing results based on such samples and thus it leaves room for bias since “if a sample is biased, a good clustering based on samples will not necessarily represent a good clustering of the dataset.” (Halkidi et al., 2001). In order to overcome such shortcomings Ng and Jiawei Han (2002) suggest the use of CLARANS (often named Randomized CLARA), which randomly defines the subsets of the initial dataset to use as samples, mitigating bias through extending the search to wider areas of the dataset, and thus producing more consistent results.

When considering alternatives to obtain fuzzy partitions, analogue methods, such as Fuzzy \( c \)-Means (Kaufman & Roussseuw, 1990) and Fuzzy \( c \)-Medoids (Krishnapuram, Joshi, Nasraoui & Yi, 2001) are applicable. Although these are out of the scope of this research project (crisp partitions are preferred for the sake of simplicity) it is noteworthy to mention that these algorithms perform well under certain conditions such as searching for spherical-shaped clusters and while handling small/medium-sized sets of data.
1.4.3.2 Hierarchical Methods

This class of methods is based on the premise of grouping elements into a tree of clusters, usually graphically represented through means of a dendrogram. Hierarchical clustering includes two general types of methods:

- agglomerative methods, with a starting point where each element belongs to its own cluster and iteratively merging clusters into successively larger dimension clusters until all elements belong to a single cluster (or until certain pre-defined stoppage conditions such as a maximum number of clusters to define are met);
- divisive methods, starting from a single cluster that comprises all elements, these methods iterate splitting it into successively smaller clusters until each element belongs to its own cluster (or until certain pre-defined stoppage conditions such as a maximum number of clusters are met).

According to Liao (2005) both kinds of “pure hierarchical methods suffer from the inability to perform adjustments once a merge or split decision has been executed”. For this reason, hierarchical methods are usually combined with other clustering techniques to address such limitation. Although the use of hierarchical clustering methods does not grant the desired level of flexibility at each iteration, their application can be worthwhile when combined with other non-hierarchical methods. For that reason, the implementation carried in this study has its foundation on the application of hierarchical agglomerative clustering (hereinafter: HAC) methods as means of a synergy to decide for a suitable number of $k$ clusters, given the need of this aprioristic element as an input for non-hierarchical clustering application. Such class of methods makes use of a similarity-dissimilarity matrix used to “gradually build clusters by putting similar entities into the same cluster”, where each of the matrix constituting cells accounts for the “degree of similarity between two entities” (Blashfield, 1976).

Among all the HAC methods described in the available literature and illustrated in Figure 5, there are four methods that are commonly recognized to be the ones of most popular application according to Anderberg (1973), Manning, Raghavan and Schütze (2008) and Everitt (2011). The first method, single-linkage clustering, proposed by Sneath (1957) and later improved by Johnson (1967), Lance and Williams (1967) and finally McQuitty (1967), uses a local merge criterion assuming that “the similarity of two clusters is the similarity of their most similar members”, focusing solely over “the area where the two clusters come closest to each other” and ignoring “other, more distant parts of the cluster and the clusters’ overall structure”. Sokal and Sneath (1963) and Cormack (1971) point single-linkage’s major drawback to be the tendency to suffer from chaining effect as “an entity is added to a cluster if it is highly similar to any member of that cluster […] causing resulting clusters to resemble long chains when plotted in a multidimensional space”, discouraging most researchers to employ this method. The second method within this category is complete-linkage clustering, which considers non-local merge criterion if “the similarity of two clusters is the similarity of their most dissimilar members […] equivalent to choosing the cluster
pair whose merge has the smallest diameter”, typically resulting into “a preference for compact clusters with small diameters over long, straggly clusters, but also causes sensitivity to outliers.” (Manning et al., 2008). Although this method can be considered as chaining-effect-proof, Lance and Williams (1967) identify its “space-diluting” properties as a major shortcoming. Put simply, given the fact that “an entity cannot join a cluster until it obtains a given similarity level with all members of a cluster, the probability of a cluster obtaining a new member becomes smaller as the size of the cluster increases” (Blashfield, 1976). Average-linkage clustering, often named group-average agglomerative clustering, as proposed by Sokal and Sneath (1963), essentially relies on the definition of cluster as a “group of entities, in which each member has a greater mean similarity with all members of the same cluster than it does with all members of any other cluster” (Blashfield, 1976) and thus circumventing the pitfalls of the previous identified methods within this category, reason why it is considered to be the “most preferable” for a number of authors such as Sokal and Rohlf (1962), Sneath (1966), Rohlf (1970) and Cunningham and Ogilvie (1972). Other authors such as Williams, Clifford and Lance (1972) criticize this method upon proving its higher likelihood to produce “nonconformist groups as the size of clusters increase”.

Figure 5. Cluster Similarity Notions in HAC

The last method among the four most popular in the HAC category, is the Ward’s minimum variance, proposed by Ward (1963) and later enhanced by El-Hamdouchi and Willett (1986), iteratively merging the clusters with the smallest residual sum of squares (hereinafter: RSS). Defined as “the squared distance of each vector from its centroid summed over all vectors”, RSS is described in the literature as “a measure of how well the centroids represent the members of their clusters” (Manning et al., 2008). The agglomeration criterion applied in Ward’s method is intimately related with the one in the average-linkage clustering. The major drawback of this method it is, as denoted by Cormack (1971), its inherent bias towards producing spherical-shaped clusters.

1.4.3.3 Density-based Methods

Density-based clustering methods define clusters as higher element density areas within the dataset (Kriegel, Kröger, Sander & Zimek, 2011). The most renowned method within this class is DBSCAN
for which the main idea is to “continue growing a cluster as long as the density (number of objects or data points) in the neighborhood exceeds some threshold” (Sander, Ester, Kriegel & Xu, 1998).

1.4.3.4 Grid-based Methods

This class of methods is named after the “quantization of the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed” (Wang, Yang, & Muntz, 1997), for which the STING algorithm is the most popular, especially in application to spatial datasets of very large dimension.

1.4.3.5 Model-based Methods

Model-based methods are built on the premise of assigning to each of the clusters a different model, pursuing the best fit of the data to the respectively assigned model. Two major approaches within this class of methods are of special relevance:

- statistical, where the determination of the number of clusters makes use of Bayesian statistical analysis, and for which methods like “AutoClass” (Cheeseman & Stutz, 1996);
- artificial neural networks, including variations such as competitive learning and self-organizing feature maps.

Disregarding the broad variety of clustering algorithms belonging to different method categories, this research project is mainly focused in the application the most renowned hierarchical (agglomerative) and partitioning methods, leaving the implementation of additional clustering methods as an opportunity for further research.

1.4.4 Cluster Validity and Evaluation Criteria

Clustering, as previously mentioned, is a process that requires both knowledge on the data at use and the methods employed. Although, even possessing aprioristic knowledge on the data does not ensure an effective choice of the “optimal” grouping scheme: the one that best reproduces the latent partitions of the dataset. Cluster analysis requires trial and error experimentation and recursive evaluation and comparison of the produced solutions to attain the desired partition scheme. This section presents a number of criteria to assess the validity of produced clustering solutions. Regardless of the dataset geometric structure, each clustering solution is intimately connected to the algorithm at use and thus the values assigned to the underlying input parameters. As mentioned in the preceding section, some clustering methods require an aprioristic knowledge of the number of clusters to determine (e.g., partitioning methods). While this requires an extensive knowledge of the dataset, since multidimensional visualization is often limited, the first aspect to address is the choice of the “optimal” number of clusters to identify. In the sequel of the choice of the “optimal” number
of clusters follows the evaluation of the obtained solutions. Theodoridis and Koutroumbas (1999) present three main inspective approaches regarding cluster validity, each of them on the foundation of different evaluation criteria: external, internal and relative.

1.4.4.1 External Evaluation Criteria

External criteria are employed to “measure the agreement between two partitions where the first is the a priori known clustering structure, and the second results from the clustering procedure” (Dudoit & Fridlyand, 2002). It is immediate the conclusion that such criteria do not fit for the study conducted in this thesis give the fact that data is unlabeled and thus, there is no information regarding an existent partition structure of the dataset.

1.4.4.2 Internal Evaluation Criteria

On the other hand, internal criteria aim to “measure the goodness of a clustering structure without external information (...) evaluating the results using quantities and features inherent in the data” (Thalamuthu, Mukhopadhyay, Zheng & Tseng, 2006). Halkidi et al. (2001) discern two main applications of internal criteria upon the clustering structure. The former regards to the validation of a hierarchy of clustering schemes having its main application for the validation of solutions resulting from the employment of hierarchical clustering methods. Considering the cophenetic matrix, $P_c$, which stores in each $P_c(i,j)$ cell a distance measure (typically the dendrogram height) at which the $x_i$ and $x_j$ elements are included in the same cluster. Sokal and Rohlf (1962) define the Cophenetic Correlation Coefficient (hereinafter: CPCC) as an index to “measure the degree of similarity between $P$ and $P_c$ (proximity matrix) matrices”. This index varies within the $[-1, 1]$ range, and can be formalized as

$$CPCC = \frac{1}{\sqrt{\left(\frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^2 - \mu_P^2 \right) \left(\frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} c_{ij}^2 - \mu_c^2 \right)}} \tag{2}$$

where:

- $N$ is the number of constituting elements of the dataset;
- $M = N \cdot (N - 1)/2$;
- $\mu_P = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} P(i,j)$;
- $\mu_c = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} P_c(i,j)$, with;
- $\mu_P$ and $\mu_c$ the respective means of $P$ and $P_c$ matrices.
The CPCC must be interpreted in terms of the level of similarity between $P_c$ and $P$ matrices, for which values close to 0 support the evidence of a high similarity between the two matrices and thus a low-quality solution. Values closer to 1 are therefore preferred.

The latest application of internal cluster evaluation criteria concerns to the validation of a single partition scheme, investigating the level of similarity between a given partition $C$ and the proximity matrix $P$. In order to perform such similarity check, Halkidi et al. (2001) suggest the use of Hubert’s $\Gamma$ statistic defined as

$$\Gamma = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} X(i,j)Y(i,j)$$

or the normalized Hubert’s $\Gamma$ statistic defined as

$$\Gamma = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{(X(i,j) - \mu_X)(Y(i,j) - \mu_Y)}{\sigma_X \sigma_Y}$$

where $X$ and $Y$ are the matrices under comparison and $\mu_X$, $\mu_Y$, $\sigma_X$ and $\sigma_Y$ are their respective means and variances. The use of the normalized statistic provides values within the $[-1,1]$ range, which are of easier interpretation.

1.4.4.3 Relative Evaluation Criteria

Finally, relative criteria are useful when “comparing a specific partition structure against other clustering schemes obtained using the same algorithm but with different parameter values” (Halkidi et al., 2001). This approach implies a reduced computational cost when compared to the previous external and internal criteria given their statistical-based nature. Relative evaluation can be performed by the evaluation of certain validity indices that allow the identification $k$, optimal number of clusters, even when applying algorithms that require the specification of this parameter as an input. Such validity indexes are computed for each run of an algorithm within a certain targeted range of $k$. As a result, it is possible to plot the obtained values of the index against the respective number of clusters considered. Each plot requires one of two different interpretations:

- for the cases where there is no evidence of an increasing (decreasing) trend of the index as a result of an increase (decrease) of $k$, the ideal number of clusters is given at the value of $k$ for which the index is maximized (minimized);
on the contrary, when a trend in the index value in terms of $k$ is not evident, the optimal number of clusters is usually given by “the value of $k$ at which a significant local change in value of the index occurs”; in graphical terms, this point represents a “knee” which usually outstands the remaining plotted structure. Additionally, the authors suggest that “the absence of a knee may be an indication that the dataset possesses no clustering structure”.

While there is a variety of validity indexes that can aid the search for the optimal number of clusters and thus to the optimal partition structure, this study comprehends a set of those which have been popularized within the relevant literature. The first here presented is the modified Hubert $\Gamma$ statistic defined as a particular case of the aforementioned Hubert $\Gamma$ statistic having $P$ (dataset proximity matrix) and $Q_{N \times N}$ (matrix containing the distance between the cluster representatives for each element of the dataset) as the matrices under comparison. A large value obtained as a result of this implementation indicates evidence towards more compact clusters. Additionally, this implementation can be done applying similar reasoning to $\Gamma$ (normalized Hubert $\Gamma$). The graphical interpretation of the plot of the Hubert $\Gamma$ statistic in terms of “clusters, the optimal number for” is located at the value where a significant “knee” outstands.

Another validity index here addressed is the Dunn index, useful for identifying “compact and well separated crisp clusters” (Dunn, 1974). It is possible to formalize the Dunn index, $D_k$, as a function of both dissimilarity and diameter, given by the equation

$$D_k = \min_{i=1, \ldots, k} \left\{ \min_{j=i+1, \ldots, k} \left( \frac{d(c_i, c_j)}{\max_{t=1, \ldots, k} \text{diam}(c_t)} \right) \right\}, \quad (5)$$

having:

- $d(c_i, c_j) = \min_{x \in c_i, y \in c_j} d(x, y)$, dissimilarity between two clusters $c_i$ and $c_j$;
- $\text{diam}(C) = \max_{x, y \in C} d(x, y)$, diameter of a cluster $C$, defined as the maximum distance between the two most distant elements belonging the same cluster.

Conversely, given a dataset comprising compact and well-separated clusters, large values of $D_k$ are expectable, implying the evidence for the existence of a partition structure where clusters are narrow (in diameter) and widely separated. Given the absence of a trend in $D_k$ regarding $k$, the optimal number of clusters is graphically represented by the value of $k$ at which the $D_k$ index is maximum. Some common known drawbacks of using $D_k$ as a validity index are both the computational effort required and its sensitivity to noise and outliers which might produce larger values of $\text{diam}(C)$.
Rousseeuw (1987) suggest the use of silhouettes as cluster validation technique of very wide application regarding the clustering methods at use. Considering the total number of observations for a given dataset, \( n \), the Silhouette index is given by the equation

\[
\text{Silhouette}_k = \frac{1}{n} \sum_{i=1}^{n} \frac{s(i)}{\max\{a(i); b(i)\}},
\]

where:

- \( a(i) = \frac{\sum_{j \in |C_i \setminus j|} d_{ij}}{n_i - 1} \),
- \( d_{ic_s} = \frac{\sum_{j \in C_s} d_{ij}}{n_s} \),
- \( b(i) = \min_{s \neq r} d_{ic_s} \), and;
- \( a(i) \) and \( d_{ic_s} \) are the average dissimilarities of the \( i \)-th element to the other elements contained in clusters \( C_r \) and \( C_s \), respectively;

Silhouette\(_k\) is defined for \( k > 1 \) clusters, and it ranges within \([-1,1]\), where high values imply that dataset elements are generally well assigned to their respective natural cluster (and thus disassociated of any neighboring cluster), supporting the evidence for an appropriate partition scheme (Charrad, Ghazzali, Boiteau & Niknafs, 2014).

Davies & Bouldin (1979) purpose a measure of the average similarity between each of the clusters identified within a partition structure, the Davies-Bouldin index, defined as

\[
\text{DB}_k = \frac{1}{k} \sum_{i=1}^{k} R_i,
\]

where:

- \( R_i = \max_{i=1,\ldots,k \atop i \neq j} R_{ij} \), and;
- \( R_{ij} = (s_i + s_j)/d_{ij} \) is the ratio between the sum of the dispersions of the clusters \( i \) and \( j \), \( s_i \) and \( s_j \), respectively, and the distance measure computed for those clusters, \( d_{ij} \).
The presented definition of $R_{ij}$ is purposed given its simple implementation in order to satisfy a set of preliminary conditions implying that $R_{ij}$ must be a nonnegative and symmetric matrix. This index presents no trend in terms of the number of clusters, reason why the optimal number of clusters is given by the value of $k$ at which $DB_k$ is minimum, given the objective of searching for a clustering scheme that minimizes the similarity between each of the existent clusters.

Contrary to the previously mentioned cluster validity indices, Sharma (1996) purposes the use of R-Squared (hereinafter: RS) as a “measure of the degree of homogeneity between groups”. RS recursively computed for each iteration of a hierarchical clustering, more specifically for each merge within agglomerative process described in the context of HAC methods. The RS for the $k^{th}$ merging iteration is defined as

$$R_k^2 = \frac{SS_b}{SS_t}, \quad (k = 1, \ldots, n - 1) \tag{8}$$

where $SS_b$ is the between group sum of squares and, since $SS_t = SS_p + SS_w$ or conversely, $SS_b = SS_t - SS_w$, i.e., the larger the distance between the clusters, the higher the homogeneity degree. That being said, when plotting the values of RS against the number of clusters under investigation, the most salient “knee” represents the optimal number of clusters. Similarly to RS, the analysis of the within group sum of squares, $SS_w$, can be useful to search for the clustering solution comprehending an optimal number of clusters for a certain employed method. Again, the target value for the desired number of clusters, $k$, is located at the most significant “knee”. As $SS_w$ tends to decrease as $k$ grows, it is possible to conclude that including additional clusters minimizes heterogeneity (or, maximizes homogeneity). That being said, the optimal number of clusters, $k^*$, represents the partition for which the homogeneity is maximized at the cost of including an additional cluster.

1.4.5 Longitudinal Data Clustering Methods

The previously presented cluster methods comprehend algorithms frequently used in classic cluster analysis problems. Notwithstanding their proven efficiency and the quality of the solutions produced as a result of their implementation, when clustering a set of objects or individuals according to their time dynamic behavior, the methods above usually fail to account for time dependencies. For that reason, this section is meant to introduce few concepts regarding longitudinal clustering.

1.4.5.1 Definition of Longitudinal Data

Longitudinal studies imply the measurement of a set of variables over time. Because of this time-dynamic profile, variables are frequently described in the literature as “variable-trajectories” or “joint-trajectories”, when considering longitudinal datasets where several variables are under observation (Genolini, Alacoque, Sentenac, & Arnaud, 2015). To attain a formal definition of a
longitudinal dataset clustering problem, there is the need to define it using appropriate notation properly. Consider \( S \) as the set of \( n \) subjects in the dataset, for which \( m \) variables are measured over \( t \) time periods. A single trajectory, measurement of the same variable over time with respect to a certain subject is formally represented as

\[
y_{iA} = (y_{i1A}, y_{i2A}, \ldots, y_{iBA}).
\]

A set of multiple single trajectories encompasses a joint-trajectory. Each joint-trajectory is composed of the set of measurements of \( m \) variables on the \( j^{th} \) period \((j = 1, \ldots, t)\) for the respective a certain subject, being formally defined as

\[
y_L = \begin{pmatrix}
  y_{i,A} \\
y_{i,B} \\
  \vdots \\
y_{i,M}
\end{pmatrix} = \begin{pmatrix}
  y_{i1A} & y_{i2A} & \cdots & y_{iBA} \\
y_{i1B} & y_{i2B} & \cdots & y_{iBB} \\
  \vdots & \vdots & \vdots & \vdots \\
y_{i1M} & y_{i2M} & \cdots & y_{iBM}
\end{pmatrix}
\]

having each row representing a single variable trajectory.

1.4.5.2 Strengths and Weaknesses of Longitudinal Data Clustering

The application of cluster analysis in the context of longitudinal studies seeks to cluster joint-trajectories of several individuals into groups of homogeneous characteristics. A variety of methods to attain this main goal have been proposed by Tarpey and Kinateder (2003) and Nagin (2005), for which the main advantage of such is the fact of enabling “the conversion of several correlated continuous variables into a single categorical variable”. On the other hand, the drawbacks of longitudinal clustering are various. Such methods are typically proposed for tackling a single variable-trajectory problem, with most joint-trajectory clustering methods turning out to be inefficient since they do not capture important underlying interactions between each of the variables under observation. Also, and as it stands for the previously presented clustering methods, “there is no reliable method to determine the ‘true’ number of clusters in a dataset” (Everitt et al., 2011). Although there is no consensus among the scientific community in this matter, such assessment is commonly done by combining several validity indices to aid a more appropriate choice of the optimal number of clusters. Examples are the Calinski-Harabasz (1974), the Davies-Bouldin (1979) and Rousseeuw's average silhouette (1987) indices.

Although some accurate missing value imputation methods have been developed over the years, Laird (1988), Little (1993), Hedeker and Gibbons (1997) and Mallinckrodt, Lane, Schnell, Peng and Mancuso (2008) emphasize the inherent loss of information that results from restricting the study to individual for which the set of measurements is fully complete. This situation is naturally applicable in the context of this research study given the existence of missing data inherent to the data.
collection process. To hold simplicity and to avoid artificial bias resulting from the imputation of missing values, the study hereby carried only comprises join-trajectories of countries for which data completion is not problematic. Finally, the use of partitioning techniques is said to highly depend on starting conditions specified at the algorithm initialization and, therefore, each run of an algorithm might produce results that approximate the former run, being a best practice to include a reasonably large enough number of runs in search for a convergent solution.

1.4.6 Text Clustering Methods and Algorithms

The purpose of the present subsection is to introduce the methods and algorithms assessed during the data cleansing process regarding some of the text attributes from the GTD dataset (more specifically provstate and city), providing a solid background for the interpretation of the results obtained in that context.

Starting with stricter methods, Key Collision stands for a family of methods built on the premise of designing a surrogate representation of a value (commonly named “key”) containing only the most meaningful part of the string, binning together several different strings given the same key for those strings (Verborgh & De Wilde, 2013). These methods are quite linear and straightforward from the computational perspective ensuring a satisfactory ratio between efficiency and speed, even when operating on millions of values, depending on the selected keying function. Using the “Fingerprint” keying function provides a satisfactory simplification level at a high pace for the context of clustering names of states, provinces and cities, due to its scarce likeliness to produce false positives. The mechanism that creates the key from a text string value is summed up by the following workflow:

- remove leading and trailing whitespace;
- transform all characters into lowercase representation;
- erase any punctuation, control characters;
- sort the tokens and discard duplicates;
- merge tokens back together;
- normalize EASCII characters using their ASCII representation (e.g., béjaïa" → "bejaia").

The advantages of this algorithm rely on the fact that whitespace is standardized, characters are lowercased, the punctuation is ignored and the strings parts order is not relevant, having all these elements no crucial role in the clustering decision. Employing the “N-Gram Fingerprint” keying function produces similar results to the previously stated algorithm, using n-grams (where the n is a user specified character length of the token) instead whitespace separated tokens as noted by Broder, Glassman, Manasse and Zweig (1997) and Dunning (1994). The “Metaphone-3” algorithm using a phonetic fingerprint keying function (for English language) was tested having the worse performance within the three mentioned fingerprinting methods. This algorithm differs from the
remaining because it transforms tokens into the way they are pronounced, being helpful to spot typos resulting from human misunderstanding or misspelling of words, assigning the same key to resembling words, falling into the same cluster.

Moving to laxer depth clustering approaches, the “Nearest Neighbor” (hereinafter: k-NN) method allows a calibration of the resilience of the algorithms given the strings to cluster. By fine-tuning the value for $k$ (or radius) it is possible to set a threshold for the distance between any pair of strings to be included in the same cluster. One of the disadvantages arising from this method is the higher computational complexity since, given $n$ text strings, there are $n(n-1)/2$ pairs of strings, resulting in the same amount of relative distances to be computed and compared. In the case of the GTD dataset in use, that amount of distances is bigger than $10^9$, given $n \approx 142,000$. “Blocking”, a technique developed to extremely reduce the number of strings to be matched against each other, consists in distributing the strings among $n$ blocks and, given an average block size, each string is just matched within its block, resulting in only $n \times m(m-1)/2$ distances to be computed. Given the GTD dataset in use, given $n \approx 142,000$ and for an average block size of $m = 6$ (default value in OpenRefine), the computational effort is significantly reduced down to 355,000 distances to be computed. In sum, regardless the above mentioned $k$-NN features, the clustering efficiency varies mostly on the algorithm implemented to evaluate the resulting distances. The implementation of the Levenshtein distance is quite straightforward in order to attain the aforementioned purpose, gauging the minimum number of editing operations required to transform a string into its cluster representative (for this reason this algorithm is also called “Edit distance”). This particularity makes such algorithm quite flawless in detecting typos and mistakes in general, although it is prone to generate manifold of false positives when the edit distance is considered large. Lastly, the “Prediction by Partial Matching” (hereinafter: PPM) algorithm implements the insights from Li, Chen et al. (2004) regarding the estimation of strings resemblance by means of the Kolmogorov Complexity. Generally, most of the text compressors run based on the assumption that, given two identical text strings $A$ and $B$, the results from compressing $A$ or compressing a concatenated version of both, say $A+B$, produce minor difference; however, if the differences between $A$ and $B$ are sizable, this may culminate in unbending differences regarding the length resulting from the analogous compression as pointed by Chang and Witten (1984) and Moffat (1990). The normalized implementation present in OpenRefine computes the distance between those two strings as

$$d(A, B) = \frac{\text{len}(A+B) + \text{len}(B+A)}{\text{len}(A+A) + \text{len}(B+B)}$$

(11)

where $\text{len}(t)$ stands for the byte length of the compressed sequence of the string $t$. The PPM compressor is empirically known to have outstanding performance in the context of text compressing, using statistical methods to make predictions on string generation. Notwithstanding, there may arise a significant amount of false positives but, those might be useful in identifying non-obvious substructures, reason why this algorithm should be used just as an alternative given poor performance from all previously stated algorithms.
2 DATA COLLECTION AND PREPARATION

As stated in the preliminary phase of this written work, the analysis hereby carried out involves two primary kinds of data: the first relates to terrorist events and the second one to financial and operational facts pertaining the hospitality industry from the countries (or regions) under the scope of the study. The purpose of this section is to present the databases used regarding the data collection process and the assumptions that are seen to be crucial for the same, and relevant considerations regarding the data cleansing techniques used for attaining superior efficiency during further analysis. Regarding the collection of the data, it is important to notice that all data was obtained as a secondary data collection process and so there might exist limitations arising from the fact that the illustrated facts were not initially meant for this analysis. Moreover, it is important to match the two kinds of data to derive useful insights and understand any existing constraints from this matching process.

2.1 Terrorism Data: Global Terrorism Database

The current section presents the first of a twofold kind of data, more precisely, data about a collection of terrorist attacks recorded and described in both qualitative and quantitative terms.

2.1.1 Origin and Purpose

The database here addressed is the Global Terrorism Database, maintained by the National Consortium for the Study of Terrorism and Responses to Terrorism (hereinafter: START), a Department of Homeland Security Center of Excellence led by the University of Maryland, U.S.A. The GTD results from the collaboration between several public and private institutions hereby listed in chronological order of collaboration: the Pinkerton Global Intelligence Service (hereinafter: PGIS), the Center for Terrorism and Intelligence Studies (hereinafter: CETIS), the Institute for the Study of Violent Groups (hereinafter: ISVG) and the previously mentioned START. The early stages of the data collection processes undertaken by the previously mentioned institutions are described by LaFree and Dugan (2007) and their ownership is summarized in Table 1.

<table>
<thead>
<tr>
<th>Dates Regarding Records in GTD</th>
<th>Responsible Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beginning</td>
<td>Ending</td>
</tr>
<tr>
<td>1/1/1970</td>
<td>12/31/1997</td>
</tr>
<tr>
<td>1/1/1998</td>
<td>3/31/2008</td>
</tr>
<tr>
<td>4/1/2008</td>
<td>10/31/2011</td>
</tr>
<tr>
<td>11/1/2011</td>
<td>12/31/2014</td>
</tr>
</tbody>
</table>

Source: Adapted after Global Terrorism Database, Codebook: Inclusion Criteria and Variables, 2017.
2.1.2 Data Collection: Methodology and Limitations

The GTD results from efforts from the aforementioned institutions that relied on publicly available, unclassified source materials including media articles, electronic news archives, books, journals and legal documents. Naturally, the selected data sources are not all of the same quality, being prioritized the one considered to be of high-quality, i.e., sources that benefit from governmental, political or corporate independence, that regularly report externally indubitable content, and those that arise from primary rather than secondary collection processes. Note that due to the insufficiency of high-quality sources in particular geographies the data collection process followed a conservative approach when documenting the attacks.

The current data collection methodology applies since START officially led the project and combines automated and manual data collection techniques aiming to “maximize the efficiency, accuracy and completion of the collection” (National Consortium for the Study of Terrorism and Responses to Terrorism, 2016). Starting from a pool of over one million articles of any topic published worldwide on daily basis, the process iteratively subsets the pool into smaller fractions accordingly to the relevance with the terrorist attacks topic. By recursively applying tailored (and refined) keyword filters to each iteration pool through a subscription to a Metabase Application Programming Interface fed with articles from the Open Source Center, including English-language translations of sources from 160 countries in more than 80 languages. The processing of the smaller subsets is then carried out through the use of refined natural language processing and machine learning techniques in order to derive superior results and discard duplicate registers, being then manually assessed according to the GTD inclusion criteria and coding rules described in the GTD Codebook.

As mentioned in the previous section the definition of terrorist attack used for the scope of this thesis is the one suggested by START on the exercise of the data collection process currently being described. This definition is crucial when defining the inclusion criteria for the collected incidents into the database itself. Putting it simple, there are three mandatory criteria that need to be fulfilled for a record to be included in the GTD:

- the incident must be intentional;
- the incident must entail some level of violence or immediate threat of violence;
- the perpetrators of the incidents must be sub-national actors – the database excludes any case of state terrorism.

Additionally, any event is eligible for inclusion into the GTD if it verifies at least two of the following three supplementary criteria:
• the act must be aimed at attaining a political, economic, religious or social goal;
• there must be evidence of an intention to coerce, intimidate, or convey some other message to one or more larger audiences than the immediate victims;
• the action must be outside the context of legitimate warfare activities.

It is also relevant to mention that the inclusion of records in the GTD relies on a single incident determination principle that consists in the aggregation of all events occurring in both the same geographic and temporal point as a single event. In opposition, i.e. if there is evidence of any discontinuance in the time or location of the attack, the incidents are compiled as multiple separate records.

Although the data collection is an ongoing process carried out by START, the very last version available at the time of this written work synthetizes facts from attacks registered during the 1970-2014 period for worldwide terrorist attacks. A brief note goes for the fact that incidents that occurred during the year of 1993 were not originally included in the GTD since they were lost prior to START assume the responsibility of compiling the existing sources. Unfortunately, the retrospective data collection efforts faced several unavoidable challenges resulting in the identification of about 15% of the estimated incidents. This fraction was then manually included according to the standards of the remaining database in order to preserve the integrity and consistency of the data. This being said, there may exist evidences of a discontinuance for that same year (1993) for any of the time series illustrated by the data to be analyzed and such perturbations (e.g., shocks on terroristic activity) shall not be accounted with severe importance due to lower data reliability.

2.1.3 Metadata and Data Cleansing Codebook

The collaboration of the mentioned institutions resulted in more than 140,000 registered terrorist incidents included in GTD at the time of the elaboration of the exploratory analysis hereby carried out. The growing sophistication of the employed text mining techniques enabled the collection of a range of 134 distinct attributes describing the nature of each incident. Selecting the right attributes in the context of the data preparation is an important milestone for the efficiency and consistency of future results, given certain aspects such as the value of the insights for them provided and the computational constraints inherent to the dimension and type of data to be used. To understand which attributes are useful for further analyses, it is vital to notice that the attributes captured during the secondary data collection illustrated in Table 1 are of distinct nature. Not all the listed categories (and thus, the corresponding attributes) are of the same relevance, reason why it is vital to make a selection of those that are valuable for achieving the purpose of this thesis. This subsection aims to identify and describe the relevant attributes regarding the scope of this research work, providing a clear understanding of the variables used in further analyses, but also clarifying the assumptions on which all the data preparation relies. In addition, the fundamental data cleansing procedures undertaken are also described next.
2.1.3.1 GTD ID and Date

Every record depicts a single incident for which it corresponds a unique identifier (eventid). For each record (and, hence, for each eventid) the chronological attribute selected is the year of occurrence of the attack (iyear).

2.1.3.2 Incident Location

Under this category, it is possible to identify several relevant attributes that provide useful information respecting to the geographic localization of the incident sites, crucial for the good understanding of the terrorism “hot-spots” distribution. The attribute region is meant to gather countries into logical groups according to their geographic location. Country and region (categorical variables) are both straightforward and relevant for inclusion in further descriptive and predictive models. Note that these attributes originally incorporate several inconsistencies due to the change of political circumstances that led to the redefinition of several borders, countries and sovereign states over time. The most frequently found insufficiencies are of three main types:

- for some specific territories events are either recorded as taking place in the respective country or in the corresponding sovereign state of which that country is part of. The most notorious of these cases in the spotlight here is certainly Northern Ireland, where for attacks hosted in that territory are often coded for the country attribute as either Northern Ireland or the United Kingdom. To keep consistency between the terrorism and the tourism data (presented in the next section), data from attacks occurring in Northern Ireland territory will be re-coded as incidents in United Kingdom;
- a specific country hosting an incident at a certain time does not longer exists as that same entity: as an illustrative example, an incident perpetrated in 1989 in Bonn would have been recorded has taken place in West Germany (FDR); a similar attack occurring in 1991 would be coded for country as Germany (DEU). Given the major political changes that occurred in Europe over the last four decades (detailed in Table 2), the existing country coding does not fit for the purpose of the analyses to be carried out in further sections: there is the need to recode terrorist incidents pertaining to dissolved (or merged) countries as perpetrated in the resulting country for the same period(s) to analyze;
- for the specific cases of the former Soviet Socialist Republics, given the dissolution of the USSR, there is the need to recode such countries according to their geographic region since part of them shall be included under “Eastern Europe” and the remaining under “Central Asia”; these changes are also detailed in Table 2.

Resolving these latent insufficiencies of the data through inspection is quite time consuming and error prone task and may lead to the loss of accuracy of the data. For those reasons, and to minimize
this accuracy loss, the aforementioned data cleansing tasks are performed only for the mentioned group of countries, since the main targets of this thesis are European countries.

Table 2. Political Changes in Europe

<table>
<thead>
<tr>
<th>Dissolved Country</th>
<th>Flag</th>
<th>Process</th>
<th>Date</th>
<th>Originated Country</th>
<th>Flag</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>East Germany (GDR)</td>
<td></td>
<td>Unification</td>
<td>3/10/1990</td>
<td>Germany</td>
<td></td>
<td>Western Europe</td>
</tr>
<tr>
<td>West Germany (FRG)</td>
<td></td>
<td></td>
<td>3/10/1990</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Union of Soviet Socialist Republics (USSR)</td>
<td></td>
<td>Independence</td>
<td></td>
<td>Georgia</td>
<td></td>
<td>Central Asia</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4/9/1991</td>
<td>Latvia</td>
<td></td>
<td>Eastern Europe</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8/21/1991</td>
<td>Russian Federation</td>
<td></td>
<td>Eastern Europe</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Independence</td>
<td>8/24/1991</td>
<td>Ukraine</td>
<td></td>
<td>Eastern Europe</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8/25/1991</td>
<td>Belarus</td>
<td></td>
<td>Eastern Europe</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Independence</td>
<td>8/27/1991</td>
<td>Moldova</td>
<td></td>
<td>Eastern Europe</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8/30/1991</td>
<td>Azerbaijan</td>
<td></td>
<td>Central Asia</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Independence</td>
<td>8/31/1991</td>
<td>Kyrgyzstan</td>
<td></td>
<td>Central Asia</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9/1/1991</td>
<td>Uzbekistan</td>
<td></td>
<td>Central Asia</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9/9/1991</td>
<td>Tajikistan</td>
<td></td>
<td>Central Asia</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9/17/1991</td>
<td>Estonia</td>
<td></td>
<td>Eastern Europe</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9/17/1991</td>
<td>Lithuania</td>
<td></td>
<td>Eastern Europe</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9/21/1991</td>
<td>Armenia</td>
<td></td>
<td>Central Asia</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10/27/1991</td>
<td>Turkmenistan</td>
<td></td>
<td>Central Asia</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>12/16/1991</td>
<td>Kazakhstan</td>
<td></td>
<td>Central Asia</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9/8/1991</td>
<td>Macedonia</td>
<td></td>
<td>Eastern Europe</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1/1/1992</td>
<td>Slovenia</td>
<td></td>
<td>Eastern Europe</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Independence</td>
<td>4/11/1992</td>
<td>Bosnia and Herzegovina</td>
<td></td>
<td>Eastern Europe</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6/3/2006</td>
<td>Montenegro</td>
<td></td>
<td>Eastern Europe</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6/3/2006</td>
<td>Serbia</td>
<td></td>
<td>Eastern Europe</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2/17/2008</td>
<td>Kosovo</td>
<td></td>
<td>Eastern Europe</td>
</tr>
<tr>
<td>Czechoslovakia (CSK)</td>
<td></td>
<td>Independence</td>
<td>1/1/1993</td>
<td>Czech Republic</td>
<td></td>
<td>Eastern Europe</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1/1/1993</td>
<td>Slovakia</td>
<td></td>
<td>Eastern Europe</td>
</tr>
</tbody>
</table>

Source: Adapted after Global Terrorism Database, Codebook: Inclusion Criteria and Variables, 2017.

The categorical attributes for province/state and city (respectively, provstate and city) present severe inconsistencies due to their non-structured text nature. The most notorious examples are of three types: designation misspelling (typos), redundant/alternative designations for the same value (e.g., “J&K” or “Jammu & Kashmir” as alternative representations of “Jammu and Kashmir”) and, the use of ambiguous designations (e.g., “15 km from South African border”). Additionally, it is also
not practical to use such categorical attributes in both clustering or regression analysis, since the immense range of possible categories is likely to generate inaccurate results with poor ability to generalize conclusions. Thusly, provstate and city are discarded from variables to include in any descriptive model. However, when combining the information contained in such attributes with the information from country, it is possible to generate data to fulfill missing values from variables such as latitude and longitude used for visualization purposes when mapping the location of the incidents, to understand the terrorism hot-spots to be analyzed in further detail. Under those circumstances, there are two inherent major challenges from the data cleansing perspective: the first, is to refine the information provided by provstate and city; the second, to use that outcome values to generate the coordinates for incidents where these data are missing. To attain this, it is a priority to tackle the inconsistencies in provstate and city attributes by applying the text mining methods (through means of string clustering, as mentioned in 0) to generate uniform designations. This procedure strongly relies on the wide set of data cleansing and reconciliation of OpenRefine, from which the regular Fingerprint algorithm was shown outstanding performance, leading to a smaller number of false positives and thus more consistent clustered values than all other algorithms tested. As a second part of this challenge, it is possible to generate approximate coordinates for the previously mentioned event sites by fetching URLs from Google Map’s API, resulting in a substantial improvement of the level of data completion, making data visualization more accurate and representative of the true panorama of the global terrorism hotspots.

2.1.3.3 Attack Information

This category of attributes gathers information regarding the nature of the attacks, enabling the distinction between the various strategies employed by terrorists to reach their objective. The main attributes to highlight within this category are success and suicide. Success informs about the completion or materialization of the attack. This attribute is selected to perceive if the success of a terrorist attack has a significant impact on the tourist decisions and their intentions to travel to a recently (or frequently) attacked country. Suicide codes the attack regarding the employment of suicidal practices from the side of the perpetrator(s) and it is included for further analyses to assess if the risk perception of tourists is affected by such practice. Additionally, Appendix B presents metadata regarding attacks classification in the GTD framework.

2.1.3.4 Perpetrator Information

This class of attributes related to the perpetrator entity (group or individual) providing information regarding their affiliation, the number of individuals involved in each incident and from those how much of them were captured in the aftermath of an attack. Assuming that, regardless of the identity of their perpetrators, terrorist attacks have, in general, impacts on the hospitality industry, it is legit to prioritize the identification of an attack claim over the identification of its perpetrator. The existence of a public responsibility claim of attack is believed to impact hospitality industry stakeholders’ perception of a terrorist threat, since such vindication naturally raises their awareness.
Hence, the attribute *claimed* codes if the responsibility for a terrorist attack was claimed (by a single or multiple perpetrators).

2.1.3.5 Casualties and Consequences

To assess the damage impact of each terrorist incident, this category provides attributes that measure the resulting mortality, injuries, and material damages. This class contains those attributes that provide the most valuable insights regarding the danger perception and terrorist threat awareness. This being said, it is crucial to select which ones to test as part of any forthcoming model to keep the simplicity of the same. In spite of the findings from Pizam and Fleischer (2002) pointing out that attacks’ frequency generally has more impact on tourism demand over their severity, it is still important to use such variables for this research work since casualties (and consequences) are generally seen as “key metrics” on terrorist attacks, from the perpetrator perspective. For this reason, the first attribute to highlight here is *nkill*, numeric variable that records the total number of fatalities (including all victims and perpetrators who died as a direct outcome of the incident). It is important to remark the fact that GTD sometimes presents fractioned values (instead of integer numbers) for this variable: in cases when several incidents are linked together, the sources from where the data collection resulted may present a cumulative number of casualties for all those incidents; in order to preserve the statistical accuracy, the aforementioned cumulative values for casualties are evenly divided for the number of linked incidents. While cleansing the data for this research work, the strategy hereby adopted is to simply round all values to the next integer value - even though this approach might raise some issues regarding data consistency, the total deviation implied is smaller than 0.01% of the actual values and having little or no impact on the analysis. Similarly, and facing the same issues regarding the data collection, *nwound* refers to the number of (both victims and perpetrators) injured as a direct result of the attack.

To capture the number of hostages or kidnapping victims, the numeric variable *nhostkid* is kept as a relevant variable to test in further analyses. For successful hijackings, this attribute reflects the total number of crew members and passengers aboard the vehicle at the time of the incident. The data collection for this attribute follows a conservative approach regarding vaguely stated figures: the number of hostages or kidnapping victims to be recorded is the lowest among the various sources in use.

Regarding material damages, and to simplify the inclusion of variables in further models, the attribute hereby selected is *propvalue*, a numerical variable that captures the value of the property damage (in USD) registered as a direct result from the incident. Note that the value of damages only includes the direct economic impact of the incident (e.g., cost of the damaged buildings, etc.) but not indirect economic costs (longer term effects on specific industries, tourism, etc.). Also, for the values left blank is important to remark them as inaccurate estimates (in USD) instead of the absence of property damage.
Table 3 summarizes the attributes that were previously considered as being of relevance for the further analytic steps of this research work, together with a brief overview of the respective metadata.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Data Type</th>
<th>Scale Type</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>eventid</td>
<td>Char</td>
<td>Nominal</td>
<td>Unique identifier.</td>
</tr>
<tr>
<td>iyear</td>
<td>Integer</td>
<td>Interval</td>
<td>Year of the attack.</td>
</tr>
<tr>
<td>country</td>
<td>Varchar</td>
<td>Nominal</td>
<td>Country where the attack was perpetrated.</td>
</tr>
<tr>
<td>region</td>
<td>Varchar</td>
<td>Nominal</td>
<td>Region of the planet where the attack was perpetrated.</td>
</tr>
<tr>
<td>latitude</td>
<td>Double</td>
<td>Interval</td>
<td>Latitude of the attack site.</td>
</tr>
<tr>
<td>longitude</td>
<td>Double</td>
<td>Interval</td>
<td>Longitude of the attack site.</td>
</tr>
<tr>
<td>nkill</td>
<td>Integer</td>
<td>Ratio</td>
<td>Number of casualties as a direct result of the attack.</td>
</tr>
<tr>
<td>nwound</td>
<td>Integer</td>
<td>Ratio</td>
<td>Number of wounded people as a result of the attack.</td>
</tr>
<tr>
<td>propvalue</td>
<td>Integer</td>
<td>Ratio</td>
<td>Estimated property damage value as a direct result of the attack (in current USD).</td>
</tr>
<tr>
<td>nhostkid</td>
<td>Integer</td>
<td>Ratio</td>
<td>Number of hostages kidnapped as a direct result of the attack.</td>
</tr>
<tr>
<td>success</td>
<td>Boolean</td>
<td>Nominal</td>
<td>&quot;1&quot; if the attack direct goals were (partly or fully) achieved; &quot;0&quot; if the attack failed to attain any of the intended direct goals.</td>
</tr>
<tr>
<td>multiple</td>
<td>Boolean</td>
<td>Nominal</td>
<td>&quot;1&quot; if the attack is part of a set of multiple events; &quot;0&quot; if the attack is an isolated event.</td>
</tr>
<tr>
<td>extended</td>
<td>Boolean</td>
<td>Nominal</td>
<td>&quot;1&quot; if the attack took place during one single day period; &quot;0&quot; if the attack lasted for longer than one day.</td>
</tr>
<tr>
<td>suicide</td>
<td>Boolean</td>
<td>Nominal</td>
<td>&quot;1&quot; if any of the perpetrators committed suicide during the attack; &quot;0&quot; if none of the perpetrators committed suicide during the attack.</td>
</tr>
<tr>
<td>claimed</td>
<td>Boolean</td>
<td>Nominal</td>
<td>&quot;1&quot; if (at least) one group/individual claimed responsibility for the attack; &quot;0&quot; if no group/individual claimed responsibility for the attack.</td>
</tr>
</tbody>
</table>

2.2 Tourism Data – World Tourism Indicators

"Tourism is officially recognized as a directly measurable activity, enabling more accurate analysis and more effective policy. Whereas previously the sector relied mostly on approximations from related areas of measurement (e.g. Balance of Payments statistics), tourism today possesses a range of instruments to track its productive activities and the activities of the consumers that drive them: visitors (both tourists and excursionists). An increasing number of countries have opened up and invested in tourism development, making tourism a key driver of socio-economic progress through export revenues, the creation of jobs and enterprises, and infrastructure development. As an internationally traded service, inbound tourism has become one of the world's major trade categories. For many developing countries it is one of the main sources of foreign exchange income and a major component of exports, creating much needed employment and development.
opportunities" (World Tourism Organization, 2016). To capture and study the effects of terrorism on tourism, there is the need of gathering suitable attributes that consistently allow portraying the behavior of travelers in response to the terroristic events previously described. This section presents the data collected regarding the activity of the hospitality industries from the various countries under the scope of the analysis. Here the collected figures are described regarding their conceptual meaning in the context for which the collection processes were undertaken, presenting the strengths and limitations of each dataset to use.

2.2.1 Origin and Purpose

For studying the impact of terrorism in the hospitality industry at a worldwide level, it is relevant to collect macroeconomic and industry data that are suitable for the segmentation purpose of this research project, profiling each country regarding the contribution of its hospitality industry for the respective economy and in terms of how does terrorism affects such contribution.

Tourism is formally perceived as a directly quantifiable activity, allowing more precise examination and an adequate policy. While beforehand the hospitality industry depended mostly on tourism-specific measurements estimates (e.g. Balance of Payments measurements), tourism today employs a scope of instruments to record its commercial activities, tracking the activities of both kinds of visitors, i.e., tourists and excursionists. An expanding number of nations have embraced tourism, assigning resources for its development, making tourism a revenue key driver for socio-economic growth, developing investment, employment and infrastructures. As a globally exchanged service, inbound tourism became one of the world's major trade industries, with special emphasis for some developing countries which invest in tourism as a primary source of foreign exchange income, having a major impact in their exports and thus in the creation of wealth and employment.

The data here presented are retrieved from the World Bank Open Data repository. Such data are gathered by UNWTO upon collection by the competent authorities of each country, following the guidelines and recommendations mentioned in section 0, and fully detailed in Appendix A. The below Table 4 presents the list of indicators that were selected as relevant for the intended purpose of this thesis. Note that other indicators might also be used although, given several limitations regarding data completion, excessive granularity and the non-structured nature of data represent a barrier for such inclusion.

<table>
<thead>
<tr>
<th>Indicator Code</th>
<th>Indicator Name</th>
<th>Indicator Short Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST_INT_ARVL</td>
<td>Number of International Tourist Arrivals (1)</td>
<td>Arrivals</td>
</tr>
<tr>
<td>ST_INT_RCPT_CD</td>
<td>Receipts in current USD (1)</td>
<td>Receipts</td>
</tr>
<tr>
<td>NY_GDP_MKTP_CD</td>
<td>Gross Domestic Product in current USD (2)(3)</td>
<td>GDP</td>
</tr>
</tbody>
</table>

2.2.2 Data collection: Methodology and limitations

Statistical information on tourism depends essentially on figures from arrivals and overnight stays, together with the balance of payments figures, not being adequate by themselves for public policy making or industry operational management due to the insufficient capacity to illustrate the whole economic reality of the hospitality industry. Although the continuous improvement in the harmonization of definitions and measurement processes reported by UNWTO, there are still countries missing cohesion to enable full comparability with the generality of the reporting countries.

2.2.2.1 Arrivals

As recommended in UNWTO guidelines for tourism statistics, the “data on inbound tourists refer to the number of arrivals, not to the number of people traveling, thus a person who makes several trips to a country during a given period is counted each time as a new arrival”. Data are collected combining a mixture of a manifold kind of sources: 1) administrative records, including immigration services and traffic controls; 2) border surveys, and; 3) accommodation surveys for which the number of guests corresponds to an estimate of the arrivals number. This indicator ideally refers to the arrivals of overnight visitors at national borders. A set of major identified limitations arising from the existing differences in the data collection methodologies are:

- for the cases that such data are inaccessible or insufficient, data from international visitors are shown, comprising figures from tourists, same-day visitors, cruise passengers and crew members – other types of travelers (see Figure 3) should not be included in this measurement as they do not qualify as visitors;
- some countries limit their reports to arrivals by air;
- others only refer to arrivals of visitors staying in hotels;
- some countries report arrivals of nationals residing abroad while others do not.

2.2.2.2 Receipts (Current US$)

Receipts captures the economic phenomenon of inbound tourism, representing the total expenditure linked to the activity of inbound visitors. Such expenditures are conventionally recorded as inflows (or credit movements) in the travel items of the Balance of Payments (hereinafter: BoP) from the International Monetary Fund (hereinafter: IMF) under the “travel receipts” label. In accordance with the UNWTO guidelines for tourism statistics, the transportation of passengers must be embodied in the measurement of tourism industries and respective products. Therefore, the ideal estimate of the expenditures associated with the inbound tourism activity would be the sum of two BoP items: the expenditure value for the travel plus the expenditure value of the passenger transport. A set of major identified limitations arising from the existing differences in the data collection methodologies are:
• for the cases that IMF is unable to capture data for the expenditure value of the passenger transport, only the expenditure value for the travel is shown;
• BoP estimates comprise expenditures associated with other types of travelers which might lead to biased comparison for some countries given the substantial number of visitors other than overnight visitors (e.g., same-day visitors, border and seasonal workers, long-term students or patients, etc.); although, this situation is generally tackled “when these [expenditures] are important enough to justify separate classification”;
• the aggregates are calculated using the World Bank’s weighted aggregation methodology which differ from the UNWTO’s aggregates.

2.2.2.3 GDP

Gross Domestic Product (hereinafter: GDP) comprises the aggregation of the value added generated by all the producing activity of an economy. For a better understanding of this indicator, it is important to break down the concept of value added as “the gross output of producers less the value of intermediate goods and services consumed in production, before accounting for the consumption of fixed capital in production” (The World Bank, 2016). Value added (and thus GDP, as it is a value added aggregation) may be valuated at one of the following valuation methods:

• basic prices, i.e. the “amount receivable by the producer, exclusive of taxes payable on products and inclusive of subsidies receivable on products”;
• producer prices, i.e. the “amount receivable by the producer inclusive of taxes on products except deductible value added tax and exclusive of subsidies on products”;
• purchaser prices, i.e. the “amount payable by the purchaser, which includes trade margins realized by wholesalers and retailers (by definition, their output) as well as transport margins (that is, any transport charges paid separately by the purchaser) and non-deductible Value Added Tax (VAT).”

The major limitations of this indicator are concerned with the multiplicity of possible valuation methods for the value added, seizing full comparability over the data from different countries, from which are relevant to remark:

• the System of National Accounts (SNA) of the United Nations recommends the value added to be valuated either at basic prices or producer prices, given evidence that “GDP estimates based on the production approach are generally more reliable than estimates compiled from the income or expenditure side” (The World Bank, 2016); nonetheless, different countries have opted for different valuation standards, the reason why there is the need to identify a group of homogeneously reporting countries, as done for the previous indicators;
• both basic prices and producer prices valuation methods exclude transport expenditures that are invoiced solely by producers;
• value added by industry is commonly measured at basic prices;
• the national accounts compiling process is affected by a set of inherent insufficiencies (usually more preeminent in developing countries) constituting serious obstacles to the production of reliable and comprehensive series of national accounts statistics, most commonly:
  o the lack of human resources (personnel, training, etc.);
  o the magnitude of unreported economic activity associated with the informal sector of the economy, representing a misevaluation of a country value added and thus its GDP.

Notwithstanding, GDP may not always be the ideal indicator to summarize the aggregated economic performance (particularly for the cases the productive process happens at the expense fixed capital consumption), and thus not ideal for assessing the contribution of tourism for such performance. A more suitable indicator for measuring the contribution of tourism for an economy GDP would be the Tourism Direct Gross Domestic Product (hereinafter: TDGDP), statistically defined as “the part of GDP attributable directly to domestic and inbound tourism” (Dupeyras & MacCallum, 2013). Nevertheless, given the unavailability of TDGDP data (only collected for some OECD countries and for a respectively small period range), the aforementioned indicators Receipts and GDP are used instead (in a later stage of this study) regardless of their inherent limitations, in order to derive an indicator of closer accuracy for measuring the contribution of tourism for each country economy.

2.2.3 Metadata and Data Cleansing Codebook

Given the aforementioned data collection inconsistencies at the country-level plus the lack of data completeness for some countries and periods, there is the need of carrying an inspective analysis of the collection criteria for each indicator. This subsection is meant to define the geographical and temporal scope of the data to be used, discarding both the countries and periods for which the comparison is not possible due to the nature of their underlying limitations. This data cleansing procedure consists of three main stages:

• identify, for each of the selected World Development Indicators, the countries that permit consistent comparison given their uniform (or similar) collection criteria, logically including specific exceptions that induce minor bias;
• cross-matching the identified countries for each indicator, to define a common geographic framework provided with the ability for generalizing further conclusions;
• identify major issues regarding data completeness, excluding both the countries and periods for which those issues present severe magnitude.

The proper identification of the inclusion criteria, given the definitions of tourism (see 0) and the metadata labels contained in the original data, is the fundamental milestones for the success of this
To minimize the impact of possible flaws arising from data collection inconsistencies, a conservative approach is used, defining the inclusion criteria as stated in Table 5.

**Table 5. World Development Indicators: Data Cleansing**

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Inclusion Criteria</th>
<th>Included Countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arrivals</td>
<td>Refers to arrivals of non-resident tourists at national borders</td>
<td>128</td>
</tr>
<tr>
<td></td>
<td>Refers to arrivals of non-resident tourists in all types of accommodation establishments</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Refers to arrivals of non-resident tourists in hotels and similar establishments</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Subtotal</td>
<td>155</td>
</tr>
<tr>
<td>Receipts</td>
<td>Compiled from data reported in the IMF’s BoP statistics database.</td>
<td>159</td>
</tr>
<tr>
<td>GDP</td>
<td>Value added at basic prices (VAB)</td>
<td>170</td>
</tr>
</tbody>
</table>

The country inclusion criteria for the *Arrivals* data considers three major groups of countries: the first group accounts to countries where the arrivals are collected at national borders, while for the remaining groups this collection is done by accommodation establishments of different kinds. Although each of these collection methods produces different figures, they are all considered given the main focus of studying the economic impact of terrorism in the hospitality industry, i.e. arrivals of tourists that do not stay in accommodation establishments are assumed to have a minor impact in a country hospitality industry. Regarding the *Receipts* indicator, and because originally several countries only collect data regarding the expenditure value for the travel, ignoring the expenditure value of the passenger transport, the selected countries are those who perform a full collection of these data, reporting it under the IMF’s BoP standards. The *GDP* indicator presents smaller granularity regarding the selection of inclusion criteria, having been selected only those countries for which the reported data is consistent with value added at basic prices. As a result of the above-mentioned data cleansing process, a set of 74 countries are selected within the scope of the analysis, for a temporal period of 20 years (1995-2014).

### 2.3 Unifying Data Sources

The purpose of this section is to summarize the main aspects of combining the previously presented kinds of data in a single dataset for further use. Additionally, the impact of the dataset merging process is assessed in terms of the quality of the data regarding generalization ability. Given the data cleansing efforts employed, the resulting terrorism dataset is composed of 20 attributes considered to be relevant while performing the exploratory analysis of the data. Such attributes describe 142,686 incidents occurred within a pool of 204 countries from 1970 to 2014. Albeit this dataset enables the worldwide description of the terrorism phenomenon, this does not verify when combined with the existent tourism data. In fact, given the limitations above regarding the collection methodology used for tourism data result in a batch of 3 attributes, recorded over the 1995-2014 period for 74 countries.
Table 6. Merged Dataset: Countries/Region Breakdown

<table>
<thead>
<tr>
<th>Region</th>
<th>Code</th>
<th>Country</th>
<th>Region</th>
<th>Code</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>North &amp; South America</td>
<td>CAN</td>
<td>Canada</td>
<td>Western Europe</td>
<td>AUT</td>
<td>Austria</td>
</tr>
<tr>
<td></td>
<td>USA</td>
<td>United States of America</td>
<td>BEL</td>
<td>Belgium</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BHS</td>
<td>Bahamas</td>
<td></td>
<td>DEU</td>
<td>Germany</td>
</tr>
<tr>
<td></td>
<td>BLZ</td>
<td>Belize</td>
<td></td>
<td>ESP</td>
<td>Spain</td>
</tr>
<tr>
<td></td>
<td>CRI</td>
<td>Costa Rica</td>
<td></td>
<td>FRA</td>
<td>France</td>
</tr>
<tr>
<td></td>
<td>HND</td>
<td>Honduras</td>
<td></td>
<td>GBR</td>
<td>United Kingdom</td>
</tr>
<tr>
<td></td>
<td>NIC</td>
<td>Nicaragua</td>
<td></td>
<td>GRC</td>
<td>Greece</td>
</tr>
<tr>
<td></td>
<td>PAN</td>
<td>Panama</td>
<td></td>
<td>IRL</td>
<td>Ireland</td>
</tr>
<tr>
<td></td>
<td>SLV</td>
<td>El Salvador</td>
<td></td>
<td>ISL</td>
<td>Iceland</td>
</tr>
<tr>
<td></td>
<td>ARG</td>
<td>Argentina</td>
<td></td>
<td>ITA</td>
<td>Italy</td>
</tr>
<tr>
<td></td>
<td>BOL</td>
<td>Bolivia</td>
<td></td>
<td>LUX</td>
<td>Luxembourg</td>
</tr>
<tr>
<td></td>
<td>BRA</td>
<td>Brazil</td>
<td></td>
<td>NLD</td>
<td>Netherlands</td>
</tr>
<tr>
<td></td>
<td>CHL</td>
<td>Chile</td>
<td></td>
<td>NOR</td>
<td>Norway</td>
</tr>
<tr>
<td></td>
<td>COL</td>
<td>Colombia</td>
<td></td>
<td>PRT</td>
<td>Portugal</td>
</tr>
<tr>
<td></td>
<td>ECU</td>
<td>Ecuador</td>
<td></td>
<td>SWE</td>
<td>Sweden</td>
</tr>
<tr>
<td></td>
<td>GUY</td>
<td>Guyana</td>
<td></td>
<td>BGR</td>
<td>Bulgaria</td>
</tr>
<tr>
<td>Southeast Asia</td>
<td>PER</td>
<td>Peru</td>
<td></td>
<td>EST</td>
<td>Estonia</td>
</tr>
<tr>
<td></td>
<td>PRY</td>
<td>Paraguay</td>
<td></td>
<td>HRV</td>
<td>Croatia</td>
</tr>
<tr>
<td></td>
<td>SUR</td>
<td>Suriname</td>
<td></td>
<td>LTU</td>
<td>Lithuania</td>
</tr>
<tr>
<td></td>
<td>URY</td>
<td>Uruguay</td>
<td></td>
<td>LVA</td>
<td>Latvia</td>
</tr>
<tr>
<td></td>
<td>CHN</td>
<td>China</td>
<td></td>
<td>MKD</td>
<td>Macedonia</td>
</tr>
<tr>
<td></td>
<td>JPN</td>
<td>Japan</td>
<td></td>
<td>POL</td>
<td>Poland</td>
</tr>
<tr>
<td></td>
<td>IDN</td>
<td>Indonesia</td>
<td></td>
<td>RUS</td>
<td>Russia</td>
</tr>
<tr>
<td></td>
<td>LAO</td>
<td>Laos</td>
<td></td>
<td>SVN</td>
<td>Slovenia</td>
</tr>
<tr>
<td></td>
<td>MYS</td>
<td>Malaysia</td>
<td></td>
<td>UKR</td>
<td>Ukraine</td>
</tr>
<tr>
<td>South Asia</td>
<td>THA</td>
<td>Thailand</td>
<td>Middle East &amp; North Africa</td>
<td>BHR</td>
<td>Bahrain</td>
</tr>
<tr>
<td></td>
<td>BGD</td>
<td>Bangladesh</td>
<td></td>
<td>EGY</td>
<td>Egypt</td>
</tr>
<tr>
<td></td>
<td>BTN</td>
<td>Bhutan</td>
<td></td>
<td>ISR</td>
<td>Israel</td>
</tr>
<tr>
<td></td>
<td>IND</td>
<td>India</td>
<td></td>
<td>JOR</td>
<td>Jordan</td>
</tr>
<tr>
<td></td>
<td>LKA</td>
<td>Sri Lanka</td>
<td></td>
<td>LBN</td>
<td>Lebanon</td>
</tr>
<tr>
<td></td>
<td>MUS</td>
<td>Mauritius</td>
<td></td>
<td>TUN</td>
<td>Tunisia</td>
</tr>
<tr>
<td>Australasia &amp; Oceania</td>
<td>NPL</td>
<td>Nepal</td>
<td>Sub-Saharan Africa</td>
<td>AGO</td>
<td>Angola</td>
</tr>
<tr>
<td></td>
<td>ARM</td>
<td>Armenia</td>
<td></td>
<td>KEN</td>
<td>Kenya</td>
</tr>
<tr>
<td></td>
<td>KGZ</td>
<td>Kyrgyzstan</td>
<td></td>
<td>LSO</td>
<td>Lesotho</td>
</tr>
<tr>
<td></td>
<td>AUS</td>
<td>Australia</td>
<td></td>
<td>SYC</td>
<td>Seychelles</td>
</tr>
<tr>
<td></td>
<td>PNG</td>
<td>Papua New Guinea</td>
<td></td>
<td>TZA</td>
<td>Tanzania</td>
</tr>
<tr>
<td></td>
<td>VUT</td>
<td>Vanuatu</td>
<td></td>
<td>UGA</td>
<td>Uganda</td>
</tr>
</tbody>
</table>

Source: Adapted after Global Terrorism Database, Codebook: Inclusion Criteria and Variables, 2017.
As a result of merging the two-fold kinds of data there is an unavoidably inherent loss of data: indeed, the terrorism data is constrained by the narrower extension of the tourism data both in terms of the countries recorded and the time-frame represented. That being said, the merged dataset encompasses only the observations concerning to the countries and years for which all the mentioned aggregate terrorism and tourism attributes are fully available. Those countries are listed in Table 6 (above), organized per geographic region. In a first stage, the analysis covers the entire terrorism dataset length, i.e. all the records resulting from the cleansing employed to GTD original data. The latest stage of this analysis solely covers the records considered after both datasets are merged, which is limited to only 74 out of an initial pool of 204 countries. Table 7 summarizes the number of occurred incidents, successful incidents, casualties and injured, aggregated per geographic region for the early terrorism dataset. Such metrics are considered to be milestones to describe terrorism incidents and thus the terrorism phenomenon in quantitative terms.

### Table 7. Terrorism Geographical Distribution: Region Aggregates (1970-2014)

<table>
<thead>
<tr>
<th>Region</th>
<th>Incidents</th>
<th></th>
<th>Total Casualties</th>
<th></th>
<th>Total Injured</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Successful</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>#</td>
<td>%</td>
<td>#</td>
<td>%</td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>Australasia &amp; Oceania</td>
<td>241</td>
<td>0.17</td>
<td>209</td>
<td>0.16</td>
<td>159</td>
<td>0.05</td>
</tr>
<tr>
<td>Central America &amp; Caribbean</td>
<td>10,345</td>
<td>7.25</td>
<td>9,980</td>
<td>7.69</td>
<td>28,700</td>
<td>9.17</td>
</tr>
<tr>
<td>Central Asia</td>
<td>572</td>
<td>0.40</td>
<td>517</td>
<td>0.40</td>
<td>1,026</td>
<td>0.33</td>
</tr>
<tr>
<td>East Asia</td>
<td>762</td>
<td>0.53</td>
<td>642</td>
<td>0.49</td>
<td>1,001</td>
<td>0.32</td>
</tr>
<tr>
<td>Eastern Europe</td>
<td>4,134</td>
<td>2.90</td>
<td>3,538</td>
<td>2.73</td>
<td>6,366</td>
<td>2.03</td>
</tr>
<tr>
<td>Middle East &amp; North Africa</td>
<td>34,668</td>
<td>24.30</td>
<td>31,510</td>
<td>24.29</td>
<td>89,001</td>
<td>28.43</td>
</tr>
<tr>
<td>North America</td>
<td>3,230</td>
<td>2.26</td>
<td>2,697</td>
<td>2.08</td>
<td>4,616</td>
<td>1.47</td>
</tr>
<tr>
<td>South America</td>
<td>18,535</td>
<td>12.99</td>
<td>17,238</td>
<td>13.29</td>
<td>28,797</td>
<td>9.20</td>
</tr>
<tr>
<td>South Asia</td>
<td>33,331</td>
<td>23.36</td>
<td>30,247</td>
<td>23.32</td>
<td>78,928</td>
<td>25.21</td>
</tr>
<tr>
<td>Southeast Asia</td>
<td>9,356</td>
<td>6.56</td>
<td>8,563</td>
<td>6.60</td>
<td>13,775</td>
<td>4.40</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>11,548</td>
<td>8.09</td>
<td>10,878</td>
<td>8.39</td>
<td>54,358</td>
<td>17.36</td>
</tr>
<tr>
<td>Western Europe</td>
<td>15,964</td>
<td>11.19</td>
<td>13,707</td>
<td>10.57</td>
<td>6,318</td>
<td>2.02</td>
</tr>
</tbody>
</table>

According to the above table, the regions which have been suffering from more intense terroristic activity during the whole 1970-2014 period length are Middle East & North Africa, South Asia and South America representing together around 60 percent of the worldwide attacks, with similar aggregate success rate and combining nearly 62 percent of casualties and 72 percent of resulting injured worldwide. When looking over the European countries, here labeled as part of Eastern Europe or Western Europe, the total number of incidents correspond to only 14 percent of the worldwide attacks and presenting similar aggregate success rate. Nonetheless, and in opposition to the previously analyzed geographies, the reported attacks solely account for 4 percent of casualties and 7 percent of resulting injured worldwide. Additionally, Figure 6 (below) enables the visual understanding of the global spread of the terrorist events captured in the merged dataset.
On the contrary, Figure 7 depicts the geographical distribution of the attacks excluded as a result of the dataset merging process (these attacks occurred either outside of the inclusion period, i.e., before 1995, or took place in countries that were excluded from the analysis). Note that some incidents location may overlap according to the precision used at the data collection stage when storing incident coordinates. Furthermore, part of this overlapping result from the missing data imputation process that generates coordinates from the existing information regarding the administrative divisions where the incidents were perpetrated.
Drilling down into the list of 74 countries for which both terrorism and tourism data are available results in a severe but unavoidable loss of data, both in the geographical and temporal dimensions. For this reason (and to serve as a matter of comparison between datasets), Table 8 summarizes the most describing and relevant metrics aggregated at the region level (likewise Table 7 does for the former dataset).

Table 8. Terrorism Geographic Distribution: Region Aggregates (1995-2014)

<table>
<thead>
<tr>
<th>Region</th>
<th>Incidents</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Successful</td>
<td>Total Casualties</td>
<td>Total Injured</td>
<td>#</td>
<td>%</td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>Australasia &amp; Oceania</td>
<td>64</td>
<td>0.23</td>
<td>58</td>
<td>0.24</td>
<td>33</td>
<td>0.07</td>
<td>48</td>
<td>0.05</td>
</tr>
<tr>
<td>Central America &amp; Caribbean</td>
<td>165</td>
<td>0.60</td>
<td>150</td>
<td>0.61</td>
<td>150</td>
<td>0.31</td>
<td>233</td>
<td>0.25</td>
</tr>
<tr>
<td>Central Asia</td>
<td>39</td>
<td>0.14</td>
<td>31</td>
<td>0.13</td>
<td>22</td>
<td>0.05</td>
<td>42</td>
<td>0.05</td>
</tr>
<tr>
<td>East Asia</td>
<td>267</td>
<td>0.96</td>
<td>229</td>
<td>0.94</td>
<td>845</td>
<td>1.75</td>
<td>7,828</td>
<td>8.56</td>
</tr>
<tr>
<td>Eastern Europe</td>
<td>3,160</td>
<td>11.41</td>
<td>2,692</td>
<td>11.03</td>
<td>5,631</td>
<td>11.68</td>
<td>8,711</td>
<td>9.52</td>
</tr>
<tr>
<td>North America</td>
<td>551</td>
<td>1.99</td>
<td>426</td>
<td>1.75</td>
<td>3,271</td>
<td>6.79</td>
<td>1,480</td>
<td>1.62</td>
</tr>
<tr>
<td>South America</td>
<td>3,678</td>
<td>13.28</td>
<td>3,377</td>
<td>13.84</td>
<td>6,390</td>
<td>13.26</td>
<td>6,623</td>
<td>7.24</td>
</tr>
<tr>
<td>South Asia</td>
<td>9,639</td>
<td>34.80</td>
<td>8,513</td>
<td>34.89</td>
<td>19,683</td>
<td>40.84</td>
<td>36,468</td>
<td>39.86</td>
</tr>
<tr>
<td>Southeast Asia</td>
<td>3,607</td>
<td>13.02</td>
<td>3,370</td>
<td>13.81</td>
<td>3,183</td>
<td>6.60</td>
<td>8,357</td>
<td>9.13</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>813</td>
<td>2.94</td>
<td>758</td>
<td>3.11</td>
<td>5,212</td>
<td>10.81</td>
<td>7,750</td>
<td>8.47</td>
</tr>
<tr>
<td>Western Europe</td>
<td>2,758</td>
<td>9.96</td>
<td>2,324</td>
<td>9.52</td>
<td>506</td>
<td>1.05</td>
<td>3,513</td>
<td>3.84</td>
</tr>
</tbody>
</table>

When comparing Table 7 and Table 8 it is possible to derive some conclusions arising from the aforementioned data reduction:

- Middle East and North Africa accounts for only 11 per cent of the terrorist incidents of the dataset (against the 24 per cent before the data reduction); unavoidably, the distribution of successful incidents varies similarly; the distribution of casualties and injuries significantly shrank (from 28 and 39 percent to 7 and 11 percent, respectively), which dues to the typically high mortality rate of the attacks occurring in these countries;

- on the other side, European countries gain more relative importance after the reduction of the dataset, encompassing now 21 per cent of the incidents (against the former 14 per cent), even having nearly half the number of countries subtracted.

When aiming attention at the previously drawn conclusions, it is also important to remark that, despite the variations caused by the loss of the 1970-1994 data, the attributes relative frequencies also vary at the region level given country shortage implied by the dataset merging process. In fact, and using Middle East & North Africa as an example, it is possible to question the ability of the latest dataset to represent the terrorism phenomenon for that region, including only 6 countries out of the 22 originally contained in the former dataset. Notwithstanding the data, reduction is not
proportional between countries (nor regions), some might be the ones who suffer from more severe impacts, losing the ability to realistically represent the respective region as a whole. Figure 8 (below) makes a clear distinction between the number of countries included after the dataset merging process for each region.

*Figure 8. The Impact of the Dataset Merging Process: Countries per Region*

Through the observation of Figure 8 is perceptible that no region remains unaffected by the aforementioned exclusion of countries. The regions for which such shrinkage represent smaller representativeness are respectively Sub-Saharan Africa, Central Asia and the Middle East & North Africa. On the other hand, South America, Western Europe, and Eastern Europe are respectively those who suffer from smaller impacts in terms of country representativeness.

### 3 EXPLORATORY ANALYSIS

Terrorism and tourism are both dynamic phenomena in terms of space and time: both present distinct characteristics throughout the various geographies of the globe, shaping each culture, country and/or the economy in a distinct manner and thus impacting them differently. Similarly, to an economy as a whole, each of its constituent industries also suffers the effects of the terrorist externalities in distinct ways: while some industries are more likely to lose with terrorism, others may see an opportunity on it. In what regards to the hospitality industry, the substitution effect arising from changes in travel destinations, accommodation or means of travel represent both gains and losses for opposing players. The present section is dedicated to the descriptive exploratory analysis of the previously merged dataset, having been employed several cleansing procedures in order to resolve data inconsistencies allowing the extraction of enhanced insights. The latest product
of the exploratory analysis culminates into a cluster analysis in order to gather country observations together into groups of similar behavior regarding tourism and terrorism performance, summarizing the profile of those countries in analysis and enabling the generalization of further conclusions.

Given both the temporal and sectional dynamic structure of the dataset in use, some difficulties or challenges apply when formalizing the clustering problem. In fact, considering both the temporal and sectional dimensions at the same time might produce results of little interest and for which the interpretation is not straightforward, at a huge implementation cost. On the other hand, ignoring any of those dimensions might result in a significant loss of complexity. A clear example of this trade-off between complexity and implementation cost is the loss of understanding of time-variant behaviors for clustered individuals (or countries) when ignoring the temporal dimension. To assess whether applying a simplistic approach produces satisfactory solutions for the main purpose of this research work, several approaches are separately carried out. The various approaches differ regarding the length and complexity of the data (in size and depth), the clustering procedure, the involved algorithms and associated assumptions. Results are then presented, and compared against each other in seeking for the methods that provide more meaningful cluster solutions. The clustering approaches considered are listed down as follows:

- **Country-aggregated cluster analysis** – the data is aggregated and summarized at the country level through the use of the arithmetic mean of each feature for the dataset period as its own representative. This is the simplest approach from the implementation point of view, yet theoretically efficient when assuming countries behave similarly over time regarding terrorism and tourism; albeit this assumption does not fully apply, the results from performing this approach might be used as aprioristic insights when fine-tuning the further approaches;
- **“Pooled” cluster analysis** – this approach does not consider each country as a section observer for each time unit in the period; in fact, each “country-year” pair is included as an independent individual or section. Although the results are likely to include several observations of a country for different years within the same cluster, this is an easy implementation to identify shifts of countries between hypothetical clusters and thus have a better understanding of such variations;
- **Longitudinal clustering analysis** – considering both the temporal and sectional dimensions of the data, using longitudinal data enables the identification of different clusters of countries according to their trajectory shapes.

### 3.1 Feature Selection

Given the considerations regarding the formalization of the clustering problem made in section 0, the first step to address is the selection of features to include as relevant variables to produce satisfactory solutions. To have an overview of which features must be accounted, an exploration of each of the collected attributes is conducted. Appendix C depicts both the boxplot diagrams (grouped at the country level and sorted per increasing order of variability) and the histogram of the
respective attribute average value per country. An evident initial remark regarding tourism related attributes is that most countries present small variability during the whole length of the sample period. Such small variability for the majority of the countries can be interpreted as an evidence of slight variations of the economic, and especially, the hospitality industry scenario. On the contrary, few are the countries for which there is a significant fluctuation of each attribute value (noticeable on the right-hand side of each boxplot). Histograms plots are helpful to perceive how different groups of countries are binned together regarding each attribute, showing a high concentration of countries on values smaller than the respective attribute average. Furthermore, a few countries may be classified as outliers given the magnitude of their distance from the remainder of the sample. The most noteworthy examples are:

- France, the country presenting a significant dominance of the international tourist arrivals for the generality of the period;
- the United States of America could easily be identified as a single-element cluster regarding the international tourism receipts; additionally, it is possible to understand the presence of 3 to 4 main hypothetic groups for this attribute;
- similar to what happens with the international tourism receipts, the United States of America lead the gross domestic product statistics, although this time followed by consistent runner-up’s such as Japan and China. The number of clusters to identify regarding the gross domestic period is similar as for the international tourism receipts.

Carrying the analogous analysis of boxplot diagrams and histograms for terrorism attributes, similar conclusions can be drawn. As expected, and for the majority of the countries, values close to zero for this class of attributes are more frequent as a natural result of the inclusion of countries where terrorism is not a common everyday practice. Additionally, most of those countries present a relatively small variability over the years for each attribute, with few exceptions happening in countries facing periods of more intense terroristic activity, being hard to identify an aprioristic number of existing clusters, since this number varies for each variable.

When addressing a clustering problem, some authors discourage the inclusion of strongly correlated variables (usually those presenting absolute correlations above 0.90) as they are usually problematic. In the limit situation of having two collinear variables (i.e., perfectly correlated, \( \rho = \pm 1 \), both of them describe the same concept or feature. That said, including highly correlated variables incur in overweighting the same concept or feature. According to Sambandam (2007), when multicollinear variables are used to produce a solution for the clustering problem, that “solution is likely to be skewed in the direction of that concept, which could be a problem if not anticipated”. The analogous reasoning applies for multiple correlated variables (and thus multicollinearity), having the overweighting effect applicable towards an unknown number of concepts or features. Aiming attention at the previous considerations, it is reasonable to inspect variables regarding their correlation prior to study their individual behavior. This analysis focus only over cross-correlation
Correlation between variables is presented in Table 9 (above), which allows the detection of the following major noteworthy relationships:

- **Total Attacks** is virtually collinear (0.9953) with **Successful Attacks**, reason why, the variable accounting for **Total Attacks** is dropped to the detriment of the newly created **Failed Attacks** (computed as the difference between **Total Attacks** and **Successful Attacks**) which is believed to represent a non-repeated feature and does not imply any significant problematic correlation with other variables;

- **Receipts** is strongly correlated (0.8849) with **GDP**; this relationship is expectable since tourism is for many countries one of the industries that contributes the most to the respective GDP. For this reason, the ratio between **Receipts** and **GDP** is computed to provide a more explanatory measure of the importance of revenues arising from tourism inflows in the whole economy of a country. Such ratio, hereinafter **Tourism Contribution in GDP**, is expected to vary with **Arrivals** (included as a time and country-variant regressor) along with fluctuations in the variables that describe the terroristic activity for each country and period.

The aforementioned changes applied to variables tackle the major existing problems of quasi-collinearity (or strong correlation) as it is possible to verify in Table 9 (bottom table). Repeating the analogous analysis regarding the variability of the newly generated variables (see bottom plots in Table 9).
it is notorious that the variability of *Tourism Contribution to GDP* is more widely distributed for the generality of the countries. Regarding *Failed Attacks*, there is a slightly wider spread of countries over values other than zero, when compared with *Total Attacks*: this is expected to have positive effects in the clustering results since more countries are differently profiled by this feature.

### 3.2 Cluster Analysis

This section encompasses the procedures carried for each of the previously purposed clustering approaches, describing their inherent assumptions, algorithms employed and presenting the respective resulting solutions.

#### 3.2.1 Country-aggregated Cluster Analysis

The present approach is the simplest among the set of tested alternatives, yet providing an easy implementation and straightforward interpretation of the generated solutions. The clustering procedure is built on the premise of analyzing countries from a general perspective, obtaining a wider picture of their tourism and terrorism panoramas over the 1995-2014 period. For that reason, data are aggregated at the country level, resulting in the average value of each of the collected attributes for the whole period length. Before engaging in the selection of an adequate clustering algorithm, there is important to bring into the spotlight the fact that aggregating the data at the country level produces a much more compact dataset (for the case, containing only one single record per country). As an underlying consequence of the aggregation of the dataset, the correlation between variables is “aggravated” as described by Clark and Avery (2010) having possibly noxious effects in the clustering results. Such issue can be addressed from two main perspectives. The former consists in dropping the variables which are verified to be problematic. This the simplest approach, requiring low computational effort and allowing for the use of a wide range of clustering algorithms; on the other hand, dropping highly correlated variables implies a loss of complexity regarding the descriptive potential of those variables, and thus that may lead to fruitless solutions. A flagrant case hereby identified is the strong correlation between the newly created variable *Failed Attacks* and several other variables describing the terroristic activity. Notice that, although this variable had been generated to tackle the excessive correlation between *Total Attacks* and *Successful Attacks* (both variables are quite similar), *Failed Attacks* describes a complete opposite feature than the one captured by *Successful Attacks*: in fact, the intention of including such variable is to assess whether tourism inflows is influenced by an effective counter-terrorist law enforcement or not. The latest perspective dwells in the application of most conservative options from the perspective of the quality of the data, allowing for the use of strongly correlated variables. The application of alternative clustering methods and algorithms, as well as the use of more appropriate distance measures, may constitute valid alternatives to tackle the limitations imposed by the high correlated...
variables. One example is the use of the Mahalanobis square distance instead of the classic Euclidean distance.

As previously stated various clustering methods and algorithms are tested to achieve a set of comparable solutions, for which the most adequate is chosen according to certain criteria. Those methods are briefly introduced in section 0 (see Section 1). The fact that the some of the available variables are not directly comparable is a key aspect to consider before undertaking any further clustering task. In fact, the diversity of measurement units employed given the different nature of each of the collected features may lead to undesirable clustering solutions where features (or variables) of higher magnitude dominate the clustering process given the superior distances computed to be decisive in the cluster formation. The use of data normalization techniques, commonly used to prevent such constraints, frequently relies on the application of one of the two most frequently used techniques:

- **Z-score standardization** rescales each variable into a normally (standard) distributed variable. This operation is formally defined as

\[
Z(x_{ij}) = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \sim \mathcal{N}(0,1)
\]  

(12)

where \(\bar{x}_j\) and \(\sigma_j\) are respectively the sample mean and standard deviation of the \(j^{th}\) variable. Regardless of its proven efficiency tackling the use of different measurement units, one of the major drawbacks of this technique is the “loss of location and scale information of the original variable” (Johor Bahru et al., 2013);

- **Min-Max normalization** rescales each observed value of a variable into the equivalent value within the \([0;1]\) range, having each variable lowest value represented by 0 and the highest represent by 1, and thus “provides an easy way to compare values that are measured using different scales or different units of measure”. As a result of setting equal bounding values across variables, any computed standard deviations are relatively smaller (when compared to Z-score application) “muzzling” the contrasting effect of existing outliers. Min-Max normalization is formally defined as

\[
MM(x_{ij}) = \frac{x_{ij} - x_{j_{min}}}{x_{j_{max}} - x_{j_{min}}}
\]  

(13)

where \(x_{j_{min}}\) and \(x_{j_{max}}\) are the corresponding minimum and maximum observed values of the \(j^{th}\) variable.
Given the previous considerations regarding each of the data normalization techniques, and to derive simpler comparison between variables, Min-Max normalization is chosen in the application for the specific purpose of this exploratory analysis. Noting the potential difficulties arising from the existence of few highly correlated variables, the use of the Euclidean distance might produce unsatisfactory solutions. For that reason, two additional distance measures are employed enabling further comparison of the results. That being said, the distance measures considered are as follows:

- the Euclidean (or $L^2$) distance, commonly described as the distance between two observations in Euclidean space, is defined for $n$ dimensions as

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2};$$  \hspace{1cm} (14)

- the Mahalanobis distance has its main application in “classification problems where there are several groups and the investigation concerns the affinities between groups” with the ultimate goal of forming “clusters of members who are similar to each other, perhaps in a hierarchical scheme” (McLachlan, 1999). It can be defined in the context of an $n$-dimensional space as

$$d_M(\bar{p}, \bar{q}) = \sqrt{(\bar{p} - \bar{q})^T S^{-1} (\bar{p} - \bar{q})}$$  \hspace{1cm} (15)

where $\bar{p} = (p_1, p_2, \ldots, p_n)$ and $\bar{q} = (q_1, q_2, \ldots, q_n)$ have the same distribution with covariance matrix $S$;

- the Manhattan distance (also known as the City-Block or Taxicab distance) between two points, is the sum of the absolute differences of their cartesian coordinates. It can be formally defined in the $n$-dimensional space as

$$d_1 (p, q) = ||p - q||_1 = |p_1 - q_1| + |p_2 - q_2| + \cdots + |p_n - q_n| = \sum_{i=1}^{n} |p_i - q_i|. \hspace{1cm} (16)$$

Each of the aforementioned distance measures, is implemented for every agglomeration method as described in 0, resulting in a set of 12 hierarchical clustering solutions. Each solution is illustrated in the form of a dendrogram plot, for which the most notable solutions are highlighted according to a set of cluster validity criteria as described in 0. Table 10 summarizes the information provided by each dendrogram and respective validity indices plot (these are presented in further detail in Appendices D, E, F, and G), highlighting the preferred solutions according to the optimal cluster number purposed for each of the considered validity criteria. It is important to notice that the investigation for the optimal number of clusters is performed between 3 and 8 clusters. Such boundaries are set to avoid solutions of non-practical application regarding the purpose of this thesis, achieving a cluster structure containing relevant groups for further analysis. Through comparison of the Cophenetic Correlation Coefficient it is possible to understand that implementing
distinct agglomeration criteria (along with the application of different distance measures) results in partition schemes of distinct accuracy. This validity criterion suggests that partition schemes resulting from single-linkage and average-linkage are generally more adequate, with dendrograms presenting a better preservation of the pairwise distances between dataset objects. Nonetheless, the internal nature of this evaluation does not account to generate clusters of a certain required homogeneity level. In fact, both simple-linkage and average-linkage provide solutions that consider several clusters containing a unique object. The analogous anomaly is also present for all the implementations considering Mahalanobis distance as a dissimilarity measure. Also, it is possible to notice that the Mahalanobis distance implementations are inefficiently evaluated by Hubert’s $P$ statistic. Given the aim of grouping countries according to their common characteristics regarding terrorism and tourism, such solutions do not fit for purpose, reason why they are dropped from this stage onwards. Although the results obtained from the implementation of the remaining algorithms present weaker clustering structure when evaluated internally through means of CPCC, the groups formed are generally suitable for the research purpose. The theoretical application of the relative cluster validity criteria purposed in 0, enables to inspect the existing clusters from two different scopes: compactness and separation of clusters. To hold simplicity and minimizing disagreement between validity criteria, the two more stable indices are selected among all the ones presented before:

- the compactness of clusters can be assessed by comparing the values obtained for the $D_k$ index, where large values suggest more compact clusters;
- the degree of separation of clusters is well portrayed by $DB_k$, which evaluates the average similarity between each of the form clusters in a specific solution, for which small values are an evidence of more dissimilar clusters (and thus more distant).

\[ \text{Table 10. Optimal Partition Schemes per HAC Method and Distance Measure Employed} \]

<table>
<thead>
<tr>
<th>HAC Method</th>
<th>Distance Measure</th>
<th>CPCC</th>
<th>( D_k^* )</th>
<th>Silhouette ( k^* )</th>
<th>( DB_k^* )</th>
<th>( \hat{P}_k^* )</th>
<th>( SS_{Wk}^* )</th>
<th>( R_{k}^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Linkage</td>
<td>Euclidean</td>
<td>0.9772</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Mahalanobis</td>
<td>0.9714</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>3</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Manhattan</td>
<td>0.9841</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Complete Linkage</td>
<td>Euclidean</td>
<td>0.9362</td>
<td>4</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Mahalanobis</td>
<td>0.9321</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Manhattan</td>
<td>0.9311</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Average Linkage</td>
<td>Euclidean</td>
<td>0.9493</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Mahalanobis</td>
<td>0.9487</td>
<td>7</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Manhattan</td>
<td>0.9766</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Ward's Minimum Variance</td>
<td>Euclidean</td>
<td>0.8539</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Mahalanobis</td>
<td>0.7502</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Manhattan</td>
<td>0.7823</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Legend: 
- Preferred solution; 
- Alternative solution(s); 
- Inaccurate validity index
That being said, it is possible to identify three outstanding clustering structures that simultaneously present satisfactory degrees of compactness and separation of clusters, as it is illustrated in Figure 9. Such implementations are: (1) CLMN5 which generates five clusters through the application of Complete-Linkage HAC with Manhattan Distance, (2) WMEU3 derives three clusters through the application of Ward’s Minimum Variance HAC using Euclidean Distance, and (3) WMMN4 containing four clusters as a result of implementing Ward’s Minimum Variance HAC with Manhattan Distance. The WMEU3 solution accounts for the highest $D_k$ value among the three highlighted solutions, encompassing the most compact clusters. In opposition, that same solution also presents the highest $DB_k$ value and thus its clusters are not as dissimilar as the ones presented in the remaining highlighted solutions. Additionally, when looking over the composition of the clusters it is notorious the presence of a unitary cluster that only includes India. This situation is recurrent among the generality of the produced solutions and thus it could be an evidence that India represents an outlier. WMEU3 considers two additional clusters of much different dimension, reason why this solution does not fit for the purpose of this study. The remaining hierarchical clustering solutions are CLMN5 and WMMN4. At this point it is possible to remark that Manhattan distance implementations generally perform better than Euclidean distance. A visual comparison of these solutions is provided in Appendix H in the form of a “tanglegram” (a visualization method for comparing dendrogram structures).

*Figure 9. Aggregated Dataset: HAC Solutions (Compactness versus Separation)*

![Dunn Index](image1)

![DB Index](image2)

Although Figure 9 depicts that CLMN5 is a clustering structure presenting slightly more compact and well separated clusters when compared against WMMN4, it is possible to make some remarks from the information portrayed in the above mentioned “tanglegram”:
both solutions are very similar in terms of their inherent clustering structure: in the tanglegram this can be easily noticed by the amount of horizontal lines linking elements (countries) from each solution, meaning they are assigned to clusters in similar way;

the main difference identified is that Russia encompasses its own cluster in CLMN5, while it is grouped together with Israel and Sri Lanka in WMMN4; in fact, Russia represents an outlier in the context of the clustering structure purposed in CLMN5 (just like India does in both of the presented solutions) and thus, when isolated in its own cluster it ameliorates the values regarding compactness and separation of clusters. For this reason, CLMN5 presents more compact and well separated clusters.

Given those considerations, it is possible to conclude that although the relative validity indexes selected for this analysis are helpful to identify an adequate number of clusters in the presence of some level of uncertainty, it is fair to conclude that such measures are somewhat sensible to outliers. In fact, although both solutions are practically similar, it is possible to assume that WMMN4 presents a more appropriate grouping scheme since the trade-off in compactness and separation from including an additional (unitary) cluster it is not sizeable.

The purposed application of partitioning cluster methods encompasses three main algorithms: k-Means, PAM and CLARA, each of these possessing comparative advantages but also drawbacks as described in 1.4.3. A similarly set of cluster validity indices as used for the different HAC methods is employed to identify the optimal partition scheme as a result of the implementation of each algorithm (see Appendices J, K and L). Note that both PAM and CLARA are not evaluate in terms of \( SS_{W_k} \) and \( R_k^2 \) since their objective function optimization does not necessary depend on the minimization of the within group sum of squares. Additionally, PAM and CLARA are evaluated both regarding the implementation of Euclidean and Manhattan distance measures given their flexibility for the use of arbitrary distance measures.

From the application of partitioning cluster methods, the solutions extracted according to the cluster validity criteria are as follows: (1) KM5 and (2) KM6 culminate into five and six clusters, respectively, both as a result of the application of k-Means using Euclidean distance, (3) PAMEU4 is the four-cluster solution derived from PAM using Euclidean distance, and finally, (4) CLARAMN6 encompasses six clusters, partition obtained through application of CLARA using Manhattan distance. Figure 10 allows the comparison of the highlighted solutions in terms of their compactness and separation metrics, similarly to what have been done for the hierarchical clustering solutions. Although there is not a consensual choice among the four encountered solutions, KM6 and CLARAMN6 are both the solutions providing more compact clusters, while PAMEU4 comprehends most dissimilar clusters. Given a relatively small difference in terms of the values of \( D_k \) index assumed by KM6, PAMEU4 and CLARAMN6, it is reasonable to assume that PAMEU4 is globally the best partition among the remaining solutions. Finally, a graphical overview of PAMEU4 solutions is portrayed at Appendix M.
Once again, when comparing the solutions obtained from the application of hierarchical and partitioning cluster methods, the lack of a consensus in the selected cluster validity criteria difficult a clear choice among WMMN4 and PAMEU4 solutions. While the former presents higher values in terms of the $D_k$ index, and thus much more compact clusters, the latter provides a set of more dissimilar and well separated clusters, as it is associated with lower $DB_k$ index values. When considering the practical application of the clustering results within the scope of the research problem it is reasonable to prefer the partition resulting from PAMEU4, given the fact it consists of more dissimilar clusters, for which differences are more evident. Additionally, the resulting clusters are of more uniform size when compared with solutions produced from hierarchical clustering methods.

3.2.2 “Pooled” Dataset Cluster Analysis

As described at the introductory stage of the present section, this approach considers every observation of a country on a certain year to be comparable against each other. In order to simplify the clustering procedure, the results from the previously carried cluster analyses are employed assuming that differences in performance are similar for the present dataset. That being said, the cluster analysis hereby undertaken consists of the application of PAM algorithm, given its sufficient performance and the quality of the results produced. Following the same heuristic employed in the former cluster analyses, Appendix N presents the cluster validity indexes evaluated for the present implementation of PAM using two different distance measures that have been observed to provide
satisfactory results. As a result of those implementations of PAM, three remarkable solutions arise:
(1) PAMEU6 consists of six clusters identified using Euclidean distance; (2) PAMMN4 and (3) PAMMN6 comprehend four and six clusters, respectively, identified through the use of Manhattan distance instead. The comparison of these solutions is done in Figure 11 (below) in terms of the most relevant validation criteria: Dunn, David-Bouldin and Silhouette indices, respectively used to measure compactness, separation and the accuracy on the classification of each observation among the identified clusters. Through analysis of Figure 11, three main conclusions are drawn:

- all solutions present $D_k$ values close to zero, from which is reasonable to accept that the compactness of clusters is quite similar across them;
- PAMMN4 has the highest $DB_k$ value, representing the solution for which the separation level is the lowest. PAMEU6 is the solution that minimizes $DB_k$ and thus maximizes separation; however, PAMMN6 performs similarly;
- PAMMN4 is the solution for which $Silhouette_k$ is maximal, therefore it comprises the best classification of the observations across the identified clusters, among all the three solutions. Although presenting $Silhouette_k$ somewhat smaller, PAMMN6 performs better than PAMEU6, which presents evidence of the worse classification among all solutions (which can play an important role within the computation of the previous cluster validity indices).

**Figure 11.** Evaluation of PAM for Pooled Dataset: Compactness, Separation and Classification

Once again, more than evaluating the aforementioned indices, it is important to understand each solution from the perspective of the research problem. Relying on the separation of clusters over
their compactness, and considering the insights regarding the classification of data points, PAMMN6 is, among the other solutions, the most meaningful partition for the present dataset. The representation of PAMMN6 solution through means of a heat map (see Appendix O) enables a distinctive perception of how countries are individually assigned to clusters through time. It is possible to notice that, although some countries are static, i.e., assigned to the same cluster throughout the whole period, others are volatile. Switching between clusters from year to year.

In order to simplify the interpretation of the present solution, countries are assigned to their respective modal cluster: this approach is more conservative than using average (given the fact that clusters are numbered just for the purpose of coding and, therefore, the mean has no true value). Such re-assignment of countries to their final cluster is illustrated in the bubble chart in Appendix O, together with the respective cluster averages and geographical distribution of the identified clusters.

### 3.3.3 Longitudinal Dataset Cluster Analysis

This section is meant for the application of the final cluster analysis methodology as proposed at the introductory stage of the present section. Such methodology relies on the application of clustering techniques specifically tailored for longitudinal data, having its underlying assumptions regarding longitudinal datasets defined in O. In order to achieve a satisfactory partition, the next cluster analysis implements the \texttt{kml3d} R-package for clustering longitudinal data. The first major concern pinpointed by Genolini et al. (2015) when setting up a longitudinal data clustering problem regards to the choice of the distance measure in use. The “\texttt{kml3d}” package implements \textit{k}-means algorithm, having been tailored for the use of arbitrary distance measures. The use of non-Euclidean distance measures in \textit{k-means} has been discussed to be controversial, giving the fact that different authors disagree that its proprieties verify for non-Euclidean spaces. For that reason, and to overcome possible misleading clustering solutions, Euclidean distance is chosen as distance measure to implement. In the context of longitudinal data clustering the Euclidean distance between two joint variable-trajectories is defined as the analogous Minkowski distance having \( p = 2 \), formally,

\[
\text{Dist}(y_{i1}, y_{i+1}) = \sqrt{\sum_{j \in X} |y_{ij} - y_{i+1,j}|^2} = \sqrt{\sum_{j \in X} |y_{ij} - y_{i+1,j}|^2}.
\]  

Additionally, the authors advert for the need of normalizing the data, given the inherent scale differences in the joint-trajectories at use. Having the application of data standardization techniques been previously discussed for classic clustering problems, it is important to note that, in the context of longitudinal datasets, “each variable-trajectory is not normalized at each time but in its entirety”. Considering \( \overline{y}_{X} \) and \( s_{X} \), respectively the mean and standard deviation of \( y_{X} \), the “\texttt{kml3d}” package applies the global z-score standardization to each \( y_{ijX} \) measurement,
The standardized joint trajectory $y'_{i\ell}$ results from the iterative application of this technique throughout all its single trajectories $y'_{iA}, y'_{iB}, \ldots, y'_{iM}$.

In order to maintain the comparability between the various clustering methodologies at use, the search for the ideal solution is done in similar fashion as for the previous applications:

- the search for the optimal number of clusters in the dataset obeys the same boundaries, i.e., between three and eight clusters;
- the algorithm ($kml3d$) is rerolled for hundred times using the same starting conditions in search for converging solutions in order to mitigate potential random effects inherent to the aleatory nature of k-means based algorithms;
- partitions presenting distinctly better performance according to the listed cluster validity criteria are investigated and interpreted.

As mentioned for the previously employed cluster analysis, not always the solution that maximizes a certain criterion is the one that produces the most relevant partition. The fact that multiple criteria are proposed in the literature, each one of them offering a distinct validation of the resulting partitions, proves the non-uniqueness of an optimal solution for the clustering problem. Figure 12 illustrates the process of searching for the optimal number of clusters according to the Calinski-Harabatz criterion.

**Figure 12. Longitudinal Data Clustering: Validity Criterion (Calinski-Harabatz)**

In theory, the partition that maximizes this index presents outstanding levels of separation and compactness of clusters. Although the figure points for a clear dominance of solutions
encompassing three clusters for the majority of the rerolling attempts, it is easy to conclude that once again this solution identifies one cluster to which only one country is assigned: in fact, similarly to the previous approaches, India is assigned to its own cluster, which reinforces its role as a possible outlier in this dataset. Note that given the constraints with data completion several countries that were lacking data were excluded from the analysis which might be one possible reason for the extreme dissimilarity of India (this is clearly noticeable to be connected with India’s high figures for terrorism-related attributes; including other countries of similar terroristic activity would probably euphemize this issue). In order to mitigate the inclusion of a single element cluster, and after thorough inspection, it was possible to understand that only partitions from the 65th rerolling onwards (as illustrated in the figure) do not face this issue; however, choosing such solution implies a clear loss of clustering quality as the drop in the validity criterion intensifies. Additionally, accepting a solution encompassing any different number of clusters will most of the times result in similar problem.

Through means of this iterative inspection, there are three main advantages of marginally increasing the additional number of clusters:

- although India is still clustered alone (and thus considered as an “outlying cluster”), the remaining clusters become more compact and the separation between them becomes more obvious;
- the risk of creating clusters containing one single element increase for each additional cluster to consider; therefore, several iterations of the algorithm have proven that adding one single cluster (to a total of 4), minimizes such risk, while maintaining an acceptable value for the Calinski-Harabatz criterion;
- finally, the majority of solutions selected from each of the previous approaches encompasses 4 clusters and thus, adopting similar number of clusters for the longitudinal data approach provides a higher level of comparability between methods and results.

That said, the selected solution, say LDKM4, is the one selected among all the possible partitions resulting from the application of the above presented longitudinal data clustering approach. The LDKM4 corresponding joint-trajectories are easily distinguishable in Appendix P, along with the cluster solution means and respective geographical distribution.

4 RESULTS DISCUSSION

This section is meant to establish the comparison of the solutions obtain using different methods while clustering a set of seventy four countries, for which both terrorism and tourism indicators were observed over twenty years, with the goal of proposing a logical grouping of countries given their inherent characteristics regarding such phenomena. Although preliminary results from each macro approach have been discussed and compared with that category of methods, it is still necessary to compare the shortlisted set of solutions to decide for those solutions that best fit the
The purpose of this thesis. The first approach employed relies on a simplistic use of summary values for indicators, allowing for the application of a wider range of methods that generally produce results of quite comparable quality. Although deterministic, and therefore entirely replicable, the application of hierarchical clustering methods generally tends to produce more unitary clusters, which naturally denotes some sensitivity for outliers. On the other hand, partition methods have shown better performance in tackling that issue. Accepting the unavoidable trade-off between cluster separation and compactness lead to the choice of the PAMEU4 solution (an application of Partitioning Around Medoids algorithm). The four encountered clusters, illustrated in Figure 13, can be described as:

- Cluster 1 – countries with low occurrence of terrorism and for which tourism has a moderate contribution for the development of the economy;
- Cluster 2 – countries with low occurrence of terrorism and for which tourism represents the major economic activity, however the inflow of tourists into such countries is of relatively small dimension;
- Cluster 3 – countries with relatively high inflow of tourists and/or for which the terroristic activity is moderate, although present;
- Cluster 4 – unitary cluster, for which the only country is India (IND), the country that presents the highest figures for the majority of the terroristic activity indicators, while performing poorly in terms of inbound tourism when compared to other countries.

*Figure 13. Cluster Solution Overview: PAMEU4*
Although this solution still isolates India (IND) as an outlier, the cardinality of the remaining clusters is moderately balanced as described in Table 11.

### Table 11. Cluster Country List: PAMEU4

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>AGO, ARG, ARM, AUS, AUT, BHR, BGD, BEL, BOL, BRA, BGR, CAN, CHL, CRI, HRV, ECU, EGY, EST, DEU, GRC, GUY, HND, IND, IRL, ISR, KEN, KGZ, LAO, LVA, LSO, LTU, LUX, MKD, MYS, NLD, NIC, NOR, PAN, PNG, PRY, PER, POL, PRT, LKA, SLV, SUR, SVN, SWE, TUN, UGA, UKR, GBR</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>BLZ, BHS, HRV, JOR, LBN, MUS, SYC, VUT</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>CHN, COL, ESP, FRA, ITA, RUS, THA, USA</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>IND</td>
</tr>
</tbody>
</table>

The application of the second approach is *a priori* more challenging than the previous one, given the introduction of the time varying dimension, which imposes limitations from the point of view of static clustering methods. Here each observation of a country on a certain period is seen an individual. The possibility that multiple observations of the same country belong to the same (or to a different) cluster illustrate the shifts between clusters as the time goes by.

Since this approach deals with a much higher number of observations (more precisely, 20 times the size of the previous sample) it is acceptable to exclude the use of hierarchical agglomerative clustering as this class of methods tends to result in more unitary (or low cardinality) clusters, as seen before. Again, the application of PAM has proven to perform best among other partitioning methods. As a final solution, each country is assigned to its modal cluster (most frequent cluster), for which six clusters have been identified and illustrated below in Figure 14:

- Cluster 1 – countries with small (or almost inexistent) terroristic activity and for which tourism is relatively small both in terms of inflow of tourists and contribution for the economy;
- Cluster 2 – countries with small to moderate terroristic activity and small inflow of tourists, however showing a clear economic dependency on tourism;
- Cluster 3 – countries with moderate terroristic activity, moderate inflow of tourists and a moderate economic dependency on tourism;
- Cluster 4 – countries with small (or almost inexistent) terroristic activity and small inflow of tourists, however showing really strong economic dependency on tourism;
- Cluster 5 – countries with intense terroristic activity but small inflow of tourists and weak economic dependency on tourism;
- Cluster 6 – countries with moderate terroristic activity and, however showing relatively high inflow of tourists still have a moderate economic dependency on tourism.
This solution presents a much better partition of the whole set of countries into more distinct groups (see Table 12), accounting for different levels of intensity of both terroristic activity, tourism inflows and economic dependency on the hospitality industry. In addition, the current partition does not isolate any country into a unitary cluster.

Table 12. Cluster Country List: PAMMN6

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>AGO, ARG, ARM, AUS, BGD, BEL, BTN, BOL, BRA, CHL, ECU, SLV, GUY, HND, ISL, IDN, IRL, ISR, JPN, KEN, KGZ, LAO, LVA, LSO, LTU, MKD, NPL, NLD, NIC, NOR, PNG, PRY, PER, SVN, LKA, SUR, SWE, TZA, UGA, URY</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>BHR, BLZ, BGR, CRI, HRV, EGY, EST, JOR, LUX, MUS, PAN, PRT, THA, TUN</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>AUT, CAN, DEU, GRC, MYS, POL, RUS, UKR, GBR</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>BHS, LBN, SYC, VUT</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>COL, IND</td>
</tr>
<tr>
<td>Cluster 6</td>
<td>CHN, FRA, ITA, ESP, USA</td>
</tr>
</tbody>
</table>

Finally, the third approach possesses superior complexity, accounting for both time and cross-sectional dimensions and capturing the variation of each indicator over time for a certain country. Employing a modified k-Means algorithm for clustering longitudinal data, this solution once again is
based on the use of partitioning cluster methods, from which result the 4 following clusters, illustrated in Figure 15.

**Figure 15. Cluster Solution Overview: LDKM4**

Such clusters are described as follows:

- Cluster 1 – countries with low occurrence of terrorism and for which tourism has a moderate contribution for the development of the economy;
- Cluster 2 – countries that, although having a moderate level of terroristic activity, are destination for a relatively high number of inbound tourists;
- Cluster 3 – countries of moderate terroristic activity along with moderate tourism inflow and moderate economic dependency on the tourism sector;
- Cluster 4 - unitary cluster, for which the only country is India (IND), the country that presents the highest figures for most the terroristic activity indicators, while performing poorly in terms of inbound tourism when compared to other countries.

Alike the first approach, employing longitudinal cluster produces a unitary cluster, issue that could not be resolved by setting any different number of clusters to partitioning the existing set of countries due to the extreme outlying profile of India (IND). In addition, there are quite a strong level of similarity between clusters defined by the solutions resulting from both approaches (see Table 13), which can be partly explain by the use of partitioning methods and the selection of four clusters.
Put simply, both PAMEU4 and LDMK4 solutions present a comparable clustering structure, with several common elements (or countries) group together in similar clusters, despite involving completely different levels of computational complexity. Such similarities, both in terms of the clusters cardinality and composition, can be interpreted as indicators of convergence into what is believed to be a quasi-optimal clustering structure. However, it is also perceptible that both solutions are disturbed by the presence of outliers that dominate one of the clusters, issue mitigated in PAMMN6, where no country is left alone in a unitary cluster. In addition, the quality of the final clustering structure is superior given the higher level of cluster separation (and thus higher heterogeneity between clusters). When looking for the cluster description of PAMMN6 partition (hereinafter mentioned as final solution, and represented in Figure 16) it is possible to understand much more features and particularities of different sets of countries, reason why this solution is preferred among all others obtained during this research work.

Figure 16. Final Solution: Geographic Overview
There are few key remarks to highlight when looking deeper into the different clusters provided by the final solution. Although Cluster 1 groups countries with small (or almost inexistent) terroristic activity such as Scandinavian and Baltic countries, along with Japan and Australia, there is a clear contrast with a conglomerate of most countries in South America, countries that have well known levels of violence and conflicts, however not often labeled as terrorism. Perhaps for this reason, these two sub-clusters are grouped together. Additional reasons that might also help explaining the lower inflow of tourism in countries belonging to this cluster relate to the geographic position of such countries: countries in the South hemisphere do not benefit from such developed and diverse transportation methods when compared with North hemisphere countries, which can possibly lead to less frequently tourism travelling. On the other side, although benefiting of the closure to Central Europe and main European transportation hubs, one of the factors that might constrain tourism inflow into the identified Scandinavian and Baltic countries might be the extremely position northern position which is naturally connected with low temperatures and unpleasant weather conditions, which are not favorable to the standards of most tourists.

Slightly more active in terms terroristic activity, and with more pronounced economic dependency on tourism, Cluster 2 presents a set of small countries located closer to the equator which are attractive for tourists during most times of the year given the favorable weather conditions with a great percentage of sunny days and, usually, a vast coast that is the main drive for tourism resorts. Clear examples are countries located in Central America and Caribbean, Portugal and Croatia in Europe, Egypt and Tunisia in North Africa and Thailand in Asia. In addition to geographic reasons that might help explain higher dependency on tourism it is also important to highlight economic factors such as the low purchasing power when compared to visitors’ origin countries.

Cluster 3 represents a class of countries that, although having a significantly higher inflow of tourists than countries in the previous cluster, their economies do not show such a strong dependence on the hospitality industry. Some countries that clearly fit this description are Germany, Austria, Poland, Russia and Canada. In addition, this cluster is also characterized for moderate level of terroristic activity, from which the major contributions belong to countries such as Malaysia, the United Kingdom (mostly impacted by the intense period of terrorist conflict in Northern Ireland promoted by the IRA), Greece, Russia and Ukraine (which has suffered severe increases in terms of terrorism indicators in the past few years).

The core of Cluster 4 is composed of The Bahamas, Seychelles and Vanuatu. These countries have quite small absolute number of inbound visitors when compared with countries from other clusters in analysis; however, these countries show an umbilical economic dependency on the hospitality industry. From the geographic perspective, the countries are very small in area and the three of them are in islands, which might be contributing factor for the underdevelopment of other economic sectors than tourism, given the lack of communication with neighboring countries, with potential impact in trading relationships of all kind. In addition to these countries, and with a sizeable level of
tourism dependency, Lebanon differs from others as a country surrounded by neighboring conflicts of both warfare and terrorism nature.

Cluster 5 comprises only two countries: India and Colombia. Although these countries are distinct geographically distinct, both in size and longitude, they both share favorable proximity to the equatorial latitude, which in theory is an attractive factor for tourists. However, the intense terrorist activities of guerrillas such as FARC, together with attacks promoted by several narco-trafficking cartels, have made Colombia a non-welcoming country for standard tourists. Similarly, in India, tourism has been mostly affected by terrorist causes from ethnic and religious natures, as well as for narco-trafficking reasons. These countries are clearly the most violent terrorism hotspots in analysis.

Finally, Cluster 6 is composed by a set of major players in world tourism: The United States of America, China, France, Italy and Spain. Although these countries have moderate terroristic activity, they still capture the biggest number of visitors around the globe and showing a somewhat relevant economic dependency on their respective hospitality industry. However, it is important noticing that this economic dependency on tourism is clearly stronger in the European countries, while USA and China are placed in this cluster mostly due to the absolute number of tourists.

In overview and put simply, the final solution breaks down broader clusters of the previously discussed solutions, allowing for a more extensive distinction of the profiled countries, and therefore more accurate classification. Although it is not possible to establish causation mechanism between terrorism and tourism activity with such descriptive methodologies (which is not in the scope of this thesis), the final solution clearly depicts different correlations between these two major socioeconomic indicators.

CONCLUSION

The main purpose of this thesis is to classify countries that have been targeted with different frequency and intensity levels of terroristic activity but also considering their respective inherent characteristics regarding inbound tourism. The data cleansing and preparation stage, as well as the conceptual definition of tourism and terrorism, were crucial milestones for having consistent data that illustrate such phenomena. Following a set of considerations regarding the distribution and variability of data, together with a robust clustering analysis that encompasses several classic methods, together with more experimental approaches, allows for the evaluation of the best performing solutions against each other, done in the previous section. As a result, Partitioning Around Medoids, a classic approach within partitioning cluster methods has shown to be an efficient way of distinguishing between 6 major groups of countries, having this classification relied in 3 major dimensions of metrics: the absolute size of the hospitality industry, the relative weight of this industry in the whole economy and, finally the level of terrorist activity. The main goals of this
written work have been successfully achieved, given fact that each of the final solution presents the highest intra-cluster homogeneity and inter-cluster heterogeneity levels from all other tested solutions, while defining a distinct and easily understandable profile of the analyzed countries.

It is also important to pinpoint the impact of inherent limitations that constrained the results of the research hereby carried out, which are mostly connected with deficiencies such as inexistence of data for some of the analyzed attributes in specific countries during significant fractions of the period in study. Likewise, the proeminence of data inconsistencies, which were fairly noticeable in the process of tourism data collection given the lack of a standardized set of data measurement practices across the countries, imposed several constraints to the end product of this written work. For that reason, it is important to mention that a relatively high number of countries where terrorism practices are frequent and intense were left out of this analysis, given the lack of data on tourism.

As for opportunities for further use of this research, the replication of this methodology with more complete (and more consistent) datasets might produce different results, which are expected to provide additional accuracy to the classification of countries. In addition, one interesting suggestion for further research is using the established groups of countries as cross-sectional variables for panel data studies, with aim to verify if the different groups suffer different economic impact from tourism sector as a response of terrorist activity.
REFERENCE LIST


66. Ng, R. T., & Jiawei Han, J. (2002). CLARANS: a method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering, 14*(5), 1003–1016.
APPENDIXES
TABLE OF APPENDIXES

Appendix A: Key Concepts and Definitions on Tourism ................................................................. 1
Appendix B: Attack Type Metadata .................................................................................................. 4
Appendix C: Key Metrics Yearly Distribution per Country ............................................................. 6
Appendix D: Aggregated Dataset Hierarchical Clustering (Simple Linkage) ..................................... 11
Appendix E: Aggregated Dataset Hierarchical Clustering (Complete Linkage) ................................. 13
Appendix F: Aggregated Dataset Hierarchical Clustering (Average Linkage) ................................. 15
Appendix G: Aggregated Dataset Hierarchical Clustering (Ward’s Method) .................................... 17
Appendix H: Aggregated Dataset: Visual Comparison of HAC Solutions ......................................... 19
Appendix I: Aggregated Dataset – WMMN4 Solution Overview ...................................................... 20
Appendix J: Aggregated Dataset k-Means Clustering ....................................................................... 21
Appendix K: Aggregated Dataset PAM Clustering ........................................................................... 22
Appendix L: Aggregated Dataset CLARA Clustering ........................................................................ 23
Appendix M: Aggregated Dataset – PAMEU4 Solution Overview .................................................... 24
Appendix N: Pooled Dataset PAM Clustering Solutions and Validation ........................................... 25
Appendix O: Pooled Dataset – PAMMN6 Cluster Solution ................................................................. 27
Appendix P: Longitudinal Dataset – LDKM4 Clustering Solution ...................................................... 30
### Appendix A: Key Concepts and Definitions on Tourism

**Table 1. Key Concepts and Definitions on Tourism**

<table>
<thead>
<tr>
<th>Concept</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business visitor</td>
<td>A business visitor is a visitor whose main purpose for a tourism trip is to undertake business activity in the place visited.</td>
</tr>
<tr>
<td>Destination (or main destination) of a trip</td>
<td>The main destination of a tourism trip is defined as the place visited that is central to the decision to take the trip. See also purpose of a tourism trip.</td>
</tr>
<tr>
<td>Domestic tourism</td>
<td>Comprises the activities of resident visitors within the country of reference, either as part of a domestic tourism trip or part of an outbound tourism trip.</td>
</tr>
<tr>
<td>Domestic visitors</td>
<td>From the perspective of the country of reference, a domestic traveler qualifies as a domestic visitor if: a) he/she is on a tourism trip, and b) he/she is a resident travelling in the country of reference.</td>
</tr>
<tr>
<td>Employment in tourism industries</td>
<td>Employment in tourism industries may be measured as a count of the persons employed in tourism industries in any of their jobs, as a count of the persons employed in tourism industries in their main job, as a count of the jobs in tourism industries, or as full-time equivalent figures</td>
</tr>
<tr>
<td>Excursionist</td>
<td>A visitor (domestic, inbound or outbound) is classified as a same-day visitor (or excursionist) if his/her trip does not include an overnight stay. Visitors from cruise ships, for example, classify as excursionists because they do not stay overnight in the country of reference.</td>
</tr>
<tr>
<td>Forms of tourism</td>
<td>There are three basic forms of tourism: domestic tourism, inbound tourism, and outbound tourism. These can be combined in various ways to derive the following additional forms of tourism: internal tourism, national tourism and international tourism.</td>
</tr>
<tr>
<td>Inbound tourism</td>
<td>Comprises the activities of non-resident visitors within the country of reference on an inbound tourism trip.</td>
</tr>
<tr>
<td>Outbound tourism</td>
<td>Comprises the activities of a resident visitor outside the country of reference, either as part of an outbound tourism trip or as part of a domestic tourism trip.</td>
</tr>
<tr>
<td>Place of usual residence</td>
<td>The place of usual residence is the geographical place where the enumerated person usually resides, and is defined by the location of his/her principal dwelling.</td>
</tr>
<tr>
<td>Purpose of a tourism trip (main)</td>
<td>The main purpose of a tourism trip is defined as the purpose in the absence of which the trip would not have taken place. Classification of tourism trips according to the main purpose refers to nine categories: this typology allows the identification of different subsets of visitors (business visitors, transit visitors, etc.). See also destination of a trip.</td>
</tr>
<tr>
<td>Concept</td>
<td>Definition</td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| Classification of tourism trips according to the main purpose | 1. Personal  
1.1. Holidays, leisure and recreation  
1.2. Visiting friends and relatives  
1.3. Education and training  
1.4. Health and medical care  
1.5. Religion/pilgrimages  
1.6. Shopping  
1.7. Transit  
1.8. Other  
2. Business and professional |
| Region | The administrative unit corresponding to the first level of territorial disaggregation of a country in terms of its political and administrative organization – for instance, the NUTS 2 level in the EU, provinces in Canada and China, states in Brazil and Mexico, etc. Consequently, the definition of a “region” for the INRouTe initiative refers to a normative criterion; no other criterion (analytical or functional) is considered. |
| Regional tourism | In order to separate visitors to a region who have their place of usual residence within this region from those who come from other regions or countries, it is recommended that three subsets of visitors to or in this region be identified:  
- Residents from countries other than the country of reference (inbound visitors to the country as a whole)  
- Residents from another region of the country of reference  
- Residents in the region of reference (being their usual environment located in such region)  
Regional tourism comprises the activities of these three subsets of visitors. If deemed appropriate and feasible, additional subsets could also be identified for analytical purposes (basically, residents of a region travelling to another part of the national territory / to other countries / to a neighbor country) |
| Same-day visitor | See Excursionist. |
| Tourism expenditure | Tourism expenditure refers to the amount paid by visitors for the acquisition of consumption goods and services, as well as valuables by visitors, for own use or to give away, for and during tourism trips. |
| Tourism industries | Tourism industries (also referred to as tourism activities) are the activities that typically produce tourism characteristic products. Tourism characteristic products are those that satisfy one or both of the following criteria:  
- Tourism expenditure on the product (either good or service) should represent a significant share of tourism expenditure (share-of-expenditure/demand condition);  
- Tourism expenditure on the product should represent a significant share of the supply of the product in the economy (share-of-supply condition). This criterion implies that the supply of a tourism characteristic product would cease to exist in meaningful quantity in the absence of visitors. |
<table>
<thead>
<tr>
<th>Concept</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>List of categories of tourism characteristic products and tourism industries</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Products</strong></td>
<td><strong>Industries</strong></td>
</tr>
<tr>
<td>1. Accommodation services for visitors</td>
<td>1. Accommodation for visitors</td>
</tr>
<tr>
<td>2. Food and beverage serving services</td>
<td>2. Food and beverage serving activities</td>
</tr>
<tr>
<td>3. Railway passenger transport services</td>
<td>3. Railway passenger transport</td>
</tr>
<tr>
<td>4. Road passenger transport services</td>
<td>4. Road passenger transport</td>
</tr>
<tr>
<td>5. Water passenger transport services</td>
<td>5. Water passenger transport</td>
</tr>
<tr>
<td>6. Air passenger transport services</td>
<td>6. Air passenger transport</td>
</tr>
<tr>
<td>7. Transport equipment rental services</td>
<td>7. Transport equipment rental</td>
</tr>
<tr>
<td>8. Travel agencies and other reservation services</td>
<td>8. Travel agencies and other reservation services activities</td>
</tr>
<tr>
<td>9. Cultural services</td>
<td>9. Cultural activities</td>
</tr>
<tr>
<td>10. Sports and recreational services</td>
<td>10. Sports and recreational activities</td>
</tr>
<tr>
<td>12. Country-specific tourism characteristic services</td>
<td>12. Other country-specific tourism characteristic activities</td>
</tr>
</tbody>
</table>

**Tourism sector**

The tourism sector is the cluster of production units in different industries that provide consumption goods and services demanded by visitors. Such industries are called tourism industries.

**Tourist**

A visitor (domestic, inbound or outbound) is classified as a tourist (or overnight visitor), if his/her trip includes an overnight stay in the place visited.

**Travel / tourism**

Travel refers to the activities of travelers who are people who move between different geographic locations, for any purpose and any duration. The visitor is a particular type of traveler and consequently tourism is a subset of travel.

**Trip**

A trip refers to the travel by a person from the time of departure from his/her usual residence until he/she returns: it thus refers to a round trip. Trips taken by visitors are tourism trips.

**Usual environment**

The usual environment of an individual, a key concept in tourism, is defined as the geographical area (though not necessarily a contiguous one) within which an individual conducts his/her regular life routines.

**Vacation home**

A vacation home (sometimes also designated as a holiday home) is a secondary dwelling that is visited by the members of the household mostly for purposes of recreation, vacation or any other form of leisure.

**Visit**

A trip is made up of visits to different places. The term tourism visit refers to a stay in a place visited during a tourism trip.

**Visitor**

A visitor is a traveler taking a trip to a main destination outside his/her usual environment, for less than a year, for any main purpose (business, leisure or other personal purpose) other than to be employed by a resident entity in the country or place visited. A visitor (domestic, inbound or outbound) is classified as a tourist (or overnight visitor), if his/her trip includes an overnight stay, or as a same-day visitor (or excursionist) otherwise.

## Appendix B: Attack Type Metadata

*Table 2. Attack Type Metadata*

<table>
<thead>
<tr>
<th>Attack Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Assassination</strong></td>
<td>An act whose primary objective is to kill one or more specific, prominent individuals. Usually carried out on persons of some note, such as high-ranking military officers, government officials, celebrities, etc. Not to include non-specific members of a targeted group. The killing of a police officer would be an armed assault unless there is reason to believe the attackers singled out a particularly prominent officer for assassination.</td>
</tr>
<tr>
<td><strong>Armed assault</strong></td>
<td>An attack whose primary objective is to cause physical harm or death directly to human beings by use of a firearm, incendiary, or sharp instrument (knife, etc.). Not to include involving the use of fists, rocks, sticks, or other handheld (less-than-lethal) weapons. Also includes involving certain classes of explosive devices in addition to firearms, incendiaries, or sharp instruments. The explosive device subcategories that are included in this classification are grenades, projectiles, and unknown or other explosive devices that are thrown.</td>
</tr>
<tr>
<td><strong>Bombing/ explosion</strong></td>
<td>An attack where the primary effects are caused by an energetically unstable material undergoing rapid decomposition and releasing a pressure wave that causes physical damage to the surrounding environment. Can include either high or low explosives (including a dirty bomb) but does not include a nuclear explosive device that releases energy from fission and/or fusion, or an incendiary device where decomposition takes place at a much slower rate. If an attack involves certain classes of explosive devices along with firearms, incendiaries, or sharp objects, then the attack is coded as an armed assault only. The explosive device subcategories that are included in this classification are grenades, projectiles, and unknown or other explosive devices that are thrown in which the bombers are also using firearms or incendiary devices.</td>
</tr>
<tr>
<td><strong>Hijacking</strong></td>
<td>An act whose primary objective is to take control of a vehicle such as an aircraft, boat, bus, etc. for the purpose of diverting it to an unprogrammed destination, force the release of prisoners, or some other political objective. Obtaining payment of a ransom should not be the sole purpose of a Hijacking, but can be one element of the incident so long as additional objectives have also been stated. Hijackings are distinct from Hostage Taking because the target is a vehicle, regardless of whether there are people/passengers in the vehicle.</td>
</tr>
</tbody>
</table>

*(table continues)*
<table>
<thead>
<tr>
<th>Attack Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hostage tacking (barricade incident)</td>
<td>An act whose primary objective is to take control of hostages for the purpose of achieving a political objective through concessions or through disruption of normal operations. Such are distinguished from kidnapping since the incident occurs and usually plays out at the target location with little or no intention to hold the hostages for an extended period in a separate clandestine location.</td>
</tr>
<tr>
<td>Hostage taking (kidnapping)</td>
<td>An act whose primary objective is to take control of hostages for the purpose of achieving a political objective through concessions or through disruption of normal operations. Differs from Barricade Incidents (above) in that they involve moving and holding the hostages in another location.</td>
</tr>
<tr>
<td>Facility/infrastructure attack</td>
<td>An act, excluding the use of an explosive, whose primary objective is to cause damage to a non-human target, such as a building, monument, train, pipeline, etc. Such include arson and various forms of sabotage (e.g., sabotaging a train track is a facility/infrastructure attack, even if passengers are killed). Facility/infrastructure can include acts which aim to harm an installation, yet also cause harm to people incidentally (e.g. an arson attack primarily aimed at damaging a building, but causes injuries or fatalities).</td>
</tr>
<tr>
<td>Unarmed assault</td>
<td>An attack whose primary objective is to cause physical harm or death directly to human beings by any means other than explosive, firearm, incendiary, or sharp instrument (knife, etc.). involving chemical, biological or radiological weapons are considered unarmed assaults.</td>
</tr>
<tr>
<td>Unknown</td>
<td>The attack type cannot be determined from the available information.</td>
</tr>
</tbody>
</table>

Source: Adapted after National Consortium for the Study of Terrorism and Responses to Terrorism (START), *Global Terrorism Database: Codebook and Inclusion Criteria*, 2015.
Appendix C: Key Metrics Yearly Distribution per Country

*Figure 1. International Tourist Arrivals (Millions) per Country*

*Figure 2. International Tourism Receipts (Thousands of Million USD) per Country*

*Figure 3. Gross Domestic Product (Current Prices, Billion USD) per Country*
Figure 4. Total Terrorist Attacks per Country

Figure 5. Successful Terrorist Attacks per Country

Figure 6. Casualties per Country
Figure 7. Injured Victims per Country

Figure 8. Estimated Material Damages (Current Prices, Million USD) per Country

Figure 9. Hostage-taking Victims per Country
Figure 10. Multiple Attacks per Country

Figure 11. Extended Attacks per Country

Figure 12. Suicide Attacks per Country
Figure 13. Claimed Attacks per Country

Figure 14. International Tourism Contribution (in % of GDP) per Country

Figure 15. Failed Terrorist Attacks
Appendix D: Aggregated Dataset Hierarchical Clustering (Simple Linkage)

Figure 16. HAC-SL (Euclidean Distance): Solution and Validation

Figure 17. HAC-SL (Mahalanobis Distance): Solution and Validation
Figure 18. HAC-SL (Manhattan Distance): Solution and Validation
Appendix E: Aggregated Dataset Hierarchical Clustering (Complete Linkage)

Figure 19. HAC-CL (Euclidean Distance): Solution and Validation

Figure 20. HAC-CL (Mahalanobis Distance): Solution and Validation
Figure 21. HAC-CL (Manhattan Distance): Solution and Validation
Appendix F: Aggregated Dataset Hierarchical Clustering (Average Linkage)

**Figure 22.** HAC-AL (Euclidean Distance): Solution and Validation

![Graph](image)

**Figure 23.** HAC-AL (Mahalanobis Distance): Solution and Validation

![Graph](image)
Figure 24. HAC-AL (Manhattan Distance): Solution and Validation
Appendix G: Aggregated Dataset Hierarchical Clustering (Ward’s Method)

Figure 25. HAC-WMV (Euclidean Distance): Solution and Validation

Figure 26. HAC-WMV (Mahalanobis Distance): Solution and Validation
Figure 27. HAC-WMV (Manhattan Distance): Solution and Validation

![Graph showing various validation indices for HAC-WMV (Manhattan Distance)](graph)

- **Dunn Index**
- **Silhouette Index**
- **Davies-Bouldin Index**
- **Hubert’s Statistic**
- **Within Group Sum of Squares**
- **R-Squared**

Legend:
- IND: INDIA
- SYC: SYRIA
- BHS: BOSNIA HERZEGOVINA
- VUT: VATICAN CITY
- BLZ: BELIZE
- MUS: MUSCAT
- HRV: HRVATSKA
- JOR: JORDAN
- EST: ESTONIA
- LUX: LUXEMBOURG
- GUY: GUAYANA
- CRI: COSTA RICA
- PAN: PANAMA
- BHR: BAHRAIN
- BGR: BULGARIA
- TUN: TUNISIA
- AGO: ANGUILLA
- UGA: UGANDA
- AUT: AUSTRIA
- MYS: MALAYSIA
- GBR: UNITED KINGDOM
- CAN: CANADA
- POL: POLAND
- NIC: NICARAGUA
- LTU: LITHUANIA
- URY: URUGUAY
- BTN: BUTAN
- SUR: SURINAME
- LVA: LETTONIA
- SLV: EL SALVADOR
- PRT: PORTUGAL
- HND: HONDURAS
- KGZ: KIRGIZSTAN
- TZA: TANZANIA
- ARM: ARMENIA
- ISL: ICELAND
- LAO: LAOS
- SVN: SLOVENIA
- IRL: IRELAND
- NLD: NETHERLANDS
- AUS: AUSTRALIA
- BEL: BELGIUM
- BRA: BRAZIL
- SWE: SWEDEN
- ARG: ARGENTINA
- NOR: NORWAY
- PER: PERU
- PNG: PAPUA NEW GUINEA
- BOL: BOLIVIA
- LSO: LESOTHO
- MKD: MACEDONIA
- CHL: CHILE
- ECU: ECUADOR
- PRY: PARAGUAY
- CHN: CHINA
- DEU: GERMANY
- ITA: ITALY
- FRA: FRANCE
- ESP: SPAIN
- USA: UNITED STATES
- NPL: NEW ZEALAND
- JPN: JAPAN
- BGD: BANGLADESH
- KEN: KENYA
- LBN: LIBAN
- GRC: GREECE
- UKR: UKRAINE
- EGY: EGYPT
- IDN: INDONESIA
- COL: COLOMBIA
- THA: THAILAND
- RUS: RUSSIA
- ISR: ISRAEL
- LKA: SRI LANKA
- ARM: ARMENIA
- ISL: ICELAND
- LAO: LAOS
- SVN: SLOVENIA
- IRL: IRELAND
- NLD: NETHERLANDS
- AUS: AUSTRALIA
- BEL: BELGIUM
- BRA: BRAZIL
- SWE: SWEDEN
- ARG: ARGENTINA
- NOR: NORWAY
- PER: PERU
- PNG: PAPUA NEW GUINEA
- BOL: BOLIVIA
- LSO: LESOTHO
- MKD: MACEDONIA
- CHL: CHILE
- ECU: ECUADOR
- PRY: PARAGUAY
- CHN: CHINA
- DEU: GERMANY
- ITA: ITALY
- FRA: FRANCE
- ESP: SPAIN
- USA: UNITED STATES
- NPL: NEW ZEALAND
- JPN: JAPAN
- BGD: BANGLADESH
- KEN: KENYA
- LBN: LIBAN
- GRC: GREECE
- UKR: UKRAINE
- EGY: EGYPT
- IDN: INDONESIA
- COL: COLOMBIA
- THA: THAILAND
- RUS: RUSSIA
- ISR: ISRAEL
- LKA: SRI LANKA
- ARM: ARMENIA
- ISL: ICELAND
- LAO: LAOS
- SVN: SLOVENIA
- IRL: IRELAND
- NLD: NETHERLANDS
- AUS: AUSTRALIA
- BEL: BELGIUM
- BRA: BRAZIL
- SWE: SWEDEN
- ARG: ARGENTINA
- NOR: NORWAY
- PER: PERU
- PNG: PAPUA NEW GUINEA
- BOL: BOLIVIA
- LSO: LESOTHO
- MKD: MACEDONIA
- CHL: CHILE
- ECU: ECUADOR
- PRY: PARAGUAY
- CHN: CHINA
- DEU: GERMANY
- ITA: ITALY
- FRA: FRANCE
- ESP: SPAIN
- USA: UNITED STATES
- NPL: NEW ZEALAND
- JPN: JAPAN
- BGD: BANGLADESH
- KEN: KENYA
- LBN: LIBAN
- GRC: GREECE
- UKR: UKRAINE
- EGY: EGYPT
- IDN: INDONESIA
- COL: COLOMBIA
- THA: THAILAND
- RUS: RUSSIA
- ISR: ISRAEL
- LKA: SRI LANKA
- ARM: ARMENIA
- ISL: ICELAND
- LAO: LAOS
- SVN: SLOVENIA
- IRL: IRELAND
- NLD: NETHERLANDS
- AUS: AUSTRALIA
- BEL: BELGIUM
- BRA: BRAZIL
- SWE: SWEDEN
- ARG: ARGENTINA
- NOR: NORWAY
- PER: PERU
- PNG: PAPUA NEW GUINEA
- BOL: BOLIVIA
- LSO: LESOTHO
- MKD: MACEDONIA
- CHL: CHILE
- ECU: ECUADOR
- PRY: PARAGUAY
- CHN: CHINA
- DEU: GERMANY
- ITA: ITALY
- FRA: FRANCE
- ESP: SPAIN
- USA: UNITED STATES
- NPL: NEW ZEALAND
- JPN: JAPAN
- BGD: BANGLADESH
- KEN: KENYA
- LBN: LIBAN
- GRC: GREECE
- UKR: UKRAINE
- EGY: EGYPT
- IDN: INDONESIA
- COL: COLOMBIA
- THA: THAILAND
- RUS: RUSSIA
- ISR: ISRAEL
- LKA: SRI LANKA
- ARM: ARMENIA
- ISL: ICELAND
- LAO: LAOS
- SVN: SLOVENIA
- IRL: IRELAND
- NLD: NETHERLANDS
- AUS: AUSTRALIA
- BEL: BELGIUM
- BRA: BRAZIL
- SWE: SWEDEN
- ARG: ARGENTINA
- NOR: NORWAY
- PER: PERU
- PNG: PAPUA NEW GUINEA
- BOL: BOLIVIA
- LSO: LESOTHO
- MKD: MACEDONIA
- CHL: CHILE
- ECU: ECUADOR
- PRY: PARAGUAY
- CHN: CHINA
- DEU: GERMANY
- ITA: ITALY
- FRA: FRANCE
- ESP: SPAIN
- USA: UNITED STATES
- NPL: NEW ZEALAND
- JPN: JAPAN
- BGD: BANGLADESH
- KEN: KENYA
- LBN: LIBAN
- GRC: GREECE
- UKR: UKRAINE
- EGY: EGYPT
- IDN: INDONESIA
- COL: COLOMBIA
- THA: THAILAND
- RUS: RUSSIA
- ISR: ISRAEL
- LKA: SRI LANKA
- ARM: ARMENIA
- ISL: ICELAND
- LAO: LAOS
- SVN: SLOVENIA
- IRL: IRELAND
- NLD: NETHERLANDS
- AUS: AUSTRALIA
- BEL: BELGIUM
- BRA: BRAZIL
- SWE: SWEDEN
- ARG: ARGENTINA
- NOR: NORWAY
- PER: PERU
- PNG: PAPUA NEW GUINEA
- BOL: BOLIVIA
- LSO: LESOTHO
- MKD: MACEDONIA
- CHL: CHILE
- ECU: ECUADOR
- PRY: PARAGUAY
- CHN: CHINA
- DEU: GERMANY
- ITA: ITALY
- FRA: FRANCE
- ESP: SPAIN
- USA: UNITED STATES
- NPL: NEW ZEALAND
- JPN: JAPAN
- BGD: BANGLADESH
- KEN: KENYA
- LBN: LIBAN
- GRC: GREECE
- UKR: UKRAINE
- EGY: EGYPT
- IDN: INDONESIA
- COL: COLOMBIA
- THA: THAILAND
- RUS: RUSSIA
- ISR: ISRAEL
- LKA: SRI LANKA
- ARM: ARMENIA
- ISL: ICELAND
- LAO: LAOS
- SVN: SLOVENIA
- IRL: IRELAND
- NLD: NETHERLANDS
- AUS: AUSTRALIA
- BEL: BELGIUM
- BRA: BRAZIL
- SWE: SWEDEN
- ARG: ARGENTINA
- NOR: NORWAY
- PER: PERU
- PNG: PAPUA NEW GUINEA
- BOL: BOLIVIA
- LSO: LESOTHO
- MKD: MACEDONIA
- CHL: CHILE
- ECU: ECUADOR
- PRY: PARAGUAY
- CHN: CHINA
- DEU: GERMANY
- ITA: ITALY
- FRA: FRANCE
- ESP: SPAIN
- USA: UNITED STATES
- NPL: NEW ZEALAND
- JPN: JAPAN
- BGD: BANGLADESH
- KEN: KENYA
- LBN: LIBAN
- GRC: GREECE
- UKR: UKRAINE
- EGY: EGYPT
- IDN: INDONESIA
- COL: COLOMBIA
- THA: THAILAND
- RUS: RUSSIA
- ISR: ISRAEL
- LKA: SRI LANKA
- ARM: ARMENIA
- ISL: ICELAND
- LAO: LAOS
- SVN: SLOVENIA
- IRL: IRELAND
- NLD: NETHERLANDS
- AUS: AUSTRALIA
- BEL: BELGIUM
- BRA: BRAZIL
- SWE: SWEDEN
- ARG: ARGENTINA
- NOR: NORWAY
- PER: PERU
- PNG: PAPUA NEW GUINEA
- BOL: BOLIVIA
- LSO: LESOTHO
- MKD: MACEDONIA
- CHL: CHILE
- ECU: ECUADOR
- PRY: PARAGUAY
- CHN: CHINA
- DEU: GERMANY
- ITA: ITALY
- FRA: FRANCE
- ESP: SPAIN
- USA: UNITED STATES
- NPL: NEW ZEALAND
- JPN: JAPAN
- BGD: BANGLADESH
- KEN: KENYA
- LBN: LIBAN
- GRC: GREECE
- UKR: UKRAINE
- EGY: EGYPT
- IDN: INDONE
Appendix H: Aggregated Dataset: Visual Comparison of HAC Solutions

Figure 28. HAC Comparison – Tanglegram
Appendix I: Aggregated Dataset – WMMN4 Solution Overview

Figure 29. WMMN4 Solution: Cluster Means

Figure 30. WMMN4 Solution: Geographical Distribution

20
Appendix J: Aggregated Dataset k-Means Clustering

*Figure 31. K-Means: Validation Criteria*

![Validation Criteria Diagram](image1)

*Figure 32. K-Means: Clustering Solutions*

![Clustering Solutions Diagram](image2)
Appendix K: Aggregated Dataset PAM Clustering

Figure 33. PAM (Euclidean Distance): Validation Criteria

Figure 34. PAM (Euclidean Distance): Clustering Solutions

The PCA accounts for 72.44% of the data variability

Figure 35. PAM (Manhattan Distance): Validation Criteria
Appendix L: Aggregated Dataset CLARA Clustering

**Figure 36.** CLARA (Euclidean Distance): Validation Criteria

Clustering Large Applications (CLARA) using Euclidean Distance

**Figure 37.** CLARA (Manhattan Distance): Validation Criteria

Clustering Large Applications (CLARA) using Manhattan Distance

**Figure 38.** CLARA (Manhattan Distance): Clustering Solution

The PCA accounts for 72.45% of the data variability

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>USA</td>
</tr>
<tr>
<td>2</td>
<td>CAN</td>
</tr>
<tr>
<td>3</td>
<td>AUS</td>
</tr>
<tr>
<td>4</td>
<td>ARG</td>
</tr>
<tr>
<td>5</td>
<td>MUS</td>
</tr>
<tr>
<td>6</td>
<td>BGR</td>
</tr>
</tbody>
</table>
Appendix M: Aggregated Dataset – PAMEU4 Solution Overview

Figure 39. PAMEU4 Solution: Cluster Means

Figure 40. PAMEU4 Solution: Geographical Distribution
Appendix N. Pooled Dataset PAM Clustering Solutions and Validation

Figure 41. PAM (Euclidean Distance): Validation Criteria

- Dunn Index
- Silhouette Index
- Davies-Bouldin Index
- Hubert Statistic

Figure 42. PAM (Euclidean Distance): Clustering Solution

The PCA accounts for 54.08% of the data variability
Figure 43. PAM (Manhattan Distance): Validation Criteria

Partitioning Around Medoids (PAM) using manhattan Distance

- Dunn Index
- Silhouette Index
- Davies-Bouldin Index
- Hubert G Statistic

Figure 44. PAM (Manhattan Distance): Clustering Solution

The PCA accounts for 54.08% of the data variability

Cluster visualization with PCA scores for each cluster.
Appendix O: Pooled Dataset – PAMMN6 Cluster Solution

Figure 45. PAMMN6 Solution: Chronological Heatmap
Figure 46. PAMMN6: Mode versus Variability
Figure 47. PAMMN6 Solution: Cluster Means

Figure 48. PAMMN6 Solution: Geographical Distribution
Appendix P: Longitudinal Dataset – LDKM4 Clustering Solution

*Figure 49. LDKM4 Joint-trajectories Clustering*
Figure 50. LDKM4 Solution: Cluster Means

Figure 51. LDKM4 Solution: Geographical Distribution