Masters Program in Geospatial Technologies

Spatial Conflict Prediction with Machine Learning
Conflict Vulnerability in the Sahel Region

Frank Guzzardo

Dissertation submitted in partial fulfilment of the requirements
for the Degree of *Master of Science in Geospatial Technologies*

NOVA IMS Information Management School
Universitat Jaume I
Westfälische Wilhelms-Universität Münster

Spatial Conflict Prediction with Machine Learning

Conflict Vulnerability in the Sahel Region

Dissertation supervised by

Professor Pinherio

Professor Sospedra

Professor Painho

March of 2022

DECLARATION OF ORIGINALITY

I declare that the work described in this document is my own and not from someone else. All the assistance I have received from other people is duly acknowledged and all the sources (published or not published) are referenced.

This work has not been previously evaluated or submitted to NOVA Information Management School or elsewhere.

Lisbon, March 1st

Frank Guzzardo

ACKNOWLEDGMENTS

First and foremost, I would like to thank my primary supervisor, Flavio Pinheiro for his support and advice throughout this research. Additionally, I would like to express my gratitude to Professor Marco Painho for his guidance and invaluable support throughout the Masters Program.

SPATIAL CONFLICT PREDICTION WITH MACHINE LEARNING

Conflict Vulnerability in the Sahel Region

ABSTRACT

In the last few decades there has been a steady increase in intrastate conflict around the globe. In response, there is a rising need for actionable information for national and international stakeholders to better forecast and mitigate the effects of intrastate conflict. The Sahel region is especially vulnerable to intrastate conflict suffering a multidimensional crisis that includes climate change, food insecurity, and the proliferation of armed conflict. This study seeks to explore the feasibility of producing a heuristic machine learning model utilizing open-source data to predict localized intrastate conflict events on a regional scale using the random forest regression algorithm. The model includes data from 2007 to 2020 selected from multiple sources to create 17 features representing real-world phenomena to predict conflict occurrence. A unified spatial data structure consisting of quadratic grid cells was used for local-level analysis. Implementing a 10-fold cross-validation method, the model performed well with an RMSE of 1.394 and an R2 of .95. There was an improvement of 76% from the baseline model.

KEY WORDS

Sahel

Conflict

Random Forest

Prediction

ACRONYMS

RMSE – root mean squared error

R2 – R (regression) squared

ACLED- Armed Conflict Location and Event Data

WFP- World Food Programme

INDEX

INDEX OF TABLES

INDEX OF FIGURES

1. INTRODUCTION

1.1 Context and Motivation

Since World War II, conflicts between states or interstate conflicts have drastically declined while interstate conflicts have become ever more common. Between 1945 and 2000, it is estimated there have been 25 interstate wars with more than 100 fatalities with a median duration of 3 months. In the same period, it is estimated that more than 127 intrastate conflicts have cost at least 1000 lives. (Fearon & Laitin, 2003) Moreover, the estimated total cost of life has been five times higher for intrastate conflicts than interstate conflicts in the same period. This decrease in conflicts between state forces can be attributed to several factors. Since the end of WWII and the subsequent establishment of the United Nations Organization (UN), there has been a generalized attempt to avoid further conflicts.

Two main theories are omnipresent in conflict studies that help explain the decline of interstate warfare. Some experts believe that this trend stems from the global increase of democracies and is based on the theory that accountable leaders are less willing to engage in potentially costly and unpopular decisions such as going to war. Another theory attributes the increase in international trade, creating an economically interdependent world where actors could possibly go against their economic interests by engaging in warfare. Although this trend is reassuring, these theories do not explain the rise of intrastate conflicts in many parts of the world. (Clauset, 2018; Backer et al., 2014)

The decrease of interstate wars, both in frequency and magnitude of the destruction, is a welcome development; however, the destruction and instability that intrastate conflicts cause should not be overlooked. Generally, intrastate conflicts can be divided into several categories: war of secession, war of succession, violence waged by terrorist or criminal organizations, state-sanctioned violence, and resource-driven conflict. Wars of succession are characterized by citizens fighting to overthrow a ruling party or government, while wars of secession can be described as citizens attempting to form their government and sovereignty outside of the political confines of the status quo. A quest for territorial gain and control of the people in those territories is a common

strategy of drug cartels and terrorist groups. State-sanctioned violence often occurs when a government attempts to stoke ethnic tensions to control or discourage dissent while preserving social hierarchy by persecuting minorities. Lastly, many conflicts between citizens and governmental entities stem from corruption, lack of economic opportunities, and struggle for control over territories and natural resources. (Newman, 2014)

Many experts believe it is critical for the global community to consider intrastate conflicts because they are more frequent and quantifiably more destructive and because they have tangible ramifications far outside the areas where they occur. Intrastate conflict zones are de facto safe havens for terrorist and criminal organizations where these groups can operate with impunity. (George et al., 2020) Additionally, high conflict zones can be a hotbed for infectious disease propagation, such as the recent Ebola crisis in West Africa. Furthermore, intrastate conflicts drive today's refugee crises far from the epicenter as people flee affected areas without much choice. These effects can be profoundly destabilizing to regions suffering high conflict occurrence and disrupt the international system in an interconnected world. (Backer et al., 2014)

Since the turn of the century, there has been a drastic increase in interest regarding intrastate conflicts. In response, numerous quantitative studies on the causes and nature of intrastate conflicts have been conducted, particularly the application of a variety of statistical methods to better identify civil war factors with varying degrees of success. Despite this pivot to the quantitative investigation of this phenomenon, there is still much ambiguity regarding the core causes of intrastate conflict. As computational power increases and large stores of diverse and open-source data become more accessible, coupled with enhanced methodology and machine learning algorithms, the potential to isolate the causes or factors which could be attributed to the intrastate conflict needs to be further investigated. Additionally, become possible to better assess fragility and vulnerability to intrastate conflict and even predict future matches with increasing accuracy. (Python et al., 2021)

One region where there has been a steady increase of intrastate conflict in the past 20 years, the Sahel, is a vast expanse of arid to semiarid terrain, which is home to over 135 million people. Located on the southern edge of the Sahara Desert, this diverse region includes roughly 19 countries with various ethnic groups, both nomadic and sedentary, subsisting on both agricultural and pastoral livelihood activities. Throughout this region, communities are facing a multidimensional crisis that includes climate change, food insecurity, and the proliferation of armed conflict, to name a few. With vast semi-autonomous regions, an array of paramilitary groups, and relatively weak governmental structures, conflict in the region is steadily increasing, and all signs suggest that this trend will only continue. (Raleigh et al., 2015)

Recent machine learning research has assessed the feasibility of producing robust models to predict conflict with favorable results. According to the literature, there is much room to build on these models and compare machine learning predictive algorithms to forecast intrastate conflict using a diverse group of datasets on a subnational level. (Ettensperger, 2020)

1.2 Objectives and Research Questions

The main objective of this research is to construct a heuristic machine learning model utilizing available data to predict the occurrence of localized intrastate conflict events on a regional scale using the random forest regression algorithm. The secondary goal of this study is to identify potential hotspots for conflict and better understand the environment that they take place. Finally, we identified a set of potential predictors for conflict in the region based on a literature review and the random forest regression algorithm.

The three objectives of this study can best be encompassed with the following research questions:

1. How can conflict vulnerability be modeled based on environmental and socio-economic factors on the local level?
2. How can conflict event occurrence be predicted using a random forest regression machine learning algorithm applied to a unified spatial data structure?

3. What socio-economic and environmental factors are the best predictors for intrastate conflict?

This study is unique because it sets out to predict conflict on a uniform grid system. Several studies have been conducted using vast amounts of environmental, political, and demographic data on state, regional, or administrative scales. Although these studies have contributed significantly to understanding intrastate conflicts and machine learning application in the field, they have arguably fallen short in predicting intrastate conflict on a truly local level. Seldomly does intrastate conflict respect administrative or national boundaries; therefore, this study attempts to deconstruct the "boundary" approach and applies a uniform grid mapping unit. The uniform grid constructed by the Research Council of Norway serves as the backbone of the study. A multitude of data, including satellite imagery and data derived from satellite imagery and empirical studies, has been extracted to investigate the ability to predict conflict in the Sahel region on a local level. (Tollefsen et al., 2012)

1.3 Thesis Organization

This document presents our research in 5 chapters. Chapter 1 introduces the reader to the topic, provides the work's motivation and contextual background, and lists the objectives and research questions to fulfill. Chapter 2 reviews related studies, attempts to identify the key factors affecting the research approach, and analyzes the theoretical background necessary to follow the present study. Chapter 3 explains the data and the methodological workflow of the thesis. Chapter 4 presents and discusses the results of the study, and Chapter 5 summarizes the main conclusions found and suggests some future works on the matter.

2. LITERATURE REVIEW

2.1 Predictors of Conflict

A thorough investigation into predictors and factors of intrastate conflict was conducted to identify relevant actors and general characteristics of intrastate conflict events. This research is intended to better select potential datasets for the analysis. This was a

critical point of the study as a cursory investigation has shown that there is little consensus between experts in conflict studies on what constitutes an "intrastate conflict," who are the actors, and a general understanding of the nature of these events. Fortunately, there has been an increase of work in conflict studies in recent years, particularly in the development of conflict predictor frameworks.

Greed vs. Grievance

In the last two decades, many intrastate conflict studies focused on the "greed" vs. "grievance" dichotomy. Political science generally deals with intrastate factors from the "motive" perspective. Several research undertakings have featured the relationship between political grievances against controlling institutions and the economic needs and desires of rebel groups and other non-state actors. Following this line of thought, a rebellion or conflict occurs when there is a sufficient level of "grievance" that people are willing to engage in violent protest or conflict. On the other side of this theory, an economic perspective considers rebellion as a profit-generating industry. (Collier et al., 2004)

In Greed and Grievance in Civil War, Collier and Hoeffler investigate the causes of intrastate conflicts utilizing a dataset of wars between 1960 and 1999. The authors argue that "rebellion" may be explained by unusually severe "grievances," such as high inequality, lack of political representation, or ethnic and religious divisions in society. On the "greed" side, they introduce the idea that causes of intrastate conflict can also be explained by unique opportunities or "greed" to build rebel organizations. Examples of the authors' atypical opportunities are extortion of natural resources, remittances or donations from diasporas, and foreign aid from hostile 3rd governments. This research finds that traditional social and political variables often used as proxy metrics for "grievances" tend to lack exploratory power. Collier and Hoefler conclude that the model that focuses on opportunity or greed generally performed better than the model based on grievances. The best predictor in their model was the primary commodity exports as a percentage of GDP. (Collier & Hoeffler, 2004)

Inequality

Buhaug et al. provide a distinction between horizontal and vertical inequality and argue that horizontal inequality or inequality between groups as a whole are generally better predictors of conflict occurrence than vertical inequality, or inequality between individuals. Simply put, intrastate conflict occurs between groups of people rather than individuals. Collier and Hoeffler's research also state that proxy for grievances includes ethnic or religious "hatred". This sentiment cannot be quantified, but it is safe to assume that it is only possible in multi-ethnic or multi-faith communities. The work further clarifies and provides evidence that diversity alone doesn't provide the environment for conflict, but ethnic polarization and the struggle for dominance do. The ethnic distinction in an area can be conducive to war when a particular group constitutes a clear majority and other groups are powerless or perceive being powerless. (Buhaug et al., 2014)

Environmental Factors

James Fearon's and David Laitin's work discuss the role of state capacity in intrastate conflict vulnerability. The authors argue that poverty, rugged terrain, and a weak central government are highly favorable conditions for rebellion and therefore produce an environment where the likelihood of intrastate conflict is increased. The study argues that poverty, environmental, and political factors are better predictors of intrastate conflict than state discrimination, ethnic and religious tension, and even economic inequality. It's worth mentioning that the authors' findings demonstrate that rebel groups' ability to hide from the government enables small groups to sustain an armed campaign against a larger government and make it more difficult for the government to intervene or eradicate the adversary. (Fearon & Laitin, 2003)

Politics

Goldstone et al.'s study examines onsets of political instability on the global scale from 1955 to 2003. Their model performed with 80% accuracy and distinguished between states that experienced instability and those that did not within a 2-year time frame. The model successfully predicted onsets of intrastate conflict and nonviolent democratic reversal based on a non-linear five-category regime measure derived from the Polity dataset. The new measure of regime types designed through the study proved to be a

much more significant predictor of intrastate conflict onset than predictors based on economic conditions, demographics, or geography. This new measure of regime type emerges as the most potent predictor of instability onsets, leading them to conclude that political institutions and not economic conditions, demography, or geography are the most important predictors of the beginning of internal political instability that may lead to intrastate conflict. The other predictors used were infant mortality, conflict in neighboring countries, and state-led discrimination. (Goldstone et al., 2010)

Food Security and Climate

Raleigh et al.'s work use the Armed Conflict Location and Event Data (ACLED) dataset to analyze the interrelations between climate change, food price, and conflict in Africa ranging between 1997 and 2010. The authors found a clear connection between the three factors, and they found that "across Africa, conflict increases the price of commodities, which, in turn, increases the rate of political violence." Furthermore, "...feedback exists between food price and political violence: higher food prices increase conflict within markets, and conflict increases food price". To explain this theory, the research found that conflict tends to impact food security conditions by destroying agricultural production, distribution, and markets, hindering economic growth, and increasing unemployment levels. It is worth mentioning that this study does not isolate intrastate conflict events from interstate conflicts, instead it uses all events in the geographic area of interest-based on the assumption that there is no standard reaction to climate crisis events or food price volatility. (Raleigh et al., 2015)

2.2 Machine Learning in Conflict Prediction

The work of Collier Hoeffler marked a tangible pivot towards quantitative conflict research. Since this seminal study, research has shown that predicting conflict utilizing machine learning algorithms has proven relatively accurate. Chris Perry states that "machine learning techniques could provide significant contributions to tactical early warning systems and conflict prevention strategies in particular if leveraged intelligently" (Perry, 2013; Collier et al. 2004)

Additionally, the Peace Research Institute Oslo (PRIO) Centre for the Study of Civil War also touches on the importance of applying full machine learning models instead of simple past outbreak histories of conflict. (Python et al., 2021) Hilaire Meire's work focuses on predictive models to forecast civil wars and whether non-parametric tree-based models are better suited for this purpose than parametric methods such as logistic regression. According to Hilaire Meire, tree-based ensemble methods generally perform significantly better than logistic regression algorithms. (Meire, 2017)

In Perry's 2013 research, he highlights that choosing appropriate machine learning methodology significantly impacts the performance and accuracy of the predictions, with some algorithms offering significant improvements. In this research, a 4-fold random forest predictor was performed to model fragility and vulnerability to conflict at the global level. This feasibility study introduced several lessons learned that can benefit future studies. The author mentions narrowing the geographic scope to limit missing data and possibly create a hierarchy of risk factors or predictors for a specific context. Perry cautions that machine learning tends to over predict outliers or conflict events that are very rare in this case. He used two methods to set a target for prediction; the first, a simple binary class indicating if one or more battles occurred in each district or not. The second was a numeric count of battles in a given district each year. The random forest algorithm performed significantly better than the Naïve Bayes predictor, and both algorithms outperformed their respective baseline metrics. This study further supports the notion that machine learning algorithms have lots of potential in conflict prediction and social sciences at large. (Perry, 2013)

Ettensperger's 2019 research into several supervised learning algorithms and artificial neural networks for global conflict prediction introduces a novel dataset including seven socio-economic and political indicators coupled with six years of conflict intensity data to build a robust predictor framework. The study tested and compared several predictive methods, including classification and regression trees, k-nearest neighbor, random forest, and several advanced artificial neural networks. Ettensperger mentions that modeling vulnerability to intrastate conflict on the global scale is still dominated by conventional regression methods that tend to be inflexible and ill-suited

to model complex non-linear interactions between features and contexts in general. The random forest regression in the study improved the prediction of the model by over 20% compared to linear regression and mixed-effects multi-level regression models. This highlights that the relationship between the socio-economic, political, demographic, dependent variables is not linear. The work further suggests that modern supervised learning algorithms should be assessed in this capacity. (Ettensperger, 2020)

The Python et al. 2021 study predicts non-state terrorism on a global scale. The research attempts to construct models for reliable and short-term predictions of non-state terrorism on a local level utilizing open-source data and the Prio-Grid unified spatial data structure. The study concludes with the assertion that structural and procedural predictors can accurately predict events that happen in the same or subsequent week. The authors highlight the importance of theoretically informed models and conclude that these models systematically outperform models using predictors based on past event occurrences. In this study, the random forest regressor performed 63 times better than the baseline predictor, and in high event, areas was the best predictor overall. (Python et al., 2021)

3. DATA AND METHODOLOGY

3.1 Data Description

A large portion of time and attention was spent ensuring that the data and the derived features were relevant predictors of conflict. A causal relationship was not a priority when selecting features. On the other hand, care to accurately represent real-world events or phenomena that could be used as proxy indicators was highly prioritized.

The area of study is the general Sahel region, a loosely defined geographic zone located where the arid Sahara desert transitions into the semi-arid Savannah, stretching across the south-central latitudes of northern Africa between the Atlantic Ocean the Red Sea. Nineteen countries that are found in the Sahel have been selected for this analysis, including Burkina Faso, Benin, Central African Republic, Cameroon, Djibouti, Eritrea, Ethiopia, Ghana, Gambia, Mali, Mauritania, Niger, Nigeria, Sudan, Senegal, Somalia,

South Sudan, Chad, and Togo. Data from 2007 to 2020 has been selected from multiple sources culminating in a random forest regression analysis utilizing 17 features representing a diverse array of real-world phenomena to predict conflict occurrence on a regional scale. Data selection was an iterative process based on the literature review and availability of open-source data.

A unified spatial data structure covering the entire globe with quadratic grid cells measuring approximately 50km x 50km at the equator called the Prio-Grid was used as the spatial mapping unit for the analysis. This spatial structure allowed for local-level analysis and can be scaled up or down dependent on the intended purpose of the research. The grid cells are fixed in space and entirely disregard political boundaries and geopolitical developments, much like the features presented in this study. The datasets used in this research are shown in Table 1. (Tollefsen et al., 2012)

Source	Derived Features	Primary Source	Units
Google Earth Engine	Maximum temperature (tmmx)	TerraClimate	Degrees Celsius
	Palmer Drought Severity Index (pdsi)	TerraClimate	Index
	Precipitation Accumulation (pr)	TerraClimate	Millimeters
PRIO-GRID	Landcover (7 classes)	Globcover 2009 dataset v.2.3.	Percentage of area
	Mountain mean	UNEP Mountain Watch Report	Proportion
	Harvest area	MIRCA2000 dataset v.1.1	Hectares
	Distance to Territorial Boundary	CShapes Dataset	Kilometers
	Distance to Capital	CShapes Dataset	Kilometers
ETH Zurich	Overlap of ethnic groups within 100km	ETH Zurich Department of Humanities, Social, and Political Sciences	Count
The Humanitarian Data Exchange, OCHA	Market Food Price Volatility Index	World Food Programme, United Nations	USD equivalency
Armed Conflict Location & Event Data Project	Conflict Occurrence Index	Armed Conflict Location & Event Data Project (ACLED)	Count

Table 1: Data Description

3.1.1 TerraClimate, University of Idaho

TerraClimate is a monthly dataset monitoring climate and climatic water balance for global terrestrial surfaces. This data was sourced from the Google Earth Engine data catalog and three bands were selected for the study including the “tmmx” band which records maximum temperature, the “pr” band which records precipitation accumulation, and the “pdsi” band which calculates the Palmer Drought Index (PDSI). The PDSI utilizes open-source temperature and precipitation data to calculate relative dryness. The spatial resolution of all three datasets is approximately 4638 meters. (Abatzoglou et al., 2018)

3.1.2 PRIO-GRID by Research Council of Norway

The Prio Grid datasets are based on a grid structure with a wide range of data aggregation. This grid system permits for compilation, management, and analysis of spatial data within a spatially-consistent framework. The structure of the dataset consists of quadratic cells measuring approximately 3025km2 each and spanning the entire terrestrial areas of the globe. (Tollefsen et al., 2015)

For this study seven landcover features were selected that were extracted from the Globcover 2009 dataset. Each class was computed using the FAO land cover classification system used by Globcover. This dataset measures the percentage of coverage of the landcover class in each cell. The 7 classes include agricultural areas, urban areas, forest areas, shrubland areas, herbaceous vegetation areas, aquatic vegetation areas, barren areas, and water areas. In addition to the landcover classes, a mountain mean feature from the UNEP Mountain Watch Report is included which measures the proportion of mountainous terrain within a given cell. (Bontemps et al., 2009)

To supplement the agricultural areas landcover feature, the sum of the harvested area of the predominant crop was introduced into the model. This data was originally sourced from the MIRCA2000 dataset. Finally, two distance features were introduced from the Prio yearly dataset; “capdist”, which provides the spherical distance in kilometers from the cell centroid to the national capital in the respective country, and “bdist3”, which provides the spherical distance (in kilometers) from the cell centroid to the territorial outline of the country the cell belongs too. Both features represent a simple metric for remoteness. (Portmann et al., 2010)

3.1.3 Georeferenced Ethnic Power Relations Politically Relevant Groups, ETH Zurich

The Georeferenced Ethnic Power Relations (GeoEPR) released by ETH Zurich, is based on the Ethnic Power Relations dataset which codes the settlement patterns of politically relevant ethnic groups in independent states from 1946 to 2021 into polygons and is published as a shapefile. (Vogt et al., 2015)

3.1.4 Monthly Food Market Price (World Food Programme)

This dataset was sourced from OCHA's Humanitarian Data Exchange and contains monthly food price data on a per market per month basis. The United Nations World Food Programme (WFP) collects data on specific food commodities including several types of maize, beans, and sorghum, to name a few. Majority of the markets are geocoded, and the datasets are updated weekly. The data collection program started to collect price data in 1992 in a few countries and has expanded over the years. Today around 3000 markets in 98 countries are being monitored.

3.1.5 Armed Conflict Location Event Data (ACLED) Project Conflict Events

ACLED is a real-time disaggregated data collection, analysis, and crisis mapping project. ACLED collects the dates, actors, locations, fatalities, and types of all reported political violence and protest events around the world. ACLED data is collected from a wide array of sources including local, regional, and national sources. From the temporal aspect, this data is recorded daily; one record per day. For example, if a battle lasted for 14 days, 14 records are recorded in the dataset. This dataset is based on the event which takes place on a specific date and at a specific location. The encoding of conflict events follows a hierarchy of events in order to prevent "double counting". For example, if there is an attack on civilians on one day during a battle this will result in only one "battle" record. Furthermore, if two events between the same actors are reported on the same day and location this will be coded as one event in the higher-level event. (Raleigh et al., 2010)

3.2 Preprocessing and Feature Engineering

3.2.1 Prio-Grid 2.0

The Prio-Grid 2.0 consists of two versions: static and yearly. Features were extracted from both. The distance to capital and distance to country border were extracted from the yearly dataset and were joined to the static dataset. Several features were dropped that were deemed redundant or irrelevant and null values were replaced with zeros. Finally, after joining the dataset to the Prio-Grid shapefile, the records were duplicated for the number of years in the study and a year column was added to make a unique identifier by concatenating the grid identifier with the year column. This unique

identifier, “gid_year” was added to each dataset in the study regardless if it was a static or yearly dataset. (Tollefsen et al., 2015)

3.3.2 Climatic Data

Three sets of satellite derived datasets were extracted from Google Earth Engine. The mean was calculated per year for the monthly raster datasets before download. Next the rasters were overlayed on top of the Prio Grid cells and the mean of all the pixels found in each grid cell were calculated using the zonal statistics tool in ArcGIS Pro. Subsequently, the mean cells were joined with the Prio Grid. Finally, a unique “gid_year” identifier was added. (Abatzoglou et al., 2018)

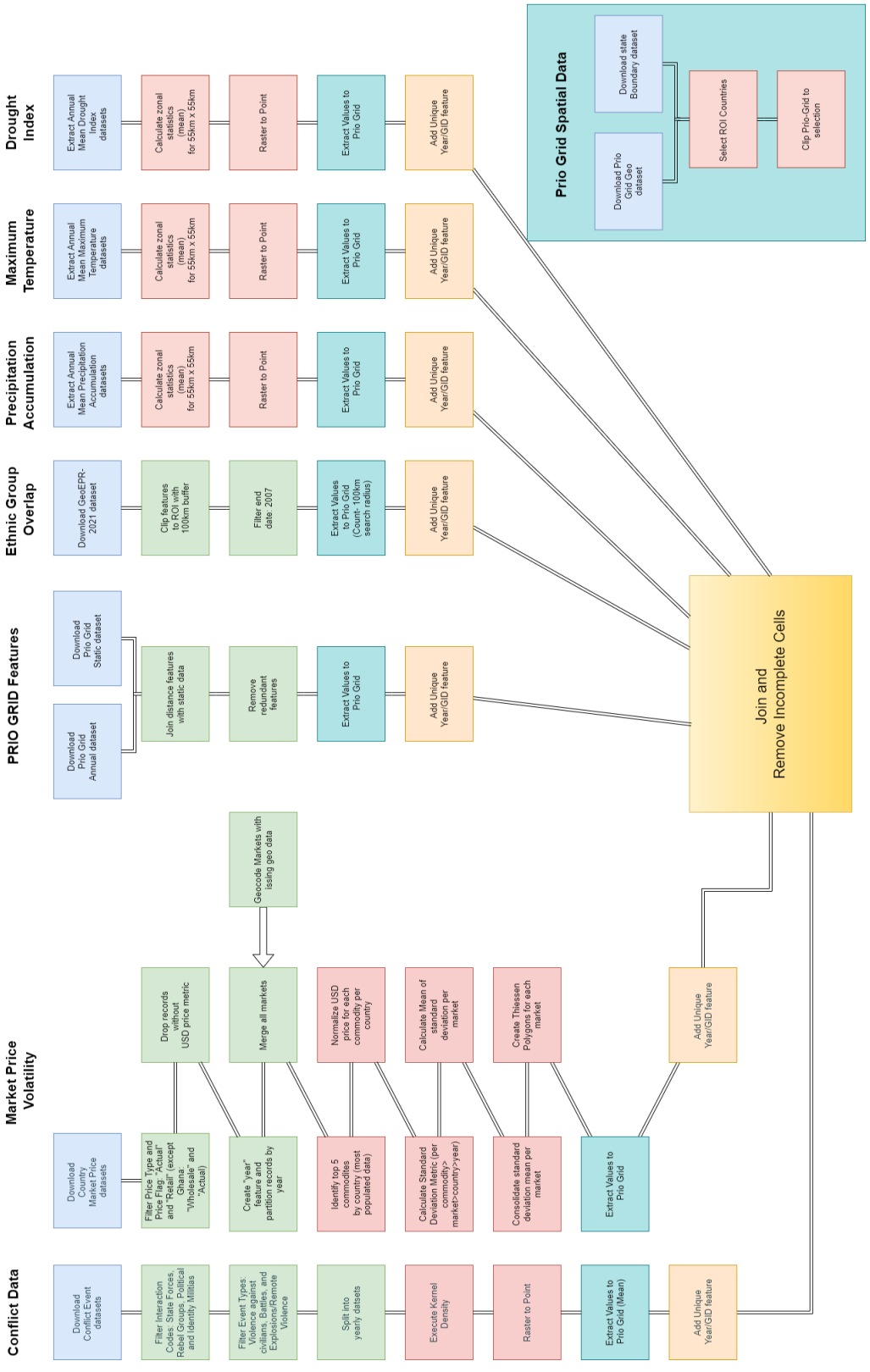

Figure 1: Pre-processing Workflow

3.3.3 Georeferenced Ethnic Power Relations Politically Relevant Groups

After downloading the polygon GeoEPR-2021 dataset, a 100-kilometer buffer of the study area was made to extract the ethnic group polygons in the area of interest and include ethnic groups that were in proximity of the study area. In essence, this was done to mitigate the border effect and ensure that the 100km search radius for the overlay function was implemented for the cells found near the edge of the study area. Next, the dataset was filtered to only include ethnic groups in the study period (2007-2020) before spatially joining it to the Prio Grid with a count function within a 100km radius. Finally, the dataset was duplicated for the number of years, and a unique "gid_year" feature was added for the final merge with all the other datasets. (Vogt et al., 2015)

3.3.4 Food Market Volatility Index

After downloading the data and conducting the initial visual exploratory analysis in ArcGIS Pro, the best proxy metric for market price volatility would be the mean, standard deviation per market. The retail price was used with consistent unit measures for all the countries except Ghana, where the "wholesale" price was used because the data collector did not record retail price in Ghanian markets. The US dollar equivalency price was used to mitigate the multiple currencies in such a vast area of interest. Next, Min-Max scaling was applied to each commodity per country over the entire study period. Only the top 5 commodities per market were included in the metric based on the assumption that the commodities with the most records were the most essential commodities in that country. This was deemed acceptable because the alternative would have been to conduct an in-depth literature review on predominant commodities in each country or region. The cost-benefit ratio proved this method too costly. Once the top 5 commodities were selected, a standard deviation was calculated for each market by country and year. A threshold was set so that only commodities with at least 8 out of 12 monthly (not necessarily consecutive) data records per market were included in the calculation. The mean of the commodity standard deviations of the top 5 commodities was calculated for a market price volatility index. Some markets did not

have data on five commodities (i.e., Djibouti), so the standard deviation was calculated for only the available commodities based on the above criteria. (Caccavale & Flamig, 2017)

Finally, the surface creation and value extraction to the Prio Grid was conducted. Three methods for extrapolating the index point data for the entire study area were considered: inverse distance weighted interpolation (IDW), nearest market, and Thiessen polygons. The IDW interpolation proved to be incapable of standardization for comparison between years and the nearest neighbor method produced similar results to the Thiessen polygon method. The Thiessen polygon method was chosen, and the resulting data were spatially joined to the Prio grid to extract the market volatility metric to each cell. Finally, a unique "gid_year" feature was added for the final merge with all the other datasets.

3.3.5 Conflict Occurrence Index

Based on the literature review, some general assumptions were applied, and subsequently, selections were made based on those assumptions. The dataset included precision variables for time and location, but this was ignored for all intents and purposes, and the inaccuracies were diminished in the feature engineering phase. Additionally, because a density kernel was applied, location precision or lack thereof was mitigated to an acceptable level. The temporal accuracy was deemed insignificant since this analysis is based on an annual temporal resolution. After an extensive literature review, it became apparent that there is no agreed-upon definition of an intrastate conflict event; therefore, several assumptions based on expert sources were made. The "event type" feature was used to filter for events that best represented violent intrastate conflict events. By the ACLED codebook, it was decided that the study would only include events categorized as violent events, which were coded as "Battles", "Explosions/Remote violence", and "Violence against civilians". Next, the "actor" codes and interaction codes were used to select violent events that include "state forces", "rebel groups", and two types of "militias" (political and identity). It's worth mentioning that "state forces" vs. "state forces" events were also filtered out. This process effectively, omitted relatively ambiguous categories such as "rioters",

"protesters", "civilians", and "other external forces", attempting to capture events that only represent the nature of violent intrastate conflict. (ACLED, 2019)

Once the data was cleaned, filtered, and parsed by year, consistent density kernels were constructed in ArcGIS Pro with special attention to standardization to ensure consistent and replicable results for comparison over the analysis time frame. The density kernel method was selected to diminish the potential effect of false positives and questionable accuracy of reporting events inherent in the dataset and unsatisfactory precision of location data. Next the kernel density surface values were converted to points. These points were subsequently spatially joined to the Prio Grid vector, applying the mean function resulting in a single continuous float value for each cell. This workflow was repeated for each year. Determining the optimal search radius proved to be an iterative process. The first iteration used the default search radius which is based on the Silverman's rule of thumb. (Silverman, 1986) This produced diverse search radiuses which were then averaged to calculate a standardized radius of 275930 meters. In the final density kernel surface iteration, the mean radius was applied to all the kernel density workflows. The cell size was determined by a trial-and-error method balancing computational practicality and acceptable resolution to not lose information. Due to the limited capacity of the hardware available for this analysis, 3000 meters, which satisfied the need for detail while also not being too costly on processing time and capacity, was used.

3.4 Exploratory Data Analysis

3.4.1 Feature Summary Statistics

This section presents an assortment of experimental techniques that were used better to understand the features' main characteristics in this study. After preprocessing the selected elements, summary statistics were calculated to understand the data distribution better. In table 2, one can see that the dependent variable, conflict occurrence index, has a low standard deviation with many low values. This is expected as the study area is relatively large compared to where conflict events were recorded. Furthermore, the conflict events were aggregated into density kernels; therefore many continuous low values were produced. Although the range of the market food price

volatility index is extensive in comparison to the conflict occurrence index, the standard deviation remains low, signifying relatively low variability in this feature. The ethnic group overlap feature's standard deviation is relatively high, illustrating high variability in the feature with 75% of cells only having seven or fewer ethnic groups present within a 100km radius and the remaining 25% of cells has an extensive range between 7 and 73 ethnic groups present within 100km of the cell. The variability of the climatic variables is to be expected as the Sahel region has a very diverse climate across the region, ranging from desert, high temperature, low to no precipitation to dense moist forest with high rainfall. One can observe that the "distance" features are spatially autocorrelated and therefore more equally distributed. From the landcover features the table shows that urban areas, areas covered by bodies of water, and mountainous areas are rare in the region.

	Mountain Mean	Agricultural Cover (%)	Aquatic Vegetation Cover (%)	Barren Cover (%)	Forest Cover (%)	Herbaceous Cover (%)	Urban Cover (%)	Water Cover (%)	Distance to border (km)	Distance to capital (km)
count	53004	53004	53004	53004	53004	53004	53004	53004	53004	53004
mean	0.100	19.990	3.920	37.226	19.933	9.414	0.047	0.304	139.386	580.023
std	0.233	29.699	10.362	44.518	28.271	20.290	0.392	2.022	109.579	343.662
min	0	0	0	0	0	0	0	0	0	5.18
25%	0	0	0	0	0	0	0	0	52.82	312.53
50%	0	1.43	0.02	2.07	4.99	0	0	0	112.49	507.93
75%	0.025	31.92	2.22	98.27	30.66	4.81	0	0	199.88	808.28
max	1	99.81	94.49	100	100	97.72	14.86	46.35	581.03	1888.88

	Conflict Index	Ethnic Group (#)	Food Price Volatility Index	Harvest Area (ha)	Drought Index	Precipitation Accumulation (mm)	Maximum Temperature (C)
count	53004	53004	53004	53004	53004	53004	53004
mean	0.000314	6.70	0.358412	8322.77	-21.36	45.81	34.6
std	0.000835	9.708	1.527689	17034.536	405.257	42.737	2.756
min	0	0	0	0	-1629	0	19.7
25%	1.50E-06	2	0.01152	0	-239.25	9	33.4
50%	3.44E-05	3	0.030469	661.98	-35	34	35.2
75%	0.000242	7	0.057727	9064.07	156	75	36.5
max	0.01593	73	21.40181	147751.3	2401	400	40.1

Table 2: Feature Summary Statistics

3.4.2 Feature Data Distribution

To better understand the distribution of the features before introducing them into the machine learning algorithm a series of histograms were produced.

Conflict Occurrence Index, Market Price Volatility Index, Ethnic Group Overlap Feature

In figure 2 it can be observed that the conflict occurrence index or dependent variable is positively skewed with numerous "0" values. The market price volatility histogram shows that the data is also positively skewed but has several clusters of markets with relatively high price volatility. The ethnic group overlap feature is bit more normally distributed however it is still positively skewed with a relatively wide range of values.

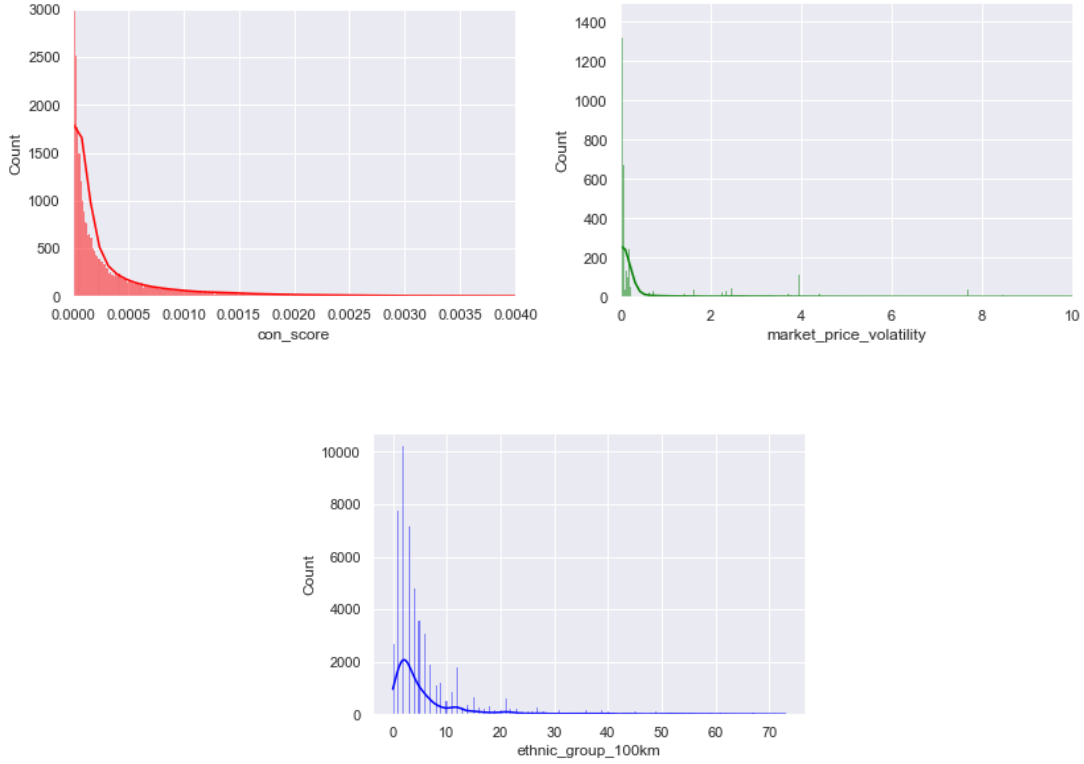

Figure 2: Top Left: Conflict occurrence index. Top Right: Market price volatility index, Bottom: Ethnic group overlap count

Climatic Features

Compared to the engineered features above, the climatic features are much more equally distributed, which is to be expected as mentioned earlier. The maximum temperature feature though is negatively skewed. The high frequency of "0" values in the precipitation accumulation feature is expected as vast arid expanses in the study area. (see figure 3)

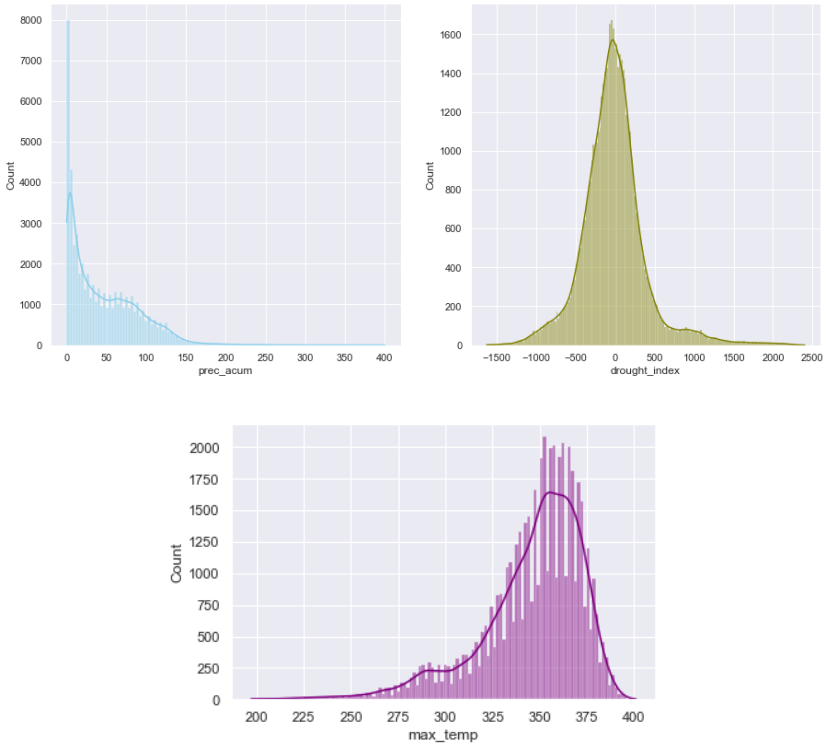

Figure 3: Top left: Precipitation accumulation. Top right: Drought index, Bottom: Maximum temperature

Landcover Features, Harvest Area, and Mountain Area Features

The histograms in figure 4 further confirm that urban, mountainous areas, and water bodies are rare in the study area. Barren groundcover feature has many records with 100% coverage and about half as many with 0% coverage. The forest ground cover and harvest area feature histograms illustrate that the study area dips into areas that are conducive to rainfed agriculture and areas with significant forest coverage resulting in many values between 0 and 40% forest coverage and 0 to 40000 hectares of harvest area.

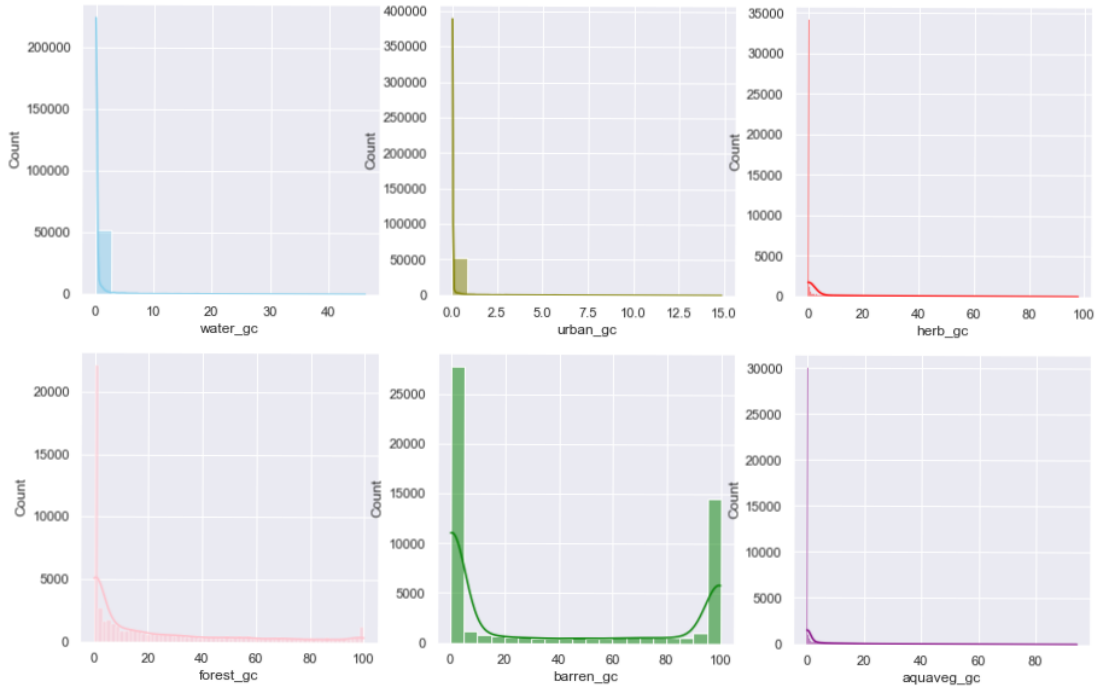

Figure 4: Top row from left to right: Water ground cover, Urban ground cover, Herbaceous ground cover. Bottom row from left to right: Forest ground cover, Barren land cover, Aquatic vegetation ground cover

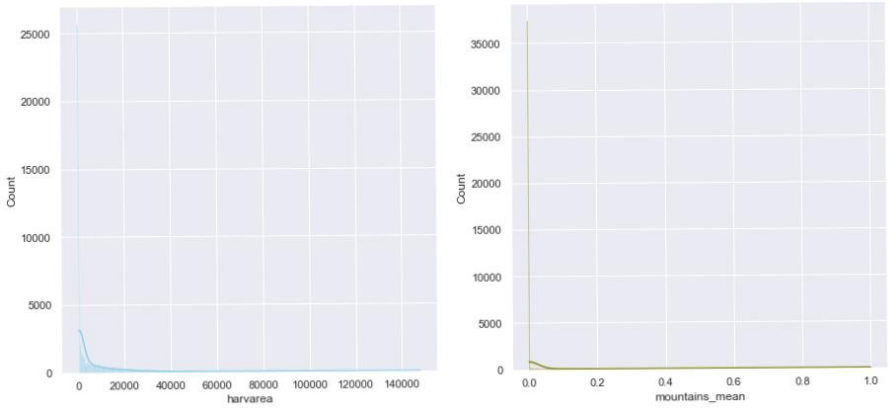

Figure 5: Left: Harvest area Right: Mountain mean

Distance (Remoteness) Features

Although positively skewed, the distance feature histograms are relatively more generally distributed than the landcover features. The data is most likely positively

skewed because there are many relatively small countries in the study area where the cells are comparably close to the territorial border and the capital city.

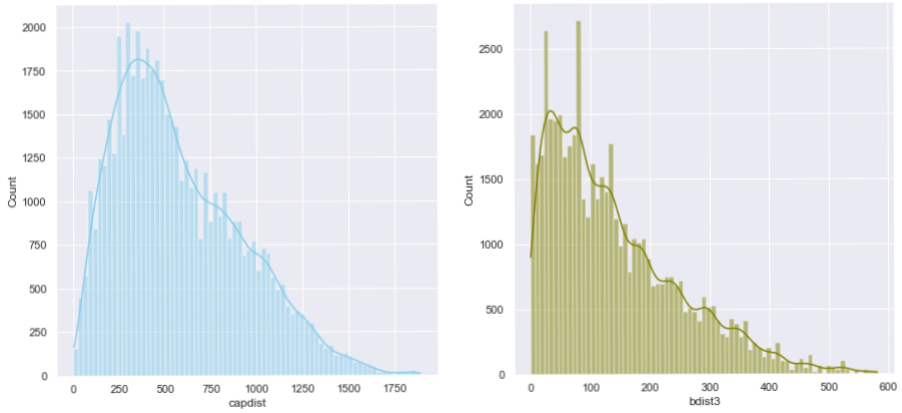

Figure 6: Right: Distance to capital. Left: Distance to border

3.4.3 Feature Correlation

To lower the risk of introducing redundant features into the model while getting a better understanding of the relationship between features, a correlation matrix was created, which can be observed in Figure 7. After an initial assessment, no redundant variables were identified; however, a few valuable relationships were observed.

First, there is a relatively strong negative correlation between precipitation accumulation and barren landcover and a high correlation between the former and maximum temperature. These relationships can best be described by desert or arid areas.

Next, a positive correlation between capital distance and barren landcover is apparent, which can be best explained assuming that most capitals in this region are located south of the Sahara or near river basins where semi-arid conditions are present.

Furthermore, it can be assumed that large cities and their suburbs have vegetation or urban coverage. Also related to the capital distance feature, precipitation accumulation is negatively correlated to this feature representing the distance to the capital. Although

not significant, from all the features market price volatility index has the highest correlation with the conflict occurrence index.

Finally, two interesting observations related to the ethnic group overlap feature can be made. First, from all the other features, harvest area in hectares has the highest correlation with ethnic groups present within 100km. This can be somewhat expected as high population areas will have the most agricultural area and most likely be where the most ethnic diversity can be found. Secondly, the ethnic group overlap negatively correlates with the food market volatility index. Perhaps an area with many ethnic groups within 100km is more developed or has larger markets with better connectivity.

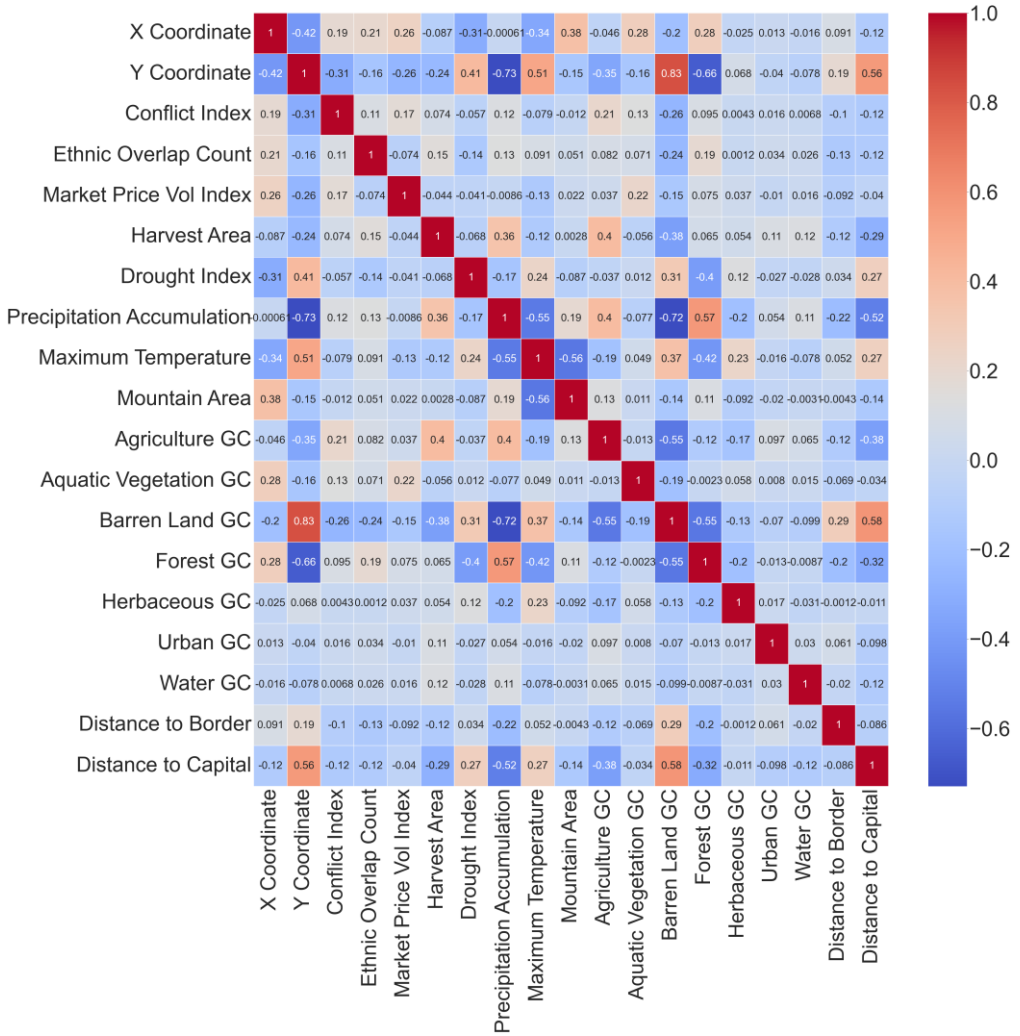

Figure 7: Correlation matrix of all features

3.5 Random Forest Regression Model

The random forest regression predictor was chosen for this analysis. It is a robust regression tool capable of performance gains between permutations, that can handle a large number of variables well, and is relatively cost-effective.

An ensemble learner, the random forest works well "out-of-the-box" or without hyperparameter tuning. It can calculate feature importance which is necessary to identify the most influential predictors of conflict events. Overfitting is generally avoided because of this algorithm's ability to randomize feature subset selection. By using a unique subset of the data for every permutation, the algorithm ensures that the decision trees are less correlated while identifying general patterns within the training data. Moreover, it splits each node in the decision trees using a random set of features. This, in effect means that no single tree sees all the data during training and testing. This is advantageous when working with noisy data such as this analysis. (Muchlinksi et al., 2016) The random forest regression model in this study was built and implemented using the Scikit Learn toolbox and was executed in Jupyter Notebooks.

The K-fold validation method with the shuffle function was utilized to ensure that each permutation consisted of a random sample of records for each permutation. The random shuffle function allowed the model to test the predictions on lots of different data segments to mitigate the fact that training and testing were done with the same dataset. After several tests, the model was implemented using three runs and 100 trees, then with 1000 trees with five folds and then with ten folds. Table 3 illustrates the parameters of the three variations of the model. Following the conventional methodology, the study paid particular attention to the ratio of training to testing data in the model. This methodology achieved enough data to train the model properly, thus providing accurate results. Each permutation used one randomly selected subset or fold of the data to test the predictions. In effect, the 5-fold model assigned 20% of the records to the testing data, while the 10-fold model assigned 10% of the records to test

the predictions. The primary hyperparameters were left at default, including, the maximum depth of a tree was set at "none," the minimum number of samples required to split an internal node was set at "2", and the function to measure the quality of the split or "criterion" was set to "squared error." These were unchanged moving forward as it was decided that the results produced with default parameters were acceptable and that hyperparameter tuning would not have significant gains in accuracy rather data manipulation and incorporating other data sources could achieve gains. (Scikit-learn n.d, para. 1)

Following convention, a baseline model for prediction was calculated to provide a metric for comparison. The goal of the baseline model was to strike a balance between simplicity and practicality and find a meaningful metric to compare the random forest regression model results with. The baseline model was calculated using the mean of training targets per fold.

	Model 1	Model 2	Model 3
Trees	100	1000	1000
Folds	5	5	10
Runs	3	3	3
Time (min)	**10.04**	**161.67**	**250.95**

Table 3: Model specifications

[1]

4. RESULTS AND DISCUSSION

4.1 Results

To evaluate this model, three metrics were calculated. First, R-squared or the coefficient was applied to measure the proportion of variance in the independent features about the conflict occurrence index or "target." This was a good model level predictor of accuracy and provided a reliable metric to assess to what extent the data fit

[1] Processor specifications: Intel(R) Core(TM) i7-8650U CPU @ 1.90GHz 2.11 GHz

the model. Additionally, the Root Mean Square Error (RMSE) was applied to both the baseline model and the random forest regression prediction results to measure the residuals for each target and predicted value. In general, the random forest model produced strong results in comparison to the baseline model. As mentioned above, three versions of the model were implemented, and their statistical and visual outputs are further evaluated in the next section. The results of the third variation proved to be satisfactory not only in comparison to the baseline but it also improved on the results of the preceding 1000 tree 5-fold model.

4.2 Evaluation of Model

To thoroughly assess the model, the results were analyzed on a per fold basis and the distribution of the results were visualized utilizing the Seaborn toolkit. It became apparent that although the results were deemed satisfactory there were significant outliers in the data and the results needed to be investigated further. In table 4 below one can observe the results for all three variations.

	Model 1	Model 2	Model 3
R Squared Mean	0.922	0.924	0.948
RMSE Baseline Mean	7.080	7.080	7.034
RMSE Predictions Mean	1.740	1.716	1.394
RMSE Difference Mean	70.514	70.875	76.439
Absolute Mean Error/Bias	-0.135	-0.137	-0.112

Table 4: Model Evaluation

In table 5, one can see that the calculated average root mean squared error (RMSE) was 1.394. The RMSE is the average of all three runs' RMSE averages. Furthermore, the standard deviation of the RMSE outputs of all 30 permutations was .33 and the average R2 of was .95. This is the average of all 3 runs' R2 averages. The standard deviation of R2 averages is .05.

Comparing with the baseline model the model has an average of 76% relative improvement with an average standard deviation of 12.67% between the 30 permutations. This metric communicates the diversity of data ranges and data variation of target values in each fold

Run #	Fold #	R Squared	RMSE Predictions	RMSE Baseline	Relative improvement in RMSE from Baseline to Predictions (%)
1	1	0.940	1.309	4.677	72.01
	2	0.867	1.862	4.814	61.32
	3	0.779	2.237	3.544	36.89
	4	0.915	1.213	3.613	66.43
	5	0.975	0.958	5.776	83.42
	6	0.990	1.157	9.333	87.6
	7	0.982	1.678	10.127	83.43
	8	0.986	1.189	8.850	86.56
	9	0.985	1.279	9.639	86.74
	10	0.970	1.837	9.954	81.54
Mean		0.939	1.472	7.033	74.6
Std Dev		0.069	0.407	2.777	16.11
2	1	0.943	1.231	4.674	73.65
	2	0.914	1.452	4.810	69.81
	3	0.866	1.576	3.565	55.8
	4	0.914	1.263	3.607	64.97
	5	0.974	0.962	5.778	83.35
	6	0.990	1.068	9.335	88.56
	7	0.985	1.556	10.127	84.63
	8	0.989	1.098	8.851	87.6
	9	0.984	1.348	9.641	86.02
	10	0.967	1.894	9.952	80.97
Mean		0.953	1.345	7.034	77.54
Std Dev		0.042	0.282	2.776	11.03
3	1	0.949	1.121	4.676	76.04
	2	0.916	1.431	4.809	70.23
	3	0.865	1.647	3.562	53.75
	4	0.901	1.307	3.622	63.92
	5	0.968	1.097	5.774	81
	6	0.992	1.009	9.334	89.19
	7	0.988	1.409	10.125	86.09
	8	0.987	1.191	8.850	86.54
	9	0.980	1.447	9.639	84.99
	10	0.965	1.985	9.955	80.06
Mean		0.951	1.364	7.035	77.18
Std Dev		0.043	0.292	2.774	11.4
Mean of all runs		**0.948**	**1.394**	**7.034**	**76.44**
Std Dev of all records		**0.051**	**0.325**	**2.678**	**12.67**

Table 5: Model Evaluation Per Run

As mentioned earlier conflict events are very rare events and even with the kernel density approach there are significant amounts of records with "0" values for the index in the study area. To properly assess the accuracy of the predictions in comparison to the target conflict occurrence index values below 50 were removed and the relatively "high conflict zones" were isolated for visual inspection. This was repeated for values under 100 to compare targets and predictions in "very high conflict" areas. It is evident

that rather large errors in forecasts were in the high “very high conflict” areas. Figures 8, 9, and 10 illustrate the relationship of predictions to targets in each subset described above.

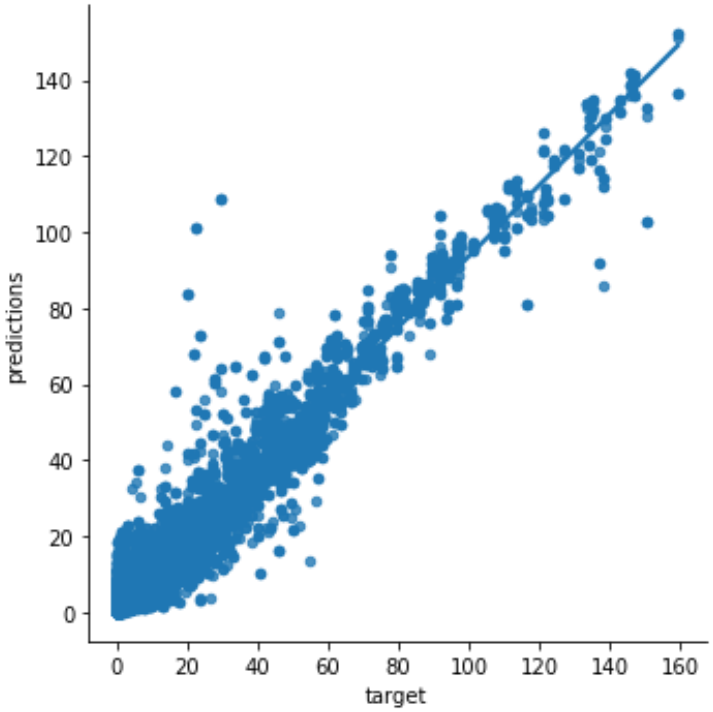

Figure 8: Comparison of predictions to targets (All)

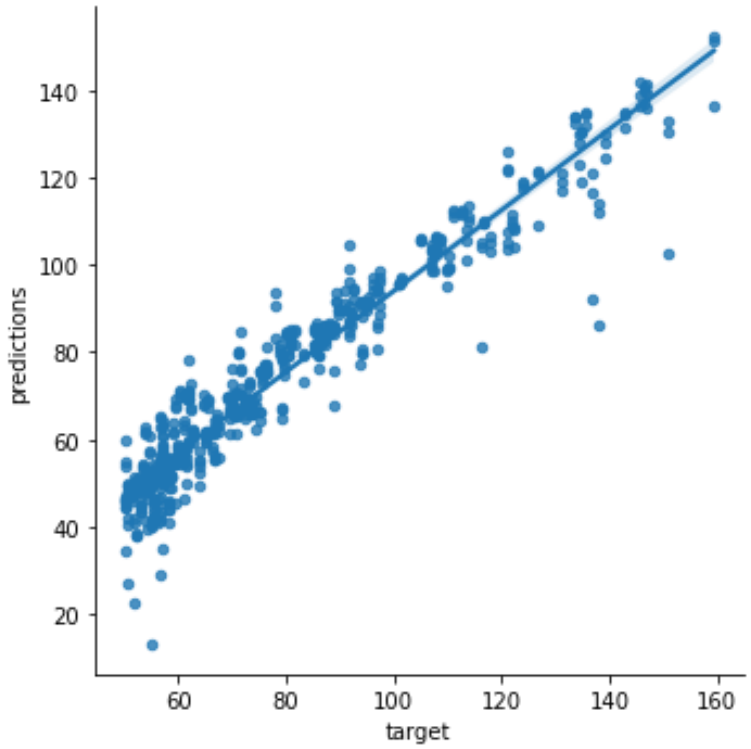

Figure 9: Comparison of predictions and targets in high conflict zones (>.005)

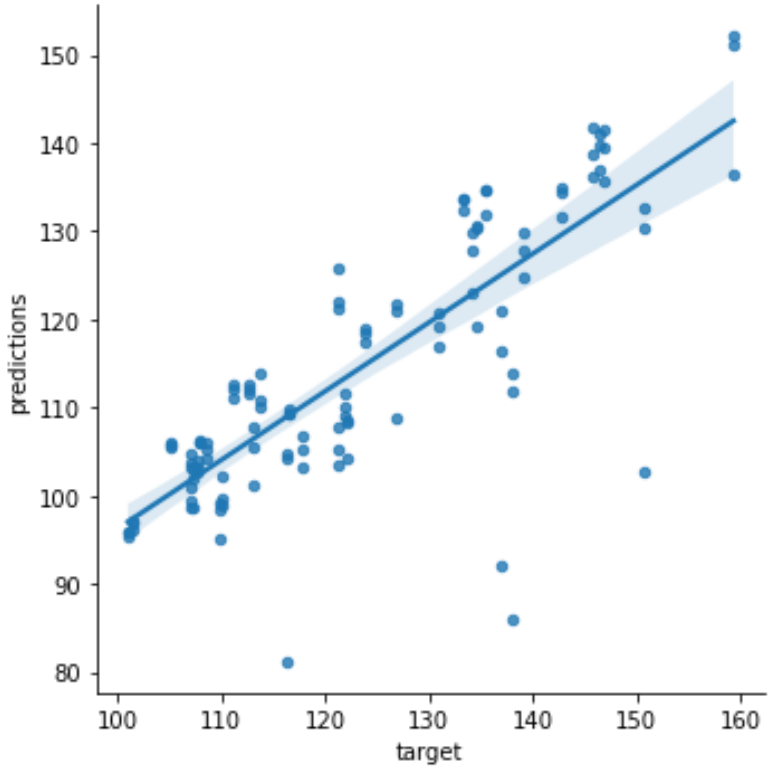

Figure 10: Comparison of predictions and targets in very high conflict zones (>.01)

Figures 11, 12, and 13 illustrate the distribution of predictions and targets for comparison. The first bar graph shows all the targets and predictions, the second, all targets above 50 with their respective predictions and the last bar graph shows the last third of targets and their accompanying predictions. In the second graph one can observe that there is a dramatic over-prediction around the 50-target value and drastic under prediction in the 60-value area. In the last graph it is evident that between the 90 and 110 value count the inverse phenomenon takes place.

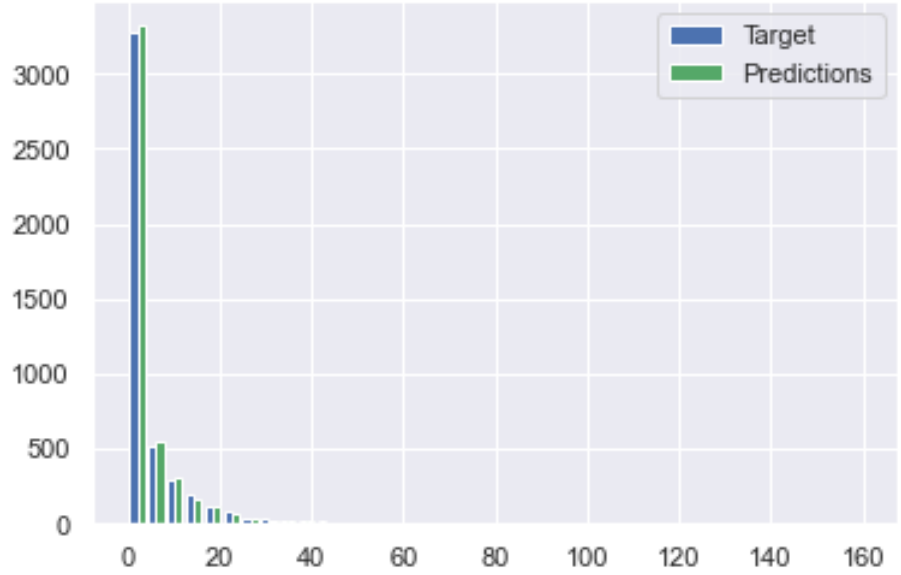

Figure 11: Comparison of predictions to targets distribution (all)

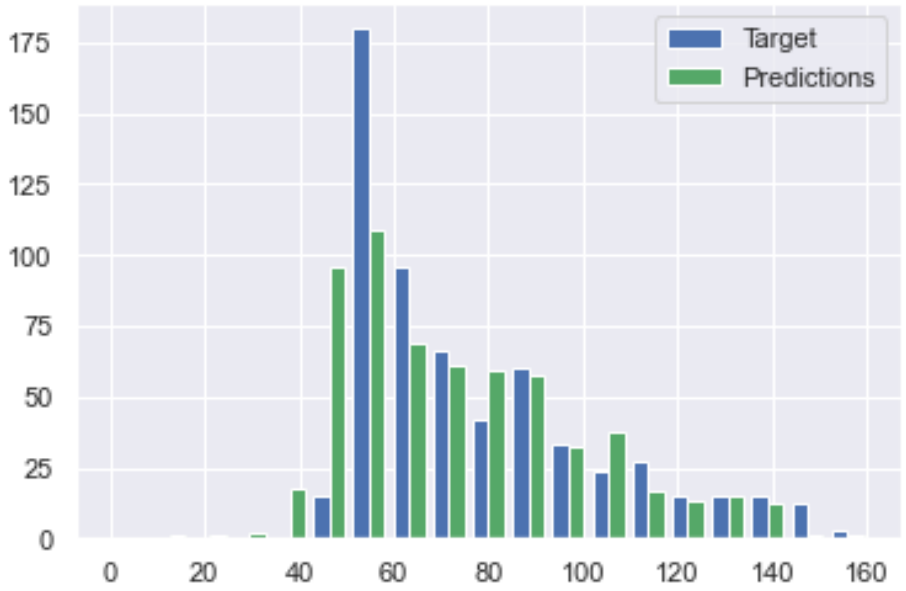

Figure 12: Comparison of predictions to targets distribution in high conflict areas (>.005)

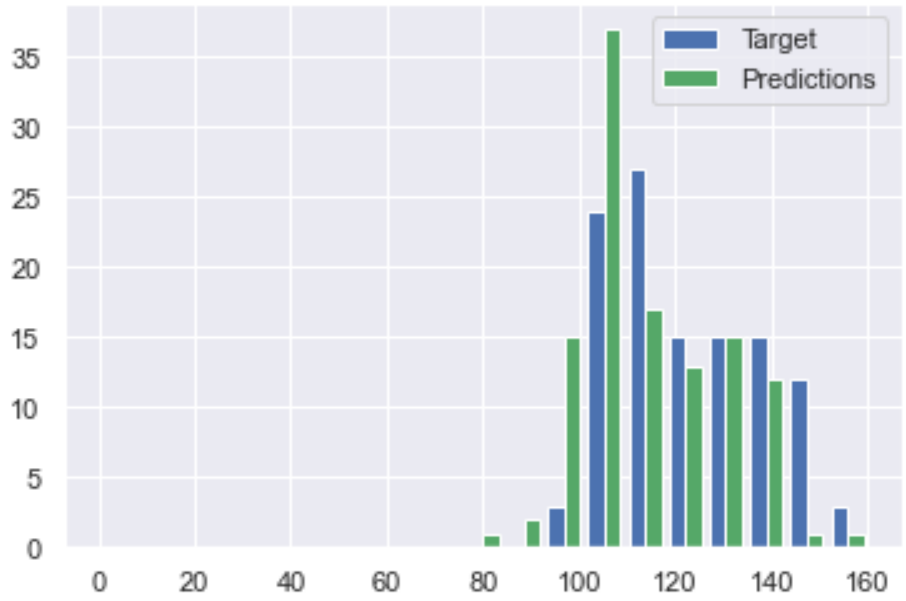

Figure 13: Comparison of predictions to targets distribution in very high conflict areas (>.01)

Target and Prediction Spatial Distribution

The plotted results in the study area in figures 14, 15, and 16 further illustrate that the model performed well in general and that the most significant errors were present in the

surrounding areas of Mogadishu, Somalia, the Lake Chad Region, and the tri-border region of Mali, Niger, and Burkina Faso. These three areas are characterized as very high conflict zones. On the other hand, the model performed relatively well in different high conflict zones such as Darfur in Sudan, South Sudan, and Nigeria.

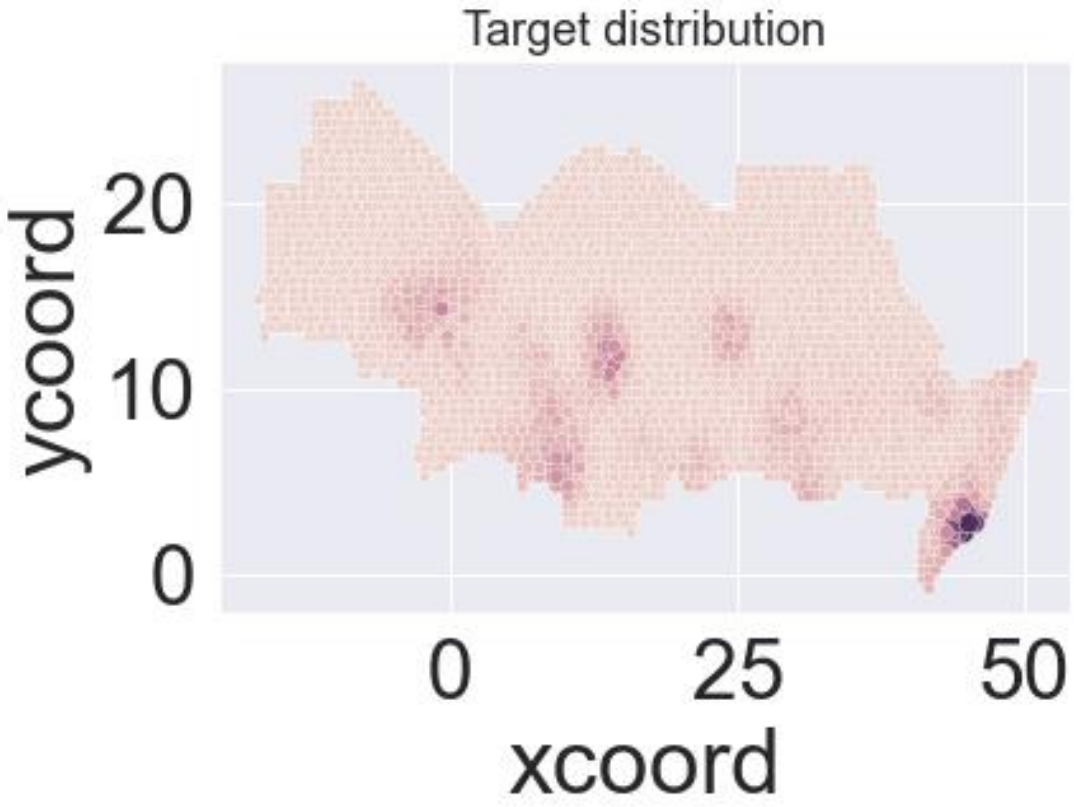

Figure 14: Target spatial distribution

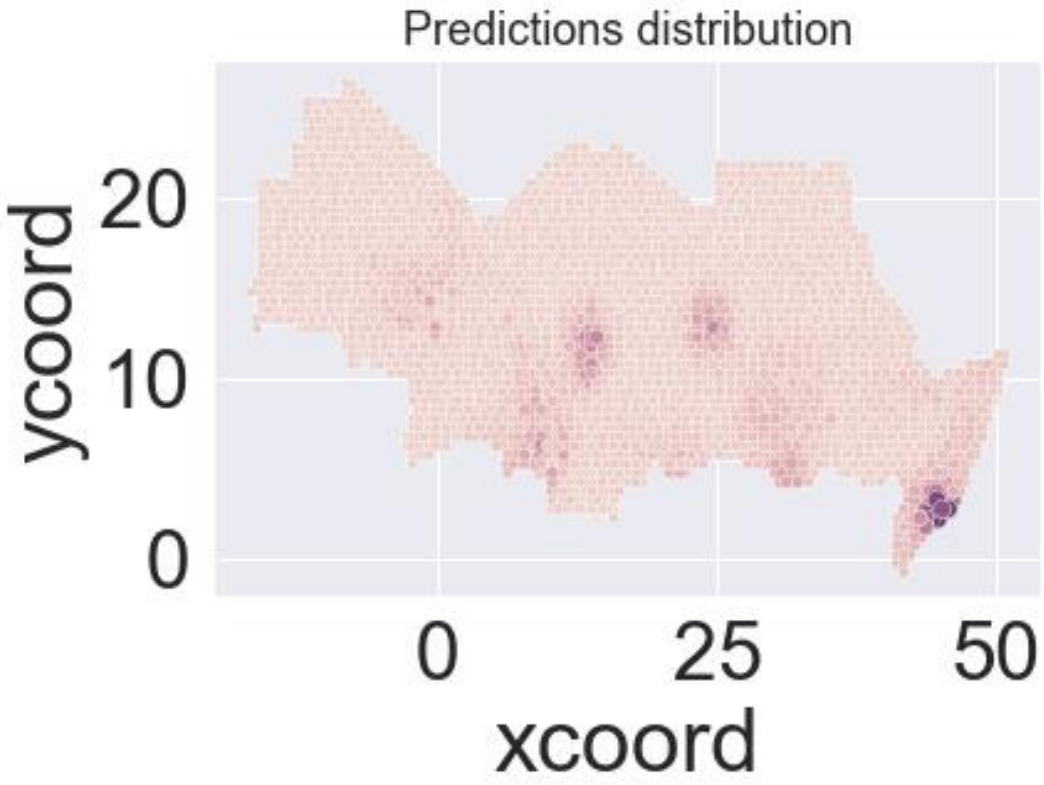

Figure 15: Predictions spatial distribution

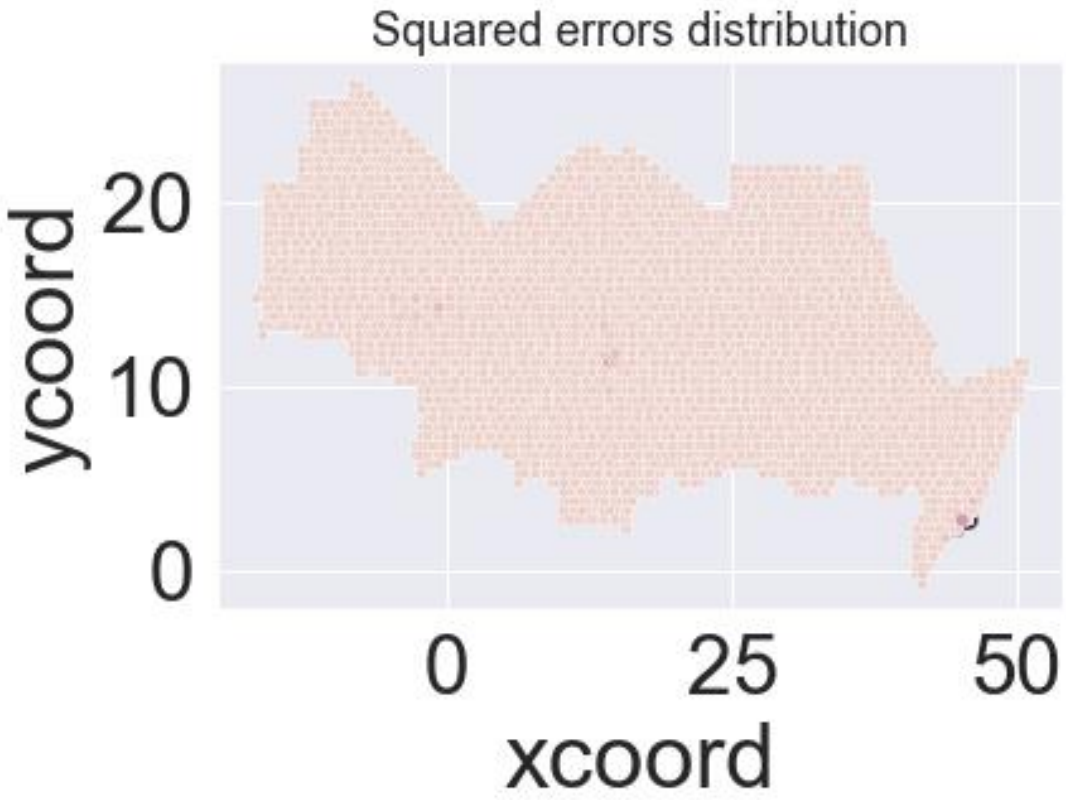

Figure 16: Squared errors spatial distribution

4.3 Evaluation of Predictors

As mentioned above, a vital advantage of the random forest regression algorithm is that influence of the independent features can be assessed and ranked from high to low importance. The evaluation of predictors supported the literature that food security (food price volatility index), ethnic polarity, and climatic features ranked high. Interestingly, the third most important feature in the model was the distance to capital which is expected as conflict events tend to take place around high population centers, and in most if not all the countries in the study, the capital is the most densely populated area.

Feature	Model 1	Model 2	Model 3
Market Food Price Volatility Index	25%	25%	25%
Ethnic Group Overlap	22%	22%	22%
Distance to Capital	15%	16%	15%
Drought Index	8%	8%	8%
Precipitation Accumulation	6%	6%	6%
Maximum Temperature	5%	5%	5%

Table 6: Predictor Importance

4.4 Discussion and Limitations

Conflict prediction using machine learning algorithms is not a new area of interest or experimentation. It has been implemented in the past utilizing a vast array of data features and algorithms. However, to date, there has not been an analysis conducted at this scale or on a unified spatial data structure for the Sahel region. Prior research mostly focused on regional or administrative zone spatial distribution and not on uniform quadratic grid cells. This is a necessary improvement as the conflict in this region is seldomly bound by national or regional boundaries. Furthermore, the kernel density and surface extraction for the food volatility index and the conflict occurrence index provided an improved metric to extrapolate food security and conflict vulnerability more accurately throughout the study area.

The random forest regressor performed well against the baseline prediction, but it is essential to mention that this algorithm is not without its shortcomings. It is well known that the random forest regressor algorithm only predicts values within the range of the data, therefore, limiting the algorithm's ability to predict values without bias. As the conflict occurrence index data is grossly unbalanced with numerous "0" values the model overpredicted many times influencing the final model's prediction metrics.

An important facet of any machine learning algorithm is the sampling method and the size and distribution of samples used for training and testing. Utilizing the randomly shuffled k-fold validation method to test this model with unseen data was an acceptable approach and supplied the model with virtually unseen data for testing. This model could benefit from a more targeted shuffle method, for instance, a country-based fold

split. By splitting the data by country, it would ensure the model does not learn patterns specific to a given country, therefore, providing the model with a more robust testing subset.

The feature engineering implemented in this study intended to quantify and model a diverse array of real-world phenomena, most notably market price volatility and occurrence of conflict events. The engineering of these features addressed several challenges including data availability, data continuity, and consistency. This was especially challenging in the development of the market price volatility index. In several countries in the study area data was incomplete and only price data was collected for a limited number of commodities. To overcome this bias the analysis implemented a strategy that chose the top 5 commodities in each country with the most records. The rationale or assumption behind this was that the commodities with the most records are the most essential staples in each country. This shouldn't pose a future challenge as WFP has increased its scope and frequency of data collection throughout the region. In some countries, there was an interruption in data collection. It is difficult to tell why this interruption occurred, but perhaps it could be partially attributed to conflict occurrence as markets could be potentially closed or inaccessible when conflict is present.

In addition to data availability, several issues regarding data quality needed to be considered and mitigated. The ACLED dataset has been critiqued in the past for its inaccuracies. (Eck, 2012) One of the challenges of this model was to diminish the effect of the inaccuracies so they wouldn't propagate throughout the model. This was done through the creation of standardized kernel density surfaces. Another critique could be the maximum temperature and accumulated precipitation features. However, they proved to be helpful in prediction as they have been engineered to reflect the extreme climatic events. Perhaps, a moving window of standard deviation could have been applied to introduce climate volatility into the model. The engineered ethnic group overlap feature, which introduces demographics into the model and serves as a proxy for diversity, could be further engineered or replaced with another dataset to strengthen

the model in line with the literature becoming a rough estimate of ethnic group prevalence in a specific search radius.

In line with Collier & Hoeffler's theory, the lone economic feature, representing food security, was the most critical indicator in the study. In that sense, prediction is fundamentally different from explanation or causality even though a particular model fits the data well one should proceed with caution to try to conclude the cause from predictive modeling.

5. CONCLUSION

The focus of this study was to identify several predictors, realistically model them, and construct a helpful prediction model for conflict occurrence utilizing random forest regression. This study builds on the foundation of prior studies in the field, borrowing both methodology and data sources and applying them differently and spatial and temporal resolutions. The features were engineered, and the values were extrapolated across the study area to represent real-world socio-economic, demographic, and climatic phenomena. The model constructed fit the data well with over 94% accuracy. Moreover, the RMSE mean the difference between the baseline and the random forest regression was significant at 76.43%. The Random Forest algorithm consistently evaluated market price volatility, ethnic group overlap, and distance to capital as the most important predictors in the model.

This simple heuristic model using open-source data illustrates that, in principle, one can build a model for any area of interest limited only by data availability. It is essential to mention that utility or usefulness doesn't necessarily translate into complexity, and this model and its results support that notion. Special attention was given to ensure that the model could be reproduced with open-source data and tools. Although ArcGIS Pro, commercial software, was used for some spatial modeling, this same workflow can be implemented in python, R, or any open-source GIS software.

Machine learning is iterative, and data can be added, refined, and engineered to improve models. This may result in faster and more robust learners resulting in more accurate

predictions in local contexts. This experimental study into conflict prediction is intended to add to the "conversation" on the matter and introduce a different spatial resolution of investigating a very complex phenomenon.

Although drivers of war have been largely static throughout human history, the nature of modern-day conflicts is changing. New quantitative spatial models to assist in understanding intrastate conflicts should continue to be tested and improved. Through ever more accurate predictions, governments and international stakeholders can be better equipped with actionable information to address the underlying issues that this study has shed light on.

5.1 Future Works

The application of predictive analytics in social sciences is passing its infancy, and there are still many questions to answer. Without a doubt, future quantitative studies in this subject area would benefit from applying and comparing different machine learning algorithms. A diverse array of ensemble algorithms should be applied as there is no gold standard predictor that performs best across different dataset configurations. According to Meire's overview of machine learning algorithms used in conflict prediction, extreme gradient boosting should be explored in this context. From the spatial aspect, a sensitivity analysis could be applied in future work so that data resolution can be better assessed and compared to the granularity of the analysis.

Furthermore, more open-source data is available and including traditional and innovative data-derived indices would only strengthen models. Perhaps, including data reflecting public fear and the perceived threat or incorporating foreign assistance or aid to assess the impact of these controversial methods for prophylactic interventions by international stakeholders and nation-states. According to the literature, including extreme weather data could also profoundly affect the model as it would represent anomalous weather phenomena such as flooding, drought, or heat waves. Furthermore, political data or more sophisticated ethnic group demographic data could also strengthen the model's predictive accuracy.

In addition to sourcing new data, further engineering of the features included in this study could make the model more realistic and perhaps provide even better predictions. For example, developing the ethnic group feature to represent accurate world demographics or utilizing the fatality count feature in the ACLED dataset to describe intensity or magnitude could add nuance to the model. As for the market food price volatility index could provide a more realistic food metric for the entire study area by creating an interpolated surface by applying the inverse weighted distance method with standardized parameters. Going a bit further and conducting a network analysis to represent a real-world network of markets and distribution routes on a local level could be explored in the future. Parsing and categorizing the relatively rich ACLED conflict events dataset by classifying actors and conflict event types more appropriately based on an in-depth literature review and perhaps model a specific type of conflict or compare the different configuration of actors and event types in other models should be explored. Additionally, conducting a more holistic analysis by exploiting the temporal autocorrelation of intrastate conflict events, one could include event types such as "riots," "protests," or even "strategic events" into the analysis could give a much more accurate depiction of intrastate conflict development.

As the results illustrate, the model performed differently in high conflict and very high conflict zones. As the study did not explicitly discriminate between conflict zones, future works could focus on conflict "hotspots" by isolating areas with a high or very high conflict occurrence index and predicting conflict occurrence in those areas only. Related to this, a population layer and a masking technique could be applied to mask out desert areas and other low population areas. This would need to be done with caution and attention to detail. It would risk omitting areas of conflict in sparsely populated areas, which could be a characteristic of some conflict zones in the Sahel region.

As mentioned earlier, sampling training and testing data based on countries in the study area to make the model more reliable is also an approach worth investigating. This method would enhance the evaluation of the model's reliability by testing the model

with unseen data and decreasing the probability of overfitting data to country-specific characteristics.

From the temporal aspect, this study was conducted using annual intervals, which diminishes the practical utility of the model. The temporal nature of conflict events is very complex, but a quarterly or monthly model could benefit stakeholders with a higher degree of actionable information. Additionally, the concept of "lag" can be introduced into the model by predicting conflict occurrence based on independent features representing the past month or week.

Finally, utilizing parallel processing techniques that are readily available today, future research could play a vital role in creating accurate or near-real-time conflict predictions. To further enhance the actionability of the forecast, one could go a step further to attempt to automate a near real-time model that could predict vulnerability to conflict for the upcoming month or week, for example. The ethics of such a prediction should not be ignored as early warning systems for conflict could be problematic if readily available to the public as they could influence perception and drive fear which in turn can play a role in actual conflict occurrence.

REFERENCES

Abatzoglou, J.T., S.Z. Dobrowski, S.A., Parks, K.C., Hegewisch. (2018). Terraclimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958-2015, Scientific Data 5:170191, doi:10.1038/sdata.2017.191

ACLED. (2019). Armed Conflict Location & Event Data Project (ACLED) Codebook, 2019.

Backer, D. A., Wilkenfeld, J., & Huth, P. K. (2014). Peace and conflict 2014. Paradigm.

Bontemps, Sophie, Defourny, Pierre and Bogaert, Eric Van. (2009). Globcover 2009. Products Description and Validation Report. European Space Agency. http://due.esrin.esa.int/files/GLOBCOVER2009_Validation_Report_2.2.pdf

Buhaug, Halvard, Cederman, Lars-Erik, and Gleditsch, Kristian Skrede. (2014). Square pegs in round holes: Inequalities, grievances, and civil war." International Studies Quarterly 58, no. 2

Caccavale, Oscar M., Flamig, Tobias (2017). Collecting prices for food security programming: The how and why of price data collection at WFP. World Food Programme

Clauset, A. (2018). Trends and fluctuations in the severity of interstate wars. Science Advances, 4(2), eaao3580. https://doi.org/10.1126/sciadv.aao3580

Collier, Paul and Hoeffler, Anke (2004). Greed and Grievance in Civil War. Oxford Economic Papers 56, no. 4

Eck K. In data we trust? A comparison of UCDP GED and ACLED conflict events datasets. (2012). Cooperation and Conflict.;47(1):124-141. doi:10.1177/0010836711434463

Ettensperger, F. (2020). Comparing supervised learning algorithms and artificial neural networks for conflict prediction: performance and applicability of deep learning in the field. Qual Quant 54, 567–601 https://doi.org/10.1007/s11135-019-00882-w

Fearon, James D., and David D. Laitin. (2003). Ethnicity, Insurgency, and Civil War. American Political Science Review 97, no. 1 (2003): 75-90.

George, J., Adelaja, A., & Weatherspoon, D. (2020). Armed Conflicts and Food Insecurity: Evidence from Boko Haram's Attacks. American Journal of Agricultural Economics, 102(1), 114–131. https://doi.org/10.1093/ajae/aaz039

Goldstone, Jack A., Bates, Robert H., Epstein, David L., Gurr, Ted Robert, Lustik, Michael B., Marshall, Monty G., Ulfelder, Jay, and Woodward, Mark. (2010). A global model for forecasting political instability. American Journal of Political Science 54, no. 1

Meire, H. (2017). Machine learning and civil war: Investigating tree-based models for predicting intrastate violence [Masters in Science & Government Analytics] John Hopkins University

Muchlinski, D., Siroky, D., He, J., & Kocher, M. (2016). Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. Political Analysis, 24(1), 87-103. doi:10.1093/pan/mpv024

Newman, E. (2014). Understanding Civil Wars: Continuity and change in intrastate conflict (1st ed.). Routledge. https://doi.org/10.4324/9781315881584

Perry, C., (2013). Machine Learning and Conflict Prediction: A Use Case. Stability: International Journal of Security and Development, 2(3), p.Art. 56. DOI: http://doi.org/10.5334/sta.cr

Portmann, Felix T., Siebert, Stefan and Döll, Petra. (2010). MIRCA2000 – Global monthly irrigated and rainfed crop areas around the year 2000: A new high resolution data set for agricultural and hydrological modeling, Global Biogeochemical Cycles, 24, GB 1011, doi:10.1029/2008GB003435.

Python, A., Bender. A., Nandi, A. K., Hancock, P. A., Arambepola R., Brandsch, J., Lucas, T. C. D. (2021). Predicting non-state terrorism worldwide. Sci. Adv.7

Raleigh, Clionadh, Linke, Andrew, Hegre, Håvard, and Karlsen, Joakim. (2010). Introducing ACLED-Armed Conflict Location and Event Data. Journal of Peace Research 47(5) 651- 660.

Raleigh, Clionadh, Choi, Hyun Jin and Kniveton, Dominic (2015). The devil is in the details: an investigation of the relationships between conflict, food price, and climate across Africa. Global Environmental Change, 32. pp. 187-199. ISSN 0959-3780

Scikit-learn (No date). sklearn.ensemble.RandomForestRegressor [website]. Retrieved 11 January 2022 (https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, London. https://doi.org/10.1007/978-1-4899-3324-9

Tollefsen, A.F., Strand, H., Buhaug, H. (2012). PRIO-GRID: A unified spatial data structure. Journal of Peace Research. doi:10.1177/0022343311431287

Tollefsen, Andreas, Forø, Karim, Bahgat, Nordkvelle, Jonas, and Buhaug, Halvard (2015). PRIO-GRID v.2.0 Codebook. Peace Research Institute Oslo.

Vogt, Manuel, Bormann, Nils-Christian, Rüegger, Seraina, Cederman, Lars-Erik, Hunziker, Philipp, and Girardin, Luc. (2015). Integrating Data on Ethnicity, Geography, and Conflict: The Ethnic Power Relations Data Set Family. Journal of Conflict Resolution 59(7): 1327–42

Spatial Conflict Prediction with Machine Conflict Vulnerability in the Sahel Region

Frank Guzzardo

Masters Program in **Geospatial Technologies**

Supported by:
Education and Culture
ERASMUS MUNDUS