2022-12

USING SUPERVISED MACHINE LEARNING METHODS TO IDENTIFY FACTORS THAT INFLUENCE THE PROBABILITY OF FUTURE TERRORIST ACTIVITIES

Boone, Ethan C.

Monterey, CA; Naval Postgraduate School

https://hdl.handle.net/10945/71437

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

USING SUPERVISED MACHINE LEARNING METHODS TO IDENTIFY FACTORS THAT INFLUENCE THE PROBABILITY OF FUTURE TERRORIST ACTIVITIES

by

Ethan C. Boone

December 2022

Thesis Advisor:	Ruriko Yoshida
Co-Advisor:	Ross J. Schuchard
Second Reader:	Sean Eskew (TRAC Monterey)

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** December 2022	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE USING SUPERVISED MACHINE LEARNING METHODS TO IDENTIFY FACTORS THAT INFLUENCE THE PROBABILITY OF FUTURE TERRORIST ACTIVITIES	**5. FUNDING NUMBERS**
6. AUTHOR(S) Ethan C. Boone	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.	**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

The Defense Counter Terrorism Center (DCTC) at the Defense Intelligence Agency (DIA) focuses on classifying and predicting terrorist activities at a global scale. To accomplish this, DCTC analysts collect, process, and analyze open-source data from across the internet, including event information as reported by traditional and social media sources. This information is often aggregated in publicly available datasets, such as the Global Terrorism Database (GTD) and the Armed Conflict Location & Event Data Project (ACLED), that require additional analytic scrutiny for the DCTC team to fully exploit the contained information. In support of these efforts, this study utilizes the ACLED dataset and geospatial data to provide a monthly prediction of violent events to the DCTC team. Two models are used for comparison: a generalized network autoregressive (GNAR) time series model and an ensemble model. The results from these machine learning models will be integrated into an interactive dashboard that displays descriptive statistical information and the predictive model results about various terrorist organizations.

14. SUBJECT TERMS terrorism, machine learning, DIA, generalized network autoregressive model, terrorist attacks, ensemble model, predictive models, feature importance, random forest, classification, regression, time series analysis	**15. NUMBER OF PAGES** 75
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

THIS PAGE INTENTIONALLY LEFT BLANK

USING SUPERVISED MACHINE LEARNING METHODS TO IDENTIFY FACTORS THAT INFLUENCE THE PROBABILITY OF FUTURE TERRORIST ACTIVITIES

Ethan C. Boone
Ensign, United States Navy
BS, Embry-Riddle Aeronautical University, Daytona Beach, 2021

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
December 2022

Approved by: Ruriko Yoshida
Advisor

Ross J. Schuchard
Co-Advisor

Sean Eskew
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The Defense Counter Terrorism Center (DCTC) at the Defense Intelligence Agency (DIA) focuses on classifying and predicting terrorist activities at a global scale. To accomplish this, DCTC analysts collect, process, and analyze open-source data from across the internet, including event information as reported by traditional and social media sources. This information is often aggregated in publicly available datasets, such as the Global Terrorism Database (GTD) and the Armed Conflict Location & Event Data Project (ACLED), that require additional analytic scrutiny for the DCTC team to fully exploit the contained information. In support of these efforts, this study utilizes the ACLED dataset and geospatial data to provide a monthly prediction of violent events to the DCTC team. Two models are used for comparison: a generalized network autoregressive (GNAR) time series model and an ensemble model. The results from these machine learning models will be integrated into an interactive dashboard that displays descriptive statistical information and the predictive model results about various terrorist organizations.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

ACLED	Armed Conflict Location & Event Data Project
AIC	Akaike information criterion
AR	Autoregressive
ARIMA	autoregressive integrated moving average
AUC	area under curve
CENTCOM	United States Central Command
CSV	comma-separated values
DIA	Defense Intelligence Agency
DOD	Department of Defense
DCTC	Defense Counter Terrorism Center
GDELT	Global Database of Events, Language and Tone
G-Econ	Geographically based Economic data
GIS	Graphic Information System
GNAR	Generalized Network Autoregressive
GTD	Global Terrorism Database
INDOPACOM	United States Indo-Pacific Command
KDE	kernel density estimation
MAPE	mean absolute percent error
MASE	mean absolute squared error
MDG	mean decrease Gini
MSE	mean square error
NN	neural network
NPS	Naval Postgraduate School

RH	rolling horizon
RF	random forest
ROC	receiver operating characteristic
SD	standard deviation
SVM	support vector machine

Executive Summary

Analysts at counter-terrorism agencies, such as the Defense Intelligence Agency (DIA) Defense Counter Terrorism Center (DCTC) team, often rely on the manual examination of disparate open-source event databases. The DCTC team is interested in the development of a prototype analytic tool that increases their productivity and provides additional insights derived from open-source event databases. Within the prototype analytic tool the framework exists to aggregate, process, and visualize the open-source event information.

To provide stakeholders with additional insights from the open-source event databases, this thesis aims to implement predictive models. These models can provide additional insights into geographic areas of concern for the stakeholders, as well as identify the models important features. The predictive models attempt to predict the number of violent events with fatalities in various areas of the world. A GNAR model and an ensemble model are used to attempt to provide these predictions.

Prior to running the predictive models, the selection of an open-source event database is required. The data sources considered for this thesis are both well known by the stakeholders for this thesis and academia. The three sources considered are the Global Terrorism Database (GTD), ACLED, and Global Database of Events, Language and Tone (GDELT). The GTD database is published annually, requires the least cleaning, and is carefully vetted. The ACLED database is updated weekly, requires slightly more cleaning than the GTD, and is also carefully vetted before publication. The GDELT database is updated approximately every 15 minutes, the most often of the three databases. However, this high frequency of updates makes this database require the most cleaning. The ACLED database was chosen as the data source for this thesis due to its balance of update frequency and cleanliness of the data.

With a data source selected, predictors were then collected. The source of predictors falls into two main categories: those engineered from the ACLED database and those taken from other external open-source databases. Two of the main external databases considered for predictors were the World Bank and the Yale G-Econ. During the exploratory analysis phase, it was discovered that the World Bank does not provide suitable predictors due to its

annual update frequency and poor coverage of some countries. Four static predictors were utilized from the G-Econ database, which consisted of distances to major bodies of water. Lastly, predictors were generated from feature engineering the ACLED database.

The first model utilized in this thesis was the GNAR model. As implied by its name, this model uses a network structure to allow the inclusion of neighboring information. Overall, this model performed quite poorly at predicting the number of violent events with fatalities. This poor performance was particularly captured by the resulting MASE values.

The ensemble model consisted of two components. Firstly, a time-series forecasting of the predictors one month ahead was completed. With those forecasted values, a prediction model was then fit. Both a classification and a prediction model were attempted. The classification problem utilized a RF and logistic regression. For the classification problem the RF outperformed the logistic regression. The regression problems utilized a RF and linear regression. For the regression problem there was no clear top performer. However, the important predictors were gathered for both the classification and regression problem. The important predictors from the RF were calculated using the mean decrease Gini (MDG). The logistic and linear regression important predictors were calculated using the t-statistic value.

Overall, the important predictors for the classification and regression problem were mainly derived from fatality or event information. Interestingly, the logistic regression was the only model that identified some geographic predictors within the top five most important predictors. This is interesting given that several previous studies identified some geographic predictors as the most important.

Lastly, the performance of the GNAR and ensemble model appears to indicate that the development of higher resolution models could improve overall predictive power, as well as provide more accurate important predictors. This insight was driven by the interpreted poor performance of the ensemble model in some regions. The regions that could likely be improved from higher resolution models are the Middle East, South Asia, Central Africa, and Central America.

Acknowledgments

What at first felt to be an impossible endeavor would not be possible without my advisory team. Dr. Yoshida, thank you for your invaluable expertise and guidance throughout this process.

LTC Schuchard, your guidance and mentorship both in the classroom and on this project were invaluable during my tenure at the Naval Postgraduate School (NPS).

To MAJ Eskew, your iterative feedback and experience were essential to keeping this thesis on track.

To the stakeholders of this project, particularly the Defense Intelligence Agency (DIA) Defense Counter Terrorism Center (DCTC) team, thank you for the opportunity to apply the skills gained at NPS to this particular problem.

To my cohort, thank you for providing the unique exposure to so many diverse backgrounds at the beginning of my career.

I would also like to thank Naval Aviation for affording me the opportunity to earn a graduate degree prior to starting flight school.

It would be remiss to not mention my family for their support during this endeavor. To my grandmother, Mimi, thank you for the many hours spent proofing over the years. Lastly, I would especially like to thank my girlfriend, Taylor, for continuing to be a pillar of support.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Introduction

This thesis aims to provide predictions of terrorist activities using existing open source information. In addition to prediction, this thesis seeks to provide the stakeholders with the importance of various predictors in providing this prediction.

1.1 Background

Initially this thesis began with the Defense Intelligence Agency (DIA) Defense Counter Terrorism Center (DCTC) team as the sole stakeholder. However, through the development of this thesis, the United States Central Command (CENTCOM) J-8 Resources and Analysis team joined as a stakeholder and United States Indo-Pacific Command (INDOPACOM) has also expressed their interest in the project. All stakeholders for this thesis are interested in the additional insight that can be derived from using statistical methods as well as machine learning to understand threats within their Areas of Responsibilities. Counter-terrorism groups, as well as those interested in terrorist activities, leverage many different tools to provide more insight and better analyze terrorist activities. The implementation and addition of new analytic tools help to ensure these organizations remain effective at providing insights into terrorist activities.

There are several open-source event databases that are known in the intelligence community. However, while analysts, such as the DCTC, are aware of the existence of these databases they are typically unable to apply advanced statistical methods themselves. The skills and methods taught in the Operations Research curriculum at the Naval Postgraduate School (NPS) allow for the implementation of statistical methods and machine learning on this problem.

The DCTC team is interested in the development of a prototype dashboard that provides statistical insights into open-source data, as well as provides predictive modeling. Iterative feedback was generated through stakeholder engagement with both the DCTC team and the stakeholders at CENTCOM J-8 Resources and Analysis. The stakeholders at CENTCOM are particularly interested in pulling as much information as possible from the Armed Conflict

Location & Event Data Project (ACLED) database. The stakeholders at INDOPACOM are interested in the methods utilized in this thesis. INDOPACOM is particularly interested in how the data was structured to ultimately feed into the predictive models.

1.2 Objective

The objective of this thesis is to provide stakeholders with an analytic tool that provides additional insight into open-source information, as well as create and deliver a predictive model. Through stakeholder engagement, it was determined that predicting the number of violent events with fatalities would be an appropriate problem to approach for the prediction models. This objective was not only beneficial for the predictions it could provide, but also for the identification of important variables associated with a rise in violence. For each stakeholders' end users, the developed prediction model allows for the identification of areas at risk for an increase in potential terrorist activities for the upcoming month. Additionally, the use of certain machine learning algorithms provides the stakeholders with predictors the model found important. The monthly feedback on the importance of predictors allows them to assess if there are any predictors they are not currently tracking but should be tracking.

An additional objective for this thesis was to aggregate information from the open-source event database that is not easily captured by the data in its raw format. Through stakeholder engagement it quickly became apparent that there was significant interest in presenting the data in a way that allowed for exploratory analysis by the end user. Allowing the end users the freedom to conduct exploratory analysis enables the development of in-house intelligence products leveraging the selected open-source event database.

1.3 Scope

The scope of this thesis consists of two main parts: the development of a dashboard and the implementation of a prediction model. For the stakeholders, the end product is the delivery of the prototype dashboard.

Stakeholder engagements identified that the scope of the predictive model should be global. This was derived from the DCTC team's concerns regarding worldwide terrorism. Based on the findings of this thesis, further research could be conducted to improve on areas in which this work fell short.

The data used in all areas of the dashboard, outside of the predictive models, is derived from the ACLED database. The selection of features shown in the dashboard is a product of the iterative feedback from the stakeholders, as well as the implementation of statistical methods taught in the Operations Research curriculum.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 2: Literature Review

The ability to provide insight into future terrorist attacks, as well as identifying predictors that influence them are of interest to counter-terrorism agencies as well as academic institutions. Researchers have utilized various machine learning methods to provide insights into future terrorist attacks and identify the important factors that influence them. Using machine learning methods can outperform naïve prediction methods.

Common types of machine learning algorithms have been applied to this particular research area. These algorithms include support vector machine (SVM), neural network (NN), and random forest (RF). These are typically applied in a classification approach in an effort to determine the areas of high and low risk within various regions. Spatial data is fed into these models, which can be gathered from various databases. Some of the most well-known databases are Global Terrorism Database (GTD), ACLED, and Global Database of Events, Language and Tone (GDELT). The differences between each of these will be discussed in Section 2.3.

2.1 Relevance

Terrorism is not a problem that's constrained to any one area of the world. The ability for counter-terrorism organizations, particularly the DCTC team, to be able to predict areas with increased terrorist activities is vital to their mission. In particular, the DCTC is particularly interested in this ability, as it would increase the efficiency of their initial analysis. This would free up more time for their analysts to focus on more intensive tasks.

2.2 Related Research

Given that terrorism is global problem, there is no shortage of research into developing models that attempt to predict it and provide additional insights. Many of these models have been centered around using spatial data and machine learning methods such as RF, SVM, logistic regressions, and NNs.

2.2.1 Data Collection and Processing

All studies that sought to provide a prediction of terrorist events required two types of data sources: terrorist event sources and predictor sources. The types of sources used by related studies are discussed below.

Terrorist Events

Both Hao et al. (2019) and Ding et al. (2017) used data from the GTD for collecting information on terrorist events for their analysis.

The location of a terrorist event is typically represented by the geographic coordinates where it occurred. However, it is difficult to group data by coordinates without some sort of organizational structure. This is a problem that researchers needed to solve in order to complete any further analysis. Hao et al. (2019) used grids to group data from the GTD. Since they were only looking at the Indochina peninsula, they sampled their data into grid sizes of 0.05×0.05 degrees (Hao et al. 2019). During the sampling process, a summary of all events within a grid was obtained. This grid format summary was the data utilized in the prediction portion of their research (Hao et al. 2019).

The grid sizes chosen by Ding et al. (2017) were initially determined at the resolution of the Geographically based Economic data (G-Econ) (1×1 deg). Through the use of Graphic Information System (GIS) they were able to resample their data to a resolution of 0.1×0.1 degrees. The structure of their data was handled in the same way that Hao et al. (2019) handled their data. Ding et al. (2017) also generated a summary of events within each grid, which was then used in their prediction models.

Predictors

Related studies collected their predictors from tabular databases to GIS databases. Research completed by Hao et al. (2019) collected predictors that fit into three categories: social, natural, and geographical elements. These are shown in Table 2.1.

Table 2.1. Databases used in Hao et al. (2019). Adapted from Hao et al. (2019, table 2).

Social	Natural	Geographic
Global Fragile States Index	G-Econ 4.0	ASTER Global DEM
GeoEPR, the Ethnic Power Relations dataset	Land surface temperature anomaly	Global urbanisation and accessibility map
World drug report	Drought index of the world	G-Econ 4.0
Adjusted population density, V4.10	Global Multi-hazard Frequency and Distribution, v1	
Version 4 DMSP-OLS night-time lights time series		

As shown in Table 2.1, Hao et al. (2019) used the most databases for collecting social predictors. They also utilized the G-Econ database for natural and geographic predictors.

The study completed by Ding et al. (2017) used G-Econ, GeoEPR, World drug report, Nighttime Lights of the World, Population Density of the World, and the Digital Elevation Model. These databases were used to derive 10 predictors that were used in their prediction model.

2.2.2 Identifying Factors

All of the discussed studies involved the collection and engineering of predictors and factors. The identification and collection of good factors are essential to conduct valuable analysis for stakeholders. Several pieces of literature focus exclusively on identifying what predictors are critical to increasing the risk of terrorist attacks. Research done by Luo and Qi (2021) identified the most important factors to increasing the risk of terrorist attacks. These factors are:

- Human Loss
- GDP Growth
- Military Expenditure
- Population Growth
- Population
- Unemployment
- Urban Population Growth
- Internal Conflict. (Luo and Qi 2021)

Other research also included other spatial-temporal information. Hao et al. (2019) found the most important predictors were, in order of importance,

1. Urban Accessibility
2. Topography
3. Average Precipitation
4. Night-Time Light
5. Distance to Major Navigable River
6. Distance to Major Navigable Lake
7. Population Density
8. Average Temperature. (Hao et al. 2019)

These above predictors accounted for 71.96% of the contribution to their prediction models results (Hao et al. 2019). There is a fairly even split between static and time series predictors in the study completed by Hao et al. (2019).

The researchers Ding et al. (2017) found important predictors that were similar to those found by Hao et al. (2019). Ding et al. (2017) found the most important factors were, in order of importance,

1. Population Density
2. Latitude
3. Longitude
4. Nighttime Lights
5. Distance to River
6. Distance to Lake
7. Major Drug Regions
8. Distance to Ocean
9. Average Temperature
10. Topography. (Ding et al. 2017)

It is interesting that six out of the ten most important predictors in Ding et al. (2017) study are static. The other time-dependent variables are information that tend to not change rapidly over time.

It is important to note that there is a mixture of static and time series predictors present in the various studies. All factors identified by Luo and Qi (2021) were time series. However, some factors from Hao et al. (2019) are constant through time. These include topography information and distances to various bodies of water. For this thesis, both static and time series information should be considered for predictors.

2.2.3 Predicting Terrorist Events

At the heart of all related studies was the goal to accurately predict terrorist events and their areas of occurrence. Related research tends to focus mainly on the prediction of terrorist events, not the time series analysis of predictors.

Time Series Prediction

A majority of related research did not focus on completing time series analysis of terrorist event data. The closest study to use some semblance of time series analysis is by Hao et al. (2019). They utilized kernel density estimation (KDE) to generate a risk metric for different time periods on the Indochina Peninsula. However, for their analysis they only focused on applying KDE from 2005-2016.

Classification Prediction

Related research typically sought to identify areas of high or low risk. The study by Ding et al. (2017) considered high risk areas to have events that resulted in casualties, whereas low risk areas resulted in no casualties. Using that definition of risk formed a classification problem that was then solved using the following methods; SVM, NN, and RF. These methods resulted in area under curve (AUC) values from 0.976 to 0.971 when applied on the validation data (Ding et al. 2017). The study leaned towards the use of a RF due to its ability to determine the importance of predictors in the classification problem.

Hao et al. (2019) only utilized a RF to determine their areas of high risk. Their classification problem had two values: one where at least one terrorist attack had occurred and zero where no terrorist attack had occurred. Once implemented in the RF, the researchers then used prediction probabilities to create a risk metric for terrorist attacks. Utilizing the resulting geospatial risk plot, they were able visually identify areas of the Indochina peninsula that are at high risk.

2.3 Event Data Sources

There are three main open-source databases that contain information on violent and non-violent events on the global scale. These three are GTD, ACLED, and the GDELT. A description of each are provided below.

> **GTD** The GTD started in 2001 at the University of Maryland (START (National Consortium for the Study of Terrorism and Responses to Terrorism) 2022). This is an open-source database that encompasses terrorist events around the world spanning from 1970 to 2020 (START (National Consortium for the Study of Terrorism and Responses to Terrorism) 2022). Updates to this database tend to occur annually. Over 200,000 curated events are included in this database. For each event there is a minimum of 45 variables. However, newer entries have more than 120 variables (START (National Consortium for the Study of Terrorism and Responses to Terrorism) 2022).
>
> **ACLED** The ACLED dataset is maintained and operated by a non-profit organization located in the United States (ACLED 2022). The data provided by the ACLED project is event based. This data is collected in real time and published weekly (ACLED 2019). Twenty-eight variables are collected for each event, which contain location, actor, and event type information (ACLED 2019).
>
> **GDELT** The GDELT database uses Google Jigsaw to monitor media in over 100 languages across the globe (Leetaru 2022). This is the largest of the two databases, with the raw comma-separated values (CSV) files taking over 2.5 TB of storage (Leetaru 2022). Their event database contains over 250 million records (Leetaru 2022). However, this data is not very uniform, given that it is just the results of constant web scraping.

The databases above are listed in order of the lowest to highest frequency of updates. Generally, as the frequency of updates increases, the overall cleanliness and immediate usability of the data decreases. This was especially important to consider, given the goal was to create a prediction model from the chosen dataset.

2.4 Areas of Improvement

As mentioned in Section 2.2.3, much of the related research focused on predicting risk levels using data sources that are not updated frequently. It is also important that a black box product is not delivered, as it would not be as useful to the DCTC team. This thesis goal is two-fold: utilize time series analysis to forecast up-to-date predictors and then use those values to solve the classification and/or regression problem. This method should provide the DCTC team with a monthly prediction, unlike previous research which would analyze past data.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 3: Methodology

The overall goal of this thesis is two-fold. The first goal is to provide the DCTC team with a tool to provide insight into areas of the world that are likely to have an increase in violent events. The second goal is to provide them with important predictors in determining an increase in violent events. From this information, they can confirm the effectiveness of current predictors, as well as introduce new predictors for the team to use in their analysis.

3.1 Data Collection

The event data for this thesis was collected from ACLED (Raleigh et al. 2010). The data from ACLED reports information about the type, agents, location, date, and other characteristics of select events. These events include political violence, demonstrations, and some politically relevant non-violent events. ACLED allows the tracking of violent and non-violent events across various types of groups. This database is updated weekly, with events going to the most recent Friday (ACLED 2019). A description of all ACLED variables, as well as their data type is explained in Table 3.1.

Table 3.1. Description of ACLED variables. Adapted from ACLED (2019, table 1)

Variable Name	Variable Type	Description
ISO	Numeric	This is a numeric code that's unique to each country.
EVENT_ID_CNTY	Character	This is an individual identifier that is comprised of a country acronym and a number.
EVENT_ID_NO_CNTY	Numeric	This is an individual numeric identifier.
EVENT_Date	Date	The day, month, and year a given event took place.

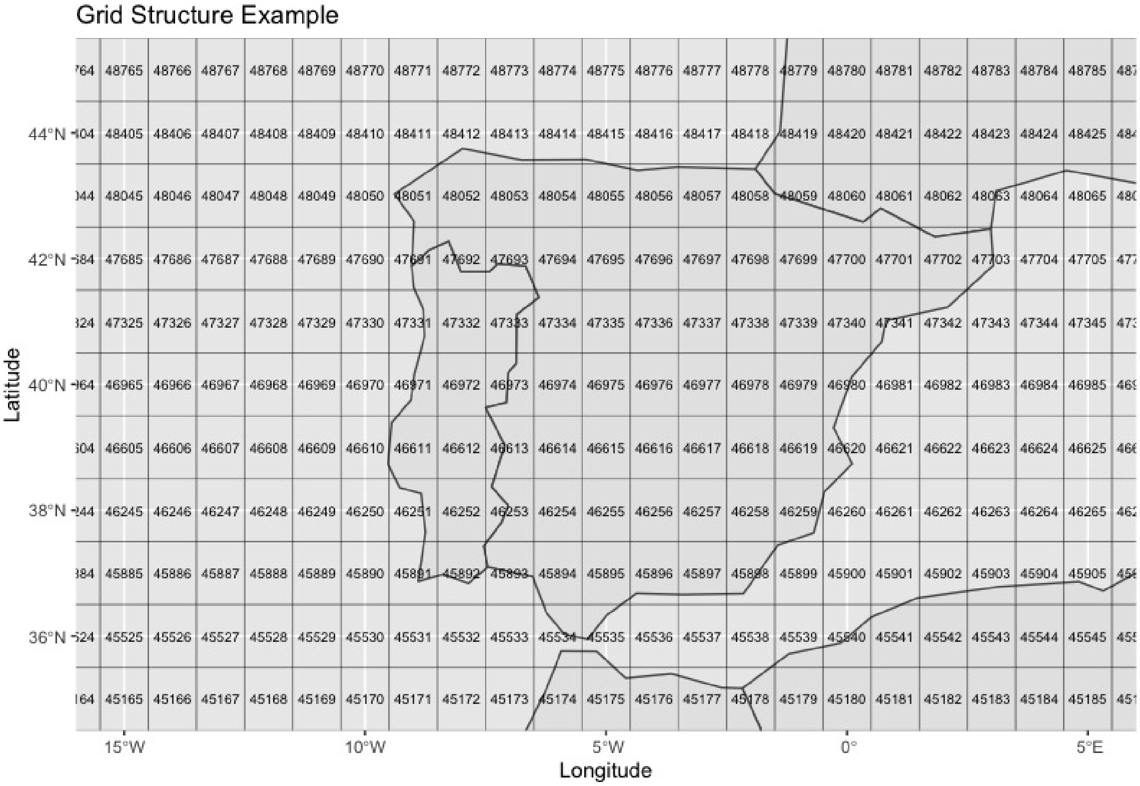
YEAR	Numeric	The year an event took place.
TIME_PRECISION	Numeric	Code indicating the certainty of an events' recorded date.
EVENT_TYPE	Character	One of six types an event can be categorized under.
SUB_EVENT_TYPE	Character	Further breakdown of an event type.
ACTOR1	Character	The name of an actor that was involved in the event.
ASSOC_ACTOR_1	Character	The name of the actor associated with or identifying ACTOR1.
INTER1	Numeric	A number that indicates the organization type of ACTOR1.
ACTOR2	Character	The name of an actor that was involved in the event.
ASSOC_ACTOR_2	Character	The name of the actor associated with or identifying ACTOR2.
INTER2	Numeric	A number that indicates the organization type of ACTOR2.
INTERACTION	Numeric	The numeric combination of INTER1 and INTER2.
REGION	Character	The region in which the event took place.
COUNTRY	Character	The country in which the event took place.
ADMIN1	Character	The largest sub-national administrative region in which the event took place.
ADMIN2	Character	The second largest sub-national administrative region in which the event took place.
ADMIN3	Character	The third largest sub-national administrative region in which the event took place.

LOCATION	Character	The village or town in which an event occurred.
LATITUDE	Numeric	The latitude of the location where the event took place.
LONGITUDE	Numeric	The longitude of the location where the event took place.
GEO_PRECISION	Numeric	A numeric value that indicates the certainty of the coded location for an event.
SOURCE	Character	The source for the event information.
SOURCE SCALE	Character	Indicates if the source scale is local, regional, or national.
NOTES	Character	Contains a short description of the event.
FATALITIES	Numeric	The count of reported fatalities for a given event.

As shown in Table 3.1, the ACLED database captures a wide range of information about events. The data is collected from various sources, which span from national to local. The data collection is completed by trained experts before events appear on the ACLED database (Raleigh et al. 2010).

3.1.1 Data Setup

The highest degree of spatial resolution present in the ACLED data is by sub-national administrative region. To better work with the data, it was decided to utilize a grid structure. Each grid would be one degree of latitude by one degree of longitude. This created 64,800 unique grids into which the data could be grouped. An example of this grid structure on Spain and Portugal is shown in Figure 3.1.



764	48765	48766	48767	48768	48769	48770	48771	48772	48773	48774	48775	48776	48777	48778	48779	48780	48781	48782	48783	48784	48785	487
404	48405	48406	48407	48408	48409	48410	48411	48412	48413	48414	48415	48416	48417	48418	48419	48420	48421	48422	48423	48424	48425	484
044	48045	48046	48047	48048	48049	48050	48051	48052	48053	48054	48055	48056	48057	48058	48059	48060	48061	48062	48063	48064	48065	480
684	47685	47686	47687	47688	47689	47690	47691	47692	47693	47694	47695	47696	47697	47698	47699	47700	47701	47702	47703	47704	47705	477
324	47325	47326	47327	47328	47329	47330	47331	47332	47333	47334	47335	47336	47337	47338	47339	47340	47341	47342	47343	47344	47345	473
964	46965	46966	46967	46968	46969	46970	46971	46972	46973	46974	46975	46976	46977	46978	46979	46980	46981	46982	46983	46984	46985	469
604	46605	46606	46607	46608	46609	46610	46611	46612	46613	46614	46615	46616	46617	46618	46619	46620	46621	46622	46623	46624	46625	466
244	46245	46246	46247	46248	46249	46250	46251	46252	46253	46254	46255	46256	46257	46258	46259	46260	46261	46262	46263	46264	46265	462
884	45885	45886	45887	45888	45889	45890	45891	45892	45893	45894	45895	45896	45897	45898	45899	45900	45901	45902	45903	45904	45905	459
524	45525	45526	45527	45528	45529	45530	45531	45532	45533	45534	45535	45536	45537	45538	45539	45540	45541	45542	45543	45544	45545	455
164	45165	45166	45167	45168	45169	45170	45171	45172	45173	45174	45175	45176	45177	45178	45179	45180	45181	45182	45183	45184	45185	451

Figure 3.1. Developed grid structure as shown over Spain and Portugal. Each grid is 1° latitude and 1° longitude.

As shown in Figure 3.1, the grids are filled out left to right and then bottom to top. The first grid cell is located at 90°S, 180°W.

With the grid structure determined, the temporal breaks needed to be determined. It was decided that the data best be split up monthly, as it would provide less variance for both the time series and prediction analysis. The decision then had to be made about what dates would be considered for this analysis. It was determined that only events starting from 2018 to present would be utilized in this analysis.

3.2 Determining Predictors

The most important task prior to beginning the analysis was the collection of predictors. Three main sources were used to collect the predictors, The World Bank, Yale G-Econ, and feature engineering from ACLED.

3.2.1 World Bank

Initially, many predictors were gathered from The World Bank database. These predictors can be found in Table 3.2.

Table 3.2. Compilation of World Bank (2022) data that was collected as predictors.

Name	Collection Frequency
Time to Electricity	Yearly
Time to Start Business (Male)	Yearly
Time to Start Business (Female)	Yearly
Arable Land (% Of Total Land)	Yearly
Agricultural Employment (Male)	Yearly
Agricultural Employment (Female)	Yearly
Air Departures Worldwide	Yearly
ATMs per 100k Population	Yearly
Fertility Rate per Woman	Yearly

The data from The World Bank is well curated, which required little cleaning to start using it. All datasets from The World Bank were collected as a CSV file, making it very easy to work with in R.

It should also be noted that there were several disadvantages of using data from The World Bank. It was determined that the ACLED data would be separated using a one month time step. However, The World Bank data, as shown in Table 3.2, is collected annually. This is undesirable because it means that nine potential predictors will only change annually for the prediction model. This could be mitigated by having a large corpus of training data that goes back many years. However, for this thesis a predictor from The World Bank could have a maximum of four different values.

It appears that a trade-off of cleanliness of the World Bank data is the speed in which new data is released. At the time of collecting data from The World Bank, the most recent year present in any of the datasets was 2019. This proved problematic, as this thesis aims to deliver a prototype that can be used immediately. Given that there are no features available in the World Bank data past 2019, it was eliminated as a source for predictors.

Lastly there were several countries that were not included in these datasets. An example of countries that were missing from a World Bank dataset is shown in Figure 3.2.

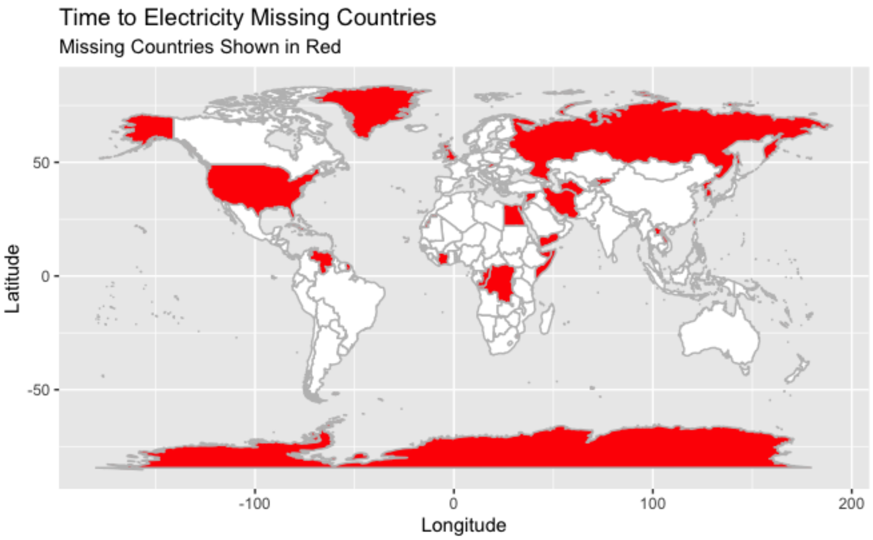

Time to Electricity Missing Countries
Missing Countries Shown in Red
50
0
-50
Latitude
-100
0
100
200
Longitude

Figure 3.2. Countries with no recorded values for their time to electricity are shaded red. This figure was generated using data from the World Bank (2022).

As shown in Figure 3.2, some of these missing countries are common operating areas of terrorist groups. For example, there is no time to electricity data for Syria, Iran, Afghanistan, and Somalia. These four countries are known for having high amounts of terrorist activity.

3.2.2 Yale G-Econ

The Yale G-Econ database contained economic and geographic information that was organized by grids that are one by one degree of latitude and longitude. The predictors that were considered for this thesis were those that related to distances to major bodies of water. This decision was made based off the important predictors derived in previous studies.

3.2.3 ACLED Data

Several predictors were also generated through feature engineering from the ACLED database. These features were all calculated when the ACLED data was grouped by month and grid. There were several columns of interest to determine predictors, which are identified in Table 3.3.

Table 3.3. ACLED variables used for feature engineering. The description of each variable is provided in Table 3.1. An X indicates the statistic type that was derived from each variable.

Variable	Count	Sum	Mean	Range	Unique Values
ACTOR1					X
ACTOR2					X
EVENT_TYPE	X				X
SUB_EVENT_TYPE	X				
FATALITIES		X	X	X	
ADMIN1					X
ADMIN2					X
ADMIN3					X

In addition to the predictors in Table 3.3, the counts for each unique event type were also considered. These predictors were then broken down monthly for each grid present in the ACLED dataset. Missing values were replaced with zeros for all predictors in Table 3.3.

3.3 Generalized Network Autoregressive (GNAR) Model

Initially, changes in event types were modeled using a GNAR time series model. This model utilizes network structures to provide better fitting Autoregressive (AR) time series models. To work with this model, the ACLED data needed to be fit into a network structure. Two files were needed to create this structure: a file that contained the event counts across time and a file that contained an adjacency matrix.

For the count of events, a rolling sum approach was taken. This means that a grid's event counts was the cumulative sum of that month plus the counts of all previous months. This was done due to the small counts of events that happen in each month. Typically, monthly values for grids rarely stray far from zero.

The adjacency matrix allowed room for interpretation. It was decided that only cells within one step of a grid would be considered adjacency. This assumption is shown in Figure 3.3.

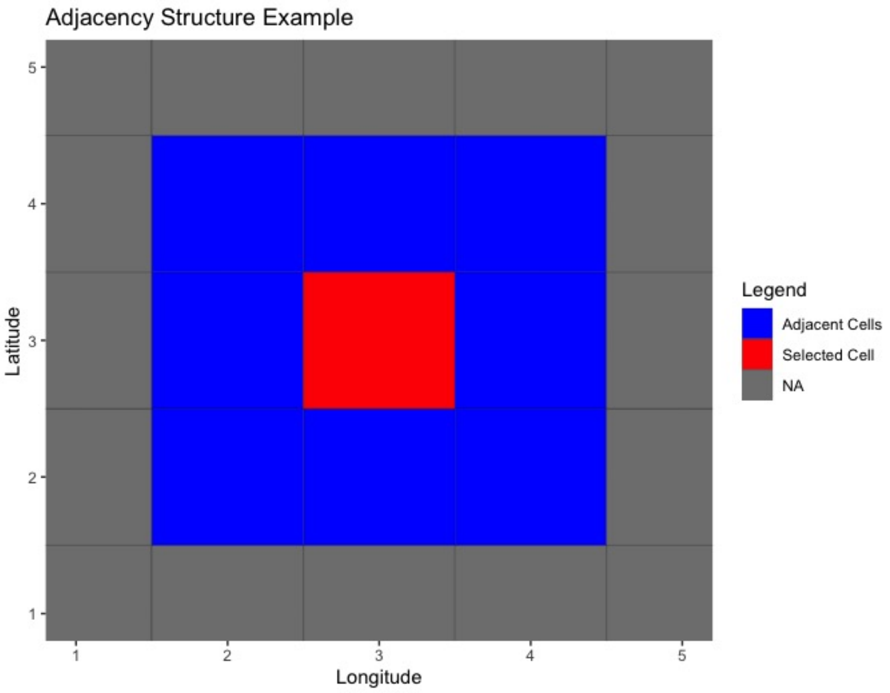


Figure 3.3. Visualization of one step adjacency structure used for the GNAR model. Blue grids indicate the cells that were considered adjacent to the red grid. Cells not considered adjacent are gray.

As shown by Figure 3.3, the selected cell is red and the adjacent grids are highlighted in blue. A one step adjacency approach was applied to this problem which included cells touching on the diagonal. This same process was followed for all grids present in the ACLED database. The grids present in the ACLED database is shown in Figure 3.4.

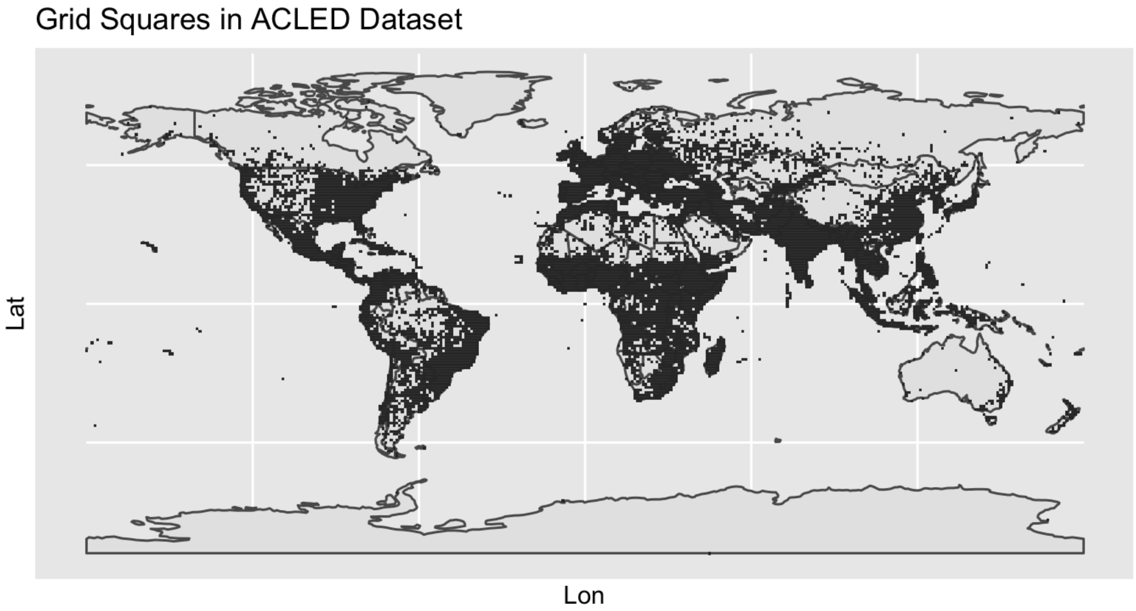


Figure 3.4. Visualization of grids that are contained in the ACLED dataset. The black grids identify those that were in the These are collected from Jan 1997 to Feb 2022. This figure was generated using data provided by Raleigh et al. (2010).

As shown in Figure 3.4, the grids collected in the ACLED dataset closely resemble a population map. This makes sense given that events are likely to only occur in populated areas. The inclusion of additional grids not captured by ACLED increases the risk of accuracy bias. For the GNAR model, there were additional neighbors added to each grid in the ACLED dataset. These grids were added in accordance with the example shown in Figure 3.3.

There are several parameters in the GNAR model to tune. These include the α and β parameters. The α parameter is a non-negative integer specifying the maximum time-lag to the model (Leeming et al. 2020). The β parameter is a vector of length α, which specifies the maximum neighbor set to the model at each of the time lags (Leeming et al. 2020).

3.4 Ensemble Model

An ensemble model was also considered for this problem set. The model consisted of using time series analysis to collect a robust set of predictors. Many of these predictors were used from Yale G-Econ database. Once the time series analysis was complete, the results were then merged onto the response variables generated from the ACLED database to be used in the prediction models.

3.4.1 Time Series Analysis

A time series analysis was completed for each predictor in each grid. A monthly time series spanning from January 2018 to February 2022 was used for this analysis. A time series model was fit using an autoregressive integrated moving average (ARIMA) model. Specifically, the auto.arima() function from the forecast package by Hyndman and Khandakar (2008) was used as the time series model. These models were fit by minimizing the Akaike information criterion (AIC).

For a more robust analysis, the forecast of each model was compared by using a rolling horizon (RH) approach. The RH could then be used to calculate performance metrics such as mean square error (MSE), mean absolute squared error (MASE), and mean absolute percent error (MAPE). The equations for each are shown in Equations (3.1) to (3.3).

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(Y_t - F_t)^2, \tag{3.1}$$

$$MASE = \begin{cases} \frac{\sum_{t=1}^{N} \left|\frac{Y_t - F_t}{Y_t - Y_{t-1}}\right|}{N} & \text{if } \sum_{t=1}^{N} |Y_t - Y_{t-1}| \neq 0 \\ \frac{\sum_{t=1}^{N} |Y_t - F_t|}{N} & \text{if } \sum_{t=1}^{N} |Y_t - Y_{t-1}| = 0, \end{cases} \tag{3.2}$$

$$MAPE = \begin{cases} \frac{\sum_{t=1}^{N} |\frac{Y_t - F_t}{Y_t}|}{N} & \text{if } \sum_{t=1}^{N} |Y_t| \neq 0 \\ \frac{\sum_{t=1}^{N} |Y_t - F_t|}{N} & \text{if } \sum_{t=1}^{N} |Y_t| = 0, \end{cases} \tag{3.3}$$

where Y_t is an observation at time t and F_t is an estimate of Y_t at time t. There were modifications made to the traditional MASE and MAPE formulas to account for instances that would render these metrics unusable. In this thesis, Equations (3.1) to (3.3) were calculated using the Metrics package by Hamner and Frasco (2018).

These metrics are used to measure various performance aspects of the fitted time series model. The MAPE describes the average of ratios of the error for each observation by a model and magnitude of the observation. MASE measures the average of ratios for errors by a given model and errors by the naïve model. Thus, if the MASE is equal to one, it means that the model is performing no better than if we had predicted the previous time steps value. A MASE of 0.5 means that we have doubled our performance compared to the naïve method.

3.4.2 Classification and Regression Analysis

The forecast values from the time series analysis would then be used to solve classification and regression problems. The response variables for both problems were based on the count of violent events with fatalities by grid and month. For the classification problem the response variable represented either an increase in the number of violent events with fatalities from the previous month or no change or a decrease in the number of events.

Classification Analysis

The classification problems used the following algorithms: RF and logistic regression. The initial classification problem attempted to determine if a grid had an increase in violent events with fatalities from the previous month. Violent events were defined in accordance with the ACLED codebook. These events included battles, violence against civilians, and explosions/remote violence.

The performance of classification algorithms was determined from the receiver operating characteristic (ROC), AUC, accuracy, sensitivity and specificity metrics. The ROC and subsequently the AUC provide useful insights into the overall performance of classification models. As defined by Yoshida (2022b), the ROC curve indicates the overall performance of a classification model. An ROC that is a straight line at a 45° angle indicates that a classification model is randomly assigning classes. An ideal ROC curve would have an immediate vertical increase, resulting in a 90° bend. As previously stated, a byproduct of the ROC curve is the AUC. A perfect ROC curve would have an AUC of one.

Another indicator of a classification model's performance comes from the accuracy, sensitivity, and specificity values. To understand these metrics, it is important to first start by understanding a confusion matrix. An example of a confusion matrix, as adapted by Yoshida (2022b) is shown in Table 3.4.

Table 3.4. Example of a confusion matrix. Adapted from Yoshida (2022b).

		Predicted	
		Positive	Negative
Actual	Positive	a	b
	Negative	c	d

The equations for the accuracy, sensitivity, and specificity as defined by Yoshida (2022b).

$$Accuracy = \frac{a+d}{a+b+c+d}, \tag{3.4}$$

$$Sensitivity = \frac{a}{a+b}, \tag{3.5}$$

$$Specificity = \frac{d}{c+d}. \tag{3.6}$$

where a is the number of true positive predictions, b is the number of false positive predictions, c is the number of false negative predictions, and d is the number of true negative predictions. These are best understood through the use of the confusion matrix shown in Table 3.4.

Regression Analysis

The regression problems used RF and linear regression. The first problem attempted to predict the number of violent events in each grid. This definition of violent events is the same found in Section 3.4.2. There were several methods used to analyze the results from this analysis. These include the calculation of time series analysis metrics across the RH time steps and spatial analysis of frequently incorrect regions. Those time series metrics are the same defined in Equations (3.1) to (3.3).

3.4.3 Determining Predictor Importance

To satisfy one of the main goals of this thesis, the predictors that are important to the number of violent events with fatalities needed to be identified. The RF models used the built-in importance feature from the randomForest package from Liaw and Wiener (2002). The logistic regression used the varImp() function from the caret package by Kuhn (2022).

The variable importance for the RF models is determined by the mean decrease Gini (MDG) (Liaw and Wiener 2002). The Gini index is calculated using Equation (3.7) (Yoshida 2022a).

$$Gini = 1 - \sum_{i=1}^{C} p_i^2. \tag{3.7}$$

In Equation (3.7), p_i is the frequency of class i that is present in the set (Yoshida 2022a). The MDG used by Liaw and Wiener (2002) is an average of the the Gini reduction from the split and unsplit data across all trees. The higher the MDG the more important the predictor.

The variable importance for the logistic and linear regression models was determined by, “the absolute value of the t-statistic for each model parameter” (Kuhn 2022). The calculation of important predictors is essential to answering the problems posed in this thesis.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 4: Analysis

This chapter focuses on the implementation of the techniques discussed in Chapter 3. There is focus on the overall performance of the GNAR and ensemble models. For the ensemble model, particular emphasis was placed on the extraction of important variables from the trained models.

4.1 GNAR Model

The GNAR model attempted to predict the number of violent events for each grid over the next four months. As outlined in Section 3.3, this model utilized a grid structure that was laid on the world.

There were three separate GNAR models fit in this analysis. The first model used an alpha order of one and a beta order of one. The second model had the same alpha value but a beta value of two. The last model was a default GNAR model, with no specified alpha or beta values. The MASE values for each model were observed to provide the initial insights into the performance of the three models.

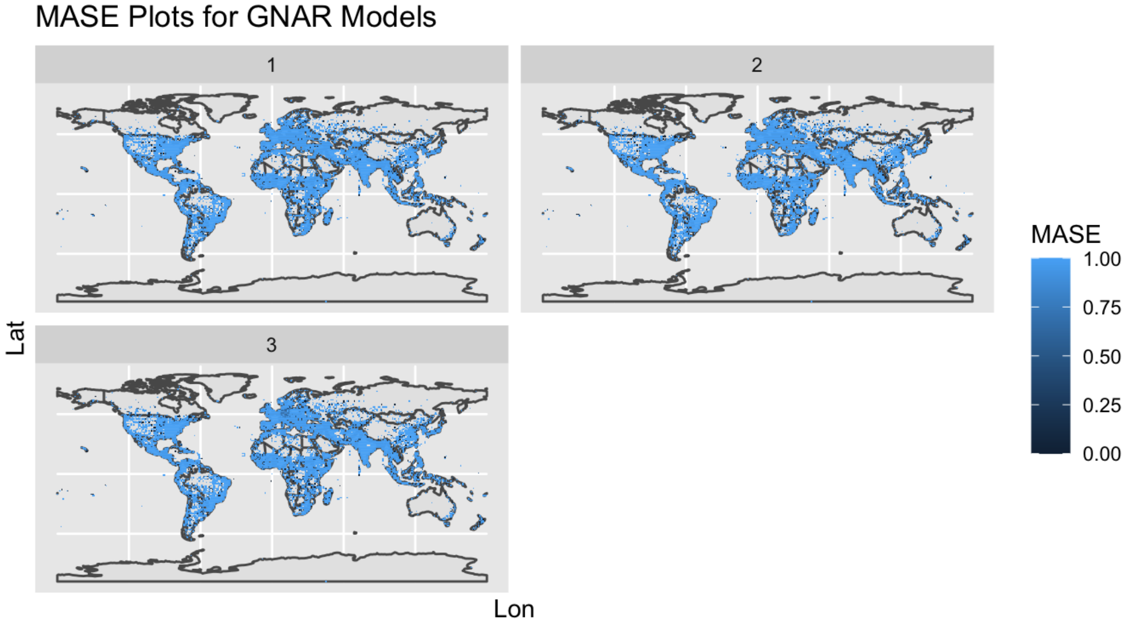


Figure 4.1. Gridded GNAR model MASE values across all models. This figure was generated using ACLED data provided by Raleigh et al. (2010).

As shown in Figure 4.1, all models have high MASE values across the globe. The fill scale is capped at one, meaning that many grids have higher MASE values. Overall, it appears in Figure 4.1 that the poor performance is not restricted to a specific region.

Table 4.1. Percentage of grids by model with MASE values greater than or equal to one.

Model	Percent of Grids with MASE ≥ 1
Model 1	91.85%
Model 2	90.93%
Model 3	88.93%

The values shown in Table 4.1 indicate that all three models exhibit little to no improvement relative to the naïve model. This is shown by the composition of grids with high MASE values for each model being 88.93% or greater across all models. Table 4.1 and Figure 4.1 indicate that there is little variation among the models performance across regions.

All GNAR models significantly under perform the naïve model. While the naïve model performed the best in terms of MAPE, it still is not a good predictor for future events. The best MAPE value for the naïve model fails to be lower than 0.2, which signifies that this model is not well suited to data.

While the GNAR model has proven to be useful in other problems, its performance relative to the naïve indicates that another approach should be investigated.

4.2 Ensemble Model

Given the poor performance of the GNAR model, it was determined that a different modeling approach should be pursued. This led to the development of an ensemble model. As outlined in Section 3.4, the ensemble model consisted of two main parts: the time series forecasting of predictors and the development of a prediction model using the forecast predictors.

4.2.1 Time Series Analysis Results

The time series analysis consisted of forecasting each of the predictors for each grid. As mentioned in Section 3.4.1, the model was fit using the auto.arima() function. Given that this function was fit to over 7,600 grids, it has a considerable run time. The long run time was mitigated by using parSapply() function from the parallel library. The broom and sweep packages by Robinson et al. (2022) and Dancho and Vaughan (2020) were also used to decrease computation time. This brought the total run time down from a week to just over one day.

The analysis of the time series forecast was completed using a RH approach. The next three month values for each predictor were forecast using the RH design. There were a total of six time-steps used in the RH for this analysis. The earliest forecast values started in June 2021 and the latest started in November of 2021. These dates were a result of when the ACLED data was pulled, which was in early 2022.

The performance metrics used to assess these forecasts were outlined in Section 3.4.1. The primary metric used for analyzing the model performance was the MASE. Initially, the MASE value was calculated for each RH. This equated to a MASE value for each of the 17 predictors, across the RH design. A plot of the most recent RH MASE values was generated in Figure 4.2.

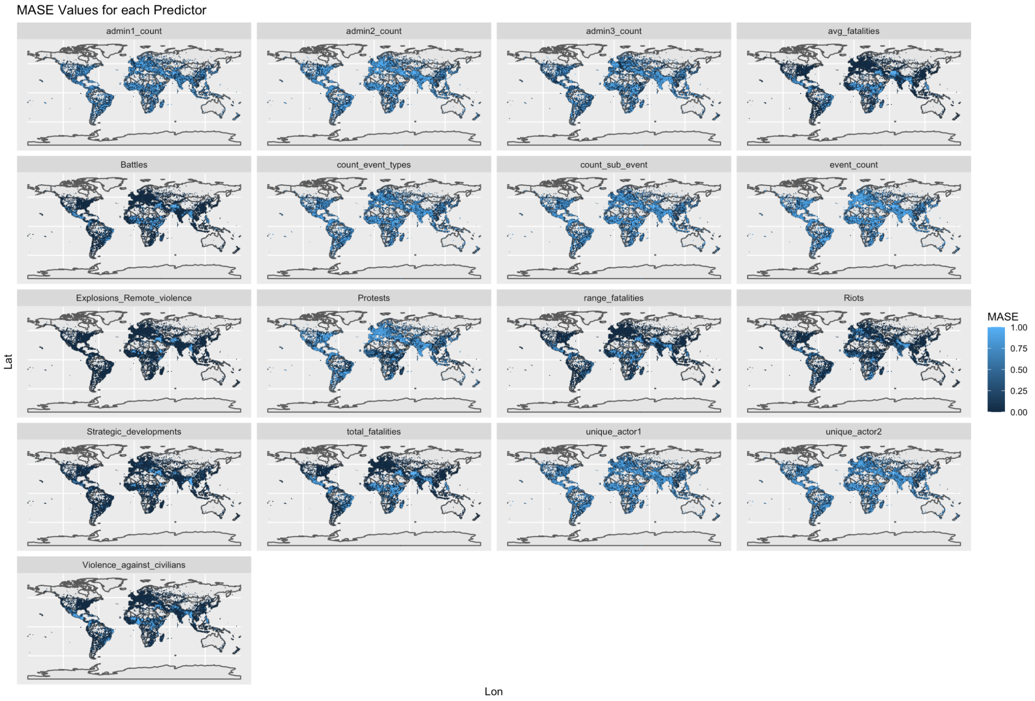


Figure 4.2. The MASE value for each grid across all predictors. A bright blue value indicates that a grid has a MASE value greater than or equal to one, signifying it's performing worse than the naïve model. This figure was generated using ACLED data provided by Raleigh et al. (2010).

The upper limit for the fill scale for Figure 4.2 was set to one, which was due to the characteristics of the MASE metric. As discussed in Section 3.4.1, a MASE value of one or higher indicates that the forecast values are no more accurate than picking the last month's value. Visually, there appear to be several predictors that are forecast poorly using the fitted model. These particular predictors are summarized in the list below.

- Unique Count of Admin1 Observations
- Unique Count of Admin2 Observations
- Unique Count of Admin3 Observations
- Unique count of Event Types
- Unique Count of Sub-Event Types
- Total Event Count
- Count of Protests
- Unique Count of Actor1 Observations
- Unique Count of Actor2 Observations.

This observation was investigated further by observing the spread of MASE values for each time series predictor.

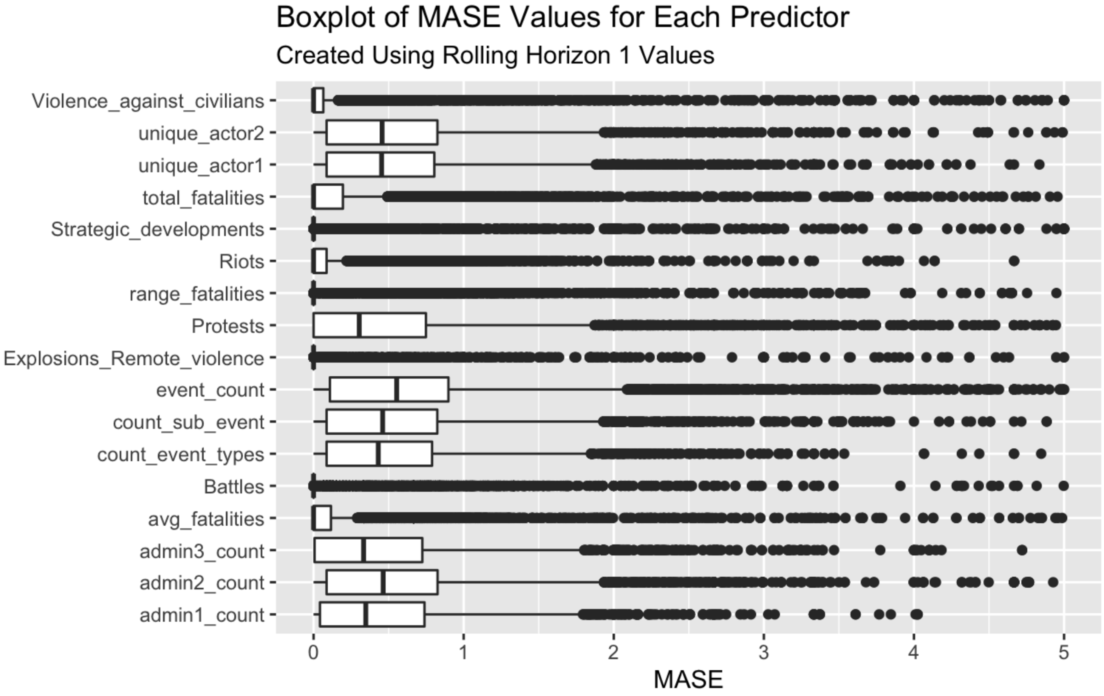


Figure 4.3. The spread of MASE values for each of the time series predictors. The third interquartile range is below a MASE value of one for all predictors. This figure was generated using ACLED data provided by Raleigh et al. (2010).

In Figure 4.3, there are several predictors whose third interquartile range stand out. These happen to be the same predictors that were mentioned in the list above. These are also the only predictors with a "maximum" MASE value above one. This "maximum" value is indicated by the whisker extending from the right side of each box. The remaining points to the right of the whisker represent outliers. For the remaining predictors, only their outliers have MASE values greater than or equal to one.

While the information from Figures 4.2 and 4.3 provides insight on the MASE values across all predictors, it is not clear how individual grids perform across all predictors. Once again, using the same RH, a tally for each grid was created for MASE values that are greater than one. High tally values would indicate grids that are frequently inaccurately forecast using the fitted model. Given that there are 17 predictors, a grid having a value of 17 would indicate that every forecast predictor was less accurate than using the naïve value.

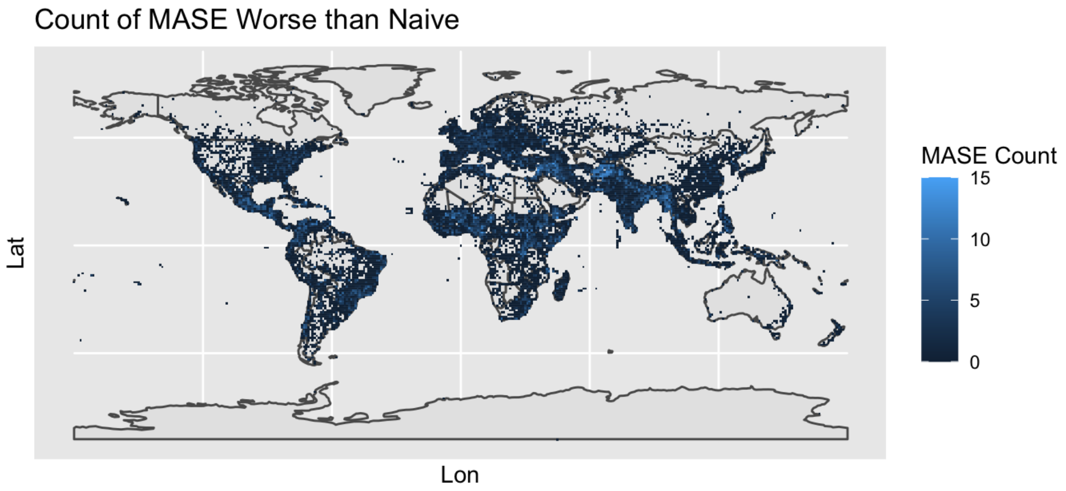


Figure 4.4. Count of predictors for each grid where the MASE value is greater than or equal to one. This figure was generated using ACLED data provided by Raleigh et al. (2010).

It appears from Figure 4.4 that there are several geographic concentration of grids where the fitted model consistently performs worse than the naïve model. These concentrations appear to be in the Middle East, Southern Asia, Central Africa, and some areas in Central

America. To better quantify these observations, a boxplot of the high MASE counts for each region was generated.

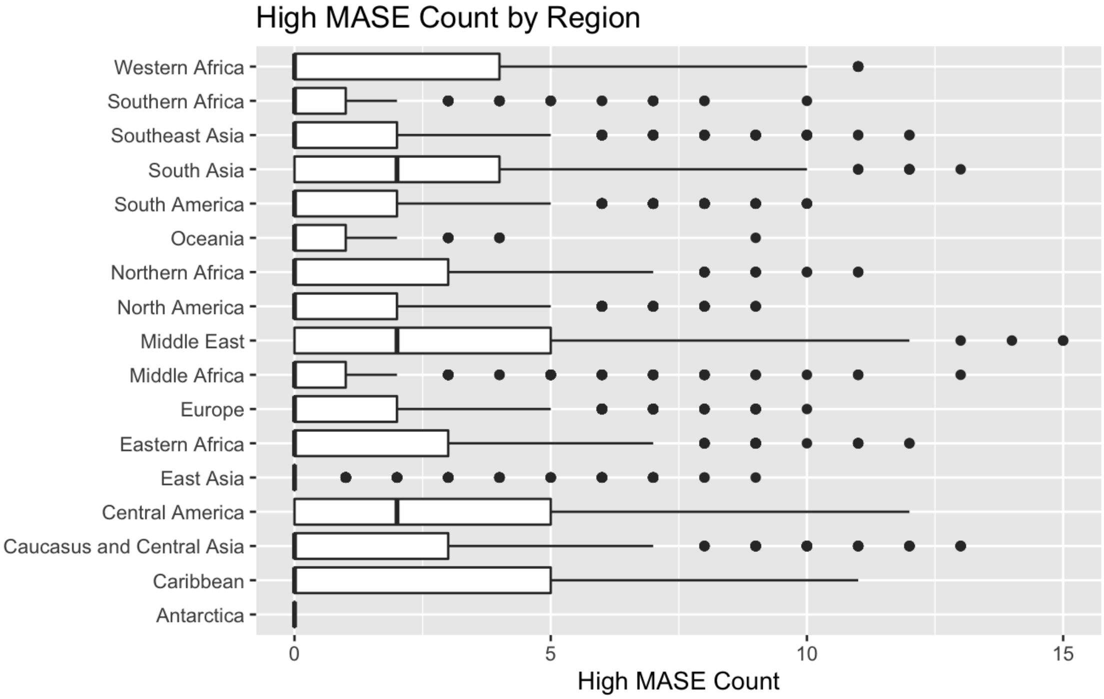


Figure 4.5. Regional breakdown of grids that have MASE values greater than or equal to one. This figure was generated using ACLED data provided by Raleigh et al. (2010).

The values shown in Figure 4.5 confirm the assumptions about which regions' forecast values perform poorly. The Middle East hosts the grid with the most occurrences of the high MASE values. Other regions that appear to have an average high MASE count greater than zero include South Asia and Central America. All other regions, as indicated by Figure 4.5, appear to have a mean number of high MASE occurrences very close to zero.

The model performance can be analyzed with more granularity by looking at the countries with the highest percentage of high MASE occurrences. This information provides more detail about which countries are not fit well by the forecast values.

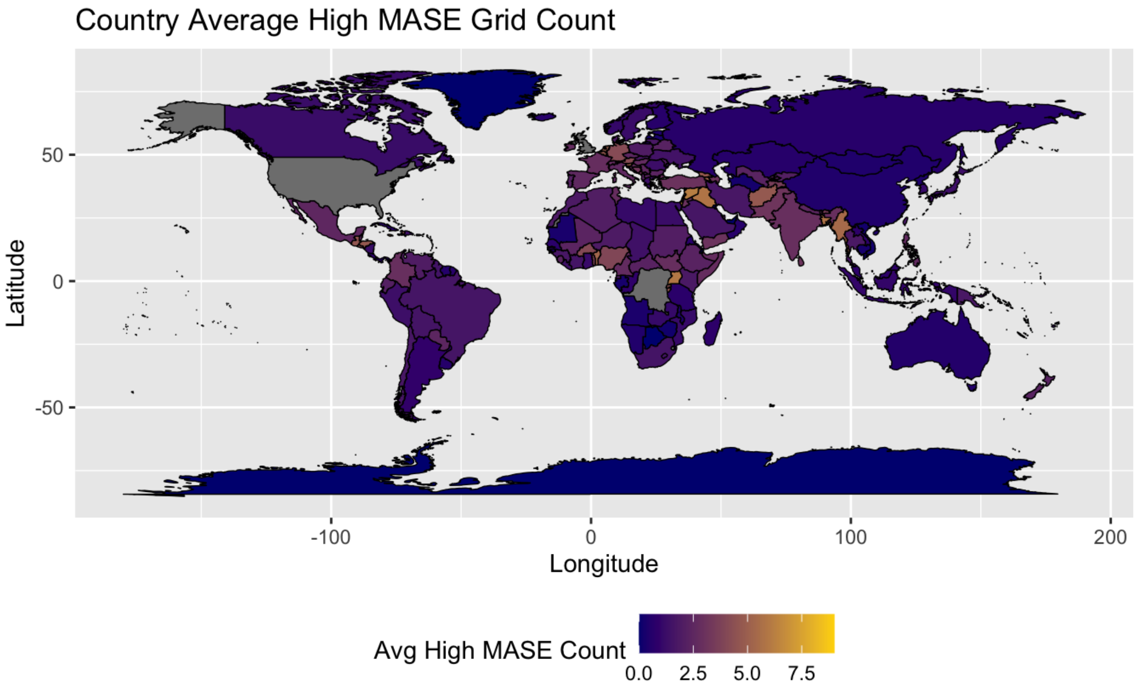


Figure 4.6. Average high MASE count for each country in time series analysis. This figure was generated using ACLED data provided by Raleigh et al. (2010).

Figure 4.6 shows that grids in several Middle Eastern and Southern Asian countries are subject to high MASE values. This suggests that perhaps a different modeling approach would better fit those areas of the world. To quantify these findings across all RHs, the top five countries with the highest average number of high MASE observations was calculated.

Table 4.2. Top five countries with highest average high MASE observations.

RH	1	2	3	4	5
1	Lebanon	Bermuda	Afghanistan	Guadeloupe	Burundi
2	Palestine	Bermuda	Lebanon	Israel	Afghanistan
3	Samoa	Afghanistan	Lebanon	Palestine	Israel
4	Liechtenstein	Afghanistan	Trinidad and Tobago	Syria	Bailiwick of Guernsey
5	Palestine	Israel	Lebanon	Burundi	Afghanistan
6	Liechtenstein	Palestine	Israel	Lebanon	Syria

There are several countries that repeat in Table 4.2. These include Afghanistan, Lebanon, Israel, and Palestine. This indicates that the time series forecasting has poor performance in several Middle Eastern countries. This is significant given the general interest by counter-terrorist agencies in the activities of terror organizations operating in this region.

4.2.2 Classification Prediction Results

From the forecast predictor values, a classification model was solved to predict if the number of violent events with fatalities would increase in a given cell. As mentioned in Section 3.4.2, the algorithms used in this analysis were RF and logistic regression. These algorithms were applied on each RH and the results compared from each.

There were two versions of the classification model that were run: one with the G-Econ predictors and the other without. The G-Econ database does not cover all the grid squares that were created from the ACLED database. Due to this fact, the determination was made that it would be beneficial to compare the results between the two versions. An initial comparison was conducted using the ROC curves for each version.

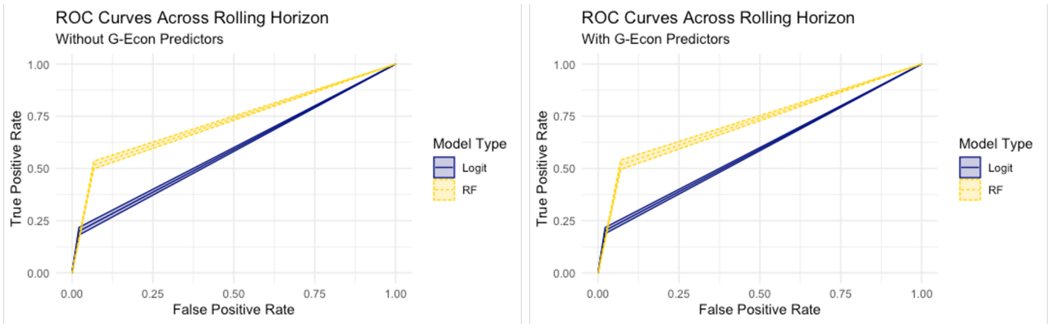


Figure 4.7. Spread of ROC curves across versions and model types. The left plot was produced without the G-Econ predictors, whereas the right plot was produced with the G-Econ predictors. These figures were generated using ACLED data provided by Raleigh et al. (2010).

An initial analysis of Figure 4.7 indicates that the RF model performs better than the logistic regression. Visually, both models perform better than randomly assigning predictions. It is also difficult to distinguish any differences between models based on the inclusion of the G-Econ predictors. This observation is best analyzed by comparing the AUC values for each model across each version. For each version, the ROC curve and the AUC was calculated for each time step in the RH design. For analysis purposes, the mean and standard deviation (SD) were calculated for each model's AUC for both versions.

Table 4.3. Summary statistics of the AUC values for each model based on the inclusion of the G-Econ predictors.

Version Type	RF		Logistic Regression	
	Mean	SD	Mean	SD
No G-Econ	0.7246	0.0090	0.5890	0.0090
G-Econ	0.7239	0.0113	0.5904	0.0076

The AUC values shown in Table 4.3 demonstrate that the inclusion of the G-Econ predictors does not have a large effect on overall model performance. For the RF model, the average AUC value decreases with the inclusion of the G-Econ predictors. The logistic regression, however, has a slight increase in the AUC value. Of the two models considered, it appears that the RF is the better performing model.

Another method to analyze classification models includes the calculation of performance metrics. Just like the ROC and AUC values, these metrics were calculated for each model with and without the inclusion of the G-Econ predictors.

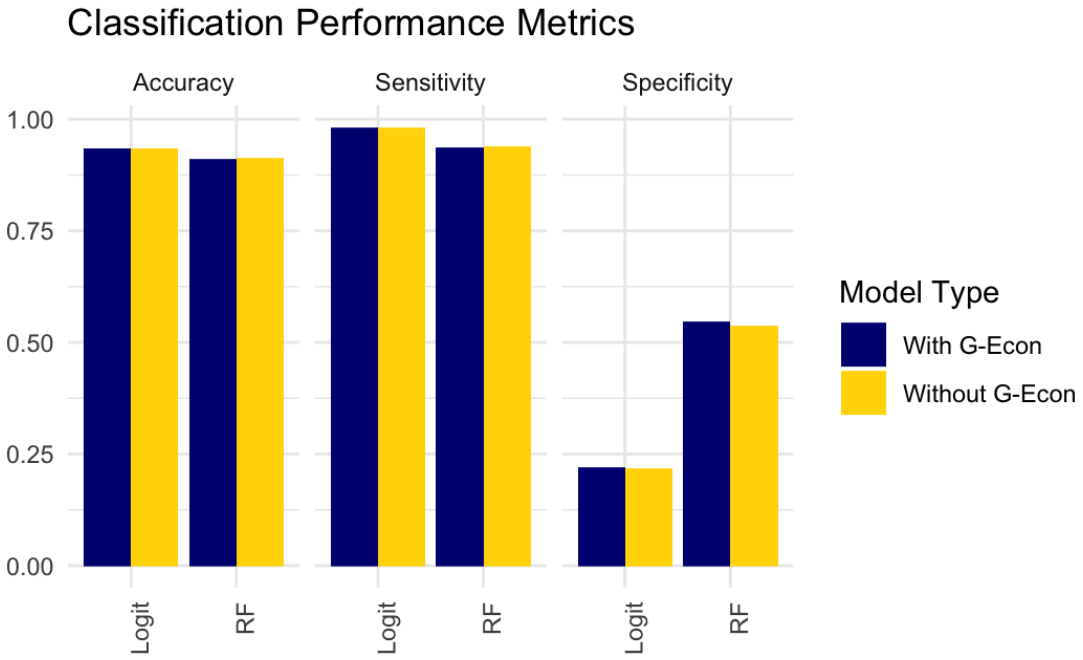


Figure 4.8. Effect of classification models performance metrics based on the inclusion of G-Econ predictors. This figure was generated using ACLED data provided by Raleigh et al. (2010).

In Figure 4.8, the values for the accuracy and sensitivity for both the RF and logistic models are nearly the same. The biggest difference between models is shown in the specificity values. In this category, the RF performs almost twice as well as the logistic regression. Figure 4.8 also suggests that the inclusion of the G-Econ predictors provides little improvement in overall model performance. Across all three metrics, it is difficult to distinguish any difference across models based on the inclusion of the G-Econ predictors.

To gain a better understanding of the geographical information contained in the predictions, some information was overlayed on a world map. To understand areas of the world that are subject to be incorrectly predicted by our models, a plot showing the percentage of time each grid was predicted incorrectly was generated. The results for the models without the G-Econ data are shown in Figure 4.9 and those models without the G-Econ data are shown in Figure 4.10.

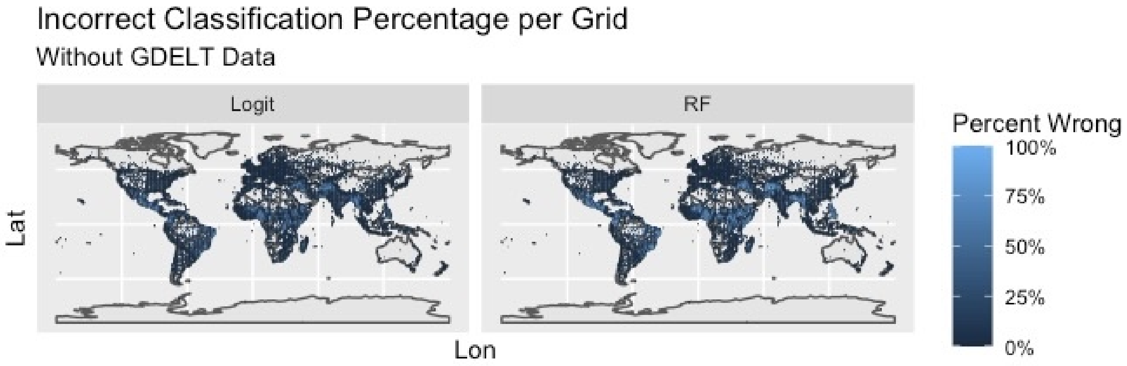


Figure 4.9. Percentage of time the classification prediction was incorrect for each grid without using G-Econ predictors. Bright blue represents the more that a grid was incorrect. This figure was generated using ACLED data provided by Raleigh et al. (2010).

The results shown in Figure 4.9 echo the same results found in Section 4.2.1. Visually, there appears to be high concentrations of inaccurate predictions in Central America, Central Africa, the Middle East, and Southern Asia. It does appear that the results from the logistic regression model perform better than those from the RF model. While both models appear to inaccurately predict in similar areas, the shading of the logistic regression in those regions is not as drastic. The same plot was generated for the model that utilized the G-Econ predictors.

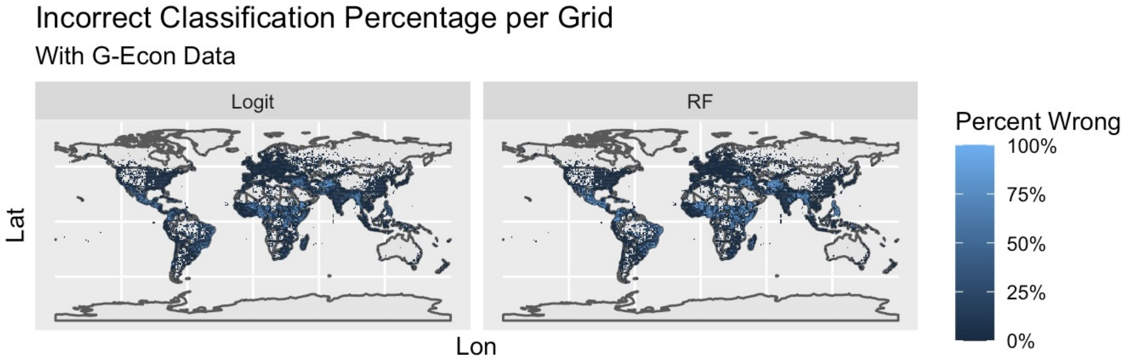


Figure 4.10. Percentage of time the classification prediction was incorrect for each grid using G-Econ predictors. Bright blue represents the more that a grid was incorrect. This figure was generated using ACLED data provided by Raleigh et al. (2010).

The concentrations of consistently incorrect grids in Figure 4.10 are very similar to those shown in Figure 4.9. Both versions of the models appear to have trouble fitting the same regions. These include the Middle East, Central Africa, Central America and Southern Asia. These observations are better evaluated by analyzing the mean, SD, and median of these values by region.

Table 4.4. Regional descriptive statistics of grid percentage of time wrongly predicted from model without using G-Econ predictors.

	RF			Logistic Regression		
Region	Mean	SD	Median	Mean	SD	Median
Central America	26.00%	30.10%	8.33%	20.8%	25.5%	8.33%
Eastern Africa	22.00%	27.80%	0.00%	15.3%	20.5%	0.00%
Western Africa	19.10%	27.50%	0.00%	14.5%	22.2%	0.00%
Middle East	15.70%	26.30%	0.00%	12.2%	20.6%	0.00%
Caucasus and Central Asia	15.20%	28.90%	0.00%	11.0%	22.6%	0.00%
South Asia	14.50%	25.50%	0.00%	9.54%	17.3%	0.00%
Middle Africa	13.50%	23.80%	0.00%	10.3%	18.3%	0.00%
Southeast Asia	12.70%	24.30%	0.00%	9.21%	18.5%	0.00%
Northern Africa	12.50%	24.10%	0.00%	7.69%	15.3%	0.00%
South America	12.00%	22.40%	0.00%	9.18%	17.2%	0.00%
Caribbean	9.87%	20.90%	0.00%	7.46%	16.6%	0.00%
North America	6.61%	18.20%	0.00%	4.71%	13.3%	0.00%
Southern Africa	4.44%	12.10%	0.00%	3.64%	9.44%	0.00%
Oceania	0.92%	5.30%	0.00%	0.61%	3.14%	0.00%
Europe	0.55%	4.76%	0.00%	0.41%	3.51%	0.00%
East Asia	0.24%	2.73%	0.00%	0.21%	2.48%	0.00%
Antarctica	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 4.4 confirms the visual assumptions about areas that are consistently misclassified. The top five regions with the highest average percentage of misclassification are the same between the RF and logistic regression models. It is also interesting that Central America is the only region with non-zero median, suggesting that both models perform particularly poorly in that region. Table 4.5 is similar to Table 4.4, but its models used the G-Econ predictors.

Table 4.5. Regional descriptive statistics of grid percentage of time wrongly predicted from model using G-Econ predictors.

	RF			Logistic Regression		
Region	Mean	SD	Median	Mean	SD	Median
Central America	26.20%	29.04%	16.70%	21.00%	25.10%	16.70%
Eastern Africa	22.50%	28.00%	16.70%	15.50%	20.40%	0.00%
Western Africa	19.30%	27.80%	0.00%	14.70%	21.90%	0.00%
Middle East	16.60%	27.00%	0.00%	12.50%	20.70%	0.00%
Caucasus and Central Asia	15.30%	29.00%	0.00%	11.60%	23.50%	0.00%
South Asia	14.50%	25.30%	0.00%	9.90%	17.60%	0.00%
Caribbean	14.50%	24.10%	0.00%	11.10%	19.90%	0.00%
Southeast Asia	13.70%	25.10%	0.00%	9.76%	19.00%	0.00%
Middle Africa	13.30%	23.10%	0.00%	10.30%	18.20%	0.00%
Northern Africa	13.00%	24.50%	0.00%	8.10%	15.90%	0.00%
South America	12.40%	22.80%	0.00%	9.09%	16.70%	0.00%
North America	6.79%	18.70%	0.00%	4.74%	13.30%	0.00%
Southern Africa	4.72%	13.00%	0.00%	3.69%	9.49%	0.00%
Oceania	1.09%	5.77%	0.00%	0.73%	3.41%	0.00%
Europe	0.54%	4.62%	0.00%	0.41%	3.26%	0.00%
East Asia	0.22%	2.70%	0.00%	0.22%	3.02%	0.00%
Antarctica	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

The metrics shown in Table 4.5 suggest that the G-Econ predictors do not meaningfully change the model performance. Compared to the median values found in Table 4.9, the median values in Table 4.5 are higher. This suggests that there are more grids present in Central America and Eastern Africa that have been incorrectly predicted.

Overall, the results of the classification problem provide some useful insights into the performance of our models. Firstly, the use of the G-Econ predictors appears to have little effect on the overall performance of the RF or logistic regression models. The RF performs the best when considering the ROC curves, AUC, and performance metrics.

Important Predictors

In addition to developing a well-performing model, another goal of this thesis is to identify those predictors that influence the probability of future terrorist actions. To do this the RF model is particularly useful, given that it inherently calculates the importance of the predictors. This feature was used to identify those predictors that are important to classification of grids that are going to have an increase in violent events with fatalities. Initially, bar plots displaying the predictors' importance for each model based on the inclusion of the G-Econ predictors were generated.

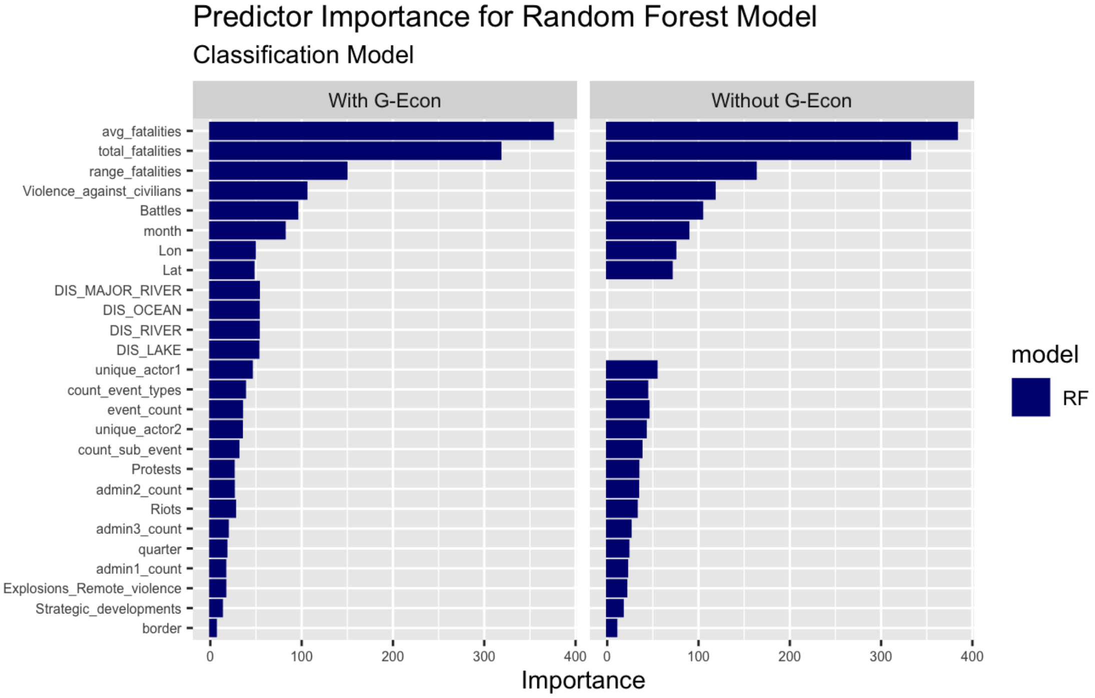


Figure 4.11. Predictor importance for the RF. Plots are separated based on the inclusion of the G-Econ predictors. This figure was generated using ACLED data provided by Raleigh et al. (2010).

In Figure 4.11 it is interesting that all four G-Econ predictors are grouped together. When they are included they appear to have the same importance values. For the rest of the predictors, the inclusion of the G-Econ predictors do not have much of an effect on their order of importance.

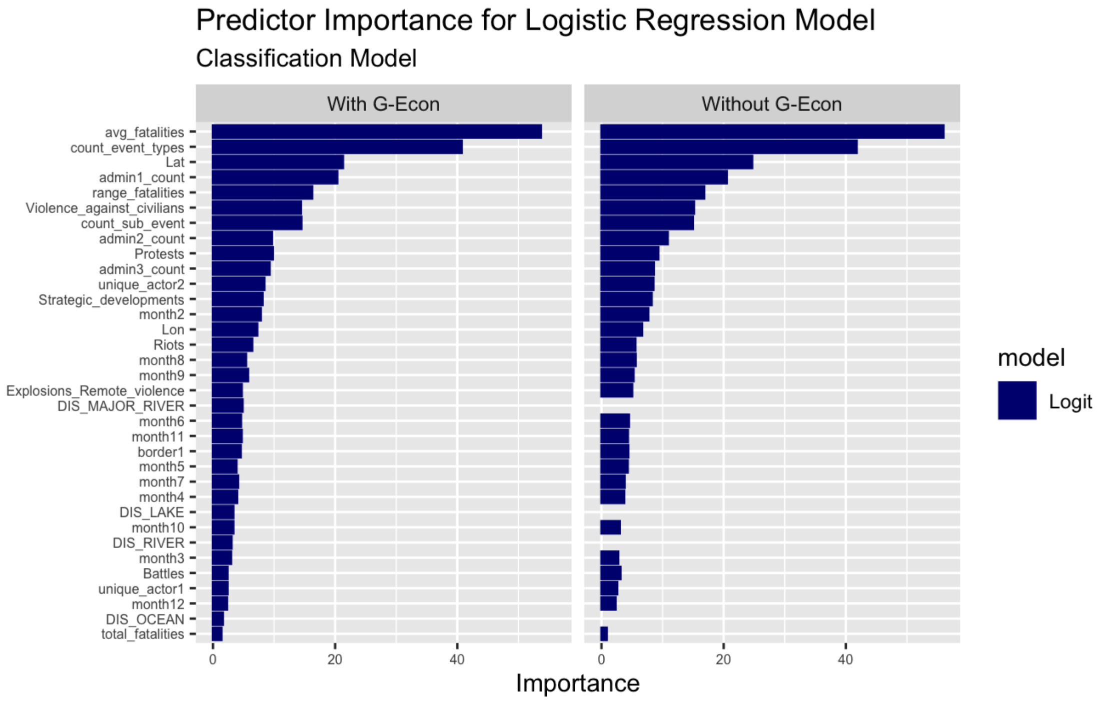


Figure 4.12. Predictor importance for the logistic regression. Plots are separated based on the inclusion of the G-Econ predictors. This figure was generated using ACLED data provided by Raleigh et al. (2010).

There is much more of a spread in the ranking of the G-Econ predictors for the logistic regression. For this model, the most important G-Econ predictor is the distance to a major river. Once again, the inclusion of the G-Econ predictors does not appear to have a large effect on the ranking of important predictors. The top five most important predictors were also included in Table 4.6. Given that the G-Econ predictors do not affect the top five performers, the table is condensed to only show the RF and logistic regression values.

Table 4.6. Classification problem top five important predictors for the RF and logistic regression models.

Rank	RF	Logistic Regression
1	Avg. Fatalities	Avg. Fatalities
2	Total Fatalities	Distinct Event Types
3	Range Fatalities	Latitude
4	Viol. Against Civ.	Distinct Admin 1
5	Battles	Range Fatalities

The most important predictors for the RF model in Table 4.6 are either statistics about fatalities or different violent event types. There is much less of a pattern present in the top five most important predictors from the logistic regression model shown in Table 4.6. The logistic regression found that statistics about fatalities, geographic information, and event information to be the most important. The identified important predictors suggest that an increase in violent events with fatalities can be attributed to the overall activity of a grid, as well as its geographic location.

4.2.3 Regression Prediction Results

A similar methodology was followed as shown in Section 4.2.2; however, the new goal is to predict the number of violent events with fatalities. As discussed in Section 3.4.2, the algorithms used for this problem are RF and linear regression. Each algorithm was applied on each of the RH results. The inclusion of the G-Econ predictors was also analyzed for both models.

Since the regression models are applied across each RH, some time series metrics can be used to analyze the results. These metrics are the MASE and MAPE. The MASE for each algorithm was displayed on the world map.

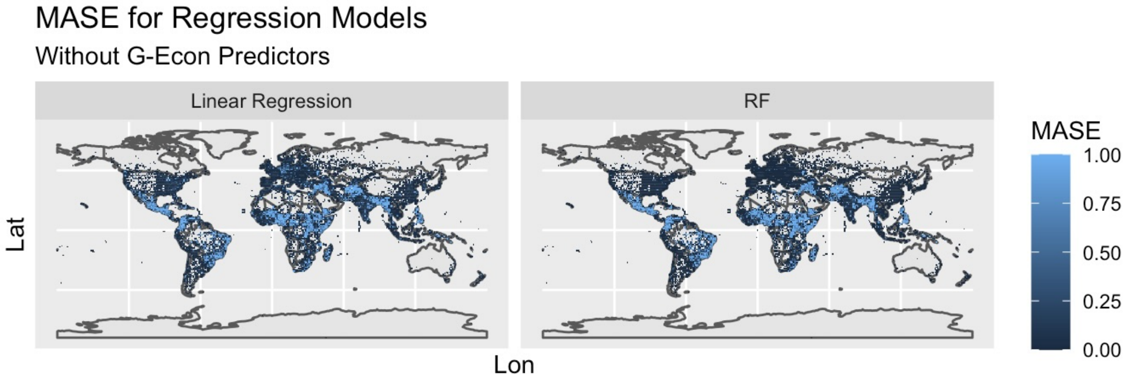


Figure 4.13. The MASE values for the regression problem using the G-Econ predictors. This figure was generated using ACLED data provided by Raleigh et al. (2010).

The same plots in Figure 4.13 was generated in Figure 4.14, but the results were generated from the models that used the G-Econ predictors.

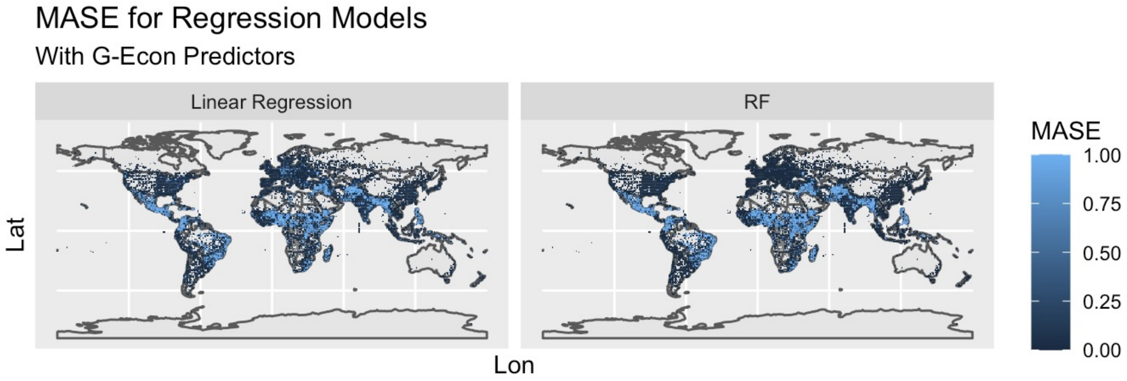


Figure 4.14. MASE values for the regression models without the inclusion of the G-Econ predictors. This figure was generated using ACLED data provided by Raleigh et al. (2010).

Measurements in terms of MASE for Figures 4.13 and 4.14 were calculated using Equation (3.2). In both versions of the models, the areas of high MASE appear to be equally dispersed. Both versions of the linear regression have areas of high MASE that are not present in the

RF models. The linear regression appears to be more susceptible to generate inaccurate predictions in areas of Europe, whereas the RF does not.

Important Predictors

The important predictors for the regression models was calculated using the same methods outlined in Section 3.4.3. The importance plots for the RF models and linear regression models are shown in Figures 4.15 and 4.16.

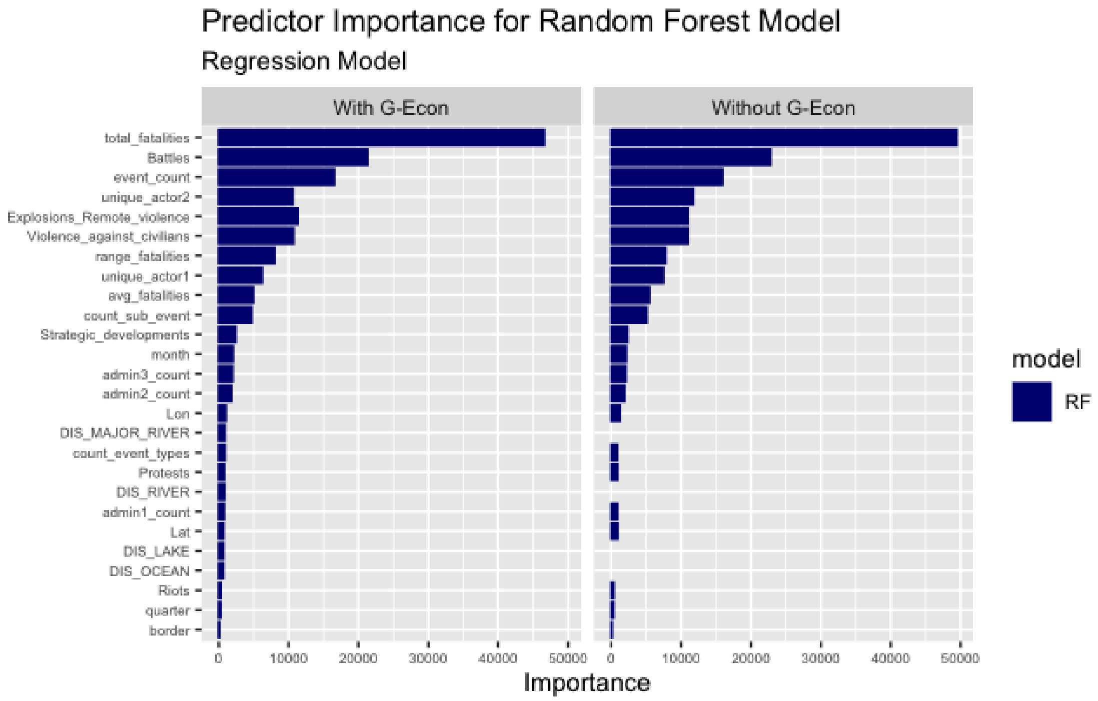


Figure 4.15. Importance plot for RF model on regression problem. Variable importance is calculated with and without the inclusion of the G-Econ data. This figure was generated using ACLED data provided by Raleigh et al. (2010).

Unlike the RF importance plots for the classification problem, the G-Econ predictors in Figure 4.15 are not grouped together.

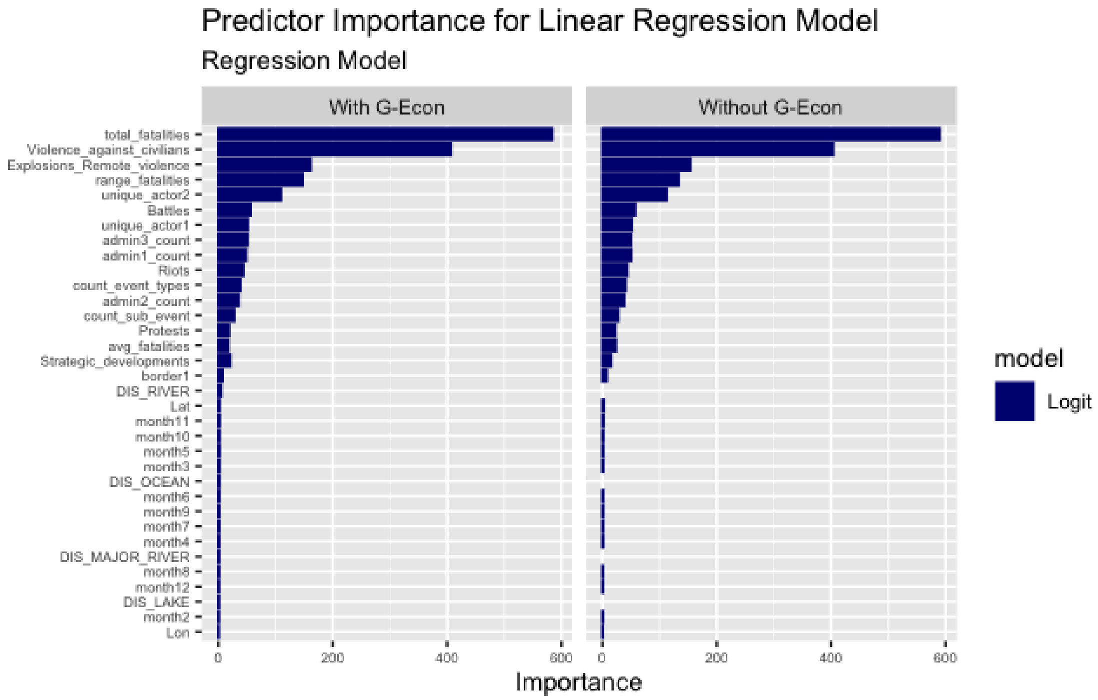


Figure 4.16. Importance plot for the linear regression model on regression problem. Variable importance is calculated with and without the inclusion of the G-Econ data. This figure was generated using ACLED data provided by Raleigh et al. (2010).

Similarly to Figure 4.15, the G-Econ predictors in Figure 4.16 are not grouped together. In both Figures 4.15 and 4.16 the G-Econ predictors are not even in the top 10 most important predictors. Table 4.7 shows the top five important predictors for the RF and linear regression models. This table does not focus on the inclusion of the G-Econ predictors because their inclusion had no effect on the top five most important predictors.

Table 4.7. Top five most important predictors for the regression problem. These are the same for the RF and linear regression models even with the inclusion of the G-Econ predictors.

Rank	RF	Linear Regression
1	Total Fatalities	Total Fatalities
2	Battles	Viol. Against Civ.
3	Event Count	Explosions and Remote Viol.
4	Distinct Actor 2	Range Fatalities
5	Explosions and Remote Viol.	Distinct Actor 2

Unsurprisingly, the predictors shown in Table 4.7 are largely centered around fatality and event statistics. The forecast total fatalities for a given grid is by far the most important predictor, as shown in Figures 4.15 and 4.16. Interestingly, the number of protests or similar non-violent activity does not have a larger effect on the prediction. Overall, this information is useful, as it helps provide context into the factors that are really driving the number of violent events with fatalities.

CHAPTER 5: Conclusion

While the developed prediction models do not perfectly identify areas of the world expected to have an increase in violent events with fatalities, they do provide insights into what contributes to an increase in such events. The predictor importance plots from Sections 4.2.2 and 4.2.3 provide this information.

The classification models provide some additional benefit from the naïve methods. This is shown by the ROC and AUC, as well as the performance metrics. The ROC and the AUC values for both models are both higher than the naïve method. This is shown by AUC values being greater than 0.5. The performance metrics also provide important information about the model performance. In particular, the sensitivity values (approximately 0.90) for both models indicate that they are very good at filtering out grids that will not have an increase in the number of violent events with fatalities. This is important, as it can quickly aid analysts with filtering out areas.

The prediction models developed in this thesis were also useful to identify predictors that affect the number of violent events with fatalities. In both the classification and regression problems, the important features were mainly derived from fatality or event information. This is important for the sponsors/stakeholders of this project, as it helps inform them of the types of potential predictors that could provide them with added value. It also enables the cross-validation of features the stakeholders find important, with the features the predictive models find important.

Overall, this thesis was able to provide important predictors from the ACLED and G-Econ datasets that influence the number of violent events with fatalities in different areas of the world. The developed models, particularly those for the classification model, can also provide added benefit to the stakeholders.

5.1 Future Work

There are two main directions for future work following this thesis. The first is a continuation of the work started in this thesis. The second is the implementation of more geospatial methods to derive more insight from the ACLED dataset.

5.1.1 Continuation of Current Work

It would be beneficial for future work to look at breaking down into regions of the world. Given that the sponsor, DIA, works with the Department of Defense (DOD) these areas could be broken down into the combatant commands. Using this approach would likely result in much better performance. With the current approach, our prediction methods are trained on global data while attempting to predict trends for grids of size one latitude and one longitude.

In fact, the determination of regional breakdown would be a significant piece of work in of itself. It could likely be the case that a model of the Middle East would not fully capture the dynamics of each country. For example, the predictors that influence an increase in violent events with fatalities in Syria could not be same that influence those events in Afghanistan. This research would likely be a repetitive task of fitting models over different areas and comparing the results.

Another area of improvement for this work is further fine tuning of the predictors. It should be investigated if the scaling of any predictors influences the performance of the model. For example, the forecast for the number of protests could be scaled with respect to the largest recorded value for that grid in the time series. This could perhaps provide better performance in both the classification and regression problems.

5.1.2 Application of Geo-Spatial Methods

Another area for future work is in the geo-spatial space. These endeavors could include analyzing how a group's area of operation changes over time. For example, events in which the Taliban instigated would be grouped by month. The area of operation for each month would be estimated using KDE. The results from each KDE could then be used in predictive models to determine what factors influence changes in their area of operation. This has the potential to be scaled to include multiple groups. When scaled, the intersections of their areas of operation can be analyzed as well to help predict areas likely to have an increase in terrorist activities.

THIS PAGE INTENTIONALLY LEFT BLANK

List of References

ACLED (2019) *Armed Conflict Location & Event Data Project (ACLED) Codebook.* https://acleddata.com/acleddatanew/wp-content/uploads/2021/11/ACLED_Codebook_v1_January-2021.pdf.

ACLED (2022) ACLED | Bringing Clarity to Crisis. https://acleddata.com/.

Dancho M, Vaughan D (2020) *sweep: Tidy Tools for Forecasting*. https://CRAN.R-project.org/package=sweep, r package version 0.2.3.

Ding F, Ge Q, Jiang D, Fu J, Hao M (2017) Understanding the dynamics of terrorism events with multiple-discipline datasets and machine learning approach. *PLOS ONE* 12(6):11, ISSN 1932-6203, http://dx.doi.org/10.1371/journal.pone.0179057.

Hamner B, Frasco M (2018) *Metrics: Evaluation Metrics for Machine Learning*. https://CRAN.R-project.org/package=Metrics, r package version 0.1.4.

Hao M, Jiang D, Ding F, Fu J, Chen S (2019) Simulating Spatio-Temporal Patterns of Terrorism Incidents on the Indochina Peninsula with GIS and the Random Forest Method. *ISPRS International Journal of Geo-Information* 8(3):133, ISSN 2220-9964, http://dx.doi.org/10.3390/ijgi8030133.

Hyndman RJ, Khandakar Y (2008) Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software* 26(3):1–22, http://dx.doi.org/10.18637/jss.v027.i03.

Kuhn M (2022) *caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret, r package version 6.0-93.

Leeming K, Nason G, Nunes M, Knight M (2020) Methods for Fitting Network Time Series Models https://cran.r-project.org/web/packages/GNAR/GNAR.pdf.

Leetaru K (2022) The GDELT Project. https://www.gdeltproject.org/.

Liaw A, Wiener M (2002) Classification and regression by randomforest. *R News* 2(3):18–22, https://CRAN.R-project.org/doc/Rnews/.

Luo L, Qi C (2021) An analysis of the crucial indicators impacting the risk of terrorist attacks: A predictive perspective. *Safety Science* 144:105442, ISSN 09257535, http://dx.doi.org/10.1016/j.ssci.2021.105442.

Raleigh, Clionadh, Linke A, Hegre H, Karlsen J (2010) Introducing ACLED-Armed Conflict Location and Event Data. *Journal of Peace Research 47(5) 651- 660* .

Robinson D, Hayes A, Couch S (2022) *broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom, r package version 1.0.0.

START (National Consortium for the Study of Terrorism and Responses to Terrorism) (2022) Global Terrorism Database 1970-2020. https://www.start.umd.edu/gtd/.

World Bank (2022) World Bank Open Data | Data. https://data.worldbank.org/.

Yoshida R (2022a) Random Forest. Unpublished, Naval Postgraduate School.

Yoshida R (2022b) Receiver operating characteristic. Unpublished, Naval Postgraduate School.

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California