Analytical Tool for Advanced Persistent Threats Considering Geopolitical Data (September 2021)

Miguel Pedro Peidro Paredes, *Carlos III de Madrid University, Computer Engineer, Puerta de Toledo, Madrid*

PhD José María De Fuentes García-Romero de Tejada, *Carlos III de Madrid University*, Associate *Professor Computer Science Department*

Abstract— **Advanced Persistent Threats show an exponential increase in their activity and level of hostility. The damage caused by the State Actors responsible for these incidents is mainly directed at the public sector, critical infrastructures and strategic companies. All this, coupled with the complexity of detecting these incidents, poses a challenge for the containment of these threats. In response we can see the effort of the main agencies responsible for building detection, prevention and response systems. These systems are mainly focused on infrastructure analysis, but we should not ignore that the motivation of these State Actors may potentially be linked to geopolitical objectives and strategies. The present study addresses the detection and analysis of these threats from a geopolitical perspective. Our results show that using a multiple linear regression model it is possible to predict the number of incidents between pairs of countries in a given period of time with an accuracy of 86.55% and an average prediction error of 1 incident for the most belligerent countries.**

Index Terms— **Advanced Persistent Threats, GDELT, BigData, Cybersecurity Incident Prediction.**

I. INTRODUCTION

DURING the last year 2020, the Spanish CCN detected 73,184 cybersecurity incidents during the first eleven months of the period analyzed. This trend far exceeds the predictions made for that period. The CCN-CERT also notes that "Spain suffers daily cyber-attacks of critical or very high danger against the public sector and strategic companies. Many of these actions come from other states, which have among their motivations to weaken national political, technological and economic capacity." [1].

It is important to highlight that these nation-state actors have large resources, highly qualified personnel, and high capabilities in cyberspace, carrying out operations of influence, cyber espionage, disruptive and control operations, with increasing complexity and impact. [2].

All this poses a hostile situation and a challenge for the detection, prevention and knowledge of the operations performed by these nation-state actors in order to mitigate their impact throughout the National Territory. The acquisition of this information should not only come from the technological infrastructure, even the CCN-CERT itself recognizes the geopolitical analysis as a crucial source of information to understand the situation in the cyberspace [3].

To date, there are no solutions that rely on geopolitical analysis to anticipate the risks posed by these threats, even demonstrating the feasibility of predicting events with BigData sources like GDELT [4], the Advanced Persistent Threat detection systems focus on the **analysis of infrastructures**.

Therefore, we consider that, demonstrating a correlation between geopolitical events and advanced incidents, can contribute to the detection and prevention of these threats.

To address this objective, this work aims in the first place to tackle the following research question: **Is geopolitical analysis a useful tool** for the detection and prevention of advanced cyber-attacks? To answer this question , this study proposes the following contributions:

- A comprehensive dataset with the collection of incidents attributed to nation-state actors, collecting the temporal context of the incidents, as well as the countries involved.
- A second dataset composed by the geopolitical context of each incident collected in the first one.
- A correlation analysis between the two previous datasets using Multiple Linear Regression models.

Miguel Pedro Peidro Paredes is with Computer Science Department, Carlos III de Madrid University, Student, Puerta de Toledo, 28270 (e-mail: 100363615@alumnos.uc3m.es).

PhD José María De Fuentes García-Romero de Tejada, Carlos III de Madrid University, Associate Professor Computer Science Department (e-mail: jfuentes@inf.uc3m.es).

In case of correlation found between these datasets, the present study proposes the following contribution:

- An analytical-predictive tool, which in addition to providing the analysis of the cyberspace situation, it predicts based on the geopolitical context analysis, the **advanced incidents** occurrence to anticipate these threats.

The remainder of this paper is structured as follows. Section II describes the background, Section III describes the methodology used in the whole process of the research development, Section IV through Section VII describes in depth the technical process carried out for the achievement of each proposed contribution and finally, Section VIII describes the related work and Section IX describes the research conclusions.

II. BACKGROUND

Although the application of Big Data solutions in conjunction with artificial intelligence algorithms or mathematical models are the most widely used tools for understanding complex problems, there are still fields in which their application is limited. Hybrid warfare and specifically cyberspace situational awareness is a clear example of this phenomenon.

Considering the high impact that Advanced Persistent Threats have on our society, we believe that the contribution to this field with the application of some of these techniques means a substantial benefit for the strategic situation of our country [5].

The creation of a tool that provides information on the cyberspace state, requires a large volume of data prior to the period to be analyzed. This preliminary requirement faces difficulties due to the **lack of standardization** of such information. One of the requirements for the analysis of advanced incidents is **their attribution.** This attribution is the result of an exhaustive study of threat behavior, which translates into collaborative efforts between Certs and research groups. Each institution involved provides the result of this attribution process individually. As an example, the data sources that we will use for the construction of the first dataset (MITRE and ThaiCert), individually publish a list of advanced groups. In this directories you can access information such as the countries associated to these groups or the techniques and tools they use among other information resulting from these attribution processes. The problem faced is that these data are not provided in a standardized and unified manner. As a result, there is no **public repository** that collects incidents attributed to State Actors.

The creation of a repository that unifies these incidents, their temporal context, incident locations, countries involved, besides being the first front in the development of an analytical-predictive tool, requires, **due to the number of incidents**, an **automated acquisition** process on the agencies reporting the attribution such as MITRE or ThaiCert.

MITRE is an American non-profit organization in charge of federally funded research and development centers (FFRDCs[1]). This agency provides information on the attributions performed from an advanced group perspective.

ThaiCert on the other hand, is the CERT appointed by the Thai Government, responsible for the management of cyber threats in Thailand. The information provided by this agency will also be used for the construction of the advanced incident dataset.

Although other agencies such as Fireeye provide data on advanced groups. MITRE and ThaiCert provides high volume of incidents with a common structure.

Regarding the geopolitical context of the incidents, the BigData GDELT collects since 1979 the largest volume of events from news, social medias, and public information sources with a global scope. All these events are included in the GDELT database using the CAMEO[2] framework.

For the representation of the geopolitical context, we will focus on the analysis of an indicator included in the events database known as the **Goldstein Index**. This metric represents the tone of an event or, in the same way, the friendliness or hostility of the country that originates the event towards the country it affects. The possible values of this metric range from -10, representing the maximum of the negative character, to 10 representing its opposite. Therefore, this metric allows us to know the **geopolitical relationship between pairs of countries** in specific periods of time.

III. METHODOLOGY

This section explains the process followed from the creation of the first dataset, through the study of the correlation between the geopolitical context and the advanced incidents, to the use of this correlation to build a tool to make predictions and study the current situation in cyberspace.

This section provides a high-level view of the project development. The technical processes followed for the construction of the datasets, the study of the correlation and the development of the analytical-predictive tool are developed in depth in the chapters dedicated to each contribution.

For the first contribution, the development of a tool capable of anticipating certain events or analyzing specific situations in a

[1] Federally funded research and development centers (FFRDCs) are public-private partnerships which conduct research and development for the United States Government.

[2] Conflict and Mediation Event Observations (CAMEO) is a framework for coding event data (typically used for events that merit news coverage, and generally applied to the study of political news and violence).

given period of time; the beginning point is the acquisition of previous events.

As shown in the Figure below, the information of **incidents** obtained from MITRE and ThaiCert composes the **first dataset**. These incidents are used to query the GDELT data source to obtain the **geopolitical context of the incidents**. This information composes the **second dataset**. Finally, the analysis of the correlation between both datasets is performed.

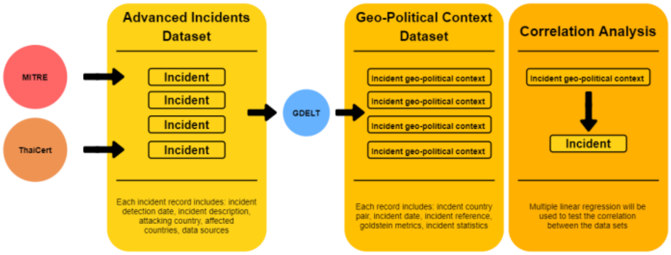
MITRE
ThaiCert
Advanced Incidents Dataset
Incident
Incident
Incident
Incident
Each incident record includes: incident detection date, incident description, attacking country, affected countries, data sources
GDELT
Geo-Political Context Dataset
Incident geo-political context
Incident geo-political context
Incident geo-political context
Incident geo-political context
Each record includes: incdent country pair, incident date, incident reference, goldstein metrics, incident statistics
Correlation Analysis
Incident geo-political context
Incident
Multiple linear regression will be used to test the correlation between the data sets

Figure 1 First Development Phase

To be considered of interest, the events must meet several requirements that will be described in depth in the following section. It is crucial to clarify that the data from both datasets will be used together to configure the analytical-predictive tool.

The first dataset will be composed of all cybersecurity incidents attributed to State Actors by the MITRE and ThaiCert sources. These sources record Advanced Persistent Threat incidents structured by groups. Those groups whose activity is not attributed to a specific country will be excluded. Regarding the incident data attributed to each advanced group, the data sources only include the references to the other sources reporting the incident. An automated acquisition procedure is required to extract, from the reference, relevant information such as time stamps or affected countries.

The second contribution aims to collect data representing the geopolitical context of the incidents in the first dataset. The objective is to analyze the correlation between the detected incidents and the geopolitical data that can be acquired from publicly available data sources. For the study of this correlation a Multiple Linear Regression model will be used, yielding the contributions of all the geopolitical variables for the study of advanced incidents.

In case of obtaining a statistically significant correlation between such data, a model will be developed to make predictions, in addition to the analytical component of the tool, which will allow consulting the advanced incident dataset to obtain statistics and visualize the state of cyberspace in a certain period of time. All this translates into obtaining a measure of risk or probability of occurrence of an advanced incident under certain conditions.

The figure below shows how the analytical component uses the first dataset to obtain incident analytics, while the predictive component queries the geopolitical context of the incidents from the GDELT data source, which in turn feeds the predictive model to obtain the prediction.

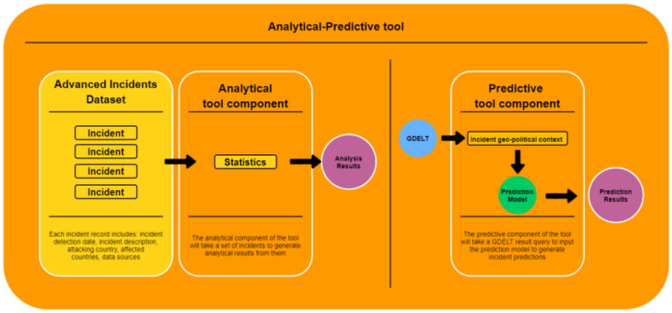
Analytical-Predictive tool
Advanced Incidents Dataset
Incident
Incident
Incident
Incident
Each incident record includes: incident detection date, incident description, attacking county, affected countries, data sources
Analytical tool component
Statistics
The analytical component of the tool will take a set of incidents to generate analytical results from them
Analysis Results
GDELT
Predictive tool component
Incident geo-political context
Prediction Model
The predictive component of the tool will take a GDELT result query to input the prediction model to generate incident predictions
Prediction Results

Figure 2 Second Development Phase

It should be noted that all the datasets generated during the development of this work will be publicly accessible, as well as all the code developed for the creation of these datasets and the code corresponding to the analytical-predictive tool. This code will be available in the servers of the Computer Security Lab (COSEC)[3] · University Carlos III de Madrid.

IV. APT-RELATED INCIDENTS DATASET

This chapter develops in depth the technical aspect of the first dataset creation, explaining the problems faced and the solutions adopted. The format of the data sets obtained can finally be found in the annex “Datasets Documentation”.

The creation of a dataset of advanced incidents attributed to nation-state actors by MITRE and ThaiCert data sources, besides being a prerequisite for the development of the proposed analytical-predictive tool, is a contribution itself as there is no public repository that collects advanced incidents including their temporal context, attribution, countries involved in the incidents, differentiating between attacking and affected countries.

The two main objectives for the creation of this dataset is to provide uniqueness to the data published by MITRE and ThaiCert, as well as generate a data source that allows us to **analyze advanced incidents**, being able to obtain as a result of such analysis the evolution of incidents between certain countries, the level of hostility of a particular country against a set of victim countries or the cyberspace situation in a given period of time. These functionalities are part of the analytical tool which is detailed in its dedicated chapter.

[3] COSEC server: https://cosec.inf.uc3m.es/

The first step for the dataset creation is the analysis of the data structure used by the MITRE and ThaiCert. MITRE data source[4] provides a set of data associated with advanced groups. Using a table format, it provides for each group a unique identifier, their attributed name, a list of associated groups if known, and a brief description of the analyzed group. In the specific pages of each group, there is a summary of the above information as well as a list of techniques and software used by that group. Finally, we can see a list of references that collect all the incidents detected and carried out by that advanced group. These references must be analyzed to extract from them time stamps and countries involved.

The analysis of the ThaiCert[5] data source follows a similar scheme, including an indicator of group activity in the last month, the advanced group name, country associated with the group (if included in the attribution), associated groups (if known), last group update date. If we access the group page we find the group description, target sectors and as in the MITRE data source a list of references corresponding to the incidents carried out by that group.

The necessary requirements to register an incident in the repository are the acquisition of the **attacking country** attributed to the advanced group, **date of incident detection, affected countries**, being necessary to detect at least one country to generate a relationship of attacker and victim between pairs of countries. It will also be necessary to include an incident in the repository including the **reference** to ensure the source of the attribution.

In the case of the ThaiCert data source, the detection date of the incident is included in conjunction with the incident reference, whereas for the MITRE data source, the automated acquisition process must be able to extract this timestamp to include the incident. This timestamp must specify at least the year of detection, with the month being possible to record if known.

A similar situation is seen in the acquisition of the country associated with the attacking group. The ThaiCert data source provides this data if included in the attribution, the MITRE data source requires an automated acquisition process for the extraction of this field. In most cases, this data can be found in the Advanced group description.

This automated acquisition procedure executes a Python script that performs a web scraping process executing the following steps:

For the MITRE data source, the first stage of the script performs the web scraping process on the general page of advanced groups. During this stage, the acquisition of the advanced group name and date of the first detection is carried. In addition, an analysis of the group description is needed for the extraction of the group country.
The second stage of the script iterates over the specific page of each group and over the references corresponding to each incident performing the web scraping process. During this stage, the description of the incident, the detection date of the incident, the affected countries extracted from the references and the reference to save the source are acquired.

For the ThaiCert data source, the first stage follows a sequence similar to the MITRE data source, without analyzing the group description for the extraction of the associated country, since the data source integrates this data natively. From this stage we obtain the name of the advanced group, the first detection date of the advanced group and the country associated to the advanced group.

The second stage also iterates over the page of each group and the references corresponding to each incident. The date of the incident is also included natively on the group specific page, so the web scraping process automatically acquires the affected countries from the incident references.

For the acquisition of timestamps on the references, the text of the reference is compared after the web scraping process using the Python "dateparser" library [10]. For the acquisition of affected countries on the references, the text is compared after web scraping against the Python library "pycountry" [11].

For the validation of these data a first manual analysis is performed finding some cases in which the list of countries affected amounts to more than 200, these errors are due to the inclusion of country lists in the source code of the references. The solution applied to this problem is the filtering of these records using an additional script for the detection of these errors and the manual correction of those.

We also found that in some results of MITRE data source analysis, the country attributed to an advanced group is not included. These records are stored in the database without associated country because the automated acquisition process is not able to extract it. Since this data is provided by multiple sources such as ThaiCert, it is possible the inclusion of these records. The procedure used to solve this problem was to update this field manually by including the **source used for attribution** in the group references field.

Considering that this dataset will be used for the acquisition of geopolitical data from the BigData GDELT source, it is interesting, in order to simplify the creation of queries, to enrich the database by adding the CAMEO Country code[6] [12] of both, the country associated to the group and the countries affected, matching these fields with the format used by the

[4] The reference to the Mitre data source is to the advanced groups section provided in the domain: "attack.mitre.org/groups/".

[5] The advanced groups section provided by ThaiCert is available in the domain: "apt.thaicert.or.th/cgi-bin/listgroups.cgi

[6] Similar to ISO-3166 Codes, Cameo Country Codes is a standard country notation composed by three capital letters, we can find this codes in the CAMEO manual provided by GDELT in the references

GDELT data source. To carry out this task, a script developed in Python has been used to perform a search on a repository of CAMEO Country codes, querying the country associated to the group and the countries affected in each incident, including the CAMEO Country codes of these countries in a new collection called "APT_Group_Incidents_CAMEO".

The result of this process is a dataset with a total number of **1,654 advanced incidents**, which can be **publicly accessed**.

These results are temporally distributed as follows:

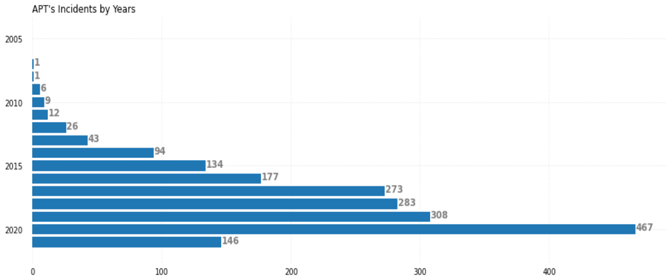

Figure 3 Advanced incidents distribution over time

The figure above shows the number of advanced incidents collected for each year ordered from 2008, being the first incident recorded, to the current year, 2021.

The distribution of incidents by attacking countries can also be observed in the following figure.

Due to the number of attacking countries, to improve the visualization of the data we will show the top 10 countries with the highest volume of incidents.

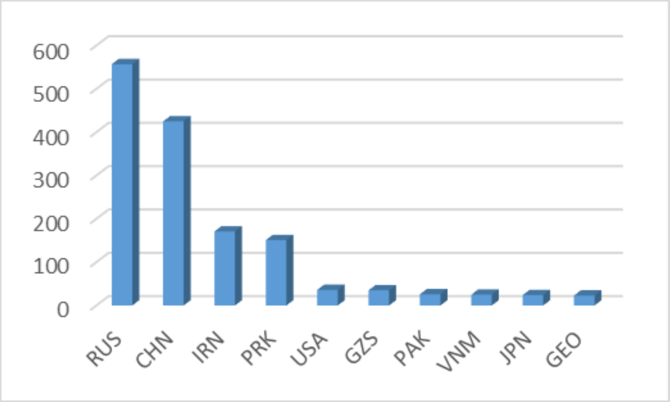

Figure 5 Top 10 Incidents by Country

We can see that there is a total predominance of Russia, China, Iran and North Korea in terms of volume of incidents. On the opposite, we see some countries with a very small number of detected incidents. These cases may be due to a reduced activity of these countries or an error in the automated information acquisition procedure, which will be discussed later.

This patterns can be observed in trends reported by multiple Certs and entities responsible for advanced cybersecurity incidents [2] [3] [1] [13].

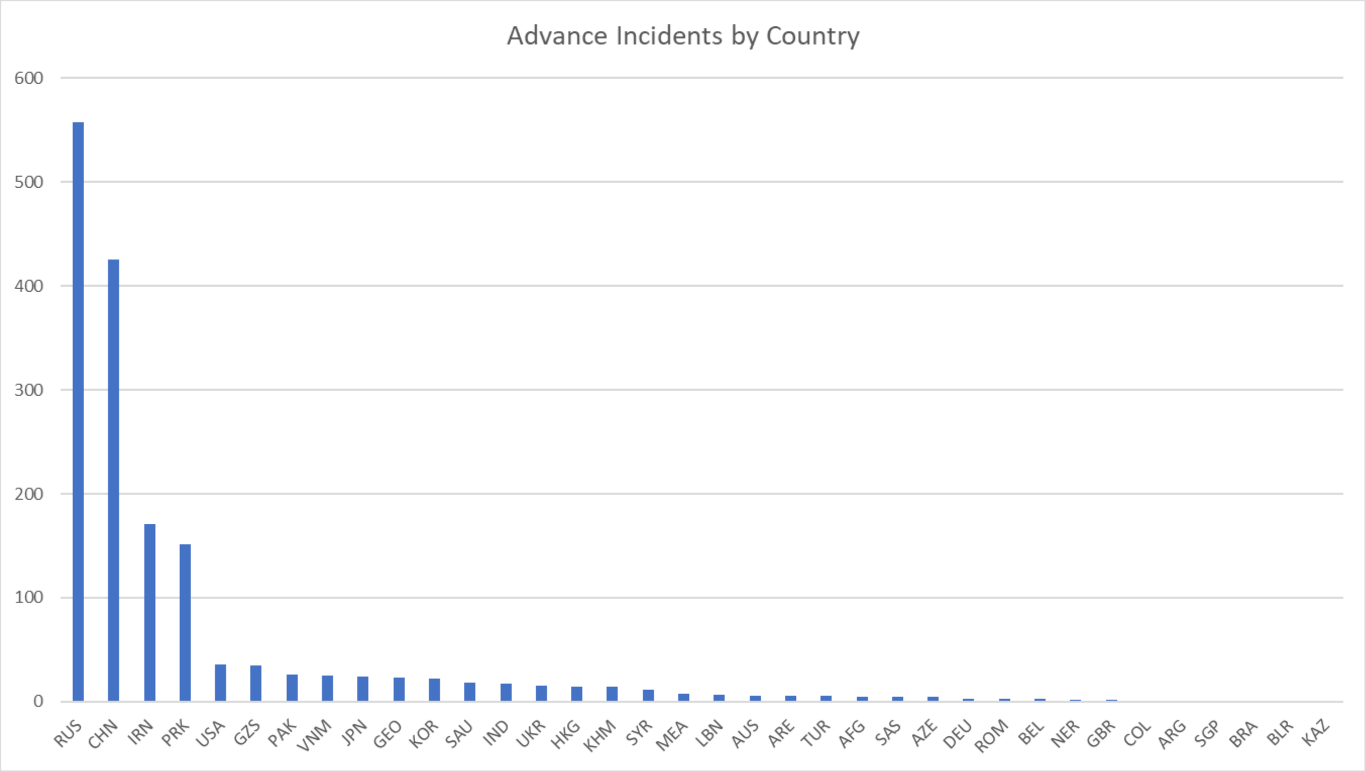

Figure 4 Incidents Distribution by Attacking Country

V. Geopolitical Context Dataset

In line with the second contribution, the new dataset aims to collect the geopolitical context of the incidents that compose the first dataset. For this purpose, we will use the BigData source GDELT. This source offers several databases that we must analyze in order to select the sources that most accurately describe the geopolitical context of the advanced incidents.

The first version of the GDELT Event Database[7] has 3.5 billion mentions included with an update interval of one day. Although this database allows the registration of events from all over the world, the analysis of this database shows that the largest number of events were registered in the United States territory due to the limited adoption in other countries.

The following figure shows the percentage of events registered by each country in the first version of GDELT Event Database:

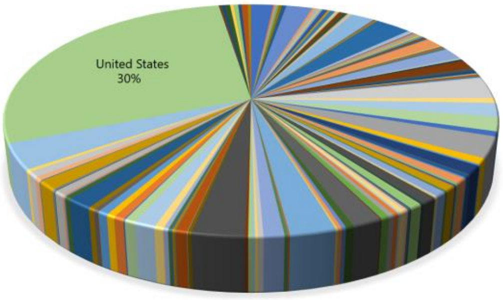

Figure 6 GDELT 1.0 Event Database total event distribution by country

As we can see, 30% of the incidents comes from United State, it is important to take this factor into account to avoid any bias caused by this phenomenon during the acquisition and use of these data.

In addition to the event database, GDELT offers the GKG[8] complement which records in addition the social context of events such as the number of victims of a catastrophe, name of people involved in the event, etc. This add-on does not include any metrics of interest in our analysis of the geopolitical context of advanced incidents.

We consider that social data do not play a key role in the geopolitical context of advanced incidents, so the use of GDELT will be limited to event databases.

In the second version, GDELT offers an updated Event Database[9]. It provides an even larger volume since the update interval is 15 minutes. The format of this data is common for both versions.

We found limitation in both versions for the construction of the dataset. In reference to the first version, the recorded events present a centralization in the United States due to the limited adoption, while the second version presents an initial date of registration from February 2015. Taking into account that the first incident collected in the advanced incident database is dated from 2008 as can be seen in the *"Figure 3 Advanced incidents distribution over time"* and the geopolitical context is intended to be collected from the fifth year prior to the detection of the incident, **the second version** of the event database **does not meet the time interval** requirement.

Therefore, a hybrid solution is proposed with the use of both versions, extracting the geopolitical data prior to 2015 from the first version of the events database, while the geopolitical data after that date will be obtained from the second version, thus taking advantage of the increased data volume and less centralization of the data.

There are two main ways to query the GDELT databases. Although all databases are publicly accessible, the most common way to query them is using Google's BigQuery. This service has several limitations:

Firstly, it only offers the second version of GDELT, and secondly, it is a limited use service and generates a cost for querying certain volumes of data. Looking at the volume of data to be analyzed, the estimate reaches 126.31TB, as can be seen in the annex *"GDELTY BigQuery Cost and Time estimation"* exceeding 125 times the service limit. The great advantage of this service is the speed of data access, which is why it is the most used procedure.

The alternative to this method is the direct download of the files that make up the selected databases and then processing them locally. This alternative would overcome the cost limitations of the service but would have a very significant impact on the execution time.

Studying the feasibility of these two forms of access, in case of using the BigQuery service we would have to adopt a hybrid solution applying both procedures, processing the first version of GDELT locally to acquire the geopolitical data prior to February 2015 together with the acquisition of the data after that date with the use of BigQuery service. This alternative requires a previous study of the target data volume in order to have control over the cost.

[7] "GDELT v1.0 Event Database" Documentation can be found at: "http://data.gdeltproject.org/documentation/GDELT-Data_Format_Codebook.pdf"

[8] "GDELT 1.0 GKG" Documentation can be found at: http://data.gdeltproject.org/documentation/GDELT-Global_Knowledge_Graph_Codebook.pdf

[9] "GDELT 2.0 Event Database" can be found at: https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/

The alternative to this procedure, as we have pointed out, would be the acquisition and processing of data from both versions locally with the consequent penalty in execution time.

Considering the benefits and limitations, both solutions have been developed, offering a choice of procedure. The following annex *"GDELTY BigQuery Cost and Time estimation"* shows us the cost and time estimation of the BigQuery service solution.

In addition to the preliminary estimation, the following annex *"BigQuery and Local Comparison"* offers a comparison of the real cost measurements, both monetary and the execution time of the BigQuery version versus Local procedure.

The result of this process will be the creation of a dataset that collects the geopolitical context over a period of 5 years prior to the incident detection including the year of detection over all country-pair relationships collected in the first advanced incident dataset.

The download process of all the files that compose both versions of the GDELT Event Database is described below.

A Python script has been developed to perform a web scraping process for the first version[10], making a download request for each source file, decompressing the file in CSV format and storing the file in the folder corresponding to its year for agility reasons in the access when processing this information.

The downloading process of the second version does not require web scraping but the processing of a master text file[11], this script performs a similar process to the previous one, making download requests and storing the source files in folders corresponding to the year.

Both scripts use concurrent processing techniques and parallelism to speed up execution. Despite this, the execution of the download of the first version takes approximately 10 hours in an environment with 24 threads and 100MB/s real download speed, while the download of the second version takes approximately 22 hours in the same environment.

Both scripts offer the possibility to update the database files starting from the last downloaded files. That way, we can update the remaining files if we have made a previous download.

The result of this process shows a disk occupation of 224 GB corresponding to the first version and 277 GB corresponding to the second version at the time of download.
As mentioned above, the representation of the geopolitical context will focus on the **Goldstein index**. Extracting the mean, which will represent the relationship between the attacking and affected country during the year under evaluation, the standard deviation and the percentage of events with a positive Goldstein index.

In addition to these metrics, the number of incidents that have been detected orchestrated by the attacking country towards the affected country will be evaluated, thus providing the evolution in the number of incidents together with the evolution of the geopolitical context.

The final format of this collection can be found in the annex "Datasets Documentation", specifically in the collections corresponding to the geopolitical context.

Several Python scripts will be developed for the creation of the geopolitical contexts dataset. First, a corresponding script for the local procedure for both GDELT versions. Secondly, a corresponding script for the processing of the GDELT second version using the BigQuery service.

The geopolitical contexts dataset also includes statistics from the first dataset. To obtain these statistics, multiple documents are generated in the MongoDB database, which will be consulted by the creation scripts of the second dataset. This way, with the analysis and storage of these statistics in a preliminary way, we save the runtime calculation of these statistics, transforming this operation into a significantly faster database query.

To calculate the total number of incidents carried out by the attacking country, the advanced incidents dataset is iteratively analyzed, and the incidents carried out by that country are counted, storing these statistics in a dictionary, which will be stored as a document in the MongoDB database. This document will have as a search key the CAMEO Country code of the attacking country.

Similar to the previous procedure, four more statistics have been calculated:

- Total incidents received by country
- Total incidents by country pair
- Total incidents by country pair per year

Focusing on the calculation of the percentage of positive and negative Goldstein mentions, the mean of this index and standard deviation, an analysis has been performed for the selection of a representative sample in order to reduce the volume of data requested to the BigQuery service, resulting in a **reduction** of the **cost** and a reduction of the local file **processing time**. The result of this process shows that the processing of **25,000 events** is sufficiently representative for all the proposed values.

[10] GDELT first version is provided in the following domain: "http://data.gdeltproject.org/events/index.html",

[11] GDELT second version is obtained from the following domain: "http://data.gdeltproject.org/gdeltv2/masterfilelist-translation.txt"

Therefore, both scripts start with the recovery of the incidents that make up the first dataset, collecting for each incident the relationship between pairs of countries and mainly the year of detection. For each relationship between the attacking country and the affected country, the five years prior to the detection of the incident plus the year of detection are analyzed. For each iteration, the attacking country, affected country, incident reference, incident year, and the previously documented statistical databases are included. In addition to these incident data, the calculation of the proposed measures for the representation of the geopolitical context is performed.

In both scripts, the processing of the first version is performed locally, so the script loads the files corresponding to the year being analyzed in each iteration. If an incident detected in the year 2018 is processed, the first year from which the metrics are extracted is the year 2013 which is part of the first version of GDELT. This way, the script loads the files of that year in a dataframe, in case they have not been previously calculated, since it is used **recycled from all calculations**. If the calculation of that year's metrics for the same pair of countries had been previously carried out, the data would be queried to the database, saving local processing or query and data processing in cases where the BigQuery service is used.

After loading the file, a filtering of the records corresponding to the attacking country and the affected country is performed. From the resulting dataframe, the calculation of the Goldstein positive events percentage, mean and standard deviation of the Goldstein index for the whole year is performed.

In case of processing the **second version locally**, the script logic is similar to the previous one, adding the selection of a representative sample of 10,000 files that make up this database, since the number of files per year averages 50,000 files, which reduced the execution time.

In the case of processing the **second version using the BigQuery service**, the technique of recycling the calculations is also applied to reduce the volume of data consulted and thus reduce the cost of the operation. A query is made corresponding to the events of the attacking country over those of the affected country during the year under analysis, corresponding to the fields "Actor1Code", "Actor2Code" and "Year" respectively, from which the same metrics previously exposed are obtained.

The operations have been optimized using only floating-point variables and the numpy library, which has a shorter execution time for these operations compared to the pandas library also used for the optimization of the data loading of local files.

All the above data is stored in a dictionary, to be loaded into the MongoDB database once all the metrics have been calculated between 2013 and 2018 corresponding to the example given. In this example, the years that would be processed using the first version would only be the year 2013 and 2014. The remaining years would be processed using the second version, which can be done locally or using the BigQuery service.

The result of this process is a collection with a total of **1,337 geopolitical contexts.**

With the validation of these data, we see that some documents lack the total number of years. This may be due to the unavailability of events that meet the query requirements, since most of the detected errors correspond to the first version of GDELT, which, as noted above, has some limitations. After removing these incomplete records using a Python script, we obtain a new collection, presenting a total of **1,253 geopolitical context documents.**

Additionally, a Python script is offered to analyze only the second version of GDELT, performing the same data processing on the years after 2014 and generating an collection of geopolitical contexts from that date onwards. This script aims to offer a much more agile retrieval method than the one developed previously. Due to the exclusive use of BigQuery for the acquisition of the data, it presents a significantly lower execution time as can be seen in the appendix *"BigQuery and Local Comparison"*.

The results of the execution of this procedure are currently incomplete, but this method may be useful if the objective is to analyze the incidents detected from the year 2020 onwards. The result of this process, although incomplete, is provided in a collection with a total of **3,427 documents**.

We can see that the comparison of execution time and cost of the version that acquires the data from the BigQuery service presents both costs and execution time significantly lower than estimated. This is due to the recycling of calculations, since it avoids consulting and processing again the calculations that have been previously done. Being repetitive calculations due to the similarity of the patterns followed by the threaten actors throughout the years. The savings are notorious.

VI. Model analysis

In this Section, the correlation results for geopolitical incidents are addressed. Thus, Section VI-A describes the correlation existing between both datasets. Section VI-B discusses the prediction results. Finally, the last Section provides with strengths and weaknesses of the research results.

A. Datasets Correlation Analysis

This section explains the process carried out to study the correlation between the data describing the geopolitical context of the incidents and their occurrence.

To study this correlation, several Multiple Linear Regression models were used, with the subsequent analysis of the Annova table to identify the contribution of the independent variables to the variable to be explained.

The first step in this process has been the generation of a spreadsheet from the complete geopolitical dataset, importing the total of the variables, which can be consulted in the annex "Datasets Documentation" to the Statgraphics statistical analysis tool.

The Multiple Linear Regression model will use as **dependent variable** the **number of incidents** between the attacking country and the affected country **during the year of detection** of the incidents.

The first Multiple Linear Regression model to be evaluated includes as independent variables all the variables of the second dataset excluding the CAMEO Country codes, the year of the incident, the incident ID and the percentage of negative incidents because it is linearly dependent on the percentage of positive incidents.
The first model shows a correlation of 78.12%. The analysis of the p-values and the weights assigned to each variable, which can be found in the annex "Initial Model", shows us that some independent variables have a low contribution to the model. We can see that both, means and standard deviations, show for some previous years a p-value higher than 0.05, which indicates that these variables have a limited correlation in some years. Analyzing the weights assigned to the variables, we see that global incident statistics such as the total number of incidents carried out by the attacking country or the total number of incidents received by the affected country are assigned very low weights, limiting the contribution of these variables to the model.

Making a new selection of independent variables, reducing these variables to the **Goldstein means**, **percentage of positive indexes** and **number of incidents in the range of five years** prior to the detection of the incident, together with the total number of incidents carried out by the attacking country to the affected country, shows a correlation of 76.74%, which is very close to the correlation of 78.12% obtained from the initial model. Taking into account the reduction of variables, we can consider it an improvement of the model. Likewise, we can continue to appreciate that the p-values of the Goldstein index means exceed the 0.05 threshold and the weights assigned to these variables are in most years very small.

The following study will be the previous model excluding the Goldstein index means of the independent variables.

Analyzing this third model, we obtain a correlation of 76.49% close to the previous one, showing that the mean of the Goldstein index has a very limited contribution to the model. This may be due to the stability of the data. Most of the countries show a mean of the Goldstein index centered at 0 with a standard deviation close to 1. Considering that these values have a range from -10 to 10, the values taken by these variables can be considered **very stable**.

Performing a holistic analysis of the dataset used and, in order to improve the correlation of these data, we perceived that the dataset included the incidents of the current year 2021, being these data a clear bias for having been collected during the beginning of the year, presenting incomplete incident statistics, as can be seen in the *"Figure 3 Advanced incidents distribution over time"*.

To solve this bias in the analysis, all the geopolitical contexts of the incidents detected during the year 2021 must be excluded. For this purpose, a new dataset excluding these records has been generated with **a total of 1026 documents.**

We will evaluate the two previously analyzed models on this new dataset.

The model that includes the averages of the Goldstein indexes shows a correlation of **86.60%**, notably higher than the equivalent model with the **bias** originated by the current year.

The model that excludes the Goldstein index averages shows a correlation of 86.55%, even with the previous model, again showing that the Goldstein index means have a very small contribution.

Considering that the correlation obtained is higher than 85%, we can conclude that the geopolitical data extracted can predict 86.55% of the number of incidents that will be carried out by one country on another in a given period of time.

All evaluated models can be consulted in the annex *"Evaluation of Multiple linear regression models",* which includes the Annova table for each analysis, as well as the final weighting model for each analysis.

B. *Prediction Analysis*

We already know that the correlation between the data representing the geopolitical context and the advanced incidents reaches 86.55% in the case of the last model evaluated.

To perform an in-depth analysis of the predictions that can be made with this model, we must first analyze the data used to configure the model, as well as the results of these predictions.

First, it should be noted that the prediction model uses both the number of incidents perpetrated by the attacking country for each year in the 5-year range prior to the detection of the incident, and the percentage of events with a positive Goldstein index during the same period.

In order to study the error incurred by the linear regression model from the same perspective of attacking countries, we have made a boxplot representing the normal distribution of the absolute value of the difference between the number of actual incidents and the number of incidents predicted by attacking countries as shown in the *"Figure 7 Prediction Error Distribution by Attacking Country":*

The area between the first and third quartile, which corresponds to the colored area, represents 50% of the error made by the prediction, while the limits established by the lines represent 25% of the error distributed on each side; points exceeding these limits are considered outliers.

Thus, we can see that the countries with the highest volume of incidents have similar error values, showing an average error of around 1.2 incidents in the prediction, and concentrating 75% of the error between 0-2 incidents error. This measure means that in 75% of the predictions corresponding to these countries, the prediction can be wrong between 0 and 2 incidents, the average error being 1.2 incidents.

The overall mean of the model is 1.626 incidents, which indicates that the prediction is more reliable for countries with a higher volume of detected incidents. On the other hand, the limit of the standard values is below 6 incidents, and some outliers can be observed above 6 incidents, establishing a maximum limit for the error in the countries with the highest volume of incidents below 6 incidents, also below the global maximum limit of 6.65 incidents as error.

On the other hand, we are also aware of the temporal distribution of the incidents detected, shown in the *"Figure 3 Advanced incidents distribution over time".*

This distribution shows a higher volume of incidents in recent years due to the exponential increase in the activity of State Actors in the cyberspace.

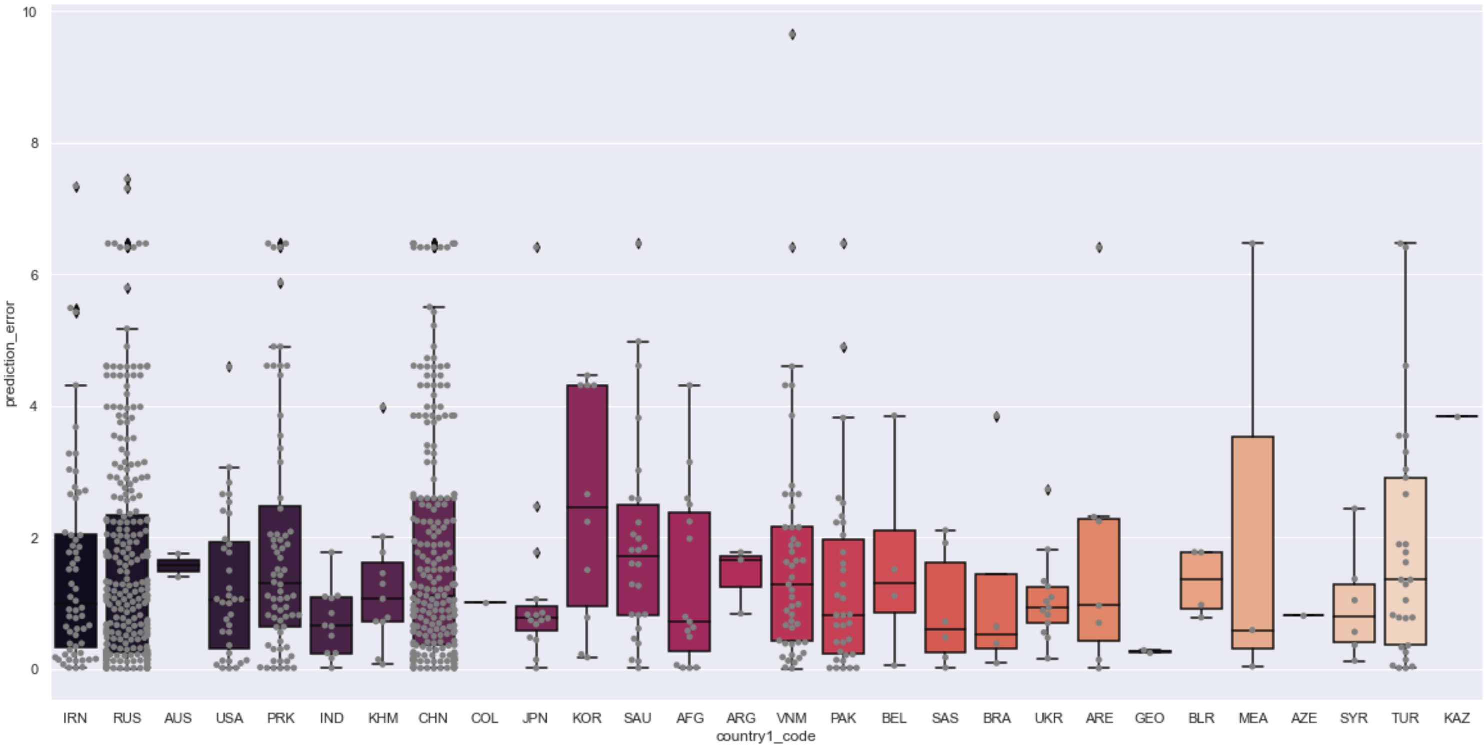

Figure 7 Prediction Error Distribution by Attacking Country

Developing the same view as above from a chronological approach, we can observe the evolution of the committed error by the predictive model over time:

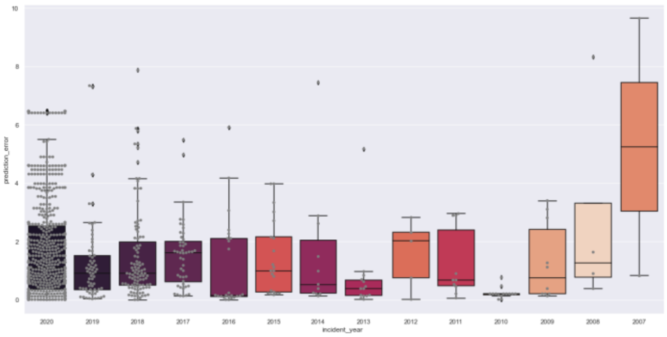

Figure 8 Prediction Error Chronologic Distribution

These results are shown in chronological order, starting from 2020 on the left to 2007 on the right.

We can clearly see the increase of the error proportionally in the year 2007, this may be due to the limitations of the first version of GDELT Event Database, from which the data used for the prediction of that year has been obtained, together with the lower volume of incidents detected in the early years.

It is interesting to note that the error in the prediction can come from both a deviation in the activity of a country over a period of time, as well as the discontinuity in the detection of such incidents. It is important to keep in mind that these data come from a process as complex as the detection of Advanced Persistent Threats. We could obtain a high error rate due to an increase in the complexity of these threats, which would reduce the incident detection rate, as well as a discontinuity in the activity of these threats.

Analyzing the prediction errors using the second version of GDELT Event Database, we obtain the following graph:

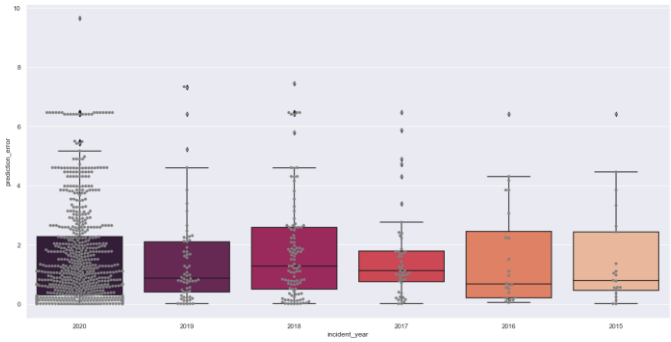

Figure 8 GDELT 2.0 Event Database Prediction Error Distribution

As can be seen, 75% of the error would be between 0 and 2 incidents, with an average of approximately one incident, a similar measure obtained in the errors of the countries with the highest incident volumes. Likewise, the maximum limit is below 5 incidents, leading to the conclusion that the use of the second version of GDELT offers a reduction in the prediction error, compared to the global average of 1.626 incidents with a maximum limit of 6.65 incidents.

It is also interesting to carry out a continent-by-continent analysis to study if the accuracy of the prediction model depends on the region.

To do so, we will plot the prediction error between pairs of countries on the geographical map to visualize the error by region generating the *"Figure 10 Prediction Error Geographic Distribution by Country Pairs"*.

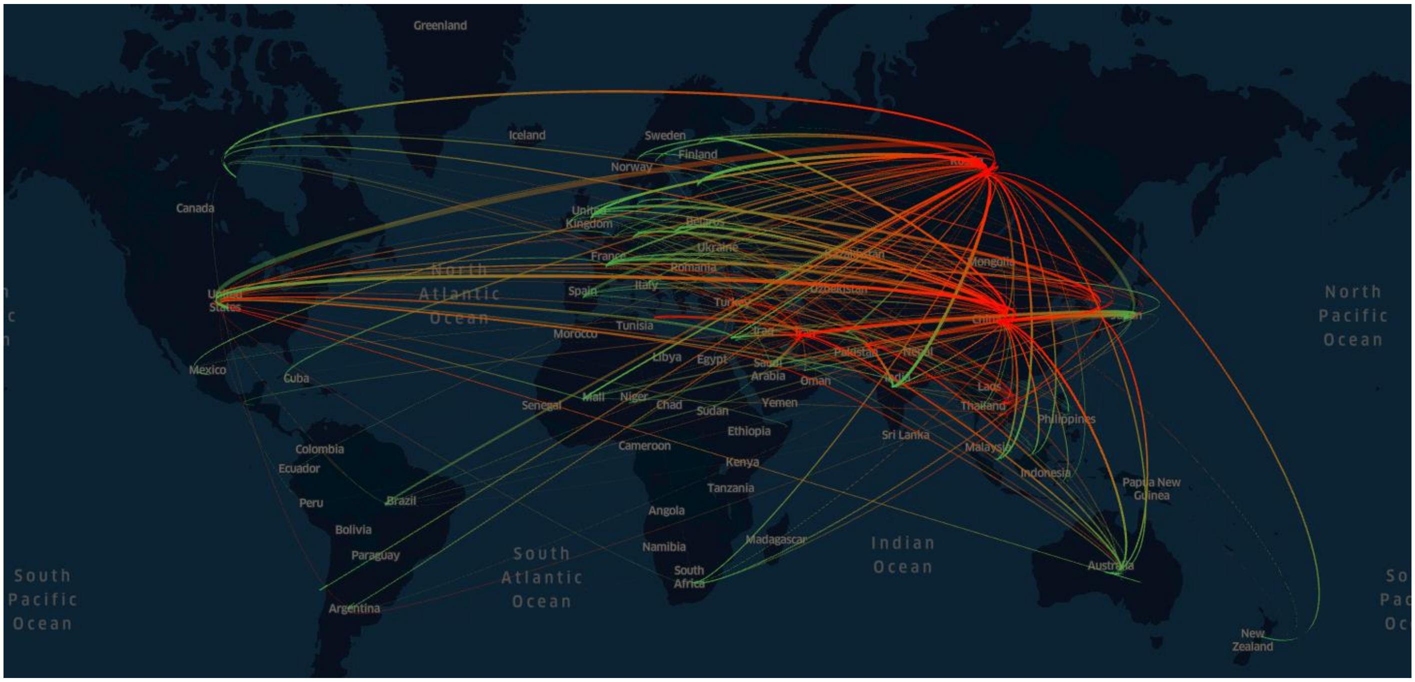

Figure 9 Prediction Error Geographic Distribution by Country Pairs

In this graph, each arc represents a relationship between pairs of countries, the red end represents the attacking country, while the green end represents the victim country, likewise, the thickness of the arc represents the error, with the thicker arcs being the largest errors made by the prediction model.

In addition, a heat map has been created showing this same information from another perspective. This map can be found in the annex *"Prediction Error HeatMap by Regions"*.

It is observed that the geographic distribution not reveal a clear pattern to determine if the prediction model obtains different results depending on the region using this view.

In addition, to visualize the error committed by the prediction in each continent, we will use a boxplot plot representing the distribution of the error committed grouped by continents corresponding to the attacking countries.

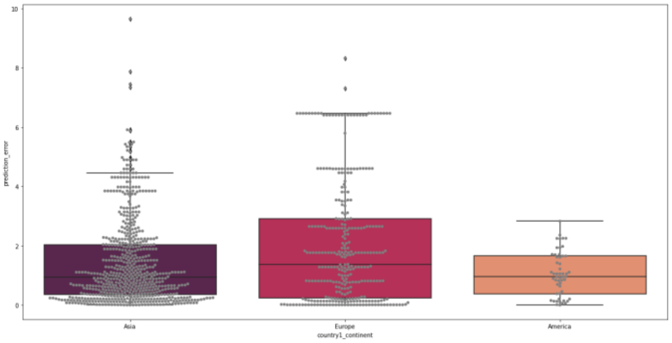

Figure 11 Prediction Error Distribution by Continent

We can see in the graph a higher error in the prediction of the attacking countries in the European region, being also visible a deviation from the maximum limit in the error, exceeding 6 incidents. This is due to the contribution in the European region of countries with a very small number of incidents, as can be seen in the figure "Figure 5 Incidents Distribution by Attacking Country", which penalizes the prediction in this area due to the error made in the prediction of State Actors with a small number of incidents. In contrast, Asia has a higher volume of incidents distributed in State Actors with a higher attribution volume than the average for European countries. Specifically, the volume of incidents recorded in Asia amounts to 654, distributed mainly among China, North Korea and Iran, while Europe has a total of 317 incidents attributed mainly to Russia.

In conclusion, the overall average error of the prediction model is 1.626 incidents, we observe a reduction of this error in the case of the countries with the highest activity, which is positive, since they will be the countries with the greatest interest in the analysis, showing an average error centered on 1.2 incidents and a maximum error limit of less than 6 incidents. Similarly, the analysis of the error in a chronological way shows that the use of the GDELT 2.0 Event Database version improves the prediction of the model, centering the error in 1 incident, with a maximum error limit lower than 5 incidents. This presents an optimistic conclusion for the application of this model, as it is reinforced by the increase in the volume of incidents and the use of the latest version of GDELT.

C. Discussion

In this section, we will analyze the strengths and weaknesses of the research, contextualizing the results obtained.

First of all, regarding the advanced incidents dataset, we can observe results similar to those reported by the CCN-Cert, both in the chronological evolution of the number of incidents and in the predominance of certain nation-state actors [2] [1] [3] [13].

Even having obtained satisfactory results contributing to the normalization and publication of said dataset, we should not ignore that due to the volume of data proposed to be analyzed and due to the automated solutions applied to solve the problem, there could be errors in the extraction of the data that make up the dataset of advanced incidents. This has been observed during the validation of this data, with the consequent correction of the errors found mainly in the extraction of countries. This can be considered an intrinsic weakness of the procedure used to solve the problem.

This error could be propagated to the second dataset, which captures the geopolitical context of such incidents. If this propagation takes place, finally taking part of the dataset, could impact in the results of the Multiple Linear Regression model. This counter can be confronted, as in this case, with an exhaustive validation of the first dataset.

Several Python scripts for the detection of possible errors in the incident dataset are provided to facilitate this task.

Regarding the second dataset, we must take into account that these data come from the GDELT BigData source, which as previously discussed, sometimes presents event classification errors. These errors can impact the representation of the geopolitical context of the incidents, but we consider that the GDELT data source is the most indicated for the acquisition of these data due to its open access nature and presenting the **largest volume of events** with a high level of decentralization in the latest version.

VII. Analytical predictive tool

The analytical-predictive tool is composed of two main components. The first component is in charge of providing the analysis of advanced incidents. The second component focuses on predicting the occurrence of advanced incidents under certain conditions.

This contribution is materialized in a reactive web application that will provide an interface to interact with both components. The backend of the web application has been developed with Python, offering a public API for querying all the datasets described in the annex "Datasets Documentation". The frontend has been developed with React in a DashBoard format.

The analytical component has been developed for the user to perform queries directly on the datasets offered. This tool allows the user to access directly to a specific data that can be extracted from these datasets. This component also allows the user to generate different views of the data without having to download and process it locally.

An additional feature allows the visualization of datasets in interactive maps. These datasets must meet the requirement of including geographic coordinates. These fields can be consulted in the annex "Datasets Documentation".

The prediction component provides an interface for predicting the number of incidents under certain conditions, determined by year ranges, attacking country and affected countries.

VIII.Related Work

This section focuses on contextualizing the present research, comparing the contributions provided with the previous state of the art.

As previously mentioned, there was no public dataset compiling advanced incidents attributed to nation-state actors. The contribution provided satisfies this lack by also including the temporal context and the relationship of countries involved in the incidents.
Regarding the second contribution and due to the close relationship between this one and the first contribution, there was no dataset that collected the geopolitical context of the advanced incidents. The solution adopted for this contribution allows the analysis of relations between countries over a period of five years prior to the detection of the incident, described by the Goldstein index of events between the countries involved in the incident.

There are several studies that apply the analysis of the GDELT data source for incident prevention, focusing mainly on the prevention of civil risks, events of social unrest or armed conflicts. At the international level, a comparison of accuracy between an HMM model trained from the GDELT events database with an approximate study period of 30 years on five major nations in Southeast Asia and a logistic regression model (LogReg) was carried out, obtaining an increase in accuracy in the prediction of the first model, higher than the prediction method with logistic regression (LogReg) [6].

At the European level, a remarkable effort has been made for the creation of a dynamic Global Conflict Risk Index (GCRI), which uses datasets from various sources such as GDELT, ICEWS and OEDA for the detection of possible triggers of violent conflicts. This study marks the feasibility of using GDELT for conflict prediction while showing the limitations of the data sources used [4]. Despite the existence of errors in the classification of news, GDELT presents greater granularity over time, greater integration of languages, larger volumes and origins of data, while ICEWS and OEDA sources show a limitation of access and discontinuity over time from 2015 respectively, coinciding with comparisons between these data sources [7] [8].

Regarding Advanced Persistent Threat detection tools, we see that this class of solutions is limited to the analysis of infrastructures. At the national level, the most prominent solution is the "Carmen" tool, managed by the CCN-CERT [9]. This tool automatically analyzes the traffic of target networks to detect data exfiltration or communications with C2 (Command & Control) systems. This tool stands out for its ability to correlate events between different critical networks in order to detect simultaneous threats or in close time periods. Regarding the use of geopolitical analysis for the detection of this kind of threat, the number of publications focused on the development of a tool that applies this approach is extremely limited. No studies have been found on the development of an analytical-predictive tool based on the analysis of geopolitical data for the detection or estimation of the risk of advanced persistent threats.

The closest studies found for the detection of any kind of cyber-attacks uses as main data source Twitter social media together with machine learning algorithms. In particular, the prediction of cyberattacks has been studied with Bayesian Networks fed with the number of raw tweets containing the name of the type of attack as a key word [14]. Another approach seen is the use of L1 Regularization for post in Twitter sentiment classification, in order to use threshold in the coefficient of determination for cyberattack prediction [15]. Finally, a study adds to the Twitter data source, other data sources such as Facebook, Instagram, Blogs, Forums and from the internal IDS, analyzing this data with Naïve Bayes Classifier, obtaining a predictive accuracy of 63.33% [16].

IX. Conclusion

Besides satisfying all the proposed contributions, the present work constitutes a remarkable enrichment allowing the student to acquire skills in BigData, optimization of costly tasks in execution time, creation of reactive applications, among others, due to the multidisciplinary nature of the project.

The first contribution of this study is the publication of a dataset with the collection of advanced incidents attributed to State Actors. This dataset finally offers 1,654 incidents in normalized format, recording both the countries and groups responsible for the incidents, as well as the countries affected and their temporal context. This dataset can be accessed publicly, being also able to download this dataset in the web application when it is deployed in production.

The second contribution includes the creation of different datasets for the representation of the geopolitical context of the advanced incidents that conform the previous dataset. These datasets provide insight into the relationship between the attacking and victim countries over a period of 5 years prior to the detection of the incidents. Obtained from the BigData GDELT source, these datasets maintain different characteristics as can be seen in the appendix “Datasets Documentation”, highlighting the dataset that gathers the complete context of 1,253 advanced incidents, being necessary the query and processing of more than **126TB of data** concerning geopolitical events in the context of advanced incidents.

The correlation analysis between the above-mentioned datasets concludes with a correlation of **86.55%** when evaluating the evolution of the incidents and the geopolitical context of the previous 5 years to the incident detection. The **error incurred** by the model was reduced to **1.2 incidents** with a **maximum limit of 6 incidents**, in the prediction of the incidents perpetrated by the countries with the highest volume of incidents, these being the most belligerent countries, as well as reducing the **error** to **1 incident** with a maximum limit of 5 incidents, with the use of the **latest version of GDELT**, both factors marking an optimistic conclusion for the use of this procedure for the detection and prevention of Advanced Persistent Threats.

Finally, the creation of a reactive web application to improve the use of the datasets offered, as well as the use of the prediction model. This web application, in addition to allowing queries on the generated datasets, offers the possibility of creating numerous views of the data, even represented geographically on interactive maps. The predictive component of the tool allows us to make predictions of the number of incidents between pairs of countries over a given period.

As a compilation, these predictions are 86.55% accurate, with an error of 1.646 incidents globally or 1.2 incidents in the case of the most belligerent countries, with a maximum error limit of 6 incidents.

In case of using the latest version of GDELT for the whole geopolitical context, the error is centered on 1 incident with a maximum limit of 5 incidents. It should be noted that future predictions will make exclusive use of the second version of GDELT. Therefore, the predictions made by the prediction model present an **error centered on 1 incident** with a **maximum limit of 5 incidents**.

As future lines, this research can consequently generate new lines of study with the analysis of different prediction models that improve the predictive capacity of the data sets generated, with the development of specific predictive models for more specific Actor-States or conflict situations, even with the enrichment of the databases generated, adding variables such as available resources, acquisitive level of the State Actors, State maturity, etc.

X. References

[1] CCN-CERT, "Ciberamenazas y Tendencias, Edición 2020," Centro Criptológico Nacional, 2020.

[2] CCN-CERT, "Ciberamenazas y Tendencias, Edicción 2018," Centro Criptológico Nacional, 2018.

[3] CCN-CERT, "Ciberamenazas y Tendencias, Edicción 2019," Centro Criptológico Nacional, 2019.

[4] European Comission, "Dynamic Global Conflict Risk Index," Publications Office of the European Union, Luxembourg, 2019.

[5] MCCD, "Amenazas Persistentes Avanzadas (APT) como medida de disuasión en el ciberespacio," Instituto Español de Estudios Estratégicos, 2020.

[6] College of Information Systems and Management, National University of Defense Technology, Changsha, Hunan 410073, China, "Predicting Social Unrest Events with Hidden Markov Models," Hindawi, Hunan, 2017.

[7] P. C. R. a. D. R. Kishi, "COMPARING CONFLICT DATA, SIMILARITIES AND DIFFERENCES ACROSS CONFLICT DATASETS," ACLED, 2019.

[8] M. D. W. &. A. B. &. J. C. &. M. D. &. C. D. &. B. Radford, "Comparing GDELT and ICEWS Event Data," ResearchGate, 2013.

[9] CCN-CERT, "Carmen, Detección de ataques avanzados / APT," Centro Criptológico Nacional, 2019.

[10] Python, "Dateparser Documentation," [Online]. Available: https://pypi.org/project/dateparser/. [Accessed June 2021].

[11] Python, "PyCountry Documentation," [Online]. Available: https://pypi.org/project/pycountry/.

[12] GDELT, "CAMEO MANUAL," [Online]. Available: http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf. [Accessed June 2021].

[13] Ministerio de la Presidencia, Relaciones con las Cortes y Memoria Democrática, "Informe Anual, Edicción 2020," Gobierno de España, 2020.

[14] Computer Engineering Rochester Institute of Technology Rochester, NY, USA, "Predicting Cyber Attacks With Bayesian Networks Using Unconventional Signals," CISCR, New York, 2017.

[15] ESIME Culhuacan, University of Warwick, "Social Sentiment Sensor in Twitter for Predicting Cyber-Attacks Using L1 Regularization," Sensors, 2018.

[16] George Onoh, Bowie State University, Maryland, "Predicting Cyber-Attacks Using Publicly Available Data," CISSE, 2018.

[17] Department of Computer Science, Durham University, Durham, UK, "Detection of advanced persistent threat using machine-learning," ELSEVIER, 2018.

[18] Department of Engineering “Enzo Ferrari”, University of Modena and Reggio Emilia, Italy, "Analysis of high volumes of network traffic for Advanced Persistent Threat detection," ELSEVIER, 2016.

[19] European Commission, Joint Research Centre (JRC), "Conflict Event Modelling: Research Experiment and Event Data Limitations," European Commission, Luxembourg, 2020.

[20] S. Sonjai, "Measuring Economic Impact of Political Protest by Using The Global Database of Events Languages and Tone (GDELT)," Chulalongkorn University, 2017.

[21] V. Prenosil, "Advanced Persistent Threat Attack Detection," Vaclav Prenosil, Masaryk University, 2015.

[22] J. M. L. Diego J. Bodas-Sagi, "Using GDELT Data to Evaluate the Confidence on the Spanish Government Energy Policy," Madrid, 2017.

[23] J. E. YONAMINE, "PREDICTING FUTURE LEVELS OF VIOLENCE IN AFGHANISTAN DISTRICTS USING GDELT," 2014.

[24] Department of Department of Political Science, Duke University, "Automated Learning of Event Coding Dictionaries for Novel Domains with an Application to Cyberspace," 2016.

[25] CCN-CERT, [Online]. Available: https://www.ccn-cert.cni.es/. [Accessed June 2021].

[26] ThaiCert, [Online]. Available: https://www.thaicert.or.th/.

[27] MITRE, [Online]. Available: https://www.mitre.org/. [Accessed June 2021].

[28] Numpy, [Online]. Available: https://numpy.org/. [Accessed June 2021].

[29] Pandas, [Online]. Available: https://pandas.pydata.org/. [Accessed June 2021].

[30] P. A. Schrodt, "Conflict and Mediation Event Observations Event and Actor Codebook," Pennsylvania State University, Pennsylvania, 2012.

[31] Fireeye, [Online]. Available: https://www.fireeye.com/. [Accessed June 2021].

[32] GDELT, "The GDELT Project," [Online]. Available: https://www.gdeltproject.org/. [Accessed May 2021].

[33] Google, "BigQuery," [Online]. Available: https://cloud.google.com/bigquery/docs.

XI. Anexes

1. GDELTY BigQuery Cost and Time estimation

Number of incidents	1654
Country pair combinations	5375
Queries per combination	5
Query size estimation (GB)	4.7
Query time estimation (seconds)	0.3
	Data Volume
GB	126312.5
	Estimated Cost
euro per TB	5
Total	631.5625
	Estimated Time
seconds	8062.5
hours	2.239583333

Table 1 BigQuery Cost and Time estimation

2. BigQuery and Local Comparison

GDELT Version			
GDELT 1.0	Local	Local	None
GDELT 2.0	Local	BigQuery	BigQuery
Comparison			
Execution time (hours)	52.3	15.1	0.3 (Incomplete)
Cost(euro)	0	389	389
Enviroment			
Disk transfer rate(MB/s)	3500		
Internet transfer rate (MB/s)	100		
Number of Threats	24		

Table 2 BigQuery and Local Comparison

3. Datasets Documentation

3.1. Incident Datasets

Collection	Fields										
	'group_name'	'group_country'	'reference'	'operation_date'	'operation_month'	'operation_year'	'operation_description'	'operation_urls'	'operation_countries'	'cameo_country'	'cameo_operation_countries'
APT_Group_Incidents	X	X	X	X	X	X	X	X	X		
APT_Group_Incidents_CAMEO	X	X	X	X	X	X	X	X	X	X	X

Table 3 Incident Datasets Field Structure

3.1.1. Fields Description

- **Group_name:** Name of the advanced group (these names are attributed by the Certs and research groups. The same group can have several associated names attributed by several entities).
- **Group_country:** Country attributed to the advanced group.
- **Reference:** Url to MITRE or ThaiCert source referring to the advanced group.
- **Operation_date:** Incident detection date as extracted from the data source (this field is in raw format).
- **Operation_month:** Month extracted from the date of detection of the incident in raw format (this data may not exist, in case of not acquiring this data, the value that the variable will take is "None").
- **Operation_year:** Year extracted from the date of detection of the incident.
- **Operation_description:** Incident description.
- **Operation_urls:** List of references of the sources reporting the incident.
- **Operation_countries:** List of countries affected by the incident.
- **Cameo_country:** CAMEO Country code of the country responsible for the incident.
- **Cameo_operation_countries:** List of CAMEO Country codes of the affected countries of the incident.

3.1.2. Datasets Description

- **APT_Group_Incidents:** Records the advanced incidents extracted from MITRE and ThaiCert data sources as described in chapter "IV. ADVANCED INCIDENTS DATASET".
- **APT_Group_Incidents_CAMEO:** Generated from the previous dataset, it includes the CAMEO format codes of the attacking country and the affected countries.

3.2. GEO-Political Context Datasets

Collection	Fields											
	'country1_code'	'country2_code'	'incident_id'	'incident_year'	'country1_snt_incidents'	'contry2_rvd_incidents'	'country1_to_country2_incidents'	'incidents_X'	'positive_goldstein_mentions_X'	'negative_goldstein_mentions_X'	'golds_mean_X'	'golds_std_X'
APT_GEO_Context	X	X	X	X	X	X	X		X	X	X	X
APT_GEO_Context_complete	X	X	X	X	X	X	X	X	X	X	X	X
APT_GEO_Context_2021_missing	X	X	X	X	X	X	X	X	X	X	X	X
APT_GEO_Context_raw	X	X	X	X	X	X	X		X	X	X	X
APT_GEO_Context_v2	X	X	X	X	X	X	X		X	X	X	X

Table 4 GEO-Political Context Dataset Field Structure

3.2.1. Fields Description

- **Country1_code:** CAMEO Country code of the country responsible for the incident.

- **Country2_code:** CAMEO Country code of the country affected by the incident.

- **Incident_id:** Reference to the incident under analysis.

- **Incident_year:** Year of incident detection.

- **Country1_snt_incidents:** Total Number of incidents perpetrated by the first country.

- **Country2_rvd_incidents:** Total Number of incidents received by the second country.

- **Country1_to_contry2_incidents:** Number of incidents from the first country to the second country.

- **Positive_goldstein_mentions_X:** Percentage of number of events between the first and second country with a positive Goldstein index during the “X“ year prior to the incident.

- **Negative_goldstein_mentions_X:** Percentage of number of events between the first and second country with negative Goldstein index during the “X“ year prior to the incident.

- **Golds_mean_X:** Goldstein mean of events between the first and second country during the “X” year prior to the incident.

- **Golds_std_X:** Goldstein Standard deviation of events between the first and second country during the “X” year prior to the incident.

- **Incidents_X:** Number of incidents detected during the “X” year prior to the incident between country_1 and country_2.

The values of the variable "X" take values between zero and five representing the metrics corresponding to the years prior to the detection of the incident. Thus, the variable "Golds_mean_5" represents the average of the Goldstein indexes of the events between the attacking country and the victim country during the fifth year prior to the detection of the incident.

3.2.2. Datasets Description

- **APT_GEO_Context:** Records the initial geo-political context extraction from GDELT as described in chapter "V. GEOPOLITICAL CONTEXT DATASET".

- **APT_GEO_Context_raw:** Generated from the previous dataset, this dataset excludes errors detected in the initial dataset.

- **APT_GEO_Context_complete:** Generated from the previous dataset, it deletes records that do not have all fields, thus excluding incomplete contexts.

- **APT_GEO_Context_2021_missing:** This dataset is a copy of "APT_GEO_Context_complete" excluding the contexts of the incidents detected in 2021.

- **APT_GEO_Context_v2:** This dataset has been generated using only the GDELT Events v2.0 version for the creation of the geo-political contexts.

3.3. Analysis Datasets

3.3.1. Analysis by Advanced Group Datasets

Collection	Fields						
	'group_name'	'country'	'country_code'	'latitude'	'longitude'	'incidents_sent'	'year_list'
APT_Total_incident_by_group	X	X	X	X	X	X	X

Table 5 Analysis by Advanced Group Datasets Field Structure

- **APT_Total_incident_by_group:** This dataset records the incidents perpetrated by the advanced groups detected.

3.3.2. Analysis by Attacking Country Datasets

Collection	Fields					
	'country'	'country_code'	'latitude'	'longitude'	'incidents_sent'	year'
APT_Total_incident_sent_by_country		X			X	
APT_Total_incident_sent_by_country_table	X	X	X	X	X	
APT_Total_incidents_sent_by_years		X			X	X
APT_Total_incidents_sent_by_years_table	X	X	X	X	X	X

Table 6 Analysis by Attacking Country Datasets Field Structure

- **APT_Total_incident_sent_by_country:** This dataset collects the incidents perpetrated by each country in a single document.

- **APT_Total_incident_sent_by_country_table:** This dataset collects all the incidents perpetrated by each country including the geographical coordinates in order to represent the information on the geographical map.

- **APT_Total_incidents_sent_by_years:** This dataset records all incidents perpetrated by country for each year.

- **APT_Total_incidents_sent_by_years_table:** This dataset collects all the incidents perpetrated by each country in each year including the geographical coordinates in order to represent the information on the geographical map.

3.3.3. Analysis by Affected Country Datasets

Collection	Fields				
	'country'	'country_code'	'latitude'	'longitude'	'incidents_received'
APT_Total_incidents_received_by_country		X			X
APT_Total_incident_received_by_country_table	X	X	X	X	X

Table 7 Analysis by Affected Country Datasets Field Structure

- **APT_Total_incidents_received_by_country:** This dataset collects the incidents received by each country in a single document.

- **APT_Total_incident_received_by_country_table:** This dataset collects all the incidents received by each country including the geographical coordinates in order to represent the information on the geographical map.

3.3.4. Analysis by Country Pairs Datasets

Collection	Fields									
	'year'	'country1'	'country1_code'	'country1_latitude'	'country1_longitude'	'country2'	'country2_code'	'country2_latitude'	'country2_longitude'	'incidents'
APT_Total_incidents_by_country_pairs			X				X			X
APT_Total_incidents_by_country_pairs_table		X	X	X	X	X	X	X	X	X
APT_Total_incidents_by_country_pairs_years	X		X				X			X
APT_Total_incidents_by_country_pairs_years_table	X	X	X	X	X	X	X	X	X	X

Table 8 Analysis by Country Pairs Datasets Field Structure

- **APT_Total_incidents_by_country_pairs:** This dataset collects the incidents sent and received by each country pair in a single document.

- **APT_Total_incidents_by_country_pairs_table:** This dataset collects all the incidents sent and received by each country pair including the geographical coordinates in order to represent the information on the geographical map.

- **APT_Total_incidents_by_country_pairs_years:** This dataset collects the incidents sent and received by each country pair for each year in a single document.

- **APT_Total_incidents_by_country_pairs_years_table:** This dataset collects all the incidents sent and received by each country pair for each year including the geographical coordinates in order to represent the information on the geographical map.

4. *Evaluation of Multiple linear regression models*

4.1. Initial Model

Independent Variables

		Standard	T	
Parameter	Estimate	Error	Statistic	P-Value
CONSTANT	-1,98634	2,01061	-0,98793	0,3232
country1_snt_incidents	0,00295428	0,00058523	5,04807	0,0000
contry2_rvd_incidents	-0,0000326161	0,000857358	-0,0380426	0,9696
country1_to_country2_incidents	0,125448	0,00818252	15,3312	0,0000
positive_goldstein_mentions_0	4,60845	1,67777	2,74677	0,0060
golds_mean_0	-0,231821	0,137594	-1,68483	0,0920
golds_std_0	0,399458	0,12771	3,12785	0,0018
incidents_1	0,442818	0,0380567	11,6357	0,0000
positive_goldstein_mentions_1	1,5361	1,94867	0,788281	0,4305
golds_mean_1	-0,0118332	0,168599	-0,0701852	0,9440
golds_std_1	0,0639229	0,156848	0,407547	0,6836
incidents_2	-0,149263	0,0436561	-3,41905	0,0006
positive_goldstein_mentions_2	-13,9159	2,08055	-6,68853	0,0000
golds_mean_2	0,500854	0,16265	3,07933	0,0021
golds_std_2	-0,651703	0,174697	-3,73048	0,0002
incidents_3	0,0639364	0,0388788	1,6445	0,1001
positive_goldstein_mentions_3	9,41383	2,15968	4,3589	0,0000
golds_mean_3	-0,349722	0,168174	-2,07952	0,0376
golds_std_3	0,32755	0,159372	2,05525	0,0399
incidents_4	-0,198536	0,0370082	-5,36465	0,0000
positive_goldstein_mentions_4	0,675938	2,36051	0,286352	0,7746
golds_mean_4	-0,395877	0,170844	-2,31719	0,0205
golds_std_4	0,262823	0,140695	1,86804	0,0618
incidents_5	-0,700852	0,0424277	-16,5187	0,0000
positive_goldstein_mentions_5	0,687521	1,86642	0,368362	0,7126
golds_mean_5	0,0241221	0,137558	0,17536	0,8608
golds_std_5	0,00498008	0,123919	0,040188	0,9679

Table 9 Initial Model - Independent Variables

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	34322,7	26	1320,1	168,39	0,0000
Residual	9611,11	1226	7,8394		
Total (Corr.)	43933,8	1252			

Table 10 Initial Model - Analysis of Variance

R-squared = 78,1237 percent
R-squared (adjusted for d.f.) = 77,6597 percent
Standard Error of Est. = 2,79989
Mean absolute error = 1,90065
Durbin-Watson statistic = 1,59203 (P=0,0000)
Lag 1 residual autocorrelation = 0,202946

Model

incidents_0 = -1,98634 + 0,00295428*country1_snt_incidents - 0,0000326161*contry2_rvd_incidents + 0,125448*country1_to_country2_incidents + 4,60845*positive_goldstein_mentions_0 - 0,231821*golds_mean_0 + 0,399458*golds_std_0 + 0,442818*incidents_1 + 1,5361*positive_goldstein_mentions_1 - 0,0118332*golds_mean_1 + 0,0639229*golds_std_1 - 0,149263*incidents_2 - 13,9159*positive_goldstein_mentions_2 + 0,500854*golds_mean_2 - 0,651703*golds_std_2 + 0,0639364*incidents_3 + 9,41383*positive_goldstein_mentions_3 - 0,349722*golds_mean_3 + 0,32755*golds_std_3 - 0,198536*incidents_4 + 0,675938*positive_goldstein_mentions_4 - 0,395877*golds_mean_4 + 0,262823*golds_std_4 - 0,700852*incidents_5 + 0,687521*positive_goldstein_mentions_5 + 0,0241221*golds_mean_5 + 0,00498008*golds_std_5

4.2. First Model Improvement

Independent Variables

		Standard	T	
Parameter	Estimate	Error	Statistic	P-Value
CONSTANT	2,66816	1,3407	1,99011	0,0466
country1_to_country2_incidents	0,131801	0,0073597	17,9085	0,0000
positive_goldstein_mentions_0	2,05203	1,54982	1,32404	0,1855
golds_mean_0	-0,148818	0,138134	-1,07735	0,2813
incidents_1	0,536892	0,0360876	14,8774	0,0000
positive_goldstein_mentions_1	1,69052	1,77314	0,9534	0,3404
golds_mean_1	-0,0387612	0,1665	-0,2328	0,8159
incidents_2	-0,18323	0,0441095	-4,15397	0,0000
positive_goldstein_mentions_2	-8,85225	1,79537	-4,93061	0,0000
golds_mean_2	0,283483	0,161663	1,75354	0,0795
incidents_3	0,045455	0,0392711	1,15747	0,2471
positive_goldstein_mentions_3	8,49499	1,74902	4,85699	0,0000
golds_mean_3	-0,33019	0,157852	-2,09177	0,0365
incidents_4	-0,200733	0,0375497	-5,34579	0,0000
positive_goldstein_mentions_4	-4,39209	1,65914	-2,64722	0,0081
golds_mean_4	-0,0501427	0,151576	-0,330809	0,7408
incidents_5	-0,722389	0,0430381	-16,7849	0,0000
positive_goldstein_mentions_5	0,716009	1,35188	0,529638	0,5964
golds_mean_5	-0,128097	0,117257	-1,09244	0,2746

Table 11 First Model Improvement - Independent Variables

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	33715,1	18	1873,06	226,19	0,0000
Residual	10218,7	1234	8,28097		
Total (Corr.)	43933,8	1252			

Table 12 First Model Inprovement - Analysis of Variance

R-squared = 76,7406 percent
R-squared (adjusted for d.f.) = 76,4014 percent
Standard Error of Est. = 2,87767
Mean absolute error = 1,94134
Durbin-Watson statistic = 1,60196 (P=0,0000)
Lag 1 residual autocorrelation = 0,19811

Model

incidents_0 = 2,66816 + 0,131801*country1_to_country2_incidents + 2,05203*positive_goldstein_mentions_0 - 0,148818*golds_mean_0 + 0,536892*incidents_1 + 1,69052*positive_goldstein_mentions_1 - 0,0387612*golds_mean_1 - 0,18323*incidents_2 - 8,85225*positive_goldstein_mentions_2 + 0,283483*golds_mean_2 + 0,045455*incidents_3 + 8,49499*positive_goldstein_mentions_3 - 0,33019*golds_mean_3 - 0,200733*incidents_4 - 4,39209*positive_goldstein_mentions_4 - 0,0501427*golds_mean_4 - 0,722389*incidents_5 + 0,716009*positive_goldstein_mentions_5 - 0,128097*golds_mean_5

4.3. Second Model Improvement

Independent Variables

		Standard	T	
Parameter	Estimate	Error	Statistic	P-Value
CONSTANT	5,68126	0,685096	8,29265	0,0000
country1_to_country2_incidents	0,127965	0,00721261	17,7418	0,0000
positive_goldstein_mentions_0	0,719674	0,867365	0,829725	0,4067
incidents_1	0,553882	0,035773	15,4833	0,0000
positive_goldstein_mentions_1	0,481841	1,02409	0,470507	0,6380
incidents_2	-0,169552	0,0438644	-3,86535	0,0001

positive_goldstein_mentions_2	-6,27496	0,923751	-6,79291	0,0000
incidents_3	0,0338523	0,0392234	0,863064	0,3881
positive_goldstein_mentions_3	4,85043	0,995701	4,87137	0,0000
incidents_4	-0,196146	0,0375918	-5,21778	0,0000
positive_goldstein_mentions_4	-4,82096	1,18806	-4,05784	0,0000
incidents_5	-0,725495	0,0428694	-16,9234	0,0000
positive_goldstein_mentions_5	-0,0197445	1,00652	-0,0196166	0,9843

Table 13 Second Model Improvements - Independent Variables

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	33602,9	12	2800,25	336,11	0,0000
Residual	10330,8	1240	8,33132		
Total (Corr.)	43933,8	1252			

Table 14 Second Model Improvements - Analysis of Variance

R-squared = 76,4854 percent
R-squared (adjusted for d.f.) = 76,2579 percent
Standard Error of Est. = 2,8864
Mean absolute error = 1,93185
Durbin-Watson statistic = 1,58238 (P=0,0000)
Lag 1 residual autocorrelation = 0,207796

Model

incidents_0 = 5,68126 + 0,127965*country1_to_country2_incidents + 0,719674*positive_goldstein_mentions_0 + 0,553882*incidents_1 + 0,481841*positive_goldstein_mentions_1 - 0,169552*incidents_2 - 6,27496*positive_goldstein_mentions_2 + 0,0338523*incidents_3 + 4,85043*positive_goldstein_mentions_3 - 0,196146*incidents_4 - 4,82096*positive_goldstein_mentions_4 - 0,725495*incidents_5 - 0,0197445*positive_goldstein_mentions_5

4.4. First Model Improvement without 2021 incidents

Independent Variables

		Standard	T	
Parameter	Estimate	Error	Statistic	P-Value
CONSTANT	5,92538	1,23365	4,80315	0,0000
country1_to_country2_incidents	0,0747093	0,0067788	11,021	0,0000
positive_goldstein_mentions_0	0,327616	1,42061	0,230617	0,8176
golds_mean_0	-0,0361908	0,131067	-0,276124	0,7824
incidents_1	0,816938	0,0345866	23,62	0,0000
positive_goldstein_mentions_1	-0,240196	1,68028	-0,14295	0,8863
golds_mean_1	-0,0179367	0,152124	-0,117908	0,9061
incidents_2	-0,152845	0,0397017	-3,84984	0,0001
positive_goldstein_mentions_2	-7,28951	1,67937	-4,34062	0,0000
golds_mean_2	0,210023	0,144406	1,45439	0,1458
incidents_3	0,0625206	0,0374424	1,66978	0,0950
positive_goldstein_mentions_3	5,73722	1,51904	3,77687	0,0002
golds_mean_3	-0,100722	0,141632	-0,711156	0,4770
incidents_4	-0,0764821	0,0372165	-2,05506	0,0399
positive_goldstein_mentions_4	-2,34004	1,42512	-1,642	0,1006
golds_mean_4	0,0165373	0,132178	0,125114	0,9004
incidents_5	-0,59801	0,0528163	-11,3225	0,0000
positive_goldstein_mentions_5	-1,30168	1,16679	-1,11561	0,2646
golds_mean_5	-0,0896224	0,0992469	-0,903024	0,3665

Table 15 First Model Improvement without 2021 incidents – Independent Variables

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	35375,5	18	1965,3	361,27	0,0000
Residual	5478,13	1007	5,44005		

Total (Corr.)	40853,6	1025			

Table 16 First Model Improvement without 2021 incidents - Analysis of Variance

R-squared = 86,5908 percent
R-squared (adjusted for d.f.) = 86,3511 percent
Standard Error of Est. = 2,33239
Mean absolute error = 1,61756
Durbin-Watson statistic = 1,72601 (P=0,0000)
Lag 1 residual autocorrelation = 0,135857

Model

incidents_0 = 5,92538 + 0,0747093*country1_to_country2_incidents + 0,327616*positive_goldstein_mentions_0 - 0,0361908*golds_mean_0 + 0,816938*incidents_1 - 0,240196*positive_goldstein_mentions_1 - 0,0179367*golds_mean_1 - 0,152845*incidents_2 - 7,28951*positive_goldstein_mentions_2 + 0,210023*golds_mean_2 + 0,0625206*incidents_3 + 5,73722*positive_goldstein_mentions_3 - 0,100722*golds_mean_3 - 0,0764821*incidents_4 - 2,34004*positive_goldstein_mentions_4 + 0,0165373*golds_mean_4 - 0,59801*incidents_5 - 1,30168*positive_goldstein_mentions_5 - 0,0896224*golds_mean_5

4.5. Second Model Improvement without 2021 incidents

Independent Variables

		Standard	T	
Parameter	Estimate	Error	Statistic	P-Value
CONSTANT	6,09383	0,640298	9,51719	0,0000
country1_to_country2_incidents	0,0735337	0,00654596	11,2334	0,0000
positive_goldstein_mentions_0	0,377417	0,836512	0,451179	0,6519
incidents_1	0,822492	0,0339182	24,2493	0,0000
positive_goldstein_mentions_1	-0,795441	0,978571	-0,81286	0,4163
incidents_2	-0,149853	0,0392665	-3,8163	0,0001
positive_goldstein_mentions_2	-5,3547	0,833119	-6,42729	0,0000
incidents_3	0,0587856	0,0372639	1,57755	0,1147
positive_goldstein_mentions_3	4,55947	0,916825	4,97311	0,0000
incidents_4	-0,071567	0,0369436	-1,9372	0,0527
positive_goldstein_mentions_4	-2,38105	1,04331	-2,28221	0,0225
incidents_5	-0,593068	0,0524798	-11,3009	0,0000
positive_goldstein_mentions_5	-1,77882	0,897602	-1,98174	0,0475

Table 17 Second Model Improvement without 2021 incidents – Independent Variables

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	35357,7	12	2946,48	543,10	0,0000
Residual	5495,84	1013	5,42531		
Total (Corr.)	40853,6	1025			

Table 18 Second Model Improvement without 2021 incidents - Analysis of Variance

R-squared = 86,5475 percent
R-squared (adjusted for d.f.) = 86,3881 percent
Standard Error of Est. = 2,32923
Mean absolute error = 1,61148
Durbin-Watson statistic = 1,71503 (P=0,0000)
Lag 1 residual autocorrelation = 0,141186

Model

incidents_0 = 6,09383 + 0,0735337*country1_to_country2_incidents + 0,377417*positive_goldstein_mentions_0 + 0,822492*incidents_1 - 0,795441*positive_goldstein_mentions_1 - 0,149853*incidents_2 - 5,3547*positive_goldstein_mentions_2 + 0,0587856*incidents_3 + 4,55947*positive_goldstein_mentions_3 - 0,071567*incidents_4 - 2,38105*positive_goldstein_mentions_4 - 0,593068*incidents_5 - 1,77882*positive_goldstein_mentions_5

5. *Prediction Error HeatMap by Regions*

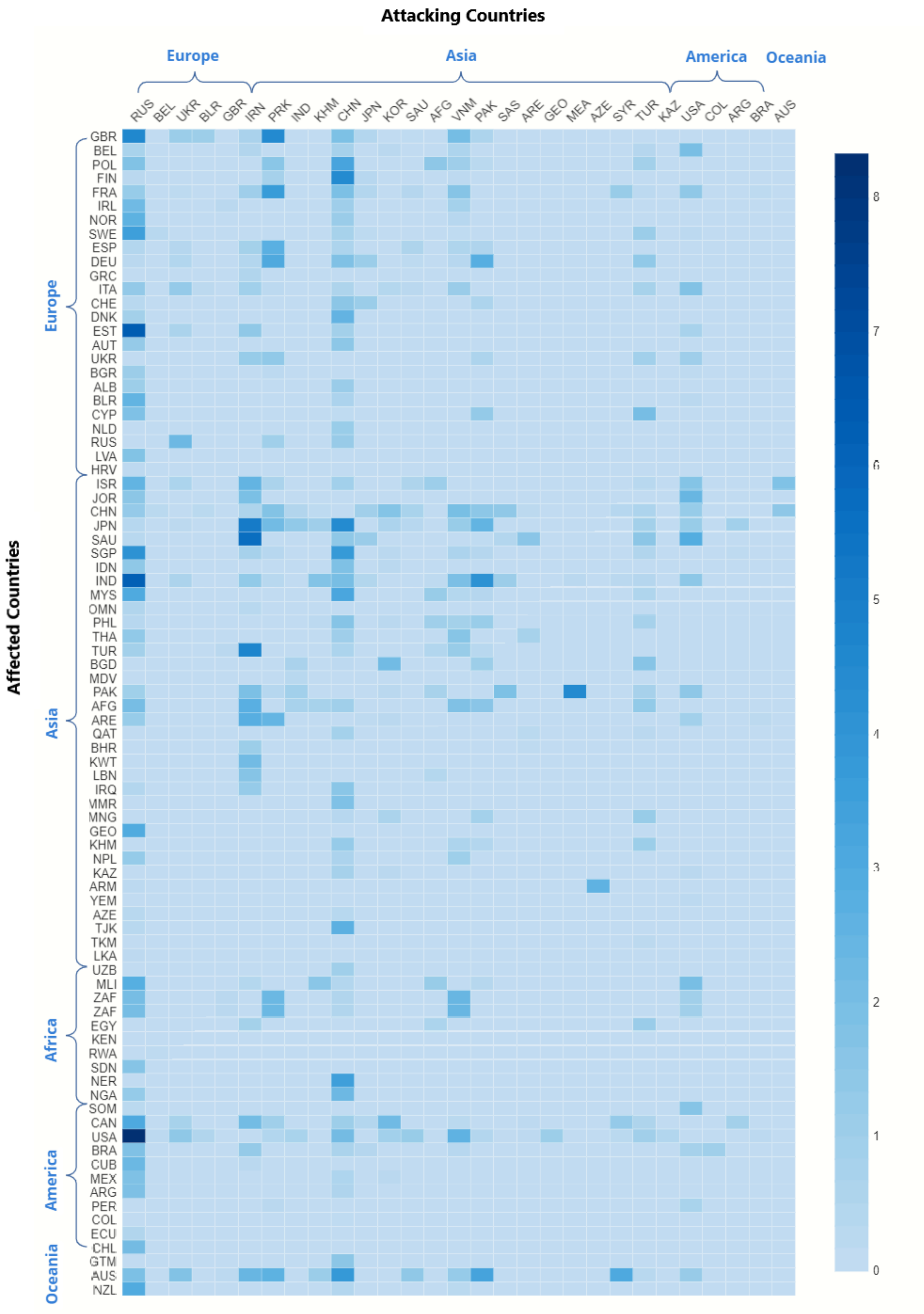

Figure 12 Prediction Error Heatmap